MSCI 718 - Assignment 3

Rishabh Karwayun and Karan Kohli

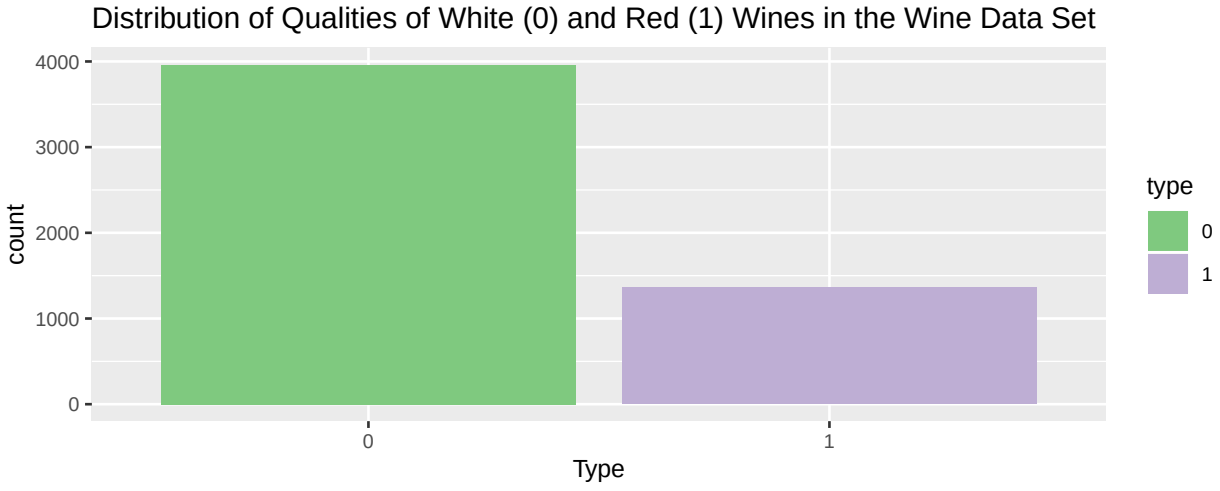
28/03/2020

Introduction

We have two datasets that are related to red and white variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are given (e.g. there is no data about grape types, wine brand, wine selling price, etc.)[1]. We aim to build a model that can predict the color of wine based on its other characteristics.

Data Exploration

Both the datasets were combined to result in a single dataset containing information for both red and white wines. There were some duplicate values in the data which were removed. The classes in the data are ordered but not balanced. For example, there are almost 4000 observations of white wine and close to 1300 observations of red wine as we can see from the barplot below.



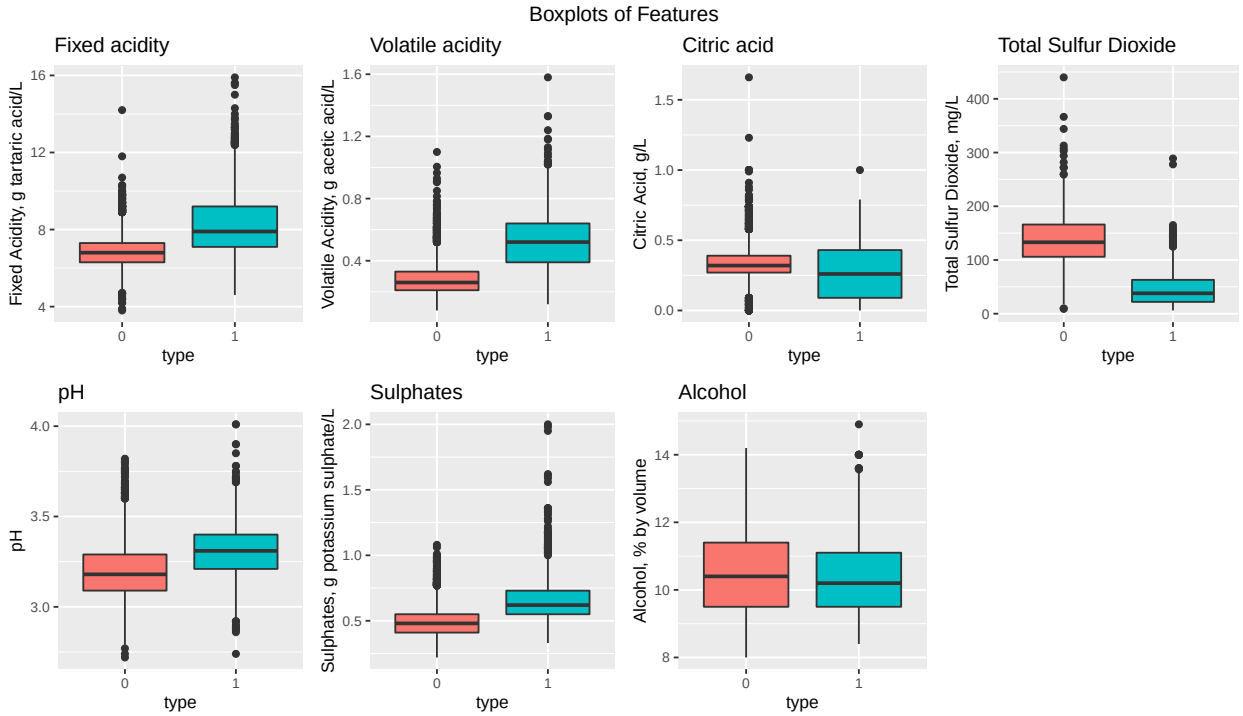
Fig(1)

In total, there are 5320 observations of 13 variables in the dataset. Out of these, 'type' is a binary variable with values as "0" signifying white wine and "1" for red wine whereas 'quality' is an ordinal variable which takes values in the form of natural numbers ranging from 3 to 9 in increasing order of better quality. Both of these are represented as factor variables. Rest of the variables are continuous and are represented using decimal values.

We observe that there are no NA values in the data. From the boxplots (Fig(2) and Appendix 8), we observe that all features apart from 'alcohol' have significant number of outlier values. In order to ensure that our

model is not influenced by outliers, we remove all observations that lie outside of 3 standard deviations from its mean. After removing the outliers, our data now has 4886 obervations.

