Battle of the Neighborhoods - Capstone Project

1. Introduction: Business Problem

About 300,000 immigrants are coming to Canada every year and a lot of them will settle in Toronto. Many of them need some help to locate the place to live. So, the idea behind this Capstone project is to create a tool which will help people moving to Toronto to choose the right location by providing data about the neighborhoods population, average income, apartment cost, school ranking and venue density. We will be comparing 2 Borrows: North York and Scarborough.

2. Data

To solve the problem described above, we need the data on geolocation of neighborhoods, its population, average income in every neighborhood, apartment cost, school ranking and venue density. These data contained in the following data sets:

• **Toronto's Census data** - is obtained from this website: https://www03.cmhc-schl.gc.ca. It contains the following features:

- Name of the Neighborhood;
- Average Household Income Before Taxes;
- Median Household Income Before Taxes;
- Average Household Income After Taxes;
- Median Household Income After Taxes;

• **Toronto Neighborhoods data** - is obtained by scraping Wikipedia page with list of postal_codes of Canada from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. It contains the following features:

- Postal Code;
- Burrows;
- Neighborhood;

• For geotagging Postal Codes in Toronto is used **Geospatial** dataset with geolocation data, obtained from https://cocl.us/Geospatial_data' by web scraping. It contains the following features:

- Postal Code;
- Latitude;
- Longitude;

• **School Ranking Dataset** - data for both, elementary and secondary schools, obtained from the http://ontario.compareschoolrankings.org/ website through web scraping. It contains the following features:

- Full name of the school;
- City;
- 2017-18 Rating;
- Rank of the school;

- Postcode of the school;

• **Rental Apartments Dataset** - is obtained from www.kaggle.com and contains the following features:

- Addresses of apartments;
- Number of bedrooms;
- Number of bathrooms;
- Does it have den?;
- Postal codes and addresses;
- Longitude and Latitude;
- Rental price

• **Venues Data Dataset** - is obtained through Foursquare API and contains:

- Counts of Venues in the Neighborhood;
- The relative proportion of each venue category;
- etc.;

When we have all of the data, we can create models, maps and identify clusters according to the following features: population, average income in every neighborhood, apartment cost, school ranking and venue density.

Methodology

Dataflow

To compare the similarities of two Borough's, we explored their neighborhoods, segmented and grouped them into clusters using **K-Means clustering algorithm**. For this we combined all data sets described in the Data section (**Neighborhoods data, Census data, School Ranking Dataset, Venues Dataset)** into one dataset by joining them using columns containing Postal code. Because we think that Venues features are inter dependable and using PCA for dimensionality reduction would be a good idea (but we didn't use it in this course), so we simply separated Venues features from Average Income and School Rating and conducted clustering on both sets separately. Also, we didn't use Apartment Rental dataset, because we found that it doesn't have enough statistics for our Boroughs.

We began assignment with scraping the Wikipedia page containing mapping Toronto's Postal Codes on Boroughs and neighborhoods. We parsed html using pandas read_html method. For the geolocation information we read csv file. After some filtering and transforming we joined both datasets and used **folium** library to create map containing both, North York (blue dots) and Scarborough (red dots) neighborhoods. We used **geopy** library to calculate coordinates of the middle point between Boroughs.

