Capstone Project

EDA on Airbnb Booking Analysis

By : G MOHAMED LUQMAN

WHAT WE ARE TALKING ABOUT ?

INTRODUCTION

Airbnb is an online platform that connects hosts renting out space in their homes with guests seeking lodging for generally cheaper prices than a hotel.

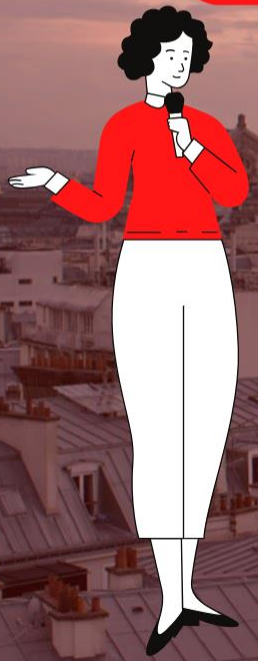
It has millions of listing, which generate alots of data. We are analyzing these data for making business decision, for looking best room type, etc.,

Agenda

1. Relationship anaylsis between variable
2. Finding Top Neighborhood Group
3. Finding total count of each room types
4. Finding Most popular Neighborhood
4. Types of room by neighbourhood regions
6. Average price for each room type
7. Rooms availabilty in different areas
8. Price Analysis
9. Listing Analysis
10. Availability Analysis
11. Profitability Analysis
12. Visual Map Analysis and WordCount Analysis

AI

CONTENT



airbnb

Understanding about the Dataset

1 id	Unique Listing Id
2. Name	Name of the Property
3. Host_id	Unique Id for each listed host
4. Host_name	Name of the host
5. Neighbourhood_Group	Location
6. Neighbourhood	Area
7. Latitude	Latitude Cordinates
8. Longtitude	Longtitude Cordinates
9. Room_Type	Listing space types
10. Price	Price in Dollars
11. Minimum_nights	Minimum nights required to stay
12. Number_of_reviews	No. of reviews written for the listing
13. Last_review	Last reviewed date for the listing
14. Reviews_Per_Month	Total review per month for the listing
15. Calculated_host _listings_count	Total no of listing against the host id
16. Availability_365	Number of days when listing is available for booking

General Information about the Data

Size of the dataset

782320

Shape of the Dataset

48895 (Rows)
&
16 (Columns)

Numerical Column (Quantitative Feature)

10

Categorical Column (Quanlitative Feature)

6

Descriptive Summary Statistics

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.720687	7.029962	23.274466	1.373221	7.143982	112.781327
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.154170	20.510550	44.550582	1.680442	32.952519	131.622289
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	0.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

Observations:-

- **reviews_per_month column has few missing values.**
- **Minimum value in price column is 0$ and max is 10,000$. (0$ means free airbnbs, impossible, so bad data)**
- **There is/are properties whose availability_365 for booking is 0 day. (explored later)**
- **Maximum value for minimum_nights column is 1250. (explored later)**

```
[ ] # A function for price_correction
    def price_correction(x):
        if x==0:
            return 100
        else:
            return x
```

```
[ ] df['price']=df['price'].apply(price_correction)
```

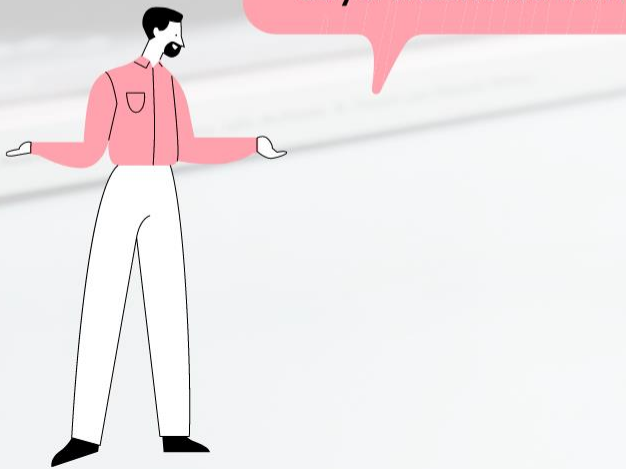
```
[ ] df['price'].apply(price_correction)

    0        149
    1        225
    2        150
    3         89
    4         80
            ...
    48890     70
    48891     40
    48892    115
    48893     55
    48894     90
    Name: price, Length: 48895, dtype: int64
```

Minimum price is zero which is not possible. we should replace price value as 10 or 100

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.743184	7.029962	23.274466	1.373221	7.143982	112.781327
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.144546	20.510550	44.550582	1.680442	32.952519	131.622289
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	10.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	1250.000000	629.000000	58.500000	327.000000	365.000000

