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

Sadia Tangim Promi

Contents

List of Figures

Introduction

Setting up shop for a new business in a new city invites a lot of challenges. One of the most important challenges a business owner or entrepreneur might face while setting up shop, is to decide on the location of the business. Depending on the choice of location, a business might thrive in a particular location, or even fail.

The most important question when opening a new outlet or setting up a new shop is whether the location is suitable for that particular business or not. Historical trends in that area, existing businesses, places or venues around that area - all are the factors which one way or another might influence the success or failure of the business. So, it is important to search for patterns in this data which might provide us valuable insights.

Foursquare is a perfect candidate for providing us with such data. Foursquare is basically a location data service which crowdsources information like shops, restaurants, businesses, different places and venues etc. with respect to their geographic location and maintains a rich database. This data is easily accessible through the Foursquare API which we can utilize to solve the problem at hand.

Business owners who want to expand their business or even an entrepreneur deciding to start a business in a particular area might benefit from this project. The aim of this project is to primarily answer the following questions which might help in their endeavor –

- 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?

Getting meaningful answers to these questions would go a long way in helping the stakeholder decide the location of their next successful business venture.

Data

Say, for our problem, a business owner wants to open a new Restaurant in the city of Toronto.

In order to decide on a suitable location for a particular business, might take help from a variety of data sources. Historical trends in that area, existing businesses, places or venues around that area - might provide us valuable insights to answer our questions.

We use two data sources for this task. Firstly, we need to have a list of probable neighborhoods around the city of Toronto. In order to obtain a list of neighborhoods around the city of Toronto, we scraped a Wikipedia Table to collect information i.e. Names of different neighborhoods in Toronto, their Postal Codes and Co-ordinates (Latitudes & Longitudes etc.

An example of a couple of records from the scraped table is as follows:

	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

Figure 1: Neighborhood Data

The second, and most important source of data is the Foursquare API. We shall utilize the neighborhood-wise information obtained from the Wikipedia table to make calls to the Foursquare API to collect information about the most popular venues in those locations. Foursquare also categorizes its venues into some categories which might be helpful for our task. We utilize the foursquare API to extract the following data fields for each venue –

- Venue Name
- Neighborhood
- Venue Location (Latitude & Longitude)
- Venue Type/Category

An example of the information obtained from the Foursquare API, after some preprocessing is shown as follows:

	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

Figure 2 : Venue Data

Methodology:

The methodology of the project can be discussed in a couple of different steps which are as follows:

Step 1: Collect Neighborhood Data for the City Toronto

The first step in the methodology is to collect the data of different Neighborhoods for the city of Toronto. In order to do this, we used python library **pandas** to scrape a Wikipedia page. This gave us a table containing the different neighborhoods.

	Postal code	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Figure 3 : Postal Code data from Wikipedia

In conjunction with a **csv** file which contained the corresponding co-ordinates for the different postal codes, the final Neighborhood data frame named **to_data** was constructed which had the following structure:

	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
5	M9A	Etobicoke	Islington Avenue	43.667856	-79.532242
6	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
7	M3B	North York	Don Mills	43.745906	-79.352188
8	M4B	East York	Parkview Hill, Woodbine Gardens	43.706397	-79.309937
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937

Figure 4: Neighborhood Data with Location

Step 2: Collecting Venue Data from Foursquare API

This step involved using **Foursquare API** to collect data from different neighborhoods. The data frame from the first step provided us with a number of neighborhood names and locations. These were used to make calls to the Foursquare API. The foursquare API returned the queries in **json** format, which were parsed and stored into a data frame called **to_venues**. The queries to the Foursquare API resulted in 2150 unique responses i.e. venues. Each venue had seven distinct attributes: **Neighborhood, Neighborhood Latitude, Neighborhood Longitude, Venue, Venue Latitude, Venue Longitude** & **Venue Category.** The resulting data frame head is as follows:

	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

Figure 5: Venue Data from Foursquare

Step 3: Data Analysis

Some basic exploratory data analysis was done in this step and the following. Exploring the data from **to_venues** we could see that there is a total of **266** unique categories of venues. The **value counts()** function was used on this to return us the venues that occurred the most. Following is a list of the top 20 most frequent venue categories:

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

Figure 6: Top Venues

We can see from the illustration that the **top 5** venues occurring frequently are: **Coffee Shop, Café, Restaurant, Park & Pizza Place.**

Since, our target is to gain insights about the restaurants, we searched for venues that were essentially restaurants. This search showed us that there were **486** restaurants of different cuisines in the data frame.