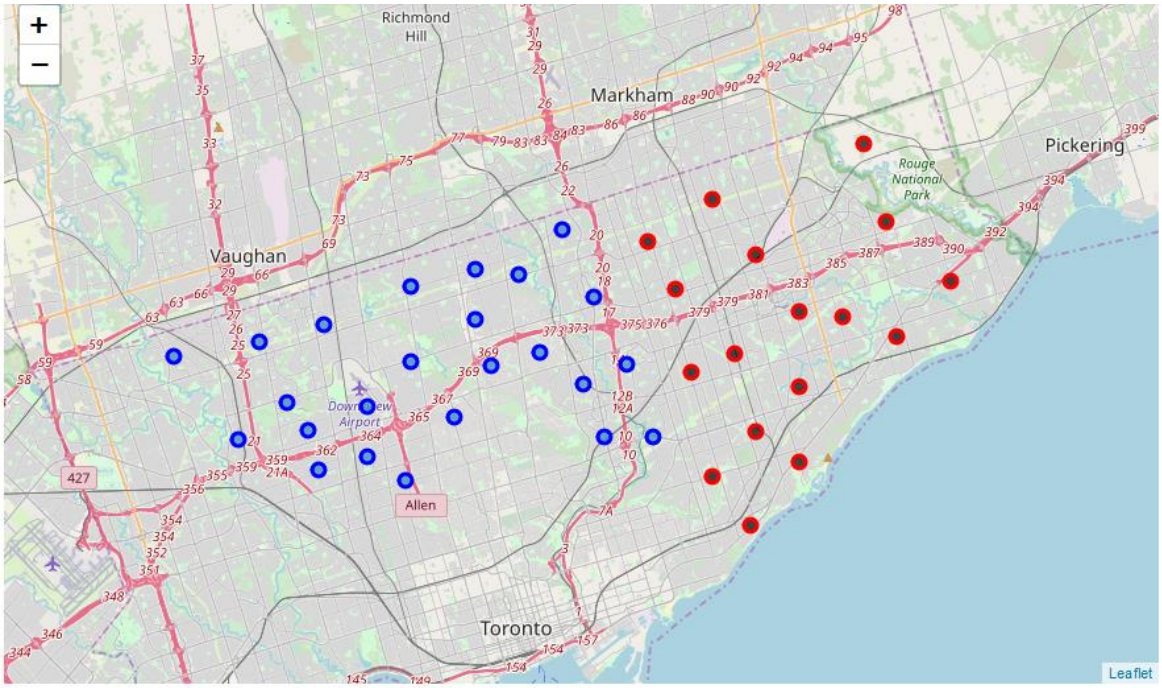

Fig. 1. Map Toronto with North York (blue dots) and Scarborough (red dots) neighborhoods.

Next, we read csv file containing data on **rental cost of apartments** in Toronto. We extracted **Postal Code** from addresses using **regular expression**, transformed the dataframe by using **onehot encoder** to make number of bedrooms feature of the dataset. Unfortunately, after filtering and transformations, dataset became depleted with ~ 120 apartments total in both Boroughs. We decided not to use approximations for the neighborhoods where data is missing as it would defeat the purpose of analysis.

Using the similar approach, we retrieved csv file containing **census data** for **average income** in neighborhoods.

To get **School Ratings** information, we had to visit every school website. On the first step, we used combination of **regular expression** and **BeautifulSoup** library to parse the landing page and retrieve list of ratings and links to schools websites, on the second run – we visited every school website and used **regular expressions** and **BeautifulSoup** library to retrieve full name of the school and its **Postal Code**. And we did it separately for elementary and secondary schools and then we concatenated data in one dataframe, which we used to calculate **Average School Rating** per neighborhood.

Using K-Means Clustering Approach on Average Income and Average School Rate for Neighborhoods

For correct comparison clusters in each of Boroughs, we used **StandardScaler** to scale the features on the common dataset containing data for both Boroughs and then applied clustering algorithm to it. We added separate column with obtained clusters identifiers. Then we separated dataframe on two – one to build North York map, second for Scarborough map.

