AIRBNB Case Study IIIT-B

By: Priyanshi Srivastava, Athira Anil & Kowshik Sharma

Methodology Document PPT 1:

In our Airbnb case study, we utilized Jupyter Notebook for initial data analysis and seamlessly transitioned to Tableau for advanced data exploration and visualization, ensuring a comprehensive and insightful examination of the dataset.

Initial Analysis using Jupiter Notebook: Data Set Used: AB_NYC_2019.csv

Number of Rows: 48895

Number of Columns: 16

```
# 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
```

```
# Data conversion and Understanding
airbnb = pd.read_csv("AB_NYC_2019.csv")
airbnb.head(5)
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_revi
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	

```
# Check the rows and columns of the dataset
airbnb.shape
```

```
(48895, 16)
```

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

```
# Calculating the missing values in the dataset
airbnb.isnull().sum()
```

```
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
```

```
# 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)
```

```
# View whether the columns are dropped
airbnb.head(5)
```

	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	

Step 1: We pruned non-informative columns such as Id, Name, and Last Review to streamline our dataset, focusing on relevant information for a more concise and meaningful analysis.

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

```
airbnb.reviews_per_month.isnull().sum()
```

```
0
```

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

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

```
len(airbnb.room_type.unique())
```

```
3
```

```
airbnb.neighbourhood_group.unique()
```

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

