Location Selection for New Business

Data Science / Machine Learning Capstone Project

Agenda

- Introduction
- Data Acquisition and Cleaning
- Data Analysis
- Predictive Modeling
    - Data Preparation
    - Linear Regression
    - Polynomial Regression
    - k-means Clustering
- Results
- Discussion
- Conclusion

Introduction

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.

Interest: 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. 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.

Data Acquisition and Cleaning

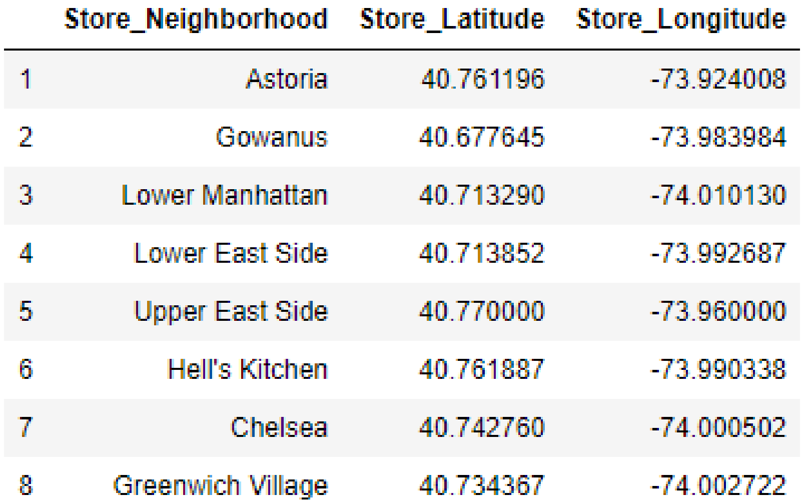
Data Acquisition

- Store location and sales information are found in Kaggle
- Neighborhood information (nearby popular venues and their categories) of all store locations is obtained from Foursquare location services

Data Cleaning

Removed data that is not needed for our analysis

- From store location file, took only store neighborhood name, latitude, and longitude

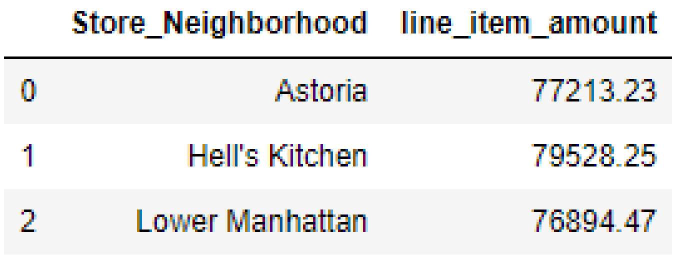


	Store_Neighborhood	Store_Latitude	Store_Longitude
1	Astoria	40.761196	-73.924008
2	Gowanus	40.677645	-73.983984
3	Lower Manhattan	40.713290	-74.010130
4	Lower East Side	40.713852	-73.992687
5	Upper East Side	40.770000	-73.960000
6	Hell's Kitchen	40.761887	-73.990338
7	Chelsea	40.742760	-74.000502
8	Greenwich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

Data Acquisition and Cleaning

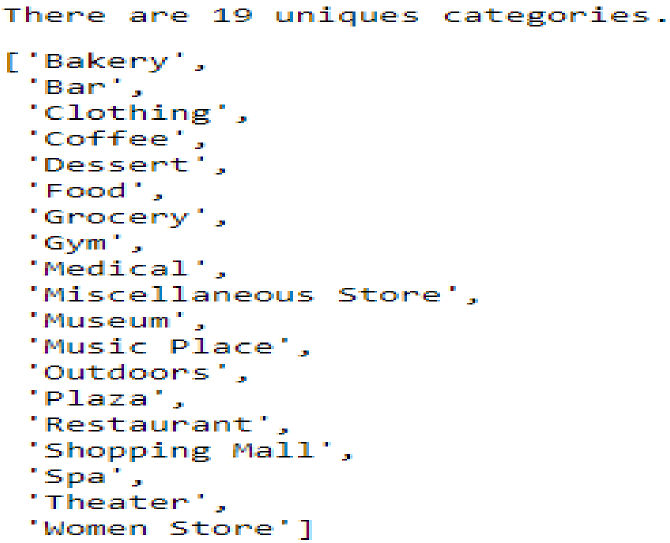
Data Cleaning

- Summarized the sales information by store



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

- In neighborhood data, grouped some of the categories into one. Eg: categories that contain the word 'Restaurant', 'Diner', 'Steak', 'Bistro', 'BBQ' are grouped into one category called 'Restaurant’.

