Assessment 2 - Individual Project Report PDF

by Sagari Muraliegaran

Submission date: 31-Mar-2025 01:31PM (UTC+0100)

Submission ID: 254310701

File name: 12762_Sagari_Muraliegaran_Assessment_2_-_Individual_Project_Report_PDF_339127_800733062.pdf (984.04K)

Word count: 3210

Character count: 17729

Introduction

Globally, adolescent depression is becoming a more serious public health issue, especially in low- and middle-income countries with inadequate mental health services. Due to the major biological, psychological and social changes that occur throughout adolescence, young people are more susceptible to mental health conditions like depression. Global estimates indicate that between 10 and 20 percent of teenagers suffer from mental health issues, however, the majority of these instances go undetected and untreated. Adolescent depression can result in substance misuse, social isolation, poor academic achievement, and long-term mental health problems as an adult.

This study is based on the research article by Shimelis Girma (*Depression and Its Determinants among Adolescents in Jimma Town, Southwest Ethiopia | PLOS One* n.d.), which examined the prevalence and contributing factors of depression among teenagers enrolled in school in Jimma town, Southwest Ethiopia, served as the basis for this report. The study examined characteristics linked to depression as determined by the PHQ-9 questionnaire using a range of sociodemographic and psychological variables.

The file cutdata.xlsx, which was provided in the Moodle page, contains a subset of the original research data used for this analysis. Age, sex, grade level, kind of school (private or public), area of residence (rural or urban), self-reported health condition, Oslo social support score, body mass index (BMI) and PHQ-9 depression score are among the factors that make up the dataset, which contains data on 546 teenagers.

The report's main objective is to examine the predictive power of the other factors in the dataset for depression scores (PHQ-9). The goal of the investigation is to ascertain whether a trustworthy prediction model can be developed and whether sociodemographic or health-related characteristics are most closely linked to teenage depression.

Good

Exploratory Analysis through Descriptive Statistics and Graphical Summaries

Summary of Categorical Variables

In this section, the dataset summary provides a brief synopsis of each variable's distribution, range, and centre values (mean and median). It guarantees that the proper statistical techniques are applied for additional analysis and assists in identifying problems such as skewness or outliers. The r code for this table can be found in Appendix 2.

Table 1 summarises the statistics for continuous variables in the dataset. These data reveal the distribution and central tendency of each variable within the study's teenage sample.

	Age	Health Status	Social Support	BMI	Depression Score
Min	14.00	1.000	3.000	15.04	0.000
1st Qu	16.00	3.000	8.000	18.40	3.250
Median	17.00	4.000	10.000	19.65	6.000
Mean	16.83	3.822	9.901	19.95	6.769
3rd Qu	18.00	5.000	12.000	21.11	10.000
Max	19.00	5.000	14.000	30.85	15.000

With an average age of 16.8, the summary table reveals that the majority of participants are between the ages of 14 and 19. With medians of 4 and 10, respectively, social support and health status are relatively favourable. Depression scores indicate mild to moderate symptoms on average, whereas BMI readings are generally within a healthy range. This summary highlights attention to individual differences and aids in directing more research.

Histograms

This section's histograms depict the distribution of teenage depression score, BMI and social support. They aid in the visualisation of patterns like skewness or clustering. Aspects such as weight status, depression intensity, and perceived support are highlighted differently in each histogram. The R codes for the three histograms can be found in Appendix 3.

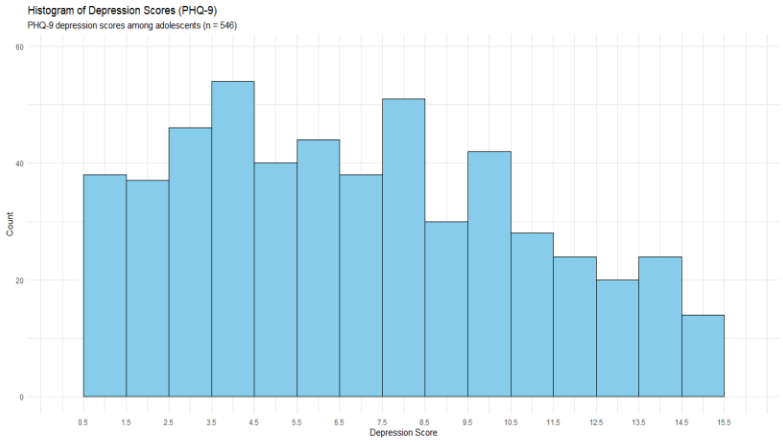


Figure 1 shows a histogram and the distribution of PHQ-9 depression scores among adolescents (n = 546). With some concentration around scores of 4-5 and 7-8, the histogram shows a rather uniform distribution of PHQ-9 depression scores. This implies that a large number of the sample's teenagers suffer from mild to moderate depression. Fewer people reported greater depression ratings, indicating a relatively positively skewed distribution. The histogram's overall indication of significant heterogeneity in depression symptoms supports the necessity of investigating potential influencing variables.

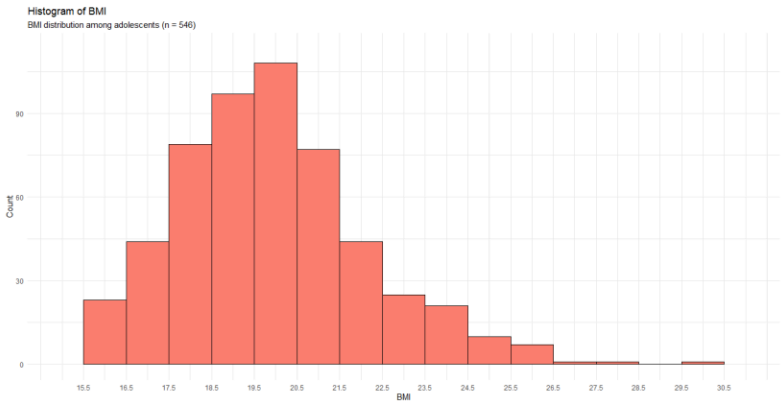


Figure 2 shows a histogram and its distribution of Body Mass Index (BMI) among adolescents (n = 546). With a distinct peak at 20-21, the histogram indicates that the majority of the sample's teenagers fall within a healthy weight range, with the majority having a BMI between 17.5 and 22.5. Fewer people had higher BMIs over 25, indicating a positively skewed distribution. This histogram emphasises that although the majority of the population has a normal BMI, a small percentage maybe overweight, which should be taken into account when comparing it to other factors like depression or health status.

