Hotel Booking Analysis

SHUBHAM, AMIT SAXENA
Data science trainees,
AlmaBetter, Bangalore

Abstract:

The present Hotel Data Set includes booking information of City and Resort Hotels from 2015-2017. The data set includes data features such as is_canceled, arrival_date_year, meal, adults, babies, children, stays_in_weekend_nights, is_repeated_guest, stays_in_week_nights, market_segment are considered for performing EDA to dig out some interesting facts related to the data set.

Our EDA analysis provides solutions to different problems such as which type of hotels are favored by visitors, the busiest months of the hotel, why visitors are canceling a hotel booking, from where most of the visitors are coming, and what meals should hotels offers to visitors and many more interesting facts to increase the profits in hospitality services.

Keywords: *EDA, data visualization, data features, matplot, seaborn, pandas, data frame.*

1. Problem Statement

This data set contains booking information for a city hotel and a resort hotel and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. The main objective is to explore and analyze the data to discover important factors that govern the bookings in the hotel, which could help in finding the loopholes to increase the overall productivity of the hospitality industry. This would also help hotels to provide the best offers and facilities to their customers.

- hotel: Type of hotel as City or Resort.
- is_canceled: The number of bookings that are canceled or not.
- arrival_date_year: Year in which visitors come to the hotel.
- market_segment: From which source visitors booked their bookings.
- meal: Type of meal preferred by visitors.
- stays_in_weekend_nights/stays_in_ week_nights: Number of visitors that stay on weekends or weekdays.
- is_repeated_guest: Number of visitors who came back again.
- adults: Number of adult visitors.
- babies: Number of baby visitors.
- children: Number of children visitors.
- deposit_type: Type of deposit as refundable or non-refundable/
- reservation_status: Type of reservation used by visitors.

2. Introduction

The project hotel data set includes the real-world data record of hotel bookings of a city and a resort hotel including details like bookings, guest details cancellations, etc. from 2015 - 2017. The main goal of the project is to understand and visualize the dataset from the hotel and customer's point of view such as reasons for booking cancellations across various parameters, the best time to book the hotel, peak season, and provide suggestions to reduce the cancellations and increase the revenue of hotels.

We will perform exploratory data analysis with python to get insight from the hotel data set.

3. Variables description

Variable	Description
arrival_date_year	Year of arrival date
country	Country of origin of visitors
customer_type	Type of booking such as Contract - When the booking has an allotment or other type of contract associated to it. Group – When the booking is associated with a group booking. Transient – When the booking is not part of a group or contract and is not associated with other transient bookings. Transient-party - When the booking is transient but is associated with at least other transient bookings.
deposit_type	Indication on if the customer deposited to guarantee the booking. No Deposit – no deposit was made. Non Refund – a deposit was made in the value of the total stay cost. Refundable – a deposit was made with a value under the total cost of the stay.
is_canceled	Value indicates if the booking was canceled (1) or not (0)
is_repeated_guest	Value indicates if the booking name was from a repeated guest (1) or not (0)
market_segment	Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators
meal	Type of meal booked such as BB – Bed & Breakfast. HB – Half board (breakfast and one other meal – usual dinner). FB – Full board (breakfast, lunch, and dinner). Undefined/SC – no meal package;
reservation_status	Reservation status Canceled – booking was canceled by the customer. Check-Out – customer has checked in but already departed. No-Show – the customer did not check in and did inform the hotel of the reason why
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/children/babies	Number of adults, children, and babies visiting hotels.

