Predictive Modelling Report 1

Data Summary

Numerical Summary

Table 1: Summary of the Variables in the dataset. (Appendix A.1)

Variable	Meaning of each Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max
y	Final Exam Scores (Response Variable)	26.00	55.00	64.00	63.17	72.00	94.00
x1	Study Hours	0.00	1.80	3.80	3.91	6.00	8.00
x2	Attendance (1=present, 0=Absent)	0.00	1.00	1.00	0.89	1.00	1.00
x3	Previous Exam Scores	16.00	53.75	68.00	66.40	81.00	99.00
x4	Sleep Hours	2.10	4.70	7.20	7.16	9.70	12.00
x5	Homework Completion (1 = Yes, 0 = No)	0.00	0.00	1.00	0.60	1.00	1.00
x6	Participation in Study Groups (0 or 1)	0.00	0.00	1.00	0.70	1.00	1.00
x7	Extracurricular Activities (Count)	0.00	1.00	2.00	2.38	4.00	6.00
x8	Class Participation	0.00	0.20	0.20	0.26	0.30	0.80

The data showcases that final exam results (y) can vary. The range is 26 to 94, with a mean of 63.17, which shows a moderate performance among the pupils. Study hours (x1) have a median of 3.8, highlighting that students spend some time studying, which indicates the differences in scores. As most students had high attendance (x2) it highlights constant class involvement for most students, therefore improving learning results. As previous exam scores (x3) range from 16 to 99, this can affect the final exam score (y) as there may be differences in prior knowledge.

Sleep hours (x4) mean was 7.16, showing that most students had sufficient rest, but if the student had excessive or insufficient sleep can affect preparation and performance. Homework completion (x5) and participation in study groups (x6), with a mean of 0.6 and 0.7, show various ways of consistent effort and collaborative learning. Students that do extracurricular activities (x7) with a mean of 2.38 and class participation (x8) with a mean of 0.26. This highlights limited engagement in activities outside academics or depending on the individual's circumstances don’t want to participate in class.

Overall, academic preparation like study hours, previous exam scores and homework completion are important when predicting success. The variety of predictors highlights the significant support that is required for various student's needs.

Graphical Summary

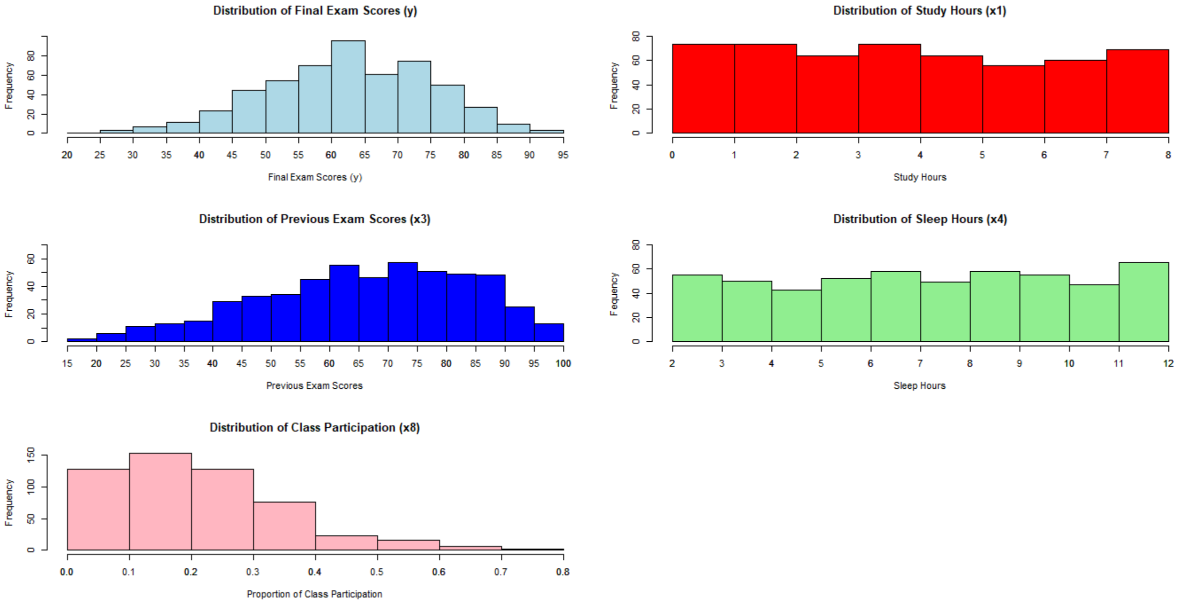


Figure 1: Distribution of Final Exam Scores (y) and the predictive variables (x1, x3, x4 and x8) (Appendix A.2)

Figure 1 shows a visual overview of the dataset's continuous variable distribution, which are shown as histograms to efficiently explain their spread. Final exam scores (y), shown in light blue, and previous exam scores (x3), shown in dark blue, both have a normal distribution which indicates that students performed near the mean. This highlights that the student's past achievements are consistent.

Study hours (x1), shown in red, and Sleep hours (x4), shown in green, have a uniform distribution, which indicates that the students are evenly distributed across levels. This variety demonstrates a range of habits from limited study to inconsistent sleep patterns to a more structured approach.

Class Participation (x8) is skewed to the left, which means that not many students participate in class. The low level of participation in class may imply a lack of emphasis on interactive learning or difficulties developing active involvement. The skewness may be due to students prioritising other academic tasks like homework or studying.

Overall, the distribution shows that most students have similar performance levels which may impact their final exam scores.

Correlation Analysis

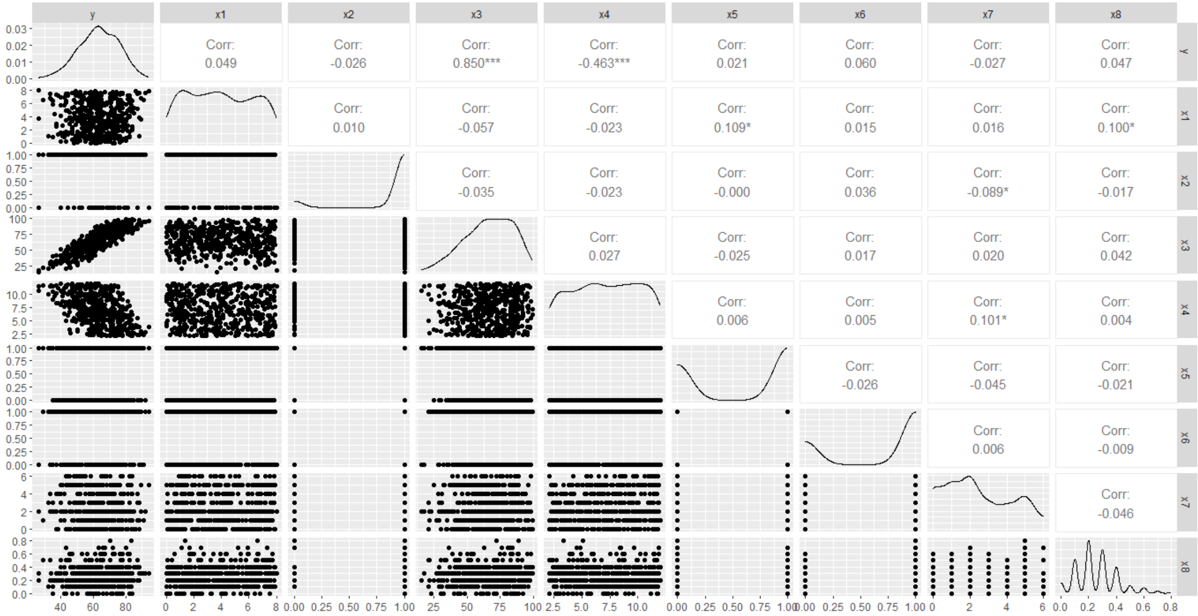


Figure 2 Pair Plot of Variables in the Dataset (Appendix A.3)

Figure 2 is a pair plot and shows the major correlations between variables. Previous Exam Scores (x3) have a strong correlation with final exam scores (y) as 0.850***. This highlights the importance of prior knowledge. Sleep hours (x4) have a negative correlation (-0.463***), highlighting that excessive sleep can affect preparation. Study hours (x1) and homework completion (x5) both have a weak positive correlation which suggests that if the student puts in consistent effort, it can be a positive and have high final grades. However, less consistent efforts can have negative results.

