R Notebook

Code ▾

This is an R Markdown (http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

— Correlations Between Variables —

To see why we get very low accuracies accross the board for this Wines dataset, let’s look at the correlation relationships between the predictors. Let’s also take a look at the class label distributions of the dataset (how balanced is it?).

Red Wines Dataset:

Hide

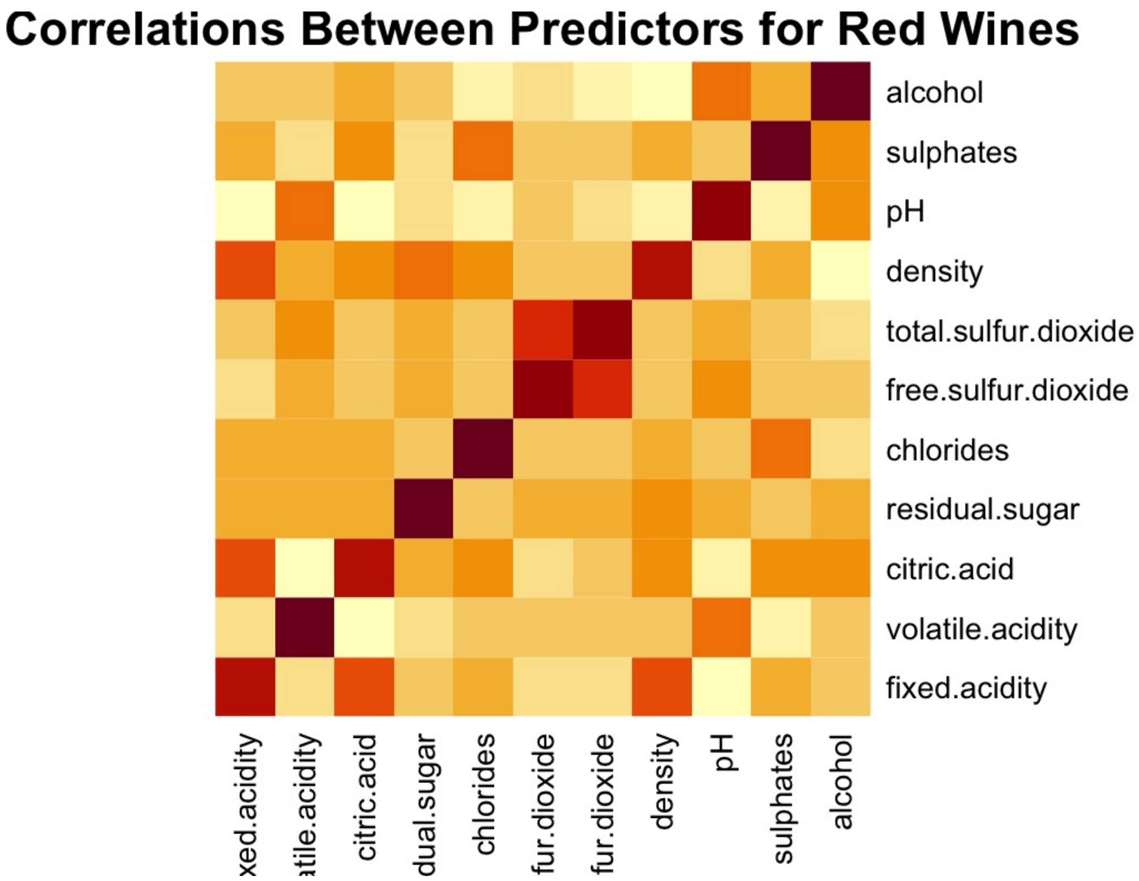
```
red <- read.csv('winequality-red.csv', header = TRUE, sep=";")
red <- na.omit(red)
red.quality <- red$quality
red[,-12] <- scale(red[,-12])

# Correlation matrix between variables
corr <- cor(red[,-12])
corr
```

```
                     fixed.acidity volatile.acidity citric.acid residual.sugar    chlorides free.sulfur.dioxide t
otal.sulfur.dioxide     density          pH
fixed.acidity           1.00000000     -0.256130895  0.67170343    0.114776724  0.093705186        -0.153794193
-0.11318144  0.66804729 -0.68297819
volatile.acidity       -0.25613089      1.000000000 -0.55249568    0.001917882  0.061297772        -0.010503827
0.07647000  0.02202623  0.23493729
citric.acid             0.67170343     -0.552495685  1.00000000    0.143577162  0.203822914        -0.060978129
0.03553302  0.36494718 -0.54190414
residual.sugar          0.11477672      0.001917882  0.14357716    1.000000000  0.055609535         0.187048995
0.20302788  0.35528337 -0.08565242
chlorides               0.09370519      0.061297772  0.20382291    0.055609535  1.000000000         0.005562147
0.04740047  0.20063233 -0.26502613
free.sulfur.dioxide    -0.15379419     -0.010503827 -0.06097813    0.187048995  0.005562147         1.000000000
0.66766645 -0.02194583  0.07037750
total.sulfur.dioxide   -0.11318144      0.076470005  0.03553302    0.203027882  0.047400468         0.667666450
1.00000000  0.07126948 -0.06649456
density                 0.66804729      0.022026232  0.36494718    0.355283371  0.200632327        -0.021945831
0.07126948  1.00000000 -0.34169933
pH                     -0.68297819      0.234937294 -0.54190414   -0.085652422 -0.265026131         0.070377499
-0.06649456 -0.34169933  1.00000000
sulphates               0.18300566     -0.260986685  0.31277004    0.005527121  0.371260481         0.051657572
0.04294684  0.14850641 -0.19664760
alcohol                -0.06166827     -0.202288027  0.10990325    0.042075437 -0.221140545        -0.069408354
-0.20565394 -0.49617977  0.20563251
                        sulphates     alcohol
fixed.acidity         0.183005664 -0.06166827
volatile.acidity     -0.260986685 -0.20228803
citric.acid           0.312770044  0.10990325
residual.sugar        0.005527121  0.04207544
chlorides             0.371260481 -0.22114054
free.sulfur.dioxide   0.051657572 -0.06940835
total.sulfur.dioxide  0.042946836 -0.20565394
density               0.148506412 -0.49617977
pH                   -0.196647602  0.20563251
sulphates             1.000000000  0.09359475
alcohol               0.093594750  1.00000000
```

