# Capstone Project - The Battle of Neighborhoods

**IN SEARCH OF A CITY**

**Introduction**

* 1. **Background**

This study attributes to the contribution towards the growth of the state and the city. This is about a conclusion for a state and a "city" for the possibility of making an investment. In case if you are a Real estate Investor and you want to make an educated investment decision, you must understand the demography of that place, population information, its growth, Cost of living or real estate.

Populations are never static, they grow or decline because of various reasons- births, deaths or migration. In this study, we will learn about where the Millennials are migrating to, we will learn about the population and its growth and other attributes. We will implement the below methodology to reach to a conclusion. We will follow the Top- down approach to start with as follows:

* Top five cities in Texas based upon real estate list price
* Analyze the city based on Avg Real estate List Price
* Study the city based on other demographic details
* Plot the city and segment the neighborhoods

Data Exploratory method has been adopted to understand the data. You can use this report as a tool to devise a strategy for making an investment or, expand a business and or if it’s a fresh start up in that city. Hence the name is “In search of a city”.

**1.2 Problem**

Let’s assume you are a real estate investor, and you have to make investments for the expansion of your business in some other state or a city, it is imperative that you must know about that area, its population, cost of living and other similar demographics details.

Hence data is acquired and processed so that a conclusion can be made about a city and its neighborhood for the possibility of making an investment.

Data includes city’s demographic details, population growth, age distribution and analysis on other attributes

**1.3 Interest**

Studying the attributes that contributes towards the growth of the state and the city, like change in population across United States and how the migration is taking place, understanding its demographics.

This study will help Real Estate investors who would be interested in finding out the demographic details based upon the migration trends or change in population in United States. This will help Real Estate investors to make more educated investment decisions.

**2. Data acquisition and cleaning**

**2.1 Data sources**

In the first step, we study the change in population for last decade for states within the United States.

National Ranking of states has been listed for the change in population by the US Census Buearu. I have downloaded the data from <https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html>, and read it through read\_csv() in pandas. Based on this data we shortlist the top 3 states that ranked for change in population for last decade telling which states are growing fast.

With some analysis done, I compare the states because of the " Cost of Living " parameter and describe why one outweighs the other.

Data has been collected for the analysis of Top five cities in that state. For that, data has been collected for average real estate list price as per postal codes. The reference of the website is  <https://www.businessinsider.com/texas-cities-everyone-is-moving-to-2019-11> .

In location-based data, some postal codes were missing their latitude and longitude details as per real estate price values. I have prepared the data by explicitly finding those details on Google. I also scraped different websites for data like Inbound migration trends, Population growth of the city. Some of the references are provided below. Some of the sources of the data are obtained from the below list:

1. U.S. Census Bureau
2. www.Simplemaps/uszips
3. <https://www.realtor.com/research/data>

I collect the data for understanding city’s demographics, population growth , and the features like age distribution on the basis of postal codes. I also used Foursquare API to study the neighborhoods using geocoder location data.

**Data Cleaning:**

Data downloaded or scraped from multiple sources for the analysis. When I uploaded the data , data frame was not recognizing most of the column names . The column names were displayed like

| **Unnamed: 1** | **Unnamed: 2** | **Unnamed: 3** | **Unnamed: 4** |
| --- | --- | --- | --- |

Hence I read the data different into the data frame. I skipped the rows and arrange the skipfooter in read\_csv() , so that data can be understood easily.

When I uploaded the data in pandas, some rows converted to NaN along with columns. I dropped the columns using drop method and NaN values using dropna() . I renamed the columns so that it become easier for you to understand the features. Data sometimes populated as strings and not integers/ float, in this case, I have to change their types and also I did the indexing and slicing so that the data is ready for plotting the graphs.

# There were rows and columns that were not required. I decided to Remove the columns and created new dataframe by fetching two columns derived from the above dataset, so that I can use it for plotting. Sorted the data by the column to get the top 3 states for the net rate change of population. Once I obtained the results for top 3 states – I compared California versus Texas – on the basis of Cost of Living by scraping the data from <https://www.bestplaces.net/cost-of-living/dallas-tx/los-angeles-ca/65000> .

