Austin, TX - Real Estate

Applied Data Science Capstone Project
Miguel Bencomo

Introduction

Over the past ten years the city of Austin, TX went from being the 14th to the 11th most populous city in the United States. This growth is largely in part to the tech market which has grown by 25% over the past five years, making Austin the home to many tech giants, including Google, Facebook, Oracle, and Apple. This growth has also impacted surrounding areas like Georgetown, Cedar Park and Round Rock. As the region grows, it is predicted between 3 and 4 million people will live in the 4 metro areas by 2040. Housing is expected to become more expensive and more compact as the population continues to grow.

Business Problem

Finding and purchasing a house is already a complicated process that involves a lot of variables. Some of those variables are price, which requires research through different websites and/or real state agents; location, the house should be in a good neighborhood with great gathering spots, parks, shops and a good school district. For this project we will look at real state data from Austin and it's surrounding areas to determine good deals for potential buyers in this growing regions based on their requirements.

Data

The data needed for this project will include:

1. Zillow website data. Using the data from this real state website we can gather information about house prices, location, school districts, etc.
2. FourSquare data. Using FourSquare API we can get information on the neighborhoods, including shops, restaurants, and venues around them.
3. City-data. Use this website to gather information about the city of Austin and its surrounding areas.

Methodology

Libraries

For this project we will use a variety of libraries which include: numpy and pandas for store and manipulate data frames and tables; matplotlib and folium for plotting graphs and maps, respectively; requests to handle html data, as well as BeautifulSoup to extract data from it; and sklearn to use clusterinf in determining areas with different price points for houses.

Extracting Data

The methodology for this project is as follows. First, we determine all the zip codes in the Austin area from http://www.city-data.com/zipmaps/Austin-Texas.html. From this website we obtain 52 different zip codes in the greater Austin area. Those zip codes include:

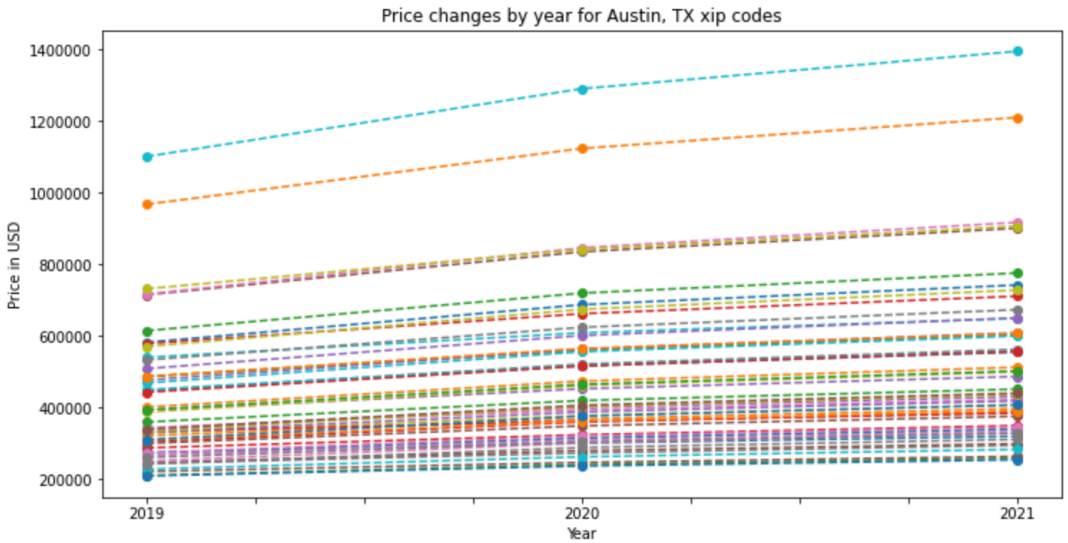
['78610', '78613', '78617', '78641', '78652', '78653', '78660', '78664', '78681', '78701', '78702', '78703', '78704', '78705', '78717', '78719', '78721', '78722', '78723', '78724', '78725', '78726', '78727', '78728', '78729', '78730', '78731', '78732', '78733', '78734', '78735', '78736', '78737', '78738', '78739', '78741', '78742', '78744', '78745', '78746', '78747', '78748', '78749', '78750', '78751', '78752', '78753', '78754', '78756', '78757', '78758', '78759']

With this information, we can then obtain the average values of a house in each zip code from zillow.com, where we can also get how the average price has change from last year (change) and the predicted change for the next year (forecast), both as a percentage.

