Location Selection for New Business

Prasad K
January 5, 2020

1. Introduction

1.1 Background

Many people can't imagine starting their day without a cup of coffee in the morning. 66% of American women drink coffee every day compared to 62% of American men. An average American drinks 3.1 cups of coffee per day.

New York City has more coffee shops and cafes than any place else in the U.S. Manhattan's daytime population is approximately 4 million, so there is still lot of potential to open new coffee outlets.

1.2 Problem and Interests

A coffee house chain has 3 coffee retail stores in New York city. They wanted to expand their business by opening more stores in various locations in the city. They pre-chose 5 possible areas to select from. The location selection depends on the population working and living in the area along with venues/activities around the location. They wanted to make a study of the stores' data and neighborhood information to determine best locations for their new stores and to predict sales in those locations.

Other retailers can also be interested in similar information.

2. Data acquisition and cleaning

2.1 Data sources

Store information and neighborhood information are the two types of datasets required for this analysis.

2.1.1 Store and sales information

I found stores and sales data in Kaggle. It contains existing 3 stores and possible 5 new stores data. The existing store data consists of store locations and sales by each transaction for one month. The possible new stores data consists of just location information.

Sample Data:

sales_outlet.csv: This file has data with physical characteristics of all 8 store locations and a warehouse information.

sales_outlet_id,sales_outlet_type,store_square_feet,store_address,store_city,store_state_province,store_telephone,store_postal_code,store_longitude,store_latitude,manager,Neighorhood
2,warehouse,3400,164-14 Jamaica Ave,Jamaica,NY,972-871-0402,11432,-73.795168,40.705226,,Jamaica
3,retail,1300,32-20 Broadway,Long Island City,NY,777-718-3190,11106,-73.924008,40.761196,6,Astoria
4,retail,1300,604 Union Street,Brooklyn,NY,619-347-5193,11215,-73.983984,40.677645,11,Gowanus
:
:

sales_receipts.csv: This file contains data for each transaction in the existing 3 stores for a month.

"transaction_id","transaction_date","transaction_time","sales_outlet_id","staff_id","customer_id","instore_yn","order","line_item_id","product_id","quantity","line_item_amount","unit_price","promo_item_yn"
7,2019-04-01,12:04:43,3,12,558,N,1,1,52,1,2.50,2.50,N
11,2019-04-01,15:54:39,3,17,781,N,1,1,27,2,7.00,3.50,N
19,2019-04-01,14:34:59,3,17,788,Y,1,1,46,2,5.00,2.50,N
32,2019-04-01,16:06:04,3,12,683,N,1,1,23,2,5.00,2.50,N
:
:

2.1.2 Neighborhood information
Foursquare location services provide information on most popular nearby venues in the neighborhood of a given location.

I used the Foursquare API https://api.foursquare.com/v2/venues/explore to get the top 100 venues within 500 meters of each store location. There are a total of 800 venues obtained for 8 locations. Sample data of the first 5 venues is shown below.

Venue	Venue_Latitude	Venue_Longitude	Venue_Category
Astoria Bier & Cheese	40.760581	-73.922542	Cheese Shop
Yoga Agora	40.761200	-73.923862	Yoga Studio
Lockwood	40.760928	-73.924028	Gift Shop
Brooklyn Bagel & Coffee Co.	40.760408	-73.921967	Bagel Shop
King Of Falafel & Shawarma	40.762041	-73.925098	Middle Eastern Restaurant

2.2 Data pre-processing / cleaning

Initially the data that we received may not be in the format we need. It may also contains some noise i.e. data not needed for our analysis. So we need to clean the data so that it can be used in data analysis and machine learning.

Stores data:

- Spelling is corrected in the column header for neighborhood
- Deleted the data for warehouse since we are only going to use the stores information

store_square_feet	store_address	store_city	store_state_province	store_telephone	store_postal_code	store_longitude	store_latitude	manager	Neighborhood
1300	32-20 Broadway	Long Island City	NY	777-718-3190	11106	-73.924008	40.761196	6	Astoria
1300	604 Union Street	Brooklyn	NY	619-347-5193	11215	-73.983984	40.677645	11	Gowanus
900	100 Church Street	New York	NY	343-212-5151	10007	-74.010130	40.713290	16	Lower Manhattan
1000	122 E Broadway	New York	NY	613-555-4989	10002	-73.992687	40.713852	21	Lower East Side
1200	224 E 57th Street	New York	NY	287-817-2330	10021	-73.960000	40.770000	26	Upper East Side
1500	687 9th Avenue	New York	NY	652-212-7020	10036	-73.990338	40.761887	31	Hell's Kitchen
1700	175 8th Avenue	New York	NY	242-212-0080	10011	-74.000502	40.742760	36	Chelsea
1600	183 W 10th Street	New York	NY	674-646-6434	10014	-74.002722	40.734367	41	Greenwich Village

- From the above dataset, I took only neighborhood name, Latitude, and Longitude, which are the only needed features for our analysis.

