METHODOLOGY DOCUMENT

(Airbnb Case Study)

Students: Le Thi Hong Ha, Pham Sy Hai

1. Problem Identification

1.1.Problem statement

For the past few months, Airbnb has seen a major decline in revenue due to the lockdown imposed during the pandemic.

Now that the restrictions have started lifting and people have started to travel more. Hence, Airbnb wants to make sure that it is fully prepared for this change.

1.2.Context (domain understanding)

Airbnb is an American company based in San Francisco, California. It operates an online marketplace for lodging, primarily homestays for vacation rentals, and tourism activities.

After all, being an online marketplace for hosting personal homestays and private apartments in the majority, the company had two types of customers. One who hosts their place and the another who books the place for a particular time is the end consumer utilizing the hosted place.

Airbnb earns commission from both ends and hence must make sure both of its customers are able to generate value from their business. They also must make the hosted place offered on their platform provide the best services at reasonable prices and lookout for the best technology to ease out the booking process for the end consumer without hassle.

1.3.Target Audience

- Presentation –I: (Technical background audience)

  Data Analysis Managers: These people manage the data analysts directly for processes and their technical expertise is basic.

  Lead Data Analyst: The lead data analyst looks after the entire team of data and business analysts and is technically sound.

- Presentation – II (Business background audience)

  Head of Acquisitions and Operations, NYC: This head looks after all the property and hosts acquisitions and operations. Acquisition of the best properties, price negotiation, and negotiating the services the properties offer falls under the purview of this role.

  Head of User Experience, NYC: The head of the user experience looks after the customer preferences and handles the properties listed on the website and the Airbnb app. Basically, the head of the user experience tries to optimize the order of property listing in certain neighborhoods and cities in order to get every property the optimal amount of traction.

1.4.Data analysis and visualization toolkit

- Python
- PowerBI
- Excel

1.5.Data sources

- Upgrad platform> AB_NYC_2019
- Airbnb website: http://airbnb.com/about

2. DataWrangling

2.1. Data collection

For this case study, we will use dataset of Airbnb listings in New York City in 2019. This data includes information about the hosts, location, price, review, minimum nights,and other attributes

We'll be using Python with the following libraries:

- Numpy- for linear algebra
- Pandas- for manipulating and preprocessing the data
- Seaborn- making pretty plots (uses matplotlib)
- Plotly- creating the map of New York City
- Matplotlib - more pretty plots (also necessary for seaborn)
- Nltk – natural language toolkit

2.2. Data understanding

- .head() to display all the columns and the first 5 rows (default) and the data inside. If you want to display more rolls, for example 20 rows, you can write " .head(20)"
- .info() to display the number of data columns, column ‘names, data types, memories usage.
- .shape to display the number of rows and columns.
- .describe() to display statistical information of data, such as min, max, mean, quartiles.

2.2.1. Column description

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

2.2.2. *Variable types: categorical, numeric, location, and time*

```
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
```

- There are 48895 rows and 16 columns in the data frame

```
airbnb.shape
```

```
(48895, 16)
```

```
airbnb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 16 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              48895 non-null  int64  
 1   name                            48879 non-null  object 
 2   host_id                         48895 non-null  int64  
 3   host_name                       48874 non-null  object 
 4   neighbourhood_group             48895 non-null  object 
 5   neighbourhood                   48895 non-null  object 
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object 
 9   price                           48895 non-null  int64  
 10  minimum_nights                  48895 non-null  int64  
 11  number_of_reviews               48895 non-null  int64  
 12  last_review                     38843 non-null  object 
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64  
 15  availability_365                48895 non-null  int64  
dtypes: float64(3), int64(7), object(6)
memory usage: 6.0+ MB
```

2.3.1. Data cleaning

- Duplicate data: there is no duplicate data by using function:
  duplicates = airbnb[airbnb.duplicated()]

