A

Minor Project Report

On

"Airbnb Booking Analysis "

Submitted in partial fulfilment for the award of the degree of

Masters of Computer Application(MCA)

from

Dr. A.P.J. Abdul Kalam Technical University,Lucknow

(Session: 2025-26)

Dewan V.S. Institute of Engineering & Technology, Meerut

Submitted To:

Mr.

HOD, BCA Deptt.

Submitted By:

Vishu Rajput

MCA 3rd Sem

Roll No: 240311014005

Page No.

TABLE OF CONTENTS

Content

DECLARATION

I, **Vishu Rajput**, bearing roll number **2403110140058**, hereby declare that the work which is being presented in the Minor Project, entitled "**Air bnb Booking Analysis**" in partial fulfilment for award of Degree of

"Masters of Computer Applications" in Department of Computer Application is submitted under the Guidance of "**Dr. Arif Md. Sattar**".

I have not submitted the matter presented in this work anywhere for the award of any other Degree.

Date: .....

Student Name: Vishu Rajput

Roll No.: 2**403110140058**

CERTIFICATE

Certified that the Project Report entitled "**Air bnb Booking Analysis**" submitted by **Vishu Rajput** bearing roll no. 2**403110140058** in partial fulfilment of the requirements for the award of the degree of Masters of Computer Applications is a record of the student's own work carried out under my supervision and guidance. To the best of our knowledge, this Minor Project work has not been submitted anywhere for the award of any other Degree.

It is further understood that by this certificate the undersigned does not endorser approve of any statement made, opinion expressed or conclusion drawn therein but approve Minor Project for the purpose for which it is submitted.

Dr. Arif Md. Sattar

(Project-Guide) (HOD, BCA Deptt.)

ACKNOWLEDGEMENT

Many people have supported me, in different ways, during the work with the minor project. I'd like to thank my guide **Dr. Arif Md. Sattar** & HOD *for their kind and active support and valuable guidance during the work process. My family has, as always, offered me their unconditional support, during my efforts in completing this Minor Project.

However, it would not have been possible without the kind support of many individuals and institution.

I would like to extend my sincere thanks to each and every members related to DVSIET

Vishu Rajput

MCA 3rd Sem

Roll No. 2**403110140058**

Abstract: -

Airbnb is a trusted community marketplace for people to browse, discover, and book unique accommodations and other hospitality services around the world. Its objective is to create a world where anyone can belong anywhere. Millions of listings were tabulated by the host from Airbnb and these millions of listings had generated a lot of data. To make sense of the enormously growing data, one of the crucial divisions for making sense out of the numbers and variables is the data analytics division of the company. Our analysis can help understand the reason for the success of Airbnb based on boroughs, by studying the reasons behind variations based on boroughs by prices (which area is expensive), reviews (which area is the rated best), and the varied type of accommodation that they provide. Ascertain the busiest hosts and determine the reason for their popularity which is done with the help of data exploration, data analysis, and data visualization.

KEYWORDS:

Data analysis, data visualization, boroughs, prices, reviews, hosts, listings.

Problem Statement:-

Airbnb is an online marketplace that connects people who want to rent out their properties to the people who are looking for hospitality facilities in that locality. It currently covers more than 100,000 cities in over 220 countries worldwide. There's a lot of competition, resulting in the occasional race to the bottom. It may rent more frequently at economic rates leading to lesser profit per stay. In this world of competition, we can predict in which country a new user will make his /her first booking and is there any noticeable difference of traffic among different areas and decipher the reason behind it and who among the hosts are in most demand hence being the busiest among the rest and what makes them the choice of the clients. These are a few of the objectives that we are aiming to analyze.

Introduction:

Airbnb is an online marketplace and hospitality service, aiding users to rent or lease accommodation not limited to bed and breakfasts, hostels, homestays, apartments, rooms, or hotels. Airbnb does not own any of the properties but collects brokerage fee and service fee percentages from both the host and the guest per booking. Airbnb was established in August of 2008 and founded in San Francisco, California. Accommodations located all over the world can be booked online using a tablet, mobile phone, PC, or Mac ("About Us – Airbnb", 2017). Joe Gebbia was having difficulties with affording their rent and as a way to make extra money, they started the business by purchasing inflatable airbeds, placing them on the floor of their apartment, renting them out to guests for $80 per night which included breakfast. Chesky and Gebbia teamed up with friend Nathan Blecharczyk to assist in progressing the business growth. Airbnb claims to have over 160,000,000 guests using it in 191 countries, 65,000 cities, have 1,400 castles listed, and over 3,000,000 properties listed.

Reason for Airbnb Pricing:

Airbnb pricing strategy is a key player in its business model. Airbnb uses a dynamic pricing strategy based on the trend, demand, and availability. The direct impact of Airbnb pricing is the number of bookings they receive. Pricing is probably the first detail that a potential guest will notice. It can be a critical deciding factor for an end-user to choose or reject a deal.

Reasons for Airbnb Pricing is:

1. Type of Property:

2. Type of Location:

3. Competitors' Price:

To get a clear picture of what the price per night should look like one needs to find similar listings in a local area. Putting the same price as competitor's does not make any sense, though we can learn a lot from it

4. Weekends: The weekend is a great reason to raise the price for a rental by 10-15 percent. Monday to Thursday are the cheapest days of the week because of lower occupancy. However, most weekends are busier than regular weekdays.

5. High Seasons: Prices for a property rental also have to change depending on the region's high and low seasons. Increase the price during indemand times. Based on our experience we found out, that the rate for an entire month is equal to the rate for just one week in peak season.

6. Special Events: Make a separate calendar for popular events in the area and make sure one changes the price well in advance. Such events like concerts, conferences, sports events announce their event date months beforehand, which helps businesses like Airbnb to restructure their prices accordingly.

7. Special Amenities: Data from the Airbnb survey says that about 97% of Americans care about the amenities at the property. The majority of countries ranked pools as a top amenity. Pet-friendly and property with free parking follow closely behind. Guests also love the functionality and service provided in an amenity.

How Airbnb Works:

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. Airbnb takes a 3% commission from bookings as well as a 6%-12% servicing fee from guests. How Airbnb works with hosts? If one has a room to rent (or an entire house), then they will have to register your listing on Airbnb and outline certain specifications like area, size, number of rooms, and other details on your listing How Does Airbnb Work for Guests? As a guest, you can go on the Airbnb website and search through dozens of filters like area (city, country), space (if you need a room or house), and several guests to accommodate. How Do Airbnb Payments Work? To streamline the process, Airbnb handles payments and accepts most major credit cards including MasterCard (MA) Airbnb Fees Airbnb takes 6-12% of guest servicing fees on top of your payment. Additionally, the site takes a 3% commission as well.

Steps Involved

EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset. EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better.

The steps we perform are below:

1. Acquiring and loading data: Data acquisition has been understood as the process of gathering, filtering, and cleaning data before the data is put in a data warehouse or any other storage solution.
   1(a) Load python libraries For data analysis, we use NumPy and Pandas. For Visualization, we use matplotlib and seaborn.
   1(b) Load Dataset we have to use the pandas library to load the dataset and to read the CSV file of the Airbnb
   1(c) understanding Data to understand our csv data we to use many techniques, like-info (), describe()..
   1(d) cleaning data - the dataset we use has many error values like null Values, missing values, outliers, etc. So we have to clean them for a better analyz e our data.
   1(e) Exploration and visualization of data - we explore data by visualizing and analy zing the data features by understanding the main concept of each attribute and what will be their result.

Challenges faced

1. The very first challenge that we faced while starting with the project was mapping out our procedures.
2. Identifying the keen difference between “Neighborhood” and “Neighborhood group” was challenging and required a few more searches on google.
3. Interpreting the column “Host listings” took some time and identifying the correct relationship between this column and other columns was another pondering task.
4. Marking the appropriate threshold for removing outliers from the given data was a challenging and anticipating task.

Conclusion

1. "Manhattan" has the most expensive bookings We can say this based on the neighbourhood vs listings and neighbourhood vs price.
2. Manhattan is also considered as the best location based on the graph of neighbourhood group vs several reviews.
3. Busiest Host = Based on the different graphs like neighbourhood vs price, neighbourhood group vs Airbnb listings, and neighbourhood group vs several reviews, the host (host id = 219517861 and his name = Sonder NYC) who has 327 listings is considered as the busiest host in NYC and he belongs to the Manhattan.

]

Data Exploration

After importing the dataset, this is what I am faced with in terms of fields:

☑	Type	Field Name	Original Field Name	Changes	Preview
☑	#	id	id		2,539, 2,595, 3,647
☑	Abc	name	name		Clean & quiet apt home by the park, Skylit Midtown Castle, THE VILLAGE OF HARLEM....NEW YORK !
☑	#	host_id	host_id		2,787, 2,845, 4,632
☑	Abc	host_name	host_name		John, Jennifer, Elisabeth
☑	Abc	neighbourhood_...	neighbourhood_group		Brooklyn, Manhattan
☑	Abc	neighbourhood	neighbourhood		Kensington, Midtown, Harlem
☑	#	latitude	latitude		40.64749, 40.75362, 40.80902
☑	#	longitude	longitude		-73.97237, -73.98377, -73.9419
☑	Abc	room_type	room_type		Private room, Entire home/apt
☑	#	price	price		149, 225, 150
☑	#	minimum_nights	minimum_nights		1, 3
☑	#	number_of_revi...	number_of_reviews		9, 45, 0
☑	📅	last_review	last_review		10/19/2018, 05/21/2019, null
☑	#	reviews_per_m...	reviews_per_month		0.21, 0.38, null
☑	#	calculated_host...	calculated_host_listings_...		6, 2, 1
☑	#	availability_365	availability_365		365, 355

As it can be observed, I have information on the hosts, geographical information and metrics such as price, number of reviews and other related fields.

Next, I will add a *clean step* to my flow to get better descriptions of my fields. From this, I can see that this dataset is already looking clean but there is room for further preparation. For example some *null* values can be observed. In the "*last_review*" and "*reviews_per_month*" columns, *null* values take ~20% of the total values. So I should **filter** them.

Quick tip:

1. Right click the *null* bar of the "*last_review*" in the Profile Pane.
2. Click Exclude.

Also in the same field, the date ranges from 2011 to 2020, 65% of the data is between 2019 and 2020 therefore, I will **filter** out the dates before 2019. Aside from that, I also noticed some potential outliers across fields, however I will leave them in.

