Milestone Report

Abstract

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 an algorithm 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 and offer guidance on whether it is a good buy or should be passed on.

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 one, but the second or third cheapest, 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? You're running out of time to impress your date.

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 good or if you should pick up a second or third bottle.

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 be given a judgement on which is likely to have the highest value. At the grocery store you can type in characteristics of a specific wine along with its price and receive a rating of good or bad buy instantaneously.

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 taste of specific people, but rather the reviewers' opinions as a whole. Type and designation were removed because the information they contained was represented in other columns. Region_2 was removed as it was blank in more than 50% of observances.

One of the main issues was 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 analysis 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 started. 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

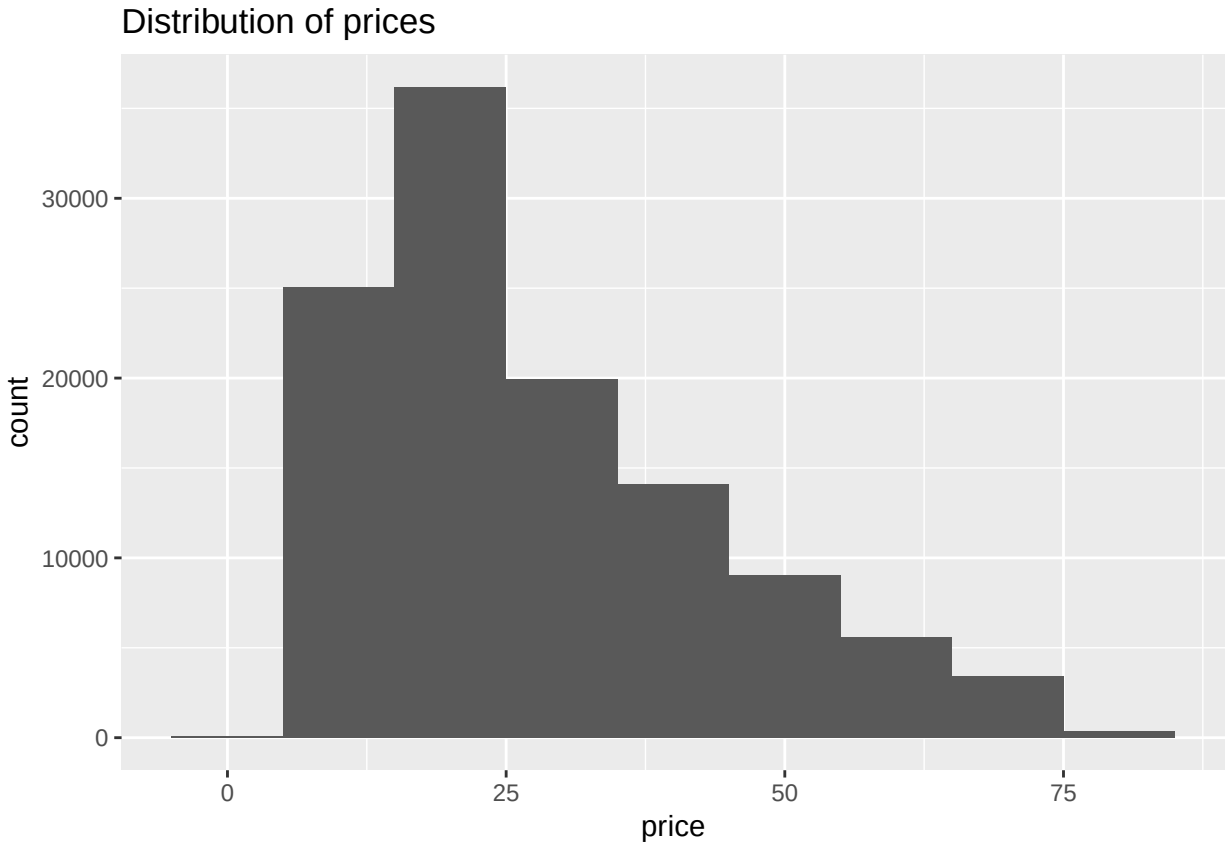
Summary Statistics

- 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

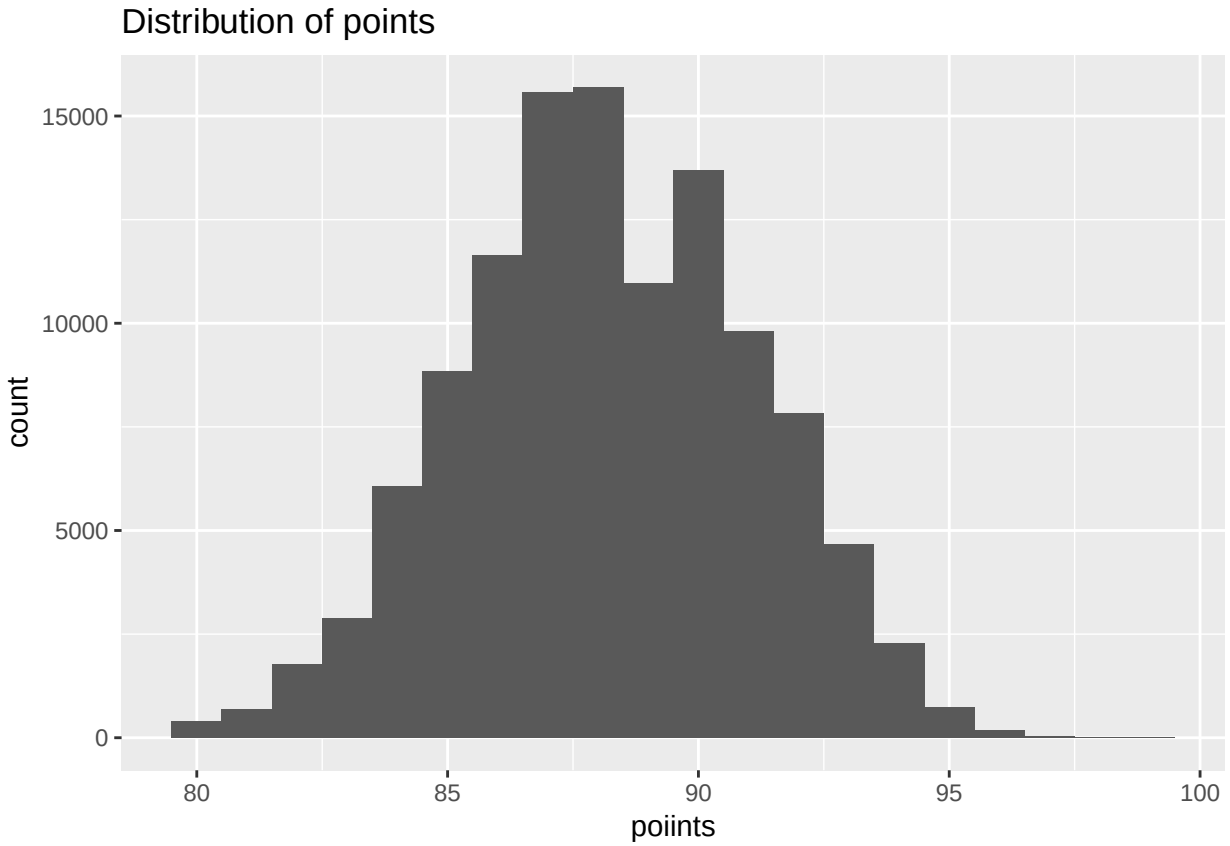
- The distribution of price is right skewed as the median price is $25 and mean price is $35.36, with a handful of prices that are far above average. The maximum price in the original dataset is $3300. The minimum price is $4.

- Outliers were removed using Tukey's method, meaning that ~7000 observations of wines with price above $79.50 were removed. When one considers that the average consumer will likely very rarely purchase a wine over $80 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 ~5% of the dataset is less alarming.