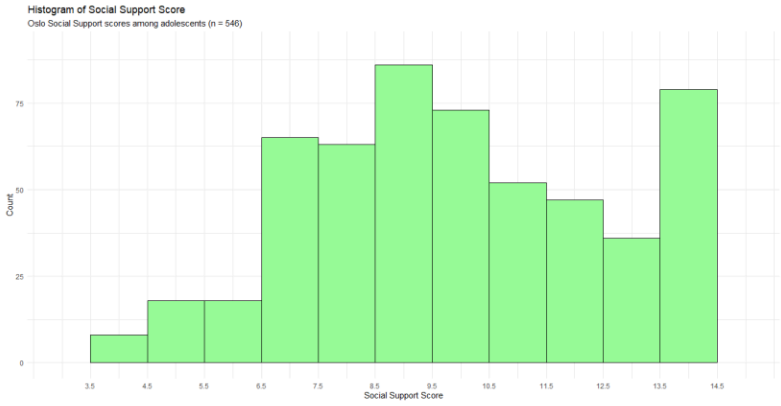


Figure 3 shows a histogram and its distribution of Oslo Social Support Scores among adolescents (n = 546). With clear peaks around scores of 9 and 14 show that most teenagers report moderate to high levels of social support. The majority of participants report feeling supported to some degree, as seen by the very small number of students who report an extremely poor support rating (below 6). There appear to be two typical degrees of perceived support in the sample, as indicated by the

somewhat bimodal distribution. When examining the relationship between social support and depression or other measures of well-being, this difference could be significant.

Boxplots

Boxplots allow the visual compassion of depression scores among groups, by displaying the median and distribution of the data. To illustrate potential disparities in depression levels among teenagers, the following charts compare PHQ-9 scores by sex, school grade and residence. The R code for the boxplots can be seen in Appendix 4.

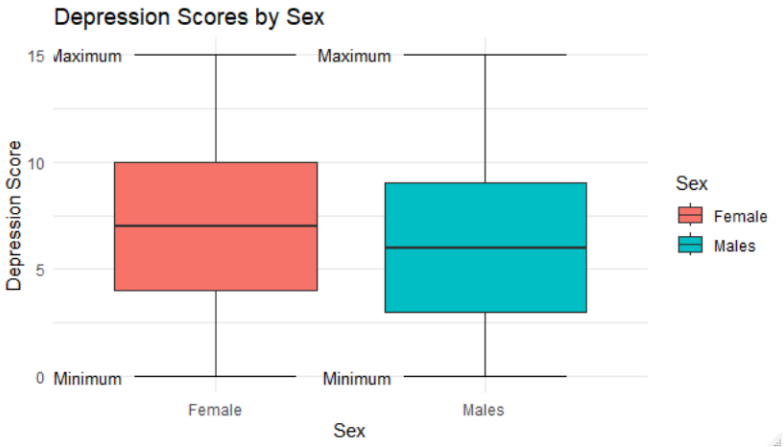


Figure 4 shows boxplots and its distribution of PHQ-9 depression scores by sex. Female and male categories are compared to examine differences in depression levels. For females, they have a higher median and somewhat larger interquartile range (IQR); the boxplot indicates that women typically have slightly greater depression scores than men. The minimum (0) and maximum (15) scores are the same for both sexes, but the distribution for females is somewhat higher, indicating that women would generally have more depressed symptoms. This boxplot comparison draws attention to a possible gender disparity in depression that might be investigated further using statistical modelling.

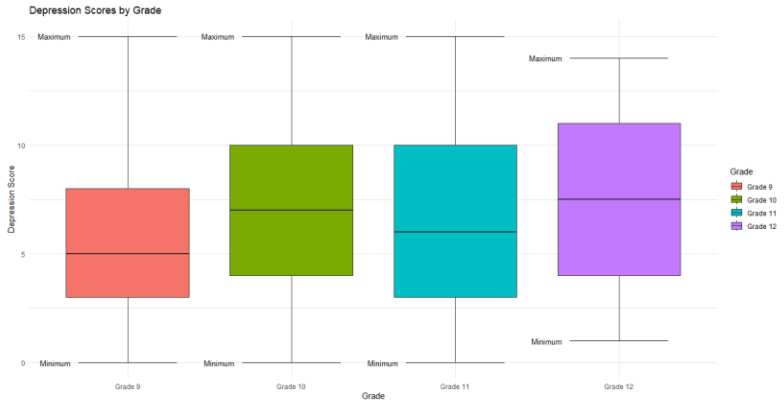


Figure 5 shows boxplots and its distribution of PHQ-9 depression scores across school grades (Grade 9 to Grade 12). According to the boxplot, depression scores often rise somewhat with grade level. When compared to other grades, students in Grade 12 exhibit a greater range of values and the highest median depression score. While Grade 9 had the lowest median and a tighter range, suggesting usually lower depression scores, Grades 10 and 11 also have quite high IQRs. This pattern could point to an increase in depression symptoms as students advance academically, possibly as a result of growing pressure and stress in higher grades.

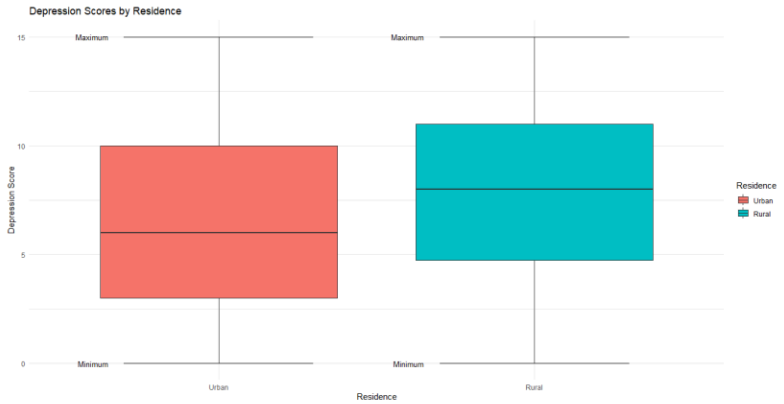


Figure 6 shows boxplots and the distribution of PHQ-9 depression scores by place of residence (Urban vs Rural). According to the boxplot, teenagers from rural areas often score higher on depression than those who live in urban areas. There appears to be greater variation in depressed symptoms among rural students, as seen by the higher median and IQR. Although the overall score range for both groups is the same (0-15), the higher central tendency among teenagers in rural areas may indicate possible discrepancies in mental health, maybe as a result of different access to services or support systems.

Correlation Matrix

In this section, it looks at the relationship between continuous variables and helps identify predictors of depression and check for multicollinearity before modelling. (See Appendix 5 for R code).

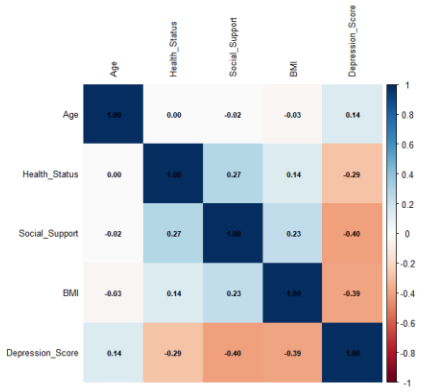


Figure 7 shows a heatmap of relationships between the continuous variables of age, depression score, BMI, social support and health status. There are moderately negative relationships between depression scores and social support (r = -0.40), health status (r = -0.29), and BMI (r = -0.39), indicating that lower depression is associated with greater support, higher BMI, and better health. These variables were added to the regression model as there are no indications of multicollinearity.

