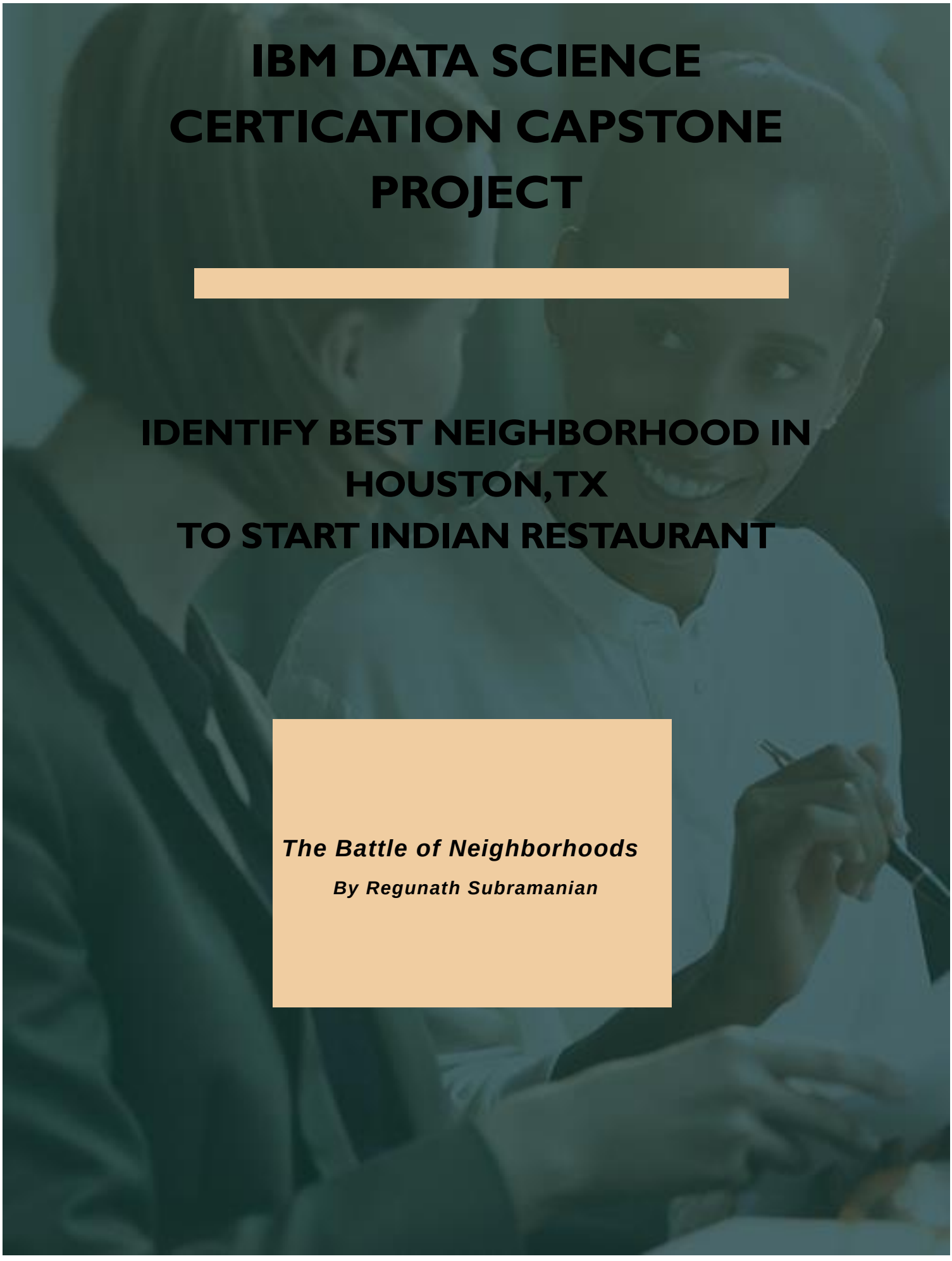


IBM DATA SCIENCE CERTICATION CAPSTONE PROJECT

IDENTIFY BEST NEIGHBORHOOD IN HOUSTON,TX TO START INDIAN RESTAURANT

The Battle of Neighborhoods

By Regunath Subramanian

TABLE OF CONTENTS

INTRODUCTION

Houston is the energy capital of the world; it's the headquarters and intellectual city for virtually all segment of the oil industry including technology, exploration, production, marketing, transmission, and supply. Houston is the fourth largest populous city in the United Status. One of my clients would like to start an Indian restaurant in Houston, TX and they would like to identify the optimal or best neighborhood. Therefore, this project will perform data analysis and try to find the most optimal neighborhood to open the Indian restaurant according to those criteria. It's obvious, that there are many additional factors, such as distance from parking places or distance from the main streets, but this analysis can be done after choosing the neighborhood, and thus will not be performed within the scope of this project. The insights derived from analysis will give good understanding of the business environment which help in strategically targeting the market. This will help in reduction of risk. And the Return on Investment will be reasonable.

1. BUSINESS PROBLEM

Due to oil prices, it is always full of highs, lows and learning for the restaurant industry in Houston. For the past few years, lower oil prices and increased competition have put significant pressure on Houston's more than 12,000 restaurants. Houston restaurant community suffered damages, including lost homes, cars and businesses due to recent hurricanes such as Hurricane Harvey. These unfortunate circumstances highlight the precarious nature of the restaurant business, where margins are slim, and most operators do not have significant sums saved for emergencies. Restaurants in Houston still face many challenges: Oil prices have not fully recovered, competition grows as more concepts open, rents are unreasonably high in many areas, and it's still a challenge to find enough qualified hospitality workers to keep restaurants fully staffed. In spite of these challenges and the ups and downs, Houston's restaurant and bar community will only get bigger and better in 2020.

Due to all these scenarios opening a new Indian restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the restaurant is one of the most important decisions that will determine whether the restaurant will be a success or a failure. The objective of this capstone project is to analyze the data and select the best locations in the city of Houston, Texas to open a new Indian restaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question to find an optimal neighborhood in the city of Huston, TX to open a new Indian restaurant.

Target Audience of this project

This project is particularly useful to restaurant entrepreneurs and investors looking to open or invest in restaurant in the city of Huston, TX.

2. DATA DESCRIPTION

The data below will be used to analyze this problem and make a recommendation to the client:

- List of neighborhoods in Huston, TX.
  - The below Wikipedia page would contain a list of neighborhoods in Huston, TX.
    - https://en.wikipedia.org/wiki/Category:Neighborhoods_in_Houston
  - Web scraping techniques is used to extract the data from the Wikipedia page, with the help of Python libraries and beautifulsoup packages.

