Airbnb in New York City - Impact of Neighborhoods

- **An Exploratory Data Analysis Project**

Kumardipta Sarkar

14.06.2021

Contents

1. Introduction

1.1 Background

<u>About Airbnb</u>

> "**Millions of Airbnb Hosts connect curious people to an endlessly interesting world.** Guests can discover the perfect place to stay for every getaway and explore new experiences while traveling, or online. Hosts can list their extra space, receive hosting tips and support, and earn money while creating memorable moments for guests."

- *This is how Airbnb describe themselves on the Google Playstore*

Airbnb is a platform provider for hosts and guests, where hosts can list their properties for the purpose of providing lodging and homestay facilities, and guests can avail these said facilities. Founded in the year of 2008, in San Francisco, California - Airbnb has come a long way such that now they have a global presence for providing their one of a kind service.

1.2 Problem and Interest

The business model of Airbnb is that it facilitates the rental process of accommodations, lodgings and homestays by providing an online marketplace. The company does not own any of the properties in the listings, they just charge a commission for each of the bookings.

Thus, one of the most important aspect would be to get an understanding of the locality of the properties and to see if and how it has any impact on its pricing or popularity.

This can be used for taking business decisions by getting an understanding of customers' and providers' behavior and performance on the platform as a result helping to guide marketing initiatives and maybe implementation of innovative additional services, etc.

2. Data

So now we move on to the data we will be requiring and using for this analysis.

- We will be using the "New York City Airbnb Open Data" available on Kaggle. The link to the database is: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

- This dataset has around 49,000 entries with 16 columns. We will not be requiring all the columns and hence we will perform data cleaning and wrangling methods to simplify the data as per our requirement

- Let us now understand the data. The columns for the original dataset and their description are as follows:

Columns	**Description**
id	id of the listing
name	title of the listing
host_id	id of the host who has listed
host_name	name of the host who has listed
neighbourhood_group	name of the borough
neighbourhood	name of the neighborhood
latitude	location latitude of the listing
longitude	location longitude of the listing
room_type	type of room / accommodation
price	price of the listing
minimum_nights	minimum number of nights to be booked for
number_of_reviews	total number of reviews for the listing
last_review	date of the last review
reviews_per_month	average reviews per month
calculated_host_listings_count	total no of listings by the host
availability_365	property available for number of days per year

- We already have latitude and longitude data of the properties in the dataset which can be used for finding the nearby venues for these properties using the **Foursquare API**

```
# let us see the size of the dataset

airbnb_df.shape
```

```
(48895, 16)
```

Once we load the data we can see that there are 48,895 rows and 16 columns. Some of the columns contain numerical data while the others contain categorical data.

```
# let us see the datatypes of the dataframe

airbnb_df.dtypes
```

```
id                                  int64
name                               object
host_id                             int64
host_name                          object
neighbourhood_group                object
neighbourhood                      object
latitude                          float64
longitude                         float64
room_type                          object
price                               int64
minimum_nights                      int64
number_of_reviews                   int64
last_review                        object
reviews_per_month                 float64
calculated_host_listings_count      int64
availability_365                    int64
dtype: object
```

For our analysis we can remove the data relating to the hosts as it will not be required. Hence, we can drop the columns host_id, host_name and calculated_host_listings_count.

Also, we can remove last_review column.

Checking to see if there are any mull values in any of the columns we see that there are 16 null entries in the name column and 10,052 null entries in the reviews_per_month column.

```
# let us find the total number of null values per column

airbnb_df.isnull().sum()
```

```
id                         0
name                      16
neighbourhood_group        0
neighbourhood              0
latitude                   0
longitude                  0
room_type                  0
price                      0
minimum_nights             0
number_of_reviews          0
reviews_per_month      10052
availability_365           0
dtype: int64
```

How should we deal with these?

We can drop the rows where the name is null. And for reviews_per_month, we can replace the empty values with 0 as logically empty reviews_per_month means no reviews have been given and hence 0 should suffice.

In the availability_365 some of the values are 0. So, the properties which are never available throughout the year will create noise for our model, hence it is better to get rid of them. So, we will remove the entries with availability_365 having value of 0

Now the finishing step for our data preparation process would be rearrange the columns.

Here price is our dependent variable and rest other parameters are independent variables. Hence, we will move the price column to the last column for easier visualization and understanding. Also, we will drop the id column as it will also not be required for the analysis. Lastly let us also rename the neighbourhood_group column as borough.

Below is a sample of our finished dataset.

```
# The prepped dataset
airbnb_df.head()
```

	name	borough	neighbourhood	latitude	longitude	room_type	minimum_nights	number_of_reviews	reviews_per_month	availability_365	price
0	Clean & quiet apt home by the park	Brooklyn	Kensington	40.64749	-73.97237	Private room	1	9	0.21	365	149
1	Skylit Midtown Castle	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	1	45	0.38	355	225
2	THE VILLAGE OF HARLEM....NEW YORK !	Manhattan	Harlem	40.80902	-73.94190	Private room	3	0	0.00	365	150
3	Cozy Entire Floor of Brownstone	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	1	270	4.64	194	89
5	Large Cozy 1 BR Apartment In Midtown East	Manhattan	Murray Hill	40.74767	-73.97500	Entire home/apt	3	74	0.59	129	200

3. Methodology

Now that we have the cleaned data we can start with our analysis.

So, our methodology for the analysis will be as follows:

- As we already mentioned, for our analysis price will be the dependent variable and we will try the understand how the other variables are affecting it. We will conduct Exploratory Data Analysis to understands the relationships and trends.
- In our dataset we have the location latitudes and longitudes for each of the properties. Using the Foursquare API, we will find the venues nearby to each of the properties and form clusters by using Kmeans Clustering. Once we have the cluster labels, we will analyze the price trends for each of the cluster and try to observe if there is any relationship or trend.

