STORYTELLING CASE STUDY: AIRBNB, NYC

By- SHRIKANT VISHWAKARMA

PRESENTATION – 1

APPENDIX

Data Sources: Presentation.1

The columns in the dataset are self-explanatory. You can refer to the diagram given below to get a better idea of what each column signifies.

Column	Description
id	listing ID
name	name of the listing
host_id	host ID
host_name	name of the host
neighbourhood_group	location
neighbourhood	area
latitude	latitude coordinates
longitude	longitude coordinates
room_type	listing space type
price	
minimum_nights	amount of nights minimum
number_of_reviews	number of reviews
last_review	latest review
reviews_per_month	number of reviews per month
calculated_host_listings_count	amount of listing per host
availability_365	number of days when listing is available for booking

Methodology Document: Presentation.1

In our case study, we utilized Jupyter Notebook for the initial data analysis and employed Tableau for in-depth data exploration and visualization.

Initial Analysis using Jupiter Notebook:

- Data Set Used: AB_NYC_2019.csv
- Number of Rows: 48895
- Number of Columns: 16

In [1]:
```
# Importing necessary libraries.

import numpy as np
import pandas as pd

import warnings
warnings.filterwarnings('ignore')

import matplotlib.pyplot as plt
%matplotlib inline

import seaborn as sns
```

In [2]:
```
# Importing dataset.

df = pd.read_csv('AB_NYC_2019.csv')
df.head()
```

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]:
```
# Checking the shape of the dataset.

df.shape
```

Out[3]:
```
(48895, 16)
```

In [4]:
```
# Checking the null and missing values in a dataset.

round(df.isnull().sum()/df.shape[0]*100,2)
```

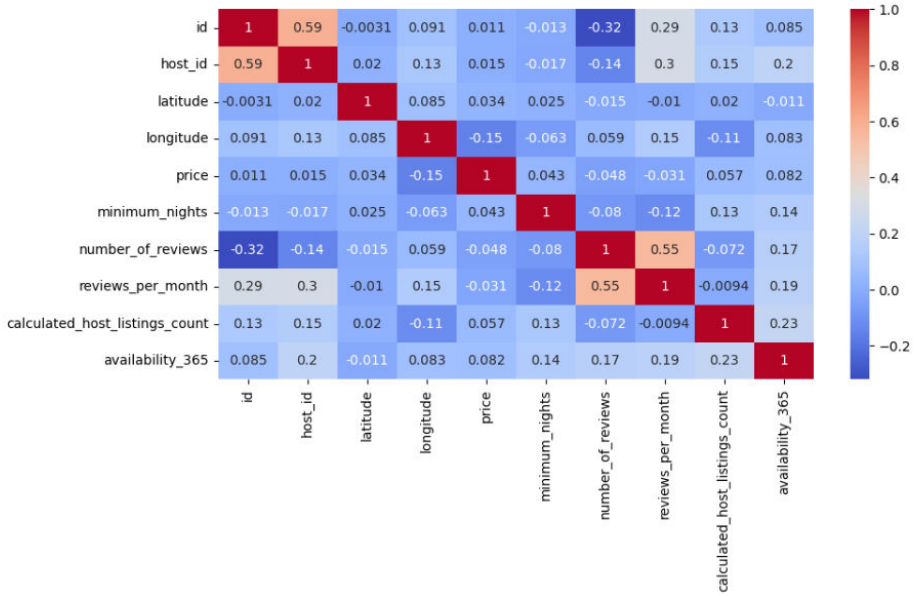
Out[4]:
```
id                                0.00
name                              0.03
host_id                           0.00
host_name                         0.04
neighbourhood_group               0.00
neighbourhood                     0.00
latitude                          0.00
longitude                         0.00
room_type                         0.00
price                             0.00
minimum_nights                    0.00
number_of_reviews                 0.00
last_review                      20.56
reviews_per_month                20.56
calculated_host_listings_count    0.00
availability_365                  0.00
dtype: float64
```

In [5]:
```
# Creating a new dataframe with only numerical variables.

df_num = df.select_dtypes(include=(int,float))
```

In [6]:
```
# Now checking the correlation among numerical variables by plotting heatmap.

plt.figure(figsize=[10,5])
sns.heatmap(df_num.corr(), annot=True, cmap= 'coolwarm')
plt.show()
```



