**Project Capstome**

**Segmenting and Clustering Neighborhoods in Houston**

Une image contenant extérieur, plusieurs

Description générée automatiquement

Table des matières

[Table of Contents 2](#_Toc62635968)

[1. Introduction 2](#_Toc62635969)

[2. Target Audience 2](#_Toc62635970)

[3. Data Overview Methodology Discussion Conclusion 2](#_Toc62635971)

[1.Introduction 2](#_Toc62635972)

[2.Target Audience 3](#_Toc62635973)

[3.Data Overview 3](#_Toc62635974)

[3.1 — Data acquisition: 3](#_Toc62635975)

[4. Methodology 5](#_Toc62635976)

[4.1 — Data Cleansing 5](#_Toc62635977)

[4.2 — Data Exploration 5](#_Toc62635978)

[4.3 — Machine Learning 7](#_Toc62635979)

[Then we will sort the venues for each neighborhood according to the frequency 7](#_Toc62635980)

[5.K-Means Clustering 7](#_Toc62635981)

[6.Conclusion 9](#_Toc62635982)

This project aims to utilize all Data Science Concepts learned in the IBM Data Science Professional Course. We define a Business Problem, the data that will be utilized and using that data, we are able to analyze it using Machine Learning tools. In this project, we will go through all the processes in a step-by-step manner from problem designing, data preparation to final analysis and finally will provide a conclusion that can be leveraged by the business stakeholders to make their decisions.

# **1.Introduction**

Houston is the most populous city in the U.S. state of Texas, fourth-most populous city in the United States, most populous city in the Southern United States, as well as the sixth-most populous in North America, with an estimated 2019 population of 2,320,268.Located in Southeast Texas near Galveston Bay and the Gulf of Mexico, it is the seat of Harris County and the principal city of the Greater Houston metropolitan area, which is the fifth-most populous metropolitan statistical area in the United States and the second-most populous in Texas after the Dallas–Fort Worth metroplex, with a population of 7,066,141 in 2019. Houston is the southeast anchor of the greater megaregion known as the Texas Triangle.

The objective of this project is to use Foursquare location data and regional clustering of venue information to determine which neighborhoods are 'similar' in term of venues. Through this project, we will cluster the neighborhoods in Houston to get this information

# **2.Target Audience**

This project is aimed towards real estate agents or company or household who desire to move to another neighborhood. The analysis will provide vital information that can be used by the target audience about the neighborhood.

# 

# **3.Data Overview**

# **3.1 — Data acquisition:**

Une image contenant texte

Description générée automatiquement

Figure 1 :https://en.wikipedia.org/wiki/List\_of\_Houston\_neighborhoods

The Wikepedia site shown above provided almost all the information about the neighborhoods. It included Coordonate; borough and the name of the neighborhoods present in Miami. Since the data is not in a format that is suitable for analysis, scraping of the data was done from this site (shown in figure2).

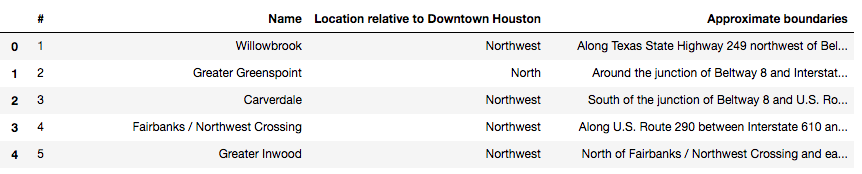


Figure 1 :data after webscapring

Source 2: Geographical Location data using Geocoder Package

With some cleaning an processing on the Neighborhood name and an for loop we cane extract the coordinate of the neighborhoods

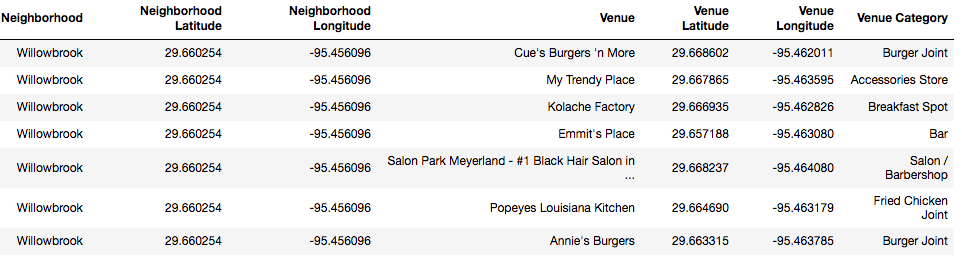
Une image contenant texte

Description générée automatiquement

Une image contenant table

Description générée automatiquement

Source 3: Venue Data using Foursquare



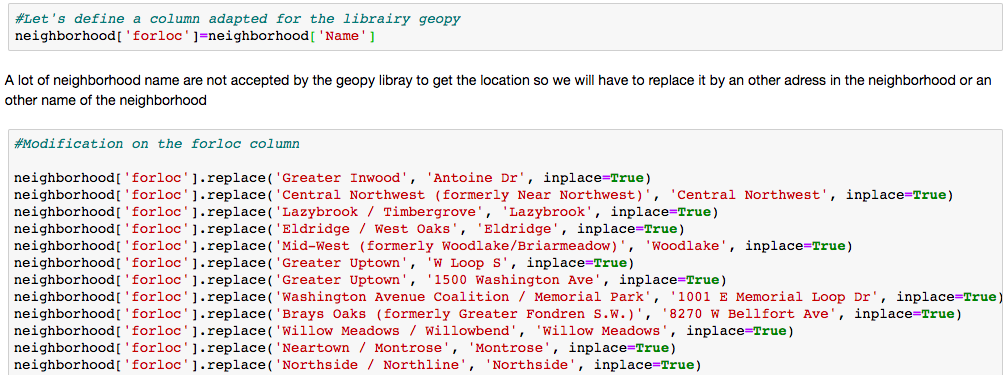
We performed a bit of data cleaning. It is seen through figure 5 (above) that the neighborhoods are grouped by the name of the neighborhood, so data clustering is made easier later on.

