FINAL REPORT ON THE BATTLE OF NEIGHBORHOODS

TORONTO VS NEW YORK

MODULE: - APPLIED DATA SCIENCE CAPSTONE

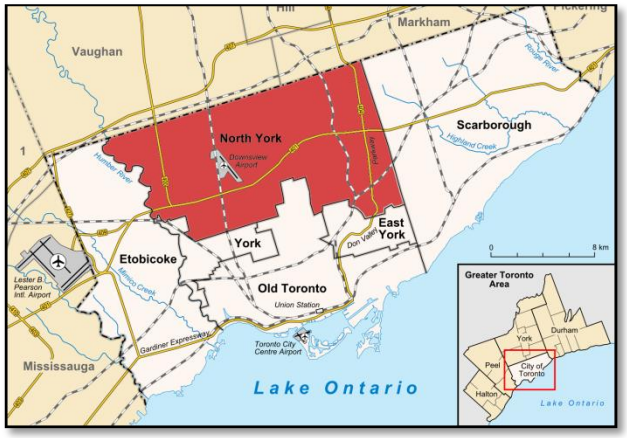
SUBMITTED BY: - KAJAL PANDA

TABLE OF CONTENTS:-

1. INTRODUCTION/BUSINESS PROBLEM:

1.1. DISCUSSION OF THE PROBLEM:

This project report on the battle of neighbourhoods titled 'The Battle of Neighbourhoods - Toronto vs New York City' gives an overview on the comparison of neighbourhoods of two different places. A person named Laurel got a job offer and she has to choose between two job locations. One is North York, Toronto and the other is Manhattan, New York City. She is confused where to move in and she is unaware of which locality provides better living standards. The project will help in understanding the similarities and dissimilarities of the neighbourhoods of North York, Toronto and Manhattan, New York city. North York is the area in the north of Toronto that includes a wide range of neighbourhoods. North York is an eclectic, multicultural district home to the hands-on Ontario Science Centre and the Aga Khan Museum etc. Manhattan is the most densely populated of New York City's 5 boroughs. Its iconic sites include skyscrapers such as the Empire State Building, neon-lit Times Square etc. With Google one can get every detailing. But this project is a small approach to deal with similar problems.



1.2. TARGET AUDIENCE:

The objective is to obtain the insights of the neighbourhoods of North York and Manhattan for Laurel. It will help her to opt for a place with better living standards. But this project can be broadly categorised to solve problems like an enterprise wanting to relocate its office, families searching for a better locality to settle in, entrepreneurs wishing to setup businesses and much more.

2. DATA SECTION:

2.1. DATA SOURCES:

The data used in this project is acquired from the respective cities Wikipedia website pages. The datasets consists of the postal codes, neighbourhood names, latitude, and longitude information for each neighbourhood. Foursquare API search feature will be used to collect neighbourhood venue information. In addition to Foursquare, various python packages are used to create maps and machine learning models to provide insights into this neighbourhood battle project.

The following datasets from these websites are used:

- Toronto neighbourhoods with boroughs:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

- Toronto Latitudes and Longitudes:
http://cocl.us/Geospatial_data

- New York city neighbourhoods with boroughs:
https://geo.nyu.edu/catalog/nyu_2451_34572

- New York city Latitudes and Longitudes:
Python Geo library

2.2. DATA CLEANING AND MANIPULATION:

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of not assigned values for Borough and Neighbourhood columns which are dropped. Unique features are considered to get the acquired result. The data frames are constructed merging postal code with respective neighbourhoods, boroughs, latitude and longitudes.

