# Evaluating Price Determinants of California Wine

Andrew Abrahamian, Victoria Hollingshead, Heesuk Jang, Hsi-sheng Wei

2022-08-02

## Introduction

According to a recent industry report from the Silicon Valley Bank, the United States premium wine industry is in the midst of an existential crisis. Although exceedingly popular in the 1990's due to their wide adoption by the boomer generation, for the last 20 years, premium wineries have experienced a steady decline in sales growth. As more consumer research surfaces, industry leaders are faced with the uncomfortable realization that the exceptional reputation and world-renown health benefits of their product are no longer appealing to the needs, style, or values of the new and emerging consumer base: millennials.[1]

In addition to the lack of adoption by millennials, the wine industry faces stiff competition from other alcohol categories, where the marginal utility far exceeds their grape-based competitor (i.e., spirits offer more ethanol per mL). With 70% of American consumers indicating that they would be switching to cheaper food and staple alternatives due to inflation fatigue[2], there is a growing impetus for the premium wine industry to solidify its position in the American household, not only to survive the economic downturn, but to emerge in the next generation.

In this study, we hope to provide valuable insight into which factors contribute the most to US wine product prices. With this information, we hope to enable the premium wine industry leaders to develop and implement data-driven business strategies and to strengthen an otherwise languid influence over the millennial consumer base. The results of this study would contribute to research efforts from a wider network of private and academic partnerships.

## Data & Methodology

The data in this study comes from the Wine Enthusiast magazine. We sourced the data from TidyTuesday's Github repository, but it was originally scraped from the Wine Enthusiast website on November 24, 2017. We performed exploratory analysis on a 30% set of our sampling frame. We build the model and generated statistics on the other 70%.

The analysis used a sample of 14514 cases, which can be evaluated against the large sample assumptions of IID data and the existence of a unique BLP. First, regarding IID data, the experts from Wine Enthusiast comprehensively taste wines across the globe every year, and our final sample contains all Californian wines reviewed in the database with each of the major regions sampled. Although the possibility of geographical clustering may not be ruled out, there is in fact much variety within the state, and observing one wine review hardly provides information about some other wine. At the time of analysis, each row in the dataset represented a unique product review of a specific bottle of wine.[3] Second, for a unique BLP to exist, there should be no perfect collinearity among the predictors. In other words, no variable is a linear combination of other variables. When analyzing nearly perfect collinearity, the strongest variance inflation factor was 1.12 on the **region** covariate, indicating that the independent covariates's contributions to the others variance in our regression are minimal.

[1] McMillan, Rob. "State of the US Wine Industry 2022" Silicon Valley Bank (2022).

[2] Terlep, Sharon. "Americans Are Showing Inflation Fatigue, and Some Companies See a Breaking Point" The Wall Street Journal (2022).

[3] Further research showed that Wine Enthusiast employs two reviewers for the California region, which may contradict our assumption of anonymous random sampled reviews.

There were five key concepts in this study: **price**, **points**, **region**, **color**, and **age**. The focus of this study was to understand how the latter four explain wine price. **Price** is defined as the cost in US dollars for a bottle of the particular wine. **Points**, the rating Wine Enthusiast reviewers assigned each tested wine between 80-100, is our primary predictor and was collected using blind tests from a panel of wine experts. Because we operationalized points as a metric value representing "wine quality," we hypothesized that there will be a positive correlation between points and price. Because California is the leading wine production area in the US, this report focused on Californian wines. We re-classified all the American Viticultural Areas (AVA) of California into six main **regions** as seen in **Figure 1**. We hypothesized that price would be positively correlated with the *Napa-Sonoma* region for its world-renown brand reputation, particularly Napa Valley. The grape varieties were combined into three **color** categories: red, rosé, and white. After reviewing the literature, we hypothesized that red wines would have a higher price, as they are generally more expensive to produce due to more expensive raw materials. **Wine ages** were calculated by subtracting the vintage year, extracted from wine titles, by 2017, the year that the dataset was created. Using general heuristics, we hypothesized that age would have a positive correlation with price, as older vintages tend to be more rare and more expensive.

