IBM Data Science Capstone Project Report
Battle of Neighborhoods-Houston

Mohab Dessouki

Problem and Background

- Houston’s fast population growth, accompanied by urban expansion, create new business opportunities in order to meet the fast-growing demand for services, products, new jobs, and infrastructure.
- However, there are some factors that needs to be considered about the business environment such as the business type, location, competition, target market, education and skills.
- The information and data science tools used in this report can be used by Investors who are looking for an overview about the city’s market environment, the tools can be adjusted to accommodate different business types and objectives.
- An example of a business venue (Starbucks) is presented for illustration.

Data Description

Neighborhoods Demographics

- The city of Houston has 147 neighborhoods. The demographic data of those neighborhoods including population, median income, median age, education, unemployment are collected from (https://www.houstoniamag.com) that originally comes from the 2010 U.S. Census and the 2015 American Community Survey. This data is used for statistical analysis and to study the trends in the neighborhood demographics data.

Neighborhoods Geographic Coordinates

- The neighborhoods geographic coordinates were obtained using “Bing maps” geocoding. The (Latitude, Longitude) information was used to visualize the neighborhood population distribution in Houston.

Top 100 Neighborhood Venues

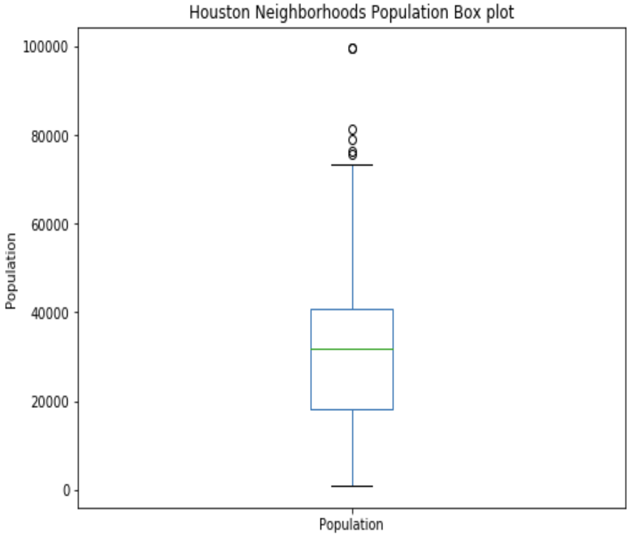
- The social networking service “Foursquare” API was used to search for the top 100 venues in each neighborhood. The information collected includes the venues’ name, latitude, longitude, and venue category.

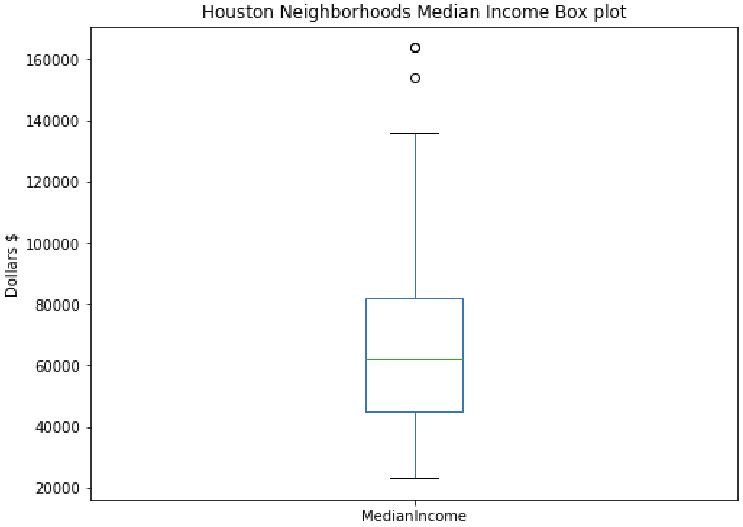
PYTHON LIBRARIES INSTALLATION

- BeautifulSoup: This package is used for parsing HTML documents. It is useful for extracting data from websites (web scraping) (https://www.crummy.com/software/BeautifulSoup/bs4/doc/, 2020).
- Geocoder: Deals with multiple different geocoding service providers such as Google, Bing, OSM (https://geocoder.readthedocs.io/, 2020).
- Numpy: An N-dimensional array object. Handles data in a vectorized manner.
- Pandas: It offers data structures and operations for manipulating numerical tables (DataFrames).
- Requests: Allows sending HTTP requests to a website server in an easy way (https://requests.readthedocs.io/en/master/, 2020).
- Matplotlib: Used for creating plots in python (https://matplotlib.org/, 2020)
- Scikit-learn: It features various classification, regression, and clustering algorithms including Kmeans (https://scikit-learn.org/stable/, 2020).
- Folium: It enables both the binding of data to a map for visualization as well as passing HTML visualization as markers on the map.

