**Selecting the best places in Toronto to put business**

**Christian Leon**

**November 5, 2019**

1. **Introduction**

**As we know now, some business has become very competitive in the lasts years, so certain characteristics about planning to open one becomes fundamental. Businesspersons can choose wrongly the place to settle their company or shop, giving negative revenue in the short term and causing the closure of their business in the medium or long term. On the other side, firms who know where to geospatially land have bigger probabilities to earn positives profits and prosper in the short and long term. Some factors that determines this are: the demand, the offer, demographical data about the population in a close radius, etc. Therefore, it is advantageous for firms to know where exactly to settle in order to compete and gain a competitive edge over the geographical space.**

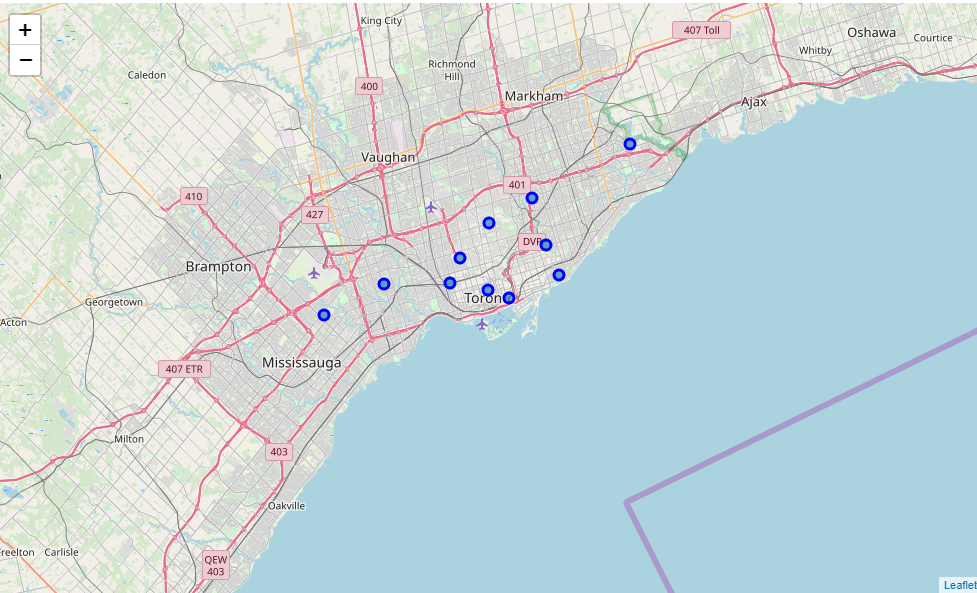
**This projects aims to help businesspersons who are interested in settle their business in a positive earning area. Obviously, a lot of firms will be interest because of the expectations they had on their firms and ideas.**

1. **Data acquisition and cleaning**

**Data was scrapped from Wikipedia and asked by a consult in Foursqueare API. All relevant data can be found in the Jupyter Notebook here. This data consists on the neighborhoods of Toronto and their principal places per neighborhood for the year 2019, so we are getting contemporaneous data, giving more relevancy to the segmentation of the model.**

**After the scrapping was completed, we started giving form to the data. First, we cleaned all the “Not Assigned” labels on the Neighborhood column, and summarize the boroughs per neighborhood in the “Borough” column. After we managed to do this, we complete the data with geospatial data (longitude and latitude) by merging the datasets given by Coursera platform so we can then call the Foursqueare API.**

**We then created a map using folium so we can visualize every neighborhood more clearly. As we can see in the notebook all the libraries were imported correctly and give the accurate location of the city of Toronto, Canada. As we can see in the figure below all the neighborhoods can be recognized by blue points all along indicating the center of each one.**

****

**Obviously, this is not all the data we need to start the segmentation algorithm, so from now on we started using the Foursquare API. First, we defined a function which input are the neighborhoods of Toronto and output is the various venues it has with their latitude and longitude, we get the venue category as well.**

**After this was done, we started to manipulate the data to get some insightful information as the top favorite venues and their categories. With this, we got the final data set before the employment of the K-means algorithm for segmentation.**

