Viktoriia Slaikovskaia

July 2020

Austin neighborhoods analysis

1 Introduction

Austin is the capital city of the US state of Texas, as well as the seat and largest city of Travis County, with portions extending into Hays and Williamson counties. It is the 11th-most populous city in the United States, the fourth-most-populous 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. It was the southernmost state capital in the contiguous United States.

1.1 Problem

Austin is an excellent example of a growing city and a unique model for relocation purposes. Sometimes it isn't easy to choose between neighbourhoods, especially if you aren't familiar with the city.

This project aims to cluster neighbourhoods by venues, crime rate, and average rent, which could help people choose between them.

1.2 Audience

This project could help those relocating to Austin or someone who needs to know more about their city.

The project will give the following information to that group of people:

- Which neighbourhoods have a high crime rate? It could be useful for someone searching for a peaceful community.
- What regions to consider to relocation by their budget?
- What neighbourhoods have venues that you like?

2. Data

2.1 Data sources and description

As the problem implies having at least 3 data sets about crimes, venues, and rent, we should find and practice on it.

Sources of data:

- Austin neighborhoods geojson, taken from GitHub, will help to create future choropleth maps for visualization purposes. It consists of primary information about neighborhoods: geometry and name.
- Criminal reports from the Austin government website. It consists of general information about crimes, mainly we need a type of crime, coordinates, and time. I'm going to use this data for crime rate segmentation showing the crime rate in the areas. It would provide us with enough information to create a choropleth map.
- Average rent per house I will scratch from here and some famous sites for renting properties. Data consists of prices, bedrooms for rent per home. That will help to understand how the cost distributed by arias.
- Foursquare API will give information about venues around each area. Data consists of venue names, coordinates, and venue category. I'm going to use this data to create clusters by their similarity and understand what we could find in each neighborhood.

2.3. Data cleaning

The main problem with the crime data set is overloaded and missing information.

At the beginning of data cleaning, data has 2242728 records.

Many columns need to be dropped by not requiring usage, for example:

Incident Number - we do not need this one because we have pandas unique index.

Occurred Date Time, Report Date Time - columns with the date and time apart is enough.

APD Sector, APD District, PRA - not informative for us, because we need only neighborhood data.

Clearance Status, Clearance Date - occurred date is enough.

Address, Zip Code, Council District, Census Tract, X-coordinate, Y-coordinate - latitude, and longitude is enough for our purposes.

We have to change value types in data and time columns, as I need recent information about what's going on in every region in the last five years, I will use dates after the year 2015. After reducing the data, only 594803 records left.

As we require to know where the crime occurred, we demand to know coordinates data—so the right choice to drop lines without longitude and latitude. 12239 rows were without coordinates.

To know what area crime appears, we will need to create a point with a crime coordinate and check if any of these areas contain this point; after looping crime data to function for finding the name of a neighborhood by location, we will have a column with neighborhood names.

In our data set, 406687 crimes committed in Austin and 175877 outside. After we located neighborhood names for every crime, we should reduce rows with crime happened outside Austin.

3. Methodology

In this project, we have four main goals:

1. Bin serious crimes into six categories. For this, we have to create a data frame with an only part I crimes.

2. Bin rent prices into three categories: low, medium, and high-priced neighborhoods.

3. Cluster venues by their similarity. Foursquare API will help me to create a data frame, and the K-means method is used for clustering.

4. Cluster neighborhoods using all 3 data sources above. Elbow method is used to find an optimal number of clusters and K-means for clustering.

3.1 Crime data

Crime data consist of many features, but the main one is the number of crimes. Creating a top 10 neighborhoods will help to find safe and unsafe zones; it's going to be useful for whom the only interest is a peaceful area.

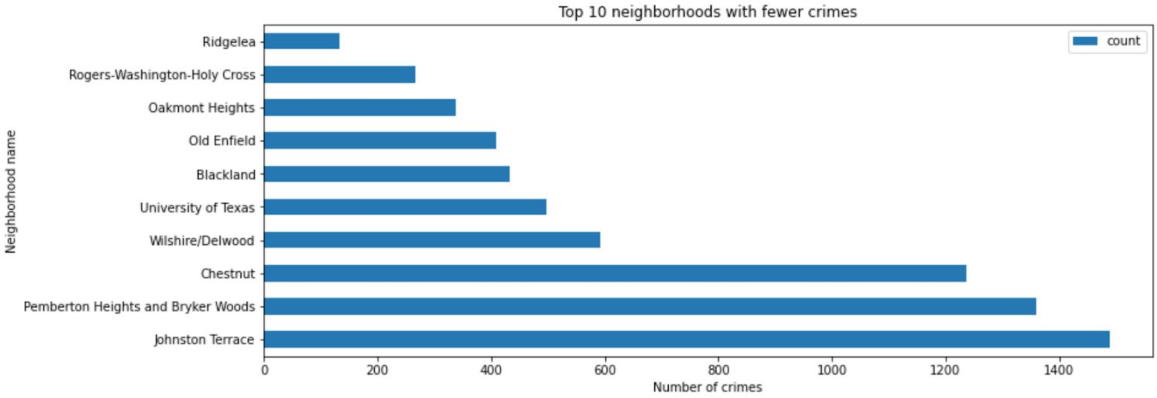


Figure 1. Top 10 neighborhoods with fewer crimes.

Top 10 quiet zones. In the second rent section, we could see how much is the calm life costs.

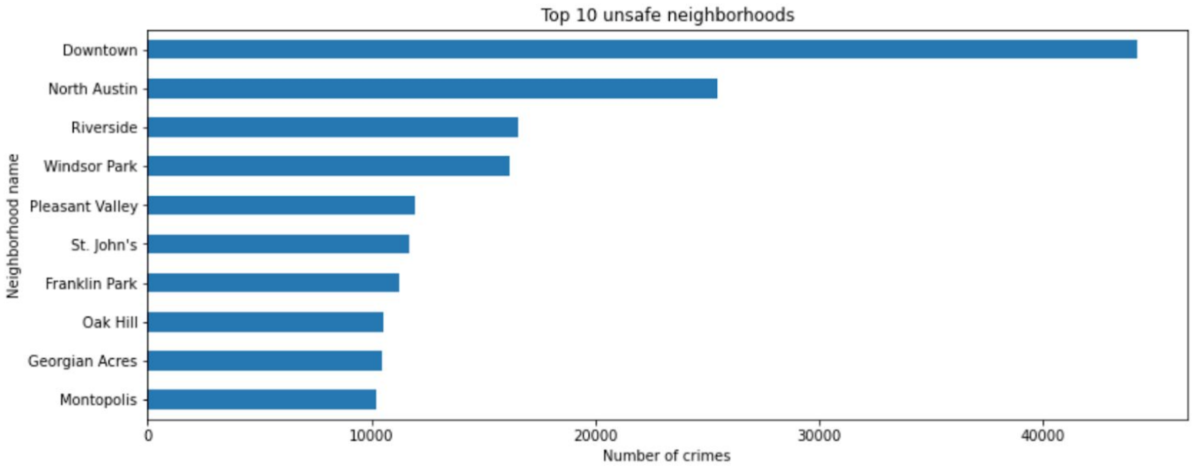


Figure 2. Top unsafe 10 neighborhoods

The Downtown shows a significant number of crimes; second place is a North Austin.

In conclusion, you could see crime distribution on the map. Downtown and North Austin are easy to find.

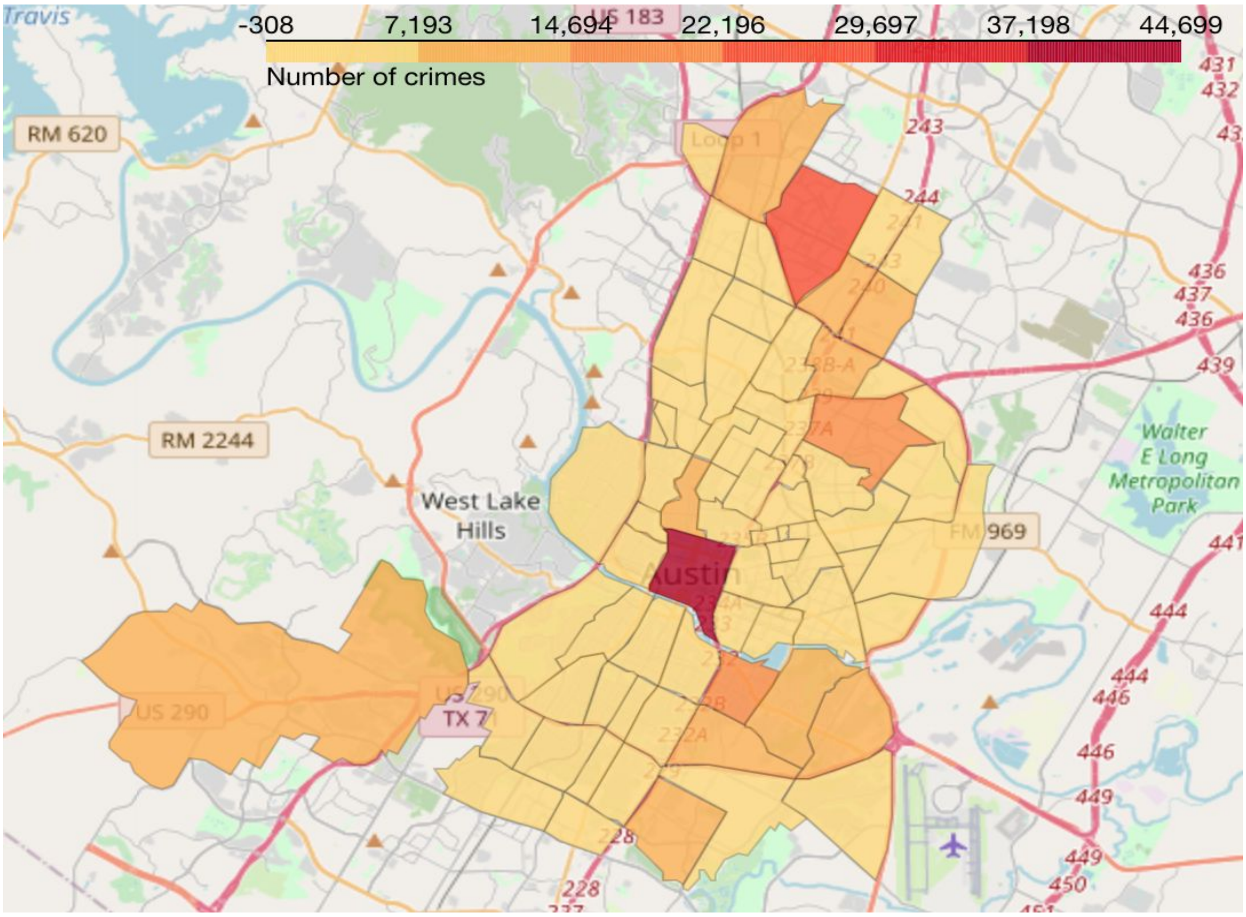


Figure 3. Choropleth map of the distribution of crimes by areas

3.1.1 Crimes after dark

One more significant mark is the number of crimes at night, showing the nightlife area for this purpose where chosen crimes appeared after 8 p.m till 6 a.m.

