IBM – Coursera

Data Science Specialization

Capstone project - Final report

Clustering and Analyzing Venues in Manhattan

Tanmay Sah, FRM – 2019

Table of content:

I. Introduction:

This report is for the final course of the Data Science Specialization. A 9-courses series created by IBM, hosted on Coursera platform.

The main goal will be exploring and analyzing the neighborhoods of New York city (Manhattan) and divide the neighborhoods into 5 clusters

The idea comes from the fact that if someone is interested in opening a restaurant, coffee shops and so on then it is important for them to know the frequency of such venues in that area. Based on this clustering it is depend on the person to decide what kind of shops he/she wants to open and in what areas

The target audience for this report are:

- Entrepreneurs- who want to invest in opening shops
- Real estate makers and planners who can decide what kind of venues to put around their products to maximize selling price.
- Houses sellers who can optimize their advertisements.
- Venture Capital funds- who are interested in finding the city venue patterns.

II. Data description:

New York city neighborhoods were chosen as the observation target due to the following reasons:

- Most popular city in the world
- Investors are always interested to invest in New York City
- The availability of geo data which can be used to visualize the dataset onto a map.

The dataset will be composed from the following main source:

- Foursquare API which provides the surrounding venues of a given coordinates.

The process of collecting and clean data:

- Find the geographic data of the neighborhoods. Both their center coordinates and their border.
- For each neighborhood, 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 in a neighborhood.

The result dataset is a 2 dimensions data frame (Figure 1):

- Each row represents a neighborhood.
- Each column, except the last one, is the occurrence of a venue type. The last column will be the standardized average price.

	Neighborhood	Accessories Store	Adult Boutique	African Restaurant	American Restaurant	Animal Shelter	Antiq Shop	...	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio	StandardizedAvgPrice
0	Battery Park City	0	0	0	3	0	0		0	1	4	0	1	0	-1.303912
1	Bedford-Stuyvesant	0	0	0	0	0	0		0	1	6	0	0	1	-0.418350
2	Boerum Hill	0	0	0	1	0	0		0	0	2	0	0	2	0.015011
3	Brooklyn Heights	0	0	0	2	0	0		0	1	4	0	0	5	-1.099479
4	Bushwick	0	0	0	1	0	0		0	0	1	0	0	2	-0.587926

Figure 1 - Final dataset

The dataset has 50 samples and more than 300 features. The number of features may vary for different runs due to Foursquare API may returns different recommended venues at different points in time.Next, we are going to start utilizing the Foursquare API to explore the neighborhoods and segment them**.**