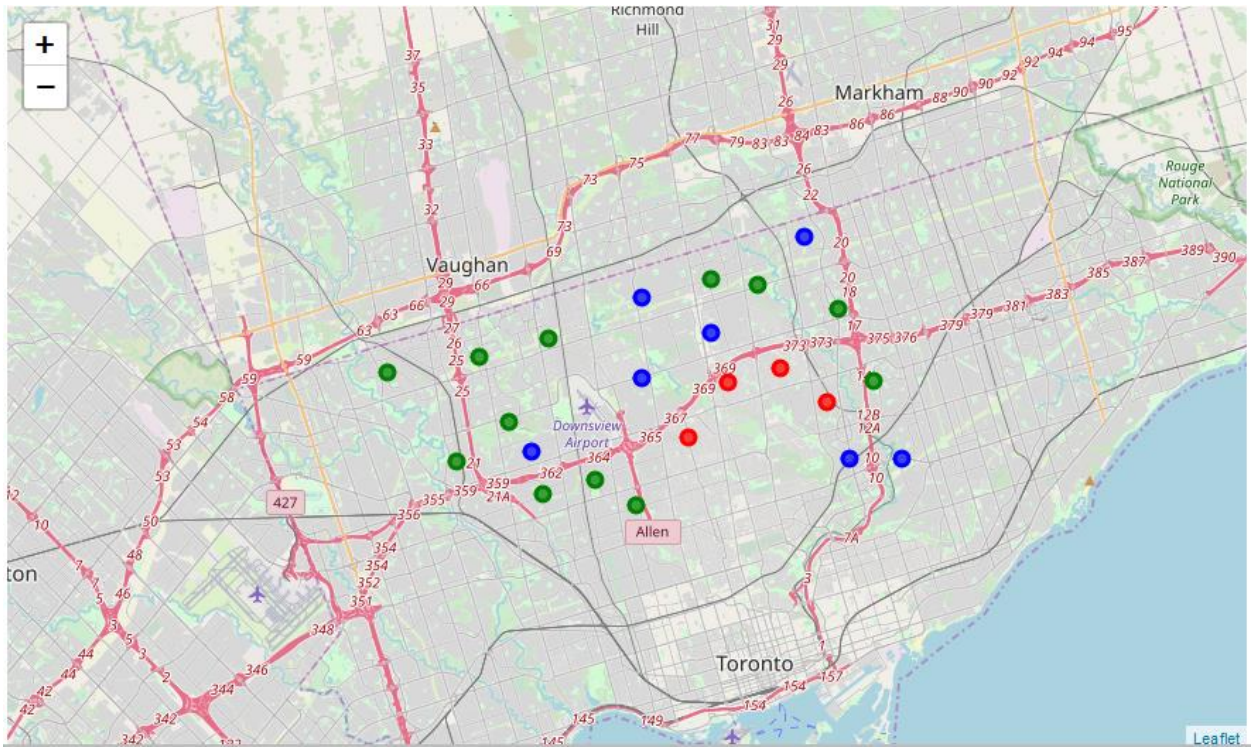

Fig. 2. North York clustering based on Average Income and Average School Rate

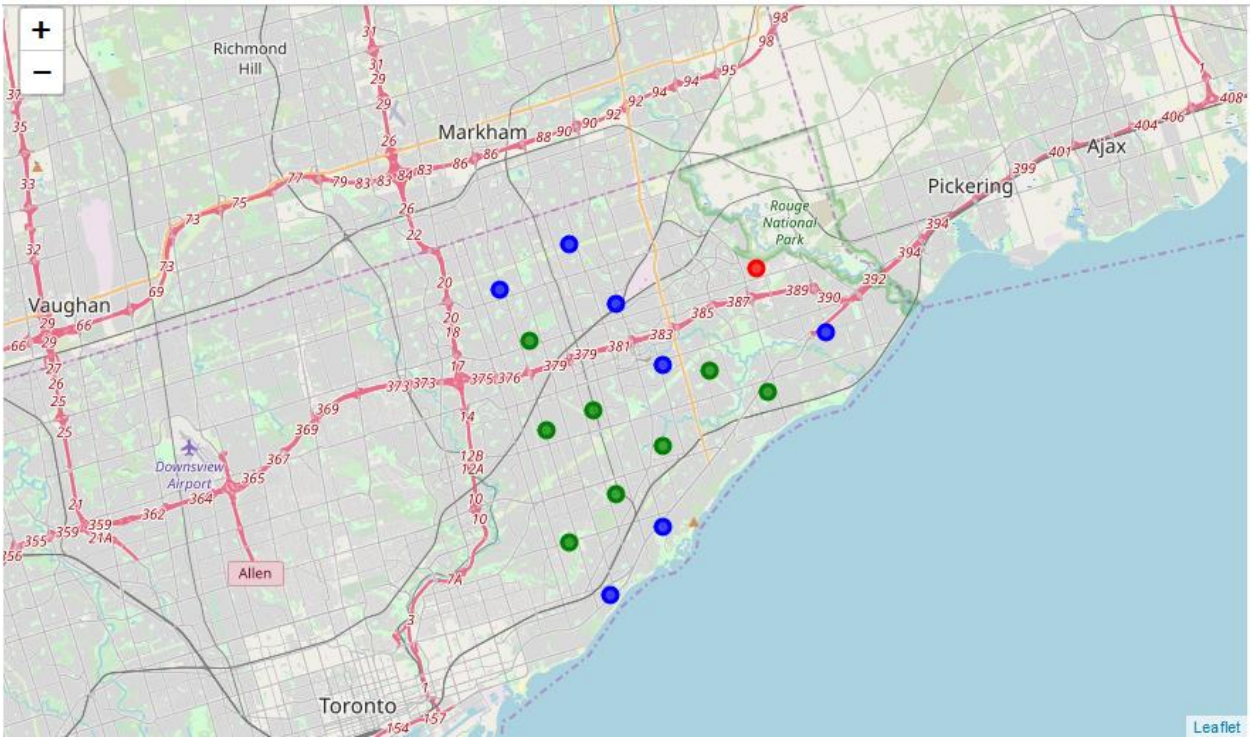

Fig. 3. Scarborough clustering based on Average Income and Average School Rate

Using Foursquare API and K-Means Clustering Approach

To explore venues in neighborhoods we used Foursquare API to obtain list of venues around Postal Code location for each neighborhood. Then we used **onehot encoder** to calculate relative frequencies of each venue in neighborhoods. Again, we used common dataset when we applied K-Means Clustering algorithm and then separated results for Boroughs.

