Finding the Perfect Neighborhood in San Antonio, TX

David Risius

29 January 2020

1. Introduction

1.1 Background

San Antonio, Texas is one of the fastest growing cities in the United States. According to the United States Census Bureau, San Antonio topped the list of the fastest growing metro areas for 2017. In previous analysis, we clustered and segmented neighborhoods in Toronto and New York city based on FourSquare venue data. San Antonio is a very different city than either New York or Toronto. For one, it is a very large city with relatively sparse population compared to the other cities. According to Wikipedia, San Antonio city consists of around 1.5 million people within a land area of 461 square miles compared to 8.5 million for 303 square miles in New York City and 2.7 million for 243 square miles in Toronto. The ethnicity of the three cities is also different. San Antonio has a large Hispanic influence with around 63% of residents of Hispanic or Latino origin. New York is around 28% Hispanic while Toronto is around 4% Hispanic with a much larger proportion of Asian (40%) and European (48%) than San Antonio or New York. Median housing prices between New York City and San Antonio are also very different. According to Zillow, the median single-family home in December, 2019 was $477K in New York compared to around $204K in San Antonio. If one can find the right neighborhood to live, San Antonio could provide a lot of value for the cost of living.

1.2 Problem Statement

Given a list of preferred criteria about a neighborhood, we would like to find an initial set of neighborhoods to begin searching for a new home in the San Antonio area. Our initial set of criteria is as follows:

1. Median Home Price: I am looking for a single-family house within the $200K−$350K range. There are multiple neighborhoods both above and below this range so these will be eventually filtered out. We would also like to find those neighborhoods where the median home price is increasing over time in case we would like to sell the home in the future
2. Good Schools: Since I have school-aged children, good schools in the neighborhood are very important.
3. Active lifestyle: Proximity to parks or other outdoor recreation is important. The ability to walk or bike versus drive to these areas is also important.
4. Diversity of Activities: I would like the neighborhood to have a wide range of venues available nearby. For instance, I wouldn't want all the top venues in the neighborhood to be gas stations or BBQ joints. A wide range of venues such as dining, shopping, and recreation would be important.

Given our previous analysis clustering and segmenting neighborhoods using FourSquare data in New York City and Toronto, how does San Antonio, Texas compare in terms of most popular types of venues? If we wanted to move to a new neighborhood in San Antonio, can we use the FourSquare data for the different clusters to inform a decision on where to start our home search?

2. Data

2.1 Data Sources

New York City and Toronto have well defined neighborhoods that helped us cluster the data. San Antonio has some established neighborhoods, however many of the areas within the city are not defined within a particular neighborhood. Therefore we can't use the same approach as we did with New York and Toronto as we would omit large portions of the city. San Antonio consists of 87 separate zip codes. For analyzing San Antonio we will these zip codes instead and will map and cluster those using the geographical center of the zip code. To get the geographic coordinates we used the website San Antonio AreaConnect which provides latitude/longitude coordinates for the various zip codes around San Antonio. We will cluster these zip codes using the Foursquare location data similar to the analysis in New York and Toronto. Based on the cluster analysis, and our defined search criteria, we will recommend areas to start searching for homes in San Antonio. First we import all the necessary packages to read the data as a Pandas dataframe and plot the geographic data on a map.

To analyze housing prices, we will use data from Zillow Research, which provides data on median home prices over time by zip code or other criteria. This data can assist of in narrowing down neighborhoods based on affordability and also show which growth over time.

For school information, there are multiple organizations that provide information and ratings on primary education. For this project, we will use TxSmartSchools.org. TXSmartSchools uses academic, financial, and demographic data to identify school districts and campuses that produce high academic achievement while also maintaining cost-effective operations. This data may assist us further narrowing our search based on proximity to quality schools.

The final dataset is from the FourSquare API, which I will use to find the most popular venues for each postal code in San Antonio. This data will help me determine the most desirable amenities in my future neighborhood.

2.2 Data Manipulation and Cleaning

The first data source used is the geographic zip code from San Antonio AreaConnect where I got the geographic coordinates for the 87 zip codes around San Antonio. This data was loaded into a Pandas dataframe in Python for further analysis. Figure 1 below shows an excerpt from the data.

