Coursera Data science Final Project

Author: Oliver Ignetik

Contents

[Introduction 1](#_Toc9873305)

[Business problem: Opening a coffee shop in Toronto 1](#_Toc9873306)

[Data 1](#_Toc9873307)

[Methodology 2](#_Toc9873308)

[Results 3](#_Toc9873309)

[Cluster 1 – restaurant and food districts 3](#_Toc9873310)

[Cluster 2 – workplace districts 4](#_Toc9873311)

[Discussion 4](#_Toc9873312)

[Conclusion 4](#_Toc9873313)

# Introduction

## Business problem: Opening a coffee shop in Toronto

This project will investigate the best location to open a new coffee shop in the city of Toronto, Ontario. This will depend on a number of factors such as:

* Locations of competing coffee vendors
* Types of venues nearby
* Demographics of the neighbourhood
* Average annual income per household
* Vicinity of workplaces

This project will be of interest to any investors considering getting into the coffee business in Toronto. This project will focus on the distribution of different venues.

# Data

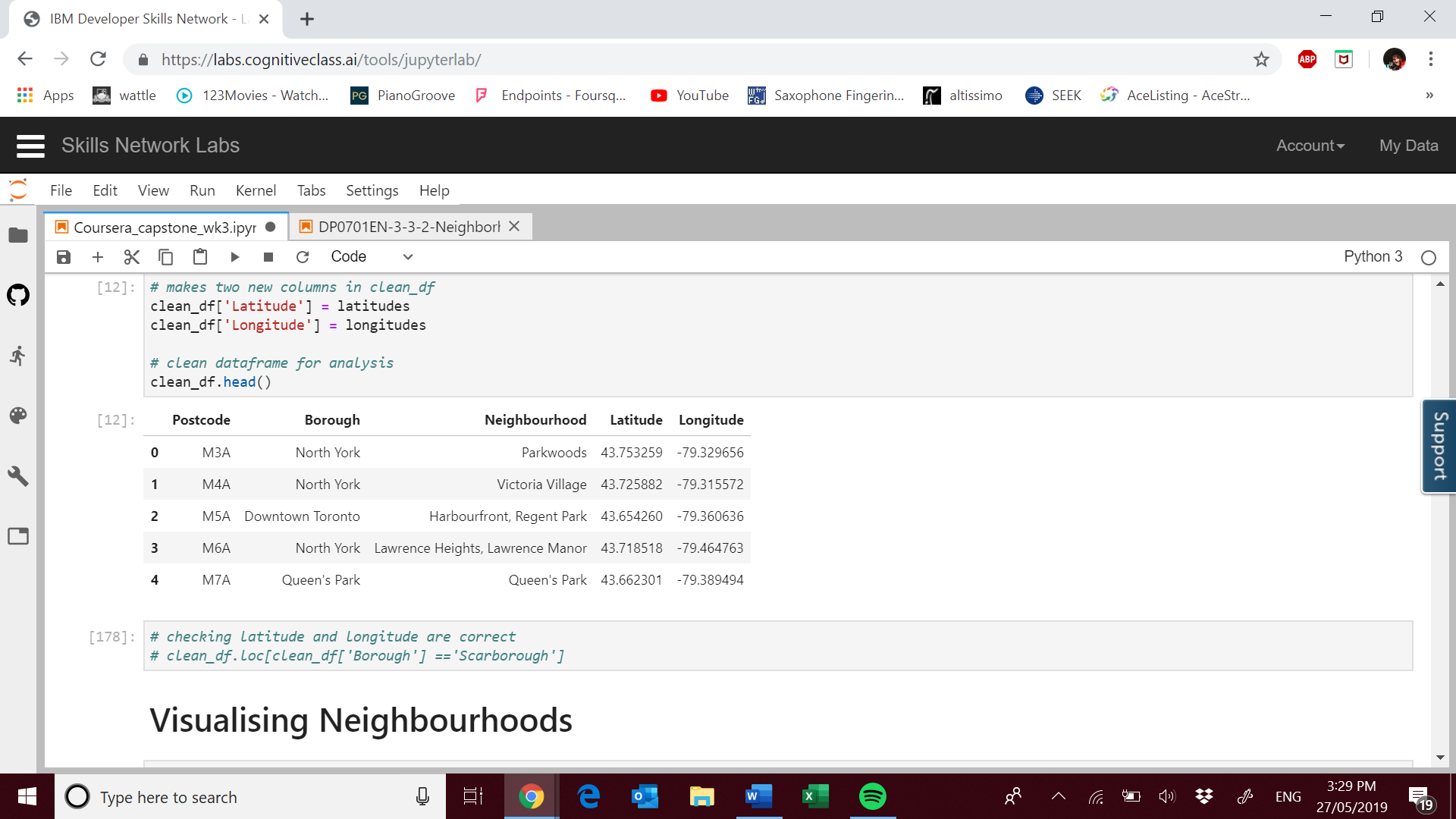
The borough and postcode of each neighbourhood was retrieved from: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