Good

Scatter Plots

In this section, scatterplots highlight the relationship between depression scores and continuous predictors. The R code for the scatterplots (Figures 8 and 9) can be seen in Appendix 6.

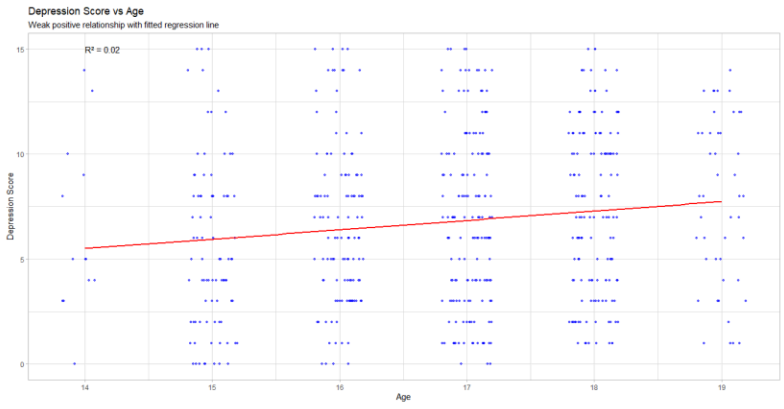


Figure 8 shows a scatter plot and its relationship between age and PHQ-9 depression score among adolescents. The R^2 value (0.02) shows the percentage of variance in depression scores that can be attributed to age, and a red regression line has been drawn to show the trend. The Scatter plot's low R^2 value of 0.02 and the regression line's small upward slope suggest a very faint positive link between age and depression score. This implies that depression scores only marginally rise with age, and that age accounts for a little portion of the variation in individuals' depression levels. Age is not a reliable indicator of depression in this population, as seen by the widely dispersed points.

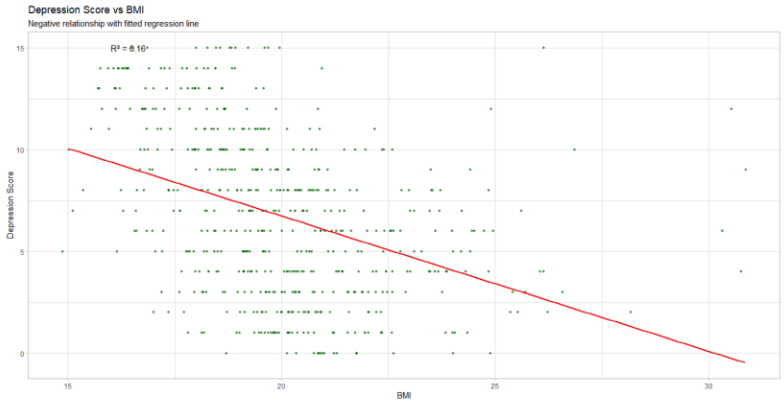


Figure 9 shows a scatterplot and the relationship between BMI and PHQ-9 depression scores among adolescents. The R^2 value (0.16) displays the percentage of variance in depression scores that can be attributed to BMI, while the red regression line denotes the linear trend.
There is a somewhat negative correlation between depression scores and BMI, according to the scatterplot. The regression line slopes downward, indicating that depression scores trend to decline as BMI rises. With an R2 value of 0.16. BMI is a comparatively greater predictor than age, accounting for almost 16% of the variation in depression scores. The data in this dataset indicates that BMI may have a significant impact on teenage mental health, even though the points are still a little dispersed.

OL Good

Statistical Methods

In this study, a full multiple linear regression model was used to examine the relationship between several predictor variables and adolescent depression scores (measured by PHQ-9). The primary model includes the following predictors: Age, Sex, Grade, School Type, Residence, Health Status, Social Support and BMI.

The model aims to explore the hypothesis that psychological, demographic and health related factors might strongly predict depression scores. In particular, we investigate the following theories:

- H_0 (null hypothesis): None if the predictor variables significantly influence depression scores (all regression coefficients = 0)
- H_1 (alternative hypothesis): At least one predictor variable has a significant effect on depression scores.

The R code for the Statistical Methods, which looks at the full multiple regression model, the final regression model, and the residual analysis, can be seen in Appendix 7.

Full Multiple Regression Model

Depression Score = 19.28922 + 0.23435(Age) – 1.03695(Sex [Males]) + 0.55723 (Grade 10) + 0.55800(Grade 11) + 1.02339(Grade 12) + 0.08400(School Type [Governmental]) +1.05995(Residence [Rural]) – 0.64696(Health Status) – 0.43304 (Social Support) – 0.50248(BMI)

Table 2 shows a coefficients table from the full multiple linear regression model predicting depression scores.

Predictor	Estimate	Std. Error	t-value	p-value
(Intercept)	19.28922	3.15719	6.110	1.92e-09
Age	0.23435	0.18796	1.247	0.21303
Sex (Male)	-1.03695	0.29932	-3.464	0.000574
Grade (Grade 10)	0.55723	0.45644	1.221	0.222698
Grade (Grade 11)	0.55800	0.55572	1.004	0.315776
Grade (Grade 12)	1.02339	0.64549	1.585	0.113453
School type (Government)	0.08400	0.37820	0.222	0.824329
Resident (Rural)	1.05995	0.38856	2.728	0.006583
Health Status	-0.64696	0.15586	-4.151	3.85e-05
Social Support	-0.43304	0.05778	-7.495	2.76e-13
BMI	-0.50248	0.06230	-8.065	4.83e-15

Show less decimal places here

Table 3 shows the summary metrics for the multiple linear regression model predicting depression scores.

Metric	Value
Residual Std. Error	3.341
Degrees of Freedom	535
Multiple R-squared	0.3261
Adjusted R-squared	0.3135
F-statistic	25.89 on 10 and 535 DF
Overall Model p-value	< 2.2e-16

Males, those with higher BMI, greater social support and better health all had lower depression levels, according to the entire regression model, but living in a rural area was associated with higher scores. There was no significant difference in age, grade or school type. Overall, the model was statistically significant and represented 32.6% of the variation in depression score.

