Battle of Neighbourhoods

By RAHUL RAJ

Introduction

- With a population just short of 3 million people, the city of Toronto is the largest in Canada, and one of the largest in North America (behind only Mexico City, New York and Los Angeles). Toronto is also one of the most multicultural cities in the world, making life in Toronto a wonderful multicultural experience for all. More than 140 languages and dialects are spoken in the city, and almost half the population Toronto were born outside Canada. It is a place where people can try the best of each culture, either while they work or just passing through. Toronto is well known for its great food.
- The objective of this project is to find the best neighborhood in Toronto to open a restaurant using Foursquare location data. In this project we'll go through the solution for this problem for avoiding or considering low risk criteria and high success rate

Data acquisition and cleaning

- For this project we need the following data:

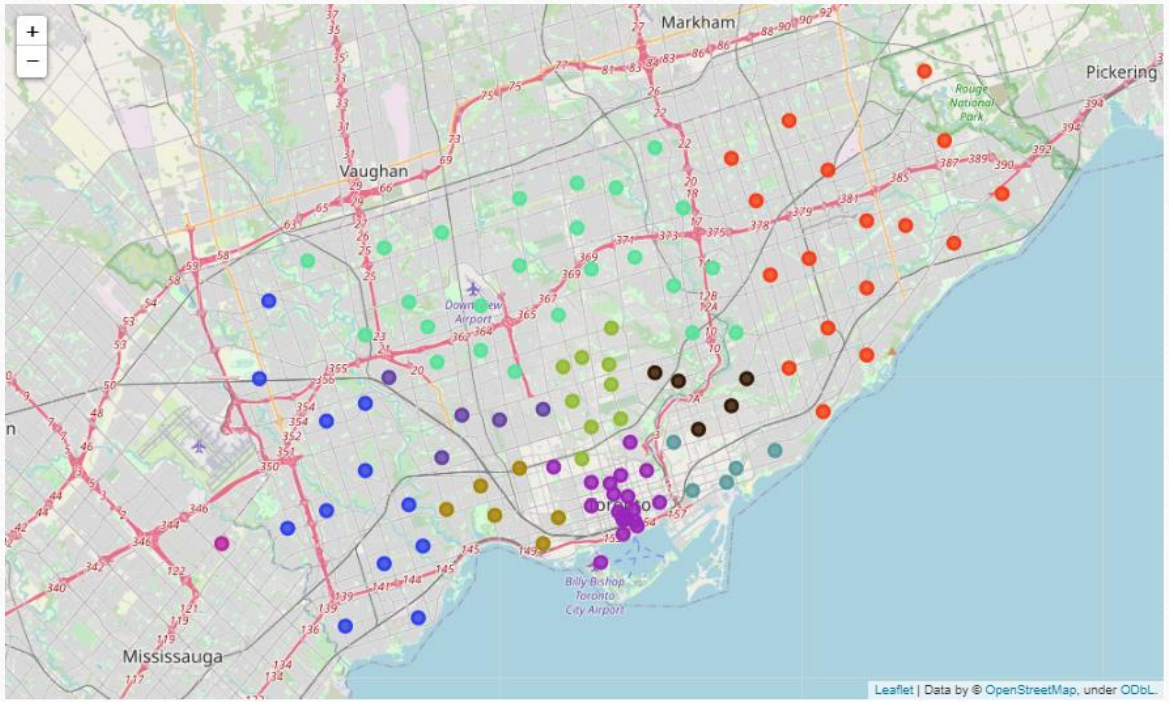
1. Toronto City data that contains Borough, Neighborhoods along with their latitudes and longitudes.
   - Source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
   - Using Beautiful Soup package, we scrap the data in tabular form.
2. Geographical Location data using Geocoder Package.
   - Source: https://cocl.us/Geospatial_data
   - The second source of data provided us with the Geographical coordinates of the neighbourhoods with the respective Postal Codes.
3. Venue Data using Foursquare API
   - Source: https://foursquare.com/developers/apps
   - From Foursquare API we can get the name, category, latitude, longitude for each venue.

