ORIE 3120 Final Project: Wines

Data Visualization and Analysis

For this project, we analyzed a dataset of wine reviews scraped from the Wine Enthusiast Magazine. It includes features such as reviewer, country, and province of origin, price, variety, vintage, winery, and ratings, on a scale from 0-100, of different wines.

The overall goal of our project is to understand what features contribute to high ratings for wines so that we can provide consumers with recommendations about what to look for when purchasing a bottle of wine. We will explore the following questions in this report to answer our overarching question.

1. How does price affect wine quality and how well can we predict wine rating from the price?
2. How do other qualitative variables, such as winery, category, country, and variety affect the rating?
3. How can temperature and precipitation help us predict rating?
4. How does reviewers' personal taste affect ratings?

Question 1: How does price affect the quality of wine and can we predict the rating from price?

Data Analysis Methods: Testing Assumptions for Linear Regression and Linear Regression

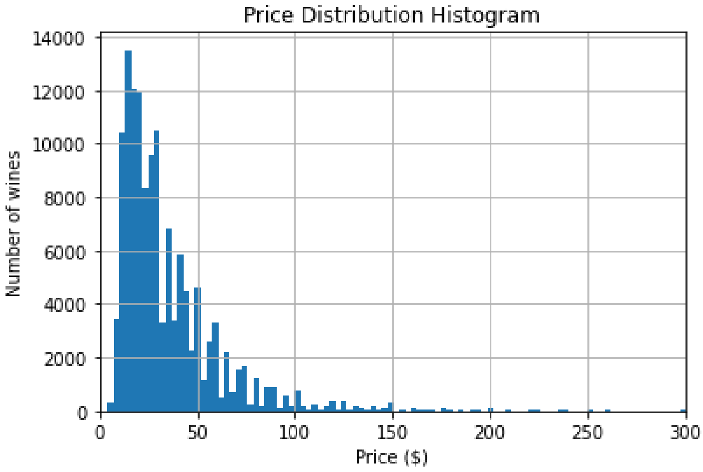


Figure 1. Distribution of Wine Price

The price of the wine can greatly vary - you can find boxed wine for $5 as well as bottles that cost thousands of dollars. However, is the price really a big factor in how good a wine is? Is it worth investing in a fancier bottle or can we still get a high-quality wine without breaking the bank? In this section, we will study the relationship between the price of the wine and the average rating per price. First, let's observe the distribution of wine prices. From The histogram above we can see that most of the wines from the dataset are in the range of $10 to $50, which is not surprising. Next, to see if there is any correlation between wine price and the ratings, we found the median rating for each unique price. That is, if for one specific price, there are multiple

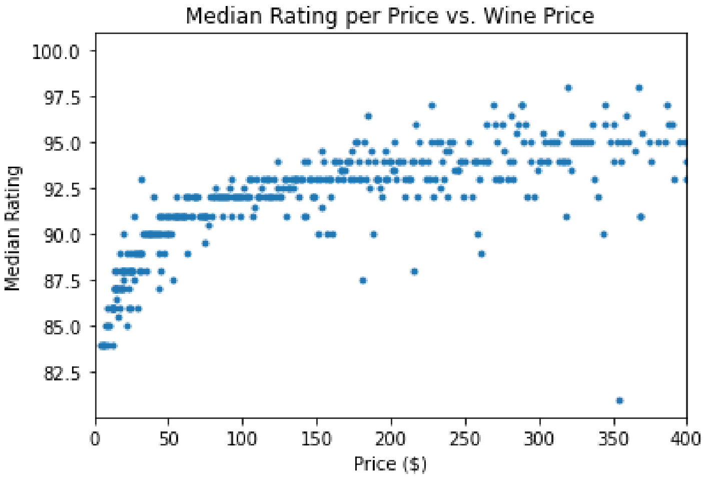


Figure 2. Median Rating per Price vs. Unique wine prices

wines in the dataset, we take the median of all the ratings of these wines. This way, we remove outliers and any noise caused by the personal preferences of the reviewers.

