**Segmenting and Clustering Neighbourhoods in Toronto**

Applied Data Science Capstone Project

Amira Kaskasoli

**Introduction: The Business Problem**

A client is looking to move to Canada, specifically Toronto, and would like advice on which neighbourhoods he should narrow his search to based on given criteria.

Toronto is a very large city spread across 630 square kilometres, and it is known for being a hub for finance, business, arts, culture and entertainment. With a population of over 6 million people and 27.5 million tourists visiting annually, it is considered to be in the top 5 liveable cities in the world, making it is a very attractive place to consider living.

Property costs in Toronto are among the highest in Canada. The client understands this and is willing to pay for the right location.

The client would like his new home to be near coffee shops, exercise facilities and his favourite type of cuisine: Italian food. House price is also a consideration for the client but at this point he is looking for information on which neighbourhoods to consider.

The client is looking to move ASAP so current data on Toronto is appropriate to review.

Using data analytics methodology and machine learning techniques such as clustering, this project will provide an answer to the client’s question regarding the most appropriate neighbourhood to live in in Toronto.

**The Data**

To solve the client’s problem, this report will be using data from various sources. The original data source will be a Wikipedia page listing the various boroughs and their corresponding neighbourhoods and postcodes in Toronto. A second data source containing geospatial data of Toronto will then be inputted and merged with the dataframe of neighbourhoods. The resulting dataframe will contain details of neighbourhoods and boroughs, alongside respective coordinates of each location. Folium will be used to visualise the data on a map of Toronto. Finally, Foursquare location data will be used to explore the neighbourhoods in Toronto and make informed statements about the data to provide to the client.

**Methodology**

The initial data source was a Wikipedia page containing a list of boroughs, neighbourhoods and postal codes in Toronto.

Python was used with web scraping techniques to load the data into a pandas dataframe. Any postal codes missing an assigned borough or neighbourhood were removed from the dataset. Next, the python Geocoder package was used to provide latitude and longitude coordinates to each neighbourhood in a second dataframe.

Once the two dataframes were merged, Folium was used to visualise a map of Toronto, with markers added to represent each neighbourhood’s location.

Foursquare location data was then used to retrieve popular venues in Toronto within a 500m radius. To do this, a Foursquare Developer Account was created to obtain the required ID and Key, and these were used to make API calls to Foursquare, passing the coordinates of neighbourhoods in the dataset using a python loop. From the resulting JSON file, the venue name, category, latitude and longitude were extracted. Using one hot coding, the category and occurrence of each venue was determined. The top 10 venues for each neighbourhood was then found by defining a function to return the most common venues.

A new dataframe was then created with each neighbourhood’s top 10 venues shown.

At this stage, the client’s requirements were taken into consideration and the dataframe was checked for any neighbourhoods which have all the following specifications in the top 10 venues: exercise facilities, Italian restaurants and coffee shops. These neighbourhoods are shown on a map using Folium as this is an easy visualisation tool for the client.

A second method of finding neighbourhoods with the required criteria was done as a more accurate method of narrowing down the neighbourhoods. This method was K Means Clustering. Five clusters were created and were analysed in turn to check for the presence of the client’s criteria.

The data was provided to the client as lists of appropriate neighbourhoods and associated venues as well as Folium maps to visualise the locations.

**Results**

The initial dataframe scraped from Wikipedia, once cleaned, contained 103 rows of postal codes in Toronto. This was visualised using Folium, as shown in Figure 1.

