Analysis of TUMO Student Data

DS116 / CS343

Rita, Hayk A, Hayk G, Gor, Aram

August 4, 2025

Outline

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. analysis.

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.

Overall Age Distribution

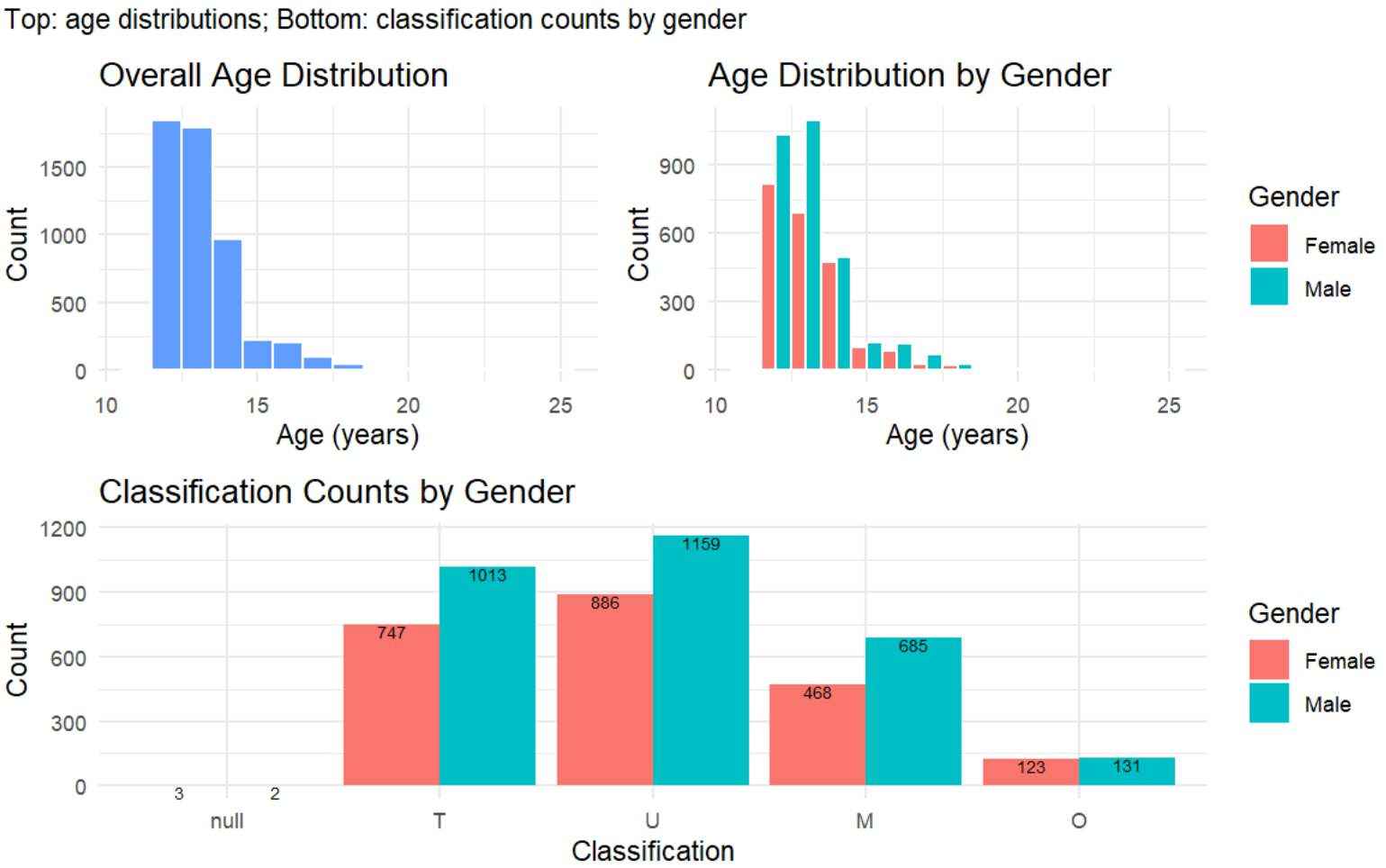

Figure: Figure: Overall Age Distribution.

