Determining Possible Locations for a Premium Barbershops in San Antonio, TX

by

Olin Kennedy

1. Introduction

1.1 Background

Premium Men's Barbershops are a recent trend in the United States and their popularity is exploding. These barbershops are characterized by tasteful and themed décor, a lengthier list of barber services such as straight razor shaves, massages, beard grooming, and often these places serve complimentary beverages, alcoholic or otherwise. From a business perspective, they have taken a low-margin business and increased margins and profitability by creating and serving a premium niche in the barbershop marketplace.

San Antonio, Texas is the 2nd largest city in Texas. The city is economically vibrant and growing, pulling in migrants from around the United States. Texas is one of the fastest growing states, both economically and population wise, in the United States and San Antonio is well placed to take advantage of both trends.

Additionally, San Antonio is typical in spatial arrangement for a US city, in that the city and its residents rely on personally owned vehicles to get from home to work, places of play, etc. The historic city center is 'conveniently' walkable, but for everywhere else, residents must drive or take a bus to get to where they need to go. A vast majority or residents live in the suburbs of San Antonio.

1.2 Problem

Premium Barbershops are no longer a secret phenomenon and successful first-movers in this arena have been able to expand into multiple locations or franchise out their business model and brand. However, there is reason to believe that the market for premium barbershops is still underserved and, that long-term demand for these services is still growing.

So, given the presence of competition, where should an entrepreneur locate his new premium barbershop? Or where should an existing premium barbershop owner choose to open a new location? Since this space in the marketplace is no longer in its infancy, all the obvious places to locate a premium barbershop are likely already served. Can we identify favorable locations using data?

1.3 Interest

The obvious business interest in this problem would come from businessmen or entrepreneurs looking to open a new premium barbershop in San Antonio, Texas.

2. Data Acquisition and Cleaning

2.1 Data Sources

In order to identify the best places to locate a premium business, we are looking for economic and social data tied to a location. Additionally, we are looking for this high-quality data to be packaged in the smallest and most discrete form available. Luckily, the US Census collects detailed data down to the ZIP code level. ZIP codes are the smallest unit of area data for which there is availability. For example, ZIP codes are smaller than counties and cities, in the context of US political organization.

Zip-codes.com is a data packaging service which is easier to use than the US Census Bureau website and pulls all their data from the US Census Bureau. They are also free, so I used them to pull demographic data for all the zip codes of San Antonio, Texas.

To determine which zip codes fall within the political boundaries of San Antonio, I did a Google search and then saved the table from the website 'zip-codes.com'.

Additionally, for mapping, I needed to find geographical boundaries of each ZIP code. I found a geoJSON of all the zip codes and their shapes/boundaries on GitHub, posted by user 'enactdev' and this was the basis for mapping much of our data.

Lastly, I used the Foursquare API to pull all the Salons and Barbershops for San Antonio, Texas.

2.2 Data Cleaning

The first task was to clean the data that I pulled open-source for the zip codes of San Antonio. I dropped data not relevant to the task at hand such as the Area codes (phone numbers) associated with that zip code and the county within which that ZIP code resided. I also decided to drop the 'population' from this column, opting to pull it later from the zip-codes.com API so I would be working with the most current population data. The most relevant data cleaning I did on this list was eliminating the 'P.O. Box' only zip codes, which are not associated with a geographic area and do not have associated demographic data attached to them. I also changed the way that ZIP codes were listed from strings to integers to make using this data as a callable reference easier when working with other datasets.

Next, I took the list of validated San Antonio zip codes and used it to slice the geoJSON I had for Texas zip codes and their boundaries. I used the mapshaper.org service to do this. I used my list of San Antonio Zip Codes to procedurally generate the JavaScript command to slice the Texas geoJSON into a smaller and more relevant San Antonio geoJSON. With this file, I can later map our data and our model.