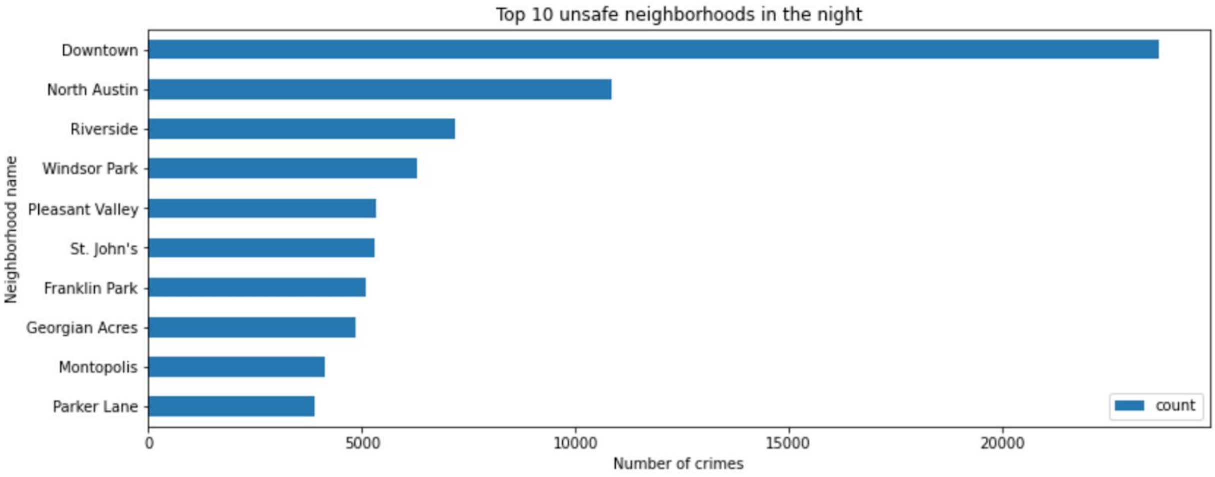


Figure 4. Top 10 unsafe neighborhoods in the night.

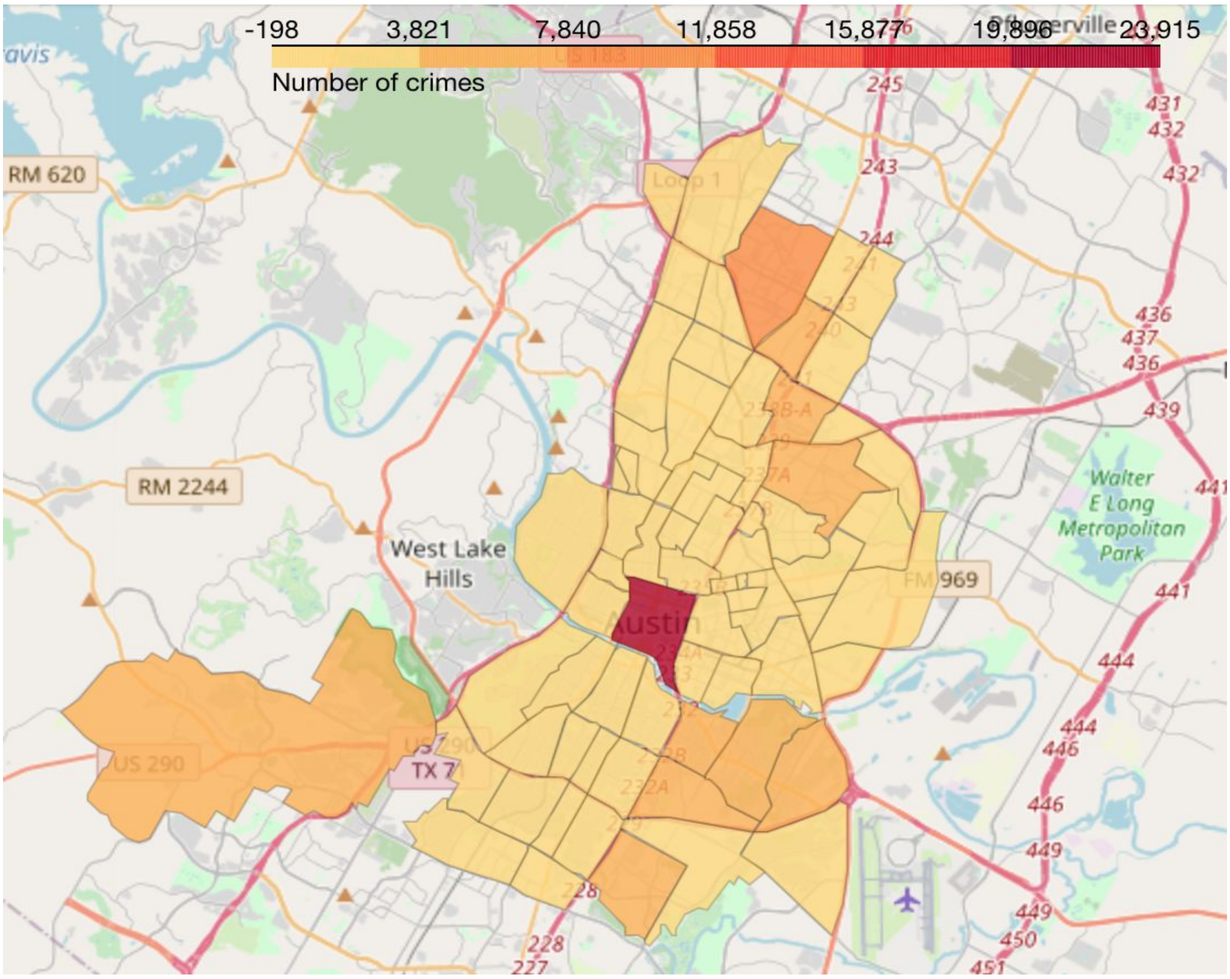


Figure 5. Choropleth map of the distribution of night crimes by areas

As we can see, again, Downtown and North Austin are the top of the unsafe neighborhoods. I wouldn't recommend those neighborhoods for people with family.

3.1.2 Family violence

Family violence occurs across the world, in various cultures, and affects people across society at all levels of economic status. However, indicators of lower socioeconomic status such as unemployment and low income are risk factors for higher domestic violence levels in several studies. By knowing areas with high rate family violence, we could suggest which regions have a lower socioeconomic status.

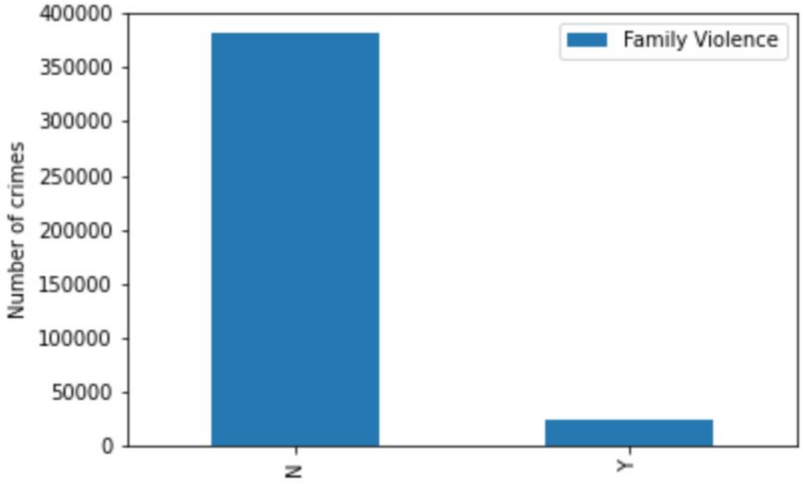


Figure 6. Number of family violence crimes in a proportion of all crimes

Percent of family violence is 6.21 from all crimes that happened in Austin.

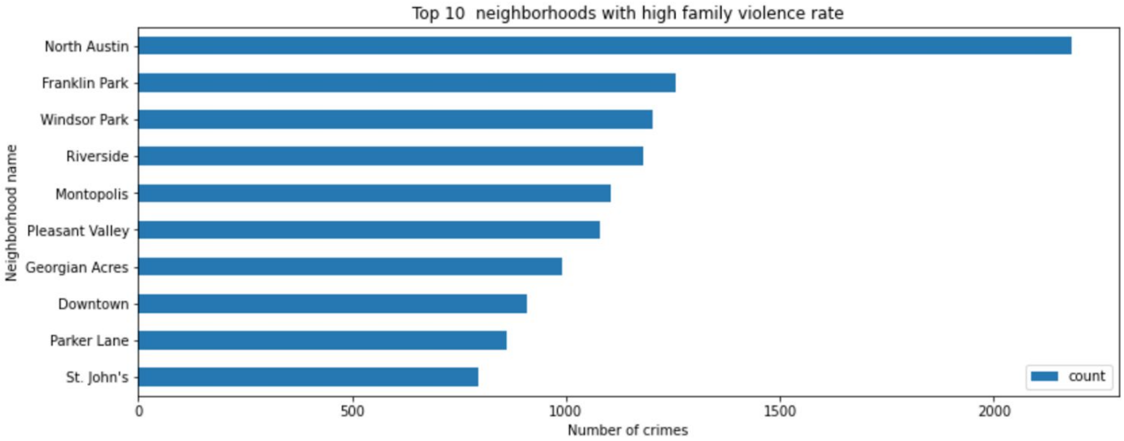


Figure 7. Top 10 neighborhoods with high family violence.

Downtown is only in 8th place, but North Austin is on the top.

