Clustering in North San Antonio and Surrounding Area

Jessica Beamer

1. Introduction

1.1 Background

San Antonio, TX is a diverse city that is seeing continued growth, and the surrounding areas are also experiencing an increase in development. It was recently ranked as one of the fastest growing cities in the United States. The intent of this project is to identify areas in north San Antonio, TX and the surrounding area that would be ideal for opening a restaurant. We will explore the current offerings and review the potential for future business endeavors. This analysis would be particularly helpful for a restaurateur who is interested in starting or expanding their business.

1.2 Business Problem

The objective is to find the ideal location for a restaurateur to open a new restaurant in north San Antonio or the surrounding area.

1.2 Interest

Developers, Franchise owners, entrepreneurs, and other restaurateurs would be interested in understanding the make-up of San Antonio and the surrounding areas.

2. Data Acquisition and Cleaning

2.1 Data Used and the Sources

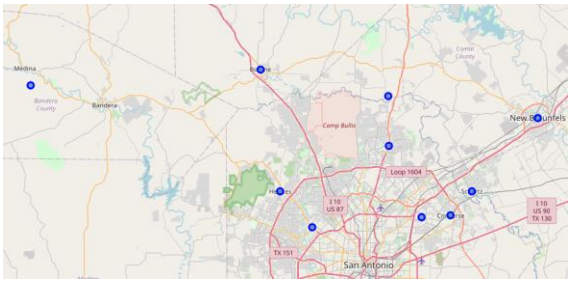
We will identify the major neighborhoods/ sub areas to explore after reviewing the San Antonio map (produced with Folium) and then create a dataframe that includes the neighborhood name, along with the latitude and longitude that are obtained with the Geocoder package.

We will be exploring the venue category by neighborhood. We will use the Foursquare API to obtain venue data related to these neighborhoods/ sub-areas.

3. Methodology

3.1 Visualizing the neighborhoods to explore

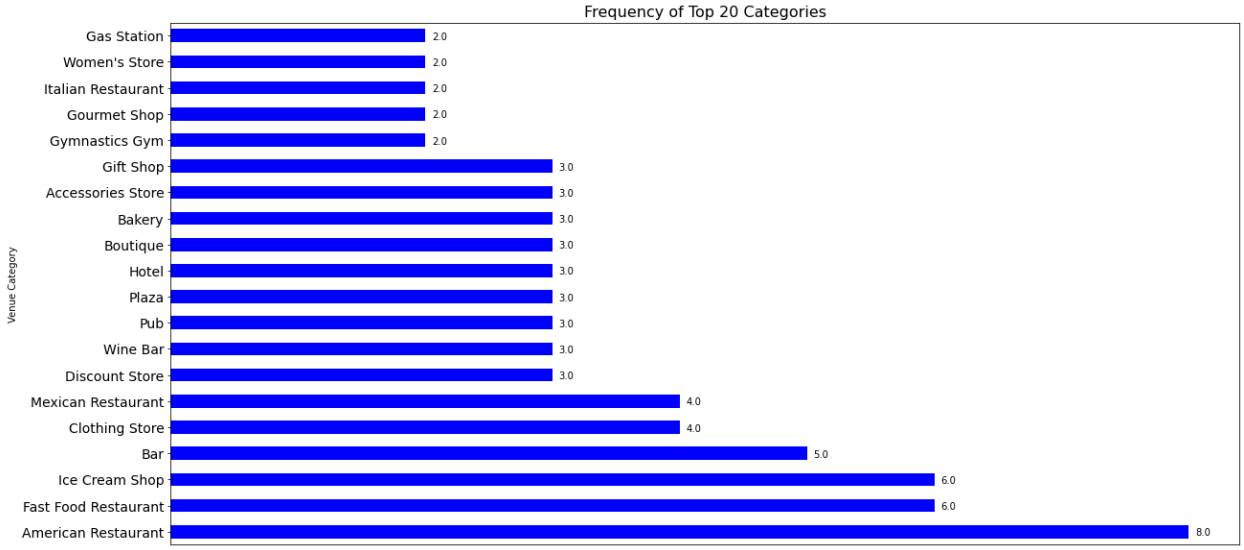
The Folium package was first used to review the smaller cities within the north San Antonio area, and it was used again after selecting the top cities to review to visualize where they are located.



3.2 Reviewing nearby venues

Foursquare API data was used to access the data of venues within a 500m radius. The below graph is a subset of the data that shows the frequency of the top 20 venue categories in the area.

This is helpful for understanding what kind of areas we're looking at and what restaurants might be most successful (or could already be saturated).



The top 10 most common venues by neighborhood were also reviewed.

