IBM Professional Data Science

## Capstone Project RePort

# Introduction

Chinatown/Penn Quarter is one of Washington DC's most popular neighbourhoods for dining because of its' central location and easy access to many of the city's biggest attractions including the Verizon Center, the National Portrait Gallery & American Art Museum and the International Spy Museum. Chinatown is just a few blocks away and the Washington Convention Center is also within walking distance. The Gallery Place-Chinatown Metro station makes this area easily accessible from across the city. Restaurants in the area offer a wide range of cuisine from contemporary American to Asian Fusion, to Italian or Latin American fare. Chinatown has approximately 20 Chinese and Asian restaurants.

To begin with I am importing all the necessary libraries and installing all the necessary packages. I will be primarily using Pandas, scikit learn, Matplotlib, Geocoders, Foursquare API, Folium

**Business Problem** – A tourist visiting an unknown city for the first time is often looking for places to eat which are popular in that area but due to unfamiliarity with the set of cuisines and preferences of local people is often confused about which are the best restaurants that he must explore.

**Clustering** is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

# DaTASET Used and ITS Source

I have primarily worked with Foursquare API to obtain data of all the venues in Chinatown, Washington DC. In addition to that I have used the same data to obtain their longitude and latitude information as well as the categories of venues. I have also used information of likes to obtain reviews of people for the venues

For this assignment, I will be using the Foursquare API to pull the following location data on restaurants in Chinatown, Washington:

1. Venue Name
2. Venue ID
3. Venue Location
4. Venue Category
5. Count of Likes

For data acquisition I am using below methods:

* Get geolocator lat and long coordinates for Chinatown, Washington
* Use Foursquare API to get a list of all venues in Chinatown
* Get venue name, venue ID, location, category, and likes

After data acquisition using foursquare and geolocator, I will do preprocessing on data to remove the categories which are not Food Joints and then categorize them into 5 major food categories based on the data. I will then using histograms create bins to identify five categories of restaurant ratings based on user likes. I will then create a k-means clustering algorithm that will group restaurants into 5 clusters.

# METHODOLOGY

In order to determine the popular restaurants in Chinatown, Washington, we will be focusing on finding major restaurant clusters in the area.

#### Obtain Location

To begin with we will be using geolocator to obtain the location – longitude and latitude coordinates of the Chinatown in Washington DC

#### Obtain Nearby Venues

* We will then use the Foursquare API to obtain all the nearby venues to the location coordinates
* The Foursquare API’s explore method will return the Location IDs, name, category, and tips information as a json

#### Read Data in Dataframe

* We will read the above data into pandas dataframe which will make it easier to process in the other libraries like scikit and folium
* Clean the data to obtain meaningful Headers
* Read Venue Categories

#### Refine the Data

* For cleaning the data we will obtain all the venue categories
* After checking all the categories we well remove the categories that are not food joints from our list
* This will reduce our target set to those venues which are eateries

#### Categorize the likes data to get Ratings

* Based on the number of likes obtained for each restaurant we will categorize them to provide ratings
* In order to do this we will use histograms to bins the likes data
* After observing and analyzing the data we will define the percentiles to create explicit bins
* These bins will then be converted into ratings category
* Eg - <10 percentile – poor, >95 percentile is best

Number of restaurants vs Likes

|  |  |  |
| --- | --- | --- |
| Percentile | Likes | Category |
| <10 | 11 | Poor |
| 25 | 38 | Below average |
| 50 | 116 | Above average |
| 75 | 194 | Good |
| >95 | 532 | Best |

#### Group the Restaurants Category

* As the data may still contains a number or categories too big to encode we will further refine this data to obtain the major categories of restaurants
* To group the data we will use a bar chart analysis to sea major restaurant types and then divide them into broad meaningful labels of restaurants like – Italian, European , Asian, etc
* We will then add the newly added label and drop the previous one

|  |  |
| --- | --- |
| Categories | New Category |
| 'American Restaurant', 'Restaurant', 'Diner', 'New American Restaurant', 'Burger Joint' | American |
| 'Hotel Bar', 'Wine Shop', 'Cocktail Bar', 'Bar' | Bars |
| 'Taco Place', 'Brazilian Restaurant', 'Southern / Soul Food Restaurant', 'Mexican Restaurant', 'Cuban Restaurant' | South American |
| 'Korean Restaurant', 'Mediterranean Restaurant', 'Asian Restaurant', 'Vegetarian / Vegan Restaurant', 'Ramen Restaurant', 'Japanese Restaurant', 'Tapas Restaurant' | Middle East and Asian Food |
| 'Italian Restaurant', 'Portuguese Restaurant', 'Pizza Place', 'French Restaurant', 'Turkish Restaurant' | European |
| 'Seafood Restaurant', 'Salad Place', 'Steakhouse', 'Juice Bar', 'Ice Cream Shop', 'Coffee Shop', 'Café' | others |

#### COnverting categorical values to codes

* We will use hot encoding methodology to encode the categorical variables
* add neighborhood column back to dataframe
* move neighborhood column to the first column

Since clustering algorithms including kmeans use distance-based measurements to determine the similarity between data points, it’s recommended to standardize the data to have a mean of zero and a standard deviation of one since almost always the features in any dataset would have different units of measurements such as age vs income.

#### CLUSTERING the Data

* Since this is an unsupervised data we will use k-means clustering to obtain clusters
* We used elbow technique to obtain the best K
* We have selected 5 clusters for this data

The way kmeans algorithm works is as follows:

-Specify number of clusters *K*.

- Initialize centroids by first shuffling the dataset and then randomly selecting *K*data points for the centroids without replacement.

- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.

- Compute the sum of the squared distance between data points and all centroids.

- Assign each data point to the closest cluster (centroid).

- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.