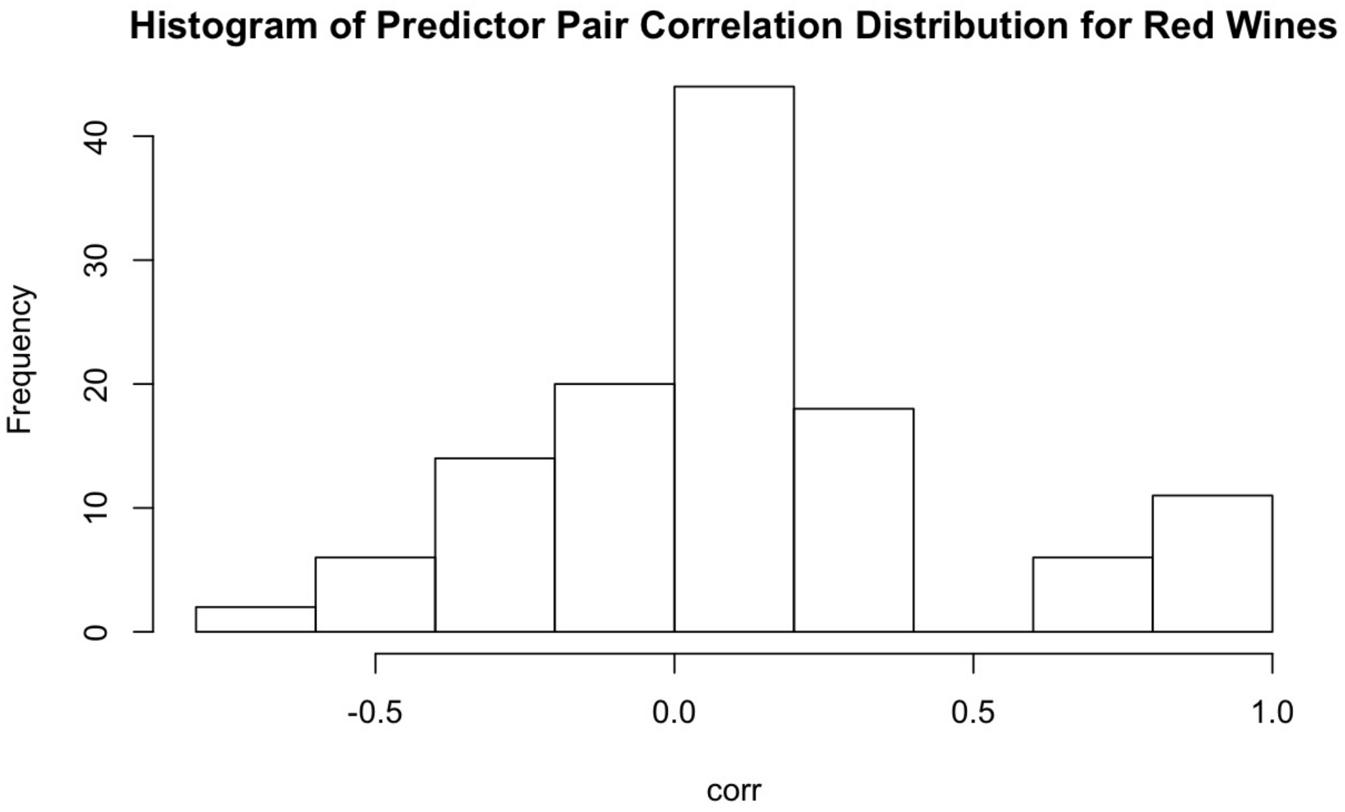
Hide

```
# Plot Heatmap
heatmap(corr, main="Correlations Between Predictors for Red Wines", Colv = NA, Rowv = NA, scale="column")
```


