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.

	PostalCode	Borough	Neighborhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
7	M8A	Not assigned	Not assigned
8	M9A	Etobicoke	Islington Avenue, Humber Valley Village
9	M1B	Scarborough	Malvern, Rouge

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.

	PostalCode	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, Humber Valley Village	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
10	M6B	North York	Glencairn	43.709577	-79.445073

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.

	Neighborhood	Shopping Mall
0	Agincourt	0.000000
1	Alderwood, Long Branch	0.000000
2	Bathurst Manor, Wilson Heights, Downsview North	0.052632
3	Bayview Village	0.000000
4	Bedford Park, Lawrence Manor East	0.000000

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.

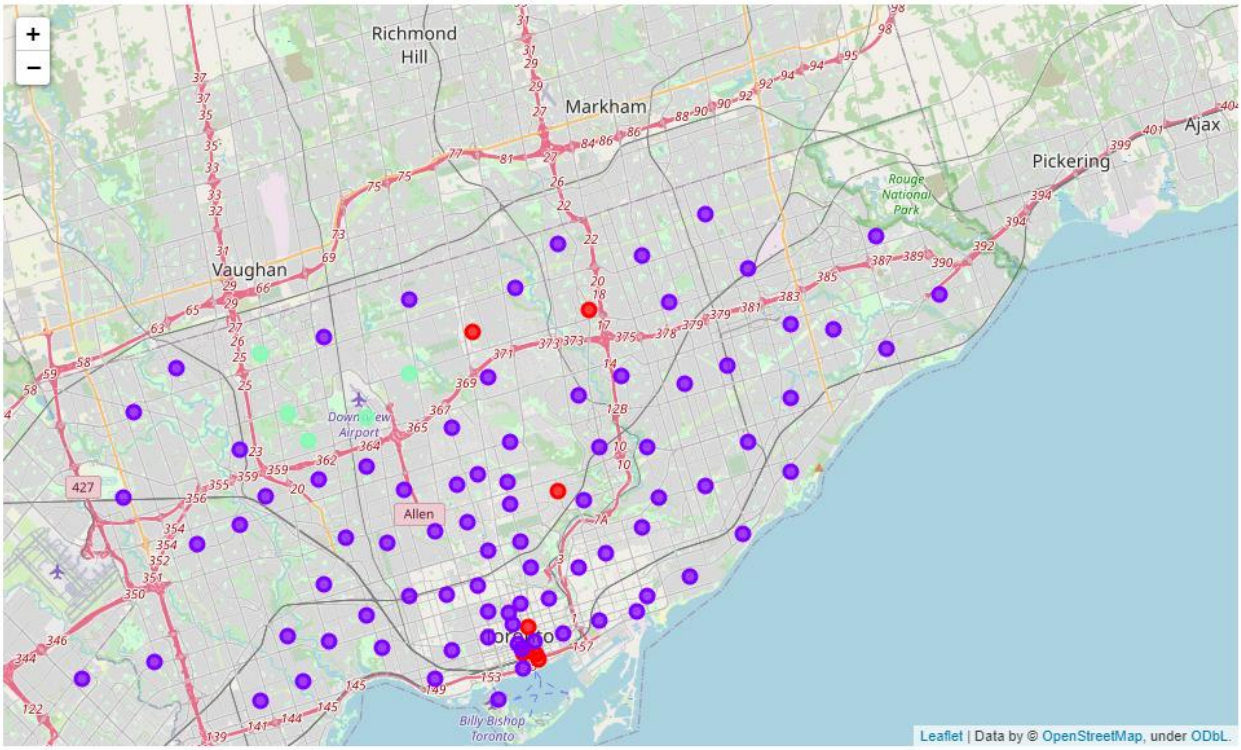

Fig 4. Segments after clustering

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 discu ssed, 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 whi ch are nearer to them, and that will attract more stores and thereby increase the profits of the stakeholders. Fro m another perspective, if the stakeholders are already well established in similar cities, they can target neighbo rhoods in cluster 0 too, as there are only moderate number of competition and an already well-established nam e will give them an edge.