**Hotel Booking Analysis**

**By- N.SANTOSH KUMAR CHOUDHURY,**

**KURESH CHANDRA TRIPATHY**

**Data science trainees,**

**Alma Better, Bangalore**

# **Hotel Bookings Exploratory Data Analysis**

## **Objective:-**

We are provided with a hotel bookings dataset. Our main objective is performing EDA on the given dataset and draw useful conclusions about general trends in hotel bookings and how factors governing hotel bookings interact with each other.

## **Data Summary:-**

We are given a hotel bookings dataset. This dataset contains booking information for a city hotel and a resort hotel. It contains the following features.

- hotel: Name of hotel (City or Resort)

- is\_canceled: Whether the booking is canceled or not (0 for no canceled and 1 for canceled)

- lead\_time: time (in days) between booking transaction and actual arrival.

- arrival\_date\_year: Year of arrival

- arrival\_date\_month: month of arrival

- arrival\_date\_week\_number: week number of arrival date.

- arrival\_date\_day\_of\_month: Day of month of arrival date

- stays\_in\_weekend\_nights: No. of weekend nights spent in a hotel

- stays\_in\_week\_nights: No. of weeknights spent in a hotel

- adults: No. of adults in single booking record.

- children: No. of children in single booking record.

- babies: No. of babies in single booking record.

- meal: Type of meal chosen

- country: Country of origin of customers (as mentioned by them)

- market\_segment: What segment via booking was made and for what purpose.

- distribution\_channel: Via which medium booking was made.

- is\_repeated\_guest: Whether the customer has made any booking before (0 for No and 1 for Yes)

- previous\_cancellations: No. of previous canceled bookings.

- previous\_bookings\_not\_canceled: No. of previous non-canceled bookings.

- reserved\_room\_type: Room type reserved by a customer.

- assigned\_room\_type: Room type assigned to the customer.

- booking\_changes: No. of booking changes done by customers

- deposit\_type: Type of deposit at the time of making a booking (No deposit/ Refundable/ No refund)

- agent: Id of agent for booking

- company: Id of the company making a booking

- days\_in\_waiting\_list: No. of days on waiting list.

- customer\_type: Type of customer (Transient, Group, etc.)

- adr: Average Daily rate.

- required\_car\_parking\_spaces: No. of car parking asked in booking

- total\_of\_special\_requests: total no. of special request.

- reservation\_status: Whether a customer has checked out or canceled,or not showed

- reservation\_status\_date: Date of making reservation status.

* Total number of rows in data: 119390
* Total number of columns: 32

## **Data Cleaning and Feature Engineering**

### (1) Removing Duplicate rows

All duplicate rows were dropped.

### (2) Handling null values

* Null values in columns company and agent were replaced by 0.
* Null values in column country were replaced by 'others'.
* Null values in column children were replaced by the mean of the column.

### (3) Converting columns to appropriate data types

* Changed data type
* of children, company, agent to int type.
* Changed data type of reservation\_status\_date to date type.

### (4) Removing outliers

* One outlier was found in the adr column. Simply dropped it.

### (5) Creating new columns

* Created new column total\_stay by adding stays\_in\_weekend\_nights+stays\_in\_week\_nights.
* Created new column total\_people by adding adults+children+babies.

## **Exploratory Data Analysis**

Performed EDA and tried answering the following questions:

* Which Year most room booking happen?
* In which month most of the bookings happened?
* Visualizing Hotel wise yearly bookings
* What is the relationship between market segment and cancellation?

# Correlation of the columns.

# Which Country have Most Babies During their Visit?

# Which type of Meal is mostly preferred by the guests during their visit?

# Draw the boxplots of the two columns of stays\_in\_weekend\_nights and stays\_in\_week\_nights in a single plot using the following one-liner code. The method we use is plots box.