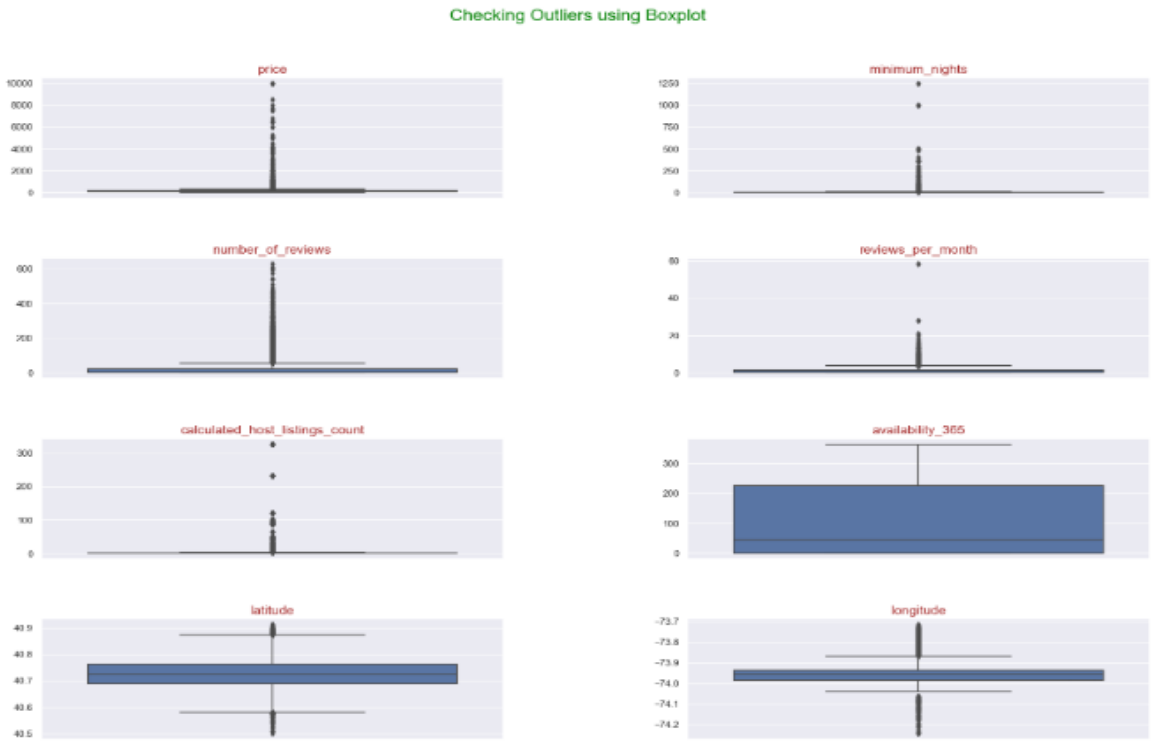
- Missing values/irrelevant data:

```
# Percentage of missing values
round((inp0.isnull().sum()/len(inp0))*100,2)
```

```
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
availability_365_categories       0.00
minimum_night_categories          0.00
number_of_reviews_categories      0.00
price_categories                  0.00
dtype: float64
```

last_review and *reviews_per_month* has 20.56% missing value.
We are just analyzing dataset and not making a model, so no need to drop columns or records

- Outliers
  - Checking the outliers for numeric variables, including:



- Handle the outliers with Categorizing

```
def minimum_nights_function(row):

    if row <= 1:
        return 'very Low'
    elif row <= 3:
        return 'Low'
    elif row < 5:
        return 'Medium'
    elif (row <= 7):
        return 'High'
    else:
        return 'very High'

inp0['minimum_nights_bins'] = inp0.minimum_nights.map(number_of_reviews_categories_function)
```

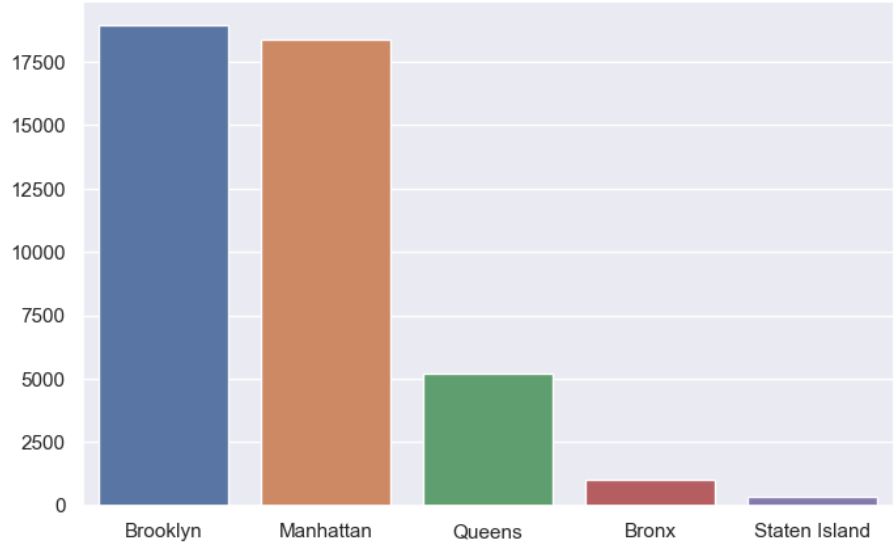
```
def number_of_reviews_categories_function(row):

    if row <= 1:
        return 'very Low'
    elif row <= 10:
        return 'Low'
    elif row <= 20:
        return 'Medium'
    elif (row <= 30):
        return 'High'
    else:
        return 'very High'

inp0['number_of_reviews_bins'] = inp0.minimum_nights.map(number_of_reviews_categories_function)
```

