1.1 Introduction:

North York is an eclectic, multicultural district home to the hands-on Ontario Science Centre and the Aga Khan Museum, with exhibits on Islamic culture in a striking modern building. In the area’s north, Black Creek Pioneer Village is an 1800s living museum. Sprawling Downsview Park includes a lake, event spaces, and a flea and farmers’ market, while Edwards Gardens has a greenhouse, fountains, and botanic gardens. In this district there is also the Earl Bales Park which is a very large park. The area is well developed and therefore many new businesses are interested in investing in this specific region therefore the market is highly competitive. *[Wikipedia]*

1.2 Problem Description

A Mediterranean restaurant named ‘Mediterrano’ is interested to open its new business unit in North York but first they have requested from us an analysis in the different neighborhoods of North York and the most common venues in order to find the optimal location to open their new restaurant. ‘Mediterrano’ it´s a very successful restaurant in Ontario and now they want to be sure that they will continue their success also in North York.

1.3 Interest

The management team of ‘Mediterrano’ are the main people of interest of this research. Other businesses which might be interested in investing in this region and belonging in the food industry belong could also be benefit from this analysis.

2. Data acquisition and cleaning

2.1 Data Sources

Our goal is to segment and cluster the neighborhoods of North York in the city of Toronto as those will be used to analyze the existing venues in the region of North York.

* 1. [Postal codes of Canada](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). This Wikipedia page was used in order to scrap the postal codes of the whole city of Toronto.
  2. Geospatial\_Coordinates.csv was provided from Applied Data Science Capstone course from IBM in coursera which contains the geospatial coordinates of the different regions of Toronto.
  3. Venue’s data was extracted using an API call from Foursquare

2.2 Data cleaning

Two tables were created using the data from i & ii sources. The first table included the postal codes, the boroughs and the neighborhoods of the city of Toronto and the second one the postal codes and the geospatial coordinates. The purpose was to combine those two tables but there were some differences in the naming of the columns so there was some data processing that was required in order to make the tables consistent and combine them.

After this step the source iii was used in order to extract the Venues data and combine it with the other datasets in order to create a homogeneous dataset which will later on be used for the k-means clustering technique.

2.3 Feature selection

The features that will be used have been collected from the location provider Foursquare.com, which we used to extract data for the different types of venues that exist in every neighborhood. The data for every neighborhood consisted of the features below:

* Neighborhood
* Neighborhood Latitude
* Neighborhood Longitude
* Venue
* Venue Latitude
* Venue Category

A picture containing table

Description automatically generated