What Makes a Good Wine?

Due Friday, November 20, 11:59 PM

Tennis Trio: Hamilton Murrah, Kellyn McDonald, Naima Turbes

Introduction

People have cultivated and appreciated wines for millennia. In fact, the oldest known winery was found in a cave outside of a village in Armenia and dates back to 4,100 BC. Since then, wine production and consumption has spread throughout the globe and become a important part of many cultures. Today, wine is largely used as a celebratory drink during activities like weddings or holidays. Fine wines are often more expensive and associated with higher societal status or class[1].

Most current wine production happens in Europe. In 2019, Italy produced 21.6 million hectoliters of wine and was the world's largest wine producer. The next two largest wine producers that year were Spain and France, producing 21.3 million hectoliters and 14.2 million hectoliters respectively [2]. Nevertheless, a 2019 survey from Forbes found that the the best ranked wine was from Napa Valley in the United States [3]. We are curious to see if this has changed.

We investigated wine ratings from wine enthusiasts around the globe using wine ranking data scraped from WineEnthusiast.com.

Our main motivation for investigating wine data is pure interest and desire to understand how to select the best wine. We are specifically interested in how the perceived quality of these wines varies based on where they were produced. Additionally, we want to understand what words wine enthusiasts use to describe highly rated wines versus lower rated wines. Finally, in an attempt to find the best wine to buy, we want to investigate and characterize what wines have the best value for their rating. We hypothesize that the best wines are also the most expensive wines, come from France and Italy, and are described using positive descriptors.

Data Description

User zackthoutt on Kaggle collected this data by scraping data from WineEnthusiast.com.

A description of the columns of our data set are as follows

- `X1`: Number corresponding to the observation. Listed in ascending order from 0
- `country`: The country that the wine is from
- `description`: A few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.
- `designation`: The vineyard within the winery where the grapes that made the wine are from
- `points`: The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score ≥ 80)
- `price`: The cost for a bottle of the wine
- `province`: The province or state that the wine is from
- `region_1`: The wine growing area in a province or state (ie Napa)
- `region_2`: Sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley), but this value can sometimes be blank
- `variety`: The type of grapes used to make the wine (ie Pinot Noir)
- `winery`: The winery that made the wine

Glimpse of data

```
## Rows: 8,000
## Columns: 11
## $ X1          <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1...
## $ country     <chr> "US", "Spain", "US", "US", "France", "Spain", "Spain", ...
## $ description <chr> "This tremendous 100% varietal wine hails from Oakville...
## $ designation <chr> "Martha's Vineyard", "Carodorum Selección Especial Rese...
## $ points      <dbl> 96, 96, 96, 96, 95, 95, 95, 95, 95, 95, 95, 95, 95, 95,...
## $ price       <dbl> 235, 110, 90, 65, 66, 73, 65, 110, 65, 60, 80, 48, 48, ...
## $ province    <chr> "California", "Northern Spain", "California", "Oregon",...
## $ region_1    <chr> "Napa Valley", "Toro", "Knights Valley", "Willamette Va...
## $ region_2    <chr> "Napa", NA, "Sonoma", "Willamette Valley", NA, NA, NA, ...
## $ variety     <chr> "Cabernet Sauvignon", "Tinta de Toro", "Sauvignon Blanc...
## $ winery      <chr> "Heitz", "Bodega Carmen Rodríguez", "Macauley", "Ponzi"...
```

Methodology

We first calculated on average which countries have the highest rated wines by finding the mean scores of each country and summarizing the data into a table. We did the same calculation for top varieties of wine.

Countries with Highest Average Wine Point Score

```
## # A tibble: 5 x 2
##   country   avg
##   <chr>   <dbl>
## 1 Austria  90.8
## 2 Germany  89.2
## 3 France   89.1
## 4 US       89.0
## 5 Italy    89.0
```

Contrary to our hypothesis, Italy and France were not the top two countries. Both Austria and Germany had on average higher point ratings for their wines than these two countries. The top 5 countries with the highest average wine score are Austria, Germany, France, US, and Italy in that order.

Varieties with the Highest Average Wine Point Score

```
## # A tibble: 5 x 2
##   variety                     mean
##   <chr>                      <dbl>
## 1 Cabernet-Shiraz             96
## 2 Cabernet Sauvignon-Shiraz  94
## 3 Blauburgunder               93
## 4 Friulano                    93
## 5 Carignan                    92.6
```

```
## # A tibble: 5 x 2
##   variety           mean
##   <chr>            <dbl>
## 1 Gewürztraminer    90.6
## 2 Grüner Veltliner  90.6
## 3 Pinot Gris        89.9
## 4 Pinot Noir        89.8
## 5 Riesling          89.8
```

We calculated the top 5 best rated varieties of wine similarly. Our first calculation returned Cabernet-Shiraz, Cabernet Sauvignon-Shiraz, Blauburgunder, Friulano, and Carignan as the top rated wine varieties.

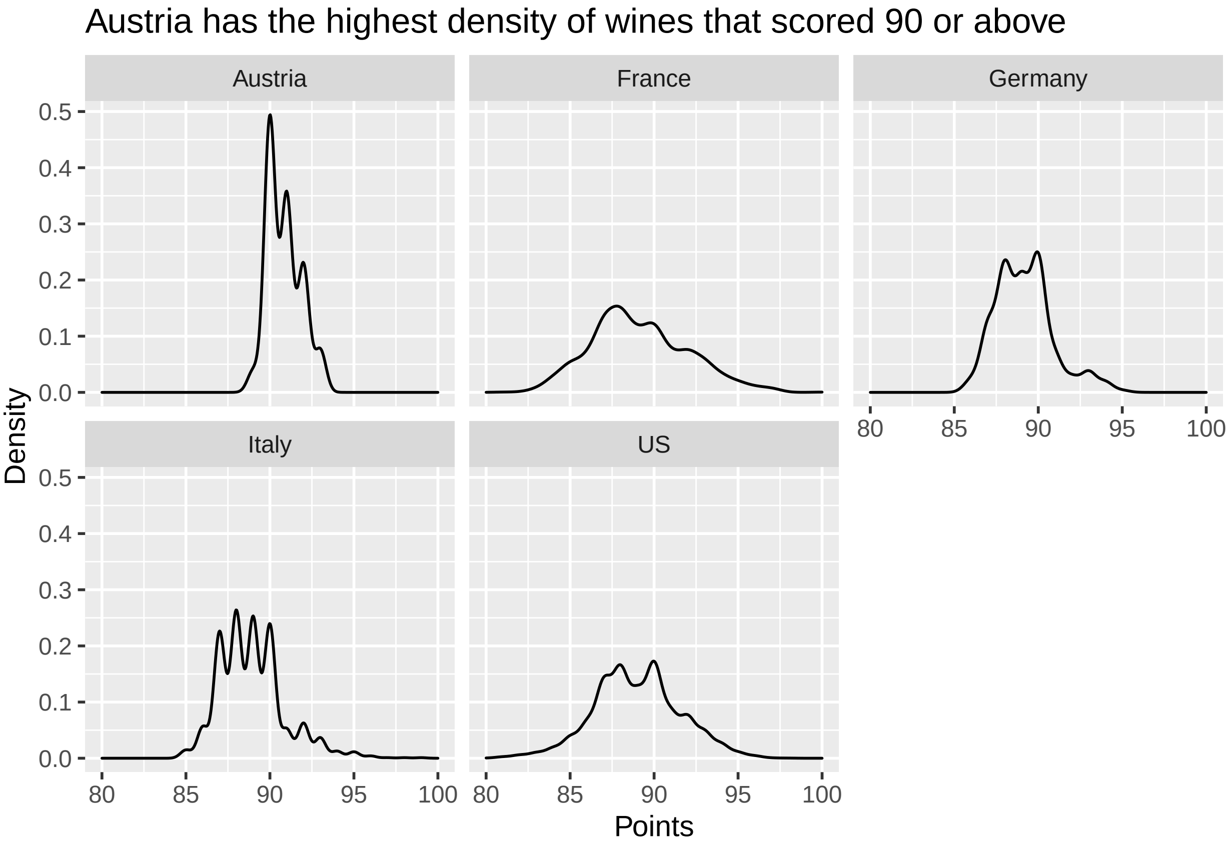
However, after observing the data, we noticed that these wine varieties had a relatively small number of wines represented in the entire data set. To try to better represent the varieties of wines with consistently top rated products, we filtered to only include wine varieties with at least 50 wine ratings represented in the data set. The top varieties presented in this calculation were Gewürztraminer, Grüner Veltliner, Pinot Gris, Pinot Noir, and Riesling.