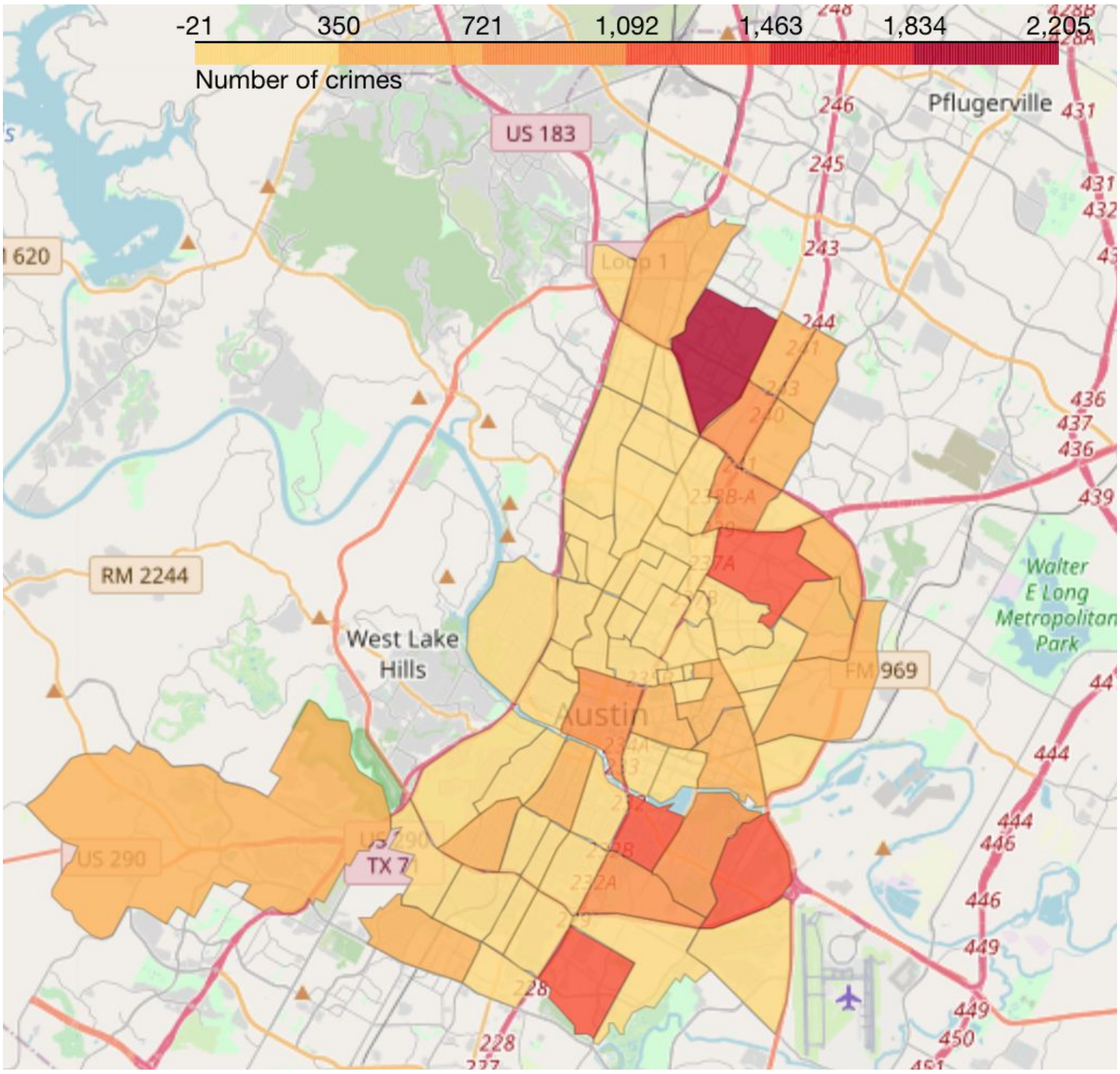


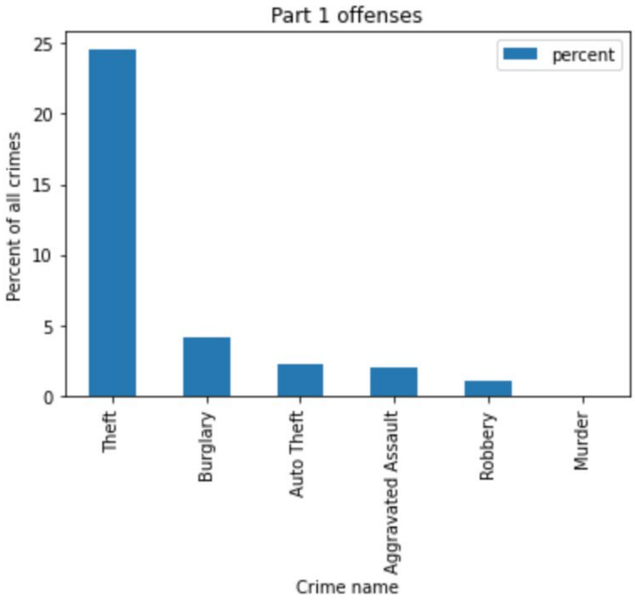
Figure 8. Choropleth map of the family violence crimes by neighborhoods

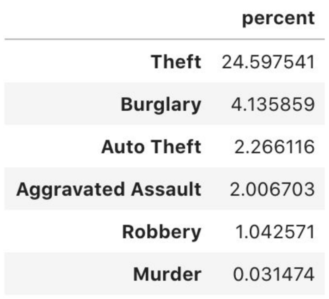
3.1.3 Part I offenses

In the traditional Summary Reporting System (SRS), there are eight crimes, or Part I offenses, murder, and nonnegligent homicide, rape, robbery, aggravated assault, burglary, motor vehicle theft, larceny-theft, and arson to be reported to the UCR Program. These offenses were chosen because they are serious crimes, they occur with regularity in all areas of the country, and they are likely to be reported to the police.

I want to use only these offenses data in my next steps because they are serious crimes. Data consists of topics described above, like family violence and crimes after dark. It's more presentable than all others.

At first, I want to know which categories data consists of.





	percent
Theft	24.597541
Burglary	4.135859
Auto Theft	2.266116
Aggravated Assault	2.006703
Robbery	1.042571
Murder	0.031474

Figure 9. The percentage ratio of part I offenses.

In a graph, the more significant part of those offenses is a theft category, almost 25 percent of all crimes.

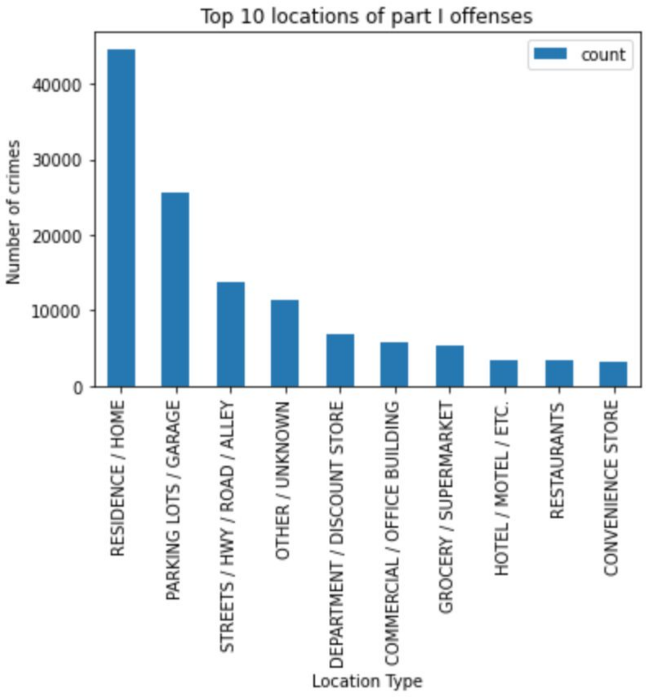


Figure 10. Top 10 locations severe crime happened.

From graph #10, we could see that crimes occurred mainly at the residences and homes.

