Final Project

August 04, 2025

Contents

DS116 / CS343

Final Project

Analysis of TUMO Student Data

Rita Chamiyan, Hayk Alekyan, Aram Barkhudaryan, Gor Arutiunian, Hayk Grigoryan

American University of Armenia
Yerevan, Armenia

August 4, 2025

Contents

Abstract

This project investigates student learning outcomes at the TUMO Center through data-driven exploratory analysis in R. Using anonymized student-level data, we analyze engagement patterns, performance trends over time, and cluster students based on behavioral and demographic features. The findings aim to inform program design and decision-making at TUMO by identifying key insights into student progress and segmentation.

Keywords:

TUMO; Student Performance; Learning Outcomes; Data Visualization; RShiny; Dashboard; Clustering; Segmentation; R; Data Science; Exploratory Data Analysis (EDA); Education Analytics.

Introduction

The TUMO Center for Creative Technologies empowers students through project-based learning in various fields, from technology to design. As participation grows, so does the importance of understanding how students learn and progress. This project applies exploratory data analysis (EDA) techniques using R to uncover trends in student performance, engagement over time, course popularity, and meaningful subgroups within the student population. The goal is to extract actionable insights that can support curriculum planning and personalized learning paths.

Literature Review

Exploratory Data Analysis (EDA) is a foundational approach in educational data mining for uncovering patterns that conventional assessments may overlook. Numerous studies highlight the utility of clustering and segmentation techniques in understanding the diversity of learning behaviors and optimizing interventions for personalized learning. In the context of this project, tools from the R ecosystem—such as ggplot2 for visualization, dplyr for data manipulation, and rstatix for statistical testing—enabled a comprehensive and reproducible workflow. Research in tech-enabled learning environments has shown that trend analysis over time, combined with segmentation based on learner profiles, can provide meaningful insights to guide curriculum and engagement strategies.

Data Collection

The dataset used in this project was provided directly by the TUMO Center for Creative Technologies. It consists of anonymized student-level records captured across different sessions, workshops, and time periods. Key features include student identifiers, demographic details (e.g., age, gender), enrollment and participation dates, workshop attendance, and module completion indicators. The data was primarily supplied in CSV format, with some parts briefly reviewed in Excel to understand initial structure and metadata annotations. All source files were placed in the Data/ directory of the project to ensure organization and reproducibility.

Data Preprocessing

Before conducting any analysis, the raw dataset provided by the TUMO Center was carefully cleaned and prepared to ensure reliability and consistency. The preprocessing stage involved several key steps:

- **Cleaning column names** to standardize and simplify the structure using consistent naming conventions.
- **Removing duplicate records** to ensure each student and workshop entry was unique and not overrepresented.
- **Handling missing values**, especially in critical fields like student IDs, age, or gender. Rows with incomplete key data were removed or imputed as appropriate.
- **Parsing and formatting date variables** such as enrollment dates using consistent `YYYY-MM-DD` formats. This allowed for accurate time-based calculations and visualizations.
- **Encoding categorical variables**, such as gender, using factor levels to support grouped summaries and statistical tests.
- **Creating derived variables**, such as total workshops attended per student or age at the time of enrollment, to provide additional insights for segmentation and trend analysis.
- **Joining and transforming data** into a tidy structure, where each variable forms a column, each observation forms a row, and each type of observational unit forms a table.

To perform these operations, a range of tidyverse-based packages were used, including `dplyr`, `janitor`, `lubridate`, `tidyr`, and `forcats`. These tools enabled efficient and reproducible data wrangling, setting the stage for meaningful exploratory analysis and visualization.

Overall Age Distribution

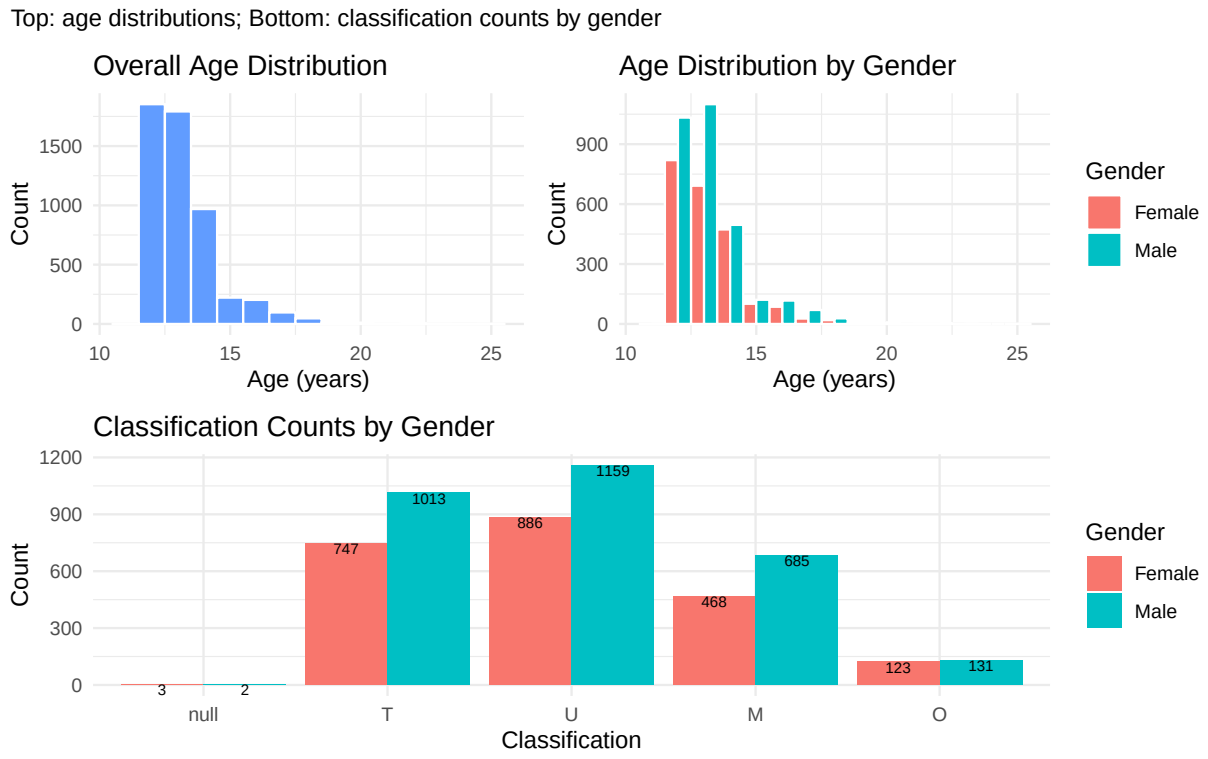


Figure 1: Age & Classification Overview

The majority of TUMO students are between 12 and 16 years old, with a peak at age 13. Male and female students share nearly identical age profiles, both peaking at 13, indicating no gender gap in age distribution. Classification T and U dominate,in contrast, M and O are smaller.

Boxplot of Age per Classification

Students in T are the youngest (median ~12), followed by U, M, and O; O shows the greatest age variability.

Age by Classification & Gender

Within each classification, male and female ages overlap heavily; no gender-specific outliers.

