1. Introduction

In one of my Coursera courses managed by IBM I was required to post a blog-post regarding my capstone project. This work is part of that certification.

Capstone project required that we include Foursquare API to complete this project. We had done some homework on given data for New York, USA and Toronto, Canada and this data for both cities was readily available. Since I wanted to explore more I decided that I would prefer to collect my own data and modify it as per my needs. This became a big challenge in itself as I figured that not all the data is coherently available for most projects. I had to assume (logically) things in order to complete/discard available data to make sense or fill any gaps in data.

As I currently reside in Houston, Teaxs, USA and it is a prominent metropolitan city I considered it for my project. What caught my attention during this research is [1]:

Houston is 4th largest city in USA.
Houston recently overtook New York as most ethnically and racially diverse city in USA.
Houston is top market for job creation in USA.
Housing is affordable compared to other big cities.
Houston's unemployment rate is far below the national level.
Houston is one of the centers of America's oil & gas industry.

At the end of this exercise I would like to visualize maps which will give price range and population range of zipcodes so some one moving into Houston can decide where he/she would like to buy a home.

2. Data

As I started to search and collect relevant data, I realized that like New York and Toronto, Houston is not divided by boroughs. It has management districts and super neighborhoods. Unfortunately most of the census data is not available based on these divisions and made my data collection extremely difficult. I finally settled on zipcode as an element on which I collected my data.

A) I collected map data from city of Houston website which consisted 213 zipcodes. I did not change this data as I decided to have a representation of complete greater Houston when drawing a map for greater Houston area. I modified remaining data for Houston only zipcodes.[2]

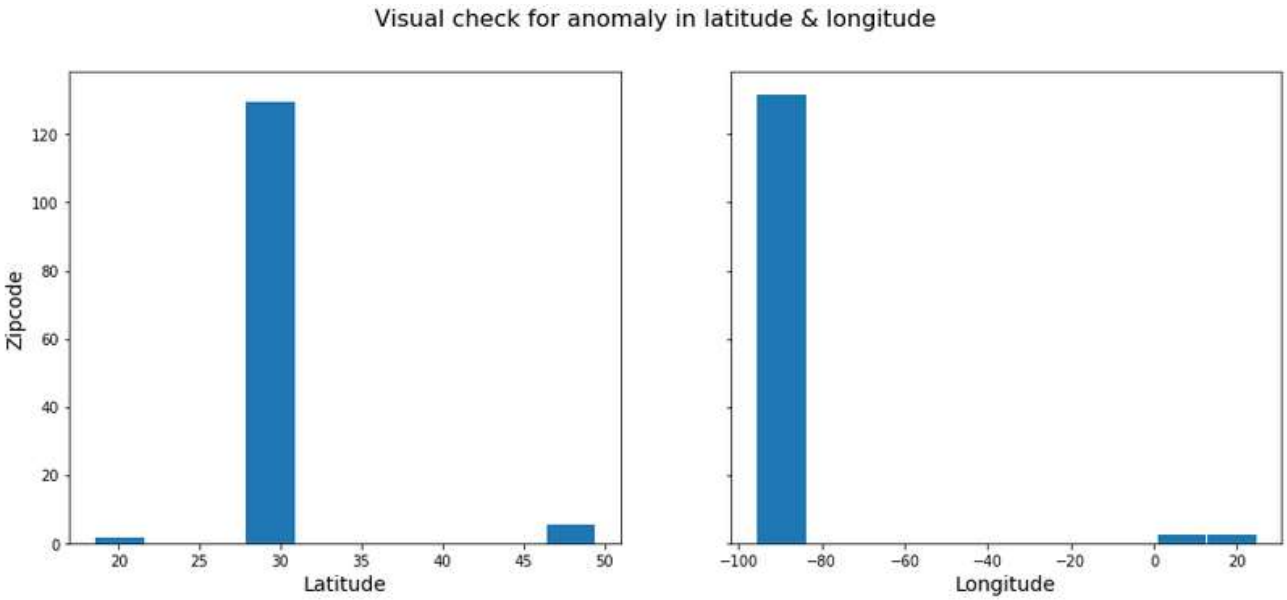
B) I wanted to include a glimpse of lifestyle for the selected zipcodes. As it is very difficult to agree on a representative factor for that I eventually decided to use median price per sqft data from house sales data. [3]

C) I also collected population data for all zipcodes. I wanted to create a choropleth map of Houston which also had relevant data for each zipcode after analysis. [4]

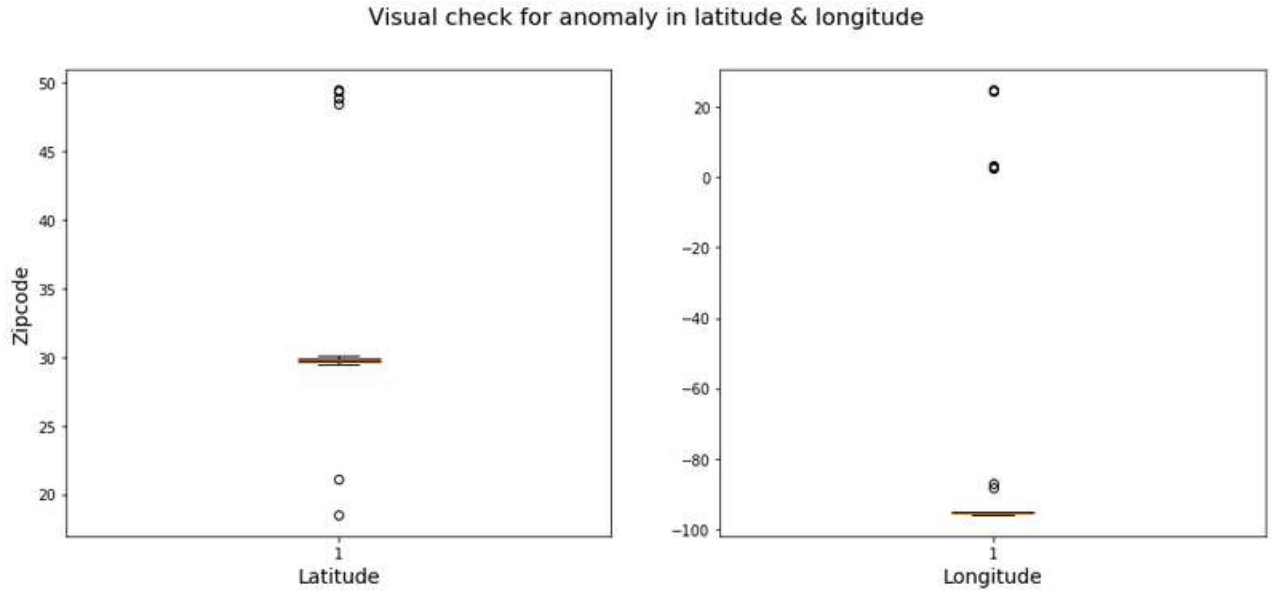
D) As required I used Foursquare API to explore venues around my zipcode locations and I used functions from lab assignments. For the zipcodes which got wrong latitude and longitude using geocoder, I used google maps "Nearby" search option to get correct latitude and longitude.

