SDS_230_Final Project (5/8/2019)

Team: Isabella Teng, Gina Zhu, Josefina Mendez, Hersh Gupta

Introduction As wine lovers, it can be difficult choosing the best wine. Wine can come from a variety of countries and can take on drastically different prices. In this project, we seek to evaluate and determine how to identify the best wine. With this goal in mind, we used wine review data from Kaggle collected in 2017 and examined several questions that a typical consumer may be interested in when shopping for wine. The data covers approximately 13000 reviews of wine from 43 total countries, and each wine review contains information such as country of origin, price, and a rating on a scale of 1-100. We were interested in using this Wine Review Data to better understand:

```
  1. Which countries are the major producers of wine and what sorts of wine
have the highest ratings in these countries?
  2. Is there any difference in mean ratings of wine produced from different
countries?
  3. Is there a correlation between price of the wine and its rating, and is
this related to country of origin?
  4. Can we estimate true wine prices for each different "category" of wine
ratings? And does price increase with rating cateogry as       expected?
  5. Can we identify what are the predictors for wine ratings?
  6. Is there any difference in ratings from wineries from specific regions
within a country?
```

Further, we broadened our scope and completed some Web Data scraping from Wikipedia to find out more information about alcohol consumption in the world – and more specifically, how much consumers in each country actually choose to drink wine over other forms of alcohol. We asked:

```
  1. What is the percentage of consumption of wine out of (Wine, Beer,
Spirits) for each country?
  2. What is the layout of alcohol consumption in the country that is most
reviewed by Kaggle?
```

The data set in use contains approximately 130k wine reviews with accompanying information such as variety, location, winery, price, and description of the wine. This data set can be accessed at: **LINK**

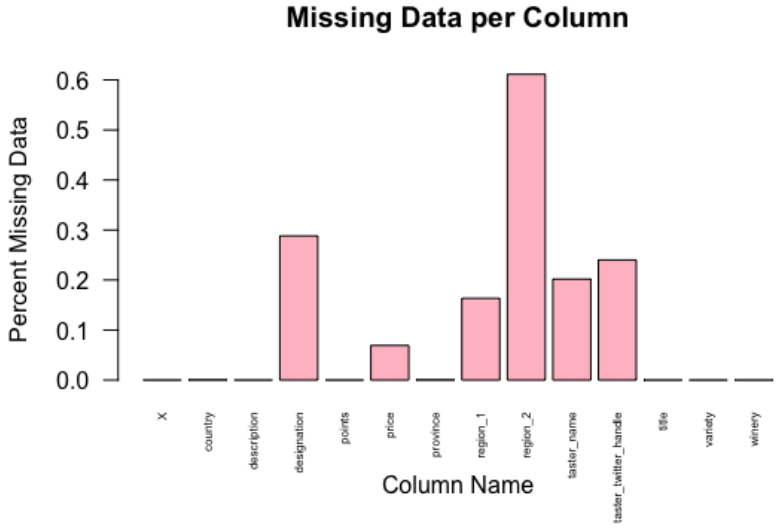
Complete List of Variables

X (*For 0-indexing*), Country (*The country where the wine is from*), Description (*Characteristic description of the wine including flavor, aroma, etc.*), Designation (*The vineyard within the winery where the grapes that made the wine are from.*), Points (*This is the methadology of rating the wine. The number of points WineEnthusiast rated the wine on a scale of 1-100*), 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 (*Some wine include an even more specific region specified within a wine growing area (ie Rutherford inside the Napa Valley), but this value can also be blank*), Taster_name (*The name of the wine reviewer*), Taster_twitter_handle (*The twitter handle of the wine reviewer*), Title (*Title of the wine review*), Variety (*The type of grapes that were used to make the wine (ex. Pinot Noir)*), Winery (*The winery that produced the wine*).

DATA CLEANING PROCESS We employed multiple techniques to produce a managable and clean dataframe, including dealing with missing data, reducing dimensionality of the original dataset, and adding numerical columns that would help us in our data analyses. To better understand our data, we plotted the percent of missing data by column and removed the columns: regions_2, X, and taster_twitter_handle. Then, since price and country were the main features we wanted to investigate, we removed rows with missing data for those columns. Another feature we cleaned was wine variety. Upon investigating, we found that there were thousands of wine varieties and chose to remove ones that were reviewed less than 2000 times. Similarly, we dropped countries as well that had less than 2000 wine reviews. Two columns we added to our dataframe were ratings_num and ratings. Following Wine Magazine's chart that gave categorical rankings to the points assigned to wines, we decided to add both a categorical and numerical ranking column so that we'd have more features to analyze. We also added an additional continuous variable of winePercent (percent of wine consumption) acquired from Web Data Scraping.

Issues We Encountered: At first we used `na.omit` to remove rows where price or country was missing, but this removed more than half of the rows, which raised concern. Switching to `complete.cases` worked instead. Other considerations and decisions we had to make was with dropping rows versus renaming values. With wine variety, we originally created a new category "Other" for varieties reviewed less than 2000 times, but this made the 'Other' category the greatest, so instead we dropped these rows, as we were more interested in the most popular wine varieties. Another concern we encountered was with originally combining the Superb and Classic ratings as there were very few Classic wines; however, this increased the interval of this last category as compared to the others, later affecting bootstrap results, so we decided to leave it as it is.



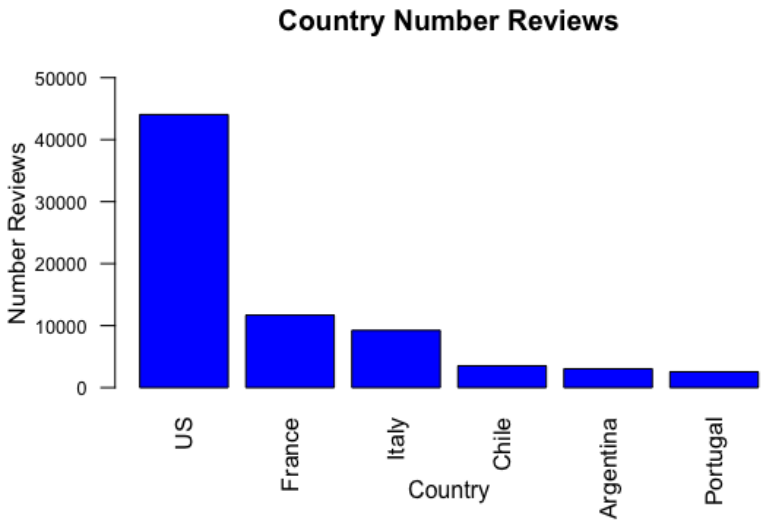
Add a column mapping scores to categorical ratings according to: SITE Acceptable (0), Good(1), Very Good (2), Excellent (3), Superb (4), Classic (5)

```
##
## Acceptable    Classic  Excellent       Good     Superb  Very Good
##       2847        116      39713      29624       5544      43072

##
##     0     1     2     3     4     5
##  2847 29624 43072 39713  5544   116
```