	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
count	4.889500e+04	4.889500e+04	48895.000000	48895.000000	48895.000000	48895.000000	48895.000000	38843.000000	48895.000000	48895.000000
mean	1.901714e+07	6.762001e+07	40.728949	-73.952170	152.743184	6.942980	23.274466	1.373221	7.143982	112.781327
std	1.098311e+07	7.861097e+07	0.054530	0.046157	240.144546	17.530294	44.550582	1.680442	32.952519	131.622289
min	2.539000e+03	2.438000e+03	40.499790	-74.244420	10.000000	1.000000	0.000000	0.010000	1.000000	0.000000
25%	9.471945e+06	7.822033e+06	40.690100	-73.983070	69.000000	1.000000	1.000000	0.190000	1.000000	0.000000
50%	1.967728e+07	3.079382e+07	40.723070	-73.955680	106.000000	3.000000	5.000000	0.720000	1.000000	45.000000
75%	2.915218e+07	1.074344e+08	40.763115	-73.936275	175.000000	5.000000	24.000000	2.020000	2.000000	227.000000
max	3.648724e+07	2.743213e+08	40.913060	-73.712990	10000.000000	365.000000	629.000000	58.500000	327.000000	365.000000

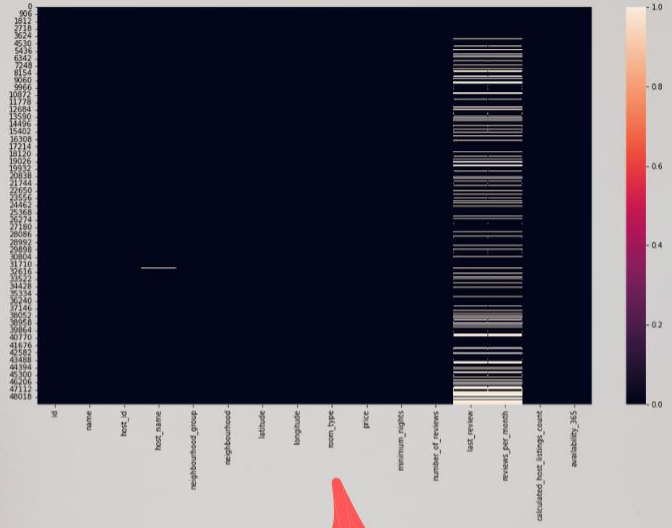


NULL Values in the Dataset

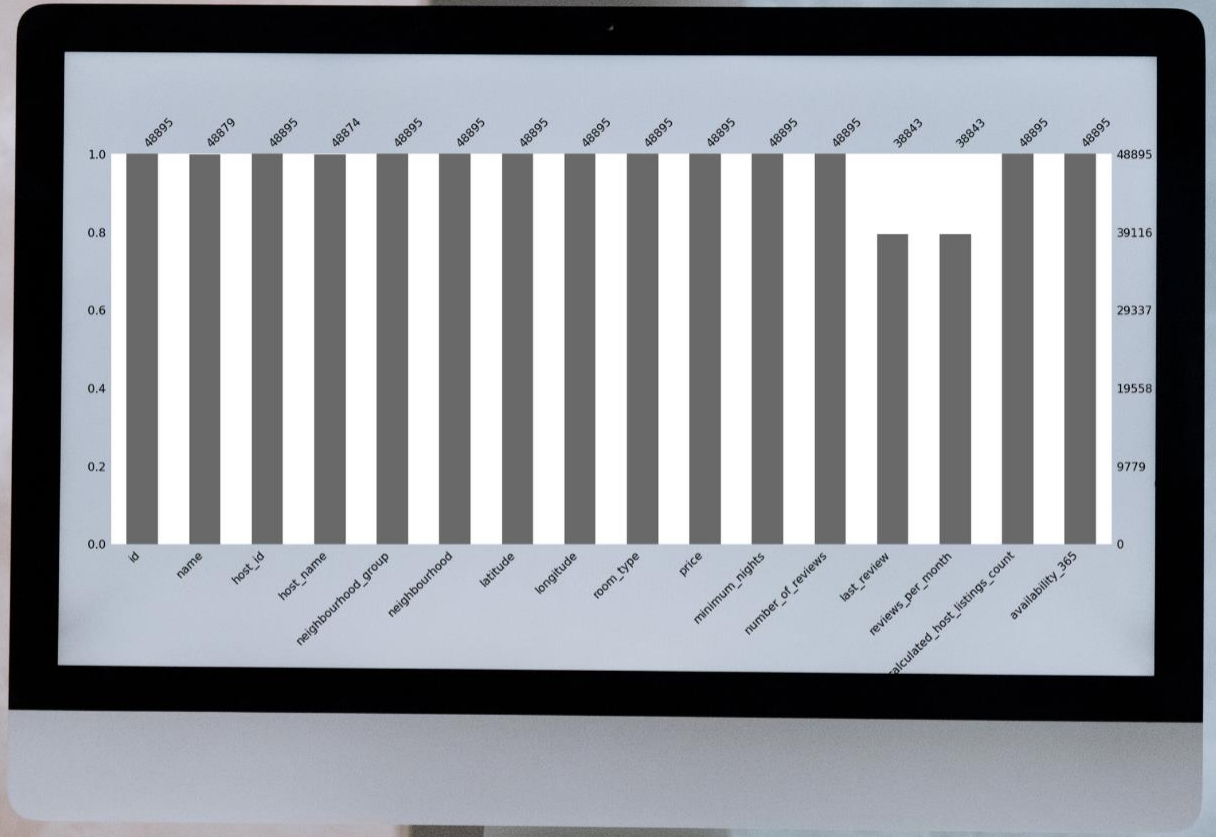
```
[ ]  # Checking for missing values

     df.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
```