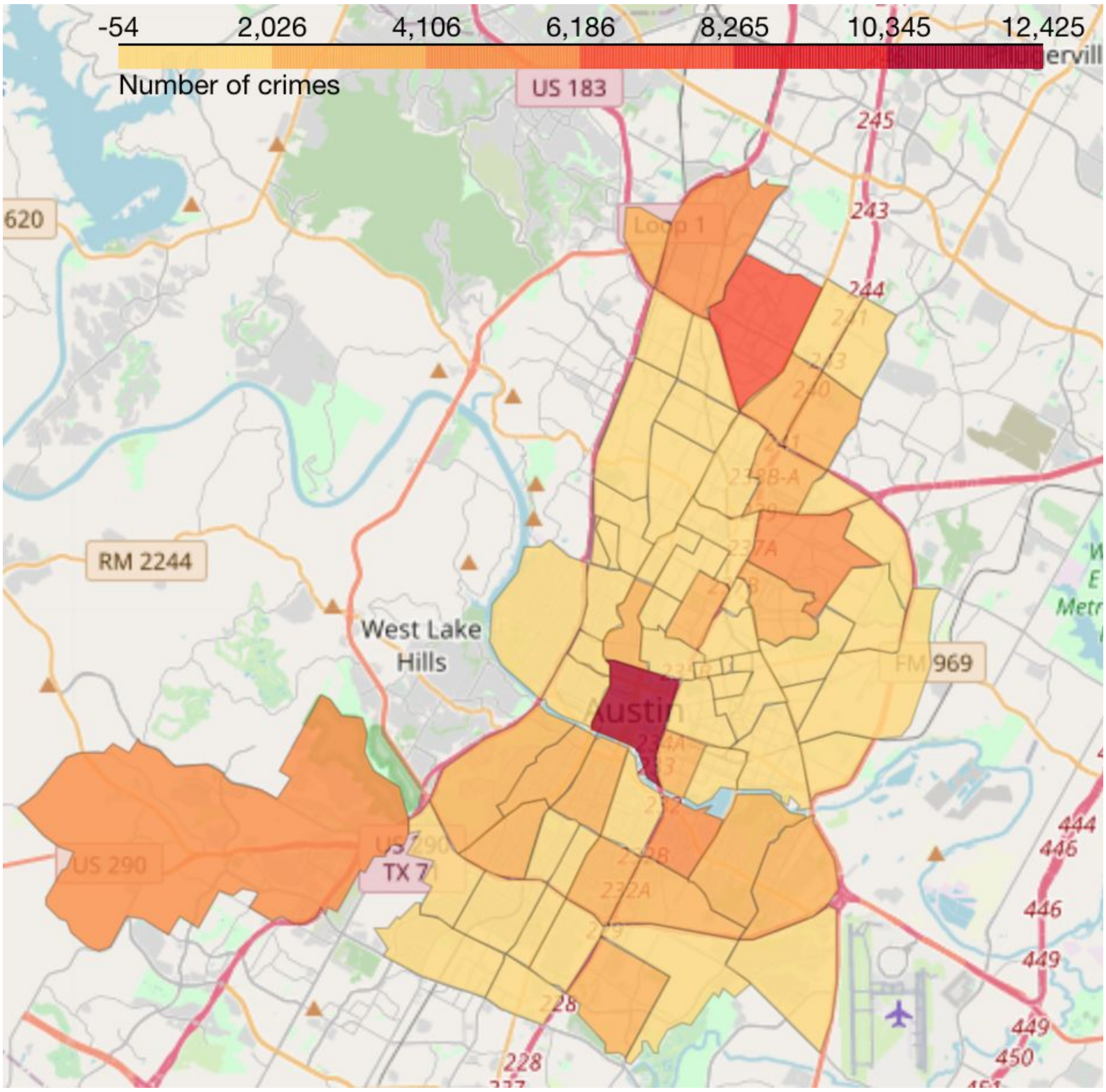


Figure 11. Choropleth map of the severe crimes by neighborhoods

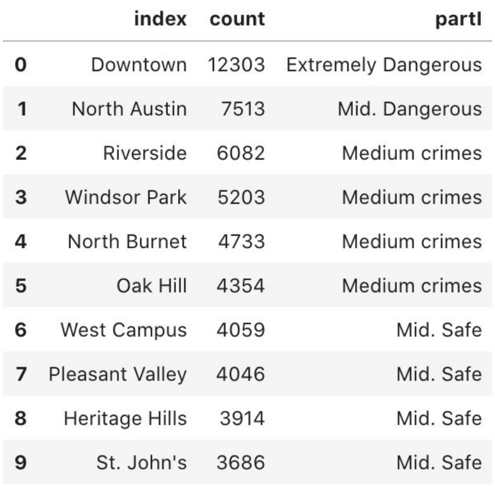
3.1.4 Bin Part I offenses

Binning crimes into categorical data will show us the crime rate in every neighborhood.

I created bins of severe crimes by cutting data into six categories:

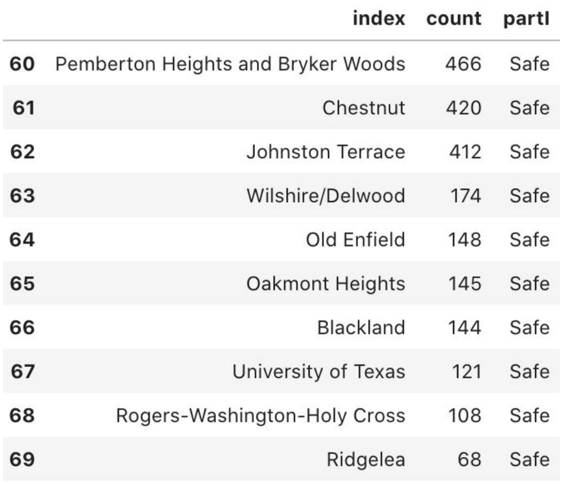
Safe, Medium Safe, Medium crimes, Medium Dangerous, Dangerous, Extremely Dangerous

This data will be used for the last neighborhood clustering.



	index	count	partl
0	Downtown	12303	Extremely Dangerous
1	North Austin	7513	Mid. Dangerous
2	Riverside	6082	Medium crimes
3	Windsor Park	5203	Medium crimes
4	North Burnet	4733	Medium crimes
5	Oak Hill	4354	Medium crimes
6	West Campus	4059	Mid. Safe
7	Pleasant Valley	4046	Mid. Safe
8	Heritage Hills	3914	Mid. Safe
9	St. John's	3686	Mid. Safe

Figure 12. Top 10 dangerous neighborhoods



	index	count	partl
60	Pemberton Heights and Bryker Woods	466	Safe
61	Chestnut	420	Safe
62	Johnston Terrace	412	Safe
63	Wilshire/Delwood	174	Safe
64	Old Enfield	148	Safe
65	Oakmont Heights	145	Safe
66	Blackland	144	Safe
67	University of Texas	121	Safe
68	Rogers-Washington-Holy Cross	108	Safe
69	Ridgelea	68	Safe

Figure 13. Top 10 safe neighborhoods

By confronting the first tops 10 neighborhoods, I could say:

The four places in unsafe areas still keep by the same regions: Downtown, North Austin, Riverside, Windsor Park. Pleasant Valley from 5th place went to 7th.

Rigelea and Rogers Washington Holy Cross still on top of safe places. The University of Texas went from 6th to 3rd, Blackland from 5th to 4th, Oakmont Heights went from 3rd to 5th place, and others just change places inside the top 10. The content of the safe category didn't change, only positions.

3.2 Rent data

Beneficial information could be a distribution of cost dependent on the number of rooms. When you have a certain amount of money what house you could allow yourself. I made a graph showing this relationship between rent and the number of rooms.

Another useful information just to know the general distribution of prices in every neighborhood. For choosing zones are suitable for your budget.

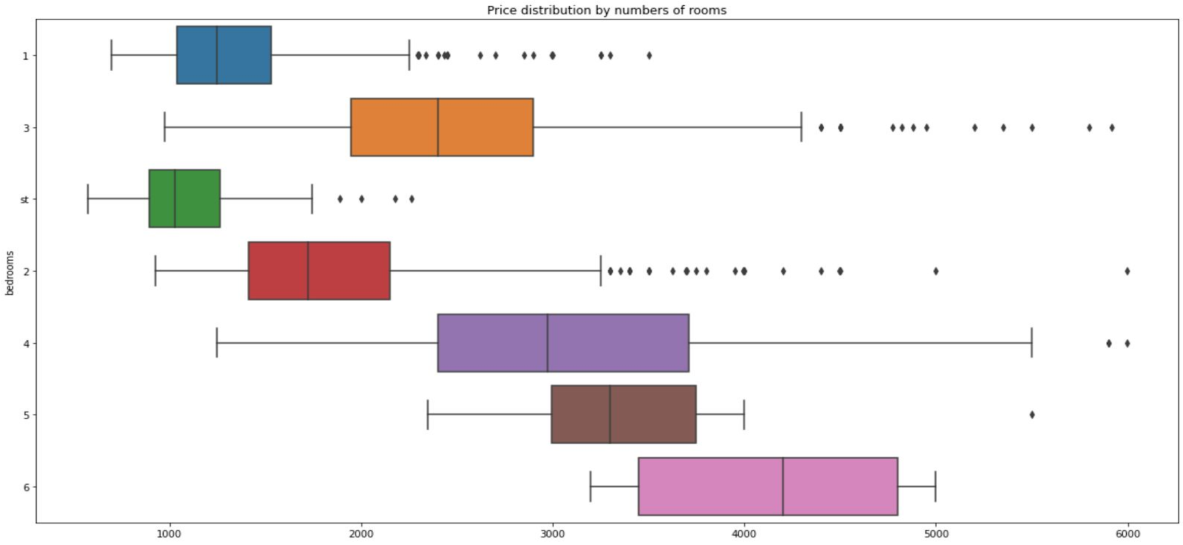


Figure 14. Price distribution by number of rooms

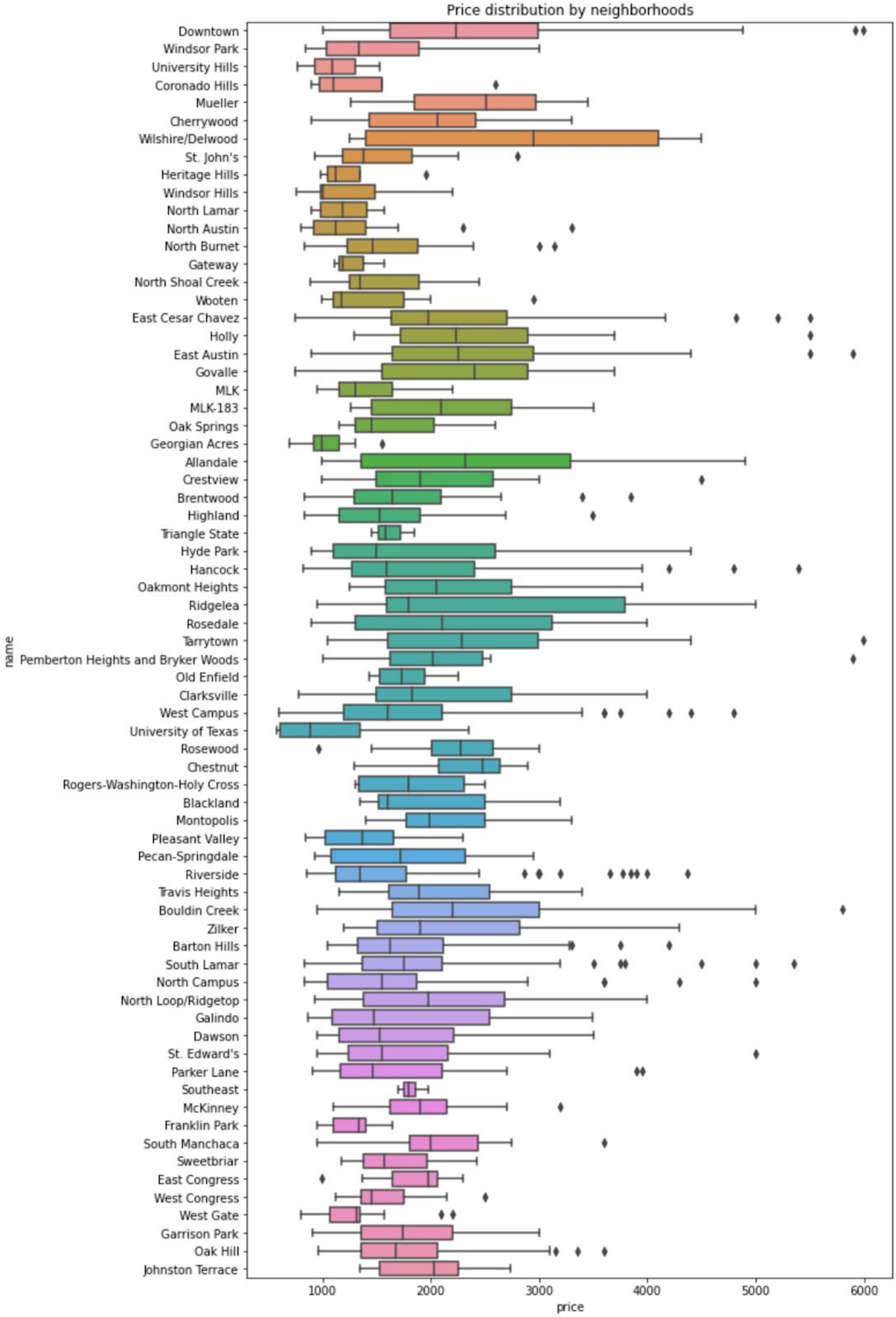


Figure 15. Price distribution by neighborhoods

Binning rent prices into categorical data shows us the median price rate in every neighborhood. For calculating, I used the median statistic function.