Neighborhood Demographic Data Boxplots

- **Minimum Score** is the lowest value in a dataset.
- Twenty-five percent of scores fall below the **Lower Quartile** value.
- The **Median** marks the mid-point of the dataset.
- Seventy-five percent of scores fall below the **Upper Quartile** value.
- **Maximum Score** is the highest value in a dataset.
- The **Interquartile Range** is the range between the Lower and upper Quartiles.

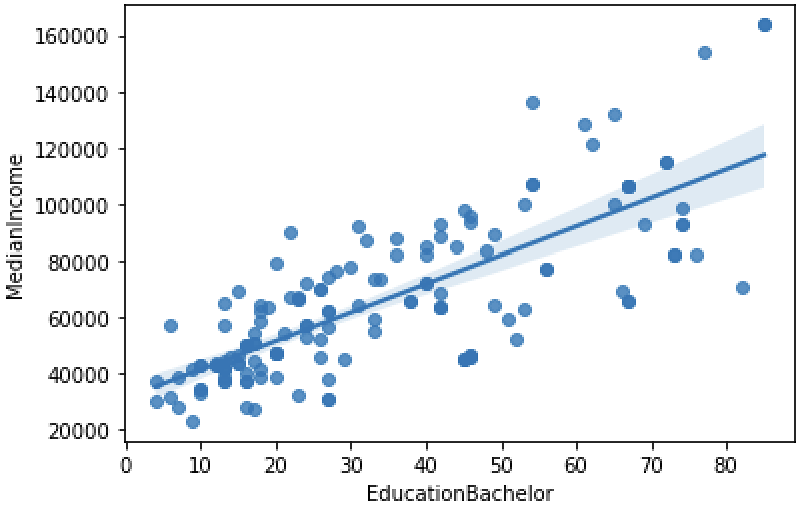


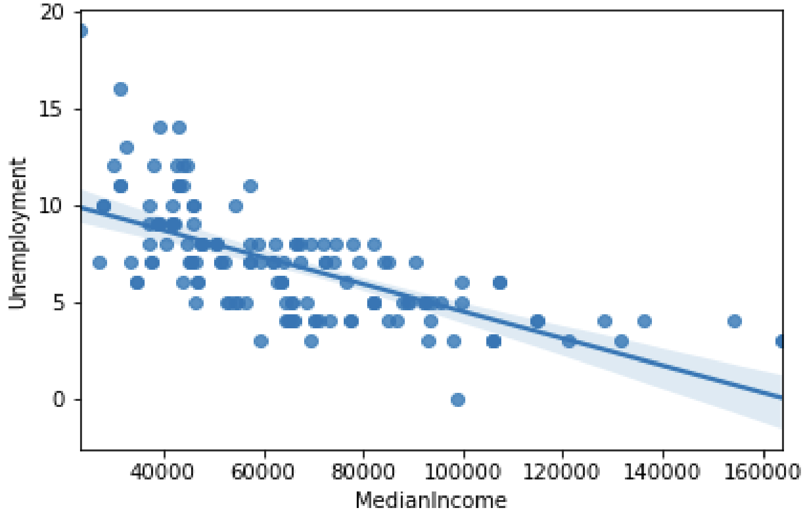


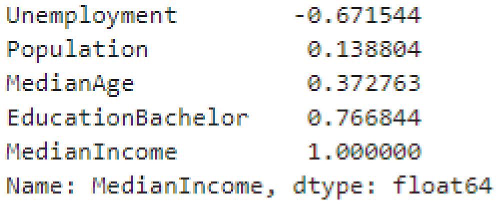
- 75% of the neighborhoods have a population of more than 18000.
- 75% of the neighborhoods have a median income higher than 45000$

Neighborhood Demographic Regression Analysis

- Income is one of the factors that indicates whether a neighborhood could be attractive for investment.
- Regression analysis between Income and other demographic variables was performed.





