Predicting Housing Price in New York using the neighborhood data and their nearby venues

Yingxian Estella Yu

May 8, 2019

Abstract:
In this analysis, the following content will be listed

- **Web scraping**: obtain apartment rental price by neighborhood
- **Data wrangling**: clean up the data using python pandas
- **Data analyze**:
  - obtain more neighborhood venue data using FourSquare data base,
  - explore the data set in pandas, matplotlib, etc.
  - explain rental price trend using machine learning skills

I. Introduction

1. Background

New York is such a unique place full of attractions from the financial capital, fashion trends, artistic and historic atmosphere, that words just simply can't describe enough. The only way to know the life in New York is simply to experience it. Yet, city life comes with a lot of dollar signs -- especially in Manhattan. According to the recent National Rent Report (Feb 2019[1]), the rental price in New York for a 1 Bedroom apartment ($2,780) is ranking 2nd across the nation, right behind the crazy San Francisco. What's more, based on the data shown in businessinsider.com [2], the asking rent has drastically increased by 33% in the window from Dec 2009 to July 2017 (in less than 9 years)!

Therefore, it's of special significance to analyze and understand the housing trend in New York. With some simple search, one can easily spot that the housing price in New York is highly correlated with its location. For example, the rent near Soho (average $5,000 - $6,000+) is higher than the Manhattan average by 52% and is certainly pricier than the rent near East Harlem (average $2,000 - $3,000+).

So, to what extend can we predict the rent in Manhattan based on the neighboring venue, an important component that contributes to the vibes in the neighborhood? Is it easier to spot an Italian restaurant than a pizza store at the pricy neighborhood? How much does a school, a mall, a supermarket potentially contribute to the housing price? We are going to figure it out in this report!

2. Project Description

Using data to analyze the following questions:

- **Why** do we want to analyze the housing price in **New York**?
- **How** is the apartment rental price vary by neighborhood?

- If you are planning to move to a new neighborhood, **what** typical venues will you be looking for?
- Do the popular venue & higher end apartment price align?

3. potential target reader

The results and analysis enclosed in this project can be closely relevant to:

- People related to rental activities in New York (landlord, tenant, real estate agent, etc.)
- Business personal: if one plans to open a new business in a certain neighborhood, which neighborhoods are more appropriate, and do the target neighborhoods have relevant venue already?
- or Anyone who's curious about data (like us! :))

II. Data Description

1. Data Source:

- Zumper (National Rent Report: February 2019):
- Rentcafe (Manhattan, NY Rental Market Trends):
- FourSquare API (venue data around each neighborhood):

2. Data Collection:

i. I found the data of **average rental price** based on different neighborhoods in **Manhattan**, which is very well tabulated on the website Rentcafe. I will first scrape and name of each neighborhood and the corresponding rental price from the website using the *Beautiful Soup API*, and then translate the data into *pandas* dataframe.

ii. Based on the name of each neighborhood, the latitude and longitude will be obtained through *GeoPy API*, which is then plotted on a map using *folium*.

iii. The *FourSquare API* was used to get the most common venues around each neighborhood, where the venue data is also added to the folium map.

iv. Using the average rental price, a heat map is plotted showing the distribution of average rent demographically.

3. Use data to solve the problem:

i. Visualize the national rent price across nation (folium, heat map)
ii. Cluster the neighborhood based on rental price, close by venues, etc.
iii. Plot rental price vs. popular venue across all neighborhood, and explore correlations
iv. Analyze a special venue

III. Methodology

1. Data exploration

The complete data analysis process is documented in Jupyter Notebook, which is archived on a GitHub repository. After the data is scraped from the internet, a data frame is generated, listing the name, average rent, and latitude, longitude information for each neighborhood. According to the web scraping result, there are **51 neighborhoods** found in Manhattan in total.

	Neighborhood	Average Rent	loc_lat	loc_lon
0	Marble Hill	1694.0	40.8762983	-73.9104292
1	Inwood	2225.0	40.8692579	-73.9204949
2	Washington Heights	2243.0	40.8401984	-73.9402214
3	Randalls and Wards Islands	2336.0	40.79144785	-73.921023713881
4	East Harlem	3334.0	40.7947222	-73.9425

Fig. 1. Dataframe from web scraping – the name of each neighborhood and the corresponding average rent are scraped from the internet, and the latitude & longitude location are obtained from the Geopy API.

