APPLIED DATA SCIENCE
CAPSTONE BY IBM/COURSERA

CAPSTONE PROJECT - THE BATTLE OF NEIGHBORHOODS - TORONTO

TABLE OF CONTENTS

INTRODUCTION/BUSINESS PROBLEM

Coffee is a rich source of antioxidants that slow down the aging process of tissues and effectively protect the body against health loss. One cup of coffee has been shown to contain even more antioxidants than a glass of grapefruit, blueberry, raspberry or orange juice. Investor already has many cafes of his own brand in the world. Now he intends to conquer a new market. Taking into account the above, the investor intends to open a new cafe in Toronto. Unfortunately, he doesn't know the city well and doesn't know where to open a café. He wants to know if there are coffee shops in all the neighborhoods. **That's why he wants to open a business where there are already cafés but they are not very popular.** As a result, the business problem is:

Where to open a new a Successful Cafe in Toronto?

I would like to do some research and recommend them the best place based on the number of cafes in different districts of Toronto. In order to solve this business problem, we intend to merge Toronto districts into a cluster in order to recommend locations.

DATA

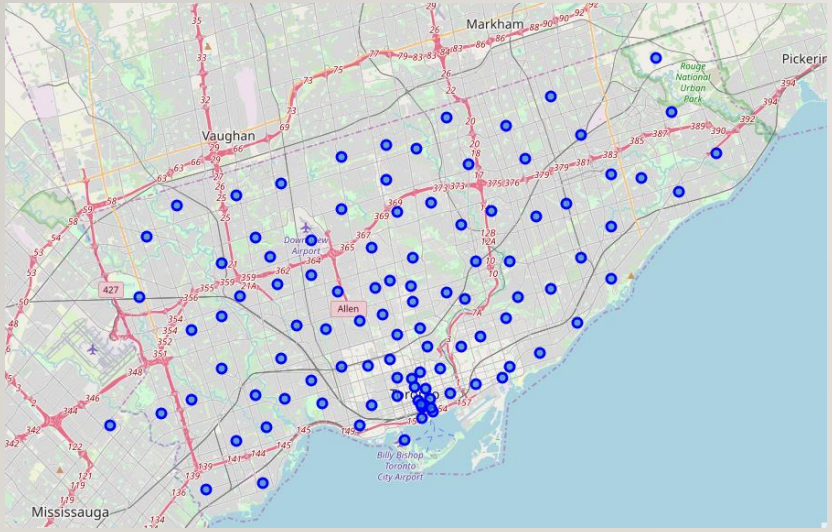
- To consider the objective stated above, we can list the below data sources used for the analysis.
- **Districts of Toronto** Wikipedia page was scraped to pull out the necessary information;
- **Coordinate data** for each Districts of Toronto obtained through Nominatim search engine for OpenStreetMap data;
- In order to investigate and target recommended locations in different locations depending on the presence of facilities and necessary objects, we will access the data through the **FourSquare API** and arrange it as a data frame for visualization. By combining data about districts in Toronto and data about amenities and essential facilities surrounding such properties from the FourSquare API, we will be able to recommend an appropriate location.

METHODOLOGY

The methodology in this project consists of two parts:

- **Data Understanding & Data Preparation**: Process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.Once the data has been collected, it must be transformed into a useable subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases.
- **Data Preparation & Data Exploration**: Process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.Once the data has been collected, it must be transformed into a useable subset unless it is determined that more data is needed. Once a dataset is chosen, it must then be checked for questionable, missing, or ambiguous cases. It is necessary to visualise the districts of Toronto.
- **Modelling**: To help people find similar neighborhoods in the safest borough we will be clustering similar neighborhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 5 for this project that will cluster neighborhoods into 5 clusters. The reason to conduct a K- means clustering is to cluster neighborhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighborhood.

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Heights, Lawrence Manor	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park	43.662301	-79.389494

