Battle of the Businesses in a Neighborhood

Applied Data Science Capstone Project

Rohan Loke

# INTRODUCTION

## Building a model to determine best local business to start in a popular Neighborhood

In addition to this, model also provides with most competitive and least competitive business

### Background

Many of the world’s most valuable companies had humble beginnings as start-ups. In the olden days, it was extremely difficult to create a large and successful business without a tremendous amount of capital to open a factory or buy a fleet of trading vessels, for instance. Today ground-breaking innovations can occur in a basement, a garage, or a college dorm. As a result, new start-ups pop up every day all around the world, each of them hoping to get acquired by a larger company or make it big in their own right. However, for every wildly successful start-up, there are thousands which fall into obscurity, which is why start-ups valued at a billion dollars or more are facetiously referred to as “unicorns”, a reference to their elusiveness.

### Problem

In this project I am going to explore top -3 local businesses in the city of Toronto (as an example). This project will help anyone analyse the top businesses in the neighbourhood / Boroughs / Cities. It will help the business owners to analyse which location is best for which type of business. These neighbourhoods that we will select will be the popular neighbourhoods. The business owners can select whether they want to start a business which has huge competition because it is popular business or they want to start a business which has less competition and gain competitive advantage. I am using Toronto as an example. In this city we will break down the analysis as follows:

### Interest

The people who are looking to start a business of their own but don’t know where to start. They can use this model to find out the best neighborhood to start the business and which business to start in that neighborhood.

# Data acquisition and cleaning

### 2.1 Data sources

The data for “Toronto” was found on Wikipedia and the table was scraped. Also the “Geospatial\_Coordinates” data was used which was obtained during the course period. These 2 datasets were used in the model.

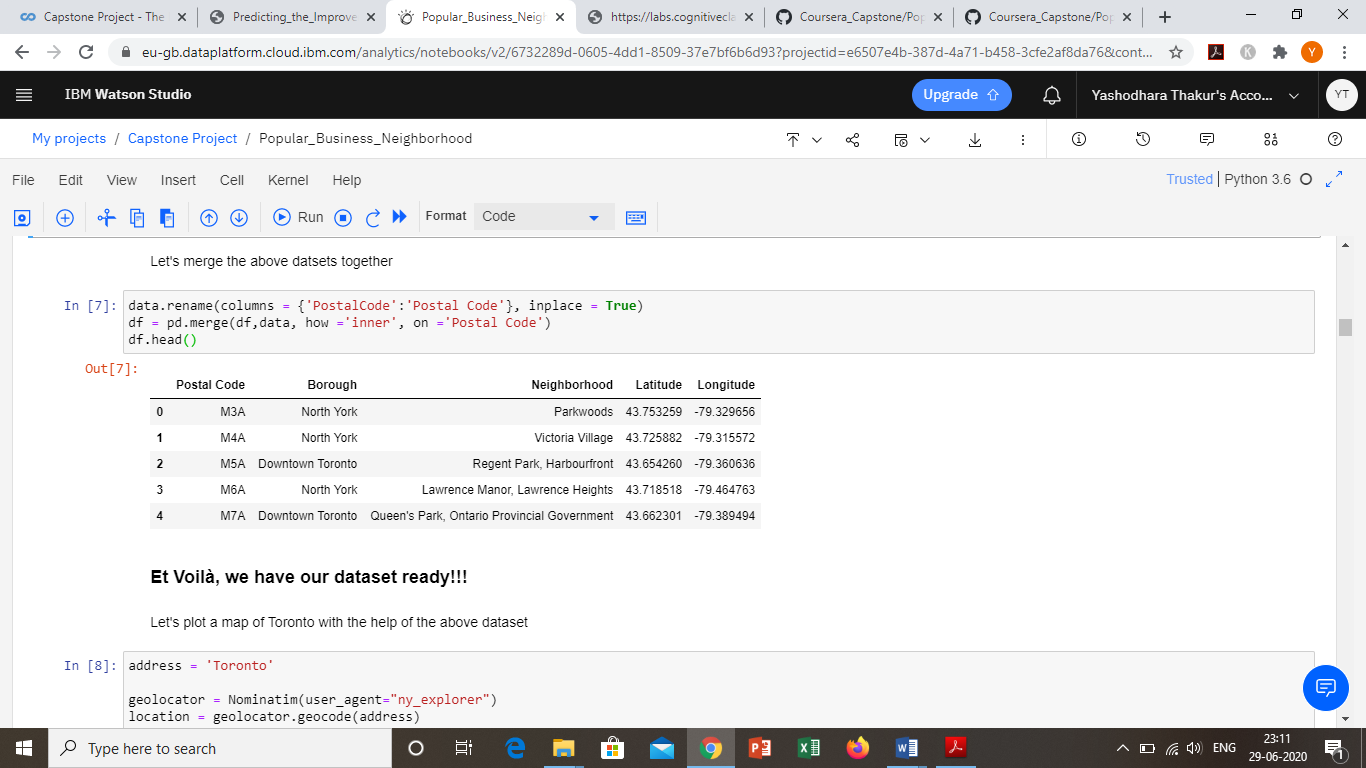
### 2.2 Data Cleaning

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values in the Toronto dataset. I decided to drop the values where Borough were ‘Not assigned’ and mask the Neighbourhood where values were ‘Not assigned’.