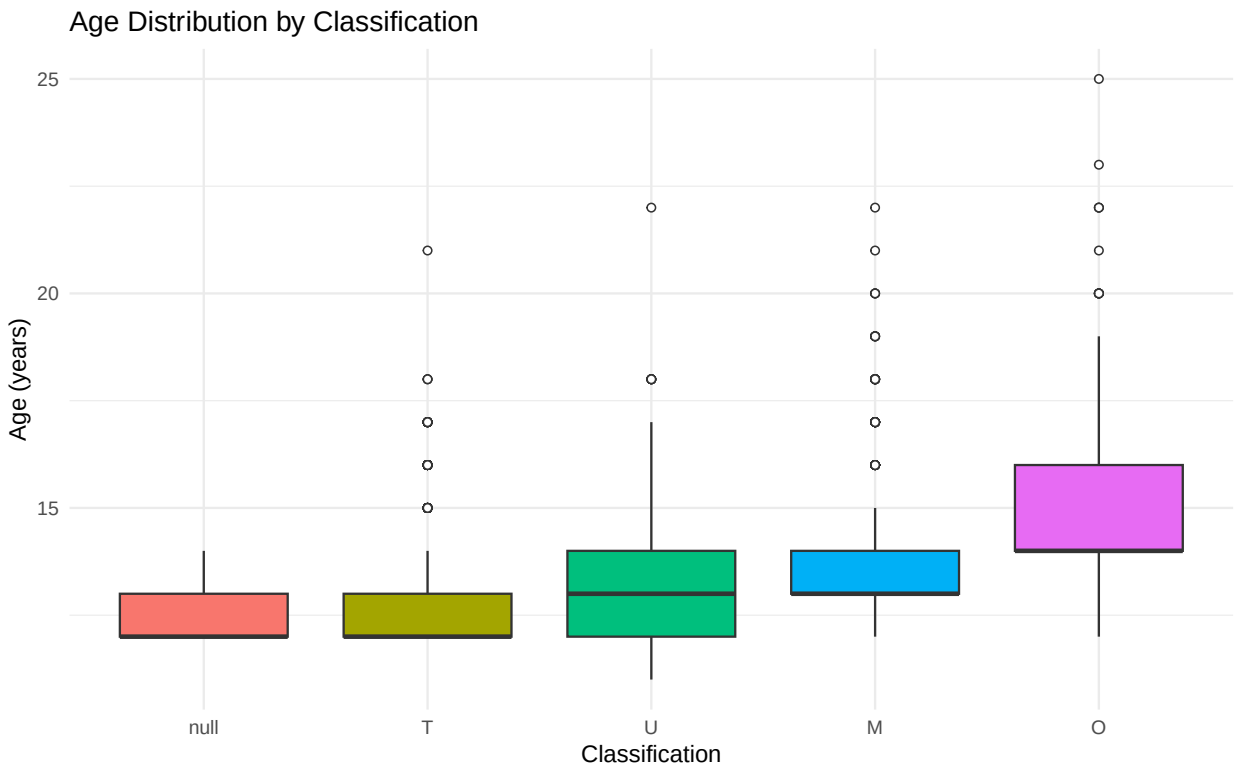


Figure 2: Boxplot of Age by Classification

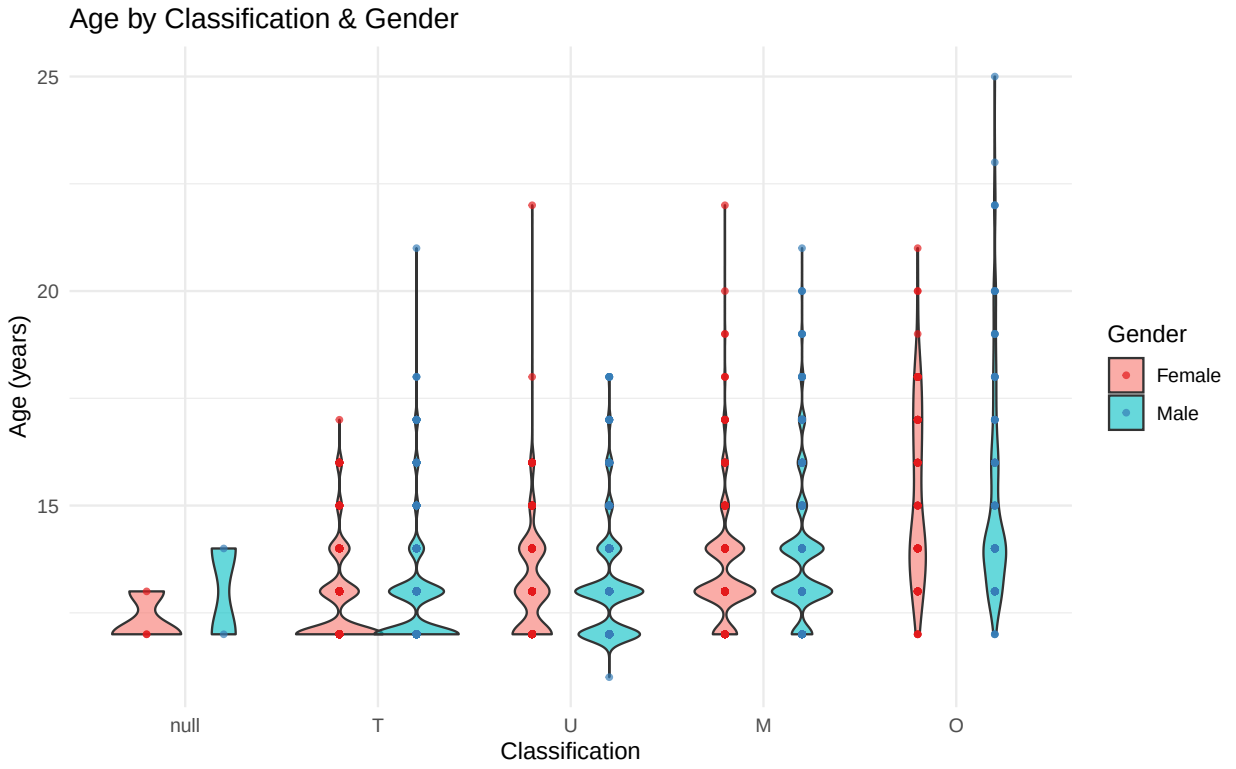


Figure 3: Violin + Jitter: Age by Classification & Gender

Student Dynamics & Learning Patterns

This section looks at how student involvement in workshops changes over time. It covers seasonal patterns, which skills are popular, when students tend to withdraw or fail. 1. Which skills are becoming more or less popular over time? 2. Do withdrawals happen more often in certain months/seasons? 3. Are there time-based patterns in failure rates (Incomplete or Participated outcomes)?

H1. Course Popularity Over Time Hypothesis: Some skills have become more popular in recent months compared to others.

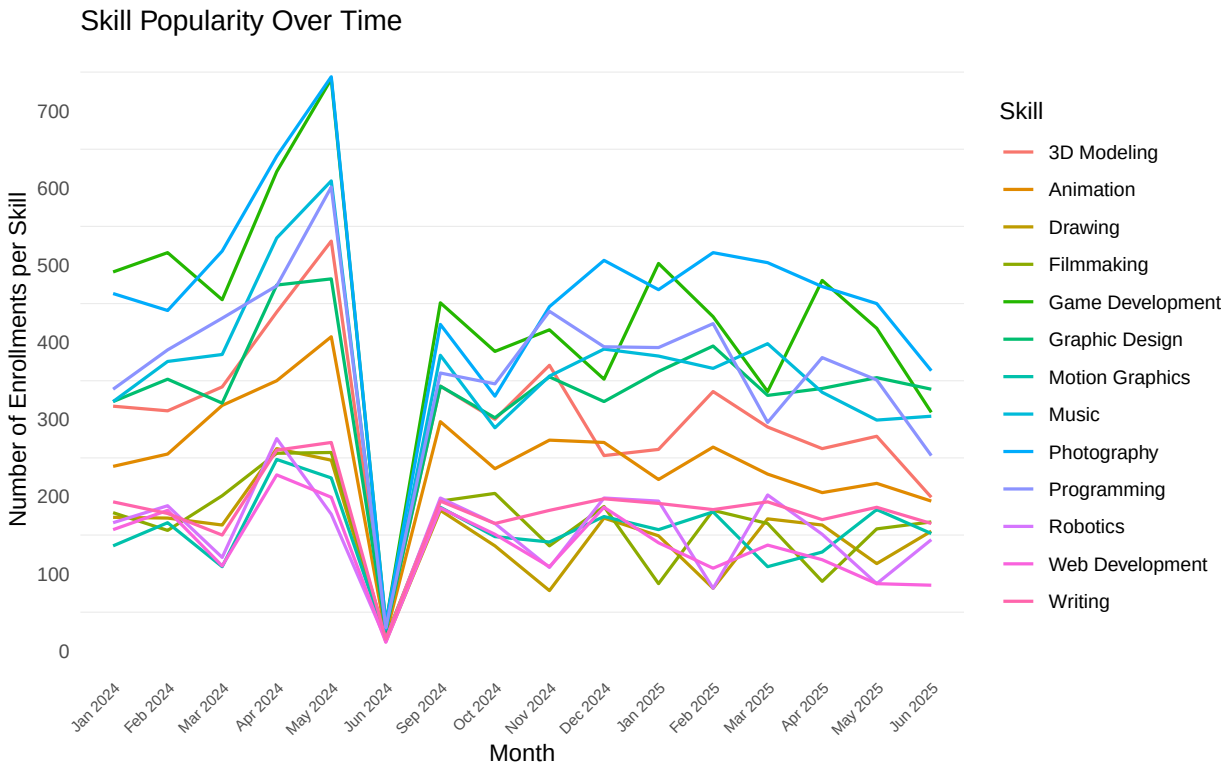
H(Null): Skill enrollments have remained constant over time — no skill shows increasing popularity. H(Alternative): At least one skill shows a significant increase in enrollments over time (i.e., skill popularity is increasing).

Workshop enrollments from January 2024 to June 2025 show clear ups and downs. In the first few months of 2024, around 3,500 students joined each month. That number jumped in April and May, going over 5,000 students — the busiest time of the year. Then in June 2024, enrollments dropped below 500. This wasn't because students lost interest, but because fewer workshops were offered that month. After summer, enrollments picked up again in September and stayed steady through April 2025, averaging around 3,200 to 3,800 per month. May and June 2025 showed another small decline. Overall, this pattern reflects the center's usual schedule: peak engagement in spring, lower activity in summer, and steady interest during the school year.

```
##
##  Pearson's Chi-squared test
##
## data:  skill_month_table
## X-squared = 59472, df = 208, p-value < 2.2e-16

## H is supported: Skill popularity changes significantly over time.
```