Visualization of Null values using Heatmap



using missingno plot can display
count of non - null values

Handling Missing Value

```
def impute_median(series):
    return series.fillna(series.median())
```

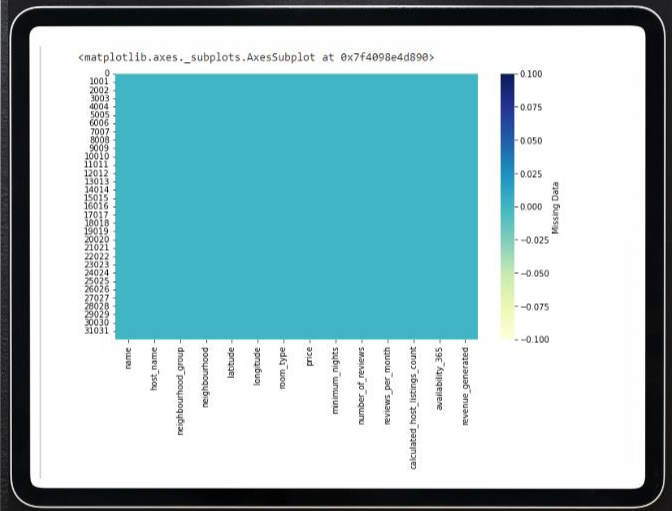
```
df.reviews_per_month=df["reviews_per_month"].transform(impute_median)
```

```
df['host_name'].fillna("unavailable",inplace = True)
df['name'].fillna("unavailable",inplace = True)
```

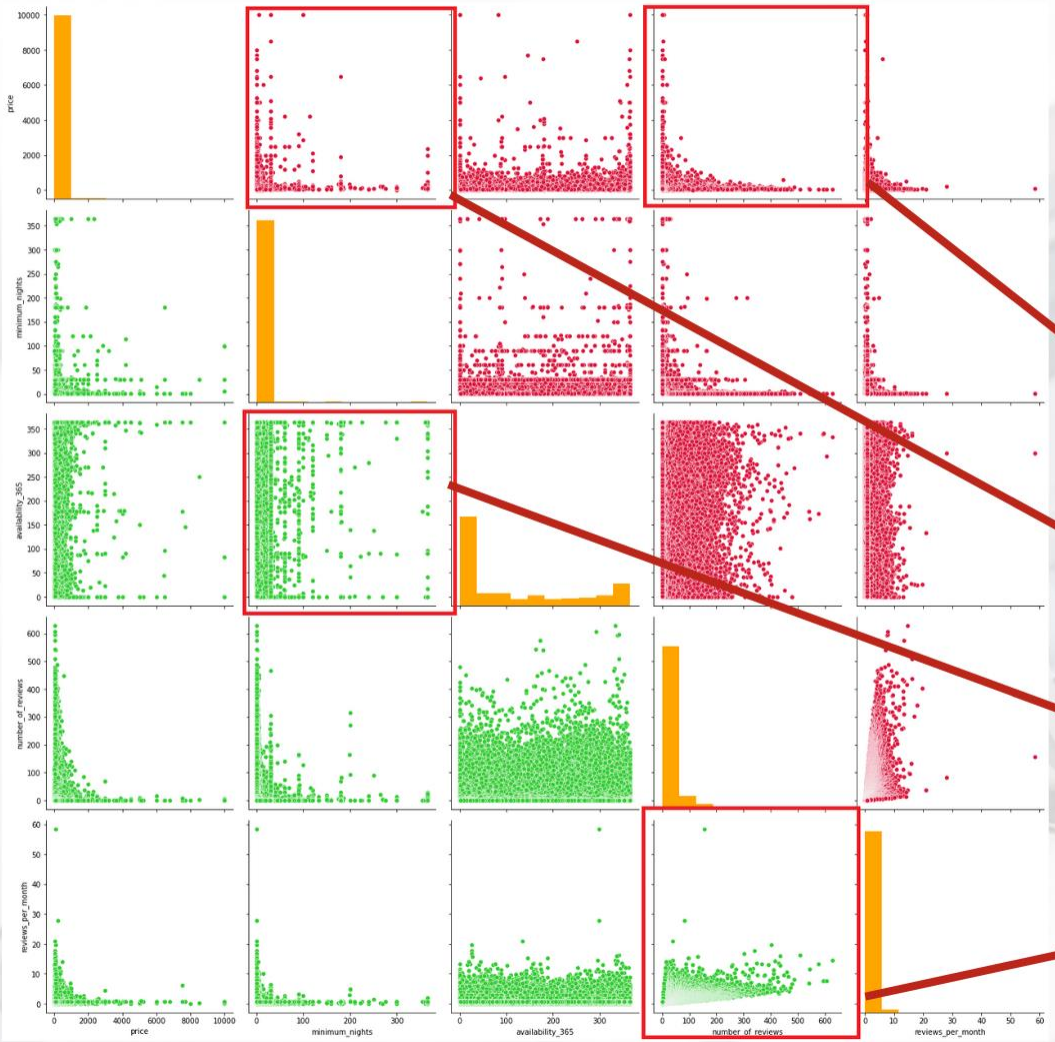
```
df.isnull().sum()
```

```
name                            0
host_name                       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
dtype: int64
```

```
plt.figure(figsize=(10,6))
sns.heatmap(df.isna(),cmap="YlGnBu",cbar_kws={'label': 'Missing Data'})
```

