final_project

ChihChin Yang(Adela), ChunYu Lo(Jack), SzuWei Fu(Alexia)

2022/3/13

1. Title Page:

ChihChin Yang(Adela), ChunYu Lo(Jack), SzuWei Fu(Alexia)

2. Introduction:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(lattice)
library(ggfortify)
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

```
white_wine <- read.table(file = 'winequality-white.csv',header = TRUE,sep=";")
red_wine <- read.table(file = 'winequality-red.csv',header = TRUE,sep=";")
```

- We find two data sets, one is red wine and the other one is white wine from the north of Portugal. Sample collection is uneven, most of the observation are normal wines and there are only few excellent or poor wines. White wine have 4898 observation of 12 variables.; Red wine have 1599 observation of 12 variables. All variables are quantitative variables.
- This dataset has several variables, including a dependent variable, wine quality. We can make a prediction linear model to predict the wine quality. We want to know which variable contributes the most to wine quality rather than just predict the dependent variable. Thus, the regression model is appropriate to this dataset.

```
# All variables are quantitative
str(white_wine)
```

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

```
str(red_wine)
```

```
## '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 ...
```

```
summary(white_wine)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

```
summary(red_wine)
```

```
##  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
```

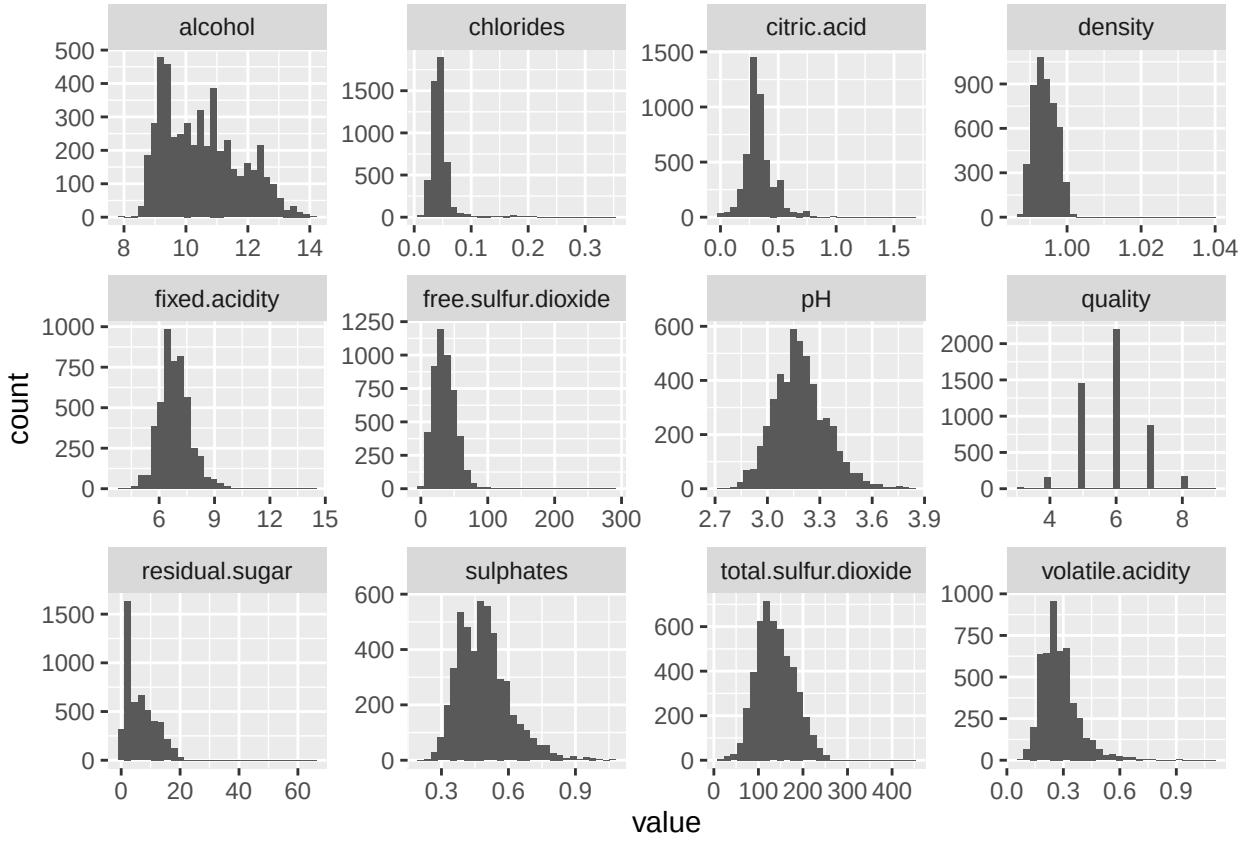
- Input variables (based on physicochemical tests):

1. fixed acidity: most acids involved with wine that are fixed or nonvolatile
2. volatile acidity: amount of acetic acid in wine

3. citric acid: can add 'freshness' and flavor to wines
4. residual sugar: amount of sugar remaining after fermentation stops
5. chlorides: amount of salt in the wine
6. free sulfur dioxide: prevents microbial growth and the oxidation of wine(free form)
7. total sulfur dioxide: free and bound forms of S02; in low concentrations
8. density: closeness of density to that of water depending on the percent alcohol and sugar content
9. pH: describes how acidic a wine is on a scale from 0 (very acidic) to 14 (very basic)
10. sulphates: Wine additive which can contribute to sulfur dioxide gas (S02) levels. Acts as an antimicrobial and antioxidant
11. alcohol: the percent alcohol content of the wine (% by volume)

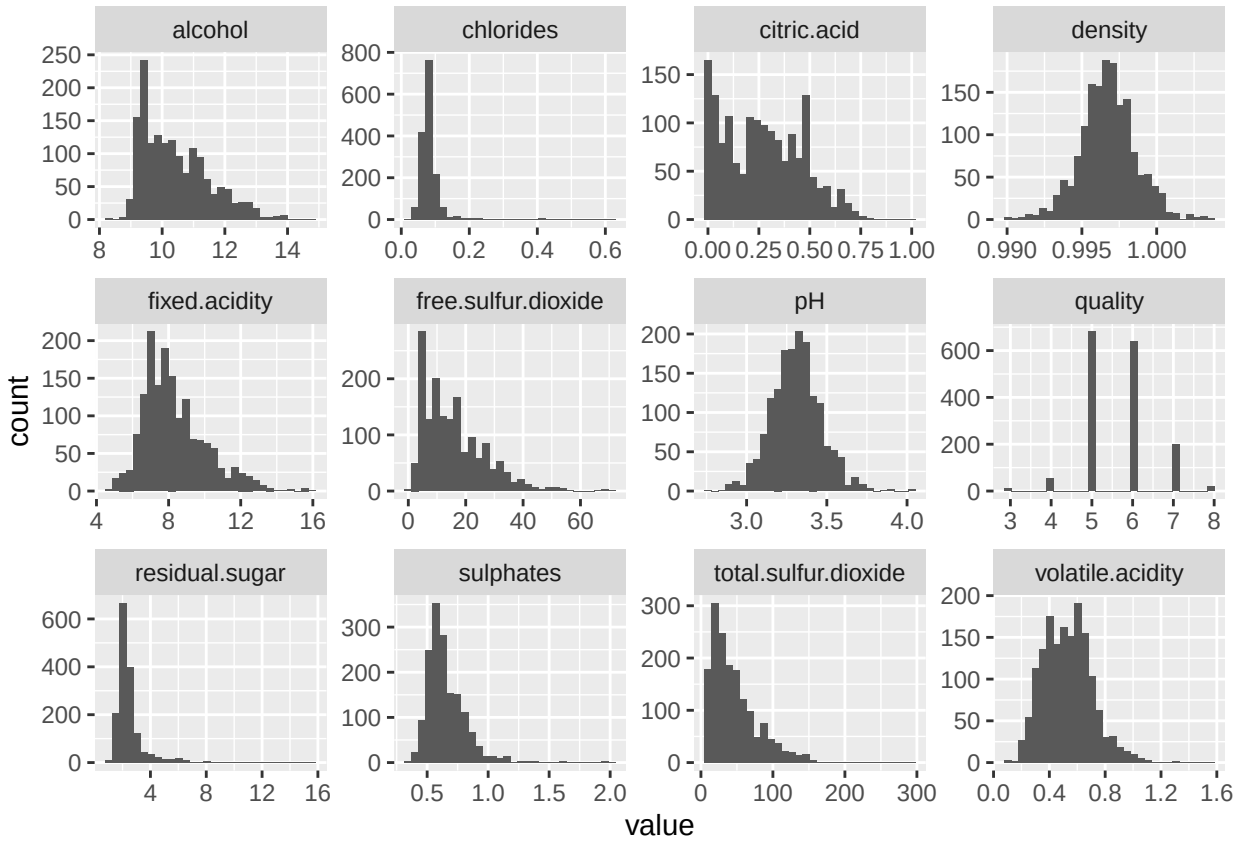
```
ggplot(gather(white_wine), aes(value)) +
  geom_histogram() +
  facet_wrap(~key, scales = "free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(gather(red_wine), aes(value)) +
  geom_histogram() +
  facet_wrap(~key, scales = "free")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