Second, multiple entries existed for similar Postal Code with different neighbourhoods. This cause their data to represent multiple samples with incomplete data. I wrote script to extract the unique Postal Code, I grouped the data according to the ‘Postal Code’ and ‘Borough’.

Now I wanted to merge the Toronto dataset with the Geospatial\_Coordinates. To do this I merged them on the basis of ‘Postal Code’.

After cleaning and merging the data set looked like this

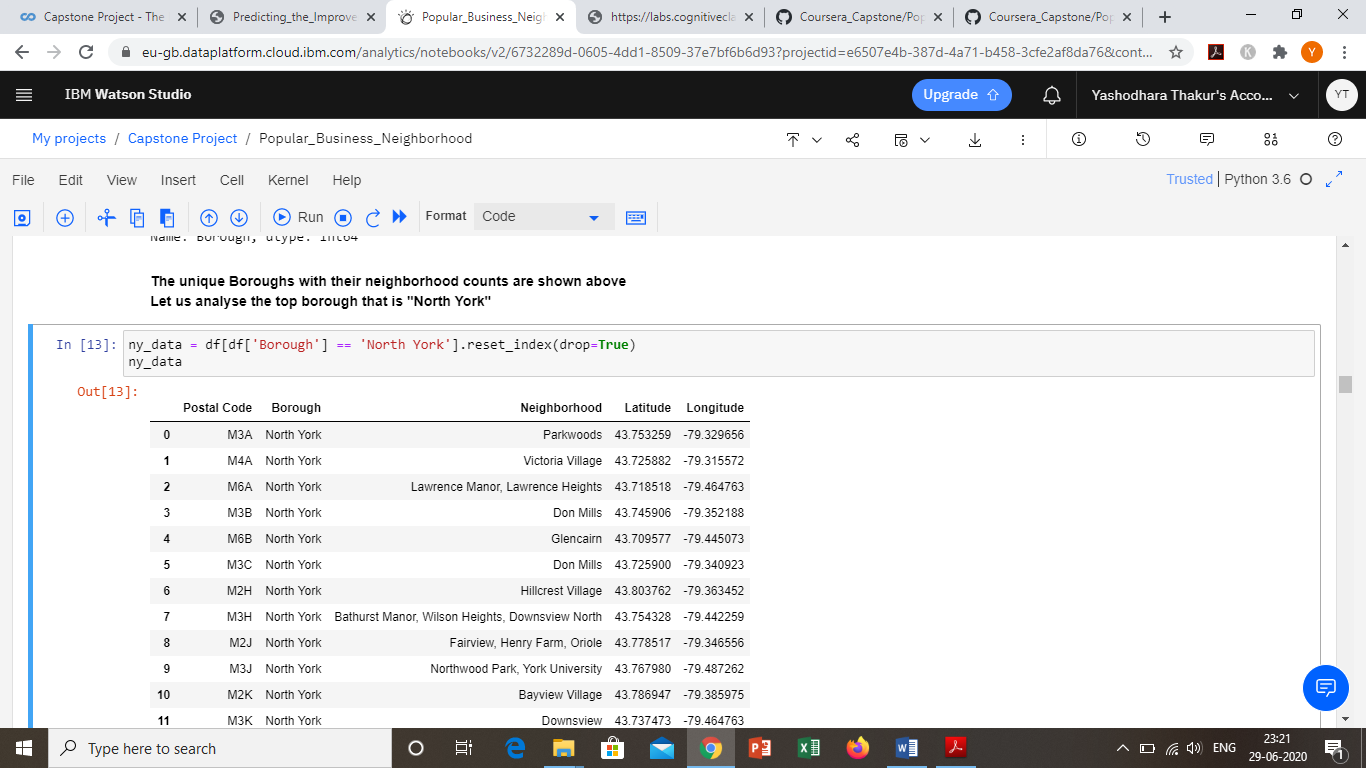


### 2.3 Feature Selection

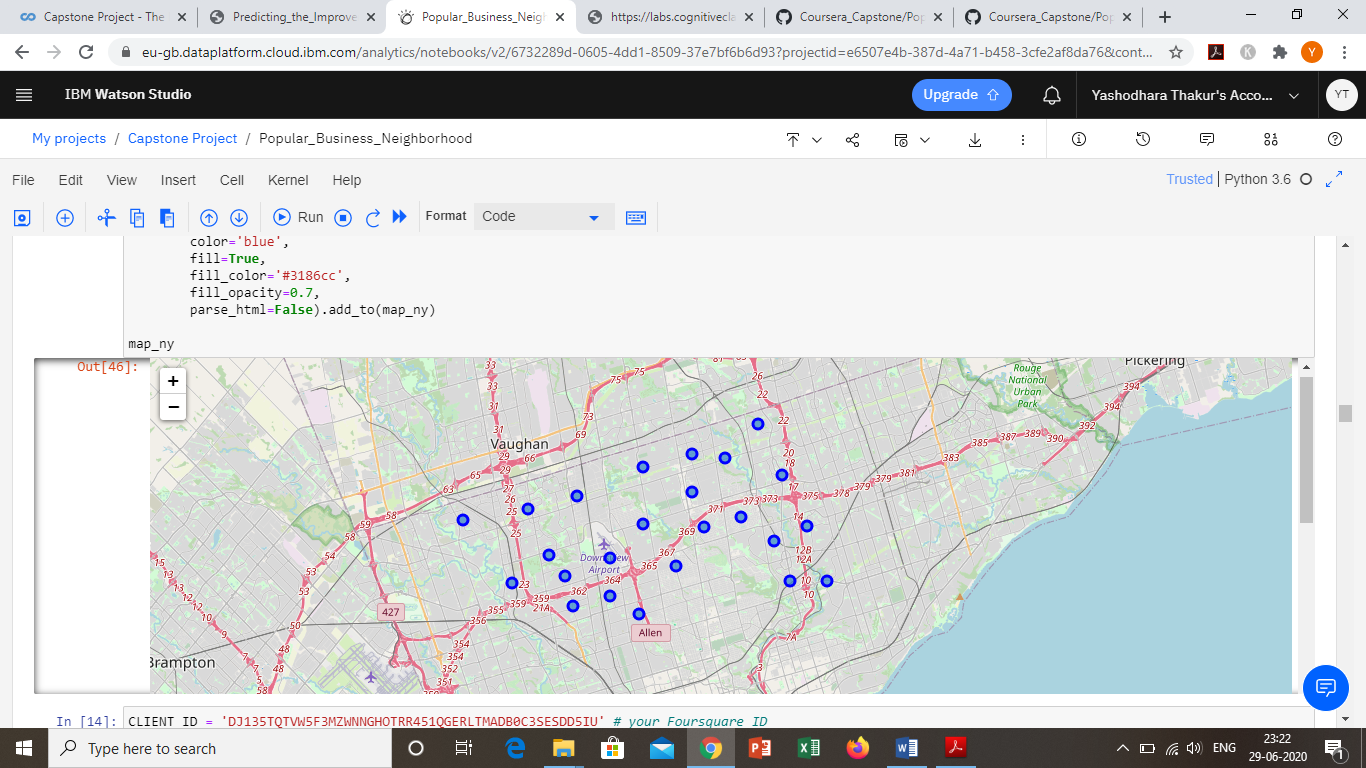