4. Analysis

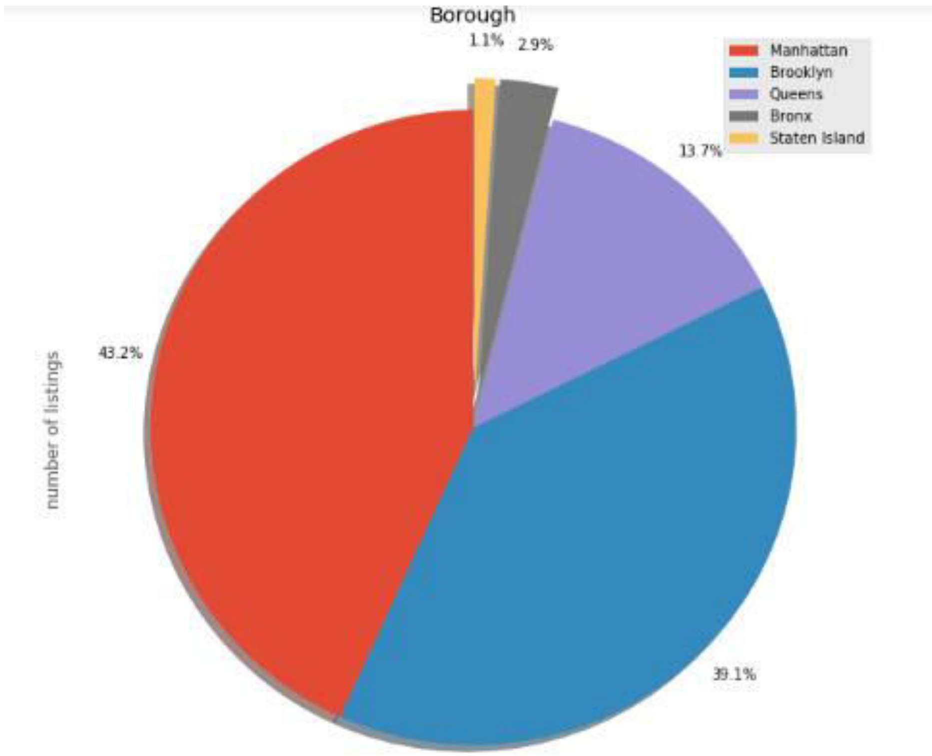
4.1 Exploratory Data Analysis

<u>Boroughs</u>

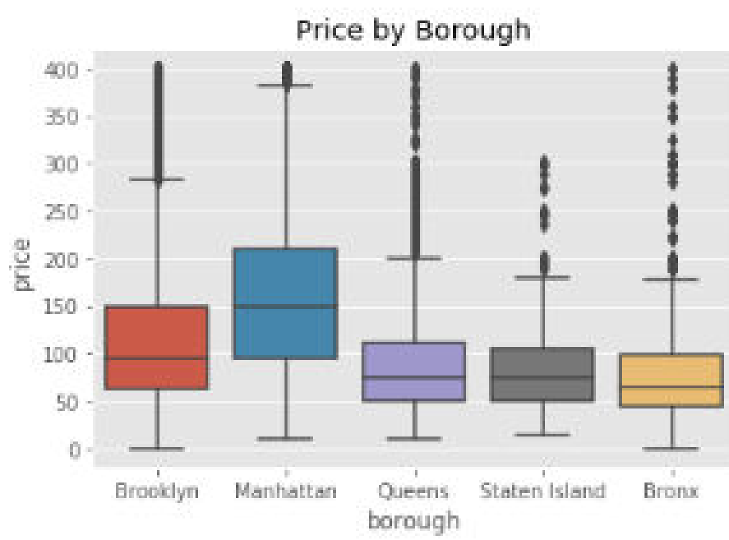
There are 5 boroughs in the City of New York, which is also evident from the table. As below we can see a table and a pie chart showing the distribution of listings per borough.

We would have assumed *Manhattan* and *Brooklyn* to have higher listings compared to the other boroughs as these two are comparatively busier and more crowded than the rest. And now we can see that our data also reflects the same.

