

**Exploratory Data Analysis on Airbnb Bookings**

**AMAN MISHRA RAJDEEP LIMBU**

**MAYUR CHOUGULE MOHIT SAKHARE**

**KIRAN MISHRA**

**CONTENTS**

**1. Introduction**

**2. Problem Statement**

**3. Data Wrangling**

**4. Data Visualization and Exploratory Data Analysis (EDA)**

**5. Limitation**

**6. Scope of Improvement**

**7. Conclusion**

**Introduction:**

Airbnb has been an American Company since 2007; it is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales.

The dataset from Airbnb is based in NY. NY is amongst the most expensive places to live in the USA. We want to perform an in-depth analysis of one of the most densely populated cities in the world. Our dataset is feature-rich, containing locations with coordinates, prices, hostname, room types, and availability throughout the season.

With these features, we’ve done exploratory data analysis and tried to extract information like the most expensive places to live in NY, is location varies with occupancy rate, what type of room people tends to choose most, is there any particular season for tourists or locale when we can follow a surge in prices or occupancy rate of properties, etc.

**Problem Statement:**

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 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, and implementing innovative additional services, and much more.

We need to 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)
* 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?

**Data Wrangling:**

The dataset contains 48895 observations with 16 features. This data file includes all needed information to find out more about hosts, geographical availability, and necessary metrics to make predictions and draw conclusions. Let us look through our features,

* Id: a unique id identifying an Airbnb listing or property
* name: name representing the accommodation
* host\_id: a unique id identifying an Airbnb host
* neighbourhood\_group: a group of area
* neighborhood: area falls under neighbourhood\_group
* latitude: coordinate of listing
* longitude: coordinate of listing
* room\_type: different room types
* price: price of listing
* minimum\_nights: the minimum nights for which the property was booked
* number\_of\_reviews: total count of reviews given by visitors
* last\_review: date of last review given
* reviews\_per\_month: number of reviews given per month
* calculated\_host\_listings\_count: total no of listing registered under the host
* availability\_365: the number of days for which a property is available in a year.

latitude and longitude have represented a co-ordinate, neighbourhood\_group, neighborhood and room\_type are columns of categorical type. last\_review is a column of date type; we will convert it as required.

The distribution of numerical columns are as follows,

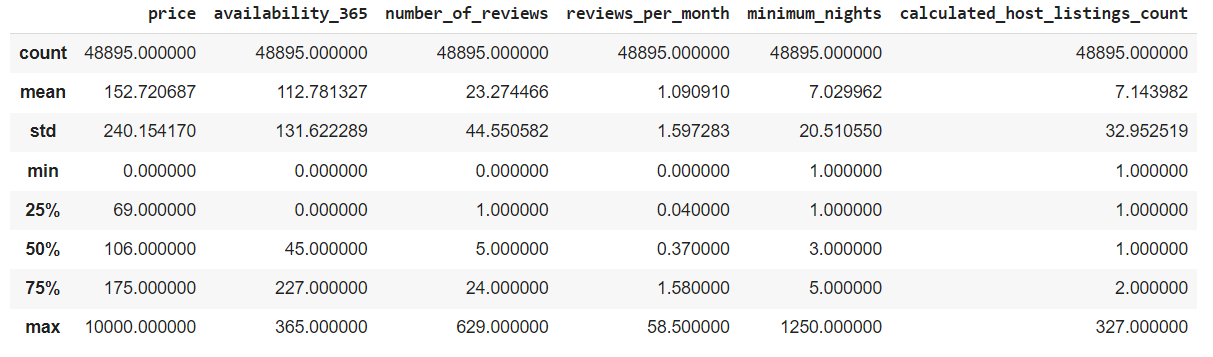


Fig 1. Statistical Distribution of Numerical Features

Other 3 important columns are,

* neighbourhood\_group: It contains 5 unique neighbourhood groups which are Manhattan, Brooklyn, Queens, Bronx & Staten Island.
* neighbourhood: It contains 211 unique neighborhoods.
* room\_type: It contains 3 unique room types which are Entire home/apt, Private room, Shared room

The distribution of our numerical columns has positive skewness.

