Airbnb booking Analysis EDA

Project Summary -

The agenda of this project/EDA will be to analyse Airbnb bookings dataset and pull fruitful information out of it. Moreover perform data cleaning, data transformation in order to get processed data.

Airbnb, Inc. is an American San Francisco-based company operating an online marketplace for short- and long-term homestays and experiences. The company acts as a broker and charges a commission from each booking. The company was founded in 2008 by Brian Chesky, Nathan Blecharczyk, and Joe Gebbia. Airbnb is a shortened version of its original name, AirBedandBreakfast.com. The company is credited with revolutionizing the tourism industry, while also having been the subject of intense criticism by residents of tourism hotspot cities like Barcelona and Venice for enabling an unaffordable increase in home rents, and for a lack of regulation.

This colab notebook will hold every informative data that will give a brief overview of Airbnb dataset insight. Post loading the desired Airbnb dataset we will have relevant information such as id of the property, name of the property, host_id of the guest, room_type like private room or entire property for rent, price of the property based on the selection either private or entire, no.of reviews making it easy for guests to know more about property before making booking etc. The dataset comprises 48895 rows × 16 columns.

We will remove all the null values, handle missing data, and set up outliers. By the end of this project we will have clear and crisp data. A concluding statement will eventually mark the completion of this project, hence leaving a powerful data insight for the audience to view and use for their projects further.

Problem Statement

Extracting useful information from a website about a particular property, guest name, room type etc, can be a tedious task. There is a lot of scattered data present over the internet and finding meaningful information out of it can be a very painful and challenging job, requiring more time and effort.

Define Your Business Objective?

In order to reduce extra efforts we can do EDA which gives clear and concise data along with visually appealing charts. The EDA provides better insight of all the necessary information that needs to be extracted saving additional time and effort put forward. The objective is very straight forward, to present the dataset in such a way that the audience could get a meaningful insight from it without going through the entire dataset.

Import Libraries

```
import matplotlib.pyplot as plt

import seaborn as sns

import pandas as pd

import numpy as np
```

Dataset Loading

```
from google.colab import drive

drive.mount("/content/drive")

data = pd.read_csv("/content/drive/MyDrive/Airbnb NYC 2019.csv")
```

Dataset Rows & Columns count

```
no_of_rows = data.shape[0]

no_of_columns = data.shape[1]
```

Duplicate Values

```
duplicates = data.duplicated().value_counts()   duplicates
```

Missing Values/Null Values

```
new = data.isnull().sum().reset_index()

visual = data.isnull().value_counts().reset_index()
```

Summary of the findings:

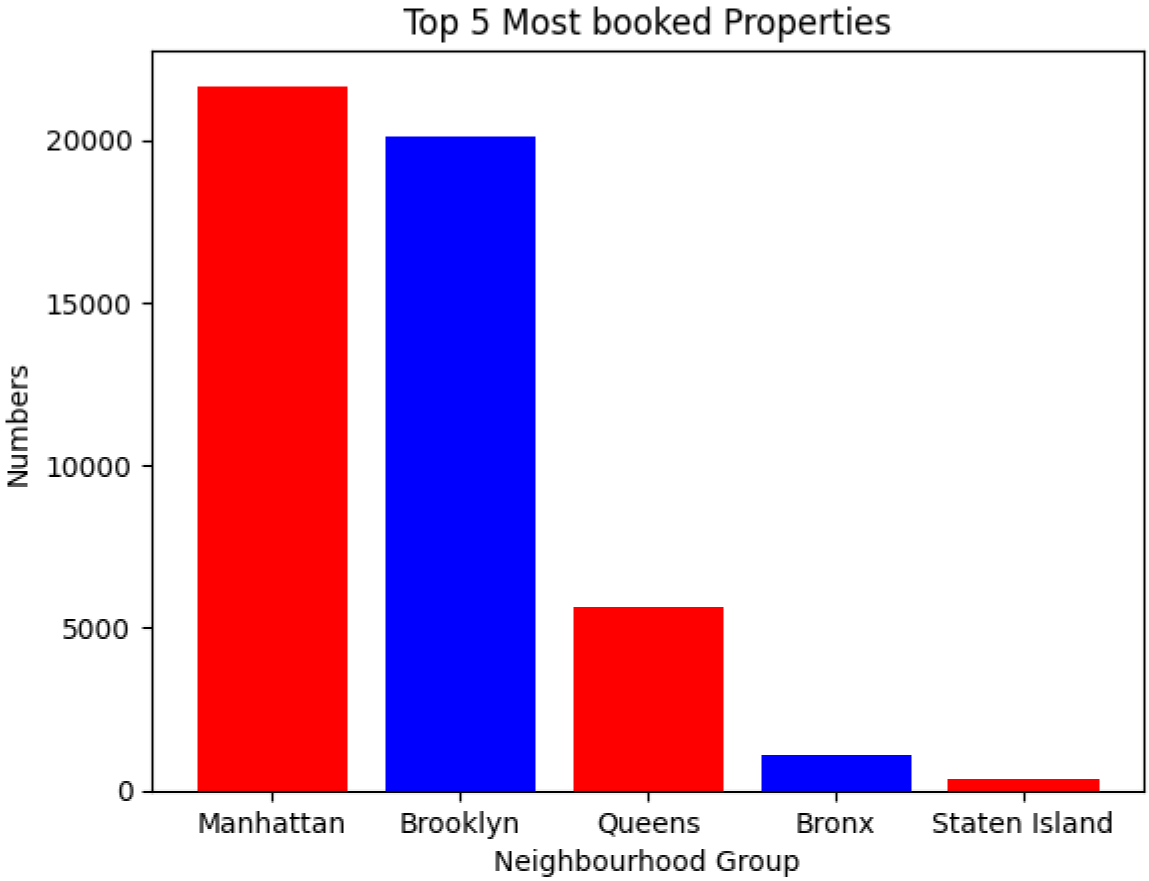
1. Based on the findings till now it can be concluded that the Airbnb dataset gives a glimpse of the kind of the data that can be fetched.
2. We can have a clear idea from the dataset, what kind of information is stored, how many rows and columns are there, how many null values are there etc.
3. Many columns in the dataset hold null values. We have tried to find the null values in each column, **last_review** and **reviews_per_month** column having the most no.of nulls whereas **name** and **hostname** column having the least.
4. **last_review** and **reviews_per_month** column has NaN in the same rows stating that these 2 columns share some sort of correlation. Moreover, the last_review column doesn't serve the utmost importance as **reviews_per_month** column is enough, so we can drop it and fill NaN in **reviews_per_month** with 0.
5. NaN in **name** and **host_name** column are very less in number hence we can replace it with **unknown** in both the columns.
6. As data cleaning requires replacing null values but replacing null values can sometimes mislead the dataset.
7. So, it must be worked upon cautiously only where it's required.

Manipulation on the basis of findings:

1. We have done data wrangling on the previous dataset provided which had some missing values.
2. As we initiated the **drop_duplicates()** method but didn't find any duplicate values in the dataset.
3. Then we performed **isnull()** to find null values and concluded that we have **4 columns** with some null values.
4. For **name** and **host_name** column replaced **NaN** with **unknown** as null values are very less in number. We could have dropped the columns but it would then remove some meaningful data as well.
5. We then replaced the null values in **reviews_per_month** column with **0** as it didn't impact the meaning of the dataset.
6. Since **last_review** has numerous null values and the info contained in it is almost there in **reviews_per_momth** column, we dropped it off.
7. We transformed the **price** column, since property can't be available just for free i.e., 0 price. So, we replaced all price 0 with mean price value.
8. Hence we can say we got a processed and clean dataset.

# Chart - 1

The chart shows the list of **top 5 most booked properties.** This gives a kind of understanding as to which property is most visited and hence helps the user in making the right decision before booking.

