Predicting the best area for a new construction building in New York

Konstantinos Bousios

February 28, 2019

1. Introduction

1.1 Background

New York is a city with the largest, populous and densely populated city in the United States and one of the world's most populated urban settlements. A strong city that has a significant impact on marketing, economy, information, art, fashion, research, technology, education and entertainment. It is an important center for international diplomacy and has been described as the cultural and economic capital of the world. For this reason, New York needs new buildings to accommodate all these needs. So, it is very important to finding areas which can better serve the above assumptions and new buildings to provide as many benefits as possible. Therefore, it is advantageous for the constructions companies to accurately predict which areas are ideal for the erection of buildings and they will have a big value. For example, this information can be used to find if there is public transport close to the new building, this provision adds value to the building.

1.2 Problem

Data that can help determine the finest area for new construction may include buildings values of the neighborhood, public transport, nearby restaurants, supermarkets, companies, and et cetera. This project aims to predict the neighborhood to build a new building which gives a big value in the real estate field based on these data.

1.3 Interest

Obviously, constructions companies would be very interested in accurate prediction of the right area for new construction, for competitive advantage and business values. Others who are interested in the right area for new construction such as people who search for a new building which provides many benefits may also be interested.

2. Data acquisition and cleaning

2.1 Data sources

The value for buildings in New York can be found from Kaggle dataset here. This dataset, however, lacks data for more information about benefits in the neighbourhoods. For example, venues in neighbourhoods. Exploring more about the neighbourhoods of New York I used the dataset from here. Also, after processing from the above datasets, I will use Foursquare API for the Borough of New York which I find out from the analysis of the other two datasets, to explore the benefits of each neighbourhood. For example, our findings into the exploration of neighbourhoods, give us information about the category of each venue and the exact location in the neighbourhood of the appropriate borough.

2.2 Data cleaning

Data downloaded from multiple sources. The processing of data from kaggle done in IBM Watson Studio. From first dataset, we kept columns for the borough, value of buildings and information about the locations. After this processing, I cleared all the nan values and created the dataset with all the data I need for my research. I had chosen the Watson studio for the processing of data from the first dataset for the reason that the dataset was too big and I had a problem with the memory to run the whole procession in local. By this choice, I achieved to create one dataset in csv form and continue the project. There were several problems with the first dataset like I had a lot of missing values and I tried to find as many possible information in the missing values for the reason, I wanted to keep the dataset with a lot of information and not minimize it, with the conviction that it will give me better conclusions. After fixing these problems, my data are ready for the analysis. For the second dataset, I created columns with values of Borough, Neighbourhood, Latitude, and Longitude from New York. After data cleaning in the first dataset, there were 27,312 samples and 8 features in the data. In the second dataset, there were 306 samples and 4 features in the data. By completing the first analysis of the two datasets, it will be followed by API Foursquare by the import of new data, to make a deeper exploration in the problem and give us the best solution.

3. Exploratory Data Analysis

3.1 Relationship between BOROCODE and ORIGINAL MARKET VALUE

First of all, from the first dataset, I made some metrics through graphic visualization to find the borough with the highest value of buildings. The borocode symbolizes the borough of New York. The explanation from the borocode is (1: Manhattan, 2: Bronx, 3: Brooklyn, 4: Queens, and 5: Staten Island). From the histogram, I had verified the assumption that Manhattan is the richest borough. Furthermore, Manhattan has a big difference in the value of the other boroughs. In figure 1 we observe optically this big difference of Manhattan.

