Applied Data Science Capstone

RISHABH TAMHANE

This report involves segmenting and clustering the neighborhoods in the city of Toronto, Canada.

1. INTRODUCTION

When a Restaurant Chain plans on opening a chain of restaurants in a city, there are many factors that need to be considered:
1) The most commonly visited category of restaurants in the region.
2) The most optimal locations for setting up the restaurants so that they don't end up competing against each other, thereby reducing the profits of the restaurant chain.
3) The locations of the restaurants the Restaurant Chain plans on competing against so as to avoid cut throat competition against closely clustered restaurants.

2. METHODOLOGY

1. Web Scraping Data about the Neighborhoods in Toronto from Wikipedia
2. Web Scraping Data about the Restaurants in the Neighborhood
3. Cleaning and Organising the Data
4. Exploratory Data Analysis
5. Conclusion

3. WEB SCRAPING DATA ABOUT THE NEIGHBORHOODS IN TORONTO FROM WIKIPEDIA

After scraping and cleaning the data obtained from Wikipedia, the following dataframe provides us information about the Postal Code, Borough, Neighborhood and their Locations.

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	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. S1. Neighborhoods in Toronto

4. WEB SCRAPING DATA ABOUT THE RESTAURANTS IN THE NEIGHBORHOOD

Foursquare API, is a local search-and-discovery API developed by Foursquare Labs Inc. The API provides personalized recommendations of places to go near a user's location. It lets users search for restaurants, nightlife spots, shops and other places in a location. The app displays personalized recommendations based on factors that include the time of day, a user's check-in history, their "Tastes" and their venue ratings.

Using the data obtained in the table S1, we can now extract restaurants in the 2 kilometer radius of each of the Pin Code Locations. Merging the restaurant details obtained from each location and deleting the copies of the same restaurant extracted from multiple locations, we get the following data frame.

	name	categories	address	cc	city	country	crossStreet	distance	formattedAddress	labeledLatLngs
0	Ted's Restaurant	Breakfast Spot	404 Old Kingston Rd.	CA	Scarborough	Canada	NaN	3194	[404 Old Kingston Rd., Scarborough ON, Canada]	[{'label': 'display', 43.78446796744
1	Africa Restaurant	Food & Drink Shop	NaN	CA	Toronto Division	Canada	NaN	1636	[Toronto Division ON, Canada]	[{'label': 'display', 43.81958724358
2	Rex's Den Restaurant	Burger Joint	1265 Military Tr	CA	Toronto	Canada	NaN	2485	[1265 Military Tr, Toronto ON, Canada]	[{'label': 'display', 43.78500731279
3	Grapevine Restaurant	Restaurant	NaN	CA	NaN	Canada	NaN	3042	[Canada]	[{'label': 'display', 43.78047180175
4	Sabby's Restaurant	Caribbean Restaurant	5780 Sheppard Ave E	CA	Toronto	Canada	Markham Rd	3276	[5780 Sheppard Ave E (Markham Rd), Toronto ON,...	[{'label': 'display', 43.79554878818

Fig. S2. Restaurants in the Neighborhoods in Toronto

5. CLEANING AND ORGANISING THE DATA

After Cleaning and Organising the Data we obtain the following tables:
**Note: Only a few rows from the tables are represented in the table.

	name	categories	address	cc	city	country	crossStreet	distance	formattedAddress	labeledLatLngs	lat	lng
0	Ted's Restaurant	Breakfast Spot	404 Old Kingston Rd.	CA	Scarborough	Canada	NaN	3194	[404 Old Kingston Rd., Scarborough ON, Canada]	[{'label': 'display', 'lat': 43.78446796744621...	43.784468	-79.169200
1	Africa Restaurant	Food & Drink Shop	NaN	CA	Toronto Division	Canada	NaN	1636	[Toronto Division ON, Canada]	[{'label': 'display', 'lat': 43.81958724358436...	43.819587	-79.184574
2	Rex's Den Restaurant	Burger Joint	1265 Military Tr	CA	Toronto	Canada	NaN	2485	[1265 Military Tr, Toronto ON, Canada]	[{'label': 'display', 'lat': 43.78500731279124...	43.785007	-79.186964
3	Grapevine Restaurant	Restaurant	NaN	CA	NaN	Canada	NaN	3042	[Canada]	[{'label': 'display', 'lat': 43.78047180175781...	43.780472	-79.205078
4	Sabby's Restaurant	Caribbean Restaurant	5780 Sheppard Ave E	CA	Toronto	Canada	Markham Rd	3276	[5780 Sheppard Ave E (Markham Rd), Toronto ON,...	[{'label': 'display', 'lat': 43.79554878818409...	43.795549	-79.232105

