Restaurant/Café location Identifier

Coursera Capstone: Battle of the Neighborhoods

By: Reem AlAbdouli

Business Problem

- New York allures international spotlight where it is one of the most sought after travel destination due to:
  - ○ Cultural, ethnic and natural diversity
  - ○ World best museums and art galleries
  - ○ Developed infrastructure & Fine educational institutions
  - ○ Being the heart of trade as economic growth & having the best technological, medical and scientific minds in the world.

- This project will focus on Manhattan because the possibilities are endless where it has:
  - ○ Dense population & Beautiful skyscrapers
  - ○ Lavish shopping &Tourist attractions & iconic historical structures
  - ○ Fine and performing arts
  - ○ Beautiful parks & Recreational facilities & some of best restaurants in the world

- Since New York is host to culinary experts from all across the globe and has one of the most competitive and diverse restaurant scenes in the world:
  - ○ Currently, not easy to casually predict if opening a certain restaurant/café in Manhattan will be successful or not.
  - ○ This is where this project makes a breakthrough in helping food business seekers to decide the best locations for their restaurant/café

So, the aim of this project is to use clustering techniques to group neighborhoods in Manhattan and analyze them which will support food business seekers to decide which neighborhood will be best suited for their business, this was rather hard without algorithms.

Data Sources

New York Neighborhood Dataset

- includes all New York Boroughs, Neighborhoods and locations (latitude, longitude)
- Will help in finding the required venues and information from Foursquare API

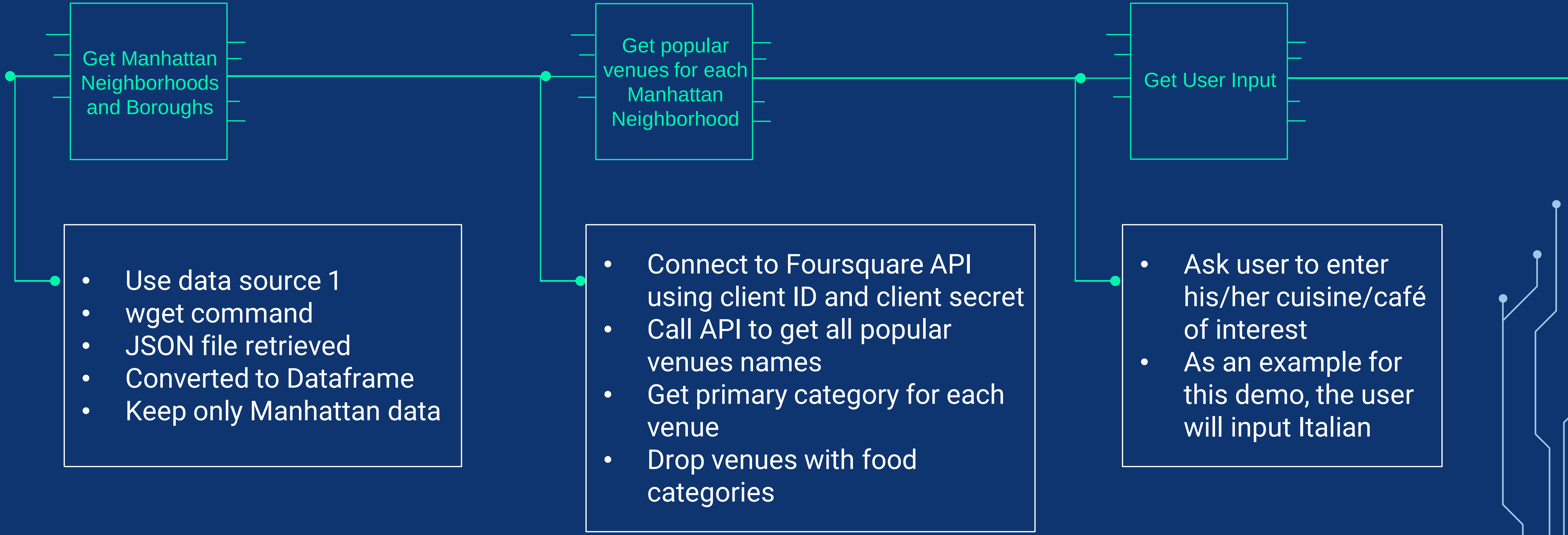
Data Source 1

Foursquare Places API

- Gives real-time access to Foursquare’s global database of rich venue data and user content.
- Will help in analyzing the neighborhoods in terms of the popular venues and food venues
- Provides features regarding venues such as rating and price which will be helpful in clustering.

