Leveraging association rule mining in travelers’ hotel selection preferences

A thesis submitted to Auckland
University of Technology in fulfilment
of the requirements for the degree of
Master of Philosophy (MPhil)

Xuan Zhu
2019

School of Computing and Mathematical Sciences

Primary Supervisor: Professor Stephen G. MacDonell
Secondary Supervisor: Dr. Paul Leong
Additional Supervisor: Associate Professor Russel Pears

Abstract

In seeking to explore and understand the hotel preferences of travelers, this study applies the association rule mining (ARM) method to the database of a hotel property management system (PMS). In particular, this study considers the associations between travelers' hotel selection behavior and demographic factors, travel types, and hotel attributes. Most of the prior research literature that has addressed travelers' hotel preferences is based on customers' responses to questionnaires. Potential deficiencies in such a research methodology include the following:

1. Questionnaire responses are based on a 'virtual' hotel booking scenario, where the potential customer ranks the importance of a list of possible influential hotel selection factors, and the resulting indications may not accurately reflect the *actual* purchasing propensity of the travelers;

2. Questionnaires typically include hotel selection factors such as "Hotel Staff Service Attitude" and "Hotel Facilities Cleanliness" which cannot be known by customers when they first book a hotel. Therefore, the conclusion of such studies cannot fully reflect travelers' preferences when booking a hotel for the first time.

This study explores customers' behavioral tendency by investigating the decisions reflected in hotel customers' actual purchase behavior as recorded in a hotel management software database (i.e., their reservations), and eliminates the 'unknowable' influencing factors of the first hotel booking choice.

This study screens, transforms, and preprocesses the original data in the PMS database, retaining and generating factors that have a potential impact on the customers' hotel selection behavior. *Lift* is used to filter the rules that are positively associated to the choice of hotels. The results show that the ARM method can effectively identify the rules of strong correlation and can be used to explain hotel customer behavior and – potentially – predict hotel booking trends. For instance, the analysis has revealed that travel purpose influences travelers' length of stay, and when traveling in pairs, customers are less price sensitive when selecting hotels. The study thus concludes that the factors considered are indeed influential in customers' hotel selection. The study further suggests the impact of excavated association rules on hotel marketing strategies and market segmentation strategies.

Table of Content

List of Tables

Attestation of Authorship

"I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of a university or other institution of higher learning, except where due acknowledgement is made in the acknowledgements."

Yours sincerely,

(Xuan Zhu)

Acknowledgements

To Professor Stephen MacDonell, Professor Russel Pears, Dr. Paul Leong, for their knowledge and patience during the time of this project.

To my beloved wife Elva, for her support and care to help me through hardship.

To grandma, wish her all the best and rest in peace.

Chapter 1: Introduction

1.1 Background to the research

Competition within the hotel industry in Hong Kong is fierce as new hotels and motels are joining the market every year. According to a government report published in February 2018 (statistics.gov.hk, 2018), the number of registered hotels in Hong Kong rose from 263 to 277 and the number of hotel rooms increased by 4,000 in the year of 2017 alone. Unofficial statistics shows that 24 new hotels with more than 6,000 rooms will be opened in 2019 (Hongkongextras.com, n.d.). A threat to the hotel industry also comes from the accelerating development of Airbnb and other lodging options. It is thus a pressing issue for hoteliers and hotel marketers to understand the requirements and preferences of hotel guests towards the selection of hotels and hotel services to develop specific and effective marketing strategies.

Attribute-centric research to measure consumer acceptance of particular marketing programs was introduced in 1977 (Martilla & James, 1977). The concept was soon applied in the hotel industry to analyze travelers' hotel selection factors. Lewis (1984) in his research sought to understand which elements from demographic factors, travelers' purpose, and hotel attributes, were determinant (those that actually cause a purchase), salient (those that are top of the mind but may not actually distinguish the hotel), or important to the guests in choosing hotels in the same city. Later studies (Ananth et al., 1992; McCleary, Weaver, & Hutchinson, 1993; McCleary, Weaver, & Lan, 1994; Chu & Choi, 2000; Choi & Chu, 2001; Lockyer, 2002; Chan & Wong, 2006; Tsai, Yeung, & Yim, 2011) identified multiple factors, such as age, gender, travel purpose, hotel security, cleanliness, location, price structure, and staff behavior, as having a significant impact on travelers' hotel-selecting behavior.

Various data analytic methods (Mccleary, Choi, & Weaver, 1998; Yavas & Babakus, 2005; Cobanoglu et al., 2003) have been applied to analyze responses collected from specifically designed questionnaire surveys. However, the methodology of using pre-designed questionnaires to collect travelers' responses towards certain hotel selection scenarios has been questioned, regarding its lack of consideration of the potential matters that will influence the hotel choice and implicit needs of hotel guests (Jones & Chen, 2011).

Moreover, some of the hotel attributes listed in such surveys are unlikely to be considered by prospective hotel guests during the reservation process under an actual hotel selection scenario, for instance, the enthusiasm of the hotel staff and the efficiency of the hotel front desk. Tsai et al. (2011) also pointed out that previous research had failed to distinguish “evaluative criteria” from “choice criteria”. Evaluation is the “decision-making” process before the purchasing behavior, when customers decide which factors are to be considered and which are more important. Choice is the outcome of “decision-making”, and reflects the actual purchase decision made by the customer (Tsai, Yeung, & Yim, 2011). Put another way, customers’ choices in actual hotel reservation scenarios can vary from their responses towards imaginary purchasing scenarios. In light of this, the research reported here focuses on actual purchasing scenarios, i.e. the “choice criteria”, to reveal patterns in travelers’ hotel selection decisions.

One of the most effective ways of studying customers “choice criteria” is to analyze the customers’ actual purchasing records, rather than their responses to hypothesized scenarios, as explained above. The development and popularization of computer technology enables hotels to shift from manually recording guest information and reservation details to software facilitation, namely by using Property Management Systems (PMS). The databases from hotel PMS provide us with a suitable data source of customers’ purchasing records. This research uses the reservation information retrieved from hotel Property Management Systems (PMS) to learn the actual purchasing behaviors of customers.

The overall objective of this research, then, is to explore the applicability of leveraging data mining methods, particularly association rule mining, in understanding travelers’ hotel selection preferences. Such an objective is important as it can provide hotels with a methodology to utilize readily available data (from PMS) to generate a hotel-specific solution to develop tailor-made marketing strategies for particular market segments. The trends and correlations revealed from association rules could help hotel marketers to adjust their market segmentation and marketing strategy on a more timely basis and create a competitive advantage.

The main research questions addressed in this research are:

1. What factors, which are potentially important for travelers' hotel selection decisions can be identified from the PMS database?
2. How does association rule mining support travelers' hotel selection preference analysis?
3. How could the trends and correlations revealed from association rules help hoteliers understand travelers' behaviors and react accordingly?

1.2 Motivations for the research

❖ Urgency of understanding hotel customer behavior

The recent popularity of online hotel booking platforms (e.g. booking.com; agoda.com) and travel review websites (e.g. tripadvisor.com; oyster.com) has made hotel reservation and price comparison much easier. Recommendations from online travel agents (OTA) and the immense volume of related user-generated information provide travelers with numerous options in making decisions for accommodation. Along with the increasing volume of hotel information available online, travelers' hotel preferences are changing quickly (Li et al., 2015). Hotels urgently need a method to understand emerging trends of customer behaviors to adjust their services and offers to meet the emerging requirements.

Questionnaire survey has traditionally been an easy-to-operate method to collect customer opinions and understand trends of their preferences, complemented recently by online-content mining as suggested by Chareyron et al. (2014) and Xiang et al. (2015). However, planning, distributing, and collecting questionnaires or crawling text information from websites requires large amounts of time and resources. A more effective and efficient option is to utilize the readily available customer behavior data in hotel property management systems.

❖ Necessity of creating a data mining method to process PMS data

The speed of data generation in the hotel sector can be prodigious; conventional methods of data processing can barely keep up with the speed of data generation – a powerful data mining method is therefore required. PMS covers operational functions from front desk, sales and planning, to accounting and reporting. The system handles numerous hotel operations including guest bookings, customer information, online reservation, posting charges and is capable of integrating or interfacing with third-party solutions such as a central reservation system, online booking engine, and a customer relationship management system. The database of a PMS is ready for access, free of extra cost, is up-to-date, and hotel-specific. If data mining methods can identify consumers' needs and expectations from the PMS database in a short period of time, the results could provide hoteliers with opportunities to prioritize tasks, better allocate their resources, and rapidly develop tailor-made marketing strategies for the target segments (Hsu et al., 1997). Once customers' requirements are clearly identified and understood, hoteliers are likely to be in a better position to anticipate and cater for their customers' desires and needs, rather than merely reacting to their dissatisfaction (Oberoi & Hales, 1990).

Many property management systems are capable of outputting reservation information in transaction databases that contain booking details and customer details. Association rule mining is one of the best-known data mining techniques to identify patterns in transaction databases such as these. Influencing factors that are identified from these data can simulate items in a *market basket analysis*, and correlations between these factors can help us discover hotel customers' purchasing habits.

During the review of prior research it was noted that rarely did studies use a PMS database as data source in terms of traveler hotel selection pattern discovery. This research is thus a first trial in addressing this gap and using such a new data source.

1.3 Outline of this thesis

Chapter 1 of this thesis introduces the background and explains the motivation for this research.

Chapter 2 provides a targeted review of prior research to identify the most studied hotel selection criteria and influencing factors, as well as data mining methods used in analyzing these factors. A comprehensive review of the application of association rule mining is also included in this chapter, to support the applicability of applying association rule mining in understanding travelers’ hotel selection patterns.

Chapter 3 explains research methodology and data mining tools employed in this research, and describes in detail the data retrieved from a candidate property management system.

Chapter 4 describes the data cleansing criteria and pre-processing steps to prepare the dataset for data mining.

Chapter 5 presents the execution of data mining along with its results. Important association rules are highlighted and interpreted in the hotel context in this chapter.

Chapter 6 reports a discussion on the outcomes of the study, suggests what might be of value in the findings, discusses possible future research directions and concludes the thesis.

Chapter 2: Literature Review

This chapter reviews previous research related to the current study's objectives. In keeping with the nature of an MPhil thesis this chapter aims to provide sufficient background to further establish the motivation for this study and to support the rationale of the hotel selection criteria examined, rather than a comprehensive review of the range of methodologies used to study hotel customers' purchasing behavior. Therefore, this chapter addresses the following:

1. Hotel selection criteria and their impact on hotel customers' lodging choice as examined in past studies, especially those relating to travelers' purpose and behavior, customer demographic factors and hotel attributes.

2. The application of the association rule mining method in studying customer purchasing behavior and customer preference profiling.

3. A brief summarization of the data mining tools that have been utilized to detect trends and discover behavior patterns of travelers in the hotel industry

2.1 Hotel selection influencing factors

Emphasizing a multi-attribute model for performance-importance analysis to understand the impact of attributes toward customer satisfaction and repetitive purchasing behavior was first proposed in 1977 (Martilla & James, 1977). The model has been leveraged in studying hotel selection factors since the 1980s. In 1984, Lewis (1984) initially proposed to relate consumer preferences and perceptions to their choice of hotels and market segments, to understand the nature of the hotel selection and decision-making process. Lewis (1984) research sought to find out which elements are salient, determinant and important for different hotels in the same city. The purpose of Lewis' study was to identify appropriate market strategies to maintain current customer loyalty and increase market share in the target segment. According to his findings, traveler's purpose proved to have a major impact on their preference towards hotel attributes, while demographic factors were important for salient factors.

Travelers' purpose and behavior

Travelers' purpose is often broadly divided into business travel and leisure travel in research on customer hotel preferences. The influence of travelers' purpose on hotel selection, especially hotel attributes selection, has been extensively studied. Security and price are repeatedly mentioned as the most influential factors for leisure travelers (Lewis, 1985; Parasuraman et al., 1988; Marshall, 1993; Chow, Garretson, & Kurtz, 1995), and leisure travelers are considered more price-sensitive than business travelers (Cobanoglu et al., 2003). Regarding preferences for hotel services, leisure travelers pay more attention to personal interaction and traditional hotel services. In contrast, for business travelers, cleanliness and location are the most important hotel attributes (Lewis & Chambers, 1989; Taninecz, 1990; McCleary et al., 1993), and they also value room and front desk attributes (Chu & Choi, 2000). A 2003 study mentions that technology amenities are also an important factor affecting hotel selection for Turkey's business travelers (Cobanoglu et al., 2003). The study of Kucukusta, Pang, and Chui (2013) also suggests that business travelers will pay more attention to unique or local-specific hotel attributes and services. Researchers also summarized some of the factors that are salient for both business and leisure travelers, such as service quality, value, food and recreation (Chu & Choi, 2000), as well as employee attitudes, location, and rooms (Yavas & Babakus, 2005).

Travelers' purpose can actually contain a finer classification than the simple division of business and leisure travelers. In Kaynak and Yavas' (1981) research, Canadian travelers were classified as vacationer, businessmen, and visitors of relatives; Cai's (2001) study, a comparison between three traveler groups, namely business, business and leisure, and leisure, was conducted. Travelers' purpose has been confirmed to have a significant impact on hotel guests' selection criteria. The hotel attributes valued by business travelers and leisure travelers are also significantly different. However, within the scope of the literature reviewed for this research, very few articles focus on another specific hotel guest group - daytime guests, also known as day-use guests. The concept of day-use rooms can be associated with the idea of renting hotel rooms on an hourly basis. The practice can be traced as far back as the concept of the hotel itself (Matthias, 2006). In Japan, renting hotel rooms by the hour is a common practice exclusively in a specific type of hotel called a love hotel, which provides short-term hotel room use for dating couples (Lin, 2008). Day-use rooms have also become a more common selection for frequent business travelers as it offers a more economical option for short-period resting for visitors arriving or departing in the early hours of the morning. An emerging trend of more hotels in Hong Kong starting to provide day-use services can be observed from the data retrieved for this study: all three subject hotels are frequently providing day-use rooms for guests; one of the hotels has dedicated rooms exclusively reserved for day-use guests. Therefore, in this study, day-use is treated as a stand-alone travel purpose to be studied side-by-side with business and leisure.

In the research conducted by Li et al. (2013), “travel type” was suggested to be an important behavioral pattern for developing customers’ preference profiles. Compiling preference profiles for travelers was believed to be crucial for hotel management to identify target customers and design suitable products (Li et al., 2013); however, the factors influencing customers or customers’ profiles are rarely studied (Tanford, Raab, & Kim, 2012). This research provides an opportunity to examine the influence of travel type, among many other factors, on travelers’ hotel choice preferences to fill this gap in knowledge. The “travel type”, referred to as “travel model” in some research, generally describes the nature of companion(s) travel with hotel visitors (Yoo, McKercher, & Mena, 2004). For example, Li et al. (2013) divided hotel visitors into three travel type groups: business, couple, and family. In this study, the categorization is based on the number of adults and children involved in a single transaction. Detailed travel type grouping is explained in Chapter 4.

Lockyer (2005b) mentioned in his study that time is one of the trigger points that can influence a purchase choice. Godinho et al. (2016) also discovered that time pressure can have a “surprising impact” on consumer behavior under certain circumstances in actual purchase environment. In the hotel industry, booking time has been an important concern for revenue management, especially advance booking and last-minute booking (Schwartz, 2006; Jang, Chen, & Miao, 2019). To evaluate the impact of booking time on travelers’ lodging options, the time between reservation date and check-in date is collected from the data source and studied here as an influencing factor.

Demographic factors

Demographic factors are commonly an important indicator of marketing segmentation. Almost all articles reviewed for this thesis that study customer behavior and hotel selection criteria contain an analysis of demographic factors. Among them, gender, nationality, and age are the most commonly studied.

McCleary, Weaver and Lan (1994) observed significant differences between male and female travelers when studying business travelers' lodging preference. In that research, women considered security facilities, room services, and low price, while men paid more attention to office amenities, office space, and availability of suite rooms. The difference in behavioral patterns between male and female business travelers was also significant; female business travelers travelled less frequently than men, but the length of their international travel is usually higher. Lockyer (2002), in similar research conducted on New Zealand business travelers, validated women's prior attention to security in their hotel selection. Male business guests, according to Lockyer's findings, perceived availability of lounge, bar, and restaurant facilities as more important influencing factors. It is worth noting that there are also discussions on the price sensitivity of men and women in both studies conducted by McCleary et al. (1994) and Lockyer (2002), but no definitive conclusions have been drawn. McCleary et al. believed that, in his study, women were more concerned about low prices than men because the average income of female respondents in the study was lower; in the study of Lockyer, males and females did not show a clear difference in price sensitivity because of the lack of respondents' income information.

Travelers of different nationalities also demonstrate different preferences in hotel selection criteria. A study that compared business travelers in the United States and South Korea showed that although business travelers in both countries considered cleanliness to be the most important hotel selection factor, American travelers gave greater consideration to safety and security, while Korean travelers paid more attention to hotel staff friendliness (McCleary, Choi, & Weaver, 1998). In addition, the study also indicated that attitudes towards the importance of hotel locations between the Korean travelers and US travelers varied. In another article comparing hotel preferences for Asian and Western tourists, researchers found that Asian travelers tend to choose hotels that can offer room facilities that have previously provided a satisfying experience; for Western tourists, the reputation of the hotel and the consistency of hotel services are more important influencing factors (Chan & Wong, 2006). In studying attitudes towards price and value, Gilbert and Tsao (2000) suggested that Chinese travelers are more sensitive to price than Western travelers. Tsai, Yeung, and Yim (2011) confirmed their finding and argued that Chinese travelers visiting Hong Kong paid high attention to the location of the hotel, especially the convenience of the hotel to tourist attractions and MTR underground railway stations. For Western travelers in Hong Kong, cleanliness of the room is the most important hotel attribute.

Lepisto and McCleary (1988) stated that using traveler age by itself as a segmentation dimension is not appropriate because no correlation was observed between the preferences for a particular type of hotel and the age group in their study. However, age has still been included in many studies on hotel selection criteria (Lewis, 1985; McCleary, Weaver, & Hutchinson, 1993; Chow, Garretson, & Kurtz, 1995; McCleary, Choi, & Weaver, 1998; Callan, 1998; Chu & Choi, 2000; Cobanoglu et al., 2003; Yavas & Babakus, 2005; Chan & Wong, 2006; Lockyer, 2005a; Lockyer, 2005b; Tsai, Yeung, & Yim, 2011; Sohrabi et al., 2012; Kucukusta, Pang, & Chui, 2013; Baber & Kaurav, 2015). However, while Ananth et al. (1992) discovered that important hotel selection factors were differentiated by age group, no similar conclusions were drawn to support or association between age groups and hotel preferences. Ananth et al. (1992) suggested in their research that elderly travelers pay extra attention to the hotel's catering services and medical services; and most importantly, the expending habits of mature travelers are likely to be more affluent. In this study, age is used as a research element in discovering travelers' hotel selection patterns, to validate providing an opportunity conclusions suggested by Ananth et al.

Hotel attributes

Hotel attributes, also referred to as hotel selection attributes, are the most often considered criteria in previous studies that have sought to understand travelers' hotel selection patterns. Hotel attributes are typically divided into different categories, which are called hotel selection factors. In previous studies, researchers frequently listed a large number of hotel attributes in questionnaires to collect importance ratings from hotel customers. Callan (1994; 1998) used 166 attributes in his research; Lewis (1987), Ramanathan & Ramanathan (2011) used 84; Wilensky & Buttle (1988) used 49. The use of large numbers of attributes has been criticized as inappropriate (Jones & Chen, 2011) because research in other fields has shown that the human brain simplifies the decision-making process when making complex choices (Belonax & Mittelstaedt, 1978), and consumer purchasing decisions are usually based on a smaller number of determinants (Myers & Alpert, 1968). Moreover, the common use of a post-purchase data in previous studies has led to confusion between pre-purchasing attributes (that can be learned before purchasing) and post-purchasing attributes (only discernible after purchasing). Therefore the respondents could not correctly distinguish between hotel selection and repeat purchase, which ultimately led to failure in identifying real determinant factors in hotel selection (Jones & Chen, 2011). Many of the hotel attributes listed as important, such as "room cleanliness", "friendliness of hotel staff", and "comfort of bed" are not perceptible by hotel guests until a hotel is selected or even checked in.

A set of 49 attributes selected from Callan's 166 attributes list and Lewis 84 attributes list are presented in table 1. The selected attributes cover the majority of the most frequently studied attributes in the reviewed articles and so are used here to distinguish pre-purchasing attributes from post-purchasing attributes based on information availability from general information channels such as hotel homepage, booking websites, and Google maps. The pre-purchasing attributes identified from the table are examined for availability in PMS databases (the primary data source of this thesis), informing whether they are to be studied in this particular research.

Table 1: Analysis of hotel selection factors and attributes as pre-purchasing or post-purchasing attributes.

