Data

Data Sources

Solving the business problem at hand will require information about the characteristics of neighbourhoods in both Toronto and New York City. There are two main sets of data for Toronto and New York City that were examined to help solve the business problem: 1) listings of neighbourhoods, and 2) listing of existing venues. For both data sets, spatial coordinates were provided as latitude and longitude.

According to the City of Toronto Web site, Toronto has 140 city-designated neighbourhoods [1]. Neighbourhood listings with coordinates are provided for Toronto in .csv format from the following link:

> https://open.toronto.ca/dataset/neighbourhoods/

For New York City, location data on neighbourhoods is provided by New York University (NYU) at the at the following link in JSON format:

> https://geo.nyu.edu/download/file/nyu-2451-34572-geojson.json

In each neighbourhood location data set, spatial coordinates are provided as latitude and longitude.

Data on venues such as restaurants, shops, and sight-seeing attractions were obtained from Foursquare (https://foursquare.com/). In particular, the `explore` function in Foursquare's RestAPI provides listings of venues within a specified radius of given latitude and longitude coordinates. Each venue is listed with a category that describes the type of venue (e.g. "Japanese Restaurant", "Movie Theatre", "Park", etc.). Those categories and their hierarchical organization are described by Foursquare's documentation here:

> https://developer.foursquare.com/docs/resources/categories

Data Preparation

Neighbourhood Location Data

Data on the names and locations of neighbourhoods from the Toronto Open Data portal was well structured and clean. Each row represented one neighbourhood. While there were several fields contained in the data set, the only fields kept were the those listed in Table 1.

Field	**Description**
AREA_NAME	Neighbourhood name
LONGITUDE	Longitude coordinate
LATITUDE	Latitude coordinate

Table 1: Toronto neighbourhood location data set fields used

The locations of Toronto neighbourhoods are plotted in Figure 1. There were 140 different neighbourhoods for Toronto

