Capstone Project:

The Battle of Neighborhoods

Soumyadeep Bhattacharjee

Introduction

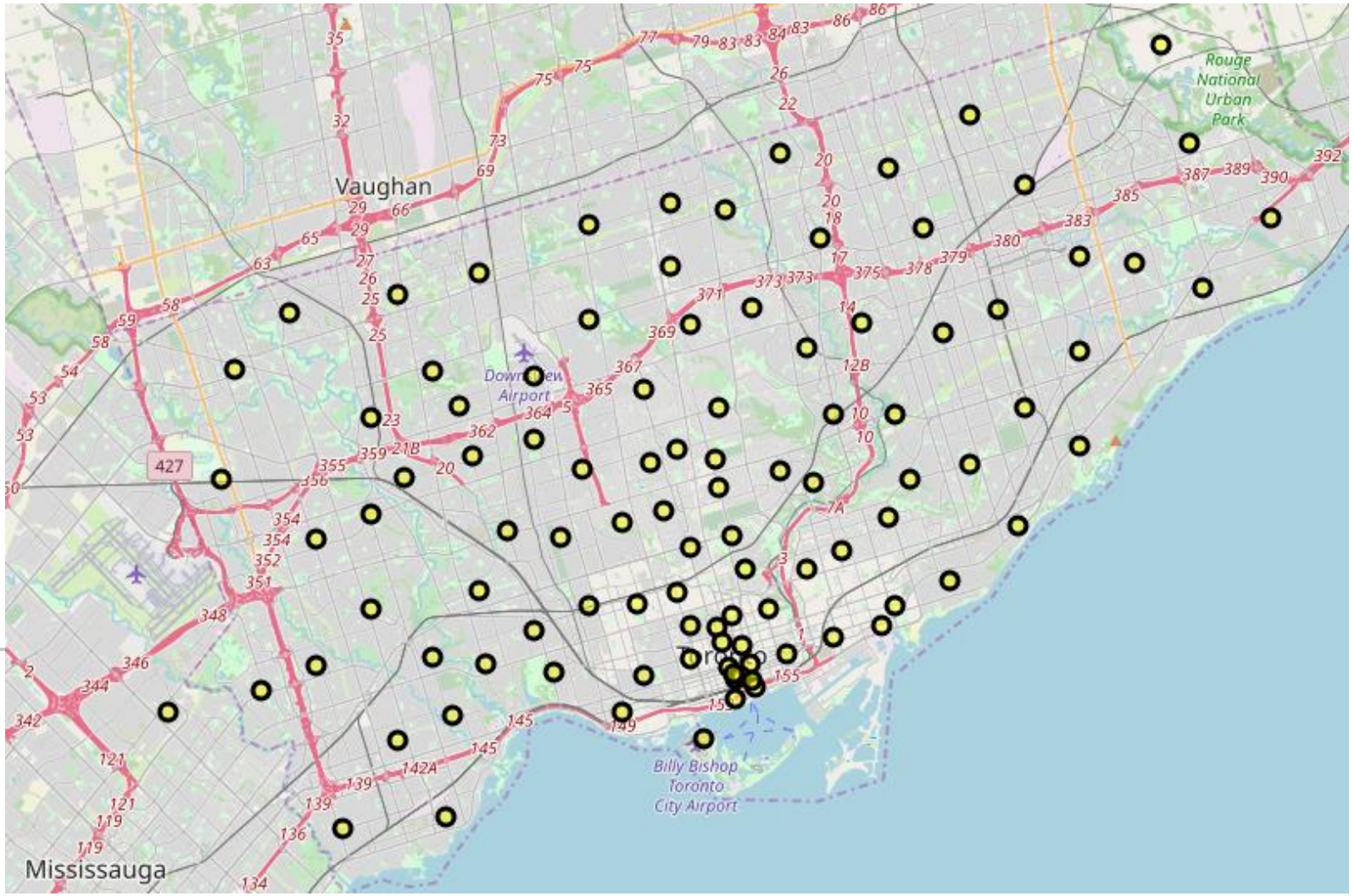
- To allow for an easier decision-making process, neighborhoods can be clustered into prospective zones based on popularity/availability of venues.