Fig. S3. Details of the Restaurants in the Neighborhoods in Toronto

	categories	count
0	Chinese Restaurant	138
1	Restaurant	109
2	Caribbean Restaurant	46
3	Korean Restaurant	39
4	Indian Restaurant	38
5	Asian Restaurant	33

Fig. S4. Different Categories of Restaurants in the Neighborhoods in Toronto

	City	count
0	Toronto	586
1	Scarborough	68
2	Mississauga	36
3	North York	34
4	Etobicoke	33
5	Vaughan	15

Fig. S5. Count of Restaurants in the Neighborhoods in Toronto

6. EXPLORATORY DATA ANALYSIS(DISCUSSION)

From the bar graph represented in Fig.S6, the most common restaurants in the neighborhoods of Toronto are:

1. Chinese Restaurants
2. Caribbean Restaurants
3. Korean Restaurants

From the bar graph represented in Fig.S7, the most restaurants in the neighborhoods of Toronto are in:

1. Toronto
2. Scarborough
3. Mississauga

From the bar graph represented in Fig.S8, the most Chinese restaurants in the neighborhoods of Toronto are in:

1. Toronto
2. Scarborough
3. Markham

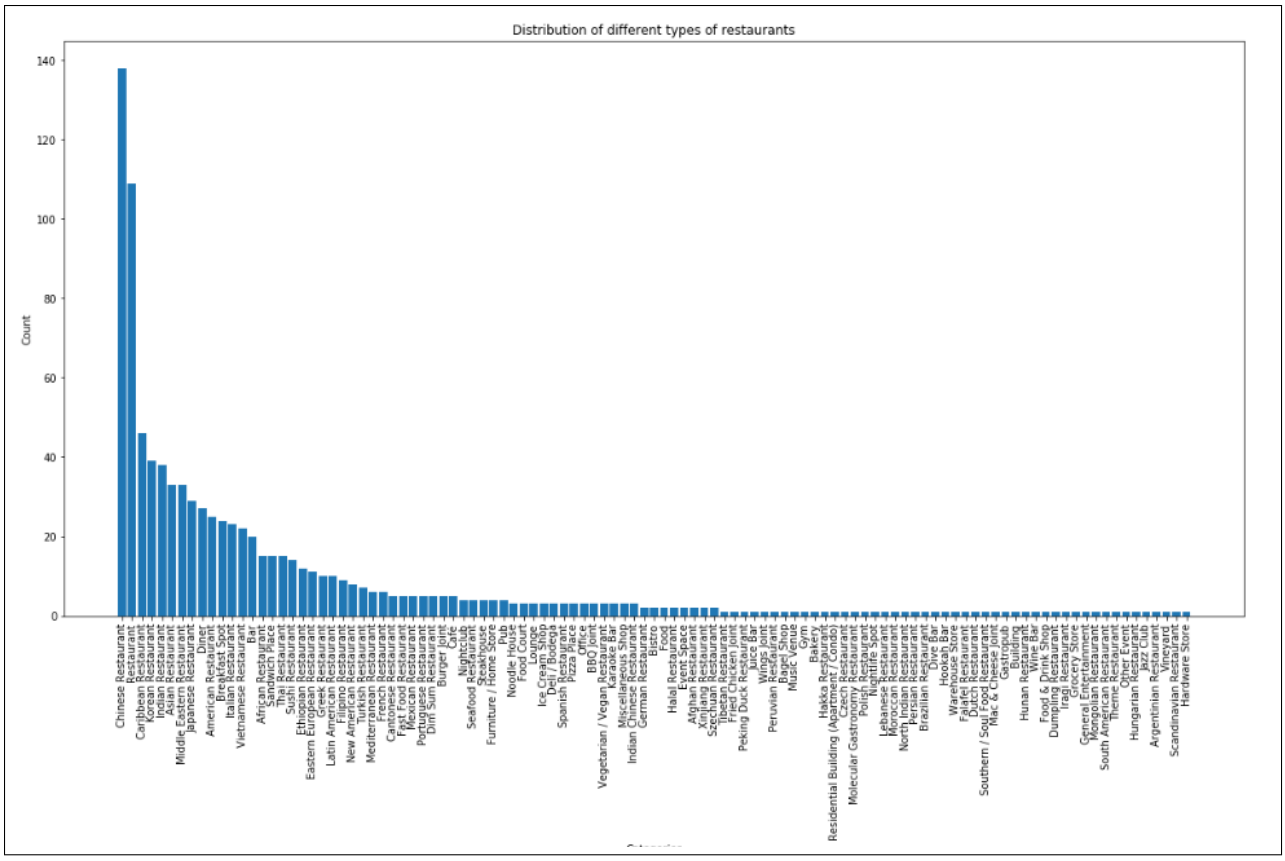

Fig. S6. Distribution of different categories of restaurants

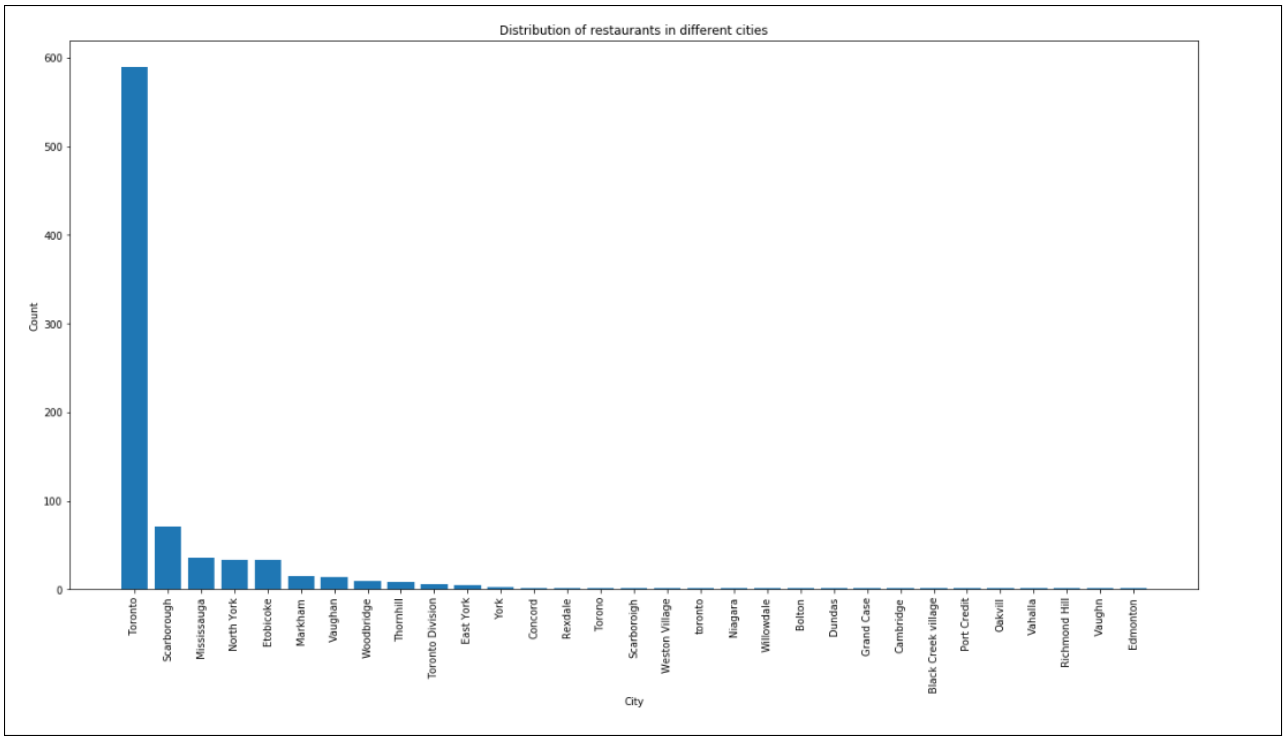

Fig. S7. Distribution of restaurants

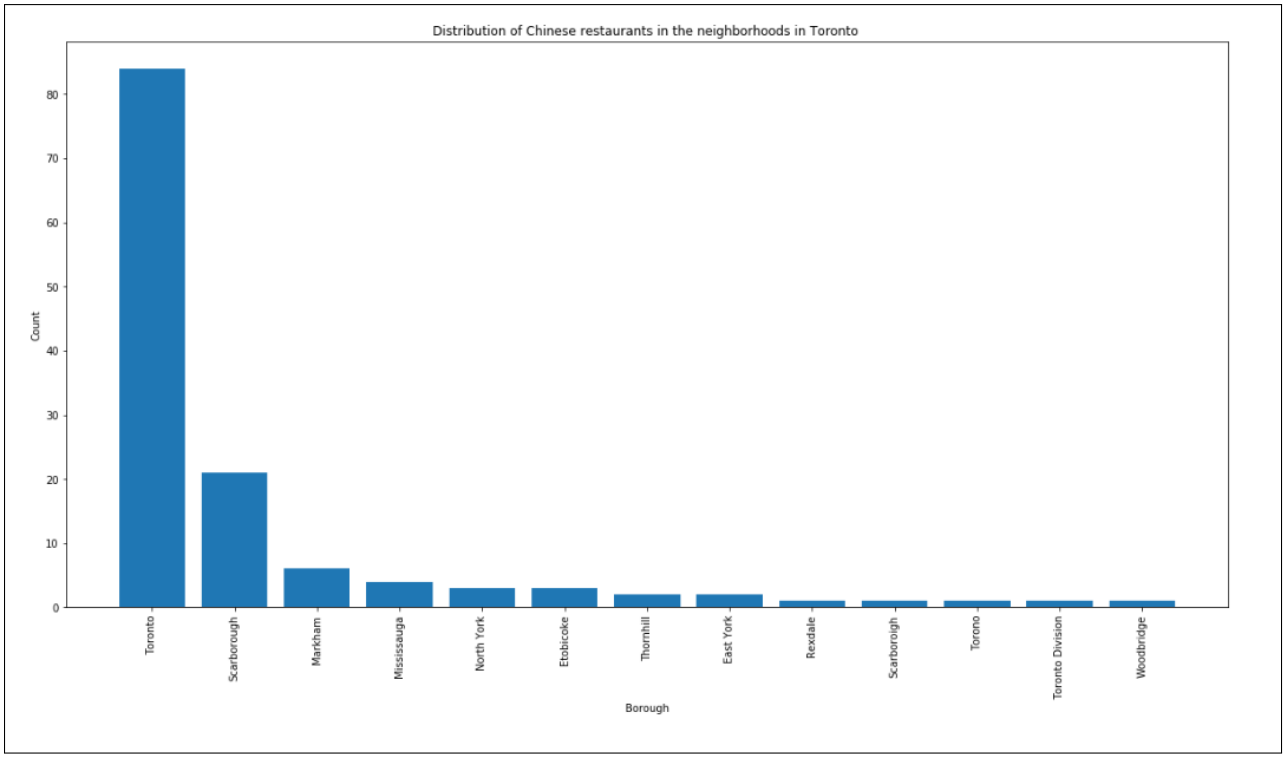

