**Analyzing neighborhoods in Houston to start a private school**

**Saranya Kamalasekaran**

**March 22,2021**

1. **Introduction: Business Problem**

The purpose of this project is to analyze neighborhoods in Houston to find the best location to build a private school. Houston is a large metropolis in Texas with 2.3 million residents, is the fourth most populous city in the United States, trailing after New York, Los Angeles, and Chicago. The city is the largest in the South and the Southwest.

In this Data Science project, we will analyze the number of private schools all over Houston and its neighborhoods, by visualizing the data points on Houston map to locate a specific area with minimum number of private school close with valuable reasoning. We are also looking for a developed area with all amenities, where most families are attracted. This project will be highly valuable for new business owners and stakeholders interested in opening a private school in Houston, Texas.

In this Data Science workflow, automated ML algorithms are used to find better recommendations for the stakeholders.

1. **Methodology**

In this project we are focusing on finding areas of Houston with a smaller number of private schools.

In the first step, we will import USA private school data and filter schools pertaining to Texas. Then, we use Houston county dataset and merge both data frames to obtain private school dataset that belongs to Houston. Once the data is ready. We **analyze and explore the density of private schools** across different counties and neighborhoods in Houston.

In second, we will **use Folium and map the private schools** to get a clear picture of how they are distributed across Houston. From this, we can identify a few promising areas with low number of private schools in general and focus our attention on those areas.

In third and final step we will focus on most promising areas. We will use Foursquare API to find top 10 venues of each neighborhood in Houston and **clustering them into groups based on their development**. We are looking for a promising area with low number of private school that matches with the cluster that designated as highly developed area since housing market will be booming in those area, it will be a perfect place for stakeholders to build a private school.

1. **Data acquisition and cleaning**

Based on the problem, factors that influence the decision are,

* number of existing private schools in Houston.
* number of neighborhoods in Houston and its outskirt.
* neighborhood top 10 venues.

**3.1 Data sources**

Following data will be needed to generate the required information.

* List of Houston Neighborhoods is extracted from <https://en.wikipedia.org/wiki/List_of_Houston_neighborhoods>.
* The private school dataset used in this project is obtained from Kaggle which is originally obtained from the US Department of Homeland Security. It contains information about all private schools with attributes regarding their geographical distribution. <https://www.kaggle.com/andrewmvd/us-schools-dataset>
* List of Houston county dataset scraped from 'https://en.m.wikipedia.org/wiki/Greater\_Houston#Counties' is used to filter private school dataset pertaining to Houston.
* The geographical location required for this project is obtained from Foursquare API.

**3.2 Data cleaning**

**Private school dataset**

The private school dataset for this project is imported are then read into a panda data frame using pd. read\_csv () method. It contains all US private school data with

attributes regarding their geographical distribution.

Among 33 attributes (columns) in this dataset, we pick 8 attributes that are necessary for this project. Dropping off redundant data, the first 5 rows after the transformation are

shown below.

****

Figure 1: Private school dataset after redundancy

For our business problem, we need dataset pertaining only Houston private school. To overcome this problem, we use Houston county dataset scraped from 'https://en.m.wikipedia.org/wiki/Greater\_Houston#Counties'.

Aggregating both Private school and Houston county data frame using merge () method. We obtain Houston Private school dataset; the first 5 rows are shown in Figure 2.

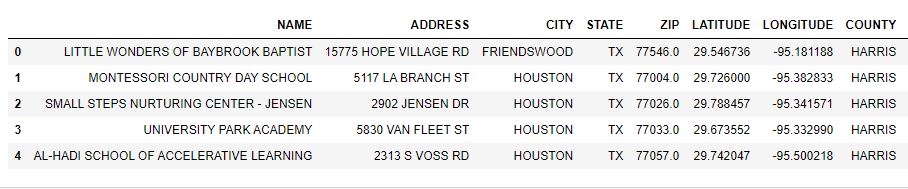
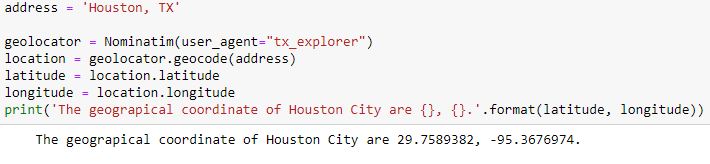


Figure 2: Houston private school dataset

**Geographical coordinates**

The geographical coordinates for Houston have been obtained from the Geopy library in python. This data is relevant for plotting the map of Houston using the Folium library. The code for getting the geographical coordinates of Houston is shown below.