Cleaning Wine Varieties Let's drop the wine grape varieties that have been reviewed less than 2000 times

```
##
##               Pinot Noir                 Chardonnay       Cabernet Sauvignon
##                    12785                      11077                     9384
##                Red Blend Bordeaux-style Red Blend                  Riesling
##                     8466                       5340                     4971
```

Let's remove countries that have less than 2000 reviewed wines

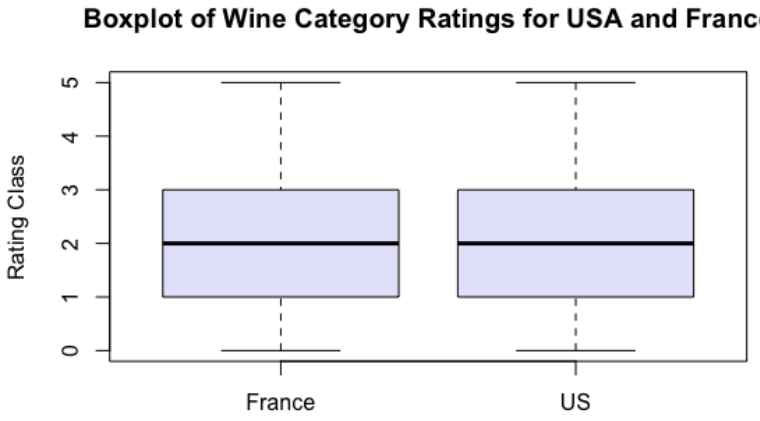


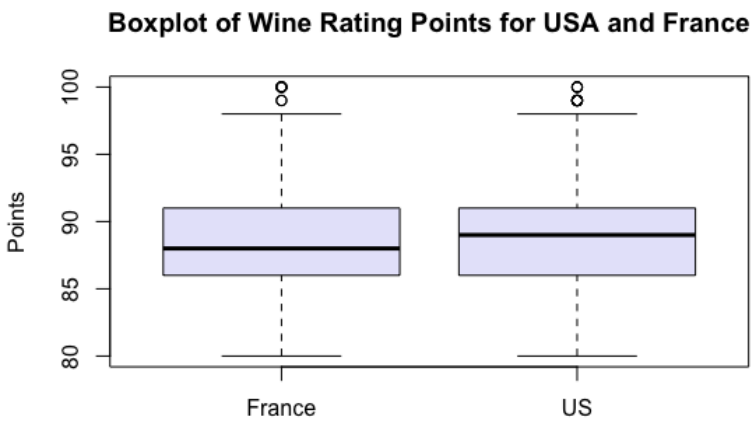
DATA SCRAPING FOR CONTINUOUS VARIABLES We wanted to add more continuous variables for our data set so we decided to scrape data on alcohol consumption per capital from the site: SITE. This Wikipedia article contains data from countries all over the world, and tracks the % of each type of alcohol is consumed in that country out of all types of alcohol. The types of alcohol the website tracked were: Beer, Spirits, Wine, Other.

```
#Defining the URL of interest
urlWine <-
"https://en.wikipedia.org/wiki/List_of_countries_by_alcohol_consumption_per_capita"

#Reading the HTML code from the website into a new object
wineConsumption <- read_html(urlWine)
```

```r
#Reading in Country information and cleaning
country <- gsub("^\\s+", "", html_text(html_nodes(wineConsumption, ":nth-
child(9) tr :nth-child(2)")))
country <- unique(country)
country <- gsub("United States", "US", country)

#Reading in Percent Wine consumption data and cleaning
winePercent <- html_text(html_nodes(wineConsumption, "tr :nth-child(7)"))
winePercent <- as.numeric(winePercent[c(1:192)])

## Warning: NAs introduced by coercion

winePercentData<- data.frame(country, winePercent)[-c(1),]

#Merging into our larger data set
wineNew<- merge.data.frame(wine, winePercentData, by = 'country', all = TRUE,
sort = FALSE)
wineNew <- left_join(wine, winePercentData)

## Joining, by = "country"

## Warning: Column `country` joining factors with different levels, coercing
## to character vector
```

DESCRIPTIVE PLOTS, SUMMARY INFORMATION, AND ANALYSIS T-Test We wanted to see if there was a difference between the mean ratings for wines produced in the US and France - the two most reviewed countries. We first decided to look at the difference in numerical ranking categories between the US and French wines, where wines are assigned a number from 0-5, which each number corresponding to a categorical rating (acceptable, good, etc.) based on the number of rating points it received. We also decided to look at the difference in the ratings based on raw points assigned to each wine where they were assigned a number from 0-100, with a higher number corresponding to a better wine. We decided to use a two-sample t-test to test the difference. Our null hypothesis is that the difference in mean numerical ratings from wines produced in the US and France is 0, and our alternative hypothesis is that the difference is not zero. We chose and alpha of 0.05 to test the significance.





T-Test by rating_num

```
## wineFRA_USA$country: France
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.188   3.000   5.000
## ------------------------------------------------------------
## wineFRA_USA$country: US
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.207   3.000   5.000

##
##  Welch Two Sample t-test
##
## data:  rating_num by country
## t = -1.9525, df = 18781, p-value = 0.05089
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0378215203  0.0000731035
## sample estimates:
## mean in group France     mean in group US
##             2.188387             2.207261
```