I created bins of rent costs by cutting data into three categories:

High price, Medium price, and Low price.

This data will be used for the last neighborhood clustering.

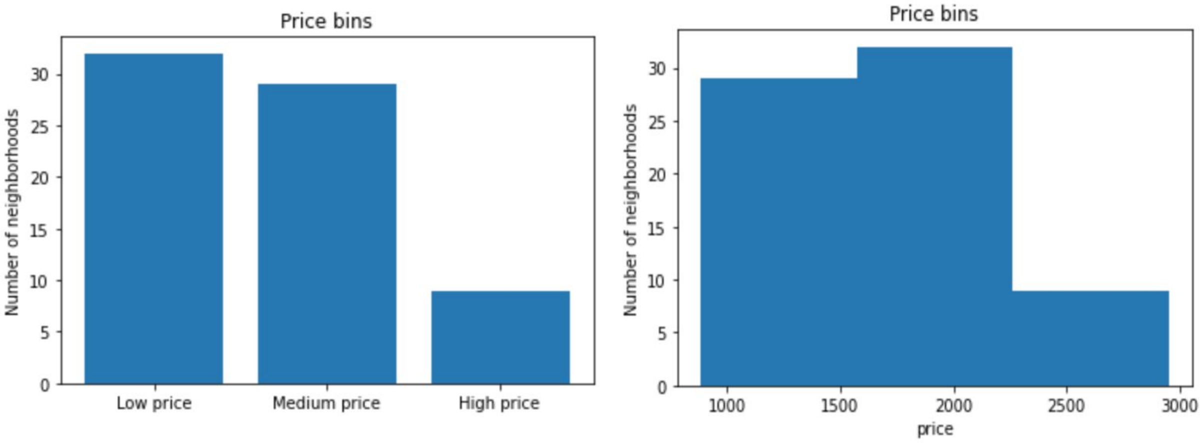
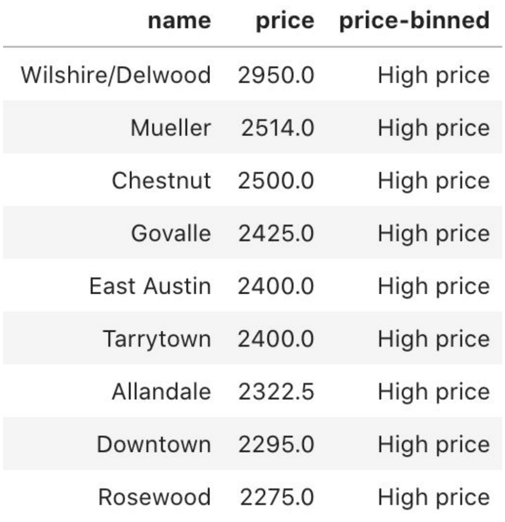


Figure 16-17. Price distribution

After segmentation rent cost, we got 32 medium priced neighborhoods, 29 low priced, and 9 high priced.

Let's look at High priced neighborhoods.

Even Downtown has a hazardous crime rate; it has quite high prices for apartments.



name	price	price-binned
Wilshire/Delwood	2950.0	High price
Mueller	2514.0	High price
Chestnut	2500.0	High price
Govalle	2425.0	High price
East Austin	2400.0	High price
Tarrytown	2400.0	High price
Allandale	2322.5	High price
Downtown	2295.0	High price
Rosewood	2275.0	High price

Figure 18. Hight priced neighborhoods

I created a choropleth rent map with popup labels with neighborhood names and the median price for the convenience of understanding.

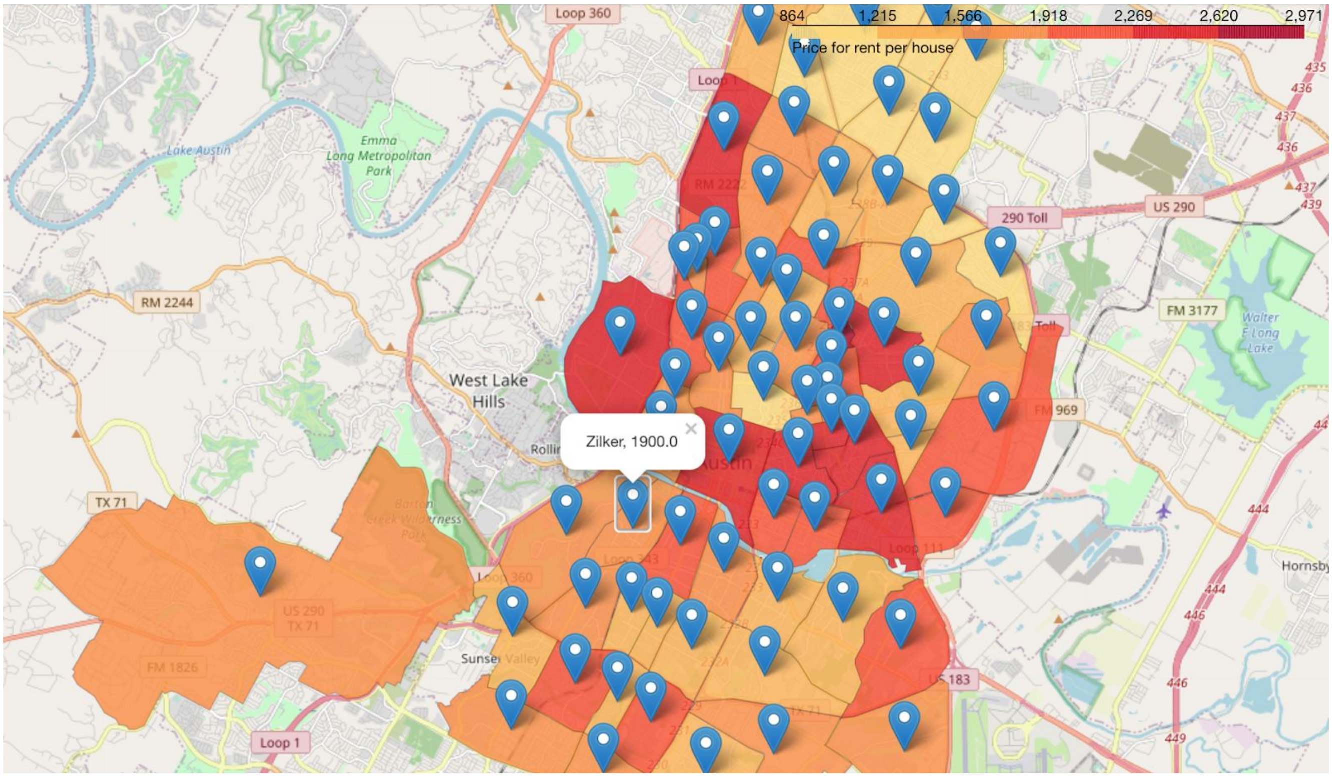


Figure 19. Choropleth rent map

3.3 Venue data

3.3.1 Exploring venues data

I applied the Foursquare API to explore the neighborhoods and segment them. I created a function to get venue categories and made a data frame limiting venue number by 100, in a radius of 1000 meters from each neighborhood center.

After running all operations in the data frame, it turned out to be 323 unique categories.

3.3.2 Clustering

I'm going to use venue data to create clusters by their similarity and understand what we could find in each neighborhood.

I will use unsupervised learning particularly the K-means algorithm to cluster the districts with five clusters. K-Means algorithm is one of the most common cluster methods of unsupervised learning.

Let's look at the map we have at the result.

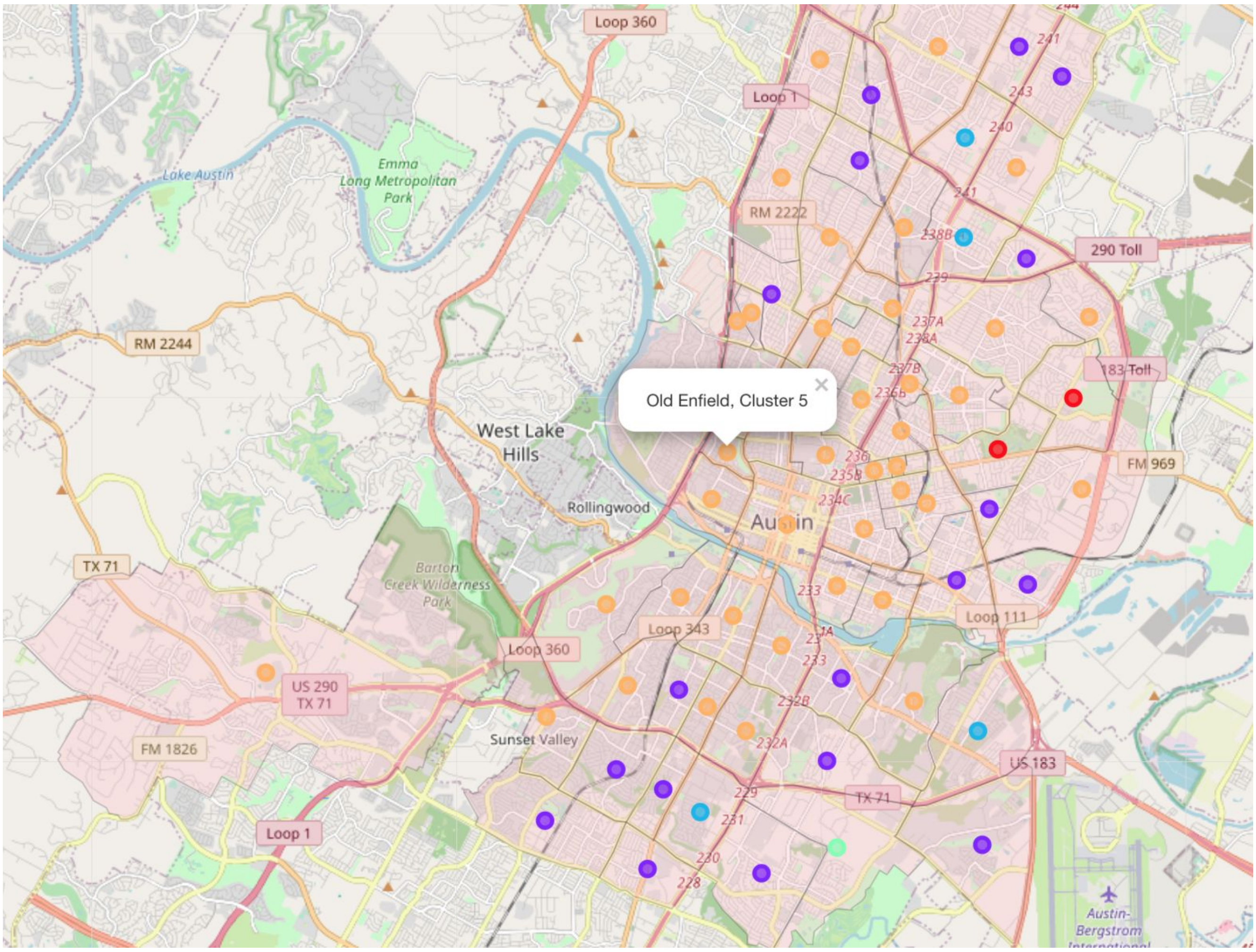


Figure 20. Choropleth map with clustered neighborhoods by venues similarity

3.3.3 Give names for every neighborhood

For understanding how to name every cluster, I will use a wordcloud visualization library for visual presentation of prevalent venues in it.