	Zipcode	City	State	AreaCode	County	Latitude	Longitude
0	78201	San Antonio	TX	210	Bexar	29.472	-98.537
1	78202	San Antonio	TX	210	Bexar	29.422	-98.466
2	78203	San Antonio	TX	210	Bexar	29.415	-98.462
3	78204	San Antonio	TX	210	Bexar	29.397	-98.500
4	78205	San Antonio	TX	210	Bexar	29.424	-98.487

Figure 1 San Antonio Neighborhood Data

The next data source is the school data from TxSmartSchools.org. This data contains much useful information about elementary through high schools. We are particularly interested in the 'Smart Score', which is the overall measure of the schools rating. Since, we are only looking around San Antonio, I filtered the data for region 20 and also only kept the middle and high schools. The schools dataset is missing postal code information so we need to figure out how to add it to merging the two datasets.

SchoolName	District Id	District Name	County Name	Region Number	Charter School	Alt Ed Type	Alt Ed Campus	Alternate Education	Disciplinary Alt Ed Program	Juvenile Justice Alt Ed	Grade Span	School Type	Composite Academic Progress Percentile (3 Year Avg)	Composite Progress Z-Score (3 Year Avg)	Composite Academic Progress Quintile (3 Year Avg)	Math Progress Z-Score (3 Year Avg)	Math Progress Z-Score	Math Progress Z-Score standard error	Reading Progress Z-Score (3 Year Avg)	Reading Progress Z-Score	Reading Progress Z-Score standard error	TEA Accountability Rating	Smart Score	Spending Index
ALAMO HEIGHTS H S	15901	ALAMO HEIGHTS ISD	BEXAR	20	N		N	N	N	N	09-12	S	9.0	-0.129	1.0	-0.185063	-0.303317	0.110026	-0.073563	-0.081283	0.063798	Met Standard	2.0	Average Spending
ALAMO HEIGHTS J H	15901	ALAMO HEIGHTS ISD	BEXAR	20	N		N	N	N	N	06-08	M	39.0	0.005	2.0	-0.004032	0.036091	0.105000	0.014587	0.034028	0.063195	Met Standard	2.0	High Spending
HARLANDALE H S	15904	HARLANDALE ISD	BEXAR	20	N		N	N	N	N	09-12	S	7.0	-0.148	1.0	-0.191388	-0.113796	0.063124	-0.104438	-0.090119	0.037749	Met Standard	2.0	Average Spending
MCCOLLUM H S	15904	HARLANDALE ISD	BEXAR	20	N		N	N	N	N	09-12	S	4.0	-0.180	1.0	-0.250308	-0.147146	0.063613	-0.109703	-0.113306	0.038534	Met Standard	1.5	High Spending
HARLANDALE ISD STEM ECHS-ALAMO COL	15904	HARLANDALE ISD	BEXAR	20	N		N	N	N	N	09-12	S	57.0	0.065	3.0	-0.057686	0.129964	0.101563	0.188013	-0.119174	0.059987	Met Standard	2.5	High Spending

Figure 2. Schools Dataset with Information on San Antonio Schools

Next I use the *Nominatum* from the *Geopy* package in Python to look up the address for each school using the lat/long coordinates and used the *re* package, using regular expressions to extract the postal code. These were added to the schools dataframe, and then I summarized the Smart Score for each postal code using the mean of all the particular schools scores in that postal code. Now the dataset was ready to merge with the neighborhood data.

	Neighborhood	Smart Score
0	78023	2.5
1	78109	5.0
2	78148	7.5
3	78150	4.5
4	78154	3.0

Figure 3 Cleaned School Data

The next dataset is the housing data from Zillow Research. This data contains monthly median housing prices by postal code (RegionName) from 1996 to December, 2019. Although, I'm interested in how much housing prices have increased over time by postal code, I'm not going to consider this in the analysis, so I will clean up the dataset