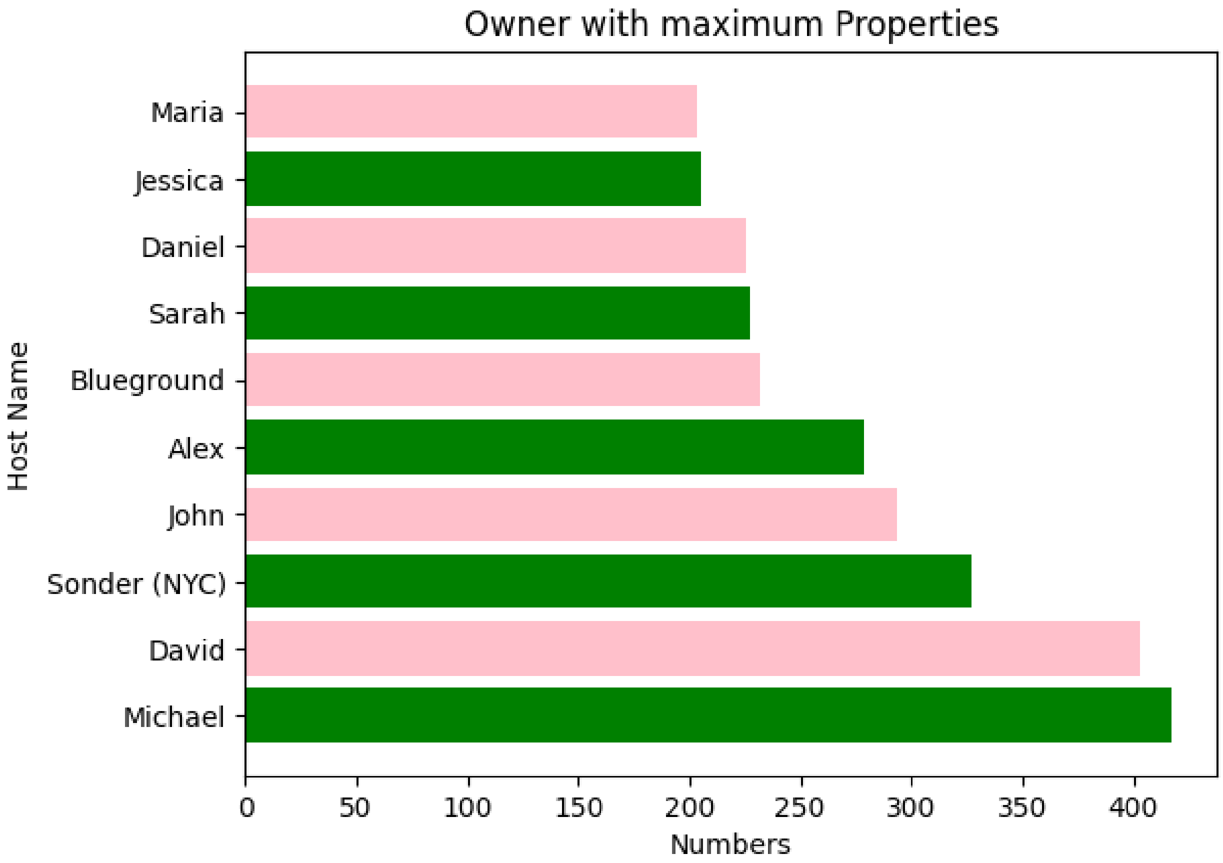


The above chart depicts the number of times property sold like:

1. **Manhattan** : 21661 times
2. **Brooklyn** : 20104 times
3. **Queens** : 5666 times
4. **Bronx** : 1091 times
5. **Staten Island** : 373 times

# Chart 2

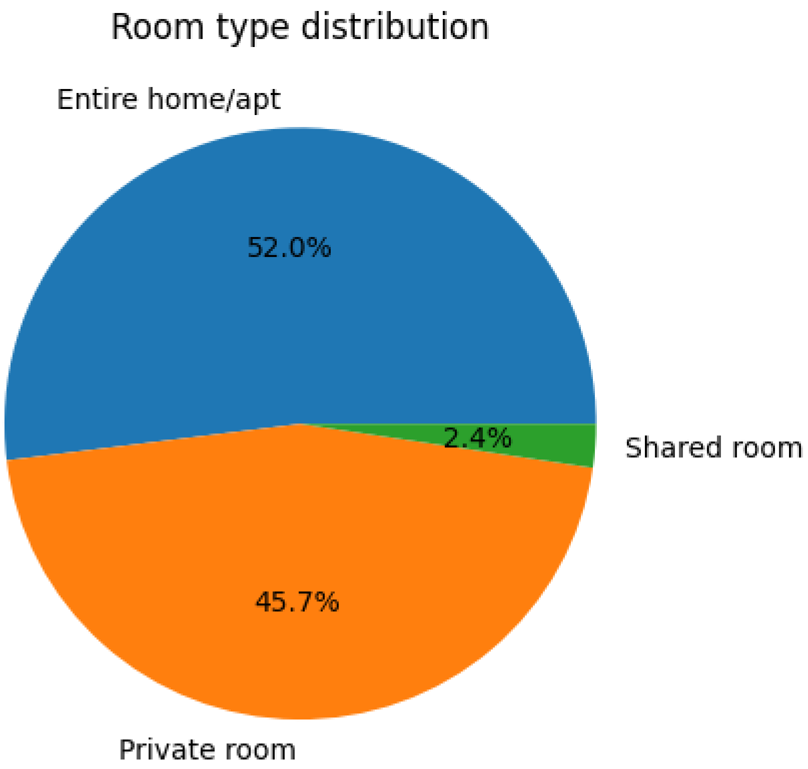
This chart is to get an overview of which host has how much property.



1. This specific chart gives an understanding of the host with maximum property available.
2. **Michael** holds maximum property i.e., **417** followed by **David 407**.

# Chart 3

The motive of selecting this specific chart is to get a clear picture of **room_type** distribution.

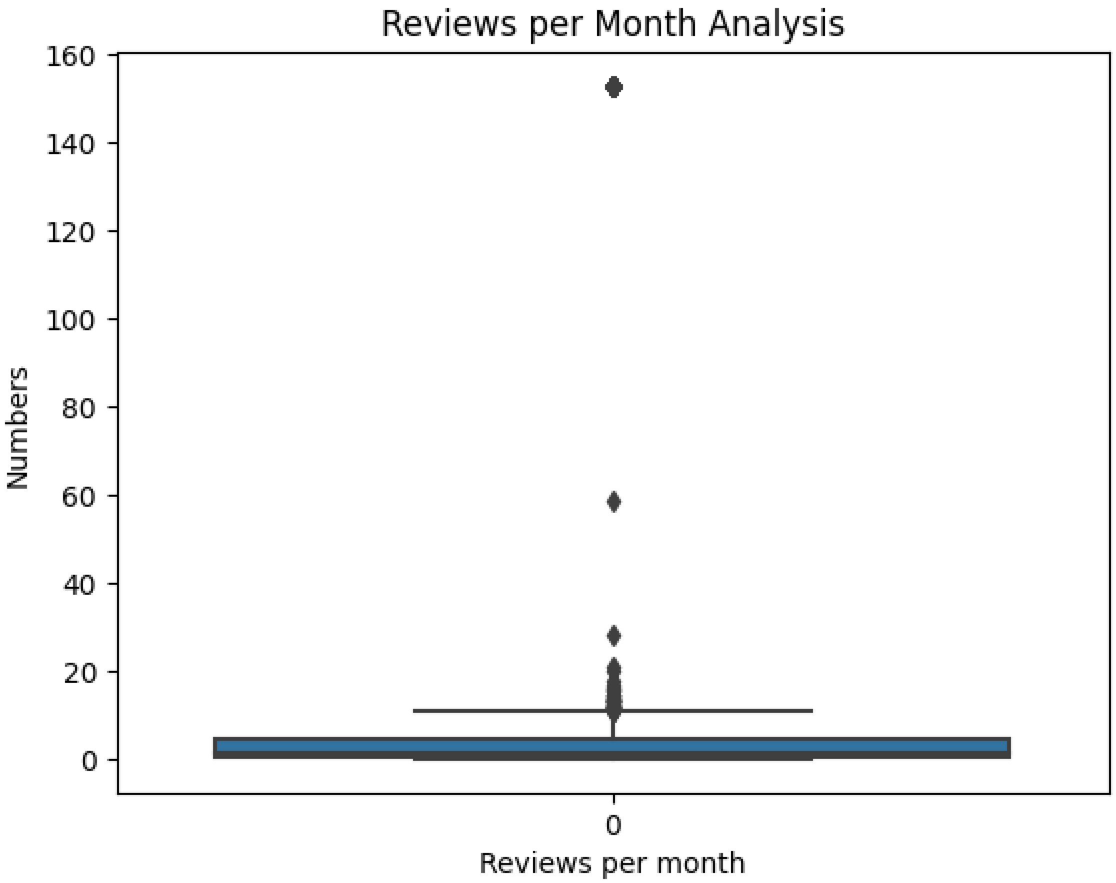


- From the above chart we can say that **% share** of different types of room_type distribution is as follows:

1. **Entire home/apt** has the maximum distribution nearly **52%.**
2. **Private room** being the second highest in terms of % distribution with **45.7%.**
3. **Shared room** having the least distribution with **2.4%.**

# Chart 4

The agenda behind selecting this type of chart is to get an overview of reviews per month distribution which in turn will give an insight of how often property is booked or liked.

