Assignment 3

Shahzeb Naveed (20789222) | Zaryab Javaid (20852202) | Muhammad Mohsin Tahir (20812155)

Introduction

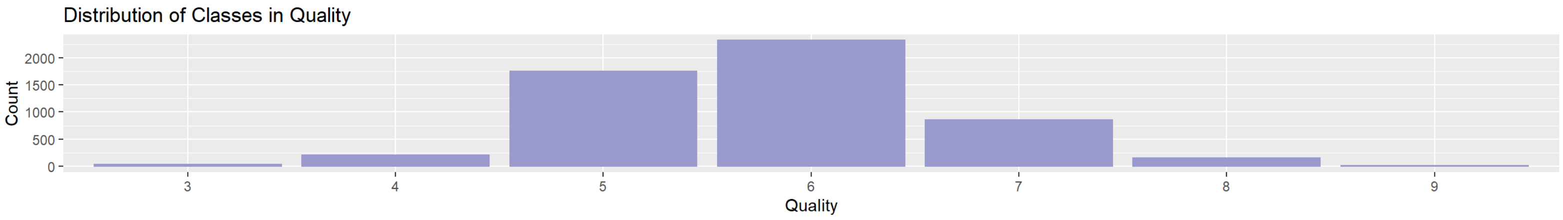
For the love of wine, and data science,
we attempt to explore, what makes it fine.

Objective

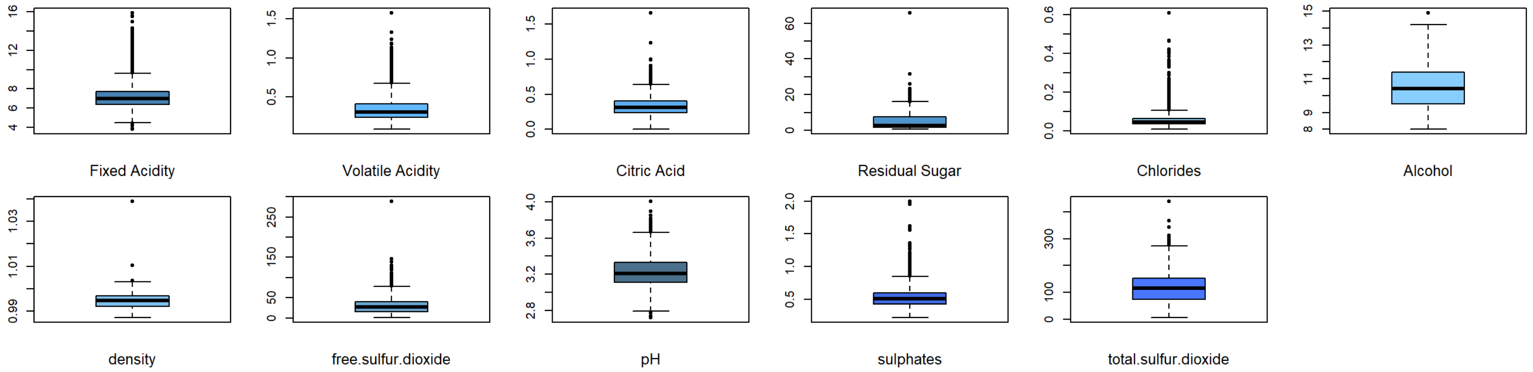
After analyzing several physiochemical properties, we aim to build a model that can predict quality of a wine based on its constituents. The two datasets employed contain information on red and white variants of the Portuguese “Vinho Verde” wine. Initially, the combined dataset had `6497` rows with all columns as integers apart from `color` which was coded as factor. For a reason to be discussed later, duplicate rows were excluded from our analysis leaving beind `5320` rows. With this aim in mind, we test the **hypothesis** that whether the chosen properties significantly determine wine’s quality.

Data Cleaning and Exploration

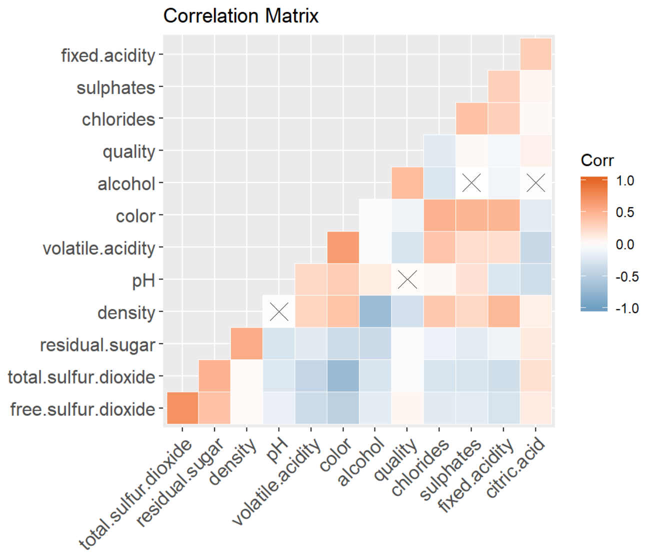
To quickly begin with the EDA process, we explore the class distribution of our outcome variable `quality` and find that the classes are imbalanced (unfortunately). Furthermore, the dataset has `0` NAs (fortunately).



To dig deeper and explore the spread of variables, we plot boxplots and find that apart from `alcohol` almost all variables have quite a large number of outliers. This, along with the fact that our classes are imbalanced, indicates the possibility that some ingredients are found in excessive quantity in low-quality and some in high-quality wines. Anyways, to ensure our model is not influenced by outliers, we removed all the outliers that had a less than 1% chance of occurring as a non-outlier.



To explore correlations in our data, we plot a **heatmap** and find that the highest postive correlation exists between `alcohol` and `quality`. `Citric.acid` and some other variables have apparently no correlation with `quality`. We ’ll keep this in mind while developing our model.



Model Building

As `quality` is an integer, we treat it as an ordinal categorical variable and employ **ordinal logistic regression** to build a prediction model. We will use `polr()` as it fits a logistic or probit regression model to an ordered factor response. But before diving into modelling, we first take a look at some of the underlying assumptions below.

Assumptions of Logistic Regression

1. **Categorical Predictor:** Quality is ordinal so this assumption holds.
2. **Large sample size:** We have `5320` rows so we are good to go!
3. **Independent Observations:** We have independent observations with all duplicates removed.
4. **No or Less Multicollinearity:** We'll come to that in a while.
5. **Linearity of predictors and logit of outcome variable:** We'll come to that in a while as well.
6. **Complete Information:** As mentioned earlier, we have `0` NAs.
7. **Incomplete Separation:** We can clearly see that there is no complete separation in the scatterplots. `Appendix (i)`

To check Multicollinearity, we calculate **Variance Inflation Factors** (VIF) `Appendix (ii)`. Based on this, we exclude 'density' as it has a VIF = 10.17 > 10 (a well-established no-go area). We removed color as well because it is really not a "determinant" of quality.

We now move on to testing the last assumption: Log-Linearity! To do this, we regress our model with log(predictor)*predictor interaction terms. When we do this `Appendix (iii)`, all the interaction variables except those of `free.sulphur.dioxide` and `pH`, are found non-signficant indicating that the assumption of linearity is met for all the other variables. ~~We have not yet discarded these two variables.~~

Feature Selection

For selecting predictors, we use **Backwards Step-wise Logistic Regression** using the `step()` function. For evaluation, we use **Akaike Information Criteria** that depends on model deviance (which is twice the **log-likelihood**) and number of predictor variables employed (thus penalizing the increase in number of predictors). Bottom-line: the smaller the AIC, the better the model.

Using this method, we find that by removing `citric.acid` from the model, we can achieve a lower AIC (`11716` in this case). `Appendix (iv)`

Evaluating Significance

Ho: We observe that p-value for `fixed.acidity` is greater than `0.05 Appendix (v)`. So, it is statistically insignifcant but when we tried removing it, it slightly increases the AIC (not good). To get a third opinion on this, we make another model with `fixed.acidity` removed and then apply **ANOVA** to see there's any improvement:

```
##   Resid. df Resid. Dev   Test    Df LR stat. Pr(Chi)
## 1      5308   11692.53                              
## 2      5308   11692.53 1 vs 2     0        0       1
```

By ANOVA, we get to know that removing `fixed.acidity` doesn't make a statistical difference. So, we keep the new model that requires less predictors.