	Store_Neighborhood	Store_Latitude	Store_Longitude
1	Astoria	40.761196	-73.924008
2	Gowanus	40.677645	-73.983984
3	Lower Manhattan	40.713290	-74.010130
4	Lower East Side	40.713852	-73.992687
5	Upper East Side	40.770000	-73.960000
6	Hell's Kitchen	40.761887	-73.990338
7	Chelsea	40.742760	-74.000502
8	Greenwich Village	40.734367	-74.002722

Sales data:

- Replaced the store number with store neighborhood name from stores dataset to easily identify the store location.
- Grouped the data by store by day to understand the sales trend

	Store_Neighborhood	transaction_date	line_item_amount
0	Astoria	2019-04-01	2571.40
1	Astoria	2019-04-02	2701.50
2	Astoria	2019-04-03	2759.05
3	Astoria	2019-04-04	2511.75
4	Astoria	2019-04-05	2669.55

- Grouped the data by store only to use in the modeling

	Store_Neighborhood	line_item_amount
0	Astoria	77213.23
1	Hell's Kitchen	79528.25
2	Lower Manhattan	76894.47

Neighborhood information from Foursquare:

- There are total of 800 venues in 203 categories in our dataset obtained in section 2.1.2 above. These are too many categories and some store locations missing many of these categories, so it may skew the results. I analyzed the data and decided to merge some of these categories into one. For example: categories that contain the word 'Restaurant', 'Diner', 'Steak', 'Bistro', 'BBQ' are grouped into one category called 'Restaurant'. As there is no base to show that coffee drinking habits of people are based on the cuisine they eat, grouped all of them into one category called 'Restaurant'. Now the categories are reduced to 23.

- After the above consolidation of categories, there are four categories in only 2 or 3 out of 8 store neighborhoods. Eg.: Children store. Removed data for these four categories reducing the categories to 19.

```
There are 19 uniques categories.

['Bakery',
 'Bar',
 'Clothing',
 'Coffee',
 'Dessert',
 'Food',
 'Grocery',
 'Gym',
 'Medical',
 'Miscellaneous Store',
 'Museum',
 'Music Place',
 'Outdoors',
 'Plaza',
 'Restaurant',
 'Shopping Mall',
 'Spa',
 'Theater',
 'Women Store']
```

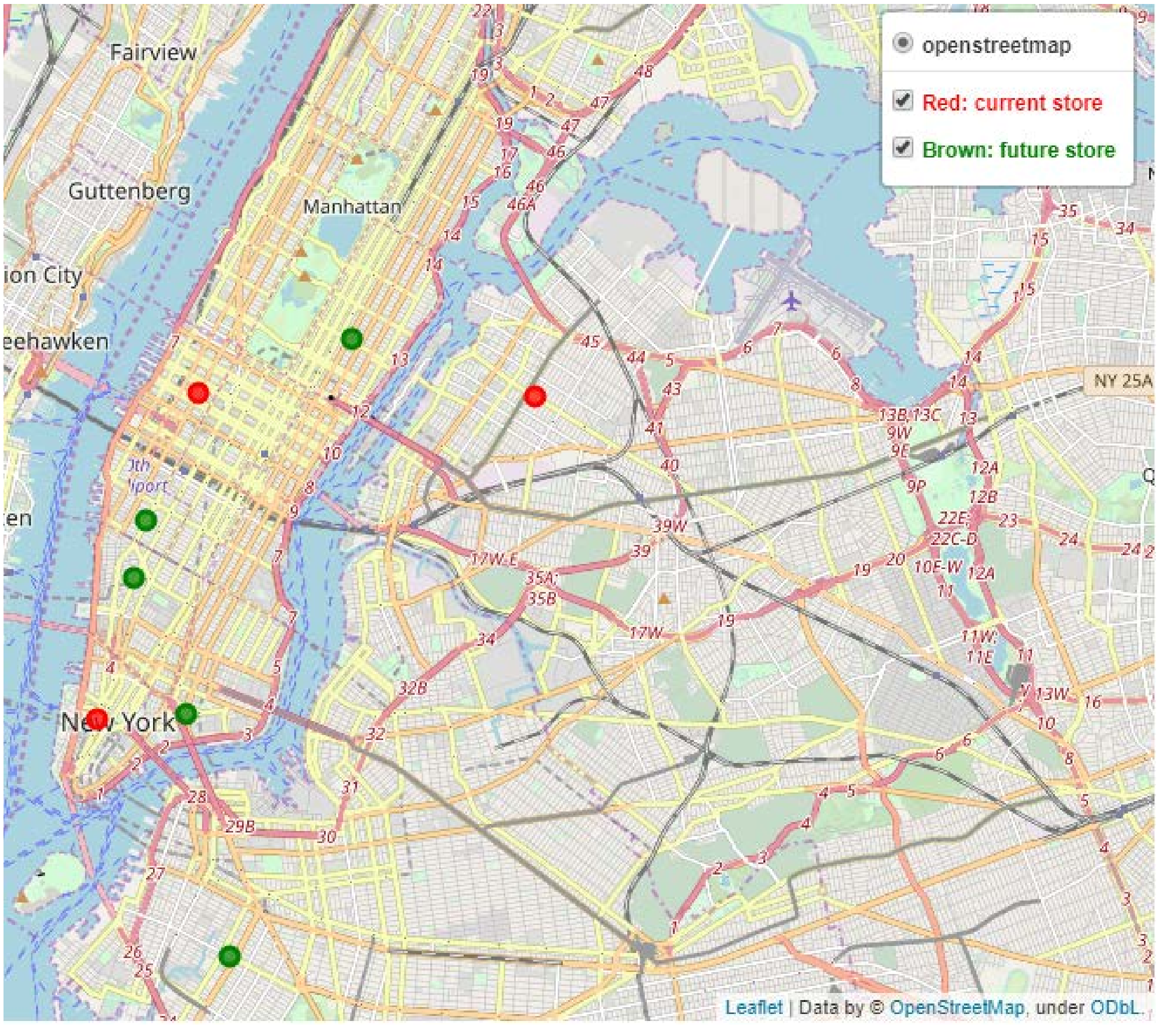
2.3 Features used in Modeling

Stores Data: Neighborhood (location name), store_latitude, and store_longitude
Sales Data: Store_Neighborhood (location name) and line_item_amount (store revenue in a month)
Foursquare Data: Venue_Categories