Data Source 2

Methodology



Get Manhattan Neighborhoods and Boroughs

- Use data source 1
- wget command
- JSON file retrieved
- Converted to Dataframe
- Keep only Manhattan data

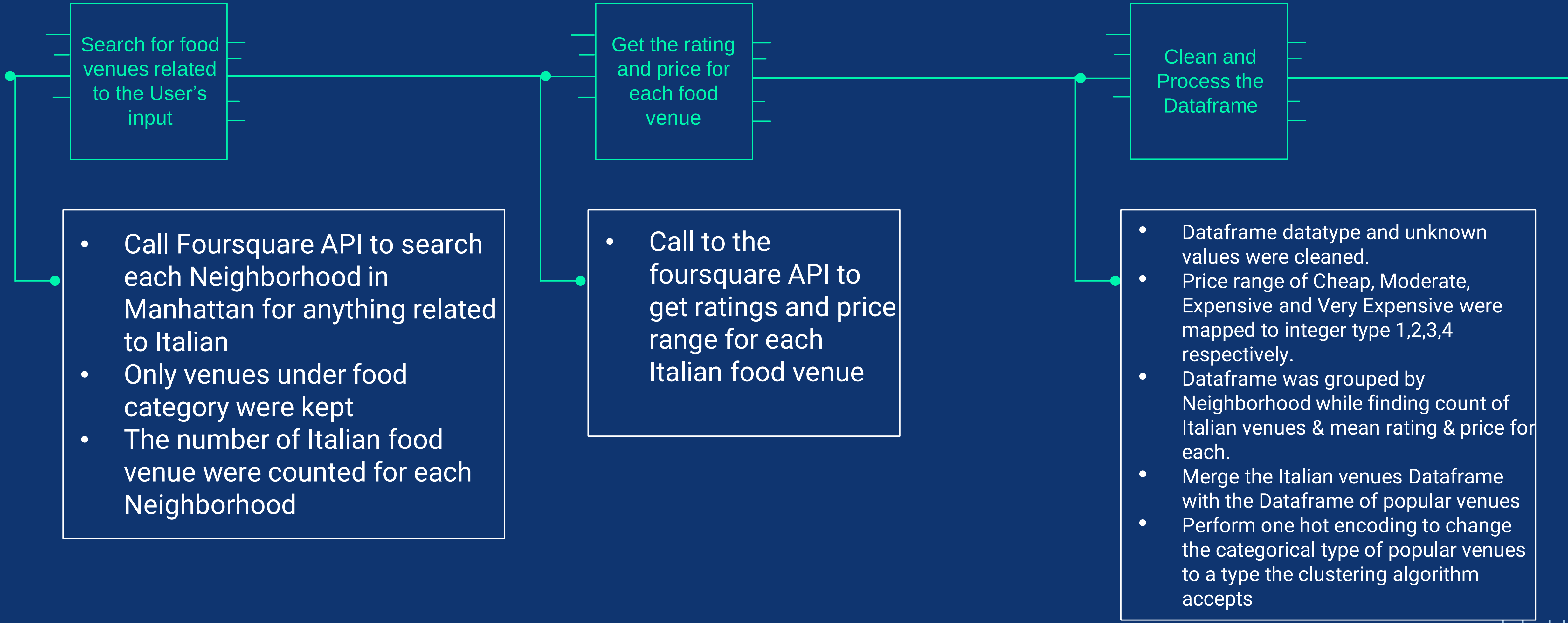
Get popular venues for each Manhattan Neighborhood

- Connect to Foursquare API using client ID and client secret
- Call API to get all popular venues names
- Get primary category for each venue
- Drop venues with food categories

Get User Input

- Ask user to enter his/her cuisine/café of interest
- As an example for this demo, the user will input Italian

Methodology



Search for food venues related to the User's input

- Call Foursquare API to search each Neighborhood in Manhattan for anything related to Italian
- Only venues under food category were kept
- The number of Italian food venue were counted for each Neighborhood