2.3 Feature Selection: ZIP Code Demographic/Economic Data

The next step in solving the business problem was to gather the demographic data by zip codes. The zip-codes.com API returns 103 data points for each zip code, so then the task was to select the relevant features about each zip code that will help us predict where a good location to open a premium barbershop is.

Broadly speaking, I selected features that could indicate or answer 3 things:

- What is the wealth of the residents and employees in each zip code? This indicates potential customers who live or work nearby with the prerequisite wealth to afford premium haircuts.
- What social features might affect the proportion of a population who is willing or able to pay for a premium barbershop experience?
- What is the population density of a given area? This would indicate more potential customers.

Based on the above, I selected the following features:

Feature	Reason
ZIP Code	This is the item being described by the other features
Latitude	Location Data
Longitude	Location Data
Land Area	Used to determine Population Density
Population	Used to determine Population Density
Average House Value	Indicator of Wealth of Residents
Income per Household	Indicator of Wealth of Residents
Business Employment	Used to determine average wealth of workers in each zip code
Business Payroll	Used to determine average wealth of workers in each zip code
Median Age Male	Age may indicate a likelihood to use premium barbershop services
Average Family Size	Family Size effects disposable income and thus likelihood to use a premium barbershop service.

I also created a population density feature for each zip code by dividing the population of a zip code by its land area. This also helps control for the fact that ZIP codes are not uniform in size. Lastly, I created an average payroll feature by dividing the number of jobs in a zip code by the payroll of that zip code.

2.4 Feature Selection: FourSquare Data

For the Foursquare data, I queried the API based up the venue category ‘Salon / Barbershops’ and by zip code. From the data returned on this query, I selected venue name, latitude, and longitude. This query returned 2215 items and not all of them were relevant. To make this data useful for our purposes, I conducted the following:

- Eliminated duplicates in the results by comparing the latitude and longitude of the results
- I eliminated salons that were not oriented towards cutting hair (i.e. tanning, makeup, nails, etc.)

- I also eliminated salons that were oriented towards woman and kids.

I did this by looking at the name of the salon. When I was unsure, I conducted a google search of the venue. My method for filtering the results was that I built a list of keywords and names by which to filter the results by Venue name.

When this was complete, I had a list of barbershops where men could be expected to get their haircuts. Then, I needed to split this list into premium venues and regular venues. I used the same methodology as above, creating a list of keywords and venue names to assign a premium label to the premium establishments. Once again, if I was unsure about a venue, I google searched it so I could classify it correctly.

The result of the foursquare analysis is two lists: a 'premium list' and an 'economy' list of barbershops. With these, we can plot the locations of the venues which serves two purposes in this study: validate the clustering and classification of neighborhoods to determine suitability of a zip code for hosting a premium barbershop venue and to show visually to business stakeholders where their competition is located on the map.

3. Methodology

3.1 Type of Model to apply

Our data, primarily zip code data, is without labels, so we do not need a classifier type of algorithm. It would also be possible to use a regression model to take the inputs of our demographic data to build a model where the target variable was a numerical score which would describe the suitability of a certain zip code. However, we still don't have labels, much less quanitifiable labels, with which to know we were fitting the coefficients of our model to predict a target variable.

Therefore, our data is best suited to clustering. We can cluster our zipcodes together based upon their similarities in the demographic data and then apply some human judgement and subjectively classify and then rate each cluster. Furthermore, with the Foursquare data, we can backwards apply that data to see if clustering the neighborhoods fit our purpose and helped us solve this problem. We will apply k-means clustering for this problem.

3.2 Hypothesized model for good places to locate a business

We have various economic, demographic, population density measurements with which we can cluster neighborhoods. Once they are clustered, we can subjectively interpret which clusters would be better for opening a premium barbershop business. Here are the rules I used to sort through clusters and determine good clusters

Population Density	Higher population densities are better than lower, indicating more potential customers in a zip code
Average House Value	Higher House Values were better, indicating richer residents in a zip code. Richer residents are more likely to be premium barbershop customers.