In [7]:
```
# Checking price column

df['price'].describe()
```

Out[7]:
```
count    48895.000000
mean       152.720687
std        240.154170
min          0.000000
25%         69.000000
50%        106.000000
75%        175.000000
max      10000.000000
Name: price, dtype: float64
```

In [8]:
```
# Checking distribution of values in Price column.

plt.figure(figsize=[15,2])
sns.boxplot(df['price'], orient='h')
plt.show()
```

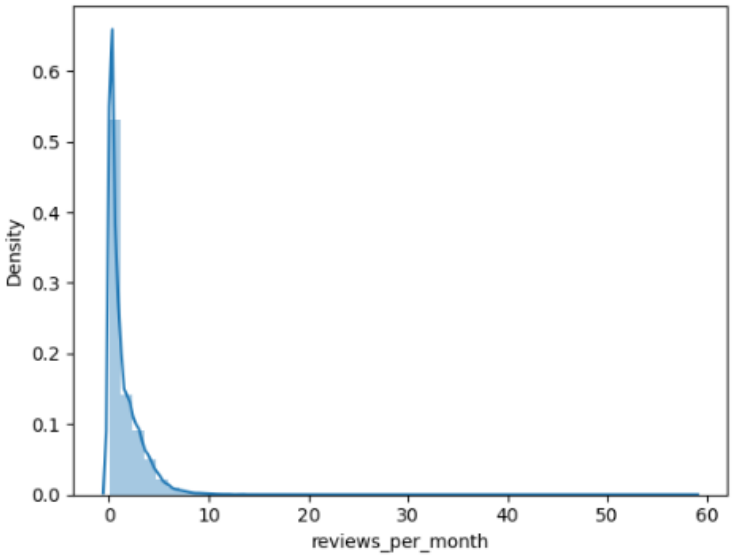
In [9]:
```
# Checking reviews_per_month column

df['reviews_per_month'].describe(percentiles=[0.25,0.5,0.75,0.99])
```

Out[9]:
```
count    38843.000000
mean         1.373221
std          1.680442
min          0.010000
25%          0.190000
50%          0.720000
75%          2.020000
99%          7.195800
max         58.500000
Name: reviews_per_month, dtype: float64
```

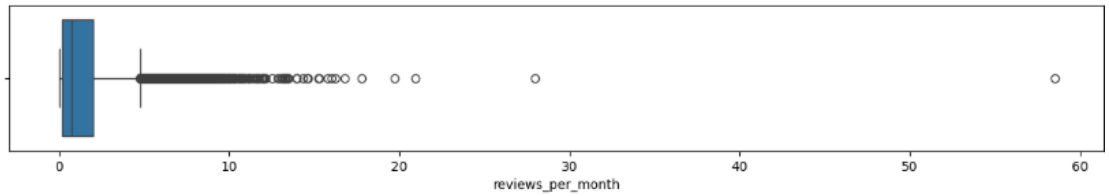
In [10]:
```
# Plotting distribution plot to know the value distribution in reviews_per_month column.

sns.distplot(df['reviews_per_month'])
plt.show()
```



In [11]:
```
# Plotting boxplot to find the outliers in reviews_per_month column.

plt.figure(figsize=[15,2])
sns.boxplot(df['reviews_per_month'], orient='h')
plt.show()
```