Out of all columns, 4 columns containing null values which are name, host\_name (looks like listing name and host\_name doesn't really matter to us for now) and last\_reviews, reviews\_per\_month (obviously, if a listing has never received a review, it’s possible and valid). So, those null values have been replaced with 0 during our analysis.

**Data Visualization and Exploratory Data Analysis (EDA):**

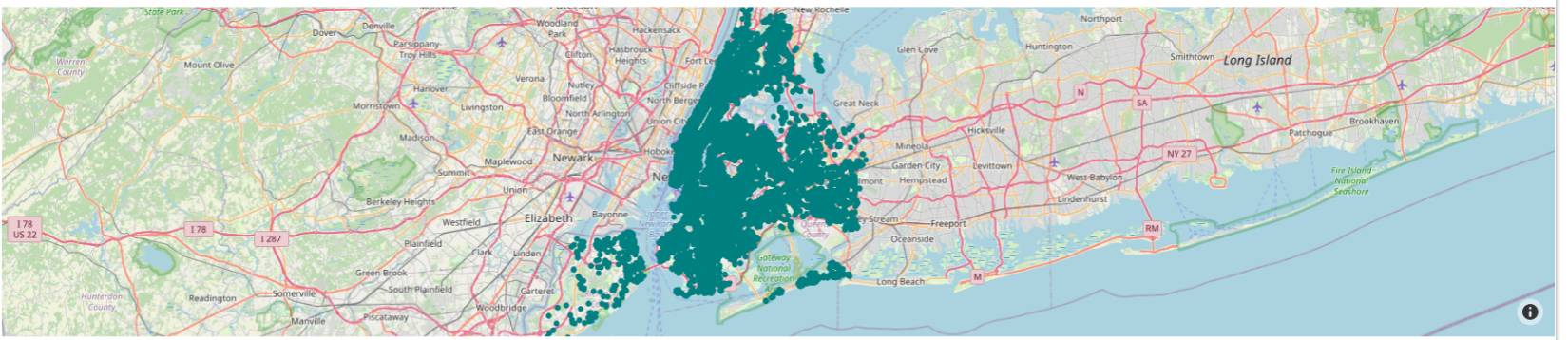
This is the exact graphical representation of the Dataset columns containing the values of longitude and latitude coordinates. The data is hence plotted on the map of NYC to show the actual representation of Airbnb listings.

Fig 2. Plotting of listings in the actual map of NYC

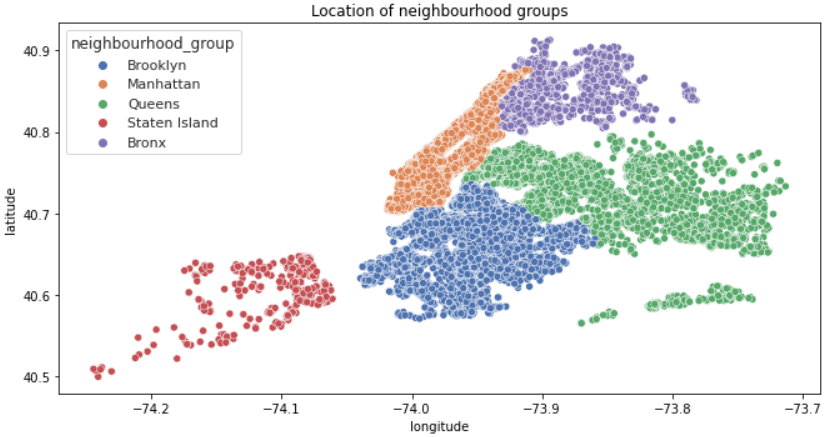
Below plot describes democratic view of properties listed also it provides a clear view of the city area.

Fig 3. Location of Neighborhood Groups

Next, we will look for the distributions of price, no. of reviews and availability throughout the area.

