New York University

University of Toronto

Capstone Project | The Choice

By Emma K

May 2019

Coursera IBM Data Science Professional Certificate

Introduction / Background

Background:

DrivenByData is a new startup company that provides its clients detailed analysis and observations backed on relevant data, to help its subscribed users in making informed decisions.

Business Problem:

There is a recent request from Charlotte who resides in Paris, France. She is an international student wanting to pursue her post graduate degree in Fine Arts abroad. She received acceptance letters from both New York University | NYU (U.S) and University of Toronto | UofT (Canada). Charlotte who's never been to either cities before, would like to make an informed decision on which school to attend based on the proximity and availability of art related venues (such as performing centres, art galleries, art stores, etc.) to the university where she will attend classes. Charlotte would also like to have a choice of neighbourhoods based on her preference to be considered for residence.

Data

Data Description:

- FourSquare explorer API to search for Art related venues around NYU and UofT in their respective cities of New York City and Toronto.

- Publicly available dataset that contains neighbourhoods data of Manhattan (and New York City) at https://cocl.us/new_york_dataset

Approach to solve the problem:

- Convert addresses into their equivalent latitude and longitude values.
- Use the Foursquare API to explore neighbourhoods in New York City and Toronto
- Use the Foursquare API explore function to get the most common venue categories in each city
- Group the neighbourhoods into clusters using k-means clustering algorithm

