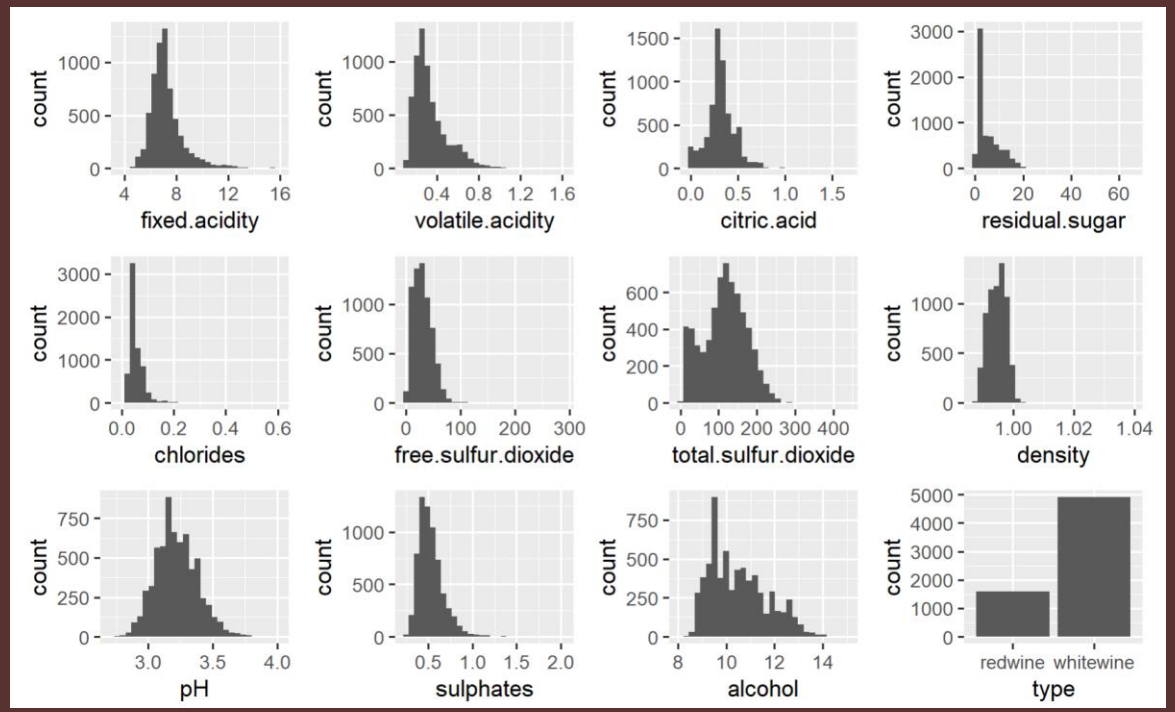
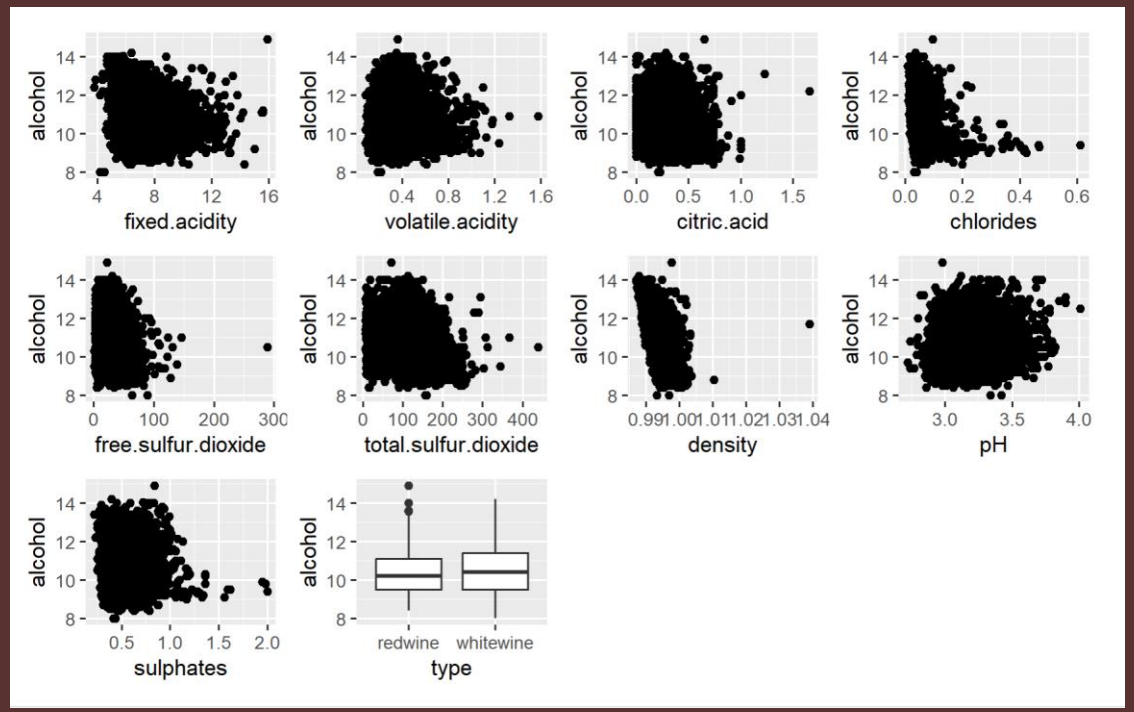
Investigating the relationship between alcohol content in wine and its physicochemical features