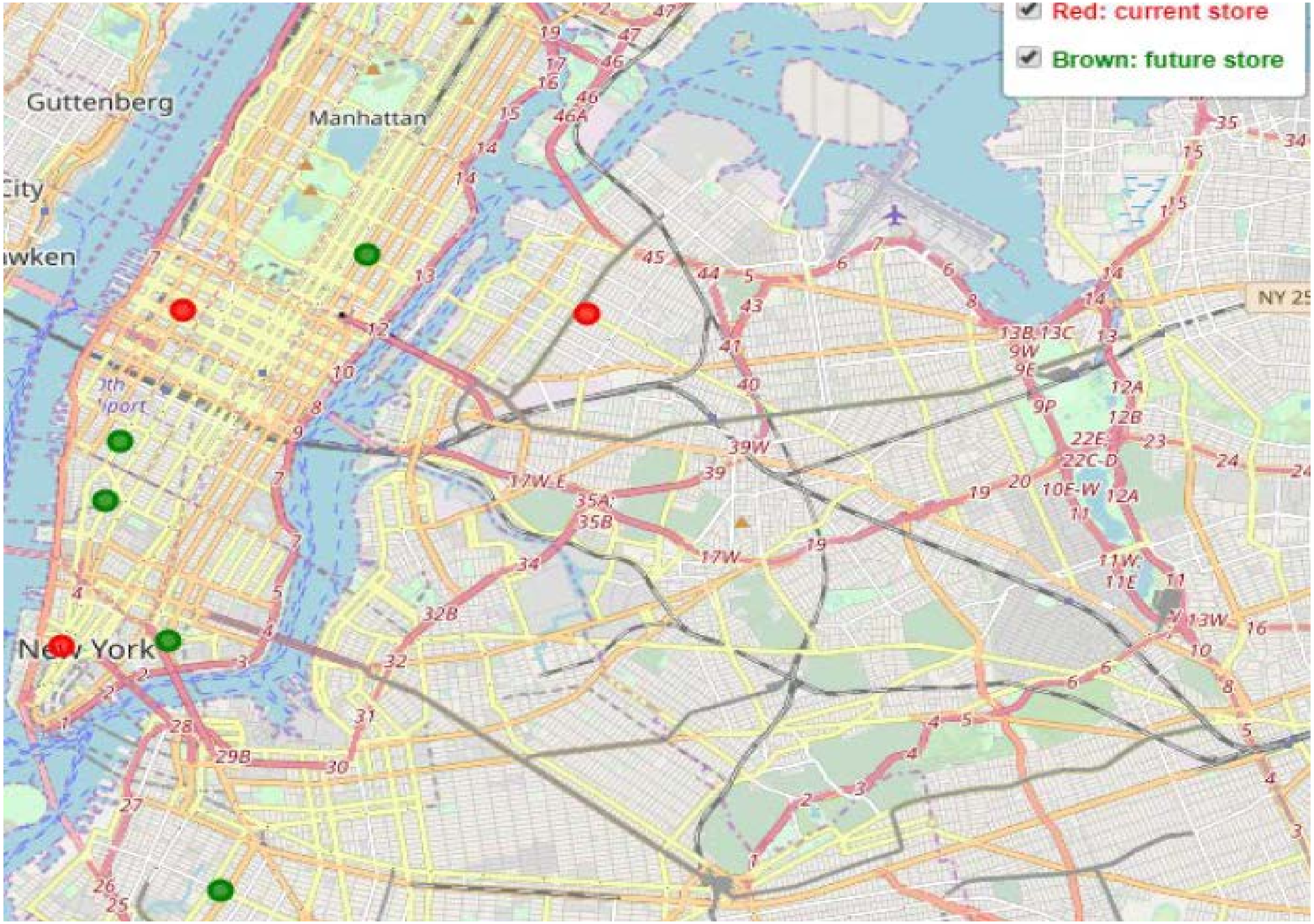


```
There are 19 uniques categories.

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

Data Analysis

Store Locations:

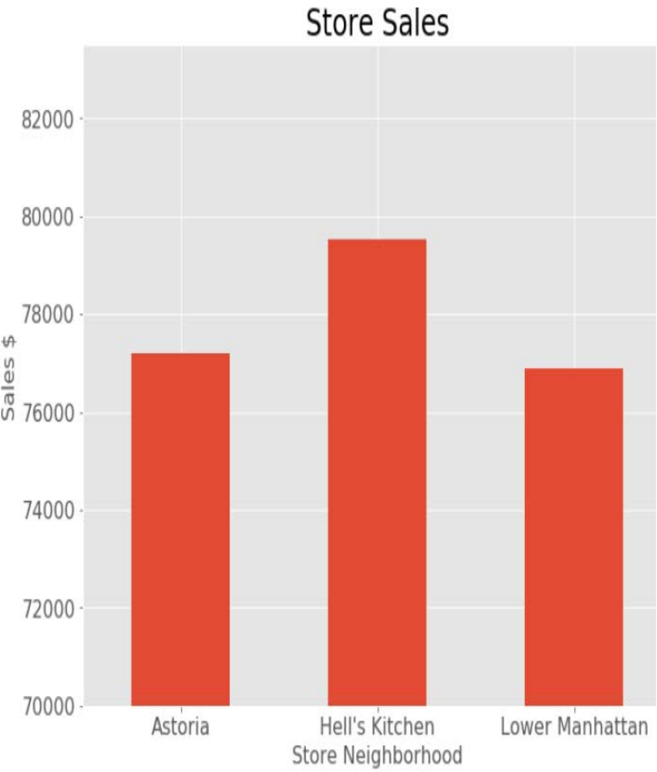


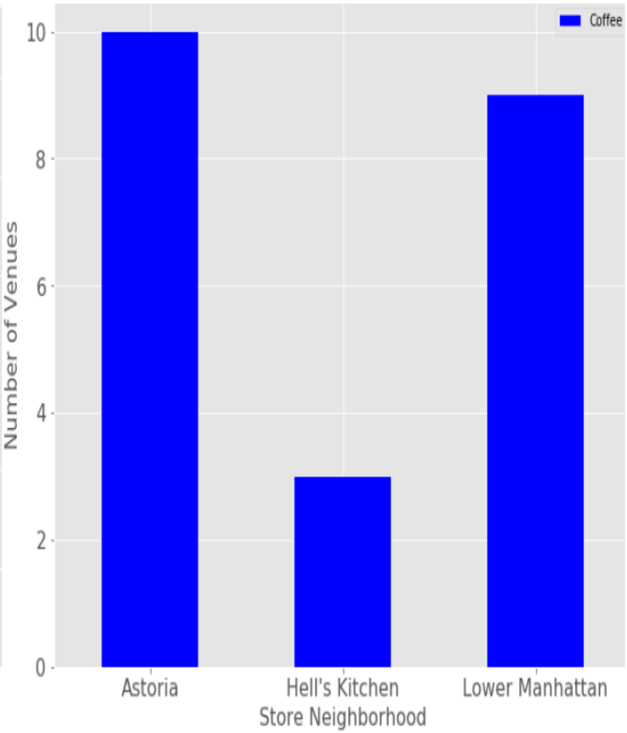
Data Analysis

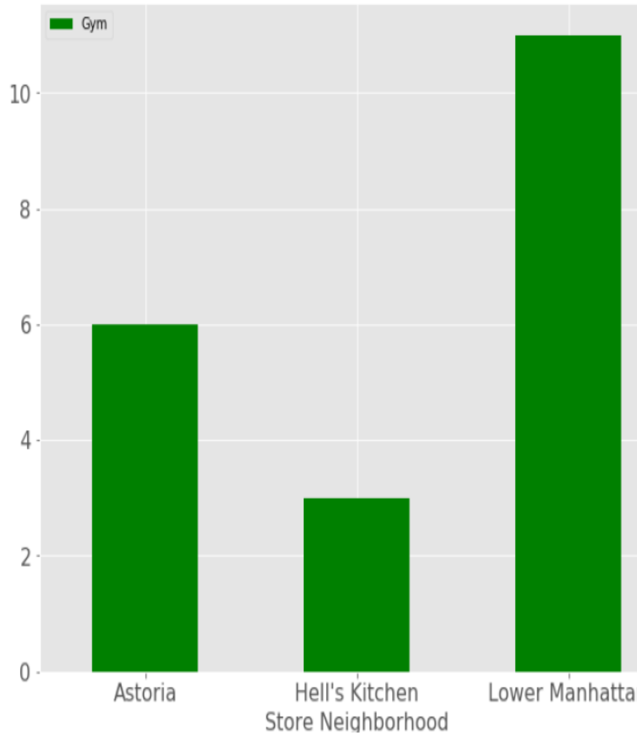
Store sales and effect of store features:

The Hell’s Kitchen neighborhood store sales are higher than the other two stores. Let’s analyze the impact of couple of features on the sales.

- Hell’s Kitchen neighborhood has less number of coffee shops than Astoria or Lower Manhattan neighborhoods, so our store in Hell’s Kitchen area performed well compared to the other two.