From the graph, we can see that there is a positive correlation between the price and the ratings of the wines. We noticed that the rating rapidly increases until the price reaches around $50, after which its growth starts to decrease. Based on this observation, our first conclusion is that the relationship between the price and the ratings is not linear. The second conclusion that we made is that a reasonable price at which you can still get a good quality wine is around $50. Since the relationship between the price and the ratings is not linear, we tested the linear regression assumptions on various transformations of price. In the end, we found that the log of the price met all the linear regression assumptions, as seen from the tests below.

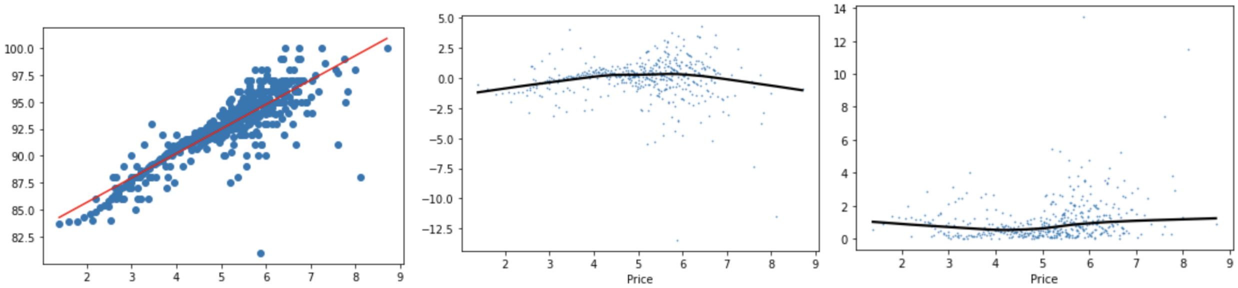


Figure 3. Testing linear regression assumptions (LOWESS lines)

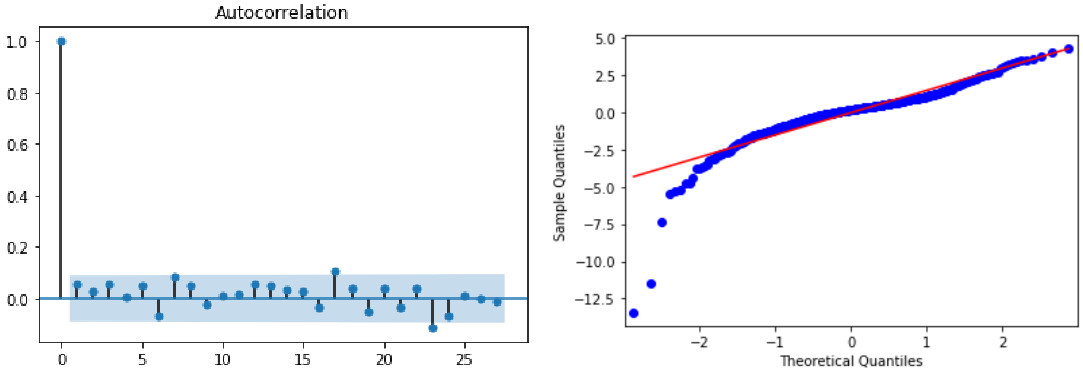


Figure 4. Autocorrelation and QQ Plots

In the center graph of Figure 3, we see that the LOWESS line is approximately flat and at the 0 mark, so the assumption that the errors have a constant mean of 0 is met. Moreover, in the rightmost graph, the LOWESS line is relatively flat and the spread in the vertical direction in the residual plot is approximately the same regardless of the x value, so the assumption that the errors have constant variance is met. In the autocorrelation plot of Figure 4, we don't observe a significantly higher autocorrelation at any lag except lag 0, so the assumption that the errors are independent of each other is met. Finally, in the QQ-plot on the right side of Figure 4, we see that the plot is relatively aligned with the line of the theoretical quantile even though it is heavy-tailed, so the assumption that the errors are normally distributed is met.