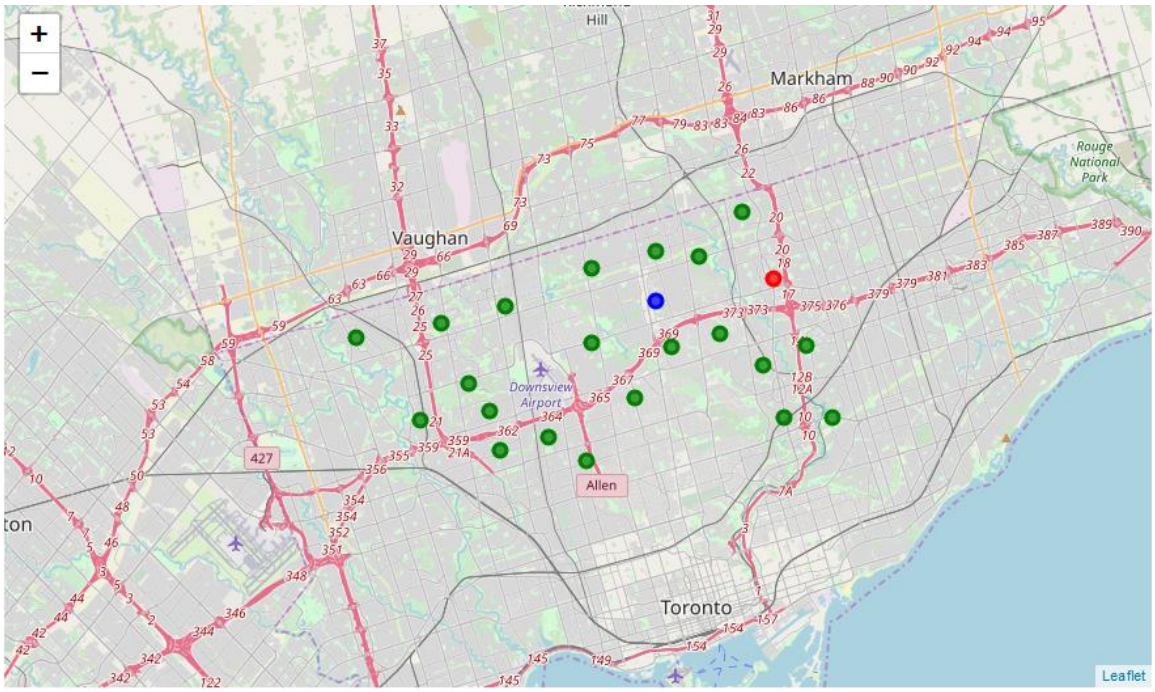

Fig. 4. North York clustering based on Venues distribution

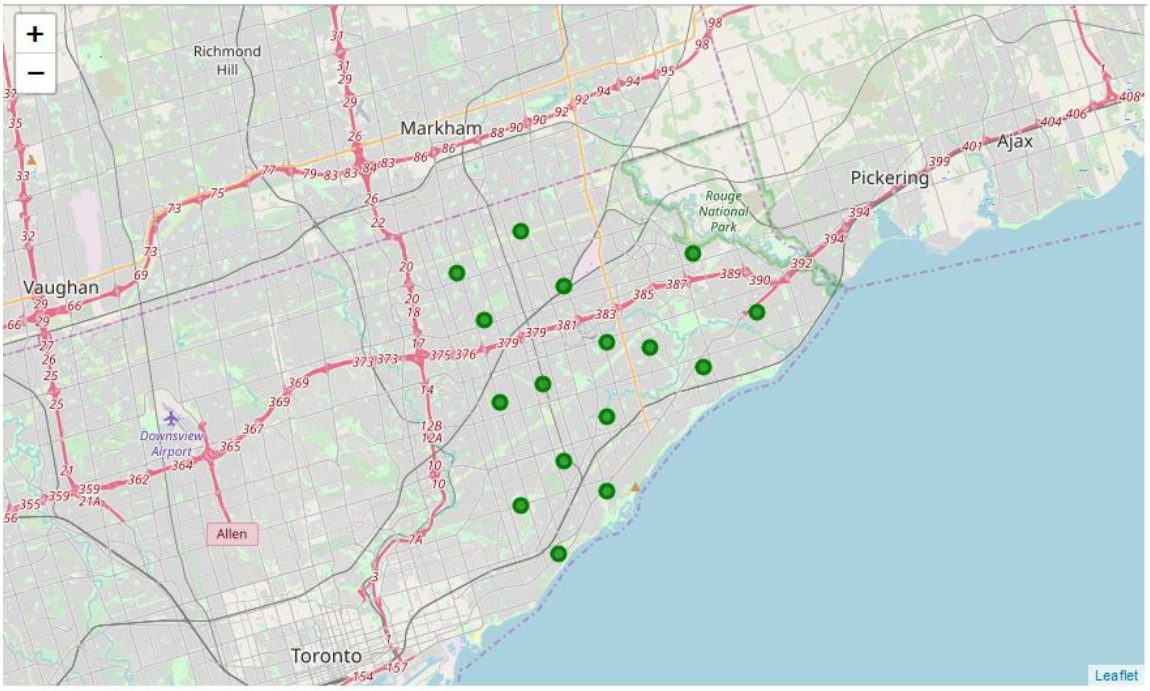

Fig. 5. Scarborough clustering based on Venues distribution

Results

North York

As could be seen from the fig. 2. North York clustering based on Average Income and Average School Rate, we can see that high-income/high-school rating cluster is located just south of the 401, between Allen Road and Don Valley Pkwy. There is no clear geographical separation between the other two clusters, with high-income/middle-school rating and with middle income-lower school rating.

As can be seen from fig 4. North York clustering based on Venues distribution, K-means analysis didn't discover much diversity in Venues distribution in North York. One cluster contains 21 neighborhood, each of other 2 contains just 1 neighborhood.

Scarborough

From the fig. 3. Scarborough clustering based on Average Income and Average School Rate we can see that high-income/high-school rating cluster consists of just one neighborhood and located near the Rouge National Park.

As can be seen from fig. 5. Scarborough clustering based on Venues distribution, according to K-means analysis Venues in Scarborough spread even more uniformly than in North York. All of neighborhoods belong here to the same cluster, because K-means algorithm was applied to dataset containing data for both Boroughs and all neighborhoods in Scarborough were classified as belonging to the same cluster as 21 of 23 neighborhoods in North York.

Discussion

Chosen for comparison Boroughs are very similar and there is no much difference between the two were discovered when K-means analysis was applied. Situation potentially could be changes if we take into account price of the houses, population, ethnicity of the population, etc.

Conclusion

Research described in this paper didn't find significant differences between 2 Borrows in Toronto: North York and Scarborough.