Methodology

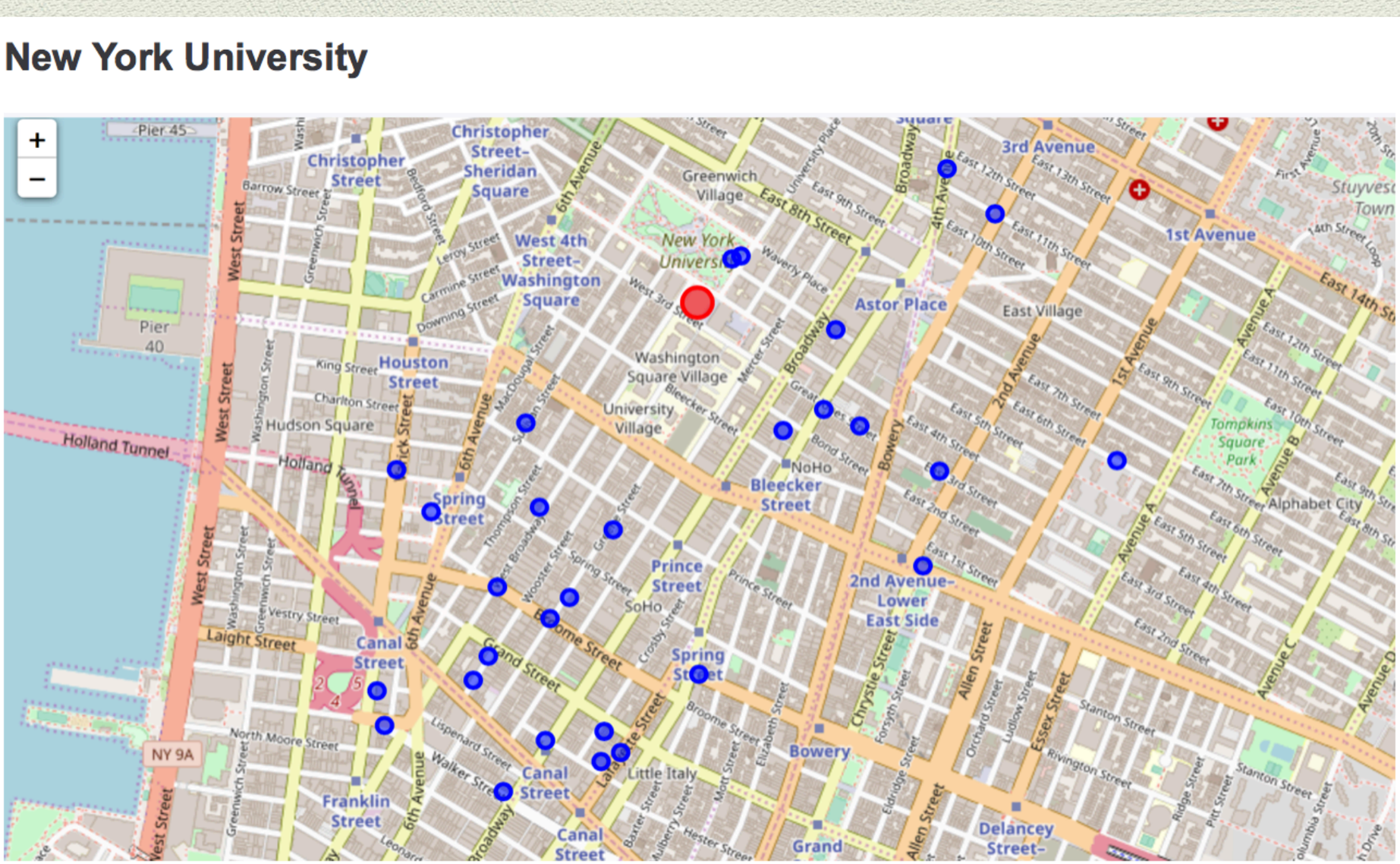
Data preparation

- Use the Nominatim module to convert an address into latitude and longitude values
- Foursquare API requests are built to obtain results with Art related venues in the vicinity of each university (radius of 1000m). The search query used was the word ‘art’ along with the geo coordinates.
- The resulting json response was analyzed based on the ‘venues’ component, then normalized into a dataframe. This was done for both universities locations.
- In addition, the dataframe was filtered on venue category and selected the following as they are of interest to our subject audience:

  - Performing Art Venues
  - Art Gallery
  - Arts & Crafts Store

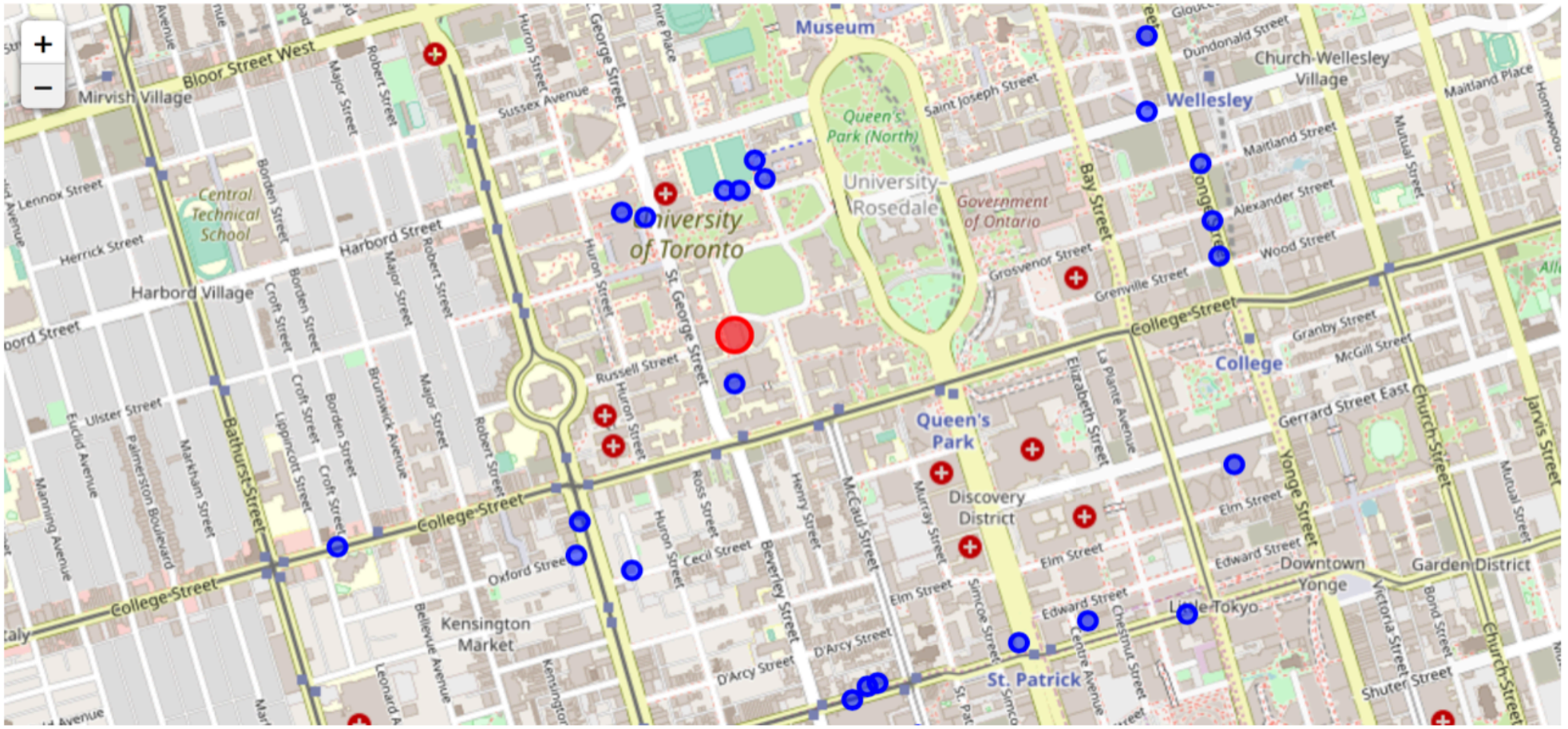
Exploratory Analysis

Red marker indicates the campus, and the blue markers are the art venues.