	RegionID	RegionName	City	State	Metro	CountyName	SizeRank	1996-04	1996-05	1996-06	1996-07	1996-08	1996-09	1996-10	1996-11	1996-12	1997-01	1997-02	1997-03
0	92271	78130	New Braunfels	TX	San Antonio-New Braunfels	Comal County	23	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	92341	78245	San Antonio	TX	San Antonio-New Braunfels	Bexar County	47	100978.0	100731.0	100679.0	100603.0	100540.0	100590.0	100707.0	100864.0	100871.0	100899.0	100866.0	100986.0
2	92336	78240	San Antonio	TX	San Antonio-New Braunfels	Bexar County	153	105809.0	105650.0	105602.0	105532.0	105570.0	105690.0	105852.0	105877.0	105931.0	106016.0	106071.0	106040.0
3	92345	78249	San Antonio	TX	San Antonio-New Braunfels	Bexar County	381	117740.0	117543.0	117548.0	117642.0	117819.0	118094.0	118342.0	118504.0	118757.0	119003.0	119245.0	119254.0
4	92350	78254	San Antonio	TX	San Antonio-New Braunfels	Bexar County	389	128487.0	128363.0	128211.0	128046.0	127939.0	128066.0	128229.0	128320.0	128359.0	128257.0	128228.0	128172.0

Figure 4 Median Housing Price Data by Postal Code

I removed most of the unnecessary columns from the housing dataset and also categorized the median houses into bins with 1 being $100K-$200K, 2 being $200K-$300K, etc. I kept columns from December 2012 and 2019 so we could compare 5-year price increase.

	Neighborhood	2012-12	2019-12	price_bins	price_labels
0	78130	174114.0	239955	(200000, 300000]	2
1	78245	126281.0	186460	(100000, 200000]	1
2	78240	138430.0	200246	(200000, 300000]	2
3	78249	158196.0	224169	(200000, 300000]	2
4	78254	163446.0	226363	(200000, 300000]	2

Figure 5 Cleaned Median Housing Price Data

We now have the three datasets cleaned and ready to join. Figure 6 below shows the joined Pandas dataset.

	Zipcode	City	State	AreaCode	County	Latitude	Longitude	2012-12	2019-12	price_bins	price_labels	Smart Score
0	78201	San Antonio	TX	210	Bexar	29.472	-98.537	88509.0	156320.0	(100000, 200000]	1	4.0
1	78202	San Antonio	TX	210	Bexar	29.422	-98.466	60016.0	129942.0	(100000, 200000]	1	3.0
2	78203	San Antonio	TX	210	Bexar	29.415	-98.462	71213.0	150560.0	(100000, 200000]	1	1.5
3	78204	San Antonio	TX	210	Bexar	29.397	-98.500	77524.0	137329.0	(100000, 200000]	1	3.0
4	78205	San Antonio	TX	210	Bexar	29.424	-98.487	184158.0	259457.0	(200000, 300000]	2	2.0

Figure 6 Final Neighborhood Dataset for Analysis

The final data set uses the FourSquare API to determine popular venues within each zip code. We will use the same method that I used when accessing the Toronto downtown venue data on GitHub. Figure 7 shows the data which consists of 5,990 rows (venues) and 7 columns.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	78201	29.472	-98.537	Original Donut Shop	29.472703	-98.534598	Donut Shop
1	78201	29.472	-98.537	Restaurant Depot	29.473163	-98.535505	Kitchen Supply Store
2	78201	29.472	-98.537	Pancake Joes	29.464605	-98.543695	Breakfast Spot
3	78201	29.472	-98.537	Taqueria Puro Jalisco	29.479385	-98.541358	Mexican Restaurant
4	78201	29.472	-98.537	Jacala Mexican Restaurant	29.468267	-98.525847	Mexican Restaurant

Figure 7: FourSquare Venue Data for San Antonio

3. Methodology

Using the FourSquare data by San Antonio postal code, I use k-means clustering with k=7 to group each neighborhood according to the most popular venues. I use one-hot encoding on the San Antonio FourSquare data to determine the mean frequency of occurrence of different venue types for each postal code. This provides multiple clusters. I analyze the clusters to determine which have the desirable characteristics such as diversity of venues and proximity to parks and entertainment. Next, I filter out the zip codes based on median home prices and smart school scores. Combining the clusters with the school and home price data provides me with a short-list of neighborhoods to focus for my future home search.

4. Data Analysis

4.1 Mapping the Neighborhoods

I used the folium library to render the San Antonio neighborhood data on a map. Figure 8 shows a map with each of the neighborhoods plotted. Using the map we can see where the downtown area is with the tightly packed circles. We can also see that as we get further from the city center, the neighborhoods get further apart. There are 87 postal codes plotted on this map which doesn't help us narrow down our search for a desired neighborhood. Through clustering and filtering, we should be able to narrow down the search to a more manageable list of postal codes.