3. Methodology:

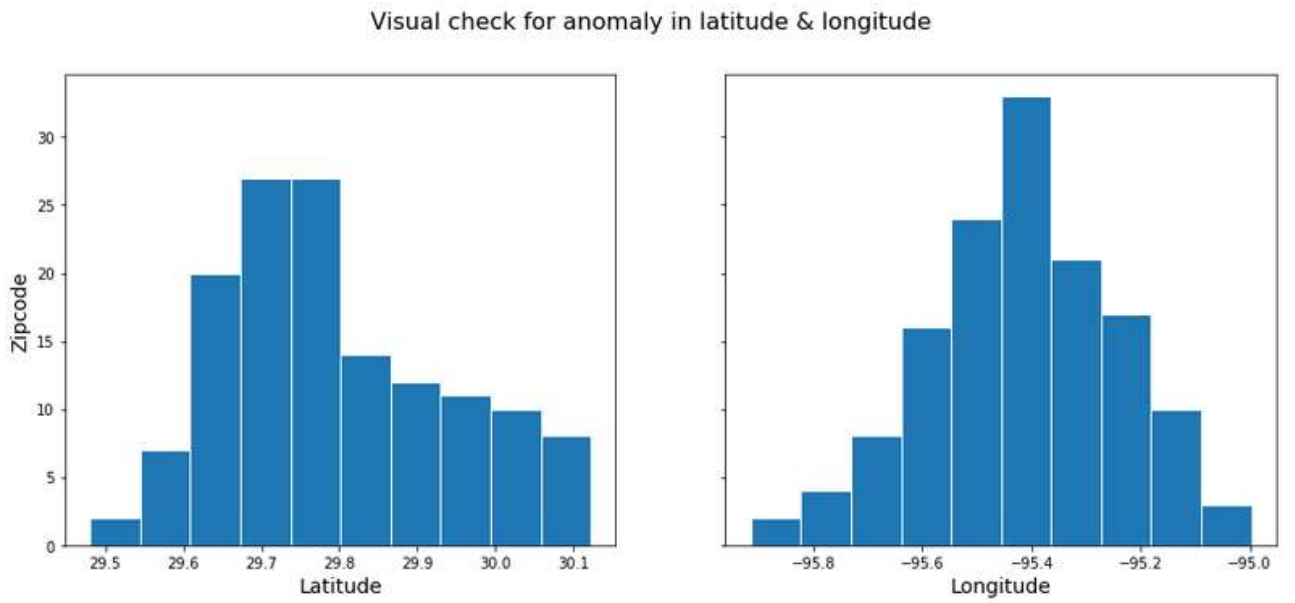
I used jupyter notebook for this project. My first task was to check my data for outliers in zipcode location data.

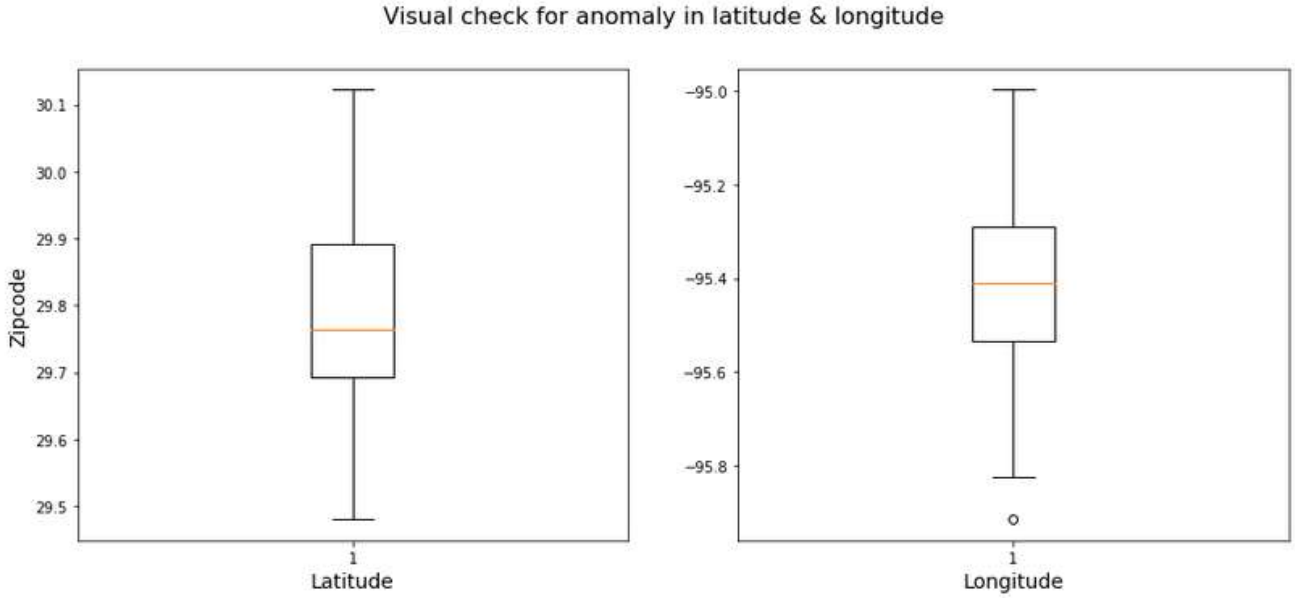


As I was working on one city, I was not expecting this variation in latitude and longitude data.



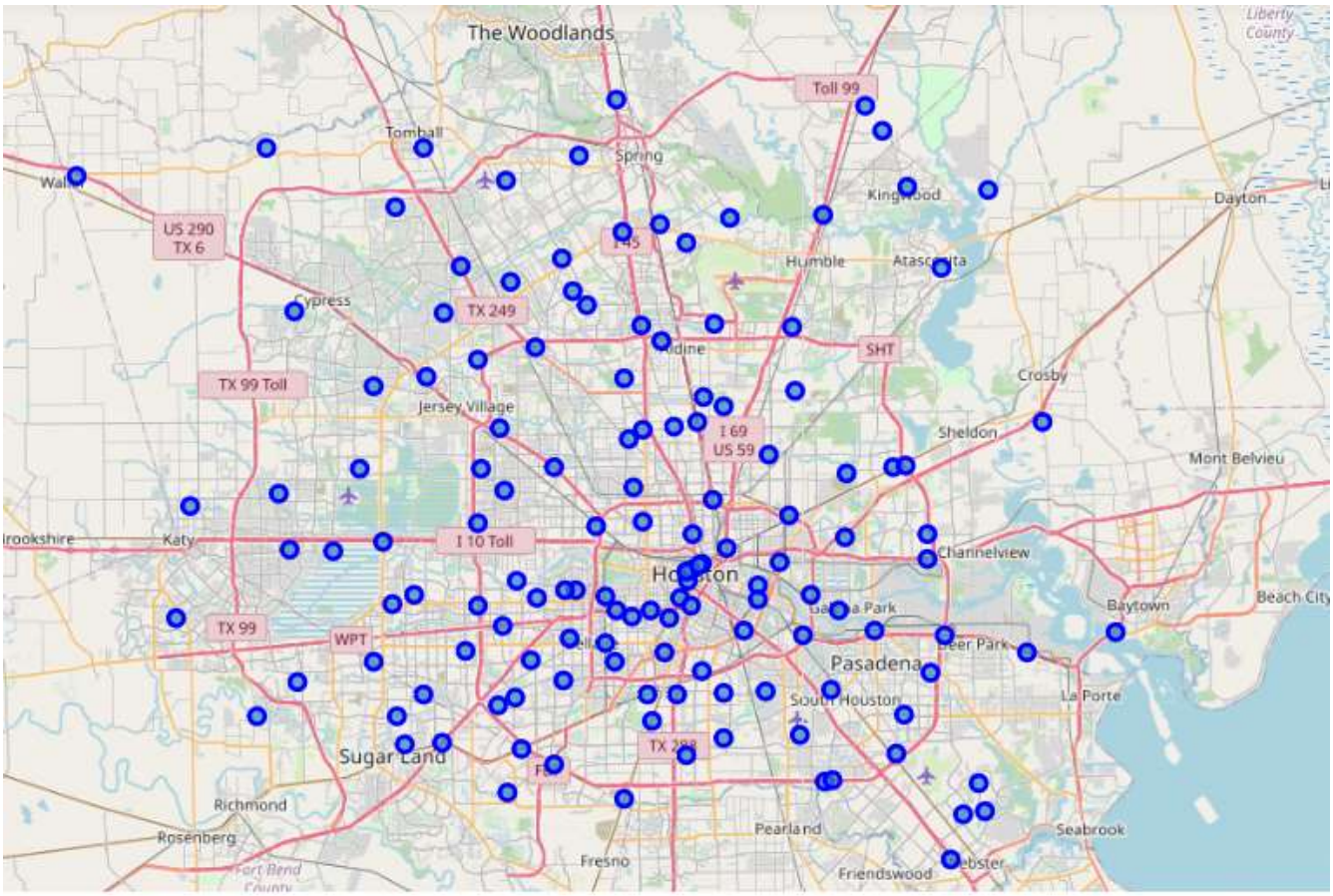
Above box plot confirms that I have few outliers. I collected latitude and longitude of these outlier zipcodes from google maps using nearby search option and replaced outliers with them. Below is the new graphs for comparison.



