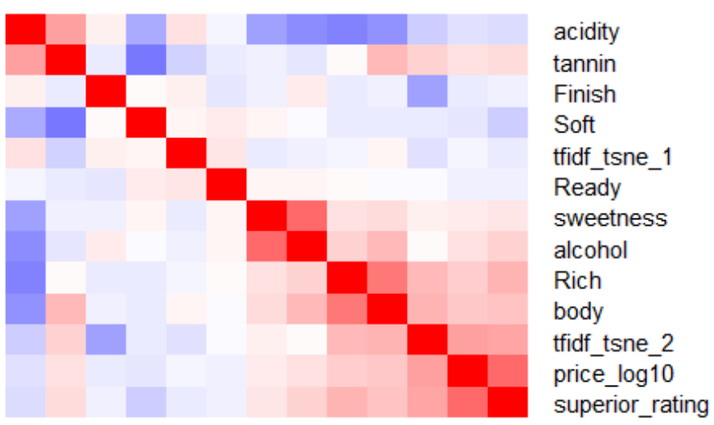
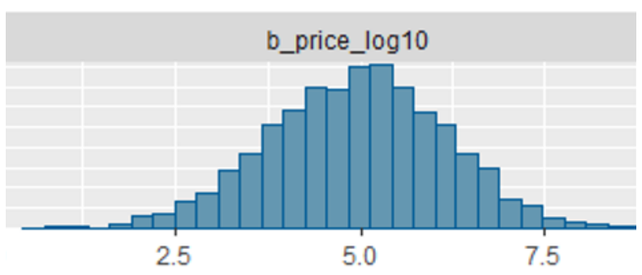
Let them drink Wine

GOOD NEWS! DECEPTIVE WINE PRICES

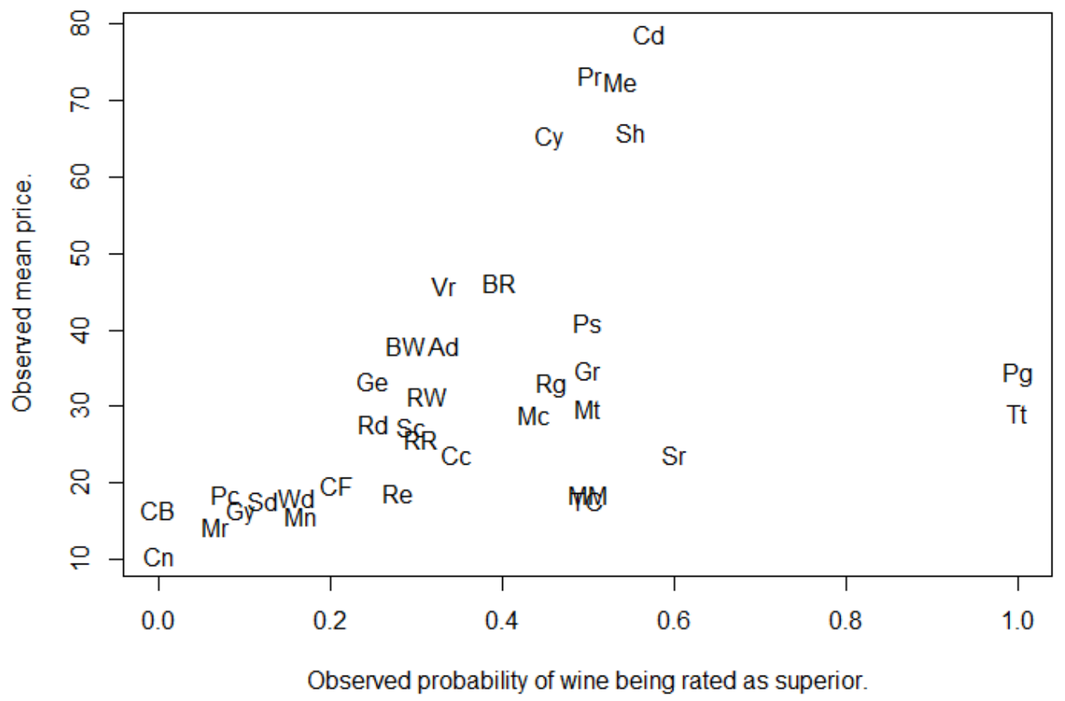
Analysis of wine reviews that uses Bayesian hierarchical logistic regression to model variety wise effects in addition to population level effects of wine characteristics on likelihood of it being rated as superior.





