The Battle of Neighborhoods: Searching for Maximum Return for Rental Properties in New York City

A multi-facet analysis paired with machine learning to segment neighborhoods and find the best rentals in New York City

Troy Brommenschenkel

8/10/20

The Coursera Applied Data Science Capstone Project

The Battle of Neighborhoods: Searching for Maximum Return for Rental Properties in New York City

A. Introduction

1. Background

New York City is the most populous city in the United States with a population of 8,253,213 and a density of 27,274.3 people per square mile. New York City (NYC) has been described as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports, and is the most photographed city in the world. With a description like that, who wouldn't want to live there?

2. Business Problem

A real estate investment firm is looking to begin purchasing properties in New York City to be used as rental properties. However, they have had mixed luck with real estate agencies leading them to what they think are the best opportunities across the five boroughs. The investment firm would like an analysis to indicate where the most desirable rental properties are located for the lowest price.

- As many unique venues within walking distance (because owning a car in NYC is too expensive).
- What does the crime rate look like by borough and neighborhood?
- Rental property median price.

Therefore, the project goal is to figure out the best locations to rent in New York City to maximize the return on investment to have access to the most unique venue types within walking distance.

3. Target Audience

This study is of interest to **everyone** who may have a curiosity in relocating to or moving within New York City to maximize their return for rent paid. The project will also be of interest to **business owners** seeking to locate their business in an area of little competition or high diversity. **Real estate investors** will additionally be interested who may wonder how data science can provide insight into neighborhoods that will maximize their investment.

B. Data Description

1. Data Requirements and Collection

For this project we will need historical rental prices for properties in New York and historical crime data for incidents in each of the boroughs. We can also leverage Foursquare Location data to compare neighborhoods with venue locations and their respective ratings. The following are data sources that were used for this project:

- **Zillow Observed Rent Index (ZORI):** The most up-to-date median rental prices for all U.S. cities segmented by zip code.
- **New York City Borough and Neighborhoods:** JSON data containing the 5 boroughs and 306 neighborhoods classified as New York City with latitude and longitude coordinates.
- **NYPD Shooting Crime Data:** Data of shootings with geo locations retrieved from NYC OpenData.
- **New York ZCTA to PUMA Data:** Mapping of postal codes within New York City to neighborhood names.
- **New York ZCTA to Borough Data:** Mapping of postal codes within New York City to each of the five boroughs.
- **Foursquare Venue Data:** The most popular venues of a given neighborhood in New York City. This information is stored inside **Foursquare Location Data**, and we will use the **Foursquare API** to access it.

2. Data Cleaning and Extraction

- The first dataset is a CSV file retrieved from Zillow containing 1743 rows and 94 columns. The data is collected each month with the median rental price separated by zip code and month for the entire United States. We will focus on the rows containing only zip codes within New York City to understand neighborhood rental prices.
- The second dataset is a JSON file containing geolocation coordinates matching all of the zip codes to the corresponding neighborhoods and boroughs.
- The third dataset is a CSV file retrieved from NYC OpenData containing all shooting crimes within New York City. The data contains information such as geo location and a flag to denote if the shooting was deemed fatal.
- The fourth dataset is a CSV file retrieved from Baruch College containing 212 rows and 2 columns mapping postal zip codes to Public Use Microdata Areas (PUMA), also known as neighborhoods.
- The fifth dataset is a CSV file containing all postal zip codes with assignment to each of the five boroughs within NYC.
- The sixth dataset is stored within Foursquare Location Data and will be accessed through the API. We will utilize postal coordinates to retrieve venues, categories, and their ratings. We will then use this data to cluster the unique categories to each of the neighborhoods to find the best mix with minimal distance between them.

The first and second datasets will be used to analyze median rent amounts for all boroughs and neighborhoods of New York City. Then we can use the New York Police Department (NYPD) shooting data to find the safest areas within New York. The fourth and fifth data sources will be combined to create a list of all available postal zip codes with both neighborhood and borough assignments to be matched with the first data source. Finally, we will use the coordinates and foursquare credentials to access the sixth data source through its API and retrieve popular venues along with their details. The venue frequency in each neighborhood will be the features of the clustering model.

C. Methodology

1. Analytic Approach

First we begin by analyzing the crime data source to get a picture of where we can focus the analysis to reduce the amount of data as well as narrow down the best area to research rental properties. Then, we approach the problem by utilizing the **k-Means** clustering technique. This approach enables the audience to see how similar neighborhoods cluster together based on the types of places that reside there. We can examine each cluster and determine the venue categories that establish each cluster.
K-Means is a common machine learning algorithm used to cluster data points based on similar characteristics. The algorithm is fast and efficient for medium and large-sized data sets and can be deployed to discover insights from unlabeled data quickly.