3. Data Analysis

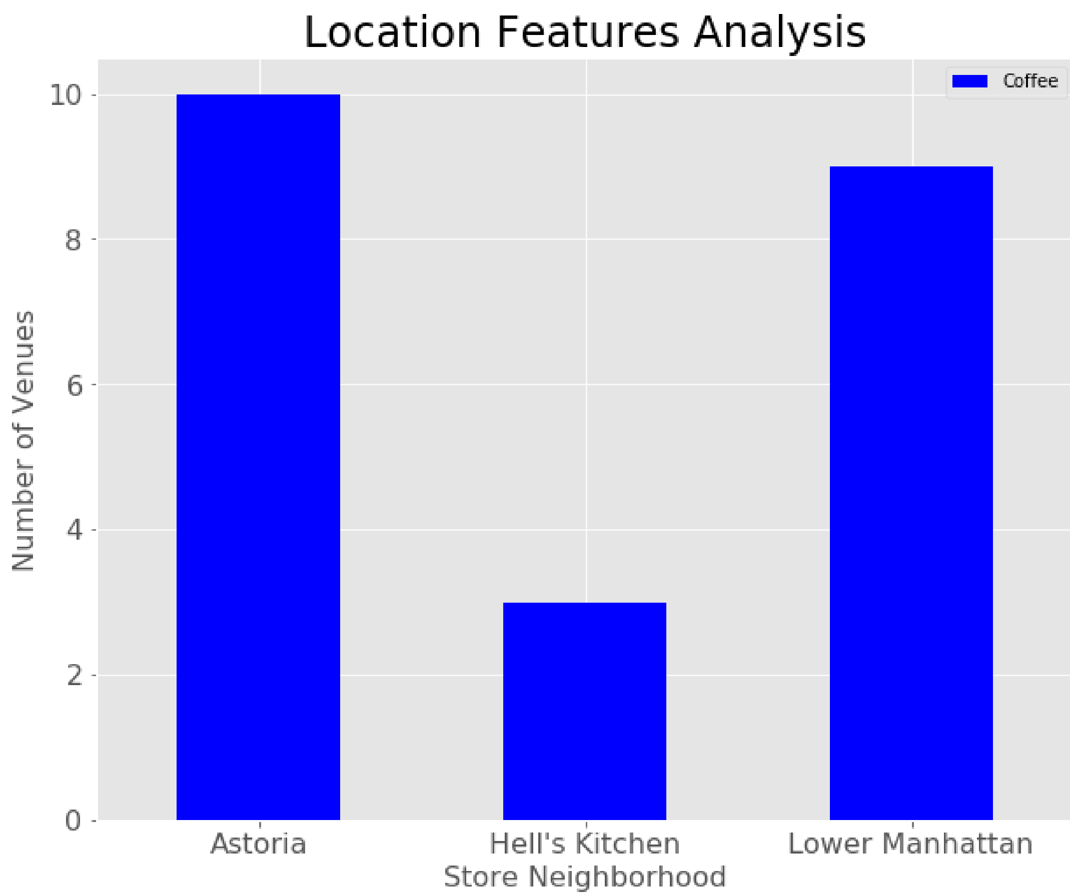
The location selection for new coffee shop depends on various factors. But if we analyze sales of the existing shops in comparison to nearby venues, we can observe few interesting points (some are obvious).