As a result, we fitted a linear regression model using the logarithm of the price with a 70-30 train and test split. This model had an adjusted R-squared value of 0.707 and a sum of squared errors of 883.32. We then found the predictions that this model makes and plotted them along with the observations to see if they fit well into the data points.

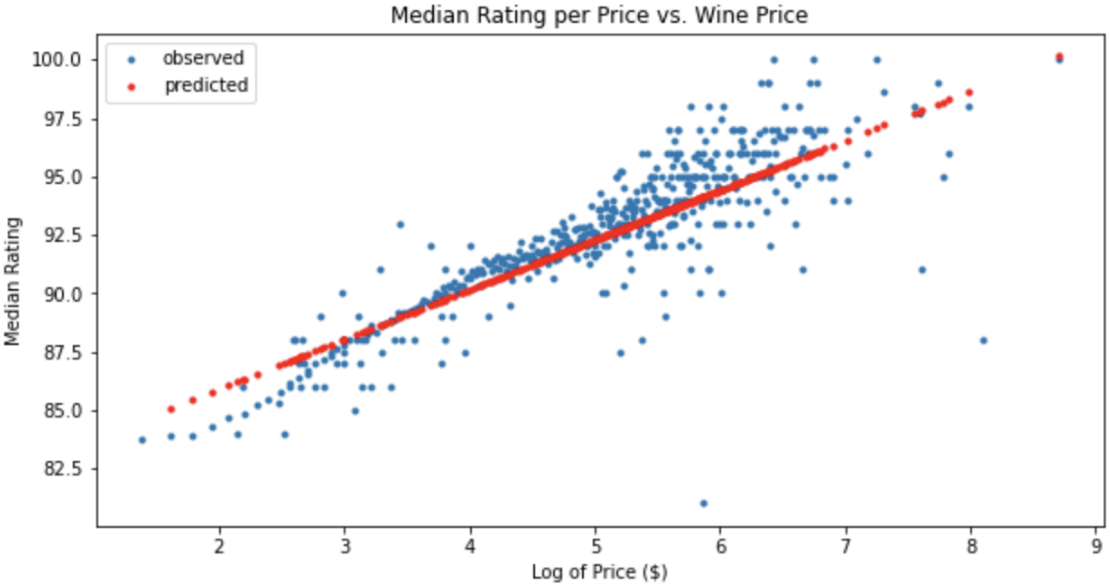


Figure 5. Predicted and observed values of median rating per price

From Figure 5, we can see that the predicted values fit well with the observed data points. Overall, we can conclude the log of the price has a linear relationship with the rating and also serves as a strong indicator of rating. One possible reason is that when reviewers look at a particularly expensive wine, they may develop an expectation that it will be of high quality. This confirmation bias phenomenon could contribute to the positive correlation between rating and price. We recommend that consumers pay attention to price when trying to select good quality wine, however, spending an extraordinary amount is not necessary. Generally, wines priced over $100 have a similar rating, so spending more than $100 may not be worth it.

Question 2: How do other qualitative variables, such as winery, category, country, and variety affect the rating?

Data Analysis Methods Used: Linear Regression

In addition to investing in the effect of price on rating, we wanted to explore the impact of qualitative features in the dataset. As seen in Figure 6, below, the origin country of wine appears to be correlated with its rating. For example, after making sure each country had at least 30 wines to form a representative sample, we see wines from England and Austria have a higher rating on average. In contrast, wines from Romania are rated much lower.

