Best venue for barber/stylists and customers in Sacramento, CA

Ruben Ochoa-Banuelos

January 2, 2020

1. Introduction

1.1 Background

The objective will be exploring if in Sacramento would be best to start a business in hair cutting/styling industry. In this notebook we will identify the best prices and ratings within Sacramento. This information with be extracted from FourSquare API and Google places API. Here, we'll identify venues that are fit for various individuals based on the information collected from the two APIs and using Data Science. Once we have the plot with the venues, any company can launch an application using the same data and suggest users such information.

1.2 Interested Audience

This information would be helpful to almost anyone, since almost everyone gets haircuts or styled hair. On the same token, this industry is also highly competitive in some cities while others may seem that there is not enough options to choose from. Therefore, we will be looking at what hair care business have the most options for customers and the least options for potential new businesses to capture that location's business or even freelance stylists to enter the market.

2. Data

2.1 Data Sources

In order to collect the necessary data to complete this objective, we must extract venue information from Foursquare's API and from Google Places API. First, we fetched venues up to 10 miles out in order to capture a decent sample size.

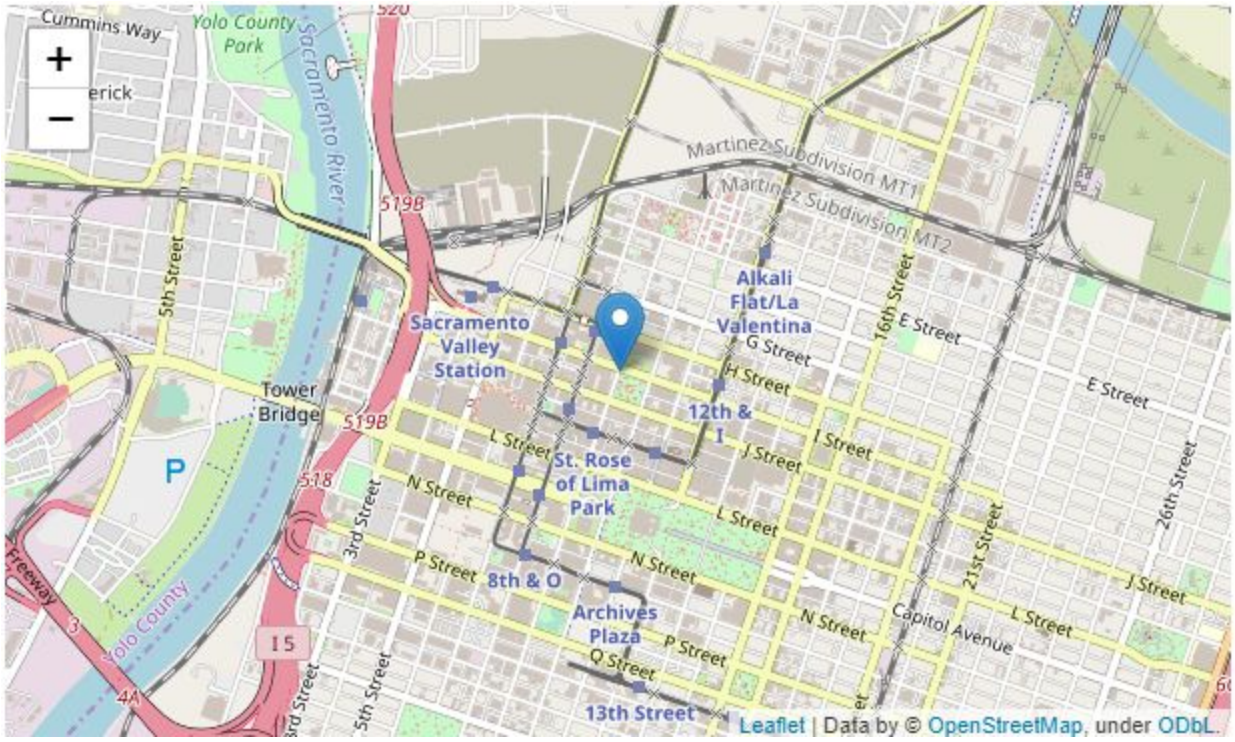

Figure 1:Target city coordinates validated using Folium

Foursquare API was fetched from this URL: (https://api.foursquare.com/v2/venues/search). The values extracted were the following:

- Name: The name of venue
- Category: The category given as defined by Foursquare
- Latitude: The latitude of venue
- Longitude: The longitude of venue

Google Places API was fetched from this
URL: (https://maps.googleapis.com/maps/api/place/nearbysearch).
The values extracted were the following:

- Name: The name of venue

- Category: The category given as defined by Google
- Rating: The average rating of the venue
- Rating total count: The total number of ratings given by consumers

2.2 Data Cleaning

The data extracted from Foursquare contained more columns than needed so we had to drop some of these columns to clean up the dataframe as they provided information that were made obsolete by the latitude and longitude column values, such as address or county. We then manually looked through and validated venues that seemed off by a simple Google search. Turns out these were mis-categorized by Foursquare API. Fetching data from Google Places API showed that only pages of 20 can be called at one time, so we had to merge that data into one data frame. The next step was to look at the data fetched from Google places API and manually look through the venues the same way we did for the Foursquare data. Once that was complete we merged the two data frame while excluding any duplicate values within the indices. After the final comb through the data for duplicates, odd information and NaN values changed to 0, only crucial columns were kept for the objective. This leaves our sample size being **54** observations.

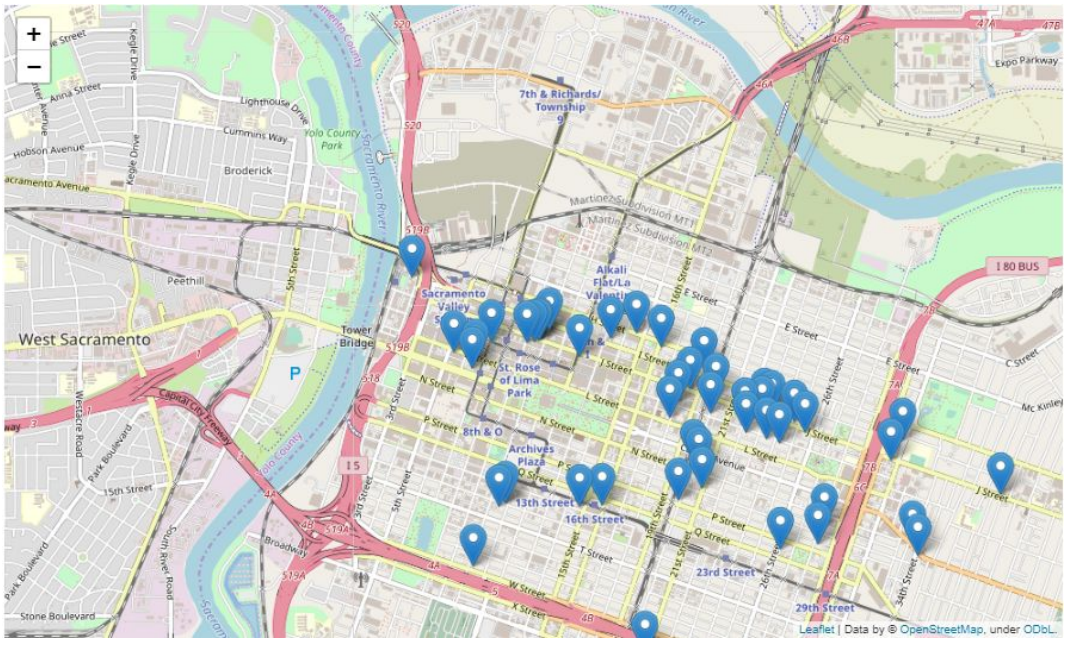

Figure 2: Shows spatial concentration of venues

name	rating	user_ratings_total
Allure Salon & Spa	4.7	27.0
The look threading salon	4.9	16.0
Salon Entro	3.4	9.0
Michael Z Salon	5.0	10.0
Double Take Hair Gallery	5.0	7.0
Lust Beauty Bar	0.0	0.0
Nina's Studio	0.0	0.0
Sola Salon Studios	4.7	36.0
Melissa Moreno Studio	0.0	0.0
Five1Nine Salon Suites	5.0	1.0
Benefit Cosmetics BrowBar Beauty Counter	4.0	4.0

