

Capstone Project - The Battle of the Neighborhoods

Applied data science capstone Project

BY

HIMANSHI SHARMA

Presentation Outline

2

- ▶ Introduction: BUSINESS PROBLEM
- ▶ Background
- ▶ Problem description
- ▶ Target audience
- ▶ Success Criteria
- ▶ DATA
- ▶ Methodology
- ▶ Analysis
- ▶ conclusion

INTRODUCTION : BUSINESS PROBLEM

In this project we will try to find an location for a restaurant. Since there are lots of restaurants in Toronto we will try to detect locations that are not already crowded with restaurants. We are also particularly interested in areas with no Indian restaurants in vicinity. We would also prefer locations as close to city center as possible, assuming that first two conditions are met.

Background

There is a restaurant contractor in one of the boroughs of Toronto (Scarborough). This contractor looking for the places such as Different types of Restaurants, Bakery, Breakfast Spot, Brewery and Café with fresh and high-quality. The contractor wants to build a restaurant for the Italian food lovers inside the borough, so that they will support more customers and also bring better "Quality of Service" to the there customers.

Problem Description

Since there are lots of restaurants in Toronto we will try to detect locations that are not already crowded with restaurants. We are also particularly interested in areas with no Indian restaurants in vicinity. We would also prefer locations as close to city center as possible, assuming that first two conditions are met.

We will use our data science powers to generate a few most promising neighborhoods based on this criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders. this report will be targeted to stakeholders interested in opening an Italian restaurant in Toronto.

Target Audience

Target audience for this project will be all the stakeholders who want establish restaurants or café in Toronto neighborhood and want a less populated area so that the competition to other stakeholders will be less in comparison to others.

Success criteria

The success criteria of this project will be a recommendation of area for new restaurant opening for the Italian food lovers on the basis of clusters in the nearby area.

DATA

8

The dataframe will consist of three columns: PostalCode, Borough, and Neighborhood. Only process the cells that have an assigned borough. number of existing restaurants in the neighborhood (any type of restaurant) number of and distance to Italian restaurants in the neighborhood, if any distance of neighborhood from city center We decided to use regularly spaced grid of locations, centered around city center, to define our neighborhoods.

Following data sources will be needed to extract/generate the required information:

centers of candidate areas will be generated algorithmically and approximate addresses of centers of those areas will be obtained using Google Maps API reverse geocoding number of restaurants and their type and location in every neighborhood will be obtained using Foursquare API coordinate of Delhi center will be obtained using Google Maps API geocoding of well known toronto location.

METHODOLOGY

9

•Part 1: Identifying Postal Codes (and then Neighborhoods)

we need to have dataframe that contain both the data frames df1 and df_geo_coor so here it is

```
df_t = pd.merge(df1, df_geo_coor, how='inner', left_on = 'Postal Code', right_on = 'Postal Code')
```

```
df_t.shape
```

```
3]: (103, 5)
```

