Battle of the Businesses in a Neighborhood
Applied Data Science Capstone Project
Yashodhara Thakur

1. INTRODUCTION

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

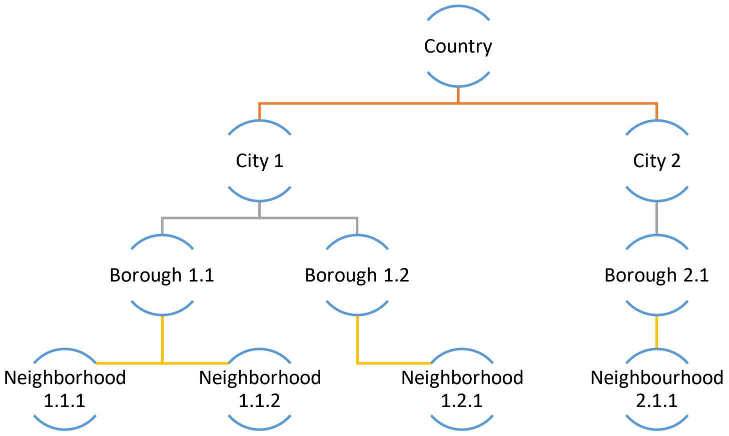
In addition to this, model also provides with most competitive and least competitive business

1.1.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.

1.2. 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:



1.3. 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.

2. 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.

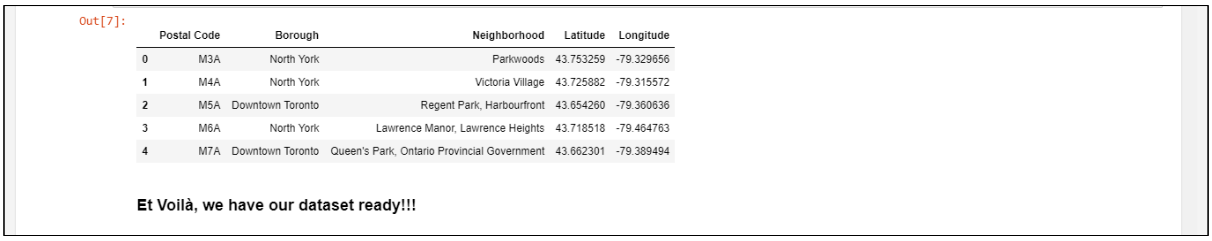
2.2 Data Cleaning

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



Out[7]:

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Et Voilà, we have our dataset ready!!!

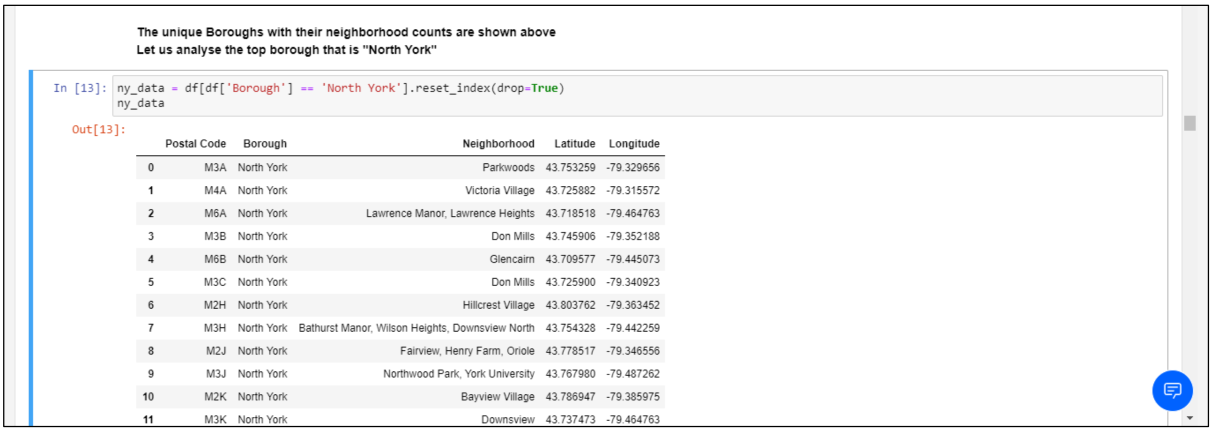
2.3 Feature Selection

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

3. 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:



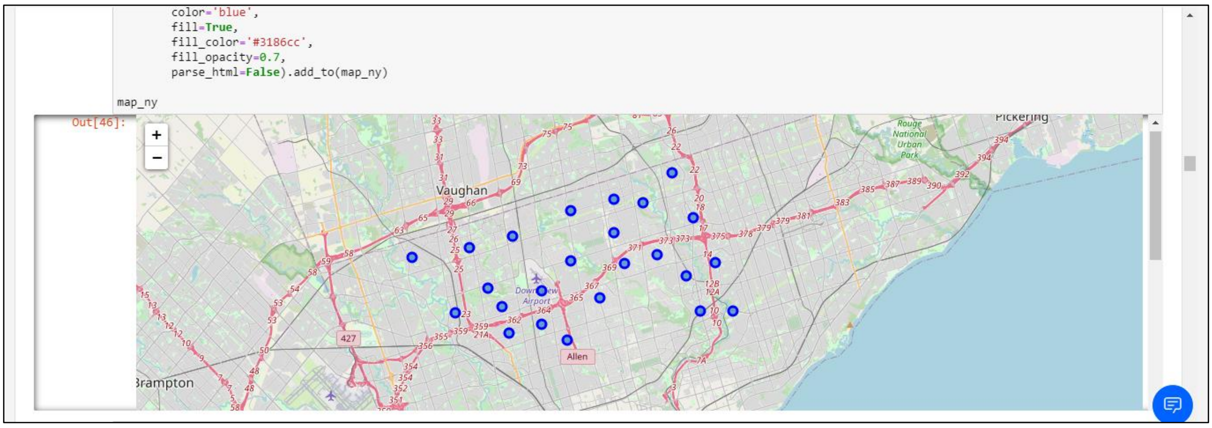
The unique Boroughs with their neighborhood counts are shown above
Let us analyse the top borough that is "North York"

```
In [13]: ny_data = df[df['Borough'] == 'North York'].reset_index(drop=True)
         ny_data
```

Out[13]:

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
3	M3B	North York	Don Mills	43.745906	-79.352188
4	M6B	North York	Glencairn	43.709577	-79.445073
5	M3C	North York	Don Mills	43.725900	-79.340923
6	M2H	North York	Hillcrest Village	43.803762	-79.363452
7	M3H	North York	Bathurst Manor, Wilson Heights, Downsview North	43.754328	-79.442259
8	M2J	North York	Fairview, Henry Farm, Oriole	43.778517	-79.346556
9	M3J	North York	Northwood Park, York University	43.767980	-79.487262
10	M2K	North York	Bayview Village	43.786947	-79.385975
11	M3K	North York	Downsview	43.737473	-79.464763

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



```
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_ny)

map_ny
```

Out[46]:

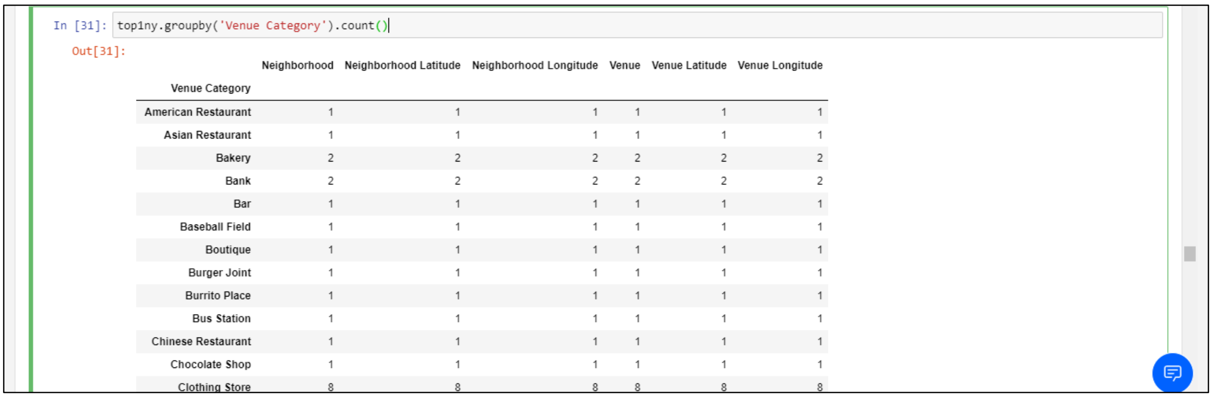
Now we wanted to find the best neighbourhoods in North York.