4. Steps involved in **Exploratory Data Analysis**

- Firstly, we load the hotel data set from our drive to the Colab network.
- Afterward, we analyze the data set by looking into the columns and rows to dig out important features that are required for exploring and analyzing the data set.
- After analyzing, we move to the Data cleaning steps to drop the unnecessary columns so that noise can be reduced in our data set.
    - Firstly, we have created a copy of our data set, so that the original data set does not get affected in further steps.
    - Afterward, we found the columns with null values and replaced them with 0.
    - The missing values in the country column are replaced with mode value that appears mostly and the missing children values are replaced with the rounded mean value for proper evaluation of the data set.
    - Lastly, we drop the rows where there is no baby, children, and adult, and convert the datatype of children, company, and agent columns from float to integer for noise-free calculation of the data set.
- Moving forward, now we start Data Visualization on our data set. We tried to find out some interesting facts about the data set.
  The following objectives are covered in our analysis:
    - First Objective: Hotel-wise yearly bookings
    - Second Objective: Finding how many Bookings were canceled or not canceled?
    - Third Objective: Visualizing the number of hotel bookings that were canceled and not canceled Yearly.
    - Fourth Objective: From which Country the most guests are coming?
    - Fifth Objective: Market Segment wise hotel bookings
    - Sixth Objective: Finding the relationship between cancelation and market segmentation
    - Seventh Objective: Finding cancellations with respect to customer types
    - Eight Objective: Busy months for Hotels
    - Nineth Objective: Finding the most booked accommodation types as Family, Couple, Single.
    - Tenth Objective: Finding the Deposit Type with respect to cancelation.
    - Eleventh Objective: Finding the Reservation Status.

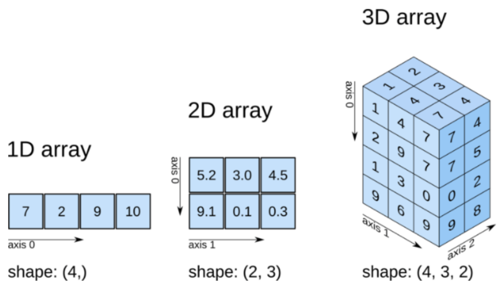
- ➔ Twelfth Objective: Finding visitors that stay on weekends and weekdays.
- ➔ Thirteenth Objective: Visualizing type of visitors with respect to hotel type.
- ➔ Fourteenth Objective: Finding meals for visitors.
- ➔ Fifteenth Objective: Finding Repeat guests in hotelsFinding meals for visitors.

5. Visualization Tools:

1. NumPy:

NumPy is a Python library that is used with arrays. It also has functions for working within the domain of algebra, Fourier transform, and matrices. The array object in NumPy is termed ndarray. The NumPy array can include float or integers value but not both. This enhances the speed of linear algebra calculations. NumPy supports basic operations like average, minimum, maximum, variance, variance, and plenty more. The NumPy array can have dimensions as

- One dimension arrays (1D) represent vectors.
- Two-dimensional arrays (2D) represent matrices.
- And higher dimensional arrays represent tensors.



2. Pandas:

Pandas library is used for exploring and analyzing data sets. It has functions for cleaning, exploring, analyzing, and manipulating data. Pandas library provides analysis of big data and makes conclusions based on statistical theories. Pandas provide solutions to the data such as

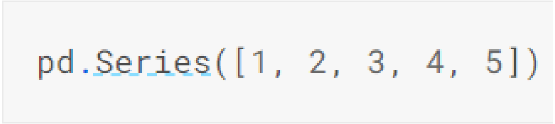
- ➔ Finds a correlation between two or more columns?
- ➔ What are the average/max/min values?

2.1 Creating data:

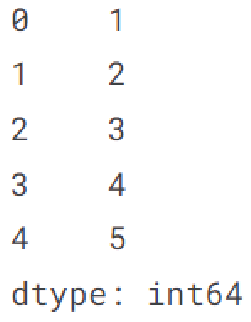
There are two core objects in pandas Series and DataFrame.

Series:

Series is a sequence of data values such as a list.



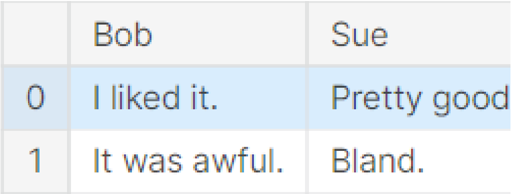
```
pd.Series([1, 2, 3, 4, 5])
```



```
0     1
1     2
2     3
3     4
4     5
dtype: int64
```

DataFrame:

DataFrame is a table. It contains an array of individual entries, each of that has a certain value. Each value corresponds to a column and a row.