Methodology

After scraping the data from Wikipedia there were Boroughs that were not assigned to any neighbourhood therefore, the following assumptions were made:

- Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned.
- More than one neighbourhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighbourhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighbourhoods separated with a comma as shown in row 11 in the above table.
- If a cell has a borough but a Not assigned neighbourhood, then the neighbourhood will be the same as the borough.

Methodology(Cont'd)



Data Pre-processing

- To analyse the respective Italian restaurant, present in that neighbourhood or not, we'll use One hot encoding technique. For each of the neighbourhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighbourhood.
- Then we grouped those rows by Neighbourhood and by taking the average of the frequency of occurrence of each Venue Category.
- After, we created a new data frame that only stored the Neighbourhood names as well as the mean frequency of Italian Restaurants in that Neighbourhood. This allowed the data to be summarized based on each individual Neighbourhood and made the data much simpler to analyse.

K-Means Clustering

- Now we'll cluster these neighbourhoods based on the frequency of Italian restaurants present. To do this we apply k-means clustering algorithm. To avoid the overfitting and underfitting of the model we need a optimum value of "k". There are many techniques like Elbow method, Silhouette score method to get the best "k" value. Here we're going to use Elbow method to get best "k" value. We'll import 'K Elbow Visualizer' from the yellow brick package. Then we fit our K-Means model above to the Elbow visualizer.

```
# Instantiate the clustering model and visualizer
model = KMeans()
visualizer = KElbowVisualizer(model, k=(2,21))

visualizer.fit(X)        # Fit the data to the visualizer
visualizer.show()
```

K-Means Clustering (Cont'd)

- This bunch of code will give this below graph

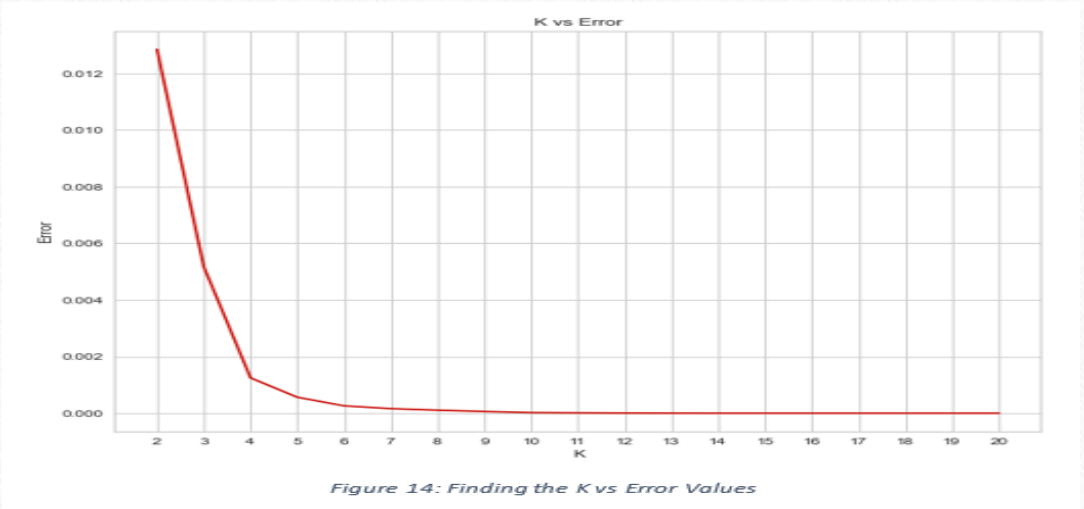
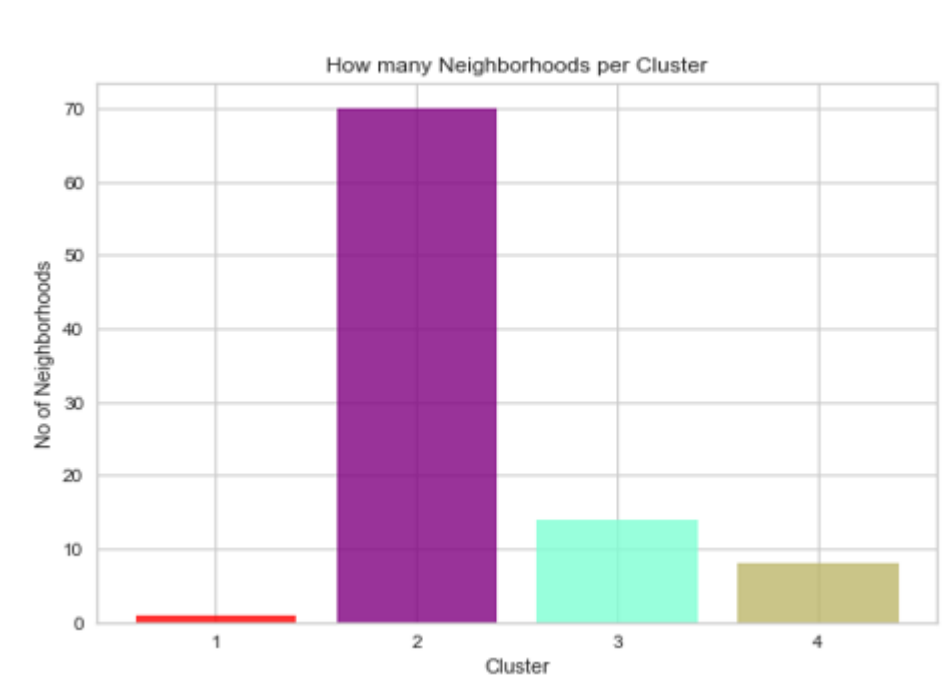
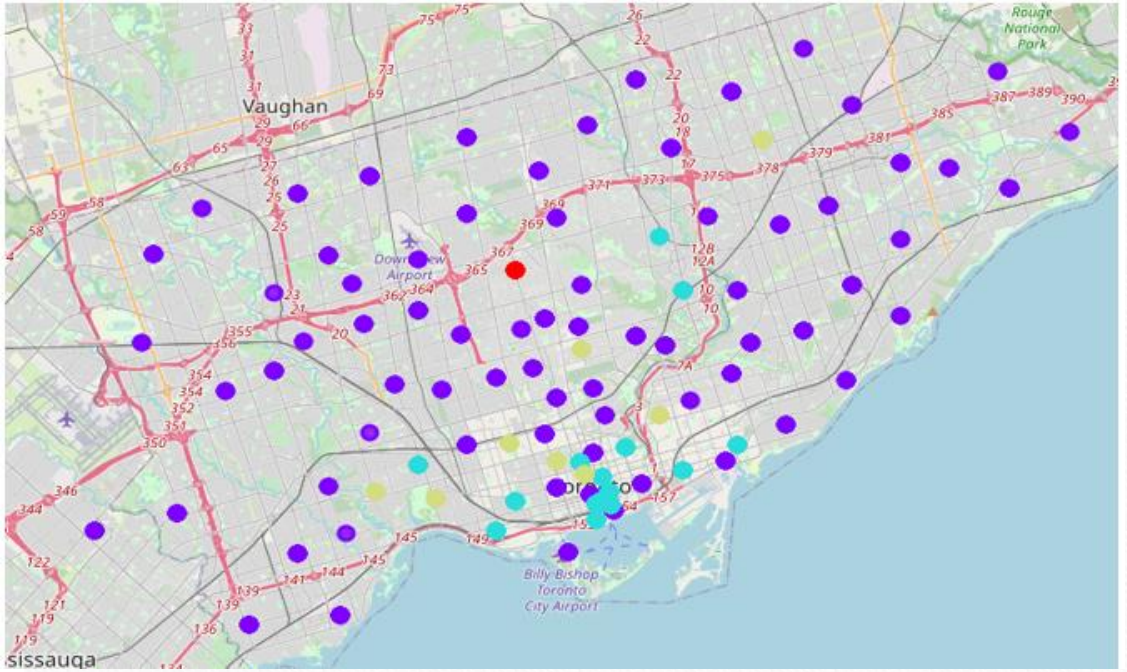


Figure 14: Finding the K vs Error Values

Result & Outcomes





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.