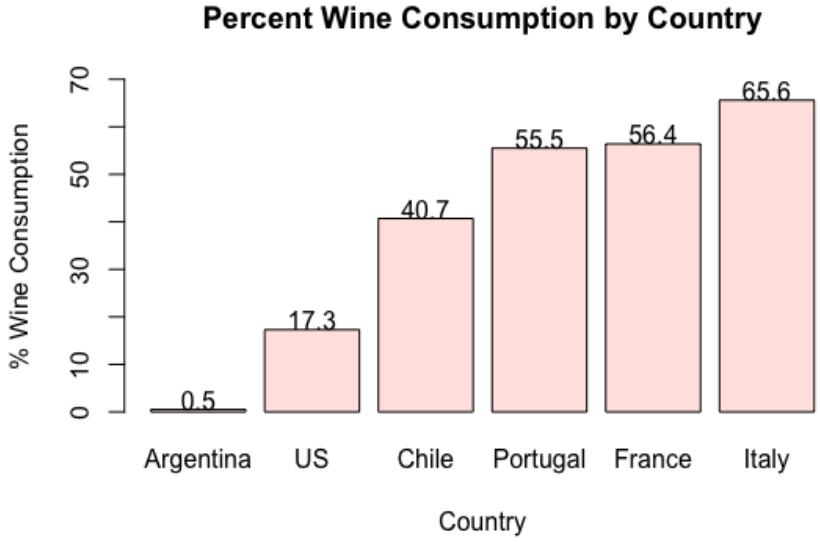
T-Test by points

```
## wineFRA_USA$country: France
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   80.00   86.00   88.00   88.65   91.00  100.00
## ------------------------------------------------------------
## wineFRA_USA$country: US
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   80.00   86.00   89.00   88.68   91.00  100.00

##
##  Welch Two Sample t-test
##
## data:  points by country
## t = -0.95133, df = 18787, p-value = 0.3414
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.09398505  0.03256425
## sample estimates:
## mean in group France     mean in group US
##             88.64845             88.67916
```

We conducted a Welch Two Sample t-test to see if there was a difference in the mean wine ratings between wines produced in France vs. the USA. Looking at the difference in means of numerical rating categories (0-5), our p-value = 0.05, which is equal to our alpha of 0.05, meaning the difference is not statistically significant. Our 95% confidence interval for the difference in means is from - 0.04 to 0, which includes 0 on the upper limit. Thus, we fail to reject our null hypothesis that the difference in mean numerical category ratings for US and

French wines is 0. Looking at the difference in means of rating points, our p-value was 0.3414, which is much greater than our alpha of 0.05. Our 95% confidence interval for the difference in means is from - 0.09 to 0.03, which very clearly includes 0. Thus, we again fail to reject our null hypothesis that the difference in mean rating in points for US and French wines is 0.

We were now interested in getting a better picture of wine consumption out of ***all alcohol types*** in these countries; data scraped from Wikapedia.



According to the Kaggle data, wines originiating from the U.S. are the most heavily reviewed. However, even though the U.S. seems to produce a higher quantity of wine, consumers in the U.S. only choose to drink wine 17.3% of the time over other types of alcohol. On the other hand, wine orginiating from Portugal had relatively low number of reviews in the Kaggle data, but wine consumption in this country is 55% out of all alcohol choices. This seems to suggest that there is not a perfect relation between a country producing "reviewable" wine, and the country's citizens consuming wine in general.

It's interesting that the the majoriy of Kaggle Reviews are wines from the United States, yet wine does not seem to be the most popular alcohol of choice for US residents. So what is?

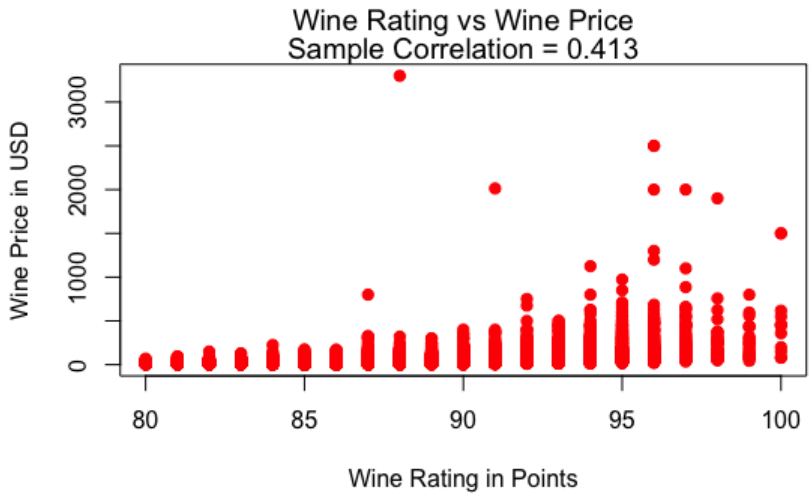
The most consumed type of alcohol in the United States is Beer, followed by Spirits (which is defined as all distilled beverages such as vodka), with wine being the least popular. Therefore,

while the majority of the wine reviewed in the Kaggle data set are produced in the United States, US citizens prefer other forms of alcohol more.

Correlation Tests We wanted to see if there was any correlation between the price and rating of wines. Are higher rated wines really more expensive as one might initially think?

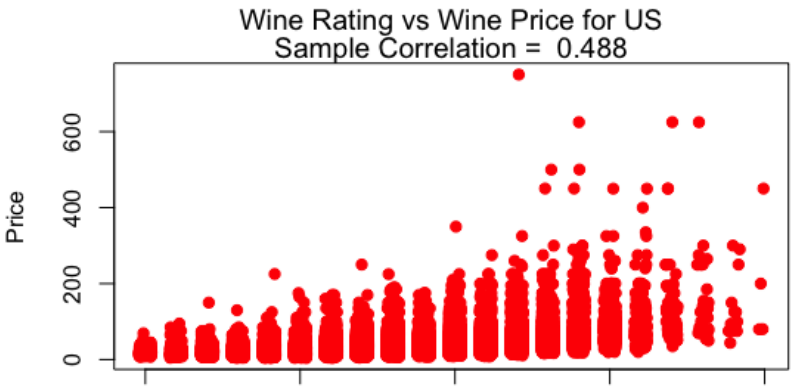
Correlation between price and rating of wines:

```
## [1] 0.4128493
```

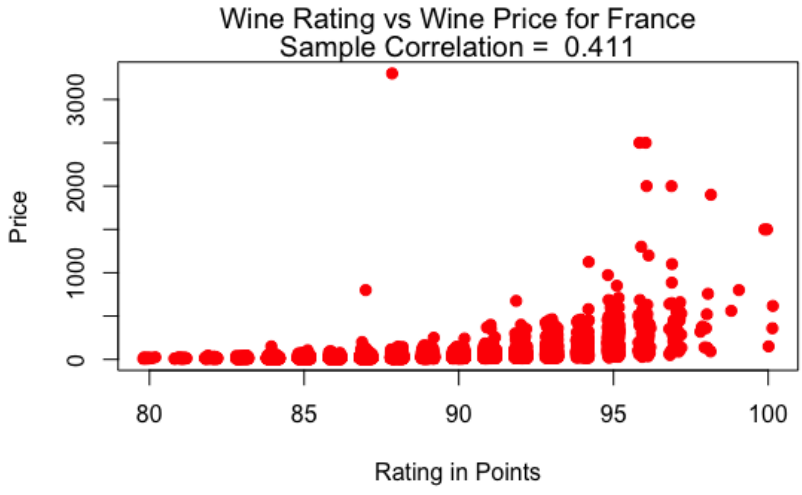


It seems that there is a moderate positive correlation (0.413) between the Wine Rating in Points and Price. Wines with higher ratings generally will be more expensive.