3. METHODOLOGY SECTION:

3.1. PROCESS STEPS:

- HTTP requests would be made to this Foursquare API server using zip codes of the city neighbourhoods to get the location information (Latitude and Longitude).
- Foursquare API search feature helps in collecting the nearby places of the neighbourhoods. Due to http request limitations the number of places per neighbourhood parameter is set to 100 and the radius parameter is set to 500.
- Folium- Python visualization library is used to visualize the neighbourhoods cluster distribution of both the cities over an interactive leaflet map.
- Extensive comparative analysis of two neighbourhoods is carried out to derive the desirable insights from the outcomes using python's scientific libraries Pandas, NumPy and Scikit-learn.
- Unsupervised machine learning algorithm K-mean clustering is applied to form the clusters of different categories of places residing in and around the neighbourhoods. These clusters from each of the neighbourhoods are analysed collectively and comparatively to derive the conclusions.

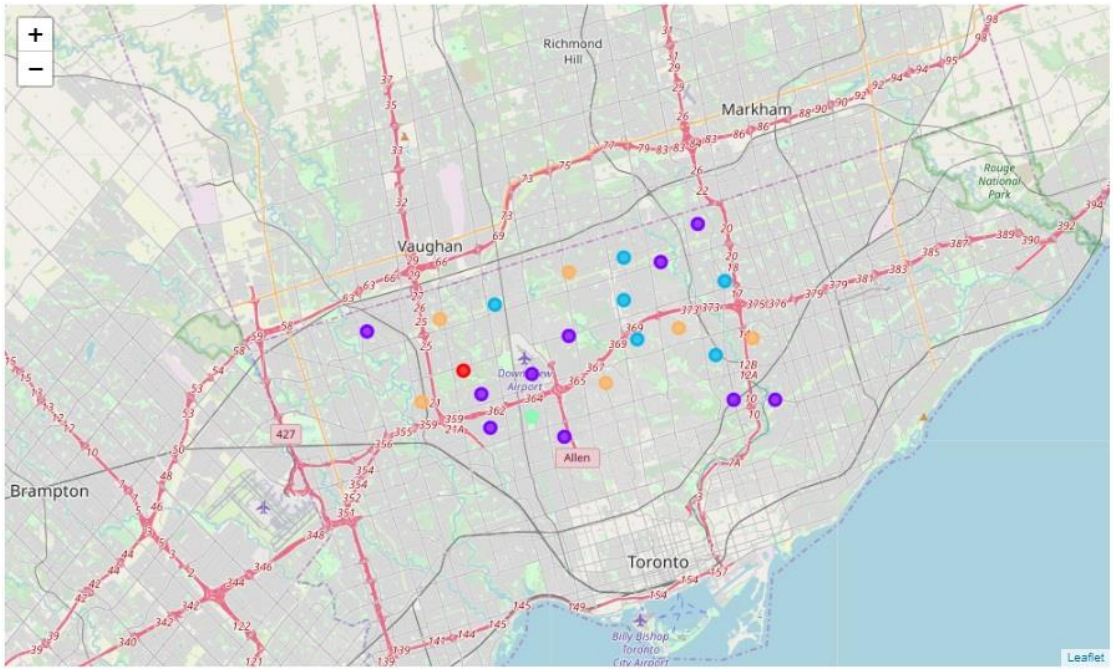
3.2. LIBRARY AND PACKAGES IMPORTED:

- Pandas - Library for data analysis
- NumPy – Library to handle vector data
- JSON – Library to handle JSON files
- Geopy – To retrieve location data
- Requests – Library to handle http requests
- Matplotlib – Python plotting module
- Sklearn – Python machine learning library
- Folium – Map rendering library

