Capstone Report

Introduction

In 2016 Americans drank about 4.24 billion bottles of wine– that's 13.3 bottles of wine for every man, woman, and child. Those statistics make the United States the biggest global wine consumer country in the world. However, choosing which glass to order or bottle to buy is often a fraught situation for many Americans. There are many factors that go into choosing a wine, including price, variety, country of origin, region of origin, and year of bottling. Weighing these many factors can make what should be a pleasurable experience stressful.

This project will create a model that aids in the process of choosing a wine. Using various data points about wines reviewed by wine experts at Wine Enthusiast magazine, including price, country of origin, region of origin, variety, winery, and points awarded, this algorithm will predict a wine's value in order to offer guidance on whether it is a good buy or should be passed over for a different wine.

Problem Setup

Estimates vary, but at most there are no more than 5,000 certified sommeliers in the United States. The remaining 99.9% of Americans have varying levels of wine knowledge and expertise, but without spending years to study wine and pass tests certifying one as an expert, Americans from the most casual of wine drinkers to wine enthusiasts have a gap in knowledge that can be closed with additional data.

Consider a scenario– you're an average American and you go on a date at a fancy Italian restaurant. You want to impress your date by choosing a great wine, but you also know that the price of multiple delicious courses of Italian food will add up, so you want to be smart about the price. The typical approach may be to simply choose a cheap bottle, perhaps the second or third cheapest on the wine list, so that it's not obvious that you're trying to save money. Another approach may be to visit a site with a database of wine reviews and type in a couple of the wines to see which one has the best reviews for the price. But what if the site you choose does not have reviews for the specific wines on the menu in front of you?

Consider a second situation– you're a recent college graduate who wants to start a wine club with friends. You and your friends are just getting started in the working world so you don't have a massive budget. You want to choose wines for your first meeting that taste great, but you're limited by your budget and the selection at your local supermarket. You procrastinate the decision so that on the day of the meeting you're rushing home from work and stop by the store. You could look up all the wines on the shelves and compare points and price, but what you really want to know is whether the first bottle you pick up is a good choice or if you should pick up a second bottle.

A third scenario– You are a purchaser at a banquet hall that requires brides and grooms to set up their wedding bar through the venue. You need to offer a selection of wines that will satisfy a diverse group of people, but are also inexpensive enough that clients will not be shocked once you add in the venue markup. The banquet hall is perpetually understaffed, so you do not have a lot of time to pick and choose to find bottles that will please wedding guests' palates and soon-to-be honeymooners' wallets.

Ultimately, what we are all looking for is to buy the best wine at the lowest price, which can be represented by a wine's value, defined for the purposes of this study as reviewed points per USD. This model will reduce the stress of both of the scenarios described above. At the restaurant you can type in a couple characteristics, such as variety, country of origin, region of origin, and year of several of the lower-priced wines and make a judgment on which bottle is likely to be the best valye. At the grocery store you can type in characteristics of a specific wine along with its price and easily determine if you should buy it.

This model has applications for the everyday wine consumer, as well as buyers for restaurants, hotels and event spaces, professionals who wish to impress clients with gifts, and others.

The Dataset

The original dataset consists of 129,971 observations of 14 variables. Each observation corresponds to a review by a wine expert for the Wine Enthusiast database. The reviewers blind sampled the wine and assigned a points value based on the taste. Points are awarded by Wine Enthusiast on a scale of 0-100. The table below gives explanations of the points.

Score	Description
80-84	Good: a solid, well-made wine
85-89	Very good: a wine with special qualities
90-94	Outstanding: a wine of superior character and style
95-100	Classic: a great wine

Other data points about the wines were added when the review was recorded. These include country of origin, a description, price, region of origin, taster name, title, and variety.

One limitation in this dataset is that it cannot explain what makes a wine good, merely point to likelihood of wine quality based on a number of mainly geographic factors. This dataset cannot offer nuance into aspects of wine making, such as weather, soil characteristics, and harvest quality amongst other determinants of wine taste. The dataset also features issues with incomplete data. Not all observations have a recorded price, country, variety and winery.

Data Cleaning

Initial Data Wrangling

The data set was downloaded from Kaggle and was fairly clean to begin with as it was scraped from a database of wine reviews. Several columns were removed as they were not pertinent to this analysis or were duplicative of data in other columns. The taster name and twitter handle were removed as we are not analyzing the tastes of specific people, but rather the reviewers' opinions as a whole. Type and designation were removed because the information contained in these fields was represented in other columns. Region_2 was removed as it was blank in more than 50% of observances.