|  |  |
| --- | --- |
| Features | Description |
| Postal code | Unique postcode in Ontario, Toronto |
| Neighbourhood | District within Toronto |
| Borough | Administrative district within Toronto |

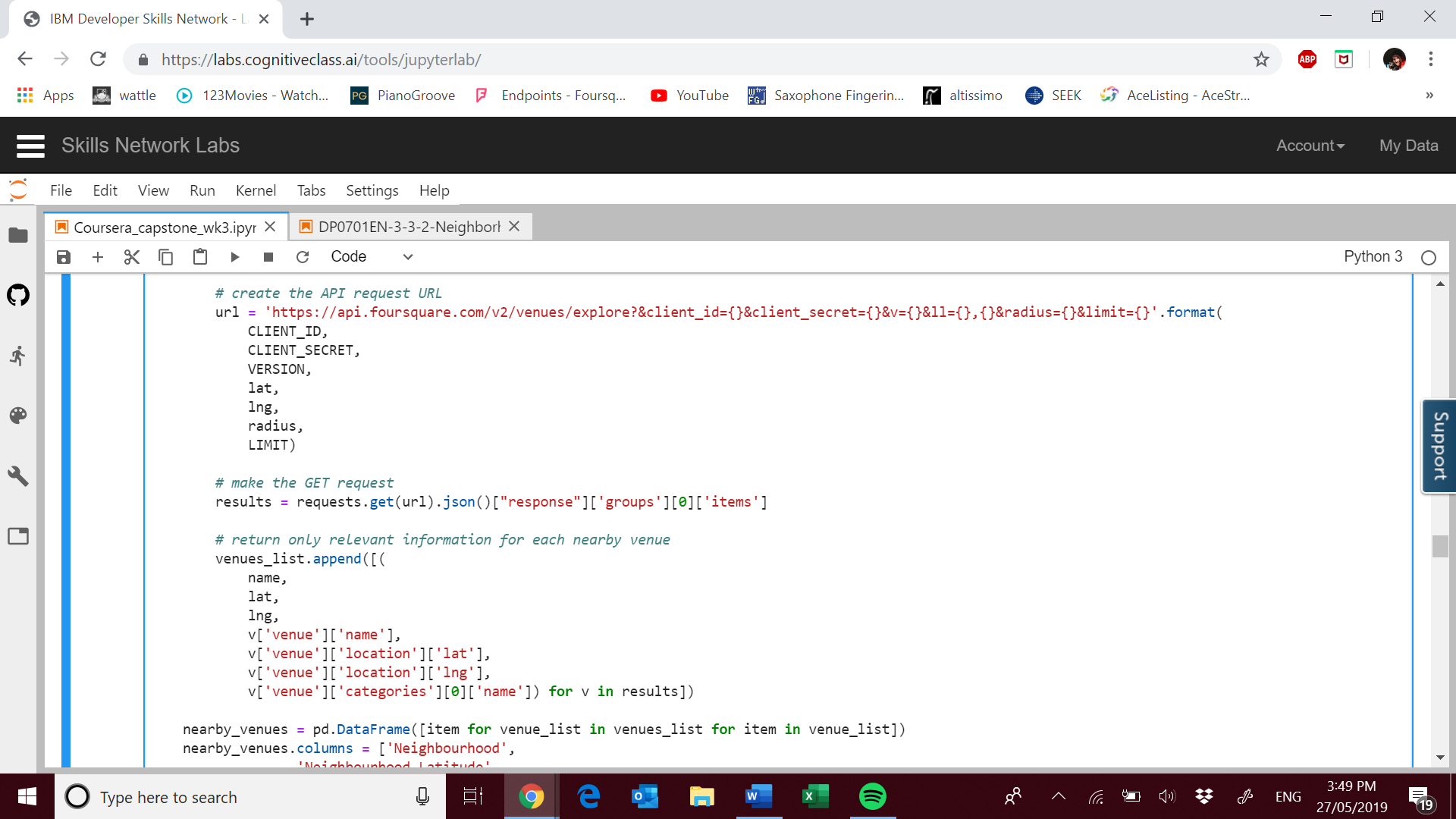
The latitude and longitude of each postcode was retrieved from: <https://cocl.us/Geospatial_data>. The fields of this csv file are discussed in the table below.

