SCI3501
ACADEMIC WRITING AND STATISTICAL TECHNIQUES FOR SCIENTIFIC RESEARCH

Analysing the Physicochemistry and Perceived Quality of *Vinho Verde* Wines from North Portugal

Compiled by: Yesahel Scicluna (141901L)

(**Grouped with:** Francesca Grech, Bettina Nardelli, Samantha Jade Sammut)

Tutor: Dr Monique Borg Inguanez

Department of Statistics and Operations Research
University of Malta

JANUARY 2023

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Two datasets concerning the red and white variants of the Portuguese *vinho verde* were obtained from *http://archive.ics.uci.edu/ml/datasets/Wine+Quality* on 12th January 2023 and analysed using RStudio 2022.12.0 (Build 353). Having a combined sample size of 6497 instances, these datasets constitute 12 variables, 6 of which have been considered in this study. A brief description of these 6 variables is presented in **Table 1** and a full list of the codes inputted into R and the respective results it outputted presented in the **Appendix** section.

Table 1. Summary description of the 6 variables associated with the *vinho verde* datasets selected for analysis in this report.

Variable name	Variable type	Units/ Levels	Variable description
Colour	Qualitative: nominal	red; white	Colour of the wine sample, whether red or white.
Acidity	Qualitative: ordinal	stronger acid; weaker acid	Acidity of the wine sample, whether stronger (pH 0-3.5) or weaker (pH 3.5-7).
Quality	Quantitative: discrete	0; 1; 2; 3; 4; 5; 6; 7; 8; 9; 10	Median of at least 3 evaluations made by wine experts, each of which graded the wine sample between 0 (very poor) and 10 (very excellent).
Alcohol	Quantitative: continuous	% vol.	Percent alcohol content of the wine sample.
Sulfates	Quantitative: continuous	g/dm^3	Concentration of potassium sulfate in the wine sample.
Density	Quantitative: continuous	g/cm^3	Mass per unit volume of the wine sample.

Vinho verde is a unique wine from the Minho region in northwest Portugal that is medium in alcohol and that is particularly appreciated for its 'fresh' taste; it accounts for 15% of the total Portuguese production. The data on this product were collected between the period of May 2004 and February 2007 using only Protected Designation of Origin samples approved by the Commission of Viticulture of the Region Vinho Verde, the official certification entity (Cortez et al., 2009). The aim of this study was to provide tentative answers to three questions relating to *vinho verde*, namely:

1) What colour *vinho verde* is better rated?
2) Do the density, alcohol content, and sulfate content of *vinho verde* have any bearing on the ratings it receives?
3) Can the colour of *vinho verde* predict whether the wine is more or less acidic in nature?

The objectives of this study were to first conduct exploratory analyses related to the above questions and to subsequently apply statistical tests and models where appropriate to corroborate our findings.

EXPLORATORY DATA ANALYSIS

Question 1:

Red *vinho verde* samples were given a mean quality score of 5.636 and white *vinho verde* samples a mean quality score of 5.878. This statistic already seems to indicate that there is no practical difference in how the two colour variants of *vinho verde* are rated. However, confirmation of this inference requires further testing, namely by the parametric Independent Samples *t* Test or otherwise by its non-parametric (and therefore less powerful) equivalent, the Mann-Whitney U test, should the criteria necessitated by the former test not be met.

In anticipation of the Independent Samples *t* Test assuming that the two colour variants have quality scores that are normally distributed and that are equal in variance, bar charts and box plots for the two wines were computed and values for standard deviation (the square root of variance) calculated. The bar charts presented in **Figure 1** may suggest that the scores of the two wines are vaguely normally distributed, though the box plot presented in **Figure 2** strongly suggests the contrary – that the score distributions are not symmetrical and that, more specifically, they are negatively skewed (i.e., skewed to the left). With regards to the second assumption, red *vinho verde* samples were found to have scores with an SD value of 0.808 and white *vinho verde* samples to have scores with an SD value of 0.886, seemingly indicating that the two sets of scores may indeed be equivalent in their variances. Confirmation of these inferences will also need further statistical testing, namely by applying the Shapiro-Wilk test for normality and the Levene test for equality of variances.

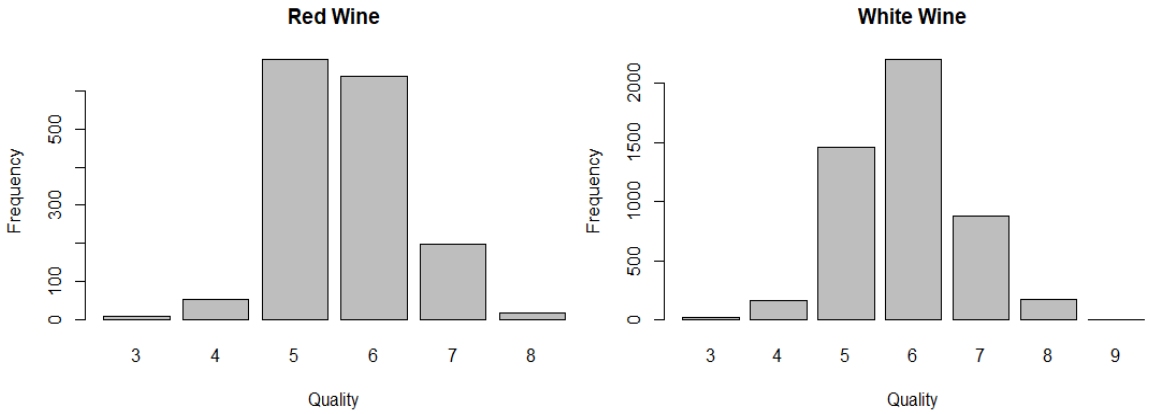


Fig. 1. Bar chart illustrating the distribution of the scores given to the quality of red and white *vinho verde* samples.

