Applied Data Science Capstone Report

Introduction

- Toronto is the provincial capital of Ontario and the most populous city in Canada. Toronto is an international center of business, finance, arts, and culture, and is recognized as one of the most multicultural and cosmopolitan cities in the world. Its economy is highly diversified with strengths in technology, design, financial services, life sciences, education, arts, fashion, business services, environmental innovation, food services, and tourism.
- The strength and vitality of the many neighbourhoods that make up Toronto, Ontario, Canada has earned the city its unofficial nickname of "the city of neighbourhoods. We've heard Toronto described as "New York City run by the Swiss". Both cities are very diverse and are the financial capitals of their respective countries.
- The project will try to answer the question “where should someone open a restaurant in the city?”.

Data Sources

- The data sources to be used in this study includes the followings:

1. The city of Toronto data.

- Data source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2. Geospace data for each neighbourhood in Toronto City.

- Data source: https://cocl.us/Geospatial_data csv file

3. Venue data.

- Data source: https://developer.foursquare.com/docs

Methodology

1. A list of neighbourhoods in the city of Toronto from the Wikipedia page, will be imported to the Notebook to generate a dataframe with 3 columns i.e. Postal code, Borough and Neighbourhood.
2. The coordination data set in term of latitude and longitude will be imported from Geospatial_data.csv file. Two more columns will be added to the existing dataframe based on the postal code i.e. Latitude and Longitude. Now we will be able to locate the Neighbourhood with the corresponding Latitude and Longitude data.
3. The map of neighbourhoods in the city of Toronto will be generated using Folium package.
4. Foursquare API will be applied to explore the neighbourhoods. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. From this, we can start analyze each neighbourhood.

Results

- By using the k-means clustering, the clustering shows that the neighbourhoods are classified into 5 clusters based on the frequency of occurrence for restaurants:
- Cluster 1: Neighbourhoods with no restaurant.
- Cluster 5: Neighbourhoods with low number of restaurants.
- Cluster 2, 4 : Neighbourhoods with moderate number of restaurants.
- Cluster 3: Neighbourhoods with high number of restaurants.

Discussion

- According to the results, the neighbourhoods in Cluster 1 have no restaurant. This may be inferred that there is no demand of the restaurants in these areas for some reason. The cluster 1 is not an interesting area for the investment.
- In Cluster 5, the neighbourhoods with low number of restaurants are present. These areas are interesting for the investment to open a new restaurant due to low competitiveness.
- Cluster 2 and Cluster 4 have moderate density of restaurants. Cluster 3 shows the highest competitiveness of restaurants in the area due to the high frequency of restaurants.

Conclusion

- The answer to the business question “In the city of Toronto, Canada, if a retailer is looking to open a new restaurant, where would you recommend that they open it? “, the answer proposed by this project is:
- The neighbourhoods in cluster 5 are the most preferred locations to open a new restaurant.