Scatterplots show clear trends for predictors such as x3, whereas weaker predictors are more scattered. Extracurricular activities (x7) and Class Participation have little impact on final exam scores where Figure 2 shows a low correlation value of -0.027 and 0.047 and the scatterplot shows no discernible patterns between x7 or x8 to y. The data points appear randomly distributed, therefore there is no strong positive or negative trend.

Overall, the findings show that academic preparation and previous knowledge are the most important indicators, whereas less relevant predictors might be overlooked in this model.

Model Selection and Validation

Model Construction

A full linear regression model using all predictor variables (x1 to x8) to predict the response variable (y, final exam scores). The regression formula is:

$$y = 35.16 + 0.4339(x1) - 0.3628(x2) - 0.5967(x3) - 2.0459(x4) + 0.9669(x5) + 1.2599(x6) + 0.0296(x7) + 0.5185(x8)$$

Table 2 summarises the regression coefficients for the model predicting final exam scores (y) based on eight predictor variables (x1 to x8). (Appendix A.4)

Predictor	Coefficient	Std. Error	t-value	p-value	Interpretation
Intercept	35.1563	0.5772	60.909	< 2e-16	Baseline score when all predictors = 0
Study Hours (x1)	0.4339	0.0394	11.022	< 2e-16	Positive effect: scores increase by 0.4339 per hour studied
Attendance (x2)	-0.3628	0.2965	-1.226	0.221	Not statistically significant
Previous Scores (x3)	0.5967	0.0051	117.128	< 2e-16	Strong positive effect; higher pervious scores improve outcomes.
Sleep Hours (x4)	-2.0459	0.0310	-65.519	< 2e-16	Negative effect; excessive sleep lower scores.
Homework (x5)	0.9669	0.1868	5.175	3.25 e-07	Completing homework improves scores by 0.9669
Study Groups (x6)	1.2599	0.1990	6.310	5.94e-10	Participation in study groups increases
Extracurricular (x7)	0.0296	0.5049	0.587	0.558	Not statistically significant
Class Participation (x8)	0.5185	0.6149	0.843	0.401	Not statistically significant

Table 3 provides key performance metrics for the regression model for x1 to x8 (Appendix A.4)

Metric	Value
Residual Standard Error	2.097
Multiple R^2	0.9715
Adjusted R^2	0.9711
F-Statistic	2230
p-value	< 2.2e-16

In summary, academic preparation (study hours, previous exam scores, homework and group study participation) has a significant impact on performance, but other elements (attendance, extracurriculars and class participation) have less influence. The model explains most of the fluctuations in exam performance.

Variable Selection

The linear regression model has undergone stepwise. The regression model formula is now:

$$y = 34.98 + 0.4376x1 + 0.5972x3 – 2.0430x4 + 0.9571x5 + 1.249x6$$

Table 4 summaries the final regression model predicting final exam scores (y) based on the most significant predictors identified through stepwise regression (both forwards and backwards). (Appendix A.5)

Predictor	Coefficient	Std. Error	t-value	p-value	Interpretation
Intercept	34.98	0.4743	73.753	<2e-16	Baseline score when predictors = 0
Study Hours (x1)	0.4376	0.0391	11.189	<2e-16	positive effect, +0.4376 per hour
Previous Scores (x3)	0.5972	0.0058	117.470	<2e-16	Strong positive effect
Sleep Hours (x4)	-2.0430	0.0311	-65.795	<2e-16	Negative effect, suggest that excessive sleep may hinder exam preparation.
Homework (x5)	0.9571	0.1846	5.133	1.03e-07	Completing homework improves scores
Study Groups (x6)	1.2494	0.1980	6.310	5.94e-10	Study group participation improves scores

Table 5 summarises the performance metrics of the final regression model. (Appendix A.5)

Metric	Value
Residual Standard Error	2.097
R^2	0.9714
Adjusted R^2	0.9711
F-statistic	3570
p-value	2.2e-16

The refined regression model, obtained through stepwise selection, identifies the most important indicators for final exam scores (y). The new formula simplifies the model while keeping the predictors that have the strongest impact.

Residual Analysis

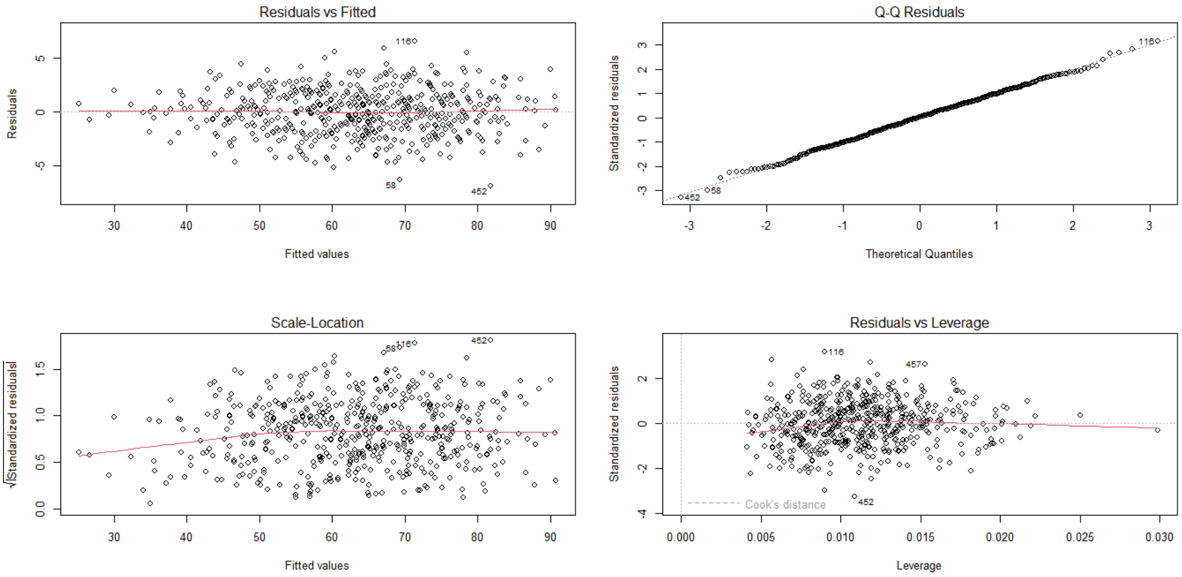


Figure 3 shows 4 different types of graphs of the residuals which evaluates the assumption of linear regression of the model predicting final exam scores. This figure looks at the stepwise regression model. (Appendix A.6)

Model Interpretation

The stepwise regression model is shown in the section, "Variable Selection” and has simplified the model by only having the relevant predictors (x1, x3, x4, x5, x6). The model excludes predictors such as attendance (x2), extracurricular activities (x7) and class participation (x8) to improve interpretability. The predictors that are in the stepwise regression model all have a role in emphasising preparation however sleep hours have a negative effect, which emphasises the importance of balance. The stepwise model is more concise and practical while maintaining precise predictions.

Residual analysis (figure 3) supports the model’s validity. The Residuals vs Fitted plot demonstrates that the linearity assumption has been satisfied, whilst the Q-Q plot suggests that the residuals are approximately normal. The Scale-Location plot indicates minor heteroscedasticity, whereas the Residuals vs Leverage plot shows no extremely influential spots. Addressing modest heteroscedasticity may help to increase dependability even more. Overall, the model emphasises the relevance of study habits, past academic achievement and collaborative learning in test performance, with some surprising findings, such as the negative influence of sleep, requiring additional investigation.