Mainly performed using Matplotlib and Seaborn library and the following graph and plots had been used:

* Bar Plot.
* Histogram.
* Scatter Plot.
* Pie Chart.
* Line Plot.
* Heatmap.
* Box Plot

**Code used to Analysis Booking**

* # return highest bookings made in a year

highest\_bookings= df\_copy.groupby(['arrival\_date\_year'])['arrival\_date\_year'].agg({'count'}).reset\_index().rename(columns={'count': "Most\_Bookings" }).sort\_values(by='Most\_Bookings',ascending=False)

* # bookings made in a year

top\_ten\_highest\_bookings=highest\_bookings[:]

top\_ten\_highest\_bookings

* #set plotsize

plt.figure(figsize=(7,8))

* #plotting of bar

sns.countplot(x=df\_copy['arrival\_date\_year'],order=df\_copy['arrival\_date\_year'].value\_counts().index)

* # return highest bookings made in a year

country\_list= df\_copy.groupby(['country'])['country'].agg({'count'}).reset\_index().rename(columns={'index': "country" }).sort\_values(by='count',ascending=False)

country\_list.drop(country\_list[country\_list['country']==0].index,inplace=True)

country\_list=country\_list[:20]

country\_list

plt.xlabel('Year')

* # set y-label in the plot

plt.ylabel('Number of Booking')

* # set title for the plot

plt.title("Year wise Booking")

ax= df.country.value\_counts().head(10).plot (kind= 'bar');

for p in ax.patches:

    ax.annotate(str(p.get\_height()), (p.get\_x() \* 1.005, p.get\_height() \* 1.005))

* import folium
* import plotly.express as px
* # searched on google

plt.figure(figsize = (18,12))

basemap = folium.Map()

guests\_map = px.choropleth(country\_list, locations = country\_list['country'],color = country\_list['count'], hover\_name = country\_list['country'])

guests\_map.show()

* # return highest bookings made in a year

Hotel\_list= df\_copy.groupby(['hotel'])['hotel'].agg({'count'}).reset\_index().rename(columns={'index': "Serial" }).sort\_values(by='count',ascending=False)

* # country\_list.drop(countrylist[country\_list['country']==0].index,inplace=True)
* # country\_list=country\_list[:20]

print(Hotel\_list)

* # Visualizing by  plotting the graph

plt.figure(figsize=(6,10))

sns.barplot(x=Hotel\_list['hotel'],y=Hotel\_list['count'])

plt.xlabel('hotel')

plt.ylabel('Number of guests',fontsize=12)

plt.title("Hotel types with number of booking")

* # return highest bookings made in a year
* bookings\_room= df\_copy.groupby(['reserved\_room\_type'])['reserved\_room\_type'].agg({'count'}).reset\_index().rename(columns={'count': "No\_of\_Bookings" })
* # bookings made in a year

booking\_room\_type=bookings\_room[:]booking\_room\_type

* booking\_room\_type['No\_of\_Bookings'].plot.pie(explode=[0.05]\*10,shadow=True,autopct='%1.1f%%',figsize=(15,20),fontsize=15,labels=None)

labels1=booking\_room\_type['reserved\_room\_type'].value\_counts().index.tolist()

labels=booking\_room\_type['No\_of\_Bookings'].value\_counts().index.tolist()

Lab3=(labels1+labels)

* plt.title('% Distribution of Customer Type')
* plt.legend(bbox\_to\_anchor=(0.85, 1), loc='upper left', labels=labels1)
* # return highest bookings made in a year

agent\_no= df\_copy.groupby(['agent'])['agent'].agg({'count'}).reset\_index().rename(columns={'count': "No\_of\_Bookings" }).sort\_values(by='No\_of\_Bookings',ascending=False)

* # bookings made in a year

booking\_by\_agent=agent\_no.iloc[0:20].sort\_values(by='agent',ascending=True)

booking\_by\_agent

* # Visualizing by  plotting the graph

plt.figure(figsize=(20,8))

sns.barplot(x=booking\_by\_agent["agent"],y=booking\_by\_agent['No\_of\_Bookings'])

plt.xlabel('agent\_code')

plt.ylabel('Number of Room Booking',fontsize=12)

plt.title("Number of Room Booking by top 10 agent")