*Factors/ Attributes	**Pre-purchasing attribute**	**Post-purchasing attribute**	**Studied in this research**	**Reason**
*Location and Image				
Location	√		**Yes**	
Good reputation	√		No	The subject hotels chosen for this research are individual hotels newly opened within two years. The reputation has not been built up. Moreover, reputation is an attribute not available from PMS database
Prestige of property/ chain	√		No	Same as "good reputation"
Historical, traditionality of hotel	√		No	Same as "good reputation"
Exterior aesthetics	√		No	Though exterior aesthetics is available from the images on hotel website and Google street view, it is not available from PMS database
Public room aesthetics		√	N/A	
*Price/ Value				
Actual price	√		**Yes**	
Price/ value		√	N/A	
Restaurant/bar price varieties		√	N/A	
*Service				
Promptness of all services		√	N/A	
Variety of services offered		√	N/A	
Professionalism of staff		√	N/A	
Quick check-in/check-out		√	N/A	
Staff friendliness		√	N/A	
Rooms made up promptly		√	N/A	
Reservation system convenience	√		No	Attribute information not available from PMS database
Management attention		√	N/A	
*Access				
Nearness to safe parking		√		
Availability throughout the year	√		**Yes**	Availability of hotel throughout the year especially during holidays can be important for travelers' planning of visits
Convenience to public transport		√	N/A	
*Security				
Security of hotel		√		
Security of area	√		**Yes**	The security condition of hotel is hardly available prior to arrival, but the security of surrounding neighborhood can be obtained through simple research. This attribute can be treated as an element of attribute "location"

***Additional Services**				
VIP rooms, sections		√	N/A	
VIP treatment		√	N/A	
Eating/ drinking options		√	N/A	
Late night food service available		√	N/A	
Elegant dining available		√	N/A	
Business priority check in		√	N/A	
Shops in hotel		√	N/A	
***Rooms**				
Cleanliness of the room		√	N/A	
Comfort of bed		√	N/A	
Size of room/ bed	√		**Yes**	The size of room/ bed can be interpreted as the "room type" in hospitality term. This is an attribute that can be easily obtained from hotel PMS database
Quality of TV/ radio		√	N/A	
Decor, furnishings of room/bath		√	N/A	
Room service availability		√	N/A	
Quietness of room		√	N/A	
Quietness of hotel		√	N/A	
Physical condition of rooms/baths		√	N/A	
Cable TV	√		No	Cable TV, Wi-Fi access, air conditioner, and in-room telephone are all standard facilities in hotel rooms today. The facility information is generally available on booking websites, however, this particular attribute has limited impact on hotel travelers' choice of hotel nowadays.
Yellow pages available		√	N/A	
Small amenities, e.g., mints, soaps		√	N/A	
***Leisure Facilities**				
Year around pool	√		No	Attribute information not available from PMS database
Sauna, steam bath, exercise	√		No	Attribute information not available from PMS database
Eating and drinking facilities	√		No	Attribute information not available from PMS database
Restaurant food quality		√	N/A	
Restaurant service quality		√	N/A	
Quality of drinks		√	N/A	
Quality of wine list		√	N/A	
Nightlife, entertainment		√	N/A	

Security has been the most frequently cited hotel selection factor considered important in previous studies (McCleary, Weaver, & Lan, 1994; Chow, Garretson, & Kurtz, 1995; Mccleary, Choi, & Weaver, 1998; Chu & Choi, 2000; Choi & Chu, 2001; Lockyer, 2002;

Tsai, Yeung, & Yim, 2011), followed by locations and hotel services (McCleary, Weaver, & Hutchinson, 1993; Chu & Choi, 2000; Choi & Chu, 2001; Chan & Wong, 2006; Lockyer, 2005a; Lockyer, 2005b; Zaman, Botti, & Thanh, 2016). From the listing in Table 1, it can be noted that the hotel security attributes are unperceivable before choosing a hotel for first visits. Attributes that hotel visitors assign high value to, such as the number and location of surveillance cameras, the arrangement of security personnel, the availability of private safes and vaults in the room, and the confidentiality of the personal information of the guests, are considered post-purchasing attributes. However, the security level of the area where the hotel is located can be learned through researching local news and police websites. Some other attributes, such as the ease of transportation, and distance from shopping and tourist attractions, are frequently part of the hotel's location information. Therefore, in this study, three hotels of distinctive and representative locations are selected to examine the impact of the location on hotel selection.

Hotel service, another commonly mentioned key selection factor, generally cannot be directly experienced prior to the hotel guests' stay in the hotel. However, due to the large number of traveler reviews obtainable from travel websites such as Tripadvisor, travelers can be well informed of the service quality of the hotel indirectly. Limited by the data sources for this study, being the PMS database this particular hotel selection factor will not appear as a primary research subject in this thesis.

In perusing Table 1 for important attributes that are obtainable before hotel reservation and available in the PMS database, the following hotel attributes were selected as the focus points of this study, to understand travelers’ preference for hotel selection: "location", "actual room price", "period of year", and "room type".

2.2 Association rule mining and other data mining methods

The concept of association rule mining was originally proposed by Hájek et al. in 1966 but was not widely applied in data mining. The popularity of association rule mining began with Agrawal's publication in 1993 (Agrawal, 1993). In early studies, association rule mining was used to explore the relationship between sales volume of items in the retail industry, which is often referred to as *market basket analysis*. A rule indicates the association between two sets of items X and Y in the form of "X ⇒ Y", where X is referred to as the antecedent and Y is referred to as the consequent. In the Agrawal study, five categories of the rules were perceived particularly important: 1. all rules that have "item A" as consequent, to find out which items are often purchased with and can increase the sales of "item A"; 2. all rules that have "item B" as antecedent, to find out when "item B" is missing from the aisle which items' sales will be affected; 3. all rules that have "item C" in antecedent and "item D" as consequent, to understand which items bought together with "item C" can increase the chance of purchasing "item D"; 4. all rules that contain items from aisle X and aisle Y, to help supermarkets and other retailers rearrange the aisles; 5. rules that suggest the best items influencing the sales of "item E" by controlling support and confidence of the rules. Association rule mining was introduced as an approach to find underlying correlations from large transaction databases and study consumer buying habits to support targeted sales and marketing programs.

Association rule mining is the discovery of association rules that satisfy a predefined minimum support and confidence for a given database (see section 3.1.1). Usage of support and confidence as thresholds to constrain rules was first introduced by Webb in 1989. The mining of association rules usually involves two steps. The first step is to find the set of items in the database that exceed the predetermined threshold (*support*); these items are called frequent or large item sets. The second step is to generate association rules from those frequent item sets with minimal *confidence* constraints. Assuming that one of the frequent item sets is I_k, $I_k = \{i_1, i_2,..., i_k\}$, one of the rules that can be generated from this frequent item set is $\{i_1,i_2,...,i_{k-1}\} \Rightarrow \{i_k\}$. The rule can be determined to be interesting or not interesting by checking the confidence. The next rule is generated by transferring the last item in the antecedent (left-hand side of the rule) to the consequent (right-hand side) and check the confidence for interestingness. The processes are iterated until the antecedent becomes empty (Kotsiantis & Kanellopoulos, 2006).

In many cases, the number of association rules generated by the algorithm is large, often in thousands or even millions. Furthermore, the association rules themselves are sometimes large and complicated. Therefore, it can be challenging for the end user to understand or verify such a large number of complex association rules, thereby limiting the usefulness of the data mining results. Strategies such as generating only "interesting" rules, removing "redundant" rules, or only generating those rules that satisfy certain other criteria (such as coverage, leverage, lift, or strength) are therefore used to limit the number of associated rules generated (Kotsiantis & Kanellopoulos, 2006).

In addition to being used in market basket analysis, association rules are used in many other domains and contests. For instance, the application of association rule mining includes stock analysis, Web log mining, medical diagnosis, bioinformatics, and as involved in this study, customer behavior analysis (Solanki & Patel, 2015).

In previous studies, association rule mining was found to be widely used in research related to customer relationship management, especially for understanding and predicting customer behavior. Prior studies have found that association rule mining is mainly used for customer retention and new customer development, especially one-to-one marketing and market basket analysis (Ngai, Xiu, & Chau, 2009). Association rule mining tools are used to explore interesting associations that are ‘hidden’ in the database, above a certain threshold (Wang et al., 2005): these thresholds suggest the strength of the customer's behavior patterns and the likelihood that the rules will reoccur (Berson et al., 2000). Selected association rules can be used to build models that predict future customer value (Wang et al., 2005).

In Ng & Liu's study (2000), association rules, feature selection and deviation analysis were integrated to develop a novel approach to measure customer loyalty and predict the likelihood of defection. The method identifies the “early signs” of client defection and actions triggered by these warnings were found to usually play an important role in the final retention of the customer. The same approach can be employed to address similar issues in sales and service-related industries. This early success in preventing client defection indicated that the maturity of association rule mining had reached the point where it was desirable and feasible to apply large-scale applications to practical problems (Ng & Liu, 2000).

A similar application of association rule mining in another study explores hotel industry data by using association rules to discover connections between records to answer questions such as "why does the length of stay of customers increase after a particular promotion?" or "why is a particular promotion more effective for a particular customer segment?" (Magnini, Honeycutt, & Hodge, 2003).

Au & Chan (2003) introduced a new algorithm called Fuzzy Association Rule Mining II (FARM II) when mining association rules from a bank account database to discover previously unnoticed customer loan habits, such as loan balance preferences for customers of different ages and marital status. In another study that also examined bank customer behavior habits, Hsieh (2004) used self-organizing map neural networks to establish a behavioral scoring model and classify bank customers into three different profit groups. The customer analysis is then completed by using the Apriori association rule inducer, and the behavior patterns of different profit groups are analyzed to inform marketing strategy.

There are many examples of using association rules to build consumer behavior models and estimate consumer behavior patterns. Casillas, Martinez-Lopez, & Martinez (2004) used fuzzy association rule mining to discover relationships between different variables to deduce patterns in examined data and then use these patterns to estimate consumer behavior. The model infers the consumer's overall opinion on a certain advertisement through the consumer's cognitive evaluation of the advertisement and overall opinion about the advertisement. A similar approach has been used to study marketing knowledge patterns and rules in electronic catalog marketing and sales management at retail malls in Taiwan (Liao & Chen, 2004). One of the ultimate goals of developing knowledge in customer behavior patterns was to help managers build better marketing strategies to increase profits. A simulated model to demonstrate the use of mined customers' behavior patterns in direct marketing introduced by Wang et al. (2005) indicated that the new approach using association rules to predict the value of future guests increases the projected average profit significantly.

In addition to exploring customer preferences and behavioral patterns at a single point in time, Song et al. (2001) developed a method to detect changes in customer behavior at different time snapshots. Customer behavior rules were generated by placing customer profile information on the left-hand side (antecedent) and customer behavior information on the right-hand side (consequent). Customer behavior changes can be detected by comparing the customer behavior rules generated at time *t* and *t+k.* A similar approach was applied in the study conducted by Wu, Chen, & Chen (2005) in understanding customer relationship management for credit card business. In their study, the association rule mining tool Weka was used to measure changes in association rules, namely to discover emerging pattern rules, unexpected change rules and added rules.

Data mining tools can effectively identify patterns in customer spending and trends in behavioral changes, which can enable management to detect potential changes in customer preferences in large databases and provide customers with the products and services they need to slow or prevent defection. The approach of mining changes in customer behavior patterns was also used in retail marketing. Chen, Chiu, & Chang (2005) integrated customer behavioral variables, demographic variables, and transactional databases in their study of retail customer behavior, and designed two extended measures for similarity and unexpectedness to analyze the degree of resemblance between patterns at different time periods.

Association rule mining has proven to be able to help develop better marketing strategies and increase profits by mining customer knowledge, building customer profiles, and predicting customer behavior in retail, banking, online shopping and other industries. However, in the scope of the literature reviewed in this thesis, no attempts have been conducted in understanding customers' hotel selection patterns by mining association rules from the hotel guest's reservation information. This study therefore explores the underlying correlations of hotel guests' booking behavior using association rule mining methods, thereby discovering their preferences in the hotel selection process and establishing a hotel customer preference profile.

Chapter 3: Research Methodology and Data Acquisition

This chapter describes and justifies the research methodology and data mining tools adopted in this study. There then follows a description of the form and nature of the data source, with justification.

3.1 Research Methodology

Ha & Park (1998) summarized the procedures followed by research in data mining and suggested that the following nine steps are generally taken by most studies:

1) Establishing an understanding of the application domain, the end user's goals and previous research.
2) Creating, selecting a target dataset, or focusing on a subset of variables or data samples for analysis.
3) Data cleansing and preprocessing, involving clearing noise and outliers, handling missing data fields, and transforming data into forms easier to process.
4) Data conversion and reduction by using features to represent data based on task objectives, or by using dimensionality reduction or transformation methods to reduce the effective number of variables in valid data.
5) Determining the data mining tasks and goals by answering questions such as: Whether is the mining verification-driven or discovery driven? What is the primary goal of the mining: finding association rules, clustering, classification, sequencing or forecasting?
6) Selecting the corresponding data mining algorithms needed to explore patterns in the data or fitting models to the data.
7) Searching for interesting patterns in a particular form, or alternatively, in a set of such representation.
8) Attempting to interpret patterns found in step 7 or return to any previous step for further iteration.
9) In the final step, the discovered knowledge is incorporated into the performance system for testing, or simply integrated and reported to interested parties.

Ha & Park's nine steps provide broad guidance for data mining in general. In this particular research, the core aim is to explore the use of a discovery-driven approach to find interesting rules in hotel PMS databases, and to create hotel customer preference profiles. Reliability of the profiles will be validated by comparing with prior knowledge. Finally, interpretation of particularly interesting patterns retrieved from the profiles will be attempted in the actual operating context of hotel selection and marketing.

3.1.1 Association rule mining

Association rule mining is a data mining method that is widely used to find correlations between item sets in transaction data. In other areas of research, association rule mining is mainly used to learn user behavior, understand user preferences, and assist in market segmentation, market development, and product portfolio planning (Ngai, Xiu, & Chau, 2009). According to the original definition of association rules by Agrawal et al. (1993), the set of all considered items contained in the transaction database is called *the items*, or *item base*. Any subset of the item base is called an *item set*. In association rule mining, a *rule* is defined as the association between two item sets X and Y of item base I in the form of

$$\mathrm{X} \Rightarrow \mathrm{Y}, \text{where } \mathrm{X}, \mathrm{Y} \ \subseteq \mathrm{I}$$

Where X is referred to as the antecedent, or Left-Hand-Side (LHS) and Y is referred to as the consequent, or Right-Hand Side (RHS). To evaluate the quality of association rules and to find interesting rules from all possible rules, concepts such as *support*, *confidence*, and *lift* are introduced as measures of significance and interest (Agrawal & Srikant, 1994; Agrawal, Imieliński, & Swami, 1993).

Support

Support is an indication of how frequently a particular item set appears in the dataset. The support of the item set X is defined as the proportion of the transactions *t* containing the item set X in the transaction set T.

$$\mathrm{Supp}\ (\mathrm{X}) = \frac{|\{\mathrm{t} \in \mathrm{T}; \mathrm{X} \subseteq t\}|}{|\mathrm{T}|}$$

In association rule mining, in order to ensure that the rules obtained are statistically significant, researchers typically set a minimum support as a threshold.

Confidence

Confidence is used to indicate the extent to which a rule has been found to be true. For the item set X and Y in the transaction set T, the confidence of rule $X \Rightarrow Y$ expresses the occurrence ratio of Y in a transaction which contains X.

$$\mathrm{Conf}\,(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X)}$$

In association rule learning, minimum support and confidence are the most commonly used constraint thresholds. However, high confidence does not necessarily suggest strong correlation between X and Y in the same transaction. Assume that the support of two **independent** item sets, X and Y, are both 0.90, the confidence of rule $X \Rightarrow Y$ is as high as 0.81. Obviously, no useful rules can be extracted between two independent subsets, therefore, to correctly understand the correlation between the item set X and Y, or the influence of the occurrence of the LHS on that of the RHS, the measure *lift* was introduced.

Lift

Lift of a rule is defined to measure the ratio between the observed $\mathrm{Supp}\,(X \Rightarrow Y)$ and the support that assumes X and Y are independent.

$$\mathrm{Lift}\,(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X) \times \mathrm{supp}\,(Y)}$$

If lift = 1, it would imply that X and Y are independent, the rule between them is neither important nor interesting; if lift > 1, it means that the occurrence of Y is dependent on X, and the larger the lift value, the greater the relevance. This makes the rules between X and Y potentially useful to predict the occurrence of RHS in future datasets. For example, a rule with $\mathrm{Lift}\,(X \Rightarrow Y) = 1.50$ means that the frequency of Y occurring in a single transaction containing X is 50% higher than the frequency at which the occurrence of X and Y are completely random. Conversely, if lift < 1, it means that the occurrence of X has a negative effect on that of Y.

In this study, attributes that can potentially influence the purchasing choices of hotel guests are extracted from the transaction database of hotel property management system as “items”. All possible association rules are mined from the dataset and filtered according to the extent to which antecedents influence the occurrence of consequents. Therefore, lift will become a significant indicator for rules screening.

3.1.2 Apriori algorithm

Apriori is an algorithm developed by Agrawal & Srikant (1994) for frequent item set mining and association rules learning. It has been recognized as the most commonly used algorithm in association rule mining due to its efficiency in handling large database (Abaya, 2012). Apriori algorithm is based on a fundamental principle, which states "all non empty subsets of a frequent item set must be frequent" (Agrawal, Imieliński, & Swami, 1993); or in other words, the support of an item set never exceeds the support of its own subset, which is also known as the anti-monotype property of support.

Apriori algorithm first generates a list of all singleton item sets with sufficient support, then builds the next level of candidate frequent item sets with combinations of these singleton item sets until the optimum length of the frequent item set is reached. Apriori scans the database multiple times during the process of generating the candidate item sets and so can generate a large number of subsets potentially leading to a significant memory requirement (Heaton, 2016). However, in comparison with other common association rule mining algorithms such as Eclat and FP-growth, although Apriori is considered to have "scalability issues and exhausts available memory much faster than Eclat and FP-Growth" (Heaton, 2016), it appears to be more effective in finding association rules for large pattern datasets (Kavitha & Selvi, 2016).

Apriori's shortcomings can be overcome by reducing the number of effective variables with transformational methods or using features to represent data to reduce the number of candidate item sets generated, and ultimately reduce the need for memory. At the same time, since the concept of Apriori is relatively easy to understand, it is a popular choice as the starting point of frequent item set study (Heaton, 2016). Considering that the database to be processed in this study contains over 100,000 transactions, Apriori was selected to ensure the efficiency and effectiveness of the data mining.

3.2 Data acquisition

The core function of a hotel's property management system (PMS) is to record and manage room availability, hotel reservations, transactions, customer information and to interface to third party systems such as door lock controls, telephone exchanges, accounting systems, distribution systems and customer relationship management (CRM) systems. There are numerous types of PMS that are diverse in database structure and operating platform, employed by different hotels. Differentiation in software functional settings and hotel operational requirements leads to variation in the format and availability of attributes and variables in a PMS database; however, critical reservation and customer information remain universal across the industry e.g. check-in/-out date, number of guests, room type, room rate, nationality of guest. In this study, data mining is only performed on such critical variables, universal to all PMS systems, as this means the methodology of this study can be applied to data sources from various PMS systems.

The hotel databases used in this study are provided by an anonymous PMS vendor in Hong Kong. The databases were extracted from the live PMS systems of three newly opened individual hotels in Hong Kong after careful selection to ensure the three hotels have similar rankings (3 to 4 stars), similar market positioning (mid-end boutique hotels), and the same price range. The intention behind these screening criteria is to ensure that the target customer groups of hotel A, B and C are as similar as possible. Therefore, the impact of factors that are not in the scope of this study, such as hotel star rating, reputation, and market positioning are eliminated to the maximum extent when connecting customer behavior patterns and their choice of hotels. The database has been pre-processed before being made available for this research by removing hotel customers' private information including their name, travel ID, contact information, and address. The raw data obtained from the PMS vendor is in a format of a transaction table with 18 attributes. A sample of the raw data can be viewed in Appendix 1.

In the second chapter of this thesis the potential importance of hotel location as a factor in customers' hotel choice was described. The location of the hotel is associated with ease of transportation, the security of the surrounding area, and the nearby attractions that can potentially influence the purchasing behavior of hotel customers with different travel purposes. Different areas of Hong Kong are known for their distinctive features, buildings, and lifestyles preserved from different historical periods. *The Outlying Islands* retain the cultural heritage before the British colonial period of Hong Kong. Residents on the islands still maintain the traditional fishing-centered lifestyle. Historical sites such as the Mazu (Chinese Sea Goddess) Temples have become famous tourist attractions (Discoverhongkong.com, n.d.). The transportation condition of the outlying islands is not ideal; a limited number of ferries that can be easily affected by weather connect the outlying islands with other parts of Hong Kong. Modern buildings and shopping centers are rarely seen on the islands. The Kowloon area centered on *Yau Ma Tei* is one of the most densely populated areas in Hong Kong and retains a large number of historic buildings from the period of the British Colonial Government. The area around Yau Ma Tei is known for its distinctive flea market and electronic merchandising malls. In *Tsim Sha Tsui* there are a large number of contemporary buildings, bars, modern tourist attractions and luxury shopping malls. Connected by the narrow water of Victoria Harbor, Tsim Sha Tsui is less than one kilometer away from the heart of the world's third largest financial centre, Central.

In this study, then, three hotels A, B and C from the representative areas of Yau Ma Tei, Tsim Sha Tsui and Outlying Islands were selected to explore the impact of hotel location on travelers' accommodation options. Hotel A is located in Yau Ma Tei and the data retrieved is between October 7, 2015 and March 12, 2018. Hotel B is located in Tsim Sha Tsui and the data retrieved is from September 27, 2015 to May 23, 2018. Hotel C is located in one of Hong Kong's outlying islands and the data retrieved is from October 25, 2016 to May 23, 2018. All three hotels were opened between June 2015 and September 2016.

The dataset extracted from the hotel PMS is in the form of transactions. The PMS supports the registration and management of multiple guest profiles under each reservation. Each transaction in the dataset consists of the customer information and reservation information of one registered guest. In other words, one reservation with *three* registered guests in the PMS contributes *three* transactions in the dataset. Data from three hotels are consolidated together, a new attribute “Hotel Code” is added to each transaction to indicate the origin. Table 2 lists the 20 attributes contained in each transaction from the consolidated dataset. For each attribute, a description of its nature, functionality, and data type are presented.

