# House Purchase

Applied Data Science Capstone Project

**IBM DATA SCIENCE PROFESSIONAL CERTIFICATE SPECIALIZATION | COURSERA**

**Akshay Bhaskaran**

## Introduction

Purchasing a house could be a painful task, especially if it is your first one, and you have many options lying in front of you. There are so many factors to consider, compare and evaluate before you decide where, what, and how your new house is going to be. This project aims to provide, at a high level, some suggestions and comparison metrics that helps in narrowing down the options to purchase a house. People who are in search of their new houses and are confused as to which among these three locations - Austin, Round Rock, and Cedar Park - to choose, this project would definitely be a good starting point for them to gain a better understanding and good insight into the nuances of the city.

## Data

One main data that's used to solve this problem would be location data that includes various venues, categories, and other details around a given area, and this data is completely available from https://developer.foursquare.com/. For getting latitude and longitude details of specific locations, I used https://opencagedata.com/. Other useful data that was used to solve this problem was data about Austin and its neighboring cities which was available from wikipedia. With respect to area-specific real-estate details like average price of the house, % of increase from last year, and predicted % in the next upcoming year was collected from https://www.zillow.com/. Then, there was a place where I needed Austin's zip code data which was collected from http://www.city-data.com/zipmaps/Austin-Texas.html. Finally, to get further insights into each zip code, a special python library was used that can be referred here: https://pypi.org/project/uszipcode/.

## Methodology

This section talks about the overall methodology involved, with example screen captures of some steps.

**Libraries and credentials:**

- Before diving into the project directly, a list of all python libraries that would be needed to better execute this project was analyzed, installed and kept ready.
- Since this project is mostly based on API's available on the internet, I'd created accounts over the respective websites, and got the credentials to access those APIs

**Shortlisting:**

- It is nearly impossible to analyze and explore all the cities in Texas, and that would become a tedious task
- So, I took Austin, TX as my mid-point and looked at major cities surrounding Austin along with their population to arrive at this data:

| | CityName | Population |
|---|---|---|
| 0 | Austin | 950,715 |
| 1 | Round Rock | 123,678 |
| 2 | Cedar Park | 75,704 |
| 3 | Georgetown | 70,685 |
| 4 | San Marcos | 63,071 |
| 5 | Pflugerville | 59,245 |
| 6 | Leander | 42,761 |
| 7 | Kyle | 39,060 |
| 8 | Hutto | 23,832 |

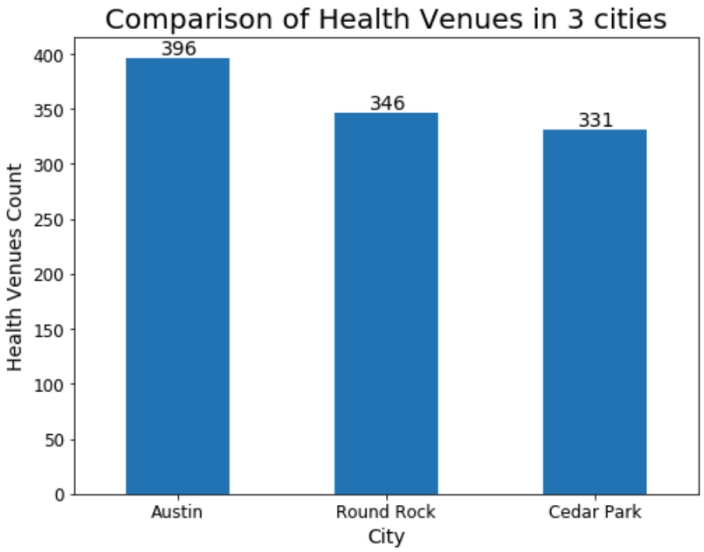
- It is still a big task to kind of analyze these 8 cities, so I further shortlisted my research to top 3 cities with in terms of population that left me with Austin, Round Rock, Cedar Park
- With this three cities fixed, the very next step I did was to get their latitude, and longitude values which acts as the major data to explore online API's with respect to different venues around these cities.

