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 Houston 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.

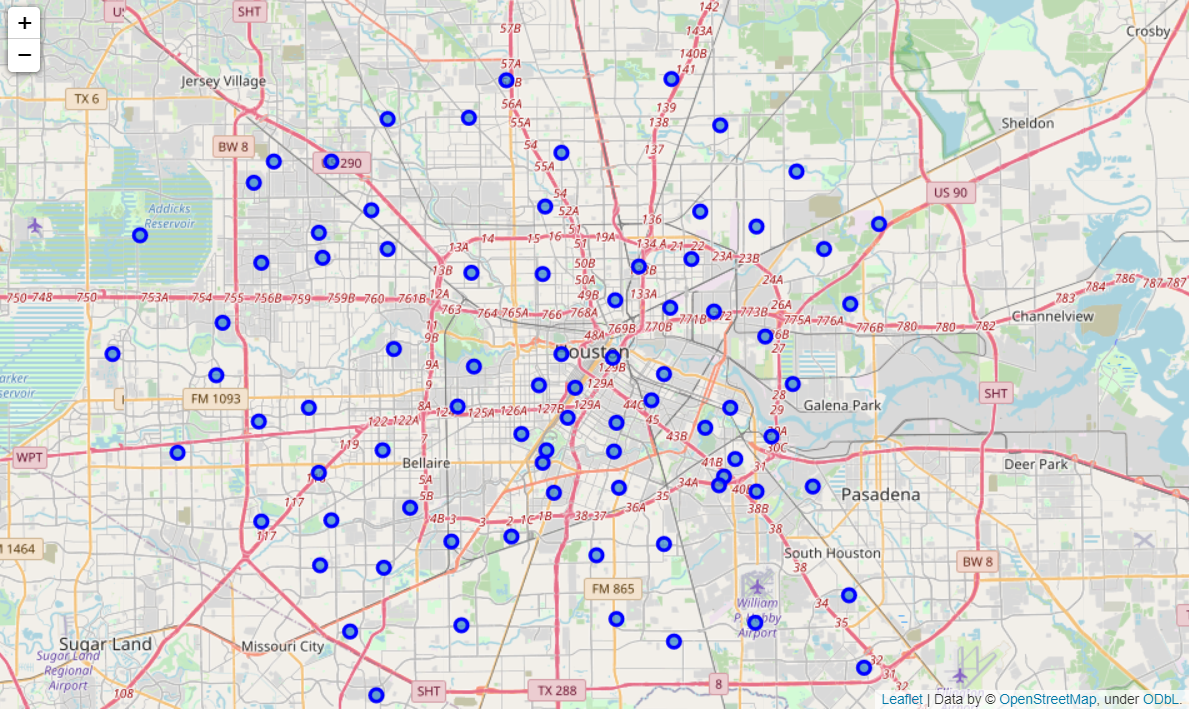
•Venue data, particularly data related to Asian Restaurants. We will use this data to perform clustering on the neighborhoods.

# 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.



# METHODOLOGY

Firstly, we need to get the list of neighborhoods in the city of Houston. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/List\_of\_Houston\_neighborhoods). We will do web scraping using Python requests and beautifulsoup packages to extract the list of Neighborhoods 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 the 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 Neighborhoods 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 Dallas. Next, we will use the Foursquare API to get the top 100 venues that are within a radius of 1000 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 Neighborhoods 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 Neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each Neighborhood by grouping the rows by Neighborhood 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 “Vegetarian/Vegan” data, we will filter the “Vegetarian/Vegan” as venue 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 Neighborhoods into 3 clusters based on their frequency of occurrence for “Vegetarian/Vegan”. The results will allow us to identify which Neighborhoods have a higher concentration of Vegetarian/Vegans while which Neighborhoods have a fewer number of Vegetarian/Vegans. Based on the occurrence of Vegetarian/Vegan Restaurants in different Neighborhoods, it will help us to answer the question as to which Neighborhoods are most suitable to open new s. Therefore, this project recommends property developers to capitalize on these findings to open new Vegetarian/Vegan Restaurants in Neighborhoods in cluster 2 with little to no competition.