Model Prediction

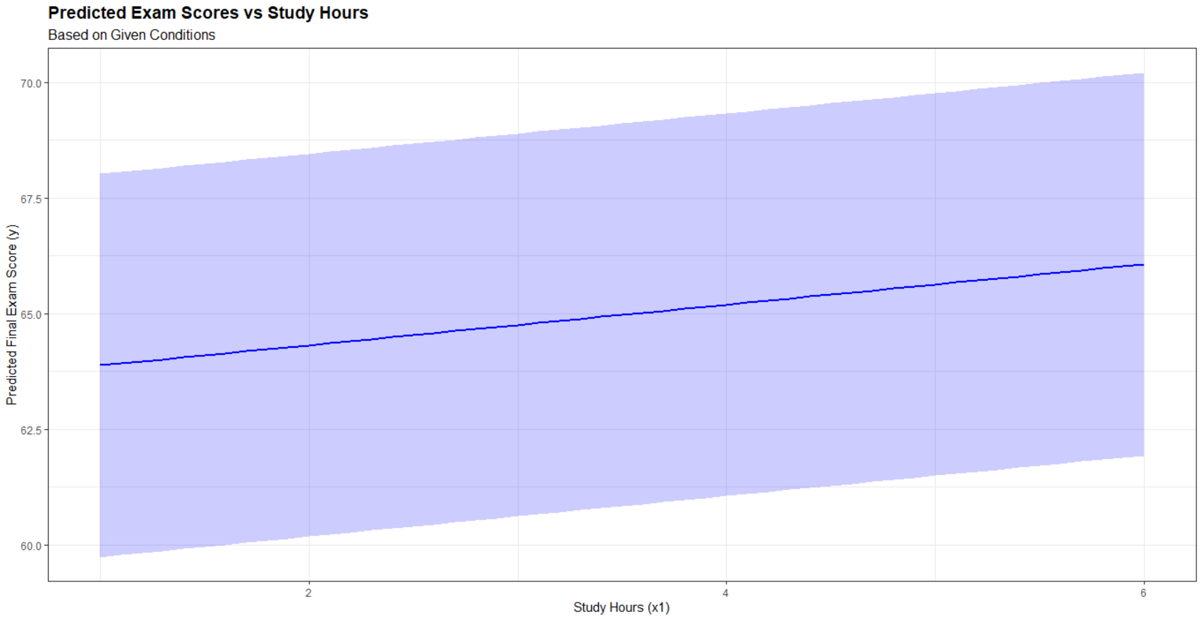


Figure 4 shows the predicted final exam scores (y) for a student studying between 1 and 6 hours per week under specific conditions. (Appendix A.7)

This figure shows how study hours (x1) affect predicted final exam scores (y) under specific conditions. The student attends more than 80% of classes, has a previous exam score of 70, sleeps 7 hours before the exam, does not participate in study groups, completes homework assignments, participates in 2 extracurricular activities and has a class participation of 0.2.

The blue line represents the predicted relationship between study hours and exam scores, and has a clear positive slope, showing that students who study are more likely to perform better on their exams. The more hours the student studies the predicted score will increase which demonstrates the importance of dedicating time to study so that performance can increase.

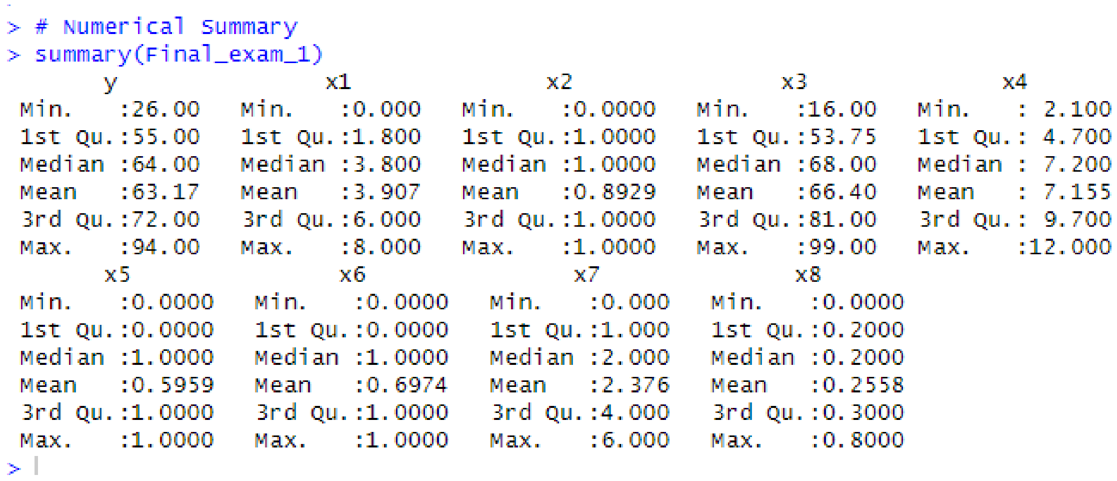
The shaded area around the line represents the prediction interval, which accounts for the model’s uncertainty. It shows a range of possible exam results for the number of hours. As study hours increase the gap widens a little which indicates greater uncertainty in predictions for individuals who study more.

Overall, the figure showed the strong positive impact of study hours on exam performance while controlling for uncertainty caused by other factors such as previous exam scores and homework completion. It visualises the model’s predictions under specific conditions.

Script

Appendix A.1

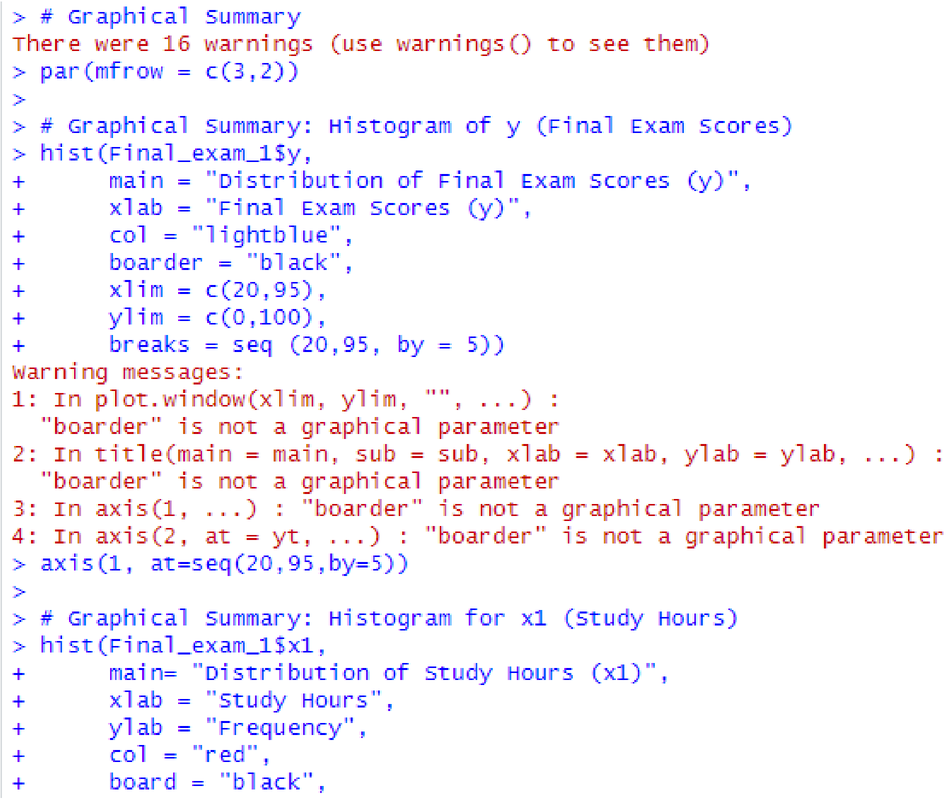
This is the r code for the Numerical Summary and the results were added to Table 1 in the Numerical Summary section of the report.



