Coursera Capstone

IBM Applied Data Science Capstone Project

Cluster the places in New York City for the tourists to get a better idea about the places it has to offer on basis of their categories

By: Shibajyoti Debbarma



**Introduction**

The City of New York is the most populous city in the United States. New York City witnessed an eighth consecutive annual number of approximately 60 million tourists in 2017 including commuters. Most of the people arriving for tourism purposes are International residents. Major destinations include places such as the Empire State Building, the Statue of Liberty on Liberty Island, Central Park, Times Square, Museums, luxury shopping along Fifth and Madison Avenues, entertainment events such as the Tribeca Film Festival and free performances on streets. New York City has an area of 783.8 km².

**Business Problem**

With such a huge population of tourists increasing every year, the people visiting the place face a very significant problem as to where to visit first and how to map their stay in the city and make it most productive in order to eat in good restaurants and enjoy the time in New York City properly. The problem can be solved by using Machine Learning Techniques in order to cluster the places according to their similarities or dissimilarities.

**Target Audience of this project**

The project is particularly focused towards the millions of tourists visiting the New York City every year which can be utilised for clustering the places according to their similarities and the rating categories.

**Data**

To solve the problem, we will need the following data parameters:

* Venue ID
* Venue Name
* Venue Location
* Likes
* Ratings Given
* Category
* Phrases

**Data Clustering Method**

The project will be utilising K-Means Clustering Algorithm for grouping the places which can be used by the tourists for their own specific usages.

**Methodology**

- The data is imported by using the Foursquare API by using Developer Tools.

- The data is then pre-processed by correctly labelling the columns.

- Visualise the data using the statistical measurements.

- Pre-process the Likes rating column by defining the ranges.

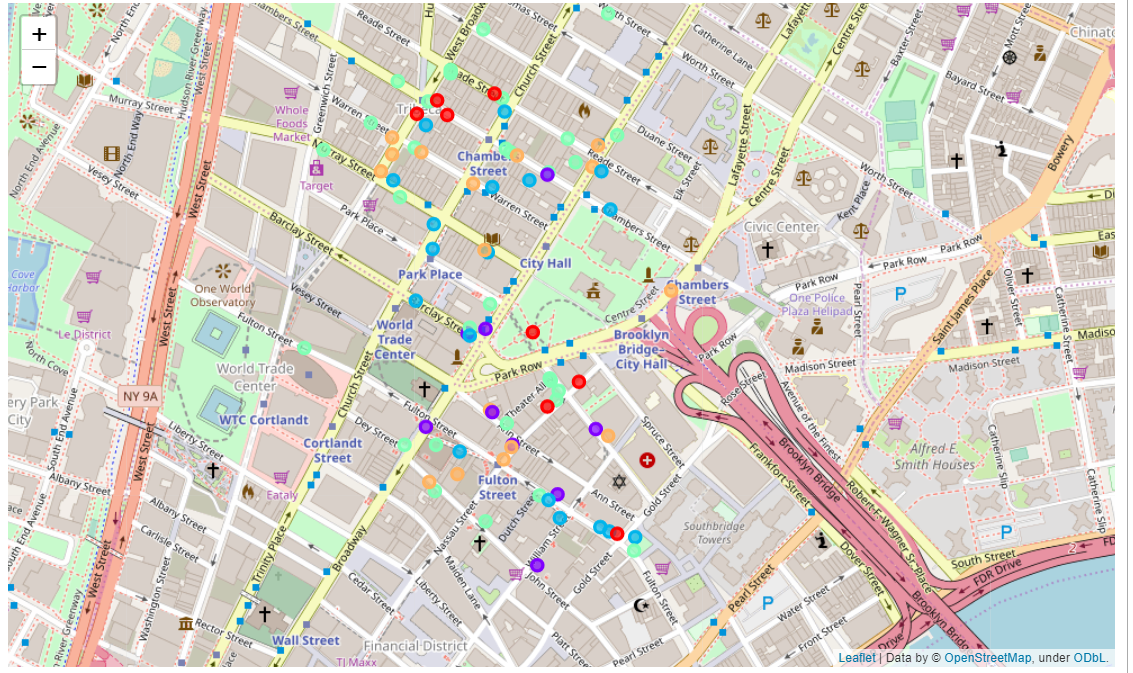
- Pre-process the categories and total likes categories using the One Hot Encoding.