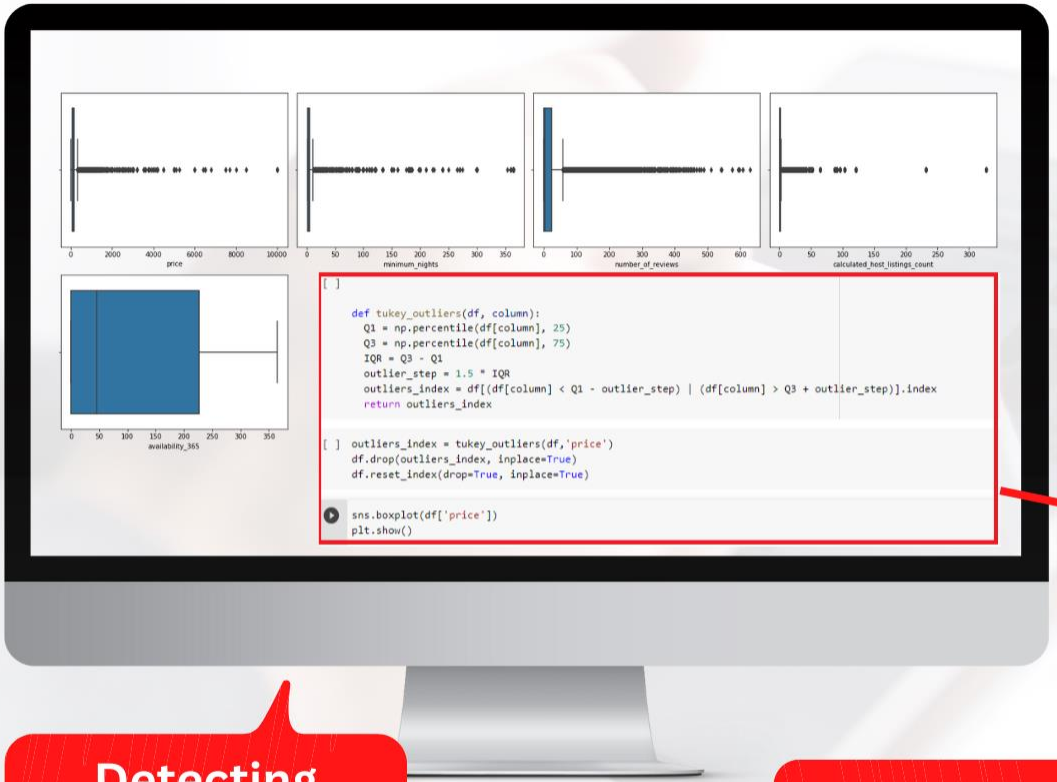


Correlation



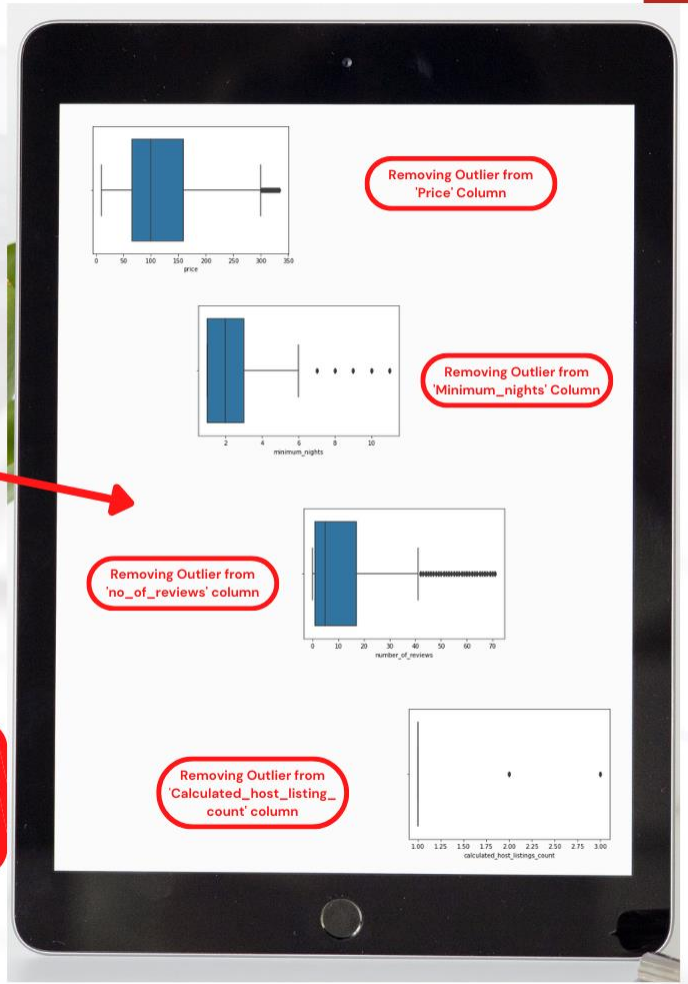
Relationship analysis between variable (pairplot)

1. There are more number of reviews where price is low
2. As minimum nights of booking increase price decreases significantly.
3. Unexpected minimum nights at 0 availability
4. Number of reviews and review per month follows linear relationship between them. Some plot has a extreme points