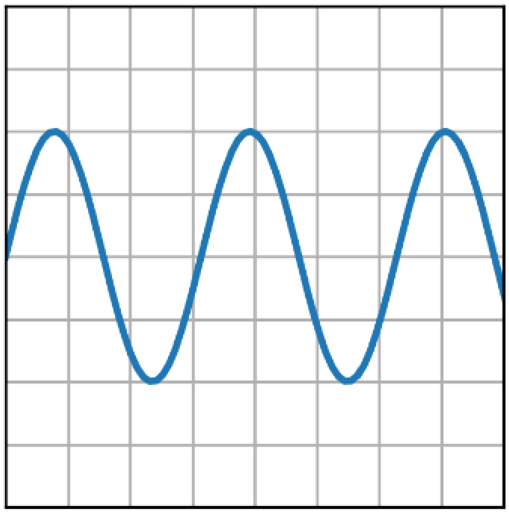
	Bob	Sue
0	I liked it.	Pretty good
1	It was awful.	Bland.

3. Matplotlib:

Matplotlib library creates static, animated, and interactive visualizations in Python. Matplotlib may be used for:

- ➔ Creating publication quality plots.
- ➔ Make interactive figures which will zoom, pan, and update.
- ➔ Customize visual style and layout.
- ➔ Export to many file formats.

The matplotlib visualization allows us visual access to very large amounts of information in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, countplot, histogram, etc.



plot(x, y)

4. Seaborn:

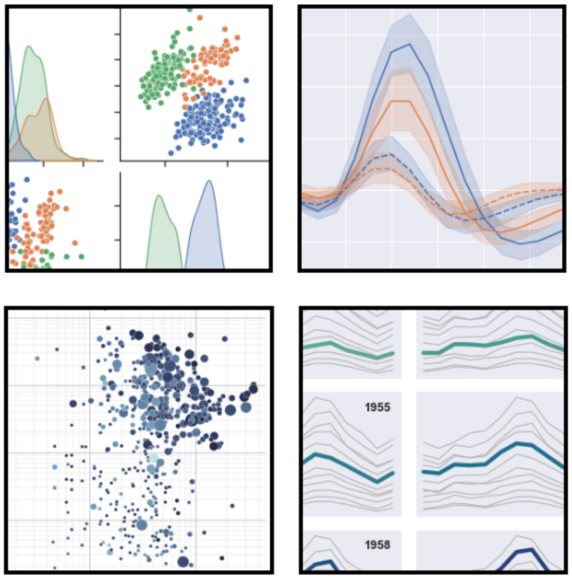
Seaborn is a Python data visualization library supported matplotlib. It generates a high-level interface for drawing attractive and informative statistical graphics. Seaborn aims to create visualization the central part of exploring and understanding data.

4.1. Types of plots in Seaborn:

Plots are used for visualizing the connection between variables. Those variables will be either completely numerical or a category sort of a group, class, or division. Seaborn plot categories are

- ➢ Relational plots: This plot is employed to grasp the relation between two variables. Categorical plots: This plot deals with categorical variables and the way they'll be visualized.
- ➢ Distribution plots: This plot is employed for examining bivariate and univariate distributions.
- ➢ Regression plots: The regression plots in seaborn are primarily intended to feature a visible guide that helps to emphasize patterns in an exceeding dataset during exploratory data analyses.
- ➢ Matrix plots: The matrix plot represents an array of scatterplots.

- Multi-plot grids: It is a useful approach to draw multiple instances of the same plot on different subsets of the dataset.



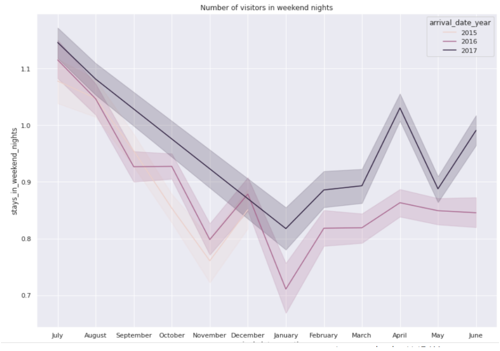
➢ seaborn.lineplot:

Draw a line plot with the possibility of several semantic groupings.