- Latitude and longitude coordinates of those neighborhoods in order to plot the map and also to get the venue data.
  - Geographical coordinates will be retrieved for neighborhoods using Python Geocoder package, which will give us the latitude and longitude coordinates of the neighborhoods.

- Venue data, particularly data related to restaurants and it would be used to perform clustering on the neighborhoods.
  - Foursquare API will be used to get the venue data for these neighborhoods. Foursquare has one of the largest databases of 105+ million places and it had 50 million monthly active users . Foursquare API will provide many categories of the venue data, especially restaurant category.

- The below data science techniques and skills will be used to analyze and solve the business problem.
  - Web scraping Wikipedia Data
  - Using with Foursquare API
  - Data cleaning
  - Data wrangling
  - Machine learning techniques such as K-means clustering
  - Map visualization by using Folium
  - Prepare a report to present the solution or/and recommendation

3. METHODOLOGY

- The below two process are used to identify the best location to open new Indian restaurant.
  - Process one: Identify best neighborhood that has more Asian restaurants including Indian. This exercise will make sure, there is enough demand and marketplace available for Asian restaurants.
  - Process two: Identify the popularity among existing Asian restaurants as this process would assist us to mitigate the risk of competition. Based on number of likes the quality of the restaurant will be calculated.
- Below methodologies followed to analyze the data to solve the business problem:
  - Get the list of neighborhoods in the city of Houston from below Wikipedia page:
    - https://en.wikipdia.org/wiki/Category:Greater_Houston
  - Perform web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this data would not contain geographical data such as latitude and longitude.
  - By using Geocoder package, the system retrieves the latitude and longitude coordinates for each neighborhood.
  - Populate neighborhood data into DataFrame by using pandas DataFrame libraries.
  - Visualize the retrieved the data by using Folium package.
  - By using Foursquare API, the below neighborhood details will be retrieved via JSON file by using Python loop. Foursquare active developer account is required in order to get the Foursquare ID and secret key that are required to build Foursquare API URL.
    - Process one: Get top 100 venues details such as name, category, latitude and longitude within a radius of 2000 meters for each neighborhood.
    - Process two: Get venue details such as Name, ID, Location, Category, Count of Likes for the specific neighborhood that was identified from the above process one.
  - Analyze the unique venue data by using K-means clustering algorithm as it identifies k number of centroids and allocates every data point to the nearest cluster while keeping the centroids as small as possible. This unsupervised machine learning algorithms would be used to solve the problem for this project.
    - Process one: Identify the neighborhood that contains most Asian restaurants by filtering them based on their frequency of occurrence.
    - Process two: Group the data into quality categorical variable so we can cluster appropriately. In addition, we will create new categorical variables for the restaurants to better group them based on Asian cuisine.

