Logo

Description automatically generated

**College of Professional Studies**

**Northeastern University San Jose**

**MPS Analytics**

**Course: ALY6000 - Introduction to Data Analytics**

**Assignment:**

MODULE 1 PRACTICE ASSIGNMENT 1

**Submitted on:**

November 6, 2022

**Submitted to:**  **Submitted by:**

Professor: BEHZAD AHMADI NIKSHITA RANGANATHAN

**ABSTRACT**

Analytics of data is vital for the success of modern companies because the analysts review key performance metrics, help in data interpretation, and use it to determine an effective approach for driving organizational performance. The importance of data analytics is growing across all industries, generating vast amounts of knowledge acumen that can offer insightful information. Data analytics is predicted to rise considerably quicker than the average rate of growth for all occupations between 2019 and 2029.

Data Analytics always helps companies to get an insight into how to develop the business & use data analytics tools to understand the customer and increase productivity. It helps them to store, study and interpret data about the latest trends in the market enabling them to foresee consumer trends, take actions and decisions based on best-available evidence and increase competitiveness.

Online trip planning, booking, and digital communications have created big data in the hospitality industry. Travelers provide details of their trip, check in to post-stay satisfaction reviews, and leave digital footprints. These pieces of information are used by the hotels to better recognize their target demographics and deliver an improved customer experience.

Hotel data analytics discusses the quantitative and qualitative methods used to boost productivity, occupancy charges, marketing tactics, and yield. The hotel data is first extracted and then categorized to identify and analyze the behavioral data and patterns; the techniques used can vary depending on the business requirements. The top hotels focus on these data analytics to help figure out the best hotel strategy for Optimal Revenue Management, Improve Inventory Management, Predict Customer Behaviour Patterns & Demand Forecasting, Effective Targeted Marketing, Competitor Rate Analysis, Review Management, and Reputation, **Better Customer Satisfaction**

**INTRODUCTION**

This project aims to provide an exploratory analysis of the booking demand information of the hotels. The data that is used within the project is sourced from [*Kaggle*](https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand) by the data owner, Jesse Mostipak.

**About this Dataset:** The dataset considered is based on Hotel Booking Demand. This dataset comprises of hotel booking details of 32 attributes and 119390 data observations for both city hotels and resort hotels from July 2015-Aug 2017. The data set contains data such as booking date, reservation status, length of stay, customer type, country, required car parking spaces, number of special requests, etc.

The data from the dataset is imported into RStudio and analyzed to understand several insights of the hotel industry.

Below are the data descriptions of each variable of the data that briefly describe the contents of the data set. The feature of the dataset is as follows:

|  |  |  |
| --- | --- | --- |
| **No** | **Feature** | **Dictionary** |
| 1 | hotel | Hotel type – Resort / City Hotel |
| 2 | is\_canceled | The value (1) signifies if the booking was canceled or not (0) |
| 3 | lead\_time | The difference in days between the booking and arrival dates |
| 4 | arrival\_date\_year | Arrival year of the customer |
| 5 | arrival\_date\_month | Arrival month of the customer |
| 6 | arrival\_date\_week\_number | Arrival week of the customer |
| 7 | arrival\_date\_day\_of\_month | The arrival date of the customer |
| 8 | stays\_in\_weekend\_nights | Number of weekend nights the guest stays in the hotel (Saturday/ Sunday) |
| 9 | stays\_in\_week\_nights | Number of weekday nights the guest stays in the hotel (Monday - Friday) |
| 10 | adults | Number of Adults staying in the hotel |
| 11 | children | Number of Children accompanying the adults during the stay |
| 12 | babies | Number of babies accompanying the adults during the stay |
| 13 | meal | Meal type –   * Undefined/SC: no meal in the package * BB: Bed & Breakfast meal package * HB: Half board meal package (breakfast or other meal) * FB: Full board meal package which includes all meals |
| 14 | country | Origin Country - Categories are characterized by ISO 3155–3:2013 format |
| 15 | market\_segment | Market Segment |
| 16 | distribution\_channel | Distribution channel used for booking |
| 17 | is\_repeated\_guest | The value (1) signifies if the guest is repeated or not (0) |
| 18 | previous\_cancellations | Number of cancellations done by the customer previously |
| 19 | previous\_bookings\_not\_canceled | Number of bookings not canceled by the customer previously |
| 20 | reserved\_room\_type | Type of room reserved by the customer |
| 21 | assigned\_room\_type | Type of room assigned to the customer |
| 22 | booking\_changes | Number of changes to the current booking |
| 23 | deposit\_type | Type of deposit by the customer to confirm the booking. The three categories are :   * No Deposit: No payment was not made * Non-Refund: Payment is non-refundable * Refundable: Payment is refundable |
| 24 | agent | Travel agency related to the booking |
| 25 | company | The company related to the booking |
| 26 | days\_in\_waiting\_list | Waiting days before the booking got confirmed with the customer |
| 27 | customer\_type | Customer type visiting the hotel. The four categories are:   * Contract: Booking on a contract basis * Group: Booking on a group basis * Transient: Booking is not linked with other transient bookings * Transient party: Booking is transient and is linked with other transient bookings |
| 28 | adr | Average Daily Rate – calculated by dividing the total of all transactions by the number of stays |
| 29 | required\_car\_parking\_spaces | Requests made by the customers for parking spaces for car |
| 30 | total\_of\_special\_requests | Special Requests made by the customers |
| 31 | reservation\_status | Status of the booking. The three categories are:   * Canceled: Customer canceled the booking * Check-Out: Customer visited and has already checked out * No-Show: The customer did not check-in |
| 32 | reservation\_status\_date | The date at which the reservation status was updated |

*Table 1: Dictionary with the Features of the Hotel Booking Demand dataset*

The purpose of this project is to provide an exploratory analysis of the data in the hotel industry. The following are the questions derived to do the data analysis which is presented in a variety of data visualizations:

1. Which category of customers preferred a particular hotel type?

2. Which type of customers prefer a particular meal type?

3. What is the Average Daily Rate (ADR) throughout a year by a company and individual?

4. How the lead times vary for the years 2015, 2016, and 2017?

5. From which countries do customers come from? How many bookings are done worldwide by the top 20 countries?

6. How many bookings done by repeated guests over three years?

7. How many customers have preferred to stay in the hotel during weekdays & weekends?

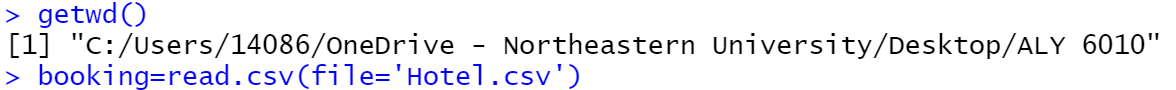
8. Which distribution channels were highly preferred for booking?

9. What is the correlation between different features?

**DATA CLEANING AND MANIPULATION**

* Importing the Hotel Booking Demand CSV file

<booking> vector contains the details of the Hotel booking by the customer.



**Figure 1-read.csv()**

* Analyzing the booking dataset

The str() function displays the various datatypes for the variables

It can be observed that the variables of the dataset are of datatype character and integer.

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

**Figure 2 – str()**

dim() views the number of columns and rows of a dataset. In this case, we have 119390 observations and 32 features.

A picture containing text

Description automatically generated

**Figure 3 – dim()**

colnames() shows the names of all the columns in the dataset.

A screenshot of a computer

Description automatically generated with medium confidence

**Figure 4 – colnames()**

* Manipulating the dataset
* Combining columns

The date, month, and year were concatenated together in one column.



**Figure 5 –Arrival date, month and year**

Then, I added the values in the columns babies and children and named it “Kids”.



**Figure 6 – Number of children and babies in one column**

* Replaced columns values by using ifelse() with conditions

If the reserved and assigned room types are the same, “Same” was added to another column (Reserved VS Assigned), and if not, “Different” was inserted.

The company column had many NULL values which indicated that no company was associated while booking the hotel. So, ifelse() was applied to substitute NULL values with “Individual”.





