Analysis of quality of wine

The dataset employed in this project pertains to red wine quality. It encompasses 12 variables and comprises a total of 1599 observations. Within this dataset, these variables represent potential factors that could impact the quality of the wine. The primary objective of this analysis is to identify the variable or variables that exert the most significant influence on wine quality. Additionally, we aim to predict the quality of the wine itself. This particular dataset was selected due to its similarity to the data we previously examined.

Exploring basic data statistics

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar  
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900  
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900  
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200  
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539  
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600  
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   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 :0.9968  
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967  
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978  
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037  
##        pH          sulphates         alcohol         quality     
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000  
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000  
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000  
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636  
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000  
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000  
```

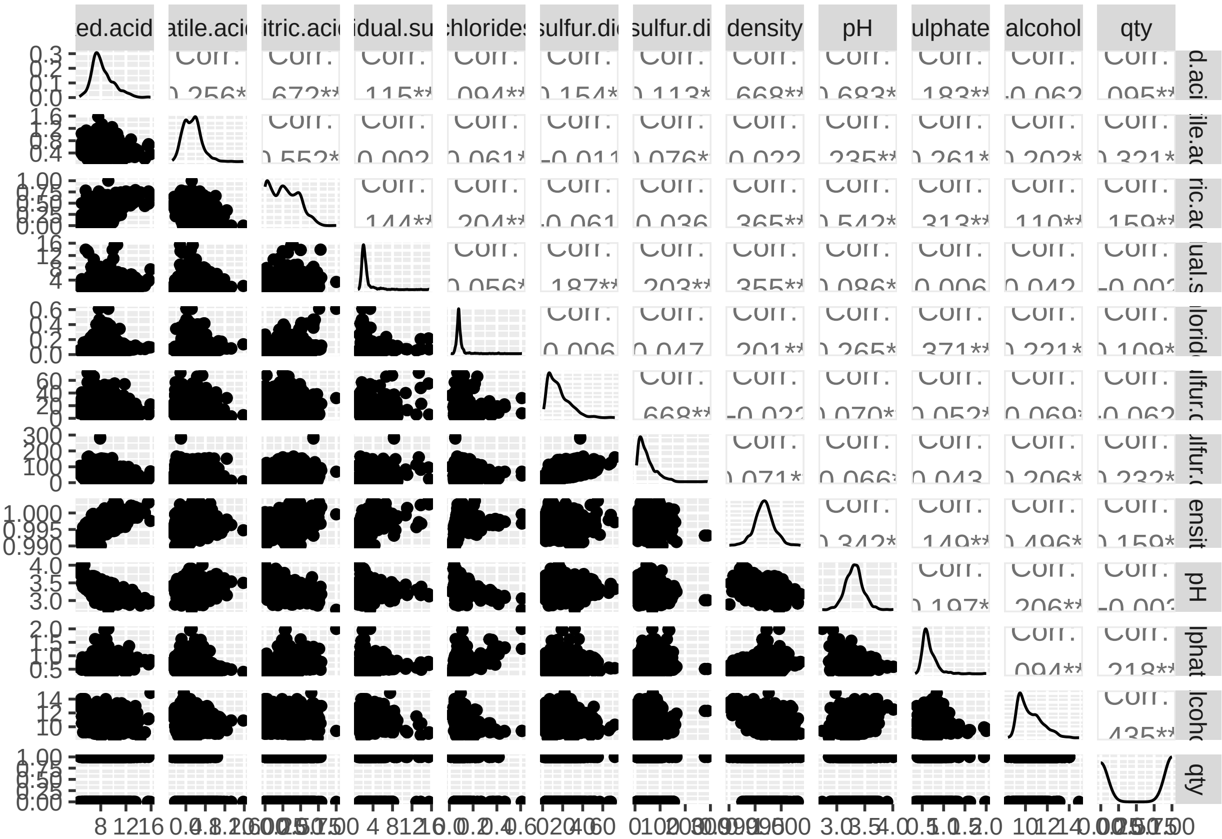
Originally, the wine quality was assessed on a scale ranging from 1 to 10. However, I have redefined the wine quality, classifying ratings less than or equal to 5 as “low,” indicated as Zero (0) in the dataset, and ratings greater than 5 as “high,” designated as One (1) in the data.

Now, I want to find which variable is mostly correlated with the wine data.

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##                      fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity           1.00000000     -0.256130895  0.67170343    0.114776724
## volatile.acidity       -0.25613089      1.000000000 -0.55249568    0.001917882
## citric.acid             0.67170343     -0.552495685  1.00000000    0.143577162
## residual.sugar          0.11477672      0.001917882  0.14357716    1.000000000
## chlorides               0.09370519      0.061297772  0.20382291    0.055609535
## free.sulfur.dioxide    -0.15379419     -0.010503827 -0.06097813    0.187048995
## total.sulfur.dioxide   -0.11318144      0.076470005  0.03553302    0.203027882
## density                 0.66804729      0.022026232  0.36494718    0.355283371
## pH                     -0.68297819      0.234937294 -0.54190414   -0.085652422
## sulphates               0.18300566     -0.260986685  0.31277004    0.005527121
## alcohol                -0.06166827     -0.202288027  0.10990325    0.042075437
## qty                     0.09509349     -0.321440854  0.15912941   -0.002160450
##                         chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity         0.093705186        -0.153794193          -0.11318144
## volatile.acidity      0.061297772        -0.010503827           0.07647000
## citric.acid           0.203822914        -0.060978129           0.03553302
## residual.sugar        0.055609535         0.187048995           0.20302788
## chlorides             1.000000000         0.005562147           0.04740047
## free.sulfur.dioxide   0.005562147         1.000000000           0.66766645
## total.sulfur.dioxide  0.047400468         0.667666450           1.00000000
## density               0.200632327        -0.021945831           0.07126948
## pH                   -0.265026131         0.070377499          -0.06649456
## sulphates             0.371260481         0.051657572           0.04294684
## alcohol              -0.221140545        -0.069408354          -0.20565394
## qty                  -0.109493996        -0.061756744          -0.23196298
##                          density           pH    sulphates     alcohol
## fixed.acidity         0.66804729 -0.682978195  0.183005664 -0.06166827
## volatile.acidity      0.02202623  0.234937294 -0.260986685 -0.20228803
## citric.acid           0.36494718 -0.541904145  0.312770044  0.10990325
## residual.sugar        0.35528337 -0.085652422  0.005527121  0.04207544
## chlorides             0.20063233 -0.265026131  0.371260481 -0.22114054
## free.sulfur.dioxide  -0.02194583  0.070377499  0.051657572 -0.06940835
## total.sulfur.dioxide  0.07126948 -0.066494559  0.042946836 -0.20565394
## density               1.00000000 -0.341699335  0.148506412 -0.49617977
## pH                   -0.34169933  1.000000000 -0.196647602  0.20563251
## sulphates             0.14850641 -0.196647602  1.000000000  0.09359475
## alcohol              -0.49617977  0.205632509  0.093594750  1.00000000
## qty                  -0.15910997 -0.003263984  0.218071663  0.43475120
##                               qty
## fixed.acidity         0.095093490
## volatile.acidity     -0.321440854
## citric.acid           0.159129408
## residual.sugar       -0.002160450
## chlorides            -0.109493996
## free.sulfur.dioxide  -0.061756744
## total.sulfur.dioxide -0.231962976
## density              -0.159109969
## pH                   -0.003263984
## sulphates             0.218071663
## alcohol               0.434751205
## qty                   1.000000000
```

```
##                  qty              alcohol            sulphates
##                 TRUE                 TRUE                FALSE
##          citric.acid        fixed.acidity       residual.sugar
##                FALSE                FALSE                FALSE
##                   pH  free.sulfur.dioxide            chlorides
##                FALSE                FALSE                FALSE
##              density total.sulfur.dioxide     volatile.acidity
##                FALSE                FALSE                 TRUE
```



I've chosen to identify variables with a correlation coefficient exceeding 0.3. Based on the correlation analysis, it appears that volatile acidity and alcohol exhibit the strongest correlations with wine quality. Alcohol demonstrates a positive correlation, while volatile acidity displays a negative correlation coefficient.

Splitting data

I partitioned the data into training and testing sets using a 60% to 40% ratio, facilitated by the caTools library function. The objective is to evaluate the influence of variables on wine quality by applying three different models. The first model in line is logistic regression.

With Logistic Regression

```
##
## Call:
## glm(formula = qty ~ volatile.acidity + alcohol, family = binomial,
##     data = tr.data)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -7.12629    0.95067  -7.496 6.58e-14 ***
## volatile.acidity -3.98603    0.47478  -8.395  < 2e-16 ***
## alcohol           0.91270    0.08961  10.185  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1287.7  on 932  degrees of freedom
## Residual deviance: 1031.8  on 930  degrees of freedom
## AIC: 1037.8
##
## Number of Fisher Scoring iterations: 4
```

```
## Confusion Matrix:
```

```
##
## preds   0   1
##     0 230  85
##     1  84 267
```

```
## Overall test accuracy (percentage) : 74.62
```

```
## [1] "True Positive Rate, TPR (percentage):"
## [1] 75.85
## [1] "False Postive Rate, FPR (percentage):"
## [1] 26.75
```

As volatile acidity and alcohol exhibit strong associations with wine quality, I employed a logistic model focusing on these variables. The results from the logistic model indicate that both volatile acidity and alcohol are statistically significant. The estimated coefficient for volatile acidity is -3.02073, signifying that when other predictors in the model remain constant, we can expect a mean decrease in log-odds with a unit increase in wine quality. Similarly, the estimated coefficient for alcohol is 1.10115, suggesting that, under constant conditions for the other predictors, a unit increase in wine quality leads to a mean increase in log-odds.

In the confusion matrix of the logistic regression, the model achieved a test accuracy of 72.11%. The true positive rate stood at 70.9%, and the false positive rate was 26.52%, which is considered favorable.

With LDA

```
## [1] "Statistics for the LDA"
## [1] "Confusion Matrix:"
##
## preds   0   1
##     0 234  86
##     1  80 266
## [1] "Model Accuracy (Percentage):"
## [1] 75.08
## [1] "True Positive Rate, TPR (percentage):"
## [1] 75.57
## [1] "False Postive Rate, FPR (percentage):"
## [1] 25.48
```

The LDA model reveals that both logistic regression and LDA produce comparable outcomes. The accuracy, true positive rate, and false positive rate in the LDA model are nearly identical to those in the logistic regression model.

With QDA

```
## [1] "Statistics for the QDA"
```

```
## [1] "Confusion Matrix:"
##
## preds   0   1
##     0 253 107
##     1  61 245
## [1] "Model Accuracy (Percentage):"
## [1] 74.77
## [1] "True Positive Rate, TPR (percentage):"
## [1] 69.6
## [1] "False Postive Rate, FPR (percentage):"
## [1] 19.43
```

In the QDA model, the accuracy stands at 71.06%, which is slightly lower than the other models. The true positive rate is 65.54%, also lagging behind the other models. However, it's noteworthy that the false positive rate is 22.68%, which is 5% and 2% lower than the other models, indicating a more favorable performance in this aspect.