- Our data is showing that Lower Manhattan has the highest number of gyms and Hell’s Kitchen neighborhood has the lowest. It is also showing that the sales are in reverse order, i.e. coffee store sales are inversely proportional to the number of fitness centers in the neighborhood.



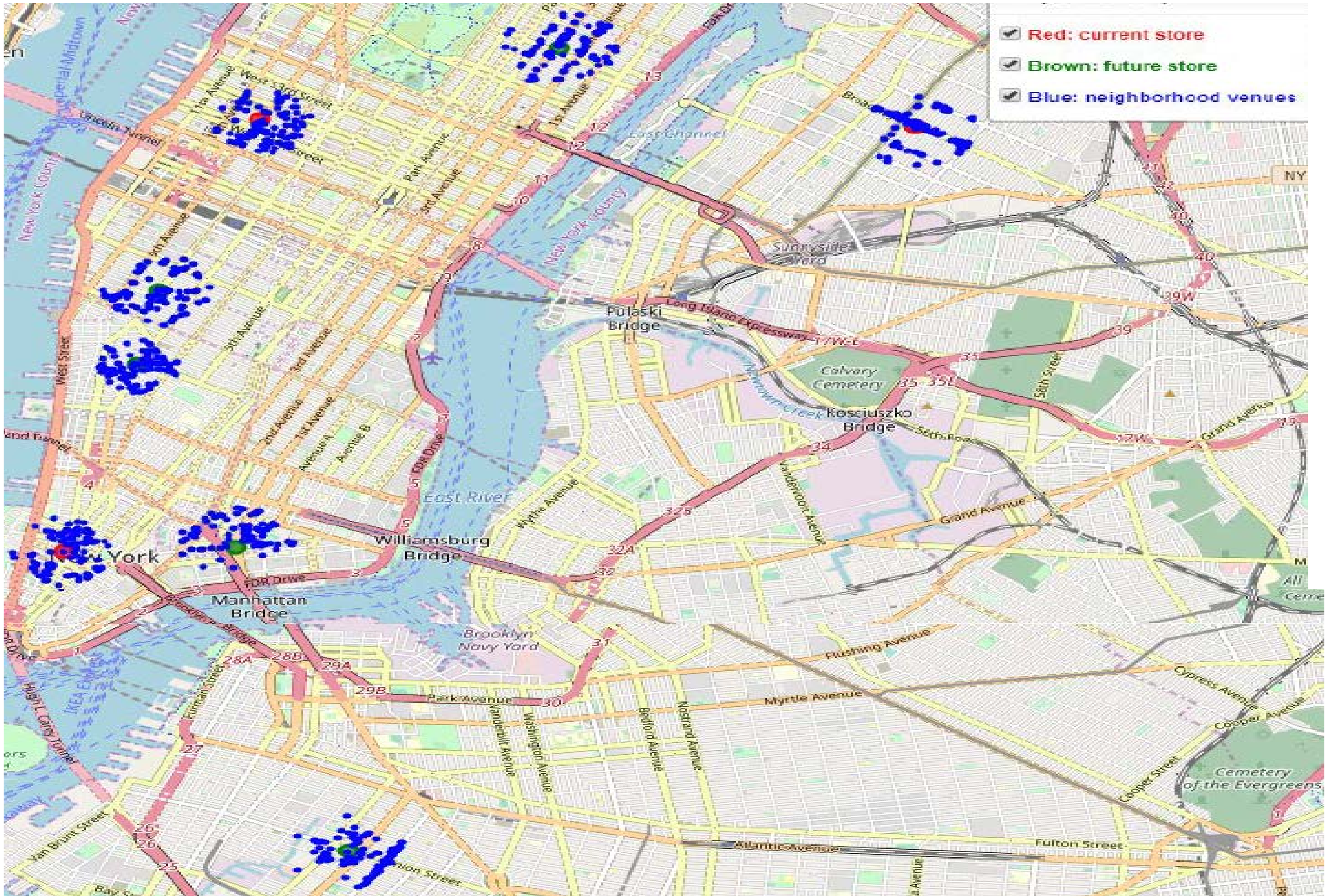




Predictive Modeling

Data Preparation:

- Get neighborhood top venues for each store location
- Group them by venue category for each store location
- Normalize the data to give equal weightage for all features



Predictive Modeling

Linear Regression:

- Build linear regression model

- Train the model with the 3 current stores data
- Test the model with the same data
  - The results are overfitting

- Train the model with 2 current stores data