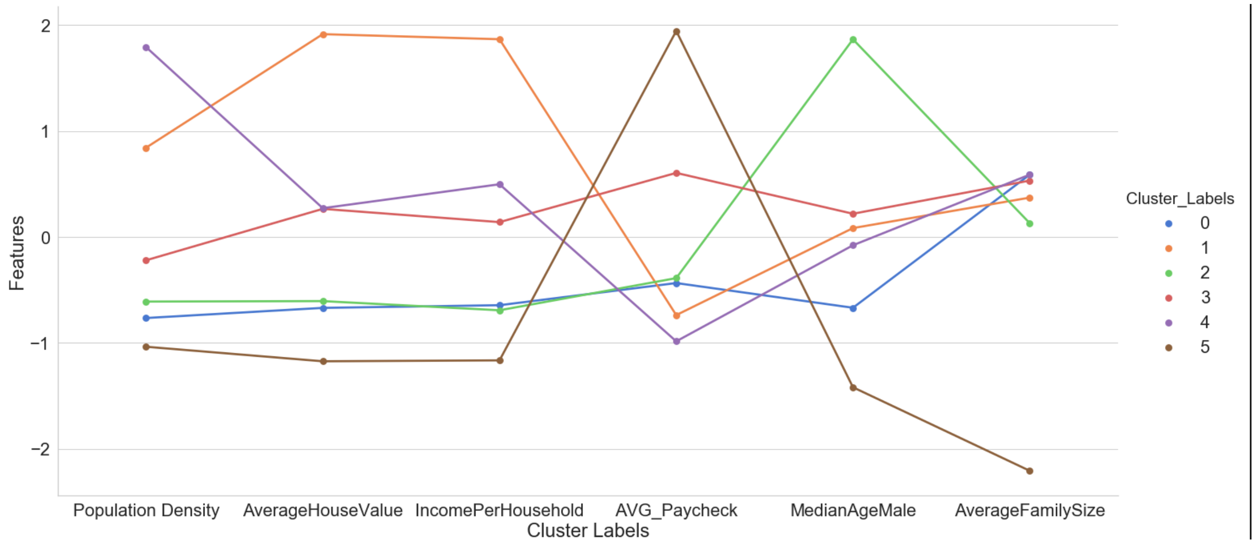
Income per Household	Higher incomes were better, indicating richer residents in a zip code. Richer residents are more likely to be premium barbershop customers.
Average Paycheck	Used as a measure of wealth of workers in each zip code. This is important because workers in a ZIP code are still potential customers of a barbershop, even if they don't live there.
Median Age Male	Age may indicate a likelihood to use premium barbershop services. The theory here is that both very young median ages and old median age indicate a population that may be willing to spend more on premium services because they don't have the responsibility and cost of child-rearing.
Average Family Size	Lower family size would be better because of the reduced responsibility and financial burden of children on our potential customers.

3.3 K-means Clustering based upon the above variables

Of the features I selected, the features can be classified into generally 3 types: resident wealth, employee wealth, and demographic measures of the residents. I also reasoned that broadly, each of these three measures could have either favorable or unfavorable metrics towards opening a premium barbershop. Based on 3 measures with 2 types of metrics, this led me to initially setting my number of clusters to 8 (2^3=8). However, using both 8 cluster and 7 clusters left me outlier clusters which I did not want. I settled on 6 clusters because, in terms of providing a useful map product to the customer on a choropleth map, this provided the best visual product for the customer while not having single zipcode outlier clusters.

Of note, I also normalized my data prior to clustering the zip codes.

Once my six clusters were formed, I plotted the average values of each cluster so I could distinguish its characteristics visually.



With the above plot, I seek to answer the question: Which clusters would be good zip codes to locate our business? And which clusters of zip codes should we not locate our business?

Best Clusters:

Cluster 3 is the only cluster with above average house values, incomes, and high value businesses present in the zip codes. Age is higher, and family size is lower, so I would associate this cluster with older, rich neighborhoods closer to the center of San Antonio. Cluster 3 is clearly the best cluster overall.