One of the main issues with the dataset was the presence of blank values for various columns. In the case of most columns, NAs and blanks were replace with the string "NA."

The title column had the wine's title including year, but also featured, alternately, the province of origin or region in parentheses. This information in parentheses was removed as it is contained in a different column.

For the purpose of this analysis, observations that did not include a price were removed. This left a total of 120,975 observations. Records not containing price were removed as the price is a necessary component of determining the value of a wine, the metric used as the base for wine quality.

Additional Columns Added

- The "value" column was added by dividing points by price (in USD). This column will serve as the determinant for the quality of a wine. A higher value indicates that the wine has a higher points per dollar value and is thus a better value.

- The "continent" column was added using the country column in the original dataset. This column was added in order to analyze trends in wine by continent.

- The "year" column was added by extracting years from the title in the original dataset. After sampling instances where the year that was extracted was earlier than 1950, it appeared that many wine titles feature years that the wine was not bottled in, rather the year is in the name of the winery. This year often indicates the year that the winery was founded. To control for this, the year column is limited to cases where the date in the title is more recent than 1950.

Description of Final Variables

After data cleaning we are left with 120,0975 observations of 12 variables.

- **X** - unique identifier
- **country** - the country of origin of the wine
- **continent** - the countinent of origin of the wine
- **province** - the province of origin of the wine
- **region** - the wine growing region of the wine
- **winery** - the winery that created the wine
- **title** - the name of the wine
- **year** - the vintage of the wine
- **variety** - the variety of the wine
- **points** - the points awarded by a Wine Enthusisast reviewer
- **price** - the price in USD of the wine per bottle
- **value** - the points awarded to the wine per USD

