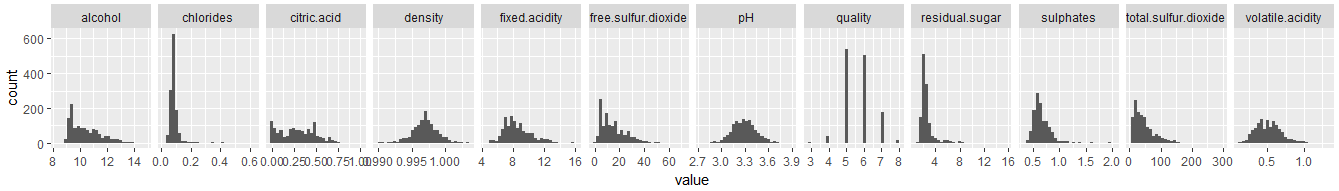
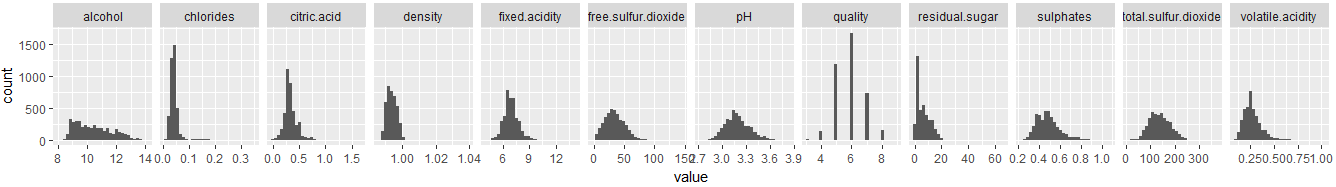
**Assignment 3**

**Regression & Model Evaluation**

Red wine (training data)



White wine (training data)



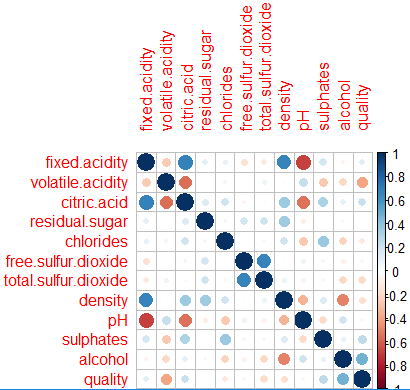
The distribution of chlorides and residual sugar are very skewed to the right.

4. Adjusted R2 for various models:

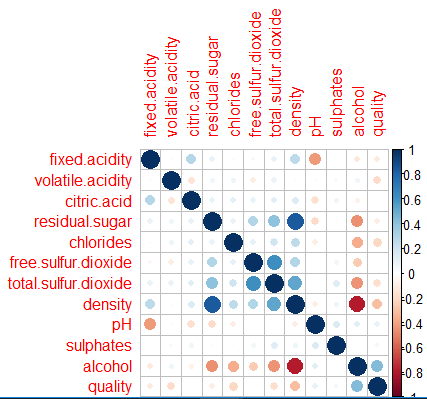
|  |  |  |
| --- | --- | --- |
| Model | Adjusted R2 | |
| Red Wine | White Wine |
| Fit | 0.3642 | 0.2973 |
| Step | 0.3651 | 0.2978 |
| Fit (without outlier) | 0.4068 | 0.2678 |
| Step (without outlier) | 0.4082 | 0.2678 |

5.

Red Wine:



White Wine:



Yes, multicollinearity exists between those variables having high correlation in above table.

We could improve by excluding one of them, and rebuild the model.

6.  
Model for red wine:

# Call:

# lm(formula = quality ~ volatile.acidity + chlorides + total.sulfur.dioxide +

# pH + sulphates + alcohol, data = redwine\_train)

#

# Residuals:

# Min 1Q Median 3Q Max

# -2.60944 -0.37929 -0.07128 0.46400 1.95003

#

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) 4.0181035 0.4416996 9.097 < 2e-16 \*\*\*

# volatile.acidity -1.0081450 0.1119893 -9.002 < 2e-16 \*\*\*

# chlorides -1.9140239 0.4336091 -4.414 1.10e-05 \*\*\*

# total.sulfur.dioxide -0.0030719 0.0005521 -5.564 3.21e-08 \*\*\*

# pH -0.3450333 0.1271282 -2.714 0.00674 \*\*

# sulphates 0.7827645 0.1199544 6.526 9.76e-11 \*\*\*

# alcohol 0.2982770 0.0180298 16.544 < 2e-16 \*\*\*

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#

# Residual standard error: 0.6477 on 1272 degrees of freedom

# Multiple R-squared: 0.3673, Adjusted R-squared: 0.3644

# F-statistic: 123.1 on 6 and 1272 DF, p-value: < 2.2e-16

Model for white wine:

# Call:

# lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +

# free.sulfur.dioxide + pH + sulphates + alcohol, data = whitewine\_train)

#

# Residuals:

# Min 1Q Median 3Q Max

# -3.3729 -0.5100 -0.0178 0.4708 3.2436

#

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# (Intercept) 1.4659409 0.3857029 3.801 0.000146 \*\*\*

# fixed.acidity -0.0763247 0.0163392 -4.671 3.09e-06 \*\*\*

# volatile.acidity -1.9206521 0.1244258 -15.436 < 2e-16 \*\*\*

# residual.sugar 0.0234435 0.0028257 8.297 < 2e-16 \*\*\*

# free.sulfur.dioxide 0.0047199 0.0007825 6.032 1.77e-09 \*\*\*

# pH 0.2211367 0.0922498 2.397 0.016570 \*

# sulphates 0.4907318 0.1099860 4.462 8.36e-06 \*\*\*

# alcohol 0.4028446 0.0114480 35.189 < 2e-16 \*\*\*

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#

# Residual standard error: 0.7691 on 3910 degrees of freedom

# Multiple R-squared: 0.2888, Adjusted R-squared: 0.2875

# F-statistic: 226.8 on 7 and 3910 DF, p-value: < 2.2e-16

Similarity:

* The influence of volatile.acidity are quite high in both model, it is negatively correlated (The higher the volatile.acidity, the lower the quality rating)
* pH, sulphates, and alcohol also factors that affect the quality

Difference:

* Chlorides and total.sulfur.dioxide are considered in red wine model, while fixed.acidity, residual.sugar, and free.sulfur.dioxide are considered in white wine model