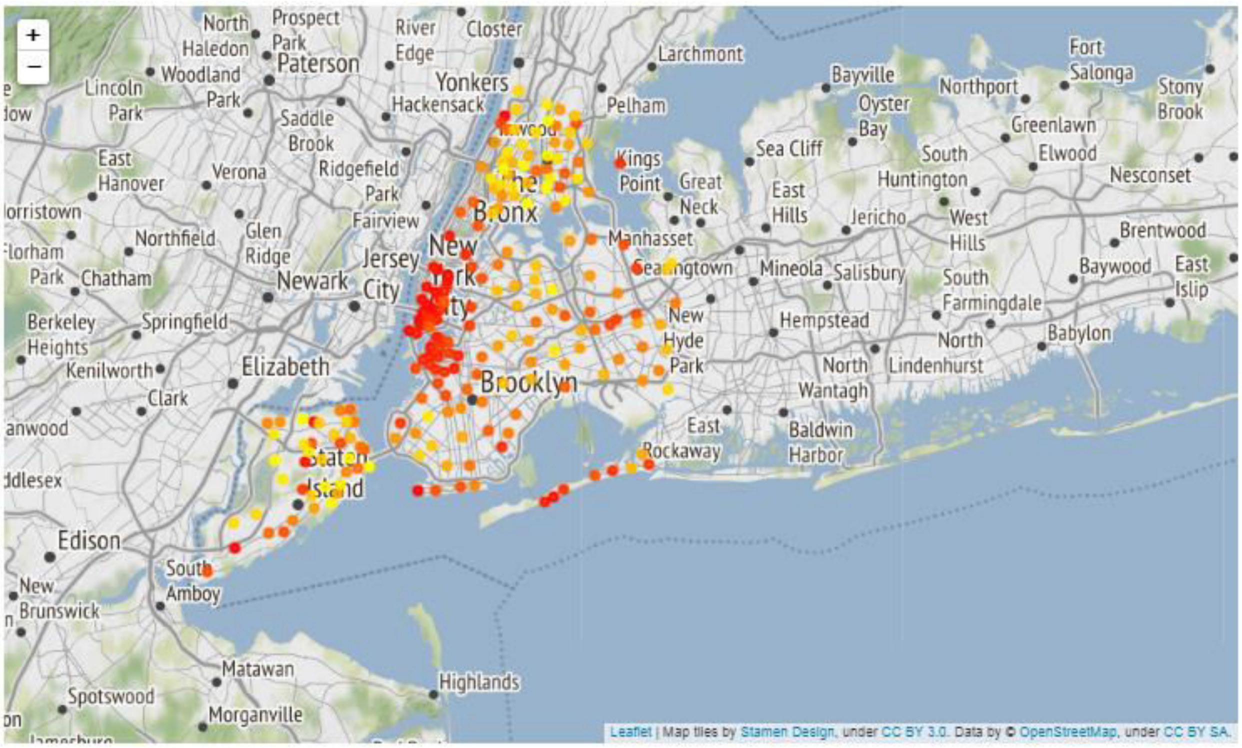
	number of listings
Manhattan	13557
Brooklyn	12259
Queens	4298
Bronx	913
Staten Island	331



Also, as we can see from the below boxplot, the median price is higher in the more popular boroughs of *Brooklyn* and *Manhattan*. This is also expected.



Neighborhood

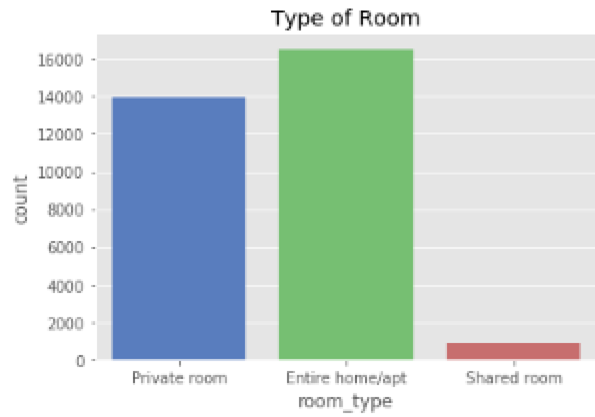


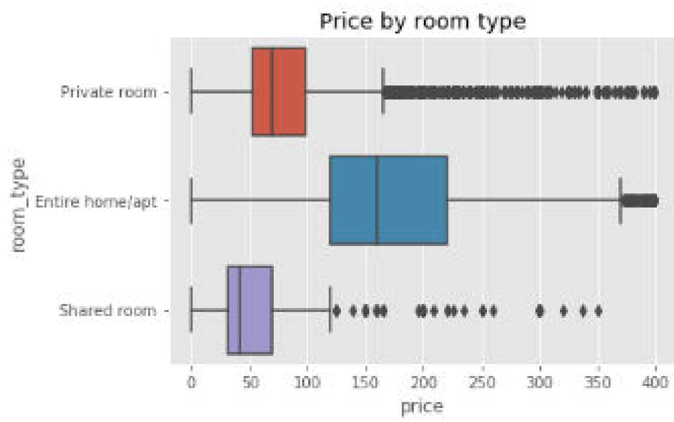
The above Folium Map of New York shows us the markers per neighborhood, color coded with respect to average price of listings from highest to lowest. Here, the color gradient is red to yellow, where red shows the highest value and yellow shows the least.

As we can see from the above map, the listings for the neighborhoods of Manhattan are concentrated in red. This shows that the prices are comparatively higher for the properties in Manhattan. This also goes in line with the fact that in real life the Manhattan area is more high cost compared to other areas of New York.

Similarly, average prices are lower in the neighborhoods of Bronx and Staten Island as can be seen from the map markers. This is also as per expectation.

Room Type

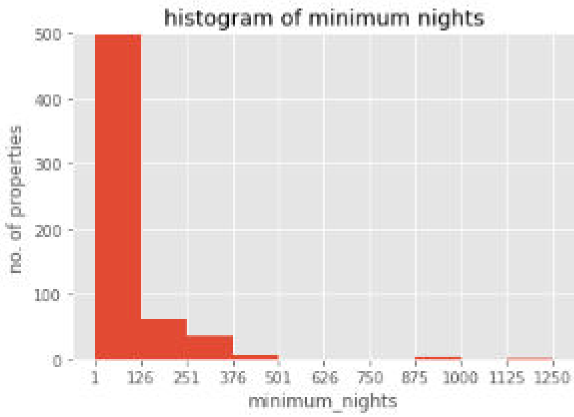




From the above figures we can see that the number of listings with shared room as room type is considerably much lesser than the listings for room types of private room or entire home/apartment. This can be due to the fact that people prefer to stay with privacy rather than staying in a shared setup in general, and this is also reflected by the listings as shown in the countplot.

