GPA Trends and Predictive Modeling

Thomas Wright

April 1st, 2025

1 Introduction

This project investigates academic performance patterns using the Academic Performance of University Student dataset (sourced from Kaggle, by Krishnansh Verma), which includes GPA records for approximately 3,000 undergraduate students tracked across four academic years. The dataset provides information on important academic indicators such as first-year GPA (`CGPA100`), final cumulative GPA (`CGPA`), secondary school GPA (`SGPA`), and academic program (`Prog.Code`).

The primary goal of the project is to understand how GPA evolves over time and to identify early academic signals that predict final academic performance. I approached the analysis with two main questions in mind: first, whether student background, specifically first-year college and secondary school achievement, meaningfully impacts university outcomes; and second, whether certain academic programs influence GPA due to differences in difficulty or grading norms. For example, I suspected that fields like electrical engineering might suppress GPA due to their rigor, but I also considered that such programs often attract high-achieving students, which could offset or complicate the effect.

Beyond prediction, I was interested in whether patterns in the data showed potential points for academic intervention. If some students struggle early or certain groups consistently underperform, it might offer insight into where support systems could be most effectively deployed. This analysis aims to balance descriptive insights (e.g., trends in GPA by year and program) with explanatory modeling, focusing on how prior academic performance and program enrollment relate to final outcomes.

2 Research Questions

This analysis is guided by three core research questions. First, how does GPA change over time throughout a student's academic journey? By examining GPA distributions from freshman to senior year, we can observe whether performance tends to improve, decline, or remain stable, and whether certain stages of the academic path are more variable than others.

Second, can a student's first-year GPA reliably predict their final cumulative GPA? Since early academic success is often used as a signal for long-term outcomes, this question tests whether that assumption holds true in this dataset. A strong predictive relationship would suggest that first-year performance is a meaningful early indicator of a student's academic trajectory.

Third, do academic programs themselves significantly influence final GPA outcomes, even after accounting for a student's academic background? This question explores whether certain majors are associated with systematically higher or lower GPAs. By using multiple regression to control for secondary school GPA, first-year GPA, and gender, we can isolate the adjusted effect of each program and determine whether the differences reflect more than just student selection.

3 Libraries

The tidyverse collection was central to the workflow. Within it, dplyr and tidyr enabled efficient and readable data wrangling, including reshaping GPA data from wide to long format to support year-by-year analysis and ggplot2 powered all of the visualizations.

To enhance specific plots, additional libraries were used. The ggridges package was chosen to create ridge density plots, which provide a compact and visually intuitive way to compare GPA distributions across multiple academic programs in a single view. RColorBrewer helped select pastel color palettes that clearly distinguish categories while maintaining readability. The knitr package was used for formatting tables, making it easier to present results in a clean and interpretable format within the report and slides. Finally, tinytex was used to ensure compatibility and successful rendering of the PDF output.

```r
library(tidyverse)
library(knitr)
library(ggridges)
library(RColorBrewer)
library(tinytex)
```

4 Data Preparation

```r
gpaData <- read.csv("academic_performance_dataset_V2.csv")

gpaData_long <- gpaData %>%
  pivot_longer(cols = c(CGPA100, CGPA200, CGPA300, CGPA400),
               names_to = "Year", values_to = "Yearly_CGPA")

gpaData_long$Year <- factor(gpaData_long$Year, levels = c("CGPA100", "CGPA200",
                                                "CGPA300", "CGPA400"),
                       labels = c("Freshman", "Sophomore", "Junior", "Senior"))
```

The original data had separate columns for each year's GPA. To analyze trends over time (e.g., plot GPA progression or year-by-year comparisons), the data needed to be reshaped from wide to long format. This makes it easier to use ggplot2 for faceted or grouped plots. The factor labeling improves readability in visualizations.

```r
gpaData <- gpaData %>%
  mutate(
    Program_Name = recode(Prog.Code,
      "BCH" = "Biochem", "BLD" = "Building Tech",
      "CEN" = "Comp E", "CHE" = "Chem E",
      "CHM" = "Ind Chem", "CIS" = "Comp Sci",
      "CVE" = "Civil", "EEE" = "EEE",
      "ICE" = "Info Comm", "MAT" = "Math",
      "MCB" = "Microbio", "MCE" = "Mech E",
      "MIS" = "Mgmt Info Sys",
      "PET" = "Petrol E", "PHYE" = "Ind Phys - E/IT",
      "PHYG" = "Ind Phys - Geo", "PHYR" = "Ind Phys - Renew"
    ),
    Program_Name = factor(Program_Name),
    Gender = factor(Gender)
  )
```

The original Prog.Code values were short abbreviations. Replacing them with readable labels makes the graphs and analysis easier to interpret. Regression models in R treat factor variables as categorical. Converting Program_Name and Gender to factors ensures that the model correctly estimates separate effects for each category. Without this step, R might treat the variables as numeric or character types, leading to incorrect model behavior.

5 GPA Distribution by Year

```r
ggplot(gpaData_long, aes(x = Yearly_CGPA, fill = Year)) +
  geom_histogram(bins = 20, alpha = 0.7) +
  facet_wrap(~ Year, scales = "free_y", ncol = 2) +
  scale_fill_manual(values = c(
    "Freshman" = "lightblue",
    "Sophomore" = "lightgreen",
    "Junior" = "lightyellow3",
    "Senior" = "lightpink"
  )) +
  theme_minimal(base_size = 14) +
  labs(x = "GPA", y = "Count")
```

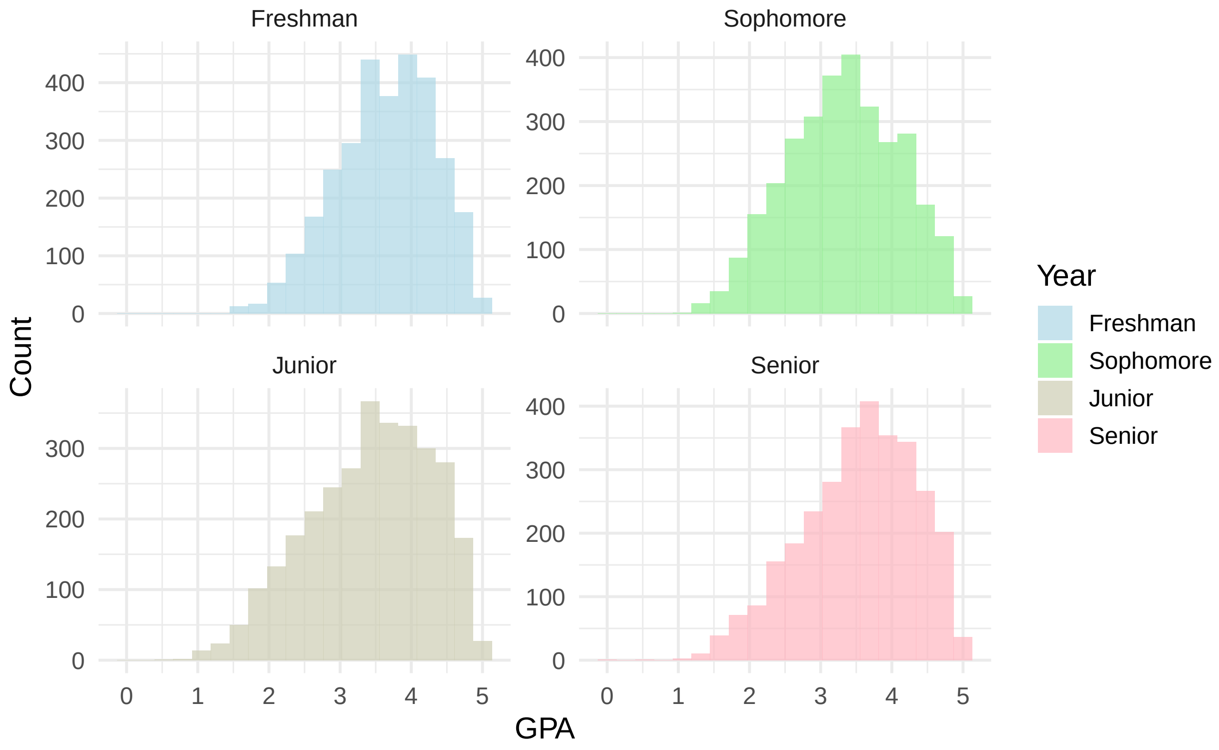


Figure 1: Histogram of GPA distributions across each academic year

To begin the exploratory analysis, I created a faceted histogram to visualize the distribution of GPA scores across each academic year: Freshman, Sophomore, Junior, and Senior. This plot makes it easy to compare the shape, spread, and central tendency of student performance at different stages of their academic careers.

By using `facet_wrap` with `scales = "free_y"`, each histogram adjusts its vertical axis independently, ensuring that variations in sample size don't distort the visual interpretation of the distributions. Color fills were applied to distinguish each year and enhance readability.

The GPA distributions are relatively consistent across years, generally centered around mid-to-high GPA values. However, Sophomore year shows a slight downward shift, with more students clustering in the 3.0–3.5 range compared to the Freshman year's peak closer to 4.0. Junior and Senior distributions appear similar to each other, maintaining this mid-range concentration.

6 GPA Progression by Year

```r
ggplot(gpaData_long, aes(x = Year, y = Yearly_CGPA, fill = Year)) +
  geom_boxplot(alpha = 0.8) +
  scale_fill_manual(values = c(
    "Freshman" = "lightblue",
    "Sophomore" = "lightgreen",
    "Junior" = "lightyellow2",
    "Senior" = "lightpink"
  )) +
  theme_minimal(base_size = 14) +
  labs(title = "GPA Trends from Freshman to Senior Year",
       x = "Year",
       y = "GPA")
```

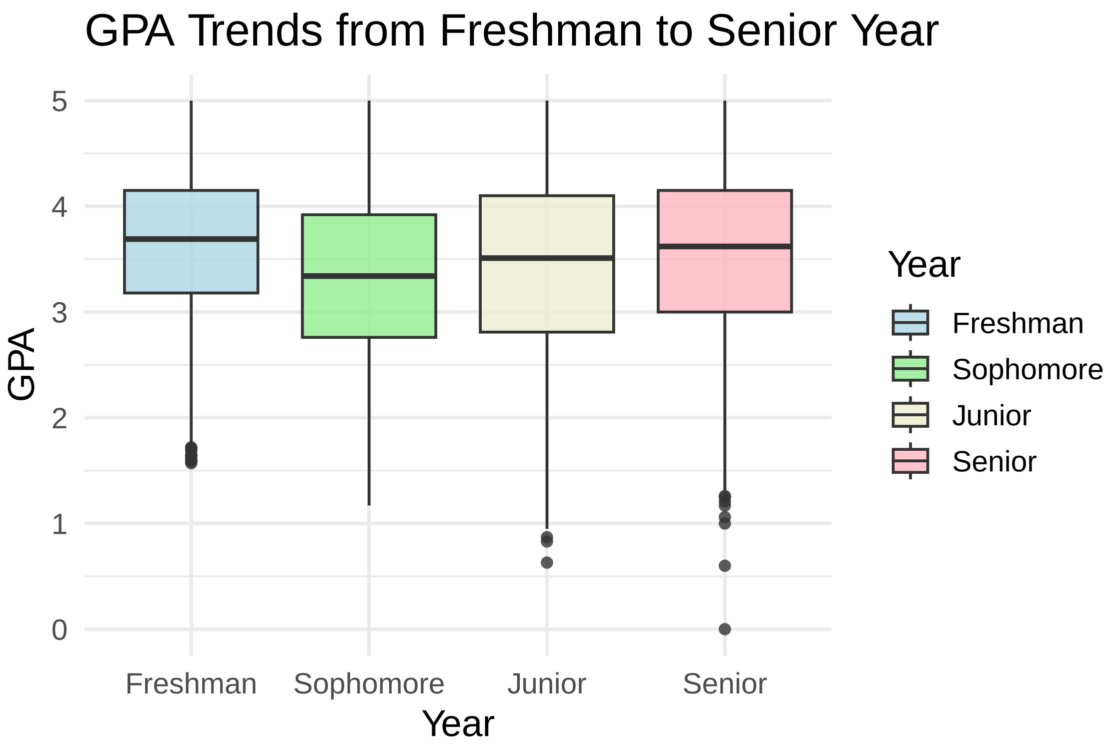


Figure 2: Boxplots showing GPA progression from Freshman to Senior year.

To complement the histograms, I created a boxplot to visualize GPA progression from Freshman to Senior year. This plot summarizes the distribution of GPA scores using medians, interquartile ranges, and potential outliers, giving a compact view of both performance trends and variability across academic years.

The median GPA remains fairly stable over time, generally staying just above 3.5. However, Sophomore year shows a noticeable dip in both the median and the lower quartile, confirming a potential "sophomore slump", often attributed to the transition into more demanding coursework or a drop in motivation after the structured support of Freshman year. GPAs appear to rebound in Junior and Senior years, with medians returning to earlier levels.

That said, the Senior year shows a slight increase in outliers on the lower end. This could reflect signs of academic burnout or loss of motivation as students approach graduation. While the overall trend remains steady, these subtle changes suggest that academic pressures and student engagement may fluctuate in meaningful ways throughout a college career.

7 Predicting Final GPA from First-Year GPA

Before fitting a full model, I wanted to explore the relationship between First-Year GPA (`CGPA100`) and Final GPA (`CGPA`) to test an early hunch: that students who start strong tend to finish strong. A simple correlation is a good starting point to assess this association.

```
cor(gpaData$CGPA100, gpaData$CGPA)
```

```
## [1] 0.7923636
```

This yields a Pearson correlation coefficient of `0.79`, indicating a strong positive relationship. In other words, students who performed well in their first year generally continued to do well through to graduation.

This early result confirmed that first-year academic performance is a meaningful predictor of final outcomes, and motivated the next step of building a linear regression model to quantify this effect more precisely.

Building on the strong correlation, I fit a simple linear regression model with first-year GPA (`CGPA100`) as the sole predictor of final GPA (`CGPA`):

```
model1 <- lm(CGPA ~ CGPA100, data = gpaData)
summary(model1)
```

```
##
## Call:
## lm(formula = CGPA ~ CGPA100, data = gpaData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47226 -0.26587  0.04027  0.30275  1.58913
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.56128    0.04164   13.48   <2e-16 ***
## CGPA100      0.80678    0.01126   71.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.422 on 3044 degrees of freedom
## Multiple R-squared:  0.6278, Adjusted R-squared:  0.6277
## F-statistic:  5135 on 1 and 3044 DF,  p-value: < 2.2e-16
```

The results confirm a strong and statistically significant relationship. Specifically, for each one-point increase in first-year GPA, final GPA increases by approximately `0.81` points. This effect is highly precise, with a standard error of just `0.011`, and statistically significant with a p-value well below `0.001`. The model's R-squared value is `0.63`, indicating that first-year GPA alone accounts for roughly 63% of the variation in final GPA across students. While the intercept of the model (0.56) is not itself meaningful, since a first-year GPA of zero is unrealistic.

This strong linear relationship is also visible in the regression plot below. The red line shows the best-fitting linear trend, surrounded by a shaded confidence band. The data points are tightly clustered around the line, especially in the mid-range of GPA scores and reinforces the model's message that early academic success tends to persist through graduation. This justified the next step of expanding the model to control for other background factors and academic program.

```r
r2_value <- round(summary(model1)$r.squared,3)
r_value <- round(sqrt(r2_value), 3)

ggplot(gpaData, aes(x = CGPA100, y = CGPA)) +
  geom_point(alpha = 0.5, color = "steelblue1") +
  geom_smooth(method = "lm", se = TRUE, color = "red", fill = "gray") +
  annotate("text", x = 1, y = 4.1, label = paste0("r = ", r_value, ", R^2 = ", r2_value),
                                                  size = 5, hjust = 0) +
  theme_minimal(base_size = 14) +
  labs(x = "First-Year GPA", y = "Final GPA")
```

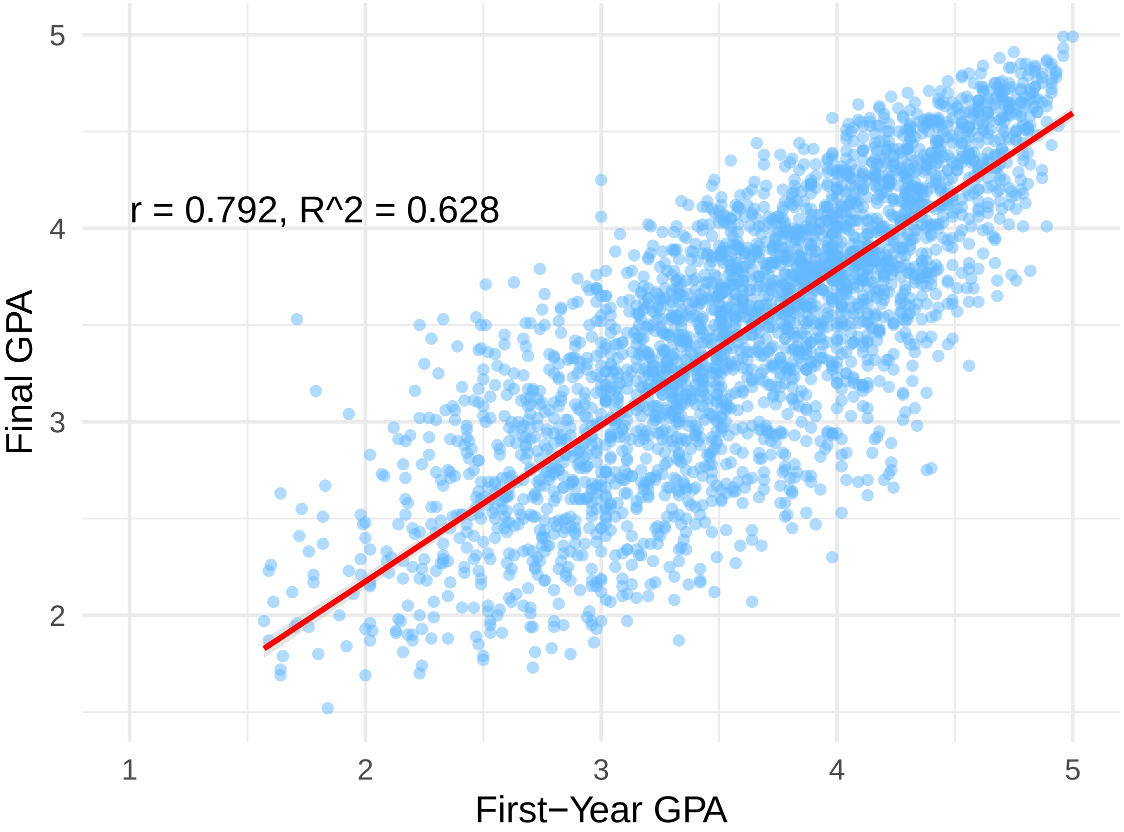
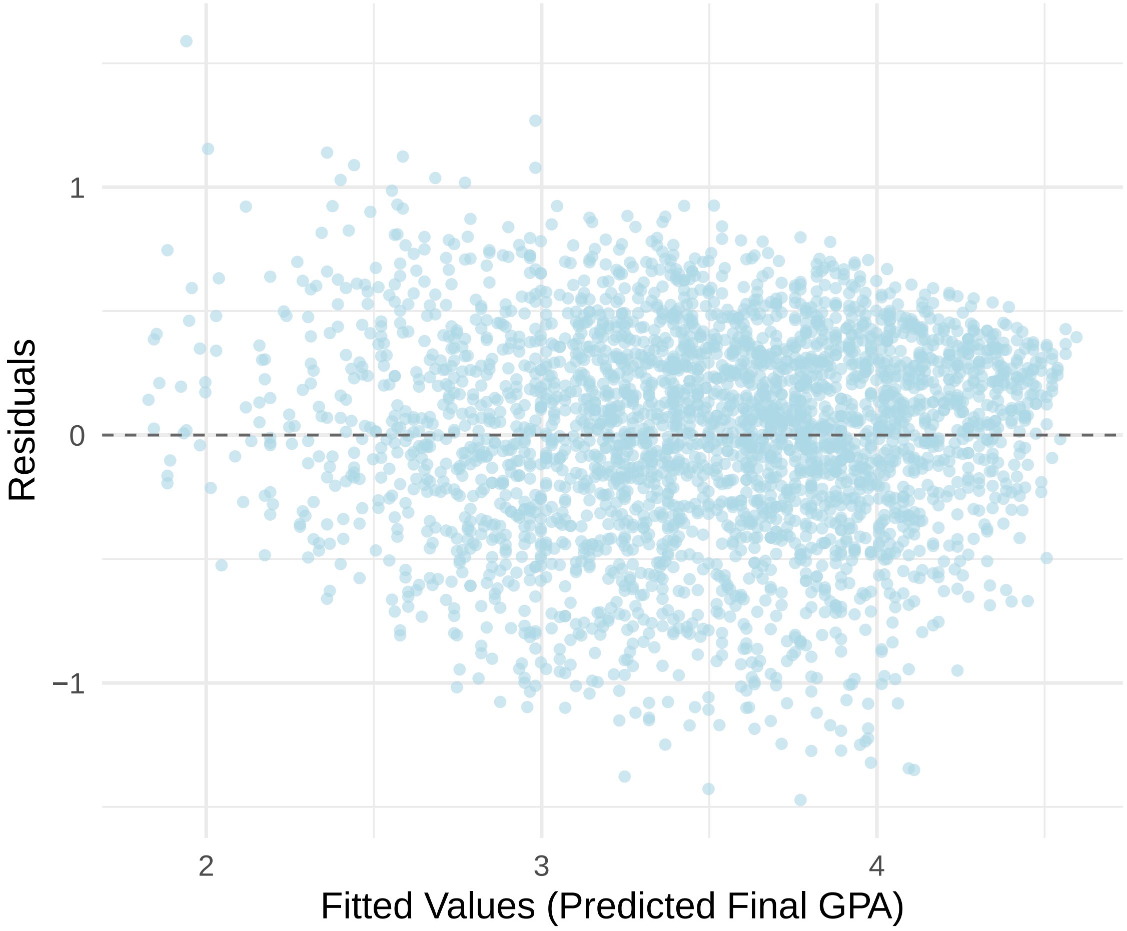


Figure 3: Regression line predicting Final GPA from First-Year GPA

To evaluate the assumptions of the linear model I generated a residual plot. This shows the residual differences between the observed and predicted GPA values plotted against the model's fitted values. Ideally, residuals should be randomly scattered around zero without a systematic pattern, which would indicate that the model's assumptions of linearity and constant variance are reasonably met. As shown in the plot, the residuals are fairly symmetrically distributed across the range of fitted values, though there is slightly more spread in the middle. This is not a serious concern, but does imply that the linear model may slightly underfit students with mid-range predicted GPAs. But overall it shows that the linear model is adequate for this initial analysis.

```
gpaData$residualsCGPA <- resid(model1)
gpaData$fittedCGPA <- fitted(model1)

ggplot(gpaData, aes(x = fittedCGPA, y = residualsCGPA)) +
  geom_point(alpha = 0.6, color = "lightblue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "gray40") +
  theme_minimal(base_size = 14) +
  labs(
    x = "Fitted Values (Predicted Final GPA)",
    y = "Residuals"
  )
```



8 Gender and Academic Performance

To build on the earlier analysis of how first-year GPA predicts final outcomes, I next explored whether gender alone has any explanatory power in predicting final cumulative GPA. While it's well known that academic performance can vary across demographic groups, I wanted to see whether there was a measurable difference in GPA between male and female students in this dataset.

The first step was to calculate some basic descriptive statistics. As shown in the summary table, male students entered with slightly higher average first-year GPAs. However, their final GPAs were notably lower than those of female students. This pattern raises the question of whether gender plays a meaningful role in shaping academic outcomes by graduation.

```
summary_table <- gpaData %>%
  select(CGPA100, SGPA, CGPA, Gender) %>%
  group_by(Gender) %>%
  summarise(
    N = n(),
    `Avg First-Year GPA` = mean(CGPA100),
    `Avg Sec. School GPA` = mean(SGPA),
    `Avg Final GPA` = mean(CGPA)
  )
kable(summary_table, caption = "Average GPA values by gender")
```

Table 1: Average GPA values by gender

Gender	N	Avg First-Year GPA	Avg Sec. School GPA	Avg Final GPA
Female	1093	3.621345	3.139222	3.680412
Male	1953	3.644368	3.108689	3.390942

To test this more formally, I fit a simple linear regression model using gender as the sole predictor of final GPA, offering a baseline view of any performance gap before introducing more complex controls in the multiple regression analysis later on.

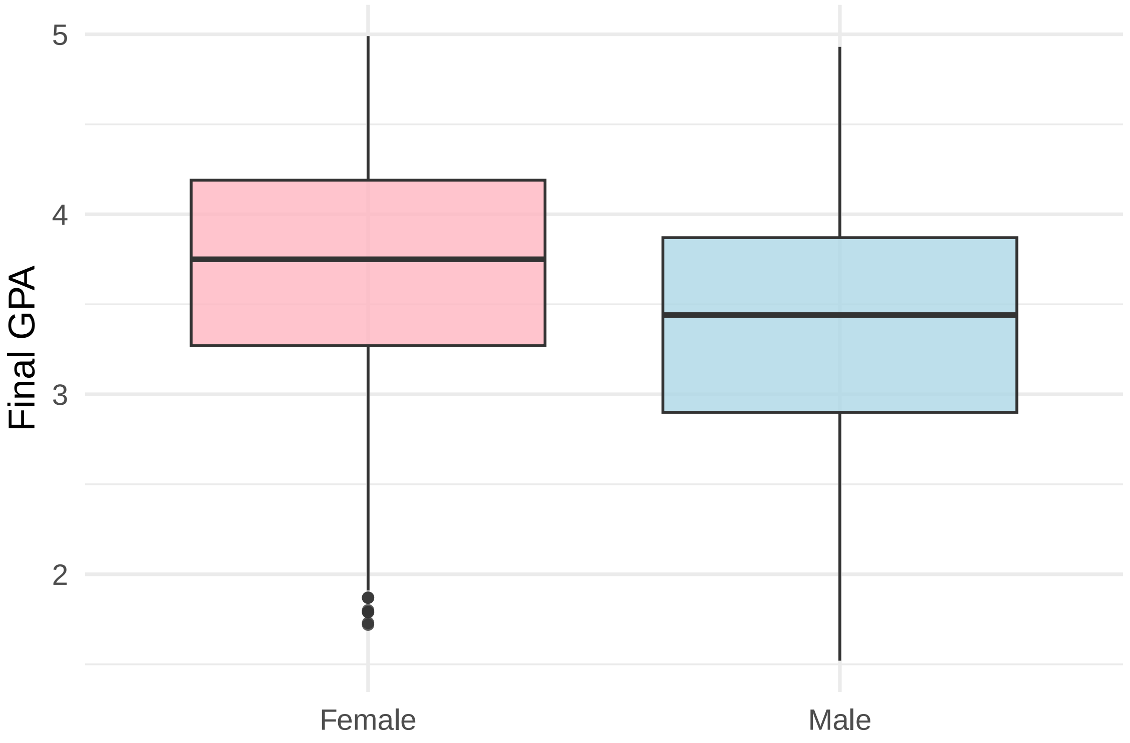
```
model3 <- lm(CGPA ~ Gender, data = gpaData)
summary(model3)
```

```
##
## Call:
## lm(formula = CGPA ~ Gender, data = gpaData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96041 -0.47094  0.05432  0.48906  1.53906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.6804     0.0205  179.56   <2e-16 ***
## GenderMale   -0.2895     0.0256  -11.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.6776 on 3044 degrees of freedom
## Multiple R-squared:  0.04032,    Adjusted R-squared:   0.04
## F-statistic: 127.9 on 1 and 3044 DF,  p-value: < 2.2e-16
```

The model indicates that gender is a statistically significant predictor of GPA, with males scoring on average 0.29 points lower than females (Estimate = -0.2895, $p < 2e-16$). The intercept, representing the mean GPA for females, is approximately 3.68. While the effect is significant, the model explains only 4% of the variance in GPA ($R^2 = 0.04$), indicating that gender, though relevant, accounts for only a small portion of the variation. This supports the decision to include gender in a multiple regression model, but that other factors, like prior academic performance and program of study, likely play a much larger role.

```r
ggplot(gpaData, aes(x = Gender, y = CGPA, fill = Gender)) +
  geom_boxplot(alpha = 0.8) +
  scale_fill_manual(values = c("Female" = "lightpink", "Male" = "lightblue")) +
  theme_minimal(base_size = 14) +
  labs(x = NULL,
       y = "Final GPA"
  ) +
  theme(legend.position = "none")
```



This boxplot displays the distribution of final GPA by gender. Female students tend to have slightly higher final GPAs on average, with a median around 3.8, compared to male students whose median GPA is closer to 3.5. The interquartile range (IQR) appears similar for both groups, indicating a comparable spread in the middle 50% of scores. This corroborates the earlier regression result showing a small but statistically significant negative effect for male gender on final GPA. Although the overall difference is modest.

9 Program-Level GPA Distributions

This was another exploratory step towards answering the third research question: whether academic programs influence GPA outcomes. I created a ridge density plot to visualize the distribution of final GPAs across each academic program. This chart provides a compact and intuitive view of how GPA varies both within and between programs.

```r
gpaData$Program_Name <- reorder(gpaData$Program_Name,
                                gpaData$CGPA, median)

pastel_colors <- colorRampPalette(brewer.pal(9, "Pastel1"))(length(unique(gpaData$Program_Name)))

ggplot(gpaData, aes(x = CGPA, y = Program_Name, fill = Program_Name)) +
  geom_density_ridges(scale = 2, alpha = 0.9) +
  scale_fill_manual(values = pastel_colors) +
  theme_minimal(base_size = 14) +
  theme(legend.position = "none") +
  labs(
    title = "Final CGPA Distribution by Program",
    x = "Final CGPA",
    y = NULL
  )
```

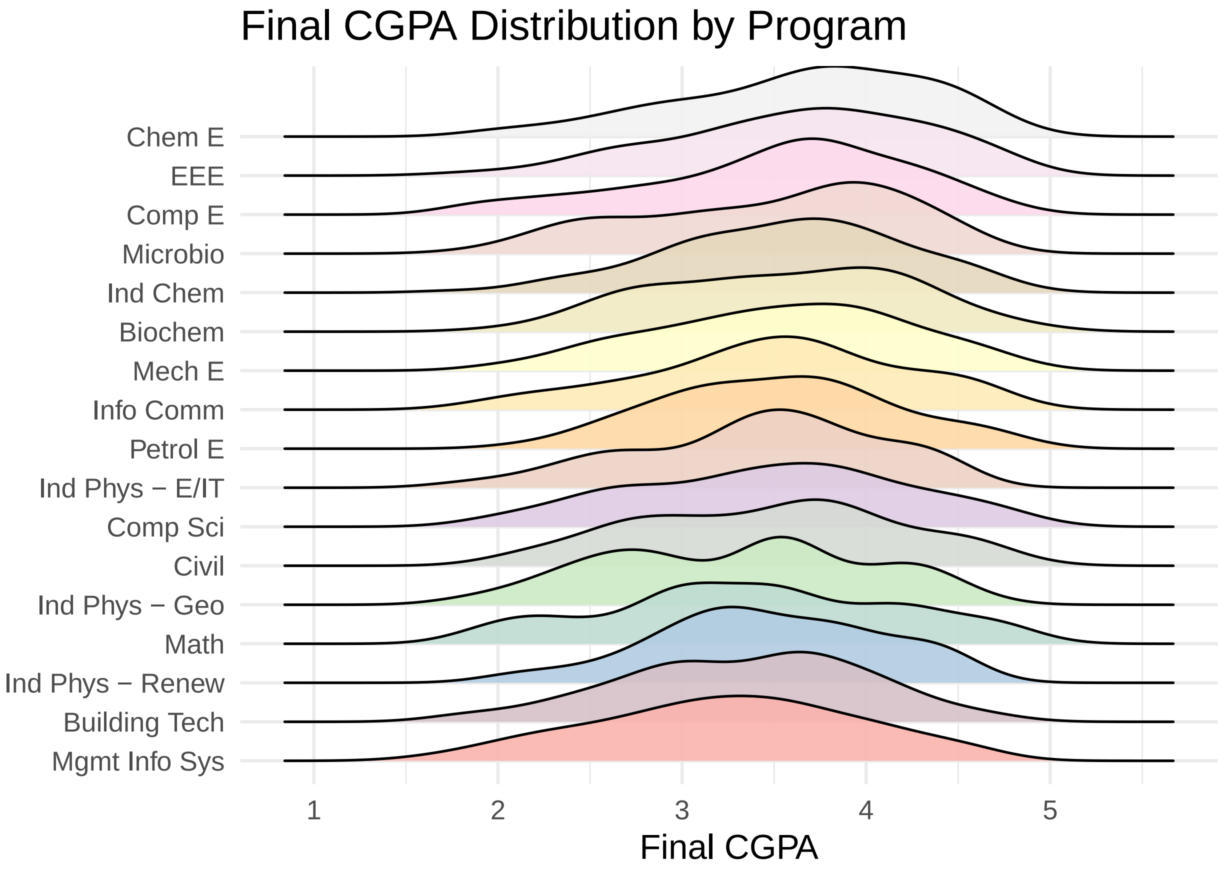


Figure 4: Density plots of final GPA for each program.

```r
gpaData %>%
  group_by(Program_Name) %>%
  summarize(Median_CGPA = median(CGPA, na.rm = TRUE)) %>%
  arrange(desc(Median_CGPA)) %>%
  kable(digits = 3, caption = "Median Final GPA by Academic Program")
```

Table 2: Median Final GPA by Academic Program

Program_Name	Median_CGPA
Chem E	3.760
EEE	3.710
Comp E	3.640
Microbio	3.625
Ind Chem	3.610
Biochem	3.595
Mech E	3.570
Info Comm	3.540
Petrol E	3.530
Comp Sci	3.515
Ind Phys - E/IT	3.515
Civil	3.500
Ind Phys - Geo	3.460
Math	3.430
Ind Phys - Renew	3.425
Building Tech	3.420
Mgmt Info Sys	3.260

To enhance interpretability, the programs are sorted by median GPA, from lowest at the bottom to highest at the top. This ordering reveals that students in Chemical Engineering have the highest median GPA (`3.76`), followed by Electrical Engineering (`3.71`) and Computer Engineering (`3.64`). Toward the lower end are programs like Management Information Systems (`3.26`), Building Technology (`3.42`), and Industrial Physics – Renewable (`3.42`). These differences are visually apparent in the density peaks and reinforce the possibility that program of study plays a role in shaping academic outcomes.

While this plot is descriptive and does not account for differences in student background, it provides a strong rationale for including program as a predictor in the multiple regression model. The goal is to determine whether these observed differences remain significant after adjusting for academic background and gender.

10 Multiple Regression: Program Effects

To more rigorously test whether academic programs influence final GPA, even after accounting for academic background, I fit a multiple linear regression model with `CGPA` as the outcome. The predictors included first-year GPA (`CGPA100`), secondary school GPA (`SGPA`), `gender`, and `academic program`. This model allowed me to estimate the unique contribution of each factor while holding the others constant, providing a clearer picture of what influences final academic outcomes.

The model is expressed as:

$$\text{CGPA} = \beta_0 + \beta_1\,\text{CGPA100} + \beta_2\,\text{SGPA} + \beta_3\,\text{Gender} + \sum_{i=1}^{I-1} \gamma_i\,\text{Program}_i + \varepsilon$$

Here, β_0 is the intercept, and the coefficients β_1, β_2, and β_3 represent the estimated effects of first-year GPA, secondary school GPA, and gender, respectively. The γ_i terms represent the effects of academic programs. The error term ε captures variation not explained by the model. This allowed for clear comparisons of program-level differences, while still adjusting for students' academic backgrounds.

Because academic programs are categorical, I used effect coding (`contr.sum`) rather than the default dummy coding. Effect coding compares each program to the overall mean rather than to a single reference group, which is more appropriate when no single program is a natural baseline. This method also enabled me to manually reintroduce the omitted category (`Chem E`) for plotting purposes and maintain a complete set of comparisons.

However, effect coding replaces program names in the model summary with generic labels (e.g., Program_Name1, Program_Name2). To retain readable program names for interpretation and reporting purposes, I also ran a second version of the same model using the default coding. This second model below (`model2_labels`) was not used for inference but simply to make the output more accessible when reviewing coefficients for each academic program.

```r
model2_labels <- lm(CGPA ~ CGPA100 + SGPA + Gender + Program_Name, data = gpaData)
summary(model2_labels)
```

```
## 
## Call:
## lm(formula = CGPA ~ CGPA100 + SGPA + Gender + Program_Name, data = gpaData)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.34393 -0.22029  0.02819  0.24405  1.73667 
## 
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                   0.41090    0.04665   8.808  < 2e-16 ***
## CGPA100                       0.89060    0.01176  75.719  < 2e-16 ***
## SGPA                          0.07659    0.01214   6.308 3.24e-10 ***
## GenderMale                   -0.23929    0.01508 -15.869  < 2e-16 ***
## Program_NameBuilding Tech     0.15789    0.04271   3.696 0.000222 ***
## Program_NameInd Phys - Renew -0.10309    0.07512  -1.372 0.170030    
## Program_NameMath             -0.10618    0.05143  -2.065 0.039046 *  
## Program_NameInd Phys - Geo   -0.22735    0.06403  -3.551 0.000390 ***
## Program_NameCivil            -0.19434    0.03616  -5.375 8.26e-08 ***
## Program_NameComp Sci         -0.21748    0.02945  -7.384 1.97e-13 ***
## Program_NameInd Phys - E/IT  -0.08008    0.04600  -1.741 0.081775 .  
## Program_NamePetrol E         -0.40568    0.03418 -11.869  < 2e-16 ***
## Program_NameInfo Comm        -0.32731    0.03227 -10.143  < 2e-16 ***
## Program_NameMech E           -0.39843    0.03614 -11.025  < 2e-16 ***
## Program_NameBiochem          -0.17384    0.03715  -4.679 3.01e-06 ***
## Program_NameInd Chem          0.11048    0.04060   2.721 0.006544 ** 
## Program_NameMicrobio         -0.09811    0.03581  -2.740 0.006185 ** 
## Program_NameComp E           -0.35326    0.03291 -10.734  < 2e-16 ***
## Program_NameEEE              -0.43384    0.03015 -14.388  < 2e-16 ***
## Program_NameChem E           -0.38981    0.03433 -11.355  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.366 on 3026 degrees of freedom
## Multiple R-squared:  0.7218, Adjusted R-squared:   0.72
## F-statistic: 413.1 on 19 and 3026 DF,  p-value: < 2.2e-16
```

The regression results reinforced the importance of early academic performance. Both CGPA100 and SGPA were strong, statistically significant predictors of final GPA. Gender also emerged as a meaningful factor, male students had lower average final GPAs than female students, even after adjusting for GPA history and program enrollment. Most importantly, the model revealed systematic differences in GPA outcomes across academic programs, suggesting that some programs are associated with higher or lower GPAs even after accounting for student's academic background. The overall model explained 72% of the variance in final GPA, improving substantially upon the 63% explained by the simple regression using only first-year GPA. This increase demonstrates that program choice and academic history together offer a much clearer picture of student outcomes.

```r
contrasts(gpaData$Program_Name) <- contr.sum
program_labels <- levels(gpaData$Program_Name)[-length(levels(gpaData$Program_Name))]
model2 <- lm(CGPA ~ CGPA100 + SGPA + Gender + Program_Name, data = gpaData)
summary(model2)
```

```r
program_gpaData <- summary(model2)$coefficients %>%
  as.data.frame() %>%
  rownames_to_column("term") %>%
  filter(str_starts(term, "Program_Name")) %>%
  mutate(
    Program = program_labels,
    CI_low = Estimate - 1.96 * `Std. Error`,
    CI_high = Estimate + 1.96 * `Std. Error`
  ) %>%
  select(Program, Estimate, `Std. Error`, CI_low, CI_high) %>%
  arrange(Estimate) %>%
  mutate(Program = factor(Program, levels = Program))
```

To visualize and interpret the effect of academic programs on final GPA, I extracted the relevant coefficients from the multiple regression model. These coefficients represent the estimated effect of each program on final GPA, after controlling for first-year GPA (CGPA100), secondary school GPA (SGPA), and gender. Instead of using default dummy coding, which compares every program to a single reference group (in this case it happens to be Chem E), I used effect coding (contr.sum). This changes the comparison. Now, each program's effect is relative to the overall average rather than a specific baseline which I believe improved the interpretability of the subsequent effect plot. Because of this, one program (Chem E) is considered omitted from the model output because it's no longer a reference program.

```r
gender_row <- summary(model2)$coefficients %>%
  as.data.frame() %>%
  rownames_to_column("term") %>%
  filter(term == "GenderMale") %>%
  mutate(
    Program = "Male vs Female",
    CI_low = Estimate - 1.96 * `Std. Error`,
    CI_high = Estimate + 1.96 * `Std. Error`
  ) %>%
  select(Program, Estimate, `Std. Error`, CI_low, CI_high)
```

The first step was to isolate all rows from the model output related to program effects. I calculated 95% confidence intervals for each program's coefficient to reflect the uncertainty around each estimate. I then assigned readable program labels and arranged the programs in order of effect size to aid interpretation in the resulting plot.

```r
chem_e_row <- tibble(
    Program = "Chem E",
    Estimate = -sum(program_gpaData$Estimate),
    `Std. Error` = NA,
    CI_low = NA,
    CI_high = NA
  )
```

Because `Chem E` was omitted from the model output due to effect coding, I manually added it back by calculating its implied effect. Under effect coding, the sum of all group-level estimates is constrained to zero. Therefore, the estimate for `Chem E` is simply the negative of the sum of all other program estimates. I left the standard error and confidence interval as `NA` since they aren't directly available from the model output and calculating them manually is somewhat complex and beyond what I felt comfortable attempting for this project. This step allowed `Chem E` to be included alongside the other programs in the final visualization for a complete comparison.

```r
program_gpaData <- bind_rows(program_gpaData, gender_row, chem_e_row) %>%
  arrange(Estimate) %>%
  mutate(Program = factor(Program, levels = unique(Program)))
```

To complete the dataset for visualization, I combined the program coefficients extracted from the model with the manually created rows for `gender` and `Chem E`. This allowed me to include all relevant effects, both automatically generated and manually added in, into a single data frame.

Finally, I re-leveled the `Program` variable so that the order of the categories matched the sorted regression estimates. This step was purely for visualization purposes. By setting the factor levels to follow the ascending order of program effects, the plot would automatically display the programs from the lowest to highest estimated impact on GPA along the y-axis. This makes the figure more intuitive and interpretable, as viewers can easily compare which programs are associated with stronger or weaker GPA outcomes. It also preserves the visual ranking established earlier in the ridge plot.

```r
ggplot(program_gpaData, aes(x = Estimate, y = Program)) +
  geom_point(color = "steelblue", size = 3) +
  geom_errorbarh(aes(xmin = CI_low, xmax = CI_high), height = 0.2, color = "gray50") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red", linewidth = 0.5) +
  theme_minimal(base_size = 13) +
  theme(plot.subtitle = element_text(size = 9)) +
  labs(
    x = "Effect on Final GPA",
    y = NULL,
    title = "Estimated Program Effects",
    subtitle = "Compared to Average Program, Controlling for Academic Background and Gender",
    caption =
"Includes gender effect for reference. Negative value indicates males scored lower, on average."
  )
```

To visualize the results of the multiple regression model, I created a coefficient plot showing the estimated effect of each academic program on final GPA (`CGPA`), controlling for first-year GPA (`CGPA100`), secondary school GPA (`SGPA`), and `gender`. Each point represents a program's estimated effect relative to the average program, and horizontal error bars show 95% confidence intervals. Because effect coding was used, the baseline comparison is the overall mean GPA, not a specific reference group.

This makes the plot easier to interpret, since values above zero indicate programs where students perform better than expected based on their academic background, while values below zero reflect underperformance

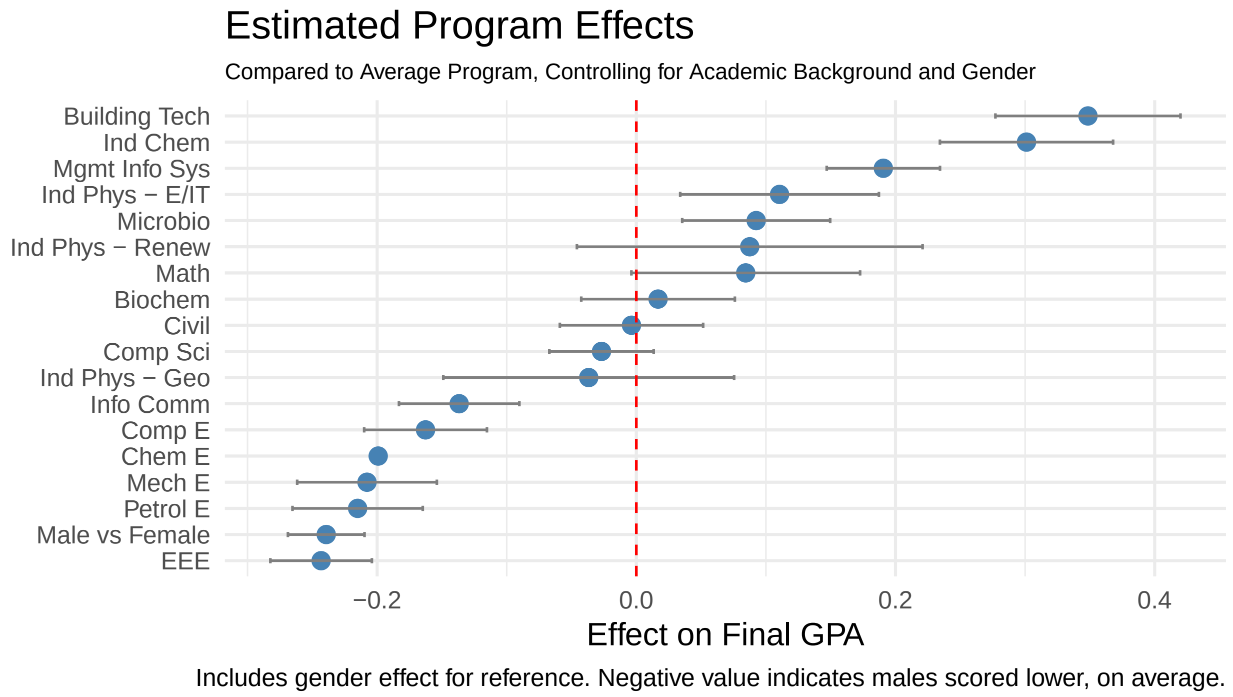


Figure 5: Estimated program effects on final GPA, controlling for background.

relative to expectations. For example, students in Building Technology and Industrial Chemistry tend to exceed predicted outcomes, even after adjusting for first-year GPA, secondary school GPA, and gender. In contrast, students in Electrical and Petroleum Engineering, while often strong performers in absolute terms, have lower final GPAs than the model would predict given their strong academic starting points. This suggests that these programs may be more rigorous or have stricter grading standards, leading to relatively lower outcomes after controlling for prior achievement. The gender effect is also included, showing that male students tend to score lower than females on average, even when accounting for academic history and program of study. See table below for details.

```
kable(program_gpaData, digits = 3, caption =
        "Estimated Effects of Academic Program on Final GPA (Relative to Mean)")
```

Table 3: Estimated Effects of Academic Program on Final GPA (Relative to Mean)

Program	Estimate	Std. Error	CI_low	CI_high
EEE	-0.243	0.020	-0.282	-0.204
Male vs Female	-0.239	0.015	-0.269	-0.210
Petrol E	-0.215	0.026	-0.265	-0.165
Mech E	-0.208	0.027	-0.262	-0.154
Chem E	-0.199	NA	NA	NA
Comp E	-0.163	0.024	-0.210	-0.115
Info Comm	-0.137	0.024	-0.183	-0.090
Ind Phys - Geo	-0.037	0.057	-0.149	0.075
Comp Sci	-0.027	0.021	-0.067	0.013
Civil	-0.004	0.028	-0.059	0.052

Program	Estimate	Std. Error	CI_low	CI_high
Biochem	0.017	0.030	-0.042	0.076
Math	0.084	0.045	-0.004	0.173
Ind Phys - Renew	0.088	0.068	-0.046	0.221
Microbio	0.093	0.029	0.035	0.150
Ind Phys - E/IT	0.111	0.039	0.034	0.187
Mgmt Info Sys	0.191	0.022	0.147	0.234
Ind Chem	0.301	0.034	0.234	0.368
Building Tech	0.349	0.036	0.277	0.420

```r
gpaData$residuals_model2 <- resid(model2)
gpaData$fitted_model2 <- fitted(model2)

ggplot(gpaData, aes(x = fitted_model2, y = residuals_model2)) +
  geom_point(alpha = 0.6, color = "lightblue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "gray40") +
  theme_minimal(base_size = 14) +
  labs(x = "Fitted Values (Predicted Final GPA)",
       y = "Residuals"
  )
```

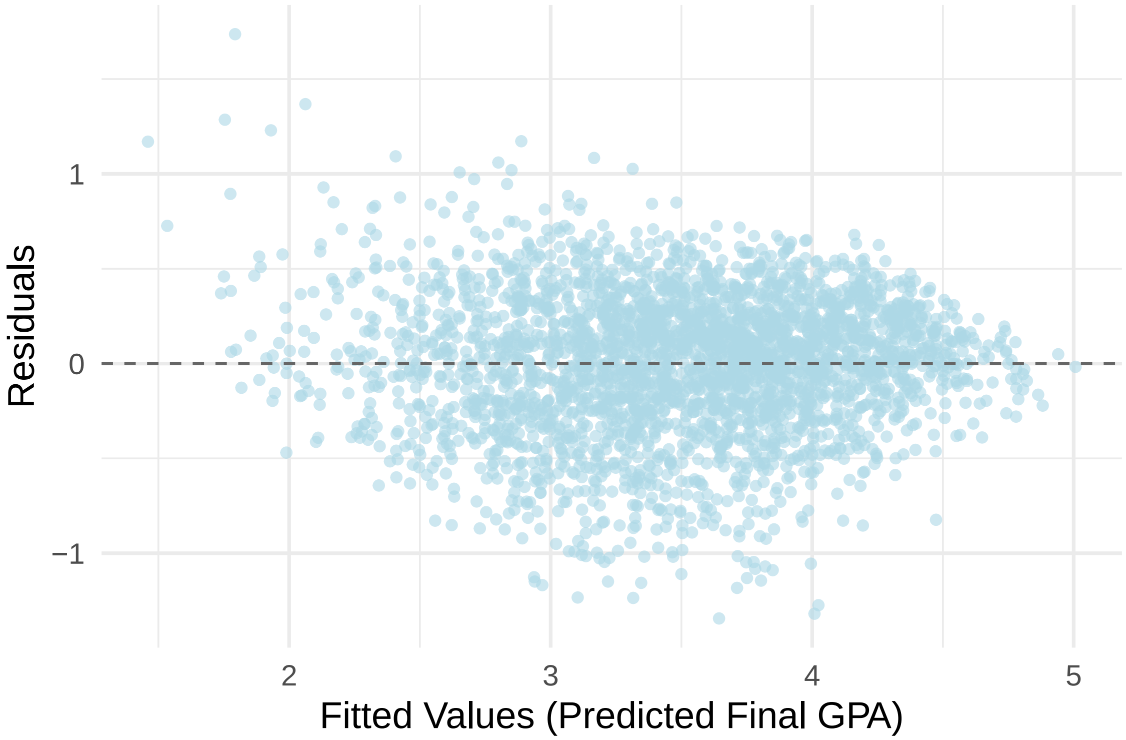


Figure 6: Residual plot assessing fit of the multiple regression model.

To assess how well the multiple regression model fits the data, I plotted the residuals against the fitted values. In this plot, each point represents the difference between a student's actual GPA and the GPA predicted by the model, after accounting for first-year GPA, secondary school GPA, gender, and academic program. The residuals are randomly scattered around the zero line, with no clear pattern or curvature. This implies that the model's linearity assumption is reasonable and that are no major systematic errors. In other words, while there is still some noise in individual predictions, as expected in any real-world data, the model captures the major relationships fairly well and the spread appears roughly consistent across the range of fitted values.

11 Conclusion

This analysis provided insight into the academic performance of university students by modeling final GPA outcomes using early academic indicators and academic programs. The findings illustrate the importance of early college performance, as First-Year GPA (`CGPA100`) alone accounted for approximately `63%` of the variance in final cumulative GPA (`CGPA`). This association reinforces the idea that early academic success is a strong predictor of long-term outcomes and may serve as a benchmark for early intervention and support.

When additional variables were introduced, including secondary school GPA (`SGPA`), `gender`, and `academic program`, the model's explanatory power increased substantially, with the adjusted R^2 rising to `0.72`. This suggests that while First-Year GPA is highly predictive of final cummulative GPA, academic background and other non-measured factors also play a meaningful role.

One notable finding was the `gender` effect. Male students, on average, had significantly lower final GPAs than their female counterparts, even after accounting for prior academic performance and program of study. This result could reflect differences in study habits, support networks, or institutional dynamics, and warrant further investigation.

The program-level analysis revealed meaningful differences in GPA outcomes across academic fields. Students enrolled in Building Technology and Industrial Chemistry had estimated GPA increases of roughly +0.3 to +0.4 points above the overall average, after controlling for other factors. In contrast, students in Electrical and Petroleum Engineering experienced decreases of approximately -0.2 to -0.25 GPA points. These differences may reflect variation in grading norms, curriculum difficulty, or academic support across departments.

These results highlight not only which factors predict success but also where disparities exist, potentially guiding institutional efforts to improve academic outcomes.

12 Limitations and Future Directions

While the analysis provides strong evidence associating early academic performance and program enrollment to final GPA, there are important limitations worth noting. First, the dataset lacks several potential influencing factors such as socioeconomic status, standardized test scores, attendance rates, or measures of academic engagement. Including these could provide a more complete picture of what drives student success.

Second, the data appears to be from a single university, which could limit the generalizability of the findings. Grading standards, academic support systems, and cultural expectations can and do differ across institutions, meaning the patterns observed here might not hold.

Third, while linear regression offers interpretable estimates, it assumes linearity and additive relationships. Future research could explore non-linear models or more complex machine learning methods to capture more complex interactions between predictors.

Finally, causality cannot be inferred from this analysis. Although strong associations are present, there may be confounders that influence the relationships identified. Future work using experimental designs could identify causal and effect for academic interventions.

13 Acknowledgments

The dataset used in this project was created by Krishnansh Verma and made publicly available on Kaggle and provided the foundation for all analysis in this report. I am grateful for the open availability of this resource, which can be accessed here: Academic Performance of University Student Dataset.