

## **Case Study Aim**

The wine quality is affected by many factors; therefore, the study is aimed to investigate relationships between varied factors, and how some of them influence the final quality of the wine.

## **Objective**

1. Find out if high alcohol percentage raises the wine quality.
2. Investigate the effect of Residual Sugar on Density of wine.
3. To study the significance of alcohol, density, and chlorides on quality of white wine.

## **Statistical Procedures**

### **Summary Statistics: Precondition**

#### **Summary of features**

Values included: Minimum, maximum, quartile 1 (Q1), median, quartile 3 (Q3), mean, standard deviation (SD), variance, range, interquartile range (IQR)

```

> summary(white_wine)
fixed_acidity   volatile_acidity   citric_acid   residual_sugar   chlorides   free_sulfur_dioxide
Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600   Min.   :0.00900   Min.   : 2.00
1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200   Median :0.04300   Median : 34.00
Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800   Max.   :0.34600   Max.   :289.00
total_sulfur_dioxide   density   pH   sulphates   alcohol   quality
Min.   : 9.0   Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
1st Qu.:108.0   1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
Median :134.0   Median :0.9937   Median :3.180   Median :0.4700   Median :10.40   Median :6.000
Mean   :138.4   Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
3rd Qu.:167.0   3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
Max.   :440.0   Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
> # standard deviation of each features
> sd(fixed_acidity)
[1] 0.8438682
> sd(volatile_acidity)
[1] 0.1007945
> sd(citric_acid)
[1] 0.1210198
> sd(residual_sugar)
[1] 5.072058
> sd(chlorides)
[1] 0.02184797
> sd(free_sulfur_dioxide)
[1] 17.00714
> sd(total_sulfur_dioxide)
[1] 42.49806
> sd(density)
[1] 0.002990907
> sd(pH)
[1] 0.1510006
> sd(sulphates)
[1] 0.1141258
> sd(alcohol)
[1] 1.230621
> sd(quality)
[1] 0.8856386

```

Features	fixed_acidity	volatile_acidity	citric_acid	residual_sugar
Minimum	3.800	0.080	0.000	0.600
Q1	6.3	0.21	0.27	1.7
Median	6.8	0.26	0.32	5.2
Mean	6.855	0.2782	0.3342	6.391
Q3	7.3	0.32	0.39	9.9
Maximum	14.2	1.1	1.66	65.8
SD	0.8438682	0.1007945	0.1210198	5.072058
Variance	0.7121136	0.01015954	0.01464579	25.72577
Range	10.4	1.02	1.66	65.2
IQR	1	0.11	0.12	8.2

Features	chlorides	free_sulphur_d ioxide	total_sulphur_ dioxide	density
Minimum	0.009	2	9.0	0.9871
Q1	0.036	23	108	0.9917
Median	0.043	34	134	0.9937

<b>Mean</b>	0.04577	35.31	138.4	0.994
<b>Q3</b>	0.05	46	167.0	0.9961
<b>Maximum</b>	0.346	289	440	1.039
<b>SD</b>	0.02184797	17.00714	42.49806	0.002990907
<b>Variance</b>	0.000477334	289.2427	1806.085	8.94552e-06
<b>Range</b>	0.337	287	431	0.05187
<b>IQR</b>	0.014	23	59	0.0043775

<b>Features</b>	<b>ph</b>	<b>sulphates</b>	<b>alcohol</b>	<b>quality</b>
<b>Minimum</b>	2.72	0.22	8	3
<b>Q1</b>	3.09	0.41	9.5	5
<b>Median</b>	3.18	0.47	10.4	6
<b>Mean</b>	3.188	0.4898	10.51	5.878
<b>Q3</b>	3.28	0.55	11.40	6
<b>Maximum</b>	3.82	1.08	14.20	9
<b>SD</b>	0.1510006	0.1141258	1.230621	0.8856386
<b>Variance</b>	0.0228012	0.01302471	1.514427	0.7843557
<b>Range</b>	1.1	0.86	6.2	6
<b>IQR</b>	0.19	0.14	1.9	1

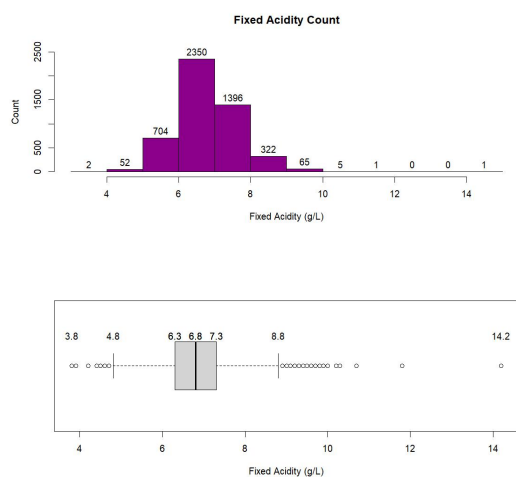
## Graphs

### Bar Plots and Box plots

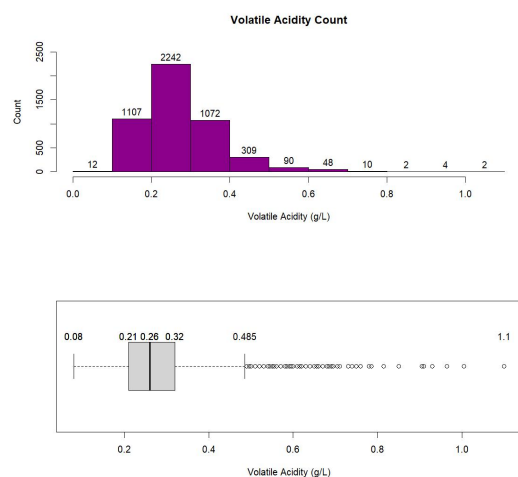
The plots show pattern of each feature, most of them are skewed to the right with limited values, such as fixed acidity, volatile acidity, citric acid, chlorides, free sulphur dioxide, total sulphur dioxide, density, and pH values. Meanwhile, features with heavy right-skewed are residual sugar and alcohol, therefore these two are not normally distributed nor symmetrical (Siegel, 2012)

The distribution of quality feature is a special case as its median and third quantile share the same value of 6, which does not match the characteristic of normal distribution. Although the left and right whiskers show the same gap, it is a non-symmetrical pattern.