4. RESULTS

- Process one results as follows:
  - Using K-mean to clustering data area with more number of Asian restaurants:

	Neighborhood	Asian Restaurant	Cluster Labels	Latitude	Longitude
0	► Baytown, Texas (1 C, 16 P)	0.000000	0	29.731875	-94.967078
1	► Bellaire, Texas (1 C, 7 P)	0.070000	0	29.696550	-95.575290
2	► Brazoria County, Texas (6 C, 16 P)	0.000000	0	29.179619	-95.410481
3	► Chambers County, Texas (7 C, 4 P)	0.000000	0	29.775887	-94.678311
4	► Clear Creek Independent School District (1...	0.000000	0	29.812750	-95.561750
5	► Conroe, Texas (1 C, 31 P)	0.010000	0	29.678734	-95.405491
6	► Fort Bend County, Texas (6 C, 11 P)	0.000000	0	29.538752	-95.448635
7	► Galveston Bay Area (8 C, 96 P)	0.000000	0	29.656876	-95.244964
8	► Galveston County, Texas (8 C, 15 P)	0.000000	0	29.507951	-95.093515
9	► Harris County, Texas (12 C, 53 P)	0.000000	0	29.658619	-95.228339
10	► Houston (25 C, 15 P, 3 F)	0.000000	0	29.760580	-95.369680
11	► Katy, Texas (2 C, 9 P)	0.000000	0	29.588706	-95.297429
12	► La Porte, Texas (8 P)	0.028571	0	29.713948	-95.278350
13	► Montgomery County, Texas (6 C, 7 P)	0.000000	0	30.300205	-95.503049
14	► Pasadena, Texas (3 C, 11 P)	0.000000	0	29.771910	-95.455933
15	► Pearland, Texas (3 C, 15 P)	0.000000	0	29.564527	-95.285006
16	► People from the Houston metropolitan area ...	0.000000	0	29.760580	-95.369680
17	► Rosenberg, Texas (1 C, 9 P)	0.000000	0	29.562258	-95.810106
18	► Santa Fe, Texas (8 P)	0.000000	0	29.661902	-95.301964
19	► Sports in the Houston metropolitan area (2...	0.000000	0	39.049830	-84.895540
20	► Sugar Land, Texas (5 C, 15 P)	0.020000	0	29.596949	-95.620896

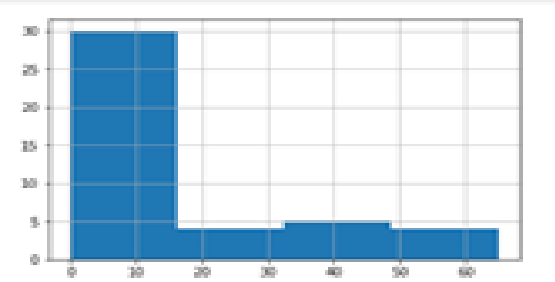
  - Using folium library to create map with markers to establish a dataset and examine the accuracy of coordinates.

  - Based on above data, Bellaire, TX neighborhood has a greater number of Asian restaurants so we to drilldown in this area and found out our competitors' strength and weakness to support our analysis.