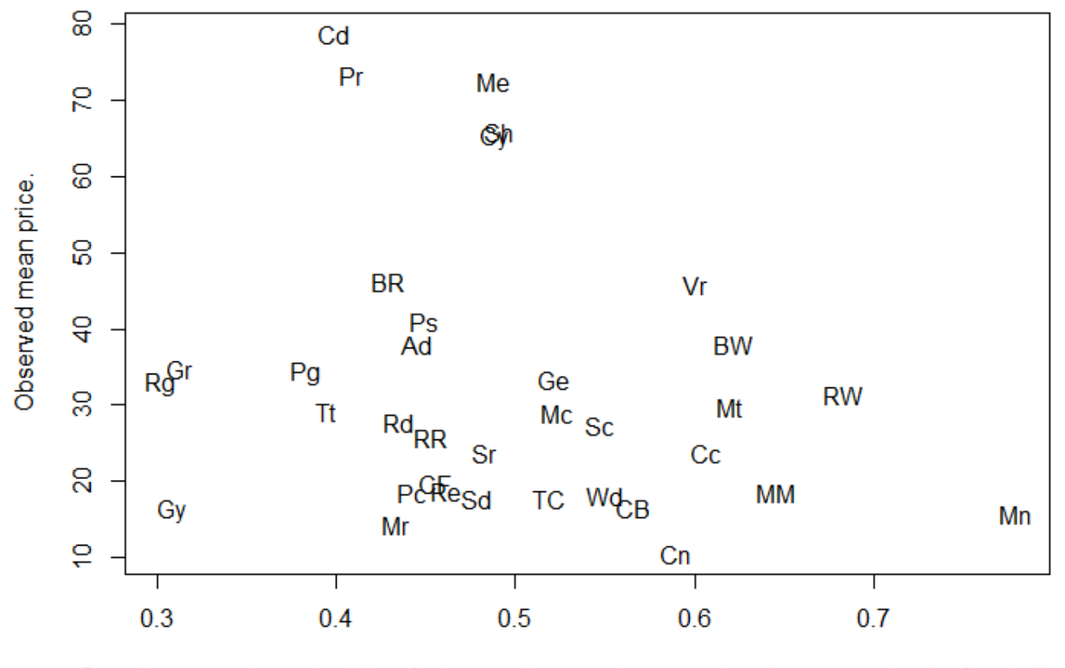
Prices were most positively correlated with wine ratings and price had greatest effect on logs odds of wine being rated as superior. effect on logs odds of wine being rated as superior.