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:



```
In [31]: top1ny.groupby('Venue Category').count()
Out[31]:
```

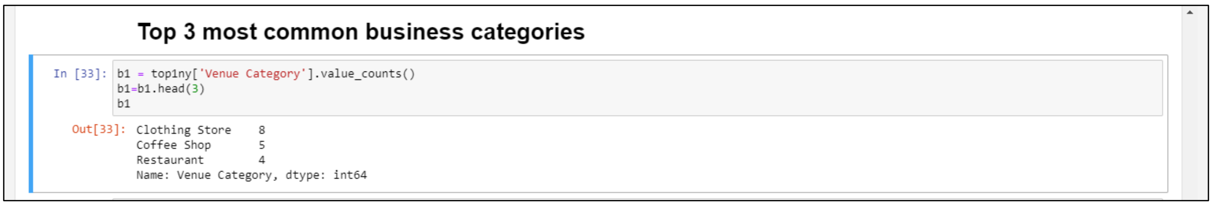
Venue Category	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude
American Restaurant	1	1	1	1	1	1
Asian Restaurant	1	1	1	1	1	1
Bakery	2	2	2	2	2	2
Bank	2	2	2	2	2	2
Bar	1	1	1	1	1	1
Baseball Field	1	1	1	1	1	1
Boutique	1	1	1	1	1	1
Burger Joint	1	1	1	1	1	1
Burrito Place	1	1	1	1	1	1
Bus Station	1	1	1	1	1	1
Chinese Restaurant	1	1	1	1	1	1
Chocolate Shop	1	1	1	1	1	1
Clothing Store	8	8	8	8	8	8

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

4. Results

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

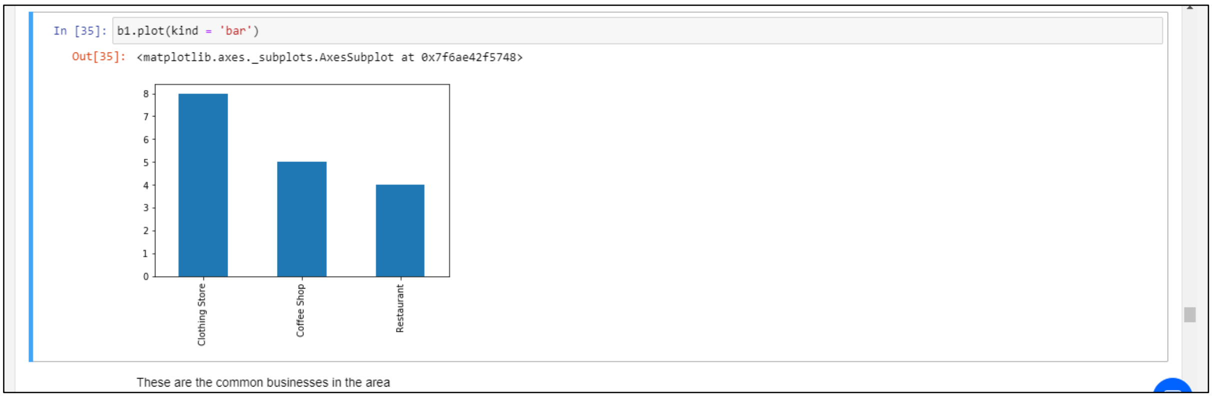
The example of the output is shown below



Top 3 most common business categories

```
In [33]: b1 = top1ny['Venue Category'].value_counts()
         b1=b1.head(3)
         b1

Out[33]: Clothing Store    8
         Coffee Shop       5
         Restaurant        4
         Name: Venue Category, dtype: int64
```