- These clusters would help in identifying the ideal neighborhood based on the use case.

Interest Groups

- Expats wishing to relocate to a new area
- Budding entrepreneurs or Businesspersons wishing to find a neighborhood to start a new Store, Mall or Restaurant
- Real Estate agencies who wish to identify zones and ascertain House prices based on the popularity of the neighborhood.

Data Acquisition



Data Acquisition

- Data regarding the List of postal codes, Boroughs and Neighborhoods of Toronto, Canada was obtained by scraping data from **Wikipedia**
- **Geospatial Data** corresponding to each postal code was obtained from the link provided in the course
- All nearby venues within 1 KM radius of each neighborhood was collected using the explore call to the **Foursquare API**

Data Cleaning

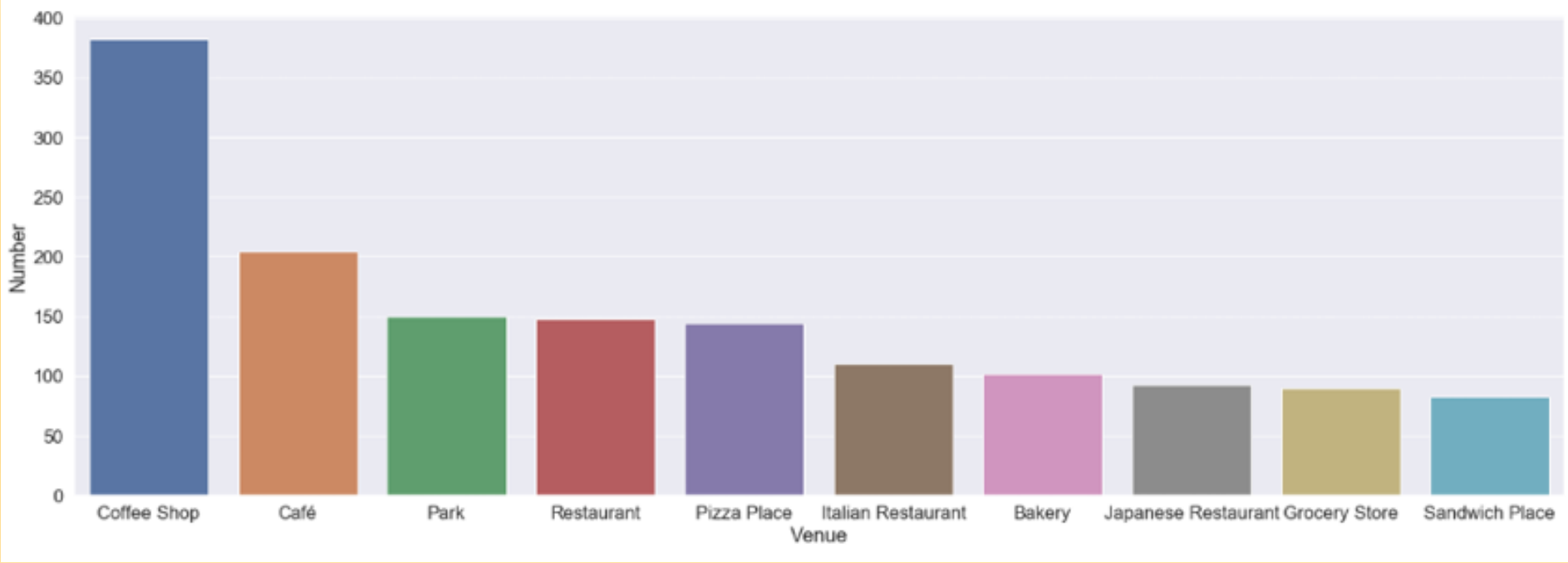
- Removing newline characters
- Replace '/' with a ',' as a separator for multiple neighborhoods
- Removing Rows where the Borough is 'Not assigned'
- Removing extra spaces

Methodology

- In order to identify the most frequent venues to choose to cluster neighborhoods, exploratory data analysis was performed on the list of venues obtained from the Foursquare API query.
- All unique venues were identified, and their frequency of occurrence was stored in a dataframe.