Cluster 1 is the second most promising Cluster. Residential Wealth is high, with AVG_paycheck being low. Family sizes are smaller and median male ages are higher. This cluster is most likely a rich suburb full of older professionals.

Cluster 4 looks like a good choice with higher incomes and home values. Similar to cluster 1, there is dip in average paycheck which probably means this is a suburb. Male ages are lower and family size is higher than in Cluster 1. These zip codes are most likely suburbs for younger, up-and-coming professionals.

Okay Clusters

Cluster 2 looks like Zip codes full of old people living on pensions with children moved out of the house. Except for Cluster 5 (outlier cluster), house value and incomes are low, while median age is through the roof. Average family size is very low. On the other hand, retired people may turn out to be a good target demographic for customers if marketed to appropriately.

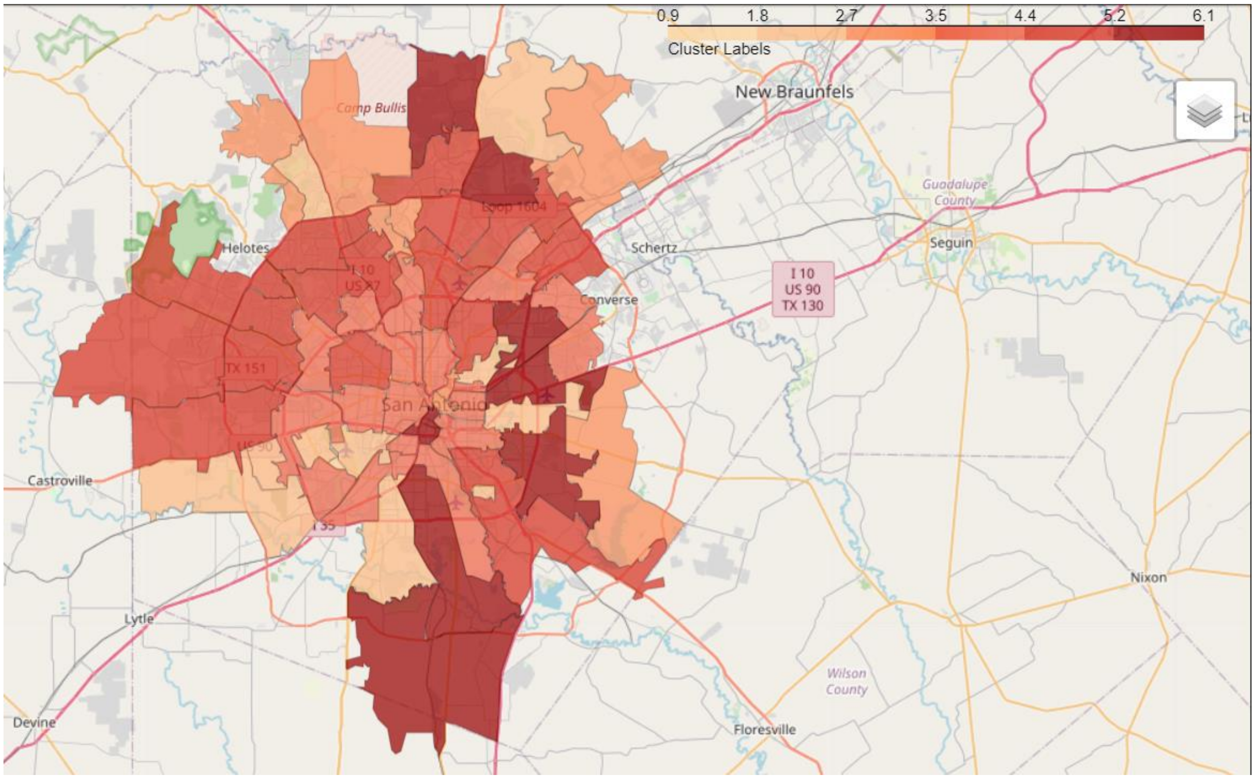
Worst Clusters:

Cluster 0 is below average on all measures economic vibrancy, and makes a poor cluster to locate a premium business.

