Applied Data Science Capstone Project

Restaurant Analysis
for the City of Austin

Author: Zack O'Keefe

Date: August 2020

Table of Contents

1. Introduction

Austin is the state capital of Texas and one of the fastest growing cities in the United States. It is a hub for technology, music, film, and an increasingly emerging food scene. The city features several diverse neighborhoods, each offering differing cultural attractions and demographics. New restaurants are opening constantly in Austin partly due to the reasonable cost of doing business compared to other large cities in the United States. With such a high volume of restaurants of differing cuisines, it would be difficult for an aspiring owner or investor to predict the success of a new venture.

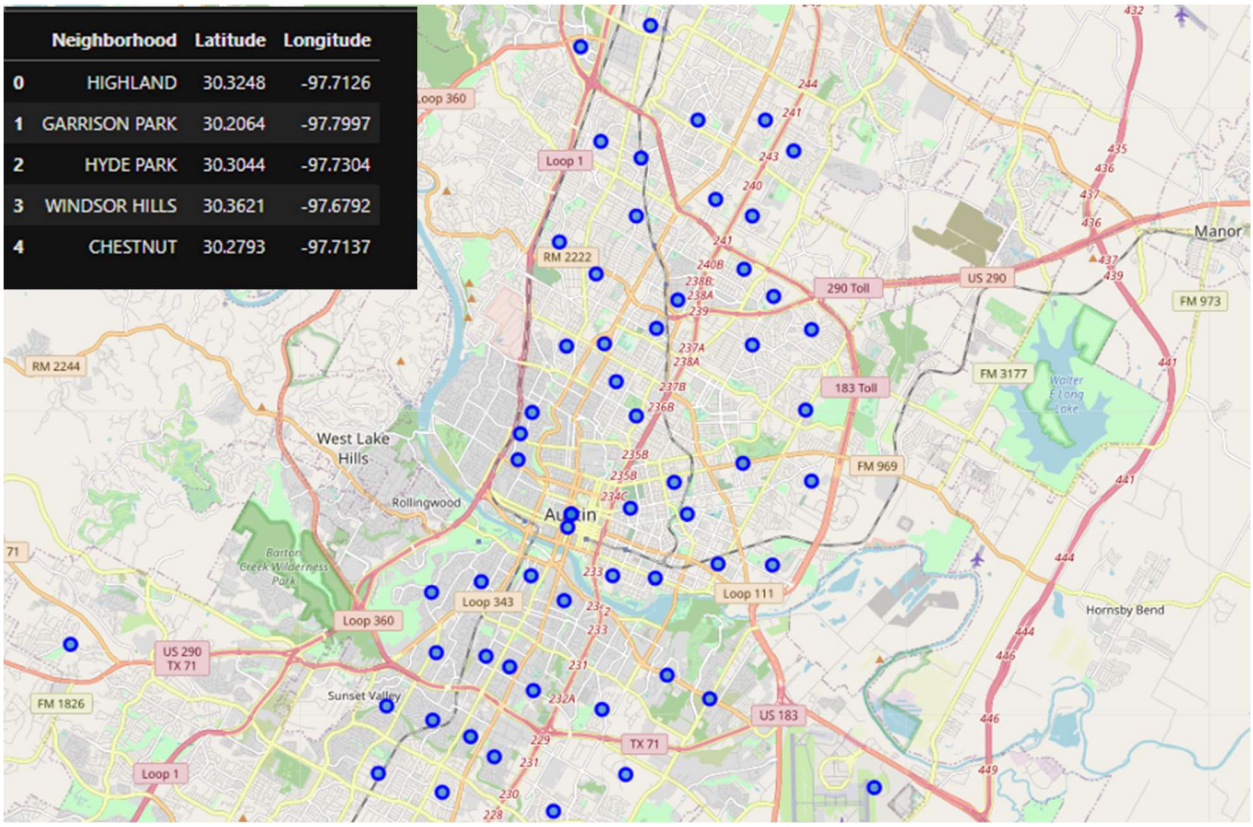
Potential restaurant investors would benefit from knowing which type of restaurants have higher ratings based on the food category and price in each neighborhood to predict the likelihood of it being successful and making it worth their investment. This analysis will explore the top venues in each neighborhood and attempt to find a relationship between restaurant success, measured by the rating, and the type of food it serves and price points in order to steer an investors decision on the best restaurant to fund for a certain area of the city.

