Capstone Project - The Battle of Neighborhoods

**IBM Applied Data Science Capstone**

Opening a New Shopping Mall in Toronto, Canada

By: Santosh Krishna

July 2020

1. **Introduction:**

Shopping malls have become a vital part of consumer behavior. With a variety of outlets for Clothing, Books, Accessories etc. and places for movies, games and other activities, malls have become an integral part of every community. They are the one-stop place for all activities, especially in urban areas. To address the needs of people, there are a plenty of shopping malls in a big city like Toronto and many more are being built. Finding a place to open a new shopping mall is important, to maximize revenues by minimizing competition and yet, it close to the good parts of a city. This would largely determine whether there would be enough footfall and revenues to the Mall.

**1.1 Problem:**

With the growing number of shopping malls, a lot of them are facing losses due to competition. One of the major reasons for this is because shopping mall locations aren’t strategically chosen enough.

The objective of this project is to analyze Toronto’s location data and find the best location in the city to open a new Mall. We would be using a variety of Data science techniques like Analysis, Visualization, Machine Learning, clustering etc. to recommend the best location to open a shopping mall in Toronto, Canada.

**1.2 Target audience:**

This project would be useful for any property developers or realtors to identify a place to open a Shopping mall in Toronto, Canada. This would help them estimate the footfall and identify the competitors well in advance.

1. **Data:**

We would be using the following publicly available data to achieve this objective:

* List of Neighborhoods (and boroughs) in Toronto
* Geo Coordinates of neighborhoods (Latitudes and Longitudes)
* Venue data of Shopping malls to identify competition and cluster neighborhoods

**2.1 How to use the Data:**

The Wiki page: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> gives list of neighborhoods and boroughs in Toronto, Canada. We will scrap this data to extract data from this page using Python and beautifulsoup package. We will then use the Geocode package to fetch the coordinates of the boroughs. After fetching the Latitude and Longitude data of neighborhoods, we will use the Foursquare API to obtain the venue data that is the shopping malls to identify the competition in a neighborhood. Foursquare is one of the biggest location based databases available.

**2.2 Steps:**

* Scraping Toronto neighborhoods data from Wiki page
* Fetch geographical coordinates using Geocoder in Python
* Visualize Toronto’s map
* Use Foursquare to fetch shopping malls data
* Group data by neighborhood and calculate mean of frequency of each category
* Narrow down the data to Shopping malls
* Use k-means clustering to cluster the locations
* Visualize the data using Folium

1. **Methodology**

This section explains the approach followed to build this model that is being used to identify the best area in Toronto to open a Shopping mall.

* 1. **Scraping Toronto neighborhoods data from Wiki page:**

The page <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> has the list of postal codes of Canada that helps to fetch the Neighborhoods and Boroughs. The BeautifulSoup package in Python is used to scrap the data from the page.

* 1. **Cleaning the data:**

The data obtained is not clean. Here are some of the cleaning techniques used on this data to make it ready for use:

* The data will consist of three columns: Postal Code, Borough, and Neighborhood
* Only processed the cells that have an assigned borough. Ignored cells with a borough that is Not assigned.
* More than one neighborhood can exist in one postal code area. These will be combined into one row with the neighborhoods separated with a comma.
* If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.
  1. **Fetching Geo coordinates using Geocoder in Python:**

The Geospatial data is fetched from the Geocoder package. The Longitudes and Latitudes are then merged to the neighborhoods data. The data is then filtered to only contain Toronto’s data

* 1. **Visualize Toronto’s Map:**

The Folium library in Python is used to visualize all the Neighborhoods of Toronto in a Map.

* 1. **Use Foursquare to fetch shopping malls data in Toronto**

The Foursquare API is one of the largest location libraries. It has all the venue data and their corresponding geo coordinates. We used to fetch top 100 venues within 2 kilometers of all the neighborhoods in Toronto.

* 1. **Retrieve relevant information**

Filter required data and fetch the unique categories. We then group the data by Neighborhood and narrow it down to only Shopping Malls.

* 1. **K-means clustering to classify the neighborhoods**

Using K-means clustering, the neighborhoods in the previous step are divided into three categories by maintaining the value of K as 3. The k-means clustering functionality then splits the neighborhoods into 3 categories. The Latitude and Longitude details of these neighborhoods are then merged to this table.

* 1. **Visualizing the data**

The Folium application is again used to visualize the three clusters on Toronto’s map. The three categories of Neighborhoods are shown in different colors.

* 1. **Analyze the clusters**

The 3 clusters are visualized and the data is displayed to identify the cluster with least competition, hence with the most probability of success, under the assumption that location is the primary differentiator for the new Shopping Mall.

1. **Results**

Most of the shopping malls are present in the central bay area of Toronto, that is in Cluster 0. There are a moderate number in a scattered manner in the outer parts of the city, that is cluster 1. Cluster 2 on the other hand has very less competition, there are only 2 shopping malls in this region - near the University of Toronto. This is a great opportunity for realtors to open a new shopping mall in this region, first due to lesser competition and good proximity to the City center.

1. **Conclusion**

* **Cluster 0** - seem to be suffering from over-competition
* **Cluster 1** – Do not have much competition, but are scattered and not in the heart of the city
* **Cluster 2** - Good area, less competition. Suggested neighborhood for a new shopping mall - Near University of Toronto