Cluster 5 is an outlier with only one observation grouped to it. This zipcode is probably a very small area downtown that has no meaningful residential area. This is worst cluster of the set.

Now we can rank order our Clusters from Best to worst as 3,1,4,2,0,5. I want to visualize this data as a choropleth map, with the shading showing the most desirable zip codes by color. But to do this, we'll have to reassign cluster labels to match our worst to best ranking because the choropleth is designed to show a continuous variable instead of discrete variables. I reassigned the best clusters with higher cluster labels and moved the range of cluster labels to be 1 to 6 for easier interpretation (i.e. 3 ->6, 1->5, etc.).

This is what the output of our ranked clusters looks like on the map. Darker shades indicate better areas to locate.

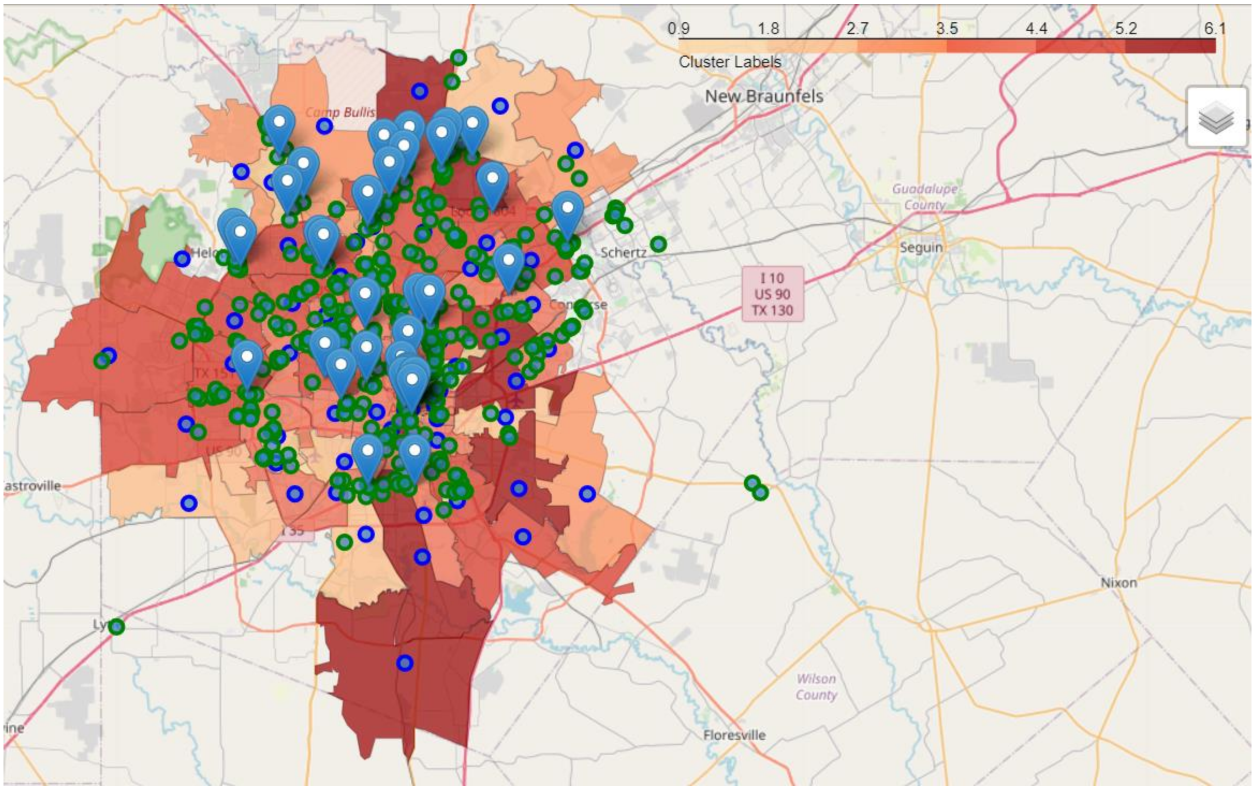


3.4 Adding location data to validate the model and check for competition

Now it is time to pull the Foursquare data and apply it. I already described the method by which I collected and cleaned the data in the data section already, so I won't bore you here with it again. The goal of plotting the foursquare data on the map is to 1) validate the idea that darker areas are indeed more favorable for premium barbershops and 2) checking visually for competition of other premium barbershops.