In [13]:
```
# Checking host_name column

df['host_name'].value_counts()
```

Out[13]:
```
host_name
Michael                 417
David                   403
Sonder (NYC)            327
John                    294
Alex                    279
                       ...
Rhonycs                   1
Brandy-Courtney           1
Shanthony                 1
Aurore And Jamila         1
Ilgar & Aysel             1
Name: count, Length: 11452, dtype: int64
```

```
In [14]: df['room_type'].value_counts(normalize=True)*100

Out[14]: room_type
         Entire home/apt    51.966459
         Private room       45.661111
         Shared room         2.372431
         Name: proportion, dtype: float64

In [15]: df['neighbourhood_group'].value_counts(normalize=True)*100

Out[15]: neighbourhood_group
         Manhattan        44.301053
         Brooklyn         41.116679
         Queens           11.588097
         Bronx             2.231312
         Staten Island     0.762859
         Name: proportion, dtype: float64

In [16]: df['neighbourhood'].value_counts(normalize=True)

Out[16]: neighbourhood
         Williamsburg          0.080172
         Bedford-Stuyvesant    0.075959
         Harlem                0.054361
         Bushwick              0.050414
         Upper West Side       0.040311
                                  ...
         Fort Wadsworth        0.000020
         Richmondtown          0.000020
         New Dorp              0.000020
         Rossville             0.000020
         Willowbrook           0.000020
         Name: proportion, Length: 221, dtype: float64
```

```
In [18]: # Now to check the unique values of other columns'
         df['room_type'].unique()

Out[18]: array(['Private room', 'Entire home/apt', 'Shared room'], dtype=object)

In [19]: len(df['room_type'].unique())

Out[19]: 3

In [20]: df['neighbourhood_group'].unique()

Out[20]: array(['Brooklyn', 'Manhattan', 'Queens', 'Staten Island', 'Bronx'],
               dtype=object)

In [21]: len(df['neighbourhood_group'].unique())

Out[21]: 5

In [22]: len(df['neighbourhood'].unique())

Out[22]: 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, host-name, last review and review-per-month have null values.
- 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.

SILDE- 1: Cover page

SLIDE- 2: Agenda

SLIDE- 3: Objective

- Provide necessary objectives that what we are going to tell/show them.

SLIDE- 4: Data Life Cycle

SLIDE- 5: Evaluating Hosts Performance

- Created a Pareto chart to evaluating the hosts performance.
- We added “Host Name” to columns and Count of “Host Id” in rows to create a bar chart and then put a line chart of Running total of Count of “Host Id” and then combine them by using Dual Axis option and also put a reference line of Average.

SLIDE- 6: Room Type with respect to Neighbourhood group

Created a Donut chart to understand the “Room Type” and “Neighbourhood group” preference of customers.

- A 2-donut charts were utilized to visually represent the proportional distribution of Room Type and Neighbourhood group, providing a clear and intuitive comparison of categories while maintaining an emphasis on overall composition.

Discover the most popular Room Types in Neighbourhood Groups.

- We created a donut chart to visualize the percentage of preferred room types in relation to the neighbourhood group.
- In the middle of donut chart indicates value distribution among Neighbourhood groups.

SLIDE- 7: Price Analysis Neighbourhood wise

- We used a box and whisker’s plot with Neighbourhood Groups in Columns and Price in Rows.

SLIDE- 8: Customer bookings with respect to Minimum nights

- We created the bins 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 neighbourhood group.

SLIDE- 9: Understanding Price variation with respect to Room Type and Neighborhood

- We created 2 bubble charts with Neighbourhood Groups and Room Type in Columns respectively and Price in Rows.
- We created treemap to understand the Price variation with respect to Room Type and Neighbourhood group.

SLIDE- 10: Customer Reviews with respect to Room Type and Neighbourhood groups

Created a Donut chart to understand the “Room Type” and “Neighbourhood group” preference of customers.

- 2 donut charts were utilized to visually represent the Number of Reviews distribution among Neighbourhood group and Room Type, providing a clear and intuitive comparison of categories while maintaining an emphasis on overall composition.

Discover the Number of Reviews with respect to Room Types in Neighbourhood Groups.

- We created a donut chart to visualize the Number of Reviews percentage with respect to room types in relation to the neighbourhood group.
- In the middle of donut chart indicates Number of Reviews distribution among Neighbourhood groups.

Data Assumption: Presentation.1

```
Categorical Variables:
    - room_type
    - neighbourhood_group
    - neighbourhood

Continous Variables(Numerical):
    - Price
    - minimum_nights
    - number_of_reviews
    - reviews_per_month
    - calculated_host_listings_count
    - availability_365
- Continous Variables could be binned in to groups too