Varieties with the Lowest Average Wine Point Score

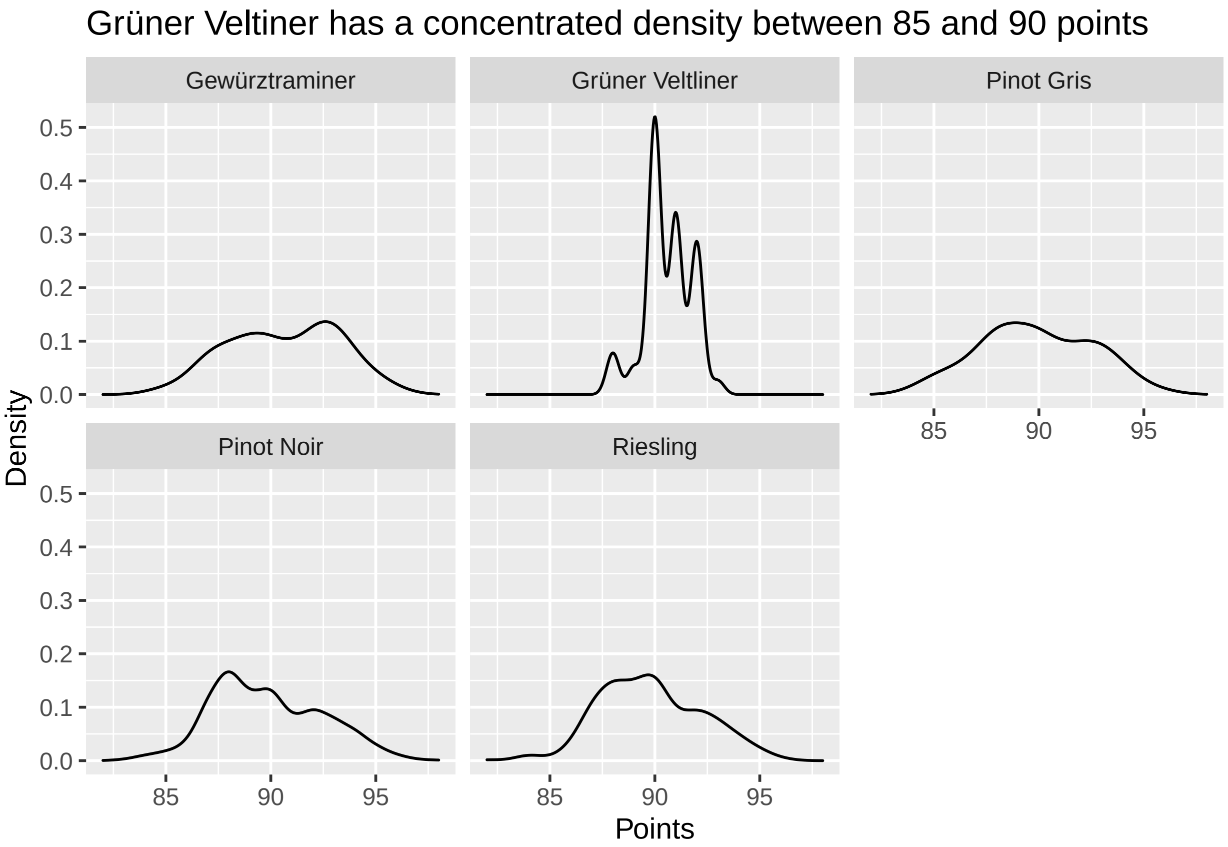
```
## # A tibble: 5 x 2
##   variety           mean
##   <chr>            <dbl>
## 1 Portuguese White  87.0
## 2 Merlot            87.7
## 3 Malbec            87.8
## 4 White Blend       87.8
## 5 Viognier          87.8
```

We followed this same methodology in order to calculate the bottom 5 rated varieties. We again filtered the data to only consider varieties with 50 or more samples in the data set. The bottom 5 varieties from this calculation were Portuguese White, Merlot, Malbec, White Blend, and Viognier.

We then created density plots for the point distribution for the top 5 varieties and top 5 countries.

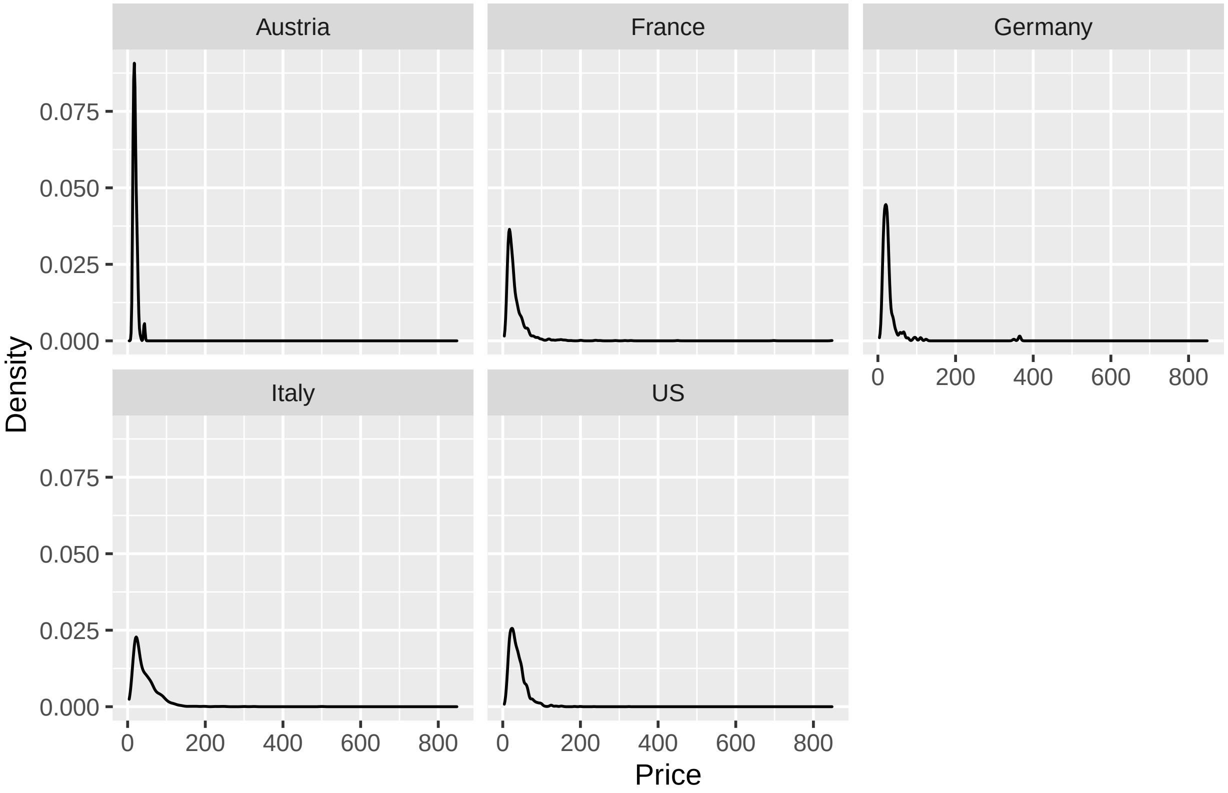


The density plots of the point distribution for the top five countries are consistent with the mean point values calculated earlier. Austria has a very high density of wine scores around 90 that likely contributes to the country's higher average wine point score.

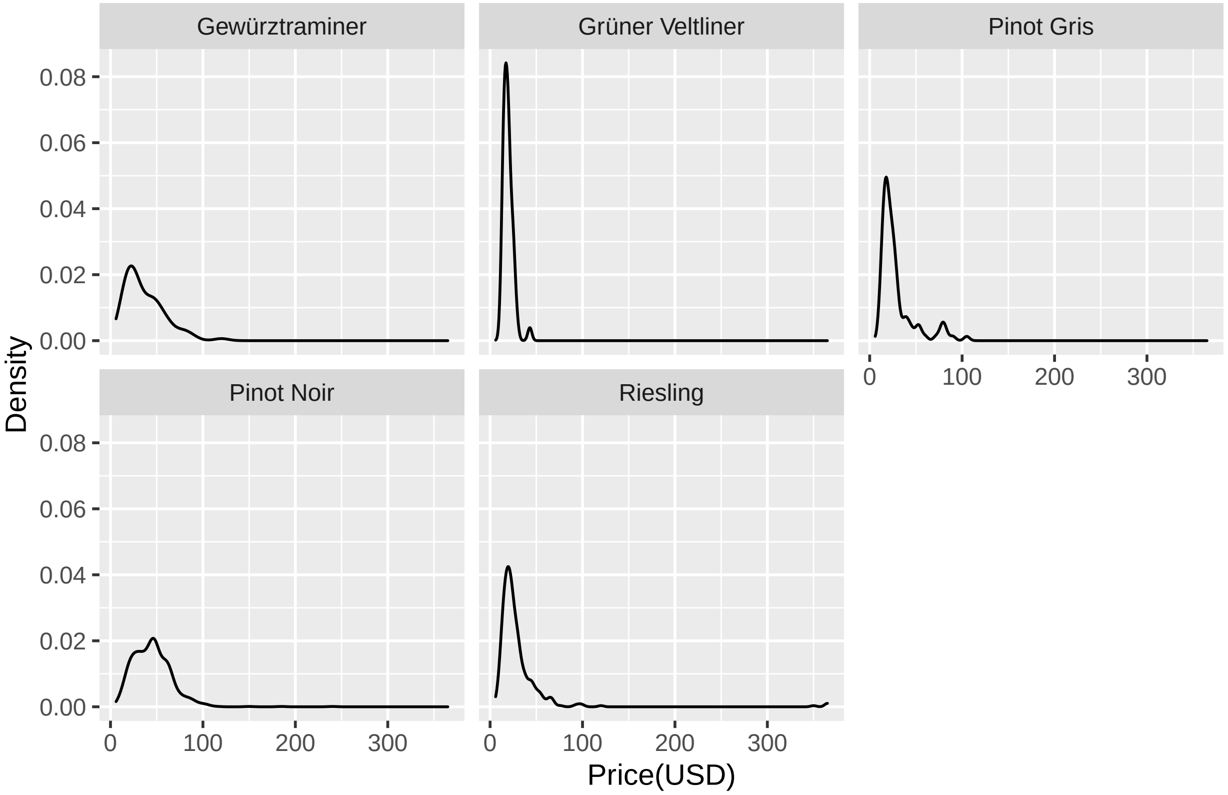


The density plots for the point values for the top five varieties of wine display an interesting trend. Gewürztraminer, Pinot Gris, Pinot Noir, and Riesling varieties show more evenly distributed point values for wines whereas Grüner Veltliner's data distribution is concentrated around the 90 point value. Unlike with Austria, however, this larger density of wines around the 90 point value did not make the Grüner Veltliner the variety with the best average point value.