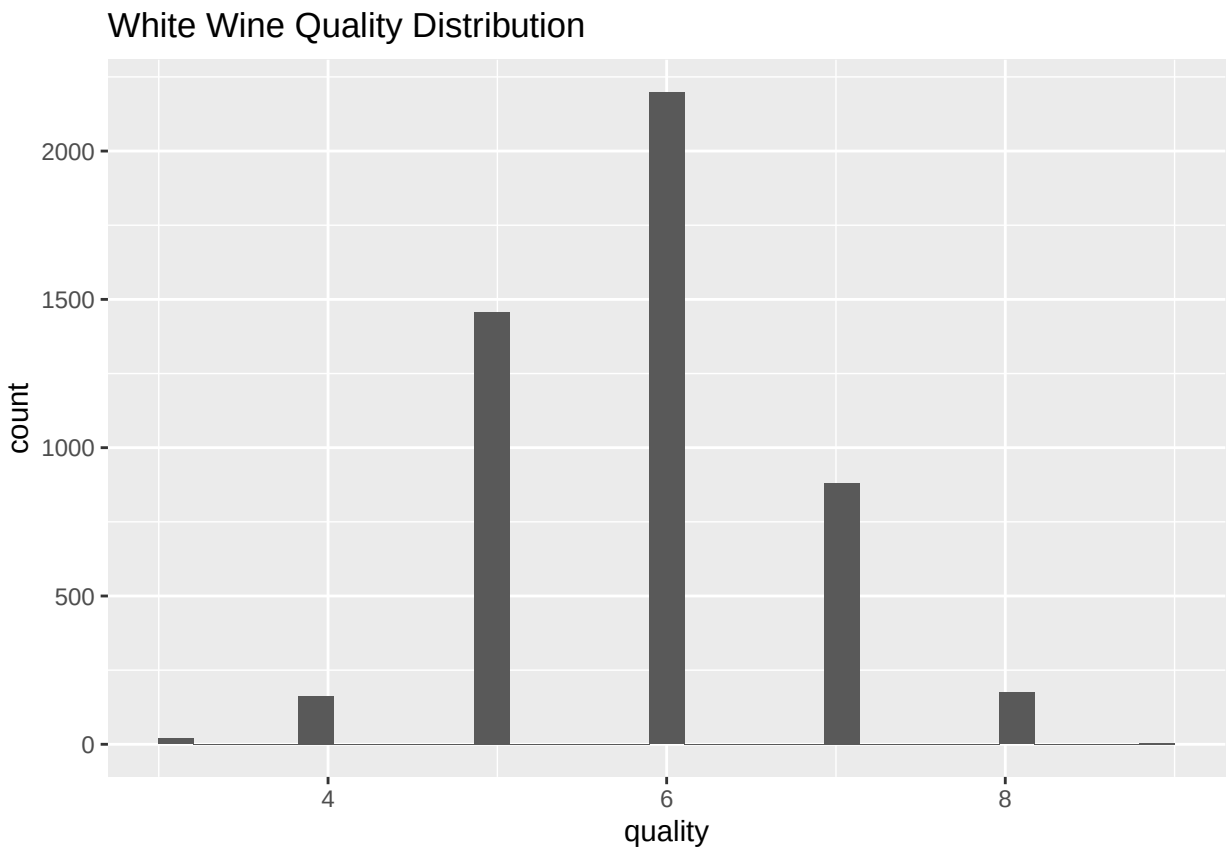
- Density and pH seems normally distributed. Majority of other variables are skewed to the right.

