IBM Data Science Capstone Project

Intro:
The purpose of this project is to collect geospatial data, using Foursquare, on a given location to aid in solving a hypothetical business problem.

Problem Description:
A successful owner of a steakhouse franchise is looking to open a new location in Toronto. The objective of this analysis is to identify the ideal location within Toronto to open the new restaurant. The success of the business will be largely dependant on the location of the restaurant within the city, therefore data will be collected on all local businesses to find the ideal blend of surrounding venues to compliment the restaurant.

Data:
To perform the analysis into finding the ideal location, geospatial data is required. Firstly data is obtained about the target location, in this case Toronto. This data includes towns, boroughs and geographical coordinates, creating the framework to dive deeper into the location analysis with Foursquare. Foursquare identifies and categorises businesses within the given towns and boroughs. The data is all aggregated together to then visual and analyse.

Methodology:
Firstly data is obtained about the target location, in this case Toronto. The data is taken from a wikipedia page containing a list of post codes in Toronto, using the web scraping library BeautifulSoup. The data is scraped as a string output and then parsed through to obtain the relevant information. This data includes towns, boroughs and geographical coordinates, creating the framework to dive deeper into the location analysis with Foursquare. The data is cleaned and stored in a dataframe, before being merged with a csv containing the longitude and latitude for the corresponding postcodes.

Next the towns, a long with their longitude and latitude values, are passed through the Foursquare API to obtain information regarding the business of nearby venues to that location, the data is stored and aggregated to a town level, including the venue category. The venue category information is then split in multiple columns and transformed into numerical values using one hot encoding technique. Using this dataset a new dataframe is created to store the 10 most common venues in each location, this builds the profile for each location.

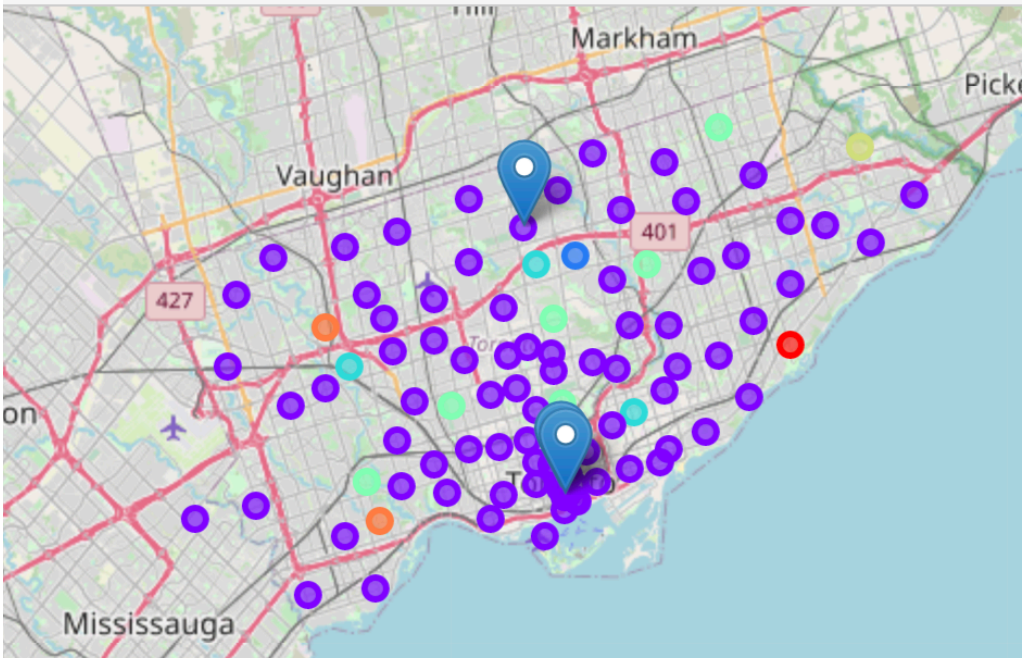
The profile for each location is then analysed through k-means to cluster into multiple segments. [The number of these segments is determined through the highest accuracy score for a range of values between 1 and 10]. These clusters create a network of locations based on their similarity in nearby venues, as previously stated in the problem description, the success of the project is paramount to the blend of surrounding businesses complimenting the restaurant, thereby attracting the right customers.

Further analysis is undertaken inspecting each cluster and labelling them accordingly to their most common venues.

Results:
The results of the clusters are as followed..
Cluster 0: Contains a single location, with a varied array of venues, mostly low spending shops.

Cluster 1: This clusters contains the majority of the locations analysed. With a wide variety of venues but predominately food and drink venues, restaurants in particular.
Cluster 2: Contains a duplicate of Cluster 0. The exact same values.
Cluster 3: Most common venue is parks followed by convenience stores and other small shopping facilities.
Cluster 4: Contains a high influence of parks, rivers, dog runs, gyms and discount store. This shows a low consumer spending budget and consists of more leisure and activities.
Cluster 5: Contains a single value, consisting of fast food venues and discount stores.
Cluster 6: Contains two values, consisting of fast food discount stores and in particular two baseball fields.

The visual distribution of the clusters is somewhat varied, there is no single cluster which sits in a particular location, apart from the clusters with a single value. The one slight exception to this is cluster 1, which is spread entirely across Toronto but does form a slight concentration of data points at the 'Lake Shore Boulevard' this would indicate the area has a high consistency with that of cluster 1s features. Thereby creating a wide variety of venues with the same business, attracting the same customers and giving them more variety in terms of competition, which further fuels the incentive of the consumer to visit the area.



Discussion:
Moving forward, the next steps would be to gather financial information regarding each location. This could include maintenance costs (rent), consumer income and spending figures and local authority future spending initiatives (where money is spent, what buildings/venues are upcoming to be built). This is information is relevant to the longevity of the proposed business plan. Additionally consumer data would also benefit our model, this could include demographic and marketing data (understand the local population's interests, in particular their appetite).

Cluster 1 could also be further refined as it holds a large majority and not all top 10 venues show consistency.

Conclusion:
To conclude, the cluster 1 produced the desired venues fit for the location of the new steakhouse restaurant, high spending venues in hospitality and food and drink. This is validated by the fact three existing steakhouse venues (that made it into the top 10 venues per location) also classified

into this cluster. However two of these locations are in the area ‘Lake Shore Boulevard’ of Toronto, which has a high concentration of cluster 1 locations, whereas the other is positioned in the outer perimeter in the area of data points. Therefore we could position the steakhouse in a similar outer position, perhaps on the other side of Toronto. Or we could choose to position it in the concentration of cluster at ‘Lake Boulevard’. I would recommend the latter because the concentration of similar venues attracts the same customers and it offers them a variety in terms of competition, which further fuels the incentive of the consumer to visit the area.