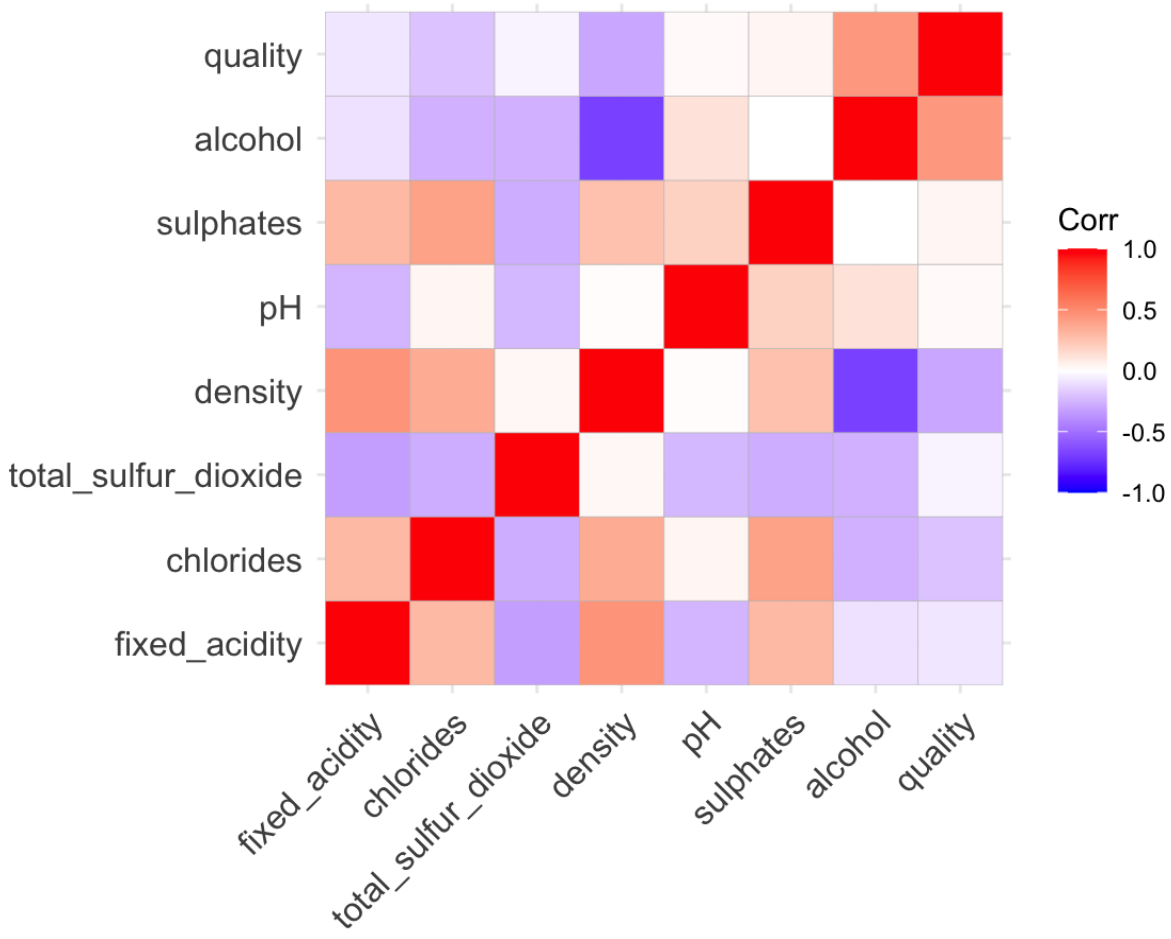
WINE QUALITY DATA ANALYSIS

About Dataset:
https://www.kaggle.com/datasets/rajyellow46/wine-quality
The reference [Cortez et al., 2009] provides information about two datasets pertaining to the red and white variations of Portuguese "Vinho Verde" wine.
Attribute Information:

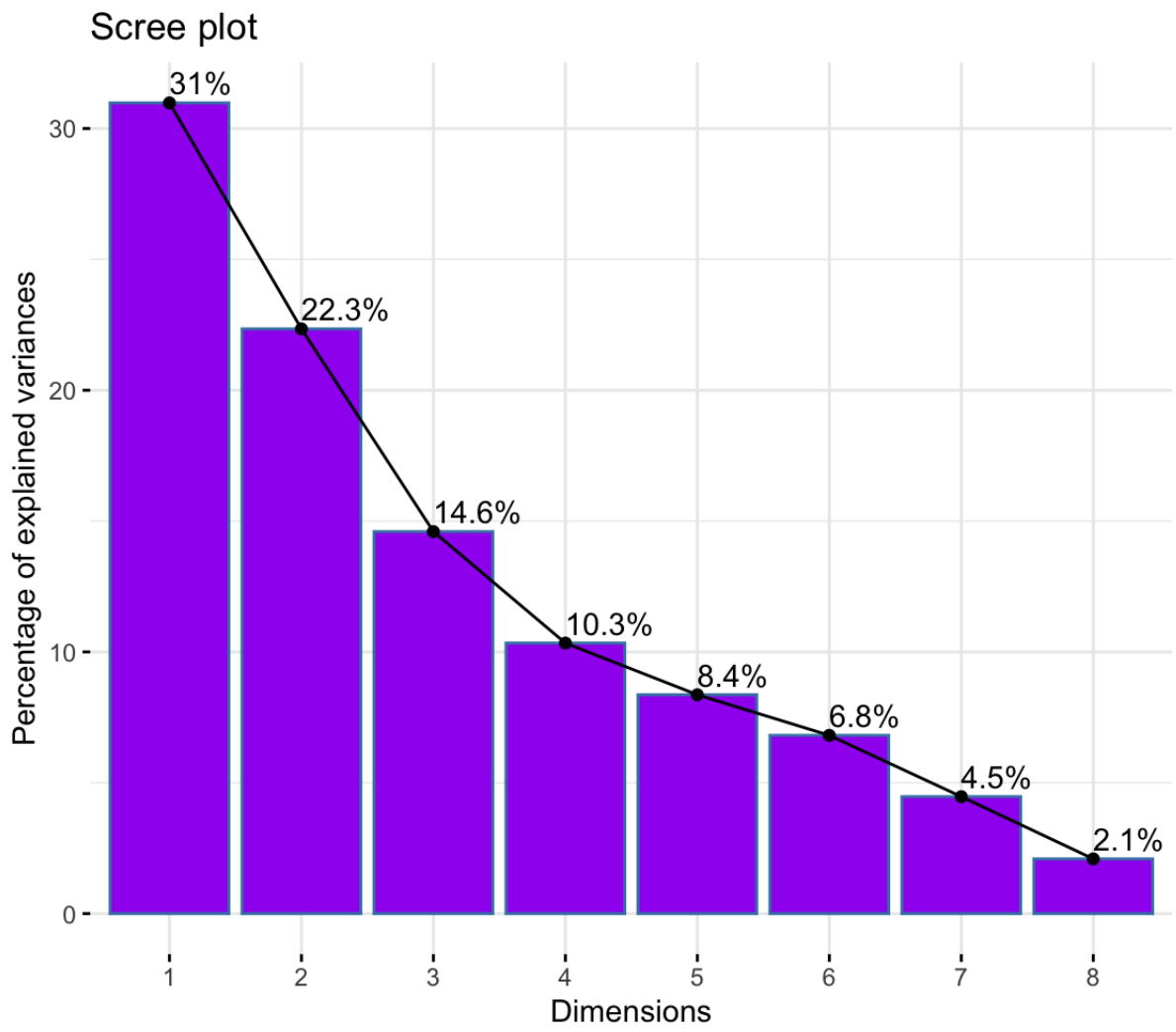
Variables in the dataset:

1 - Type
2 - fixed acidity
3 - chlorides
4 - total sulfur dioxide
5 - density
6 - pH
7 - sulphates
8 - alcohol
9 - quality



There is a high negative correlation between alcohol and the density. The more the density the less is the alcohol and vice versa.
There is no correlation between the pH value and density. No matter how high the density is, the pH remains unchanged.
The quality of the alcohol is directly proportional to the alcohol content. The better the quality, more is the alcohol.