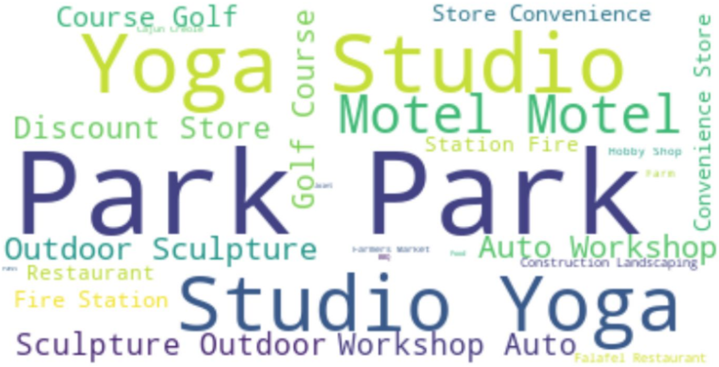


Figure 21. Cluster #1 word cloud

Cluster #1 Most popular venues appear to be park zones and sports venues. These neighborhoods would be suitable for someone who loves outdoor activities and love sports.

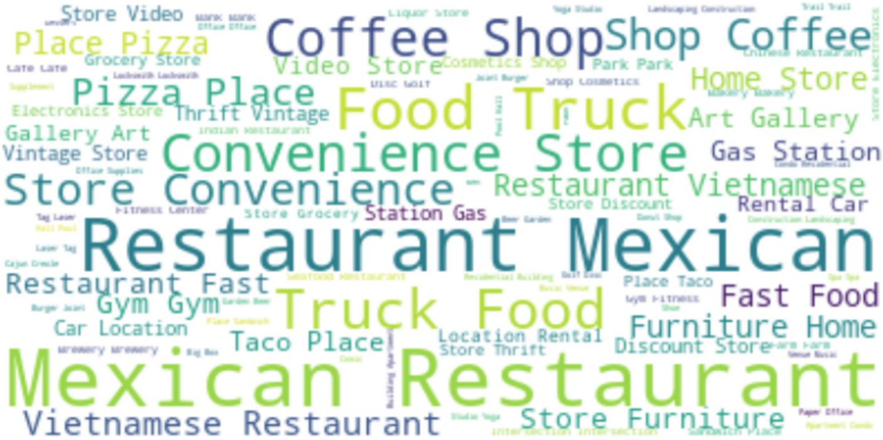


Figure 22. Cluster #2 word cloud

Cluster #2 Most popular venues appear to be multi-social venues. In these neighborhoods, you could find plenty of places you’d like. These neighborhoods would be suitable for someone who loves to go to different places to meet with friends.

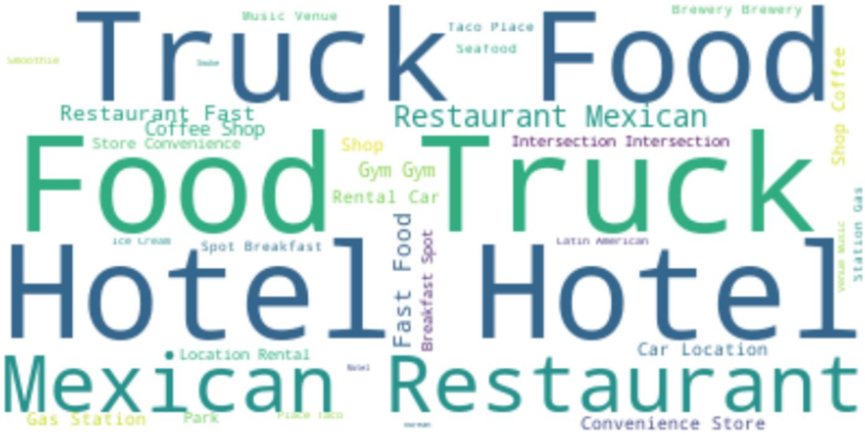


Figure 23. Cluster #3 word cloud

Cluster #3 Most popular venues here are Food Trucks and Hotels. Food trucks are the hallmark of the city. These neighborhoods would be suitable for someone who loves good food or came to Austin for a short period.

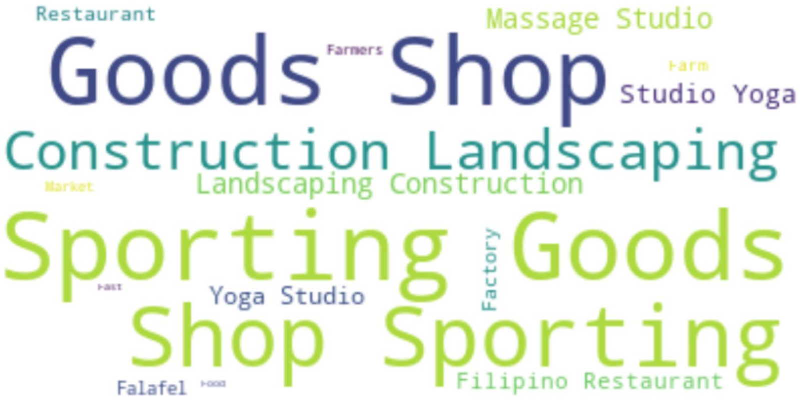


Figure 24. Cluster #4 word cloud

Cluster #4 consists of 1 neighborhood. Sporting goods shop the main venue here. Right area if you don't like many people gathering around.

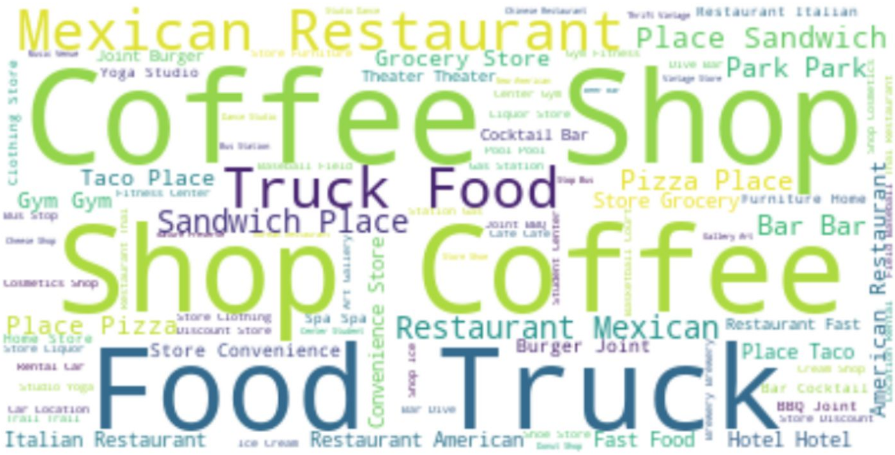


Figure 25. Cluster #5 word cloud

Cluster #5 Most popular venues appear to be coffee and food venues. These neighborhoods would be suitable for someone who can't imagine his life without coffee and love to visit food places.

3.4 Clustering neighborhoods

For the last clustering, we will use 3 data frames: median price neighborhood - data frame with a categorical median price for every region; offenses - data frame with categorical data about the part I offenses that are more severe than usual crime data; venue clusters.

To apply a machine learning algorithm, we should convert categorical data into numerical using one-hot encoding.

I will use the K-Means method because it’s an unsupervised learning method. We only want to try to investigate the structure of the data by grouping the data points into distinct subgroups.

3.4.1 Choosing the number of clusters

Before choosing the number of clusters, I would like to find an optimal amount for my purpose using the elbow method.

The KElbowVisualizer implements the "elbow" method to help data scientists select the optimal number of clusters by fitting the model with a range of values for K. If the line chart resembles an arm, then the "elbow" (the point of inflection on the curve) is a good indication that the underlying model fits best at that point. In the visualizer, "elbow" will be annotated with a dashed vertical line.

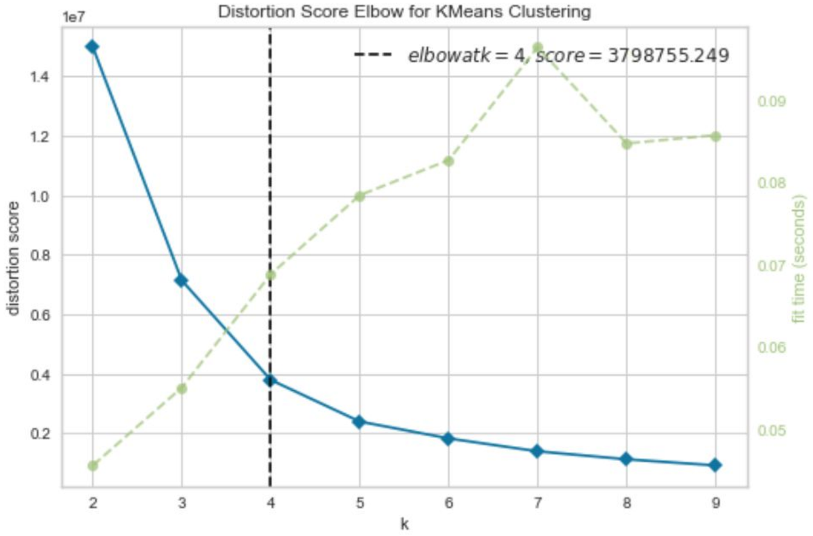


Figure 26. Distortion score for choosing optimal k

By running the elbow method, we could see that the optimal number of clusters is 4.