Syntax:

seaborn.lineplot(data=None, x=None, y=None, hue=None, size=None, style=None, units=None, palette=None)

The relationship of x and y are often shown for various subsets of the information using the hue, size, and gnificence parameters. Using redundant semantics (style and hue for the identical variable) can help make graphics more accessible.

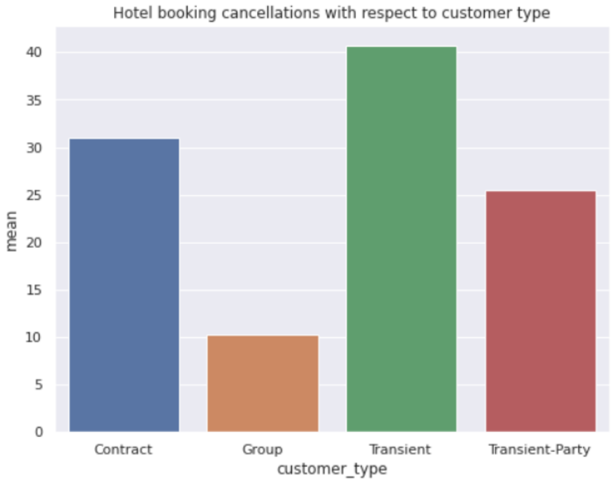


➢ seaborn.barplot:

A bar plot represents an estimate of the central tendency for a numeric variable with the peak of every rectangle and provides some indication of the uncertainty around that estimate using error bars. Bar plots include 0 within the quantitative axis range, and they are a decent choice when 0 may be a meaningful value for the quantitative variable.

Syntax:

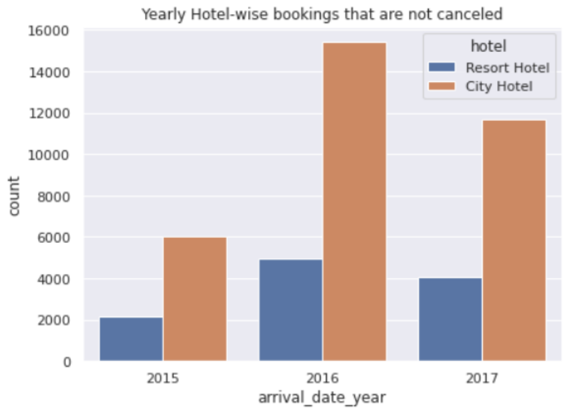
seaborn.barplot(data=None, x=None, y=None, hue=None, order=None)



- **seaborn.countplot:**

  The count plot can be regarded as a histogram across a categorical, instead of quantitative, variable. The basic API and options are identical to barplot(), so we can compare counts across nested variables.

  Syntax:
  seaborn.countplot(data=None, x=None, y=None, hue=None, order=None)

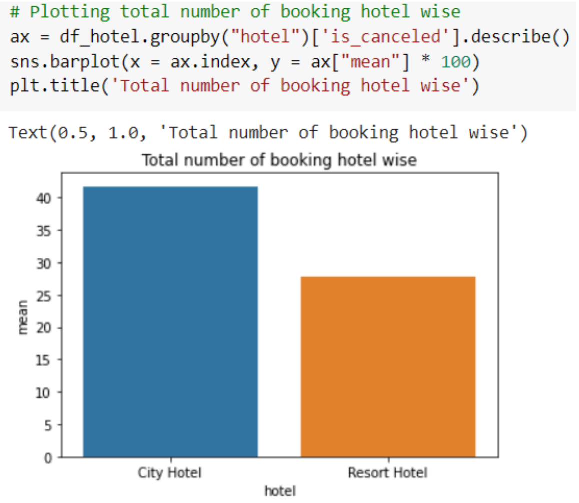


6. Statistical Tools:

6.1 Mean:

The statistical mean is an arithmetic mean that adds up all numbers in a data set and then divides the total by the number of data points.
An arithmetic mean is calculated using the following equation:

$$A := \frac{1}{n}\sum_{i=1}^{n} a_i$$



```
# Plotting total number of booking hotel wise
ax = df_hotel.groupby("hotel")['is_canceled'].describe()
sns.barplot(x = ax.index, y = ax["mean"] * 100)
plt.title('Total number of booking hotel wise')
```

