Methodology Document

Step 1) Data understanding and Cleaning

i) With the help of Python we Completed null value analysis and its imputation.
ii) Dropped The Name and Host Name Columns at Its not Required for the Analysis.
iii) Check the outliers and duplicated Values in the dataset
iv) Extract Year value from the Last review column
v) Impute null values

Null values removal and imputation

```
air.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
```

```
# check for the null values
air.isnull().sum()
```

```
id                                  0
host_id                             0
neighbourhood_group                 0
neighbourhood                       0
latitude                            0
longitude                           0
room_type                           0
price                               0
minimum_nights                      0
number_of_reviews                   0
reviews_per_month                   0
calculated_host_listings_count      0
availability_365                    0
Year                            10052
dtype: int64
```

```
# convert the last_review column into date time
air['last_review'] = pd.to_datetime(air['last_review'])
```

```
# drop the last_review column
air = air.drop(['last_review'], axis = 1)
```

```
# Extracting the year out of last_review
air['last_review'] = pd.to_datetime(air['last_review'])
air['Year'] = air['last_review'].dt.year
```

```
# also drop the name and host_name features which are not much more important for furthe analysi
air = air.drop(['name', 'host_name'], axis = 1)
```

```
# We can replace the nulls in review_per_month with a 0 for better understanding
air.fillna({'reviews_per_month':0},inplace= True)
air['reviews_per_month'].isna().sum()
```

```
0
```

```
air.isnull().sum()
```

```
id                              0
host_id                         0
neighbourhood_group             0
neighbourhood                   0
latitude                        0
longitude                       0
room_type                       0
price                           0
minimum_nights                  0
number_of_reviews               0
reviews_per_month               0
calculated_host_listings_count  0
availability_365                0
Year                            0
dtype: int64
```

Step 2). After Step 1, exported the CSV file to the local server and load it on Power Bi for analysis and take out the insights. The following actions were perform

i) Make a new column name as price range
ii) Column 'last review' date format corrected with DD/MM/YYYY

Step 3) Data Analysis using Power Bi

Insights after analyzing the dataset

i) Most popular locations are Brooklyn and Manhattan and least popular are Queens, Bronx and Staten Island.
ii) Price ranges are high in Manhattan, Brooklyn and Staten Island as compared to Queens and Bronx.
iii) Popularity and volume of Entire home /apt and private rooms is high in Manhattan and Brooklyn. So the insight is need to focus on increase in other locations also like Bronx, Queens and Staten Island.
iv) Popularity of sharing rooms are very less but they mostly available in 365 days. The insight is to make available popular room type like entire home/apt and private rooms in all areas.
v) Price range popularity is 0-300 so need to think about the offers where popularity is very less or less popular areas
vi) Queens distribution area is more as compared to other groups but having less popularity. The insight is to utilize the available area and implement the availability of rooms and types of rooms should be increased
vii) In the Bronx, Queens and Staten Island volume of Entire home/apt and private rooms are very less need to increase the volume of those properties.
viii) Neighborhood group vice availability in 365 days is most of the time Staten Island is available in 365 days and least time Brooklyn is available
ix) Most of the customers are like to take a package of 2 nights and 1 night both are almost 38 % each.

Step 4) Presentation

While Preparing the presentation we try to incorporate the best practices and pyramid principle.

Presentation 1 – Analytics:

1. Use technical terms in the presentation
2. Presented the more details of methodology and techniques

Presentation 2 – Head:

1. Use more business terms in the presentation
2. Presented more and more business languages
3. Include the recommendations to improve business