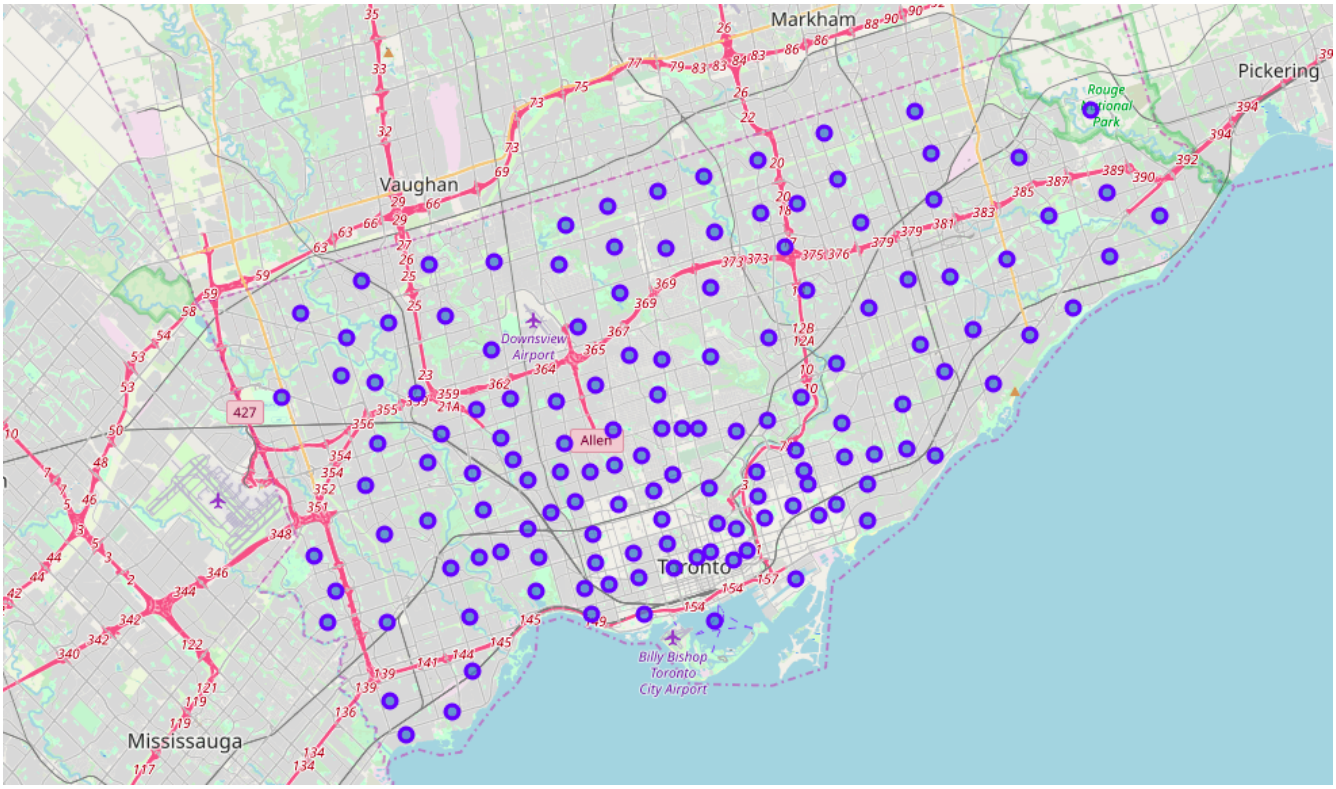

Figure 1: Map of Toronto neighbourhood locations.

For New York City, the data from NYU was also well structured in JSON format and required no cleaning. However, the JSON format was adapted to a Pandas DataFrame to allow for analysis. While there were several data fields for each neighbourhood listed, only the fields listed in Table 2 were retained.

Field	**Description**
borough	Borough name
name	Neighbourhood name
coordinates[0]	Longitude coordinate
coordinates[1]	Latitude coordinate

Table 2: New York City neighbourhood location data fields kept

There were 306 neighbourhoods in the New York City data set. A map of New York City neighbourhood locations is shown in Figure 2.

