IBM Professional Data Science Capstone Project

Battle of The Neighbourhoods

(Submitted by Prateek Dang)

Index

Introduction

This final report is the part of the IBM Professional Data Science project, where the learners are left to solve a real-world problem with data science using the help of the Foursqaure API and/or other data sources.

The Problem

Traveling has always been a part of every life form on this planet. Humans specifically travel for many reasons like vacation, moving cities/countries for work or perhaps for education. Traveling is always going to be an integral part of our lives. While traveling to somewhere new may sound very exciting, it also comes with several cons. One of the most common one being finding the right place to purchase or utilize a commodity or service from, and this is the problem we are going to be solving here.

We are going to utilize the various data science and machine learning tools available at our disposal to work with this problem and try to find a solution to it. For the sake of this project we are going to be working with some of the most common type of shops in the area of Toronto, Canada and by the end we will be able to tell the most common shops in various neighbourhoods in Toronto.

Target audience

Though this project focuses only on the area of Toronto, the methodologies used can be applied to any location on this planet (results may vary). Therefore, the target audience for this particular project can be one of the following:

- An Immigrant to Toronto from a foreign country.
- A student from any other place in Canada.
- Tourists visiting Toronto.

The Data

The data used for this project is derived from the following three sources:

- https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M (This contains all the postal code of the various Neighbourhood in Toronto, along with their respective Boroughs.)[1]
- http://cocl.us/Geospatial_data (This contains all the latitude and longitude data for each of the neighbourhood which is needed for visualization.)[2]
- Json file of all the near-by shops of every neighbourhood derived by the help of the Foursquare API.[3]

1) **The neighbourhoods of Toronto**
   This data set is derived from the Wikipedia link given above and it contains the following columns:
   - Postcode
   - Borough
   - Neighbourhood

	Postcode	Borough	Neighborhood
1	M1A	Not assigned	Not assigned
2	M2A	Not assigned	Not assigned
3	M3A	North York	Parkwoods
4	M4A	North York	Victoria Village
5	M5A	Downtown Toronto	Harbourfront

Fig: Neighbourhoods of Toronto

As it can be seen from the snapshot above, there are some discrepancies in our data since some of the Boroughs and Neighbourhoods are not assigned any value. Thus, there is a need a for cleaning this data. The following steps were taken:

- Dropping the rows where the borough is not assigned.
- Since many neighbourhoods have the same borough, the entire data set was grouped on the basis of the Postcode and Borough.

The resultant data:

	Postcode	Borough	Neighborhood
3	M3A	North York	Parkwoods
4	M4A	North York	Victoria Village
5	M5A	Downtown Toronto	Harbourfront
6	M6A	North York	Lawrence Heights
7	M6A	North York	Lawrence Manor

Fig: Resultant data after the above steps

2) **Combining with the Geospatial data**

Our neighbourhood data is now ready to be combined with its latitude and longitude values, the geospatial data consisted of the following:

- Postcode
- Latitude
- Longitude

	Postcode	lattitude	longitutde
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Fig: Geospatial data

This data was then merged with the neighbourhood dataset with the help of the Postcode column (it being common in both Geospatial and Neighbourhood data set. The result of that looked something like this:

	Postcode	Borough	Neighborhood	lattitude	longitutde
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

Fig: Neighbourhood and geospatial data combination

3) **Json data of the nearby-shops**

Lastly it was now time to gather the data of the near by shops of the various neighbourhood.

This was possible with the help of the Foursquare API and the geospatial data. The following code snippet shows how this data was derived:

```
def getNearbyHotels(names, latitudes, longitudes, radius=1000):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        #print(name)

        # create the API request URL
        url=f'https://api.foursquare.com/v2/venues/search?&ll={lat},{lng}&query=shops&client_id={client_id}
        &client_secret={client_secret}&v={version_}'

        # make the GET request
        results = requests.get(url).json()["response"]['venues']

        # return only relevant information for each nearby venue
        for i in range(len(results)):
            if results[i]['categories']== []:
                pass
            else:
                venues_list.append([(
                    name,
                    lat,
                    lng,
                    results[i]['categories'][0]['name']) ])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue Category']

    return(nearby_venues)
```

Fig: deriving the nearby-shops data

The final data set contains the following columns:

- Name of the neighbourhood(s).
- Their latitude and Longitude values
- Category of the shop

The data contained many different types of shops and some shops didn't have any categories assigned and so it was time to perform some more cleaning on the data.

