JJ Goh
Ishan Supernekar
Tim Tan
Brian Kim

Wine Data Set Write up

Intro:

Our group's motivation for our project is: "to gain a deeper insight on which wine variables deem a high wine score/rating on WineEnthusiast." As a group, we thought that if went to random people on the street and even avid wine college enthusiasts, we would expect to hear several different reasons why a wine would be highly rated. This dataset gave us the opportunity to create models/plots that would closely examine the relationship our several variables with points. Additionally, this information is extremely useful for wine companies and distributors. If a wine corporation can map out the few variables that affect wine points the heaviest, they can make sure to incorporate those variables into their wines, whether they are a certain variety of wine or from a specific region.

Data Cleaning:

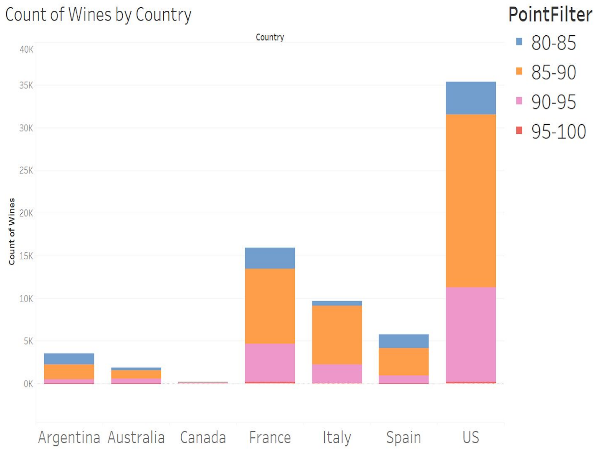
For our data cleaning, we took out five of our features (x1, description, region2, taster_twitter handle, designation) because we felt that these features were either too specific or redundant. With the remaining features, we omitted every missing value which left us with about 77,267 observations in our data set. We then filtered all of the wines to have prices below $1000. Then, we factorized all of the categorical values which left us with far too many features to handle.

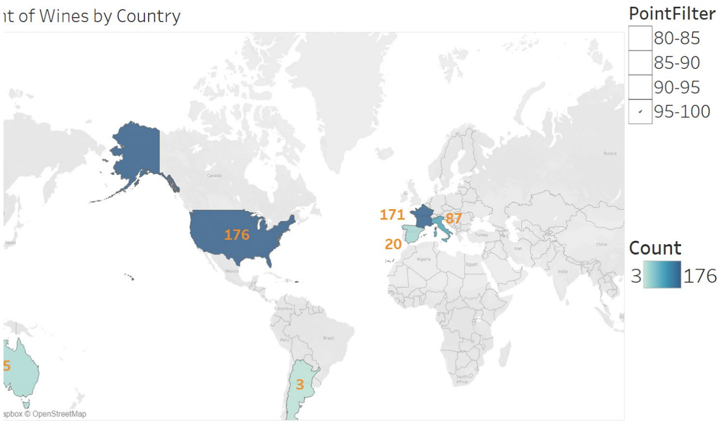
In order to combat the large number of features, we lumped together categorical variables with over 20 factors. For province, we mutated a new category of province_ordered with the function, fct_infreq(variety). Then, we lumped the factors into only the top 10 categories with the function, fct_lump(province_ordered, 10). Here, we used the variety_ordered category in order to keep the top ten factors of provinces. We then repeated this process for the categories, variety, region1, and winery. With the factor levels lumped in our data, this leaves about 70 variables to handle for our dataset.

Next, we extracted the years out of the description category. We used regex to look for series of numbers from 0-9 from the descriptions. After creating a years category, we converted it from character to numeric. Then, we removed all the missing values and numbers in which there were more than 4 from the year category and filtered out from years below 1821. We had to filter the older years because we saw that some wine bottles had a single number to represent a version of a wine. This seemed like the year was before the year 100 which was not possible. Finally, we created a new data frame with the numerical categories, lumped categories, and the year. We also split the training and testing set as 75/25.

