**Applied Data science capstone project**

1. **Introduction**

Toronto is no doubt a multicultural and diverse metropolis, attracting people from all North America and the world. Data from financial post claim that the average Canadian moves 4 to 6 times in their life time while statistics from Statistical Canada also proves that Toronto has one of the largest population of immigrants in North America. Both these evidence show that there are a huge number of people, either from within or outside of Toronto, are looking into settling down in this fascinating city every day. Since Coffee is unquestionably one of the most favourite beverages and definitely an essential to many people, the quantity and quality of coffee shops nearby are thus a significant factor in deciding where to stay. Hence this project aims to analyse the number of coffee shops and their performances within each neighbourhood. This data would help those who are planning to settle down in Toronto to make better informed choices when choosing their new homes. People on business trips etc to Toronto could also benefit from this information when choosing their short stay locations.

1. **Data**
2. Wikipedia page to obtain Toronto’s neighbourhood boroughs information. This data source contains a list of all boroughs and their corresponding postal codes and neighbourhood within.

Source: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

1. URL file and Geocoder package for geographical information. This url file contains latitude and longitudes of very postal codes of Toronto. This data will be combined with data 1 to obtain the location data of Toronto neighbourhoods

Source: <http://cocl.us/Geospatial_data>

1. Geojson file. A Geojson file with geographical data of Toronto boroughs’ boundary. This is used to plot the choropleth map.

Source: https://raw.githubusercontent.com/ag2816/Visualizations/master/data/Toronto2.geojson

1. Foursquare to obtain information on Chinese restaurants within each venue. This includes quantities and ratings of available restaurants.
2. **Methodology**
   1. **Data retrieving and cleaning**

Information of the boroughs and neighbourhoods in Toronto is imported from Wikipedia page and all rows with missing information are dropped. This data frame is then merged with the data frame that contained geographical information, i.e. the latitude and longitudes.

A screenshot of a cell phone

Description automatically generated

With the above data, a Toronto map indicating all the neighbourhoods can be plotted for better visualisation. This map is plotted using folium.

A close up of a map

Description automatically generated

A function is defined to interact with Foursquare API to obtain the top 100 venues within 1000 radii from the given geographical location of each neighbourhood. The retrieved data frame is filtered using masking techniques. Since there might be various category names for various shops, filtering is done based on strings that contain either coffee or café.

Leveraging the above information, performance data such as ratings and likes are acquired through interaction with Foursquare API using premium calls. This information is stored within a data frame and is merged with the location data. Unnecessary columns are removed, duplicates rows are dropped, and index is reset.

A screenshot of a cell phone

Description automatically generated

* 1. **Analysis**

First, the quantity of coffee shops within each neighbourhood is calculated using the group\_by.count attribute and sorted in descending order. Using this information and the Geojson file downloaded with wget, a choropleth map is plotted using folium.

A screenshot of a cell phone

Description automatically generated

Second, the performance/quality can be analysed using the average ratings and likes of coffee shops within each neighbourhood. These is done using group\_by.mean attribute. Since both of these parameters indicate how good the coffee shop is, they can be combined to give an indication of the overall performance. They are normalized before combining to prevent biasness. The top 5 neighbourhoods are then determined using nlargest attribute and masking technique. Due to the limited amount of premium calls available (500 calls per day), this analysis is only done to category under cafés.

A screenshot of a cell phone

Description automatically generated

Finally, the neighbourhoods are clustered using k-means clustering to divide them into tier. Features for clustering are the performance and the count of coffee shops.

1. **Results & Discussion**
   1. **Quantity of coffee shops**

Average number of coffee shops within each neighbourhood is 6.5. The choropleth map below shows the number of coffee shops in each neighbourhood. It is obvious that the central area has significantly higher amount of coffee shops. Specifically, Regent Park, Harbourfront, Berzy Park and Station A claims the top spot with 19 coffee shops each. While St.James Town and Davisville North came next with 17 coffee shops each. These neighbourhoods are especially suitable for coffee lovers who loves to explore new coffees since there are more options to discover.

A close up of a map

Description automatically generated

* 1. **Quality of coffee shops**

Quality/performance of coffee shops is done via analysis of “Likings” and “Ratings” as discussed in section 3.2 Analysis. The bar chart below shows the overall performance of each neighbourhoods in terms of coffee shops. The top 5 neighbourhoods are therefore “Garden District, Ryerson”, “Kensington Market, Chinatown, Grange Park”, “Little Portugal, Trinity”, “Richmond, Adeaide, King” and “University of Toronto, Harbord” which are colour codes as red. High number of likes and high ratings indicates the superior quality and popularity of the coffee shops. Thus, distance to these neighbourhoods can be considered for coffee addicts looking to get a new home in Toronto.

A screenshot of a cell phone

Description automatically generated

* 1. **Neighbourhood Clustering**

The neighbourhoods are clustered into tiers using the quality and quantity scores previously determined. They are grouped into 5 tiers, with tier 1 being most suitable neighbourhoods for coffee addicts. This scatter plot below depicts the clustering result.A close up of a piece of paper

Description automatically generated

Table below summarised the neighbourhoods within each tier for easy reference.

|  |  |
| --- | --- |
| Tier | Neighbourhoods |
| 1 (Best) | A screenshot of a cell phone  Description automatically generated |
| 2 | A screenshot of a cell phone  Description automatically generated A screenshot of a cell phone  Description automatically generated |
| 3 | A screenshot of a cell phone  Description automatically generated A screenshot of a cell phone  Description automatically generated |
| 4 | A screenshot of a cell phone  Description automatically generated |
| 5 (Worst) | A screenshot of a cell phone  Description automatically generated |

1. **Conclusion**

This project is based only on information extracted from Foursquare. Thus it is limited by the source of information and may not be comprehensive. However the results are still useful to provide some insights to people planning to get a new home in Toronto. From the analysis, the best coffee neighbourhoods are “Garden District, Ryerson”, University of Toronto, Harbord” and “Kensington Market, Chinatown, Grange Park”. When deciding on where to locate their new homes, coffee addicts could definitely consider the aforementioned areas, or neighbourhoods close to those areas.