The majority of wines in all top countries are cheap



The majority of wines in all top varieties are cheap



The density plots for the top five varieties and top five countries all look generally similar. The data is left skewed and most of the wines are less than 100 dollars. One noticeable difference is that the Gruner Veltliner variety and Austria have cheaper wines compared to the other top varieties or countries.

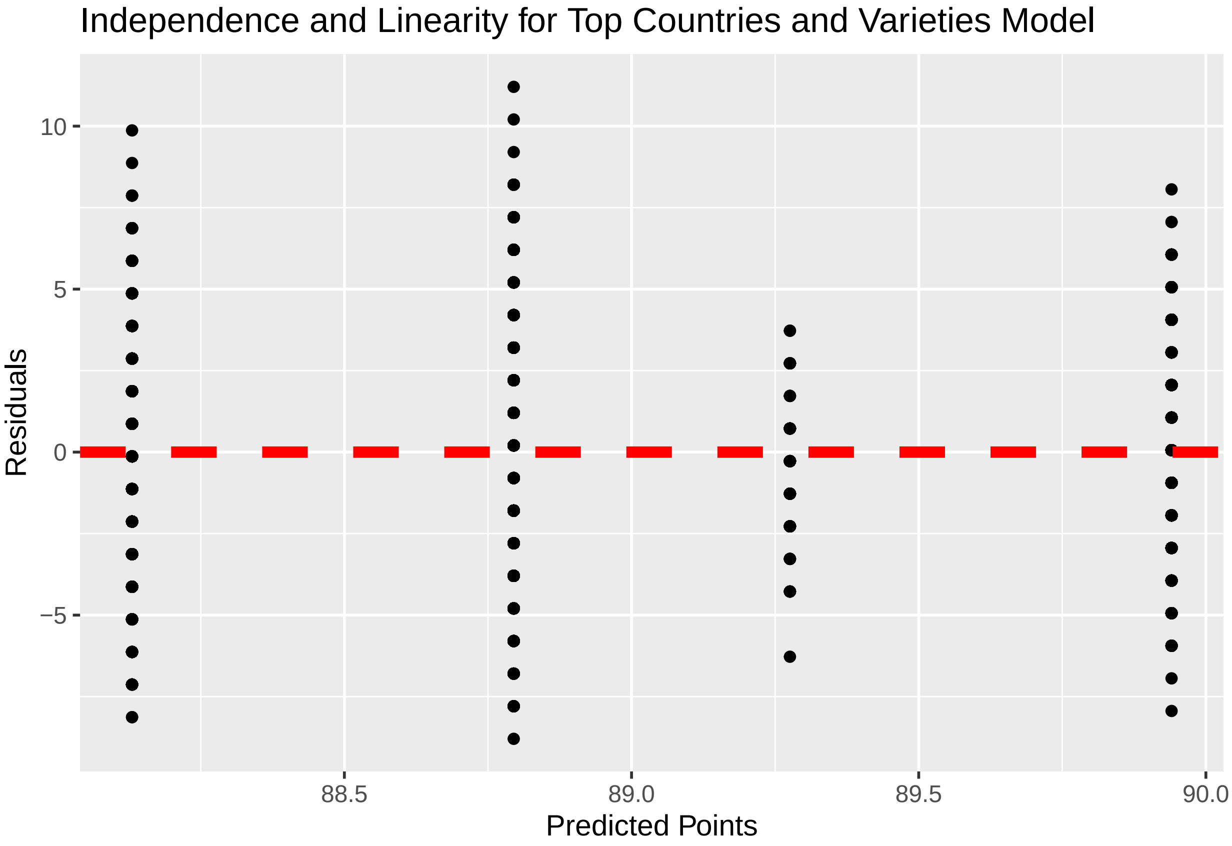
We conducted the following tests below to evaluate our hypotheses.

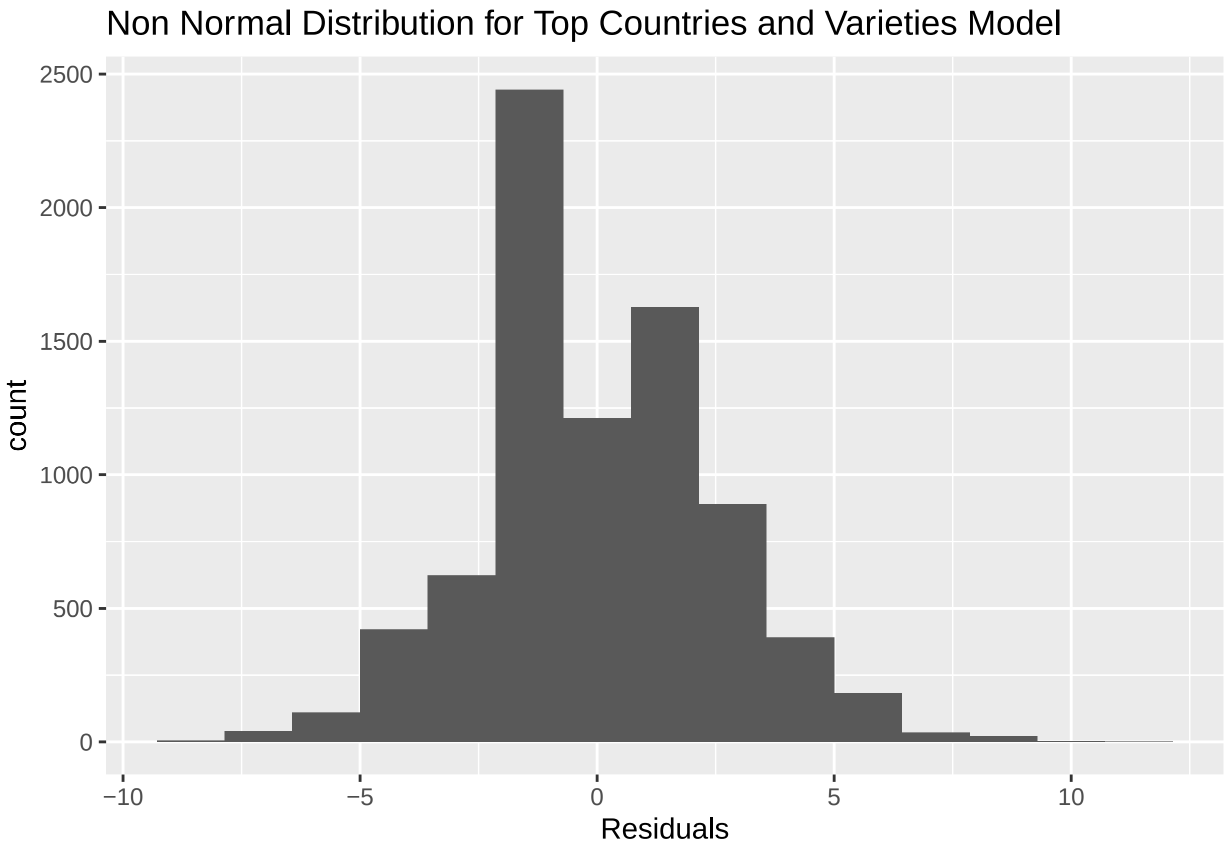
1. We created a linear model for the predicted points a wine received based on the price of a wine.