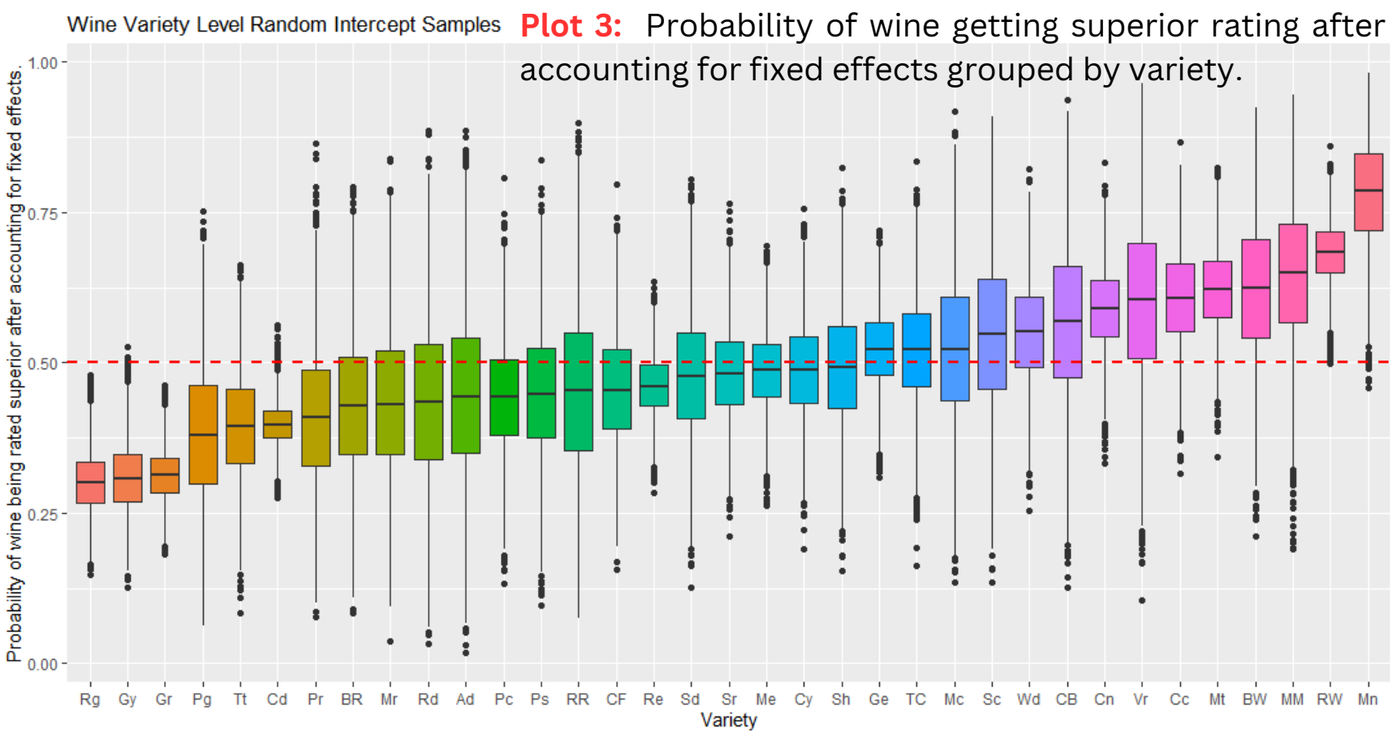
Plot 1: Unadjusted relationship between price (y axis) and the observed probability of superior rating for each wine variety (x axis)



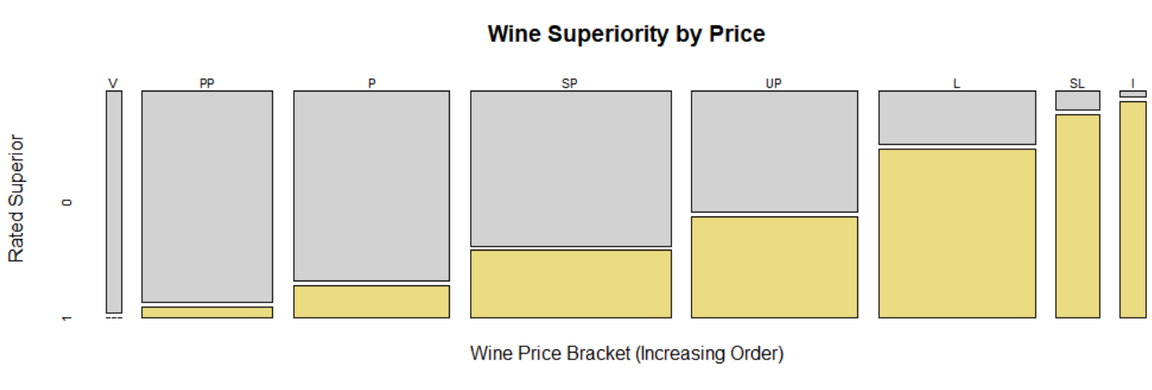
From plot 1 to plot 2, the probability of high-priced wines like Cd (Champagne Blend) dropped while that of cheaper ones like Mn (Melon) increased. Even varieties like Cn (Cabernet Sauvignon) and CB (Chenin Blanc-Chardonnay) with very low odds of being ranked as superior in plot1 were assigned higher probabilities (55-60%) by the model in plot2 likely because they have characteristics beyond price that are similar to that of superior wines. Thus, price is an important factor influencing observed probabilities. It's possible that some varieties like Cd may be overpriced. Long story short, you don't have to break the bank to have great wine. There are plenty of affordable options. Here, the model recommends Melon (Mn), Malbec-Merlot (MM) and Rhône-style White Blend (RW) wine varieties as great affordable options that may even be better at times, than high-priced alternatives.