A **folium map** is then generated, showing the location of each neighborhood. In addition, a heat map is generate based on the average rental price, with the red region showing a region with high average rent, and blue indicating a relatively low price.

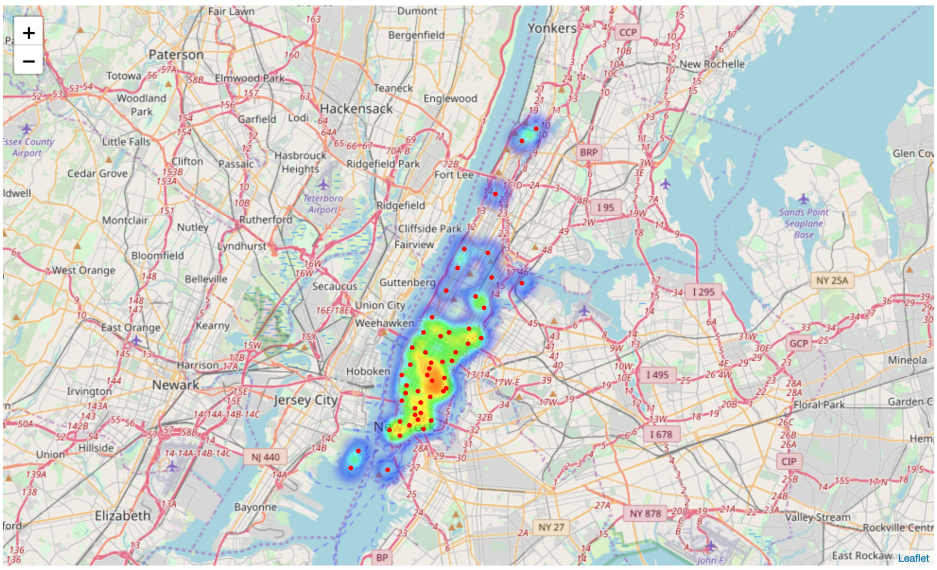

Fig. 2. Folium map plotted based on the dataframe. In addition, a heat map is generated, showing the regions with higher rental price in red and lower price in blue.

In order to get more insights to each of the neighborhood and to find possible correlations between the average price and neighborhood environment, the surrounding venue data was retrieved using the **FourSqare** API in each neighborhood. A searching zone is set centering at the latitude and longitude coordinates of each neighborhood with radius of 500m. The first 100 venue results are retrieved and stored in a dataframe, and the results are shown in Fig. 3.

	Neighborhood	Average_Rent	Venue	lat	lon	Category
0	Marble Hill	1694.0	108 Marblehill	40.876605	-73.909430	Housing Development
1	Marble Hill	1694.0	Marble Hill	40.876111	-73.911111	Neighborhood
2	Marble Hill	1694.0	St. Johns Roman Catholic Church	40.876174	-73.909795	Church
3	Marble Hill	1694.0	CTown Supermarkets	40.876218	-73.908541	Supermarket
4	Marble Hill	1694.0	Wine & Liquors	40.874535	-73.909832	Wine Shop

Fig. 3. Using the FourSquare API, the venues in each neighborhood are fetched. The name of each venue, geological location and its category are cleaned and tabulated.

Although 100 venues are requested in each region, the number of venues available in ea ch neighborhood might vary. Thus, the actual number of obtained venues are reported and plotte d below. **4695 venues** are found in total, of which **415 of them are in unique** categories.

