Searching the Optimal Location for a Restaurant in Toronto using Data Science

Sadia Tangim Promi

Introduction

- Most important question when opening a new business?
  - Where should I open shop?
- Data Science is here to help!

Factors that might help answer the question?

- Historical trends in the geographical area
- Existing Businesses
- Places or Venues around the area etc.

Introduction

- Foursquare API:
  - Crowd-sourced Location Data Service
  - Rich archive of location and venue data

This data can be utilized to answer the following questions:

- What are the hot locations for the targeted business?
- What is the nature of existing businesses around their area?
- What would be a great location to start the business?
- If he/she has a location of choice, what is the likelihood that his business might thrive in that location?

Data

- Neighborhood data for the city of Toronto:
  - Collected from Wikipedia
    - Postal Code
    - Borough
    - Neighborhood
    - Latitude
    - Longitude

	Postal code	Borough	Neighborhood	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

Data

- Venue Data from Foursquare API:
  - Venue name
  - Neighborhood
  - Venue Location
  - Venue Type/Category

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

Methodology

- Step 1: Collect Neighborhood Data for the city Toronto:
  - Scraped from Wikipedia
- Step 2: Collect Venue Data from Foursquare API:
  - Use neighborhood data to call Foursquare API and get information on venues.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
1	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
2	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
3	Victoria Village	43.725882	-79.315572	Tim Hortons	43.725517	-79.313103	Coffee Shop
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

Step 3: Data Analysis

Coffee Shop	185	Clothing Store	36
Café	100	Gym	34
Restaurant	66	Bar	32
Park	50	Fast Food Restaurant	29
Pizza Place	48	Sushi Restaurant	29
Italian Restaurant	47	American Restaurant	29
Hotel	43	Pub	27
Japanese Restaurant	42	Bank	26
Sandwich Place	41	Grocery Store	25
Bakery	38	Breakfast Spot	24

