IBM DATA SCIENCE

CAPSTONE PROJECT – FINAL SUBMISSION

BACKGROUND, PROBLEM & INTEREST

BACKGROUND

- The final assessment set by the Coursera Applied Data Science Capstone centres around leveraging Foursquare location data to explore and compare neighbourhoods or cities of my choice; and to come up with and solve a problem using Foursquare location data.

PROBLEM

- The task can be split into two sections: neighbourhood/city comparison, and problem solving hereafter named Question 1 and Question 2 respectively.
- **Question 1 objective:** To compare the neighbourhoods of Downtown Toronto and Ottawa and determine how similar or dissimilar they are.
- Using the Foursquare API I will explore the most common venue categories in Downtown Toronto and Ottawa, then use this feature to group the neighbourhoods into clusters – using K means. After which I will use the Folium library to visualise the neighbourhoods in both Toronto and Ottawa along with their emerging venue clusters. This information will benefit the Government of Canada as they are attempting to establish the diversity of venue types in these locations.
- **Question 2 objective:** A restaurant owner is looking to open a new Italian restaurant in Toronto, the objective is to recommend the best area in which a new restaurant could be located.
- Using the Foursquare API I will explore the Italian restaurants in each neighbourhood in Toronto. After which I will use the Folium library to visualise the restaurants to inform the owner of the current distribution. The spatial distribution is highly important from a competition point of view as an area highly saturated in italian cuisine will prove detrimental to their business. Therefore, the owner will be looking for an area that has none/few Italian restaurants at present.

INTEREST

- As mentioned in the problem section of this report the interest can also be split into two sections.
- **Question 1 interest:** The target audience of this analysis are the analysts within the Government of Canada. The stakeholders are the wider Government of Canada.
- **Question 2 interest:** The target audience is the restaurant owner. The stakeholders are the bank – who are lending the owner the money to build his new restaurant.

DATA SOURCES

QUESTION 1:

NEIGHBOURHOOD/CITY COMPARISON

- Csv of Toronto location data containing postal codes, boroughs and neighbourhoods – from a previous assessment in the IBM Data Science Specialization
- Csv of Ottawa geospatial data containing postal codes, city, neighbourhood, latitude and longitude - https://github.com/ccnixon/postalcodes/blob/master/CanadianPostalCodes.csv
- Geospatial coordinates of Toronto containing latitude and longitude - from a previous assessment in the IBM Data Science Specialization
- Foursquare API location data for venues in Toronto and Ottawa – including venue latitude, longitude, category and name - https://api.foursquare.com/v2/ . Limit 100. Radius 500.

QUESTION 2:

PROBLEM SOLVING

- Csv of Toronto location data containing postal codes, boroughs and neighbourhoods - from a previous assessment in the IBM Data Science Specialization
- Geospatial coordinates or Toronto containing latitude and longitude - from a previous assessment in the IBM Data Science Specialization
- Foursquare API location data for italian restaurants in Toronto – including name, address, latitude, longitude, distance, postal code, city, state and neighbourhood - https://api.foursquare.com/v2/ . Limit 100. Radius 10000.

DATA CLEANING

QUESTION 1:

NEIGHBOURHOOD/CITY COMPARISON

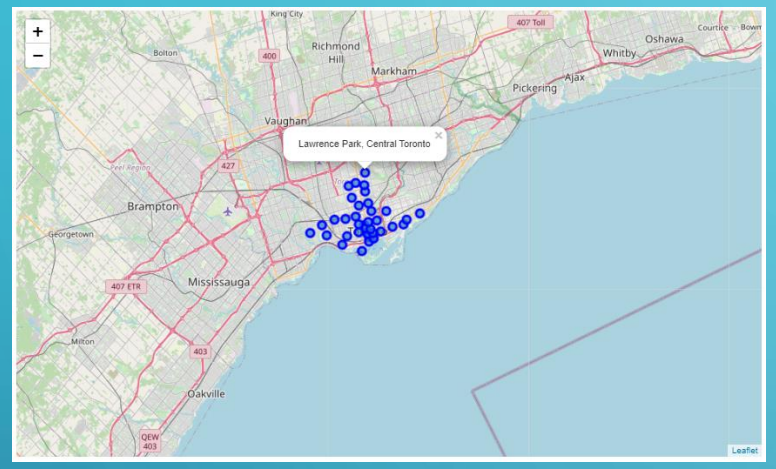
- The Toronto is separated across two csv files.
- The process of data cleaning involved joining these sources and inputting them into the same pandas dataframe.
- After which all boroughs that contained the value “Not assigned” were dropped from the dataframe and filtered to the Downtown Toronto borough.
- The venue data from the Foursquare API was filtered to only include venue name, category, latitude, longitude in Toronto and Ottawa respectively. The Ottawa data required no cleaning.

