# Hotel Booking Demand Business Analytics Using EDA And Machine learning Techniques

Thanh Tran x Nghia Nguyen x Mohsin Qureshi

May 07, 2023

LUT University

# Table of Contents

# 1. Introduction

This report outlines our procedures, methodology and findings of our data mining project. The chosen dataset contains information about booking transactions from two hotels covering customer demographics, cancellation, charge rates, selling channels, etc…

Our goal was to identify patterns and trends in the data, understand the dynamics of the hotel demand, and use them to answer some business questions on three topics: customer segmentation, pricing strategies and demand distribution.

To achieve this, we applied exploratory data analysis and visualization to reveal patterns in the data and developed classification and clustering models to uncover some hidden insights with intelligent features.

# 2. Data summary

## 2.1. Definitions

The dataset consists of 119, 390 booking transactions of two anonymous hotel namely “City Hotel” and “Resort Hotel”.

Thirty-two attributes are available in the dataset:

| Column Name | Description |
|---|---|
| **hotel** | Hotel (Resort Hotel or City Hotel) |
| **is_canceled** | Value indicating if the booking was cancelled (1) or not (0) |
| **lead_time** | Number of days that elapsed between the entering date of the booking into the PMS and the arrival date |
| **arrival_date_year** | Year of arrival date |
| **arrival_date_month** | Month of arrival date |
| **arrival_date_week_number** | Week number of year for arrival date |
| **arrival_date_day_of_month** | Day of arrival date |

| | |
|---|---|
| **stays_in_weekend_nights** | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel |
| **stays_in_week_nights** | Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel |
| **adults** | Number of adults |
| **children** | Number of children |
| **babies** | Number of babies |
| **meal** | Type of meal booked |
| **country** | Country of origin |
| **market_segment** | Market segment designation |
| **distribution_channel** | Booking distribution channel |
| **is_repeated_guest** | Value indicating if the booking name was from a repeated guest (1) or not (0) |
| **previous_cancellations** | Number of previous bookings that were cancelled by the customer prior to the current booking |
| **previous_bookings_not_canceled** | Number of previous bookings not cancelled by the customer prior to the current booking |
| **reserved_room_type** | Code of room type reserved |
| **assigned_room_type** | Code for the type of room assigned to the booking |
| **booking_changes** | Number of changes/amendments made to the booking |
| **deposit_type** | Indication on if the customer made a deposit to guarantee the booking |
| **agent** | ID of the travel agency that made the booking |
| **company** | ID of the company/entity that made the booking or responsible for paying the booking |
| **days_in_waiting_list** | Number of days the booking was in the waiting list before it was confirmed to the customer |
| **customer_type** | Type of booking, assuming one of four categories |
| **adr** | Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights |

| | |
|---|---|
| **required_car_parking_spaces** | Number of car parking spaces required by the customer |
| **total_of_special_requests** | Number of special requests made by the customer |
| **reservation_status** | Reservation last status, assuming one of three categories |
| **reservation_status_date** | Date at which the last status was set. |

## 2.2. Statistical analysis

### 2.2.1. Data summary

The dataset covers a roughly three-year timeframe from Oct 2014 to Sep 2017. The number of records from "City Hotel" nearly doubled that of "Resort hotel". Cancellation rate was 37% according to the total number of bookings and the number of cancelled bookings.

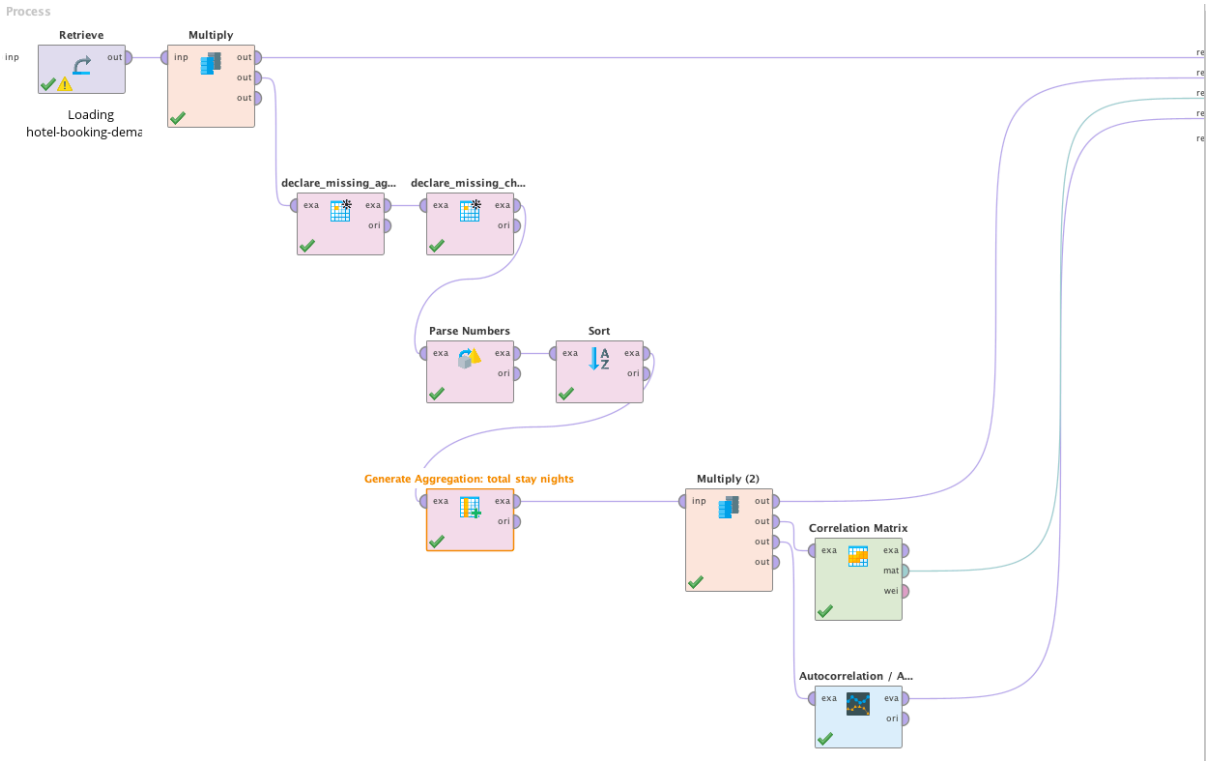
The average lead time was 107 days with values ranging from 0 to more that 700 days. Demand of 2016 was highest, and it nearly tripled that of 2015. Overall demand had seasonal components which were low during winter months from November to January, high during summer and autumn. Highest demands were spotted in August.

The customers were from 178 different countries. Portuguese customers accounted for over 40% of bookings. Bookings from repeated customers accounted for only 3.2% of the total.

Average daily rate ranged from -6.38 to 5400 euros with the average of 101.8 euros and deviation of 50.5 euros.

### 2.2.1. Process

To perform statistical analysis, the data was imported to the workspace. Missing values were declared for attributes where missing values existed. After that, some misunderstood nominal attributes were parsed to its right numeric type. In order to facilitate the time series analysis, the dataset was sorted by datetime attribute. A new aggregated field was created by summing the two columns "stays_in_weekend_nights" and "stays_in_week_nights" to compute the total nights of the stays. Eventually, Pearson correlation and autocorrelation were examined.