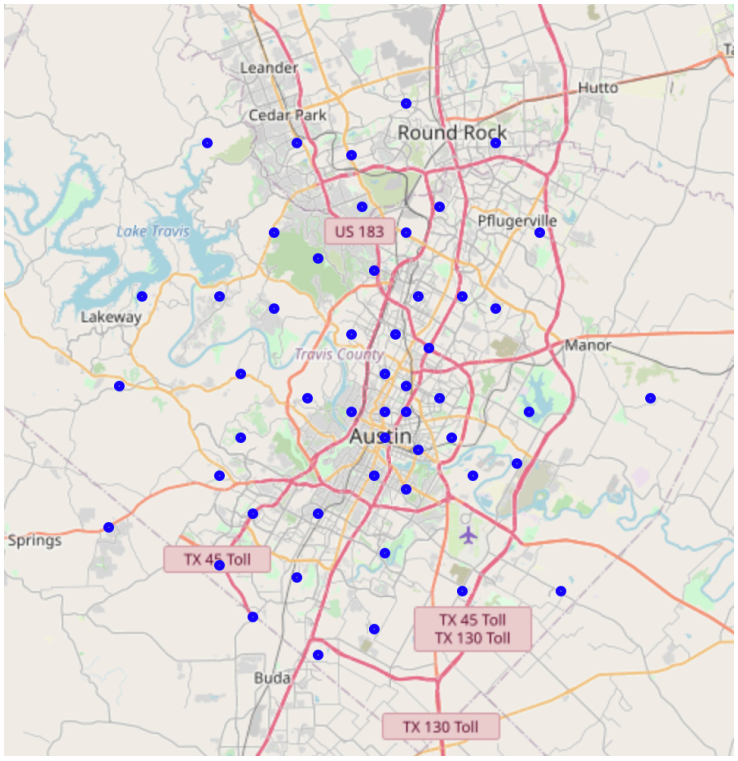
	Zip code	Avg price	Change	Forecast
0	78610	300935	8.5	6.3
1	78613	368184	12.7	7.9
2	78617	240601	12.5	7.5
3	78641	324149	11.3	7.9
4	78652	400637	14.9	7.5

From this table we can actually determine the price for last year and the predicted price for next year where we can check that the prices of houses are going up as the trends

seem to indicate. This happens to be true for all zip codes in the Austin area, as it can be seen in the graph below, where each line represents a different zip code.



We now need to get the latitude and longitude for all the zip codes using python's uszipcode library. This library contains a lot more information that just coordinates however, we are just interested in the latitude and longitude of each zip code. The coordinates can then be plotted to visually locate each zip code.



Based on the coordinates of each particular zip code we can use FourSquare API to determine the number of venues within a 2 mile radius. From FourSquare we obtain the name of the venue, its coordinates, and the venue category. An example is given below for the 78610 zip code.

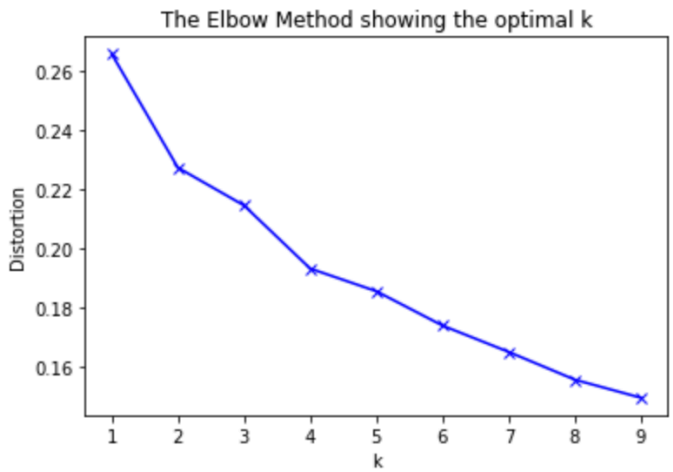
	name	categories	lat	lng
0	Starbucks	Coffee Shop	30.090111	-97.816899
1	Chili's Grill & Bar	Tex-Mex Restaurant	30.084372	-97.823409
2	Hat Creek Burger Co.	Restaurant	30.083872	-97.824578
3	Dan's Hamburgers	Burger Joint	30.085976	-97.822469
4	Chick-fil-A	Fast Food Restaurant	30.083136	-97.823442

