```r
winedata <- read.csv("https://awagaman.people.amherst.edu/introstats/projects/GroupDDataF19.csv")
```

Predicting Alcohol Content in Red wine

Group D, Mary Kate McGranahan, Caroline Useda, and Steven Yu

Abstract

Our project goal was to analyze if the variables that make up wine such as the amount of citric acid, residual sugar, and chlorides have a significant effect on the alcohol content of the wine. We hypothesized that higher quality wine will have lower alcohol content. We constructed a multiple regression model check the effect of our explanatory variables on alcohol content. We found that our model fit very well with an 0.667 r-squared which meant 66.7% of the variance in alcohol content in the wines can be explained by our multiple regression model. We then constructed and ANOVA to see if the alcohol of wine was the same amongst quality groups. Our ANOVA found that higher quality wine actually had higher alcohol content, disproving our hypothesis.

Introduction

In this project, we are analyzing data about red wine, including variables such as alcohol content, quality as rated by wine experts, residual sugar, fixed acidity, citric acid, and more. With these variables in mind, we chose to assess the effect other variables have on the percent of alcohol, as well as if the amount of alcohol in a wine is associated with it's quality. To do so, we completed a multiple regression model for our data, as well as an ANOVA to check for a difference in mean alcohol content across low, medium, and high levels of quality. We began our multiple regression predicting alcohol content using 10 out of the 11 variables left, excluding quality since it's qualitative. When analyzing this model, we found 2 variables (free sulfur dioxide and volatile acidity) had particularly high p values. After noting this, and generating added variable plots for both free sulfur dioxide and volatile acidity, we chose to remove those variables from the model. This left us with a model predicting alcohol content from pH, residual sugar, density, fixed acidity, citric acid, chlorides, total sulfur dioxide, and sulphates.

After completing our multiple regression model, we then began our ANOVA to check for differences in means across quality levels. While quality could have been rated anywhere from 1-10, the values in our data set only ranged from 3 to 8. Thus, to set the levels for our ANOVA, we used values from 2.9 to 4.5 as "Low", 4.5 to 6.5 as "Medium", and 6.5 to 8.1 as "High". We then proceeded to complete the ANOVA, and found there to be a difference in means, and then went on to find level the difference occurred in.

Data

The data that we are using includes various chemical aspects of red vinho verde wines, from the vinho verde region of Portugal, based on multiple factors that could affect it.The observational units are the different red wines used. There are 1,599 observations. We are trying to generalize our results to the population of red vinho verde wines. We are unsure of how many individuals are in the population, but is likely to be a very large number.

The response variable is alcohol which is measured in the percent proportion of alcohol in the wine. The explanatory variables are pH which describe how acidic or basic a wine is on the PH scale from 0 to 14. Residual sugar is the amount of sugar remaining after fermentation stops (g / dm^3). Chlorides are the amount of salt in the wine (g / dm^3). Citric acid is the amount of acetic acid in wine (g / dm^3). The fixed acidity is amount of acids in the wine that do not evaporate readily, the predominant fixed acids in wine are tartaric, malic, citric, succinic, this data measures tartaric acid (tartaric acid - g / dm^3). The volatile acidity is the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste (acetic acid - g / dm^3). Free Sulfur Dioxide is the free form of SO2 that exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine (mg / dm^3). Total Sulfur Dioxide is amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the smell and taste of wine (g / cm^3). The density of wine is close to that of water depending on percent alcohol and sugar content . Sulphates are a wine additive that can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant (potassium sulphate - g / dm^3). Quality is a score between 0-10 based on sensory data from a median of at least 3 evaluations made by wine experts.