Exploratory Data Analysis

- The top 10 most frequent locations were plotted as a bar graph:



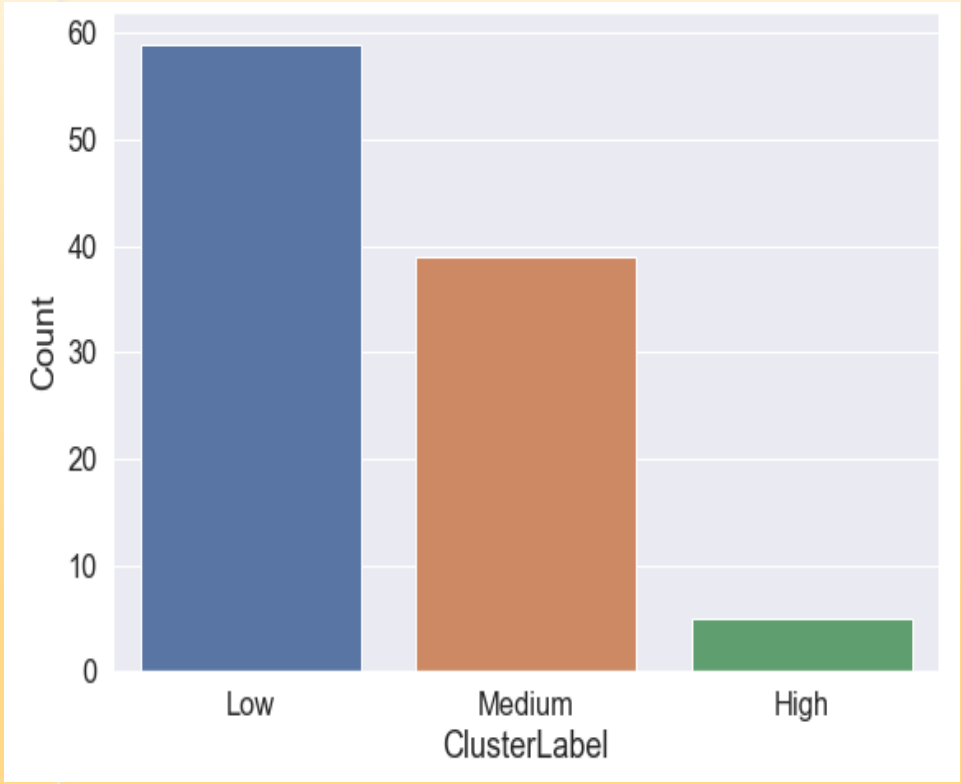
Clustering using K-Means Algorithm

- Cluster similar neighborhoods based on the venues/amenities available around a 1 Kilometer radius using K - means clustering algorithm

- We will use a cluster size of 3. Each of the clusters would signify the popularity level of a neighborhood:
    - Low
    - Medium
    - High