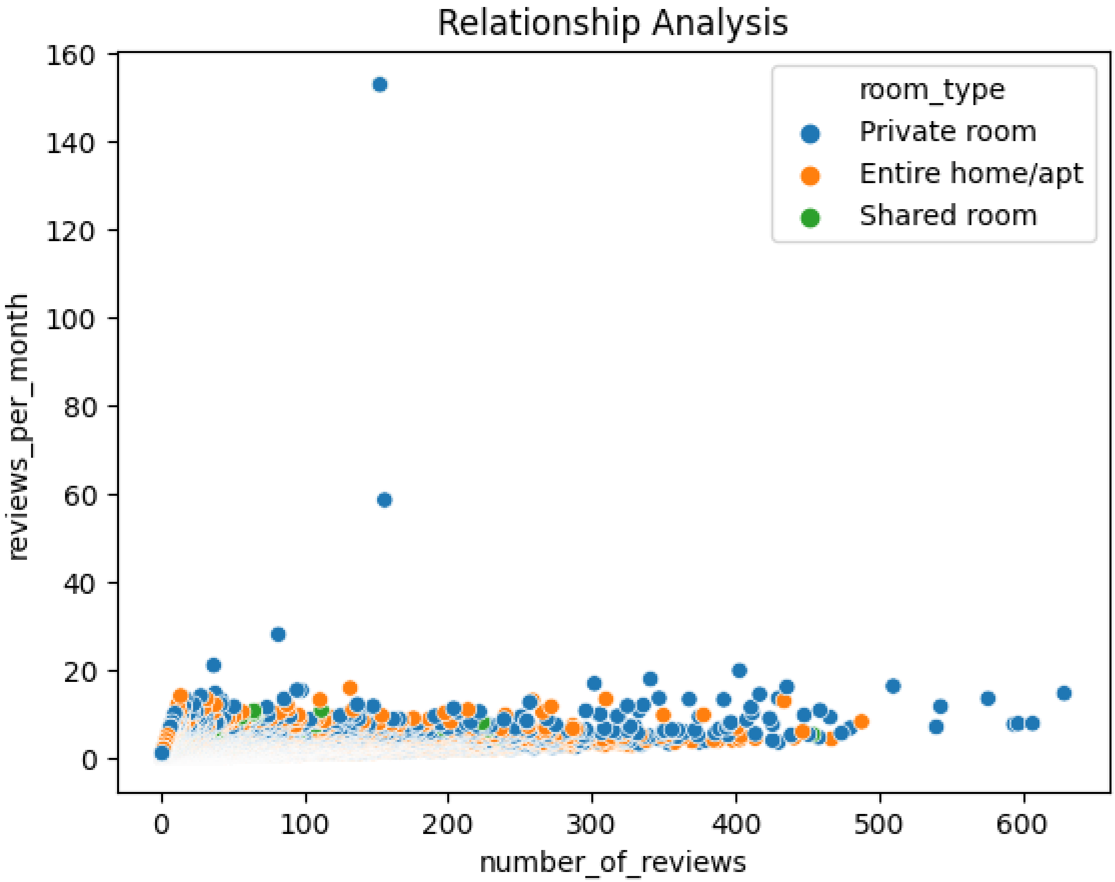


- We can typically verify the following points from chart:

1. The thin line just below the thick line depicts **minimum value** i.e., **0.01**.
2. The thick line represents the **median value** i.e., **1.22**.
3. The first point above the single line depicts the **mean value** i.e., **32.48**.
4. The last point at the top most of the chart represents **maximum value** i.e., **152.72**.

# Chart 5

The idea behind choosing scatterplot chart is to drive a relationship b/w **reviews_per_month** vs **number_of_reviews** based on **room_type**.

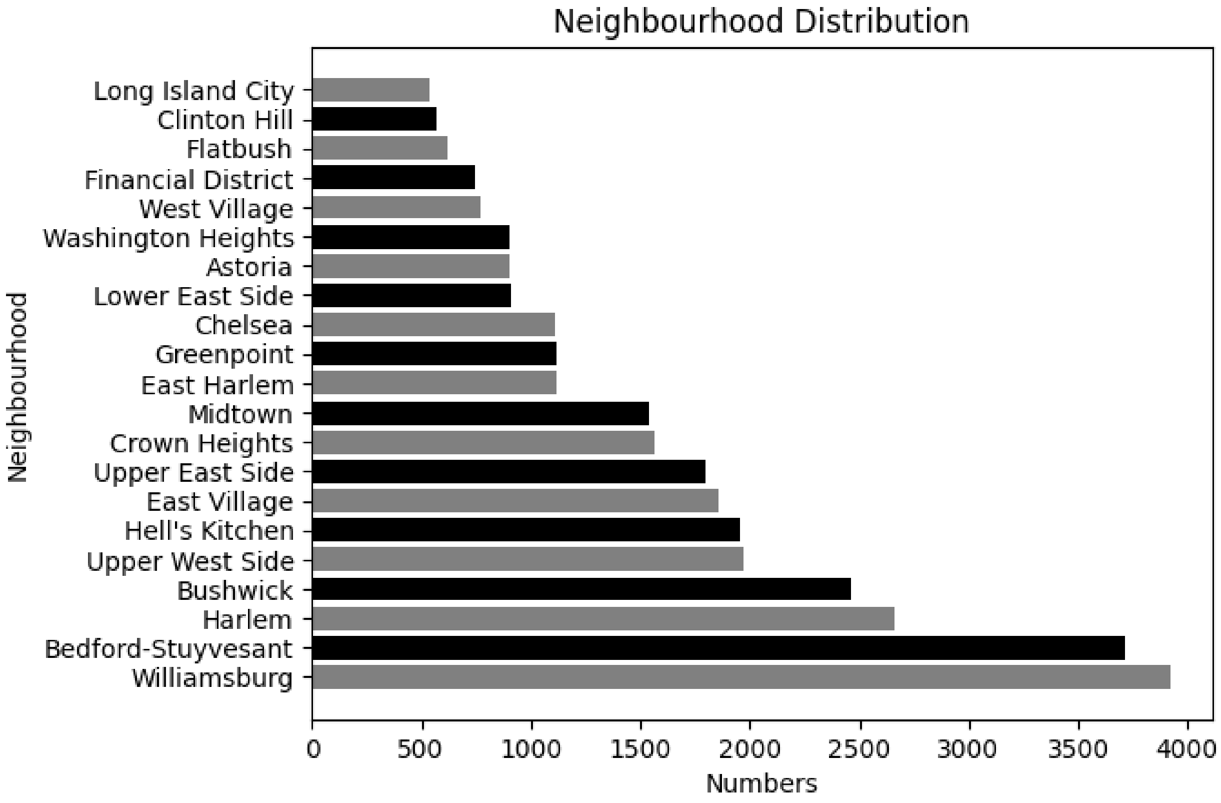


- From the above plot we can drive following insights:

1. **Entire home** type room has maximum review_per_month vs number_of_reviews depicting its **mostly liked** property type by customers.
2. Followed by Entire home type room, **Private room** is **next most liked** property from customers view point.
3. **Shared room** is **least liked** room type as it can be clearly seen the density of review_per_month and number_of_reviews is almost negligible.

Chart 6

The above chart serves the motive of understanding neighbourhood distribution, which neighbour is most likely visited. How many unique neighbours are there in dataset etc.

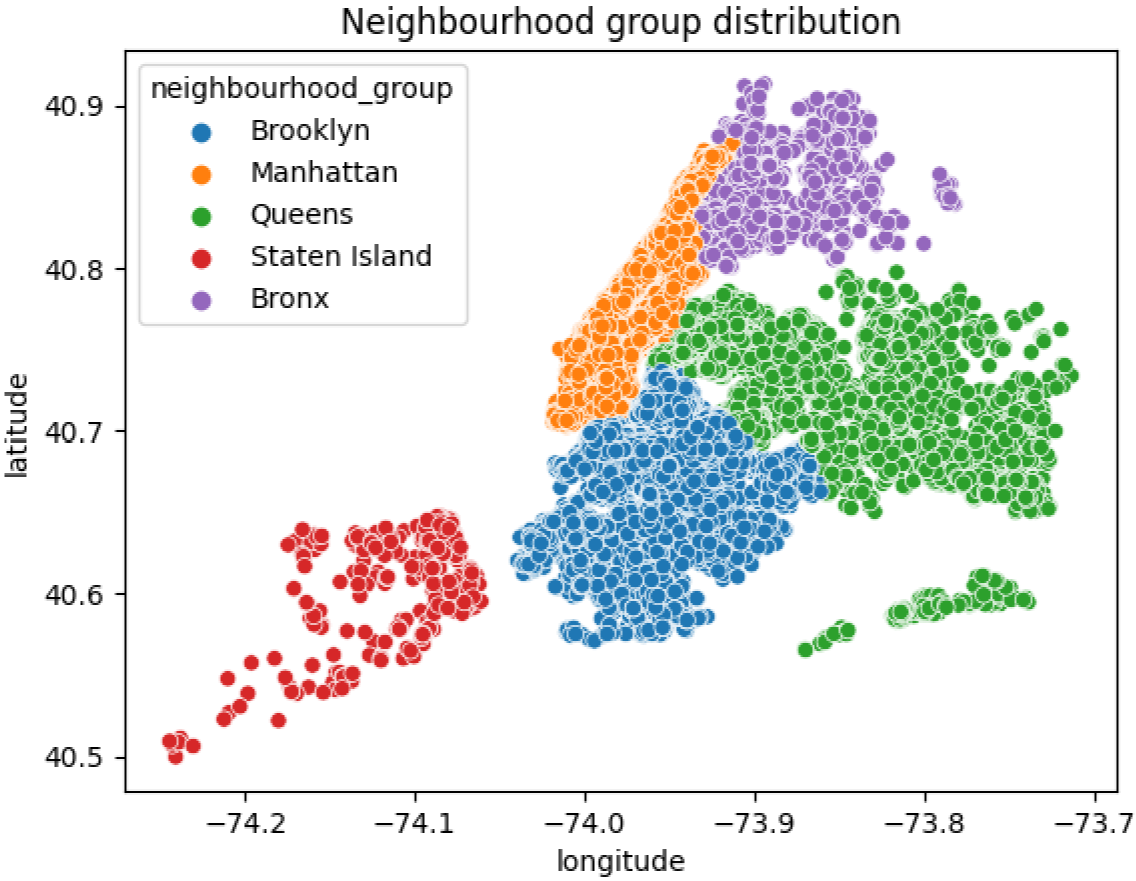


- We can forcast following insight from above data:

1. There are in total **221** unique **neighbours**.
2. Out of 221 unique neighbours we extracted out the top **21** most visited **neighbours** to understand the tendency of occupancy.
3. **Williamsburg** has made most no.of visits i.e., **3920** followed by **Bedford-Stuyvesant** i.e., **3714**.

