Capstone Project - Restaurants of Austin

Bargav Cheruku

October 6th, 2019

Table of contents

1. Introduction

1.1 Problem

The business problem is to get the count, pricing and rating of restaurants by zip codes in Austin. Map them based on each variable independently and then group them into clusters based on price and rating collectively to provide these details to residents and newcomers.

1.2 Background and Interest

Austin is one of top growing cities, many families and businesses are relocating to the city. Some of the key information that the potential newcomers need is count, pricing and rating information of venues particularly restaurants in different neighborhoods of the city so that they can choose an area that fits their needs.

2. Data

2.1 Data Collection

The required data that is needed for this problem is to first get all zip codes in Austin. Then get the restaurants in each zip code and finally get pricing and rating for each restaurant.

Opendatasoft will be used to get zip codes of Austin city along with latitude and longitude information and google places api will be used to get the restaurants in each zip code along with their pricing and rating.

2.2 Data Cleaning

It is convenient to download the zip codes data from opendatasoft in csv format. The csv is loaded into a pandas data frame and columns Zip, City, State, Longitude and Latitude are retained by dropping the columns that are not needed.

For each zip code in above data frame, restaurants within 2500 m from the epicenter are queried. Restaurant name, location, price and rating are stored along with zip code information to a new data frame. Finally, the restaurants with no information for pricing and rating, and duplicates are dropped to get the clean data.

3. Methodology

Methodology part is divided into 4 sections. In first section the data frame is analyzed by count, second section is based on pricing, third section uses rating for analysis, and finally pricing and rating are collectively used for clustering.

Initially zip codes centers are mapped using Folium to visualize spreading of zip codes. Zip codes around the center of the city are closer and they are farther while moving away from the center of city.