**Figure 7 – ifelse()**

* Dropping columns

Columns agent, arrival week, arrival date, arrival month, reservation status date, market segment, children, babies, assigned and reserved room type were eliminated from the dataset.

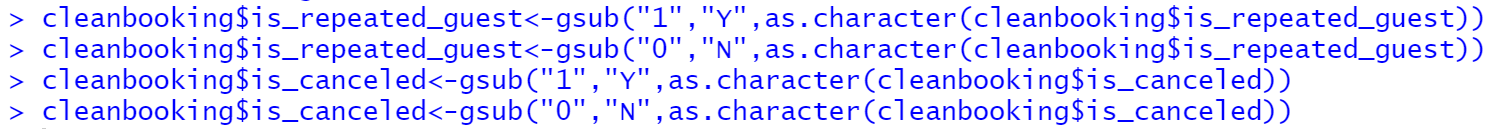
Text

Description automatically generated

**Figure 8 – names() function and ! operator**

* Replacing more column values

In both the columns, is repeated guest and is canceled, I switched the values “1” to “Y” and “0” to “N” with the help of gsub().



**Figure 9 – gsub()**

* Deleting rows with inaccurate data

Country and Meal columns had some NULL and Undefined values.

Also, some observations had no adults and no kids which seemed incorrect.

All these unclear records were removed.



Graphical user interface, text

Description automatically generated with medium confidence

**Figure 10 – grepl() and subset()**

* Removing duplication

I checked for identical rows in the dataset and found some which were eliminated by distinct(). Distinct() returned all the unique rows. We can confirm using anyDuplicated() that there are no more duplicate rows in the booking dataset.

Text

Description automatically generated with medium confidence

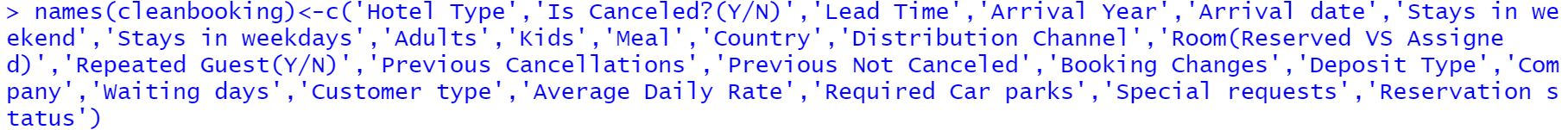
A picture containing text

Description automatically generated

**Figure 11 – duplicated(),distinct() and anyDuplictated()**

* Revising the column names

New column names are given for a better understanding.



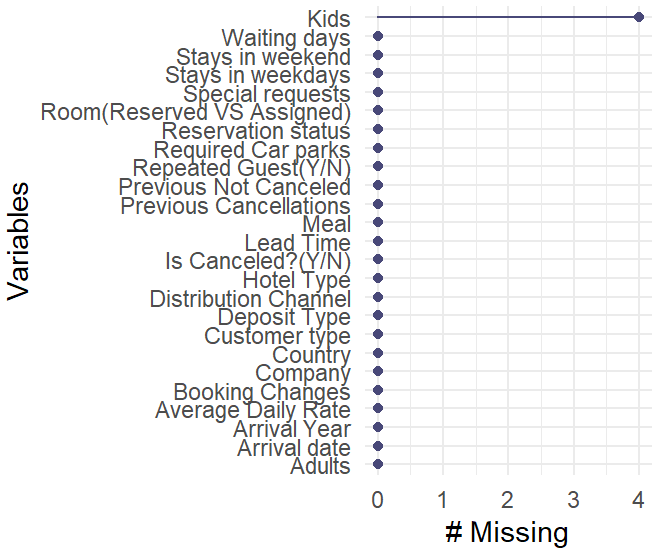
**Figure 12 –names()**

* Visualization of <NA> values

gg\_miss\_which is a function under the naniar package and is used to check the column that has the blanks.

miss\_var\_summary() summarizes the missingness under each variable. The Kids column has 4 missing values

gg\_miss\_var() shows a plot with the number of missing values.