There are also other things I will do to make the map more useful for the customer, such as adding zip code and cluster labels and a layer for the economy barber shops as well.

4. Results



4.1 Generalized Observations

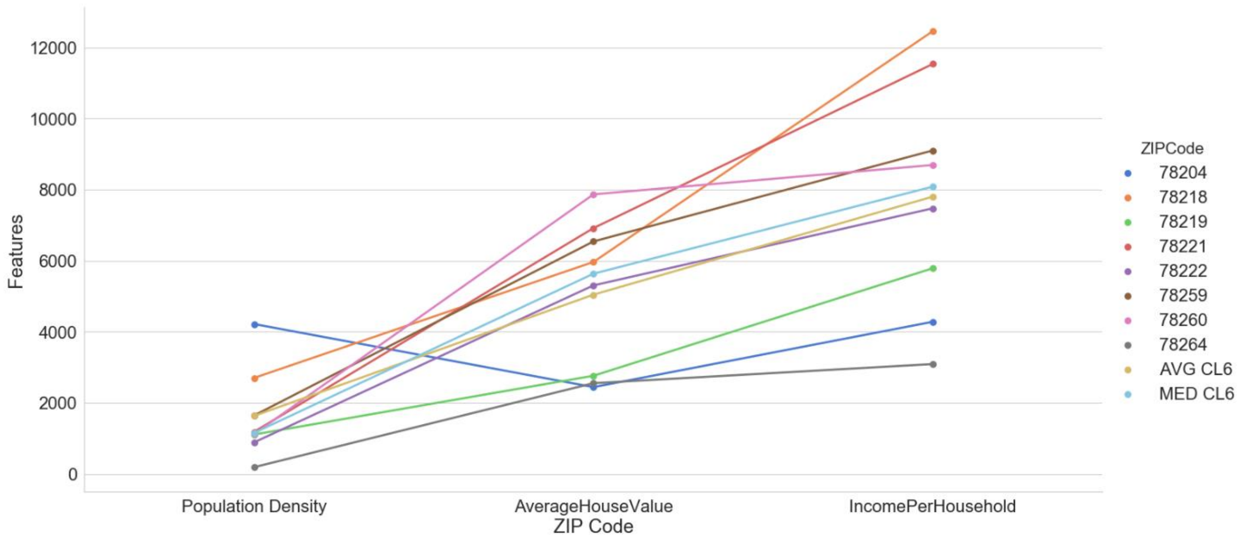
Cluster 5 by far houses the highest number of premium barbershops which means that, subjectively, the k-means clustering seemed to work. Furthermore, a simple majority of the premium barbershops are adjacent to Cluster 6 zip codes. This goes to show that applying some basic demographic and economic data at scale and then working backwards to interpret the Clusters produced a map model that provides useful insight against the business problem. This largely validates our methodology in that we could classify ripe areas for premium barbershops and then see that premium barbershops are indeed located in the ‘ripe areas’.

There is only one instance of a premium barbershop not being fully adjacent to a cluster 5 or cluster 6 area: The northwestern-most premium barbershop which is adjacent to two cluster 3 areas.

Additionally, the most promising zip code for starting a new premium barbershop appears to be 78222 (shown in the picture above), which is the along the San Antonio outer loop. It's a cluster 6 area adjacent to a cluster 4 and cluster 5. Also, the second most promising Zip code would be 78250, which is a cluster 5 zip code adjacent to 5 other cluster 5 zip codes. We will investigate further.

