Intro

Background

The housing market in the United States accounts for, on average, 15-18% of the GDP and this can be split into residential investment (3-5%) and spending on housing services (12-13%) (Housing's Contribution to GDP). It can be deduced that real estate brings a lot of value to the overall US market and that people are investing in real estate – whether it be by renting or home ownership. A problem that arises when thinking about renting or buying a home is, where to live? There are multiple factors that could make a location right for you but, how much time would you be willing to spend to gather all this information? The location is based on zip codes and the factors considered are median home value, crime, and venue types. The purpose is to cluster zip codes together based on crime and venue types to give insight on where a good place would be to live. Someone who would be directly interested in this project would be anyone looking to move to a new location. For the purposes of this project though, I will be focusing on the implementation of this for one of the largest cities in the US, Houston, TX.

Data Acquisition and Cleaning

Data Sources

The data was gathered from various websites. The zip codes for Houston were scraped from here, the median home value came from Zillow, the crime statistics came from the Houston Government, and the venue data was retrieved from Foursquare. The Foursquare API requires latitude and longitude data therefore, the GeoPy package in python was used to get longitude and latitude data for each zip code.

Data Cleaning

The crime statistics were downloaded as an Excel file and then imported as a DataFrame in python. All columns from the table were removed except for the zip code and description of the crime. The number of crimes for a given zip code was realized by counting the number of times a zip code appeared. There were some zip codes that were missing from the crime data and these zip codes were assigned the normalized mean crime count value.

The information that is relevant from Foursquare is the venue name, latitude, longitude, and venue category. This information is retrieved from the json output of Foursquare and put into a panda's DataFrame. To perform analysis on the venue types, pandas get_dummies function was used to create a DataFrame with an indicator variable of 1 instead of categorical variables. This made all the unique venue types a column and if the zip code has one of the venue types, a 1 would appear under the venue type.

Methodology

Initially, I completed the project without using crime data and I found that the insight gained was not entirely clear. After some research, it became evident that other data should be used to cluster zip

codes and crime would be a contributing factor to whether someone moves to an area. Also, I originally used normalized median home values when performing the clustering of zip codes and I found that this was not providing clusters as anticipated. I believe that when median home values were used as a feature, it became the dominant feature, and this is what was causing clustering issues. Therefore, the median home value is to be used as more of a filter than a feature in the dataset. The median home value is added back to the DataFrame after clustering is performed. Overall, this project itself is more of an exploratory data analysis feat. With the use of unsupervised learning techniques, I am looking to gain insight into how zip codes across a city differ.

The first analysis that was completed was creating a map that plotted all the zip codes, using latitude and longitude, in Houston using the folium package. This proved to be important because, originally, some latitude and longitude information gathered from GeoPy put zip codes in places that was not Houston, TX. I found that this was an issue with the "user_agent" that was used in the Nominatim() function and ultimately, I used the foursquare user_agent.