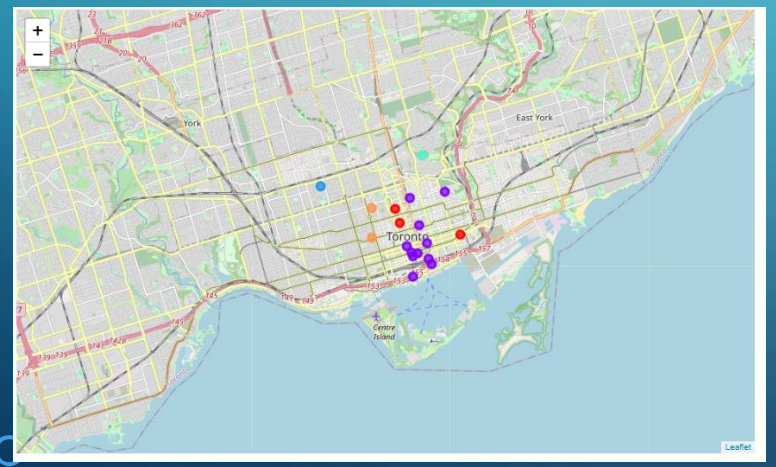
QUESTION 2:

PROBLEM SOLVING

- No additional cleaning was required as this question uses the same Toronto dataset as Question 1.
- It is important to note that in this case the Toronto data was not limited to the Downtown borough.
- The Foursquare API data for this question was filtered to Toronto with an additional search query equal to italian.

K MEANS CLUSTERING – TORONTO (QUESTION 1)



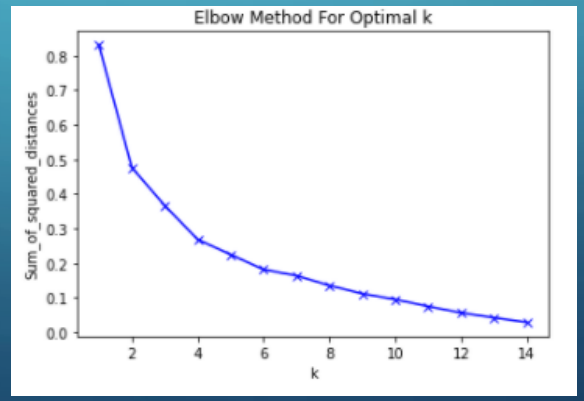


Observations:

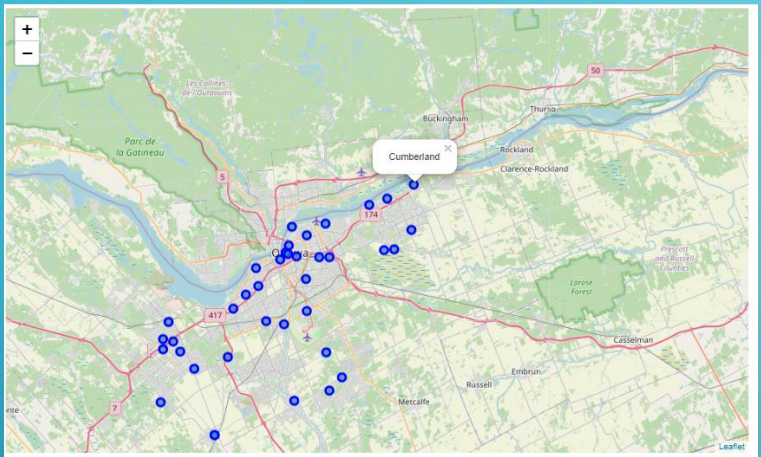
- The dominant cluster is 2. The most common venue in this cluster is a coffee shop which makes sense in a metropolis.
- The performance of the k-means clustering model has separated cluster 1 and cluster 2. However, on closer analysis of the data the most common venue in both clusters is a coffee shop, so they should in fact be a single cluster.
- The remaining clusters have performed well, separating out venues such as parks, airport, grocery stores and cafes.