The geolocator library is also used to obtain Houston neighborhood coordinates, then converted into panda data frame for further analysis. The top 5 rows of Houston neighborhoods and its coordinate is shown below.

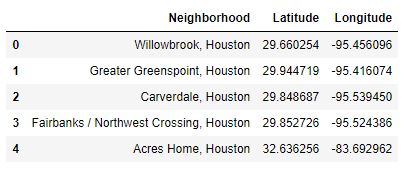


Figure 3: Houston neighborhood coordinates obtained using geolocator

**Venue data**

The venue data has been extracted using the Foursquare API. This data contains venue recommendations for all neighborhoods in Houston and is used to study the popular venues of different neighborhoods as well as build the unsupervised learning model to cluster neighborhoods. The venue recommendations of all neighborhoods were obtained with a limit of 200, that is, maximum of 200 venue recommendations per neighborhood and a radius of 10km around the neighborhood’s geographical coordinates. Figure 4 shows the top 10 rows depicting the results obtained after cleaning the data from Foursquare API.

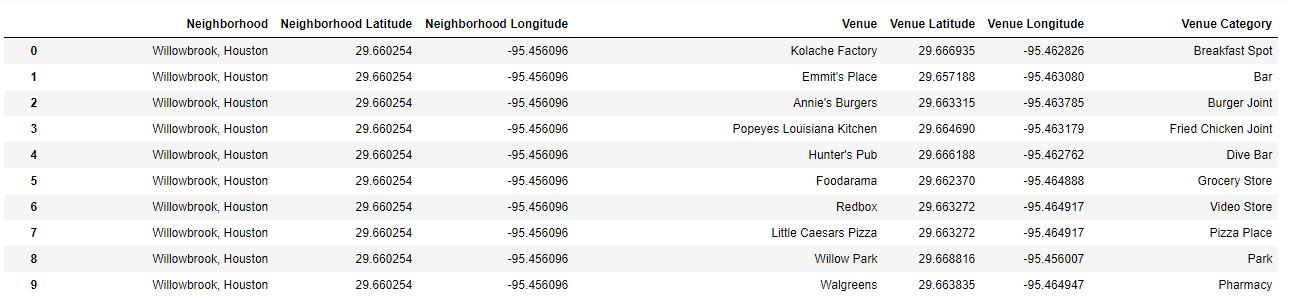
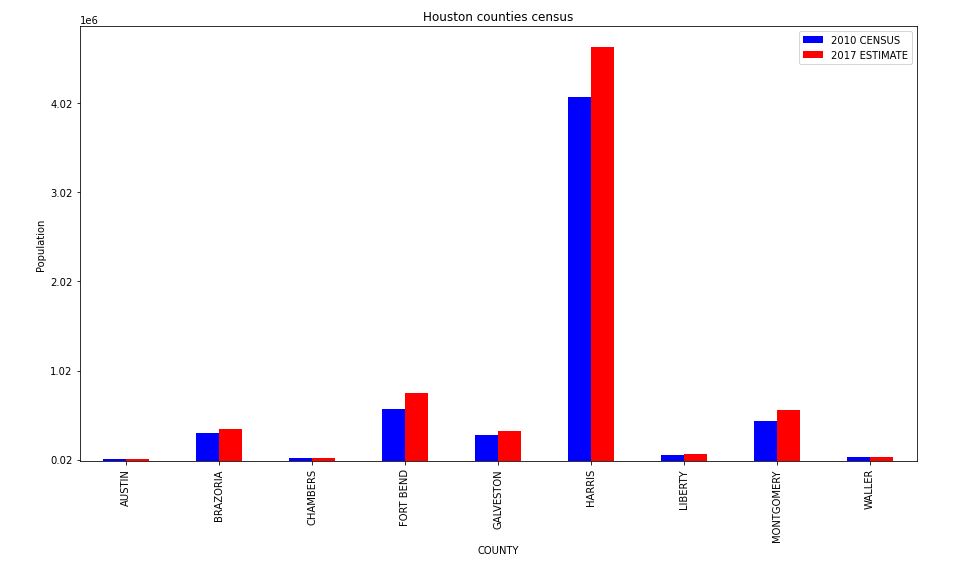
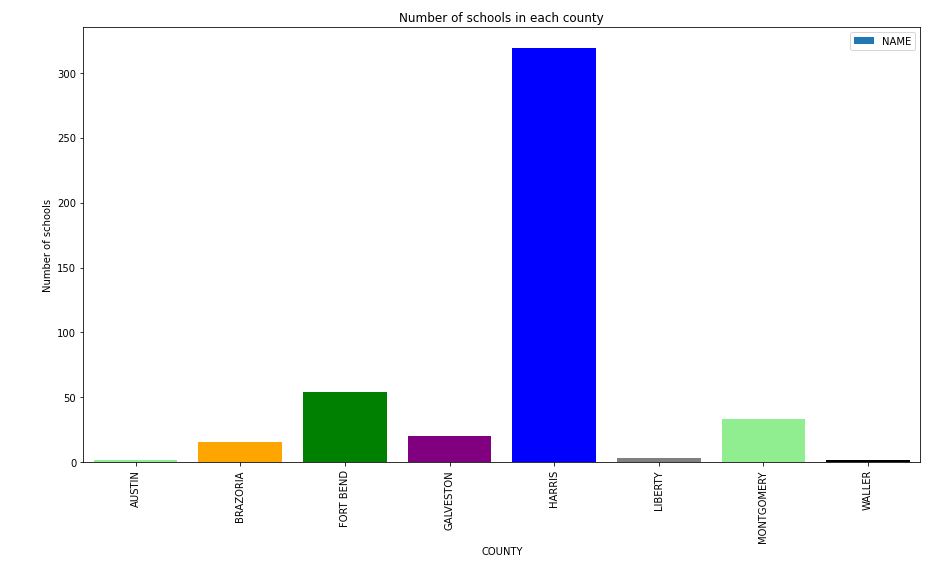