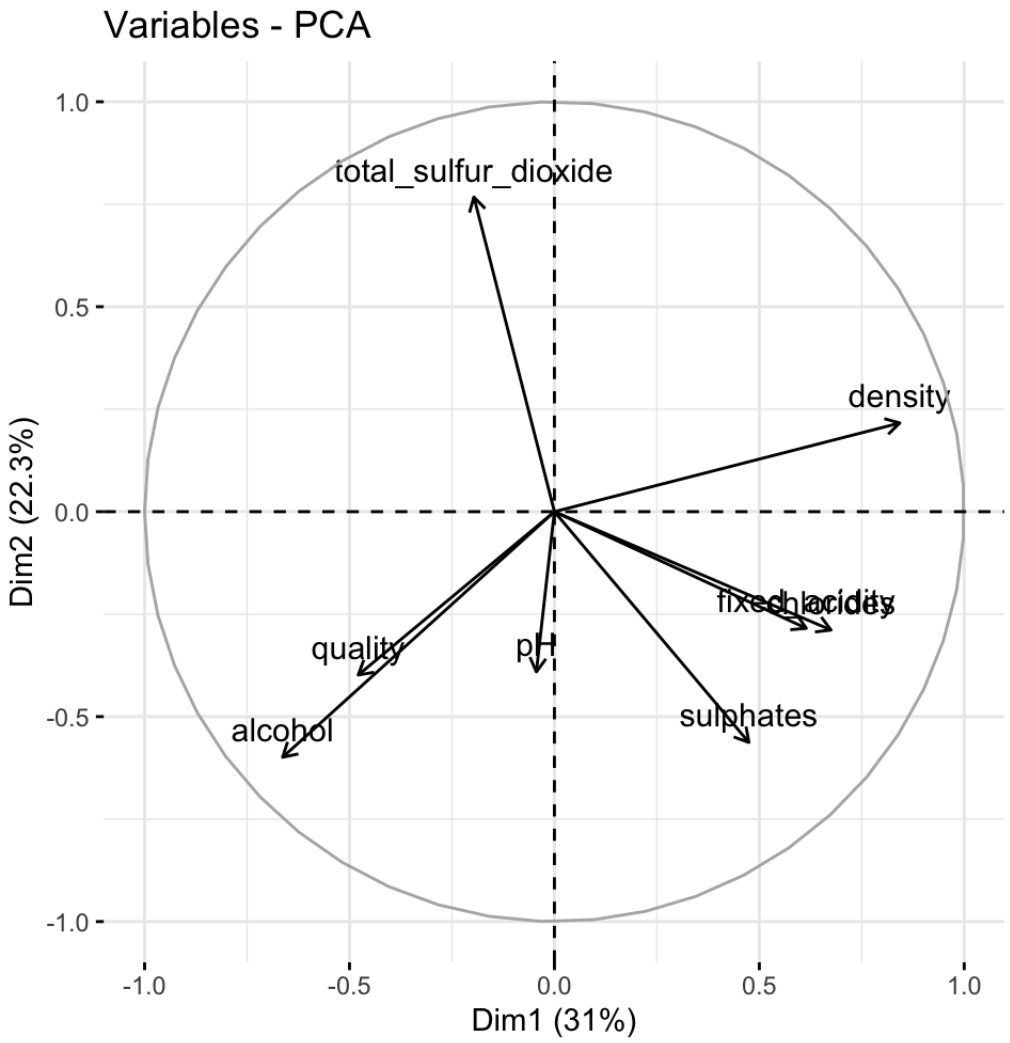
Principal Component Analysis:



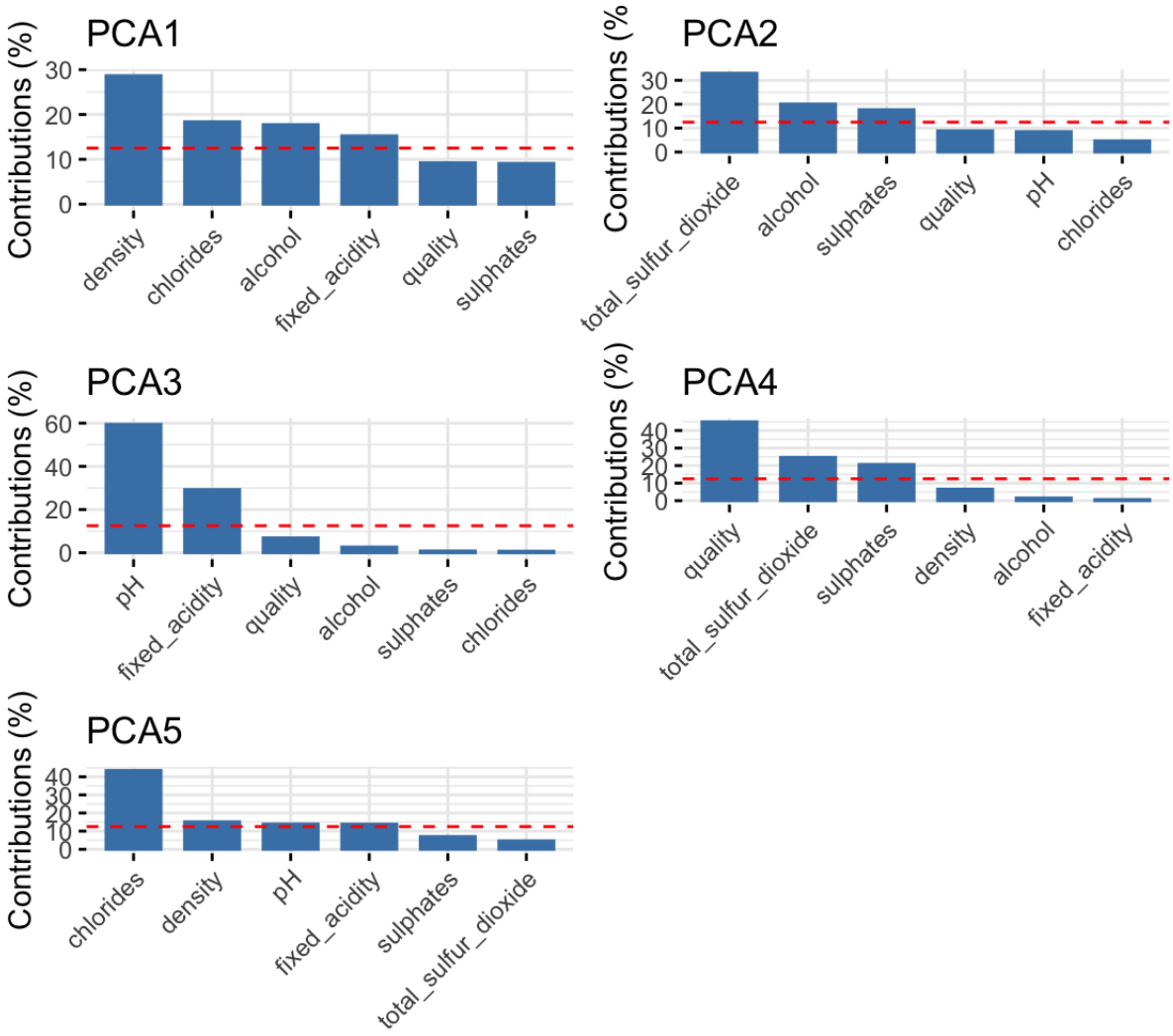
The first 5 PCA's approximately more than 85% of the total variation of the data. Hence, it is observed that the first 5 PCA's capture the crux of the dataset successfully.

```
> eigenvalues
    PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8
0.30980 0.22344 0.14598 0.10338 0.08364 0.06812 0.04471 0.02093
>
```

The results of the loading matrix show that the majority of the variation in the dataset is captured by the first principal component (PC1), which is primarily driven by the levels of fixed acidity, chlorides and density. The second principal component (PC2) explains a comparatively smaller proportion of the variance, and is primarily driven by the levels of total sulphur dioxide. The third principal component (PC3) explains another small proportion of the variance, and is primarily driven by the fixed acidity. The fourth principal component (PC4) explains yet another small proportion of the variance, and is primarily driven by total sulphur dioxide and sulphates.

