Wine Quality

Ankush Morey

6/19/2021

Part A. Background and data exploration

```
df <- read.csv('C:\\Users\\ankus\\Desktop\\Ankush\\BANA\\Statistical Models\\Case_Study\\winequality-re
```

Summary Statistics

```
dim(df)
```

```
## [1] 1599   12
```

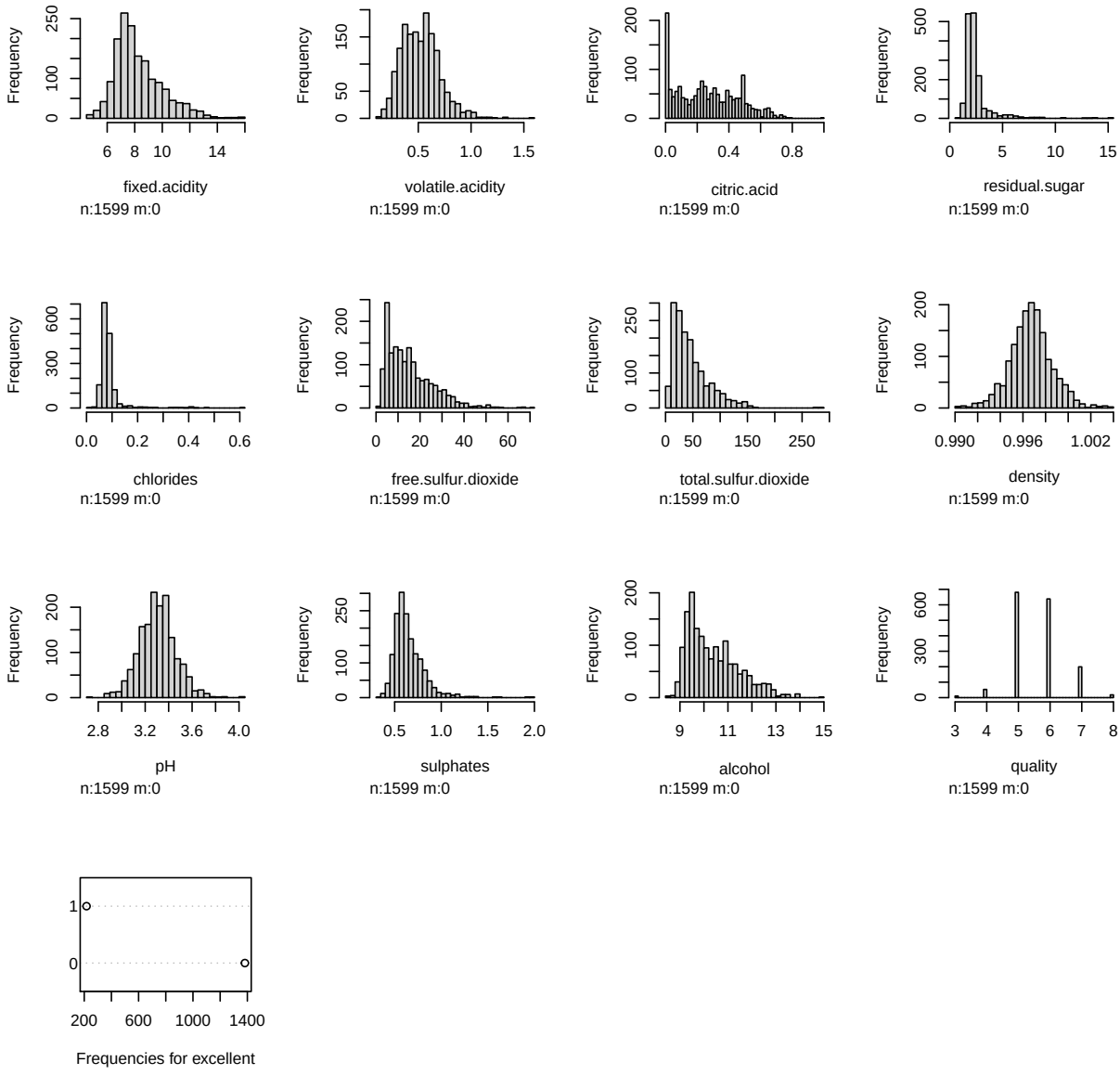
```
#str(df)
summary(df)
```

```
##  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
```

Part B. Visualization and initial models for a binary response

###Distribution of variables #### Histograms

```
df$excellent <- as.factor(ifelse(df$quality>=7,1,0))
hist.data.frame(df)
```



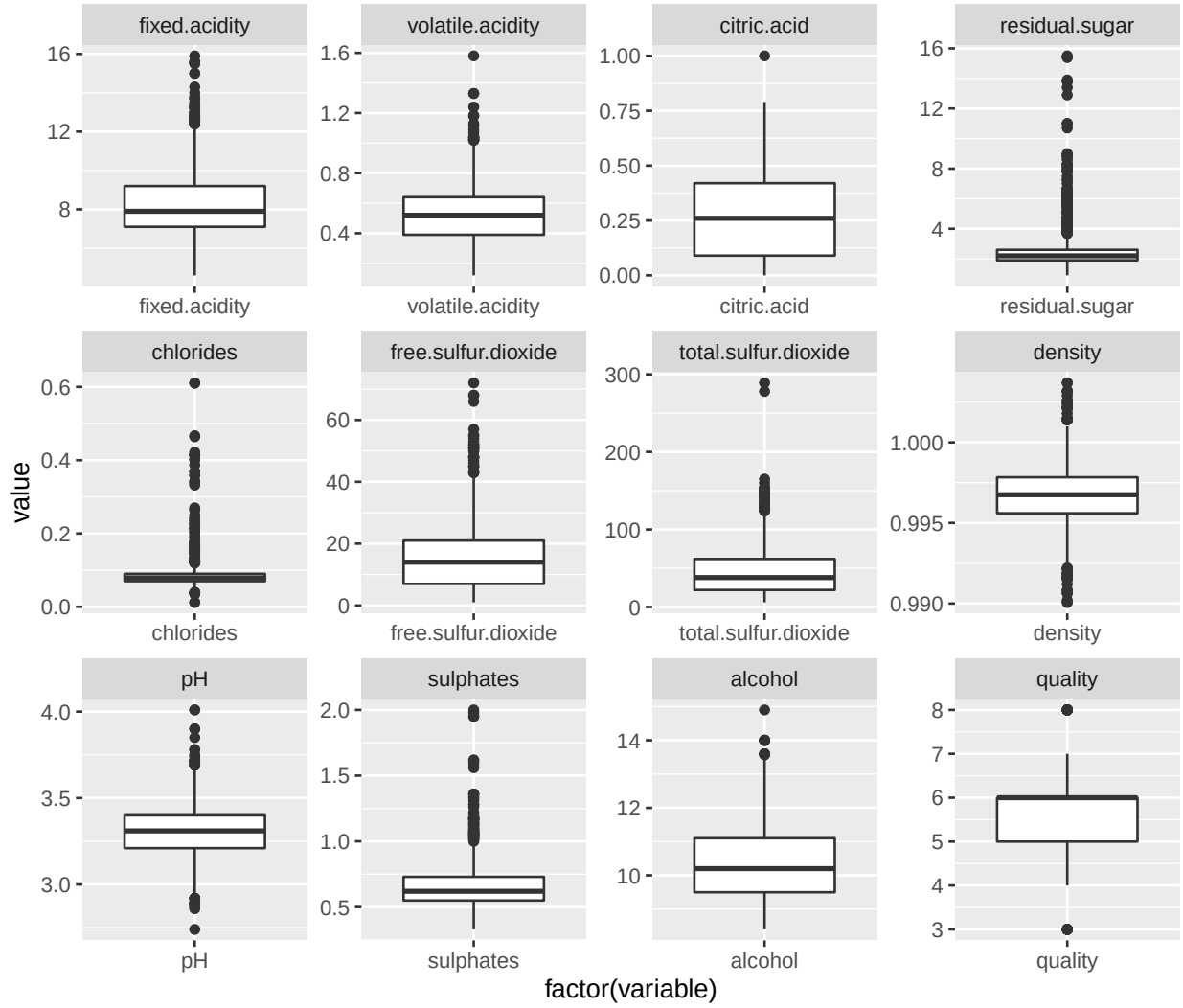

```
#table(df$excellent)
```

```
meltData <- melt(df)
```

Boxplots

```
## Using excellent as id variables
```

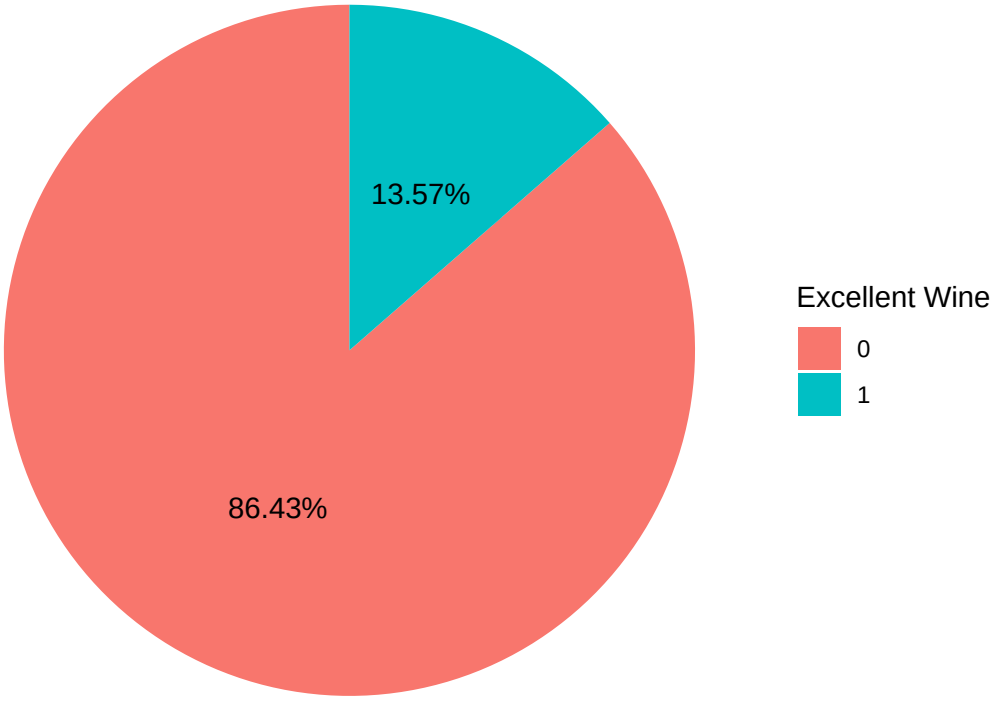
```
p <- ggplot(meltData, aes(factor(variable), value))
p + geom_boxplot() + facet_wrap(~variable, scale="free")
```




```
ggplot(df, aes(x = "", fill = factor(excellent))) +
  geom_bar(width = 1) + coord_polar("y", start=0) +
  geom_text(data = as.data.frame(table(df$excellent)),
            aes(group = Var1, label = scales::percent(Freq / sum(Freq),accuracy = 0.01),
                y = Freq, fill = Var1),
            position = position_stack(vjust = 0.5)) + theme_void() +
  guides(fill=guide_legend(title="Excellent Wine"))
```

