Wine Classification Case Study

Isabella Sturm

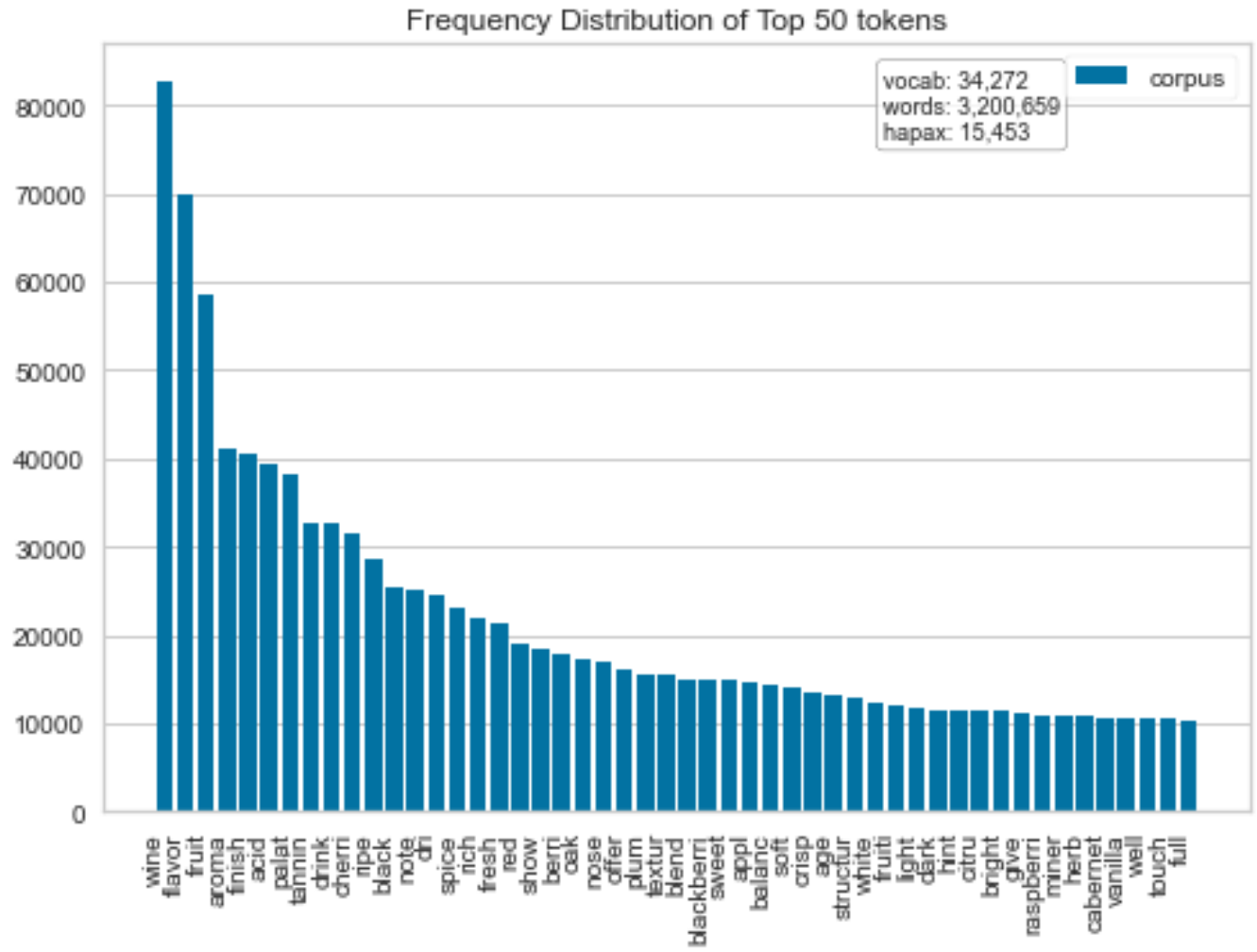
DSC 550 Final Project

2021-06-04

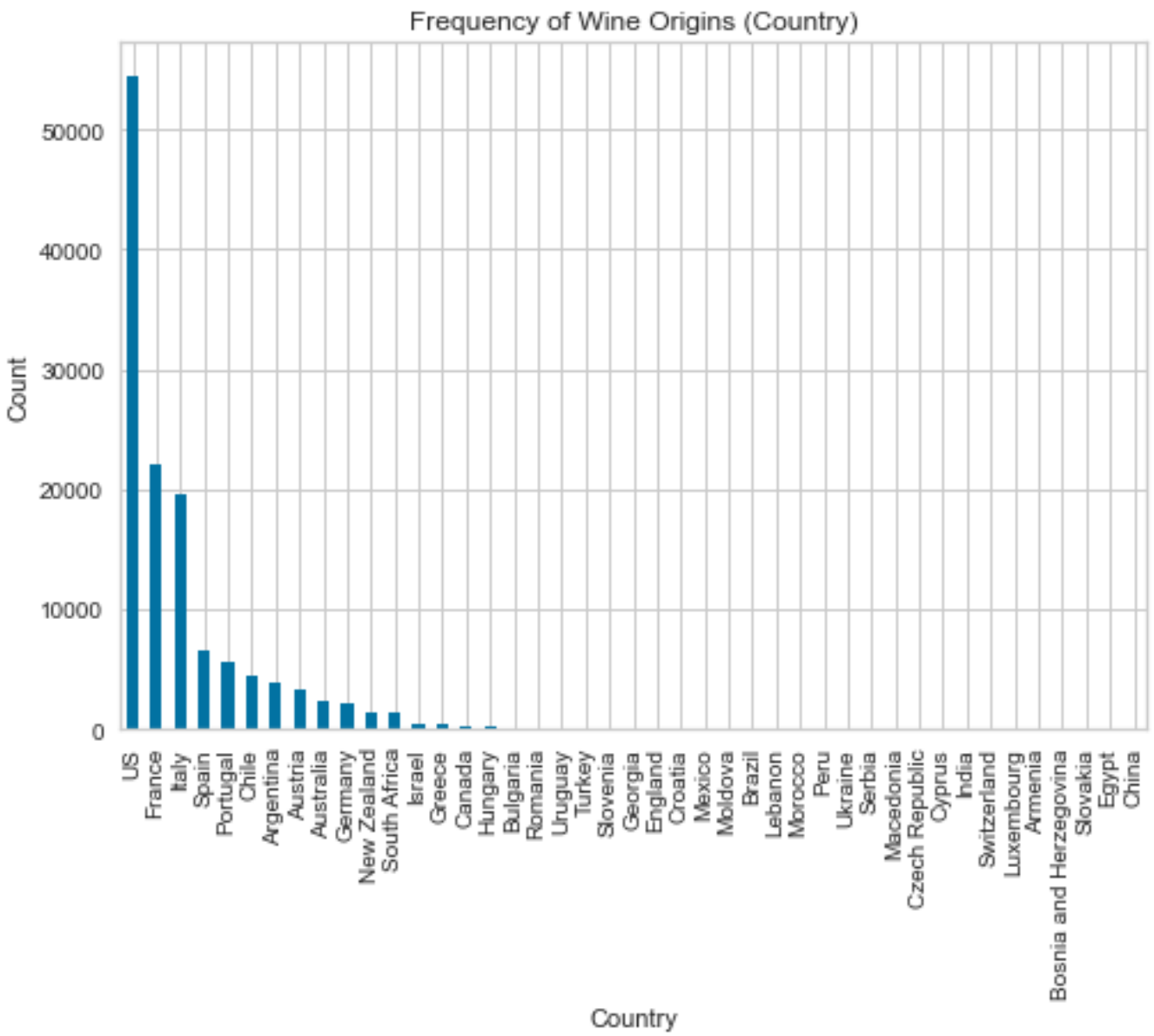
Narrative

There are thousands of known wine grape varieties and many counties or regions that produce wine. The prices can range from as low as one dollar to more than a half million dollars. There are many factors that go into the price of a wine and the quality of the wine including the wine variety used, the region it was used, the age of the wine, and the process used during growing, soaking, fermenting and the overall production of the wine. The goal of my analysis is to use price, region, variety and wine reviews/descriptions to determine quality (I will use points) of the wine. For enjoyers of wine, this analysis is useful to determine which wines or types of wines are worth their prices.

