Data Methodology

Step 1: Storyboarding

- Went through the data to get familiarized with it and noted down important fields.
- Made a mind map of the various slides of the presentation.
- Made a rough template based on this mind map.

Step 2: Data Wrangling using Python

Data Cleaning on AB_NYC_2019 :

- Finding the percentage of null values in the columns.

```
100*df.isnull().mean()
```

```
id                                 0.000000
name                               0.032723
host_id                            0.000000
host_name                          0.042949
neighbourhood_group                0.000000
neighbourhood                      0.000000
latitude                           0.000000
longitude                          0.000000
room_type                          0.000000
price                              0.000000
minimum_nights                     0.000000
number_of_reviews                  0.000000
last_review                       20.558339
reviews_per_month                 20.558339
calculated_host_listings_count     0.000000
availability_365                   0.000000
dtype: float64
```

- The percentage of null values is more in last_review and review_per_month column. As last_review would not affect our analysis, we imputed the missing values in reviews_per_month column with zero.

- Name and Host name have very few missing values, so we dropped them.

- After treating the missing values

```
: df1.isnull().mean()
```

```
: id                                0.0
  name                              0.0
  host_id                           0.0
  host_name                         0.0
  neighbourhood_group               0.0
  neighbourhood                     0.0
  latitude                          0.0
  longitude                         0.0
  room_type                         0.0
  price                             0.0
  minimum_nights                    0.0
  number_of_reviews                 0.0
  last_review                       0.0
  reviews_per_month                 0.0
  calculated_host_listings_count    0.0
  availability_365                  0.0
  dtype: float64
```

- Binning of continuous variables for segmented analysis like price, minimum nights, no_of_reviews, Reviews per month column, calculated host listings count and availability 365 columns.

```
# creating BINS for continuous columns
```

```
#Bins for Price column

bins=[-1,100,1000,5000,10001]
price_range=["low","medium","high","very-high"]
```

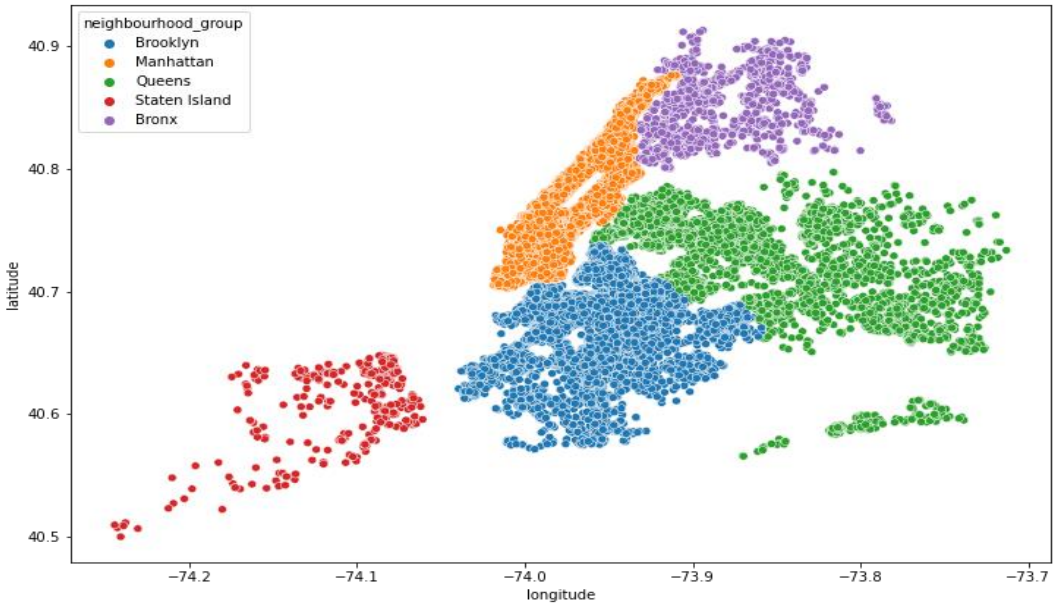
```
df["price_category"]=pd.cut(df["price"],bins,labels=price_range)
```

- We have used the above mentioned method to create bins for all the continuous columns.
- Did univariate & Bivariate analysis. Found some interesting insights that are provided in the PPT.

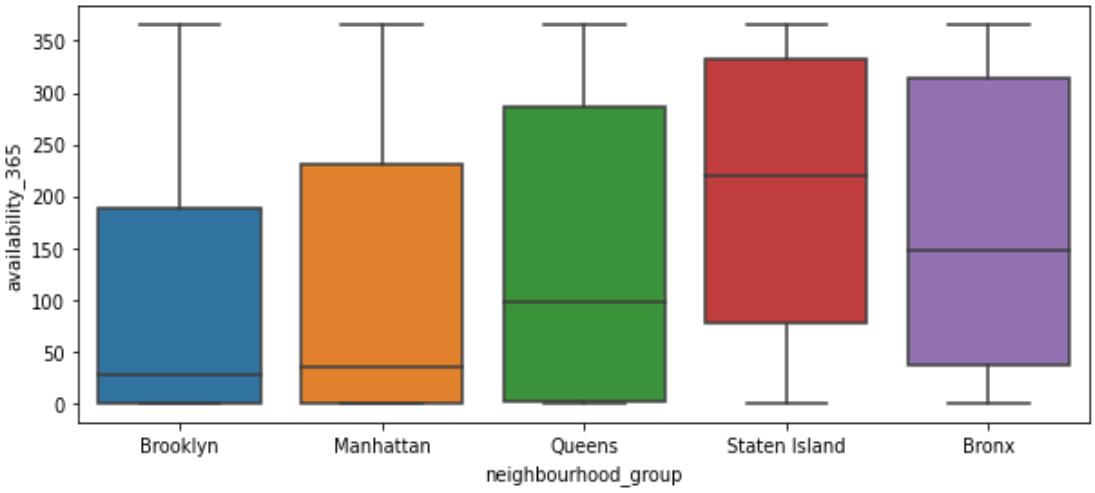
Step 3: Analysis using Python & Tableau

Below are few of the analysis plots from the presentation :

