Capstone: Find the best neighborhood in Toronto to open a Restaurant Supply Store

Applied Data Science Capstone

IBM Data Science Professional Certificate

INTRODUCTION:

- A friend of mine who runs a leading Restaurant Supply Store has found out that I am studying data science and has asked for help in trying to determine which neighborhood in Toronto he should open his new store in.
  Example Company:
  http://www.bramainc.com/about-brama

- I begin with an interview with my friend to determine the requirements.

Problem Description:

➢ **Which neighborhood should my friend open his new Restaurant Supply store in Toronto?**
He wants to ensure steady and sustainable business.

Requirements:

1. Store needs to be strategically located inside the biggest concentration of restaurants in Toronto area.

2. Confirm any assumption by means of modeling and testing the data. Specifically, visually cluster common restaurants in Toronto by neighborhood.

3. Additionally determine that a good number people can frequent these restaurants with sustainable frequency inside these neighborhoods.

a.) Is the neighborhood populous?

b.) Is the neighborhood average salary close to the Canadian National Average?

❑ My friend wants to be able to judge which neighborhoods also may be poised to grow in restaurant numbers in coming years.

❑ Locating his new store according to these requirements will ensure the following:

- lowest cost for delivery
- shortest travel time to his store for his clients
- overall lower run costs
- increase in overall business
- overall greater customer satisfaction

Data

- You can follow along in my Capstone Notebook located here:

- https://github.com/vir007/Coursera_Capstone/blob/master/Week%204_5/Week_5.ipynb

- **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

  - ***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

Data cont'd:

- ***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. ***

Data Cont.

- ***1.5 Toronto list of Restaurants or Venues that could potentially use Restaurant Equipment***
- 4SQUARE API
  https://api.foursquare.com

- ***1.5.1 Get all the Venues in Toronto.***
- ***1.5.2 Only add Restaurants as Venue Categories***
- Use this list to Extract Restaurants and only include Restaurants in our Data Set.

- ***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**

Methodology cont’d:

- **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.

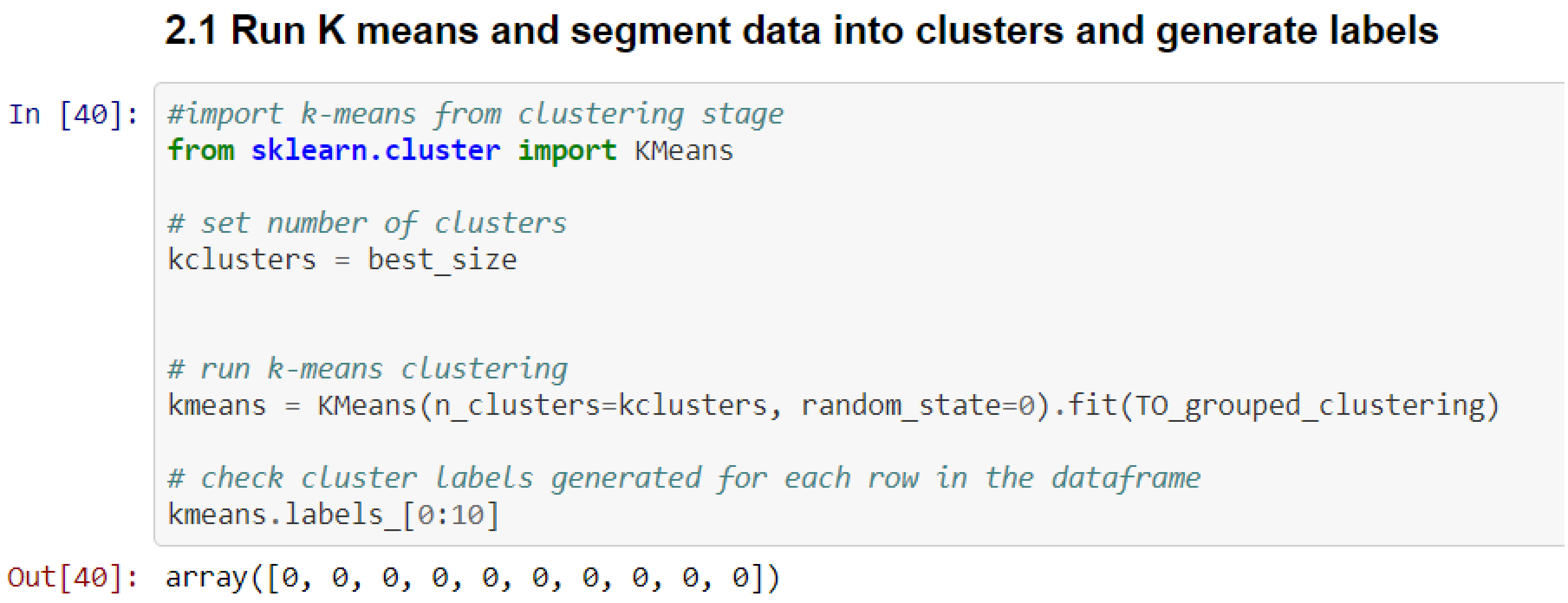
- A backgrounder on Silhouette analysis.
- “We can use Silhouette analysis to evaluate each model. A Silhouette coefficient is calculated for observation, which is then averaged to determine the Silhouette score.