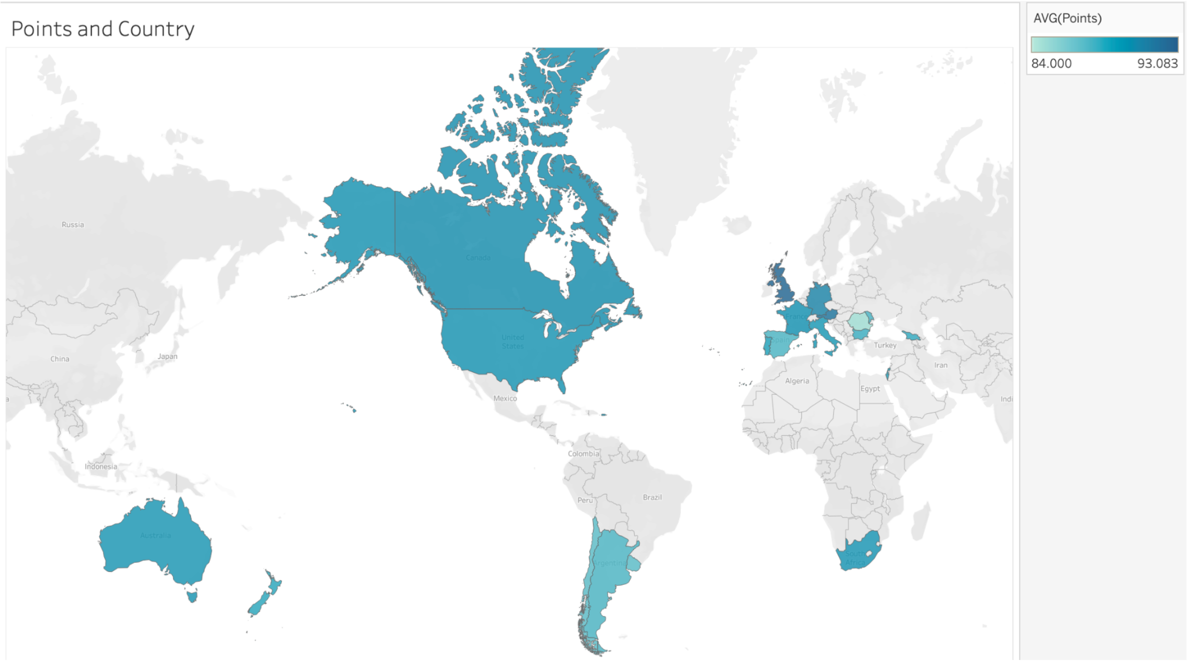
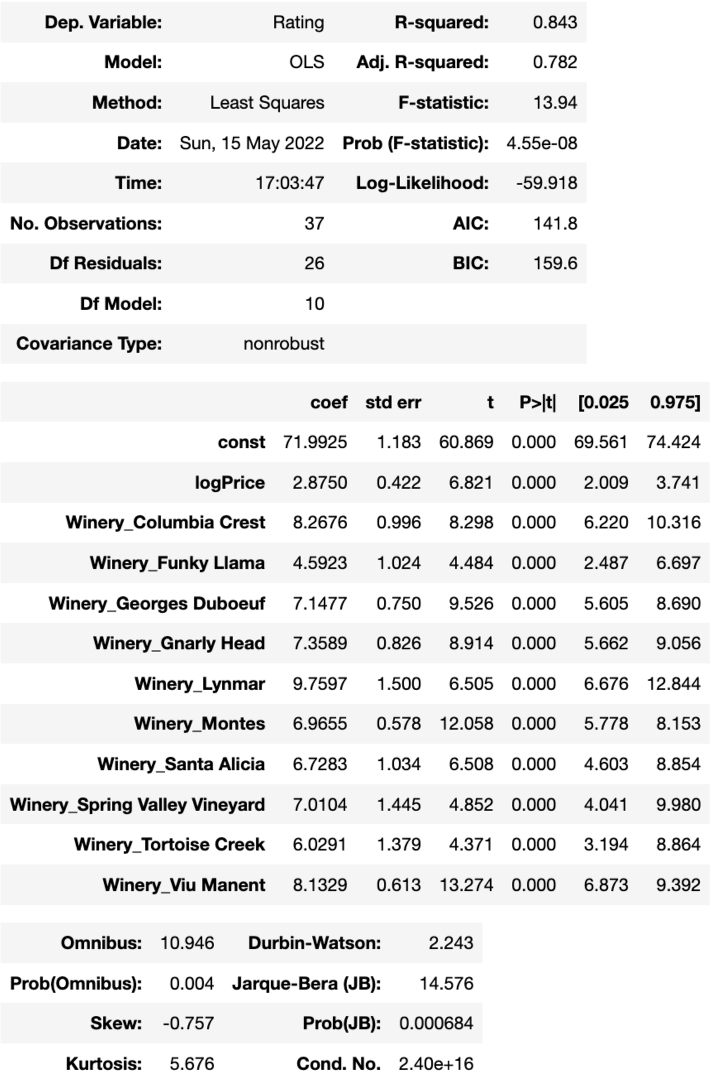