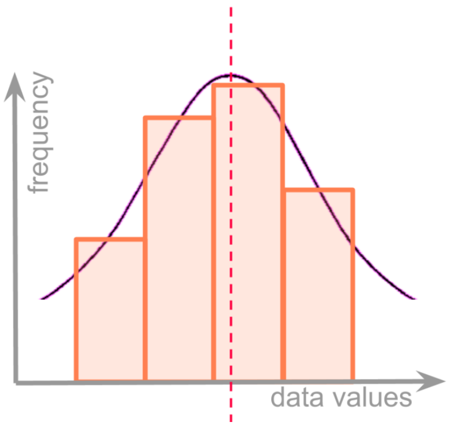
```
Text(0.5, 1.0, 'Total number of booking hotel wise')
```

6.2 Mode:

The mode is the value that comes mostly in a set of data values. In other words, the mode is the value at which the probability mass function takes its maximum value. The mode is calculated using the following equation:

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

where $l =$ lower limit of the modal class,
$h =$ size of the class interval (assuming all class sizes to be equal),
$f_1 =$ frequency of the modal class,
$f_0 =$ frequency of the class preceding the modal class,
$f_2 =$ frequency of the class succeeding the modal class.



7. Conclusion:

7.1 Hotels perspective:

- Overall, People prefer City hotels over Resort Hotels. So, the hospitality industry can create new City hotels and can provide offers to visitors in Resort hotels to engage more visitors.
- People for a shorter duration of stay prefer City hotels over Resort hotels and for longer stays go for Resort hotels.
- From May to August the hotels seem busy. During this period hotels can increase their price for increasing their profits. Although, at the end and start of the year very few visitors arrive, in this period hotels can provide promotional offers to attract visitors.
- Couples with no children and babies prefer to stay in both types of hotels. But couples with babies prefer to stay in Resort hotels. So, additional couple services can be provided to attract visitors towards hotels. Also, services like baby care can be provided in Resort hotels.
- Visitors mostly prefer online/offline travel agents and tour operators for their hotel bookings. So, hotel advertising can be processed from this channel to increase the range of hotels.
- More than 70% of people prefer BB meals, so menu prices can be increased. As most people prefer to do breakfast in hotels. At the same time, additional offers for lunch and dinners should be added to attract visitors.
- The majority of bookings that are canceled are from a no deposit type that does not require any amount to be deposited, due to this a high cancelation rate is observed. Also, it is interesting to note that refundable deposits had fewer cancellations than non-refundable deposits. Logically one would have assumed that non-refundable deposits have fewer cancellations as hotel rates are usually higher.
- The majority of bookings are transient (booking that is not a part of a contract or group). This shows that Booking online is becoming increasingly consumer friendly. Hotels can advertise and provide offers through this channel to increase their sales.
- Countries like Germany, Italy, Ireland, Belgium, Brazil, and the Netherlands have a very less number of visitors. Since they are in the top 10 countries from where visitors are coming so these country visitors have the potential to increase further if a proper advertising channel is established to engage them.
- The most painful analysis in this data set is a very less number of people repeated, this shows people are not satisfied with the hotel services. Proper feedback should be taken at checkouts. Plus proper connectivity channels should be established with the visitors and hotels should send promotional offers and reminders to the visitors regularly to retain them.

7.2 Visitor's perspective:

- People planning big parties in resort and city hotels can plan at the start or end of the year as there is no rush and they can get heavy discounts in this period.
- People can use online/offline travel agents and tour operators for booking their tickets since they seem to be the most trustworthy while booking hotels and people can also get better deals in the future through this channel.
- People can go for BB meals in hotels since 77% of visitors preferred it.

8. References:

- W3Schools
- Analytics Vidhya
- GeekforGeeks
- Matplotlib, Seaborn documentation
- Science Direct