Baseline Analysis Plot(Tableau):

Next, we used Tableau in order to illustrate the distribution of points throughout each country. As observed, every country maintains a similar point distribution with 85-90 points as a majority count of wine and with 95-100 points as minority wine in each country. What's interesting to note is that the USA has by far the most wines in the dataset, however, has close to an identical count of 95-100 rated wines as France; even though France has roughly half the wines in the dataset. This comparison is measured in the map chart(see figure:) This could lead to further inspection whether French reviews are heavily biased or the wine in France justifies a very high rating. Here are the plots created:

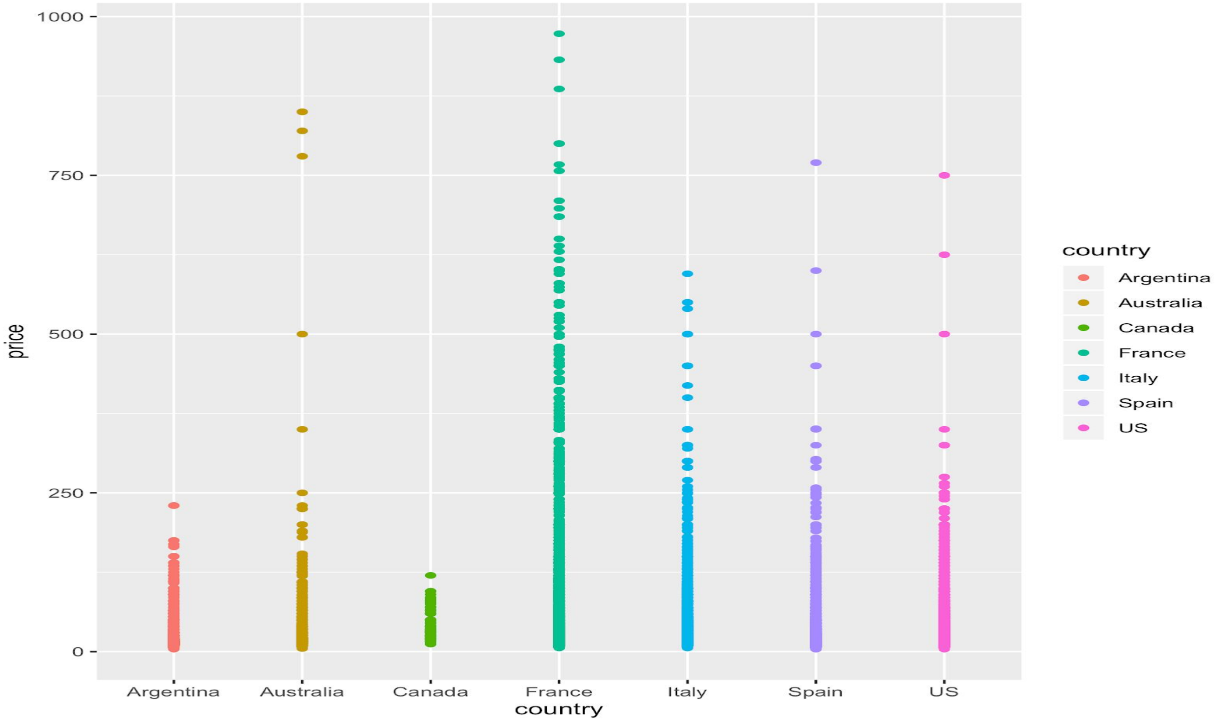




Backward Stepwise:

After fitting a backwards stepwise model and summarizing it, it is obvious that price is the variable that affected points the most. It can be assumed that price and points have a high correlation and studies has shown that critic's scores had a statistically significant effect on wine prices (GWS, 2018). The next variable that affected points the most is wine taster name (Michael Schachner) followed by province (California).

Ggplot price v. Country:

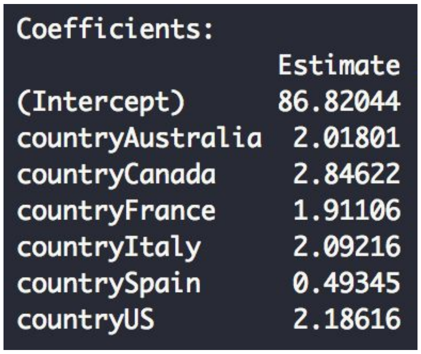


A plot of price against country was plotted to show the distribution of the prices (refer to appendix) of wines relative to their country. It is noticeable that wines from Europe are relatively more expensive than wines from the rest of the world, with the exceptions of some outliers from the US and Australia.