2. Data

The City of Austin neighborhood data was exported from data.austintexas.gov as a csv file and represents the boundaries of the City of Austin Neighborhood Planning Areas (NPA) and is public domain. The neighborhood data returned 95 samples and was later trimmed down to 60 samples after removing duplicates and names that were not recognized by the geolocator. Venue data was gathered from the Foursquare location data by making several API calls.

A geolocator was used to find single point coordinates for each neighborhood instead of having a multipolygon feature class. One neighborhood's coordinates did not return in the correct location, so I had to override the latitude and longitude for it using Google Maps. Once this was completed, I used the coordinates to make the Foursquare API calls for venue information including the name, type of venue, rating, price, etc. After merging this data into a single data frame, I began my analysis looking for patterns that lead to high rated restaurants in each neighborhood to pass on to interested investors.

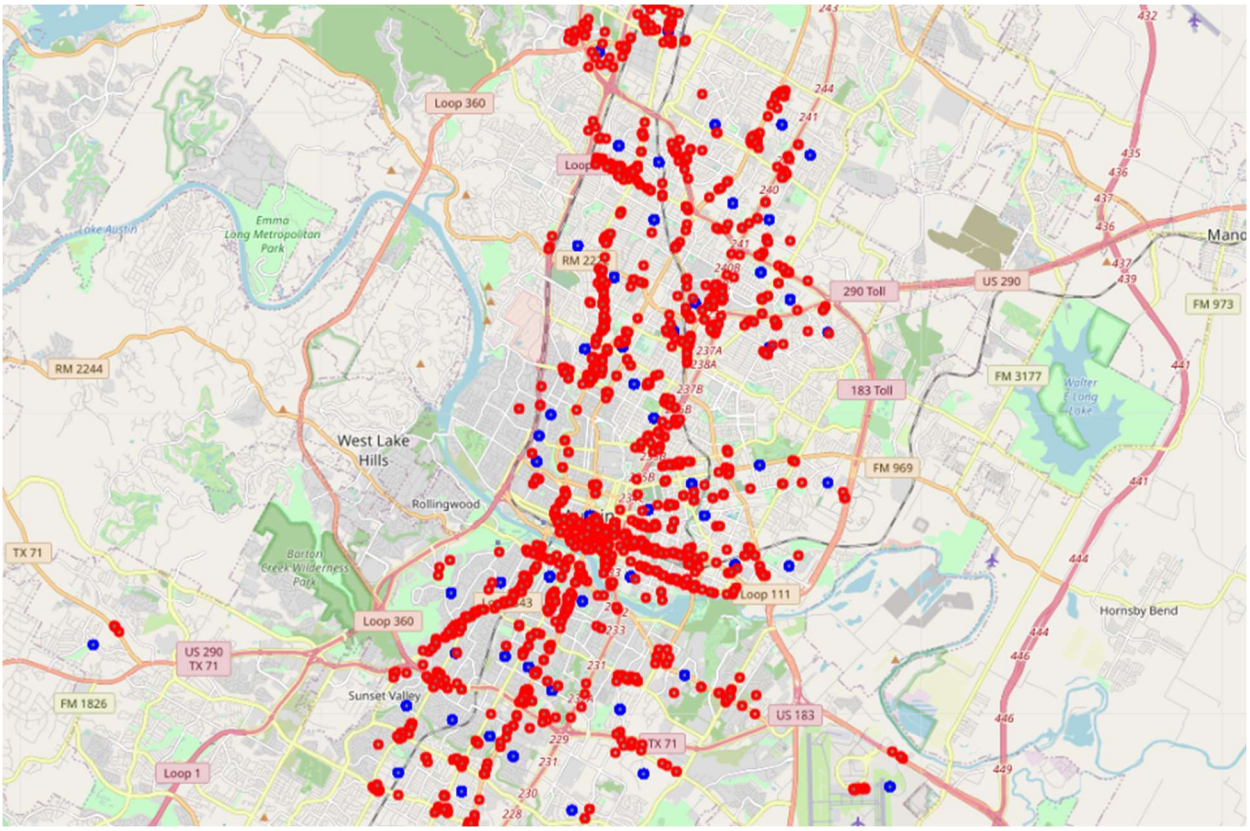
The neighborhood data frame, named df_austin, had 95 features after importing the data. The neighborhoods were looped through a geolocator to get the initial coordinates and then the data frame was cleaned to remove duplicates and modify some of the coordinates. The new version of the data frame consisted of 60 features, each with a unique latitude and longitude. The figure below shows the resulting data frame with the first five neighborhoods and a map of all 60 neighborhoods in Austin that were used for analysis in this project.