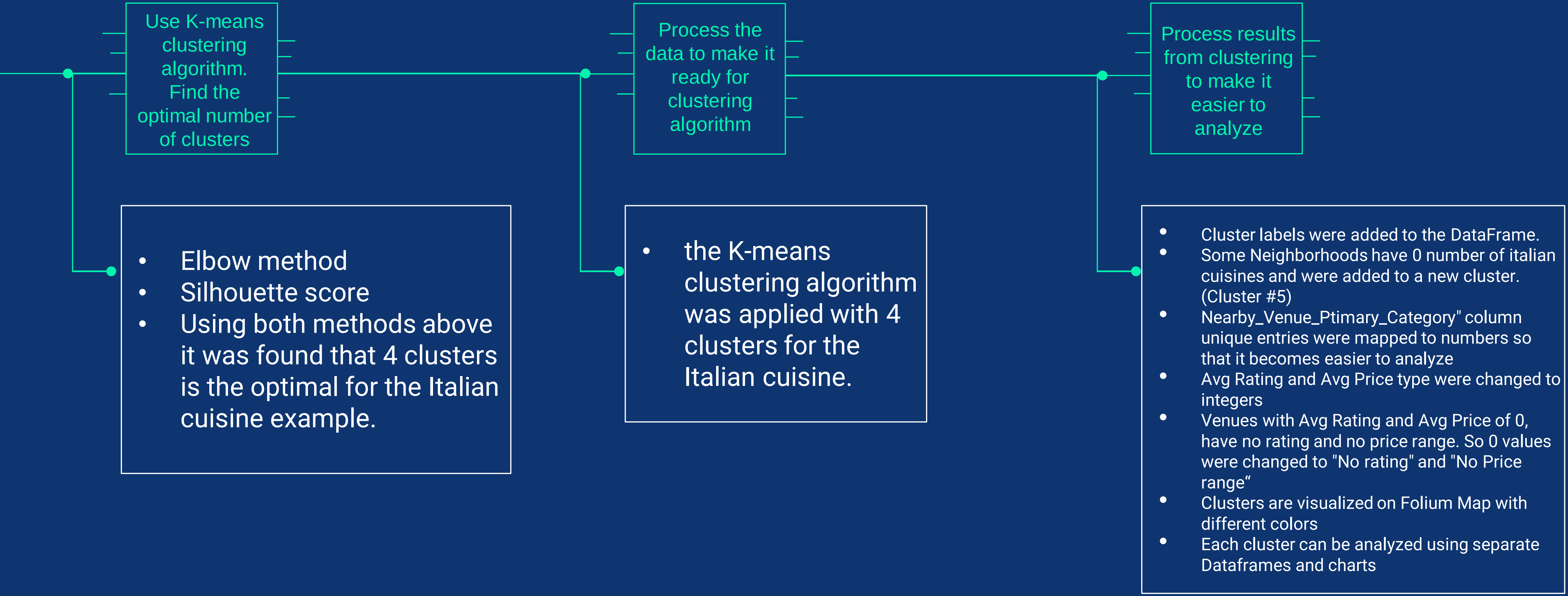
Get the rating and price for each food venue

- Call to the foursquare API to get ratings and price range for each Italian food venue

Clean and Process the Dataframe

- Dataframe datatype and unknown values were cleaned.
- Price range of Cheap, Moderate, Expensive and Very Expensive were mapped to integer type 1,2,3,4 respectively.
- Dataframe was grouped by Neighborhood while finding count of Italian venues & mean rating & price for each.
- Merge the Italian venues Dataframe with the Dataframe of popular venues
- Perform one hot encoding to change the categorical type of popular venues to a type the clustering algorithm accepts

Methodology



Use K-means clustering algorithm. Find the optimal number of clusters

- Elbow method
- Silhouette score
- Using both methods above it was found that 4 clusters is the optimal for the Italian cuisine example.