Residual Analysis

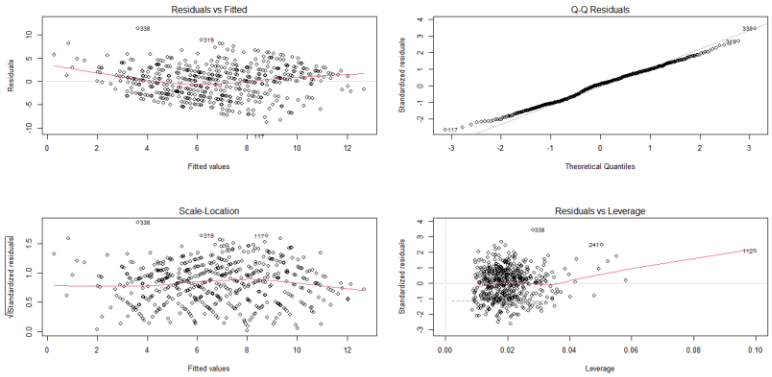


Figure 10 shows residual analysis for the multiple linear regression model predicting depression score. The plots suggest that most of the linear regressions are satisfied. Although a little curvature suggests minor non-linearity, residuals show up randomly distributed about zero, confirming linearity. With very slight variations in the tails, the Q-Q plot displays the residuals' approaching normalcy. Although a modest funnel shape suggests mild heteroscedasticity, the Scale-Location plot shows rather steady variance. While most findings are not very powerful, some could be and should be investigated further.

Shapiro-wilk normality test	
w	0.99401
p-value	0.03001

Given the large sample size and near-normal Q-Q plot, the Shapiro-Wilk test (p = 0.030) indicates a small non-normality in the residuals, but this is unlikely to have an impact on the validity of the model.

Good

Final Multiple Regression Model

To get to the final Multiple Regression Model, a stepwise regression with AIC was used to remove variables that did not significantly contribute to the model's development and prevent overfitting. By penalising superfluous predictors, AIC ensures a balance between accuracy and complexity and encourages model simplicity. Stepwise regression was suitable for choosing the significant variables since the correlation matrix did not exhibit multicollinearity. A more effective and understandable model of teenage depression was made possible by this data-driven approach.

Depression Score = 16.1022 + 0.4659(Age) – 1.0291(Sex[Males]) + 1.0435(Residence[Rural]) – 0.6361(Health Status) – 0.4315(Social Support) – 0.5135(BMI) | AIC = 1323.03

A number of variables showed a strong correlation with depression scores in the final model. Students from rural locations had greater depression levels than those from urban areas, while males scored lower than females. Lower depression scores were associated with higher BMI, better health and more social support, suggesting their protective function. Age was still included in the model, but it had a negligible and non-statistically significant impact. Using stepwise regression and the Akaike Information Criterion (AIC), the model was optimised by reducing overfitting and choosing the most relevant predictors.

Results and Conclusions

According to the investigation, many variables have a strong correlation with adolescent depression scores. These variables may have protective benefits, as seen by the lower depression scores among males, those with higher BMI, better self-reported health and stronger social support. Students who lived in the rural area, on the other hand, scored higher, suggesting a potential disadvantage in terms of mental health outcomes.

Using stepwise regression with AIC, the final multiple linear regression model was chosen. It was statistically significant ($p < 2.2e\text{-}16$) and explained approximately 32.6% of the variation in depression scores ($R^2 = 0.3261$). This suggests that additional unmeasured factors may contribute to teenage depression, even if it also shows that the model predicts depression scores with a moderate level of accuracy.

The model is useful for determining important risk and protective factors, but the results should be taken cautiously. In conclusion, at least one predictor significantly influences depression scores in this sample, rejecting the null hypothesis in light of the entire model's significance.

Reflection

To better understand and convey the data, a number of visualisations were made during the investigation. To comprehend distributions and group differences, simple plots like boxplots and histograms were created first. These were then modified for impact and clarity; for instance, histograms were improved by modifying colour schemes, axis intervals and bin sizes, and labels were added to make the lowest and maximum values visible. Annotations showing the upper and lower extremes were added to boxplots to better illustrate the variability among characteristics, including sex, grade and residency. For the scatter plots, a fitted regression line and R2 values were added to help with understanding. The analysis had limits, even of the visualisations helped identify important trends. Only around 32.6% of the variation in depression scores could be explained by the model, suggesting that many affecting factors may go unmatched. Furthermore, the cross-sectional structure of the data restricts the capacity to draw conclusions about causality and residual diagnostics revealed minor departures from normalcy.

OL Good

Reference

Depression and Its Determinants among Adolescents in Jimma Town, Southwest Ethiopia | PLOS One [Online]. Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0250927 [Accessed: 31 March 2025].

Appendix of R code

I have issues downloading my R Markdown in Word, PDF and as HTML, so this is the best option. I do apologise, as I know this is not how it should be presented, but I have no other option.

```
---
title: "Cutdata"
author: "Sagari Muraliegaran"
date: "2025-03-29"
output:
  pdf_document: default
  html_document:
    df_print: paged
  word_document: default
---
```

Appendix 1 – Clean up

Appendix 1 - This section is just making sure all the necessary packages are installed and the data set is labelled properly for the next process.

```
{r}
# Install all required packages (only run once per machine)
install.packages(c("readxl", "dplyr", "ggplot2", "corrplot",
                   "GGally", "scales", "tidyr", "ggthemes"))

# Load required packages
library(readxl)   #for reading Exel files
library(dplyr)    #for data manipulation
library(ggplot2)  #for visualisations
library(corrplot) #for correlation matrix
library(GGally)   #for correlation heatmaps
library(scales)   #for scaling plots
library(tidyr)    #for reshaping data
library(ggthemes) #for cleaner themes

# Load data
data1 <- cutdata <- read_excel("Data Analytics/Communicating and Presentating
Results/CW/Individual Project/cutdata.xlsx")

# Check structure
str(data1)

# Rename columns for clarity
colnames(data1) <- c("ID", "Age", "Sex","Grade", "School_Type", "Residence",
"Health_Status", "Social_Support", "BMI", "Depression_Score")
```

```r
# Convert categorical variables to factor
data1$Sex <- factor(data1$Sex, levels = c(0,1), labels = c("Female", "Males"))
data1$School_Type <- factor(data1$School_Type, levels = c(0,1), labels = c
("Private", "Governmental"))
data1$Residence <- factor(data1$Residence, levels = c(0,1), labels = c("Urban",
"Rural"))
data1$Grade <- factor(data1$Grade, levels = 1:4, labels = c("Grade 9", "Grade 10",
"Grade 11", "Grade 12"))
```

```
Warning messages:
1: package 'quanteda' was built under R version 4.4.3
2: package 'topicmodels' was built under R version 4.4.3
Error in install.packages : Updating loaded packages
Warning: package 'readxl' was built under R version 4.4.3Warning: package 'dplyr'
was built under R version 4.4.3
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

Warning: package 'ggplot2' was built under R version 4.4.3Warning: package
'corrplot' was built under R version 4.4.3corrplot 0.95 loaded
Warning: package 'GGally' was built under R version 4.4.3Registered S3 method
overwritten by 'GGally':
  method from
  +.gg   ggplot2
Warning: package 'scales' was built under R version 4.4.3Warning: package 'tidyr'
was built under R version 4.4.3Warning: package 'ggthemes' was built under R
version 4.4.3tibble [546 × 10] (S3: tbl_df/tbl/data.frame)
 $ ID        : num [1:546] 1 2 3 4 5 6 7 8 9 10 ...
 $ AGE       : num [1:546] 15 16 16 15 15 18 16 16 16 16 ...
 $ SEX       : num [1:546] 0 1 0 1 0 0 0 0 0 0 ...
 $ GRADE     : num [1:546] 1 1 1 1 1 2 1 1 1 1 ...
 $ SCHOOLTYPE: num [1:546] 1 1 1 0 1 1 1 1 1 1 ...
 $ RESIDENCE : num [1:546] 0 0 0 0 0 0 0 0 0 0 ...
 $ HEALTH    : num [1:546] 4 4 5 4 4 4 4 3 4 5 ...
 $ OSLO3     : num [1:546] 10 9 7 4 10 5 11 11 14 10 ...
 $ BMI       : num [1:546] 20.3 19.1 19.6 16.3 23.9 ...
 $ DEPSCORE  : num [1:546] 0 4 3 12 5 13 8 8 5 3 ...
```