Figure 4: Houston neighborhood venues

1. **Preliminary data analysis**

Visualizing the difference in population of each county in span of 7 years, it’s evident that there wasn't much of development in Austin, Liberty, and Waller county in the span of 7 years. These counties might have a slow pace in development in future. Our best shot is to go with Brazoria, Fort Bend, Harris, Galveston, and Montgomery counties, with significant increase in population.

****

Let’s look at the number of private schools and its distribution all over Houston, to support our idea.****

Notice**,** that the density of school is high in the very same counties (Brazoria, Fort Bend, Harris, Galveston, and Montgomery counties) where the population growth is significant. Using folium, a map is plotted to show how the private schools are spread across Houston and the color signifies different county. This is shown in Figure.

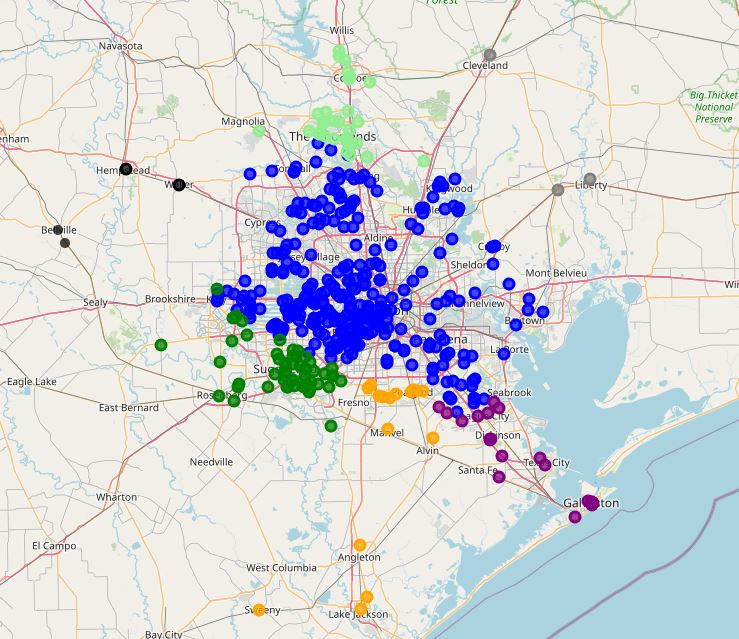
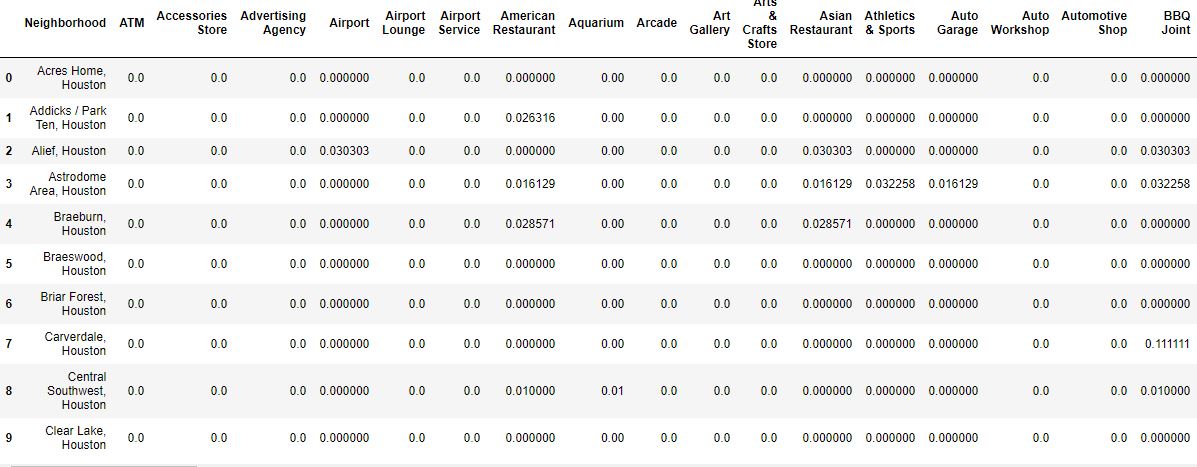