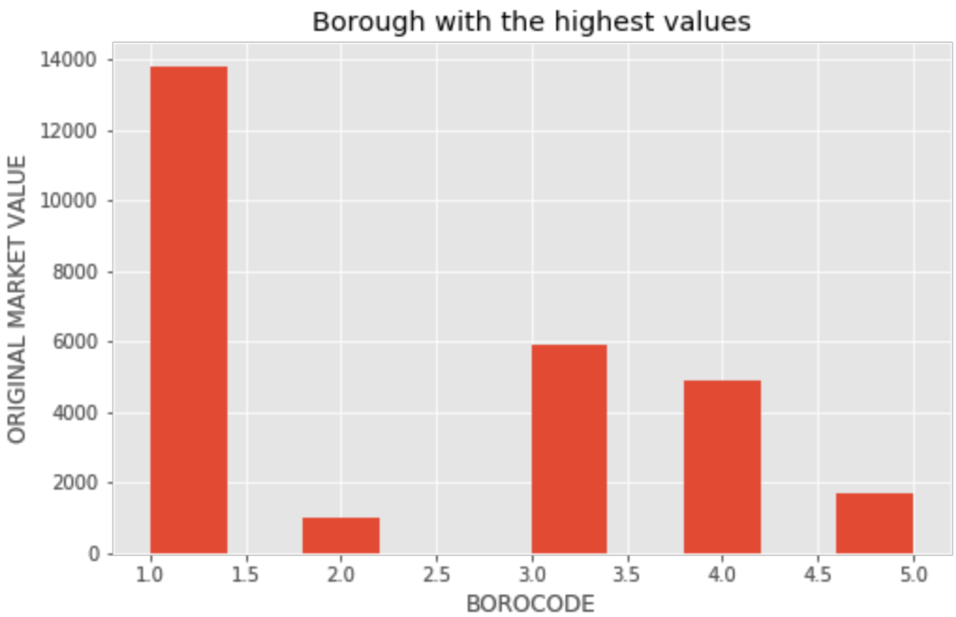


Figure 1. Borough with the highest value.

3.2 Relationship between Postcode and ORIGINAL MARKET VALUE

Secondly, from the first observation which I found that Manhattan is the richest borough, I started to explore more the areas in Manhattan. By the help of the two variables, postcode, and the original market value, I calculated the means of the original market value base on the postcodes. Then I found the top 15 postcodes with the highest mean market value and I created a bar plot to depiction them.