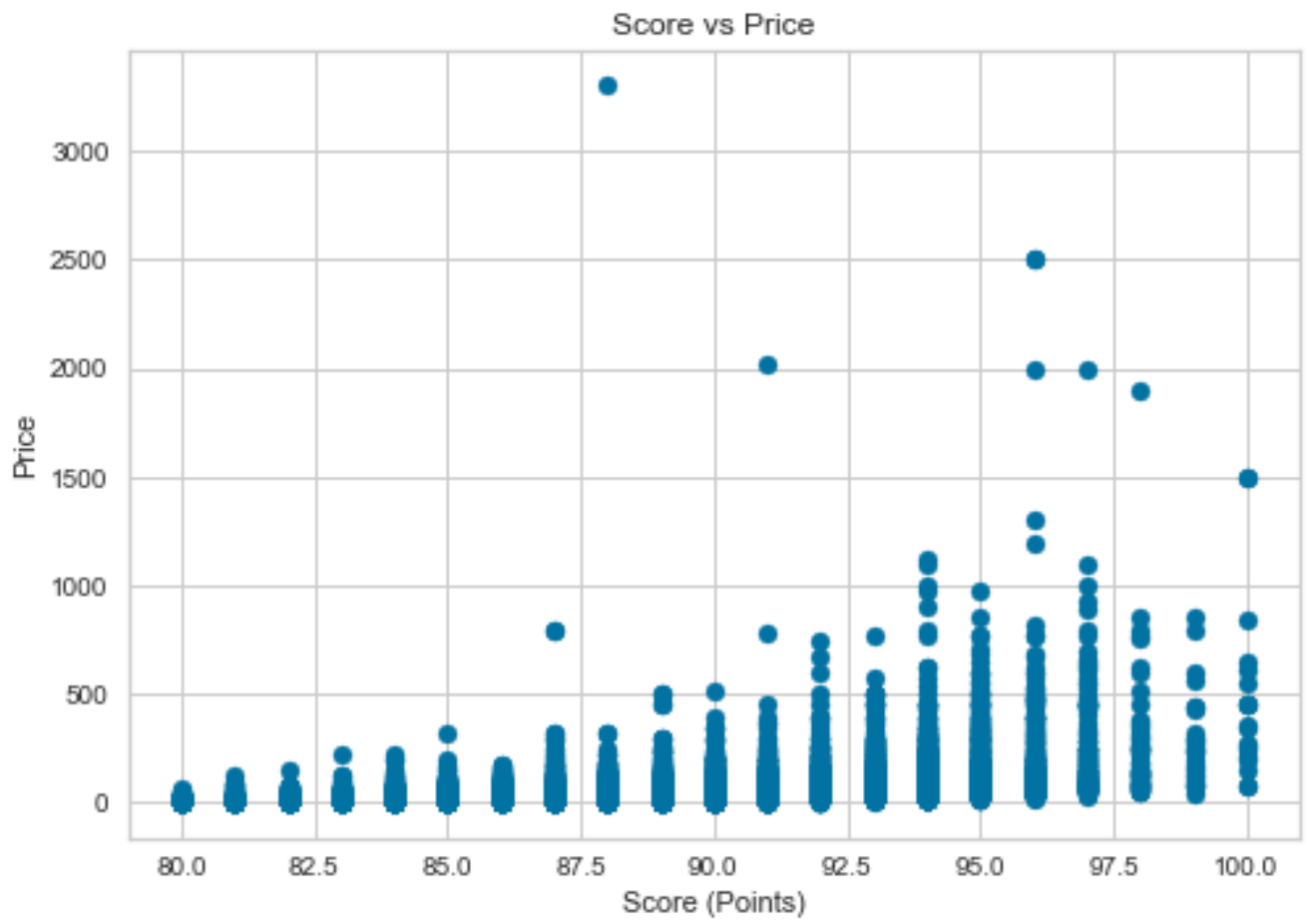
Graph Analysis



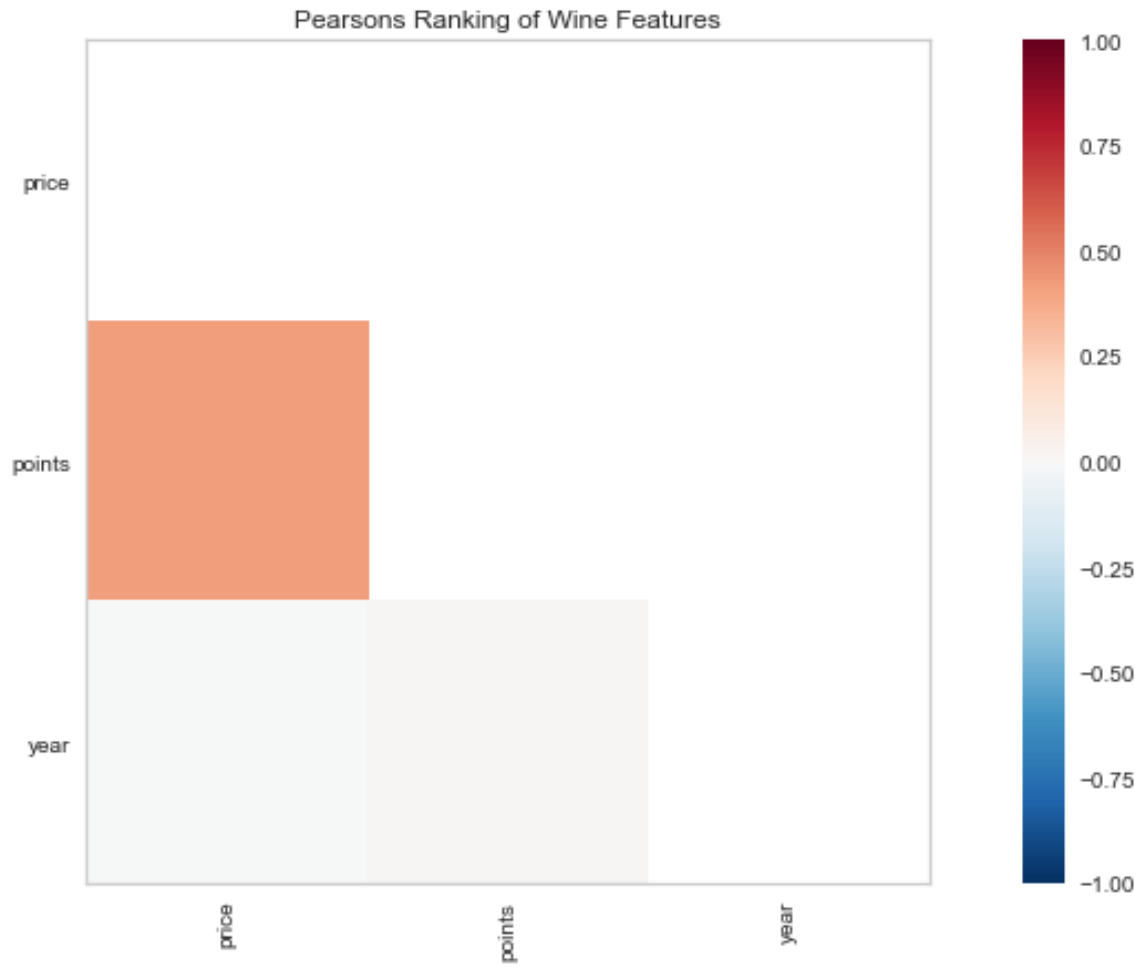
This graph shows the top 50 words found in the descriptions of the wine. This is useful for showing the top features of wine (aspects of the wine most commonly described) that come up most often and can be used in further analysis to show which features equate to the highest quality of wine.