3.4.2 Clustering

After running the K-mean method, I got 4 clusters:

The first cluster consists of 14th neighborhoods:

Georgian Acres, Heritage Hills, Pleasant Valley, St. John's, St. Edward's, West Campus, Hancock, Barton Hills, South Lamar, Zilker, Montopolis, East Cesar Chavez, Bouldin Creek, Mueller.

The second cluster consists of 19th neighborhoods:

University of Texas, University Hills, North Austin, Gateway, West Gate, Windsor Park, North Shoal Creek, North Burnet, Hyde Park, Highland, Dawson, North Campus, Rosewood, Downtown, Allandale, East Austin, Tarrytown, Chestnut, Wilshire/Delwood.

The third cluster consists of 19th neighborhoods:

Windsor Hills, Coronado Hills, Wooten, North Lamar, MLK, Franklin Park, Riverside, Oak Springs, West Congress, Parker Lane, Galindo, Sweetbriar, Garrison Park, Southeast, Crestview, South Manchaca, Johnston Terrace, Rosedale, Govalle.

The fourth cluster consists of 18th neighborhoods:

Triangle State, Blackland, Brentwood, Oak Hill, Pecan-Springdale, Rogers-Washington-Holy Cross, Ridgeley, Old Enfield, Clarksville, Travis Heights, McKinney, North Loop/Ridgetop, East Congress, Pemberton Heights and Bryker Woods, Oakmont Heights, Cherrywood, MLK-183, Holly.

On the map (figure 27), you can see different colored circle markers showing belongings to a particular cluster.

4. Results

In the result section, I created a choropleth rent map, with popup label with the following information:

- Neighborhood name,
- Cluster,
- Rent price category,
- Common venue categories

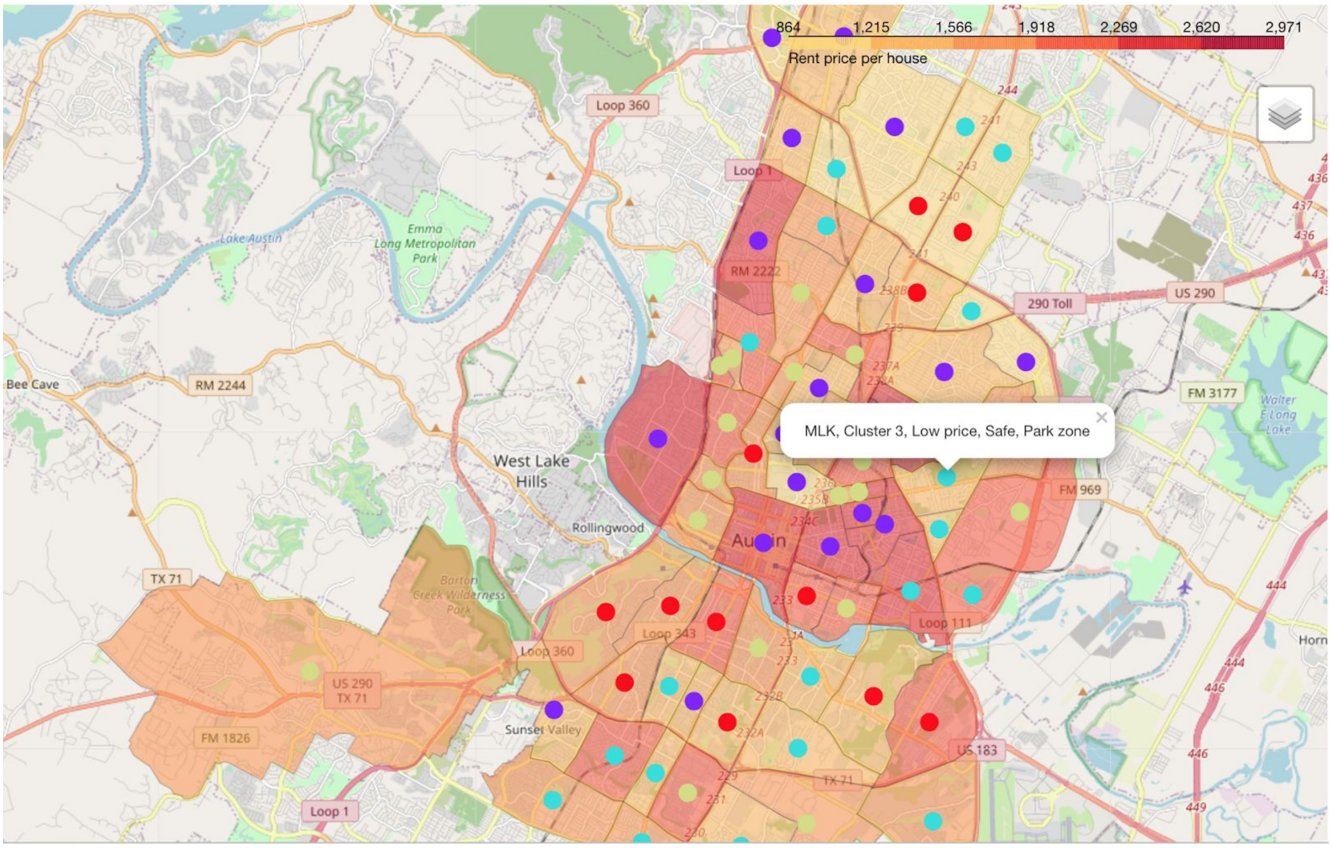


Figure 27. Choropleth rent map with clustered neighborhoods markers

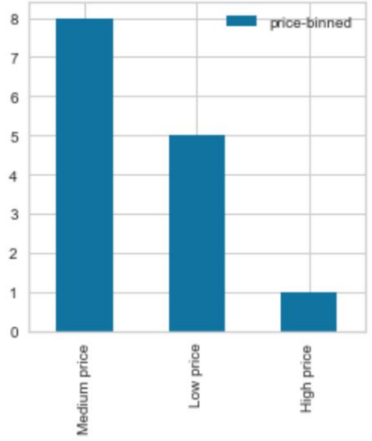


Figure 28. Price categories graph in cluster #1

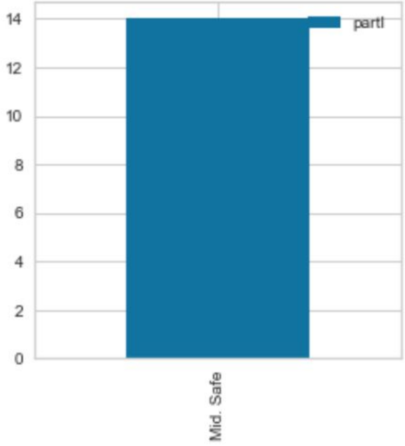


Figure 29. Crime rating in cluster # 1

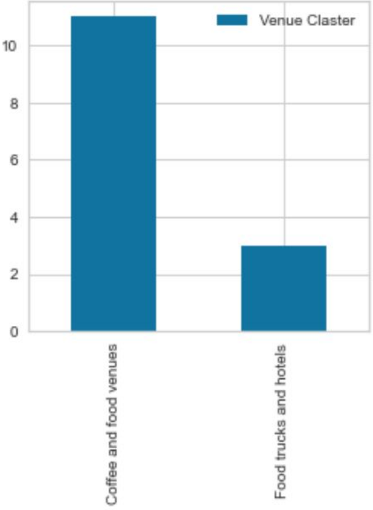


Figure 30. Categories of venues in cluster #1

Cluster #1 has medium safe neighborhoods, a variety of rent prices, coffee, food venues, and food trucks and hotel venues.

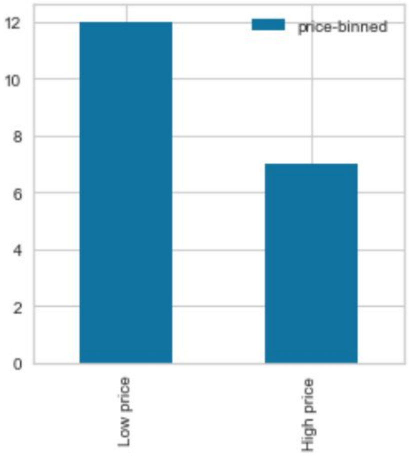


Figure 31. Price categories graph in cluster #2

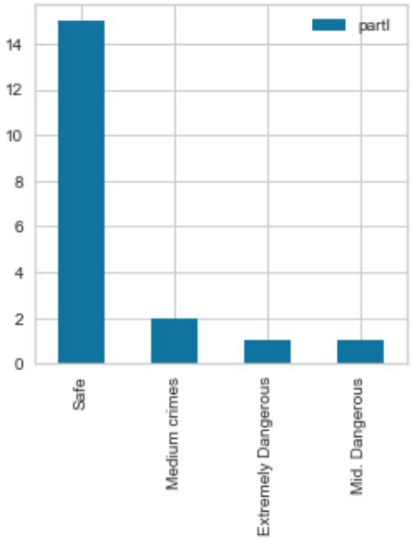


Figure 32. Crime rating in cluster #2

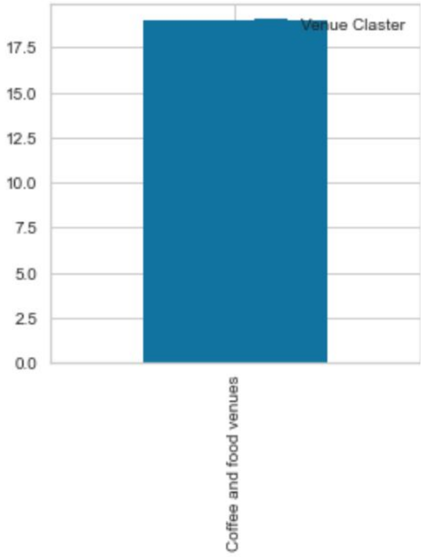


Figure 33. Categories of venues in cluster #2

Cluster #2 generally has safe neighborhoods (79 percent), but you have to be careful choosing this because it consists of extremely dangerous, medium dangerous, and an average rate of crimes. Rent prices are either low or high (proportions are 63:37); you could find coffee and food venues.

