Final Capstone Project (Week 2)

Martin Deal IBM Capstone Project

Table of Contents

Introduction and Business Problem

A theoretical residential home builder is looking to set up a new location in the Houston Texas Area. This analysis will help decide where in this area might be a good location to consider based on neighborhoods.

We will be analyzing neighborhoods in the Houston area to try and determine if some similarities and characteristics between them exists. Since the builder has not provided details about what market he is looking to serve (custom vs. high volume, luxury vs. not, etc.), an exact location cannot be determined. However, we can make observations about the different types of neighborhoods and present these back to the builder so they can make a final decision.

Data

Data is from several sources. The neighborhood information, with some real estate information, is from Houstonia Magazine. The article is tiled "Neighborhoods by the Numbers 2017" and was published in March 2017. It can be found at "https://www.houstoniamag.com/home-and-real-estate/2017/03/neighborhoods-by-the-numbers-real-estate-data-2017". The information will have median home pricing and home value growth by percent for specified time periods. The geocoordinates of the neighborhoods is derived by using their zip codes. Using the site zipinfo.com, I will download a csv file with all zip codes and corresponding latitude and longitude. Lastly, Foursquare will be used to pull in the venues data for the neighborhoods and this can be found at Foursquare.com.

The real estate data will be standardized and combined with the venue data to group the neighborhoods using **KMeans**. Once completed, analysis will be performed by mapping of the clusters of neighborhoods in the Houston area to visually see how they relate to each other. Also, a review of the top venues in each neighborhood cluster should provide observations about characteristics and similarities. Certain clusters may prove more suited for a new location versus others and this information will be provided to the homebuilder.

