House Market and Venues in Austin

Author: Rahul Ravi

1. Introduction

Austin is the capital city of the U.S. state of Texas, as well as the seat and largest city of Travis County, with portions extending into Hays and Williamson counties. Incorporated on December 27, 1839, it is the 11th-most populous city in the United States, the fourth-most-populous city in Texas, and the second-most-populous state capital city (after Phoenix, Arizona). It was also the fastest growing large city in the United States in 2015 and 2016.

At the 2000 United States Census, there were 656,562 people, 265,649 households, and 141,590 families residing in the city (roughly comparable in size to San Francisco, Leeds, UK; and Ottawa, Ontario, Canada). The population density was 2,610.4 inhabitants per square mile (1,007.9/km2). There were 276,842 housing units at an average density of 1,100.7 per square mile (425.0/km2).

Austin is considered to be a major center for high tech. Thousands of graduates each year from the engineering and computer science programs at the University of Texas at Austin provide a steady source of employees that help to fuel Austin's technology and defense industry sectors. The region's rapid growth has led Forbes to rank the Austin metropolitan area number one among all big cities for jobs for 2012 in their annual survey and WSJ Marketwatch to rank the area number one for growing businesses. The proliferation of technology companies has led to the region's nickname, "Silicon Hills," and spurred development that greatly expanded the city.[1]

Austin’s population has been growing at an average rate of 55,500 people every year or 155 people every day since 2010. The biggest source of new residents for the city come from other parts of Texas, followed by California, Florida, New York, and Illinois. The city also leads in job creation, Economists at PriceWaterhouseCoopers expect the city to create more jobs than any other metropolitan area in the country in 2020, despite the pandemic.[2] When we think of it by the investor, we expect from them to prefer the neighborhoods where there is a lower real estate cost and the type of business they want to start is less intense. If we think of the city residents, they may want to choose the regions where real estate values are lower. At the same time, they may want to choose the neighborhood according to the social place’s density. However, it is difficult to obtain information that will guide investors in this direction.
When we consider all these problems, we can create a map and information chart where the real estate index is placed on Austin and each district is clustered according to the venue density.

2. Data Description

We will be using the following sources to get our required data from:

- Foursquare API to get the most common venues of given neighborhood of Austin.
- An excel file containing a list of Austin neighborhoods and the associated median home prices downloaded from the following zillow.
- A json file from the Official City of Austin Data portal containing a list of neighborhoods and their planning areas.

2.1 Get geographic coordinates of Austin for map initialization

- By using geopy, the geographical coordinate of Austin is 30.2711286, -97.7436995.

2.2. Get data from Geojson boundary file and housing data for each neighborhood

- Retrieved from a public dataset [3]
- Boundaries are used to plot a choropleth map that shows the number of Asian restaurants
  nearby for each neighborhood
- Central coordinates are used to add markers for each neighborhood with pop-up text on the
  map
- Clean the GeoJSON data to extract latitude and longitude for each neighborhood

	planning_a	shape_area	label	shape_leng	_feature_i	geometry	Center_point	lat	long
0	ALLANDALE	65792689.5531	Allandale	42253.1072105	1.0	MULTIPOLYGON (((-97.73974 30.32808, -97.73962 ...	POINT (-97.74517 30.34030)	30.340301	-97.745169
1	BARTON HILLS	88901714.7947	Barton Hills	48353.9339254	2.0	MULTIPOLYGON (((-97.79627 30.23398, -97.79767 ...	POINT (-97.78837 30.25202)	30.252021	-97.788367
2	BOULDIN CREEK	33258999.9494	Bouldin Creek	25667.3403756	3.0	MULTIPOLYGON (((-97.75962 30.24211, -97.76031 ...	POINT (-97.75563 30.25170)	30.251705	-97.755626
3	BRENTWOOD	44207756.068	Brentwood	29612.4036976	4.0	MULTIPOLYGON (((-97.73692 30.31449, -97.73757 ...	POINT (-97.73245 30.33062)	30.330625	-97.732451
4	CENTRAL EAST AUSTIN	26970983.8606	Central East Austin	22198.5298969	5.0	MULTIPOLYGON (((-97.71925 30.27073, -97.71903 ...	POINT (-97.72415 30.26974)	30.269742	-97.724153

- Extract Neighborhood median housing price data from dataset [4]