Linear Regression:

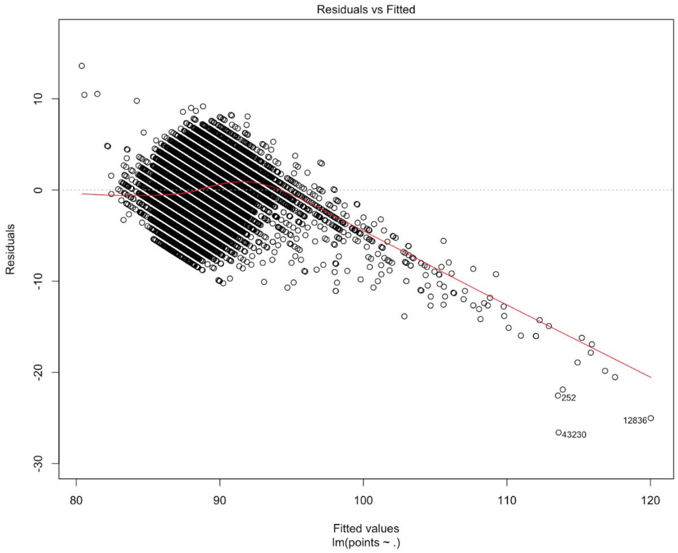
A simple linear regression was used to predict points against country (refer to appendix). Something interesting to note here is that wines from Australia, Canada, and US will score about the same points and even more against European wines. For example, if a wine from Canada is selected for review, it will score 2.846 points more compared to 1.911 points more if the wine is from France. However, something to take note here is that the amount of wines reviewed from Canada is significantly a lot less compared to the rest of the world. This might be one factor that is causing the high discrepancy in points. If everything else is ignored, this result can eliminate the preconceived notion that wines from Europe are ‘superior’ to New World Wines (wines not from Europe).

OLS:



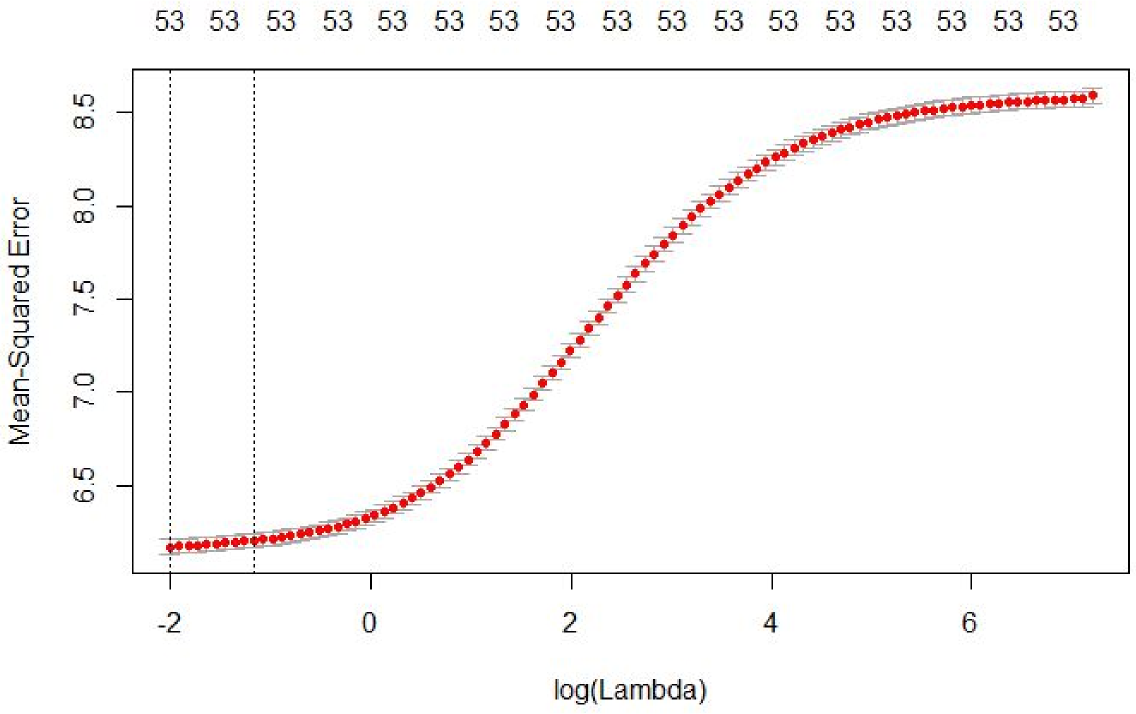
```
Coefficients:
                 Estimate
(Intercept)      86.82044
countryAustralia  2.01801
countryCanada     2.84622
countryFrance     1.91106
countryItaly      2.09216
countrySpain      0.49345
countryUS         2.18616
```

The OLS model fitted has an R Squared value of 0.301. This means that this model is capturing 30.1% of the noise/variance of the training data. In other worlds, this OLS model is able to explain 30.1% of the variability around the mean of the training data. While R squared is the measure of fit, RMSE is the absolute measure of fit and predicts how accurately the model predicts response. In this model, RMSE scores about 2.45. The lower the number, the better the fit. In this Lastly, MAE is the average magnitude of errors in our prediction, and likewise to RMSE, the lower the number the better the fit. The residuals of the OLS model is located in the appendix.



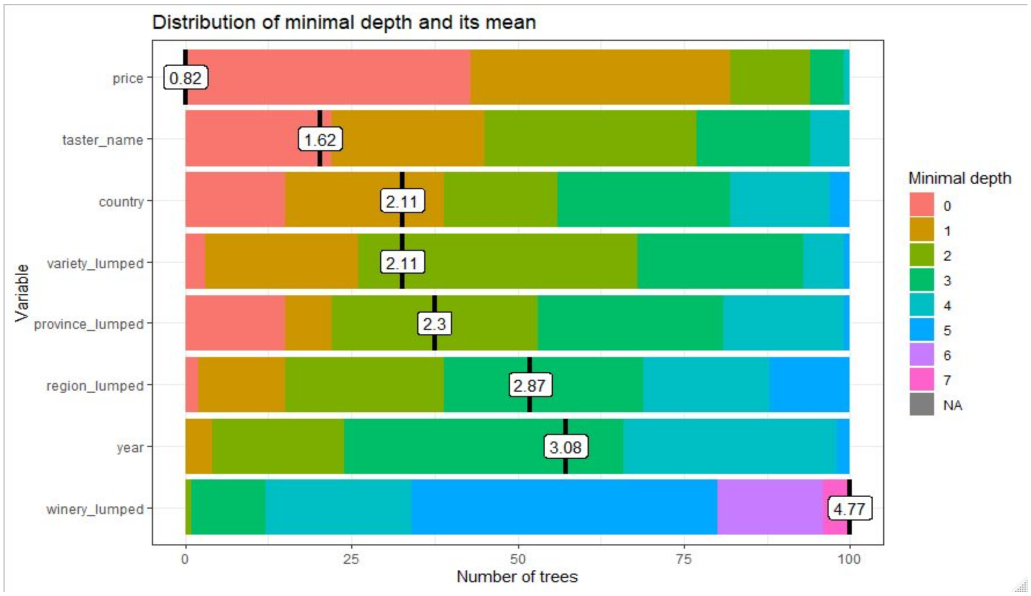
Lasso / Ridge / Enet Model:

With these models, we wanted to compare which would do better in predicting the score with taking other variables into consideration as shown within the wine dataset. Initially because we already knew that lasso was a variable selector, we hypothesized that lasso would be able to eliminate the unwanted variables within the data and come up with better R2, MAE, and RMSE values in comparison to the Ridge and Enet models. The lasso model was able to reduce the unimportant variables and ended up with around 39 variables of importance. Within the Ridge model and Enet model, although lasso ended up with the better values of R2, MAE, and RMSE, the Ridge and Enet model's numbers were very similar to lasso. From this we can predict that although the variable selection did help to make the R2, MAE, and RMSE more reliable, the results were marginal and didn't affect the data as much as we hoped it would.



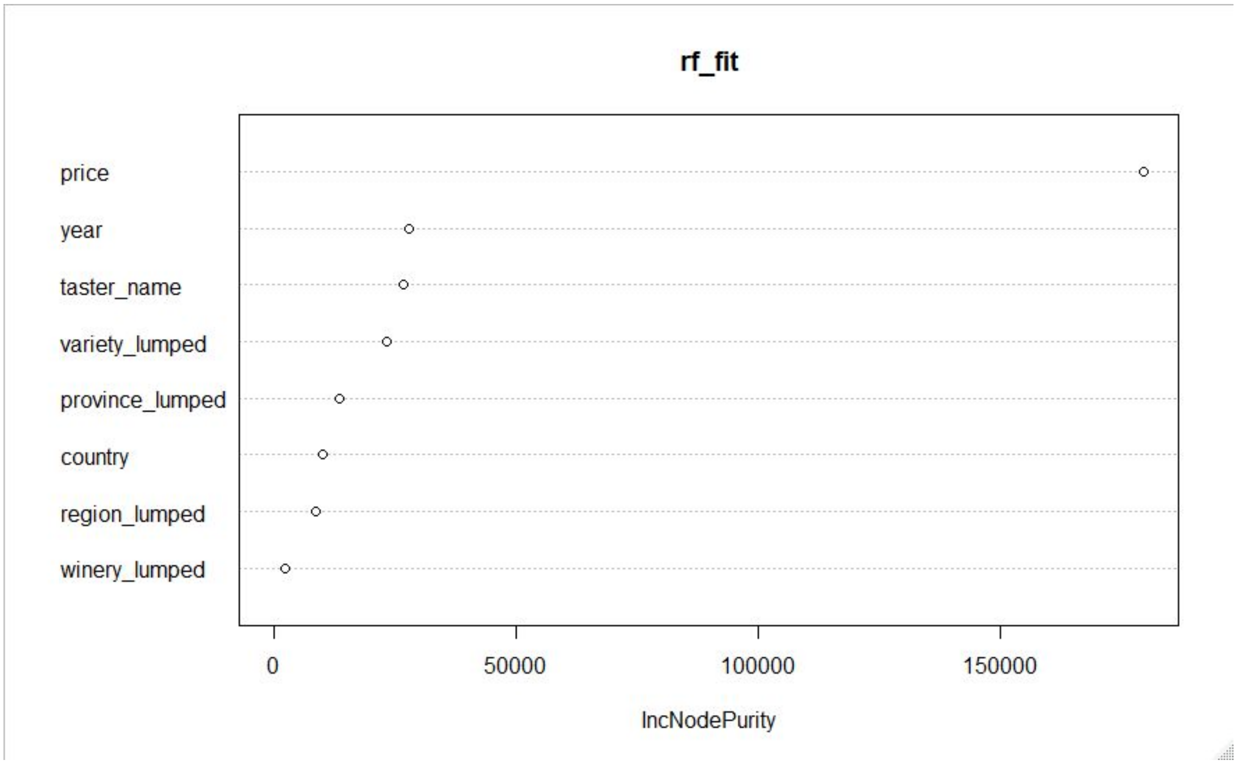
Random Forest Model:

For our Random Forest model, we set the amount of trees to 100 and mtry to 3. We set this to a low amount due to the high amount of variables and factors within the dataset which required a lot more computational power than we had. After looking at our random forest plot,



we can see that there should be about 100 trees set for the model. If we had better computational power, we would want to see how the tree would fare with more trees.

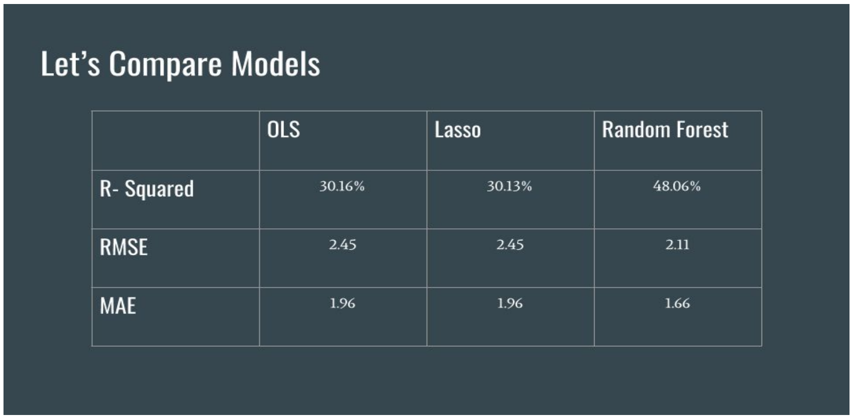
Our first plot is the plot_min_depth_distribution plot. We can see that price is our most important factor with .82 as the mean for that factor. Since price has been shown to be very influential in our previous plots, this only confirms what we had already known. Taster_name is our second important factor with 1.62 as the mean. The next five factors (country, variety_lumped, province_lumped, region_lumped, year) seem to stay very close together with means ranging from 2.11 to 3.08. Finally, we see that winery_lumped is the least influential of all of our factors with a mean of 4.77. That being said, we still can say that it still has somewhat an effect of the points of the wine.



Our next plot is the variable importance plot, where we see the same outcome as our plot_min_depth_distribution plot. The only key difference is that we see year as the second important variable which bumps the categories down 1 in the order of importance.

Now, we will look at the performance of the random forest model. Overall, we have a strong performance over the training set. We have an R squared of 48.06%, a RMSE of 2.11, and MAE of 1.66. When performing on the testing set, we have an R squared of 48.18, a RMSE of 2.15, and a MAE of 1.69. We can confidently say that this model is performing well over the other models we used.

Let’s Compare Models:



Let’s Compare Models

	OLS	Lasso	Random Forest
R- Squared	30.16%	30.13%	48.06%
RMSE	2.45	2.45	2.11
MAE	1.96	1.96	1.66

From the chart, is evident, Random forests performed the best by far in R2, and fared better in RMSE and MAE. Meaning it was the most effective in capturing the noise of the overall data, and contained the lowest amount of average error of magnitude than the rest of the models. In all Aspects, it seemed this was the most robust model.

Conclusion:

Through our analysis of several models, we believe we have found the most important variables that predict points. By far, the heaviest weighted variable was *price*. Through our findings we expected a .03 point increase per dollar on a wine bottle. Meaning a 200 dollar increase would justify a 6 point increase. The next most important variable on price was taster name. This means depending on who tastes your wine has a profound impact your individual wine bottle. However, from a business perspective, this variable renders useless as companies can’t simply have a specific reviewer tasting their wines, at least ethnically. Our last important variable was wine produced in the Province of California. It is very possible that the reputation of Napa Valley wines are holding credibility and tells us information that New World wines (outside of EU) perform well. These variables were pulled from our backward stepwise and are quite consistent with the previously mentioned models like Random Forests. We also do realize there is a high interdependency with points and price. Every feature of a wine, whether it be a luxurious variety of grape or a well regarded province, will be expensive and lead to a higher price. Since every luxurious factor comes back to price, it may be interesting to conduct this project by removing price totally. This can help isolate the rest of the variables as we can later compare to price to gain a better insight on a “best-value” wine.

Our business recommendation for a wine company would be export high quality ingredients for point maximization. As price is the highest weighted variable, and previously mentioned, every luxurious component of a wine like variety and region will cost more to produce / import. WineEnthusiast doesn’t just look for the best tasting bottle of wine when they review wines. They look at the value of the wine at the price of the bottle. We would advise wine companies to try to get the best value out of their wines with solid pricing rather than trying to create the best tasting wine yet.

Works Cited

Score, The Global Wine. “How Do Critic Scores Affect Wine Prices? A Study of Napa Valley Wines.” *Medium*, The Global Wine Score (GWS), 28 Sept. 2018, medium.com/the-global-wine-score/how-do-critic-scores-affect-wine-prices-a-study-of-napa-valley-wines-226d8d08adb7.