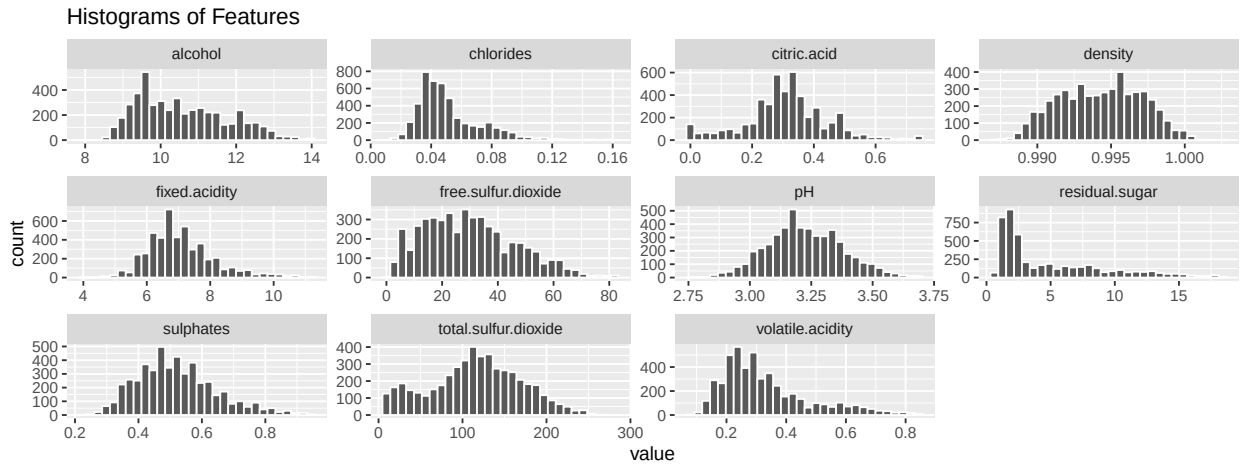
From the summary of data (Appendix 1), we can see that'chlorides' has a mean value of 0.05 with standard deviation as 0.02. pH has a mean value of 3.22 with a standard deviation of 0.01. 'volatile.acidity' has 0.33 and 0.15 as mean and standard deviation respectively.



Fig(2)

From the histograms below (Fig(3)), we see that most of the variables are normally distributed with some negative skewness being observed in the variables like residual.sugar, volatile.acidity and chlorides.

To explore correlations in our data, we plot a **heatmap** (Appendix 4) and find that the highest correlation exists between `chlorides` and `type` (0.54) and `total.sulfur.dioxide` and `type`. `Alcohol` seems to have very little correlation with `type`.



Fig(3)

Assumptions and Building a Model:

In the given data, 'type' and 'quality' can be considered as outcome variables and rest of the features can be considered as predictor variables. It would be interesting to figure if the type of the wine (red or white) can be predicted based on its features. We can do regression to create such model. Since the outcome variable (type) is binary and not continuous, we are using Logistic Regression to create a prediction model. The analysis is interpretable with White Wine (type = '0') as the default case.

We first create an initial model while considering all the features (we do not consider 'quality' to be a feature hence we do not consider it in our analysis). To make our model efficient we eliminate variables that do not contribute significantly to the model. We use the Backward Step method for this. Reason behind choosing this method is that all subset method grows exponentially in complexity based on the number of variables. Since we have a sizeable number of features, we proceed with backward step model. We observe that the following variables: 'sulphates' was found to be non-significant, hence we remove this feature from our consideration. We can look at the model and step results in the Appendix 2.

Now, we move on to check the assumptions of the Logistic Regression Model:

Assumptions:

1) **Multicollinearity**: We inspected the VIF (Variance Inflation Factor) to investigate multicollinearity. The largest value of VIF was 10.384382 for 'density', which is more than 10. Hence, we remove 'density' feature from our model[2] and calculate VIF values again. The largest value of VIF in the new model is 2.21192 for 'total.sulfur.dioxide'. The lowest tolerance (1/VIF) was 0.4520959, which is more than 0.1 (which indicates a serious problem) and 0.2 (which indicates a poential problem). Therefore, we proceed ahead with the assumption that there is no collinearity in the data.
2) **Linearity of Logit**: Based on the test for linearity (Appendix 5), we observe that features 'volatile.acidity', 'chlorides', 'total.sulfuric.dioxide' and 'pH' have p-values less than 0.05. But from the summary of our model (Appendix 3), we observe that these variables have a significant contribution in explaining the variance of the data [3]. This can also be seen from the correlation heatmap (Appendix 4). We understand that is a limitation, but we do not remove these variables and proceed with the assumption of linearity.
3) **Independence Of Errors**: The Durbin-Watson test for independent errors was not significant at the 5% level of significance (d=1.65). As d is close to 2 (which would indicate no autocorrelation detected)[4], and we assume that the data was sampled independently, therefore we do not reject the null hypothesis (that the errors are independent), and continue with the assumption of independence met.
4) **Categorical Outcome**: Our outcome variable (type) is categorical.
5) **Independent Observations**: All duplicate values were removed hence all observations are independent.
6) **Large Sample Size**: There are 4886 observations in the dataset.
7) **Complete Separation**: From the scatter plots in Appendix 7, we can see that there is no complete separation in the data and hence this assumption does not get violated.

Checking for Outliers and Influential Points

We found 23 residuals are above or below 1.96 standard deviations. As this represents approximately 0.5% of the observations, which are expected if the residuals are normal (5% of data is expected to be outside of 2 standard deviations), hence we do not consider any of these observations as outliers and continued with all observations included in the model.

We find that no observation is having a Cook's distance of more than '1'. Hence, we conclude that there are no influential cases in our model.

Model Analysis and Interpretation

Hypothesis

H0: Coefficients of predictor variables in the Logistic Regression model are zero.

Ha: Coefficients of predictor variables in the Logistic Regression model are non-zero.

Based on the calculated co-efficients and their p-values, we reject Null Hypothesis and conclude that our model is significant. The model is summarized in Appendix 3. To interpret the co-efficients, we converted them to odd-ratio using exponentials.