Next, this data frame was grouped by **Neighborhood** to obtain the total count of different restaurants within each neighborhood. For example –

	Neighborhood	Restaurant
0	Agincourt	1
1	Alderwood, Long Branch	0
2	Bathurst Manor, Wilson Heights, Downsview North	3
3	Bayview Village	2
4	Bedford Park, Lawrence Manor East	11

Figure 7: Restaurant Counts

The results were stored in a data frame called **rest_counts**, sorting which we can get a list of neighborhoods with the most Restaurants:

	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

Figure 8: Neighborhoods with the most Restaurants

So, we can conclude that **Commerce Court, Victoria Hotel, First Canadian Place, Underground City, Church and Wellesley** etc. are the neighborhoods which contain the greatest number of restaurants. These are essentially hot-spots for restaurants.

Step 4: Plot Restaurants in a map to Visualize Hotspots

The python library **folium** was utilized to generate a visualization of the different locations which contained the greatest number of neighborhoods. The visualization is as follows:

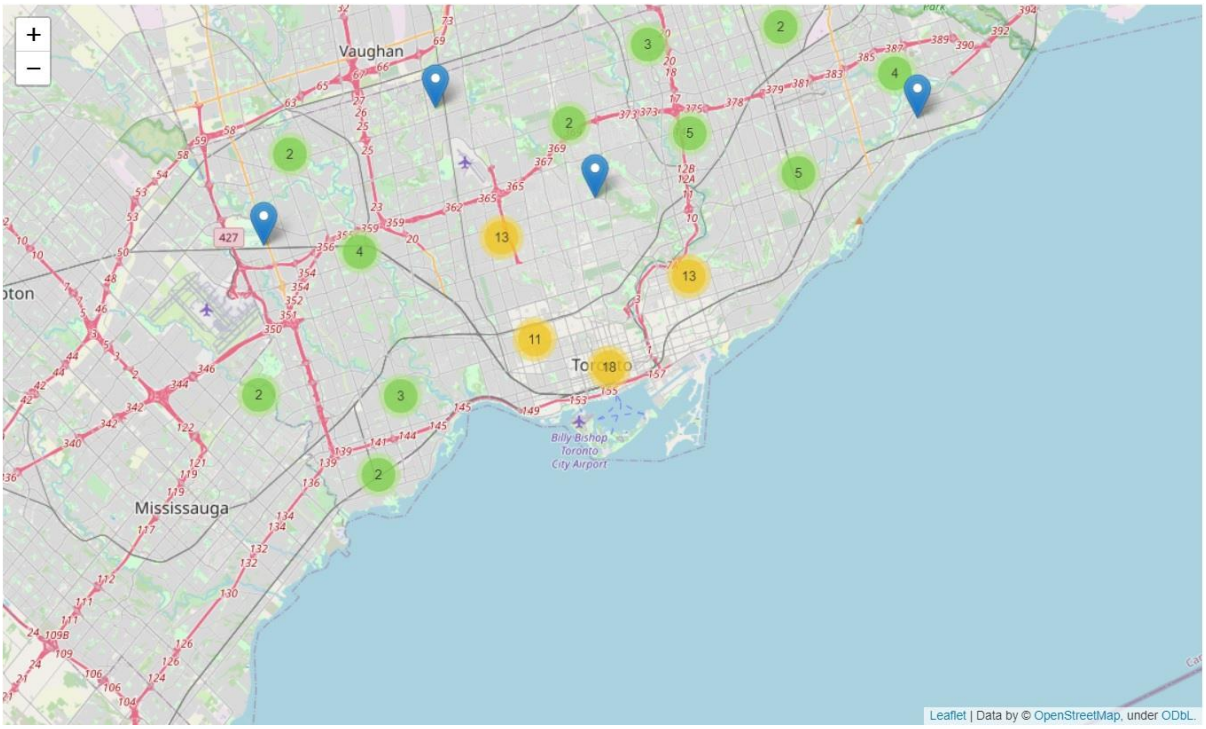

Figure 9: Visualization of Restaurant Hostposts

As we can see from the map, the yellow marked regions contain the greatest number of restaurants and are essentially hotspots for restaurants.

Step 5: Data Preparation for Machine Learning

The data frame was encoded to one-hot encoding to contain 267 feature columns, one for each category:

	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

Figure 10: One-Hot Encoded Data Frame

This data frame was then grouped by **Neighborhood** and the values of the categories summed. The **top 5** categories for each Neighborhood were selected to create a feature list. The restaurants were later filtered out of this list since we want to train the classifier with restaurants as target data. After discarding duplicates and data from the restaurants, 108 features remained. Here is a couple of features from the feature list:

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

Figure 11: Features in the Feature List

Step 6: Create a Machine Learning Classifier to Predict the likelihood of a Restaurant

Since 108 features seemed too much, it was decided to minimize the feature set so that it would only contain important and meaningful features. **Correlation** was taken from the

data frame features with respect to the target column **'Restaurants'**. Here are a few positively correlated features:

	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

Figure 12: Positively Correlated Features

While there were plenty of significantly positively correlated features, the negatively correlated features were insignificant:

	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

Figure 13: Negatively Correlated Features

So, it was decided that the features with a correlation value greater than **0.20** with respect to the **'Restaurants'** column would be selected to train the machine learning classifiers. This resulted in a reduced feature set of 37 features.

As we can see, the most positively correlated features are: **Café, Coffee Shop, Hotel, Gastropub**, **Gym** etc. We created regression plots for these features to verify their correlation with the number of restaurants.