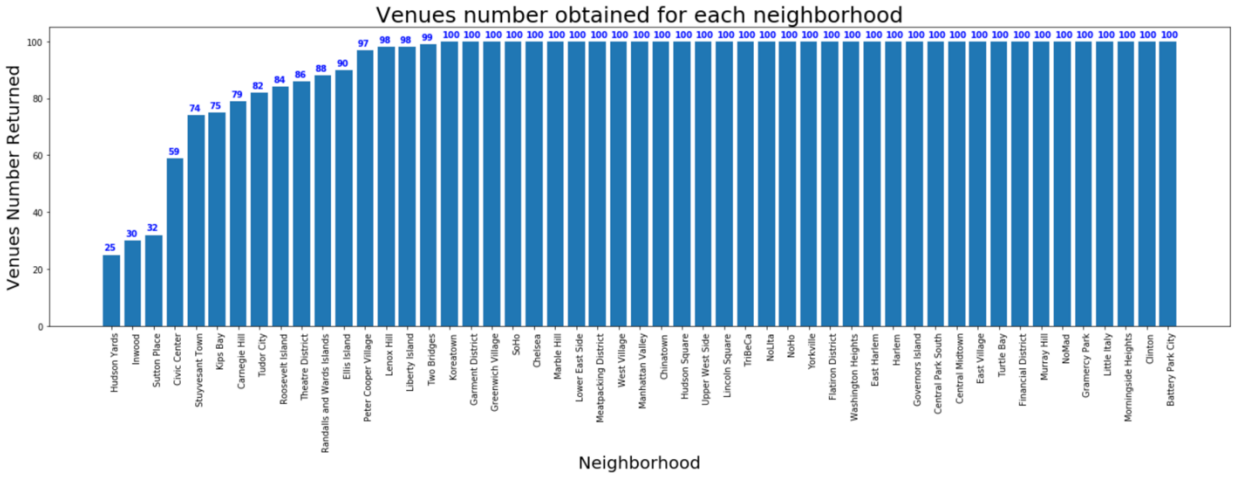

Fig. 4. Total number of venues returned in each neighborhood varies. With a limit of 100 venues set across all neighborhoods, some of the neighborhoods returns venues smaller than the target limit.

The venues are then one-hot coded based on their categories, and the number of venues in each category are group by the neighborhoods (Fig. 5.).

	Manhattan_Neighborhood	Average_Rent	lat	lon	ATM	Accessories Store	Acupuncturist	Advertising Agency	African Restaurant	Alternative Healer	...	Warehouse	Wa Sh
0	Battery Park City	5603.0	40.7110166	-74.0169369	0	0	0	0	0	0	...	0	
1	Carnegie Hill	4271.0	40.7841972	-73.954339	0	0	0	0	0	0	...	0	
2	Central Midtown	3913.0	40.7622684	-73.9795443	0	0	0	0	0	0	...	0	
3	Central Park South	3786.0	40.7646364	-73.9737661	0	0	0	0	0	0	...	0	
4	Chelsea	4359.0	40.7464906	-74.0015283	0	0	0	0	0	0	...	0	

5 rows × 419 columns

Fig. 5. One-hot coding for each unique category, and results are group by neighborhood.

Furthermore, the most common venues in Manhattan are found, and the top 50 venue categories are shown in a bar chart. Although 400+ unique categories are returned from the API, the top 100 categories are in fact covering more than 70% of the total number of venues. Therefore, the top 100 categories will be used for the linear regression analysis in the following section.

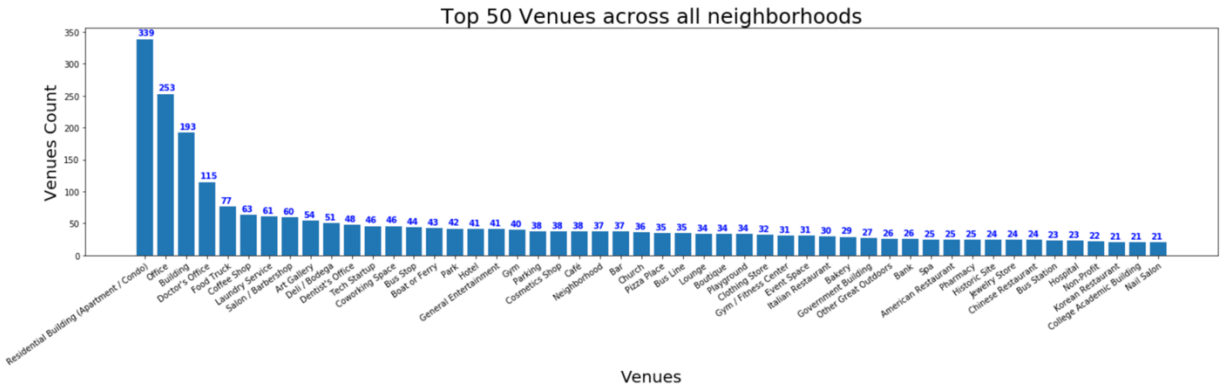

Fig. 6. The popularity of each unique venue type is obtained, and the first 50 most popular venues in Manhattan are plotted