**Exploratory Analysis:**

- For the selected three cities, I did an exploratory analysis to better understand the city's facilities that plays a major factor when deciding to buy a house
- This analysis was based on 5 major categories:
    - **Health** : places like hospital, pharmacies, clinics, emergency rooms etc.
    - **Food** : this includes grocery shops, restaurants, coffee shops etc.
    - **Education** : which has elementary schools, middle schools, high schools, colleges/universities etc.
    - **Entertainment** : movie theaters, bars, other night life activities
    - **Average price** : average price of a house in that particular city, along with last year rate, and forecast prediction
- The reason to begin my analysis with these top 5 categories is because these are the first and foremost things that anyone would like to evaluate before getting a house in a particular city
- For each of the category listed above, I used FourSquare API's to get an overall count of different venues that fall into a single category. For example, all hospitals, all ER's, all dentists, all eye-doctors, all pharmacies are all combined into a single category called health, and its corresponding count was taken.
- The results were plotted in bar graphs format to easily understand and appreciate the differences between each cities.
- For the final analysis on average price of the house, I tried to use a simple line graph to show the previous year's house rate, current average price, and a projected prediction for 2020.

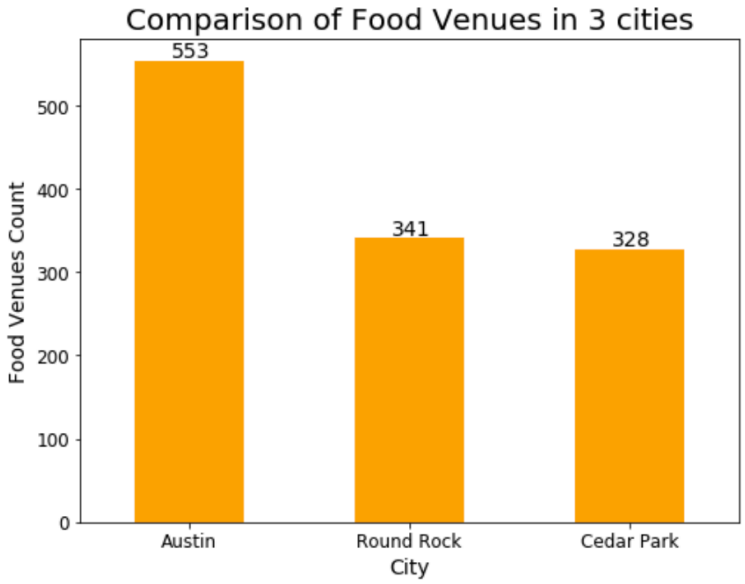
**a. Health**

Though health is a single term, there were a lot of different venues that was taken into account including, but not limited to - hospitals, pharmacies, emergency rooms, dentists, orthopedics, eye-doctors, clinics, specialized doctors, and more. An overall aggregate of these venues was plotted, and compared, for three cities.



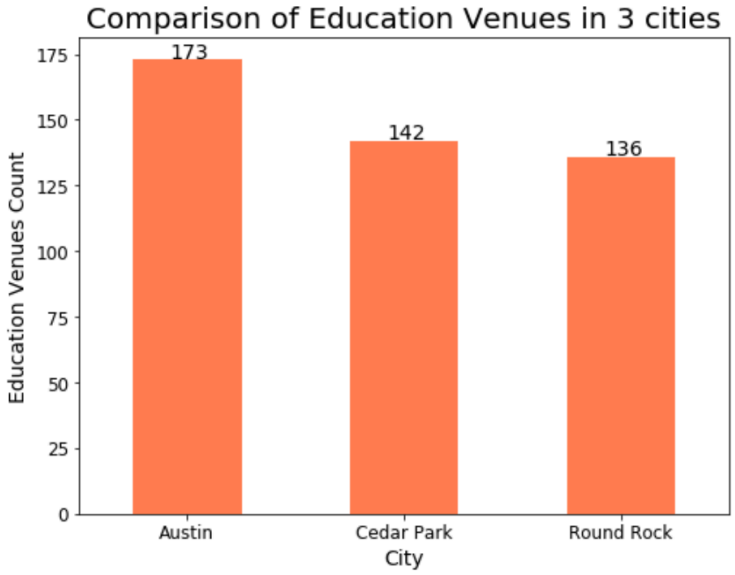

**b. Food**

Food is the second next important thing after health (at least for me). So, this exploration was based upon all categories that falls under food - like restaurants, coffee shops, sandwich shops, taco shops, burger joints, steakhouses, juice parlors, grocery stores, condiments stores etc. And again, the overall aggregate was plotted against three cities.