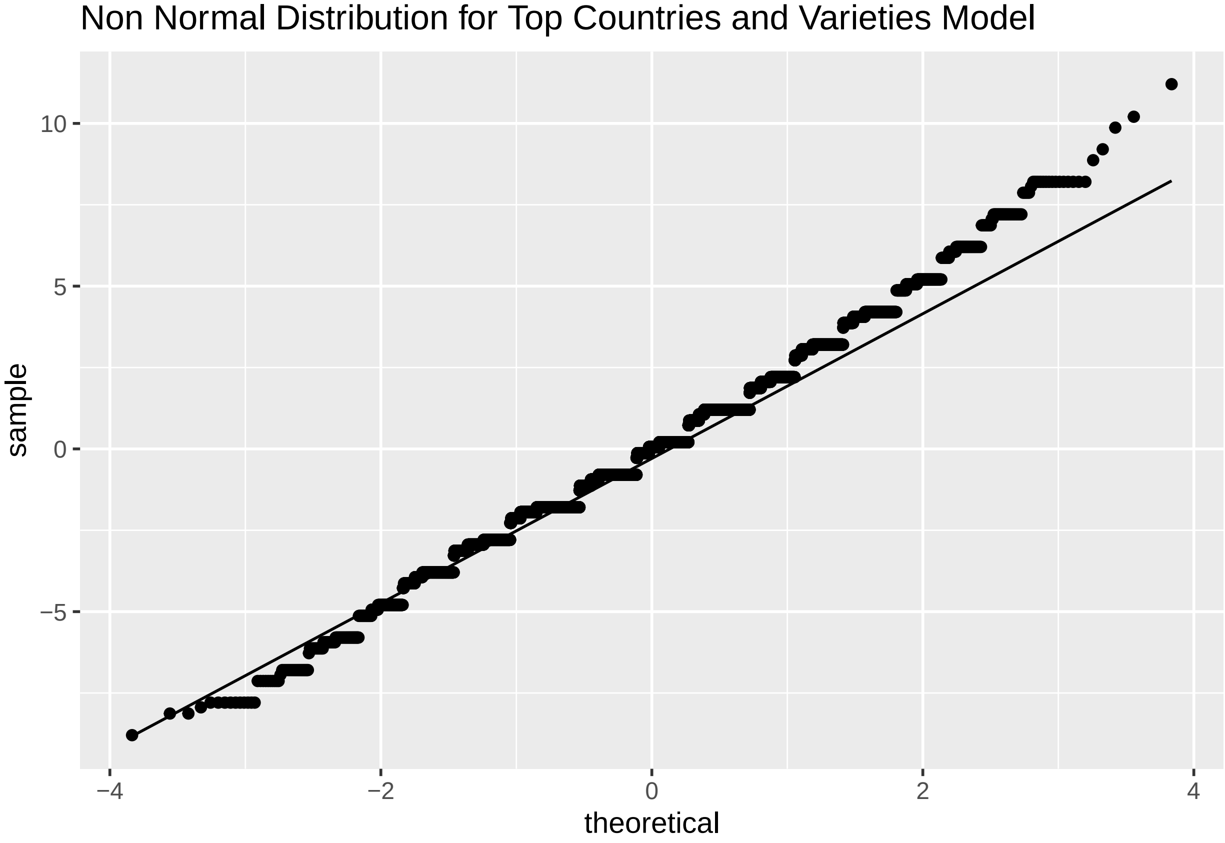
We evaluated the data below to characterize how appropriate it is to create a linear model for the data.

```
## # A tibble: 2 x 2
##   term        estimate
##   <chr>          <dbl>
## 1 (Intercept)  87.7
## 2 price         0.0321
```

```
## # A tibble: 3 x 5
##   term           estimate std.error statistic  p.value
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     88.1      0.0621   1418.   0.
## 2 top_countries    0.664    0.0705      9.42 5.56e-21
## 3 top_varieties    1.15     0.0717     16.0  1.63e-56
```


Independence and Linearity for Top Countries and Varieties Model





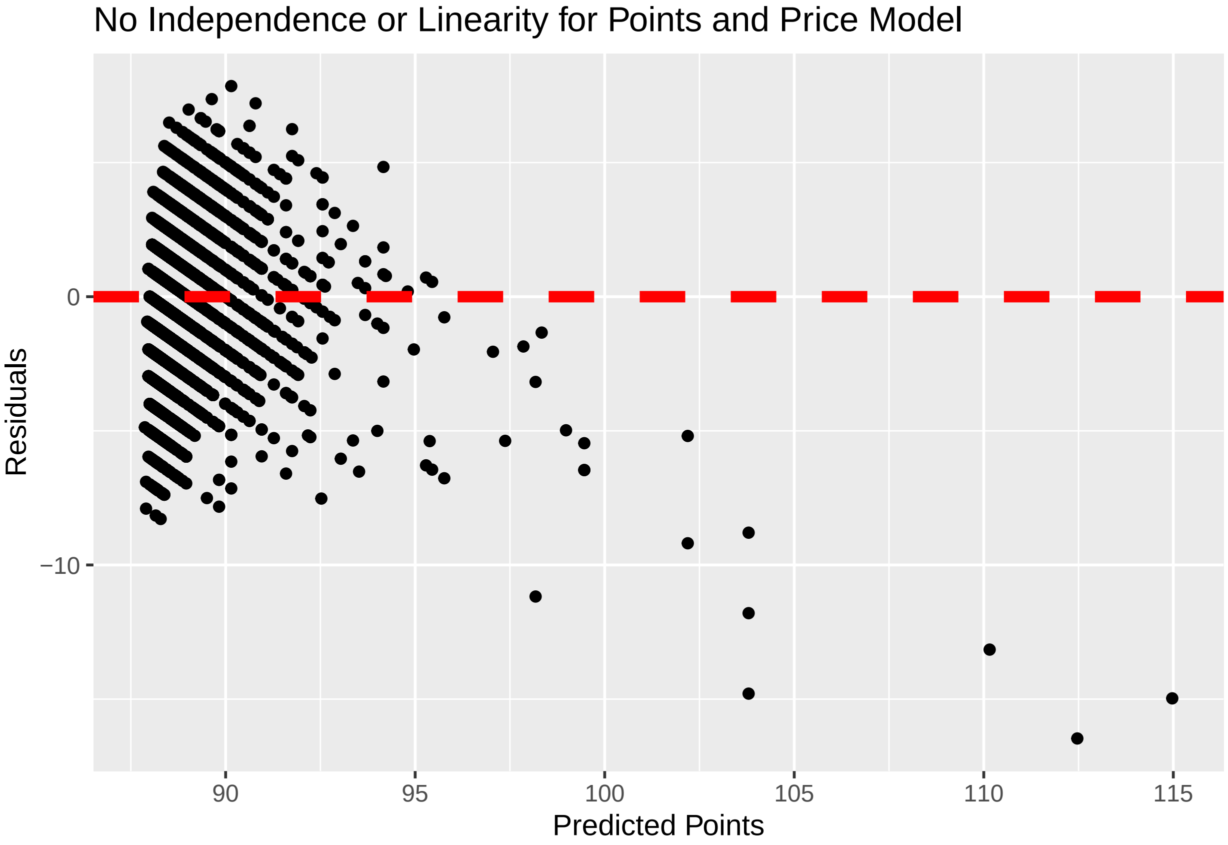
We can assume independence for this data due to the way that the data was randomly collected. We have symmetrically distributed data around the y=0 line to support linearity. We also have equal variance because each vertical slice in the graph has a similar spread of data.

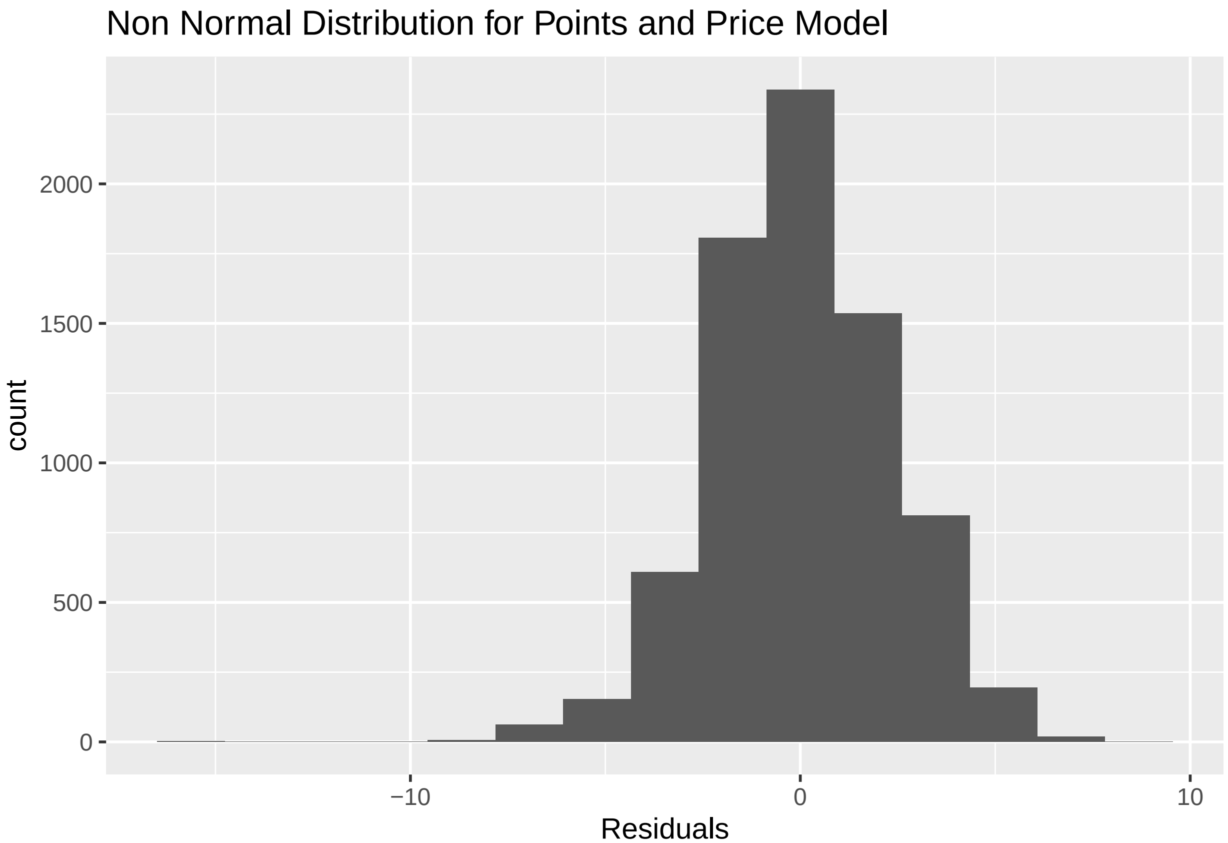
We do not have data to support normality as seen where the right portion of the final graph does not follow the line for normally distributed data.