|  |  |
| --- | --- |
| Features | Description |
| Postal code | Unique postcode in Ontario, Toronto |
| Latitude | Angular distance from equator |
| Longitude | Angular distance from Greenwich median |

The data from both these sources was filtered and stored in the data frame named clean\_df.



The Foursquare API was used to retrieve information about the venues nearby each postcode. Foursquare requests are in the form of a JSON file. The information of value was extracted from the JSON file by accessing the relevant fields as shown below:



Once the data had been retrieved from the JSON field the rest of the data analysis was completed entirely within the Jupyter Notebook environment.

# Methodology

1. The location data was collected from the sources outlined in the [Data](#_Data) section.
2. Raw csv and html files were converted in dataframes in pandas for further manipulation. This information was collated in the data frame clean\_df.
3. Foursquare was utilized to retrieve the nearby venues for each unique postcode in clean\_df and stored in the data frame Toronto\_venues.
4. Using one hot encoding the venues were regrouped into neighbourhoods ready for use in the clustering algorithm.
5. K-means algorithm was used to cluster the neighbourhoods depending on the types of venues. The number of clusters was set to 5.
6. The output of the k-means algorithm along with the most common venues in each neighbourhood were stored in the data frame toronto\_merged.
7. Using the folium package the clusters were visualized on a map of toronto to allow for transparency of the findings in the report.

# Results

After observing the results of the clustering algorithm it was uncovered that two of the clusters had ideal distributions of venues in each of the neighbourhoods. Figure 1 shows the neighbourhoods from cluster 1 and cluster 2 as dark blue and red respectively. The most prominent feature of both of these clusters was that the frequency of coffee venues in these neighbourhoods was very low.