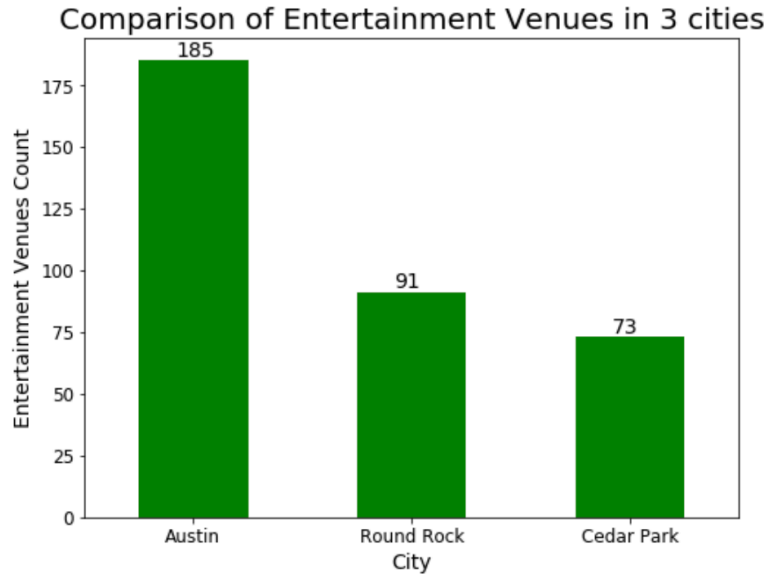
**c. Education**

Education is much important as food is, and in this category I've tried to touch all different venues possible like elementary schools, middle schools, high schools, colleges, universities, tutor academies, coaching classes etc.




**d. Entertainment**

Okay!! You now have good health, ate good food, and got great education. Don't you want some entertainment in life? Yes, I'd definitely prefer that, so here comes the comparison of all entertainment venues in three cities like movie theaters, parks, disco halls, clubs, bars, nightclubs, comedy clubs, drama clubs etc.

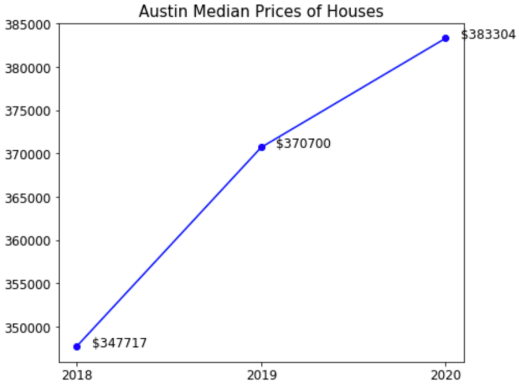


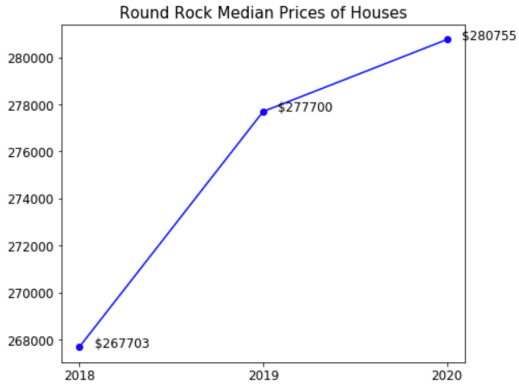
**e. Average Price**

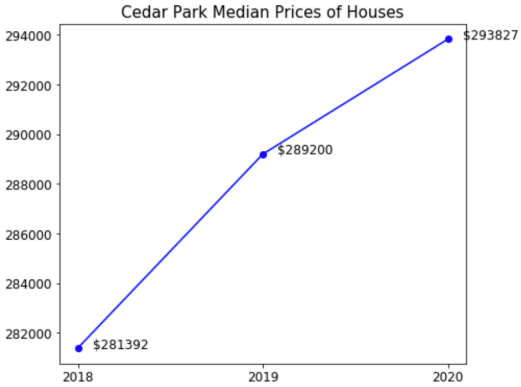
Once all our initial checks are done, we'd now definitely be interested to look at a particular city's real-estate history. We would like to know the average prices of house in the previous year (2018), the current average price of the house (2019), and forecasted (predicted) price of the house in the upcoming year (2020). For this, we use Zillow's home-values site along with some parsing to get the values for all three cities.

| | City | Last Year Value | Current Median Value | Next Year Forecast |
|---|---|---|---|---|
| **0** | austin | 6.2% | 370700 | 3.4% |
| **1** | round-rock | 3.6% | 277700 | 1.1% |
| **2** | cedar-park | 2.7% | 289200 | 1.6% |

This table looks good, but it would be great if we could get some ballpark values instead of "%"'s that would give the user a good exact view of the market scenario. So, applying a little math here and plotting the graphs.