Now that I am done with my preparation, I can put an output block on my flow and save the updated version to be used in Tableau Desktop for the data analysis.

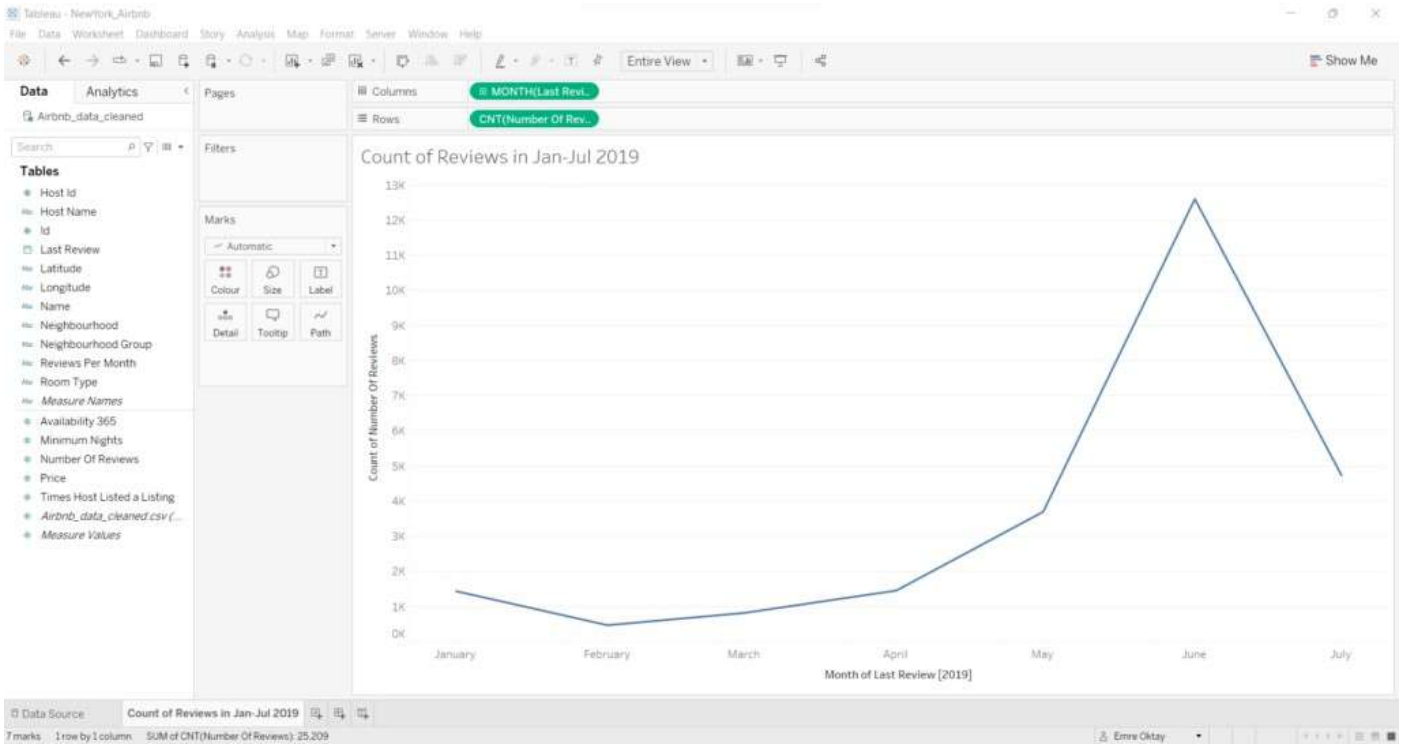
Data Analysis

I am lucky enough that our dataset contains a good range of data types. I have:

- **Date field** = I can create **line plot** to see changes in activity
- **Latitude / Longitude** = I can create a **map**s.
- **Borough names** = Tableau has the ability to recognize certain area descriptions (more predominantly in the US), can be used within the **map**).
- Room Type, Price, Number of Reviews = I can **categorize**, do **ranking**.

Line Plot

First I am going to get started working with the date data. I want to see the changes in the number of reviews for all listings over the months within 2019. For this, I will plot a line graph as follows:

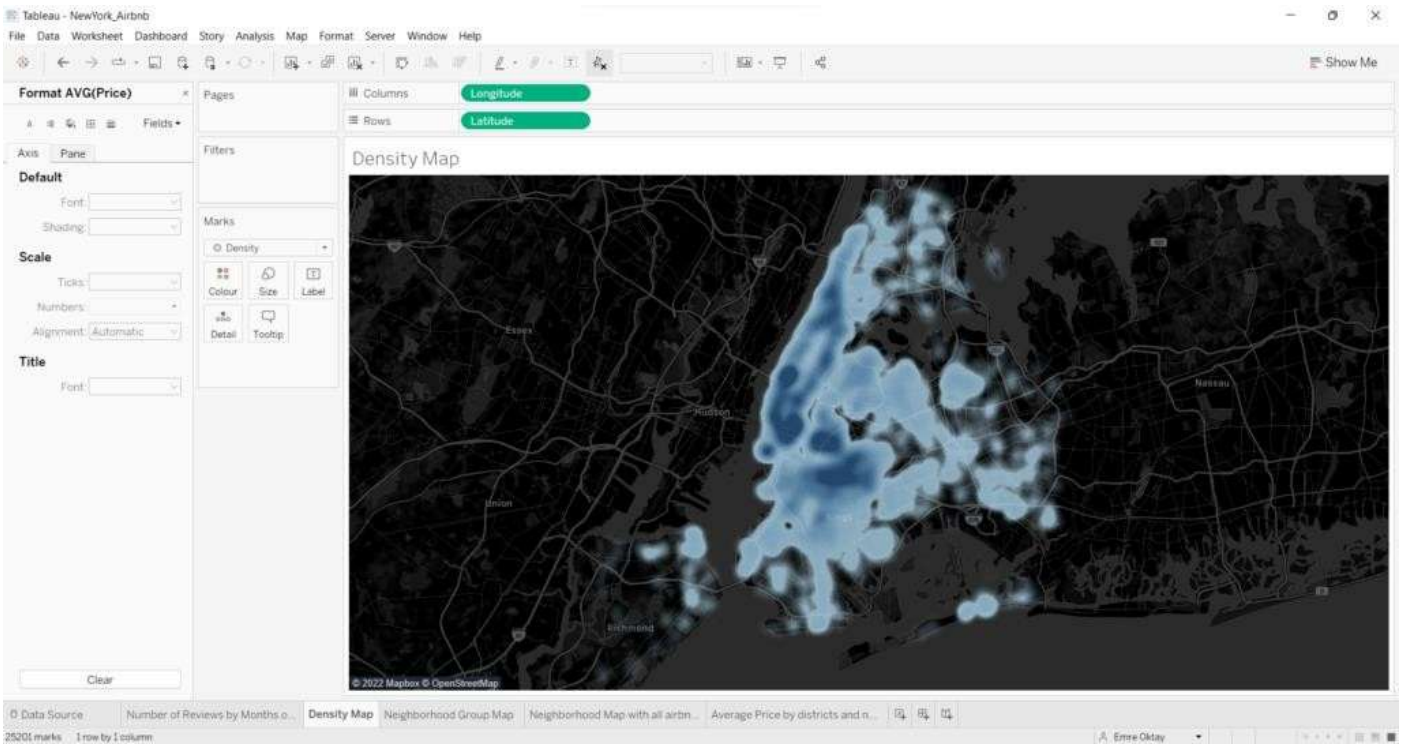


<u>*QUICK OBSERVATION:*</u>

As it can be observed, there is a sharp increase in reviews from May (~3000) to June (~12000) with a sharp decrease by July (~4700). This can be due to the summer season coming up in the year and the influx of tourists staying in New York through the summer. Since the dataset ends on July, we cannot see how the trend is for the end of the year.

Creating Maps

Next, I can create a map with all the geolocations of the listings provided in the dataset. Since I have the specific locations for each listing, I also have the option of creating a **density map**. The neighborhood classification for the listings also will allow us to create a **map with borders**. As for the density map, it is fairly straight forward to create. But for it to look nicer, some map formatting might be beneficial. This is the final look of the **density map**:



I have chosen a dark theme for the look of the map since I believe that the density maps transparency works best in this format where city/district names are more readable. You can play with the opacity and size settings with the Color and Size buttons in the Marks Card to your liking, low opacity means more readability.

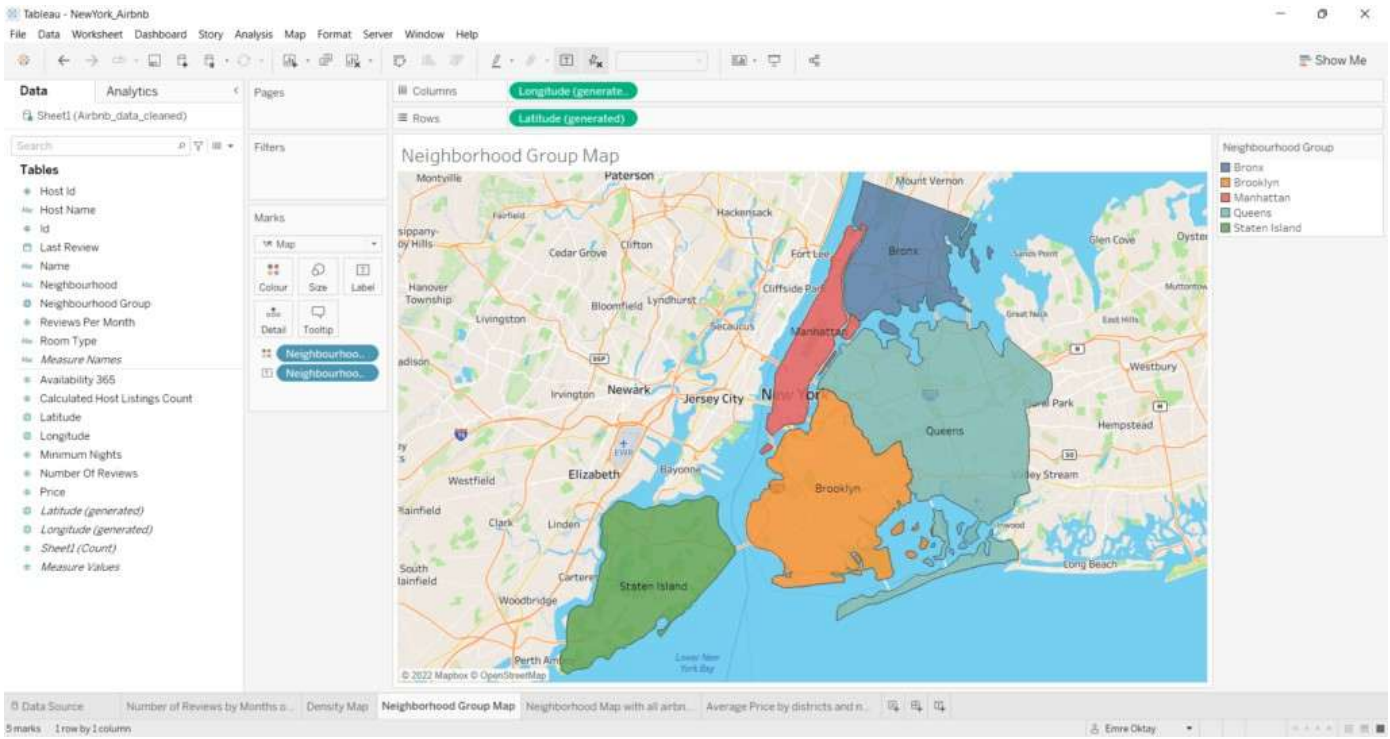
Quick Tip:

1. Place *Latitude* and *Longitude* in the columns and rows and convert them to ***dimensions***.
2. Select “density” option in the Marks card.
3. Go to “Map” –> “Background Maps” and select “Dark”.
4. Go to “Map” –> “Map Layers” in the toolbar at the top, check “Streets, Highways, Routes” and “County Names”.