**Text

Description automatically generated**

**A picture containing text, receipt, screenshot

Description automatically generated**

**Figure 13 –<NA> value representation**

* Eliminating the <NA>s

Text

Description automatically generatedThe number of rows reduced from 119390 to 85518.

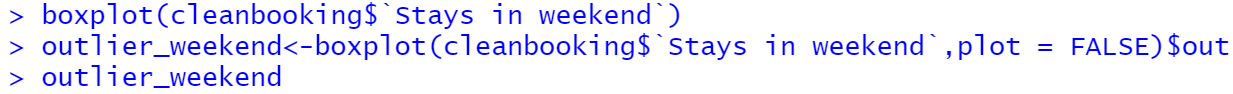


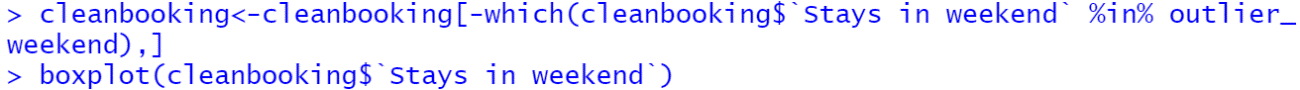
**Figure 14 - Code for obtaining clean Netflix dataset without <NA> values**

* Removing the outliers

Outliers are the values that do not belong to a pattern that the rest of the dataset follows.

Outliers are usually detected using boxplots. I removed all the outliers from many columns using which().

****

****

**Figure 15 – which()**

* Changing datatypes

as.factor() changes the data type of the columns from one datatype to factor.

Text, letter

Description automatically generated

**Figure 16 – as.factor()**

**DATA ANALYSIS**

* Skew and Kurtosis of variables

Skewness and Kurtosis are a part of the moments package.

Kurtosis of Average Daily rate, lead time, and stays in the weekend are less than 3 which indicates that their distribution is flat and less peaked (Platykurtic). On the hand, stays in weekdays has a positive kurtosis value i.e more than 3 (Leptokurtic).

Average daily rate and stays in weekend columns have values between -0.5 and 0.5 which signifies a symmetric distribution. Whereas the other two columns (Lead time and stays in weekdays have a moderately skewed distribution.

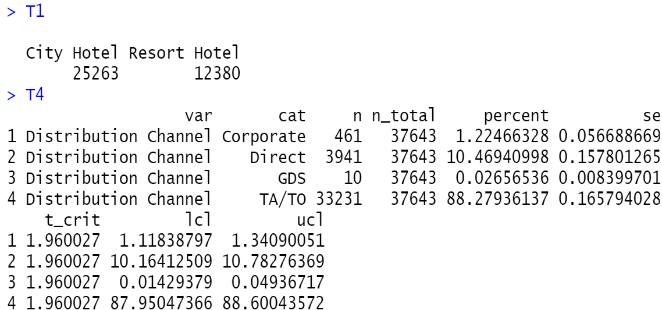
Text

Description automatically generated

**Figure 17 – skewness() and kurtosis()**

Frequency Tables

I created frequency tables for Hotel type, Meal, Country, Distribution channel and Deposit type utilizing table() and freq\_table()



Text

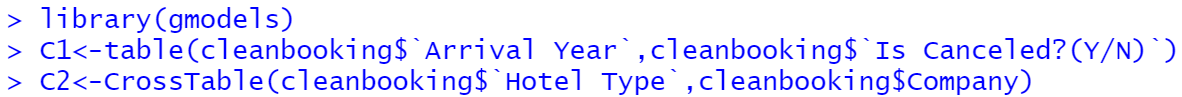
Description automatically generated

**Figure 18 –table(), freq\_table () and examples**

* Crosstables

Crosstable() is a function in gmodels that generates crosstabulations between variables.

I also used table() here to





Table

Description automatically generated

Text

Description automatically generated

**Figure 19 –table(), Crosstable() and examples**

* Histograms

I made histograms using hist() and plotly.

The first histogram shows the distribution of lead time.