- Output variable (based on sensory data):

12. quality (score between 0 and 10)

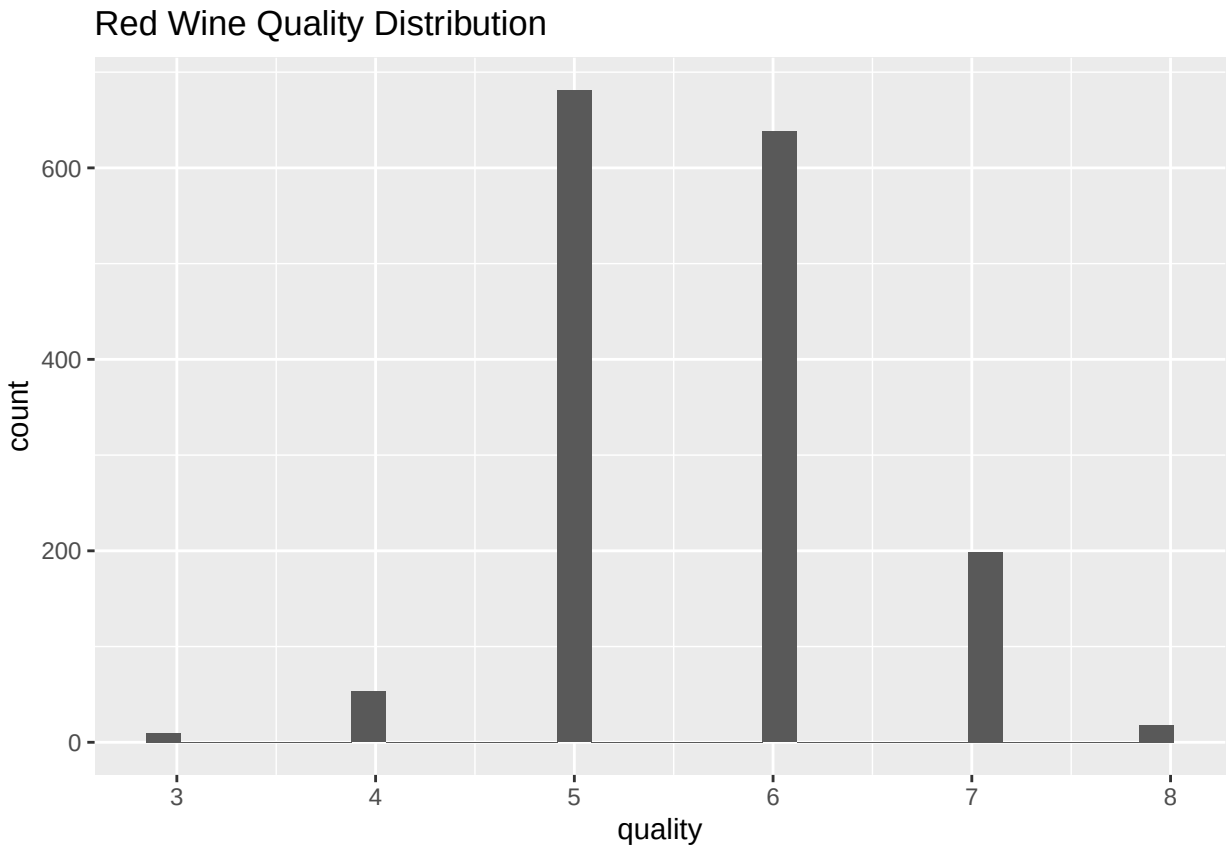
```
ggplot(white_wine, aes(quality))+
  geom_histogram()+
  ggtitle("White Wine Quality Distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(red_wine, aes(quality))+
  geom_histogram()+
  ggtitle("Red Wine Quality Distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- The dependent variable is discrete variable (quality rating ranges from 3 to 8, mostly between 5 and 6). It has only integer but it's continuous, not categorical, because rational number also has a meaning. Therefore, we could make a linear model without any conversion.