Table 2: list of attributes from the original PMS database

Attributes	**Code as in raw data**	**Description of attributes**	**Data type and range**	**Mandatory field in PMS**	**Manual input/selection at reservation**
Hotel Code	-	Attribute to indicate the origin of the transaction	A, B or C	-	-
Stay ID	STAYID	Stay ID is a code assigned by the system automatically to mark every booking operation regardless of whether the booking is confirmed or not.	Integer, auto-number	Yes	No
Confirmation ID	CONFIRMATION	Confirmation ID is generated by the system simultaneously when the booking is successfully confirmed. *Transactions with same confirmation ID indicate that guests involved in these transactions are registered in the same reservation.*	Integer, auto-number	Yes	No
Travel agent name	TA NAME	Travel agent name is selected from a preset list of travel agents or online travel agents in the system. The list of travel agents is maintained by the sales department and input at the reservation by booking agent.	Item from preset list	No	Yes
Reservation status	STATUS	The status of the reservation implies the booking status according to system record at the moment of data retrieval.	One value from the following: C - reservation cancelled; I - inhouse, customer currently stays in the hotel; R - reserved, booking confirmed and customer yet to check in; O - customer checked out.	Yes	No
Check-in date	CHECK-IN DATE	The date that customer checked in or plans to check in	Date in format yyyy-mm-dd	Yes	Yes

Check-out date	CHECK-OUT DATE	The date that customer checked out or plans to check out	Date in format yyyy-mm-dd	Yes	Yes
Booking date	BOOKING DATE	The date that the booking is made	Date in format yyyy-mm-dd	Yes	No
Room type	ROOM TYPE	Codes that represent rooms with specific features e.g. room size, bed size, views, etc. The PMS vendor is not allowed to disclose detailed descriptions of room type codes for confidentiality reasons, therefore, the code is used to distinguish between difference room types in the raw data.	Item from preset list	Yes	Yes
Room rate code	RATE CODE	The rate code can be understood as a formula for calculating the room price based on variables such as room type, promotion, booking time, length of stay, etc. By setting the rate code, the hotels are enabled to sell room-nights at various prices in accordance to situation, and thus achieving the price variation strategy.	Item from preset list	Yes	Yes
Average daily price	PRICE	Average daily price of the booking	Numeric value calculated by rate code	Yes	No
Booking source	BOOKING SOURCE	Booking source generally marks the channel through which the reservation is made: hotel websites? walking up to the front desk? or booked by company? Booking source codes are customized and defined according to the requirement of each hotel. The categorization and definition of booking source codes in Hotel A, B, and C resembles each other which makes data processing simpler for this study as demonstrated in next section.	Item from preset list	Yes	Yes
Number of adults	ADULTS	Number of adults is provided by the customer or travel agent at the time of booking.	Integer	Yes	Yes

Number of children	CHILD	Number of children is provided by the customer or travel agent at the time of booking.	Integer	Yes	Yes
Guest ID (only available in Hotel A and B)	GUEST ID	The PMS allows the creation of profiles for hotel guests. The profiles can be either *temporal* (without Guest ID) or *permanent* (with Guest ID). Repeated bookings from the same guest can be linked to the *permanent profile*, provided the necessary information to identify the guest is given at the time of booking.	Integer, auto-number	No	Yes
Share code	SHARE CODE	The share code system is closely linked to the guest profiles and billing logic of the PMS system. Hotel generally requires front desk to register all guests involved in each reservation. Multiple guest profiles can be created under each booking. One guest will be registered as Master guest (share code M) and primary bill taker, the other registered guests are simultaneously marked as Phone guests (share code P) and secondary bill takers. Automatic room rate posting, in-hotel expenses posting and other billing functions linked to these guest profiles are achieved through the share code system.	One value from the following:M - master guest;P - phone guest.	Yes	No

Gender	GENDER	Each registered guest can choose to register their gender as M (male), F (female), or U (Unspecified).	One value from the following:
			M - male;
			F - female;
			U - unspecified.
				Yes	Yes

Nationality	NATIONALITY	Marks the nationality of the guest provided the information is given.	Item from list of all nations	No	Yes
Date of birth	DATE OF BIRTH	Marks the guest's birthday provided the information is given	Date in format yyyy-mm-dd	No	Yes

Company ID (Only available in Hotel B and C)	COMPANY ID	Similar to Guest ID, PMS allows the creation of profiles for companies that have contract with the hotel. Bookings from these companies are usually for their staffs or members. These companies are not Travel Agents or Online Travel Agents, but possibly be group-buying agents. The company names and nature are not disclosed in the data set due to confidentiality reasons.	Integer, auto-number	No	Yes

Large amount of information can be withdrawn from this raw data obtained from PMS, however, its value for understanding customers' hotel choices is limited in its original format. Some information in the raw dataset requires pruning to reduce the redundancy, some other information can be converted to hotel or customer attributes that hold more value for this research. For example, "date of birth" value can be translated into the age of the customers when they check in, an important demographic factor for understanding customer behavior. The data cleansing, conversion, integration, and other data preparation processes are described in the next chapter.

Chapter 4: Data cleansing and data processing

As reported in this chapter the data is cleansed and further processed to remove or minimize possible errors and to increase the data quality for mining. Original attributes from raw datasets are transformed to meet the format requirement of Apriori. Additional attributes that are potentially important to answer the research questions are generated from existing information, with justification.

4.1 Data cleansing

A variety of factors can affect the quality of datasets collected in the real world, with data entry mistakes one of the most significant sources of errors (Maletic & Marcus, 2009). Studies have shown that as much as 40% of collected data are 'dirty' in one way or another (Fayyad et al., 2003). If data is not cleaned and corrected, the usefulness of data mining and data warehousing based on them will be reduced significantly. Therefore, data cleansing is an indispensable part of data acquisition (Müller & Freytag, 2005) and one of the important prerequisites for successful database knowledge discovery (KDD) (Maletic & Marcus, 2009).

Only high quality data is suitable for processing and analysis. Müller & Freytag (2005) suggested a set of criteria to examine the quality of a set of data and determine whether it is processable and interpretable. These criteria include:

- Validity

  Validity measures the extent to which the data conforms to a defined business rule or constraint. Validity can be ensured or at least enhanced when using data capture systems designed with modern database technology. Invalid data is primarily present in legacy environments when constraints are not implemented in software, or where inappropriate data capture techniques (e.g. spreadsheet) are used where it is not possible to constrain what values can be entered by users.
- Completeness

  Completeness measures the extent to which all required values are known. Using data cleansing methods is almost impossible to resolve incompleteness: there are few feasible methods to infer facts that were not captured when the relevant data were originally recorded. Imputation method may be helpful in improving data completeness and is used commonly for this. However, in this study, some information regarding hotel customers' nationality and birthday were not recorded during the data entry process, and these data are unlikely to be recovered by data cleansing.

- Consistency

  Consistency describes the degree to which a set of measures are equivalent across the system. Inconsistency occurs when two data items in a dataset contradict each other. For example, in this study, inconsistency happens when the number of adults registered in a room differs from the headcount of guests at check in. Fixing inconsistencies is not always possible: it requires a variety of strategies - for example, deciding which data was more recently recorded, which data sources are probably the most reliable, or just trying to find the truth by testing two data items.
- Accuracy

  Accuracy represents the degree to which a measure is consistent with a standard or true value. In general, it is difficult to achieve accuracy through data cleansing because it requires access to external data sources that contain real values, which are usually not available.
- Uniformity

  Uniformity measures the degree to which the dataset is specified using the same unit of measure across the system. In datasets retrieved from different locales, data may be recorded by different units, such as in Euros and dollars, pounds and kilograms, gallons and liters. These data must be converted to a single metric by arithmetic transformation to maintain a high level of uniformity.

'Dirty' data can be caused by a variety of reasons. To clean the data these errors must be identified and corrected. Maletic & Marcus (2009) suggest that the data cleansing process consists of three phases:

- Defining and determining error types
- Searching and identifying error instances
- Correcting the uncovered errors

Returning to the current study, the raw data obtained from the PMS vendor is cleaned in line with the quality metrics suggested by Müller & Freytag following the three phases above. As noted in Table 3, each attribute is examined for possible errors with respect to validity, completeness, consistency, accuracy and uniformity. Solutions are proposed to eliminate the errors and clean the data.

Finally, redundant data that is not relevant to the aims of this study needs to be discarded. Some of the transactions in the dataset contain irrelevant information in terns of addressing the research questions; these transactions can be indentified and removed by examining:

1. Reservation status: The original dataset includes all successfully confirmed reservations, regardless of whether they are cancelled after confirmation. Cancelled reservations are unsuccessful purchases and do not directly indicate hotel guests' preference for hotel choices. It is undeniable that cancelled reservations hold certain value to reflect hotel customers' preferences indirectly, for example, reasons of cancellation may imply aspects of service improvement. However, to successfully analyze the cancelled reservations, it is necessary to obtain additional information such as the reasons for cancellation and the customers' alternative choices, which are not available from this data source. Therefore, in this study, the cancelled transaction (reservation status = C) are discarded.

2. Booking source: This study is interested in the purchase behavior involved in external reservations only, which are reservations booked through external channels and create sales for the hotel, excluding internal complimentary bookings for owners and employees or free accommodation offered to reviewers during a trial period. Therefore, this study uses the "booking source" attribute to identify internal reservations (Table 4).

Table 3: Data cleansing - potential error examination and solutions

Attributes	Potential error types					Searching for error instances	Solutions
	Validity	Completeness	Consistency	Accuracy	Uniformity		
Hotel code	-	-	-	-	-	-	-
Stay ID	-	-	-	-	-	-	-
Confirmation ID	-	-	-	-	-	-	-
Travel agent name	Items from preset list, no potential validity errors	As the hotel relies on this attribute to regularly verify the number of reservations with travel agencies and online travel agencies, the completeness of this attribute is unlikely to be in error. The missing values represents transactions (or bookings) that are not reserved through TA or OTA	Possible consistency errors may occur between this attribute and "Booking source"	Same reason as given to errors regarding completeness. The accuracy of this attribute is unlikely to be in error.	-	Compare this attribute with the "Booking source" attribute and look for discrepancies. If the value of this attribute is null, and the value of "Booking source" attribute is TA or OTA, the booking source value is perceived as error; if the value of this attribute is not null, the error is then discovered by identifying discrepancy between the nature of travel agent and booking source value.	The consistency errors identifies are corrected by assuming "Travel agent name" attribute has 100% completeness and accuracy.
Reservation status	-	-	-	-	-	The reservation status will simultaneously change at cancellation, check-in and check-out. No errors of any sort are likely to occur.	-

Check-in date	Constraint on data type, unlikely to have validity error	Mandatory field, impossible to have completeness error	No consistency errors are possible in this data set	Possible errors due to manual input	-	Though there is very slight possibility of accuracy errors due to manual input, there are no effective means to identify or correct them. In addition, the check-in/-out dates are most focused information for hotel and guests, so the possibility of errors is negligible.	-
Check-out date	Constraint on data type, unlikely to have validity error	Mandatory field, impossible to have completeness error	No consistency errors are possible in this data set	Possible errors due to manual input	-	Though there is very slight possibility of accuracy errors due to manual input, there are no effective means to identify or correct them. In addition, the check-in/-out dates are most focused information for hotel and guests, so the possibility of errors is negligible.	-
Booking date	Constraint on data type, unlikely to have validity error	Mandatory field, impossible to have completeness error	No consistency errors are possible in this data set	Auto-generated value, no accuracy errors can occur.	-	-	-
Room type	Items from preset list, no potential validity errors	Mandatory field, impossible to have completeness error	No consistency errors can be identified to this attribute	Possible errors due to manual input	-	Similar to check-in/-out date, the room type attribute are confirmed with hotel customers, possibility of errors is negligible.	-

Room rate code	Items from preset list, no potential validity errors	Mandatory field, impossible to have completeness error	No consistency errors can be identified to this attribute	Possible errors due to manual input	-	Similar to check-in/-out date, the room rate code attribute are confirmed with hotel customers, possibility of errors is negligible.	-
Average daily price	Constraint on data type, unlikely to have validity error	Mandatory field, impossible to have completeness error	No consistency errors can be identified to this attribute	Auto-generated value, no accuracy errors can occur given that the rate code is correctly selected.	The unified default currency unit in the system is HKD. No errors identified	-	-
Booking source	Items from preset list, no potential validity errors	Mandatory field, impossible to have completeness error	Possible consistency errors may occur between this attribute and "Travel agent name"	Possible errors due to manual input	-	As of discrepancies with Travel agent names, refer to strategies suggested for attribute "Travel agent name"; as of other errors caused by manual input mistakes, they are unlikely to be identified.	Refer to "Travel agent name"
Number of adults	Constraint on data type, unlikely to have validity error	Mandatory field, impossible to have completeness error	By counting transactions with same Confirmation ID, it reveals the number of registered guests under the same booking. Consistency errors may occur between the guest count from "Confirmation ID" and the attribute	Possible errors due to manual input	-	Compare this attribute with the guest count from "Confirmation ID" and look for discrepancies. The smaller of the two values is always considered error. If the Number of adults value is smaller, it is assumed to be caused by input error; if the guest count from	Ensure the value to be equal to the larger of the two guest count values.

			"Number of adults"			"Confirmation ID" is smaller, it is assumed that some guests were not registered in the system.	
Number of children	Constraint on data type, unlikely to have validity error	Mandatory field, impossible to have completeness error	No consistency errors can be identified to this attribute	Possible errors due to manual input	-	Accuracy errors are unlikely to be identified.	-
Guest ID	Auto-generated number in permanent guest profiles. No potential validity errors	Bookings can be linked to the permanent profile only when necessary information to identify the guest is given at the time of booking. Possible to have completeness errors, but lost information is unlikely to be recovered.	No consistency errors can be identified to this attribute	It is assumed that guest identity validation when linking the profile to the booking would eliminate majority of accuracy errors.	-	Completeness errors are unlikely to be identified or recovered.	-
Share code	Auto-generated value, no potential validity errors	Mandatory field, impossible to have completeness error	No consistency errors can be identified to this attribute	No accuracy errors can be identified to this attribute	-	-	-
Gender	Items from preset list, no potential validity errors	Mandatory field, impossible to have completeness error	No consistency errors can be identified to this attribute	Possible errors due to manual input	-	Accuracy errors are unlikely to be identified.	-
Nationality	Items from preset list, no potential validity errors	Possible to have completeness errors, but lost information is unlikely to be	No consistency errors can be identified to this attribute	Possible errors due to manual input	-	Accuracy errors are unlikely to be identified.	-

		recovered.					
Date of birth	Constraint on data type, unlikely to have validity error	Possible to have completeness errors, but lost information is unlikely to be recovered.	No consistency errors can be identified to this attribute	Possible errors due to manual input	-	The guest must be at least 18 years old to be registered. Calculate the guest age on the date of check in, all "Date of birth" values indicating the guest to be below 18 year old are considered errors. Other accuracy errors are unlikely to be identified.	Discard error values.
Company ID	Auto-generated number in company profiles. No potential validity errors.	Possible to have completeness errors, but lost information is unlikely to be recovered.	No consistency errors can be identified to this attribute	Possible errors due to manual input	-	Accuracy errors are unlikely to be identified.	-

4.2 Data transformation and information generation

Data preparation is a necessary step before actual data mining, ensuring that data is correctly initialized as input for the chosen data mining algorithm (García, Luengo, & Herrera, 2015). In addition to the data cleansing, other data preparation processes include data transformation, data integration, data normalization, missing data imputation, and noise identification. These processes help improve data accuracy, incorporate and adjust data, unify and scale data, handle missing data, and detect and manage noise respectively (García, Luengo, & Herrera, 2015).

In this study, classical data transformation that generates new attributes under human-supervision is a necessary pre-processing step. The original dataset contains limited information; potential preference-influencing attributes proposed in Chapter 2 require certain conversion and integration of existing attributes to be performed. Therefore, in this section, the cleaned data are transformed and integrated to generate further relevant information. Based on the findings of the review of prior researches, potential influencing factors for hotel selection are: travelers' purpose and behavior, travelers' demographic factors, and hotel attributes. From the existing attributes in the dataset, new attributes under the categories of these influencing factors are derived as follows:

1. Booking channel

Travel agencies and online travel agencies play an important role in hotel bookings. Hotel brands, especially individual hotel brands, rely on their promotion to generate business and gain visibility. However, commissions are charged by TAs and OTAs for every booking received through them. In comparison, direct-booking customers bring higher profits to the hotels. As a result, hotel management are seeking to understand the preferences of direct-booking customers to develop their loyalty to the brand. Booking channel is therefore emerging as an interesting attribute for this study.

By examining the "travel agent name" of the transaction, it is possible to distinguish whether the reservation is made through an online travel agency (OTA) or a traditional travel agency (TA). An online travel agency, also known as booking website, is defined as an online platform that provides travel planning sources and booking capabilities. Some of the most widely known OTAs are booking.com, agoda.com, and expedia.com. A transaction in which the “travel agent name” value is null is considered to be a reservation made directly through the hotel (Direct).

2. Length of stay

The attribute “length of stay,” in relation to understanding customers’ hotel booking behavior can either be the consequent or the antecedent. Identifying the factors that influence customers’ length of stay may help hotels predict their future purchasing behaviors. On the other hand, studying length of stay as an antecedent in a well-arranged trip might reveal its impact on other choices in hotel reservation. The length of stay can be calculated by the following formula:

"Check-out Date" - "Check-in Date" = "Length of stay"

3. Advance booking days

Customers’ booking habit is another potentially important behavioral attribute of their hotel booking preference. Rules associated with this attribute may help answer questions like “which hotel guests will plan ahead and book earlier?” or “which guests will walk into the hotel and book a room for the same day?”

The number of days booking in advance can be calculated by the following formula:

"Check-in date" - "Booking date" = "Advance booking days"

4. Repeated stays

Loyal guests are of high value to a hotel as they bring more profit (Kandampully & Suhartanto, 2000). Loyal customers may recommend the hotel to friends and family; and they usually book directly with the hotels so that no commission fees are paid to TAs and OTAs. Understanding the behavior patterns of returning guests can help the hotel to provide tailored promotions and further enhance customer loyalty; at the same time, it enables the hotel to target customers who have more potential to become loyal. In the PMS, if a permanent profile is created for a customer at the time of registration, the system will assign this customer a unique guest ID. Future reservations of the same customer can be linked to the profile and guest ID. Therefore, by counting the number of transactions with the same guest ID it is possible to calculate the number of repeated stays of the guest within a certain period of time.

5. Age

Age may be an important demographic factor in this study. Birthday information is provided by guests on a voluntary basis at registration. Customers' age at check-in can be calculated with the formula:

"Check-in date" - "Date of birth" = "Age at check-in"

The age values are rounded down to integers in years. All "date of birth" values that result in ages less than 18 are considered either errors or irrelevant (for marketing purposes) during the data cleansing. Therefore, all age values are greater than or equal to 18.

6. Travel type

Travel type refers to the nature and size of the traveler group. Travel type has been discussed previously as having an impact on hotel customer selection preferences (Li et al., 2013). Business travelers, single travelers, couples and families have shown differences in attribute emphasis when choosing hotels (Yoo, McKercher, & Mena, 2004). In this study, travel types are determined by the number and nature of guests in a single reservation.

As suggested at data cleansing, inconsistency may occur between the number of registered guests in the same reservation and the value of attribute "number of adults", where the former value can be calculated by counting the number of transactions sharing the same confirmation ID. The discrepancies between the two values can be caused by one or more of the following situations:

- Not all guests are registered in the PMS system at check-in. This is a common phenomenon considering potential confusion and/or lack of staff at the hotel front desk during busy hours. Incomplete guest registration will result in a positive difference between the "number of adults" value and the number of registered guests.

- Errors may occur when inputting information into the PMS. The "number of adults" information is provided by the customer or travel agent at the time of the booking process and is manually entered into the system by the booking agent. Mistakes in manual input cannot be eliminated.
- Travel plans may change with additional or absent guests at the time of check-in. The changes in “number of adults” may not be recorded in the system by mistake.

In this property management system, rate code takes the number of adults as a variable when calculating the room rate. Assuming that when the recorded “number of adults” is greater than the actual number, the guests are likely to notice the over-high room rate and require correction of the mistake; and when the recorded “number of adults” is less than the actual number, the guests may choose to ignore the difference in price and remain silent. If such an assumption is true, then:

When *"number of adults" value* ≥ *number of registered guests*, *"number of adults" value* is true;

When *"number of adults" value* < *number of registered guests*, the *number of registered guests* is true.

Therefore, in the process of data cleansing, the larger value between the "number of adults" and the number of registered guests is recognized as the "actual number of adult guests" involved in the reservation. Errors in accuracy are inevitable under such assumptions, but these errors cannot be eliminated in the absence of additional information, so the inaccuracy is tolerated in this study.

After sorting out the number of adults included in each reservation, the travel types are defined as follows:

Travel type = *Single,* when number of adults = 1, number of children = 0;

Travel type = *Pair,* when number of adults = 2, number of children = 0;

Travel type = *Group*, when number of adults ≥ 3, number of children = 0;

Travel type = *Family,* when number of adults ≥1, number of children ≥ 1.

7. Travel purpose

The two most commonly studied travel purposes in prior research are leisure and business. In this research, "day use" is introduced as an additional travel purpose to be analyzed. In hotel industry, "day use" refers to reservations that check in and out on the same day. Day-use reservations usually serve specific purposes, such as dating, waiting for late-night airlines, and short breaks during shopping or traveling. All transactions with same check-in date and check-out date in the dataset are considered as "*day use*" purposes.

Travelers usually do not state their travel purposes at the time of booking, and the hotels generally do not investigate or record this information. However, this does not mean that a supposition of a traveler's purpose cannot be deduced from the information available in the data set. The booking source attribute in the transaction can help identify the source of the booking and thereby enable speculation on possible travel purposes (Table 4). The company ID attribute in the transaction can help identify bookings from companies that have contracts with the hotel. Since no additional information is available to ascertain the nature of these companies, this study assumes that all bookings with company ID are for *business* purposes. Business travelers booked through other channels cannot be identified in the absence of more detailed information, so it is assumed that all other bookings are for *leisure* purposes.