The venue data frame, named austin_venues, was created by running each neighborhoods' coordinates through a loop that made a call to the Foursquare API and returned restaurants within a certain radius of the coordinates. The data frame consists of 1836 features and includes the neighborhood name, coordinates, venue/restaurant name and coordinates, the venue category, and a venue ID.

Next, I needed to retrieve the rating and price data for each restaurant which required premium calls. The free account only allowed 500 premium calls a day, so I had to run a loop 500 calls at a time for multiple days until I had the data appended to the data frame for all 1836 rows. I then created a new row called Category that combined the Venue Category column with the Price column, which would be the eventual criteria for clustering. At first there were around 195 different combinations of venue categories and prices, so I tweaked the price column to only have two different price levels instead of the original four, bringing the combinations down by almost half (some of the combinations did not have data tied to it so it was not exactly half). The first five rows of the data frame can be seen below, along with a map of all 60 neighborhoods in blue with the 1836 venues in red.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue ID	Rating	Price	Category
HIGHLAND	30.324836	-97.712608	Vivo Restaurant	30.325447	-97.708553	Mexican Restaurant	54483f08498ea61e416adbd9	8.3	1.0	Mexican Restaurant : 1.0
HIGHLAND	30.324836	-97.712608	Lucky's Puccias	30.329249	-97.709804	Food Truck	4c6f1cb7d274b60c11b8d50d	8.4	1.0	Food Truck : 1.0
HIGHLAND	30.324836	-97.712608	Pappadeaux Seafood Kitchen	30.325366	-97.704845	Seafood Restaurant	4b1fe27cf964a5203f2a24e3	8.5	1.0	Seafood Restaurant : 1.0
HIGHLAND	30.324836	-97.712608	Arpeggio Grill	30.331051	-97.715317	Middle Eastern Restaurant	4a090254f964a5202c741fe3	8.4	1.0	Middle Eastern Restaurant : 1.0
HIGHLAND	30.324836	-97.712608	Cho Sun Gal Bi Korean BBQ & Sushi Bar	30.329763	-97.706639	Korean Restaurant	49bfa36df964a52017551fe3	8.6	1.0	Korean Restaurant : 1.0



3. Methodology

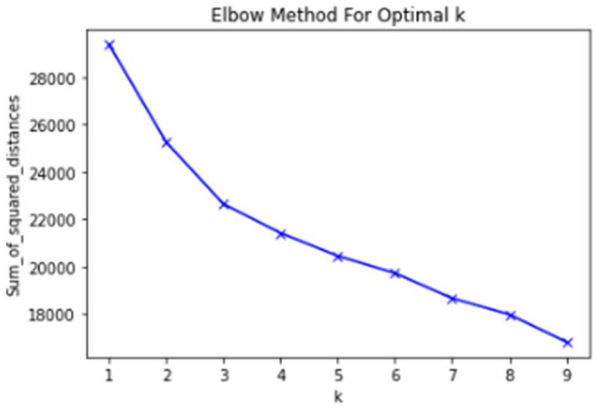
The Foursquare API calls for the venues in each neighborhood were limited to 100 results for this study and the radius was set to a 1,000 meter radius around each neighborhood coordinate. The 100 result limitation was done in order to keep the total results to a manageable number, since premium calls were required to retrieve the rating and price information and could only do 500 per day.

Now that my main data frame has been populated, I began to transform it into a format where I could perform clustering. I created a venues data frame version that contained only the neighborhood names, category, and rating. I grouped this data by neighborhood and then category, finding the mean rating for each of these new rows. From there I pivoted the data frame, leaving unique neighborhood names as the index, all of the category combinations as the new columns, and populated the average rating for each row under the corresponding category and the rest filled in with zeros. This new data frame is shown below.

The rating data from Foursquare is on a range 0.0 to 10.0 and the price tiers are from 1.0 (least pricey) to 4.0 (most pricey). As discussed in the data section above, each of the four price tiers were originally appended to each of the cuisine types to build the category column and resulted in too many different combinations. To get better results, I edited the price data to only show two tiers by changing the 2.0 tier to 1.0 and the 3.0 tier to 4.0. Now we ended up with only a 1.0 for low priced restaurants and 4.0 for high priced restaurants.