We also plotted a bar graph to represent the above data

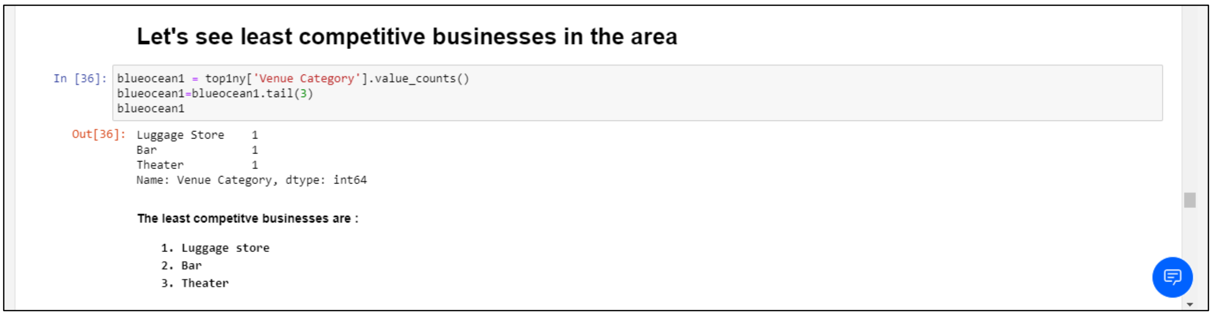


```
In [35]: b1.plot(kind = 'bar')

Out[35]: <matplotlib.axes._subplots.AxesSubplot at 0x7f6ae42f5748>
```

These are the common businesses in the area

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



Let's see least competitive businesses in the area

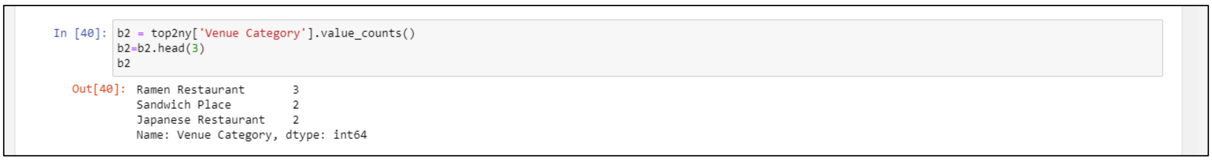
```
In [36]: blueocean1 = top1ny['Venue Category'].value_counts()
         blueocean1=blueocean1.tail(3)
         blueocean1

Out[36]: Luggage Store    1
         Bar              1
         Theater          1
         Name: Venue Category, dtype: int64
```

The least competitve businesses are :

1. Luggage store
2. Bar
3. Theater

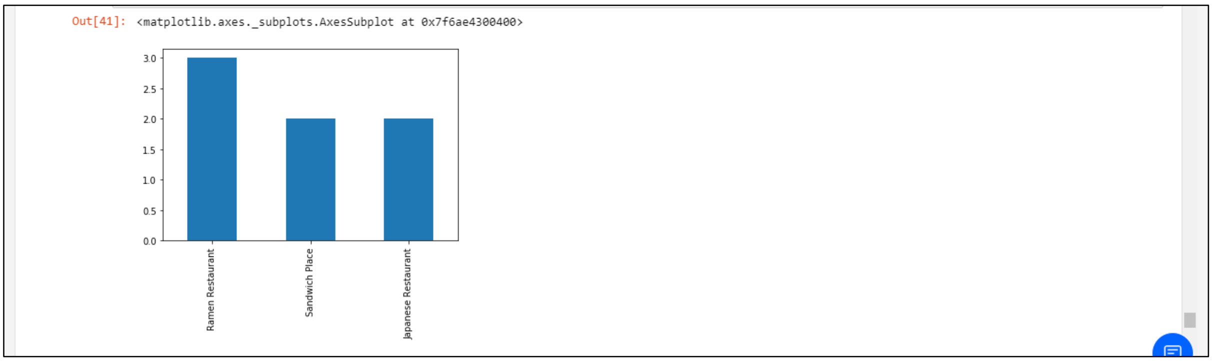
The same analysis was done with the 2nd best neighbourhood and the results were as follows:

Popular businesses in 2nd neighbourhood



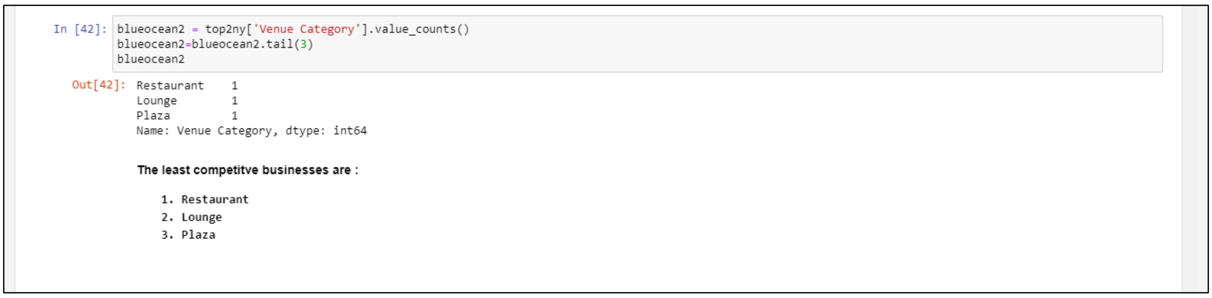
```
In [40]: b2 = top2ny['Venue Category'].value_counts()
         b2=b2.head(3)
         b2

Out[40]: Ramen Restaurant      3
         Sandwich Place        2
         Japanese Restaurant   2
         Name: Venue Category, dtype: int64
```

Bar graph



```
Out[41]: <matplotlib.axes._subplots.AxesSubplot at 0x7f6ae4300400>
```

Least competitive businesses

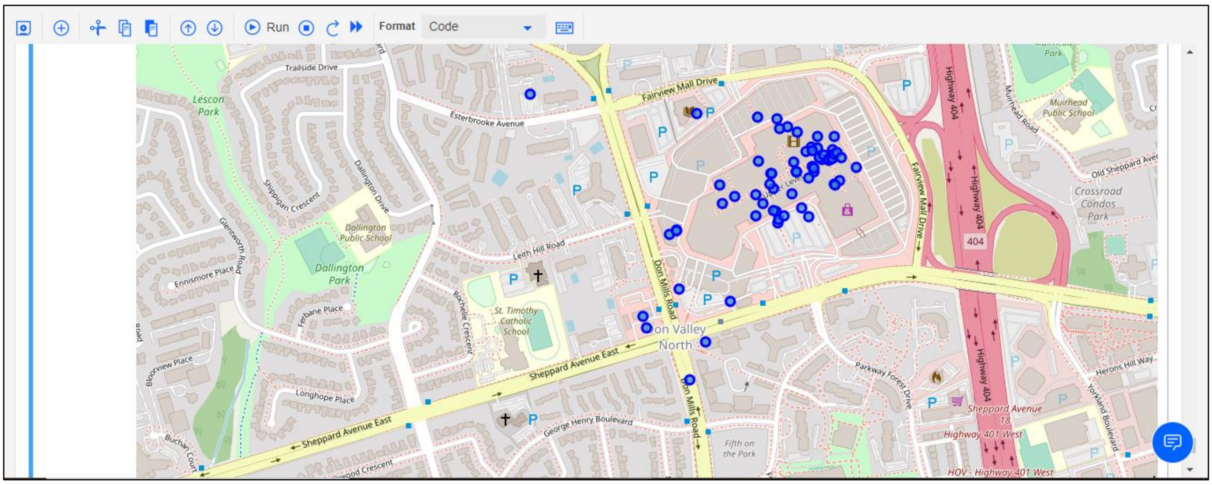


```
In [42]: blueocean2 = top2ny['Venue Category'].value_counts()
         blueocean2=blueocean2.tail(3)
         blueocean2

Out[42]: Restaurant    1
         Lounge        1
         Plaza         1
         Name: Venue Category, dtype: int64
```

The least competitve businesses are :

1. Restaurant
2. Lounge
3. Plaza

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



5. 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.

6. 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.