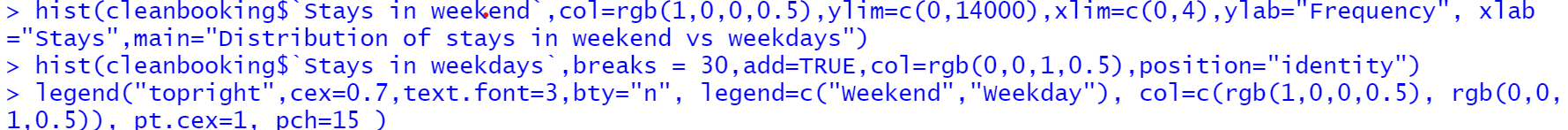


Chart, histogram

Description automatically generated

**Figure 20 –Histogram A**

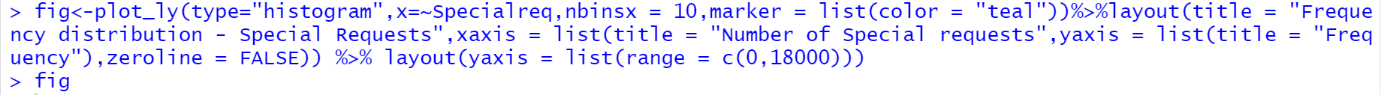
Next one depicts the frequency information of the hotel stays in weekends and weekdays by the customer.



Chart, bar chart

Description automatically generated

**Figure 21 –Histogram B**

The last histogram gives an overview of the number of special requests by customers.

Chart, bar chart

Description automatically generated

**Figure 22 –Histogram C**

* Cleanbooking dataset

Cleanbooking is the dataset after finishing the cleaning and manipulation process. We have 37643 rows and 25 columns

glimpse() is part of the dplyr package and we can see the preview of columns of the dataset with the help of this function.

skim() is useful for getting a statistical summary of the features

Text, letter

Description automatically generated

Table

Description automatically generated with medium confidence

**Figure 23 – glimpse() and skim()**

Table

Description automatically generated with medium confidence

**DATA VISUALIZATION**

* Chart, bar chart

  Description automatically generatedPlot 1

**Figure 24 – Stacked Bar graph A**

Firstly, the data analysis on the hotel industry shows that the highest number of “Transient” category of customers have booked the city type of hotel compared to the resort type. Next is the “Transient-party” type and followed by the customer category of “Contract” and the least being the “Group” type of customers. Surprisingly, corporate or group guests use company discounts and benefits to stay at the hotel but have not booked rather guests who are traveling individually, on their own have booked the hotels more. We need to check reliability & quality test on data due to this reason.

* Chart, bar chart, treemap chart

  Description automatically generatedPlot 2

**Figure 25 – Stacked Bar graph B**

According to the data analysis, the most of customers who booked the city hotel preferred “Bed & Breakfast” compared to the customers of the Resort hotel. It is also noted that no customers preferred “Full Board” which needs a customer satisfaction survey on their opinions on the services they are willing for and accordingly the industries should focus on providing flexible packages to attract more customers to book “Full board”. Customers have ordered “No Meal” as well which indicates a possibility of many restaurants nearby the hotels.

* Chart, histogram

  Description automatically generatedPlot 3

**Figure 26 – Histogram with density curve**

The density of the Average Daily Rate (ADR) is comparatively more than the hotel booked through individuals. The analysis shows that ADR in companies was monitored closely to check the property’s overall performance, its revenue growth over time, and the competition which is lacking in individual bookings so they need to set revenue goals that will increase the profits, focus on targeted marketing strategies to boost the average daily rate.

* Chart, bubble chart

  Description automatically generatedPlot 4

**Figure 27 – Bubble Chart**

In the year 2016, data shows that approximately 550 customers booked with a lead time of 1 day which might have affected the quality of service provided by the hotels as they don’t have sufficient time for labor planning and service implementation. In the year 2016, 200 customers booked with a lead time being 18 days which is again a threat as cancellations can go high. In the years 2015 & 2017, lead time showed that very less customers booked the hotels.