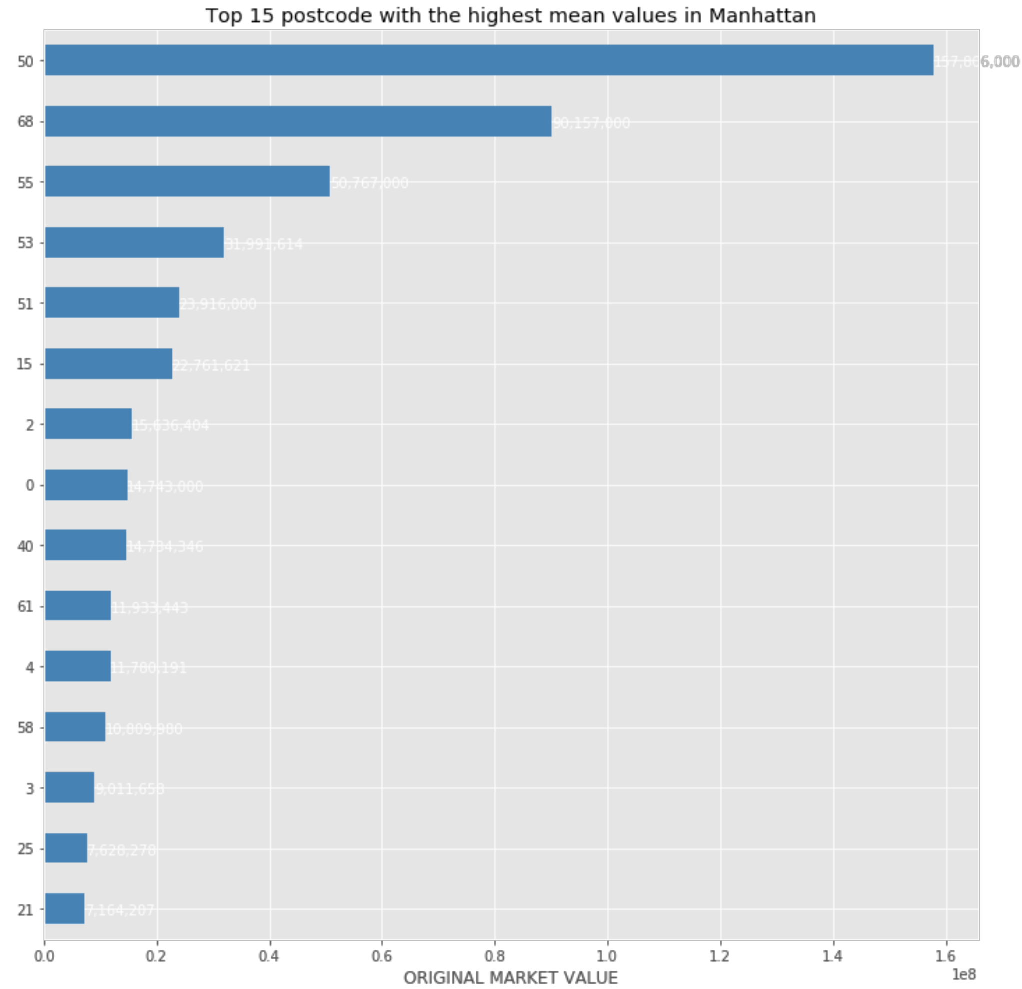
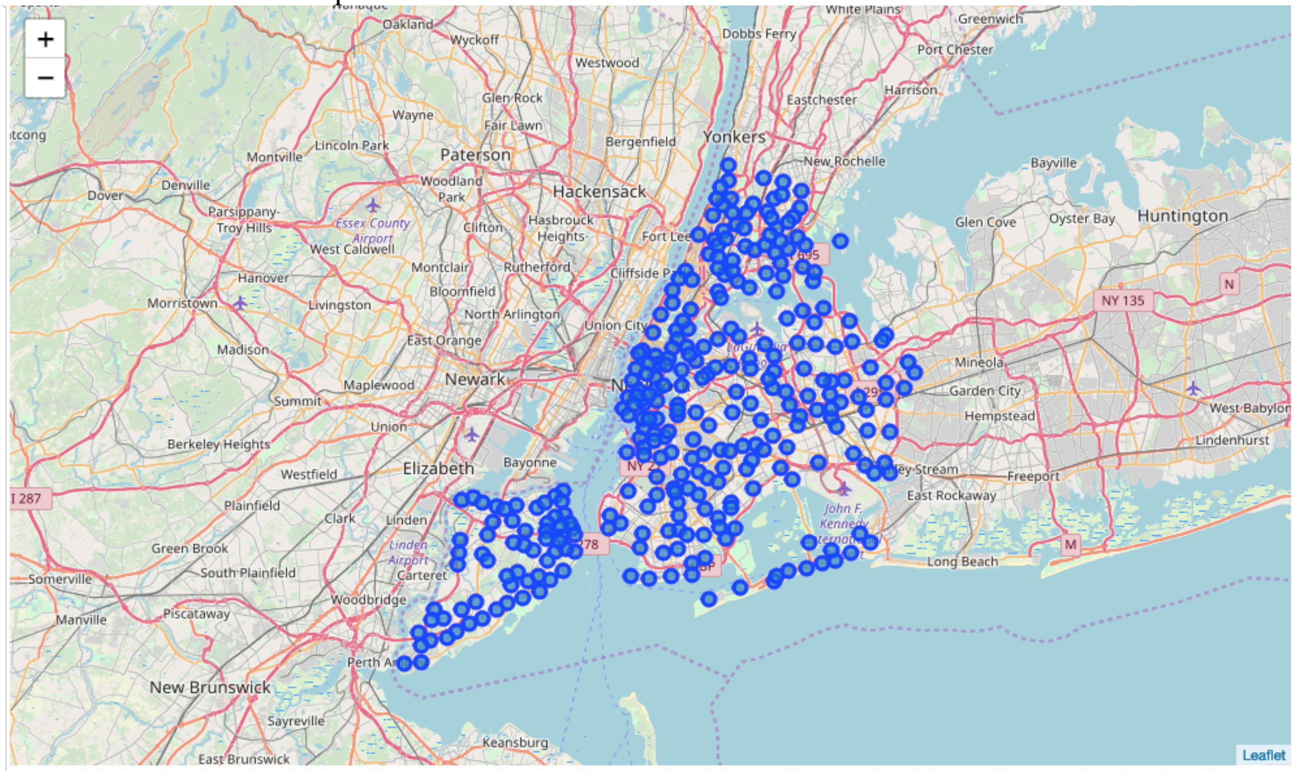


Figure 2. Top 15 postcode with the highest mean values in Manhattan.