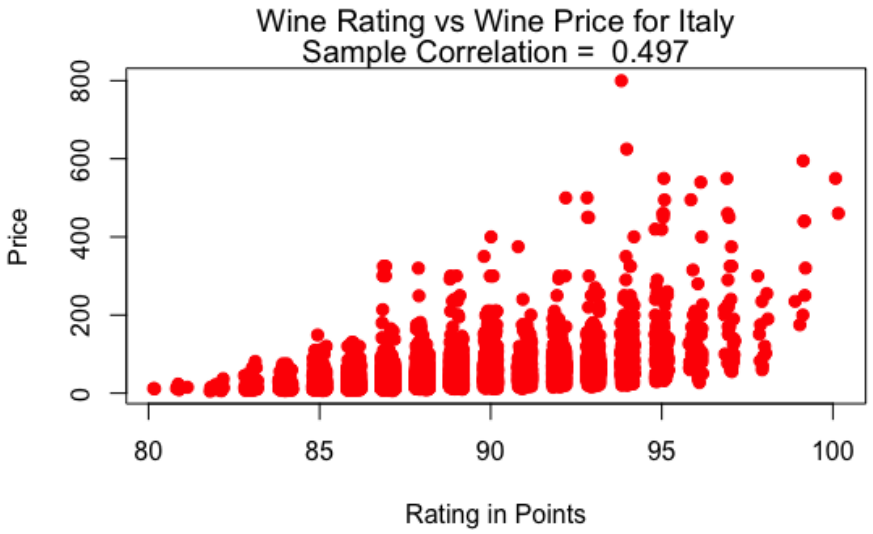
We were interested in breaking down this correlation more. Does this correlation between price and rating of wines change depending on country? We decided to look at the top 3 wine producing countries in this data set, which were the US, France, and Italy.




```
## [1] 0.4880604
```




```
## [1] 0.4107404
```

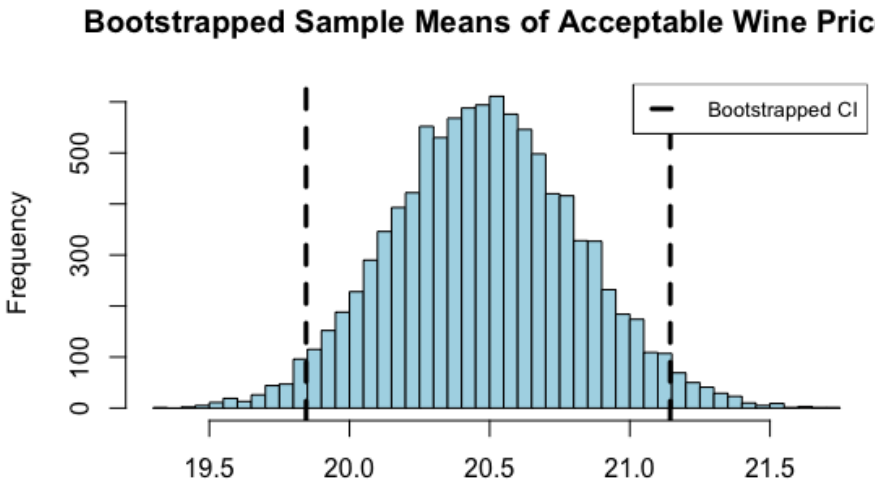



```
## [1] 0.4968985
```

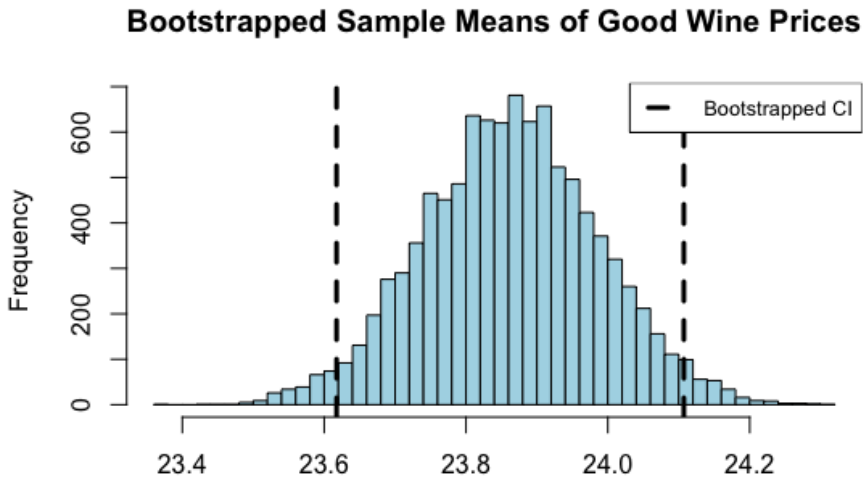
The top 3 wine producing countries in this dataset all have moderate positive correlations around the value of 0.4 - 0.5. All countries except France have slightly stronger positive correlations than the overall correlation we did for price vs. rating when considering all countries. So generally for any country, the higher the wine rating in points, the more expensive the wine is.

Bootstraped Means of Prices for Different Wine Ranking Categories We were interested in estimating the true wine prices for each of the score categories - Acceptable, Good, Very Good, Excellent, Superb, and Classic - by conducting a bootstrap. We wanted to see if there were differences in the estimated true mean and 95% confidence intervals for the prices of each category while taking into account any underlying variation in the data. We did this with the data from the US, France, and Italy - the 3 countries with the greatest number of wines reviewed. We would like to see if the estimate of the true mean price increases with the rating category of the wine as we would expect.