Interpretation:

```
##                         odds_ratio    2.5 %   97.5 %
## (Intercept)          1.000000e-30 -79.9513 -58.5724
## fixed.acidity        7.493793e+00   1.6856   2.3729
## volatile.acidity     1.287699e+05   9.7235  13.9458
## citric.acid          2.013080e-01  -3.9083   0.6521
## residual.sugar       9.214345e-01  -0.2015   0.0246
## chlorides            9.731232e+31  62.2161  86.0002
## free.sulfur.dioxide  1.053951e+00   0.0251   0.0804
## total.sulfur.dioxide 9.412318e-01  -0.0705  -0.0515
## pH                   1.401299e+06  11.9287  16.5818
## alcohol              1.273487e+00  -0.0011   0.4885
```

Using confidence intervals, we can see that the intercept is between -79.96 and -58.57, which does not overlap one. This means there is a significant difference between the odds of wine being White and Red in general, at the 5% level of significance. Also, we can conclude that none of the intervals for features overlap '1', indicating that all the features have some impact on the wine type, at 5% level of significance. Odds ratio is a measure of effect size of the feature on the outcome. Based on the Odds ratio, we observe that 'chlorides' has a significant impact on the outcome.

From the Residuals vs Fitted graph in Appendix 6, we can see that there is a separation based on color but the linearity of model can be seen and the homosedastic nature of the data can be visualised. Also, we observe that most of the points lie on the fitted line which indicates a decent fit of the data.

Conclusion and Future Work:

We conclude that quality of a wine can be significantly predicted by 'chlorides', 'pH' and 'volatile.acidity' along with some impact by rest of the variables that were included in the final model. As no confidence interval has 1 lying in its range, direction of the Odd's Interval can be considered reliable. Our model achieved an AIC value of 470.

A possible future work could be separating the model into train and test sets and finding the prediction accuracy of our model on unseen test data. Also as we observed in the data, the spread of the observations amongst wine types is not symmetrical. If possible, more data should be collected, or sampling could be done so as to have symmetry amongst different wine types which may lead to a more reliable model.

Appendix

References

1: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

2: https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression/

3: http://logisticregressionanalysis.com/758-understanding-logistic-regression-output-part-2-which-variables-matter/

4: https://www.lexjansen.com/wuss/2018/130_Final_Paper_PDF.pdf

Contributions

Both of us explored the dataset and brainstormed together to get to the final choice of model. Then, Karan Kohli worked on the coding part of making the data and getting regression model while Rishabh worked on compiling the work and writing it into the report.

Appendix 1

```
## 'data.frame':    4866 obs. of  13 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.4 7.9 7.3 7.8 7.5 6.7 5.6 8.5 ...
##  $ volatile.acidity    : num  0.7 0.76 0.66 0.6 0.65 0.58 0.5 0.58 0.615 0.28 ...
##  $ citric.acid         : num  0 0.04 0 0.06 0 0.02 0.36 0.08 0 0.56 ...
##  $ residual.sugar      : num  1.9 2.3 1.8 1.6 1.2 2 6.1 1.8 1.6 1.8 ...
##  $ chlorides           : num  0.076 0.092 0.075 0.069 0.065 0.073 0.071 0.097 0.089 0.092 ...
##  $ free.sulfur.dioxide : num  11 15 13 15 15 9 17 15 16 35 ...
##  $ total.sulfur.dioxide: num  34 54 40 59 21 18 102 65 59 103 ...
##  $ density             : num  0.998 0.997 0.998 0.996 0.995 ...
##  $ pH                  : num  3.51 3.26 3.51 3.3 3.39 3.36 3.35 3.28 3.58 3.3 ...
##  $ sulphates           : num  0.56 0.65 0.56 0.46 0.47 0.57 0.8 0.54 0.52 0.75 ...
##  $ alcohol             : num  9.4 9.8 9.4 9.4 10 9.5 10.5 9.2 9.9 10.5 ...
##  $ quality             : Ord.factor w/ 7 levels "3"<"4"<"5"<"6"<..: 3 3 3 3 5 5 3 3 3 5 ...
##  $ type                : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...

##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.900   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2400   1st Qu.: 1.800
##  Median : 6.900   Median :0.2900   Median :0.3100   Median : 2.800
##  Mean   : 7.088   Mean   :0.3306   Mean   :0.3115   Mean   : 5.036
##  3rd Qu.: 7.600   3rd Qu.:0.3950   3rd Qu.:0.3800   3rd Qu.: 7.600
##  Max.   :11.100   Max.   :0.8450   Max.   :0.7400   Max.   :18.400
##
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   : 1.00       Min.   :  6          Min.   :0.9871
##  1st Qu.:0.03700   1st Qu.:17.00       1st Qu.: 81          1st Qu.:0.9920
```

```
##  Median :0.04600   Median :29.00       Median :118          Median :0.9944
##  Mean   :0.05172   Mean   :30.14       Mean   :116          Mean   :0.9943
##  3rd Qu.:0.06000   3rd Qu.:41.00       3rd Qu.:154          3rd Qu.:0.9965
##  Max.   :0.16000   Max.   :83.00       Max.   :278          Max.   :1.0026
##
##        pH          sulphates         alcohol      quality  type
##  Min.   :2.770   Min.   :0.2200   Min.   : 8.00   3:  15   0:3787
##  1st Qu.:3.120   1st Qu.:0.4300   1st Qu.: 9.50   4: 176   1:1079
##  Median :3.220   Median :0.5000   Median :10.40   5:1577
##  Mean   :3.225   Mean   :0.5196   Mean   :10.57   6:2145
##  3rd Qu.:3.330   3rd Qu.:0.5900   3rd Qu.:11.40   7: 808
##  Max.   :3.700   Max.   :0.9400   Max.   :14.05   8: 140
##                                                   9:   5
```

```
##                      min   max range median  mean std.dev
## fixed.acidity        3.9  11.1   7.2    6.9   7.1     1.1
## volatile.acidity     0.1   0.8   0.8    0.3   0.3     0.1
## citric.acid          0.0   0.7   0.7    0.3   0.3     0.1
## residual.sugar       0.6  18.4  17.8    2.8   5.0     4.3
## chlorides            0.0   0.2   0.2    0.0   0.1     0.0
## free.sulfur.dioxide  1.0  83.0  82.0   29.0  30.1    16.3
## total.sulfur.dioxide 6.0 278.0 272.0  118.0 116.0    54.4
## density              1.0   1.0   0.0    1.0   1.0     0.0
## pH                   2.8   3.7   0.9    3.2   3.2     0.2
## sulphates            0.2   0.9   0.7    0.5   0.5     0.1
## alcohol              8.0  14.1   6.1   10.4  10.6     1.2
```

```
## [1] "NA Values:"
```

```
## [1] 0
```