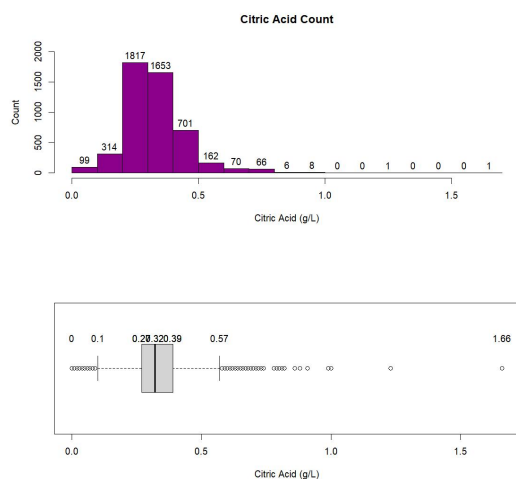
## Fixed Acidity



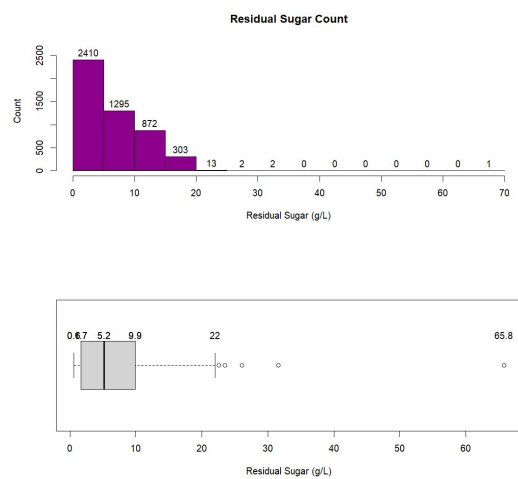
## Volatile acidity



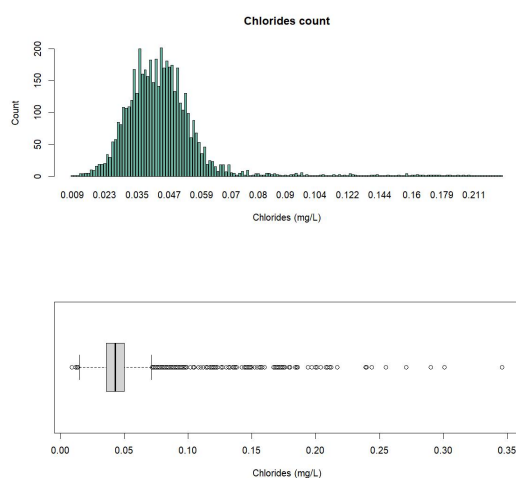
## Citric acid



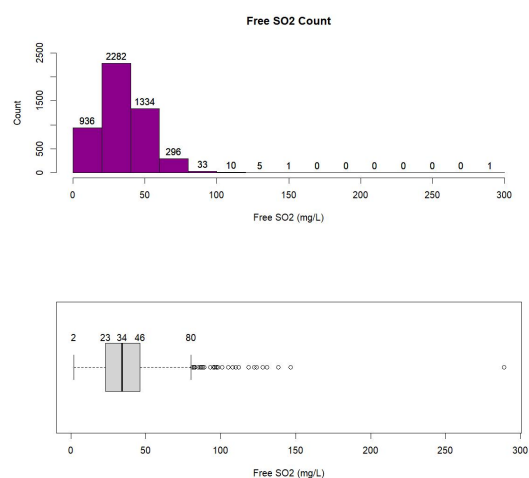
## Residual sugar



## Chlorides

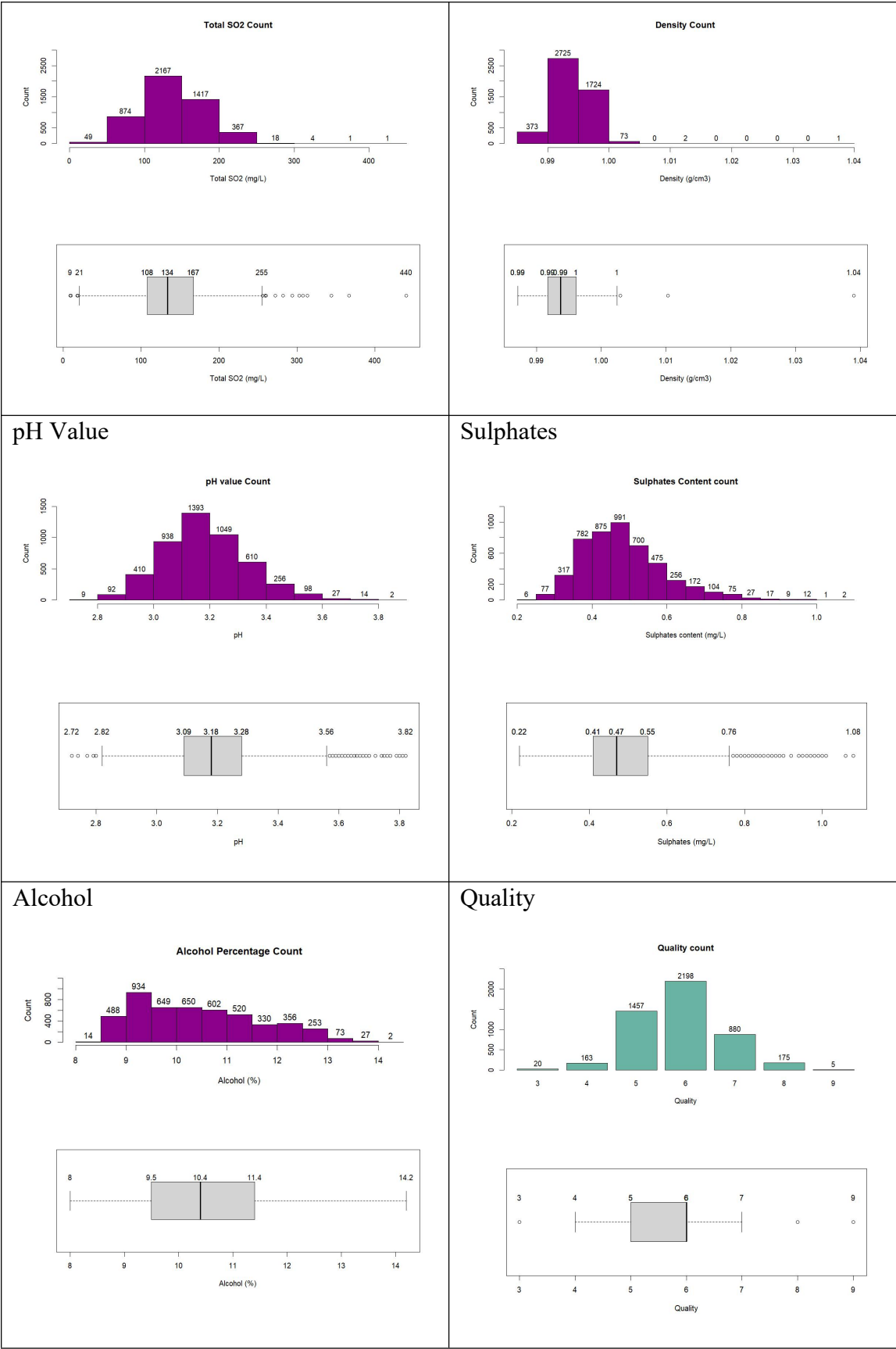


## Free sulphur dioxide



## Total sulphur dioxide

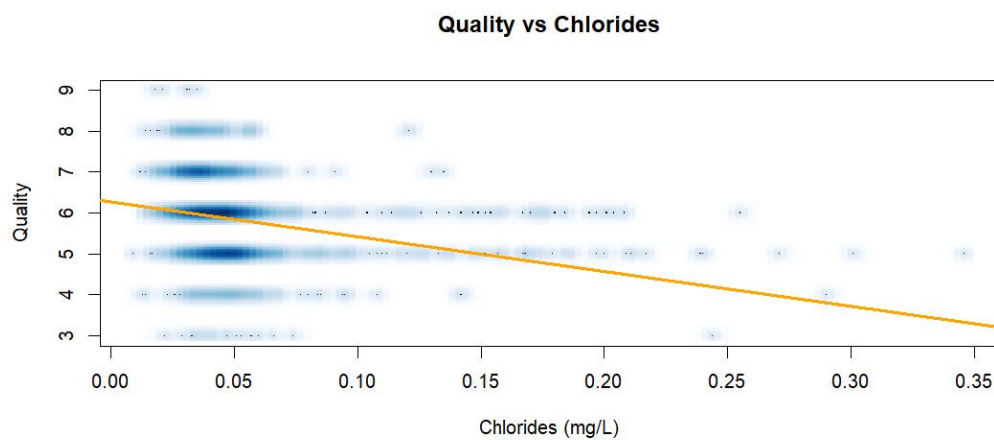
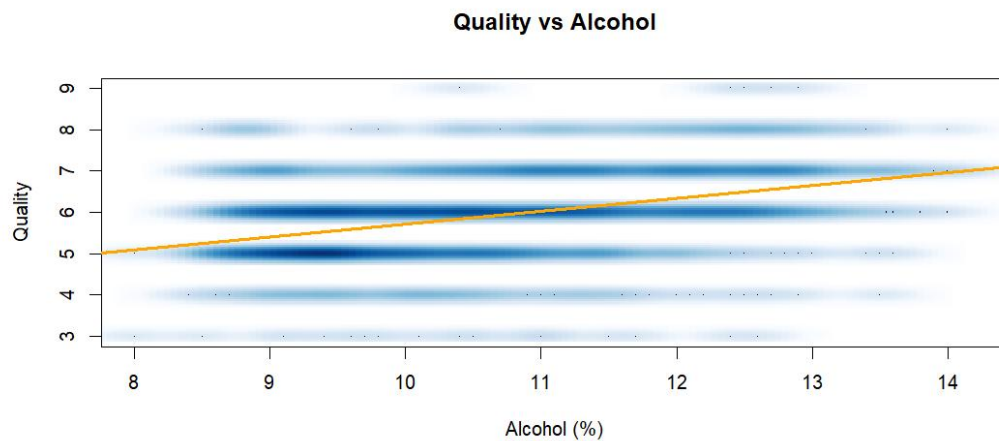
## Density

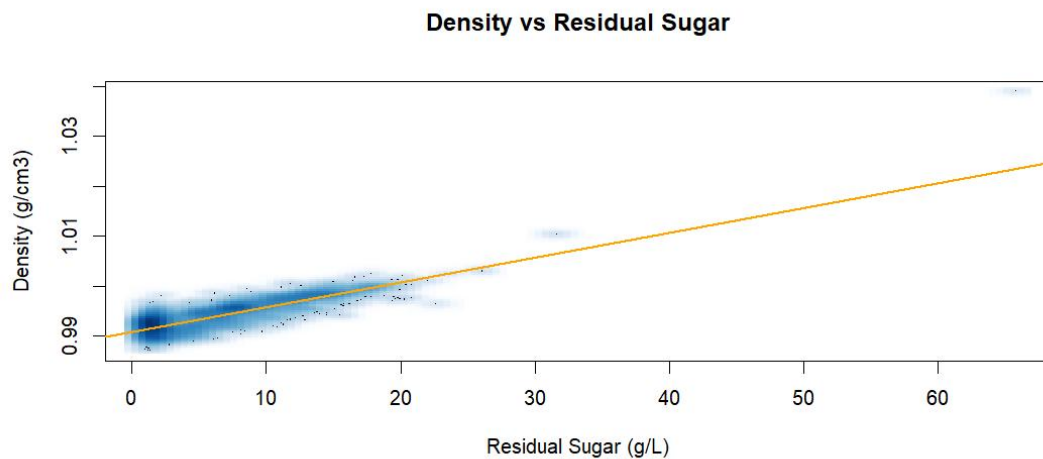


Scatter Plot

Before the study begins, scatter plots for Quality versus Alcohol and Density versus Residual Sugar are plotted to visualize their relationship.

The graph shows an increasing trend in quality as the alcohol percentage goes higher, same goes for density and residual sugar amount. Meanwhile, a decreasing pattern in quality and chlorides content, where the quality is higher at lower number of chlorides.





## Two Sample z-test: Wine quality vs Alcohol Percentage

### Requirement/Assumptions

According to The Open Educator, 2024

- Population sample size is large,  $n > 30$
- Standard deviation of population is known
- Samples are normally distributed

### Modelling

Since the quality variable is not normally distributed, which does not fit one of the assumptions above. This could pose the risk of inaccuracy and misleading results. Hence, data normalization, a data preprocessing procedure is needed to be carried out in order to provide a set of clean data through remodeling them into a standard scale (GeeksforGeeks, 2021).

There are 2 ways to transform the non-normal distributed data to be closer to the bell-shaped curve: transformation with square root or log (Sainani, 2012). There method taken to normalize these features is by using natural log transformation as it works only on positive numbers which match Wine Quality columns. However, the downside of this method is the challenges in result interpretation as the log-transformed data is not intuitive for users to understand (Esteban, 2023).

If the distribution after log transformation remains not symmetrical, the initial values would be taken for calculation of z-score and p-value. The Shapiro-Wilk test

can also be used to support the observation of normality, where the test measures how well the sample quantiles fit a standard normal quantile (King & Eckersley, 2019).

Hypothesis

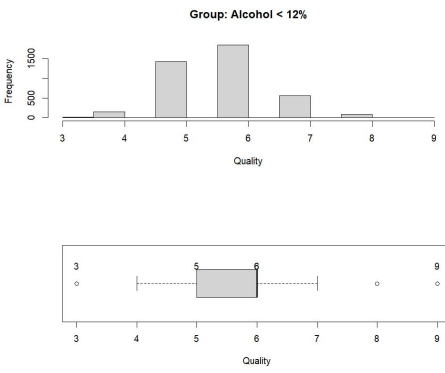
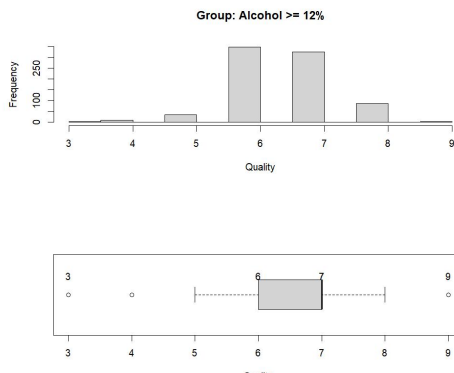
- $H_0$ : Wine with high alcohol percentage of greater than or equals to 12% do not have higher mean quality than alcohol percentage lower than 12%.
- $H_1$ : Wine with high alcohol percentage of greater than or equals to 12% have higher mean quality than alcohol percentage lower than 12%.

Steps

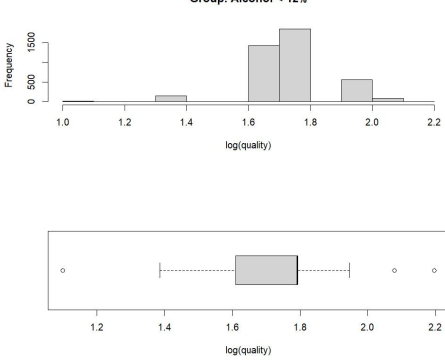
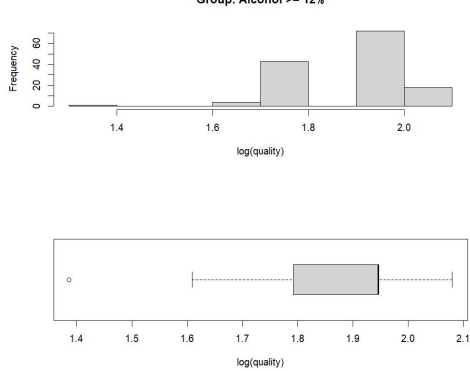
To carry out the hypothesis testing:

1. Separate the quality into 2 groups by the alcohol percentage of at least 12% and less than 12%.
2. Save the values as new variable of each group after log transformation.
3. Calculate mean and variance for the 4 set of data:
  - a) Before transformation: at least 12%, less than 12%
  - b) After transformation: at least 12%, less than 12%
4. Plot histogram and box plot for each of them to compare the distribution patterns before and after log transformation.
5. Conduct Shapiro-Wilk test
6. Calculate z-score and p-value for each group

Before and after normalization for two groups:

	Quality with alcohol < 12 %	Quality with alcohol >= 12 %
Before		



After	<p>Group: Alcohol &lt; 12%</p>  <ul style="list-style-type: none"> <li>The outliers were not removed. Distribution is not symmetrical, although range of data shifted right, landed on the middle of scale.</li> </ul>	<p>Group: Alcohol &gt;= 12%</p>  <ul style="list-style-type: none"> <li>The number of outliers reduced to 1 from 3.</li> <li>The range spread wider and shifted to the right, but distribution is not normally distributed.</li> </ul>
Shapiro-Wilk test	<p>Not normalized: W = 0.87399, p-value &lt; 2.2e-16</p> <p>Normalized: W = 0.86603, p-value &lt; 2.2e-16</p>	<p>Not normalized: W = 0.85865, p-value &lt; 2.2e-16</p> <p>Normalized: W = 0.81877, p-value = 9.066e-12</p>
Conclusion	Both are not normal as p-value < 0.05	Both are not normal as p-value < 0.05

The formula for z-score is calculated by:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

# Correlation Analysis with Pearson correlation: Residual Sugar vs Density

## Requirement / Assumptions

According to Lani (2013) and Bobbitt, (2021), the assumptions for Pearson correlation are as follow:

- Variables are continuous
- Outliers are absent
- Variables are normally distributed
- The variables have linear relationship
- The observation values come in a pair

## Modelling

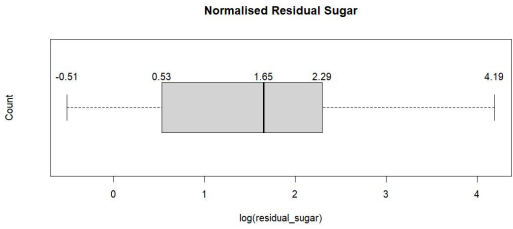
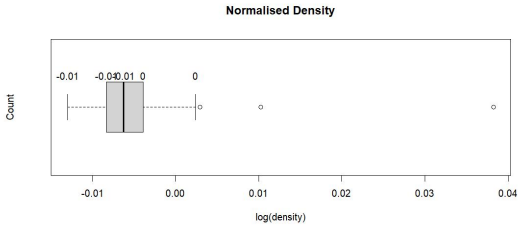
### Hypothesis

- $H_0$ : Residual sugar content is not significant to the density of wine.
- $H_1$ : Residual sugar content is significant to density of wine.

### Steps

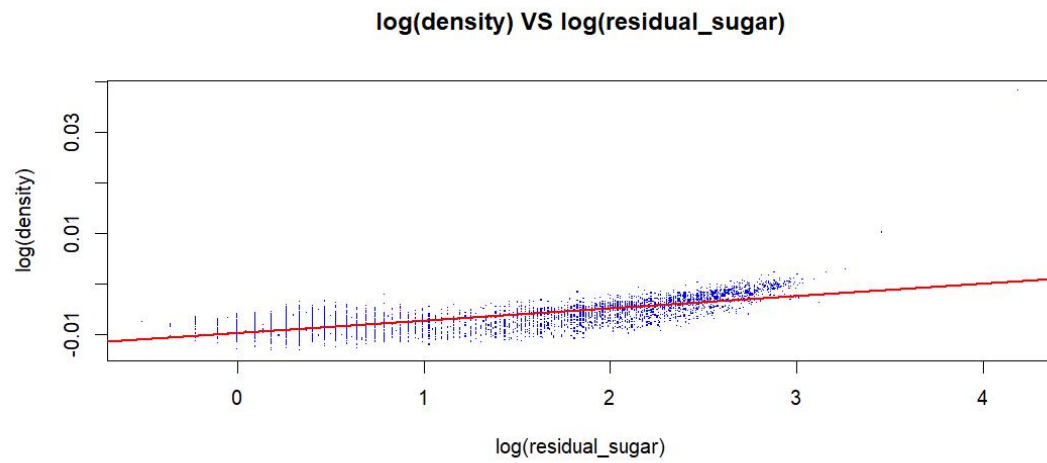
The residual sugar content and wine density box plot showed that the variables are not normally distributed and have outliers, therefore normalization process using log transformation is needed.

After normalization:

<p>Residual Sugar</p> 	<p>Density</p> 
<p>The range span from -0.51 to 4.19. Although the plot is not symmetrical, all outliers are removed. Hence, overall improvement is made.</p>	<p>There is difference in before and after log transformation is not clear, as outliers are still present, and the overall plot remains right skewed.</p>

With the new values after applying log transformation, correlation test can be run to get the p-value and correlation coefficient.

The scatter plot and line of best fit after applying the log transformation:



# Multiple linear regression model: Quality vs Alcohol Percentage, Density, and Chlorides

## Requirement / Assumptions

- Linear relationship between the responding variable and the independent variables.
- Difference between predicted and observed values are normally distributed
- Multicollinearity does not happen
- Variance of errors is consistent across all levels of independent variables.

(Assumptions of Multiple Linear Regression, n.d.)

## Modelling

### Steps

- $H_0$ : Alcohol, density, and chlorides are not sufficient predictors of white wine quality.
- $H_1$ : Alcohol, density, and chlorides are sufficient predictors of white wine quality.

### Steps

To ensure the collinearity among independent variables does not occur, a correlation matrix is calculated to give a full view of correlation coefficients between all variables. As the heat map shows, no coefficients value of more than 0.8 among the independent variables: alcohol, density, and chlorides.



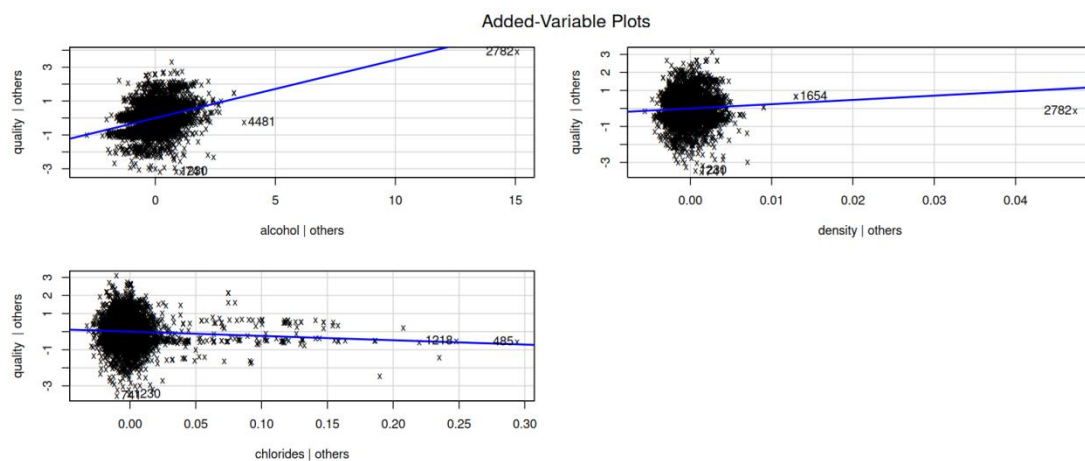
The Variance Inflation Factor (VIF) is used for detecting multicollinearity by measuring correlation and the strength between independent variables in the regression model (Bobbitt, 2019). The definition of VIF is as follow (Variance Inflation Factor (VIF), 2024):

- VIF = 1: no correlation between variables
- $1 < \text{VIF} < 5$ : moderate correlation
- $\text{VIF} > 5$ : highly correlated

From the test, result shows all independent variables included in the regression are moderately correlated with one another. Hence, the variables are less likely to cause interpretation and fitting issues in the MLR model.

```
> vif(model)
alcohol    density chlorides
2.746249   2.559278   1.150988
```

Added variable plots is created to visualize the relationship between multiple variables



Model for quality estimation based on alcohol percentage, density and chlorides is:

$$\text{quality} = b_0 + b_1 * \text{alcohol} + b_2 * \text{density} + b_3 * \text{chlorides}$$

Confidence Interval of MLR model coefficient at 95%

```
> confint(model)
                2.5 %      97.5 %
(Intercept) -33.2308382 -9.0694819
alcohol      0.3131389  0.3730946
density     11.7636320 35.5781015
chlorides    -3.4754054 -1.2891068
```

# Results

## 2 samples z-test

```
> # Before Log
> z_score <- (mean_high-mean_low)/sqrt((var_low/length(low_pct))+(var_high/length(high_pct)))
> z_score
[1] 25.83832
> p_val <- 2*pnorm(q=z_score, lower.tail=FALSE)
> p_val
[1] 3.291514e-147
>
```

The z-score obtained is 25.83832, and p-value is 3.291514e-147, that is much lower than the significance level of 0.05. Hence, the null hypothesis is rejected and concludes that wine with alcohol percentage of 12% and above has higher mean quality than those lower than 12%.

## Pearson correlation analysis

The values from calculations:

Before normalization	After normalization
<p>Pearson's product-moment correlation</p> <p>data: density and residual_sugar t = 107.87, df = 4896, p-value &lt; 2.2e-16 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: 0.8304732 0.8470698 sample estimates: cor 0.8389665</p>	<p>Pearson's product-moment correlation</p> <p>data: log(residual_sugar) and log(density) t = 81.004, df = 4896, p-value &lt; 2.2e-16 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: 0.7445333 0.7684792 sample estimates: cor 0.75676</p>
<pre>&gt; coef(fit) (Intercept) residual_sugar 0.9908653878 0.0004947244</pre>	<pre>&gt; coef(fit) (Intercept) log(residual_sugar) -0.009648201 0.002466801</pre>

Strong positive correlation is shown in the correlation coefficient value of both normalized, with value of 0.75676, and non-normalized data, of value 0.8389665. The p-value is 2.2e-16, which is smaller than the significant interval at 5%. Hence, the null hypothesis is rejected and concludes that the residual sugar content is significant to wine density, that means changes in the amount of residual sugar would influence wine density.

## Multiple linear regression model

```
Call:
lm(formula = quality ~ alcohol + density + chlorides, data = white_wine)

Residuals:
    Min       1Q   Median       3Q      Max
-3.5904 -0.5209 -0.0050  0.4832  3.0653

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.15016     6.16220  -3.432 0.000604 ***
alcohol       0.34312     0.01529  22.439 < 2e-16 ***
density      23.67087     6.07373   3.897 9.86e-05 ***
chlorides    -2.38226     0.55760  -4.272 1.97e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7946 on 4894 degrees of freedom
Multiple R-squared:  0.1955,    Adjusted R-squared:  0.195
F-statistic: 396.3 on 3 and 4894 DF,  p-value: < 2.2e-16
```

The p-value of the model is 2.2e-16, which displayed an extremely high significance that at least one of the independent variables highly influence the response variable. The p-value of each independent variable is displayed under Pr(>|t|) column and shows that all predictors are significantly affect quality of wine as the values are much lower than 0.05. Hence, the null hypothesis is rejected, and concludes that alcohol, density, and chlorides are sufficient predictors of white wine quality. The model equation is denoted as:

$$\text{quality} = -21.15 + 0.34 * \text{alcohol} + 23.67 * \text{density} - 2.38 * \text{chlorides}$$

1619 words

## Reference

1. The. (2024). The Open Educator - 7. Two Sample Z-Test. Theopeneducator.com.  
<https://www.theopeneducator.com/doi/hypothesis-Testing-Inferential-Statistics-Analysis-of-Variance-ANOVA/Two-Sample-Z-Test>
2. Right Skewed vs. Left Skewed Distribution. (2024). Investopedia.  
<https://www.investopedia.com/terms/s/skewness.asp>
3. Siegel, A. F. (2012). Histograms. Elsevier EBooks, 35–64.  
<https://doi.org/10.1016/b978-0-12-385208-3.00003-1>
4. Sainani, K. L. (2012). Dealing With Non-normal Data. PM&R, 4(12), 1001–1005.  
<https://doi.org/10.1016/j.pmrj.2012.10.013>

5. GeeksforGeeks. (2021, December 16). How to Normalize Data in R?  
GeeksforGeeks; GeeksforGeeks. <https://www.geeksforgeeks.org/how-to-normalize-data-in-r/>
6. Esteban, J. (2023, May 9). Best Tips and Tricks: When and Why to Use Logarithmic Transformations in Statistical Analysis. Medium; Medium.  
<https://juandelacalle.medium.com/best-tips-and-tricks-when-and-why-to-use-logarithmic-transformations-in-statistical-analysis-9f1d72e83cfc#:~:text=Logarithmic%20transformations%2C%20while%20powerful%20tools,and%20negative%20numbers%20is%20undefined.>
7. Scatter plot in R. (2020, April 22). RCODER. <https://r-coder.com/scatter-plot-r/#:~:text=You%20can%20create%20a%20scatter,matrix%2C%20with%20the%20pairs%20function.&text=In%20addition%2C%20in%20case%20your,the%20group%20with%20different%20color.>
8. R: Kolmogorov-Smirnov Tests. (2024). Mit.edu.  
[https://web.mit.edu/~r/current/arch/i386\\_linux26/lib/R/library/stats/html/ks.test.html](https://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/stats/html/ks.test.html)
9. Assumptions of Multiple Linear Regression. (n.d.). Statistics Solutions.  
<https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-multiple-linear-regression/>
10. Bobbitt, Z. (2020, December 23). How to Plot Multiple Linear Regression Results in R. Statology. <https://www.statology.org/plot-multiple-linear-regression-in-r/>
11. Multiple Linear Regression in R - Articles - STHDA. (2018, March 10).  
Sthda.com. <http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/>
12. King, A. P., & Eckersley, R. J. (2019). Inferential Statistics IV: Choosing a Hypothesis Test. Elsevier EBooks, 147–171. <https://doi.org/10.1016/b978-0-08-102939-8.00016-5>



## Appendix

```
# preparation for reading the csv file ----

# get directory
getwd()

# set path variable
path <- "/home/winequality-white.csv"

# read CSV file
white_wine <- read.csv(path, header=TRUE, sep=";")
white_wine

# get a list of column names
features <- colnames(white_wine)
features

# show first 5 rows
head(white_wine, 5)

# rename the columns by replacing '.' to '_' ----
colnames(white_wine)[1] <- "fixed_acidity"
colnames(white_wine)[2] <- "volatile_acidity"
colnames(white_wine)[3] <- "citric_acid"
colnames(white_wine)[4] <- "residual_sugar"
colnames(white_wine)[5] <- "chlorides"
colnames(white_wine)[6] <- "free_sulfur_dioxide"
colnames(white_wine)[7] <- "total_sulfur_dioxide"
colnames(white_wine)[8] <- "density"
colnames(white_wine)[9] <- "pH"
colnames(white_wine)[10] <- "sulphates"
colnames(white_wine)[11] <- "alcohol"
colnames(white_wine)[12] <- "quality"

fixed_acidity <- white_wine$fixed_acidity
volatile_acidity <- white_wine$volatile_acidity
citric_acid <- white_wine$citric_acid
residual_sugar <- white_wine$residual_sugar
chlorides <- white_wine$chlorides
free_sulfur_dioxide <- white_wine$free_sulfur_dioxide
total_sulfur_dioxide <- white_wine$total_sulfur_dioxide
density <- white_wine$density
pH <- white_wine$pH
sulphates <- white_wine$sulphates
```

```

alcohol <- white_wine$alcohol
quality <- white_wine$quality

# descriptive statistic section ----
# see the summary of each column
summary(white_wine)

# get range, variance, standard deviation, IQR
now <- alcohol
var(now)
sd(now)
max(now) - min(now)
IQR(now)

# histogram, boxplots -----

# fixed_acidity
# Plot histogram and label text
h <- hist(fixed_acidity,
          main="Fixed Acidity Count",
          xlab="Fixed Acidity (g/L)",
          ylab="Count",
          ylim = c(0, 2500),
          col="darkmagenta",
          freq=TRUE)
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))

# Plot boxplot and label left/right whiskers, median Q1 /Q3
boxplot(
  x=fixed_acidity,
  xlab = "Fixed Acidity (g/L)",
  data=white_wine,
  horizontal = TRUE,
  axes = TRUE)
text(x=fivenum(fixed_acidity), labels =fivenum(fixed_acidity), y=1.25)
text(x=boxplot.stats(fixed_acidity)$stats, labels =boxplot.stats(fixed_acidity)$stats,
y=1.25)

# volatile_acidity
# Plot histogram and label text
h <- hist(volatile_acidity,
          main="Volatile Acidity Count",
          xlab="Volatile Acidity (g/L)",
          ylab="Count",
          ylim = c(0, 2500),

```

```

        col="darkmagenta",
        freq=TRUE)
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))

# Plot boxplot and label left/right whiskers, median Q1 /Q3
boxplot(
  x=volatile_acidity,
  xlab = "Volatile Acidity (g/L)",
  data=white_wine,
  horizontal = TRUE,
  axes = TRUE)
text(x=fivenum(volatile_acidity), labels =fivenum(volatile_acidity), y=1.25)
text(x=boxplot.stats(volatile_acidity)$stats, labels
=boxplot.stats(volatile_acidity)$stats, y=1.25)

# citric_acid
# Plot histogram and label text
h <- hist(citric_acid,
  main="Citric Acid Count",
  xlab="Citric Acid (g/L)",
  ylab="Count",
  ylim = c(0, 2000),
  col="darkmagenta",
  freq=TRUE)
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))

# Plot boxplot and label left/right whiskers, median Q1 /Q3
boxplot(
  x=citric_acid,
  xlab = "Citric Acid (g/L)",
  data=white_wine,
  horizontal = TRUE,
  axes = TRUE)
text(x=fivenum(citric_acid), labels =fivenum(citric_acid), y=1.25)
text(x=boxplot.stats(citric_acid)$stats, labels = boxplot.stats(citric_acid)$stats,
y=1.25)

# residual_sugar
# Plot histogram and label text
h <- hist(residual_sugar,
  main="Residual Sugar Count",
  xlab="Residual Sugar (g/L)",
  ylab="Count",
  ylim = c(0, 2600),
  col="darkmagenta",

```

```

      freq=TRUE)
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))

# Plot boxplot and label left/right whiskers, median Q1 /Q3
boxplot(
  x=residual_sugar,
  xlab = "Residual Sugar (g/L)",
  data=white_wine,
  horizontal = TRUE,
  axes = TRUE)
text(x=fivenum(residual_sugar), labels =fivenum(residual_sugar), y=1.25)
text(x=boxplot.stats(residual_sugar)$stats, labels =
boxplot.stats(residual_sugar)$stats, y=1.25)

# chlorides
# Plot bar chart and label text
ch_tab <- table(chlorides)
ch_bar <- barplot(
  ch_tab,
  xlab="Chlorides (mg/L)",
  ylab="Count",
  col="#69b3a2",
  main = "Chlorides count")

# Plot boxplot and label left/right whiskers, median Q1 /Q3
boxplot(
  x=chlorides,
  xlab = "Chlorides (mg/L)",
  data=white_wine,
  horizontal = TRUE,
  axes = TRUE)
text(x=fivenum(chlorides), labels =fivenum(chlorides), y=-1.25)
text(x=boxplot.stats(chlorides)$stats, labels = boxplot.stats(chlorides)$stats, y=1.25)

# free_sulfur_dioxide
# Plot histogram and label text
h <- hist(free_sulfur_dioxide,
  main="Free SO2 Count",
  xlab="Free SO2 (mg/L)",
  ylab="Count",
  ylim = c(0, 2500),
  col="darkmagenta",
  freq=TRUE)
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))

```

```

# Plot boxplot and label left/right whiskers, median Q1 /Q3
boxplot(
  x=free_sulfur_dioxide,
  xlab = "Free SO2 (mg/L)",
  data=white_wine,
  horizontal = TRUE,
  axes = TRUE)
text(x=boxplot.stats(free_sulfur_dioxide)$stats, labels =
boxplot.stats(free_sulfur_dioxide)$stats, y=1.25)

# total_sulfur_dioxide
# Plot histogram and label text
h <- hist(total_sulfur_dioxide,
  main="Total SO2 Count",
  xlab="Total SO2 (mg/L)",
  ylab="Count",
  ylim = c(0, 3000),
  col="darkmagenta",
  freq=TRUE)
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))

# Plot boxplot and label left/right whiskers, median Q1 /Q3
boxplot(
  x=total_sulfur_dioxide,
  xlab = "Total SO2 (mg/L)",
  data=white_wine,
  horizontal = TRUE,
  axes = TRUE)
text(x=quantile(total_sulfur_dioxide), labels =quantile(total_sulfur_dioxide), y=1.25)
text(x=boxplot.stats(total_sulfur_dioxide)$stats, labels =
boxplot.stats(total_sulfur_dioxide)$stats, y=1.25)

# density
# Plot histogram and label text
h <- hist(density,
  main="Density Count",
  xlab="Density (g/cm3)",
  ylab="Count",
  ylim = c(0, 3000),
  col="darkmagenta",
  freq=TRUE)
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))

# Plot boxplot and label left/right whiskers, median Q1 /Q3
boxplot(

```

```

x=density,
xlab = "Density (g/cm3)",
data=white_wine,
horizontal = TRUE,
axes = TRUE)
text(x=quantile(density), labels =quantile(round(density,2)), y=1.25)
text(x=boxplot.stats(density)$stats, labels = round(boxplot.stats(density)$stats, 2),
y=1.25)

```

```

# pH
# Plot histogram and label text
h <- hist(pH,
          main="pH value Count",
          xlab="pH",
          ylab="Count",
          ylim = c(0, 1500),
          col="darkmagenta",
          freq=TRUE)
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))
abline(v = mean(pH), col = "red", lwd = 2)

```

```

# Plot boxplot and label left/right whiskers, median Q1 /Q3
boxplot(
  x=pH,
  xlab = "pH",
  data=white_wine,
  horizontal = TRUE,
  axes = TRUE)
text(x=quantile(pH), labels =quantile(pH), y=1.25)
text(x=boxplot.stats(pH)$stats, labels = round(boxplot.stats(pH)$stats, 2), y=1.25)

```

```

# Sulphates
# Plot histogram and label text
h <- hist(sulphates,
          main="Sulphates Content count",
          xlab="Sulphates content (mg/L)",
          ylab="Count",
          ylim = c(0, 1200),
          col="darkmagenta",
          freq=TRUE)
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))

# Plot boxplot and label left/right whiskers, median Q1 /Q3
boxplot(
  x=sulphates,

```

```

xlab = "Sulphates (mg/L)",
data=white_wine,
horizontal = TRUE)
text(x=quantile(sulphates), labels =quantile(round(sulphates,2)), y=1.25)
text(x=boxplot.stats(sulphates)$stats, labels = round(boxplot.stats(sulphates)$stats,
2), y=1.25)

```

# Alcohol

# Plot histogram and label text

```

h <- hist(alcohol,
          main="Alcohol Percentage Count",
          xlab="Alcohol (%)",
          ylab="Count",
          ylim = c(0, 1200),
          col="darkmagenta",
          freq=TRUE)
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))

```

# Plot boxplot and label left/right whiskers, median Q1 /Q3

```

boxplot(
  x=alcohol,
  xlab = "Alcohol (%)",
  data=white_wine,
  horizontal = TRUE,
  main = "Alcohol",
  axes = TRUE)
text(x=fivenum(alcohol), labels =fivenum(alcohol), y=1.25)

```

# Quality

# Plot barchart and label text

```

quality_tab <- table(quality)
quality_bar <- barplot(
  quality_tab,
  ylim = c(0, 2500),
  xlab="Quality",
  ylab="Count",
  col="#69b3a2",
  main = "Quality count")

```

```

text(x = quality_bar,
     y = quality_tab + 0.5,
     labels = quality_tab,
     adj=c(0.5, -0.5))

```

# Plot boxplot and label left/right whiskers, median Q1 /Q3

```

boxplot(
  x=quality,
  xlab = "Quality",
  data=white_wine,
  horizontal = TRUE,
  axes = TRUE)
text(x=quantile(quality), labels =quantile(quality), y=1.25)
text(x=boxplot.stats(quality)$stats, labels = round(boxplot.stats(quality)$stats, 2),
y=1.25)

```

```

# ggplot: heat map -----

```

```

# Load the required packages

```

```

library(ggplot2)

```

```

library(corrplot)

```

```

library(ggcorrplot)

```

```

library(lattice)

```

```

library(reshape2)

```

```

# correlation matrix: figures

```

```

cor_matrix <- round(cor(white_wine),2)

```

```

melted_corr_mat <- melt(cor_matrix)

```

```

# create a basic correlation heatmap using ggplot

```

```

ggplot(
  data <- melted_corr_mat, aes(x=Var1, y=Var2, fill=value) +
  geom_tile(color="white") +
  labs(x="", y="", title="Heatmap of features for wine quality") +
  theme(axis.text.x = element_text(angle=45,hjust=1)) +
  geom_text(aes(label=value),color="white", size=3)
)

```

```

# scatter plot -----

```

```

# alcohol-quality

```

```

smoothScatter(alcohol,
  quality,
  main="Quality vs Alcohol",
  xlab="Alcohol (%)",
  ylab="Quality",)
abline(lm(quality ~ alcohol, data = white_wine), col = "orange",lwd = 3)

```

```

# Quality-chlorides

```

```

smoothScatter(chlorides,
  quality,
  main="Quality vs Chlorides",
  xlab="Chlorides (mg/L)",

```



```

        ylab="Quality"),
abline(lm(quality ~ chlorides, data = white_wine), col = "orange",lwd = 3)

# residual sugar – density
smoothScatter(
  residual_sugar,
  density,
  main = "Density vs Residual Sugar",
  xlab="Residual Sugar (g/L)",
  ylab="Density (g/cm3)")
abline(lm(density ~ residual_sugar,
  data = white_wine),
  col = "orange",
  lwd=2)

# models -----
# z test – subset data -----

# low alcohol subset
low_pct <- white_wine[alcohol < 12, "quality"]

# high alcohol subset
high_pct <- white_wine[white_wine$alcohol >= 12,"quality"]

# z test: no normalisation ----
# Low alcohol %
var_low<-var(low_pct) # variance of low alcohol
mean_low<-mean(low_pct) # mean of low alcohol

hist_low <- hist(low_pct, # Plot histogram
  main="Group: Alcohol < 12%",
  xlab = "Quality",
  freq=TRUE)

box_low <- boxplot(low_pct, # Plot boxplot
  horizontal=TRUE,
  xlab = "Quality")

# High alcohol %
var_high<-var(high_pct) # variance of high alcohol
mean_high<-mean(high_pct) # mean of high alcohol

hist_high <- hist(high_pct, # Plot histogram
  main="Group: Alcohol >= 12%",
  xlab = "Quality")

```

```

box_high <- boxplot(high_pct, # Plot boxplot
  horizontal=TRUE,
  xlab = "Quality")
text(x=fivenum(high_pct), labels =fivenum(high_pct), y=1.25) # values of quantiles

shapiro.test(low_pct) # Shapiro-Wilk test to check normality

# Calculate z-score for data with no normalisation
z_score <-(mean_high-
mean_low)/sqrt((var_low/length(low_pct))+(var_high/length(high_pct)))
z_score

# Calculate p-value
p_val <- 2*pnorm(q=z_score, lower.tail=FALSE)
p_val

# z test: after normalisation ----
# Low alcohol %
a_less_12 <- log(low_pct) # log transform low alcohol group
mean_less_12 <- mean(a_less_12) # mean of logged low alcohol group
var_less_12 <- var(a_less_12) # variance of logged low alcohol group

hist_log_low <- hist(a_less_12, # Plot histogram
  main="Group: Alcohol < 12%",
  xlab = "log(quality)")

box_log_low <- boxplot(a_less_12, # Plot boxplot
  horizontal=TRUE,
  xlab = "log(quality)")
shapiro.test(a_less_12) # Shapiro-Wilk test to check normality

# high alcohol %
a_more_12 <- log(high_pct) # log transform high alcohol group
mean_more_12 <- mean(a_more_12) # mean of logged high alcohol group
var_more_12 <- var(a_more_12) # variance of logged high alcohol group

hist_log_high <- hist(a_more_12, # Plot histogram
  main="Group: Alcohol >= 12%",
  xlab = "log(quality)")
box_log_high <- boxplot(a_more_12, # Plot boxplot
  horizontal=TRUE,
  xlab = "log(quality)")
shapiro.test(a_more_12) # Shapiro-Wilk test to check normality

```

```

# Calculate z-score for data with after normalisation
z_log_score <- (mean(a_more_12) -
mean(a_less_12)) / sqrt((var(a_more_12) / length(a_more_12)) + (var(a_less_12) / length(a_
less_12)))
z_log_score

# Calculate p-value
p_val_log <- 2 * pnorm(q=z, lower.tail=FALSE)
p_val_log

# Pearson correlation analysis ----

# log(residual_sugar): Box plot to see the pattern
boxplot(log(residual_sugar), horizontal=TRUE,
        xlab="log(residual_sugar)",
        ylab="Count",
        main="Normalised Residual Sugar")
text(x=boxplot.stats(log(residual_sugar))$stats,
     labels = round(boxplot.stats(log(residual_sugar))$stats, 2),
     y=1.25)

# log(density): Box plot to see the pattern
boxplot(log(density), horizontal=TRUE,
        xlab="log(density)",
        ylab="Count",
        main="Normalised Density")
text(x=boxplot.stats(log(density))$stats,
     labels = round(boxplot.stats(log(density))$stats, 2),
     y=1.25)

# Scatter plot of normalised residual_sugar vs density
plot(x=log(residual_sugar), log(density),
     main="log(density) VS log(residual_sugar)",
     pch='.',
     col="blue")
abline(fit <- lm(log(density) ~ log(residual_sugar), data = white_wine),
      col = "red",
      lwd=2)

# get the intercept and gradient
coef(fit)
# (Intercept) log(residual_sugar)
# -0.009648201      0.002466801
# log(density) = 0.002466801 * log(residual_sugar) - 0.009648201

```

```

# Correlation test for normalised
cor.test(log(residual_sugar),
         log(density),
         method = "pearson")

# Scatter plot of residual_sugar vs density
plot(y=density, x=residual_sugar, pch='.')
abline(fit<-lm(density ~ residual_sugar, data = white_wine),
       col = "blue",
       lwd=2)

# get the intercept and gradient
coef(fit)
# (Intercept) residual_sugar
# 0.9908653878  0.0004947244
# density = 0.0004947244*residual_sugar + 0.9908653878

# correlation test
cor.test(density,
         residual_sugar,
         method = "pearson")

ggplot(white_wine, aes(x = variable1, y = variable2)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Scatter plot of Variable 1 vs Variable 2",
       x = "Variable 1",
       y = "Variable 2")

# Multiple Linear regression ----
library(car) # use car library

# build model for quality vs alcohol, density, chlorides
model <- lm(quality ~ alcohol + density + chlorides, data = white_wine)
model
summary(model)

# Variance Inflation Factor (VIF): check multicollinearity
vif(model)

# coefficients of variables
coef(model)
# quality = 0.34311688*alcohol + 23.6708667*density -2.3823*chlorides -21.1502

# confidence interval of model at 95% for all predictors

```

```
confint(model)
```

```
# Added-Variable Plots: Plot interaction between variables
```

```
avPlots(model)
```

```
abline(model,  
  col = "blue",  
  lwd=2)
```