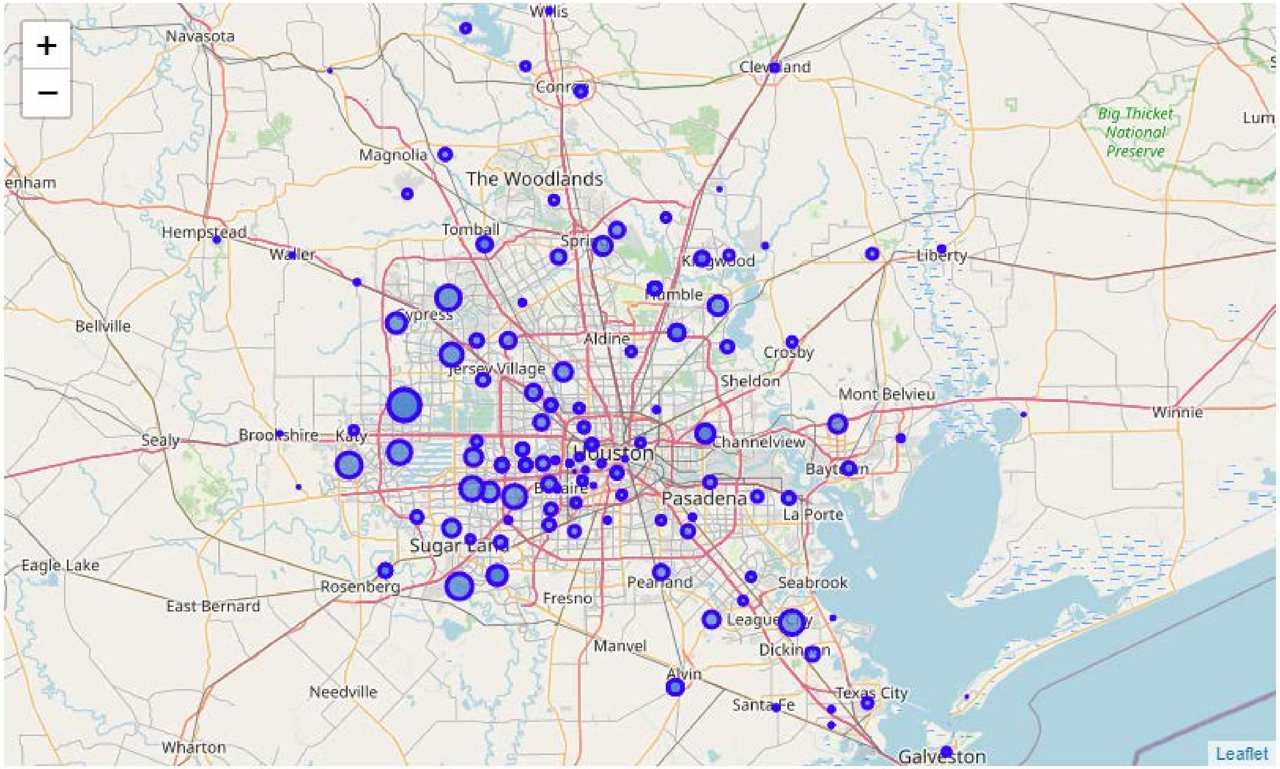


```
Unemployment        -0.671544
Population           0.138804
MedianAge            0.372763
EducationBachelor    0.766844
MedianIncome         1.000000
Name: MedianIncome, dtype: float64
```

- Income increases with education.
- unemployment decrease with increasing education

