A GUIDE TO CHOOSE YOUR FAVORITE NEIGHBORHOOD IN HOUSTON (TEXAS)!

IBM DATA SCIENCE CAPSTONE PROJECT

OCTOBER 2020

PROBLEM STATEMENT

- Moving from one place to another is very stressful specifically when you do not have en
- Finding the right location to live is not an easy task since there are many parameters which affect the decision to choose the right neighborhood.
- Accessing a tool that allows you to compare different neighborhoods in a large city, can provide valuable insights regarding each area and ultimately helps you in your decision-making process.

BACKGROUND & OBJECTIVE

- The Greater Houston metropolitan area is the fifth most populous metropolitan area in the US. The diverse industry and affordable housing make this city more attractive to the people who would like to relocate to a place which offers the amenities of a large city with lower cost.
- The objective of this work is to analyze different suburbs around Houston (Texas) and provide some insight regarding each neighborhood.

DATA INTEGRATION

- The data are extracted from different sources:
  1. geographical data related to Houston suburbs. This data was extracted from the "US zip code latitude and longitude" web page (https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/).
  2. List of 26 suburb cities was manually extracted from various pages and, stored in a csv file
  3. Foursquare location data used to find various information regarding each suburb
- The first and second set of data were joined to integrate the zip codes, suburb cities and location information. This generates the initial dataframe.

DATA EVALUATION AND CLEANING

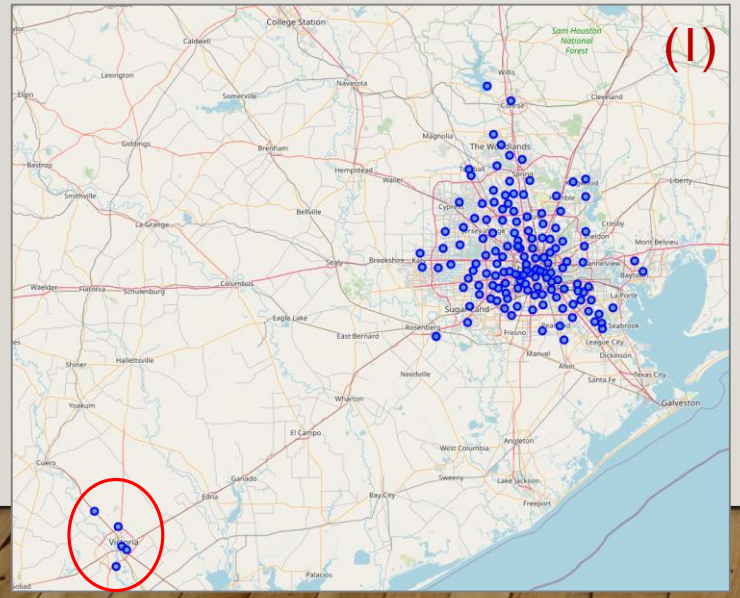
- The unnecessary columns are dropped and the resulting dataframe (DF) is examined to make sure there is not NULL values.
- Many of the cities only include 1 or 2 data points, therefore the DF was subset for only the cities with more than 3 data points.
- The resulted DF, includes 7 cities and 225 zip codes.