Store Locations

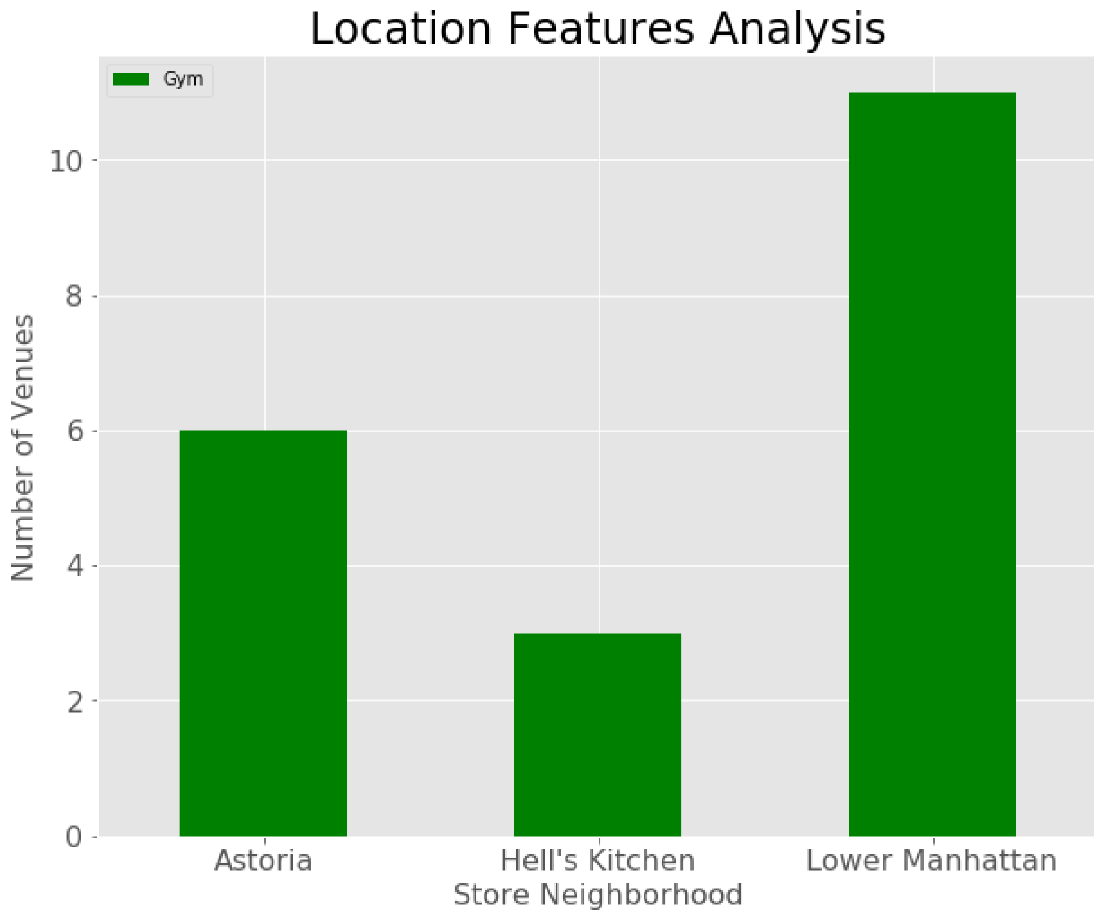


Store sales: Notice that the coffee shop in the Hell's Kitchen neighborhood has the highest sales.

Effect of number of Coffee shops in the neighborhood on the coffee shop sales:
The number of coffee shops in the Hell's Kitchen neighborhood is low compared to Astoria or Lower Manhattan neighborhoods, so the sales of Hell's Kitchen coffee shop is higher than the other two.



Effect of number of Fitness centers in the neighborhood on the coffee shop sales:
The number of fitness centers in the Hell's Kitchen neighborhood is low compared to Astoria or Lower Manhattan neighborhoods, but the sales of Hell's Kitchen coffee shop is high. This relationship is showing that people that visit gym may not like to drink coffee when they go for exercise.



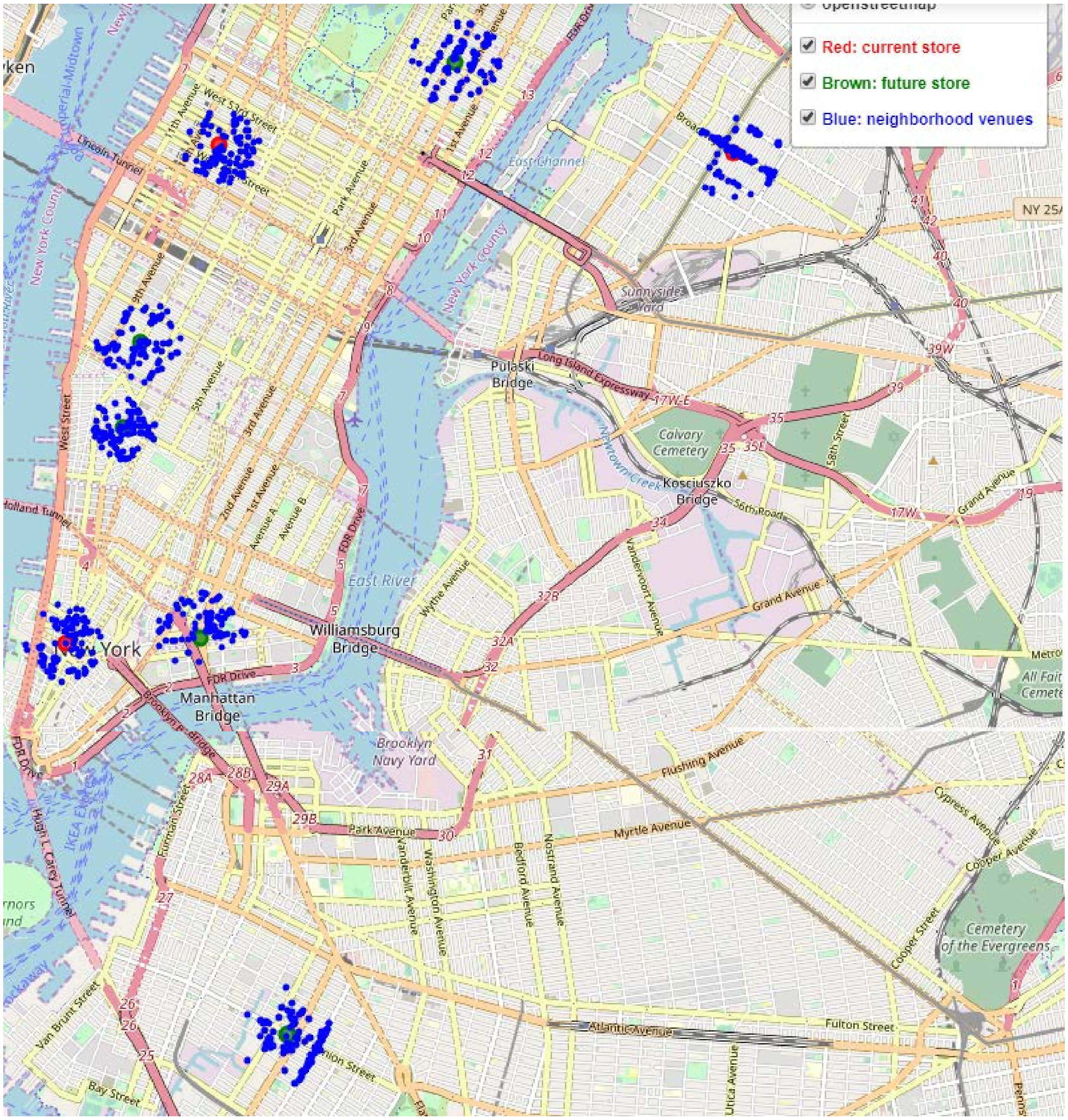
Though we can analyze the features individually, all of them together will have an impact on the coffee shop sales. I used the regression models and clustering as described in below section to predict the sales and to group the neighborhoods.

