**REPORT**

Arnav Gupta

BTech Computer Science Engineering , 2nd Year

23114010

**INTRODUCTION**

The dataset I used is attached in the repository as a csv file . The dataset is about a **hotel booking** analysis .

**Procedure**

* I loaded the dataset . Here I have attached 2 csv files
* This is because original.csv is the file containing the actual dataset . While hotel\_bookings.csv is the file containing the dataset after the preprocessing . This was necessary beacuse the data preprocessing I have performed replaces the data inplace , so if anyone wants to refer to the original dataset , I have given the provision .
* Upon inspecting the structure of dataset and its analysis , I found that there are certain columns like **"lead\_time","adults","babies","adr"** etc. which are numeric columns and certain other columns like **"meal" , "country"** etc. are categorical columns .
* There was missing data , outliers , duplicate rows present in the dataset .
* null values - present in **lead\_time , adults , children**
* duplicate rows - 94 duplicate rows found

**Preprocessing**

* I handled the null values by replacing them with their Median . This was done beacuse the columns contained Integer values , so replacement by mean would not have been feasible .
* I handled the duplicate rows by using the df.drop\_duplicates function .
* I identified the outliers using the Interquartile method . Here , the values lying above Q3+1.5IQR and below Q1-1.5IQR were considered as Outliers .
* For handling the outliers , I used the technique of Winsortization . With this technique , the outlier values were replaced by the nearest non-outlier value .
* For fixing the typos in categorical columns :
* I replaced the entries "Resort hotel" with "Resort Hotel"
* I also converted the entries of all columns to uppercase
* I also stripped the entries of any intervening white spaces

**Exploratory Data Analysis**

For all the EDA steps performed , I will describe the process and my observations :

* Performing the summary statistics on my report , I found out that :

Feature Mean Median Mode Skewness Variance

Cancellation Rate 0.277 0.0 0.0 0.996 0.2007

Lead Time 52.10 48.0 48.0 0.247 1133.29

Week Number 28.93 29.0 29.0 0.043 1.6059

Weekend Stays 1.29 1.0 2.0 -0.045 0.9567

Weekday Stays 3.45 3.0 5.0 0.061 3.7493

Adults per Booking 1.99 2.0 2.0 -0.103 0.1158

Children per Booking 0.18 0.0 0.0 2.789 0.2799

Booking Changes 0.14 0.0 0.0 2.102 0.1191

Average Daily Rate (ADR) 125.28 123.0 195.0 0.274 1189.09

Parking Requests 0.14 0.0 0.0 2.029 0.1236

* Skewness Interpretation:

Most features are slightly skewed, except for children, booking changes, and parking requests, which are highly right-skewed. Adults and weekend stays are almost symmetric.

* Variance Highlights:

Lead time and ADR have the highest variance, indicating a wide range of values.

Adults per booking and week number have the lowest variance, suggesting consistency.

**Frequency Distribution**

* **Hotel Type:**
* Resort Hotel (412 bookings) , City Hotel (392 bookings)
* **Meal Plans:**
* BB (603), HB (169), FB (32)
* **Top Countries of Guests:**
* Portugal (491), Spain (103), UK (78), Ireland (28), France (16), USA (14)
* Other countries contribute in smaller numbers.
* **Market Segments:**
* Online TA (454), Offline TA/TO (178), Direct (146), Groups (15), Corporate (10), Complementary (1)
* **Distribution Channels:**
* TA/TO (636), Direct (158), Corporate (10)
* **Reserved Room Types:**
* Type A (398), Type D (177), Type E (102), Others in smaller proportions
* **Assigned Room Types:**
* Type A (284), Type D (185), Type E (122), Others in smaller proportions
* **Deposit Type:**
* No deposit (804)
* **Customer Type:**
* Transient (705), Contract (50), Transient-Party (47), Group (2)

**The histogram and boxplots show that :**

* The outliers have been removed .
* The week of visit is roughly between 27 – 29 indicating that the data was taken from that time
* The lead\_time varies evenly between 0-100 days
* Most of the families had 2 adults and 0 children .
* The Average Daily Rate varied between 60 and 200
* Most of the families did not need any parking space

**The inference derived from the correlation matrix involves:**

* Lead Time and cancellaation rate has a weak positive correlation which indicates that longer booking times leads to more cancellation which is true also . (0.087)
* Booking changes have a weak negative correlation with cancellation . This shows that booking with any kind of modifications are less likely to be cancelled . (-0.159)
* Required paring spaces and cancellations are negatively correlated meaning if a guest requires parking space then he is less likely to cancel the booking
* Weekend stays and weekday stays are moderately correlated meaning people tend to extend their stay to weekdays.
* Lead time is correlated with both types of stays indicating that people book in advance for longer stays.
* Adults have a weak positive correlation with ADR indicating that more adults have a higher Average Daily Rate
* Children and ADR are positively correlated meaning that families with children book expensive rooms.
* ADR and week number are moderately correlated referring to the seasonal price variations
* Children and adults have a positive impact on ADR
* Lead Time and ADR are negatively correlated indicating that later the booking , more the price of room.

**Key Inferences from the Pairplot**

* Seasonal variables like lead\_time show right skewed distribution indicating that few high values pull mean above median.
* Some of the variables have horizontal and vertical lines showing they are more of categorical type .
* Adr and lead\_time have a few extreme values due to their huge range of distributon.
* Most scatter plots do not have linear trends showing that linear regression is not suitable
* The distribution in week\_no is not constant showing that bookings are very much affected by season . There is always a tourist season

**Inferences from Barplots, Violinplots, and Boxplots**

* (Barplot)Class imbalance in categorical features is present . In some variables , one of the variable significantly outnumber the other variables
* (Boxplot) adr and lead\_time indicate that some bookings have extremely high prices or lead\_times .
* (Violinplots) These plots for stays in weekend nights and weekday nights reveal that there is high variablility in how long guess stay which might depend on factors like booking type and season .
* (Violinplots and Boxplots) They indicate that adr is higher for cancelled bookings . Means that expensive bookings have a higher trend of getting cancelled.
* (Boxplots) ADR suggests that room prices vary over time indicating seasonal variation

**Inferences from the Heatmap and Grouped Variable Data**

* Weak correlation of is\_cancelled with lead\_time indicates that advance bookings get cancelled less.
* People often tend to extend their weekend stays over weekdays .
* If a family requires parking space then they are less likely to cancel a booking
* If a family has more children then ADR is higher.