**Capstone Project: The Battle of Neighborhoods**

**Opening a Sporting Goods Shop in Toronto**

By: M Farshchin

**A close up of a ball

Description automatically generated**

**Introduction/Business Problem:**

Toronto is one of the largest metropolitan cities in Canada. As the most populous city in Canada, this city is an attractive destination for many businesses. Toronto consists of 10 Boroughs and 103 neighborhoods. This project aims at exploring Toronto neighborhoods and finding potential locations for opening a **sporting goods shop** in a shopping mall or shopping plaza. For this purpose, a data-driven approach is applied to make an informed decision. As discussed in the following sections, different sources are used to obtain data about each neighborhood and K-Mean clustering method is used to cluster similar neighborhoods. The results of this study could be used by distributors of sports goods that are interested in opening new shops in the Toronto.

**Data Description:**

For this problem the following data sources are used:

* Data about Toronto neighborhoods are scraped from the following Wikipedia page:

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

This dataset provides information about different postal areas of Toronto along with their related Borough and neighborhoods.

* Geographical coordinates of each postal code is obtained from the following csv file:

<https://cocl.us/Geospatial_data>

This dataset provides geographical coordinates (Latitude and Longitude) of different postal areas of Toronto.

* Foursquare API is used to get all venues for each neighborhoods of Toronto.

**Methodology:**

In order to find the potential locations for opening a new **sport goods shop** inToronto, the following criterion are considered:

1. Neighborhoods with higher density of sport venues are more attractive to open the new shop because there is more demand in such neighborhoods.
2. Neighborhoods that already have at least one sport goods shop are not suitable because there will be more competition.
3. In this project we assume that the customer is interested in opening the sport goods shop in a shopping mall or shopping plaza. Therefore, the neighborhoods with more shopping malls or shopping plazas are more attractive because they give more options.

**Preprocessing the data and exploratory data analysis:**

First step of the project was to collect and organize the required data. As mentioned in the Data Description section, information about different neighborhoods of Toronto was collected and processed into a data frame. A snapshot of the first few rows of the resulting data frame is shown in the following image. This data frame includes information such ‘Postalcode’, ‘Borough’, ‘Latitude’, and ‘Longitude’ of each neighborhood of Toronto. Exploring this data frame shows that Toronto has 10 Boroughs and 103 Neighborhoods.

A screenshot of a cell phone

Description automatically generated

Having this information, the following map of Toronto with neighborhoods superimposed on the top is created.

A close up of a map

Description automatically generated

In the next step of this project, the **Foursquare API** is used to collect information about the different venues of different neighborhoods of Toronto. The following figure shows the main categories of Foursquare data base along with their category id:

A screenshot of a cell phone

Description automatically generated

Since we are interested in finding potential locations for opening a Sporting Goods Shop in a shopping mall or shopping plaza, we need to obtain information about venues in two main categories of **Outdoors & Recreation** and **Shop & Service**.

The following figure shows some of the subcategories of the Outdoors & Recreation:

A screenshot of a cell phone

Description automatically generated

From this category, we get all venues that are in a radius of about one mile (~1610 m) from the center of each neighborhood. In addition, since we are interested in only sport venues in this category, we only keep the sport venues. A snapshot of the resulting data frame is shown below.

A screenshot of a cell phone

Description automatically generated

To get a better understanding of the obtained venues, the top 10 venues are extracted and analyzed. The results are shown below.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

The resulting figures show that Gyms and Gym/Fitness centers are the most popular venues in Toronto.

Regarding the shopping venues, since our assumption is that the customer is looking for available shopping malls and shopping plazas to open the new shop, in the next step, we get the information about availability of these venues in neighborhoods of Toronto. In addition, our goal is to avoid neighborhoods that already have a ‘Sport Goods Shop’, therefore we get venues for this category as well.

Now we summarize all information into the following data frame (only first few rows are shown).

A screenshot of a cell phone

Description automatically generated

**Analysis and K-Means Clustering:**

After preprocessing the data, we are ready to proceed with clustering the data. In the first step of the process we remove the neighborhoods with one or more ‘Sport Goods Shop’ from the data frame. The resulting data frame will have 78 neighborhoods. A snapshot of the first few rows of this data frame is shown below.

A screenshot of a cell phone

Description automatically generated

Now we proceed with clustering the data. The following figure visualizes the resulting clusters on the Toronto map.

A picture containing text, map

Description automatically generated

**Results and discussion:**

To make a better understanding of the clusters and make better judgment of the results, we develop the following scatter plot which shows the number of shopping venues (available shopping plazas or shopping malls to open a sporting goods shop) against the number of sport venues for each neighborhood. Different clusters in this figure are shown with different colors.

A close up of a map

Description automatically generated

Preferable cluster

From this figure, it seems that neighborhoods in **Cluster 7** are superior to other neighborhoods because compared to other neighborhoods, these neighborhoods have more sport venues while they have at least 2 shopping plaza or shopping mall to open a **Sporting Goods Shop**.

The following figures shows the details of this neighborhoods and their locations on the Toronto Map.

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

In this study we explored different neighborhoods of Toronto and used data to find the potential locations for opening a sporting goods shop. We used data collection tools such as Foursquare API, descriptive data analysis, data visualization, and machine learning techniques such as K-Means clustering to solve our problem. Form the results we could see that we can use data science tools to help solve problems and make more informed decisions.