2. Exploratory Data Analysis

- Shooting Crime Data

We begin by analyzing the data about shooting crimes within NYC. The data extends historically to 2006, but we filter it to everything since 2014 to match the historical rental data we have. Each shooting is represented by a unique incident key and contains location, date, borough, and if the shooting was fatal.

	KEY	DATE	BORO	STATISTICAL_MURDER_FLAG	Latitude	Longitude	Lon_Lat	LAW_CAT_CD	CRIME
0	201575314	2019-08-23	QUEENS	False	40.697805	-73.808141	POINT (-73.80814071699996 40.697805308000056)	F	SHOOTING
1	205748546	2019-11-27	BRONX	False	40.818700	-73.918571	POINT (-73.91857061799993 40.81869973000005)	F	SHOOTING
2	193118596	2019-02-02	MANHATTAN	False	40.791916	-73.945480	POINT (-73.94547965999999 40.791916091000076)	F	SHOOTING
3	204192600	2019-10-24	STATEN ISLAND	True	40.638064	-74.166108	POINT (-74.16610830199996 40.63806398200006)	F	SHOOTING
4	201483468	2019-08-22	BRONX	False	40.854547	-73.913339	POINT (-73.91333944399999 40.85454734900003)	F	SHOOTING

Figure 1. First five rows of shooting data set

After filtering, the resulting data now contains 8,935 rows. In order to get a better perspective, we visualize each shooting on a map depicting the fatality flag with red. As we can see there are much fewer shootings in Staten Island, but is that a true representation based on the population?

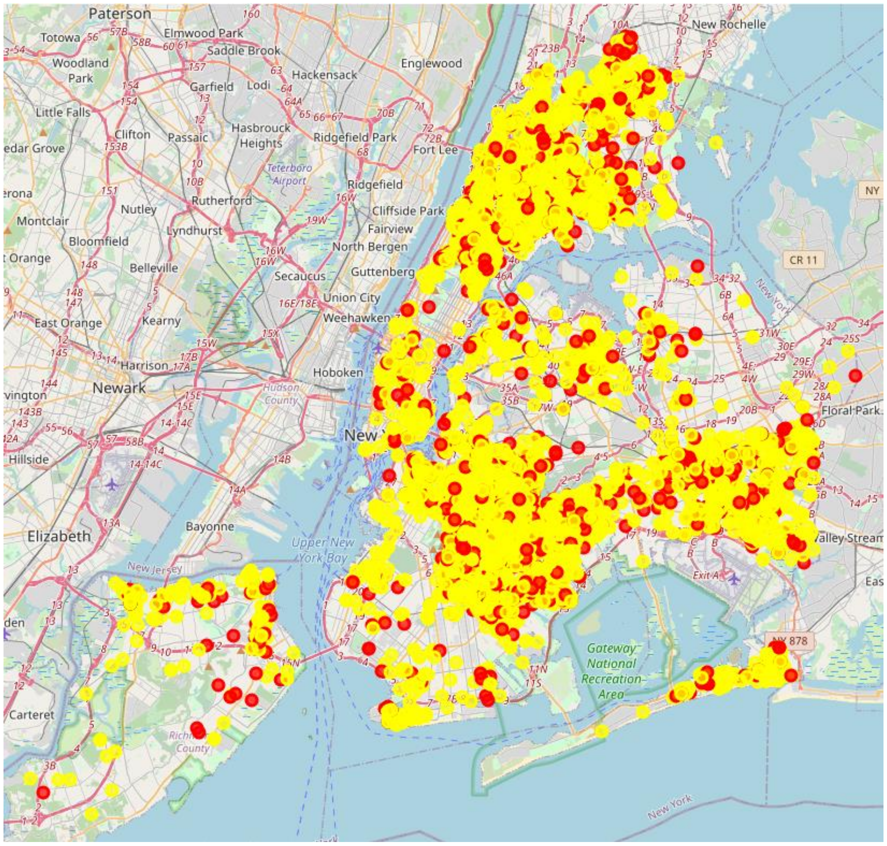

Figure 2. Map of all shootings since 2014 with red showing fatalities

To prevent drawing incorrect conclusions, we need to understand what the difference in population and area is for each borough. We engineer new metrics for crime density and murder density by dividing the number of crimes by the population density and then rank each borough by year.