4. Predictive Modeling

I built and trained the linear regression and polynomial regression models with neighborhood features to predict sales for new store locations, . I also developed a model for clustering to compare the results with regression models.

4.1 Data preparation for modeling

1) First, popular neighborhood venues are obtained for each store location from Foursquare and cleaned the data as explained in data cleaning section above.



2) The venue categories are sorted and grouped by store location.

Store_Neighborhood	Bakery	Bar	Clothing	Coffee	Dessert	Food	Grocery	Gym	Medical	Miscellaneous Store	Museum	Music Place	Outdoors	Plaza	Restaurant	Shop
Astoria	6	20	0	10	2	10	2	6	1	3	0	0	0	0	35	
Chelsea	8	9	2	8	4	7	1	8	2	2	2	0	1	2	24	
Gowanus	3	15	1	8	2	13	3	13	1	4	1	1	0	1	27	
Greenwich Village	4	16	0	7	4	5	2	2	1	7	0	5	2	0	33	
Hell's Kitchen	4	16	1	3	1	9	0	3	0	2	0	2	1	1	38	
Lower East Side	3	15	0	7	5	5	2	3	1	6	2	1	2	1	40	
Lower Manhattan	2	8	4	9	2	9	2	11	0	4	2	1	6	8	19	
Upper East Side	2	10	1	5	2	11	2	10	1	5	2	0	0	2	34	

3) Normalized the data to have equal weightage for each category. If you think some venue types have more effect on coffee sales in the area than other venue types, then you can add more weightage to those categories.

Store_Neighborhood	Bakery	Bar	Clothing	Coffee	Dessert	Food	Grocery	Gym	Medical	Miscellaneous Store	Museum	Music Place	Outdoors	
Astoria	0.060000	0.200000	0.000000	0.100000	0.020000	0.100000	0.020000	0.060000	0.010000	0.030000	0.000000	0.000000	0.000000	0.
Chelsea	0.085106	0.095745	0.021277	0.085106	0.042553	0.074468	0.010638	0.085106	0.021277	0.021277	0.021277	0.000000	0.010638	0.
Gowanus	0.030612	0.153061	0.010204	0.081633	0.020408	0.132653	0.030612	0.132653	0.010204	0.040816	0.010204	0.010204	0.000000	0.
Greenwich Village	0.040816	0.163265	0.000000	0.071429	0.040816	0.051020	0.020408	0.020408	0.010204	0.071429	0.000000	0.051020	0.020408	0.
Hell's Kitchen	0.040404	0.161616	0.010101	0.030303	0.010101	0.090909	0.000000	0.030303	0.000000	0.020202	0.000000	0.020202	0.010101	0.
Lower East Side	0.030303	0.151515	0.000000	0.070707	0.050505	0.050505	0.020202	0.030303	0.010101	0.060606	0.020202	0.010101	0.020202	0.
Lower Manhattan	0.020202	0.080808	0.040404	0.090909	0.020202	0.090909	0.020202	0.111111	0.000000	0.040404	0.020202	0.010101	0.060606	0.
Upper East Side	0.021277	0.106383	0.010638	0.053191	0.021277	0.117021	0.021277	0.106383	0.010638	0.053191	0.021277	0.000000	0.000000	0.

4.2 Linear Regression

1) Build *linear regression* model (see the code for details)

2) Train the model: We have sales data for only 3 current stores, viz. Astoria, Hell's Kitchen, and Lower Manhattan, so trained the model with neighborhood data for all 3 locations.