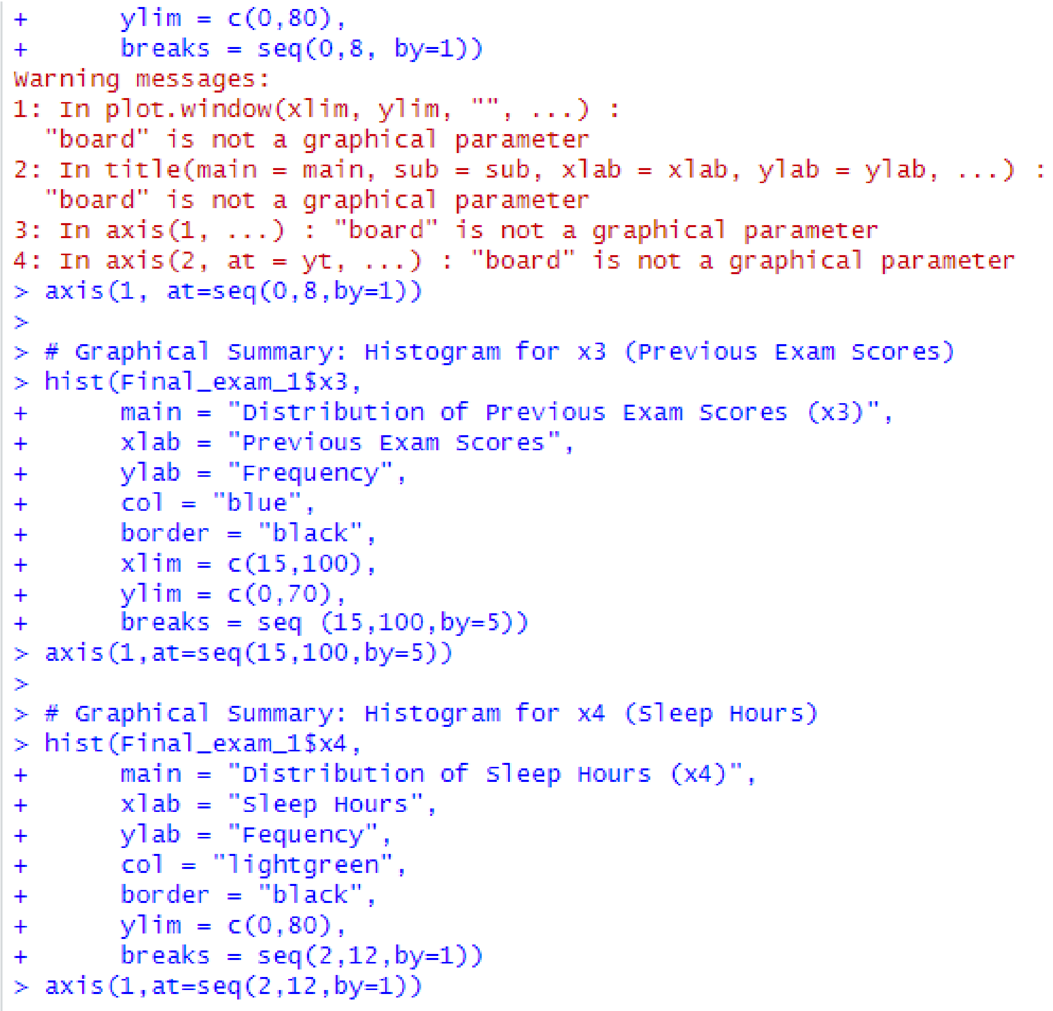
```
> # Numerical Summary
> summary(Final_exam_1)
       y               x1              x2                x3               x4        
 Min.   :26.00   Min.   :0.000   Min.   :0.0000   Min.   :16.00   Min.   : 2.100  
 1st Qu.:55.00   1st Qu.:1.800   1st Qu.:1.0000   1st Qu.:53.75   1st Qu.: 4.700  
 Median :64.00   Median :3.800   Median :1.0000   Median :68.00   Median : 7.200  
 Mean   :63.17   Mean   :3.907   Mean   :0.8929   Mean   :66.40   Mean   : 7.155  
 3rd Qu.:72.00   3rd Qu.:6.000   3rd Qu.:1.0000   3rd Qu.:81.00   3rd Qu.: 9.700  
 Max.   :94.00   Max.   :8.000   Max.   :1.0000   Max.   :99.00   Max.   :12.000  
       x5                x6               x7              x8        
 Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.2000  
 Median :1.0000   Median :1.0000   Median :2.000   Median :0.2000  
 Mean   :0.5959   Mean   :0.6974   Mean   :2.376   Mean   :0.2558  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:0.3000  
 Max.   :1.0000   Max.   :1.0000   Max.   :6.000   Max.   :0.8000  
> 
```

Appendix A.2

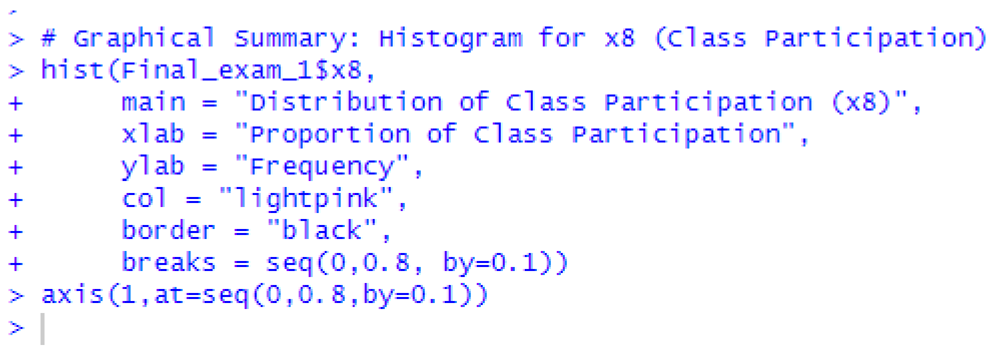
This is the r code for the Graphical Summary Section to make histograms of some predictors.



```
> # Graphical Summary
There were 16 warnings (use warnings() to see them)
> par(mfrow = c(3,2))
> 
> # Graphical Summary: Histogram of y (Final Exam Scores)
> hist(Final_exam_1$y,
+      main = "Distribution of Final Exam Scores (y)",
+      xlab = "Final Exam Scores (y)",
+      col = "lightblue",
+      boarder = "black",
+      xlim = c(20,95),
+      ylim = c(0,100),
+      breaks = seq (20,95, by = 5))
Warning messages:
1: In plot.window(xlim, ylim, "", ...) :
  "boarder" is not a graphical parameter
2: In title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...) :
  "boarder" is not a graphical parameter
3: In axis(1, ...) : "boarder" is not a graphical parameter
4: In axis(2, at = yt, ...) : "boarder" is not a graphical parameter
> axis(1, at=seq(20,95,by=5))
> 
> # Graphical Summary: Histogram for x1 (Study Hours)
> hist(Final_exam_1$x1,
+      main= "Distribution of Study Hours (x1)",
+      xlab = "Study Hours",
+      ylab = "Frequency",
+      col = "red",
+      board = "black",
```