After cleaning the data there were 103 samples and 5 columns. There was no redundancy of features, hence all the features were selected.

# Exploratory Data Analysis

During this analysis I found out that there were 10 boroughs and 103 neighbourhoods. Further analysing the data I found that “North York” was the top borough of all because it had the maximum neighbourhoods than all.

Then I created a cluster which consist of only “North York” data as the below: 

I also created a clusterd map of “North York” for visual representation.



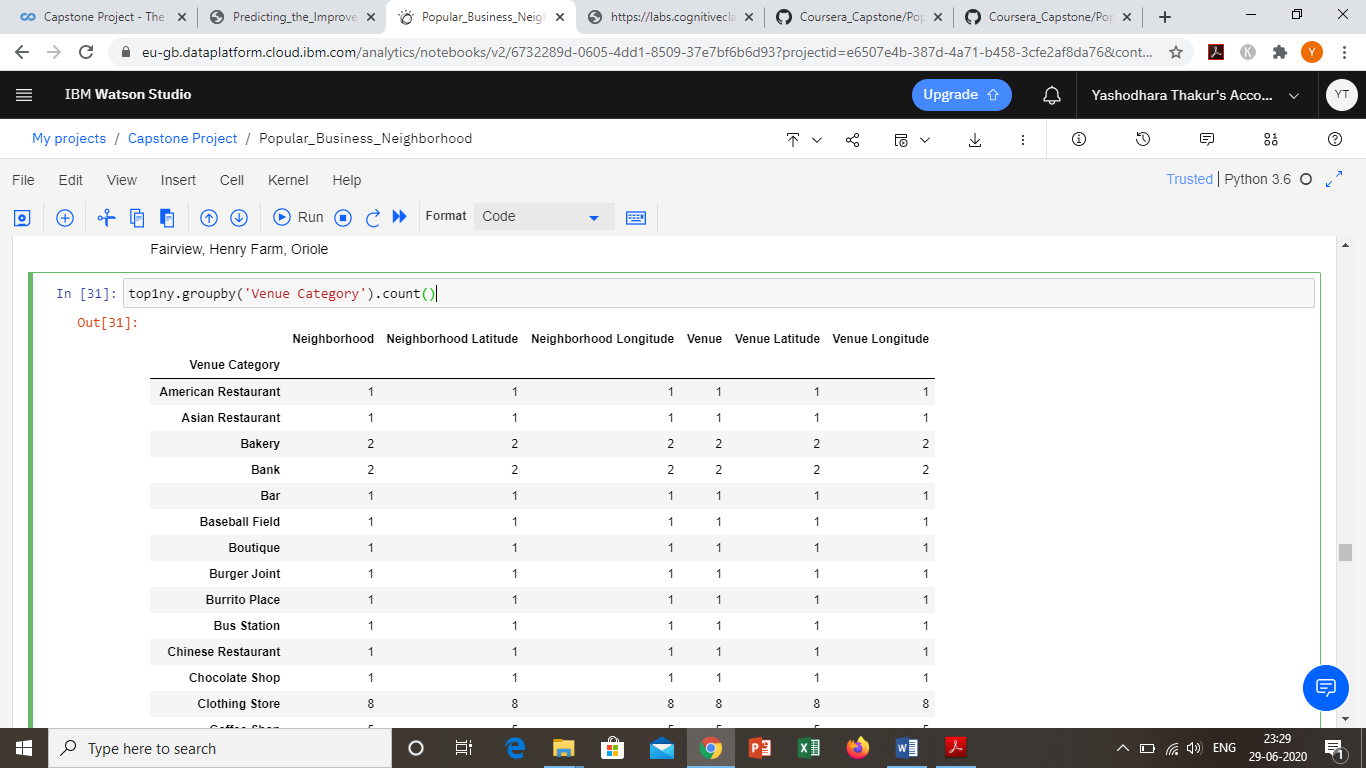
Now we wanted to find the best neighbourhoods in North York.