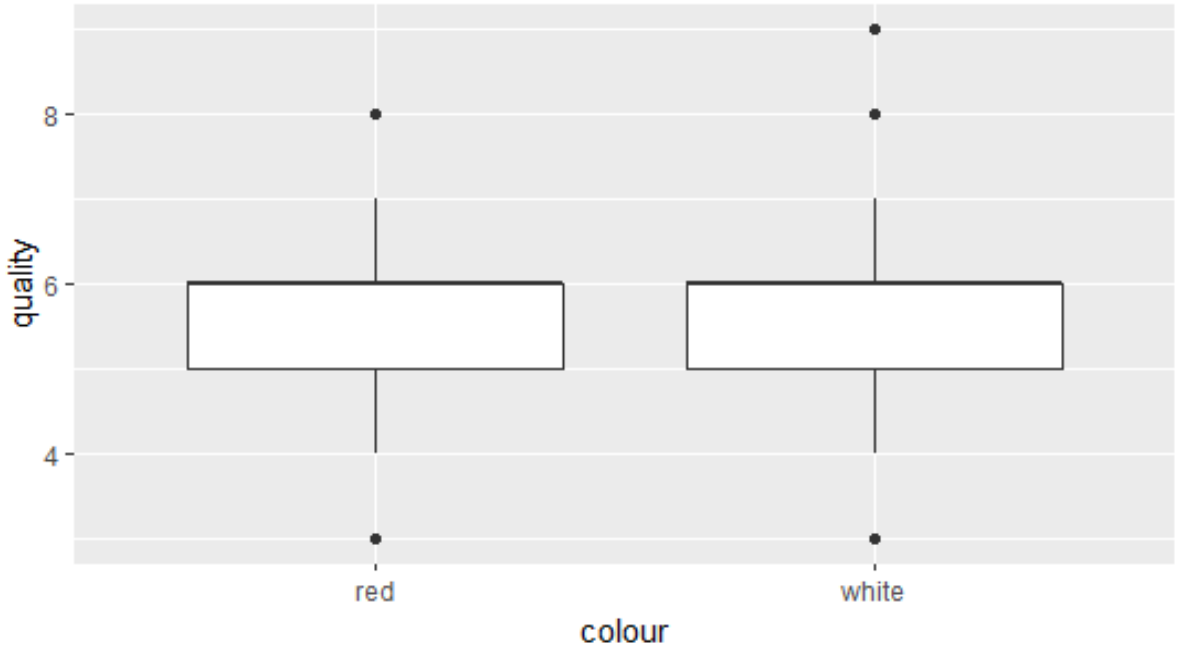


Fig. 2. Box plot illustrating the distribution of the scores given to the quality of red and white *vinho verde* samples.

Question 2:

Scatter plots were computed to visualise the relationships that may potentially exist between the three physicochemical parameters here being considered (i.e., wine density, alcohol content, and sulfate content) and the quality scores that were given to the *vinho verde* samples; these scatter plots are presented in **Figure 3**. The data points appear widely dispersed in most of the scatterplots, suggesting weak correlations throughout; it is only in the diagrams where density and alcohol are plotted against each other that the data points appear to condense along a line more clearly, suggesting a stronger correlation between the two variables. Still,

confirmation of these inferences requires further testing, namely by measuring the parametric Pearson correlation coefficient or otherwise by measuring its non-parametric (and therefore less powerful) equivalent, the Spearman's rank correlation coefficient, should the criteria necessitated by the former not be met. In this regard, the scatterplots suggest that not only are the relationships between the four parameters monotonal in nature, as assumed and therefore necessitated by Spearman's coefficient, but better yet linear, as assumed and therefore necessitated by the Pearson's coefficient.

In addition to these tests, a multiple linear regression model needs to be computed to estimate the extent to which the selected physicochemical properties determine the quality score given to *vinho verde*; this too assumes and requires linearity. As discussed above, the scatter plots seemingly suggest that a strong correlation exists between density and alcohol content. On this account, it can already be suspected that one of these covariates will need to be discarded when it comes to computing the model, given that it operates under the assumption that multicollinearity is absent.