I used the relevant data that shows **National Ranking of Regions and States-Change, 2010 to 2019 in Numbers** by US census and analyzed it to derive the results.

For the second step, I captured some data about the real estate's average list price for top 5 cities in the recommended state. I did the indexing and slicing so that the data is ready for plotting the graphs.

In the next step – I upload the data that shows the avg list price of city's real estate based upon its location coordinates and postal codes. I did search for some missing coordinates to make the data more accurate, using Google tool.

**2.3 Feature selection**

After data cleaning, and examining the meaning of each feature, it was clear that there was some redundancy in the features.

These two features contained very similar information for a year and for a decade . I was more interested in finding the ranking of states for 10 years. Some information was in numbers and percent both, about the same data.

Table: Simple feature selection during data cleaning

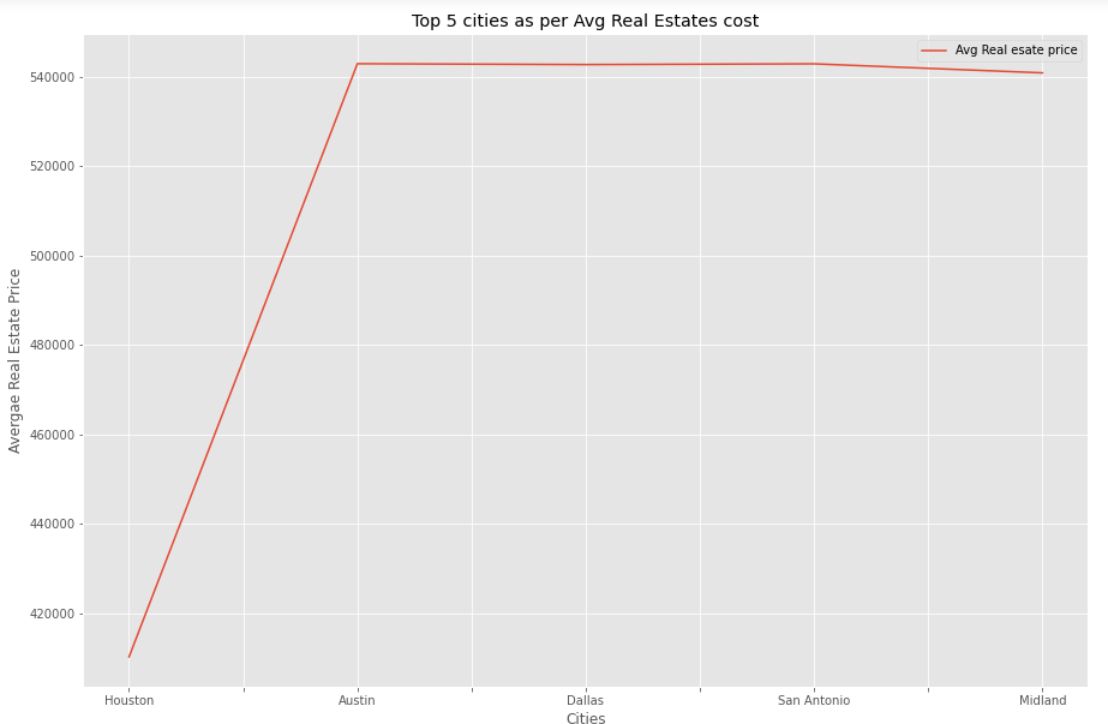
|  |  |  |
| --- | --- | --- |
| **Kept Features** | **Dropped Features** | **Reason for dropping features** |
| |  | **Population Estimate** | | --- | --- |   Of States , | **National Ranking of Regions and States-Change, 2010 to 2019-percent,**  **National Ranking of Regions and States-Population Estimate(July 1, 2019),** | Same Information in percent than numbers .  Only 1 year information. |
| **National Ranking of Regions and States-Change, 2010 to 2019-Number** | **National Ranking of Regions and States-Change, 2010 to 2019-percent** | Ranking of regions and states for 2010-2019 |
| **National Ranking of Regions and States-Change, 2010 to 2019-Number** | **Change, 2010 to 2019-Percent, July 1, 2019,**  **Change, 2010 to 2019-Number, July 1, 2019** | Slightly different features that depict the same overall values |
|  |  |  |