Process the data to make it ready for clustering algorithm

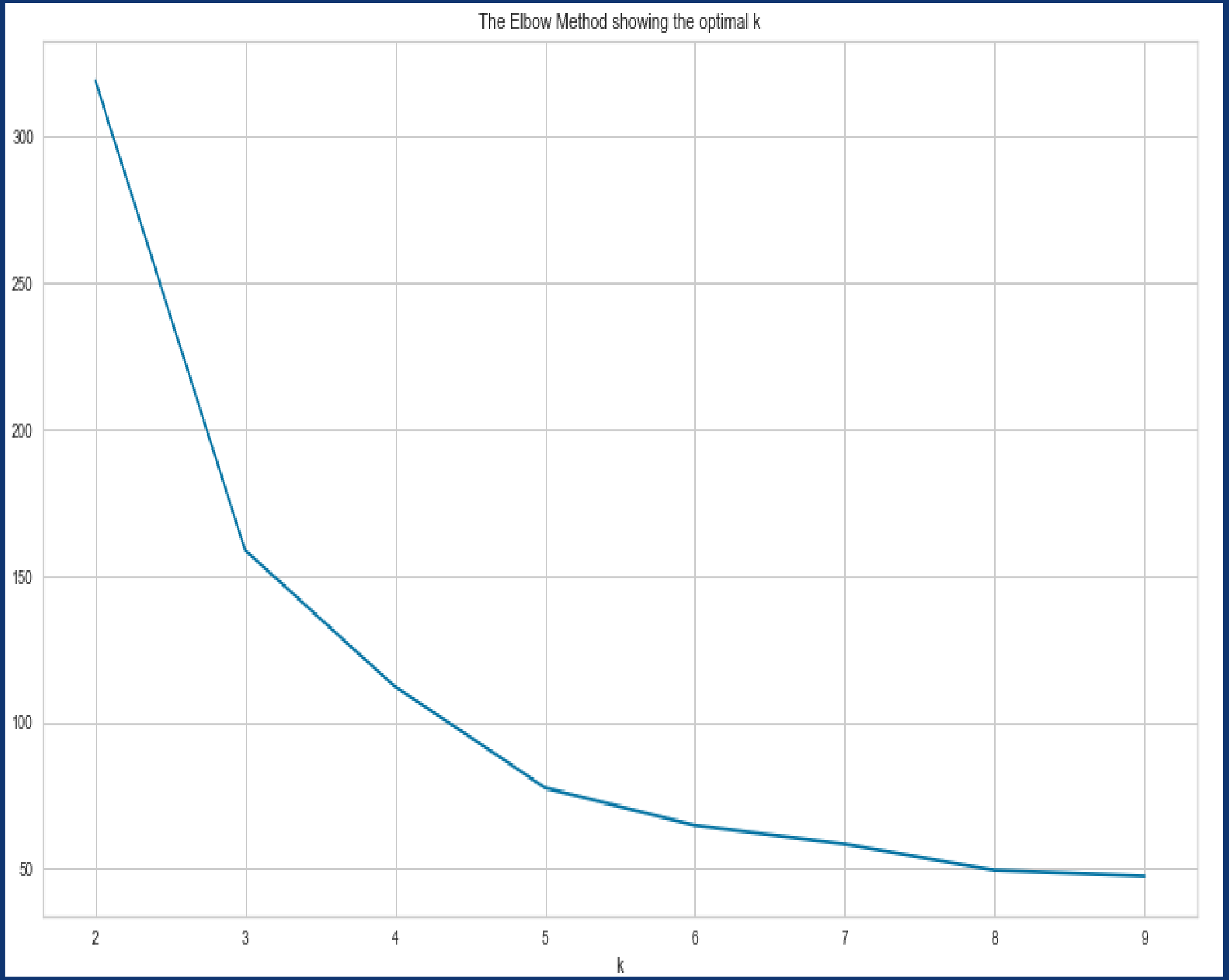
- the K-means clustering algorithm was applied with 4 clusters for the Italian cuisine.

Process results from clustering to make it easier to analyze

- Cluster labels were added to the DataFrame.
- Some Neighborhoods have 0 number of italian cuisines and were added to a new cluster. (Cluster #5)
- Nearby_Venue_Ptimary_Category" column unique entries were mapped to numbers so that it becomes easier to analyze
- Avg Rating and Avg Price type were changed to integers
- Venues with Avg Rating and Avg Price of 0, have no rating and no price range. So 0 values were changed to "No rating" and "No Price range“
- Clusters are visualized on Folium Map with different colors
- Each cluster can be analyzed using separate Dataframes and charts

