**ANALYSIS OF HOTEL BOOKING CANCELLATION**

**INTRODUCTION**

In the hospitality industry, booking cancellations greatly impact demand-management decisions, hindering the production of accurate forecasts essential for effective revenue management. To address cancellations, hotels often enforce strict cancellation policies and adopt overbooking strategies, which can negatively affect revenue and social reputation, ultimately harming business performance. Overbooking can result in the hotel denying service provision, leading to a poor customer experience and damaging the hotel’s reputation and immediate revenue. This can also result in future revenue loss as dissatisfied customers may not return. Conversely, rigid cancellation policies, particularly non-refundable ones, can reduce the number of bookings and revenue, as significant discounts may need to be applied to attract guests (Antonio et. al., 2017).

**PROBLEM STATEMENT**

The following project utilizes a hotel booking dataset containing various details surrounding two types of hotels namely City and Resort hotels.

In recent years, both the City Hotel and Resort Hotel have seen substantial increases in cancellation rates. This has led to challenges such as decreased revenue and underutilized rooms. Consequently, their primary goal is to lower cancellation rates to improve revenue generation. This report analyzes hotel booking cancellations and other factors that indirectly affect their business and annual revenue.

**Assumptions**

1. From 2015 to 2017, no significant events had a notable impact on the data considered.
2. The data is up-to-date and can be effectively utilized to assess potential hotel strategies.
3. There are no unexpected challenges with the hotel adopting any of the suggested approaches.
4. The recommended solutions are not already being implemented by the hotels.
5. Booking cancellations are the most significant factor influencing revenue generation.
6. Cancellations result in unoccupied rooms for the originally reserved period.
7. Clients generally cancel their hotel reservations within the same year they made them.

**Research Questions**

1. What factors influence hotel reservation cancellations?
2. How can we reduce hotel reservation cancellations?
3. How can hotels be assisted in making pricing and promotional decisions?

**Hypothesis**

1. Higher prices lead to more cancellations.
2. Longer waiting lists result in more frequent cancellations.
3. Most clients make their reservations through offline travel agents.

**DATA ANALYSIS**

The dataset, available on Kaggle, contains hotel demand information and originates from the article "*Hotel Booking Demand Datasets*" by Nuno Antonio, Ana Almeida, and Luis Nunes, published in the Journal “*Data in Brief - Volume 22*” in February 2019 (Elsevier Inc., 2018).

It includes records of hotel bookings made between July 1, 2015, and August 31, 2017, encompassing both arrivals and cancellations. The dataset combines data from two types of hotels: city and resort. Each dataset comprises 31 variables, with the resort hotel data containing 40,600 observations and the city hotel data containing 72,300 observations, resulting in a total of 119,300 observations. To protect privacy, all identifying information related to the hotels and customers has been removed.

The table below illustrates the information of all the thirt-two columns included in the dataset.

A screenshot of a computer

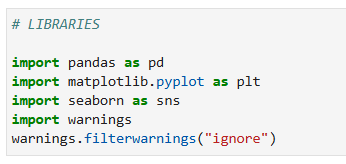
Description automatically generated

The following section will skim through the four main stages of a data analytics project namely; Data Collection, Data Exploration, Data Pre-Processing, Data Analysis, Data Interpretation.

**Data Collection**

Import Necessary Libraries

Firstly, all the necessary libraries have been imported which will aid in data exploration, manipulation, and visualisation. These are commonly used in data analysis tasks and provide great functionality in handling and displaying data in Python. The chosen IDE is JupyterLab operated though anaconda.



Importing the Dataset

The following code is used to import the dataset from a CSV file and store it into a data frame, named *df\_hotel*, followed by displaying the first 10 rows of the dataset, in order to understand the structure and content of the dataset.

A screenshot of a computer code

Description automatically generated

A screenshot of a calendar

Description automatically generated

**Data Exploration**

Exploratory Data Analysis involves analyzing and visualizing data to understand its key characteristics, uncover patterns, and identify relationships between variables. This method helps to explore datasets, discover patterns, find outliers, and identify variable relationships (EDA, 2024).

Descrptive and Summary Statistics

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

As it can be seen in the above figures, the code above provides a summary or information about the dataset. The code *df\_hotel.info()* is used to print the information on the columns, data types, and memory usage. It can be noted that there are exactly 119,389 entries with 32 attributes as explained in the above sections. This is further followed by a statistical summary of numeric and categorical variables.

It can be noted that the dataset consists of 20 numerical variables and 12 categorical variables. The variable *reservation\_status\_date* is of object type when in fact should be of date type. Furthermore, variable *average daily rate (adr)* contains negative values and the maximum value shown is considerable high; implying the data contains certain outliers which need to be further explored.

Displaying Missing and Unique Values

A close-up of a computer screen

Description automatically generated



A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

As seen from the above figures, the code represents the total number of rows and columns, missing values, and unique values (categorical variables in detail). It can be noted that there is about 112,593 (94%) missing *Company* values, 16340 (14%) missing *Agent* values, 488 (0.4%) missing *Country* values, and 4 (0.003%) missing *Children* values. These need to be further explored and handled appropriately.

**Data Pre-Processing**

Data pre-processing or cleaning is essential for producing accurate and meaningful analysis results. It involves identifying and correcting errors and inconsistencies to ensure the data's quality and reliability (DataCamp, 2023).

Convert Data Type

A close-up of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

As discussed above, *reservation\_status\_date* attribute should not be of object type. Hence, it was converted to date type, and the resulting dataset has been displayed to confirm the changes.