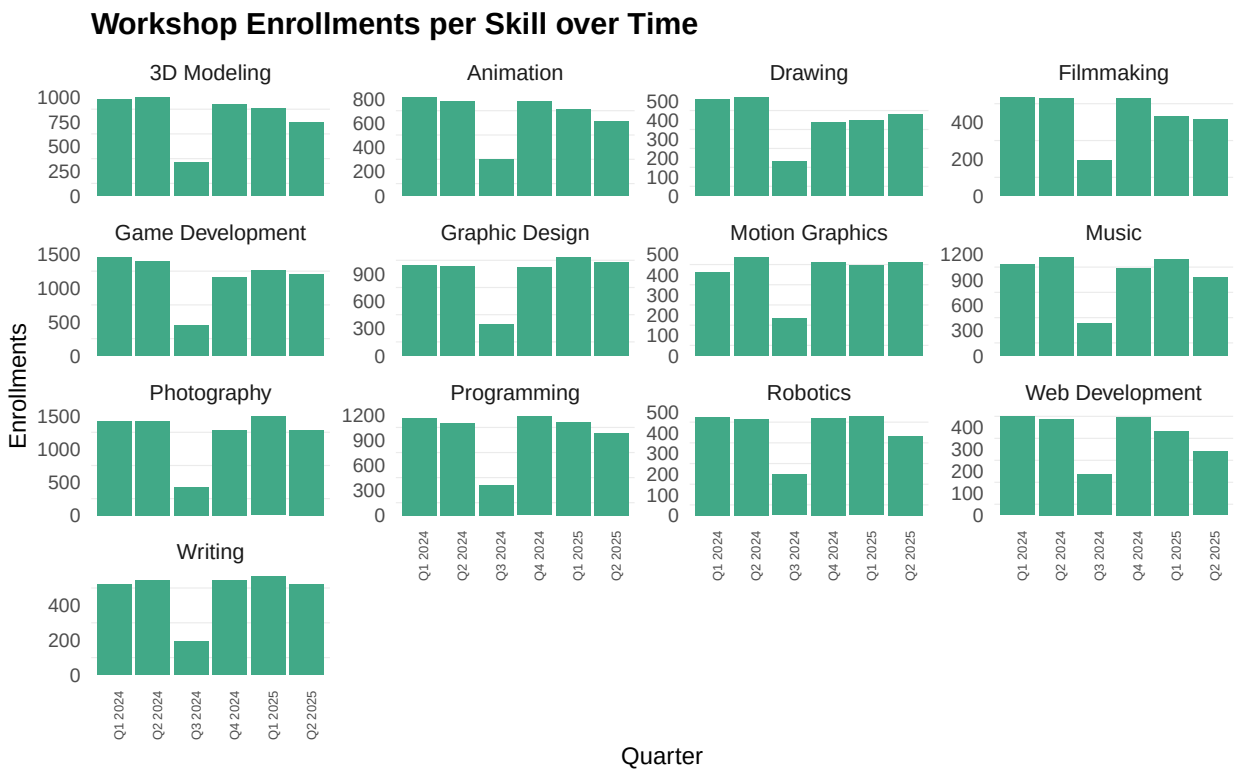
Skill popularity over Time

To see if student interest in different workshop skills changed over time, a Chi-squared test was used. The result was very clear: the p-value was far below 0.05, meaning there was a real difference. Some skills became more popular at certain times, while others dropped off. This can happen because of when skills are offered, school schedules, or changing trends.



This is shown in the line graph "Skill Popularity Over Time." Programming, Photography, Music, and Game Development stayed popular throughout the year. They often had over 500 students in peak months. Skills like Robotics, Writing, and Web Development had fewer enrollments and more ups and downs. All skills dropped in June 2024 due to fewer workshop offerings. By 2025, each skill showed its own trend. These insights help decide which skills to offer more often or promote more.

Skill Enrollment by Quarter



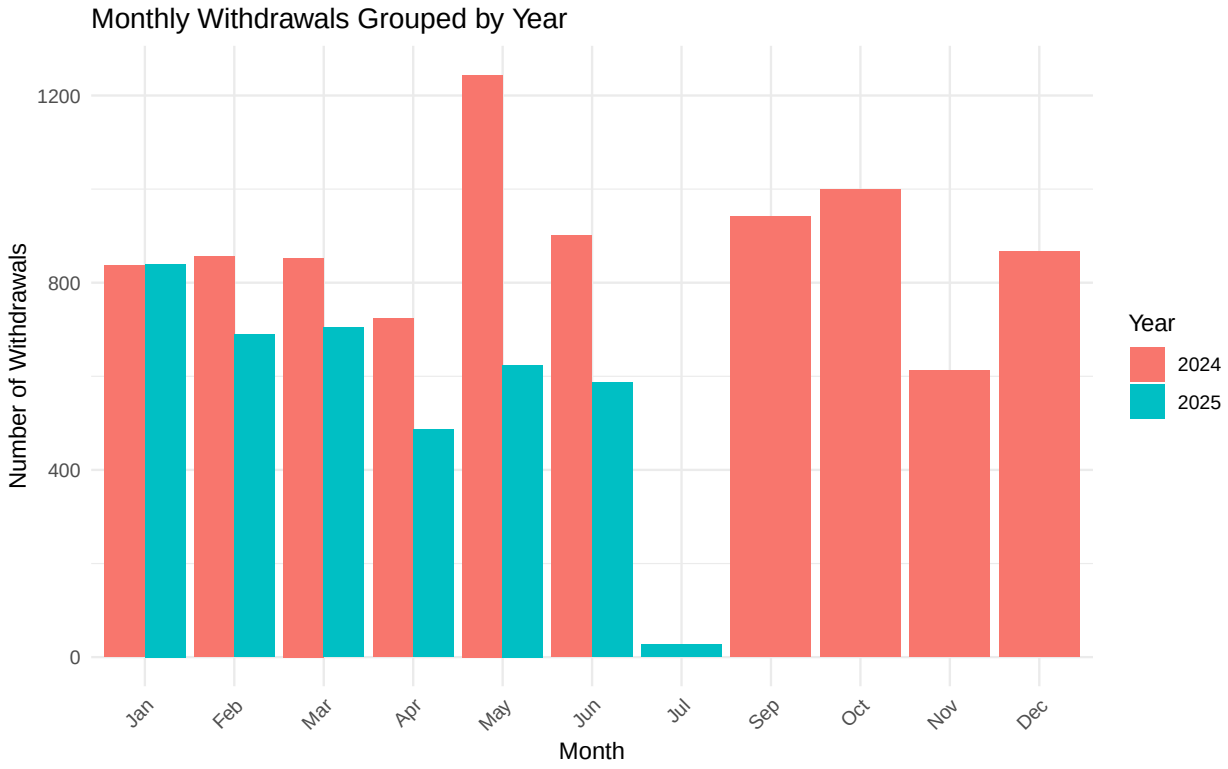
Looking at enrollments by quarter helps organize the trends more clearly. The biggest drop happened in Q2 2024 for all skills, again because of fewer offerings. Things picked up in Q3 and Q4, and by 2025, enrollment levels were stable again. Programming, Photography, Music, and Game Development had the most students each quarter — over 1,000 in many cases. Robotics, Web Development, and Writing had fewer students, but those numbers stayed steady. This shows that even smaller subjects have a loyal audience.

H2. Withdrawal Trends Over Time Hypothesis: Withdrawals are more frequent in certain months (e.g. May, September). This part looked at whether students are more likely to withdraw from workshops during certain months. A Chi-squared test showed a clear result: withdrawals are not evenly spread out. The data showed a big spike in May 2024, with over 1,200 students leaving. Other high months included September, October, and December — times that match busy academic periods like exams or school restarts.

H (Null): Withdrawals are evenly distributed across all months — no specific months have significantly higher withdrawal counts. H (Alternative): Withdrawals are not evenly distributed — some months (e.g., May, September) show higher withdrawal counts.

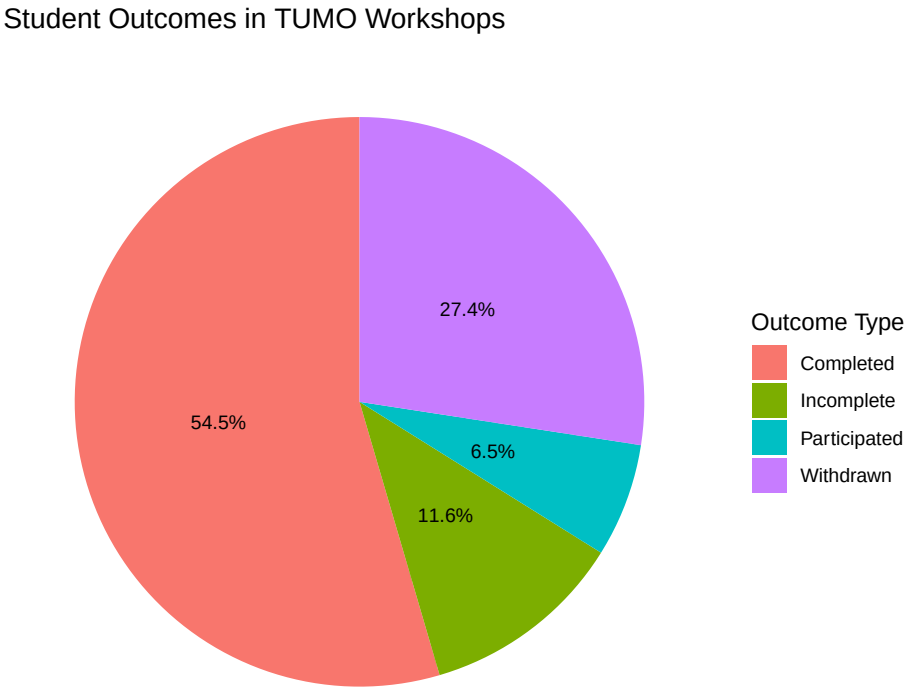
```
##
##  Chi-squared test for given probabilities
##
## data:  withdrawal_test$n
## X-squared = 2519.1, df = 10, p-value < 2.2e-16

## H  is supported: Withdrawals are not evenly distributed across months.
```

Monthy Withdrawals: 2024 VS 2025



July had almost no withdrawals, which matches the center's summer break. In 2025, the same pattern continued but with slightly fewer total withdrawals. This shows that external factors, like school schedules, affect when students drop out. To help students stay engaged, the center could offer more flexible options or support in months like April and August before the usual withdrawal spikes.

