**Problem Description:**

This work explains how the neighborhoods of Toronto have potential for Food Truck business development in several neighborhoods.

**Requirements:**

Toronto being the financial capital of Canada is one of the densely populated cities in the world. This city sees a lot of diversity resulting from the movement of a lot of immigrants from several parts of the world for work and settlement. The current food truck business is booming in many neighborhoods. So for stakeholders who are interested in low investment business will get benefits from this analysis. The Foursquare API is used for the project which will get which neighborhoods have a potential for this business and population density to have a successful earnings.

The purpose of this whole exercise is for submission of the capstone project for the **"IBM Data Science"** course on Coursera as well as to showcase my data science skills in the real-world application.

**Data:**

The data used for this project are manipulated in a way which gives a list of potential neighborhoods that have potential for food truck business. The Toronto neighborhood is made using the Wikipedia webpage. Foursquare database is used to sort neighborhoods based on their current food truck business traffic. The Wellbeing Toronto webpage is used to get and compare the resulting neighborhood from K-Mean analysis to further sort them population density-wise. This will give the most favorable neighborhood for the business and stakeholders can make an informed decision about investment.

1. List of postal codes of Canada Wiki: [**https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M**](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) for access to neighborhood data of Toronto region.
2. Geographical coordinates of the neighborhoods: [**http://cocl.us/Geospatial\_data**](http://cocl.us/Geospatial_data) for getting the longitude and latitude data for the neighborhoods.
3. The population density of Toronto city is graphed and displayed in the map [http://map.toronto.ca/wellbeing/#eyJ0b3Itd2lkZ2V0LWNsYXNzYnJlYWsiOsSAcGVyY2VudE9wYWNpdHnElzcwfSwiY3VzxIJtYcSTYcSXxIBuZWlnaGJvdXJob29kc8S2fcSrxIHEg8SFxIfEicSLdGFixYXEmCLEo3RpdmVUxZBJZMSXxYnEhMWPYi1pbmRpY2HEgnLFhcWIYWdzTWFwxLYiesWCbcSXMTHErHjEly04ODM2NDQ1LjAzNzk4OTTErMSnOjU0MjE0OTIuMcaNxow1NcWIxaTFpsWoxarFksSAxZjFq2lvbsSXMsSsc8WkZ2xlxbTErnLEk8SfVGltZcWcxKjErMWWxrrGpCI3xbRzZcaxY3RlZEnFpcWnxanEg8a8OseCMyLErMagx47FqsWcc0HFpVfEucS7dMWSW8SAxIfFnjfHlMSsd8efaHTFucSsxJPGpseFUG%2FEjnLEpcaQZmFsx4V9XcWHxYjGv2XHhce2yILFhsSsxrTGtnTHjMahx49yTcWDxrLHksaubsawxrLGn8eNxqLEg1PHhmXHiMSDxbTIgVR5xJrFnsiTyJXHlcSBxrnGu8WdOjfFhw%3D%3D](http://map.toronto.ca/wellbeing/#eyJ0b3Itd2lkZ2V0LWNsYXNzYnJlYWsiOsSAcGVyY2VudE9wYWNpdHnElzcwfSwiY3VzxIJtYcSTYcSXxIBuZWlnaGJvdXJob29kc8S2fcSrxIHEg8SFxIfEicSLdGFixYXEmCLEo3RpdmVUxZBJZMSXxYnEhMWPYi1pbmRpY2HEgnLFhcWIYWdzTWFwxLYiesWCbcSXMTHErHjEly04ODM2NDQ1LjAzNzk4OTTErMSnOjU0MjE0OTIuMcaNxow)
4. Foursquare database: [**https://Foursquare.com**](https://foursquare.com/) to be used in order to explore the desired neighborhood data for various restaurant details and access the JSON files. This data shall be utilized to map the Food Trucks in various locations.

**Problem solving strategy**

The idea is to analysis Toronto region for its potential in Food Truck business development in terms of the population spread in each neighborhood. I will specifically compare the number of food trucks and population in the city as well as list down the 4 most common venues in city’s neighborhood wise. The outcome of this study will help tourists and new immigrants have an overview of the common venues in the city and which might further help them in their decision of travel or immigration choice.

**Step-wise approach of problem-solving:**

**Step-1:** Web scraping of the neighborhood data from postal codes of Canada Wiki-link. Clean the data by removing the missing values and store the data in a python Dataframe consisting of three columns namely: PostalCode, Borough, and Neighborhood

**Step-2:** Take the help of long-lat data from the geospatial data wikilink and append the geographical coordinates in the above dataframes to get new respective dataframes for further analysis.

**Step-3:** Getting location data using the Foursquare API. It will be used to retrieve information of the common venues in Toronto neighborhoods. The API will return a JSON file which will be further converted into a Python Dataframe.

**Step-4:** List down the 4 most common venues for city.

**Step-6:** Lastly we will discuss the results based on the above findings and provide a snapshot of neighborhood which will help stakeholders with their choice.