3. Exploratory Data Analysis (analyze and visualize by Python &PowerBI)

3.1. Univariate Analysis

3.1.1. Neighbourhood groups

```
# Neighbourhood Groups
plt.figure(figsize=(8,5))
sns.barplot(x = airbnb.neighbourhood_group.value_counts().index , y = airbnb.neighbourhood_group.value_counts().values)
plt.show()
```

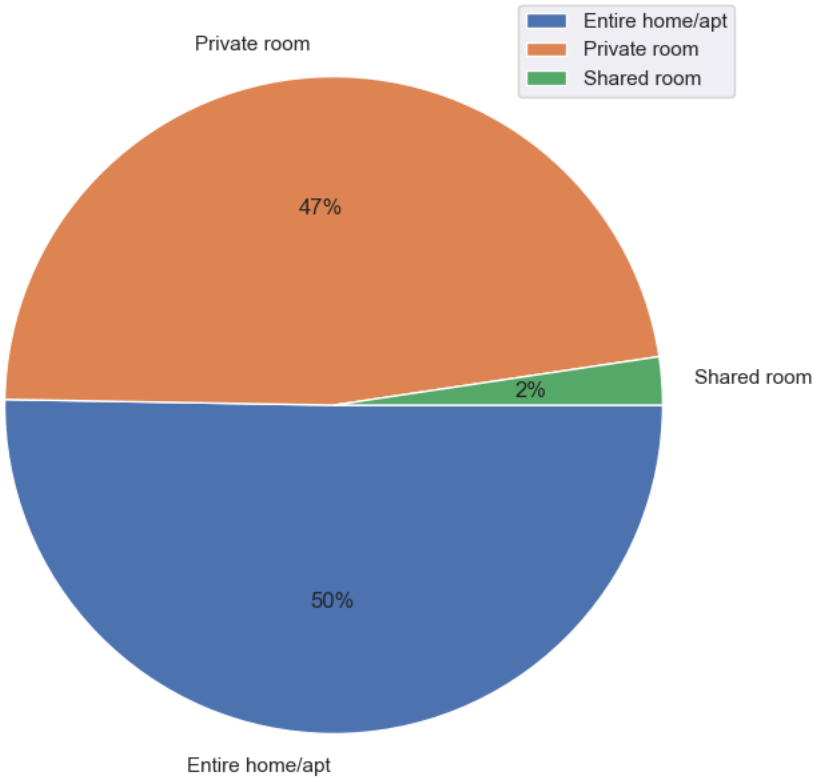


Insights

- Manhattan and Brooklyn have the highest number of listings, at 21661 and 20104.
  -> This can be attributed to the fact that both of those neighbourhoods have more of the tourist attractions, so people would typically want to stay close to what they are seeing.

3.1.2. Room type

```
plt.figure(figsize=(8,8))
plt.pie(x = airbnb.room_type.value_counts(normalize= True) * 100,
        labels = airbnb.room_type.value_counts(normalize= True).index,
        counterclock=False, autopct='%1.0f%%')
plt.legend()
plt.show()
```