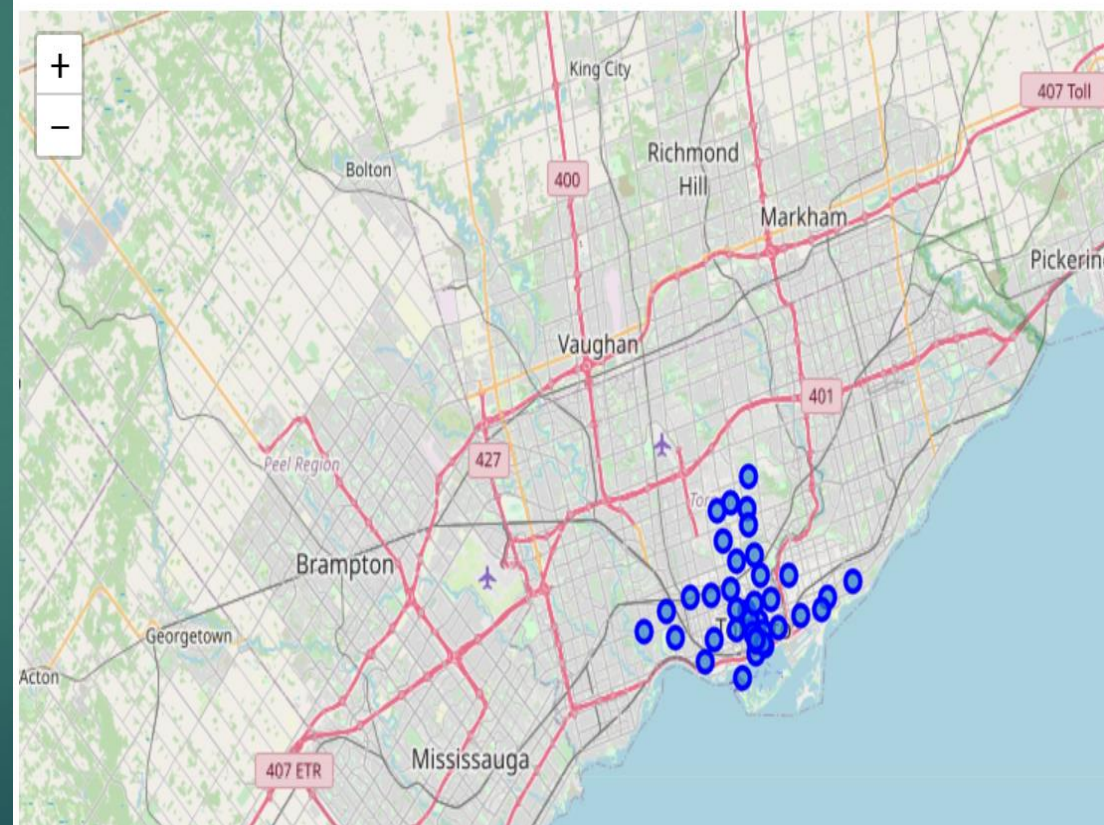
```
df_t.head()
```

```
9]:
```

	Postal Code	Borough	Neighbourhood	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	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

```
df_t.to_excel("part_22.xlsx")
```

The geographical coordinate of Toronto city are 43.6534817, -79.3839347.



Part 2: Connecting to Foursquare and Retrieving Locational Data for Each Venue in Every Neighborhood

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 1000 meter. It means that we have asked Foursquare to find venues that are at most 1000 meter far from the center of the neighborhood.

Part 3: Processing the Retrieved Data and Creating a DataFrame for All the Venues inside the Scarborough

- ▶ When the data is completely gathered, we will perform processing on that raw data to find our desirable features for each venue. Our main feature is the category of that venue. After this stage, the column "Venue's Category" will be One-hot encoded and different venues will have different feature-columns. After On-hot encoding we will integrate all restaurant columns to one column "Total Restaurants" and all food joint columns to "Total Joints" column.

Part 3: Processing the Retrieved Data and Creating a DataFrame for All the Venues

```
scarborough_venues.head()
```

	Postal Code	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Summary	Venue Category	Distance
0	M1W	Steeles West	43.799525	-79.318389	Mr Congee Chinese Cuisine 龍粥記	This spot is popular	Chinese Restaurant	72
1	M1W	Steeles West	43.799525	-79.318389	Agincourt Bakery	This spot is popular	Bakery	759
2	M1W	Steeles West	43.799525	-79.318389	Little Sheep Mongolian Hot Pot 小肥羊	This spot is popular	Hotpot Restaurant	972
3	M1W	Steeles West	43.799525	-79.318389	Phoenix Restaurant 金鳳餐廳	This spot is popular	Chinese Restaurant	147
4	M1W	Steeles West	43.799525	-79.318389	Price Chopper	This spot is popular	Grocery Store	16

