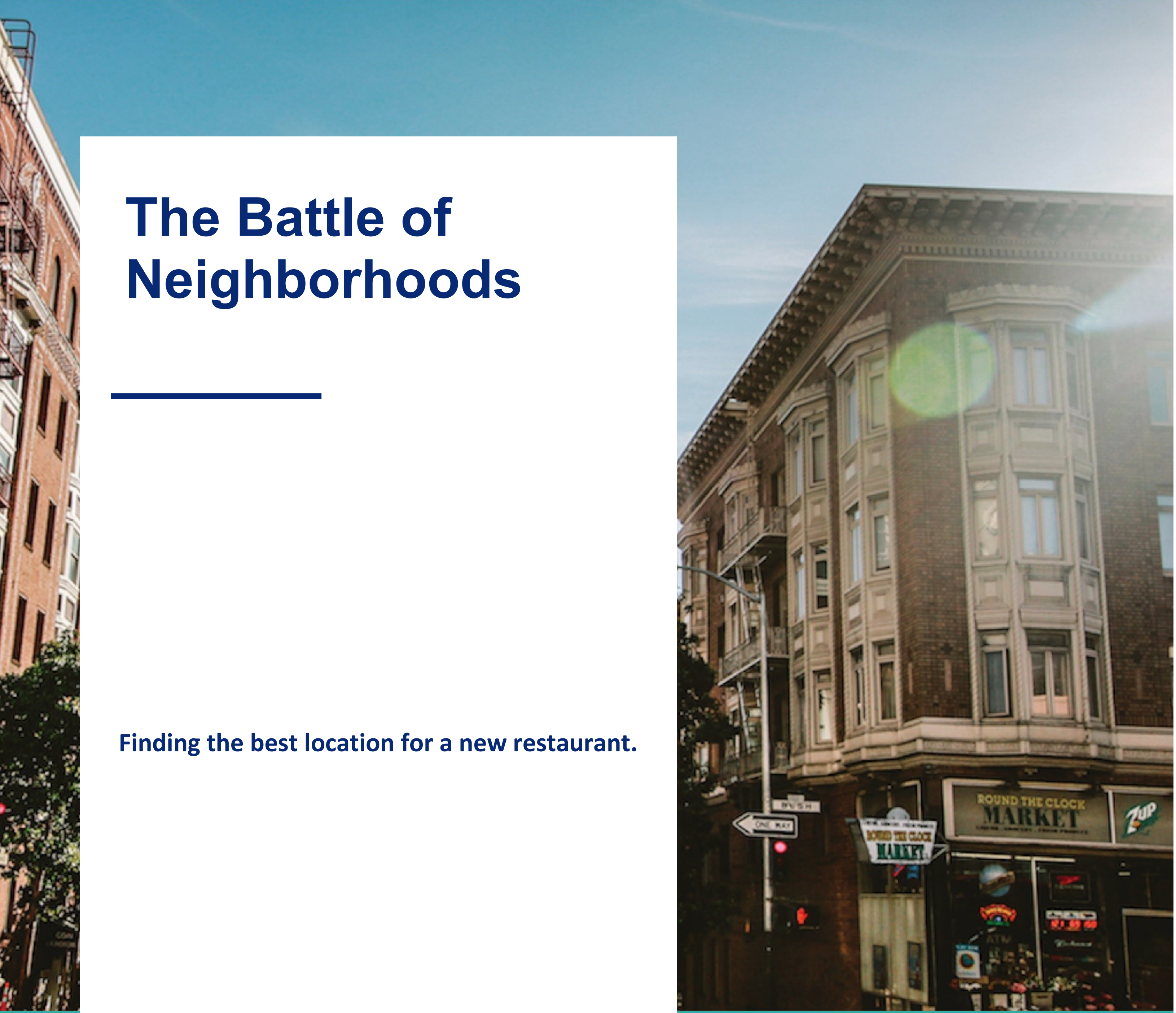


The Battle of Neighborhoods

Finding the best location for a new restaurant.