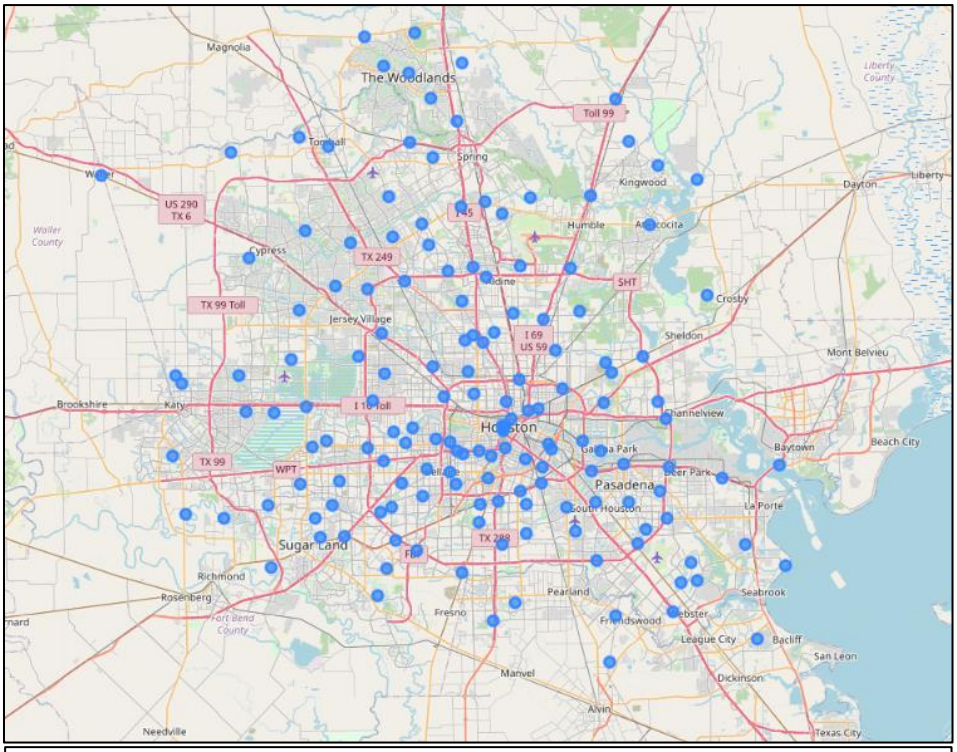

Figure 1. Zip Codes Plotted

The following analysis was to see the top 10 venues per zip code. This was done for two reasons, initially, it was done out of curiosity and then later it was found to be useful for classifying the different clusters.

	ZipCode	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	77002	Hotel	Burger Joint	Park	Gym	Bar	Fried Chicken Joint	Food Truck	Theater	Mexican Restaurant	Coffee Shop
1	77003	Bar	Hotel	Gym	Mexican Restaurant	Park	Lounge	Burger Joint	Theater	Pizza Place	Food Truck
2	77004	Nightclub	Vietnamese Restaurant	Bar	Coffee Shop	Lounge	Mexican Restaurant	Art Gallery	Beer Garden	Wine Bar	Pizza Place
3	77005	Mexican Restaurant	Burger Joint	Coffee Shop	Pizza Place	Italian Restaurant	Sandwich Place	Café	Spa	Gym	Bakery
4	77006	Zoo Exhibit	Science Museum	Bar	Coffee Shop	Art Museum	Art Gallery	Sculpture Garden	Trail	Food Truck	Garden

Figure 2. Top Ten Venue Categories

The KMeans clustering method was used because this method of clustering worked well for this dataset. Initially, the number of clusters was chosen randomly to see how the zip codes were being clustered. Then the inertia was used and plotted vs the number of clusters to determine what may be a decent cluster number for the dataset. There is also a way to determine the optimal k value by finding the maximum distance between the linear line, that joins k=1 and the maximum k, and the inertia curve. This was method of finding the optimal k-value was gained from a YouTube video from Bhavesh Bhatt. To visualize how the zip codes were being clustered, the folium package was used again where each cluster label has its own color marker. The Foursquare search radii that were examined were 500, 1500, and 2500 meters.

Results

It is important to note that there is no significance to the chosen colors and they only indicate cluster labels.

Radius of 500 Meters

There are 261 unique venue categories when using a search radius of 500 meters in the Foursquare API. The optimal k value is 6. The color-coded zip codes are plotted on an interactive folium map and can be seen below.