Workshop Outcome Types



The pie chart shows that 57.7% of students completed their workshops. But 12.3% were Incomplete and 6.8% Participated without finishing. Another 23.2% withdrew. That means over 42% of students didn't finish their workshops.

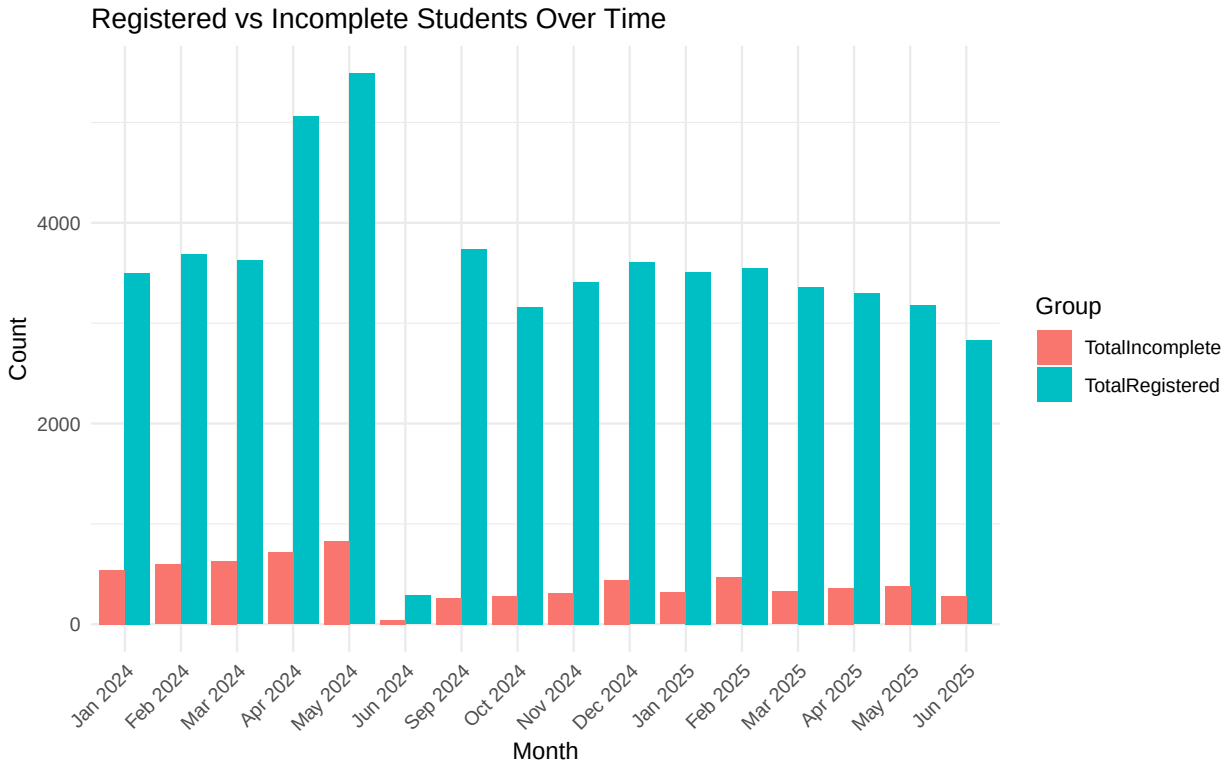
H3. Failure Trends Over Time Hypothesis: The frequency of unsuccessful outcomes (Incomplete, Complete) varies seasonally. This section studied whether failure — meaning "Incomplete" or "Participated" statuses — happens more in some seasons than others. A Chi-squared test showed a strong difference. The test result was significant, with a p-value much lower than 0.05.

H (Null): The distribution of unsuccessful outcomes (Incomplete, Complete) is the same across all seasons (or months). H (Alternative): At least one season/month has a significantly higher failure rate (i.e., Incomplete or Participated).

Chi-square test: Failure types vs Season

```
##
##  Pearson's Chi-squared test
##
## data:  failure_season_table
## X-squared = 537.25, df = 3, p-value < 2.2e-16

## H is supported: Failure outcomes vary significantly by season.
```

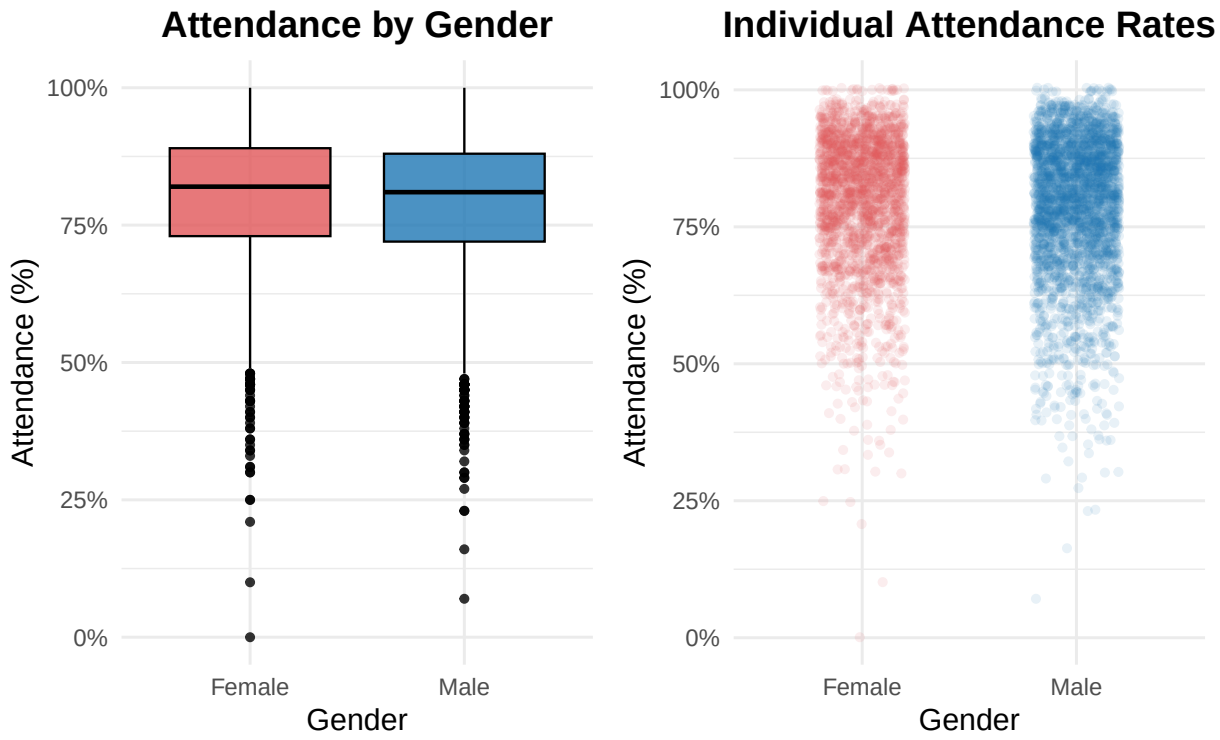
Registered VS Incomplete Students Over Time



The bar chart "Registered vs Incomplete Students Over Time" highlights when these failures happened. The biggest spikes were in April, May, and September 2024, which match school transitions or exam seasons. In contrast, failure rates were low in June, July, and August — periods with fewer or no workshops. These seasonal patterns suggest that timing matters a lot. The center could reduce failure by offering shorter or more flexible workshops during high-risk months and by checking in with students more often during spring and early fall.

Student behavior follows a clear seasonal rhythm. Participation rises in spring, drops in summer, and stays steady during the school year. Skill interest changes over time. Withdrawals and failure rates rise during academic stress periods like May and September. And new students mostly join early in the year or after summer, while returning students drive most of the engagement later on.