Boxplot of Age per Classification

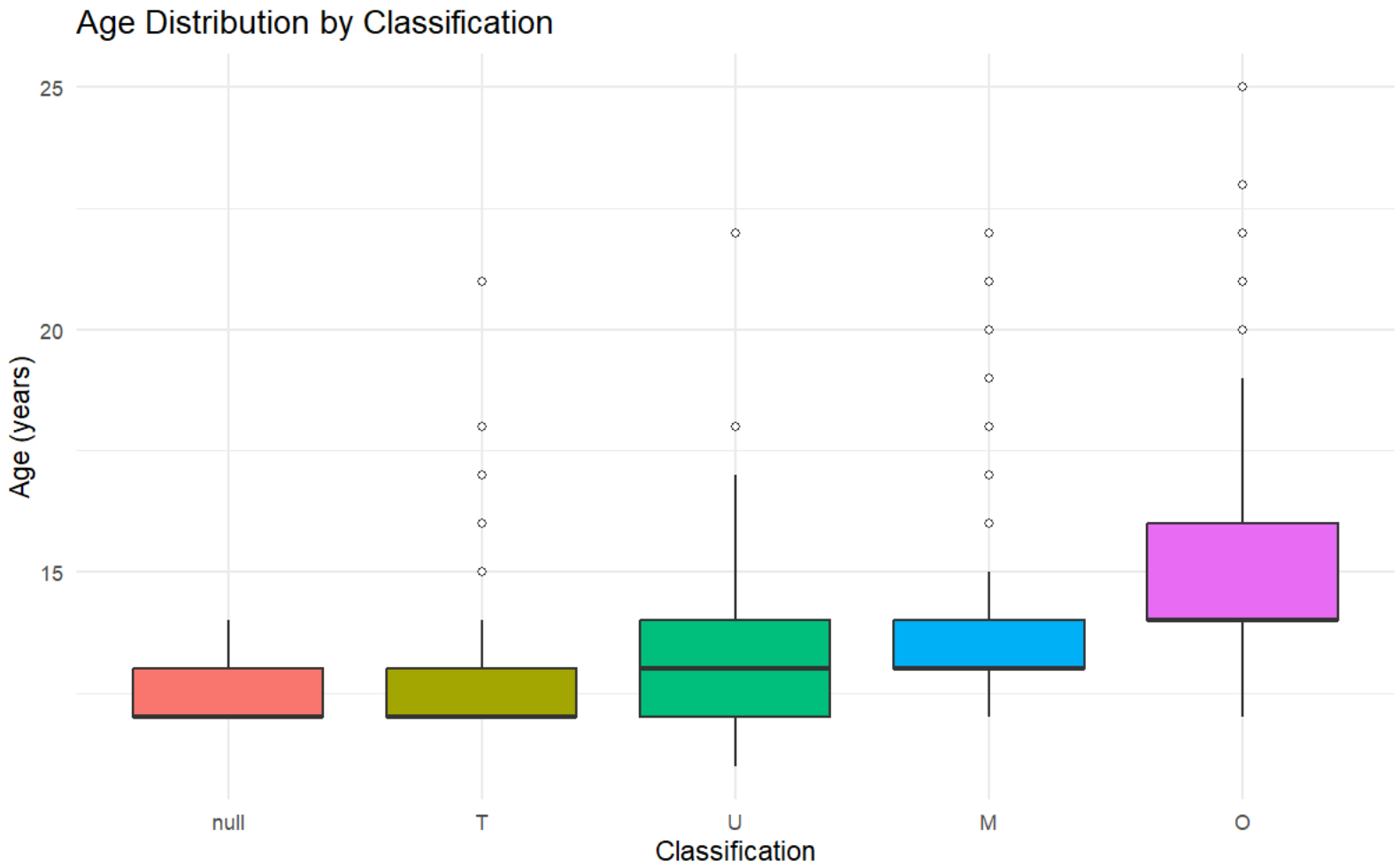

Figure: Figure: Boxplot of Age per Classification.

Age by Classification & Gender

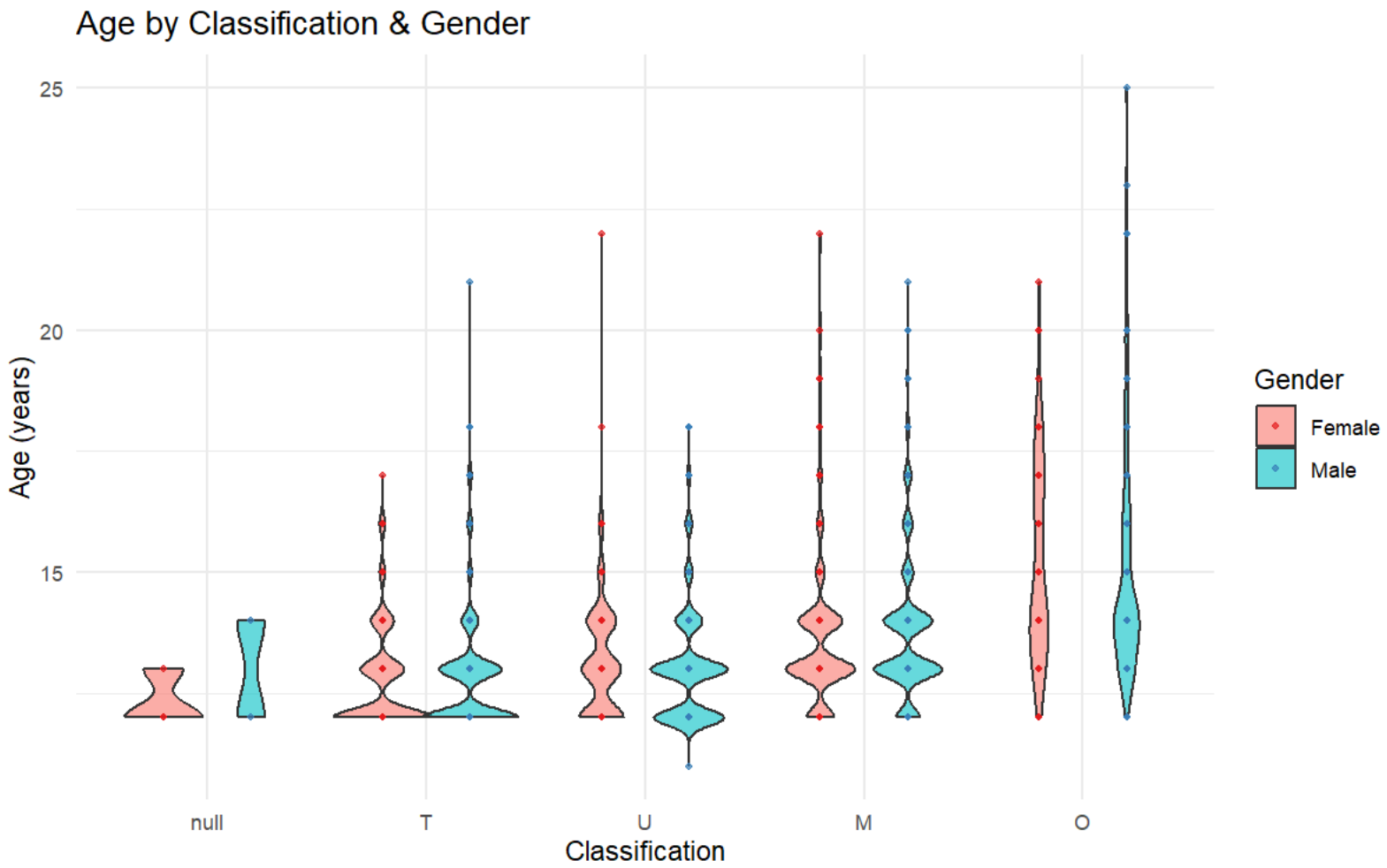

Figure: Figure: Age by Classification & Gender.

Student Dynamics & Learning Patterns

- Trends in enrollment, completion, withdrawal over time.

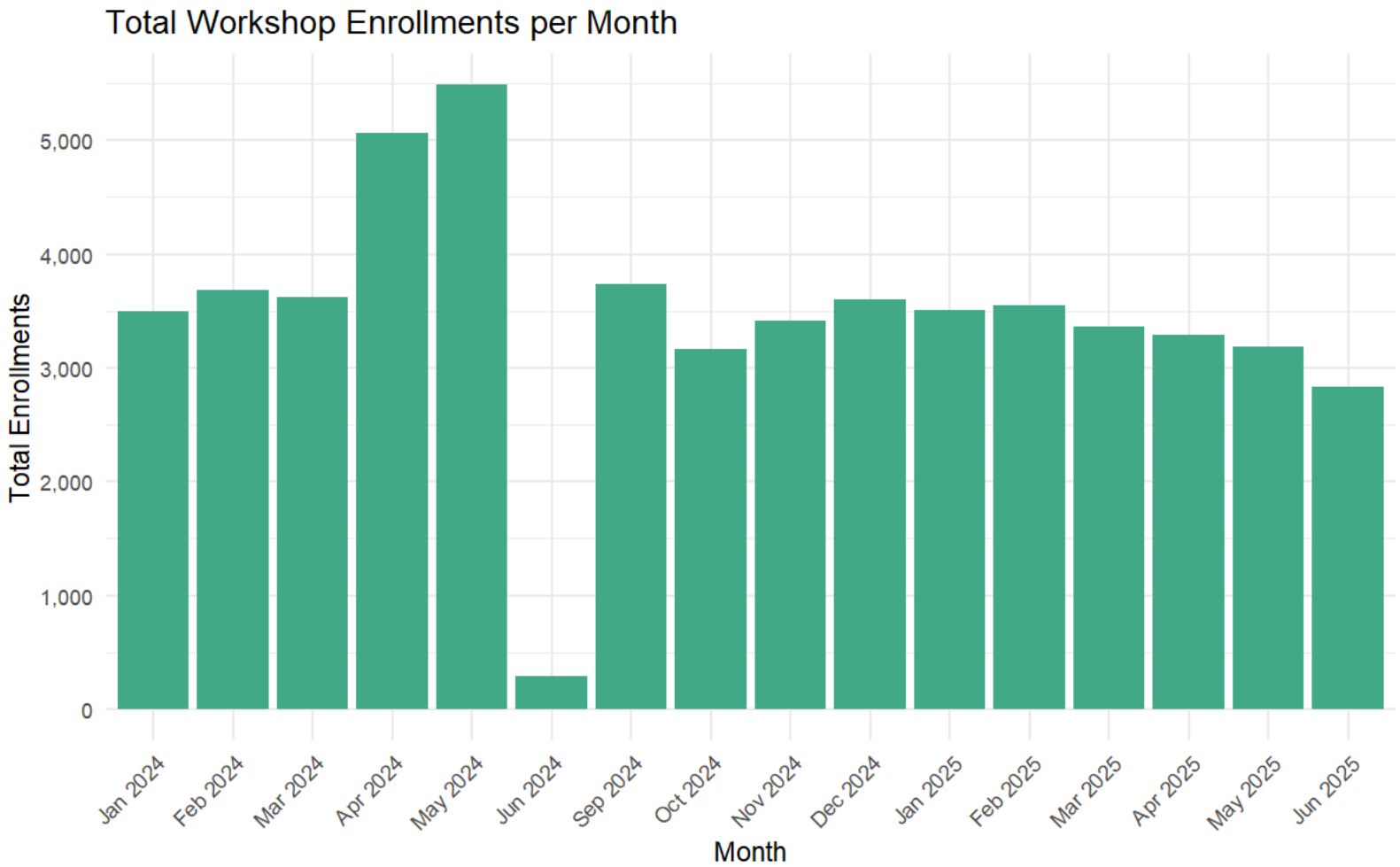

Figure: Figure: Student Dynamics & Learning Patterns.