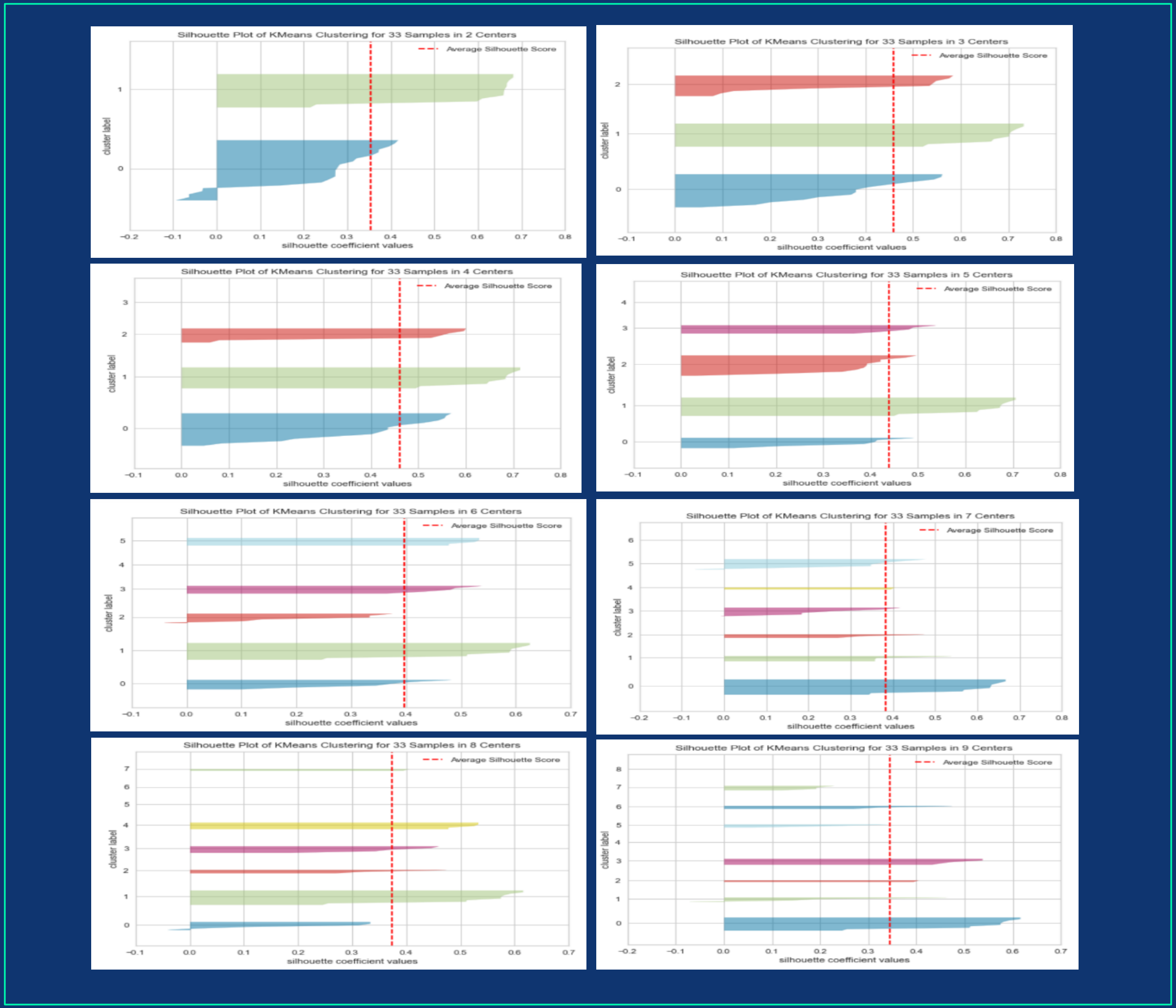
Results

- Finding Optimum number of clusters using elbow method.
- It shows 4 clusters is the optimum number.