Appendix 2

```
## Start:  AIC=255.84
## type ~ (fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol + quality) - quality
##
##                       Df Deviance    AIC
## - sulphates            1   232.79 254.79
## <none>                     231.84 255.84
## - pH                   1   235.13 257.13
## - citric.acid          1   235.96 257.96
```

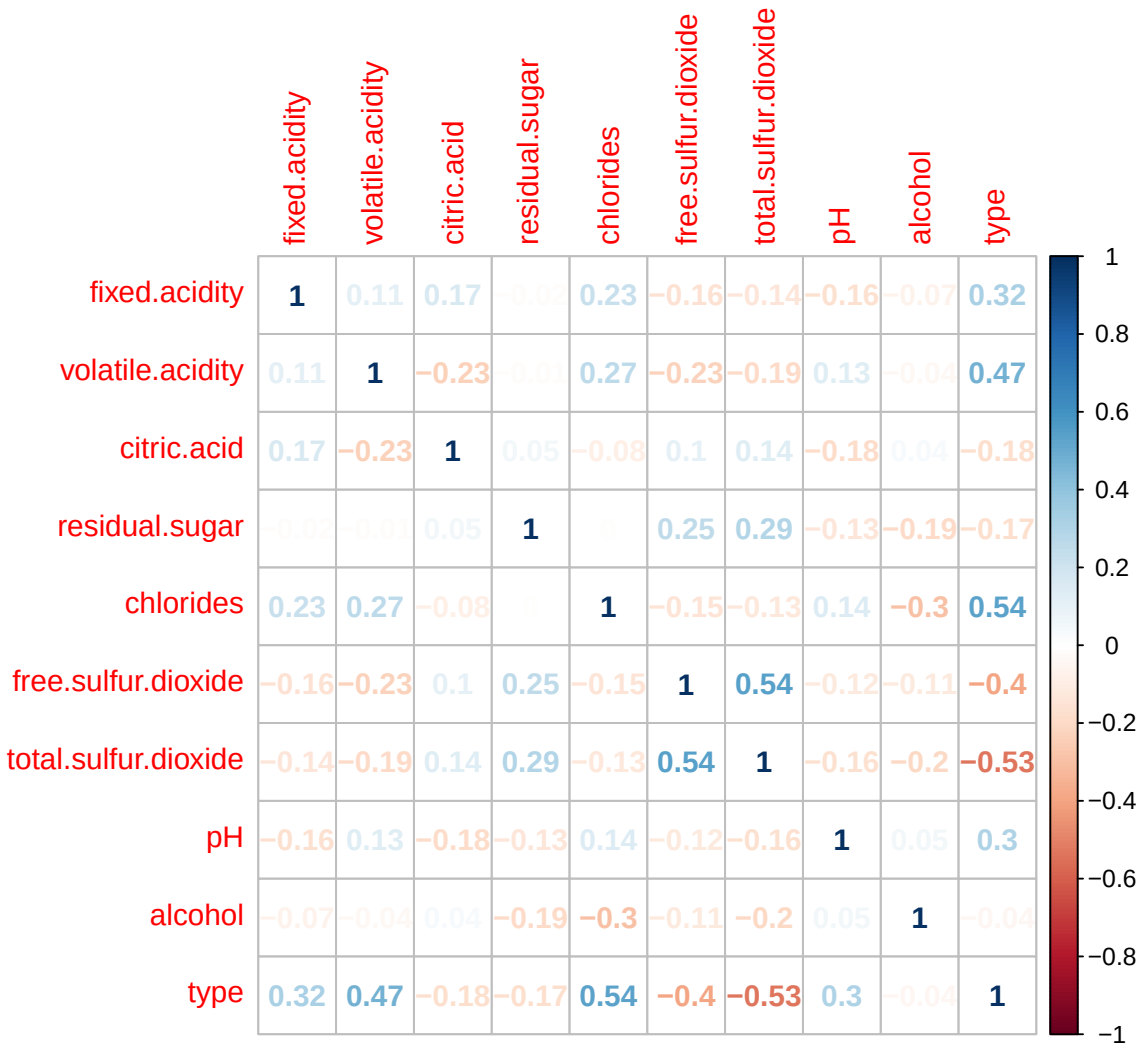
```
## - fixed.acidity         1   239.19 261.19
## - volatile.acidity      1   245.93 267.93
## - free.sulfur.dioxide   1   248.46 270.46
## - chlorides             1   259.33 281.33
## - alcohol               1   322.30 344.30
## - residual.sugar        1   334.58 356.58
## - total.sulfur.dioxide  1   359.88 381.88
## - density               1   385.89 407.89
##
## Step:  AIC=254.79
## type ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + alcohol
##
##                        Df Deviance    AIC
## <none>                      232.79 254.79
## - pH                    1   236.10 256.10
## - citric.acid           1   237.19 257.19
## - fixed.acidity         1   241.01 261.01
## - volatile.acidity      1   246.37 266.37
## - free.sulfur.dioxide   1   248.82 268.82
## - chlorides             1   260.93 280.93
## - total.sulfur.dioxide  1   360.02 380.02
## - residual.sugar        1   368.29 388.29
## - alcohol               1   381.07 401.07
## - density               1   450.00 470.00

##
## Call:  glm(formula = type ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + alcohol, family = binomial(link = "logit"),
##     data = wines)
##
## Coefficients:
##          (Intercept)         fixed.acidity      volatile.acidity
##           -2.783e+03            -9.235e-01             5.349e+00
##          citric.acid        residual.sugar             chlorides
##           -3.773e+00            -1.037e+00             4.630e+01
##  free.sulfur.dioxide  total.sulfur.dioxide               density
##            8.448e-02            -5.607e-02             2.784e+03
##                    pH               alcohol
##           -3.602e+00             3.301e+00
##
## Degrees of Freedom: 4865 Total (i.e. Null);  4855 Residual
```

```
## Null Deviance:       5149
## Residual Deviance: 232.8     AIC: 254.8
```