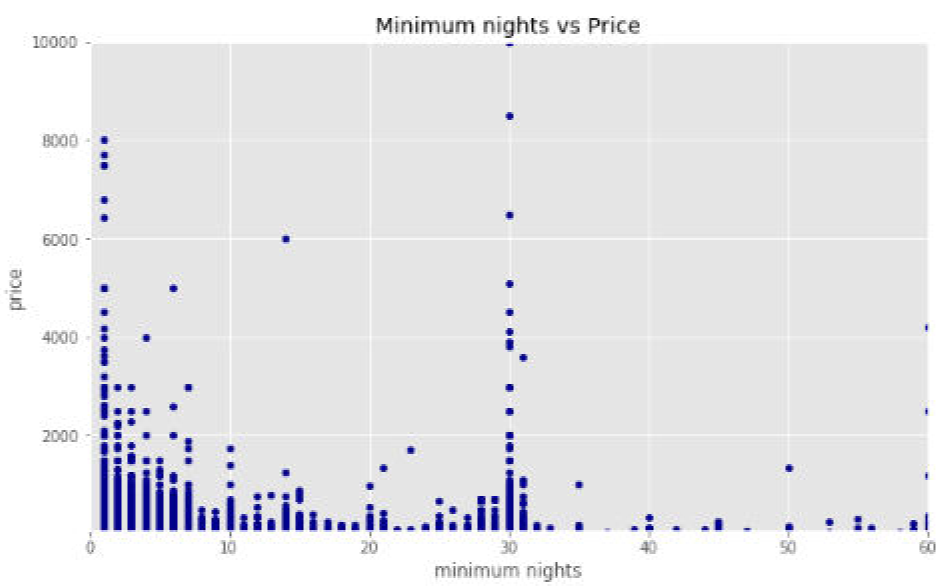
Also, we have made boxplots comparing the price and the room types. Here also we can observe that the price range and median price of entire home/apartment is also much higher than that of both private or shared rooms, which is also very obvious.

Minimum Nights



As we can see in the above figure, the highest number of properties are listed for minimum nights of 1 - 126. The values for the number of properties in the first bin was very high (more than 30000 in this case) so we have put a limit till 500 for better visualization.

So, this result is also as per expectation. This is because the model of Airbnb is to rent the properties for short term basis most of the times.

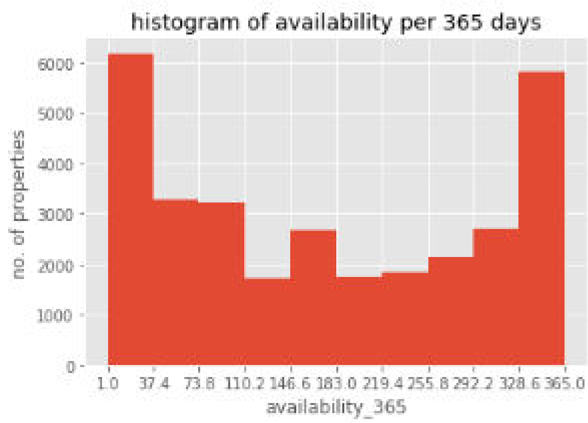


As per the above figure we can see that there is not much correlation between the price and minimum nights.

But we can observe one trend that is the most variance in price for minimum nights is in the case of minimum nights equal to 1 and 30. This means that these are the most popular bookings for minimum nights, thus having most varied types of options starting from lowest prices to prices as high as 10000.

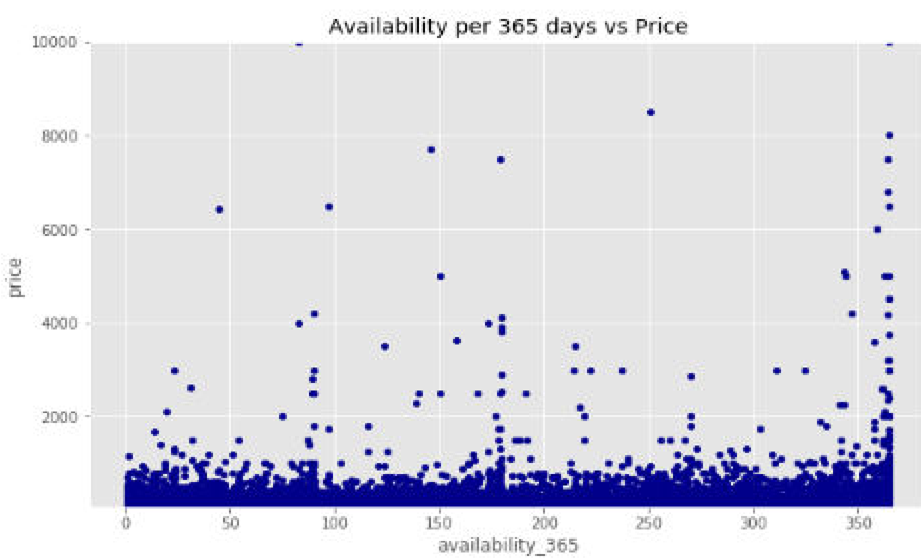
Availability 365

The analysis we did for *minimum_nights*, we can do a similar analysis *availability_365*.