Table 4: Travel purpose speculation based on booking source

Hotel Code	Booking Source	Description	Internal booking	Travel purpose
A	AIR	Airline company bookings for pilots and attendants		**Business**
A	COM	Complimentary	Yes	Leisure
A	COR	Corporate		**Business**
A	DAY	Day Use		Leisure
A	DIR	Hotel Direct		Leisure
A	EMA	Email		Leisure
A	FAX	Fax		**Business**
A	GDS	Global Distribution System		Leisure
A	GRP	Group		Leisure
A	HBS	Hotelbeds.com (online distribution platform)		Leisure
A	HSE	House Use (internal use)	Yes	Leisure
A	IND	Industry		**Business**

A	LTA	Local Travel Agency		Leisure
A	MGT	Management Referral	Yes	Leisure
A	MIC	MICE (meetings, incentives, conferences, and exhibitions)		**Business**
A	OPR	Open Room (complimentary trial stays before hotel opening)	Yes	Leisure
A	OTA	On Line Travel Agent		Leisure
A	OTH	Others		Leisure
A	OWN	Owner Referral	Yes	Leisure
A	SAL	Sales department		Leisure
A	SCE	Secret Escapes (online booking agent)		Leisure
A	SIT	Siteminder (booking channel managing platform)		Leisure
A	SYN	SynXis (online central reservation system)		Leisure
A	TEL	Phone		Leisure
A	WHO	Wholesale		Leisure
A	WKI	Walk In		Leisure
A	ZOO	Travelzoo (online booking agent)		Leisure
B	COMP	Complimentary	Yes	Leisure
B	CORP	Corporate		**Business**
B	GDT	Guest Direct		Leisure
B	HSE	House Use	Yes	Leisure
B	OTA	Online TA		Leisure
B	TA	Travel Agent		Leisure
B	WEB	Hotel Website		Leisure
C	FO	Front Office		Leisure
C	IBE	Hotel Website		Leisure
C	OWN	Owner Booking	Yes	Leisure
C	SO	Sales Office		Leisure
C	WIN	Walk-in		Leisure

8. Holiday status

Holidays are known to have a substantial impact on the tourism and hospitality industry. Resort hotels and airport hotels have peak seasons during the holidays, but business hotels are busier during weekdays (Jeffrey & Barden, 2000). Fluctuations in room rates brought about by holidays may also affect booking choices for different hotel guests.

Hong Kong celebrates both Western festivals and Chinese traditional festivals. The specific dates of public holidays can be retrieved from the Hong Kong Government website (https://www.gov.hk). By comparing the public holidays with the check-in/-out date of each transaction, a new attribute "holiday status" is established to indicate whether the travel is carried out during holidays. The judgment of the holiday status is implemented in two steps:

- The first step is to retrieve all the public holiday dates within the date range from which the original PMS transaction data is obtained. If the public holiday date is immediately next to a weekend, the weekend is also considered as a holiday date.
- The second step is to compare holiday dates with the check-in date and check-out date in each transaction. If any day between the arrival and departure date (excluding the day of departure as the guest will check out in the morning and not spend the night in the hotel) is a holiday, the transaction is marked as "*holiday*", otherwise the transaction is marked as *"non-holiday"*.

In addition to creating new attributes, some attributes in the original dataset need to be converted to be more meaningful to this research. The attributes that need to be converted are Room Type and Average Daily Price.

Room Type

Hotel customers' choice of room type may be influenced by factors including the size of the traveler group, purpose of the trip, and budget of the travelers. Studying their choices provides another perspective to understand the association between these influencing factors and customers' purchasing behavior. Due to the absence of room type description provided in this database, basic room features such as room size, bed size, number of beds, and the view of the room are not available. However, by comparing the average price of each room type with the price level of the hotel, it is possible to estimate the positioning of each room type: premium, standard, or economy. The median price of each room type is calculated and compared with the lower and higher tertile room rate of the hotel. If the median price of the room type is less than the lower tertile of the hotel room rate, the room type is marked as *economy*; if the median price of the room type is greater than the higher tertile of the hotel room rate, the room type is marked as *premium*; all other room types are marked as *standard*.

Average daily price

The room rate of the same room type can fluctuate with seasons, holidays, room demands, time of booking, and even the economical and political situation of the city. Studying customers' purchasing behaviors at different price levels could enable hotels to understand their sensitivity to price fluctuation. Similar to the conversion of the attribute "room type", the lowest one third of average daily prices of each room type are considered "*low*" rates; the second third of prices of each room type are considered "*medium*"; and the highest one third of prices for each room type are marked as "*high*" rates.

After the process of new attribute generation and data transformation, the dataset comprises 28 attributes. As shown in Table 5, fifteen now redundant attributes are discarded, retaining thirteen attributes that are potentially valuable for understanding travelers' hotel choice preferences (Table 5). In the next section, data reduction is conducted on the remaining 13 attributes to prepare the database for association rule mining with Apriori.

Table 5: interesting attributes for association rules mining

Attributes	**Interesting**	**Influencing factor category**
Hotel code	Yes	Hotel attributes
Stay ID		
Confirmation ID		
Travel agent name		
Reservation status		
Check-in date		
Check-out date		
Booking date		
Room type	Yes	Hotel attributes
Room rate code		
Average daily price	Yes	Hotel attributes
Booking source		
Number of adults		
Number of children		
Guest ID		

Share code		
Gender	Yes	Demographic factors
Nationality	Yes	Demographic factors
Date of birth		
Company ID		
Booking channel	Yes	Traveler behavior and purpose
Length of stay	Yes	Traveler behavior and purpose
Advance booking days	Yes	Traveler behavior and purpose
Repeated stays	Yes	Traveler behavior and purpose
Age	Yes	Demographic factors
Travel type	Yes	Traveler behavior and purpose
Travel purpose	Yes	Traveler behavior and purpose
Holiday status	Yes	Traveler behavior and purpose

4.3 Data reduction

By this point the transaction data from the three hotels have been integrated, cleaned, and transformed, and the number of potentially interesting attributes for rules mining is compressed down to thirteen. There is one more preparation step before the database is ready for data mining. Apriori, the algorithm chosen by this research to discover frequently occurring item sets and association rules from the transactions, needs to establish candidate item sets by scanning the database multiple times. In order to ensure the efficiency of the algorithm and reduce its demand for computer memory, the number of effective variables, that is, the number of items in the "item base", is suggested to be compressed (Kavitha & Selvi, 2016). The numerical attributes, namely “*length of stay*”, “*advance booking days*”, “*repeated stays*”, and “*age*”, need to be converted into discrete or nominal attributes with a finite number of intervals or categories. At the same time, the numerous values of particular nominal attributes such as *“nationality”* need to be reduced by discretization or categorization. In order to control the total number of items in the item base for association rule mining, three or four values are suggested to be retained for each attribute.

Length of stay

In order to ensure balance between and representativeness of the "length of stay" categories to be used for data reduction, the proportions of transactions with different lengths of stay are calculated, and presented in the table below.

Length of stay (days)	0	1	2	3	4	5	6	7 and more
Percentage	0.4	49.0	24.7	14.8	6.3	2.3	1.1	1.4

Since the transactions with "length of stay = 0" are in fact defined with a specific travel purpose "day use", to prevent potential interference in association rules mining. This is given an independent category as "*Day*". Nearly half of the transactions have "length of stay = 1", which forms the second category "*Single*". Considering that public holidays in Hong Kong, identified in section 4.1, generally last for 2-3 days, a third category "*Holiday*" is defined as "length of stay = 2 or 3", and the last category "*Extended*" contains all length of stay values equal to or greater than 4.

Advance booking days

Similarly, the proportions of transactions with different advance booking days are shown in the table below.

Advance booking (days)	0	1	2	3-6	7-20	21 and more
Percentage	14.1	8.2	5.3	14.3	23.3	34.9

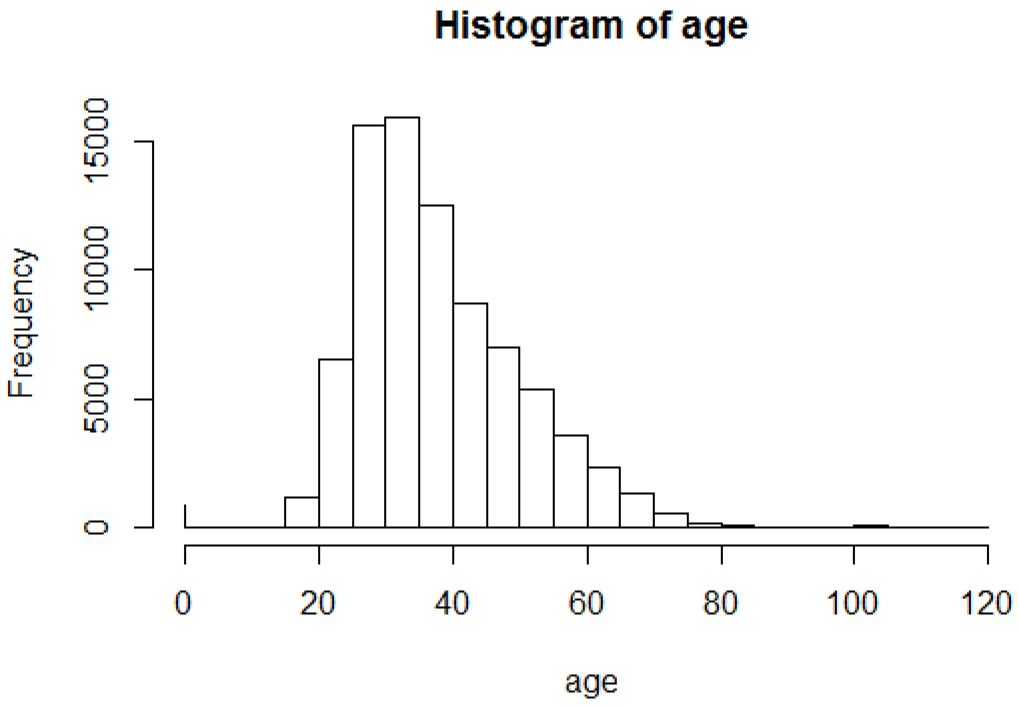
In determining the division of value intervals for this attribute, the conventional reward system employed by the hotel industry to encourage advance booking is used as a reference. Hotels generally provide extra discounts and rewards for guests who make their reservations 7 days or 21 days in advance, therefore 7 and 21 are used as endpoints in the interval division of the attribute. Since same-day reservation has been considered important for learning the purchasing behavior of hotel guests in previous studies (Jang, Chen, & Miao, 2019), the numerical attribute "advance booking days" is eventually converted into a nominal attribute with four intervals: "*Same day*" (advance booking days = 0), "*Short*" (advance booking between 1-6 days), "*7 days*" (advance booking between 7 and 20 days), and "*21 days*" (advance booking equal to or more than 21 days).

Repeated stays

By observing the database it can be seen that the majority (77.3%) of the transactions are not associated with repeat guests. Among the remaining 22.7% of the transactions, 12.7% are for repeat guests who book the same hotel twice, and 10.0% for guests with 3 or more repeated bookings. Therefore, values of this attribute are categorized in three groups: "1" (non-repeat guest), "2" (repeated guest with 2 stays in the same hotel), and "3 more" (repeated guests with 3 or more stays in the same hotel).

Age

By plotting the histogram of hotel customers' age contained in the database, it can be observed that the age distribution is mainly concentrated in the interval of 20-50 years old. This study therefore uses four age groups to ensure balance between each age group, namely 18-29 years old, 30-39 years old, 40-49 years old, and 50 years old and above.



Nationality

Mainland Chinese visitors account for 78% of all inbound visitors in Hong Kong (Statistics.gov.hk, 2018). Understanding the behavior and preference of Mainland Chinese tourists would be therefore particularly interesting and meaningful for this study. In this dataset, it can be observed that transactions with "nationality = CN (Mainland China) and HK (Hong Kong)" make up 49.8% of the total transactions. All visitors that do not come from Mainland China and Hong Kong are considered "foreign" visitors and their nationalities are categorized as "OTH" (other countries).

All attributes and items generated from the data preparation processes are now ready for rule mining with Apriori, and are shown in Table 6. The following sections report the discovery and analysis of interesting association rules that may reveal travelers' hotel choices and purchasing preferences.

Table 6: list of attributes and items for association rule mining

Influencing factor category	Attribute	Item
Hotel attributes	Hotel code	Hotel code=A
		Hotel code=B
		Hotel code=C
	Room type	Room type=economy
		Room type=standard
		Room type=premium
	Average daily price	Average daily price=low
		Average daily price=medium
		Average daily price=high
Demographic factors	Gender	Gender=male
		Gender=female
	Nationality	Nationality=CN
		Nationality=HK
		Nationality=OTH
	Age	Age=18-29
		Age=30-39
		Age=40-49
		Age=50 above

Traveler behavior and purpose	Booking channel	Booking channel=TA
		Booking channel=OTA
		Booking channel=DIRECT
	Length of stay	Length of stay=day
		Length of stay=single
		Length of stay=holiday
		Length of stay=extended
	Advance booking days	Advance booking days=same day
		Advance booking days=short
		Advance booking days=7 days
		Advance booking days=21 days
	Repeated stays	Repeated stays=1
		Repeated stays=2
		Repeated stays=3 more
	Travel type	Travel type=single
		Travel type=pair
		Travel type=group
		Travel type=family
	Travel purpose	Travel purpose=day use
		Travel purpose=business
		Travel purpose=leisure
	Holiday status	Holiday status=holiday
		Holiday status=nonholiday

Chapter 5: Association rule mining

The database has been prepared for data mining using Apriori with the completion of data preparation. There are at most 13 items in each transaction, corresponding to 13 attributes that are potentially influencing the customer's hotel selection. Appendix 2 shows a sample of the database at the conclusion of data preparation. All attributes of potential interest are grouped into three hotel selection influencing factor categories: hotel attributes, demographic factors, and customer behaviors.

When using Apriori for association rule mining the format of the association rule is expressed as $X \Rightarrow Y$, where X (antecedent) and Y (consequent) are both item sets from the item base I. Restricting either or both sides of the rule enables the discovery of association rules with specific meanings. For example, by controlling X for rule mining, it is possible to identify item sets that often occur in transactions where X is present; in this research such rules may suggest frequent behavior patterns of a particular type of hotel customer. Four types of antecedent- or consequent-controlled association rule mining are conducted in this research as they may answer questions that are particularly important or interesting for hoteliers:

1. Mine all interesting rules with demographic factors as antecedent. The association rules discovered under such restrictions may reveal hotel attribute preferences and behavior patterns of customers of different gender, nationality and age group.

2. Mine all interesting rules with travel purpose or customer behavior attributes as antecedent. These rules can inform the discovery of correlations between customer behaviors and hotel attribute preferences in different travel circumstances.

3. Mine all interesting rules with hotel attributes as consequent. Rules discovered under these restrictions are likely to indicate which customers are selecting a certain hotel under particular traveling situations.

4. Mine all interesting rules with traveler behavior attributes as consequent. These rules might reveal underlying behavioral tendencies of different types of hotel guests.

5.1 Antecedent control: Demographic factors

In Apriori, when applying specific item restrictions to the left-hand side item set (the antecedent), the LHS item set will be limited to contain only one item. The right-hand side item set (the consequent) in Apriori is predefined to contain only one item as well. Therefore, by controlling the antecedent, Apriori can discover association rules from frequent item sets of *two* items. Support and confidence thresholds were set to 0.05 and 0.10 respectively to allow the algorithm to produce as many association rules as possible for screening. Selection of these values depends on a number of factors: size of datasets, number of attributes, desirable of accuracy. For certain items with particularly low support (e.g. travel purpose=day use), lower support thresholds were used to mine potentially interesting rules. The rules are filtered and sorted by lift, and only rules that have a lift no less than 1.10 are retained.

Some correlations among the demographic factors of the hotel guests are noticeable as they implies a few customer segments that most frequently visited the hotels studied in this research. First, there is a significant correlation between nationality and the gender of the hotel customers. Among the customers of the three hotels surveyed, female travelers are more likely to be from either Mainland China or foreign countries, while Hong Kong is the top ranking nationality in terms of males travelers who stayed in one of the hotels.

Rules	Supp	Conf	Lift
{gender=F} ⇒ { nationality=OTH}	0.21	0.54	1.21
{gender=F} ⇒ {nationality=CN}	0.15	0.38	1.23
{gender=M} ⇒ {nationality=HK}	0.16	0.27	1.42

Strong correlations have also been found between the gender and the age of the hotel customers. It appeared from the rules that female customers have a higher frequency of being between 18 and 29 years of age, while male customers are more frequently falling into the 40-49 age group.

Rules	Supp	Conf	Lift
{gender=F} ⇒ {age=18-29}	0.13	0.32	1.39
{gender=M} ⇒ {age=40-49}	0.12	0.20	1.16

Finally, correlations between the traveler's nationality and their age suggest variation in the age distribution of hotel guests from different regions. It can be concluded from the following rules that travelers from mainland China are frequently younger generations aged below 30, Hong Kong local customers are more concentrated in the age group of 40-49, and travelers from abroad are more likely to be over 30 years old, especially over 50 years old.

Rules	Supp	Conf	Lift
{nationality=CN} ⇒ {age=18-29}	0.10	0.33	1.43
{nationality=HK} ⇒ {age=40-49}	0.04	0.23	1.30
{nationality=OTH} ⇒ {age=50 over}	0.09	0.21	1.33
{nationality=OTH} ⇒ {age=40-49}	0.09	0.20	1.14
{nationality=OTH} ⇒ {age=30-39}	0.16	0.35	1.11

In addition to revealing patterns in hotel customers' age, gender, and nationality, association rules mined with demographic factors as antecedent also describe different hotel selection tendency and hotel booking behaviors in relation to their gender, nationality and age.

I) Gender

Some interesting variation is found between female and male customers in terms of hotel selection and traveling behaviors. In the study of hotel choices for different genders, female guests are more likely to choose hotels A and B, which are located in Yau Ma Tei and Tsim Sha Tsui respectively, with convenient transportation and close to shopping. Correspondingly, male travelers are more inclined to choose hotel C. One reasonable hypothesis about the emergence of this difference is that women are more likely than men to be interested in hotel locations, transportation convenience and shopping, in addition to the "security" factor suggested by McCleary et al. (1994) and Lockyer (2002). However, additional information is needed to prove this hypothesis.

Rules	Supp	Conf	Lift
{gender=F} ⇒ {Hotel code=A}	0.21	0.55	1.31
{gender=F } ⇒ {Hotel code=B}	0.17	0.45	1.16
{gender=M} ⇒ {Hotel code=C}	0.18	0.31	1.62

Male and female hotel customers also show significant differences in booking behavior. Among them, in the preference of travel type, women show higher tendency of staying with another companion in the same room, while men tend to travel alone. In previous studies, female hotel customers paid high attention to safety and showed greater sensitivity to price compared to male customers (McCleary et al., 1994). Their cautiousness and frugality might well explain their choice of travel type here, as staying with another companion can not only improve security, but also reduce expenses by splitting the room rate.

Rules	Supp	Conf	Lift
{gender=F} ⇒ {travel type=PAIR}	0.30	0.77	1.13
{gender=M} ⇒ {travel type=SINGLE}	0.18	0.30	1.37

In addition, the association rules also suggest that women tend to stay longer in hotels, while more than half of the male hotel guests only spend one night in hotels.

Rules	Supp	Conf	Lift
{gender=F} ⇒ {length of stay=Holiday}	0.19	0.49	1.23
{gender=M} ⇒ {length of stay=Single}	0.33	0.56	1.15

Another potentially interesting rule indicates that when making hotel reservations, women tend to plan early and make the booking far ahead of the arrival date. The strong correlation was observed between female hotel customers and 21-day advance booking, with confidence of 0.41 and lift of 1.16.

Rules	Supp	Conf	Lift
{gender=F} ⇒ {advance booking days=21 Days }	0.16	0.41	1.16

II) Nationality

Apart from there being evidence of differences in hotel choices associated with gender differences, hotel guests of different nationalities also exhibit different hotel booking preferences. The first thing to note is the significant difference in the way bookings are made by guests from different regions. Strong correlations are observed between mainland China guests and TA booking, as well as between local customers from Hong Kong and direct bookings with the hotel. Lifts of these two rules are up to 1.89 and 2.04, which indicate significant positive correlation between the antecedent and consequent. Foreign customers are more likely to use an online booking platform for hotel reservations, with a high confidence of 0.80.

Rules	Supp	Conf	Lift
{nationality=CN} ⇒ {booking channel=TA}	0.08	0.27	1.89
{nationality=HK} ⇒ {booking channel=DIRECT}	0.06	0.29	2.04
{nationality=OTH} ⇒ {booking channel=OTA}	0.36	0.80	1.12

Hong Kong, mainland China and foreign guests also showed diverse behavioral tendencies towards booking time. From the association rules, it can be concluded that Hong Kong tourists tend to book hotel rooms on the same day or a short term ahead of check in, which is perhaps comprehensible as there is minimal need for them to plan the trip ahead and hotel stays may be improvised. Travelers from mainland China will make reservations more in advance, while foreign travelers tend to make reservations more than 21 days ahead. Taking into account the difficulty of abroad travel planning, the variance in booking time patterns between hotel customers of different nationalities is highly reasonable. In general, the further or the more difficult the journey the hotel guests have to travel to Hong Kong, the longer the duration of their planning and decision process. Mainland China travelers have various choices of relatively convenient public transportation to visit Hong Kong; therefore, they tend to plan the travel and book the hotels several days ahead. Foreign travelers, on the other hand, might make bookings several weeks in advance, not only for advance booking discounts but also for assurance of accommodation during the trip.