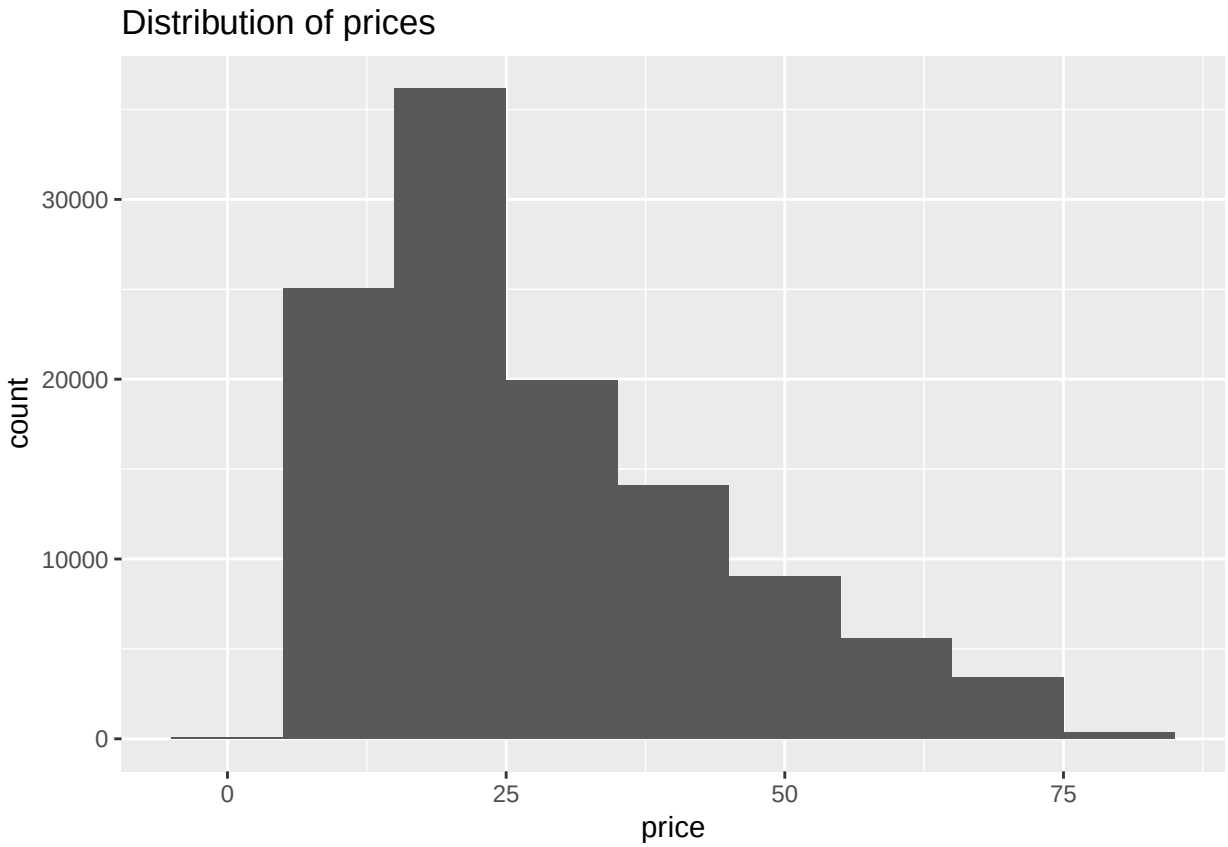
Exploratory Data Analysis

Summary Statistics

- All wine-producing continents are represented, with North America the most-represented continent with 54,589 observations. Asia is least-represented with 652 observations.
- The wines reviewed come from 43 unique countries. The most-represented country is the United States with 54,265 observations and least-represented is Slovakia with 1.
- For American wines, province typically corresponds to state. The wine-producing powerhouse state of California is the most represented of the 423 worldwide provinces in the dataset.
- The Napa Valley of California (4475 observations) is the most commonly reviewed region of 1205 regions.
- There are 698 distinct varieties of wine with the popular Pinot Noir (12,787) and Chardonnay(11,080) as the most commonly occurring wine varieties.

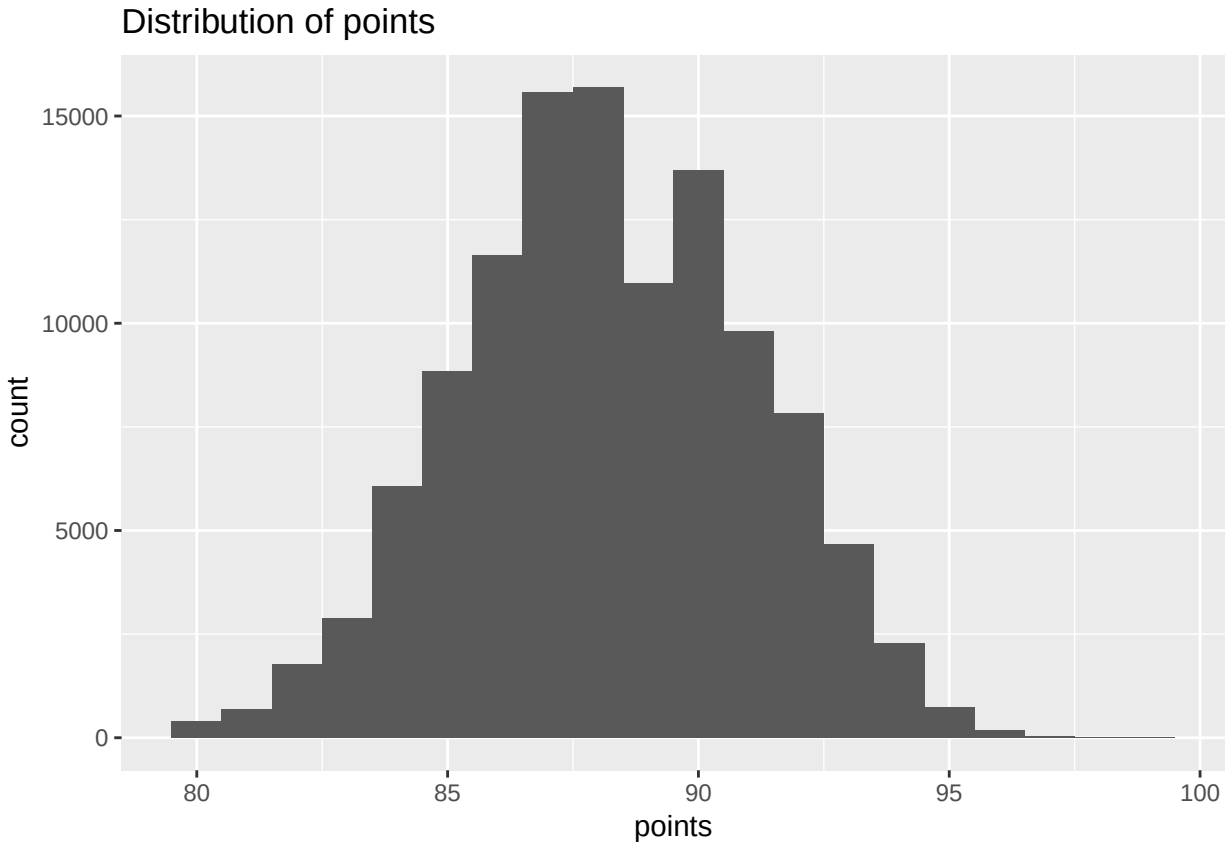
Price

- The distribution of price is right skewed as the median price is $25 and mean price is $35.36, with a handful of prices that are far above average. The maximum price in the original dataset is $3300. The minimum price is $4.