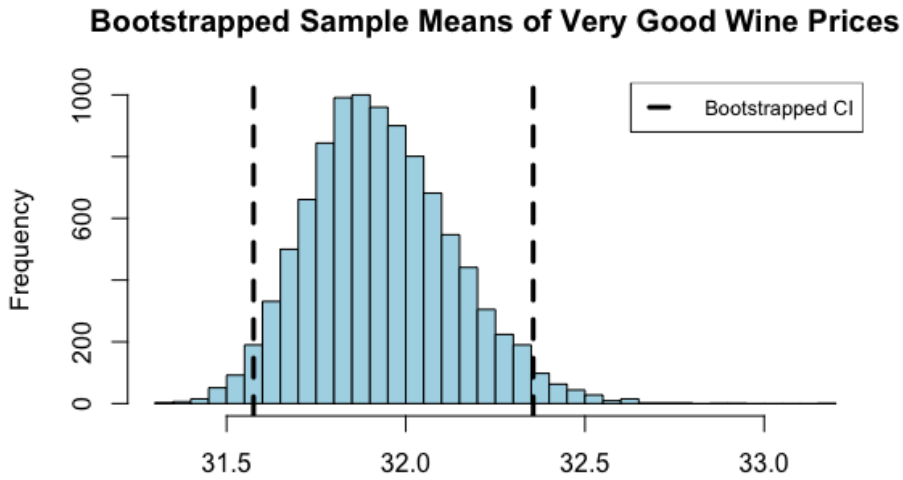
```
##
## Acceptable    Classic  Excellent       Good     Superb  Very Good
##       1331         87      23398      14257       3881      22016
```



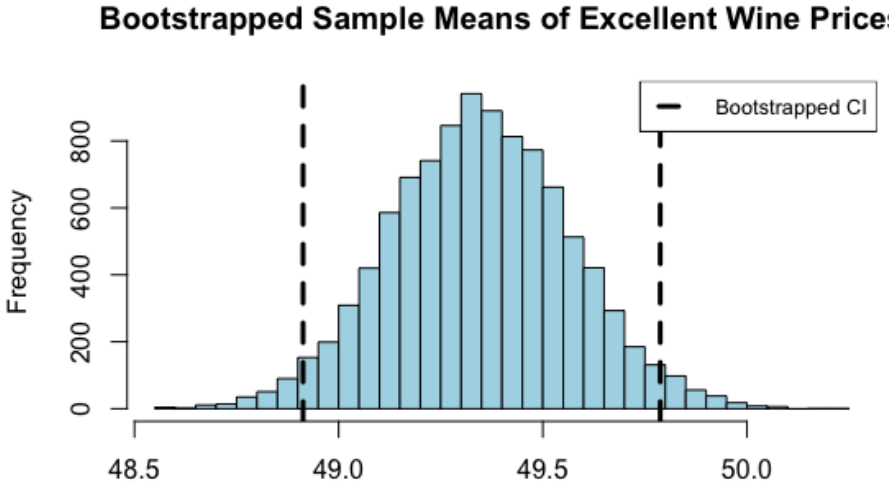
```
##     Min.
##    19.85
```



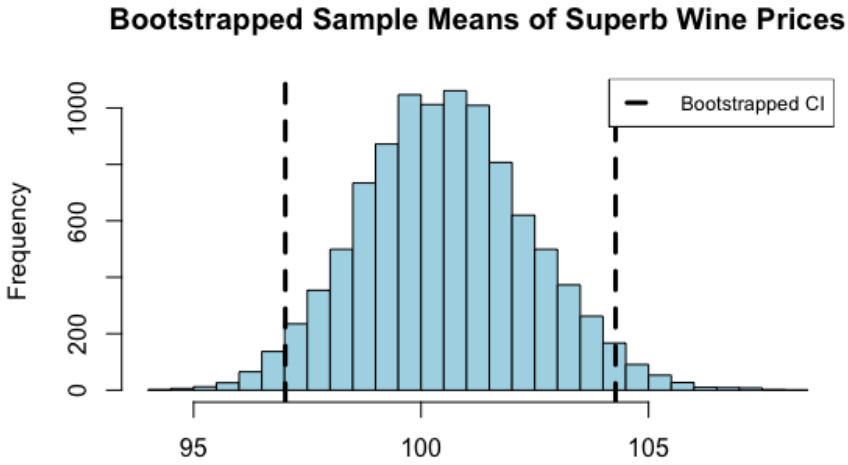

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.62   23.74   23.86   23.86   23.98   24.11
```



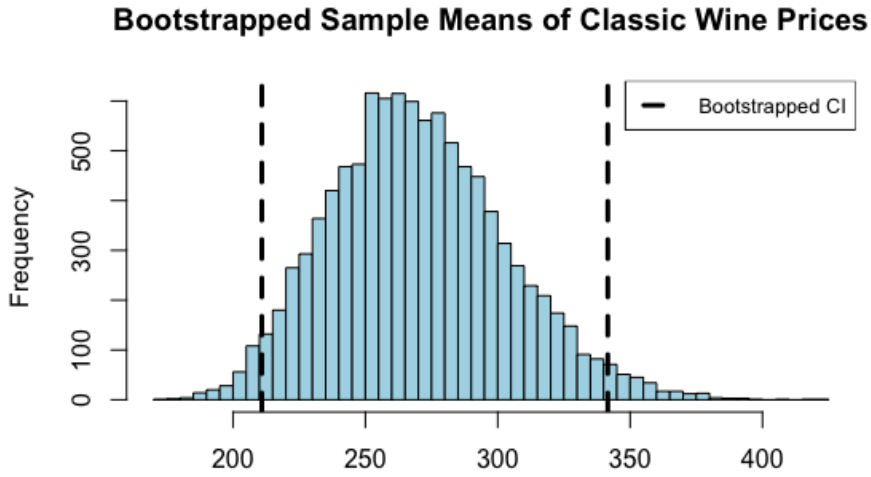

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   31.58   31.77   31.97   31.97   32.16   32.36
```



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   48.91   49.13   49.35   49.35   49.57   49.79
```




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   97.02   98.83  100.65  100.65  102.46  104.27
```