Exploratory Analysis

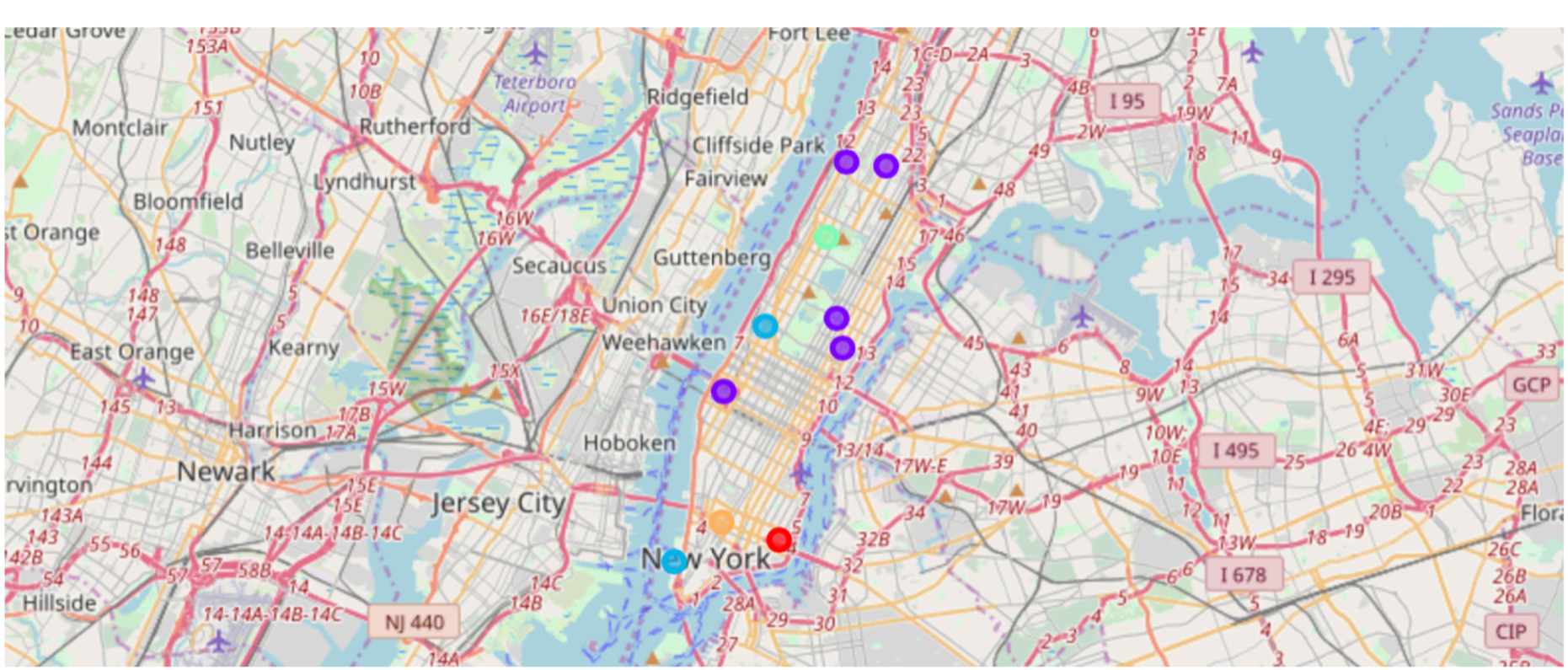
Red marker indicates the campus, and the blue markers are the art venues.



Machine Learning

Clustering is a technique for finding similar grouping in a data, also referred to as clusters. It attempts to group by similarity, but not driven by a known goal/purpose. Clustering is an unsupervised learning, as you don't have prescribed labels in the data.

The algorithm was run for 5 clusters, here's a visualization of all on the map where each colour represents one cluster.



Results

Comparing the surrounding neighbourhoods of both NYU and UofT, based on the exploratory analysis, one can easily conclude:

- there is no shortage of art related venues, galleries and art supply stores around both campuses.
- NYU has a total of 18 and UofT has 16, very comparable.
- UofT campus has 6 art venues within 500 m radius compared to NYU with only 4.
- 4 out of 6 venues are within the UofT campus and only 1 out of 5 was in NYU campus.
- This insight indicates that our choice towards NYU is the better one for the client.

To address our client's request, we also had to provide an analysis of neighbourhoods with greater options of having nearby art venues, galleries and art supply stores, for her to reside in. This is where clustering of the New York city was very helpful. The only cluster that had all three types of art venues: Performing Art Venue, Art Gallery and Arts supply store was Cluster #2 with 5 different neighbourhood choices for our client.

Conclusion

This analysis depended on the accuracy of the data returned by Foursquare API.

The client's request for having art related venues nearby was assumed to be venues of categories belonging to either a Performing Arts, Art Gallery and Arts & Crafts stores. The results could have been different if other venue categories were considered.

A similar analysis on the Toronto neighbourhoods could also be done to improve the level of confidence in the recommendation.

The analysis could be improved by adding data on demographics and rent prices for the resulting neighbourhoods.