Wine Analysis

Elliott Trott

25/10/2022

Introduction

Data taken from Kaggle describes a collection of variables measured for a total of 1359 wines. Performing some basic statistical analysis and visualizing some of the outputs, it is the key aim of this report to highlight methods to reveal relationships between selected variables simply and quickly.

The Data at a Glance

```
summary(WineData)
```

```
##  fixed acidity    volatile acidity  citric acid     residual sugar
##  Min.   : 4.600   Min.   :0.1200   Min.   :0.0000   Min.   : 0.900
##  1st Qu.: 7.100   1st Qu.:0.3900   1st Qu.:0.0900   1st Qu.: 1.900
##  Median : 7.900   Median :0.5200   Median :0.2600   Median : 2.200
##  Mean   : 8.311   Mean   :0.5295   Mean   :0.2723   Mean   : 2.523
##  3rd Qu.: 9.200   3rd Qu.:0.6400   3rd Qu.:0.4300   3rd Qu.: 2.600
##  Max.   :15.900   Max.   :1.5800   Max.   :1.0000   Max.   :15.500
##    chlorides       free sulfur dioxide total sulfur dioxide    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9967
##  Mean   :0.08812   Mean   :15.89       Mean   : 46.83       Mean   :0.9967
##  3rd Qu.:0.09100   3rd Qu.:21.00       3rd Qu.: 63.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.74   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.21   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.31   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.31   Mean   :0.6587   Mean   :10.43   Mean   :5.623
##  3rd Qu.:3.40   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.01   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

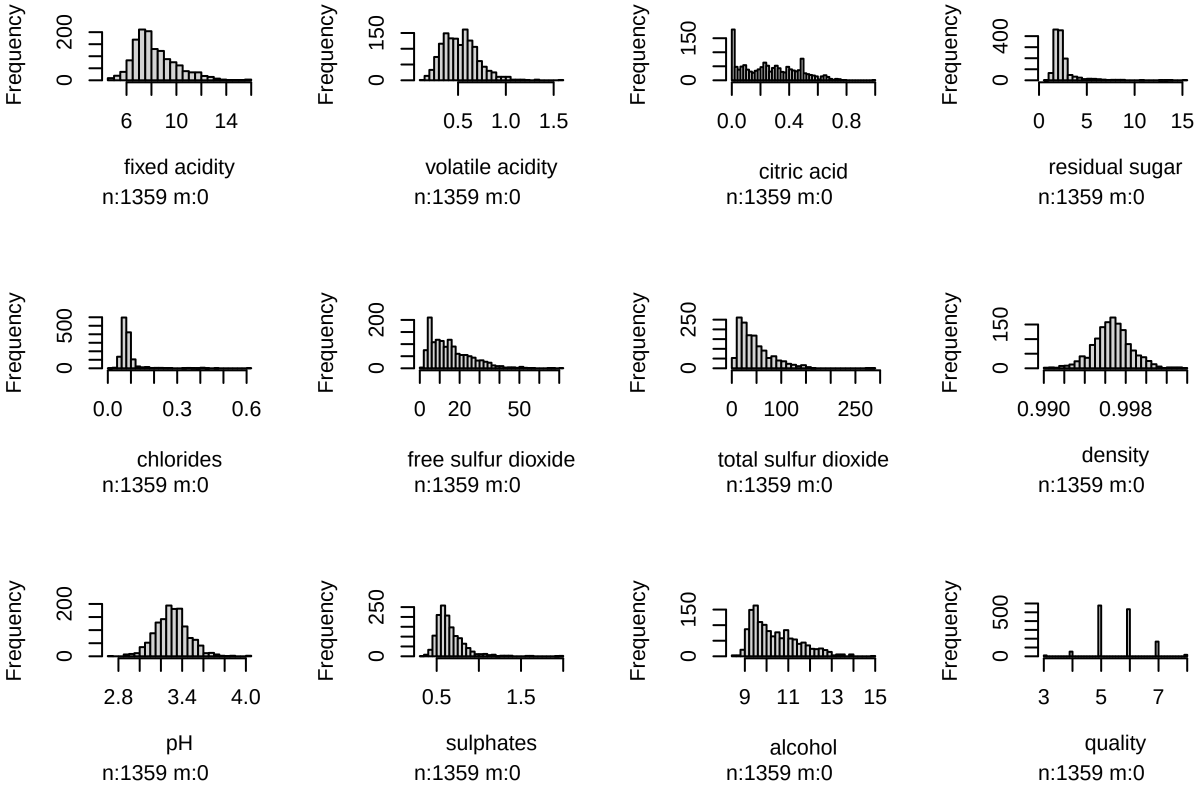
The list featured above provides a brief statistical summary of the employed data set, displaying minimum and maximum values alongside interquartile ranges by column.