Rules	Supp	Conf	Lift
{nationality=CN} ⇒ {advance booking days=Short}	0.11	0.34	1.24
{nationality=CN} ⇒ {advance booking days =7 Days}	0.08	0.26	1.11
{nationality=HK} ⇒ {advance booking days =Same Day}	0.05	0.27	1.96
{nationality=HK} ⇒ {advance booking days =Short}	0.07	0.35	1.25
{nationality=OTH} ⇒ {advance booking days =21 Days}	0.22	0.50	1.41

The length of stay patterns of hotel guests also vary with nationality. Customers from Hong Kong and mainland China tend to stay in the hotel for only one day, with Hong Kong visitors showing a more pronounced tendency; while foreign customers are more frequently staying at the hotel for 2-3 nights.

Rules	Supp	Conf	Lift
{nationality=CN} ⇒ {length of stay=Single}	0.17	0.56	1.15
{nationality=HK} ⇒ {length of stay=Single}	0.16	0.85	1.73
{nationality=OTH} ⇒ {length of stay=Holiday}	0.24	0.54	1.37

Travelers of different nationalities also show diverse preferences for the three hotels. Not surprisingly, hotels A and B, with better transportation and easier access to tourist attractions and shopping, are more popular with travelers from mainland China and foreign countries. Hong Kong customers, however, showed a strong preference towards hotel C on the outlying island. After all, if a Hong Kong traveler decides to stay at a hotel rather than at home, possible reasons could be that they missed the last ferry and were trapped on an outlying island or were on a staycation.

Rules	Supp	Conf	Lift
{nationality=CN} ⇒ {hotel code=A}	0.16	0.51	1.21
{nationality=HK} ⇒ {hotel code=C}	0.11	0.60	3.11
{nationality=OTH} ⇒ {hotel code=B}	0.23	0.52	1.34
{nationality=OTH} ⇒ {hotel code=A}	0.21	0.47	1.12

III) Age

Hotel travelers of different age groups also show different preferences in hotel attribute selection and hotel booking behavior. Young travelers aged between 18 and 29 years old, are more likely to travel in pairs. Correspondingly, travelers aged 40 and above tend to have the hotel room for only themselves.

Rules	Supp	Conf	Lift
{age=18-29} ⇒ {travel type=PAIR}	0.19	0.80	1.18
{age=40-49} ⇒ {travel type=SINGLE}	0.05	0.31	1.38
{age=50 over} ⇒ {travel type=SINGLE}	0.04	0.25	1.12

Hotel guest age is also observed to be correlated with the booking time. The highlight is that 40-49 year old customers tend to book hotels 1-6 days ahead their arrival, while customers over 50 tend to plan their travel and book hotels 21 days in advance.

Rules	Supp	Conf	Lift
{age=40-49} ⇒ {advance booking days=Short}	0.06	0.31	1.12
{age=50 over} ⇒ { advance booking days=21 Days}	0.06	0.42	1.19

Another possibly interesting age-related rule suggests that hotel guests aged 18-29 tend to choose lower rates, in other words, young travelers are usually more sensitive to the price of hotel rooms.

Rules	Supp	Conf	Lift
{age=18-29} ⇒ {average daily price=Low}	0.09	0.38	1.13

5.2 Antecedent control: Traveler behavior and purpose

I) Travel purpose

Travel purpose was confirmed in previous studies to have an impact on customers' hotel booking preferences (Lewis, 1985; McCleary et al., 1993; Kucukusta, Pang, & Chui, 2013), and this finding is further validated by the association rules discovered in this study.

The first noteworthy association exists between the travel purpose and the booking channel chosen by the travelers. Strong correlations are observed between day-use travelers and direct booking, with confidence up to 0.84 and lift of 5.90. Business travelers, according the rules, are significantly more willing to make bookings via travel agencies or directly with the hotel. In fact, less than 1% of business travelers are booking hotels from online reservation platforms.

Rules	Supp	Conf	Lift
{travel purpose=business} ⇒ {booking channel=TA}	0.01	0.50	3.48
{travel purpose=business} ⇒ {booking channel=DIRECT}	0.01	0.49	3.45
{travel purpose=day use} ⇒ {booking channel= DIRECT}	0.01	0.84	5.90

When comparing the booking habits of business travelers and day-use travelers, they were also found to have significant differences in the tendency to book in advance. Business travelers prefer to book hotel rooms more than a week prior to arrival, while unsurprisingly day-use travelers are predominantly making the reservation upon arrival on the same day.

Rules	Supp	Conf	Lift
{travel purpose=business} ⇒ {advance booking days=7 Days}	0.01	0.31	1.35
{travel purpose=day use} ⇒ {advance booking days=Same Day}	0.01	0.93	6.69

In terms of price preferences, day-use hotel customers show a preference for low prices, while business travelers show a slight tendency towards medium and low prices. Hotel guests of both travel purposes avoid booking hotel rooms at higher prices, which indicates that rising room prices may have a higher impact on the booking decisions of business travelers and day-use travelers compared to those of leisure travelers.

Rules	Supp	Conf	Lift
{travel purpose=business} ⇒ {average daily price=Medium}	0.01	0.42	1.26
{travel purpose=business} ⇒ {average daily price=Low}	0.01	0.37	1.10
{travel purpose=day use} ⇒ {average daily price=Low}	0.01	0.53	1.57

Another possibly interesting finding regarding the behavior patterns relevant to travel purposes is that the travel type of both business guests and day guests are more likely to be Single, meaning that they prefer not to share the room with other companions.

Rules	Supp	Conf	Lift
{travel purpose=business} ⇒ {travel type=SINGLE}	0.01	0.32	1.47
{travel purpose=day use} ⇒ {travel type=SINGLE}	0.01	0.50	2.26

There is another association rule for day guests that reveals an interesting pattern suggesting all day-use reservations are made on non-holidays (confidence=1.00). Usually such extreme relevance is due to special reasons, such as the emergence of two items that are perfectly correlated (or mutually exclusive). A reasonable explanation for the extreme association in this rule can be that hotels will forbid bookings for day-use reservations during holidays due to overwhelming accommodation demands to maximize profit.

Rules	Supp	Conf	Lift
{travel purpose=day use} ⇒ { holiday status=NONHOLIDAY }	0.01	1.000	1.15

II) Travel types

After examining the preferences of hotel guests for different travel purposes in the hotel booking choices, another potential influencing factor studied is travel type. From empirical studies in previous research, differences in hotel selection preferences were observed between family travelers and couples (Li et al., 2013), where Asian couples were found to pay close attention to value for money while family travelers showed no particular preference. From the association rules discovered in this study, multiple underlying correlations were found between travelers of different travel types and behavior patterns.

Rules shown below indicate that single travelers are more likely to stay in hotels for shorter periods of time compared to family travelers or group travelers, who typically spend two to three days in the same hotel.

Rules	Supp	Conf	Lift
{travel type=FAMILY} ⇒ {length of stay=Holiday}	0.03	0.51	1.28
{travel type=GROUP} ⇒ {length of stay=Holiday}	0.03	0.52	1.32
{travel type=SINGLE} ⇒ {length of stay=Single}	0.15	0.67	1.37

The association rules also reveal an association between travel type and booking time patterns: the rules suggest that the larger the traveler groups are, the earlier the reservations are made. Single travelers tend to make reservations on the day of check-in or short-term before check-in, while hotel guests traveling in groups of more than three prefer to plan their travel at least 21 days in advance.

Rules	Supp	Conf	Lift
{travel type=FAMILY} ⇒ {advance booking days=7 Days}	0.02	0.28	1.20
{travel type=FAMILY} ⇒ {advance booking days=21 Days}	0.02	0.40	1.14
{travel type=GROUP} ⇒ {advance booking days=21 Days}	0.02	0.49	1.39

{travel type=SINGLE} ⇒ {advance booking days=Same Day}	0.05	0.23	1.66
{travel type=SINGLE} ⇒ {advance booking days=Short}	0.07	0.32	1.14

It is worth noting that customers of different travel types also show preferences in booking channels: more than 80 percent of group travelers chose to book hotels through OTA, while travelers with children prefer traditional travel agencies, and single travelers are significantly inclined to make reservations directly with the hotel.

Rules	Supp	Conf	Lift
{travel type=FAMILY} ⇒ {booking channel=TA}	0.01	0.22	1.53
{travel type=GROUP} ⇒ {booking channel=OTA}	0.04	0.81	1.14
{travel type=SINGLE} ⇒ { booking channel=DIRECT}	0.07	0.30	2.09

As to the difference in preference for each hotel, more than 90% of family travelers choose Hotel A, travelers in pairs or groups tend to choose Hotel B, and single travelers are more inclined to choose Hotel C.

Rules	Supp	Conf	Lift
{Travel type=FAMILY} ⇒ {hotel code=A}	0.05	0.91	2.15
{Travel type=GROUP} ⇒ {hotel code=B}	0.03	0.52	1.35
{Travel type=PAIR} ⇒ {hotel code=B}	0.31	0.46	1.18
{Travel type=SINGLE} ⇒ {hotel code=C}	0.11	0.47	2.48

III) Holiday status

In addition to travel purpose and travel type, mining while controlling other travel behaviors such as booking channels, advance booking days, repeated stays, length of stay, and booking channels as antecedent did not reveal particularly interesting association rules. However, hotel guests who travel during the holidays show some particular behavior patterns. The rules indicate that travelers on holidays tend to book hotel rooms at an extended period of time in advance. Considering the surge in demand for hotel rooms during the holidays, it is not surprising that hotel customers decide to book well in advance to guarantee accommodation during their trip to Hong Kong. Moreover, travelers are typically spending 2 to 3 days during holidays in the same hotel, and the prices they paid for the hotel rooms tend to be higher than average. However, considering that rising price and discouraging short term bookings are common practices of hotels in Hong Kong during holidays, these rules do not necessarily show autonomous choices or behavior patterns of hotel customers, but rather conformable to hotel policies. In mining how hotel customers behave during holidays, it may be more valuable to discover which customers are willing to travel on holidays despite the higher price.

Rules	Supp	Conf	Lift
{Holiday status=HOLIDAY} ⇒ {average daily price=High}	0.07	0.58	1.77
{Holiday status=HOLIDAY} ⇒ {advance booking days=21 Days	0.07	0.52	1.47
{Holiday status=HOLIDAY} ⇒ {length of stay=Holiday}	0.06	0.45	1.13

Travelers of different gender, nationality, age, travel purpose, and travel type are verified to behave differently in terms of hotel selection and hotel reservation by controlled association rule mining. It seems that rule mining with demographic factors and traveler behaviors as antecedent could enable hoteliers to discover differences in behavior patterns and to predict how customers are likely to react towards certain promotion and marketing programs. Some rules discovered in this section show strong preferences of travelers of certain attributes, which can be used to either increase customer loyalty by providing favorable packages or to maximize profits by attracting more customers with the same attributes.

5.3 Consequent control: Traveler behavior and hotel attribute

A potential flaw of the above approach is evident when studying travelers’ behavior patterns during holidays. Behavior patterns revealed by association rules may sometimes arise due to the influence of hotel policies and events. It is therefore important to discover potentially valuable information regarding these patterns from alternative perspectives, e.g., mining rules that has “holiday status=holiday” as a consequent, and to investigate which customers are more willing to travel during holidays given the surging room price. Similarly, mining rules by controlling the consequent may find answers to some interesting questions for hoteliers, for example: Which customers are more likely to choose hotel A? Which customers are more likely to be loyal and repeatedly stay in the same hotel? Which customers are less price-sensitive and willing to book premium rooms? The rules mined by restricting the antecedent have maximum length of two items, while rules mined with consequent control have no limit in length. Therefore, the answers obtained from association rules for the above questions will be more specific.

The consequent-controlled rules were mined at minimum support of 0.05, minimum confidence of 0.10, and max length of 4 to ensure a sufficient number of statistically significant rules were generated (although for rarely occurring consequent item sets such as “{room type=premium}”, the support benchmark is 0.01). The rules were sorted and screened by lift, which is an important indicator reflecting the positive correlation between antecedent and consequent. After filtering and pruning redundant rules, 1207 association rules that have lift greater than or equal to 1.20 were acquired.

However, the rules generated are not all valid or meaningful. Among the three categories of hotel selection influencing factors, certain hotel attributes (namely hotel code, room price, and room type) are decisions made by the customers after their selection process, which means hotel attributes are always consequent of their purchasing behavior rather than antecedent. Under this premise, the rules with hotel attributes in antecedent item sets are removed and 607 valid rules remain.

Despite the relatively large number of remaining rules, some possibly interesting patterns start to emerge. Particular controlling consequent item sets generate extensive numbers of rules such as "{hotel code=B}", "{length of stay=Single}", "{length of stay=Holiday}", "{repeated stay=2}", "{repeated stay=3 more}", "{advance booking days=21 days}" and "{advance booking days=Short}". Further screening regarding these consequent item sets is explained further in this section.

(A) Answering potentially interesting questions

Having acquired a number of association rules that potentially reveal factors that drive particular hotel customers' behavior and choices, answers to some potentially valued questions posed by hoteliers can be drawn from these rules.

I) Hotel choice

There are three mid-ranged individual hotels studied in this research, each located at a distinct and representative area of Hong Kong with various points of interest and different transportation situations. Hotel choice is a complicated and multi-factor decision making process that typically require large data sample and more sophisticated combination of advanced data mining methods to understand the mechanism in depth (Li et al., 2013). However, with the association rule mining approach applied in this research it is possible to identify certain types of hotel customers that have a higher tendency to choose one of the three hotels.

Further rule screening is required as extensive numbers of rules have been generated for each hotel code. It can be done by neglecting rules with attributes that are less informative in explaining this question. For example, "repeated stay" as a traveler behavior attribute reflects the frequency of the same guests returning to the same hotel. The records of repeated stay numbers are stand-alone in each hotel and are significantly affected by the number of loyal customers in each hotel and how PMS operators register the guests. This attribute can be potentially useful in telling the behavior patterns of loyal customers, e.g. booking channel preference, length of stay patterns, and advance booking patterns, but it is unreliable when drawing conclusions such as "repeated customers prefer hotel A rather than hotel B or C", as hotel B or C might have fewer loyal customers. Therefore, rules that contains the "repeated stay" attribute in antecedent item set are removed for this question.

"**Gender=F**", "**nationality=CN**", and "**booking channel=TA**" appear to be predominantly frequent attributes in customers who prefer hotel A. At least one of these items appears in every antecedent item set of hotel A rules. In addition to these attributes, it appears that travelers under 30 years old and those who travel in pairs will also consider hotel A as their choice of accommodation.

Antecedent	Consequent	Supp	Conf	Lift
{gender=F}	{hotel code=A}	0.21	0.55	1.31
{nationality=CN}	{hotel code=A}	0.16	0.51	1.21
{booking channel=TA}	{hotel code=A}	0.08	0.54	1.28
{gender=F, nationality=CN}	{hotel code=A}	0.09	0.63	1.49
{booking channel=TA, nationality=CN}	{hotel code=A}	0.05	0.61	1.46
{gender=F, advance booking days=21 Days}	{hotel code=A}	0.09	0.55	1.32
{booking channel=TA, holiday status=NONHOLIDAY}	{hotel code=A}	0.07	0.54	1.29
{booking channel=TA, travel type=PAIR}	{hotel code=A}	0.05	0.58	1.37

{nationality=CN, age=18-29}	{hotel code=A}	0.05	0.54	1.29
{gender=F, holiday status=NONHOLIDAY}	{hotel code=A}	0.19	0.55	1.31

In hotel B rules, "**nationality=OTH**", "**booking channel=OTA**", "**travel purpose=Leisure**", "**holiday status=NONHOLIDAY**", "**travel type=Pair**", and "**length of stay=Holiday**" are the most frequently occurring items. There are 33 rules of which the antecedent item set is the combination of two or three of these items, and other rules have at least one of the above items in their antecedent item sets (as below). From the rules it is noticeable that, apart from the most dominant attributes above, customers between 30-39 years of age and those who tend to book hotel rooms 21 days ahead are more likely to choose hotel B, while no significant difference in preference towards hotel B can be observed between male and female.

Antecedent	Consequent	Supp	Conf	Lift
{booking channel=OTA, **gender=F**}	{hotel code=B}	0.15	0.51	1.33
{booking channel=OTA, **gender=F**, travel purpose=Leisure}	{hotel code=B}	0.15	0.51	1.33
{**gender=F**, travel type=PAIR}	{hotel code=B}	0.15	0.49	1.27
{**gender=F**, travel type=PAIR, travel purpose=Leisure}	{hotel code=B}	0.15	0.50	1.30
{booking channel=OTA, **gender=F**, holiday status=NONHOLIDAY}	{hotel code=B}	0.13	0.52	1.33
{booking channel=OTA, **advance booking days=21 Days**}	{hotel code=B}	0.13	0.50	1.28
{booking channel=OTA, **gender=F**, travel type=PAIR}	{hotel code=B}	0.13	0.56	1.44
{**gender=M**, nationality=OTH}	{hotel code=B}	0.13	0.54	1.39
{**gender=M**, nationality=OTH, travel purpose=Leisure}	{hotel code=B}	0.13	0.54	1.40
{**advance booking days=21 Days**, travel type=PAIR}	{hotel code=B}	0.12	0.48	1.23
{**advance booking days=21 Days**, travel type=PAIR, travel purpose=Leisure}	{hotel code=B}	0.12	0.49	1.26
{booking channel=OTA, **age=30-39**}	{hotel code=B}	0.12	0.51	1.31
{booking channel=OTA, **age=30-39**, travel purpose=Leisure}	{hotel code=B}	0.12	0.51	1.31
{**age=30-39**, travel type=PAIR}	{hotel code=B}	0.11	0.50	1.30
{**age=30-39**, travel type=PAIR, travel purpose=Leisure}	{hotel code=B}	0.11	0.51	1.33
{booking channel=OTA, **gender=M**, nationality=OTH}	{hotel code=B}	0.11	0.59	1.52
{**gender=M**, nationality=OTH, holiday status=NONHOLIDAY}	{hotel code=B}	0.11	0.54	1.39
{booking channel=OTA, **advance booking days=21 Days**, holiday status=NONHOLIDAY}	{hotel code=B}	0.11	0.51	1.32
{booking channel=OTA, **advance booking days=21 Days**, travel type=PAIR}	{hotel code=B}	0.11	0.53	1.37

Frequently occurring attributes of hotel customers who prefer hotel C are "**gender=M**", "**nationality=HK**", "**travel type=Single**", "**booking channel=Direct**", and "**length of stay=Single**". These attributes are perceived to have significant positive correlation with the customers' choice of hotel C, judging by the confidence and lift of the rules.

Antecedent	Consequent	Supp	Conf	Lift
{gender=M}	{hotel code=C}	0.18	0.31	1.62
{nationality=HK}	{hotel code=C}	0.11	0.60	3.11
{travel type=SINGLE}	{hotel code=C}	0.10	0.47	2.48
{length of stay=Single}	{hotel code=C}	0.17	0.34	1.77
{booking channel=Direct}	{hotel code=C}	0.06	0.45	2.33
{length of stay=Single, gender=M}	{hotel code=C}	0.16	0.48	2.51
{gender=M, nationality=HK}	{hotel code=C}	0.11	0.70	3.67
{gender=M, travel type=SINGLE}	{hotel code=C}	0.10	0.57	2.95
{length of stay=Single, nationality=HK}	{hotel code=C}	0.10	0.61	3.18
{length of stay=Single, gender=M, nationality=HK}	{hotel code=C}	0.10	0.71	3.73
{length of stay=Single, travel type=SINGLE}	{hotel code=C}	0.09	0.60	3.15
{length of stay=Single, gender=M, travel type=SINGLE}	{hotel code=C}	0.09	0.68	3.56
{booking channel=Direct, **holiday status=NONHOLIDAY**}	{hotel code=C}	0.06	0.45	2.37
{nationality=HK, travel type=SINGLE}	{hotel code=C}	0.06	0.76	3.98
{gender=M, nationality=HK, travel type=SINGLE}	{hotel code=C}	0.06	0.81	4.25
{booking channel=Direct, gender=M}	{hotel code=C}	0.06	0.56	2.95
{booking channel=Direct, length of stay=Single}	{hotel code=C}	0.05	0.62	3.23

Comparing the influencing attributes of customers in regard to their preference for hotel selection between hotel A, B and C, it is evident that each hotel attracts particular types of travelers: hotel A appears to be popular with mainland China female travelers who book through travel agencies, hotel B attracts online-booking foreign leisure travelers who travel to Hong Kong in pairs on non-holiday days and typically stay for 2-3 days, while hotel C is most favored by Hong Kong local male travelers who like to spend one day by themselves on the outlying island, and they typically book directly with the hotel. Understanding these travelers preferences for hotels could benefit the hotel managers and their marketing staff in two ways: 1) by analyzing the customer profiles of those who are particularly attracted by their hotel they can understand their strength and develop promotion schemes and rewards to increase the loyalty of these customers; 2) by comparing the features and attracted customers of their own hotel and the competitor hotels they can develop awareness in weaknesses and challenges, and potentially understand the hotel features needed to improve to develop new markets.

II) Price sensitivity

Lewis (1984) suggested that female guests are more price sensitive than male guests in selecting hotels. McCleary et al. (1994) and Lockyer (2002) also explored differences in price sensitivity between genders in their research but both failed to draw definitive conclusions. From the association rules mined in this study, some possible interesting patterns have been found in terms of cautiousness level in price selection, which may argue that demographic factors and other attributes may indeed influence price sensitivity of particular types of customers.

To explore this particular tendency, rules that show positive correlation with higher-than-average room price, lower-than-average room price, and premium rooms are collected and examined.

Antecedent	Consequent	Supp	Conf	Lift
{holiday status=HOLIDAY}	{average daily price=High}	0.07	0.58	1.77
{booking channel=OTA, holiday status=HOLIDAY}	{average daily price=High}	0.05	0.59	1.79
{**travel purpose=LT**, holiday status=HOLIDAY}	{average daily price=High}	0.07	0.59	1.78
{booking channel=OTA, **advance booking days=21 Days**, **gender=M**}	{average daily price=High}	0.06	0.41	1.25
{booking channel=OTA, **advance booking days=21 Days**}	{average daily price=High}	0.11	0.40	1.22

It can be observed from the rules that item "holiday status=Holiday" and "booking channel=OTA" are frequently occurring as factors that drive customers to make high-price bookings. Noticeably, leisure travelers and male customers are more willingly to pay higher price on accommodation under certain circumstances.