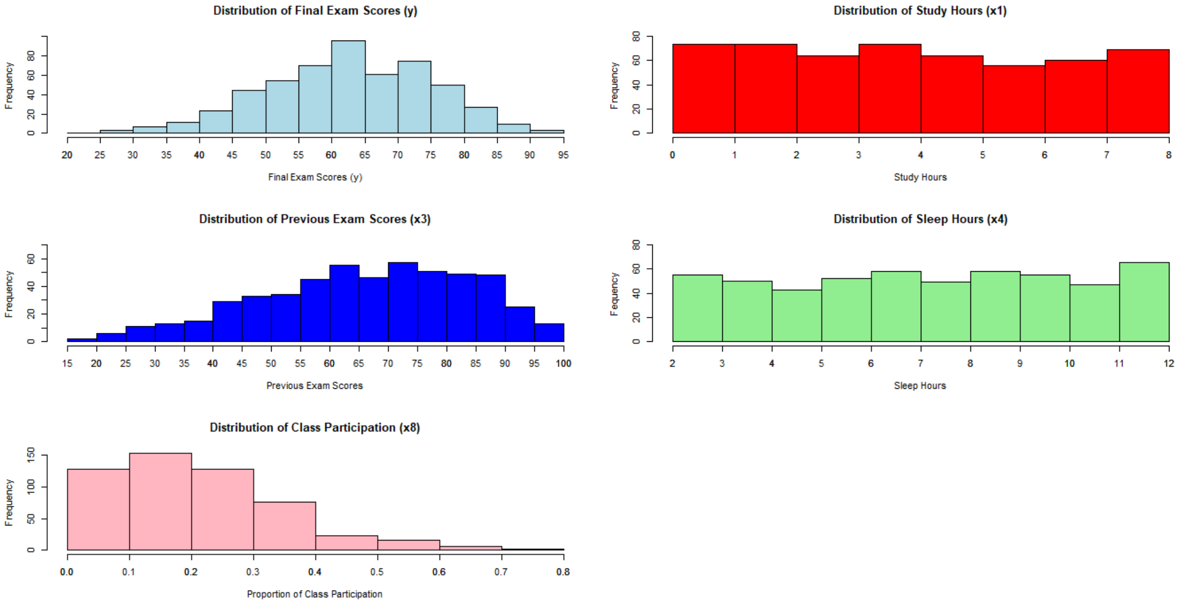


```
+       ylim = c(0,80),
+       breaks = seq(0,8, by=1))
Warning messages:
1: In plot.window(xlim, ylim, "", ...) :
  "board" is not a graphical parameter
2: In title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...) :
  "board" is not a graphical parameter
3: In axis(1, ...) : "board" is not a graphical parameter
4: In axis(2, at = yt, ...) : "board" is not a graphical parameter
> axis(1, at=seq(0,8,by=1))
>
> # Graphical Summary: Histogram for x3 (Previous Exam Scores)
> hist(Final_exam_1$x3,
+       main = "Distribution of Previous Exam Scores (x3)",
+       xlab = "Previous Exam Scores",
+       ylab = "Frequency",
+       col = "blue",
+       border = "black",
+       xlim = c(15,100),
+       ylim = c(0,70),
+       breaks = seq (15,100,by=5))
> axis(1,at=seq(15,100,by=5))
>
> # Graphical Summary: Histogram for x4 (Sleep Hours)
> hist(Final_exam_1$x4,
+       main = "Distribution of Sleep Hours (x4)",
+       xlab = "Sleep Hours",
+       ylab = "Fequency",
+       col = "lightgreen",
+       border = "black",
+       ylim = c(0,80),
+       breaks = seq(2,12,by=1))
> axis(1,at=seq(2,12,by=1))
```



```
> # Graphical Summary: Histogram for x8 (Class Participation)
> hist(Final_exam_1$x8,
+       main = "Distribution of Class Participation (x8)",
+       xlab = "Proportion of Class Participation",
+       ylab = "Frequency",
+       col = "lightpink",
+       border = "black",
+       breaks = seq(0,0.8, by=0.1))
> axis(1,at=seq(0,0.8,by=0.1))
>
```

This code produced the figure:



Distribution of Final Exam Scores (y)

Distribution of Study Hours (x1)

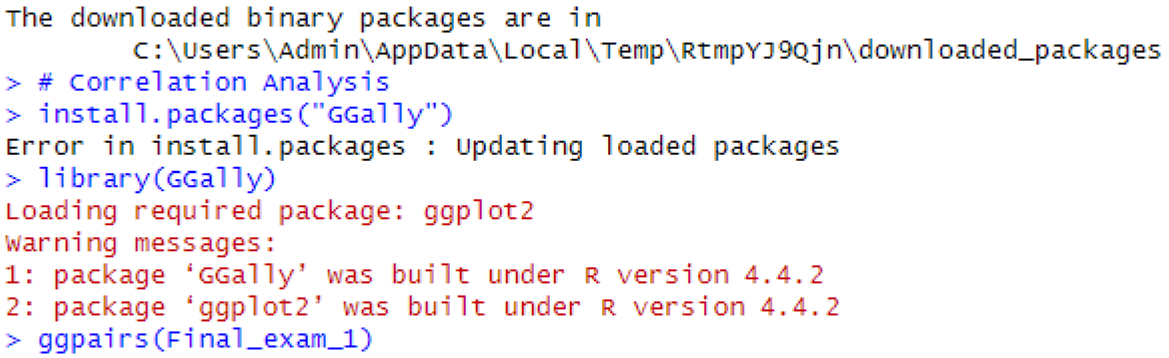
Distribution of Previous Exam Scores (x3)

Distribution of Sleep Hours (x4)

Distribution of Class Participation (x8)

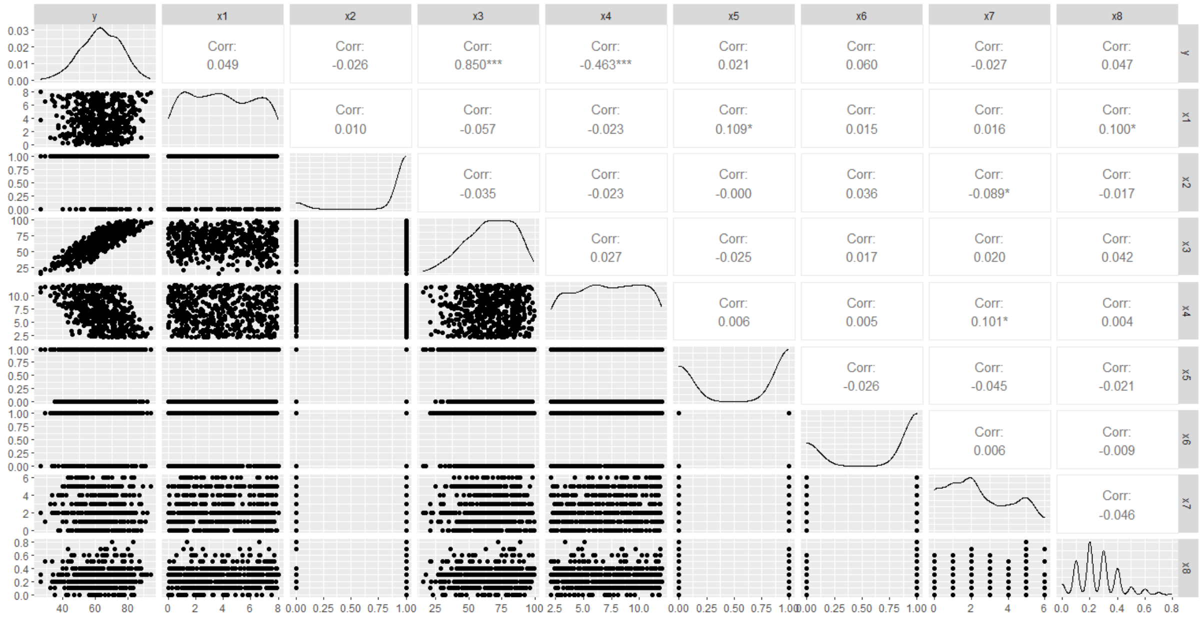
Appendix A.3

This is the r code for the section Correlation Analysis.



```
The downloaded binary packages are in
        C:\Users\Admin\AppData\Local\Temp\RtmpYJ9Qjn\downloaded_packages
> # Correlation Analysis
> install.packages("GGally")
Error in install.packages : Updating loaded packages
> library(GGally)
Loading required package: ggplot2
Warning messages:
1: package ‘GGally’ was built under R version 4.4.2
2: package ‘ggplot2’ was built under R version 4.4.2
> ggpairs(Final_exam_1)
```