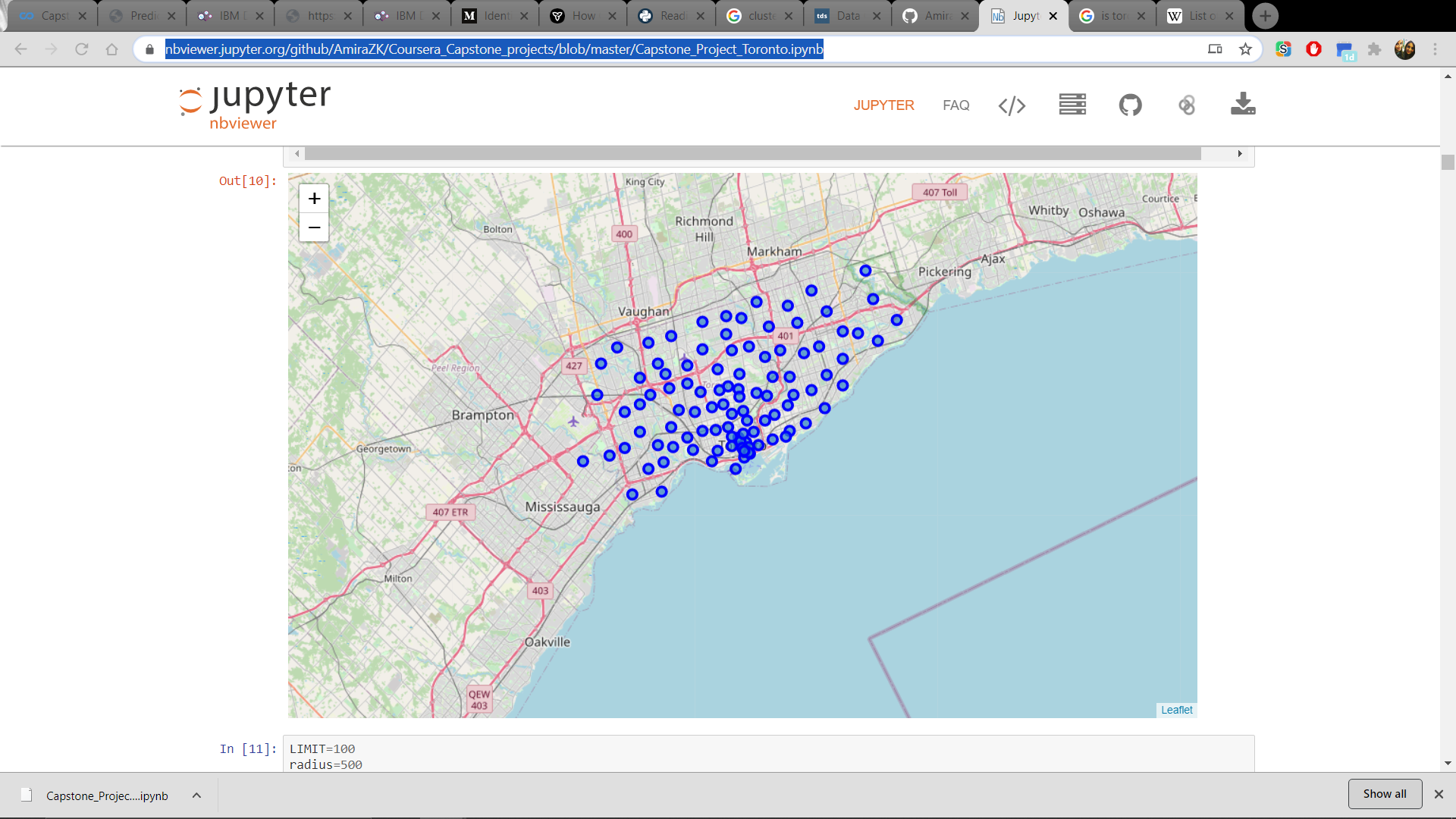


Figure 1: Mapped Neighbourhoods in Toronto

The initial analysis of venues for each neighbourhood showed that three neighbourhoods all had Italian Food and Gym listed as top 2 venues as shown in Figure 2.