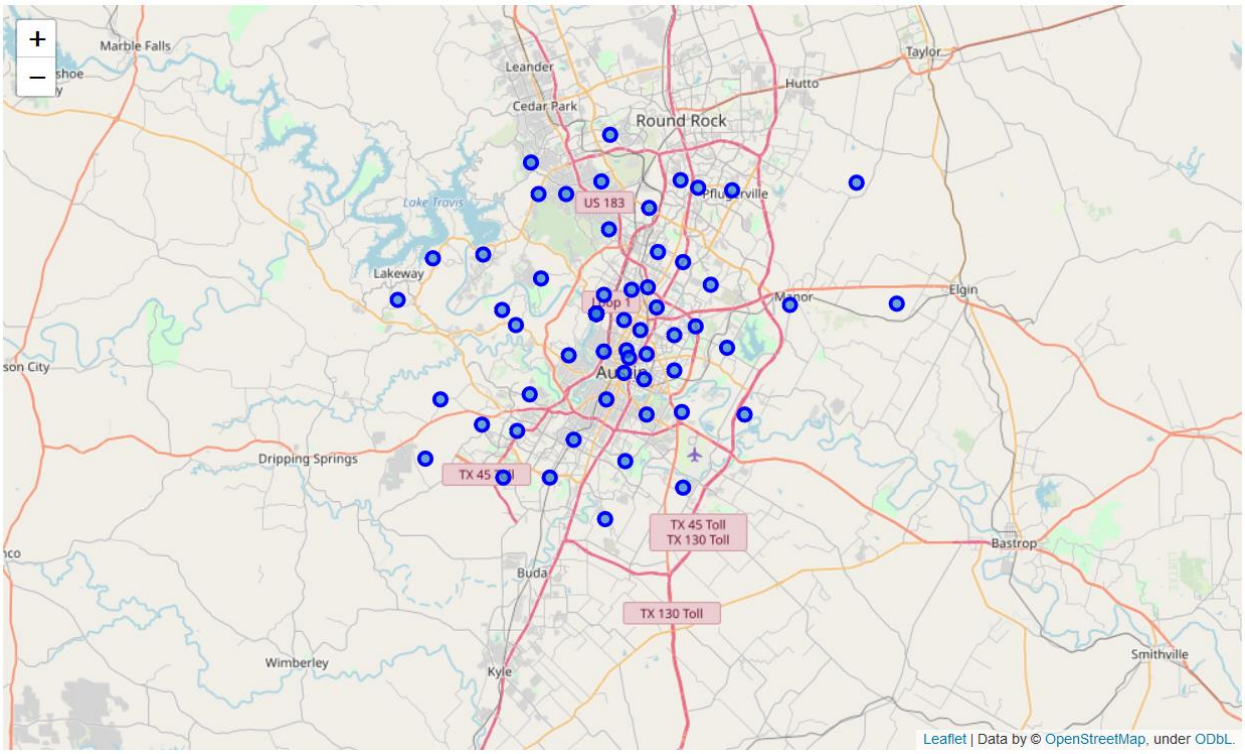

Figure 1. Austin city zip codes

Data queried using places api is stored into a csv file to reduce repetitive querying and processing. Following is head of the data frame that is used for analysis. There are a total of 1130 restaurants within 46 zip codes in the cleaned data set.

	Zip	Zip Latitude	Zip Longitude	Restaurant	Latitude	Longitude	Price	Rating
0	78727	30.425652	-97.71419	Tacodeli Gracy Farms	30.407702	-97.713286	1	4.6
1	78727	30.425652	-97.71419	Tomodachi Sushi	30.425470	-97.716920	2	4.6
2	78727	30.425652	-97.71419	Silver Grill Cafe	30.425090	-97.716169	1	3.8
3	78727	30.425652	-97.71419	P. Terry's Burger Stand	30.414557	-97.705283	1	4.4
4	78727	30.425652	-97.71419	Biryani Pot	30.417239	-97.704026	2	3.9

Table 1. Cleaned data frame

3.1 Count

Data is grouped by zip code, zip code's latitude, zip code's longitude and count of all restaurants in the zip code.

	Zip	Zip Latitude	Zip Longitude	0
0	78701	30.271270	-97.74103	58
1	78702	30.265158	-97.71879	22
2	78703	30.290907	-97.76277	26
3	78704	30.246309	-97.76087	48
4	78705	30.292424	-97.73856	48

Table 2. Data frame grouped by count

3.1.1 Count Map

The range of count is 0 to 60, it is divided into 6 equal parts. Each zip code is mapped with color that matches to count of restaurants in the zip code. Map shows that the zip codes near center of city and zip codes near north and south west with more offices and residences have higher count of restaurants and the count decreases as we move further away from those. This corelates well with the fact that the downtown and places with denser population that have more offices and residences also have higher number of restaurants.

