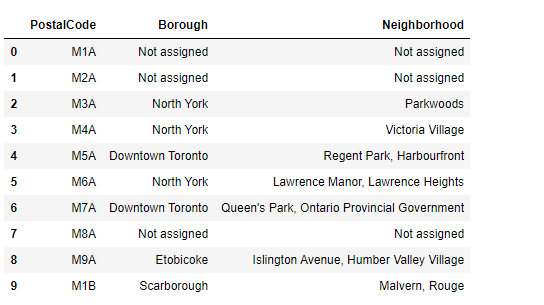
**COURSERA CAPSTONE PROJECT – THE BATTLE OF NEIGHBORHOODS**

**INTRODUCTION**

The aim of this project is to select an ideal location for opening a shopping mall in the city of Toronto, Canada. This report is targeted towards stakeholders who wants to invest in a shopping mall in the city. Using Data Science methodologies, we will find the perfect location where there are very few malls in the vicinity, to avoid competitions. We will use Clustering methodology to cluster the neighborhoods to find a location that is highly optimal for opening the mall.

**DATA DESCRIPTION**

The Geographical data to analyze the data is scrapped from the internet as the structured format of the data was not available. It is taken from the Wikipedia page using BeautifulSoup Package in Python. It consists of columns like Postal Code, Boroughs and Neighborhood which would be ideal for our analysis.

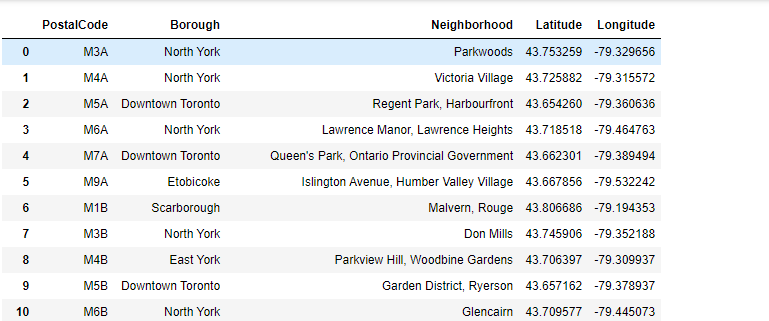


*Fig 1. Scrapped data from Wikipedia*

Further, the latitude and longitude information is added to get a clear picture of the location of neighborhoods and the venues in the neighborhoods are got through Foursquare API.

**DATA CLEANING AND TRANSFORMATION**

There were quite a few places in which the boroughs were unassigned, those rows were dropped as there are no useful information that can be obtained. The Neighborhoods which were unassigned were changed to the respective borough names. The data was joined with the latitude and longitude data and a sample of the cleaned data frames is show in figure 2.



*Fig 2. Data after cleaning*

After the data is cleaned, the venues in the neighborhoods is got through the Foursquare API for analysis. We then filter the data to get only the count of the shopping malls in the respective neighborhoods. Once this is done, we will have Neighborhood and the list of venues in them. We further transform the data such that each neighborhood has the mean of count of all of its venues to make it easier for analysis.



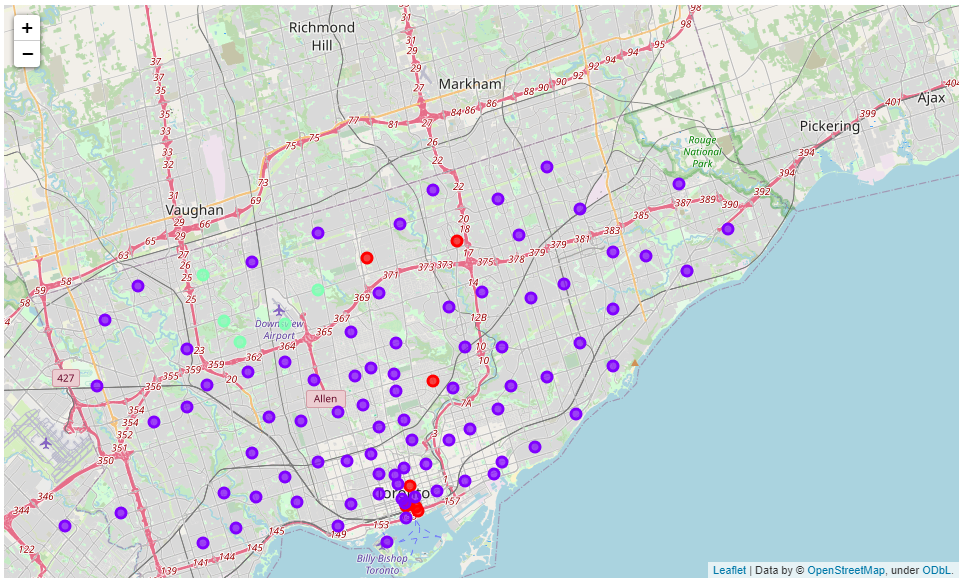
*Fig 3. Neighborhoods and Shopping malls(mean)*

**METHODOLOGY**

The data is segmented using k-means clustering in Python. This segments the neighborhood based on its similarity. The neighborhoods are divided into 3 clusters. After modelling the algorithm, it is further combined with its corresponding Neighborhood, Boroughs and Coordinates to interpret the data.

**RESULTS**

The clusters are formed as shown in figure 4, where Purple circles indicates cluster 1, Green circles indicates cluster 2, and Red circles indicated cluster 0. As it can be been seen the cluster 1 has very few to no shopping malls present in it, cluster 0 has moderate number of shopping malls whereas cluster 2 has many. On further analysis of clusters, it can been seen that the Boroughs like Scarborough, Etobicoke and Central Toronto are highly sparse.



*Fig 4. Segments after clustering*

**CONCLUSION**

As It can be seen from the model, building a shopping mall in the neighborhoods in cluster 2 is highly optimal as there are very few malls present and so, there are little to no competition in these areas. As previously discussed, targeting boroughs like Scarborough, Etobicoke and Central Toronto is highly recommended as not only in the sense that there are little competition in these areas but also people are more likely to visit the mall which are nearer to them, and that will attract more stores and thereby increase the profits of the stakeholders. From another perspective, if the stakeholders are already well established in similar cities, they can target neighborhoods in cluster 0 too, as there are only moderate number of competition and an already well-established name will give them an edge.