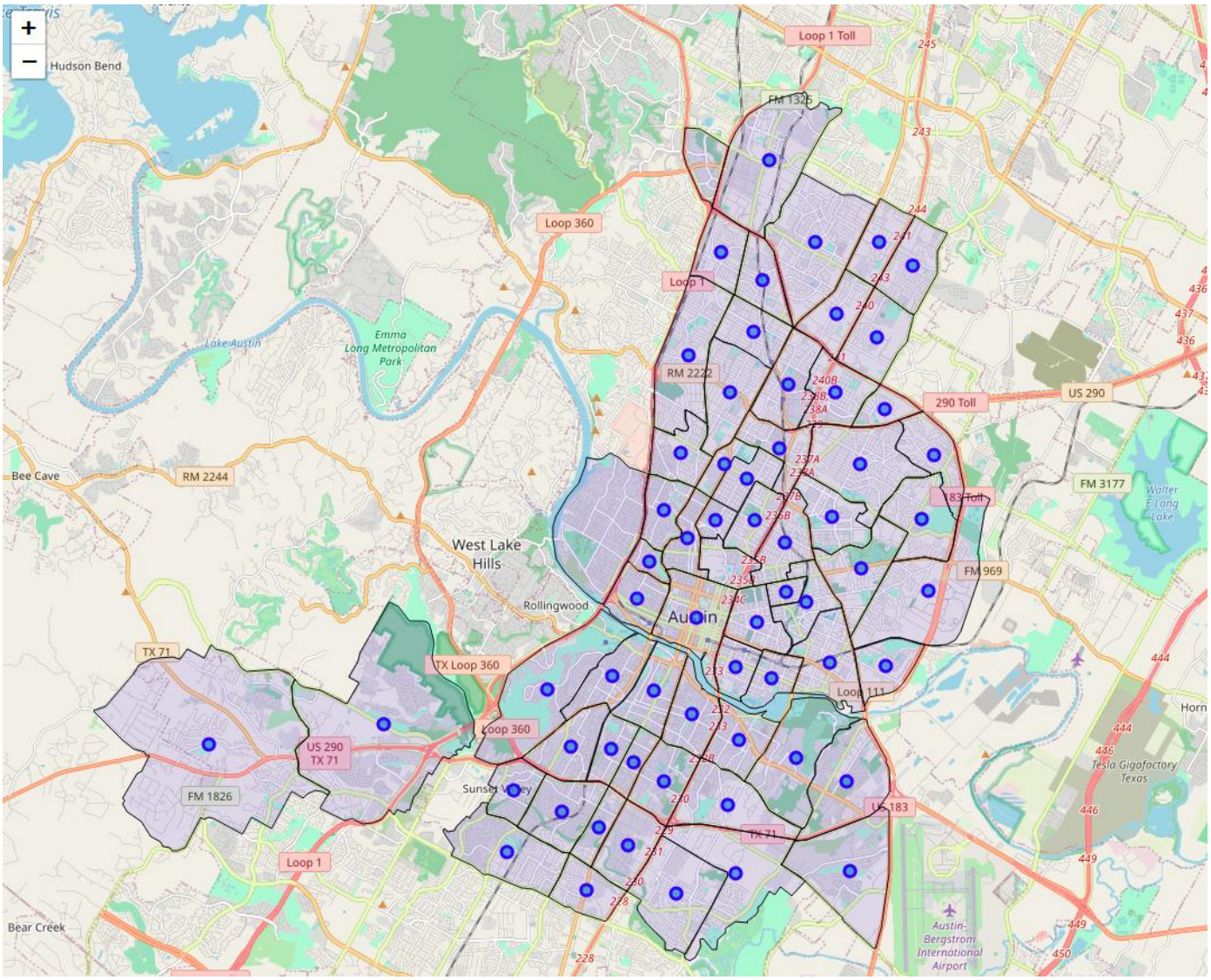
	Region Name	Region Type	Type	Current
0	Allandale	neighborhood	All Homes	621300
1	Barton Creek	neighborhood	All Homes	1384600
2	Barton Hills	neighborhood	All Homes	690100
3	Bouldin Creek	neighborhood	All Homes	710300
4	Brentwood	neighborhood	All Homes	493800

- Clean data and Merge results with the data frame containing Neighborhood coordinates to arrive at the source data set that we shall be using for analysis

	Neighborhood	Median-HousePrice	Latitude	Longitude
0	Allandale	621300	30.340301	-97.745169
1	Barton Hills	690100	30.252021	-97.788367
2	Bouldin Creek	710300	30.251705	-97.755626
3	Brentwood	493800	30.330625	-97.732451
4	Central East Austin	483800	30.269742	-97.724153

2.3 Render the map with cleaned data to ensure it is reliable

An example of a rendered map by folium with neighborhood boundaries and markers:



2.4 Return a table of Asian restaurants near each neighborhood

- Call the FourSquare API using its explore feature
- Query a list of venues in each neighborhood in Austin defined by FourSquare API
- Set the radius of exploring to 750 meters and set the number of returned venues to maximum (100 venues per call)
- Store necessary values for analysis only which include venue names, venue location, and venue categories. Then, drop duplicate entries to avoid data overlapping when clustering in the modeling phase.
- After cleaning the dataframe, there are 316 unique categories of venues near all the neighborhoods in Austin. An example of part of the dataset returned by FourSquare API after merging and cleaning:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Allandale	30.340301	-97.745169	Barley Swine	30.341256	-97.738458	New American Restaurant
1	Allandale	30.340301	-97.745169	Yard Bar	30.342881	-97.738871	Bar
2	Allandale	30.340301	-97.745169	Lick Ice Creams Burnet Road	30.341143	-97.738408	Ice Cream Shop
3	Allandale	30.340301	-97.745169	Bufalina Due	30.341030	-97.738422	Pizza Place
4	Allandale	30.340301	-97.745169	Three Little Pigs	30.340192	-97.738426	Food Truck

3. Methodology

Now that we have collected, initially understood, and cleaned the data required. It is time to analyze and prepare data for visualization and modeling.

So far, we have initially formatted the following data:

- Dataframe df_merged contains names of Austin neighborhoods, median house prices and their geographic coordinates.
- Dataframe austin_venues contains neighborhoods with all nearby venues as well as returned values from FourSquare API

The purpose of using the data:

- Display the number of nearby venues for each neighborhood on a choropleth map
- Mark neighborhoods on the map with clusters to reveal insights to categories of venues
- Mark map with clusters to reveal housing price trends for each neighborhood in Austin.

3.1 Data preparation

- To display venue density on a choropleth map, create a dataframe venue_counts with a count of venues near each neighborhood
- Based on returned data from FourSquare API, there are 316 unique venue categories near neighborhoods in Austin.

- To reveal characteristics of nearby venues, compute the mean of each venue category. This is done by first get dummies for each category and then take a mean on occurrence. Below is part of the dataframe with the mean of each venue category:

	Neighborhood	ATM	Adult Boutique	Advertising Agency	African Restaurant	American Restaurant	Antique Shop	Arcade	Argentinian Restaurant	Art Gallery	...	Video Game Store	Video Store	Vietnamese Restaurant	Whisky Bar	Wine Bar
0	Allandale	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.00	0.00	...	0.0	0.027778	0.000000	0.0	0.00
1	Barton Hills	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.00	0.00	...	0.0	0.000000	0.000000	0.0	0.00
2	Bouldin Creek	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.00	0.00	...	0.0	0.000000	0.010000	0.0	0.00
3	Brentwood	0.0	0.0	0.0	0.0	0.027778	0.027778	0.0	0.00	0.00	...	0.0	0.013889	0.013889	0.0	0.00
4	Central East Austin	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.01	0.00	...	0.0	0.000000	0.000000	0.0	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

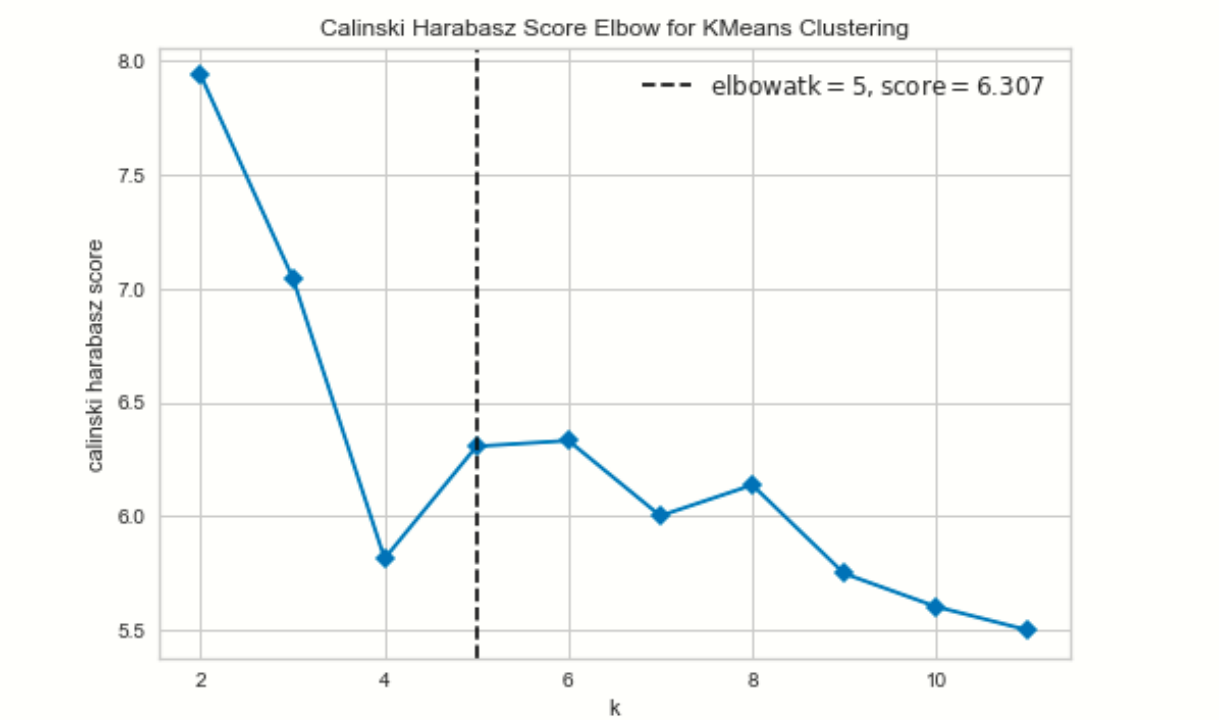
- Sort the mean occurrence in descending order and encode them for clustering based on the top 10 common venue categories per neighborhood. Part of the dataframe up to this step:

	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	Allandale	Food Truck	Pizza Place	Gym	Mexican Restaurant	Ice Cream Shop	Bus Stop	Baseball Field	Beer Bar	Supermarket	Storage Facility
1	Barton Hills	Pool	Trail	Taco Place	Other Great Outdoors	Spa	Scenic Lookout	Gym	Coffee Shop	Farmers Market	Farm
2	Bouldin Creek	Coffee Shop	Food Truck	Ice Cream Shop	New American Restaurant	Salon / Barbershop	Mexican Restaurant	Italian Restaurant	Hotel	Thai Restaurant	Gym / Fitness Center
3	Brentwood	Burger Joint	Gas Station	Mexican Restaurant	Coffee Shop	Thrift / Vintage Store	Taco Place	Pet Store	Bookstore	Thai Restaurant	Pizza Place
4	Central East Austin	Food Truck	Bar	Mexican Restaurant	Cocktail Bar	Coffee Shop	Dive Bar	Yoga Studio	Restaurant	Grocery Store	BBQ Joint

Given that the data we have is well-defined and are separated without overlapping, we can use K-Mean clustering to robustly segment neighborhoods based on their 10 most common venue categories.

3.2 Modeling and evaluation

Before moving on to clustering, we first visualized the elbow using the Calinski Harabasz method to determine an optimum number of clusters for modeling. I set the range of K from 2 to 12
The result of the elbow visualization can be seen below:



After fitting the K-Means clustering model to the dataset, we proceed towards adding clustered labels as a new column. We now have a new dataframe with their top 10 common Asian restaurant categories and their cluster labels.

	Cluster Labels	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	0	Allandale	Food Truck	Pizza Place	Gym	Mexican Restaurant	Ice Cream Shop	Bus Stop	Baseball Field	Beer Bar	Supermarket	Storage Facility
1	0	Barton Hills	Pool	Trail	Taco Place	Other Great Outdoors	Spa	Scenic Lookout	Gym	Coffee Shop	Farmers Market	Farm
2	0	Bouldin Creek	Coffee Shop	Food Truck	Ice Cream Shop	New American Restaurant	Salon / Barbershop	Mexican Restaurant	Italian Restaurant	Hotel	Thai Restaurant	Gym / Fitness Center
3	0	Brentwood	Burger Joint	Gas Station	Mexican Restaurant	Coffee Shop	Thrift / Vintage Store	Taco Place	Pet Store	Bookstore	Thai Restaurant	Pizza Place
4	0	Central East Austin	Food Truck	Bar	Mexican Restaurant	Cocktail Bar	Coffee Shop	Dive Bar	Yoga Studio	Restaurant	Grocery Store	BBQ Joint

Now we can merge this data set with the earlier dataframe containing neighborhood coordinates as well as median housing price data to arrive at a merged dataframe as depicted in the screenshot

	Neighborhood	Median-HousePrice	Latitude	Longitude	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	9 C
0	Allandale	621300	30.340301	-97.745169	0	Food Truck	Pizza Place	Gym	Mexican Restaurant	Ice Cream Shop	Bus Stop	Baseball Field	Beer Bar	Supe
1	Barton Hills	690100	30.252021	-97.788367	0	Pool	Trail	Taco Place	Other Great Outdoors	Spa	Scenic Lookout	Gym	Coffee Shop	F
2	Bouldin Creek	710300	30.251705	-97.755626	0	Coffee Shop	Food Truck	Ice Cream Shop	New American Restaurant	Salon / Barbershop	Mexican Restaurant	Italian Restaurant	Hotel	Res
3	Brentwood	493800	30.330625	-97.732451	0	Burger Joint	Gas Station	Mexican Restaurant	Coffee Shop	Thrift / Vintage Store	Taco Place	Pet Store	Bookstore	Res
4	Central East Austin	483800	30.269742	-97.724153	0	Food Truck	Bar	Mexican Restaurant	Cocktail Bar	Coffee Shop	Dive Bar	Yoga Studio	Restaurant	