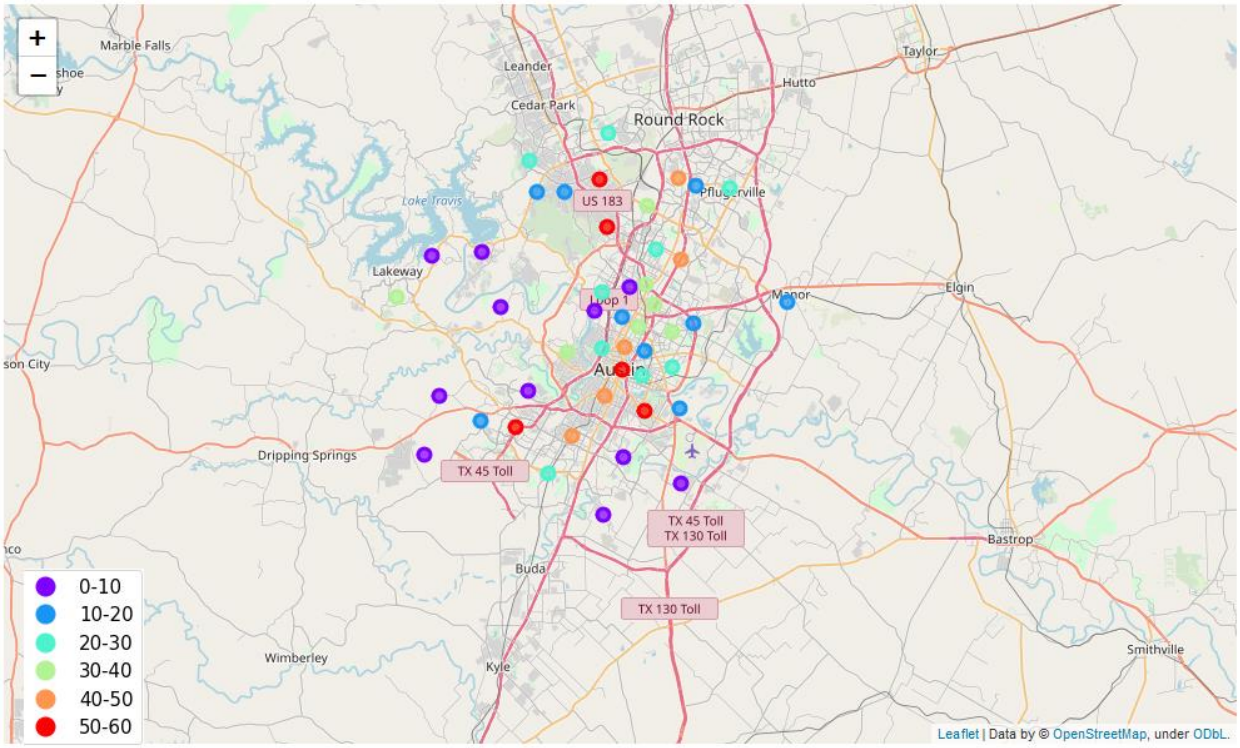

Figure 2. Zip codes colored by restaurant count

3.2 Pricing

Data is grouped by zip code, zip code's latitude, zip code's longitude and mean pricing of all restaurants in the zip code. Pricing for a venue is divided into 4 levels in google places where higher the level, more expensive is the restaurant. Each restaurant has one of the four levels for pricing, we use level itself as pricing for simplifying analysis and take the mean to show overall pricing for the neighborhood.

	Zip	Zip Latitude	Zip Longitude	Price
0	78701	30.271270	-97.74103	1.982759
1	78702	30.265158	-97.71879	1.909091
2	78703	30.290907	-97.76277	1.769231
3	78704	30.246309	-97.76087	1.625000
4	78705	30.292424	-97.73856	1.520833

Table 3. Data frame grouped by mean pricing

3.2.1 Pricing Map

The grouped data is mapped using folium. Each zip code center is colored by mean pricing. Since most restaurants in all zip codes have pricing levels 1 and 2, mean pricing is in that range.

Again, it clearly shows that the restaurants near downtown and within few well-off suburbs have higher pricing. This gives a very clear picture of the distribution of pricing of restaurants.