	Year	BORO	KEY	STATISTICAL_MURDER_FLAG	Population	Area	Pop_dens	CrimeDens	MurderDens	CrimeRank	MurderRank
5	2019	BRONX	49081	266	1438000	42.47	33859.194726	1.449562	0.007856	3	2
6	2020	BRONX	33229	505	1438000	42.47	33859.194726	0.981388	0.014915	3	2
12	2019	BROOKLYN	58660	372	2601000	69.50	37424.460432	1.567424	0.009940	2	1
13	2020	BROOKLYN	39067	809	2601000	69.50	37424.460432	1.043889	0.021617	2	1
19	2019	MANHATTAN	54060	145	1632000	22.82	71516.213848	0.755912	0.002028	5	5
20	2020	MANHATTAN	33529	274	1632000	22.82	71516.213848	0.468831	0.003831	5	5
26	2019	QUEENS	44725	158	2299000	108.10	21267.345051	2.102989	0.007429	1	3
27	2020	QUEENS	30283	302	2299000	108.10	21267.345051	1.423920	0.014200	1	3
33	2019	STATEN ISLAND	9058	26	474101	58.69	8078.054183	1.121310	0.003219	4	4
34	2020	STATEN ISLAND	6247	52	474101	58.69	8078.054183	0.773330	0.006437	4	4

Figure 3. Boroughs ranked by crime and murder density

From this information we can discern Manhattan has the lowest crime density when compared to the number of people living there (figure 3). We can see Staten Island has the lowest in terms of actual numbers, but it also has the lowest population compared to the other boroughs, which is revealed by our engineered metric.

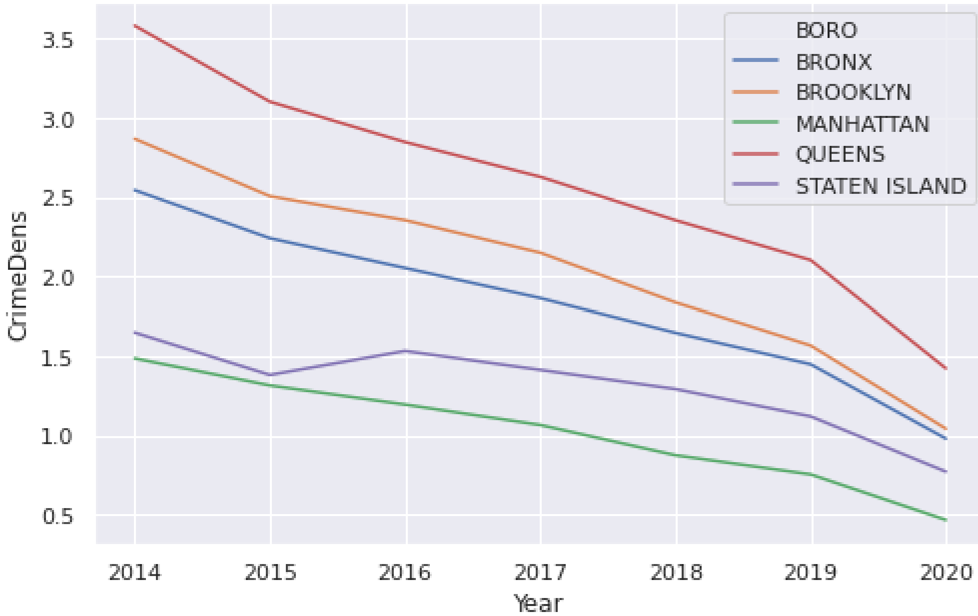

Figure 4. Line plot of crime density by year and borough

With this information, we will focus the remainder of the analysis on finding the best rental neighborhoods in Manhattan as they are deemed the safest. Unfortunately, Manhattan also contains the highest median rent prices in all of New York.

- Neighborhood Analysis

```
(62, 99)
```

	RegionID	ZipCode	SizeRank	CityState	2014-01	2014-02	2014-03	2014-04	2014-05	2014-06	...	2021-02	2021-03	2021-04	2021-05	2021-06	City	State	Zip	Neighborhood	Borough
0	61639	10025	1	New York, NY	2889.0	2904.0	2919.0	2935.0	2949.0	2964.0	...	2883.0	2865.0	2848.0	2834.0	2820.0	New York	NY	10025	Upper West Side & West Side	Manhattan
1	61637	10023	3	New York, NY	3014.0	3024.0	3033.0	3043.0	3052.0	3061.0	...	2881.0	2861.0	2841.0	2824.0	2807.0	New York	NY	10023	Upper West Side & West Side	Manhattan
2	61616	10002	7	New York, NY	2699.0	2715.0	2730.0	2746.0	2761.0	2777.0	...	2734.0	2709.0	2685.0	2664.0	2643.0	New York	NY	10002	Chinatown & Lower East Side	Manhattan
3	62037	11226	11	New York, NY	1691.0	1694.0	1696.0	1698.0	1701.0	1704.0	...	1967.0	1957.0	1947.0	1938.0	1928.0	New York	NY	11226	Flatbush & Midwood	Brooklyn
4	61630	10016	16	New York, NY	3120.0	3132.0	3144.0	3156.0	3168.0	3181.0	...	2977.0	2951.0	2925.0	2903.0	2881.0	New York	NY	10016	Murray Hill, Gramercy & Stuyvesant Town	Manhattan

Figure 5. First five rows of the merged location and rental data

Merging the **New York City Borough and Neighborhoods** location data with the **Zillow Rent Index**, we find that Zillow data only contains 62 out of the 95 ZIP codes within NYC. To better understand how the boroughs compare to each other in terms of rent prices, we visualize them as a line chart. As we can clearly see, there is a separation in terms of the median rent price range between all five boroughs (figure 4). Now we can overlay both the median rent prices and match them against the shooting crime data to visualize if

lower rent is also tied to higher crime rates using the Folium module. Based on the map, we can confirm that Manhattan, in fact, does contain less shootings than many of the other boroughs and neighborhoods.