# 

# **4. Methodology**

# **4.1 — Data Cleansing**

After all the data was collected and put into data frames, cleansing and merging of the data was required to start the process of analysis. When getting the data from Wikipedia, there were without coordinate that were not assigned to any neighborhood.



Une image contenant texte

Description générée automatiquement

Retrieve the foursquare data and clean the venue

Une image contenant texte

Description générée automatiquement

# 

# **4.2 — Data Exploration**

Now after cleansing the data, the next step was to analyze it. We then created a map using Folium where each Neighborhood was located in.

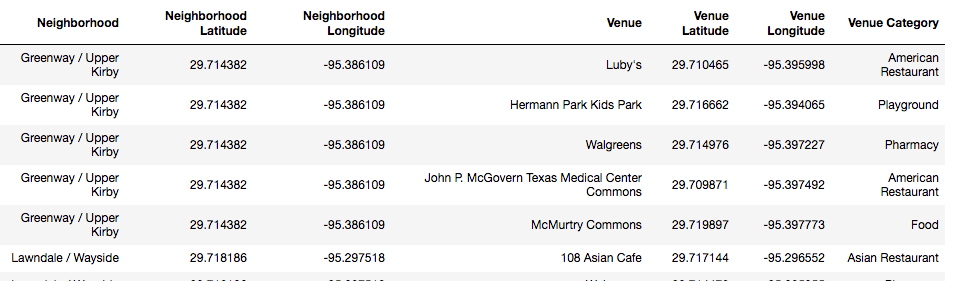
Une image contenant carte

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

Next, we used the Foursquare API to get a list of all the Venues in Houston which included Parks, Schools, Café Shops, Asian Restaurants etc. Getting this data was crucial to analyzing the number similarity between neighborhoods. We then merged the Foursquare Venue data with the Neighborhood data which then gave us the nearest Venue for each of the Neighborhoods.



# **4.3 — Machine Learning**

Then to analyze the data we performed a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This technique is called **One hot encoding**. For each of the neighbourhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighbourhood.

Une image contenant table

Description générée automatiquement

Then we grouped those rows by Neighborhood and by taking the average of the frequency of occurrence of each Venue Category.

Une image contenant table

Description générée automatiquement

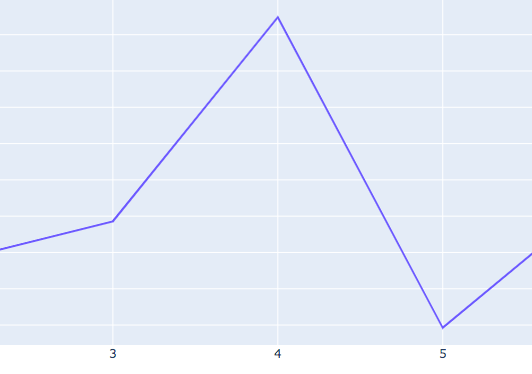
## Then we will sort the venues for each neighborhood according to the frequency

Une image contenant table

Description générée automatiquement

# **5.K-Means Clustering**

To make the analysis more interesting, we wanted to cluster the neighbourhoods based on the neighbourhoods that had similar venues. To do this we used **K-Means**clustering. To get our optimum K value that was neither overfitting or underfitting the model, we used the **Silhouette score** Technique. In this technique, we ran a test with different number of K values and measured the average silhouette score and then chose the best K value. The best K value is chosen at the point in which the line has the sharpest turn. In our case, we had the the ottmum at K = 4. That means we will have a total of 4 clusters.



This technique of optimization is not so perfect specifically because all my data are not continuous, but we will keep it to get our cluster.

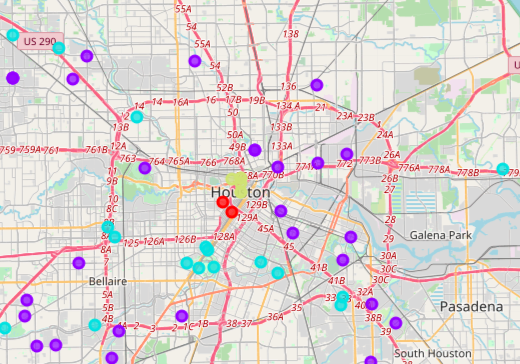
Une image contenant table

Description générée automatiquement

Let’s visualize our cluster on the map and color each color

Une image contenant texte

Description générée automatiquement



# **4.4 — Data Analysis**

Cluster 0:

In this neighborhood we can find a lot of coffee bar and Mexican restaurant

****

**Cluster 1:**

Lot of parks, Mexican restaurants and fast food and construction and landscaping store

Une image contenant herbe, arbre, ciel, extérieur

Description générée automatiquement

Braseswood neighborhood, Houston

Cluster 2:

Clothing stores, zoo and Italian restaurant

Cluster 3:

A lot of hotel, coffee shop, park, bar, typical restaurant soul food restaurant. It seems to be the good place for visiting the south and to eat typical food and enjoy the large space of the South of the US

Une image contenant extérieur, arbre, ciel, bâtiment

Description générée automatiquement