ANAYLSIS

12

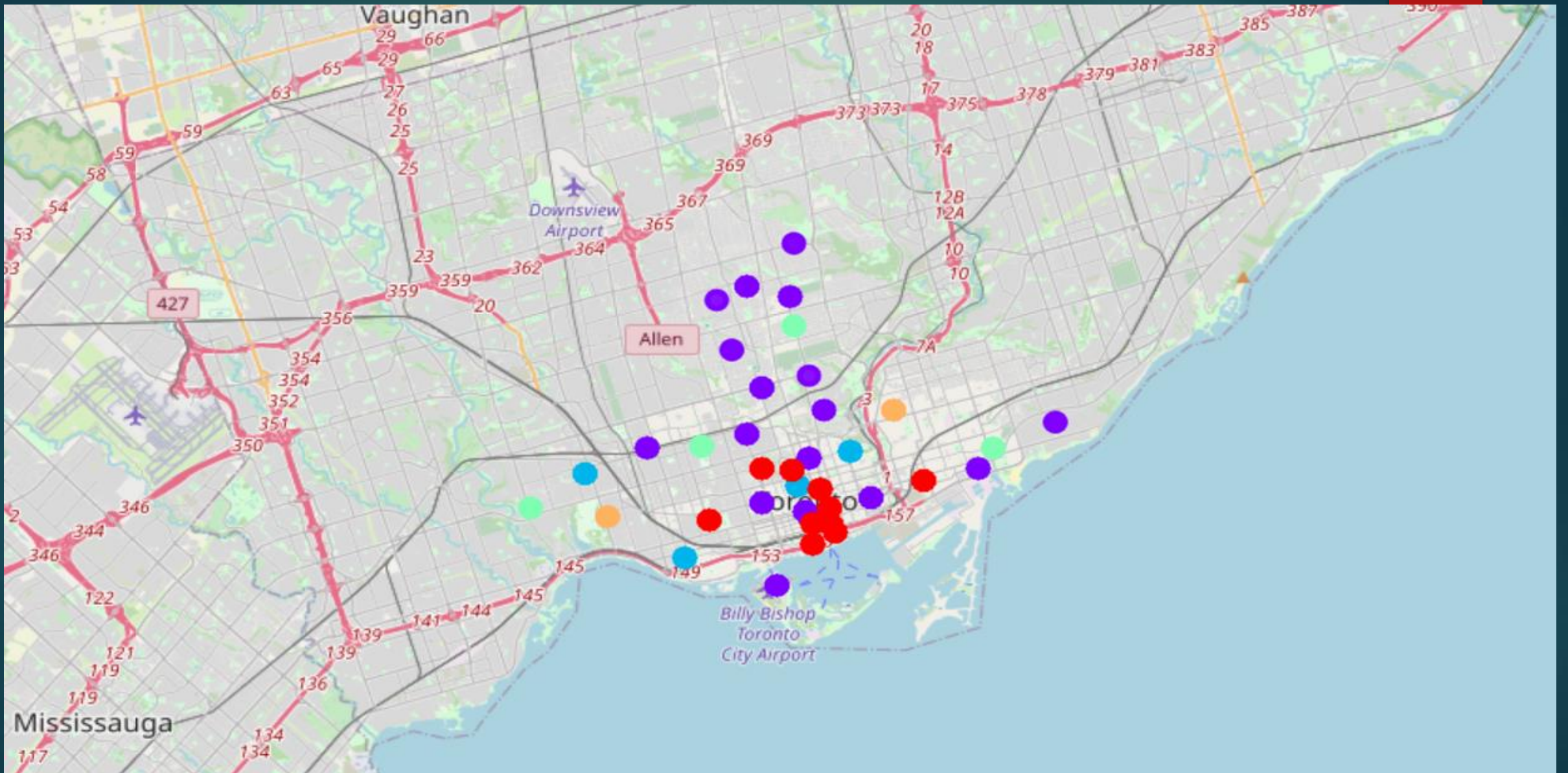
Now, the dataset is fully ready to be used for machine learning (and statistical analysis) purposes.

```
toronto_denc_venues.groupby('Neighbourhood').count()
```

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighbourhood						
Berczy Park	57	57	57	57	57	57
Brockton, Parkdale Village, Exhibition Place	22	22	22	22	22	22
Business reply mail Processing Centre, South Central Letter Processing Plant Toronto	15	15	15	15	15	15
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	15	15	15	15	15	15
Central Bay Street	64	64	64	64	64	64
Christie	17	17	17	17	17	17
Church and Wellesley	76	76	76	76	76	76
Commerce Court, Victoria Hotel	100	100	100	100	100	100
Davisville	33	33	33	33	33	33
Davisville North	10	10	10	10	10	10

Part 4: Applying one of Machine Learning Techniques (K-Means Clustering)

Now, we focus on the centers of clusters and compare them for their "Total Restaurants" and their "Total Joints". The group which its center has the highest "Total Sum" will be our best recommendation to the contractor. {Note: Total Sum = Total Restaurants + Total Joints.} This algorithm although is pretty straightforward yet is strongly powerful. Run k-means to cluster the neighborhoods in Toronto into 5 clusters and then we will plot the map with the five different colour clusters, here this dark blue cluster is cluster 0 ,orange cluster is cluster 1 then red colour cluster is cluster 2, sky blue colour cluster is cluster 3 and light green cluster is cluster 4.



OBSERVATIONS

15

Most of Italian restaurants are in Cluster 0 and lowest in Cluster 4 and cluster 2 areas. Also, there are good opportunities to open restrurant in parkside drive and bayview avenue bloor street areas as the competition seems to be low. Looking at nearby venues, it seems Cluster 4 might be a good location as there are not a lot of italian restaurants in these areas. Therefore, this project recommends the entrepreneur to open an authentic Burmese restaurant in these locations with little to no competition. Nonetheless, if the food is authentic, affordable and good taste, I am confident that it will have great following everywhere.

Thank You!!!!