After some thorough testing and consideration, the following categories were picked for performing our analysis:

- Automotive shop
- Bank
- Clothing Store
- Doctor's Office
- Women's store
- Furniture/ home store
- Saloon/barber shop

Some may argue that why restaurants weren't considered since they are one of the most common "shops" used on a day-to-day basis, the reason for that is the restaurants are of many different kinds with many categories thus it would have led to plenty of noise in the data.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Category
Rouge,Malvern	43.806686	-79.194353	Automotive Shop
Rouge,Malvern	43.806686	-79.194353	Doctor's Office
Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	Automotive Shop
Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	Doctor's Office
Guildwood,Morningside,West Hill	43.763573	-79.188711	Automotive Shop
Guildwood,Morningside,West Hill	43.763573	-79.188711	Doctor's Office
Woburn	43.770992	-79.216917	Automotive Shop
Woburn	43.770992	-79.216917	Doctor's Office
Cedarbrae	43.773136	-79.239476	Automotive Shop

Fig: neighbourhood data with the various shop categories

The Venue Category (shop categories) consists of categorical values which had to be converted into numeric values and the repeated entries were grouped. The following is the result of those operations:

	key_0	Neighborhood Latitude	Neighborhood Longitude	Automotive Shop	Bank	Clothing Store	Doctor's Office	Furniture / Home Store	Salon / Barbershop	Women's Store
0	Adelaide,King,Richmond	43.650571	-79.384568	0.333333	0.000000	0.00	0.000000	0.333333	0.333333	0.0
1	Agincourt	43.794200	-79.262029	0.500000	0.000000	0.00	0.500000	0.000000	0.000000	0.0
2	Agincourt North,L'Amoreaux East,Milliken,Steel...	43.815252	-79.284577	0.333333	0.333333	0.00	0.333333	0.000000	0.000000	0.0
3	Albion Gardens,Beaumond Heights,Humbergate,Jam...	43.739416	-79.588437	1.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.0
4	Alderwood,Long Branch	43.602414	-79.543484	1.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.0
5	Bathurst Manor,Downsview North,Wilson Heights	43.754328	-79.442259	0.500000	0.000000	0.25	0.250000	0.000000	0.000000	0.0

Fig: the final data set

The snapshot above shows a portion of the final data set that is used for the analysis.

As it can be noticed the categories of the shops now have their own column each and there are no redundant entries of any neighbourhood, each row consists of the mean of the frequencies of shops for each Neighbourhood (denoted by key_0).

The Methodology

This project makes use of Python, Folium (for plotting map data), Matplotlib (for visualising the data set) and Sci-kit Learns' KMeans algorithm (for making the clusters).

As discussed above the objective of this project is to determine the locations where a particular category of shop might be located all over the city of Toronto, so we start off by plotting the latitudes and longitudes of the various boroughs as seen below:

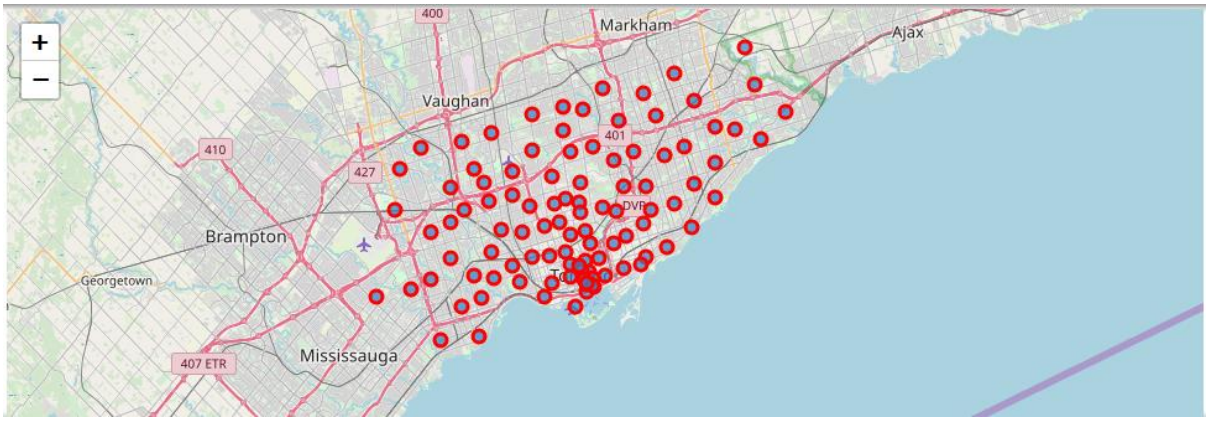

Fig: Borough plotted on the map of Toronto with Folium

After all the scrapping and cleaning we then arrive at our final data that consists of the frequencies of the various categories of shops all over Toronto along with the Neighbourhood names and their latitude and longitude values, the final data looks something like this:

key_0	Neighborhood Latitude	Neighborhood Longitude	Automotive Shop	Bank	Clothing Store	Doctor's Office	Furniture / Home Store	Salon / Barbershop	Women's Store
Rouge,Malvern	43.806686	-79.194353	1	0	0	0	0	0	0
Rouge,Malvern	43.806686	-79.194353	0	0	0	1	0	0	0
Rouge,Malvern	43.806686	-79.194353	1	0	0	0	0	0	0
Rouge,Malvern	43.806686	-79.194353	0	0	0	1	0	0	0
Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497	1	0	0	0	0	0	0