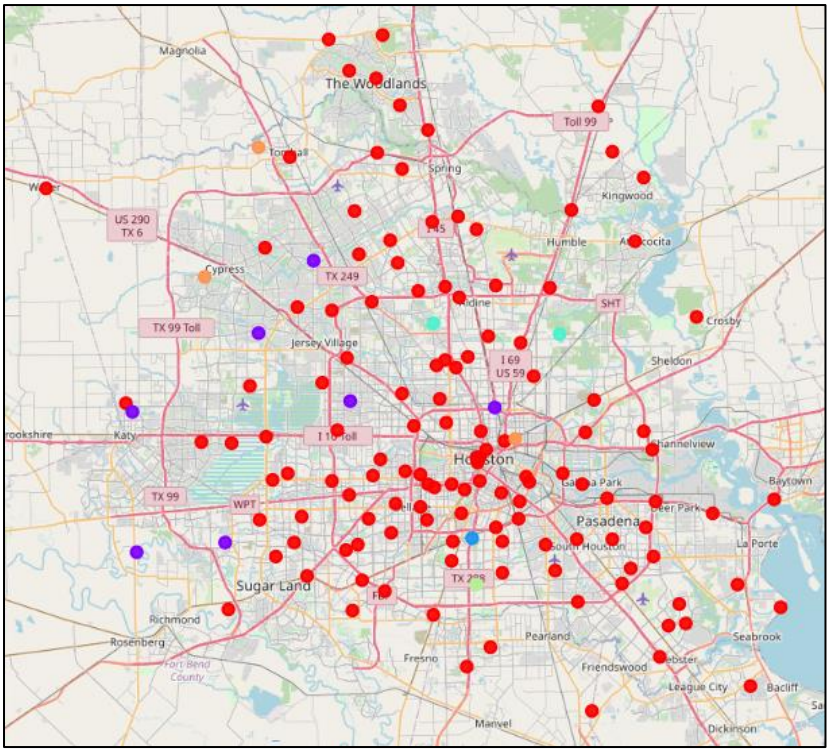

Figure 3. Clustered Zip Codes using Search Radius of 500 Meters

Radius of 1500

Using a search radius of 1500 meters in the Foursquare API, there are 368 unique venue categories. The optimal k value is 7. The color-coded zip codes are plotted and can be seen below.

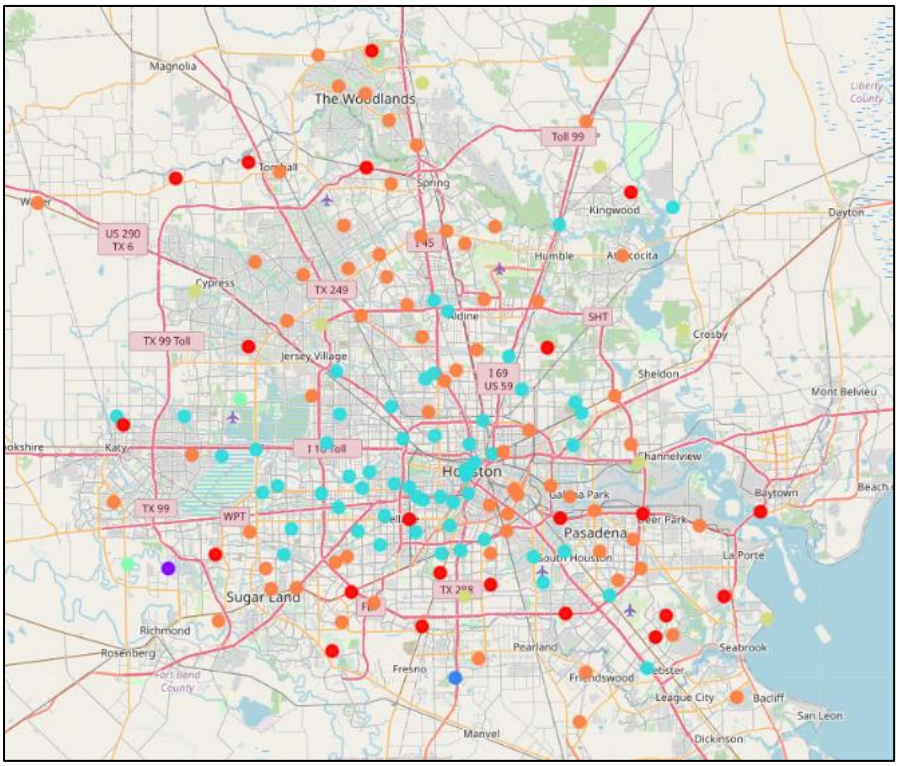

Figure 4. Clustered Zip Codes for Radius of 1500 Meters

Radius of 2500

There are 391 unique categories when using a search radius of 2500 meters in the Foursquare API. The optimal k value is 8. The color-coded zip codes are plotted on an interactive folium map and can be seen below.