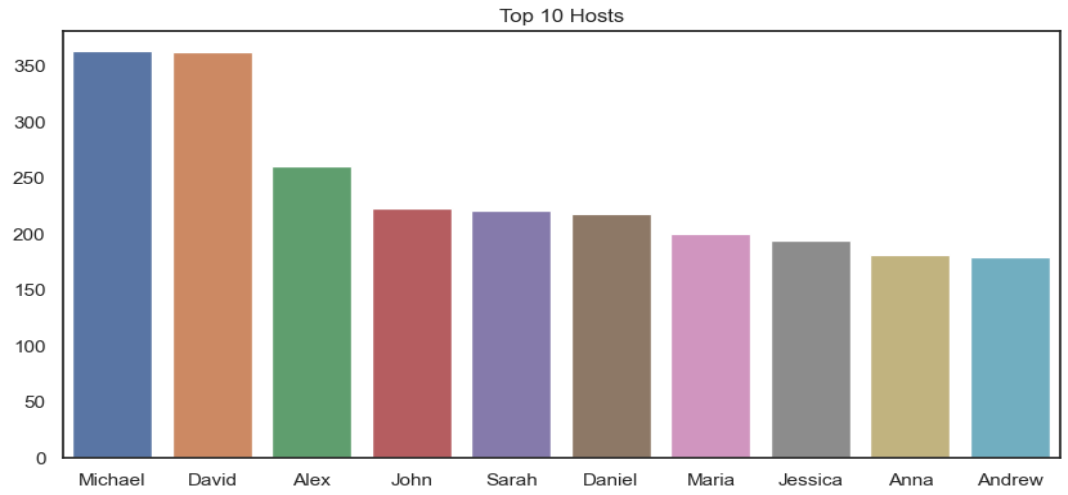


Insights

- Entire home/apt and private rooms are the most common, which may be because the demand for shared rooms is significantly lower.

3.1.3. Top host

```
# Top 10 host's
plt.figure(figsize=(10,5))
sns.barplot(x = airbnb.host_name.value_counts().index[:10],
            y = airbnb.host_name.value_counts().values[:10]).set_title('Top 10 Hosts')
plt.show()
```

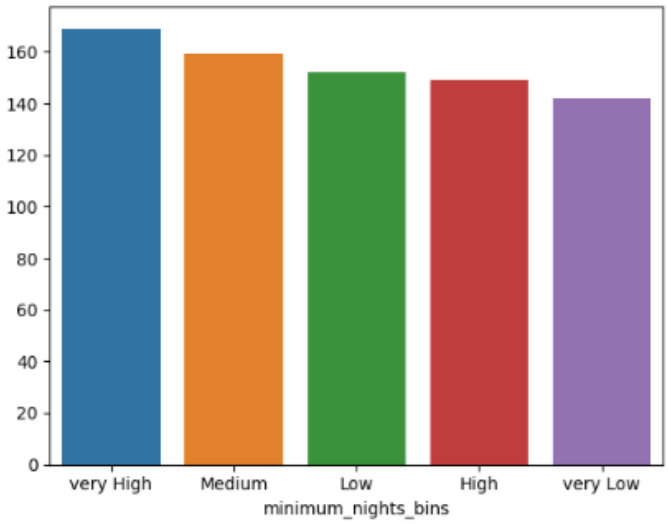


3.1.4. Minimum nights and Number of reviews

```
x1=inp0.groupby('number_of_reviews_bins').price.mean().sort_values(ascending= False)
x1
```

```
number_of_reviews_bins
very High    238.863454
High         178.419158
Low          152.733029
very Low     142.022877
Medium       117.290840
Name: price, dtype: float64
```

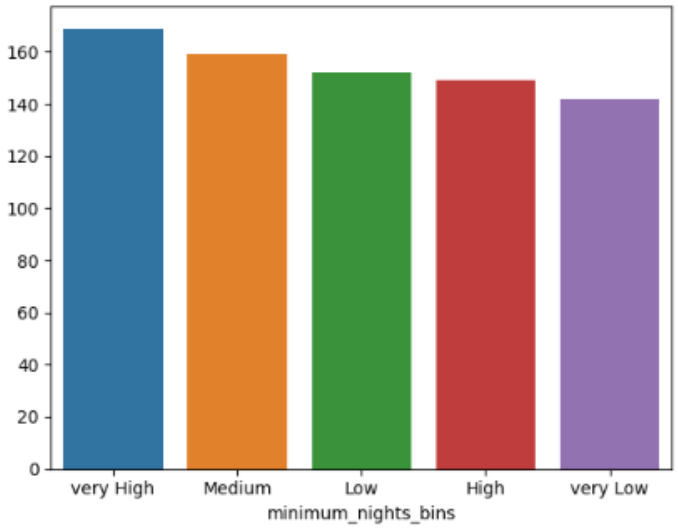
```
sns.barplot(x = x1.index,y = x1.values)
plt.show()
```



```
x1=inp0.groupby('minimum_nights_bins').price.mean().sort_values(ascending= False)
x1
```

```
minimum_nights_bins
very High    169.057003
Medium       159.333596
Low          151.948515
High         149.013879
very Low     142.022877
Name: price, dtype: float64
```

```
sns.barplot(x = x1.index,y = x1.values)
plt.show()
```



