7COM1079-0901-2025 - Team Research and Development Project

Final report title: **Correlation Between Study Hours per Week and GPA**

Group ID: 7COM1079-Group-A-68

Dataset number: DS072

Prepared by: Girish Chowdary Mupparaju [24107364]

Kaviya Saravanan [24152410]

Hemasri Kodamala [24141189]

Navyashree Kodanda [24141020]

Rohith Gandham [24130010]

University of Hertfordshire

Hatfield, 2025

Table of Contents

1. Introduction

1.1 Problem statement and research motivation

Understanding factors that influence academic performance is important for guiding effective study behaviours. Many students assume that spending more hours studying will automatically increase their GPA, yet research suggests the relationship between study time and academic achievement is often weak or inconsistent. Investigating this relationship in the given dataset helps clarify whether study hours genuinely contribute to improved GPA outcomes.

1.2 The data set

The dataset contains information for 55 students, including GPA, weekly study hours, sleep duration, time spent on leisure activities, and gender. For this study, GPA is treated as the dependent interval variable and weekly study hours as the independent interval variable. Only these two variables are analysed in the correlation framework.

1.3 Research question

Is there a correlation between GPA and study hours per week?

1.4 Null hypothesis and alternative hypothesis

Null hypothesis (H0): There is no correlation between GPA and study hours per week ($\rho = 0$).

Alternative hypothesis (H1): There is a correlation between GPA and study hours per week ($\rho \neq 0$).

A two-tailed test is used due to mixed prior evidence and to avoid assuming a positive direction in advance.

2. Background research

2.1 Research papers

Research investigating the relationship between study time and academic performance consistently reports weak or inconsistent correlations. Credé and Phillips (2011) found that study habits and time-management behaviours explain relatively little variance in academic outcomes, suggesting that study duration alone is not a strong predictor of performance. Nonis and Hudson (2010) similarly reported that students who studied more hours did not necessarily achieve higher grades, highlighting the influence of study quality and cognitive engagement. Plant et al. (2005) demonstrated that deliberate, focused study activities—not total time spent—were associated with higher achievement. These studies collectively indicate that the link between weekly study hours and GPA may be limited, making it necessary to evaluate this relationship empirically within the given dataset.

2.2 Why RQ is of interest.

Although many students believe that increasing study hours will directly improve GPA, existing evidence challenges this assumption. The literature shows a research gap in understanding the conditions under which study time truly predicts academic performance. This dataset allows investigation of this relationship in a real student sample, contributing practical insight into whether study effort aligns with academic outcomes. Understanding the strength and nature of this relationship can help both students and educators develop more realistic expectations and emphasise study quality rather than study quantity.