Handling Missing Values

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

As mentioned previously, the columns *Company* and *Agent* have a very high number of null values which can be difficult to deal with. As these columns are not necessarily required for the purpose of this project, they will be dropped from the dataset.

The remaining two columns, namely *Country* and *Children* have a lower number of null values, which will not have a big impact on the analysis. Hence, the rows containing null values will just be dropped from the dataset.

Outlier Detection

A screenshot of a computer

Description automatically generatedA screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

The column *adr* contains outlier values, indicated by a skewed box plot distribution. To address this, the Interquartile Range (IQR) between the first quartile (Q1) and third quartile (Q3), was calculated to identify the bounds. The results indicated 3,883 upper outliers and no lower outliers. Values outside these bounds were eliminated, thus ensuring a dataset free of outliers, ready for analysis.

**Data Analysis and Visualization**

This section includes various visualizations designed to further analyze the reasons for hotel cancellations and to validate the hypotheses and assumptions.

Count of Reservation Status

A red and blue pie chart

Description automatically generated

The bar graph illustrates the percentage of reservations that were canceled versus those that were not. It's evident that a substantial number of reservations remain active. However, 37% of clients canceled their reservations, significantly impacting the hotels' revenue.

Count of Reservation Status for Each Hotel

A graph of blue squares

Description automatically generated

City hotels have more bookings compared to resort hotels, possibly because resort hotels are more expensive. However, cancellations in City hotels (Cancelled - 42%, Not Cancelled – 58%) are higher than those of Resort Hotels (Cancelled – 28%, Not Cancelled – 72%).

Calculate the Average Daily Rate for Each Hotel

A graph of a hotel

Description automatically generated

The line graph indicates that on some days, the average daily rate for city hotels is lower than that of resort hotels, and on other days, it is even lower. It is clear that weekends and holidays are likely to cause an increase in resort hotel rates.

Analysis of Reservation Status on a Monthly Basis

A graph of blue bars

Description automatically generated

A grouped bar graph was created to analyze the months with the highest and lowest reservation levels based on reservation status. It shows that both confirmed and canceled reservations peak in August, while January has the most canceled reservations.

Analysis of Average Daily Rate of Cancelled Bookings on a Monthly Basis

A graph of different colored rectangular objects

Description automatically generated with medium confidence

As shown, ADR in August was the lowest, whilst ADR in January was the highest. The above bar graph indicates that cancellations are most frequent when prices are highest and least frequent when prices are lowest. Thus, the cost of accommodation is the primary factor driving cancellations.

Top 10 Countries with Cancelled Bookings

A pie chart with different colored circles

Description automatically generated

The top country is Portugal with the highest number of cancellations, amounting to 70%. This is followed by the UK (GBR), Spain (ESP), and France (FRA).

Analysis of Market Segment

A screenshot of a computer

Description automatically generated

There are four ways for hotel guests to book their reservations: Direct, Groups, Online, or Offline Travel Agents. Approximately 47% of clients book through online travel agencies, while 27% come from groups. Only 4% of clients make direct bookings by visiting the hotels. As shown above, most of the cancellations are from those booked via online agencies, amounting to 46%.

Analysis of ADR for Cancelled and Non-Cancelled Bookings

A graph showing a number of red and blue lines

Description automatically generated

The graph also shows that reservations are more likely to be canceled when the average daily rate is higher, reinforcing the analysis that higher prices lead to higher cancellation rates.

**Data Interpretation**

1. Adjust Pricing Strategies:

* As prices increase, so do cancellation rates. To reduce cancellations, hotels should implement dynamic pricing strategies, potentially lowering rates for specific locations and times. Offering targeted discounts and special promotions to consumers during off-peak periods can also help retain bookings.

1. Flexible Booking Options:

* Since the cancellation rate is higher for resort hotels compared to city hotels, resorts should consider offering discounts on room prices during weekends and holidays. Additionally, providing flexible cancellation policies and allowing guests to change booking dates without penalties can reduce cancellations.

1. Targeted Marketing Campaigns:

* Given that cancellations peak in January, hotels could launch seasonal marketing campaigns to boost revenue during this month. Focused promotions on special packages, events, or discounted rates can attract more guests and decrease cancellations.
* Launch geographical campaigns aimed at countries with high cancellation rates, such as Portugal, the UK, Spain, and France. Tailor the marketing messages to address the specific needs and preferences of guests from these regions.

1. Enhance Direct Booking Incentives:

* Encourage direct bookings by offering exclusive discounts, loyalty points, or additional perks for guests who book through the hotel’s website or in person. Enhancing the online booking experience can also make direct bookings more appealing and reduce cancellations.

1. Improve Quality of Services and Facilities:

* Improving the quality of hotels and services, especially in Portugal, can help reduce the cancellation rate. Investing in upgrading hotel facilities and enhancing service quality can increase guest satisfaction and loyalty, making them less likely to cancel.
* Actively seek guest feedback and address any issues promptly to maintain a positive reputation and attract more bookings.

1. Engage with Customers:

* Engage with guests through personalized communication, offering tailored suggestions and keeping them well-informed about their stay. Building stronger relationships with guests can reduce the likelihood of cancellations.
* Utilize guest feedback and reviews to continuously improve services and address potential concerns that may lead to cancellations.

REFERENCES

<https://www.sciencedirect.com/science/article/pii/S2352340918315191#section-cited-by>

<https://www.kaggle.com/datasets/mojtaba142/hotel-booking/data>

<https://www.datacamp.com/blog/what-is-data-analysis-expert-guide>

<https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/>

Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model - <https://ieeexplore.ieee.org/document/8260781>