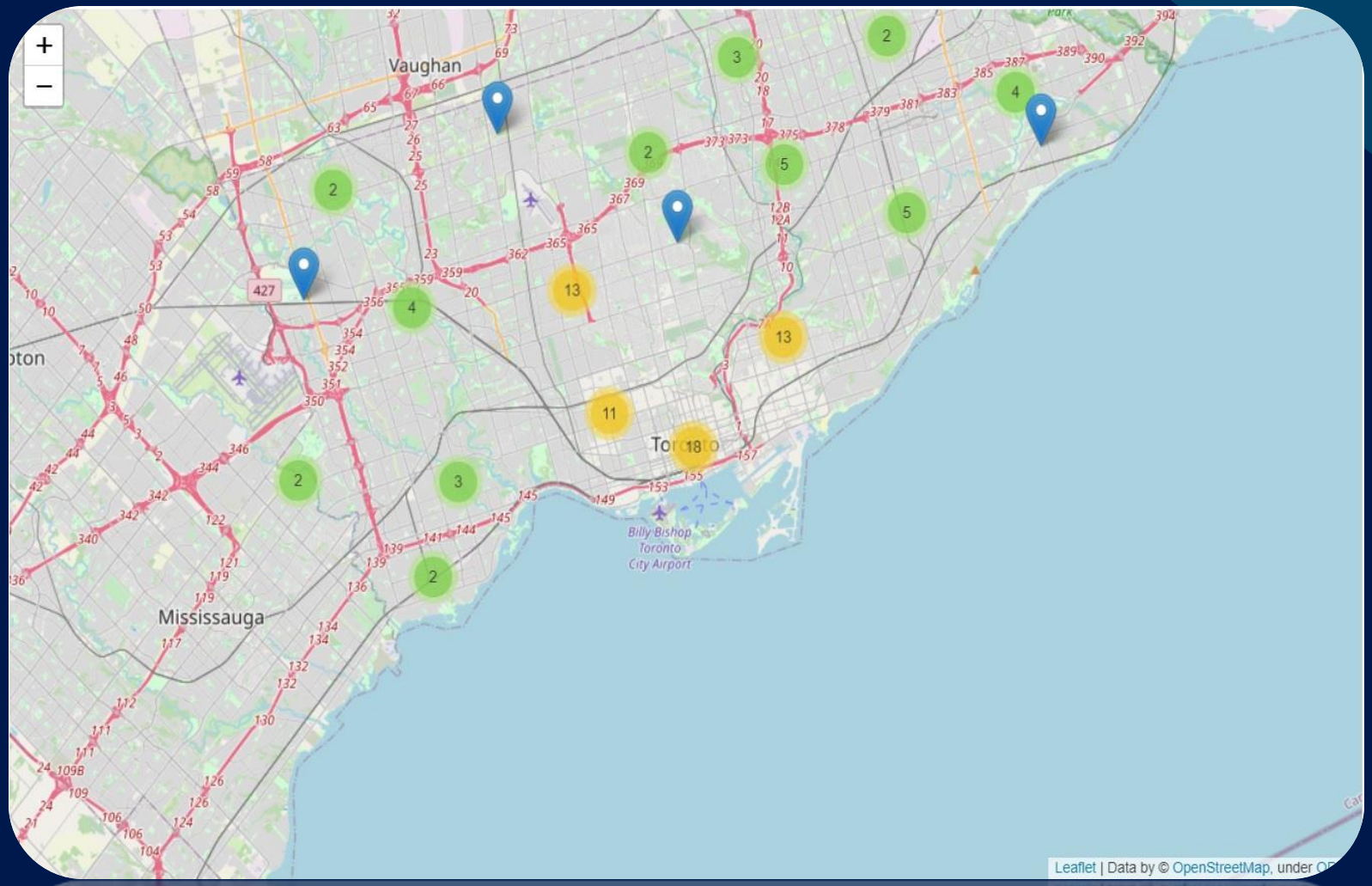
Top Venues

Step 3: Data Analysis

	Neighborhood	Latitude	Longitude	Restaurant
43	Commerce Court, Victoria Hotel	43.648198	-79.379817	30
87	First Canadian Place, Underground city	43.648429	-79.382280	30
89	Church and Wellesley	43.665860	-79.383160	26
39	Toronto Dominion Centre, Design Exchange	43.647177	-79.381576	26
27	Richmond, Adelaide, King	43.650571	-79.384568	24
8	Garden District, Ryerson	43.657162	-79.378937	24
83	Stn A PO Boxes	43.646435	-79.374846	21
12	St. James Town	43.651494	-79.375418	21
21	Central Bay Street	43.657952	-79.387383	19
34	Little Portugal, Trinity	43.647927	-79.419750	16

Neighborhoods with the most Restaurants

Step 4: Visualize Hot-spots



Visualization of Restaurant Hotspots

Step 5: Data Preparation for Machine Learning

- One-hot Encoding
- Group By ‘Neighborhood’ sums
- Select top 5 Features from each neighborhood
- Exclude ‘Restaurant’ Data

	Yoga Studio	Accessories Store	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0

5 rows × 267 columns

```
['Neighborhood',
 'Accessories Store',
 'Airport Lounge',
 'Airport Service',
 'Airport Terminal',
 'Aquarium',
 'Athletics & Sports',
 'Auto Garage',
 'Auto Workshop',
 'Bakery']
```

Step 6: Create Machine Learning Classifier to predict likelihood of Restaurant

- Calculate correlations
- Take only positively correlated features with correlation value > 0.20

