|  |  |
| --- | --- |
| A large building  Description automatically generated  The Battle of Neighborhoods: Opening a Restaurant in a New City  Segmenting neighborhoods in Toronto to find the most conducive locations for starting a restaurant using Data Science. | Applied Data Science Capstone Project, by IBM.  Usman Aftab Khan |

Introduction

##### I. Background | Business Problem

# 

***Mirch Masala*** (fictitious) is a restaurant renowned for bringing the mouth-watering taste of Pakistani and Indian cuisine to your tables. Among the wide range of courses on the menu, their specialty lies in Biryani, Karahi, and Chicken Tikka.

from left: Chicken Tikka, Chicken Biryani, and Chicken Karahi

Their first outing in New York City was an enormous hit. New York City is recognized as an epitome of cultural diversity, due to the immensely large population of immigrants from all over the globe. This allows it to be one of the topmost cities among other metropolitan and cosmopolitan hubs. Another city, of similar characteristics and stature, lies across the border — the City of Toronto, Canada. Home to a dense population of **6.2 million**, Toronto is Canada’s largest city and shares similar multicultural traits as New York City. And so, **the restaurant has set its eyes on expanding its business across the border in the city of Toronto**.

Our project’s objective is to figure out conducive locations in the city that are ideal for opening a restaurant. To ensure the success of our project**, the team requires insight into the demographics and the neighboring businesses**. For instance:

* the number of restaurants present in each neighborhood,
* the most popular restaurants (both similar and different cuisines),
* the traffic of our target demographic,
* and the frequency of our target audience.

This project gives **Data Scientists/Analysts** the opportunity to apply their knowledge of data science and categorically go through different processes. Defining the business problem, requirement elicitation, retrieving and utilizing data from external sources, parsing and cleaning the data, and analytical assessment through Machine Learning algorithms and tools. The evaluation from the final analysis leads to a conclusion which can then be leveraged by stakeholders. As this project has a lot of aspects to be considered, **it is open for discussion** and targeted towards the **entrepreneurs** and **stakeholders.**

##### II. Description of Data

The data required will be a combination of **CSV files** that have been prepared for the purpose of the analysis.

* **1st set of Data:**  
  Recent most updated record of traffic signal — vehicle and pedestrian volumes in Toronto. This data is typically collected between 7:30 a.m. and 6:00 p.m. at intersections where there are traffic signals.
* **2nd set of Data:**   
  The list of neighborhoods in Toronto represented by postal codes and their boroughs. We will be using the **Geocoder** Python package to retrieve the postal code’s coordinates.
* **3rd set of Data:**   
  The most common venues of a given neighborhood in Toronto. This information is stored inside **Foursquare Location Data**, and we will use **Foursquare API** to access it.

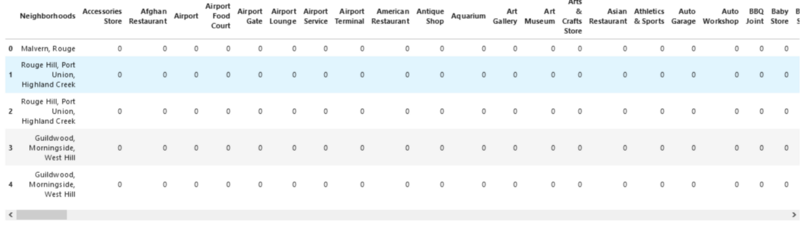
To recap, we will use the 1st set of Data to analyze the pedestrian/vehicle volume. Then, we load the 2nd set of Data to obtain the exact coordinates for each neighborhood based on their respective postal code, allowing us to explore and map the city. By using those coordinates and Foursquare credentials, we will access the 3rd set of Data sources through Foursquare API, and retrieve the popular venues along with their details, especially for restaurants (irrespective of their cuisine).

Methodology

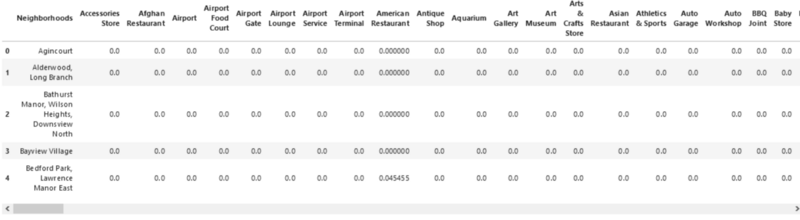
##### I. Analytical Approach

#### **One-Hot Encoding**

We begin our pursuit by approaching the problem using a technique called **one-hot encoding.** This technique can transform data from categorical form to numerical form for Machine Learning algorithms. Each venue was turned into a frequency at how many of those venues were in each candidate neighborhood.