- Test the model with third store data
  - The mean absolute error is 1422.95 which is about 1.84% error rate

Polynomial Regression:

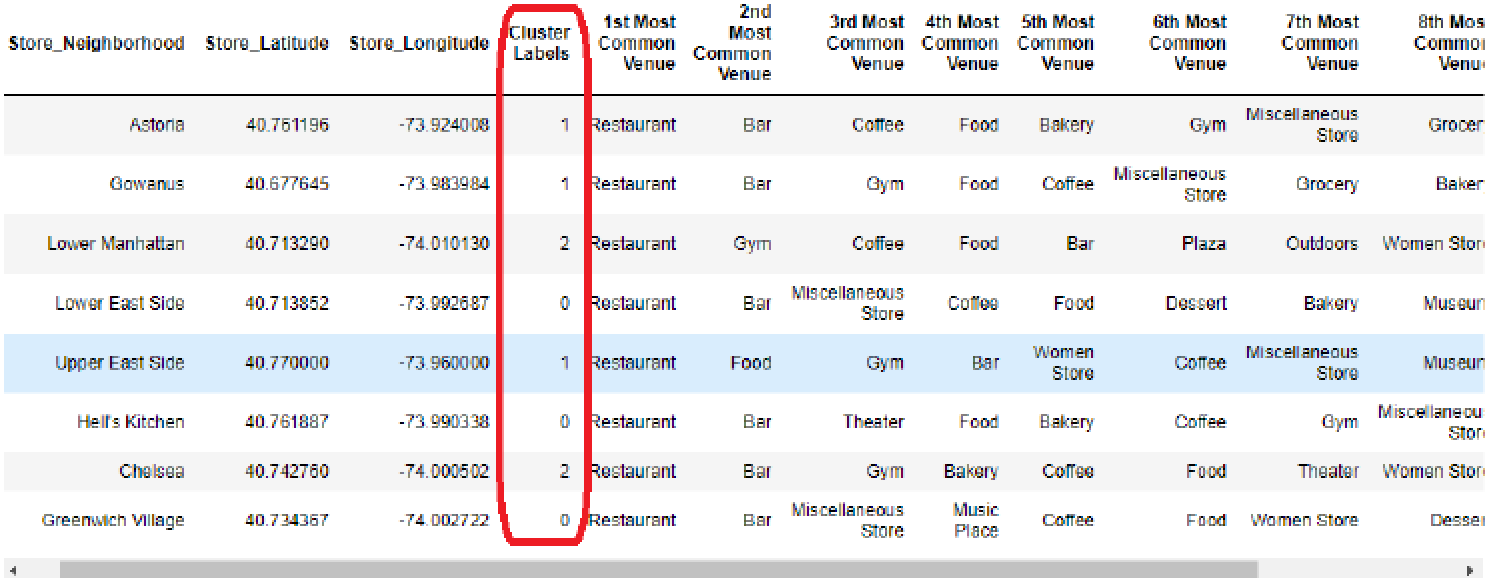
- Build second degree polynomial regression model
- Train the model with 2 current stores data
- Test the model with third store data
  - The mean absolute error is 237.48 which is about 0.31% error rate

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

Predictive Modeling

k-means Clustering:

- Build k-means clustering model with k = 3
- Use normalized data from all stores to train the model

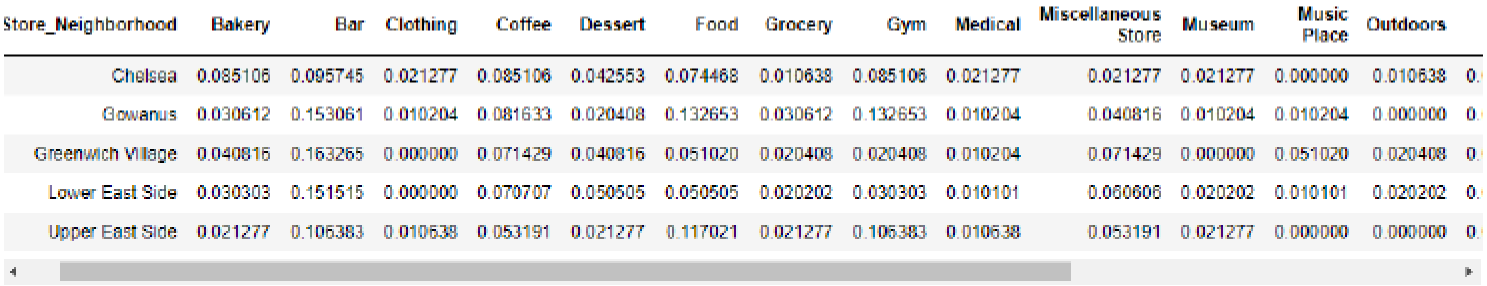


Store_Neighborhood	Store_Latitude	Store_Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Mos Commo Venu