3. Visualisation

3.1 Histogram of GPA (Normality Assessment)

The histogram of GPA includes both a kernel density estimates and an overlaid normal distribution curve. The overall shape is approximately bell-shaped, with no strong skew or extreme deviations. The similarity between the empirical density curve and the theoretical normal curve suggests that GPA is reasonably close to a normal distribution. This visual assessment supports the suitability of applying a parametric correlation test in the analysis.
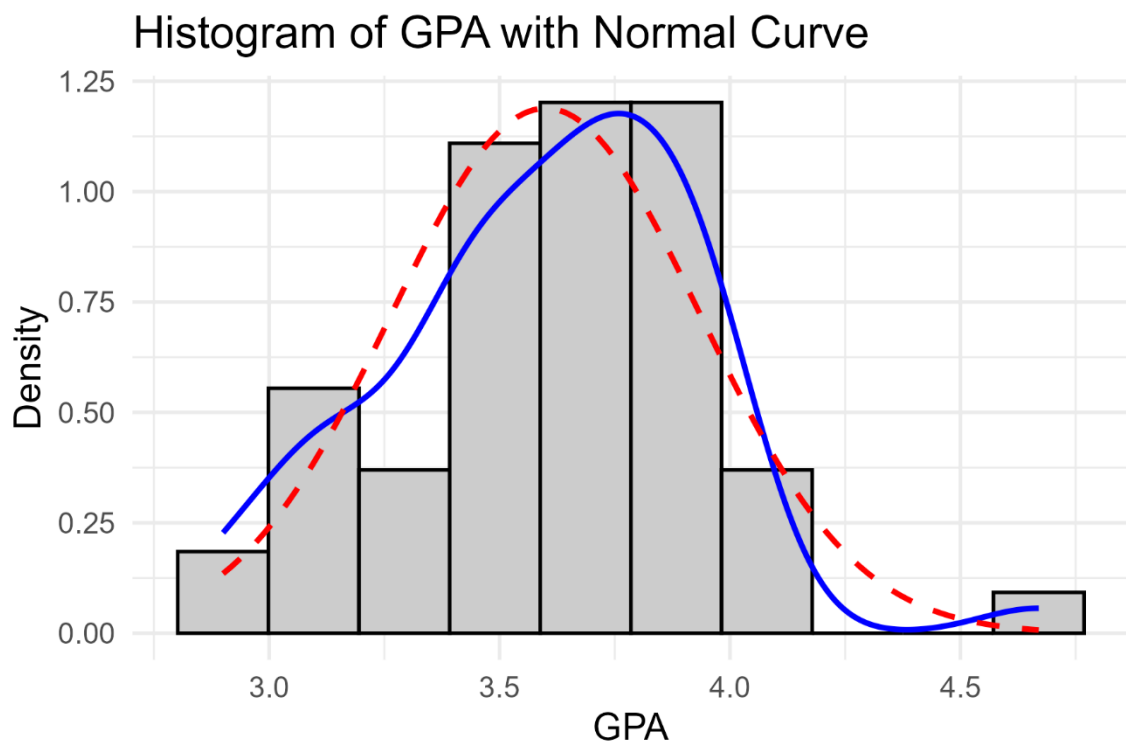

Figure 1. **Histogram of GPA with density curve and normal distribution overlay**

3.2 Scatterplot of GPA vs Study Hours per Week

The scatterplot shows the distribution of GPA across different levels of weekly study hours. The points are widely dispersed, and there is no clear upward or downward pattern. The fitted line remains almost flat, indicating the absence of a meaningful linear trend. This visual evidence suggests a weak or negligible association between the two variables, which is confirmed in the statistical analysis.
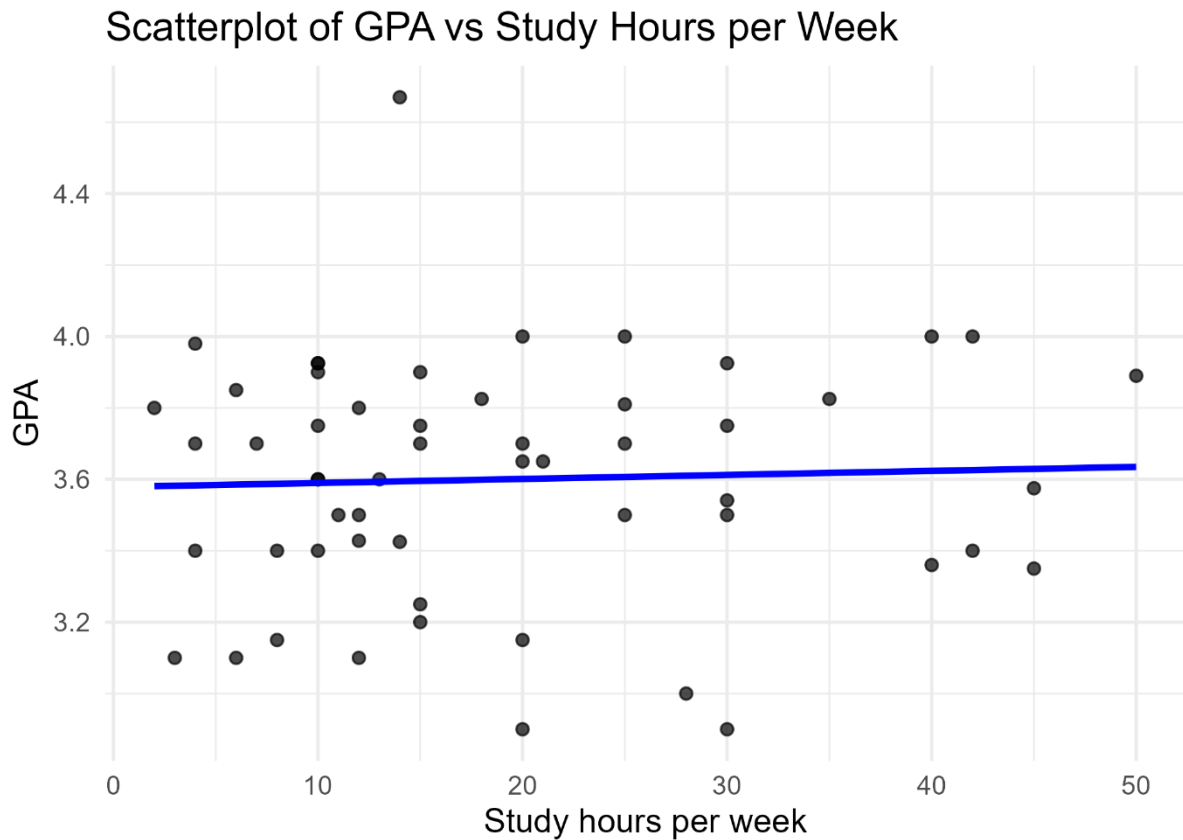
Figure 2. **Scatterplot of GPA against study hours per week**

3.3 Key observations summary

- GPA appears approximately normally distributed, supporting the use of a parametric test.

- The scatterplot shows no visible linear relationship between study hours and GPA.

- The visuals suggest that variation in study hours does not correspond to meaningful changes in GPA.

4. Analysis

4.1 Statistical test used

A Shapiro–Wilk test was performed to assess whether GPA met the normality assumption required for a parametric correlation analysis. The result indicated no significant deviation from normality (W = 0.9609, p = 0.07096). Based on this, Pearson's correlation was selected as the appropriate statistical test for examining the relationship between study hours per week and GPA. This decision aligns with standard guidance that Pearson's correlation is suitable when the dependent variable is approximately normally distributed.

4.2 Hypothesis decision

The Pearson correlation test produced a very weak correlation coefficient (r = 0.0416) with a non-significant p-value (p = 0.76297). As the p-value exceeds the conventional 0.05 threshold, the null hypothesis cannot be rejected. This means there is no statistical evidence to suggest that weekly study hours are linearly associated with GPA in this sample. The results indicate that study hours alone do not meaningfully explain variation in students' GPA, consistent with the lack of pattern observed in the visualisations.

5. Evaluation

5.1 What went well

The group established a clear workflow early in the project, which helped maintain consistent progress. Using GitHub for version control ensured that contributions were transparent and allowed members to work independently without conflicts. The analysis and visualisation stages were completed efficiently due to good coordination and early agreement on responsibilities. Communication improved throughout the project, and deadlines were met without requiring last-minute work.

5.2 Points for improvement

Workload distribution was uneven at times, with some members requiring additional prompting to complete their sections. Earlier engagement with the statistical requirements and template format would have reduced rework later in the project. The group could also improve by holding more frequent progress reviews to ensure that everyone understands requirements and stays aligned with the academic standards expected.

5.3 Group's time management

Despite initial delays from a few members, the structured task plan and GitHub workflow helped the group stay on track. Completing the analysis and visualisation phases early allowed more time for refining the report. Time management improved significantly as the submission deadline approached.

5.4 Project's overall judgement

Overall, the project was successful in answering the research question and producing a report that meets the module requirements. The dataset was analysed using appropriate methods, and the final document demonstrates a sound understanding of correlation analysis and academic reporting conventions. The experience strengthened the group's technical and collaborative skills.

5.5 GitHub log commentary

The GitHub log in Appendix B shows regular commits across the duration of the project, reflecting active participation and transparent workflow. Key commits document major milestones such as dataset upload, script development, visualisation, and report assembly. The log provides clear evidence of how the project evolved and supports fair evaluation of individual contributions.

1. Commit Message: CAD33E5 ANALYSIS: implement full Analysis.R and generate plots + result CSVs

Explanation: Implemented the core Analysis.R script that reads the dataset, computes summary statistics, runs normality and correlation tests, and outputs plots and CSV results. This commit created the reproducible analysis pipeline central to the project.

2. Commit Message: 33C7243 ANALYSIS: clean Analysis.R + add R_RUN instructions; regenerate outputs

Explanation: Polished the analysis scripts, added run instructions and reproducible execution steps. This ensured other group members and markers can reproduce results without path issues and enabled consistent regenerations of outputs.

3. Commit Message: 57C5E4C CLEANUP: remove old/duplicate diagnostic plots from report/figures

Explanation: Cleaned repository and removed unapproved/duplicate figures, ensuring the report complies with the module rules (only GPA histogram + main scatter) and avoids confusion during assessment.

6. Conclusions

6.1 Results explained

The results show that GPA is approximately normally distributed, which justifies the use of Pearson's correlation. The correlation analysis found a very weak relationship between weekly study hours and GPA, with the coefficient close to zero and the p-value well above the 0.05 threshold. This indicates that, in this dataset, study hours do not meaningfully explain variation in GPA.

6.2 Interpretation of the results

These findings suggest that simply increasing study hours does not necessarily lead to improved academic performance. This aligns with previous research indicating that study quality, strategy, and engagement are often more influential than total time spent. For this group of students, differences in GPA appear to be influenced by factors other than study duration, highlighting the limits of using study hours as a reliable predictor of achievement.

6.3 Limitations and future work

The study is limited by the small sample size and the narrow range of variables included. Factors such as motivation, study environment, course difficulty, and learning strategies were not measured but may significantly affect GPA. Future work should investigate these additional variables and explore non-linear or multivariate approaches to better understand academic performance.

7. References

Nonis, S.A. & Hudson, G.I. (2010) 'Performance of college students: The role of study time and strategies', Journal of Education for Business, 85(4), pp. 229–234.

Credé, M. & Phillips, L.A. (2011) 'A meta-analytic review of study habits and academic performance', Educational Psychology Review, 23(2), pp. 123–148.

Plant, E.A., Ericsson, K.A., Hill, L. & Asberg, K. (2005) 'Why study quality matters: Deliberate practice and academic performance', Learning and Instruction, 15(4), pp. 365–382.

8. Appendices

Appendix A – R code

```r
# To reproduce: from project root run in R:
# install.packages(c("tidyverse","broom","ggpubr","readr"))
# source("scripts/Analysis.R")
# ---- Setup ----
      required_pkgs <- c("tidyverse", "broom", "ggpubr", "readr")
      missing_pkgs <- setdiff(required_pkgs,
      rownames(installed.packages()))
      if(length(missing_pkgs)) {
        message("Missing packages: ", paste(missing_pkgs, collapse = ",
      "),
                "\nInstall them before running (uncomment
      install.packages line if desired).")
        # install.packages(missing_pkgs) # uncomment if you want
      automatic install
      }
      library(tidyverse)
      library(broom)
      library(ggpubr)
      library(readr)


      data_file <- NULL
      for(p in possible_paths) {
        if(file.exists(p)) { data_file <- p; break }
      }
      if(is.null(data_file)) {
        stop("Data file not found in expected locations. Please set
      'data_file' variable to the correct path.")
      }

      out_plot_dir <- "plots"
      out_data_dir <- "data/results"

      dir.create(out_plot_dir, showWarnings = FALSE, recursive = TRUE)
      dir.create(out_data_dir, showWarnings = FALSE, recursive = TRUE)
```

```r
# ---- Read data ----
df <- read_csv(data_file, show_col_types = FALSE)

# ---- Clean & coerce ----
df <- df %>%
  mutate(
    gpa = as.numeric(gpa),
    studyweek = as.numeric(studyweek),
    sleepnight = as.numeric(sleepnight),
    out = as.numeric(out),
    gender = as.factor(gender)
  ) %>%
  filter(!is.na(gpa) & !is.na(studyweek))

# ---- Summary stats ----
summary_stats <- tibble(
  n = nrow(df),
  mean_gpa = mean(df$gpa, na.rm = TRUE),
  sd_gpa = sd(df$gpa, na.rm = TRUE),
  median_gpa = median(df$gpa, na.rm = TRUE),
  iqr_gpa = IQR(df$gpa, na.rm = TRUE),
  min_gpa = min(df$gpa, na.rm = TRUE),
  max_gpa = max(df$gpa, na.rm = TRUE)
)
write_csv(summary_stats, file.path(out_data_dir,
"summary_stats.csv"))

# ---- Normality test for GPA (Shapiro-Wilk) ----
# Note: Shapiro is sensitive for large N; interpret alongside
histogram.
shapiro_res <- shapiro.test(df$gpa)
shapiro_tidy <- broom::tidy(shapiro_res)
write_csv(as_tibble(shapiro_tidy), file.path(out_data_dir,
"shapiro_gpa_result.csv"))

# ---- Histogram of GPA with kernel density + normal curve ----
p_hist_gpa <- ggplot(df, aes(x = gpa)) +
  geom_histogram(aes(y = ..density..),
                 bins = 10,
```

```r
                        fill = "grey80",
                        color = "black") +
    geom_density(color = "blue", linewidth = 1) +
    stat_function(
      fun = dnorm,
      args = list(mean = mean(df$gpa, na.rm = TRUE),
                  sd = sd(df$gpa, na.rm = TRUE)),
      color = "red",
      linewidth = 1,
      linetype = "dashed"
    ) +
    labs(
      title = "Histogram of GPA with Normal Curve",
      x = "GPA",
      y = "Density"
    ) +
    theme_minimal(base_size = 14)


ggsave(filename = file.path(out_plot_dir,
"hist_gpa_normalcurve.png"),
       plot = p_hist_gpa, width = 6, height = 4, dpi = 300)


# ---- Scatterplot: Study hours vs GPA (main plot) ----
p_scatter <- ggplot(df, aes(x = studyweek, y = gpa)) +
  geom_point(alpha = 0.7, size = 2) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Scatterplot of GPA vs Study Hours per Week",
       x = "Study hours per week",
       y = "GPA") +
  theme_minimal(base_size = 14)


ggsave(filename = file.path(out_plot_dir, "scatter_gpa_study.png"),
       plot = p_scatter, width = 7, height = 5, dpi = 300)


# ---- Correlation test selection based on Shapiro result ----
shapiro_p <- shapiro_res$p.value


if(shapiro_p > 0.05) {
  corr_method <- "pearson"
```

```
} else {
  corr_method <- "spearman"
}

corr_test <- cor.test(df$studyweek, df$gpa, method = corr_method)
corr_tidy <- broom::tidy(corr_test)
corr_tidy$method <- corr_method
write_csv(corr_tidy, file.path(out_data_dir,
"correlation_test_result.csv"))

# ---- Console summary (concise) ----
cat("---- Analysis Summary ----\n")
cat("Sample size (n):", summary_stats$n, "\n")
cat(sprintf("GPA mean = %.3f; sd = %.3f; median = %.3f\n",
            summary_stats$mean_gpa, summary_stats$sd_gpa,
summary_stats$median_gpa))
cat(sprintf("Shapiro-Wilk W = %.4f, p = %.6f\n",
shapiro_res$statistic, shapiro_res$p.value))
cat("Chosen correlation method:", corr_method, "\n")
cat(sprintf("Correlation estimate = %.4f; p = %.6f\n",
corr_test$estimate, corr_test$p.value))
cat("Histogram saved to:", file.path(out_plot_dir,
"hist_gpa_normalcurve.png"), "\n")
cat("Scatterplot saved to:", file.path(out_plot_dir,
"scatter_gpa_study.png"), "\n")
cat("CSV results saved to:", out_data_dir, "\n")
```
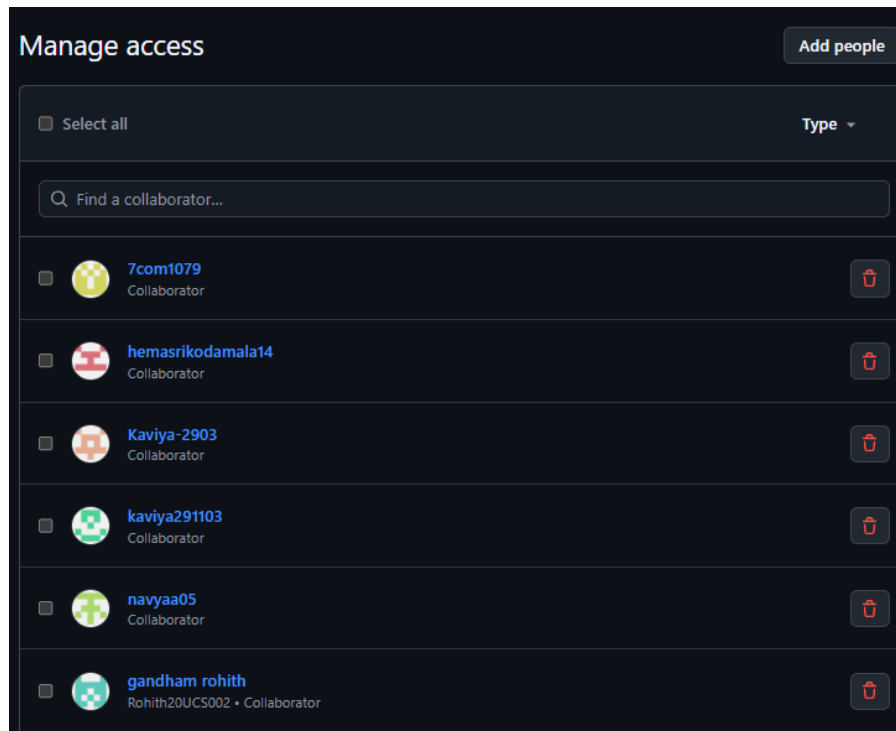
Appendix B – GitHub screenshots

B1. COLLABORATORS

Figure B1. Collaborators and module team access
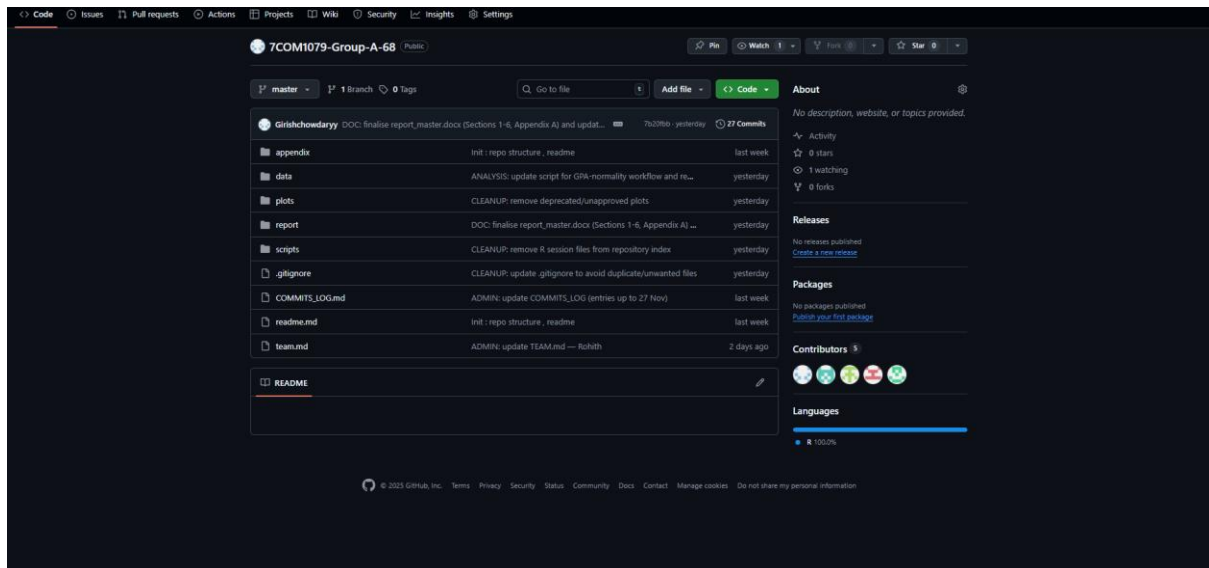
B2. REPOSITORY STRUCTURE

Figure B2. Repository Structure
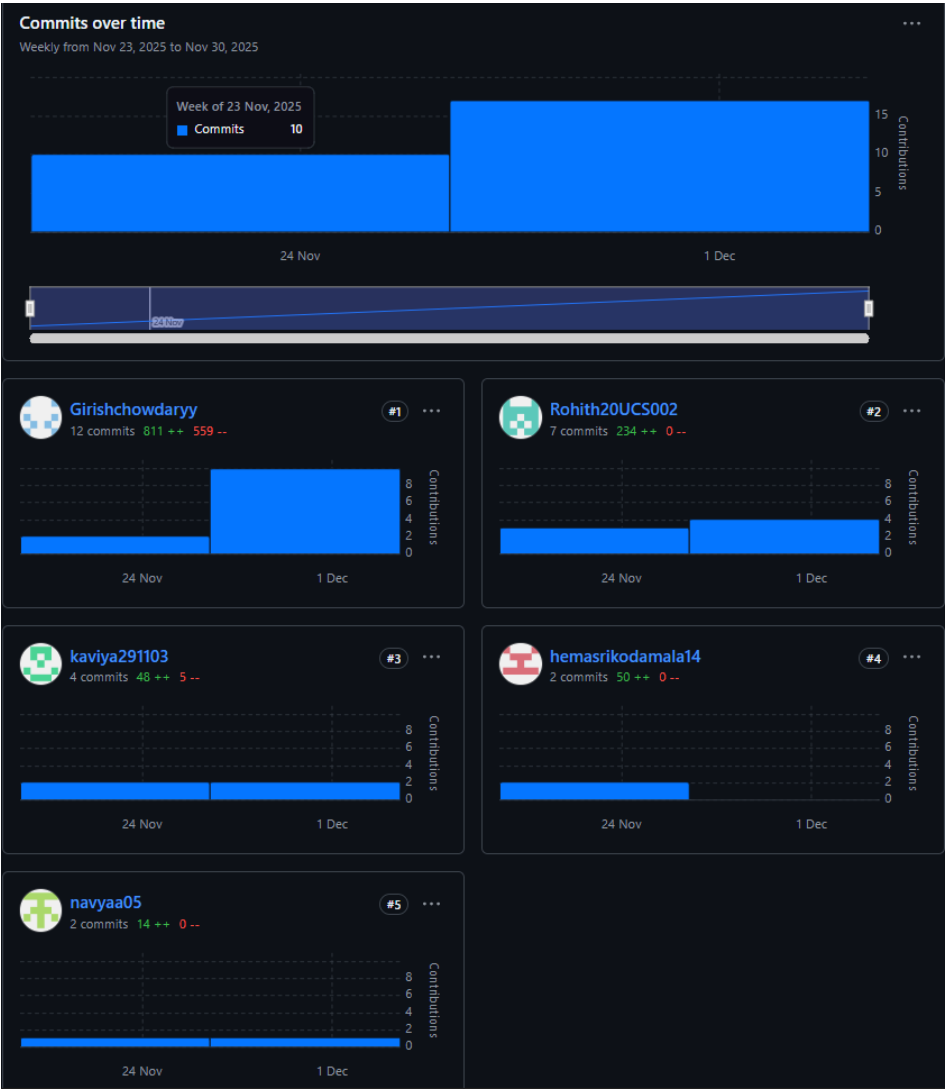
14

B3. CONTRIBUTORS GRAPH

Figure B3. Contributions over time
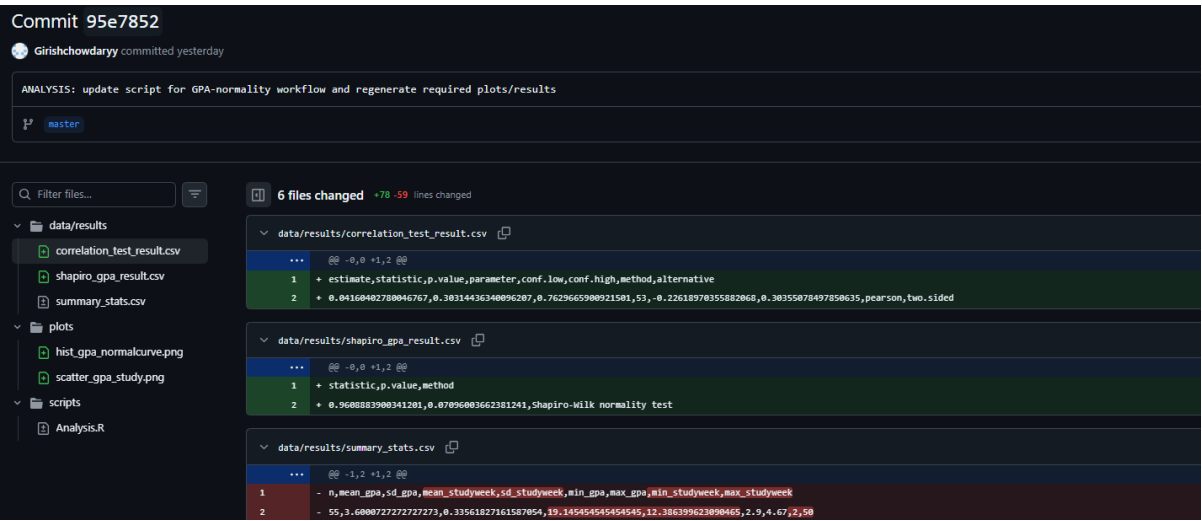
B4. EXAMPLE COMMIT DIFF

Figure B4. Example commit showing changes to analysis outputs

B5. FULL COMMIT LOG

102e50e (HEAD -> master, origin/master, origin/HEAD) DOC: update Appendix B with screenshots and full commit log

692e286 DOC: update Appendix B with new screenshots and commit log

7b20fbb DOC: finalise report_master.docx (Sections 1-6, Appendix A) and update TOC

57c5e4c CLEANUP: remove old/duplicate diagnostic plots from report/figures

dbef8bc DOC: add final figures to report/figures

6e58409 CLEANUP: update .gitignore to avoid duplicate/unwanted files

95f5dc4 CLEANUP: remove R session files from repository index

b164738 CLEANUP: add R session files to .gitignore

6d23b50 CLEANUP: remove deprecated/unapproved plots

95e7852 ANALYSIS: update script for GPA-normality workflow and regenerate required plots/results

86c3a22 DOC: finalise report with Sections 5 and 6, appendices, and formatting

b12f64e DOC: add draft evaluation (Section 5) and conclusions (Section 6)

950aadc DOC: add figures_preview.md for Section 3

52520a7 DOC: finalise Section 2 (literature + research gap)

ceced70 DOC: add full Harvard references for literature review

e6c5a21 DOC: refine Section 4 interpretation

b1354b9 ADMIN: update TEAM.md — Rohith

33c7243 ANALYSIS: clean Analysis.R + add R_RUN instructions; regenerate outputs

af2e519 ADMIN: add GitHub activity screenshots for Appendix B

0c6e191 PLOTS: finalize captions and high-res images; update visuals draft

ca19f8a DOC: complete Section 4 analysis draft

78dec3e DOC: complete literature review (Sections 2.1 and 2.2)

50eab08 DOC: complete Section 3 visualisation draft

9fda89f ADMIN: update COMMITS_LOG (entries up to 27 Nov)

4338015 DOC: add draft_intro (Sections 1.1-1.4)

cad33e5 ANALYSIS: implement full Analysis.R and generate plots + result CSVs.

6046236 added gpa.csv and Analysis.R

284306e Init: repo structure, readme

f06bafc Init: Repo structure