IBM Data Science Capstone Project Report

Battle of Neighborhoods-Houston

Mohab Dessouki

1. PROBLEM & BACKGROUND

The city of Houston is the fourth populous city in the United States, the most populous city in the state of Texas (www.houstontx.gov, 2020). The Houston-The Woodlands-Sugar Land Metropolitan Statistical Area covers 9,444 square miles, and area larger than five states, including New Hampshire, New Jersey, Connecticut, Delaware, and Rhode Island (www.houston.org). Houston has one of the most fast-growing and most diverse populations anywhere in the world (https://www.houston.org/why-houston, 2020). Houston is a hub for foreign trade and a great target for foreign investment. The Houston region offers a dynamic infrastructure to support fast growing industries including manufacturing, life sciences, aerospace, and energy. Houston is home to the largest medical complex in the world. Twenty-two Fortune 500 companies are headquartered in the Houston region. Houston is known as the Energy Capital of the World with more than 500 Oil and Gas Exploration and Production companies. Houston has been resilient to several global economic shocks thanks to its diversified and strong economy. The Houston region offers a diverse and highly skilled labor force of more than three million workers. Houston is a top 10 city for attracting millennials and about one-third of Houstonians 25 years or older is a college graduate (www.houston.org).

Houston is considered one of the most attractive cities for investment according to the "City attractiveness for investment" criteria described in (Snieska & Zykiene, 2015). 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 (http://www.bisworld.info). 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.

2. DATA DESCRIPTION

1. 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.

2. 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.

3. 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. Each neighborhood is clustered based on the most common venue categories present in it. This information can give investors an overview about the neighborhoods business types and the geographic concentration of a venue category.

3. PYTHON LIBRARIES INSTALLATION

1- 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).
2- Geocoder: Deals with multiple different geocoding service providers such as Google, Bing, OSM (https://geocoder.readthedocs.io/, 2020).
3- Numpy: An N-dimensional array object. Handles data in a vectorized manner.
4- Pandas: It offers data structures and operations for manipulating numerical tables (DataFrames).
5- Requests: Allows sending HTTP requests to a website server in an easy way (https://requests.readthedocs.io/en/master/, 2020).
6- Matplotlib: Used for creating plots in python (https://matplotlib.org/, 2020)
7- Scikit-learn: It features various classification, regression, and clustering algorithms including Kmeans (https://scikit-learn.org/stable/, 2020).
8- Folium: It enables both the binding of data to a map for visualization as well as passing HTML visualization as markers on the map.
9-

4. DATASET DOWNLOAD AND EXPLORATION

4.1 Getting the Neighborhood Demographic Data

The Houston neighborhoods demographic data was collected using a HTML request to URL (https://www.houstoniamag.com/home-and-real-estate/2017/03/neighborhoods-by-the-numbers-real-estate-data-2017). data was parsed using BeautifulSoup. The HTML file included the whole webpage information. The demographic data was included in two tables. The python find_all() method was used to create objects that only includes the two tables. The demographic data includes 7 variables (Neighborhood name, Zip code, Population, Median Income, Median Age, Unemployment, and Education). A list for each variable was created and added to a DataFrame called df_Houston. The data was cleaned by removing the special characters from it and converting the values from string to its proper type. A DataFrame with 147 Neighborhoods demographic data was created.

4.2 Neighborhood Demographic Data Boxplots

Boxplots provide summary statistics used to explain the distribution of numerical data and skewness through displaying the data quartiles. The box plot describes 6 main properties in dataset (Minimum Score, Lower Quartlle, Median, Upper Quartile, Maximum Score, Interquartile Range). **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. It can be inferred from the box plot shown in figure.1 that around 75% of the neighborhoods have a population of more than 19000. Figure. 2 shows that 75% of the neighborhoods have a median income higher than 45000$. Figure. 3 shows that 75% of the neighborhoods have a median age of less than 38 years. Which indicates a relatively youthful population. Figure. 4 shows that Figure 4 Seventy-five percent of the neighborhoods have more than 80% of their population hold a bachelor’s degree.

