Capstone Project

Predicting rating of restaurant in California, Nevada, Utah and Arizona

Cristian Vildosola June 2020

Objective

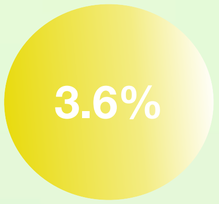
Our approach to predict

Our main goal is to develop a machine learning model able to predict the rating of a restaurant based on different variables. Being able to predict a business to succeed help us to determinate the best possible locations and amount of positive reviews in order to find the formula of a successful restaurant.

Predicting the rating of a restaurant

Of restaurants fails their first year

Restaurants close before their 5 year anniversary



Market growth in 2019

The most common cause of failure are location and user reactions in social platforms.

Data Acquisition

Where do we get the data?

In order to obtain the data, we used the API offered by Yelp, specifically the business search.

Then we choose the following variables:

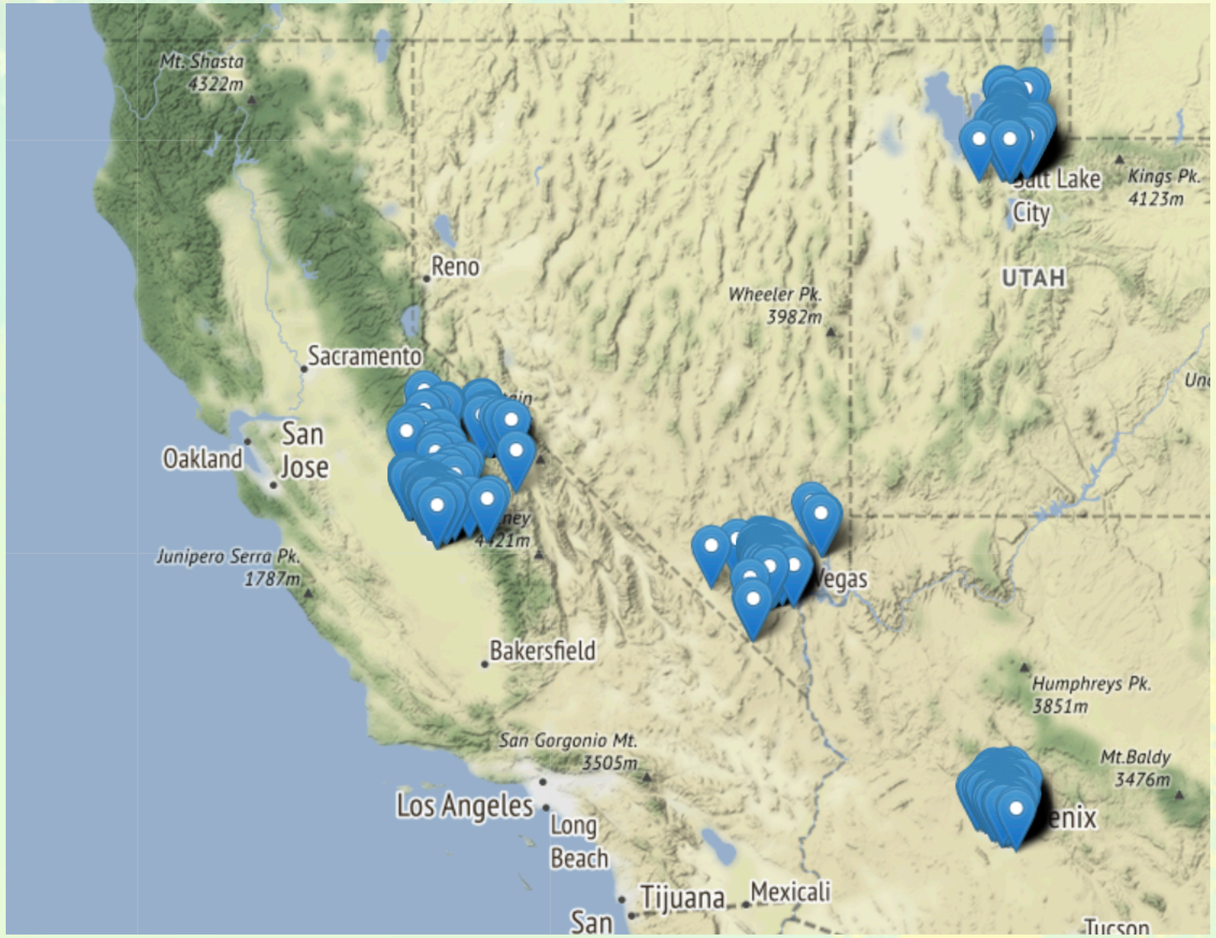
- City
- Price
- Reviews
- Rating
- Latitude
- Longitude