- The research question is which critical factors can influence the wine quality.
- The goal is to model wine quality based on physicochemical tests.

3. Initial Hypotheses:

Model assumtion:

$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{11} X_{11} + \varepsilon$

Initial hypothese:

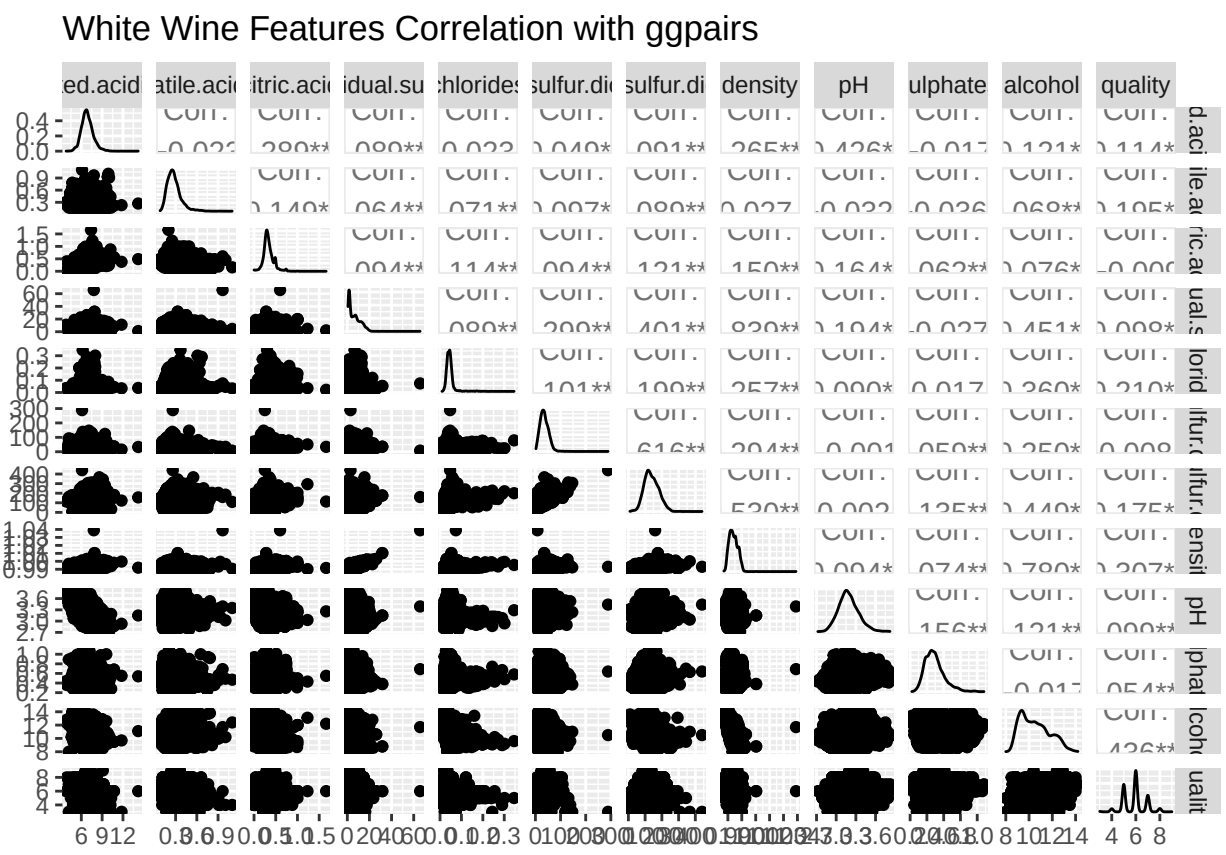
$H_0 = \beta_1 = \beta_2 = ... = \beta_{11} = 0$

$H_1 = \beta_i \neq 0$ for some $i = 1, 2, ...11$

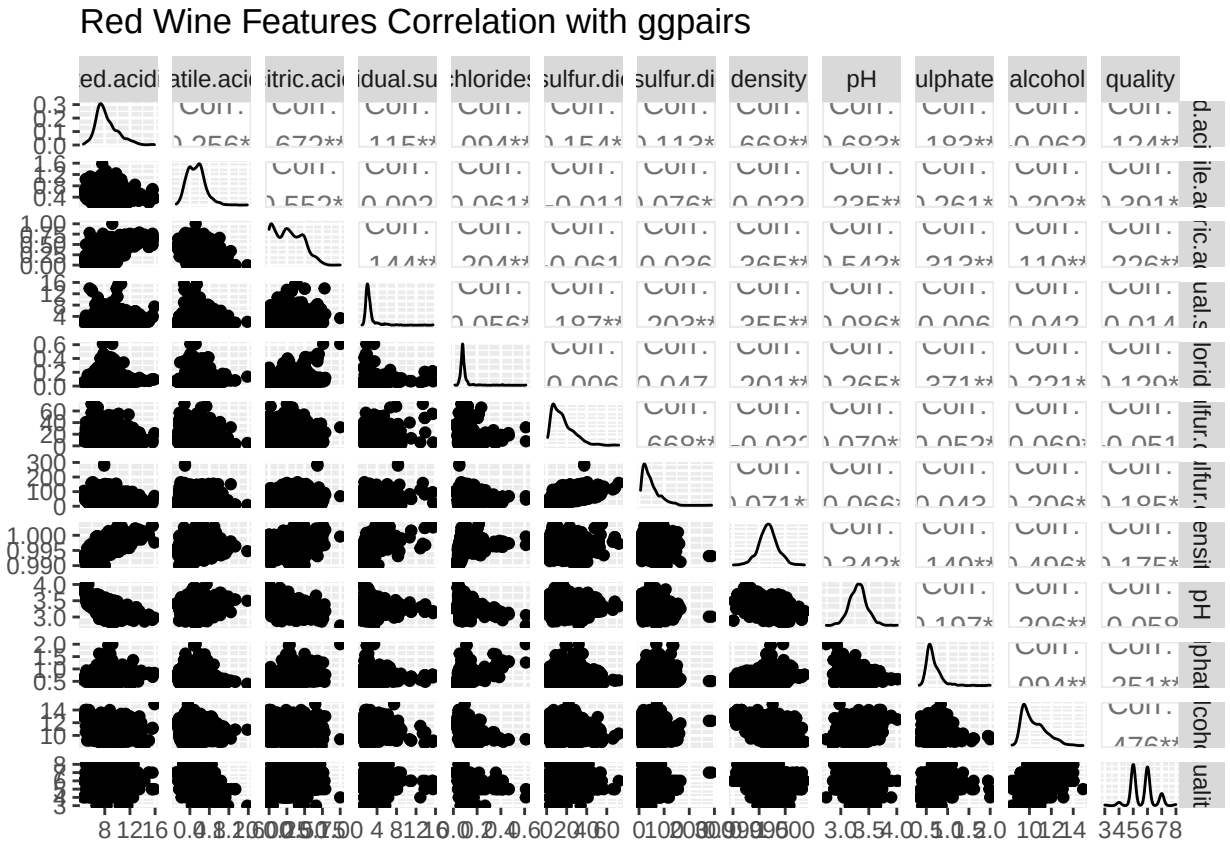
Before we look at the data, we believe that model satisfy assumtion (1)Normality (2)Independence (3)Homogeneity

4. Exploratory Data Analysis:

```
ggpairs(white_wine, title="White Wine Features Correlation with ggpairs")
```




```
ggpairs(red_wine, title="Red Wine Features Correlation with ggpairs")
```



- White wine: The strongest association with quality is alcohol(0.436). The lowest association with quality are citric acid(-0.009) and free sulfur dioxide(0.008).
- Red wine: The strongest association with quality is alcohol(0.476). The lowest association with quality is residual sugar(0.014).

Multiple linear regression

```
lm_white_wine <- lm(quality ~ ., data = white_wine)
summary(lm_white_wine)
```

```
##
## Call:
## lm(formula = quality ~ ., data = white_wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8348 -0.4934 -0.0379  0.4637  3.1143
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.502e+02  1.880e+01   7.987 1.71e-15 ***
## fixed.acidity         6.552e-02  2.087e-02   3.139  0.00171 **
```

```
## volatile.acidity     -1.863e+00  1.138e-01 -16.373  < 2e-16 ***
## citric.acid           2.209e-02  9.577e-02   0.231  0.81759
## residual.sugar        8.148e-02  7.527e-03  10.825  < 2e-16 ***
## chlorides            -2.473e-01  5.465e-01  -0.452  0.65097
## free.sulfur.dioxide   3.733e-03  8.441e-04   4.422 9.99e-06 ***
## total.sulfur.dioxide -2.857e-04  3.781e-04  -0.756  0.44979
## density              -1.503e+02  1.907e+01  -7.879 4.04e-15 ***
## pH                    6.863e-01  1.054e-01   6.513 8.10e-11 ***
## sulphates             6.315e-01  1.004e-01   6.291 3.44e-10 ***
## alcohol               1.935e-01  2.422e-02   7.988 1.70e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7514 on 4886 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2803
## F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```

```
reduce_lm_white_wine <- step(lm_white_wine, direction = "both")
```

```
## Start:  AIC=-2788.44
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##
##                       Df Sum of Sq    RSS     AIC
## - citric.acid          1     0.030 2758.4 -2790.4
## - chlorides            1     0.116 2758.4 -2790.2
## - total.sulfur.dioxide 1     0.323 2758.7 -2789.9
## <none>                             2758.3 -2788.4
## - fixed.acidity        1     5.562 2763.9 -2780.6
## - free.sulfur.dioxide  1    11.039 2769.4 -2770.9
## - sulphates            1    22.339 2780.7 -2750.9
## - pH                   1    23.948 2782.3 -2748.1
## - density              1    35.044 2793.4 -2728.6
## - alcohol              1    36.020 2794.3 -2726.9
## - residual.sugar       1    66.152 2824.5 -2674.4
## - volatile.acidity     1   151.345 2909.7 -2528.8
##
## Step:  AIC=-2790.39
## quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##
##                       Df Sum of Sq    RSS     AIC
## - chlorides            1     0.105 2758.5 -2792.2
## - total.sulfur.dioxide 1     0.315 2758.7 -2791.8
## <none>                             2758.4 -2790.4
## + citric.acid          1     0.030 2758.3 -2788.4
## - fixed.acidity        1     5.749 2764.1 -2782.2
## - free.sulfur.dioxide  1    11.096 2769.4 -2772.7
## - sulphates            1    22.444 2780.8 -2752.7
## - pH                   1    23.971 2782.3 -2750.0
## - density              1    35.066 2793.4 -2730.5
## - alcohol              1    36.540 2794.9 -2727.9
```

```
## - residual.sugar        1    66.160 2824.5 -2676.3
## - volatile.acidity      1   156.805 2915.2 -2521.6
##
## Step:  AIC=-2792.2
## quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + total.sulfur.dioxide + density + pH +
##     sulphates + alcohol
##
##                        Df Sum of Sq    RSS     AIC
## - total.sulfur.dioxide  1     0.320 2758.8 -2793.6
## <none>                              2758.5 -2792.2
## + chlorides             1     0.105 2758.4 -2790.4
## + citric.acid           1     0.019 2758.4 -2790.2
## - fixed.acidity         1     6.157 2764.6 -2783.3
## - free.sulfur.dioxide   1    11.036 2769.5 -2774.7
## - sulphates             1    22.570 2781.0 -2754.3
## - pH                    1    25.297 2783.8 -2749.5
## - alcohol               1    36.536 2795.0 -2729.8
## - density               1    36.823 2795.3 -2729.2
## - residual.sugar        1    70.134 2828.6 -2671.2
## - volatile.acidity      1   158.543 2917.0 -2520.5
##
## Step:  AIC=-2793.63
## quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + density + pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS     AIC
## <none>                              2758.8 -2793.6
## + total.sulfur.dioxide  1     0.320 2758.5 -2792.2
## + chlorides             1     0.110 2758.7 -2791.8
## + citric.acid           1     0.013 2758.8 -2791.7
## - fixed.acidity         1     6.270 2765.1 -2784.5
## - free.sulfur.dioxide   1    13.826 2772.6 -2771.2
## - sulphates             1    22.303 2781.1 -2756.2
## - pH                    1    25.460 2784.2 -2750.6
## - alcohol               1    36.300 2795.1 -2731.6
## - density               1    39.920 2798.7 -2725.3
## - residual.sugar        1    72.942 2831.7 -2667.8
## - volatile.acidity      1   167.753 2926.5 -2506.5
```

```
reduce_lm_white_wine$call
```

```
## lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + density + pH + sulphates + alcohol,
##     data = white_wine)
```

```
summary(reduce_lm_white_wine)
```

```
##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + density + pH + sulphates + alcohol,
##     data = white_wine)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8246 -0.4938 -0.0396  0.4660  3.1208
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.541e+02  1.810e+01   8.514  < 2e-16 ***
## fixed.acidity        6.810e-02  2.043e-02   3.333 0.000864 ***
## volatile.acidity    -1.888e+00  1.095e-01 -17.242  < 2e-16 ***
## residual.sugar       8.285e-02  7.287e-03  11.370  < 2e-16 ***
## free.sulfur.dioxide  3.349e-03  6.766e-04   4.950 7.67e-07 ***
## density             -1.543e+02  1.834e+01  -8.411  < 2e-16 ***
## pH                   6.942e-01  1.034e-01   6.717 2.07e-11 ***
## sulphates            6.285e-01  9.997e-02   6.287 3.52e-10 ***
## alcohol              1.932e-01  2.408e-02   8.021 1.31e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7512 on 4889 degrees of freedom
## Multiple R-squared:  0.2818, Adjusted R-squared:  0.2806
## F-statistic: 239.7 on 8 and 4889 DF,  p-value: < 2.2e-16
```