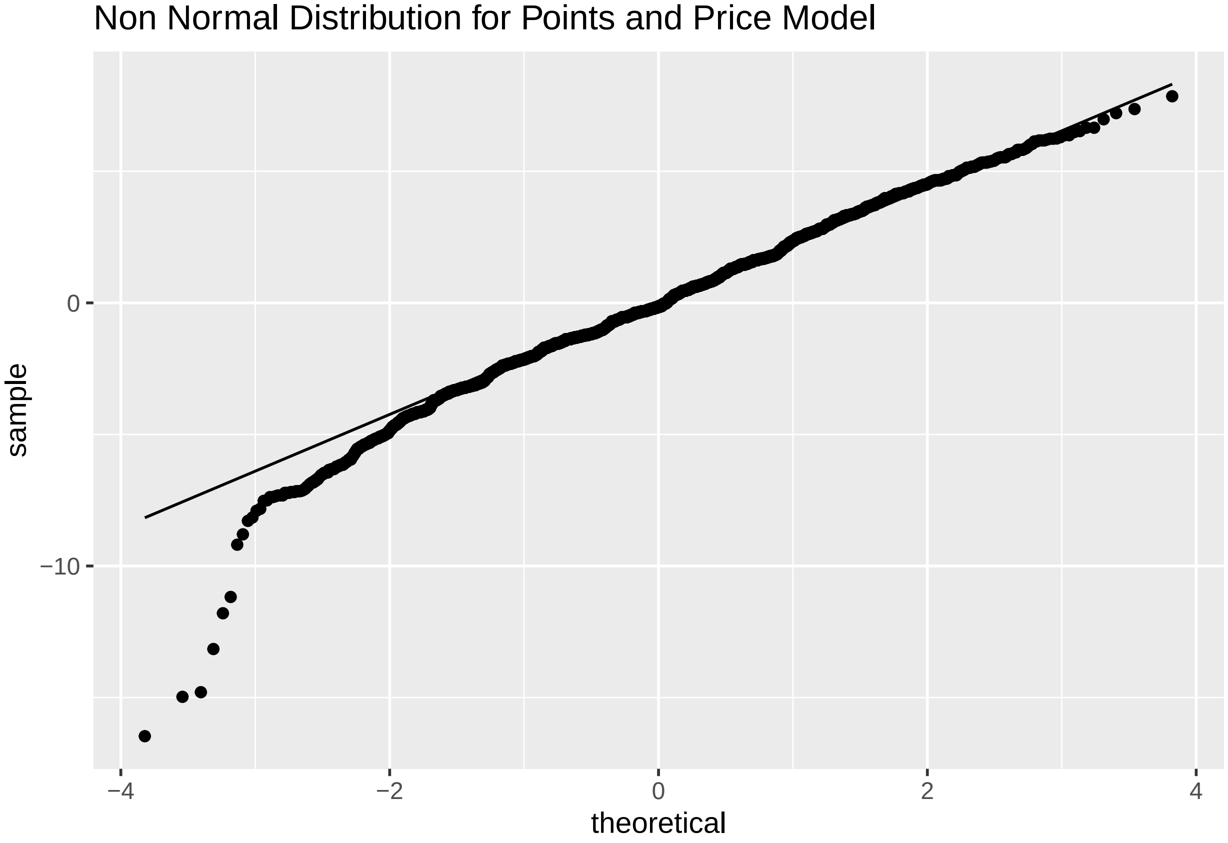
2. We created another linear model to predict the number of points that a wine would receive depending on whether or not it was made in a top country and whether it was a top variety or not.

We evaluated the data below to characterize how appropriate it is to create a linear model for the data.

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  87.7     0.0392      2238.  0.
## 2 price         0.0321  0.000812      39.6 3.77e-311
```







We can assume independence for this data due to the way that the data was randomly collected. We have equal variance because there is a each vertical slice in the graph has a similar spread of data.

Nevertheless, we do not have symmetrically distributed data around the y=0 line to support linearity. Additionally, we do not have evidence to support normality as seen where the left portion of the final graph does not follow the line for normally distributed data.

3. We used a CLT based t test to determine whether there is a statistically significant difference between the highest rated country's average score and the average wine score for all countries.

We know that it is appropriate to use a CLT based test because the samples are independent of each other (taken randomly) due to the way that the data was collected. Moreover, n>30 in this sampling.

4. We used text analysis on the description variable to pick out key word descriptors in order to determine if there is there a relationship between the descriptor and the score. We visualized this relationship.
5. We calculated a weighted score for wines considering the number of points received and expense of the wine. We created a transformed linear model to observe this relationship on a graph.

Results

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  87.7     0.0392      2238.  0.
## 2 price         0.0321  0.000812      39.6 3.77e-311
```

predicted points = -440.214 + 5.351(price)

For every dollar increase in price, holding all other variables constant, we can expect that there will be a 0.03211381 increase in points.

```
## # A tibble: 3 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     88.1      0.0621   1418.   0.
## 2 top_countries    0.664    0.0705      9.42 5.56e-21
## 3 top_varieties    1.15     0.0717     16.0  1.63e-56
```

predicted points = 88.13 + 0.66(topcountry) + 1.145(topvariety)

This is a linear model that shows the predicted number of points for a particular wine given whether or not the wine is from a top country and whether or not the wine is from a top variety.

88.13 is the predicted point value of a wine assuming all other variables are 0. This means that if the wine was not in a top country and not in a top variety, this is the estimated point value.

While holding all other variables constant, a wine from one of the five top countries is expected to have a predicted point value that is 0.66 points higher than a wine from any other country.

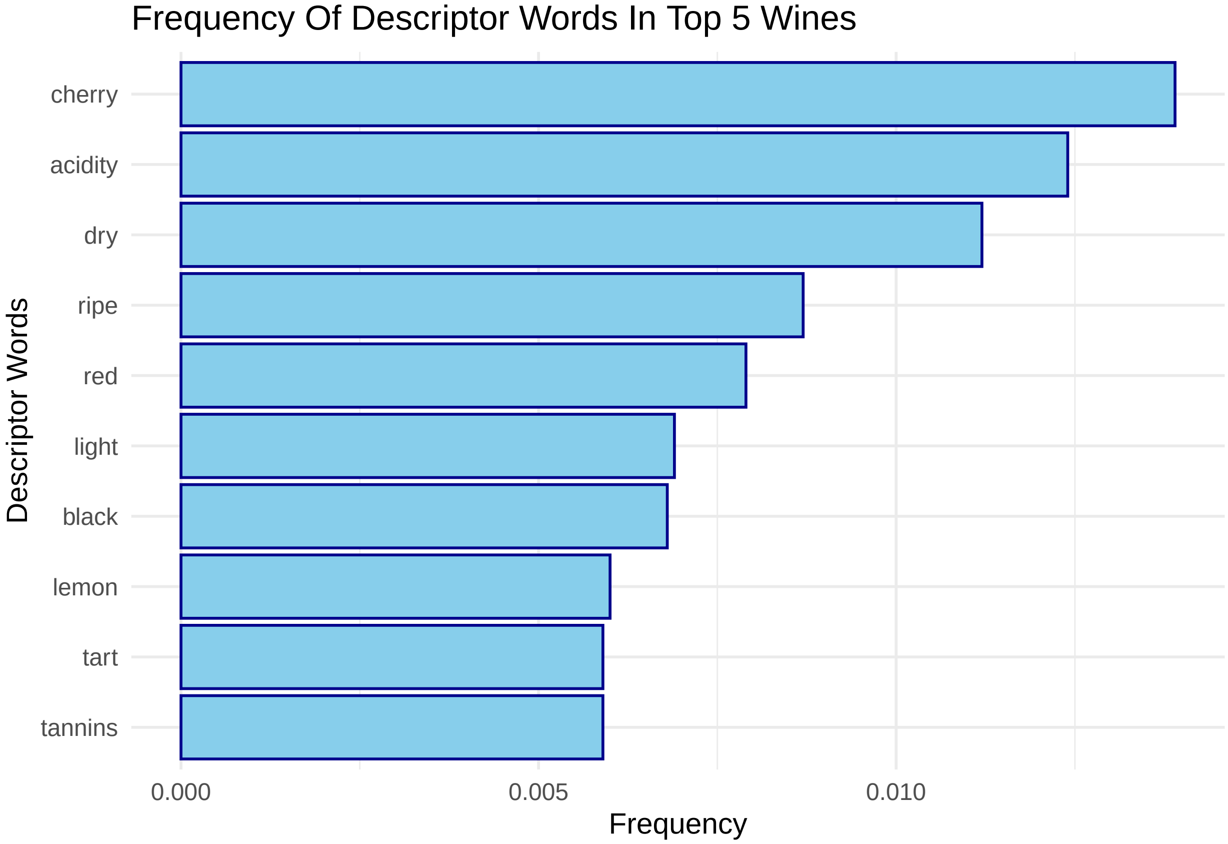
While holding all other variables constant, a wine from one of the five top varieties is expected to have a predicted point value that is 1.145 points higher than a wine from any other variety

