Explaining the problem:

Choosing a good wine on variety or name alone is difficult. With the vast amount different wines in existence grown from every corner of the world, it is almost impossible to select a good wine without prior knowledge or recommendation guide. We use reviewer ratings as a predictor of quality. The goal of this analysis is to see what characteristics of wines lead to high ratings using qualitative factors such as region, variety and quantitative factors like price.

Approach

The dataset we used is “winemags” reviewer ratings for wine. This dataset reviewed over 130k different wines from across the globe. It contained information such as price, rating, country, variety, and winery. My goal was to help determine which qualities correlated to ratings.

Data Engineering

The price variable had missing values. To solve this I determined that using the median price would be the best approach as that would prevent skewing the data.

Because the amount of wines coming from different countries was so large and would overfit the model, I further sub categorized the regions into larger geographic entities.

Furthermore, because there were so many different variety of wines I decided it would be best to categorize the wines into larger categories of Red, White, and ..

From there in order to fit the geographic categories into a correlation matrix, I created binary values of 0 and 1 for each geographic region and wine type.

Because of the large range of points and price I created columns for mean center points and Z score to better assist in plotting

Findings

There is a linear relationship between price and rating. A 0.4 correlation.

While the price is the strongest factor in the prediction of rating, we can conclude that wines from certain region score higher even in the same price range.

As we can see wines originating from Europe and North America score higher across the entire range of prices.

Recommendations

Recommendations