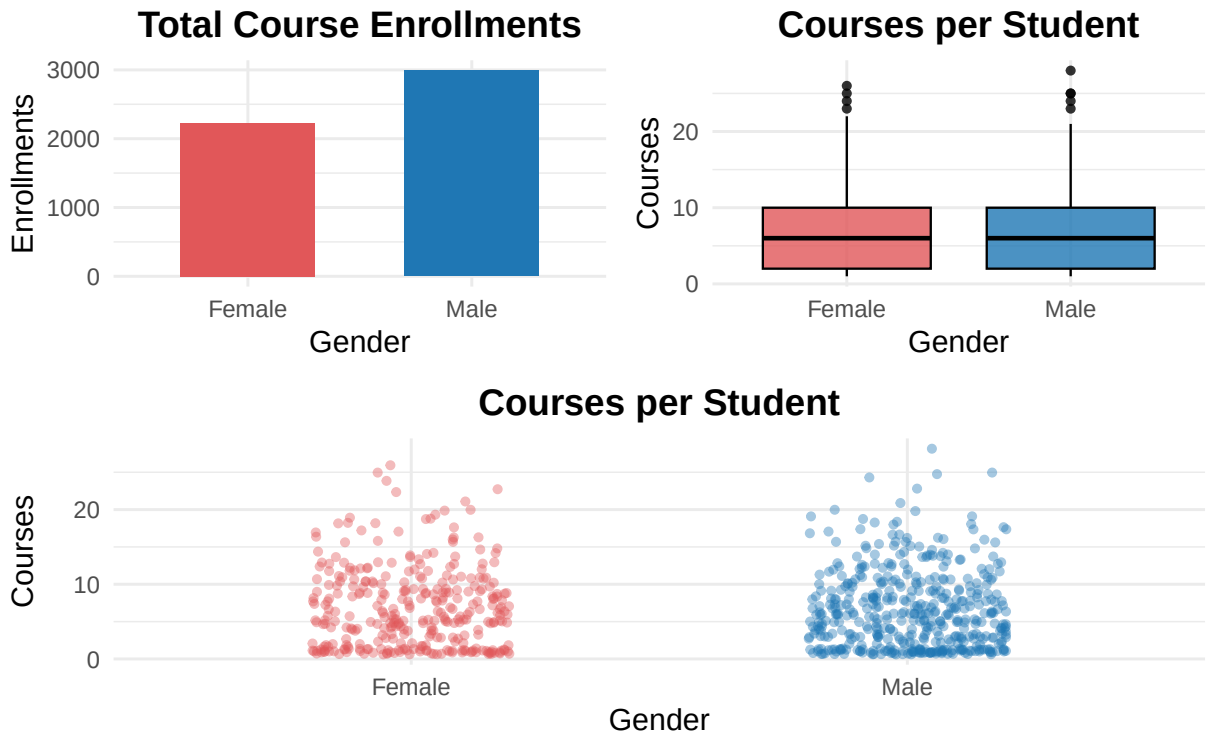
Attendance by Gender



These two graphs show attendance percentages for students, grouped by gender. On the left part, you can see boxplots by which we can see that both medians of the genders are high, nearly 80-82%, so despite the gender, all the students have a high attendance rate; however, outliers are also a lot, and many students have a 25-50% participation rate. Females have a slightly higher median, which is why we can conclude that female students have more consistent attendance. Overall, there is no extreme gap between genders. On the right is the scatterplot for the individual attendances. Each dot represents one student, colored by gender. This gives a deeper look at how individual attendance varies. The densest part again was between 75-98 percent, and here we can see the attendance of each individual, unlike the box plot, which shows the summary of the attendance rates.

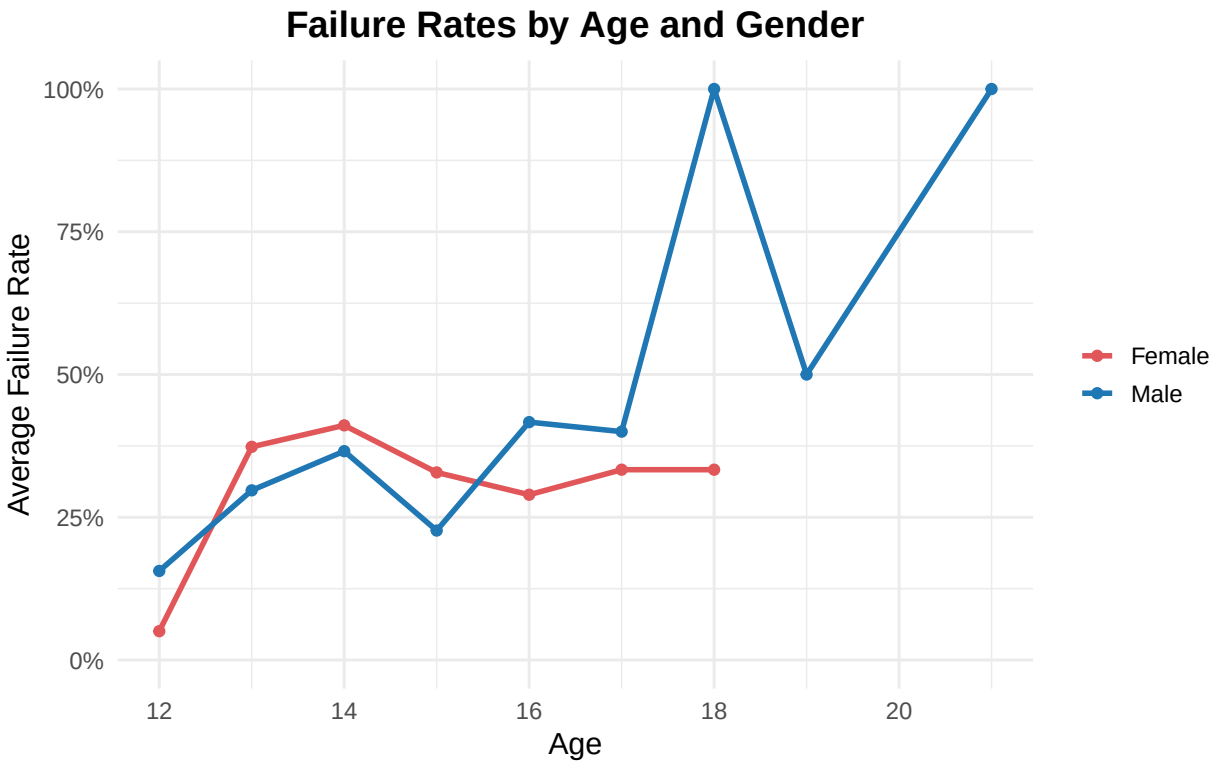
There's no obvious gap between male and female groups — both show very similar behavior.

Gender-Based Course Enrollment Analysis

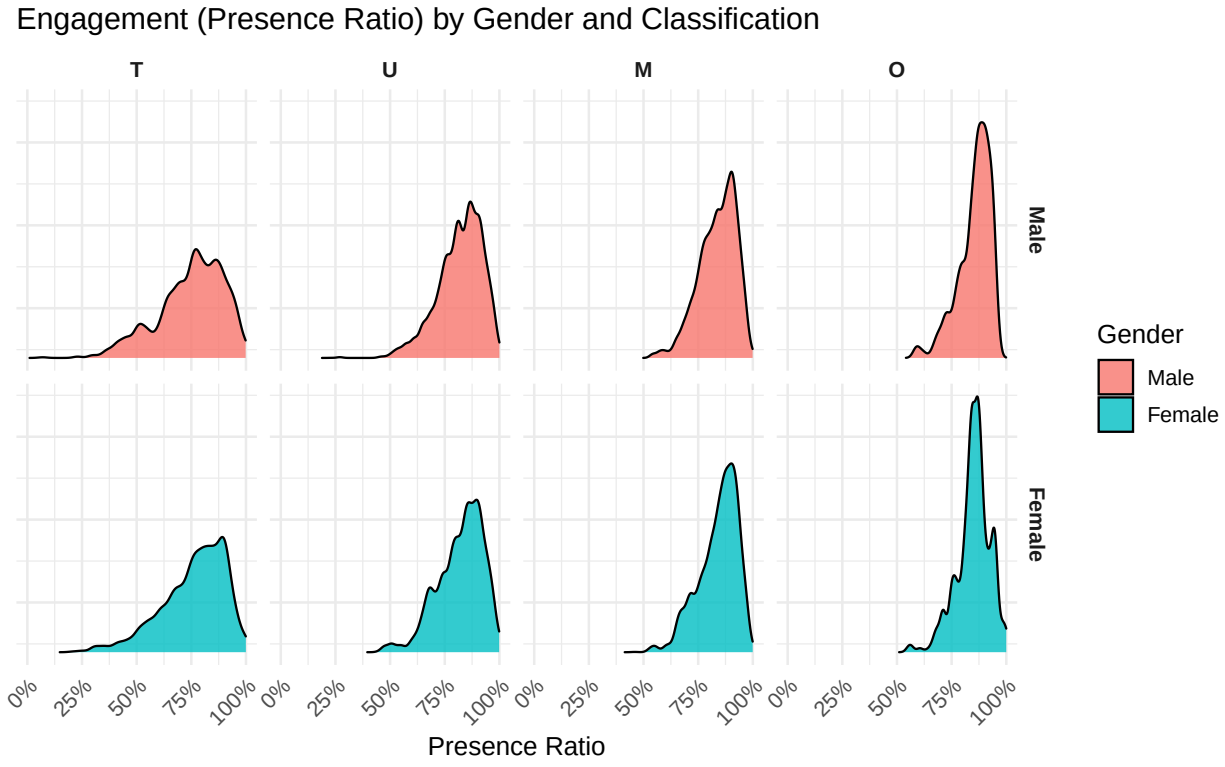


Here are three graphs: a bar chart, a box plot, and a scatterplot for the Course enrollments of the students. In the first graph, the bar chart, we can see that Males have a higher total enrollment (3000 total courses) than Female students (2200 total courses). The box plot shows that medians for the individual enrollments are nearly equal, with a tiny female domination. Each student, on average, took six courses. By scatterplot, we get a more detailed picture. A significant number of the students take between 1 and 2 courses, while others take 2-15 classes. Only a few students take 15 or more courses. Both genders have a similar concentration of students in the mid-range. Male students have slightly more variation, including more students who take over 20 courses. Overall, the patterns are comparable, with no huge gender differences.