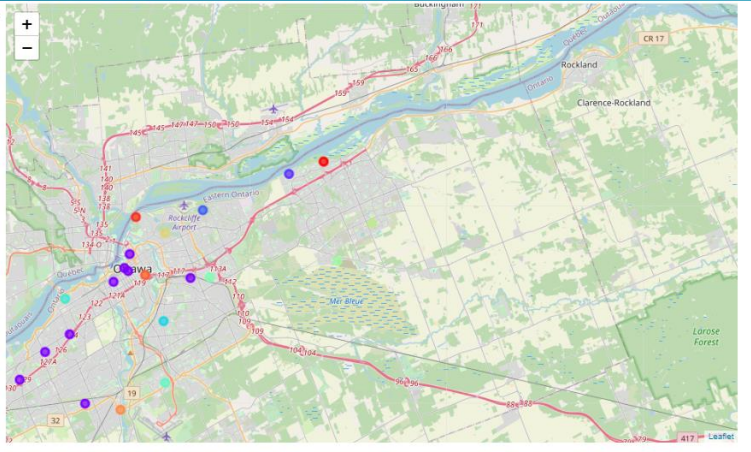
Recommendations:

- The low value and clear elbow of k in the Toronto dataset does demonstrate a relatively robust model.
- The above comment is justified by the success of the k-means algorithm in the Toronto dataset and its categorization of coffee shops.
- I recommend using k-means algorithm when analysing the Toronto dataset



K MEANS CLUSTERING – OTTAWA (QUESTION 1)



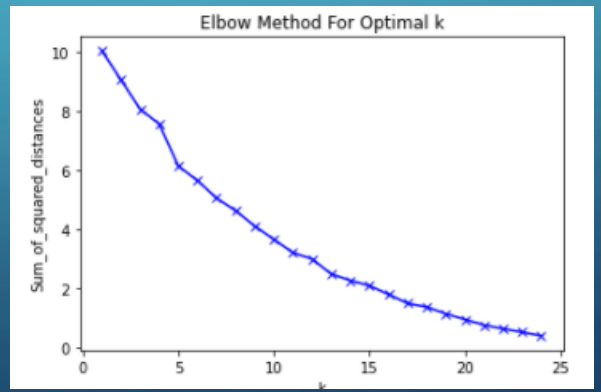


Observations:

- The dominant cluster is again 2 in this dataset. However, the performance of the model is poor in comparison to the Toronto dataset. Cluster 2 contains a mixture of venue categories ranging from coffee shops, gyms, pools and restaurants.
- The remaining clusters have performed substantially better, successfully categorising arenas, bus stations, construction & landscaping, and theme parks.