Astoria	40.761196	-73.924008	1	Restaurant	Bar	Coffee	Food	Bakery	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	Museur
Upper East Side	40.770000	-73.960000	1	Restaurant	Food	Gym	Bar	Women Store	Coffee	Miscellaneous Store	Museur
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

Results

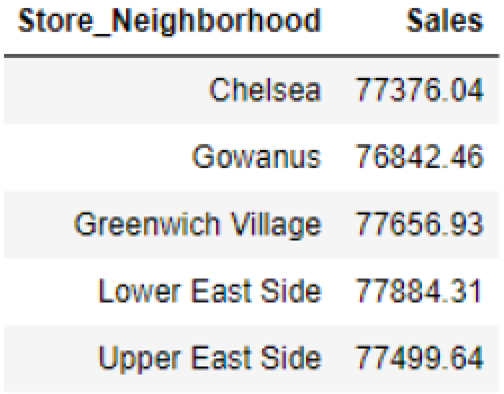
Polynomial Regression Model:

- Test the model with normalized data of the 5 new store locations



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.

- Predicted sales are:



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

Results

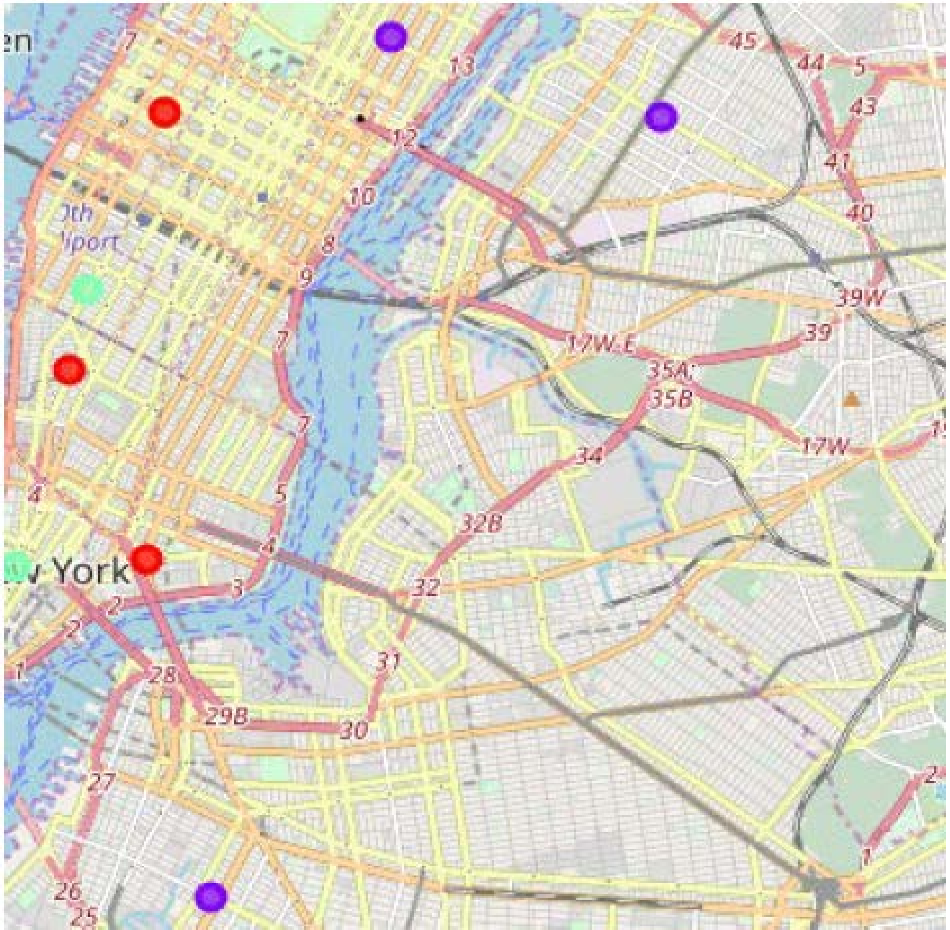
K-means Clustering:

- The clustering results are:

  Cluster 0: Lower East Side, Hell's Kitchen, Greenwich Village