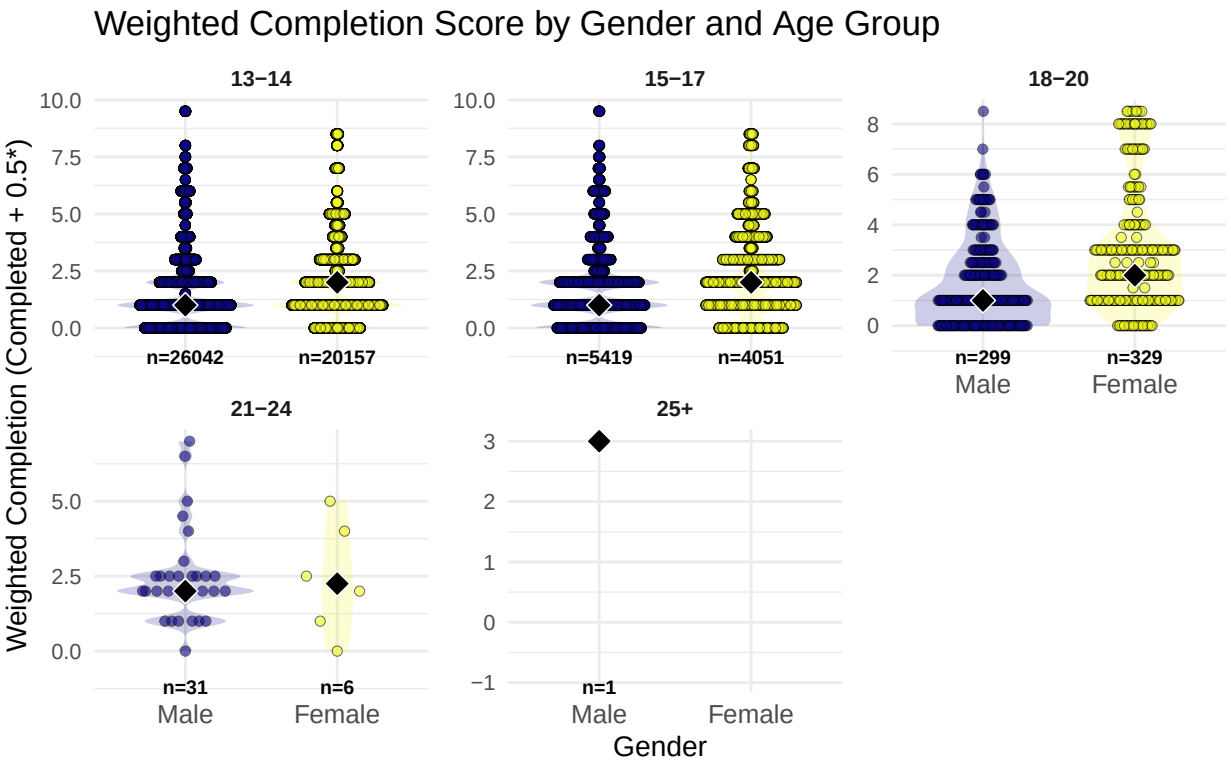
Failure Rates by Age & Gender



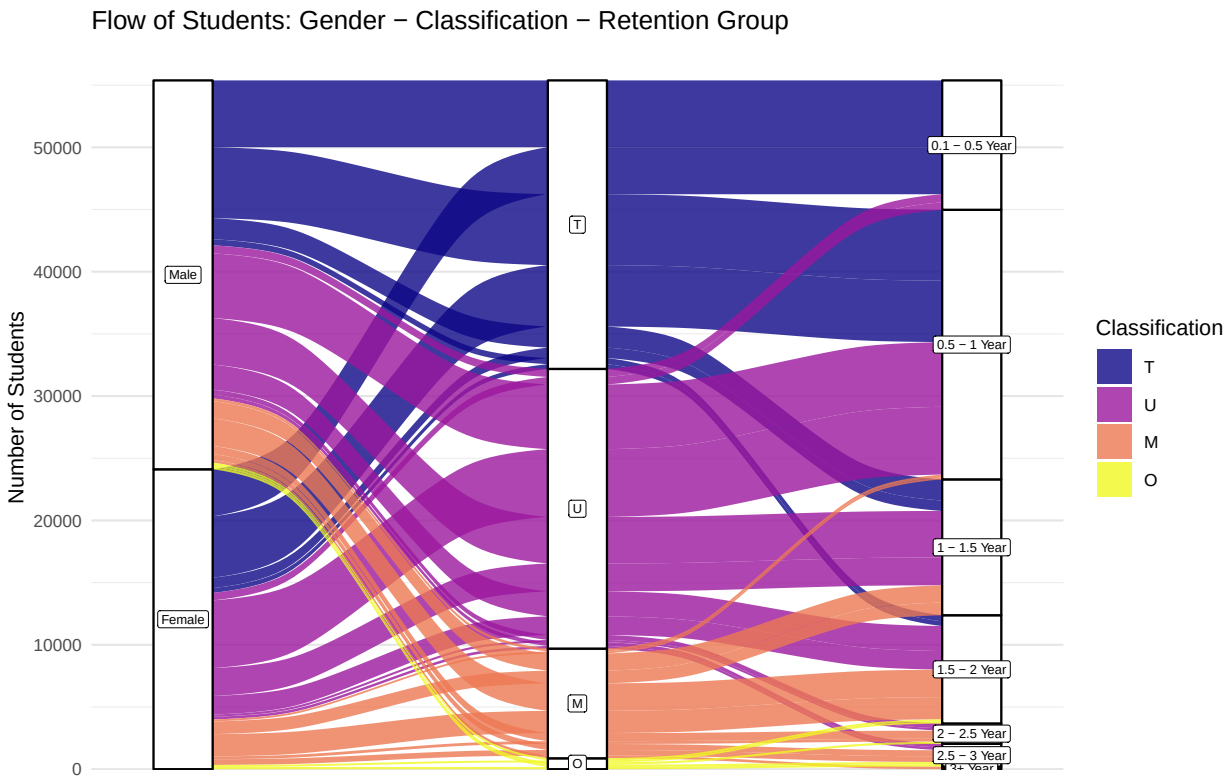
This graph shows how average failure rates change across different ages and how those patterns differ by gender. Here we can see that failure rates increase with age, especially for male students. For younger students (12–15), failure rates are fairly close for both genders. Starting from age 16 and older, male failure rates spike significantly. At age 18 and 21, failure rate for males hits 100%, meaning every male student at those ages failed. Female failure rates stay more stable and lower compared to males. Even at older ages, it remains below 40%. So in all ages female students have more concentrated and consistent behaviour in studying than the male students.



The plot shows that most students—both male and female—have high presence ratios (around 75–100%), indicating strong engagement. Gender differences are minimal within each classification, as the distributions look similar. Some classifications (like "M" and "O") have tighter peaks near high attendance, suggesting more consistent engagement, while others show slightly more spread. Very few students fall into low-engagement territory.



The plot shows weighted completion scores (completed + 0.5) stratified by gender and age group. For the larger cohorts (13–14 and 15–17), males and females have broadly overlapping distributions with medians near each other, indicating little gender gap in completion when sample sizes are substantial. In the 18–20 group, females appear to have a slightly higher central tendency, but the difference is modest. The smaller samples in 21–24 and especially 25+ (notably the single male in 25+) make those panels unstable—interpret those with caution. Overall, completion performance is fairly comparable across gender within age bands, with no dramatic disparities.



The alluvial shows how students flow from gender into classification and then into retention cohorts. Most students (both male and female) end up in the "U" and "T" classifications, with substantial representation in mid-to-longer retention groups (e.g., 0.1–0.5 year and 0.5–1 year), suggesting moderate persistence. There's no stark gender imbalance in the major flows, though the thickness of bands can hint at slight differences in how genders distribute across classifications. Lower classifications (like "M" and "O") feed more into shorter retention spans, implying those groups tend to drop off sooner.

This part is identifying age groups where both genders have enough data (at least two students and more than one unique weighted completion score) to make a meaningful comparison, then running Wilcoxon rank-sum tests within each of those age groups to see if the distribution of weighted completion scores differs by gender. In short: it's a nonparametric subgroup analysis by age to detect gender gaps in weighted completion while guarding against spurious findings from multiple comparisons.

```
## # A tibble: 4 x 10
##   age_group .y.          group1 group2    n1    n2 statistic         p     p.adj
##   <fct>     <chr>        <chr>  <chr>  <int> <int>     <dbl>     <dbl>     <dbl>
## 1 13-14     weighted_co~ Male   Female 26042 20157    2.17e8 7.57e-238 3.03e-237
## 2 15-17     weighted_co~ Male   Female  5419  4051    9.07e6 1.73e- 49 3.46e- 49
## 3 18-20     weighted_co~ Male   Female   299   329    3.21e4 1.85e- 14 2.47e- 14
## 4 21-24     weighted_co~ Male   Female    31     6    9.35e1 1   e+  0 1   e+  0
## # i 1 more variable: p.adj.signif <chr>
```

The Wilcoxon tests show highly significant gender differences in weighted completion score for the 13–14, 15–17, and 18–20 age groups (adjusted p's 0.001, "****"), but no evidence of a difference in 21–24 (adjusted p = 1, "ns")—likely because that cell is tiny (n=31 vs n=6). Given the large sample sizes in the younger students, these results indicate a real shift in the distribution of completion scores between males and females there.

