**Introduction**

**Business Problem Description.**

**I have been hired by FIFA (International Federation of Association Football) to decide the city with the best conditions (accommodation, venue, transportation) to host for the World Cup semi-finals and final in Summer 2019. The 2 cities running to host the event are New York in USA and Toronto in Canada.**

**There are expected more than 400,000 people in a 5 days event that have to be accommodated, transported and entertained meeting FIFA standards.**

**Data used to solve the problem**

**The data used is coming from the following sources:**

* **New York City Data (CSV file).**

**From the Data Base I’m taking information of all Boroughs and Neighborhood in New York City including Coordinates.**

* **Toronto Data (Wikipedia).**

**From this web page I’m extracting a table that contents all information regarding Boroughs, Neighborhoods.**

* **Toronto Coordinates (CSV file)**

**Using a CSV file with all coordinates to match up with Borough and Neighborhoods.**

* **Foursquare.com**

**From here we obtain all information on venues for both cities and use this information to compare them.**

* **FIFA (Web Page)**

**From FIFA web page we get the guide to the bidding process for the FIFA World Cup, with this information we can determine how this organization based the selection of the cities to host the tournament.**

**Methodology**

**Here I describe the methods used for assessing both Cities and obtaining the winner.**

* **Cleaning the Data**

**In the first place I had to obtain and clean the information for both cities. The information is coming in different formats:**

**For Toronto the information was coming from a table located in the following web page:** [**https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M**](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)**, to do so, I had to use BeautifulSoup to extract the data from the web page.**

**I goal for this part of the program was to create a Pandas Data Frame that included all Postal Codes, Borough and Neighborhoods with their coordinates for the City of Toronto. To do so, once I pulled Toronto’s data I created 3 lists with all Postal Codes, Borough and Neighborhoods and cleaned up all “bad data”(N/A, Null) obtained from the web page and joined all 3 cleaned lists to create my Data Frame, after this process I incorporated the coordinates to the Data Frame by opening a CSV file that had the data.**

**Second Step on this process was to open and clean up New York’s data.**

**This was a very simple step, because I had a CSV file available with the data. So to get it I just applied:**

**pd.read\_csv(r’File\_Location\NewYork\_Data.csv’)**

* **Exploring the Data:**

**I start exploring data using Foursquare data. First I access the development Foursquare environment giving the Client\_ID, Client\_Secret and Version and then I get all venues for Toronto and New York, using a function created to explore all Neighborhoods in both cities. The result obtained are shown in the result section.**

* **Creating Clusters**

**After exploring the data and obtaining al venues, I proceed to create cluster for both cities. These clusters are going to help me to assess each city according to the standards used by the client in this case FIFA.**

**I used the method K-Means to do the clustering, because it gives me a segmentation that will be easier to evaluate in each city and apply the weight matrix to give the results.**

**See Results section.**

* **Evaluation and comparison of each Cluster/City:**

**I use the cluster results to evaluate all cluster in each city and weight them. The clusters are:**

**Food & Beverage**

**Accommodation**

**Transportation**

**Stadium**

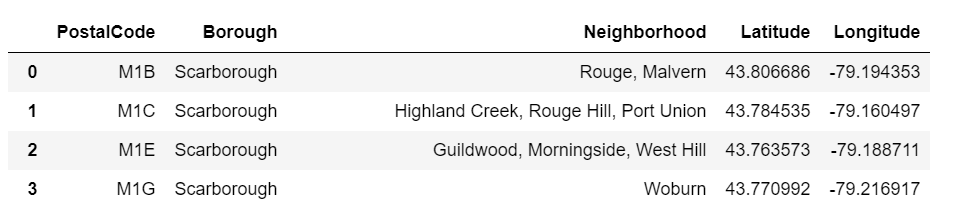
**and the weight of every Cluster is 20%, 20%, 30%, 30% respectively.**