Correlations Between Predictors for Red Wines

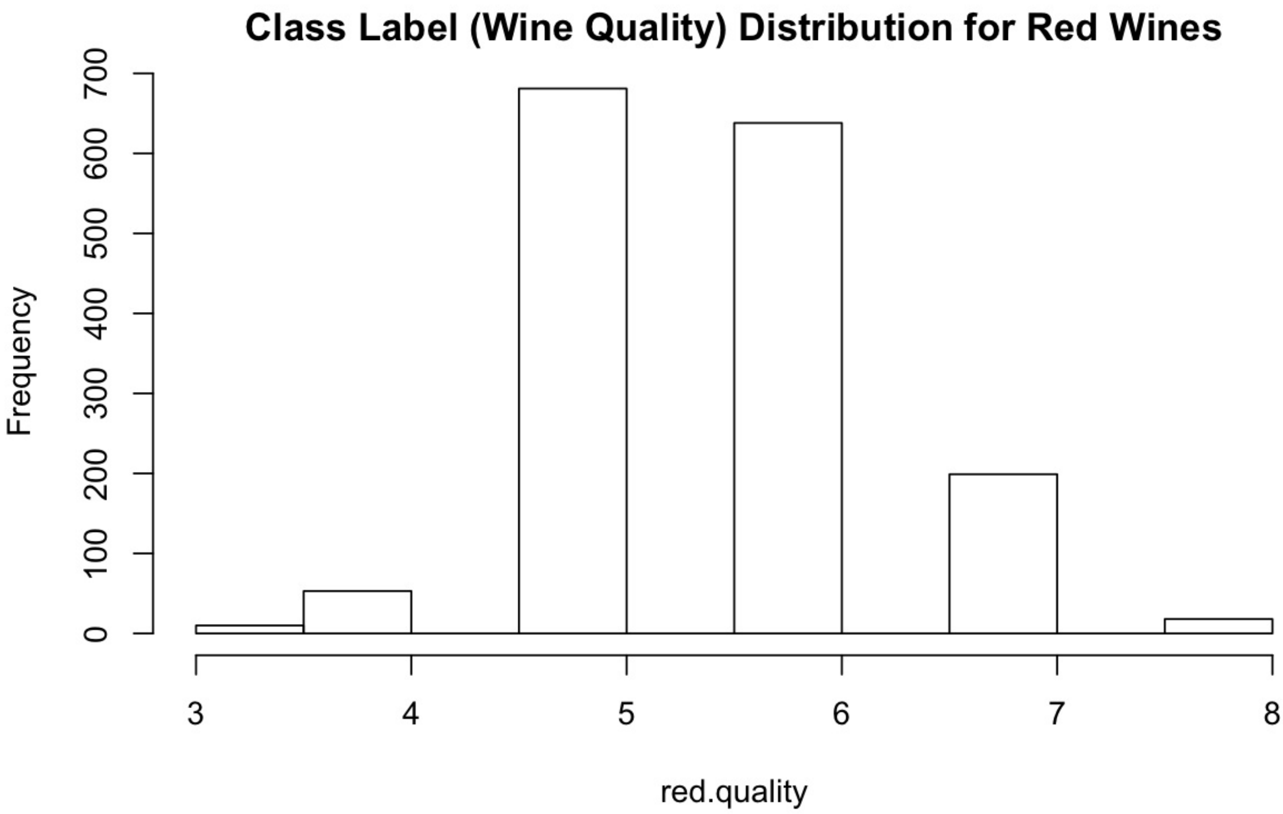
Hide

```
# Plot histogram
hist(corr, main='Histogram of Predictor Pair Correlation Distribution for Red Wines')
```