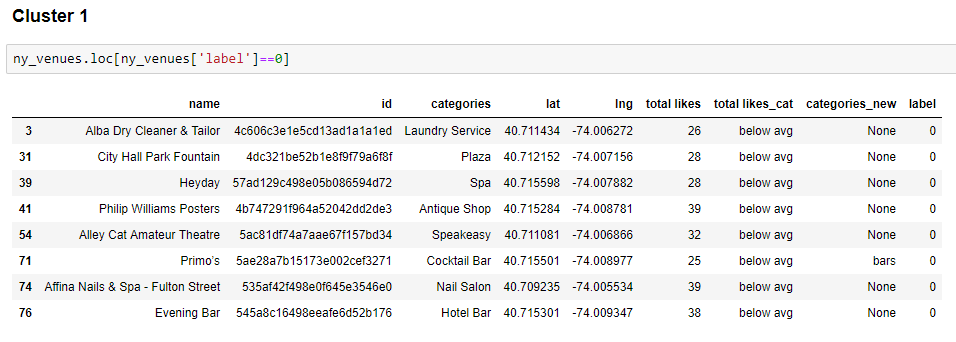
- Perform Clustering using K-Means.

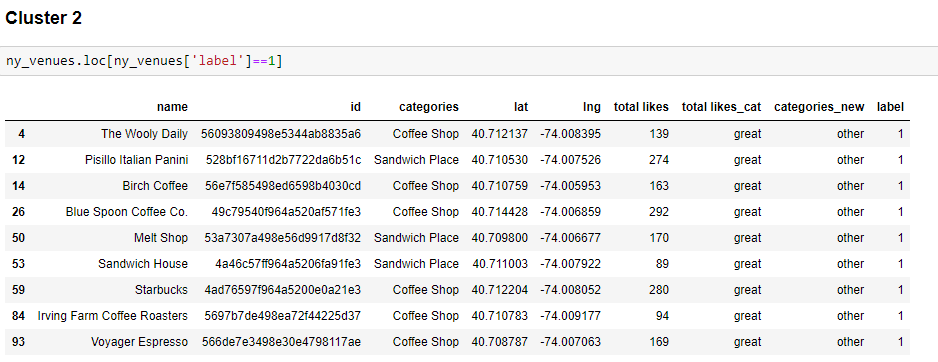
- Visualise the clusters using folium map.