Regression plot for **Café:**

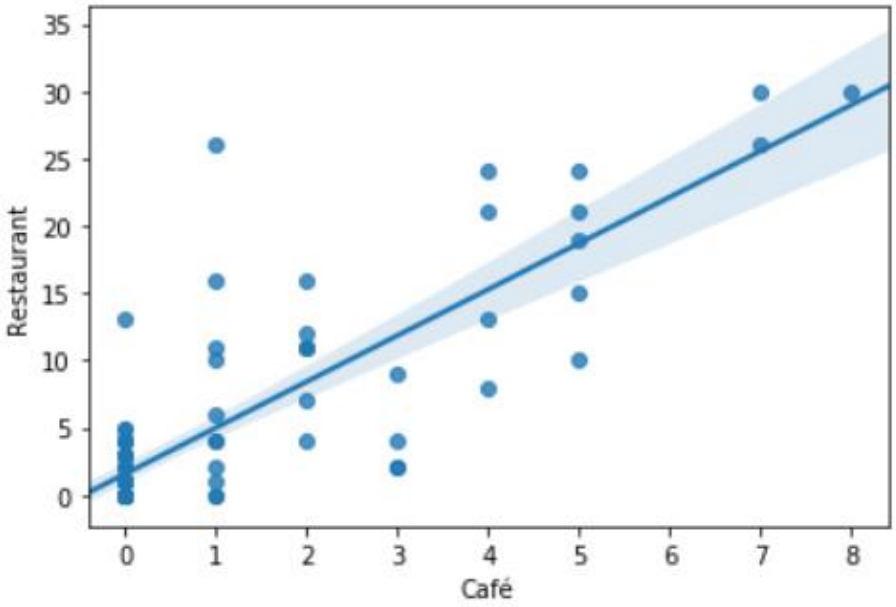

Figure 14: Reression Plot for Cafe

Regression plot for **Coffee Shop:**

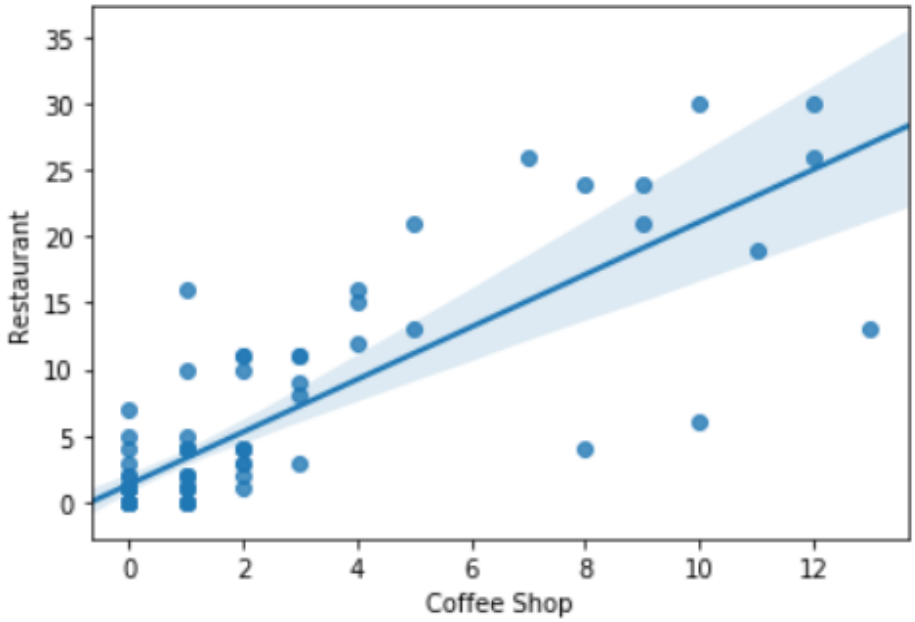

Figure 15: Regression Plot for Coffee SHop

Regression plot for **Hotel:**