https://www.yelp.com/dataset/

Exploring the data

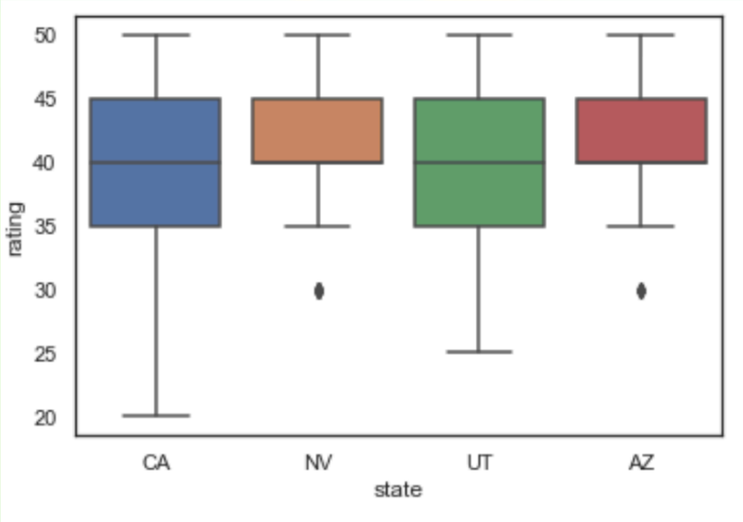
We got data from 4 states of United States, these are: California, Nevada, Utah and Arizona.

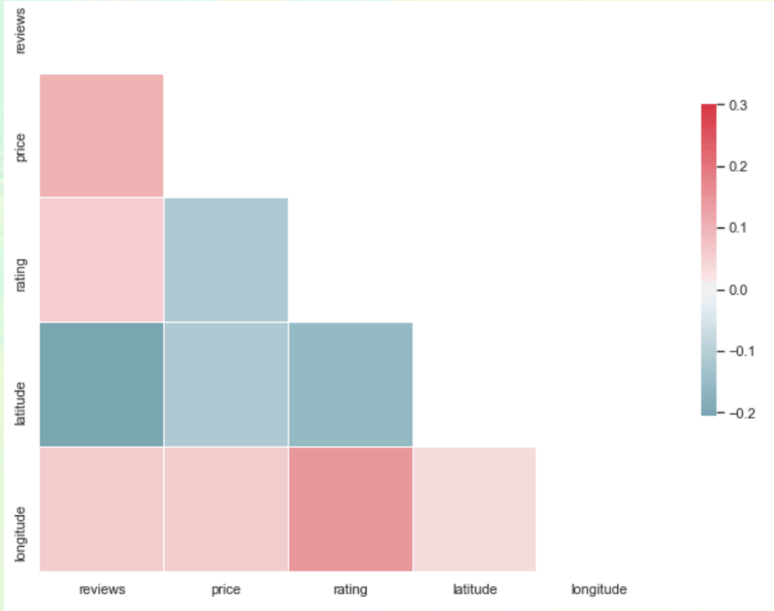
For each place we have around 500 restaurants, so in total we got almost 2000 places to explore.



Exploring the data

In the next plot we can explore the distribution of the rating and the correlation of all the variables selected.