### 2.2.2. Data integrity

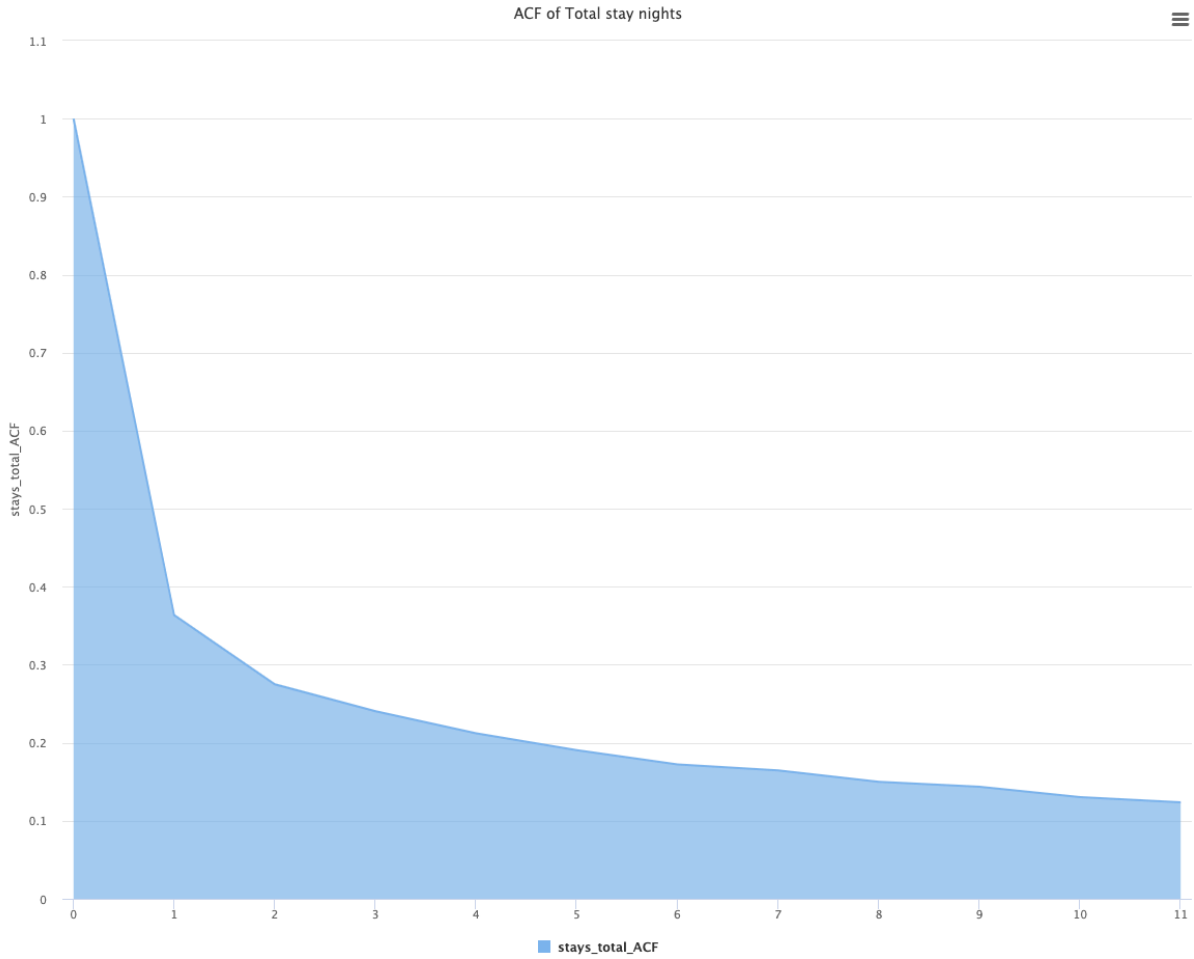
Imported data in RapidMiner showed up without missing values. However, there are some cases with missing values mis-understood by RapidMiner, such as “NULL”, “NA” were recognized as a valid string value for some attributes. Therefore, some pre-processing was required to declare such values as missing values to acknowledge the data completeness of the dataset. After pre-processing, there was just a few cases in the “children” column considered as missing values. Even though “NULL” values presented in the “agent” and “company” columns, it was reasonable since there were possibilities that private customers placed bookings without any middle entity such as travel agent or company. Therefore, those cases were not considered as data completeness issues.

Additionally, there were no outliers detected.

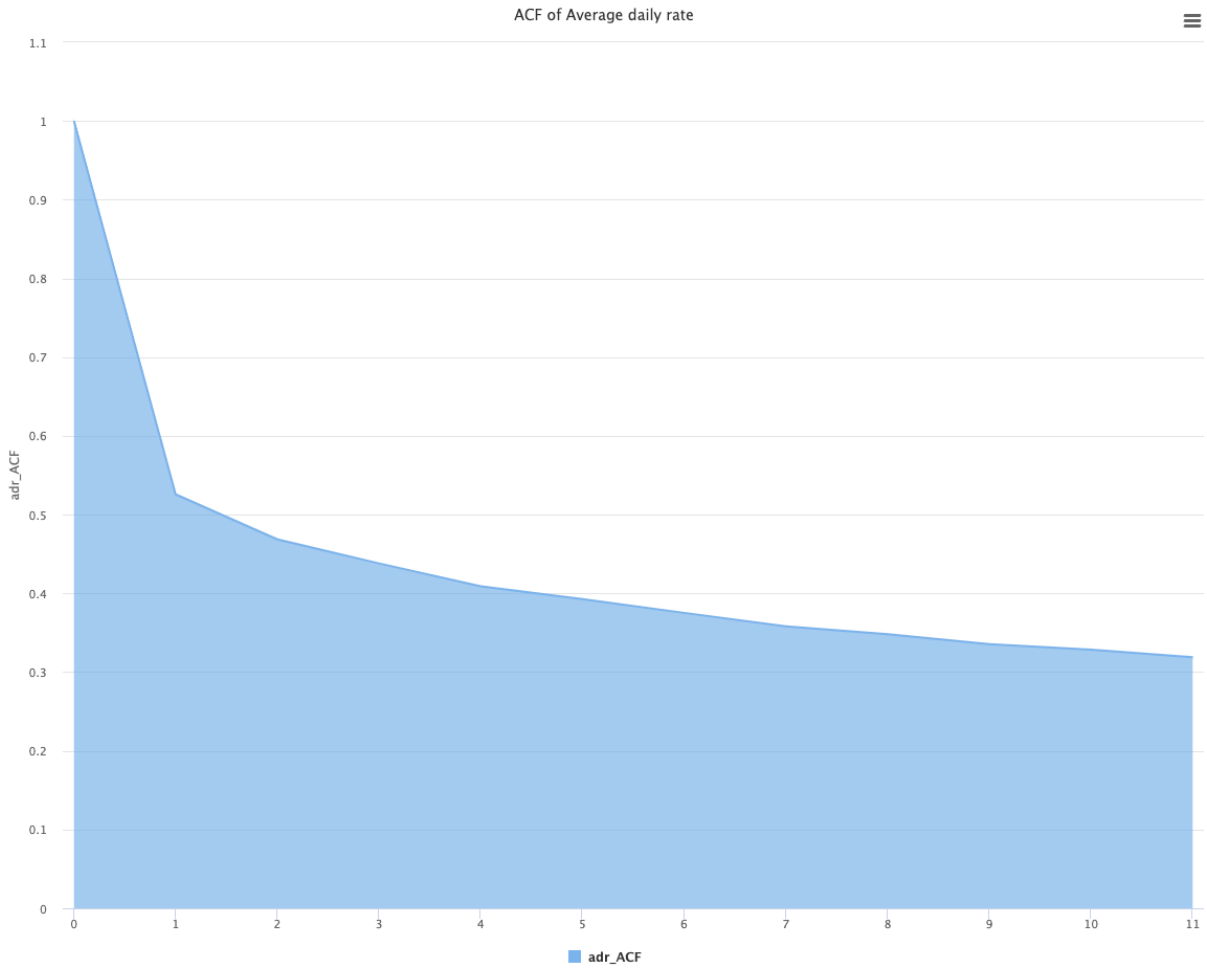
### 2.2.3. Data correlation

To highlight interesting relationship of attributes, Pearson correlation and autocorrelation was studied. Pearson correlation suggested a moderate positive relationship between average daily rate and the number of children with a magnitude of 0.325. Relationship between the number of stays in weeknights and weekend nights is an obvious one and the magnitude was 0.499.

This dataset can be considered as a timeseries, autocorrelation is a relevant to check for relationship of each attribute with its past values. Autocorrelation graph of total stay nights was decayed rapidly indicating a weak or no correlation between the lagged values of the time series. That meant the timeseries was stationary and there was little or no predictive power in the past values for forecasting future values. Stationarity is an important property for timeseries analysis, and it makes modelling convenient and effective.




In the other hand, average daily rate seemed to have relatively high autocorrelation suggesting that the timeseries has some degree of persistence or memory, meaning that the past values can be useful in predicting future values. This could make the time series more difficult to model and forecast accurately.




# 3. Business questions:

Some business questions have been raised as a roadmap for analytics goals and procedures.

## 3.1. Customer segmentation: EDA & Clustering analysis

- Where are the customers from?
- Bookings per market segment
- How long per stay customers booked?
- Cancelations per month
- Can we develop a clustering model to segment customers?

### 3.2. Pricing strategies: EDA

- How does the price vary? Over time? Seasonal? Room types? Lead-time?
- What are the effective marketing channels?
- The price by market segment and room type

### 3.3. Demand distribution: EDA

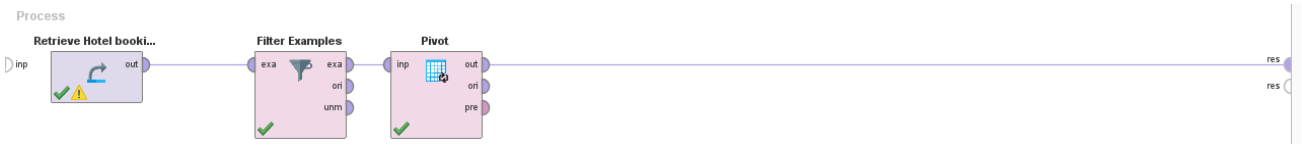
- In which months/seasons the overall demand is high/low?
- How does demand vary per room types/ customer types?
- Which travel agency made the highest number of bookings?