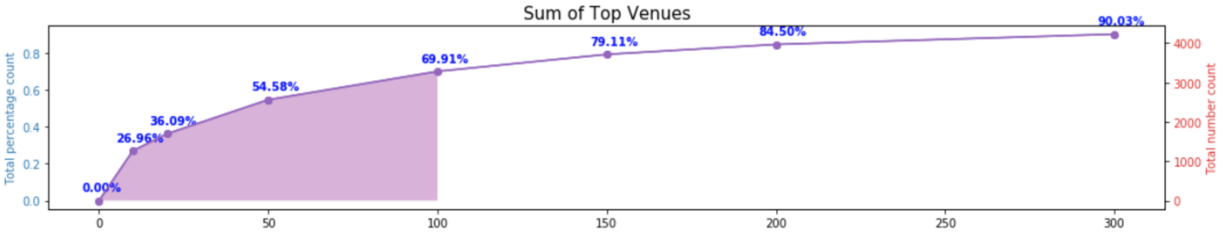

Fig. 7. Although 400+ unique venues are returned, the popularities vary a lot across category. Taking into account the top 100 unique categories can help us to cover over 70% of the total venues return.

2. Simple linear regression

Considering that the venue category data is very sparse (many "0"s in the one-hot data frame), and that the venues may have intra-correlations among one anther (e.g. Chinese/Korean/Italian restaurants are all under a superset of restaurants), multivariant linear regression may not provide accurate regression result. Simple 1D linear regression is performed, regressing the number of appearances of each of the top 100 venue categories across all 51 neighborhood rental prices.

Although the resultant R^2 scores are still fairly small, some obvious peaks are observed in certain categories. This shows that **some regions** (e.g. tech startups, gyms, American restaurants, etc.) **influences the rental prices more than others**.

Furthermore the slope of the regression results are also plotted for the top 100 most popular venue categories, showing that some categories are showing a much higher positive/negative influence compared to other categories. For example, the neighborhood having **more pharmacies** nearby might have a tendency of a **higher rental price**.

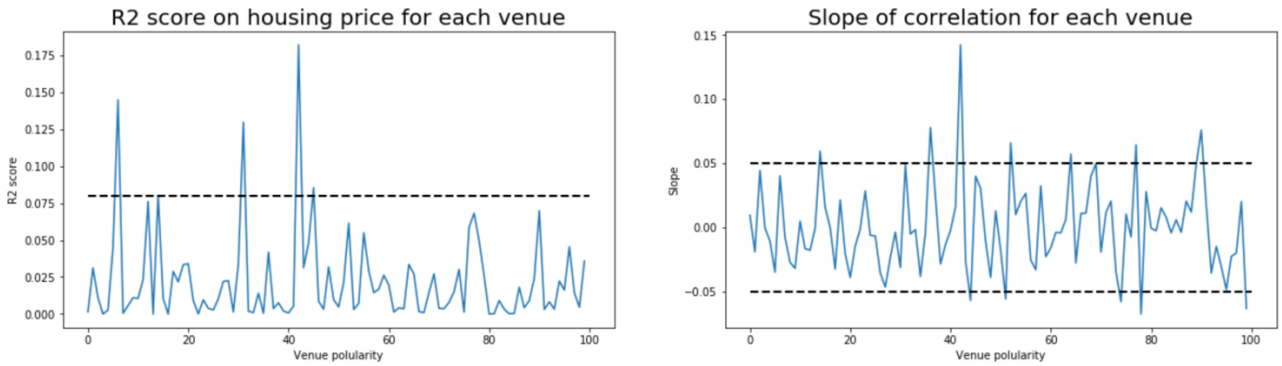

Fig. 8. Simple linear regression on each venue categories across all neighborhood. Although the R^2 score are small, some helpful insights can be obtained. For example, some categories are more likely to contribute to the price variation than other categories, and some positive/negative influences are more impactful than the other.

3. Kmeans clustering

We then perform a **K-means clustering** (unsupervised machine learning method) to categories the different neighborhoods based on the most common venue types. With the top 5 most common venues generated for each neighborhood, the **elbow method** was used in order to determine the appropriate number of clusters.