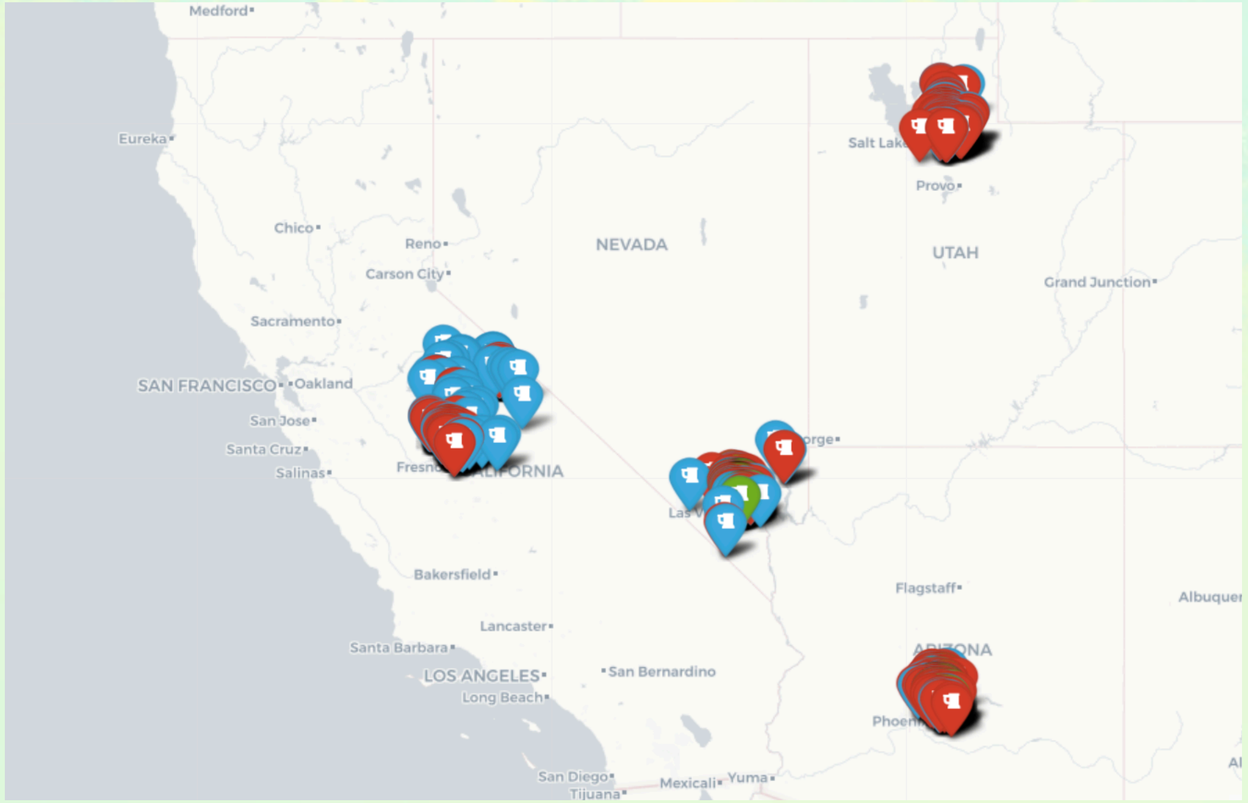


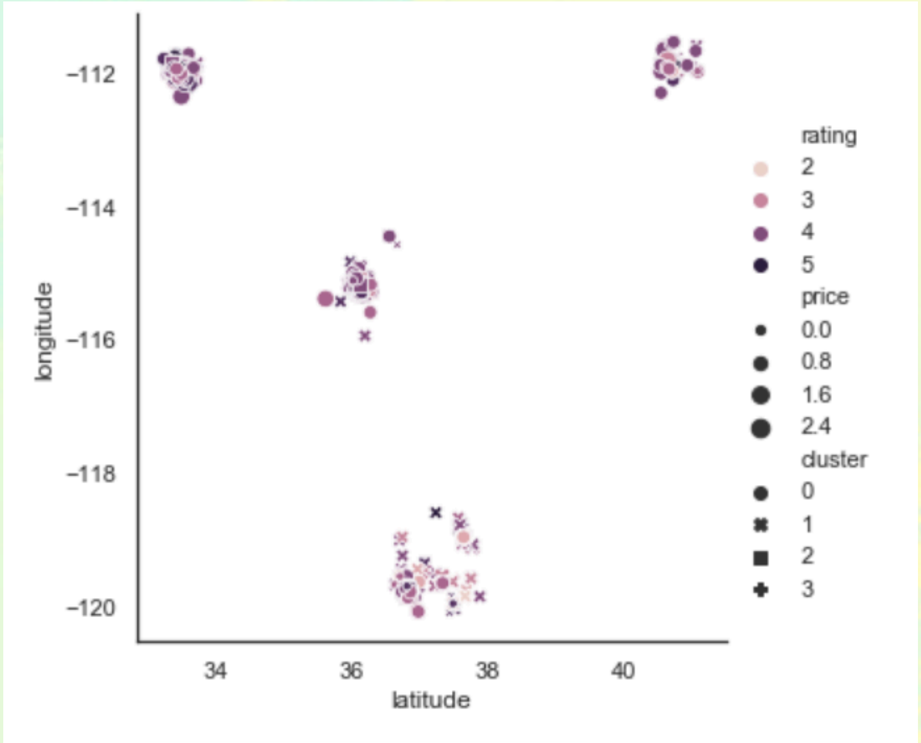


With this information we can say that the data is very uniform because all the states has similar behaviors, but the correlation of the