STAT467 PROJECT

GROUP 6

Mustafa Ugur Yalcin - 2502458
Melih Can Kanmaz - 2502169
Tevfik Oguz - 2442051

CONTENT

- About the dataset
- Inference of Mean Vector
- Comparisons of Several Multivariate Means
- Principal Component Analysis and PCR
- Factor Analysis and Factor Rotation
- Discrimination and Classification
- Clustering
- Canonical Correlation Analysis

PROJECT AIM

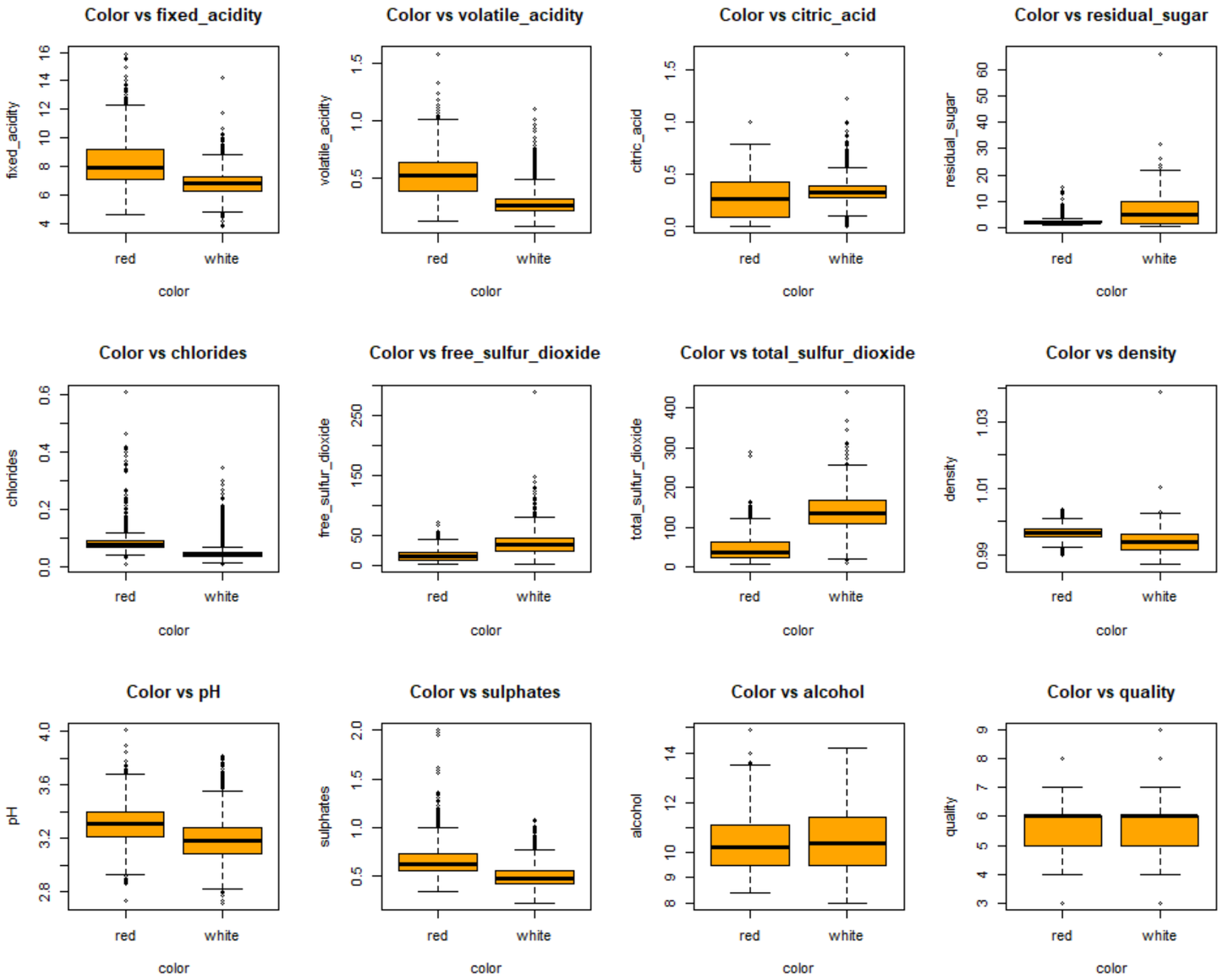
- Does the quality differ across wine colours?
- What factors influence the quality of wine?
- What factors affect the alcohol content?

ABOUT THE DATA

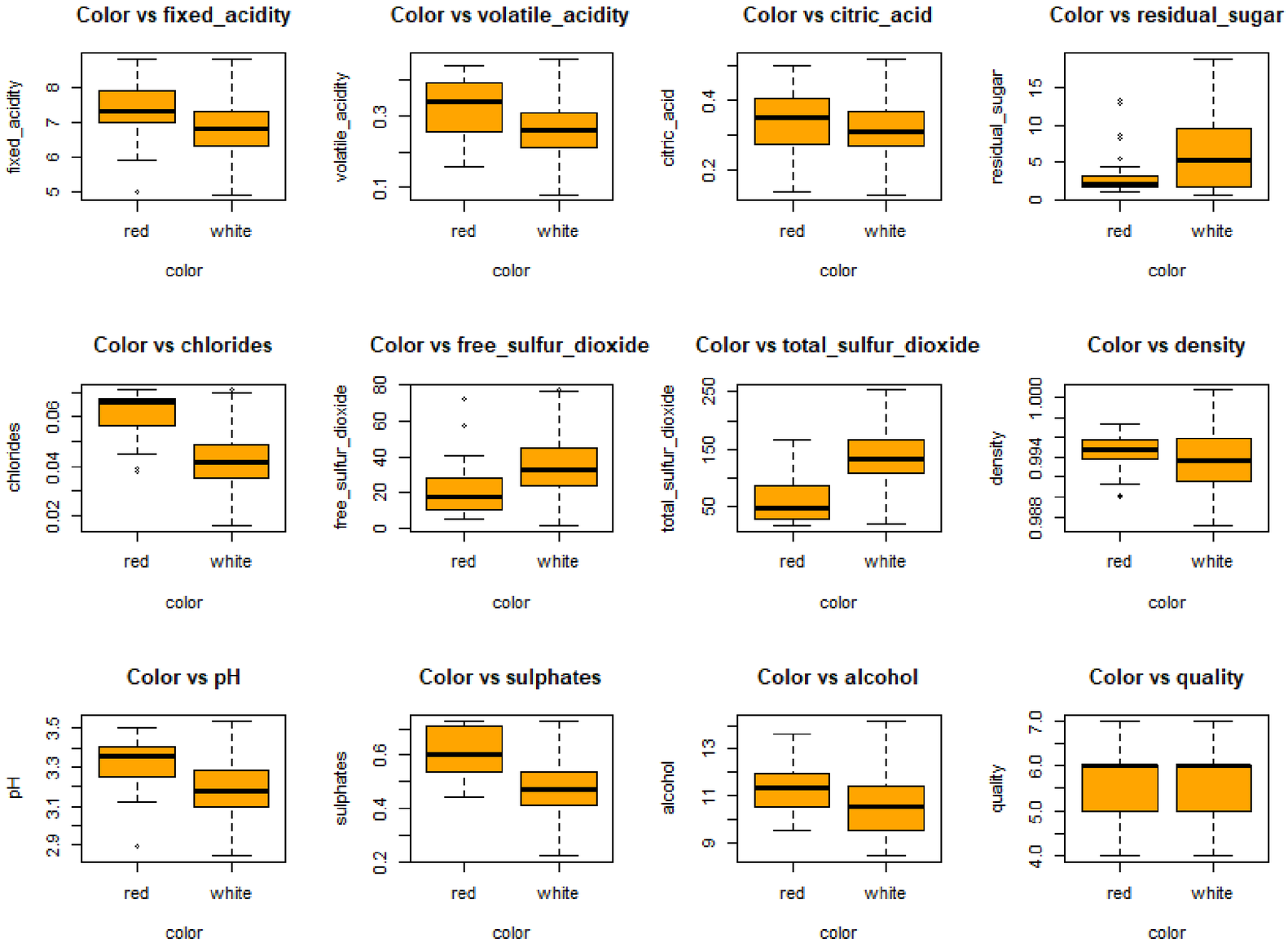
SUMMARY OF THE DATA

```
> summary(data)
 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.    :  1.00
 1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800   1st Qu.:0.03800   1st Qu.: 17.00
 Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000   Median :0.04700   Median : 29.00
 Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443   Mean   :0.05603   Mean   : 30.53
 3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100   3rd Qu.:0.06500   3rd Qu.: 41.00
 Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800   Max.   :0.61100   Max.   :289.00
 total_sulfur_dioxide    density             pH          sulphates         alcohol          quality
 Min.    :  6.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
 1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000
 Median :118.0        Median :0.9949   Median :3.210   Median :0.5100   Median :10.30   Median :6.000
 Mean   :115.7        Mean   :0.9947   Mean   :3.219   Mean   :0.5313   Mean   :10.49   Mean   :5.818
 3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000
 Max.   :440.0        Max.   :1.0390   Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :9.000
    color
 Length:6497
 Class :character
 Mode  :character
```