Pie Chart

```
## Warning: Ignoring unknown aesthetics: fill
```

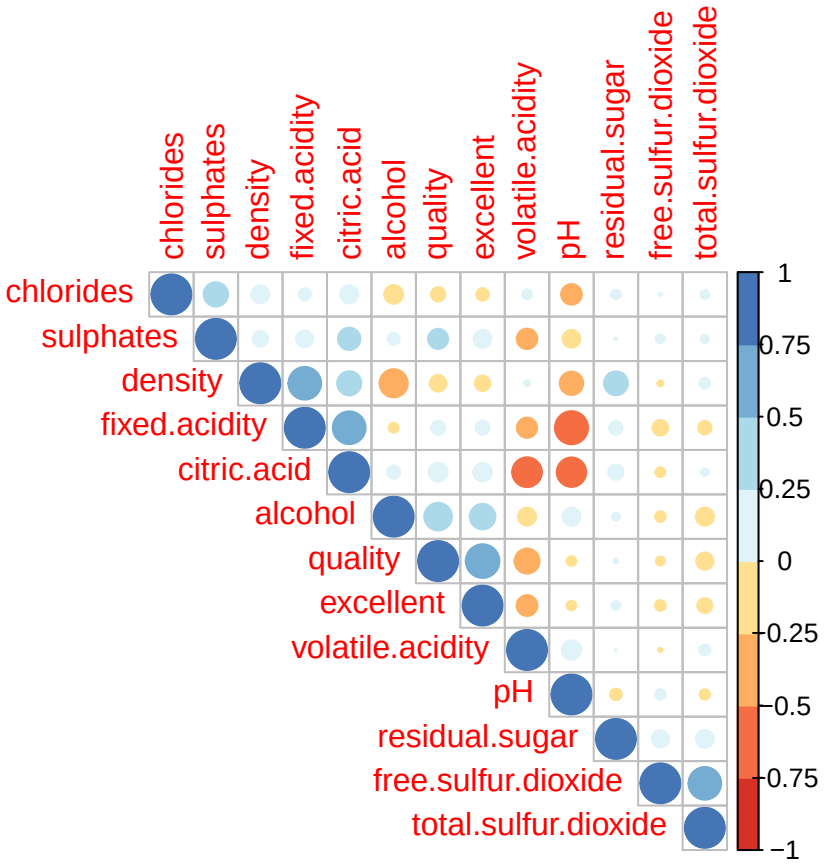



```
df$excellent<-as.numeric(df$excellent)
M <-cor(df)
cor(df,as.numeric(df$excellent))
```

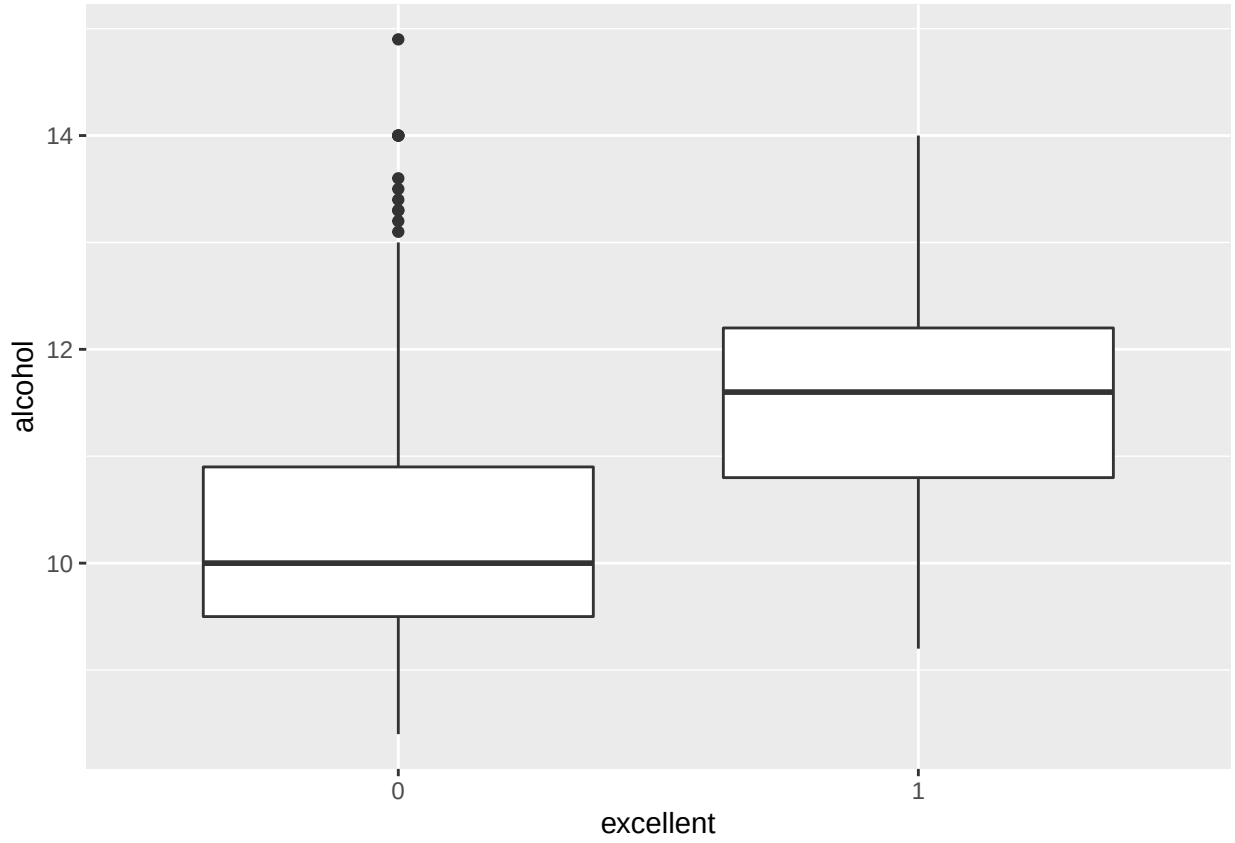
Association with response

```
##                              [,1]
## fixed.acidity         0.12006104
## volatile.acidity     -0.27071153
## citric.acid           0.21471559
## residual.sugar        0.04777895
## chlorides            -0.09730764
## free.sulfur.dioxide  -0.07174730
## total.sulfur.dioxide -0.13951655
## density              -0.15045968
## pH                   -0.05728334
## sulphates             0.19948521
## alcohol               0.40731485
## quality               0.71019625
## excellent             1.00000000
```

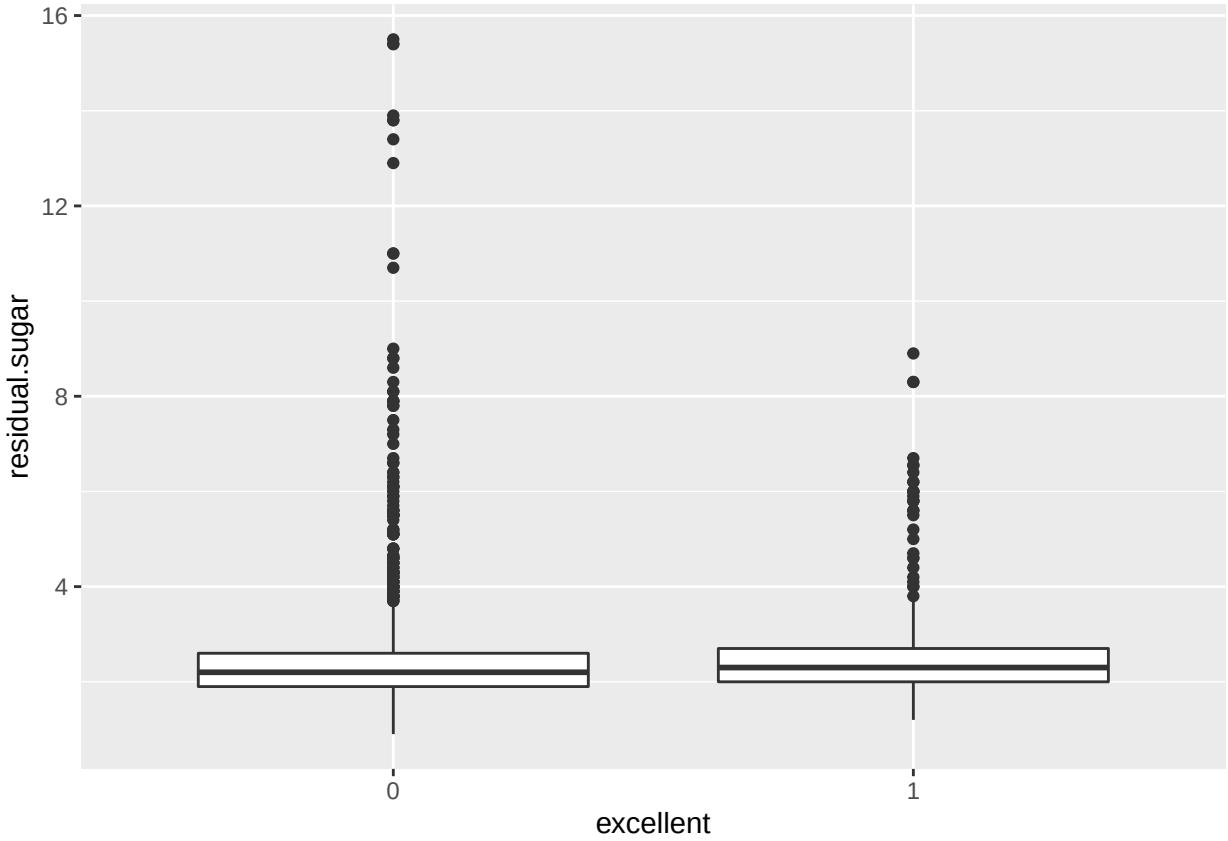
```
corrplot(M, type="upper", order="hclust",col=brewer.pal(n=8, name="RdYlBu"))
```




```
df$excellent <- as.factor(ifelse(df$quality>=7,1,0))
ggplot(data=df, mapping = aes(x =excellent, y = alcohol)) +
  geom_boxplot()
```




```
ggplot(data=df, mapping = aes(x =excellent, y = residual.sugar)) +
  geom_boxplot()
```




```
df1<-df
df1$quality<-NULL

mod_linear <- lm(as.numeric(excellent)~.,df1)
summary(mod_linear)
```