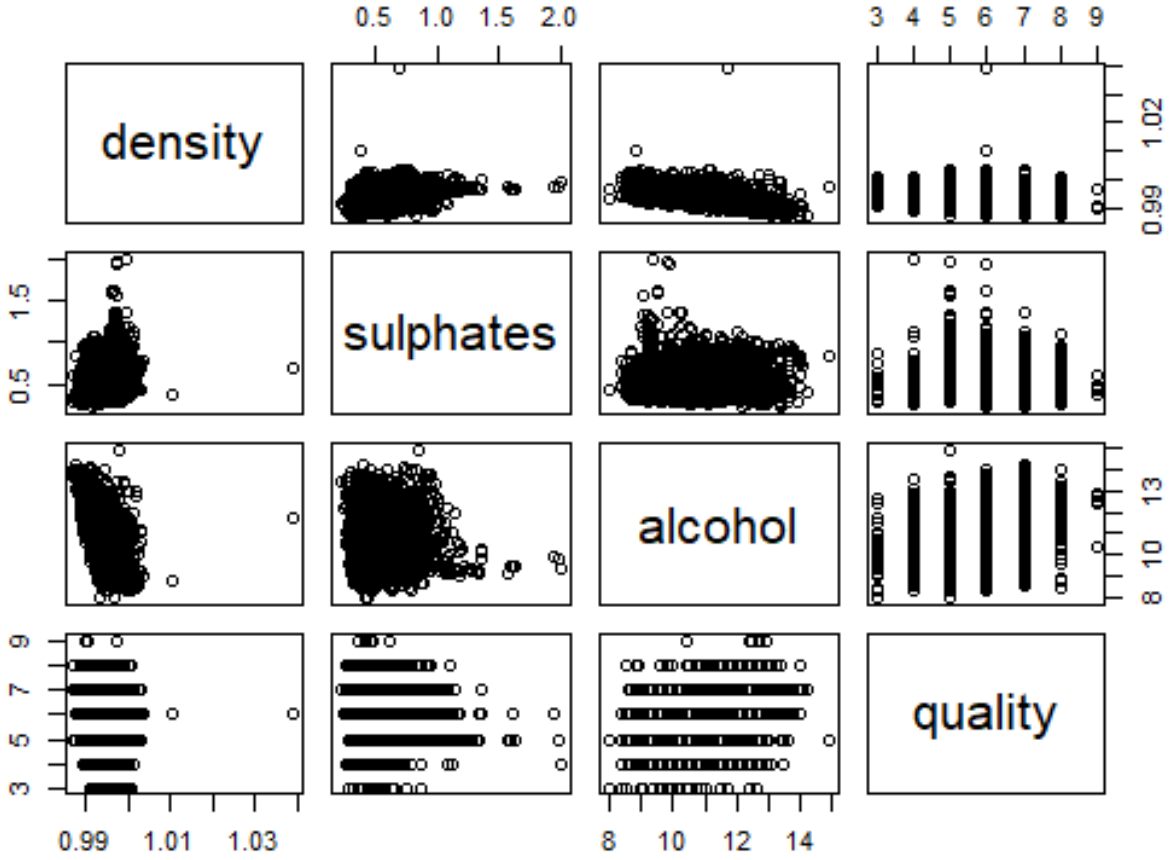


Fig. 3. Scatter plots illustrating the relationships and strengths thereof of between the density, alcohol content, and sulfate content of the *vinho verde* wine samples and the quality scores given to them.

Question 3:

Pie charts were computed in order to compare the frequencies of *vinho verde* samples recorded in the dataset that differ by colour and by extent of acidity; these are presented in **Figure 4**. From these pie charts, one might infer that white *vinho verde* samples are largely stronger acids and that red *vinho verde* samples are largely weaker acids, but the association between the two categorical variables needs to be confirmed through a Pearson's χ^2-test.

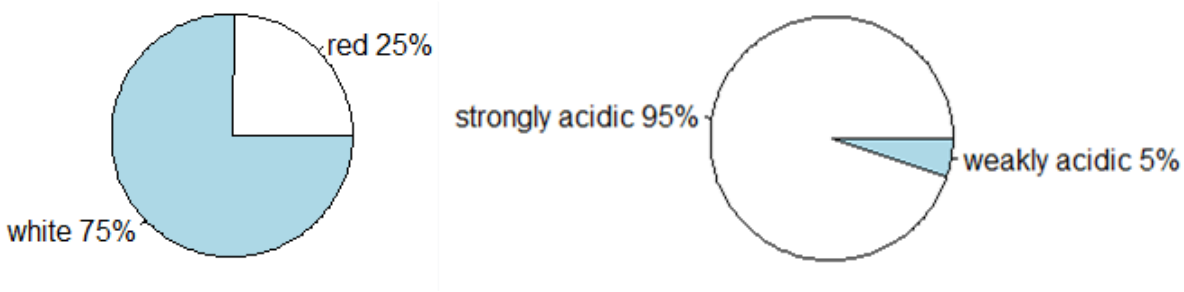


Fig. 4. Pie charts illustrating the proportion of red and white *vinho verde* samples recorded in the dataset as well as the proportion of *vinho verde* samples that are strongly acidic (pH 0-3.5) and weakly acidic (pH 3.5-7).

STATISTICAL TESTING AND MODELLING

Question 1:

Taking into consideration a 0.05 level of significance, the Shapiro-Wilk test for normality was applied to test the following hypotheses:

H_0*: The quality scores attributed to red and white* vinho verde *wines follow a normal distribution.*

H_1*: The quality scores attributed to red and white* vinho verde *wines do not follow a normal distribution.*

The following values were outputted for the red and white wines, respectively: W = 0.85759, *p*-value < 2.210^{-16}; W = 0.88904; *p*-value < 2.2×10^{-16}.

Since both the resulting *p*-values are less than 0.05, there is enough evidence to reject the null hypothesis and accept the alternative hypothesis, which means that it can be assumed that the quality scores attributed to either colour variant of *vinho verde* wine do not follow a normal distribution.

Furthermore, again taking into consideration a 0.05 level of significance, Levene's test for equality of variances was applied to test the following hypotheses:

H_0*: The variances of the quality scores attributed to red and white* vinho verde *wines are equal.*

H_1 (two-tailed)*: The variances of the quality scores attributed to red and white* vinho verde *wines are not equal.*

The following values were outputted for the red and white wines, respectively: test statistic = 0.62133, *p*-value = 0.4306.

Since the resulting *p*-value is greater than 0.05, there is not enough evidence to reject the null hypothesis and accept the alternative hypothesis, which means that it can be assumed that the quality scores attributed to the two colour variants of *vinho verde* wine are equal in variance.

Given that the quality scores do not follow a normal distribution, the Mann-Whitney U test was applied in lieu of the Independent Samples *t* Test to test for differences in the average scores attributed to the two *vinho verde* variants. A 0.05 level of significance was used to test the following hypotheses:

H_0*: The median quality scores attributed to red and white* vinho verde *wines are equal.*

H_1 (two-tailed) *: The median quality scores attributed to red and white* vinho verde *wines are not equal.*

The following values were outputted: W = 3311514, *p*-value < 2.2×10^{-16}.

Since the resulting *p*-value is less than 0.05, there is enough evidence to reject the null hypothesis and accept the alternative hypothesis, which means that it can be assumed that there is a significant difference in the median quality score attributed to the red and white *vinho verde* wines. Taking into consideration the descriptive statistics reported in the previous section, rejecting this null hypothesis implies that white *vinho verde* wines are, statistically speaking, significantly better rated than their red counterparts. This result is highly unexpected considering that the two variants have similar mean scores of 5.636 and 5.878 and equal median scores of 6.000 (see the box plot in **Figure 2**).