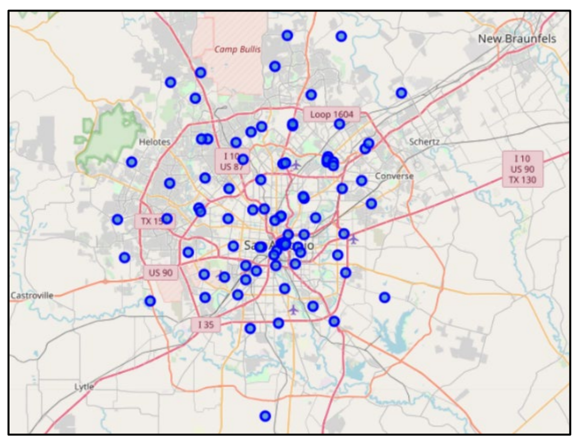

Figure 8: Plotted San Antonio Postal Codes

4.2 K-Means Clustering the Neighborhoods

In the methodology section, I explained how I used one-hot encoding to determine the mean frequency of popular venues for each of the postal codes. Figure 9 shows an excerpt of the dataset after one-hot encoding. We use this data to define the clusters using k-means clustering. We use this dataset to build a table of the most popular venues within each postal code. Figure 10 shows an excerpt of the data.

	Neighborhood	Zoo Exhibit	Accessories Store	Airport Terminal	American Restaurant	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	Ar Entertainn
0	78201	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	
1	78202	0.00	0.000000	0.000000	0.010000	0.000000	0.000000	0.01	0.010000	
2	78203	0.00	0.000000	0.000000	0.014493	0.000000	0.014493	0.00	0.000000	
3	78204	0.00	0.000000	0.000000	0.020000	0.000000	0.010000	0.00	0.000000	
4	78205	0.00	0.000000	0.000000	0.020000	0.000000	0.000000	0.01	0.010000	
5	78206	0.00	0.000000	0.000000	0.020000	0.000000	0.020000	0.01	0.000000	
6	78207	0.00	0.000000	0.000000	0.000000	0.000000	0.013889	0.00	0.013889	
7	78208	0.00	0.000000	0.000000	0.020000	0.000000	0.010000	0.00	0.000000	
8	78209	0.00	0.010000	0.000000	0.040000	0.000000	0.000000	0.01	0.000000	

Figure 9: One-hot endoding on the FourSquare Dataset

	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	78201	Mexican Restaurant	Discount Store	Fast Food Restaurant	Convenience Store	Pizza Place	Burger Joint	Coffee Shop	Ice Cream Shop	Gym / Fitness Center	Breakfast Spot
1	78202	Hotel	Steakhouse	Sandwich Place	BBQ Joint	Theater	Ice Cream Shop	Cocktail Bar	Coffee Shop	Concert Hall	Sports Bar
2	78203	Hotel	Mexican Restaurant	BBQ Joint	Coffee Shop	Burger Joint	Park	Sports Bar	Southern / Soul Food Restaurant	Steakhouse	Juice Bar
3	78204	Mexican Restaurant	Gas Station	Grocery Store	BBQ Joint	Seafood Restaurant	Park	Beer Garden	Trail	Fast Food Restaurant	Brewery
4	78205	Hotel	Bar	Steakhouse	Theater	Plaza	Sandwich Place	Ice Cream Shop	Cocktail Bar	Dessert Shop	Beer Garden

Figure 10: Most popular venues by postal code

When we cluster the San Antonio neighborhoods using the FourSquare venue data, we find that the neighborhoods generally cluster geographically with the downtown generally clustering together (orange), the west, south, and east immediately outside of the downtown area (light blue), north side of downtown (red), and a few clusters outside of loop 1604 (blue). Analyzing these clusters, we can make some generalizations on the most popular venues for each. I summarize each of the clusters in the table below. I experimented with the number of clusters to use. If I use k=4, there isn't much too distinguish between the clusters. With k=6, we get a good mix of clusters, however there are two clusters that have only one postal code. I initially was going to remove these however they are pretty unique and worth separating.