**Results**

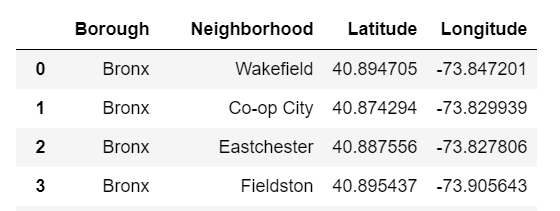
* **Cleaning the Data:**

**I obtained at the end of this process 2 tables with all Neighborhood and coordinates for each city.:**

**Toronto:**

****

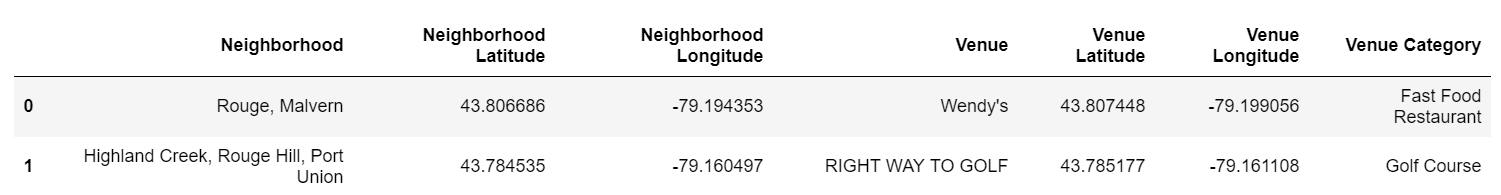
**New York:**

****

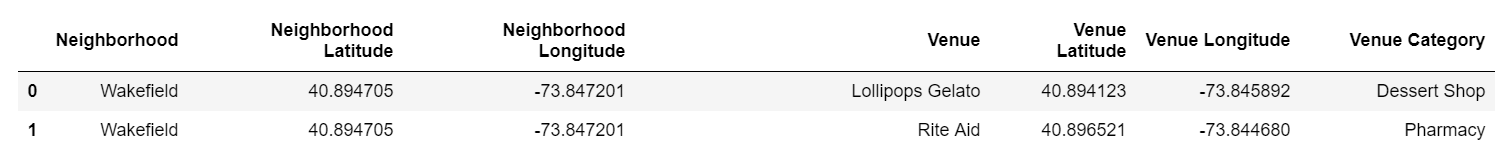
* **Exploring the Data:**

**The results obtained after this process is one table for each city where we are able to see the Neighborhoods and Venues together in the same table alone with their coordinates .**

**Toronto:**

****

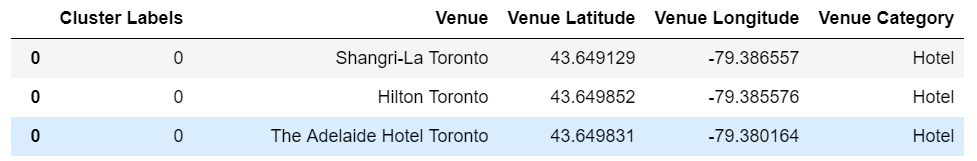
**New York:**



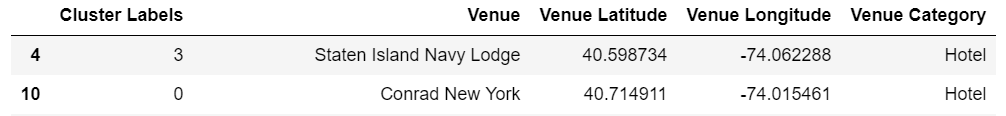
* **Creating Clusters:**

**The idea is to assign clusters according to every venue in the city according to the segmentation used to do the qualification. Hence: Food & Beverage, Accommodation, Transportation, Stadiums. Results obtained:**

**Toronto:**



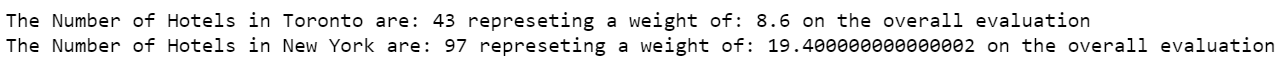
**New York:**



* **Evaluation and Comparison:**

**I did the comparison between Categories:**

**Example of one of them:**

****

**The winner City will be one with more cumulative points overall:**



**Discussion**

* **To have a better results definitely more relevant information should be incorporated to this assessment. Some of the information that would be great to have is related to finance matter. Amount these additional data we have: predicted costs of the competition, estimated revenue from the sale of tickets and hospitality packages, estimated revenue from the sale of media and marketing rights, IT and location of the International Broadcasting Centre. Probably the results could change after incorporating these financial variables that are important on the decision making.**
* **The scope of this work was simple and since I was only looking to apply the methods learned during the course I did not incorporated the data mention before, but to have a thorough analysis with accurate results we should do it.**

**Conclusion**

**The winner of the contest is:**

**NEW YORK CITY**

**We can conclude that since I was evaluating only infrastructure the result reflects the wider infrastructure present in New York than in Toronto.**

**As mention in the Discussion section, this may change incorporating more relevant data.**