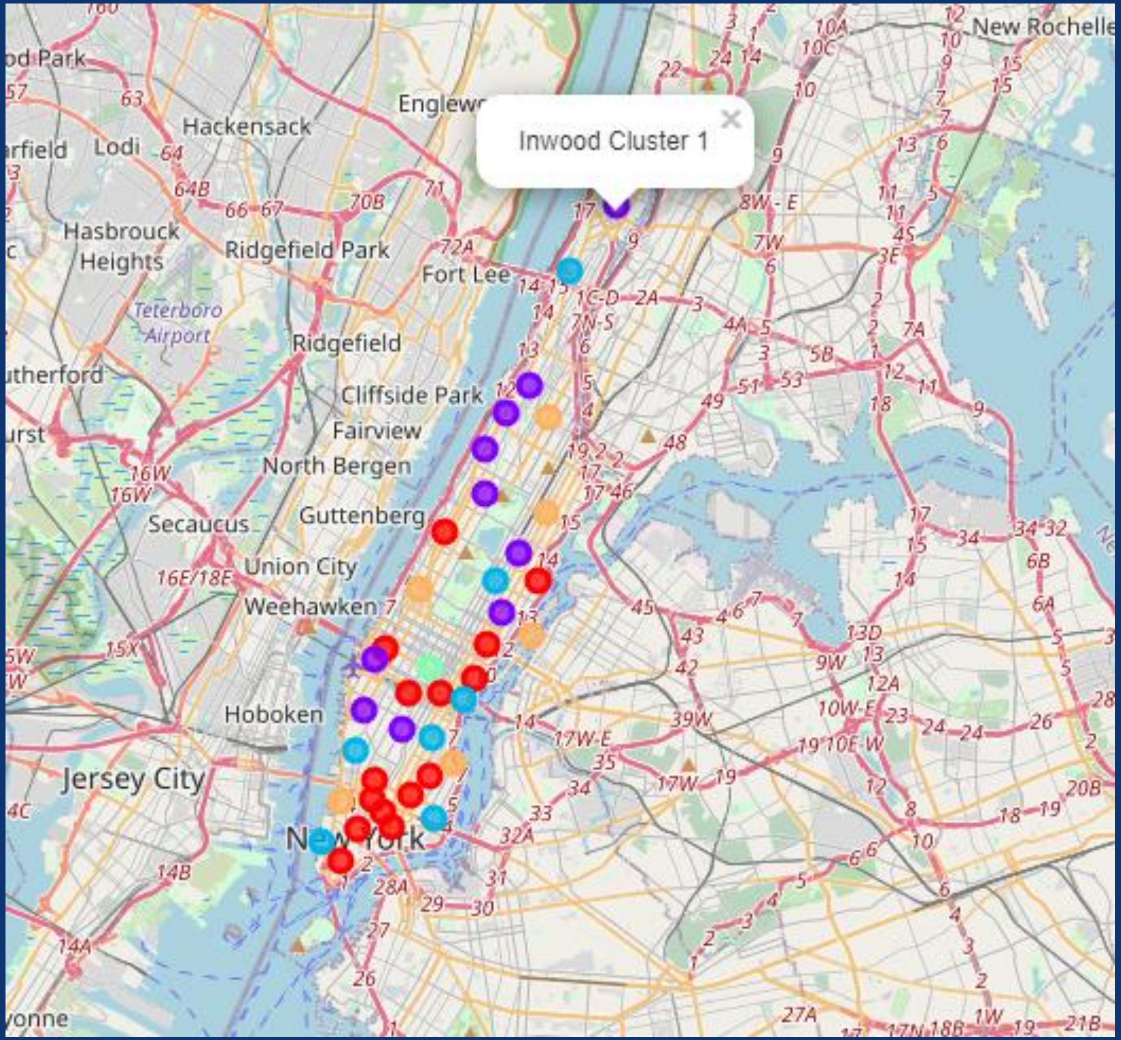
Results

- Finding optimum number of clusters using silhouette score
- Silhouette visualizer was used to compare Silhouette score for different number of clusters
- 2 Clusters, 3 Clusters, 4 Clusters, 5 Clusters and 7 Clusters all have Silhouette scores above average Silhouette score which make them candidate for the optimal number of clusters. However, fluctuation in size (thickness) of the silhouette plot representing each cluster also is a deciding point. So, 2 Clusters, 3 Clusters, 5 Clusters and 7 Clusters have more fluctuation in size as compared to 4 Clusters. For the plot with 4 Clusters, the thickness is more uniform than the plot with 2 Clusters, 3 Clusters, 5 Clusters and 7 Clusters. Thus, one can select the optimal number of Clusters as 4.
- 6 Clusters, 8 Clusters and 9 Clusters are not considered optimal number of Clusters because of the presence of Clusters with below-average silhouette scores and the fluctuations in the size of the silhouette plots



Results

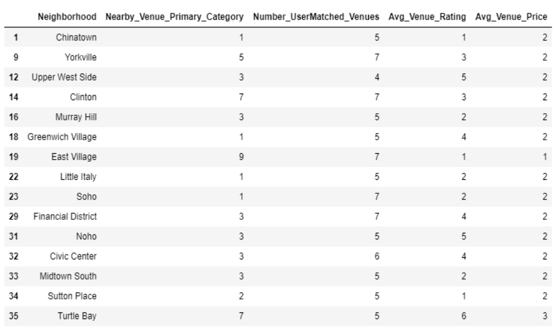
- After running kmeans clustering with 4 clusters, Follium map was used to display the clusters.
- Follium Map centered around manhattan showing different clusters with different colors.
- If a marker is clicked it shows the neighborhood name and cluster number.



Results

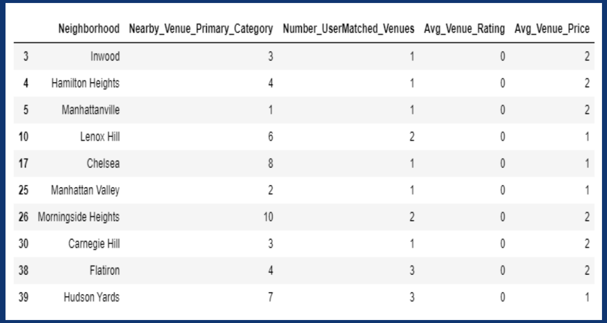
Dataframes of 5 clusters to help analyze

Cluster 1



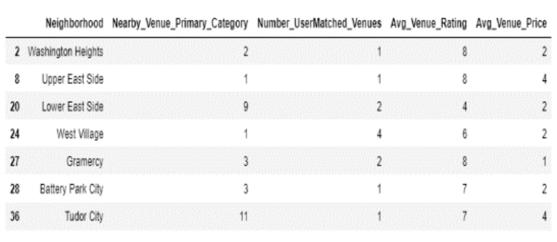
	Neighborhood	Nearby_Venue_Primary_Category	Number_UserMatched_Venues	Avg_Venue_Rating	Avg_Venue_Price
1	Chinatown	1	5	1	2
9	Yorkville	5	7	3	2
12	Upper West Side	3	4	5	2
14	Clinton	7	7	3	2
16	Murray Hill	3	5	2	2
18	Greenwich Village	1	5	4	2
19	East Village	9	7	1	1
22	Little Italy	1	5	2	2
23	Soho	1	7	2	2
29	Financial District	3	7	4	2
31	Noho	3	5	5	2
32	Civic Center	3	6	4	2
33	Midtown South	3	5	2	2
34	Sutton Place	2	5	1	2
35	Turtle Bay	7	5	6	3

