**Coursera Capstone**

**IBM Applied Data Science Capstone**

***Opening a New indian restaurant in newyork***

***by :aditya***

***january 2020***



**Introduction**

This final project explores the best locations for Indian restaurants throughout the city of New York. Punjabi style with creamy curries is one of the popular dish. New York is a major metropolitan area with more than 8.4 million (Quick Facts, 2018) people living within city limits. Most of the Indian immigration into the United States occurred during the late18th and early 19th century with over two million immigrants between 1900 and 1910.In 2014, the borough of Queens had a population of 144,896 indianwhich was the highest population of Indians in any New York borough. Other boroughs with significant indian population were Brooklyn and Manhattan.This report explores which neighborhoods and boroughs of New York City have the most as well as the best Indian restaurants. Additionally, I will attempt to answer the questions “Where should I open an Indian Restaurant?” and “Where should I stay If I want great Indian food?”

**Business Problem**

The objective of this capstone project is to analyse and select the best locations in the city to open a indian restaurant .

**Data**

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

Data will be collected from https://cocl.us/new\_york\_dataset and cleaned and processed into a dataframe.

• FourSquare be used to locate all venues and then filtered by Indian restaurants. Ratings, tips, and likes by users will be counted and added to the dataframe.

• Data will be sorted based on ranking

• Finally, the data be will be visually assessed using graphing from various Python libraries

**Sources of data and methods to extract them**

This link(https://cocl.us/new\_york\_dataset ) contains a list of neighbourhoods in newyork, The **neighborhoods in New York City** are located within the five [boroughs](https://en.wikipedia.org/wiki/Boroughs_of_New_York_City) of the [City of New York](https://en.wikipedia.org/wiki/New_York_City). Their names and borders are not officially defined, and they change from time to time . We will use web scraping techniques to extract the data from the link , with the help of Python requests . Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (link), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

**Methodology**

Firstly, we need to get the list of neighbourhoods in the city of new york. which is obtained from the link. We will do web scraping using Python requests to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Kuala Lumpur.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “indian restaurant” data, we will filter the “indian restaurant ” as venue category for the neighbourhoods.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 3 clusters based on their frequency of occurrence for “indian restaurant”. The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new shopping malls.

**Results**

Due to limited amount of data on indian restaurant k mean clustering is impossible

**Limitations and Suggestions for Future Research**

In this project, we only consider one factor i.e. frequency of occurrence of indian restaurant, there are other factors such as population and income of residents that could influence the location decision of a restaurant. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new shopping mall. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

**Conclusion**

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, and tried performing machine learning by clustering the data into 3 clusters based on their similarities.

**References**

Category: newyork data set Retrieved from

https://cocl.us/new\_york\_dataset

Foursquare Developers Documentation. *Foursquare*. Retrieved from <https://developer.foursquare.com/docs>