**Predicting the best Neighborhood to establish a new Italian restaurant in Houston, TX**

Marcus Cobb

May 2, 2020

**1. Introduction/Business Problem**

**1.1 Background**

Houston, Texas is the largest city in Texas and is one of the largest Cities in the US and it has a total area of 637.4 square miles. Houston is divided into 88 different Super Neighborhoods that represents different areas of the city. Determining the best location to build/invest in a new restaurant will be critical to how well the restaurant preforms. Before investing in a new restaurant, you need to understand the demographics of the city, in order to determine an ideal location.

**1.2 Problem**

Data can might contribute to determining the ideal Super Neighborhood to establish a new Italian Restaurant are the number of competitor’s (other Italian restaurants) per Super Neighborhood, the number of total restaurants per Super Neighborhood, the number of Other Venues per Super Neighborhood. This analysis of this data will help predict what Super Neighborhood will have the lowest number of competitors and have a larger number of potential customers (ie. Tourists).

**2. Data**

**2.1 Data sources**

* The geo data for the Super Neighborhoods boundaries will come from <https://cohgis-mycity.opendata.arcgis.com/datasets/coh-super-neighborhoods>, which is a geojson file of the 88 Super Neighborhoods in Houston.
* The restaurant data will be acquired through the Foursquare API for each Super Neighborhood. The restaurant data will be divided into two categories Italian Restaurants (competitors) and all other restaurants (these will still take away potential customers, but will not directly compete with an Italian Restaurant).
* The Other Venues (or things to do) will also be acquired through the Foursquare API for each Super Neighborhood. An increase number of attractions in a given Super Neighborhood will increase traffic of potential customers to include both tourist and residents. The number of attractions in a given neighborhood and also weighting the reviews of these attractions will help understand which neighborhoods will potentially have more people on a daily basis.

**2.2 Data Cleaning**

The geojson file for the super neighbor hoods was cleaned and converted into a panda’s data frame. The geojson did not have center points for the neighborhoods, so I used Google Earth to load the layer and find the center points. I added the center point’s lat/lons to a csv file with the neighborhood names. I combined the center point csv with the data frame made from the geojson file. After querying the Foursquare API, I cleaned the results json and converted it in to a data frame. I then filtered the Foursquare results into three different categories (Italian Restaurants, Other Restaurants, and Other Venues). I then did a count of each category and added it to the data frame made from the geojson file.

**3. Methodology**

**3.1 Data Exploration**

Houston, TX is comprised of 88 Super Neighborhoods. Out of the 88 Super Neighborhoods I divided the data into three data frames based on the number of Italian Restaurants in each neighborhood. The three data frames are Neighborhoods with no Italian restaurants (40 Neighborhoods), Neighborhoods with only 1 Italian restaurant (23 Neighborhoods), and Neighborhoods with 2 or more Italian restaurants (25 Neighborhoods).



