Locating a good area for Businesses

Zonghao Zou

January 8, 2020

1. Introduction

1.1 Background

There are many types of businesses out there. Technology companies, services companies, or other types you can think of. Some of those businesses are easier to start than the others. Food and drink services usually can be considered as one of the easiest businesses to start. It is this project's goal to understand and study if you like to start a food or drink services business, where would be the best location to open the store. To simplify the problem, I will focus on a very specific food and drink service franchise, Starbucks. However, it is important to note that this can be easily expanded to other types of businesses or franchises

1.2 Reasons for selecting Starbucks

Starbucks is one of the most prevalent franchises in United States. It was found in 1971, Seattle, Washington. As of early 2019, the company operates over 30,000 locations worldwide. Out of all the cities have Starbucks, New York city is the one with most Starbucks located. I want to select a franchise that has a rather large base of dataset, so that I can perform some analysis on it.

1.3 Questions trying to answer

Since many people go to Starbucks, a nature question follows, if a customer wants to experience a good Starbucks service, which area of the city will they visit? To answer this question, it is similar with answering, if a person wants to open a Starbucks coffeeshop, where would be the best location?

This project will answer that question based on several factors, such as latitude, longitude, as well as ratings. All those factors will be further explained in the Data Cleaning session.

1.4 Interest

This could serve any person who is wishing to open up a Starbucks store in New York city, or any customers who want to experience a better service.

2. Data Acquisition and Cleaning

2.1 Data Acquisition

I used Foursquare as the source for my data. More specifically, I set version = '20180604', limit =500, search_query = 'Starbucks', radius =100000, and address = 'New York, NY'. I searched for large amount of data sets near within 100km of New York city. This search is done via venue search.

After acquiring those data, I also performed an ID search on each individual data points to gain more specific information of those stores.

2.2 Data Cleaning

After I acquired those data from Foursquare, I manipulated the desired json file. I selected response and venues out of my acquired json file, and further normalized the data with json_normalize function. I, then filtered the according database in acquiring the following information of my data.

	name	categories	address	cc	city	country	crossStreet	distance	formattedAddress	labeledLatLngs	lat	lng	neighborhood	postalCode	state
0	Starbucks	Coffee Shop	195 Broadway	US	New York	United States	at Dey St	412	[195 Broadway (at Dey St), New York, NY 10007,...	[{'label': 'display', 'lat': 40.710922, 'lng':...	40.710922	-74.010284	NaN	10007	NY
1	Starbucks	Coffee Shop	38 Park Row	US	New York	United States	at Beekman St	139	[38 Park Row (at Beekman St), New York, NY 100...	[{'label': 'display', 'lat': 40.71159756, 'lng...	40.711598	-74.006726	NaN	10038	NY
2	Starbucks Reserve	Coffee Shop	250 Vesey St	US	New York	United States	2nd Fl	810	[250 Vesey St (2nd Fl), New York, NY 10281, Un...	[{'label': 'display', 'lat': 40.71417, 'lng': ...	40.714170	-74.015434	Battery Park City	10281	NY
3	Starbucks	Coffee Shop	125 Chambers St	US	New York	United States	at W Broadway	402	[125 Chambers St (at W Broadway), New York, NY...	[{'label': 'display', 'lat': 40.715534, 'lng':...	40.715534	-74.009030	NaN	10007	NY
4	Starbucks	Coffee Shop	233 Broadway	US	New York	United States	at Barclay St	181	[233 Broadway (at Barclay St), New York, NY 10...	[{'label': 'display', 'lat': 40.71220388, 'lng...	40.712204	-74.008052	NaN	10279	NY

In the last column not shown above, I have a list of unique ID for each Starbucks. I looped through each Starbucks store to acquire their according information. I only consider their ratings, so I went to their response, venue, and rating section to acquire each ratings. I further attached the ratings onto the original dataframe.