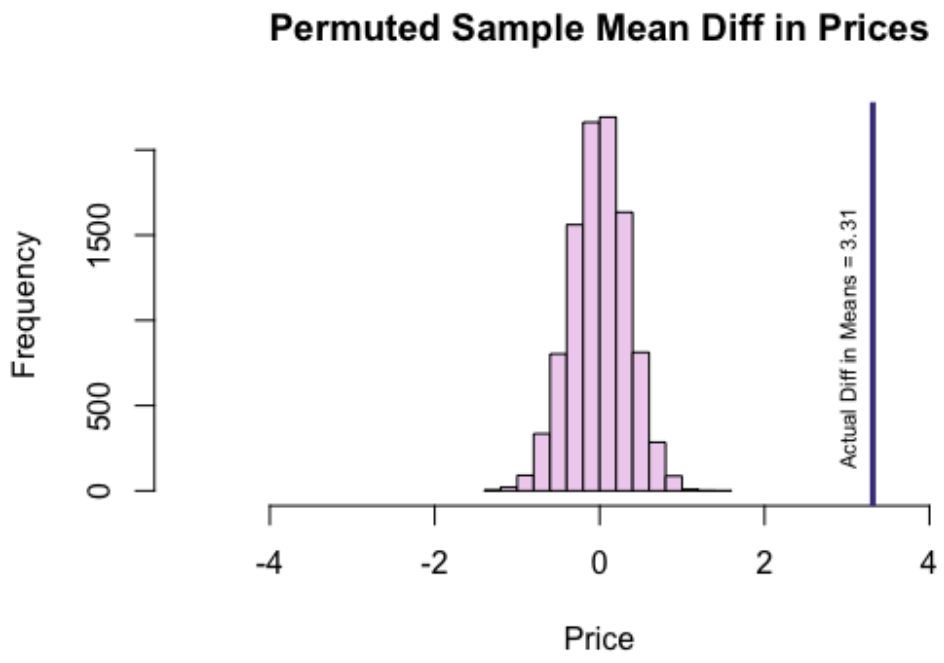

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   210.9   243.6   276.2   276.2   308.9   341.6

##   Rating.Category Mean.Price X95.Pct.Confidence.Interval
## 1      Acceptable      20.66                20.38 - 20.94
## 2            Good      22.91                  22.81-23.01
## 3       Very Good      30.29                30.15 - 30.44
## 4       Excellent      47.21                47.02 - 47.39
## 5          Superb      99.33               97.67 - 100.99
## 6         Classic     278.40                248.4 - 308.1
```

Constructing bootstrap confidence intervals for each of the score categories gave us an idea of how the true mean prices differed from category to category. The results followed our initial hypothesis that the estimated true mean price would increase with a higher rating category for the wine. None of the bootstrapped 95% confidence intervals for each rating category overlap, signaling that the true mean prices are indeed different for each rating category. It is also seen that the mean price for the wines increases as the rating gets better.

Permutation Test for Mean of Prices for Acceptable and Good While the bootstrap confidence intervals for estimates of the true means above do not seem to overlap, we wanted to more rigorously evaluate whether there was a statistically significant difference in the means for groups that appeared most similar in their mean prices. We found above that the confidence intervals are most similar for prices in the "Acceptable" and "Good" rating categories. To compare whether "Good" wines do in fact have higher prices on average than "Acceptable" wines, we construct and evaluate the following hypothesis test:

$$H_0: \mu_{Good} - \mu_{Acceptable} = 0$$

$$H_a: \mu_{Good} - \mu_{Acceptable} \neq 0$$



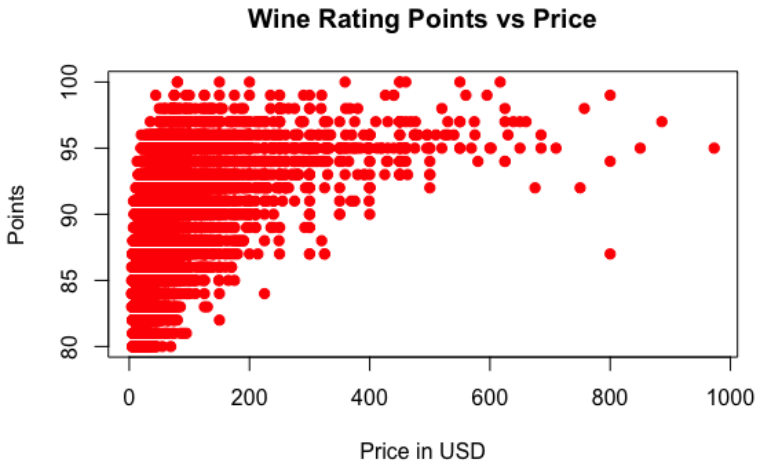
Two-sided p value:

```
## [1] 0

##
##  Welch Two Sample t-test
##
## data:  price by rating
## t = -11.606, df = 2458.5, p-value < 2.2e-16
```

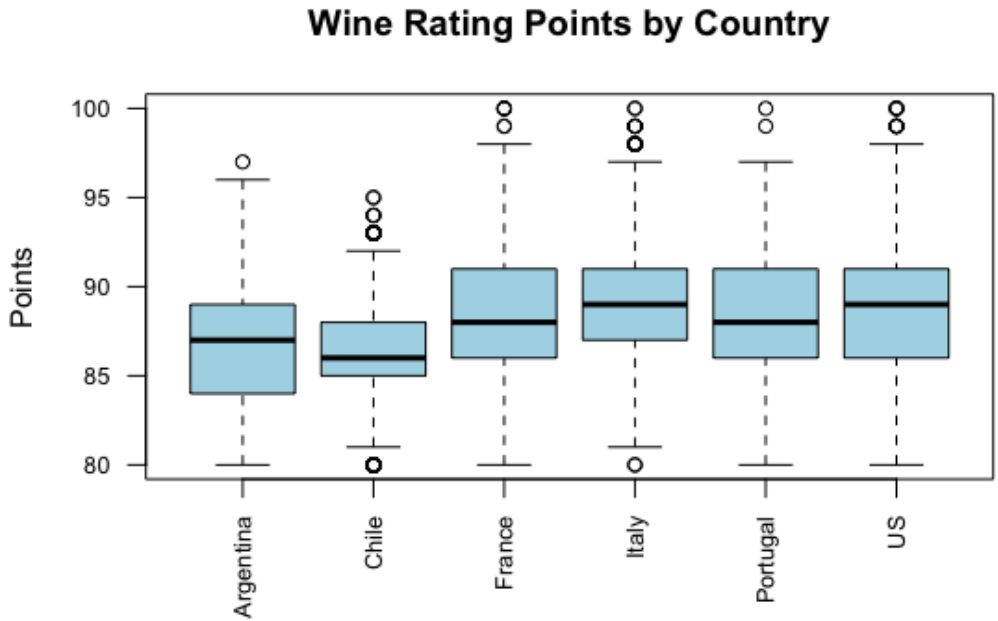
```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.873565 -2.753816
## sample estimates:
## mean in group Acceptable       mean in group Good
##                 18.57455                 21.88824
```

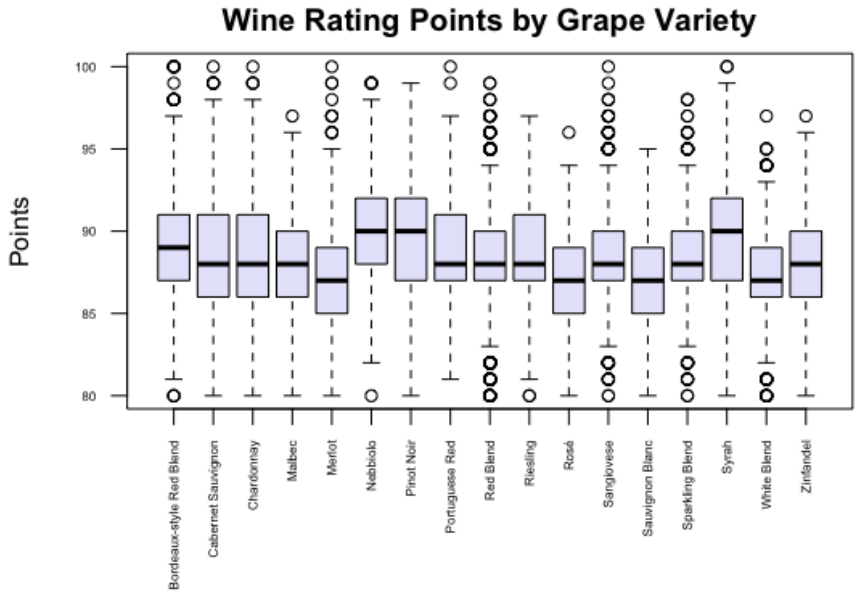
The estimated two-sided p-value for the difference in mean is 0. From computing the binomial estimate of the p-value and a confidence interval, we note that estimated p-value is exactly 0 with the 95% confidence interval from 0 to 0.0003688199. Comparing with a t-test that reports the p-value as < 0.00000000000000022, we can safely conclude our actual perm test p-value is extremely small and <<< 0.05, hence we reject our null hypothesis that there is no difference between the "Acceptable" and "Good" means, and conclude that there is a statistically significant difference among prices between the two groups.

Multiple Regression for Predictors of Wine Rating We were interested in what were the factors that could significantly predict the rating of a wine. Is the country where the wine is produced a good predictor? The price? Or perhaps grape variety? We set out to investigate this doing a multiple regression for predictors of wine rating.