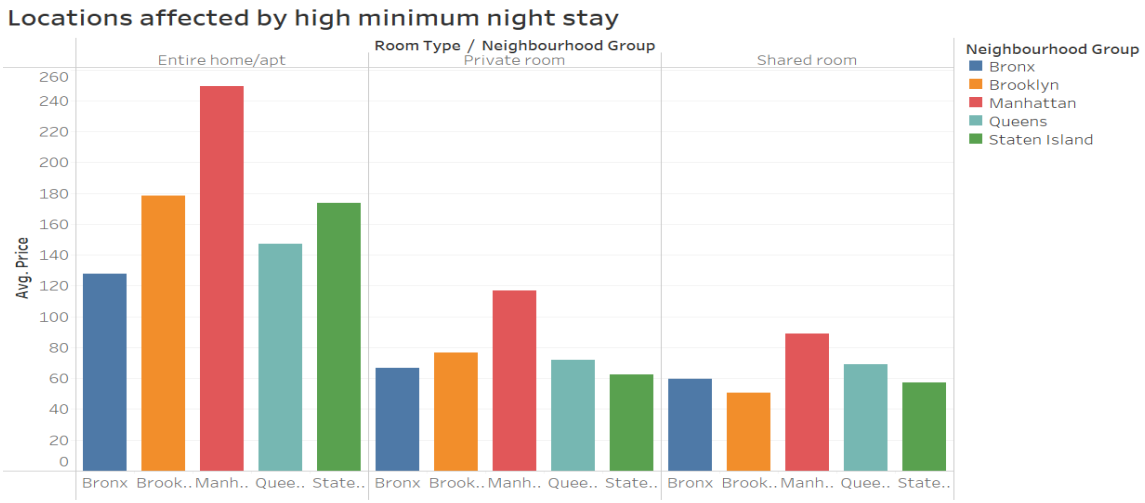
- Checked **total spread** of Airbnb listings in New York.



- Checked the **availability of room** with respect to Neighbourhood group

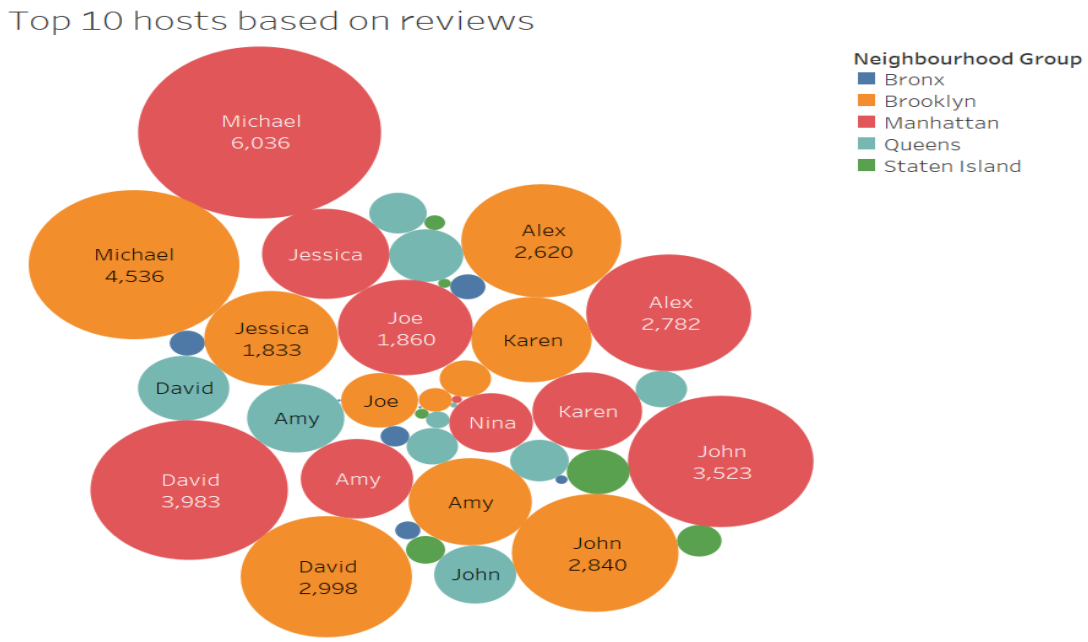


- **Average Price** for each Neighbourhood Group broken down by Room Type.



Average of Price for each Neighbourhood Group broken down by Room Type. Color shows details about Neighbourhood Group.

- **Top 10 hosts** based on reviews



Host Name and sum of Number Of Reviews. Color shows details about Neighbourhood Group. Size shows sum of Number Of Reviews. The marks are labeled by Host Name and sum of Number Of Reviews. The view is filtered on Host Name, which keeps 10 of 11,452 members.

Step 4: Presentation

- Made the presentations adhering to best practices and pyramid principle.
- Added recommendations for the respective departments.

Tools used

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