Fig. S8. Distribution of Chinese restaurants in the Neighborhood in Toronto

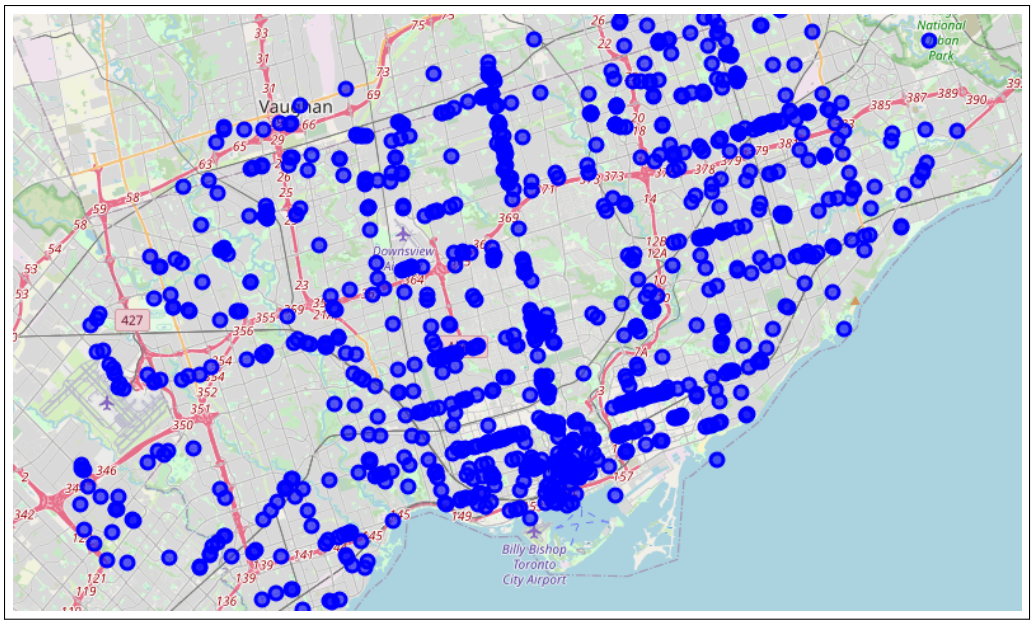

Fig. S9. Map Plot of the restaurants in the Neighborhoods in Toronto

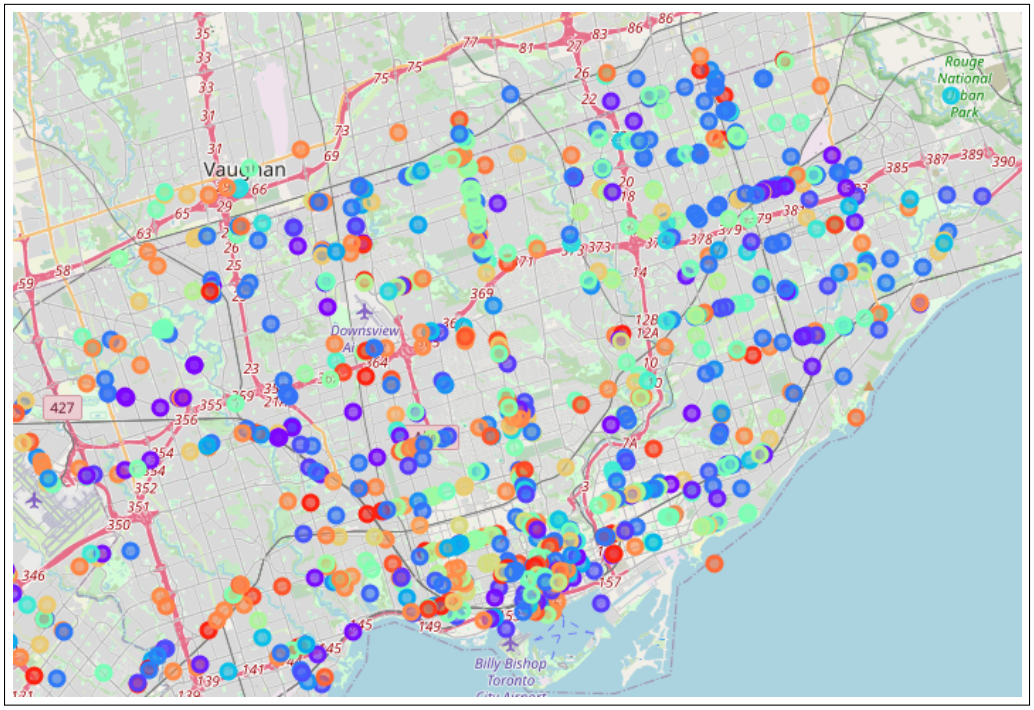

Fig. S10. Map Plot of classification of the restaurants based on the categories

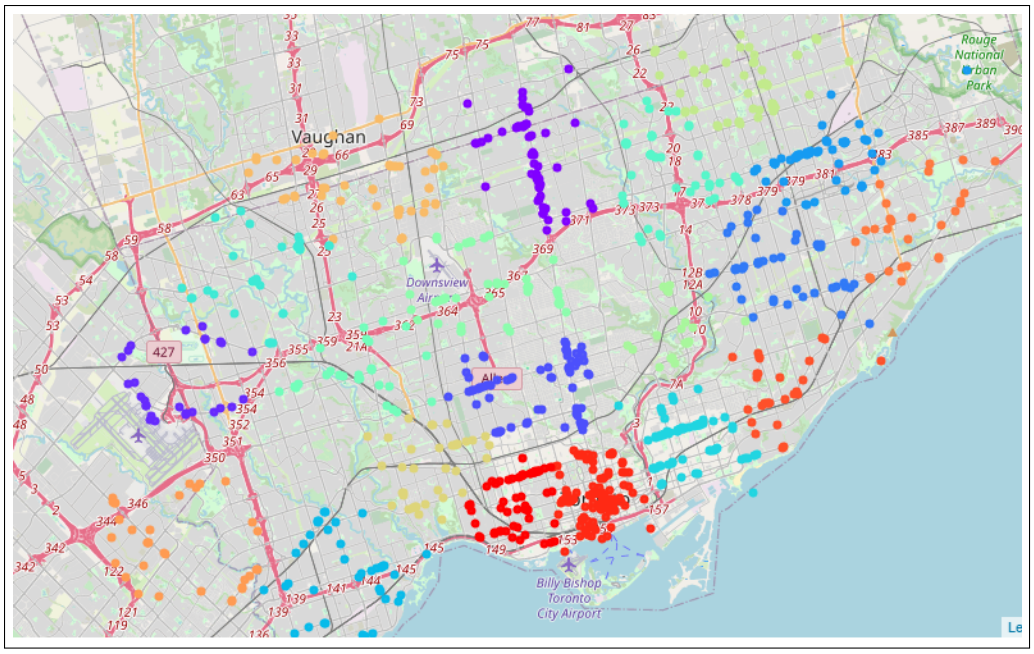

Fig. S11. Map Plot of classification of the restaurants based on the categories

7. CONCLUSION

1. Let us say I have a client who is looking to set up a restaurant in Canada. I have to make a recommendation about which is the most common category of restaurants in the neighborhood.