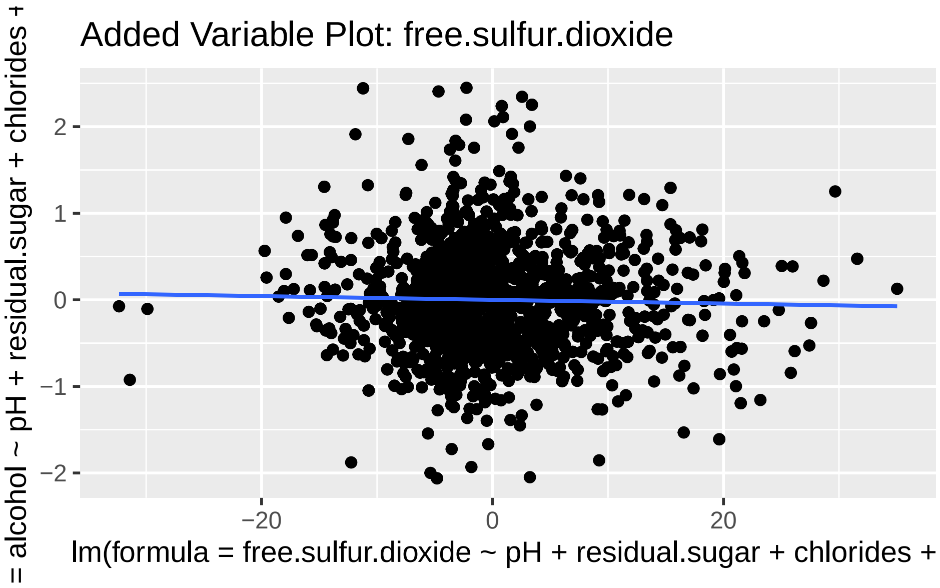
Results

```
winemod <- lm(alcohol ~ pH + residual.sugar + chlorides + citric.acid + fixed.acidity + volatile.acidit
msummary(winemod)
```

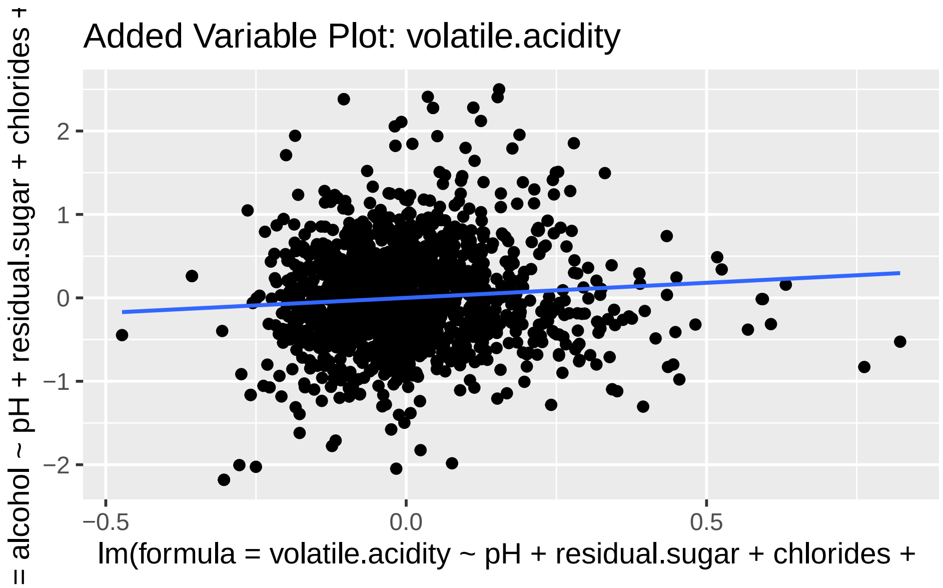
```
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            6.072e+02  1.308e+01  46.419  < 2e-16 ***
## pH                     3.762e+00  1.551e-01  24.263  < 2e-16 ***
## residual.sugar         2.844e-01  1.229e-02  23.135  < 2e-16 ***
## chlorides             -1.462e+00  3.956e-01  -3.696 0.000227 ***
## citric.acid            8.306e-01  1.379e-01   6.024 2.11e-09 ***
## fixed.acidity          5.324e-01  2.064e-02  25.796  < 2e-16 ***
## volatile.acidity       3.608e-01  1.144e-01   3.154 0.001638 ** 
## free.sulfur.dioxide   -2.143e-03  2.057e-03  -1.042 0.297517    
## total.sulfur.dioxide  -2.296e-03  6.881e-04  -3.336 0.000868 ***
## density               -6.174e+02  1.342e+01 -45.998  < 2e-16 ***
## sulphates              1.247e+00  1.037e-01  12.020  < 2e-16 ***
## 
## Residual standard error: 0.614 on 1588 degrees of freedom
## Multiple R-squared:  0.6701, Adjusted R-squared:  0.668 
## F-statistic: 322.5 on 10 and 1588 DF,  p-value: < 2.2e-16
```