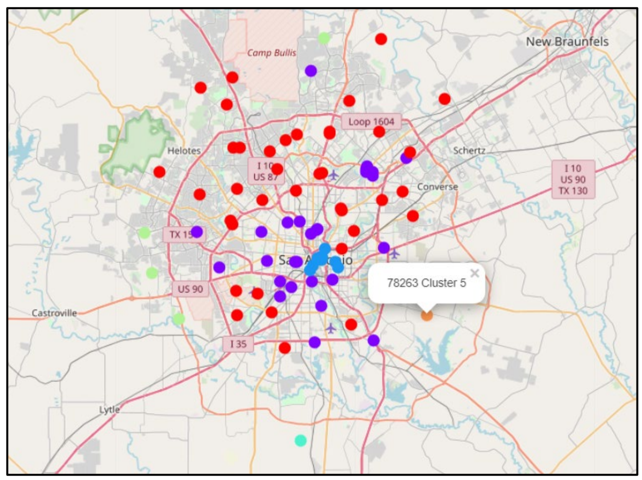

Figure 11: Clustered San Antonio Neighborhoods

Cluster	Color	Characteristics
0	Red	Restaurants of all types, coffee shops, ice cream shops
1	Purple	Fast food and Mexican restaurants, discount stores, bars
2	Blue	Hotels, high-end restaurants, plazas, museums (this is the tourist district)
3	Light Blue	This only contains 1 postal code in south. Massage, studio, zoo, fish market, flea market
4	Light Green	On far west and far north sides. Parks, pools, golf, seafood, pharmacies
5	Orange	Gym, zoo, seafood, flea market, restaurants (this could probably be combined with cluster 3

4.3 Finding the Right Neighborhood to Start the Home Search

We now have our neighborhoods clustered, however this information alone will not help us pick a neighborhood to start our housing search. Next we will use the median house pricing data and school data to filter the data to a smaller subset of neighborhoods. I originally used this data in the clustering but there are several postal codes with missing school data and we don't want to have to exclude neighborhoods because of this. To filter based on median housing price, I set the median housing price between $200K and $350K. Next I review the neighborhoods on the map. Using the filtered data, now I can review the map to find those areas with desirable schools and good locations. Figure 13 shows the final data set of neighborhoods, from this list, I will narrow down to five areas to start my housing search.