Question 2:

Taking into consideration a 0.05 level of significance, the energy test of multivariate normality was applied to test the following hypotheses:

H_0*: The data pertaining to* vinho verde *density, alcohol content, sulfate content and quality score follow a multivariate normal distribution.*

H_1: The data pertaining to vinho verde *density, alcohol content, sulfate content and quality score do not follow a multivariate normal distribution.*

The following values were outputted: E-statistic = 69.196, *p*-value < 2.210^{-16}.

Since the resulting *p*-value is less than 0.05, there is enough evidence to reject the null hypothesis and accept the alternative hypothesis, which means that it can be assumed that the data pertaining to *vinho verde* density, alcohol content, sulfate content and quality score do not follow a multivariate normal distribution.

Given that these data do not follow a multivariate normal distribution, this being an assumed prerequisite when it comes to calculating a value for the Pearson correlation coefficient, a value for Spearman's rank correlation coefficient was calculated instead. As already indicated in the previous section, cursory inspection of the scatter plots in **Figure 3** suggests that this non-parametric test's assumption of monotonicity is met. A 0.05 level of significance was used to test the following hypotheses per pair of quantitative variables:

H_0: The two quantitative variables are independent of one another.

H_1: The two quantitative variables are not independent of one another; a relationship exists between them that can be modelled by a monotonic function.

The matrix of correlation coefficients and *p*-values outputted by the test are presented in **Table 2**. These results indicate that alcohol content is moderately correlated to a statistically significant extent with the quality score (ρ= 0.45, *p*-value= 0.0000); that density is somewhat less correlated with the quality score, but still to a statistically significant extent (ρ= -0.32, *p*-value= 0.0000); and that sulfate content is practically not correlated with the quality score at all, but that this correlation is still statistically significant (ρ= 0.03, *p*-value= 0.0162). In practice, this indicates that the persons rating the *vinho verde* samples gave somewhat consistently higher scores to wines that were higher in alcohol content; higher scores to lighter wines (i.e., to ones of lower density), although somewhat less consistently; and higher scores to wines of highly variable sulfate contents.

Moreover, as suspected, these results indicate a strong and statistically significant correlation between wine density and alcohol content (ρ= -0.70, *p*-value= 0.0000). Given that density is less strongly correlated with quality score (ρ =-0.32 as opposed to ρ= 0.45), it will be disregarded in fitting the multiple linear regression model discussed hereunder. This is because one of the assumptions of this statistical model is the absence of multicollinearity among the covariates.

Table 2. Matrix of Spearman rank correlation coefficients (ρ) and associated *p*-values obtained for the density, alcohol content, sulfate content, and quality scores of the *vinho verde* wine samples.

	Density	**Sulfates**	**Alcohol**	**Quality**
Density	ρ= 1.00	ρ= 0.27, p= 0.0000	ρ= -0.70, p= 0.0000	ρ= -0.32, p= 0.0000
Sulfates	ρ= 0.27, p= 0.0000	ρ= 1.00	ρ= 0.00, p= 0.7119	ρ= 0.03, p= 0.0162
Alcohol	ρ= -0.70, p= 0.0000	ρ= 0.00, p= 0.7119	ρ= 1.00	ρ= 0.45, p= 0.0000
Quality	ρ= -0.32, p= 0.0000	ρ= 0.03, p= 0.0162	ρ= 0.45, p= 0.0000	ρ= 1.00

A multiple linear regression model of the below general equation was fitted onto the data, considering the *vinho verde* quality scores as the response (Y) variable and sulfate content and alcohol content as the explanatory ($X_{1,2}$) variables:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \varepsilon$$

Taking into consideration a 0.05 level of significance, the fitness of the model was evaluated by testing for the following three pairs of hypotheses:

H_0: $\beta_{0,1,2}$ are statistically equal to zero.
$H_{1\ (two\text{-}tailed)}$: $\beta_{0,1,2}$ are statistically not equal to zero.

The values outputted by the test are summarised in **Table 3**.

Table 3. Results outputted upon attempting to fit a multiple linear regression model with *vinho verde* quality as the response variable and sulfates and alcohol as the explanatory variables.

	Coefficient estimate	**Std. error**	**t value**	**p-value**
Intercept	2.280158	0.092678	24.603	$< 2x10^{-16}$
Sulfates	0.233749	0.065175	3.586	0.000338
Alcohol	0.325400	0.008131	40.018	$< 2x10^{-16}$

Since the resulting *p*-values were all less than 0.05, there is enough evidence to reject the null hypothesis in all three cases, which means that the estimated coefficients are significantly different from 0. Therefore, the fitted model takes the following equation:

$$\widehat{Quality} = 2.280158 + 0.233749(Sulfates) + 0.325400(Alcohol)$$

Inputting a wine sample's sulfate concentration in g/dm^3 and alcohol content as % vol into the above equation should therefore allow for the quality score given to the same wine sample to be calculated. Furthermore, the adjusted R^2 value obtained is 0.1988. This is a rather low value that implies that the sulfate and alcohol content of *vinho verde* wines can only account for 19.88% of the variation in the quality scores the wines obtain; sulfate and alcohol content alone are therefore not very good predictors of *vinho verde* quality.

The above inferences are only valid if the assumptions related to multiple linear regression are satisfied. In light of this, the residuals were tested for their normality, independence, and homoscedasticity and outlier diagnostic measures applied, as detailed hereunder.

Provided that the sample size exceeded 5000 instances, the Anderson-Darling normality test was applied in lieu of the Shapiro-Wilk test for normality. Taking into consideration a 0.05 level of significance, the following hypotheses were tested:

H_0: The residuals of the fitted model follow a normal distribution.

H_1: The residuals of the fitted model do not follow a normal distribution.

The following values were outputted: A = 29.757; *p*-value < 2.2×10^{-16}.