Linear model

```
##
## Call:
## lm(formula = as.numeric(excellent) ~ ., data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92396 -0.17446 -0.04220  0.05006  0.99838
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.643e+01  9.792e+00   3.721 0.000206 ***
## fixed.acidity         3.400e-02  1.199e-02   2.836 0.004620 **
## volatile.acidity     -1.783e-01  5.595e-02  -3.187 0.001468 **
## citric.acid           8.683e-02  6.799e-02   1.277 0.201774
## residual.sugar        2.520e-02  6.931e-03   3.635 0.000286 ***
```

```
## chlorides            -6.556e-01  1.937e-01  -3.385 0.000730 ***
## free.sulfur.dioxide  -5.526e-04  1.003e-03  -0.551 0.581817
## total.sulfur.dioxide -6.658e-04  3.367e-04  -1.977 0.048159 *
## density              -3.668e+01  9.994e+00  -3.670 0.000251 ***
## pH                    1.725e-02  8.852e-02   0.195 0.845516
## sulphates             3.515e-01  5.282e-02   6.655 3.89e-11 ***
## alcohol               7.618e-02  1.224e-02   6.226 6.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2994 on 1587 degrees of freedom
## Multiple R-squared:  0.2416, Adjusted R-squared:  0.2363
## F-statistic: 45.96 on 11 and 1587 DF,  p-value: < 2.2e-16
```

```
mod_log <- glm(excellent~.,family = binomial,data = df1)
summary(mod_log)
```

Logistic Model

```
##
## Call:
## glm(formula = excellent ~ ., family = binomial, data = df1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9878  -0.4351  -0.2207  -0.1222   2.9869
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.428e+02  1.081e+02   2.247 0.024660 *
## fixed.acidity         2.750e-01  1.253e-01   2.195 0.028183 *
## volatile.acidity     -2.581e+00  7.843e-01  -3.291 0.000999 ***
## citric.acid           5.678e-01  8.385e-01   0.677 0.498313
## residual.sugar        2.395e-01  7.373e-02   3.248 0.001163 **
## chlorides            -8.816e+00  3.365e+00  -2.620 0.008788 **
## free.sulfur.dioxide   1.082e-02  1.223e-02   0.884 0.376469
## total.sulfur.dioxide -1.653e-02  4.894e-03  -3.378 0.000731 ***
## density              -2.578e+02  1.104e+02  -2.335 0.019536 *
## pH                    2.242e-01  9.984e-01   0.225 0.822327
## sulphates             3.750e+00  5.416e-01   6.924 4.39e-12 ***
## alcohol               7.533e-01  1.316e-01   5.724 1.04e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  870.86  on 1587  degrees of freedom
## AIC: 894.86
##
## Number of Fisher Scoring iterations: 6
```

Part C. Feature Selection and Model Building

```
nullmodel=glm(excellent~1, family=binomial, data=df1)
fullmodel=glm(excellent~., family=binomial, data=df1)
```

Stepwise

```
mod_step_f_aic <- step(nullmodel, scope=list(lower=nullmodel, upper=fullmodel), direction='forward')
```

Forward selection wit AIC

```
## Start:  AIC=1271.92
## excellent ~ 1
##
##                        Df Deviance    AIC
## + alcohol               1   1027.9 1031.9
## + volatile.acidity      1   1130.6 1134.6
## + citric.acid           1   1197.2 1201.2
## + sulphates             1   1218.1 1222.1
## + total.sulfur.dioxide  1   1232.5 1236.5
## + density               1   1233.2 1237.2
## + chlorides             1   1239.5 1243.5
## + fixed.acidity         1   1248.4 1252.4
## + free.sulfur.dioxide   1   1261.1 1265.1
## + pH                    1   1264.6 1268.6
## + residual.sugar        1   1266.7 1270.7
## <none>                      1269.9 1271.9
##
## Step:  AIC=1031.89
## excellent ~ alcohol
##
##                        Df Deviance     AIC
## + volatile.acidity      1   948.48  954.48
## + citric.acid           1   975.02  981.02
## + sulphates             1   975.76  981.76
## + fixed.acidity         1   987.08  993.08
## + pH                    1   991.30  997.30
## + total.sulfur.dioxide  1  1013.65 1019.65
## + density               1  1019.13 1025.13
## + free.sulfur.dioxide   1  1023.45 1029.45
## <none>                     1027.89 1031.89
## + chlorides             1  1026.32 1032.32
## + residual.sugar        1  1026.50 1032.50
##
## Step:  AIC=954.48
## excellent ~ alcohol + volatile.acidity
##
##                        Df Deviance    AIC
```

```
## + sulphates             1   917.26 925.26
## + fixed.acidity         1   932.01 940.01
## + total.sulfur.dioxide  1   936.34 944.34
## + pH                    1   937.82 945.82
## + citric.acid           1   941.11 949.11
## + density               1   941.30 949.30
## + free.sulfur.dioxide   1   944.00 952.00
## <none>                      948.48 954.48
## + residual.sugar        1   946.90 954.90
## + chlorides             1   947.52 955.52
##
## Step:  AIC=925.26
## excellent ~ alcohol + volatile.acidity + sulphates
##
##                         Df Deviance    AIC
## + total.sulfur.dioxide  1   899.55 909.55
## + fixed.acidity         1   905.71 915.71
## + free.sulfur.dioxide   1   910.42 920.42
## + chlorides             1   911.20 921.20
## + pH                    1   911.70 921.70
## + citric.acid           1   914.11 924.11
## + density               1   914.82 924.82
## + residual.sugar        1   915.24 925.24
## <none>                      917.26 925.26
##
## Step:  AIC=909.55
## excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide
##
##                        Df Deviance    AIC
## + chlorides             1   890.90 902.90
## + fixed.acidity         1   893.48 905.48
## + residual.sugar        1   894.74 906.74
## + pH                    1   895.19 907.19
## <none>                      899.55 909.55
## + citric.acid           1   897.60 909.60
## + density               1   898.65 910.65
## + free.sulfur.dioxide   1   899.31 911.31
##
## Step:  AIC=902.9
## excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides
##
##                        Df Deviance    AIC
## + fixed.acidity         1   883.17 897.17
## + pH                    1   883.36 897.36
## + residual.sugar        1   884.50 898.50
## + citric.acid           1   884.85 898.85
## <none>                      890.90 902.90
## + density               1   889.04 903.04
## + free.sulfur.dioxide   1   890.74 904.74
##
## Step:  AIC=897.17
## excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + fixed.acidity
```

```
##
##                        Df Deviance    AIC
## + residual.sugar        1   878.99 894.99
## <none>                      883.17 897.17
## + density               1   881.60 897.60
## + pH                    1   881.79 897.79
## + free.sulfur.dioxide   1   882.71 898.71
## + citric.acid           1   882.72 898.72
##
## Step:  AIC=894.99
## excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + fixed.acidity + residual.sugar
##
##                        Df Deviance    AIC
## + density               1   872.08 890.08
## <none>                      878.99 894.99
## + pH                    1   877.59 895.59
## + free.sulfur.dioxide   1   878.15 896.15
## + citric.acid           1   878.86 896.86
##
## Step:  AIC=890.08
## excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + fixed.acidity + residual.sugar + density
##
##                        Df Deviance    AIC
## <none>                      872.08 890.08
## + free.sulfur.dioxide   1   871.33 891.33
## + citric.acid           1   871.78 891.78
## + pH                    1   872.01 892.01
```

excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar + density

```
mod_step_b_aic <- step(fullmodel)
```

Backward selection wit AIC

```
## Start:  AIC=894.86
## excellent ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##
##                         Df Deviance    AIC
## - pH                     1   870.91 892.91
## - citric.acid            1   871.32 893.32
## - free.sulfur.dioxide    1   871.64 893.64
## <none>                       870.86 894.86
## - fixed.acidity          1   875.67 897.67
## - density                1   876.34 898.34
## - residual.sugar         1   880.02 902.02
## - chlorides              1   880.85 902.85
```

```
## - volatile.acidity      1   882.52 904.52
## - total.sulfur.dioxide  1   884.49 906.49
## - alcohol               1   904.51 926.51
## - sulphates             1   915.26 937.26
##
## Step:  AIC=892.91
## excellent ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + sulphates + alcohol
##
##                         Df Deviance    AIC
## - citric.acid            1   871.33 891.33
## - free.sulfur.dioxide    1   871.78 891.78
## <none>                       870.91 892.91
## - density                1   877.94 897.94
## - fixed.acidity          1   878.81 898.81
## - residual.sugar         1   880.40 900.40
## - chlorides              1   881.53 901.53
## - volatile.acidity       1   882.72 902.72
## - total.sulfur.dioxide   1   885.36 905.36
## - sulphates              1   915.50 935.50
## - alcohol                1   916.76 936.76
##
## Step:  AIC=891.33
## excellent ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + sulphates + alcohol
##
##                         Df Deviance    AIC
## - free.sulfur.dioxide    1   872.08 890.08
## <none>                       871.33 891.33
## - density                1   878.15 896.15
## - residual.sugar         1   881.27 899.27
## - chlorides              1   881.76 899.76
## - fixed.acidity          1   883.87 901.87
## - total.sulfur.dioxide   1   885.36 903.36
## - volatile.acidity       1   892.88 910.88
## - sulphates              1   915.82 933.82
## - alcohol                1   921.78 939.78
##
## Step:  AIC=890.08
## excellent ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + total.sulfur.dioxide + density + sulphates +
##     alcohol
##
##                         Df Deviance    AIC
## <none>                       872.08 890.08
## - density                1   878.99 894.99
## - residual.sugar         1   881.60 897.60
## - chlorides              1   882.47 898.47
## - fixed.acidity          1   884.45 900.45
## - total.sulfur.dioxide   1   890.34 906.34
## - volatile.acidity       1   894.16 910.16
## - sulphates              1   917.01 933.01
```

```
## - alcohol               1   922.50 938.50
```

excellent ~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + total.sulfur.dioxide + density + sulphates + alcohol

```
mod_step_f_bic <- step(nullmodel, scope=list(lower=nullmodel, upper=fullmodel), direction='forward',k=lo
```