Appendix 3

```
##
## Call:
## glm(formula = type ~ . - quality - sulphates - density, family = binomial(link = "logit"),
##     data = wines)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4880  -0.0499  -0.0143  -0.0017   5.7753
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -68.738458   5.440540 -12.634  < 2e-16 ***
## fixed.acidity          2.014075   0.174952  11.512  < 2e-16 ***
## volatile.acidity      11.765782   1.074406  10.951  < 2e-16 ***
## citric.acid           -1.602919   1.160471  -1.381  0.16720
## residual.sugar        -0.081824   0.057552  -1.422  0.15510
## chlorides             73.655478   6.051462  12.172  < 2e-16 ***
## free.sulfur.dioxide    0.052546   0.014081   3.732  0.00019 ***
## total.sulfur.dioxide  -0.060566   0.004849 -12.490  < 2e-16 ***
## pH                    14.152910   1.183972  11.954  < 2e-16 ***
## alcohol                0.241759   0.124646   1.940  0.05243 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5149.2  on 4865  degrees of freedom
## Residual deviance:  450.0  on 4856  degrees of freedom
## AIC: 470
##
## Number of Fisher Scoring iterations: 9
```

Appendix 4

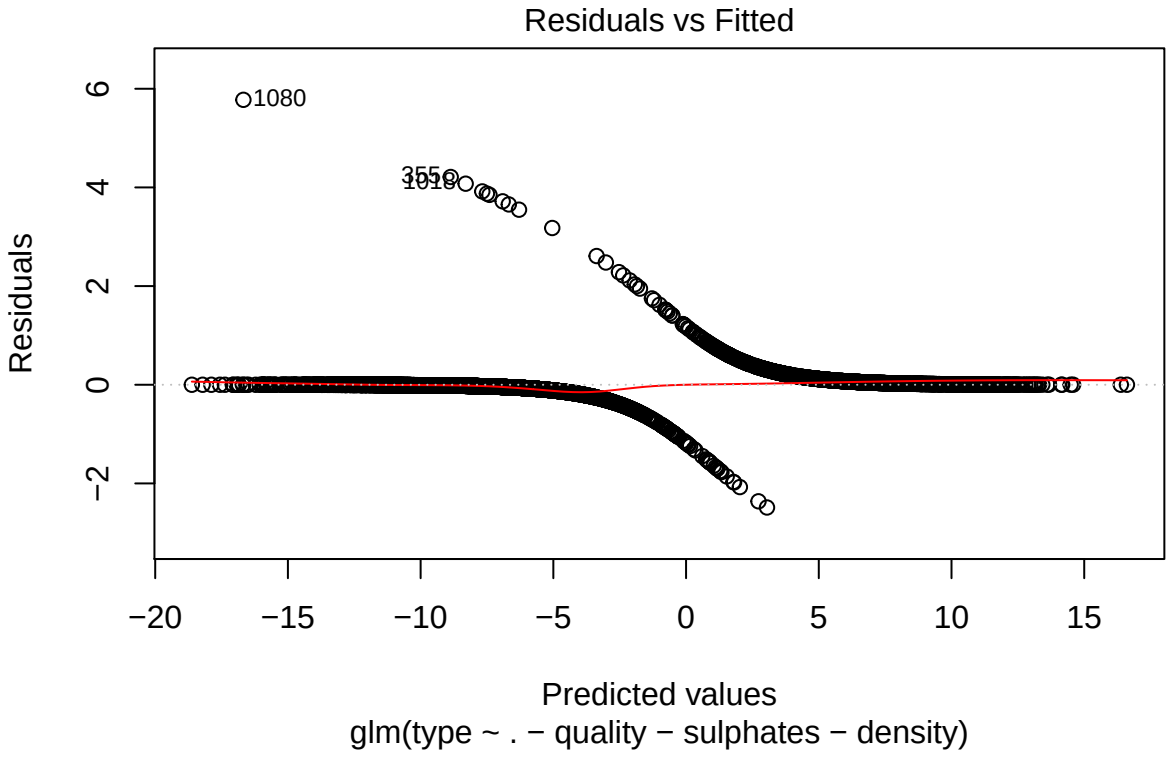


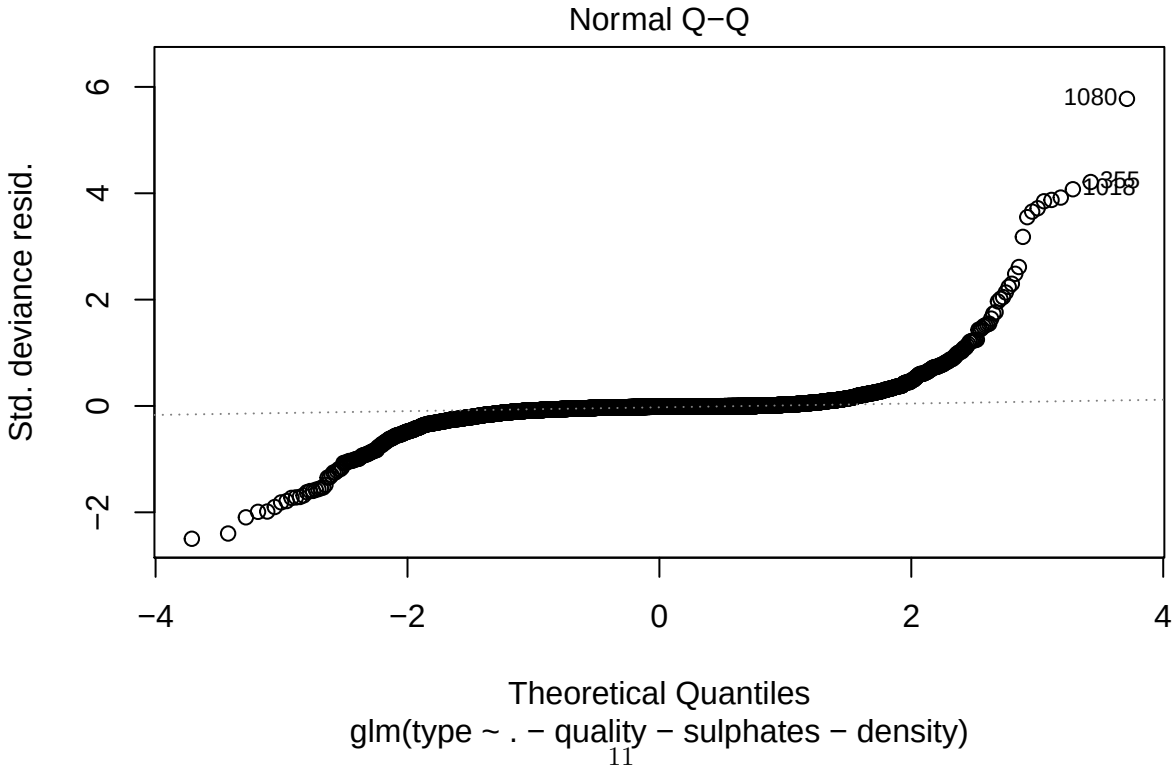
Appendix 5

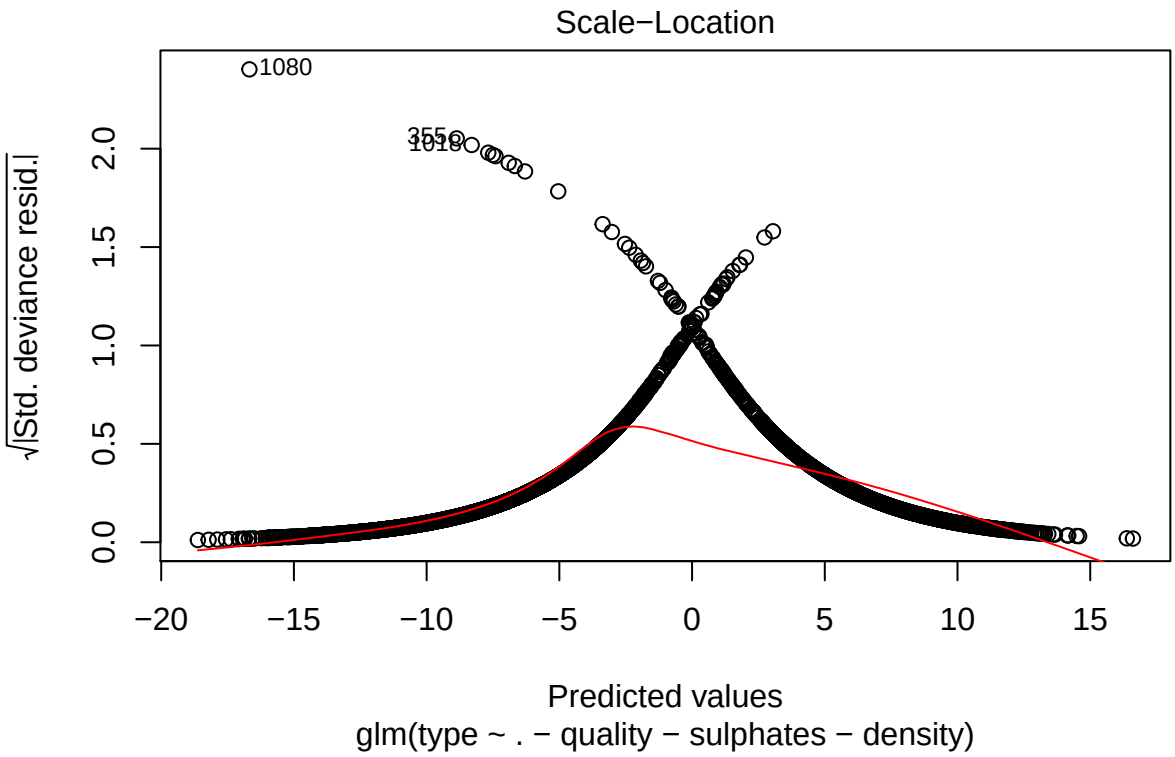
```
##
## Call:
## glm(formula = type ~ . - quality - sulphates - density, family = binomial(link = "logit"),
##     data = wines_test)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

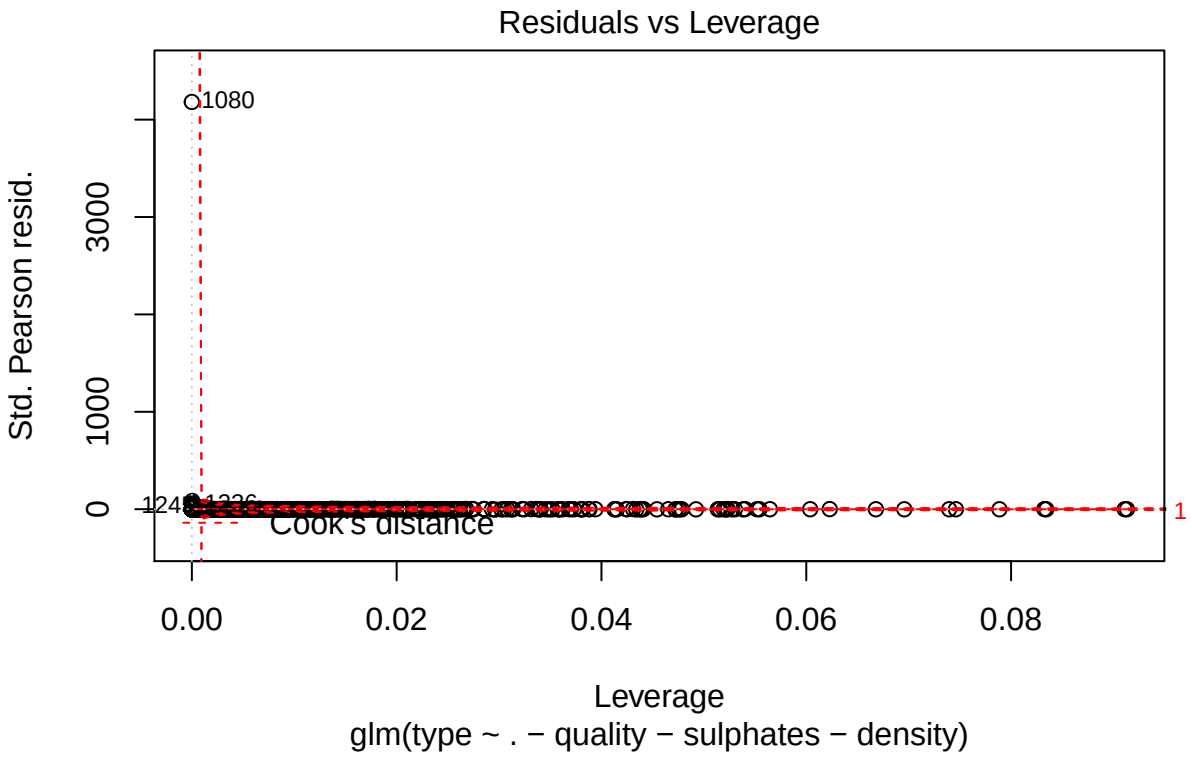
```
## -2.2718  -0.0333  -0.0093  -0.0011   5.7872
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -352.53749  148.92274  -2.367 0.017921 *
## fixed.acidity          -8.42958    5.04271  -1.672 0.094596 .
## volatile.acidity       12.77801    1.63535   7.814 5.56e-15 ***
## citric.acid             1.18712    1.55043   0.766 0.443871
## residual.sugar         -0.00275    0.53842  -0.005 0.995925
## chlorides            -365.12376   54.47030  -6.703 2.04e-11 ***
## free.sulfur.dioxide     0.13932    0.19503   0.714 0.475002