  Cluster 1: Astoria, Gowanus, Upper East Side

  Cluster 2: Lower Manhattan, Chelsea

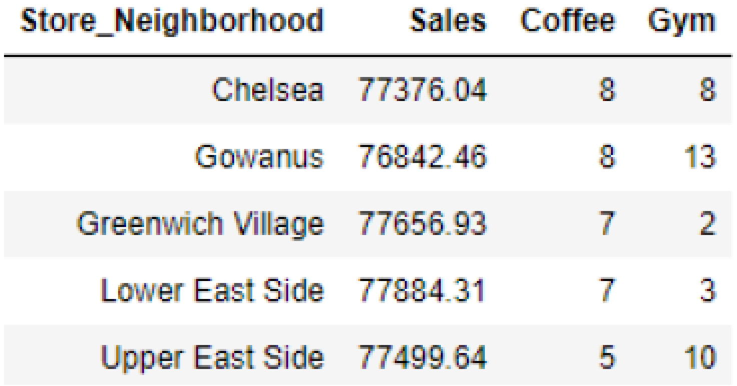


Discussion

Comparing Polynomial Regression and K-means Clustering models:

- Polynomial regression model predicted that the sales of the Lower East Side and Greenwich Village stores will have highest sales compared to other new proposed stores.
- K-means clustering model also grouped Lower East Side and Greenwich Village stores along with Hell's Kitchen area store into one cluster
- The results of both the polynomial regression model and k-means clustering model matched

We can also observe that the number of coffee shops and gyms are lower in the Lower East Side and Greenwich Village neighborhoods which may resulted in higher coffee sales. This observation matches the previous observation with the current stores.



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

Conclusion

This project provides information on:

- Acquiring data from customer datasets and from Foursquare location services
- Data cleaning and normalization
- Build various machine learning models
- Train and test the models
- Predict results and clustering the data

The results will get better with dataset that is large enough and with additional features like population living, working, and visiting the neighborhood of the store locations.

Additional analysis that can be done:

- 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.