To do that I used foursquare. I used foursquare to get the venues for all the neighbourhoods. Once this was done I had to get the top 2 neighbourhoods with maximum venues.

After computing the top 2 neighbourhoods were: [Fairview, Henry Farm, Oriole] and [Willowdale and Willowdale East].

Now we had to find the popular businesses in these neighbourhoods.

To do that we first made 2 dataframes which consist the information of these 2 neighbourhoods and then grouped them according to the “Venue Category”, which looked something like this:

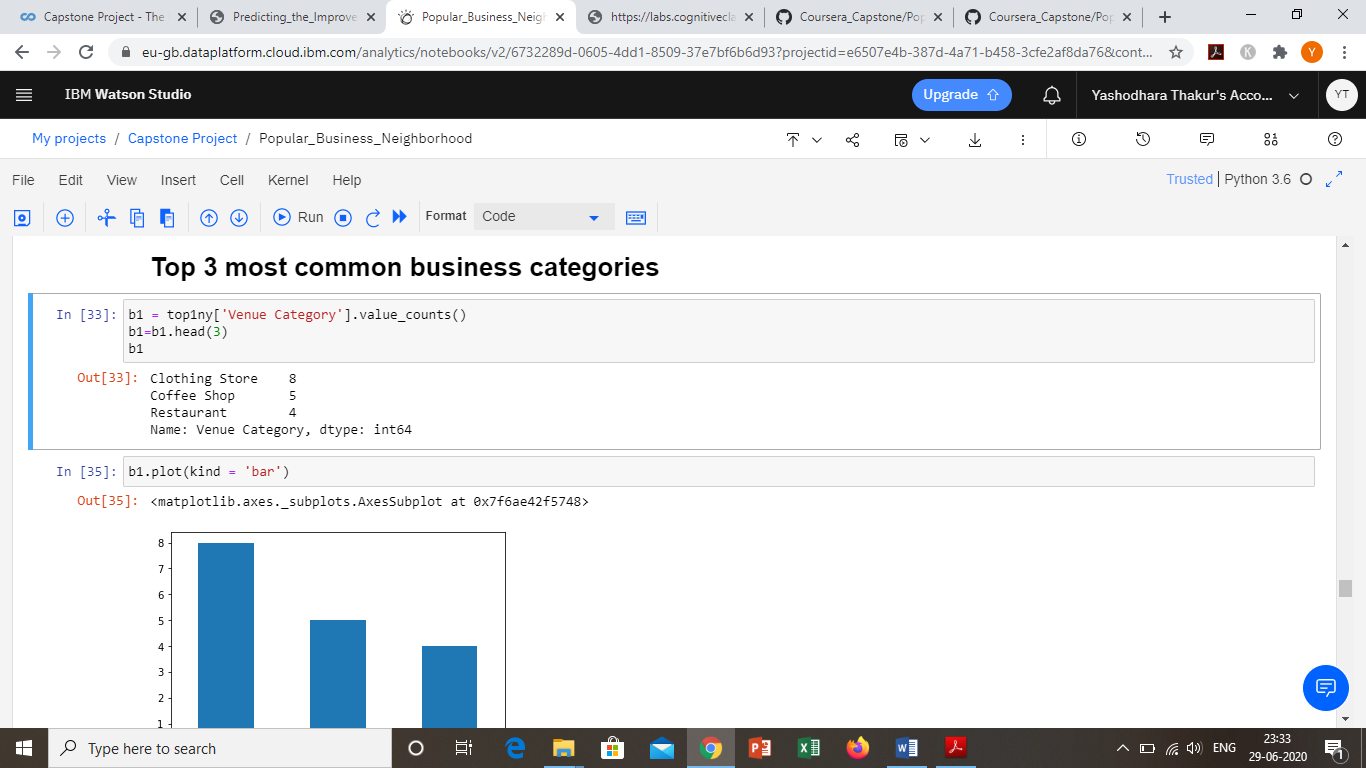


We also saw how many unique local businesses were currently present in the neighbourhood.