In constructing the MLR, the output showed us that explanatory variables volatile acidity and free sulfur dioxide had higher p-values than the rest of the variables.

```
plotAddedVar(winemod,"free.sulfur.dioxide")
```



```
plotAddedVar(winemod, "volatile.acidity")
```



The added variable plots we constructed for these variables displayed horizontal lines, showing that they had almost no effect on predicting the alcohol content in wine in the presence of the other variables. We then chose to take these out.

```
winemod2 <- lm(alcohol ~ pH + residual.sugar + chlorides + citric.acid + fixed.acidity + total.sulfur.di
msummary(winemod2)
```

```
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.004e+02  1.296e+01  46.325  < 2e-16 ***
## pH                    3.768e+00  1.537e-01  24.519  < 2e-16 ***
## residual.sugar        2.826e-01  1.223e-02  23.109  < 2e-16 ***
## chlorides            -1.151e+00  3.833e-01  -3.002  0.00273 **
## citric.acid           6.172e-01  1.167e-01   5.290 1.40e-07 ***
## fixed.acidity         5.364e-01  2.048e-02  26.190  < 2e-16 ***
## total.sulfur.dioxide -2.547e-03  5.095e-04  -4.998 6.42e-07 ***
## density              -6.104e+02  1.329e+01 -45.913  < 2e-16 ***
## sulphates             1.169e+00  1.017e-01  11.497  < 2e-16 ***
##
## Residual standard error: 0.616 on 1590 degrees of freedom
## Multiple R-squared:  0.6675, Adjusted R-squared:  0.6658
## F-statistic:   399 on 8 and 1590 DF,  p-value: < 2.2e-16
```

$\hat{alcohol} = 607 + 3.76(pH) + .284(residual.sugar) - 1.46(chlorides) + .831(citric.acid) + .532(fixed.acidity) - .00230(total.sulfur.dioxide) - .0617(density) + 1.25(sulphates)$

Looking at our final MLR regression equation, if all the explanatory variables are fixed, then for each change of one unit in pH, predicted alcohol increases 3.76 units. If all variables are fixed, for every one-unit increase in the residual sugar, the predicted alcohol will lead to an increase in 0.284. When examining the coefficient of fixed acidity, we can see that for every increase in fixed acidity, the predicted alcohol will increase by 0.532 units, and for a one-unit change in sulfates, there will be a 1.25 unit increase in alcohol predicted. For every unit change in chlorides, it leads to a decrease in alcohol by 1.46 units. We also see a decrease in predicted alcohol when looking at the total sulfur dioxide of 0.00230 and 0.0617 for density.

For our ANOVA, our null hypothesis was that the mean percent alcohol per portion is the same for all quality groups. Our alternative hypothesis was that the mean percent alcohol per portion is not the same for all quality groups. In order to run the ANOVA, we had to separate the quality variable into three quality groups, low, medium, and high.

```
winedata <- mutate(winedata, qualitygrp=cut(quality, breaks=c(2.9, 4.5,6.5,8.1),labels=c("low","medium"
winemod <- lm(alcohol ~ qualitygrp, data = winedata)
```

After performing an ANOVA on alcohol and quality group, we have sufficient evidence to reject our null hypothesis.

```
anova(winemod)
```

```
## Analysis of Variance Table
##
## Response: alcohol
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## qualitygrp    2  301.16 150.580  158.78 < 2.2e-16 ***
## Residuals  1596 1513.60   0.948
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the ANOVA was $2e^{-16}$, therefore if we set an alpha level at .05, the p-value is far below the alpha level, indicating that we can reject the null hypothesis. In rejecting our null hypothesis, we conclude that the mean percent alcohol per portion is not the same for all quality groups.