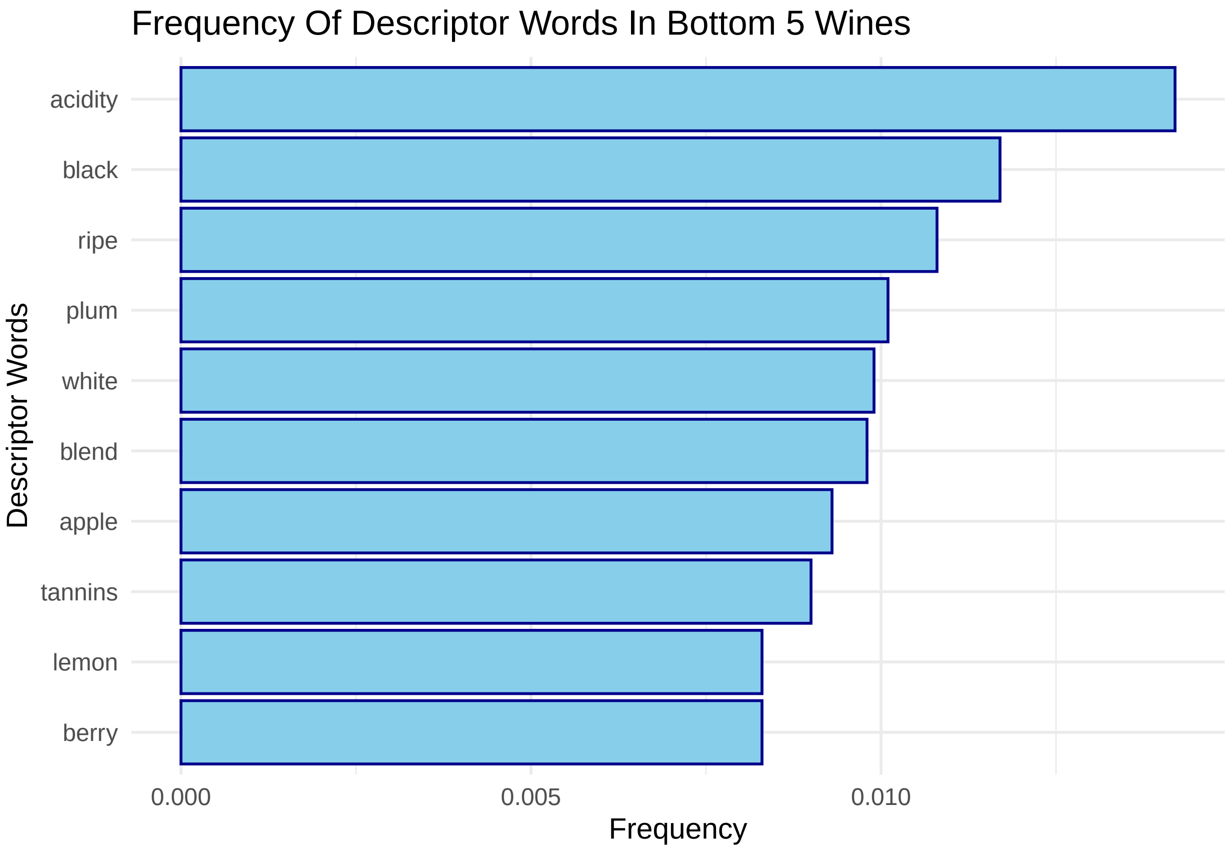
```
## 
##  One Sample t-test
## 
## data:  wine_data$points
## t = -72.288, df = 7999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 90.93939
## 95 percent confidence interval:
##  88.82544 88.93706
## sample estimates:
```

```
## mean of x
##  88.88125
```

H_0 = the true mean of the points is equal to 90.939 H_1 = the true mean of the points is NOT equal to 90.939 $\alpha = 0.05$ p value: 2.2e-16 decision: reject the null hypothesis

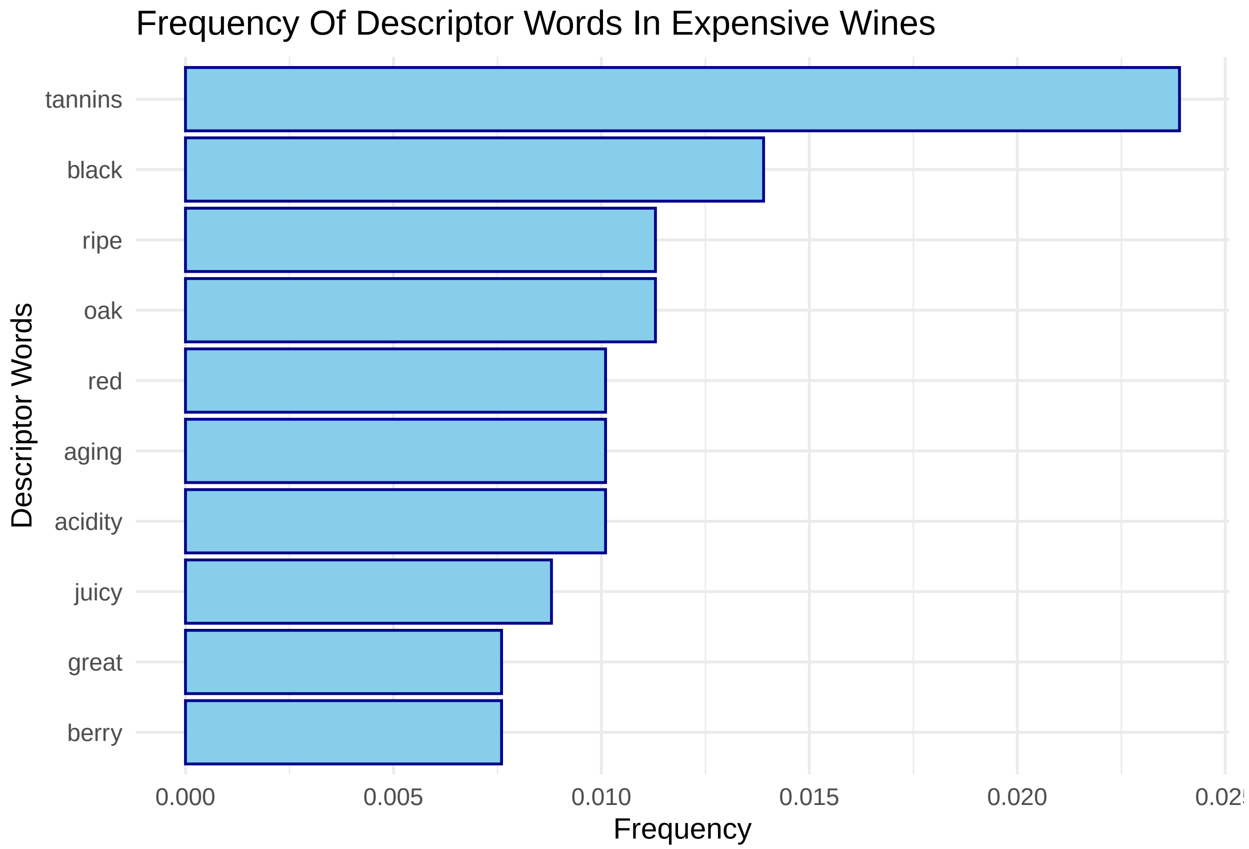
Here, we see that the overall mean of all wine scores is 88.88125 and has a 95% confidence interval or 88.82544 to 88.93706. This shows us that our observed average score of 90.79 for the top rated wine country (Austria) has statistical significance in its difference from the mean.