Insights

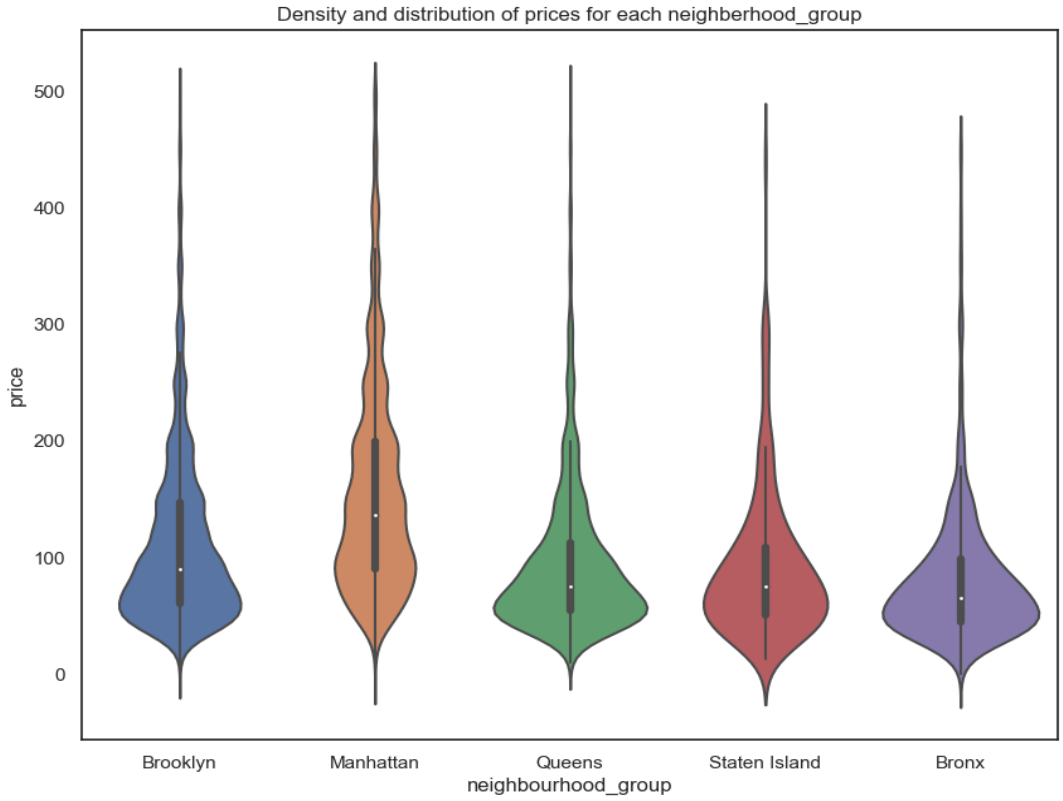
The average price is higher but not significantly where there is a very high minimum nights or number of reviews

3.2. Multivariate Analysis

3.2.1. Distribution of prices for each neighbourhood group

```
sub_6=airbnb[airbnb.price < 500]
#using violinplot to showcase density and distribtuion of prices
viz_2=sns.violinplot(data=sub_6, x='neighbourhood_group', y='price')
viz_2.set_title('Density and distribution of prices for each neighberhood_group')
```

```
Text(0.5, 1.0, 'Density and distribution of prices for each neighberhood_group')
```