Out of all the information available in the current dataframe, there are only a few of them I care about, name, address, distance, latitude, longitude, and rating. I selected those important features and put them into a new dataframe, shown below:

	name	address	distance	lat	lng	Ratings
0	Starbucks	38 Park Row	139	40.711598	-74.006726	6.8
1	Starbucks	233 Broadway	181	40.712204	-74.008052	7.4
2	Starbucks	1 Pace Plz	204	40.711142	-74.004796	6.8
3	Starbucks	291 Broadway	236	40.714855	-74.005936	6.8
4	Starbucks	120 Church St	282	40.713839	-74.009026	6.9
5	Starbucks	130 Fulton St	323	40.710280	-74.008080	6.4
6	Starbucks	125 Chambers St	402	40.715534	-74.009030	6.6
7	Starbucks	111 Worth St	410	40.715714	-74.003154	6.8
8	Starbucks	195 Broadway	412	40.710922	-74.010284	6.9
9	Starbucks	100 William St	494	40.708412	-74.007400	6.5
10	Starbucks	55 Liberty St	532	40.708740	-74.009510	6.5

I also removed outliers such as ‘Starbucks Reserve’. Those stores have a much higher rating than usual Starbucks stores. Therefore, adding them will influence the accuracy of my analysis.

2.3 Feature Selection

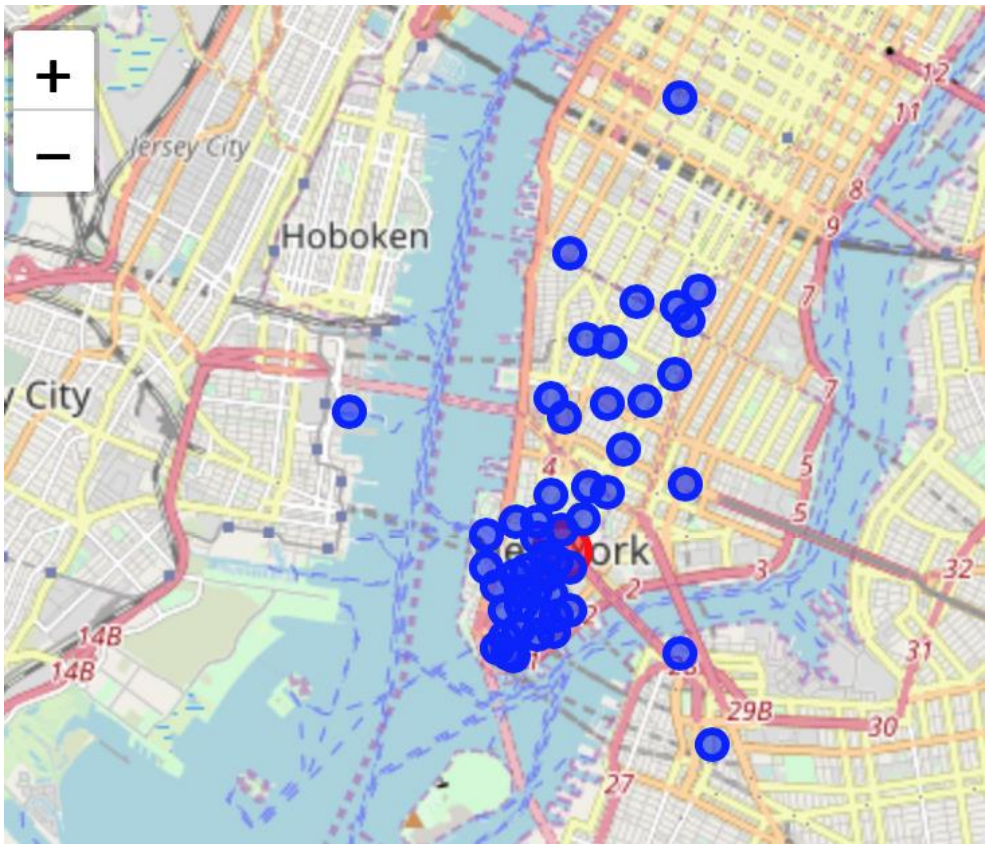
Even though I tried to acquire as much as stores as possible, Foursquare only provided me with 50 Starbucks stores. Furthermore, I had to cut two ‘Starbucks Reserve’. I have 48 stores to work with. There are 4 main features affecting my analysis, distance to New York, latitude, longitude, and rating.