# Chart 7

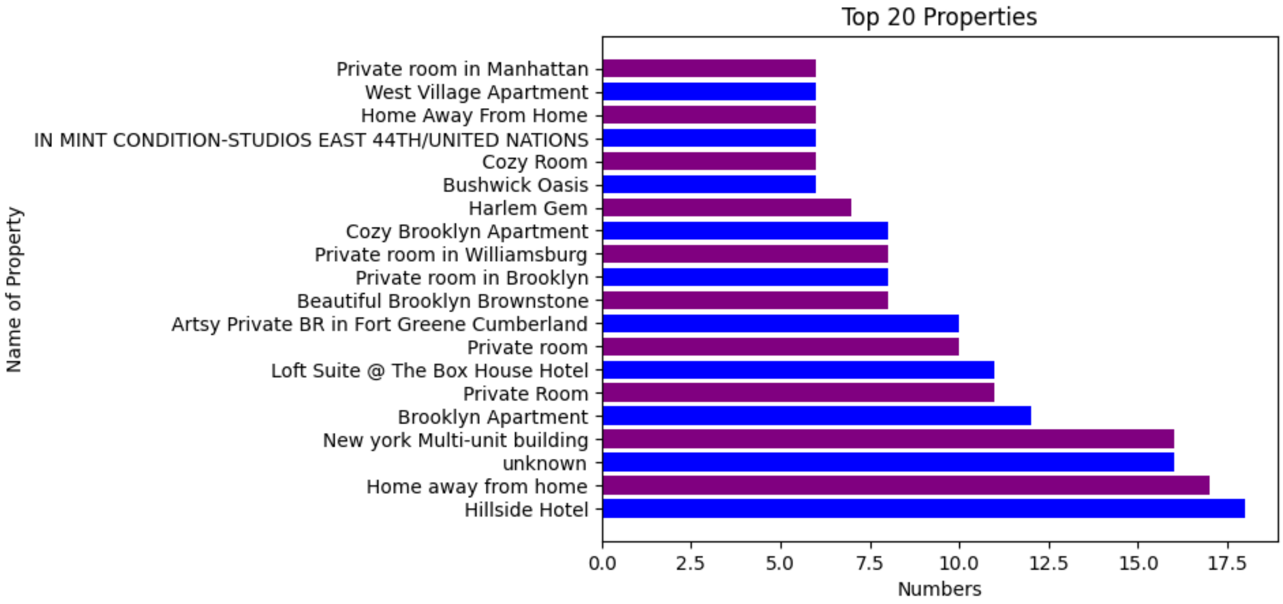
The theme behind selecting scatterplot is to grasp the footfall of occupancy in **neighbourhood_group** based on latitude vs longitude scale.



1. Detailed scatter plot gives an estimated observation about density of **neighbourhood_group** based on longitude vs latitude.
2. **Brooklyn** and **Manhattan** are the most dense neighborhood_group, followed by **queens**.

# Chart 8

The logic of driving bar chart is to conceive the insight of **Property_name** being majorly used or we can say popular amongst customers.

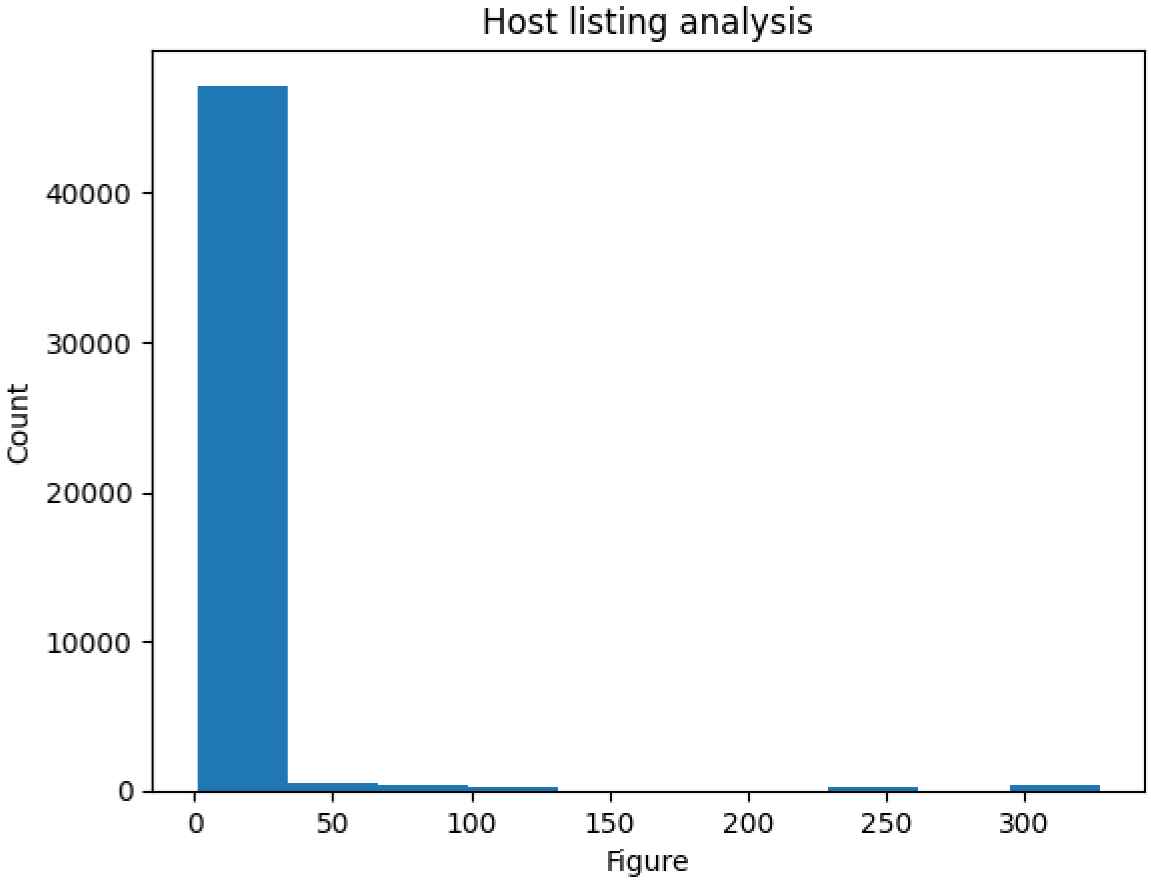


- We can deduce following kind of reference from above chart:

1. **Hillside Hotel** is highest in number i.e, with count 18, which clearly gives an indication that it is the most popular property amongst customers.
2. Followed by Hillside Hotel **Home away from home** shows maximum popularity.

# Chart 9

Selecting the above chart conveys a visual of how host listing is done in NYC. listing tells how frequent is the property taken into consideration.

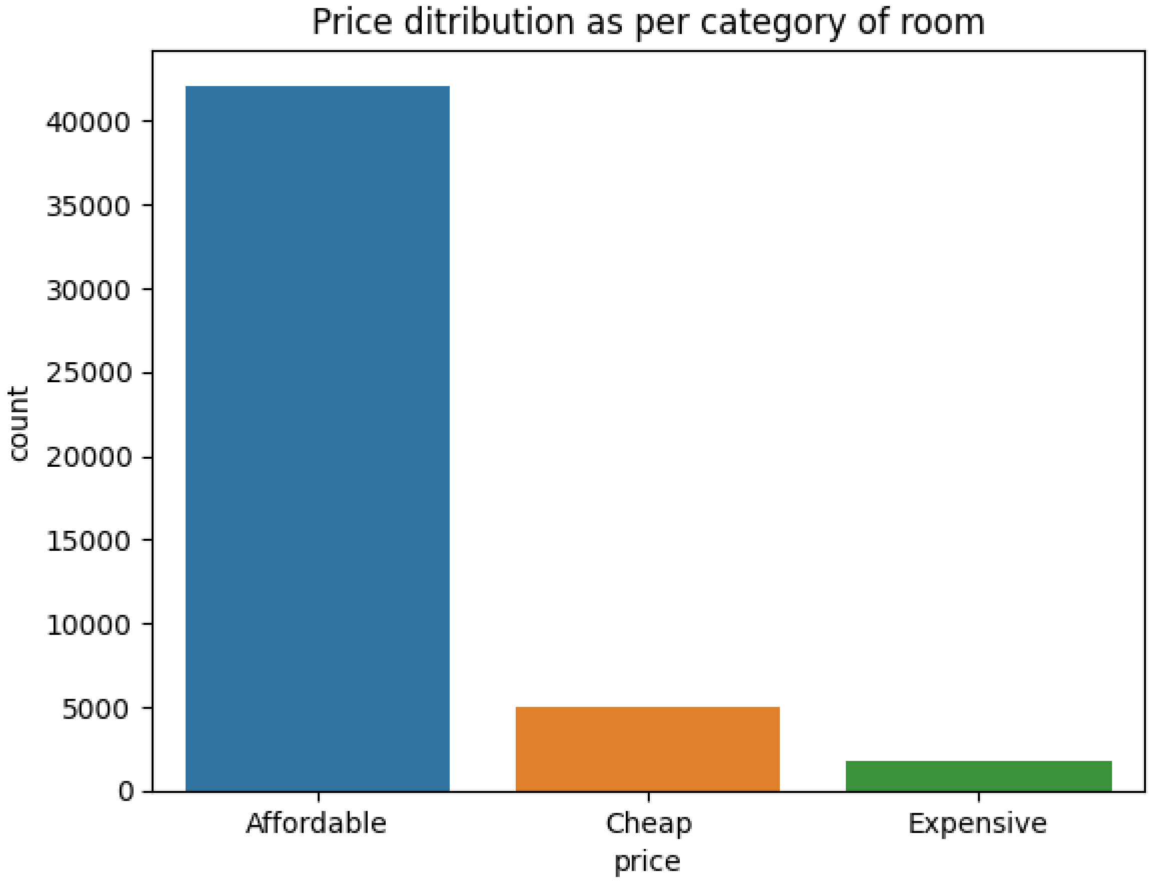


- We can deduce following insight from above graph:

1. The listing is **maximum** in the range **0-35** with alone constituting for more than **44000.**
2. From **35 - 135** a very small chunk of listing can be noticed.
3. A small block of listing in range **235-260** and then **290-327** can be rectified.