Appendix 2 – Descriptive Statistics – Descriptive Summary

This is the Descriptive Summary code and outcome, which is shown in Table 1

```{r}
# Simple summary of Continuous variables
summary(data1[, c("Age", "Health_Status", "Social_Support", "BMI",
"Depression_Score")])
```

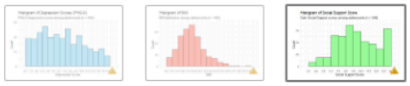
```
      Age        Health_Status   Social_Support        BMI
 Min.   :14.00   Min.   :1.000   Min.   : 3.000   Min.   :15.04
 1st Qu.:16.00   1st Qu.:3.000   1st Qu.: 8.000   1st Qu.:18.40
 Median :17.00   Median :4.000   Median :10.000   Median :19.65
 Mean   :16.83   Mean   :3.822   Mean   : 9.901   Mean   :19.95
 3rd Qu.:18.00   3rd Qu.:5.000   3rd Qu.:12.000   3rd Qu.:21.11
 Max.   :19.00   Max.   :5.000   Max.   :14.000   Max.   :30.85
 Depression_Score
 Min.   : 0.000
 1st Qu.: 3.250
 Median : 6.000
 Mean   : 6.769
 3rd Qu.:10.000
 Max.   :15.000
```

Appendix 3- Graphical Summary: Histogram

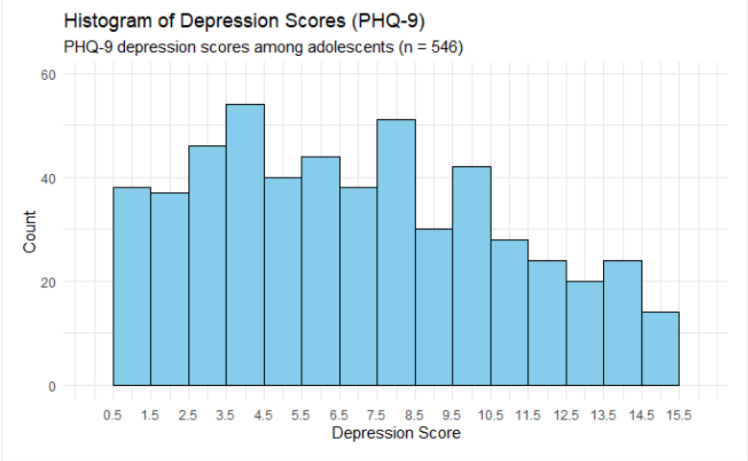
```{r}
# Depression Score Histogram
ggplot(data1, aes(x = Depression_Score)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black", boundary = 0.5,
closed = "left") +
  scale_x_continuous(
    breaks = seq(0.5, max(data1$Depression_Score) + 0.5, by = 1),  # center bins
    limits = c(0, ceiling(max(data1$Depression_Score)) + 1)
  ) +
  scale_y_continuous(
    limits = c(0, max(table(data1$Depression_Score)) + 5)
  ) +
  theme_minimal() +
  labs(
    title = "Histogram of Depression Scores (PHQ-9)",
    subtitle = "PHQ-9 depression scores among adolescents (n = 546)",
    x = "Depression Score",
    y = "Count"
  )
```

```
# BMI Histogram
ggplot(data1, aes(x = BMI)) +
  geom_histogram(binwidth = 1, fill = "salmon", color = "black", boundary = 15.5,
closed = "left") +
  scale_x_continuous(
    breaks = seq(15.5, 30.5, by = 1),  # centered intervals
    limits = c(15, 31)
  ) +
  scale_y_continuous(
    limits = c(0, max(table(round(data1$BMI))) + 5)
  ) +
  theme_minimal() +
  labs(
    title = "Histogram of BMI",
    subtitle = "BMI distribution among adolescents (n = 546)",
    x = "BMI", y = "Count")
```

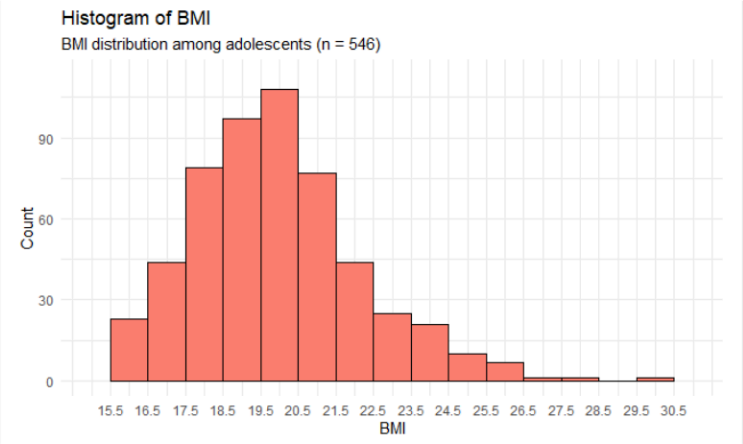
```r
# Social Support Histogram
ggplot(data1, aes(x = Social_Support)) +
  geom_histogram(binwidth = 1, fill = "palegreen", color = "black", boundary = 3.5,
closed = "left") +
  scale_x_continuous(
    breaks = seq(3.5, max(data1$Social_Support) + 0.5, by = 1),  # centered bins
    limits = c(3, ceiling(max(data1$Social_Support)) + 1)
  ) +
  scale_y_continuous(
    limits = c(0, max(table(data1$Social_Support)) + 5)
  ) +
  theme_minimal() +
  labs(
    title = "Histogram of Social Support Score",
    subtitle = "Oslo Social Support scores among adolescents (n = 546)",
    x = "Social Support Score",
    y = "Count"
  )
```



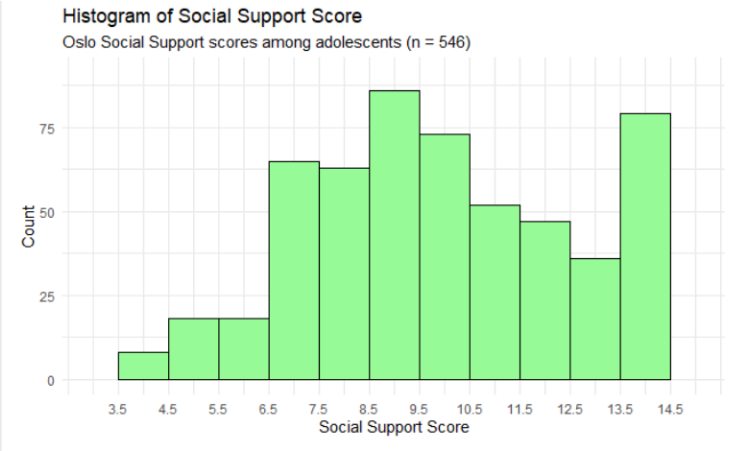
Here is the code for all three histograms.



This histogram is shown as figure 1 in the report.



This histogram is shown as Figure 2 in the report.



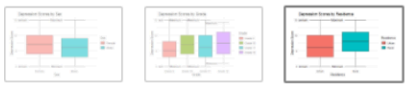
This histogram is shown as Figure 3 in the report.

Appendix 4 – Graphical Summary: Box plots

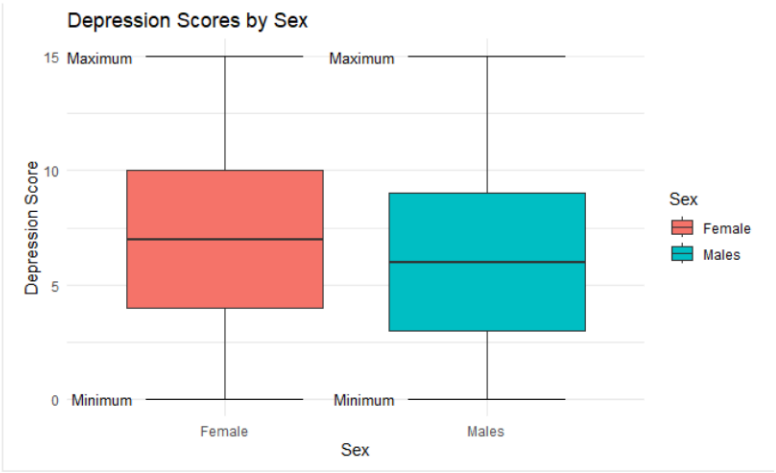
```r
{r}
# Summary stats by sex
stats <- data1 %>%
  group_by(Sex) %>%
  summarise(min = min(Depression_Score),
            max = max(Depression_Score))

# Boxplot with min/max lines and labels
ggplot(data1, aes(x = Sex, y = Depression_Score, fill = Sex)) +
  geom_boxplot() +

  geom_segment(data = stats, aes(x = as.numeric(Sex) - 0.3, xend = as.numeric(Sex)
+ 0.3,
                                 y = min, yend = min), color = "black") +
  geom_segment(data = stats, aes(x = as.numeric(Sex) - 0.3, xend = as.numeric(Sex)
+ 0.3,
                                 y = max, yend = max), color = "black") +

  geom_text(data = stats, aes(x = as.numeric(Sex) - 0.35, y = min, label =
"Minimum"), hjust = 1, size = 3.5) +
  geom_text(data = stats, aes(x = as.numeric(Sex) - 0.35, y = max, label =
"Maximum"), hjust = 1, size = 3.5) +

  theme_minimal() +
  labs(title = "Depression Scores by Sex", x = "Sex", y = "Depression Score")
```

```r
# Summary stats by Grade
grade_stats <- data1 %>%
  group_by(Grade) %>%
  summarise(min = min(Depression_Score),
            max = max(Depression_Score))

# Plot
ggplot(data1, aes(x = Grade, y = Depression_Score, fill = Grade)) +
  geom_boxplot() +
  geom_segment(data = grade_stats, aes(x = as.numeric(Grade) - 0.3, xend = as
.numeric(Grade) + 0.3,
                                       y = min, yend = min), color = "black") +
  geom_segment(data = grade_stats, aes(x = as.numeric(Grade) - 0.3, xend = as
.numeric(Grade) + 0.3,
                                       y = max, yend = max), color = "black") +
  geom_text(data = grade_stats, aes(x = as.numeric(Grade) - 0.35, y = min, label =
"Minimum"), hjust = 1, size = 3.5) +
  geom_text(data = grade_stats, aes(x = as.numeric(Grade) - 0.35, y = max, label =
"Maximum"), hjust = 1, size = 3.5) +
  theme_minimal() +
  labs(title = "Depression Scores by Grade", x = "Grade", y = "Depression Score")

# Summary stats by Residence
res_stats <- data1 %>%
  group_by(Residence) %>%
  summarise(min = min(Depression_Score),
            max = max(Depression_Score))
```

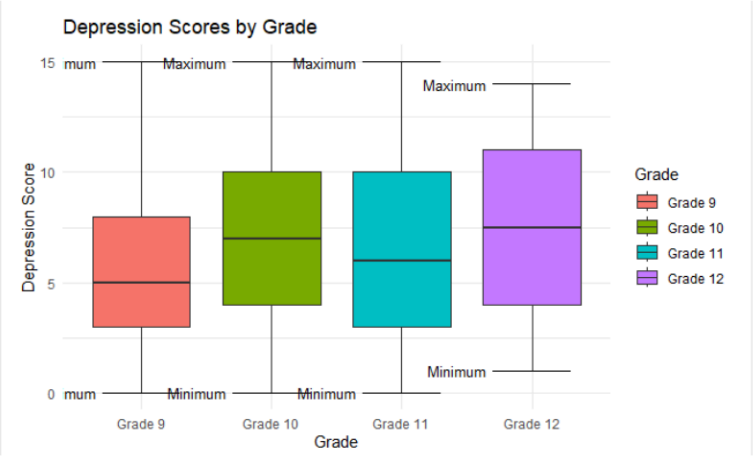
```
# Plot
ggplot(data1, aes(x = Residence, y = Depression_Score, fill = Residence)) +
  geom_boxplot() +
  geom_segment(data = res_stats, aes(x = as.numeric(Residence) - 0.3, xend = as
.numeric(Residence) + 0.3,
                                     y = min, yend = min), color = "black") +
  geom_segment(data = res_stats, aes(x = as.numeric(Residence) - 0.3, xend = as
.numeric(Residence) + 0.3,
                                     y = max, yend = max), color = "black") +
  geom_text(data = res_stats, aes(x = as.numeric(Residence) - 0.35, y = min, label
= "Minimum"), hjust = 1, size = 3.5) +
  geom_text(data = res_stats, aes(x = as.numeric(Residence) - 0.35, y = max, label
= "Maximum"), hjust = 1, size = 3.5) +
  theme_minimal() +
  labs(title = "Depression Scores by Residence", x = "Residence", y = "Depression
Score")
```



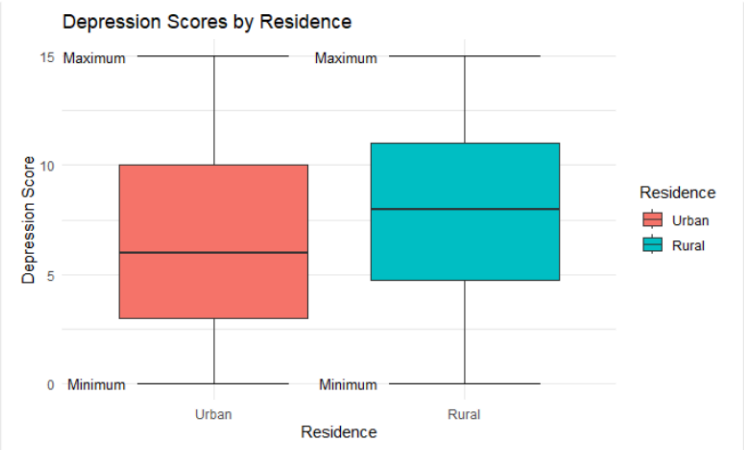
This was the code for all three boxplots.



This is figure 4 of the report.



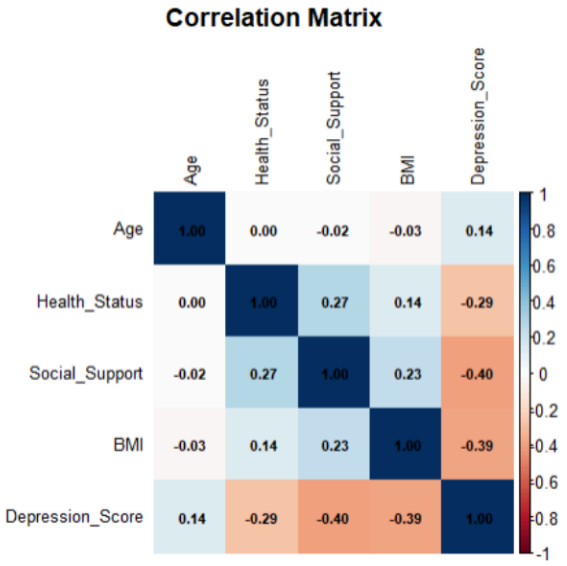
This is figure 5 of the report.



This is figure 6 of the report.

Appendix 5- Graphical Summary: Correlation Matrix

```r
# Select numeric variables
numeric_vars <- data1 %>%
  select(Age, Health_Status, Social_Support, BMI, Depression_Score)

# Calculate correlation matrix
cor_matrix <- cor(numeric_vars)

# Plot correlation heatmap
corrplot(cor_matrix, method = "color", addCoef.col = "black",
         tl.cex = 0.8, number.cex = 0.7,
         tl.col = "black", title = "Correlation Matrix",
         mar = c(0, 0, 1, 0))
```



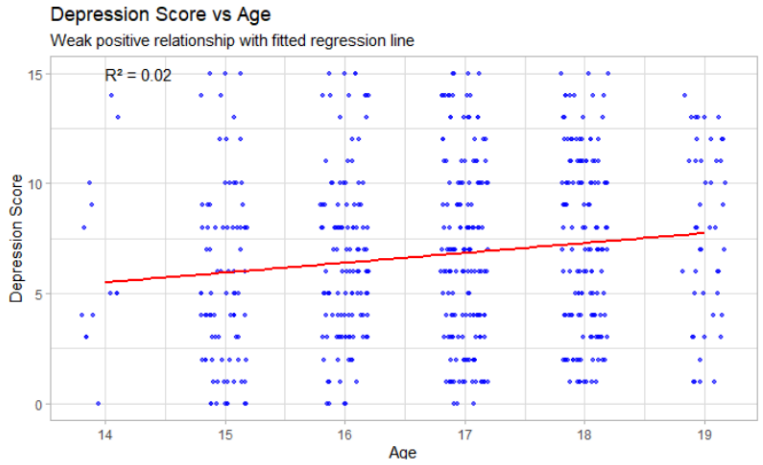
This is figure 7 of the report.

Appendix 6- Graphical Summary: Scatterplots

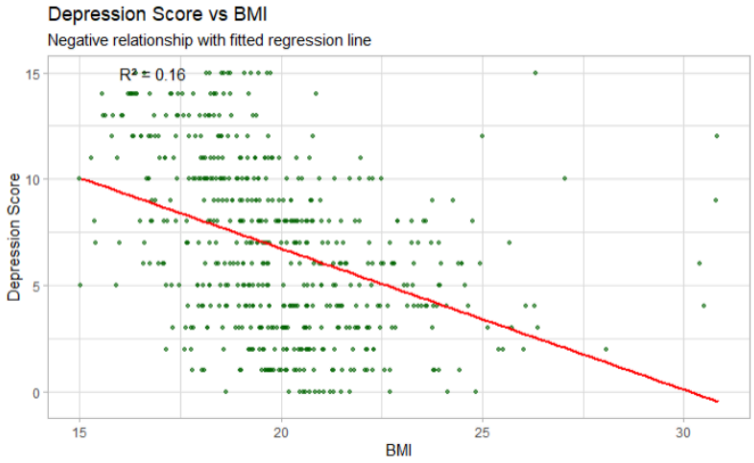
```r
```{r}
# Depression Score vs Age
# Fit linear model to get R²
model_age <- lm(Depression_Score ~ Age, data = data1)
r2_age <- round(summary(model_age)$r.squared, 2)

# Create the scatterplot
ggplot(data1, aes(x = Age, y = Depression_Score)) +
  geom_jitter(width = 0.2, height = 0, alpha = 0.6, color = "blue", size = 1.2) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  annotate("text", x = 14, y = 15, label = paste("R² =", r2_age), size = 4, hjust = 0) +
  labs(
    title = "Depression Score vs Age",
    subtitle = "Weak positive relationship with fitted regression line",
    x = "Age",
    y = "Depression Score"
  ) +
  theme_light()

# Depression Score vs BMI
# Fit linear model to get R²
model <- lm(Depression_Score ~ BMI, data = data1)
r2 <- round(summary(model)$r.squared, 2)
```

```r
# Create the scatterplot
ggplot(data1, aes(x = BMI, y = Depression_Score)) +
  geom_jitter(width = 0.2, height = 0, alpha = 0.6, color = "darkgreen", size = 1.2) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  annotate("text", x = 16, y = 15, label = paste("R² =", r2), size = 4, hjust = 0) +
  labs(
    title = "Depression Score vs BMI",
    subtitle = "Negative relationship with fitted regression line",
    x = "BMI",
    y = "Depression Score"
  ) +
  theme_light()
```
```

This is the r code for the 2 scatter plots.