Lavanya Kudli – EDA, Model A (AIC, BIC), Model B, remedial measures

Anav Vora – Model A (Leaps, PCA, Lasso, Ridge), and diagnostics

Exploratory Data Analysis





Model A – training (greedy algorithm with AIC criterion)

```
Start:  AIC=-6871.24
alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
    chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
    density + pH + sulphates + type

                      Df Sum of Sq    RSS     AIC
- total.sulfur.dioxide  1       0.5 1379.4 -6871.3
<none>                              1378.9 -6871.2
- chlorides             1       4.4 1383.3 -6856.7
- free.sulfur.dioxide   1       6.0 1384.9 -6850.7
- citric.acid           1      22.9 1401.9 -6787.5
- volatile.acidity      1      28.4 1407.3 -6767.4
- sulphates             1      98.4 1477.4 -6514.9
- type                  1     234.9 1613.9 -6055.6
- pH                    1     541.1 1920.1 -5152.7
- fixed.acidity         1     838.2 2217.1 -4405.3
- residual.sugar        1    1566.9 2945.8 -2928.4
- density               1    3778.0 5157.0   -18.2
```

```
Step:  AIC=-6871.33
alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
    chlorides + free.sulfur.dioxide + density + pH + sulphates +
    type

                      Df Sum of Sq    RSS     AIC
<none>                              1379.4 -6871.3
+ total.sulfur.dioxide  1       0.5 1378.9 -6871.2
- chlorides             1       4.3 1383.7 -6857.2
- free.sulfur.dioxide   1      12.8 1392.2 -6825.3
- citric.acid           1      22.5 1401.9 -6789.3
- volatile.acidity      1      27.9 1407.3 -6769.3
- sulphates             1      97.9 1477.4 -6516.9
- type                  1     368.0 1747.4 -5644.5
- pH                    1     544.3 1923.7 -5144.9
- fixed.acidity         1     860.4 2239.9 -4354.1
- residual.sugar        1    1604.9 2984.3 -2862.8
- density               1    4257.2 5636.7   442.1
```

Testing RMSE: 0.4543

Model A – training (greedy algorithm with BIC criterion)

```
Start:  AIC=-6789.89
alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
    chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
    density + pH + sulphates + type

                      Df Sum of Sq    RSS     AIC
- total.sulfur.dioxide  1       0.5 1379.4 -6796.8
<none>                              1378.9 -6789.9
- chlorides             1       4.4 1383.3 -6782.1
- free.sulfur.dioxide   1       6.0 1384.9 -6776.2
- citric.acid           1      22.9 1401.9 -6713.0
- volatile.acidity      1      28.4 1407.3 -6692.8
- sulphates             1      98.4 1477.4 -6440.4
- type                  1     234.9 1613.9 -5981.0
- pH                    1     541.1 1920.1 -5078.2
- fixed.acidity         1     838.2 2217.1 -4330.7
- residual.sugar        1    1566.9 2945.8 -2853.8
- density               1    3778.0 5157.0    56.4
```

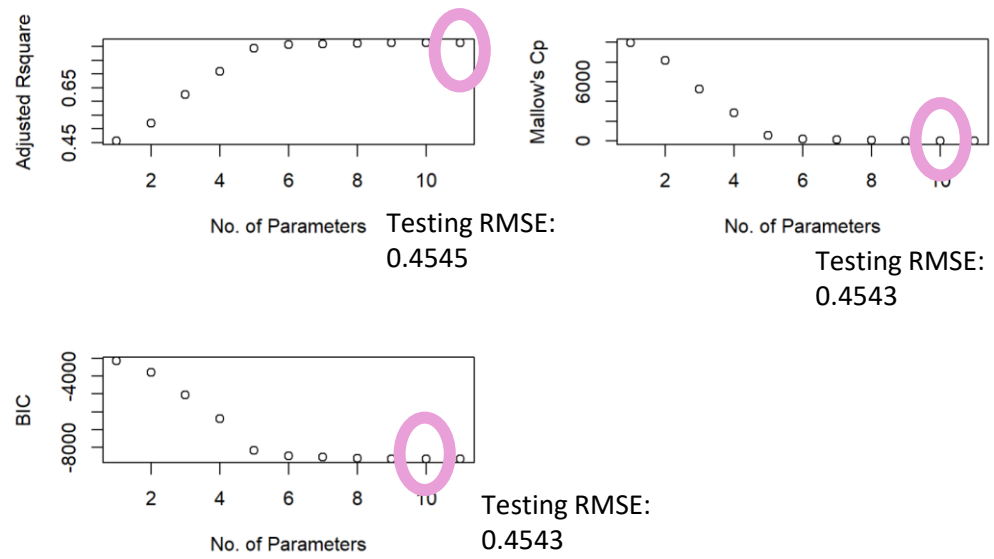
```
Step:  AIC=-6796.76
alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
    chlorides + free.sulfur.dioxide + density + pH + sulphates +
    type

                      Df Sum of Sq    RSS     AIC
<none>                              1379.4 -6796.8
+ total.sulfur.dioxide  1       0.5 1378.9 -6789.9
- chlorides             1       4.3 1383.7 -6789.4
- free.sulfur.dioxide   1      12.8 1392.2 -6757.5
- citric.acid           1      22.5 1401.9 -6721.5
- volatile.acidity      1      27.9 1407.3 -6701.5
- sulphates             1      97.9 1477.4 -6449.1
- type                  1     368.0 1747.4 -5576.7
- pH                    1     544.3 1923.7 -5077.1
- fixed.acidity         1     860.4 2239.9 -4286.3
- residual.sugar        1    1604.9 2984.3 -2795.0
- density               1    4257.2 5636.7   509.9
```

Testing RMSE: 0.4543

Model A – leap and bound & other shrinkage methods

Leap and bounds

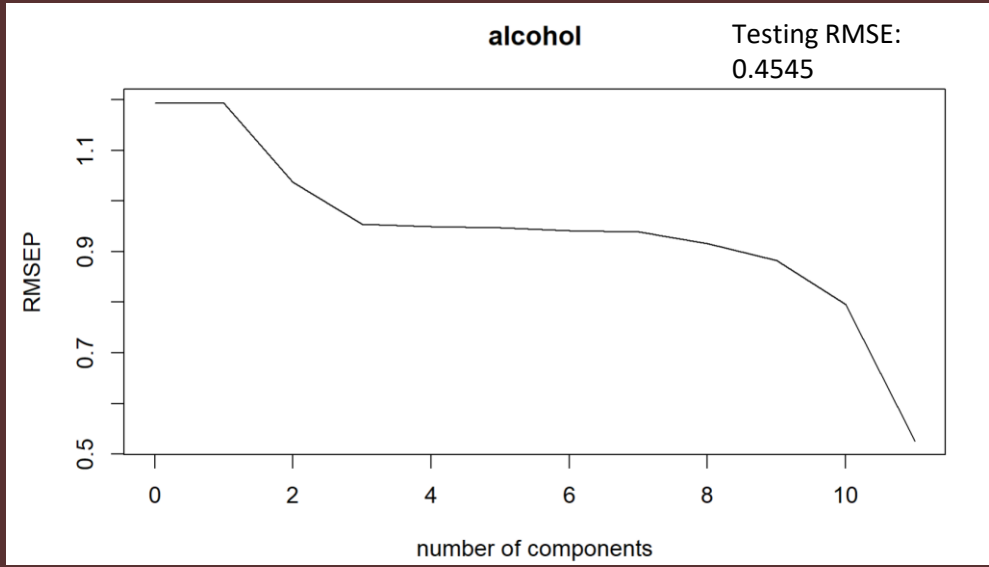


Testing RMSE: 0.4545

Testing RMSE: 0.4543

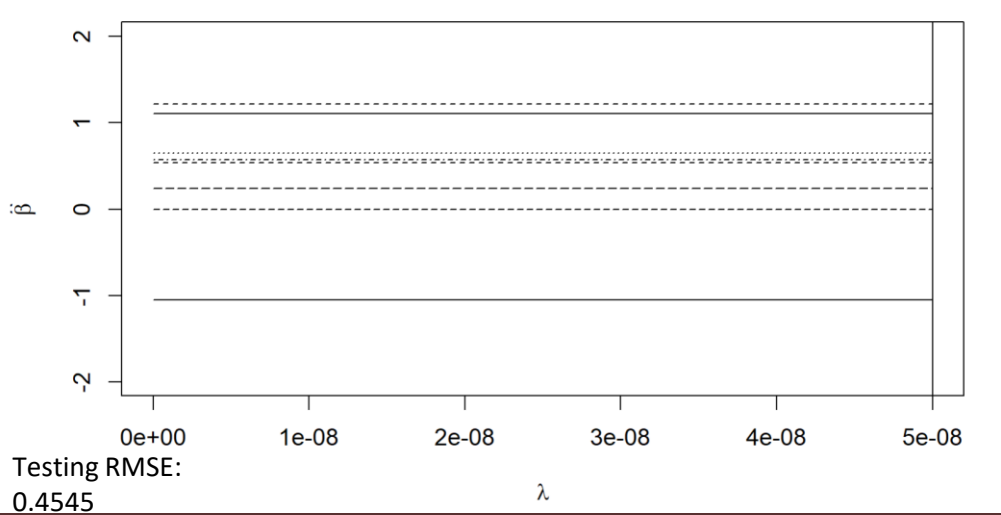
Testing RMSE: 0.4543

PCR



Testing RMSE: 0.4545

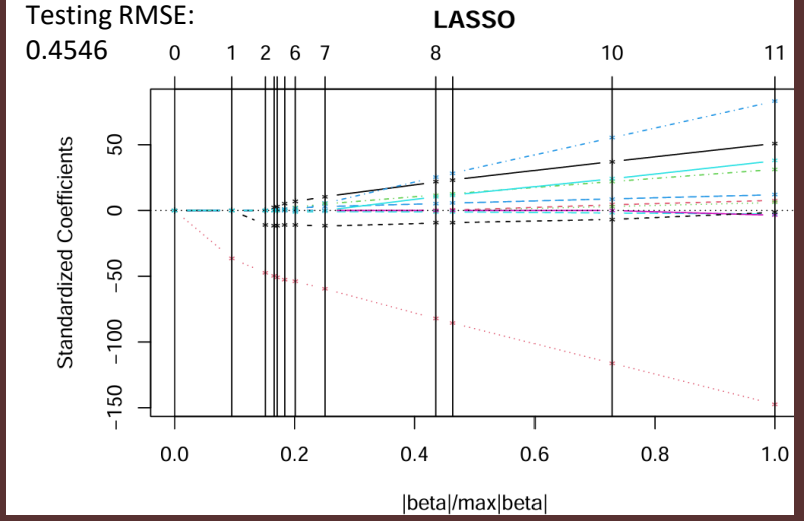
Ridge



Testing RMSE: 0.4545

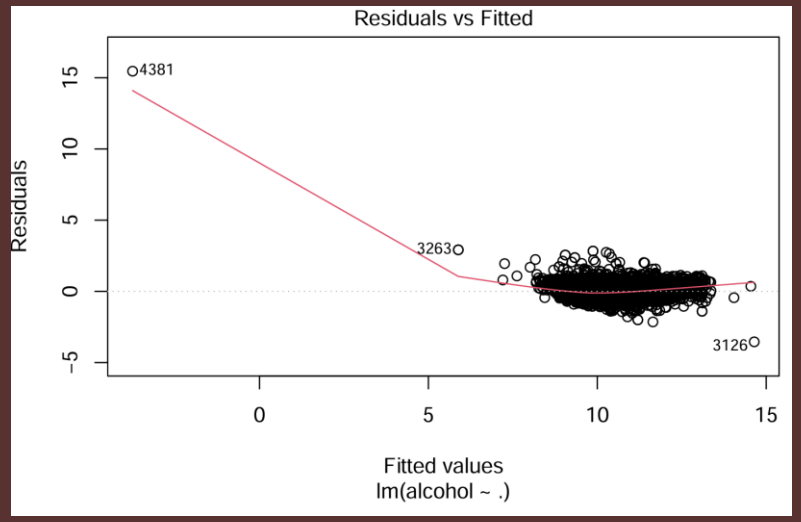
Lasso

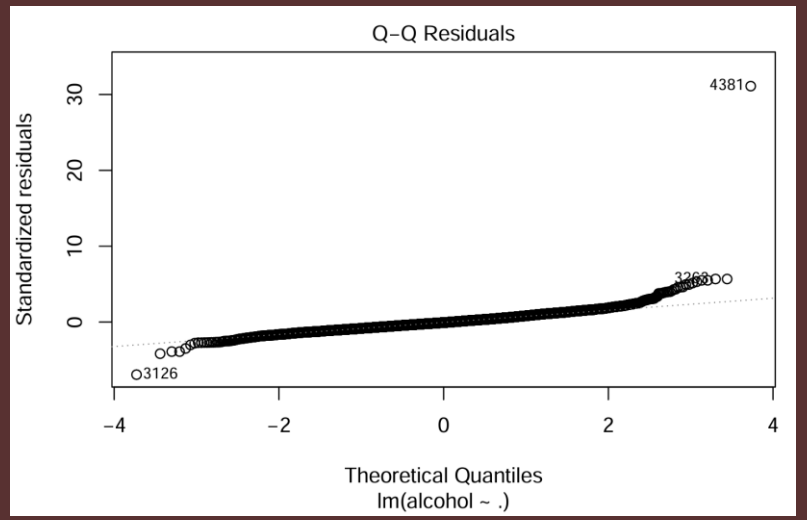
Testing RMSE: 0.4546

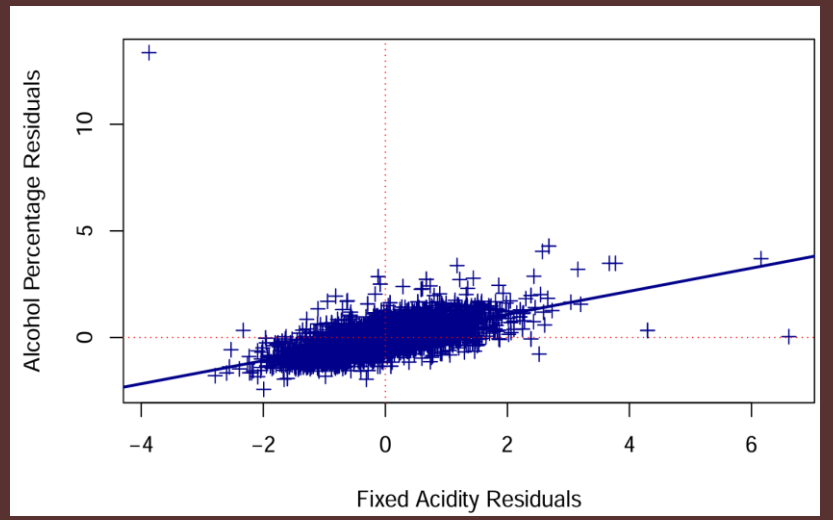


Model A – diagnostics

Leverage	Outliers	Influential points	Variance assumption	Normality	Linearity	Correlation
• 298 high leverage points • 21 bad high leverage points	• Bonferroni CV was -4.43, • 12 studentized residuals have value > 4.43	• 1 pt with Cook's distance of 6.78	• BP test : p-value < 0.05 (null rejected), variance not constant	• KS test: p-value < 0.05; null rejected; normality failed	• Added variable plots show uniformly scattered points	• No strong correlation







Model B – model selection and remediation

```
Full_Model = lm(alcohol~.,data = Wine_Data)
summary(Full_Model)
```

```
##
## Call:
## lm(formula = alcohol ~ ., data = Wine_Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5559 -0.2892 -0.0361  0.2549 15.6752
##
## Coefficients:
##                         Estimate Std. Error  t value Pr(>|t|)
## (Intercept)           6.745e+02  4.874e+00  138.392  < 2e-16 ***
## fixed.acidity         5.432e-01  8.474e-03   64.109  < 2e-16 ***
## volatile.acidity      6.502e-01  5.531e-02   11.756  < 2e-16 ***
## citric.acid           5.320e-01  5.437e-02    9.784  < 2e-16 ***
## residual.sugar        2.404e-01  2.776e-03   86.606  < 2e-16 ***
## chlorides            -1.013e+00  2.294e-01   -4.415 1.03e-05 ***
## free.sulfur.dioxide  -2.954e-03  5.252e-04   -5.625 1.93e-08 ***
## total.sulfur.dioxide -2.499e-04  2.224e-04   -1.124    0.261
## density              -6.827e+02  5.014e+00 -136.159  < 2e-16 ***
## pH                    2.721e+00  5.226e-02   52.058  < 2e-16 ***
## sulphates             1.095e+00  5.059e-02   21.645  < 2e-16 ***
## type                  1.210e+00  3.598e-02   33.645  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5037 on 6485 degrees of freedom
## Multiple R-squared:  0.822,  Adjusted R-squared:  0.8217
## F-statistic:  2722 on 11 and 6485 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + density + pH + sulphates +
##     type
## Model 2: alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + type
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   6486 1645.5
## 2   6485 1645.2  1   0.32052 1.2634 0.2611
```

```
## Analysis of Variance Table
##
## Model 1: alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     free.sulfur.dioxide + density + pH + sulphates + type
## Model 2: alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + density + pH + sulphates +
##     type
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   6487 1650.4
## 2   6486 1645.5  1    4.8719 19.203 1.194e-05 ***
```

Additionally, using the Box-Cox transformation to fix issues of non-constant variance and normality did not work.

Conclusion, both model A and B pick the model without total sulfur dioxide.