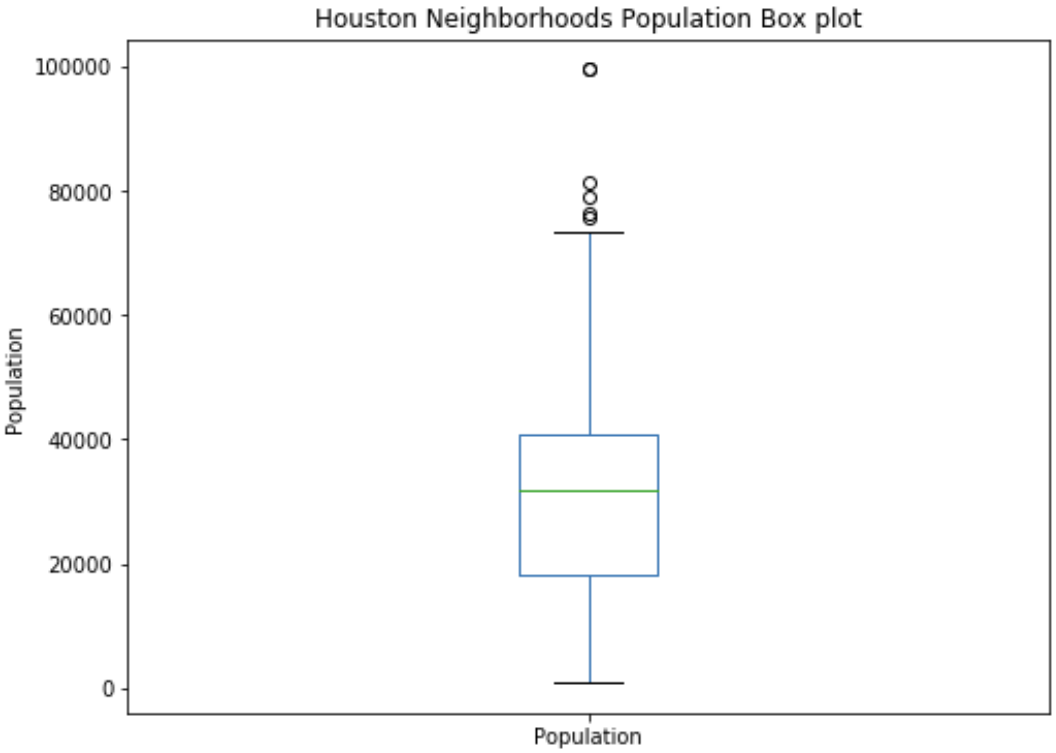

Figure 1 The Population box plot shows that 75% of the neighborhoods have a population higher than 19000.

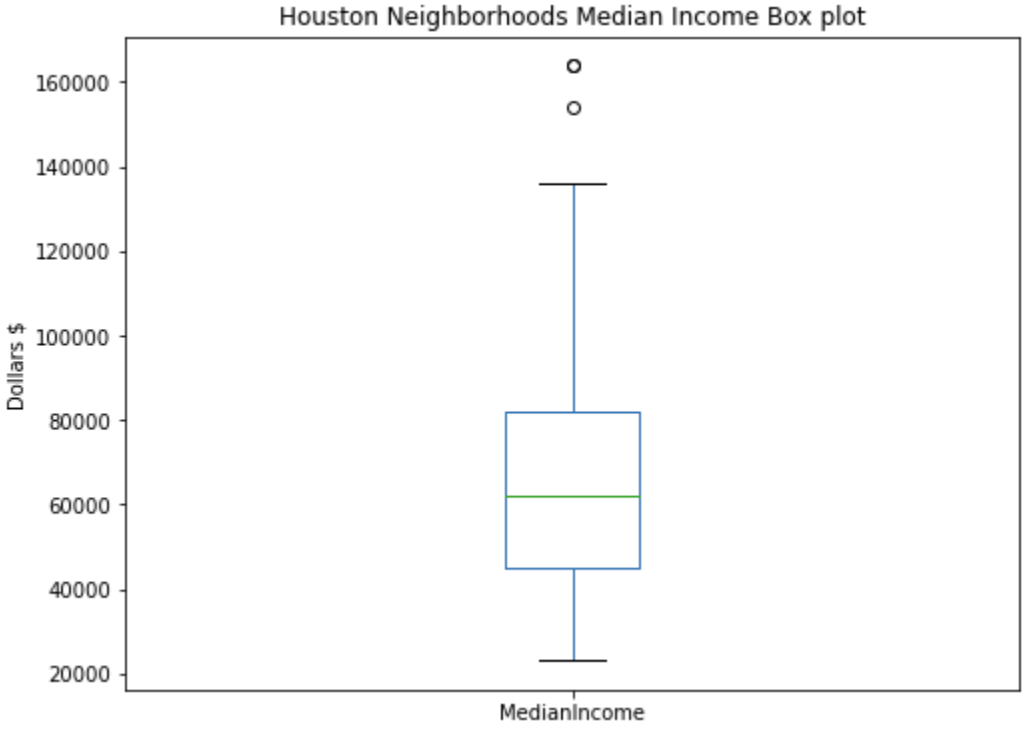

Figure 2 The Income box plot shows that 75% of the neighborhoods have a median income higher than 45000$.