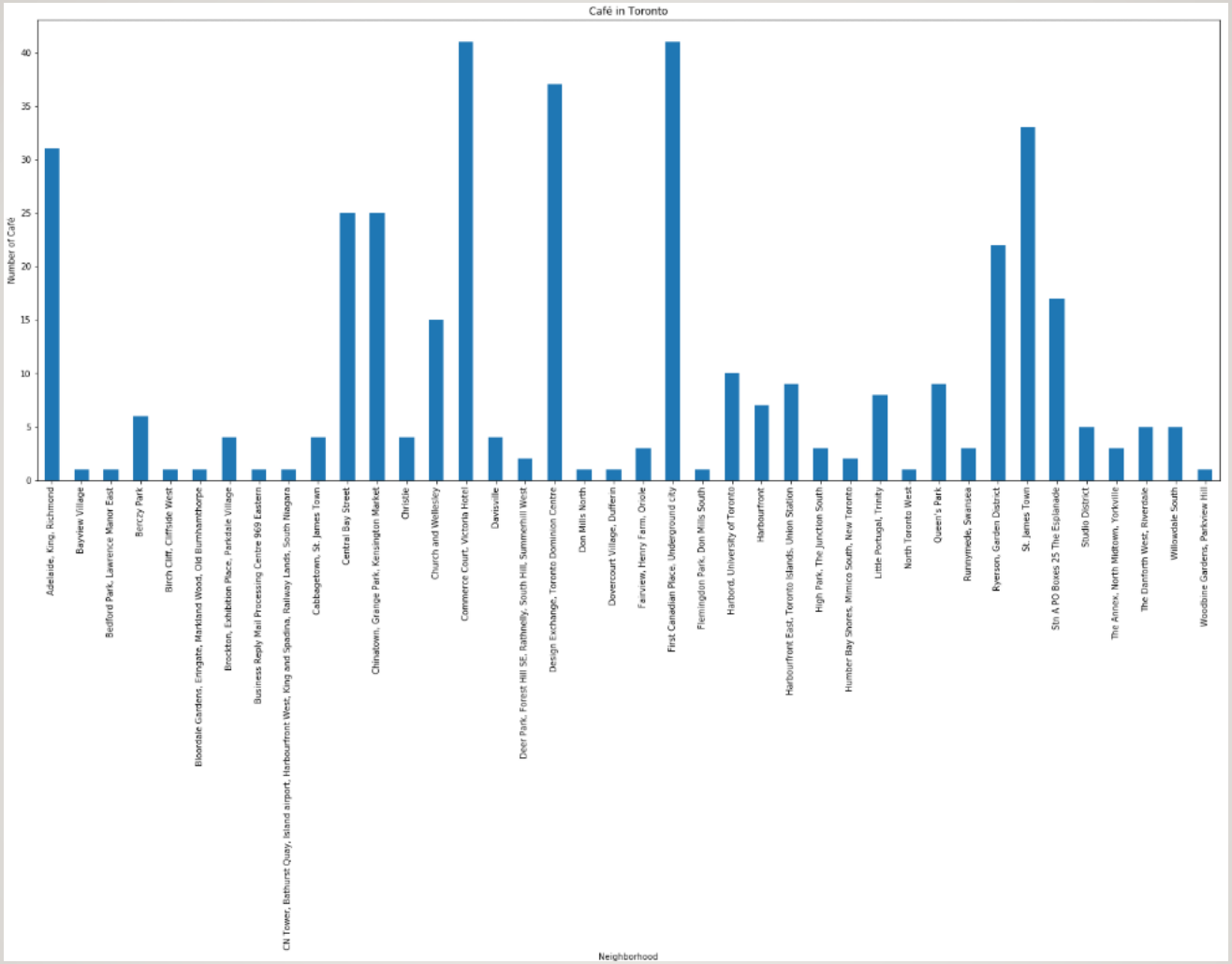


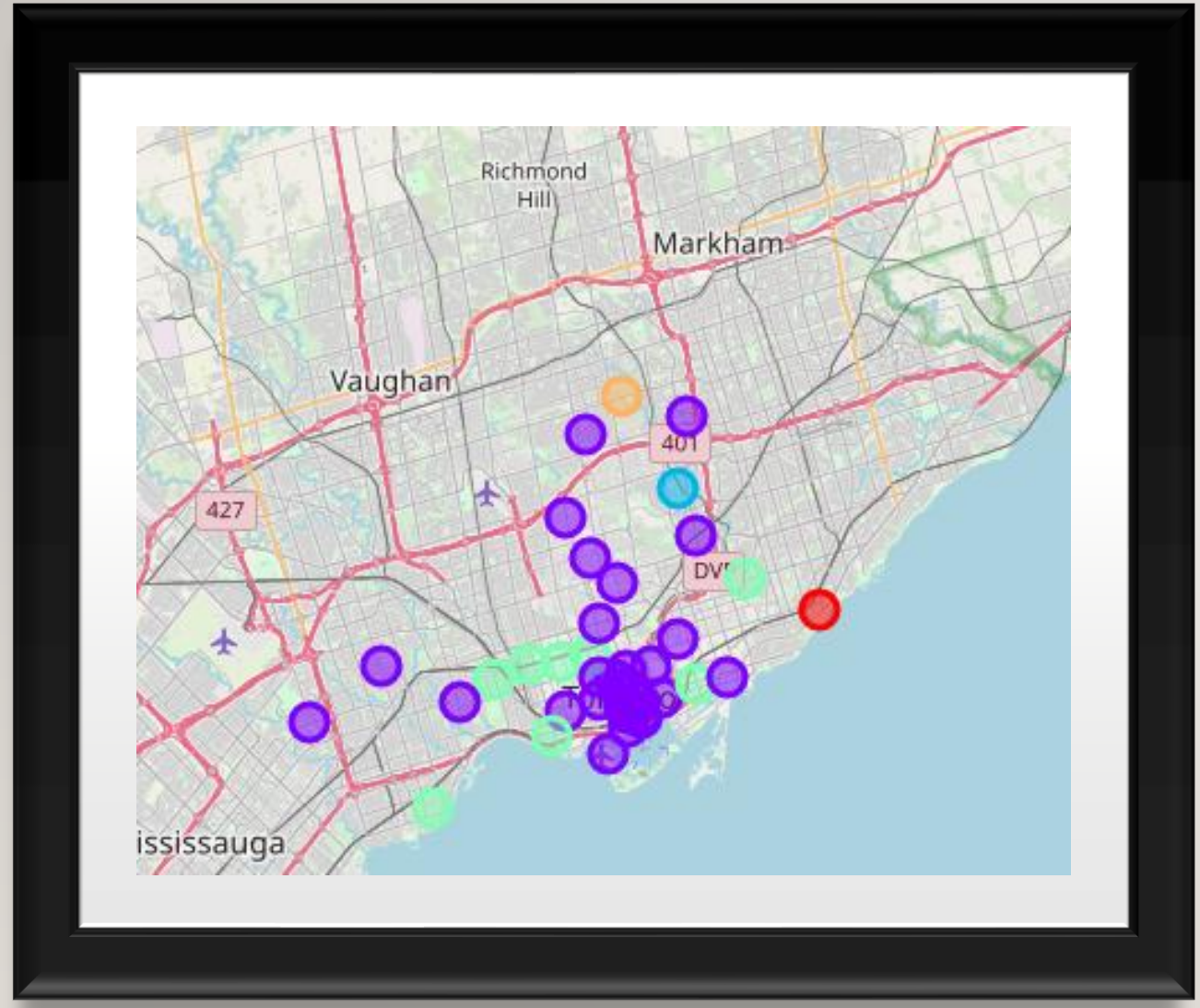
DATA UNDERSTANDING & DATA PREPARATION

- Scrape the Wikipedia page and gathering data into a Pandas dataframe
- Data cleaning
- Use geopy library to get the latitude and longitude values of Torronto
- Generating a map of Toronto and plotting the Neighborhood data on it

MODELLING

- Finding all the cafes within a 500 meter radius of each neighborhood.
- Perform one hot ecoding on the venues data.
- Grouping the venues by the neighborhood and calculating their mean.
- Performing a K-means clustering (Defining K = 5)


Café in Toronto
Number of Café
0
5
10
15
20
25
30
35
40
Adelaide, King, Richmond
Bayview Village
Bedford Park, Lawrence Manor East
Berczy Park
Birch Cliff, Cliffside West
Bloordale Gardens, Eringate, Markland Wood, Old Burnhamthorpe
Brockton, Exhibition Place, Parkdale Village
Business Reply Mail Processing Centre 969 Eastern
CN Tower, Bathurst Quay, Island airport, Harbourfront West, King and Spadina, Railway Lands, South Niagara
Cabbagetown, St. James Town
Central Bay Street
Chinatown, Grange Park, Kensington Market
Christie
Church and Wellesley
Commerce Court, Victoria Hotel
Davisville
Deer Park, Forest Hill SE, Rathnelly, South Hill, Summerhill West
Design Exchange, Toronto Dominion Centre
Don Mills North
Dovercourt Village, Dufferin
Fairview, Henry Farm, Oriole
First Canadian Place, Underground city
Flemingdon Park, Don Mills South
Harbord, University of Toronto
Harbourfront
Harbourfront East, Toronto Islands, Union Station
High Park, The Junction South
Humber Bay Shores, Mimico South, New Toronto
Little Portugal, Trinity
North Toronto West
Queen's Park
Runnymede, Swansea
Ryerson, Garden District
St. James Town
Stn A PO Boxes 25 The Esplanade
Studio District
The Annex, North Midtown, Yorkville