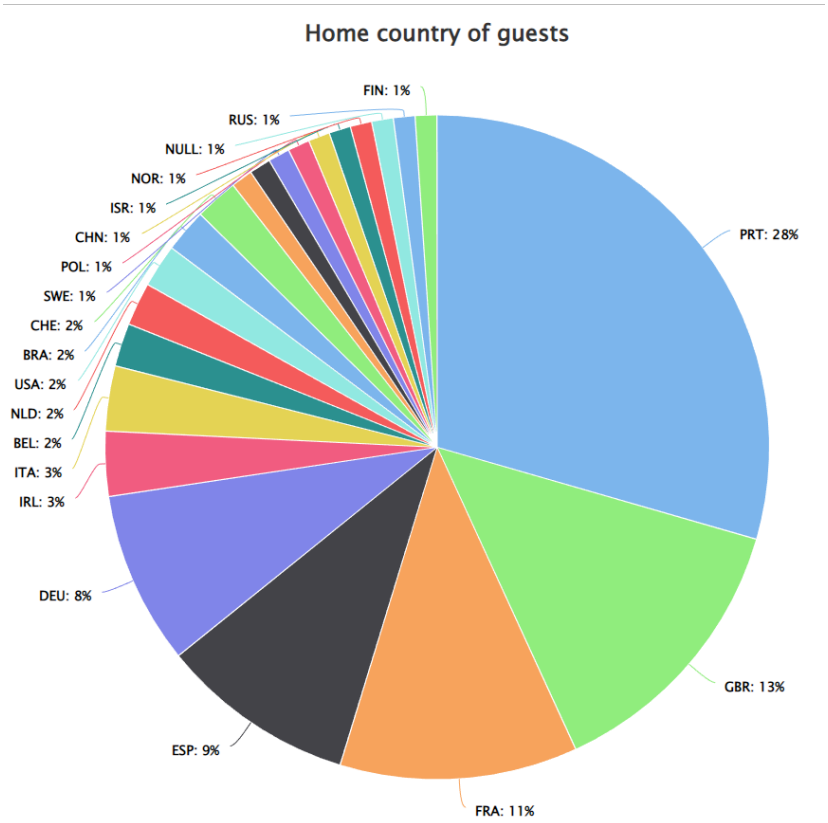
# 4. Procedures & Results

## 4.1. Customer segmentation

### 4.1.1. Exploratory Data Analysis

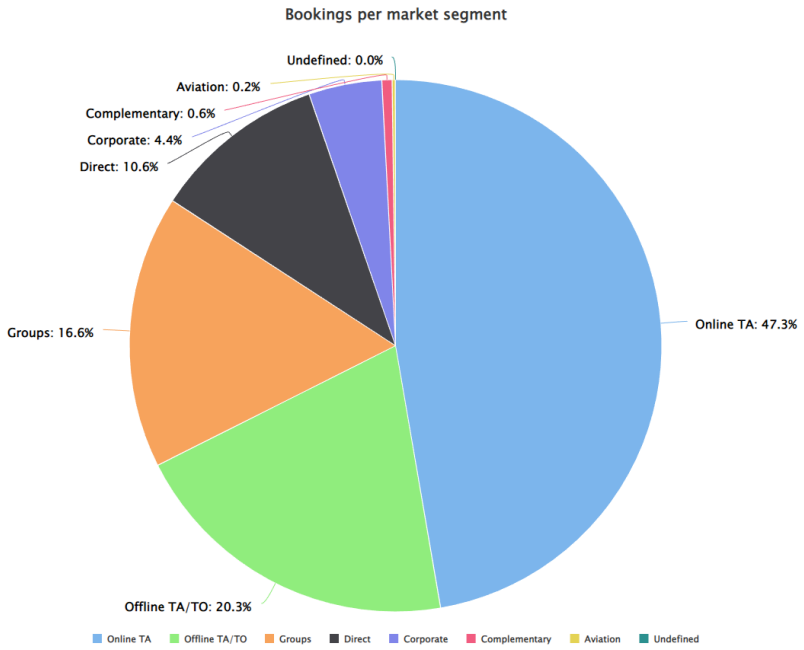
- Where are the most guests coming from?





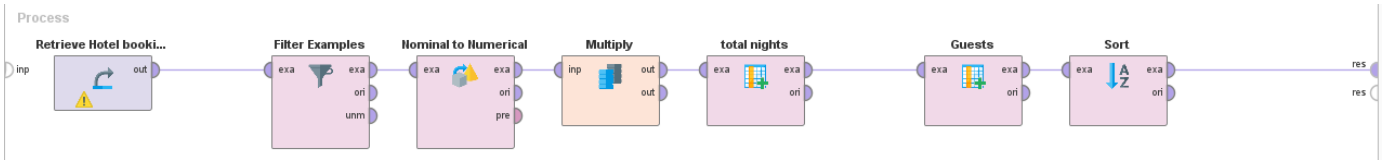

According to the figure above, majority of guests are from Portugal and various other European countries. In this analysis, the dataset was filtered based on the "is_canceled" column to include only cases where the cancellation did not occur. The pivot operation was then applied to calculate the count and percentage of occurrences for each country.

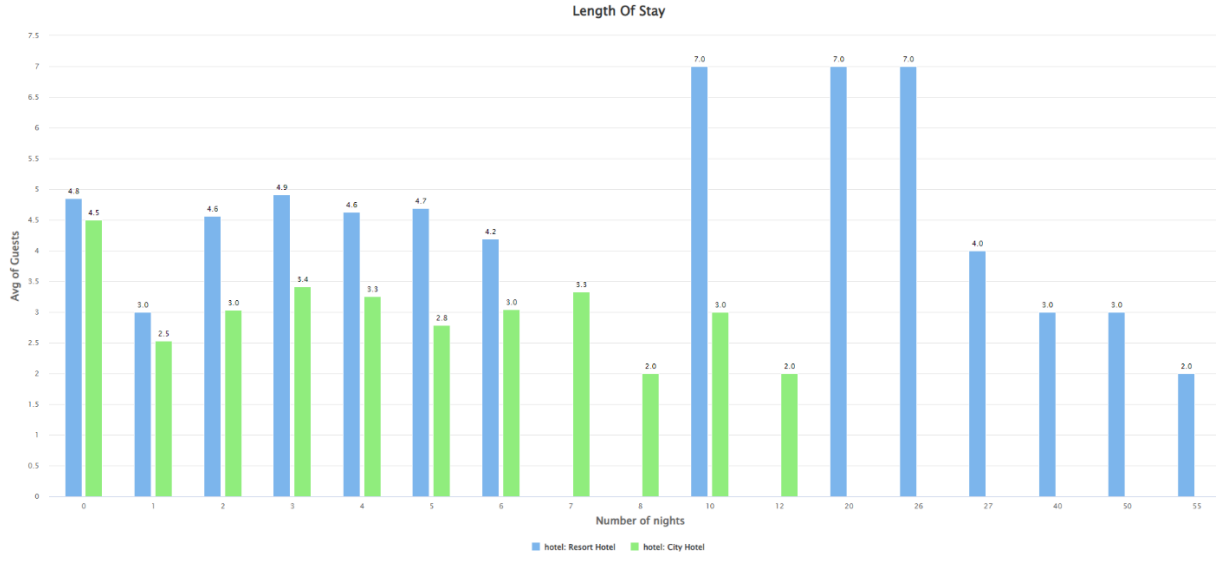
- What are the bookings by market segments?




The figure above indicates that "Online Travel Agents" hold a market share of almost 50%. Two-thirds of the remaining market are occupied by "Offline Travel Agents or Tour Operators" and "Groups".

- How long do people stay at the hotels?