We can also estimate the number of 1st Most Common Venue in each cluster. Thus, we can create a data frame which may help us to find proper label names for each cluster as in the example below

	Cluster Labels	1st Most Common Venue	Counts
0	0	American Restaurant	1
1	0	Burger Joint	1
2	0	Café	1
3	0	Coffee Shop	5
4	0	Cosmetics Shop	1
5	0	Discount Store	1
6	0	Fast Food Restaurant	1
7	0	Food Truck	5
8	0	Gym	1
9	0	Hotel	4
10	0	Italian Restaurant	1
11	0	Mexican Restaurant	1
12	0	Park	2
13	0	Pool	1
14	0	Rental Car Location	1
15	0	Sandwich Place	2
16	0	Sporting Goods Shop	1
17	1	Sporting Goods Shop	1
18	2	American Restaurant	1

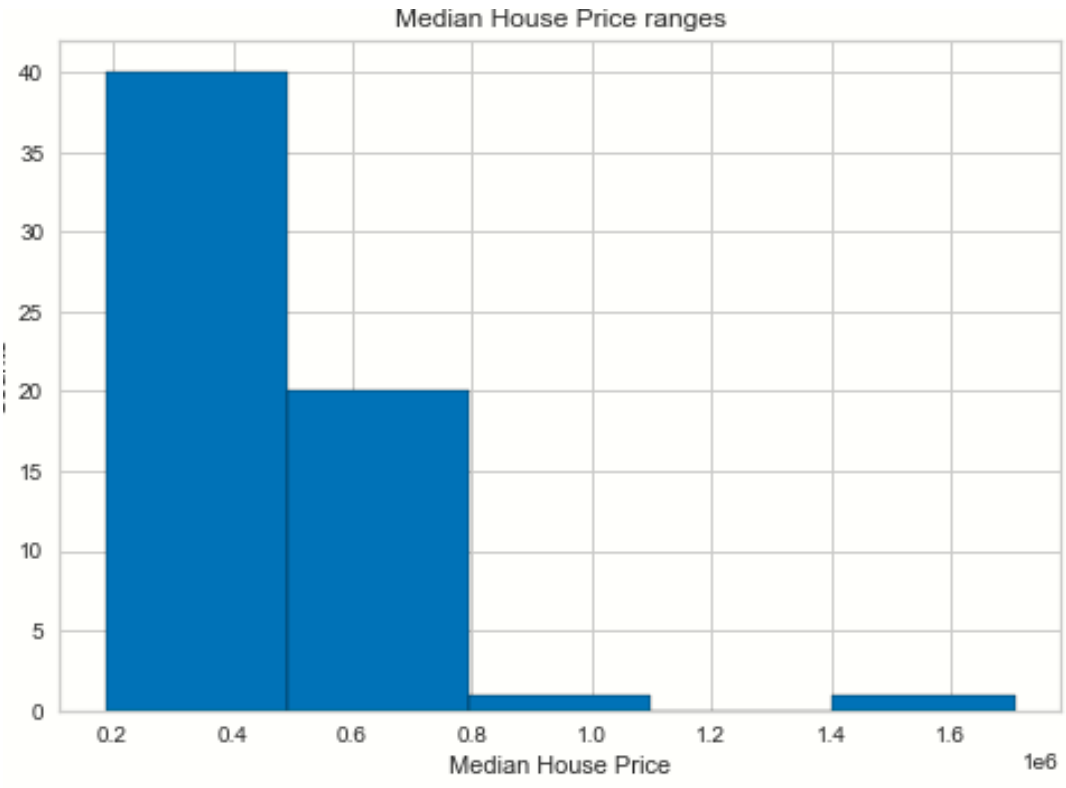
When we examine above data, we can label each cluster as follows:

- Cluster 0: "Food venues and misc social venues"
- Cluster 1: "Sporting Goods"
- Cluster 2: "Stores and residential venues"
- Cluster 3: "Park"
- Cluster 4: "Gym"

We can now assign those new labels to existing label of clusters:

	Clusters	Labels
0	0	Food venues and misc social venues
1	1	Sporting Goods
2	2	Stores and residential venues
3	3	Park
4	4	Gym

One of our goals is to map housing prices for each neighborhood. We can divide our dataset into five bins and visualize using a histogram