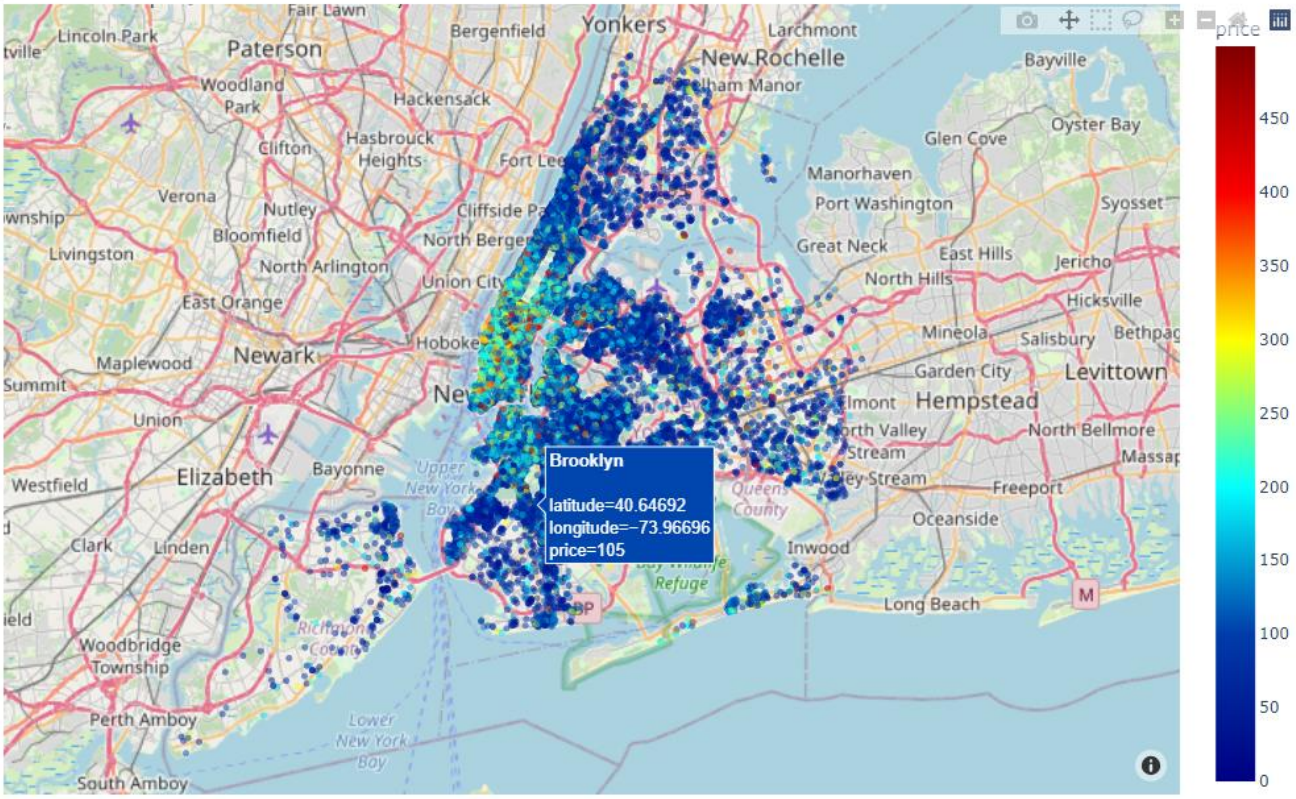


Insights

- Manhattan has the highest range of prices for the listings with $150 price as average observation, followed by Brooklyn with $90 per night.
- Queens and Staten Island appear to have very similar distributions, Bronx is the cheapest of them all. This distribution and density of prices were completely expected;
    - for example, as it is no secret that Manhattan is one of the most expensive places in the world to live in, where Bronx on other hand appears to have lower standards of living.

Location & Price Map

```
fig = px.scatter_mapbox(airbnb500,
                        lat="latitude",
                        lon="longitude",
                        hover_name="neighbourhood_group",
                        hover_data=["price"],
                        color = "price", opacity=0.5,
                        color_continuous_scale="jet",
                        title="Airbnb prices in New York City",
                        zoom=9, height=600)
fig.update_layout(mapbox_style="open-street-map")
fig.update_layout(margin={"r":0,"t":0,"l":0,"b":0})
fig.show()
```