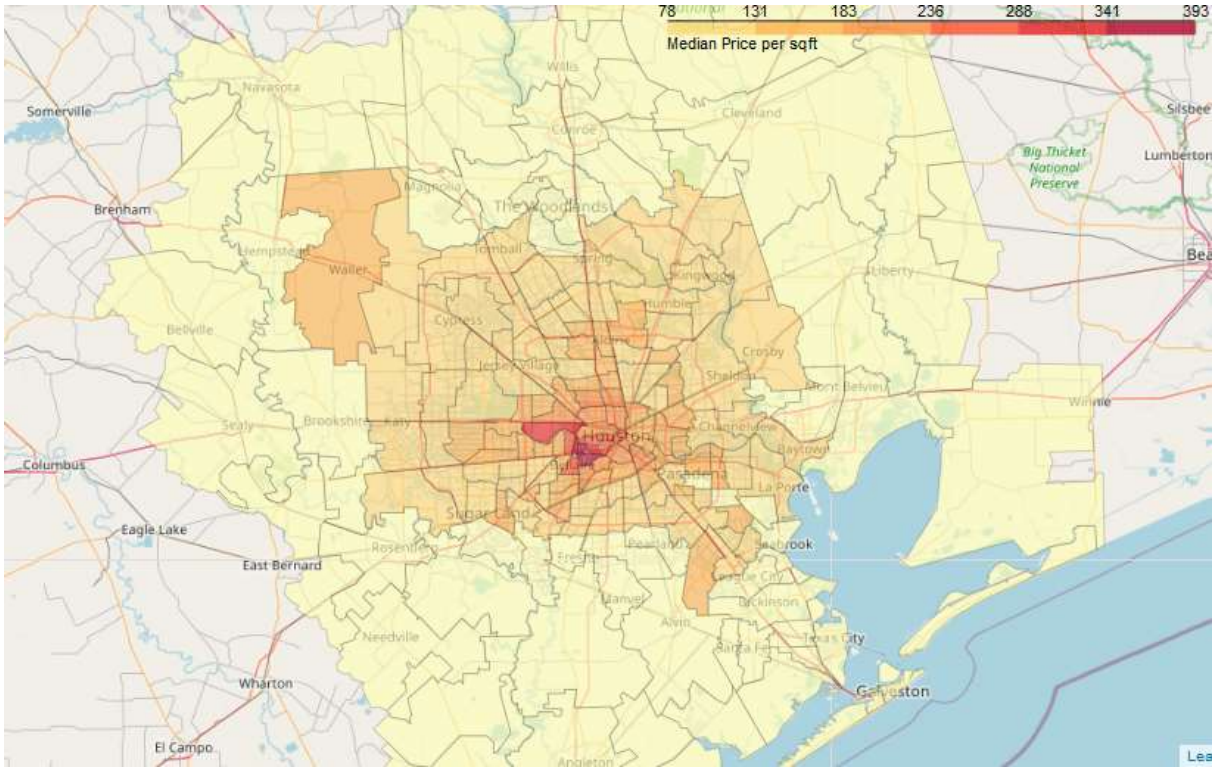


Looking closely we can see that now the spread is very tight and on expected lines as we expect city zipcodes latitudes and longitudes to be very close.

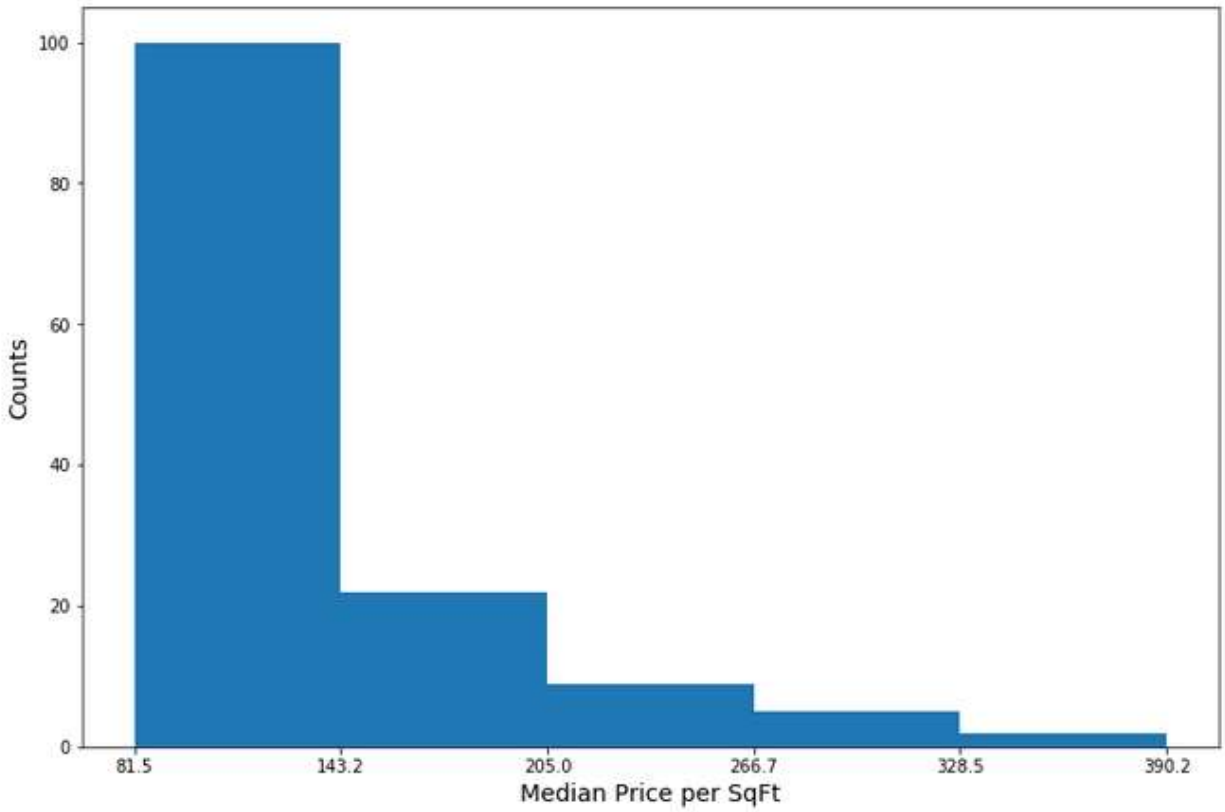
And this is how zipcodes location looks on the map. Here I used folium library to generate this map. Zipcodes are superimposed using latitude and longitude retrieved earlier.



I also collected median price per sqft data from home sales to get an idea of kind of neighborhood it is. Below is a choropleth map of this data.