QUICK OBSERVATION:

From the density map, it can be seen that most of the listings exists in the Manhattan area of New York, particularly the southern parts. Another very dense area is upper side of Brooklyn. The least listings exist in the Staten Island area.

On top of the density map, I think it is a good idea to create a **Neighborhood Group / Borough Map** where the boroughs show clear boundaries. This could be used for filtering purposes in our dashboard at the end. This is the look of my Borough Map that I have created:

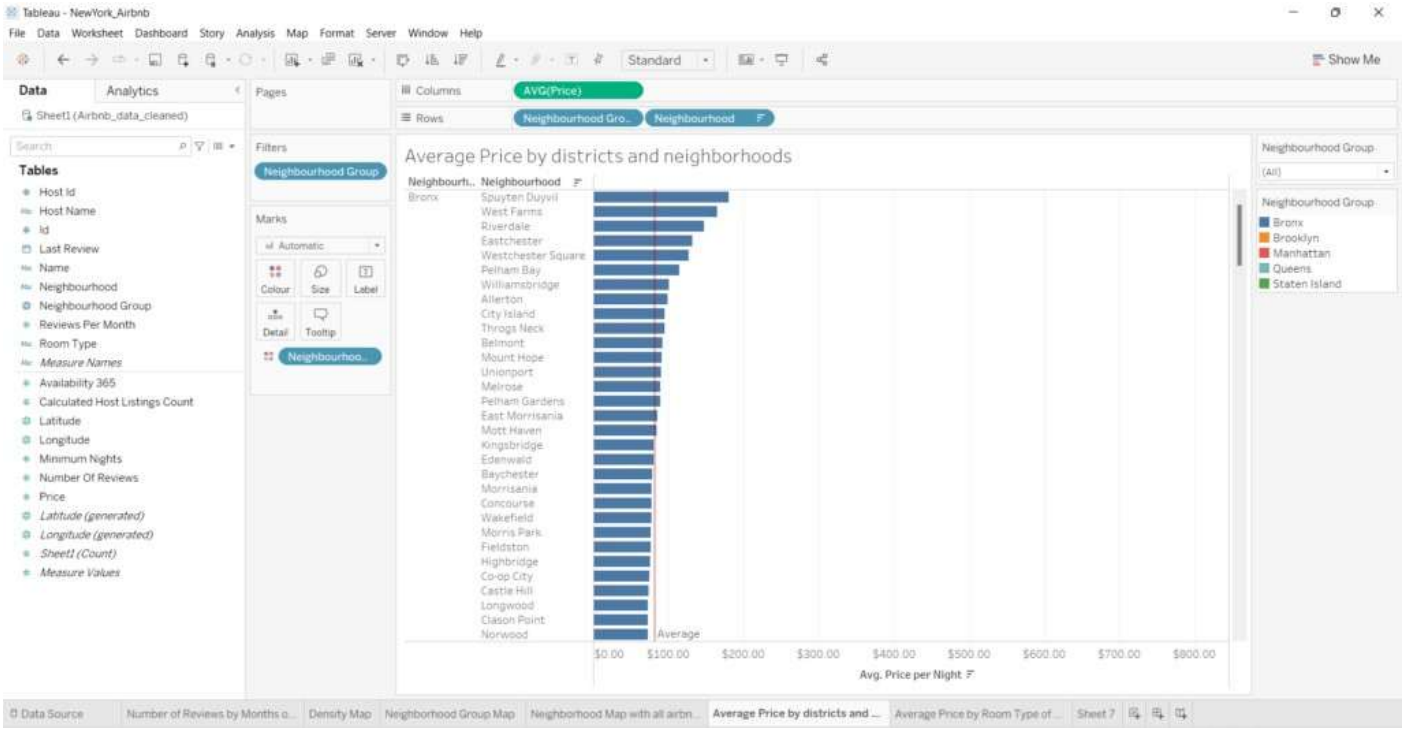


For this I used a different map background of "*Streets*". The Borough Map includes five distinct values therefore I can use color to categorize. Another important point is that the boroughs need to be assigned a geographical role. This way, Tableau can automatically draw shapes based on the borders of the five neighborhood groups of New York. The *latitude* and *longitude* data used in the view is generated from assigning the geographical role of the "Neighborhood Group" field.

Bar Charts

After I created my maps, having **bar charts** for comparison purposes is a good addition to the final dashboard. This is where users can really educate themselves on what are the average price ranges for each room type and smaller neighborhoods within the boroughs.

I will create two separate sheets for supplying the information. First. I want to plot the average prices of each Airbnb location broken down by the individual neighborhoods that exist in a borough. This is my resulting **bar chart** for comparing prices within boroughs:



I have filtered and color coded the boroughs in this sheet as well so it matches the colors of the Borough Map. The filter allows the user to select desired borough or all. In addition, I have also added an **Average Line** for each borough so that comparing average prices of specific neighborhoods with the average price per borough is easier.

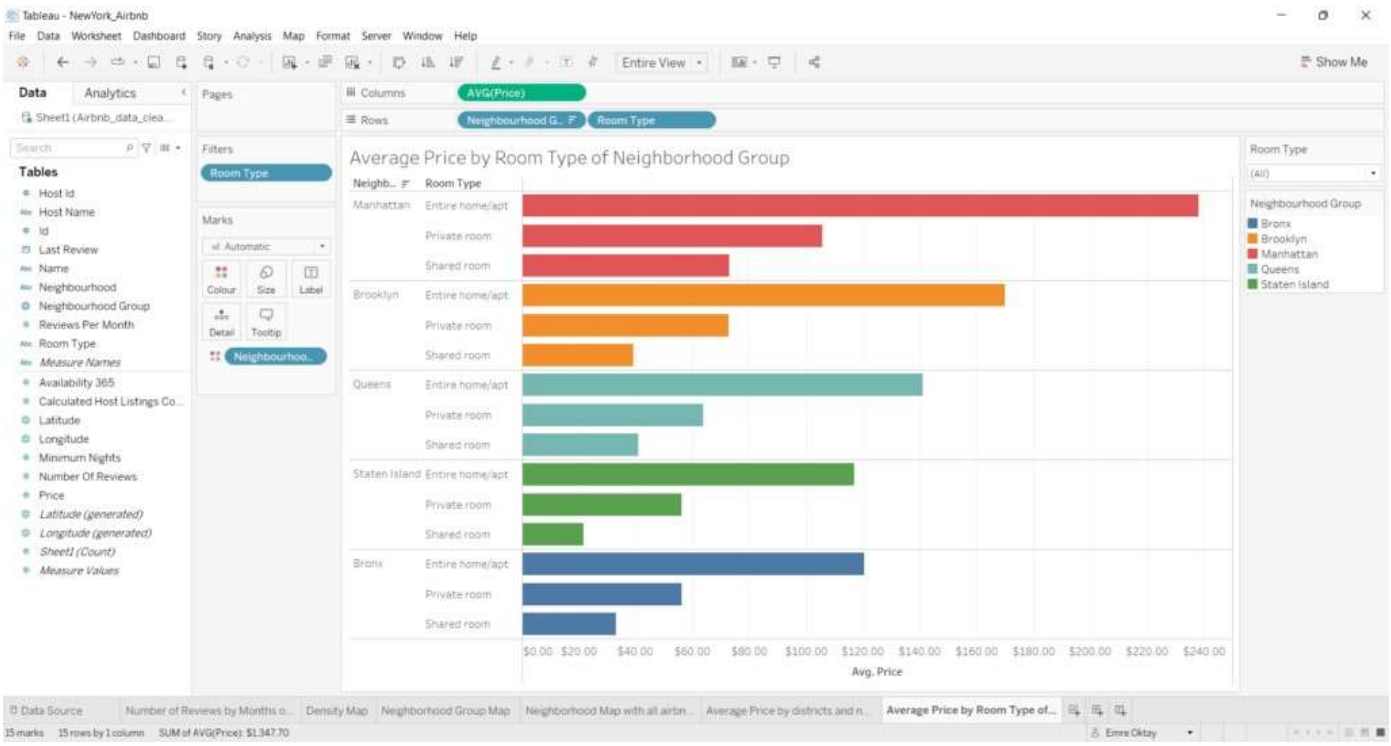
Quick Tip:

1. Click the Analytics Pane
2. Drag and Drop “Average Line” into the view.
3. Make sure you do ‘*by pan*e’.

<u>QUICK OBSERVATION:</u>

Manhattan area has the highest price average ($197.98) with Bronx having the lowest ($81.73).