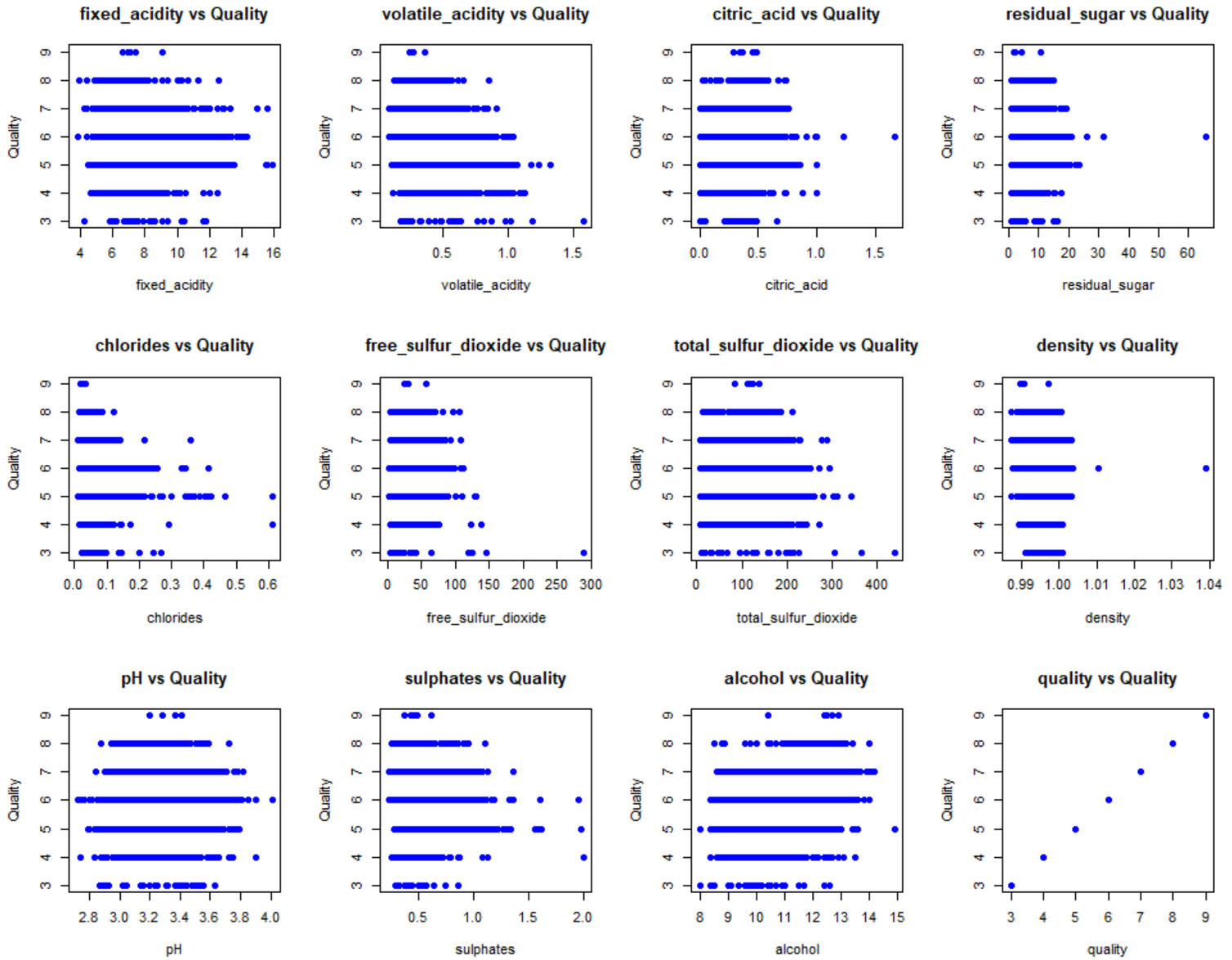
DISTRIBUTIONS OF COLOR



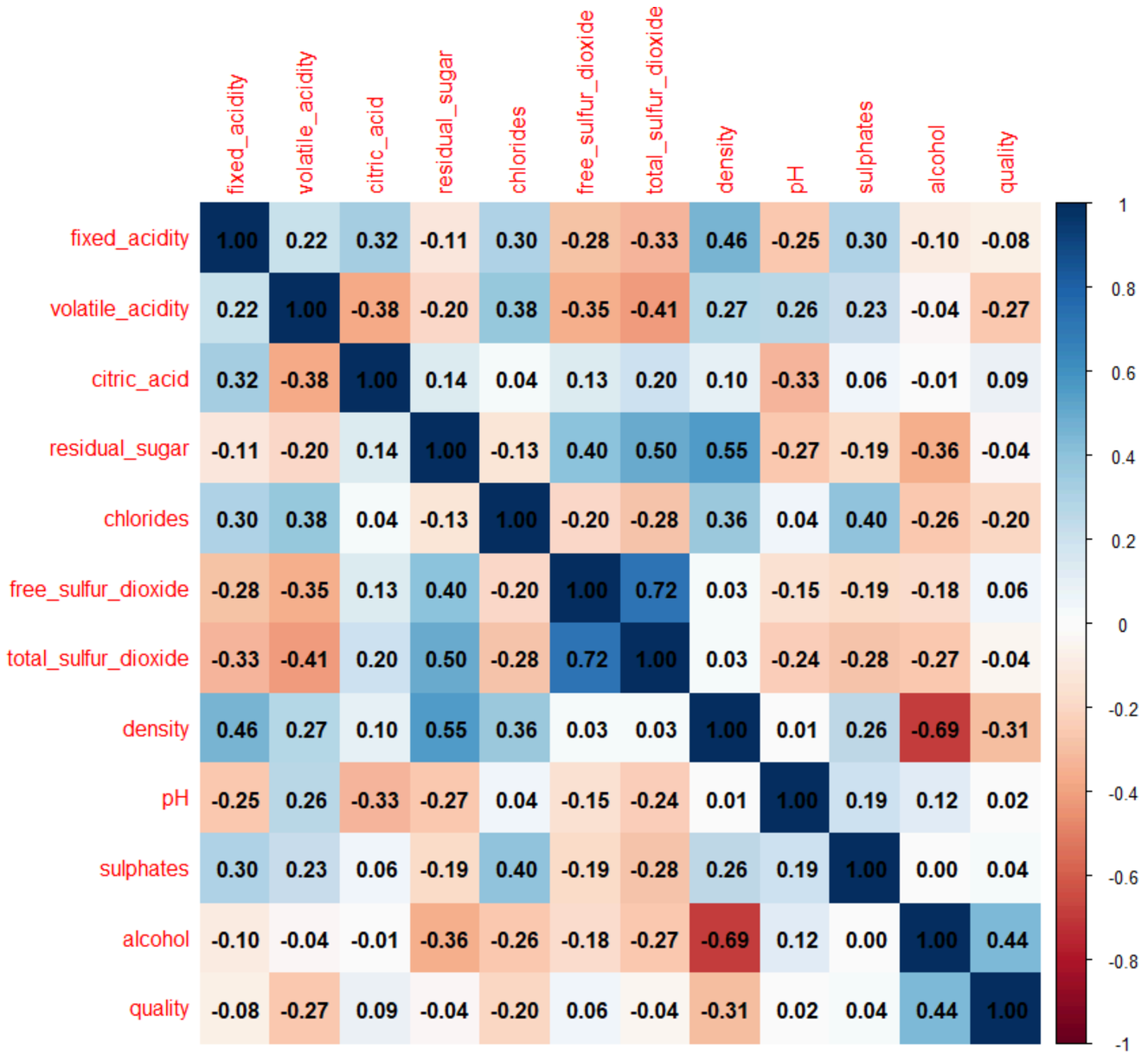
REMOVING OUTLIERS



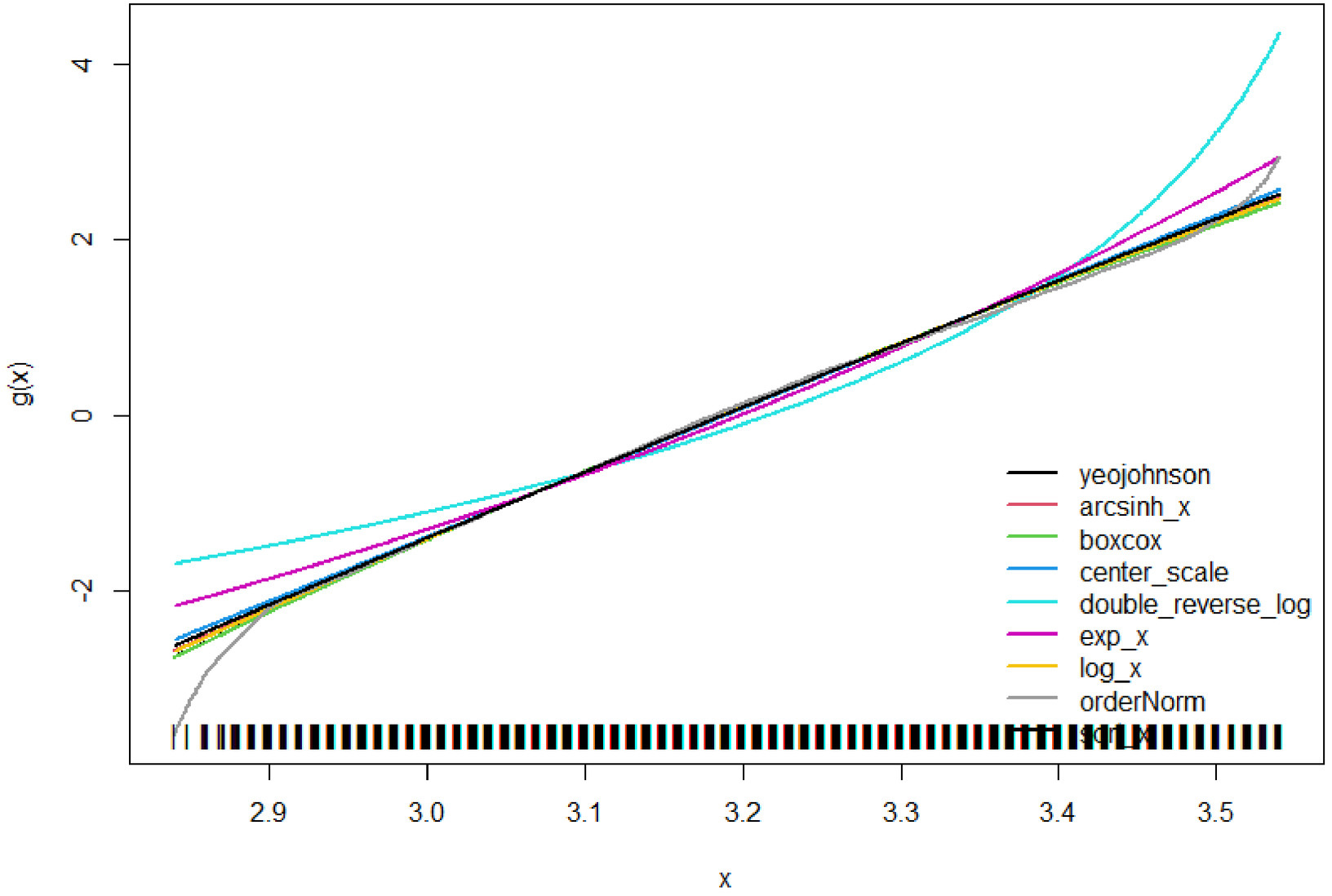
DISTRIBUTIONS OF QUALITY



CORRELATION OF VARIABLES



INFERENCE OF MEAN VECTOR

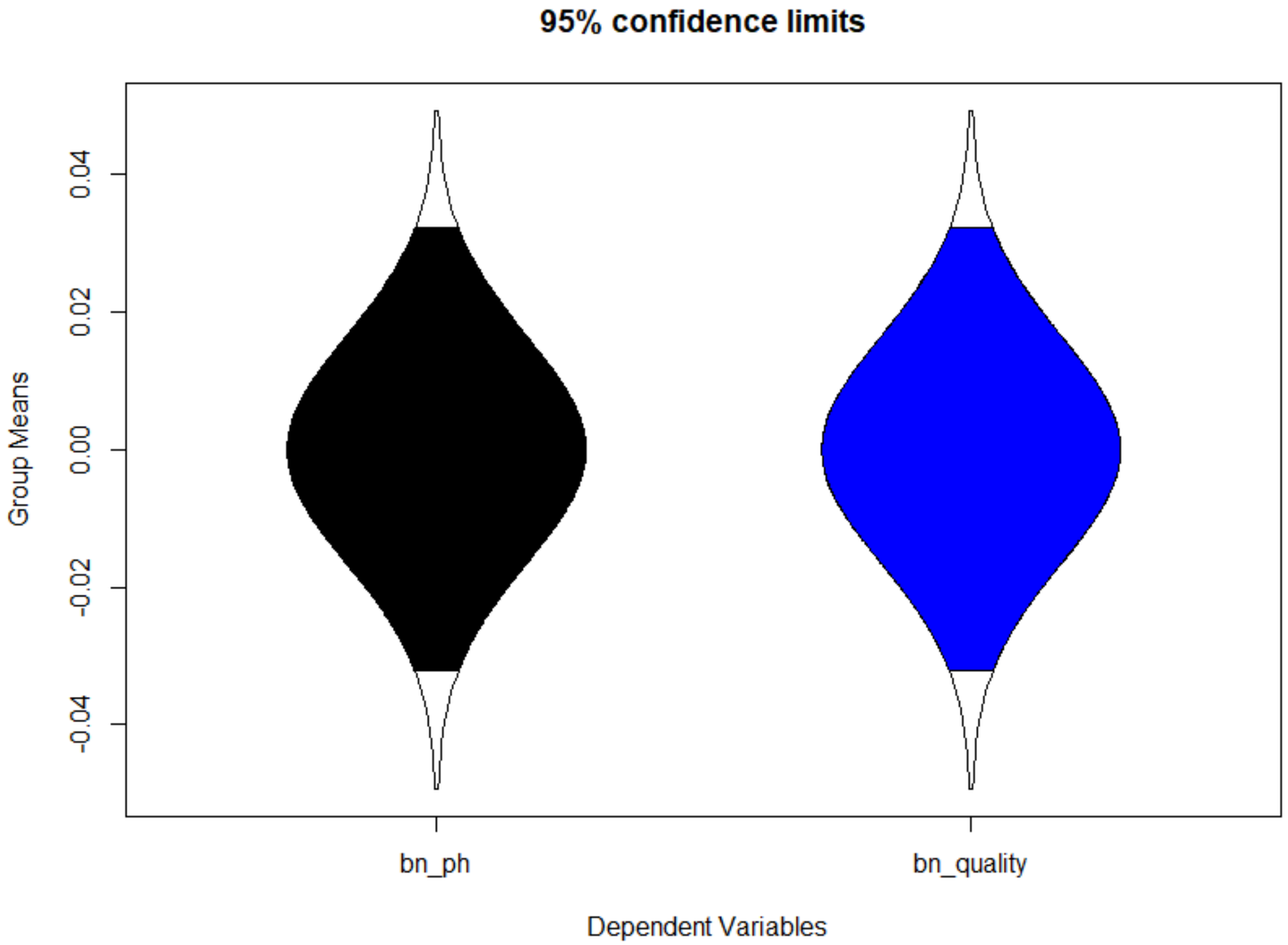


INFERENCE OF MEAN VECTOR

NORMALITY ASSUMPTION

```
> test$multivariateNormality
               Test         Statistic                p value Result
1 Mardia Skewness  138.822630619252 5.04302634059072e-29      NO
2 Mardia Kurtosis -2.91040498967062  0.00360960693973933      NO
3               MVN              <NA>                 <NA>      NO
> test<-mvn(y,mvnTest = "mardia")
> test$univariateNormality
               Test    Variable Statistic   p value Normality
1 Anderson-Darling    bn_ph        2.1183  <0.001       NO
2 Anderson-Darling bn_quality   271.7797  <0.001       NO
>
```

VIOLIN GRAPHS



INFERENCE OF MEAN VECTOR

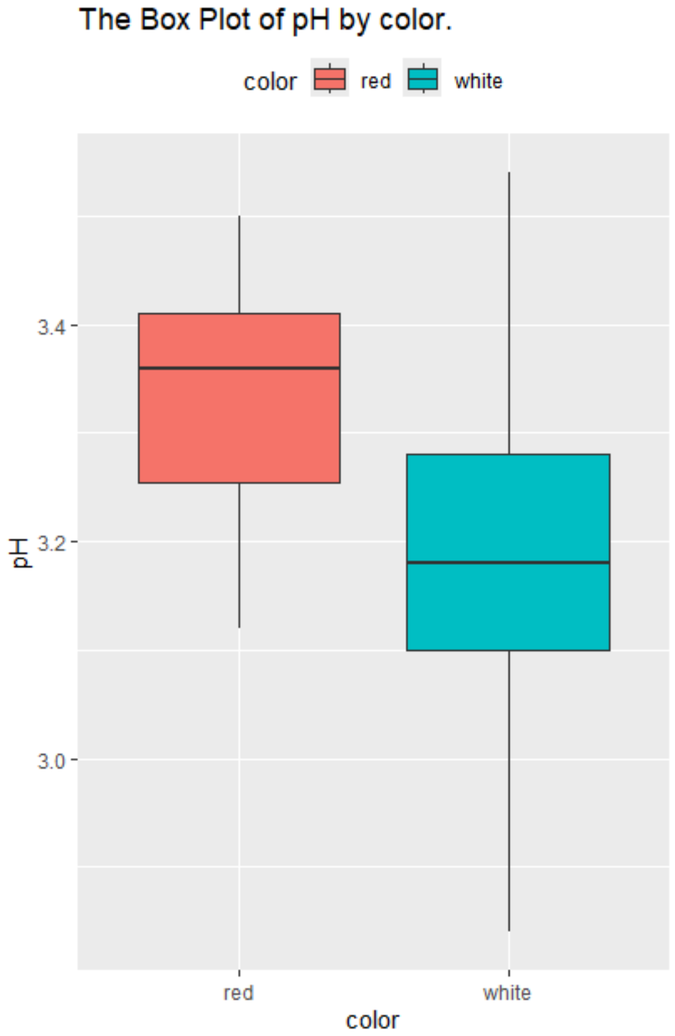
HOTELLING'S T TEST

```
> #Hotelling T test
> HotellingsT2(y,mu=mu0)

        Hotelling's one sample T2-test

data:  y
T.2 = 3532.6, df1 = 2, df2 = 3703,
p-value < 2.2e-16
alternative hypothesis: true location is not equal to c(1,1)
```

COMPARISONS OF SEVERAL MULTIVARIATE MEANS



NORMALITY TESTS

```
> subset_data1 %>% group_by(color) %>%  shapiro_test(log_pH,log_quality)
# A tibble: 4 x 4
  color variable    statistic        p
  <chr> <chr>           <dbl>    <dbl>