Hypothesis Testing

Ho: Co-efficients of selected predictors are zero

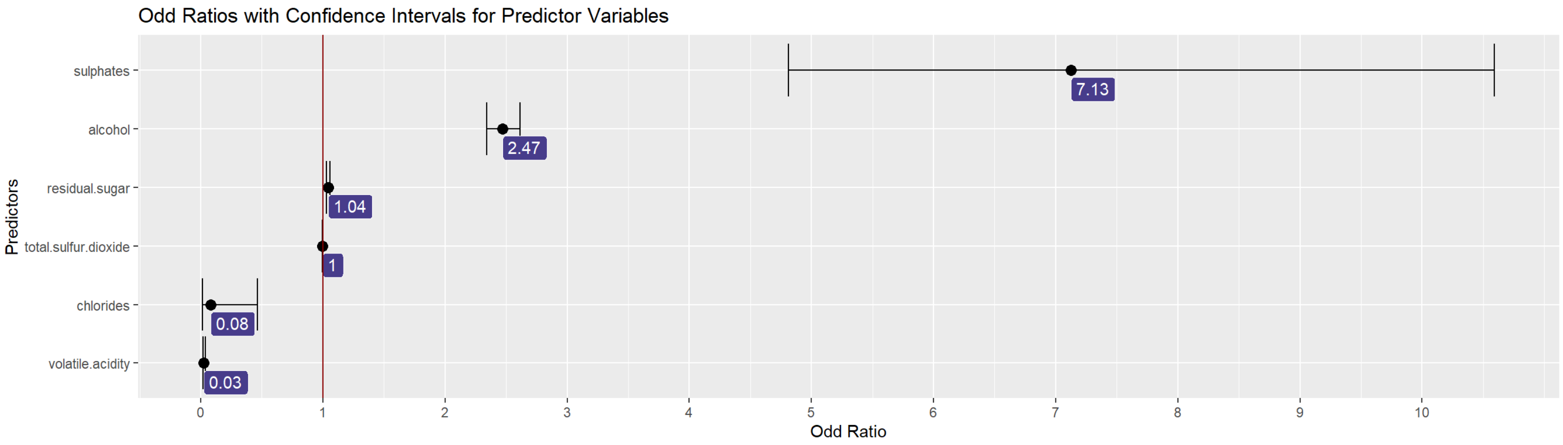
Ha: Co-efficients of selected predictors are non-zero

Based on the calculated co-efficients and their p-values, we reject Null Hypothesis and conclude that our model is significant. Our final model gives a **Hosmer and Lemeshow's R^2 for Goodness-of-Fit** of `0.14` and is summarised below.

```
## Coefficients:
##                         Value Std. Error t value
## volatile.acidity     -3.68059  0.1920681 -19.163
## residual.sugar        0.04394  0.0069213   6.349
## chlorides            -2.50068  0.8868677  -2.820
## total.sulfur.dioxide -0.00262  0.0006063  -4.322
## sulphates             1.96379  0.2012844   9.756
## alcohol               0.90526  0.0285225  31.738
## 
## Intercepts:
##     Value    Std. Error t value
## 3|4   3.1930   0.4002     7.9793
## 4|5   5.3482   0.3619    14.7778
## 5|6   8.3780   0.3628    23.0921
## 6|7  11.0028   0.3797    28.9798
## 7|8  13.3455   0.3960    33.6996
## 8|9  16.8420   0.5930    28.4002
## 
## Residual Deviance: 11692.53 
## AIC: 11716.53
```

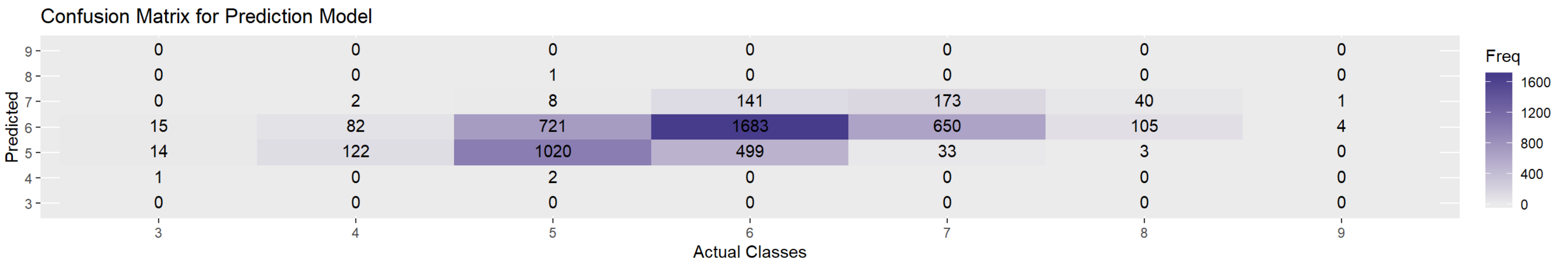
Model Interpretation and Insights

To interpret the co-efficients, we converted them to **odd-ratio** using exponentials, which are plotted below. For example, we would say that by keeping all other variables constant, when `residual.sugar` increases one unit, it is `1.04` times more likely to be in a higher category of `quality`. Furthermore, using confidence intervals, we can conclude that none of the intervals cross `1`, indicating that the direction of odd-ratios of all co-efficients is reliable. Thus, `sulphates` and `alcohol` have a high-positive impact on `quality` and, `chlorides` and `volatile.acidity` have a high-negative impact.



Testing Prediction Accuracy

For evaluating in-sample accuracy of our prediction model, we form a **confusion matrix** by `predict()` -ing using our original dataset. **Residual Plot** was also visualized and homoscedasticity was observed `Appendix (vi)` .



Predicted \ Actual	3	4	5	6	7	8	9
9	0	0	0	0	0	0	0
8	0	0	1	0	0	0	0
7	0	2	8	141	173	40	1
6	15	82	721	1683	650	105	4
5	14	122	1020	499	33	3	0
4	1	0	2	0	0	0	0
3	0	0	0	0	0	0	0

We can observe from above that accuracy for our model is `54.06` %. The low accuracy can be well-understood from the confusion matrix that classes with low frequency `(3,4,8,9)` were rarely predicted. This was due to the imbalance in the distribution of our classes.

Gap Analysis and Future Work

Class Imbalance: As already exhibited, the imbalance in our classes had a very negative impact on the accuracy of our model. Methods such as clustering or re-sampling may be be evaluated for improved accuracy (other than collecting more data for low-quality and high-quality wines, of course).

Conclusion

We conclude that quality of a wine can be significantly predicted by `sulphates` , `alohol` , `chlorides` , `volatile.acidity` and along with some impact by other variables as well that were included in our final model. The direction of odd-ratios is reliable because none of the confidence intervals cross `1` . The overall fit of the model as determined by Hosmer and Lemeshow’s R^2 for Goodness-of-Fit is `0.14` . The prediction accuracy is found to be `54.06` %.