Results

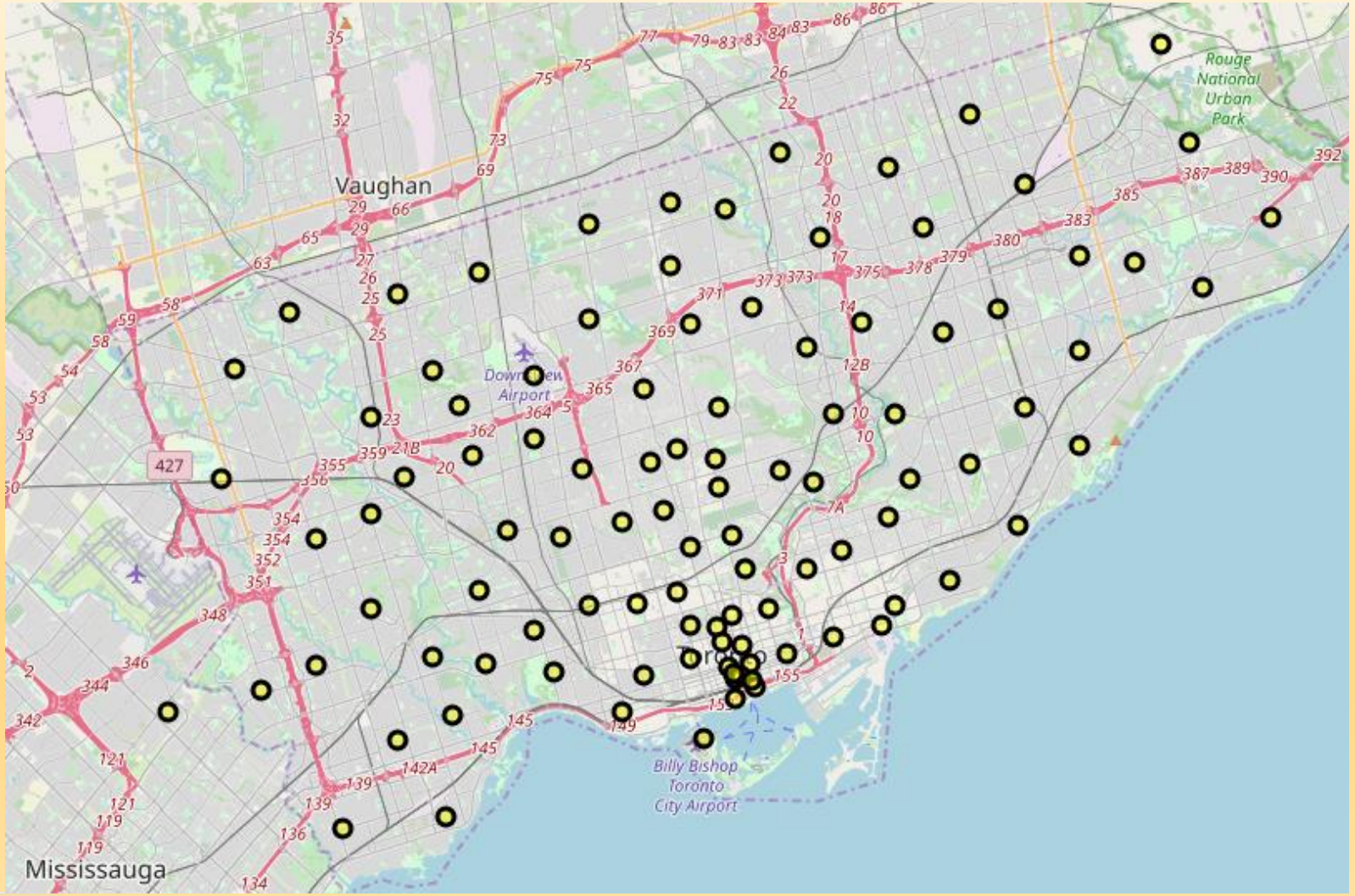
We can visualize the number of each Clusters in a bar graph



Cluster	Label	Color
0	Low Availability	Blue
2	Medium Availability	Orange
1	High Availability	Green

Results

- These are the existing neighborhoods marked on the map of Toronto



Results

- We can now visualize the clustered Neighborhoods based on the color coding:

  Blue: Low
  Orange: Medium
  Green: High

Discussion

- If a person wishes to live in a quiet neighborhood with less footfall, Cluster 0 (Low) could be the ideal place.
- Similarly, a new Restaurant can be opened in any of the Cluster 0 neighborhoods as there is a scarcity of Restaurants in this zone.
- If one wants to relocate to a more happening neighborhood, Cluster 1(High) would be his choice.

Conclusion

- This project aims to provide a solution to a User to get a better analysis of the neighborhoods with respect to the availability of venues.
- In future, this idea can be extended to many domains which leverage geographical data.
- Can be developed as a Web or Mobile Application while keeping the basic idea similar to what was described