**Exploratory Data Analysis**

**Analysis for top 3 states:** Data had been populated for National Ranking of Regions and States-Change, in population from 2010 to 2019, issued by US census. After sorting and indexing and filtering the top rows , top 3 states were displayed. This suggests the growth of the state in terms of population.

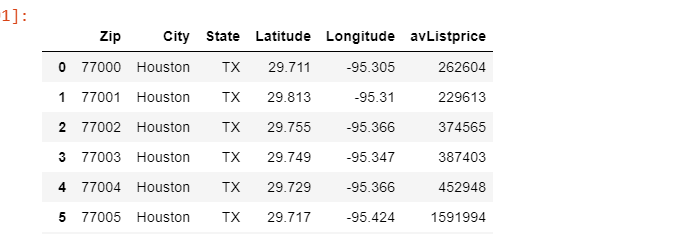
However, Populations are never static, they grow or decline because of various reasons- births, deaths or migration. Based on some extensive analysis, some predictions were made. the results obtained from the external website. Comparisons for top 2 states cost of living was done and we found the cost of living in the second outweighs the first one. Texas was the winning state declared in this case.

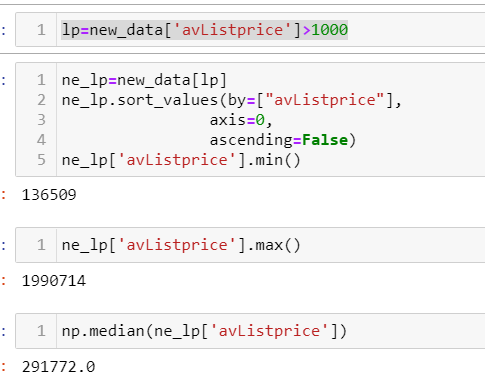
We studied Top 5 growing cities in Texas and uploaded the data for the same. Data consists of cities and its avg real estate price. The plotting was done to explore which city falls on the lower scale of average list price, any yet considered as the fastest growing one. The plot is shown as below:



The above analysis shows that Houston is on the lower side on the scale of Average Real estate Price. Houston is the largest city in Texas and the one of the fastest growing city in United States. Hence, we chose the city as Houston for further analysis.