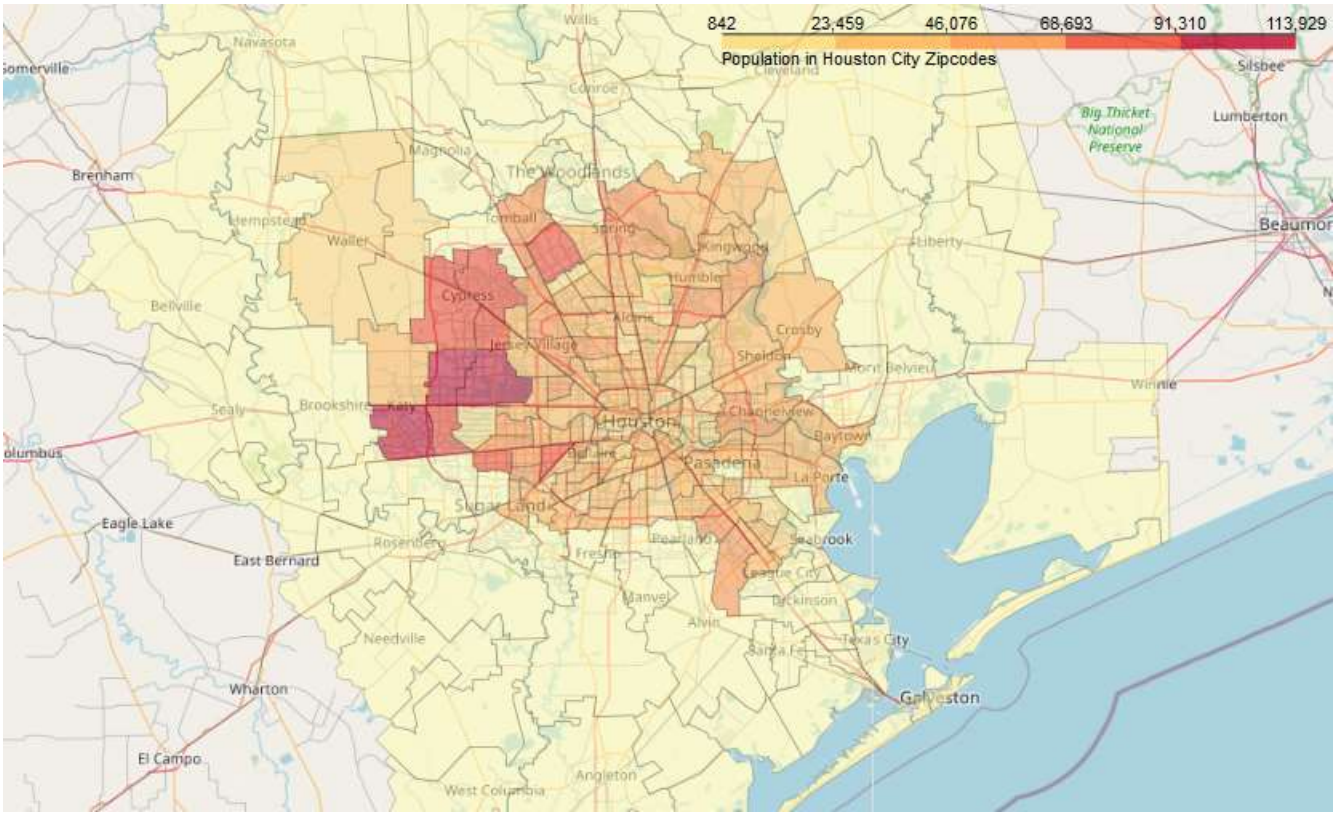
I further examined price data to define 3 ranges for price data by visualizing it as a histogram. Frequency of price data helps us decide our ranges.



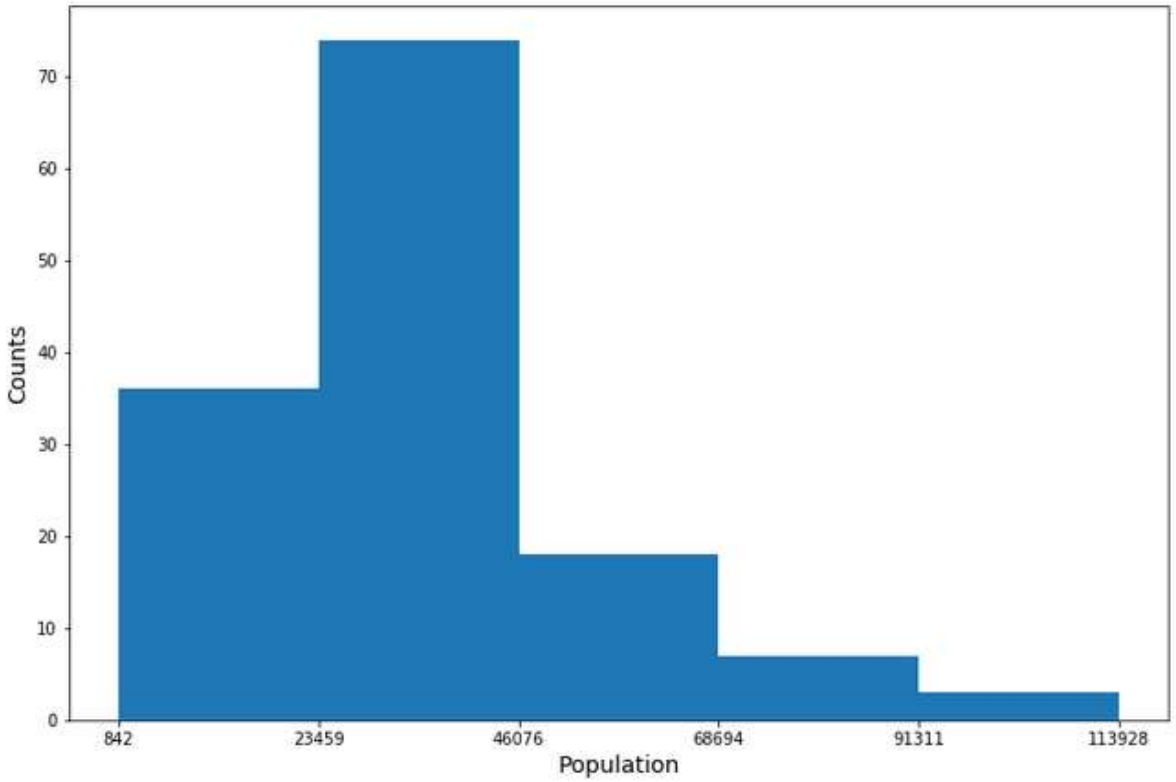
I decided to merge last 3 bins as 'high range' as there were not too many data points in those 3 bins. Below are all ranges I defined:

Low Price Range (81.5 to 143.2)
Medium Price Range (143.2 to 205)
High Price Range (205 to 390.2)

Similarly I looked into population data for all zipcodes.



I further examined population data to define 4 ranges for price data. By visualizing it as another histogram. Here also we can decide our ranges with help of frequency of data.



I decided to merge last 2 bins as 'high range' as there were not too many data points in those 2 bins. Below are all ranges I defined:

Low Population
Medium Low Population
Medium High Population
High Population

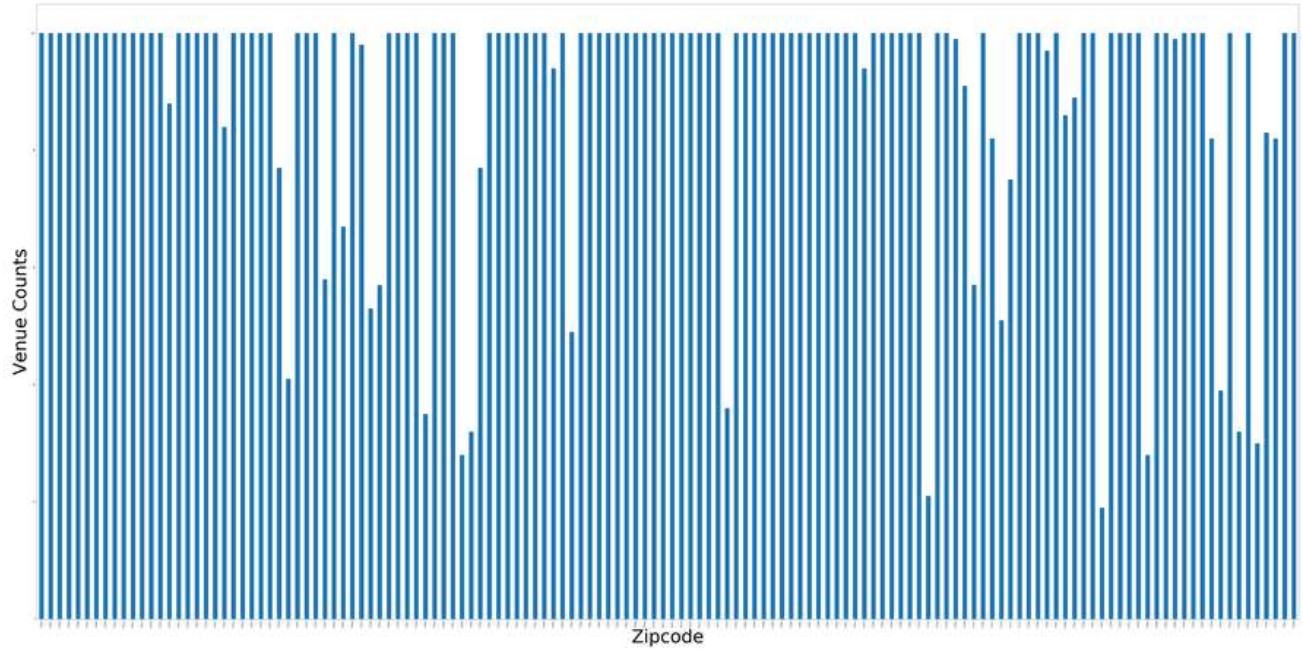
After combining all datasets my data-frame consisted of zipcode, latitude, longitude, price range & population range.