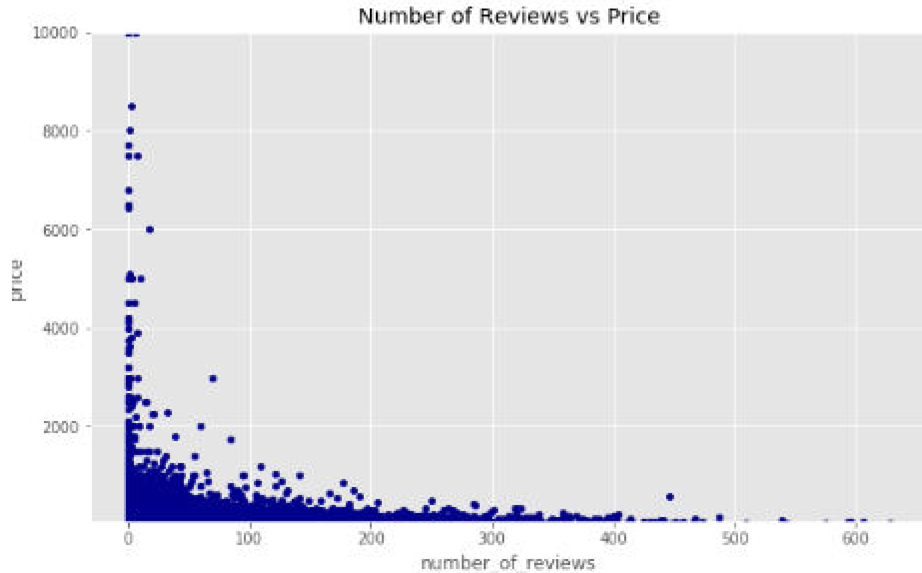
From the above plot, we can see that the number of properties available for number of days per year show extreme polarities. Either it is available almost throughout the year or it is available for only a day or a week maybe.

This behavior can be explained by understanding the offerings of Airbnb. Just as Airbnb offers regular spaces for lodging and home staying, they also have certain exclusive destination themed properties. Hence the availability of the exclusive properties maybe for a very limited time like say for a day or for a week maybe, whereas the regular lodging options may be available throughout the year.



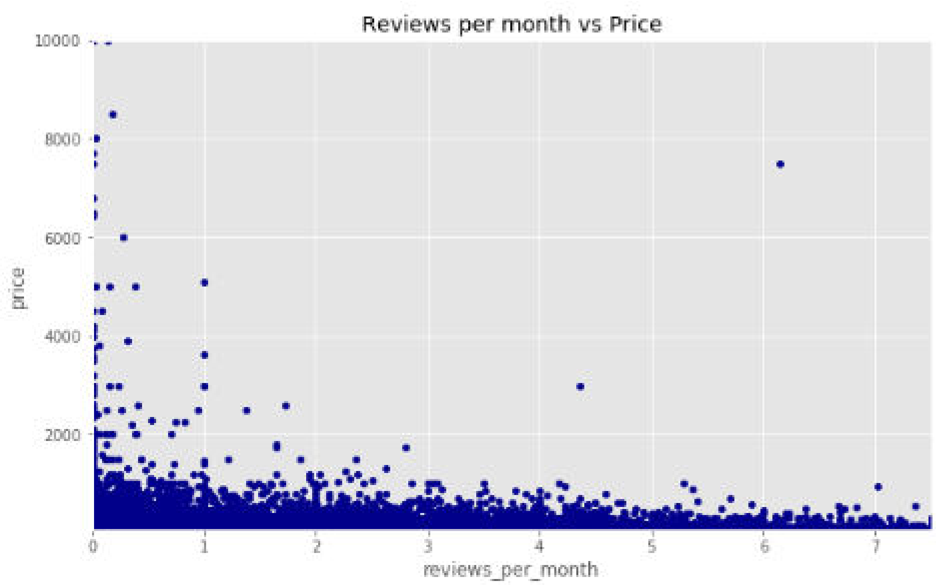
Prices do not show much correlation with availability per 365 days.

Number of Reviews & Reviews per Month



Comparing price with number of reviews we do not see much correlation.

But one conclusion we can draw from here that is the high-priced properties have less reviews. This may be due to the fact they may not be in everyone's budget and hence only few people may have availed these properties and so the number of reviews is also less.



Reviews per month also shows a similar trend to that of number of reviews.

No distinct correlation observed

Name

Wow this is quite an interesting visual!

We have plotted the *WordCloud* of all the words present in names of the listed properties and we have superimposed it on a map of New York.

As we can see from the *WordCloud*, analyzing all the names of the listed properties the most common words are Manhattan, Private and Room.

This shows that the most common listings are for **Private Rooms in Manhattan**, which was also obvious from our previous analysis. This is really great!

4.2 Clustering with Nearby Venues

In this section, we will use the latitudinal and longitudinal data of each of the properties and utilizing the *Foursquare API* we will find the top 10 nearby venues for each of the properties.

Once we have the top 10 nearby venues we will use *KMeans clustering* to separate the properties in cluster based on inter-cluster similarity and intra-cluster dissimilarity.

Once the clusters are formed we will analyze their clustering trend and try to understand if the clusters have any relationship with price.

<u>Foursquare API</u>

> The Foursquare Places API is a geolocating and geotagging API service provided by Foursquare. This API provides location-based experiences with diverse information about venues, users, photos, and check-ins and all these are provided on real-time basis.

<u>KMeans Clustering</u>

> KMeans Clustering method is a simple clustering method which is very useful for processing unlabled data. The main concept behind clustering is segregating data into clusters based on parameters by maximizing the inter-cluster similarity and intra-cluster dissimilarity.

The version of Foursquare API that we will be using is the Personal version and has a limit of 99500 api calls per day. If we use the entire dataset then we may risk using up all our available api calls.

Hence, we will use the nearby venue clustering only for the properties in the borough of Manhattan. But there is a total of 13557 datapoints for Manhattan. This is also very high, so let us create a sample of 500 rows from the Manhattan datapoints.