# RESULTS

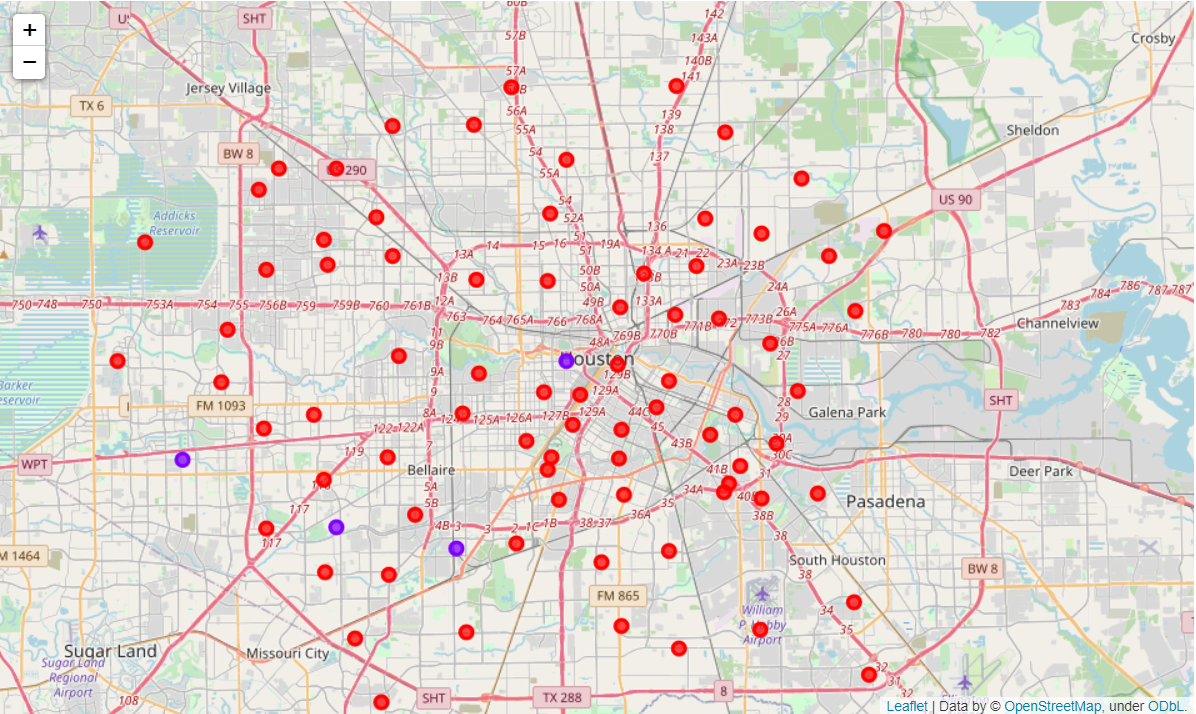
The results from the k-means clustering show that we can categorize the Neighborhoods into 3 clusters based on the frequency of occurrence for “Vegetarian/Vegan Restaurants”:

• Cluster 0: Neighborhoods with a very less number of Vegetarian/Vegan Restaurants

• Cluster 1: Neighborhoods with a moderate concentration of Vegetarian/Vegan Restaurants

• Cluster 2: Neighborhoods with a high concentration of Vegetarian/Vegan Restaurants (None returned in this dataset)

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color.



# DISCUSSIONS

As observations noted from the map in the Results section, the Vegetarian/Vegan Restaurants are very sparsely concentrated in the southwest area of Houston city, with the higher number in cluster 1and moderate number in cluster 0 . This represents a great opportunity and high potential areas to open new Vegetarian/Vegan Restaurants as there is very little to no competition from existing establishments. Meanwhile, Vegetarian/Vegan Restaurants in cluster 1 are likely suffering from more competition due to oversupply and high concentration of Vegetarian/Vegan Restaurants. From another perspective, the results also show that the oversupply of Vegetarian/Vegan Restaurants mostly happened in the suburb area of the city, with the central area still have very few Vegetarian/Vegan Restaurants. Therefore, this project recommends my friend to capitalize on these findings to open new Vegetarian/Vegan Restaurants in Neighborhoods in cluster 0 with little to no competition.

# LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

In this project, we only consider one factor i.e. frequency of occurrence of Vegetarian/Vegan Restaurants, there are other factors such as population and income of residents that could influence the location decision of a new shopping mall. However, to the best knowledge of this researcher, such data are not available to the Neighborhood 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

Property developers with unique selling propositions to stand out from the competition can also open new Vegetarian/Vegan Restaurants in Neighborhoods in cluster 1 with moderate competition. Lastly, property developers are advised to avoid Neighborhoods in cluster 2 which already have a high concentration of Vegetarian/Vegan Restaurants and suffering from intense competition.

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to my friend regarding the best locations to open a new Indian Vegetarian restaurant. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The Neighborhoods in cluster 0 are the most preferred locations to open a new Indian Vegetarian Restaurant.

# REFERENCES

•Category: Neighborhoods in Houston,Tx .Wikipedia. Retrieved from <https://en.wikipedia.org/wiki/List_of_Houston_neighborhoods>

•Foursquare Developers Documentation. Foursquare. Retrieved from https://developer.foursquare.com/docs/