Forward selection with BIC

```
## Start:  AIC=1277.3
## excellent ~ 1
##
##                         Df Deviance    AIC
## + alcohol                1   1027.9 1042.6
## + volatile.acidity       1   1130.6 1145.3
## + citric.acid            1   1197.2 1212.0
## + sulphates              1   1218.1 1232.8
## + total.sulfur.dioxide   1   1232.5 1247.2
## + density                1   1233.2 1247.9
## + chlorides              1   1239.5 1254.3
## + fixed.acidity          1   1248.4 1263.2
## + free.sulfur.dioxide    1   1261.1 1275.8
## <none>                       1269.9 1277.3
## + pH                     1   1264.6 1279.4
## + residual.sugar         1   1266.7 1281.5
##
## Step:  AIC=1042.64
## excellent ~ alcohol
##
##                         Df Deviance     AIC
## + volatile.acidity       1   948.48  970.62
## + citric.acid            1   975.02  997.15
## + sulphates              1   975.76  997.89
## + fixed.acidity          1   987.08 1009.21
## + pH                     1   991.30 1013.44
## + total.sulfur.dioxide   1  1013.65 1035.78
## + density                1  1019.13 1041.26
## <none>                      1027.89 1042.64
## + free.sulfur.dioxide    1  1023.45 1045.58
## + chlorides              1  1026.32 1048.45
## + residual.sugar         1  1026.50 1048.63
##
## Step:  AIC=970.62
## excellent ~ alcohol + volatile.acidity
##
##                         Df Deviance    AIC
## + sulphates              1   917.26 946.77
## + fixed.acidity          1   932.01 961.52
## + total.sulfur.dioxide   1   936.34 965.85
## + pH                     1   937.82 967.33
## <none>                       948.48 970.62
```

```
## + citric.acid           1   941.11 970.62
## + density               1   941.30 970.81
## + free.sulfur.dioxide   1   944.00 973.51
## + residual.sugar        1   946.90 976.41
## + chlorides             1   947.52 977.03
##
## Step:  AIC=946.77
## excellent ~ alcohol + volatile.acidity + sulphates
##
##                        Df Deviance    AIC
## + total.sulfur.dioxide  1   899.55 936.44
## + fixed.acidity         1   905.71 942.59
## <none>                      917.26 946.77
## + free.sulfur.dioxide   1   910.42 947.30
## + chlorides             1   911.20 948.09
## + pH                    1   911.70 948.58
## + citric.acid           1   914.11 951.00
## + density               1   914.82 951.71
## + residual.sugar        1   915.24 952.13
##
## Step:  AIC=936.44
## excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide
##
##                       Df Deviance    AIC
## + chlorides            1   890.90 935.17
## <none>                     899.55 936.44
## + fixed.acidity        1   893.48 937.74
## + residual.sugar       1   894.74 939.01
## + pH                   1   895.19 939.45
## + citric.acid          1   897.60 941.86
## + density              1   898.65 942.91
## + free.sulfur.dioxide  1   899.31 943.57
##
## Step:  AIC=935.17
## excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides
##
##                       Df Deviance    AIC
## + fixed.acidity        1   883.17 934.81
## + pH                   1   883.36 935.00
## <none>                     890.90 935.17
## + residual.sugar       1   884.50 936.14
## + citric.acid          1   884.85 936.49
## + density              1   889.04 940.68
## + free.sulfur.dioxide  1   890.74 942.38
##
## Step:  AIC=934.81
## excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + fixed.acidity
##
##                       Df Deviance    AIC
## <none>                     883.17 934.81
## + residual.sugar       1   878.99 938.01
## + density              1   881.60 940.62
```

```
## + pH                   1   881.79 940.81
## + free.sulfur.dioxide  1   882.71 941.73
## + citric.acid          1   882.72 941.73
```

excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide + chlorides + fixed.acidity

```
mod_step_b_bic <- step(fullmodel,k=log(nrow(df1)))
```

Backward selection wit BIC

```
## Start:  AIC=959.39
## excellent ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##
##                        Df Deviance    AIC
## - pH                    1   870.91 952.06
## - citric.acid           1   871.32 952.47
## - free.sulfur.dioxide   1   871.64 952.79
## - fixed.acidity         1   875.67 956.82
## - density               1   876.34 957.49
## <none>                      870.86 959.39
## - residual.sugar        1   880.02 961.17
## - chlorides             1   880.85 962.00
## - volatile.acidity      1   882.52 963.67
## - total.sulfur.dioxide  1   884.49 965.64
## - alcohol               1   904.51 985.66
## - sulphates             1   915.26 996.40
##
## Step:  AIC=952.06
## excellent ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + sulphates + alcohol
##
##                        Df Deviance    AIC
## - citric.acid           1   871.33 945.10
## - free.sulfur.dioxide   1   871.78 945.55
## - density               1   877.94 951.71
## <none>                      870.91 952.06
## - fixed.acidity         1   878.81 952.58
## - residual.sugar        1   880.40 954.17
## - chlorides             1   881.53 955.30
## - volatile.acidity      1   882.72 956.49
## - total.sulfur.dioxide  1   885.36 959.13
## - sulphates             1   915.50 989.27
## - alcohol               1   916.76 990.53
##
## Step:  AIC=945.1
## excellent ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + sulphates + alcohol
```

```
##
##                         Df Deviance    AIC
## - free.sulfur.dioxide    1   872.08 938.47
## - density                1   878.15 944.54
## <none>                       871.33 945.10
## - residual.sugar         1   881.27 947.67
## - chlorides              1   881.76 948.15
## - fixed.acidity          1   883.87 950.27
## - total.sulfur.dioxide   1   885.36 951.75
## - volatile.acidity       1   892.88 959.28
## - sulphates              1   915.82 982.22
## - alcohol                1   921.78 988.17
##
## Step:  AIC=938.47
## excellent ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + total.sulfur.dioxide + density + sulphates +
##     alcohol
##
##                         Df Deviance    AIC
## - density                1   878.99 938.01
## <none>                       872.08 938.47
## - residual.sugar         1   881.60 940.62
## - chlorides              1   882.47 941.48
## - fixed.acidity          1   884.45 943.46
## - total.sulfur.dioxide   1   890.34 949.36
## - volatile.acidity       1   894.16 953.17
## - sulphates              1   917.01 976.03
## - alcohol                1   922.50 981.52
##
## Step:  AIC=938.01
## excellent ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + total.sulfur.dioxide + sulphates + alcohol
##
##                         Df Deviance     AIC
## - residual.sugar         1   883.17  934.81
## - fixed.acidity          1   884.50  936.14
## <none>                       878.99  938.01
## - chlorides              1   890.52  942.16
## - total.sulfur.dioxide   1   895.86  947.50
## - volatile.acidity       1   912.55  964.19
## - sulphates              1   918.23  969.87
## - alcohol                1  1027.14 1078.78
##
## Step:  AIC=934.81
## excellent ~ fixed.acidity + volatile.acidity + chlorides + total.sulfur.dioxide +
##     sulphates + alcohol
##
##                         Df Deviance     AIC
## <none>                       883.17  934.81
## - fixed.acidity          1   890.90  935.17
## - chlorides              1   893.48  937.74
## - total.sulfur.dioxide   1   897.23  941.49
## - volatile.acidity       1   915.58  959.85
## - sulphates              1   920.83  965.10
```

```
## - alcohol               1  1040.67 1084.93
```

excellent ~ fixed.acidity + volatile.acidity + chlorides + total.sulfur.dioxide + sulphates + alcohol

```
model_step_s <- step(nullmodel, scope=list(lower=nullmodel, upper=fullmodel), direction='both')
```