This is figure 8 of the report.



This is figure 9 of the report.

Appendix 7 – Statistical Methods

Here is the r code of the statistical methods that was used in the report.

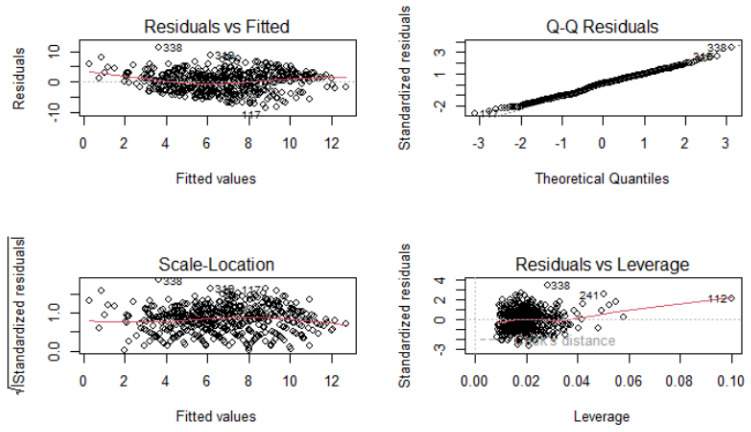
```r
# Fit the full linear regression model
model_full <- lm(Depression_Score ~ Age + Sex + Grade + School_Type + Residence +
                   Health_Status + Social_Support + BMI, data = data1)

# View model summary
summary(model_full)

par(mfrow = c(2, 2))
plot(model_full)

shapiro.test(residuals(model_full))

step(model_full, direction = "both")
```

These were the outcomes of the codes. Tables 2 and 3 are where created with the use of this data.

```
Call:
lm(formula = Depression_Score ~ Age + Sex + Grade + School_Type +
    Residence + Health_Status + Social_Support + BMI, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7816 -2.6729  0.2869  2.3840 11.3731

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            19.28922    3.15719   6.110 1.92e-09 ***
Age                     0.23435    0.18796   1.247 0.213025
SexMales               -1.03695    0.29932  -3.464 0.000574 ***
GradeGrade 10           0.55723    0.45644   1.221 0.222698
GradeGrade 11           0.55800    0.55572   1.004 0.315776
GradeGrade 12           1.02339    0.64549   1.585 0.113453
School_TypeGovernmental 0.08400    0.37820   0.222 0.824329
ResidenceRural          1.05995    0.38856   2.728 0.006583 **
Health_Status          -0.64696    0.15586  -4.151 3.85e-05 ***
Social_Support         -0.43304    0.05778  -7.495 2.76e-13 ***
BMI                    -0.50248    0.06230  -8.065 4.83e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.341 on 535 degrees of freedom
Multiple R-squared:  0.3261,    Adjusted R-squared:  0.3135


        Shapiro-Wilk normality test

data:  residuals(model_full)
W = 0.99401, p-value = 0.03001
```

```
Step:  AIC=1323.03
Depression_Score ~ Age + Sex + Residence + Health_Status + Social_Support +
    BMI

                Df Sum of Sq    RSS    AIC
<none>                      6003.6 1323.0
+ School_Type    1      1.72 6001.8 1324.9
+ Grade          3     30.16 5973.4 1326.3
- Residence      1     83.61 6087.2 1328.6
- Sex            1    132.06 6135.6 1332.9
- Age            1    184.48 6188.0 1337.5
- Health_Status  1    187.35 6190.9 1337.8
- Social_Support 1    626.61 6630.2 1375.2
- BMI            1    769.34 6772.9 1386.9

Call:
lm(formula = Depression_Score ~ Age + Sex + Residence + Health_Status +
    Social_Support + BMI, data = data1)

Coefficients:
   (Intercept)             Age         SexMales  ResidenceRural
       16.1022          0.4659          -1.0291          1.0435
 Health_Status  Social_Support             BMI
       -0.6361         -0.4315         -0.5135
```

This is figure 10 in residual analysis:



Assessment 2 - Individual Project Report PDF

ORIGINALITY REPORT

6%	3%	2%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Kent at Canterbury Student Paper	2%
2	www.grafiati.com Internet Source	1%
3	www.nku.edu Internet Source	1%
4	Submitted to Intercollege Student Paper	1%
5	Submitted to Southern New Hampshire University - Continuing Education Student Paper	<1%
6	www.frontiersin.org Internet Source	<1%
7	Julian J. Faraway. "Linear Models with R", Routledge, 2025 Publication	<1%
8	docplayer.net Internet Source	<1%
9	Hung-Chun Huang, Takashi Nagai, Mita Lovalekar, Christopher Connaboy, Bradley C. Nindl. "Physical Fitness Predictors of a Warrior Task Simulation Test", Journal of Strength and Conditioning Research, 2018 Publication	<1%
10	journals.plos.org Internet Source	<1%

11	Bryan F.J. Manly, Jorge A. Navarro Alberto. "Randomization, Bootstrap and Monte Carlo Methods in Biology", CRC Press, 2020 Publication	<1 %
12	T. Vanwalleghem, J. Poesen, A. McBratney, J. Deckers. "Spatial variability of soil horizon depth in natural loess-derived soils", Geoderma, 2010 Publication	<1 %
13	bdnj.co.uk Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		

Assessment 2 - Individual Project Report PDF

GRADEMARK REPORT

FINAL GRADE

72/100

GENERAL COMMENTS

PAGE 1

QM **Good** (Owen Lyne)

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

QM **Good** (Owen Lyne)

PAGE 7

QM **Good** (Owen Lyne)

PAGE 8

QM **Show less decimal places here** (Owen Lyne)

PAGE 9

QM **Good** (Owen Lyne)

QM **Visualise predictions** (Owen Lyne)

Visualise predictions, such as scatterplot of predicted against actual

PAGE 10

QM **Good** (Owen Lyne)

PAGE 11

PAGE 12

PAGE 13

GRADING FORM: MAST5954 ASST2 2025

SAGARI MURALIEGARAN	**72**

INTRODUCTION 10 MARKS

Good **7**

EXPLORATORY 20 MARKS

Good, thorough exploration with numerical summaries, plots and correlations examined. **16**

METHODS 20 MARKS

Multiple linear regression is appropriate. Diagnostics examined sensibly. Appropriate variable selection. To improve model, could try quadratic and/or interaction terms. **13**

RESULTS 20 MARKS

Sensible model chosen and parameters interpreted appropriately. Could also try to visualise predictions, such as with actual vs predicted plot.

13

REFLECTION 20 MARKS

Good, if somewhat brief, discussion of plots and limitations.

13

R CODE 10 MARKS

Clear, complete.

10