Histogram of Predictor Pair Correlation Distribution for Red Wines

Hide

```
# Plot histogram between class labels
class_dist <- hist(red.quality, main='Class Label (Wine Quality) Distribution for Red Wines')
```



Hide

```
class_dist$counts
```

```
 [1]  10  53   0 681   0 638   0 199   0  18
```

White Wines Dataset:

Hide

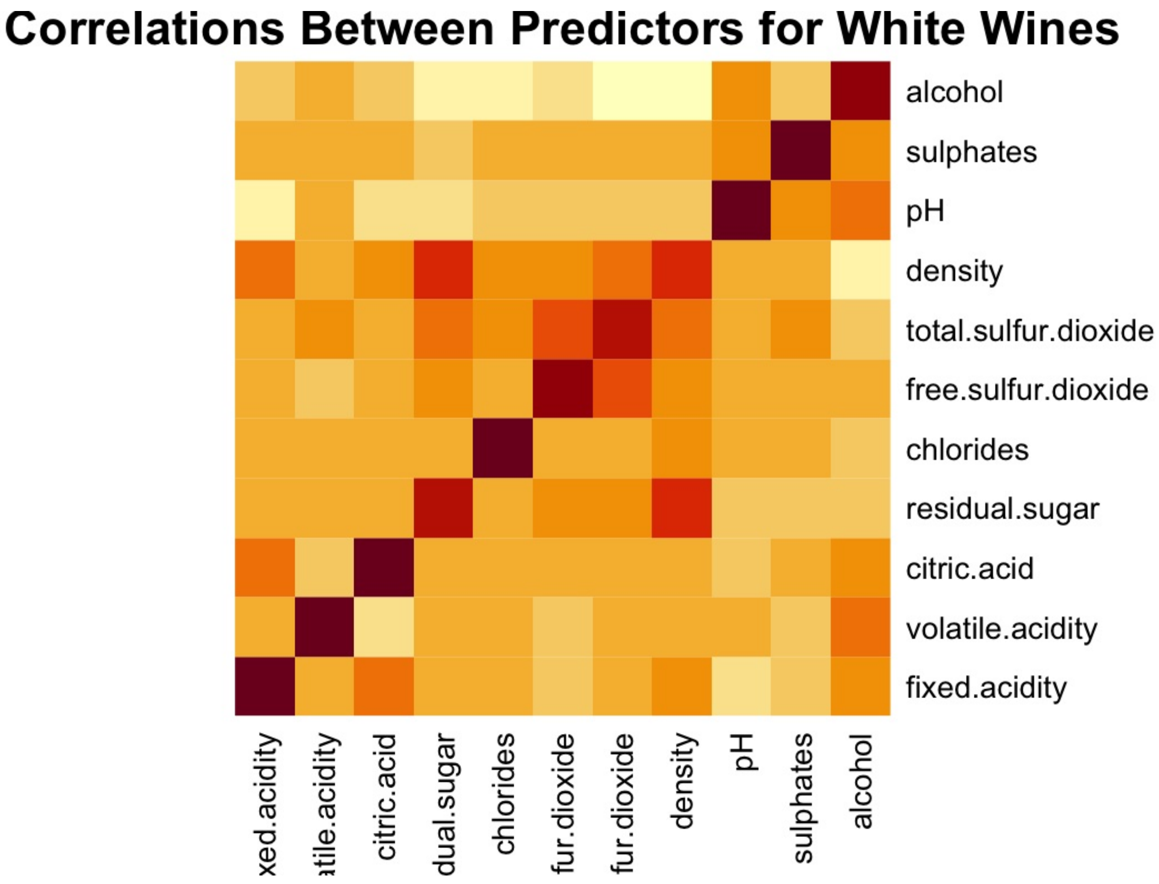
```
white <- read.csv('winequality-white.csv', header = TRUE, sep=";")
white <- na.omit(white)
white.quality <- white$quality
white[,-12] <- scale(white[,-12])

# Correlation matrix between variables
corr <- cor(white[,-12])
corr
```

```
                     fixed.acidity volatile.acidity citric.acid residual.sugar   chlorides free.sulfur.dioxide to
tal.sulfur.dioxide     density
fixed.acidity          1.00000000      -0.02269729  0.28918070     0.08902070  0.02308564       -0.0493958591
0.091069756  0.26533101
volatile.acidity      -0.02269729       1.00000000 -0.14947181     0.06428606  0.07051157       -0.0970119393
0.089260504  0.02711385
citric.acid            0.28918070      -0.14947181  1.00000000     0.09421162  0.11436445        0.0940772210
0.121130798  0.14950257
residual.sugar         0.08902070       0.06428606  0.09421162     1.00000000  0.08868454        0.2990983537
0.401439311  0.83896645
chlorides              0.02308564       0.07051157  0.11436445     0.08868454  1.00000000        0.1013923521
0.198910300  0.25721132
free.sulfur.dioxide   -0.04939586      -0.09701194  0.09407722     0.29909835  0.10139235        1.0000000000
0.615500965  0.29421041
total.sulfur.dioxide   0.09106976       0.08926050  0.12113080     0.40143931  0.19891030        0.6155009650
1.000000000  0.52988132
density                0.26533101       0.02711385  0.14950257     0.83896645  0.25721132        0.2942104109
0.529881324  1.00000000
pH                    -0.42585829      -0.03191537 -0.16374821    -0.19413345 -0.09043946       -0.0006177961
0.002320972 -0.09359149
sulphates             -0.01714299      -0.03572815  0.06233094    -0.02666437  0.01676288        0.0592172458
0.134562367  0.07449315
alcohol               -0.12088112       0.06771794 -0.07572873    -0.45063122 -0.36018871       -0.2501039415
-0.448892102 -0.78013762
                                pH   sulphates     alcohol
fixed.acidity        -0.4258582910 -0.01714299 -0.12088112
volatile.acidity     -0.0319153683 -0.03572815  0.06771794
citric.acid          -0.1637482114  0.06233094 -0.07572873
residual.sugar       -0.1941334540 -0.02666437 -0.45063122
chlorides            -0.0904394560  0.01676288 -0.36018871
free.sulfur.dioxide  -0.0006177961  0.05921725 -0.25010394
total.sulfur.dioxide  0.0023209718  0.13456237 -0.44889210
density              -0.0935914935  0.07449315 -0.78013762
pH                    1.0000000000  0.15595150  0.12143210
sulphates             0.1559514973  1.00000000 -0.01743277
alcohol               0.1214320987 -0.01743277  1.00000000
```

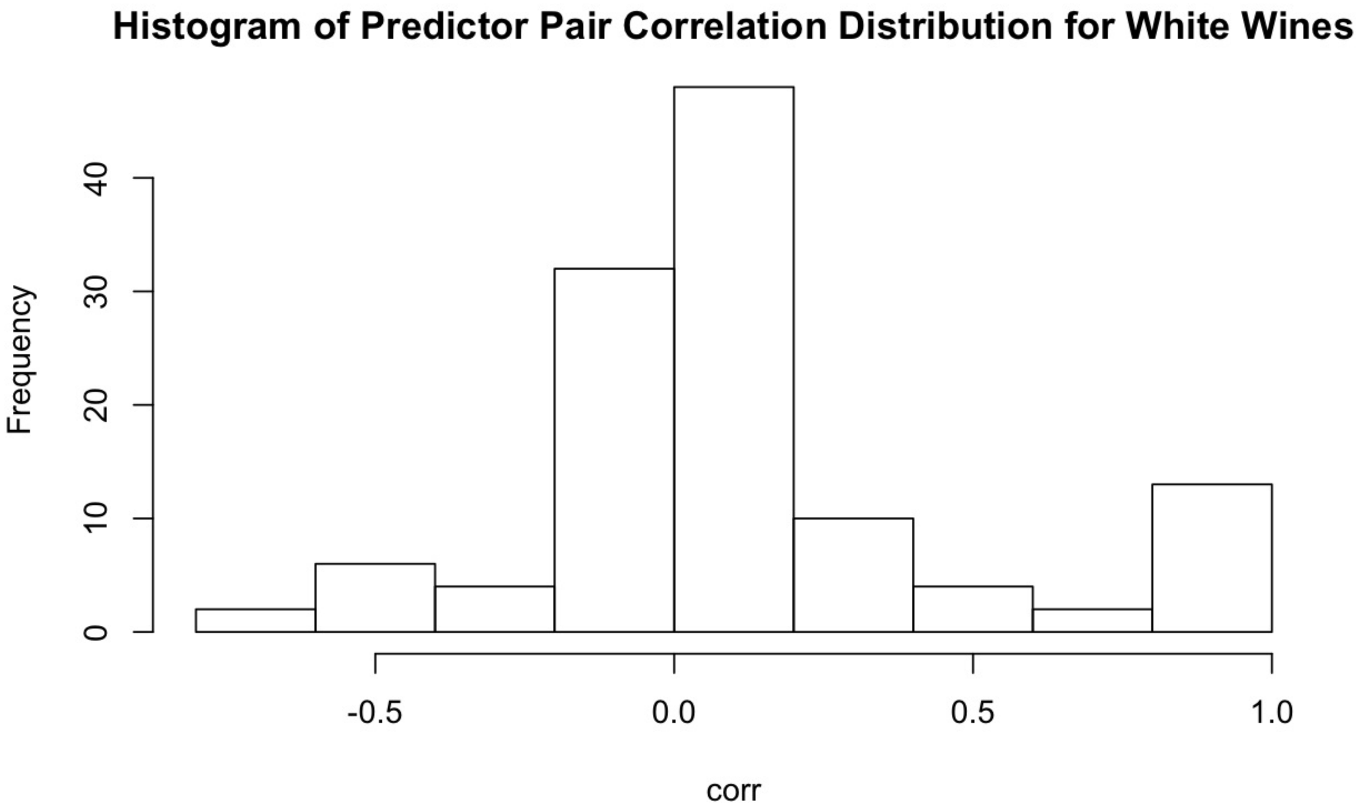
Hide

```
# Plot Heatmap
heatmap(corr, main="Correlations Between Predictors for White Wines", Colv = NA, Rowv = NA, scale="column")
```