The postcode with the number 50 in figure 2 is 10120.0 and mean value of 158 million dollars. The second postcode with the highest value is 10281.0 and mean value of 90 million dollars. The third postcode with the highest value is 10281.0 and mean value of 50 million dollars. The fourth postcode with the highest value is 10151.0 and mean value of 32 million dollars. The fifth postcode with the highest value is 10123.0 and mean value of 24 million dollars. The sixth postcode with the highest value is 10017.0 and mean value of 23 million dollars. The seventh postcode with the highest value is 10001.0 and mean value of 16 million dollars. The eighth and ninth postcodes with the highest value are 10000.0 and 10044.0 and mean value of 15 million dollars. The tenth and eleventh postcodes with the highest value are 10170.0 and 10004.0 and mean value of 12 million dollars. The twelfth postcode with the highest value is 10165.0 and mean value of 10 million dollars. The thirteenth postcode with the highest value is 10003.0 and mean value of 9 million dollars. The fourteenth and fifteen postcodes with the highest value are 10027.0 and 10023.0 and mean value of 7 million dollars.

3.3 Mapping visualization of New York

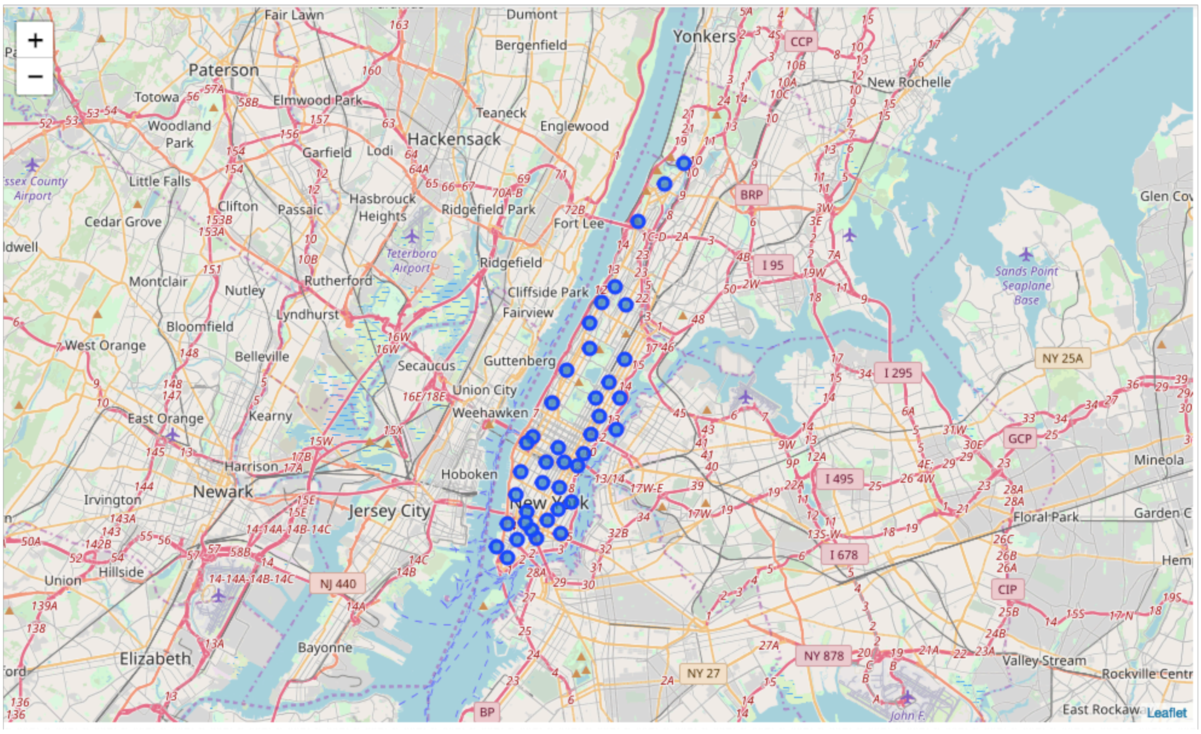
The research continued with the second dataset in which I collected exactly data for the location based on the latitude and longitude of New York for the creation of the map of this state of America in the map 1.



Map 1. New York map.

3.4 Mapping visualization of Manhattan

With the completion of the data analysis of the first dataset and the result that Manhattan is the highest value borough of New York state I focused on mapping the Manhattan (Map 2). By this section, I completed the analysis for the second dataset. After that, I continued with the API Foursquare for extra exploration for the neighborhoods.



Map 2. Manhattan map**.**

3.5 Foursquare API

In the last section for the data, I used API Foursquare to explore the neighborhoods and segment them. Giving as parameters the neighborhood's latitude and longitude values. The API Foursquare gave to us results about the venues in Manhattan, then I categorized those results and created a new data frame for deeper exploration of the neighborhoods of Manhattan. Furthermore, I took the mean of the frequency of occurrence of each category for better conclusions. After this calculation, I proceeded to export of the top ten venues for each neighborhood. with the completion of this whole process, and with the collection and study of all the past data I am in a position to apply my model to lead to my final conclusions.

4. Predictive Modeling

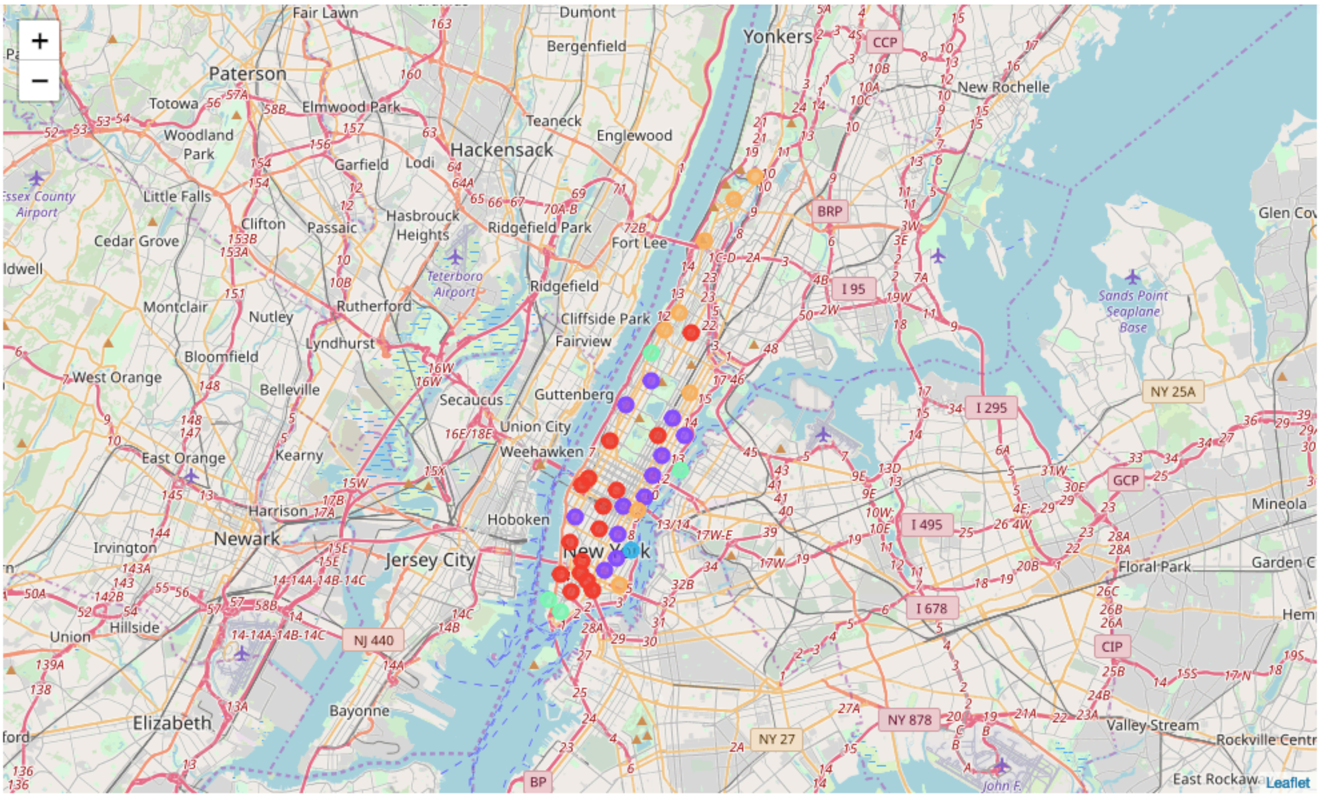
There is a type of model, clustering, that can be used to predict the best areas for the construction of a new building. Clustering means it is the process of grouping the database items into clusters. All the members of the cluster have similar features. Members belong to different clusters has dissimilar features. Cluster analysis divides data into meaningful or useful groups (clusters). Cluster analysis is very useful in spatial databases. For example, by grouping feature venues as clusters can be used to create thematic maps which are useful in geographic information systems. Therefore, in this study, I carried out a K-means clustering modeling.