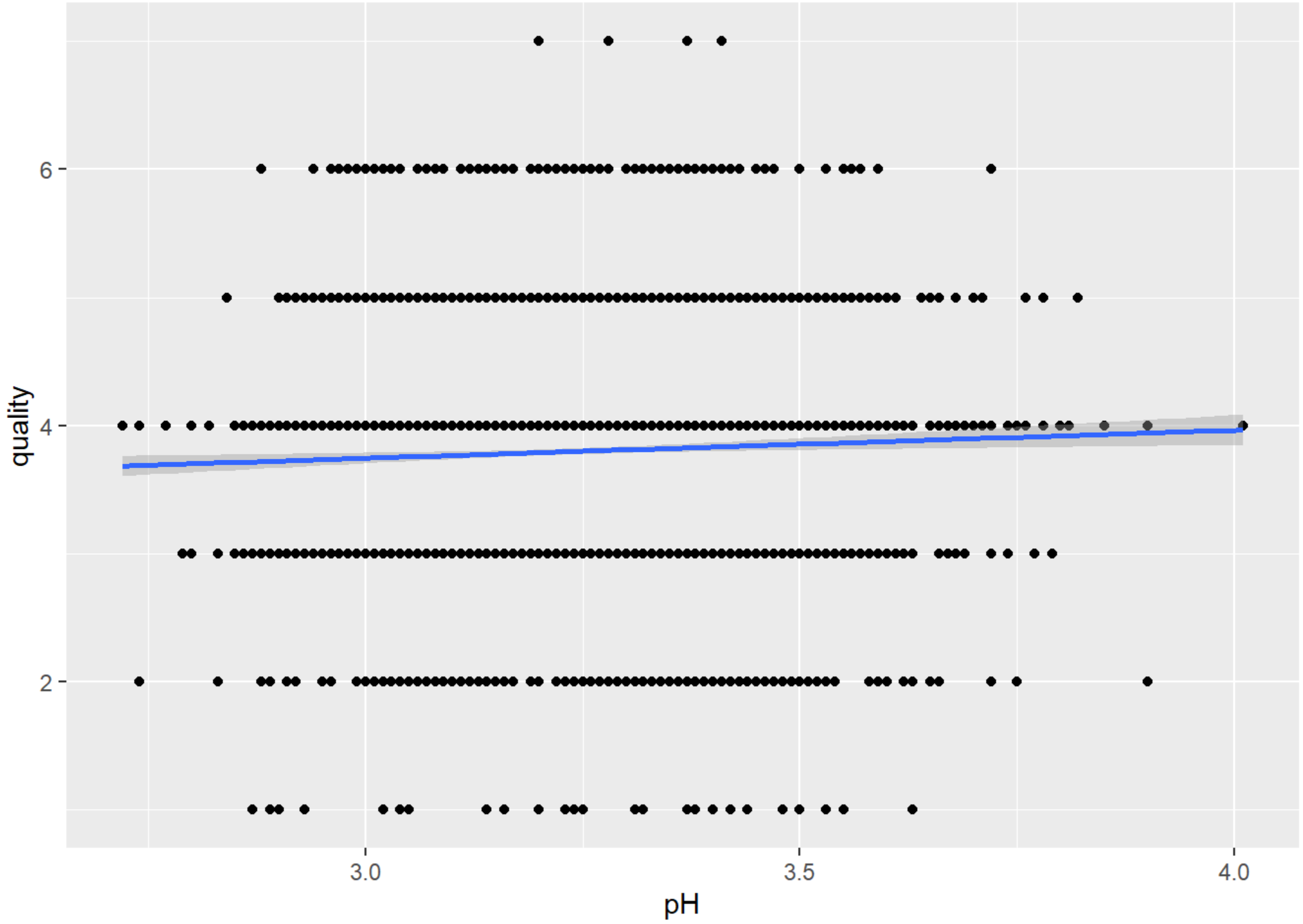
The outliers in our ingredients, along with the fact that our classes are imbalanced, indicates the possibility that some ingredients are found in excessive quantity in low-quality wine and some in high-quality wines. Also, there are not as many high-quality wines and low-quality wines as there are medium-quality wines.