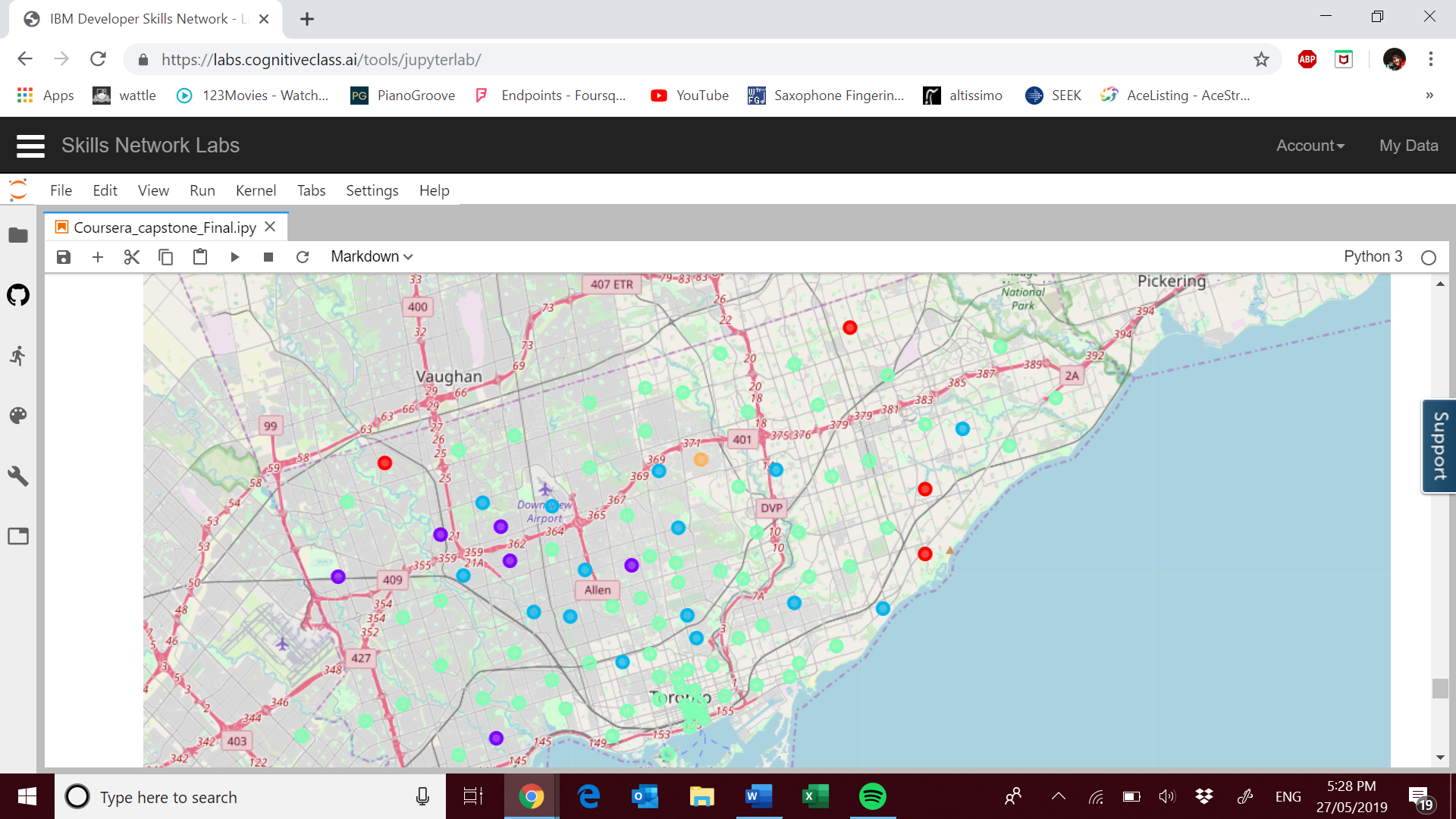


Figure - visualization of clusters

Figure 2 shows the most prevalent veneus in cluster 1 (dark blue markers). As can be seen in Figure 2 it is evident that in each of the neighbourhoods of this cluster the prevalence of coffee vendors is low. However there are a number of dessert shops, donut shops and diners.