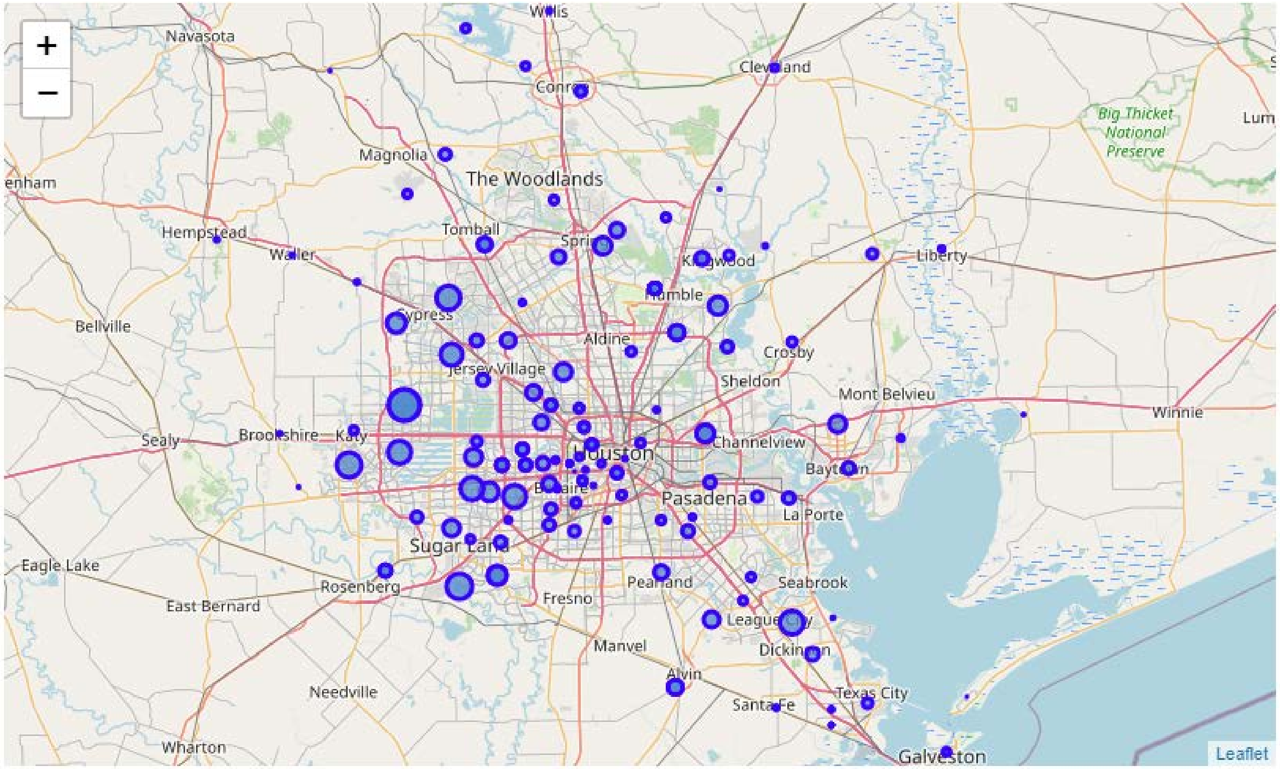
THE CITY OF HOUSTON (Population)

- The folium library can be used to visualize the neighborhoods locations on the Houston map.
- The demographic data can also be described on the map.
- It could be useful to know the population distribution as well as income on the map.
- The marker size represents the relative size of neighborhood population.