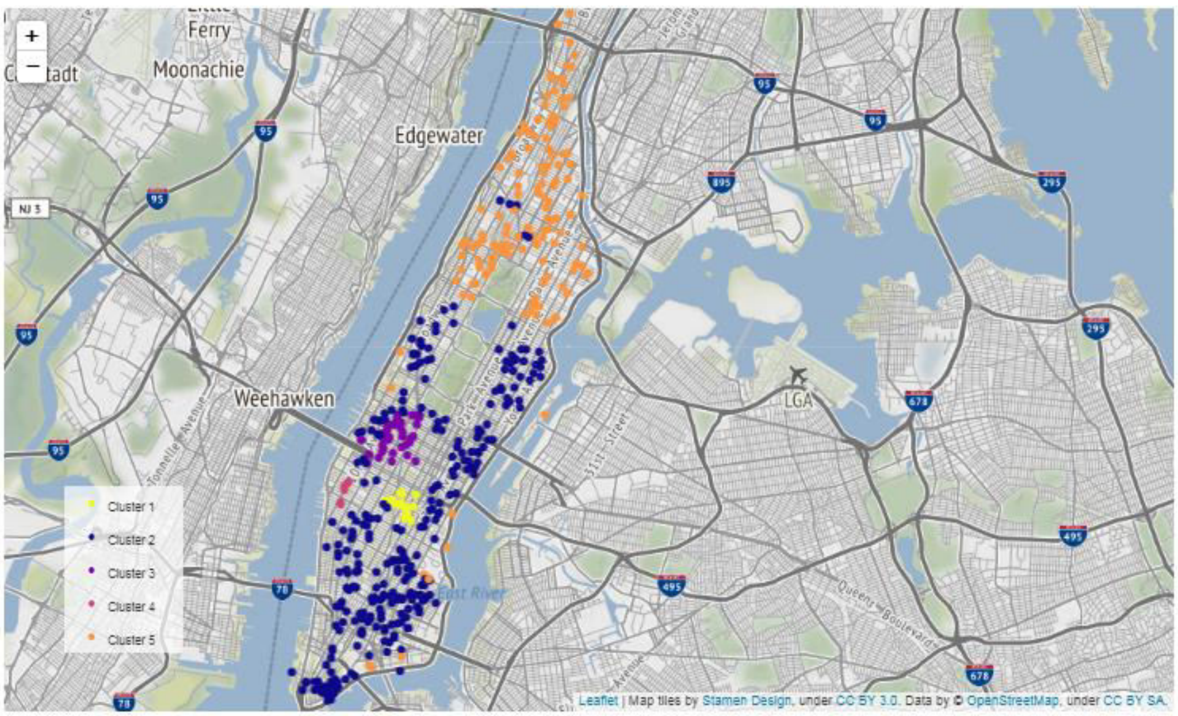
Once we have the selected 500 selected dataset, we use the Foursquare API to find the nearby venues for each of the properties.

Then we use one hot encoding for the nearby venues per property and then we group them by their means.

Finally, we find the most common venues in order for each of the properties.

	name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	"The Green Room": Harlem Brownstone	Coffee Shop	Pizza Place	Café	Deli / Bodega	Bar	American Restaurant	Bank	Pharmacy	Sandwich Place	Donut Shop
1	"Treehouse" in the East Village with Private P...	Wine Bar	Vegetarian / Vegan Restaurant	Korean Restaurant	Ice Cream Shop	Pizza Place	Bar	Vietnamese Restaurant	Cocktail Bar	Coffee Shop	Japanese Restaurant
2	(UES) Entire Apartment Near Central Park	Italian Restaurant	Coffee Shop	Japanese Restaurant	Gym	Sushi Restaurant	Spa	Deli / Bodega	Gym / Fitness Center	Thai Restaurant	Bar
3	*NO GUEST SERVICE FEE* Beekman Tower One Bedro...	French Restaurant	Italian Restaurant	Bakery	Bar	Coffee Shop	Japanese Restaurant	Turkish Restaurant	Gym / Fitness Center	American Restaurant	Thai Restaurant
4	*NO GUEST SERVICE FEE* Beekman Tower Studio Su...	French Restaurant	Thai Restaurant	Bar	American Restaurant	Bakery	Italian Restaurant	Coffee Shop	Hotel	Japanese Restaurant	Sushi Restaurant

Then we merge the nearby venues table and the original dataset of 500, and we use Kmeans Clustering on this data



From The above map visualization, we can see the properties distributed to 5 clusters and each of the properties being shown on the map using markers of corresponding color.

We can observe a segregation among the clusters on the basis of location. Let us now analyze the nearby values and try to understand the basis of the clusters.

1st Cluster

As we can observe from the below table, the first cluster has Asian restaurants mostly in the nearby venues. This means that these particular properties will be convenient from people travelling from Asia or of Asian descent.

As we will be able to see later from the Average price and Count per cluster lineplot, this is a very niche category hence very few properties fall in this cluster. Hence the number of properties aligning with this cluster is also less. Moreover, since these properties will attract mostly tourists the average price is also comparatively higher.

	name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	price
1	Fabulous 3BR/3BA NoMad Midtown LOFT	Korean Restaurant	Hotel	Gym / Fitness Center	Japanese Restaurant	Hotel Bar	Dessert Shop	Italian Restaurant	Bakery	Pizza Place	Café	800
8	Elegant Studio-Loft in Flatiron / NoMad	Korean Restaurant	Hotel	Gym / Fitness Center	Indian Restaurant	American Restaurant	New American Restaurant	Pizza Place	Spa	Italian Restaurant	Vegetarian / Vegan Restaurant	210
26	Prime 1 bedroom Doorman Gym RoofDeck 5221	Korean Restaurant	Hotel	Coffee Shop	Japanese Restaurant	Gym / Fitness Center	Bakery	Salad Place	Cosmetics Shop	Sushi Restaurant	Hotel Bar	260
32	East 29th Street, Luxury 1bd in NOMAD	Korean Restaurant	Hotel	Gym / Fitness Center	Spa	Japanese Restaurant	Pizza Place	Bakery	Dessert Shop	American Restaurant	Hotel Bar	219
55	Gilded Age Bohemia	Indian Restaurant	Hotel	American Restaurant	Korean Restaurant	Gym / Fitness Center	Pizza Place	Spa	Wine Shop	Italian Restaurant	Juice Bar	185

2nd Cluster

In the second cluster we observe that the nearby joints are mostly eateries or social spots. This means these properties will be in busy commercial areas and will attract most people who are travelling for business purpose and would require temporary accommodation.