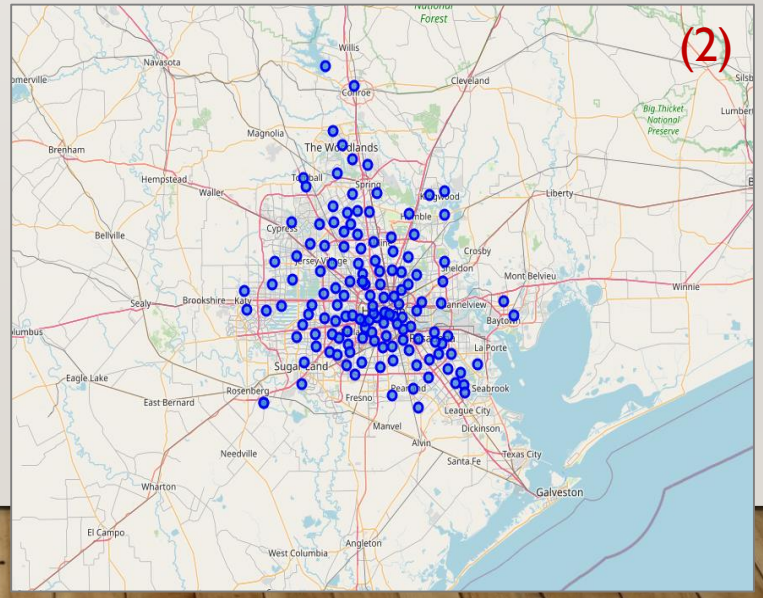
	Zip	City	Latitude	Longitude
0	77373	Spring	30.056394	-95.389610
1	77005	Houston	29.717529	-95.428210
2	77735	Tomball	30.095391	-95.628023
3	77088	Houston	29.879213	-95.450280
4	77066	Houston	29.959439	-95.496940
...	...	...	...	...
245	77008	Houston	29.798777	-95.409510
246	77238	Houston	29.833990	-95.434241
247	77086	Houston	29.920981	-95.495560
248	77248	Houston	29.833990	-95.434241
249	77026	Houston	29.794370	-95.333950

250 rows × 4 columns

DATA EVALUATION AND CLEANING

- The location data are demonstrated on the map.
  - Map (1) shows that there is a suburb which is far away from the Houston area and therefore it was removed from the dataset
  - Map (2) shows the location of the data in our final DF.





EXPLORING HOUSTON & SUBURBS

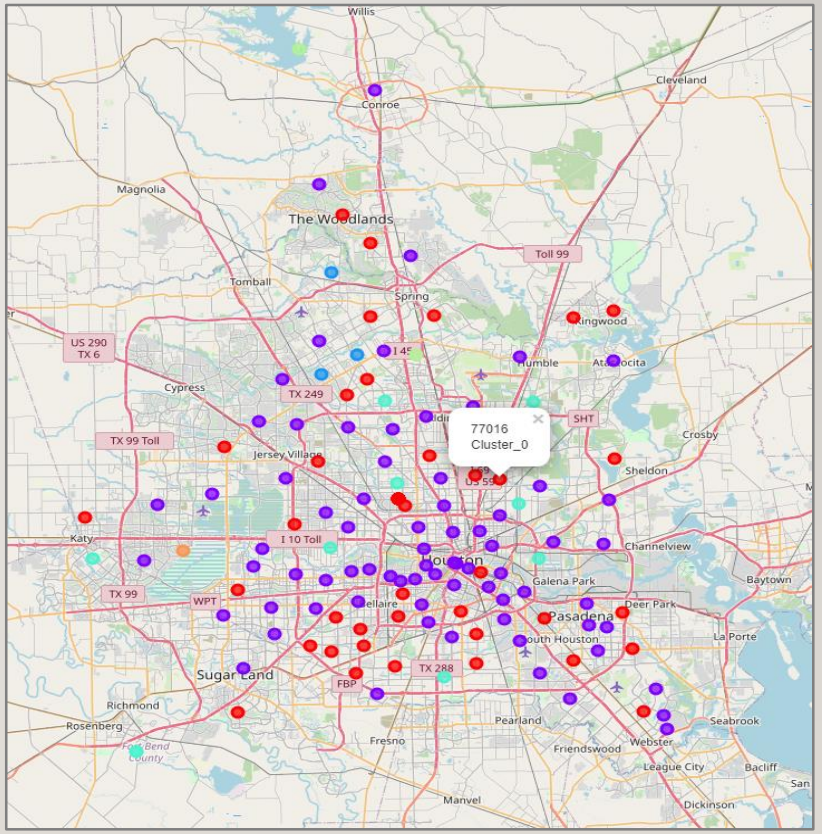
- Foursquare application is used to extract the venues and category of the venues for all available data points (at the radius of 1000 meter)
- The resulted data frame includes ***7186*** data records from Houston and suburb
- In the next step the unique venue categories are set as the input features.
- Then, by taking the mean of the frequency of each venue category, the data will be grouped for each single zip code. This decreases the data points to ***217***.