# Chart 10

Reason of selecting the countplot chart is to get a clear overview of how price distribution is done based on room type.

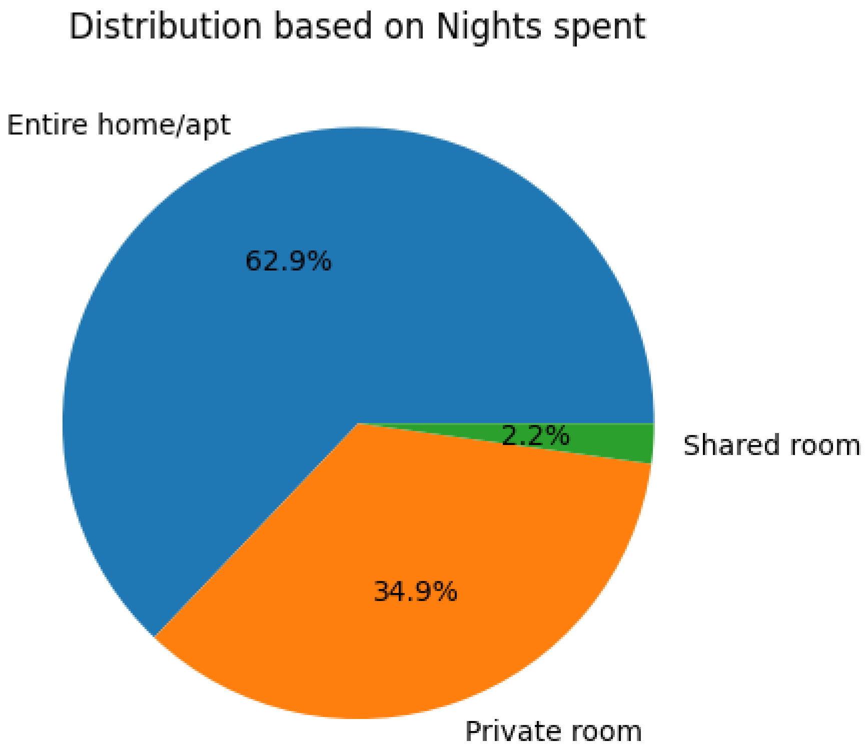


- Following points can be highlighted from above pictorial representation:

1. **Affordable property** is **pretty huge** in number.
2. **Cheap property** is **less** than affordable but slightly more than expensive.
3. **Expensive property** is very **less** in number.

# Chart 11

The above pie chart pictorial representation depicts distribution of nights spent by customers in different types of room.

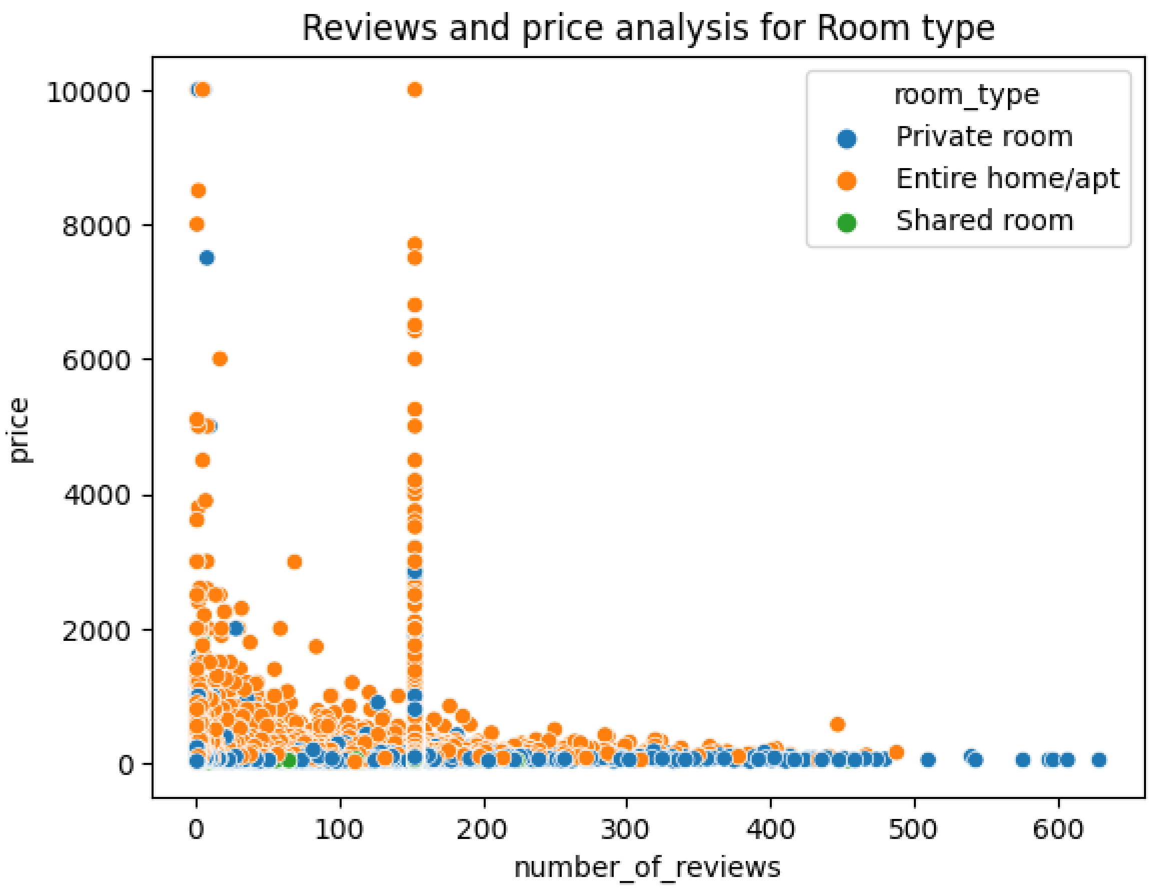


- We can navigate following breakup from above findings:

1. Nights spent by customers maximum in the case of the entire **room type** constitute for nearly **62.9%.**
2. **Private room type** has a share of **34.9%** in terms of nights spent.
3. Nights spent is least for **shared room type** with merely **2.2%.**

Chart 12

Scatter chart plot narrates, number of reviews vs price distribution analysis for particular room type.

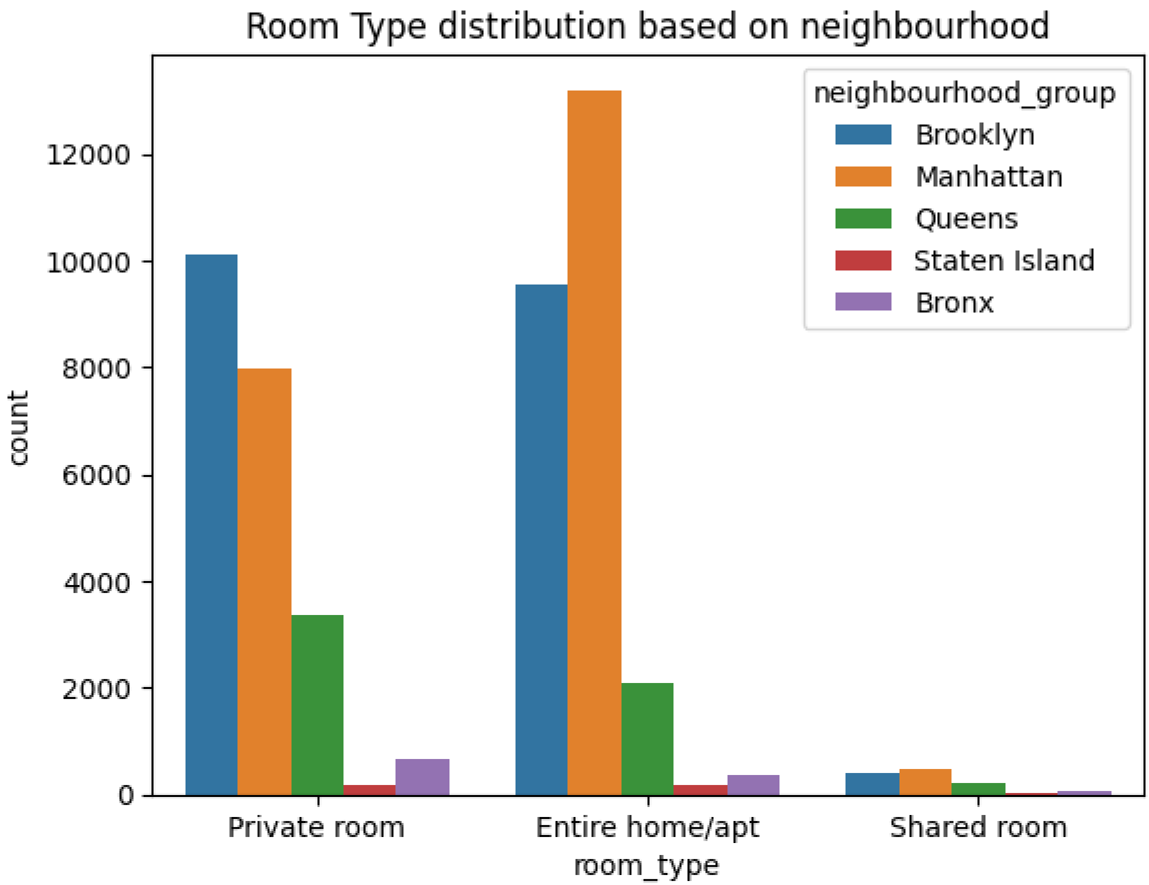


- Following theory can be penned down from above analysis:

1. **Private room** has the has very less price range and on other hand has maximum no.of reviews.
2. No.of reviews for **Entire home** is comparatively less than that of private room type moreover price is also high.
3. For **Shared room** price is pretty less but reviews are also quite negligible.