Points

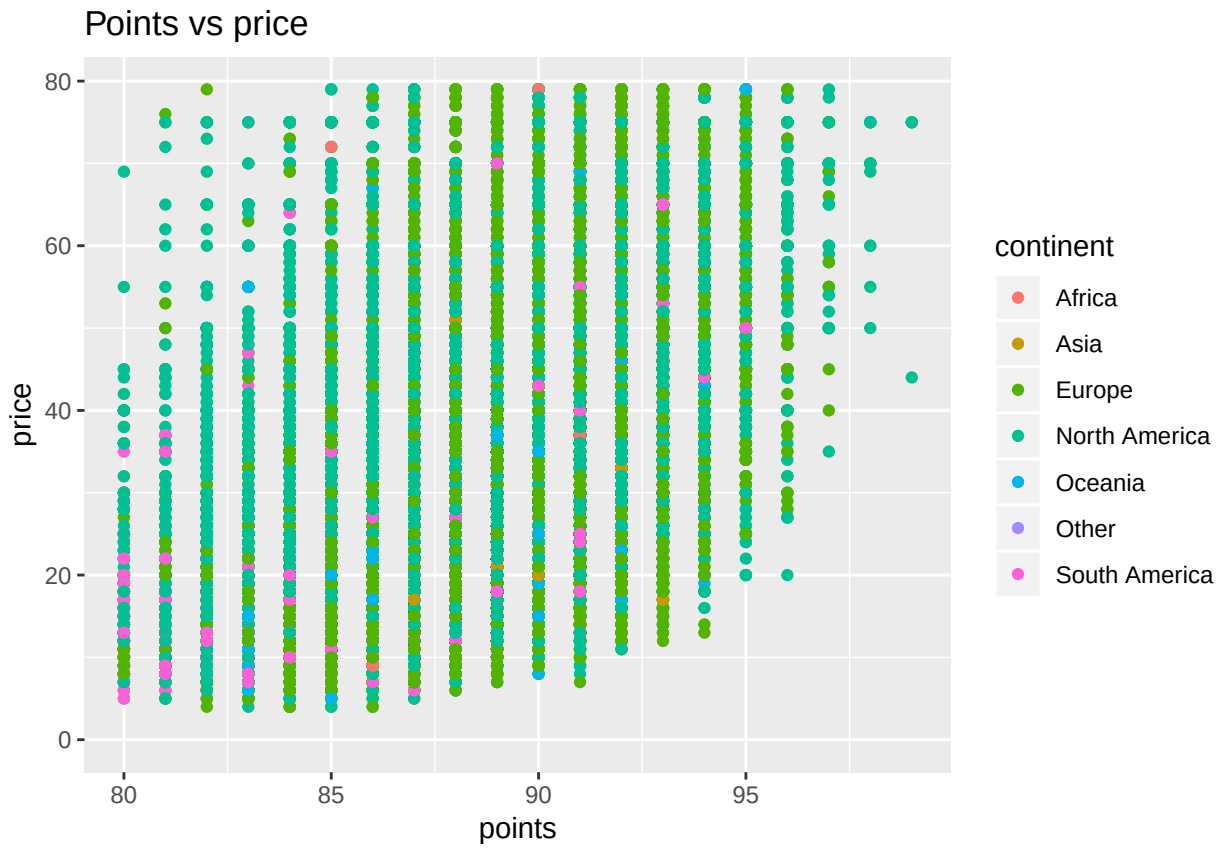
- Points are awarded by Wine Enthusiast on a scale of 0-100. This dataset only contains wines with point assignments of 80 and up, indicating that all of the wines can at least be considered to be "good." The median of points is 88 and mean 88.42, leading one to think that the distribution may be normal. However, an Anderson-Darling normality test reveals a p-value less than .05, indicating that the distribution cannot be considered normal.