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 3. ***

Methodology cont’d:

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



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

```
In [40]: #import k-means from clustering stage
         from sklearn.cluster import KMeans

         # set number of clusters
         kclusters = best_size


         # run k-means clustering
         kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(TO_grouped_clustering)

         # check cluster labels generated for each row in the dataframe
         kmeans.labels_[0:10]

Out[40]: array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

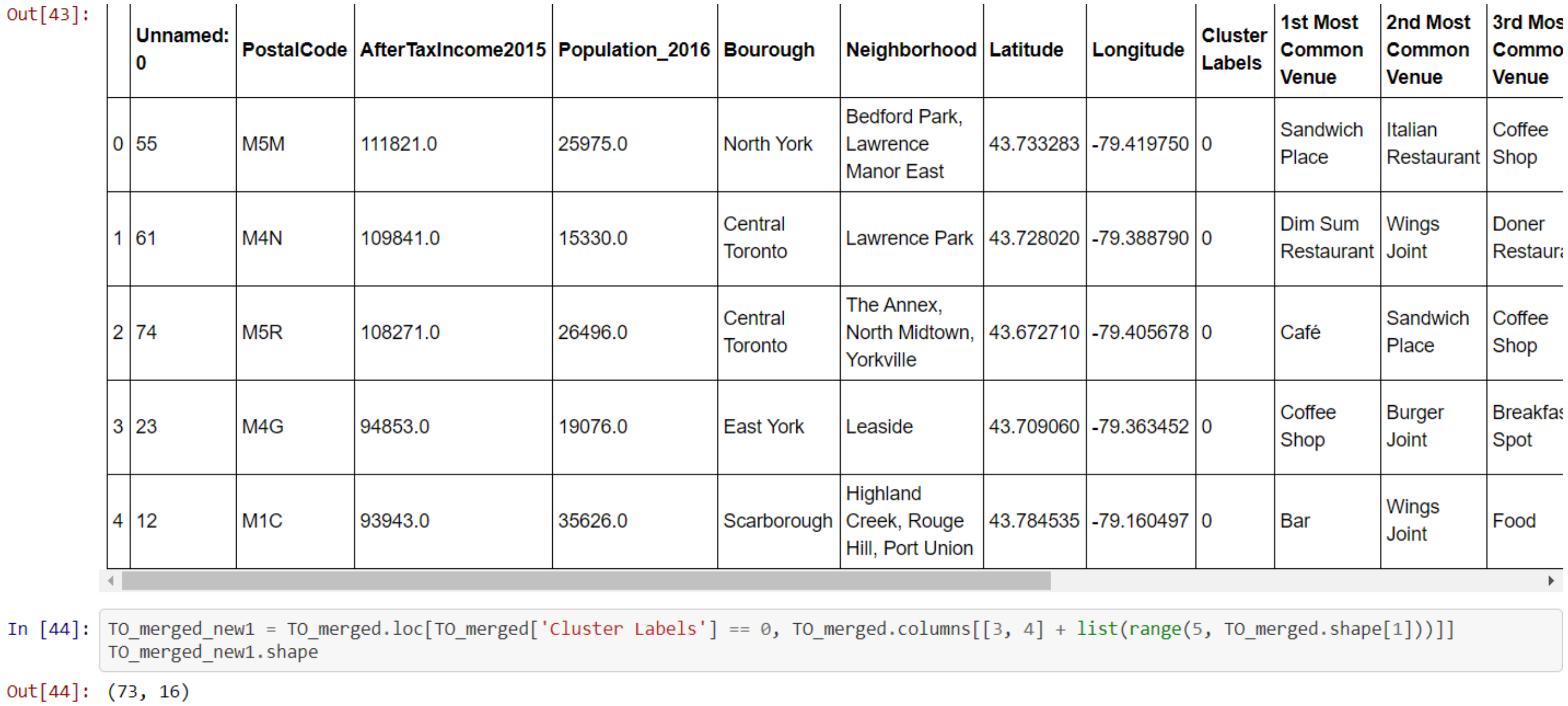
Methodology cont’d:

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 1 with a shape of (73, 16)



Out[43]:

	Unnamed: 0	PostalCode	AfterTaxIncome2015	Population_2016	Bourough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Mos Commo Venue
0	55	M5M	111821.0	25975.0	North York	Bedford Park, Lawrence Manor East	43.733283	-79.419750	0	Sandwich Place	Italian Restaurant	Coffee Shop
1	61	M4N	109841.0	15330.0	Central Toronto	Lawrence Park	43.728020	-79.388790	0	Dim Sum Restaurant	Wings Joint	Doner Restaura
2	74	M5R	108271.0	26496.0	Central Toronto	The Annex, North Midtown, Yorkville	43.672710	-79.405678	0	Café	Sandwich Place	Coffee Shop
3	23	M4G	94853.0	19076.0	East York	Leaside	43.709060	-79.363452	0	Coffee Shop	Burger Joint	Breakfas Spot
4	12	M1C	93943.0	35626.0	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497	0	Bar	Wings Joint	Food

In [44]:

```
TO_merged_new1 = TO_merged.loc[TO_merged['Cluster Labels'] == 0, TO_merged.columns[[3, 4] + list(range(5, TO_merged.shape[1]))]]
TO_merged_new1.shape
```

Out[44]:

```
(73, 16)
```

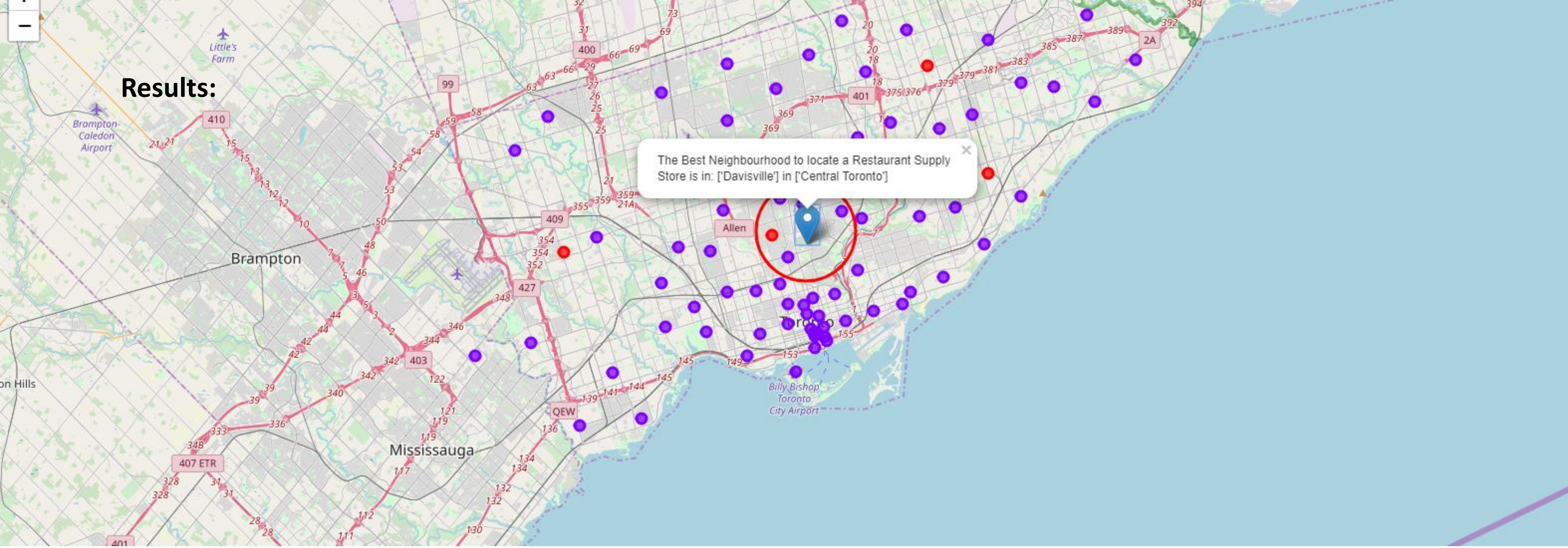
Methodology cont’d:

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.

*** Key Observation: This is the optimum location for a new Restaurant Supply Store.***
3.1 Plot the clusters on a Map of the Toronto and Super Impose the best location of a Store¶

Results:



The Best Neighbourhood to locate a Restaurant Supply Store is in: ['Davisville'] in ['Central Toronto']

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

Results Cont’d:

- **4.2 Exact Address of desired Location**

The exact Address to locate would be: lat: 43.6998426260274, 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 K-Means dataset we used Silhouette analysis to tell us there was a lot of similarity between neighborhoods and the most common restaurants contained with in. 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.