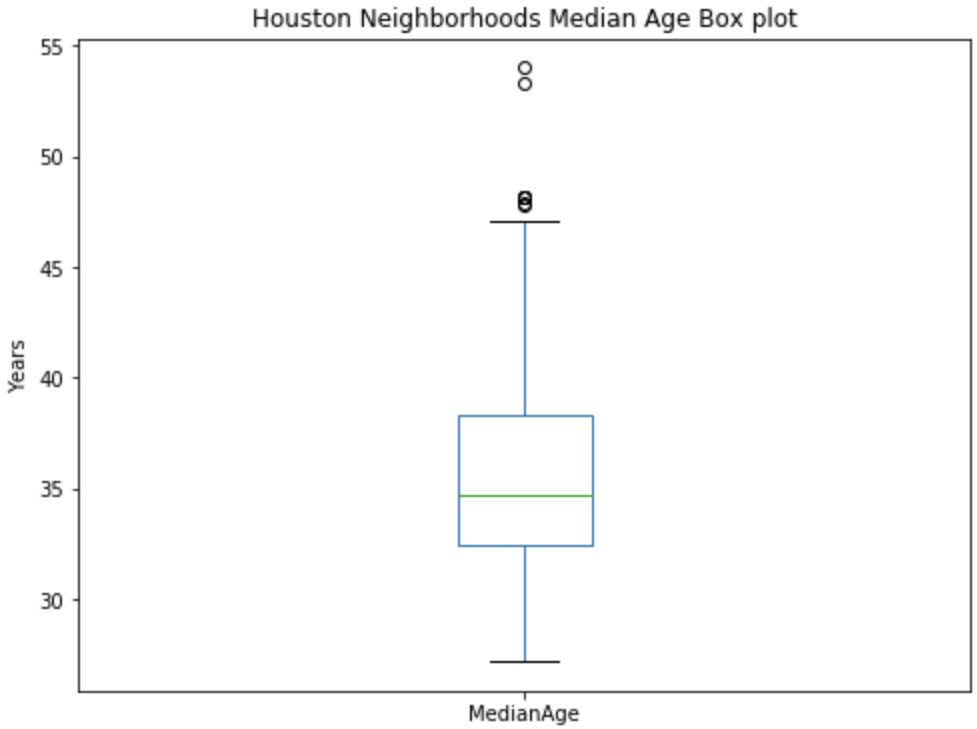

Figure 3 The Age box plot shows that 75% of the neighborhoods have a median age lower than 38 years.

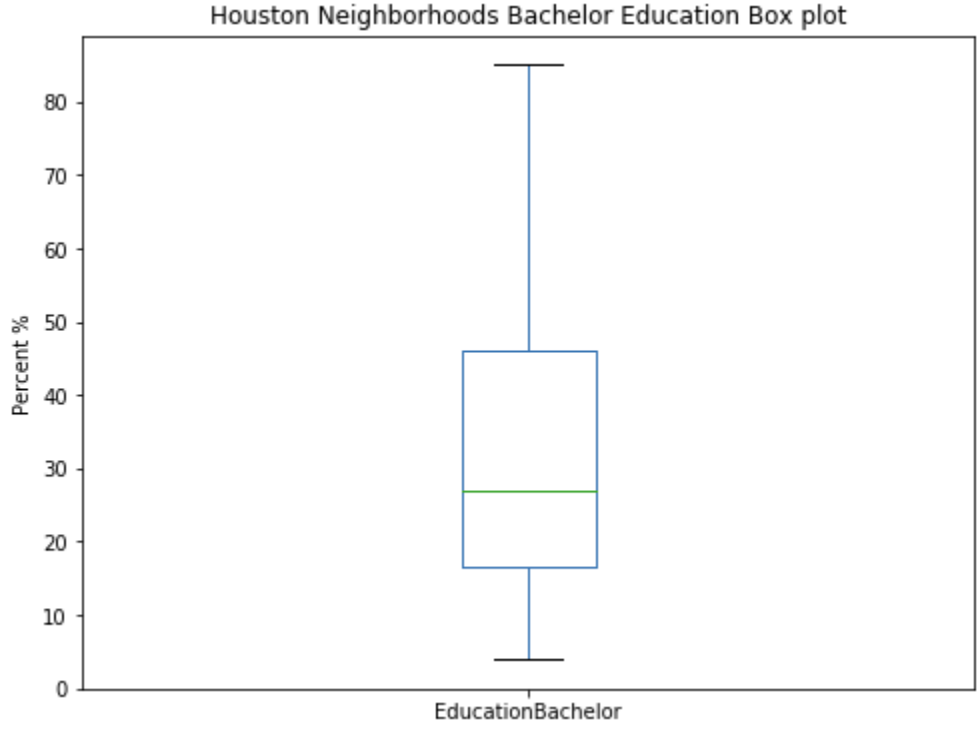

Figure 4 Seventy-five percent of the neighborhoods have more than 80% of their population hold a bachelor's degree.

4.3 Neighborhood Demographics 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. Table. 1 shows a summary of the regression analysis. We can observe that income increases with education and unemployment decreases with increasing education. The relationships are shown in Figures 5 and 6.