This graph shows the most common countries the wine in the data set comes from. It is clear the data will need to be further cleaned (but also largely biased to only the large wine regions) to remove outliers of smaller countries who don't produce much of the wine in the world.



This graph is useful for showing how the score given to a wine relates to the price of the wine. Generally, it seems that you can get a pretty quality wine for a very affordable price.

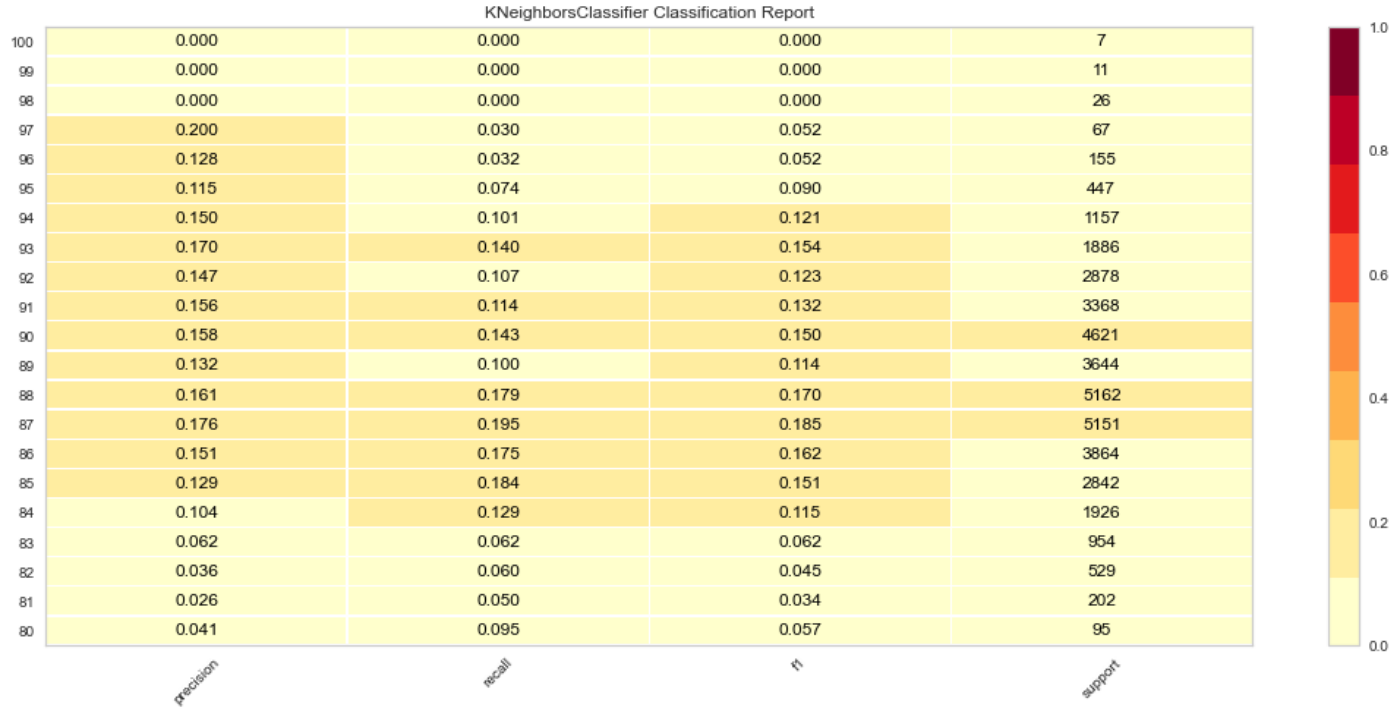


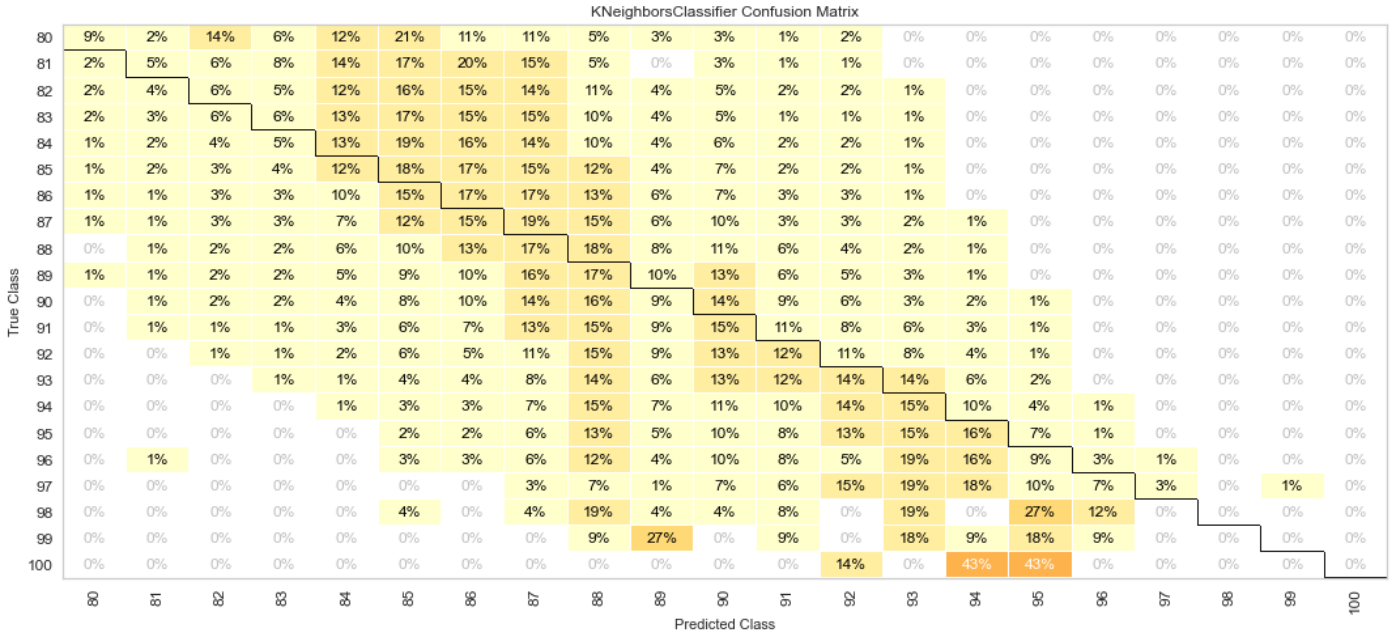
This graph shows the correlation between three features of the wine – price, points (score) and year (vintage). From this graph, it seems there is no correlation between year and price or points (score) but there is a pretty strong (medium) positive correlation between points and price.

Model Selection & Evaluation

I attempted to use K Nearest Neighbors to classify the wine data but the model trained did not show very strong results for the success of the classification, as shown in the Classification

Report and Confusion Matric shown below:





I chose KNN thinking that a classification model would be best for determining the quality of the wine based on similar features in the different wines. Also, I was thinking the scores are not very continuous and therefore it would make more sense to use a classification model over other types of model that are trained for continuous data.

Conclusion

Overall, there is more analysis that needs to be done, but it does seem there is a correlation between the score of the wine and the price of the wine but that is not always the case. However, it does seem that the vintage of the wine, based on the dataset we have here, does not really affect the quality of the wine or the price of the wine. However because our data consists of primarily newer vintages the starting dataset is skewed.

More text analysis could and should be done to figure out what qualities in the wine, as noted by the reviewers contribute to the quality of the wine. Also, some challenges I faced that should be explored further are other models that might work better for the data and features selected.