```
with(winedata, pairwise.t.test(alcohol, qualitygrp, p.adj="holm"))
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  alcohol and qualitygrp
##
##        low    medium
## medium 0.77   -
## high   <2e-16 <2e-16
##
## P value adjustment method: holm
```

We constructed a Holm's test to determine what groups had the most significant difference in means. The Holm's test showed that there is a significant difference between the alcohol content in high quality wines and the alcohol content in low and medium quality wines. According to the box plots we constructed to check and make sure there was a difference in means before we ran the ANOVA, higher quality wines have higher alcohol content.

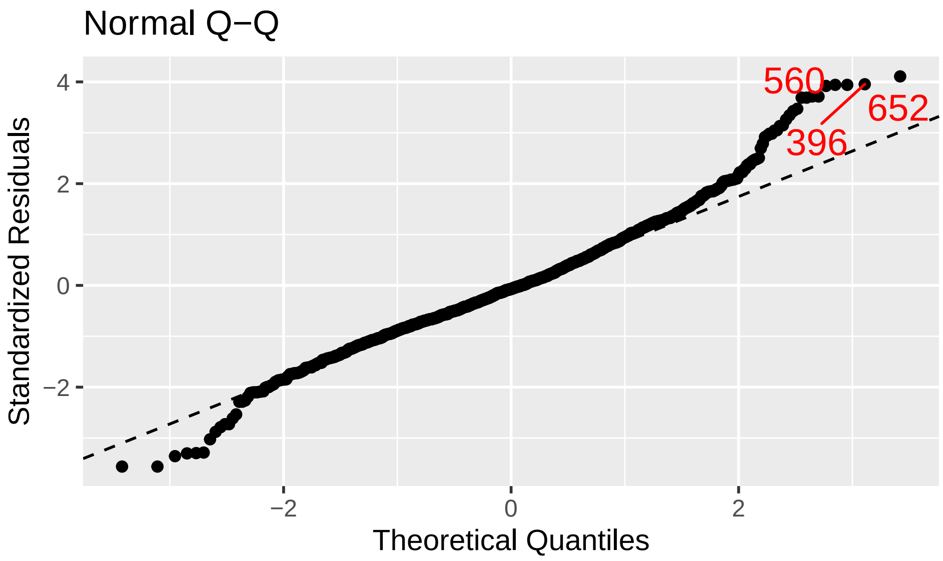
Diagnostics

MLR:

Prior to constructing our regression model, we had to check the straight enough condition for all of our variables. We found that although there was not a strong association in any of the plots, the plots were linear enough to satisfy the straight enough condition for multiple regression. To satisfy independence and randomization we had to assume this to be a representative and independent sample of vinho verde red wines. To check the nearly normal condition we constructed a Q-Q plot.

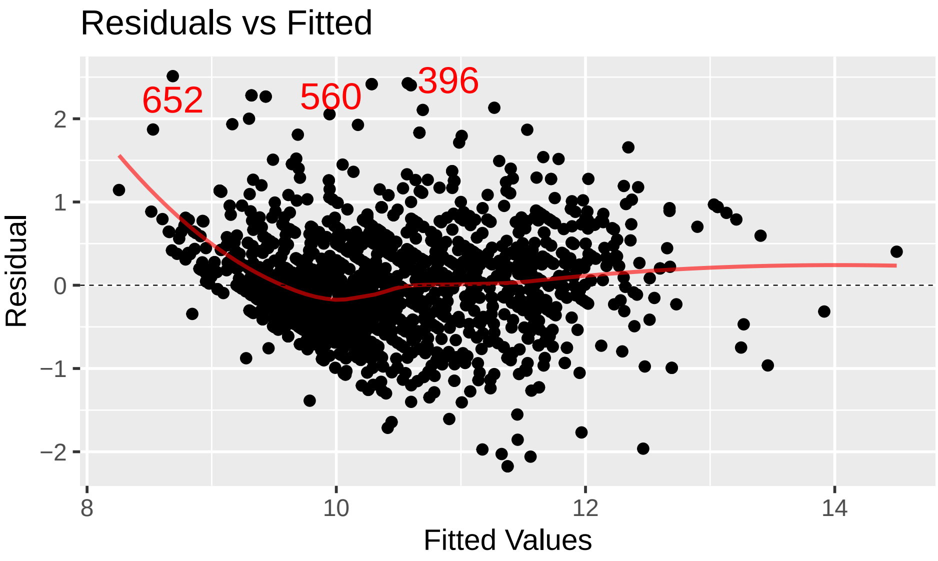
```
mplot(winemod2, which=2)
```

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