NOVEMBER 10

IBM Data Science Professional Certificate

Authored by: Jolene Zhao

Introduction

Business Problem

Being one of the largest city with an growing population, there is no doubt about the business opportunities in Toronto, Canada. Multiculturalism is seen through the various neighborhoods. Downtown Toronto being the hub of interactions between ethnicities brings many opportunities for entrepreneurs to start or grow their business. It is a place where people can try the best of each culture, either while they work or just passing through.

The objective of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the 'best' neighborhood in Toronto to open a restaurant. Through this project, we will find the most suitable location for an entrepreneur to open a new Chinese restaurant in Toronto.

"Good food, good mood"

Target Audience

This project aims entrepreneurs who want to open a new Chinese restaurant in Toronto. The analysis will provide vital information that can be used by the target audience.

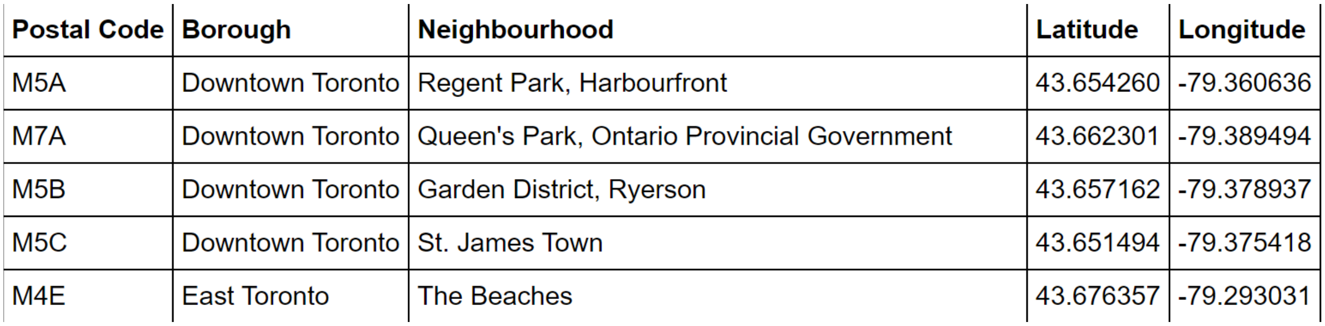
Data

The data that will be required will be a combination of CSV files that have been prepared for the purposes of the analysis from multiple sources which will provide the list of neighborhoods in Toronto (via Wikipedia), the Geographical location of the neighborhoods (via Geocoder package) and Venue data pertaining to Chinese restaurants (via Foursquare). The Venue data will help find which neighborhood is best suitable to open a Chinese restaurant.

- Data Source 1: Toronto Neighborhoods via Wikipedia
- Data Source 2: Geographical Location using Geocoder Package
- Data Source 3: Venue Data using Foursquare

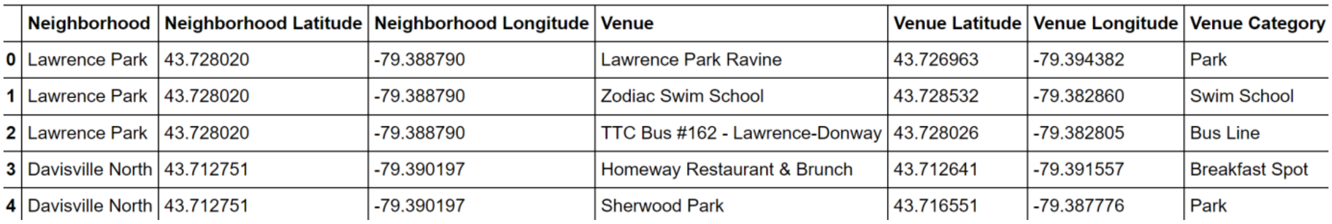
Methodology

After all the data was collected from data sources, data cleaning is conducted. Following week 3 assignment, we construct the following table for the geographical data.