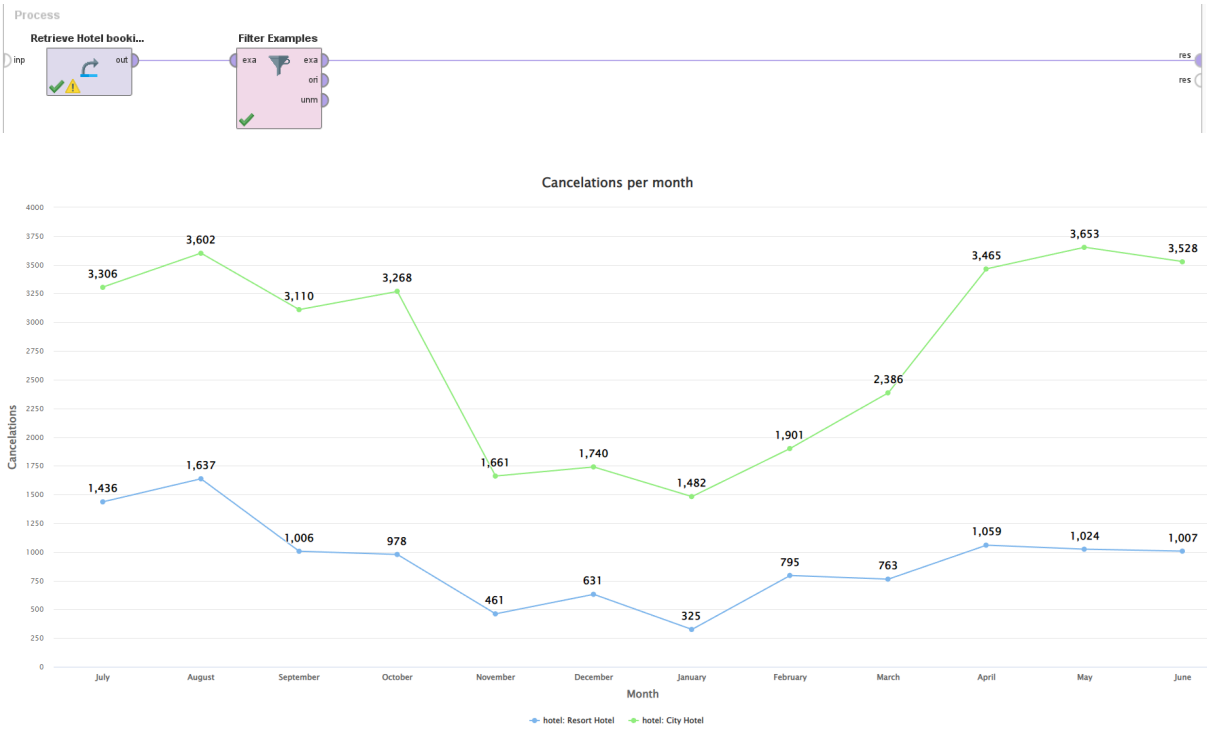


Based on the presented figure, it can be inferred that the average number of guests staying in the City Hotel and Resort Hotel remain stable for stays ranging from 1 to 6 days. However, there are certain exceptional cases where customers have stayed for extended periods, ranging from 20 to 55 days. Although such instances are infrequent, there are a few notable occurrences, such as only two cases of customers staying up to 55 days in the Resort Hotel.

In this case, the dataset was processed by removing outliers where the number of children exceeded 10. This step aimed to eliminate extreme values that could potentially skew the analysis. After removing the outliers, the type of the "children" column was transformed from a nominal categorical variable to a numerical variable. This transformation enabled the calculation of numerical operations on the column. Next, two aggregations were performed on the dataset. The first aggregation involved calculating the sum of the total nights, considering both weekdays and weekends. This aggregation aimed to provide an overall measure of the duration of stays. The second aggregation involved calculating the sum of the total guests, including adults, children, and babies. This aggregation aimed to capture the total number of individuals accommodated. Finally, the dataset was sorted in descending order based on the total nights column. This sorting allowed for the identification of the records with the longest durations of stay.

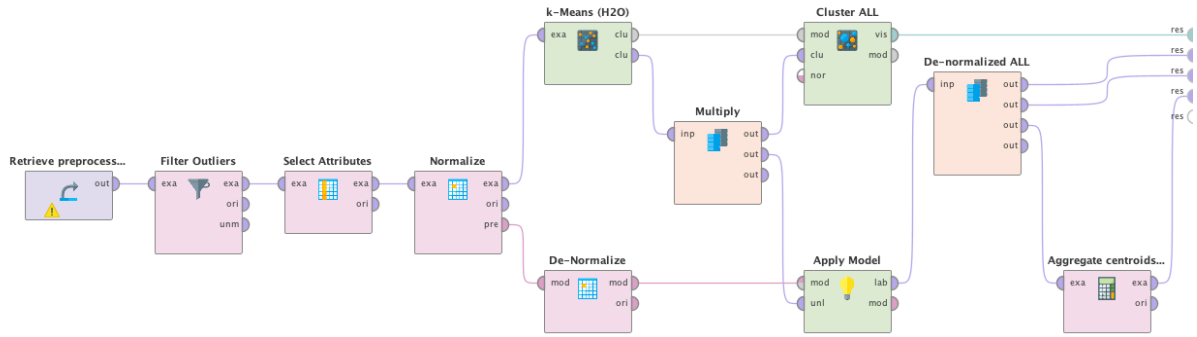
- Which months have the highest number of cancelations?



As per the presented figure, it can be observed that the months of November, December, and January typically experience the lowest number of cancellations. The number of cancellations is relatively stable, hovering around 3300 from July to November, after which the curve begins to decline from November through January. Subsequently, the curve steepens again in February and remains relatively stable from April onwards. In this analysis, the dataset was filtered based on the "is_cancelled" column to include only cases where the cancellation did occur.

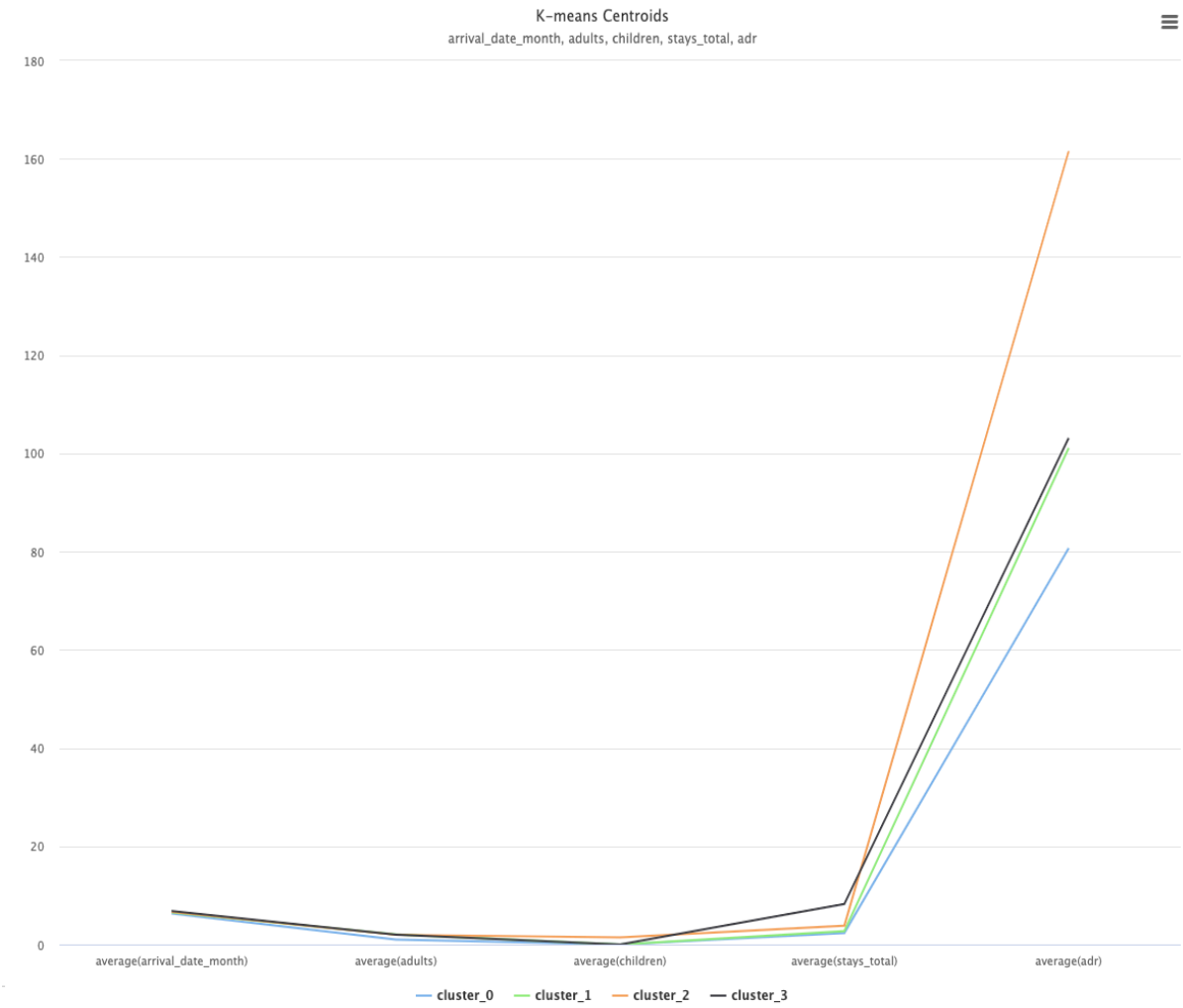
### 4.1.2. Clustering Model

To better understand what type of customer groups in the dataset, customers of all bookings were clustered based on the number of adults and children, number of total nights, the arrival month and the average daily rates.