# **Analysis on Houston city**

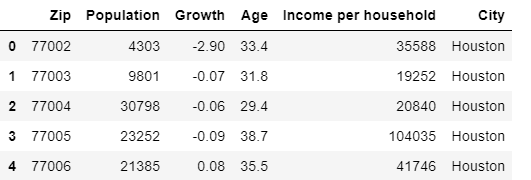


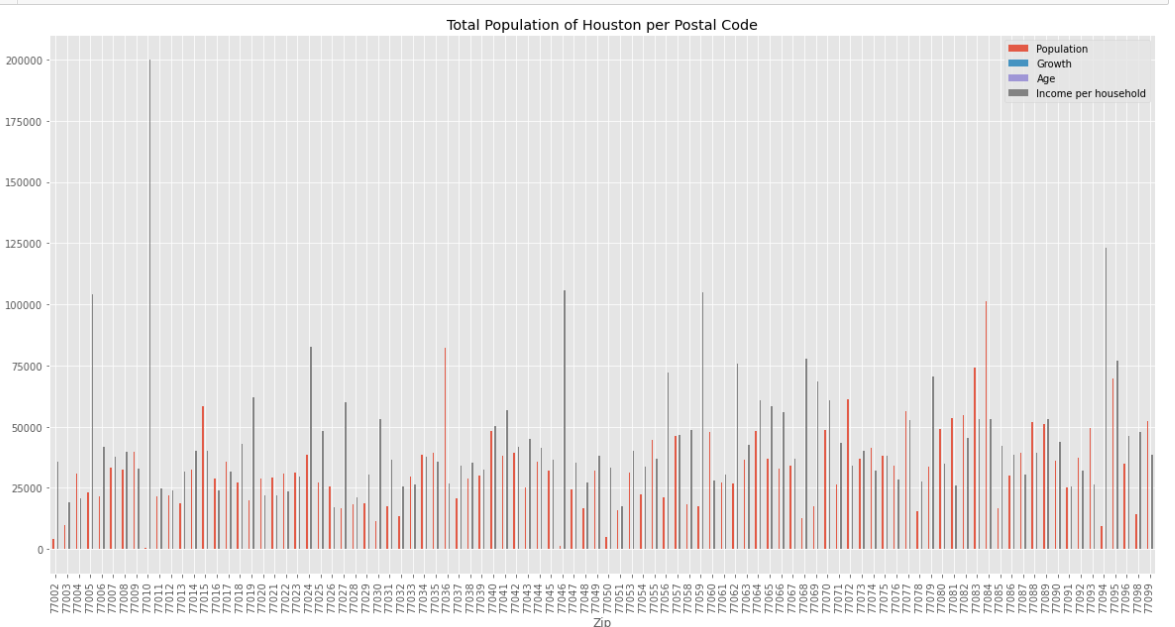


Uploaded the data that shows the avg list price of Houston city's real estate based upon its location coordinates and postal codes. I used statistical functions to find out min, max and median list price of the Real estate. The maximum average list price is 1990714 dollars, and the minimum is 136509 dollars. The median range is 291772. This data gives you an estimate of the real estate properties as per the postal codes based on data provided for Houston.

# **Analysis by Postal code and other attributes**

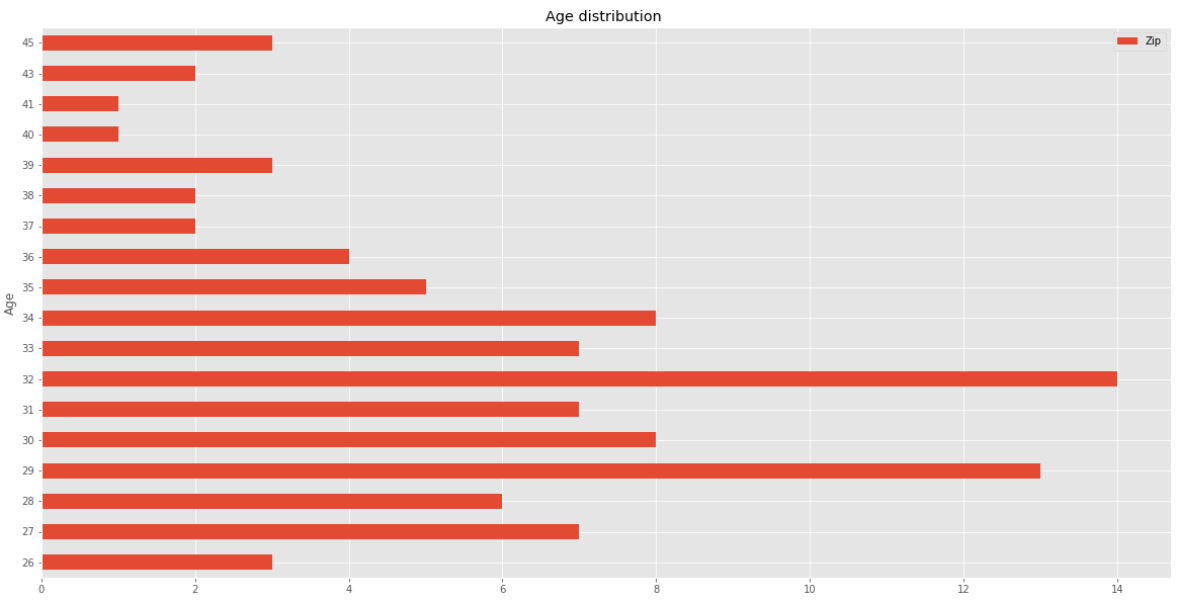
Analyze the attributes like population, growth, age, Income per household as per the postal code of Houston City- which in return give us the information and insight about the neighborhoods.