This code produced the figure:

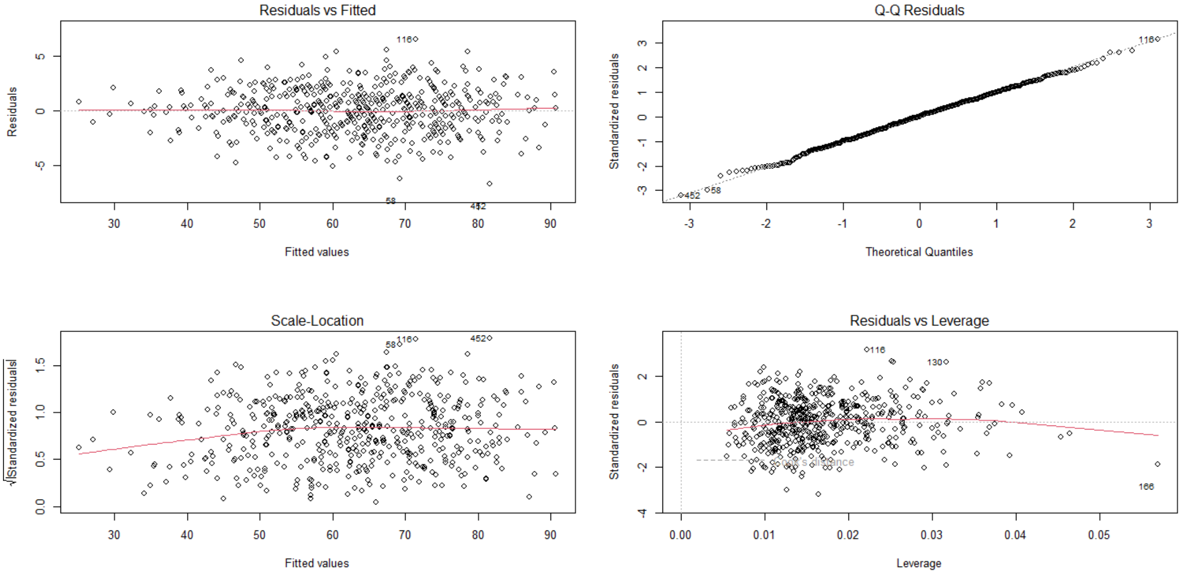


Appendix A.4

This r code is for the section Model Construction. This looks at the linear regression model with more predictors being changed.

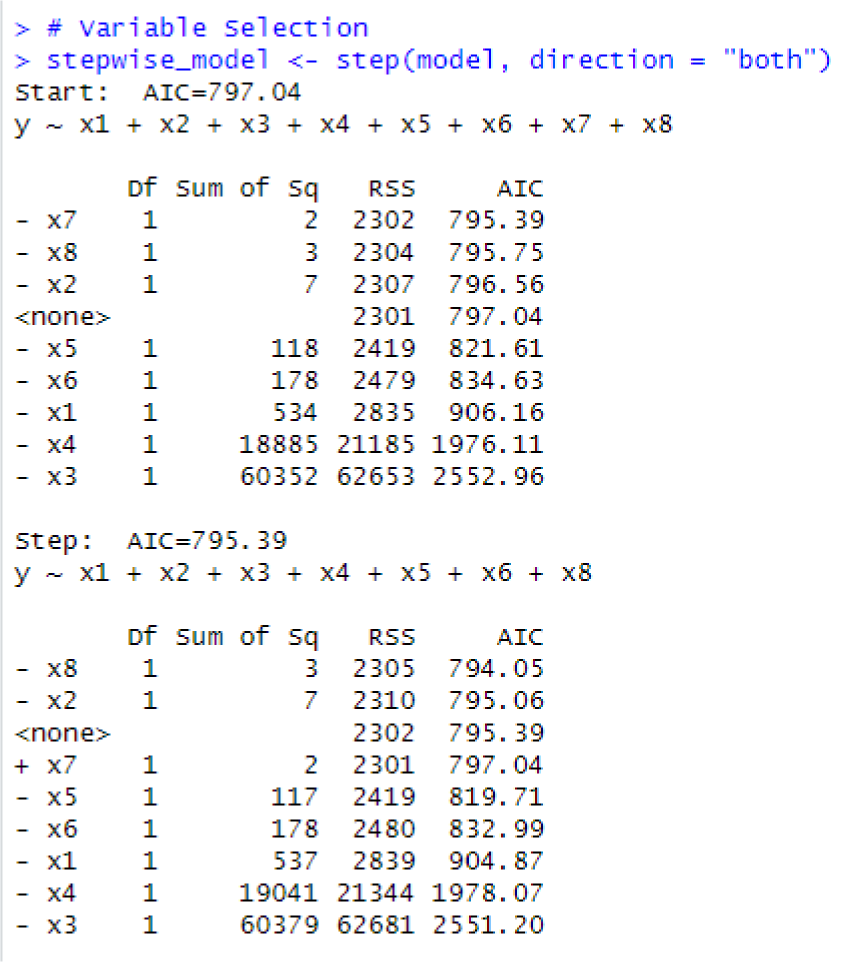
```
> # Model Construction
> model <- lm(y ~ x1+x2+x3+x4+x5+x6+x7+x8, data = Final_exam_1)
> summary(model)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8, data = Final_exam_1)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.6680 -1.4358  0.0364  1.4249  6.5629 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 35.156334   0.577193  60.909  < 2e-16 ***
x1           0.433853   0.039361  11.022  < 2e-16 ***
x2          -0.362380   0.295645  -1.226    0.221    
x3           0.596714   0.005095 117.128  < 2e-16 ***
x4          -2.045885   0.031226 -65.519  < 2e-16 ***
x5           0.966931   0.186839   5.175 3.25e-07 ***
x6           1.259974   0.198225   6.356 4.50e-10 ***
x7           0.029618   0.050489   0.587    0.558    
x8           0.518556   0.618485   0.838    0.402    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.097 on 523 degrees of freedom
Multiple R-squared:  0.9715,	Adjusted R-squared:  0.9711 
F-statistic:  2230 on 8 and 523 DF,  p-value: < 2.2e-16

> par(mfrow = c(2,2))
> plot(model)
```

This r code produces a figure, which is not shown in the report.