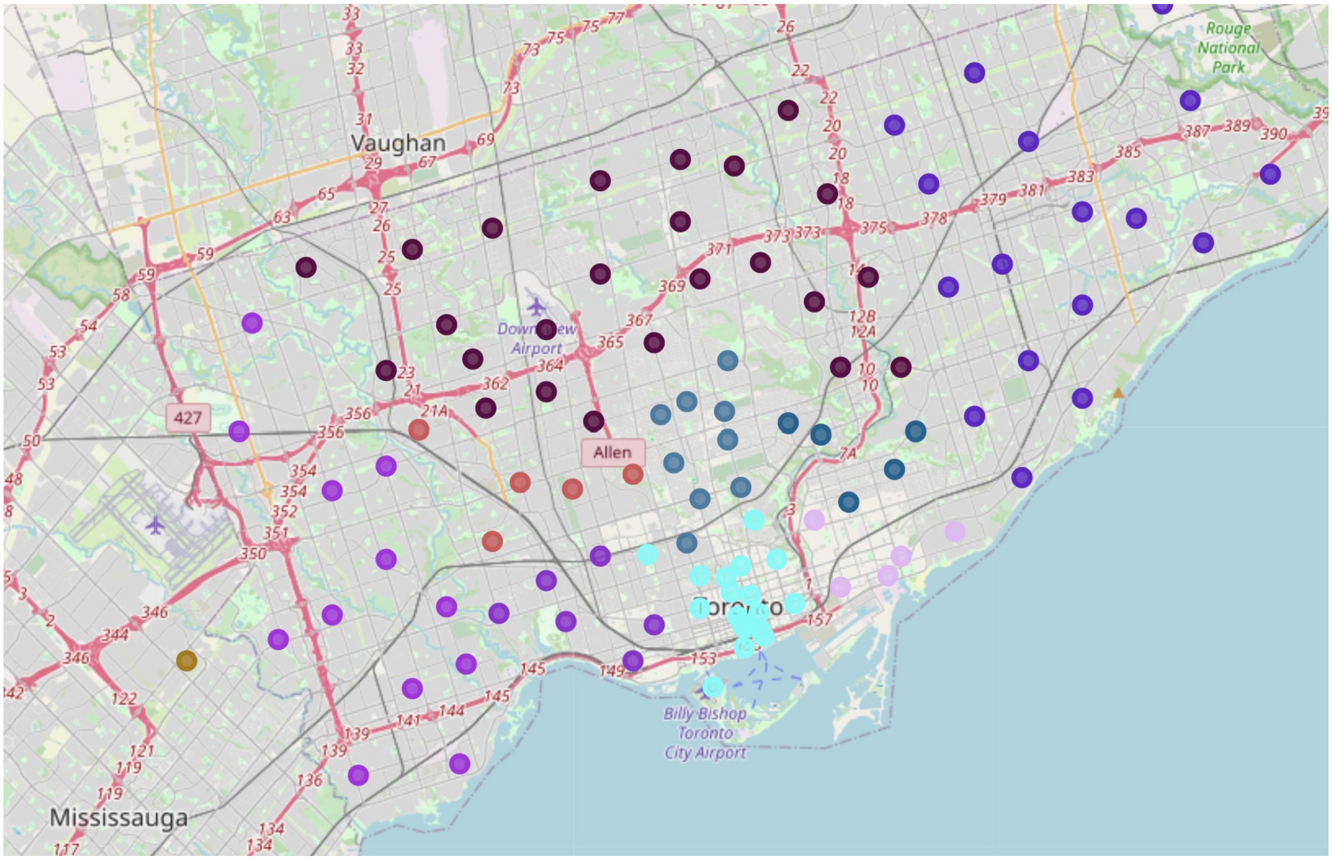
Postal Code	Borough	Neighbourhood	Latitude	Longitude
M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
M4E	East Toronto	The Beaches	43.676357	-79.293031

After, we pull the venue data from the Foursquare API and merge with the table above to provide us with the local venue within a 500-meter radius shown below.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Lawrence Park	43.728020	-79.388790	Lawrence Park Ravine	43.726963	-79.394382	Park
1	Lawrence Park	43.728020	-79.388790	Zodiac Swim School	43.728532	-79.382860	Swim School
2	Lawrence Park	43.728020	-79.388790	TTC Bus #162 - Lawrence-Donway	43.728026	-79.382805	Bus Line
3	Davisville North	43.712751	-79.390197	Homeway Restaurant & Brunch	43.712641	-79.391557	Breakfast Spot
4	Davisville North	43.712751	-79.390197	Sherwood Park	43.716551	-79.387776	Park

Now after cleansing the data, the next step is to analyze it. We create a map using Folium and color-coded each neighborhood depending on what Borough it was located in.

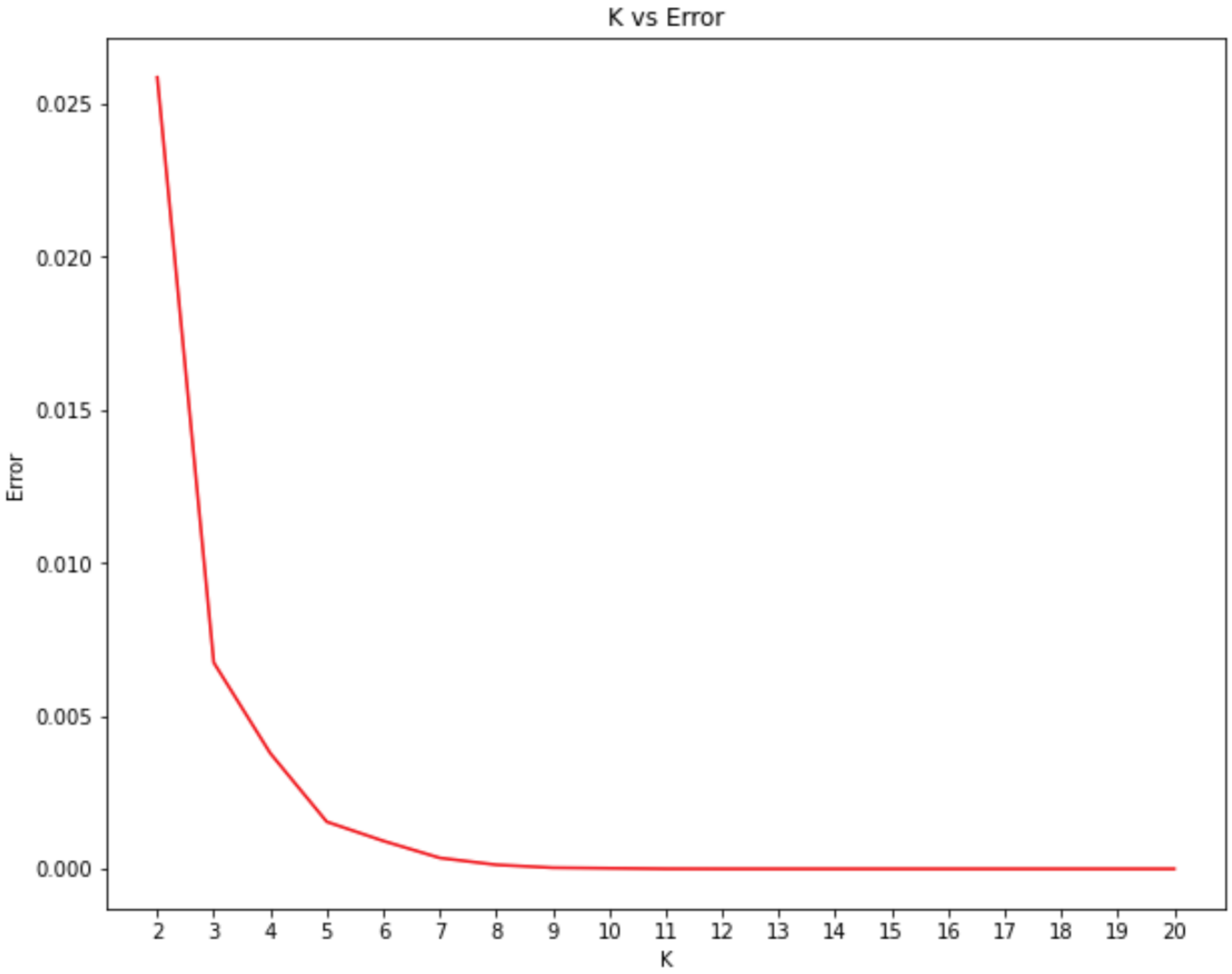


We then merge the neighborhood data with the Foursquare Venue data. The resulting dataset provides us the information on the nearest Venue for each of the Neighborhoods.

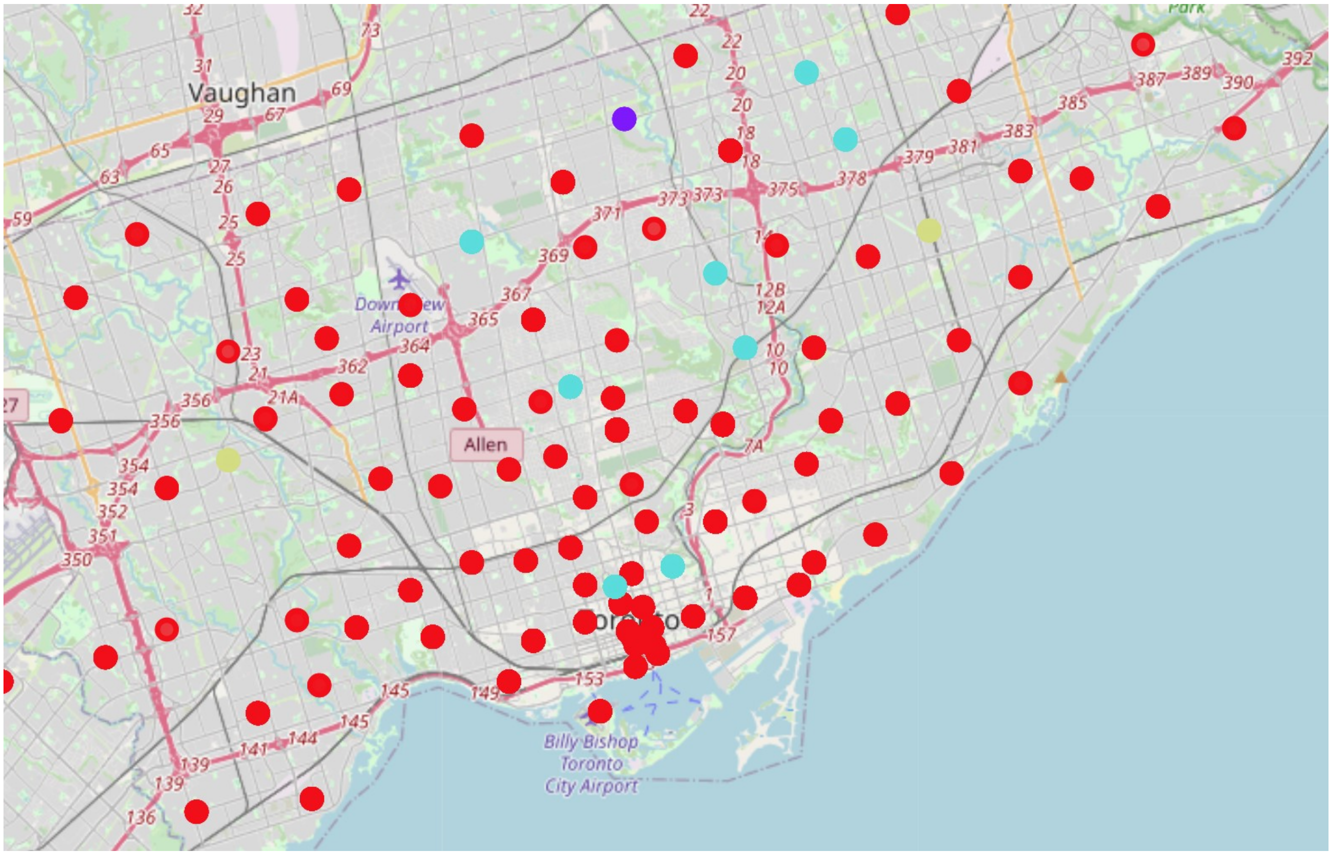
Then to analyze the data we performed a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This technique is called One hot encoding. For each of the neighborhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighborhood.

We then use K-Means Clustering. By using this approach, we wanted to cluster the neighborhoods based on the neighborhoods that had similar averages of Chinese Restaurants in that Neighborhood.