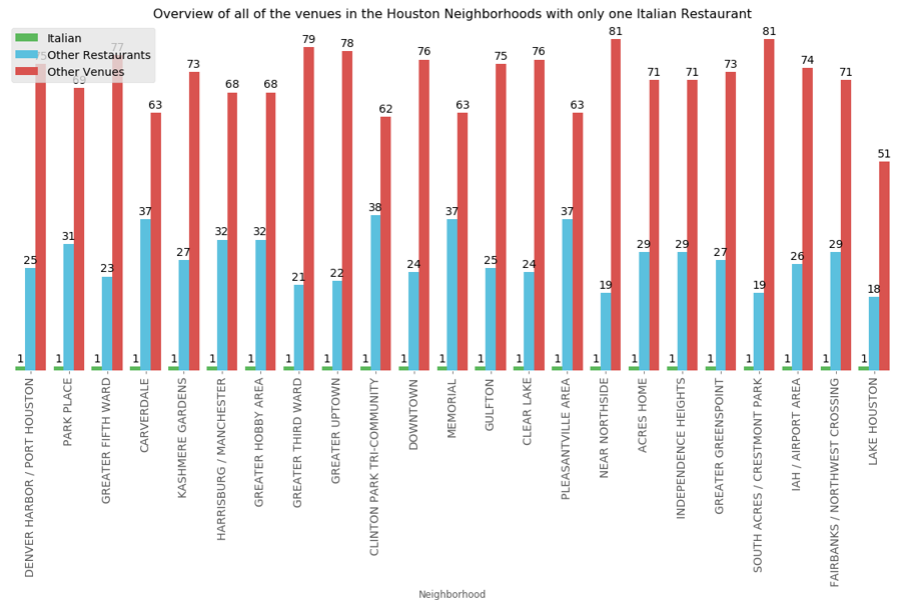


**3.2 Relationship between the target variable and other Variables**

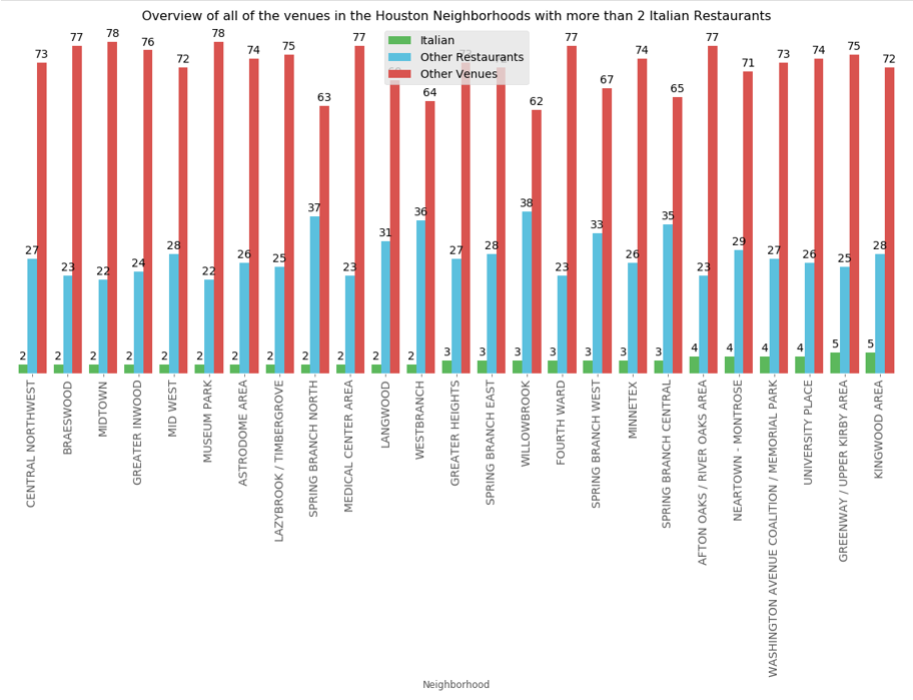
The data for Other Venues and other Restaurants were compared to Italian restaurants to determine if there is a correlation between the number and Other Venues and Other Restaurants to Italian restaurants in Houston Neighborhoods. The Box plot depicted below shows a correlation that a Italian restaurant is more likely in Neighborhoods with more Other Venues and less Other Restaurants.



**3.3 Neighborhoods with 1 Italian Restaurant, Variable comparison**



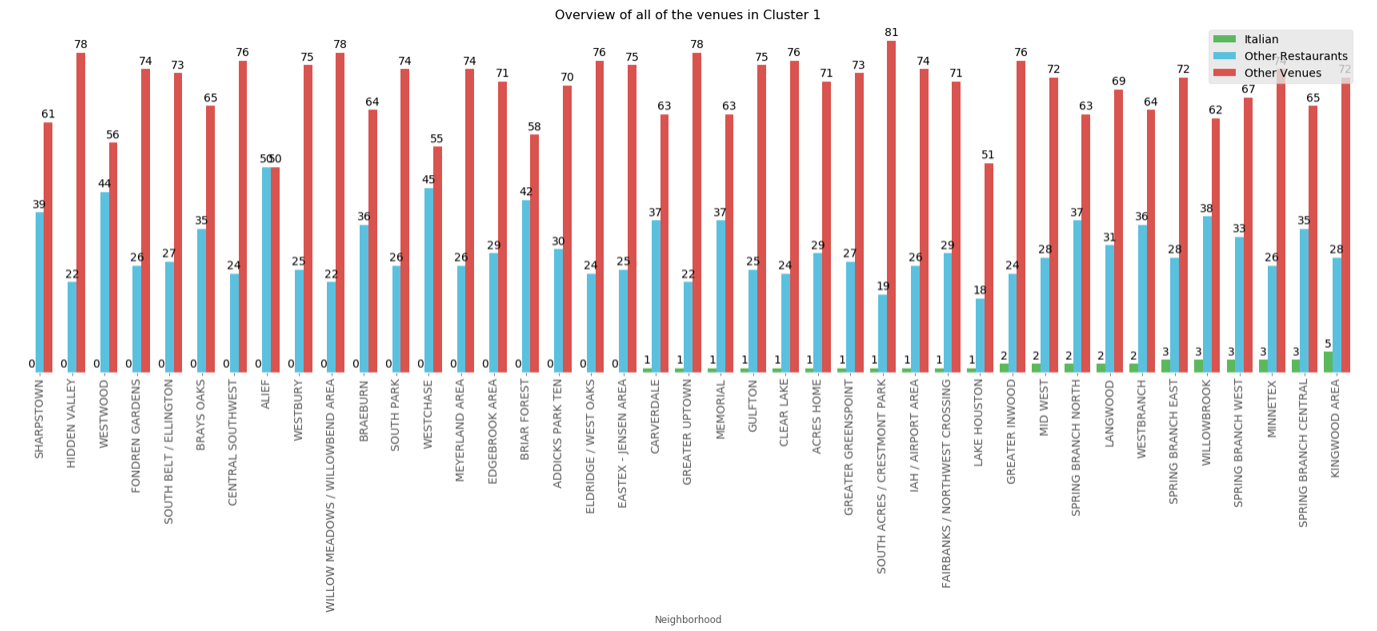
**3.4 Neighborhoods with more than 2 Italian Restaurants, Variable comparison**