Cluster 2



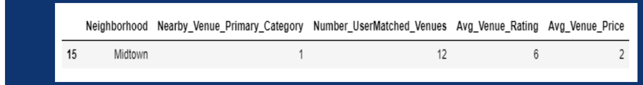
	Neighborhood	Nearby_Venue_Primary_Category	Number_UserMatched_Venues	Avg_Venue_Rating	Avg_Venue_Price
3	Inwood	3	1	0	2
4	Hamilton Heights	4	1	0	2
5	Manhattanville	1	1	0	2
10	Lenox Hill	6	2	0	1
17	Chelsea	8	1	0	1
25	Manhattan Valley	2	1	0	1
26	Morningside Heights	10	2	0	2
30	Carnegie Hill	3	1	0	2
38	Flatiron	4	3	0	2
39	Hudson Yards	7	3	0	1

Cluster 3



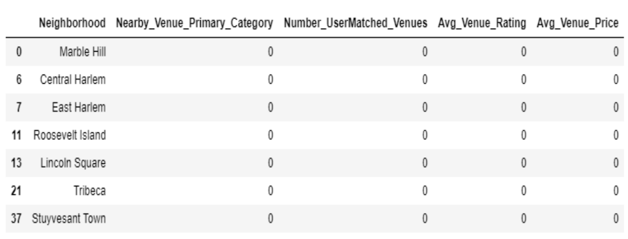
	Neighborhood	Nearby_Venue_Primary_Category	Number_UserMatched_Venues	Avg_Venue_Rating	Avg_Venue_Price
2	Washington Heights	2	1	8	2
8	Upper East Side	1	1	8	4
20	Lower East Side	9	2	4	2
24	West Village	1	4	6	2
27	Gramercy	3	2	8	1
28	Battery Park City	3	1	7	2
36	Tudor City	11	1	7	4

Cluster 4



	Neighborhood	Nearby_Venue_Primary_Category	Number_UserMatched_Venues	Avg_Venue_Rating	Avg_Venue_Price
15	Midtown	1	12	6	2