3. Data Exploring

3.1 Plotting each store on the map.

By plotting each store, we can take a look at where each of the stores are located and how dense around certain areas. This gives us a great opportunity to understand our data better.

The location plot is shown below:

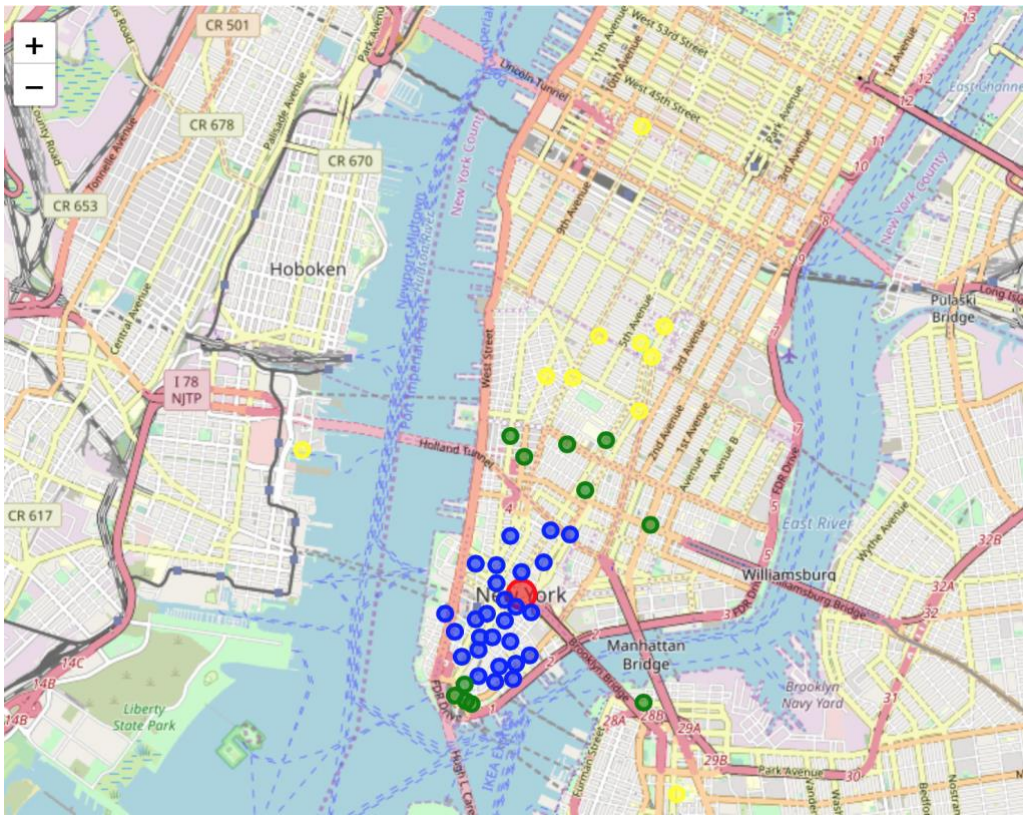


We can see that stores are much more condensed near the center of New York city.

3.2 Considering distance

We can further observe our data using the distance feature: how close are they to New York. I set up three sections, less than 1000 meters, between 1000 and 2000, greater than 2000. (colored by blue, green and yellow)

Plot is shown below:



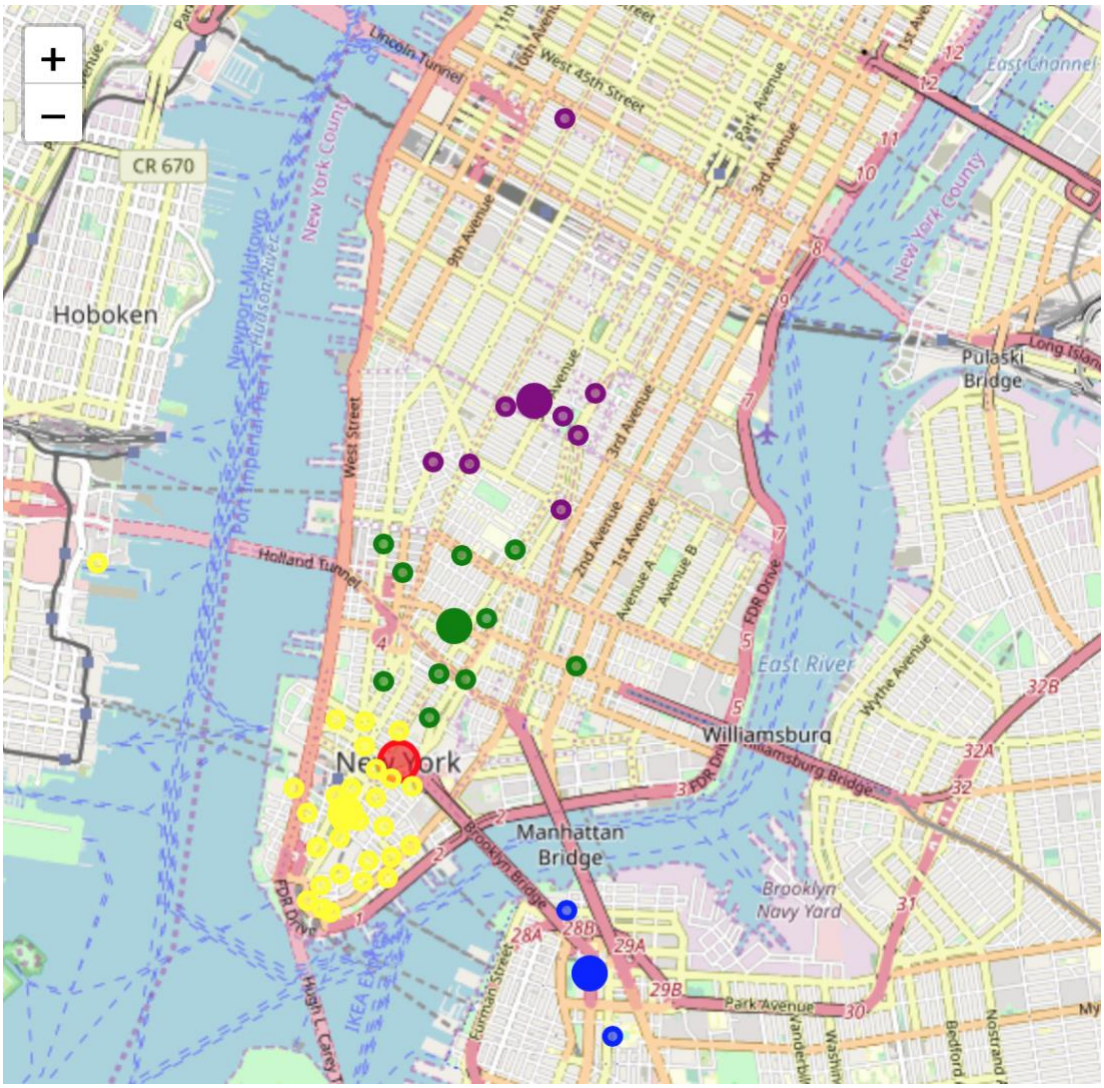
We can see that great amount of data points are within 1000 meters radius of the New York city.

4. **Clustering**

4.1 Clustering by latitude and longitude (general case)

The simplest idea to locate an area that is best for Starbucks is the perform a K means clustering on all the available datapoints. Here we chose k = 4 and run for 20 iterations (Details of my code can be found on the code section).