As per the analysis done above, the average Income Household of the city is 45000 dollars and it also shows the postal codes where the income household is above 100000 dollars.

# **Analysis about the millennials moving to Houston** In this analysis, the median age group will be reflected to show the age distribution in city- Houston. I grouped the age together and using the count method I plotted the below graph:



The above analysis shows the median age group in Houston is around 32 and the maximum population is in the age bracket of (27-34), and this age group is generally called as Millenials.

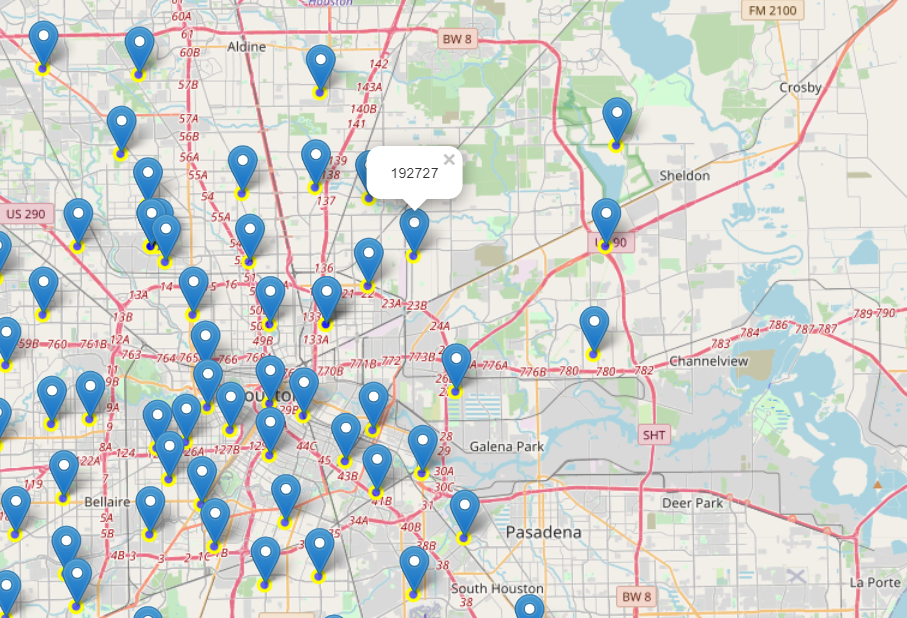
Millennials are playing an increasingly important role in local economics – as companies look for young talent in areas such as software development, cybersecurity, marketing and finance. It’s a generation that is inclined to move more. This can put the Houston area one of the markets for millennial homebuyers.

**Plot the Map of Houston**

I used Geocoder Python package, and Nominatim- (a geocoding software for Open Street Maps ). You know that Geopy can only make requests to Nominatim and using Nominatim. We need to establish a connection to APIs by setting up the geocoder.I imported the geocoder and initiated it .



The result above shows the name of the county which in this case is Harris County.



Since you know that Markers are the items used for marking a location on a map and Folium gives a folium.Marker() class for plotting markers on a map. Hence, the Houston map with the latitude and longitude markers describing the geography is displayed as below and when you click on the marker it will display the avg. list price of the real estate.