4.1 K-means Clustering model

The K-Means node clusters the data set into distinct clusters. The method defines a fixed number of clusters, iteratively assigns records to clusters, and adjusts the cluster centers until further refinement can no longer improve the model. Instead of trying to predict an outcome, k-means uses a process known as unsupervised learning to uncover patterns in the set of input fields. Records are grouped so that records within a group or cluster tend to be similar to each other, but records in different groups are dissimilar. K-Means works by defining a set of starting cluster centers derived from data. It then assigns each record to the cluster to which it is most similar, based on the record's input field values. After all, cases have been assigned, the cluster centers are updated to reflect the new set of records assigned to each cluster. The records are then checked again to see whether they should be reassigned to a different cluster, and the record cluster iteration process continues until either the maximum number of iterations is reached, or the change between one iteration and the next fails to exceed a specified threshold. So, by applying the K-means algorithm can be executed in the following steps:

- Partition of objects into k non-empty subsets.
- Identifying the cluster centroids (mean point) of the current partition.
- Assigning each point to a specific cluster.
- Compute the distances from each point and allot points to the cluster where the distance from the centroid is minimum.
- After re-allotting the points, find the centroid of the new cluster formed.

I had set five K-means Clustering which grouped the top 10 venues for each neighborhood into clusters and defines a cluster center for each cluster. These Clusters centers are the centroids of each cluster and are at a minimum distance from all the points of a particular cluster. Henceforth, the top ten venues will be at minimum distance from all the neighborhoods within a cluster. I created a map (map 3) of the whole process to have a visualize, how it looks like until now.



Map 3. Clusters map.

4.2 Results for each cluster

By examining each cluster separate we have the below results.

1) From cluster one in figure 3 we observe fifteen different neighborhoods with the top 10 common venues which are symbolized in the map 3 with the red color.

	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
1	Chinatown	Chinese Restaurant	Bubble Tea Shop	American Restaurant	Dim Sum Restaurant	Vietnamese Restaurant	Cocktail Bar	Hotpot Restaurant	Bakery	Noodle House	Salon / Barbershop
6	Central Harlem	African Restaurant	Seafood Restaurant	Fried Chicken Joint	American Restaurant	French Restaurant	Gym / Fitness Center	Chinese Restaurant	Cosmetics Shop	Bookstore	Library
8	Upper East Side	Italian Restaurant	Exhibit	Art Gallery	Bakery	Coffee Shop	Juice Bar	Boutique	French Restaurant	Gym / Fitness Center	Hotel
13	Lincoln Square	Gym / Fitness Center	Theater	Plaza	Italian Restaurant	Concert Hall	Café	French Restaurant	Indie Movie Theater	Opera House	Performing Arts Venue
14	Clinton	Theater	Italian Restaurant	American Restaurant	Gym / Fitness Center	Coffee Shop	Hotel	Wine Shop	Spa	Gym	Indie Theater
15	Midtown	Clothing Store	Hotel	Theater	Coffee Shop	Steakhouse	Bakery	Spa	Cocktail Bar	Bookstore	American Restaurant
18	Greenwich Village	Italian Restaurant	French Restaurant	Sushi Restaurant	Clothing Store	Chinese Restaurant	Indian Restaurant	Seafood Restaurant	Café	Boutique	Caribbean Restaurant
21	Tribeca	Italian Restaurant	American Restaurant	Park	Café	Spa	Gym	Boutique	Coffee Shop	Wine Bar	Greek Restaurant
22	Little Italy	Bakery	Café	Yoga Studio	Ice Cream Shop	Sandwich Place	Salon / Barbershop	Seafood Restaurant	Chinese Restaurant	Bubble Tea Shop	Clothing Store
23	Soho	Clothing Store	Boutique	Women's Store	Shoe Store	Men's Store	Italian Restaurant	Art Gallery	Coffee Shop	Mediterranean Restaurant	Seafood Restaurant
24	West Village	Italian Restaurant	Cosmetics Shop	New American Restaurant	Wine Bar	Jazz Club	Gastropub	Park	French Restaurant	Bakery	American Restaurant
32	Civic Center	Gym / Fitness Center	Italian Restaurant	Bakery	Cocktail Bar	French Restaurant	Sporting Goods Shop	Coffee Shop	Gym	Park	Sandwich Place
33	Midtown South	Korean Restaurant	Coffee Shop	Cosmetics Shop	Japanese Restaurant	Hotel Bar	Hotel	Italian Restaurant	Gym / Fitness Center	Bakery	Boutique
38	Flatiron	Italian Restaurant	Gym / Fitness Center	American Restaurant	Gym	Yoga Studio	Clothing Store	Cycle Studio	Japanese Restaurant	Dessert Shop	Cosmetics Shop
39	Hudson Yards	Italian Restaurant	Coffee Shop	Theater	Gym / Fitness Center	Restaurant	Café	Hotel	American Restaurant	Art Gallery	Gym