```
dim(WineData)
```

```
## [1] 1359   12
```

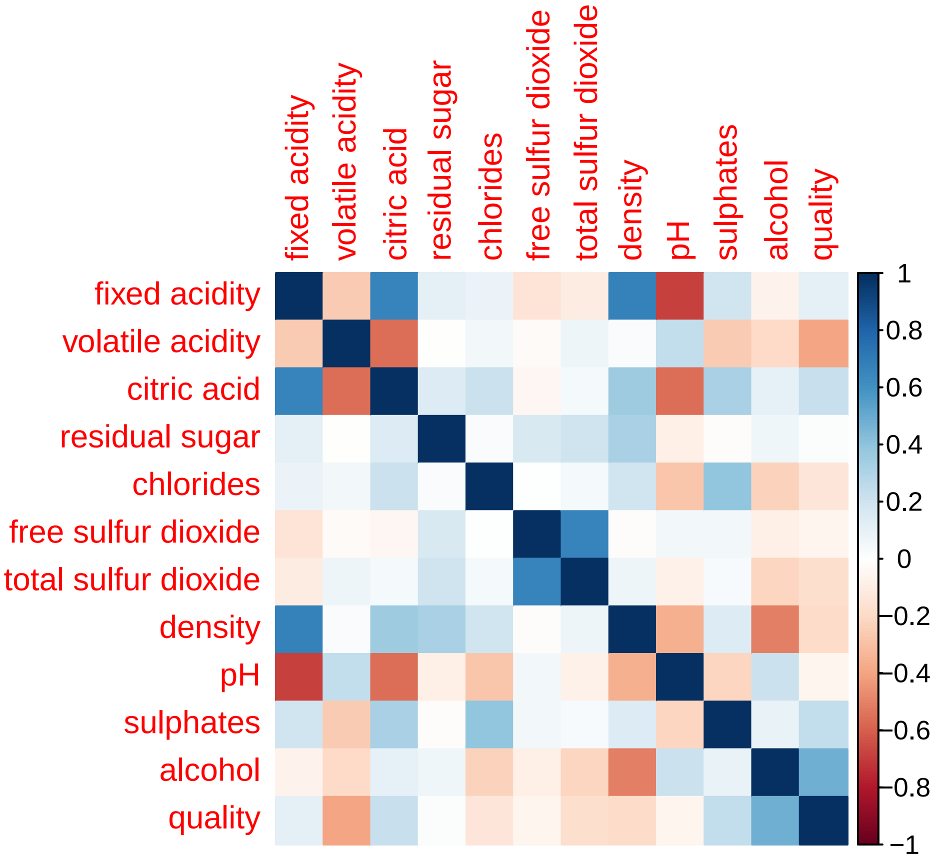
Outputs above describe the dimensions of the data set in consideration. Removal of duplicates using package dplyr and function distinct led to dimensions being shrank from 1599x12 to 1359x12. Therefore, it can be said, there are a total of 1359 distinct wines within this data sample.

```
hist.data.frame(WineData)
```



Histograms above are for each column within the data frame, better illustrating the distribution of wines across variables. It is evident that the variable quality is qualitative, other variables in the sample show distributions which appear to be quantitative. Distribution for 'quality' appears to be fairly normal, the majority of data points residing within the central 5-7 range. The variance of this particular column 0.678. Alcohol content is a variable that shall be considered and studied throughout this report, the distribution shows a negative skew, as percentage increases, frequency decreases, with most wines in the sample residing within 9-10% alcohol concentration.