Stepwise selection

```
## Start:  AIC=1271.92
## excellent ~ 1
##
##                        Df Deviance    AIC
## + alcohol               1   1027.9 1031.9
## + volatile.acidity      1   1130.6 1134.6
## + citric.acid           1   1197.2 1201.2
## + sulphates             1   1218.1 1222.1
## + total.sulfur.dioxide  1   1232.5 1236.5
## + density               1   1233.2 1237.2
## + chlorides             1   1239.5 1243.5
## + fixed.acidity         1   1248.4 1252.4
## + free.sulfur.dioxide   1   1261.1 1265.1
## + pH                    1   1264.6 1268.6
## + residual.sugar        1   1266.7 1270.7
## <none>                      1269.9 1271.9
##
## Step:  AIC=1031.89
## excellent ~ alcohol
##
##                        Df Deviance     AIC
## + volatile.acidity      1   948.48  954.48
## + citric.acid           1   975.02  981.02
## + sulphates             1   975.76  981.76
## + fixed.acidity         1   987.08  993.08
## + pH                    1   991.30  997.30
## + total.sulfur.dioxide  1  1013.65 1019.65
## + density               1  1019.13 1025.13
## + free.sulfur.dioxide   1  1023.45 1029.45
## <none>                     1027.89 1031.89
## + chlorides             1  1026.32 1032.32
## + residual.sugar        1  1026.50 1032.50
## - alcohol               1  1269.92 1271.92
##
## Step:  AIC=954.48
## excellent ~ alcohol + volatile.acidity
##
##                        Df Deviance     AIC
## + sulphates             1   917.26  925.26
## + fixed.acidity         1   932.01  940.01
## + total.sulfur.dioxide  1   936.34  944.34
## + pH                    1   937.82  945.82
## + citric.acid           1   941.11  949.11
```

```
## + density               1   941.30  949.30
## + free.sulfur.dioxide   1   944.00  952.00
## <none>                      948.48  954.48
## + residual.sugar        1   946.90  954.90
## + chlorides             1   947.52  955.52
## - volatile.acidity      1  1027.89 1031.89
## - alcohol               1  1130.57 1134.57
##
## Step:  AIC=925.26
## excellent ~ alcohol + volatile.acidity + sulphates
##
##                         Df Deviance     AIC
## + total.sulfur.dioxide  1   899.55  909.55
## + fixed.acidity         1   905.71  915.71
## + free.sulfur.dioxide   1   910.42  920.42
## + chlorides             1   911.20  921.20
## + pH                    1   911.70  921.70
## + citric.acid           1   914.11  924.11
## + density               1   914.82  924.82
## + residual.sugar        1   915.24  925.24
## <none>                      917.26  925.26
## - sulphates             1   948.48  954.48
## - volatile.acidity      1   975.76  981.76
## - alcohol               1  1106.33 1112.33
##
## Step:  AIC=909.55
## excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide
##
##                         Df Deviance     AIC
## + chlorides             1   890.90  902.90
## + fixed.acidity         1   893.48  905.48
## + residual.sugar        1   894.74  906.74
## + pH                    1   895.19  907.19
## <none>                      899.55  909.55
## + citric.acid           1   897.60  909.60
## + density               1   898.65  910.65
## + free.sulfur.dioxide   1   899.31  911.31
## - total.sulfur.dioxide  1   917.26  925.26
## - sulphates             1   936.34  944.34
## - volatile.acidity      1   950.51  958.51
## - alcohol               1  1075.29 1083.29
##
## Step:  AIC=902.9
## excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides
##
##                         Df Deviance     AIC
## + fixed.acidity         1   883.17  897.17
## + pH                    1   883.36  897.36
## + residual.sugar        1   884.50  898.50
## + citric.acid           1   884.85  898.85
## <none>                      890.90  902.90
## + density               1   889.04  903.04
## + free.sulfur.dioxide   1   890.74  904.74
```

```
## - chlorides             1   899.55  909.55
## - total.sulfur.dioxide  1   911.20  921.20
## - sulphates             1   934.54  944.54
## - volatile.acidity      1   936.08  946.08
## - alcohol               1  1041.17 1051.17
##
## Step:  AIC=897.17
## excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + fixed.acidity
##
##                         Df Deviance     AIC
## + residual.sugar         1   878.99  894.99
## <none>                       883.17  897.17
## + density                1   881.60  897.60
## + pH                     1   881.79  897.79
## + free.sulfur.dioxide    1   882.71  898.71
## + citric.acid            1   882.72  898.72
## - fixed.acidity          1   890.90  902.90
## - chlorides              1   893.48  905.48
## - total.sulfur.dioxide   1   897.23  909.23
## - volatile.acidity       1   915.58  927.58
## - sulphates              1   920.83  932.83
## - alcohol                1  1040.67 1052.67
##
## Step:  AIC=894.99
## excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + fixed.acidity + residual.sugar
##
##                         Df Deviance     AIC
## + density                1   872.08  890.08
## <none>                       878.99  894.99
## + pH                     1   877.59  895.59
## + free.sulfur.dioxide    1   878.15  896.15
## + citric.acid            1   878.86  896.86
## - residual.sugar         1   883.17  897.17
## - fixed.acidity          1   884.50  898.50
## - chlorides              1   890.52  904.52
## - total.sulfur.dioxide   1   895.86  909.86
## - volatile.acidity       1   912.55  926.55
## - sulphates              1   918.23  932.23
## - alcohol                1  1027.14 1041.14
##
## Step:  AIC=890.08
## excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + fixed.acidity + residual.sugar + density
##
##                         Df Deviance    AIC
## <none>                       872.08 890.08
## + free.sulfur.dioxide    1   871.33 891.33
## + citric.acid            1   871.78 891.78
## + pH                     1   872.01 892.01
## - density                1   878.99 894.99
## - residual.sugar         1   881.60 897.60
## - chlorides              1   882.47 898.47
```

```
## - fixed.acidity         1   884.45 900.45
## - total.sulfur.dioxide  1   890.34 906.34
## - volatile.acidity      1   894.16 910.16
## - sulphates             1   917.01 933.01
## - alcohol               1   922.50 938.50
```

```
excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
chlorides + fixed.acidity + residual.sugar + density
```

```
## excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + fixed.acidity + residual.sugar + density
```

```
summary(mod_step_f_bic)
```

Chi square test

```
##
## Call:
## glm(formula = excellent ~ alcohol + volatile.acidity + sulphates +
##     total.sulfur.dioxide + chlorides + fixed.acidity, family = binomial,
##     data = df1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6332  -0.4386  -0.2313  -0.1270   3.0411
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -13.283277   1.316313 -10.091  < 2e-16 ***
## alcohol                0.988930   0.086710  11.405  < 2e-16 ***
## volatile.acidity      -3.345588   0.619146  -5.404 6.53e-08 ***
## sulphates              3.323236   0.520367   6.386 1.70e-10 ***
## total.sulfur.dioxide  -0.012355   0.003596  -3.436 0.000591 ***
## chlorides             -8.802351   3.413297  -2.579 0.009913 **
## fixed.acidity          0.133235   0.047596   2.799 0.005121 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  883.17  on 1592  degrees of freedom
## AIC: 897.17
##
## Number of Fisher Scoring iterations: 6
```

```
dev=deviance(mod_step_f_bic)-deviance(fullmodel)
```

Deviance

```
diff_df = mod_step_f_bic$df.residual - fullmodel$df.residual
pchisq(dev,diff_df,lower=F)
```

Difference between degrees of freedom

```
## [1] 0.03079029
```

```
anova(mod_step_f_bic,fullmodel,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + fixed.acidity
## Model 2: excellent ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1592     883.17
## 2      1587     870.86  5   12.309  0.03079 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod_step_f_aic,fullmodel,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
##     chlorides + fixed.acidity + residual.sugar + density
## Model 2: excellent ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1590     872.08
## 2      1587     870.86  3   1.2116   0.7502
```

```
drop1(fullmodel,test='Chi')
```

```
## Single term deletions
##
## Model:
## excellent ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##                       Df Deviance    AIC    LRT  Pr(>Chi)
## <none>                     870.86 894.86
## fixed.acidity          1   875.67 897.67  4.807 0.0283520 *
## volatile.acidity       1   882.52 904.52 11.660 0.0006387 ***
## citric.acid            1   871.32 893.32  0.457 0.4991231
## residual.sugar         1   880.02 902.02  9.153 0.0024826 **
```

```
## chlorides             1   880.85 902.85  9.983 0.0015801 **
## free.sulfur.dioxide   1   871.64 893.64  0.780 0.3771127
## total.sulfur.dioxide  1   884.49 906.49 13.625 0.0002231 ***
## density               1   876.34 898.34  5.475 0.0192895 *
## pH                    1   870.91 892.91  0.050 0.8224870
## sulphates             1   915.26 937.26 44.391 2.689e-11 ***
## alcohol               1   904.51 926.51 33.643 6.620e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Chi square test shows p-value less than significance, we reject the null hypothesis, confirming the models are different. drop1 chi square test exactly the same model as AIC excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar + density

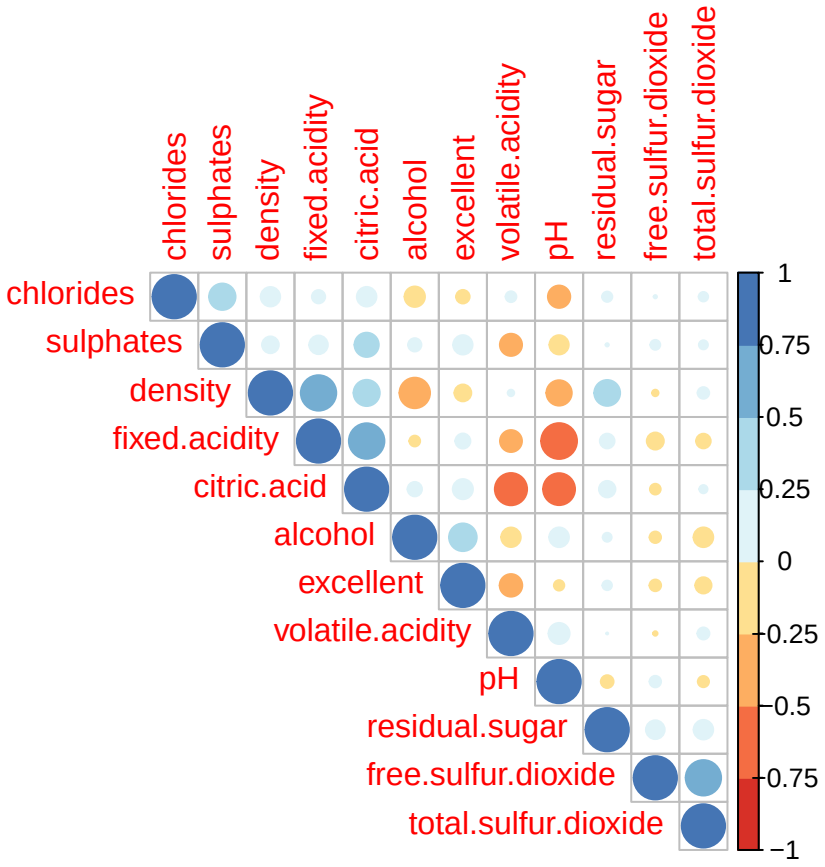
```
df_cor<-df1
df_cor$quality<-NULL
df_cor$excellent=as.numeric(df_cor$excellent)
m=cor(df_cor)
m
```

Collinearity

```
##                       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
## excellent               0.12006104     -0.270711532  0.21471559    0.047778946
##                        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
## excellent            -0.097307638        -0.071747296          -0.13951655
##                           density          pH    sulphates     alcohol
## fixed.acidity         0.66804729 -0.68297819  0.183005664 -0.06166827
## volatile.acidity      0.02202623  0.23493729 -0.260986685 -0.20228803
```