**Data Wrangling**

A lot of hard work went into creating the working data set.  
I had to combine the following disparate data sources. The order of events went like this

**1. Load all the Data from all the various sources.**

**1.1 Toronto neighborhoods broken down by postal code**

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>  
Here I used BeautifulSoup to scrape the wiki page to extract a working list of Toronto Neighborhoods sorted by postal code.

**1.1.1 Load Toronto geospatial coordinates and merge to Toronto Postal Code Data**

<http://cocl.us/Geospatial_data>  
Next, I joined geo spatial to the Toronto Data.

**1.2 Toronto neighborhoods populations broken down by postal code**

<https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Tables/File.cfm?T=1201&SR=1&RPP=9999&PR=0&CMA=0&CSD=0&S=22&O=A&Lang=Eng&OFT=CSV>  
Use Pandas to grab the csv

**1.2.1 Merge Toronto Neighbourhood populations data with Toronto Postal Code data**

Next, I joined population data to the Toronto Data.

**1.3 Toronto neighborhoods average after tax income broken down by postal code**

Here we must manually download these from Stats Canada and load them.  
<https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/search-recherche/change-geo.cfm?Lang=E&Geo1=FSA>  
See: to\_geo\_space.csv

**1.3.1 Merge Toronto Neighbourhood income data with Toronto Postal Code data**

Next, I joined income data to the Toronto Data.  
At this time I also saved a copy of the data set as my friend had asked for it in his list of requirements.  
See: TO\_Affluence.csv

**1.4 What is the Canadian National Average After Tax Income**

Here I must also manually download this from Stats Canada and load them.  
<https://www150.statcan.gc.ca/n1/daily-quotidien/180313/dq180313a-eng.htm>  
Canadian families and unattached individuals had a median after-tax income of $57,000 in 2016.

***Key Observation: Of the 103 Toronto Neighborhoods gathered only 55.3% or 57 Neighborhoods are above the median after-tax income. 37.8% or 39 Neighborhoods are below he median after-tax income. 6.7% or 7 neighborhoods did not register as it appears their populations are too low. It appears that the greatest concentration of affluence is near central Toronto. We decided to keep all neighborhoods in the dataset regardless of income of population as the majority were close enough.***

**1.5 Toronto list of Restaurants or Venues that could potentially use Restaurant Equipment**

FOURSQUARE API [https://api.foursquare.com](https://api.foursquare.com/)

**1.5.1 Get all the Venues in Toronto.**

**1.5.2 Only add Food Trucks as Venue Categories**

Use this list to food truck and related venues.

**1.5.3 OneHot encode and count restaurants**

Prepare the data for clustering

***Combine all of those into a working Data Set to cluster and geo spatial map of the results showing the best neighborhood to open a Restaurant Supply Store***

Combining all of these disparate data sets will clearly demonstrate the following:

* which neighborhoods in Toronto have clusters of like Restaurants
* how populated each neighborhoods is
* the average after tax income is all of these neighborhoods
* which neighborhood should he target to open his new store.

**Methodology:**

**Choice of Algorithms**

I chose K-Means Clustering.

<https://towardsdatascience.com/clustering-algorithms-for-customer-segmentation-af637c6830ac>

A backgrounder on K-Means clustering “K-means clustering is an iterative clustering algorithm where the number of clusters K is predetermined and the algorithm iteratively assigns each data  
point to one of the K clusters based on the feature similarity.”

***Key Observation: And for my project feature similarity means restaurant similarity in Neighborhoods***

**Choosing the correct number of clusters.**

<https://www.jeremyjordan.me/grouping-data-points-with-k-means-clustering/>  
Here I use Silhouette analysis to determine the optimum number of clusters to use.

A backgrounder on Silhouette analysis.

“We can use Silhouette analysis to evaluate each model. A Silhouette coefficient is calculated for observation, which is then averaged to determine the Silhouette score.  
The coefficient combines the average within-cluster distance with average nearest-cluster distance to assign a value between -1 and 1. A value below zero  
denotes that the observation is probably in the wrong cluster and a value closer to 1 denotes that the observation is a great fit for the cluster and clearly separated from other clusters. This coefficient essentially measures how close an observation is to neighboring clusters, where it is desirable to be the maximum distance possible from neighboring clusters. We can automatically determine the best number of clusters, k, by selecting the model which yields the highest Silhouette score.”

***Key Observation: My highest score was 2.***

**2.1 Run K means and segment data into clusters and generate labels**

**2.2 Merge the Toronto data with geo coordinates data and make sure it's the right shape**

Here I reshape the Toronto data so that it’s shape matches the clustered data.