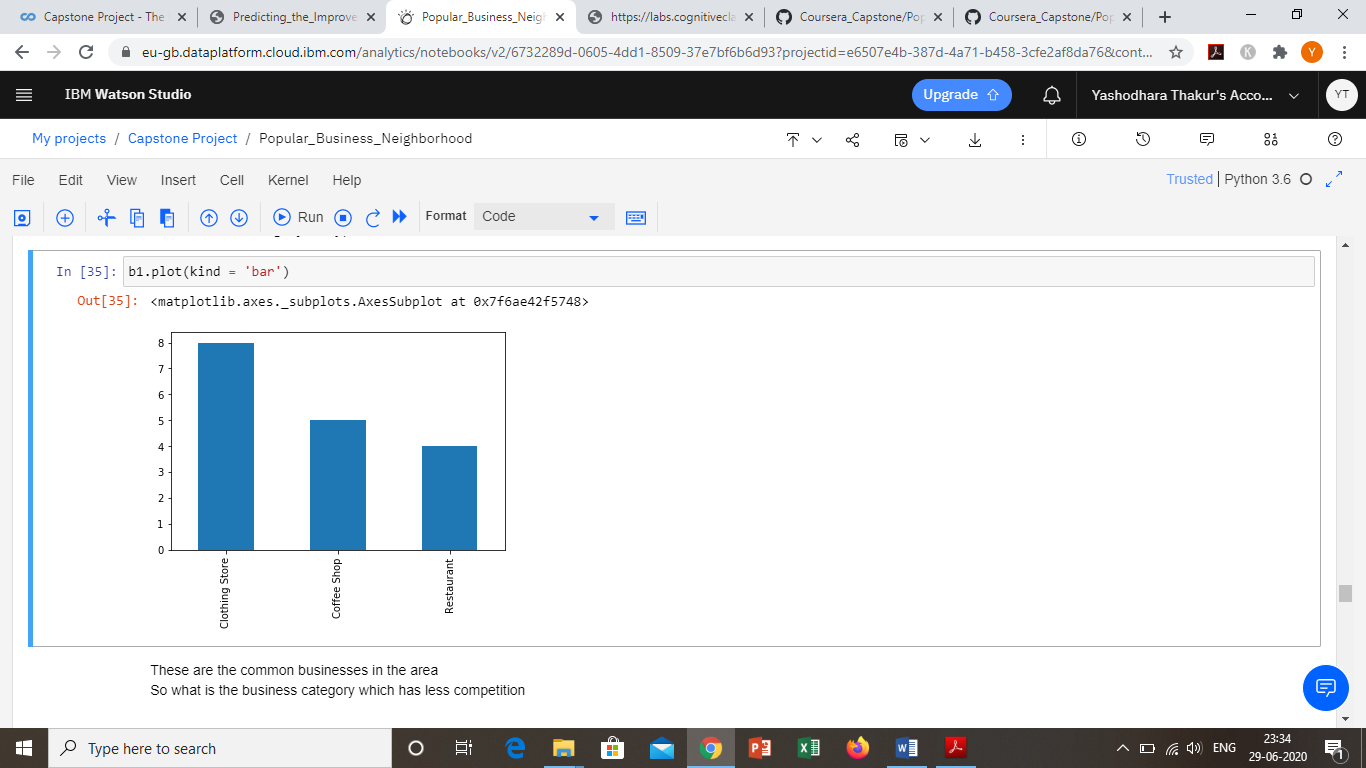
# Results

Now we analysed the data to find the Top 3 most common business categories

The example of the output is shown below

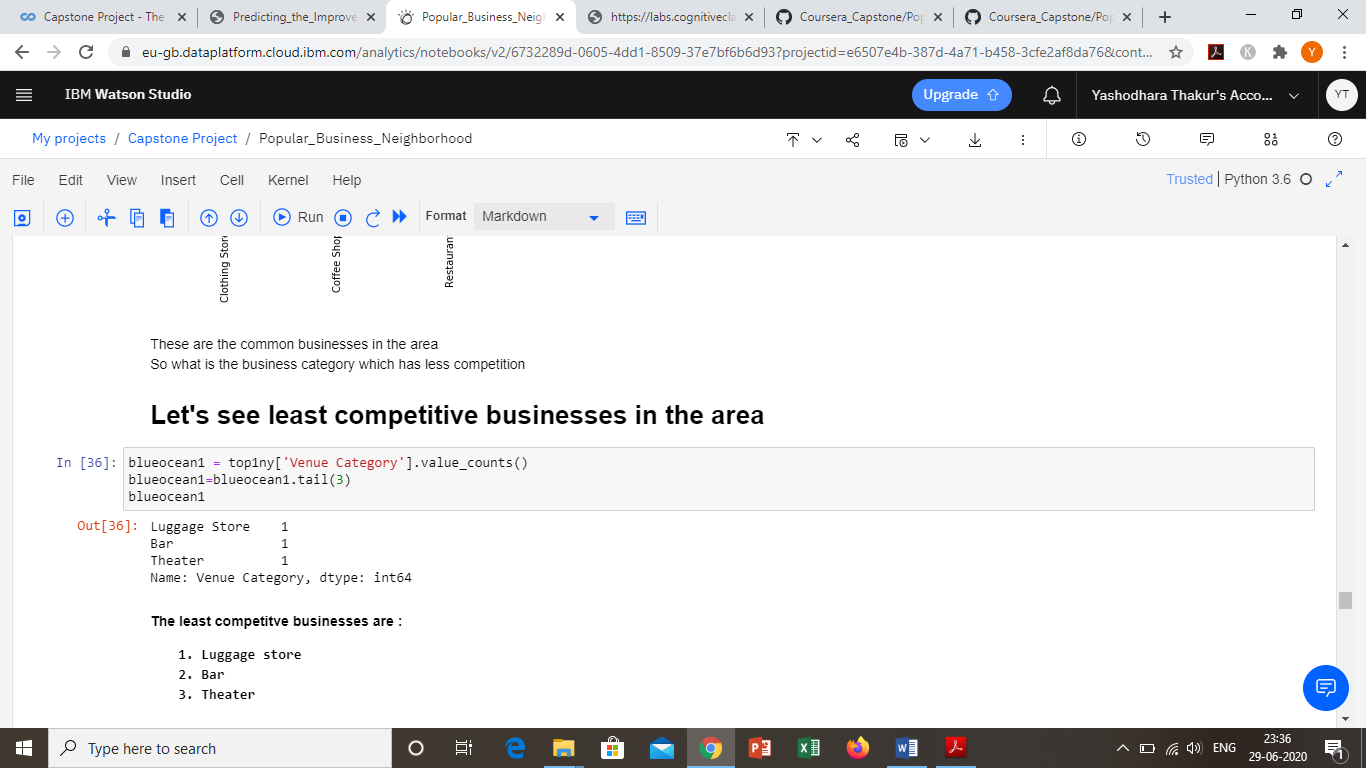