- The summary is up there. R-squared is 0.2818, not good. t-test says all variables are significant.

```
confint(reduce_lm_white_wine)
```

```
##                            2.5 %        97.5 %
## (Intercept)          1.186219e+02  1.895906e+02
## fixed.acidity        2.805009e-02  1.081578e-01
## volatile.acidity    -2.102826e+00 -1.673455e+00
## residual.sugar       6.856184e-02  9.713263e-02
## free.sulfur.dioxide  2.022619e-03  4.675411e-03
## density             -1.902537e+02 -1.183288e+02
## pH                   4.915984e-01  8.968285e-01
## sulphates            4.325178e-01  8.244984e-01
## alcohol              1.459485e-01  2.403771e-01
```

- The confidence interval for all variables doesn't include zero. That means all variables are credible.

```
lm_red_wine <- lm(quality ~ ., data = red_wine)
summary(lm_red_wine)
```

```
##
## Call:
## lm(formula = quality ~ ., data = red_wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68911 -0.36652 -0.04699  0.45202  2.02498
##
## Coefficients:
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.197e+01  2.119e+01   1.036   0.3002
## fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357
## volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
## citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150
## residual.sugar        1.633e-02  1.500e-02   1.089   0.2765
## chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
## free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *
## total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
## density              -1.788e+01  2.163e+01  -0.827   0.4086
## pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *
## sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
## alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

```
reduce_lm_red_wine <- step(lm_red_wine, direction = "both")
```

```
## Start:  AIC=-1375.49
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS     AIC
## - density               1     0.287 666.70 -1376.8
## - fixed.acidity         1     0.389 666.80 -1376.5
## - residual.sugar        1     0.498 666.91 -1376.3
## - citric.acid           1     0.646 667.06 -1375.9
## <none>                              666.41 -1375.5
## - free.sulfur.dioxide   1     1.694 668.10 -1373.4
## - pH                    1     1.957 668.37 -1372.8
## - chlorides             1     8.391 674.80 -1357.5
## - total.sulfur.dioxide  1     8.427 674.84 -1357.4
## - sulphates             1    26.971 693.38 -1314.0
## - volatile.acidity      1    33.620 700.03 -1298.8
## - alcohol               1    45.672 712.08 -1271.5
##
## Step:  AIC=-1376.8
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS     AIC
## - fixed.acidity         1     0.108 666.81 -1378.5
## - residual.sugar        1     0.231 666.93 -1378.2
## - citric.acid           1     0.654 667.35 -1377.2
## <none>                              666.70 -1376.8
## + density               1     0.287 666.41 -1375.5
## - free.sulfur.dioxide   1     1.829 668.53 -1374.4
## - pH                    1     4.325 671.02 -1368.5
```

```
## - total.sulfur.dioxide  1     8.728 675.43 -1358.0
## - chlorides             1     8.761 675.46 -1357.9
## - sulphates             1    27.287 693.98 -1314.7
## - volatile.acidity      1    35.000 701.70 -1297.0
## - alcohol               1   119.669 786.37 -1114.8
##
## Step:  AIC=-1378.54
## quality ~ volatile.acidity + citric.acid + residual.sugar + chlorides +
##     free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates +
##     alcohol
##
##                         Df Sum of Sq    RSS     AIC
## - residual.sugar         1     0.257 667.06 -1379.9
## - citric.acid            1     0.565 667.37 -1379.2
## <none>                               666.81 -1378.5
## + fixed.acidity          1     0.108 666.70 -1376.8
## + density                1     0.005 666.80 -1376.5
## - free.sulfur.dioxide    1     1.901 668.71 -1376.0
## - pH                     1     7.065 673.87 -1363.7
## - chlorides              1     9.940 676.75 -1356.9
## - total.sulfur.dioxide   1    10.031 676.84 -1356.7
## - sulphates              1    27.673 694.48 -1315.5
## - volatile.acidity       1    36.234 703.04 -1295.9
## - alcohol                1   120.633 787.44 -1114.7
##
## Step:  AIC=-1379.93
## quality ~ volatile.acidity + citric.acid + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol
##
##                         Df Sum of Sq    RSS     AIC
## - citric.acid            1     0.475 667.54 -1380.8
## <none>                               667.06 -1379.9
## + residual.sugar         1     0.257 666.81 -1378.5
## + fixed.acidity          1     0.133 666.93 -1378.2
## + density                1     0.028 667.03 -1378.0
## - free.sulfur.dioxide    1     2.064 669.13 -1377.0
## - pH                     1     7.138 674.20 -1364.9
## - total.sulfur.dioxide   1     9.828 676.89 -1358.5
## - chlorides              1     9.832 676.89 -1358.5
## - sulphates              1    27.446 694.51 -1317.5
## - volatile.acidity       1    35.977 703.04 -1297.9
## - alcohol                1   122.667 789.73 -1112.0
##
## Step:  AIC=-1380.79
## quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol
##
##                         Df Sum of Sq    RSS     AIC
## <none>                               667.54 -1380.8
## + citric.acid            1     0.475 667.06 -1379.9
## + residual.sugar         1     0.167 667.37 -1379.2
## + density                1     0.031 667.51 -1378.9
## + fixed.acidity          1     0.007 667.53 -1378.8
## - free.sulfur.dioxide    1     2.394 669.93 -1377.1
```