Clustering results can be shown below:

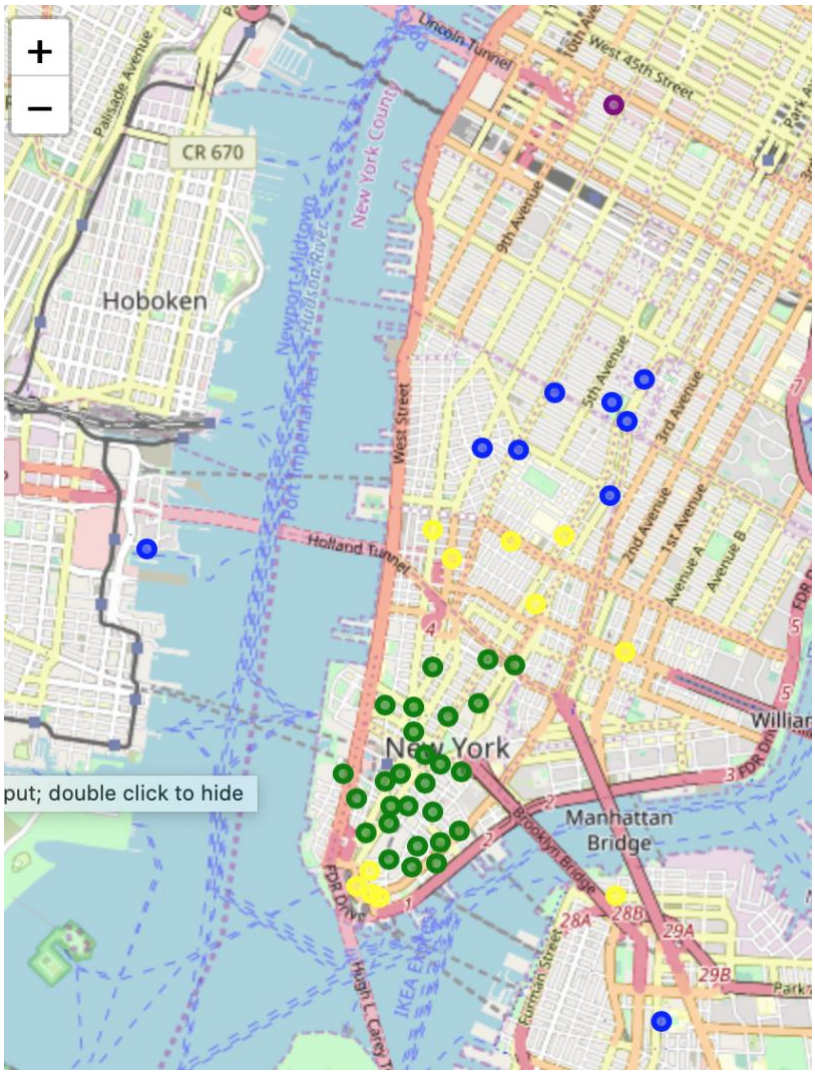


Each final centroid has been marked with larger dots. After I clustered each coffeeshop, I acquired an average rating from each cluster to see which cluster on average has the highest rating. The ratings are: 6.45,6.34,6.78, and 6.9 (corresponding to blue green, yellow, and purple).

Just from this simple model, we can already identify the area that has the highest rating, the purple section of the city.

4.2 Clustering by distance and rating

This is an attempt to find relationship between distance and ratings. I clustered the coffeeshop based on the distance away the New York and its rating. Similarly, I used K means clustering with k=4, iterations =20.