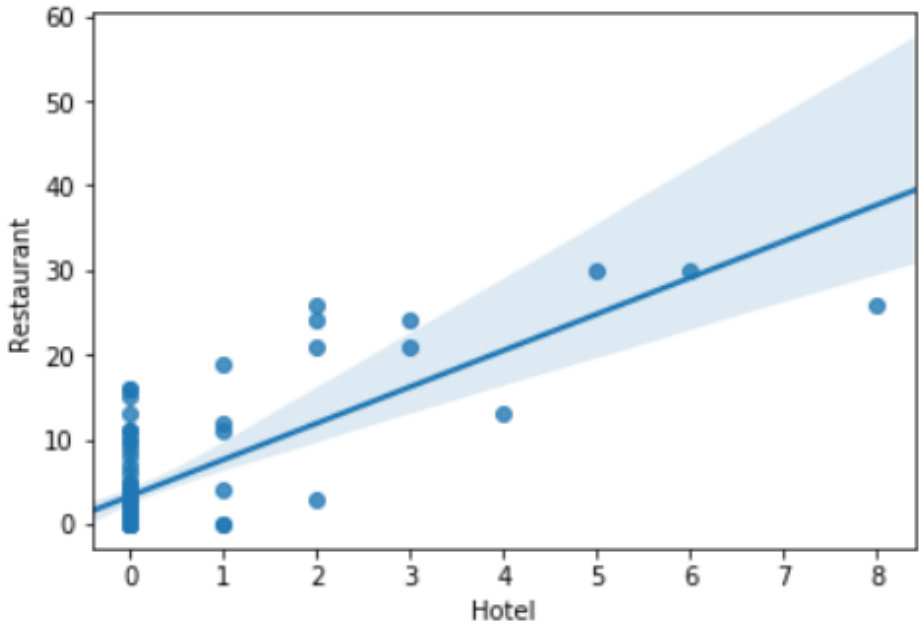

Figure 16: Regression plot for Hotel

Regression plot for **Gastropub:**

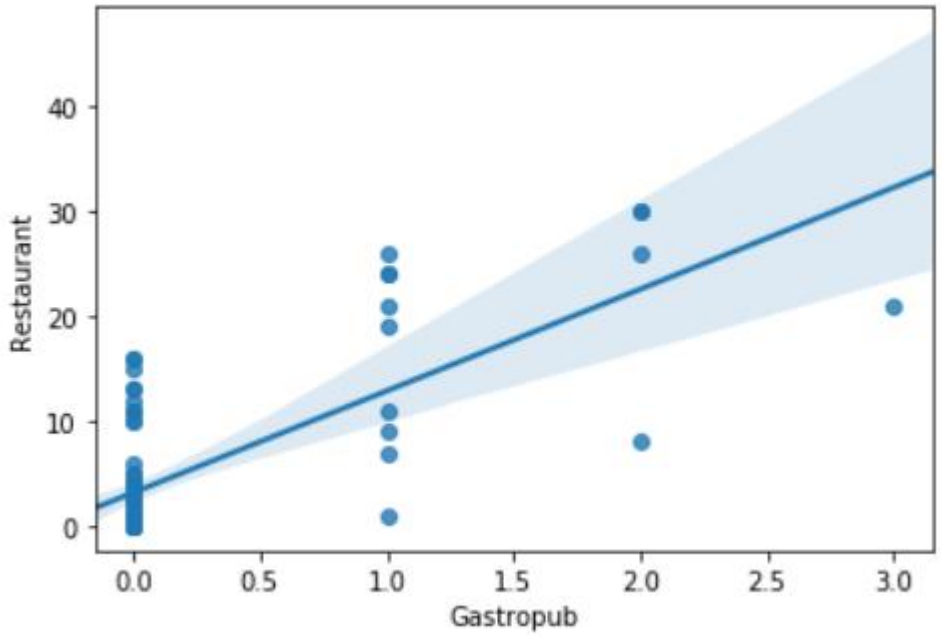

Figure 17: Regression Plot for Gastropub

Regression plot for **Gym:**

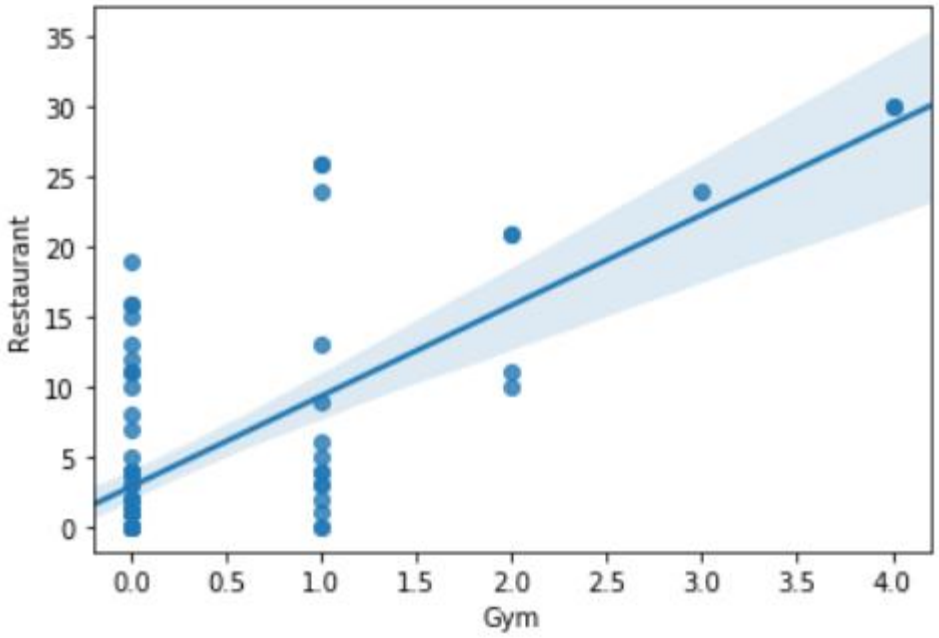

Figure 18: Regression Plot for Gym

Before passing the data to the model, the data was normalized using **StandardScaler** and the **'Restaurant'** column was converted into a binary class to approach the problem as a binary classification task.