Based on the above analysis on 5 different categories, looks like Austin is our best bet! So, the end user now can fix Austin, TX as their final city to start looking at houses, instead of getting confused with Round Rock, or Cedar Park.

**Researching Austin:**

**a. Collecting data:**
It is not enough to just say get your next house in Austin, which is like pushing a blind-folded man into a forest to pick a pinecone. We're going to do further statistical analysis into various areas inside Austin to see which ones are the best to purchase houses.

- As a first step, we collect all the zip codes that fall under Austin, TX
- Then, by using the python library "uszipcode", we search each zip code to get more details on every one of them like but not limited to area_code_list, median_house_price, common_city_list, county name, housing_units, occupied housing_units, population, latitude, longitude, radius in miles, water area, post office city etc.
- We format these details into a neat data frame to make it easier for analysis. Also, we drop some unwanted columns from the database to keep our attention focussed.

**b. Machine learning:**

- Once we have all the data in the format of a data frame, it is time for us to apply some machine learning algorithms on it. Here, we are going to do an un-supervised learning by applying "**Clustering Algorithm**". The reason for choosing clustering algorithm is - it is very useful in location-specific data like this to kind of narrow down and cluster different groups of data together, and helps in getting useful insights from each cluster.
- After setting the cluster_size = 4, and running it on the data frame, we now get the overall data frame split into 4 separate clusters.
- We are analyzing each cluster to see if we are able to get any useful, valuable insights from the data
- While analyzing each cluster, I tried to sort the median price of a house in ascending order, and this is what I've observed:

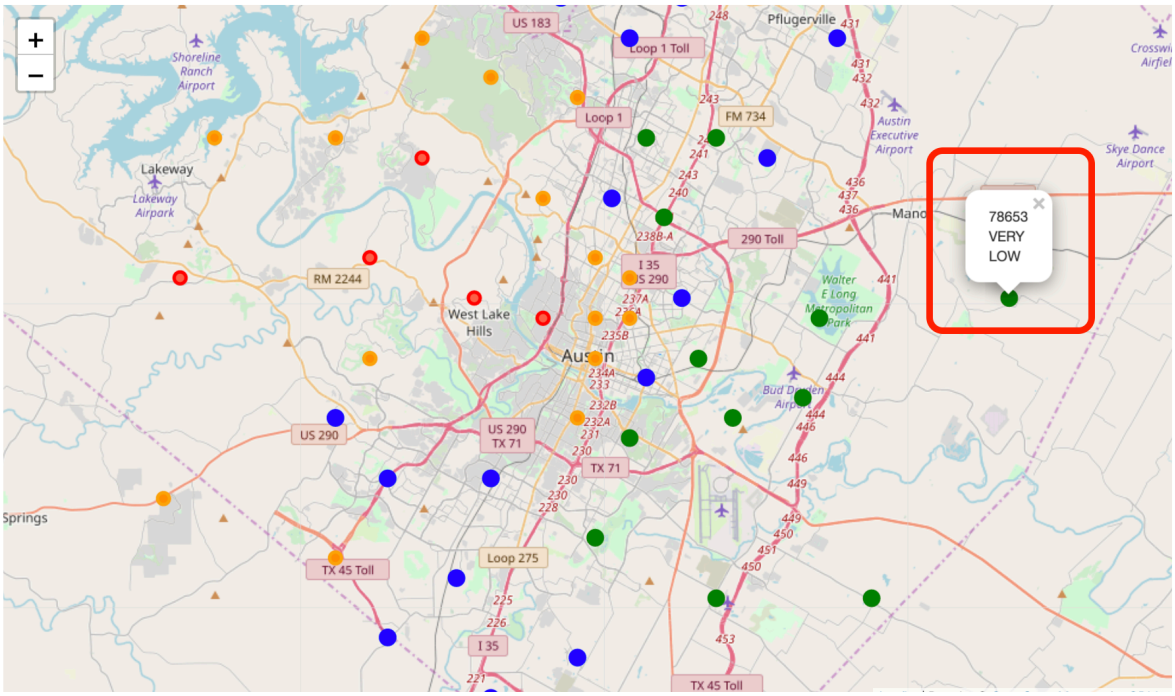
| Cluster number | Median Price of House (Minimum range) | Median Price of House (Maximum range) |
|---|---|---|
| 0 | $156,900 | $259,400 |
| 1 | $460,200 | $641,200 |
| 2 | $279,200 | $422,300 |
| 3 | $72,100 | $163,900 |