Neighborhood	African Restaurant : 1.0	American Restaurant : 1.0	American Restaurant : 4.0	Argentinian Restaurant : 1.0	Asian Restaurant : 1.0	Asian Restaurant : 4.0	BBQ Joint : 1.0	BBQ Joint : 4.0	Bagel Shop : 1.0	...	Tapas Restaurant : 1.0	Tapas Restaurant : 4.0	Tex-Mex Restaurant : 1.0	Thai Restaurant : 1.0	Thai Restaurant : 4.0	Turkish Restaurant : 1.0
ALLANDALE	0.0	0.00	0.0	0.0	0.0	0.0	0.00	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
BARTON HILLS	0.0	0.00	0.0	0.0	0.0	0.0	0.00	0.0	0.0	...	0.0	7.3	8.4	0.0	0.0	0.0
BOULDIN CREEK	6.9	0.00	0.0	0.0	6.8	0.0	7.60	6.8	0.0	...	0.0	0.0	0.0	0.0	9.1	0.0
BRENTWOOD	0.0	8.30	0.0	0.0	5.8	0.0	0.00	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
CENTRAL EAST AUSTIN	0.0	6.65	0.0	0.0	0.0	0.0	8.25	0.0	8.8	...	0.0	0.0	0.0	0.0	0.0	8.2

Next, I began the KMeans clustering analysis. I used the elbow method to find the optimal k. For this analysis I used four clusters after developing the plot below that compares the number of k's to the sum of squared distances. There is a slight elbow around a k value of three and four, so I ran the analysis with both and settled on four after reviewing both of the results.



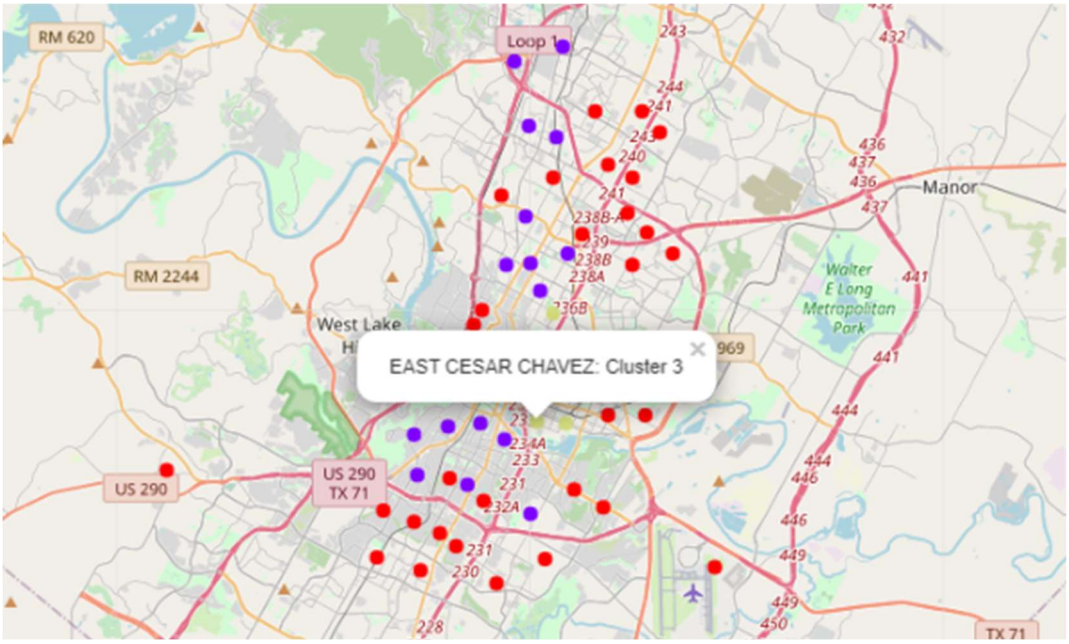
After running the clustering, I ended up with a resulting array with cluster labels. I appended the cluster labels to the latest data frame and removed the neighborhood column. I then grouped the rows by the cluster labels and found the average rating for each category of cuisine and price range.