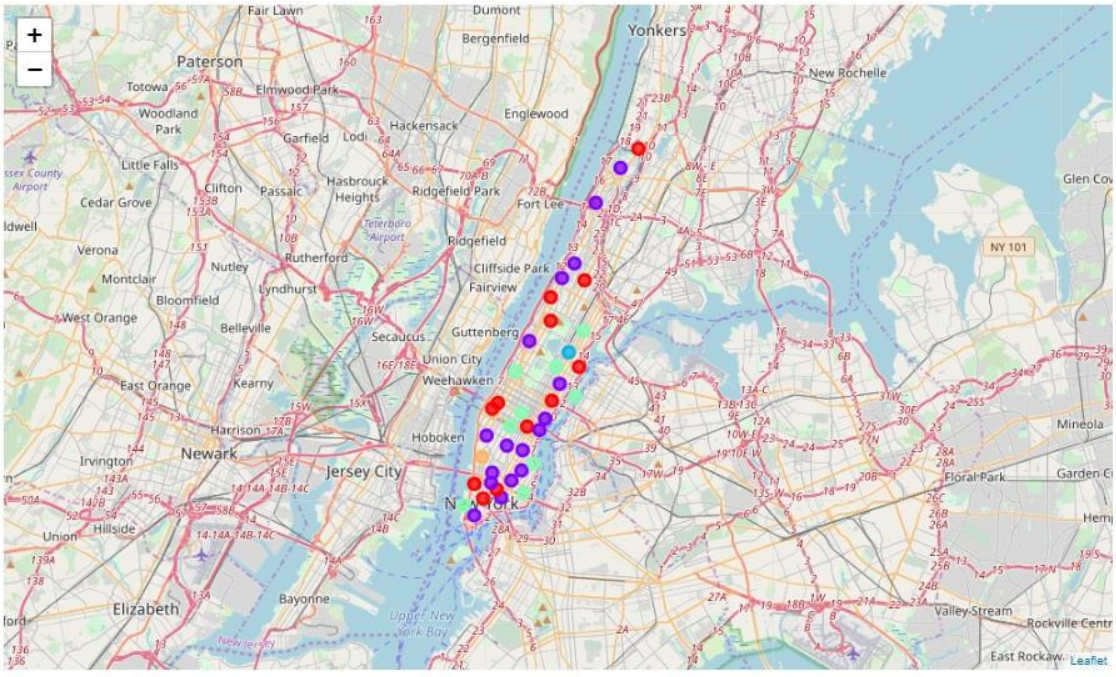
```
import numpy as np
import pandas as pd
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

import json

!conda install -c conda-forge geopy --yes
from geopy.geocoders import Nominatim

import requests
from pandas.io.json import json_normalize

import matplotlib.cm as cm
import matplotlib.colors as colors

from sklearn.cluster import KMeans

!conda install -c conda-forge folium=0.5.0 --yes
import folium
```

4. RESULT SECTION:

MAP OF NORTH YORK BOROUGH OF TORONTO NEIGHBOURHOOD CLUSTERS INTO 5 CLUSTERS:



MAP OF MANHATTAN BOROUGH OF NEW YORK CITY NEIGHBOURHOOD CLUSTERS INTO 5 CLUSTERS:



5. DISCUSSION SECTION:

Toronto has 11 boroughs and 103 neighbourhoods. The geographical coordinate of Toronto, CA are 43.7170226, -79.4197830350134. The geographical coordinate of North York, Toronto, CA are 43.7543263, -79.44911696639593. North York has 204 distinct venues in 110 categories. Most common venues include coffee shops, Asian and Chinese restaurants, clothing stores etc.

New York City has 5 boroughs and 306 neighbourhoods. The geographical coordinate of New York City, New York are 40.7308619, -73.9871558. The geographical coordinate of Manhattan, New York City, New York are 40.7896239, -73.9598939. Manhattan has 2857 distinct venues in 336 categories which is humongous in number. Most common venues are book stores, pizza place, coffee shops, gyms, bakeries, grocery stores and much more.

Many of the neighbourhoods are homogenous and are very similar to each other. Both North York and Manhattan borough consists of neighbourhood cluster that contain majority of the neighbourhoods. Manhattan borough had a significant more number of neighbourhoods and venues than North York.

6. CONCLUSION SECTION:

In the comparison between North York and Manhattan, Manhattan excels as it has more number of distinctive venues in more different categories than North York. Manhattan will be the best job location for Laurel to get a better neighbourhood and standard place for accommodation.

Like this problem, we can compare any neighbourhood for relocating, job accommodation, buying houses, setting up businesses etc. This project can be broadly categorized into many other intrigue problems. And the solution can have codes like this project including different types of clustering and classifications.