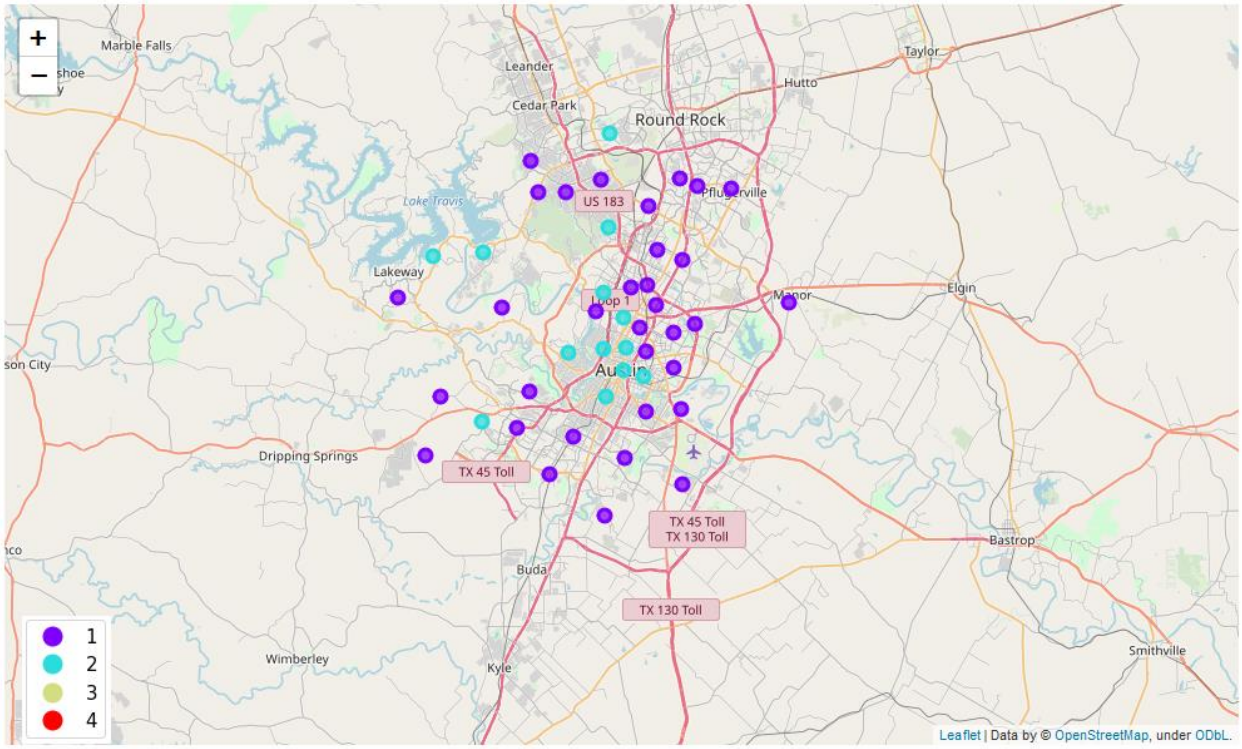

Figure 3. Zip codes colored by mean pricing

3.3 Rating

Data is grouped by zip code, zip code's latitude, zip code's longitude and mean ratings of all restaurants in the zip code. Rating for a restaurant has a range of 0 to 5 but the range of mean ratings in grouped data is 3.3 to 4.5. The difference between zip codes based on rating is less than it was based on pricing or count.

	Zip	Zip Latitude	Zip Longitude	Rating
0	78701	30.271270	-97.74103	4.348276
1	78702	30.265158	-97.71879	4.386364
2	78703	30.290907	-97.76277	4.192308
3	78704	30.246309	-97.76087	4.285417
4	78705	30.292424	-97.73856	4.281250

Table 4. Data frame grouped by mean rating

3.3.1 Rating Map

Small padding is added to the range to set it from 3.2 to 4.6 and it is divided into 7 equal parts. Zip codes are mapped with a color based of mean rating and the range described above. Though the difference based on rating is not as segregating as it was based on pricing it too shows that zip codes near downtown, north and east of the city have more rating than the ones on the west and south.