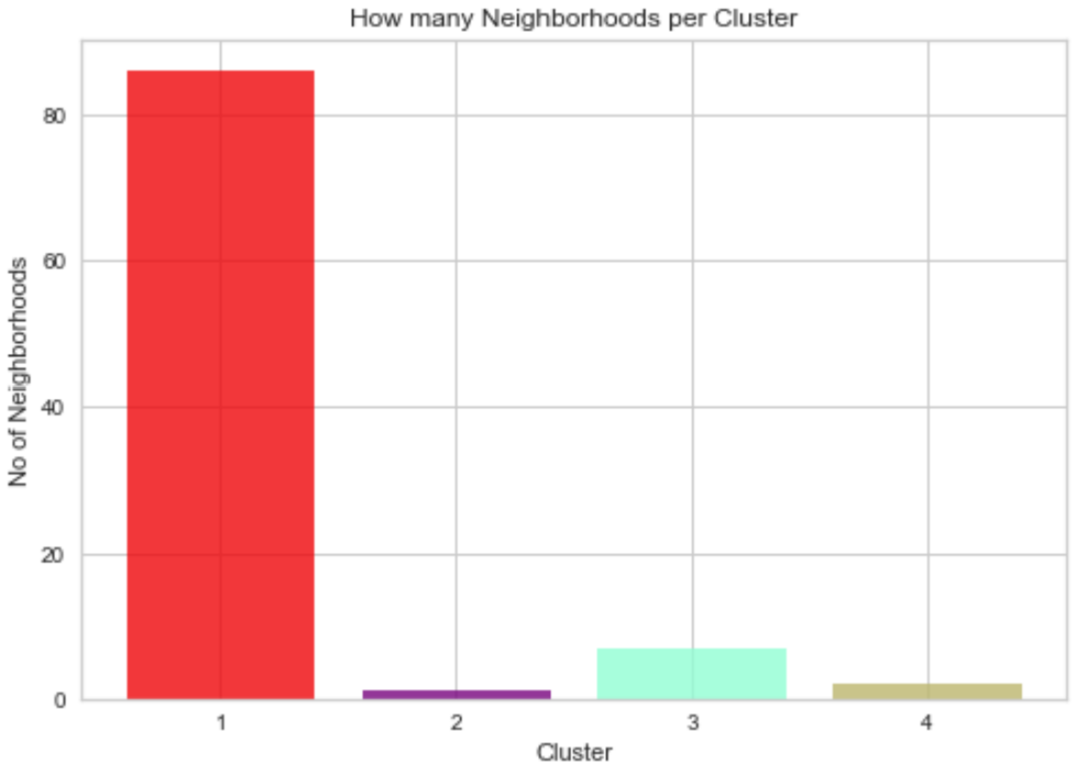
To obtain the optimum K value that is neither overfitting or underfitting, we employ the Elbow Point Techniques. In this technique, we run a test with different number of K values and measured the accuracy and then chose the best K value. The best K value is chosen at the point in which the line has the sharpest run. In our case, we had the Elbow Point at K=4. That means we will have a total of 4 clusters.