## Cluster 1 – restaurant and food districts

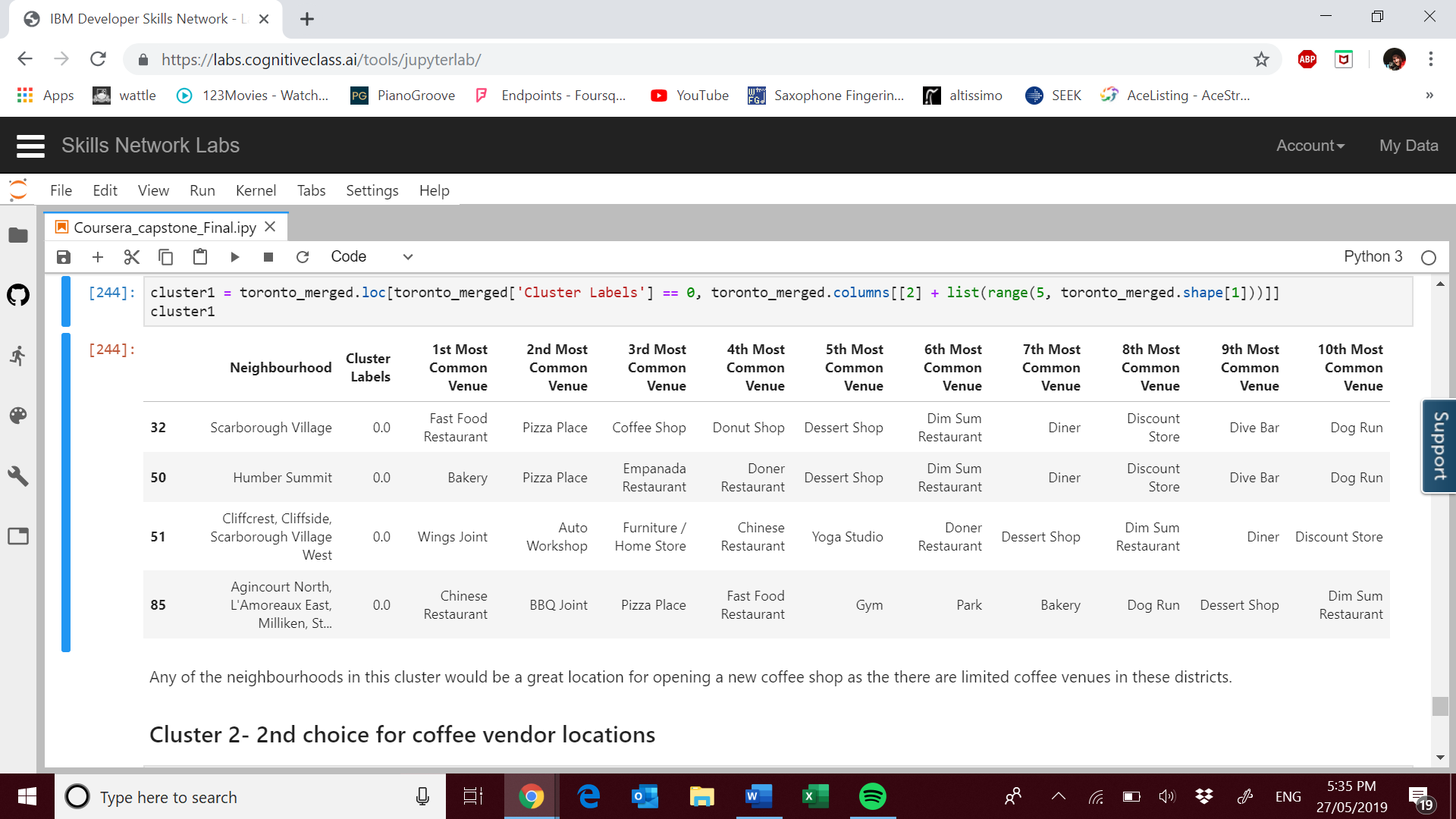


Figure - Cluster 1 most common venues

Figure 3 shows the most prevalent venues in cluster 2 (red markers). As was the case with cluster 1 the prevalance of coffee vendors is very low which is ideal for the business of a new vendor opening in the area. By inspection of the types of venues in these areas it is obvious that compared to cluster 1 there are more workplace venues nearby which may be benificial for prospective clients.