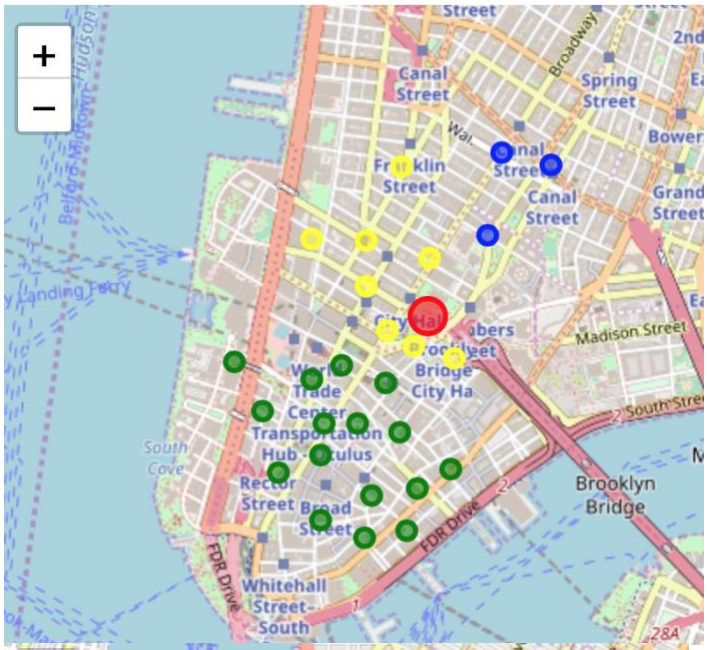


This clustering shows a rather interesting result. The rating score for each cluster is: 6.82, 6.75, 6.34, 8.1 (corresponding to blue, green, yellow, and purple). Since the last cluster only has one store, its validity as a good location is low. However, it is surprising to find roughly the same areas have good ratings, in this case blue, and green. In the previous clustering, purple and yellow.

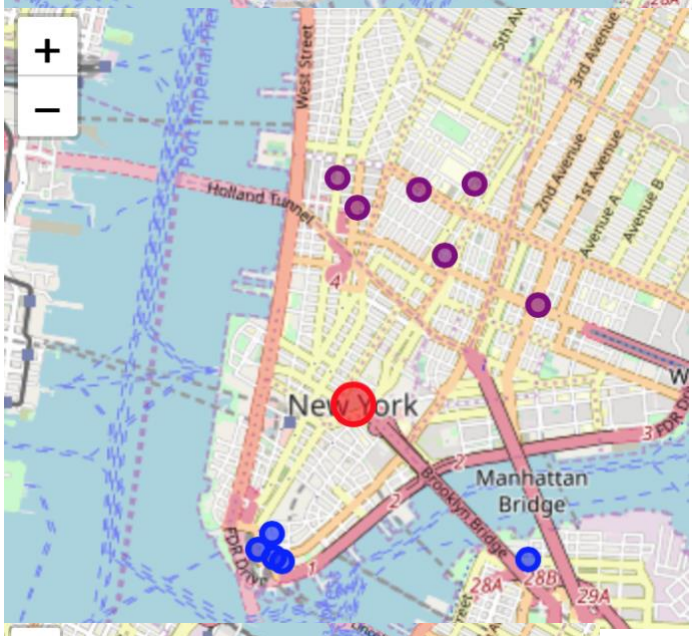
4.3 Clustering by latitude and longitude within clusters of distance

We can also dive in for a more specific look on where is the best area for opening a Starbucks coffeeshop. We further performed clustering analysis on 3.2.

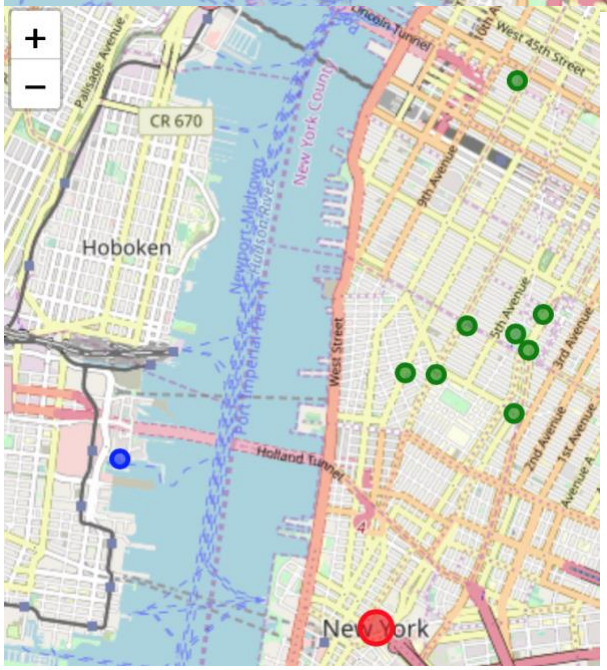
For each section, I performed individual clustering. The results are shown below.