	Manhattan_Neighborhood	Average_Rent	lat	lon	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Battery Park City	5603.0	40.7110166	-74.0169369	Residential Building (Apartment / Condo)	Building	Office	Park	Gym
1	Carnegie Hill	4271.0	40.7841972	-73.954339	Doctor's Office	Residential Building (Apartment / Condo)	Building	School	Church
2	Central Midtown	3913.0	40.7622684	-73.9795443	Office	Food Truck	Event Space	Building	Convention Center
3	Central Park South	3786.0	40.7646364	-73.9737661	Office	Food Truck	Building	Hotel Bar	Cosmetics Shop
4	Chelsea	4359.0	40.7464906	-74.0015283	Residential Building (Apartment / Condo)	Laundry Service	Building	Doctor's Office	Office

Fig. 9. Top 5 most common venues in each venue.

The number of clusters ranging from 1 to 51 (the total number of neighborhoods) are tried, and the summed square error from all the points to their corresponding centers are compared. Although a rather smooth curve was obtained, we can see that the "elbow" is somewhere between 2-4. In the following categorization analysis, the **number of clusters are set to 4**.

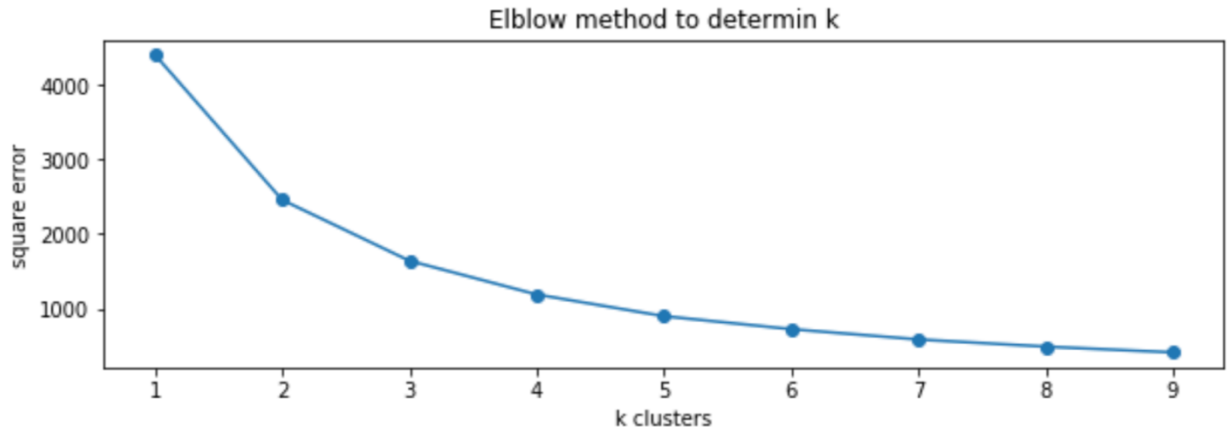

Fig. 10. Elbow method can be used to determine the number of clusters *k*.

	Cluster Labels	Manhattan_Neighborhood	Average_Rent	lat	lon	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
24	0	Liberty Island	3718.0	40.6897882	-74.0451182096341	Boat or Ferry	Monument / Landmark	Harbor / Marina	History Museum	Scenic Lookout
10	0	Ellis Island	3718.0	40.6992693	-74.0393071	Boat or Ferry	Harbor / Marina	History Museum	Monument / Landmark	American Restaurant
0	1	Battery Park City	5603.0	40.7110166	-74.0169369	Residential Building (Apartment / Condo)	Building	Office	Park	Gym
23	1	Lenox Hill	4332.0	40.7664366	-73.9590168	Residential Building (Apartment / Condo)	Laundry Service	Coffee Shop	Doctor's Office	Salon / Barbershop
49	1	West Village	4524.0	40.7352405	-74.0046133971969	Boutique	Residential Building (Apartment / Condo)	Clothing Store	Cosmetics Shop	Building

Fig. 11. Dataframe showing the name, location and most common venues of each neighborhood, with the cluster label appended.

After clustering is performed, the cluster labels are put back to the dataframe, and the top 5 most common venues are shown in the bar chart below. Based on the numbers and categories of the popular venues, we can provide names for each cluster: **1)** Island/Coastal region, **2)** Multifunctional region, **3)** residential region, and **4)** business region.

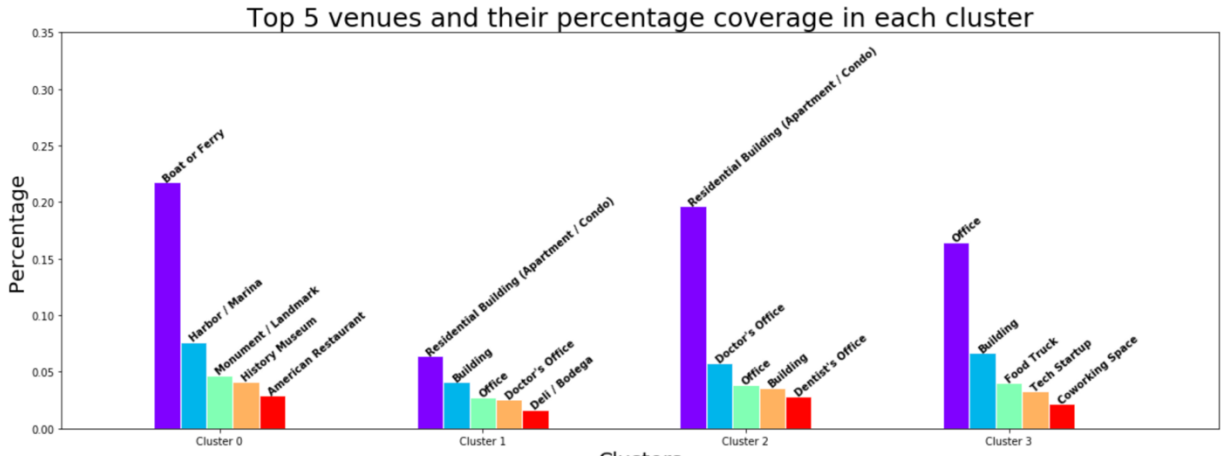

Fig. 12. A summary of the top 5 most common venues in each cluster. We can then provide each cluster with its name based on these characteristics.

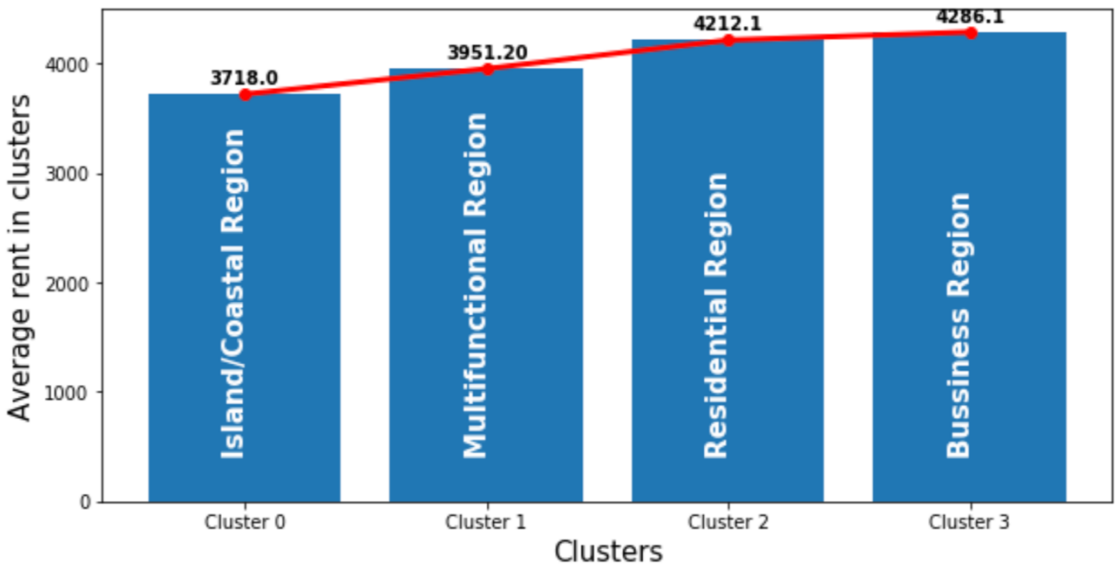

Fig. 13. The clustered region also shows a slight increasing trend of average rental price, with the relatively remote island regions having the lowest rental price across the cluster, and the busy business regions with the highest average rental price.

The average rental price is also calculated for each cluster. The clustering results are rather reasonable, predicting that the most remote region (1. The island region) has a lower average rental price, and the most commercialized region (4. The business region) has the highest price among all. The results also make sense demographically as well. As can be seen form the folium map showing the clustered neighborhoods, islands are categorized as a unique cluster, business regions' locations agree with our common sense, and multifunctional regions mostly locate in between residential and business regions.