After normalization, the data was split into a training set, and a testing set using a **train-test split** ratio of **0.2**. Which means, 80% of the data was used for training purpose and the rest 20% for testing purpose.

Four machine-learning classifiers – **Logistic Regression, k Nearest Neighbors Classifier, Decision Tree Classifier** & **Support Vector Machine Classifier** were used to tackle the binary classification problem. Of all these models, the **Logistic Regression Classifier** performed the best, with a training score of 83.78% and a test score of **84.21%**. The Logistic Regression classifier should be a good choice here, since it is able to output a probability/likelihood of a particular location as it outputs sigmoid probability distributions.

A more detailed dive into the classifiers performance and implications might be found in the **Results** and the **Discussion** sections that follow.

Step 7: Clustering

The k-Means clustering algorithm might be applied to the data when the data is dealt in an unstructured-unsupervised fashion. In order to apply k-Means, the similar preprocessing of the previous step is followed. The only difference is that we do not ignore the features coming from the different **Restaurant** categories, which increases the training feature size to **60,** up from 37.

The k-Means algorithm is run with a target cluster size of 5. The clustered data is visualized in a geospatial plot using the python library **folium.** The geospatial plot is as follows:

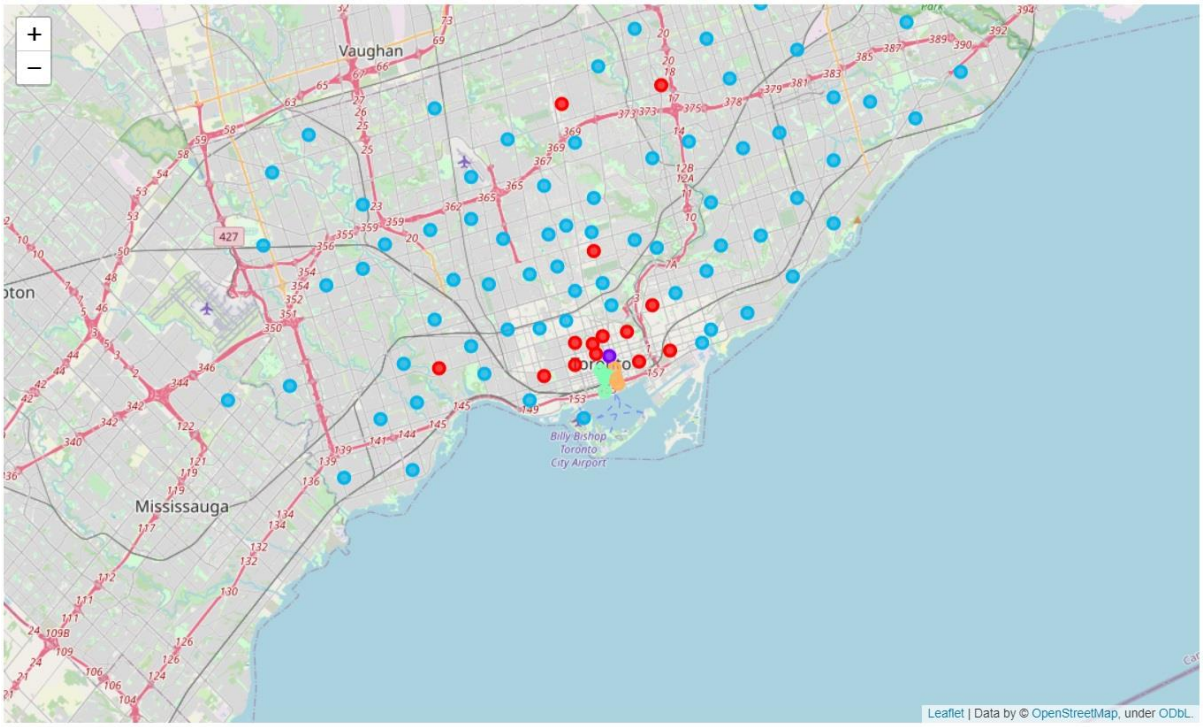

Figure 19: Clustering Results

The data frame was grouped by the clusters and then taking the mean of the column containing **Restaurant Count** provided interesting results:

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

Figure 20: Restaurant Count Means by Clusters

To further explore the matter, the value counts for each cluster were calculated:

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

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

Figure 21: Value Counts by Clusters

Which shows that the **Cluster 2** is the cluster with the greatest number of Neighborhoods, however, these neighborhoods have fewer restaurants.

Results:

We can summarize the findings of our data science project in a few key areas. These are discussed as follows:

Neighborhoods with the most restaurants:

We obtain a summary of the Neighborhoods with the most Restaurants. The summary of these neighborhoods can be found in **Figure 8.** We can see that, **Commerce Court, Victoria Hotel, First Canadian Place, Underground City, Church and Wellesley** are the locations where a restaurant is more likely to thrive. However, competition should also be higher in these areas. From the geospatial plot in **Figure 9,** we also get a similar indication of the regions of interest for Restaurants.

The top features that correlate with the number of restaurants:

The top features that positively correlate with the number of restaurants are as follows:

- Café
- Coffee Shop