```
## citric.acid           0.36494718 -0.54190414  0.312770044  0.10990325
## residual.sugar        0.35528337 -0.08565242  0.005527121  0.04207544
## chlorides             0.20063233 -0.26502613  0.371260481 -0.22114054
## free.sulfur.dioxide  -0.02194583  0.07037750  0.051657572 -0.06940835
## total.sulfur.dioxide  0.07126948 -0.06649456  0.042946836 -0.20565394
## density               1.00000000 -0.34169933  0.148506412 -0.49617977
## pH                   -0.34169933  1.00000000 -0.196647602  0.20563251
## sulphates             0.14850641 -0.19664760  1.000000000  0.09359475
## alcohol              -0.49617977  0.20563251  0.093594750  1.00000000
## excellent            -0.15045968 -0.05728334  0.199485209  0.40731485
##                       excellent
## fixed.acidity         0.12006104
## volatile.acidity     -0.27071153
## citric.acid           0.21471559
## residual.sugar        0.04777895
## chlorides            -0.09730764
## free.sulfur.dioxide  -0.07174730
## total.sulfur.dioxide -0.13951655
## density              -0.15045968
## pH                   -0.05728334
## sulphates             0.19948521
## alcohol               0.40731485
## excellent             1.00000000
```

```
corrplot(m, type="upper", order="hclust",col=brewer.pal(n=8, name="RdYlBu"))
```



We see multicollinearity in data. pH and fixed.acidity have high negative correlation. Volatile.acidity and citric.acid are correlated. Citric.acid and fixed.acidity are also correlated. Density and fixed.acidity are also correlated. Citric.acid and pH also display negative correlation. Total.sulfur.dioxide and free.sulfur.dioxide are also correlated. From these variables, we have the following pairs in our model:

excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar + density

Density and fixed acidity.

Removing density, as fixed acidity is correlated with citric.acid, density, and pH, thus it captures information of all these variables. Final model:

excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar

```
mod_logf <- glm(excellent~alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +chlorides + fi
summary(mod_logf)
```

```
##
## Call:
## glm(formula = excellent ~ alcohol + volatile.acidity + sulphates +
##     total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar,
##     family = binomial, data = df1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7860  -0.4383  -0.2277  -0.1233   3.0696
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -13.205847   1.323132  -9.981  < 2e-16 ***
## alcohol                0.970575   0.087193  11.131  < 2e-16 ***
## volatile.acidity      -3.417247   0.622217  -5.492 3.97e-08 ***
## sulphates              3.424071   0.524620   6.527 6.72e-11 ***
## total.sulfur.dioxide  -0.013450   0.003574  -3.763 0.000168 ***
## chlorides             -9.536839   3.531569  -2.700 0.006924 **
## fixed.acidity          0.114692   0.048568   2.361 0.018203 *
## residual.sugar         0.134013   0.061511   2.179 0.029355 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  878.99  on 1591  degrees of freedom
## AIC: 894.99
##
## Number of Fisher Scoring iterations: 6
```

We have alcohol and residual sugar in our final model.

```
pred<- predict(mod_logf, newdata = df1, type="response")
conf_sym <-table(df$excellent, (pred > (0.5)), dnn=c("Truth","Predicted"))
conf_sym
```

Confusion Matrix with cutoff p= 0.5

```
##      Predicted
## Truth FALSE TRUE
##     0  1337   45
##     1   146   71
```

```
Specificity_log <- conf_sym[1]/(conf_sym[1]+conf_sym[3])
Specificity_log
```

Specificity (True Negative Rate)

```
## [1] 0.9674385
```

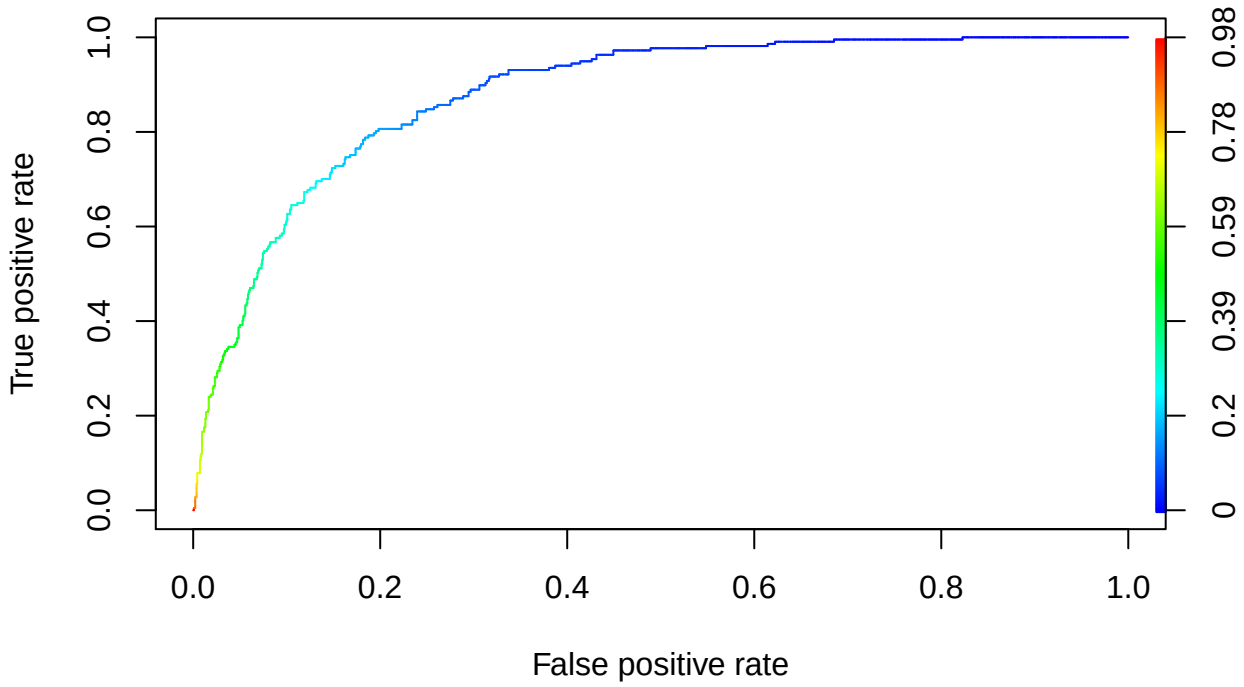
```
Sensitivity_log <- conf_sym[4]/(conf_sym[2]+conf_sym[4])
Sensitivity_log
```

Sensitivity (True Positive Rate)

```
## [1] 0.3271889
```

We need to lower the threshold to decrease the number of False Negatives.

```
predic <- prediction(pred, df$excellent)
perf <- performance(predic,"tpr","fpr")
plot(perf,colorize=TRUE)
```



ROC Curve
AUC

```
# auc.tmp <- performance(predic,"auc")
# auc <- as.numeric(auc.tmp@y.values)
# auc
unlist(slot(performance(predic, "auc"), "y.values"))
```

```
## [1] 0.8794441
```

```
new_df1<-(df[1,])
predict(mod_logf,new_df1,type='response')
```

Prediction on 1st bottle

```
##           1
## 0.009605323
```

```
predict(mod_logf,newdata=new_df1,type="link") ### Linear predictor
```

```
##         1
## -4.635786
```

```r
#predict(mod_logf,newdata=new.ind,type="response") ### Probability
predict(mod_logf,newdata=new_df1,type="response",se=T) ### Probability
```

```
## $fit
##           1
## 0.009605323
##
## $se.fit
##         1
## 0.0024676
##
## $residual.scale
## [1] 1
```

```r
### Predict the value with its standard error
conf_interval_res <-c(0.009605323-1.96*0.0024676,0.009605323+1.96*0.0024676  )
conf_interval_res
```

```
## [1] 0.004768827 0.014441819
```

```r
predict(mod_logf,newdata=new_df1,type="link",se=T)
```

```
## $fit
##         1
## -4.635786
##
## $se.fit
## [1] 0.2593907
##
## $residual.scale
## [1] 1
```

```r
ilogit(c(-4.635786-1.96*0.2593907,-4.635786+1.96*0.2593907))
```

```
## [1] 0.005799358 0.015869176
```

```r
ilogit(c(-4.635786))
```

```
## [1] 0.009605324
```

```r
new_df2<-(df[268,])
predict(mod_logf,new_df2,type='response')
```

Prediction on 268th bottle

```
##       268
## 0.7527199
```

```
predict(mod_logf,newdata=new_df2,type="link") ### Linear predictor
```

```
##      268
## 1.113171
```

```
#predict(mod_logf,newdata=new.ind,type="response") ### Probability
predict(mod_logf,newdata=new_df2,type="response",se=T) ### Probability
```

```
## $fit
##       268
## 0.7527199
##
## $se.fit
##        268
## 0.03712494
##
## $residual.scale
## [1] 1
```

```
### Predict the value with its standard error
conf_interval_res <-c(0.7527199-1.96*0.03712494,0.7527199+1.96*0.03712494  )
conf_interval_res
```

```
## [1] 0.6799550 0.8254848
```

```
predict(mod_logf,newdata=new_df2,type="link",se=T)
```

```
## $fit
##      268
## 1.113171
##
## $se.fit
## [1] 0.1994542
##
## $residual.scale
## [1] 1
```

```
ilogit(c(1.113171-1.96*0.1994542,1.113171+1.96*0.1994542))
```

```
## [1] 0.6731003 0.8181854
```

Part D. Link Function and Dispersion Parameter

Probit Model

```
mod_pro <- glm(excellent~alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +chlorides + fix
summary(mod_pro)
```

```
##
## Call:
## glm(formula = excellent ~ alcohol + volatile.acidity + sulphates +
##     total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar,
##     family = binomial(link = "probit"), data = df1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8169  -0.4548  -0.2105  -0.0852   3.2546
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -7.318889   0.706628 -10.357  < 2e-16 ***
## alcohol               0.535149   0.046935  11.402  < 2e-16 ***
## volatile.acidity     -1.848170   0.333863  -5.536 3.10e-08 ***
## sulphates             1.868117   0.293831   6.358 2.05e-10 ***
## total.sulfur.dioxide -0.007023   0.001889  -3.718 0.000201 ***
## chlorides            -4.813620   1.796623  -2.679 0.007378 **
## fixed.acidity         0.058708   0.026829   2.188 0.028653 *
## residual.sugar        0.072674   0.034060   2.134 0.032865 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  875.41  on 1591  degrees of freedom
## AIC: 891.41
##
## Number of Fisher Scoring iterations: 7
```