```
##
##  Pearson's product-moment
correlation
##
## data:  winePred$price and
winePred$points
## t = 150.82, df = 74064, p-value
< 2.2e-16
## alternative hypothesis: true
correlation is not equal to 0
## 95 percent confidence interval:
##  0.4792001 0.4902195
## sample estimates:
##       cor
## 0.4847291
```





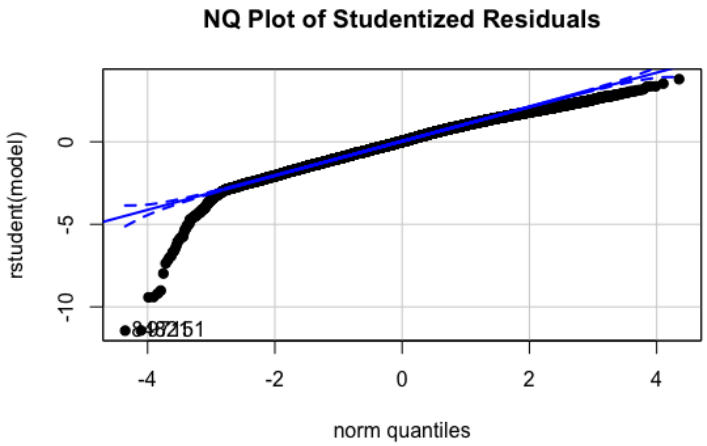
There is a moderate positive correlation (0.48) between the price and rating points of a wine, where the higher the price of the wine, generally the higher the rating in points. This is observed in the scatterplot and the correlation coefficient. From the boxplots, we can see that the mean wine rating does seem to vary by country and by grape variety, but it's not clear how significant these differences are just from looking at the plots. Thus, we went ahead and tried to fit a model to see if the wine's country of origin, price of wine, or grape variety were good predictors of its rating.

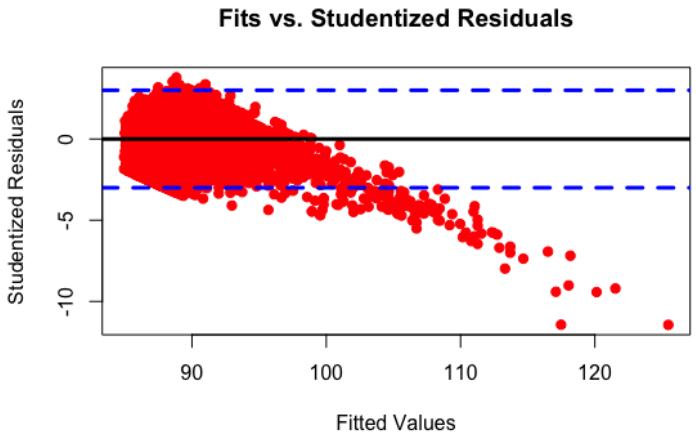
```
## Call:
## lm(formula = points ~ country + price + variety, data = winePred)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.4815  -1.7845   0.0629   1.9717  10.1610
##
## Coefficients:
##                             Estimate Std. Error  t value Pr(>|t|)
## (Intercept)               85.5406402  0.0701720 1219.014  < 2e-16 ***
## countryChile               0.1739923  0.0725845    2.397   0.0165 *
## countryFrance              1.4810145  0.0637223   23.242  < 2e-16 ***
## countryItaly               1.8160485  0.0719424   25.243  < 2e-16 ***
## countryPortugal            0.2780422  0.1561025    1.781   0.0749 .
## countryUS                  1.5872313  0.0591369   26.840  < 2e-16 ***
## price                      0.0389506  0.0002869  135.765  < 2e-16 ***
## varietyCabernet Sauvignon -0.2003066  0.0510150   -3.926 8.63e-05 ***
## varietyChardonnay         -0.0027263  0.0476296   -0.057   0.9544
## varietyMalbec              0.6984233  0.0734614    9.507  < 2e-16 ***
## varietyMerlot             -0.9110668  0.0651364  -13.987  < 2e-16 ***
## varietyNebbiolo            0.4341986  0.0816380    5.319 1.05e-07 ***
## varietyPinot Noir          0.5521442  0.0474119   11.646  < 2e-16 ***
## varietyPortuguese Red      2.0788924  0.1603856   12.962  < 2e-16 ***
## varietyRed Blend          -0.1098346  0.0557598   -1.970   0.0489 *
```

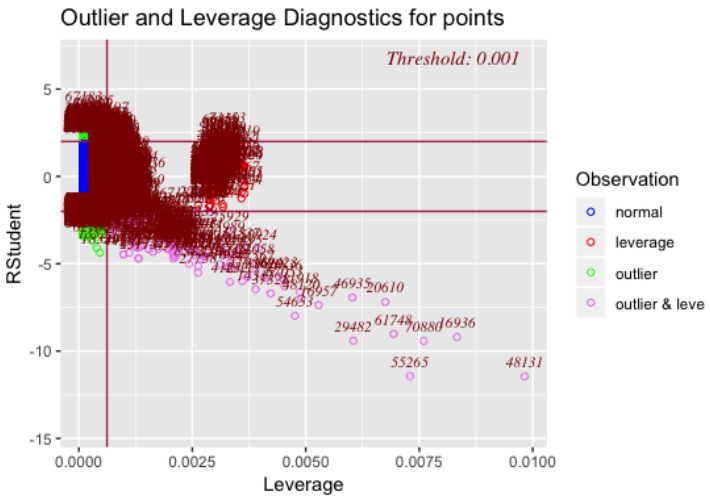
```
## varietyRiesling              0.6783382  0.0670542   10.116  < 2e-16 ***
## varietyRosé                 -0.7006230  0.0632825  -11.071  < 2e-16 ***
## varietySangiovese           -0.4666677  0.0776856   -6.007 1.90e-09 ***
## varietySauvignon Blanc      -0.2800654  0.0593329   -4.720 2.36e-06 ***
## varietySparkling Blend       0.1141423  0.0832849    1.371   0.1705
## varietySyrah                 0.7581467  0.0597986   12.678  < 2e-16 ***
## varietyWhite Blend          -0.6703146  0.0806693   -8.309  < 2e-16 ***
## varietyZinfandel            -0.4503719  0.0667279   -6.749 1.50e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.679 on 74043 degrees of freedom
##   (13 observations deleted due to missingness)
## Multiple R-squared:  0.274,  Adjusted R-squared:  0.2738
## F-statistic:  1270 on 22 and 74043 DF,  p-value: < 2.2e-16
```

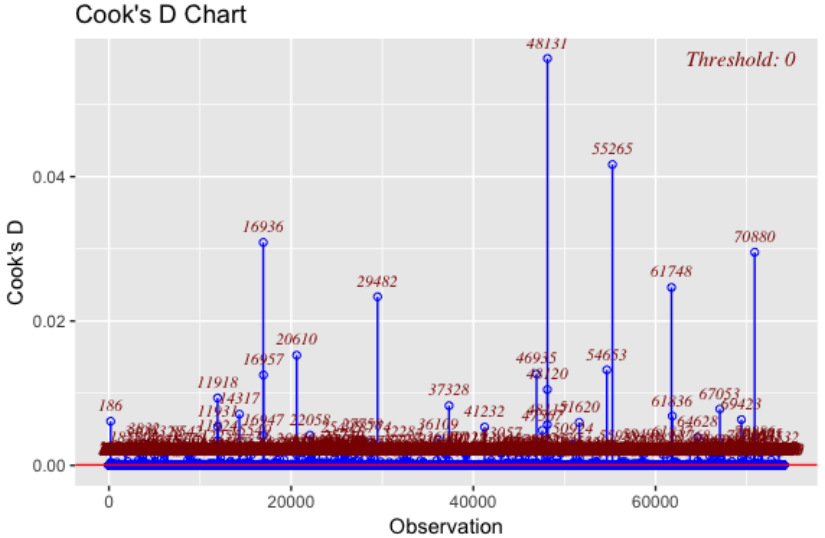
We can see that many of the coefficients in this model are significant, though which is the best predictor is not clear. To find the best predictor, we will do a best subset regression.

Based on the best subset regression, if we were to choose one predictor for the wine rating by points, we would choose price. But before we can claim that this predictor, or any of the predictors are significant, we must check our assumptions and look at our residual plots.









Looking at the residual plots, the data does not seem to be normally distributed. On the plot of studentized residuals, there is a tail at the left end of the plot that strongly deviates from the line. This makes me think there is almost another "group" of points that are significantly different from most of the data in someway. It does not look like a transformation will fix this and normalize the data, so we will take a deeper look at the data and just move forward with this in mind.

We decided to look at the points that had residuals less than -4, for that is when the tail of the plot of studentized residuals dropped away from the rest of the points. We tried looking for any trend with these observations and it seems that most of the wines in this group are from the US or France, and are a Pinot Noir or Chardonnay (refer to Rmd file for the numbers). But what probably made these observations weird was that they almost all had high point ratings in the mid to high 90's, but had wildly different prices. For example, 2 wines, one from France and one from the US were both rated a 95, but the French wine was priced at $200, whereas the US wine was only $50. So with this huge variation in prices among very similar point ratings, this is probably why our data is not normal. In the future we may choose to separate out these wines or do more specific subsetting.

Conclusions and Summary In this report, we have managed to gain some insight on how to choose the best wine to buy based on the analysis of Wine Review Data 2017 obtained from Kaggle, as well as learned a little bit more about global alcohol consumption by Web data scrapping from Wikipedia.

- The countries producing the highest amount of rated wines are the United States, France and Italy.
- There is a statistsically differece in means between the ratings of the top two rated countries (the US and France).
- While Kaggle reveals that wines produced in the US are the most reviewed, US citizens prefer to consume Beer and Spirits over wine.
- There is a moderate positive correlation between price of wine and its point rating, which can be broken down by country of origin.
- True mean price of wine increases with higher rating cateogry for the wine.
- It initally seemed that price would be an appropriate predictor for rating by points, however analysis was complicated by a wide range of observations, and wide price variety among highly rated wines. Grouping variables or undergoing specific subsetting might prove more fruitful in future analysis.