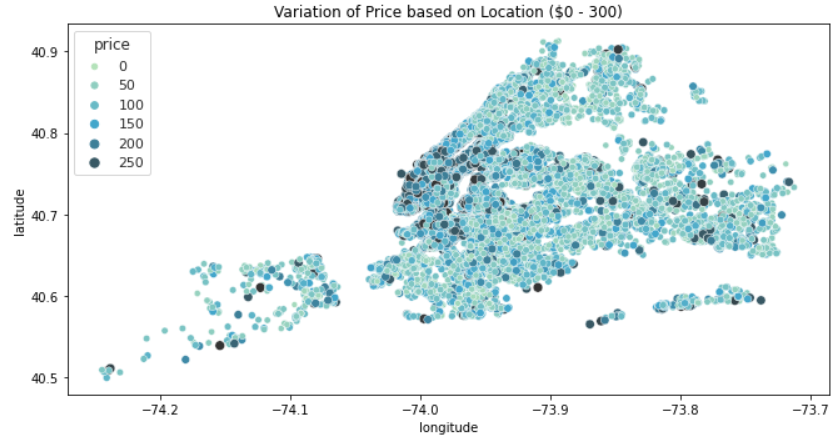
Below plot describes the distribution of property prices over filtered data of price range of maximum $ 300, as it can be seen within $ 175, 75% of property lies (fig 1). It can be noticed property prices are generally high in the southern area of Manhattan and northern area of Brooklyn. The south of Manhattan and north of Brooklyn belongs to the expensive areas of New York. (Source: [timeout.com](https://www.timeout.com/newyork/news/these-are-the-new-most-expensive-neighborhoods-in-nyc-010621) ). The average price of Manhattan & Brooklyn together is 161.98 where rest of the city average price is only 98.47.

Fig 4. Variation of Price based on Location ($0 - 300)

Third plot provides the distribution of rating count. Trend in rise of review count can be followed in the outskirts of NY. Staten Island tops the list with 30.94 average reviews per property, up next Queens with 27.74 avg reviews.