Fig: Final dataset

Let's perform some analysis on the data and visualize what category of shops has the highest frequency in Toronto, to do so we:

- We first find the sum of frequencies of each of the categories of shop
- We then plot a bar-graph of the data using matplotlib to visualize and compare the result

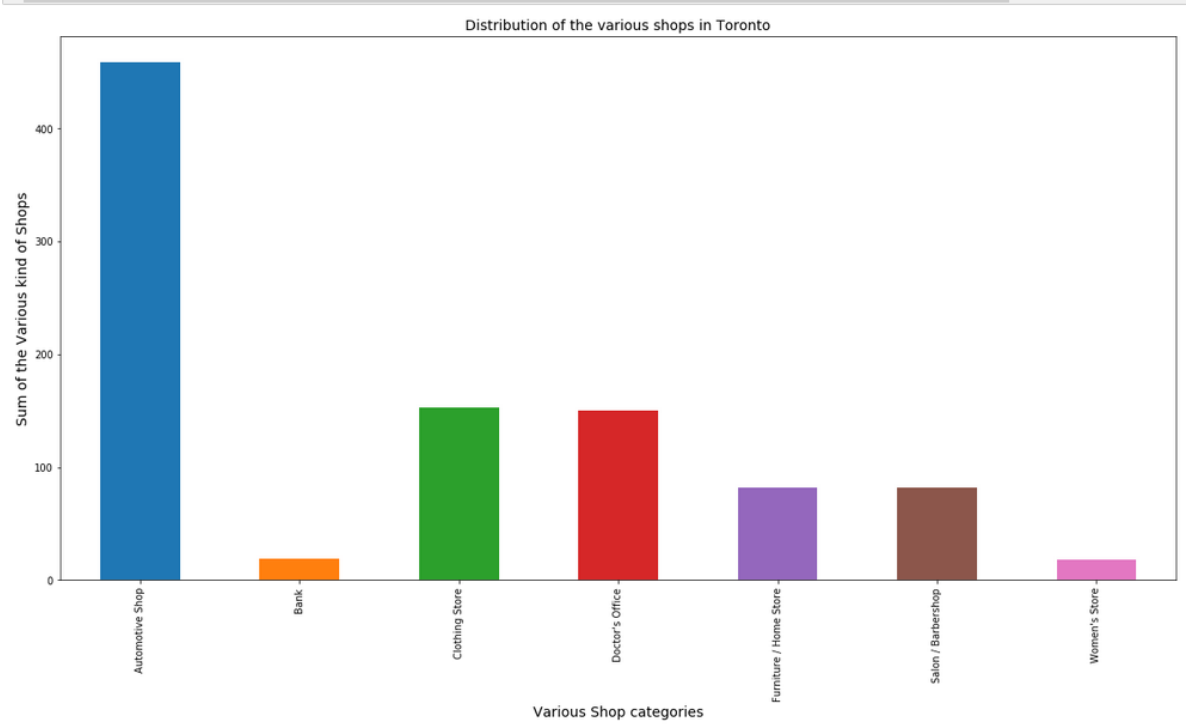

Fig: Bar graph generated with matplotlib

From the bar figure above we can clearly see that the Automotive Shops are in a huge number in Toronto, while banks and women's stores are quite few.

Now we have everything we need to create clusters to specify the localities with similar types of shops.

For this we use **KMeans** algorithm, Implementing the algorithm is really simple thanks to sci-kit learn library.

KMeans is an Unsupervised Machine learning algorithm that is used form clusters, this algorithm works with unlabelled data. More info on KMeans algorithm can be found on the link below:

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

The code snippet below shows how easily the algorithm was implemented:

```
#Importing KMeans from Sklearn.clusters
from sklearn.cluster import KMeans
#creating a KMeans object that will give us 5 clusters from our data
km=KMeans(n_clusters=5)
km.fit(cat_shops.iloc[:,3:])
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=5, n_init=10, n_jobs=None, precompute_distances='auto',
    random_state=None, tol=0.0001, verbose=0)
```