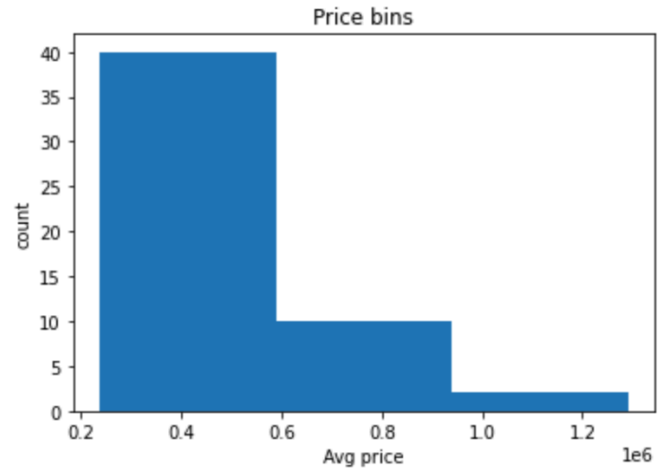
For each zip code we can determine the frequency of each category and obtain the most common venues.

	Neighborhood	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	78610	Pizza Place	Hotel	Fast Food Restaurant	Burger Joint	Price	American Restaurant	Fried Chicken Joint	Big Box Store	Gym / Fitness Center	Boutique
1	78613	Price	Mexican Restaurant	Pizza Place	BBQ Joint	Burger Joint	Coffee Shop	Fast Food Restaurant	Pharmacy	Donut Shop	Sandwich Place
2	78617	Racetrack	Market	Music Store	Garden Center	Recording Studio	Price	Pool Hall	Nail Salon	Noodle House	Print Shop
3	78641	Bar	Park	Lake	Burger Joint	Tex-Mex Restaurant	Paintball Field	Liquor Store	Gun Range	Sandwich Place	Harbor / Marina
4	78652	Mexican Restaurant	Price	Pizza Place	Music Venue	Gas Station	Locksmith	Park	Hardware Store	Dive Bar	Thrift / Vintage Store

Clustering

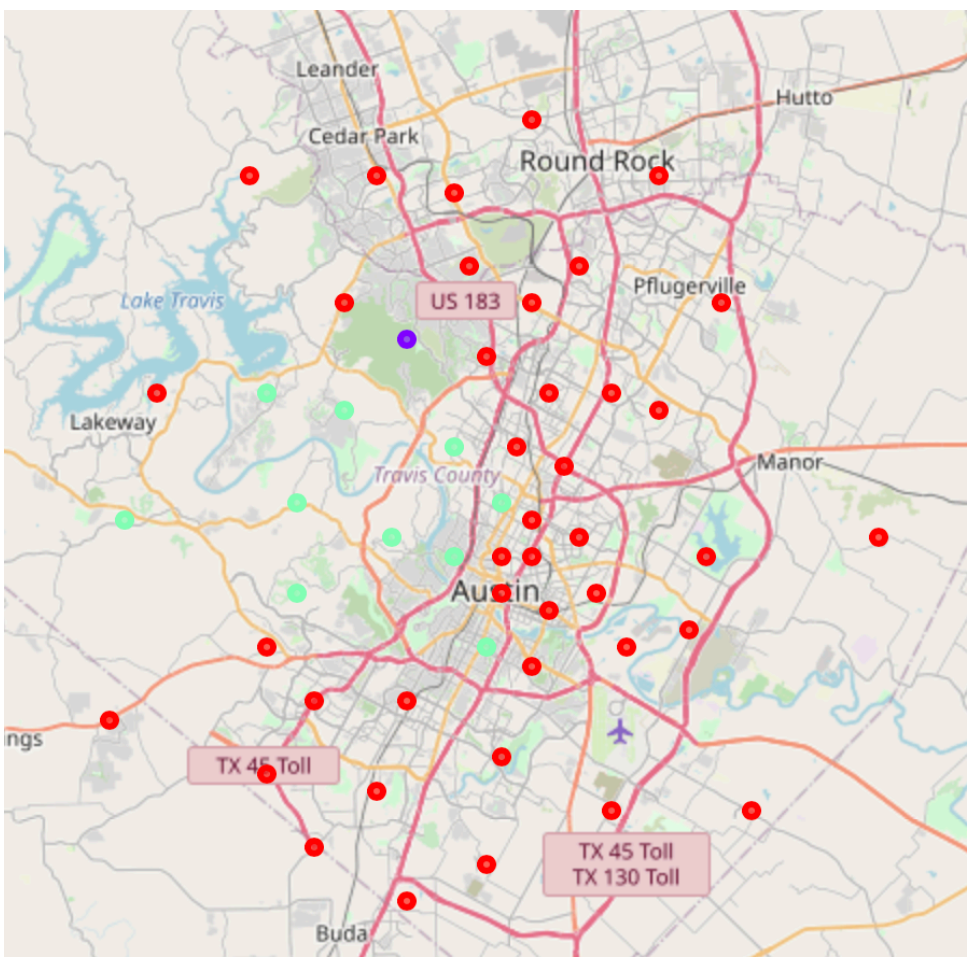
Using unsupervised machine learning, we use a clustering algorithm which is very useful with location-specific data which we want to group and analyze. We can the analyze each group or cluster and obtain insight about them. For this particular case We used 3 clusters based on the elbow method for determining an optimal k. Once we run the algorithm each zip code will be labeled with a cluster number. Each cluster represents a different price point, which we can plot easily.



