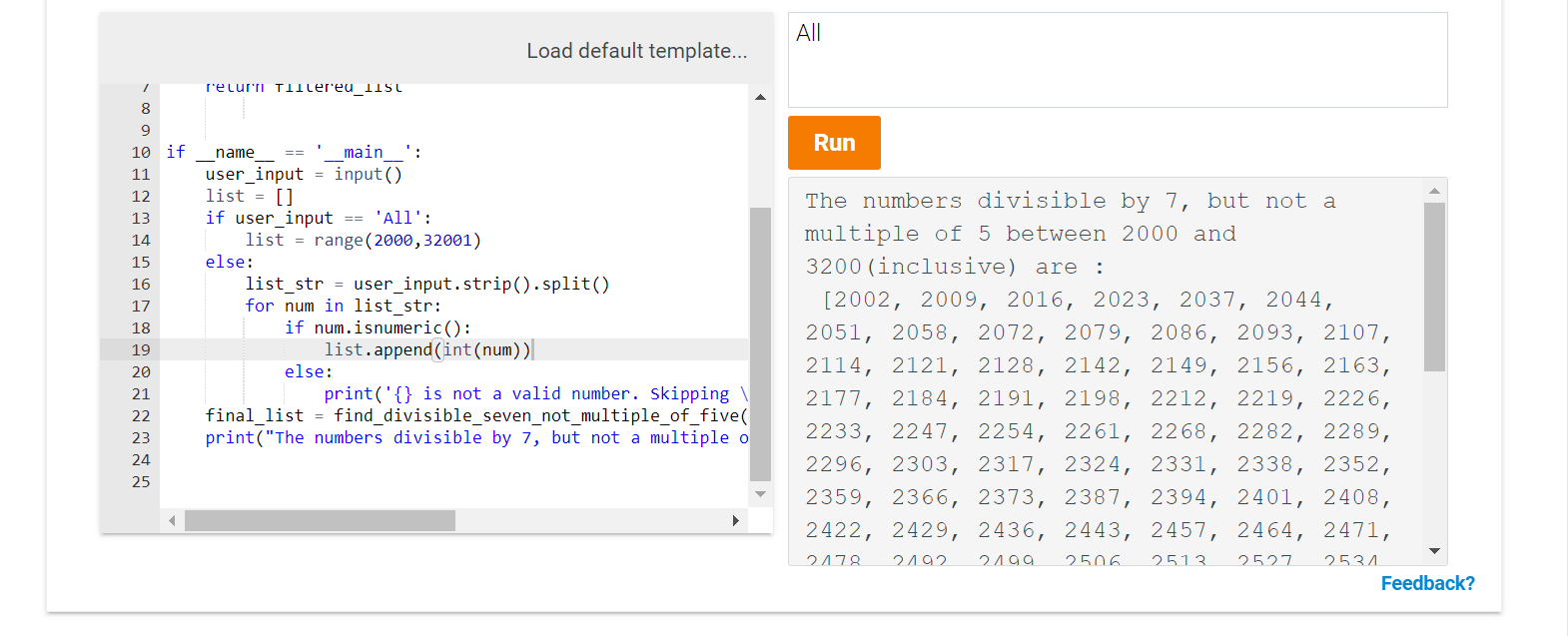
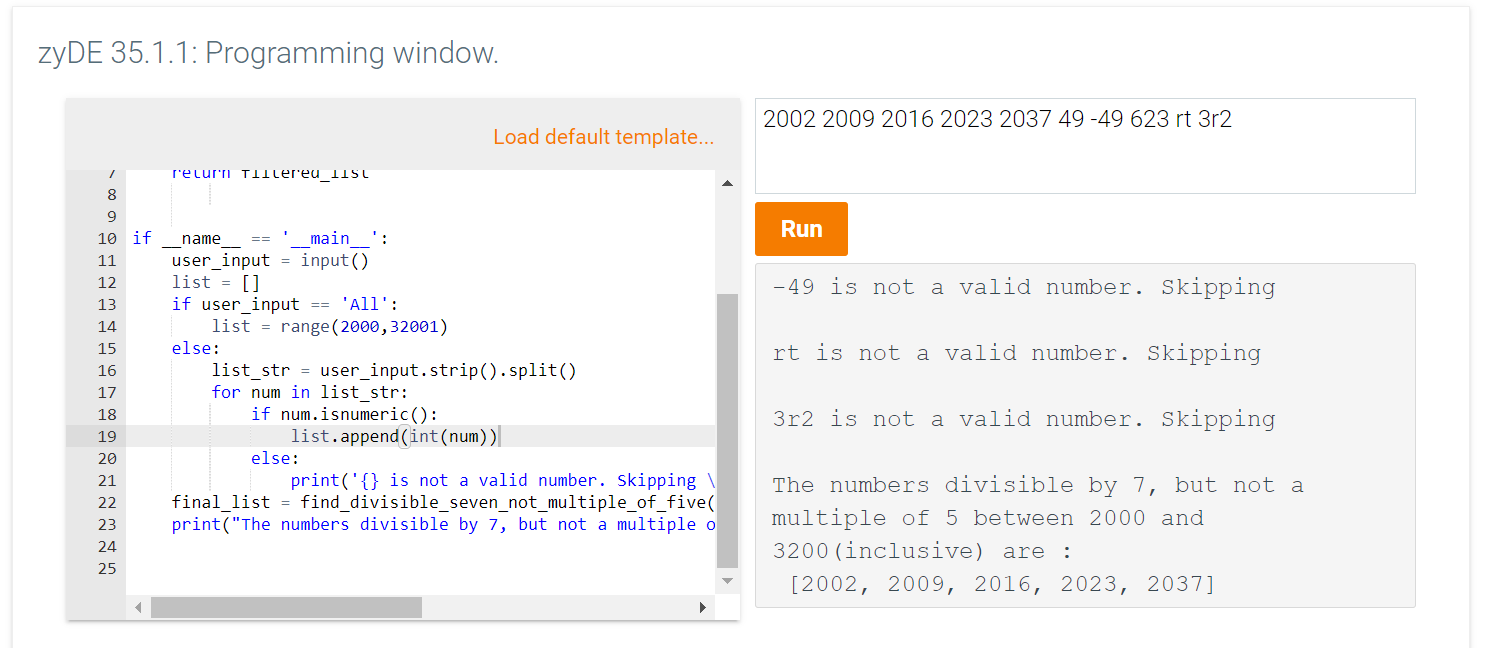
**Problem 1**

Output for option ‘All’

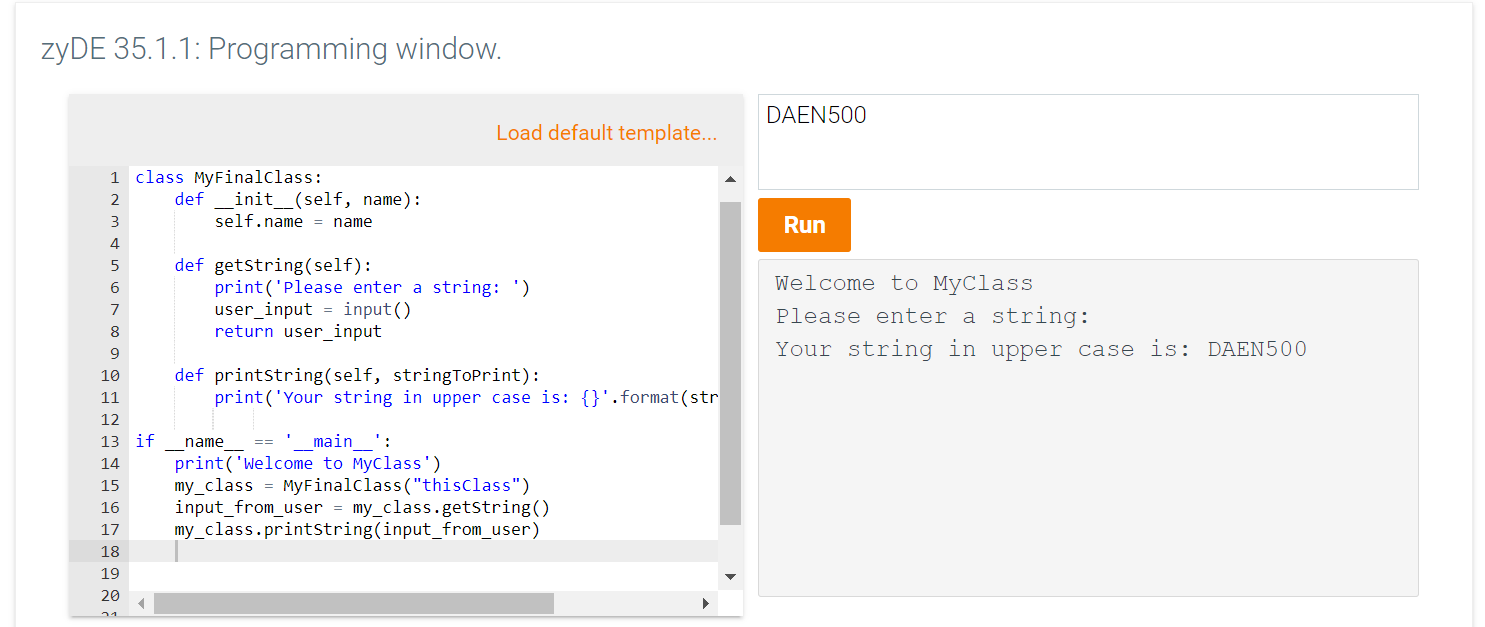


Output for option with user defined input with incorrect numbers

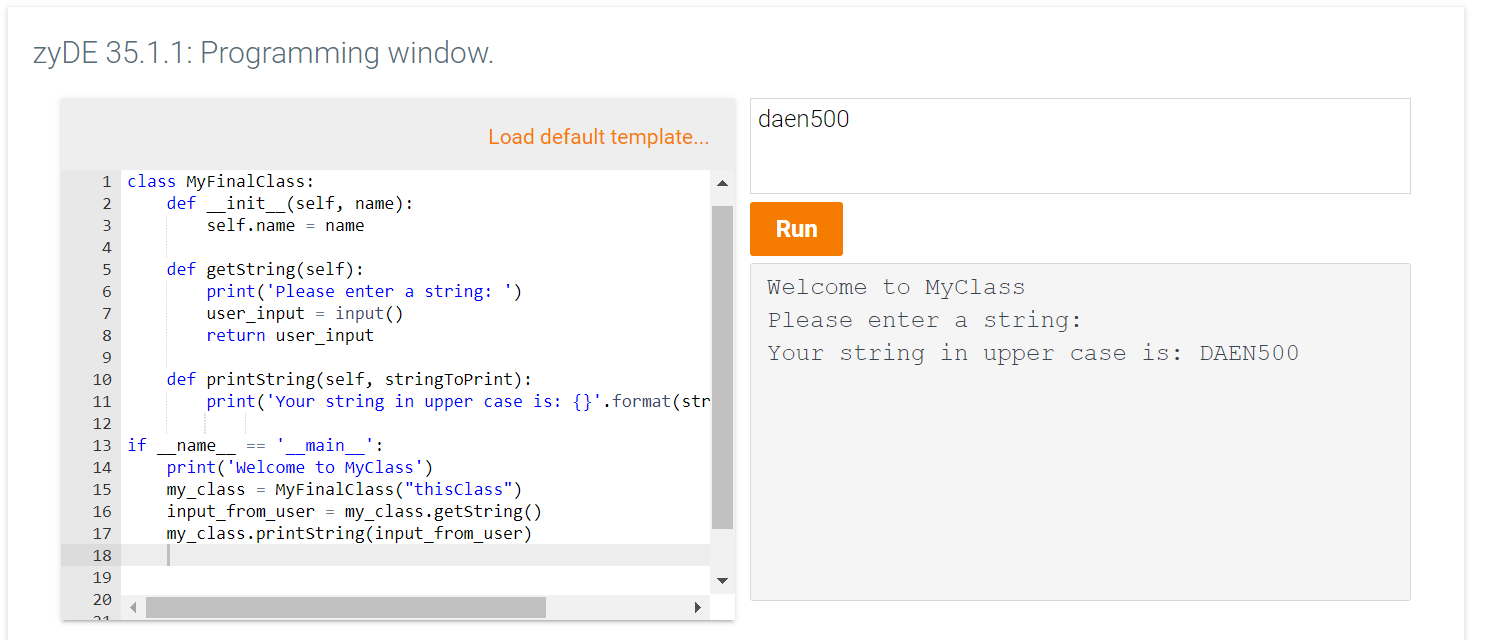


**Problem 2**

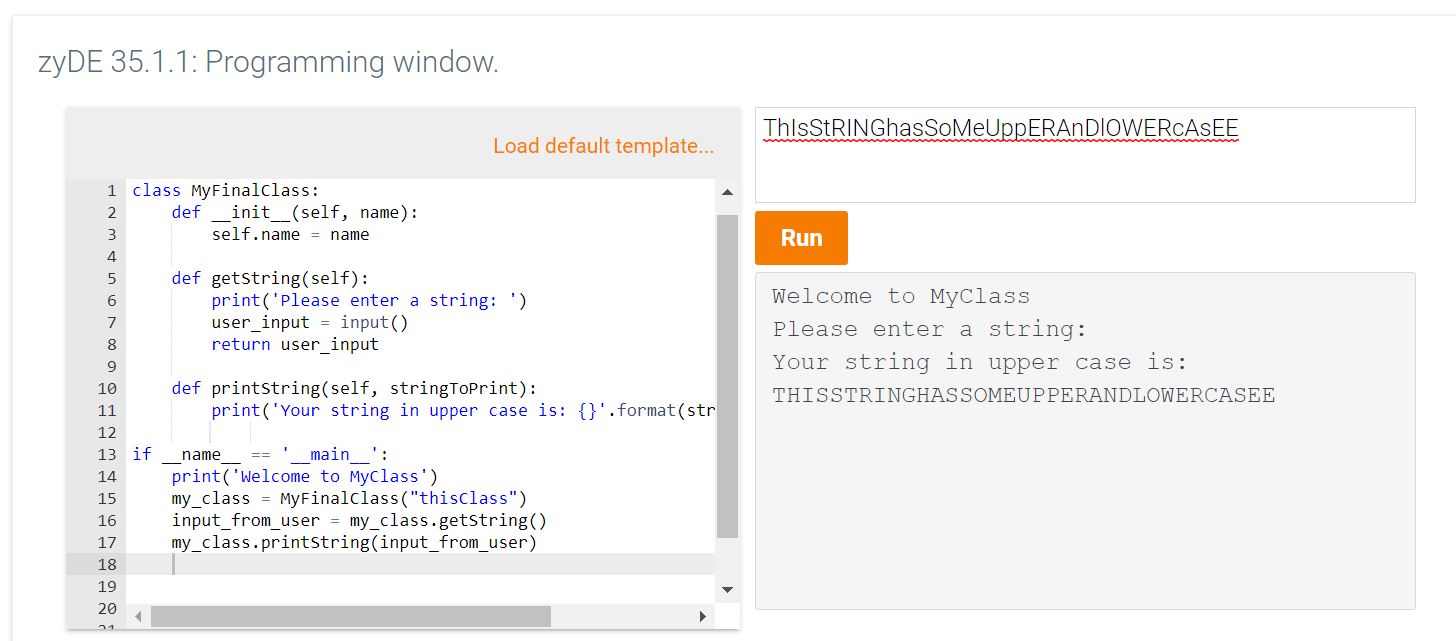
Output for all upper-case input



Output for all lower-case input



Output for mix of upper and lower case



**Problem 3**

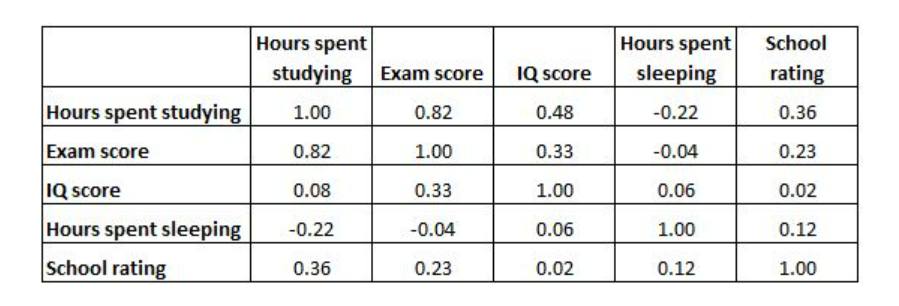
R Programming



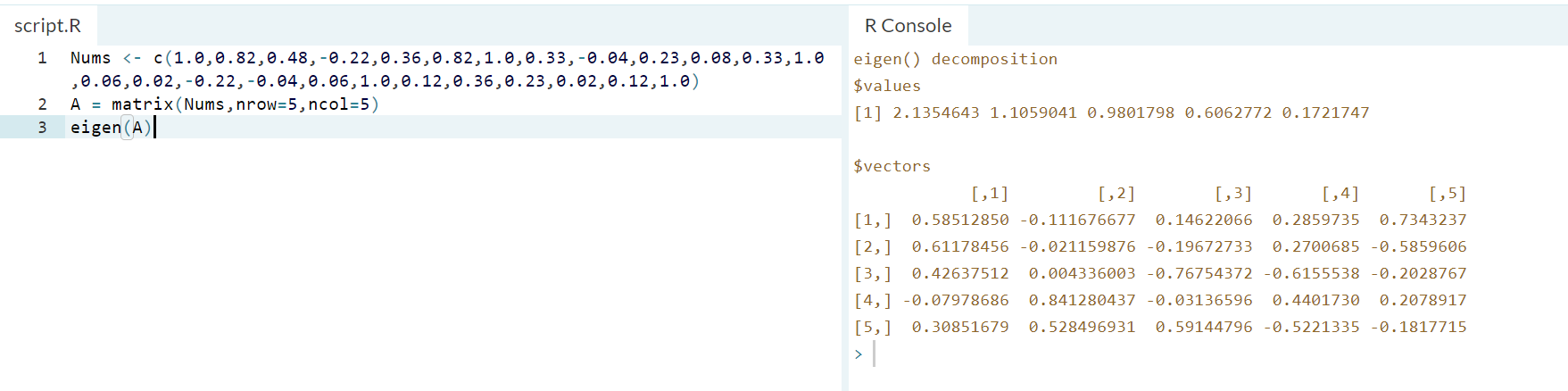
**Problem 4**

1. What is a component?

A component is a new variable created by combining multiple predictor variables in a model together. The variables that are combined to create a new one is usually highly correlated, and therefore it makes sense to combine them. An example could be number of years of a car and total mileage in determining the price of a car. Both number of years and mileage are highly correlated, and inversely proportional to the price (the more years and mileage, the less value for the car). Therefore, we could combine both variables into a new category called ‘years\_and\_mileage’.

1. Principal Component Analysis (PCA) is a statistical method that aims to reduce the number of variables by using components. The idea is to simplify the analysis with a smaller number of variables by combining highly correlated with weighted combination of predictor variables into components while using as much of the original data as possible. It shows which variables account for most of the data variability in the data set. Creating components involves rotating the data by creating new axis in the direction of the principal components. The principal components have the most variability in the data, and the second has less and it is perpendicular to the first component. Any subsequent component represents a new axis of data with less variability than the previous one. Components are calculated by finding the correlation matrix of the data, and then computing the eigenvectors and eigenvalues.
2. The following correlation matrix shows the correlation coefficients between several variables related to education (source: statology.org)

From R, we can find the principal components by finding the eigenvectors of the previous matrix:



From the above we can observe that the eigenvalues obtained from the correlation matrix (with assumed standardized data already) are 2.13, 1.10, 0.98, 0.60, and 0.17. The $vectors lists each variable (corresponding to Hours spent studying, Exam scores, IQ score , Hours spent sleeping, and School rating) in a row, and each eigenvector in a column. The first eigenvector has values of 0.58, 0.61,0.42, -0.07, and 0.30. The first principal component has similar weights for each variable with Exam score contributing to largest weight. It is a good indicator for how successful a student can be. It also accounts for about 43% of the total variability in the data. Based on the eigenvalues calculated by R, we can observe that the last 3 values are less than 1. However, the third value of 0.98 is close enough to one. If we were to drop the last 2 from the model since they are less than 1 and therefore, they do not have that much impact nor do they contain a lot of the variability in the data, we will still account for about 84% of the variability of the data which is good enough, reducing the dimensionality of the data as well. As a result, the final principal components are:

**Z1 =** 0.5851 *HoursSpentStudying* + 0.6117 *ExamScore* + 0.4263 *IQScore* -0.0797 *HoursSleeping* + 0.3085 *SchoolRating*

**Z2 =** -0.1111 *HoursSpentStudying* - 0.0211 *ExamScore* + 0.0043 *IQScore* +0.8412 *HoursSleeping* + 0.5284 *SchoolRating*

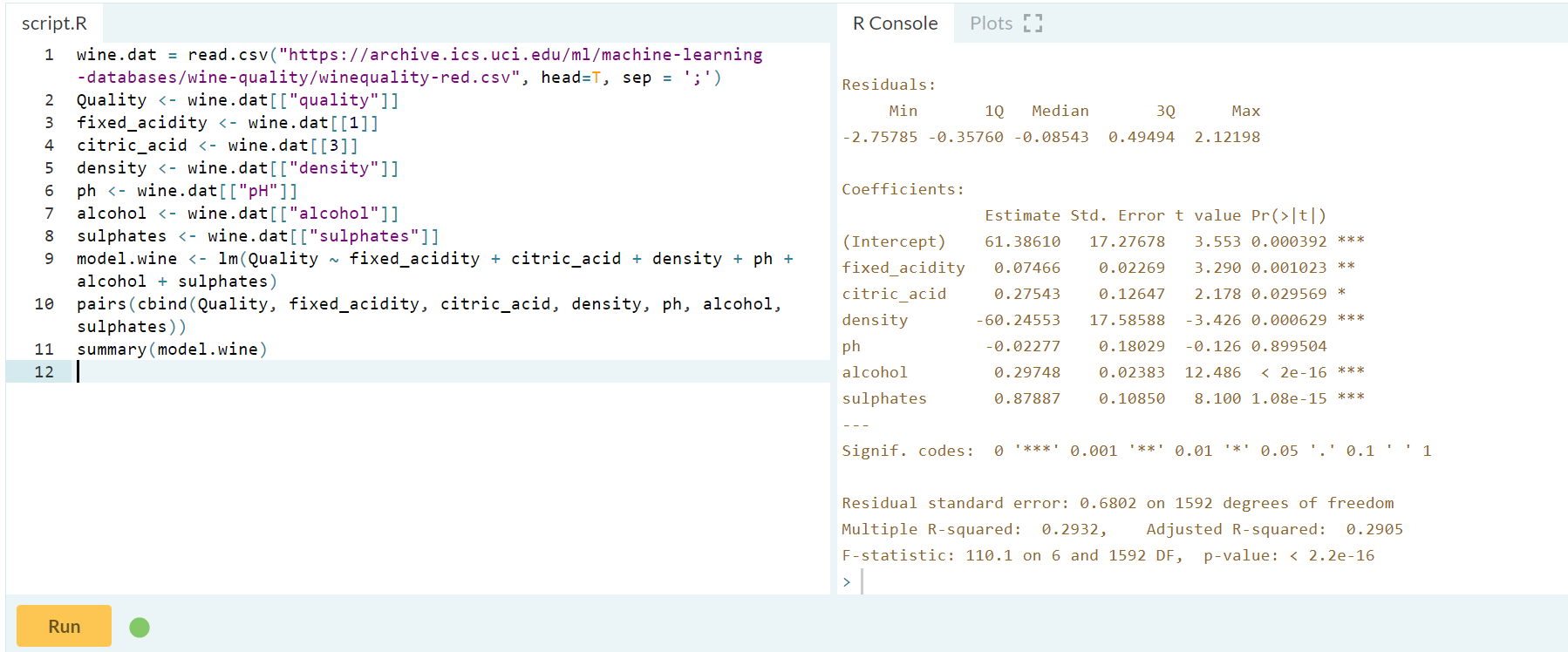
**Z3 =** 0.1462 *HoursSpentStudying* - 0.1967 *ExamScore* - 0.7675 *IQScore* -0.0313 *HoursSleeping* + 0.5914 *SchoolRating*

**Problem 5**

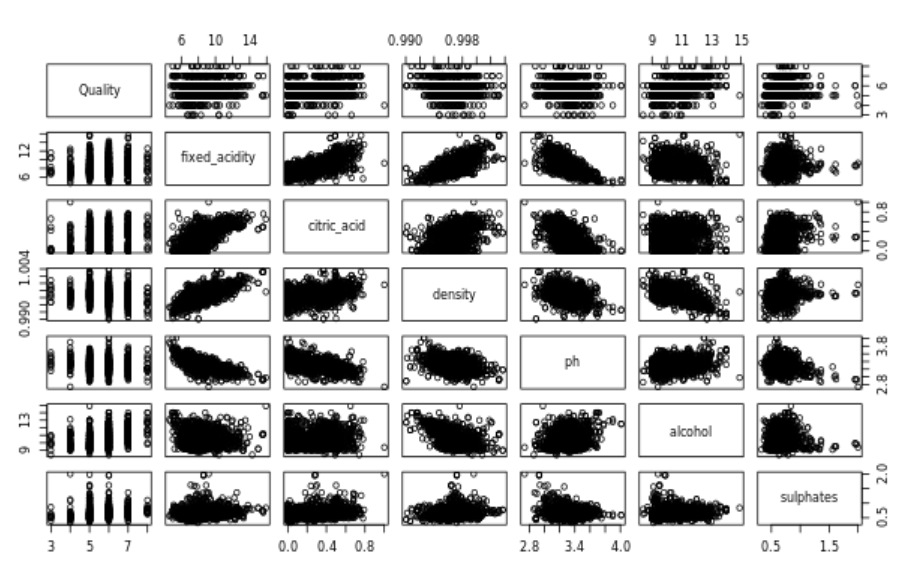
1. The difference between multiple regression and logistic regression relies mainly on the fact that all predictor variables in multiple linear regression must be numbers representing quantitative data in any range, and the response variable can take on any quantitative range as well. It implies that there is a linear relationship between the predictor and response variable. In contrast, for logistic regression, the response variable is binary, represented by either 1 or 0, accounting whether an event took place or not, or classifying a value into either one of the categories. Linear regression should be used for regression problems involving multiple variables that are not binary, and while trying to find the relationship between them by finding the line of best fit. An example is a data set of variables impacting the price of used cars such as mileage, and years. In contrast, logistic regression should be used while handling classification like problems. For example, given different variables like cholesterol level, determining whether a patient is prone or not to a heart attack. The only possible outcomes are yes or no, which is binary (0 or 1). Another example is whether someone can pass a blood alcohol test based on blood alcohol content. The outcomes are pass, or no pass (binary).
2. To demonstrate the previous, I decided to run multiple linear regression, and logistic regression on a data set from the UCI website using R. The data is about the different factors that influence the quality of red wine from Portugal. For the multiple linear regression model, the independent variables were fixed\_acidity, citric\_acid, density, ph, alcohol, and sulphates. The dependent variable ‘Quality’ has a range of values from 3 to 8. For the logistic regression, the same data set was used. However, the Quality variable was transformed into possible values of 0 (value < 6, then set to zero) and 1 (value > 8 then set to 1) to represent the categories of ‘Good’ and ‘Bad’ quality wines. The predictor variables for this model were the alcohol and sulphates.

**Linear Regression:**

By running the lm command in R after importing the data, we get the results of the linear regression model using the predictor variables described above.

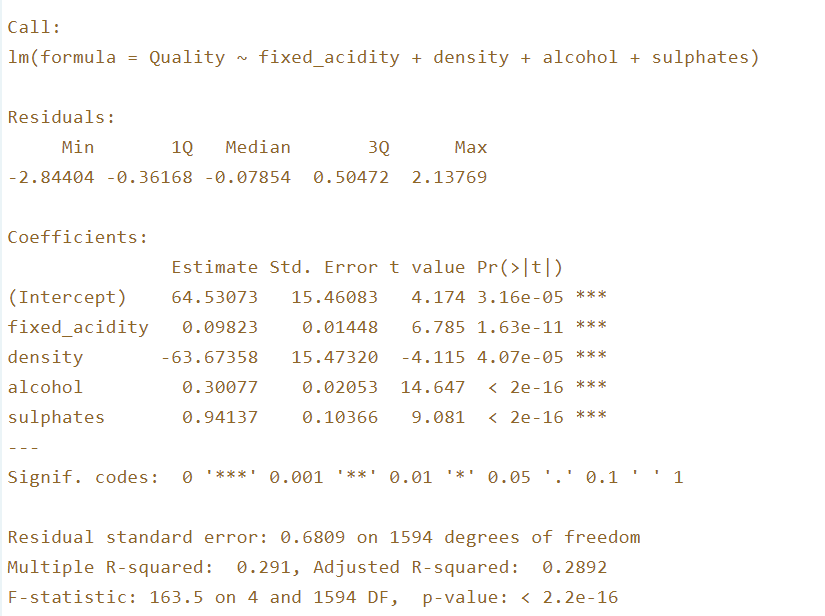


The below shows the scatter box matrix all variables combinations:

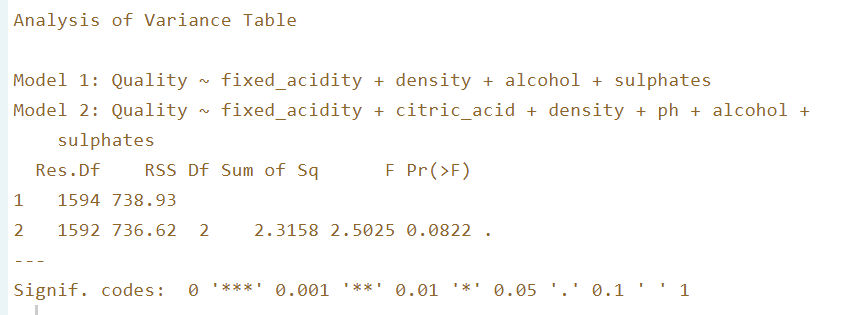


By looking at the Coefficients table, we can observe that most of the p values of the predictor variables are below the significance level of 0.05, except for pH (0.89). Although lower than 0.05, citric acid p value is also a high (0.02) in comparison to the other variables. This observation suggests we can explore and alternative model dropping these 2 variables. By doing a general linear F-test we can see if this assumption is justified. In R, we can run the anova command to compare both models, the results are the following:

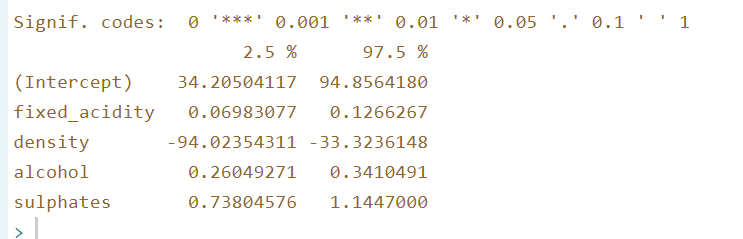
**New model**



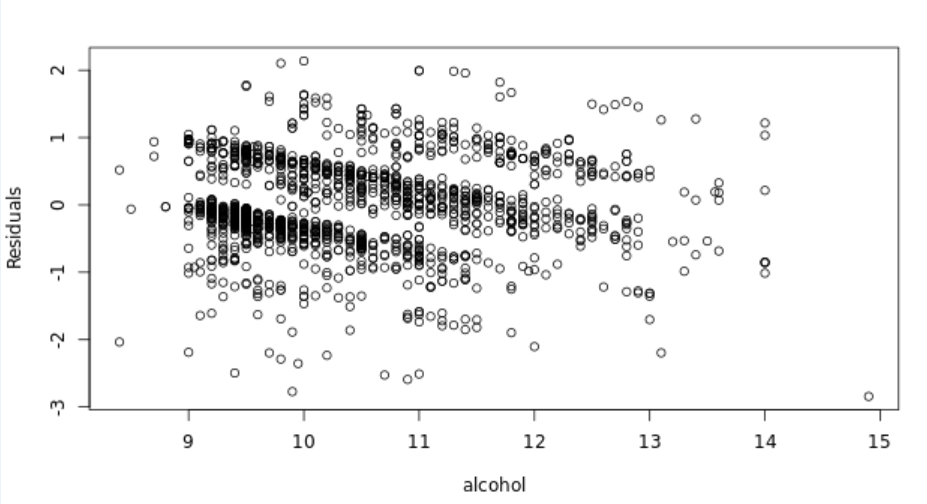
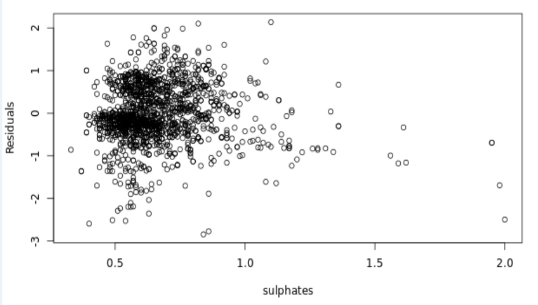
**ANOVA test comparing both models:**

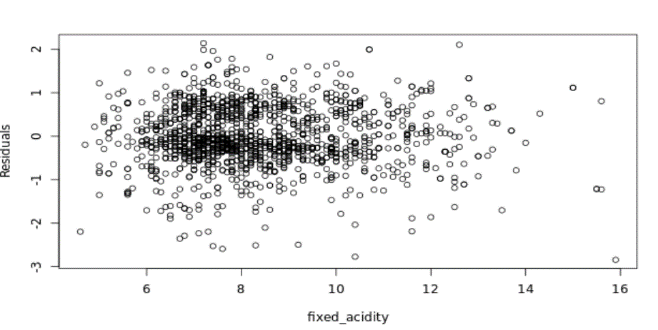


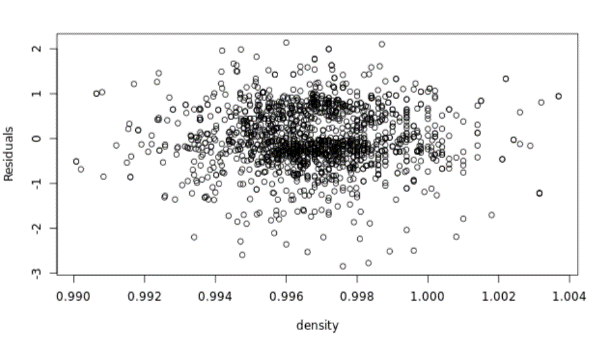
A p value of 0.08 is higher than the significance level of 0.05, and therefore insufficient evidence is presented to reject the null hypothesis, suggesting that dropping both pH and citric acid from the original model is justified. The confidence interval at a 95% for the new model is shown below:

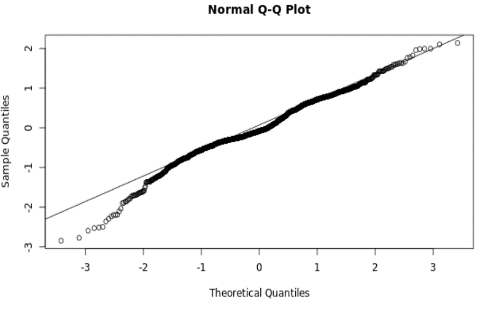


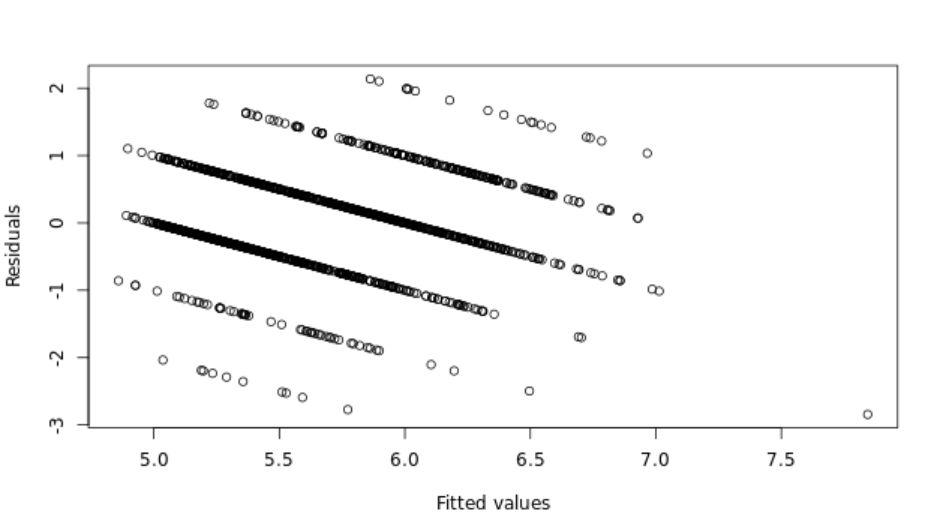
The model has a relatively high standard error with a value of 0.68, and a low R squared value of 0.28. This means that this model using the predictor variables fixed acidity, density, alcohol, and sulphates only accounts for less than 30% of the data variance of the wine quality data set. The following residual plots for each of the variables demonstrate that some of the linearity assumptions might not hold:







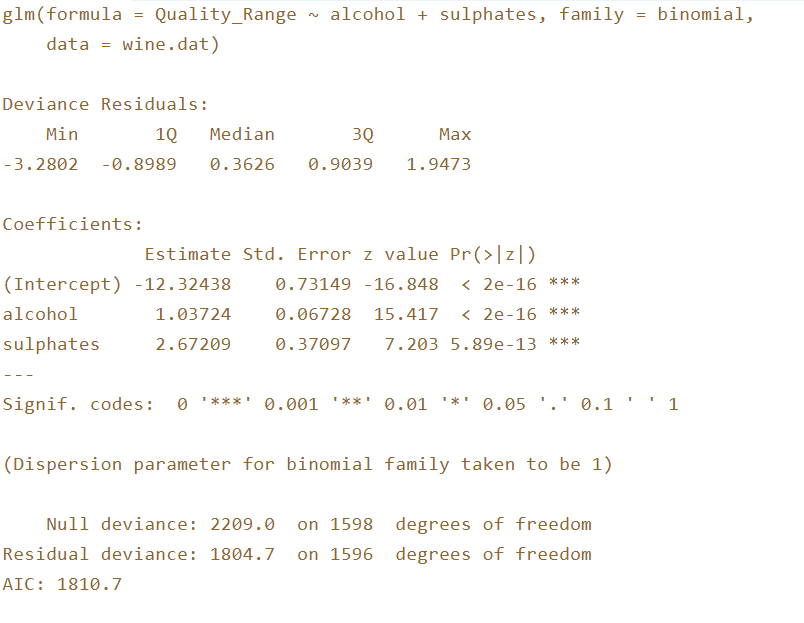




From the previous graphs we can observe that the assumption of the mean of the residuals should be approximately zero for each fixed value on the horizontal axis does not hold for some variables like sulphates. In the normal QQ plot, the plotted points lie reasonably close to the diagonal line, so the "normality" assumption appears to hold in this case. Overall, the model could be further improved by studying the possible existing of interaction terms, or some polynomials of some variables.

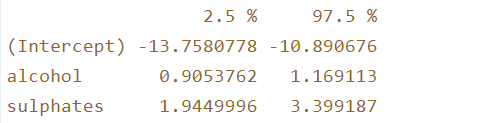
**Logistic Regression:**

By running the glm command in R, we can create a logistic model for the response variable quality and the predictor variables alcohol and sulphates.

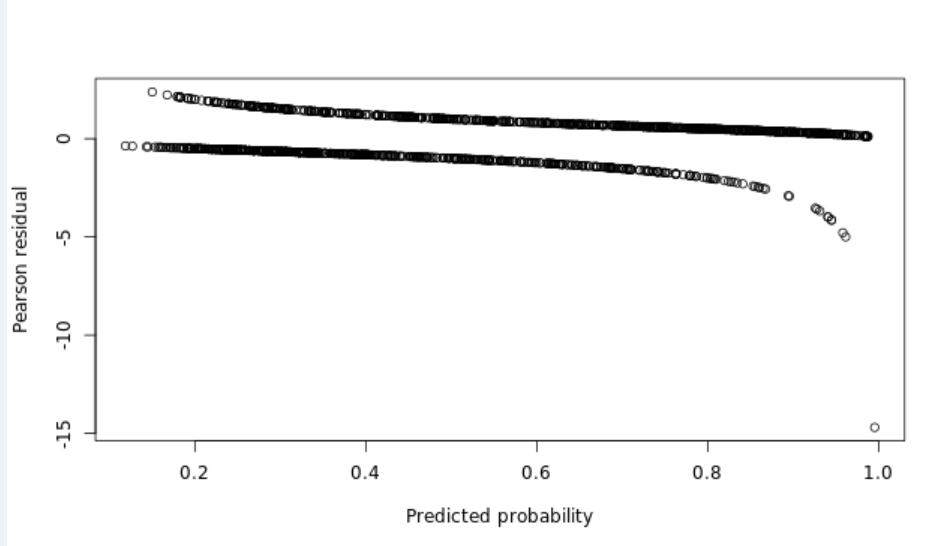


The estimate for b₀ is -12.32, while the estimate for b₁ and b2 are 1.03 and 2.67 respectively. This suggests that a wine with a higher alcohol content and higher sulphates levels might increase the chances of a wine to be classified as a ‘good’ quality one. The Wald test shows that sufficient evidence exists to reject the null hypotheses and suggests that both β₁ and b2 are significantly different from zero (based on the p values from the above results).

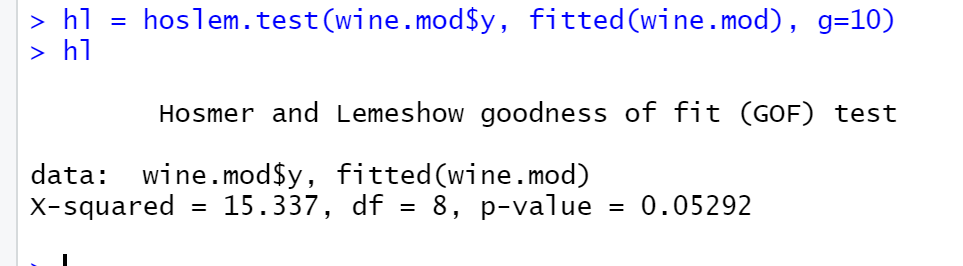
The 95% Wald confidence interval for the model is:



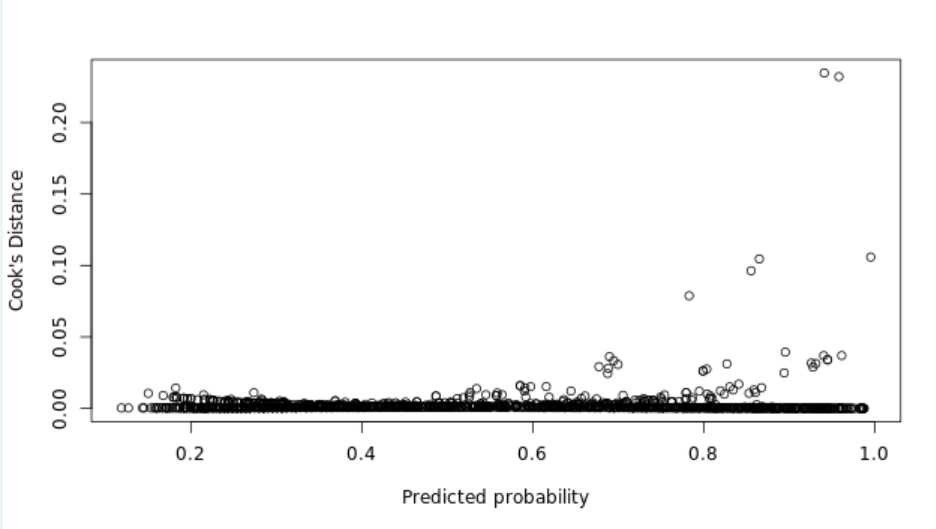
By plotting the Pearson residuals against the predicted probability determined by our model show some unusual observations. High probability values should show positive residuals, but there are a couple of high probability values with negative Pearson residuals. This suggests that these observations had high alcohol and sulphate contents, but were still classified as ‘bad’ quality wine.



The Hosmer-Lemeshow test yields a P-value of 0.0529. This value is slightly larger than a significance level of 0.05. Thus, insufficient evidence exists to reject the null hypothesis that the model fits suggesting the model does fit the data. However, the value is very close to the significant level. This could be the result of outliers in the data.



The Cook’s distance plot can show some of these possible outliers. The Cook’s distance indicates how much an observation is likely to change the parameter estimates.



We can observe a couple of values with large Cook’s distance and high predicted probabilities. For these values, the model suggests they should be classified as ‘good’ but, they are not. Further analysis can be done to improve the suggested model.