Fig: Implementing KMeans algorithm

You might notice the parameter 'n_clusters' and it's value is set to 5, what it does is that algorithm will create 5 different clusters from our unlabelled data. The output generated by the algorithm is in the form of a numpy array, as can be seen below:

```
km.labels_
```

```
array([3, 2, 2, 1, 1, 2, 2, 2, 3, 2, 1, 1, 0, 2, 3, 3, 2, 1, 2, 3, 3, 1,
       3, 2, 2, 2, 1, 3, 2, 2, 3, 1, 3, 4, 2, 1, 1, 2, 2, 2, 2, 0, 1, 2,
       3, 4, 2, 2, 2, 3, 3, 3, 1, 2, 2, 1, 1, 2, 2, 1, 1, 2, 2, 4, 4, 3,
       4, 3, 2, 2, 1, 2, 1, 1, 4, 3, 3, 2, 2, 1, 3, 2, 4, 3, 3, 3, 3, 1,
       0, 3, 1, 1, 4, 2, 4, 1, 1, 2, 2, 2, 4, 1, 2])
```

Fig: Output of the algorithm

The Result

To draw some results from our project let's first visualize our formed clusters on the map of Toronto with the help of Folium

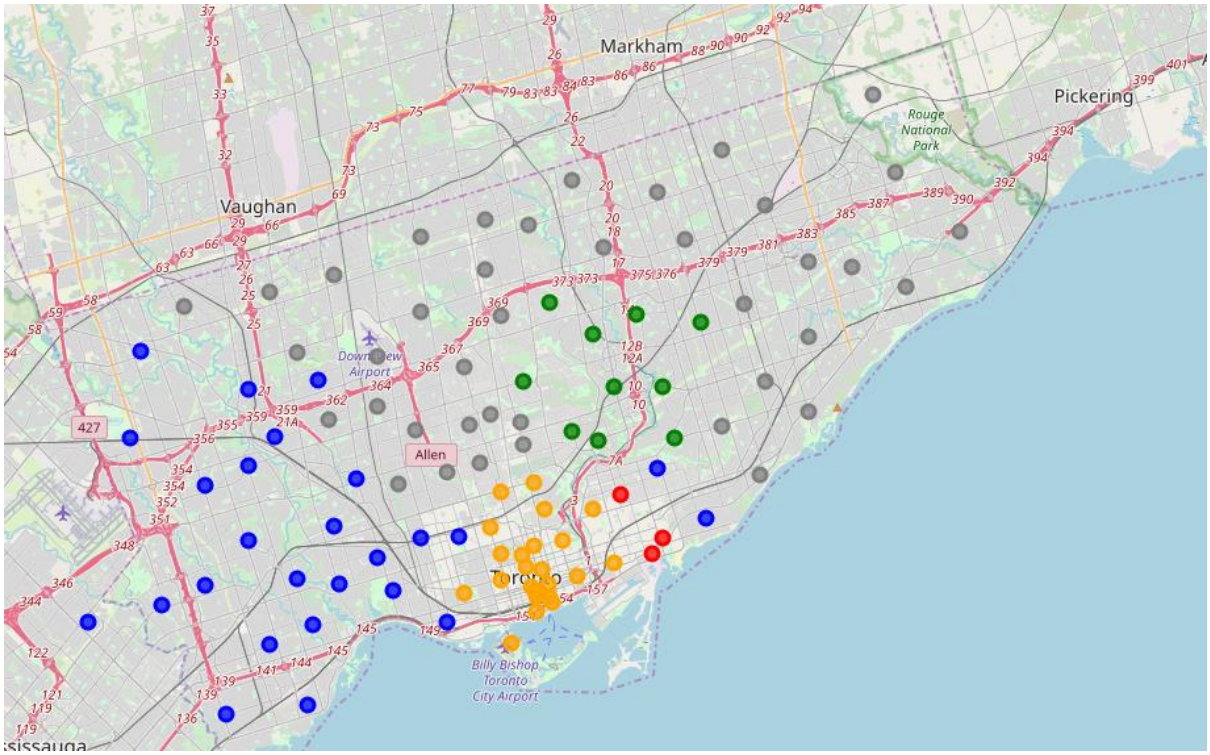

Fig: Visualization of clusters

We can see that the clusters are well spread out, lets dive into the story these clusters convey:

- Cluster number 1 (Red Marker)
  - ➢ This Cluster consists of Automotive shops and saloons/barber shops
- Cluster number 2 (Blue Marker)
  - ➢ This cluster consists of mostly just Automotive shops
- Cluster number 3 (Gray Marker)
  - ➢ This cluster consists of again Automotive shops, Doctor's office and Banks
- Cluster number 4 (Orange Marker)
  - ➢ This cluster consists of yet again Automotive shops, Furniture / Home store and some Saloon/ Barber shop
- Cluster number 5(Green Marker)

- This cluster, once again consists of Automotive shops, Clothing Store and Women's Stores

Conclusion

The results from the clusters definitely coincide with the bar graph, as it can be seen that Automotive Shops are the most popular and spread all over Toronto.

We can also see that the other categories of the shops are also well spread throughout the city.

In the end of this capstone project report battle of the neighbourhoods I would like to say that I enjoyed every second spent working on this and I also gained tons of knowledge from this course.