```
mplot(winemod2, which =1)
```

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

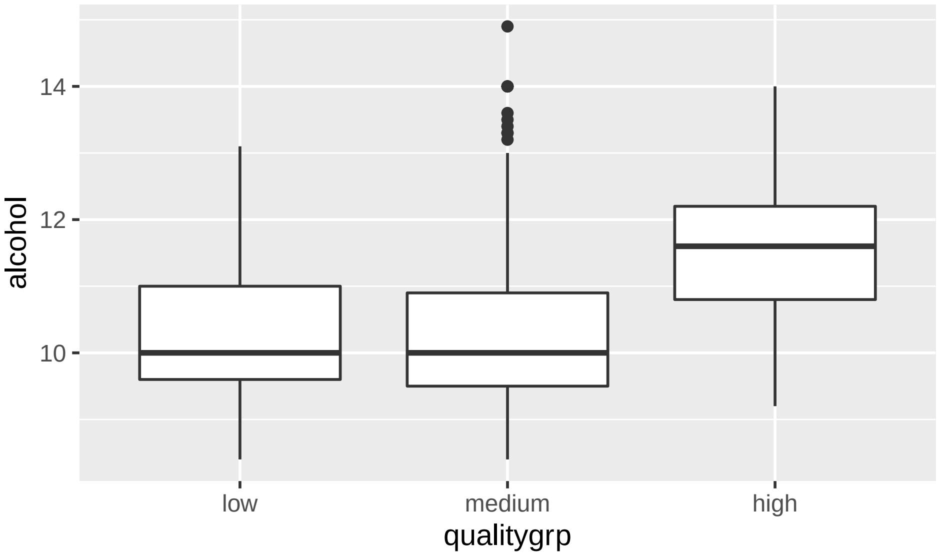


The width of the plot is not perfectly even but is fairly consistent so we decided to proceed with caution.

ANOVA:

To be able to conduct an ANOVA for mean alcohol content across varying levels of quality, we first needed to make quality into a qualitative variable. Quality can vary from 1-10 in general, but our data only ranged from 3-8. To make this a qualitative variable, we used grouped quality into low, medium and high levels. Values from 3-4 were low, 5-6 medium, and 7-8 high. After completing this step, we were able to make a boxplot of alcohol content across quality groups, which showed us than ANOVA was reasonable.

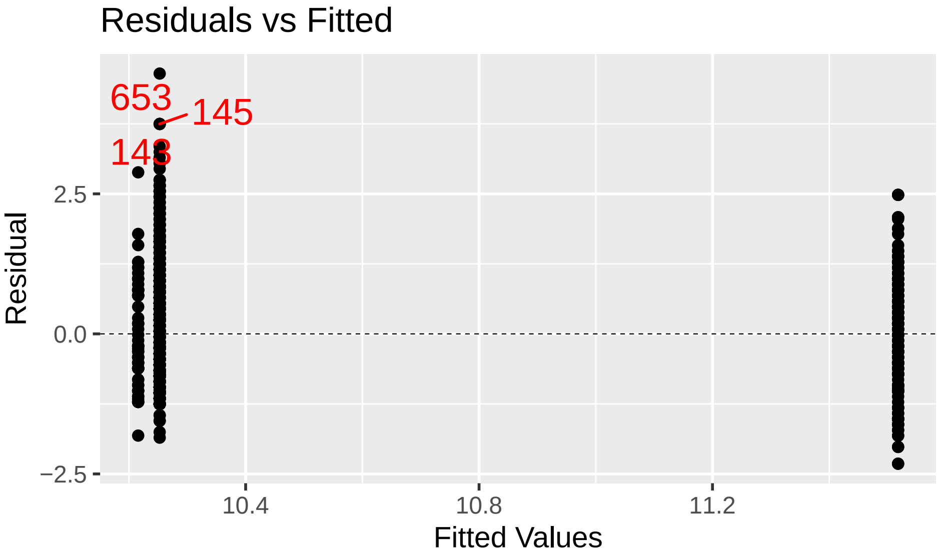
```
gf_boxplot(alcohol~qualitygrp, data=winedata)
```



To proceed with ANOVA, we also need to assume all samples were independent both of each other but in the way they were collected. Additionally, we constructed both a Q-Q plot and Residual v. Fitted Plot for alcohol~qualitygrp.

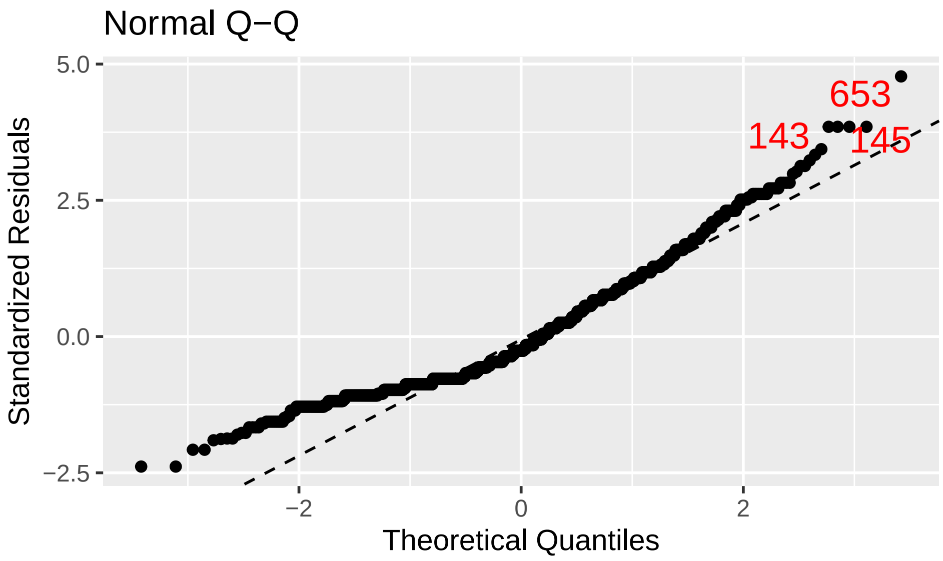
```
mplot(winemod, which = 1)
```

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