Figure 3: Final Data aggregated from both API's

3. Methodology

This project aims to identify ratings and characteristics of of the barbershop/salon market in Sacramento. This would enable potential participants to see market concentration and possible patterns in branding and how it relates to their rating. This will also show anyone in the market looking for venues with products or services that are reliable and worth visiting.

The first step was to extract data from the **Foursquare API** for the venue names up to 10 miles. We also extracted average rating and rating count from **Google Places API** and combined this with the Foursquare API data into one dataframe.

I decided to keep some venues that were relating to wigs or hair replacement. There was one duplicate so I dropped that from the dataframe. Even though some venues are also in the 'Health' market, I'm sure individuals would buy the product/services based on the looks of the end result(eg. wig or hair replacement color, length, style).

After cleaning the data, the next step was to was to see the concentration of venues on a map and a plot graph to show distribution of ratings for visual analysis.

Next we noticed **similarities** with the naming of the venues, so we decided to group these based on **keywords** (ie. hair, salon, etc.) in their name for later analysis.

For analysis we will see if the keywords have any significant relation to the rating total and the rating itself.

Finally we will discuss our findings in our conclusion.

4. Exploratory Analysis

We decided to plot the 'ratings' against the number of venues to visually inspect the distribution of ratings and noticed something worth investigating.

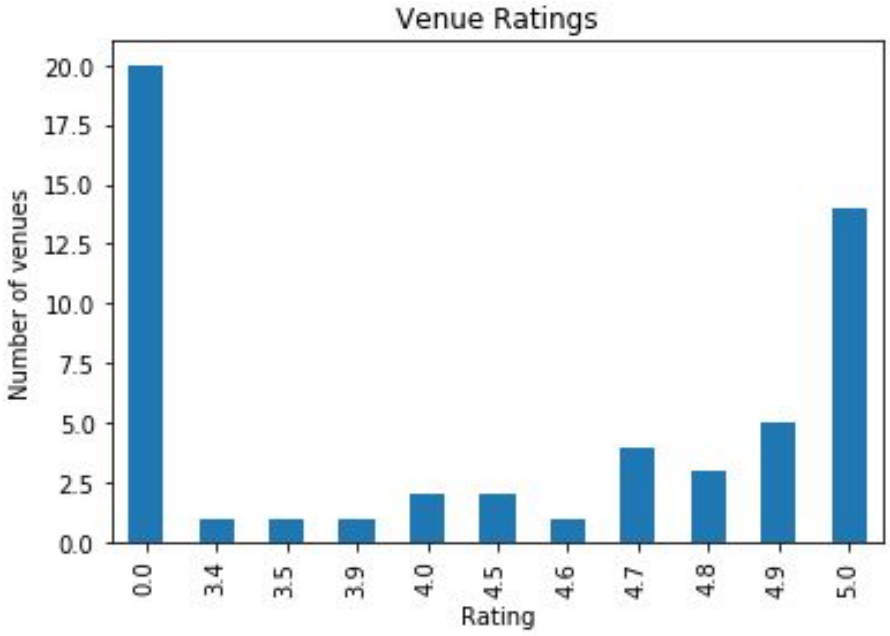

Figure 4: Distribution of ratings

Very quickly we notice that the majority of the distribution are at the extremes. it looks like it's either 5 or 0 (no rating). Let's see if there is any pattern or reason for why majority of the venues are so heavily distributed at the extremes.