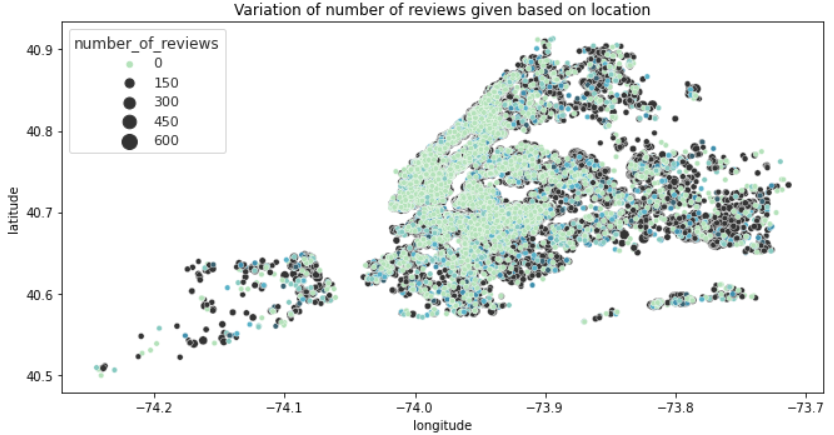


Fig 5. Variation of number of reviews given based on location

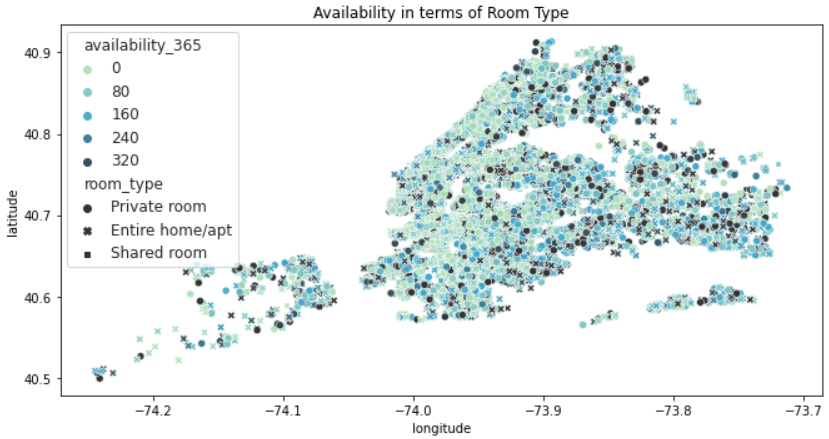
Fourth plot, creates an image of occupancy rate over the properties. Though availability based on room type is spread well, still a pattern can be followed where the heart of NY stays the busiest or booked for most of the time., with an average 106 days of availability within a year. This shows a sign of being a tourist favorite area.

Fig 6. Availability in terms of room type

With the analysis on revenue generated by the properties in different neighborhoods, it can be understood that Brooklyn and Manhattan stand within the most urban and active area, in terms of listing count and pricing both. Shared rooms are the cheapest and also has lowest count in every neighborhood, whilst Manhattan has the greatest number of Entire home/apt category, but Brooklyn has the greatest number of Private room category.

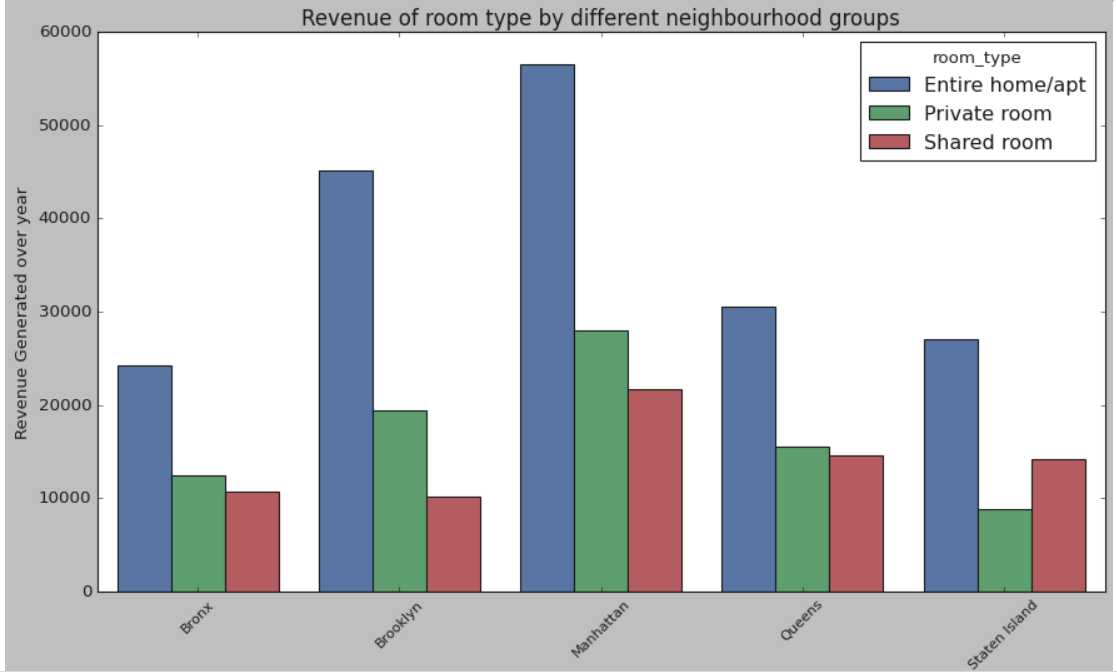


Fig 7. Revenue Generated vs Neighborhood Group

Brooklyn and Manhattan contribute the majority of share in the total listings of the property in the NYC region having values of 41% and 44% respectively. The pie diagram shows that the Staten and Bronx are the regions which contributes the least. So, various marketing and business strategies are suggested to the Airbnb for the increase of the in the number of bookings made in this region.