Features:

Store_Neighborhood	Bakery	Bar	Clothing	Coffee	Dessert	Food	Grocery	Gym	Medical	Miscellaneous Store	Museum	Music Place	Outdoors	
Astoria	0.060000	0.200000	0.000000	0.100000	0.020000	0.100000	0.020000	0.060000	0.010000	0.030000	0.000000	0.000000	0.000000	0
Hell's Kitchen	0.040404	0.161616	0.010101	0.030303	0.010101	0.090909	0.000000	0.030303	0.000000	0.020202	0.000000	0.020202	0.010101	0
Lower Manhattan	0.020202	0.080808	0.040404	0.090909	0.020202	0.090909	0.020202	0.111111	0.000000	0.040404	0.020202	0.010101	0.060606	0.

Target values:

Store_Neighborhood	line_item_amount
Astoria	77213.23
Hell's Kitchen	79528.25
Lower Manhattan	76894.47

3) Then tested the model with the same data. The predicted values are same as the actual

values resulting in overfitting, which is not good.

```
([[77213.23],
  [79528.25],
  [76894.47]])
```

4) This time, trained the model with first two records and then used the last one for testing.
The predicted value is:

```
([[78636.18125991]])
```

```
Mean absolute error: 1422.95
% difference:  [[1.84288529]]
Residual sum of squares: 2024790.29
Variance score: nan
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\metrics\regression.py:543: UndefinedMetricWarning: R^2 score is not well-def
ined with less than two samples.
  warnings.warn(msg, UndefinedMetricWarning)
```

4.3 Polynomial Regression

1) Build second degree *polynomial regression* model (see the code for details)

2) Trained the model with first two samples, which are Astoria and Hell's Kitchen stores

```
x_train =  [[0.04040404 0.16161616 0.01010101 0.03030303 0.01010101 0.09090909
  0.         0.03030303 0.         0.02020202 0.         0.02020202
  0.02020202 0.01010101 0.38383838 0.         0.02020202 0.14141414
  0.01010101]
 [0.06       0.2        0.         0.1        0.02       0.1
  0.02       0.06       0.01       0.03       0.         0.
  0.         0.         0.35       0.01       0.02       0.01
  0.01      ]]
x_test =  [[0.02020202 0.08080808 0.04040404 0.09090909 0.02020202 0.09090909
  0.02020202 0.11111111 0.         0.04040404 0.02020202 0.01010101
  0.06060606 0.08080808 0.19191919 0.03030303 0.03030303 0.01010101
  0.05050505]]
y_train =  [[79528.25]
 [77213.23]]
y_test =  [[76894.47]]
```

3) Tested the model with the last store data, which is Lower Manhattan
The predicted value is:

```
([[76656.99089166]])
```

```
Mean absolute error: 237.48
% difference:  [[-0.30883769]]
R2-score: nan
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\metrics\regression.py:543: UndefinedMetricWarning: R^2 score is not well-def
ined with less than two samples.
  warnings.warn(msg, UndefinedMetricWarning)
```

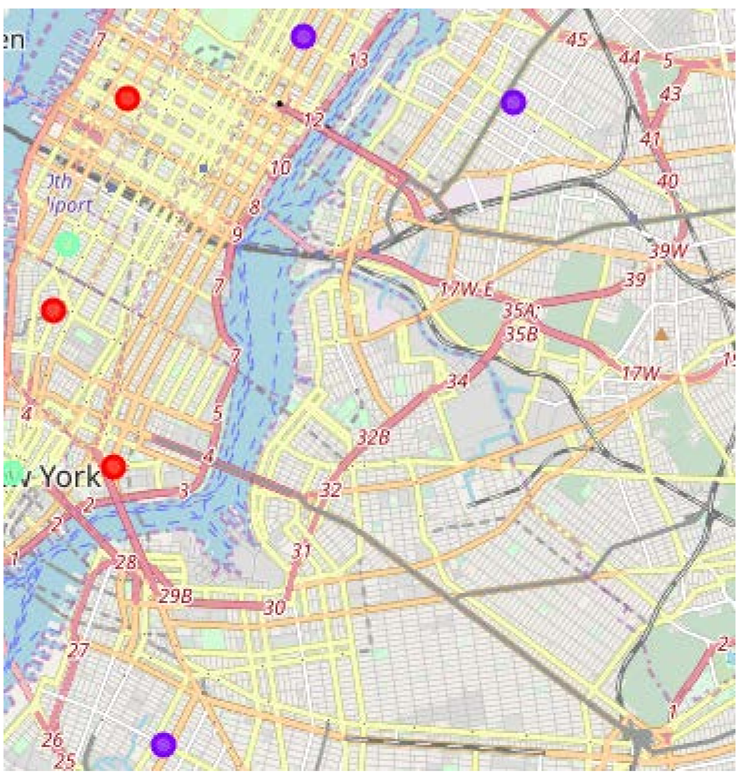
The results show that the polynomial regression is the better model than the linear regression.

4.4 k-means Clustering

1) Build *k-means clustering model* with k = 3 (see the code for details)

2) Use the normalized features from section 4.1 above to cluster the stores into 3 groups.