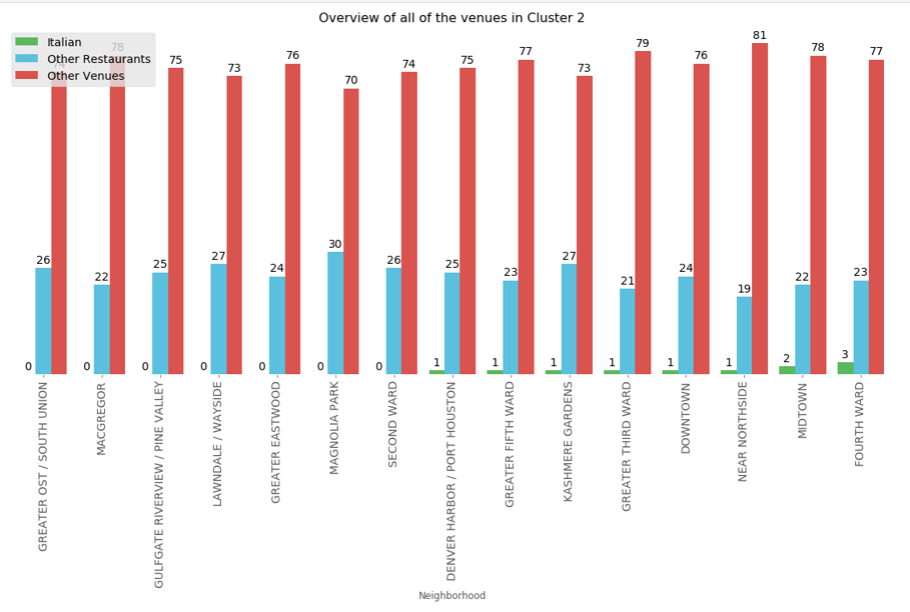
**3.5 K\_means clustering**

I created 5 Clusters of the Super Neighborhoods in Houston, based on the most common venues in each neighborhood. Each neighborhood in a cluster will be similar to the other Neighborhoods in that cluster and will help determine which neighborhoods will support a new Italian Restaurant, based on the similar neighbor in the cluster.

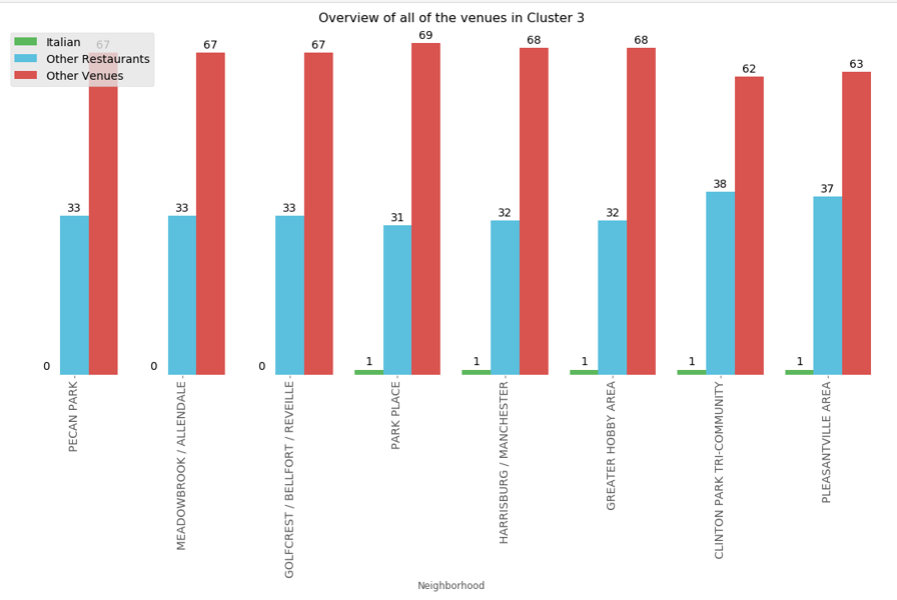
Cluster 1 is the largest cluster and almost half of the Neighborhoods with in Cluster 1 have no Italian Restaurants. The 2 most common venues in cluster 1 are Burger Joint and Mexican Restaurants.