## total.sulfur.dioxide   -0.32025    0.08380  -3.821 0.000133 ***
## pH                    207.89707   99.18334   2.096 0.036074 *
## alcohol                 1.76324    8.41572   0.210 0.834045
## logFixedAcidity         3.21704    1.69642   1.896 0.057910 .
## logVolatileAcidity    -14.72944    5.24721  -2.807 0.004999 **
## logCitricAcid           3.35945    3.22122   1.043 0.296989
## logResdiualSugar       -0.05620    0.20348  -0.276 0.782390
## logChlorides         -282.58254   35.48685  -7.963 1.68e-15 ***
## logFreeSO2              -0.01581    0.04477  -0.353 0.723992
## logTotalSO2              0.04649    0.01478   3.146 0.001656 **
## logpH                  -89.97762   45.18135  -1.991 0.046428 *
## logAlcohol              -0.39218    2.47001  -0.159 0.873844
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4851.59  on 4754  degrees of freedom
## Residual deviance:  306.35  on 4736  degrees of freedom
##   (111 observations deleted due to missingness)
## AIC: 344.35
##
## Number of Fisher Scoring iterations: 10
```

Appendix 6

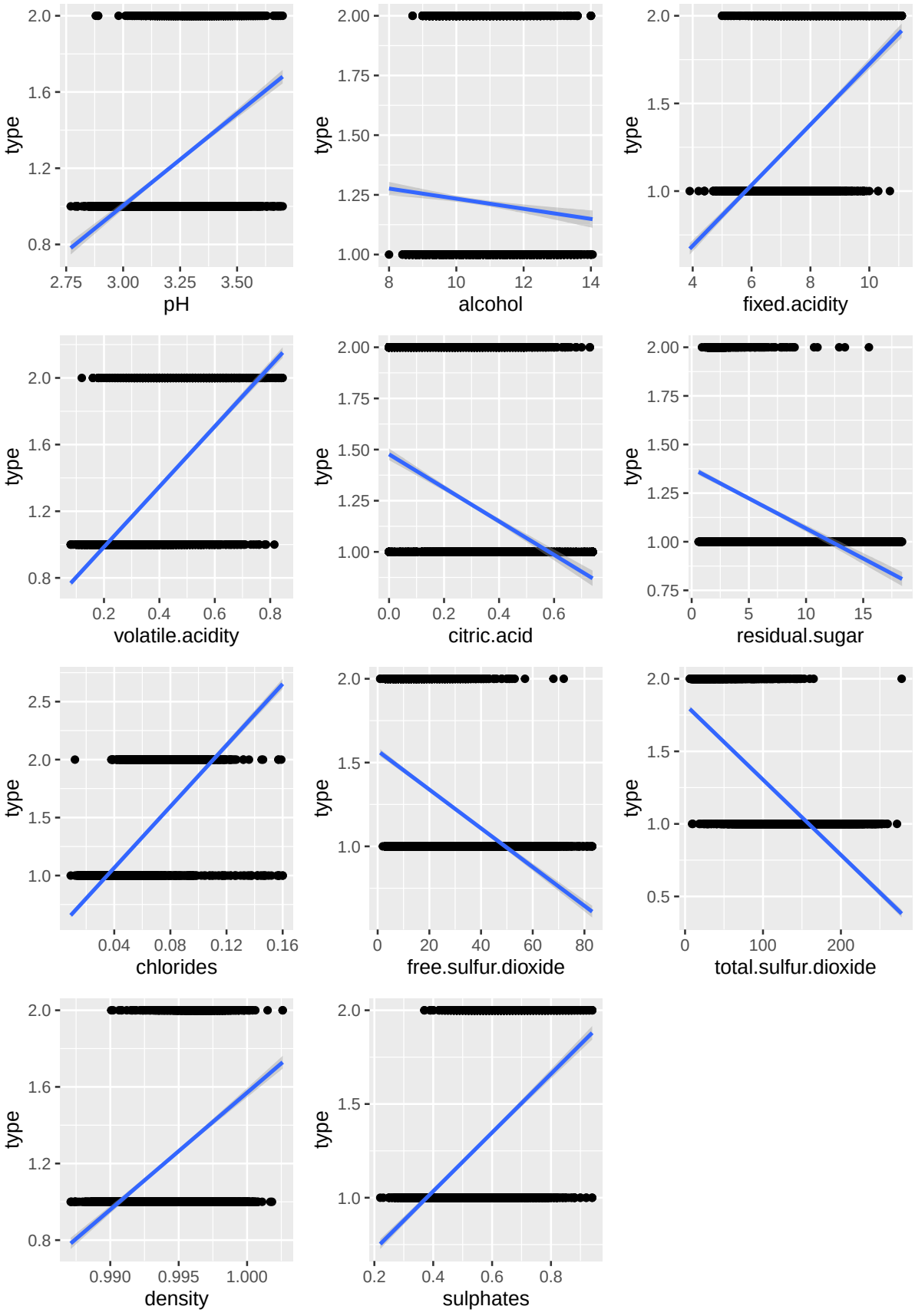

Residuals vs Fitted
Residuals
Predicted values
glm(type ~ . − quality − sulphates − density)
1080
355
1018


Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(type ~ . − quality − sulphates − density)
1080
355
1018

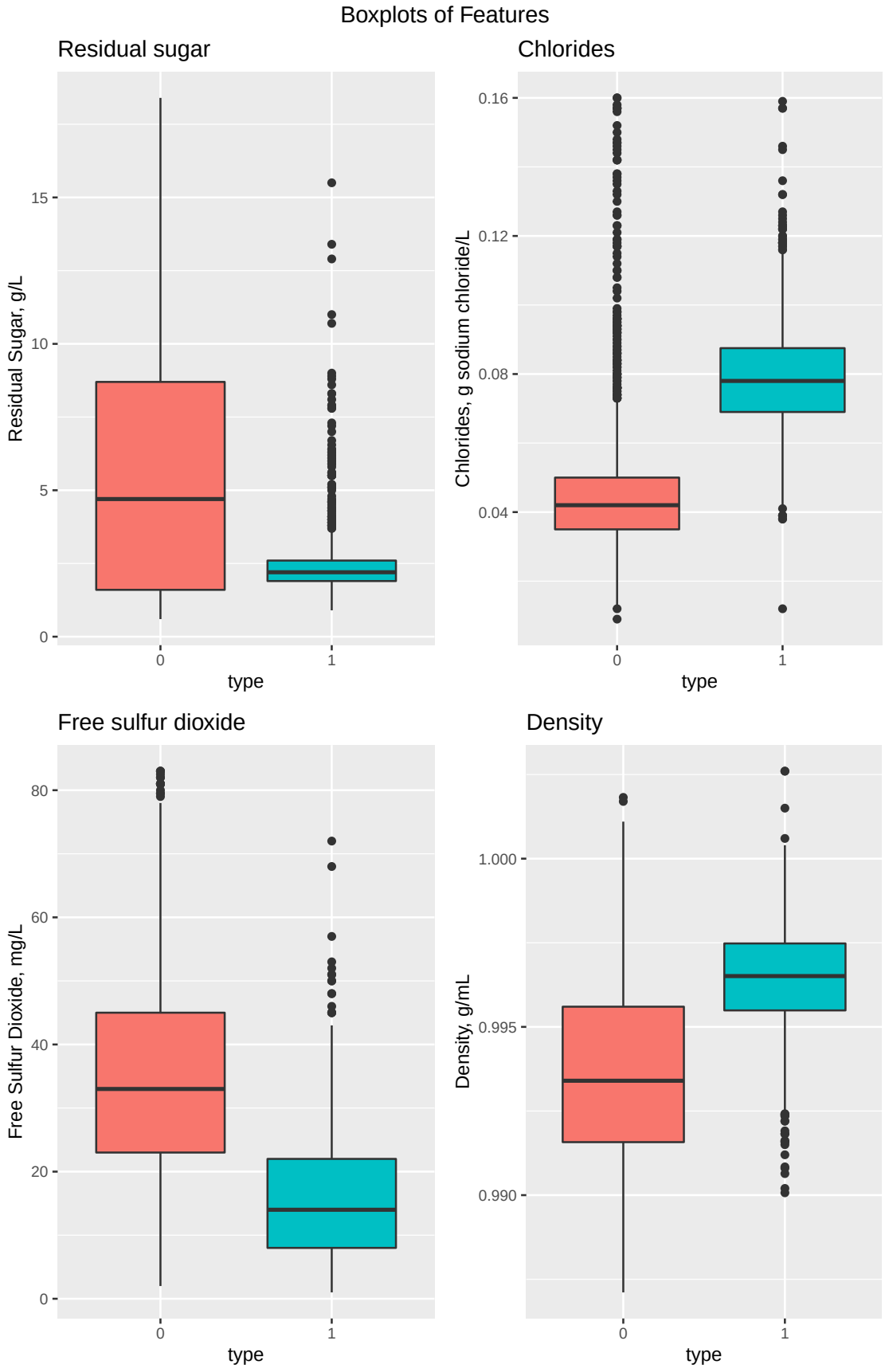

Scale−Location
1080
355
1018
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
2.0
−20
−15
−10
−5
0
5
10
15
Predicted values
glm(type ~ . − quality − sulphates − density)


Residuals vs Leverage
1080
Std. Pearson resid.
3000
1000
0
Cook's distance
1
0.00
0.02
0.04
0.06
0.08
Leverage
glm(type ~ . − quality − sulphates − density)

Appendix 7


type
pH
alcohol
fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
sulphates

Appendix 8


Boxplots of Features
Residual sugar
Residual Sugar, g/L
0
5
10
15
0
1
type
Chlorides
Chlorides, g sodium chloride/L
0.04
0.08
0.12
0.16
0
1
type
Free sulfur dioxide
Free Sulfur Dioxide, mg/L
0
20
40
60
80
0
1
type
Density
Density, g/mL
0.990
0.995
1.000
0
1
type