Store_Neighborhood	Store_Latitude	Store_Longitude	Cluster Labels	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 Mos Commo Venu
Astoria	40.761196	-73.924008	1	Restaurant	Bar	Coffee	Food	Bakery	Gym	Miscellaneous Store	Grocer
Gowanus	40.677645	-73.983984	1	Restaurant	Bar	Gym	Food	Coffee	Miscellaneous Store	Grocery	Baker
Lower Manhattan	40.713290	-74.010130	2	Restaurant	Gym	Coffee	Food	Bar	Plaza	Outdoors	Women Stor
Lower East Side	40.713852	-73.992687	0	Restaurant	Bar	Miscellaneous Store	Coffee	Food	Dessert	Bakery	Museun
Upper East Side	40.770000	-73.960000	1	Restaurant	Food	Gym	Bar	Women Store	Coffee	Miscellaneous Store	Museun
Hell's Kitchen	40.761887	-73.990338	0	Restaurant	Bar	Theater	Food	Bakery	Coffee	Gym	Miscellaneou Stor
Chelsea	40.742760	-74.000502	2	Restaurant	Bar	Gym	Bakery	Coffee	Food	Theater	Women Stor
Greenwich Village	40.734367	-74.002722	0	Restaurant	Bar	Miscellaneous Store	Music Place	Coffee	Food	Women Store	Desser



5. Results

5.1 Polynomial Regression

Used the *polynomial regression* model to predict the store sales for the new future locations.

Features:

Store_Neighborhood	Bakery	Bar	Clothing	Coffee	Dessert	Food	Grocery	Gym	Medical	Miscellaneous Store	Museum	Music Place	Outdoors	
Chelsea	0.085106	0.095745	0.021277	0.085106	0.042553	0.074468	0.010638	0.085106	0.021277	0.021277	0.021277	0.000000	0.010638	0.
Gowanus	0.030612	0.153061	0.010204	0.081633	0.020408	0.132653	0.030612	0.132653	0.010204	0.040816	0.010204	0.010204	0.000000	0.
Greenwich Village	0.040816	0.163265	0.000000	0.071429	0.040816	0.051020	0.020408	0.020408	0.010204	0.071429	0.000000	0.051020	0.020408	0.
Lower East Side	0.030303	0.151515	0.000000	0.070707	0.050505	0.050505	0.020202	0.030303	0.010101	0.060606	0.020202	0.010101	0.020202	0.
Upper East Side	0.021277	0.106383	0.010638	0.053191	0.021277	0.117021	0.021277	0.106383	0.010638	0.053191	0.021277	0.000000	0.000000	0.

Results:

Store_Neighborhood	Sales
Chelsea	77376.04
Gowanus	76842.46
Greenwich Village	77656.93
Lower East Side	77884.31
Upper East Side	77499.64

5.2 k-means Clustering

The model clustered the stores as follows:
Cluster 0: Lower East Side, Hell's Kitchen, Greenwich Village
Cluster 1: Astoria, Gowanus, Upper East Side
Cluster 2: Lower Manhattan, Chelsea

6. Discussion

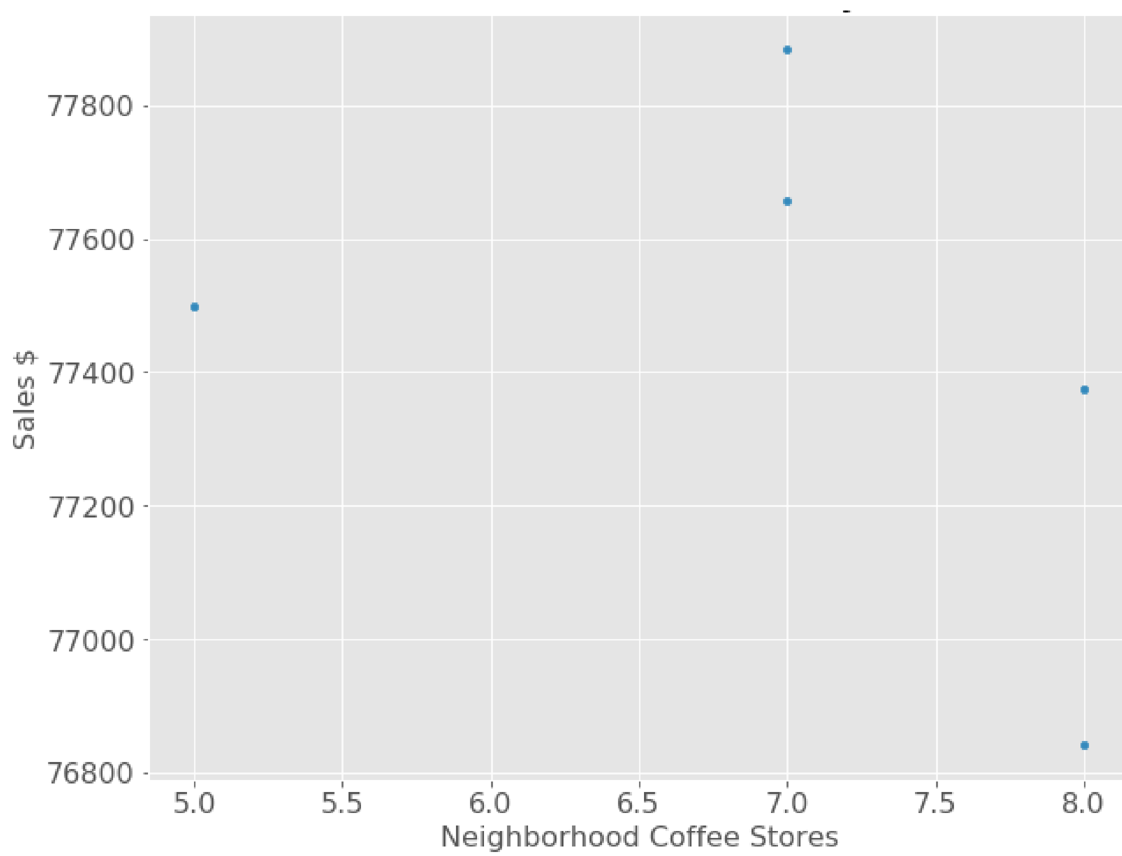
6.1 Polynomial Regression Results

Store_Neighborhood	Sales	Coffee	Gym
Chelsea	77376.04	8	8
Gowanus	76842.46	8	13
Greenwich Village	77656.93	7	2
Lower East Side	77884.31	7	3
Upper East Side	77499.64	5	10