variables it seems to be not strong enough.

Clustering the data

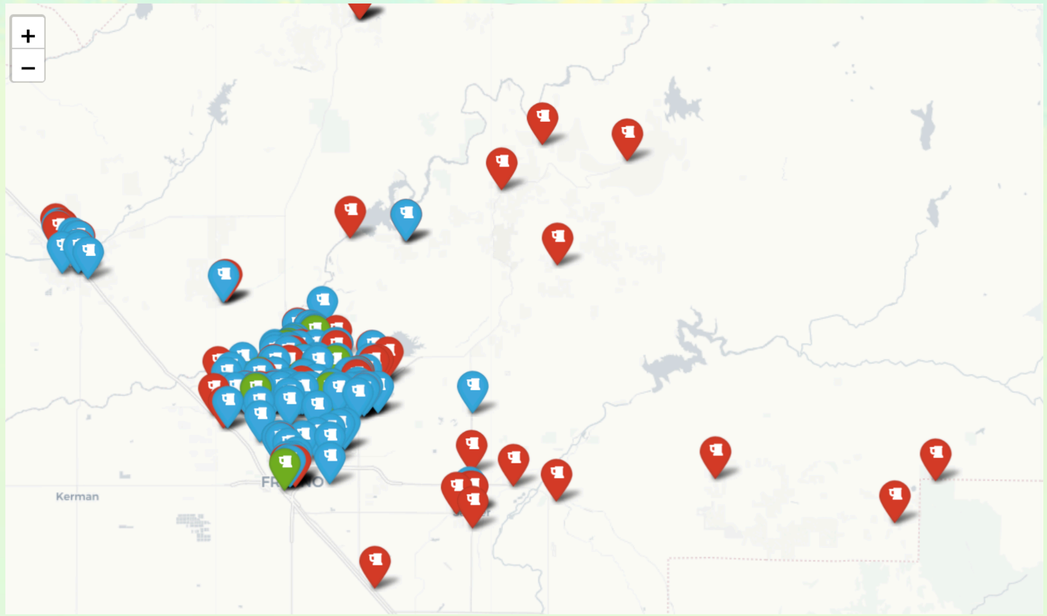
In our first cluster attempt we tried to show how the restaurants are distributed, in the following map it can be seen.