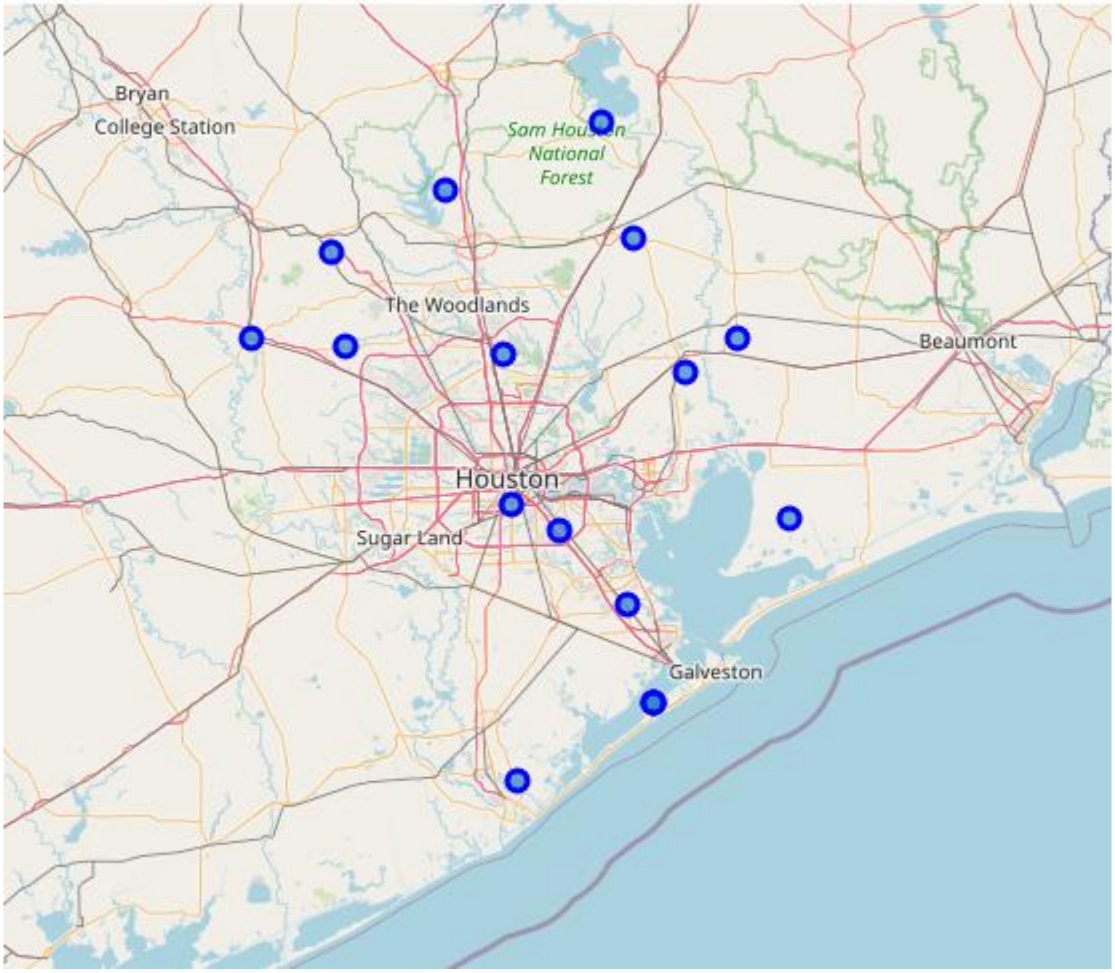
Data Cleaning

In the real estate data, I decided to drop the % owner occupied and Average days on market as I didn't think these to be relevant to the problem. After cleaning and combining the real estate and zip code data, I had a usable DataFrame. Now I could work on getting the Foursquare data ready.

Methodology

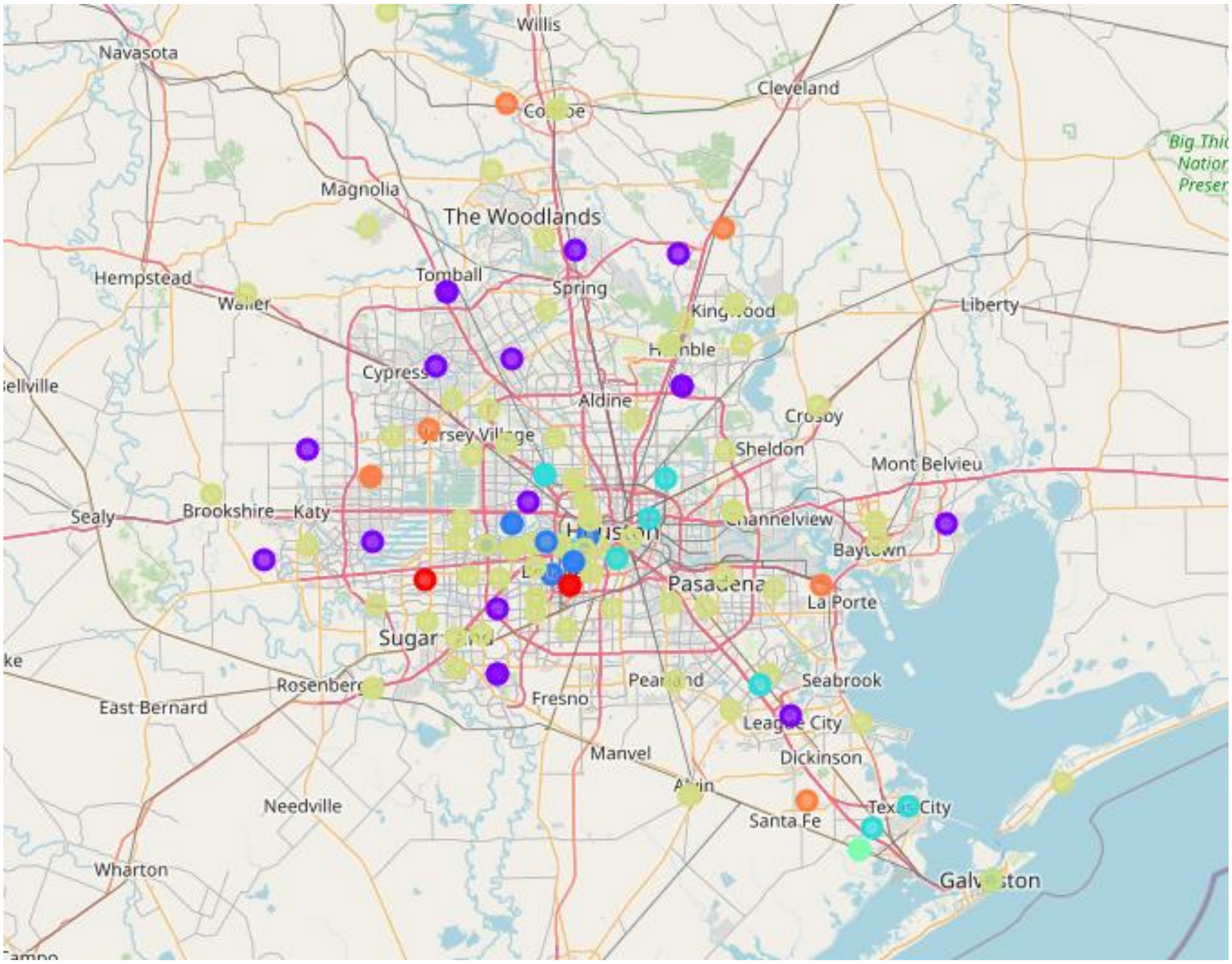
I decided to use KMeans to cluster the neighborhoods around Houston into similar groups. From this, we can look at the similarities and differences of the different clusters and decide what to pursue or not. In anticipation that there will be multiple areas, the maps will further help distinguish favorable areas to start more research for the builder.