The first map on the left is three clustering on all the stores within 1000 meters of New York city. Scores are: 6.5,6.7,6.9 (blue, green, and yellow). From this we can get a sense that slightly below New York is a better location for a Starbucks store

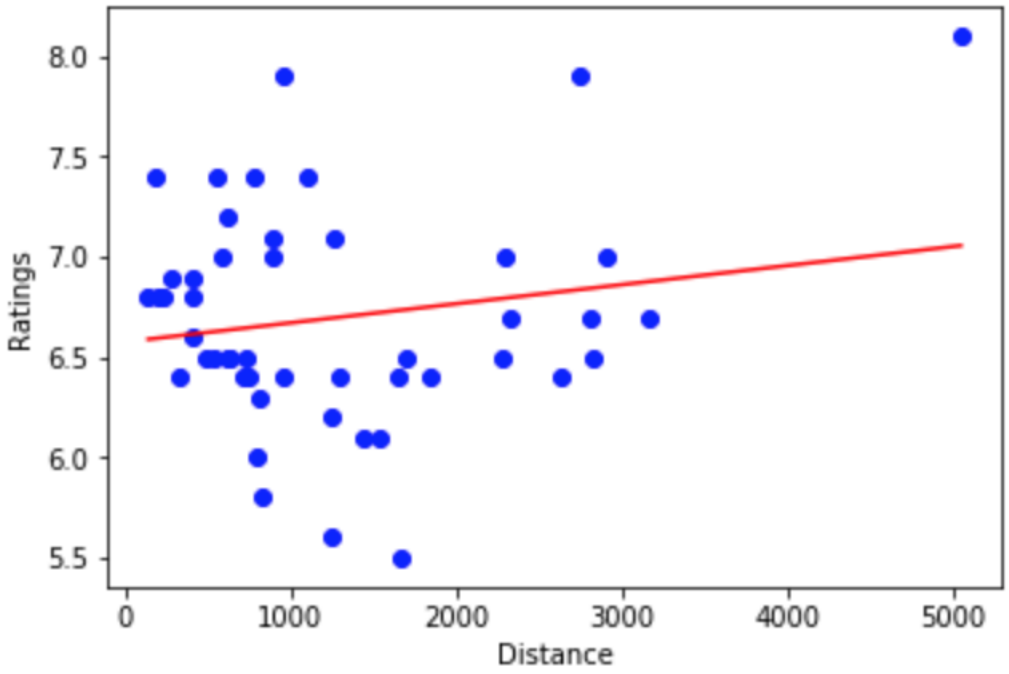


The second map is clustering of stores between 1000 and 2000. Scores are: 6.56, 6.12 (blue, purple)



The third map on the left has two clustering. Scores are: 7.9, 6.9 (blue, green)

4.4 Regression analysis on the relationship between distance and rating



We can see that there is some correlation between how far away you are from the city and how good your rating is.

5. **Conclusion**

To answer the questions we proposed, in New York city, if we wish to open a Starbucks store or if we wish to locate a good area for great services, there are primary two areas you could consider: one is within 1000 meters of the New York city, slightly below the city hall, another is roughly 2200 meters above New York city. Both areas are great shown by the analysis. As for customers' enjoyment, it is better to avoid the area just above New York city, it appeals to have a low rating regardless which analysis is performed.

6. **Further discussion**

There are some data information that could better help me answer the question such as the sales of each Starbucks store. It is possible that near the New York city, the sales are much higher, therefore much more profitable for opening stores.

Expense of coffeeshop is also a great data information to have. It can help me better understand how distance places a role in affecting the revenues of the stores.

It would be great if I could get more data from foursquare, but for some reasons, I can only get 50 store locations. All of those data can better help me with my analysis and to better answer the question.