Appendix A

Introduction

Data summary

```
## 'data.frame':    6497 obs. of  13 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 ...
##  $ color               : Factor w/ 2 levels "red","white": 1 1 1 1 1 1 1 1 1 1 ...
```

```
## [1] "red"   "white"
```

Planning the analysis

Model 1

```
## 
## Call:
## glm(formula = color ~ fixed.acidity + volatile.acidity + citric.acid + 
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide + 
##     density + pH + sulphates + alcohol + quality, family = binomial(), 
##     data = wine_data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -5.6178   0.0012   0.0188   0.0582   6.8678  
## 
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)    
## (Intercept)           1.876e+03  1.868e+02  10.043  < 2e-16 ***
## fixed.acidity         4.005e-01  2.334e-01   1.716   0.0861 .  
## volatile.acidity     -6.722e+00  1.061e+00  -6.337 2.34e-10 ***
## citric.acid           2.617e+00  1.185e+00   2.209   0.0272 *  
## residual.sugar        9.562e-01  1.012e-01   9.449  < 2e-16 ***
## chlorides            -2.201e+01  3.984e+00  -5.524 3.31e-08 ***
## free.sulfur.dioxide  -6.080e-02  1.456e-02  -4.177 2.96e-05 ***
## total.sulfur.dioxide  5.229e-02  4.990e-03  10.479  < 2e-16 ***
## density              -1.875e+03  1.904e+02  -9.846  < 2e-16 ***
## pH                    1.959e+00  1.424e+00   1.376   0.1689    
## sulphates            -2.693e+00  1.249e+00  -2.156   0.0311 *  
## alcohol              -1.792e+00  2.795e-01  -6.412 1.43e-10 ***
## quality              -4.339e-01  2.041e-01  -2.126   0.0335 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7250.98  on 6496  degrees of freedom
## Residual deviance:  424.23  on 6484  degrees of freedom
## AIC: 450.23
##
## Number of Fisher Scoring iterations: 9
```

Model 2

```
##
## Call:
## glm(formula = color ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + sulphates + alcohol + quality, family = binomial(),
##     data = wine_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.6432   0.0011   0.0179   0.0561   6.3901
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.706e+03  1.358e+02  12.558  < 2e-16 ***
## fixed.acidity         1.765e-01  1.698e-01   1.040   0.2985
## volatile.acidity     -7.081e+00  1.031e+00  -6.866 6.60e-12 ***
## citric.acid           2.389e+00  1.169e+00   2.043   0.0411 *
## residual.sugar        8.881e-01  8.728e-02  10.176  < 2e-16 ***
## chlorides            -2.350e+01  3.853e+00  -6.099 1.07e-09 ***
## free.sulfur.dioxide  -5.788e-02  1.452e-02  -3.986 6.73e-05 ***
## total.sulfur.dioxide  5.193e-02  4.958e-03  10.473  < 2e-16 ***
## density              -1.697e+03  1.357e+02 -12.504  < 2e-16 ***
## sulphates            -2.904e+00  1.204e+00  -2.412   0.0159 *
## alcohol              -1.594e+00  2.347e-01  -6.790 1.12e-11 ***
## quality              -4.073e-01  2.007e-01  -2.029   0.0424 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7250.98  on 6496  degrees of freedom
## Residual deviance:  426.15  on 6485  degrees of freedom
## AIC: 450.15
##
## Number of Fisher Scoring iterations: 9
```

Model 3

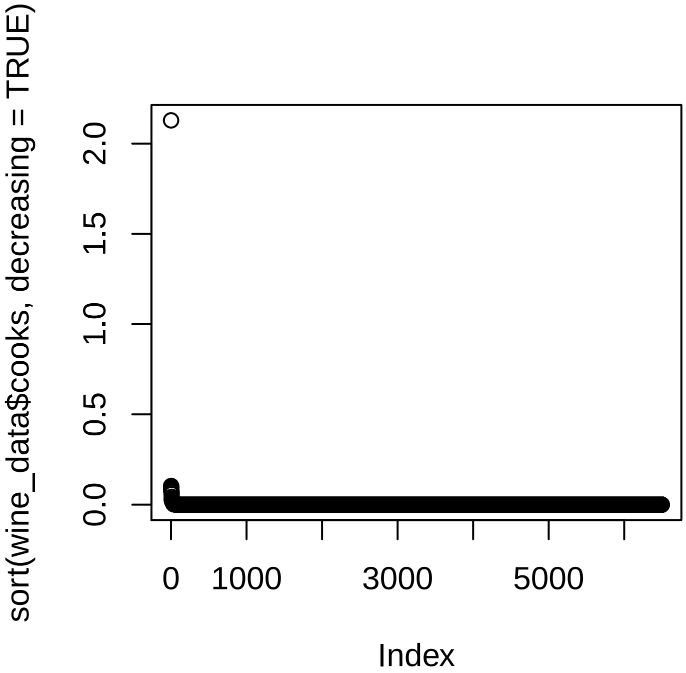
```
##
```

```
## Call:
## glm(formula = color ~ volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + sulphates + alcohol + quality, family = binomial(),
##     data = wine_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.6466   0.0010   0.0173   0.0553   6.1056
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.645e+03  1.211e+02  13.582  < 2e-16 ***
## volatile.acidity     -7.102e+00  1.030e+00  -6.897 5.33e-12 ***
## citric.acid           2.831e+00  1.090e+00   2.598  0.00938 **
## residual.sugar        8.716e-01  8.626e-02  10.104  < 2e-16 ***
## chlorides            -2.438e+01  3.772e+00  -6.464 1.02e-10 ***
## free.sulfur.dioxide  -5.860e-02  1.460e-02  -4.014 5.97e-05 ***
## total.sulfur.dioxide  5.224e-02  4.991e-03  10.467  < 2e-16 ***
## density              -1.636e+03  1.204e+02 -13.585  < 2e-16 ***
## sulphates            -3.056e+00  1.193e+00  -2.562  0.01042 *
## alcohol              -1.560e+00  2.293e-01  -6.804 1.01e-11 ***
## quality              -4.107e-01  1.999e-01  -2.055  0.03988 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7250.98  on 6496  degrees of freedom
## Residual deviance:  427.23  on 6486  degrees of freedom
## AIC: 449.23
##
## Number of Fisher Scoring iterations: 9
```