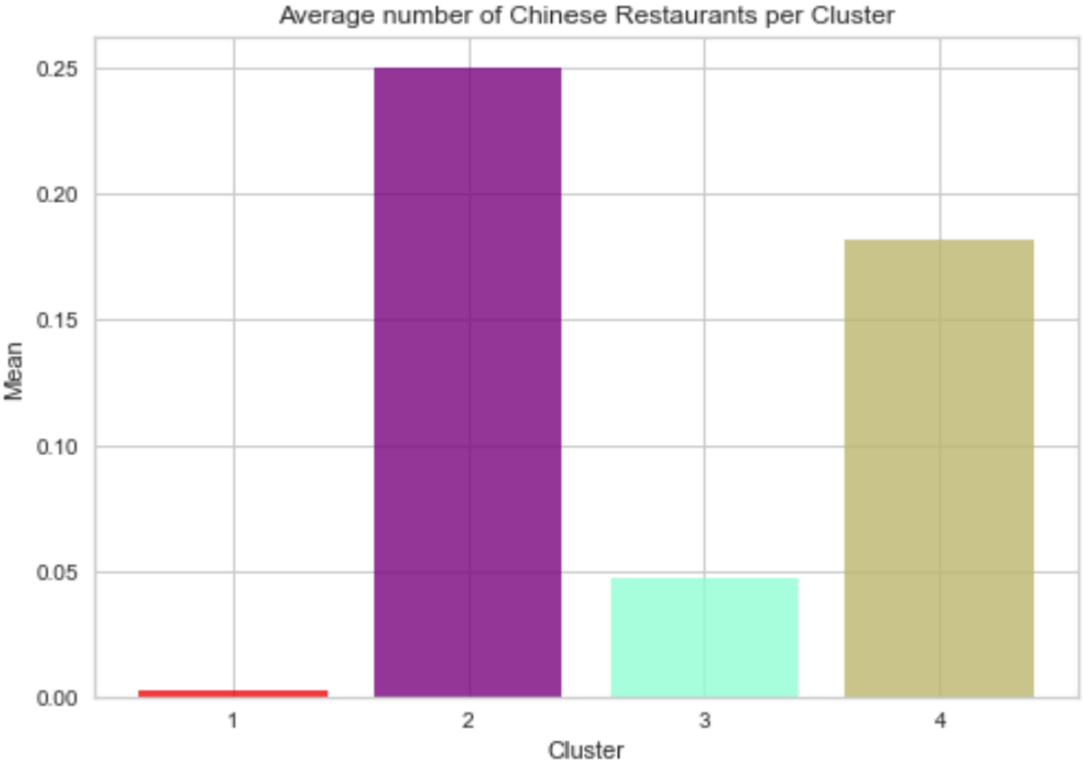


After determining the number of clusters, we develop k-mean clustering analysis. Then we create a map using the Folium package in Python and each neighborhood was colored based on the cluster label.



We then perform the clustering analysis. More specifically, we generate the following two visual analytics.





Among the four clusters, the first one (red) has the highest number of neighborhoods. However, this cluster also has the lowest number of

Chinese restaurants. On the contrary, we find that the second cluster (purple) has the lowest number of neighborhoods but has the highest number of Chinese restaurants.

Results and Discussions

Most of the Chinese Restaurants are in cluster 2 represented by the purple. The Neighborhoods located in the North York area that have the highest average of Chinese Restaurants are Bedford Park and Lawrence Manor East. Even though there is a huge number of Neighborhoods in cluster 1, there is little to no Chinese Restaurant. We see that in the Downtown Toronto area (cluster 3) has the second last average of Chinese Restaurants. Looking at the nearby venues, the optimum place to put a new Chinese Restaurant in Downtown Toronto as there are many Neighborhoods in the area but little to no Chinese Restaurants, therefore, eliminating any competition. The second-best Neighborhoods that have a great opportunity would be in areas such as Adelaide and King, Fairview, etc. which is in Cluster 2. Having 70 neighbourhoods in the area with no Chinese Restaurants gives a good opportunity for opening a new restaurant. Some of the drawbacks of this analysis are — the clustering is completely based on data obtained from the Foursquare API. Also, the analysis does not take into consideration of the Chinese population across neighbourhoods as this can play a huge factor while choosing which place to open a new Chinese restaurant. This concludes the optimal findings for this project and recommends the entrepreneur to open an authentic Chinese restaurant in these locations with little to no competition.

Conclusion

In conclusion, to end off this project, we had an opportunity on a business problem, and it was tackled in a way that it was similar to how a genuine data scientist would do. We utilized numerous Python libraries to fetch the information, control the content and break down and visualize those datasets. We have utilized Foursquare API to investigate the settings in neighborhoods of Toronto, get a great measure of data from Wikipedia which we scraped with the Beautifulsoup Web scraping Library. We also visualized utilizing different plots present in seaborn and Matplotlib libraries. Similarly, we applied AI strategy to anticipate the error given the information and utilized Folium to picture it on a map.

Places that have room for improvement or certain drawbacks give us that this project can be additionally improved with the assistance of more information and distinctive Machine Learning strategies. Additionally, we can utilize this venture to investigate any situation, for example, opening an alternate cuisine or opening of a Movie Theater and so forth. Ideally, this task acts as an initial direction to tackle more complex real-life problems using data science.