Methodology Document

AIRBNB CASE STUDY IIIT-B

- Pallav Rajput

1. Storyboarding

- Went over the data to become aware with it, made notes in key fields, and performed preliminary analysis using Jupiter Notebook and Tableau for data analysis and visualization.

```
import warnings
#warnings.filterwarning("ignore")
import numpy as np, pandas as pd, matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

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

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	19-10-2018	0.21
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	21-05-2019	0.38
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	05-07-2019	4.64

```
air.shape
```

```
(48895, 16)
```

```
[10]: air.isnull().sum()
```

```
[10]: 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
```

Collapse Output

```
[17]: air.fillna({'reviews_per_month':0}, inplace=True)
```

```
[18]: air.isnull().sum()
```

```
[18]: 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                 0
      calculated_host_listings_count    0
      availability_365                  0
      dtype: int64
```

```
[25]: air.nunique()
```

```
[25]: id                                48895
      name                              47896
      host_id                           37457
      host_name                         11452
      neighbourhood_group                   5
      neighbourhood                       221
      latitude                          19048
      longitude                         14718
      room_type                             3
      price                               674
      minimum_nights                      109
      number_of_reviews                   394
      last_review                        1764
      reviews_per_month                   938
      calculated_host_listings_count       47
      availability_365                    366
      dtype: int64
```

```
[26]: cols = ['id','name','host_id']
      # has high no. of unique values and irrelevant to our study.
```

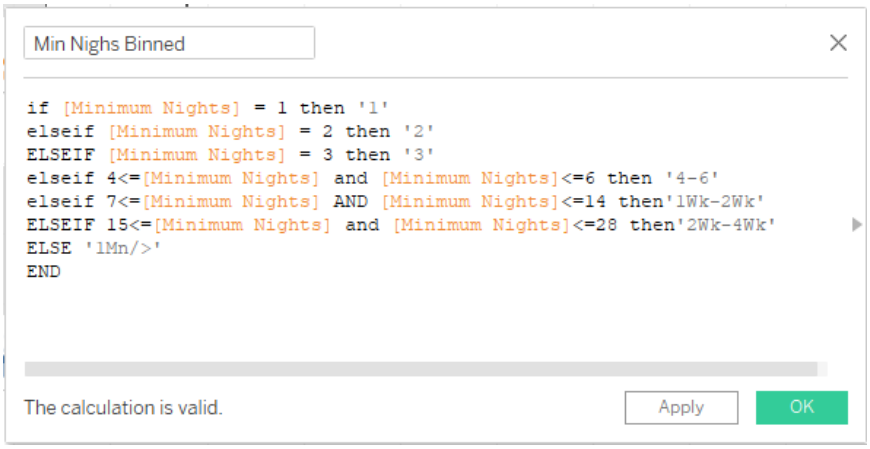
```
[27]: air.drop(cols, axis=1, inplace=True)
```

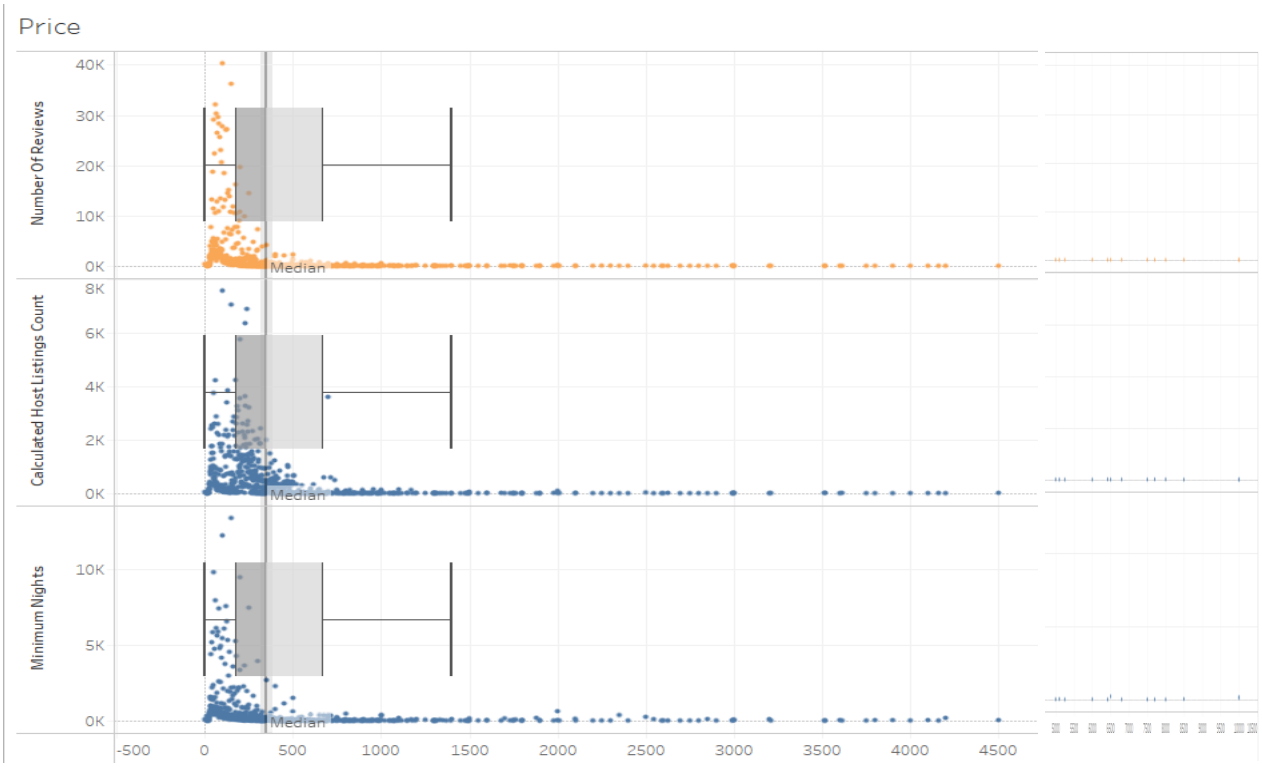
```
[24]: air.minimum_nights.describe()
```