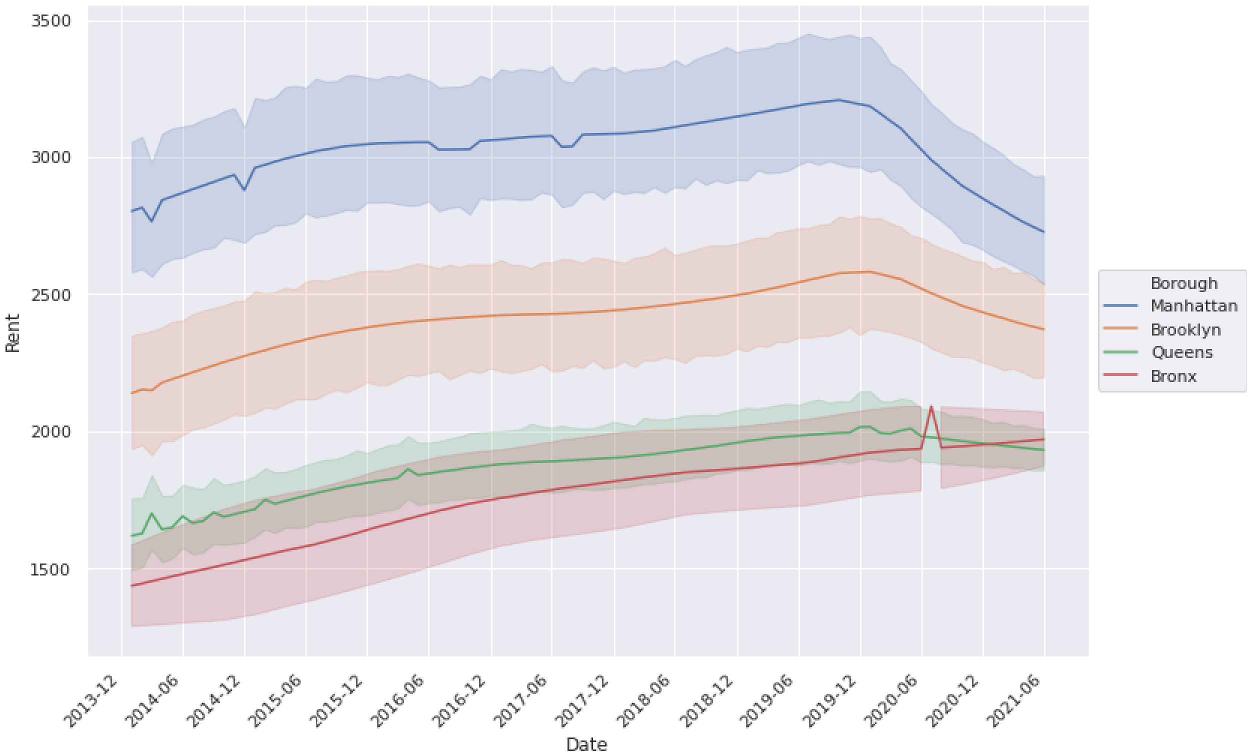

Figure 6. Median rental prices by date and borough

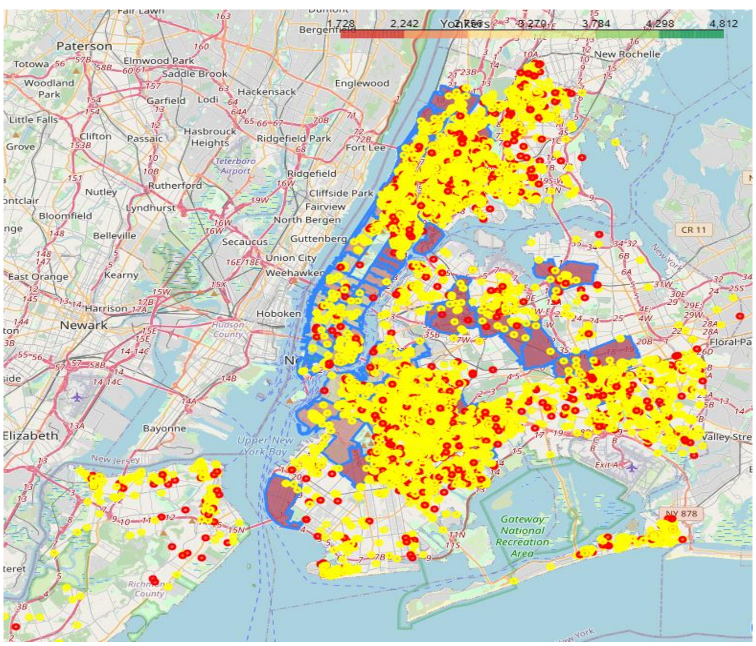

Figure 7. Crime and rent data overlaid on NYC map separated by ZIP code

Given the coordinates information, we can use the Foursquare API to access the sixth data source, explore the neighborhoods, and retrieve the top 30 venues within 0.25 miles (short walking distance) for all neighborhoods within NYC.