It can be expected that this category will attract the most number of guests and hence we can observe from the Average price and Count per cluster lineplot that most of the properties fall under this cluster and the pricing is also moderate.

	name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	price
0	MURRAY HILL LUXURY 2 BEDROOMS	Coffee Shop	Deli / Bodega	Park	Japanese Restaurant	Sushi Restaurant	Bank	Seafood Restaurant	Salad Place	Gym	Gym / Fitness Center	280
2	Your New York Penthouse	Italian Restaurant	Spa	Sushi Restaurant	Japanese Restaurant	Pet Store	Wine Shop	Coffee Shop	Bookstore	Grocery Store	Salad Place	225
6	Amazing two bedroom with the terrace/73A.	Coffee Shop	American Restaurant	Italian Restaurant	Spa	Sandwich Place	Hotel	Bar	Gym	Bakery	Juice Bar	190
7	Sunny bedroom in Soho/Greenwich village	Italian Restaurant	Dessert Shop	Pizza Place	Café	Cosmetics Shop	Sushi Restaurant	Indian Restaurant	Indie Movie Theater	Vietnamese Restaurant	Gourmet Shop	78
10	Doorman Penthouse One Bedroom Laundry 5196	Italian Restaurant	Café	Coffee Shop	Mediterranean Restaurant	Pizza Place	Bagel Shop	Park	Bank	Cocktail Bar	Shoe Store	179

3rd Cluster

The third cluster shows mostly Theatres and social hangouts spots as nearby venues. Hence these properties will likely attract more leisurely and recreational people.

The Average price and Count per cluster lineplot will show us that the number of properties in this cluster is not very high but the price is moderate.

	name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	price
48	Up to 4 people-Only steps away from Times Squa...	Theater	Hotel	Gym	Pizza Place	Plaza	Bakery	Burger Joint	Juice Bar	Food Truck	Cocktail Bar	379
52	LUXURY 3 BR WITH DOORMAN~1600 BROADWAY	Theater	Hotel	Plaza	Steakhouse	Sushi Restaurant	American Restaurant	Coffee Shop	Ice Cream Shop	Juice Bar	Gym	560
61	High Tower Luxurious 1 Bedroom in Times Square	Theater	Coffee Shop	Hotel	American Restaurant	Burger Joint	Gym / Fitness Center	Bakery	Juice Bar	Bar	Performing Arts Venue	169
87	Elegant Private Room in Midtown West	Theater	Hotel	American Restaurant	Bakery	Coffee Shop	Deli / Bodega	Burger Joint	Gym / Fitness Center	Taco Place	Pizza Place	129
119	Luxury 1-Bedroom Apartment in Midtown Gym+Pool	Theater	Coffee Shop	Hotel	Sushi Restaurant	Mexican Restaurant	Steakhouse	Pizza Place	Burger Joint	Taco Place	Sandwich Place	239

4th Cluster

This cluster has art gallery as the most common nearby venue. It is slightly less priced than the third cluster and has comparatively less number of properties.

These properties will tend to attract people who are into artistic activities and thus will be a niche category.