Figure 6. Average Rating Across Countries

To investigate the effect of other qualitative variables in our dataset on rating, we built several linear regression models using the log of the price + country, log of the price + category (red, rose, dessert, sparkling, or white wine), log of the price + winery, and log of the price + variety. Before building the model, we made sure that each combination of the log of price and either country, province, winery, or variety had at least 30 ratings, and we then took the average rating for each of these records. After, to incorporate the categorical variables into our model we encoded them as dummy variables such that each categorical variable became its own column with either a 0 or 1 for its value. In total, we had 4 models that we compared to see which would give us the best results.



Dep. Variable:	Rating	R-squared:	0.843
Model:	OLS	Adj. R-squared:	0.782
Method:	Least Squares	F-statistic:	13.94
Date:	Sun, 15 May 2022	Prob (F-statistic):	4.55e-08
Time:	17:03:47	Log-Likelihood:	-59.918
No. Observations:	37	AIC:	141.8
Df Residuals:	26	BIC:	159.6
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	71.9925	1.183	60.869	0.000	69.561	74.424
logPrice	2.8750	0.422	6.821	0.000	2.009	3.741
Winery_Columbia Crest	8.2676	0.996	8.298	0.000	6.220	10.316
Winery_Funky Llama	4.5923	1.024	4.484	0.000	2.487	6.697
Winery_Georges Duboeuf	7.1477	0.750	9.526	0.000	5.605	8.690
Winery_Gnarly Head	7.3589	0.826	8.914	0.000	5.662	9.056
Winery_Lynmar	9.7597	1.500	6.505	0.000	6.676	12.844
Winery_Montes	6.9655	0.578	12.058	0.000	5.778	8.153
Winery_Santa Alicia	6.7283	1.034	6.508	0.000	4.603	8.854
Winery_Spring Valley Vineyard	7.0104	1.445	4.852	0.000	4.041	9.980
Winery_Tortoise Creek	6.0291	1.379	4.371	0.000	3.194	8.864
Winery_Viu Manent	8.1329	0.613	13.274	0.000	6.873	9.392

Omnibus:	10.946	Durbin-Watson:	2.243
Prob(Omnibus):	0.004	Jarque-Bera (JB):	14.576
Skew:	-0.757	Prob(JB):	0.000684
Kurtosis:	5.676	Cond. No.	2.40e+16

Figure 7. OLS regression results (log of price + winery)

The first model we looked at is the log of price + country. Adding the countries improves the adjusted R-squared value to 0.734, and all the countries had significant coefficients. However, when we evaluated the model on the test data we found that the sum of squared errors increased to 2340.022. Similarly, when we fitted the log of price + variety and log of price + category models, we found the SSE increased to 5608.40 and 1825.54, respectively. It is likely that adding these features causes overfitting to the training data and therefore decreases the performance of the model on the test set.