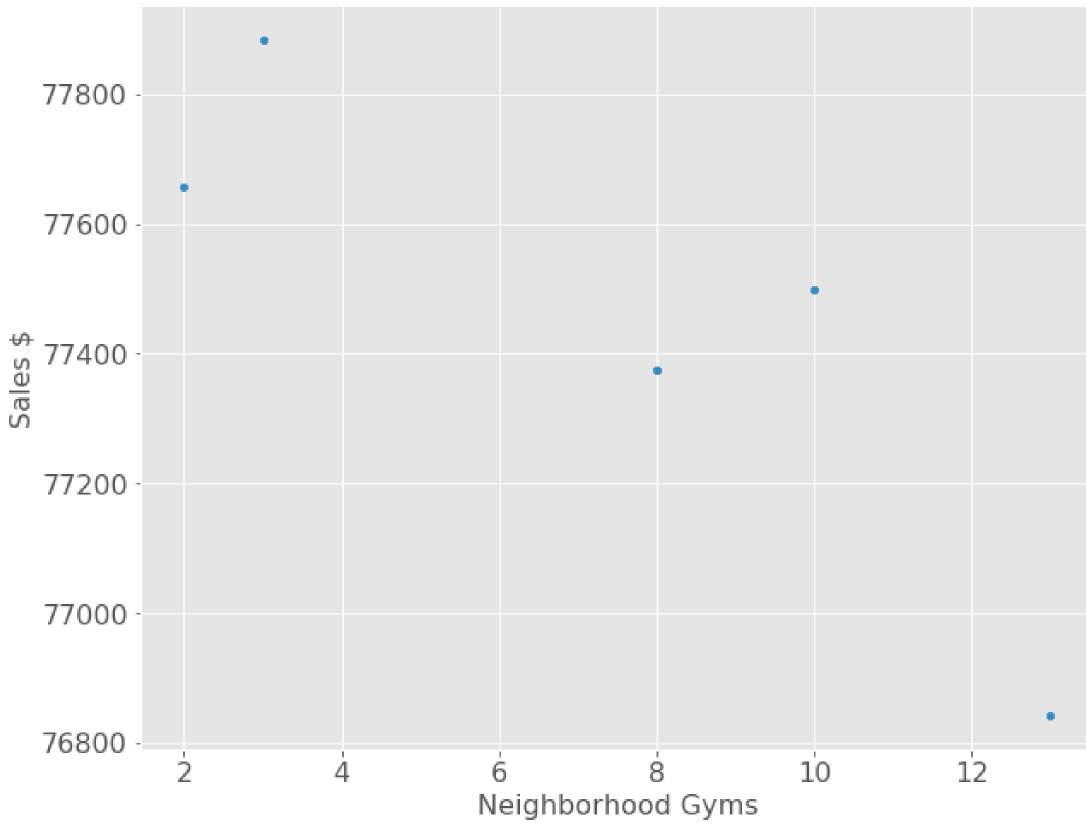
Based on the predicted sales from *polynomial regression* model, **Lower East Side or Greenwich Village neighborhoods are the best locations to start a new coffee store**.

Let's compare the results with our previous observations on the effect of number of coffee stores and fitness centers in the neighborhood on the coffee store sales.

Neighborhood Coffee Stores: Except one outlier, it is showing that the sales will decrease as the number of neighborhood coffee stores increase. This agrees with our previous observation on the the training data.



Neighborhood Gyms: This is also showing trend that the sales will decrease as the number of neighborhood fitness centers increase. This concurs our observation on the training data.



6.2 k-means Clustering

The clustering model grouped the Lower East Side, Hell's Kitchen, and Greenwich Village stores into one cluster based on the neighborhood features. This outcome is matching with the *Polynomial regression* model outcome, which is **Lower East Side and Greenwich Village coffee stores will earn the top sales along with Hell's Kitchen neighborhood store**.

7. Conclusion

This study shows how to obtain neighborhood data from Foursquare and analyze it. It also teaches how to prepare the data by normalizing it, how to build a model, train, test, and predict results with new dataset or clustering the data.

Though the dataset contains just couple of samples to train *Polynomial regression* model, the results matched with the results of *k-means clustering* model. There are not many retail store level datasets available for public use, but found this small data set in Kaggle. The results will definitely get better with dataset that is large enough and with additional features like population living, working, and visiting the neighborhood of the store locations.

Since the sales data by employee is available in the sales receipts file, we can also do the analysis on the employee performance. Similarly, the sales file has sales transactions by customer id and customer file contains their date of birth. Using this information, we can build coffee drinking profiles of customers by age groups. We can continue to do many different types of analyses based on the requirement.