From the above histogram, we can define the Median House Price ranges as below:

< $200,000: "Very Low"
$200,000 - $400,000: "Low"
$400,000 - $600,000: "Medium"
$600,000 - $800,000: "Medium-High"
$800,000 - $1,000,000: "High"
> $1,000,000 USD: "Very High"

	Neighborhood	Median-HousePrice	Cluster Labels	Ranges_labels
0	North Burnett	187600	0	Very Low
1	Franklin Park	237500	2	Low
2	Pleasant Valley	239400	0	Low
3	Mckinney	246900	1	Low
4	Southeast	259300	4	Low

We can now show the top three venues for each neighborhood on the map. To begin with we group each neighborhood by the count and type of each of the top 3 venues

	Neighborhood	Venue text
0	Allandale	3 Food Truck, 2 Gym, 2 Mexican Restaurant
1	Barton Hills	2 Pool, 2 Trail, 1 Coffee Shop
2	Bouldin Creek	6 Coffee Shop, 6 Food Truck, 4 Ice Cream Shop
3	Brentwood	5 Burger Joint, 3 Coffee Shop, 3 Gas Station
4	Central East Austin	14 Food Truck, 9 Bar, 6 Mexican Restaurant

4. Results
4.1 Examine each cluster

First, we examine each cluster to see the distribution of neighborhoods and observe the top 3 most common venues for neighborhoods in the same cluster.

Example below

Cluster 1:

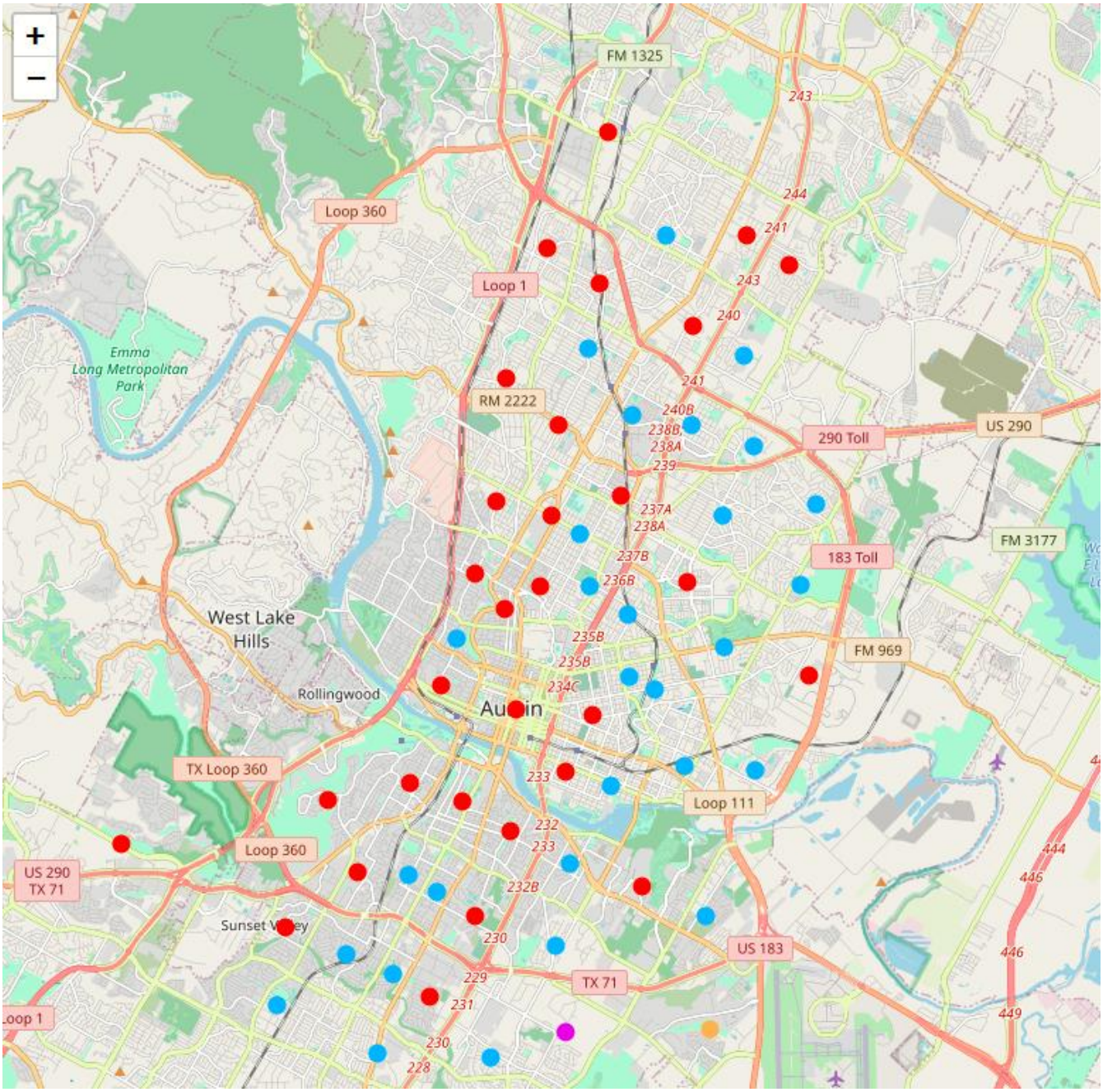
	Neighborhood	Median-HousePrice	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	Allandale	621300	0	Food Truck	Pizza Place	Gym	Mexican Restaurant	Ice Cream Shop	Bus Stop	Baseball Field	Beer Bar	Supermarket
1	Barton Hills	690100	0	Pool	Trail	Taco Place	Other Great Outdoors	Spa	Scenic Lookout	Gym	Coffee Shop	Farmers Market
							New					