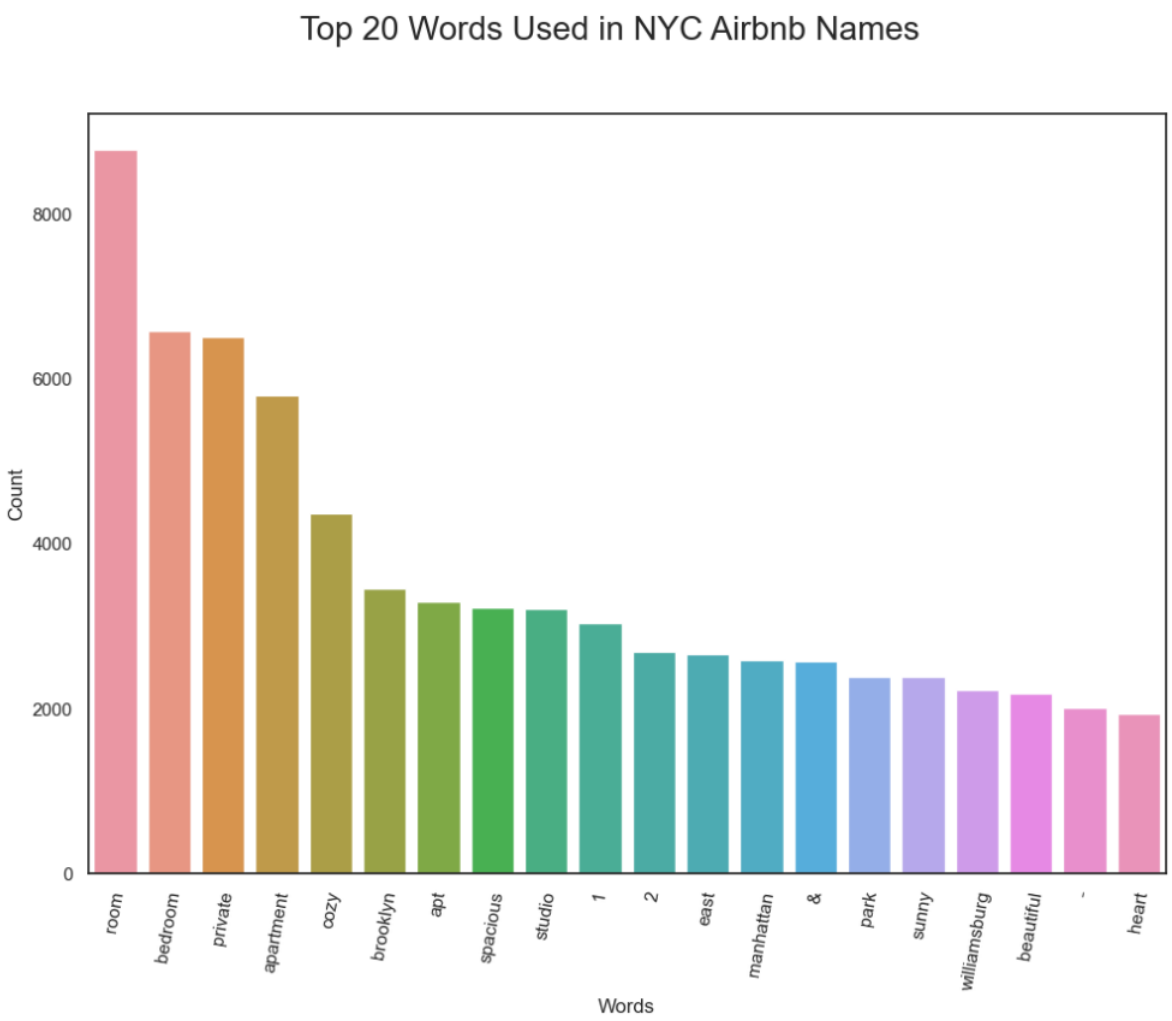


- There is definitely clustering of higher prices in downtown Manhattan.
- There are also noticeable clusters in Upper Brooklyn and Upper Manhattan.
- Location could provide a good signal of price.

3.3. Natural Language Processing

3.3.1. Word Count

```
# Now let's plot!
fig, ax = plt.subplots()
fig.set_size_inches(11.7, 8.27)
fig.suptitle('Top 20 Words Used in NYC Airbnb Names', fontsize=20)
sns.barplot(x='Words', y='Count', data=words_df, ax=ax)
plt.xticks(rotation=80)
plt.show()
```



Insights

Looking at the top 20 words, we can see come clear trends. Hosts are using simple and specific words that allow users to find their property quicker with a quick search.

- **"bedroom", "private", "apartment", and "studio** help homeowners get found, and then they can try to pull travelers in with nice pictures and descriptions.
- **"cozy", "spacious", "sunny", and "beautiful"** are the descriptive words for the place quality that appear often and impress customers.

PowerBI visualization

Average of latitude, Average of longitude, Max of price and Average of minimum_nights by ...