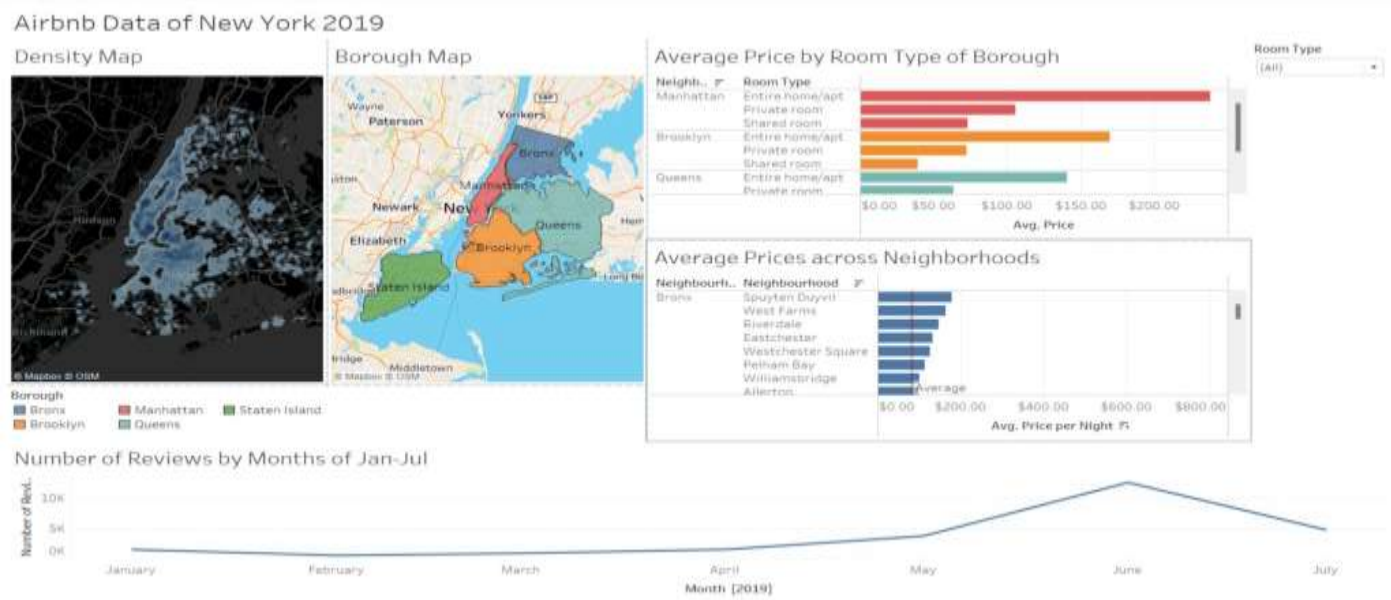
I will proceed to creating a very similar view. Only this time, I want to compare average prices for each room type within each borough. Here is the resulting bar chart:



The boroughs are color-coded again for categorization. As for the filter, the user can select the desired room type or all to do comparisons between the boroughs.

The Dashboard

I believe I have made all the charts I need to create an informative dashboard from the maps and charts I have created about the Airbnb data of New York. For this, I make a straight forward dashboard without dazzling visual elements. Just a simple environment to communicate my findings. Here is a screenshot of my resulting dashboard and the link to the dashboard in my Tableau Public account:



The borough map is used as a **filter** in the dashboard to only show results of the selected borough across the charts. It displays all if none are selected.

This dashboard incorporates every view I have created. It outlines which areas of New York is more densely populated with Airbnb locations, displays the number of reviews done for each listing in the year of 2019 and describes and compares average prices for room types between boroughs and compares average prices of listings made in each neighborhood of the boroughs.

CODING

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one of a kind service that is used and recognized by the whole world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

Explore and analyze the data to discover key understandings (not limited to these) such as :

- What can we learn about different hosts and areas?
- What can we learn from predictions? (ex: locations, prices, reviews, etc)
- Which hosts are the busiest and why?
- Is there any noticeable difference of traffic among different areas and what could be the reason for it?

Firstly, we will mount the drive and download the dataset.

```
# Mount the google drive in google colab.
from google.colab import drive drive.mount('/content/drive')
```

Now, Let's begin our data analysis by loading the Python packages/libraries.

```
# Import the main libraries
import pandas as pd import
numpy as np import
matplotlib.pyplot as plt import
seaborn as sns import
plotly.express as px
```

Defining the path in which the dataset is present. To do so, we have to use pd.read_csv() function which imports the CSV file into DataFrame.

```
df=pd.read_csv('/content/drive/MyDrive/AlmaBetter/Alma Projects/Airbnb booking anal
ysis-Tanu Rajput/Airbnb NYC 2019.csv')
 df.shape
df.head()
```

Getting/Viewing the information of the dataset.

```
# We can get the information of dataset by using "df.info()" function.
 df.info()
```

observation-1:----------

--> We get to see the different types of data variables, its datatypes present in the data set. We can also get the overview of null values roughly.

--> Also it shows the size of the dataset we are working with. There are *16 columns* ***(mix of categorical and numeric data) and*** *48895 rows.*

--> We get to see the different types of data variables, its datatypes present in the data set. We can also get the overview of null values roughly.

--> Also it shows the size of the dataset we are working with. There are *16 columns* ***(mix of categorical and numeric data) and*** *48895 rows.*

Using .describe() function to see some statistical information of the data

```
df.describe()
```

Cleaning the Data

In .info() cell we can see the nullvalues, but we want the exact number of null values present in each column.

So, ".isnull().sum()" function will show us the total number of nulls(NaN) in each column of dataset.

```
#looking forward to find out first that which columns have null values
df.isnull().sum()
```

Here we can see that,we have 2 columns which have more than 10000 null values. Having so many null values is not appropriate for exploring data or to analyse of any form. So, we have to clean the data. We observe that null values are present in irrelevant columns so we directly clean the data by dropping certain columns that is not needed for analysis i.e. 'name' ,'Host_name', 'last_reviews','latitude','longitude'....

Dropping the irrelevant columns.

```
#drop unwanted columns...
df.drop(['host_name','name','latitude','longitude','last_review'], axis = 1,inplace
= True)
```

Replacing all null values in "reviews_per_month" column with zero

```
# replacing all null values in review_per_month with 0
df.fillna({'reviews_per_month':0,}, inplace=True)

# Verifying the changes df.isnull().sum()
```

Detecting Outliers in the numerical dataset.

```
# lets observe normally the outliers or extreme values in statistical description
df.describe()
```

checking outliers of numerical columns using seaborn boxplot

```
columns = [ 'price', 'minimum_nights', 'number_of_reviews', 'calculated_host_listin
gs_count', 'availability_365'] n = 1 plt.figure(figsize=(20,15))
 for column in
columns:
plt.subplot(4,4,n)
  n = n+1
sns.boxplot(df[column])
plt.tight_layout()   plt.show()
```

--> We can clearly tell that,columns, viz., price, minimum_nights, calculated_host_listings_count has outliers or extreme values.

--> And, for availability_365 there is no single outlier in the column.

Handling Data..

We can handle the data by removing outliers or setting proper limit!.

(a) Removing outliers for "Price" column

For removing Outliers ,we are using Quantile method which is effective!

```
high_limit = df['price'].quantile(0.99986) print(high_limit)
low_limit = df['price'].quantile(0.0015) print(low_limit) new_df =
df[(df['price'] < high_limit) & (df['price'] > low_limit)] new_df


sns.boxplot(new_df['price']) plt.show()
```

(b) Removing Outliers for "minimum_nights" column

For removing Outliers ,we are using Quantile method which is effective!

```
high_min_limit = new_df['minimum_nights'].quantile(0.9999)
print(high_min_limit) low_min_limit =
new_df['minimum_nights'].quantile(0.0)
print(low_min_limit)
new_df1 = new_df[new_df['minimum_nights'] < high_min_limit]  new_df1


sns.boxplot(new_df1['minimum_nights']) plt.show()
```

After **cleaning data** we have **48796 rows and 11 columns**

```
#let's see the description again new_df1.describe()
```

Now, lets **explore**, **analyze** and **vizualize** the **cleaned dataset** and get some insights out of it...

```
# First, Let us see the first five rows of the "CLEANED DATASET" using '.head()' fu
nction     head = new_df1.head() head



# For seeing the last five rows of the cleaned data, we use ".tail()" function
tail = new_df1.tail() tail


# statistical information. new_df1.describe()
```

Lets find out the uniqueness of categorical columns....

(i) Unique values of neighbourhood_group

```
print(df.neighbourhood_group.unique())
len(df.neighbourhood_group.unique())
```

(ii) Unique values of "neighbourhood"

```
df.neighbourhood.unique()
len(df.neighbourhood.unique())
```

(iii) Unique values of "room_type"