We also plotted a bar graph to represent the above data



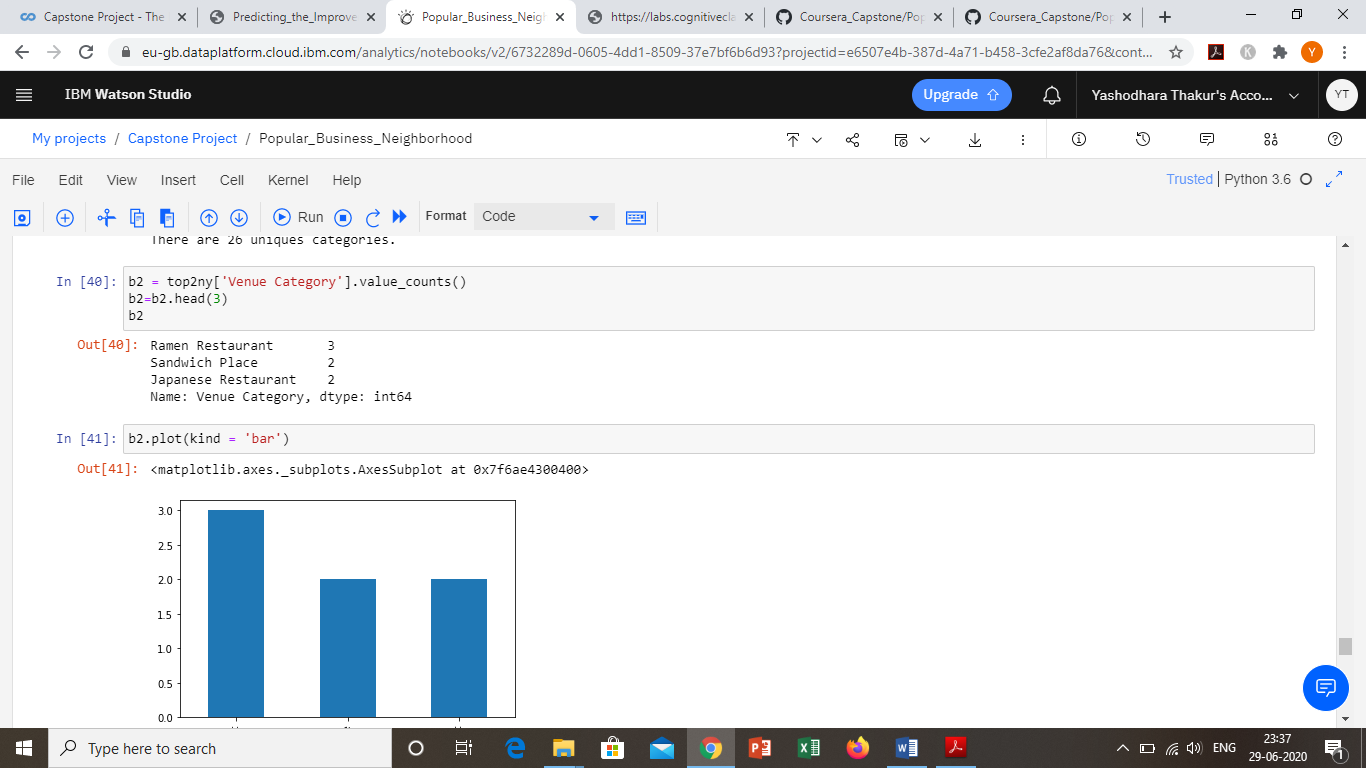
After analysing the popular local businesses we analysed the businesses with less competition and had a better scope.