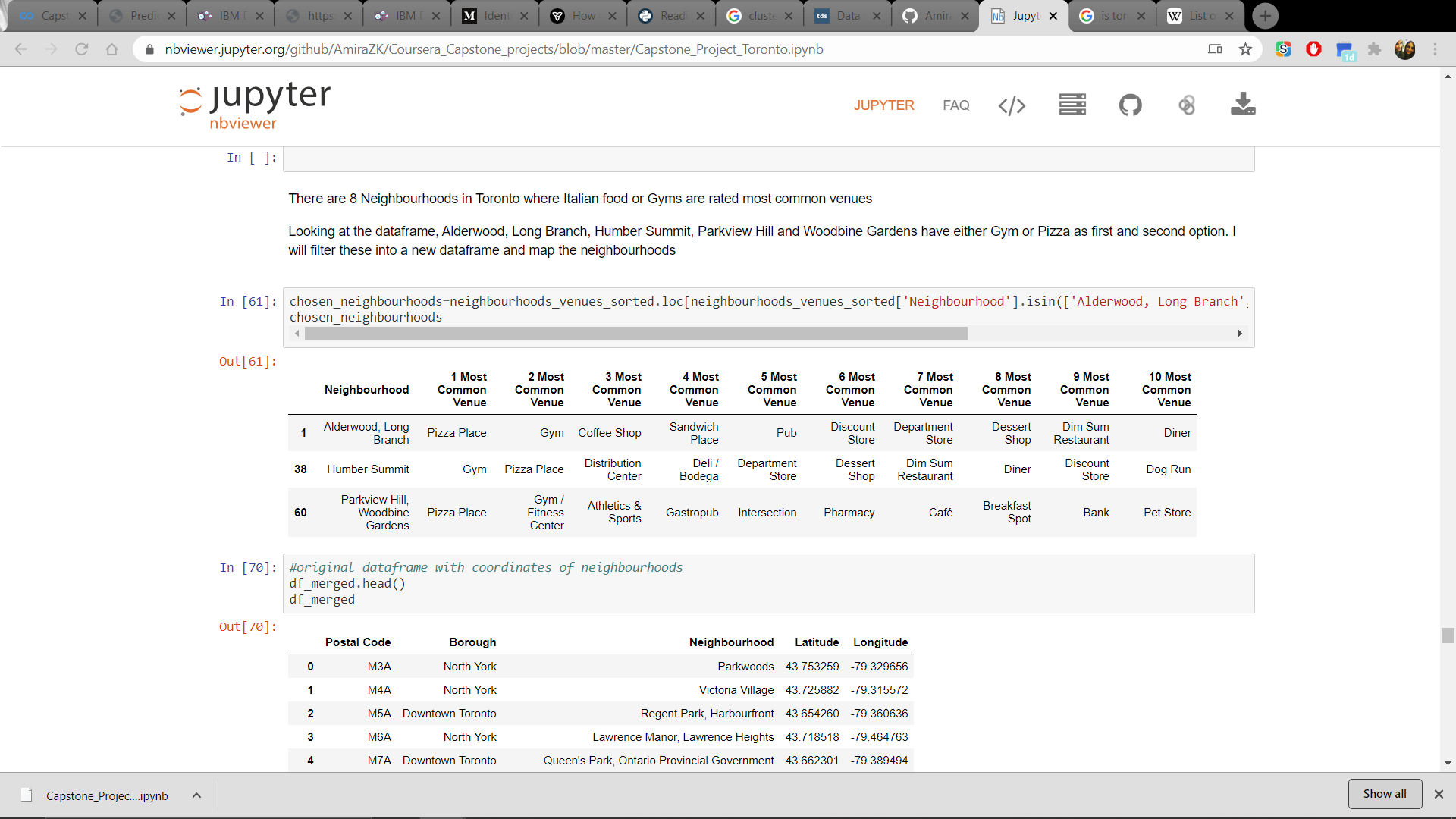


Figure 2: Chosen Neighbourhoods based on top venues

K-means clustering was used to categorised all of the neighbourhoods into 5 main clusters based on common venues within each neighbourhood.

* Cluster 1: 1 neighbourhood, did not specifically contain any of client’s requirements
* Cluster 2: 86 neighbourhoods, with ‘Italian, ‘Café’, ‘Coffee Shop’, ‘Gym’, ‘Yoga Studio’, ‘Athletics & Sports’ appearing frequently across the dataset.
* Cluster 3: 1 neighbourhood with Yoga being a most common venue
* Cluster 4: 2 neighbourhoods, ‘Yoga Studio’ and ‘Coffee Shop’.
* Cluster 5: 10 neighbourhoods, ‘Park’, ‘Yoga Studio’ and ‘Coffee Shop’.

The clustered neighbourhoods were visualised using Folium and can be seen in Figure 3 below.