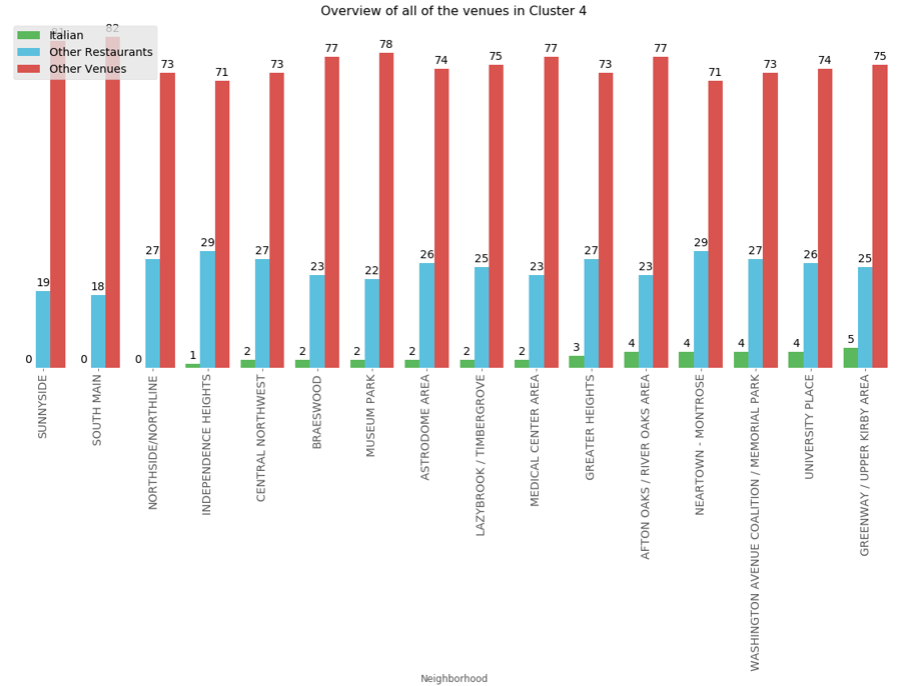
Cluster 2 is a smaller cluster with almost half of the Neighborhoods in Cluster 2 having no Italian Restaurants. The 2 most common Venues in Cluster 2 are Park and Coffee Shop.



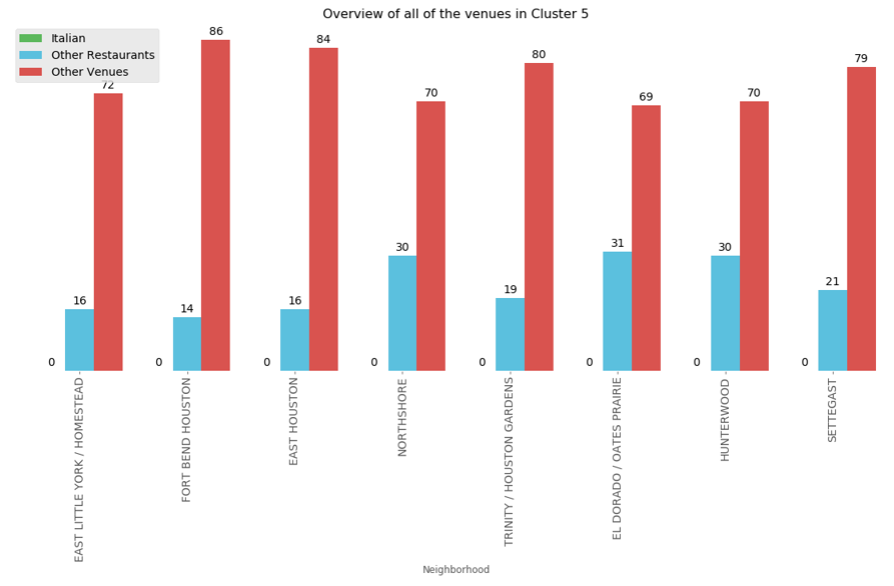
Cluster 3 is the smallest of the 5 clusters and the majority of the clusters have Italian Restuarts, but the Neighborhoods that do have Italian Restaurants only have one Italian Restaurant. The most common venues in Cluster 3 are Mexican Restaurants and Burger Joints.