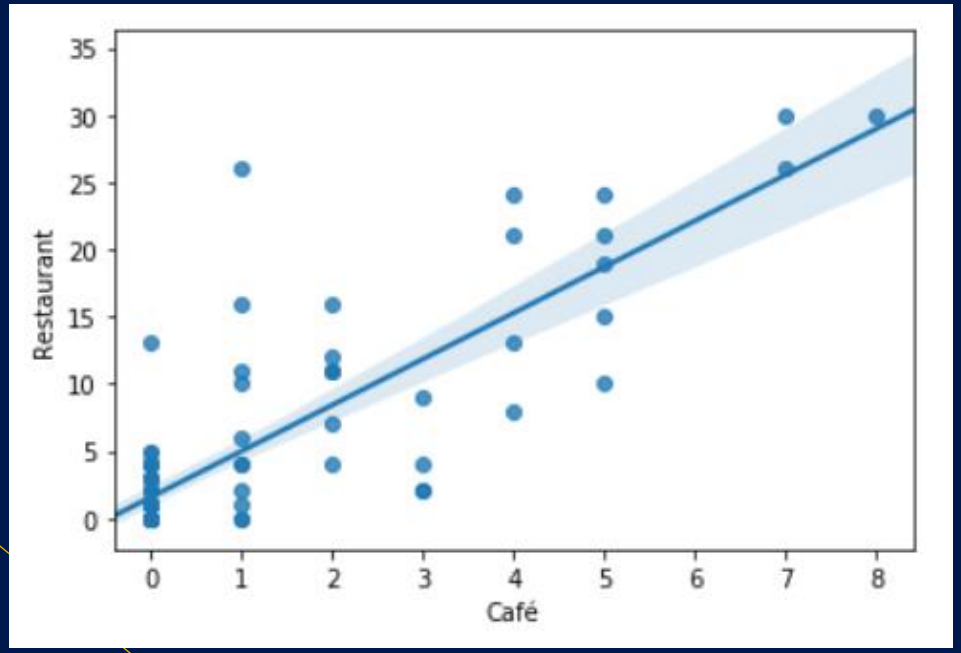
	index	Restaurant
110	Restaurant	1.000000
25	Café	0.832903
31	Coffee Shop	0.830975
67	Hotel	0.743790
54	Gastropub	0.719778
59	Gym	0.680451
17	Bookstore	0.676022
38	Deli / Bodega	0.667530
15	Beer Bar	0.654870
73	Lounge	0.572245

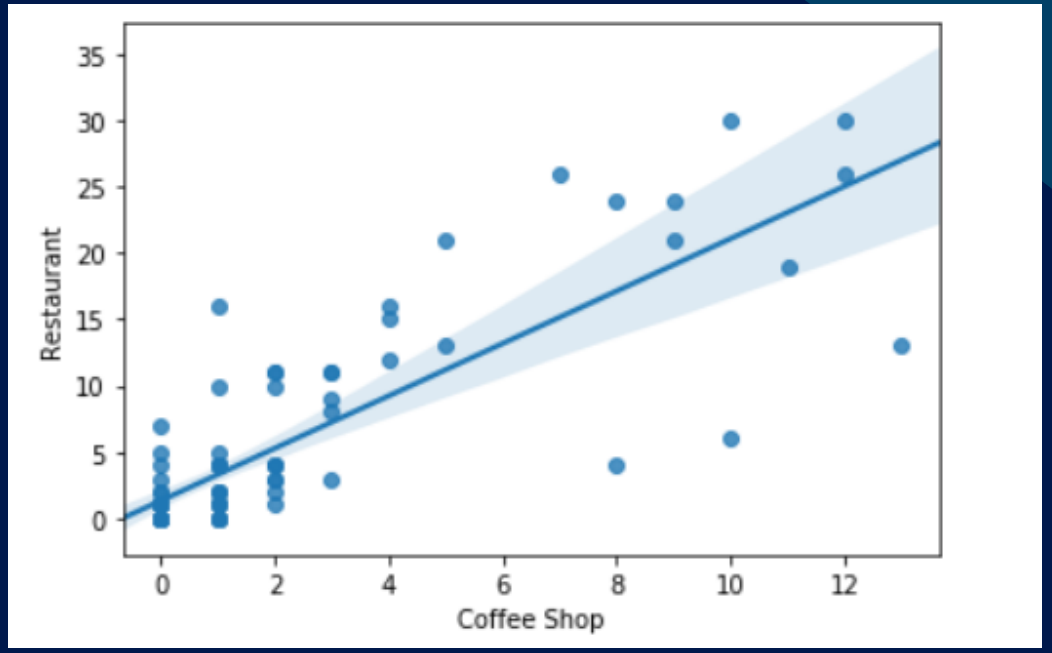
Positively Correlated

	index	Restaurant
0	Latitude	-0.352364
24	Bus Line	-0.145087
69	Intersection	-0.140361
34	Construction & Landscaping	-0.125458
7	Athletics & Sports	-0.125264
94	Rental Car Location	-0.110904
90	Pool	-0.109486
77	Metro Station	-0.092145
66	Hockey Arena	-0.082418
2	Accessories Store	-0.082418

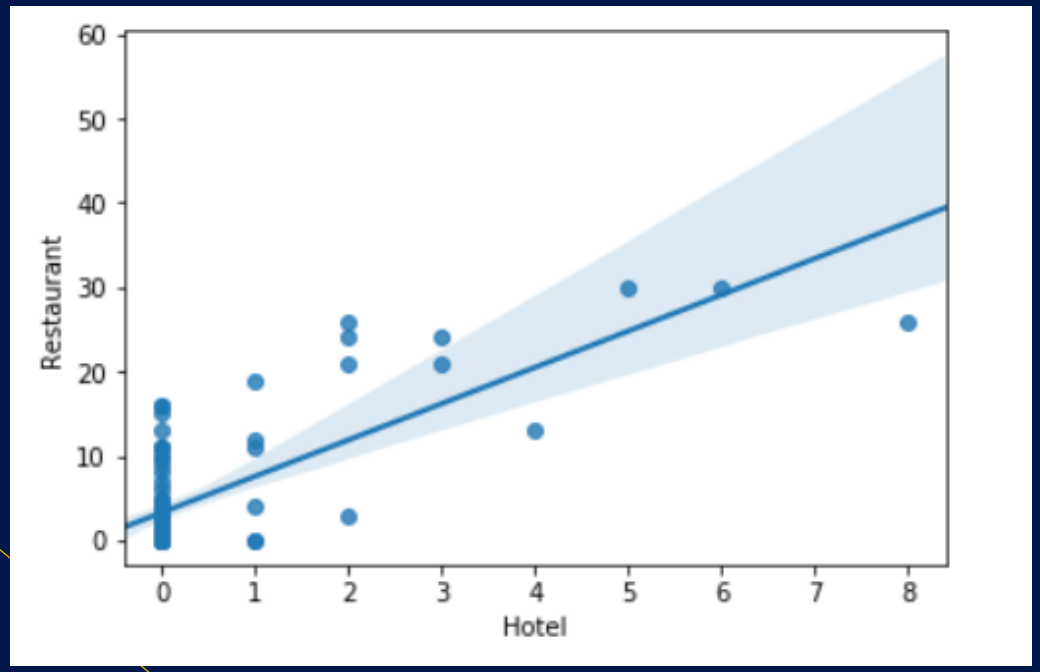
Negatively Correlated

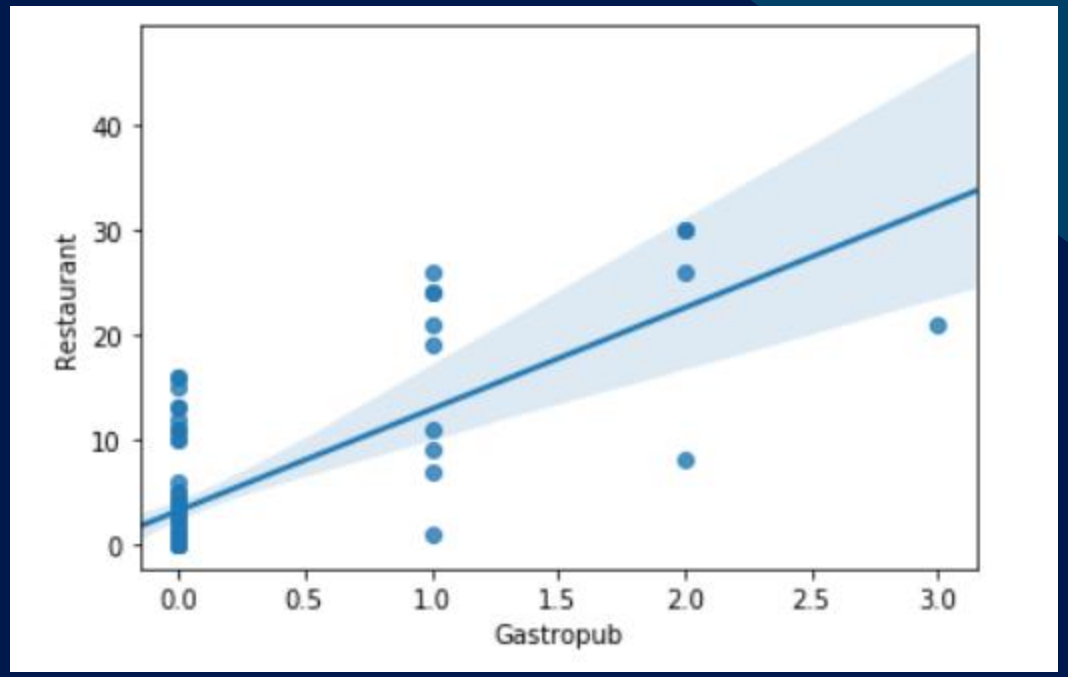
Regression Plots





Regression Plots





Step 6: Create Machine Learning Classifier to predict likelihood of Restaurant

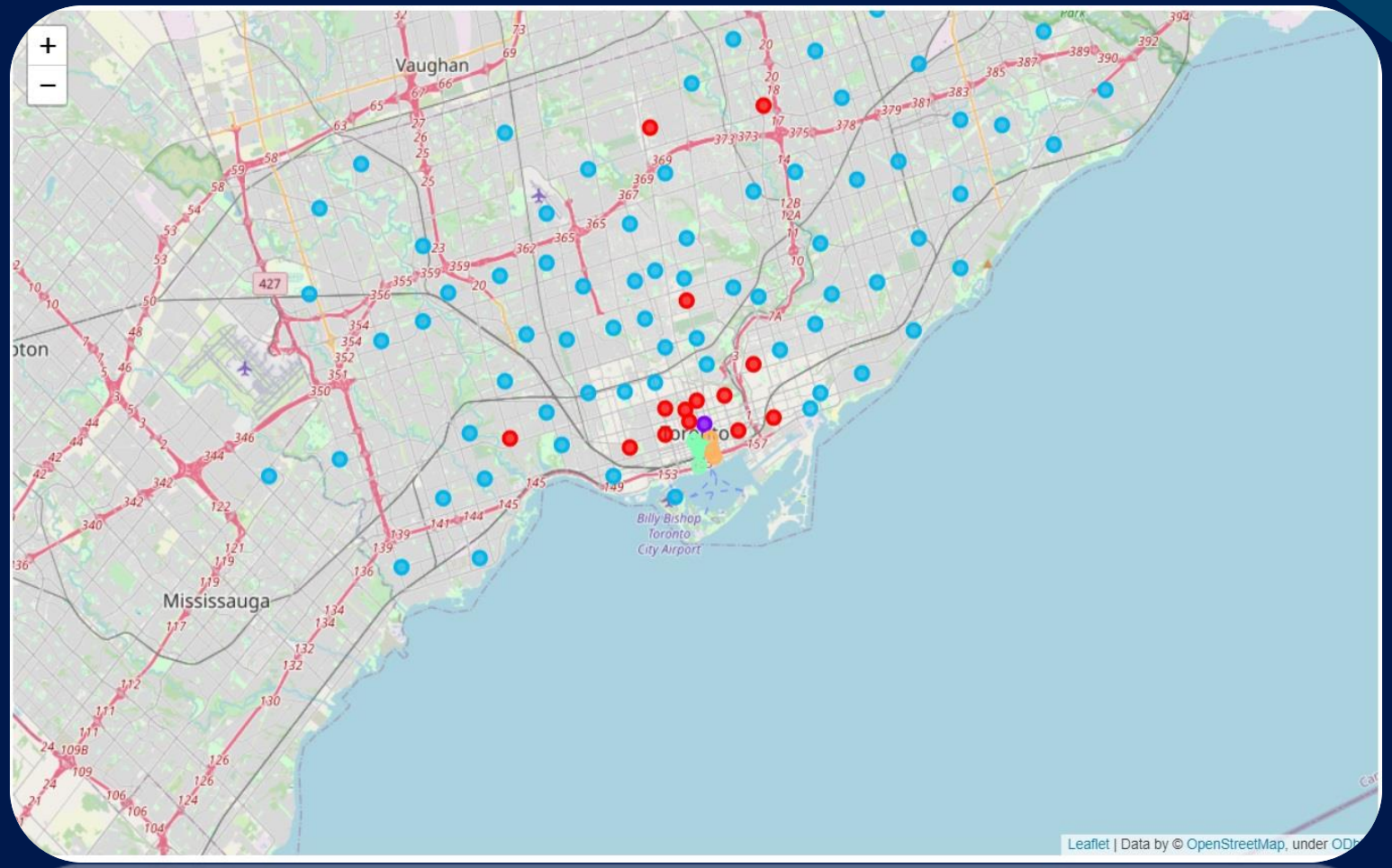
- Normalize data using Standard Scaler
- Train-test split ratio of 0.2
- Four Classifiers:
  - Logistic Regression
  - K Nearest Neighbors
  - Decision Tree Classifiers
  - Support Vector Machines

Classifier Results

Model	Training Score	Testing Score		
Table 1: Classifier Results		Jaccard Index	Log Loss	F1-Score
Logistic Regression	0.8378	0.8421	5.4534	0.88
kNN	0.7837	0.6842	10.9070	0.75
Decision Tree	0.8648	0.7894	7.2713	0.8333
SVM	0.8108	0.7368	9.0891	0.8

Best Classifier: Logistic Regression

Step 7: Clustering – Kmeans Clustering



Step 7: Clustering – Kmeans Clustering

- The data frame was grouped by the clusters and then taking the mean of the column containing **Restaurant Count** provided interesting results.
- To further explore the matter, the value counts for each cluster were calculated.
- From these results, it can be identified that **Cluster 2,** while having a lion's share of the Neighborhoods, contains very few restaurants. We can infer that Neighborhoods around this cluster are less likely to host a restaurant.

```
Cluster
0    12.500000
1    24.000000
2     1.585714
3    24.600000
4    18.000000
Name: RestCnt, dtype: float64
```

```
df['Cluster'].value_counts()
```

```
2    70
0    14
3     5
4     3
1     1
Name: Cluster, dtype: int64
```

Results and Discussion

- Results have already been discussed in previous slides
- We can conclude that there are certain locations which are more likely to be successful with a restaurant
- There are certain locations which should be avoided
- The Logistic Regression Classifier can give the probability whether or not a restaurant is appropriate at a location

Conclusion

- Data Science and Modern tools can help decide on the location
- These tools can answer questions and deliver insights to the stakeholders.
- Data from more sources shall make the pipeline more robust.

Thank You.