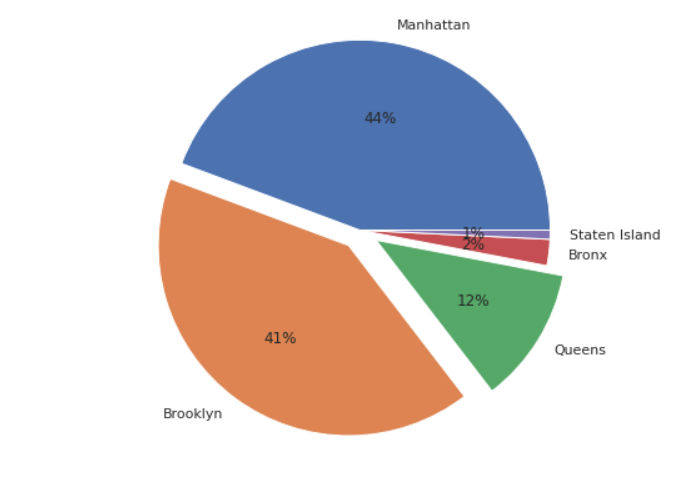


Fig 8. Percentage Listings of Properties

Sonder(NYC) followed by Blueground are one the top hosts which had their maximum properties listed in the Airbnb across NYC.

The most reviewed Neighbourhood can be seen as Brooklyn followed by Manhattan.

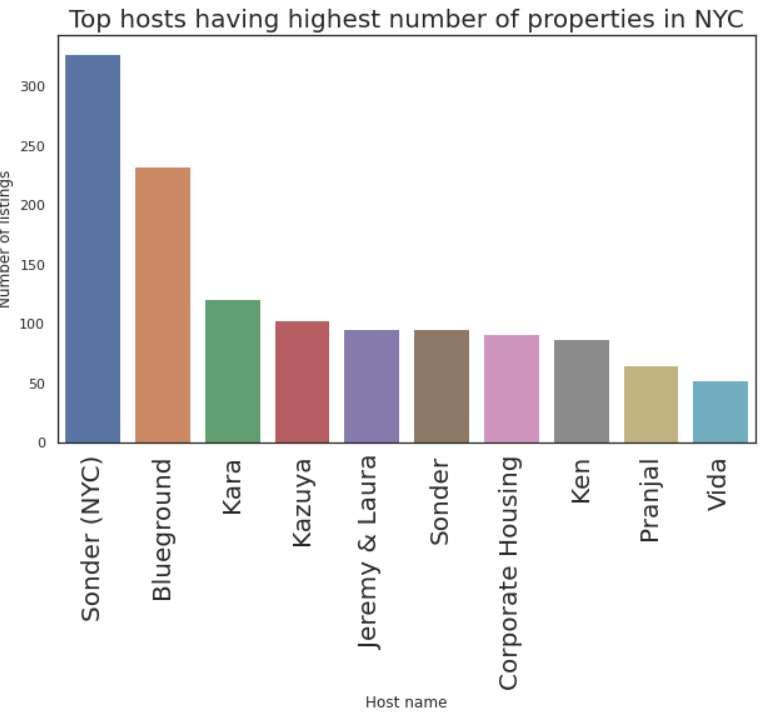
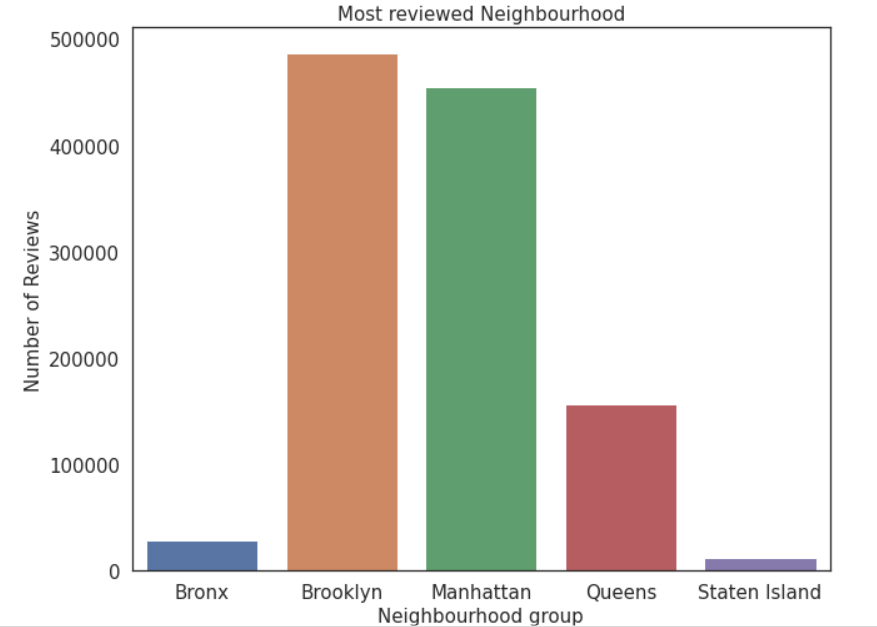
 