Table 1 A summary of the regression analysis between income and other demographic variables

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

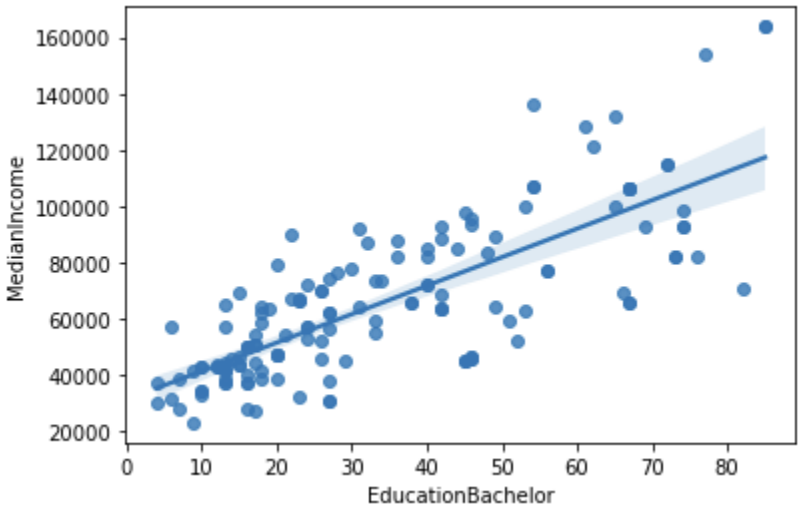

Figure 5 The median income of a neighborhood is positively correlated to the percentage of bachelor's degree population

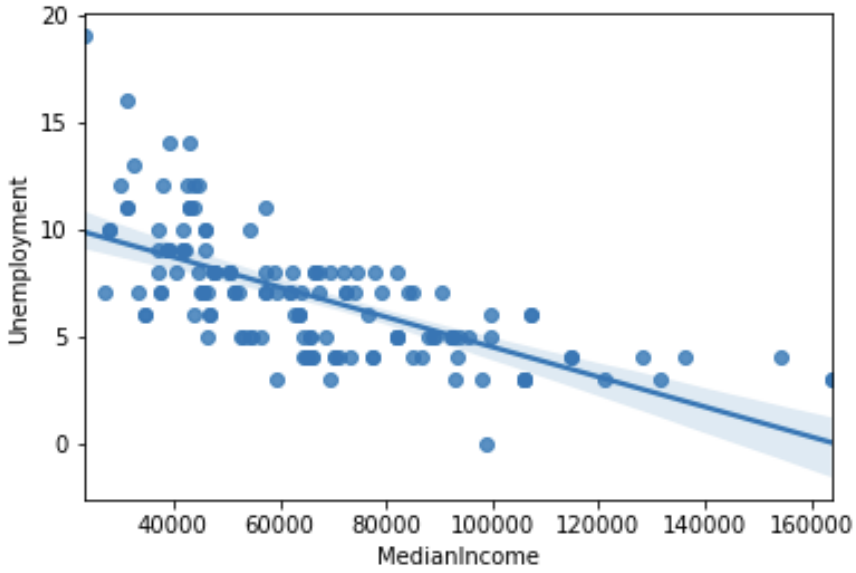

Figure 6 The unemployment seams to decrease with the increase in the education percentage

5. EXPLORING NEIGHBORHOODS IN THE CITY OF HOUSTON

In this section, we will look at the demographic distribution on the Houston map as well as the top venues at each neighborhood. First, we need to get the geographic coordinates of each neighborhood. The Bing geocoder was used to get the Latitude and Longitude of each neighborhood in Houston. The coordinate data was added to the df_Houston panda data frame. 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. Figure. 7 and 8 show the location of the neighborhoods, the marker size represents the relative size of population and income respectively in both figures. Similarly, a variable that can describe both the neighborhood population as well as income can indicate the economic strength and business activity in a neighborhood. The marker size Figure. 9 represents the multiplication product of population and income.