```
mplot(winemod, which = 2)
```

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



These showed no issues, therefore satisfying both the nearly normal condition as well as equal variance.

Conclusion

We used two methods to analyze the data about vinho verde red wine. After fitting a multiple regression model, with percent alcohol as the predictor, we found that density, chlorides, and total sulfur dioxide had negative effects on the alcohol content in red wine, while pH, residual sugar, citric acid, fixed acidity, and sulphates had a positive effect. We then went on to complete an ANOVA for difference in mean alcohol content across low, medium, and high levels of quality. From that data, we concluded that there is an association between high alcohol content and high quality rating, showing that "better" wine, in our data, has more alcohol. Therefore, our hypothesis was wrong, higher quality wine does not have lower alcohol content. This has practical applications in its ability to help winemakers determine what ingredients and what quantity of them they want to include in their wine. Alcohol is a result of fermentation, so knowing what ingredients they want to put in prior to storing the wine given what alcohol level they are looking to achieve is important. If we could do further research on this topic, we would like to find data on how the wine is stored, such as the type of container it is in and temperature it is stored. According to our research those variables are known to have a large effect on how much alcohol is formed in the fermentation process and it would be interesting to see if they could be a good addition to our model.
The limitation of this data is that it only applies to red vinho verde wine. This wine is from a very specific region and is also very young wine. The explanatory variables we used in our regression model could have different effects in more aged wine or wine from different regions so these findings cannot be generalized to a larger population.

Data Source

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

https://daily.sevenfifty.com/taking-control-of-alcohol-levels-in-wine/