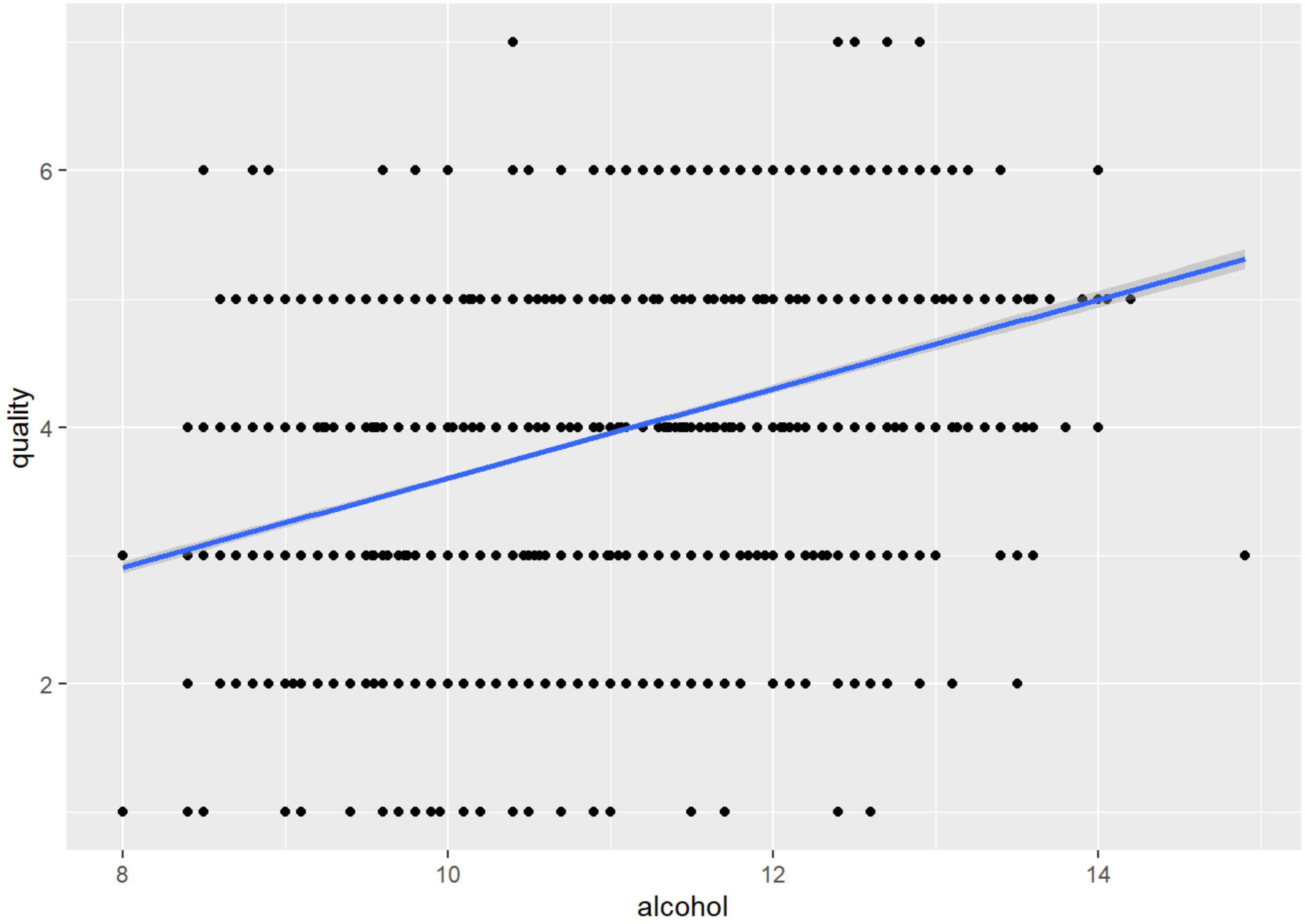
Appendix

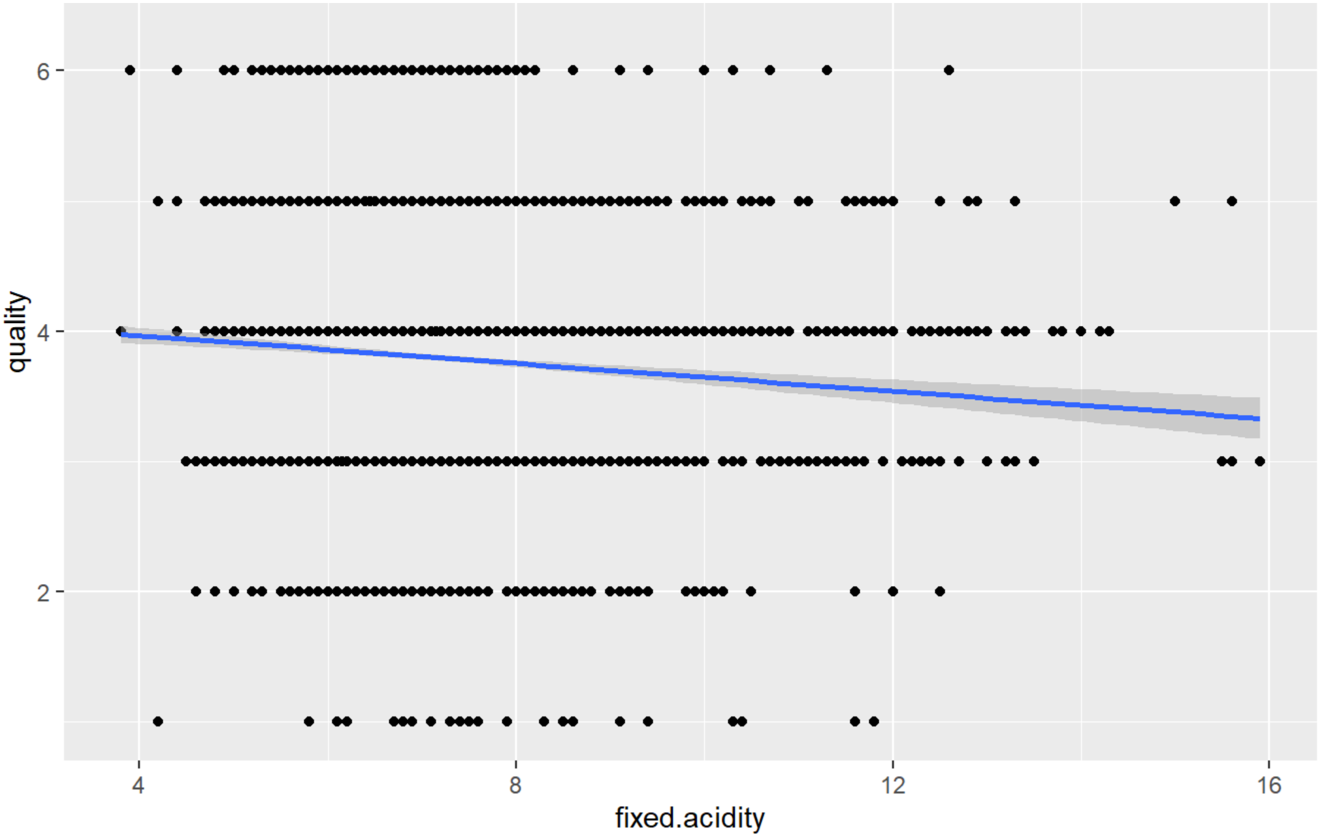
Work Distribution

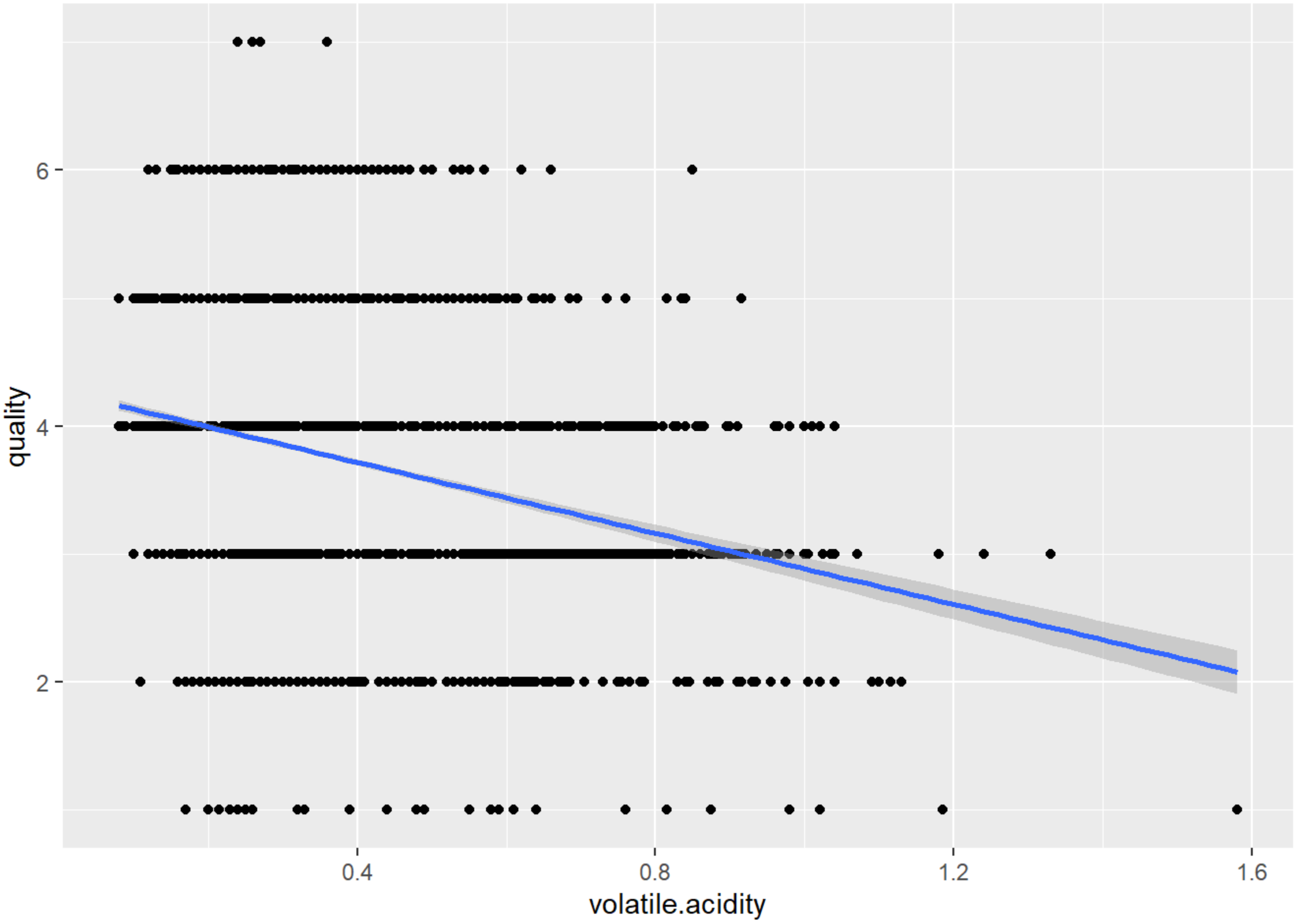
Member	Contribution
Shahzeb Naveed	Plots, Reporting, EDA
Zaryab Javaid	EDA, Data Cleaning
Mohsin Tahir	Regression Modelling

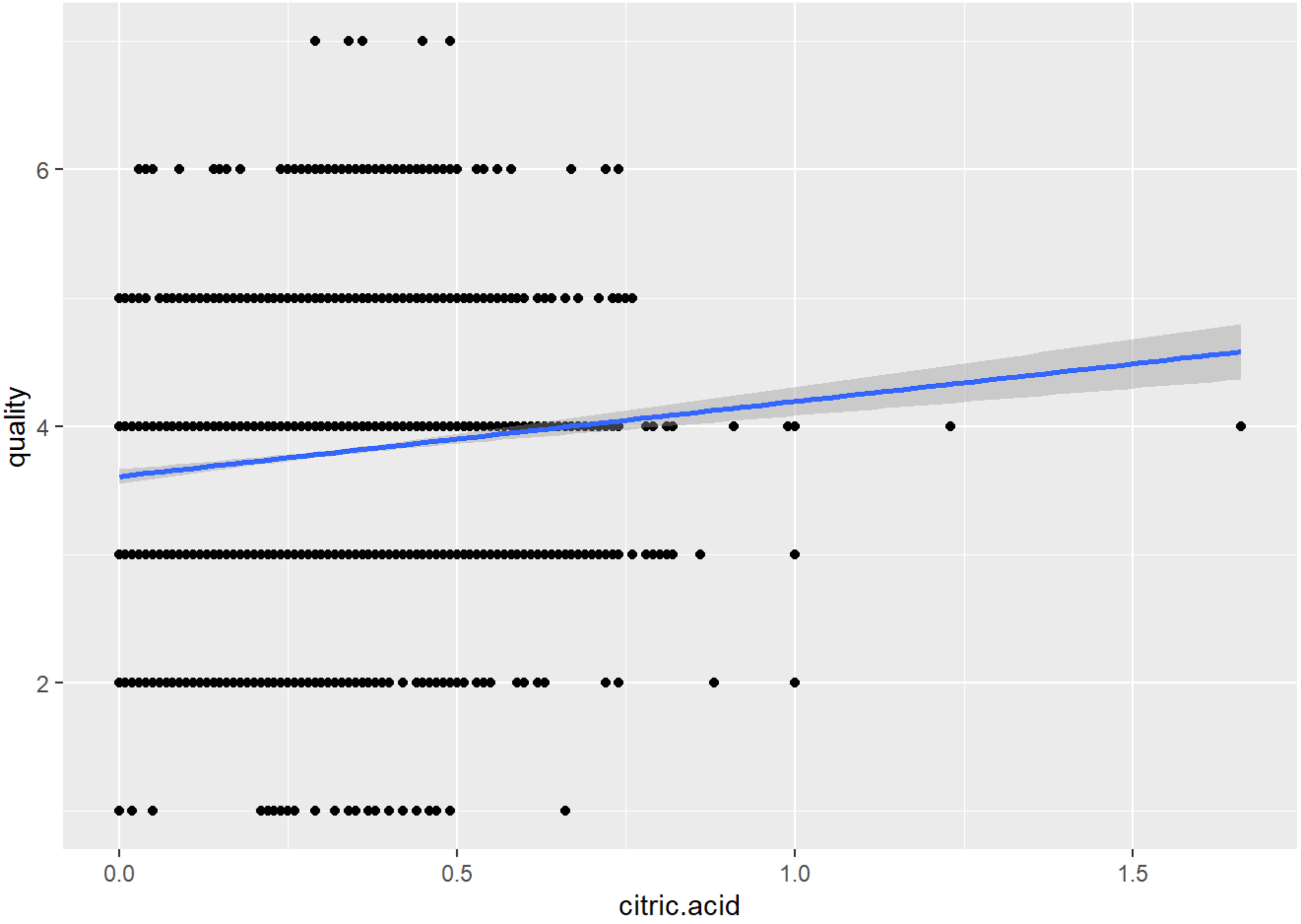
Appendix (i)

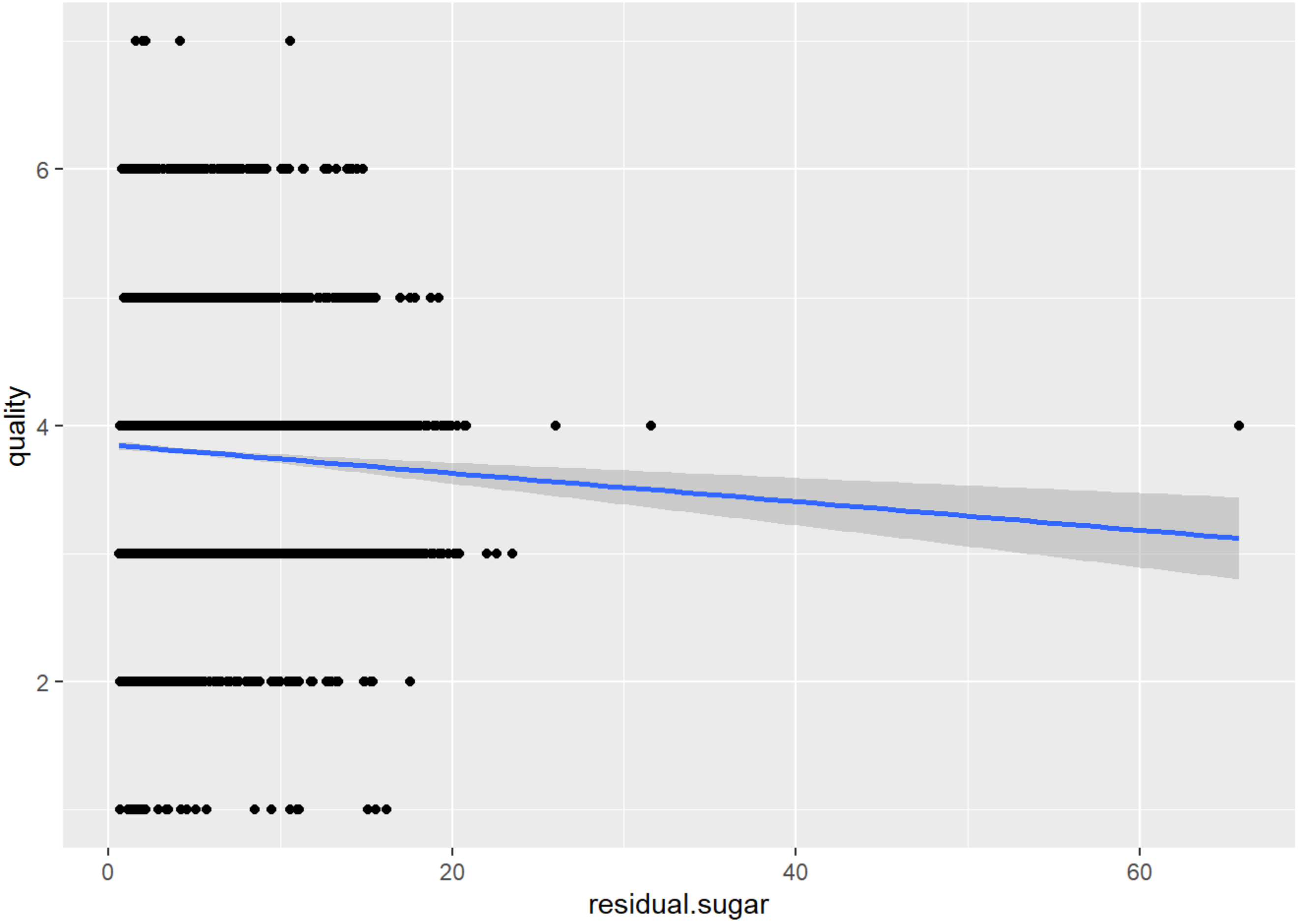

quality
6
4
2
3.0
3.5
4.0
pH

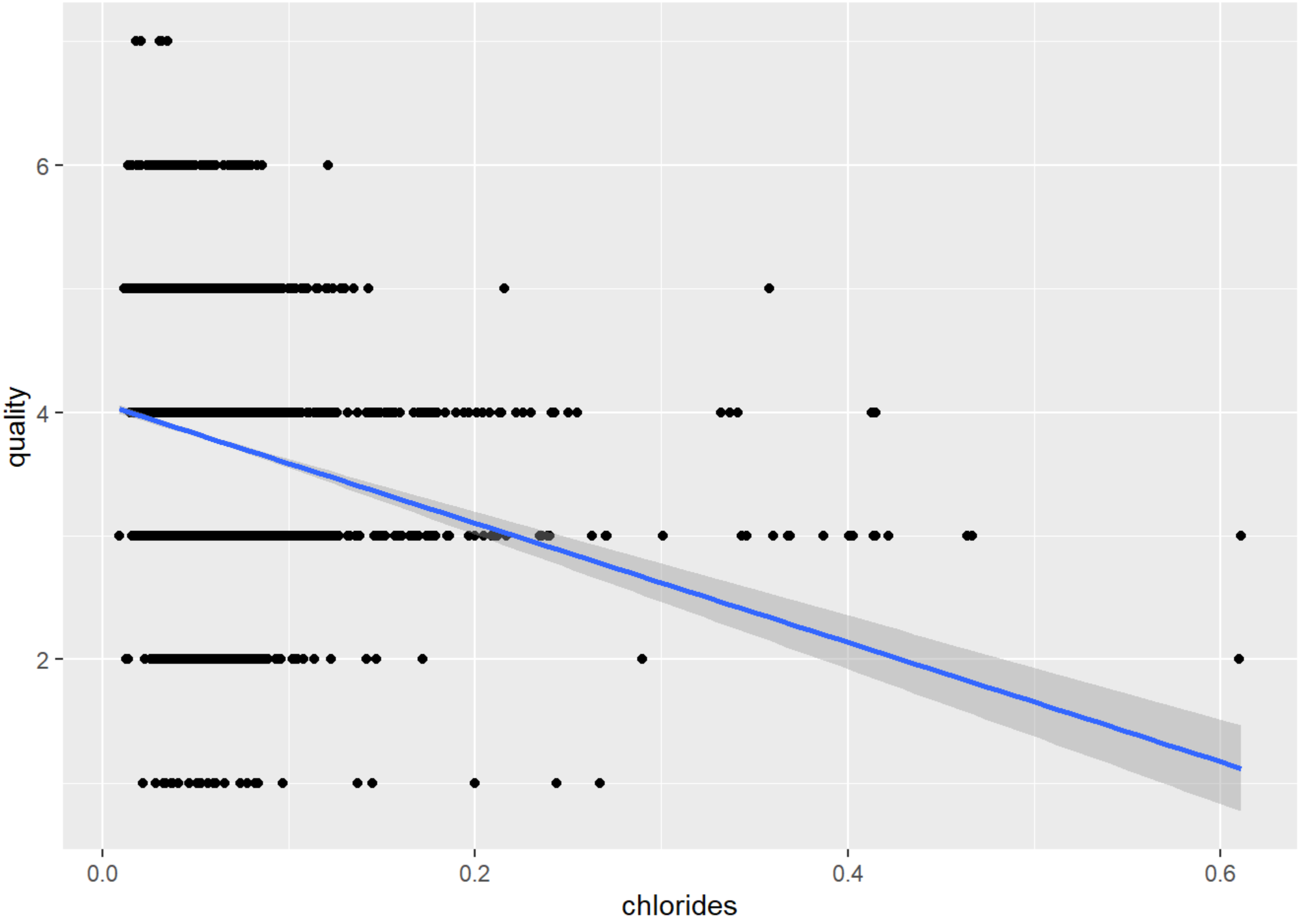

quality
6
4
2
8
10
12
14
alcohol

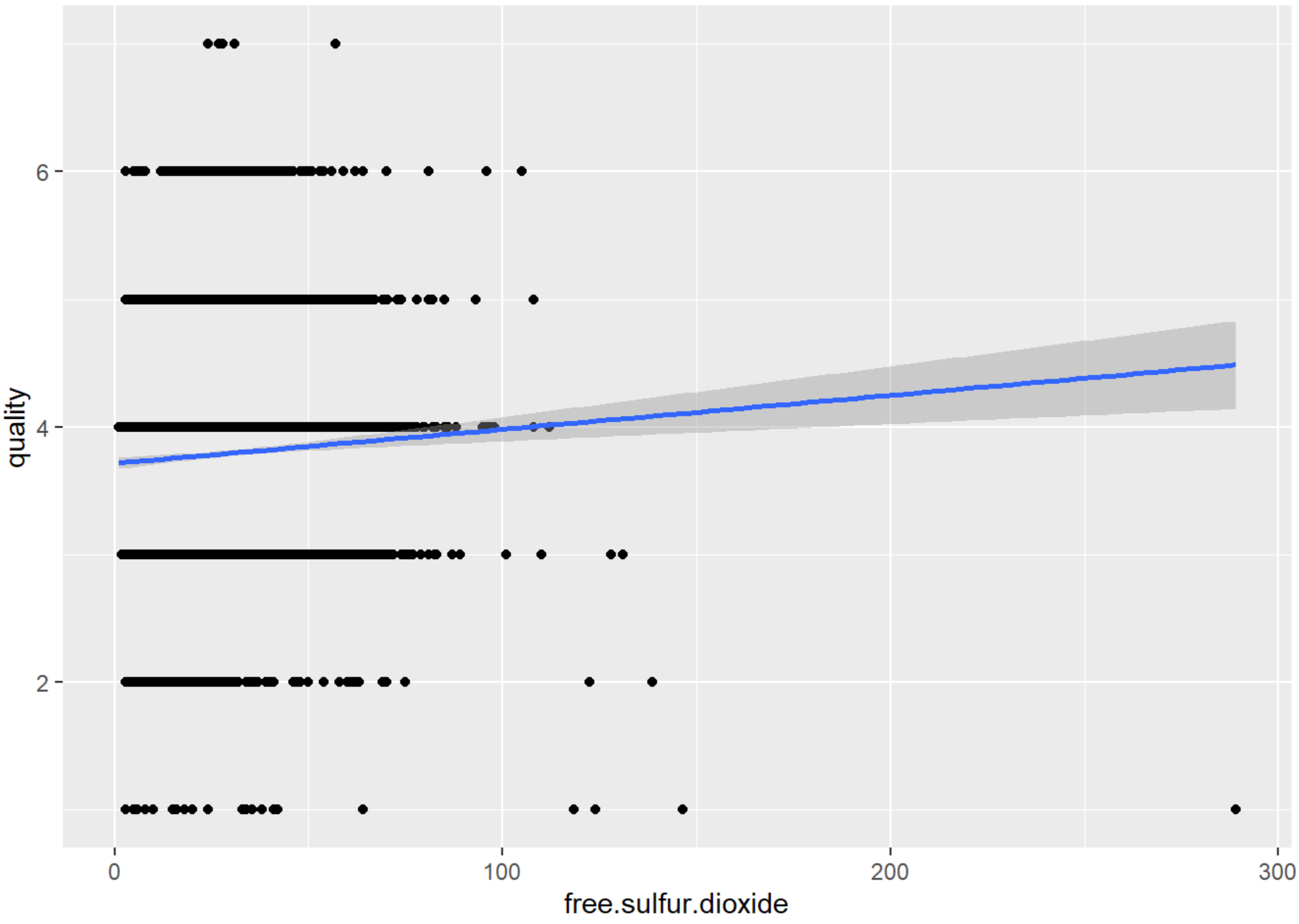

quality
6
4
2
4
8
12
16
fixed.acidity

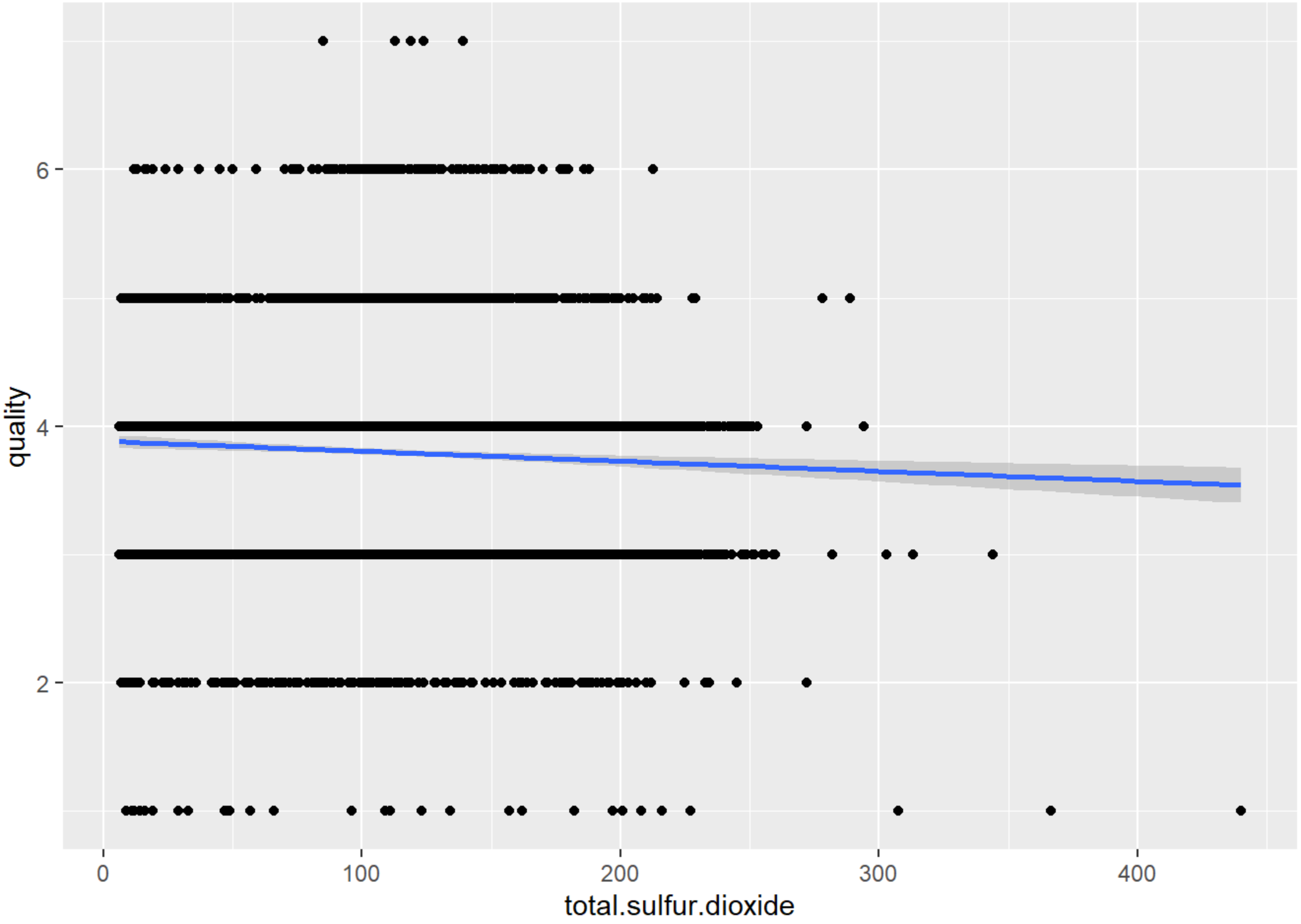

quality
6
4
2
0.4
0.8
1.2
1.6
volatile.acidity

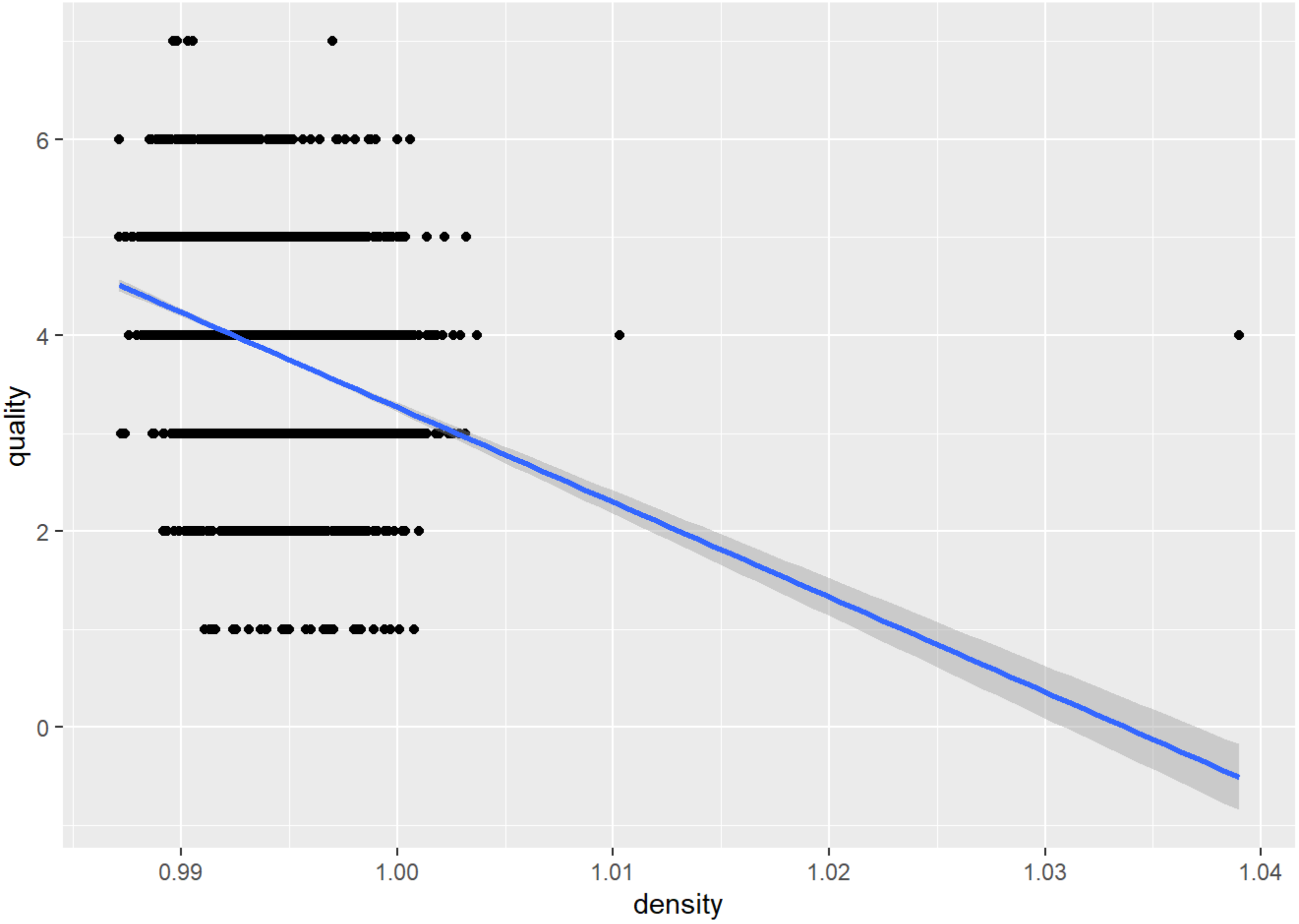

quality
6
4
2
0.0
0.5
1.0
1.5
citric.acid

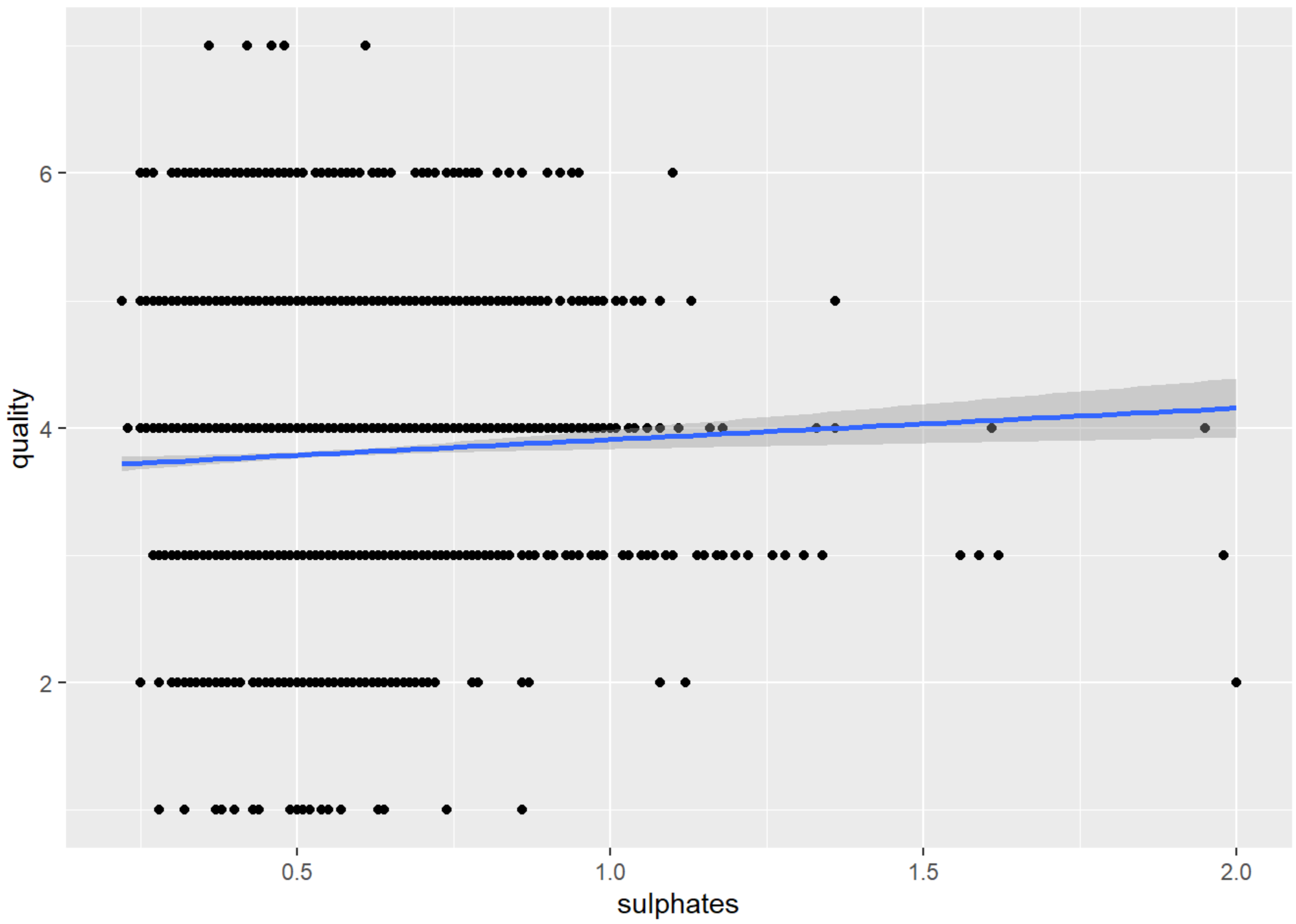

quality
6
4
2
0
20
40
60
residual.sugar


6
4
2
quality
0.0
0.2
0.4
0.6
chlorides


6
4
2
quality
0
100
200
300
free.sulfur.dioxide


quality
6
4
2
0
100
200
300
400
total.sulfur.dioxide


quality
6
4
2
0
0.99
1.00
1.01
1.02
1.03
1.04
density



Appendix (ii)

```
##       fixed.acidity     volatile.acidity       residual.sugar
##            4.483849             2.114120             5.232880
##           chlorides              density  free.sulfur.dioxide
##            1.455882            10.170734             3.598477
## total.sulfur.dioxide                  pH                color
##            6.570343             2.076801             7.231698
##           sulphates              alcohol
##            1.781381             2.577716
```

Appendix (iii)

```
##
## Call:
## glm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + density + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + color + sulphates + alcohol, family = binomial(), data = final)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8361   0.0505   0.0707   0.0987   0.8752
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -64.564099 196.414364  -0.329 0.742372
## fixed.acidity         -0.807737   0.218956  -3.689 0.000225 ***
## volatile.acidity      -5.319662   0.957892  -5.554  2.8e-08 ***
## residual.sugar        -0.032271   0.093590  -0.345 0.730237
## chlorides            -12.013227   3.742827  -3.210 0.001329 **
## density               88.301371 199.369972   0.443 0.657837
## free.sulfur.dioxide   -0.036657   0.010816  -3.389 0.000701 ***
## total.sulfur.dioxide   0.008427   0.006846   1.231 0.218290
## pH                    -4.446619   1.582166  -2.810 0.004947 **
## color                  2.865665   1.093708   2.620 0.008789 **
## sulphates              3.374793   1.959029   1.723 0.084945 .
## alcohol                0.299832   0.296180   1.012 0.311380
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 370.51  on 5319  degrees of freedom
## Residual deviance: 304.15  on 5308  degrees of freedom
## AIC: 328.15
##
## Number of Fisher Scoring iterations: 9
```

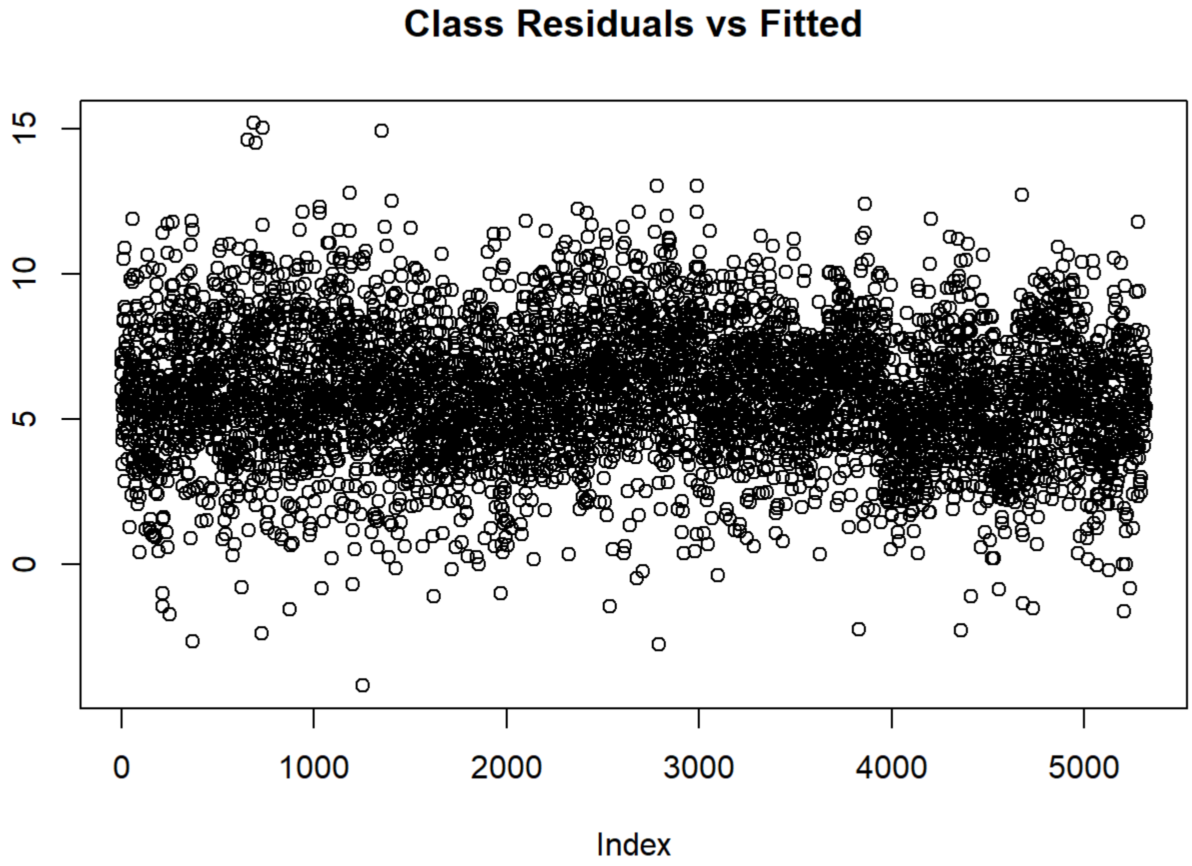
Appendix (iv)

```
## Start:  AIC=11719.3
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + total.sulfur.dioxide + sulphates + alcohol
##
##                        Df   AIC
## - citric.acid           1 11717
## - fixed.acidity         1 11718
## <none>                    11719
## - chlorides             1 11724
## - total.sulfur.dioxide  1 11736
## - residual.sugar        1 11758
## - sulphates             1 11815
## - volatile.acidity      1 12027
## - alcohol               1 12790
##
## Step:  AIC=11717.31
## quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + total.sulfur.dioxide + sulphates + alcohol
##
##                        Df   AIC
## - fixed.acidity         1 11716
## <none>                    11717
## - chlorides             