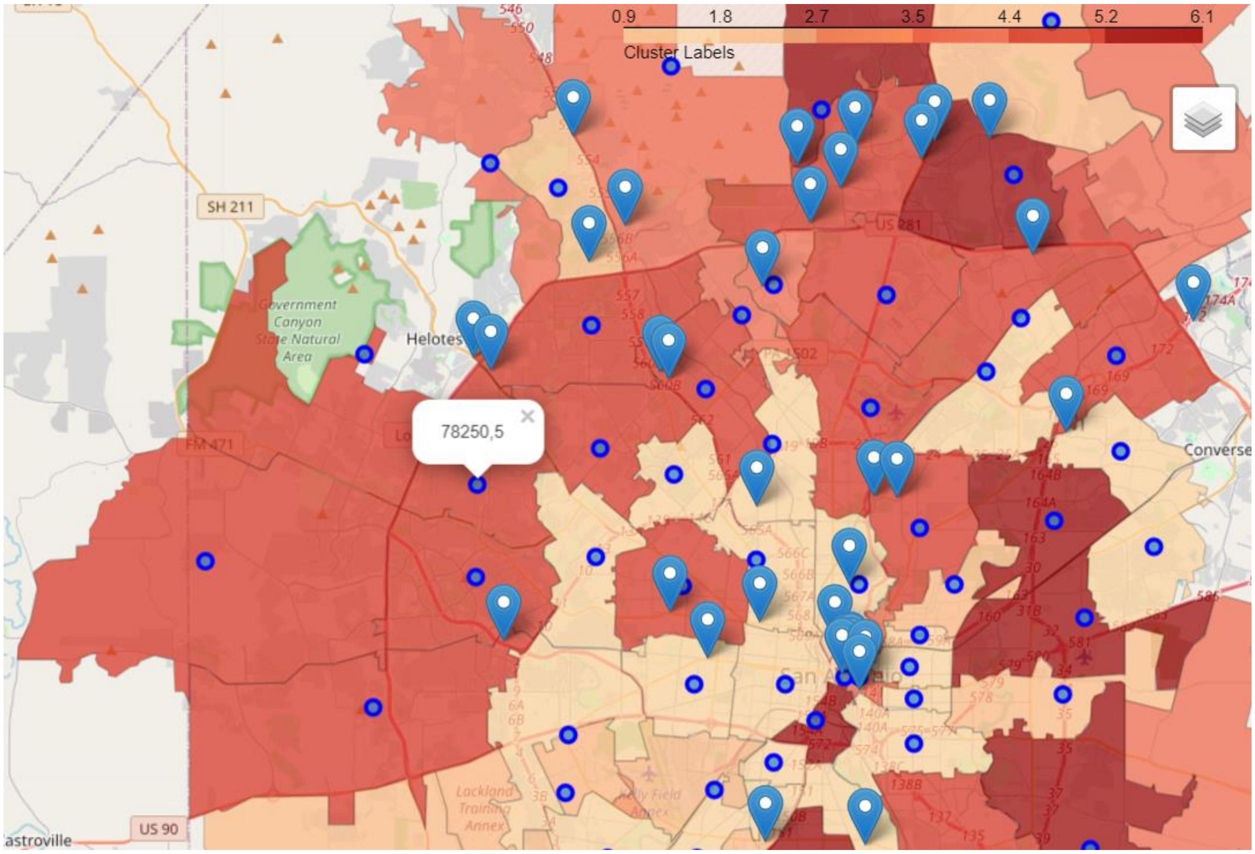
4.2 Unexpected observations and further investigative results (78222 and 78250)