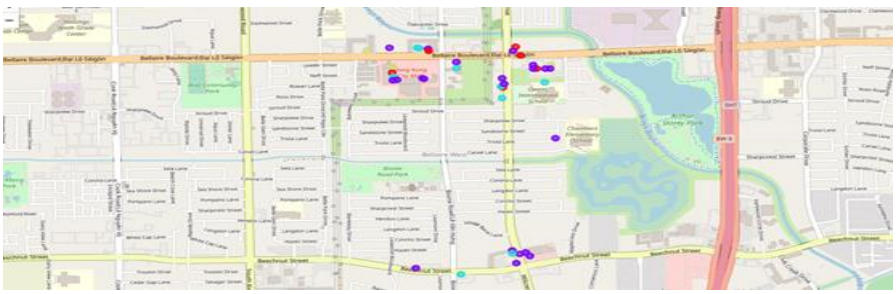
- Process two results as follows:
  - Using K-mean to clustering algorithm, generate clusters of restaurants for Bellaire, TX neighborhood with the characteristics of quality of food.

name	id	categories	lat	lng	total likes	total likes_cat	categories_new
Crawfish & Noodles	4bae9302f964a52051c23be3	Vietnamese Restaurant	29.70423	-95.5802	59	avg avg	Asian Food
Tony Thai Restaurant	4b2dab2ef964a52034da24e3	Thai Restaurant	29.70292	-95.56938	34	below avg	Asian Food
Com Tam Kieu Giang	4c66c9f9e75ac928aa77f8da	Vietnamese Restaurant	29.7023	-95.57845	4	poor	Asian Food
Pho Danh 2	52a4bf1511d278ee52857b20	Vietnamese Restaurant	29.70234	-95.57849	8	poor	Asian Food
Kim Son	4b8077a1f964a520077530e3	Vietnamese Restaurant	29.70317	-95.5685	65	avg avg	Asian Food
Pho Danh	4b916eadf964a5200ebc33e3	Vietnamese Restaurant	29.7023	-95.57849	11	poor	Asian Food
Nam Giao	4b744122f964a52035d02de3	Vietnamese Restaurant	29.70179	-95.5715	7	poor	Asian Food
Nguyễn-ngộ	4bc0e1f4461576b010f97a32	Vietnamese Restaurant	29.70441	-95.57685	20	poor	Asian Food
Thien Thanh	4b6daa97f964a52093842ce3	Vietnamese Restaurant	29.70446	-95.57697	17	poor	Asian Food
Pho Duy	51a3aa7e498e80da50870908	Vietnamese Restaurant	29.70076	-95.57158	9	poor	Asian Food
Lee's Sandwiches	4b6db576f964a52072882ce3	Vietnamese Restaurant	29.70424	-95.57623	40	below avg	Asian Food
Si Mê Restaurant	4d9ac3e73f785481881fd1d1	Vietnamese Restaurant	29.70163	-95.57166	0	poor	Asian Food
Bo Ne Houston	4bb6480af562ef3b0ce92f97	Vietnamese Restaurant	29.68782	-95.57418	9	poor	Asian Food
Xuan Huong Restaurant	590c28080f013c66ce1aa1a6	Vietnamese Restaurant	29.70296	-95.57448	3	poor	Asian Food
Mai's Baguettes	5bce720add8442002c7e4bf4	Vietnamese Restaurant	29.70191	-95.56886	0	poor	Asian Food
Pho Nguyen	4c3b5a7616cb2d7fd7ba02a9	Vietnamese Restaurant	29.68935	-95.57092	2	poor	Asian Food

  - The below bar chart visualizes our total likes based on a histogram.



  - Using folium library to create map Asian restaurant location in Bellaire, TX neighborhood:



5. DISCUSSION

As observations noted from the map in the results section, most of the Asian restaurants are concentrated in the Bellaire, TX neighborhood and this represents a great opportunity and marketplace to open new Indian restaurant. However, there would be an intensive competition from other Asian restaurants due to large numbers of Asian restaurants. This analysis is performed on limited data. This may be right or may be wrong. But if good amount of data is available there is scope to come up with better results. It can be done more detailed analysis by adding other factors such as transportation, demographics of inhabitants.

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 Indian restaurant. Foursquare proved to be a good source of data but frustrating at times and 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.

6. CONCULSION

In this project, we consider below two factors:

1. Number of Asian restaurants in each Great Houston neighborhoods.
2. Popularity of the Asian restaurants in the high concentrated area to check the competition.

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 based on their similarities, and lastly providing recommendations to the relevant stakeholders and investors regarding the best location to open a new Indian restaurant.

To answer the business question that was raised in the introduction section, the answer proposed by this project is: Bellaire, TX neighborhood is the most preferred location to open a new Indian restaurant. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Indian restaurant.