	ZipCode	Latitude	Longitude	Price Range	Population Range	MedianPriceSF	Population
0	77002	29.740709	-95.381591	High Price Range	Low Population	244.672455	12031
1	77003	29.754109	-95.374487	Medium Price Range	Low Population	197.732754	10241
2	77004	29.734089	-95.371666	Medium Price Range	Medium Low Population	190.075321	37407
3	77005	29.726717	-95.425076	High Price Range	Medium Low Population	388.503194	28104
4	77006	29.724805	-95.391550	High Price Range	Low Population	240.759593	22162
5	77007	29.766289	-95.362096	High Price Range	Medium Low Population	210.180180	35689
6	77008	29.799128	-95.414949	High Price Range	Medium Low Population	218.165221	33055
7	77009	29.789324	-95.371431	High Price Range	Medium Low Population	225.451467	38564
8	77010	29.761664	-95.372057	Low Price Range	Low Population	142.463798	842
9	77011	29.750646	-95.312696	Medium Price Range	Low Population	152.329749	18824
10	77012	29.739085	-95.313902	Low Price Range	Low Population	142.463798	20828

I segmented zipcodes utilizing FourSquare API. I explored all zipcodes utilizing their latitude and longitudes. I used a limit of 5000 venue and radius of 5000 meter for each zipcode (at the end it really did not matter). I wanted to get as many venues as possible from one zipcode but limiting factor is Foursquare account type which decides how many venues you can get for a given location. For this reason at the end of this exercise I had maximum of only 100 venues for one location after exploring them using FourSquare API. Below is a snapshot of venue details from Foursquare API. In total 12,510 venue were fetched by this API for all zipcodes.

	ZipCode	Zip Latitude	Zip Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	77002	29.740709	-95.381591	The Breakfast Klub	29.738440	-95.380558	Breakfast Spot
1	77002	29.740709	-95.381591	Tacos A Go-Go	29.738405	-95.379805	Taco Place
2	77002	29.740709	-95.381591	Jinya Ramen Bar	29.742754	-95.379648	Noodle House
3	77002	29.740709	-95.381591	Beer Market Co.	29.741913	-95.379609	Beer Garden
4	77002	29.740709	-95.381591	Holman Draft Hall	29.740245	-95.379784	Beer Bar

Below table shows that irrespective of our venue limit and radius defined in the function maximum venue data will be only 100 per zipcode. Also if we use a lower limit and radius we might get a less number of venues as they become controlling factor in place of FourSquare API account type. In this analysis In this case I got 343 unique venue categories.



Because of my choice of a large limit for venue most zipcodes were able to reach a venue count of 100. It is worth noting that in spite of that few zipcodes returned with a relatively low number of venue count. Upon a closer look in few of these zipcodes reaveals that not all the time latitude and longitude returned by geocoder is representative of these zipcodes for our purpose. Sometime these latitude and longitude are located in very less populated area, in creeks, lake or in a plantation area. This explains a low venue count for following zipcodes. (77029, 77038, 77044,77048,77049,77061, 77078, 77336, 77375, 77447,77484, 77520, 77532, 77546). This can be remedied by individually selecting a latitude and longitude that is a better representation of populated areas of a zipcode.

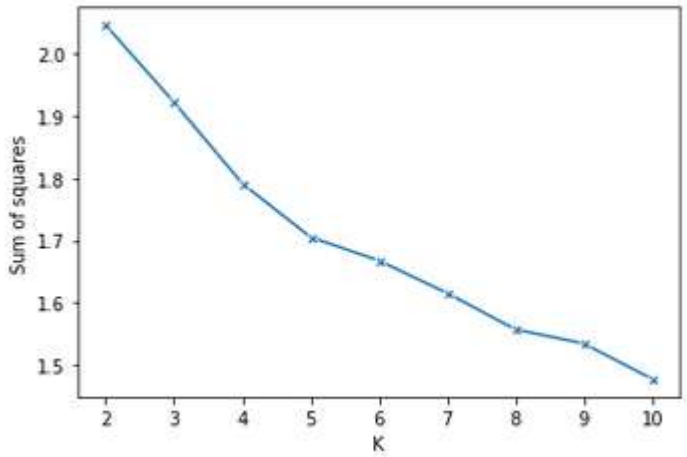
ZipCode	Zip Latitude	Zip Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
77002	100	100	100	100	100	100
77003	100	100	100	100	100	100
77004	100	100	100	100	100	100
77005	100	100	100	100	100	100
77006	100	100	100	100	100	100
77007	100	100	100	100	100	100
77008	100	100	100	100	100	100
77009	100	100	100	100	100	100
77010	100	100	100	100	100	100

Below is a table showing top 10 most common venue for zipcodes. We can see that zipcodes have some common venue categories.

	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	Coffee Shop	Park	Science Museum	Hotel	Breakfast Spot	Vietnamese Restaurant	Bar	Pizza Place	Theater	Beer Bar
1	77003	Coffee Shop	Park	Vietnamese Restaurant	Bar	Beer Garden	Pizza Place	Hotel	Wine Bar	Taco Place	Breakfast Spot
2	77004	Coffee Shop	Park	Hotel	Vietnamese Restaurant	Beer Garden	Bar	Science Museum	Pizza Place	Breakfast Spot	American Restaurant
3	77005	Café	Grocery Store	Italian Restaurant	Mexican Restaurant	Ice Cream Shop	Seafood Restaurant	Food Truck	Sushi Restaurant	American Restaurant	Indian Restaurant
4	77006	Coffee Shop	Zoo Exhibit	Café	Science Museum	Breakfast Spot	Wine Bar	Trail	Italian Restaurant	American Restaurant	Bar

Segmentation:

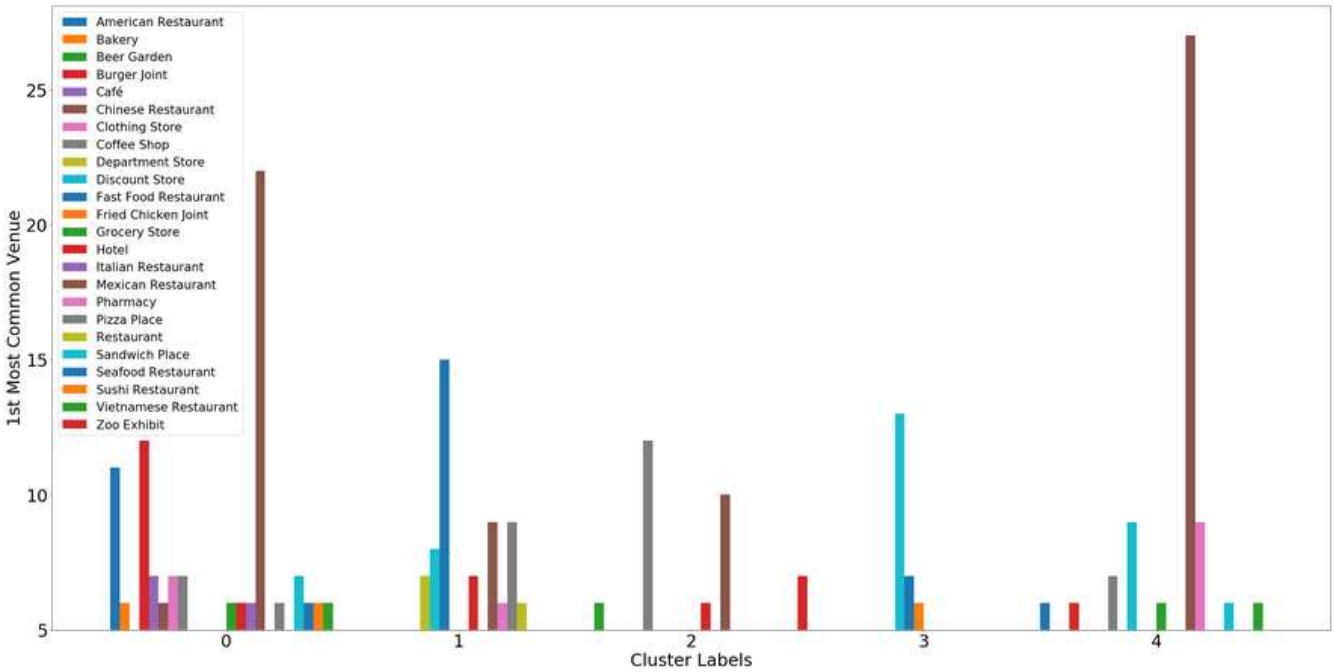
I decided to use K-means clustering to segment my zipcodes based on 1st most common Venue. K-means algorithm is one of the most common technique of unsupervised learning to divide the data into non-overlapping clusters. After analyzing sum of squares as seen below I decided to use 5 clusters for my dataset.