```
## - pH                   1     7.073 674.61 -1365.9
## - total.sulfur.dioxide  1    10.787 678.32 -1357.2
## - chlorides            1    10.809 678.35 -1357.1
## - sulphates            1    27.060 694.60 -1319.2
## - volatile.acidity     1    42.318 709.85 -1284.5
## - alcohol              1   124.483 792.02 -1109.4
```

```
reduce_lm_red_wine$call
```

```
## lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol, data = red_wine)
```

```
summary(reduce_lm_red_wine)
```

```
##
## Call:
## lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol, data = red_wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68918 -0.36757 -0.04653  0.46081  2.02954
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.4300987  0.4029168  10.995  < 2e-16 ***
## volatile.acidity     -1.0127527  0.1008429 -10.043  < 2e-16 ***
## chlorides            -2.0178138  0.3975417  -5.076 4.31e-07 ***
## free.sulfur.dioxide   0.0050774  0.0021255   2.389    0.017 *
## total.sulfur.dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
## pH                   -0.4826614  0.1175581  -4.106 4.23e-05 ***
## sulphates             0.8826651  0.1099084   8.031 1.86e-15 ***
## alcohol               0.2893028  0.0167958  17.225  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6477 on 1591 degrees of freedom
## Multiple R-squared:  0.3595, Adjusted R-squared:  0.3567
## F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16
```

- The summary is up there. R-squared is 0.3595, not good. t-test says all variables are significant.

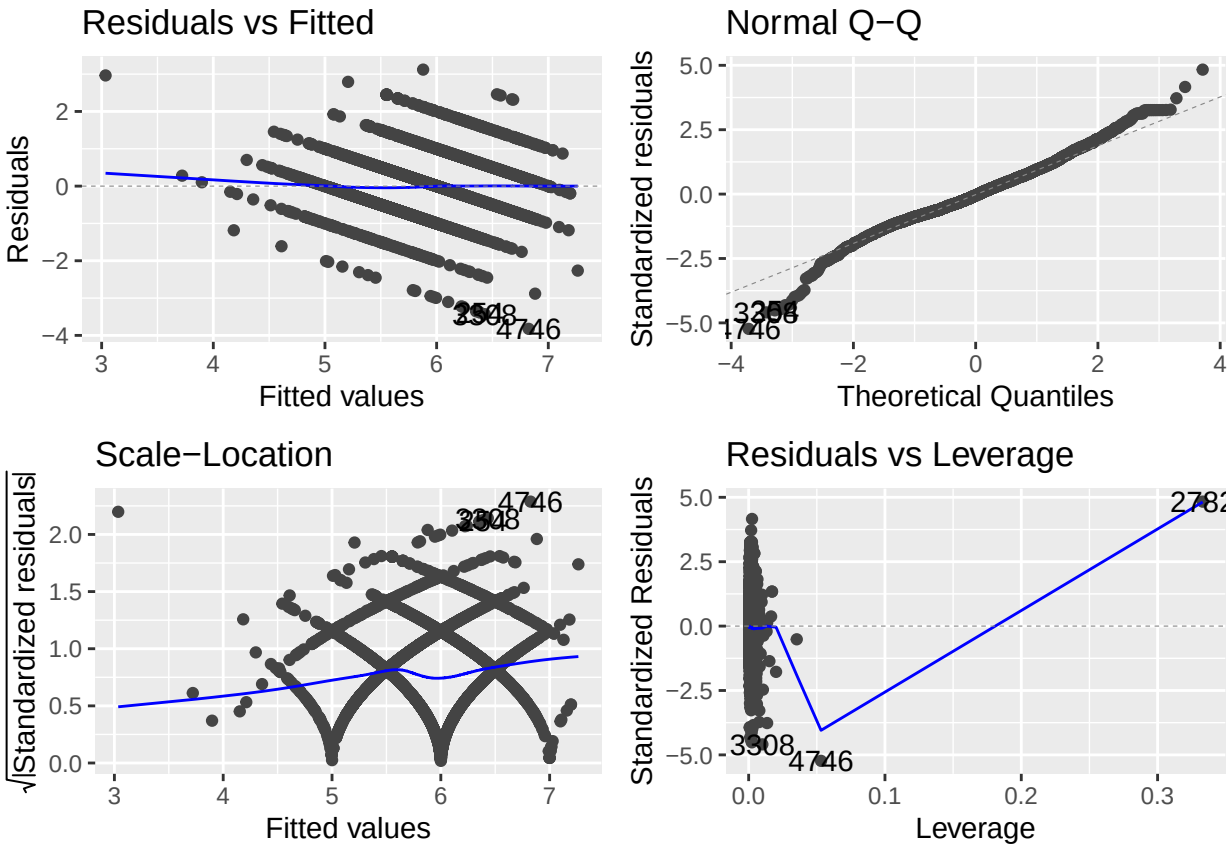
```
confint(reduce_lm_red_wine)
```

