Constructing a New Park in the City of Toronto

IBM Data Science Professional Certificate – Applied Data Science Capstone

Yu Hou

December 4, 2020

Business Problem

- Background: Toronto is an international centre of business, finance, arts, and culture, and is recognized as one of the most cosmopolitan cities
- Task: To select a location (preferably with low density of parks) to build a new park as a place for residents' daily activities
- Method: This project will predict the location of the new park from the perspective of data and algorithms, and provide suggestions for the location of the new park by visualizing the current parks in Toronto
- Target Group: Government, or other construction agencies

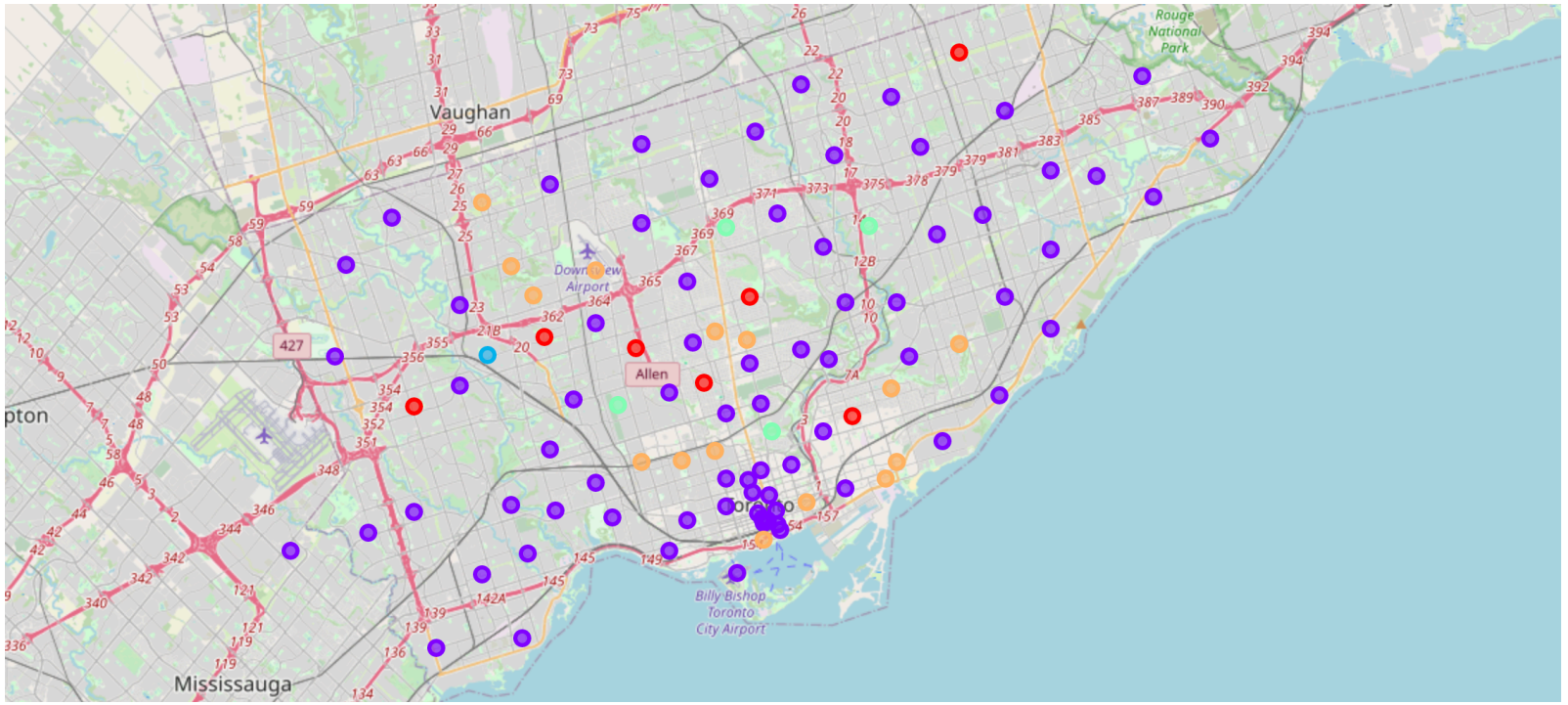
Data Source

- The information of neighborhoods, boroughs, postal codes of Toronto can be extracted from this wikipedia webpage: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Moreover, we need data of latitude and longitude to pin down each venue in Toronto, which can be retrieved from this website: https://cocl.us/Geospatial_data. This website provides a csv file that can be directly imported into notebook
- Most importantly, the venue data can be acquired using Foursquare API, a platform that provides geographical data.

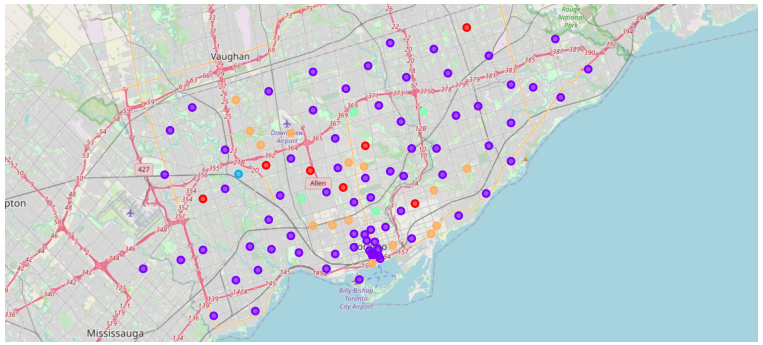
Methodology

- **Web Scrapping**: Utilize *requests* and *pandas* package to scrap and clean the data. After cleaning data, we can have 103 valid rows of data
- **Latitude and Longitude Data:** By using *geocoder* library, we can append the geographical data to the data frame
- **Data Preprocessing:** After initializing Foursquare API, we can retrieve data of specific location (i.e., parks) to check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we can group rows by neighborhood and take the mean of occurrence of each category
- **Modeling**: we perform clustering algorithm on the data by using k-means clustering, which is a very widely used unsupervised machine learning algorithm and can be applied to our case to characterize the pattern of park distribution in Toronto.

Results


Vaughan
Mississauga
Allen
427
Billy Bishop
Toronto
City Airport
Rouge
National
Park

Discussion



- Cluster 0 (in red): 7 neighborhoods are in cluster 0, showing moderate number of parks
- Cluster 1 (in purple): 73 neighborhoods are in cluster 1, showing the highest number of parks
- Cluster 2 (in blue): only 1 neighborhood is in cluster 2, showing the lowest number of parks.
- Cluster 3 (in green): 4 neighborhoods in cluster 3, showing low number of parks.
- Cluster 4 (in orange): 15 neighborhoods are in cluster 4, showing high number of parks.

Discussion

- Based on the results of our modeling, we can observe that cluster 2 shows the lowest number of parks, and includes only 1 neighborhood, which is Weston. This represents this neighborhood can be a target area. If the government or other construction agencies have interests in building a new park, this site can be considered as a beneficial location.

Conclusion

- In this project, we explored the potential site for building a new park in Toronto. Suppose that the City of Toronto is preparing to build a new park as a place for residents' daily activities. This project predicted the location of the new park from the perspective of data and algorithms and provide suggestions for the location of the new park by visualizing the current parks in Toronto. By applying k-means clustering algorithm, we identified a neighborhood with lowest number of parks – Weston, which can be considered as a suitable and potential site for the construction.