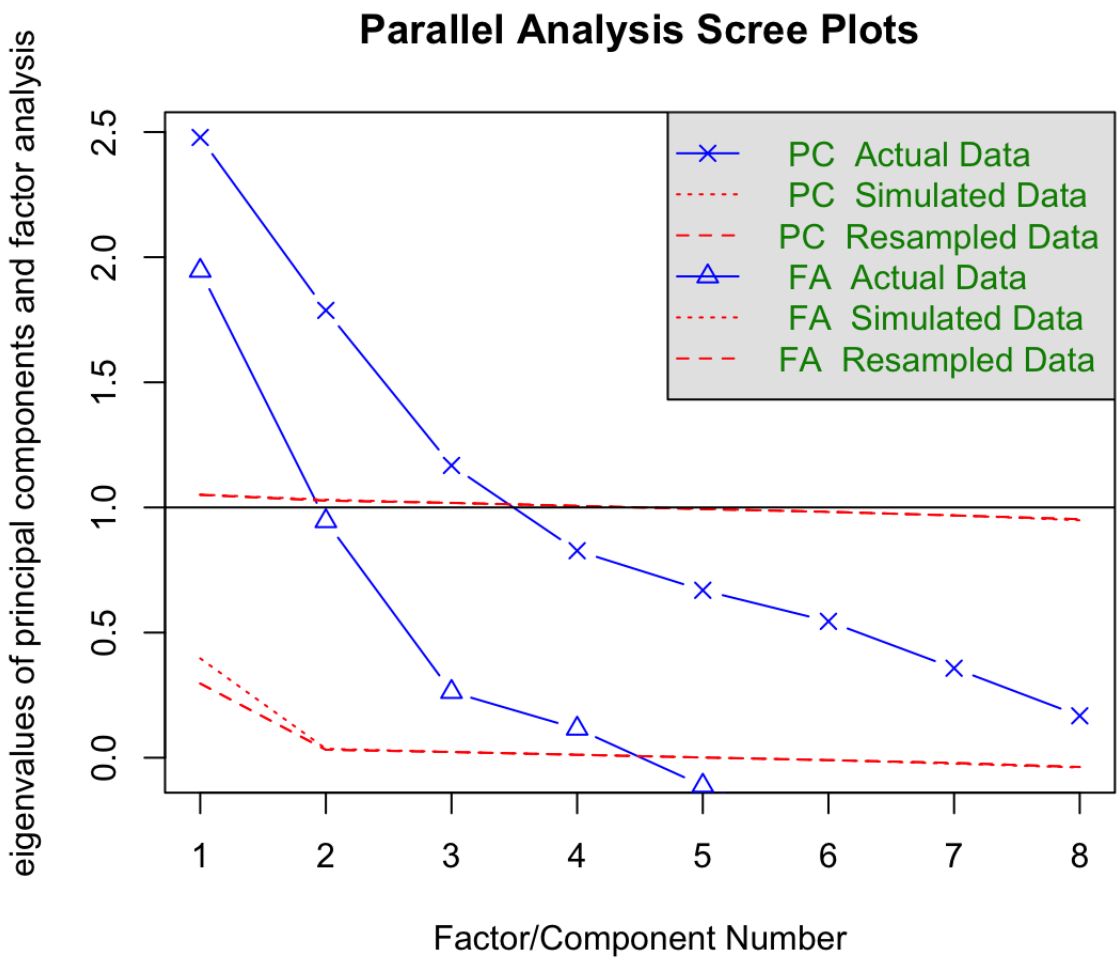


From this image, we can observe that fixed acidity and chlorides are positively correlated to one another. Alcohol and quality are positively correlated as well. Total sulphur dioxide is negatively correlated to all the components present in the image.