Since the resulting *p*-value is less than 0.05, there is enough evidence to reject the null hypothesis and accept the alternative hypothesis, which means that it can be assumed that the residuals fail to follow a normal distribution.

Taking into consideration a 0.05 level of significance, the Durbin-Watson test was applied to test the following hypotheses:

H_0: There is no correlation between the residuals of the fitted model, that is, the residuals are independent of one another.

H_1: There is a correlation between the residuals of the fitted model, that is, the residuals are not independent of one another.

The following values were outputted: DW = 1.6245; *p*-value < 2.2×10^{-16}.

Since the resulting *p*-value is less than 0.05, there is enough evidence to reject the null hypothesis and accept the alternative hypothesis, which means that it can be assumed that the residuals fail to be independent of one another.

Taking into consideration a 0.05 level of significance, the studentised Breusch-Pagan test was applied to test the following hypotheses:

H_0: The residuals of the fitted model are homoscedastic, that is, they display constant variance.

H_1: The residuals of the fitted model are not homoscedastic, that is, they do not display constant variance.

The following values were outputted: BP = 23.419; *p*-value = 8.215×10^{-6}.

Since the resulting *p*-value is less than 0.05, there is enough evidence to reject the null hypothesis and accept the alternative hypothesis, which means that it can be assumed that the residuals fail to be homoscedastic. This inference was corroborated by visual inspection of a scatter diagram in which the residual values are plotted against the fitted values; this diagram is presented in **Figure 5**.

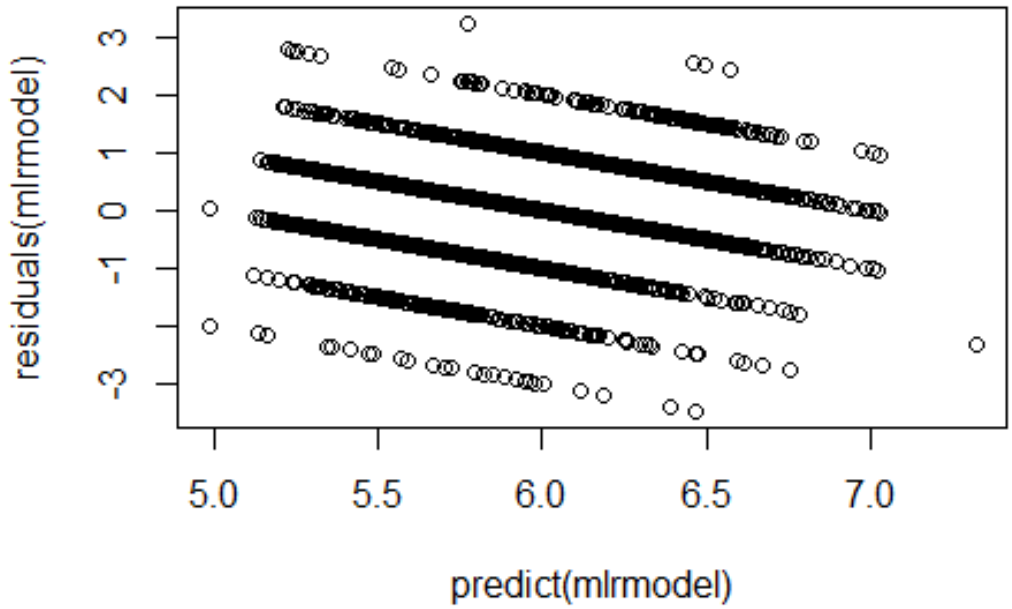


Fig. 5. Scatter diagram in which the residual values of the fitted multiple linear regression model are plotted against the predicted values of the same model, suggesting that the residuals do not display constant variance.

276 potential outliers were identified using Mahalanobis distances, an additional 137 were identified using Leverages, and none were identified using Cook's distances. It is likely that these 413 outliers are inadvertently exerting excess influence on the fitted model.

Considering that the assumptions related to residual normality, residual independence, residual homoscedasticity, and influential outliers were not satisfied, the fitted model has limited validity. Inferences on how the sulfate and alcohol content of *vinho verde* wines predict quality depending on the fitted model should therefore be made with added caution.

Question 3:

Taking into consideration a 0.05 level of significance, Pearson's χ^2-test was applied to test the following hypotheses:

H_0: There is no association between vinho verde*'s colour and acidity (the two categorical variables are independent of one another).*

H_1: There is an association between vinho verde*'s colour and acidity (the two categorical variables are not independent of one another).*

The following values were outputted: χ-squared = 6651.6, p-value < 2.2×10^{-16}.

Since the resulting p-value is less than 0.05, there is enough evidence to reject the null hypothesis and accept the alternative hypothesis, which means that it can be assumed that *vinho verde* colour and acidity are associated. Taking into consideration the pie charts in **Figure 4**, rejecting this null hypothesis implies that white *vinho verde* wines are to a statistically significant extent more strongly acidic and red *vinho verde* wines are to a statistically significant extent less strongly acidic.

CONCLUSION

Statistical analysis suggested that white *vinho verde* wines are significantly better rated than their red counterparts. However, on a simple scale from 1-10, the two variants scored the same average rating of 6. Further analysis revealed that the sulfate and alcohol content of *vinho verde* wines only account for approximately 20% of the variation in the scores the two wine variants obtain, implying that sulfate and alcohol content alone are not good predictors of *vinho verde* perceived quality. However, considering that multiple assumptions related to the multiple linear regression model that was fitted were not satisfied, this inference is expected to have limited validity. In addition, a final statistical test revealed that white *vinho verde* wines are significantly more strongly acidic in nature and that red *vinho verde* wines are significantly less strongly acidic in nature.