Before merging the two DataFrames, I performed scaling on the real estate columns and then merged the neighborhood DataFrame with the Venues DataFrame. There were some neighborhoods with null venues. I plotted these to visually see where they were and what impact it may have on analysis.



These null neighborhoods were dropped and analysis continued.

The remaining neighborhoods were then mapped and analyzed.



Results and Discussion

I used 7 clusters in the KMeans model. There were distinct differences between the clusters and grouped the neighborhoods well. Below is a simple summary of each cluster with some observations and comments for the client.

Cluster 0: These neighborhoods are represented by the Red dots on the map. The most popular venue data for these is park or recreation center. The remaining venues are oriented toward residential customers. The home value appreciation for these neighborhoods were good, with all over 30% for the 2010 - 2016 period.

Cluster1: These neighborhoods are represented by Purple dots on the map. They are grouped due the the first or second most popular venue is oriented toward industrial customers. This indicates that many of the neighborhoods are located in or near an industrial area. The home appreciation for the period 2010 - 2016 ranged anywhere from just over 10% to over 100%.

Cluster 2: These neighborhoods are represented by dark Blue dots on the map. These are most of the luxury neighborhoods in the area with the average 2016 home price over $800,000 and all but 2 neighborhoods well over 1 million. The venue data confirms this.

Cluster 3: These neighborhoods are represented on the map by light Blue dots. These neighborhoods show strong growth over the 2010 - 2016 period. All are over 60% growth with the exception of one at 47%. Several had growth rates over 100%. The venue data shows that most of the popular are oriented toward residential customers.

Cluster4: These neighborhoods are represented by Teal dots on the map. There are only three neighborhoods in this group and they all share the same venue data. The most popular is a bar followed by Yoga studio and Doctor's office. This indicates that these neighborhoods are close to some type of retail center. The growth ranged from just under 2% to over 70% for period 2010 - 2016.

Cluster 5: These neighborhoods are represented by light Brown on the map. This is the largest group of neighborhoods. The growth for these has a large range from little growth to nearly 100%. The venue data, while diverse, does indicate that they are located near retail centers. Most of the these neighborhoods had moderate home values in 2016.

Cluster 6: These neighborhoods are represented by Orange dots on the map. Growth for all these neighborhoods was good with all over 30% and one over 100%. The venue data for all has home service as the most popular. All are located further out from the center of the metro area. From the data, I assume that many venues are oriented toward residential services.

My recommendation to the residential builder would be as follows. If they want to build luxury homes, start looking in the neighborhoods from cluster 2 as a start point. If they are looking to build moderate customer or planned neighborhoods, clusters 0, 3, 5, and 6 might be good areas to start looking. Cluster 3 are all located close to center of metro area and may be a good place to look for urban style living while clusters 0, 5 and 6 are generally further out and might be better for suburban living. Clusters 1 and 4 seem to be located near industrial or major retail area and should be investigated for further insight.

Conclusion

We were able to pull in data from several sources and run an analysis to try and determine a good location for a residential home builder to locate. We were able to analyze most of the neighborhoods by some real estate metrics and using Foursquare to present the top 10 venues in each neighborhood. Based on this analysis, we have provided the home builder with several neighborhoods that could be targets for further study, depending on type and style of homes they want to build.