Cluster 2:

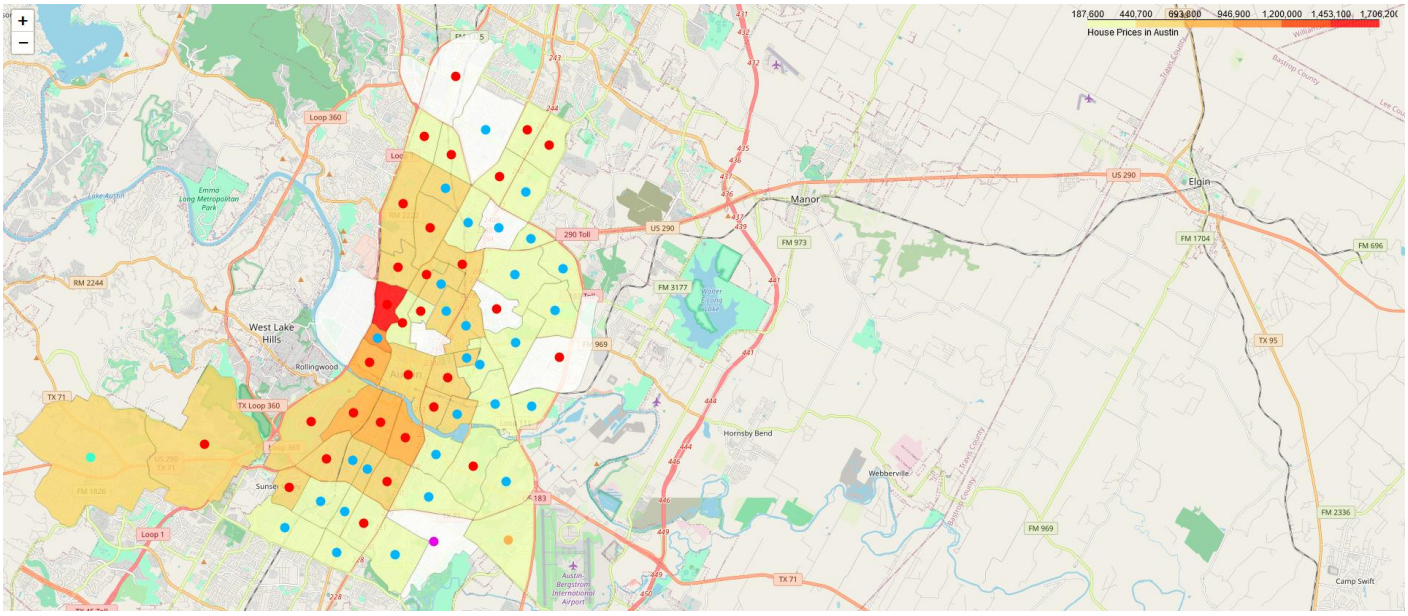
	Neighborhood	Median-HousePrice	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Venue text
26	Mckinney	246900	1	Sporting Goods Shop	Construction & Landscaping	Sandwich Place	Farm	Escape Room	Ethiopian Restaurant	Event Service	Event Space	Eye Doctor	Factory	2 Sporting Goods Shop, 1 Construction & Landsc...