REFERENCES

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553. https://doi.org/10.1016/j.dss.2009.05.016

APPENDIX

Below is a full record of the codes inputted into RStudio to generate the data analysis presented in this report:

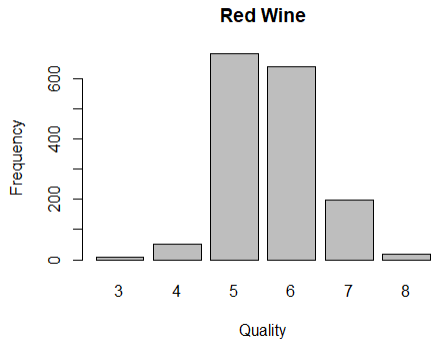
Preparing the data for analysis

```
## Importing the data

library(readxl)
winequality_red_edited <- read_excel("C:/Users/zache/Desktop/R Assignment/
0. Raw data/winequality.red.edited.xlsx")
View(winequality_red_edited)

library(readxl)
winequality_white_edited <- read_excel("C:/Users/zache/Desktop/R Assignmen
t/0. Raw data/winequality.white.edited.xlsx")
View(winequality_white_edited)

## Removing the variables that will not be considered

winequality_red_edited$'fixed acidity'<-NULL
winequality_red_edited$'volatile acidity'<-NULL
winequality_red_edited$'citric acid'<-NULL
winequality_red_edited$'residual sugar'<-NULL
winequality_red_edited$'chlorides'<-NULL
winequality_red_edited$'free sulfur dioxide'<-NULL
winequality_red_edited$'total sulfur dioxide'<-NULL

winequality_white_edited$'fixed acidity'<-NULL
winequality_white_edited$'volatile acidity'<-NULL
winequality_white_edited$'citric acid'<-NULL
winequality_white_edited$'residual sugar'<-NULL
winequality_white_edited$'chlorides'<-NULL
winequality_white_edited$'free sulfur dioxide'<-NULL
winequality_white_edited$'total sulfur dioxide'<-NULL

## Creating a merged data set

winequality_merged<-rbind(winequality_red_edited, winequality_white_edited
)
View(winequality_merged)

## Reordering the columns

winequality_merged<-winequality_merged[,c(1,3:5,2,6)]

## Designating colour as a categorical variable
```

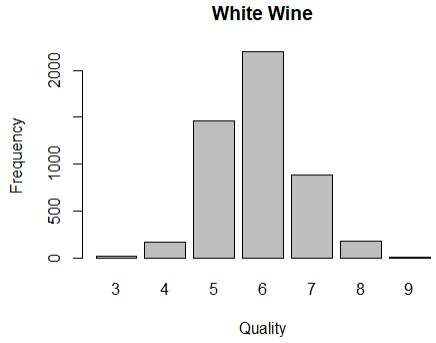
```
winequality_red_edited$colour <- factor(winequality_red_edited$colour)
levels(winequality_red_edited$colour)<-c("red", "white")

winequality_white_edited$colour <- factor(winequality_white_edited$colour)
levels(winequality_white_edited$colour)<-c("red", "white")

winequality_merged$colour <- factor(winequality_merged$colour)
levels(winequality_merged$colour)<-c("red", "white")

## Converting pH into a categorical variable, 'acidity'

acidity<-cut(winequality_merged$pH,breaks=c(-Inf, 3.5, Inf),labels=c("stro
ngly acidic","weakly acidic"),right=FALSE)
winequality_merged<-cbind(winequality_merged, acidity)
winequality_red_edited$'pH'<-NULL
winequality_white_edited$'pH'<-NULL
winequality_merged$'pH'<-NULL
```

Exploratory analysis related to question 1

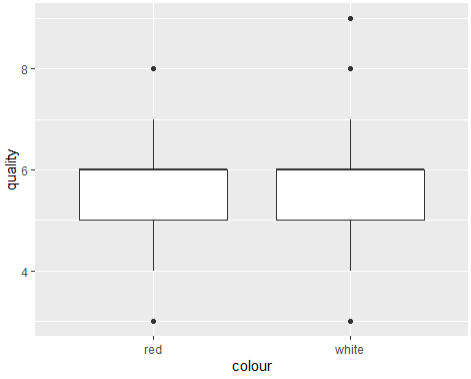
```
## Checking the means and standard deviations of the scores for red and wh
ite wines

summary(winequality_red_edited$quality)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000

sd(winequality_red_edited$quality)

## [1] 0.8075694

summary(winequality_white_edited$quality)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.878   6.000   9.000

sd(winequality_white_edited$quality)

## [1] 0.8856386

## Checking the distribution of scores for red and white wines

frequency_redwine_quality <- table(winequality_red_edited$quality)
redwine_barchart<-barplot(frequency_redwine_quality, main="Red Wine", xlab
="Quality", ylab="Frequency",
names.arg=levels(winequality_red_edited$quality))
```




```
frequency_whitewine_quality <- table(winequality_white_edited$quality)
whitewine_barchart<-barplot(frequency_whitewine_quality, main="White Wine"
, xlab="Quality", ylab="Frequency",
names.arg=levels(winequality_white_edited$quality))
```

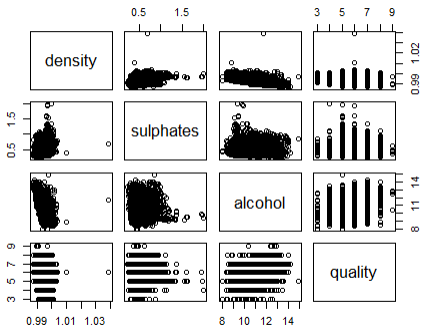



```
## Comparing the average scores of red and white wines

library(ggplot2)
score_boxplot <- ggplot(winequality_merged, aes(x=colour, y=quality, na.rm
= TRUE)) + geom_boxplot(na.rm = TRUE)
score_boxplot
```