Figure: Distribution of private schools

It is also evident that private schools are crowded on Southwest part of Houston. Let us focus our attention to the lower density area, to the Northeast part of Houston which is also a part of Harris county where the growth is high as mentioned before. In the following section, analysis on Houston neighborhoods’ venues is performed to support our hypothesis.

1. **Feature selection**

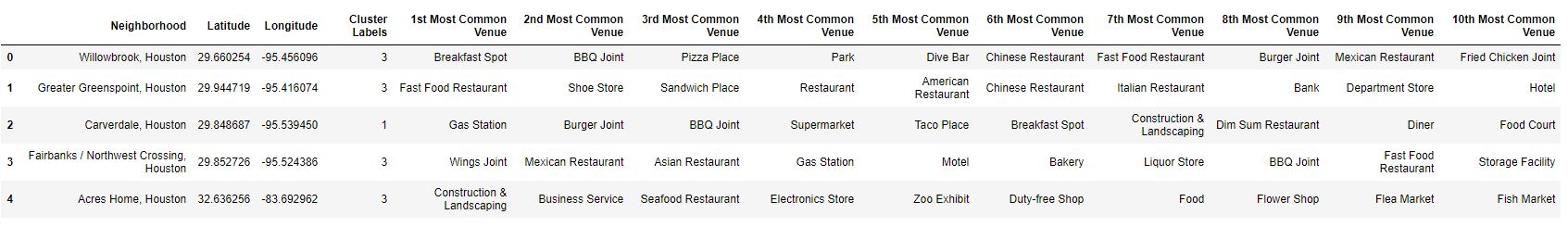
Feature selection is carried out on the venue data from Foursquare API to perform unsupervised learning (clustering) model. We use one-hot encoding process to convert venue data into numerical value by creating column for each unique category. If a neighborhood venue belongs to that category, it assigns a value of 1 for that row in that specific category column and if the venue does not belong, the value is assigned to 0. This process is repeated for all venues in each neighborhood and the result was a sparse matrix containing the neighborhood name and all unique category columns with either 1 or 0. The dataframe rows are then grouped by taking the mean of frequency of occurrence of each category. The result is shown below with only top 10 rows.



One-hot Encoding

Notice that most of the values are 0 since not all neighborhood has venues belonging to each unique category. This data is then used for the unsupervised model with Neighborhood name dropped.

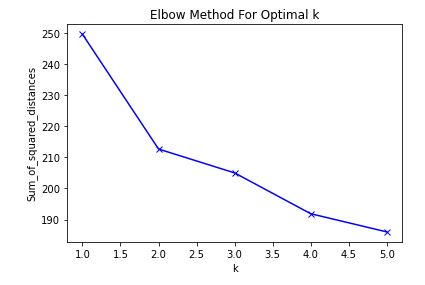
A dataframe containing top 10 venues of all neighborhood is also created, which later combined with the unsupervised learning model for further analysis. The top 5 rows are shown.