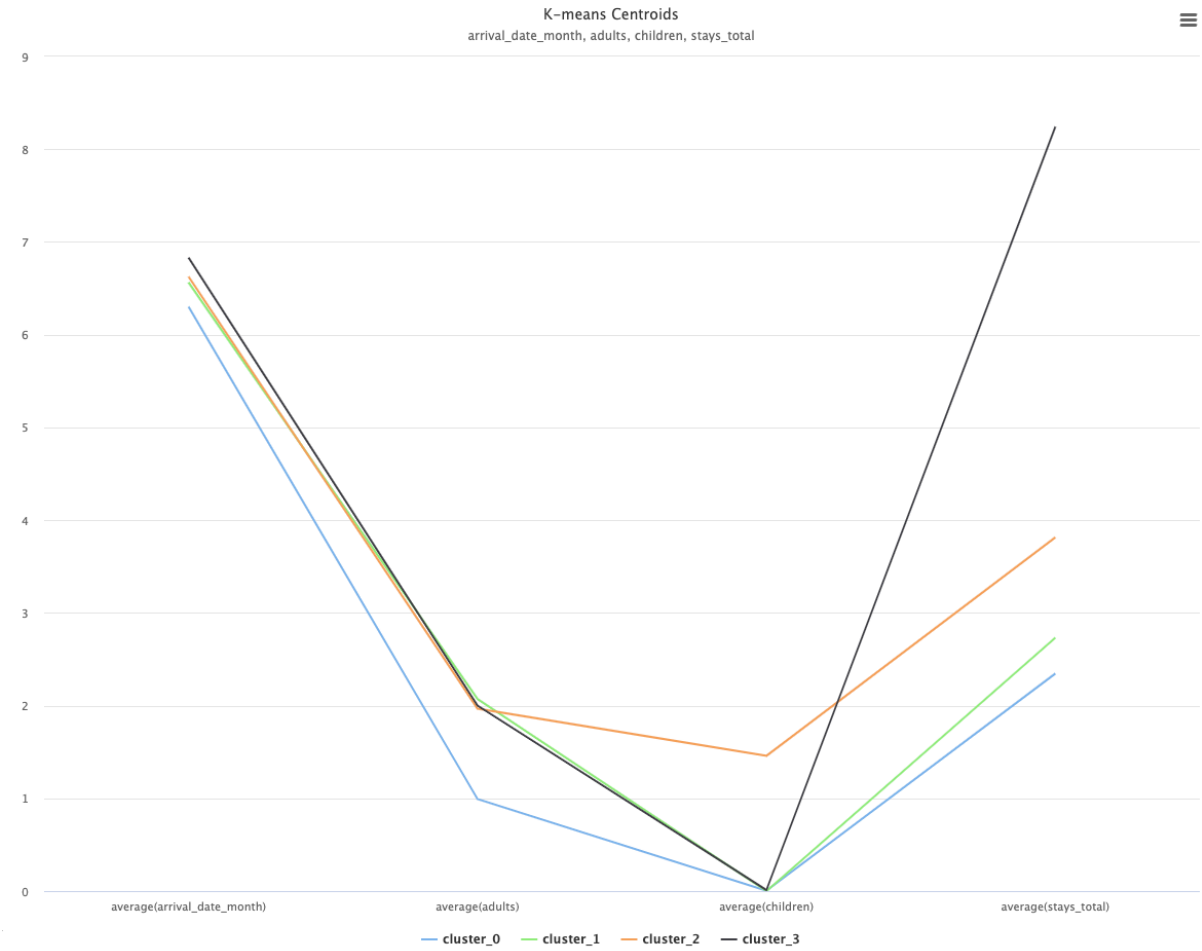

Firstly, we loaded the pre-processed dataset, filter out outliers such as the only case where daily rate over 5000 euros, the only case with 10 children and cases having negative daily rate. Then, the mentioned attributes were selected and normalized to prepare for training clustering model. We clustered all bookings with those mentioned attributes using k-means technique where we allow the model to estimate the number of k with a max of 5. Finally, in order to view the results in its original scale, the clustered data was denormalized and centroids were re-computed.

As a result, the optimal number of clusters was four. Below are plots of centroids of the four clusters.



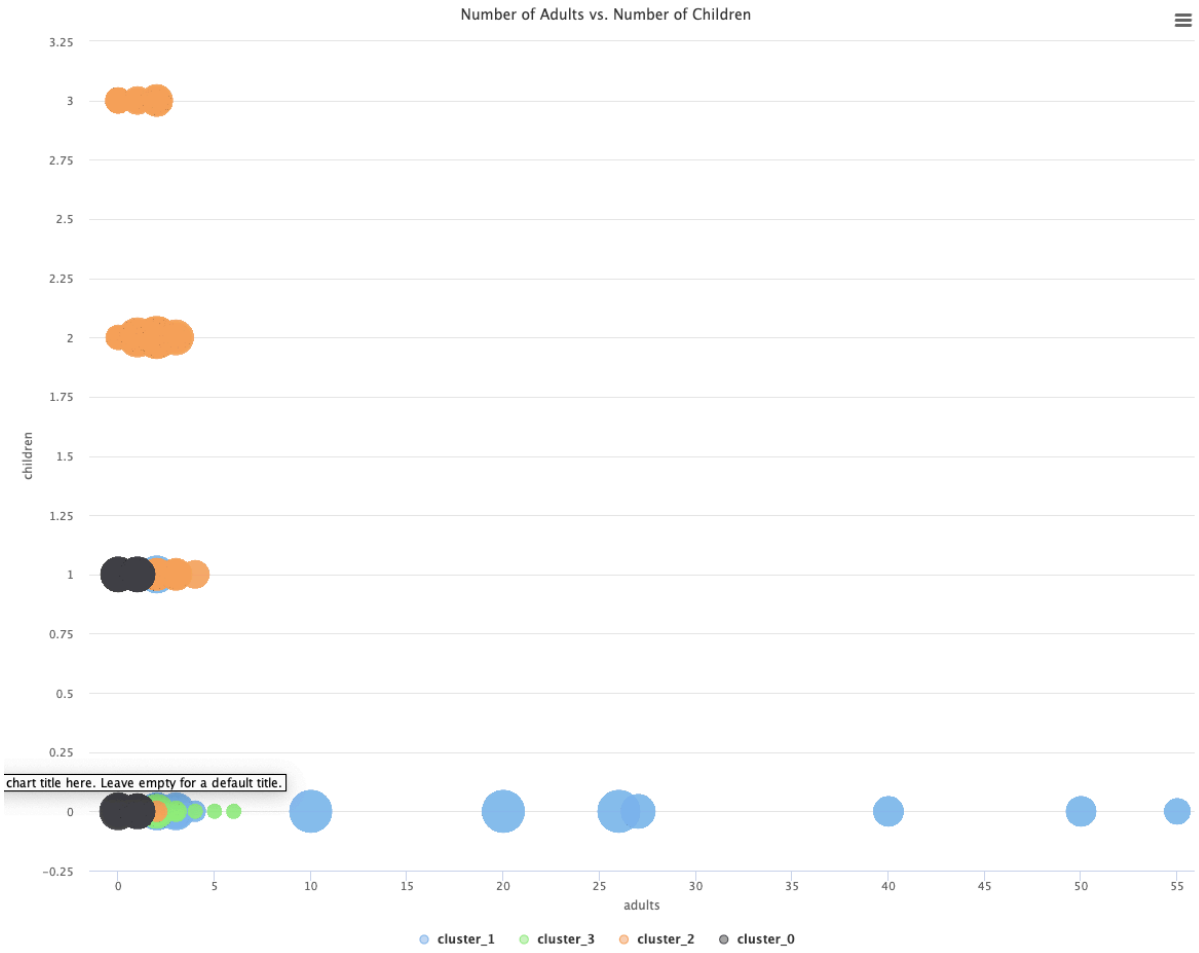

The above plot depicts centroids of all attributes included in the clustering models. Even though the other attributes were not showing well in this graph due to big difference in scales of values, it is clear that the average daily rates of bookings in cluster 0 were relatively smaller than that of other clusters. Cluster 1 and cluster 2 seemed to share similar rates, while cluster 3 peaks at 160 euros.

In order to uncover the patterns in the other attributes, the average daily rate was excluded from the graph.