1. **Methodology**

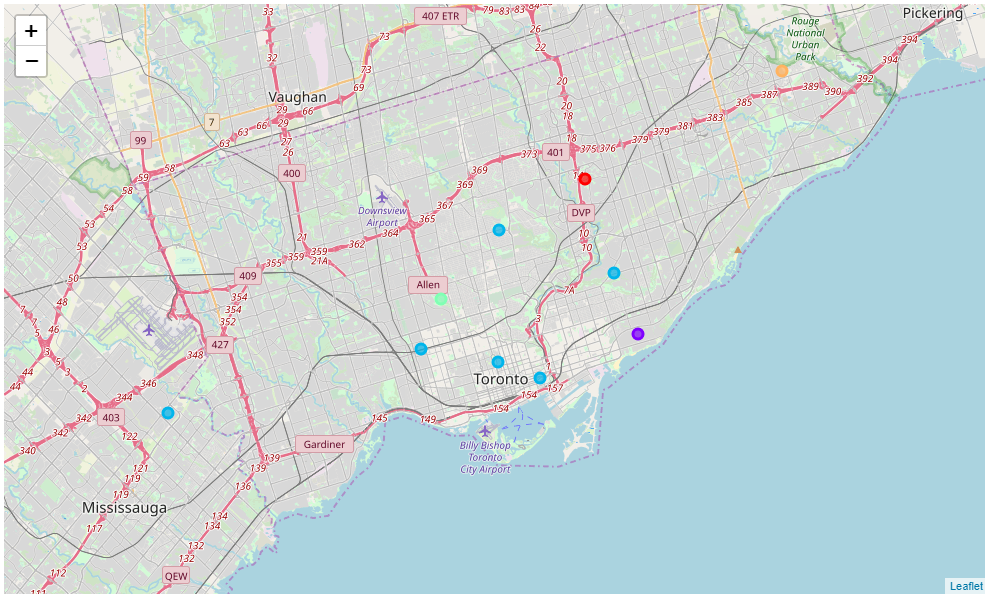
**We want to categorize each neighborhood so we can advise companies to which place to settle their business. As we saw in the data section, some of those variables are key to get the job done. Such as venue category, the other kind of business in the area, and the favorite places by the people living in the area.**

**For this, we’ll use the K-means algorithm so we can categorize each neighborhood in an optimum way. If we would like to suggest specific kind of firms, we could have use a classification algorithm but is not the case. We will make a conclusion about the results using the most likely business to prospere in each neighborhood.**

**Some of the benefits of this algorithm are its relatively simplicity to implement, scalation to large data sets, guarantess convergence, can warm-start the positions of centroiuds, easily adapts to new examples and generalize to clusters of different shapes and sizes. But one fundamentally con is the election of k manually as we can see in the notebook we use k = 5 to persist in having more categories and be more consistent with our conclusions. At least we don’t get outliers values that could potentially distortion the categorize of each neighborhood. We can confirm this by looking at the data.**

1. **Results**

**The following image suimmarize our results.**

****

**As we can see the 5 clusters were created, now I would like to make some inference about each cluster:**

* **Cluster 1: As we can see in the notebook the first common venue is a park and the next in the row are food places, so we would usually recommend people or small businessperson to go for food target environments for this cluster (North York)**
* **Cluster 2: In the second cluster we can see that the common venues refer more to service firms such as health, electronic devices, Trail, and a mix of food. We would highly recommend probably not to invest in this area because of the high variance in the kind of business here. (East Toronto)**
* **Cluster 3: The biggest cluster of them all, which englobes Downtown Toronto, Queen’s Park, East York, West Toronto, Central Otronto and Mississauga seems to highlight recreational areas, such as parks, pubs, gym, theaters. We would highly recommend entertainment firms to go to this cluster immediately. This all are fairly close to the University of Toronto which makes more sense since more population transits those areas every day.**
* **Cluster 4: York which is our 4th cluster seems like more and sport and eat area, business like sports centers and food would highly function here.**
* **Cluster 5: Finally, Scarborough seems fairly equal to the second cluster, probably this indicates that we should use only 4 clusters for next investigations.**

**As we can see the tastes of people determine the likelihood of success of a firm. We can use the results discovered to make recommendations to new firms who would like to settle in new areas.**

1. **Discussion**

**As we managed to identificate clusters, the lack of more demographic data about the people living there could have been insightful for more decisions, data about old business operating in the areas would have been perfect so we could have seen an historical progression of the popularity of some firms over Toronto. But for now this a nice first approach.**

1. **Conclusion**

**As we noticed the k-means algorithm can be somewhat discrete with the k value. But at least, results give insight about which business could open in different areas. It has a lot of room for improvement, but for now it seems a great approach. Experiments could be done changing some parameters of the k-mean algorithm and a different data set.**