```
predp<- predict(mod_pro, newdata = df1, type="response")
conf_symp <-table(df$excellent, (predp > (0.5)), dnn=c("Truth","Predicted"))
conf_symp
```

Confusion Matrix with cutoff p= 0.5

```
##      Predicted
## Truth FALSE TRUE
##     0  1342   40
##     1   149   68
```

```
Specificity_pro <- conf_symp[1]/(conf_symp[1]+conf_symp[3])
Specificity_pro
```

Specificity (True Negative Rate)

```
## [1] 0.9710564
```

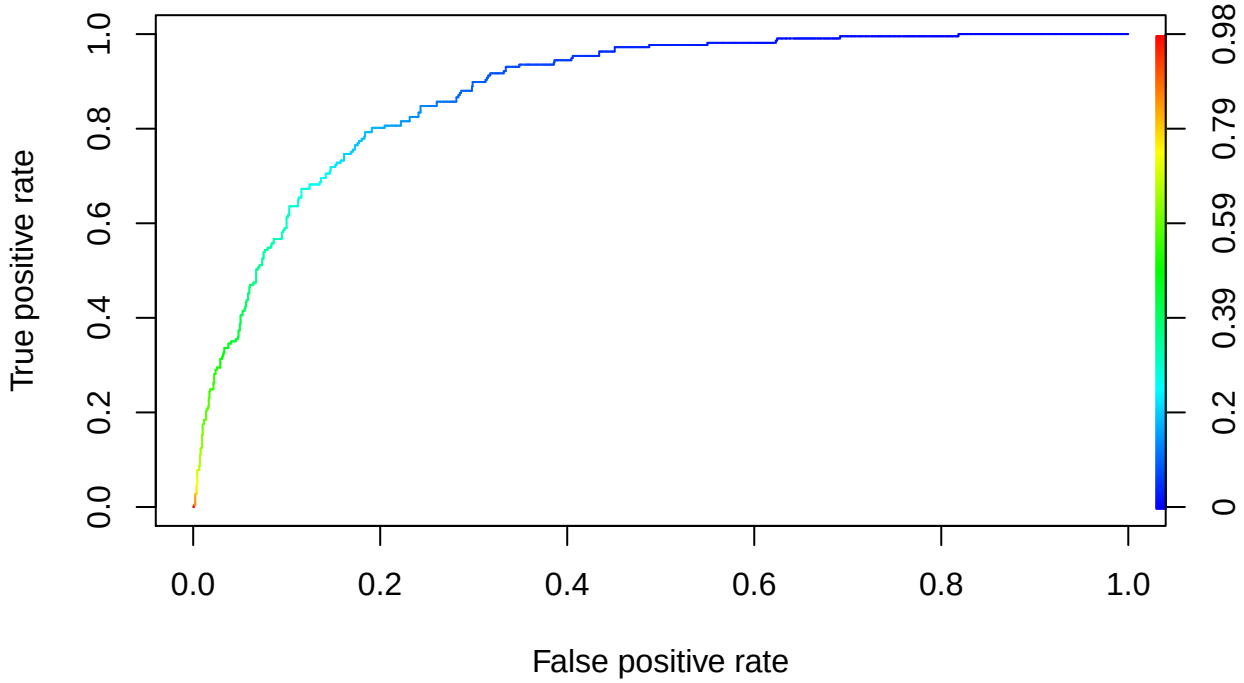
```
Sensitivity_pro <- conf_symp[4]/(conf_symp[2]+conf_symp[4])
Sensitivity_pro
```

Sensitivity (True Positive Rate)

```
## [1] 0.3133641
```

We need to lower the threshold to decrease the number of False Negatives.

```
predic_p <- prediction(predp, df$excellent)
perf_p <- performance(predic_p,"tpr","fpr")
plot(perf_p,colorize=TRUE)
```



ROC Curve
AUC

```
# auc.tmp <- performance(predic,"auc")
# auc <- as.numeric(auc.tmp@y.values)
# auc
unlist(slot(performance(predic_p, "auc"), "y.values"))
```

```
## [1] 0.8795174
```

Complementary Log Log Model

```
mod_clog <- glm(excellent~alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +chlorides + fi
```

```
## Warning: glm.fit: algorithm did not converge
```

```
summary(mod_clog)
```

```
##
## Call:
## glm(formula = excellent ~ alcohol + volatile.acidity + sulphates +
##     total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar,
##     family = binomial(link = "cloglog"), data = df1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2554  -0.4496  -0.2628  -0.1528   2.9399
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -10.105161   1.000262 -10.103  < 2e-16 ***
## alcohol                0.730512   0.064102  11.396  < 2e-16 ***
## volatile.acidity      -3.081715   0.525537  -5.864 4.52e-09 ***
## sulphates              2.837673   0.418751   6.777 1.23e-11 ***
## total.sulfur.dioxide  -0.013364   0.003226  -4.142 3.44e-05 ***
## chlorides             -8.032478   2.943072  -2.729  0.00635 **
## fixed.acidity          0.073086   0.041639   1.755  0.07922 .
## residual.sugar         0.094323   0.057407   1.643  0.10037
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  900.35  on 1591  degrees of freedom
## AIC: 916.35
##
## Number of Fisher Scoring iterations: 25
```

```
predc<- predict(mod_clog, newdata = df1, type="response")
conf_symc <-table(df$excellent, (predc > (0.5)), dnn=c("Truth","Predicted"))
conf_symc
```

Confusion Matrix with cutoff p= 0.5

```
##      Predicted
## Truth FALSE TRUE
##     0  1343   39
##     1   151   66
```

```
Specificity_clog <- conf_symc[1]/(conf_symc[1]+conf_symc[3])
Specificity_clog
```

Specificity (True Negative Rate)

```
## [1] 0.97178
```

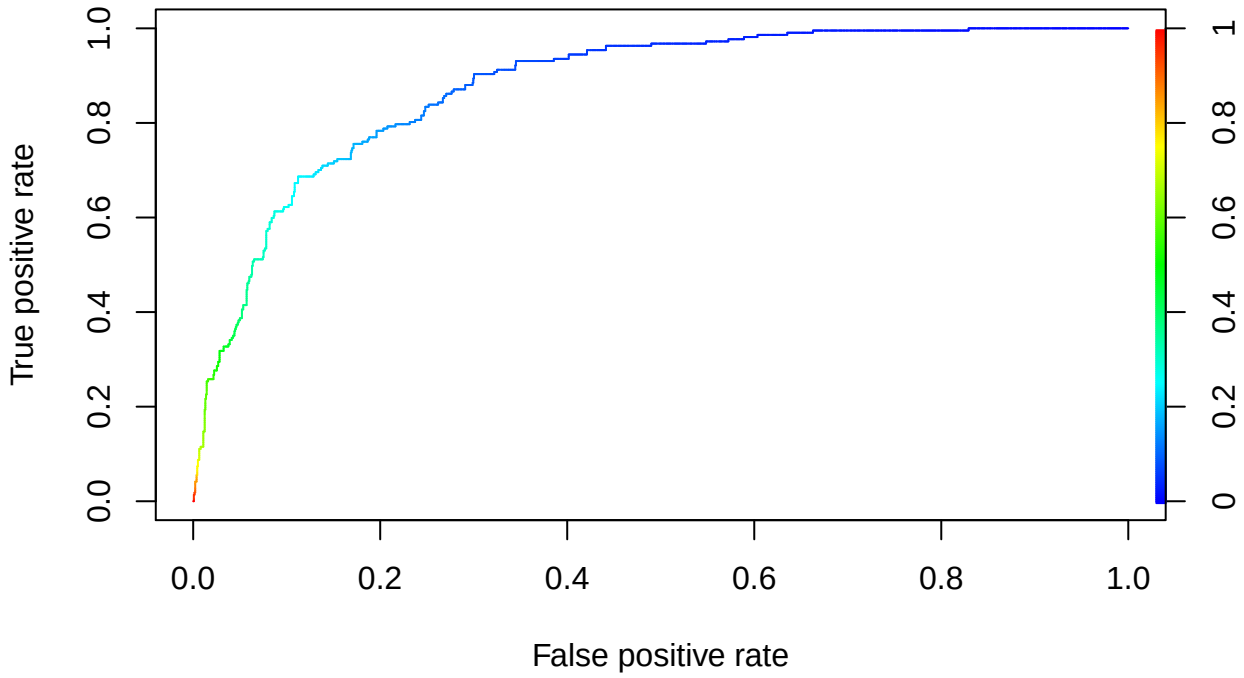
```
Sensitivity_clog <- conf_symc[4]/(conf_symc[2]+conf_symc[4])
Sensitivity_clog
```

Sensitivity (True Positive Rate)

```
## [1] 0.3041475
```

We need to lower the threshold to decrease the number of False Negatives.

```
predic_c <- prediction(predc, df$excellent)
perf_c <- performance(predic_c,"tpr","fpr")
plot(perf_c,colorize=TRUE)
```



ROC Curve
AUC

```
# auc.tmp <- performance(predic,"auc")
# auc <- as.numeric(auc.tmp@y.values)
# auc
unlist(slot(performance(predic_c, "auc"), "y.values"))
```

```
## [1] 0.8773633
```

```
BIC(mod_logf)
```

```
## [1] 938.0058
```

```
BIC(mod_pro)
```

```
## [1] 934.4298
```

```
BIC(mod_clog)
```

```
## [1] 959.3655
```

##Dispersion?