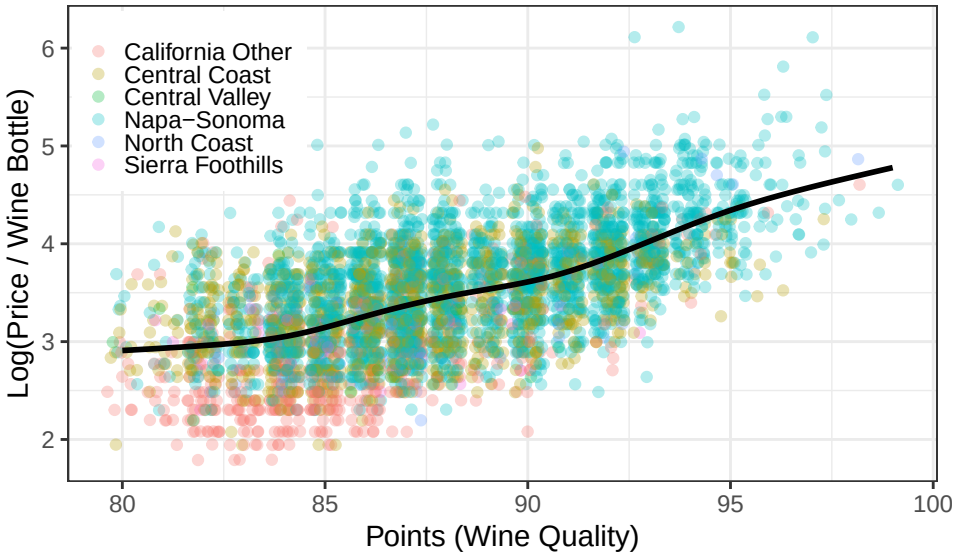


Figure 1: Wine Price as a Function of Quality and Region of Origin

In addition, to build a robust OLS regression model with interpretable results, we made four key modeling decisions: determining the sampling frame, transforming **points** as the outcome variable, re-categorizing samples in our **region** variable, and withdrawing the **age** covariate.

To meet our IID criteria, we restricted our sample to only unnamed reviewers removing 80% of reviews from the dataset, leaving 26244 observations in our sample. We made this decision to minimize any dependence between our sample in our sample distribution. At the time of analysis, we assumed that all unknown reviewers were unique and had an equal opportunity to be selected from this sample. We then restricted our sample to wines produced in California with a correct vintage year format in their description and subsequently removed duplicated reviews, resulting in a dataset with 14514 samples.

To improve linearity and normality, we applied the log transformation on price, which was heavily right-skewed with the mean price of $38 from the range of $5 to $625 per bottle.

Though *Napa-Sonoma* is a part of the *North Coast* region, it was a distinct category in our sample and this feature of the data was retained. We also combined *South Coast* with the *California Other* region due to its low sample size (n=10).

We discarded the age covariate from our final model despite its significance. We noted that the coefficient on Age, -0.003, was a different sign than its Pearson correlation coefficient of 0.04. However, the magnitude of robust standard error of the coefficient on Age, 0.002, is close in size to its coefficient.

Also, $R^2$ does not change between model 3 and model 4, indicating that age explains little to no variation in the natural log of price.We believe that the difference in signs reflects random variation around zero and that the above reasons justify withdrawing age from our final model.

## Results

**Table 1** shows the results of four representative regressions. The coefficient on points was statistically significant across all models, with point estimates ranging from 0.08 to 0.1. As we transformed price with a natural log for the regression models, this coefficient represents a percent increase in price given a one-unit increase in points. To give the appropriate context, consider two hypothetical red wines from the *Napa-Sonoma* region. Applying model 3, if one wine is rated 10 points higher than the other, that wine's price will be 82% higher holding all else equal.