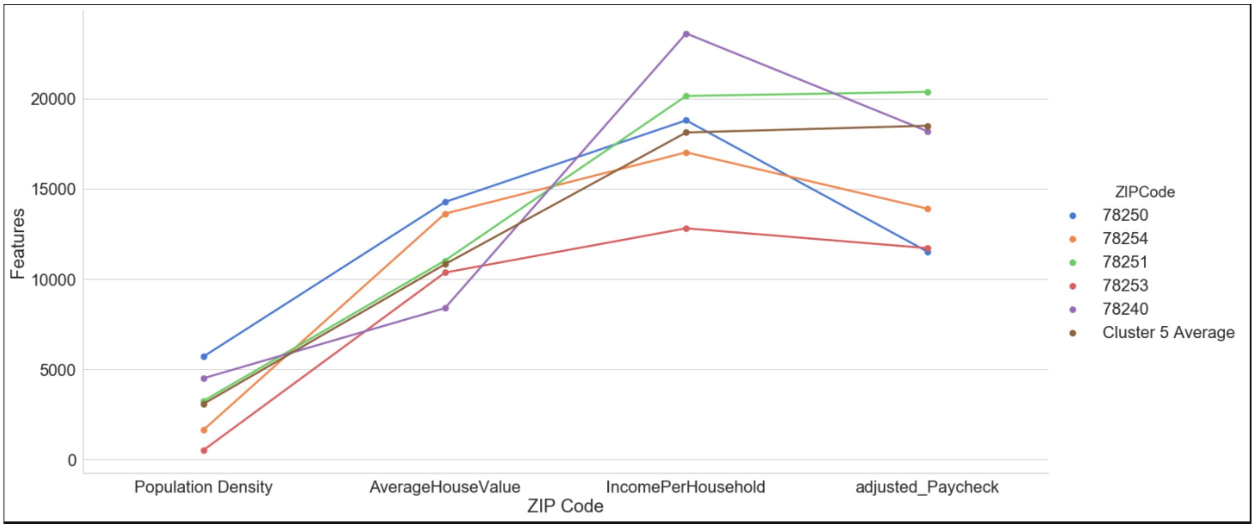
There are very few premium barbershops in our darkest clusters (which is meant to highlight suitability for a premium barbershop). However, many premium barbershops appear to be located on the edge of these dark clusters so that hints that these areas may still be associated with customers. **It** may also be the case that not all the dark clusters are not well classified and that there is high variability within the cluster.



Investigating further, it turns out that within Cluster 6, there are 3 values way below the mean that drag down the average mean and median values of the clusters. While Cluster 6 overall is still very rich and favorable for premium barbershop formation, zip codes 78219,78204, and 78264 are not. Additionally, we can tell from the above graph that 78222 is not, in fact, a good place to open a premium barbershop.

Now let’s examine the western area of San Antonio around zip code 78250.





Only one of the 5 zip codes has below average values for Cluster 5. Because 78250 is in the middle of our cluster 5 zip codes, it looks like a great place to open a premium barbershop.

The mitigating factor with this location though is that there is some competition in the adjacent area, but none in the immediate zip code itself.

5. Discussion

5.1 Insights Gained

It turned out that Average House Value, and Income per Household were the best predictors of where existing premium barbershops would be located. Afterall, this what set up Cluster 5 from the other clusters. Average family size did not appear to have any meaningful effect on cluster formation. It turns out that no matter the demographic of people living in a zip code, family size seems to be pretty consistent overall.

5.2 Recommendations for Clients

78218 and 78250 and their surrounding areas are the best places to open a premium barbershop. Of course, this analysis does not include street level knowledge, knowledge of local business rents, or availability of commercial zoned space for rent.

6. Conclusion

This data-driven study set out to solve a problem for premium barbershop entrepreneurs: Where should I set up my next shop? Along the way, this study also was able to provide answers for, “Where is my competition located?” The results of this study have narrowed down the search area that an entrepreneur would need to do from a city of 61 zip codes to 2 prospective zip codes, saving time and providing value to the entrepreneur. Of course, this study doesn’t account for all variables that go into opening a business in a certain location, but it does make finding the right place easier.