Location Varibles:
    - latitude
    - longitude

Time Varibale:
    - last_review
```

PRESENTATION - 2

APPENDIX

Data Sources: Presentation.2

The columns in the dataset are self-explanatory. You can refer to the diagram given below to get a better idea of what each column signifies.

Column	Description
id	listing ID
name	name of the listing
host_id	host ID
host_name	name of the host
neighbourhood_group	location
neighbourhood	area
latitude	latitude coordinates
longitude	longitude coordinates
room_type	listing space type
price	
minimum_nights	amount of nights minimum
number_of_reviews	number of reviews
last_review	latest review
reviews_per_month	number of reviews per month
calculated_host_listings_count	amount of listing per host
availability_365	number of days when listing is available for booking

Methodology Document: Presentation- 2

SILDE- 1: Cover page

SLIDE- 2: Agenda

SLIDE- 3: Objectives & Background

- Provide necessary objectives that what we are going to tell/show them.
- Provide background that why we are going to analyze and present the respective insights.

SLIDE – 4: Strategic Host Acquisition: Maximizing Impact in Key Markets

First visual- Created a Pareto chart to understand the hosts performance.

- We added “Host Name” to columns and Count of “Host Id” in rows to create a bar chart and then put a line chart of Running total of Count of “Host Id” and then combine them by using Dual Axis option and also put a reference line of Average.

Second visual- Created a Bubble chart to understand the host performance with respect to neighbourhood group.

- We added “Host name” in columns, then “neighbourhood group” in rows and Count of “Host Id” in values and then select a bubble chart.

SLIDE – 5: Market Trends: Where Demand Meets Expansion Potential

First visual- Created a Donut chart to understand the “Room Type” preference of customers.

- A donut chart was utilized to visually represent the proportional distribution of Room Type, providing a clear and intuitive comparison of categories while maintaining an emphasis on overall composition.

Second visual- Created a Donut chart to understand the “Neighbourhood group” preference of customers.

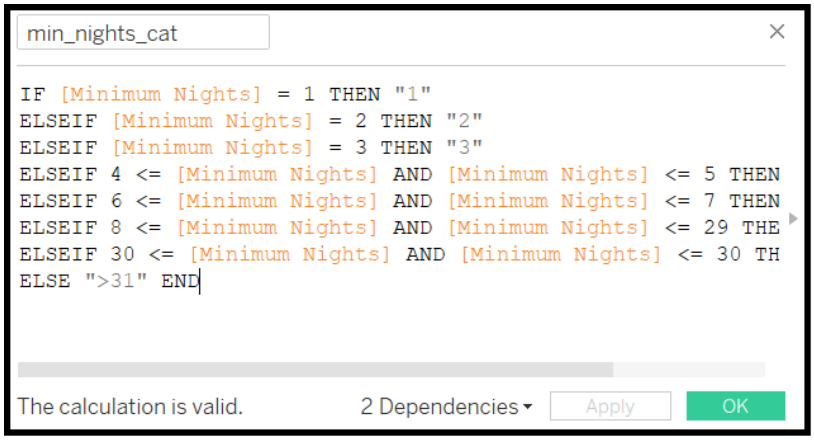
- A donut chart was utilized to visually represent the proportional distribution of “Neighbourhood group”, providing a clear and intuitive comparison of categories while maintaining an emphasis on overall composition.

SLIDE- 6: Discover the most popular Room Types in Neighbourhood Groups.

- We created a donut chart to visualize the percentage of preferred room types in relation to the neighbourhood group.
- In the middle of donut chart indicates value distribution among Neighbourhood groups.

SLIDE- 7: Strategic Analysis of Booking Trends Based on Minimum Night Requirements.

- We created the bins 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 neighbourhood group.

SLIDE- 8: Strategic Insights on Neighborhood Availability and Pricing Dynamics

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

SLIDE- 9: Key Price Range for Maximizing Bookings

- 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.

SLIDE- 10: Impact of Room Types and Neighborhoods on Pricing Strategies

- We created Highlights Table chart by taking Room Type in rows & Neighbourhood Group in column.
- We took the average price in colour Marks card to highlight the different Room Type in different colours and increase its size for better visuals.

SLIDE- 11: Short Stays, High Impact: Majority of Reviews Focus on 1-3 Nights

- We created a horizontal bar chart by taking minimum_nights_bins in rows and sum of Number of reviews in columns.
- We added Neighbourhood group to the colours Marks card to highlight the different Neighbourhood group in different colours.

SLIDE- 12: Conclusion & Recommendation

- Based on the insights provided, we summarize key findings/conclusion and offer practical recommendations.

TOOLS USED:

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

Data Assumption: Presentation.2

```
Categorical Variables:
    - room_type
    - neighbourhood_group
    - neighbourhood

Continous Variables(Numerical):
    - Price
    - minimum_nights
    - number_of_reviews
    - reviews_per_month
    - calculated_host_listings_count
    - availability_365
- Continous Variables could be binned in to groups too

Location Varibles:
    - latitude
    - longitude

Time Varibale:
    - last_review
```