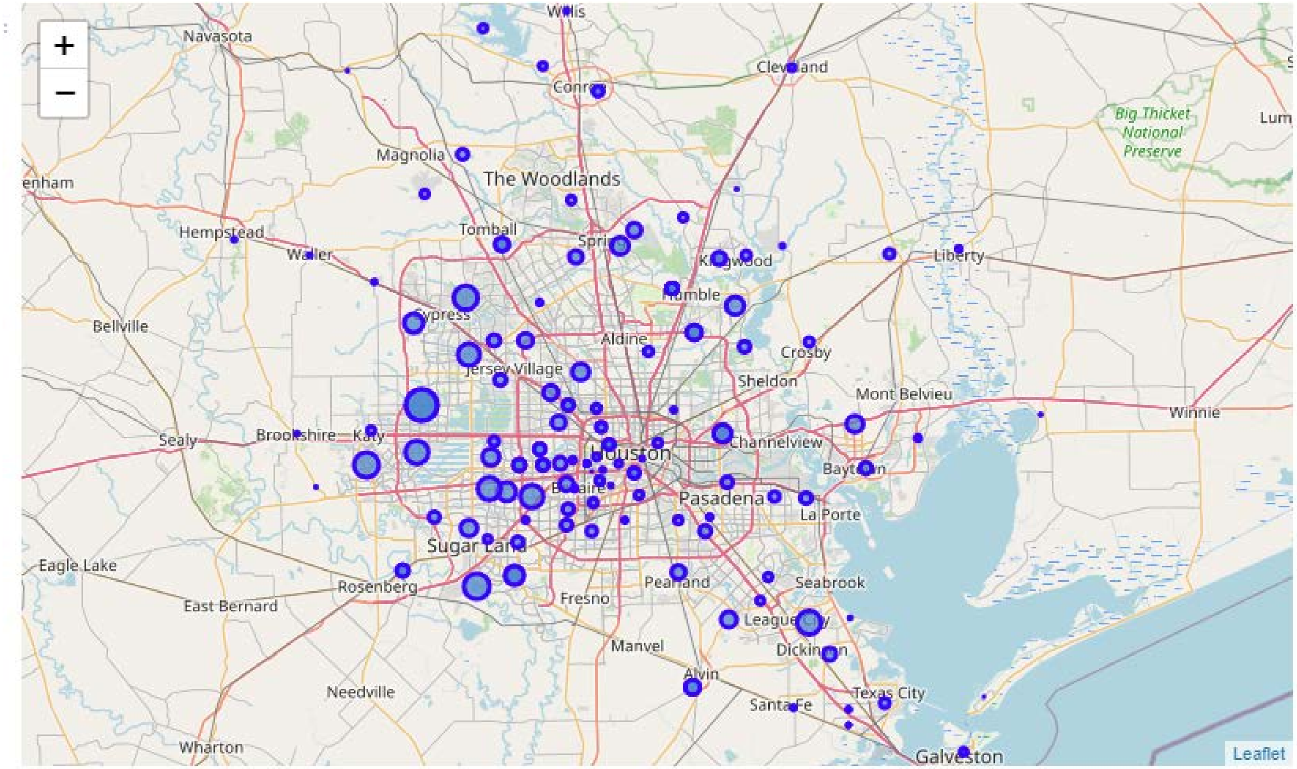
THE CITY OF HOUSTON (Income)

- The marker size represents the relative size of neighborhood median household income.



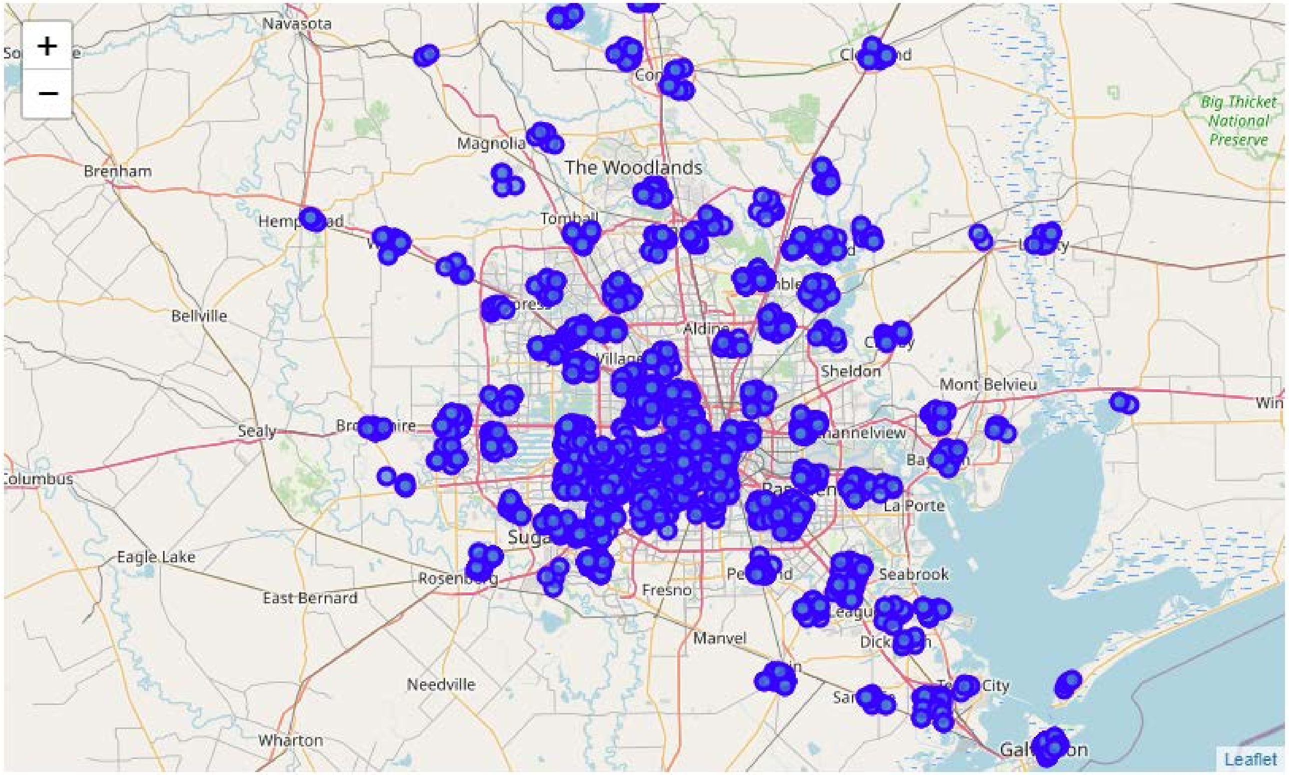
THE CITY OF HOUSTON (Population*Income)