A snippet of One-Hot encoding in action

Then, we grouped those rows by **Neighborhood** and by taking the **average** frequency of occurrence for each venue category.

A snippet of grouped neighborhoods by the avg of frequency occurrence of each venue

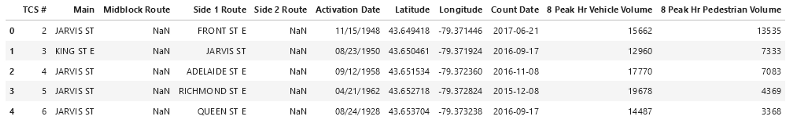
**K-Means Clustering**

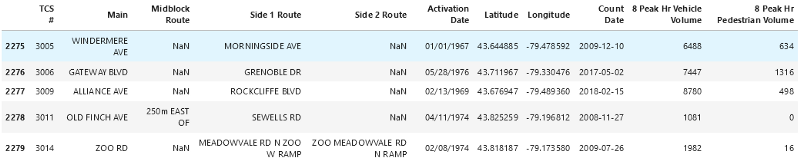
To see how the demographics exist for similar neighborhoods, and to make the analysis visually interesting, we use a clustering technique called **k-means**. **k-means** is a common machine learning algorithm used to cluster data points based on similar characteristics. The algorithm is fast and efficient for a medium and large-sized database and is useful to quickly discover insights from unlabeled data. By observing and examining each cluster, we can then determine those categories that distinguish them from one another.

#### II. Data Analysis

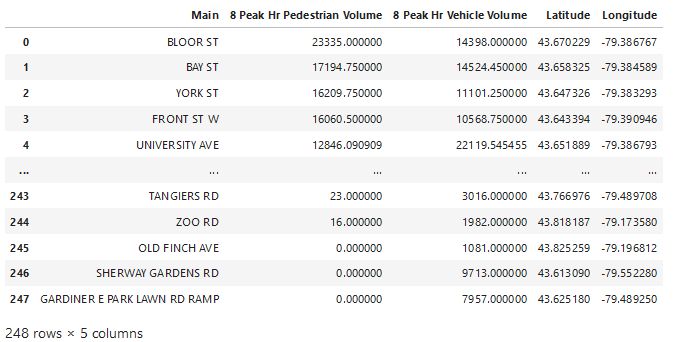
#### **Vehicle & Foot Traffic**

#### We begin by analyzing the data about the pedestrian and vehicle volumes.

A snippet of the first five pedestrian and vehicle volumes.

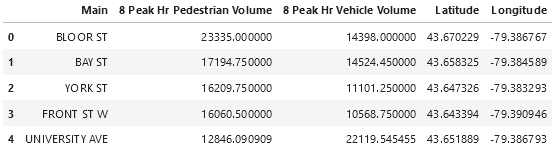
A snippet of the last five pedestrian and vehicle volumes.

The column **Main** comprises the main street name. This column has one distinctive quality — **the same name appears several times**, thus indicating it contains intersections. We can group by the street name and aggregate this either by summing those value up or averaging it. We will choose to average it for the sake of simplicity. This returns **248 main roads**.



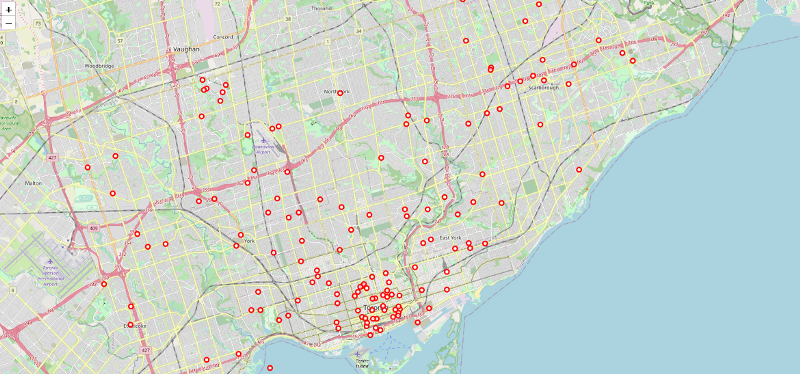
A snippet of pedestrian and vehicle volumes after being grouped by the main road.

We want our target neighborhood candidates to be active for the business’ longevity. Hence, we will filter out the main roads. In this example, we only show the roads with the average **pedestrian volume above 1200** or **vehicle volume above 12000** during peak hour **(approximately above 70%)**. This gives us exactly **139 main roads**.