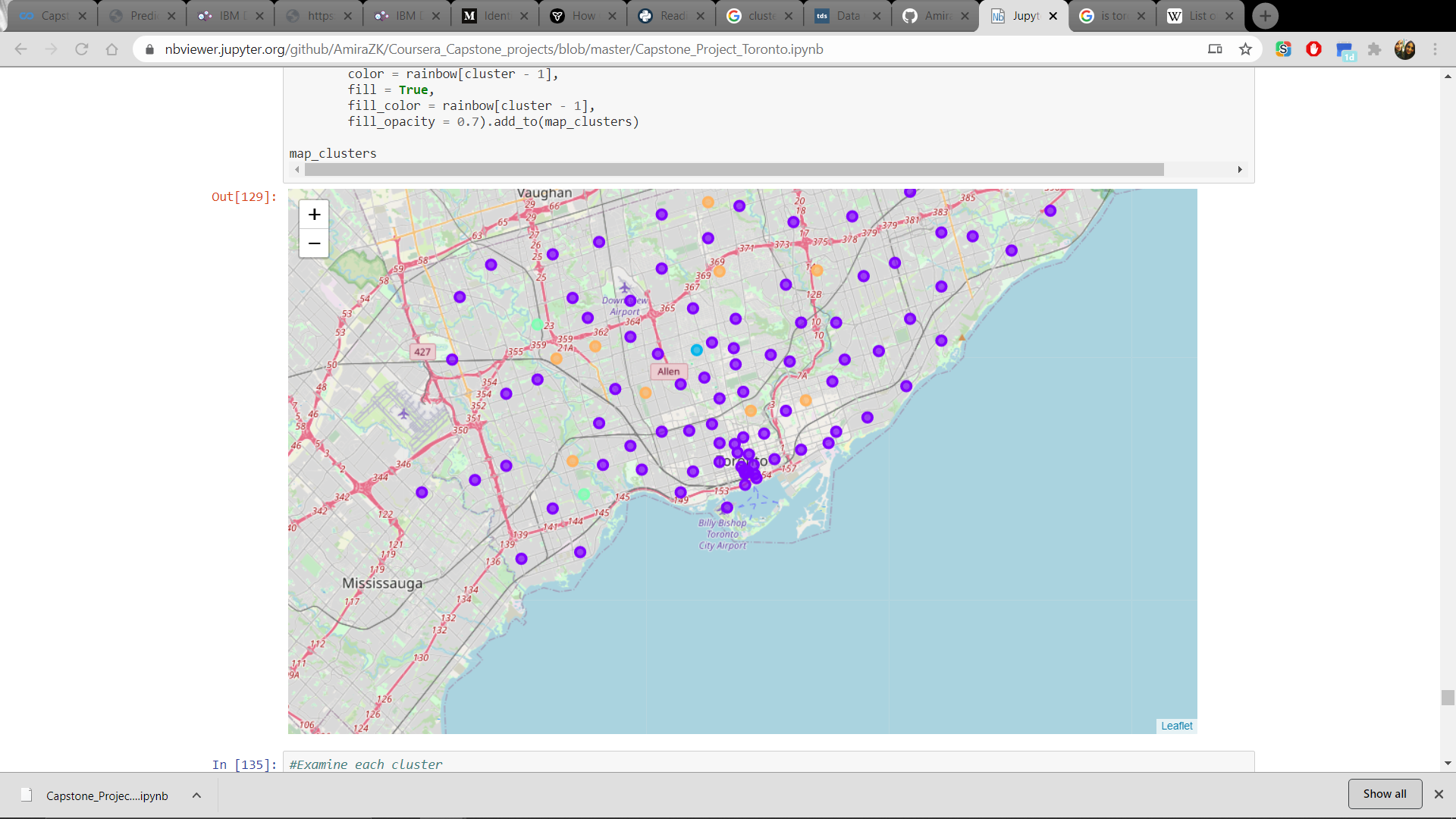


Figure 3: K Means Clustered Neighbourhoods

**Discussion**

In this project, I analysed all neighbourhoods in Toronto based on location and most common venues within each. I used k-means clustering to cluster the neighbourhoods into five distinct clusters based on their similarities to each other and dissimilarities to the other clusters. I then looked at each cluster individually and created a dataframe to show the distinct common venues found from the Foursquare API. I then used this information with the client’s stated requirements in order to recommend the best cluster of neighbourhoods for the client to choose from.

I visualised the clusters on a map in colour coding, so the client is able to clearly see the location of each neighbourhood as well as which cluster it is in, to enable an informed decision on which neighbourhood to move to.

**Conclusion**

The results of k-means clustering showed that Cluster 2 had the best potential of meeting the client’s expectations, as the specified requirements of Italian food, exercise facilities and coffee shops were met across the neighbourhoods within the cluster. This was visualised in purple on the folium map. Although this shows a large number of neighbourhoods, the client will be able to view the location of each neighbourhood on the map and determine which are suitable to look for a house in.

In order to better improve this project, a data source of historic house listings in Toronto could have been used to assess the housing market in each location. This would have allowed for analysis of the most affordable neighbourhoods and would have supported information to provide to the client. Unfortunately, this information was not readily available at this time.

Another interesting future direction for this project would have been to consider Toronto’s position in the current global situation, specifically COVID-19. If the data had been readily available, this project could have determined and mapped the highest current rates of infections in specific neighbourhoods in Toronto, as well as used data science methods to perhaps predict the increasing or decreasing rates.

References

Jupyter Notebook on Github: <https://nbviewer.jupyter.org/github/AmiraZK/Coursera_Capstone_projects/blob/master/Capstone_Project_Toronto.ipynb>

List of Postal Codes Toronto:

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>