The Danforth West, Riverdale
Willowdale South
Woodbine Gardens, Parkview Hill
Neighborhood



RESULTS

- Cluster 0 shape = (1, 12)
- Cluster 1 shape = (30, 12)
- Cluster 2 shape = (1, 12)
- Cluster 3 shape = (8, 12)
- Cluster 4 shape = (1, 12)

RESULTS AND DISCUSSION

- The purpose of my experiment was to point out the right neighborhood to open a café in Toronto. Based on the experiments, districts where cafés already exist were selected at the beginning. There are 40 of them, and then, to broaden the scope of clustering, all the popular places in the districts where the café already exists were found. It was indicated that "The biggest proportion of Café among other venues in a district in Toronto is 7 % in Woodbine Gardens, Parkview Hill." In some cases, cafés have not been identified as a popular place at all.
- However, taking into account the grouping performed, I can recommend cluster 0 and 4 because there the cafes occupy the farthest place according to popularity. We must remember that in this experiment the distance from the centre was not taken into account. The investor can of course make a different choice, because the data is already prepared.

CONCLUSION

Different applications of this analysis are available based on a different methodology and possibly different data sources. The stakeholder problem has been resolved. This project helps the investor to better understand the area in relation to the most common places in the area. It is always helpful to use technology to be one step ahead, i.e. learn more about places before opening a new coffee shop in the district. The future of this project involves considering other factors, such as the cost of living in the areas concerned, in order to draw up a short list of neighbourhoods based on a predefined budget.