Cluster Labels	African Restaurant : 1.0	American Restaurant : 1.0	American Restaurant : 4.0	Argentinian Restaurant : 1.0	Asian Restaurant : 1.0	Asian Restaurant : 4.0	BBQ Joint : 1.0	BBQ Joint : 4.0	Bagel Shop : 1.0	...	Tapas Restaurant : 1.0	Tapas Restaurant : 4.0	Tex-Mex Restaurant : 1.0	Thai Restaurant : 1.0	Thai Restaurant : 4.0	Turkish Restaurant : 1.0
0	NaN	7.500000	9.10	NaN	6.400000	NaN	7.5625	NaN	NaN	...	NaN	NaN	6.5	6.000000	NaN	NaN
1	7.0	7.026154	8.15	NaN	7.306667	8.4	6.7250	6.8	6.80	...	8.4	7.9	8.4	7.112500	8.4	NaN
2	NaN	7.970833	7.80	NaN	NaN	NaN	8.2500	8.3	8.20	...	NaN	NaN	NaN	7.550000	NaN	7.7
3	NaN	5.916667	8.00	8.0	7.500000	NaN	7.5950	NaN	7.55	...	NaN	NaN	NaN	7.666667	NaN	8.2

4. Results

Using the data frame with the average rating for each cuisine and price grouped by clusters, I sorted the ratings to show the ten highest rated categories for each cluster. This was used to display the final results and serve as an important data deliverable to pass on to clients. I also generated a list of all the neighborhoods in each cluster as well as a map, which can be used to figure out what cluster a specific neighborhood is in and then view the list to get insight on the highest rated restaurants by cuisines and price points in the area of interest.

Cluster Labels	1st Ranked Category	2nd Ranked Category	3rd Ranked Category	4th Ranked Category	5th Ranked Category	6th Ranked Category	7th Ranked Category	8th Ranked Category	9th Ranked Category	10th Ranked Category
0	Café : 4.0	Donut Shop : 4.0	American Restaurant : 4.0	Middle Eastern Restaurant : 1.0	Taco Place : 4.0	Food Stand : 1.0	Restaurant : 4.0	Mexican Restaurant : 4.0	Restaurant : 1.0	Fast Food Restaurant : 4.0
1	Ramen Restaurant : 1.0	Burger Joint : 4.0	Cuban Restaurant : 4.0	Café : 4.0	Bakery : 4.0	Steakhouse : 4.0	Hawaiian Restaurant : 1.0	Italian Restaurant : 4.0	Gluten-free Restaurant : 1.0	Greek Restaurant : 1.0
2	Café : 4.0	Restaurant : 4.0	Seafood Restaurant : 4.0	Chinese Restaurant : 1.0	Israeli Restaurant : 1.0	Cajun / Creole Restaurant : 1.0	Churrascaria : 1.0	Donut Shop : 1.0	Mexican Restaurant : 4.0	Deli / Bodega : 1.0
3	Southern / Soul Food Restaurant : 1.0	Irish Pub : 1.0	Seafood Restaurant : 1.0	Latin American Restaurant : 4.0	Vegetarian / Vegan Restaurant : 1.0	Japanese Restaurant : 1.0	Turkish Restaurant : 1.0	Ramen Restaurant : 1.0	Café : 1.0	Breakfast Spot : 1.0



5. Discussion

Based off the results data frame, investors can get a glimpse of the most successful restaurants as determined from their cuisine and price point. For example, if a potential investor is looking at supporting a restaurant in the East Cesar Chaves neighborhood of Austin, they can find that it is in cluster 3 and the highest scoring restaurant type is low cost southern/soul food. If you are looking for an opportunity in downtown Austin, which is a part of cluster 1, a low cost ramen restaurant would be a good option.

There are some basic observations found by looking at the 10 highest ranked categories for each of the four clusters. For cluster 0, four out of the first five highest ranked categories are high priced restaurants, making it appear that the neighborhoods in this cluster prefer higher end restaurants. Cluster 3 could draw the opposite assumption. Four out of the first five categories are low priced restaurants.

6. Conclusion

Austin is a diverse city with a lot of culture and continues to grow at a high rate. There are many restaurants in each neighborhood differing in cuisine, type, and price. This analysis focused on grouping the neighborhoods into clusters, based on the type of restaurant and whether it is on the cheap or expensive end of the price spectrum. The results of this analysis can serve as a guide to a potential investor or owner of a new restaurant either to choose what type of restaurant to open and at what price

point in a certain neighborhood or based on a certain restaurant concept, choose an appropriate location that will increase the chance of success.

The study could be expanded in the future by increasing the Foursquare API call for venues from 100 to a larger number. Another area to consider to potential improve on the results would be to remove the neighborhoods that have too small of a sample size of restaurants, since this could skew the average rating values in these areas.

This insight can help investors make smart decisions and hopefully lead to restaurants with high ratings that customers can enjoy all around Austin. This same analysis could easily be mimicked in our cities around the world.