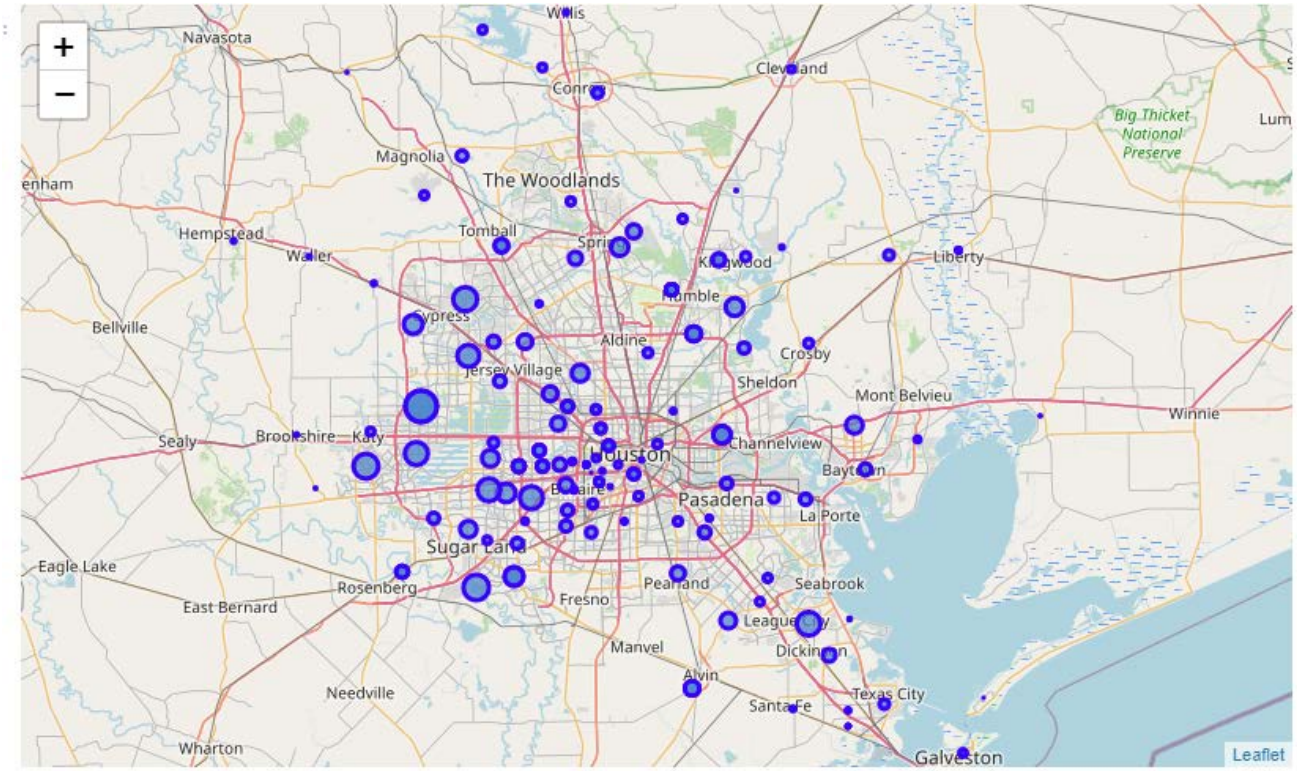

Figure 7 Each marker represent a neighborhood location, the size of the marker represents the relative size of the population in each neighborhood.

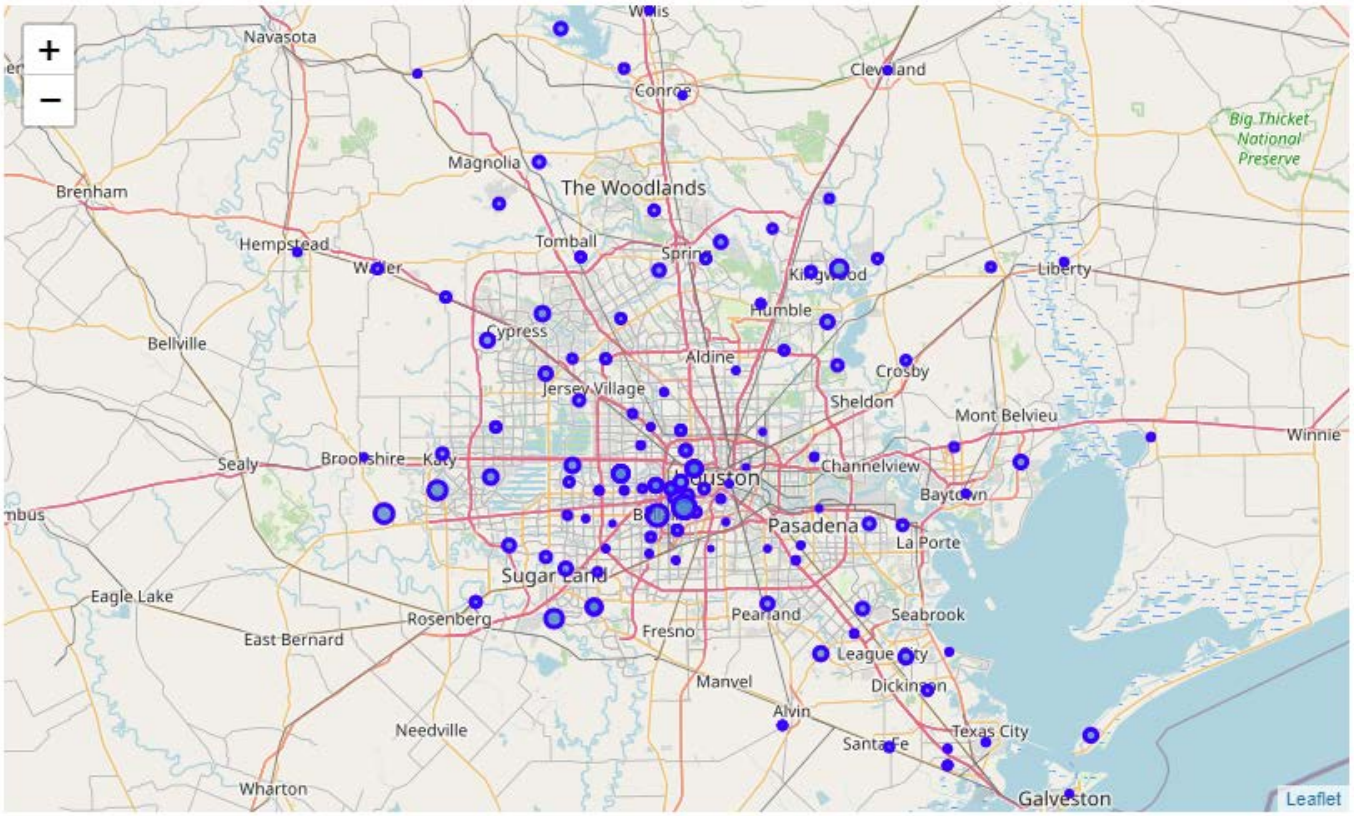

Figure 8 Each marker represent a neighborhood location, the size of the marker represents the relative size of the median income in each neighborhood.