1 11723
## - total.sulfur.dioxide  1 11735
## - residual.sugar        1 11756
## - sulphates             1 11813
## - volatile.acidity      1 12095
## - alcohol               1 12807
##
## Step:  AIC=11716.53
## quality ~ volatile.acidity + residual.sugar + chlorides + total.sulfur.dioxide +
##     sulphates + alcohol
##
##                        Df   AIC
## <none>                    11716
## - chlorides             1 11722
## - total.sulfur.dioxide  1 11733
## - residual.sugar        1 11755
## - sulphates             1 11812
## - volatile.acidity      1 12096
## - alcohol               1 12830
```

Appendix (V)

```
##                              Value   Std. Error    t value       p value
## volatile.acidity     -3.680594811 0.1920680588 -19.162972  7.544023e-82
## residual.sugar        0.043944326 0.0069212601   6.349180  2.164657e-10
## chlorides            -2.500683839 0.8868676528  -2.819681  4.807142e-03
## total.sulfur.dioxide -0.002620399 0.0006062559  -4.322266  1.544347e-05
## sulphates             1.963793250 0.2012843657   9.756313  1.733431e-22
## alcohol               0.905261810 0.0285225223  31.738491 4.577244e-221
## 3|4                   3.192979759 0.4001568320   7.979321  1.471406e-15
## 4|5                   5.348236534 0.3619094681  14.777830  2.036220e-49
## 5|6                   8.378000222 0.3628078594  23.092113 5.556908e-118
## 6|7                  11.002758103 0.3796697579  28.979812 1.182151e-184
## 7|8                  13.345507455 0.3960133694  33.699639 5.851558e-249
## 8|9                  16.841969138 0.5930228029  28.400205 2.010367e-177
```

Appendix (vi)


Class Residuals vs Fitted
Index