**Capstone Project - The Battle of Neighbourhoods (week 2)**

**Introduction/ Business Problem**

**1. Introduction**

**1.1 Background**

Data Science problems always target an audience and are meant to help a group of stakeholders solve a problem. The idea of this study is to help people planning to open a new restaurant in New York to chose the right location by providing data about the top venues visited by people on the same regions. We analyze the neighbourhood and the venues visited by them. Sorting the data by the top visited venues and giving them a simple idea to choose a venue to start their restaurant.

**1.2 Problem**

Data that might contribute to determining the most visited places of people in a particular region including the type of places they visited. The effect of places visited and the corresponding neighbourhood. This project aims to help a person who wishes to start a restaurant in a particular region by providing him the data of various factors contributing to the success of his business.

**2. Data acquisition and cleaning**

**2.1 Data sources**

First and foremost Borough, Neighborhood, Latitude, Longitude where the basic data collected to analyse data about the newyork city. The first data was imported as a json file and used to analyze. Then using the foursquare url we got the top venues, venues in the city, Name of the venue, the count of no of times visited and many more. The dataset for the city of newyork is got from <https://geo.nyu.edu/catalog/nyu_2451_34572?cm_mmc=Email_Newsletter-_-Developer_Ed%2BTech-_-WW_WW-_-SkillsNetwork-Courses-IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork-21253531&cm_mmca1=000026UJ&cm_mmca2=10006555&cm_mmca3=M12345678&cvosrc=email.Newsletter.M12345678&cvo_campaign=000026UJ&cm_mmc=Email_Newsletter-_-Developer_Ed%2BTech-_-WW_WW-_-SkillsNetwork-Courses-IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork-21253531&cm_mmca1=000026UJ&cm_mmca2=10006555&cm_mmca3=M12345678&cvosrc=email.Newsletter.M12345678&cvo_campaign=000026UJ>

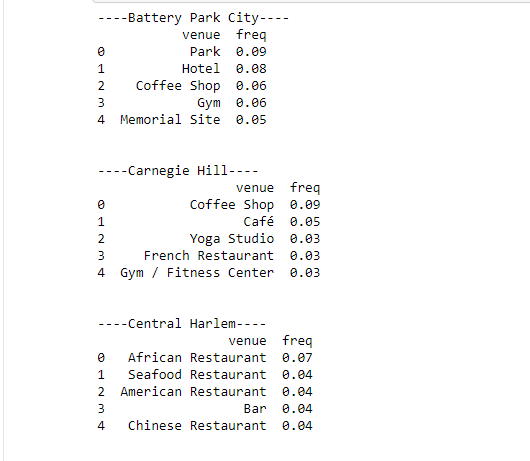
**2.2 Data Cleaning**

Data downloaded or scraped from multiple sources were combined into one table. And then they were put into dataframe to analyze data for a particular region.

**2.3 Feature selection**

After data cleaning, there were nearly 13000 samples and more than 40 features. After examining the meaning of each feature, the useful and main features were used like the neighborhood, latitude and longitude, venues, count of no of times visited according to region were selected and used. Few important features were selected and used.

**Relationship between Neighbourhood and Stores.**



**Conclusion:**

I was able to achieve somewhat suggestion for predicting the correct restaurant classified according to the region in the classification problem. However, there was still significant variance that could not be predicted by the models in this study. I think the models could use more improvements on capturing regions climatic conditions and population traits. Models in this study mainly focused on neighbourhood features. However, population and annual income also contribute to a restaurants customers and income. These data are obviously more difficult to extract and quantify, but if optimized, could bring significant improvements to the models.

**Future Direction:**

Can extract even more features like income of people in a particular region and population. The effect of population, income of people in a region to the profit of a restaurant in a region to bring even more accuracy in a model.