**Predicting the location of Venues in Toronto, CA**

Brandon Morris

June 2021

**1. Introduction and Background**

It is very difficult to find popular venues of a certain category in a city. I want to examine clusters of venues in different neighborhoods. In my project, I will leverage the Foursquare location data to find the top 10 venues in Downtown Toronto (radius of 1000 meters) based on postal code.

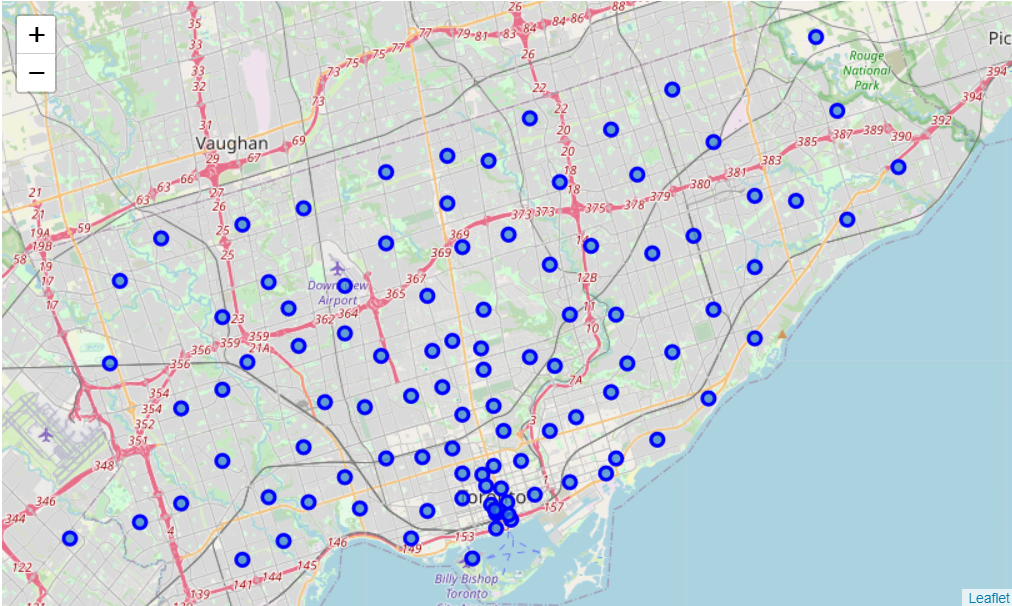
**2. Data**

I will use beautifulsoup to scrape a Wikipedia page that has postal code data from Toronto. I will use a foursquare api to get venue data. I will use one-hot encoding and k-means to cluster the neighborhoods. I will use folium to visualize the 5 clusters to see where similar venues are located throughout Toronto, CA.

Below is a pandas data frame that shows the table scraped from the Wikipedia page showing Toronto postal code data. I added Latitude and Longitude by joining the table with data from this csv (https://cocl.us/Geospatial\_data).



Below is a map showing all the locations in the dataset:



**3. Methodology**

I used the one hot encoding technique to show what venues were represented in each neighborhood. I then grouped the rows by neighborhood to show the mean of the frequency of occurrence of each category.

Below is a table that shows a subset of the data:



After finding the mean of the frequency, I created a function that sorted the venues in descending order along with other columns that show the top number of venues.

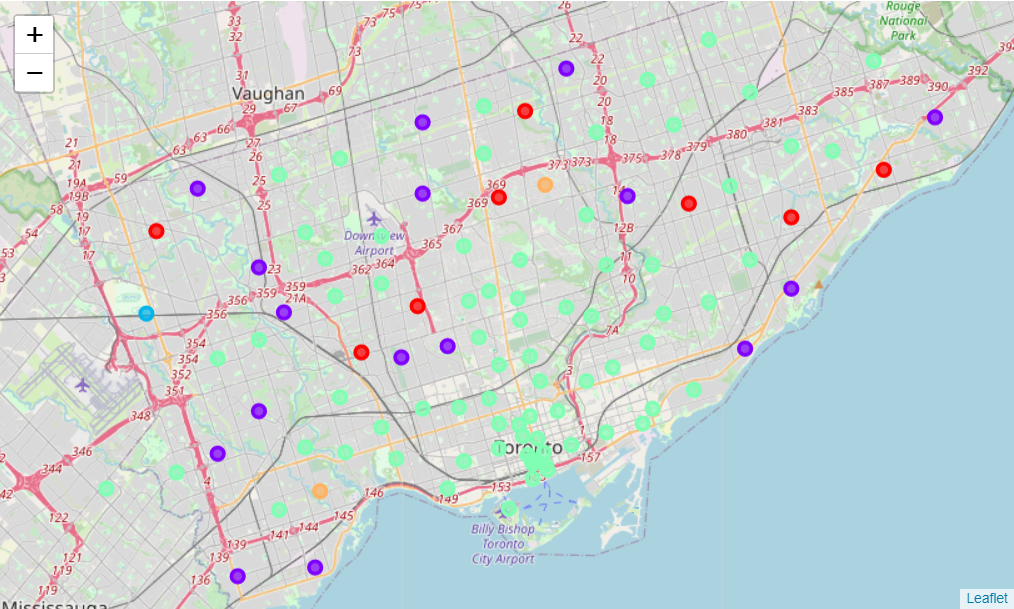


After doing this, I used kmeans to separate the data into five clusters. I then joined the result with the original data set:



**4. Results**

This map below show the neighborhoods split into 5 clusters. Each cluster shows neighborhoods with similar venues. There are two outliers in the data shown in blue and orange.



**5. Discussion**

Looking at the map, it seems most of the areas closer to downtown Toronto (nodes in green) are very similar to each other. This makes sense as city centers have many attractions and restaurants. One thing I should have done is build a formula to find the perfect number of clusters. This would have found the ideal number instead of looking to split the data in a predetermined number of groups. Using the postal code data also wasn’t ideal due to how different neighborhoods can have the same code. It made the data slightly messy. I also wish to have done some accuracy scores to further bolster my findings.

**6. Conclusion**

To conclude, the kmeans algorithm was very effective in producing clusters based on venues in each neighborhood. The maps do an amazing job of showing which areas are similar. This data can be helpful for a small business owner trying to find out where to place their business. It would give them a chance to analyze areas where similar venues are to compete with those venues. Usually, very unique venues like restaurants with food that is not local tend to cluster in cities. A city planner could also use this data to see the layout of a city at a glance. The placement of parks and other types of public offerings can be worked around certain neighborhoods.