```
len(airbnb.neighbourhood_group.unique())
```

```
5
```

```
len(airbnb.neighbourhood.unique())
```

```
221
```

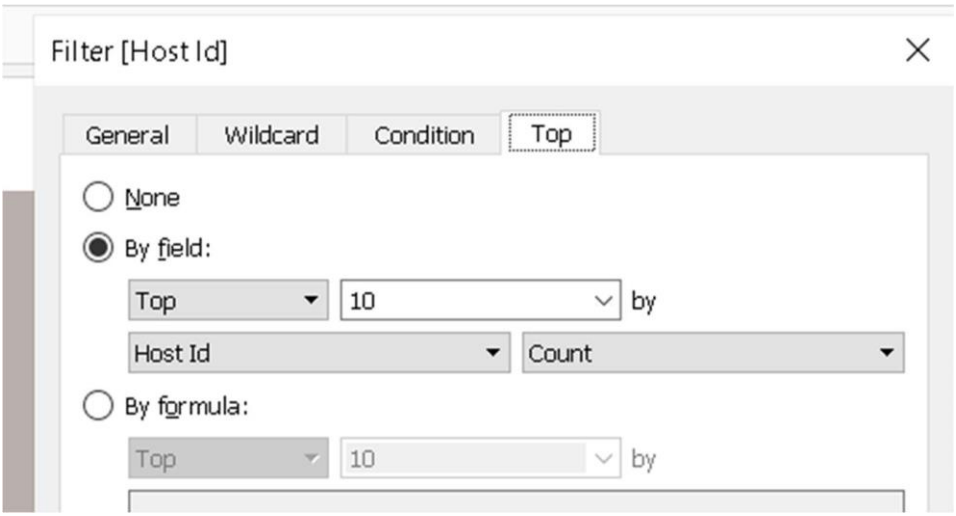
Step 2: Data Wrangling:

- Checked the Duplicate rows in our dataset and no duplicate data was found.
- Checked the Null Values in our dataset. Columns like name, hostname, last review and review-per-month have null values.
- We've dropped the column name as missing values are less and dropping it won't have significant impact on analysis.
- Checked the formatting in our dataset.
- Identified and review outliers.

Data Analysis and Visualizations using Tableau:

We have used tableau to visualize the data for the assignment. Below are the detailed steps used for each visualization.

1) **Top 10 Host:**
   - We identified the top 10 Host Ids, Host Name with count of Host Ids using the tree map.



2) **Preferred Room type with respect to Neighborhood group:**
   - We created a pie chart for understanding the percentage of room type preferred w r t neighborhood group.
   - We added Room Type to the colors Marks card to highlight the different Room Type in different colors and count of Host Id to the size.

3) **For Variance of price with Neighborhood Groups:**
   - We used a box and whisker's plot with Neighborhood Groups in Columns and Price in Rows.
   - We changed the Price from a Sum Measure to the median measure.

4) **For Variance of price with Neighborhood Groups:**
   - We used a box and whisker's plot with Neighborhood Groups in Columns and Price in

Rows.

- We changed the Price from a Sum Measure to the median measure.

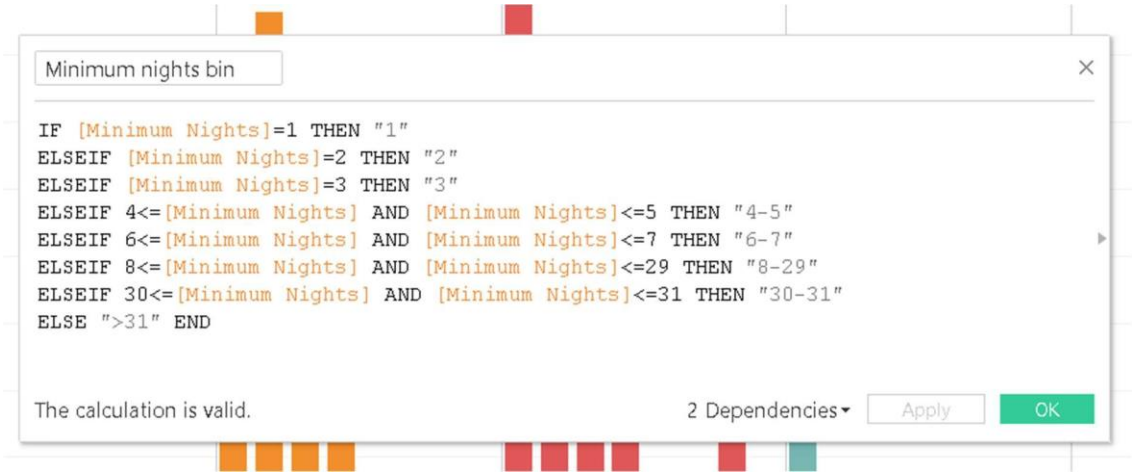
5) Average price of Neighborhood groups:

• We created a bubble chart with Neighborhood Groups in Columns and Price column in Rows.

• We added the Neighborhood Groups to the colors Marks card to highlight the different neighborhood Groups in different colors. Also Put Avg price in Label.

6) Customer Booking w r t minimum nights:

- We created the bin for Minimum nights as shown below.



- The bins were used to display the distribution of minimum nights based on the number of ids booked for each neighborhood group.

7) **Popular Neighborhoods:**

- We took neighborhood in rows and sum of reviews in column and took neighborhood groups in color.
- We used filter to show Top 20 neighbors as per the sum of reviews.

8) Neighborhood vs Availability:

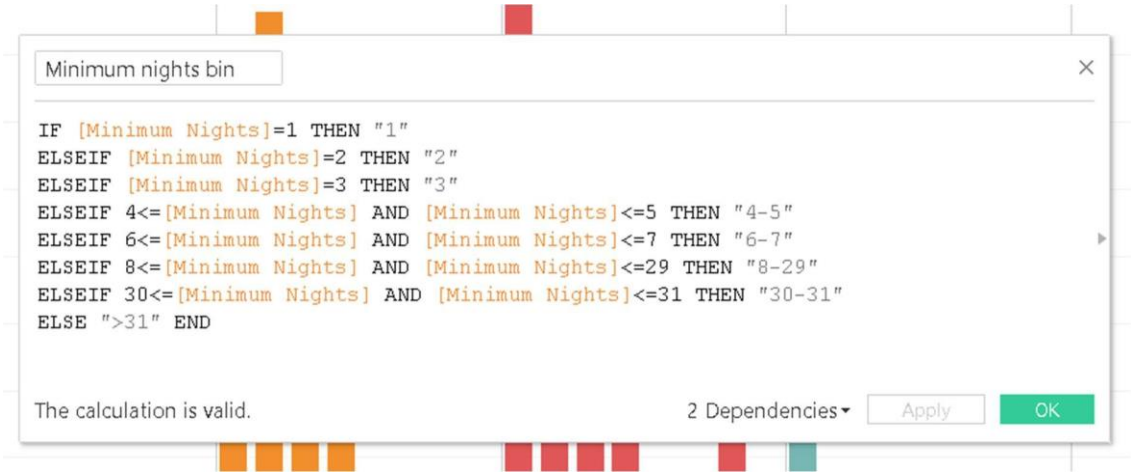
- We created a dual axis chart using bar chart for availability 365 and line chart for price for top 10 neighborhood group sorted by price.

Methodology Document PPT 2:

1) **Room type with respect to Neighborhood group:**
   - We created a pie chart for understanding the percentage of room type preferred w r t neighborhood group.
   - We added Room Type to the colors Marks card to highlight the different Room Type in different colors and count of Host Id to the size.

2) Customer Booking with respect to minimum nights:

- We created the bin for Minimum nights as shown below.



- The bins were used to display the distribution of minimum nights based on the number of ids booked for each neighborhood group.

3) Neighborhood vs Availability:

- We created a dual axis chart using bar chart for availability 365 and line chart for price for top 10 neighborhood group sorted by price.

4) Price range preferred by Customers:

- We have taken pricing preference based on volume of bookings done in a price range and no of Ids to create a bar chart. We have created bin for Price column with interval of $20.

5) Understanding Price variation w.r.t Room Type & Neighborhood:

- We created Highlights Table chat by taking Room Type in rows & Neighborhood Group in column.
- We took the average price in color Marks card to highlight the different Room Type in different colors.

6) **Price variation w r t Geography:**

- We used Geo location chart to plot neighborhood, neighborhood Group in map to show case the variation of prices across.

7) **Popular Neighborhoods:**

- We took neighborhood in rows and sum of reviews in column and took neighborhood groups in color.
- We used filter to show Top 20 neighbors as per the sum of reviews.

8) **Tools used:**

- Data cleaning and preparation: Jupiter notebook – Python
- Visualization and analysis: Tableau
- Data Storytelling: Microsoft PPT.

Recommendations

➢Airbnb can concentrate on promoting shared rooms with discounts to increase bookings and also acquire
more private listings.

➢Ample amount and variety of visuals have been used in the presentations for the stake-holders.

➢Data collection team should collect data about review scores so that it can strengthen the later analysis.

➢A clustering machine learning model to identify groups of similar objects in datasets with two or more
variable quantities can be made.

Thank You!