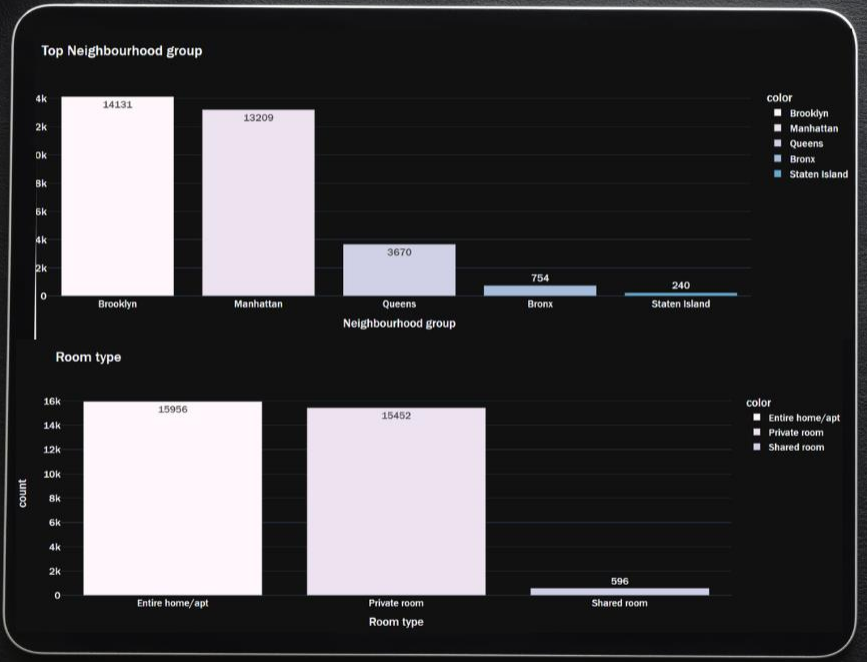
Detecting Outlier

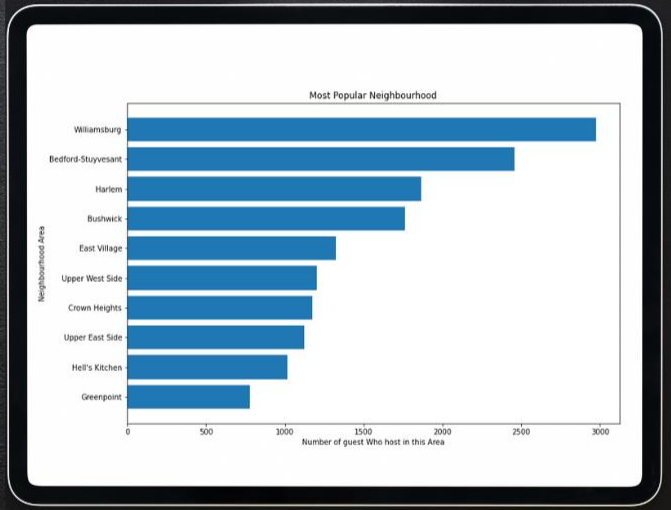
Removing Outliers




DATA
VISUALIZATION

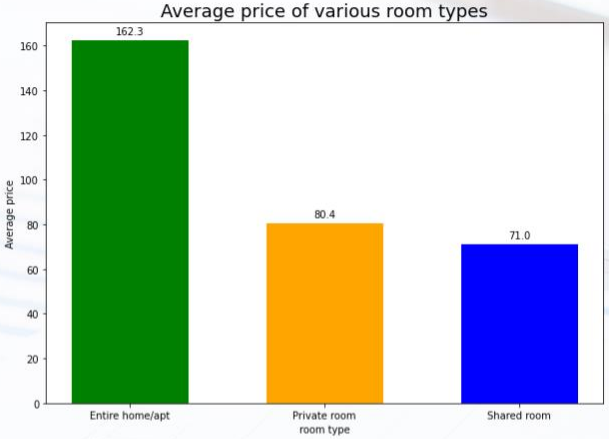
Univariate Analysis

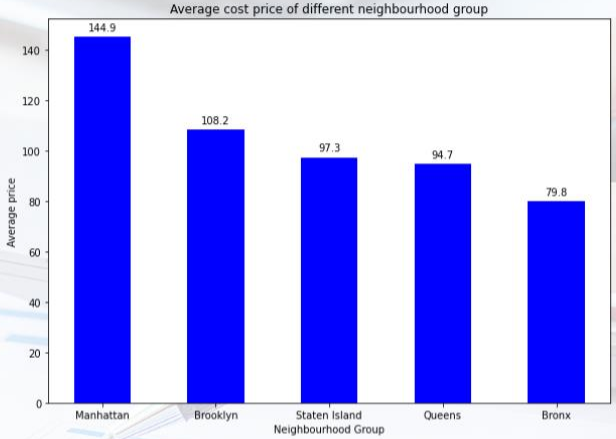


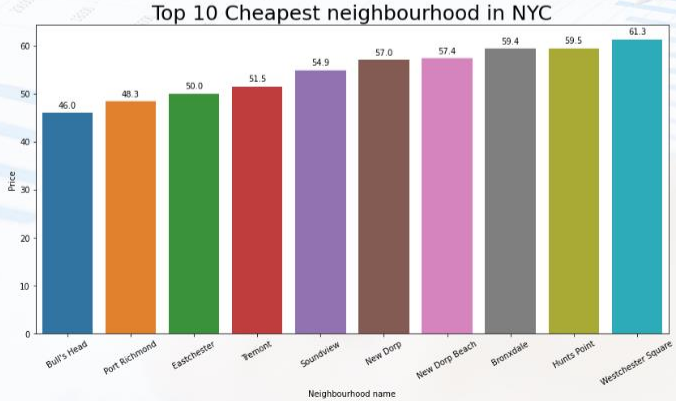



Average price for each room type
price
Room Type
71.0
80.4
162.3
Shared room
Private room
Entire home/apt
Price
availability_365
Manhattan
Brooklyn
Queens
Staten Island
Bronx