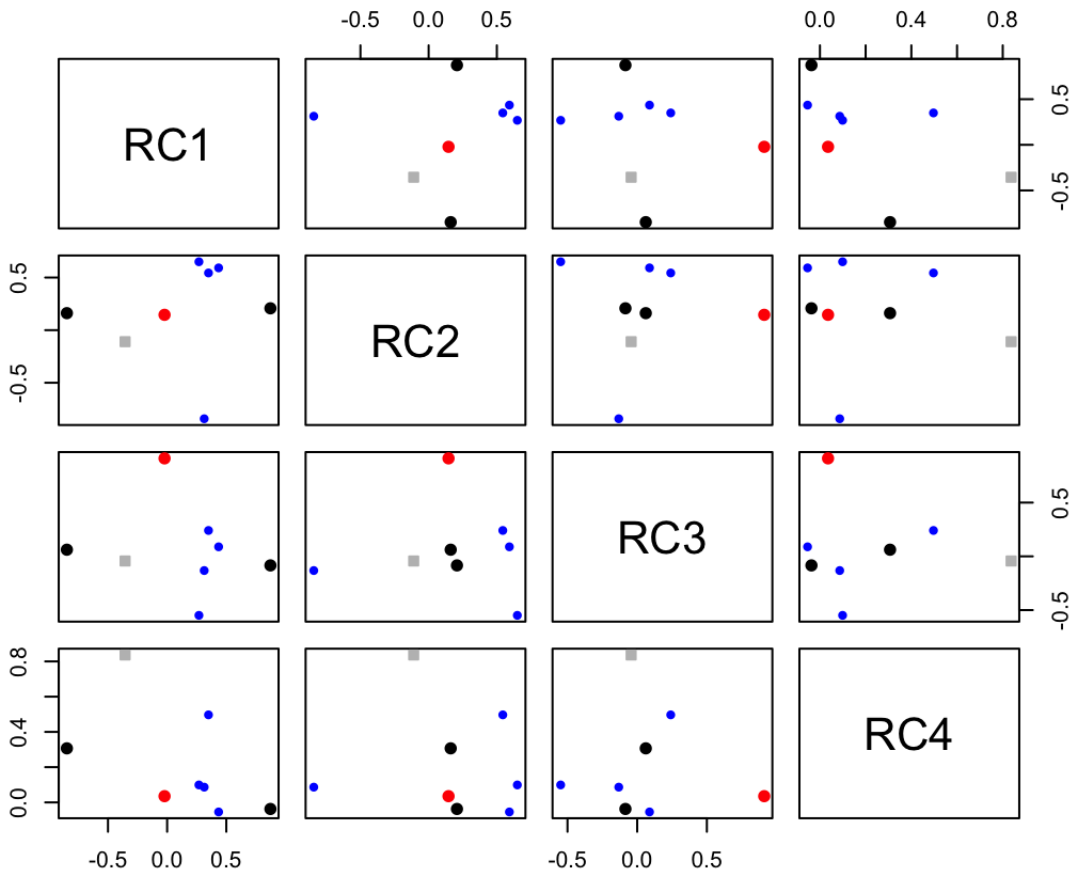


PCA1 is driven by density, chlorides and alcohol. The density measures the amount of sugar and other extras in the liquid (Less sugar = more alcohol = lighter weight).
PCA2 is driven by total sulphur dioxide. In wine making, the use of sulphur dioxide is essential. The Free SO2 and the pH of your wine determine how much SO2 is available in the active, molecular form to help protect the wine from oxidation and spoilage.
PCA3 is driven by pH.