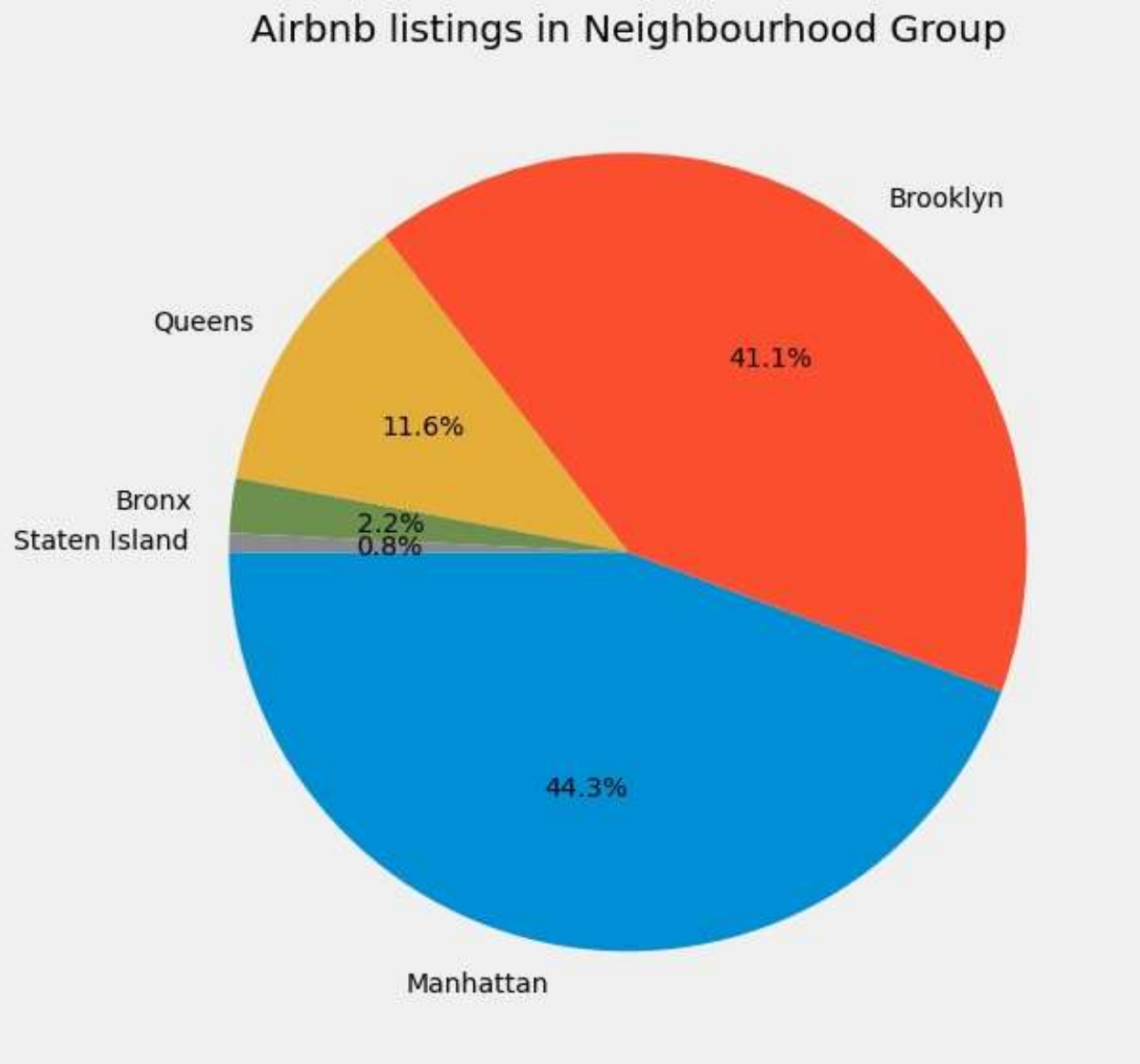
```
df.room_type.unique()
```

VISUALIZATION_____

1) Number of Airbnb listings in different neighbourhood_group

```
neigh_listings_counting = new_df1.neighbourhood_group.value_counts()
count_of_neighbourhood_group = pd.DataFrame(neigh_listings_counting)
count_of_neighbourhood_group.reset_index(inplace=True)
count_of_neighbourhood_group.rename(columns={'index':'neighbourhood_group', 'neighb
ourhood_group':'Listings_Count'}, inplace=True) print(count_of_neighbourhood_group)


# Visualization of different neighbourhood group by count
plt.style.use('fivethirtyeight') plt.figure(figsize=(15,9)) plt.title("Airbnb
listings in Neighbourhood Group") g =
plt.pie(new_df1.neighbourhood_group.value_counts(), labels=df.neighbourhood_gro
up.value_counts().index,autopct='%1.1f%%', startangle=180) plt.show()
```



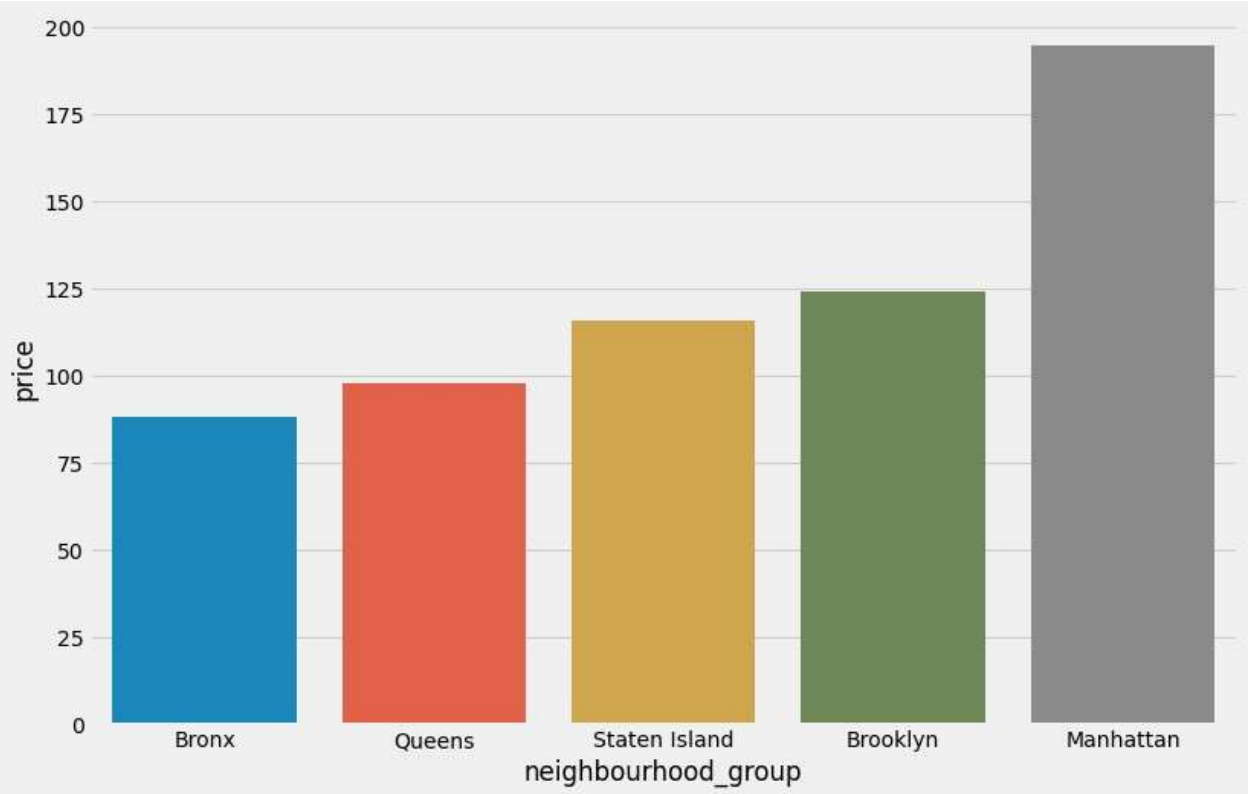
--> We can see that 'Manhattan' has most number of Airbnb listings/housings.

--> And "Staten island" has least number of Airbnb listings.

2) Average Price of Airbnb listings in different neighbourhood_group

```
avgprice_neigh_group = new_df1.groupby(["neighbourhood_group"])['price'].aggregate(
np.mean).reset_index().sort_values('price')
print(avgprice_neigh_group)
```

```
plt.figure(figsize=(12,8)) sns.barplot(x='neighbourhood_group', y = 'price',
data = avgprice_neigh_group) plt.show()
```



As we can see from the table and barplot,

--> Manhattan receives highest average price of $194.8 because of its highly demand. Its clearly shows that the highly rated neighbourhood_group(location) is to be costly maybe its constant supply, higher the demand, higher the price is!. Manhattan has the most expensive rentals compared to the other neighbourhood_group

--> Bronx receives lowest average price of $88.0

We saw average price of listings in neighbourhood_group

Now we will see the average price of listings for "neighbourhood"

```
avgprice_of_neigh =new_df1.groupby(["neighbourhood"])['price'].aggregate(np.mean).r
eset_index().sort_values('price') avgprice_of_neigh
```

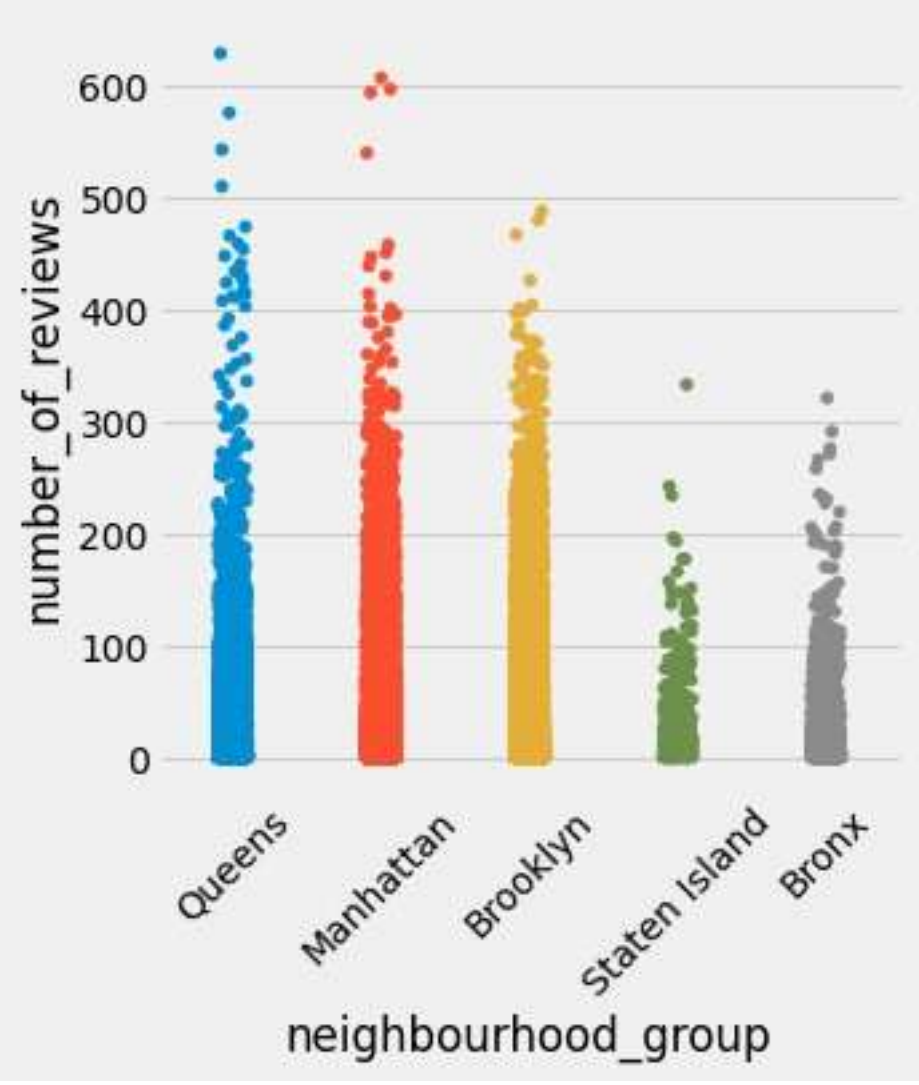
--> Fort Wadsworth have highest average price for listings which is $800