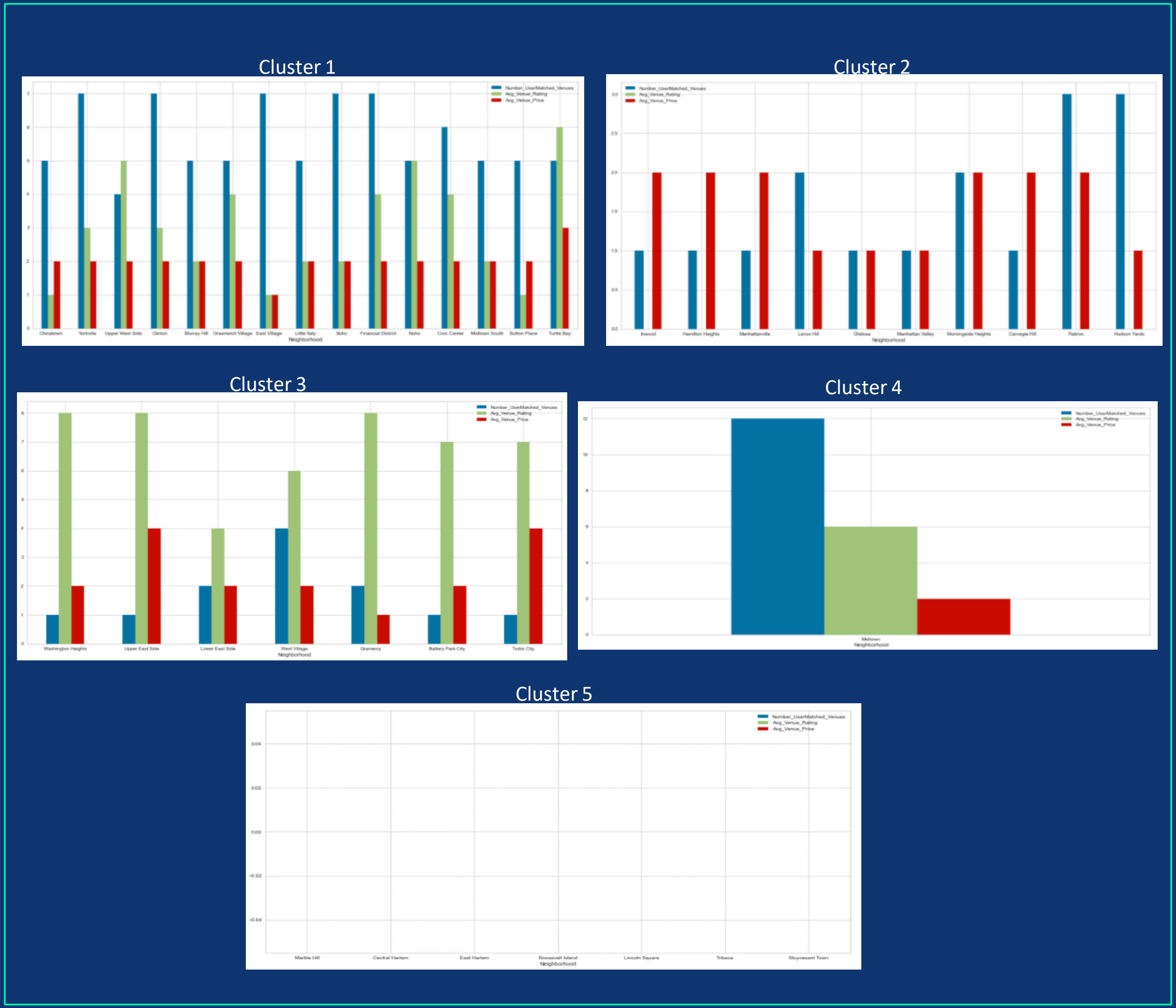
Cluster 5



	Neighborhood	Nearby_Venue_Primary_Category	Number_UserMatched_Venues	Avg_Venue_Rating	Avg_Venue_Price
0	Marble Hill	0	0	0	0
6	Central Harlem	0	0	0	0
7	East Harlem	0	0	0	0
11	Roosevelt Island	0	0	0	0
13	Lincoln Square	0	0	0	0
21	Tribeca	0	0	0	0
37	Stuyvesant Town	0	0	0	0

Results

- Showing chart for each cluster
- Each chart includes price range, avg rating and number of Italian venues.



Cluster 1

Cluster 2

Cluster 3

Cluster 4

Cluster 5

Discussion

- As noticed from the map, cluster tables and charts, Neighborhoods ['Marble Hill', 'Central Harlem', 'East Harlem', 'Roosevelt Island', 'Lincoln Square', 'Tribeca', 'Stuyvesant Town'] have 0 number of Italian cuisines. These Neighborhoods were grouped into a new cluster which is cluster #5.
- Clusters 1-4 are identified according to the number venues matched to the user's entry in the neighborhood, their average rating, average price and their surrounding popular spots.
- However, it is noticed that the popular spots in each Neighborhoods are kind of similar in each cluster.
- As seen in charts:
    - ○ Cluster #1 grouped Neighborhoods that have high number of Italian cuisine with Moderate price range and average rating
    - ○ Cluster #2 grouped Neighborhoods that have low number of Italian cuisine with Moderate price and no rating
    - ○ Cluster #3 grouped Neighborhoods that have low number of Italian cuisine with above Moderate price and high rating
    - ○ Cluster #4 grouped Neighborhoods that have very high number of Italian cuisines with Moderate price and Moderate rating
    - ○ Cluster #5 grouped Neighborhoods that have 0 number of Italian cuisines

Discussion

- The choice of which cluster to open up the Italian cuisine depends on the user's target for the Italian restaurant. So if the user is thinking of opening up a luxurious Italian restaurant then the price range will be above Moderate and so Cluster #2 or #5 might be good candidates.
- However, there should be more information on the neighborhoods such as demographics, social and economic characteristics of the people which will help to know people's interests and tendencies. So, there will be more insight and in this way it will be known, for example, if there is demand for Italian cuisine in Cluster #5 or not.

Conclusion

- New York is the core of the best restaurants globally with various cuisines and international culinary experts.
- Opening up a food business in Manhattan may be a hard decision as there is a strong of competition.
- With the help of an algorithm, identifying the best location for food business will be much simpler.
- This project is aimed to help food business seekers to decide which Neighborhood in Manhattan is best suited for their business.
- Only two data sources were used and in my opinion there is room for improvement with the use of additional data sources. For example, the
    - Use of Foursquare Places Databases will add more features regarding the venues that are not available in the Foursquare Places AP such as service quality, whether the food venue is crowded, whether food is worth the price and whether the food venue is trendy.
    - Using data about the demographic, social and economic characteristics of the people in Manhattan would’ve helped to analyze and understand people’s interests and tendencies. Adding relevant data will improve the clustering of Neighborhoods and enhance the identification of the best Neighborhood for the food business seeker.

Thank you