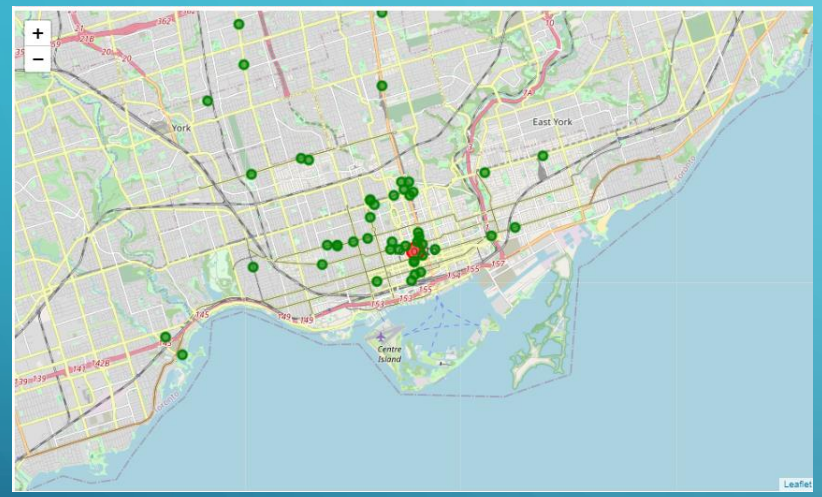
Recommendations:

- The high value of k in the Ottawa dataset promotes overfitting so the k-means method of clustering is not the greatest for predictive analysis of this dataset.
- I recommend using a different clustering algorithm on the Ottawa dataset.



ITALIAN RESTAURANT LOCATION (QUESTION 2)

The spatial distribution is highly important from a competition point of view as an area highly saturated in Italian cuisine will prove detrimental to their business. Therefore, the owner will be looking for an area that has none/few Italian restaurants at present.



Observations:

- The distribution of Italian restaurants in Toronto is concentrated in the city centre, with venues reducing in frequency away from the centre of the city
- The specific locations of Italian restaurants in Toronto does seem to be isolated to a few specific streets – demonstrated by the linear patterns.
- The sporadic distribution of Italian restaurants on the outskirts of the city does indicate that other owners have taken a risk of building away from the city centre

Recommendations:

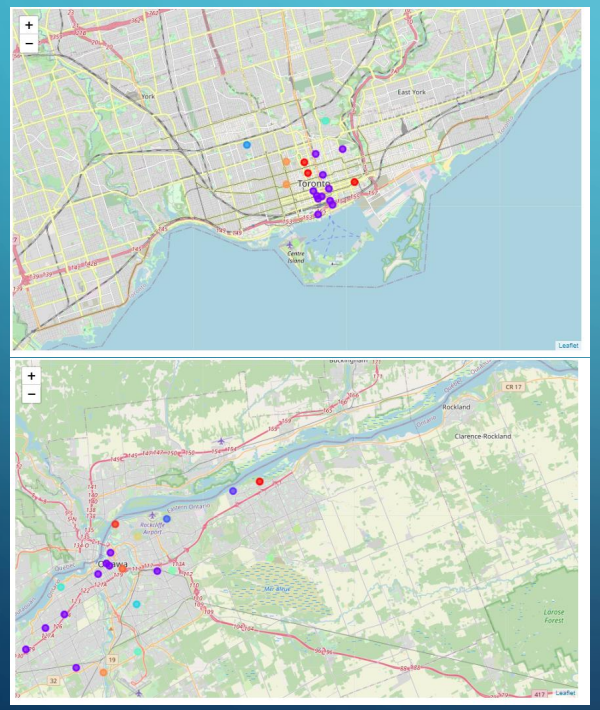
- The location in which there is the most Italian restaurant competition is clearly the city centre. As the restaurant owner wants to minimise his competition building the new restaurant in this location would not be recommended.
- Building on the outskirts of the city is a more risky endeavour are you could become isolated, and have minimal traffic into the restaurant.
- I recommended choosing a location outside of the city centre and perform some additional analysis such as:
    - Population movement – are there areas of high intensity on the outskirts of the city? Where are the choke points?
    - Local population concentration – if the owner wants to only attract local customers.
- Combing the additional recommended analysis will give the owner a more complete picture of the optimum location to build his restaurant.

CONCLUSIONS

QUESTION 1:

NEIGHBOURHOOD/CITY COMPARISON

- The neighbourhoods of Downtown Toronto and Ottawa a very dissimilar shown by the k-means algorithm.



QUESTION 2:

PROBLEM SOLVING

- The best location for the owner of the Italian restaurant with his given criteria of little competition is on the outskirts of the city centre.