--> Bull's Head have cheapest price,i.e, $47

3) Lets see which neighbourhood_group contains the listings with most reviewed

```
#Lets check out which neighbourhood group contains the listings with most reviewed.
!
most_reviewed_property = new_df1.groupby(['id','neighbourhood_group'])['number_of_reviews'].mean().reset_index().sort_values('number_of_reviews',ascending=False,ignore_index=True)
most_reviewed_property
```

```
i=sns.catplot(x='neighbourhood_group',y = 'number_of_reviews', data = most_reviewed_property)
i.set_xticklabels( rotation=45)
plt.show()
```



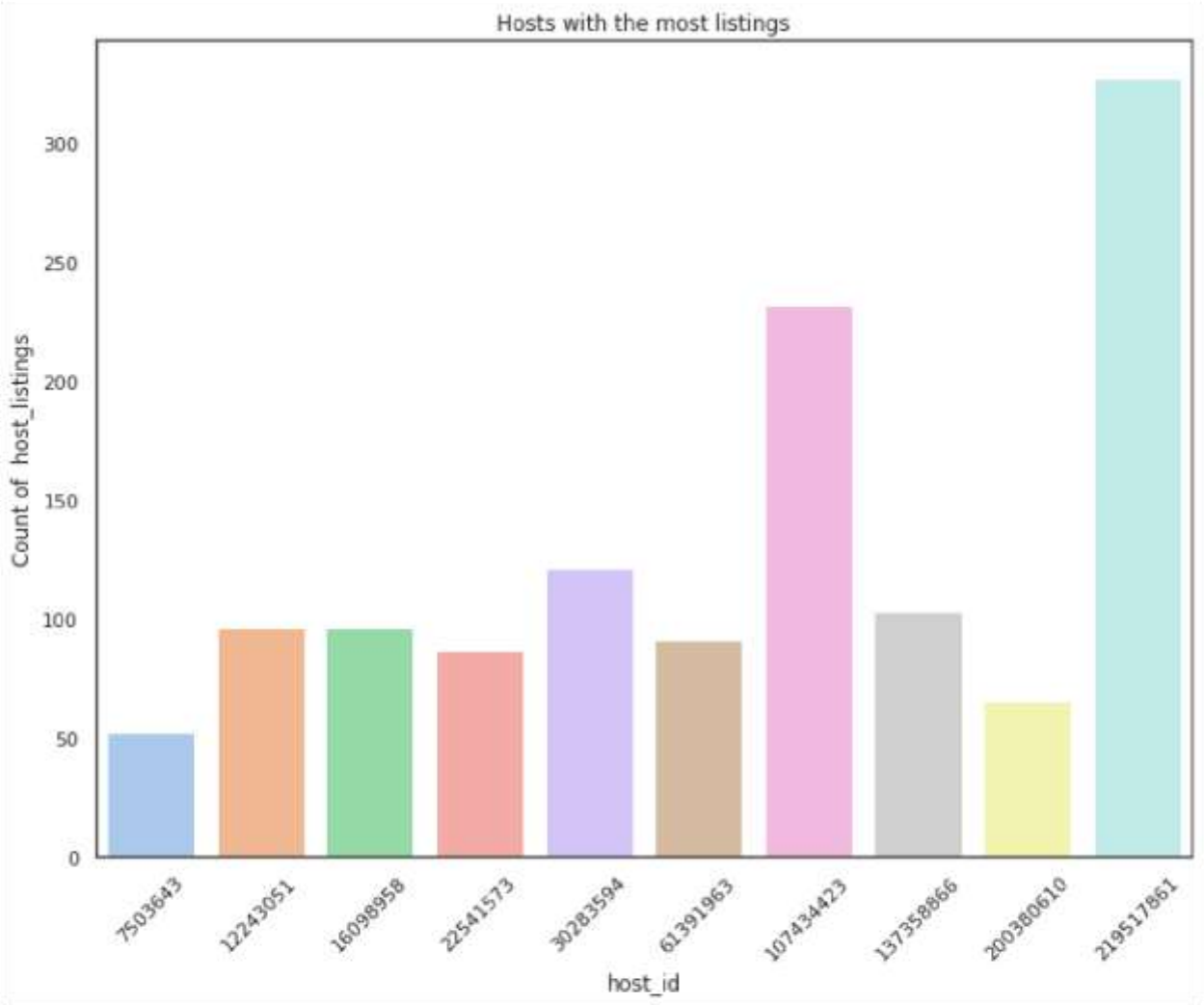
--> As we can see, Queens has the listings with most and high number of reviews followed by Manhattan.

-->Bronx has the listings with less number of reviews.

4) Which Host has the highest number of Airbnb listings?

```
# Let's check the hosts who have more number of Airbnb listings
top10_host=new_df1.host_id.value_counts().head(10)
top10_host_df=pd.DataFrame(top10_host) top10_host_df.reset_index(inplace=True)
top10_host_df.rename(columns={'index':'host_id', 'host_id':'Count'}, inplace=True)
top10_host_df
```

```
sns.set(rc={'figure.figsize':(10,8)})                    sns.set_style('white')
d_1=sns.barplot(x="host_id", y="Count", data=top10_host_df,palette='pastel')
d_1.set_title('Hosts with the most listings') d_1.set_ylabel('Count of
host_listings') d_1.set_xlabel('host_id')
d_1.set_xticklabels(d_1.get_xticklabels(), rotation=45)
```



--> We can see that there is a good distribution between top 10 hosts with the most listings.

Now, **Just out of curiousity!!!---Let's Know More**

```
# I want to find out that how much the average money(/ or total sum amt from his al
l listings) do the top host,i.e, host_id =219517861 earns?
x = new_df1.loc[new_df1['host_id']==219517861]
avg_money  = x['price'].mean()  # average money he earns!
print(avg_money)
sum_amt = x['price'].sum()  # total sum amount  from all his listings "if" all his
listings were booked for 1 night
print(sum_amt)  #--------(1)
```

```
#Now, we will do some MATH to calculate how much money did he get after deducting t
he 3% airbnb commission per listing booked!
```

```
commission = sum_amt*0.03
print(commission) #---------(2)



actual_amt = sum_amt - commission   #Subtracting (1) with (2).
print(actual_amt)
```

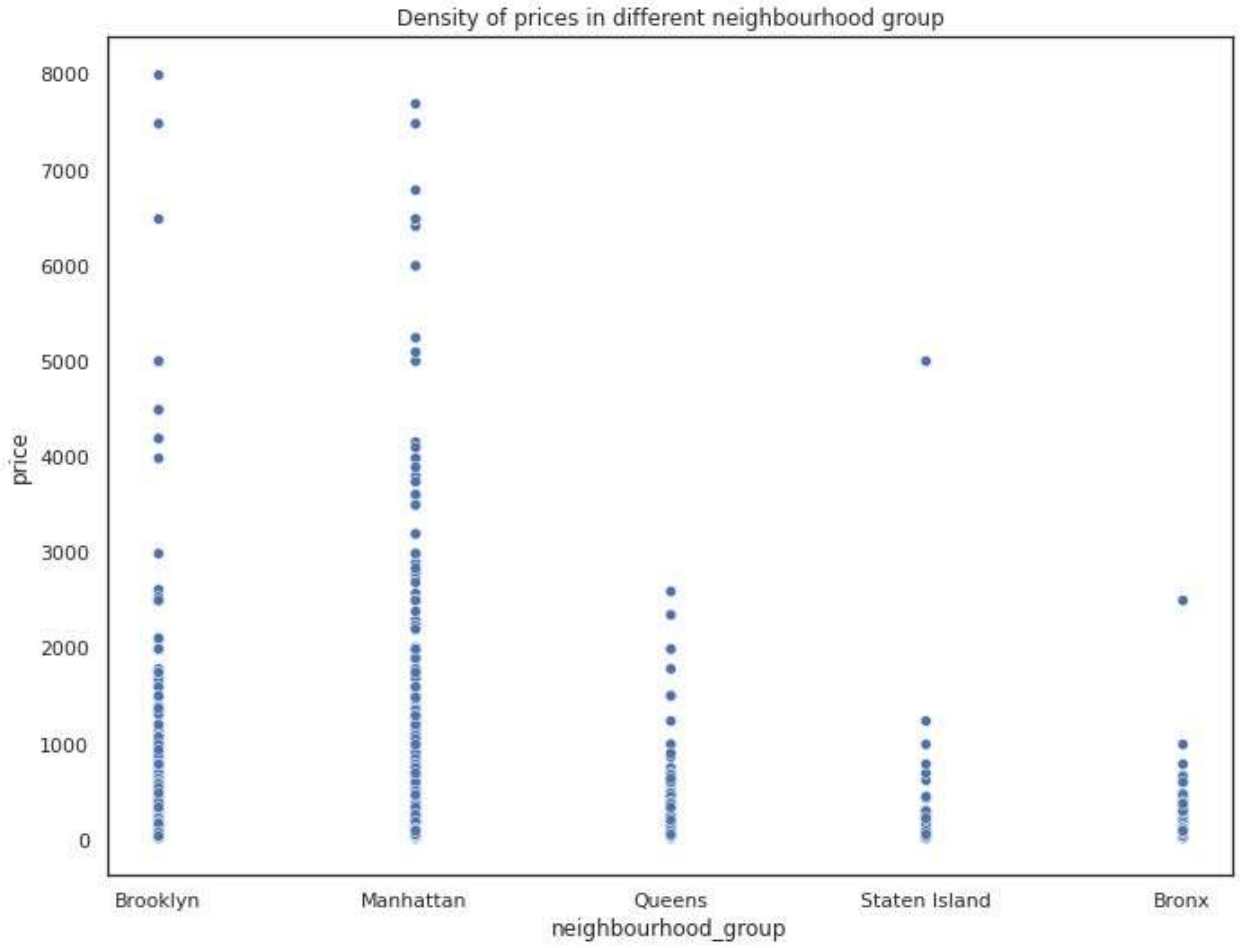
-->The host with host_id number = 219517861 who has 327 listings, the average money he earns with all his listings is $253.19

--> means he earns approx. total $82795 per 1 minimum_night if all his listings were booked!