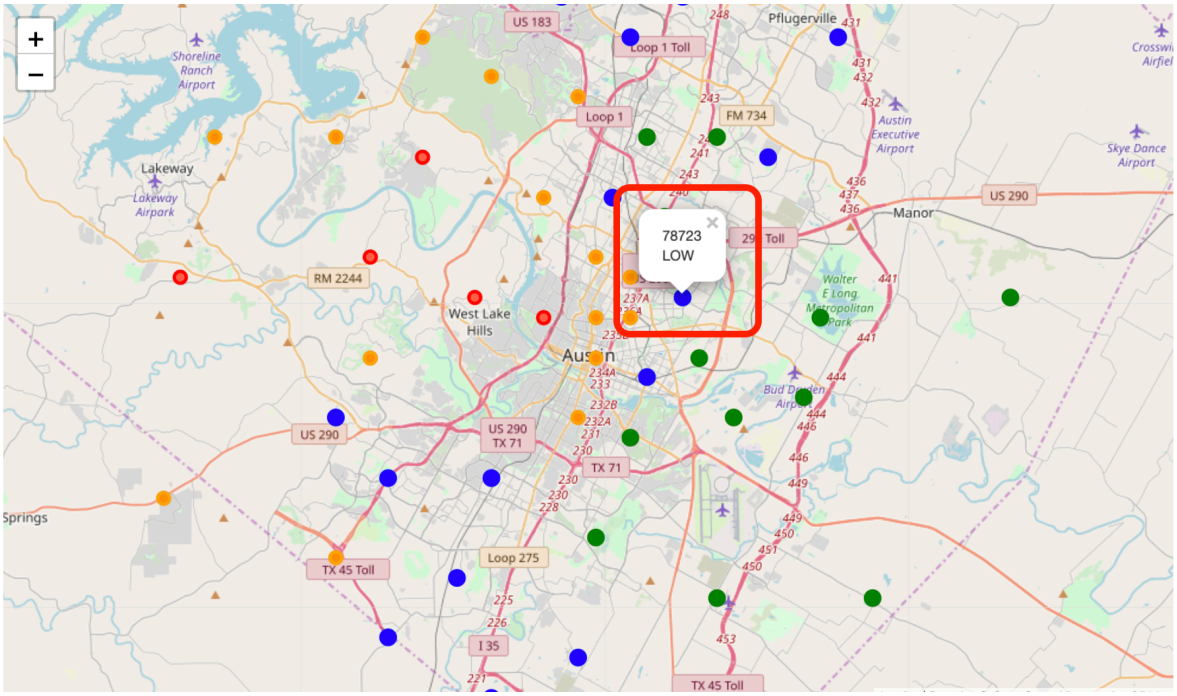
- As we can clearly see, the clustering had separated and provided us with the bucket range of median price of the houses in each zip codes in Austin.
- I've also given a category name for these bucket ranges as follows:

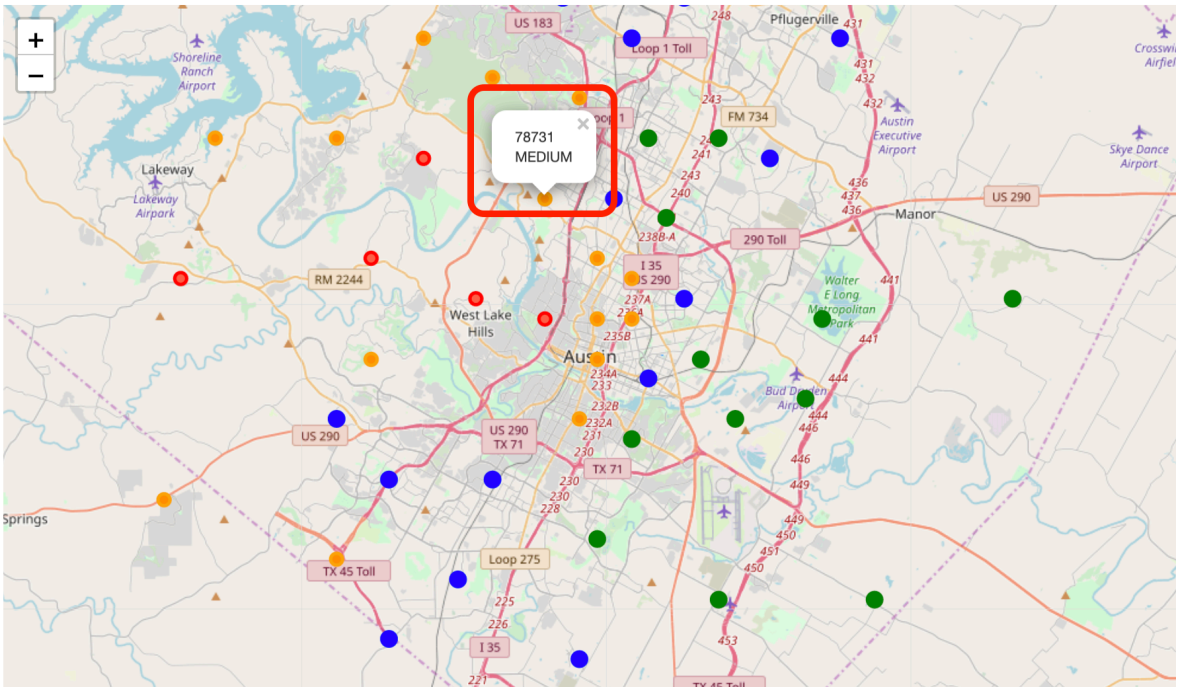
| Cluster number | Median Price of House (Minimum range) | Median Price of House (Maximum range) | Budget |
|---|---|---|---|
| 0 | $156,900 | $259,400 | **LOW** |
| 1 | $460,200 | $641,200 | **HIGH** |
| 2 | $279,200 | $422,300 | **MEDIUM** |
| 3 | $72,100 | $163,900 | **VERY LOW** |

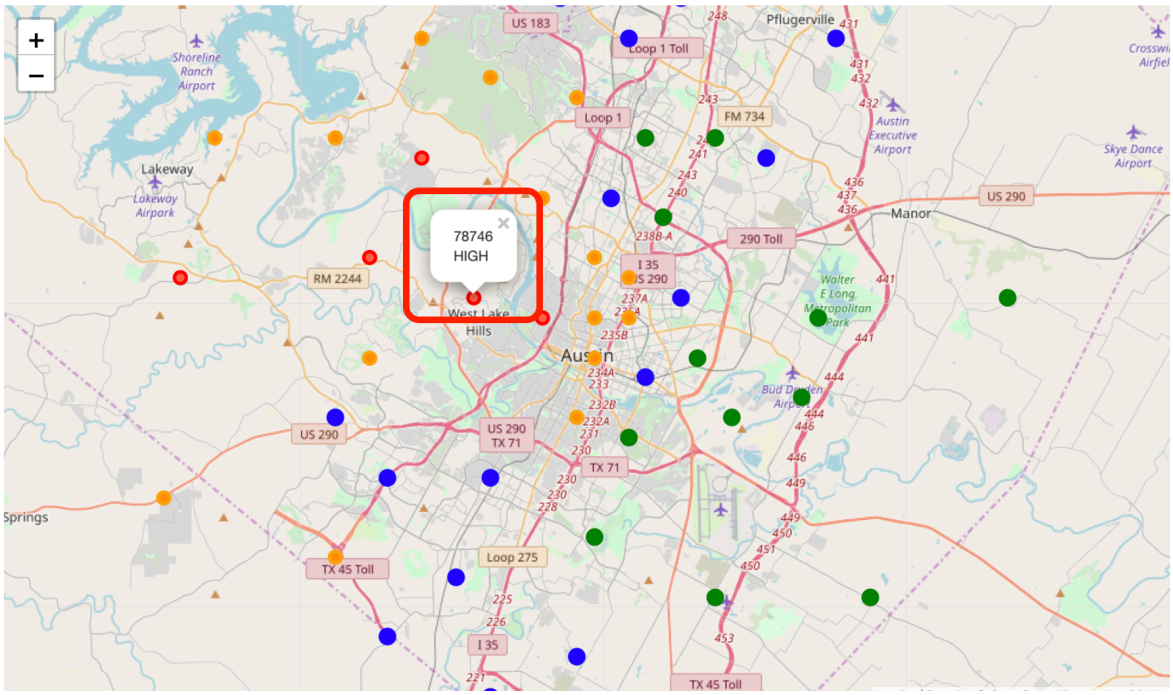
- Now, a new data frame is created along with this "Budget information", and its corresponding zip codes, latitude and longitude. This data frame will be used to plot a

