IBM
Data Science Project Week 5th

INTRODUCTION

This is a capstone project for IBM Data Science Professional Certificate. In this project, I will create a hypothetical scenario where an entrepreneur wants to build a Pub, so the young people in the region can have a pleasant place to go and have fun in the Toronto Area.
As the Pubs is a place to have fun with friends this entrepreneur might think of opening its business in areas where there is not places to have fun. With the purpose in mind, finding the location to is one of the most important decisions for this entrepreneur and I am designing this Project to help him find the most suitable location.

BUSINESS PROBLEM

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Pub in Toronto, Canada. By using data science methods and tools along with machine learning algorithms such as clustering, this project aims to provide solutions to answer the business question : In Toronto, if na entrepreneur wants to open a Pub, where should they consider opening it?

TARGET AUDIENCE

An entrepreneur wants to find the best places to create a pub for people to have fun.

DATA

To solve this problem, we will need below data:

- List of neighborhoods in Toronto, Canada
- Latitude and Longitude of these neighborhoods
- Venue data related to Pubs. This will help us find the neighborhoods that are more suitable to open a Pub

EXTRACTING THE DATA

- Scrapping of Toronto neighborhoods via Wikipedia
- Getting Latitude and Longitude data of these neighborhoods via Geocoder package
- Using Foursquare API to get venue data related to these neighborhoods

METHODOLOGY

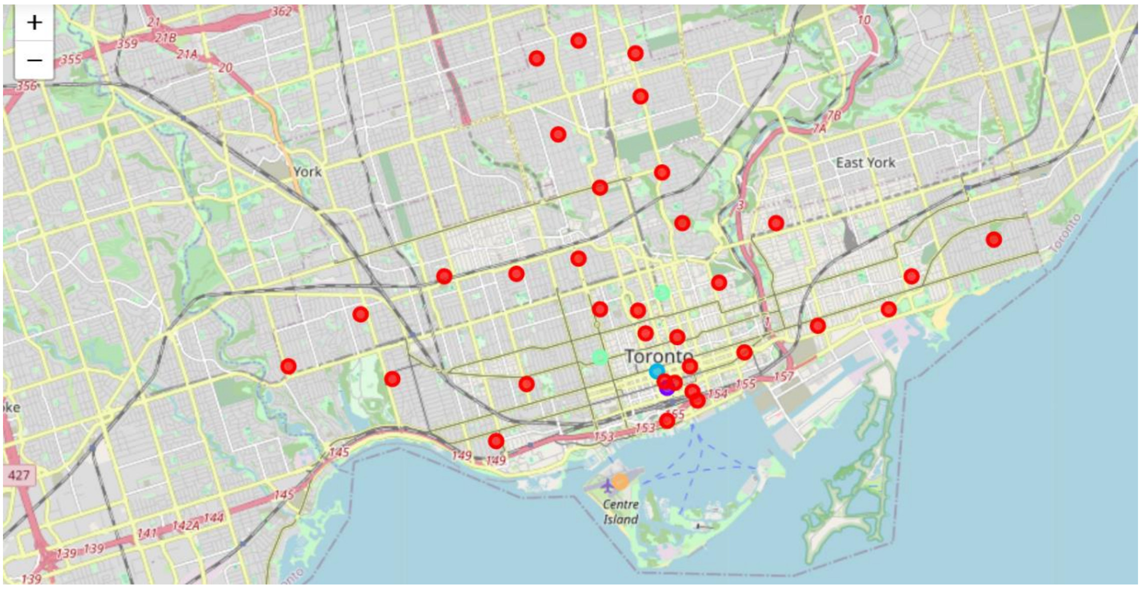
First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by extracting the list of neighborhoods from Wikipedia:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
I did the web scraping by utilizing pandas HTML table scraping method as it is easier and more convenient to pull tabular data directly from a web page into the data frame.

However, it is only a list of neighborhood names and postal codes. I need to get their coordinates to utilize Foursquare to pull the list of venues near these neighborhoods. To get the coordinates, I tried using Geocoder Package but it was not working so I used the CSV file provided by IBM team to match the coordinates of Toronto neighborhoods. After gathering these coordinates, I visualize the map of Toronto using Folium package to verify whether these are correct coordinates. Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude, and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the Pick up top 10 venues category. This is to prepare clustering to be done later. I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I have clustered the neighborhoods in Toronto into

5 clusters based on the top 10 venues categories. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the Pub.

RESULT

CLUSTERS



The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based how similiar each region is.

Cluster 0 – Areas with the most pubs and others good places to go.
Cluster 1 – Area with a Pub
Cluster 2 – Area without pubs
Cluster 3 - Area without pubs
Cluster 4 – Area without a pub (Airport Area)

RECOMMENDATIONS

Cluster 0 Have a lot of Pubs and a good places to go to have fun, only in the Clusters 2,3,4 don't have Pubs and only in the Kensington Market, Chinatown, Grange Park and Richmond, Adelaide, King are the best places to have a Pub because CN Tower is a Airport Area, and is isn't a good place to have some kind of Pub.