--> And after deducting airbnb's commission, he get the actual amount which is $80311.15

5) Density of distribution of prices in different neighbourhood_group

```
price_distribution = sns.scatterplot(data=new_df1, x='neighbourhood_group', y='price',) price_distribution.set_title("Density of prices in different neighbourhood group") plt.show()
```

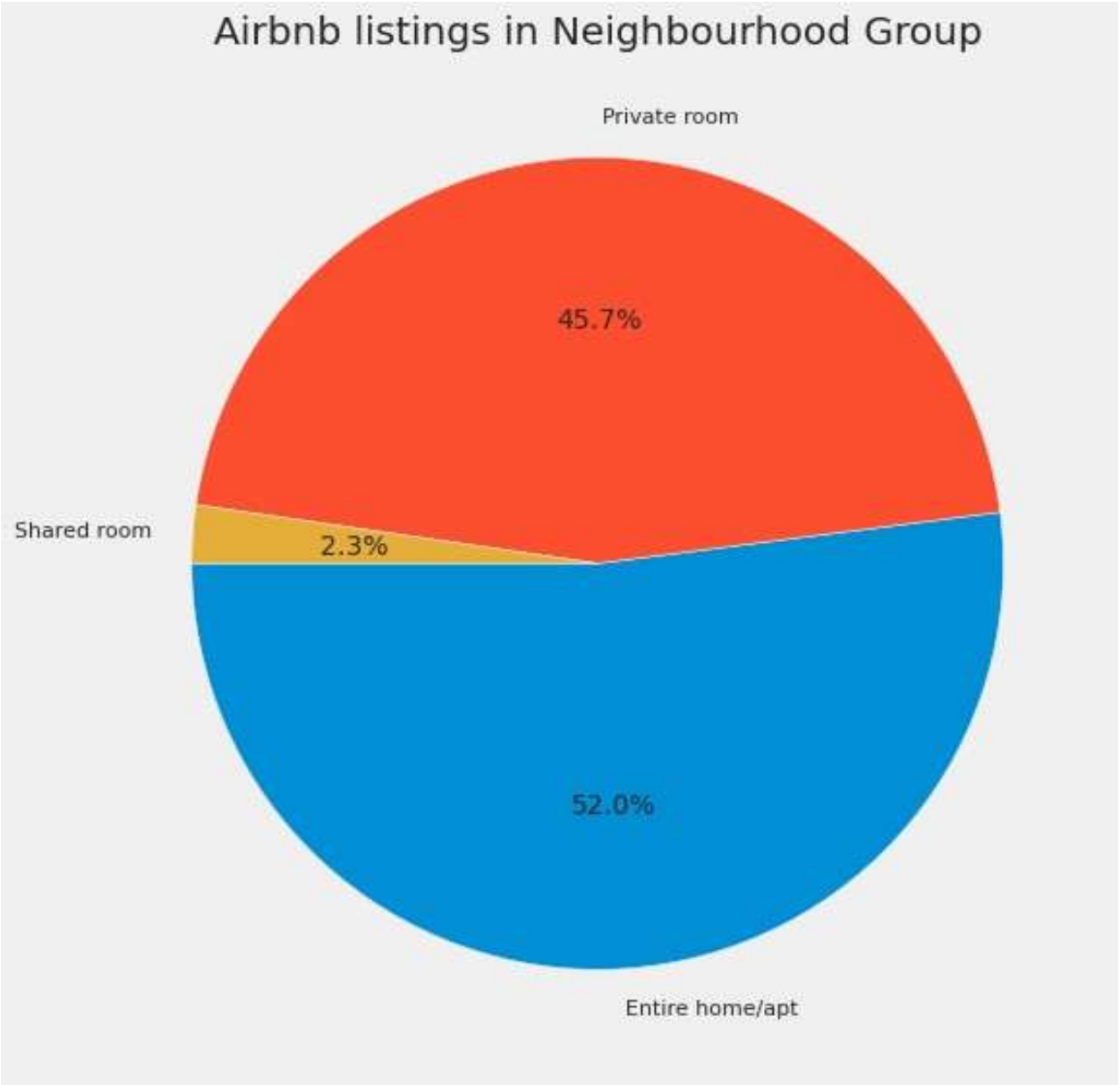


--> So, Brooklyn has the highest density of price distribution while Bronx has the lowest.

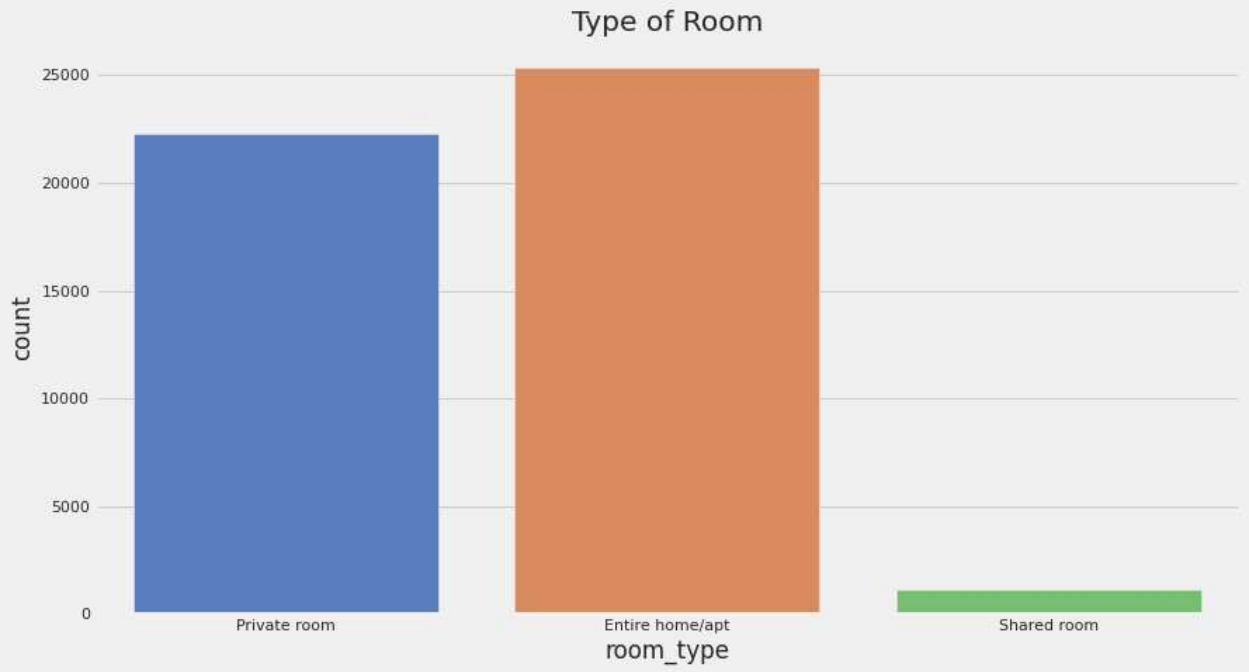
6) Lets count the different room type

```
counting_room_type = new_df1.room_type.value_counts()
print(counting_room_type)
```

```
plt.style.use('fivethirtyeight')
plt.figure(figsize=(15,9))
plt.title("Airbnb listings in Neighbourhood Group")
g = plt.pie(new_df1.room_type.value_counts(), labels=df.room_type.value_counts().in
dex,autopct='%1.1f%%', startangle=180)
plt.show()
```

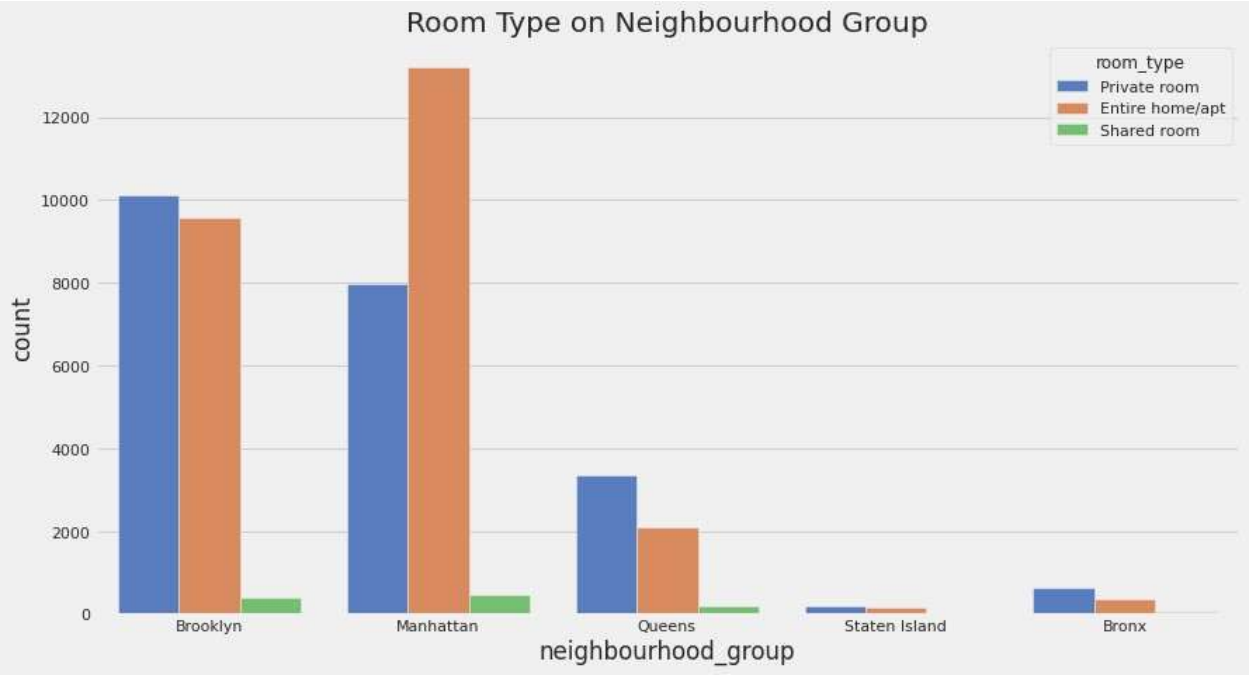



```
plt.figure(figsize=(13,7))
plt.title("Type of Room")
sns.countplot(new_df1.room_type, palette="muted")
fig = plt.gcf()
plt.show()
```



7) Count of different room_type in each neighbourhood_group

```
plt.figure(figsize=(13,7))
plt.title("Room Type on Neighbourhood Group")
sns.countplot(df.neighbourhood_group,hue=df.room_type, palette="muted")
plt.show()
```



1.Brooklyn have the high number of private room space
HERE WE CAN SEE, that

2.Manhattan have the high number entire home/apt. room type

3.Queens have the highest private room spaces which is much lesser than brooklyn and manhattan.

4.Staten island have much less number of room_types. and there is almost negligible shared room type

5.In Bronx, private room is higher than the Entire room/apt. and also there is no shared room type.