* Map

  Description automatically generatedPlot 5

**Figure 28 – Map**

The data analysis based on World Visitors Choropleth Map demonstrates the analysis of the number of bookings and their origin country along with the top 25 countries. More visitors are from Portugal being 9188 customers, the next UK being 5088 then France with 4761 customers. Next is Denmark and Spain with nearly 2280. More Customers from the Western part of Europe who have booked the hotels show that they travel a lot.

* Chart, scatter chart

  Description automatically generatedPlot 6

**Figure 29 – Dot plot**

The data shows that there are very few repeated clients only 0.26% which shows the hotels should focus on loyalty aspects and create a better strategy to enhance loyalty. They also must create better packages & services to attract customers based on the satisfaction survey to check if their expectations are fulfilled. These could be done on the Western Europe customers who travel a lot as they are the main customers for both hotels. In the year 2016, many non-repeated customers booked the hotels in the month of February, and in the year 2017, they booked in the month of April.

* Chart, bar chart

  Description automatically generatedPlot 7

Chart, waterfall chart

Description automatically generated

**Figure 30 – Animated Grouped bar graph**

Based on the analysis, we could understand that in the city hotels, 39.6% of them did not stay during the weekend, almost 29.8% of them stayed for one- or two days during the weekend, and the percentage drastically reduced to less than 1% as very few stayed for 3-4 days during the weekend. In Resort hotels, 31.7% of the customers did not stay during the weekend and 21.9% of the customers stayed for a day during the weekend. 41.6% of the resort hotel customers stayed 2 days during the weekend and almost less than 3% of them stayed 3- or 4-days weekend. Both the city & resort hotels should find out the reason for this through surveys and improve weekend stays as it is quite normal practice and is not evident in this case.

In city hotels, 7184 customers are the highest to stay for 2 days on weekdays, and the next highest being for 1- or 3- days on weekdays. Similarly, in resort hotels, 2963 customers are the highest who stayed for 5 days on weekdays. Long-term stays of around 7-8 days are not preferred by the customers as obviously people might have visited during holidays to stay on such weekdays.

* Chart, sunburst chart

  Description automatically generatedPlot 8

**Figure 31 – Donut chart**

Finally, the analysis shows that In terms of distribution channels, TA / TO has shown to be 88.3%, followed by the direct channel with the hotel being 10.5%. corporate & GDS bookings are less than 2%. In this case, the use of TA/TO is the strongest distribution channel which may be online and could have provided discounts & special offers during bookings.

* Chart, timeline, treemap chart

  Description automatically generatedPlot 8

Blue color and red color show negative and positive correlations respectively. The plot is according to the Pearson correlation coefficient

**CONCLUSION**

The hotel industry can benefit from using data analytics, as this is the key to marketing approach, building customer loyalty, and improving productivity. Data allows the hotel industry to personalize preferences & experiences for their guests, recognize the correct time to attract customers, introduce fancy hotel pricing tactics, and develop their customer base. The hotel industry will be able to get many ways in which data analytics makes a positive impact on the hotel industry.

Key takeaway points:

1. Transient customers preferred both the hotel types

2. Most customers have requested for Bed & Breakfast meal package in both the hotels

3. Average Daily Rate is a bimodal distribution that can be caused due to merging of data of two different groups-City and Resort hotel

4. Lead time is an important factor for the hotel industry as a longer lead time helps understand the profile of the customers and helps build revenue by giving satisfaction to the customers

5. Maximum number of customers are from European countries (Portugal, Spain, the UK, France, and Germany)

6. Only 0.26% of the customers are repeated guests

7. Around 88.3% of the guests use travel agents to book the trips

**REFERENCES**

Mostipak, J. (2019, November). Hotel booking demand. Kaggle. Retrieved November 11, 2022, from https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

Bluman, A. G. (2018). Elementary Statistics, 10th ed. McGraw Hill.

Kabacoff, R. I. (2011). R in action: Data analysis and graphics with R. Manning Publications Co.

Low-code data app development. Plotly. (n.d.). Retrieved October 28, 2022, from https://plotly.com/