```
CorWineData <- cor(WineData)
corrplot(CorWineData, method = 'color')
```



Correlation matrix for each of the variables. The figure featured above denotes, briefly, the correlation between assessed variables. The column ‘quality’ will be used as the ‘y’ (or dependent variable), relationships between independent variables and this one will be the main focus of this report.

From the above correlation matrix, it is evident that volatile acidity may have the strongest negative correlation with our y variable, with alcoholic contents (as a percentage) appearing to have the strongest positive correlation. Variables ‘alcohol’ and ‘quality’ are the deepest blue, approaching 1. Further statistical tests are required to prove this, and return more information about other relationships within the data set.

```
table(WineData$quality)
```

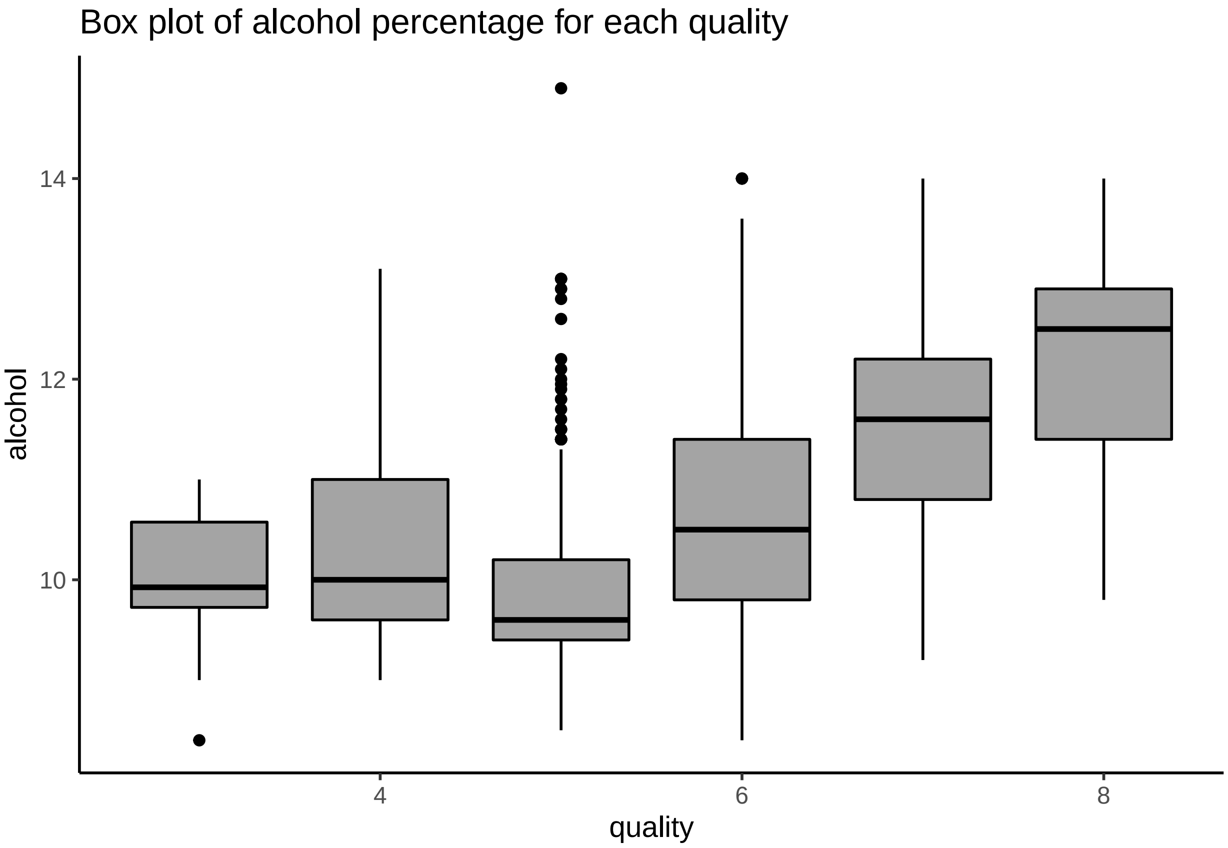
```
##
##   3   4   5   6   7   8
##  10  53 577 535 167  17
```

This table shows the distribution of wines within the column ‘quality’. Quality ‘5’ is the most frequent within the data set, 42.5% of wines across the study demonstrated this value. To determine what traits make a wine ‘high quality’, the study will concentrate primarily on the qualities displayed by wines within the quality = 8 column.

Alcohol Contents

To describe the relationship between alcohol contents and quality, a series of plots was afforded. The first being a box plot, denoting alcohol concentration (y) against quality (x).

```
PWineAlcoholBox + ggtitle("Box plot of alcohol percentage for each quality")
```



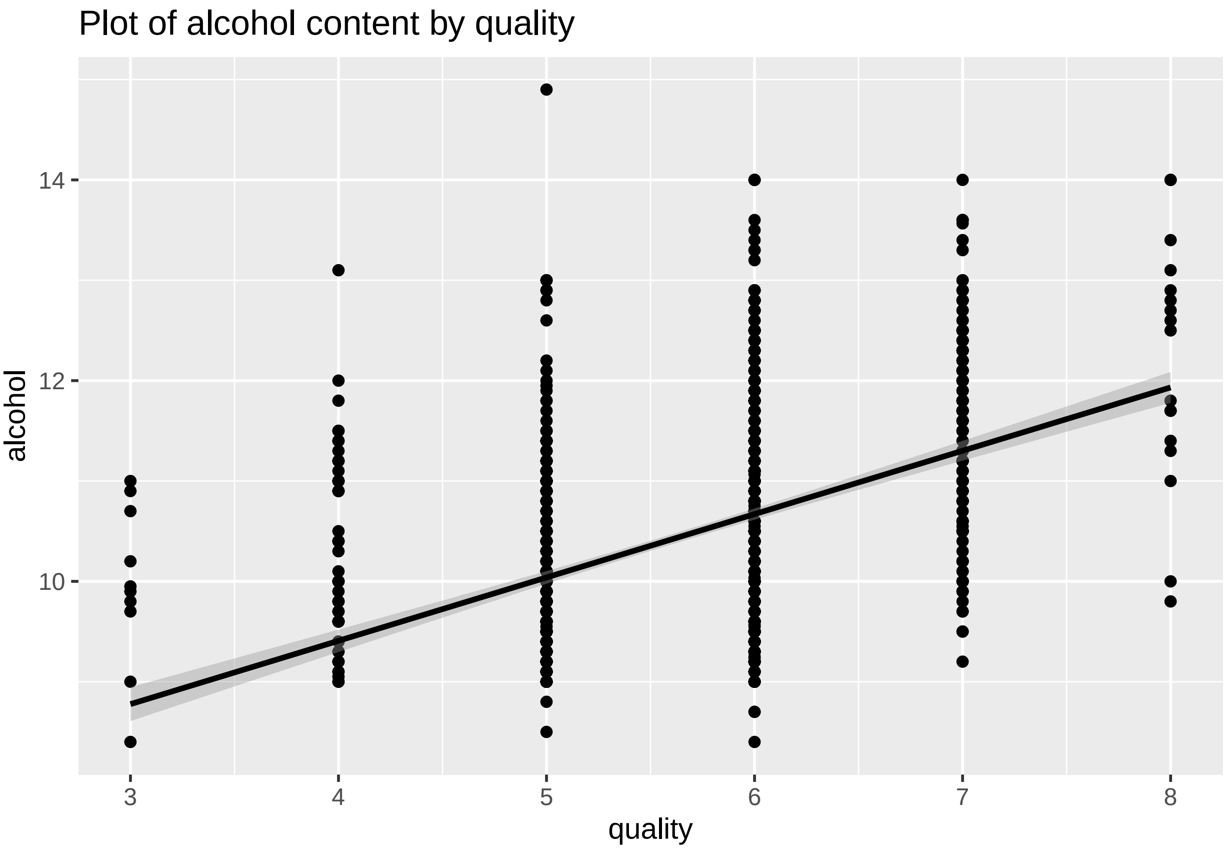
Plot above demonstrates the distribution of alcohol content for each wine by quality. The plot indicates the perceived quality of wine increases as the alcohol content rises. Although there is a trough as the mean average alcohol content of wines in category 5 is lower than those in the previous category, positive correlation is observed from this point.

Standard deviations are overlapping for wines of higher quality, meaning the difference at each stage could be labelled as insignificant. Further statistical tests are needed to clarify the relationship between wine quality and alcohol contents in this study.

The original scatter plot returned no fresh information that could otherwise be considered useful. Further investigation led to the inclusion of linear regression in tandem with the plot. This figure is provided below.

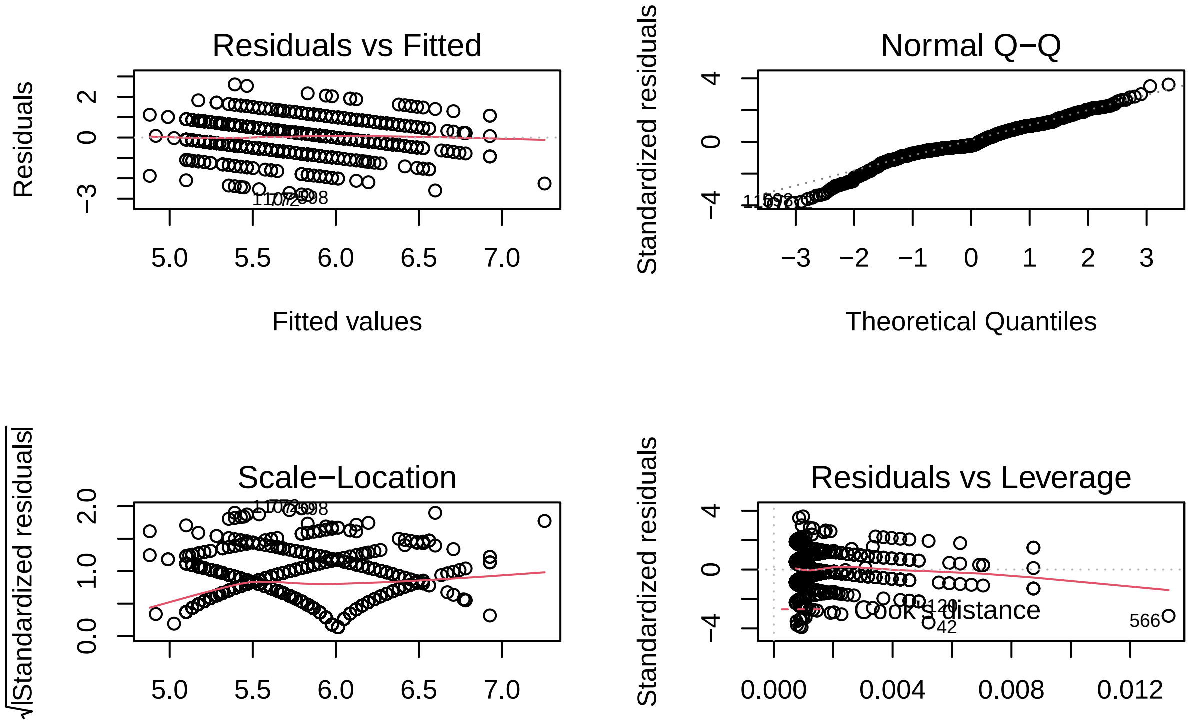
```
ggplot(WineData, aes(x = quality, y = alcohol)) + geom_point() +
  geom_smooth(method = "lm", col = "black") + ggtitle("Plot of alcohol content by quality")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



The addition of a regression line proves the earlier theory that there is a general trend of quality increasing with an increasing percentage of alcohol contents. To assess the use of regression for alcohol content, more tests need to be completed. Four outputs are provided below, describing the continuity of errors in the sample.

```
par(mfrow=c(2,2))
plot(alcohol.qual.lm)
```



The above figures displays outputs from the regression of alcohol contents against wine quality. The data does not fit the assumption of homoscedasticity perfectly.

Density

Another variable which warranted some assessment is density. The density of water is 997 kg/m3 (this would show up as 0.997 in out data frame). All wines residing within the sample are around this figure, therefore, some manipulation was required.

```
density.mean.qual.8
```

```
## # A tibble: 1 x 1
##   average.density
##             <dbl>
## 1           0.995
```

```
density.mean.qual.3
```

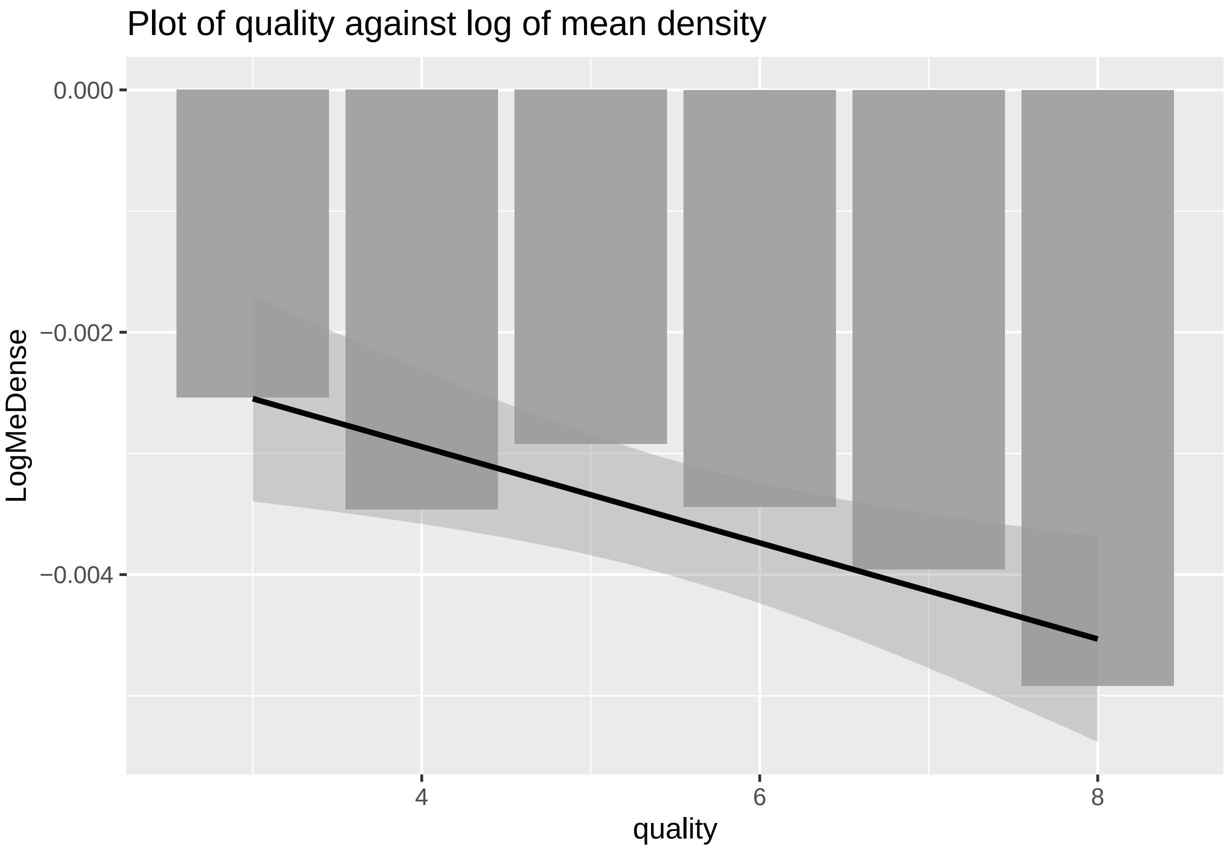
```
## # A tibble: 1 x 1
##   average.density
##             <dbl>
## 1           0.997
```

Above is the statistical output for the mean density by quality 8 and quality 3, namely, the highest and lowest qualities of wine residing within the data set.

A graphical representation of mean densities by quality is provided below.

```
PlotLogDensity + ggtitle("Plot of quality against log of mean density")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



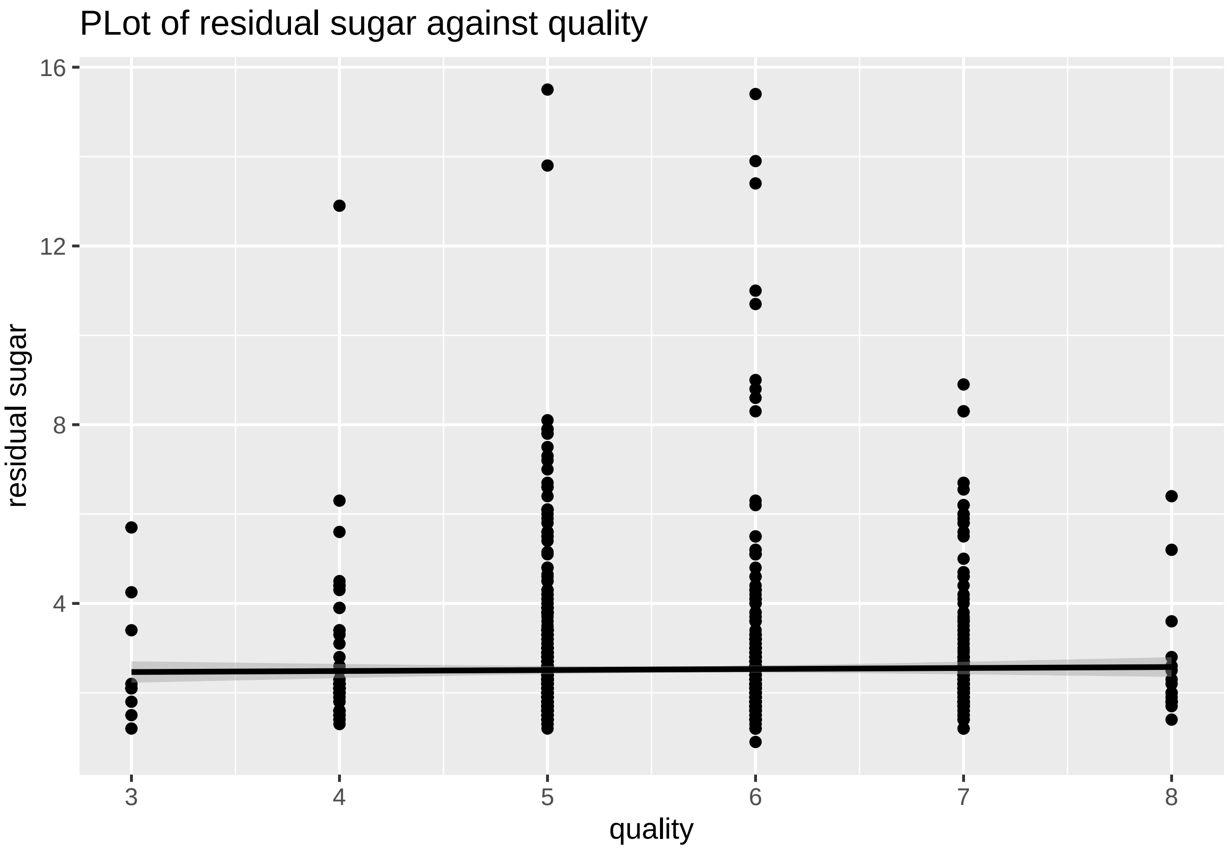
Graphical output of log density against by quality of wine. A notable trend of decreasing log density with increasing quality would suggest less dense wines are perceived as higher quality. More testing needs to be completed to assess the significance of density on quality. Logarithms were used for mean density, without taking a logarithm of each of the y values, the trend was not as evident, all values were clustered together around 0.995-0.997. Less dense wines appear to be favoured as better quality.

Residual Sugar

As an example of variables that do not show any relationship with each other, a study was completed over residual sugar. The output is afforded below.

```
ggplot(WineData, aes(x=quality, y=`residual sugar`)) + geom_point() + geom_smooth(method = "lm", col =
  ggtitle("PLot of residual sugar against quality")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



It is evident from this output there is no relationship between residual sugar and quality. The regression line is consistent throughout the plot with no positive or negative skew.

Conclusions

In conclusion, a higher alcohol contents related to a better quality in general. This was also the case with a decreasing density. The density of ethanol is 789 kg/m3 (which would be represented as 0.789 in the data frame), With an increasing concentration of ethanol in the sample, the density of each wine will decrease, it is clear the two variables relate to each other in this way.

Further investigation is possible over this data set, this report purely highlights a couple of methods that can be adopted for a range of data sets, including visualization and some basic analysis.