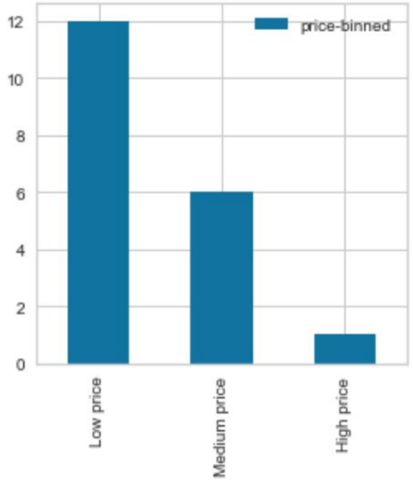


Figure 34. Price categories graph in cluster #3

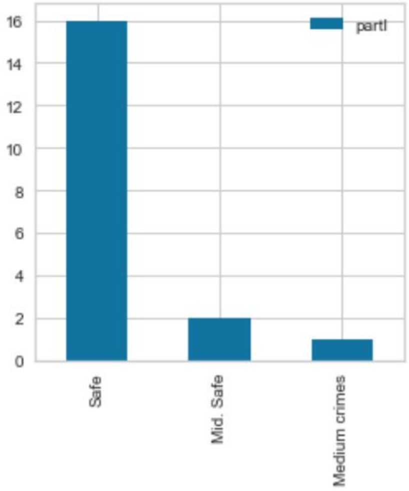


Figure 35. Crime rating in cluster #3

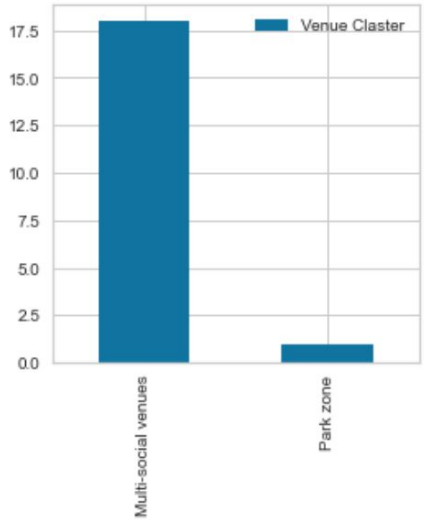


Figure 36. Categories of venues in cluster #3

Cluster #3 has multi-social venues, 85 percent of safe neighborhoods, a variety of rent prices, multi-social venues, and park zone.

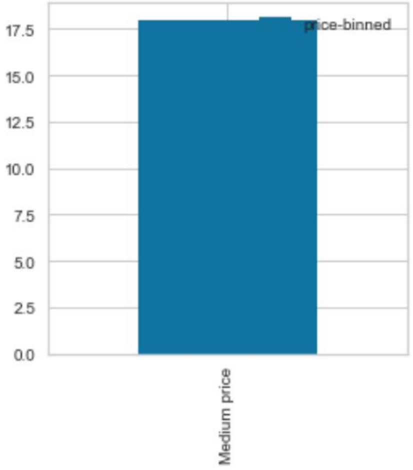


Figure 37. Price categories graph in cluster #4

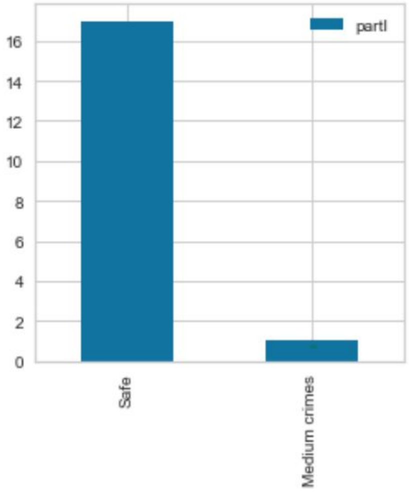


Figure 38. Crime rating in cluster #4

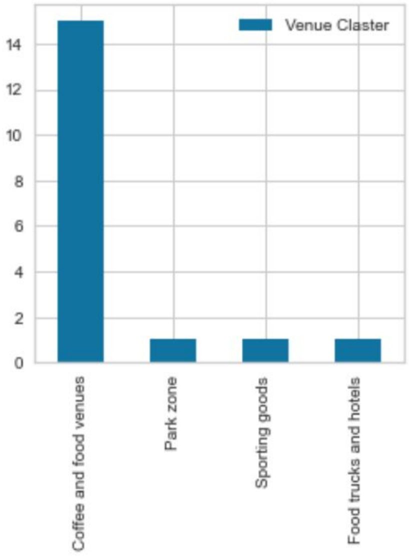


Figure 39. Categories of venues in cluster #4

Cluster #4 has a medium price property, 95 percent safe neighborhoods, 83 percent of coffee and food venues.

If someone decided which cluster they are considering to move in, I created a map with cluster layers to show or hide interesting data.

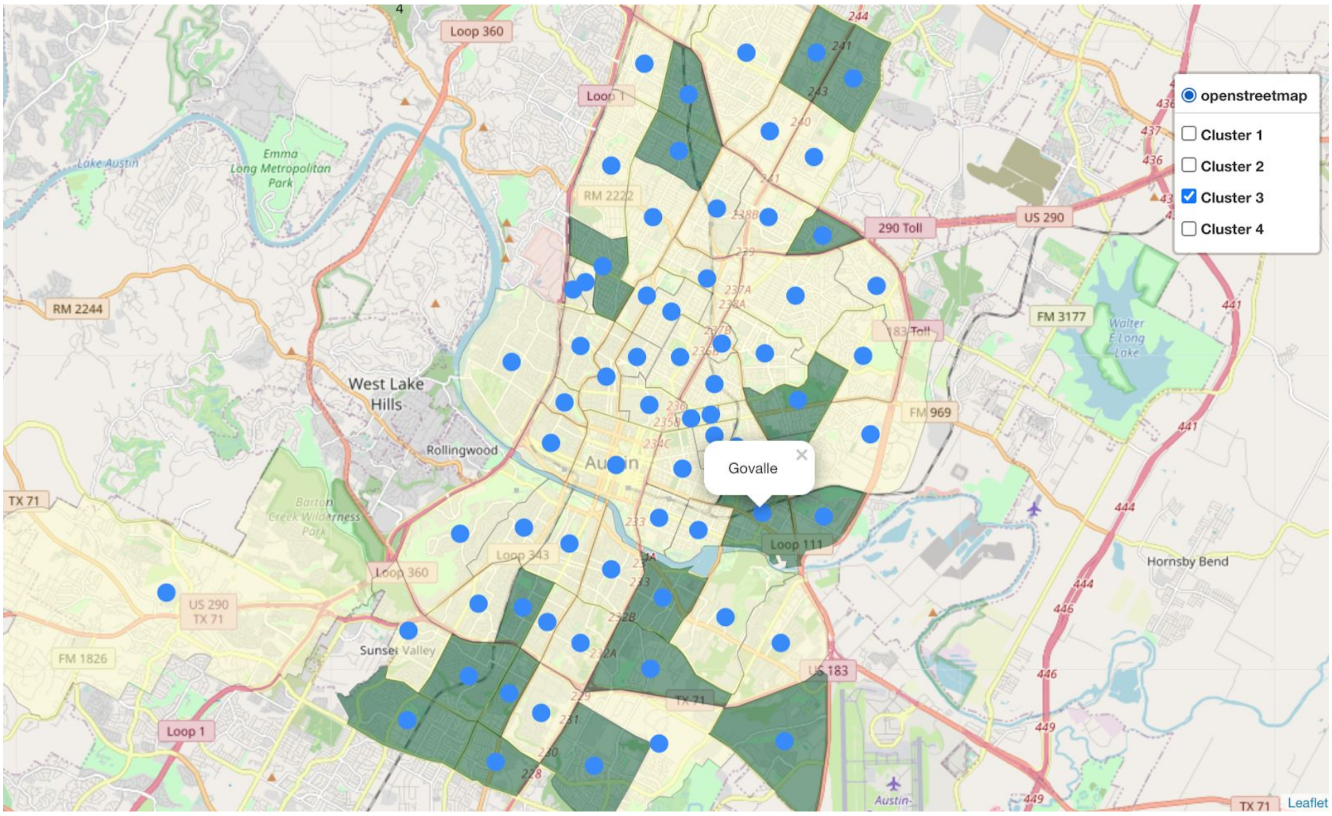


Figure 40. Choropleth map with cluster layers

5. Discussion

Choosing a neighborhood for relocation purposes isn't easy. I hope my research could help people to find the best area for them.

You could start with a topic that you interested in.

For example, your main goal is to find a peaceful region. Neighborhoods in the clusters #2 (15 areas), #3(16 areas), and #4(17 areas) are suitable for that mission. So you will have plenty of options. But always keep in mind that cluster #2 has Downtown and North Austin, which are the most dangerous regions in Austin.

The Downtown neighborhood included in high-priced areas and has a hazardous crime rate. I recommend keeping in mind those two criteria if you are choosing this part of the city.

North Austin has a meddle dangerous crime rate and has the first place at the family violence rate, but unlike Downtown has a low rent price.

If your main goal is to find a low price region, neighborhoods in the clusters #1 (12 areas), #2(12 areas), and #3(5 areas) are suitable for that mission.

Only 9 neighborhoods are high-priced (figure 18). I suggest excluding those if you don't have enough money to afford a property like that.

If your primary concern is to find a place with multi-social venues, with plenty of options, then cluster #3 is for you.

If your primary concern is to be around coffee venues or have a short path to food trucks, I would consider cluster #1(14 areas), #2(17 areas), and #4(15 areas).

In cluster #1, you could find medium safe properties, venues with coffee and food trucks, and rent prices in all budget from low to high.

In cluster #2, you could coffee venues here, safe regions, except Downtown and North Austin, price either low or high.

In cluster #3, only one cluster with multi-social venues, generally safe, and rent prices in all budget from low to high.

In cluster #4, you could find medium price properties, with a safe crime rate, venues mainly coffee.

For better orientation on the map between clusters, you could use my last map (figure 40) by easily show or hide clusters that you are interested in.

6. Conclusion

As mentioned before, Austin is an excellent example of a growing city and a unique model for relocation purposes.

This project aimed to cluster neighborhoods by venues, crime rate, and average rent; It could help those relocating to Austin or someone who needs to know more about the city.

For that purpose I

1. Bined serious crimes into six categories from part I offenses. After I caught sight of hazardous areas.

2. Bined rent prices into three categories low, medium, and high-priced neighborhoods. After binning, I found 9 communities with a high-priced rent category.

3. Clustered venues by their similarity. Foursquare API helped me to create a data frame, and the K-means method was used for clustering.

4. Clustered neighborhoods using all three primary data sources above. Elbow method served to find an optimal number of clusters, which is 4, and K-means was used for clustering.

For better orientation on the map between clusters, you could use my last map.