Fig 9. Top Hosts in the ListingsFig 10. Most reviewed Neighbourhood

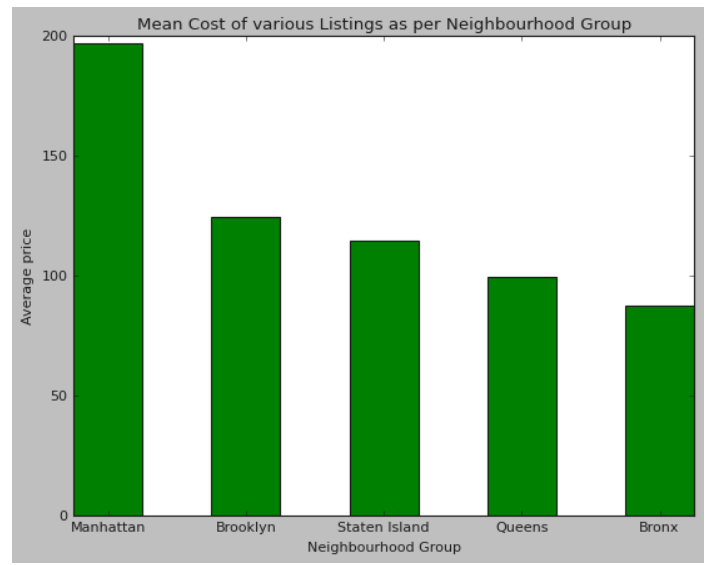
Manhattan seems to be the costliest neighbourhood to live with an average cost of $200.

Fig 11. Mean Cost of various listings as per Neighbourhood Groups

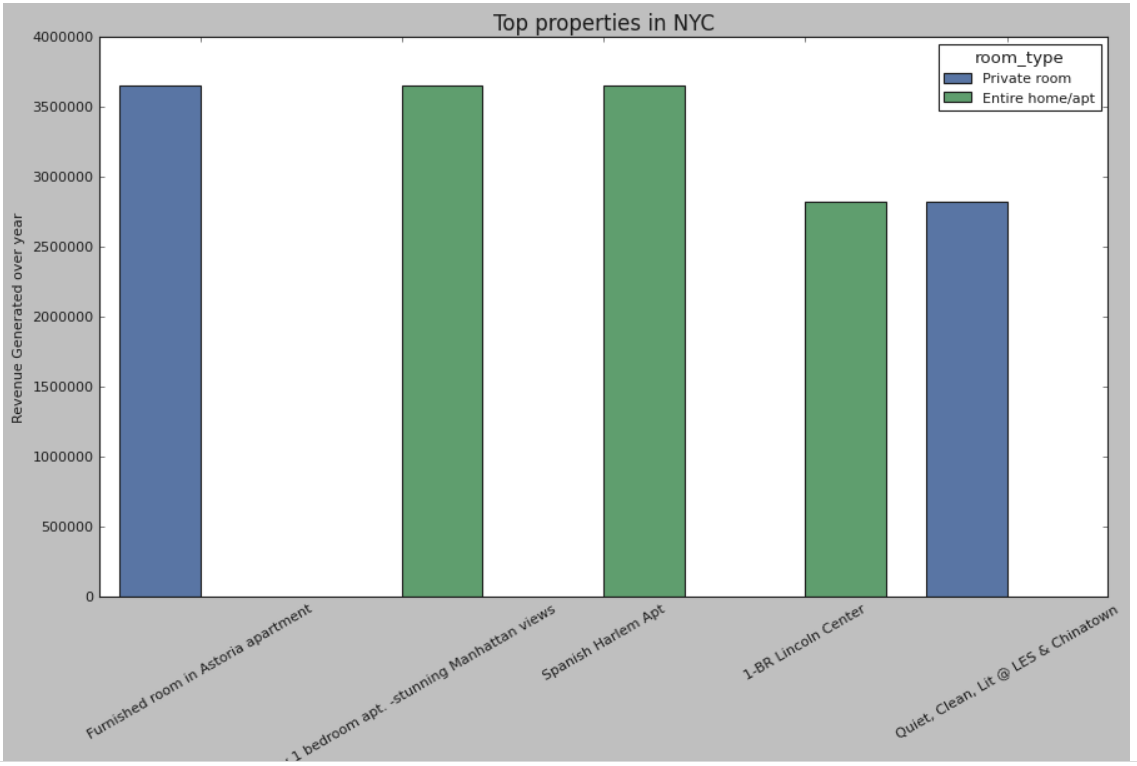
The below mentioned listings are the ones which have generated the highest revenue in a year.

Fig 12. Top Revenue generated properties in NYC

**Limitation:**

Though the dataset is significantly feature-rich, shares less correlation and contains enough sample to perform regression on price prediction, the correlation with target price is also low. So, it will result as high MSE.

To better analyze regarding the quality of the properties, it would be interesting if we had a breakdown of sentiments with property valuations.

Host user ratings aren't available; it would've been better to rank our hosts based on user satisfaction and ratings. Also, in those cases. The dataset can do further analysis to view how guests tend to rate in terms of price or room type or whether rating decides the property's valuation. Usually, a low-rated property tends to lower its cost.

The exact number of guest counts is also missing; it is just assumed that the guests by col: last\_review. A new host may never have been rated, but that doesn't mean no guest has ever stayed there.

**Scope of Improvement:**

As the dataset has few qualifying attributes to value a property, more features can be added like bedroom, bathroom, property age (it might be one of the most important ones), the tax rate applicable, distance to nearest airport, hospital, or schools.

In the presence of ratings, hosts can be classified and ranked, and special discounts or offers can be given to the highest-rated hosts following the marketing strategy.

Time series analysis can predict occupancy rates based on the tourist season.

**Conclusion:**

From the entire analysis, it can be concluded that,

* Most visitors don’t prefer shared rooms. They tend to visit private rooms or entire home.
* Manhattan and Brooklyn are the two distinguished, expensive & posh areas of NYC.
* Some properties are having Minimum Nights to stay more than 365 Days which can be favorable among Students, Low-Income Employees & Immigrants.
* Though property’s location is highly related to deciding its price, a property in popular area doesn’t mean it will stay occupied in most of the time.
* It’d be a better if we had avg guest ratings of a property, that would be beneficial in understanding the property more and could also be a factor in deciding price. A low-rated property tends to lower their price.
* Focus should be given on improving the bookings of the shared-rooms, by providing better offers and implementing new marketing strategy.
* More focus should be given towards improving the sale in Staten Island & Bronx region, by focusing on increasing the number of listings, providing better convenience and increasing the count of shared room-type as it appears to be the profitable segments.

THE END