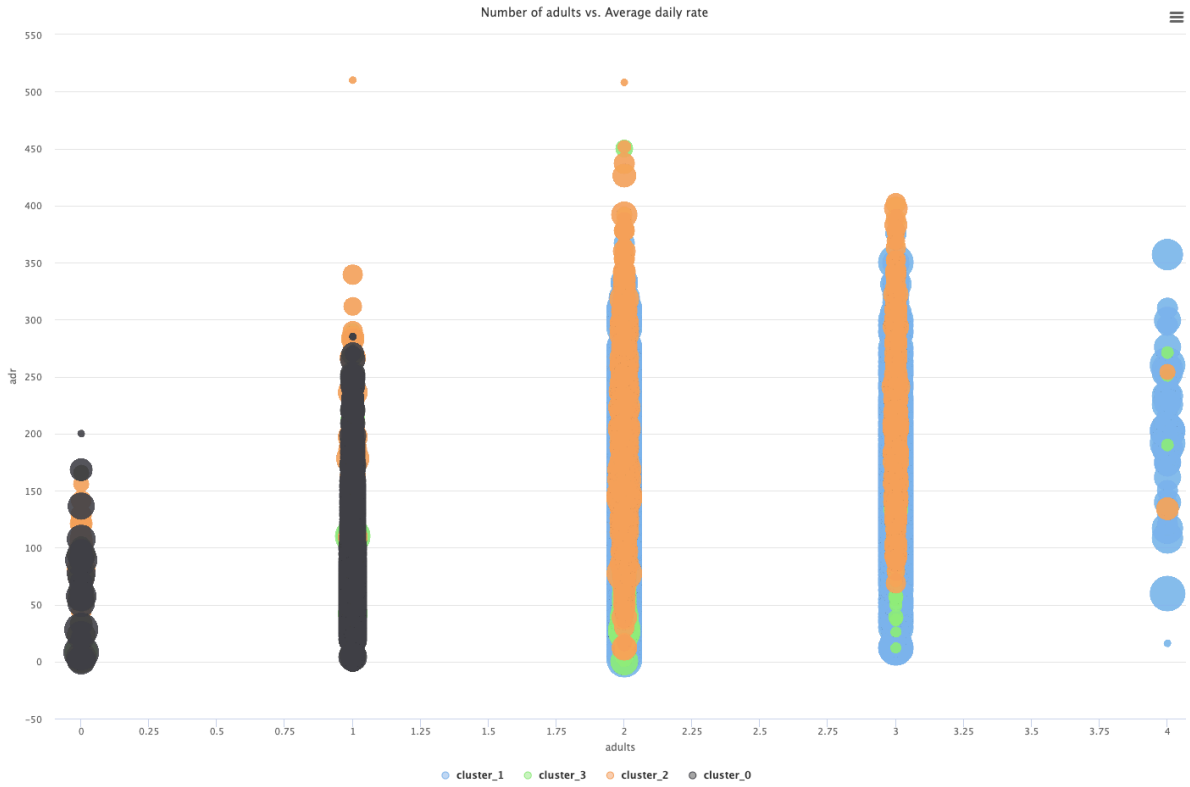



The above plot suggests that in an average scale, cluster 0 included booking transactions of mostly individual or solo travellers in the company of no ones and spent short amount of time at the corresponding hotel with an average of just over 2 nights. Cluster 1 and cluster 3 had similar customer segment in terms of arrival month, number of adults and number children. The clusters both had the average number of adults at 2 and no children. However, the clusters part way at the total number of nights spent at the property, such that cluster 1 included customers visiting for a short trip of under 3 nights, while cluster 3 was a partition for customers with usually a stay with an extended period of time of an average of over 8 nights. Cluster 2 grouped customers as family with some adults and children and they spent an average number of under 4 nights. The arrival month attribute had little power at distinguish clusters.

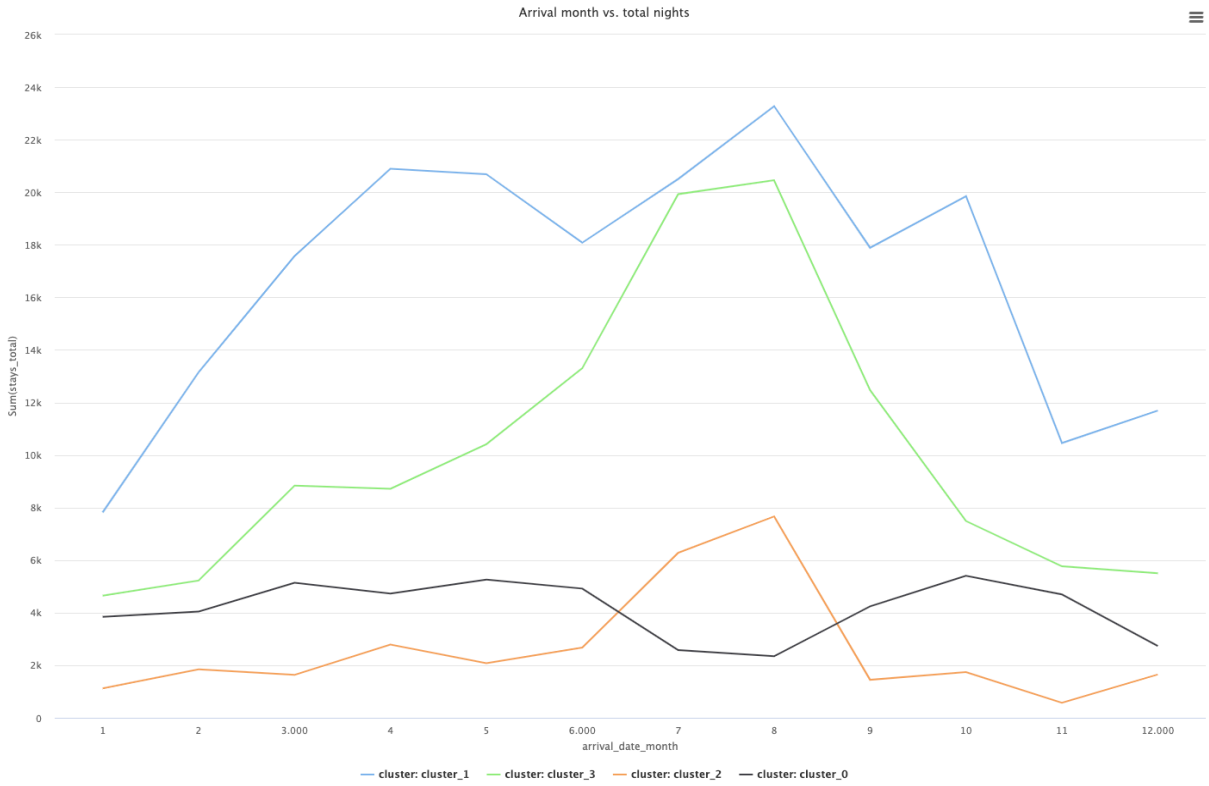
Below are scatter plots of attributes with colours corresponding for clusters and the size of the data points corresponding for the total number of nights.




The scatter plot demonstrates the number of adults on the x-axis and the number of children on the y-axis. It shows that both cluster 1 and cluster 3 included groups of customers without children, big groups of adult customers were clustered into the cluster 1 and small groups of adult customers such as couples were clustered into the cluster 3. Additionally, big groups of customers tended to have a longer stay than smaller groups. Cluster 0 were for solo travellers or travellers with small family with one child only, and their stays were usually long. Big families with children were classified into the cluster 2 and their stays varied.




The above plot scatters the number of adults on the x-axis together with the average daily rate on the y-axis. Cluster 2 with big families having children was more generous with the booking having relative higher charges. The trend was very reasonable, because big families might need bigger rooms which usually involve bigger charge. Charge rate for transactions of alone travellers or couples was relatively lower. From the charge, there was some relationship between the daily rate and the length of the stay as well. If we focus on the tip of each column, it is noticeable that rate was higher for short trips with smaller sized data points.