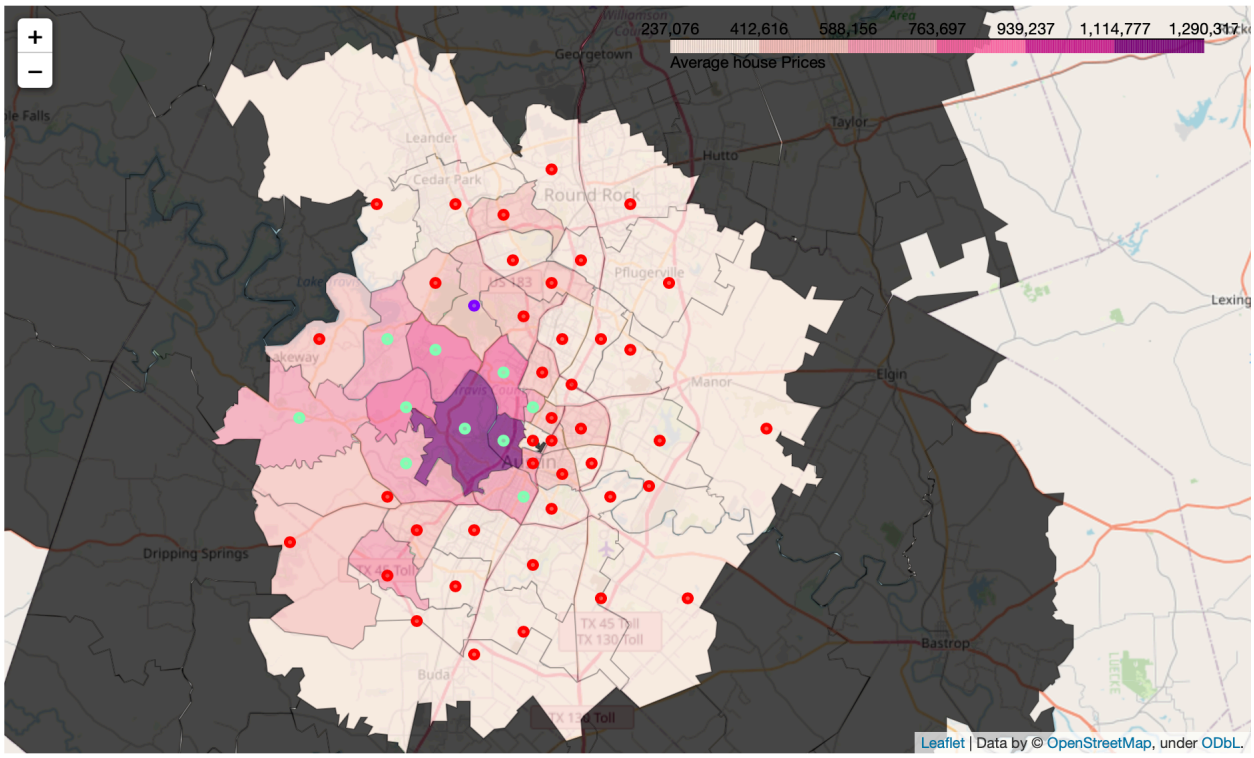


- Cluster 0: Low price point ($237,076 - $588,156)
- Cluster 1: Intermediate price point ($588,156 - $939,236)
- Cluster 2: High price point ($939,236 - $1,290,317)

We can more easily see this by visualizing the price points in map.



In addition a choropleth map visualizing the average house pricing with the clusters on top.



Results

Based on the analysis performed the downtown areas and expanding towards the west are very high, you can clearly visualized in the map above while in the suburbs and the neighborhoods away from the city center have lower prices Almost all low price neighborhoods are close to restaurants, pubs, sports facilities etc.

Conclusions

It can be concluded that prices have rising in the greater Austin area and its surroundings. Most houses are on the lower price point in most of the area, however if a buyer want to be closer to downtown, with hotels, bars, night life and in general closer to a wider variety of venues, the price points will go up significantly.