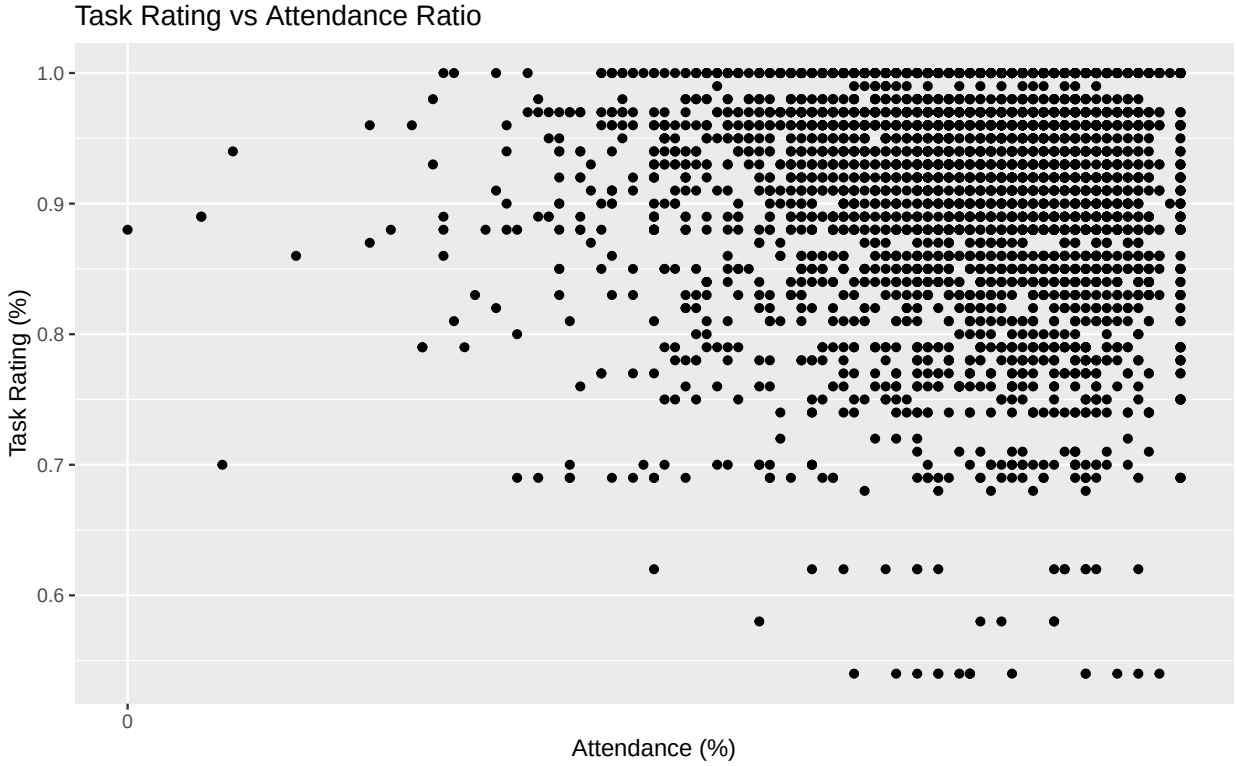
```
##
##  Pearson's Chi-squared test
##
## data:  .
## X-squared = 3313.2, df = 4, p-value < 2.2e-16
```

The chi-squared test shows a very strong association between classification and withdrawal status (X^2 = 3313.2, df = 4, p < 2.2e-16), so you can reject the null of independence—withdrawals are not evenly distributed across classifications. In other words, some classifications have disproportionately higher or lower withdrawal rates.

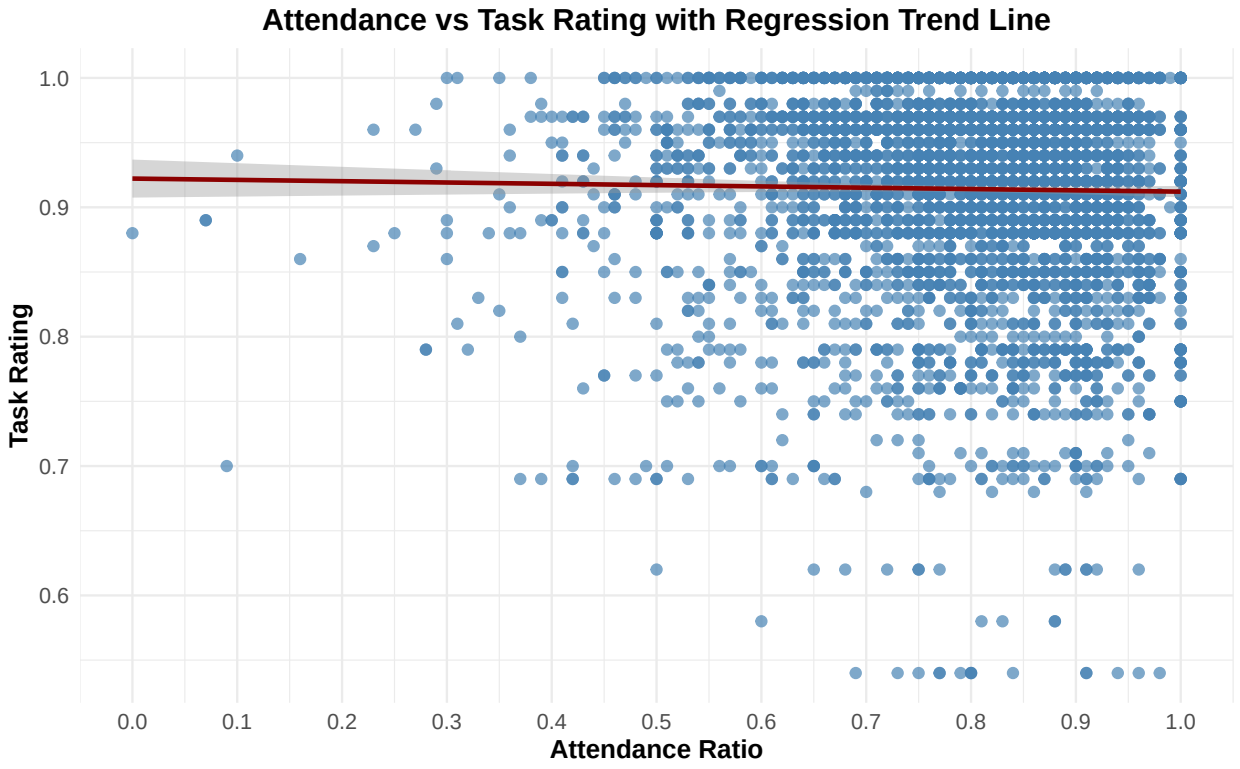
Hypothesis 6: Attendance correlates with student performance

Assumption: Students with higher attendance are more likely to perform better (complete courses) compared to students with low attendance. Rationale: Students who are actively attending classes may engage more with the material, leading to better performance.

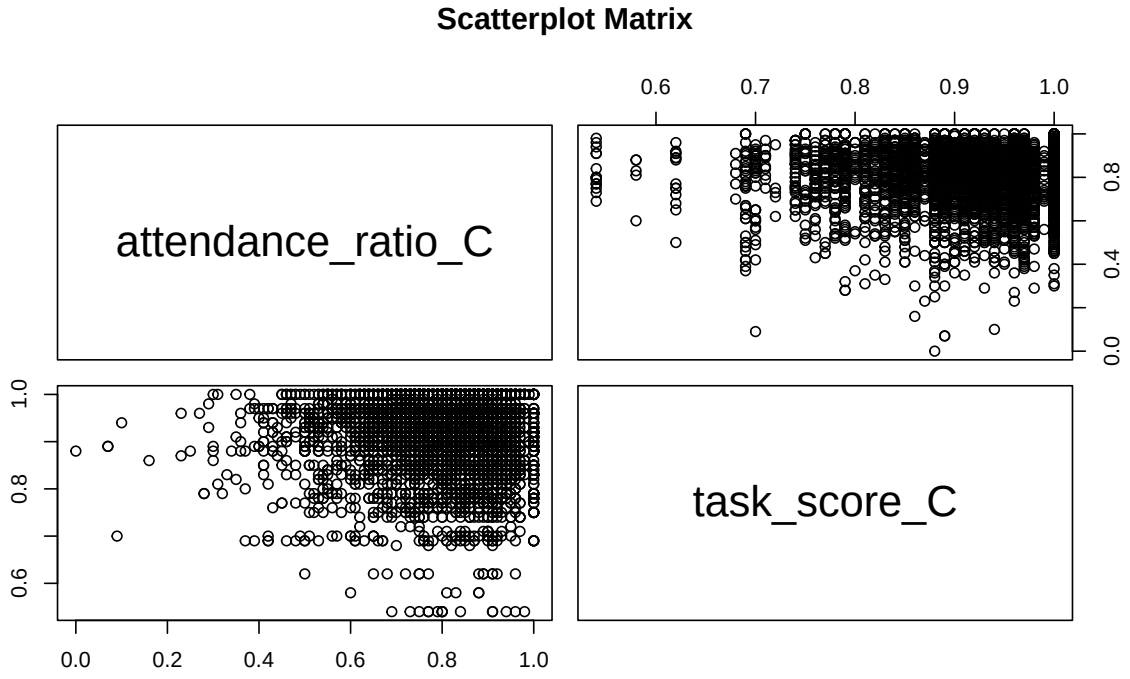
This section explores whether students who attend workshops more regularly tend to perform better in their tasks. To measure this, we compared each student's attendance ratio with their task rating — the proportion of tasks they successfully completed. A first look at the relationship is provided by a scatterplot of Task Rating vs Attendance Ratio.



This plot shows most students clustered in the upper-right corner, with both high attendance and high task ratings. However, there is noticeable spread, indicating that some students perform well even with lower attendance. To explore the pattern further, a regression line is added to the scatterplot. ### Scatterplot: Presence Rate & Task Rating Trend



The nearly flat slope suggests a weak linear relationship.



