IBM DATA SCIENCE CAPSTONE PROJECT :

PROJECT : FINDING A SUITABLE NEIGHBORHOOD in HOUSTON to open an Indian Vegetarian Restaurant

INTRODUCTION:

A family friend recently moved to Houston,TX from Dallas.TX for a new job. He has decided to help his wife move her business of running an Indian Vegetarian Restaurant near downtown Dallas to a suitable location in Houston such that the business can be successful in terms of year round customer footfall and reach to a wider customer base.

Since I am a Data Scientist , my friend has approached me to help him find a suitable location in Houston to lease a commercial space for starting an Indian Vegetarian restaurant.

BUSINESS PROBLEM:

The objective of my efforts will be to derive Knowledge and Wisdom from Data and Information which is publicly available about the geography of Houston and the business establishments that are currently operating in Houston. The Main thrust will be to find information about various neighborhoods in Houston city and then Analyze them to effectively find out which of them are potential candidates to open an Indian Restaurant in Houston.

DATA:

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

•List of neighborhoods in Houston. Tx,USA. This defines the scope of this project which is confined to the city of Houston, the city in Texas,USA.

•Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.

•Venue data, particularly data related to Asian Restaurants. We will use this data to perform clustering on the neighborhoods.

SOURCES OF DATA AND METHODS TO EXTRACT THE DATA :

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_Houston_neighborhoods) contains a list of neighborhoods in Houston, with a total of 88 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods. After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases 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 (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Houston MAP :

This map represents all the 88 neighborhoods of the city Houston. We then try to make clusters by using the Machine learning technique(K-means clustering) and group all the neighborhoods into different clusters.