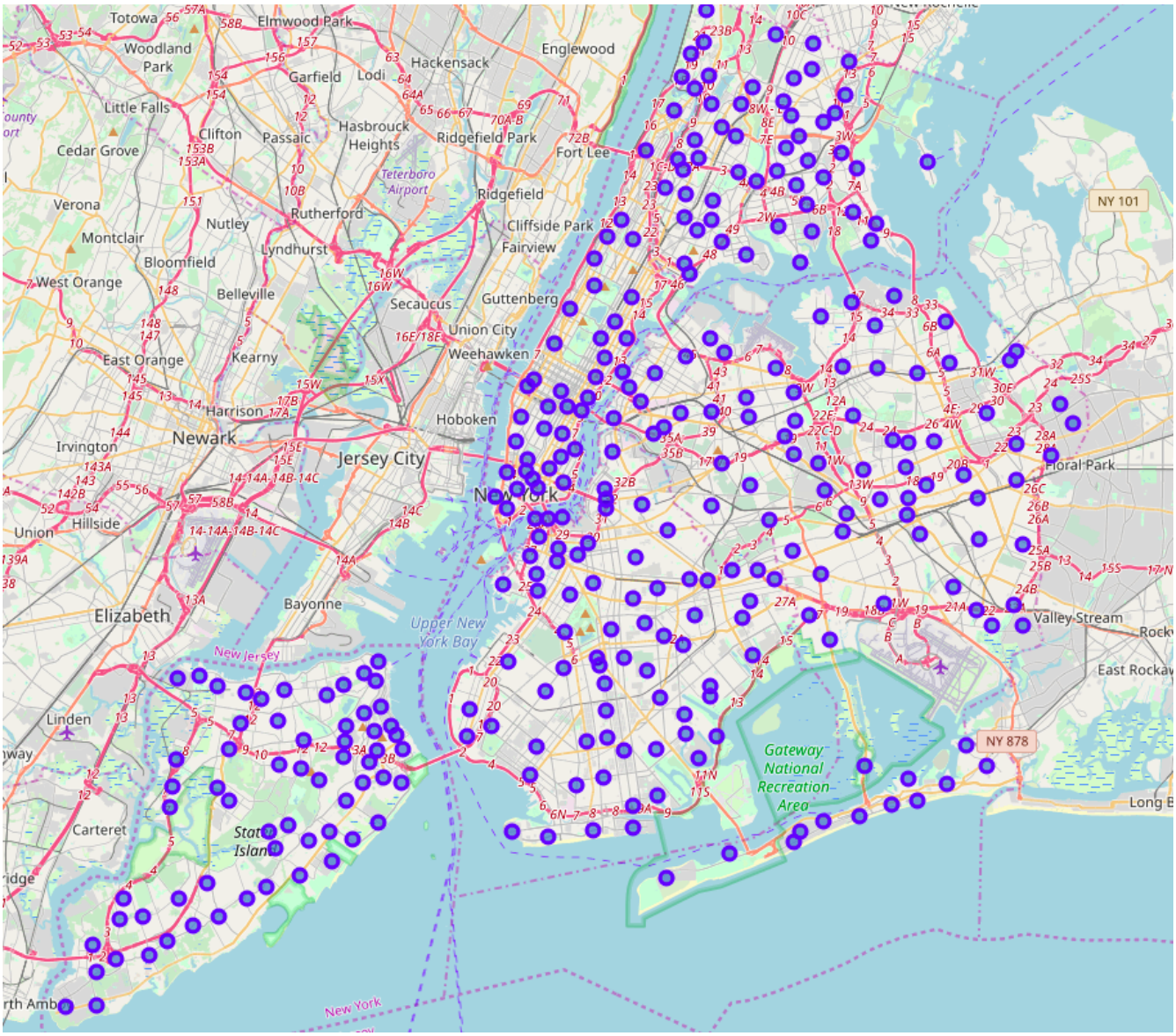

Figure 2: Map of New York City neighbourhood locations

Neighbourhood Venues Data

Data on venues within each neighbour were obtained via the Foursquare by calling the `explore` RestAPI function for each neighbourhood in both Toronto and New York City. The latitude/longitude of each neighbourhood was used as the center point for each query, and venues within 500 m of the center point were requested. Note that some neighbourhoods are separated by less than 500 m, hence there was some overlap in the returned venues. While it could be said that a particular venue belongs to one neighbourhood and not another, the aim of the analysis is to characterize neighbourhoods based on the venues that are within quick reach of a neighbourhood, not based on what is contained only within their boundaries.

Data returned from the `explore` RestAPI function is in JSON format. The following fields for each venue were extracted:

Field	**Description**
name	Venue name
location.lat	Latitude coordinate of venue
location.lng	Longitude coordinate of venue
categories[0].name	Venue category (first entry)

Table 3: Foursquare venue data fields used

Overall, there were 2102 venues found for Toronto, and 10301 venues for New York City.

The key features used in further analyses are the venue categories. For each venue, only one category was returned for each venue and was listed as the primary category. Between both Toronto and New York City, the venues returned belonged to one of 462 unique categories. In early analysis, this large number of categories proved to be problematic when attempting to cluster the neighbourhoods using the k-means algorithm. It was determined that using less categories as the feature set could improve the clustering algorithm and the overall data analysis. Upon further inspection of the categories, it was found that many were too specific. For instance, a venue categorized as a "Cantonese Restaurant" is just a more specific type of Chinese restaurant. This high level of specificity resulted in many venue categories having few members. Hence, the feature set of the data was very sparse. Instead of using the listed primary venue categories as the features, it was determined that generalizing the categories would allow for better grouping of the data into fewer features that were more usable.

Foursquare lists a hierarchy of categories that allows for generalizing the categories for each venue. It can be obtained through the Foursqaure `category` RestAPI function. With this hierarchy, the parent categories for each venue's listed category were found. For this analysis, the parent category used was the second category level down from the top level. For instance, the category "Cantonese Restaurant" was mapped to a parent category of "Asian Restaurant", which is one level down from the top level category "Food". After adding the parent categories to each venue entry in the data set, the number of categories to deal with was reduced to 299.