# Chart 13

The agend of deploying countplot on the above stat is to get an estimated distribution of room type based on particular neighborhood groups.

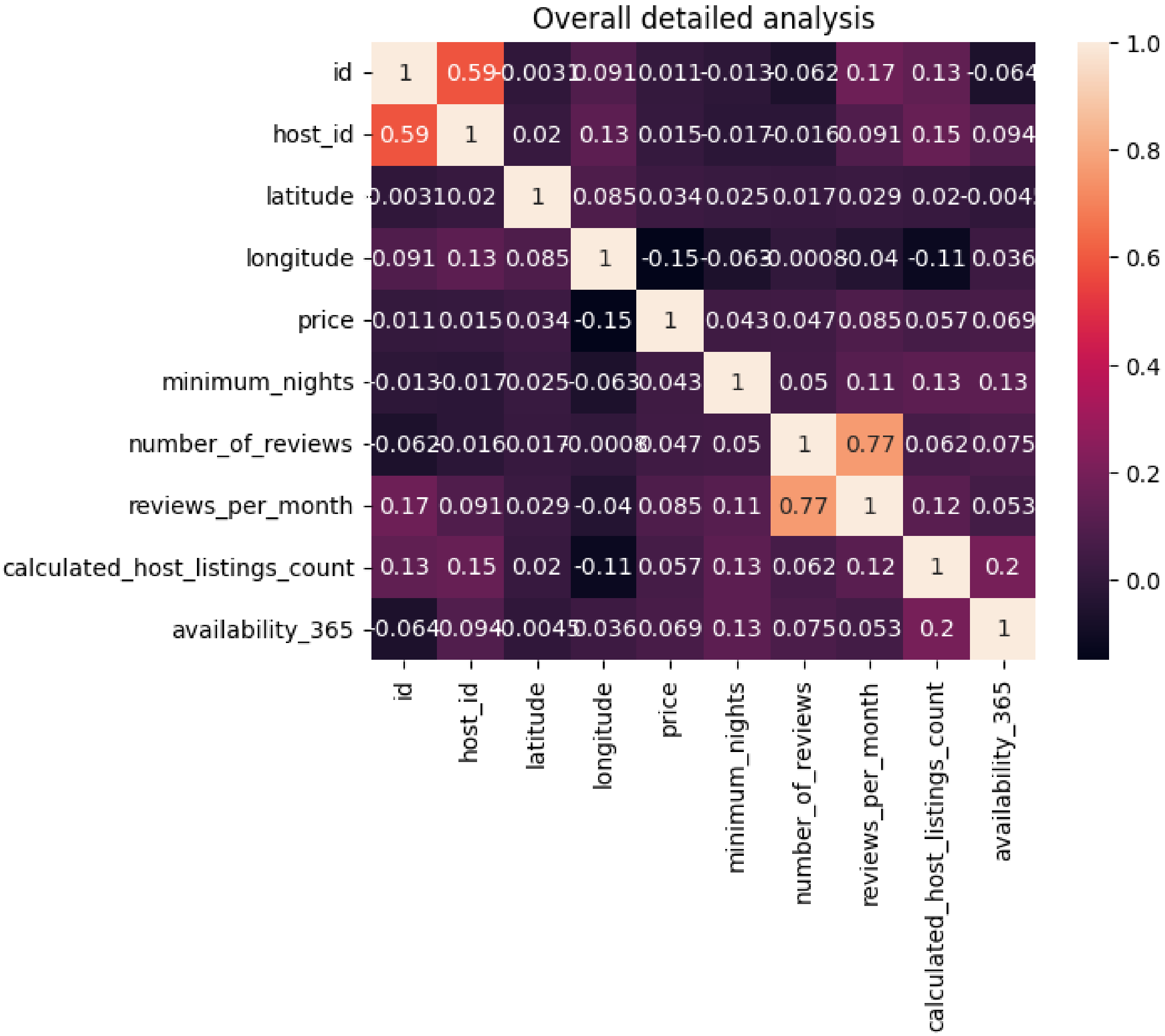


- The findings driven from above data can be described as follows:

1. **Private room** type is most in **Brooklyn** followed by **Manhattan** but least in **Staten Island**.
2. Overall **shared room** distribution is **quite less** for all neighbourhood groups.
3. **Entire home** type is very much prevalent in **Manhattan** followed by **Brooklyn** with least being in **Staten Island**.

# Chart 15

The selection of heatmap chart is to give a clear insight of how each and every series in the data set are linked to each other, how collaborative are the bonds and how loose are they.

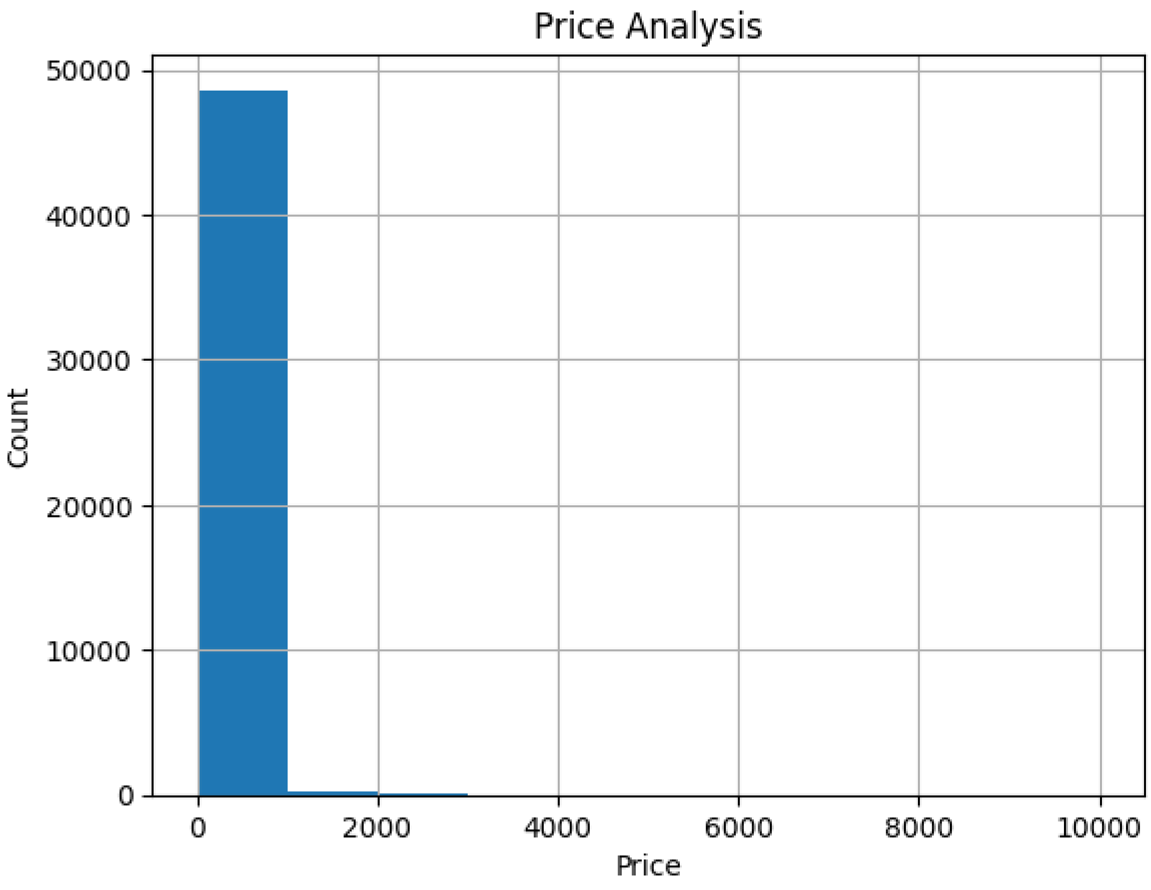


- Following output can be depicted from analysis shared above:

1. The **one's with lightest color** and indicated with numeric value 1 shares the strongest bond i.e., the series when compared with itself serves the deep collaboration.
2. As the color keeps darkening, the strength of the bonds keeps on deteriorating and the numeric values justifies it all.
3. The series who's joint venture depicts the darkest color and have least numeric value claims to have the weakest bond.

# Chart 16

The above histogram is a univariate analysis of just price, how it ranges in terms of count.