A snippet of the first 5 rows after being filtered.

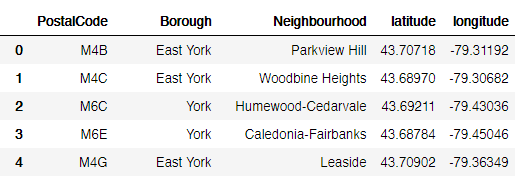
Finally, we can visualize the roads using the **Folium Python** module from the given coordinates. The map shows a glimpse of the busiest roads in the city, where many are located around downtown.



In the next section, we will explore the neighborhoods inside Central Toronto, East York, and York as the selected boroughs.

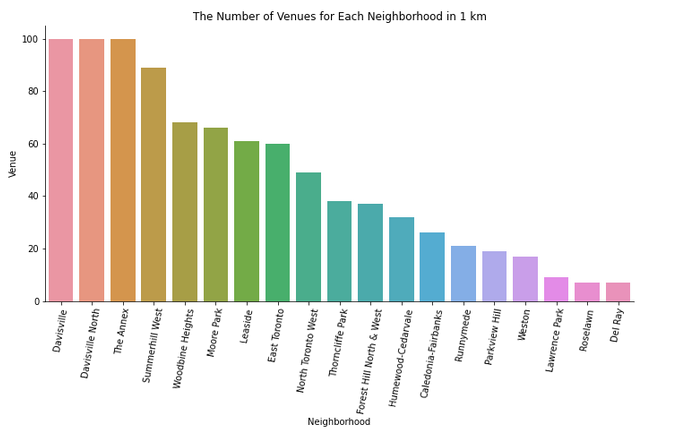
#### **Neighborhoods Analysis**

We have built a neighborhood data frame that comprises **103 postal codes**, **10 boroughs** with **neighborhood names** inside each borough, and **their coordinates**. We even used our One-Hot encoding technique to obtain the frequency of occurrences of the venues located in each neighborhood. Since we are specifically interested in neighborhoods inside **Central Toronto**, **East York**, and **York** only, we proceed by filtering the data frame resulting in **3 boroughs** and **19 neighborhoods**.



A snippet of the first five neighborhoods of the selected boroughs

As we have the coordinates information, we can then head over to Foursquare API and use it to access the data, analyze and explore the neighborhoods, and get **the top 100 venues** within a **radius of 1 km** for each. It returns a filtered result, with **905** **venues** with **172 unique venue categories**.

A snippet of the first five venues returned after filtering

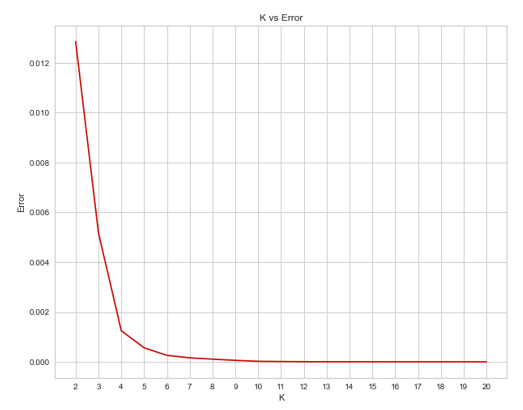
As evident from the graph above, many neighborhoods returned **above** 50 venues, such as Davisville and Davisville North with 100 venues each. However, many returned **below** 50 venues, such as Thorncliffe Park with 38 venues, and Parkview Hill with 19 venues. For each neighborhood, we can create the top 10 venues based on occurrences as follows.

A snippet of the first five rows of the neighborhood’s top 10 venues.

The data frame above is an indication that we have the same venue categories returned to different neighborhoods. We can use this idea to **cluster the neighborhoods based on their venues** representing services and amenities.

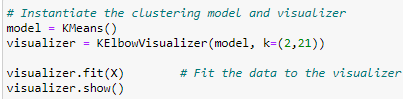
**Clustering the Neighborhoods**

We will run the **k-Means** algorithm to build a clustering model with a different number of clusters (**k**). The features will be **the mean of the frequency of occurrence of each venue category.** Using **Elbow Point**, we can get our optimum k-value. In this technique, we run a test with a different number of k-values, measure the accuracy, and then choose the k-value at the point in which the line has the **sharpest turn**. An optimum value is that which is neither overfitting nor underfitting the model.