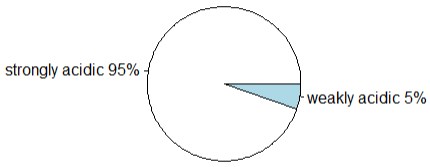
Exploratory analysis related to question 2

```
## Checking for any potential relationships between the different quantita
tive variables

pairs(winequality_merged[1:4])
```

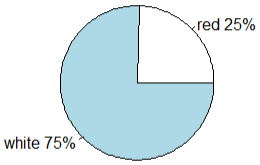


Exploratory analysis related to question 3

```
## Checking the frequency of stronger/weaker wines and red/white wines

summary(winequality_merged$acidity)

## strongly acidic   weakly acidic
##            6159             338

summary(winequality_merged$colour)

##   red white
##  1599  4898

## Expressing the above frequencies as percentages

slices<-summary(winequality_merged$acidity)
lbls<-levels(winequality_merged$acidity)
prcnt<-round(slices/sum(slices)*100)
lbls<-paste(lbls, prcnt)
lbls <- paste(lbls,"%",sep="")
pie(slices, labels=lbls)
```




```
slices2<-summary(winequality_merged$colour)
lbls2<-levels(winequality_merged$colour)
prcnt2<-round(slices2/sum(slices2)*100)
lbls2<-paste(lbls2, prcnt2)
lbls2<- paste(lbls2,"%",sep="")
pie(slices2, labels=lbls2)
```



Data analysis related to question 1

```
## Testing if the red and white wine scores are normally distributed

by(winequality_merged$quality, winequality_merged$colour, shapiro.test)
```

```
## winequality_merged$colour: red
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.85759, p-value < 2.2e-16
##
## ------------------------------------------------------------
## winequality_merged$colour: white
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.88904, p-value < 2.2e-16
```

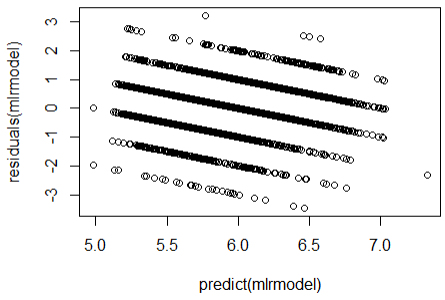
```
## Testing whether the red and white wines are equal in variance

library(lawstat)
levene.test(winequality_merged$quality,winequality_merged$colour,location=
'mean')

##
##  Classical Levene's test based on the absolute deviations from the mean
##  ( none not applied because the location is not set to median )
##
## data:  winequality_merged$quality
## Test Statistic = 0.62133, p-value = 0.4306

## Testing if the average red and white wine scores are different

wilcox.test(quality ~ colour, data=winequality_merged, exact=FALSE)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  quality by colour
## W = 3311514, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Data analysis related to question 2

```
## Testing if density, alcohol, sulfates and score exhibit multivariate no
rmal distribution

library(energy)
mvnorm.etest(winequality_merged[,1:4],R=200)

##
##  Energy test of multivariate normality: estimated parameters
##
## data:  x, sample size 6497, dimension 4, replicates 200
## E-statistic = 69.196, p-value < 2.2e-16

## Testing for correlations between density, alcohol, sulfates and score

library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units

corr_data<-as.matrix(winequality_merged[c(1:4)])
rcorr(corr_data, type="spearman")

##           density sulphates alcohol quality
## density      1.00      0.27   -0.70   -0.32
## sulphates    0.27      1.00    0.00    0.03
## alcohol     -0.70      0.00    1.00    0.45
## quality     -0.32      0.03    0.45    1.00
##
## n= 6497
##
##
## P
##           density sulphates alcohol quality
## density           0.0000    0.0000  0.0000
## sulphates 0.0000            0.7119  0.0162
## alcohol   0.0000  0.7119            0.0000
## quality   0.0000  0.0162    0.0000

## Fitting a multiple linear regression model

y<-winequality_merged[,4]
x<-winequality_merged[,2:3]
mlrmodel<-lm(y~.,x)
summary(mlrmodel)

##
## Call:
## lm(formula = y ~ ., data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4667 -0.4953 -0.0349  0.5072  3.2282
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.280158   0.092678  24.603  < 2e-16 ***
## sulphates   0.233749   0.065175   3.586 0.000338 ***
## alcohol     0.325400   0.008131  40.018  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7817 on 6494 degrees of freedom
## Multiple R-squared:  0.199,  Adjusted R-squared:  0.1988
## F-statistic: 806.7 on 2 and 6494 DF,  p-value: < 2.2e-16
```

```
## Testing if the residuals exhibit normal distribution

library(MASS)
mlrresiduals<-studres(mlrmodel)
library(nortest)
ad.test(mlrresiduals)

##
##  Anderson-Darling normality test
##
## data:  mlrresiduals
## A = 29.757, p-value < 2.2e-16

## Testing if the residuals are independent of each other

library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

dwtest(mlrmodel)

##
##  Durbin-Watson test
##
## data:  mlrmodel
## DW = 1.6245, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0

## Visualising the homoscedasticity of the residuals

plot(predict(mlrmodel),residuals(mlrmodel))
```



```r
## Testing the homoscedasticity of the residuals

library(lmtest)
bptest(mlrmodel)

##
##  studentized Breusch-Pagan test
##
## data:  mlrmodel
## BP = 23.419, df = 2, p-value = 8.215e-06

## Checking for outliers using Mahalanobis distances

m_dist<-mahalanobis(x, colMeans(x), cov(x))
cutof_mah<-qchisq(0.95, 2, lower.tail = TRUE, log.p = FALSE)
cutof_mah

## [1] 5.991465

out_mah<-which(m_dist>cutof_mah)
out_mah

##   [1]   14   15   16   18   20   23   28   43   44   70   80   82   84
87   89
##  [16]   92   93  107  111  115  129  143  145  152  162  170  182  198
199  202
##  [31]  211  227  241  246  250  259  268  269  270  272  277  278  279
282  290
##  [46]  336  339  340  341  347  348  349  351  354  362  366  370  372
373  377
##  [61]  378  379  391  416  424  431  435  439  445  452  456  466  468
475  478
##  [76]  482  483  484  485  489  492  493  502  503  504  505  507  516
521  523
##  [91]  545  571  587  589  615  618  624  640  653  690  693  724  755
774  775
## [106]  796  803  809  822  853  920  923  927  947  983 1052 1054 1071
1094 1099
## [121] 1115 1121 1127 1133 1151 1158 1159 1166 1167 1168 1208 1229 1261
1270 1271
## [136] 1289 1290 1320 1368 1371 1372 1373 1404 1406 1407 1408 1409 1410
1413 1414
## [151] 1430 1476 1478 1517 1523 1571 1589 2301 2358 2359 2452 2454 2466
2574 2616
## [166] 2699 2726 2838 2842 2843 2883 2893 2986 2994 3064 3190 3203 3341
3462 3552
## [181] 4003 4041 4194 4234 4237 4268 4348 4350 4393 4396 4472 4473 4474
4493 4517
## [196] 4522 4530 4545 4598 4608 4656 4657 4683 4686 4750 4806 4825 4844
4884 4891
## [211] 4901 4903 5058 5068 5076 5082 5083 5084 5103 5107 5114 5117 5119
```

```
5224 5255
## [226] 5271 5273 5276 5310 5328 5335 5336 5354 5364 5373 5451 5501 5504
5510 5515
## [241] 5516 5518 5531 5598 5599 5600 5612 5665 5729 5749 5795 5839 6001
6046 6062
## [256] 6091 6103 6145 6152 6160 6182 6196 6217 6245 6258 6281 6296 6356
6357 6389
## [271] 6392 6415 6418 6465 6486 6487
```

```
length(out_mah)
```

```
## [1] 276
```

```
## Checking for outliers using Leverages
```

```
cutof_lev<-2*3/(length(y))
cutof_lev
```

```
## [1] 0.0009235032
```

```
leverages<-as.data.frame(hatvalues(mlrmodel, type='rstandard'))
out_lev<-which(leverages>cutof_lev)
out_lev
```

```
##   [1]   14   15   16   18   20   23   28   43   44   70   80   82   84
87   89
##  [16]   92   93  107  111  115  129  143  145  152  162  170  182  198
199  202
##  [31]  210  211  227  241  244  245  246  250  259  265  268  269  270
272  277
##  [46]  278  279  282  290  336  339  340  341  342  347  348  349  350
351  354
##  [61]  357  362  364  366  370  372  373  376  377  378  379  391  416
424  431
##  [76]  435  439  445  452  456  466  468  475  478  482  483  484  485
489  492
##  [91]  493  502  503  504  505  506  507  516  521  523  531  536  545
571  587
## [106]  589  607  615  618  624  640  653  690  693  724  755  774  775
796  803
## [121]  806  808  809  822  829  833  834  853  897  899  911  920  923
926  927
## [136]  939  947  966  971  972  983 1003 1007 1008 1017 1039 1052 1054
1071 1094
## [151] 1099 1101 1108 1115 1119 1121 1127 1133 1147 1151 1158 1159 1166
1167 1168
## [166] 1193 1208 1210 1218 1229 1261 1268 1270 1271 1288 1289 1290 1320
1368 1371
## [181] 1372 1373 1403 1404 1406 1407 1408 1409 1410 1413 1414 1416 1430
1433 1476
## [196] 1478 1517 1523 1571 1586 1587 1589 2301 2357 2358 2359 2452 2454
```

```
2465 2466
## [211] 2467 2468 2479 2574 2616 2636 2699 2726 2772 2790 2828 2838 2842
2843 2880
## [226] 2883 2885 2893 2894 2920 2921 2933 2986 2987 2992 2994 3012 3026
3064 3190
## [241] 3203 3341 3407 3409 3414 3419 3422 3448 3462 3552 3595 3597 3598
3606 3657
## [256] 4003 4020 4041 4194 4234 4237 4252 4268 4280 4286 4348 4350 4372
4393 4396
## [271] 4414 4417 4472 4473 4474 4483 4489 4493 4517 4522 4526 4530 4531
4545 4559
## [286] 4584 4590 4598 4607 4608 4656 4657 4679 4683 4686 4722 4750 4752
4806 4807
## [301] 4825 4844 4884 4891 4901 4903 4967 4970 4973 5022 5036 5058 5068
5076 5082
## [316] 5083 5084 5099 5103 5107 5114 5116 5117 5119 5129 5139 5140 5224
5241 5242
## [331] 5255 5259 5264 5265 5271 5272 5273 5276 5310 5328 5335 5336 5354
5364 5373
## [346] 5385 5415 5429 5443 5451 5458 5501 5504 5507 5510 5515 5516 5518
5519 5522
## [361] 5531 5577 5598 5599 5600 5612 5665 5729 5749 5767 5795 5839 5903
5912 5950
## [376] 6001 6006 6009 6031 6046 6062 6080 6088 6089 6091 6103 6110 6145
6152 6160
## [391] 6182 6196 6217 6245 6258 6281 6296 6337 6356 6357 6386 6387 6389
6392 6402
## [406] 6415 6418 6437 6463 6465 6467 6486 6487

length(out_lev)

## [1] 413

## Checking for outliers using Cook's distances

cook<-cooks.distance(mlrmodel, type='rstandard')
which(cook>=1)

## named integer(0)
```

Data analysis related to question 3

```
## Running a chi-squared test on the acidity and colour variables

X2_data<-matrix(c(6159,338,1599,4898),nrow=2,byrow=TRUE)
colnames(X2_data) <- c("strongly acidic","weakly acidic")
rownames(X2_data) <- c("red","white")
chisq.test(X2_data,correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  X2_data
## X-squared = 6651.6, df = 1, p-value < 2.2e-16
```