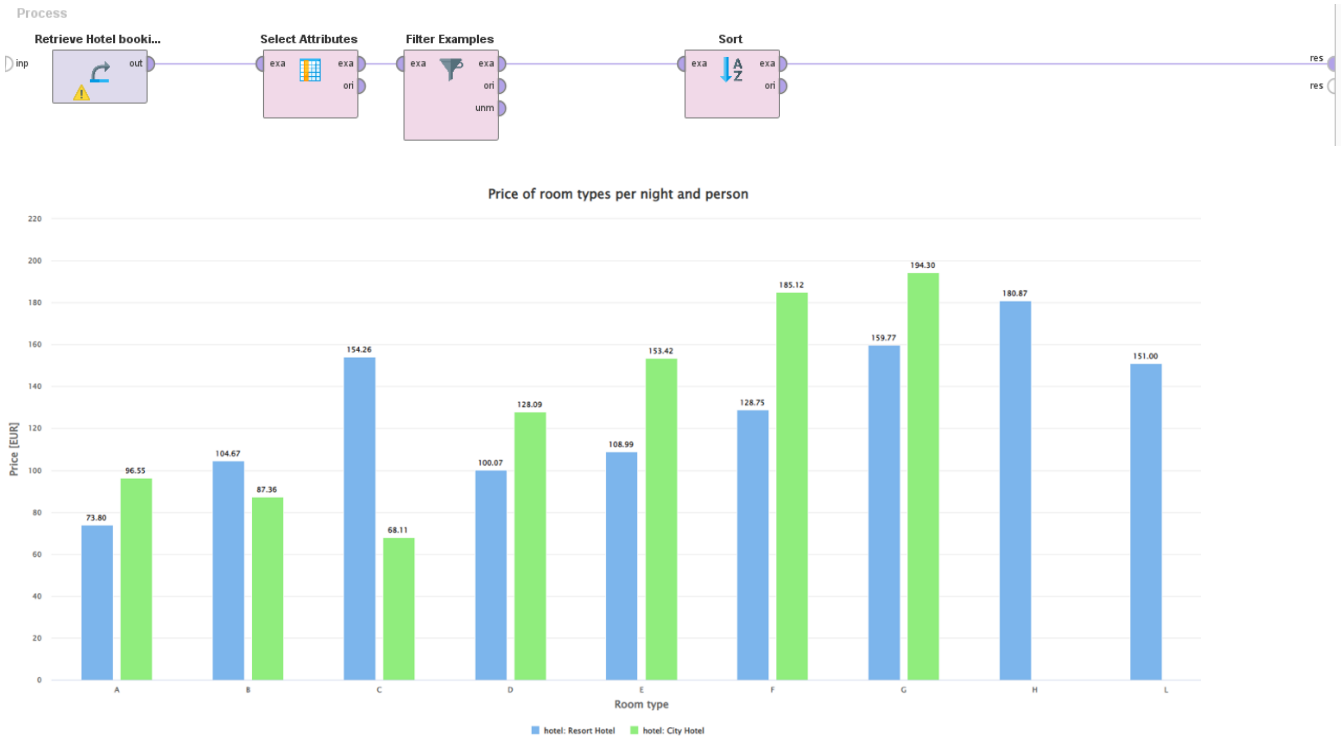

Next, we studied the customer's demand over time of a year. Peak season was in July and August when demands rose for all customer groups, except for Cluster 0 having mostly alone travellers who had a tendency to avoid high-demand season. Biggest demands belonged to the customer groups in cluster 1 with more adults. Cluster 3 with smaller groups of adult customers was interested mostly in the peak season, and thereby, lowering their demands in other months of the year. Demands of family customer groups was the lowest overall.

In summary about customer segmentation, even though information about demographics of customers was a lack, we achieved segmentation by clustering customers based on their booking transactions. By that, we acknowledged different types of customer groups, their budgets and demands. The dataset allows us to cluster the customer base into 4 segments. The first segment (cluster 0) was for alone travellers with lower budget and high demands during off-season. The second segment (cluster 1) was the most important due to its biggest demands and this segment included mostly bigger groups of adults who had a tendency to pay a long visit. The third segment (cluster 2) might be the focused area for the hotels if they wanted to capture more interest of family customer groups who usually pay more generously. The last segment (cluster 3) belonged to smaller groups of adults such as couples whose demands rose rapidly during peak season, but charge rate was relatively lower than others.

## 4.2. Pricing strategies

### 4.2.1. Exploratory Data Analysis

- How much do guests pay for a room per night?

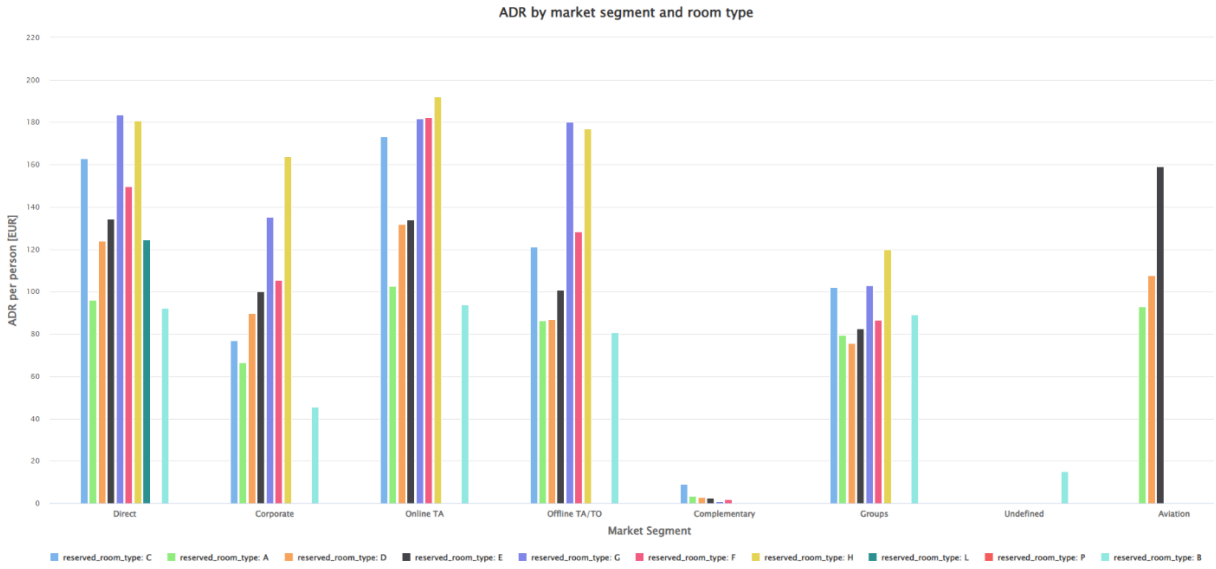



The figure presented displays the average price per room categorized by its type and standard deviation. It is important to note that rooms labelled with the same letter may not be identical across different hotels due to data anonymization. Upon inspection of the figure, it can be inferred that customers at the city hotel are not offered room types H and L.

In this case, a subset of the dataset was created by selecting only specific attributes, namely "adr", "hotel", "is_canceled", and "reserved_room_type". This step aimed to focus the analysis on these particular variables of interest. After selecting the desired attributes, the dataset was further filtered based on the "is_canceled" column. Only cases where the cancellation did not occur were included in the filtered dataset. This filtering allowed for the isolation of non-canceled bookings, enabling analysis specific to those cases. Finally, the filtered dataset was sorted in ascending order based on the "reserved_room_type" column. This sorting

arrangement allowed for the organization of the data according to the room types that were reserved, providing a structured view of the bookings based on this criterion.

- What is the price breakdown per market segment and room type?




As per the analysis presented in the explanatory data section, aviation holds a market share of only 0.2%. Nevertheless, it is noteworthy that aviation segment offers the highest price for room type E in comparison to all other segments.

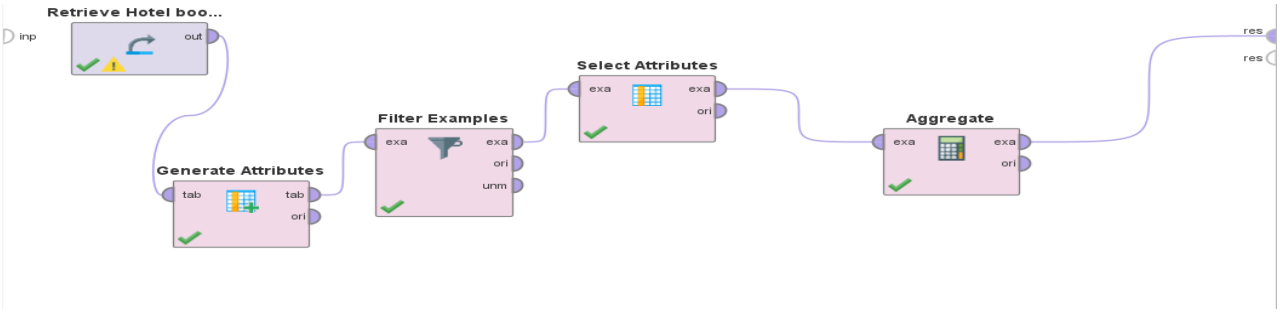
- How does the price vary over the period?

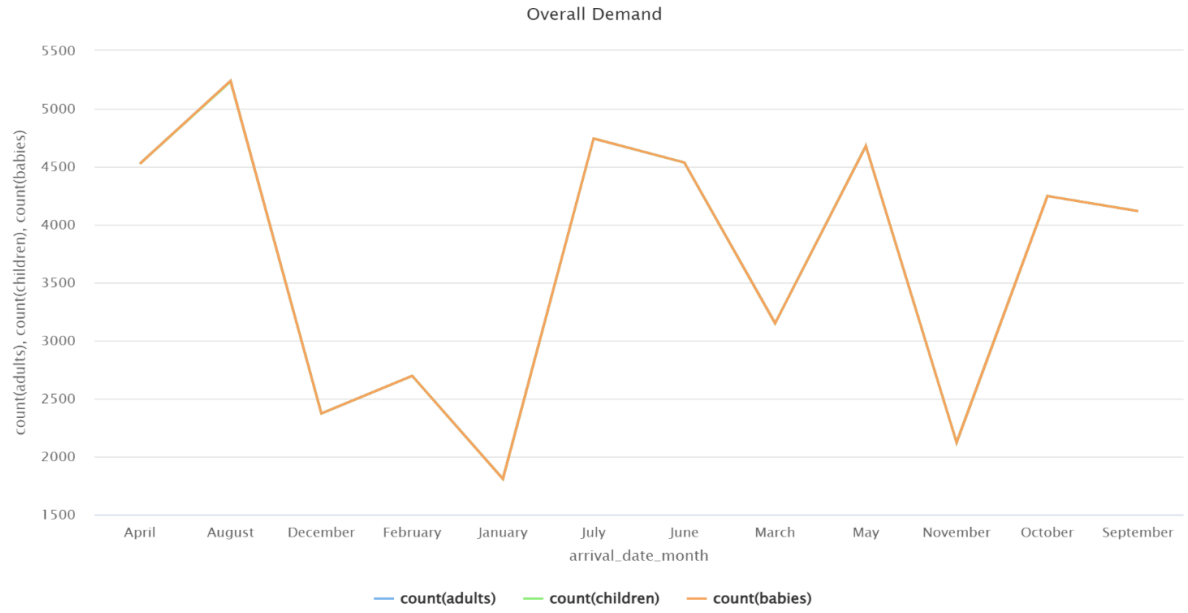

The figure above provides a clear indication that prices in the Resort hotel exhibit a significant increase during the summer months. In contrast, the price of the city hotel is relatively stable with less variation, peaking in the spring and autumn seasons when it is at its most expensive. Indeed, it is reasonable to expect that the price paid from November to January is comparatively low, which may explain the lower number of cancellations during this period. Customers may be more inclined to cancel reservations when they have paid higher prices, and thus lower prices during these months may result in a more stable booking pattern.