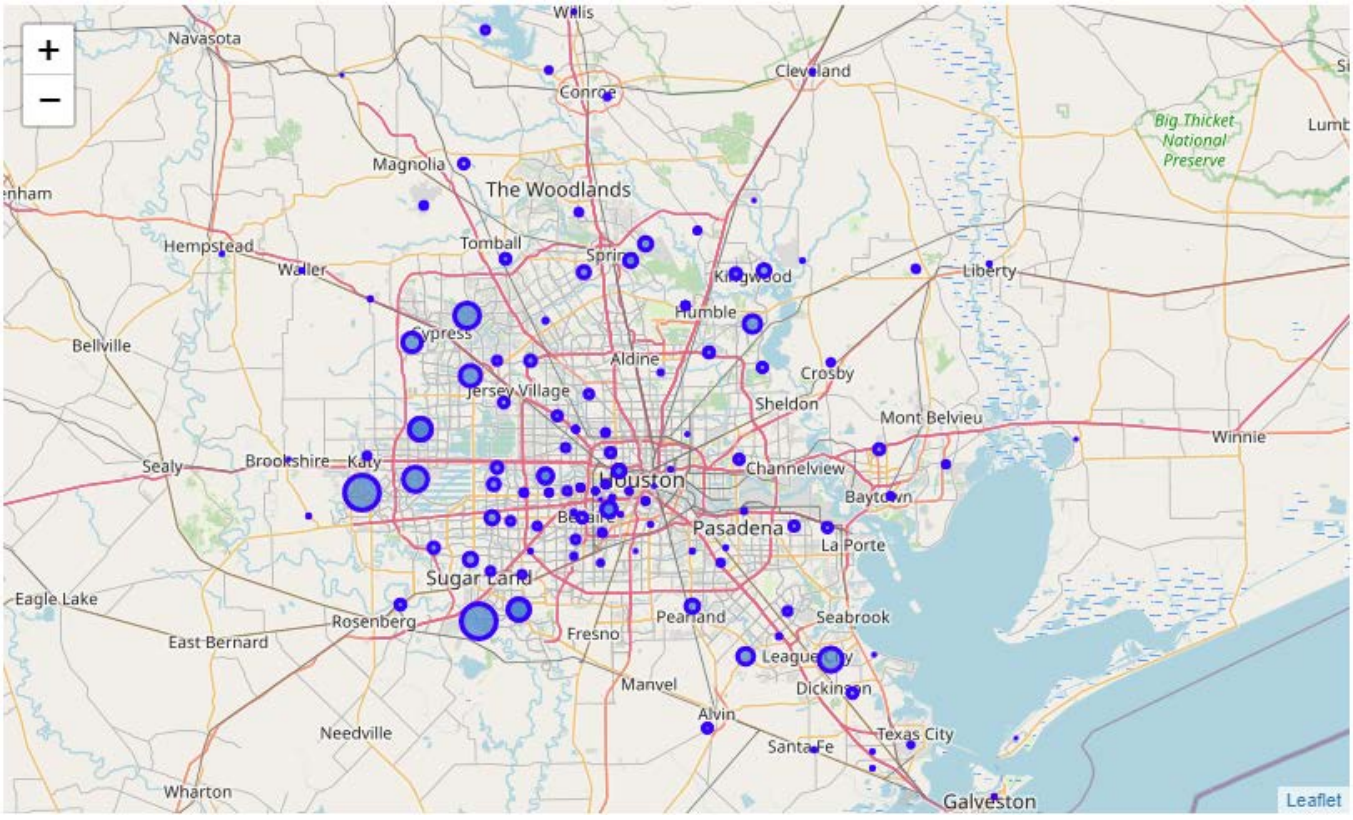

Figure 9 Each marker represent a neighborhood location, the size of the marker represents the magnitude of the multiplication product of the population and income.

6. 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 in Figure 10.

