COMPARING NEIGHBORHOODS
Author: Yan Almeida

- **Introduction/Business Problem**

Defining where to live or open a new business is always a challenge. Many important features are determinant when making this choice, and when it comes to the venues around the picked place, it's probably one of the most important ones, since it determines how easy it will be for you to work, to fight competition, to accommodate your family, to find clients etc.

With this in mind, the goal of this project is to provide an efficient way of comparing different locations based on the venues we can find around them, providing an easy way for the final user to pick a good place. This way, a possible client will be able to either find a fine place to buy a house and settle his family or start a new business in a place with less competition and more possibility of development (e.g. a restaurant close to parks, museums and corporate buildings).

- **Datasets:**

In order to show the comparison of neighborhoods, we are going to use data from two different cities, showing how this project provides a good way of defining where to live: maybe you really like New York, but have not been able to find a good neighborhood with a lot of entertainment venues to live in. Or maybe you live in Toronto and want to open a new American Restaurant in a neighborhood that is similar to New York's neighborhoods. Data from Toronto comes from Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) and Cocl.us (http://cocl.us/Geospatial_data), and data from New York is provided by the course itself. Besides that, data from FourSquare is also used in order to obtain information on the venues around the different neighborhoods.

- **Methodology**

In order to work with our data, we started organizing it into dataframes. After that, a function using FourSquare API was built in order to find venues around the neighborhoods, and then we applied KMeans clustering method, using Elbow Analysis to find the optimum number of clusters. Then, using a function to see the most common venues around a neighborhood, we were able to count them using Pandas and plot bar charts to show the most common venues for each cluster. After this, we used folium to plot a choropleth map showing each Neighborhood and its respective cluster.

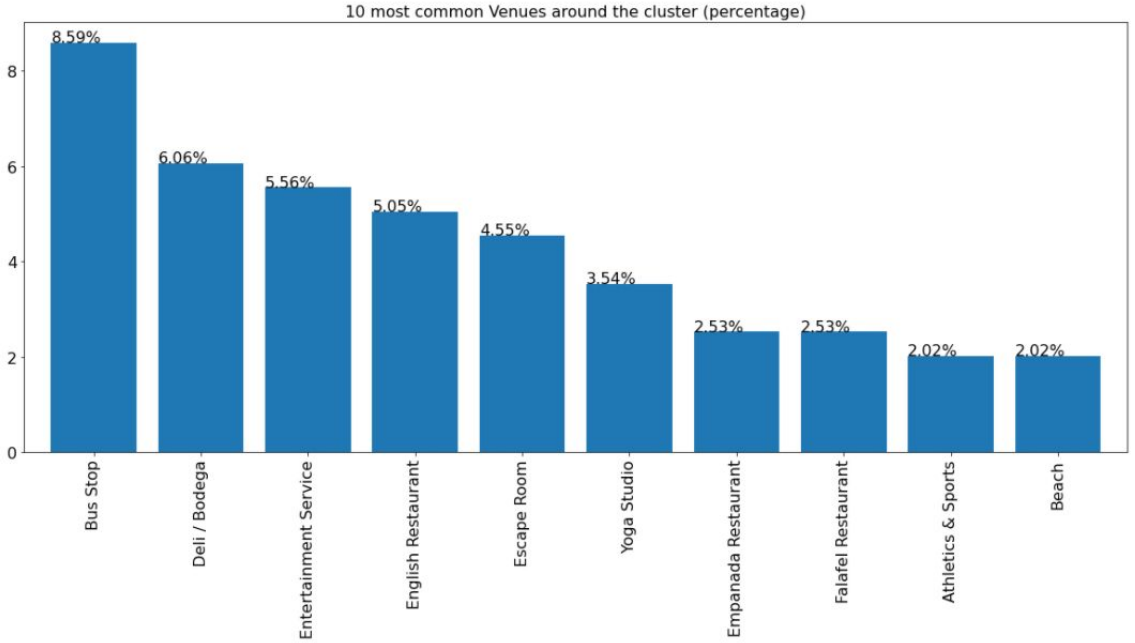
Libraries:

- Pandas
- NumPy
- Sklearn
- Geopy
- JSON
- Requests
- Matplotlib
- Folium

- **Results**

After running the built algorithm, we can see that 8 different clusters are formed, and through the bar charts we can also see that some clusters show more options of entertainment, others show venues like pharmacies and stores as most common and so on. This way, it's possible to see that the algorithm really helps us to visualize the venues we are most likely going to find around a neighborhood based on the cluster it belongs to. Using this tool, we could, for example, determine that neighborhoods which belong to cluster number 5 are a good option if you are moving and do not have a car, since the most common venues around the neighborhoods of this cluster are bus stops, as we can see below:

Cluster number 5

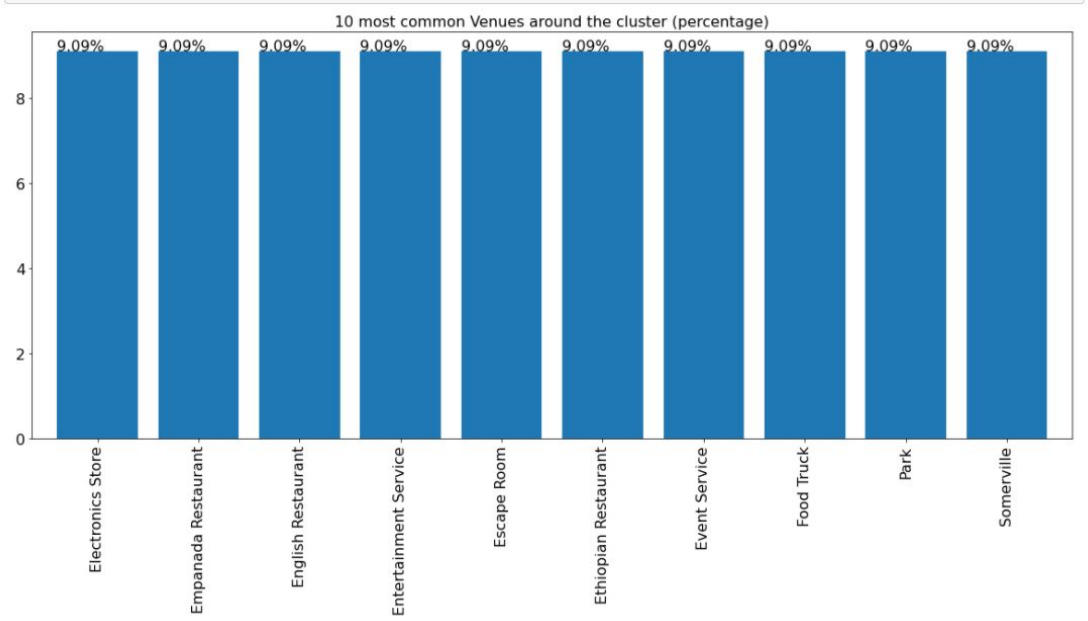


This way. We can say that the project works and is actually helpful when taking this kind of decision.

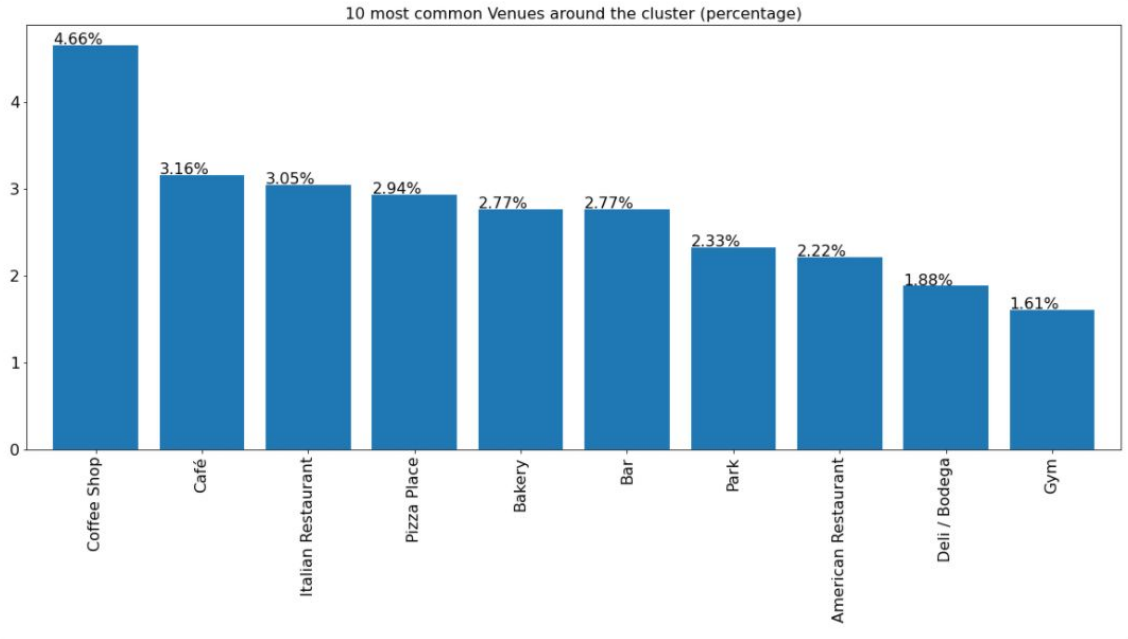
- **Discussion**

Based on the results, it's important to say that we can extract a lot of information from the bar charts plotted. For instance: we can see that some clusters present more balanced number of venues than others, as we can see below:

Cluster number 4



Cluster number 3



Besides that, the algorithm helps people to restrict their options to specific clusters, what allows them to focus only on the neighborhoods of that cluster and then compare the prices etc.

It's also important to highlight that the project can still be improved with new ideas of data cleaning and division.

- **Conclusion**

It's possible to conclude that the project is a success and can be used by different clients in order to solve the problem we proposed in the 'Introduction' section. With that said, it's also important to reiterate that the project can still be improved with new features and ideas.