Rules for lower room prices are examined for corresponding attributes namely travel purpose and gender, and the findings are again potentially of interest.

Antecedent	Consequent	Supp	Conf	Lift
{length of stay=Single, advance booking days=Short, travel purpose=Leisure}	{average daily price=Low}	0.07	0.41	1.21
{gender=M, nationality=CN, holiday status=NONHOLIDAY}	{average daily price=Low}	0.06	0.41	1.21
{length of stay=Single, gender=F, holiday status=NONHOLIDAY}	{average daily price=Low}	0.05	0.40	1.18

It appears that, although leisure travelers and male guests tend to be more generous when traveling on holidays or planning for journeys far ahead, they do not necessarily discarded low price rooms. Mainland Chinese males in particular, are more willing to spend less money on accommodation during non-holidays. Leisure travelers on shortly-planned one-day journeys also favor lower-priced rooms more. Female customers, as suggested by Lewis (1984), are somewhat price cautious during non-holidays but the tendency perceived is not strong (lift=1.18).

Rules for premium room bookings were examined next to further validate the difference in money-spending habits. No correlation is observed between female customers and premium rooms, but the two rules shown suggest that male travelers are more likely to pay for a luxury accommodation experience if it is a single experience (repeated stay=1) for one night or booked via OTA.

Antecedent	Consequent	Supp	Conf	Lift
{length of stay=Single, repeated stay=1, gender=M}	{room type=Premium}	0.01	0.11	1.23
{booking channel=OTA, repeated stay=1, gender=M}	{room type=Premium}	0.03	0.10	1.20

From the patterns perceived from the association rules above, it can be concluded that male travelers are more willing to pay extra money for a luxurious hotel experience or to book hotels at a high price when circumstances require, but it does not necessarily mean that male customers will give up the chance of pursuing low price rooms. Female customers, on the other hand, are not particularly focused on discounts or low room rates, but they are significantly more cautious in high price purchases.

Leisure travelers, whose attitude towards money spending on hotel rooms resembles that of male travelers, are willingly to pay extra for available rooms during holidays, but not particularly interested in premium rooms.

Pair travelers show unique patterns: on short stays or shortly-planned travels during non-holidays, they are keen to pursue low price rooms, but if such stays are for one time only, they do not mind selecting premium rooms. Considering that pair travelers are possibly couples on vacation, such patterns represent valuable messages to hotel marketers on how to attract pair travelers, who place high value on the quality of their accommodation experience but also have budget concerns.

Antecedent	Consequent	Supp	Conf	Lift
{advance booking days=Short, travel type=PAIR, holiday status=NONHOLIDAY}	{average daily price=Low}	0.07	0.41	1.22
{advance booking days=Short, travel type=PAIR , repeated stay=1}	{room type=Premium}	0.01	0.10	1.20
{length of stay=Single, travel type=PAIR, holiday status=NONHOLIDAY}	{average daily price=Low}	0.11	0.41	1.21
{length of stay=Single, travel type=PAIR, repeated stay=1}	{room type=Premium}	0.02	0.10	1.20

More patterns regarding price sensitivities are drawn from the association rules mined by controlling consequent item set as "{average daily price=Low}". Mainland China travelers compared to Hong Kong or foreign customers are more interested in low-price offers during non-holidays.

Antecedent	Consequent	Supp	Conf	Lift
{length of stay=Single, nationality=CN, holiday status=NONHOLIDAY}	{average daily price=Low}	0.06	0.42	1.23
{gender=M, nationality=CN, holiday status=NONHOLIDAY}	{average daily price=Low}	0.06	0.41	1.21

As to the age group that is most price sensitive compared to other groups, it turns out that younger travelers (aged 18-29) are more attracted to economic accommodation options during non-holidays.

Antecedent	Consequent	Supp	Conf	Lift
{age=18-29, holiday status=NONHOLIDAY}	{average daily price=Low}	0.08	0.41	1.20
{age=18-29, travel type=PAIR, holiday status=NONHOLIDAY}	{average daily price=Low}	0.07	0.41	1.20

Non-holiday bookings are particularly frequent in low average price rules. Online bookings and foreign nationals are more frequently occurring in premium room bookings. However, both short advance booking time (and same day reservation) and one-night stays are repeatedly appearing in rules on both occasions. Such patterns may point out a direction for hotels that are keen to increase premium room occupancy rate during non-holidays.

Antecedent	Consequent	Supp	Conf	Lift
{**advance booking days=Same Day**}	{average daily price=Low}	0.06	0.41	1.21
{**length of stay=Single**, **advance booking days=Short**}	{average daily price=Low}	0.07	0.41	1.20
{**length of stay=Single**, **advance booking days=Short**, holiday status=NONHOLIDAY}	{average daily price=Low}	0.07	0.42	1.24
{**advance booking days=Same Day**, holiday status=NONHOLIDAY}	{average daily price=Low}	0.05	0.42	1.24
{booking channel=OTA, **length of stay=Single**, **advance booking days=Short**}	{average daily price=Low}	0.05	0.42	1.24
{booking channel=OTA, **length of stay=Single**, **repeated stay=1**}	{room type=Premium}	0.02	0.11	1.27
{booking channel=OTA, **length of stay=Single**, nationality=OTH}	{room type=Premium}	0.01	0.11	1.26
{**length of stay=Single**, **repeated stay=1**, nationality=OTH}	{room type=Premium}	0.01	0.11	1.24
{**length of stay=Single**, **repeated stay=1**, gender=M}	{room type=Premium}	0.01	0.11	1.23
{**length of stay=Single**, **repeated stay=1**, travel type=PAIR}	{room type=Premium}	0.02	0.10	1.20
{booking channel=OTA, **advance booking days=Short**, **repeated stay=1**}	{room type=Premium}	0.01	0.10	1.20
{**advance booking days=Short**, **repeated stay=1**, travel type=PAIR}	{room type=Premium}	0.01	0.10	1.20

III) Loyalty

All hotels expend effort to increase customer loyalty, as maintaining an existing customer costs substantially lower than obtaining a new one, and loyal customers improve hotel profitability by repeat and referral business (Kandampully & Suhartanto, 2000). Understanding which customers are more likely to return to the same hotel for repeated stays therefore holds substantial value to hotel managers. Association rule mined by controlling repeated stay behaviors as consequent in this section might provide an approach to understanding the underlying patterns.

The rules suggest that mainland Chinese travelers and female travelers are more likely to return to the same hotel for a second stay, especially female travelers between 30-39 years of age and Chinese travelers between 18-39 years of age. Foreign travelers over 50 years of age also show similar patterns.

Antecedent	Consequent	Supp	Conf	Lift
{gender=F}	{repeated stay=2}	0.04	0.11	1.32
{nationality=CN}	{repeated stay=2}	0.03	0.11	1.26
{gender=F, nationality=CN}	{repeated stay=2}	0.02	0.13	1.59
{gender=F, age=30-39}	{repeated stay=2}	0.02	0.12	1.37
{nationality=CN, age=18-29}	{repeated stay=2}	0.01	0.12	1.45
{nationality=CN, age=30-39}	{repeated stay=2}	0.01	0.12	1.40
{nationality=OTH, age=50 over}	{repeated stay=2}	0.01	0.12	1.40
{nationality=OTH, age=50 over, travel purpose=Leisure}	{repeated stay=2}	0.01	0.12	1.40

Comparing to other travel types, young couples from mainland China are more frequently returning to the same hotel they stayed in their last visit to Hong Kong.

Antecedent	Consequent	Supp	Conf	Lift
{nationality=CN, travel type=PAIR}	{repeated stay=2}	0.02	0.11	1.36
{nationality=CN, age=18-29, travel type=PAIR}	{repeated stay=2}	0.01	0.12	1.46

Hong Kong travelers, especially Hong Kong couples traveling for leisure, show higher loyalty levels by repeatedly staying in the same hotel more than three times. However, a less desired pattern also emerges from the rules indicating that loyal Hong Kong travelers prefer to book their accommodation from online booking platforms. Hotel managers generally encourage hotel guests, and especially returning guests, to book directly with the hotel to avoid the substantial commission fees charged by travel agents or OTAs. The patterns revealed from these rules could remind hotel managers to pay more attention to promoting direct booking methods for their hotels, including but not limited to telephone, email, social network, or hotel websites.

Antecedent	Consequent	Supp	Conf	Lift

{nationality=HK}	{repeated stay=3 more}	0.02	0.12	1.81
{nationality=HK, travel type=PAIR}	{repeated stay=3 more}	0.01	0.13	2.09
{nationality=HK, travel purpose=Leisure}	{repeated stay=3 more}	0.02	0.12	1.90
{booking channel=OTA, nationality=HK}	{repeated stay=3 more}	0.02	0.15	2.27
{booking channel=OTA, nationality=HK, travel type=PAIR}	{repeated stay=3 more}	0.01	0.15	2.28
{booking channel=OTA, nationality=HK, travel purpose=Leisure}	{repeated stay=3 more}	0.02	0.15	2.27

Moreover, from both repeated stay patterns it can be noticed that pair travelers and leisure travelers are more likely to become loyal customers, regardless of their nationality.

Antecedent	Consequent	Supp	Conf	Lift
{nationality=OTH, age=50 over, travel purpose=Leisure}	{repeated stay=2}	0.01	0.12	1.40
{nationality=CN, travel type=PAIR}	{repeated stay=2}	0.02	0.11	1.36
{nationality=CN, age=18-29, travel type=PAIR}	{repeated stay=2}	0.01	0.12	1.46
{nationality=HK, travel type=PAIR}	{repeated stay=3 more}	0.01	0.13	2.09
{nationality=HK, travel purpose=Leisure}	{repeated stay=3 more}	0.02	0.12	1.90
{nationality=HK, travel type=PAIR, travel purpose=Leisure}	{repeated stay=3 more}	0.01	0.14	2.18

Noticeably, single travelers from outside mainland China and Hong Kong are keen to return to the same hotel repeatedly.

Antecedent	Consequent	Supp	Conf	Lift
{nationality=OTH, travel type=SINGLE}	{repeated stay=3 more}	0.01	0.14	2.16

Male travelers of the age group 40-49, when compared to female travelers, stay in the same hotel more frequently.

Antecedent	Consequent	Supp	Conf	Lift
{gender=M, age=40-49}	{repeated stay=3 more}	0.01	0.10	1.56
{gender=M, age=40-49, travel purpose=Leisure}	{repeated stay=3 more}	0.01	0.10	1.57

From the patterns discovered from these rules, it can be concluded overall that local leisure travelers, and couples in particular, are the most potentially loyal customers of hotels. Male travelers aged between 40-49 years old are also among the most persistent hotel guest groups. Younger mainland Chinese female travelers and elder foreign travelers are highly likely to return to the same hotel as they stayed in their last trip; therefore targeted promotions with offers of frequent customer discount and maybe free upgrade before they plan for the next travel could be highly effective in increasing their loyalty.

IV) Length of stay

Understanding the stay length patterns of customers is potentially important for hoteliers to plan marketing schemes and dynamic pricing strategies at different seasons of the year. From a productivity and labor consideration, hotels prefer long-term stays, as arrivals and departures require more labor at the front desk to handle the check ins/outs, and housekeeping attendants generally take longer to clean check out rooms. Long terms stays also guarantee occupancy and therefore hotels have fewer rooms to worry about selling. As a matter of fact, many hotels enforce policies to accept only long bookings during holidays to maximize the occupancy rate throughout peak seasons. By understanding the underlying patterns of customers' lengths of stay, hotel managers should be able to target particular market segments to attract potential long term staying guests. Furthermore, by understanding the behavior models of short-term guests, hotel managers can adjust promotion and room price accordingly to boost occupancy during lower seasons.

By observing the rules acquired for extended stays of more than two days, "repeated stay=1" and "advanced booking days=21 days" are particularly prominent as antecedent factors. By excluding these two factors it can be found that female hotel customers show a significant tendency to staying for longer periods, and in contrast, male travelers are more frequently staying for only one night.

Antecedent	Consequent	Supp	Conf	Lift
{gender=F}	{length of stay=Extended}	0.05	0.13	1.21
{gender=F}	{length of stay=Holiday}	0.19	0.49	1.23
{gender=F, nationality=OTH}	{length of stay=Holiday}	0.12	0.57	1.45
{gender=F, nationality=OTH, travel purpose=Leisure}	{length of stay=Holiday}	0.12	0.57	1.45
{gender=F, nationality=OTH, travel type=PAIR}	{length of stay=Holiday}	0.10	0.59	1.48
{gender=F, travel type=PAIR}	{length of stay=Holiday}	0.15	0.49	1.25
{gender=F, travel type=PAIR, travel purpose=Leisure}	{length of stay=Holiday}	0.14	0.49	1.25
{gender=F, age=30-39}	{length of stay=Holiday}	0.07	0.50	1.26

{gender=F, age=30-39, travel type=PAIR}	{length of stay=Holiday}	0.05	0.50	1.27
{gender=F, age=30-39, travel purpose=Leisure}	{length of stay=Holiday}	0.07	0.50	1.26
{gender=M, age=40-49}	{length of stay=Single}	0.07	0.61	1.24
{gender=M, nationality=HK}	{length of stay=Single}	0.13	0.86	1.75
{gender=M, nationality=CN}	{length of stay=Single}	0.10	0.65	1.32
{gender=M, nationality=HK, travel purpose=Leisure}	{length of stay=Single}	0.13	0.87	1.78
{gender=M, nationality=CN, travel purpose=Leisure}	{length of stay=Single}	0.10	0.65	1.33
{gender=M, travel type=SINGLE}	{length of stay=Single}	0.13	0.72	1.47
{gender=M, travel type=SINGLE, travel purpose=Leisure}	{length of stay=Single}	0.12	0.73	1.48

Travelers from different nations also show diverse patterns in their lengths of stay at hotels: foreign travelers are more frequently observed to spend more than two days at hotels, while Hong Kong travelers are typically staying for only one night.

Antecedent	Consequent	Supp	Conf	Lift
{nationality=OTH}	{length of stay=Holiday}	0.24	0.54	1.37
{nationality=OTH, travel type=PAIR}	{length of stay=Holiday}	0.19	0.56	1.42
{nationality=OTH, travel purpose=Leisure}	{length of stay=Holiday}	0.24	0.54	1.37
{nationality=OTH, age=18-29}	{length of stay=Holiday}	0.05	0.58	1.45
{nationality=OTH, age=30-39}	{length of stay=Holiday}	0.09	0.56	1.40
{nationality=OTH, age=30-39, travel type=PAIR}	{length of stay=Holiday}	0.07	0.58	1.45
{nationality=OTH, age=30-39, travel purpose=Leisure}	{length of stay=Holiday}	0.09	0.56	1.41
{nationality=OTH}	{length of stay=Extended}	0.08	0.17	1.55
{booking channel=OTA, nationality=OTH}	{length of stay=Extended}	0.06	0.17	1.55
{booking channel=OTA, nationality=OTH, travel purpose=Leisure}	{length of stay=Extended}	0.06	0.17	1.55
{nationality=HK}	{length of stay=Single}	0.16	0.85	1.73
{nationality=HK, travel type=PAIR}	{length of stay=Single}	0.09	0.85	1.74
{nationality=HK, travel purpose=Leisure}	{length of stay=Single}	0.15	0.86	1.75
{nationality=HK, travel type=PAIR, travel purpose=Leisure}	{length of stay=Single}	0.09	0.86	1.76
{nationality=HK, travel type=SINGLE, travel purpose=Leisure}	{length of stay=Single}	0.06	0.86	1.76

Traveling for leisure as an influencing factor has appeared in rules for both longer stays and shorter stays, therefore it is uncertain how travel purpose can affect length of stay patterns. Similarly, “travel type = Pair” also frequently occurs in both rules, which suggests nationality and age are probably more strongly influencing factors in these cases. However, single travelers show strong correlation with single day stays.

Antecedent	Consequent	Supp	Conf	Lift
{travel type=SINGLE}	{length of stay=Single}	0.15	0.67	1.37
{travel type=SINGLE, travel purpose=Leisure}	{length of stay=Single}	0.14	0.68	1.39
{gender=M, travel type=SINGLE}	{length of stay=Single}	0.13	0.72	1.47
{gender=M, travel type=SINGLE, travel purpose=Leisure}	{length of stay=Single}	0.12	0.73	1.48

By analyzing the hotel customers' choices in length of stay, the conclusion can be drawn that female travelers in their 30s are most willing to stay for longer periods of time in Hong Kong hotels, while male travelers in their 40s are frequently staying for only one day. Couples and leisure travelers from foreign countries, compared to those from Hong Kong and mainland China, are more likely to spend two nights or more for their stays in Hong Kong.

V) Advance booking

Advance booking is favored by hoteliers as it enables them to more effectively manage and forecast their revenues. Hotels usually apply promotional prices or attach additional benefits to advance bookings to encourage hotel customers to book ahead. Advertisement and promotion emails are often used to deliver hotels' marketing messages to potential customers. However, how to avoid flooding their inbox with unwanted junk mails and eventually irritating potential customers is a particular interesting subject to be studied. One widely applied approach is targeted marketing, or targeted promotion, by predicting the interest of, and sending tailored promotions to, targeted customers (Cahill, 1997). Understanding which customers are more acceptive of advance booking can be a first step in targeting the right market segment for promotion.

By excluding the holiday status attribute "non-holiday", which is predominantly frequently occurring in short advanced bookings and same day bookings, it is noticeable that nationality, travel type, and length of stay are particularly important in influencing customers' advance booking behavior. In particular, the longer the customer intends to stay at the hotel, the earlier they are tending to make the reservation; the further the customers must travel to the destination city, the more likely they are to book in advance. The patterns identified from the rules can be integrated with conclusions drawn from other analyze to provide guidance to hotel marketing personnel: from our previous analysis it was noticed that female travelers between 30-39 years old are more likely to stay for an extended period of time, and from the rules mined for advance booking it can be observed that extended stay is positively correlated with 21 days advance booking. By combining these findings together it can be confidently predicted that female travelers in their 30s are more likely to welcome promotions for early bookings for their next trips. Similarly, single day bookings are positively correlated with both short advance booking and same day booking behaviors; targeting the customer group that is particularly interested in spending one night at hotels may encourage them to book hotel rooms ahead rather than to book upon arrival.

Antecedent	Consequent	Supp	Conf	Lift
{nationality=OTH}	{advance booking days=21 Days}	0.22	0.50	1.41
{length of stay=Extended}	{advance booking days=21 Days}	0.07	0.61	1.73
{nationality=HK}	{advance booking days=Short}	0.07	0.35	1.25
{nationality=CN}	{advance booking days=Short}	0.10	0.34	1.24
{length of stay=Single}	{advance booking days=Short}	0.18	0.36	1.30
{nationality=HK}	{advance booking days=Same Day}	0.05	0.27	1.96
{travel type=SINGLE}	{advance booking days=Same Day}	0.05	0.23	1.66
{length of stay=Single}	{advance booking days=Same Day}	0.11	0.23	1.64

Female foreign travelers and male mainland Chinese travelers show particular interest towards extended long advance bookings and short advance bookings respectively. Older leisure travelers appear to prefer making plans well ahead of their travels.

Antecedent	Consequent	Supp	Conf	Lift
{gender=F, nationality=OTH}	{advance booking days=21 Days}	0.11	0.54	1.54
{gender=F, nationality=OTH, travel purpose=Leisure}	{advance booking days=21 Days}	0.11	0.54	1.54

{age=50 over, travel purpose=Leisure}	{advance booking days=21 Days}	0.06	0.42	1.20
{gender=M, nationality=CN}	{advance booking days=Short}	0.06	0.36	1.28
{gender=M, nationality=CN, travel purpose=Leisure}	{advance booking days=Short}	0.05	0.36	1.28

Leisure travel purpose and pair travel type are frequently observed in rules of both short advance booking and long advance booking, indicating that they are not as influential as gender, nationality or length of stay.

(B) Identifying behavior inducing factors

One reason that association rule mining is valuable to marketing personnel in the retail industry is that correlations between items discovered in market basket analysis can help improve product assortment decisions in retail corporations (Brijs et al., 1999). One of the most common cross-selling practices in retail industry is to sell items in combination or a bundle, for instance, toothbrushes and toothpastes, buckets and mop. The selection of bundled items takes into account the positive *radiation effect* of one item on other items in the combination, which can be identified from frequently occurring item sets in the same transactions (Brijs et al., 2004). Product bundling is often used for imperfectly competitive products that are "bundled" to more frequently purchased items. Product bundling sometimes results in customer dissatisfaction when getting items they consider to be unwanted or unnecessary. Association rule mining can be used to find "item C" that increases the sales of "item B" when sold together with "item A". Instead of bundling all three items together, if proper promotion is given to customers who purchase item A to successfully cross sell item C, it is very likely that item B is purchased at the same time without making the customer feel enforced or compelled to purchase.

A similar strategy can be applied in hotel promotion programs. With in depth understanding of customer behavior patterns, targeted promotions can be developed for certain customer groups to influence their booking behavior or even choice of hotel. The attributes that induce desired customer behaviors are referred as *inducing factors* in this study.

It is simple to identify the behavior inducing factors from the rules obtained by consequent control. Given the interesting rule $X \Rightarrow Y$ and S as a non empty subset of X, test the confidence and lift for interestingness of the new rule $S \Rightarrow Y$ (there is no need to test the support because “the support of an item set never exceeds the support of its own subset”). If the confidence or lift of the new rule is below the threshold, the **complement set** of S in X is considered as an inducing factor of S towards Y. For example, no positive correlation is observed between “travel purpose=Leisure” and “length of stay=Extended”; however, interesting rule *{repeated stay=1, travel purpose=Leisure}* ⇒ *{length of stay=Extended}* is obtained from the rule mining. It can therefore be said that “repeated stay=1” is the inducing factor of leisure travelers to stay for an extended period of time.

It is nearly impossible to identify all inducing factors due to the enormous number of candidate subsets for testing. In this study, only factors that have the potential to alter customers' original behavior patterns are considered interesting; that is, with the occurrence of the inducing factor, customers' preference will shift from behavior A to behavior B. Awareness of these inducing factors could enable hotels to develop new markets and even turn weakness into strength.

I) Choice of hotel

From our previous analysis it is observed that hotel A, B and C typically attract different types of hotel travelers. Hotel A is particularly favored by female, mainland Chinese travelers and those who stay for an extended length of time. Hotel B is most popular with foreign travelers and those who book from online platforms. Hotel C typically attracts male, Hong Kong travelers who travel alone and stay for only one night.

Exceptions are observed when certain influencing factors occur:

1. when traveling in pairs, female travelers are more frequently observed to choose hotel B rather than hotel A, where travel type=Pair can be addressed as an inducing factor for female travelers to select hotel B. Moreover, female leisure travelers aged between 18-39 years old are also more inclined to choose hotel B over hotel A.

Antecedent	Consequent	Supp	Conf	Lift
{gender=F, **travel type=Pair**}	{hotel code=B}	0.15	0.49	1.27

{gender=F, **age=18-29, travel purpose=Leisure**}	{hotel code=B}	0.06	0.48	1.24
{gender=F, **age=30-39, travel purpose=Leisure**}	{hotel code=B}	0.05	0.53	1.38

2. male travelers and overnight travelers from foreign countries in pairs are displaying higher preference towards hotel B instead of hotel C as shown in the rules below.

Antecedent	Consequent	Supp	Conf	Lift
{gender=M, **nationality=OTH, travel type=Pair**}	{hotel code=B}	0.10	0.59	1.53
{length of stay=Single, **nationality=OTH, travel type=Pair**}	{hotel code=B}	0.05	0.61	1.56

3. Mainland Chinese travelers, who typically prefer hotel A, appear to be more interested in hotel C if the travelers are male and on leisure trips.

Antecedent	Consequent	Supp	Conf	Lift
{**gender=M**, nationality=CN}	{hotel code=C}	0.05	0.35	1.81
{**gender=M**, nationality=CN, **travel purpose=Leisure**}	{hotel code=C}	0.05	0.36	1.86

II) Price sensitivity

In general, male travelers are less sensitive to price than female travelers when making accommodation choices. Young travelers aged below 30 are more inclined to economic hotel options. However, travelers in pairs display diverse preferences under different circumstances. It appears that, despite their cautiousness towards room price, pair travelers typically seek high-value accommodation quality and premium room type, especially for short term one-time experiences. From the rules below it can be concluded that discounts on premium rooms types could be particularly effective with pair travelers.

Antecedent	Consequent	Supp	Conf	Lift
{travel type=PAIR, **advance booking days=Short, holiday status=NONHOLIDAY**}	{average daily price=Low}	0.07	0.41	1.22
{travel type=PAIR, **age=18-29, holiday status=NONHOLIDAY**}	{average daily price=Low}	0.07	0.41	1.20
{travel type=PAIR, **length of stay=Single, holiday status=NONHOLIDAY**}	{average daily price=Low}	0.11	0.41	1.21
{travel type=PAIR, **advance booking days=Short, repeated stay=1**}	{room type=Premium}	0.01	0.10	1.20
{travel type=PAIR, **length of stay=Single, repeated**	{room type=Premium}	0.02	0.10	1.20

stay=1}				

III) Loyalty

Understanding the inducing factors that can potentially drive one-time hotel guests to become loyal customers would hold high value for hotel managers. By comparing the correlation between customers of different loyalty levels, a few interesting attributes are found that induce different behaviors between hotel guests in terms of repeated stay times.

1. Female travelers are more inclined to stay only once in the same hotel, especially those who book for more than two days. However, the frequency of female customers returning to the same hotel dramatically increases if they are only staying overnight. It is therefore possible to attract female customers to return by offering compatible promotions on overnight reservations. Moreover, on short advance bookings, female travelers show a higher tendency to return to the same hotel, while given long planning time, they are more interested in selecting a different hotel. Therefore, it is advised to send promotional information to female customers shortly before the promotions start.

Antecedent	Consequent	Supp	Conf	Lift
{gender=F}	{repeated stay=1}	0.32	0.82	1.25
{length of stay=Holiday, gender=F}	{repeated stay=1}	0.17	0.88	1.33
{**length of stay=Single**, gender=F}	{repeated stay=2}	0.02	0.16	1.89
{**length of stay=Single**, gender=F}	{repeated stay=3 more}	0.01	0.10	1.56
{advance booking days=21 Days, gender=F}	{repeated stay=1}	0.14	0.88	1.33
{**advance booking days=Short**, gender=F}	{repeated stay=2}	0.01	0.14	1.63

2. Similarly, behavior patterns in terms of returning stays of foreign travelers resemble those of female travelers; with shorter length of stay and shorter advance booking time, foreign travelers are more likely to repeatedly reserve the same hotel.

Antecedent	Consequent	Supp	Conf	Lift
{nationality=OTH}	{repeated stay=1}	0.37	0.83	1.26
{advance booking days=21 Days, nationality=OTH}	{repeated stay=1}	0.20	0.88	1.34
{length of stay=Holiday, nationality=OTH}	{repeated stay=1}	0.21	0.88	1.33
{**advance booking days=Short**, nationality=OTH}	{repeated stay=2}	0.01	0.13	1.49
{**length of stay=Single**, nationality=OTH}	{repeated stay=2}	0.02	0.15	1.75

Chapter 6: Conclusions

This chapter summarizes the findings from association rule mining in terms of the research questions. Limitations of the study and potential future research directions are also discussed in this chapter.

6.1 Summary

This study set out to explore the feasibility of leveraging data mining methods when applied to the transaction database of hotel property management systems to learn hotel selection criteria for travelers, and to answer three research questions:

- What factors, which are potentially important for travelers' hotel selection decisions, can be identified from the PMS database?
- How does association rule mining support travelers' hotel selection preference analysis?
- How could the trends and correlations revealed from association rules help hoteliers understand travelers' behaviors and react accordingly?

Previous research has suggested a variety of factors that influence the choices of travelers when making hotel reservations. The validated factors include demographic factors, travel purpose, travel type, and hotel attributes such as reputation, location, security, services, and attitudes of hotel staff. Some of the proposed hotel selection influencing factors in these studies are unobtainable by new customers before reservations are made, leading to confusion in discussion between hotel selection factors and hotel customer retention factors.

In articles reviewed for this study, questionnaires based on simulated hotel reservation scenarios, to rank the importance of multiple influencing factors, were typically used as the primary source of customers' hotel choice decisions. However, it has to be understood that hotel selection in real life is a more complicated decision-making process no matter how many influencing factors are listed in a questionnaire. Questionnaires based on simulated hotel-selection scenarios encourage respondents to conduct conscious and logical analysis to evaluate and compare the importance of each factor to arrive at their ranking or trade-off decisions. In fact, studies in other fields have shown that consumer purchase preferences are partially or even completely not associated with conscious processes (Martin & Morich, 2011; Erb, Bioy, & Hilton, 2002). In other words, by examining hotel guests' decision making processes on the basis of simulated hotel booking scenarios to infer their real hotel booking preferences is not a well-justified approach. In contrast, consumers' final purchase decisions made as the result of their selection process truly reflect their hotel reservation preferences regardless of consciousness during decision making. Thus a hotel PMS database as a record of consumers' booking decisions is a more reliable source from which to study their hotel selection criteria in this context.

In this study, 13 attributes that could potentially impact hotel choice decisions, including customer demographic factors, travel purpose, booking behaviors and hotel attributes, were extracted from the transaction database of a hotel PMS. Some attributes were not directly derived from the PMS raw data, but were generated based on a data conversion process or on reasonable inference. Controlled association rule mining confirms that all attributes independently or collaboratively influence customers' hotel booking behavior.

Association rule mining as a discovery-driven data mining approach has proven to be very effective in exploring consumer behavior patterns from a transactional dataset as in the PMS database provided for this study. Hotel selection factors are transformed into "items" in hotel customers' reservation "transactions" that can be processed by the Apriori algorithm used in this research. Predefined levels of support and confidence act as thresholds to identify frequent item sets from these transactions, and lift is used to discover the most interesting correlations between influencing factors and customer behaviors.

Inspired by Song et al. (2001), the attributes that potentially influence travelers' hotel selection decisions at the time of reservation are classified into three categories, namely traveler demographic factors, traveler behavior information and hotel information. More specific questions in regard to travelers' behavior patterns in hotel reservation can be addressed by conducting rule mining with antecedent or consequent control. Finally, more than 600 potentially interesting association rules are obtained to reflect underlying trends in the preferences of hotel customers. Some particularly interesting patterns have been discovered that could provide valuable information in helping hotel management understand the strengths and weaknesses of their hotels. Reasonable discoveries have been addressed to answer some the most studied questions in the hotel industry, e.g. "which customers favors a particular type of hotel?" and "which customers are more likely to be more loyal and repeatedly stay in the same hotel?" Moreover, by identifying particular *behavior inducing factors* from association rules this study provides a promising approach for hotels to purposefully increase favorable customer behaviors.

Among the discovered association rules, some interesting correlations emerged displaying observable differences in hotel selection choices made by different hotel guest groups under various booking scenarios. Some of these findings, now summarized, could be reasonably utilized for developing marketing and promotion strategies with specific purpose.

- As proposed by other studies, significant behavioral differences are validated between male and female hotel customers, especially the difference in their preference for length of stay, travel type, and price sensitivity. It can be noted from the rules that male travelers usually stay in the hotel for a short period of time, mostly for only one night. Female travelers typically spend a longer time on each trip. Unlike male travelers who frequently travel alone, female travelers have shown a perceivable pattern of traveling in pairs, possibly due to security and economic reasons. It can also be observed that male hotel guests show greater willingness to spend more money on accommodation, while female guests show more caution towards price. Patterns also suggest that women are more accustomed to making long-term travel plans and securing hotel bookings several weeks before departure.
- There is also evidence of significant differences in the hotel selection behaviors of travelers from different nations, especially in the choice of booking method, advance booking patterns, and length of stay. Mainland Chinese travelers are more dependent on travel agencies for hotel reservations, while foreign travelers are more accustomed to using online booking platforms. The rules also suggest that the farther the travelers are from the destination, the more they tend to make hotel reservations in advance and stay for a longer period of time at the destination.
- Young travelers, especially those under the age of 30, are found to frequently travel in pairs and are usually more sensitive to price. Travelers over the age of 50 are more likely to travel alone and are less cautious in spending patterns.
- Differences in hotel selection patterns and behavior patterns for business travelers and leisure travelers are particularly evident. Leisure travelers are easily affected by other influencing factors and showing various behavior patterns. Business travelers, on the other hand, show more steadfast tendencies towards hotel choices: they typically make

hotel reservations 7 days in advance, do not deliberately pursue low prices or premium accommodation options, frequently travel alone, and are more dependent on travel agencies to make hotel bookings.

Association rules mined using consequent control revealed guest groups that are typically attracted by a certain hotel, as well as customers that are more frequently behaving with certain patterns, e.g. book in advance, stay for extended days, and travel on holidays. Based on the behavior patterns and preferences learned, hotels could have better knowledge of their current market, strengths and weaknesses. The rules should enable hotel managers and marketing staff to target specific customer groups and develop tailored marketing strategies in order to raise market share, increase customer loyalty and develop new markets. In addition, a specific type of influencing factors - an inducing factor - can be found in association rules. The emergence of these factors can significantly increase the frequency occurrence of certain customer behaviors such as advance booking, repeated reservation, and extended stay. Understanding these behavior-inducing factors may provide an approach for hotels to direct customer behaviors towards these interest as suggested in the analytics of the last chapter. However, the effectiveness of the approach needs to be further verified in practice.

In addition, the behavior patterns and hotel choice preferences identified in this study could be utilized by more than just hotel managers. Travelers' hotel preference information learned from the association rules could also be instructive for other customers in filtering and screening desirable hotels. With hotel choice patterns derived from previous customers' reservation data, online booking platforms and search engines could provide additional screening options for users to identify hotels that are most popular for customers with same travel purpose or travel type, supporting more effective, customized recommendation.

6.2 Limitations

Data mining using the database of hotel property management system has key advantages in reflecting the actual final results of hotel customers' decision-making processes and avoiding potential subjective differentiations in simulated booking scenarios. However, information contained in a PMS database is restrained by limitations in data collection.

Some potentially important influencing factors in hotel selection, such as the customer's lifestyle, income level and education level, and customer reviews, are missing from a PMS database. In addition, the information accuracy of the database can be affected by various factors, including software design, proficiency of information input, and human error. For example, duplicate guest profiles for the same customer may lead to errors in repeated stay information; mistakes in registration may cause inaccuracies in the number of guests. These mistakes will have an impact on the validity of any data mining results. Moreover, different hotels have diverse definitions of room type, price code, and market segmentation; therefore, substantial effort is required to preprocess the data when mining patterns from multiple hotels. Finally, hotel attributes such as geographic location, nearby attractions, traffic conditions, brand reputation and competitor information require additional access and cannot be obtained directly from the database, adding to the complexity of research such as that conducted and reported here.

Using reservation records in learning hotel customers' selection criteria tends to minimize the impact of customers' access to hotel information through other channels before booking. This study excludes attributes that are unlikely to be obtained by travelers prior to arrival such as the enthusiasm of hotel staff, the quality of furnishings, and the condition of bedding and shower facilities. However, customer reviews on travel websites and online booking platforms have become an important source of such information. Despite the fact that customer evaluations can themselves be highly subjective and potentially doubtful in authenticity, studies have shown that customer evaluations and "word of mouth" can have an important and even critical impact on travelers' hotel choices (Duan et al., 2013; Schuckert, Liu, & Law, 2015). It is therefore asserted here that even more accurate and comprehensive conclusions could be drawn from a combined analysis of customer review data and hotel reservation data.

Another limitation of this study comes from the limited choice of hotels. The three hotels studied in this thesis are similar in scale, rating, and price level. Therefore, conclusions on customers' hotel selection patterns towards hotel attributes are relatively weak. If the study were to investigate the impact of multiple hotel attributes, including location, brand reputation, star rating, nearby attractions, traffic condition, and area security level by controlling variables, it would require a much larger sample base of hotels in order to draw comprehensive conclusions.

The choice of support and confidence thresholds in this study is set at a relatively low level, which may lead to a plethora of rules presented in this thesis. High support and confidence can limit the number of rules generated from mining; however, high thresholds will screen out potentially interesting rules containing rare items. As an eclectic choice, relatively low thresholds of support and confidence are applied in this study to allow enough rules to be generated, and lift is used to screen out rules that reflect distinctive correlations between attributes and customer behaviors. The excessive amount of rules reveal many underlying patterns in customers' hotel choices, but may as well decrease the value in terms of decision making.

6.3 Future directions

1. Dynamic association rule mining

Awareness of changes in customer behavior patterns is potentially valuable to hoteliers. Hotels are encouraged to react dynamically to the ever-changing trends in customer preferences to stay competitive. Association rule mining has proven to be useful in analyzing dynamic databases to explore changes in customer behaviors (Shenoy et al., 2003). In research in other fields, emerging and disappearing trends of customer behaviors have been identified by comparing association rules generated at time k and $k+t$ to find correlations that have strengthened or weakened (Wu, Chen, & Chen, 2005). By exploiting a similar methodology in hotel reservation data it would be possible to learn customer reactions towards specific events, such as an epidemic outbreak (e.g. SARS outbreak in 2004), an economic crisis (e.g. global crisis in 2008), and develop response solutions before the next strike. Dynamic association rules can also be used in examining customers' behavior change towards specific promotion programs or marketing campaigns. It would be interesting to leverage dynamic association rule mining in a PMS database to discover pattern changes over time in future research.

2. Negative association rule mining

Typical association rules only consider items that occur frequently in the same transactions, and these rules are called positive association rules. Negative association rules also consider the same item base, but only look at items that are absent from the transactions (Kotsiantis & Kanellopoulos, 2006). Negative association rules are useful in market basket analysis and can be used to identify conflicting products or complementary products (Kotsiantis & Kanellopoulos, 2006). By using negative association rules in the data of this study, it could be possible to identify hotel attributes that are particularly unattractive to customers and customer groups that typically avoid certain behaviors. Negative association rules could therefore provide information to help hotel management understand customer preferences and behavior patterns from a different dimension, and allow hotels to develop corresponding strategies to improve hotel service and marketing plans.

References

Abaya, S. A. (2012). Association rule mining based on Apriori algorithm in minimizing candidate generation. *International Journal of Scientific & Engineering Research*, *3*(7), 1-4.

Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *Acm sigmod record* (Vol. 22, No. 2, pp. 207-216). ACM.

Ananth, M., DeMicco, F. J., Moreo, P. J., & Howey, R. M. (1992). Marketplace lodging needs of mature travelers. *Cornell Hotel and Restaurant Administration Quarterly*, *33*(4), 12-24.

Baber, R., & Kaurav, R. P. S. (2015). CRITERIA FOR HOTEL SELECTION: A STUDY OF TRAVELLERS. *Pranjana: The Journal of Management Awareness*, *18*(2).

Belonax Jr, J. A., & Mittelstaedt, R. A. (1978). EVOKED SET SIZE AS A FUNCTION OF NUMBER OF CHOICE CRITERIA AND INFORMATION VARIABILITY. *Advances in consumer research*, *5*(1).

Brijs, T., Swinnen, G., Vanhoof, K., & Wets, G. (1999, August). Using association rules for product assortment decisions: A case study. In *KDD* (Vol. 99, pp. 254-260).

Brijs, T., Swinnen, G., Vanhoof, K., & Wets, G. (2004). Building an association rules framework to improve product assortment decisions. *Data Mining and Knowledge Discovery*, *8*(1), 7-23.

Cahill, D. J. (1997). Target marketing and segmentation: valid and useful tools for marketing. *Management Decision*, *35*(1), 10-13.

Cai, L. A., Lehto, X. Y., & O'leary, J. (2001). Profiling the US-bound Chinese travelers by purpose of trip. *Journal of Hospitality & Leisure Marketing*, *7*(4), 3-16.

Callan, R. J. (1994). Development of a Framework for the Determination of Attributes used for Hotel Selection–-Indications from Focus Group and In-Depth Interviews. *Hospitality Research Journal*, *18*(2), 53-74.

Callan, R. J. (1998). Attributional analysis of customers' hotel selection criteria by UK grading scheme categories. *Journal of Travel Research*, *36*(3), 20-34.

Casillas, J., Martinez-Lopez, F., & Martinez, F. (2004). Fuzzy association rules for estimating consumer behaviour models and their application to explaining trust in internet shopping. *Fuzzy Economic Review*, *9*(2), 3-26.

Chan, E. S., & Wong, S. C. (2006). Hotel selection: When price is not the issue. *Journal of Vacation Marketing*, *12*(2), 142-159.

Chareyron, G., Da-Rugna, J., & Raimbault, T. (2014, October). Big data: A new challenge for tourism. In *2014 IEEE International Conference on Big Data (Big Data)* (pp. 5-7). IEEE.

Chen, M. C., Chiu, A. L., & Chang, H. H. (2005). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, *28*(4), 773-781.

Choi, T. Y., & Chu, R. (2001). Determinants of hotel guests' satisfaction and repeat patronage in the Hong Kong hotel industry. *International Journal of Hospitality Management*, *20*(3), 277-297.

Chow, K. E., Garretson, J. A., & Kurtz, D. L. (1995). An exploratory study into the purchase decision process used by leisure travelers in hotel selection. *Journal of Hospitality & Leisure*

Marketing, *2*(4), 53-72.

Chu, R. K., & Choi, T. (2000). An importance-performance analysis of hotel selection factors in the Hong Kong hotel industry: a comparison of business and leisure travellers. *Tourism management*, *21*(4), 363-377.

Cobanoglu, C., Corbaci, K., Moreo, P. J., & Ekinci, Y. (2003). A comparative study of the importance of hotel selection components by Turkish business travelers. *International journal of hospitality & tourism administration*, *4*(1), 1-22.

Discoverhongkong.com. (n.d.). *Outlying Islands.* Retrieved April 23, 2019, from http://www.discoverhongkong.com/eng/see-do/great-outdoors/outlying-islands/index.jsp

Duan, W., Cao, Q., Yu, Y., & Levy, S. (2013, January). Mining online user-generated content: using sentiment analysis technique to study hotel service quality. In *2013 46th Hawaii International Conference on System Sciences* (pp. 3119-3128). IEEE.

Erb, H. P., Bioy, A., & Hilton, D. J. (2002). Choice preferences without inferences: Subconscious priming of risk attitudes. *Journal of Behavioral Decision Making*, *15*(3), 251-262.

Fayyad, U. M., Piatetsky-Shapiro, G., & Uthurusamy, R. (2003). Summary from the KDD-03 panel: data mining: the next 10 years. *ACM Sigkdd Explorations Newsletter*, *5*(2), 191-196.

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (pp. 59-139). New York: Springer.

Gilbert, D., & Tsao, J. (2000). Exploring Chinese cultural influences and hospitality marketing relationships. *International Journal of Contemporary Hospitality Management*, *12*(1), 45-54.

Godinho, S., Prada, M., & Garrido, M. V. (2016). Under pressure: An integrative perspective of time pressure impact on consumer decision-making. *Journal of International Consumer Marketing*, *28*(4), 251-273.

Ha, S. H., & Park, S. C. (1998). Application of data mining tools to hotel data mart on the Intranet for database marketing. *Expert Systems with Applications*, *15*(1), 1-31.