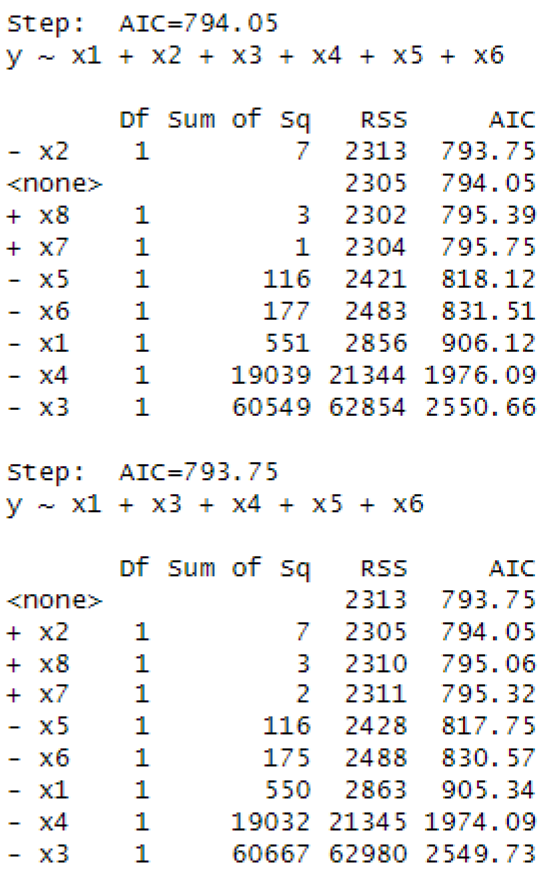


<u>Appendix A.5</u>

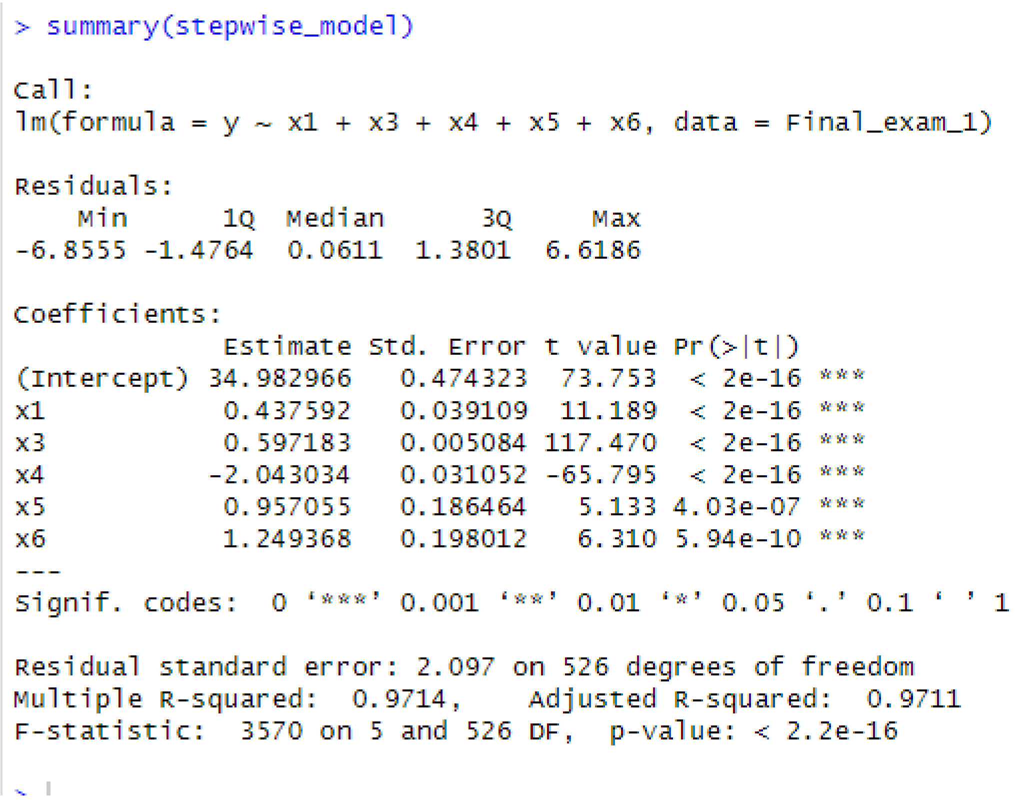
This r code is done for the section Variable Selection, so a stepwise regression model is produced. This is the r code and its output.



```
> # Variable Selection
> stepwise_model <- step(model, direction = "both")
Start:  AIC=797.04
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8

       Df Sum of Sq   RSS     AIC
- x7    1         2  2302  795.39
- x8    1         3  2304  795.75
- x2    1         7  2307  796.56
<none>               2301  797.04
- x5    1       118  2419  821.61
- x6    1       178  2479  834.63
- x1    1       534  2835  906.16
- x4    1     18885 21185 1976.11
- x3    1     60352 62653 2552.96

Step:  AIC=795.39
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x8

       Df Sum of Sq   RSS     AIC
- x8    1         3  2305  794.05
- x2    1         7  2310  795.06
<none>               2302  795.39
+ x7    1         2  2301  797.04
- x5    1       117  2419  819.71
- x6    1       178  2480  832.99
- x1    1       537  2839  904.87
- x4    1     19041 21344 1978.07
- x3    1     60379 62681 2551.20
```



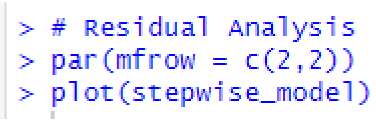
```
Step:  AIC=794.05
y ~ x1 + x2 + x3 + x4 + x5 + x6

       Df Sum of Sq   RSS     AIC
- x2    1         7  2313  793.75
<none>               2305  794.05
+ x8    1         3  2302  795.39
+ x7    1         1  2304  795.75
- x5    1       116  2421  818.12
- x6    1       177  2483  831.51
- x1    1       551  2856  906.12
- x4    1     19039 21344 1976.09
- x3    1     60549 62854 2550.66

Step:  AIC=793.75
y ~ x1 + x3 + x4 + x5 + x6

       Df Sum of Sq   RSS     AIC
<none>               2313  793.75
+ x2    1         7  2305  794.05
+ x8    1         3  2310  795.06
+ x7    1         2  2311  795.32
- x5    1       116  2428  817.75
- x6    1       175  2488  830.57
- x1    1       550  2863  905.34
- x4    1     19032 21345 1974.09
- x3    1     60667 62980 2549.73
```



```
> summary(stepwise_model)

Call:
lm(formula = y ~ x1 + x3 + x4 + x5 + x6, data = Final_exam_1)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.8555 -1.4764  0.0611  1.3801  6.6186 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.982966   0.474323  73.753  < 2e-16 ***
x1           0.437592   0.039109  11.189  < 2e-16 ***
x3           0.597183   0.005084 117.470  < 2e-16 ***
x4          -2.043034   0.031052 -65.795  < 2e-16 ***
x5           0.957055   0.186464   5.133 4.03e-07 ***
x6           1.249368   0.198012   6.310 5.94e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.097 on 526 degrees of freedom
Multiple R-squared:  0.9714,	Adjusted R-squared:  0.9711 
F-statistic:  3570 on 5 and 526 DF,  p-value: < 2.2e-16

>
```

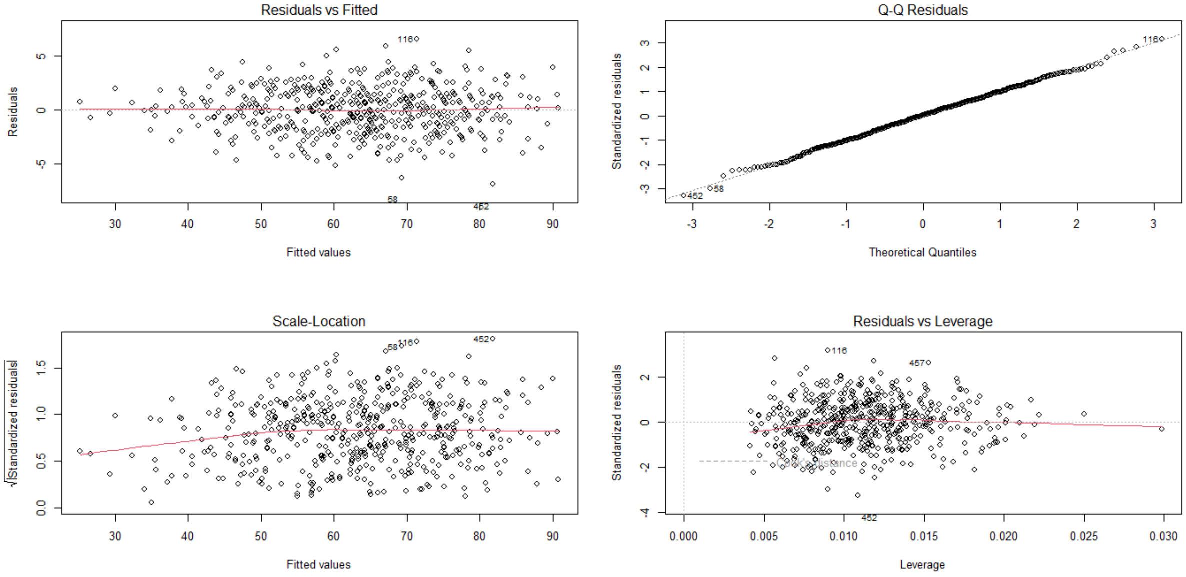
Appendix A.6

This R code is for Residual Analysis Section.