```
##                             2.5 %       97.5 %
## (Intercept)           3.6397950327  5.220402364
## volatile.acidity     -1.2105517115 -0.814953689
## chlorides            -2.7975745007 -1.238053133
## free.sulfur.dioxide   0.0009082351  0.009246504
## total.sulfur.dioxide -0.0048293390 -0.002135152
## pH                   -0.7132464077 -0.252076480
## sulphates             0.6670845326  1.098245733
## alcohol               0.2563585503  0.322246956
```

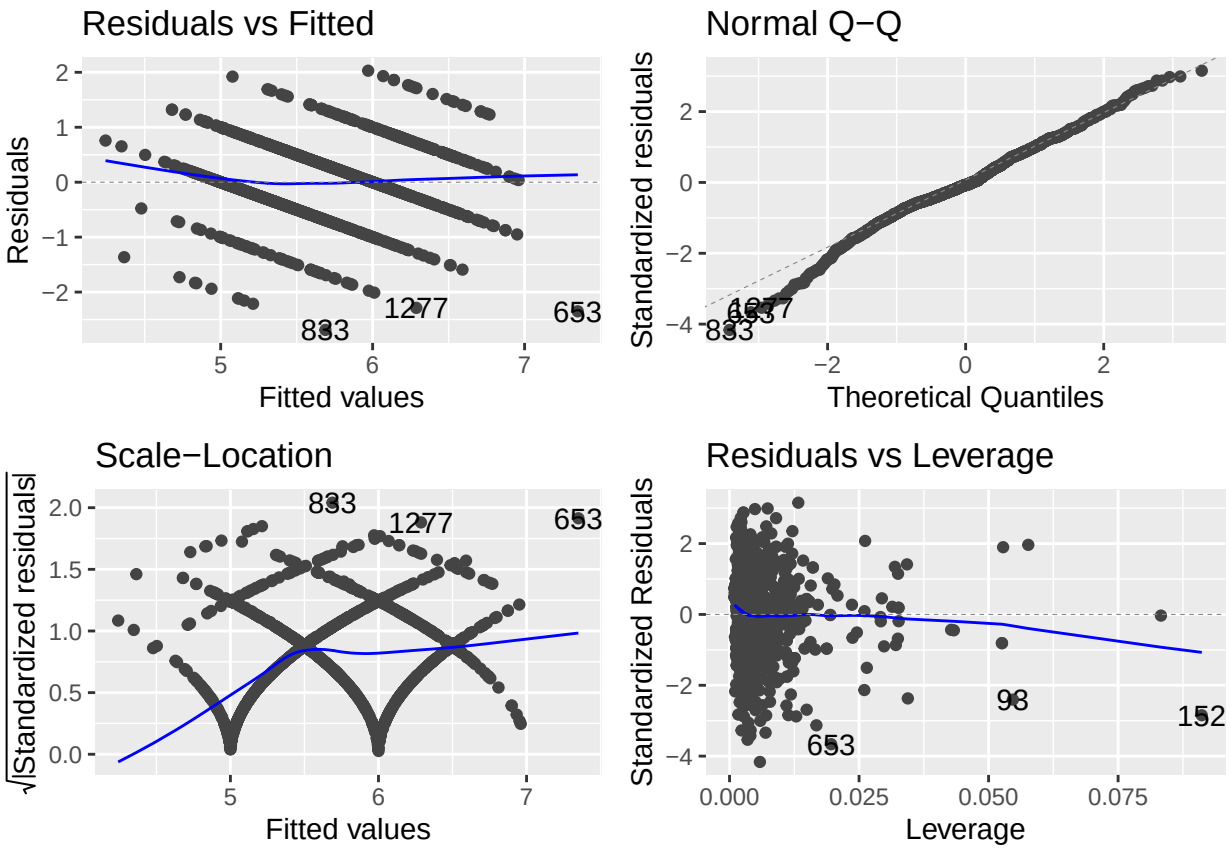
- The confidence interval for all variables doesn't include zero. That means all variables are credible.
- Conclusion: We figured out from the result, alcohol is the most powerful variables to wine quality. But it's risky to interpret the result as 'more alcohol, more quality'. We all know that's not true. It's appropriate to interpret it as it'd be better to add more alcohol within normal range to might improve wine quality. On the other hand, we can take care less pH and free sulfur dioxide or something.

Model testing

```
# Assumtion
autoplot(reduce_lm_white_wine)
```




```
autoplot(reduce_lm_red_wine)
```



Because the dependent variable is discrete the residual vs fitted plot seems ot be quite different from the one from normal regression.

```
# Residual normality test
shapiro.test(reduce_lm_white_wine$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  reduce_lm_white_wine$residuals
## W = 0.9894, p-value < 2.2e-16
```

```
shapiro.test(reduce_lm_red_wine$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  reduce_lm_red_wine$residuals
## W = 0.99137, p-value = 4.321e-08
```

```
# Residual independence test
durbinWatsonTest(reduce_lm_white_wine)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1        0.189286      1.621314       0
##  Alternative hypothesis: rho != 0
```

```r
durbinWatsonTest(reduce_lm_red_wine)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1250136      1.749967       0
##  Alternative hypothesis: rho != 0
```

```r
# Residual variance homogeneity test
ncvTest(reduce_lm_white_wine)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 16.73906, Df = 1, p = 4.2889e-05
```

```r
ncvTest(reduce_lm_red_wine)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 20.10007, Df = 1, p = 7.3494e-06
```

The p-value is very small, so the null hypothesis is rejected, indicating that y does not conform to the normal distribution, does not conform to the assumption of independence, and does not conform to the assumption of homogeneity.

5. Data-drivin Hypotheses:

$H_0 = \beta_1 = \beta_2 = ... = \beta_{11} = 0$

$H_1 = \beta_i \neq 0$ for some $i = 1, 2, ...11$

6. Discussion:

We find out there is not only one input variable strongly related with the quality. We figured out from the result, that alcohol is the most powerful variable in wine quality. But it's risky to interpret the result as "more alcohol, more quality". It's appropriate to interpret it as it'd be better to add more alcohol within the normal range to might improve wine quality.

The limitation of this research is the quality mostly distributed in 4 to 8. My recommendation for this two data sets is the quality can be distributed in 0 to 10. Thus, we can clearly know whether this model is a good one.

7. References:

Cortez, P. (2010, July). Data mining with neural networks and support vector machines using the R/rminer tool. In Industrial conference on data mining (pp. 572-583). Springer, Berlin, Heidelberg.

Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences, 225, 1-17.

Lee, K., Lam, M., Pedarsani, R., Papailiopoulos, D., & Ramchandran, K. (2017). Speeding up distributed machine learning using codes. IEEE Transactions on Information Theory, 64(3), 1514-1529.

Mei, S., & Zhu, X. (2015, February). Using machine teaching to identify optimal training-set attacks on machine learners. In Twenty-Ninth AAAI Conference on Artificial Intelligence.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

8. Appendix: