# **The Battle of Neighborhoods: New York City hotel Analysis for Investors**

## [**1. Introduction**](https://utinokati.com/en/details/land-market-value/area/Tokyo/)

### **1.1 Background**

New York City received a ninth consecutive annual record of approximately 65.2 million tourists in 2018, counting not just overnight visitors but anyone visiting for the day from over 50 miles away, including commuters. Overall the city welcomed 37.9 million visitors who stayed overnight in 2018, of which 13.6 million were international. Major destinations include the Empire State Building, Ellis Island, the Statue of Liberty on Liberty Island, Broadway theatre productions, Central Park, Times Square, Coney Island, the Financial District, museums, and sports stadiums. Other major visitor activities include luxury shopping along Fifth and Madison Avenues; entertainment events such as the Tribeca Film Festival; Randalls Island music festivals such as Governors Ball, Panorama and Electric Zoo; and free performances in Central Park at Summerstage and Delacorte Theater. Many New York City ethnic enclaves, such as Jackson Heights, Flushing, and Brighton Beach are major shopping destinations for first and second generation Americans .

With this exquisite growth in tourists , New York city must surely have a great chain of hotels , lodges to be ready to accomodate this tremendous income.People are acknowledging this uptick in the economy by tourism and are thinking of investing in these hotel chains to be more rich.

### **1.2 Business Problem**

Hotel accomodation is a booming industry with the latest rise in tourism worldwide, especially in America. Over 20% of the total tourists worldwide are aged between 19-25. And around 80% of them prefer to spend less on accomodation by opting for best hotels. This means the demand for hotels is only going to grow further and more people would want to get their legs in the game.

1)How should a new business person decide where to open a hostel?

2)What factors should he look at before investing?

3)Which neighborhood venues affect a user's rating for location of hostel?

At the same time, it is difficult for a travellers, especially first-timers, to select a hostel from among many options. Hostel reviews are subjective and differ from person-to-person and one cannot solely depend on them to make a decision. It is especially important to consider other aspects like price and neighborhood, which can greatly influence one's experience of the city/country. I will try to answer the following questions

1)How does price vary with location?

2)Where are most of the hotels located ?

3)Where are least of the hotels located ?

4)Which is the most costly locality?

5)How does different ammenities affect the charge of hotel?

6)Which amenities are more common near hotels?

This recommendation system should help in answering these questions and give some suitable locations

### **1.3 Target Audience**

This project will serve two groups of audience:

Travellers: Help them make an informed decision while choosing a hostel by providing an in-depth analysis of hotels and their neighborhood. Business Person: Provide useful information and models which can help them where to open their first/next hotel.

## **2. Analytic Approach**

will be taking two approaches in the project.

Firstly, I will use exploratory data analysis(EDA) to uncover hidden properties of data and provide useful insights to the reader, both future traveller and investor.

Secondly, I will use prescriptive analytics to help a business person decide a location for new hotel. I will use clustering (KMeans)

## **Data Requirements**

The city that will be analysed in this project is New York.

Following are the datasets used in the project along with the reasons for choosing them:

[New York Hostel Dataset](https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data):Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019.

[Foursquare API](https://foursquare.com/): This API will help me get the venues around the hotel which I will use for EDA and clustering.

[New York Land Price](https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page): I will scrape this website to get land prices of various neighborhoods in Tokyo

Since there is a quota limit on Foursquare, I'll be downloading the data and storing it locally as csv for future use. So we do not have to pull same data again when we restart the data.

I have also downloaded the excel files from rolling sales website for the land prices in New York namely

Firstly, I will use the list of hostels from Hotel dataset and use Foursquare API to get venues around the Hostel. I will then use EDA to explore the neighborhood and how it affects the price of the hostel. I will also use the combined dataset to cluster similar hostels as per pricing and neighborhood.

Second, I will combine the above data with the land price for the area in which the Hotel is situated and them develop clustering and regression models to predict where a new hotel should be opened and how much should it be priced at.

 [ ]:

2.1 New York Hostel Dataset

The original dataset on Kaggle has the following columns:

\* means included in actual dataset

\*id -Hotel id

\*name- Hotel name

host\_id – Host id

host\_name – Host name

\*neighbourhood\_group – Neighborhood group

\*neighbourhood – Neighborhood name

\*latitude -- Latitude

\*longitude -- Longitude

room\_type – Type of Room

\*price – Price per Night

minimum\_nights – Minimum nights offered

number\_of\_reviews – Number of reviews

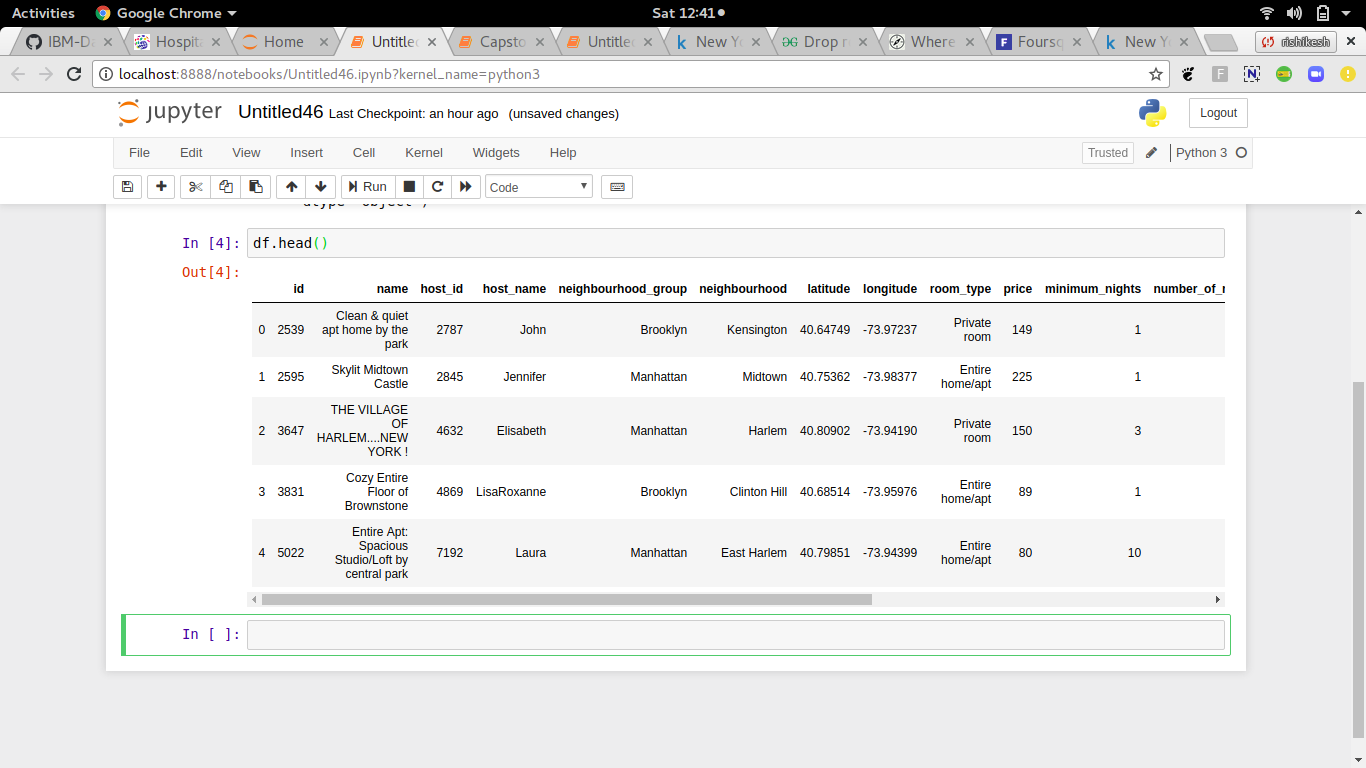
last\_review – Date of Last review

reviews\_per\_month – Number of reviews per month

calculated\_host\_listings\_count – Host lising

availability\_365 -- availabilty

Below is a snapshot of the dataset:

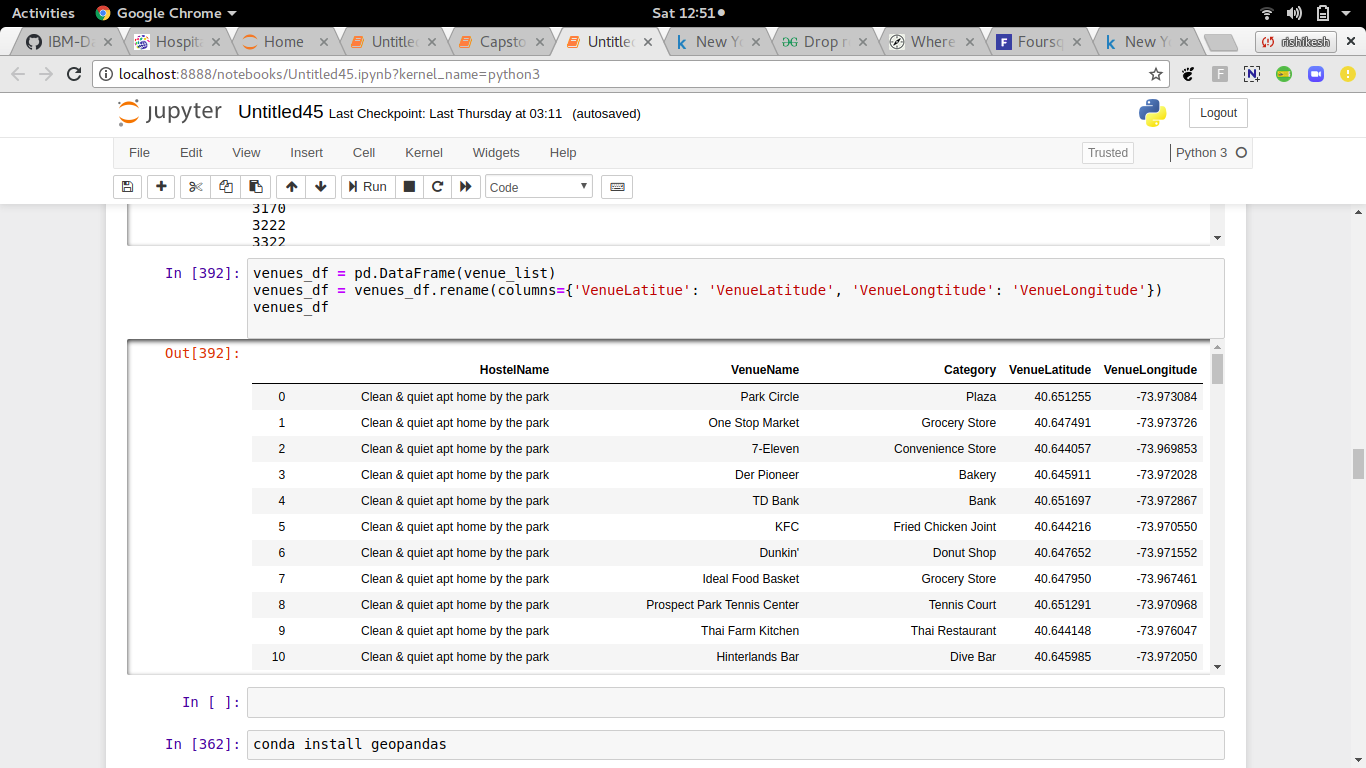


2.2 Hotel Neighborhood:

This dataset contains all the neighborhoods or venues within 500m radius of a Hotel. It has the following columns:

* HostelName: Name of the hostel
* VenueName: Name of the venue
* Category: It is the primary category of the venue, for example, Café, Train Station, Restaurant.
* VenueLatitude, VenueLongitude: Coordinates of the venue.

Below is a snapshot of the dataset:



2.3New York Land Price:

This dataset contains the locality name and the average price of land per square feet of New York neighborhoods from a government website

I have downloaded following excel files:

rollindsales\_bronx.xls

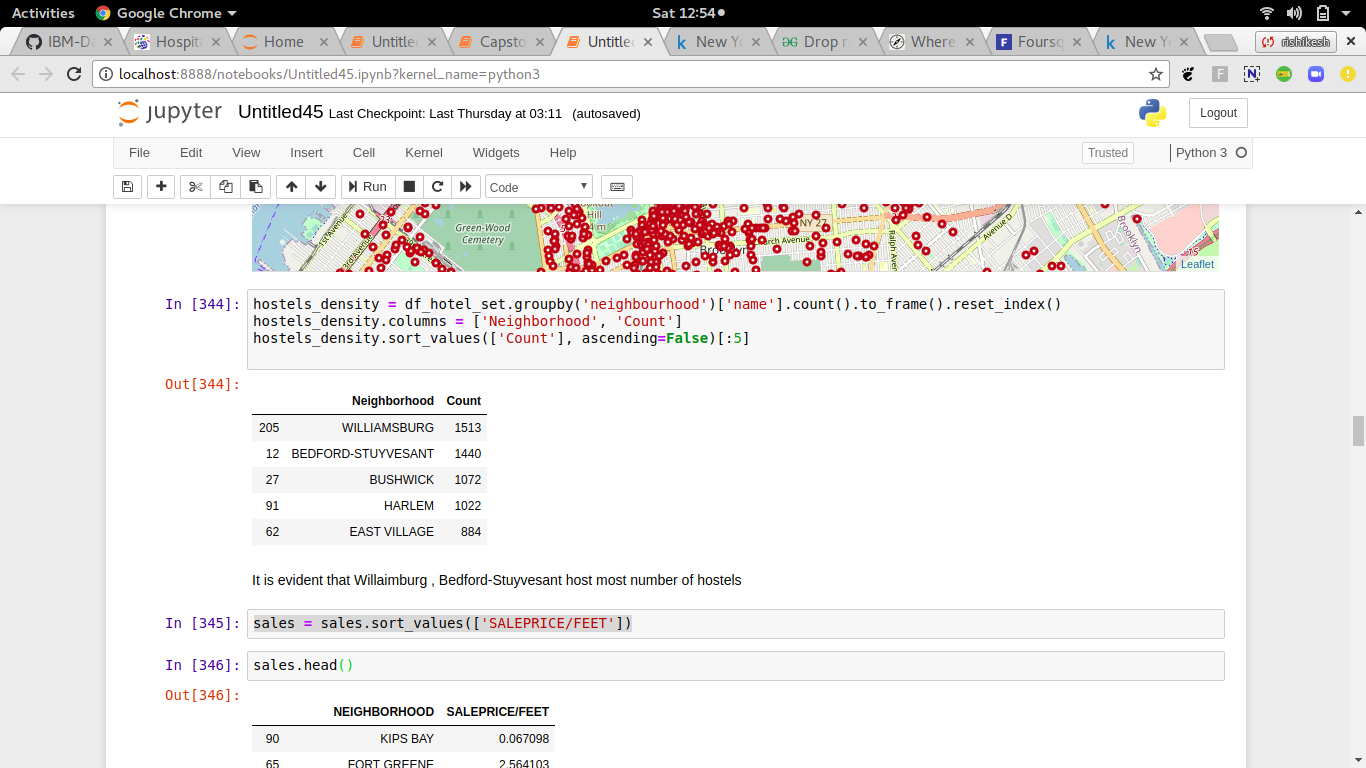
rollindsales\_brooklyn.xls

rollindsales\_statenisland.xls

rollindsales\_queens.xls

rollindsales\_manhattan.xls

Below is a snapshot:



3. Methodology

3.1 Data Collection:

* The New York Hotels dataset is freely available on Kaggle .
* We used Foursquare API to get the venues around the hotels.
* We scraped [https://rollingsales.com](https://utinokati.com/) to get land prices of various neighborhoods in New York.

3.2 Analytic Approach:

I took two approaches in the project.

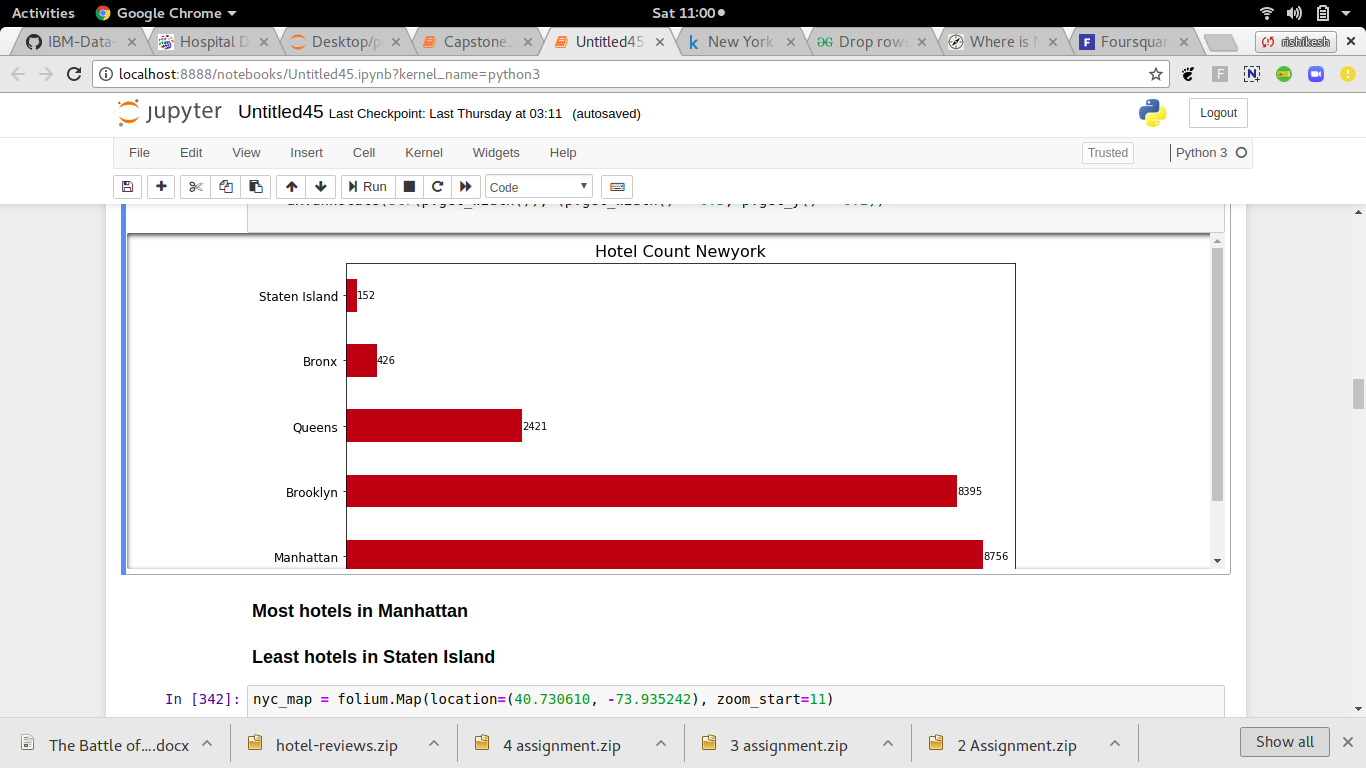
Firstly, I used exploratory data analysis(EDA) to uncover hidden properties of data and provide useful insights to the reader, both future traveller and investor. I used the list of hostels from Hostel dataset and use Foursquare API to get venues around the Hostel. I will then use EDA to explore the neighborhood and how it affects the price of the hostel. I will also use the combined dataset to cluster similar hostels as per pricing and neighborhood.

Secondly, I used prescriptive analytics to help a business person decide a location for new hostel. I will use clustering(K-Means). I combined the above data with the land price for the area in which the Hostel is situated and them develop clustering models to predict where a new hostel should be opened.

4. Analysis

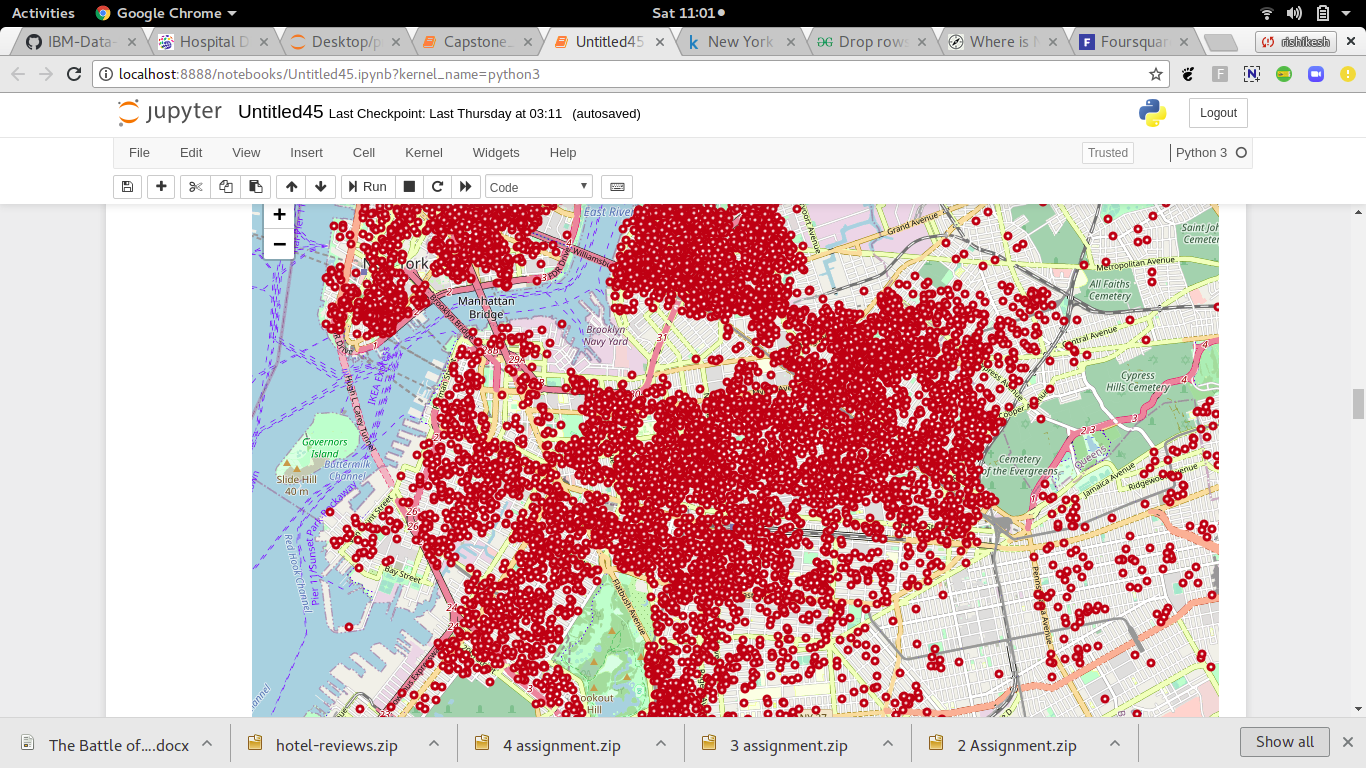
4.1 Exploratory Data Analysis

First we tried to see how many hostels were present in our dataset per neighborhood.

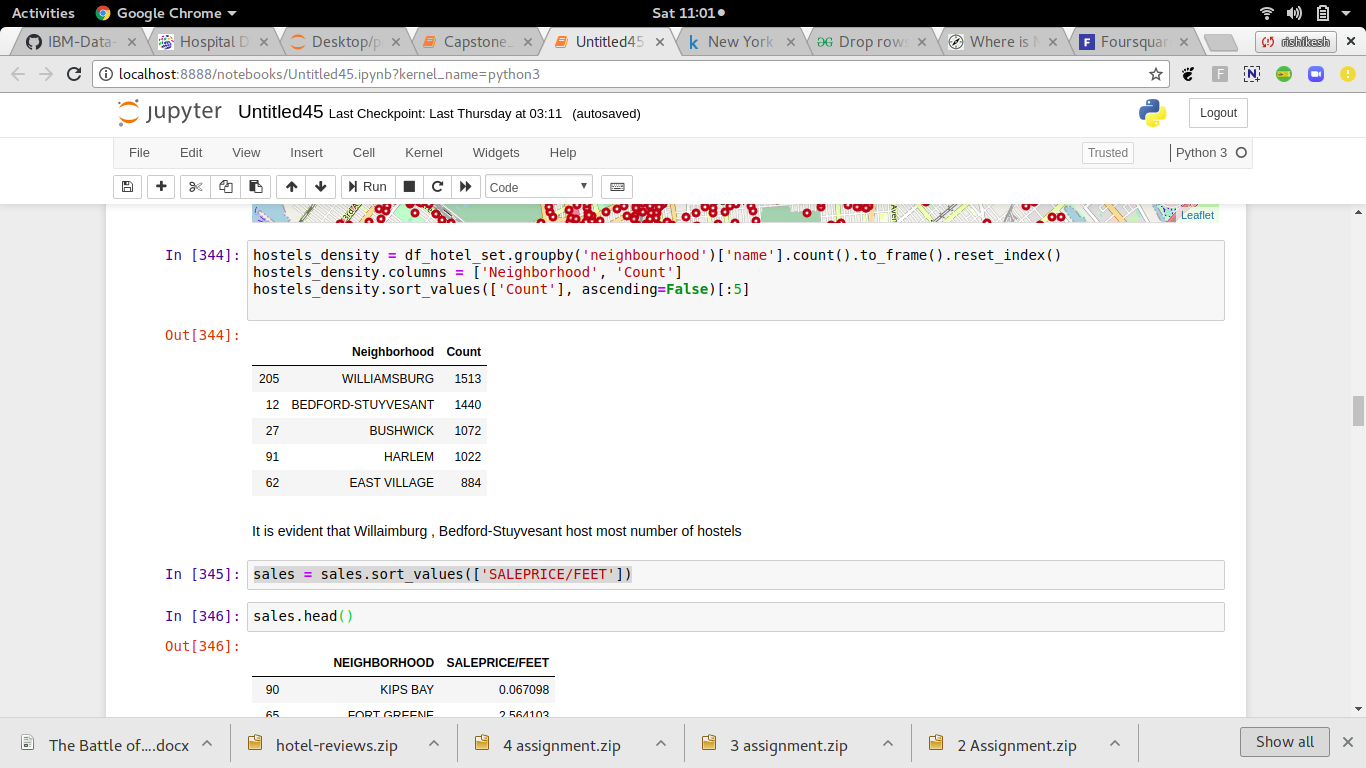


Manhattan has the most number of hotels which is understandable since it is one of the largest Neighborhood in the city and hence might attract a lot of visitors. Also, the two airports in Tokyo make it an important hub in the pacific rim.

Next, we visualized all the hostels on map using Folium and Open Street Maps. Below is the geo-visualization of the hotels in NYC. As we can see, many hostels are located near the centre with density reducing as we move away from it. In the report we have visualized the same map based on multiple criteria and clusters.

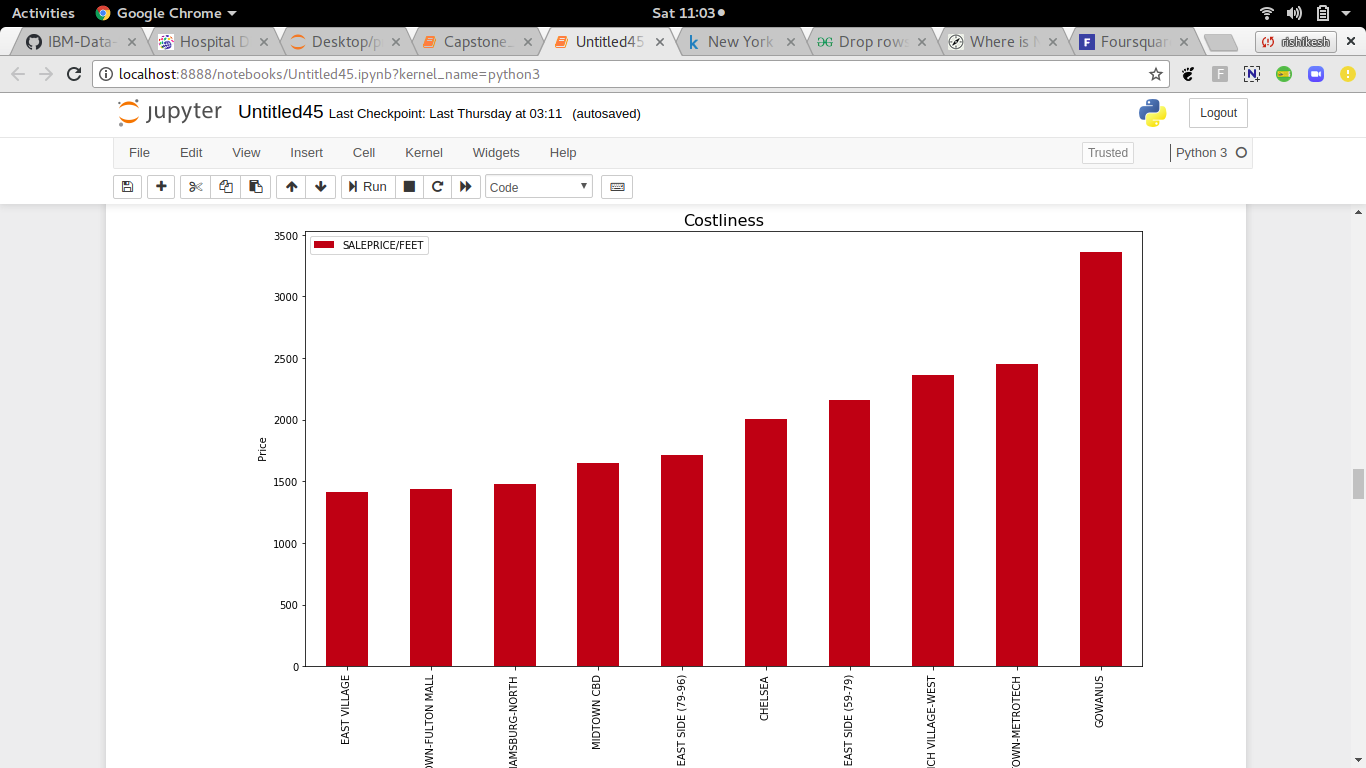


For a person interested in opening a new hostels, it is essential to identify the right area to do so. The main factors to consider while making such a decision is demand, supply and cost. Demand and supply usually go hand-in-hand in such cases i.e. we can say that the more the hostels in a region, the higher the demand. It is evident from the above map that **Manhattan** and **Brooklyn** are host to many hotels. Hotel density is highest in these neighborhoods. However, if we compare the land prices in these regions, they are amongst the most costly localities to buy property.

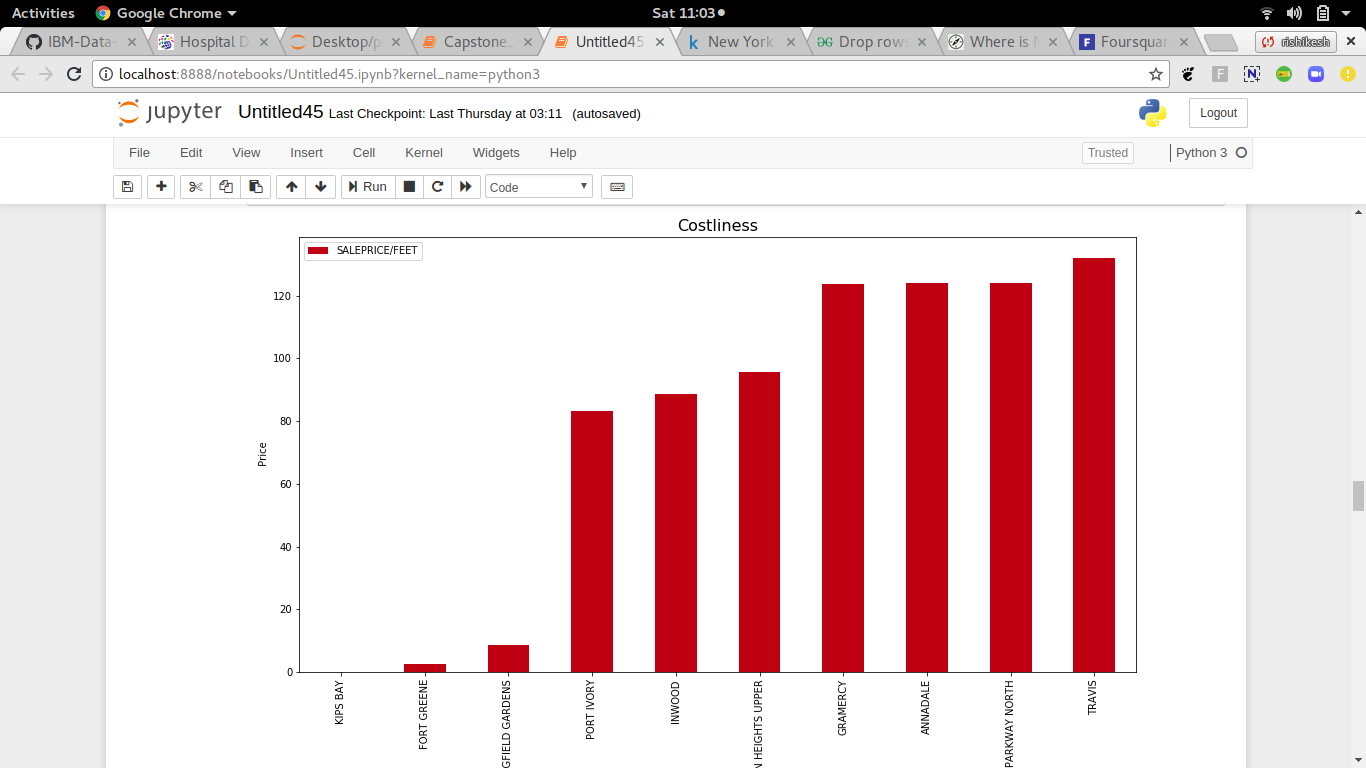


**Williamsburg** seems to be popular since it ranks first in the list of number of hostels and at the same time, it is the 4th cheapest neighborhood in our list.

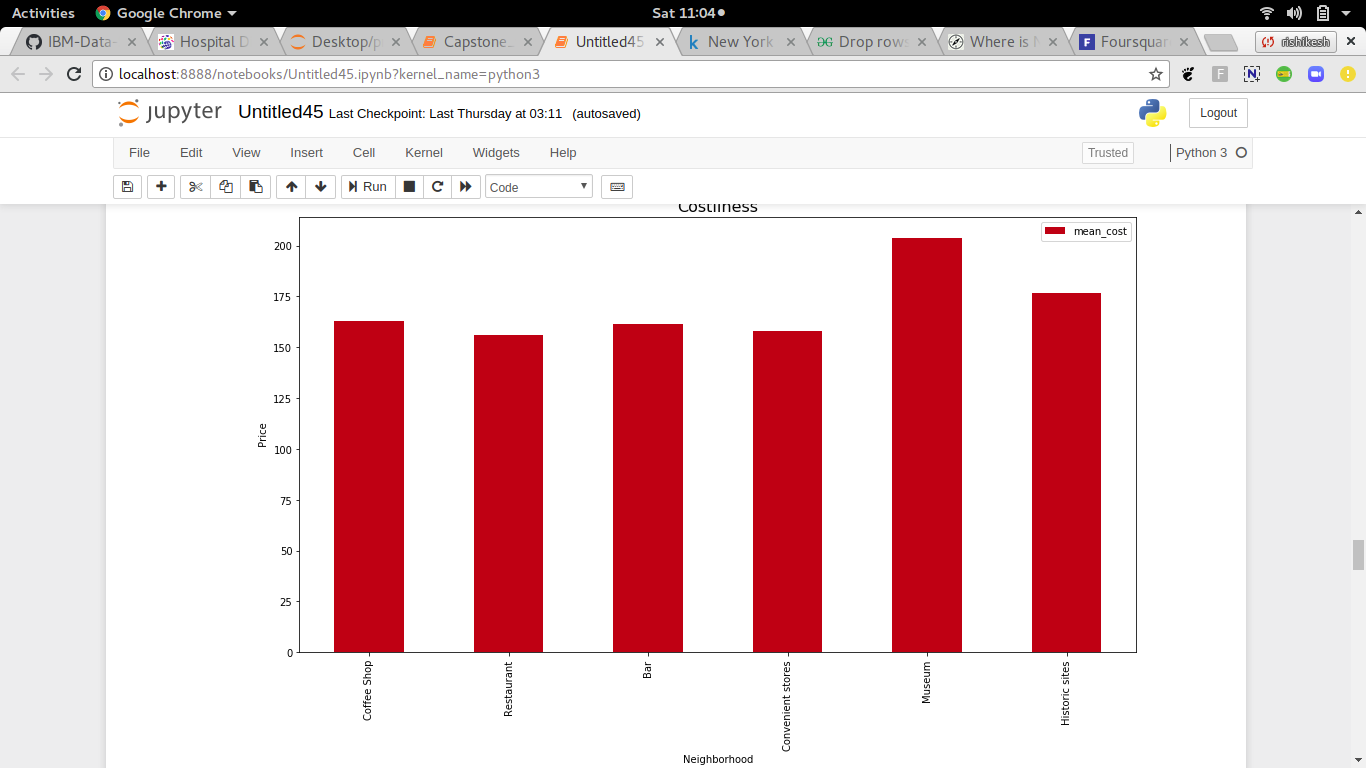
Next, we checked if the distance of a hostel from the city centre has any effect on the price.



Gowanus is the most costly neighbourhood in New York .



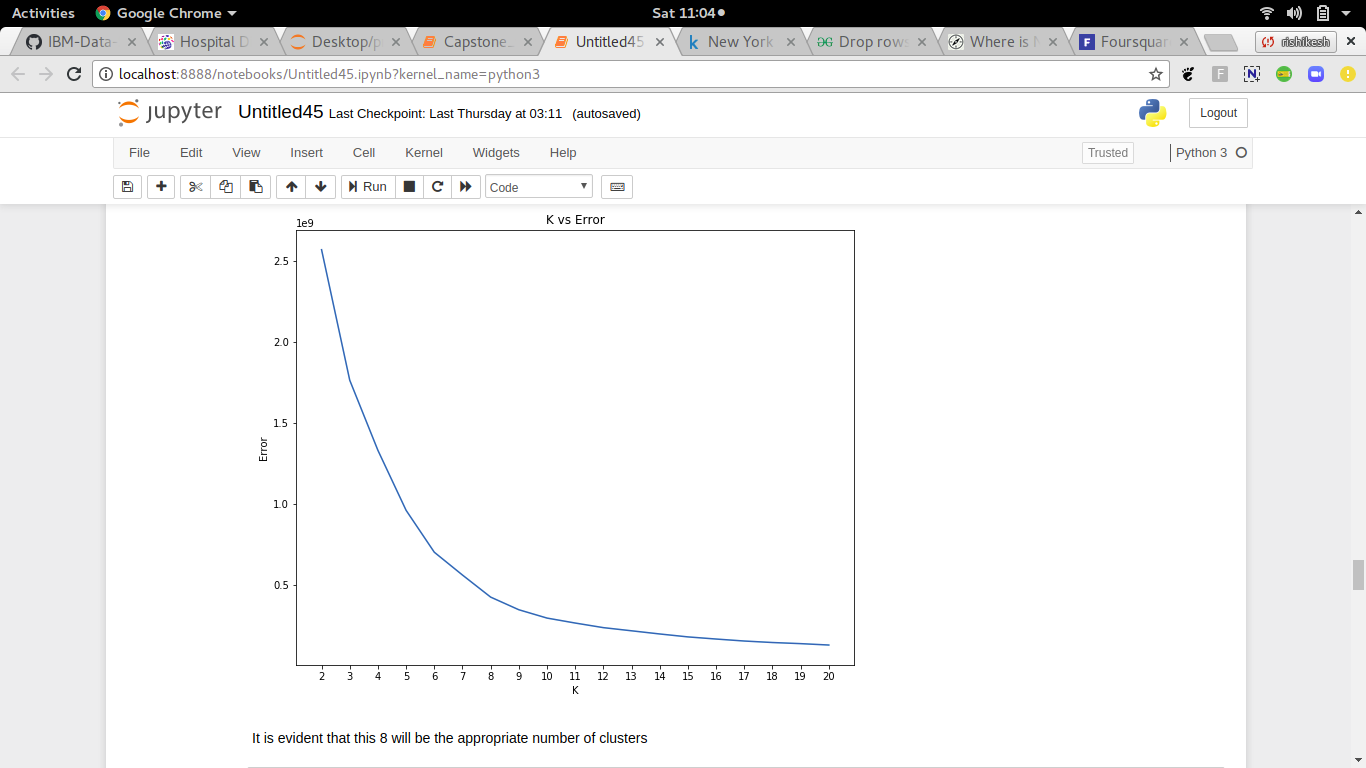
Fort Greene is the least costly neighbourhood , since it is an industrial area. It has bound to be less costly .



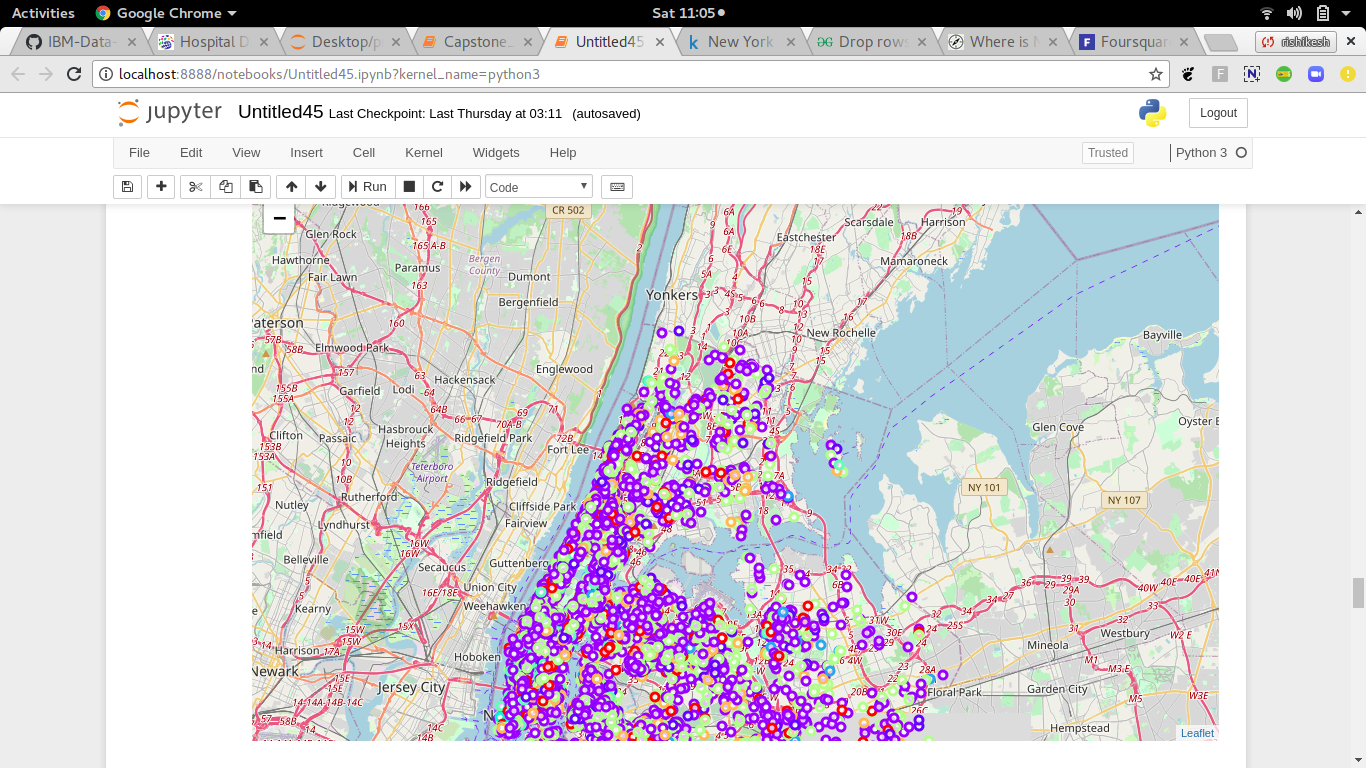
Hotels close to Museums and Historic sites are more costly as they are the tourist attractions in the city.

4.2 Clustering

We performed clustering twice based on different set of parameters. First, we clustered using the different rating scores, distance from city centre, and starting price. We used K-Means clustering algorithm and found out the K by using the elbow method. The K on our case is 8, since the error doesn’t decrease much after this point.



**The important clusters int the point of view of an investor**



**Cluster 0:**

Very High Cost. Close to City centre. Better infrastructure .

**Cluster 1:**

High cost. Near from city centre. High Value for Money score.

**Cluster 2:**

Low Cost. Average overall score between 8.5 and 9. Low atmosphere rating.

**Cluster 3:**

Costliest. Close to city centre. More facilities and staff score.

**Cluster 4:**

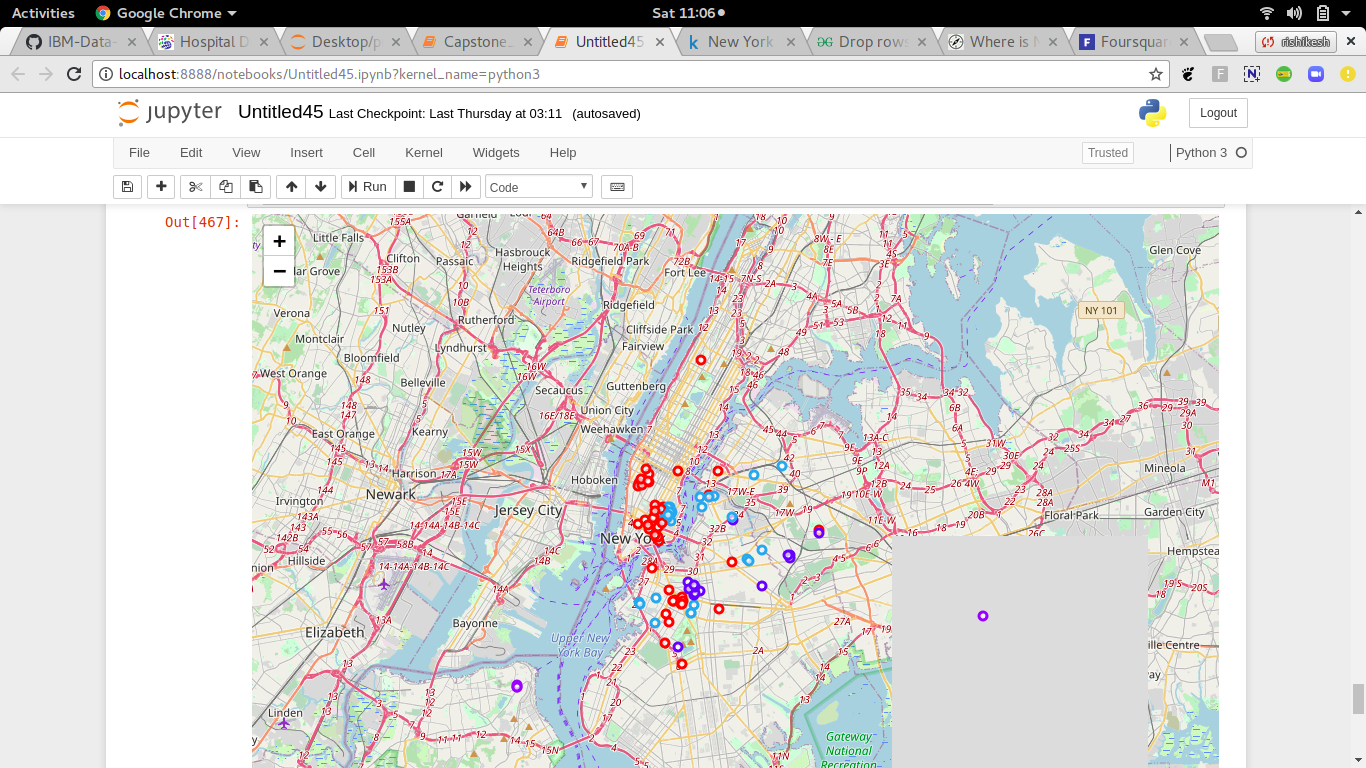
High Cost. All ratings moderately high

**Cluster 6:**

Very low cost. Far away from city centre. High value for money and industrial region .

**Cluster 7 and 5 are more or less the average of these clusters**

Second, we clustered the hostels based on the venues in its vicinity. This time, we fixed K in our K-Means algorithm to be 4 since otherwise each hostel would be assigned a unique cluster which defeats the purpose.



We can examined each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, we can then assign a name to each cluster.

|  |  |
| --- | --- |
| ***Cluster Number*** | ***Cluster Properties*** |
| ***1*** | Restaurants and Coffee Shops |
| ***2*** | Bus Stops and Grocery Stores |
| ***3*** | **Pizza joints, Fast food places** |
| ***4*** | **Bars and Restaurants** |

5. Result

We got a glimpse of the hotel scene in New York and were able to find out some interesting insights which might be useful to travellers as well as people with business interests. Let's summarize our findings:

* Most hostels are located in **WilliamsBurg**.
* Williamsburg seems to be an interesting locality since it is close to Manhattan and 4th cheapest neighborhood.
* The starting price of hostels does not vary much depending on its distance from the city centre.
* Most of the hotels in Gowanus are costly.

* Proximity to a Museum or a historic site positively affects the hostel rating.
* Hostels rated highly for being value for money are comparatively cheaper and are located away from the city centre

6. Discussion

According to the above analysis, opening a hotel in Harlem is the most appropriate option. It already contains the highest number of hotels from our dataset. This means that these wards must attract lot of tourists, no other reason appropriately justifies why they house so many hotels inspite of being very costly. Harlem is a neighbour of Gowanus and is almost 41% cheaper than Gowanus. It also contains decent amount of hostels but not as many as Willaimsburg, which leaves space for new opportunities.

The clusters will help tourists identify alternate hostels in case their hostel is not available for some reason. For example, I wanted to book a spot at Retrometro Backpackers. However, when I went to the website, it said that all the beds were booked and there was no availability during my dates. From our analysis, Retrometro belongs to Cluster 4. So I checked which other hostels are in that cluster.

Some drawbacks of the analyses are that our prescription to new business person’s for opening a new hostel is solely based on neighborhood and land price and not on other factors like how much the business person is willing to invest, what facilities will he provide, how will he price the hostel. Without this data, it is difficult to predict the success of the investment. However, in our analysis, we have ignored this since we don’t have such data and it would be difficult to farm it for a small exploratory study like ours. Hence, our analysis only helps a business person identify a region to open a hostel but doesn’t guarantee its success.

Also, it would have been beneficial if we had additional features such as crime rate in the locality, and average number of tourists in the locality. This would give us a more complete picture of the neighborhood of a hostel resulting in better analysis.

7. Conclusion

In the above study, we explored and analysed various aspects of the hotels scene in New York using data science. We used an existing dataset and combined it with data collected from Foursquare API as well as data scraped from a website. We performed EDA and clustering on these datasets in our pursuit of solutions. We were able to find satisfactory answers to the questions we posed before the study.

The study is based on limited data, but it is nevertheless a significant step in shedding light on the hotel scene in New York city . This study can be repeated easily for other cities of the world.