City	
Houston	6290
Humble	252
Katy	164
Pasadena	211
Spring	247
Sugar Land	22

	Zip_Code	Latitude	Longitude	ATM	Accessories Store	Advertising Agency	Airport	Airport Lounge	Airport Service	Airport Terminal	Antique Shop	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Arts & Entertainment	Astrologer
0	77000	29.711257	-95.304936	0.037037	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.000000	0.0	0.0
1	77001	29.813142	-95.309789	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.000000	0.0	0.0
2	77002	29.755578	-95.365310	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.000000	0.0	0.0
3	77003	29.749278	-95.347410	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.000000	0.0	0.0
4	77004	29.728779	-95.365700	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.05	0.0	0.000000	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
212	77503	29.695028	-95.157980	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.000000	0.0	0.0
213	77504	29.648780	-95.188130	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.040000	0.0	0.0
214	77505	29.650492	-95.146320	0.000000	0.016949	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.016949	0.0	0.0
215	77506	29.705678	-95.202160	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.000000	0.0	0.0
216	77508	29.569927	-95.106637	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00	0.0	0.000000	0.0	0.0

CLUSTERING

- The 7 most frequent venues for each zip code is considered as the feature inputs and based on these features we will evaluate the similarities of different suburbs, through clustering.
- *K-means* is used as an unsupervised clustering method to partition the data into ***6 clusters***
- Different colors on the map represents the clusters



DISCUSSION
CLUSTER EVALUATION

- Cluster-1
  - Is the largest cluster and includes 125 data point.
  - This cluster is distributed evenly within the Houston and suburb cities. The red dots on the map represents the data points in this cluster.
  - The most frequent categories among all the data records in the first three columns are: *Spa*, *Bar* and *Restaurant* respectively.
- Cluster-2
  - Has 75 data points which are shown as purple dots on the map
  - Evenly distributed within Houston and other cities.
  - The most frequent categories of the first three columns in this cluster are: *Restaurants, Sandwich/pizza place and bar*.

DISCUSSION
CLUSTER EVALUATION

- Cluster-3
  - Includes only 3 data points, which are located towards the North East, and far from the Houston downtown (shown with blue dots on the map).
  - The most dominant categories in this cluster are related to sport and recreation such as Golf course, pool, lake and Yoga studios.
- Cluster-4
  - Has10 data points and is shown with turquoise dots on the map.
  - The data points are distributed in some zip codes within Houston as well as other suburbs in all directions.
  - The most frequent venue categories for all data set are: *Construction & Landscaping, park, Fried Chicken and burrito place*.

DISCUSSION
CLUSTER EVALUATION

- Cluster-5
  - Only 3 data points fall into this cluster, which is shown with light green color on the map.
  - The data points are located towards the north of Houston. The most frequent categories in this cluster includes: *Home Services, Boat or Ferry, park and Yoga Studio*.
- Cluster-6
  - This cluster includes only one single data point, which is located in the western boundary of Houston and is shown with Orange dot on the map.
  - The most common categories for this zip code are: *Park, Yoga Studio, Disc Golf, Field and Farmers market*.

CONCLUSION

- The distribution of the clusters shows that within almost all the suburbs we still can find good number of similar venues such as restaurant, bars, sport facilities, etc. Therefore, any Houston suburb provides many options for these categories.
- For more accurate analysis, more data points are required especially from the suburb area.
- Finally, this analysis will be more comprehensive if other data such as crime rate, school ranking, house prices, etc. for each individual zip code (neighborhood) can be integrated to the current data set.