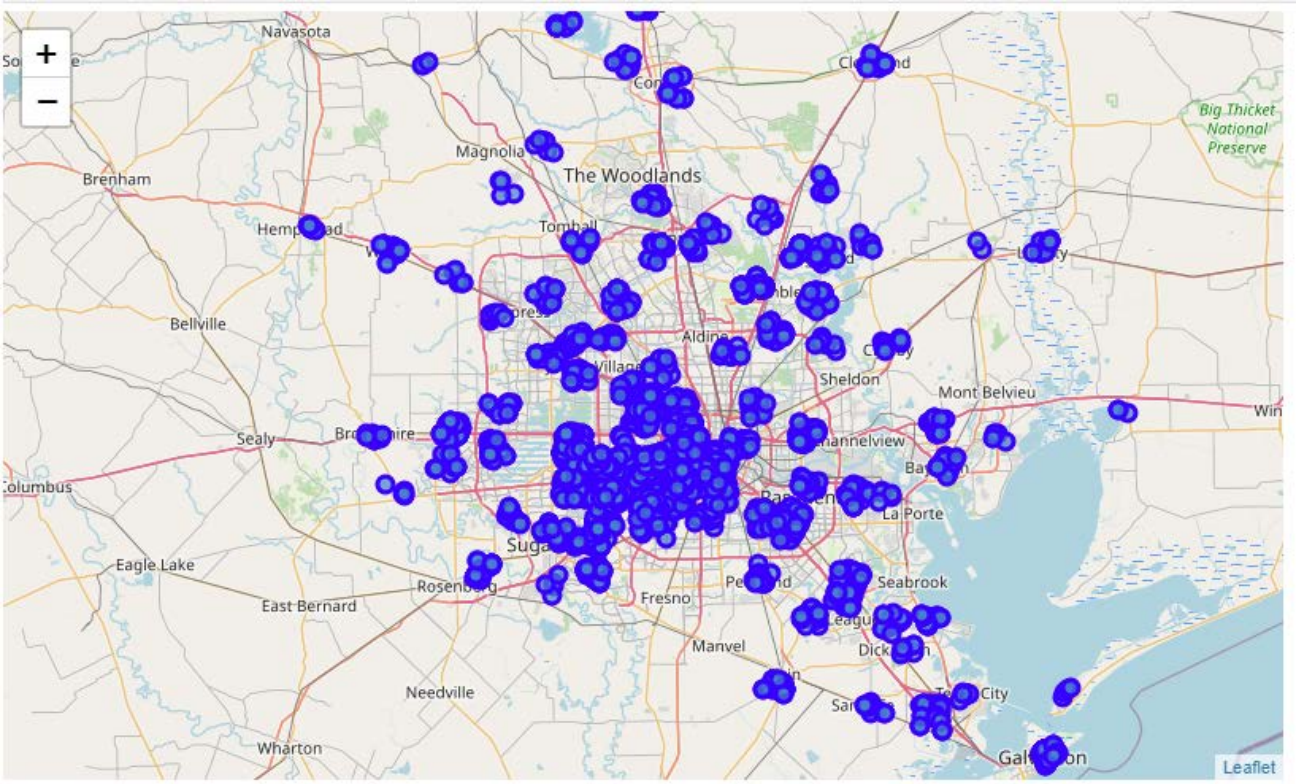

Figure 10 The venues locations are represented with a blue marker.

6.1 Frequency of a Venue Category in a Neighborhood

A categorical variable was assigned to each unique venue category. The mean of the frequency of occurrence of each category was calculated for each neighborhood. The most common venue types for each neighborhood were obtained. An example of the most 5 common venues is shown in Table 2.

Table 2 The top 5 common venues and their frequency in the Heights neighborhood

```
----Heights/Greater Heights----
                venue  freq
0  Mexican Restaurant  0.05
1  Italian Restaurant  0.04
2         Coffee Shop  0.03
3          Donut Shop  0.03
4                Park  0.03
```

6.2 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 in Figure 11.

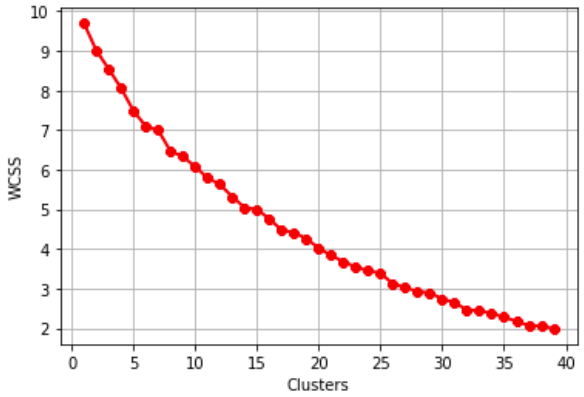

Figure 11 The WCSS is decreasing with increasing the number of clusters.

Although the optimum number of clusters seams 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. Figure 12 shows the color-coded neighborhoods clusters.