1 red   log_pH          0.837 5.83e- 6
2 red   log_quality     0.815 1.61e- 6
3 white log_pH          0.996 1.73e- 8
4 white log_quality     0.835 6.35e-52
> library(heplots)
> boxM(Y = cbind(subset_data1$log_pH,subset_data1$log_quality), group = factor(subset_da
ta1$color))

        Box's M-test for Homogeneity of Covariance
        Matrices

data:  cbind(subset_data1$log_pH, subset_data1$log_quality)
Chi-Sq (approx.) = 7.9875, df = 3, p-value
= 0.04627
> leveneTest(log_pH ~ color, data = subset_data1)
Levene's Test for Homogeneity of Variance (center = median)
         Df F value  Pr(>F)
group     1  3.2002 0.07371 .
       3703
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

MANOVA

```
> m1 <- manova(cbind(log_pH,log_quality) ~ color, data = subset_data1)
> summary(m1)
            Df  Pillai approx F num Df den Df    Pr(>F)
color        1 0.01269   23.791      2   3702 5.412e-11 ***
Residuals 3703
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary.aov(m1)
 Response log_pH :
              Df Sum Sq  Mean Sq F value   Pr(>F)
color          1 0.0837 0.083700  46.401 1.12e-11 ***
Residuals   3703 6.6797 0.001804
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response log_quality :
              Df Sum Sq   Mean Sq F value Pr(>F)
color          1  0.009 0.0094775  0.5515 0.4577
Residuals   3703 63.633 0.0171842
```

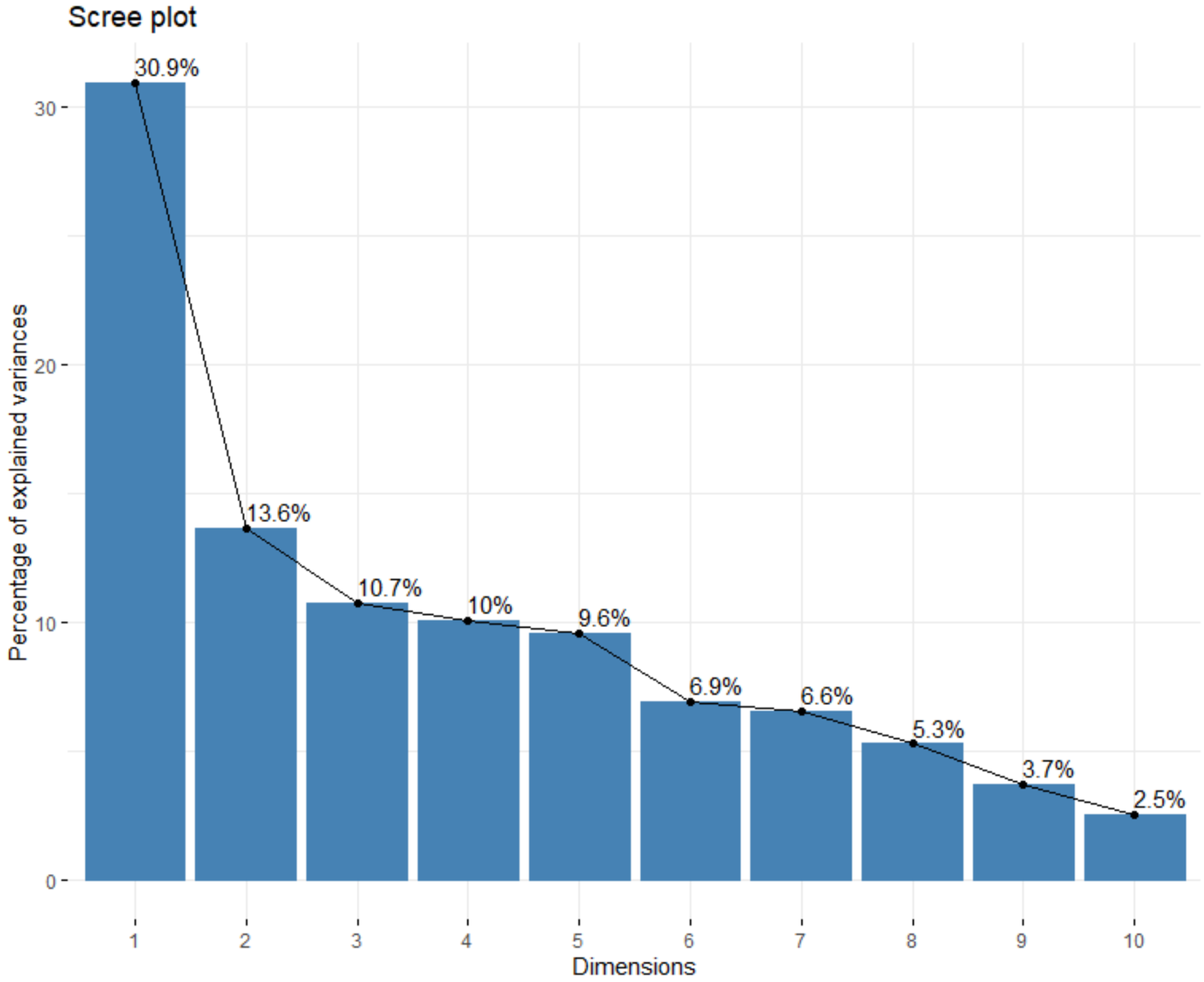
PRINCIPAL COMPONENTS ANALYSIS

```
> cor(scaled_data)
                      fixed_acidity volatile_acidity citric_acid residual_sugar
fixed_acidity          1.000000000     -0.038485191  0.26043580     0.04671636
volatile_acidity      -0.038485191      1.000000000 -0.12265398     0.07314759
citric_acid            0.260435797     -0.122653976  1.00000000    -0.01897515
residual_sugar         0.046716357      0.073147589 -0.01897515     1.00000000
chlorides              0.106754560      0.012152361  0.02005761     0.25504336
free_sulfur_dioxide   -0.042415793     -0.039255844  0.05114958     0.35113924
total_sulfur_dioxide   0.050937391      0.120747029  0.05999404     0.41218759
density                0.219015982      0.008326949  0.03351907     0.83427178
pH                    -0.360461196     -0.013102201 -0.09090630    -0.18255852
sulphates              0.002514512      0.035174451  0.06378779    -0.04183721
alcohol               -0.069948466      0.107407499  0.01991982    -0.49110790
quality               -0.053708662     -0.123667412  0.03199637    -0.11020532
> cov(scaled_data)
                      fixed_acidity volatile_acidity citric_acid residual_sugar
fixed_acidity          1.000000000     -0.038485191  0.26043580     0.04671636
volatile_acidity      -0.038485191      1.000000000 -0.12265398     0.07314759
citric_acid            0.260435797     -0.122653976  1.00000000    -0.01897515
residual_sugar         0.046716357      0.073147589 -0.01897515     1.00000000
chlorides              0.106754560      0.012152361  0.02005761     0.25504336
free_sulfur_dioxide   -0.042415793     -0.039255844  0.05114958     0.35113924
total_sulfur_dioxide   0.050937391      0.120747029  0.05999404     0.41218759
density                0.219015982      0.008326949  0.03351907     0.83427178
pH                    -0.360461196     -0.013102201 -0.09090630    -0.18255852
sulphates              0.002514512      0.035174451  0.06378779    -0.04183721
alcohol               -0.069948466      0.107407499  0.01991982    -0.49110790
quality               -0.053708662     -0.123667412  0.03199637    -0.11020532
```

PRINCIPAL COMPONENTS ANALYSIS

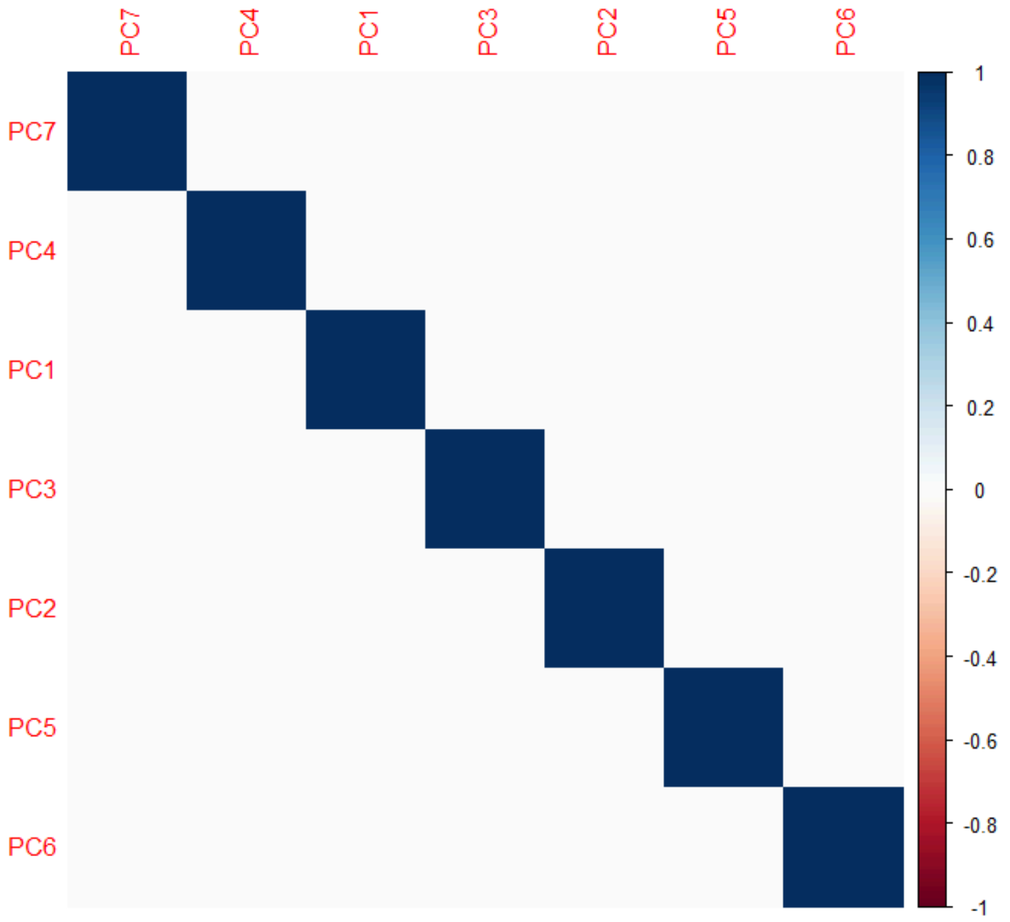
```
> summary(pca1)
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8
Standard deviation     1.8434 1.2240 1.0865 1.0514 1.02498 0.87300 0.84942 0.76205
Proportion of Variance 0.3089 0.1362 0.1073 0.1005 0.09551 0.06928 0.06559 0.05279
Cumulative Proportion  0.3089 0.4451 0.5524 0.6529 0.74842 0.81770 0.88329 0.93608
                           PC9    PC10    PC11
Standard deviation     0.63802 0.52881 0.12790
Proportion of Variance 0.03701 0.02542 0.00149
Cumulative Proportion  0.97309 0.99851 1.00000
```

PRINCIPAL COMPONENTS ANALYSIS

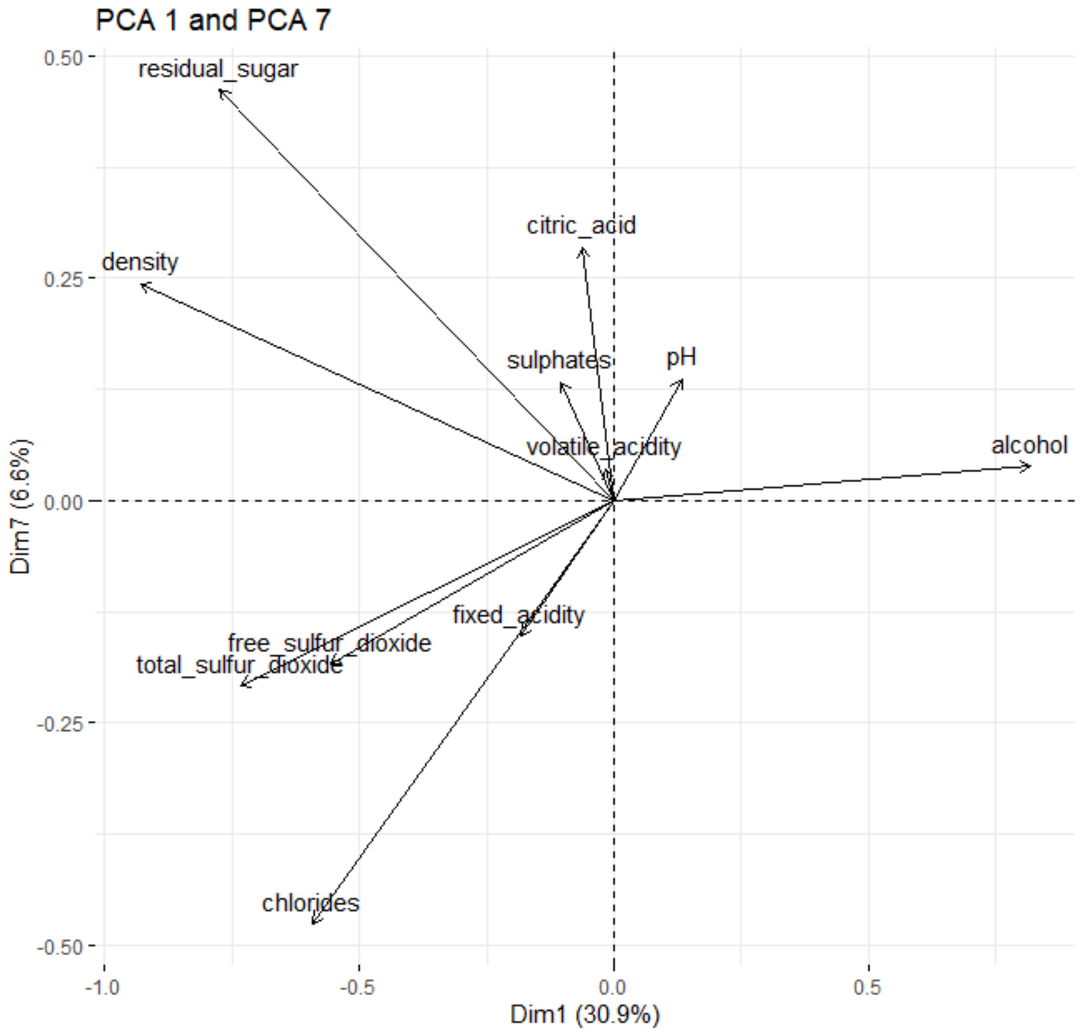


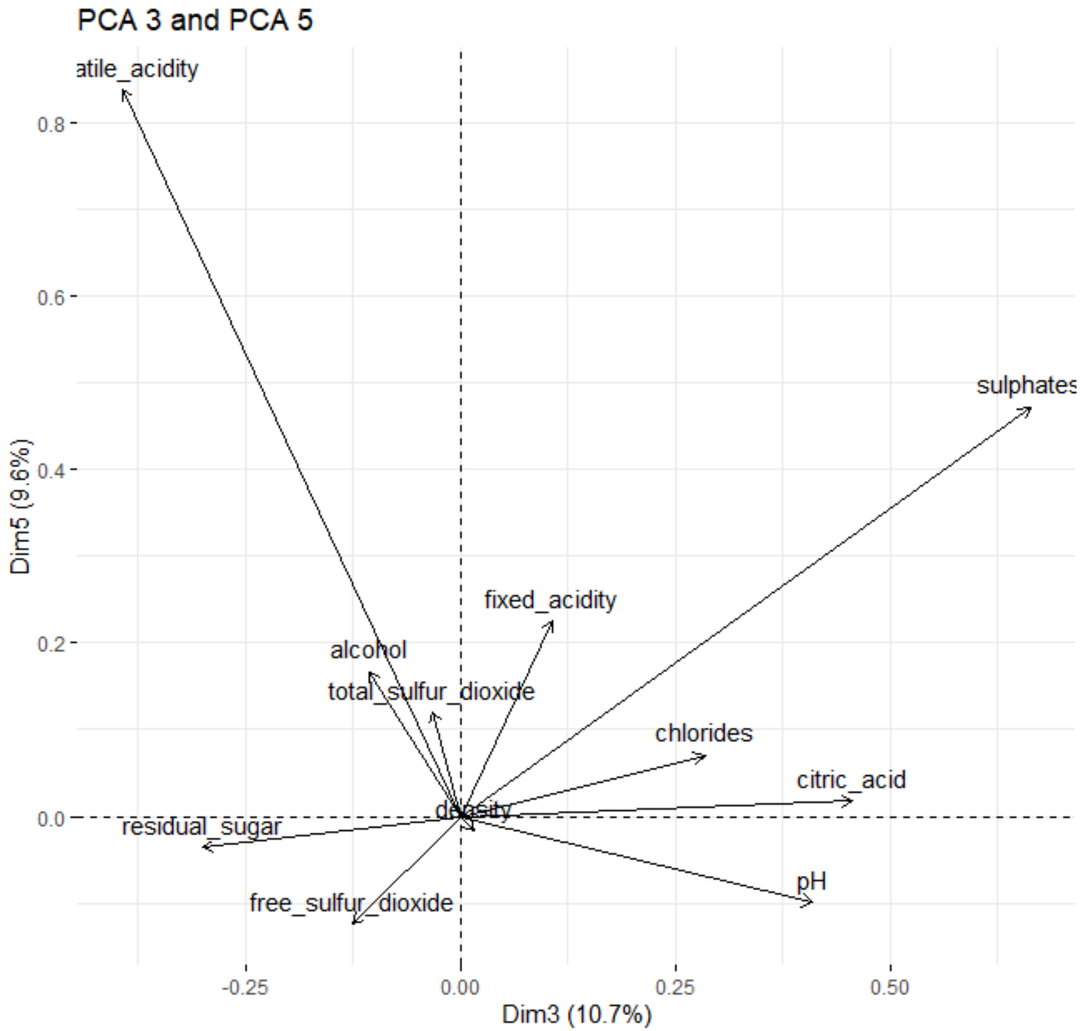
PRINCIPAL COMPONENTS ANALYSIS

```
> head(pca)
              PC1         PC2      PC3
38   -0.09207777   0.5036059 2.277253
155 -1.39337917 -0.9525258 2.736982
156 -1.38954992 -0.9271605 2.707217
157 -1.39337917 -0.9525258 2.736982
158 -1.38954992 -0.9271605 2.707217
310  0.35424132   0.2469293 2.315242
             PC4         PC5          PC6
38   -3.356240 2.7571187 -0.4209366
155 -1.004639 3.1083479  1.4162984
156 -1.097483 3.0720185  1.5147305
157 -1.004639 3.1083479  1.4162984
158 -1.097483 3.0720185  1.5147305
310 -2.991603 0.4658695  1.5681166
              PC7
38   -0.52178595
155 -0.03462939
156 -0.08623480
157 -0.03462939
158 -0.08623480
310  0.12036952
```



PRINCIPAL COMPONENTS ANALYSIS





PRINCIPAL COMPONENTS REGRESSION

```
> summary(lmodel)

Call:
lm(formula = quality ~ ., data = ols.data)

Residuals:
     Min       1Q   Median       3Q      Max
-3.11588 -0.66977  0.02529  0.62446  2.33707

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.096e-16  1.505e-02   0.000    1.000
PC1          1.600e-01  8.167e-03  19.587  < 2e-16 ***
PC2          4.218e-03  1.230e-02   0.343    0.732
PC3          9.750e-03  1.386e-02   0.704    0.482
PC4          1.993e-01  1.432e-02  13.916  < 2e-16 ***
PC5         -6.405e-02  1.469e-02  -4.361 1.33e-05 ***
PC6         -1.295e-01  1.725e-02  -7.508 7.46e-14 ***
PC7          1.393e-01  1.772e-02   7.861 4.94e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9162 on 3697 degrees of freedom
Multiple R-squared:  0.1621,    Adjusted R-squared:  0.1605
F-statistic: 102.2 on 7 and 3697 DF,  p-value: < 2.2e-16
```

FACTOR ANALYSIS AND ROTATION METHODS

There are a lot of rotation methods that can be used for factor analysis but we prefer Varimax (most preferred) and Quartimax since the estimated weights for the factor scores are unmatched for the other methods.

ANALYSIS METHODS

For our dataset, we've been researching for the best analysis method related to factor analysis such as principal axis, maximum likelihood, generalized weighted least squares methods.

	ML3	ML2	ML1
SS loadings	3.44	2.28	1.31
Proportion Var	0.26	0.18	0.10
Cumulative Var	0.26	0.44	0.54
Proportion Explained	0.49	0.32	0.19

BIC = 12703.87 Model fits the data.

Maximum Likelihood Method

	[,1]	[,2]	[,3]
SS loadings	3.29	2.22	1.42
Proportion Var	0.25	0.17	0.11
Cumulative Var	0.25	0.42	0.53
Proportion Explained	0.47	0.32	0.20

BIC = 14696.07 Model fits the data.

Ordinary Least Squares Method

VARIABLES AND PATTERNS

	ML3	ML2	ML1	h2	u2	com
fixed_acidity	-0.61	0.15	0.78	1.00	0.0050	2.0
volatile_acidity	-0.63	0.14	-0.23	0.47	0.5298	1.4
citric_acid	0.14	0.05	0.52	0.29	0.7078	1.2
residual_sugar	0.46	0.72	0.08	0.74	0.2582	1.7
chlorides	-0.49	0.24	-0.04	0.30	0.6978	1.5
free_sulfur_dioxide	0.57	0.20	0.04	0.36	0.6359	1.3
total_sulfur_dioxide	0.76	0.25	0.12	0.66	0.3427	1.3
density	-0.30	0.93	0.18	1.00	0.0049	1.3
pH	-0.28	0.03	-0.54	0.37	0.6277	1.5
sulphates	-0.48	0.13	-0.01	0.25	0.7511	1.1
alcohol	-0.05	-0.74	-0.02	0.56	0.4446	1.0
quality	0.10	-0.30	0.04	0.10	0.8969	1.3
color*	0.94	-0.14	0.13	0.93	0.0741	1.1

ML1 estimator is related to Fixed Acidity, Citric Acid, pH. So, we can correspond this factor loading as a Tendency to Acidity.

ML2 is related to Residual Sugar, Density, Alcohol, Quality. So, we can correspond this factor loading as a Sweetness Profile.

ML3 is related to Volatile Acidity, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Sulphates, Color. So, we can correspond this factor loading as a Wine Stability.

BEST AND WORST FEATURE THAT IS EXPLAINED BY FACTOR ANALYSIS

	ML3	ML2	ML1	h2	u2	com
fixed_acidity	-0.61	0.15	0.78	1.00	0.0050	2.0
volatile_acidity	-0.63	0.14	-0.23	0.47	0.5298	1.4