Hahsler, M., Grün, B., & Hornik, K. (2007). Introduction to arules–mining association rules and frequent item sets. *SIGKDD Explor*, *2*(4), 1-28.

Hájek, P., Havel, I., & Chytil, M. (1966). The GUHA method of automatic hypotheses determination. *Computing*, *1*(4), 293-308.

Heaton, J. (2016, March). Comparing dataset characteristics that favor the Apriori, Eclat or FP-Growth frequent itemset mining algorithms. In *SoutheastCon 2016* (pp. 1-7). IEEE.

Hongkongextras.com. (n.d.). *New, future and renamed hotels (January 2018 to end 2020).* Retrieved March 28, 2019, from http://www.hongkongextras.com/_new_hotels.html

Hsieh, N. C. (2004). An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert systems with applications*, *27*(4), 623-633.

Jang, Y., Chen, C. C., & Miao, L. (2019). Last-minute hotel-booking behavior: The impact of time on decision-making. *Journal of Hospitality and Tourism Management*, *38,* 49-57.

Jeffrey, D., & Barden, R. R. (2000). An analysis of daily occupancy performance: a basis for effective hotel marketing?. *International Journal of Contemporary Hospitality Management*, *12*(3), 179-189.

Jones, P., & Chen, M. M. (2011). Factors determining hotel selection: Online behaviour by leisure travellers. *Tourism and Hospitality Research*, *11*(1), 83-95.

Kandampully, J., & Suhartanto, D. (2000). Customer loyalty in the hotel industry: the role of customer satisfaction and image. *International journal of contemporary hospitality management*, *12*(6), 346-351.

Kavitha, M., & Selvi, S. T. (2016). Comparative Study on Apriori Algorithm and Fp Growth Algorithm with Pros and Cons. *International Journal of Computer Science Trends and Technology (I JCS T)–Volume*, *4*.

Kaynak, E., & Yavas, U. (1981). Segmenting the tourism market by purpose of trip: A profile analysis of visitors to Halifax, Canada. *International Journal of Tourism Management*, *2*(2), 105-112.

Kotsiantis, S., & Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 71-82.

Kucukusta, D., Pang, L., & Chui, S. (2013). Inbound travelers' selection criteria for hotel spas in Hong Kong. *Journal of Travel & Tourism Marketing*, *30*(6), 557-576.

Lepisto, L. R., & McCleary, K. W. (1988). The effect of multiple measures of age in segmenting hotel markets. *Hospitality Education and Research Journal*, *12*(2), 91-98.

Lewis, R. C. (1984). The basis off hotel selection. *Cornell hotel and restaurant administration quarterly*, *25*(1), 54-69.

Lewis, R. C. (1985). Predicting hotel choice: The factors underlying perception. *Cornell Hotel and Restaurant Administration Quarterly*, *25*(4), 82-96.

Lewis, R. C., & Chambers, R. E. (1989). *Marketing leadership in hospitality. Foundations and practices*. Van Nostrand Reinhold.

Li, G., Law, R., Vu, H. Q., & Rong, J. (2013). Discovering the hotel selection preferences of Hong Kong inbound travelers using the Choquet Integral. *Tourism Management*, *36*, 321-330.

Li, G., Law, R., Vu, H. Q., Rong, J., & Zhao, X. R. (2015). Identifying emerging hotel preferences using emerging pattern mining technique. *Tourism management*, *46*, 311-321.

Liao, S. H., & Chen, Y. J. (2004). Mining customer knowledge for electronic catalog marketing. *Expert Systems with Applications*, *27*(4), 521-532.

Liao, S. H., Chu, P. H., Chen, Y. J., & Chang, C. C. (2012). Mining customer knowledge for exploring online group buying behavior. *Expert Systems with Applications*, *39*(3), 3708-3716.

Lin*, H. S. (2008). Private love in public space: love hotels and the transformation of intimacy in contemporary Japan. *Asian Studies Review*, *32*(1), 31-56.

Lockyer, T. (2002). Business guests’ accommodation selection: the view from both sides. *International Journal of Contemporary Hospitality Management*, *14*(6), 294-300.

Lockyer, T. (2005a). The perceived importance of price as one hotel selection dimension. *Tourism Management*, *26*(4), 529-537.

Lockyer, T. (2005b). Understanding the dynamics of the hotel accommodation purchase decision. *International Journal of contemporary hospitality management*, *17*(6), 481-492.

Magnini, V. P., Honeycutt Jr, E. D., & Hodge, S. K. (2003). Data mining for hotel firms: Use and limitations. *Cornell Hotel and Restaurant Administration Quarterly*, *44*(2), 94-105.

Maletic, J. I., & Marcus, A. (2009). Data cleansing: A prelude to knowledge discovery. In *Data mining and knowledge discovery handbook* (pp. 19-32). Springer, Boston, MA.

Marshall, A. (1993). Safety tops guest's priority list; sell security as No. 1 amenity. *Hotel & Motel*

Management, *208*(11), 21-21.

Martilla, J. A., & James, J. C. (1977). Importance-performance analysis. *Journal of marketing*, *41*(1), 77-79.

Martin, N., & Morich, K. (2011). Unconscious mental processes in consumer choice: Toward a new model of consumer behavior. *Journal of Brand Management*, *18*(7), 483-505.

Matthias, B. (2006). *The hotel as setting in early twentieth-century German and Austrian literature: checking in to tell a story*. Harvard University Press.

Mccleary, K. W., Choi, B. M., & Weaver, P. A. (1998). A comparison of hotel selection criteria between US and Korean business travelers. *Journal of Hospitality & Tourism Research*, *22*(1), 25-38.

McCleary, K. W., Weaver, P. A., & Hutchinson, J. C. (1993). Hotel selection factors as they relate to business travel situations. *Journal of Travel Research*, *32*(2), 42-48.

McCleary, K. W., Weaver, P. A., & Lan, L. (1994). Gender-based differences in business travelers' lodging preferences. *Cornell Hotel and Restaurant Administration Quarterly*, *35*(2), 51-58.

Müller, H., & Freytag, J. C. (2005). *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. Für Informatik.

Myers, J. H., & Alpert, M. I. (1968). Determinant buying attitudes: meaning and measurement. *Journal of Marketing*, *32*(4_part_1), 13-20.

Ng, K., & Liu, H. (2000). Customer retention via data mining. *Artificial Intelligence Review*, *14*(6), 569-590.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, *36*(2), 2592-2602.

Oberoi, U., & Hales, C. (1990). Assessing the quality of the conference hotel service product: towards an empirically based model. *Service Industries Journal*, *10*(4), 700-721.

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). Servqual: A multiple-item scale for measuring consumer perc. *Journal of retailing*, *64*(1), 12.

Ramanathan, U., & Ramanathan, R. (2011). Guests' perceptions on factors influencing customer loyalty: An analysis for UK hotels. *International Journal of Contemporary Hospitality Management*, *23*(1), 7-25.

Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and tourism online reviews: Recent trends and future directions. *Journal of Travel & Tourism Marketing*, *32*(5), 608-621.

Schwartz, Z. (2006). Advanced booking and revenue management: Room rates and the consumers' strategic zones. *International Journal of Hospitality Management*, *25*(3), 447-462.

Shenoy, P. D., Srinivasa, K. G., Venugopal, K. R., & Patnaik, L. M. (2003, April). Evolutionary approach for mining association rules on dynamic databases. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 325-336). Springer, Berlin, Heidelberg.

Sohrabi, B., Vanani, I. R., Tahmasebipur, K., & Fazli, S. (2012). An exploratory analysis of hotel selection factors: A comprehensive survey of Tehran hotels. *International Journal of Hospitality Management*, *31*(1), 96-106.

Solanki, S. K., & Patel, J. T. (2015, February). A survey on association rule mining. In *2015 Fifth International Conference on Advanced Computing & Communication Technologies* (pp. 212-216). IEEE.

Song, H. S., kyeong Kim, J., & Kim, S. H. (2001). Mining the change of customer behavior in an internet shopping mall. *Expert Systems with Applications*, *21*(3), 157-168.

Statistics.gov.hk. (2018). *Hong Kong in figures 2018 edition*. Retrieved from https://www.statistics.gov.hk/

Tanford, S., Raab, C., & Kim, Y. S. (2012). Determinants of customer loyalty and purchasing behavior for full-service and limited-service hotels. *International Journal of Hospitality Management*, *31*(2), 319-328.

Taninecz, G. (1990). Business traveller survey. *Hotel and Motel Management*, *57*(June), 29-32.

Tsai, H., Yeung, S., & Yim, P. H. (2011). Hotel selection criteria used by mainland Chinese and foreign individual travelers to Hong Kong. *International journal of hospitality & tourism administration*, *12*(3), 252-267.

Webb, G. I. (1989). A machine learning approach to student modelling. In *Proceedings of the Third Australian Joint Conference on Artificial Intelligence* (pp. 195-205).

Wilensky, L., & Buttle, F. (1988). A multivariate analysis of hotel benefit bundles and choice trade-offs. *International Journal of Hospitality Management*, *7*(1), 29-41.

Wong, K. W., Zhou, S., Yang, Q., & Yeung, J. M. S. (2005). Mining customer value: From association rules to direct marketing. *Data Mining and Knowledge Discovery*, *11*(1), 57-79.

Wu, R. C., Chen, R. S., & Chen, C. C. (2005, July). Data mining application in customer relationship management of credit card business. In *29th Annual International Computer Software and Applications Conference (COMPSAC'05)* (Vol. 2, pp. 39-40). IEEE.

Xiang, Z., Schwartz, Z., Gerdes Jr, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction?. *International Journal of Hospitality Management*, *44*, 120-130.

Yavas, U., & Babakus, E. (2005). Dimensions of hotel choice criteria: congruence between business and leisure travelers. *International Journal of Hospitality Management*, *24*(3), 359-367.

Yoo, J. J. E., McKercher, B., & Mena, M. (2004). A cross-cultural comparison of trip characteristics: International visitors to Hong Kong from Mainland China and USA. *Journal of Travel & Tourism Marketing*, *16*(1), 65-77.

Zaman, M., Botti, L., & Thanh, T. V. (2016). Weight of criteria in hotel selection: An empirical illustration based on TripAdvisor criteria. *European journal of tourism research*, *13*(1), 132-138.

Appendix 1 – Sample of raw transaction data from PMS

STAYID	CONFIRMATION	TA NAME	STATUS	CHECK-IN DATE	CHECK-OUT DATE	BOOKING DATE	ROOM TYPE	RATE CODE	PRICE	BOOKING SOURCE	ADULTS	CHILD	GUEST ID	SHARE CODE	GENDER	NATIONALITY	DATE OF BIRTH(YYYY-M-D)
2	1		C	2015/10/7	2015/10/8	2015/10/7	AA	HSE	0	HSE	1	0		M	M		0-0-0
8	7		O	2015/10/7	2015/10/9	2015/10/7	AB	HSE	0	HSE	1	0	6	M	M	HK	0-0-0
10	8		O	2015/10/9	2015/10/11	2015/10/7	AC	COMP	0	COM	1	0	12	M	F	CN	0-0-0
10	8		O	2015/10/9	2015/10/11	2015/10/7	AC	COMP	0	COM	1	0	11	P	F	MO	1936/10/13
104	9		O	2015/10/9	2015/10/11	2015/10/7	AC	COMP	0	COM	1	0	13	M	M	CN	0-0-0
155	25		C	2015/10/9	2015/10/10	2015/10/9	AD	OPEN	500	OPR	1	0		M	M		0-0-0
156	26		C	2015/10/9	2015/10/10	2015/10/9	AE	OPEN+	500	OPR	1	0		M	M		0-0-0
157	27		C	2015/10/9	2015/10/10	2015/10/9	AB	BAR	1500	DIR	1	0		M	F		0-0-0
158	28		C	2015/10/9	2015/10/10	2015/10/9	AF	BAR+	1000	DIR	1	0		M	M		0-0-0
159	29		C	2015/10/9	2015/10/10	2015/10/9	AF	COMP	0	COM	1	0		M	M		0-0-0
160	30		C	2015/10/9	2015/10/10	2015/10/9	AB	HSE	0	HSE	1	0		M	M		0-0-0
169	33		O	2015/10/10	2015/10/11	2015/10/10	AB	BAR+	2000	WKI	2	1	16	M	M	CN	1982/10/30
169	33		O	2015/10/10	2015/10/11	2015/10/10	AB	BAR+	2000	WKI	2	1	17	P	F	CN	1989/7/16
167	31		O	2015/10/12	2015/10/17	2015/10/10	AB	BAR+	0	OWN	1	0		M	F	HK	1956/5/19
199	52		O	2015/10/13	2015/10/14	2015/10/13	AC	COMP	0	COM	1	0	19	M	M	HK	1985/6/8
199	52		O	2015/10/13	2015/10/14	2015/10/13	AC	COMP	0	COM	1	0	20	P	F	HK	1957/2/17
200	53		O	2015/10/13	2015/10/16	2015/10/13	AB	HSE	0	HSE	1	0	21	M	M	HK	1966/10/19
202	54		O	2015/10/13	2015/10/15	2015/10/13	AB	HSE	0	HSE	1	0		M	M	HK	0-0-0
143	14		O	2015/10/14	2015/10/15	2015/10/7	AB	COMP	0	COM	1	0	22	M	M	ID	1971/10/21
414	73		O	2015/10/15	2015/10/17	2015/10/15	AF	HSE	0	HSE	1	0		M	M	HK	0-0-0
304	60		O	2015/10/16	2015/10/17	2015/10/14	AB	HSE	0	HSE	1	0		M	M	HK	0-0-0
447	80		O	2015/10/17	2015/10/18	2015/10/16	AC	COMP	0	COM	2	2	27	M	M	HK	1973/9/1
447	80		O	2015/10/17	2015/10/18	2015/10/16	AC	COMP	0	COM	2	2	26	P	F	HK	1975/7/26
447	80		O	2015/10/17	2015/10/18	2015/10/16	AC	COMP	0	COM	2	2	24	P	M	HK	1971/1/13
448	81		O	2015/10/17	2015/10/18	2015/10/16	AC	COMP	0	COM	2	1		M	M	HK	1971/1/13
412	72		O	2015/10/19	2015/10/20	2015/10/15	AB	COMP	0	COM	1	0	22	M	M	ID	1971/10/21
572	100		C	2015/10/19	2015/10/20	2015/10/19	AC	HSE	0	HSE	2	0		M	F	HK	0-0-0
572	100		C	2015/10/19	2015/10/20	2015/10/19	AC	HSE	0	HSE	2	0		P	M	HK	0-0-0
577	103		O	2015/10/19	2015/10/20	2015/10/19	AB	HSE	0	HSE	1	0		M	M	HK	1967/11/5
411	71		O	2015/10/20	2015/10/21	2015/10/15	AB	COMP	0	COM	2	0	29	M	F	HK	1977/1/24
411	71		O	2015/10/20	2015/10/21	2015/10/15	AB	COMP	0	COM	2	0	30	P	M	HK	1967/7/7
466	92		O	2015/10/20	2015/10/21	2015/10/18	AF	COMP	0	OWN	2	0	36	M	F	HK	1967/9/25
467	93		O	2015/10/20	2015/10/21	2015/10/18	AC	COMP	0	OWN	2	0	37	M	M	HK	1966/9/19
468	94		O	2015/10/20	2015/10/21	2015/10/18	AB	COMP	0	OWN	2	0	31	M	M	HK	1962/12/29
468	94		O	2015/10/20	2015/10/21	2015/10/18	AB	COMP	0	OWN	2	0	32	P	F	HK	1963/11/1
668	104		O	2015/10/20	2015/10/24	2015/10/20	AA	HSE	0	HSE	2	0	34	M	F	HK	0-0-0
668	104		O	2015/10/20	2015/10/24	2015/10/20	AA	HSE	0	HSE	2	0	35	P	M	HK	0-0-0
669	105		O	2015/10/20	2015/10/21	2015/10/20	AB	HSE	0	HSE	1	0	6	M	M	HK	0-0-0
678	114		O	2015/10/20	2015/10/21	2015/10/20	AB	HSE	0	HSE	1	0	33	M	F	HK	1973/10/4
677	113		O	2015/10/21	2015/10/23	2015/10/20	AB	HSE	0	HSE	1	0	21	M	M	HK	1966/10/19
881	121		O	2015/10/21	2015/10/26	2015/10/21	AB	COMP	0	COM	1	0	38	M	F	HK	1979/2/10
884	122		O	2015/10/21	2015/10/23	2015/10/21	AB	COMP	0	COM	1	0	40	M	M	HK	1984/2/28

Appendix 2 – Sample of database after data preparation

Hotel code	Booking channel	Length of stay	Advance booking days	Room type	Average daily price	Repeated stay	Gender	Nationality	Age	Travel type	Travel purpose	Holiday status
A	Direct	Single	Same Day	Standard	High	3 more	M	HK	18-29	SINGLE	Leisure	NONHOLIDAY
A	Direct	Single	Same Day	Standard	High	1	M	CN	40-49	SINGLE	Leisure	NONHOLIDAY
A	Direct	Single	Same Day	Standard	High	1	M	CN	30-39	PAIR	Leisure	NONHOLIDAY
A	Direct	Single	Same Day	Standard	High	1	M	CN	40-49	PAIR	Leisure	NONHOLIDAY
A	TA	Single	Short	Standard	High	1	M	HK	40-49	PAIR	Leisure	NONHOLIDAY
A	TA	Single	Short	Standard	High	1	F	CN	40-49	PAIR	Leisure	NONHOLIDAY
A	Direct	Single	Short	Standard	High	1	M	OTH	40-49	SINGLE	Leisure	NONHOLIDAY
A	Direct	Single	Same Day	Standard	High	1	F	CN	30-39	PAIR	Leisure	NONHOLIDAY
A	Direct	Single	Same Day	Standard	High	1	M	HK	30-39	PAIR	Leisure	NONHOLIDAY
A	Direct	Single	Same Day	Standard	High	2	M	HK	30-39	PAIR	Leisure	NONHOLIDAY
A	Direct	Single	Same Day	Standard	High	1	F			PAIR	Leisure	NONHOLIDAY
A	TA	Single	Same Day	Standard	High	1	M	CN	30-39	FAMILY	Leisure	NONHOLIDAY
A	TA	Single	Same Day	Standard	High	1	F	CN	30-39	FAMILY	Leisure	NONHOLIDAY
A	OTA	Single	Same Day	Standard	High	2	M	OTH	18-29	PAIR	Leisure	NONHOLIDAY
A	OTA	Single	Same Day	Standard	High	1	M	OTH	40-49	PAIR	Leisure	NONHOLIDAY
A	TA	Single	Same Day	Standard	High	1	M	CN	30-39	PAIR	Leisure	NONHOLIDAY
A	TA	Single	Same Day	Standard	High	1	M	CN	30-39	PAIR	Leisure	NONHOLIDAY
B	OTA	Holiday	Short	Premium	Low	1	F	OTH	30-39	PAIR	Leisure	NONHOLIDAY
B	OTA	Holiday	Short	Economic	Low	1	F	OTH	50 over	PAIR	Leisure	NONHOLIDAY
B	OTA	Holiday	21 Days	Economic	Medium	1	F	OTH	30-39	PAIR	Leisure	NONHOLIDAY
B	TA	Holiday	7 Days	Standard	Low	1	M	OTH	30-39	PAIR	Leisure	NONHOLIDAY
B	TA	Holiday	7 Days	Standard	Low	1	M			PAIR	Leisure	NONHOLIDAY
B	TA	Extended	21 Days	Standard	Medium	1	M	OTH	18-29	PAIR	Leisure	NONHOLIDAY
B	TA	Extended	21 Days	Standard	Medium	1	F	OTH	18-29	PAIR	Leisure	NONHOLIDAY
B	Direct	Holiday	Short	Standard	Low	1	F	OTH	50 over	PAIR	Leisure	NONHOLIDAY
B	Direct	Holiday	Short	Standard	Low	1	M	OTH	18-29	PAIR	Leisure	NONHOLIDAY
B	Direct	Single	Short	Economic	Medium	1	M	HK	40-49	PAIR	Leisure	NONHOLIDAY
B	Direct	Holiday	7 Days	Premium	Low	1	M	OTH	40-49	PAIR	Leisure	NONHOLIDAY
B	Direct	Holiday	7 Days	Premium	Low	1	M			PAIR	Leisure	NONHOLIDAY
B	Direct	Single	Same Day	Standard	Medium	1	M	OTH	40-49	SINGLE	Leisure	NONHOLIDAY
B	Direct	Single	Same Day	Standard	Low	1	F	HK	30-39	SINGLE	Leisure	NONHOLIDAY
C	OTA	Single	21 Days	Standard	Medium		M	CN	30-39	SINGLE	Leisure	NONHOLIDAY
C	OTA	Single	Same Day	Standard	Medium		M	HK	50 over	SINGLE	Leisure	NONHOLIDAY
C	TA	Single	Short	Standard	Low		M	CN	40-49	PAIR	Leisure	NONHOLIDAY
C	Direct	Single	Same Day	Standard	Medium		M	HK	40-49	PAIR	Leisure	NONHOLIDAY
C	OTA	Single	Short	Standard	High		M	HK	18-29	SINGLE	Leisure	HOLIDAY
C	Direct	Single	Short	Standard	High		M	HK	50 over	PAIR	Leisure	HOLIDAY
C	Direct	Single	Same Day	Economic	Low		M	HK		SINGLE	Leisure	NONHOLIDAY
C	Direct	Single	Short	Standard	Low		M	HK	40-49	PAIR	Leisure	NONHOLIDAY
C	OTA	Single	Short	Standard	Medium		M	HK	18-29	SINGLE	Leisure	NONHOLIDAY
C	OTA	Single	Same Day	Standard	Medium		M	HK	50 over	SINGLE	Leisure	NONHOLIDAY
C	OTA	Single	7 Days	Standard	High		M	HK		PAIR	Leisure	NONHOLIDAY