	Neighborhood	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	Allerton	Pizza Place	Deli / Bodega	Discount Store	Supermarket	Intersection	Martial Arts School	Dessert Shop	Donut Shop	Pharmacy	Chinese Restaurant
1	Annadale	Pizza Place	Dance Studio	Deli / Bodega	Food	Sushi Restaurant	Restaurant	Train Station	Filipino Restaurant	Event Space	Fabric Shop
2	Arden Heights	Pizza Place	Deli / Bodega	Pharmacy	Coffee Shop	Playground	Bus Stop	Field	Event Space	Fabric Shop	Factory
3	Arlington	Bus Stop	Deli / Bodega	American Restaurant	Coffee Shop	Grocery Store	Pizza Place	Women's Store	Filipino Restaurant	Fabric Shop	Factory
4	Arrochar	Italian Restaurant	Deli / Bodega	Bus Stop	Pizza Place	Athletics & Sports	Food Truck	Bagel Shop	Cosmetics Shop	Liquor Store	Mediterranean Restaurant

Figure 8. Dataframe of neighborhoods with their most common venues

From figure 8, it is clear pizza places are the most common in the first few neighborhoods. However, we want to know which neighborhoods within Manhattan, specifically, contain the most unique venues to provide the most options to renters within a short distance. From the reduced list of only Manhattan, it appears Murray Hill contains the most unique venues followed closely by Chelsea as shown in figure 9.

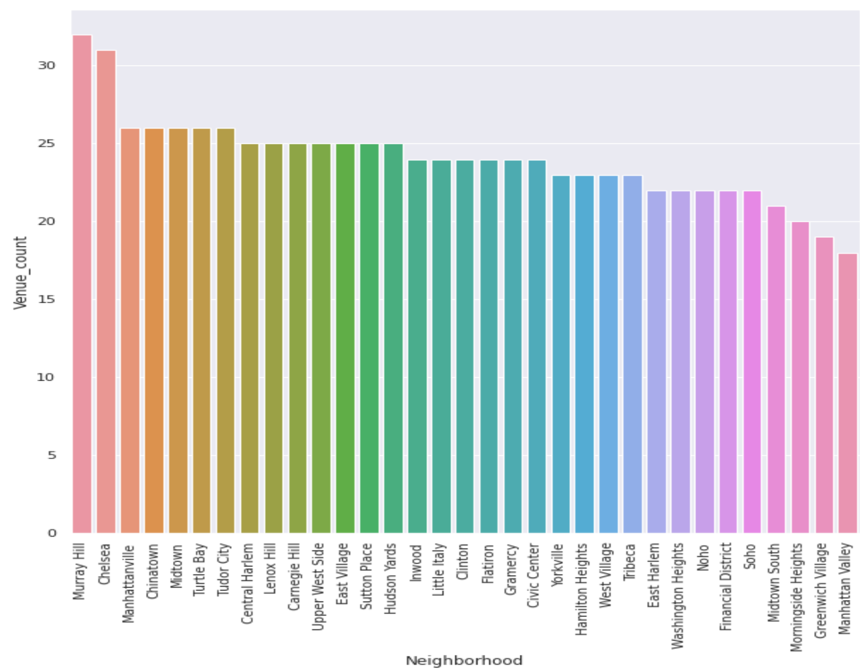

Figure 9. Count of unique venues by neighborhood

With this information we can create a new list of the most common venues within each neighborhood. We are now ready to begin clustering each of the neighborhoods to find how they compare with each other.

3. Clustering the Neighborhoods

Now that we have a focus on Manhattan, we can begin clustering each of the neighborhoods with the **k-Means** algorithm to find venues that reside there. This will provide us with the final piece of information, the most unique places within the shortest distance. From the rental data frame we can see there are 33 neighborhoods from the original 62 that are located within the Manhattan borough.

We will run the k-Means algorithm to build a clustering model with a different number of clusters (k). The features will be the mean of the frequency of occurrence of each venue category. Using Silhouette Score, we can measure and plot the clustering performances.