Plot 2: Posterior probability of wines getting rated as superior per variety after accounting for fixed effects v/s observed prices.

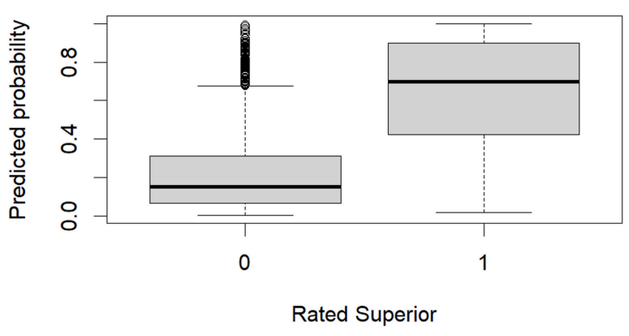




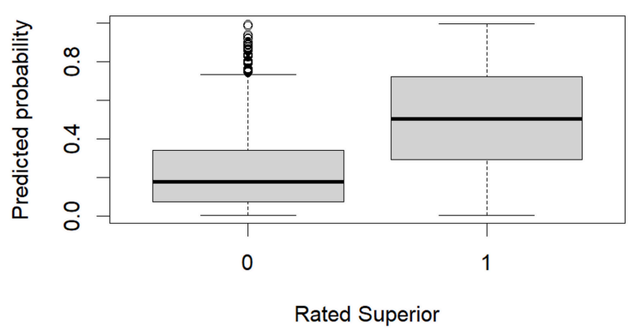
Plot 3: Probability of wine getting superior rating after accounting for fixed effects grouped by variety.



TRAIN SET: Accuracy = 0.8128, F1 Score = 0.7437



TEST SET: Accuracy = 0.7325, F1 Score = 0.5664



BACKGROUND & METHOD

- *Dataset of 2500 wine reviews, assigned ratings, prices, and presence indicators for flavor/appearance/mouthfeel characteristics.*
- *NLP methods employed to get scores for key wine judging characteristics (acidity, sweetness, tannin, alcohol, body). Topic modelling done to explore. Feature engineering added new features like price bracket, alcohol, tanning, body, acidity, sweetness, tfidf_tsne_1. Price log scaled to reduce skewness. Min max normalization before model fitting.*
- *Likelihood of superiority of wine ratings modeled using the Bayesian Hierarchical Logistic Regression approach. Grouping variable = variety. Predictors = price_log10 + tannin + alcohol + Rich + price_log10 * body + price_log10 * tfidf_tsne_1_norm. Response variable = superior_rating. Data distribution: Binomial (success = wine superior), Priors: Student t distribution for random effects term and Normal distribution for population parameters.*
- *Model was tested on new test data found online. Some overfitting was observed.*

CONCLUSION

- *Of all the fixed effects in the model, has the highest average effect (comparatively highest coefficient of 4.96).*
- *Derived features had larger effect compared to the single indicators like "Rich" reinforcing value of having derived key wine judgement criteria incorporating more than one indicator from text.*
- *The baseline tendency is for wines to not get rated as being "superior" (population fixed effect intercept estimate = -10.55).*
- *Barring interaction terms, fair evidence against zero effect for all other predictors (posteriors away from 0).*
- *Some wine varieties get ranked high more often than others (plot 3) but large variance + outliers suggest high subjectivity.*
- *Removing effect of price revealed that some cheaper wines may be just as good as more expensive ones.*