4.1 Clustering like Venues

We decided to cluster the venues based on similar keywords in their brand name to compare number rating responses and how well they were rated. Maybe we can find out if there's any correlation between brand name and their rating count and value. The other category are the names that do not contain a common keyword relating to the other three. If a venue contains more than one keyword, the first keyword listed is counted under the first keyword's count.

	name	rating	user_ratings_total	Group
0	Allure Salon & Spa	4.7	27.0	1
1	The look threading salon	4.9	16.0	1
2	Salon Entro	3.4	9.0	1
3	Michael Z Salon	5.0	10.0	1
4	Double Take Hair Gallery	5.0	7.0	4
5	Lust Beauty Bar	0.0	0.0	2
6	Nina's Studio	0.0	0.0	3
7	Sola Salon Studios	4.7	36.0	1
8	Melissa Moreno Studio	0.0	0.0	3
9	Five1Nine Salon Suites	5.0	1.0	1
10	Benefit Cosmetics BrowBar Beauty Counter	4.0	4.0	3

Figure 5: Column added Cluster venues

	Name	Average Rating
0	Salon	4.2
1	Beauty	1.4
2	Studio	2.9
3	Barber/Haircut	3.2
4	Other	2.5

Figure 6: Cluster by similar keyword in name

One of the first things people look for when choosing barber shop/salon is the average ratings and how many people rated the location. So we decided to start off to see what groups have the highest average rating.

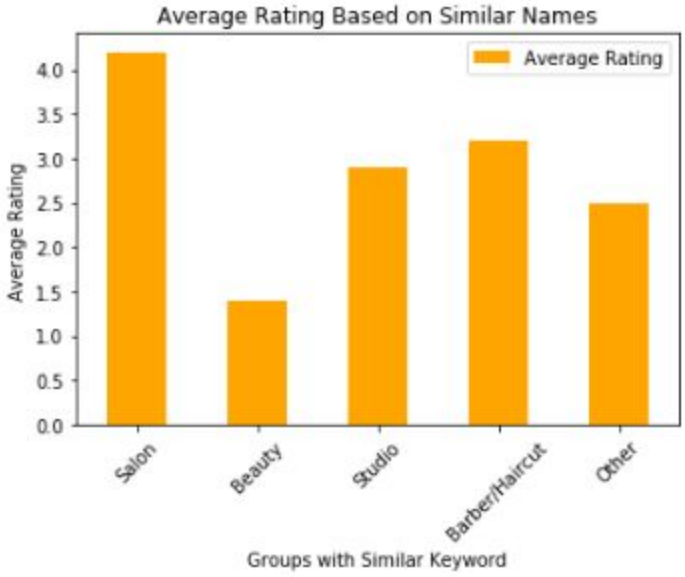

Figure 7: Average ratings per group

Ratings with low count and high rating can seem unreliable to consumers. So we decided to take a look at the Total number of ratings and noticed a clear lead in the number of reviews and the previous leader taking a step back into the second/third position with a little more than a third of the leader “Barbershop/haircut” group.