## 4.3. Demand distribution

### 4.3.1. Exploratory Data Analysis

- In which months/seasons the overall demand is high/low?




Based on the provided Line graph, it can be observed that the highest overall demand for hotel bookings occurred in the month of August, with more than 5,000 guests. July and May closely followed with 4,742 and 4,677 guests, respectively. The months of June and April had 4,535 and 4,524 guests, respectively. Meanwhile, October recorded a total of 4,246 guests.

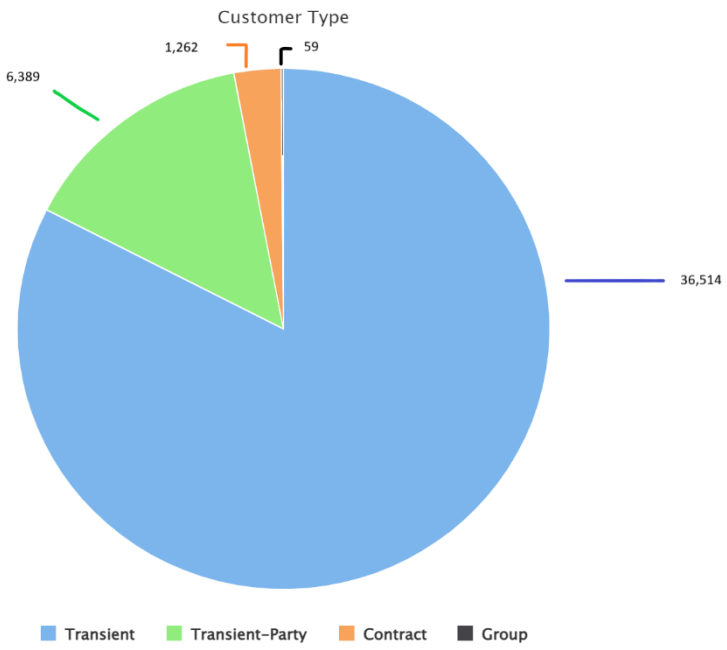
In contrast, the month with the lowest demand for hotel bookings were January, with only 1,807 guests in total. This was followed by November with 2,122 guests, December with 2,371 guests, February with 2,696 guests, March with 3,149 guests, and September with 4,116 guests in total.

It is worth noting that the level of demand for hotel bookings appears to be seasonal, with the summer months of August, July, and May being the busiest, while the winter months of January, November, and December are the least busy.

The line graph provided also confirms the seasonal demand patterns mentioned above, depicting the overall demand fluctuation over the course of a year. The graph shows a clear increase in demand during the summer months, with a peak in August and a steady decline towards the end of the year. Conversely, the winter months, particularly January and February, show a notable decrease in demand, with a gradual increase as the year progresses.

The line graph provides a more comprehensive view of the overall demand trends throughout the year, supporting the observations made from the bar chart. This information can be helpful for hotel managers in adjusting their staffing and inventory levels according to seasonal demand patterns.

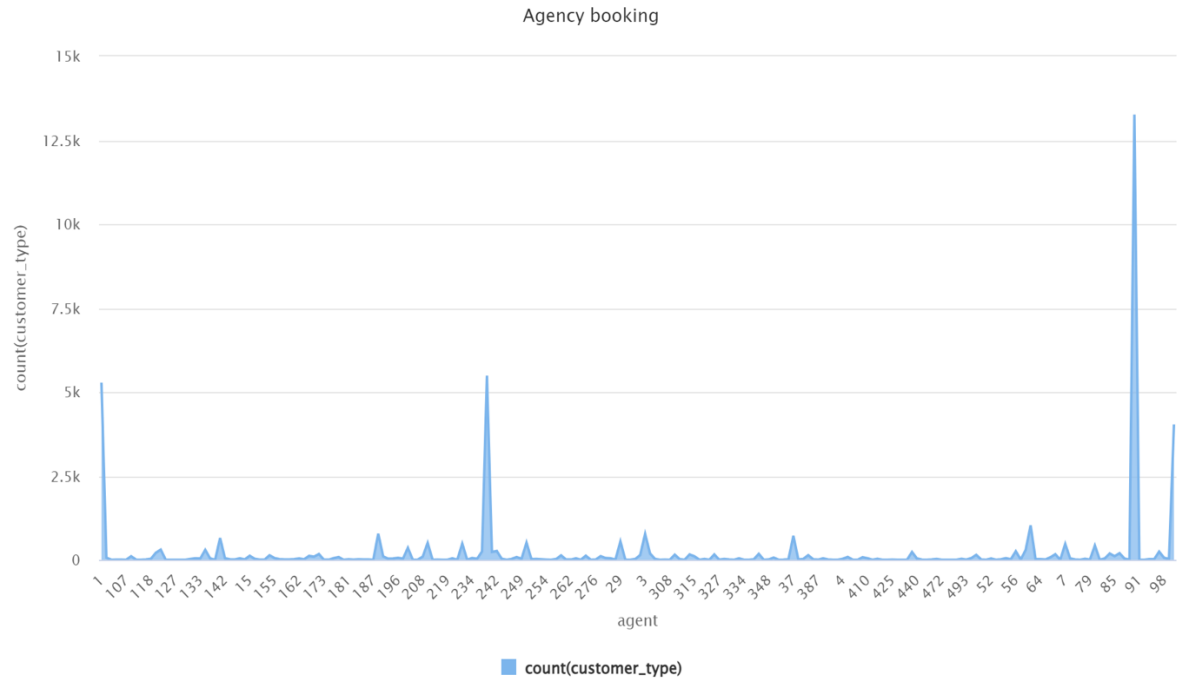
- How does demand vary per customer types?




The above pie chart provides valuable insights into the demand for hotel rooms across different customer types. The data highlights that the majority of customers, with a count of 36,514, prefer to book rooms as transient guests. Transient parties follow closely behind with 6,389

customers, while contract customers rank third with a count of 1,262. Interestingly, group customers have the lowest demand, with only 59 bookings.

The high demand for transient customers suggests that most customers prefer to book their hotel rooms individually for a short duration rather than through a group or contract arrangements. This information can be useful for hotel managers in determining their target customer base and developing effective marketing strategies to attract specific customer types. By understanding the varying demand for different customer types, hotels can tailor their services and amenities to better meet the needs and preferences of their target customers.




- Which travel agency made the highest number of bookings?

The above visualization provides us with valuable insights into the top-performing travel agencies in terms of hotel bookings. From the graph, we can see that travel agency with ID 91 secured the first position with the highest number of bookings, totally 13,264.

Additionally, agency ID 240 and agency ID 1 also performed well, securing the second and third positions with a total of 5,484 and 5,280 bookings, respectively.

By identifying the top-performing travel agencies, hotel managers can strengthen their partnerships with these agencies and collaborate to increase their bookings further.

Furthermore, targeted marketing strategies can also be developed to attract more customers through these agencies, leading to increased revenue for the hotels.

# 5. Conclusion

In this report, we have outlined our procedures, methodology, and findings of our data mining project. The chosen dataset contains information about booking transactions from two hotels covering customer demographics, cancellation, charge rates, selling channels, etc. Our goal was to identify patterns and trends in the data, understand the dynamics of hotel demand, and use them to answer some business questions on three topics: customer segmentation, pricing strategies, and demand distribution.

To achieve this goal, we used RapidMiner tool for data pre-processing and modelling. We applied exploratory data analysis and visualization techniques as well as classification and clustering models with intelligent features. Through these techniques, we were able to uncover some hidden insights that can inform decision-making in the hospitality industry.

Our findings demonstrate the value of applying data mining techniques to hotel booking data for business analytics purposes. Overall, this report provides valuable insights into the dynamics of hotel demand and offers practical recommendations for improving customer segmentation and pricing strategies.

We hope that our findings will be useful for stakeholders in the hospitality industry who are looking to optimize their operations and improve their bottom line.