neighbourhood_group
Room type by region
Number of Room types
room_type
Entire home/apt
Private room
Shared room
7462.0
5461.0
286.0
6634.0
7328.0
169.0
1479.0
2085.0
106.0
120.0
118.0
2.0
261.0
460.0
33.0
Region
Bivariate Analysis

EXPLORATORY DATA ANALYSIS







Price Analysis

Observation:

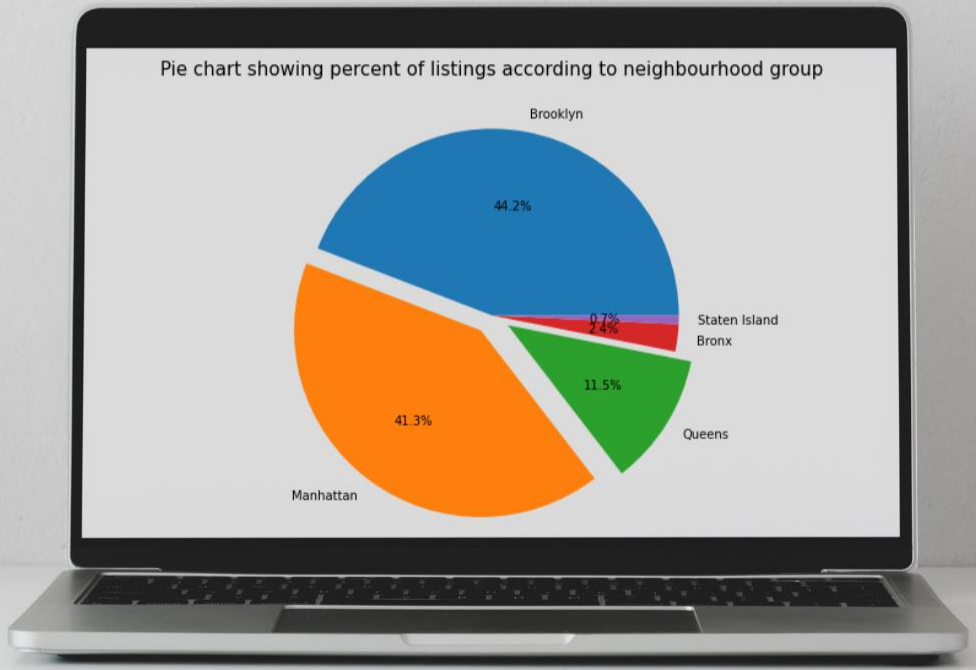
Order of costliest neighbourhood group on basis of average price and price distribution => Manhattan > Brooklyn > Staten Island > Queens > Bronx.