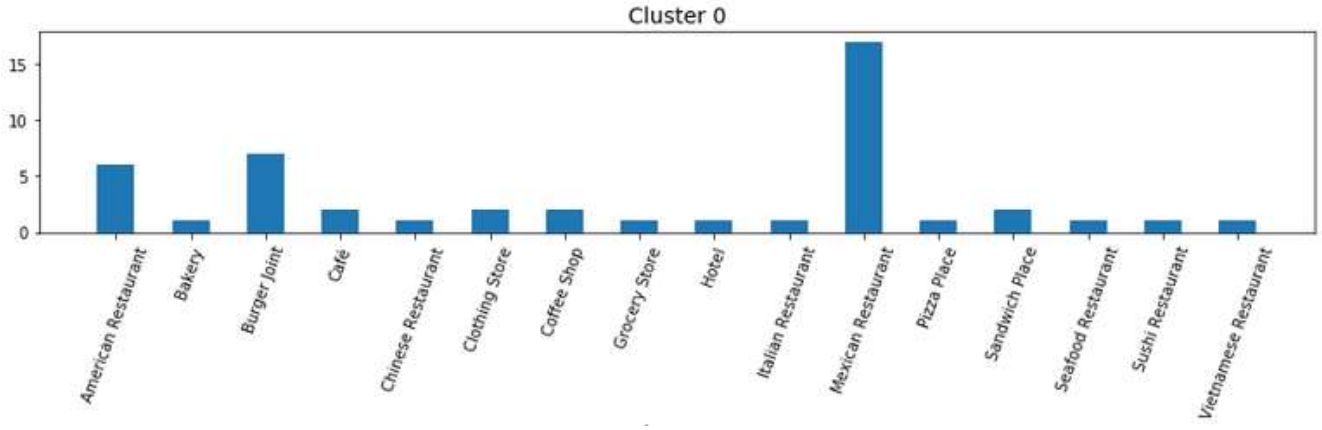
Below is resulting dataframe with cluster labels:

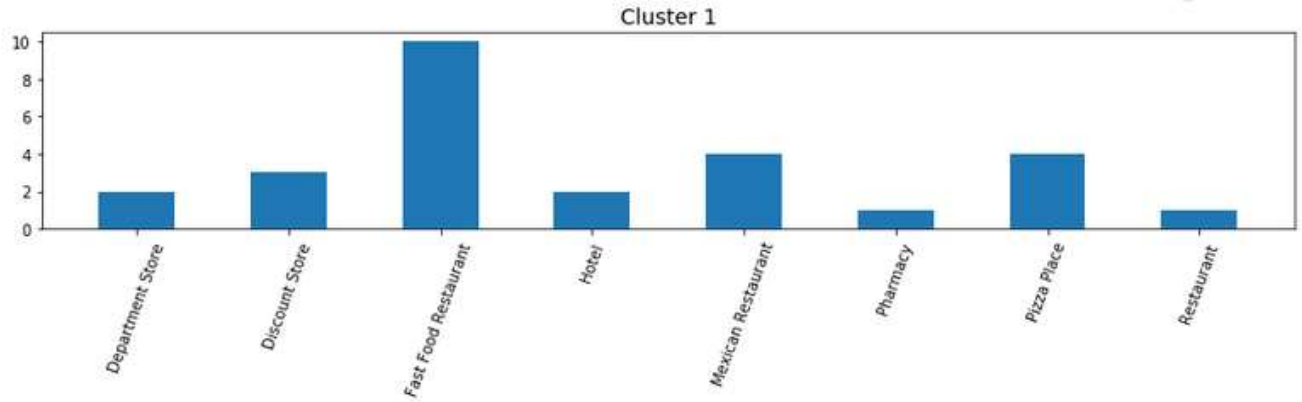
	ZipCode	Latitude	Longitude	Price Range	Population Range	MedianPriceSF	Population	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th I Com Vi
0	77002	29.740709	-95.381591	High Price Range	Low Population	244.672455	12031	2	Coffee Shop	Park	Science Museum	Hotel	Breakfast Spot	Vietna Resta
1	77003	29.754109	-95.374487	Medium Price Range	Low Population	197.732754	10241	2	Coffee Shop	Park	Vietnamese Restaurant	Bar	Beer Garden	Pizza
2	77004	29.734089	-95.371666	Medium Price Range	Medium Low Population	190.075321	37407	2	Coffee Shop	Park	Hotel	Vietnamese Restaurant	Beer Garden	
3	77005	29.726717	-95.425076	High Price Range	Medium Low Population	388.503194	28104	0	Café	Grocery Store	Italian Restaurant	Mexican Restaurant	Ice Cream Shop	Sea Resta
4	77006	29.724805	-95.391550	High Price Range	Low Population	240.759593	22162	2	Coffee Shop	Zoo Exhibit	Café	Science Museum	Breakfast Spot	Win

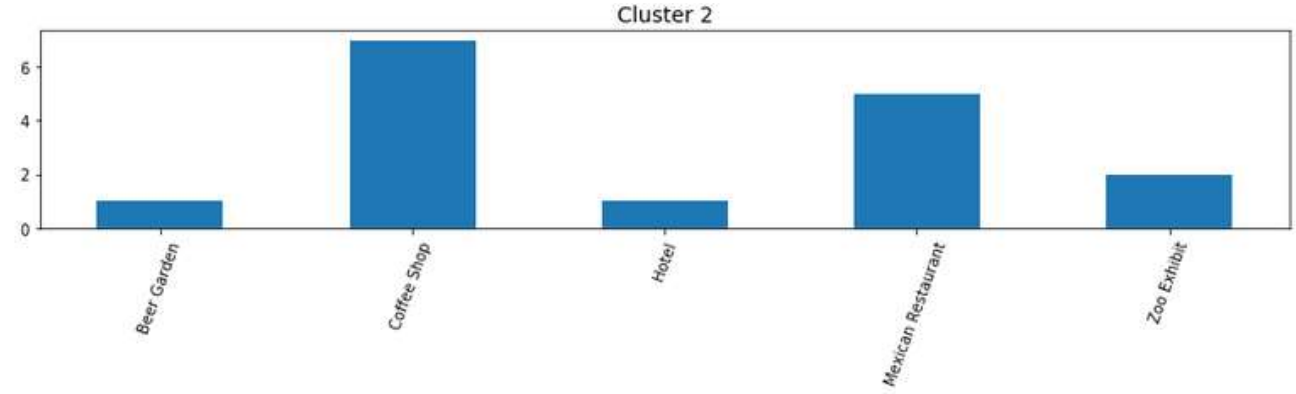
In order to be able to decide right descriptive label for all clusters I decided to create a bar chart. Below is the chart:

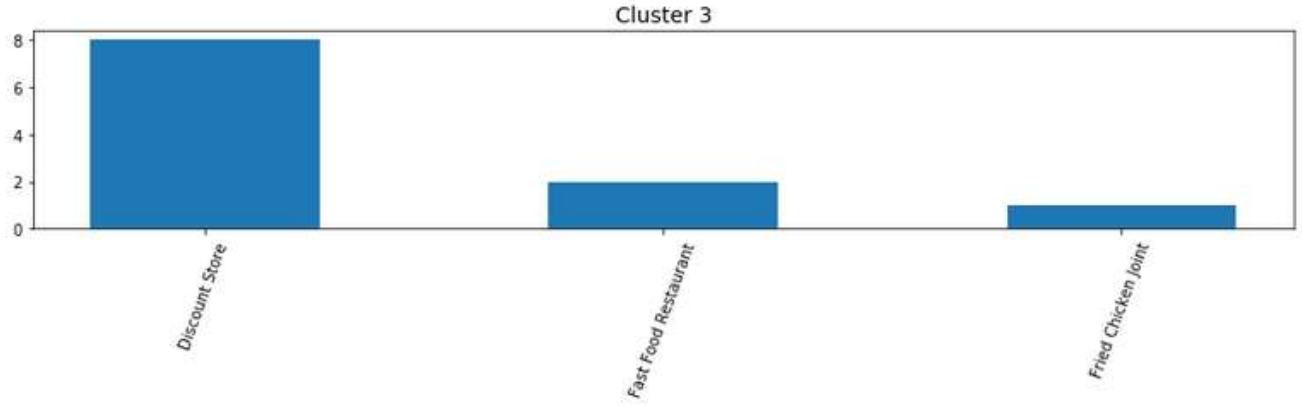


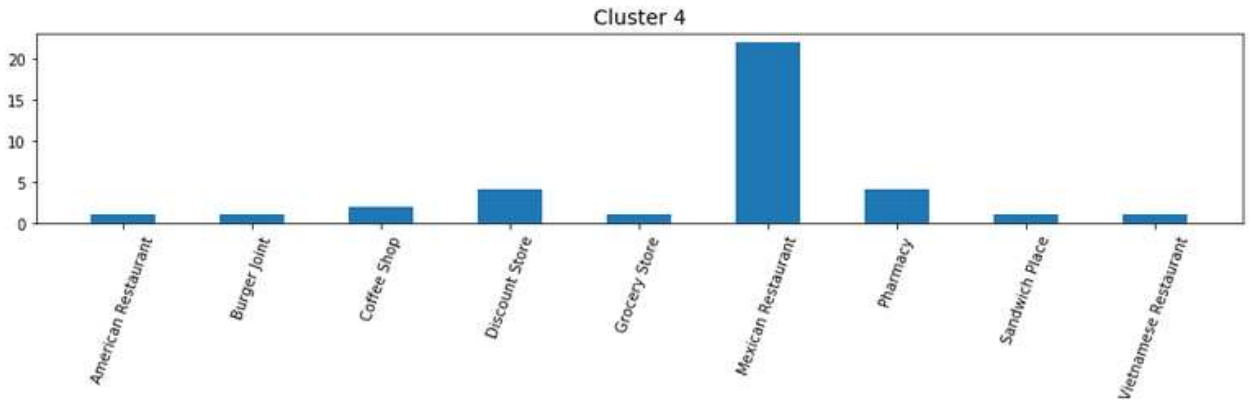
Because of less uniform spread of venue categories in above bar chart, I decided to create individual bar chart for each cluster. This helped me to look more closely into all clusters and give them a proper description. Below are individual bar charts for each cluster.











Based on these charts we can describe our clusters as follows:

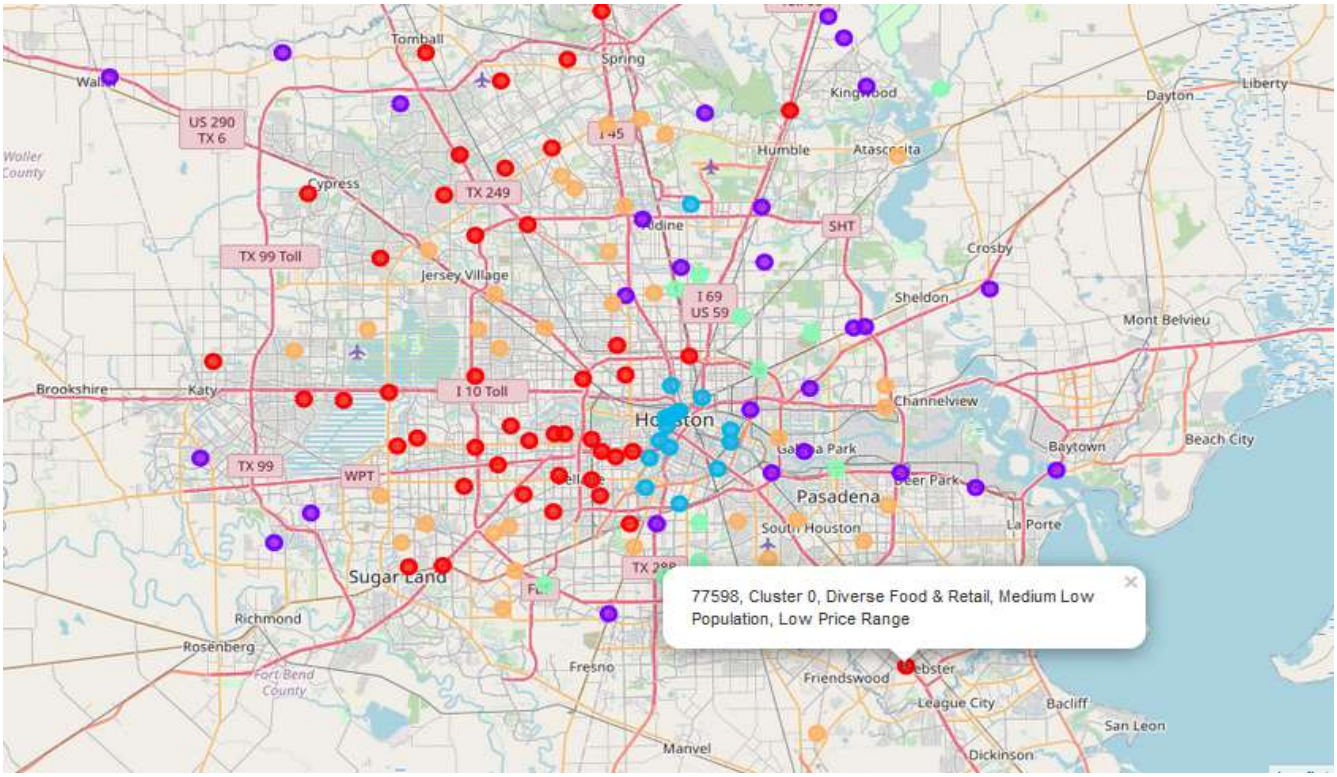
Cluster 0: Diverse Food & Retail
Cluster 1: Fast Food & Malls
Cluster 2: Social Venue
Cluster 3: Discount Store
Cluster 4: Mexican Food & Pharmacy

Below is the final outcome in the form of a table showing cluster description at the end