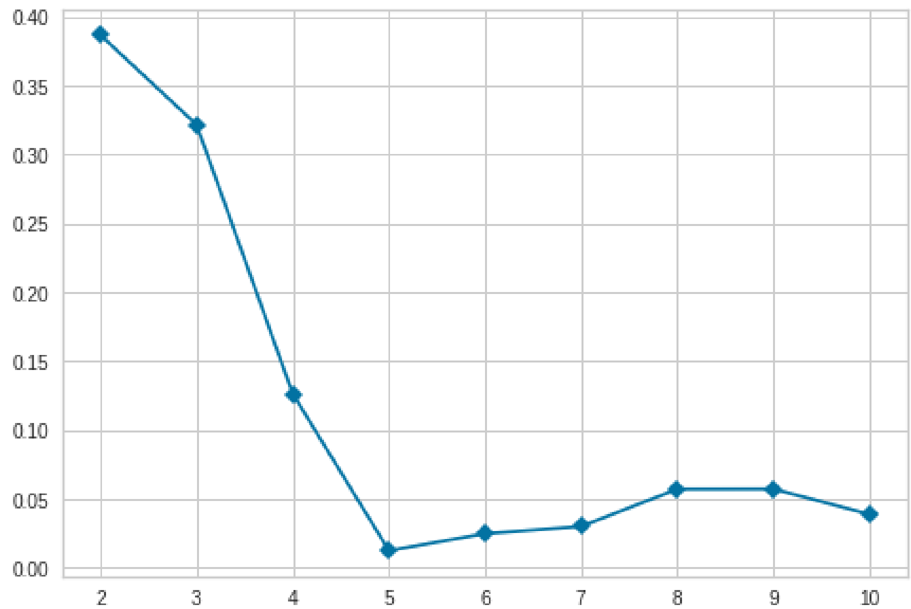

Figure 10. Silhouette Score for k-Means clustering

	Borough	Neighborhood	Latitude	Longitude	ClusterLabels	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	Venue_count
16	Manhattan	Murray Hill	40.748303	-73.978332	1	Korean Restaurant	Japanese Restaurant	Coffee Shop	Hotel	Bar	Burger Joint	Bank	Supermarket	Bagel Shop	Salon / Barbershop	32
17	Manhattan	Chelsea	40.744035	-74.003116	1	Hotel	Coffee Shop	Ice Cream Shop	Theater	Seafood Restaurant	Women's Store	Bus Stop	Tapas Restaurant	Beer Bar	Taco Place	31
8	Manhattan	Upper East Side	40.775639	-73.960508	1	Hotel	Pizza Place	Italian Restaurant	Falafel Restaurant	Bar	Sushi Restaurant	Bookstore	Miscellaneous Shop	Burrito Place	Shoe Store	27
1	Manhattan	Chinatown	40.715618	-73.994279	1	Chinese Restaurant	Spa	Sandwich Place	Yoga Studio	Dessert Shop	Spanish Restaurant	Bubble Tea Shop	Salon / Barbershop	Roof Deck	Cocktail Bar	26
36	Manhattan	Tudor City	40.746917	-73.971219	1	Park	Gym	Salad Place	Taco Place	Garden	Sushi Restaurant	Spanish Restaurant	Bridge	Seafood Restaurant	Café	26
35	Manhattan	Turtle Bay	40.752042	-73.967708	1	Karaoke Bar	Wine Bar	Italian Restaurant	Coffee Shop	Duty-free Shop	Grocery Store	Greek Restaurant	Residential Building (Apartment / Condo)	Thai Restaurant	Cocktail Bar	26
5	Manhattan	Manhattanville	40.816934	-73.957385	1	Coffee Shop	Seafood Restaurant	Bar	Park	Bike Trail	Climbing Gym	Chinese Restaurant	Latin American Restaurant	Supermarket	Boutique	26
15	Manhattan	Midtown	40.754691	-73.981669	1	Hotel	Bookstore	Park	Clothing Store	Ramen Restaurant	Szechuan Restaurant	Steakhouse	Sporting Goods Shop	Spa	Smoke Shop	26
30	Manhattan	Carnegie Hill	40.782683	-73.953256	1	Pizza Place	Café	Gym	Bookstore	Gym / Fitness Center	Coffee Shop	Restaurant	Nail Salon	Ramen Restaurant	Karaoke Bar	25
12	Manhattan	Upper West Side	40.787658	-73.977059	1	Bar	Bakery	American Restaurant	Italian Restaurant	Bagel Shop	Pub	Theater	Juice Bar	Greek Restaurant	Museum	25

Figure 11. Top 10 neighborhoods in Manhattan by unique venue count

From figure 11, the top ten neighborhoods are sorted by the number of unique venues within each neighborhood in addition to the most common types of venues. This is the final list to be used to compare rental prices to determine the best areas. However, in order to match these neighborhoods to their corresponding rent prices, we will utilize the geopy package to provide postal codes for the latitude and longitude coordinates.