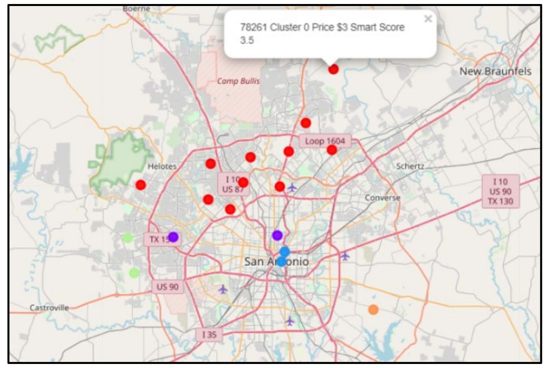

Figure 12: Map of final filtered neighborhoods

	Zipcode	2012-12	2019-12	price_bins	price_labels	Smart Score	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
4	78205	184158.0	259457.0	(200000, 300000]	2	2.0	2	Hotel	Steakhouse	Bar	Theater	Sandwich Place	Plaza	Dessert Shop	Restaurant	Mexican Restaurant	Concert Hall
14	78215	123270.0	227805.0	(200000, 300000]	2	NaN	2	Hotel	Bar	Burger Joint	Restaurant	Cocktail Bar	New American Restaurant	Bakery	Coffee Shop	Sandwich Place	Mexican Restaurant
15	78216	156180.0	233606.0	(200000, 300000]	2	2.0	0	Hotel	Mexican Restaurant	Department Store	American Restaurant	Clothing Store	Cosmetics Shop	Seafood Restaurant	Toy / Game Store	Sporting Goods Shop	Fast Food Restaurant
29	78230	198422.0	278380.0	(200000, 300000]	2	6.0	0	Mexican Restaurant	Coffee Shop	Burger Joint	Sandwich Place	Chinese Restaurant	Sushi Restaurant	Bar	Grocery Store	Fast Food Restaurant	Gym
30	78231	234642.0	326961.0	(300000, 400000]	3	NaN	0	Pizza Place	Pharmacy	Gym / Fitness Center	Gas Station	Video Store	Coffee Shop	Convenience Store	Spa	Cosmetics Shop	Fast Food Restaurant
31	78232	199577.0	280971.0	(200000, 300000]	2	3.0	0	Mexican Restaurant	Convenience Store	Burger Joint	Coffee Shop	Italian Restaurant	Fast Food Restaurant	Ice Cream Shop	Taco Place	Chinese Restaurant	Pizza Place
39	78240	138430.0	200246.0	(200000, 300000]	2	4.5	0	Mexican Restaurant	Video Store	Pizza Place	Chinese Restaurant	Sandwich Place	Discount Store	Salon / Barbershop	Pharmacy	Café	Park
46	78247	142041.0	210761.0	(200000, 300000]	2	7.0	0	Convenience Store	Fast Food Restaurant	Sandwich Place	Gas Station	Video Store	Italian Restaurant	BBQ Joint	Taco Place	Asian Restaurant	Gym / Fitness Center
48	78249	158196.0	224169.0	(200000, 300000]	2	12.0	0	Convenience Store	Fast Food Restaurant	Pizza Place	Sandwich Place	Ice Cream Shop	Coffee Shop	Mexican Restaurant	Sushi Restaurant	Department Store	Tex-Mex Restaurant
52	78253	206318.0	263034.0	(200000, 300000]	2	10.0	4	Video Store	Real Estate Office	Park	Pharmacy	Theater	Food Service	Food Truck	Football Stadium	Food Court	Food & Drink Shop
53	78254	163446.0	226363.0	(200000, 300000]	2	5.5	0	Salon / Barbershop	Convenience Store	Farm	Sandwich Place	Thrift / Vintage Store	Grocery Store	Donut Shop	Pool	Pizza Place	Pharmacy
58	78259	213546.0	286174.0	(200000, 300000]	2	10.5	0	Mexican Restaurant	Sandwich Place	Nightclub	Burger Joint	Pizza Place	Grocery Store	Fast Food Restaurant	Cosmetics Shop	Bar	Smoothie Shop
60	78261	251233.0	313051.0	(300000, 400000]	3	NaN	0	Brewery	Home Service	Construction & Landscaping	Zoo	Food Court	Fish Market	Flea Market	Flower Shop	Fondue Restaurant	Food
62	78263	177750.0	252585.0	(200000, 300000]	2	5.5	5	Construction & Landscaping	Gym	Zoo	Food & Drink Shop	Fish & Chips Shop	Fish Market	Flea Market	Flower Shop	Fondue Restaurant	Food

Figure 13: Final dataset of filtered neighborhoods

Now, I'd like to take this list and come up with a list of areas for my initial housing search. This step is completely subjective, looking through the map and data to find my initial six top search areas.

Postal Code	Median Home	School	Venues
78212	$190K	8.5	Bars, Parks, Restaurants
78231	$327K	NA	Gym, Coffee shops
78249	$224K	12.0	Restaurants, shopping
78251	$197K	10.5	Restaurants
78253	$263K	10.0	Park, theater
78254	$226K	5.5	Personal care, shopping

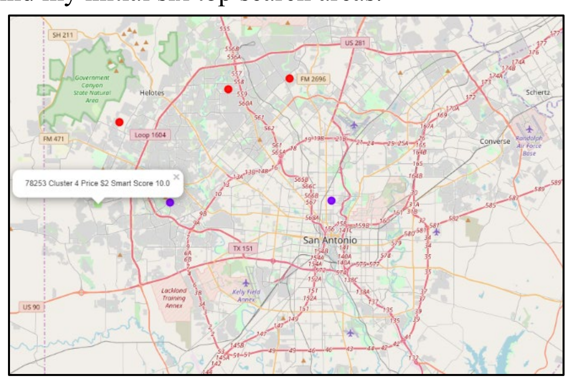

Figure 14: Final list of housing areas

4. Conclusions

In this project, I used multiple data sources to determine initial search areas for finding a new house. Using the FourSquare data along with other research data on schools and home prices can assist us to find the right areas to live based on personal and family preference. Some other important aspects for this research that we did not consider in the research were traffic and commute times, which are pretty significant in San Antonio. Most of the neighborhoods I chose were on the edge of the city where the commute to downtown can run upwards of an hour.

One interesting conclusion that I did not expect was that clustering the FourSquare venue data can tell us quite a bit about housing prices in an area. When I filtered the median housing prices, most of the purple (cluster 1) clusters filtered out. For future research, one might attempt to model median housing prices in a particular neighborhood based on the FourSquare, school scores and other data sources. They could help realtors and home buyers/sellers better determine housing prices to set.

The complete notebook with all code and data files can be find on my GitHub page.