```
[24]: count    48895.000000
      mean         7.029962
      std         20.510550
      min          1.000000
      25%          1.000000
      50%          3.000000
      75%          5.000000
      max       1250.000000
      Name: minimum_nights, dtype: float64
```

2. Data Wrangling / Binning

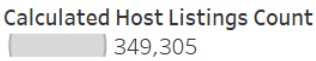
- The univariate analysis using Tableau was conducted on fields to examine their distributions, unique values, missing values, and outliers.
- Assuming null values belonged to the category, a grouped field for Minimum Number of Days was created.

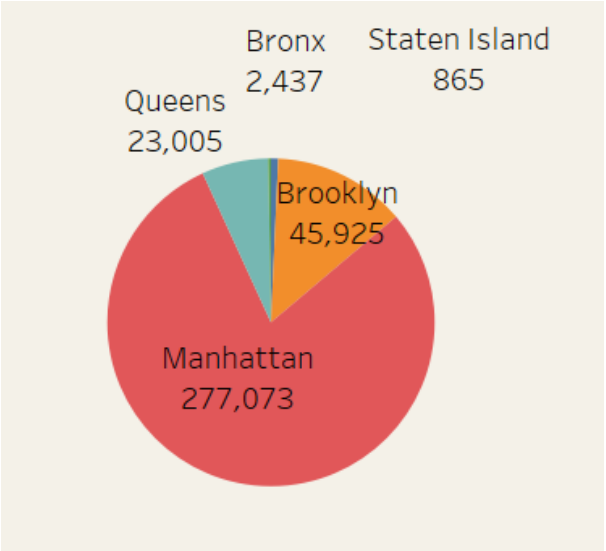


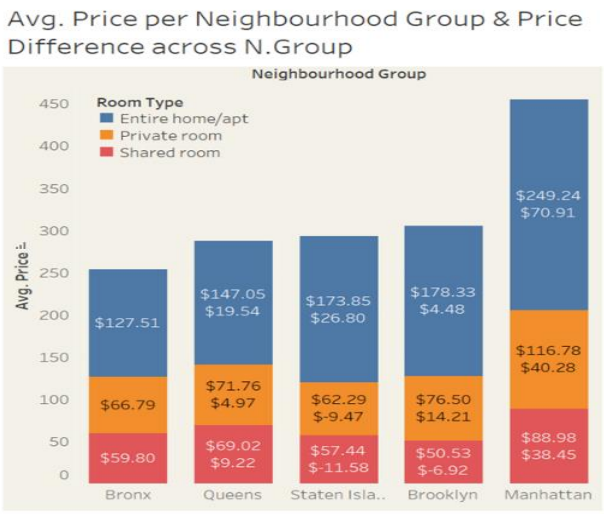


3. Analysis

- Check the overall distribution of listing in the neighborhood groups. Found Manhattan has the majority of listing.







- Of all the room types, Manhattan is also the most expensive.
- The price along the price difference (left to right) paid more across the neighborhood group is displayed in the slacked bar graph.
- The cheapest shared rooms are in Brooklyn.

4. **Presentations**

PPT1 - <u>presentation on data analysis, its procedures, and technical aspects intended for data analysis managers and lead data analysts</u>

- Highlighted the **objective** of the case study
- Presented in **overview of data lifecycle** from importing data, missing value treatment and creating features. Where we have used both python and tableau.
- Presented analysis **using box plot** on price with respect to minimum nights, host listing and number of reviews.
- Using **a pie chart** showed the distribution of total host listing count.
- Using **tableau calculated field**, **binned** the minimum night values.
- Using **Bar chart** listed 20 most listed host a long with their pricing.
- Using work of with **multiple values on X axis** listed the reviews and host listing count.
- **A tree map** is used to list the top reviewed hosts with respect to the neighborhood group.
- **SWOT** analysis.
- **Appendix data set description and variable categories.**
- **Conclusion**.

PPT2 - for the Head of User Experience and the Head of Acquisitions and Operations. presentation on host purchases and real estate. best properties, pricing differences, and consumer preferences while looking at cities and neighborhoods in general.

- Highlighted the objective of the case study and the background of the situation
- We presented a **graphical map** of New York City with spread of listing
- A **stacked bar graph** is used to list the average price across all neighborhood groups and with different room types.
- The **stacked bar graph is used will dual axis chart** that is we presenting the reviews per month across all minimum night's categories
- **Text table** is used to list the price average across all neighborhood groups and minimum nights.
- **Side by side chart** is used to representing availability.
- **Bubble chart** is used to show the highest number of hosts according to listing count.
- Recommendations.

Thank You

Pallav Rajput

+91-9888143002

ribhu.s18@gmail.com