	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	Boerne, TX	Boutique	American Restaurant	Café	Burger Joint	Restaurant	Bistro	Hotel	Mexican Restaurant	Snack Place	Gourmet Shop
1	Bulverde, TX	Wine Bar	Gift Shop	Airport	American Restaurant	Discount Store	Cocktail Bar	Coffee Shop	Construction & Landscaping	Convenience Store	Cosmetics Shop
2	Converse, TX	Gas Station	BBQ Joint	Business Service	Construction & Landscaping	Convenience Store	Locksmith	Italian Restaurant	Dive Bar	Ice Cream Shop	Gymnastics Gym
3	Helotes, TX	Discount Store	Pharmacy	Smoke Shop	Fast Food Restaurant	Ice Cream Shop	Burger Joint	Bar	Gas Station	Video Store	American Restaurant
4	Leon Valley, TX	Fast Food Restaurant	Ice Cream Shop	Pawn Shop	Automotive Shop	Other Repair Shop	Pharmacy	Chinese Restaurant	Shoe Store	Mexican Restaurant	Mobile Phone Shop
5	New Braunfels, TX	Bar	Pub	German Restaurant	American Restaurant	Market	Hotel	Plaza	Cocktail Bar	Gourmet Shop	Clothing Store
6	Schertz, TX	Pizza Place	Baseball Field	Weight Loss Center	Gymnastics Gym	Bakery	Clothing Store	Cocktail Bar	Coffee Shop	Construction & Landscaping	Convenience Store
7	Stone Oak, TX	Accessories Store	Clothing Store	Women's Store	American Restaurant	Sandwich Place	Cosmetics Shop	Fast Food Restaurant	Gastropub	Massage Studio	Juice Bar
8	Windcrest, TX	Discount Store	Golf Course	Plaza	Frozen Yogurt Shop	Food Truck	Chinese Restaurant	Clothing Store	Cocktail Bar	Coffee Shop	Construction & Landscaping

3.3 Clustering

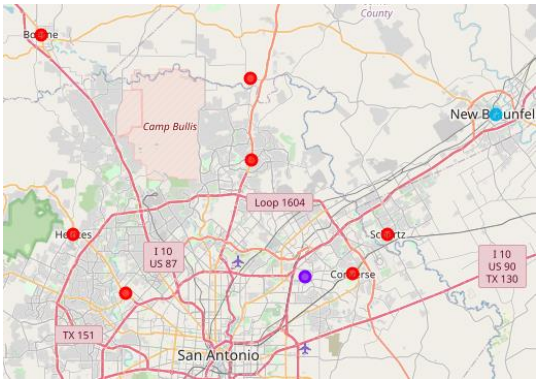
K-means clustering, which is an unsupervised machine learning technique, was used to divide the areas into three clusters. The dataframe with the restaurant venues was transformed using one-hot encoding. The process of one-hot is used for converting categorical variables into a form that can be used in a machine learning algorithm. Including a sample of the resulting df and supporting code.

```
# one hot encoding
sanantonio_onehot = pd.get_dummies(sanantonio_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
sanantonio_onehot['Neighborhood'] = sanantonio_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [sanantonio_onehot.columns[-1]] + list(sanantonio_onehot.columns[:-1])
sanantonio_onehot = sanantonio_onehot[fixed_columns]

sanantonio_onehot.head()
```

	Neighborhood	ATM	Accessories Store	Airport	American Restaurant	Antique Shop	Asian Restaurant	Automotive Shop	BBQ Joint	Bakery	Bank	Bar	Baseball Field	Beer Garden	Bistro	Boutique	Breakfast Spot
0	Stone Oak, TX	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	Stone Oak, TX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Stone Oak, TX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Stone Oak, TX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Stone Oak, TX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

4. Results

Three clusters were created. As shown in the map below, one cluster is in red, one is in blue, and one is in purple.



Cluster 1 contains the most areas, and the neighborhoods in this group appear to have more restaurants in their top 10 common venues.

Cluster 2 contains fewer restaurants, but it contains dessert and drink locations too.

Cluster 3 appears to be somewhat similar to Cluster 3, in that it does not have many restaurants. However, it contains an American restaurant.

5. Discussion

Based on the results, entrepreneurs and restaurateurs have several options, depending on their area of expertise. If they are not set on one particular cuisine though, my recommendation would be to open an American food restaurant in cluster 2 (New Braunfels area).

We saw from the initial exploratory analysis of all of the San Antonio and surrounding area results that American food restaurants are the most popular and fall into the top-rated categories. Cluster 1 has many restaurants already, and this might indicate that it could already be saturated with restaurants. Cluster 3 has fewer restaurants, but it does contain an American restaurant in its top venues. This leaves Cluster 2 (New Braunfels area) as the most desirable place to open an American food restaurant.

6. Conclusion

The San Antonio area continues to grow and develop. While outside the scope of this project, further analysis could include expanding the geographic constraints to include the entirety of San Antonio and surrounding areas, as opposed to just the north side. This could potentially help with increasing the size of the clusters. Using the tools learned in this course, we were able to create clusters and provide the recommendation for a restaurateur to open an American cuisine restaurant in the New Braunfels area.