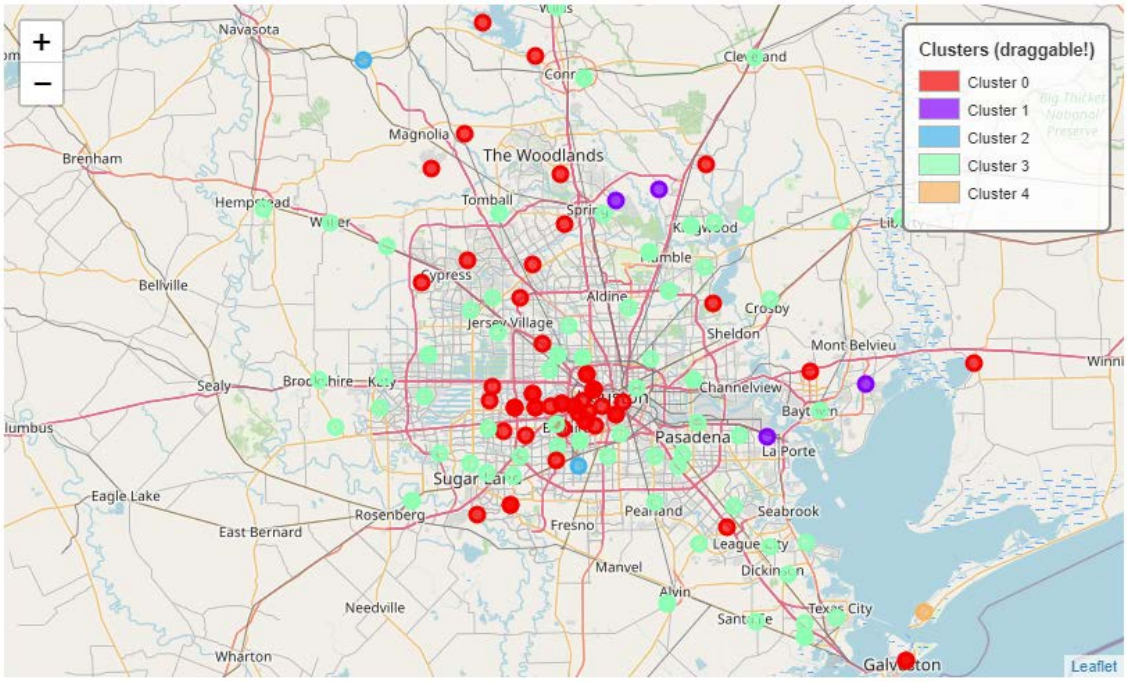

Figure 12 The color-coded clusters locations.

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

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

Neighborhoods with high population and median income could be attractive for new investments. Figure 13 shows the neighborhoods locations 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.

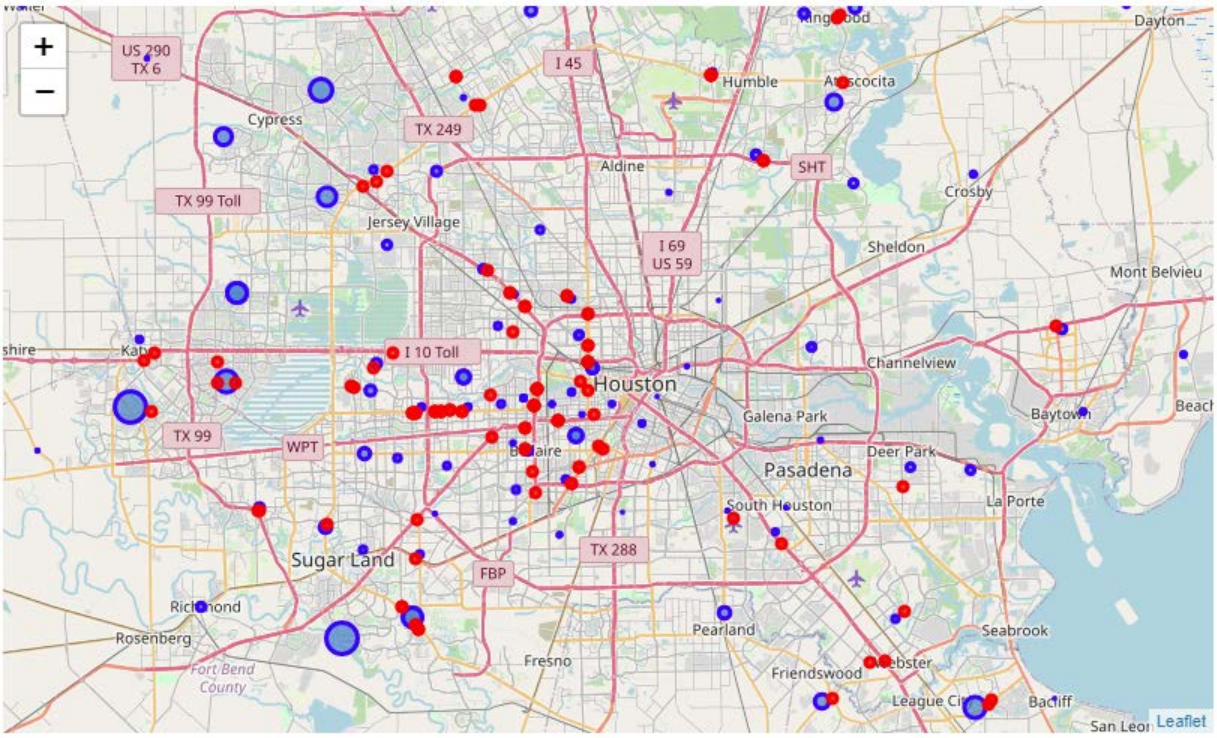

Figure 13 The blue marker represents the neighborhoods locations, and the size of the marker represent the magnitude of the population and income. The red markers represent the locations of existing Starbucks locations. According to this map, we can see that there are investment opportunities in the wealthy and populated areas represented by large markers that don't have a close by Starbucks venue.

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}.

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

8. 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.

9. 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.