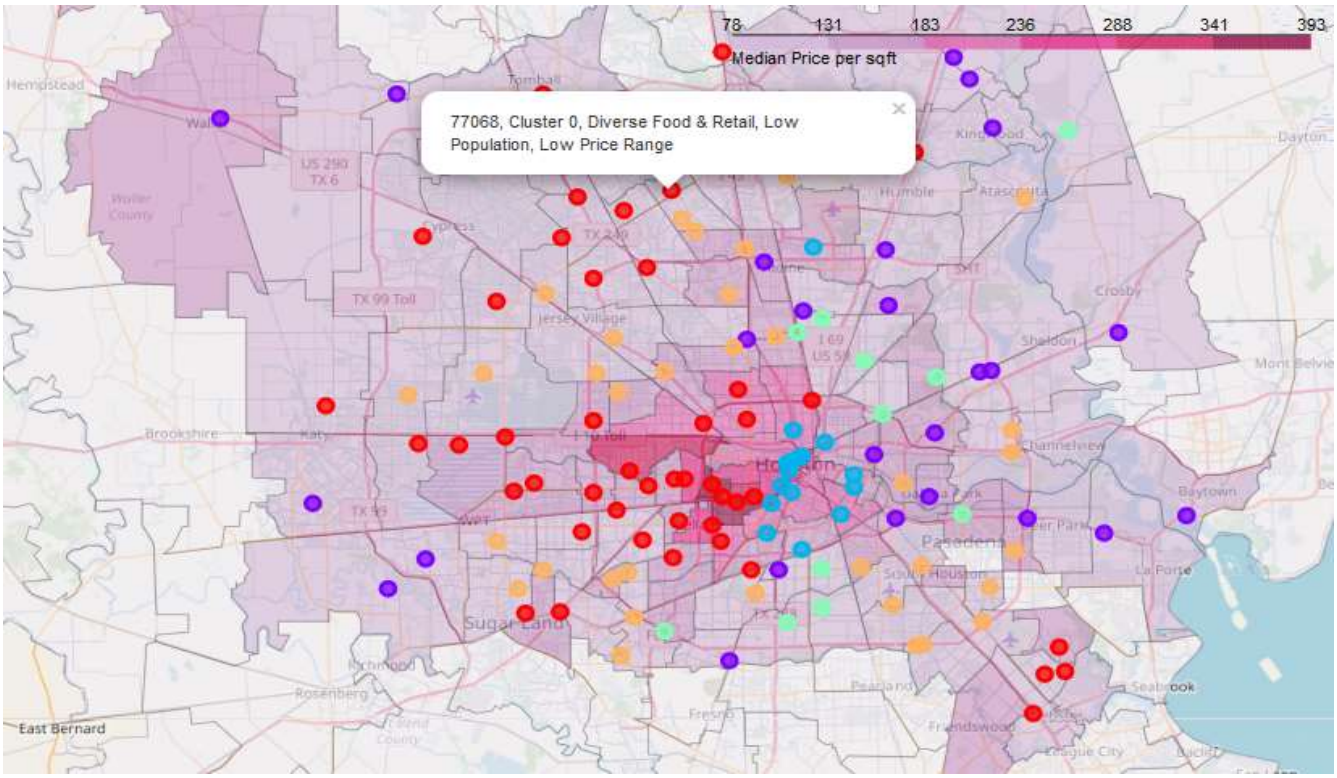
ınPrice SF	Population	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Cluster Description
44.672455	12031	2	Coffee Shop	Park	Science Museum	Hotel	Breakfast Spot	Vietnamese Restaurant	Bar	Pizza Place	Theater	Beer Bar	Social Venues
97.732754	10241	2	Coffee Shop	Park	Vietnamese Restaurant	Bar	Beer Garden	Pizza Place	Hotel	Wine Bar	Taco Place	Breakfast Spot	Social Venues
90.075321	37407	2	Coffee Shop	Park	Hotel	Vietnamese Restaurant	Beer Garden	Bar	Science Museum	Pizza Place	Breakfast Spot	American Restaurant	Social Venues
88.503194	28104	0	Café	Grocery Store	Italian Restaurant	Mexican Restaurant	Ice Cream Shop	Seafood Restaurant	Food Truck	Sushi Restaurant	American Restaurant	Indian Restaurant	Diverse Food & Retail
40.759593	22162	2	Coffee Shop	Zoo Exhibit	Café	Science Museum	Breakfast Spot	Wine Bar	Trail	Italian Restaurant	American Restaurant	Bar	Social Venues

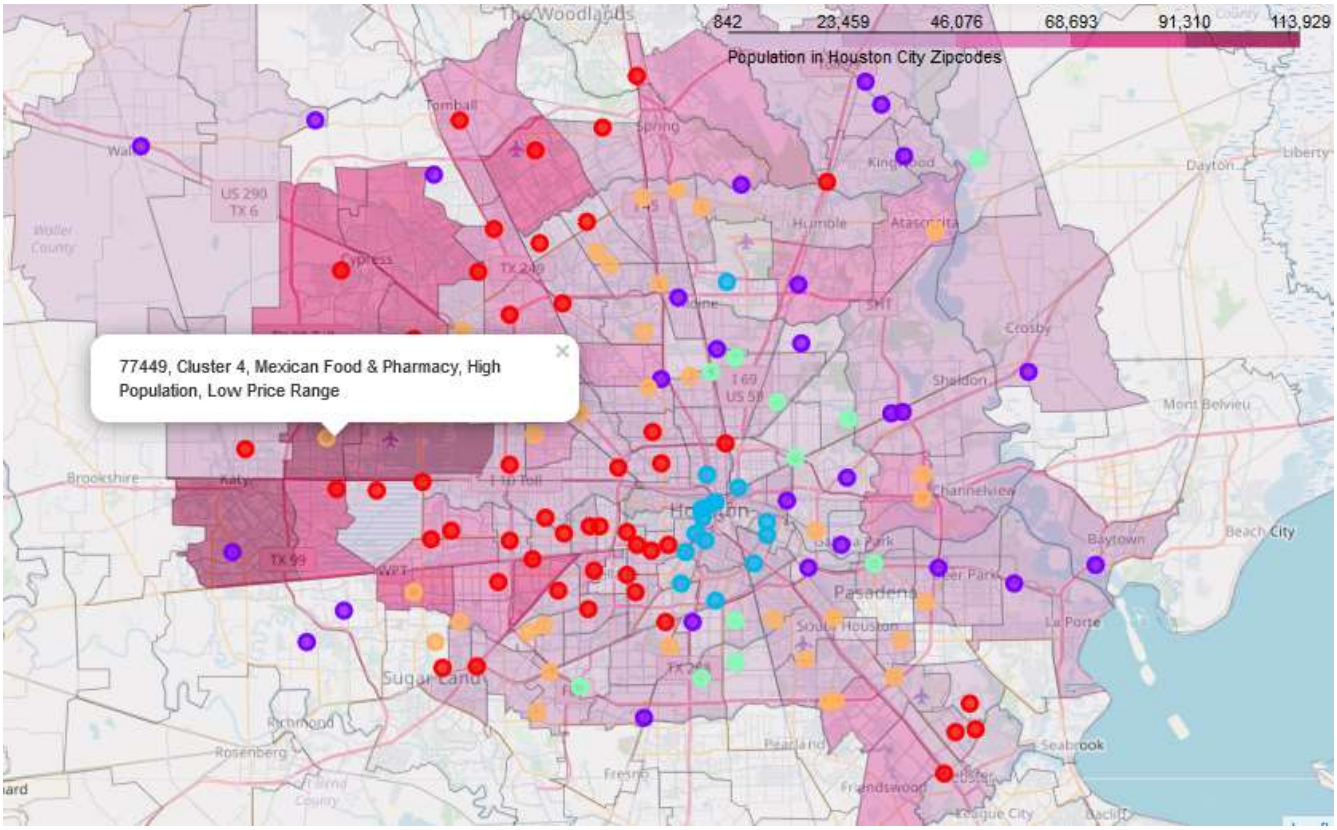
4. Results:

These cluster with their labels can be visualized in a map as seen below. Label shows zipcode, cluster number, cluster description, population range and price range for that particular zipcode.



One of my objective was to to visualize my clusters on choropleth map for price and population data. Below are two choropleth maps which also show label showing zipcode, cluster number, cluster description, population range and price range for that particular zipcode.





5. Discussion:

Houston is very big city and it is a complex city to properly divide it in a defined way as it is not divided in boroughs. It is divided in super neighborhoods and administrative districts but census data is hardly available based on these divisions.

I used population and median price per sqft from house sales data to give insight about lifestyle of a zipcode. This is a very subjective matter and can be handled differently by individuals. If we can develop a defined set of parameters to look into it we can have a much better qualitative analysis on these zipcodes. Also, in this project even though I read and collected data from different websites, I saved this data to a file and utilized it further. Hence my data collection is not real time/dynamic. This can be incorporated in future studies.

K-means algorithm was used for clustering in this project. It was only used because it is a prominent technique for unsupervised learning. As there are different complex techniques for clustering available, they can be utilized to provide a totally different set of clustering. In order to minimize my sum of square error I used k=5 which is open to scrutiny.

This project can further be advanced by taking into consideration a business requirement. An example can be: How to decide where to open an Italian restaurant in Houston.

6. Conclusion:

From this project, an individual moving into Houston can get an insight where he/she should buy a home. Depending on his/her interest they have a choice between populous area/high-low cost housing/desired cluster of venue categories.

7. References:

1. https://www.noradarealestate.com/blog/houston-real-estate-market/

2. http://data.houstontx.gov/dataset/zip-codes-in-the-region

3. https://www.realtor.com/research/data/

4. http://www.city-data.com/zipmaps/Houston-Texas.html