Unlike with country, variety, and category, when we fit a model using the log of price and winery, the model performance significantly improved compared to using just the log of price. The adjusted R-squared value increased from 0.707 to 0.782, while the MSE decreased from 883.32 to 228.67. Overall, this suggests that winery is a very important feature along with price when predicting rating, and this makes sense considering a winery's climate and practices in growing and harvesting grapes will shape the outcome of the wine.

The wineries in Figure 7, above, all have significant coefficients, however, the Lynmar, Columbia Crest, and Viu Manent have the most positive coefficients suggesting that wines from these locations tend to be rated very well. When we corroborate these results with reviews on Google, we find that the previously mentioned wineries are in fact highly rated. Moreover, Funky Llama wines generally had the lowest reviews among all the listed wineries.

From this analysis, we can conclude that the winery is also a very important feature to consider when trying to buy high-quality wines. In particular, recommend trying wines from either Lynmar, Columbia Crest, or Viu Manent.

Question 3: How can temperature and precipitation help us predict rating?

Data Analysis Methods Used: Forecasting and Linear Regression

It is common knowledge that temperature and precipitation affect the growth of grapes, which in turn determines the quality of the wine. Given that we knew the vintage of each wine we decided to explore how the average rating of wines changed each year and if this rating was dependent on temperatures and precipitation from that year. We focused on one particular region, Napa Valley in California, for which we were able to obtain weather data from 2008 to 2018 from the University of California Weather Database. Then, we first built a simple exponential smoothing model, which had an SSE of 2.052. However, as seen from Figure 8 it does follow the pattern of the data very well.

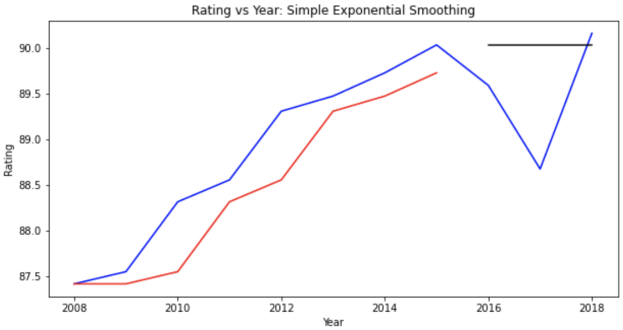


Figure 8. Forecasting with simple exponential smoothing

Our next approach was to fit a Holt-Winters model, for which we had an additive trend to account for the increasing slope and additive seasonality to account for the constant width and height of the seasonal periods. We used a seasonality value of 2, which we determined from the data. In Figure 8, the average rating from 2008 to 2009 doesn't increase very much but increases much more from 2009 to 2010. This two-year pattern repeats up to 2015. After fitting the Holt-Winters model, we see that although this approach fits the data before 2015 well, it does not accurately represent the pattern of the data afterward, and gives us an SSE of 11127673.05. It is possible that some weather anomaly occurred in 2017, which affected the ratings of the wines, and incorporating this information could help us better forecast wine quality. According to a paper by Edward Oczkowski, previous literature has commonly used the following model to predict rating:

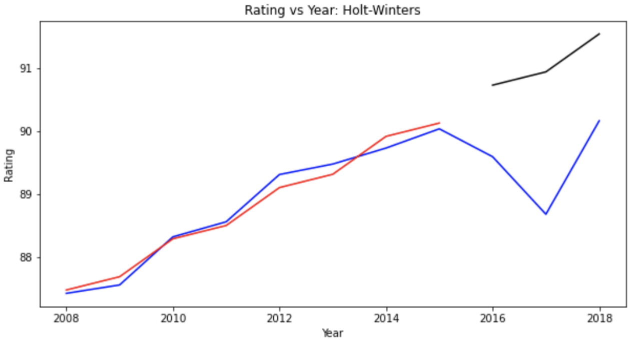


Figure 9. Forecasting with Holt-Winters

$$(Rating)_i = \alpha_0 + \alpha_1(Rain)_i + \alpha_2(Diff)_i + \alpha_3(Temp)_i + \alpha_4(Temp)_i^2 + \varepsilon_i$$

In this model, *Rating* is a score from 0 to 100, *Rain* is the average of monthly rain in mL during harvest months (January to March), *Diff* is the difference of between maximum and minimum temperatures in Celsius over the growing season (October to March), and *Temp* is (maximum+minimum)/2 over the growing season (Oczkowski, 2016). Using Napa Valley's weather data we fit a linear regression model using the features above.