Points

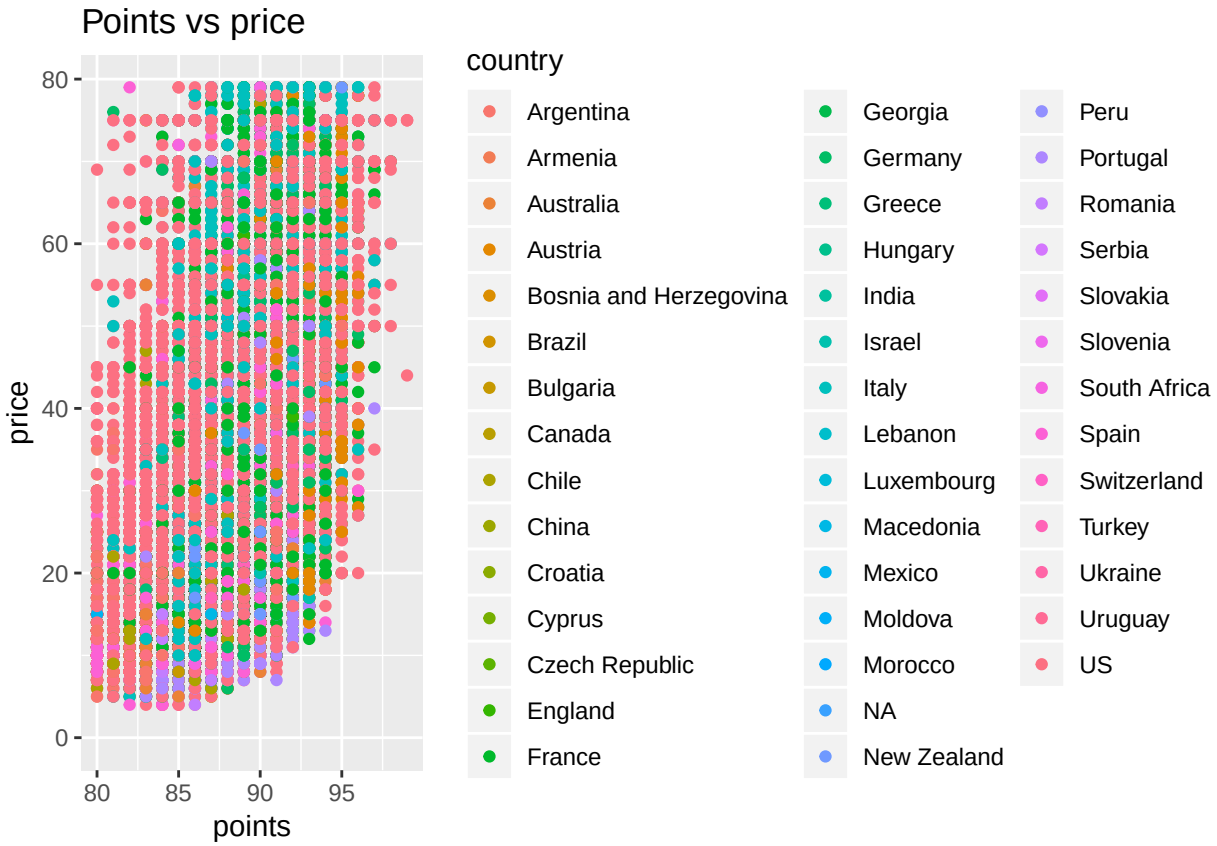
- Points are awarded by Wine Enthusiast on a scale of 0-100. This dataset only contains wines with point assignments of 80 and up, indicating that all of the wines can at least be considered to be "good." The median of points is 88 and mean 88.42, leading one to think that the distribution may be normal.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