8) Average Price of each Room_type

```
avgprice_of_room_type = new_df1.groupby('room_type')['price'].mean().reset_index().
sort_values('price')
avgprice_of_room_type
```

	room_type	price
2	Shared room	71.796082
1	Private room	89.000314
0	Entire home/apt	210.081124

```
sns.barplot(x='price', y='room_type', data=avgprice_of_room_type)
plt.title('Average price of Airbnb based on room type')
plt.show()
```

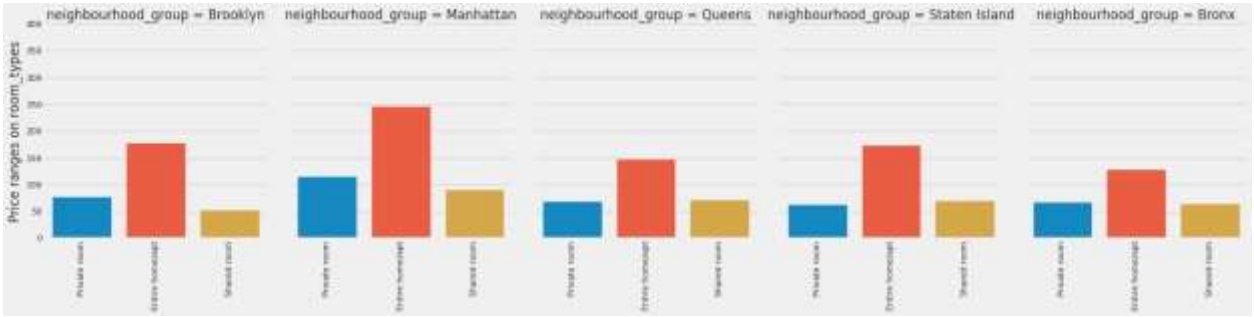
- So we can see that the most expensive room type is Entire home/Apt followed by private and share room.
- We can easily understand that the price is high for Entire room as it has high space area , and have different facilities provided by host for the convenient of guests
- Average Price of Private room is like about 50% cheaper than Entire room

9) Price range of each room types in different neighbourhood_group

```
room_type_price_of_neigh_grp =  new_df1.groupby(['neighbourhood_group','room_type']
)['price'].mean().unstack()
room_type_price_of_neigh_grp
```

room_type	Entire home/apt	Private room	Shared room
neighbourhood_group			
Bronx	127.506596	66.978462	64.222222
Brooklyn	177.422259	76.615940	52.055696
Manhattan	246.544296	115.640035	89.409664
Queens	147.368245	68.849213	70.963351
Staten Island	173.846591	62.518717	69.142857

```
price_range_room_type_in_neigh_grp = sns.catplot(x="room_type", y="price", col="neighbourhood_group",
                data=new_df1, saturation=.8,
                kind="bar",ci=None,  aspect=.9)
(price_range_room_type_in_neigh_grp.set_axis_labels("", "Price ranges on room_types")
  .set_xticklabels(["Private room", "Entire home/apt", "Shared room"],rotation=90)
  .set(ylim=(0, 400))
  .despine(left=True))
```



10) Minimum nights guests stays at Airbnb

```
min_nights  = new_df1.minimum_nights.value_counts().head(10)
min_nights1 = pd.DataFrame(min_nights)
min_nights1.reset_index(inplace=True)
min_nights1.rename(columns={'index':'min_nights','minimum_nights':'number of property' }, inplace=True)
min_nights1
```

```
gg=new_df1.drop(["id","host_id"],axis=1,inplace=True)
```

	min_nights	number of property
0	1	12686
1	2	11684
2	3	7990
3	30	3750
4	4	3300
5	5	3027
6	7	2053
7	6	752
8	14	558
9	10	478

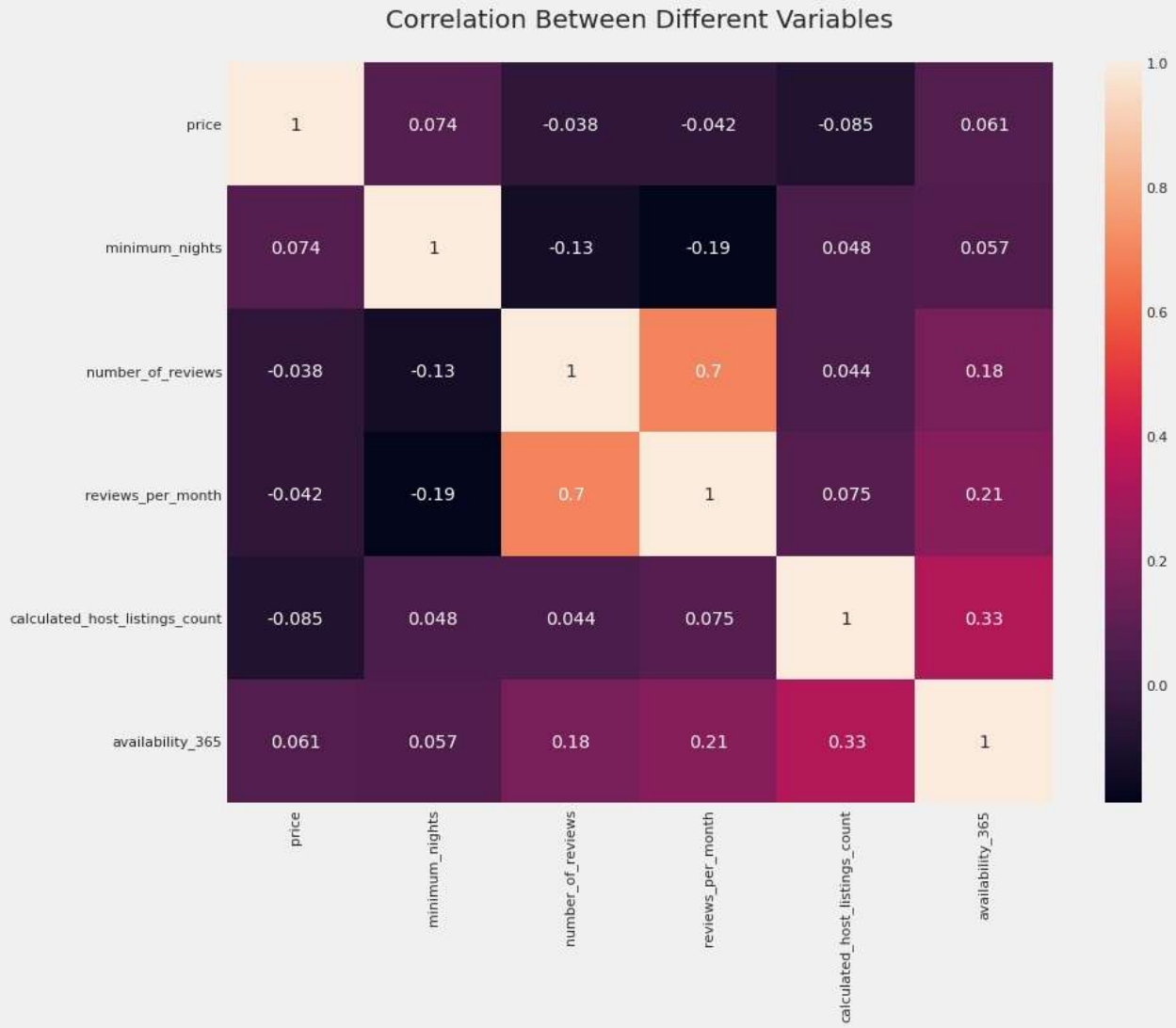
We can see that around 12686 Airbnb listings have minimum_night of 1

11) Correlation between the columns

```
# removing id and host id from correlation matrix because it will give biased value

gg
```

```
corr = new_df1.corr(method='kendall')
plt.figure(figsize=(13,10))
plt.title("Correlation Between Different Variables\n")
sns.heatmap(corr, annot=True)
plt.show()
```


Correlation Between Different Variables
price
minimum_nights
number_of_reviews
reviews_per_month
calculated_host_listings_count
availability_365
1
0.074
-0.038
-0.042
-0.085
0.061
0.074
1
-0.13
-0.19
0.048
0.057
-0.038
-0.13
1
0.7
0.044
0.18
-0.042
-0.19
0.7
1
0.075
0.21
-0.085
0.048
0.044
0.075
1
0.33
0.061
0.057
0.18
0.21
0.33
1
1.0
0.8
0.6
0.4
0.2
0.0

CONCLUSION

After exploring and analysing through data and visualization, we obtained some interesting insights into the Airbnb domain....

1. The neighbourhood group **"Manhattan"** has the most expensive bookings compared to the other neighbourhood group. We can say this based on the neighbourhood vs listings and neighbourhood vs price, the chart and graph cleary shows us that Manhattan has the highest number of Airbnb bookings and expensive because it receives average price of $194.8.!
2. **Manhattan** is also considered as the **best location** based on the graph of neighbourhood group vs number of reviews. Why Manhattan is best, most expensive and most trafficed location? So, we did some research and found out that why Manhattan is the best location, because it is closest to the famous city hotspots like Time Square, Empire Street, Central Park, etc and have very convenient transportation services.
3. **Busiest Host** = Based on the different graphs like neighbourhood vs price, neighbourhood group vs Airbnb listings and neighbourhood group vs number of reviews, **the host(host id = 219517861 and his name = Sonder NYC)** who has 327 listings is considered as the busiest host in NYC and **he belongs to the Manhattan**. We get to understand that Manhattan and brooklyn provides the most houses and rooms. Hence, we can imagine that not only the host who has 327 listings but if we consider the top 10 busy hosts, they all are from the area of Manhatten and Brooklyn.

4. According to our analysis, we **noticed** some **difference in traffic among different areas. Manhattan**, **Brooklyn** and **some part of Queens have the high traffic of airbnb bookings because** they got some famous city hotspots for exploring, travelling or for business purposes, and also they have very convenient transportation. Also there could be many reasons such as clean and hygienic houses and rooms, best amenities provided by hosts, and many more...

My Team Contribution

Team Member's Name, Email and Contribution: .

1) Vishu Rajput
Contribution:

a) Data Wrangling
b) Top hosts with highest number of listings
c) Counting and plotting of different room type
d) Price distribution in different neighborhood group
e) How much the top host earns.
f) Conclusion
g) PPT making.

2) Sachin Kumar
Contribution:

a) Data Wrangling
b) Top hosts with highest number of listings
c) Counting and plotting of different room type
d) Price distribution in different neighborhood group
e) How much the top host earns.
f) Conclusion
g) PPT making.