final map of Austin indicating the very low, low, medium, and high budget areas along with their respective zip codes.




78731
MEDIUM
US 183
Loop 1 Toll
Pflugerville
Shoreline Ranch Airport
Lakeway
Lakeway Airpark
FM 734
Austin Executive Airport
Skye Dance Airport
US 290
Manor
290 Toll
I 35
RM 2244
West Lake Hills
Walter E Long Metropolitan Park
Austin
Bud Dworsky Airport
US 290
TX 71
Springs
TX 45 Toll
Loop 275
I 35


78746
HIGH
US 183
Loop 1 Toll
Pflugerville
Shoreline Ranch Airport
Lakeway
Lakeway Airpark
Loop 1
FM 734
Austin Executive Airport
Skye Dance Airport
US 290
Manor
290 Toll
I 35
RM 2244
West Lake Hills
Walter E Long Metropolitan Park
Austin
Bud Dworsky Airport
US 290
TX 71
Springs
TX 45 Toll
Loop 275
I 35

# Results

Based on the above analysis performed, the following results can be finalized

i. Out of three cities - Austin, Round Rock, Cedar Park - Austin has a very good scope for purchasing a new house. This result is supported by various factors like the number of venues in health, education, food, and entertainment category. Also, the % of increase in the median value of houses in Austin and the projected forecast for the upcoming year makes it great and stand first in the real-estate market.
ii. In Austin, there are houses available in four different budget categories - very low, low, medium, and high budget. So, these are your zip code - budget reference tables to keep handy when hunting a house in Austin!

| | |
|---|---|
| VERY LOW | 78742 |
| VERY LOW | 78617 |
| VERY LOW | 78724 |
| VERY LOW | 78744 |
| VERY LOW | 78725 |
| VERY LOW | 78719 |
| VERY LOW | 78721 |
| VERY LOW | 78741 |
| VERY LOW | 78753 |
| VERY LOW | 78653 |
| VERY LOW | 78758 |
| VERY LOW | 78664 |
| VERY LOW | 78752 |

| | |
|---|---|
| LOW | 78641 |
| LOW | 78754 |
| LOW | 78660 |
| LOW | 78747 |
| LOW | 78745 |
| LOW | 78728 |
| LOW | 78723 |
| LOW | 78610 |
| LOW | 78748 |
| LOW | 78702 |
| LOW | 78652 |
| LOW | 78729 |
| LOW | 78613 |
| LOW | 78727 |
| LOW | 78681 |
| LOW | 78736 |
| LOW | 78749 |
| LOW | 78717 |
| LOW | 78757 |

| | |
|---|---|
| MEDIUM | 78722 |
| MEDIUM | 78751 |
| MEDIUM | 78750 |
| MEDIUM | 78705 |
| MEDIUM | 78759 |
| MEDIUM | 78734 |
| MEDIUM | 78737 |
| MEDIUM | 78726 |
| MEDIUM | 78704 |
| MEDIUM | 78739 |
| MEDIUM | 78756 |
| MEDIUM | 78701 |
| MEDIUM | 78735 |
| MEDIUM | 78732 |
| MEDIUM | 78731 |

| | |
|---|---|
| HIGH | 78738 |
| HIGH | 78733 |
| HIGH | 78730 |
| HIGH | 78746 |
| HIGH | 78703 |

# Conclusion

After performing different exploratory and statistical analysis on the location data for three different cities Austin, Round Rock, Cedar Park - it can be concluded that Austin is the top-most city with respect to all our aspects used in the comparison. Inside of Austin, it is possible to cluster and group zip codes together based on the median price of houses that gives a greater insight into various areas and their budgets in Austin.