



Coursera Capstone Project: IBM Applied Data Science

ABDUL SUBAHAN MOHAMED GHOUSE

Overview

- ▶ Introduction
- ▶ Objectives
- ▶ Data
 - ▶ Neighbourhoods
 - ▶ Geocoding
 - ▶ Venue Data
- ▶ Methodology
 - ▶ Folium
 - ▶ One hot encoding
 - ▶ Top 10 most common venues
 - ▶ K-Means Clustering
- ▶ Results
- ▶ Discussion
- ▶ Conclusion

Introduction

- ▶ When people are finding the best possible cities to live in, they would need proper means to compare cities in the most accurate way with verified information. This would allow them to make sound decisions of whether the neighbourhood suits their taste.
- ▶ Tourists are another target group who scout for amenities in neighbourhoods and the living conditions in the city.
- ▶ When we are debating on the best place to stay in, we have lots of factors that would be determining the liveability of the city.

Comparisons

- ▶ Firstly, we would look at the **Overall Comparison**. This would be determined using the same factors of the city which would give us a general overview of the cities.
- ▶ Secondly, we would look into **Crime Rates**. We would look into the crime rates of the individual cities and pit them against the statistics giving the national average.
- ▶ **Cost of Living and Salary** would be compared for the different cities which are important selling points for migrants especially.
- ▶ **School** comparison would allow migrant families to find the best schools in the vicinity.
- ▶ Lastly, we would look into **neighbourhood** comparison which would provide the best place to live within any city.


Objectives



- ▶ To apply machine learning(k-means) and data science principles to compare between New York City and Toronto cities to help individuals, families and tourists to make a decision on where to move or relocate to.
- ▶ Through a combination of Folium, Foursquare API and K-means algorithm, we would be able to obtain different cluster of neighbourhoods with its associated venues.

Data

- ▶ Neighborhood and venues of the city
- ▶ <https://api.foursquare.com/v2/venues/search>
- ▶ Downloaded Canada postal code and its latitude and longitude from <http://download.geonames.org/export/zip/> and converted as CSV and uploaded
- ▶ Venue Data
- ▶ From the location data obtained after Web Scraping and Geocoding, the venue data is found out by passing in required parameters to the FourSquare API and creating another DataFrame to contain all the venue details along with respective neighbourhoods.



From the datasets, we can access information on the following:

- ▶ State
- ▶ Country
- ▶ Cities
- ▶ Latitude and Longitude
- ▶ Population Density
- ▶ Amenities
- ▶ Neighbourhood
- ▶ County