* #plot Box & ploat graph for analysis data on highest adr

fig, axes = plt.subplots(1, 2, figsize=(10,6))

grp\_by\_room = df\_copy.groupby('assigned\_room\_type')

sns.countplot(ax = axes[0], x = df\_copy['assigned\_room\_type'])

sns.boxplot(ax = axes[1], x = df\_copy['assigned\_room\_type'], y = df\_copy['adr'])

plt.show()

* # set plot size
* plt.figure(figsize=(20,8))
* #pltting lineplot on x- months & y- booking counts

sns.lineplot(x=bookings\_by\_months\_df1['arrival\_date\_month'],y=bookings\_by\_months\_df1['Counts'])

* # set title for the plot

plt.title('Number of bookings across each month')

* #set x label

plt.xlabel('Month')

* #set y label

plt.ylabel('Number of bookings')

* #yearly booking by hotel types

plt.figure(figsize = (18,12))

sns.set\_style('whitegrid')

matplotlib.rcParams['font.size'] = 16

matplotlib.rcParams['figure.figsize'] = (11, 7)

matplotlib.rcParams['figure.facecolor'] = '#00000000'

sns.countplot (x= 'arrival\_date\_year', data= df, hue= 'hotel').set\_title ('yearly bookings')

plt.figure(figsize = (18,12))

plt.title("Booking Canceled or not by market segment")

sns.countplot(x='market\_segment',data=df\_copy ,hue='is\_canceled')

* # scatter plot

plt.figure(figsize = (15,5))

a = df.groupby("market\_segment")['is\_canceled'].describe()

sns.scatterplot(x=a.index, y=a["mean"] \* 10)

* #ploting the graph for all columns in a heat map

plt.figure(figsize=(35,20))

sns.heatmap(df.corr(),annot=True)

plt.title('Co- relation of the columns')

* # return highest bookings made in a year

highest\_babies= df\_copy.groupby(['country'])['country'].agg({'count'}).reset\_index().rename(columns={'count': "babies" }).sort\_values(by='babies',ascending=False)

* # Booking done by Country with highest number of babies.

top\_ten\_highest\_babies=highest\_babies[:10]

top\_ten\_highest\_babies

* #set plotsize

plt.figure(figsize=(10,8))

* #plotting of bar

sns.countplot(x=df\_copy['country'],order=top\_ten\_highest\_babies['country'].value\_counts().index)

* #  set x-label in plot

plt.xlabel('country')

* # set y-label in the plot

plt.ylabel('babies')

* # set title for the plot

plt.title("Most Babies During Thier Visit from Country")

* # Visualizsing the by pie chart.

df\_copy['meal'].value\_counts(). plot.pie(explode=[0.08]\*5, autopct='%1.1f%%', shadow=True, figsize=(10,14),fontsize=20)

plt.title('Pie Chart for Most Preffered  meal')

**Conclusion: -**

* Customer visiting from Country Portugal has most number of Babies.
* Most of Booking done in year 2016 i.e. 56,707 number of booking.
* Meal is mostly preferred by the guests during their visit is BB type which is 77.8% of all type of meal.
* Most number of booking coming from Country Portugal.
* Most number of booking for stays\_in\_week\_nights.
* High number of booking happens in Western Europe country.
* Most number of booking done in month of AUGUST.
* City Hotel has highest number of booking i.e. 79330 numbers.
* TA type of market-segment has most number of cancellations.
* ‘A ‘Type room has highest number of booking i.e. 72%.
* Agent no-9 is most valuable agent.
* Most demanded room type is A, but better adr rooms are of type H, G and C also.
* Hotels should increase the no. of room types A and H to maximize revenue
* Bookings made via complementary market segment andadults have on average high no. of special request.
* For customers, generally the longer stays (more than 15 days) can result in better deals in terms of low adr.

**Challenges: -**

1. There was a lot of duplicate data.
2. Data was present in wrong datatype format.
3. Choosing appropriate visualization techniques to use was difficult.
4. A lot of null values were there in the dataset.

.