Based on figure 10, it appears only 2 clusters provides the best results. Therefore, we will only have 2 cluster neighborhoods for all of Manhattan.

D. Results

Finally, we can visualize our clusters.

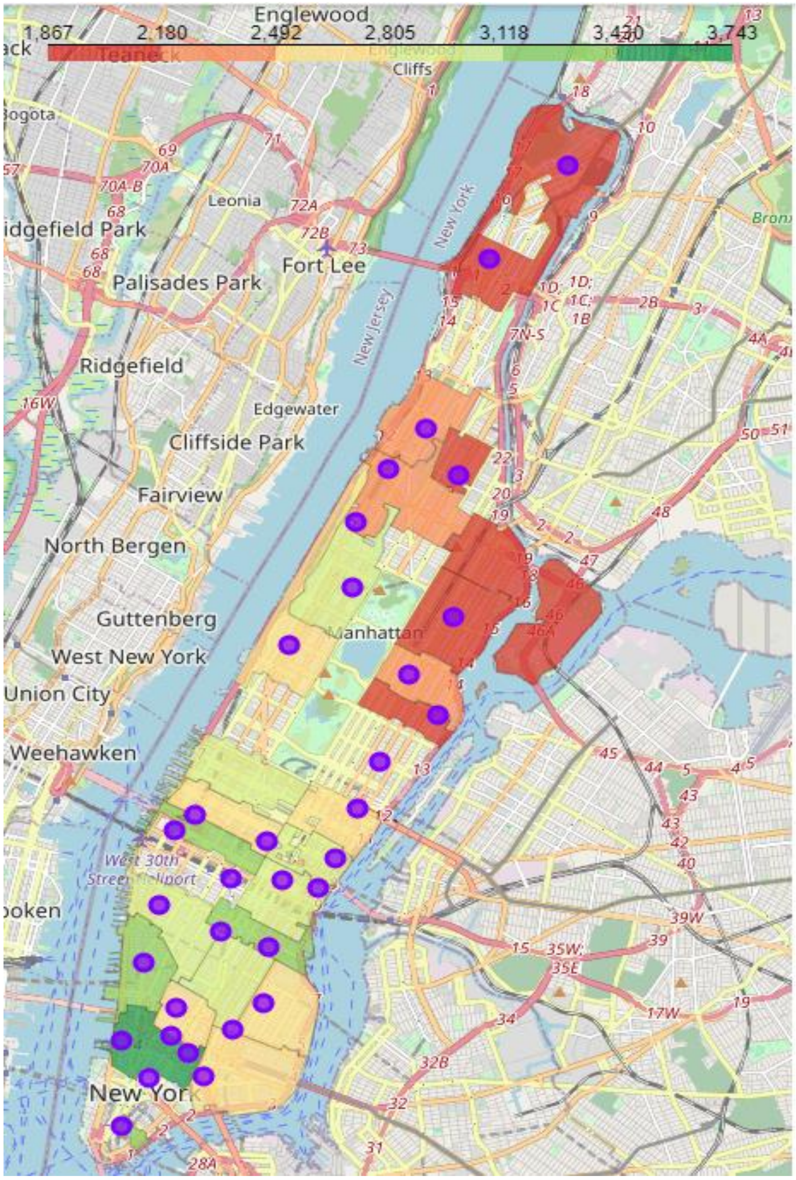

Figure 12. Neighborhood Clusters

As a result, we can examine the venues listed inside each cluster in addition to the rent price overlaid and define the discriminating venue categories that distinguish them.