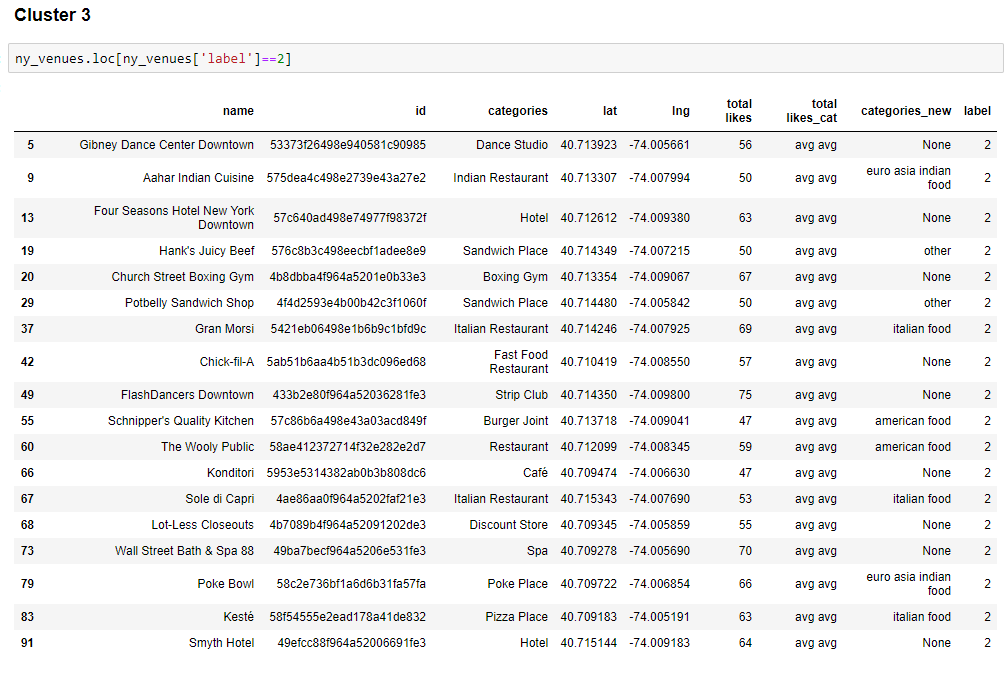
**Results**

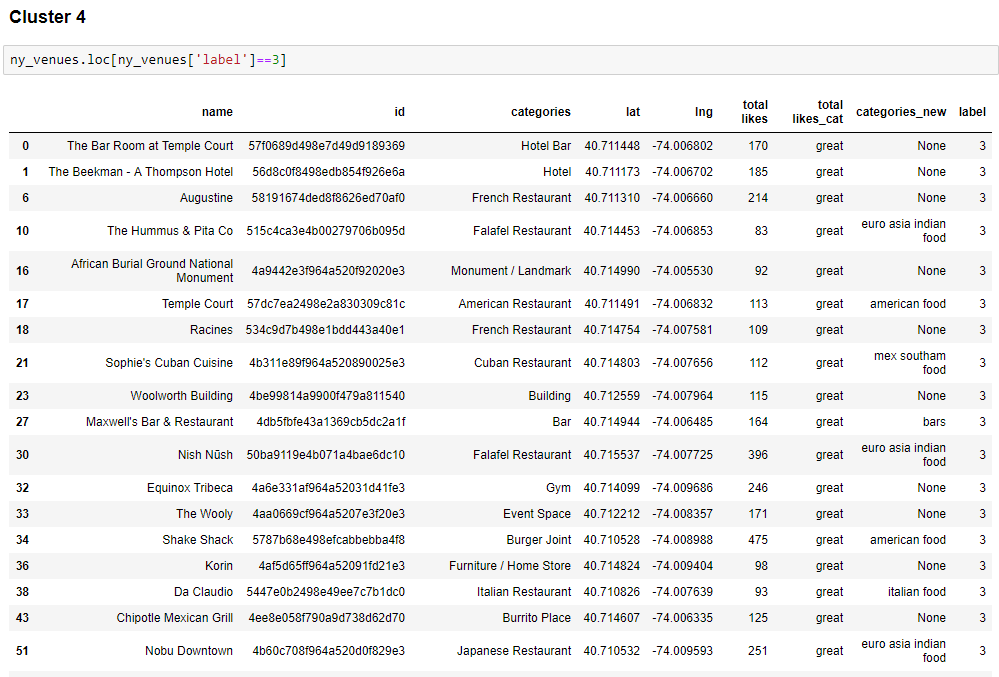
The New York City places are clustered into 5 clusters:

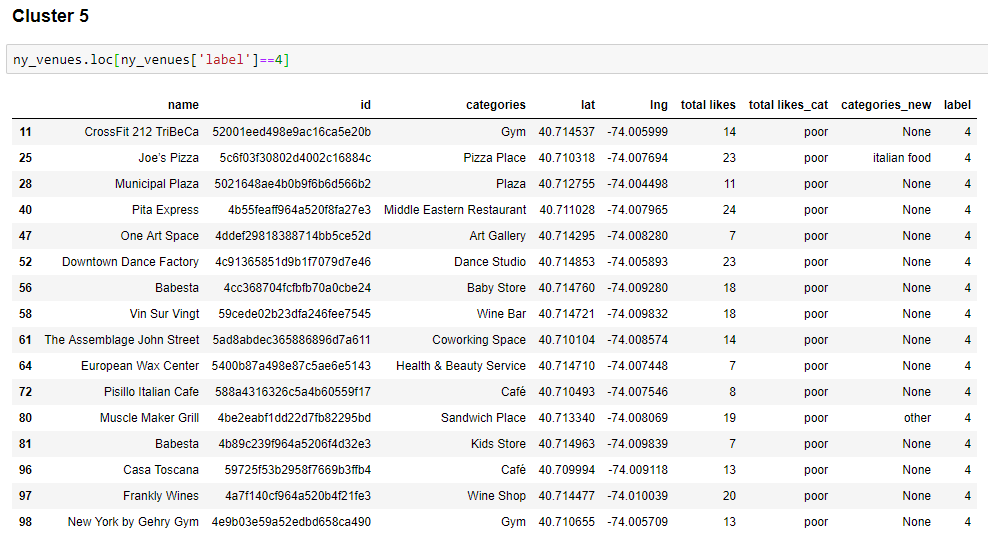


Five unique Clusters classified by the K-Means Algorithm:









**Discussion**

The result with the clustering quite makes sense to the tourists visiting New York City by clustering the place into 5 clusters. The New York City places are clustered into 5 groups on the most popular venues