- The marker size represents the magnitude of the Population and income product.



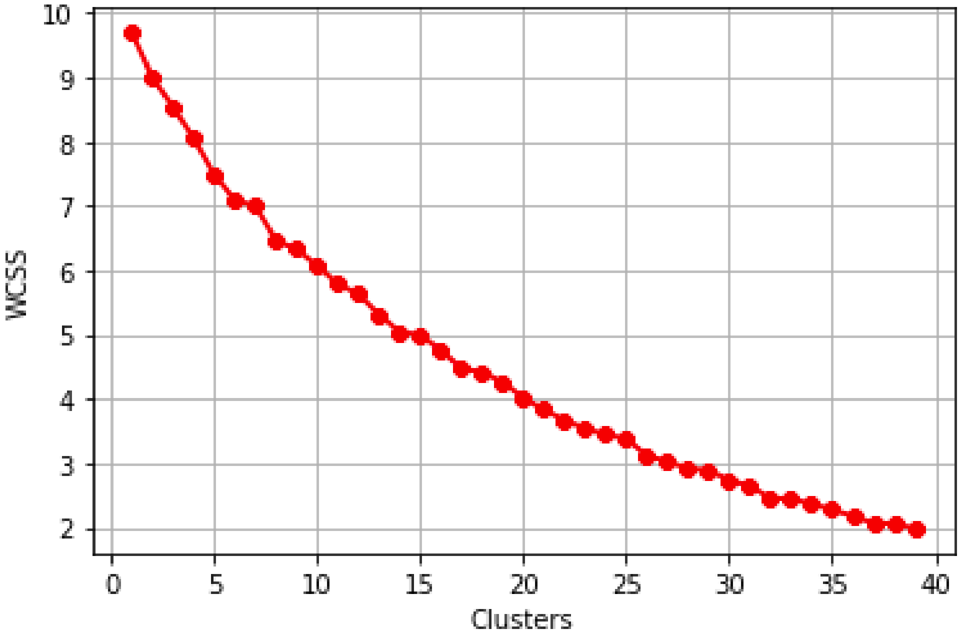
UTILIZING FOURSQUARE API TO EXPLORE NEIGHBORHOOD VENUES

- Foursquare is a social networking service provider, it is used to get information about the location and properties of top businesses and attractions in each neighborhood.
- The neighborhood coordinates were used to explore the top 100 venues within a range of 2000 meters.
- The data includes the venue name, latitude, longitude, and category. A total of 7338 venues and 350 unique business types were found within the search area.
- The geographic distribution of the venues is shown