4.2 Render a choropleth map with clustered markers that describes:

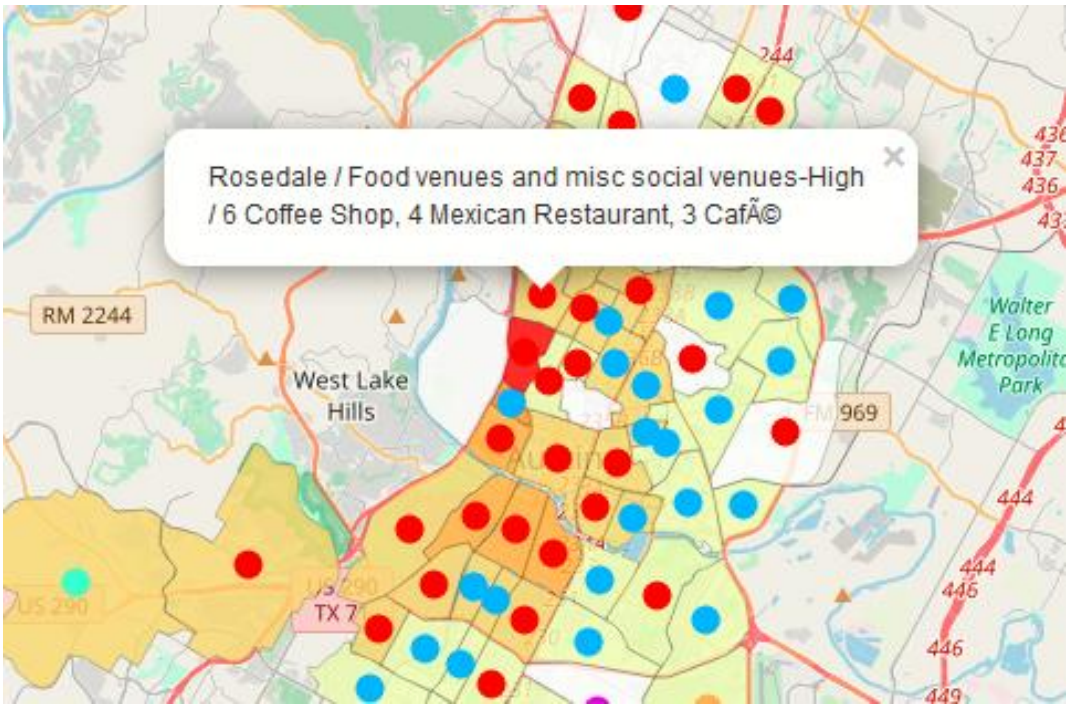
The distribution of venues among neighborhoods with different coloring for each cluster of neighborhoods



Now we can proceed to overlay our housing price color gradient on top of the above map to give us a more accurate picture of our goal- Housing mart and venue analysis for each Austin Neighborhood



The popup tooltip for each neighborhood centroid gives us information on the Neighborhood being analyzed, cluster category, range of Housing Prices and counts for the top 3 venue categories in that neighborhood



4. Discussion

Austin is one of the fastest growing cities in the United States with a burgeoning population. As such the population densities across the various neighborhoods can vary. Due to the high complexity several different clustering or classification methods can be used.

In my study I used the K-means clustering algorithm. On testing for the Elbow using the Calinski Harabasz Method we can arrive at an optimal k value of 5. This study was conducted using open source datasets that may be incomplete or outdated. Furthermore, due to certain mismatches in both the data sets a small number of neighborhoods were excluded from analysis. For more accurate results, a more recent dataset can be used which will result in more accurate results. Also, more details on each neighborhood can be explored.

We also explored the median house prices for each neighborhood to further analyze which neighborhood would offer a good combination of affordability as well as being close to social venues which would contribute to the attractiveness for any potential buyer. In future studies this data can be sourced from more recent data.

I ended the study by examining the neighborhoods in each cluster and plotting the results on a map of Austin. This depicts the venues in each neighborhood as well as the approximate cost of purchasing a house. This can assist both future investors and future residents.

5. Conclusion

As more people move to Austin, TX the city is experiencing a real estate boom along with a rapidly expanding social scene. Investors, residents and maybe even city planning officials can benefit greatly from access to results from data analysis.

5. References:

[1]. Austin, Texas, Wikipedia. Retrieved from: https://en.wikipedia.org/wiki/Austin,_Texas
[2] Texas- Hottest Real Estate Market in 2020? Retrieved from: https://www.marketcurrentswealthnet.com/features/texas-hottest-real-estate-market-in-2020/
[3]. Boundaries: Austin Neighborhood Planning Areas, data.austintexas.gov, the official City of Austin open data portal. Retrieved from: https://data.austintexas.gov/dataset/Boundaries-Austin-Neighborhood-Planning-Areas/nz5f-3t2e/data
[4]. Austin Home Prices & Values,Zillow. Retrieved from https://www.zillow.com/austin-tx/home-values/
[5] Queried venue data for each neighbood using the foursquare API, FourSquare API. Retrieved from: https://developer.foursquare.com/docs/api-reference/venues/explore/