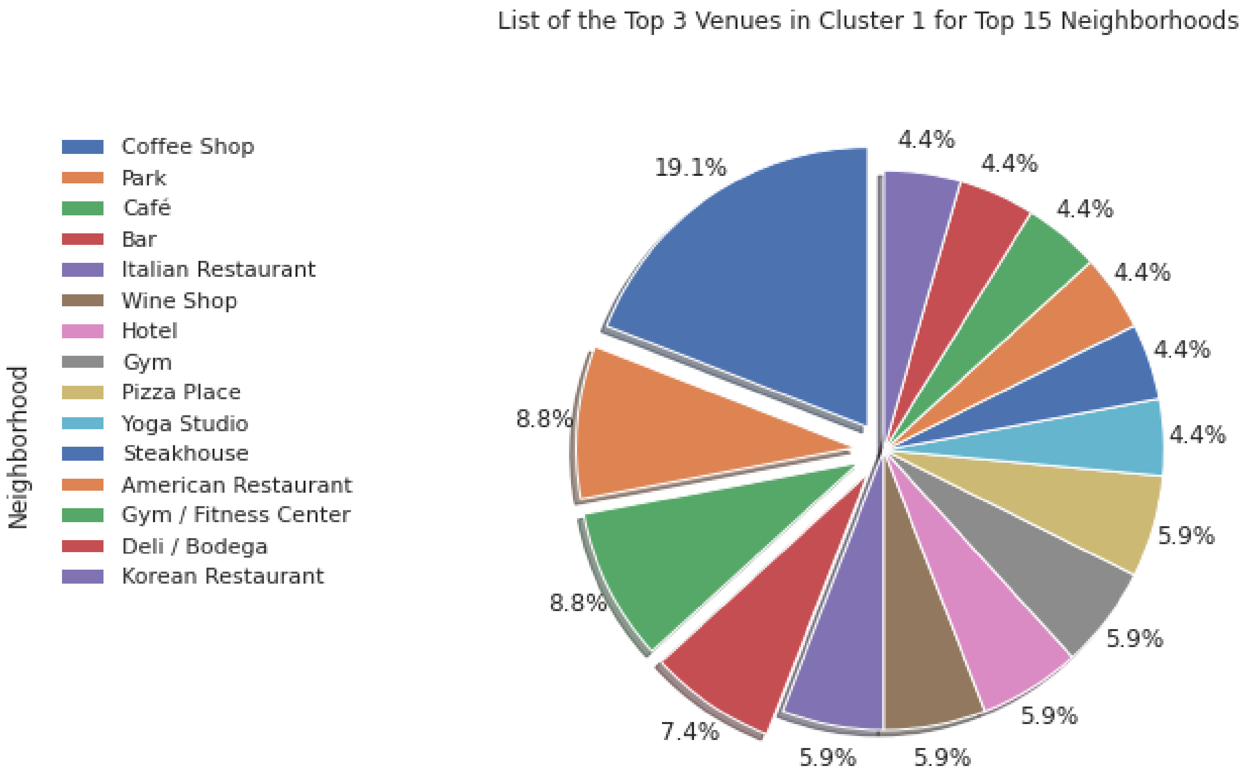

Figure 13. List of top 3 venues from the top 15 neighborhoods

However, to understand the best return for rent paid, we can take the median rent and divide it by the number of unique venues within the neighborhood to find the best relationship between venues and rent.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Venue_count	Rent	rent_per_venue
2	Inwood	Wine Bar	Deli / Bodega	Park	24	1867.0	77.791667
6	Central Harlem	African Restaurant	Bar	Chinese Restaurant	25	2087.0	83.480000
8	Yorkville	Coffee Shop	Deli / Bodega	Café	24	2044.0	85.166667
4	Manhattanville	Coffee Shop	Seafood Restaurant	Bar	26	2254.0	86.692308
9	Lenox Hill	Gym	Taco Place	Thai Restaurant	25	2171.0	86.840000
27	Carnegie Hill	Pizza Place	Café	Gym	25	2214.0	88.560000
13	Murray Hill	Korean Restaurant	Japanese Restaurant	Coffee Shop	32	2881.0	90.031250
3	Hamilton Heights	Coffee Shop	Yoga Studio	Café	23	2183.0	94.913043
7	East Harlem	Thai Restaurant	Mexican Restaurant	Latin American Restaurant	22	2101.0	95.500000
15	Chelsea	Hotel	Coffee Shop	Ice Cream Shop	31	2966.0	95.677419

Figure 14. Dataframe of the top 10 neighborhoods based on “rent per venue”

E. Discussion

The project’s main goal is to determine the best neighborhoods with the best return for rent which also correlates to opportunities for purchasing properties to serve as rentals. Tagging certain neighborhoods as “the best” can vary dependent upon opinion, but we can analytically determine the most value is determines by considering the following criteria:

1. **Safety**

- Based on the data gathered from the New York Police Department, there are locations deemed safer than others. From our analysis of shooting data specifically, we determined Manhattan to be the safest out of the five boroughs based on both the crime and murder density derived from population density.

2. **Rental Prices**

- The price of rent is, unfortunately, the **highest within Manhattan**, but that metric is skewed slightly based on the highest rent values in a small number of neighborhoods.
- Further analysis reveals there are rental deals to be had within Manhattan, but you must know where to look.

3. **Neighborhood Venues**

- A high number of unique venues appears to play a slight role in the amount of rent a neighborhood demands.
- All the neighborhoods of Manhattan were classified in the same cluster in terms of venue types. So it becomes more important to discern the difference in neighborhoods based on the median **rental cost per venue**.
- Therefore, the top three **recommended neighborhoods include Inwood, Central Harlem, and Yorkville** to their best value per venue.

F. Conclusion

Finding the best location to rent within NYC can be challenging, especially when the city is synonymous with having one of the highest costs of living within the United States. However, we can quickly gain meaningful insights into the city and its neighborhoods with openly available data.

Using the real estate investment firm and New York City as an example, I hope this project gives you a foundational understanding of how to deal with similar cases in the future. There is a laundry list of improvements that could be made to this project as well as additional sources of data to be included. What would you have done differently?

Thank you,

Troy Brommenschenkel

LinkedIn

GitHub Profile