## Cluster 2 – workplace districts

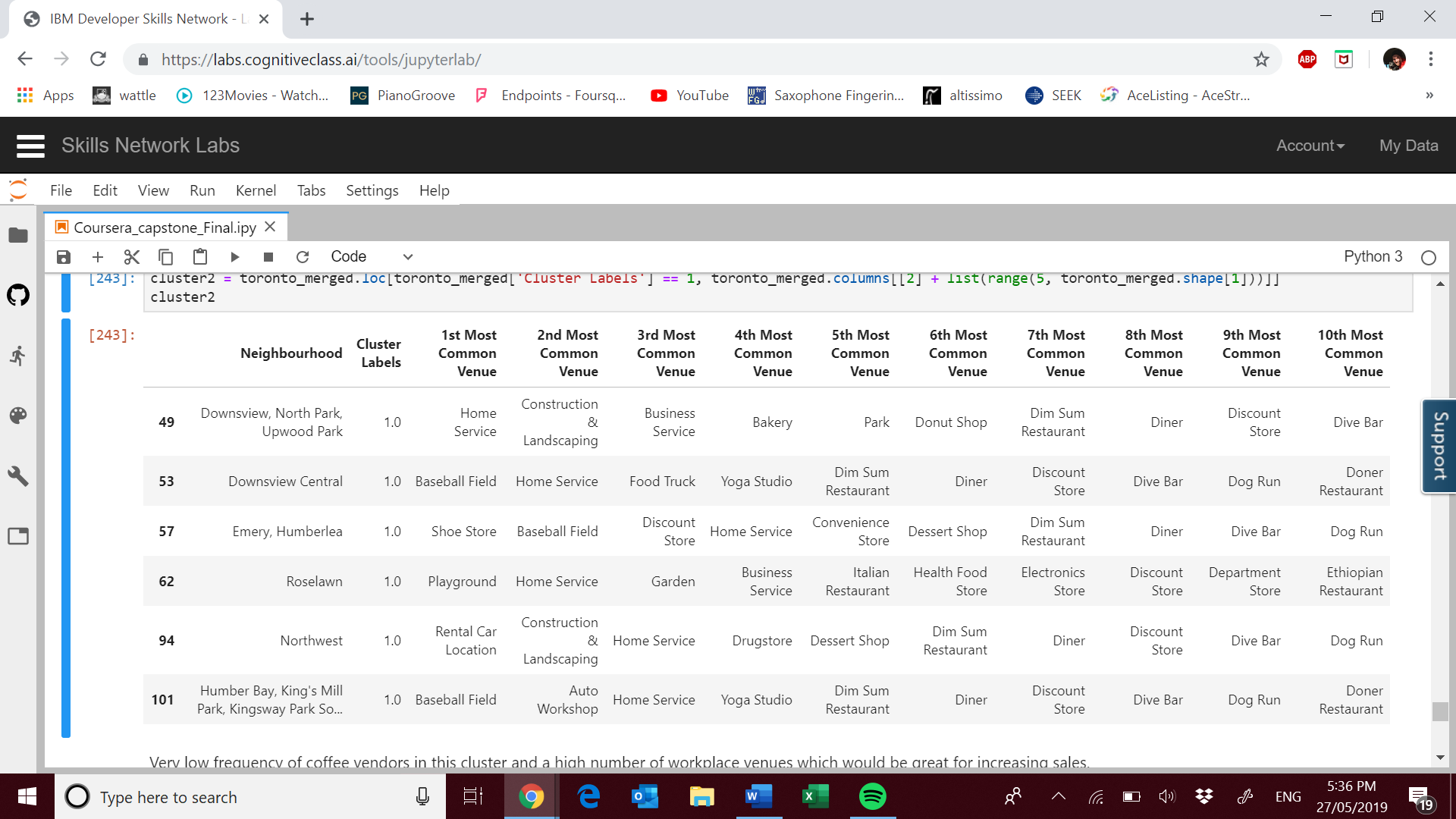


Figure - Cluster 2 most common venues

# Discussion

The clusters identified in the analysis are favorable for opening a new coffee vendor. Although there are some defining features of each cluster and a few districts that should be avoided.

As discussed in the [Cluster 1 – restaurant and food districts](#_Cluster_1_–) section Scarborough Village contains a greater number of coffee vendors then the rest of the cluster. The increased number of coffee vendors in this postal code area may not be catastrophic to business prospects in this area.

Through closer inspection of the venues in cluster 1 donut shops, dessert shops and diners are quite common venues. These venues are all known to serve coffee which could decrease potential revenue in the area. Thus it would seem that this cluster is not in fact less favorable for prospective coffee owners to open a new shop.

In the [Cluster 2 – workplace districts](#_Cluster_2_–) section it was noted that there seems to be a high number of workplace venues in this cluster. This would mean more business for the stakeholders as workers will want somewhere close by to get a coffee during their lunch break. Thus cluster 2 is the most ideal group of neighborhoods to open a new coffee shop in Toronto.

There are several improvements that could be made in the future to the recommendations made in this project. These recommendations do not consider the proximity of the new shop to coffee vendors in adjacent neighbourhoods which are in different clusters. Another improvement to the model could be the inclusion of annual average income in each nieghbourhood. By targeting areas with higher incomes, the stakeholders could maximize their profits.

# Conclusion

Using a k-means clustering algorithms the nieghbourhoods of the city Toronto were grouped into 5 different clusters. By analysis of the types of venues in each it was found that the neighbourhoods in cluster 2 are ideal. The neighbourhoods in cluster 2 have a low frequency of coffee vendors and food outlets and are known for a high number of workplaces.