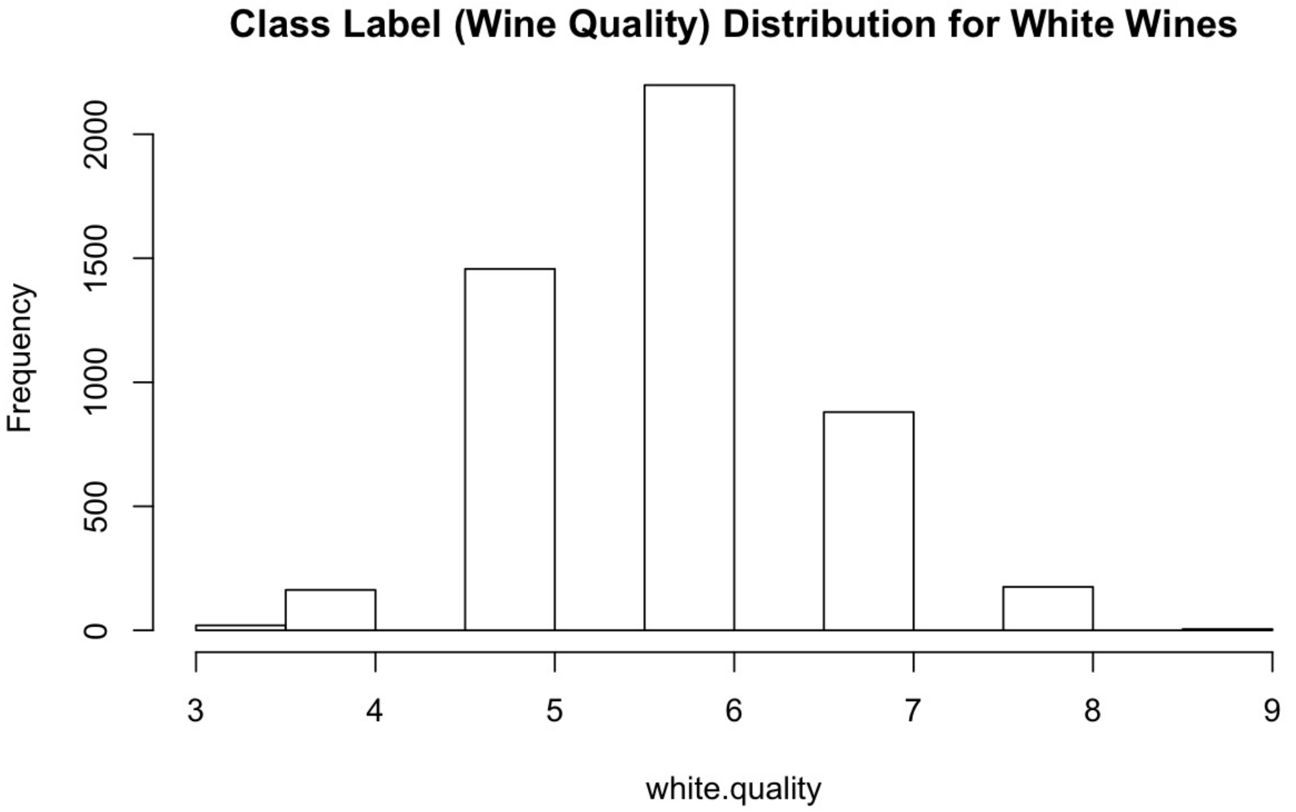
Hide

```
# Plot histogram
hist(corr, main='Histogram of Predictor Pair Correlation Distribution for White Wines')
```



Hide

```
# Plot histogram between class labels
class_dist <- hist(white.quality, main='Class Label (Wine Quality) Distribution for White Wines')
```



Hide

```
class_dist$counts
```

```
[1]   20  163    0 1457    0 2198    0  880    0  175    0    5
```

We can see that in the Wines dataset, there is very little correlation between the different predictors, so it is very difficult to find patterns among the variables, because there is not that many predictors (only 12 of them) and they are not very strong indicators. From the heat maps, we can see that most of the non-diagonal entries have a lighter color, which shows a weak relationship between the variables. Looking at the histograms, most of these predictor pairs have a correlation closer to zero, and very few of them greater than 0.3. This contributes to low accuracies accross different models, since we can’t utilize any inter-predictor patterns to better learn about the data.

Moreover, we can also see that the class labels (wine qualities) are extremely unbalanced, as the bulk of the labels are 5, 6, or 7, yet there are very few examples that are 4 or lower or 8 or above. The models, in turn, will also spit out class labels that are equally unbalanced, as the predictors will also predict mostly 5s, 6s, or 7s, while very rarely predicting any other quality value.

Reference: https://ai.plainenglish.io/estimating-wine-quality-with-machine-learning-ai-72-accuracy-8a5ff0bab3b2 (https://ai.plainenglish.io/estimating-wine-quality-with-machine-learning-ai-72-accuracy-8a5ff0bab3b2)