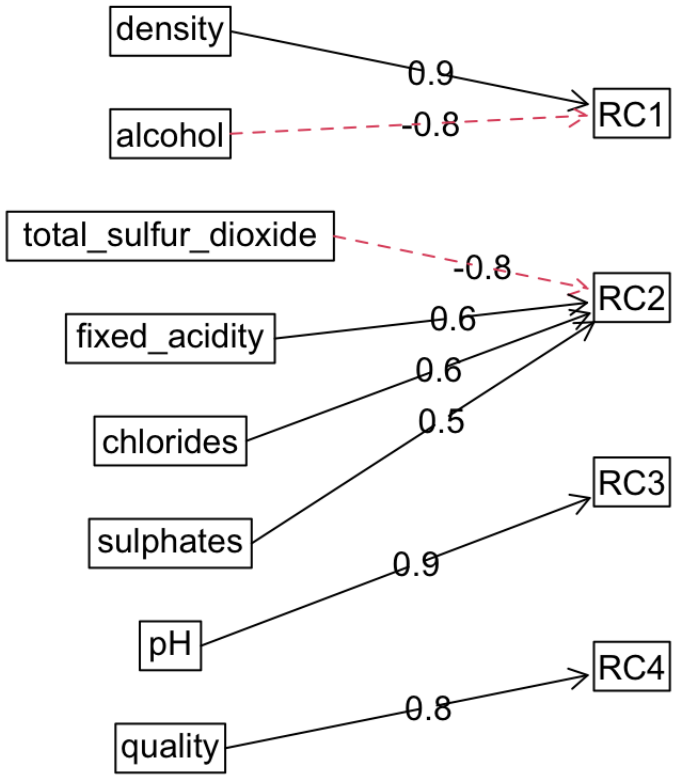
Exploratory Factor Analysis:



The number of components suggested are 4 by studying the graph. The point where the red line cuts the blue line is referred to as the number of factor.


Principal Component Analysis
RC1
RC2
RC3
RC4
-0.5
0.0
0.5
0.0
0.4
0.8
0.5
-0.5
0.5
-0.5
0.5
-0.5
0.8
0.4
0.0
-0.5
0.0
0.5
-0.5
0.0
0.5

Components Analysis



In RC1, the density is impacting it in a positive way whereas the alcohol is impacting in a negative way (hence the red line).
In RC2, the fixed acidity, chlorides and sulphates are impacting positively whereas the total sulphur dioxide impacts negatively.
In RC3, only pH is impacting in a positive manner.
In RC4, only quality is impacting in a positive manner.

Logistic Regression:

```
Call:
glm(formula = Type ~ ., family = "binomial", data = winequality)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7207   0.0013   0.0261   0.0832   6.7603

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           5.963e+02  5.557e+01  10.732  < 2e-16 ***
fixed_acidity        -1.035e+00  1.152e-01  -8.981  < 2e-16 ***
chlorides            -4.329e+01  3.014e+00 -14.362  < 2e-16 ***
total_sulfur_dioxide  6.204e-02  2.904e-03  21.364  < 2e-16 ***
density              -5.589e+02  5.553e+01 -10.065  < 2e-16 ***
pH                   -8.573e+00  7.045e-01 -12.169  < 2e-16 ***
sulphates            -7.089e+00  7.643e-01  -9.274  < 2e-16 ***
alcohol              -4.895e-01  1.300e-01  -3.765 0.000167 ***
quality               4.252e-01  1.223e-01   3.478 0.000506 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7218.33  on 6462  degrees of freedom
Residual deviance:  848.63  on 6454  degrees of freedom
AIC: 866.63

Number of Fisher Scoring iterations: 8
```

All variables are significant as the P values is less than 0.05.
In the estimate column, we can observe that if 1 unit is increased in X (coefficient) then expected change in Y (type of wine), in this case, one unit change in quality, then 4.252e-01 change in type of wine.
We can see the model gives us the AIC of 866.63.

Conclusion:
According to the data analysis, there is no association between pH and density, but there is a negative correlation between alcohol and density. Alcohol quality and alcohol content are directly inversely related. More than 85% of the variety in the data is captured by the first five PCA main components. Fixed acidity, chlorides, and density are the main factors affecting the first principal component, while total sulphur dioxide affects the second. The fourth primary component is driven by total sulphur dioxide and sulphates, while the third is driven by fixed acidity. The PCA offers four suggestions for components. Alcohol has a negative effect in RC1 whereas density has a positive effect in the correlation analysis.
In RC2, total sulphur dioxide has a negative effect, while fixed acidity, chlorides, and sulphates have beneficial effects. In RC3, only pH has a favourable effect, while in RC4, only quality does.