AIRBNB CASE STUDY

By: Anuja Phadtare

For this case study I have used Jupyter notebook for the data inspecting, data cleaning and initial analysis of the data and Tableau for data analysis and visualization.

- **Details of Initial Analysis using Jupiter Notebook:**
  - Initially I have inspected the data. Details of which are given below:
    Data Set Used: AB_NYC_2019.csv
    Number of Rows: 48895
    Number of Columns: 16
  - Further, removed the columns like Id, Name, Last Review which were not that much of a relevant information for our analysis.
  - Checked for the Duplicate rows in our dataset and no duplicate data was found.
  - Checked the Null Values in our dataset and found that columns like name, host-name, last review and review-per-month have null values.
  - Dropped the column 'name' as missing values were less and dropping it wouldn't have significant impact on analysis.
  - Checked the formatting in our dataset.
  - Identified and reviewed the outliers.

*Snapshots of the Jupyter notebook given below.

In [1]:
```
# Import the necessary libraries
import warnings
warnings.filterwarnings("ignore")
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

In [2]:
```
# Data conversion and Understanding
airbnb = pd.read_csv("AB_NYC_2019.csv")
airbnb.head(5)
```

Out[2]:

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_revie
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

In [3]:
```
# Check the rows and columns of the dataset
airbnb.shape
```

Out[3]: (48895, 16)

- The dataset contains 48895 rows and 16 columns
- Now we have to check whether there are any missing values in the dataset

In [4]:
```
# Calculating the missing values in the dataset
airbnb.isnull().sum()
```

Out[4]:
```
id                                    0
name                                 16
host_id                               0
host_name                            21
neighbourhood_group                   0
neighbourhood                         0
latitude                              0
longitude                             0
room_type                             0
price                                 0
minimum_nights                        0
number_of_reviews                     0
last_review                       10052
reviews_per_month                 10052
calculated_host_listings_count        0
availability_365                      0
dtype: int64
```

In [5]:
```
# Now we have the missing values, there are certain columns that are not efficient to the dataset
airbnb.drop(['id','name','last_review'], axis = 1, inplace = True)
```

In [6]:
```
# View whether the columns are dropped
airbnb.head(5)
```

Out[6]:

	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	reviews_per_month
0	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	0.21
1	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	0.38
2	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN
3	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	4.64
4	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	0.10

In [7]:
```
airbnb.reviews_per_month.isnull().sum()
```

Out[7]: 10052

In [8]:
```
# Now reviews per month contains more missing values which should be replaced with 0 respectively
airbnb.fillna({'reviews_per_month':0},inplace=True)
```

In [9]:
```
airbnb.reviews_per_month.isnull().sum()
```

Out[9]: 0

In [10]:
```
# There are no missing values present in reviews_per_month column
# Now to check the unique values of other columns'
airbnb.room_type.unique()
```

Out[10]: array(['Private room', 'Entire home/apt', 'Shared room'], dtype=object)

In [11]:
```
len(airbnb.room_type.unique())
```

Out[11]: 3

In [12]:
```
airbnb.neighbourhood_group.unique()
```

Out[12]:
```
array(['Brooklyn', 'Manhattan', 'Queens', 'Staten Island', 'Bronx'],
      dtype=object)
```

In [13]:
```
len(airbnb.neighbourhood_group.unique())
```

Out[13]: 5

In [14]:
```
len(airbnb.neighbourhood.unique())
```

Out[14]: 221

In [15]:
```
airbnb.to_csv(r'C:\Users\Prasad\Downloads\Airbnb NYC_Case study-Anuja Phadtare\AB_NYC_2019.csv',index=False, header=True)
```

❖ **Data Analysis and Visualizations using Tableau:**

I have used Tableau to visualize the data for the assignment. Below are the detailed steps used for each visualization.

Methodology Document for presentation no.1:

1. **Top 10 Hosts:**
   Here I identified the top 10 Host Ids, Host Name with count of Host Ids using the tree map. I used the filter option here to find out the top 10 Host Ids.

2. **Preferred Room type with respect to the Neighbourhood groups:**
   Here I used the pie chart for understanding the data. I took the percentage of room type preferred with respect to the Neighbourhood group. Then added Room Type to the 'Color' mark to highlight the different Room Type in different colours and 'Count' of Host Id to the 'Size' mark card.

3. **For Variance of price with Neighbourhood Groups:**
   Now I used a box and whisker's plot with the Neighbourhood Groups in Columns and Price in Rows. Then changed the Price from a Sum Measure to the median measure.

4. **Average price of Neighbourhood groups:**
   For this criteria, I created a bubble chart with Neighbourhood Groups in Columns and Price column in Rows. Then added the Neighbourhood Groups to the 'Color' mark card to highlight the different neighbourhood Groups in different colors and average price in the 'Label' mark.

5. **Customer Booking with respect to Minimum nights:**
   Further, created the bin for Minimum nights as shown below:

The bins display the distribution of minimum nights based on the number of bookings for each neighbourhood group.

6. **Popular Neighbourhoods:**
   Considering the Neighbourhood in rows and sum of reviews in column and dragged the 'Neighbourhood groups' in 'colour' mark. I used the filter to show Top 20 neighbours as per the sum of reviews.

7. **Neighbourhood vs Availability:**
   For this I created a dual axis chart using bar chart for availability 365 and line chart for price for top 10 neighbourhood group sorted by price.

<u>Methodology Document for presentation no.2:</u>

1. **Room type with respect to Neighbourhood group:**
   For this, I created a pie chart to understand the percentage of room type preferred with respect to neighbourhood group. Room Type was added to the 'Color' mark card to highlight the different Room Type in different colours and count of Host Id to the 'Size' mark card.

2. **Customer Booking with respect to minimum nights:**
   In this, bin for Minimum nights was created as shown below:

The bins display the distribution of minimum nights based on the number of bookings for each neighbourhood group.

3. **Neighbourhood vs Availability:**
   A dual axis chart using bar chart for availability of 365 days and line chart for price for top 10 neighbourhood group sorted by price was created.

4. **Price range preferred by Customers:**
   I have taken the pricing preference based on the volume of bookings done in a price range and number of Ids to create a bar chart. A bin for Price column with interval of $20 was created in this.

5. **Understanding Price variation with respect to Room Type & Neighbourhood:**
   For this, Highlights Table chart by taking Room Type in rows & Neighbourhood Group in column was created. Average price was dragged to the 'Color' marks card to highlight the different room type in a different colour.

6. **Price variation with respect to Geography:**
   Here I used the Geo location chart to plot Neighbourhood and Neighbourhood Group in map to showcase the variation of prices across.

7. **Popular Neighborhoods:**
   Here I took the Neighbourhood in rows and sum of reviews in column and dragged the Neighbourhood Groups in 'color' mark card for colour variation. I also used the filter to show Top 20 neighbours as per the sum of reviews.