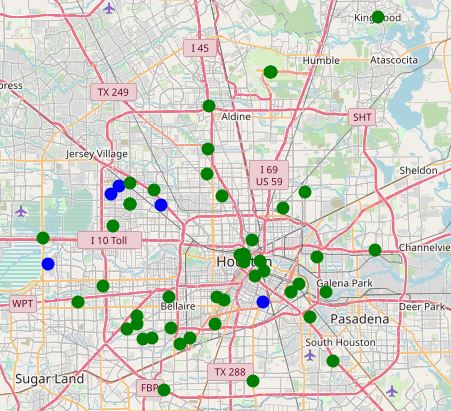
Top 10 most common venues for neighborhoods

1. **Unsupervised Learning (K-means clustering)**

We run unsupervised clustering model on Neighborhood data to differentiate them based on development. To obtain optimal result, first we must determine the number of clusters (k value). We pick value of k in range 1 to 6 and for each k value, we will initialize k-means and use the inertia attribute to identify the sum of squared distances of samples to the nearest cluster center. As k increases, the sum of squared distance tends to zero. Imagine we set k to its maximum value n (where n is number of samples) each sample will form its own cluster meaning sum of squared distances equals zero. Below is a plot of sum of squared distances for k in the range specified above. If the plot looks like an arm, then the elbow on the arm is optimal k.



In the above graph, we can interpret that optimal k value is 2. So, the data will be clustered into 2 different groups as developed and developing neighborhoods. Using Folium, the clustered neighborhoods are displayed on map as shown below.

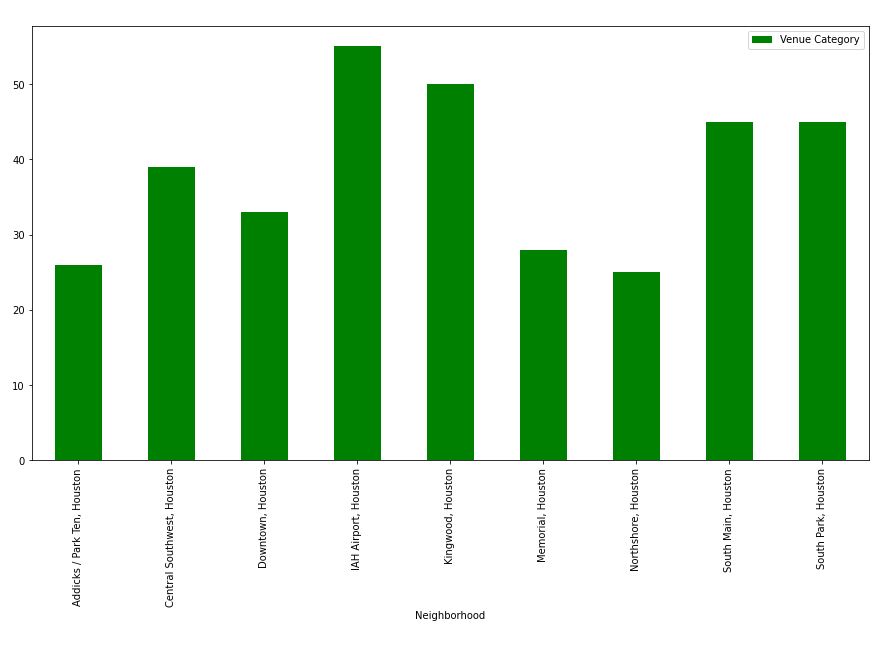


Clustering of Neighborhoods

In the map, we can see the green dots are clustered into developed neighborhoods and blue dots clustered as emerging neighborhoods.

## Results and Discussion

As mentioned before focusing our interest on North-east part of Harris county and interpreting the neighborhoods data based on development (venue categories) as in the figure below, we can pinpoint a particular neighborhood to open the new private school.

 In the above figure, the top two places that has maximum number of venue categories (IAH airport and Kingwood) are in Harris county. In this, IAH airport is considered as biased since it’s the congregating place for travelers. We pick Kingwood as our area of interest to build the new private school.

The result is highly based on county census, density of private school and neighborhood development. Other data that can interpret the result are the average income and race of the population living in the neighborhood, further analysis can be done to support our result.

## Conclusion

Purpose of this project was to identify Houston areas close to center with low number of private schools to aid stakeholders in narrowing down the search for optimal location for a new private school. By analyzing private school data and visualizing density distribution, we have first identified general area (north-east part of Houston), and then with collection of top 10 venues data using Foursquare which satisfy some basic requirements regarding developed neighborhood. Clustering of those venues was then performed to create major zones of interest (containing greatest number of potential locations) to be used as starting points for final exploration by stakeholders.

Final decision on optimal private school location will be made by stakeholders based on specific characteristics of neighborhood and locations in recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.