Staten Island was having least no. of listing but it is not the cheapest.

For buying a property to get business with airbnb, Brooklyen and queens are most prefered neighbourhood group because of less saturation level compaired to Manhatten

EXPLORATORY DATA ANALYSIS

Listing Analysis



Result:

1.Manhatten & Brooklyn are having high no. of listing.

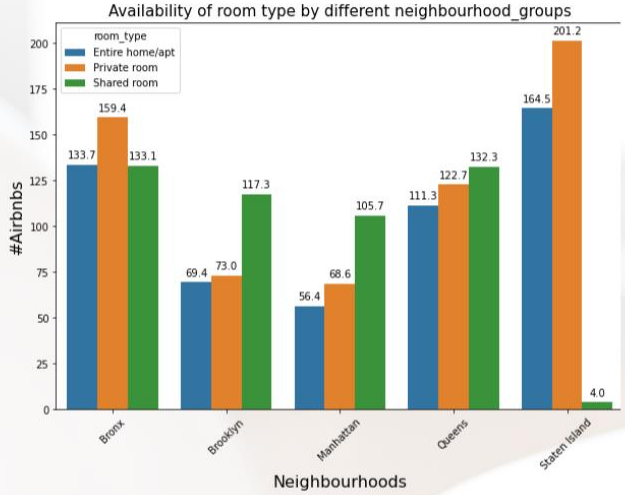
2. Staten island and Bronx have low no. of listing.

Inference:

If we want to do advertisement or marketing for selling flats we should focus on Manhattan and Brooklyn.

EXPLORATORY DATA ANALYSIS

Availability analysis



Result:

1.Private room has highest mean availibility..

2. Entire home has least mean availibility..

Inference:

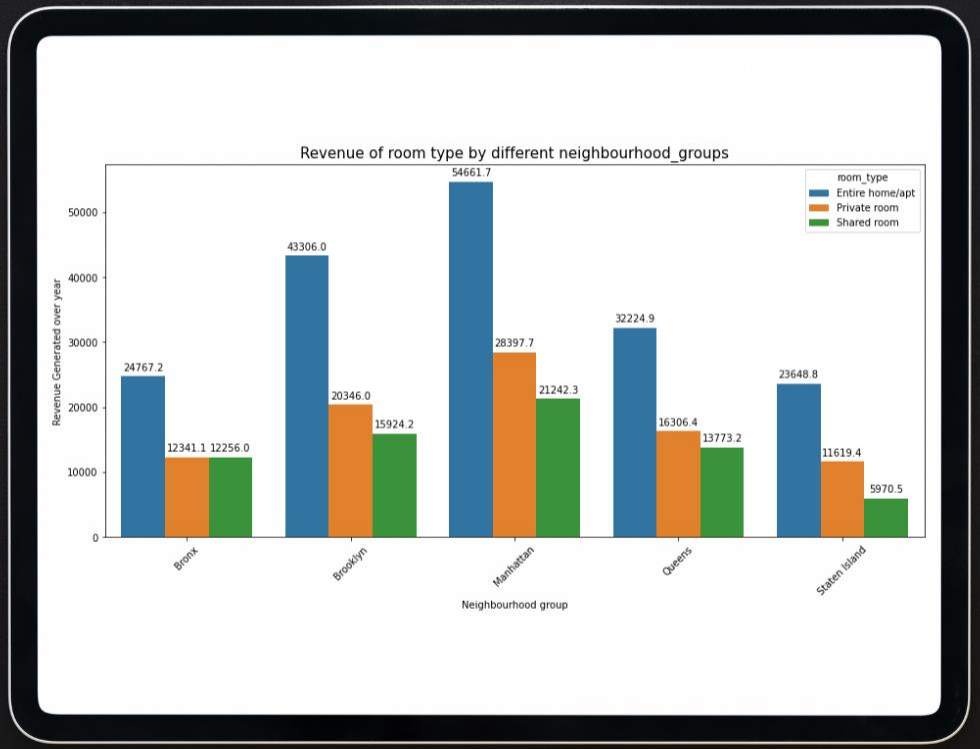
Manhatten and Brroklyn have less availibility then compared to other neighbourhood groups which is good for the host having these propeties.