- Hotel
- Gastropub
- Gym
- Bookstore
- Deli/Bodega
- Beer Bar
- Lounge etc.

Figure ** Shows a more detailed view of the features that correlate with the number of restaurants.

Binary Classifier Results:

Four different classifiers – **Logistic Regression, k Nearest Neighbors, Decision Tree Classifier** and **Support Vector Machine Classifier** were tried out to classify the data into a decision of whether or not a neighborhood might contain a restaurant. A summary of the results from the different models are tabulated in **Table 1**.

Table 1: Classifier Results

Model	Training Score	Testing Score		
		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

As we can see from the table, **Logistic Regression** is the best performing model.

Clustering Results:

We can find the results of the clustering shown in a geospatial plot in **Figure 19. Figure 20 and 21** Respectively show the mean value of Number of restaurants across the classes and also the number of Neighborhoods in each class. 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.

Discussion:

This project primarily focused on exploring an optimal location for starting a restaurant in the city of Toronto. Different steps like analysis, visualization and modelling have been performed in order to approach the problem and reach an optimal solution. The steps in Analysis have suggested us a number of locations or hot-spots where restaurant businesses might thrive. At the same time, correlations and co-existence of businesses of different nature have been studied at length. This provides us a data driven insight to come to certain statistical hypotheses that which surrounding conditions might have an impact on the feasibility of opening a restaurant in a particular area.

Visualizing the data in geospatial plots give us a better understanding as to Geographically speaking, in which locations might be suitable for setting up a new restaurant. The clustering approach identifies potential no-zones which are most likely residential areas and therefore not suitable for opening up a restaurant.

Finally, the machine learning classifier we designed is capable of providing a probability score between 0~1 of whether or not it is feasible to open a restaurant in a particular area based on a number of features. This enables us to be able to tell with the help of data science whether or not a particular location might be suitable for opening a restaurant or not.

Conclusion:

The resolution of deciding on the perfect location of your next business is not an easy task. However, as shown in this project, with the power of data science and modern tools, it is not impossible to come to a meaningful conclusion. Our project shows the potential locations and neighborhoods, around which a new business could thrive, and particular locations, which one would be better off without. The system is also capable of answering questions to the stakeholders by statistical inference, visual and analytical tools, and help them decide on the future of their endeavors.