Figure 3. Cluster 1.

2) From cluster two in figure 4 we observe twelve different neighborhoods with the top 10 common venues which are symbolized in the map 3 with the purple color.

	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
9	Yorkville	Italian Restaurant	Coffee Shop	Bar	Gym	Pizza Place	Deli / Bodega	Japanese Restaurant	Mexican Restaurant	Sushi Restaurant	Park
10	Lenox Hill	Italian Restaurant	Sushi Restaurant	Coffee Shop	Gym / Fitness Center	Pizza Place	Burger Joint	Gym	Sporting Goods Shop	Mexican Restaurant	Café
12	Upper West Side	Italian Restaurant	Bar	Indian Restaurant	Vegetarian / Vegan Restaurant	Burger Joint	Coffee Shop	Bakery	Wine Bar	Mediterranean Restaurant	Sushi Restaurant
16	Murray Hill	Coffee Shop	Hotel	Italian Restaurant	Salon / Barbershop	Sandwich Place	Spa	Bar	Japanese Restaurant	Gym	French Restaurant
17	Chelsea	Coffee Shop	Italian Restaurant	Ice Cream Shop	American Restaurant	Nightclub	Bakery	Seafood Restaurant	Theater	Hotel	Tapas Restaurant
19	East Village	Bar	Ice Cream Shop	Wine Bar	Cocktail Bar	Mexican Restaurant	Ramen Restaurant	Chinese Restaurant	Coffee Shop	Pizza Place	Speakeasy
25	Manhattan Valley	Coffee Shop	Pizza Place	Mexican Restaurant	Spa	Thai Restaurant	French Restaurant	Indian Restaurant	Deli / Bodega	Yoga Studio	Café
27	Gramercy	Italian Restaurant	Cocktail Bar	American Restaurant	Wine Shop	Bagel Shop	Coffee Shop	Restaurant	Pizza Place	Mexican Restaurant	Thrift / Vintage Store
30	Carnegie Hill	Pizza Place	Coffee Shop	Cosmetics Shop	Café	Yoga Studio	Wine Shop	Bar	Bookstore	French Restaurant	Spa
31	Noho	Italian Restaurant	French Restaurant	Cocktail Bar	Bookstore	Grocery Store	Gift Shop	Mexican Restaurant	Sushi Restaurant	Hotel	Coffee Shop
34	Sutton Place	Gym / Fitness Center	Italian Restaurant	Furniture / Home Store	Juice Bar	Indian Restaurant	Dessert Shop	Bakery	American Restaurant	Boutique	Coffee Shop
35	Turtle Bay	Italian Restaurant	Sushi Restaurant	Hotel	Wine Bar	Coffee Shop	Japanese Restaurant	Café	Steakhouse	Park	French Restaurant

Figure 4. Cluster 2.

3) From cluster three in figure 5 we observe one neighborhoods with the top 10 common venues which are symbolized in the map 3 with the blue color.

	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
37	Stuyvesant Town	Bar	Playground	Park	Basketball Court	Farmers Market	Coffee Shop	German Restaurant	Cocktail Bar	Boat or Ferry	Pet Service

Figure 5. Cluster 3.

