Methodology

AIRBNB Case Study – Sonal Khot, Abhinandan Gupta, Amruta Patil

Methodology Document PPT 1:

In this case study, we utilized Jupyter Notebook for the initial data analysis and preprocessing, and Tableau for in-depth data analysis and visualization to derive actionable insights.

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

Number of Rows: 48895

Number of Columns: 16

```
##Import the neccessay files
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

import warnings
warnings.filterwarnings("ignore")
```

```
## Data Understanding
```

```
data=pd.read_csv("AB_NYC_2019.csv")
data.head()
```

	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	

```
## Data Description
```

```
data.shape
```

```
(48895, 16)
```

```
## dataset have 48895 rows and 16 columns
```

```
data.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
```

```
## Here we coe to know that columns['name','host_name','last_review','review_per_month'] having null values
```

```
## Lets drop some columns which are having null values and also not neccessary for dataset for anylasis
```

```
data.drop(["id","name","last_review"],axis=1,inplace=True)
```

```
data.head()
```

	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviev
0	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	19-10-2018	
1	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	21-05-2019	
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	05-07-2019	
4	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	19-11-2018	

```
## Let's replace the misiing values from the column reviews_per_month as '0'.As we need this column for anylasis
```

```
data.fillna({'reviews_per_month':0},inplace=True)
```

```
## Let's check whether he columns is having still any missing values
```

```
data.reviews_per_month.isnull().sum()
```

```
0
```

```
data.head()
```

	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	review
0	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	19-10-2018	
1	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	21-05-2019	
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	05-07-2019	
4	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	19-11-2018	

```
## Lets check for unique values in dataset
```

```
data.room_type.unique()
```

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

```
len(data.room_type.unique())
```

```
3
```

```
data.neighbourhood_group.unique()
```

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

```
len(data.neighbourhood_group.unique())
```

```
5
```

```
len(data.neighbourhood.unique())
```

```
221
```

```
## Checked the Duplicate rows in our dataset and no duplicate data was found
```

Step 2: Data Wrangling:

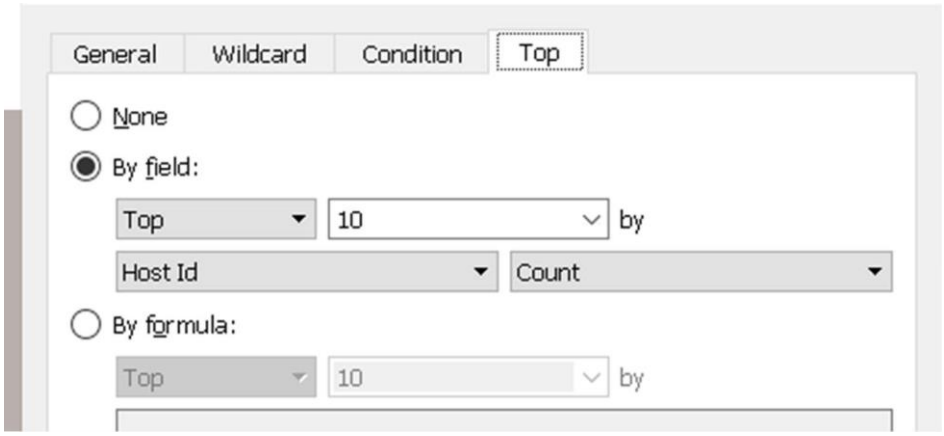
- **Duplicate Check:** Verified the dataset for duplicate rows, and none were found.
- **Null Value Check:** Identified null values in columns such as *name*, *host-name*, *last review*, and *review-per-month*.
- **Handling Missing Values:** Dropped the *name* column as its missing values were minimal and would not significantly affect the analysis.
- **Data Formatting:** Ensured consistent formatting across the dataset.
- **Outlier Identification:** Reviewed and addressed any outliers in the data to ensure accurate analysis.

Data Analysis and Visualizations using Tableau:

We used **Tableau and Excel** to perform data analysis and create visualizations for this assignment. Below are the detailed steps taken for each visualization.

PPT 1

1. Overview of NYC Airbnb Listing – created table by adding the Room type and Neighborhood in Rows and Columns while considered total count for value in table A in tableau and Average of the Prices in Table B for Value in Excel Worksheet
2. Room Type listing Added Room type in Rows and Neighbouhood in Columns with the Count of the listing as value and selected packed bubble for Visualization
3. Profit and Revenue insights – we used side by side bars with text table to compare neighborhoods as per room type to find the highest revenue generator
4. Pricing Analysis - We used a box and whisker's plot with Neighbourhood Groups in Columns and Pricein Rows. We changed the Price from a Sum Measure to the median measure.
5. Average Price of Neighborhood - We created a bubble chart with Neighborhood Groups in Columns and Price column in Rows.
6. Popularity of Localities and Properties – We created bar chart and sorted it to highlight the top 10 best performing neighbourhood and filter top 10 in given way

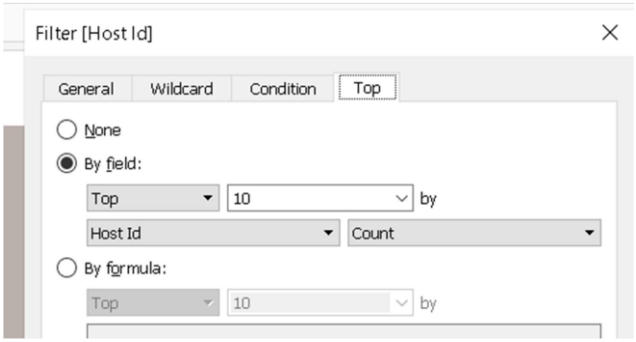


7. Minimum Night Insights – Created Groups for the Night by Right click to click option > Create > group then added the required numbers in a group and applied the same.

8. Neighborhood vs Availability - We created a dual-axis chart using bar chart for Availability 365 and a line chart for price for the top 10 neighborhood groups sorted by price.

Methodology Document PPT 2:

1. Host Acquisition Strategy - We identified the top 10 Host Ids, Host Name with count of Host Ids using the tree map



2. Property Insights by Neighbourhood – Created a Bubble chart and text chart using neighbourhood and the count
3. Room Type & Pricing Strategy – In this, we used stacked bars to show the Average price for each room type which is given on the bar as per each neighbourhood also showed the Room type Majority using the highlighting tables to highlight the highest neighbourhood in room types
4. Popularity of Properties - We took neighbourhood in rows and sum of reviews in column and took neighbourhood groups in colour. We used filter to show Top 20 neighbours as per the sum of reviews.
5. Optimizing Less Popular Properties – Used a side-by-side bars to analyse the Neighbourhoods and Room type as per the sum of reviews received per month

Tools used:

- Data cleaning and preparation: Jupyter notebook – Python
- Visualization and analysis: Tableau and Excel
- Data Storytelling: Microsoft PPT