Methodology

► **Exploratory Data Analysis**

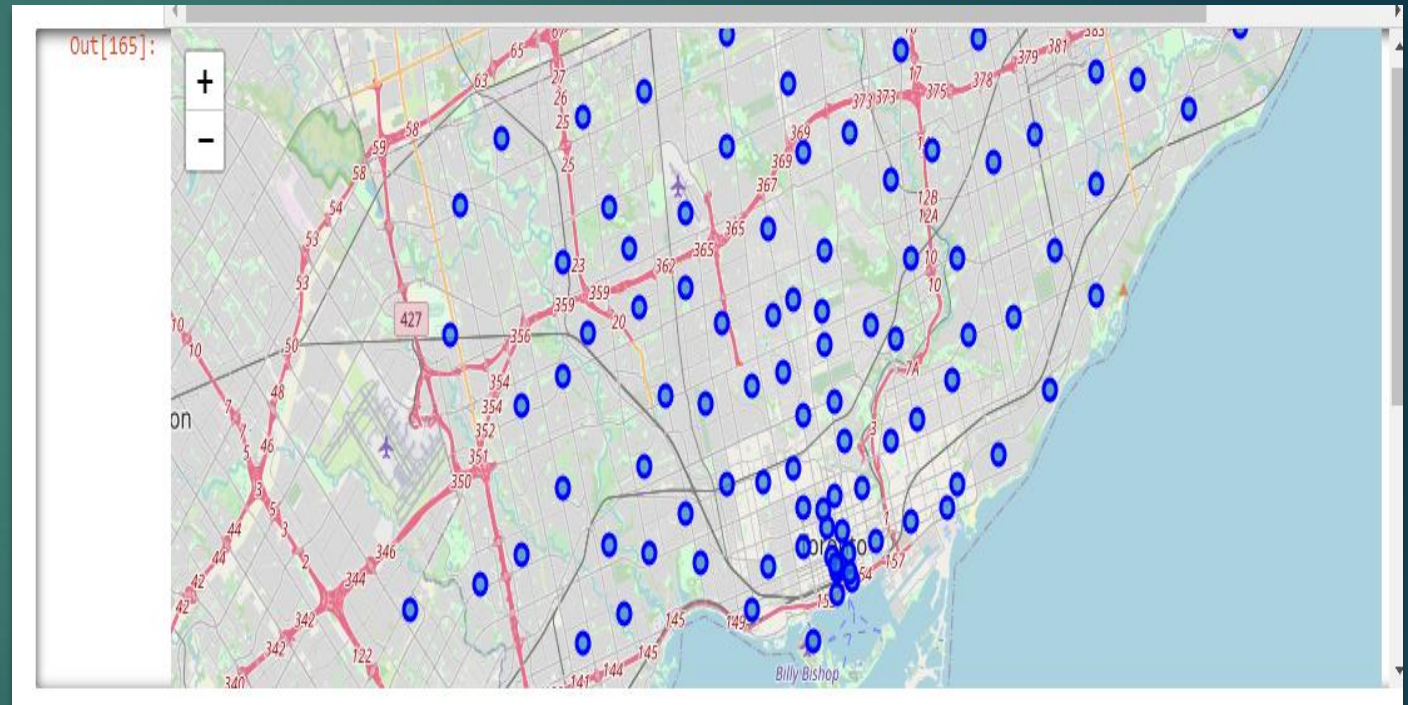
- We first need to import the data from the New York City and Toronto cities into pandas, plotting and viewing using Folium and geolocator.
- We now need to slice and segment the data. There are lots of columns, but we need to pick only what we need and drop the unwanted columns. We need to merge the data frames so that we can look at the cities at a glance.
- We can then plot and view the data to see the layout and distribution of cities and neighbourhoods on the map.

Methodology

- ▶ With the neighbourhood maps, we are going to study these neighbourhoods using Foursquare APIs.
- ▶ For each of the neighbourhood, Foursquare search engine returns a list of the top common venues.
- ▶ Based on venue information, the neighbourhoods can be clustered with some similarities.
- ▶ We would form a GeoDataFrame by using a Shapely geometry object which is formed through merging the 'Latitude' and 'Longitude' columns.

Methodology

- ▶ **Folium**
- ▶ Folium builds on the data wrangling strengths of Python and the mapping strengths of the leaflet.js library.
- ▶ All cluster visualization are done with the help of Folium which in turn generates a Leaflet Map using OpenStreetMap technology.



Methodology

- ▶ **One-hot encoding**
- ▶ One hot encoding is a process by which categorical variables are converted into a form that could be provided to Machine Learning Algorithms to do better prediction. For the K-means clustering algorithm, all unique items under Venue Category are one-hot encoded.
- ▶ **Top 10 most common venues**
- ▶ There are a high variety of venues and hence only the top 10 common venues are selected and a new DataFrame is made, which is then used to train the K-mean Clustering algorithm.

Conclusion

- ▶ The analysis has shown that Folium- Python Library is quick and effective in building an interactive data visualization and Foursquare API for the neighbourhood data collection. It is more appropriate to cluster neighbourhood cities data based on known and accepted machine learning techniques like K-means algorithm.
- ▶ Results have to very well refined and cleaned such that it fits into the scope of study. Such results would then of interest to people who want to migrate to the cities and for tourists who are looking for amenities in their vacation environment.

Conclusion



There is more room for improvement like the expansion of cities to expand the geographical setting.



Crime data can be used in more in-depth way to provide sound decisions to choosing a location to relocate to.



The methods used here may not be vigorous enough, but the approach here does not steer away from the prime focus of testing the usefulness of neighbourhood data analysis.