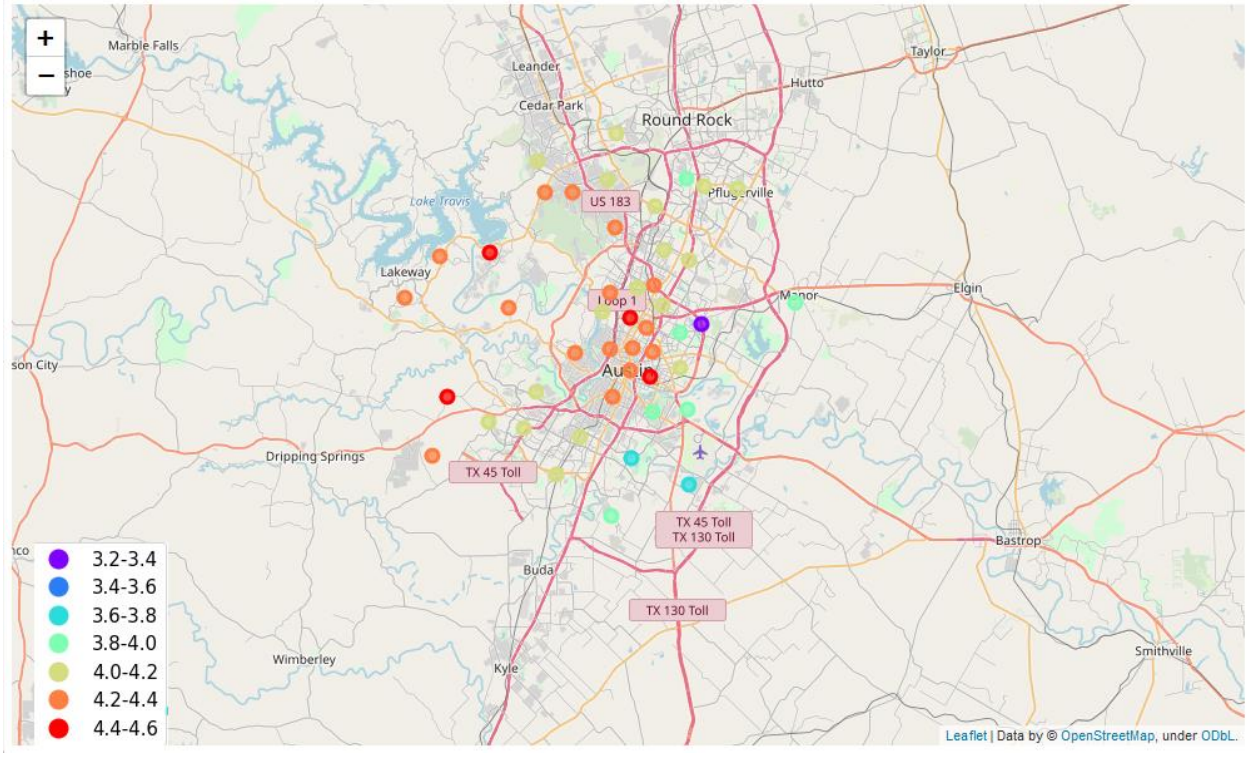

Figure 4. Zip codes colored by mean rating

3.4 Clustering

Data frames created from grouping by pricing and rating are merged to create a single data frame with both mean pricing and rating for each zip code that is used for clustering.

	Zip	Zip Latitude	Zip Longitude	Price	Rating
0	78701	30.271270	-97.74103	1.982759	4.348276
1	78702	30.265158	-97.71879	1.909091	4.386364
2	78703	30.290907	-97.76277	1.769231	4.192308
3	78704	30.246309	-97.76087	1.625000	4.285417
4	78705	30.292424	-97.73856	1.520833	4.281250

Table 5. Data frame merged mean pricing and rating

3.4.1 K-means Clustering

Zip code, zip code latitude and zip code longitude columns are dropped from the data frame to create a new data frame with only pricing and rating columns that are passed to the k-means fit. Number of clusters for k-means is set to 5 which is close to number of pricing levels 4 and rating range 0 to 5.

```
array([1, 1, 1, 1, 4, 4, 4, 4, 2, 0, 3, 0, 3, 4, 0, 0, 1, 1, 3, 1, 0, 3,
       3, 4, 0, 0, 2, 0, 4, 2, 0, 4, 4, 4, 4, 0, 1, 0, 0, 1, 4, 0, 0, 0,
       4, 2])
```

Table 6. k-means clusters labels

4. Results

Zip codes are mapped with cluster labels as color. Clustering results are comparable to results from grouping by count, pricing and rating independently. Clearly the zip codes closer to downtown, north and east side of the city are clustered in similar groups and those farther away from downtown, east and south side of city are grouped into alike clusters. Clusters show that well-off neighborhoods are similarly grouped which consist of restaurants that have good ratings and higher price levels.