The result of the analysis is shown below

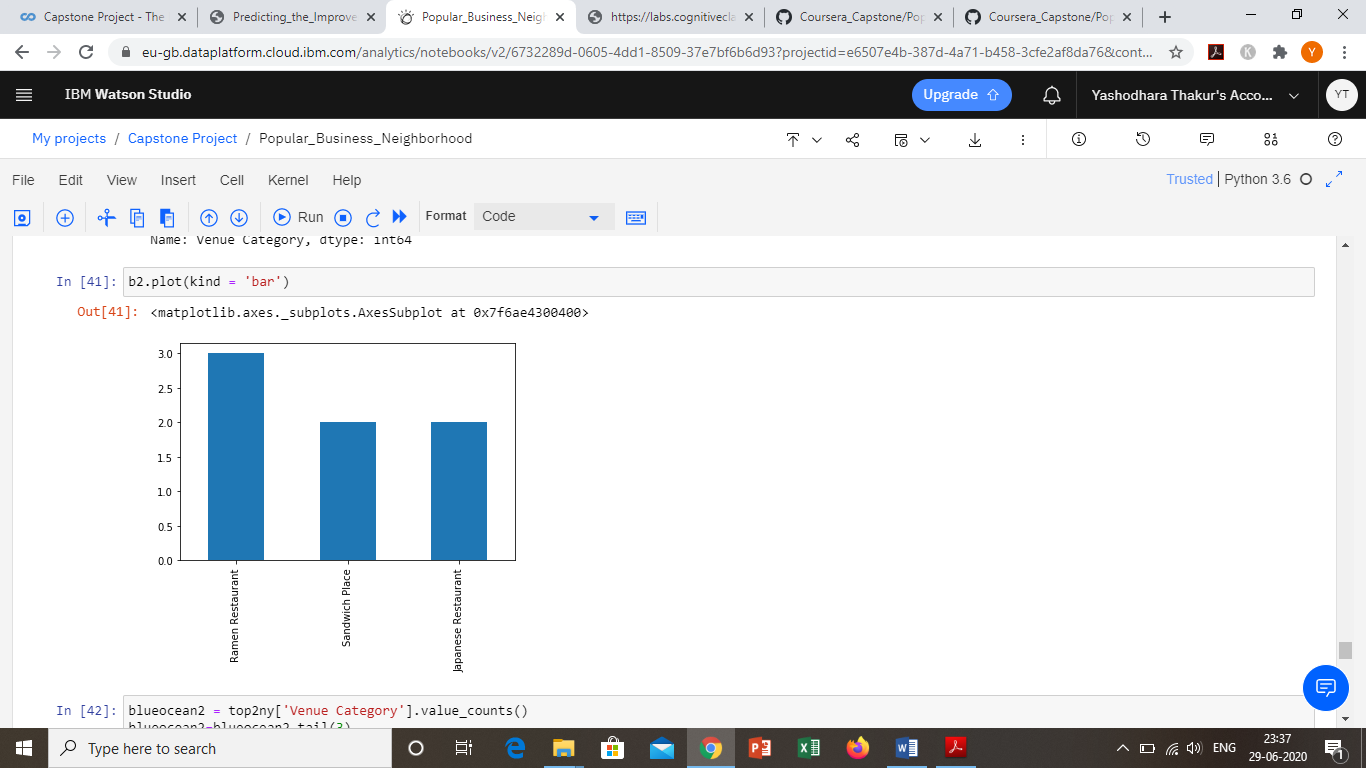


The same analysis was done with the 2nd best neighbourhood and the results were as follows:

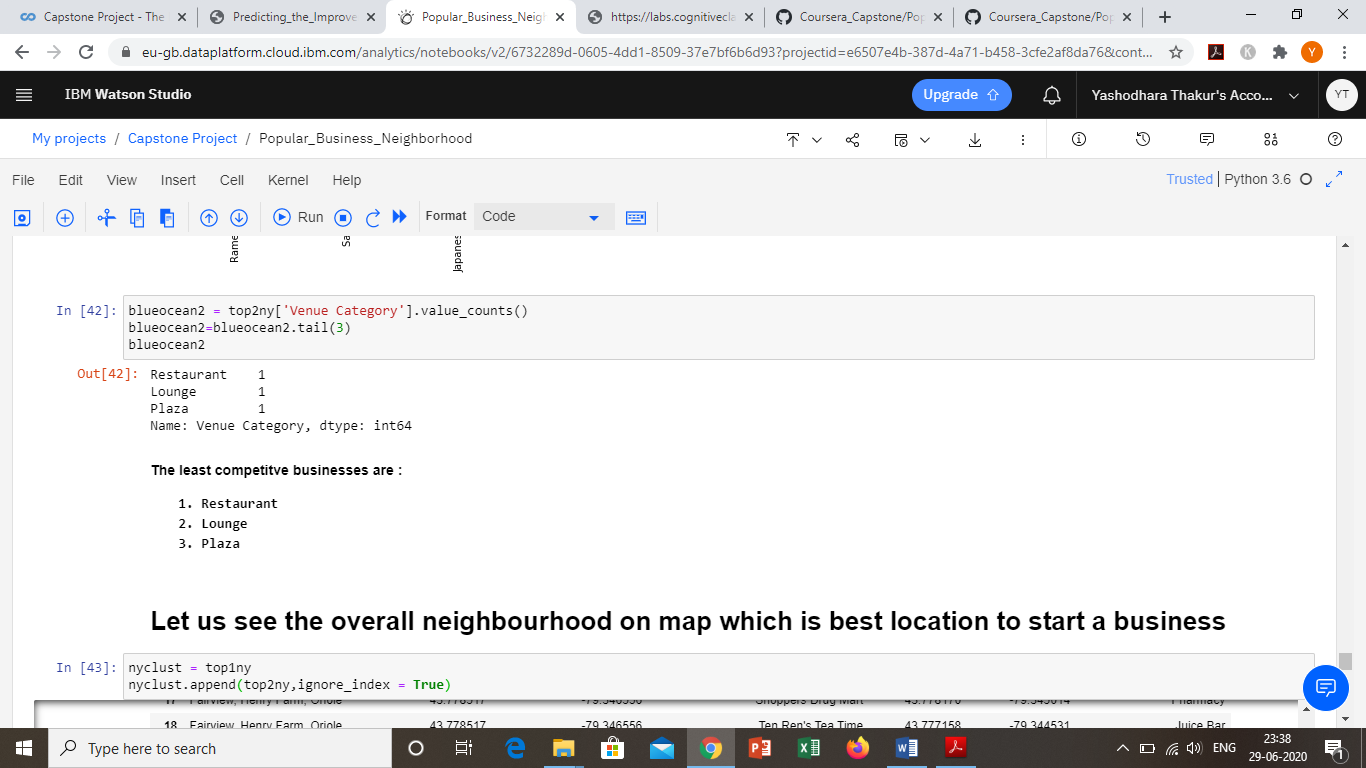
Popular businesses in 2nd neighbourhood



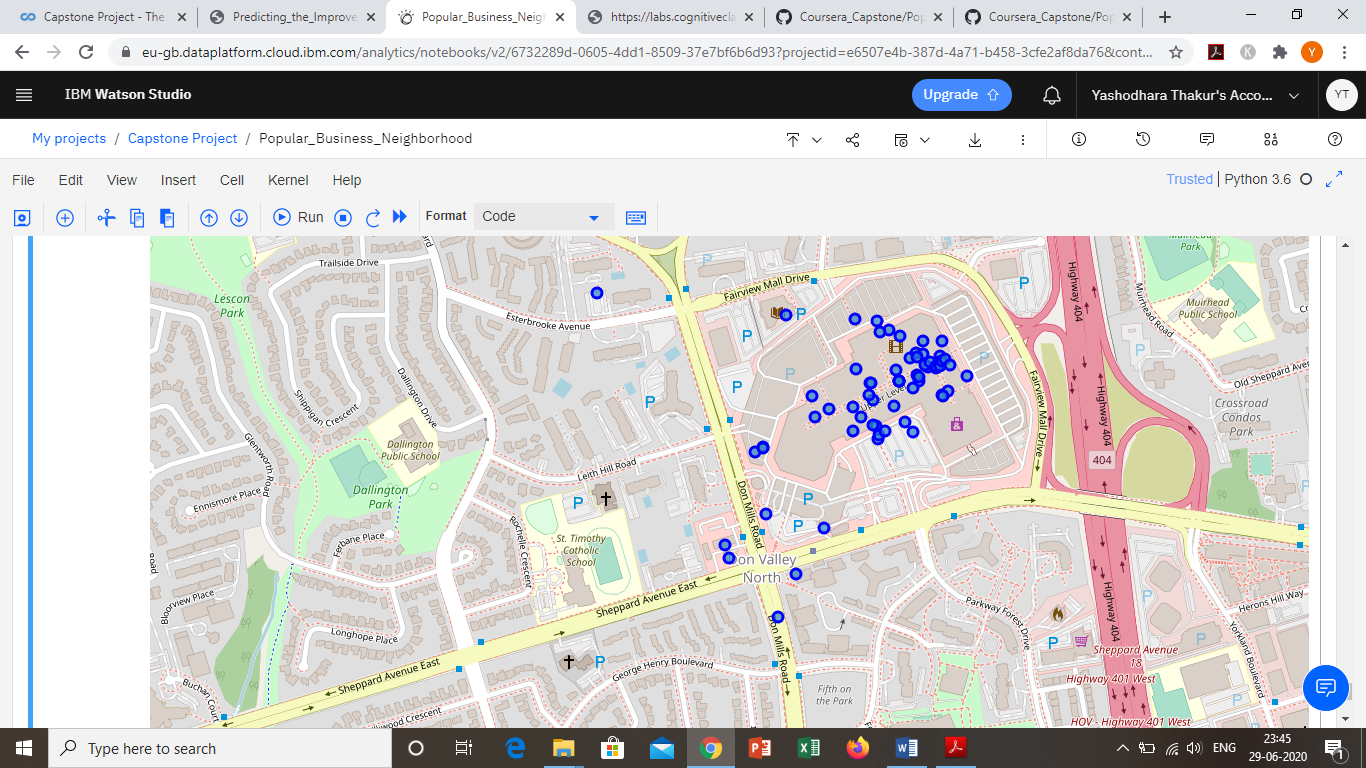
Bar graph



Least competitive businesses



Now at last we append both the datasets as one to find the best neighbourhood cluster for business and plot it on the map as shown below



# Conclusion

In this study, I analysed the different neighbourhoods in the city of Toronto. Also analysed the popular boroughs and in those boroughs we further analysed the neighbourhoods. We further analysed these neighbourhoods to find the best neighbourhood to start a business. Not only that we also found out the most competitive and the least competitive local business categories in these neighbourhoods. This model will be very helpful for people who want to start their own business and want to do competitive analysis.

# Future Directions

I have prepared this model on the basis of neighbourhoods / boroughs for a particular city. In future this model can be used to find the best city to start a business as well as a best country. Also further competitive analysis is possible using this model by adding and analysing the current businesses information.