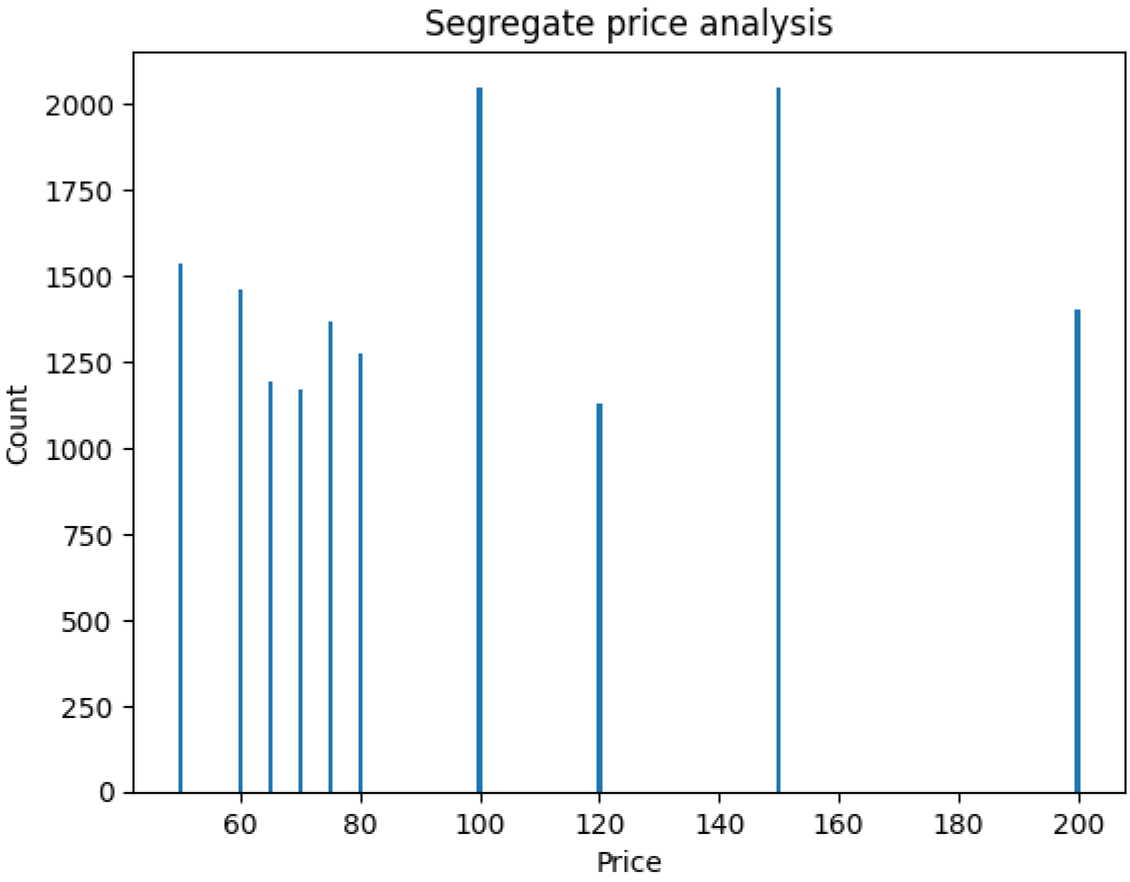


- The chart is a simple representation of price distribution all alone.

1. It simply defines how the price **range varies.**
2. Price range between **1 to 1000** is pretty **huge** in number.
3. **Above 1000** we can see a **small section** of the price range.

# Chart 17

The idea behind selecting barplot is to get an insight of segregate top 10 price distribution.

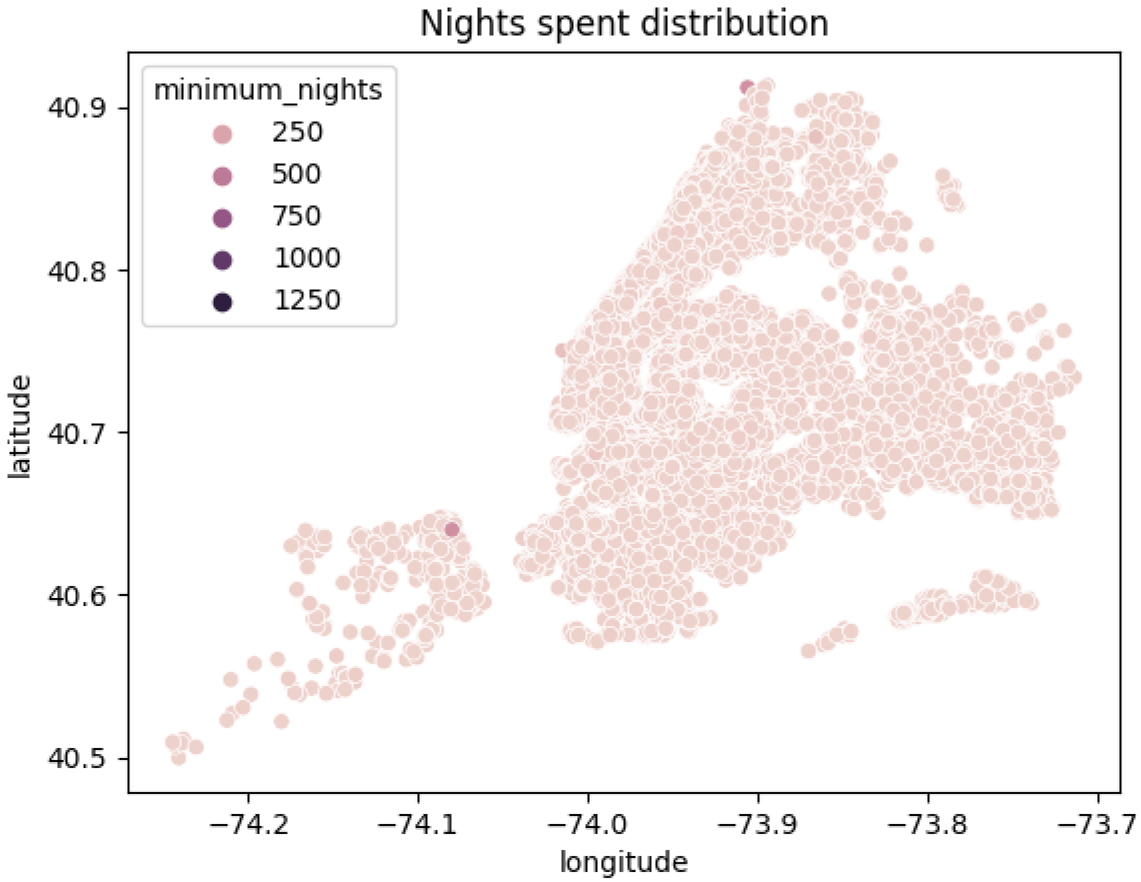


- The above distribution gives following output:

1. Property priced at **100** is **maximum** in number with average count **2051.**
2. Price of **150** is **2047** in count making it second highest in number.

#Chart 18

The scatter plot depicts how nights are spent based on longitude vs latitude.

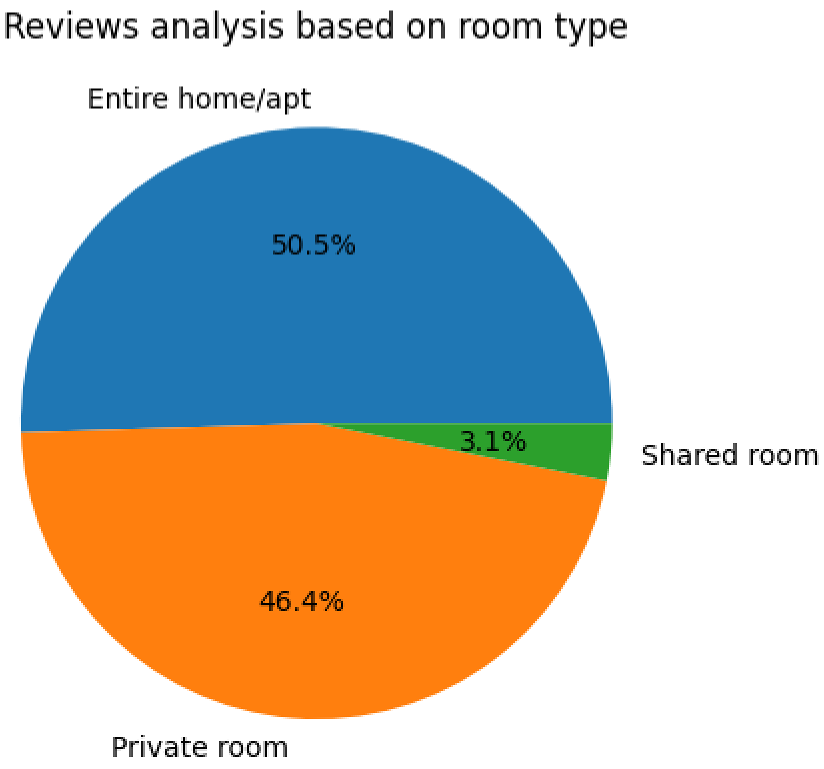


- The above data interprets following info:

1. Longitude vs latitude scales shows that **nights spent** above **500** are **very less** in number.
2. Maximum strength can be sited below **250.**

Chart 19

The pie chart give % share of reviews per month carry forward by each room type.