Table 1: Estimated Regressions

| | Output Variable: Natural Log of Price | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Points | 0.10*** | 0.09*** | 0.08*** | 0.08*** |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| Region: Central Coast | | 0.44*** | 0.42*** | 0.42*** |
| | | (0.02) | (0.02) | (0.02) |
| Region: Central Valley | | 0.10* | 0.04 | 0.05 |
| | | (0.05) | (0.04) | (0.04) |
| Region: Napa-Sonoma | | 0.61*** | 0.57*** | 0.57*** |
| | | (0.02) | (0.02) | (0.02) |
| Region: North Coast | | 0.30*** | 0.33*** | 0.34*** |
| | | (0.05) | (0.05) | (0.05) |
| Region: Sierra Foothills | | 0.29*** | 0.20*** | 0.21*** |
| | | (0.03) | (0.03) | (0.03) |
| Wine Color: Rose | | | −0.46*** | −0.47*** |
| | | | (0.03) | (0.03) |
| Wine Color: White | | | −0.35*** | −0.35*** |
| | | | (0.01) | (0.01) |
| Age of Wine | | | | −0.003* |
| | | | | (0.002) |
| Constant | −5.56*** | −4.58*** | −4.12*** | −4.10*** |
| | (0.13) | (0.12) | (0.12) | (0.12) |
| Observations | 10,152 | 10,152 | 10,152 | 10,152 |
| $R^2$ | 0.35 | 0.44 | 0.51 | 0.51 |
| Residual Std. Error | 0.49 (df = 10150) | 0.46 (df = 10145) | 0.43 (df = 10143) | 0.43 (df = 10142) |
| *Note:* | *p<0.05; **p<0.01; ***p<0.001 | | | |
| | $HC_0$ robust standard errors in parentheses. | | | |

Across all models, the coefficients on the categorical region variable was statistically significant except for wines from *Central Valley*. However, model 3 has a statistically significant F-statistic (F = 490.6, p < 0.001) relative to model 1, meaning that region should be included in our models to explain the variation in price. Applying model 3, the effect region of origin has on prices is large relative to points, with statistically significant coefficient sizes ranging from 0.2 to 0.57. Going back to our two hypothetical red wines example: now assume that one wine is from *Napa-Sonoma*, the other is from the reference region *California Other*, and both are rated as 90-point wines. The former's price will on average be 77% higher than the latter, holding all else equal.

## Limitations

Our outcome variable is the natural log of wine price in dollars, which is a reliable numeric measure. We used Wine Enthusiast reviewers' ratings, (points) as the primary predictor, and the database only contains wines with scores higher than 80, thus the actual range of that variable was 80-100. Limited range in points prevents us from evaluating wines that would have otherwise been assigned low scores. We will not know the influence of these low-scoring wines on price. Our population is restricted to presumably well-performing wines. At the same time, we use points as a metric variable in order to support our modeling, but the consensus between Wine Enthusiast and other major expert ratings was found to be moderate or low[4], which may affect the reproducibility of our results with other publications.

For our other predictors, literature research suggests that *South Coast* wines have distinctive features. However, our training dataset only featured 10 samples from the *South Coast*. To mitigate any problems with small sample size, we combined this region into the *California Other* region. Age was calculated by subtracting vintage from 2017. It is an assumption that 2017 was the year at the time of tasting, which we were unable to verify, and this issue may contribute to the barely significant result we found between age and price.

In our regression models, wine price is specified as the outcome variable while points, regions, colors, and ages as the predictors. Those predictors are authentic and temporally precede the outcome variable, thus there is no apparent omitted variable bias that affects both sides of the model and reverse causality is unlikely. However, one potential outcome variable on the right-hand side is region, since each region has its unique terroir conditions like climate, soil, and terrain, which determine the type of grape that can grow there and profoundly contribute to the complex flavor of wine and its rating. Specifically, we expect the existence of a bias towards zero, which means that the true coefficient of regions on price could be greater than the current observed value.

## Conclusions

From this study, we can conclude that points, region, and color have a statistical significant influence and explanatory effect on the overall price of wine. In reaction to the diminishing purchasing power of the average American in the short term, and the waning millennial consumer base in the long term, we can use the results of this study to recommend particular products qualities to minimize these risks to the wine industry. We recommend the vineyards and wineries prioritize producing rosé and white wines because, per our model and holding all else equal, they are 45.7% and 34.9% less expensive than red wines respectively. Likewise, wines produced from the *California Other* region are less expensive than all other regions in California. In all, selecting for these characteristics can maximize wine quality while minimizing cost for consumers, encouraging adoption from a fast-growing millennial consumer base.

Further quality ratings research would be required to make this recommendation as robust as possible. As outlined in our limitations, for example, our study did not account for potential marketing or operations mechanisms that may artificially inflate prices (without discrete value added). Research to understand how wine industry leaders can increase the efficiency of their production operations or effectiveness of their marketing channels will also be beneficial for our audience.

[4]Stuen et al. "An Analysis of Wine Critic Consensus: A Study of Washington and California Wines" Journal of Wine Economics (2015)