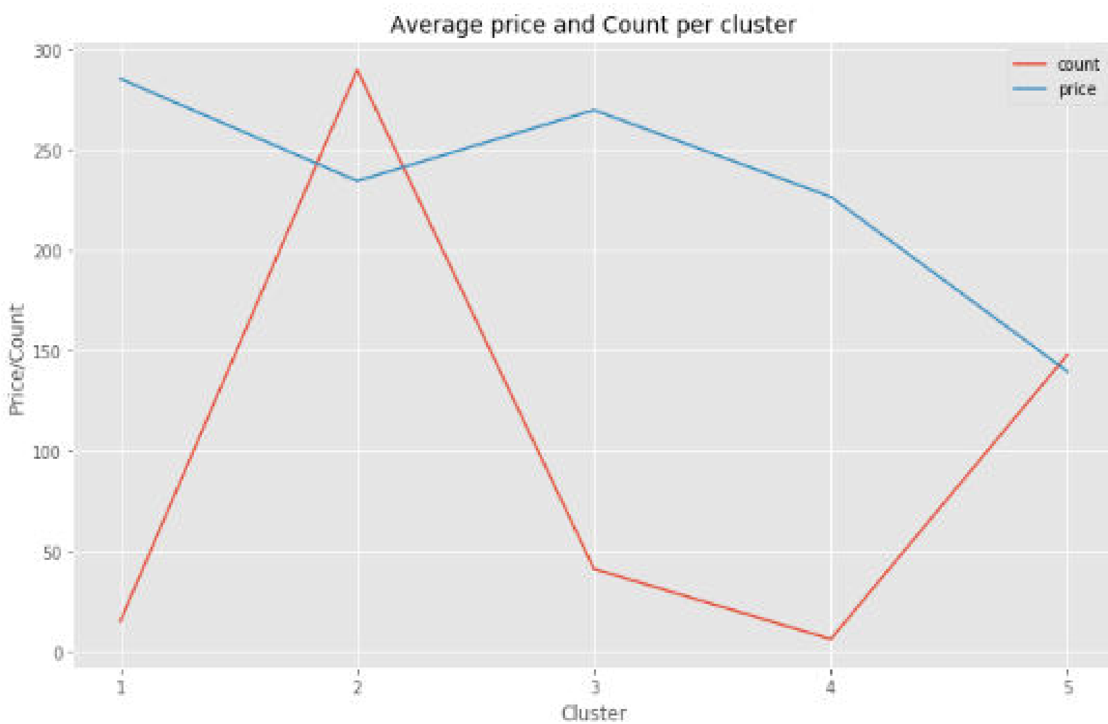
	name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	price
3	Chelsea Chic	Art Gallery	Gym / Fitness Center	Park	Coffee Shop	Italian Restaurant	Wine Shop	Restaurant	Food Truck	Performing Arts Venue	Playground	215
17	Private Room + Outdoor Space in Modern Chelsea...	Art Gallery	Park	Theater	Café	Indie Theater	Tapas Restaurant	Bakery	French Restaurant	Scenic Lookout	Gym	105
95	Chelsea Hudson yards Highline adorable apartment	Art Gallery	Gym / Fitness Center	Park	Coffee Shop	Restaurant	Cocktail Bar	Food Truck	Wine Shop	Playground	Tapas Restaurant	170
260	Design XL large one bedroom apartment in Chelsea	Art Gallery	Park	Coffee Shop	Cocktail Bar	Gym / Fitness Center	Italian Restaurant	Wine Shop	Bakery	Café	Playground	350
312	Chelsea 1 Bedroom Apartment in a doorman/elev ...	Art Gallery	Café	Ice Cream Shop	Park	Tapas Restaurant	Chinese Restaurant	Nightclub	Seafood Restaurant	Sandwich Place	Bakery	170

5th Cluster

For this cluster if we see the nearby venues we can notice that most of the areas are popular spots for people of color. The number of properties within this cluster is also quite significant and the prices are comparatively lower. This can also give us an idea about the socio-economic composition of Manhattan.

	name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	price
4	Oasis in Harlem	African Restaurant	French Restaurant	Sandwich Place	Gym	Seafood Restaurant	Pizza Place	Bar	Movie Theater	Tapas Restaurant	Library	150
5	Large luxury apartment. NYC	Southern / Soul Food Restaurant	Coffee Shop	Deli / Bodega	Pizza Place	American Restaurant	Juice Bar	Fried Chicken Joint	Chinese Restaurant	Theater	Gym / Fitness Center	280
9	3 bedroom duplex next to Central Pk	Pizza Place	Deli / Bodega	Yoga Studio	Café	Coffee Shop	Mexican Restaurant	Bar	Sushi Restaurant	Supermarket	Bubble Tea Shop	275
14	An Upper-Manhattan room of your own!	Donut Shop	Pizza Place	Mexican Restaurant	Park	Chinese Restaurant	Sandwich Place	Bar	Deli / Bodega	Grocery Store	Latin American Restaurant	51
15	A cozy Red Room with private bathroom	African Restaurant	Mobile Phone Shop	Southern / Soul Food Restaurant	Burger Joint	Grocery Store	Jazz Club	Wine Bar	Boutique	Cocktail Bar	Coffee Shop	160

Below is the distribution of number of properties and average price per cluster.



5. Results and Discussion

Now that we are done with our comprehensive and thorough study of the *Airbnb* dataset we get a very holistic idea about the properties and listings of Airbnb for the city of New York.

We had done some **exploratory data analysis** and performed a **clustering activity** based on venues nearby to the properties.

The results we get from our *exploratory data analysis* are as follows:

- In case of **boroughs**, Manhattan followed by Brooklyn have the most number of listings. Also, with respect to price the properties in Manhattan tend to have higher price, thus soliciting the fact that Manhattan is the urban core of New York
- The above result about boroughs was strengthened by our study of **neighborhood** with respect to average price of listings. We noticed the most concentration of high priced neighborhoods in Manhattan
- Studying the **types of rooms,** we observed that the with respect to numbers the *shared rooms* were the least and with respect to price the *entire home/apartment* were the highest. This reinforces the fact that most of the guests prefer privacy rather than sharing. Also, it highlights the obvious fact that rooms will be cheaper than entire homes or apartments
- Most of the properties are available for booking for **minimum nights** of a day or a week.
- For **availability across the year**, we observe most properties either available throughout the year or available for just a day or a week, thus giving us a better idea about the destination type properties being offered by Airbnb
- From the WordCloud of the **names** of all the listings, we were able to observe that the most popular listings were for ***Private Rooms in Manhattan***

For the clustering activity we divided the properties basis nearby venues into 5 clusters and conducted a comprehensive study of the listings for each cluster, trying to explain the count and average price of the properties per cluster.

6. Conclusion

The purpose of this project was to conduct an exploratory data analysis on the listings of Airbnb for the city of New York. This was done by comparing the various independent parameters to the price and number of listings and trying to establish logical explanations and valid justifications of the results observed.

Also, we did a cluster analysis type study on the listings with respect to their most common nearby venues. We tried to explain how each of the cluster justified its corresponding average price of listings and also its popularity basis the number of listings pertaining to each such clusters.

With these observations, stakeholders can plan accordingly with regards to more optimal pricing of the listings. It also does help understand the preferences and choices of both guests' and hosts' which will definitely help with maybe implementing additional services or also to guide marketing initiatives.