Capstone project report

# Introduction

New York City is the US’s top city for small business for the second year in a row, according to Biz2Credit’s annual study of the Top Small Business Cities in America, which analyzed the financial performance of 27,000 small businesses and their local market economic conditions.

A Canadian restaurant owner who has multiple restaurants in Toronto is planning to expand to the US. Seeing the potential of NYC, he decides to open a new restaurant in this city. But there is a question: Where should the new restaurant be located?

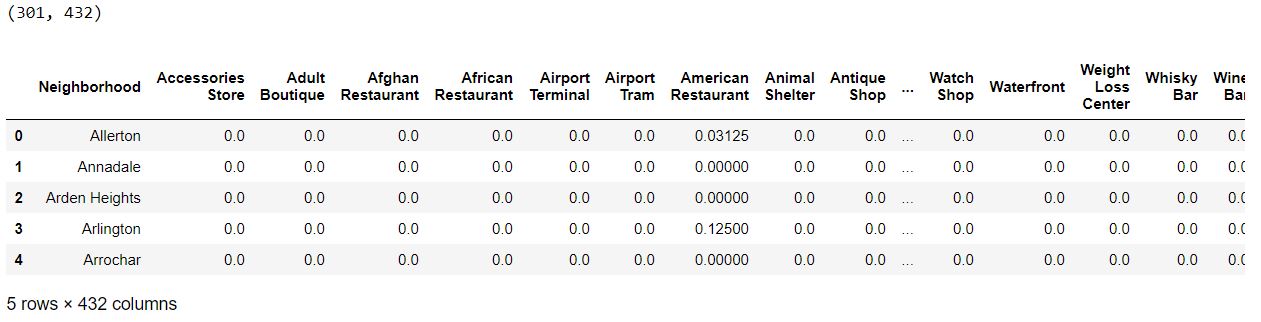
The main purpose of this report is showing a Data Science approach to solve that question.

# Data

Based on the purpose of the project, New York City neighborhoods and client’s restaurants in Toronto were chosen as the observation target. We collect the data using FourSquare API which provides the surrounding venues of a given coordinate.

The process of collecting and clean data:

* Find the geographic data of the NYC neighborhoods and client’s restaurants in Toronto.
* For each neighborhood or restaurant, pass the obtained coordinates to FourSquare API. The “explore” endpoint will return a list of surrounding venues in a pre-defined radius.
* Count the occurrence of each venue type, then apply one hot encoding to turn each venue type into a column with their occurrence as the value.
* Then merge NYC neighborhoods dataset with client’s restaurants dataset. The resulting dataset is a 2 dimensions dataframe:



# Methodology

The assumption is that a neighborhood in NYC which is similar to the one that contains the client’s restaurant in Toronto is a good place to open a new restaurant. Thus, the clustering technique will be used to analyze the dataset. In the end, we will find good neighborhoods in NYC which suits the client’s purpose.

Python data science tools will be used to help analyze the data.

In order to have the first insight into our data, we calculate the top 10 venues in each NYC neighborhood and each client’s restaurant.

**Client’s restaurant:**

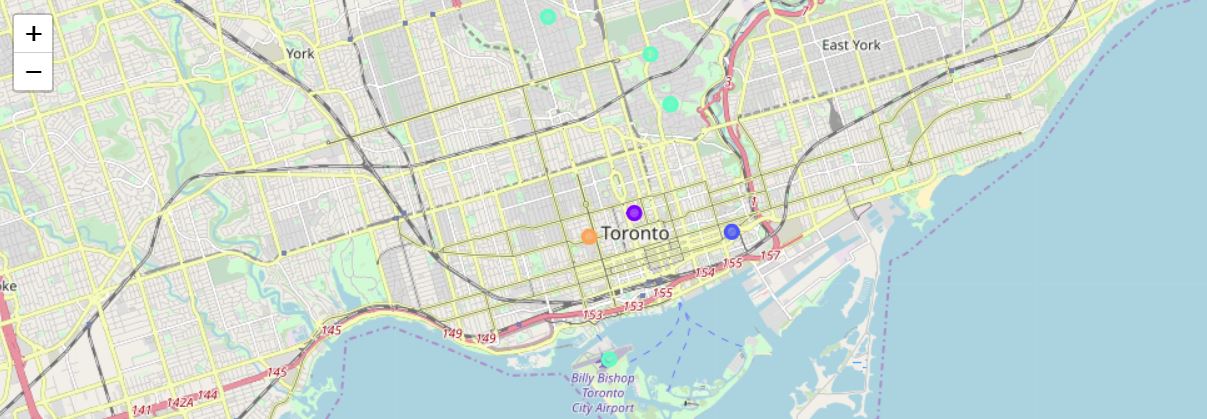


**NYC neighborhoods:**

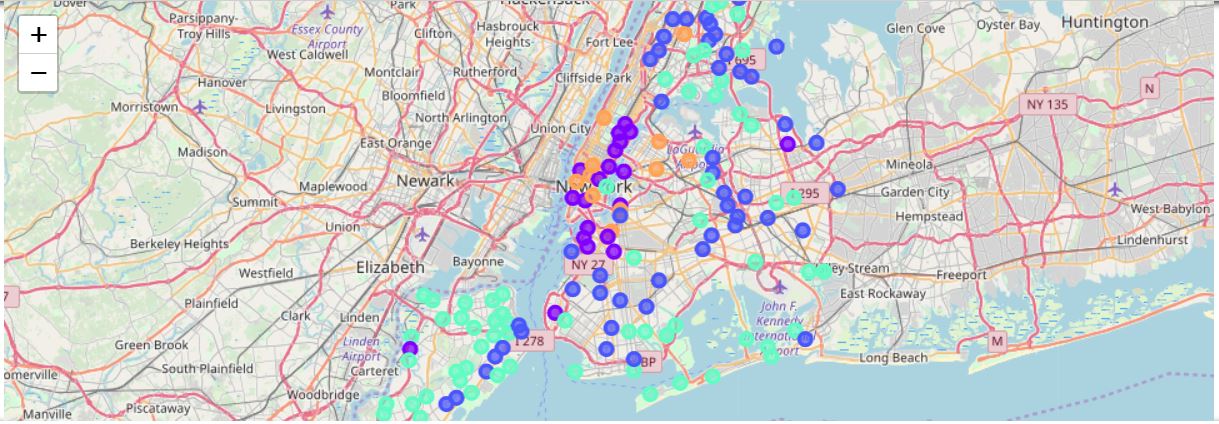
****

Then we use k-means clustering technique to find similar groups of neighborhoods. And plot the groups that contain the client’s restaurant.

**Toronto:**



**NYC:**

****

# Results

Based on 2 plots about the similar groups of client’s restaurant and NYC neighborhoods, we have a conclusion that if neighborhoods A in NYC is in the same cluster of client’s restaurant B in Toronto, then the client should open a new restaurant in A with the same business strategy that he applied in B.

# Conclusion

It’s unfortunate that the analysis couldn’t produce a precise model or showing any strong coefficient correlation for any venue type. But we can still get some meaningful and logical insights from the result.

Doing this project helps to practice every topic in the specialization, and thus, equipping learners with Data Science methodology and tools using Python libraries. Also doing a real project certainly helps one learns so much more outside the curriculum, as well as realizes what more to research into after completing the program. And as this report shows, there are surely a lot of things to dig into.

Toward the person that went through this project, many thanks for the time and patient.