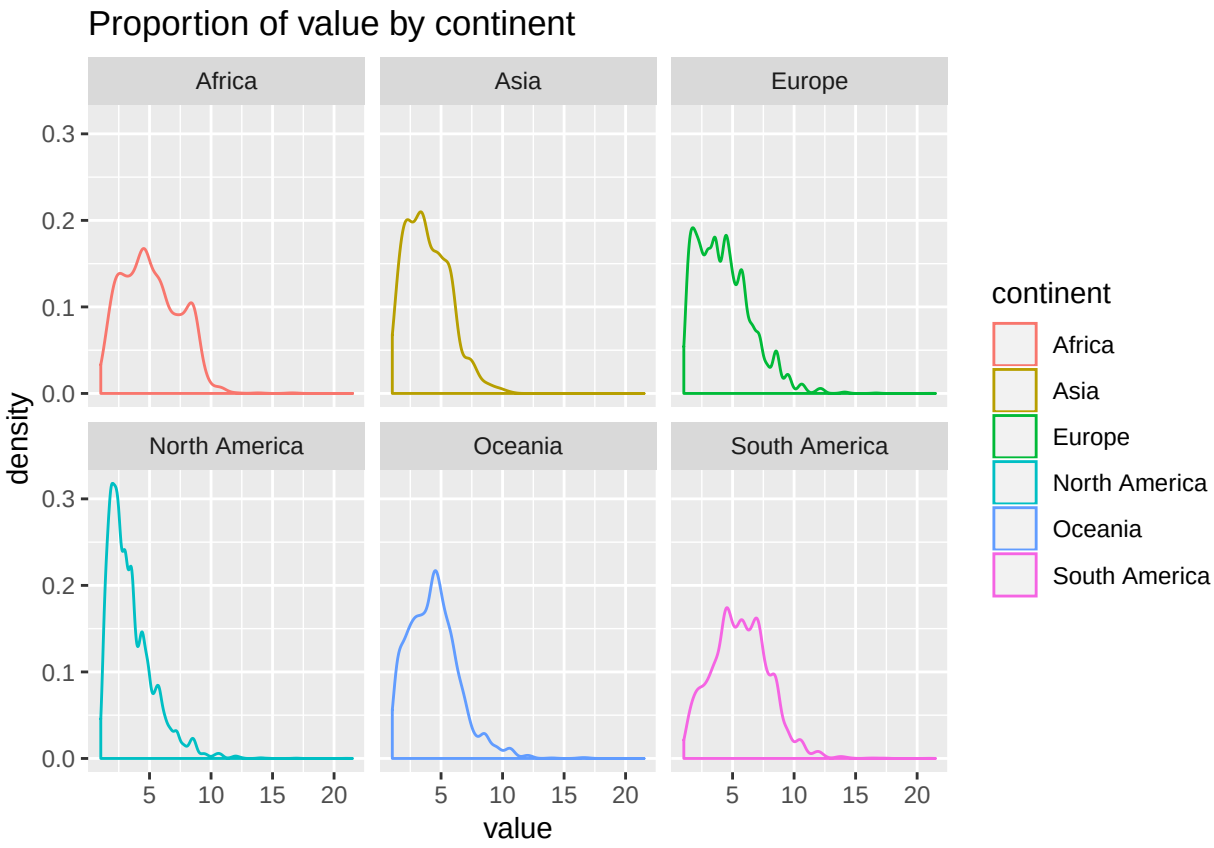
Value

Value is a function of points/price, thus a higher value indicates that the wine is a good purchase. When working with a dataset with a drastically large range of prices (from 4 USD to 3300 USD), and a points scale that only varies by 20 points, value is an imperfect measure. For example, a 4 USD wine with a score of 88 has a value of 22, while a wine with a price of 3300 USD, even with a perfect score of 100, only has a value of .03.

Pearson's correlation test indicates that the correlation value of points and price is 0.416. This indicates a weak to moderate correlation between the two.



Continent appears to be correlated with value. Certain continents have greater proportion of wines with higher values than others. South America has the highest median value at 5.5 points/USD and North America the lowest– 2.93 points/USD. Not coincidentally, South America has the lowest median price at 15 USD per bottle and North America the highest at 30 USD per bottle.



Predictive Model

Choosing a Predictive Model

Several types of machine learning methods were explored in an attempt to build the best model for predicting wine value based on factors present in wine reviews– chiefly linear regression, random forest, and gradient boosting machine.