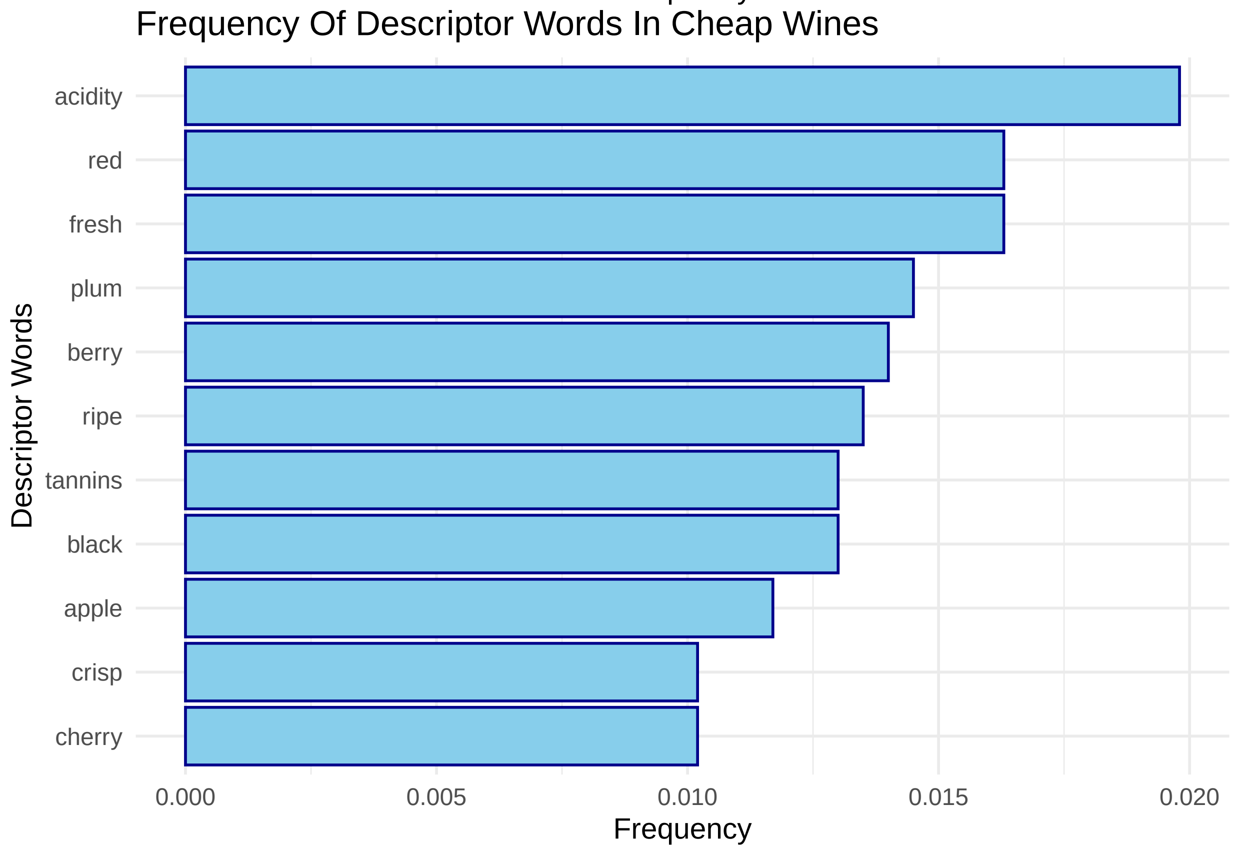


There are a surprising amount of similarities in the most commonly used descriptor words for the top and bottom 5 wine varieties. We see that tannins and lemon both fall near the bottom of the graph for both groups, but that the words acidity and ripe fall in the top 4 for both graphs as well.

On the other hand, we see some differences in the data. For example, we see that the word cherry is very commonly used in the descriptions of the top 5 wines (nearly 1.5% of all words used was the cherry!). We also see dry and red rounding out the top 5.



Frequency Of Descriptor Words In Expensive Wines

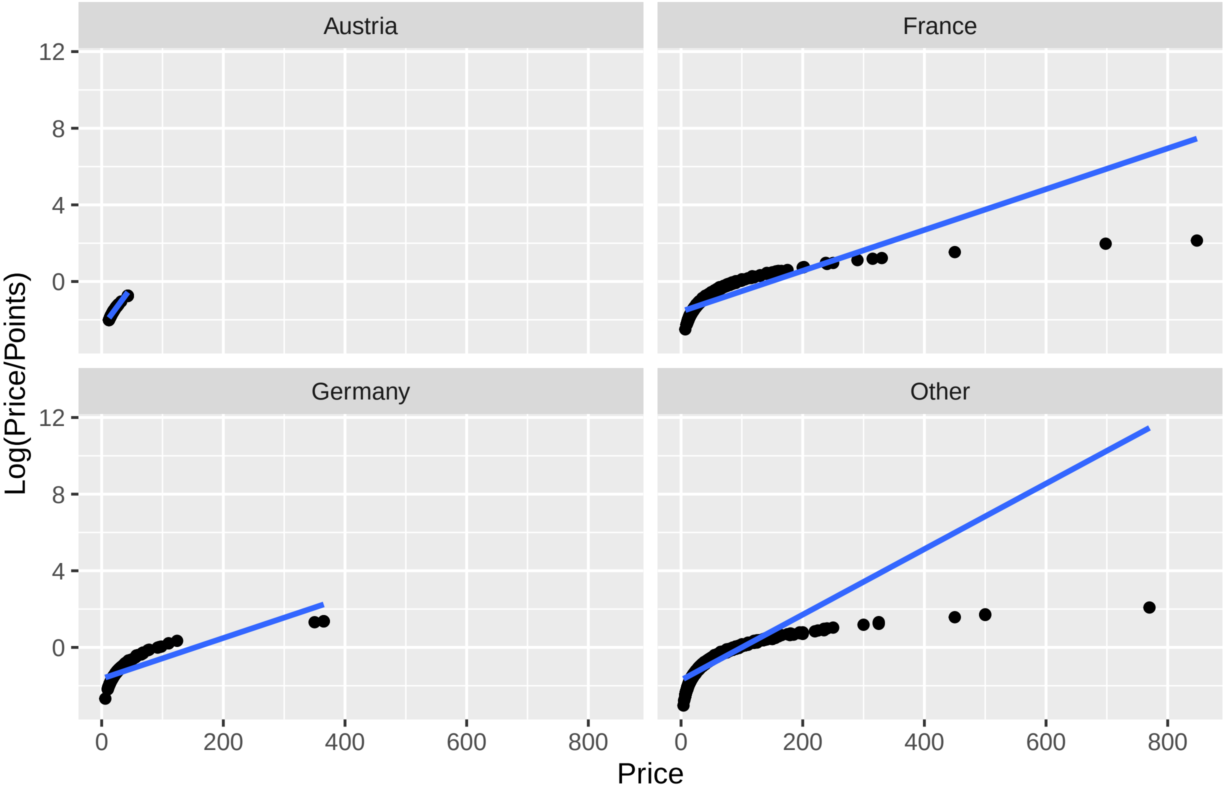


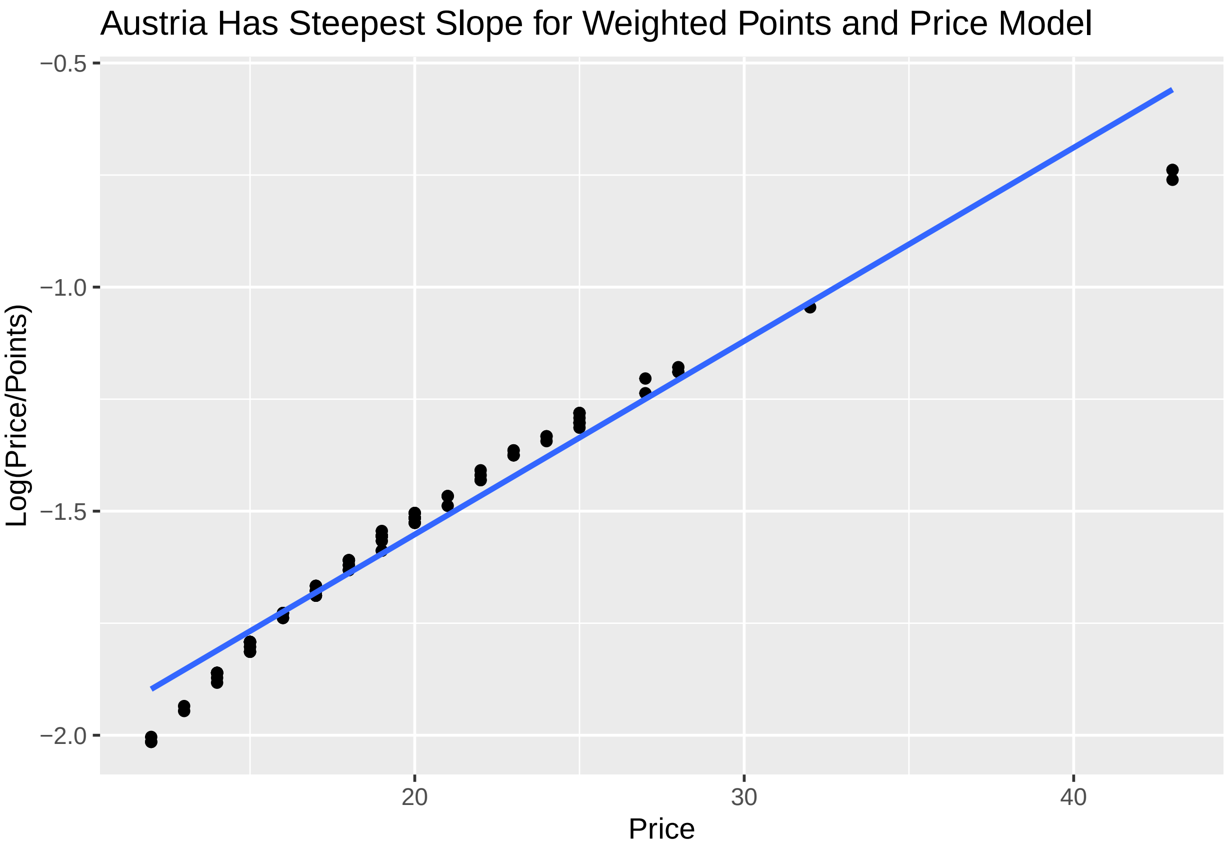
Frequency Of Descriptor Words In Cheap Wines