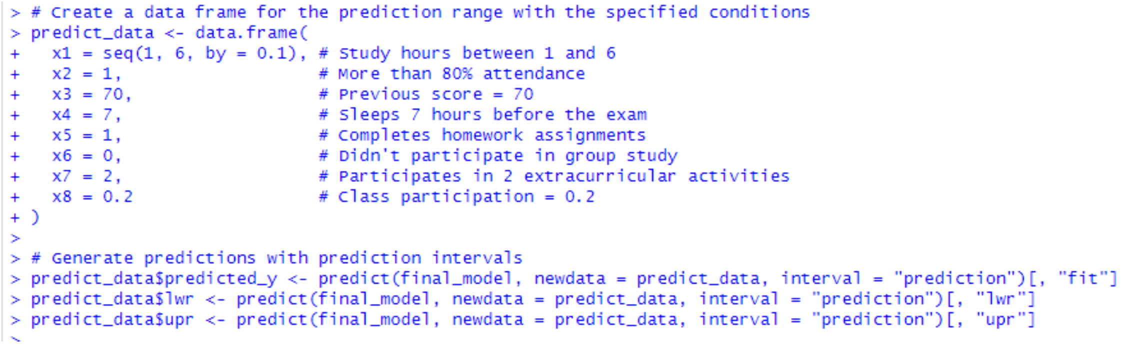
```
> # Residual Analysis
> par(mfrow = c(2,2))
> plot(stepwise_model)
```

This r code produces a figure which is shown as figure 3 in the report.

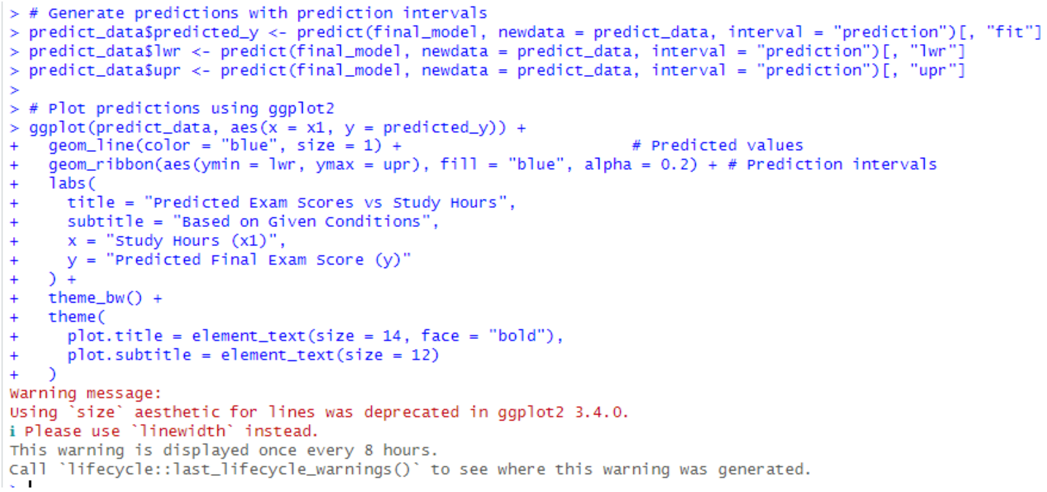


Appendix A.7

This is the r code for Model Prediction.

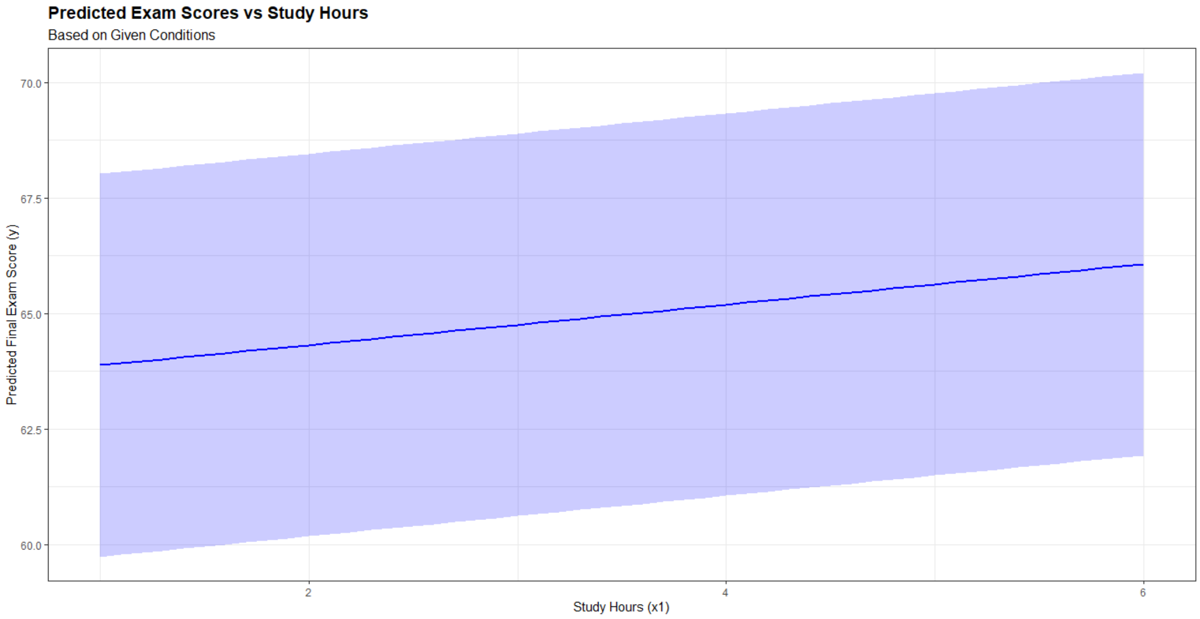


```
> # Create a data frame for the prediction range with the specified conditions
> predict_data <- data.frame(
+   x1 = seq(1, 6, by = 0.1), # Study hours between 1 and 6
+   x2 = 1,                   # More than 80% attendance
+   x3 = 70,                  # Previous score = 70
+   x4 = 7,                   # Sleeps 7 hours before the exam
+   x5 = 1,                   # Completes homework assignments
+   x6 = 0,                   # Didn't participate in group study
+   x7 = 2,                   # Participates in 2 extracurricular activities
+   x8 = 0.2                  # Class participation = 0.2
+ )
>
> # Generate predictions with prediction intervals
> predict_data$predicted_y <- predict(final_model, newdata = predict_data, interval = "prediction")[, "fit"]
> predict_data$lwr <- predict(final_model, newdata = predict_data, interval = "prediction")[, "lwr"]
> predict_data$upr <- predict(final_model, newdata = predict_data, interval = "prediction")[, "upr"]
```



```
> # Generate predictions with prediction intervals
> predict_data$predicted_y <- predict(final_model, newdata = predict_data, interval = "prediction")[, "fit"]
> predict_data$lwr <- predict(final_model, newdata = predict_data, interval = "prediction")[, "lwr"]
> predict_data$upr <- predict(final_model, newdata = predict_data, interval = "prediction")[, "upr"]
>
> # Plot predictions using ggplot2
> ggplot(predict_data, aes(x = x1, y = predicted_y)) +
+   geom_line(color = "blue", size = 1) +                        # Predicted values
+   geom_ribbon(aes(ymin = lwr, ymax = upr), fill = "blue", alpha = 0.2) + # Prediction intervals
+   labs(
+     title = "Predicted Exam Scores vs Study Hours",
+     subtitle = "Based on Given Conditions",
+     x = "Study Hours (x1)",
+     y = "Predicted Final Exam Score (y)"
+   ) +
+   theme_bw() +
+   theme(
+     plot.title = element_text(size = 14, face = "bold"),
+     plot.subtitle = element_text(size = 12)
+   )
Warning message:
Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.
This warning is displayed once every 8 hours.
Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

The r code produced a figure:


Predicted Exam Scores vs Study Hours
Based on Given Conditions
Predicted Final Exam Score (y)
Study Hours (x1)