citric_acid	0.14	0.05	0.52	0.29	0.7078	1.2
residual_sugar	0.46	0.72	0.08	0.74	0.2582	1.7
chlorides	-0.49	0.24	-0.04	0.30	0.6978	1.5
free_sulfur_dioxide	0.57	0.20	0.04	0.36	0.6359	1.3
total_sulfur_dioxide	0.76	0.25	0.12	0.66	0.3427	1.3
density	-0.30	0.93	0.18	1.00	0.0049	1.3
pH	-0.28	0.03	-0.54	0.37	0.6277	1.5
sulphates	-0.48	0.13	-0.01	0.25	0.7511	1.1
alcohol	-0.05	-0.74	-0.02	0.56	0.4446	1.0
quality	0.10	-0.30	0.04	0.10	0.8969	1.3
color*	0.94	-0.14	0.13	0.93	0.0741	1.1

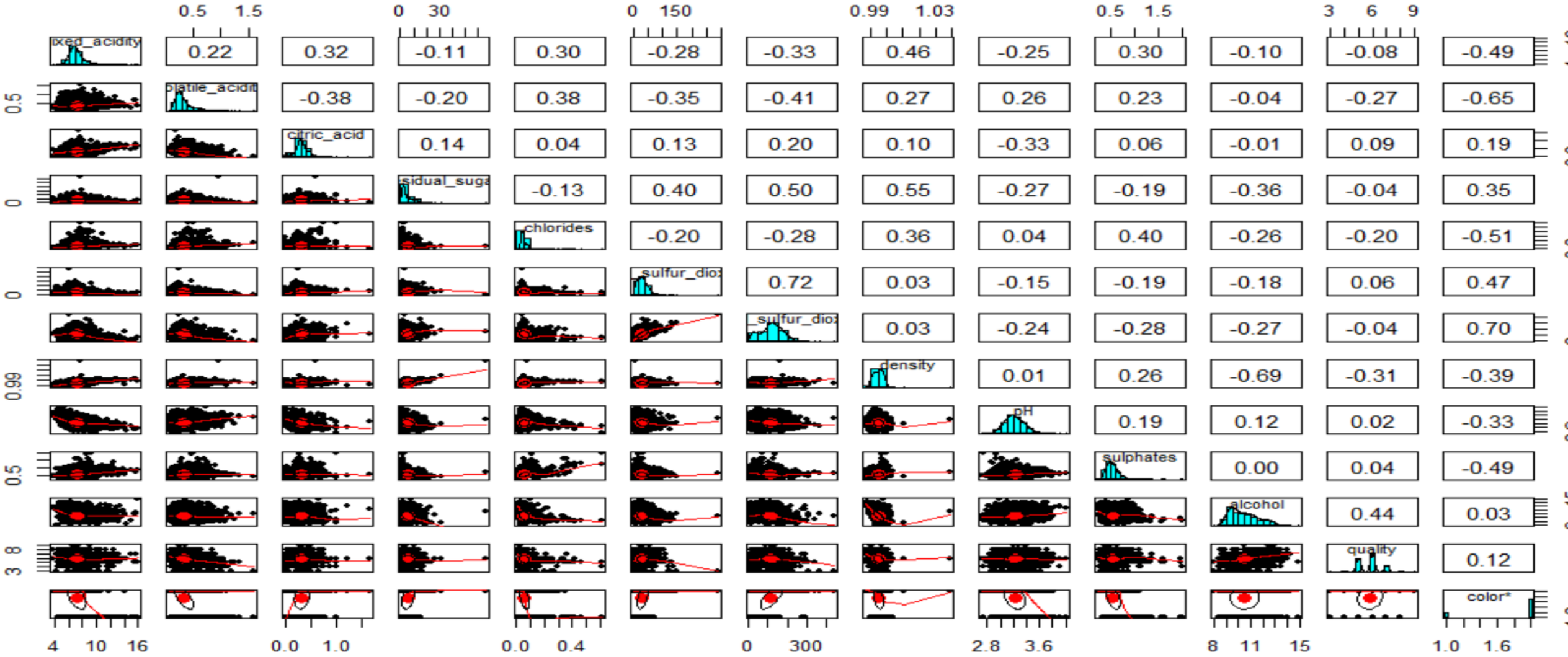
Since h^2 value of Color of Wine is equal to 0.93, it says 93% of the total variance in the variable is explained by the factors.

As for the worst, Quality variable has a 0.89 value of u^2, it says 89% of the total variance in the variable is failed to explained by the factors.

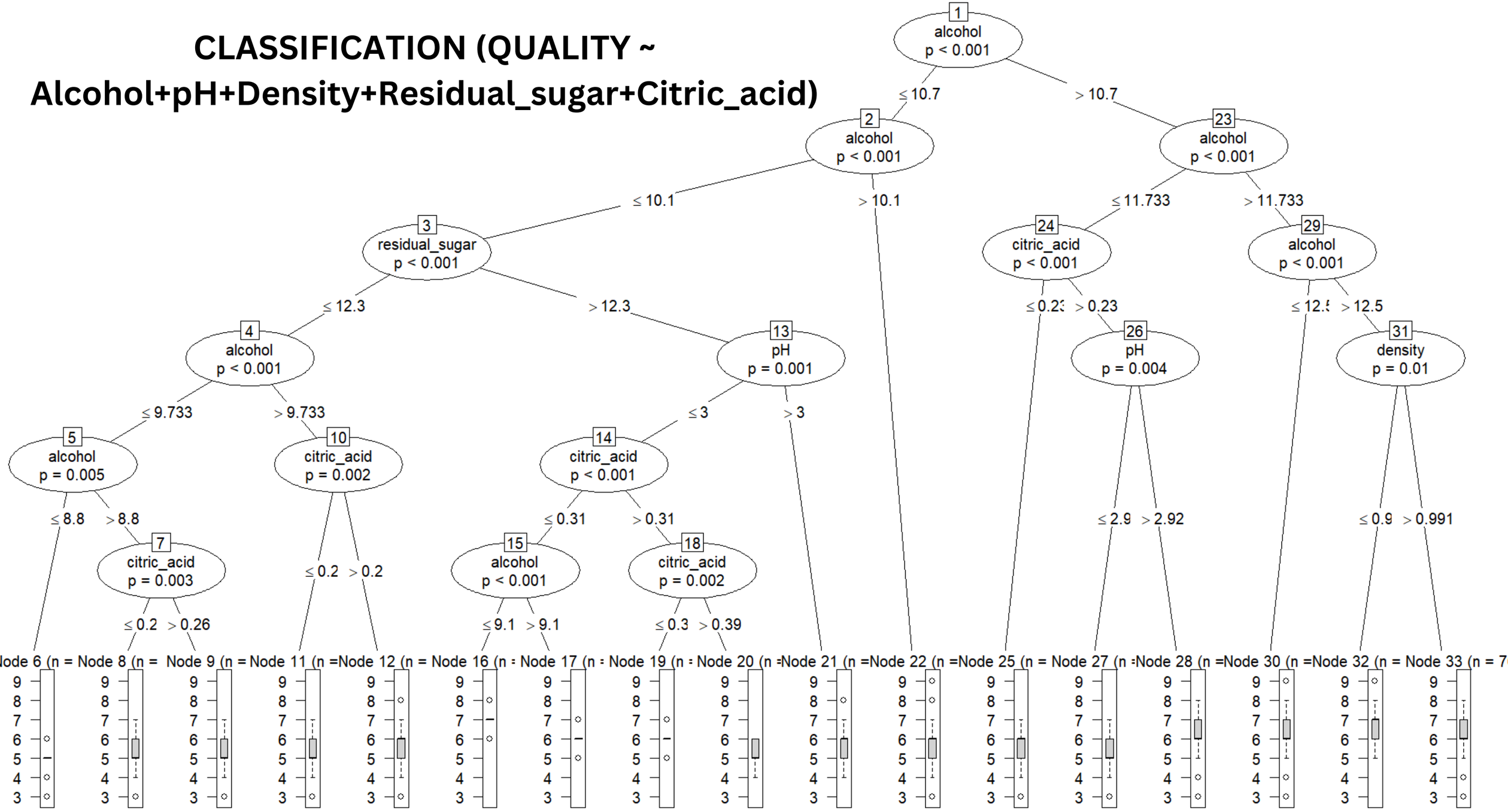
DISCRIMINATION AND CLASSIFICATION

We aim to visualize these variables and their relationships through a graphical representation, choosing a pair plot for clarity.

VARIABLES AND THEIR RELATIONS

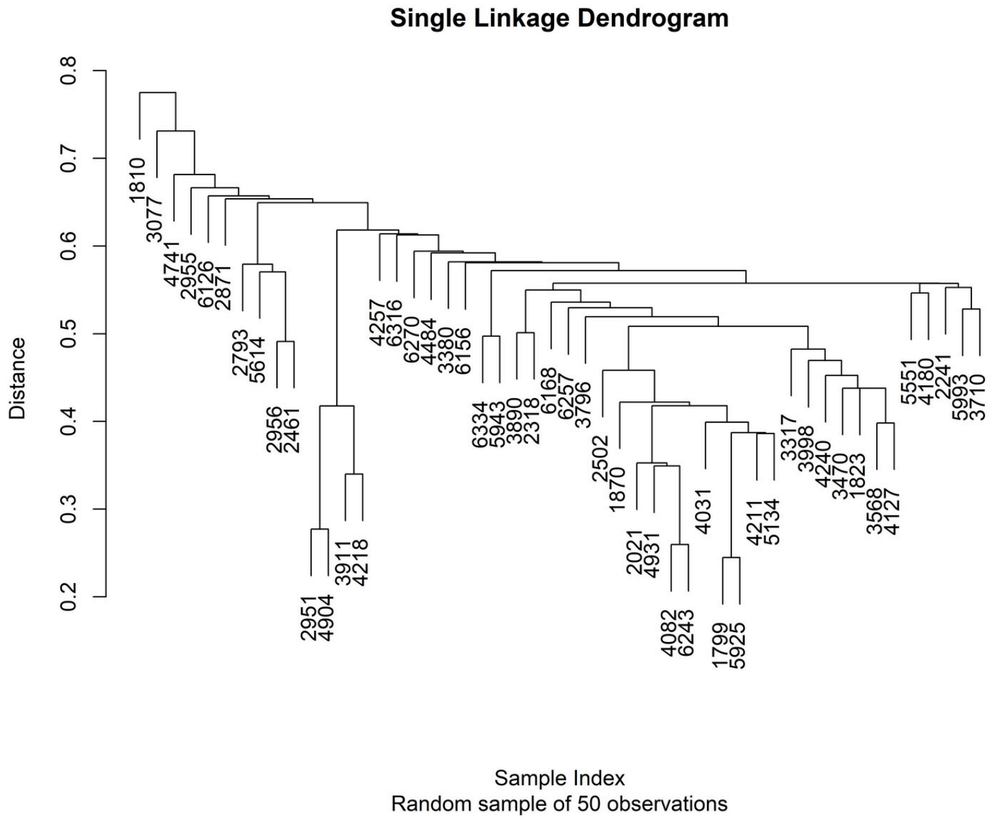


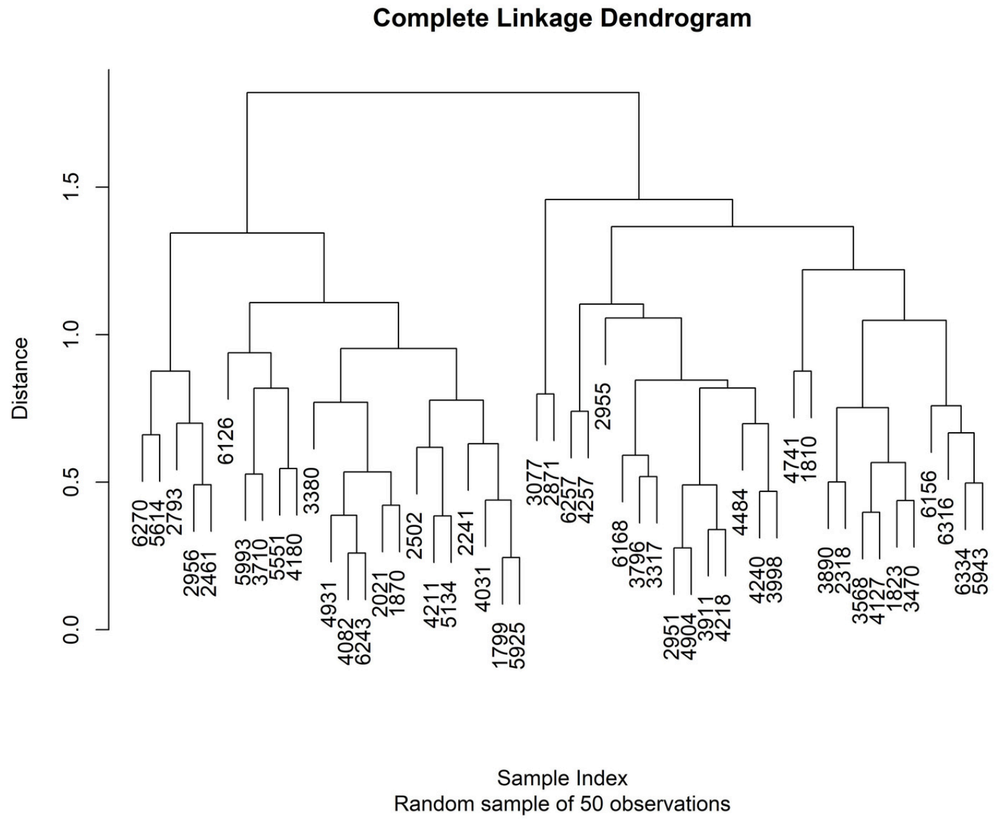
CLASSIFICATION (QUALITY ~ Alcohol+pH+Density+Residual_sugar+Citric_acid)



CLUSTERING

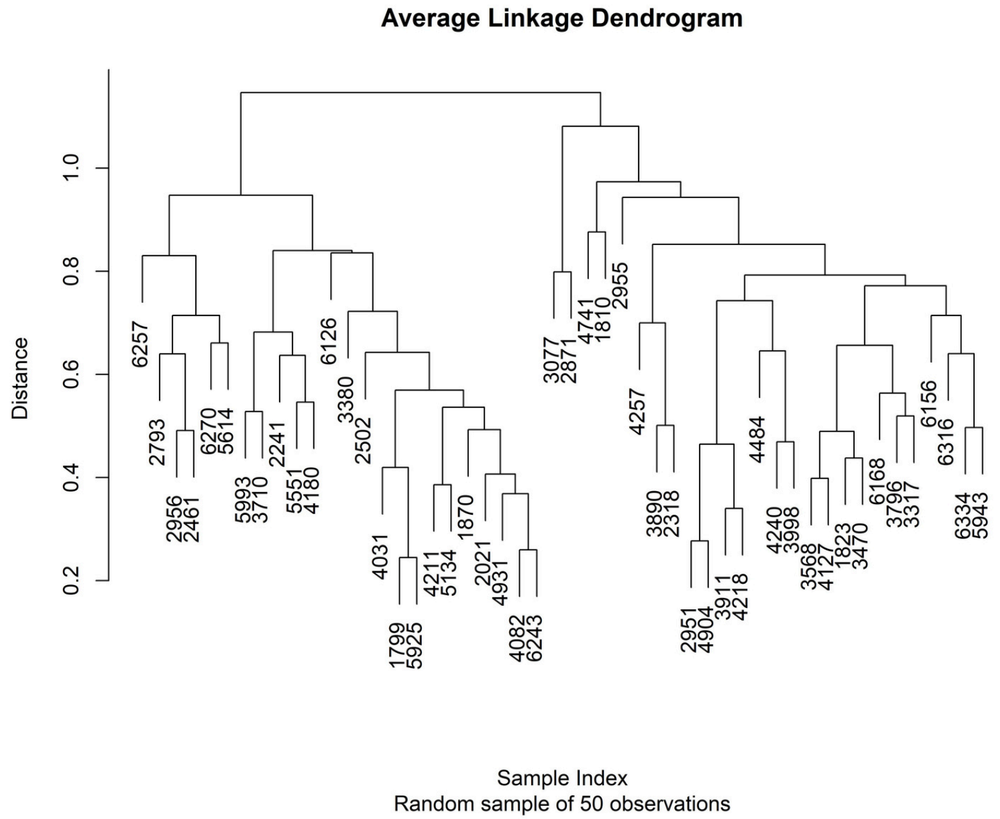
Hierarchical

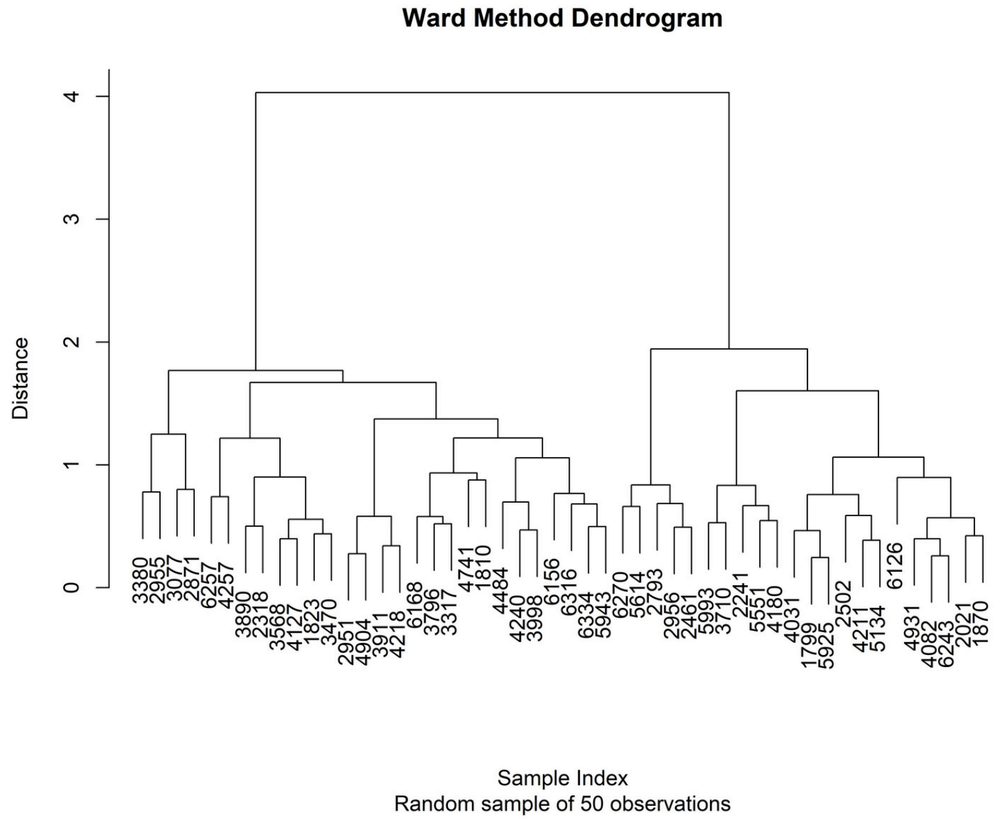




CLUSTERING

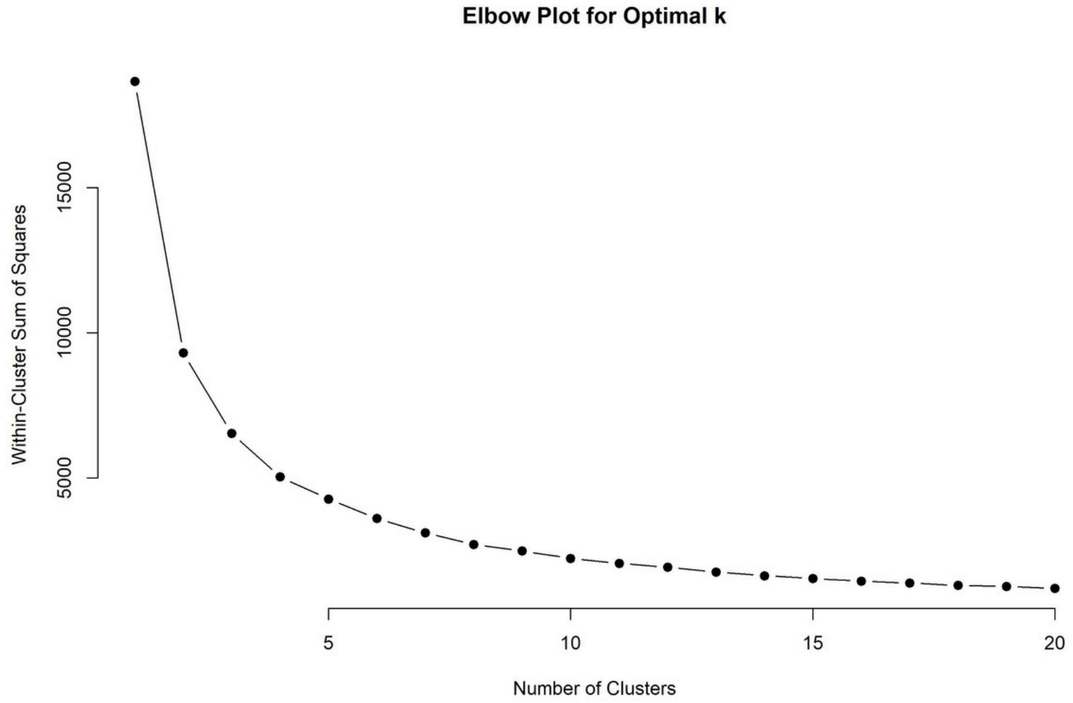
Hierarchical

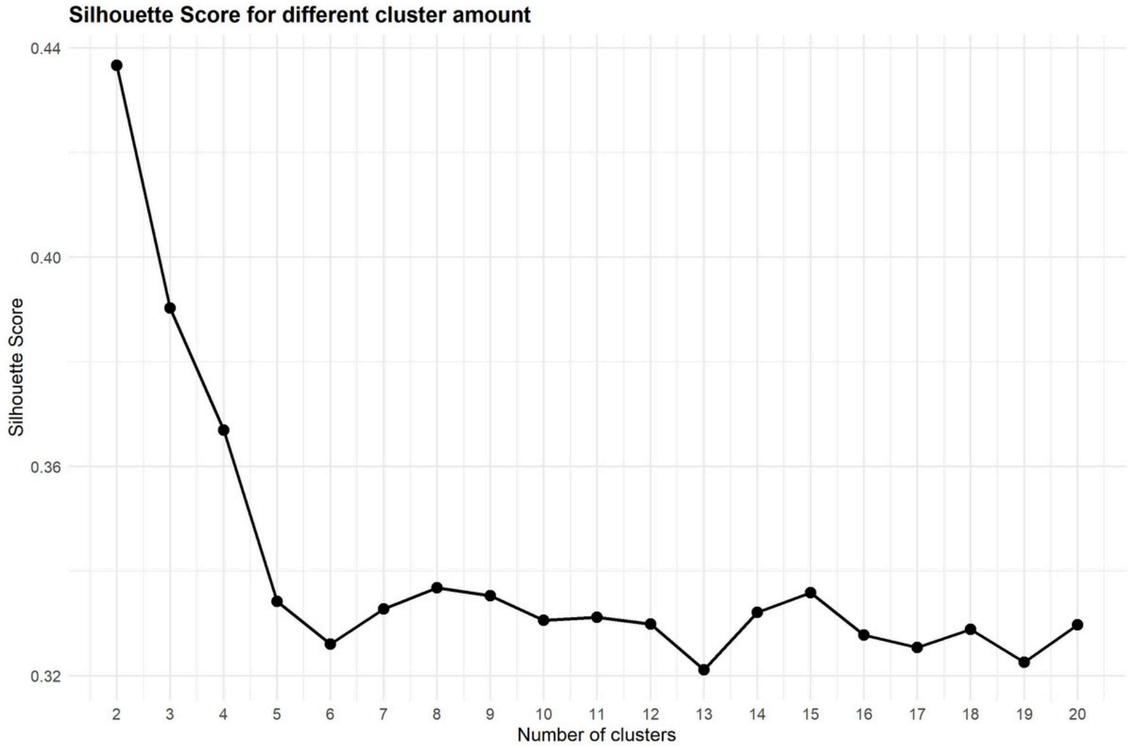




CLUSTERING

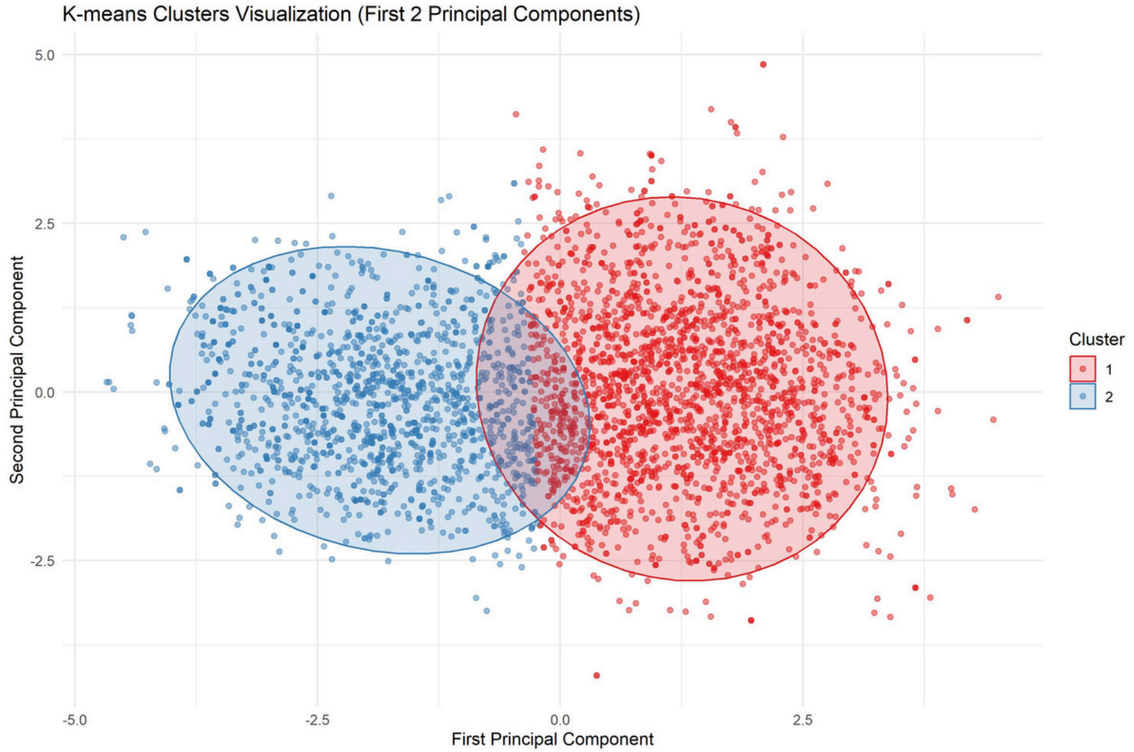
K-Means

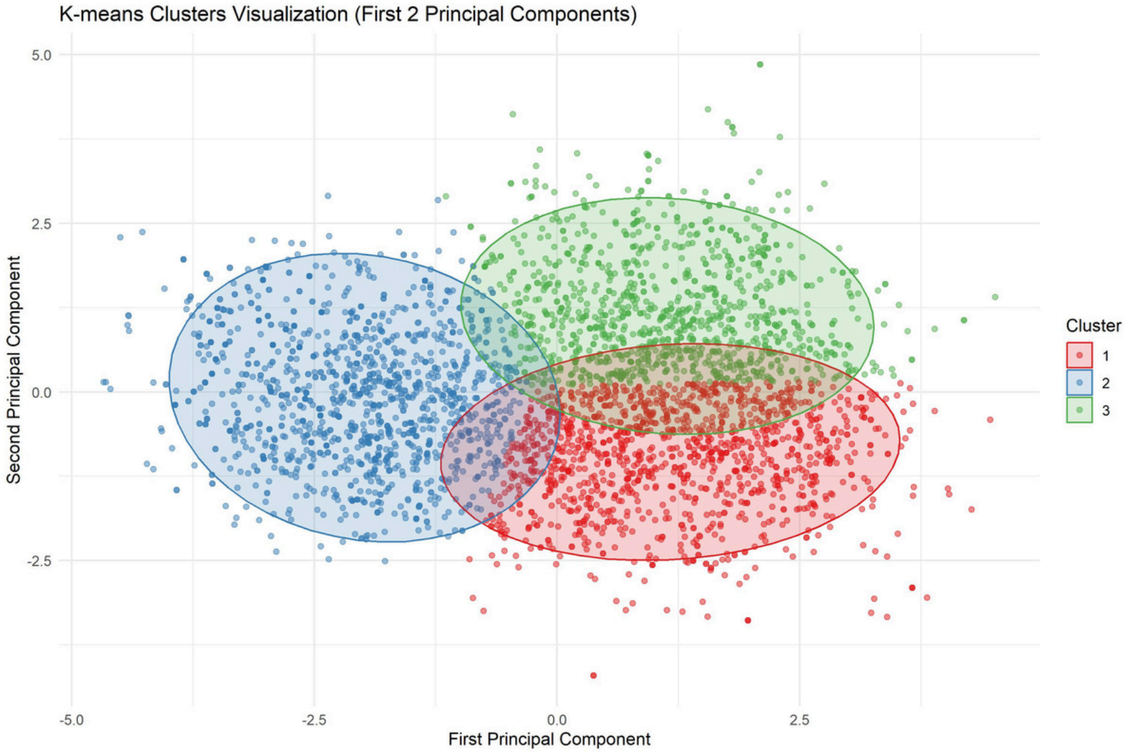




CLUSTERING

K-Means





CCA

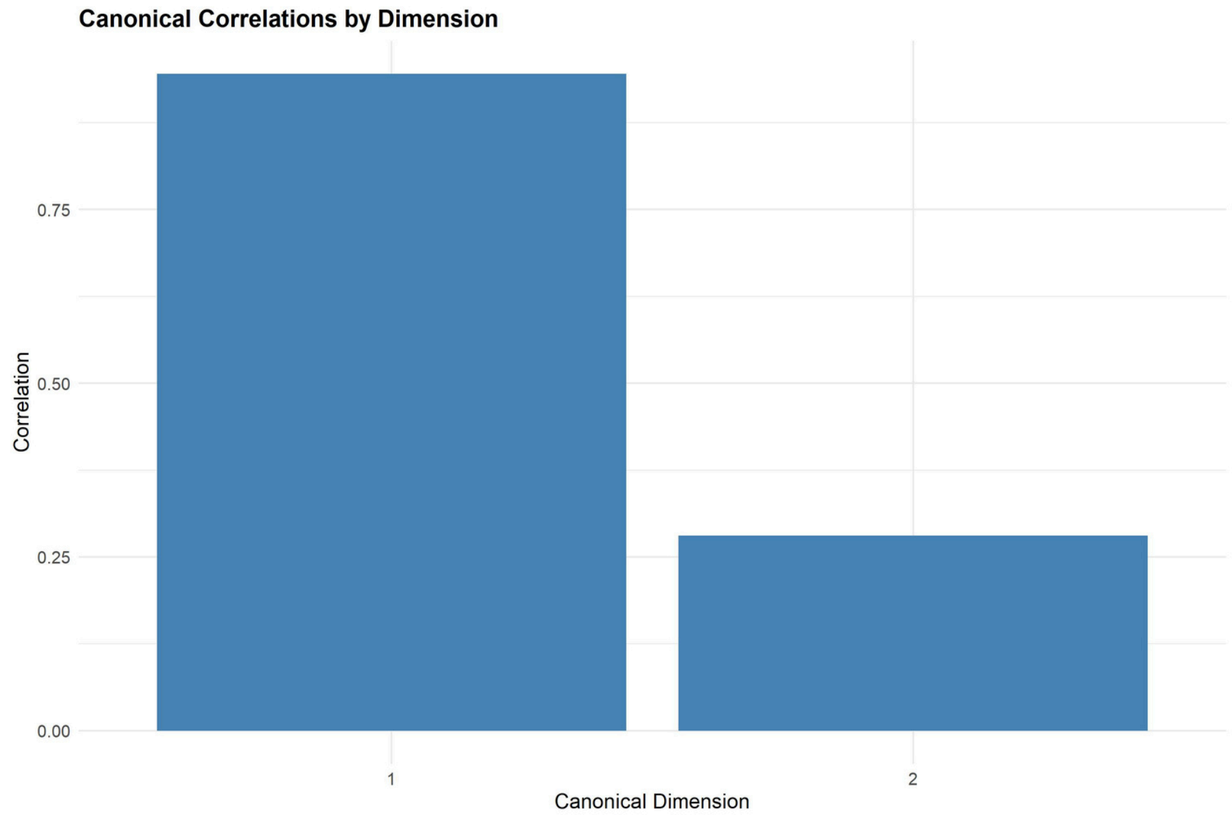
Features of the Wine

- fixed_acidity
- volatile_acidity
- citric_acid
- residual_sugar
- chlorides
- free_sulfur_dioxide
- total_sulfur_dioxide
- density
- pH
- sulphates

Responses

- alcohol
- quality

CCA




```
> p.asym(cca_result$cor, n, p, q, tstat = "Wilks")
Wilks' Lambda, using F-approximation (Rao's F):
               stat   approx df1  df2 p.value
1 to 2:  0.09821539 843.9263  20 7704       0
2 to 2:  0.92121920  36.6112   9 3853       0
```

CCA

```
$xcoef
                              [,1]          [,2]          [,3]          [,4]          [,5]          [,6]          [,7]          [,8]          [,9]          [,10]
fixed_acidity        -8.255103e-03 -1.613168e-03 -4.991130e-03 -7.220327e-03  7.415668e-03  1.406704e-03 -0.0040074578  0.0156053883 -1.592585e-02 -2.698613e-03
volatile_acidity     -1.771809e-02  1.333880e-01  3.150766e-02  9.546204e-02  3.886881e-02  7.734632e-02 -0.0539888400 -0.0374460624 -2.660643e-02 -5.373661e-02
citric_acid          -7.183468e-03  7.115340e-04  2.045824e-01  7.785046e-03 -4.112579e-03 -1.603061e-02 -0.0086390207 -0.0092121358  4.127932e-03  1.596504e-02
residual_sugar       -3.390414e-03 -3.025775e-03 -2.310143e-04  4.001277e-03 -9.464951e-04 -1.631095e-03  0.0001805842  0.0036279364 -1.831122e-03 -4.856156e-04
chlorides            -7.998196e-02  2.626506e-01 -4.098336e-04  1.124304e-01 -1.628610e+00 -2.726232e-02 -0.3334550568  0.9679406469 -2.292681e-01 -4.497120e-02
free_sulfur_dioxide   1.835910e-05 -4.263284e-04 -5.428822e-06 -1.325776e-04 -4.203248e-05  1.087511e-03 -0.0007888072 -0.0001652701  6.645632e-05  1.115670e-04
total_sulfur_dioxide  1.648582e-05  6.881169e-05  1.939561e-06  1.268899e-05  1.477944e-05  1.064496e-05  0.0005112672  0.0001618843 -4.701803e-05 -1.737873e-05
density               1.016806e+01  3.151752e+00  6.034626e-01 -3.634155e+00  4.546440e+00 -2.452156e-02 -2.5466400443 -7.5487995238  2.431802e+00  9.190540e-01
pH                   -3.726593e-02 -2.993412e-02 -2.489056e-03  7.957038e-03 -1.880053e-02 -2.993513e-03  0.0108116414 -0.0211720623 -1.247475e-01  1.953263e-02
sulphates            -1.792747e-02 -5.960948e-02 -1.879533e-03 -9.359193e-03 -1.430191e-02 -9.023040e-03  0.0088364935 -0.0095647395  1.195664e-02 -1.542182e-01

$ycoef
                [,1]        [,2]
alcohol -0.0134215501  0.006316683
quality -0.0004153857 -0.023409717

$xcenter
      fixed_acidity   volatile_acidity        citric_acid     residual_sugar          chlorides  free_sulfur_dioxide total_sulfur_dioxide            density
         6.83976449         0.26691382         0.32189182         6.16900880         0.04281004        34.10623706       135.25582298         0.99387597
                 pH          sulphates
         3.19189700         0.48480072

$ycenter
  alcohol   quality
10.578972  5.881988
```

CCA

```
$corr.X.xscores
                        [,1]        [,2]         [,3]        [,4]        [,5]         [,6]         [,7]        [,8]         [,9]         [,10]
fixed_acidity         0.07338023  0.14498328  0.060791109 -0.43992468  0.45748974 -0.034959352 -0.253310835  0.61973202 -0.305596996 -0.155638052
volatile_acidity     -0.10414682  0.66749145  0.019175134  0.53540429  0.14892270  0.285641730 -0.109399488 -0.08627773 -0.170039463 -0.323246789
citric_acid          -0.02241931 -0.08173228  0.967321500 -0.16663961  0.07110988 -0.048441487 -0.035472343  0.13803126 -0.037283475  0.015277399
residual_sugar        0.50823649 -0.37915069  0.003624151  0.65781724  0.18593128 -0.084762099 -0.096550036  0.32028349 -0.107438630  0.001495136
chlorides             0.53549616  0.19594209  0.027994471  0.01199054 -0.67445539  0.005375436 -0.164148490  0.36860724 -0.201577699 -0.124717822
free_sulfur_dioxide   0.27614093 -0.46041238  0.069624392  0.11467862 -0.01868219  0.818080986 -0.004239598  0.13036353  0.007788292  0.084196635
total_sulfur_dioxide  0.48508772 -0.09544151  0.095567342  0.17795312  0.04116799  0.479396905  0.607918114  0.30218374 -0.137943475 -0.057933992
density               0.85507690 -0.17850153  0.035005624  0.28436558  0.13091466 -0.063645111 -0.104665759  0.19976490 -0.271363817 -0.095931846
pH                   -0.06750050 -0.12280381 -0.019533275 -0.01150834 -0.30201761 -0.003752170  0.152320260 -0.59225290 -0.713503439  0.074619631
sulphates             0.06280753 -0.25720792  0.072728135 -0.15505425 -0.13849144 -0.011551518  0.059669555 -0.17571764 -0.056060641 -0.918517742

$corr.Y.xscores
              [,1]         [,2]          [,3]         [,4]          [,5]          [,6]         [,7]         [,8]          [,9]         [,10]
alcohol -0.9450419  0.004979632 -5.165005e-16 1.957909e-15 -1.843174e-15 -6.103785e-16 1.283771e-15 3.504828e-15 -2.466439e-15 3.563178e-16
quality -0.4024938 -0.253958967 -2.471846e-16 8.254339e-16 -4.955094e-16 -1.017174e-16 1.963456e-16 8.853033e-16 -9.753260e-16 3.117138e-16

$corr.X.yscores
                        [,1]        [,2]
fixed_acidity         0.06935831  0.04069379
volatile_acidity     -0.09843860  0.18735095
citric_acid          -0.02119052 -0.02294055
residual_sugar        0.48038040 -0.10641970
chlorides             0.50614599  0.05499687
free_sulfur_dioxide   0.26100584 -0.12922817
total_sulfur_dioxide  0.45850040 -0.02678845
density               0.80821072 -0.05010166
pH                   -0.06380085 -0.03446847
sulphates             0.05936509 -0.07219291

$corr.Y.yscores
              [,1]        [,2]
alcohol -0.9998426  0.01774137
quality -0.4258334 -0.90480158
```

CONCLUSION

THANK YOU

REFERENCES

- Dataset: https://www.kaggle.com/datasets/ghassenkhaled/wine-quality-data
- Alice, M. (2016, July 21). Performing Principal Components Regression (PCR) in R | R-Bloggers. R-bloggers. https://www.r-bloggers.com/2016/07/performing-principal-components-regression-pcr-in-r/
- Bobbitt, Z. (2020, December 1). Principal components analysis in R: Step-by-Step example. Statology. https://www.statology.org/principal-components-analysis-in-r/
- Canonical Correlation Analysis | R Data Analysis Examples. (n.d.). https://stats.oarc.ucla.edu/r/dae/canonical-correlation-analysis/
- GeeksforGeeks. (2020, August 25). MANOVA Test in R Programming. GeeksforGeeks. https://www.geeksforgeeks.org/manova-test-in-r-programming/
- GeeksforGeeks. (2024, May 14). What is Canonical Correlation Analysis? GeeksforGeeks. https://www.geeksforgeeks.org/what-is-canonical-correlation-analysis
- Lecture Notes