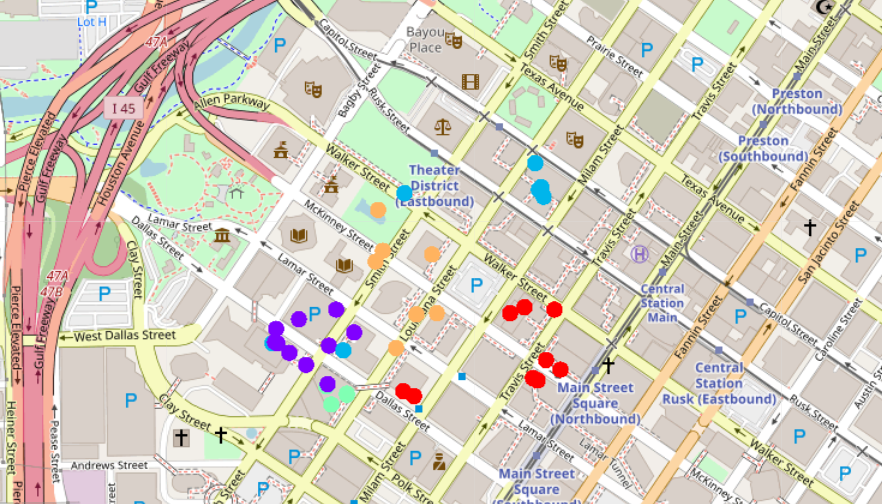
Houston Neighborhood Analysis Using FOUR Square API

Using Foursquare API, I generated the url , using which I defined the locations and venues based upon the categories . In the next step , I explored the neighborhoods where you can find the most common venues . Almost 63 unique categories of the venues were returned. Further we analyzed each neighborhood and found some most common venues as per the neighborhood like Hotels, Burger Joint, Theater Concert hall, coffee shops, parks etc.



I ran k-means to cluster the neighborhood into 5 clusters. Merged to tables to add latitude/longitude for each

neighborhood and then plotted the map:



# Results

* National Ranking of states for Change in population(2010-2019)- 1. California 2. Texas 3. Florida
* Comparison of Cost of living (Texas versus California)- Texas is a good option over the California
* Top five cities in Texas- Austin, Midland, San Antonio, Dallas, and Houston.
* Analyze the city based on Avg Real estate List Price- Houston was on the lower side on the scale of Avg. List Price
* Study the city based on other demographic details-
* Age distribution-(27-34)
* Income per Household-Max- 200000 and avg Income per household is 45000
* Plot the city and segment the neighborhoods-The common venues in the neighborhoods are Hotels, Burger Joint, Theater Concert hall, coffee shops, parks etc.

# **Conclusions**

This study was based upon a conclusion about a "city" and its neighborhood for the possibility of making an investment. We concluded that the Texas is one of the fastest growing states and Houston is among the top cities and is on the lower range of Avg Real estate price.

We learned that the avg Income per household is around 45000 dollars. The range of the listed real estate price in city Houston is( max- 1990714 dollars , minimum - 136509 and the median range is 291772).

Millennials are attracting to Texas and the age distribution of age group in Houston is(27-34).This is one of the factors for economic growth. This can put the Houston area one of the markets for millennial homebuyers.

We plotted the map of Houston and imposed the postal codes on that with the real estate price. This will provide you the insight on the geography of city - Houston

In this case the concert halls and Performing arts venue comes under the 3rd and 4th category of most common venues. Let’s assume that you have to open a Yoga center/Dance Academy , you will choose a location where you get the population of similar interests. This information can be drawn from the analysis of the neighborhoods. For Example, Harris County, in Houston can be such examples of the locations.

References: 1. <https://www.afire.org/summit/migrating-trends/> 2. <https://www.mpamag.com/commercial/how-migration-patterns-are-helping-cre-developers-238246.aspx> 3. <https://worldpopulationreview.com/us-cities/houston-tx-population> 4. <https://www.businessinsider.com/texas-cities-everyone-is-moving-to-2019-11> 6. <https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html> 7. [https://www.wikipedia.org](https://www.wikipedia.org/) Tools:  
8. [https://www.google.com](https://www.google.com/) 9. [https://www.stackoverflow.com](https://www.stackoverflow.com/)