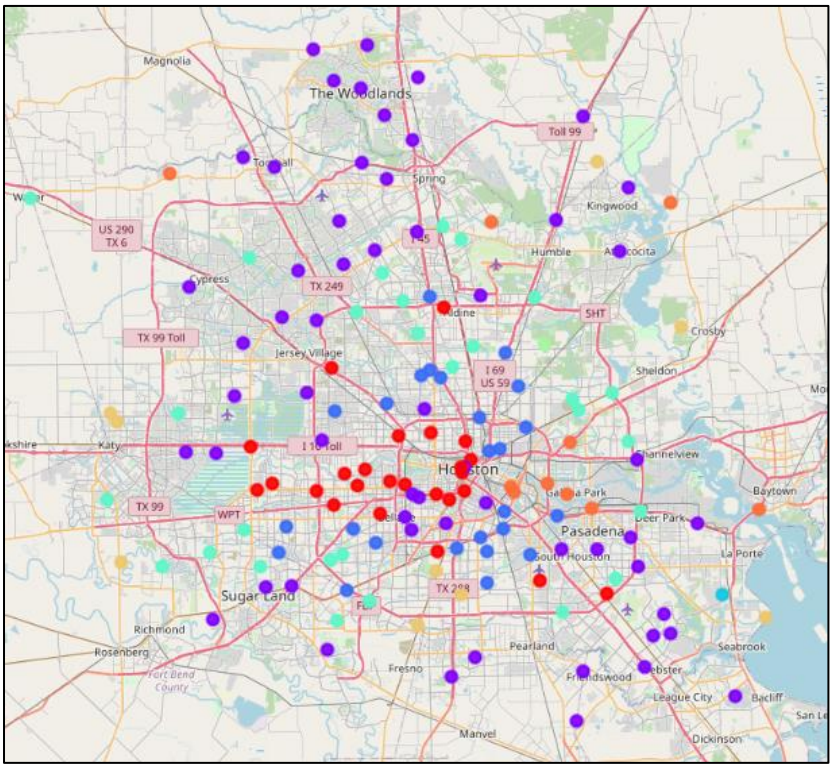

Figure 5.Clustered Zip Codes for Radius of 2500 Meters

Discussion

From looking at the information in the results section and examining the images, we can tell that the search radius used in the Foursquare call has a significant effect on how the clusters are formed. For example, in Figure 3, radius of 500 meters, there does not seem to be definitive clusters. The red cluster is by far the majority for the entire city and that is not interesting. Expanding the search radius to 1000 meters yields more engaging results. At this radius, zip codes appear to form more distinct properties. The best results for this research appear when using a radius of 2500 meters. When looking at Figure 5, the red zip codes signify inner city type environment and then, the more interesting part of this is that all the other clusters signify different types of suburb environments.

One of the most intriguing things I found during this is that low-income areas can be identified by the venue categories in the zip codes. Although this information is not exactly all that surprising, it still indicates how companies manage to keep people poor. Some of the top venue categories that appear in the dark blue and teal cluster labels in Figure 5 are fast food restaurants, discount stores, and gas stations. The way that I determined these were low-income areas is by using the median home value. I assumed that a low median home value would be equivalent to a low-income area. Although this assumption is not perfect, it fits for the purposes of this research.

Conclusion

In this study, the zip codes of Houston, TX were clustered using venue category information from Foursquare and crime data from the Houston government website. The median house cost can be used to filter out zip codes based on the house cost. This insight can be used to assist in finding the best place to purchase a house based on your venue preferences and budget.