A scatterplot matrix shows similar results — a loose pattern with a high concentration of students achieving strong results. For a more detailed statistical view, we examined both linear and monotonic correlations.

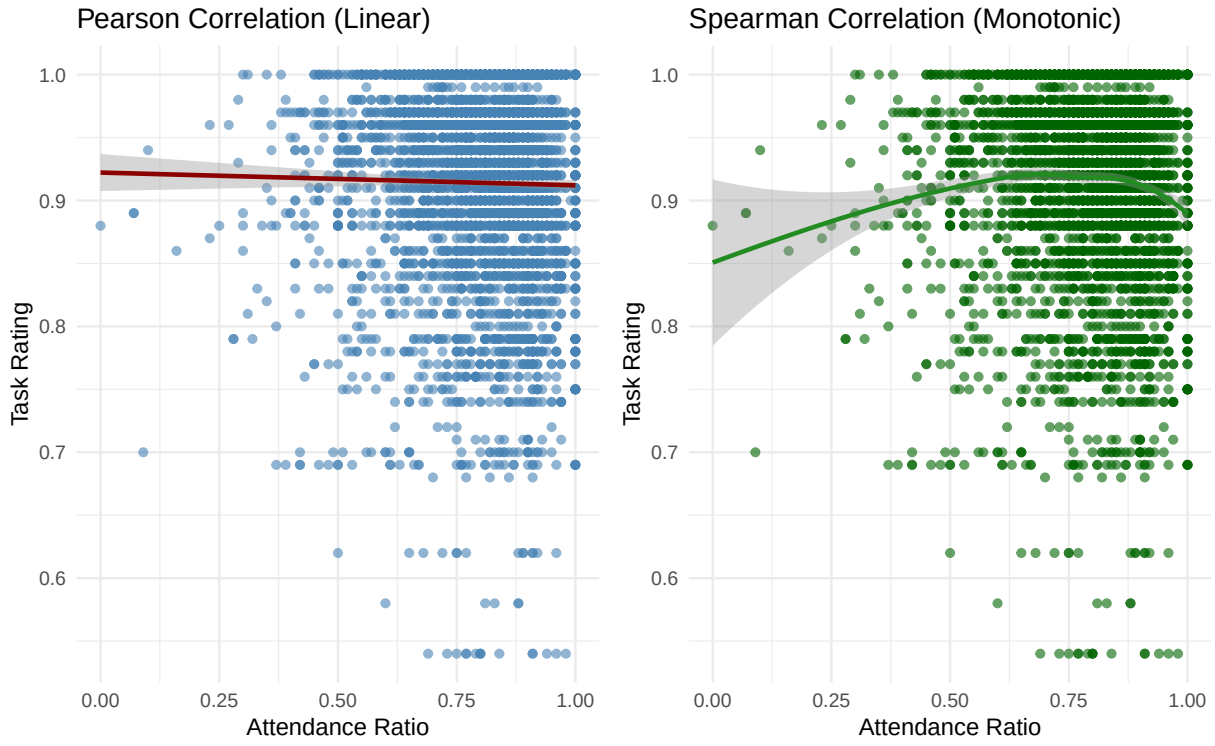
The Pearson correlation testshowed a non-significant relationship (r = -0.017, p = 0.28), while the Spearman correlation revealed a weak but statistically significant trend (= -0.055, p < 0.001). To visualize this comparison, we included two plots showing trend lines — one for linear ### Linear Relationship: Presence Rate VS Task Rating

```
##
##  Spearman's rank correlation rho
##
## data:  merged_C$attendance_ratio_C and merged_C$task_score_C
## S = 13193615048, p-value = 0.0003643
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.05486129
```

Monotonicity: Presence Rate VS Task Rating

```
##
##  Pearson's product-moment correlation
##
## data:  merged_C$attendance_ratio_C and merged_C$task_score_C
## t = -1.0881, df = 4216, p-value = 0.2766
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04691147  0.01343134
## sample estimates:
##         cor
## -0.01675532
```

Linearity and Monotonicity Plots



These plots suggest that while attendance alone is not a strong predictor of performance for most students, those with very low attendance — especially below 40% — are more likely to have reduced task ratings. Although the relationship is subtle, the Spearman result confirms that very low attendance is linked to lower success.

Hypothesis 7: Withdrawn students show different behavioral patterns than those who fail or complete courses.

Assumption: They have distinct engagement traits (e.g., lower attendance or fewer tasks completed). Rationale: Withdrawals may stem from personal or motivational issues, reflected in measurable behavior.

T-Test: Withdrawn vs Completed

```
##
##  Welch Two Sample t-test
##
## data:  attendance_ratio_C by primary_outcome_C
## t = 6.3982, df = 2091.9, p-value = 0.0000000001934
## alternative hypothesis: true difference in means between group Completed and group Withdrawn is not
## 95 percent confidence interval:
##  0.02077441 0.03913800
## sample estimates:
## mean in group Completed mean in group Withdrawn
##               0.8026398               0.7726836
```

T-Test: Withdrawn vs Failed

```
##
##  Welch Two Sample t-test
##
## data:  attendance_ratio_C by primary_outcome_C
## t = 5.8457, df = 1558.9, p-value = 0.000000006131
## alternative hypothesis: true difference in means between group Failed and group Withdrawn is not equa
## 95 percent confidence interval:
##  0.02520169 0.05065455
## sample estimates:
##    mean in group Failed mean in group Withdrawn
##               0.8106117               0.7726836
```

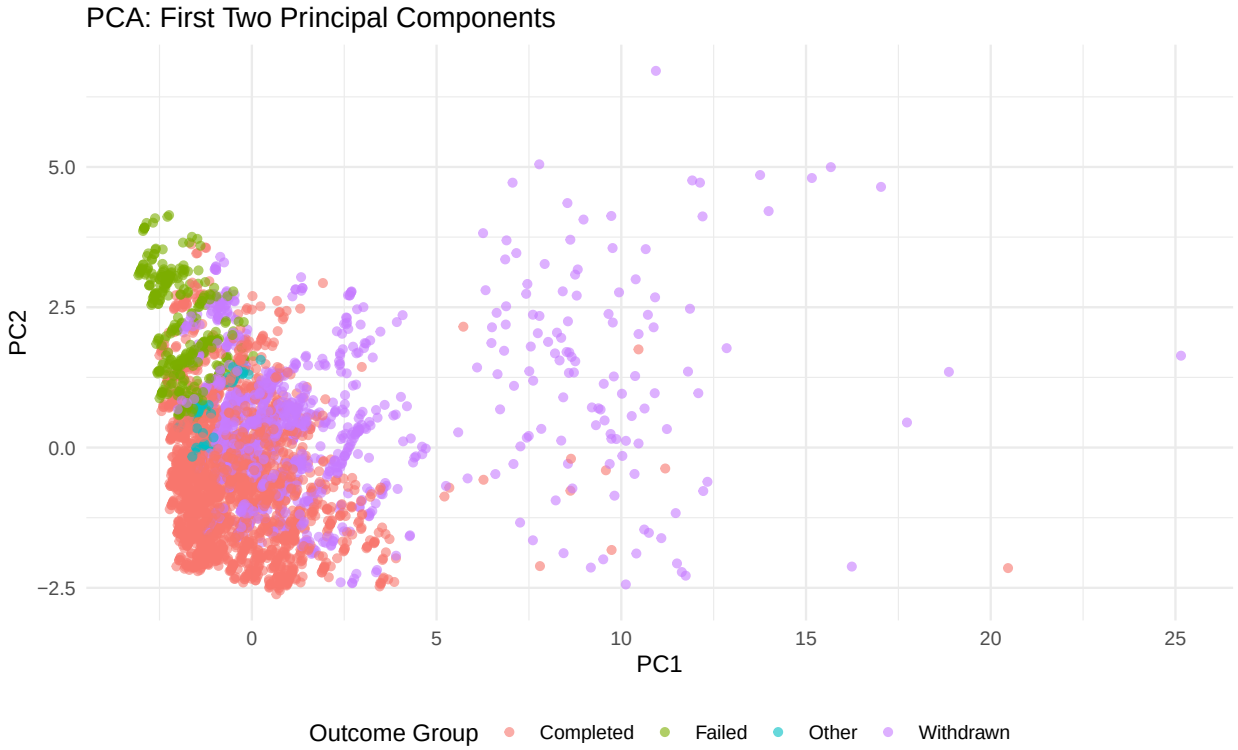
We grouped students by their main outcome: Completed, Failed, or Withdrawn. A two-sample t-test showed that students who withdrew had significantly lower average attendance than both the Completed and Failed groups. This confirms that low attendance is a clear pattern among withdrawn students. To explore behavioral differences further, we used Principal Component Analysis (PCA) to reduce the data and visualize patterns across all numeric variables.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) reduces high-dimensional data into fewer dimensions by transforming correlated variables into uncorrelated components, where each component captures the maximum possible variance in the data.

We use prcomp() with centering and scaling to ensure equal weighting across features.

PCA Plot: First Two Principal Components



The PCA scatterplot shows students positioned by their first two principal components, with colors indicating their outcome group. Withdrawn students tend to cluster away from others, suggesting they have distinct engagement patterns. These findings suggest that withdrawal is not a random outcome. Students who withdraw often exhibit early warning signs — particularly low attendance — that separate them from other groups. Recognizing this allows educators and mentors to intervene earlier and offer support to at-risk students.