While the resulting model does not closely follow the graph of ratings, it does have a similar trend, and our resulting SSE is 1.379, which suggests that the weather does have some effect on the rating, but not as much as the price or winery. As previously discussed, this is likely due to the fact that the reviewer is aware of the price and winery of the wine which can make them biased toward a higher or lower rating. The reviewer would not be aware of the conditions the wine is grown in, so unless the weather is very bad one year and it can be tasted in the wine, it may not affect the rating very much. According to the Napa Valley Vintage Reports, for example, 2017 was a particularly difficult year due to the high precipitation levels, which explains the dip in the graph (Vintners, 2022). Overall, we conclude that it is difficult to forecast the rating simply from the weather since the weather can be unpredictable and vintners have methods of adjusting to the unexpected conditions.

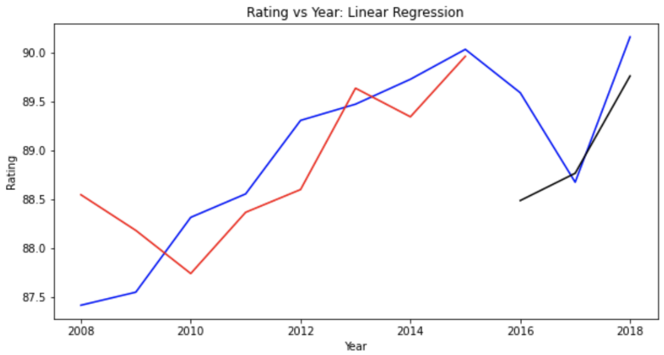


Figure 10. Forecasting with Linear Regression

Question 4: How does reviewers' personal taste affect ratings?

Data Analysis Methods Used: Linear Regression

Finally, after taking into account the important effect that price and winery have on a wine's quality, we wanted to investigate how much the individual reviewers' personal tastes and score of

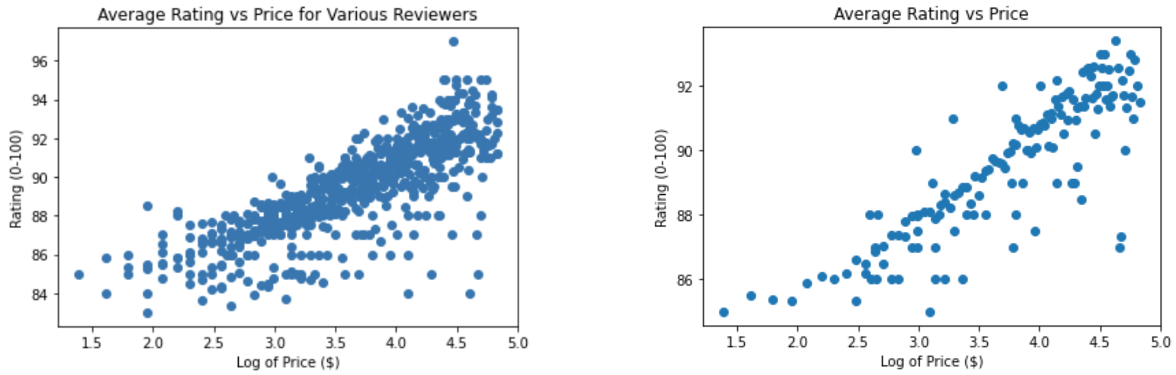


Figure 11. Average rating of wines vs price grouped by price and reviewer(left); grouped by price (right)

the wines impact their perceived quality. In order to do this, we built seven linear regression models.