Model 4

```
##
## Call:
## glm(formula = color ~ volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + sulphates + alcohol + quality, family = binomial(),
##     data = wine_data_noinf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.5609   0.0016   0.0188   0.0494   3.4771
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.962e+03  1.394e+02  14.074  < 2e-16 ***
## volatile.acidity     -6.956e+00  1.093e+00  -6.362 2.00e-10 ***
## citric.acid           2.983e+00  1.216e+00   2.452 0.014190 *
## residual.sugar        7.906e-01  7.110e-02  11.121  < 2e-16 ***
## chlorides            -2.574e+01  3.820e+00  -6.739 1.60e-11 ***
## free.sulfur.dioxide  -5.600e-02  1.551e-02  -3.611 0.000306 ***
## total.sulfur.dioxide  5.493e-02  5.466e-03  10.049  < 2e-16 ***
## density              -1.949e+03  1.384e+02 -14.086  < 2e-16 ***
## sulphates            -2.053e+00  1.172e+00  -1.752 0.079799 .
## alcohol              -2.036e+00  2.641e-01  -7.707 1.29e-14 ***
## quality              -5.199e-01  2.190e-01  -2.373 0.017623 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7250.41  on 6495  degrees of freedom
## Residual deviance:  368.68  on 6485  degrees of freedom
## AIC: 390.68
##
## Number of Fisher Scoring iterations: 9
```

Outliers and Influential points



Interpreting the Model

Confidence Interval of Odds ratio

```
## Waiting for profiling to be done...
```

```
##                          Oddsratio        2.5 %       97.5 %
## (Intercept)                    Inf          Inf          Inf
## volatile.acidity     9.526481e-04 1.036844e-04 7.499641e-03
## citric.acid          1.974942e+01 1.896827e+00 2.235326e+02
## residual.sugar       2.204785e+00 1.932662e+00 2.556088e+00
## chlorides            6.610597e-12 3.868617e-15 1.689055e-08
## free.sulfur.dioxide  9.455379e-01 9.182411e-01 9.752586e-01
## total.sulfur.dioxide 1.056464e+00 1.045753e+00 1.068461e+00
## density              0.000000e+00 0.000000e+00 0.000000e+00
## sulphates            1.283302e-01 1.139180e-02 1.129989e+00
## alcohol              1.306034e-01 7.571276e-02 2.135518e-01
## quality              5.945924e-01 3.854991e-01 9.095391e-01
```

Checking Assumptions of the Model

Multi-collinearity using VIF

```
##     volatile.acidity          citric.acid       residual.sugar 
##             1.519223             1.349094             3.318063 
##            chlorides  free.sulfur.dioxide total.sulfur.dioxide 
##             1.310094             1.917688             2.110683 
##              density            sulphates              alcohol 
##             6.364786             1.253554             4.111340 
##              quality 
##             1.809617
```

Linearity using Logit

Independence erros using Durbin-Watson test

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2037865      1.592426       0
##  Alternative hypothesis: rho != 0
```

Cross Validation of Model

Cross validation

```
## Generalized Linear Model 
## 
## 6496 samples
##   10 predictor
##    2 classes: 'red', 'white' 
## 
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 1 times) 
## Summary of sample sizes: 5847, 5846, 5846, 5847, 5846, 5846, ... 
## Resampling results:
## 
##   Accuracy   Kappa    
##   0.9953818  0.9875347
```

```
## Confusion Matrix and Statistics
## 
```

```
##           Reference
## Prediction  red white
##      red   1582    12
##      white   17  4885
##
##                Accuracy : 0.9955
##                  95% CI : (0.9936, 0.997)
##     No Information Rate : 0.7538
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.988
##
##  Mcnemar's Test P-Value : 0.4576
##
##             Sensitivity : 0.9894
##             Specificity : 0.9975
##          Pos Pred Value : 0.9925
##          Neg Pred Value : 0.9965
##              Prevalence : 0.2462
##          Detection Rate : 0.2435
##    Detection Prevalence : 0.2454
##       Balanced Accuracy : 0.9935
##
##        'Positive' Class : red
##
```