```r
sigma.squared <- sum(residuals(mod_logf,type='pearson')^2)/(nrow(df1)-8)
sigma.squared
```

```
## [1] 0.8277813
```

```r
summary(mod_logf)
```

```
##
## Call:
## glm(formula = excellent ~ alcohol + volatile.acidity + sulphates +
##     total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar,
##     family = binomial, data = df1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7860  -0.4383  -0.2277  -0.1233   3.0696
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -13.205847   1.323132  -9.981  < 2e-16 ***
## alcohol               0.970575   0.087193  11.131  < 2e-16 ***
## volatile.acidity     -3.417247   0.622217  -5.492 3.97e-08 ***
## sulphates             3.424071   0.524620   6.527 6.72e-11 ***
## total.sulfur.dioxide -0.013450   0.003574  -3.763 0.000168 ***
## chlorides            -9.536839   3.531569  -2.700 0.006924 **
## fixed.acidity         0.114692   0.048568   2.361 0.018203 *
## residual.sugar        0.134013   0.061511   2.179 0.029355 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  878.99  on 1591  degrees of freedom
## AIC: 894.99
##
## Number of Fisher Scoring iterations: 6
```

```r
summary(mod_logf,dispersion=sigma.squared)
```

```
##
## Call:
## glm(formula = excellent ~ alcohol + volatile.acidity + sulphates +
##     total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar,
##     family = binomial, data = df1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7860  -0.4383  -0.2277  -0.1233   3.0696
##
## Coefficients:
```

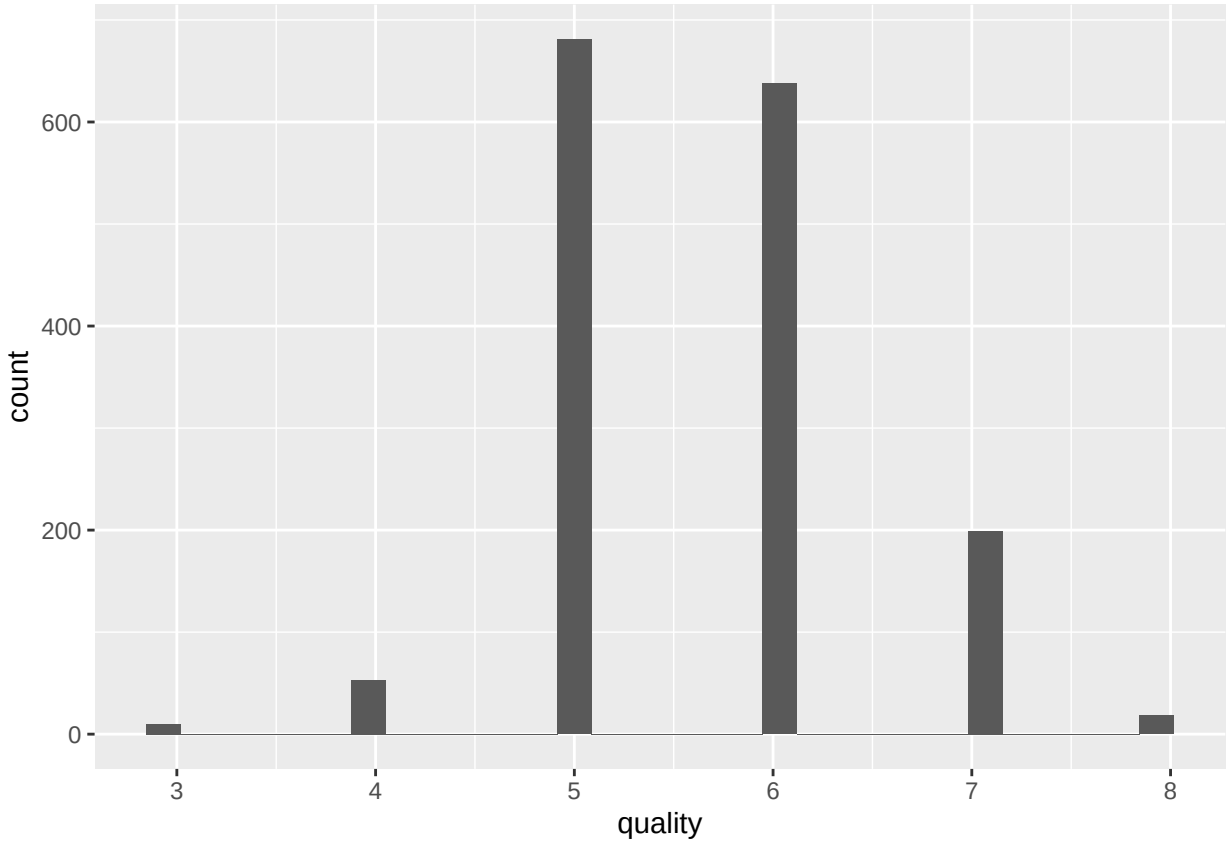
```
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -13.205847   1.203819 -10.970  < 2e-16 ***
## alcohol                0.970575   0.079330  12.235  < 2e-16 ***
## volatile.acidity      -3.417247   0.566108  -6.036 1.58e-09 ***
## sulphates              3.424071   0.477312   7.174 7.30e-13 ***
## total.sulfur.dioxide  -0.013450   0.003252  -4.136 3.53e-05 ***
## chlorides             -9.536839   3.213109  -2.968  0.00300 **
## fixed.acidity          0.114692   0.044189   2.596  0.00945 **
## residual.sugar         0.134013   0.055964   2.395  0.01664 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 0.8277813)
##
##     Null deviance: 1269.92  on 1598  degrees of freedom
## Residual deviance:  878.99  on 1591  degrees of freedom
## AIC: 894.99
##
## Number of Fisher Scoring iterations: 6
```

The Dispersion parameter is less than 1, implying there is no correlation between the wines. The parameter estimates remains same, whereas the Std. Error and p-values have changed. The inference here is, with this new model

Part E. Modeling the wine quality as a multinomial variable with order

```
ggplot(df,aes(quality)) + geom_histogram()
```

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




```
table(df$quality)
```

```
##
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

There are 6 categories, wherein quality 3 and quality 8 are sparse. We can merge quality 3 and 4. Also we can merge quality 7 and 8. Final number of categories = 4. i.e. Category 4, 5, 6 and 7.

```
df2<-df
df2$excellent<-NULL
df2$quality<-as.factor(df2$quality)
levels(df2$quality) <- c("4", "4", "5", "6", "7", "7")
```

```
table(df2$quality)
```

```
##
##   4   5   6   7
##  63 681 638 217
```

Kendal Tau Correlation

```r
df2_cor<-df2
df2_cor$quality<-as.numeric(df2_cor$quality)
kendal <- cor(df2_cor,df2_cor$quality,method='kendall')
kendal1 <- as.data.frame(kendal)
kendal1$V2 <- abs(kendal1$V1)
kendal1[with(kendal1, order(-V2)), ]
```

```
##                              V1         V2
## quality               1.00000000 1.00000000
## alcohol               0.38009700 0.38009700
## volatile.acidity     -0.30171854 0.30171854
## sulphates             0.29969762 0.29969762
## citric.acid           0.16748100 0.16748100
## total.sulfur.dioxide -0.15696617 0.15696617
## chlorides            -0.14847201 0.14847201
## density              -0.13610251 0.13610251
## fixed.acidity         0.08847661 0.08847661
## free.sulfur.dioxide  -0.04544953 0.04544953
## pH                   -0.03382797 0.03382797
## residual.sugar        0.02619190 0.02619190
```

```r
levels(df2$quality)
```

```
## [1] "4" "5" "6" "7"
```

```r
mod_multi_log <- vglm(ordered(quality)~ volatile.acidity + alcohol+sulphates +  total.sulfur.dioxide + c
summary(mod_multi_log)
```

```
##
## Call:
## vglm(formula = ordered(quality) ~ volatile.acidity + alcohol +
##     sulphates + total.sulfur.dioxide + chlorides + citric.acid,
##     family = cumulative(parallel = TRUE), data = df2)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept):1         4.025410   0.669854   6.009 1.86e-09 ***
## (Intercept):2         7.728112   0.667126  11.584  < 2e-16 ***
## (Intercept):3        10.545462   0.702618  15.009  < 2e-16 ***
## volatile.acidity      3.394545   0.369992   9.175  < 2e-16 ***
## alcohol              -0.840836   0.055510 -15.147  < 2e-16 ***
## sulphates            -2.815713   0.344737  -8.168 3.14e-16 ***
## total.sulfur.dioxide  0.008332   0.001613   5.164 2.42e-07 ***
## chlorides             4.887978   1.261429   3.875 0.000107 ***
## citric.acid          -0.033615   0.327197  -0.103 0.918173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]),
## logitlink(P[Y<=3])
##
## Residual deviance: 2923.395 on 4788 degrees of freedom
```

```
##
## Log-likelihood: -1461.697 on 4788 degrees of freedom
##
## Number of Fisher scoring iterations: 6
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##     volatile.acidity              alcohol            sulphates
##          29.80107779           0.43134987           0.05986203
## total.sulfur.dioxide            chlorides          citric.acid
##           1.00836667         132.68497372           0.96694403
```

```
predicted_df <- as.data.frame(predict(mod_multi_log,type='response'))
pred_multi <- colnames(predicted_df)[apply(predicted_df,1,which.max)]
table(df2$quality,as.numeric(pred_multi))
```

Confusion Matrix

```
##
##       4   5   6   7
##   4   0  47  16   0
##   5   2 507 169   3
##   6   0 213 387  38
##   7   0   9 141  67
```

```
predict(mod_multi_log,new_df1,type='response')
```

Prediction on new values

```
##           4         5         6          7
## 1 0.0813388 0.7008397 0.2014508 0.01637067
```

```
predict(mod_multi_log,new_df2,type='response')
```

```
##               4          5        6         7
## 268 0.000677359 0.02607724 0.288303 0.6849424
```

```
a<-predict(mod_multi_log,new_df2,type='link')
a
```

```
##     logitlink(P[Y<=1]) logitlink(P[Y<=2]) logitlink(P[Y<=3])
## 268          -7.296631           -3.59393         -0.7765793
```

```
ilogit(a[1])
```

```
## [1] 0.000677359
```

```
ilogit(a[2])
```

```
## [1] 0.0267546
```

```
ilogit(a[3])
```

```
## [1] 0.3150576
```

```
1-ilogit(a[1])-ilogit(a[2])-ilogit(a[3])
```

```
## [1] 0.6575105
```

Here we can see that we have very close results if the wine is excellent or not using Multinomial and Logistic regression. Please refer the word file for detailed report with explanation.