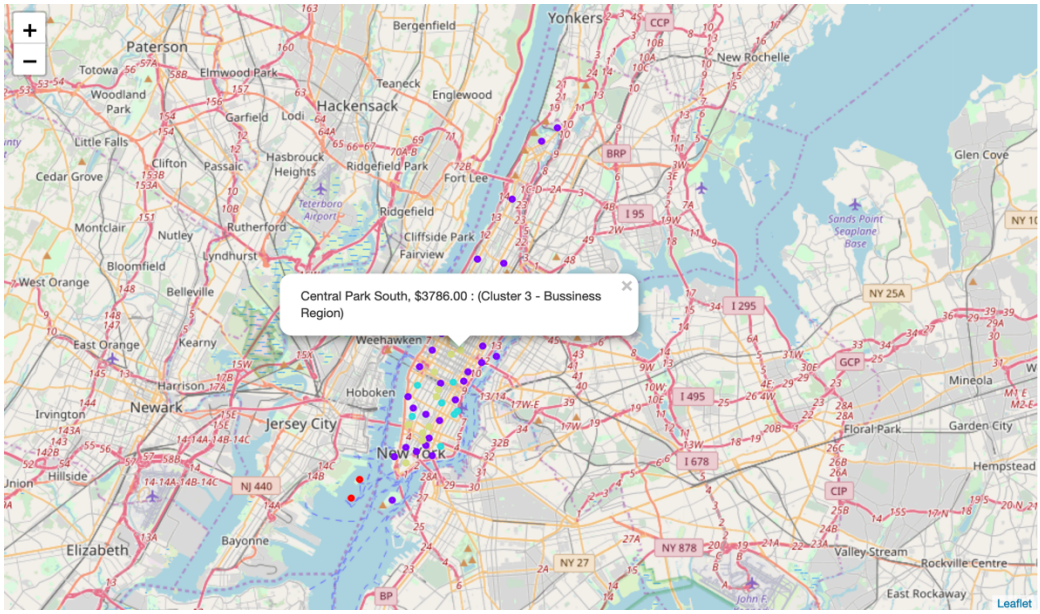

Fig. 14. The clustered neighborhoods are shown on folium map, with labels indicating the neighborhood name, average rental price, and the cluster it belongs to. For example, Central Park South will be considered as a business region (based on the surrounding venues).

IV. Result and Discussion

As mentioned before, **New York** is a big city with a **high population density in a narrow area**. Different aspects of analysis can be made across these 51 districts and can vary drastically in each analysis. As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield high quality results for this metropole.

I used the Kmeans algorithm as part of this clustering study. Since the venues are very packed and diversed in almost all districts, a sharp elbow transition is not obvious from the plot generated. Based on the figure, though, I set the optimum k value to 4. Surprisingly, the clustered results make sense, and the 51 districts can be group to **4 distinct regions** -- **1)** Island/Coastal region, **2)** Multifunctional region, **3)** residential region, and **4)** business region. These regions made sense geographically as well, as shown in the folium map -- with the 2 islands marked as the "Island region", the famous tourist districts (e.g. SoHo, Financial district, etc.) marked as "business Region", etc.

However, in each of the 51 districts, only 1 geological coordinate was used. For more detailed and accurate guidance, the data set can be expanded, and the details of the neighborhood or street can also be drilled.

I also performed a simple regression analysis on each unique type of venue in the data set, which though sparse, provided some of the guidelines for predicting the positive/negative effects in rental price. Since the venues covered in each district vary a lot, the 1D analysis does not provide prediction with high confidence. Yet, it shows, for example, a district with more **jewelry stores** generally have a **higher rental price**. In future studies, these analyses can be performed in a cluster basis instead, since the districts in the same cluster tend to share more similar venues. Furthermore, one can relate the rental price with time, and plot the price trend throughout different period of time in the year as well.

I ended the study by visualizing the data and clustering information on the New York map on folium. In future studies, web or app applications can be carried out to direct investors.

V. Conclusion

As a result, this study can be useful for a person looking for a new rental place, as not only the rental price and neighborhood venues have certain correlation, but the nearby venues also determine whether a place is ideal to live in. For this reason, people can achieve better outcomes through their access to the platforms where such information is provided.

Furthermore, it can also benefit for business personals who are looking for apartments or stores to invest in. Not to mention this type of analytic methods can also be generalized to analyze data sets in many different cities across the world.