Value

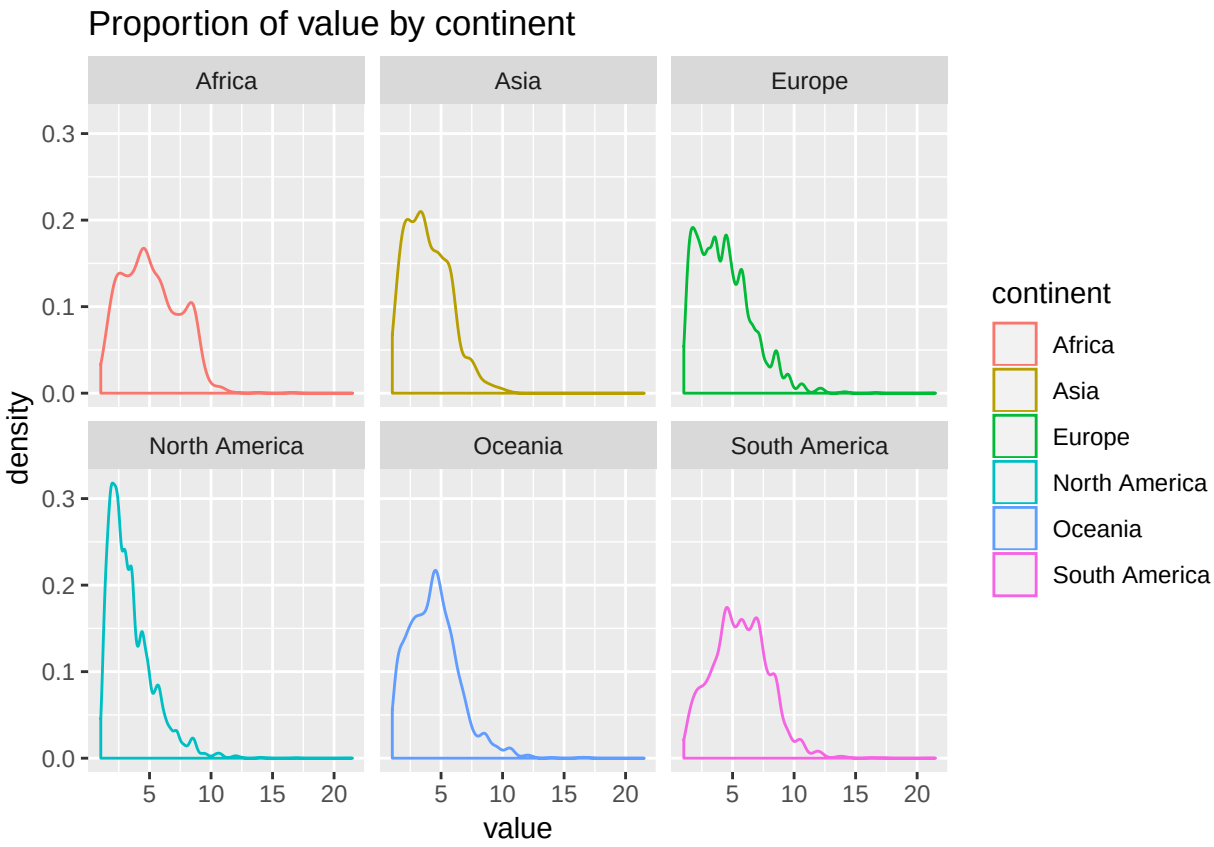
Value is a function of points/price, thus a higher value indicates that the wine is a good purchase. When working with a dataset with a drastically large range of prices (from 4 USD to 3300 USD), and a points scale that only varies by 20 points, value is an imperfect measure. For example, a 4 USD wine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

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 high values than others.

Perhaps better place to put the difference of means? Not really sure how to explain why I have this plot.



Future Approach

There are several additional types of analysis to perform to determine which factors may have an affect on wine value.

- In furthering the analysis, various attributes of the wine will be examined to determine a more convincing correlation to value. For instance, the plot below indicates that value may be correlated to the continent of origin of the wine. Other variables, such as year, country, and region will be examined as well.

- Adding an additional column for red vs white may provide insight into wine value.
- I would like to apply machine learning to the description field in the original dataset to determine if certain characteristics are correlated with a higher value. This could also help people determine a wine that not only has a high value, but is likely to appeal to their own personal palate or pair with the food that eat along with their wine.

As a final deliverable, I will apply machine learning techniques to the dataset to create a model that can predict wine value and recommend whether a wine should or should not be bought. I will split my data to use 75% as a training set and use the remaining 25% of the data to test the accuracy of the model.