**2.3 Add the KMeans Labels**

Determine the largest cluster in this case it was cluster number 2 with a shape of  
(76, 15)

**3. Cluster 2 Contains the highest cluster density. We need to find the geographic centroid for this cluster. This is the optimum location for a new Restaurant Supply Store.**

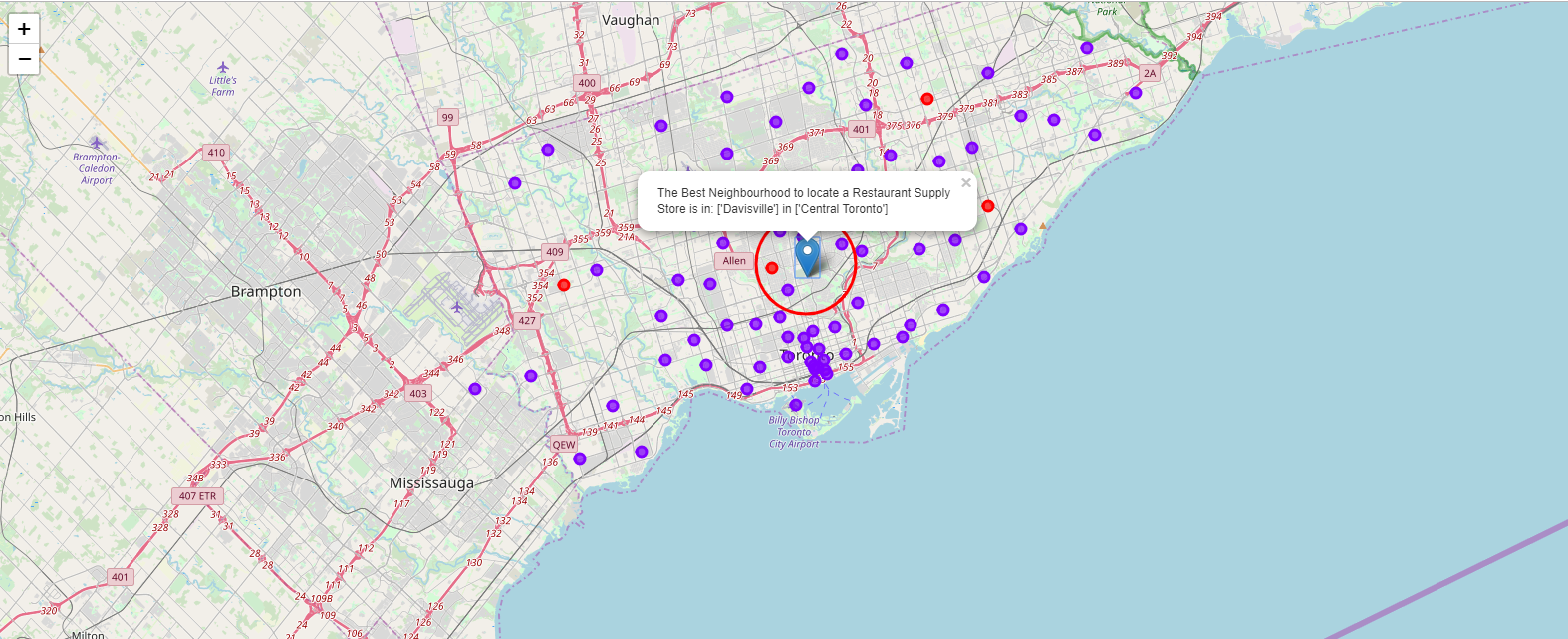
Here we take the average latitude and longitude to be the centroid.

**3.1 Install opencage to reverse lookup the coordinates**

Opencage allows me to reverse lookup the geo coordinates.  
***Key Observation: This is the optimum location for a new Restaurant Supply Store.***

**Results:**

**4.1 Plot the clusters on a Map of the Toronto and Super Impose the best location of a Store**



**4.2 Exact Address of desired Location**

Based on a reverse Lookup The exact Address to locate would be: 268 Balliol Street, ON M4S 1C2, Canada or lat: 43.6991598, lng: -79.3878871

**Discussion:**

**5.1 Explaining the results**

As we built our list of neighborhoods with Restaurant venues exclusively we discovered most neighborhoods were similar and the greatest concentration of restaurants was in Central Toronto and downtown Toronto. This might seem obvious but it would also appear that these are some of the most affluent neighborhoods in Toronto so there appears to be correlation. By Locating in the general vicinity of the Exact location my friend could be geographically centered in this cluster and poised to service his restaurant customer base with greatest efficiency.

When we built our our K-Means dataset we used Silhouette analysis to tell us there was a lot of similarity between neighborhoods and the most common restaurants contained within. Really there was only 2 types of cluster or neighborhoods in greater Toronto. The vast majority of those were in 1 cluster. So Toronto restaurants might be many but they are very homogeneously located near the center of Toronto.

Of the 103 Toronto Neighborhoods gathered only 55.3% or 57 Neighborhoods are above the median after-tax income. 37.8% or 39 Neighborhoods are below he median after-tax income. 6.7% or 7 neighborhoods did not register as it appears their populations are too low. It appears that the greatest concentration of affluence is near central Toronto. We decided to keep all neighborhoods in the dataset regardless of income of population as the majority were close enough.

**Conclusion:**

I feel confident with the recommendation I have given my friend as it is backed up with demonstrated data analysis. While nothing can ever be 100% certain he will certainly be better informed than he was prior to asking for my help.

Much more inference can be obtained with more work. A potential side business for my friend might be assisting new restaurant owners where they might locate a new restaurant, who their competition is and who their clientele might be.