We first wanted to visualize the trend of the data when we plot the average score vs price grouped by Price and Reviewer (Figure 11, on the right). Each data point represents the average scores by every reviewer at each price(s) of wine he or she tasted. We then wanted to compare that to a visualization of the trend of the data when we plot the average score vs price grouped by just price (Figure 11, on the left). Based on the visualizations, the trends are similar which is expected because of the trends we found in question 1. Both graphs show a proportional relationship between score and price. The graph on the left which is grouped by both price and reviewer shows a wider range of average scores as you can see by the y-axis. We want to explore more about this difference and find out if some reviewers are harsher and tend to give lower scores to wines of the same price, or if some reviewers are more generous and tend to give out higher scores to wines of the same price.

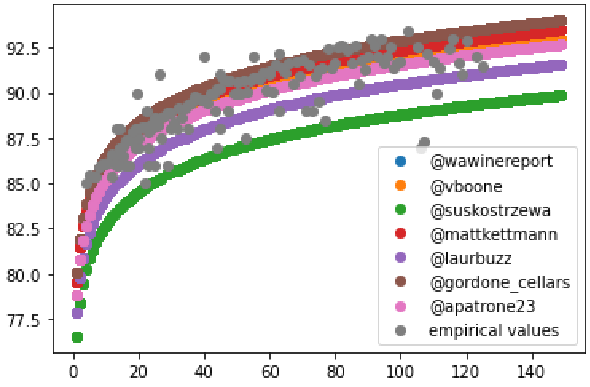


Figure 12. Comparing reviewer ratings across various prices

In order to investigate the different reviewers' impact on the ratings, we built linear regression models for each reviewer using the log of the average price and each reviewer's Twitter handle as a covariate. For the models, we used the same table as figure 8 above, so it is grouped by price and reviewer. Then, similar to our model in question 2, in order to incorporate the categorical variables, the Twitter handles, into our model we encoded them as dummy variables.

The models ended up with an adjusted R-squared value in the range of 0.633 and 0.712, and the constant and log of price coefficient values were very similar across all seven models. The main discrepancy lies in the coefficients for reviewers, which were all significantly different with p-values less than 0.05. @gordone_cellars was the one with the most positive coefficient, which was 0.8813, suggesting that he is the most lenient reviewer out of all of them. On the other hand, @suskostrzewa was by far the most critical with a coefficient of -3.368. These differences can be seen in Figure 12, above, where we randomly generated prices from 0 to 150 dollars and then predicted the rating for each reviewer based on the linear regression models. Most of the reviewers are quite similar and close to the empirical values, however, users such as @laurbuzz and @suskostrzewa appear to have much lower ratings for different prices.

To conclude, reviewers' personal taste does in fact have some effect on the overall rating of the wines. Since ideally, we would want similar ratings for the same wine for an objective measure of quality, in future studies, it might be good to control for these discrepancies when running analyses.

References

Oczkowski, E. (2016). The effect of weather on wine quality and prices: An Australian spatial analysis. *Journal of Wine Economics*, *11*(1), 48–65. https://doi.org/10.1017/jwe.2015.14

Schuring, R. (2021). *Wine Reviews*. Kaggle. Retrieved May 16, 2022, from https://www.kaggle.com/datasets/roaldschuring/wine-reviews

University of California. (2021, December). *California Weather Data: Set Dates and Variables--UC IPM*. How to manage pests. Retrieved May 15, 2022, from http://ipm.ucanr.edu/calludt.cgi/WXSTATIONDATA?STN=NAPA.C

Vintners, N. V. (2022). *Napa Valley Vintage Reports*. Napa Valley Vintners. Retrieved May 16, 2022, from https://napavintners.com/napa_valley/vintage_charts.asp