Linear regression was used due to limitations with a tree-based model. The main issue that makes this dataset ill-suited for a tree-based model is that there are several independent variables with a large number of levels. For instance, region_1 has 1205 levels and variety 698. A random forest model cannot support a categorical variable with such high numbers of levels. One frequent solution to this issue is to use one-hot encoding. This was not a proper solution for this dataset, as having so many levels to certain variables ends up reducing the importance of the variable overall, due to there being small numbers of occurrences of many levels of the variables. This issue applies to the gradient boosting machine method as well. With some adjustments, a linear regression model proved to be quite effective at predicting wine value.

Building a Linear Regression Model

An initial linear regression model was tuned by splitting the dataset into a 75-25 split training and testing sets. Using adjusted R-squared as metric of goodness-of-fit, the model

```
model <- lm(value ~ points + price + year + variety + region_1, data = train)
```

was fine tuned to have an R-squared of 0.6126 and adjusted R-squared of 0.6045.

When attempting to predict values for the testing set based on the model, issues arose due to the nature of the dataset. Namely, certain uncommon values for variety, region_1, and province that were randomly sorted into the testing set were not present in any observations in the training set. To properly evaluate a predictive model, cross validation was applied.

Cross Validation Linear Regression Model

The caret package was used to perform a 10-fold cross validation linear regression model.

```
model_cv <- train(
  value ~ points + price + year + province + variety + region_1, model_set,
  method = "lm",
  na.action = na.pass,
  trControl = trainControl(
    method = "cv", number = 10,
    verboseIter = TRUE
  )
)
```

For this analysis:
* R-squared = 0.6449 * Adjusted R-squared = 0.6387 * RMSE = 1.265

Running a Modified Model

This R-squared value of this first model indicates that it explains 64.49% percentage of variance in the data. I think we can do better than that. One issue with the dataset that was identified early on is that there are a fair number of major price outliers. Recall that the median price value for the dataset is 25 USD with a maximum price of 3300 USD. As outliers can often have an outsized effect on linear regression models, I trained a second model that had outliers removed.

Using Tukey's method for outlier removal, wines with prices above 79.50 USD were removed from the dataset. When one considers that the average consumer will likely very rarely purchase a wine over 80 USD in a retail setting and more rarely still a $200 bottle in a restaurant (assuming a very modest, average restaurant mark up of 2.5 times), the removal of data is not alarming. It is also reasonable to assume that professionals using this predictive model are unlikely to spend that much on a bottle of wine as well.

Following the same workflow as the first model, a linear regression model was fitted to this modifed dataset.

```
modified_cv <- train(
  value ~ points + price + year + province + variety + region_1, mod_model,
  method = "lm",
  trControl = trainControl(
    method = "cv", number = 10,
    verboseIter = TRUE
  )
)
```

For this model: * R-squared = 0.7904 * Adjusted R-squared = 0.7955 * RMSE = 0.9198

Removing approximately 5% of the data to increase the percentage of variance in the data by ~15% is a good trade-off when considering the needs of the client that this model is designed to serve.

Future Research

When considering further research and additional features to add to the analysis, several improvements come to mind. One additional feature that would be useful for many people is the inclusion of a color - ie white or

red wine. Even someone who has had little experience likely has a preference for either white or red wine. When looking for a reccomendation, being able to filter for wine color would be helpful for many.

Another useful feature is wine suggestion based on food type. One would hope that when the bottle of wine is consumed, it is accompanied by food. The pairing of specific wines with food follows many generally conceived theories, and is also the main purpose of a sommelier, a professional wine taster.

An additional future method of study might focus on an analysis that is separated by country or continent. In some cases, this would solve the issues that arise with having too many levels for certain independent variables, as region_1, province and variety are all characteristics that vary less from country to country or continent to continent. Region_1 and province are easily explained as they are geographical, and wine variety because certain types of grapes grow better in certain places and are ingrained in local cultures.

Conclusion

In summary, I have built a model that uses data to boost confidence in wine decisions for everyone from a once-a-year drinker to a value-conscious connoisseur. The dater on a budget can order the cheapest bottle on the menu with confidence that it is likely to have a high points to dollar ratio. The recent grad purchasing for wine club can feel good about quickly passing over the first bottle on the shelf to pick up the next. The events space manager can assure their boss that wedding-goers, brides and grooms will be happy with the wines they swiftly selected for their bar menu.

Ultimately, wine is a matter of personal taste, so when given the option to buy a wine that you have previously enjoyed at a price that you are willing to pay, go for it. However, in the cases where you are new to or unfamiliar with the world of wine, or faced with unknown choices, you can turn to data to increase the likelihood of maximizing your spend on a bottle of wine.