EXPLORATORY DATA ANALYSIS

Profitability Analysis

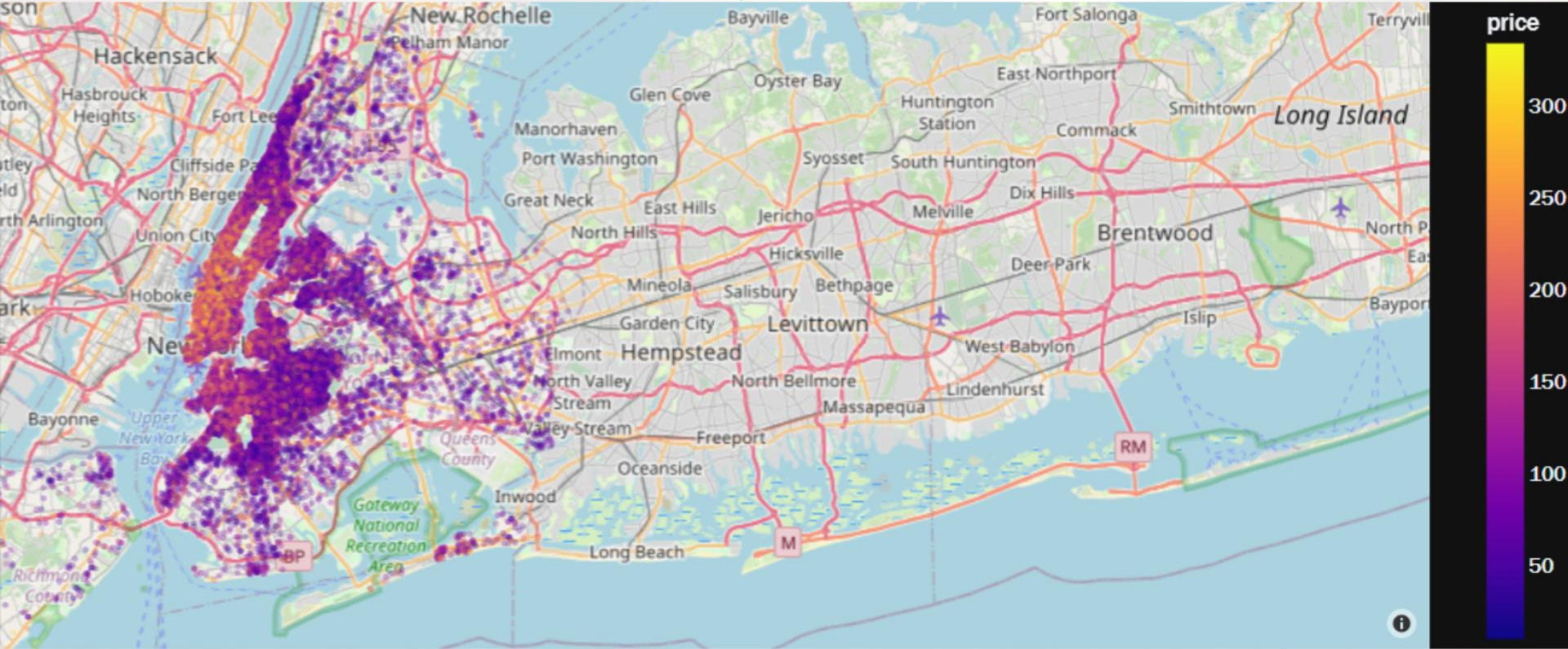
Observation:

As we can see that in every neighbourhood "Entire home" has generated most revenue. So buying a property "Entire home" and renting it, is a profitable business irrespective of neighbourhood type.



Visualization on Maps using Latitude and longitude

Visual Analysis using Satellite Map



Maps of Room Type

Word Cloud of the Most Keyword Used

CONCLUSION

1. Entire home/apt is highly expensive.
2. Manhattan living cost is highest, Bronx living cost is lowest
3. Cheapest neighbourhood is Bulls head.
4. Manhattan have the highest no. of listing.
5. Private room has the highest availability;
6. the Entire home has least availability
7. Manhattan, Brookyln & some parts of Queens have a high traffic of Booking
8. Most visitors don't prefer shared rooms, They tends to visit private or entire home/ apt.
9. Williamsburg is the neighbourhood with highest number of listings.
10. Private rooms,Apartment,Brookyln,Manhattan are most number time used keywords
11. Queens is the neighbourhood which has received maximum number of reviews, bronx was least.