4) From cluster four in figure 6 we observe four different neighborhoods with the top 10 common venues which are symbolized in the map 3 with the green color.

	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
11	Roosevelt Island	Sandwich Place	Park	Liquor Store	Athletics & Sports	Bus Station	Supermarket	Farmers Market	Metro Station	Outdoors & Recreation	School
26	Morningside Heights	Coffee Shop	Park	Bookstore	Food Truck	American Restaurant	Burger Joint	Sandwich Place	Deli / Bodega	Tennis Court	College Cafeteria
28	Battery Park City	Coffee Shop	Park	Hotel	Italian Restaurant	Wine Shop	Cupcake Shop	Food Truck	BBQ Joint	Food Court	Department Store
29	Financial District	Coffee Shop	Hotel	Wine Shop	Steakhouse	Gym	Bar	Café	Pizza Place	Food Truck	Italian Restaurant

Figure 6. Cluster 4.

5) From cluster five in figure 7 we observe eight different neighborhoods with the top 10 common venues which are symbolized in the map 3 with the yellow color.

	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	Marble Hill	Discount Store	Coffee Shop	Diner	Seafood Restaurant	Tennis Stadium	Gym	Bank	Sandwich Place	Donut Shop	Supplement Shop
2	Washington Heights	Café	Bakery	Mobile Phone Shop	Shoe Store	Pizza Place	Deli / Bodega	Sandwich Place	Supermarket	Park	Tapas Restaurant
3	Inwood	Mexican Restaurant	Café	Lounge	Pizza Place	Bakery	Deli / Bodega	Wine Bar	Pharmacy	American Restaurant	Park
4	Hamilton Heights	Mexican Restaurant	Coffee Shop	Café	Deli / Bodega	Pizza Place	Liquor Store	Indian Restaurant	Sushi Restaurant	Park	Sandwich Place
5	Manhattanville	Deli / Bodega	Italian Restaurant	Mexican Restaurant	Seafood Restaurant	Fried Chicken Joint	Beer Garden	Bike Trail	Sushi Restaurant	Supermarket	Burger Joint
7	East Harlem	Mexican Restaurant	Bakery	Deli / Bodega	Latin American Restaurant	Thai Restaurant	Seafood Restaurant	Taco Place	Street Art	Steakhouse	Pizza Place
20	Lower East Side	Coffee Shop	Chinese Restaurant	Café	Shoe Store	Bakery	Cocktail Bar	Latin American Restaurant	Pizza Place	Ramen Restaurant	Art Gallery
36	Tudor City	Mexican Restaurant	Park	Greek Restaurant	Café	Pizza Place	Dog Run	Diner	Hotel	Deli / Bodega	Spanish Restaurant

Figure 7. Cluster 5.

4.3 Results

From our findings, I conclude that the best borough of Manhattan for new construction is the neighborhoods from the cluster one (figure 3). Cluster 1 gives us information for those neighborhoods which have the most benefits and in combination with the high market value of Manhattan the new buildings will be extremely competitive in the real estate market.

5. Conclusion

In this study, I analyzed a problem to find the best neighborhood for constructing a new building in New York. I identified the market value per borough in the state of New York and then I analyzed the location of the neighborhoods. In the last step, I used API Foursquare for the exploration of the venues in neighborhoods of Manhattan, because Manhattan is the borough with the highest market values in the buildings. I built a K-means clustering model to predict the neighborhood for new building construction. This model can be very useful in helping construction companies to make the right investment for a new building or people who want to find the best neighborhood to moved.

6. Future directions

The model predicts well with those parameters. However, I think the model could use more improvements on capturing more information for the neighborhoods. For example, if there were data for police stations or fire stations the metrics maybe are different. Another example is a parameter for the health field which provides a benefit in an emergency situation, with this information maybe the recommended neighborhoods are different. More data, especially data of different types, would help improve model performances significantly. The Model in this study mainly focused on individual features. In addition, new trends for new kind of venues which adopt new investors maybe change the hotspots in neighborhoods. For example, If a new mall opens in a different neighborhood of our predictions, the new location will create a new movement in the market and the value of the new location will increase. These interactions data are obviously more difficult to extract and quantify, but if optimized, could bring significant improvements to the model.