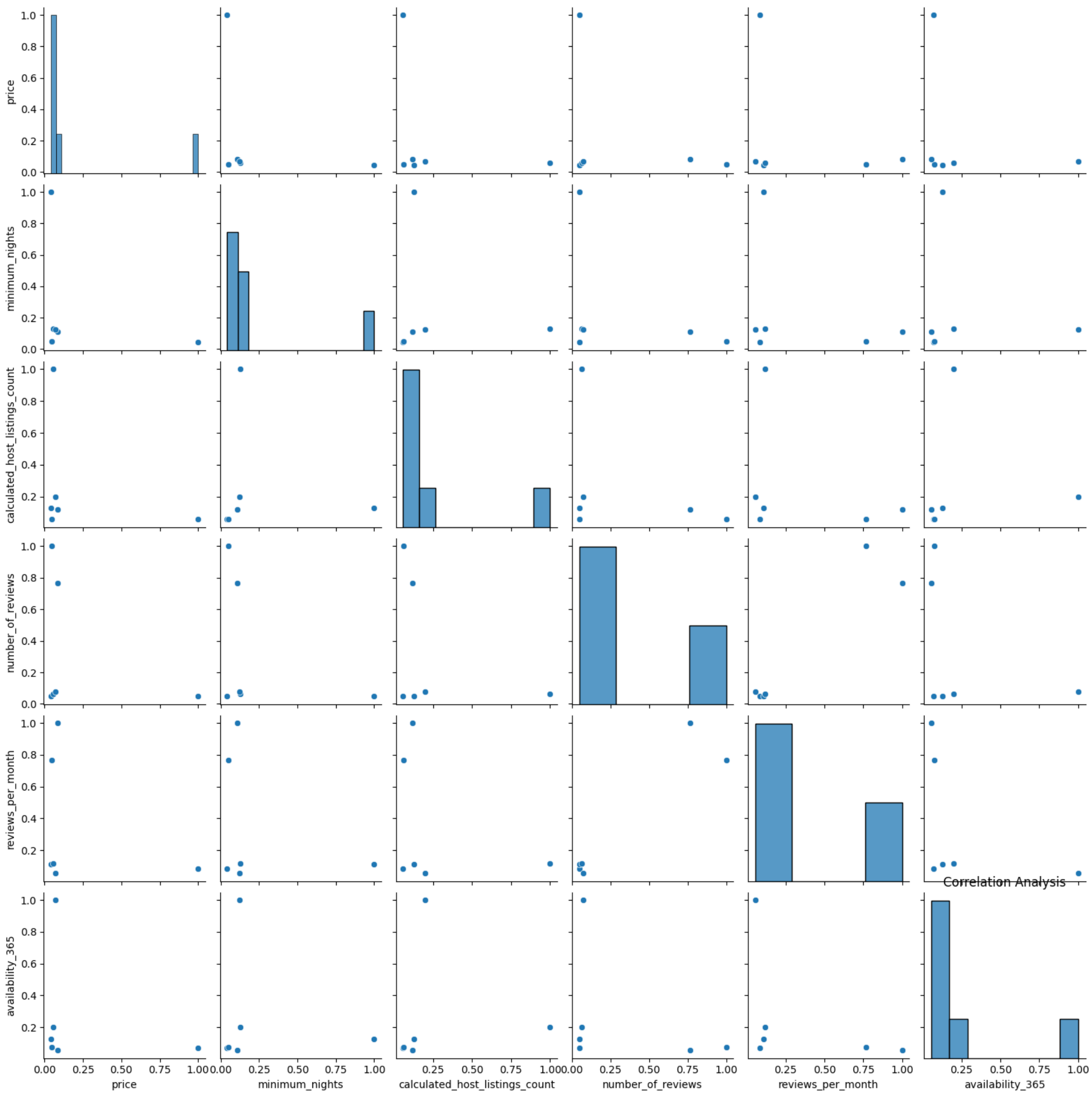


- Following understanding can be gained from above visualisation:

1. **Entire home** ceases to secure maximum **reviews % per month** which is **50.5%.**
2. Reviews per month % share for **Private room** is **46.4%.**
3. **Shared room** tends to gain the least reviews per month i.e., **3.3%.**

# Chart 20

The motive of selecting pair plot is to get a summarized insight of relationship shared between different type of columns in the dataset.



- Following analysis can be drawn from the above data:

1. Out of all the columns we have extracted out the ones which are numeric in nature and add some meaningful data to the analysis.
2. The relationship charts give clear highlights of how each column in the dataset shares a relation with another.
3. The peaks in the chart signifies a strong relation when a specified column intersects with itself.

Solution to Business Objective

- We have understood the dataset with proper conviction and dedication, tried to speculate every fruitful information from the dataset and reshaped it in form of charts.
- The charts conveys the important information that can simply be visualised and meaningful insight can be driven out of it, saving business time and complexity associated with analysing raw dataset.
- This EDA has been designed in such a way that the client just by reviewing all the charts and the comments can get an understanding about the complete dataset and need not to refer here and there to know about the dataset.

Conclusion

- There are certain key insights that can be summed up which may help business in a constructive way:

1. Amongst all the neighbourhood_group **Manhattan** has gathered **highest occupancy** i.e., **21661** times while **Staten Island** being the **least occupied** i.e., **373** times. The sign clearly points out that Manhattan is doing exceptionally well while Staten Island needs special attention.
2. **Michael** has the highest number of property ownership i.e, **417** followed by **David *i.e., *403** and **sonder** 327 respectively.
3. The room_type distribution can be seen amongst the three i.e., **Entire home** with **52% share**, followed **by Private room** with **45.7% share** and **Shared room** with **2.4% share**.
4. A **univariate boxplot** that depicts how **reviews** are shared **per month**, giving **minimum review** count to be **0.01**, **median** to be **1.22**, mean value being **32.48** and maximum as **152**.
5. We have a unique neighbours count accounting for **221** out of which **Williamsburg** has been visited the most i.e., **3920** times, followed by **Bedford-Stuyvesant** i.e., **3714**.
6. A **scatterplot** with **room_type analysis** based on reviews_per_month vs number_of_reviews. **Entire home *room type is *most liked** property. **Private room** is the next **most liked** property from a customer's view point. **Shared room** is the least **liked** room type.
7. Another scatterplot depicting the footfall of occupancy in **neighbourhood_group** based on **latitude** vs **longitude** scale. **Brooklyn ** and **Manhattan** is the most dense neighborhood_group, followed by **Queens**.

8. We have a barplot signifying property names in demand by customers. **Hillside Hotel** is **highest** in number i.e, with count **18,** which clearly gives an indication that it is the most popular property amongst customers. Followed by Hillside Hotel **Home away from home** shows maximum popularity.
9. **NYC hosting** can be seen from histogram. The listing is **maximum** in the range **0-35** with alone constituting for more than **44000.** From **35 - 135** a very **small chunk** of listing can be noticed. A small block of listing in range **235-260** and then **290-327** can be rectified.
10. The **count chart** portrays how price distribution helps in identifying different types of room. For ex. **Affordable property** is **pretty huge** in number. **Cheap property** is **less** than affordable but slightly more than expensive. **Expensive property** is **very less** in number.
11. Percentage distribution of **Nights spent** as per room type can be justified by a pie chart. **Nights spent** by customers maximum in the case of the entire **room type** constitute for nearly **62.9%. Private room** type has a share of **34.9%** in terms of nights spent. **Nights spent** is **least** for shared room type with merely **2.2%.**
12. **Scatter** chart plot narrates, number of reviews vs price distribution analysis for particular room type. **Private rooms** have a **very less** price range and on the other hand have **maximum no.of reviews**. **No.of reviews** for **Entire home** is comparatively **less** than that of private room type moreover **price** is also **high**. For **Shared room** price is pretty less but **reviews** are also **quite negligible**.
13. Room type vs neighbourhood group exploration can be seen from a barchart. **Private room** type is **most** in **Brooklyn** followed by Manhattan but **least** in **Staten Island**. Overall shared room distribution is quite less for all neighbourhood groups. **Entire home**

types are **very much** prevalent in **Manhattan** followed by **Brooklyn** with **least** being in **Staten Island**.

14. Histogram depicts price analysis in terms of count. It simply defines how the price range varies. Price range between** 1 to 1000** is pretty** huge** in number. Above **1000** we can see a small section of the price range.
15. **Longitude vs latitude** scale shows us the night's distribution. It defines that **nights spent** above **500** are **very less** in number. **Maximum strength** can be sited below **250**.
16. **Segregated price** distribution gives an overview of **top 10** properties with most nights spent. Property priced at **100** is **maximum** in number with average count **2051**. Price of **150** is **2047** in count making it second highest in number.
17. Reviews per month vs room_type insight can be driven from pie plot. **Entire home** ceases to secure maximum **reviews % per month** which is **50.5%**. Reviews per month % share for **Private room** is **46.4%**. **Shared room** tend to gain the least reviews per month i.e., **3.3%**.
18. To give a clear insight of how each and every series in the data set are linked to each other, how collaborative are the bonds and how loose they are a heatmap has been formulated. The one's with **lightest color** and indicated with numeric **value 1** shares the **strongest bond** i.e., the series when compared with itself serves the deep collaboration. As the **color** keeps darkening, the **strength** of the **bonds** keeps on **deteriorating** and the numeric values justifies it all. The series who's **joint venture** depicts the **darkest color** and have **least numeric** value claims to have the **weakest bond**.
19. Eventually a pair plot gives a summarized insight of the relationship shared between different types of columns in the dataset. Out of all the columns we have extracted out the **ones** which are **numeric in**

nature and add some meaningful data to the analysis. The relationship charts give clear highlights of how each column in the dataset shares a relation with another. The **peaks** in the chart signifies a **strong relation** when a specified column intersects with itself.

20. Overall analysis predicts the distribution and relationship that each and every data in incorporated with.