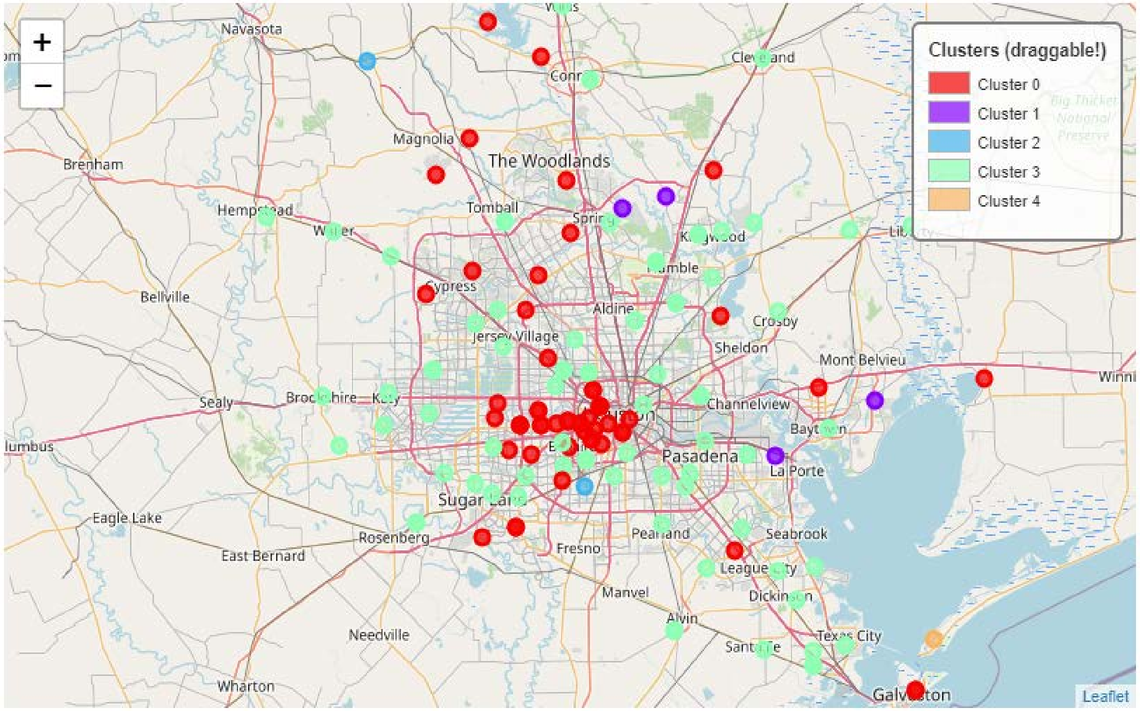
Clustering and Segmenting Neighborhoods (K-means)

- The K-means is an iterative clustering method that aims to partition the dataset to defined distinct clusters.
- It is an unsupervised algorithm that search for similarities within a dataset and creates clusters of datapoints that share similar attributes.
- The optimum number of clusters can be determined from the relationship between the WCSS (sum of squares of the distances of each data point in all clusters to their respective centroids).
- The WCSS-Clusters relationship is shown



Clustering and Segmenting Neighborhoods (K-means)