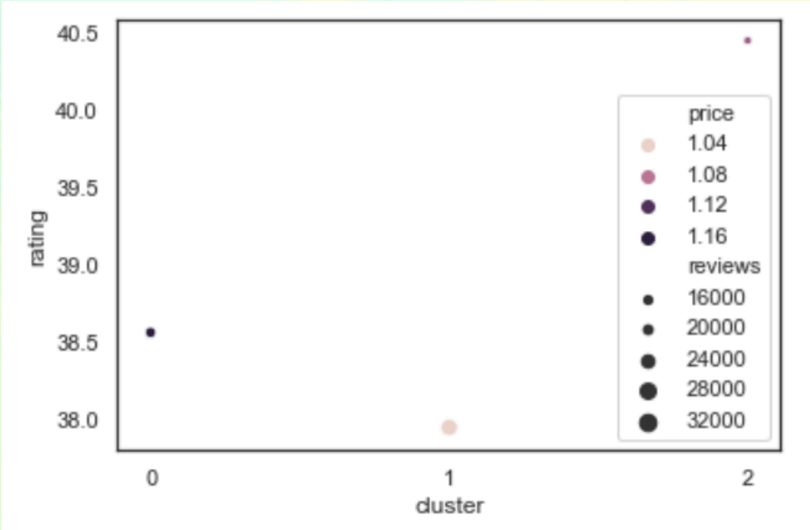




Clustering Again

In our second cluster attempt, we didn’t use the whole data, instead we selected the data just for the restaurants located in California.

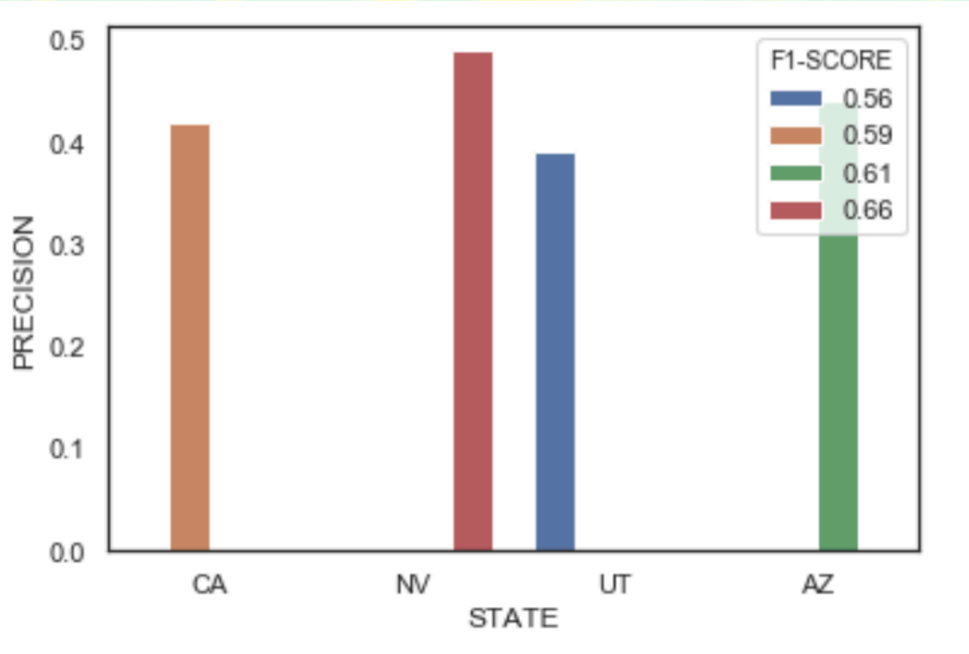




We repeated this process for all the cities, in the map, we see that clusters are mainly represented by amount of reviews, the colors of each cluster are represented by, red, blue green and pink for respectively 0, 1 and 2.

Predicting

For predicting the rating of each restaurant, first we used a multiple linear regression and then a support vector machine algorithm, so the results were:



	STATE	PRECISION	RECALL	F1-SCORE	RSS	VARIANCE_SCORE
0	CA	0.42	0.96	0.59	30.17	0.03
1	NV	0.49	1	0.66	20.44	-0.03
2	UT	0.39	1	0.56	11.77	0.02
3	AZ	0.44	1	0.61	12.59	-0.04

As the plot shows, there is not relevant prediction in the model, so our predictions isn’t good enough.

Conclusions and next steps

From the experiment we can visualize a good clustering classification of the users, in which we can get the best restaurant by price in the state clustering, by the other side the city clustering allow us to get the restaurants by the amount of reviews and the rating. The downside of the analysis is in terms of predicting the rating of each place. The model didn't perform well, we knew this when we didn't recognize cor relationships between the variables, also the number of data for each variable make the model more complex to recognize the any pattern within the data.

The next steps, could be take more data and try to separate it in more equal sized partitions to test the variables, another point to upgrade is maybe choose more related to the problem and try bigger data sets.