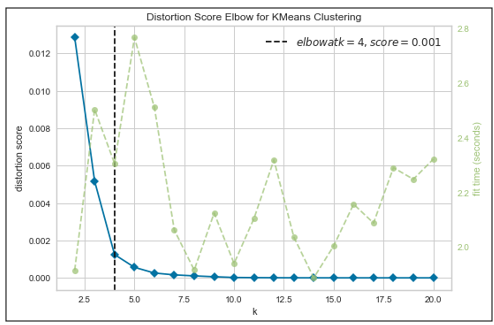


Finding the K vs Error values

Apparently, our optimum value is 4. However, for technical reassurance, we import a visualizer called **KElbowVisualizer**, from the **Yellowbrick** **package**. We fit our k-means model above to the visualizer to obtain the optimum value.



A snippet of the code in action



The model gives us this result, and we get the Elbow Point at k=4. This indicates that we will have a total of **4 cluster neighborhoods** in the end.

Note: We just integrated a model that would fit the error and calculate the distortion score. Moreover, in **k-means clustering**, objects that are similar based on a certain variable are put into the same cluster.



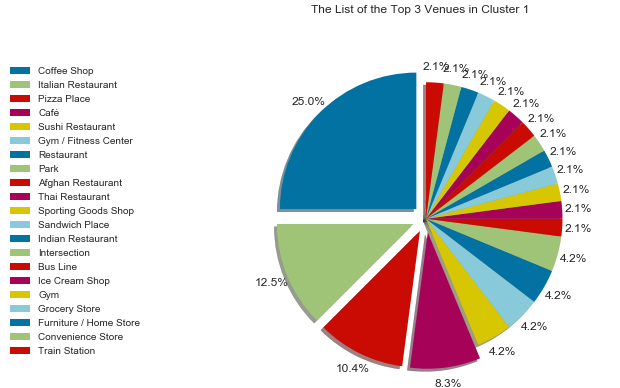
A snippet of the table with a cluster label for each neighborhood at k=4

### Results

Finally, we will visualize the resulting clusters using **Folium Python.**



As a result, we can examine venues listed inside each cluster and define the discriminating venue categories that distinguish them.





1. **Cluster 0: “Gas Station Venues”**  
   The first cluster contains one neighborhood only, with the gas station as the first most common venue.
2. **Cluster 1: “Coffee Shop and Restaurant Venues”**The second cluster holds sixteen neighborhoods, with the coffee shop, restaurant, and cafe venues appear to be the most common ones.
3. **Cluster 2: “Pharmacy Venues”**The third cluster includes one neighborhood with pharmacy as the most occurrence venue category.
4. **Cluster 3: “Park and Store Venues”**The fourth cluster has one neighborhood with the park, convenience store, and grocery as the majority venues.

### Discussion

The project’s main goal was to determine which location would be suited best for opening a restaurant in Toronto. We can evaluate which locations would be the most conducive ones, by looking at the following criteria:

#### I. Demographics & Accessibility

* **Vehicle** and **foot traffic** are factors of high significance when it comes to choosing a location for any business (restaurant, in our case). Demographics have shown the busiest and most active main roads in the city. Many of them are located around downtown. Then, we consider focusing on Central Toronto, York, and East York at first. However, this would become excessive and futile if those people are not our target demographic. Therefore, a better understanding of our target audience is required, and so a discussion with the team should be scheduled.
* **Accessibility** is also another important factor to be taken into consideration. Given that a discussion with the team took place and we know our target demographic, we will have picked a few candidate locations. Knowing how and why your customers will get to your location is crucial — factors like street visibility, parking slot, and location convenience are contributors to accessibility. Thus, further discussion with the team should take place.

#### II. Neighboring Businesses

* Neighboring businesses can have an impact on profitability; both positively and negatively.
* **Cluster 1** has the most restaurants in their neighborhoods. These businesses can be in different categories. However, they could still be competitors in the market and contend with the products you serve. Therefore, **cluster 1 is not recommended.**
* **Cluster 0, 2, and 3 are recommended neighborhoods** to inspect further. However, it will be a wise option to consider other businesses or amenities surrounding the area to complement your offerings. For instance, if we target people who spend their morning or afternoon outside, cluster 3 might be considered a good choice since it has “park” as the most common venue.

### Discussion

Finding the best location to start a business can be challenging and quite an uphill task due to many uncertainties. The abundance of data in this time and age, thanks to the digitization of society where so much human activity is now in the digital realm, along with advanced Machine Learning algorithms have made it easier for us to gain meaningful insights into the city of our choice and its pertaining neighborhoods. This helps everyone; entrepreneurs, business owners, and stakeholders to make decisions backed up by research and facts.

Thank you,

Usman Aftab Khan

[LinkedIn Profile](https://www.linkedin.com/in/usman-khan-458755163/)