   Since, there are 139 Chinese Restaurants, they are the most common restaurants in the neighborhoods.

2. Now the client is planning to set up a chain of 20 Chinese restaurants in and around Toronto. So, I have to make a recommendation about the location of these restaurants so that they don't cannibalise each other

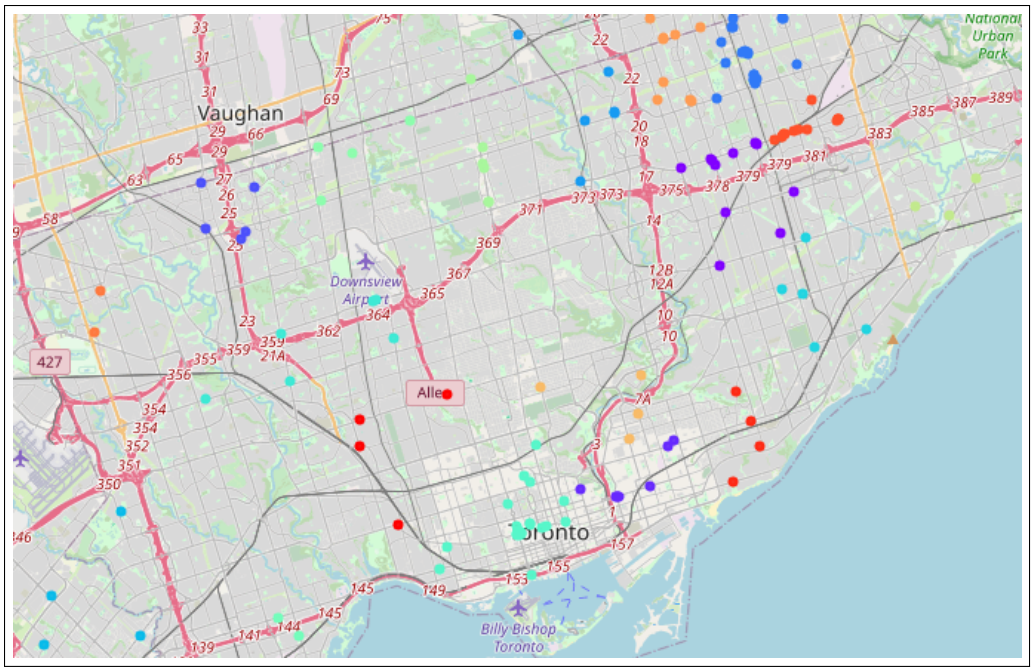

Fig. S12. Map Plot of Clustering of the Chinese Restaurants

Using K-Means Clustering which is an Unsupervised Machine Learning Algorithm, I clustered the Chinese Restaurants in 20 clusters. Now, I can recommend the Restaurant Chain to set up restaurants in each of these clusters, so that while catering to the customers, it will reduce the chances of cannibalisation amongst the restaurants belonging to the same restaurant chain. This approach also ensures that we cover a larger customer base while being aware of the competition prevalent in each area. It also makes sure that more restaurants are set up in the areas where Chinese Restaurants are more commonly visited than other areas.

8. REFERENCES

1. Foursquare API
2. IBM Applied Data Science Capstone
3. K-Means Clustering