Here we are separating the wines by expensive (>= $200) and cheap (<= $10). Again, we see some immediate similarities between the two. For example, we see tannins, red, black, ripe, and berry falling at different positions for each throughout the graph. There don't seem to be any similarities as to where they appear in the lists (i.e. appears in the top 3 for both), but it is interesting to note that they are present in both. Acidity makes sense as to why it appears as it is present in all wines, although it is interesting that it is the most common descriptor for cheaper wines by a fair margin. Another thing to note is that both categories have the word red fairly often (presumably referencing red wine) and we know from our top wine varieties that red wine tends to be rated more highly. We expected the cheap wines to reference white more often, however this was not the case.

On the expensive side, we see the words oak, aging, and berry. Perhaps it is the case that these expensive wines are more likely to be stored and aged in higher quality barrels (oak may be the preferred substance) than cheaper ones. The presence of aging makes sense since there tends to be a correlation between older wines and higher prices. Berry is intriguing though. Recall that berry was one of the descriptors present in the bottom 5 wines from above. We find it very interesting that there is this correlation between expensive wines and lowly rated wines. Additionally, both black and ripe appear in the same places in the expensive chart as they do in the bottom 5 wine chart. While both of these words are also present in the cheaper wine charts, this is expected as we hypothesized that cheaper wines would correlate higher with lower rated wines. This leads us to the conclusion that more expensive wines do not necessarily imply higher scoring wines. There may even be correlations to the contrary.

Similarly, we see that cherry is 10th most commonly occurring word for the cheap wines and it is also the most commonly occurring word for the top 5 wines. This is a very interesting correlation and, when paired with our observations from the previous paragraph, make us begin to question whether paying more for wine is really worth it overall. There are certainly outliers on both sides, but it seems that there tends to be a correlation between cheap wine and highly rated wine, as well as one between expensive wine and lowly rated wine.


Positive Relationship Between Weighted Points Value and Price



Discussion

Our analyses showed that the highest rated wines are from Austria, a positive relationship between price and point ratings, and also some small differences in the words used to describe highly rated and poorly rated wines.

Top Countries and Varieties

This investigation gave specific evidence against our hypothesis that Italy and France would be the countries with the highest rated wines. Instead, both Austria and Germany had on average higher point ratings for their wines than these two countries. This information was especially surprising considering Italy and Spain were the top wine producers in 2019 and Forbes rated Napa Valley as the source of the best ranked wine in the same year. Because only “wine experts” gave data for the data set used in this investigation, we think that the point system more closely resembles a niche set of wine connoisseurs than the general public. This might be a reason for these differences. Moreover, we recognize that this list of top 5 countries may not be an accurate representation of where the best wines come from, but rather a representation of what location on average has the wines with the highest point ratings. This calculation rewards countries with one highly rated wine and disadvantages countries that have wines represented in the data set with very high point values and more very poorly rated wines. The density plots for price and points further convey Austria as a source of better wine. Austria clearly has a larger proportion of higher rated wines than the other top 5 countries yet at similar price ranges. Also, Austria’s wines were generally cheaper than the other countries as seen by the price density distribution graph. We added in an analysis of varieties when we were exploring the data set. The density plots for top varieties showed that the Gewürztraminer wine variety had a more evenly distributed range of wines scores and prices, but a higher wine score on average. This might mean that while choosing a Gewürztraminer could give you a very highly rated wine, you could also have a lower rated wine. Whereas the less evenly distributed Grüner Veltliner variety primarily has wines with scores of at least 90.

Relationship Between Price and Points

In addition to the calculation of the top countries and varieties, the linear models for predicted price both give evidence for a positive strong relationship between price and points. The linear model that predicted the number of points a wine would receive with a price factor had a p value of 3.77e-311 for the price factor. At an alpha level of 0.05, this gives strong evidence against a null hypothesis that there is no relationship between price and points. The linear model with top countries and top variety factors gave p values of 5.56e-21 and 1.63e-56, respectively. At an alpha level of 0.05, this gives strong evidence against a null hypothesis which is that there is no relationship between being in the top variety or country and the point value. This means that when choosing a wine to maximize points awarded by wine connoisseurs, it makes the most sense to choose a wine from one of these top countries and varieties. It should be noted that in the data the first linear model showed was not normally distributed and did not have linearity. The data for the second linear model was not normally distributed. This makes analysis using a linear model less effective. Our weighted model graphically depicts the trend that as points goes up, so does price. We used a log transformation on the graph because the data was very right skewed and this transformation best displayed the relationship between the weighted variable and the price in a linear fashion. Although Austria has a relatively low sample size, the line fitted to Austria's wine data has the steepest slope which is consistent with the conclusion that Austria has the best wine for its price.

Text Analysis

In our analysis of the text description, we found that we see that tannins and lemon both fall near the bottom of the graph for both groups, but that the words acidity and ripe fall in the top 4 for both graphs as well. We hypothesize this is the case because these are things that are present in most wines, regardless of score. Lemon is a bit of an outlier in the sense that it is not a flavor that is commonly thought of when considering wines. This could be indicative that our sample size may not be as representative as we initially believed. Where we begin to see differences however, are right at the top of the graphs. We see that the word cherry is very commonly used in the descriptions of the top 5 wines (nearly 1.5% of all words used was cherry!). We also see dry and red rounding out the top 5. On the other graph, we notice that black, plum, and white are very common descriptors as well as apple and berry rounding out their top 10. From this, we can hypothesize that a highly rated wine is more likely to be a dry red with hints of cherry. While a lowly rated wine is likely to be a fruity white wine, or potentially a very dark red (where we imagine the black descriptor comes from) again with fruity hints outside of cherries. When comparing words used to describe cheap wines and expensive wines, we found that words used to describe expensive wines were also also found in poorly rated wines. On the other hand, words used to describe cheap wines were also found in highly rated wines. This indicates to us that when selecting a wine, you should still consider price, but price is not the only factor that determines good wines. You should also take into account country, variety, and previous wine ratings. In order to test these hypotheses in a future analysis we could count these words and perform a CLT based test to determine if the differences in probabilities of words in the top 5 wines or worst 5 wines are statistically different. Furthermore, in the future it would be interesting to assign a either a positive or negative value to words used in the description of wines and perform a statistical test to quantify whether there is a significant difference in positive or negative words used to describe different categories of wine. We would be also interested in combining this wine rating data with data on wine purchases to compare how and if consumer decisions align with the trends we found in our data investigation.

References

[1] “Where Did Wine Come From? The True Origin of Wine.” Wine Folly, 19 Oct. 2020, winefolly.com/deep-dive/where-did-wine-come-from/.

[2] Conway, Published by Jan, and May 13. “Leading Countries in Global Wine Production, 2019.” Statista, 13 May 2020, www.statista.com/statistics/240638/wine-production-in-selected-countries-and-regions/.

[3] Huen, Eustacia. “The World's 30 Best Wines In 2019.” Forbes, Forbes Magazine, 26 Mar. 2019, www.forbes.com/sites/eustaciahuen/2019/03/25/wine-3/.

Appendix

In working on our text and sentiment analysis, we quickly realized that we had to do some filtering beyond just a standard stop word anti join. To do this, we would run our code to retrieve the top 20 most common words that appeared in the descriptions for our different categories, and then analyze them to see if they all made sense. After repeating this many times, we felt that we had a representative sample of descriptors that were significant. Some of them were removed based off of the specific filter. For example, if we were filtering for Pinot Noirs, we didn't need to see the word Pinot in our top 20 descriptors. Additionally, we chose words like generous that didn't seem to tell us much and other words like wine and flavors that can reasonably be expected to be present in every description. There was certainly a bit of arbitrary choice involved here, but we felt all of the words taken out were not relevant to the comparisons we were making. Below you will find a list of the words we removed.

wine, flavors, palate, drink, aromas, offers, finish, notes, long, generous, hints, nose, fresh, fruit, full, character, ready, 100, merlot, texture, end, medium, bodied, malbec, portuguese, chardonnay, blanc, viognier, note, show, vineyard, style, shows, make, pinot, sauvignon, body, bottling, alongside, dense, dried, concentrated, years, vineyard, structure, age, fruits