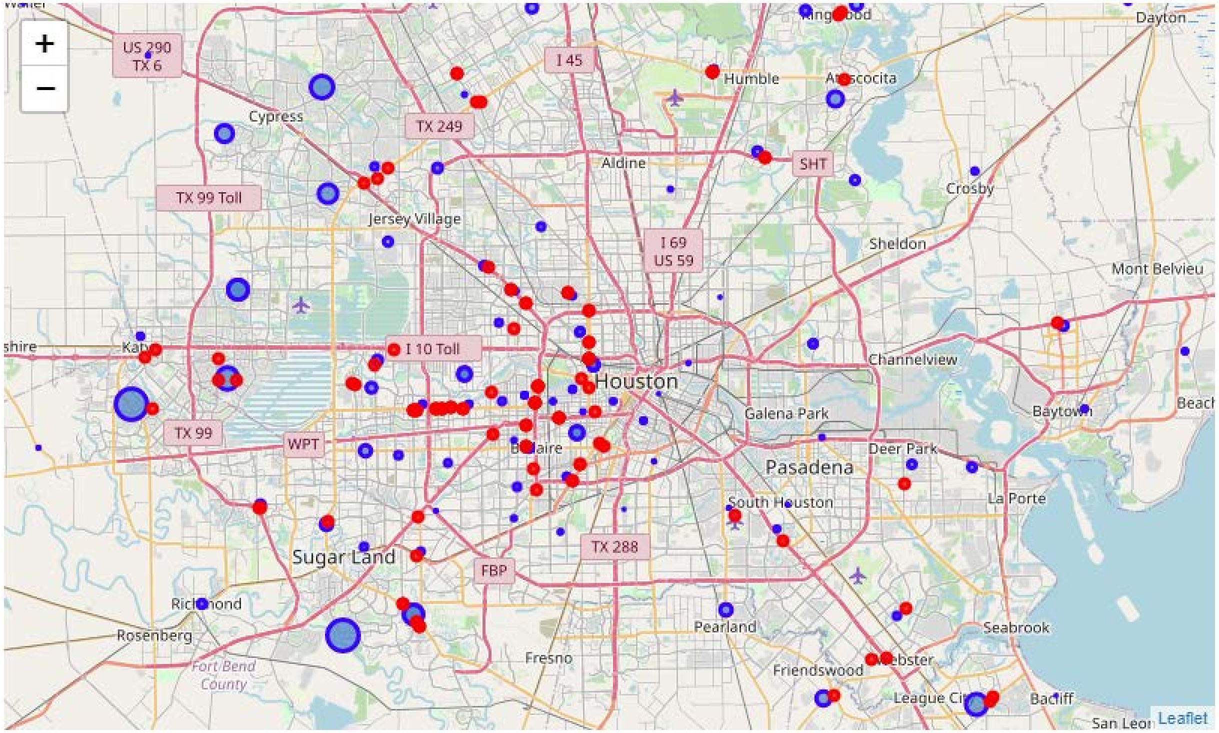
- Although the optimum number of clusters seems to be around 40 clusters.
- However, we will only use 5 clusters for simplicity.
- Using too many clusters could lead to clusters with only one neighborhood.
- The color-coded neighborhoods clusters is shown



Clusters	Most common venue Category
0	Mexican/American Restaurants
1	Construction/Lanscaping/Home services
2	Gas Stations
3	Fast Food and International Restaurants
4	Harbor/Marina

EXPLORE NEW LOCATIONS THAT COULD BE ATTRACTIVE FOR A NEW VENUE (Starbucks)

- The neighborhoods locations are represented in blue markers, and the size of the marker represent the magnitude of the population and income (the multiplication product of the Income and population of a neighborhood).
- The red markers represent the locations of existing Starbucks locations. According to this map, we can see that there are investment opportunities in the wealthy and populated areas represented by large markers that don't have a close by Starbucks venue.



Neighborhood	Population	Median Income
Sugar Land South	81466	128362
Katy-Southwest	79117	131694
Cypress North	76437	98191
Cypress South	60324	93884
Katy-North	99450	66550
Copperfield Area	69927	88555

According to this map, the neighborhoods that could be attractive for investments are {Sugar Land South, Katy-Southwest, Cypress North, Katy-North, Cypress South, Copperfield Area}.

CONCLUSIONS

- Houston neighborhoods demographic data statistical analysis shows that 75% of the neighborhoods have more than 45,000$ median household income.
- 75% of the neighborhoods have a population of more than 18,000.
- 75% of the neighborhoods have a median age less than 38 years.
- Neighborhoods with high bachelor's degree education have higher median income and lower unemployment rate.
- Houston population and wealth are more concentrated at the west side of the city.
- Although existing venues are more concentrated inside the City of Houston loop near the Downtown area, Houston suburbs are good areas for investments given their high population and income.
- There is no clear relationship between existing venue counts and neighborhood population, suggesting the need for development plans to provide services to neighborhoods proportional to their population.
- The WCSS K-means optimization method way too many clusters, that could defeat the purpose of interpreting the results. Only 5 clusters were selected to drive a conclusion in terms neighborhood clustering and segregation based on venue categories.
- Clusters 0 and 3 are the most common clusters. Restaurants and fast food are the most common business venues in these clusters.
- Starbucks coffee shop has potential investment opportunities in neighborhoods with high population and income that doesn't have current existing venues near them.

FUTURE DIRECTION

- Include more factors in the analysis such as neighborhood crime rates, rent cost, areas with high traffic, corporate locations.
- Increase the area search around a neighborhood to collect more venues data.
- Increase the maximum search limit of venues in a neighborhood.
- Cluster and segment neighborhoods based on venue category and demographics.
- Automate the process of recommending locations based on the user’s criteria.