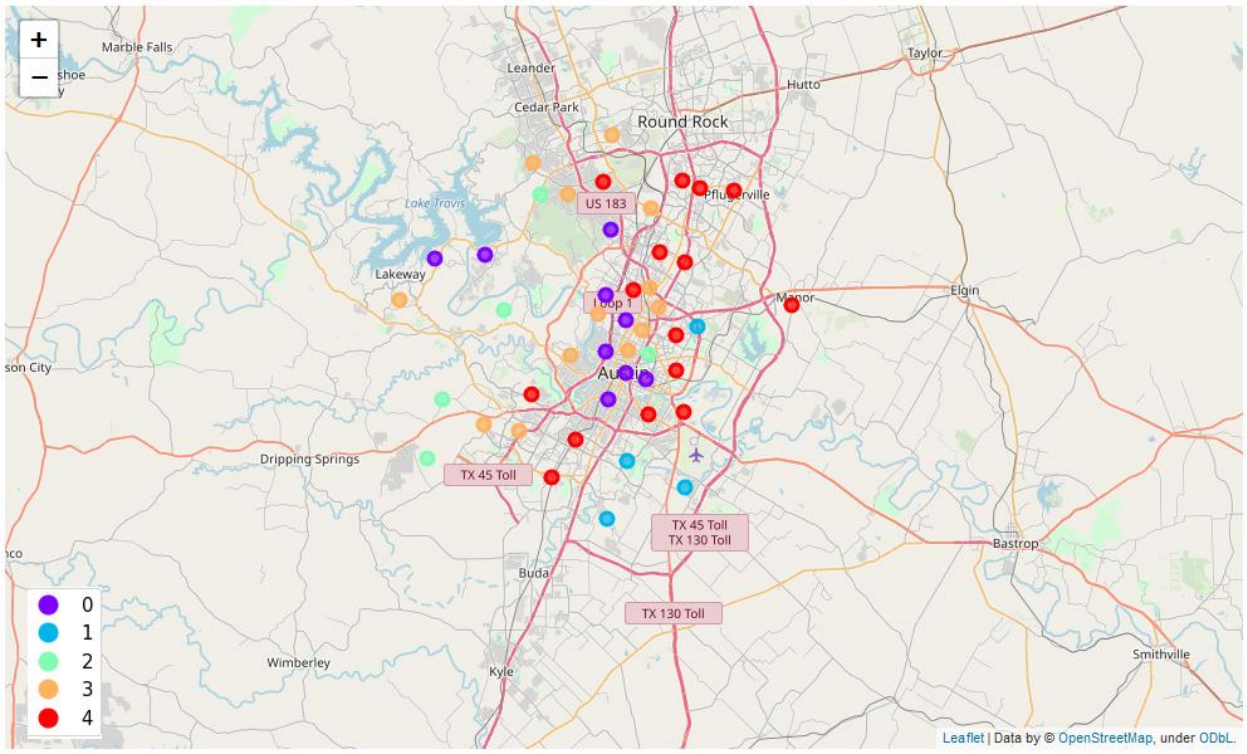

Figure 5. Zip codes colored by k-means clusters

5. Discussion

Quality and quantity of data collected was enough to do the analysis and get the results needed to answer the problem. There was good number of restaurants within each zip code to provide effective average for price and rating. This allowed to generate good maps that clearly grouped the zip codes based off count, pricing and rating.

Division among zip codes based on count gives residents who dine out frequently to choose a neighborhood with higher number of restaurants. Grouping by pricing provides residents option to select a neighborhood that fits their budget and grouping by rating provides all with data needed to filter out ones that aren't good. Finally clustering based on both pricing and rating provides users exactly neighborhoods that fits their budget and has good ratings.

All groupings and clustering show consistently that the downtown area which is quite popular for several things is also the best option for restaurant goers. Second option are the neighborhoods that are towards north and south east which are newly developing neighborhoods. And the ones on east side farther away from city are to be avoided as they have consistently ranked lower on every grouping and clustering.

6. Conclusion

New relocators and current residents looking for several good restaurants, downtown, couple of neighborhoods in north and south east are the best neighborhoods, which are also good neighborhoods based on several other factors.

7. References

- [1] OpenDataSoft- https://public.opendatasoft.com
- [2] Google Places API - https://maps.googleapis.com/maps/api/place