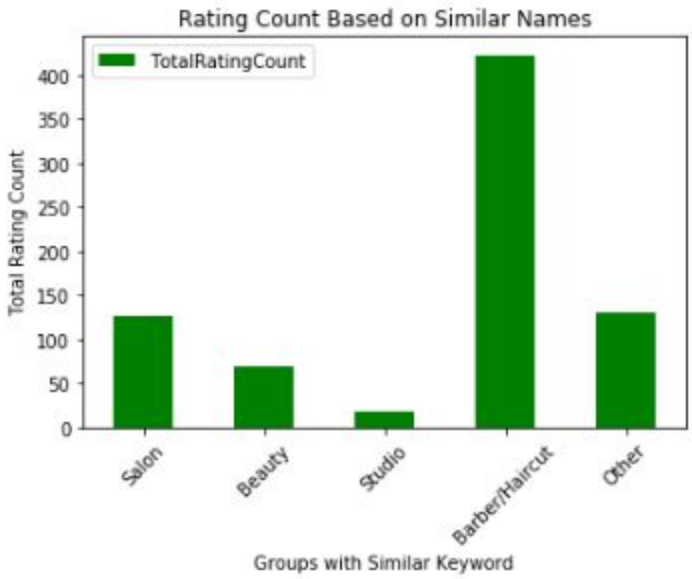

Figure 8: Total rating count per group

This graph indicates that the group with the most participation for ratings and when all groups are normalized, the barbershop/haircut group scores overall best. Which can mean to a potential consumer that these venues are a safer and more satisfactory choice. For a potential participant looking to enter the market, it looks like venues with these keywords in the brand, tend to receive more ratings and seem safer and satisfactory as indicated earlier. It would be difficult to quantify and test for what the consumer thinks is a "safe" choice. We decided to see it might be that this type of group is still dominates in the market, as indicated by venue count belonging to this group, and is still growing. We will test this with a predictive model using **k-Nearest Neighbors**.

4.2 k-Nearest Neighbors

For this method of Machine Learning we took our cleaned data frame and removed the non-numerical value columns to see if we can predict, with high accuracy, the next venue type to enter the market. Our next step was to normalize the data during the preprocessing stage so that we can construct our test and training sets by splitting 20% and 80% respectively. From there we established our k value at 3 and then we created our "neigh" model by fitting the training splits for testing. The next step we predicted the next ten values and defined this as our "Yhat". We then calculated the accuracy of our training and testing sets, as shown below.

```
Train set Accuracy:  0.9534883720930233
Test set Accuracy:  0.8181818181818182
```

However, we wanted to make sure that we chose the best k value for our model so we ran a for-loop to test every k value from 1-10. The clear winner was actually the first three runs as shown below in the output array and in the plot.

```
array([0.81818182, 0.81818182, 0.81818182, 0.63636364, 0.63636364,
       0.63636364, 0.63636364, 0.54545455, 0.54545455])
```

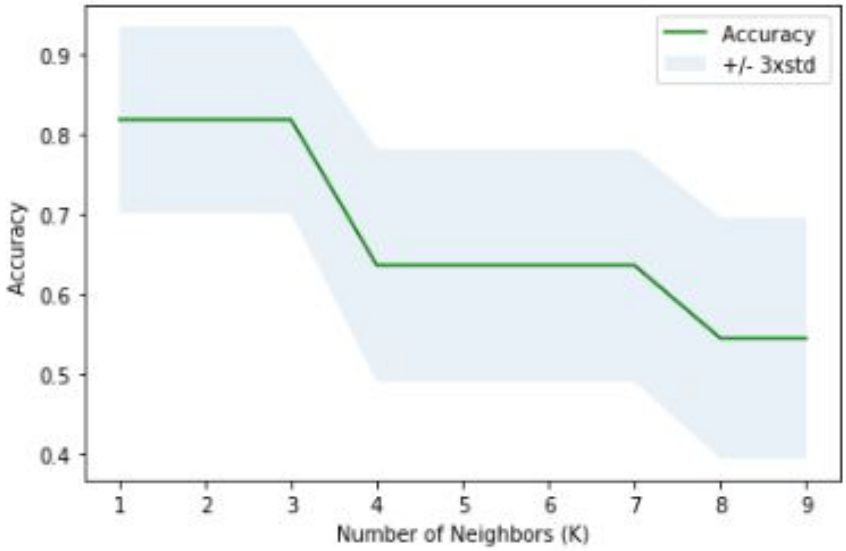

Figure 9: Visualization of k-value accuracy

5. Conclusion

Overall, we have identified the highest rated venues for consumers and potential employees wanting to enter the market with a high rated employer. We also found that some of the venues with higher ratings have very few rating counts that might suggest low traffic. We have found a few interesting things when taking a deeper look into this data. For example, in a market of haircut/stylists, venues with more traditional nomenclature have more ratings, higher ratings when normalized and have the most venues per group. In the end, we would have like to have more insightful data to check the impacts of these ratings on foot traffic and potential revenue.