Cluster 4 is primarily comprised of Neighborhoods with Multiple Italian Restaurants. Only three of the Restaurants in Cluster 4 do not have any Italian Restaurants. The most common venues in Cluster 4 consist of Trail, Zoo Exhibit, Parks, and Grocery stores.



Cluster 5 is a small cluster and none of the Neighborhoods in cluster 5 have any Italian Restaurants. The most common venues in Cluster 4 are Discount Stores and Mexican Restaurants.

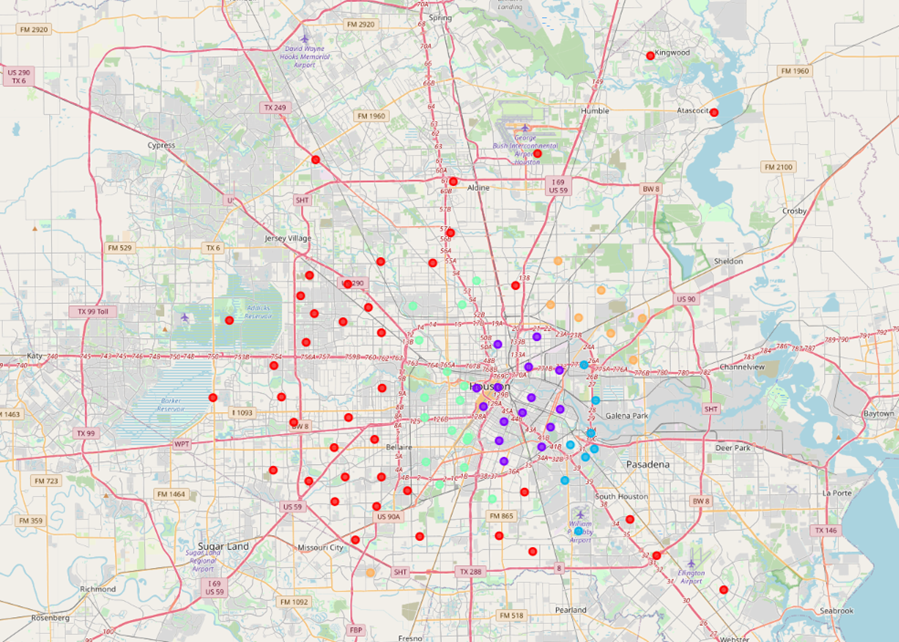


**4. Results**

I combined the K\_means Clusters of the most common Venues with the count of each venue made at the beginning of this project. I created the all of the above charts using these tables.



I then plotted all of the clusters on a map to display where the clusters are geographically. In the map below we can see all 5 clusters plotted on the Folium map. Red is Cluster 1, Purple is Cluster 2, Blue is Cluster 3, Green is Cluster 4, and Orange is Cluster 5



**5. Discussion**

Houston, TX is a large city comprised of 88 Super Neighborhoods. These neighborhoods are all unique and have different economies. However, by grouping similar neighborhoods in clusters we found the Neighborhoods that have similar economies. It is obvious which clusters will support similar venues.

I started this project by discovering the center points of each Super Neighborhood and then using those center points to query the Foursquare API. Then I had to clean the data to put them in to data to visualize it with bar charts.

I then conducted K-means clustering of the data frame to find the similar Super Neighborhoods, I set the K value to 5. This divided the 88 Neighborhoods into 5 clusters of similar Neighborhoods.

I ended the project by plotting the clusters on a map and depicting bar charts of each Neighborhood.

**6. Conclusion**

In Conclusion we can see that the two Super Neighborhoods that are most likely to respond well to a new Italian Restaurant are Sunnyside and Southmain. These two neighborhoods are both within Cluster 4 and they both have a low number of other restaurants and a high number of other venues.

The reason why I’m not recommending Northside, which is also in Cluster 4 and has no Italian Restaurants, is because of the high number of other Restaurants in the neighborhood.