Skill Popularity over Time

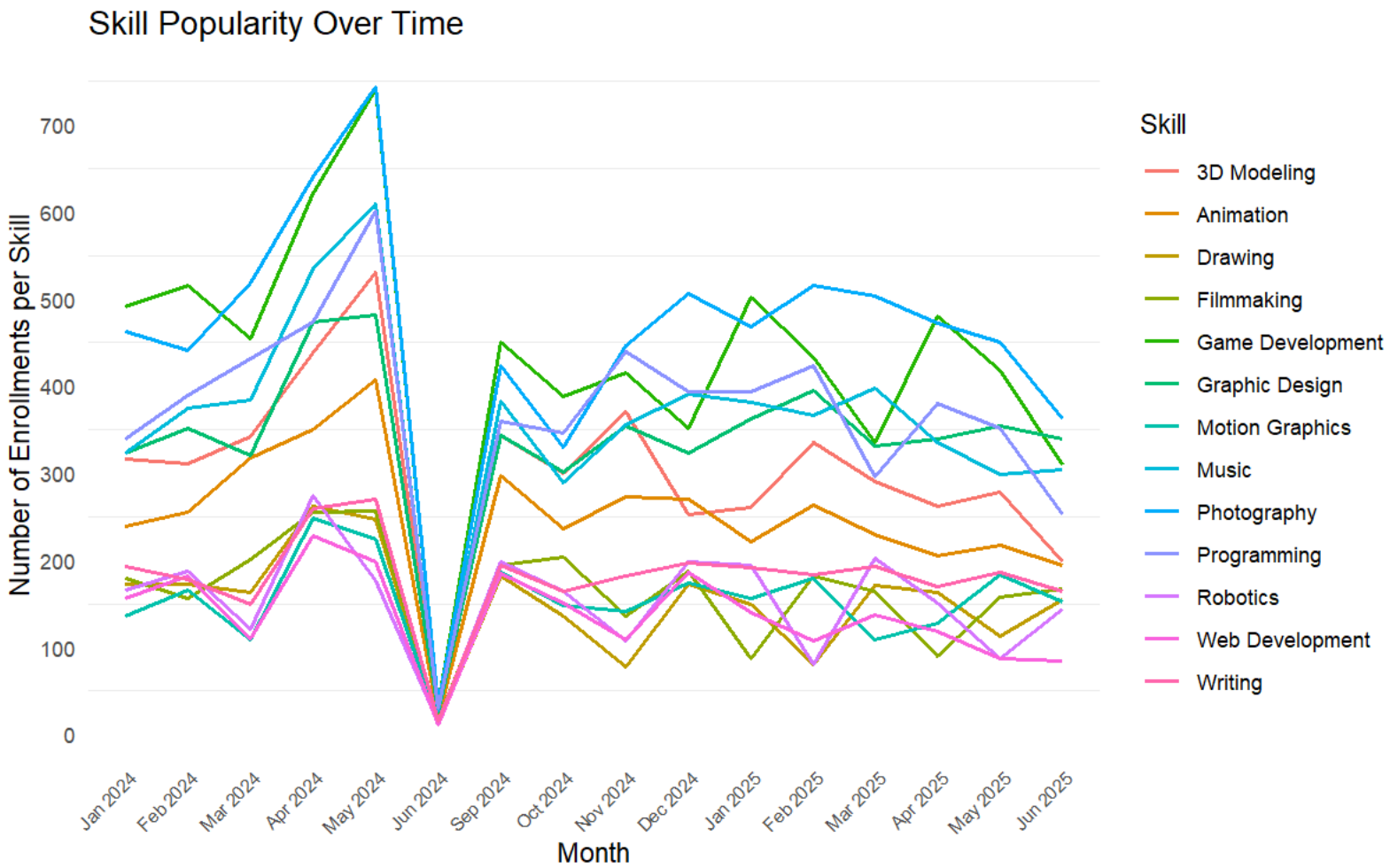

Figure: Figure: Skill Popularity over Time.

Skill Enrollment by Quarter

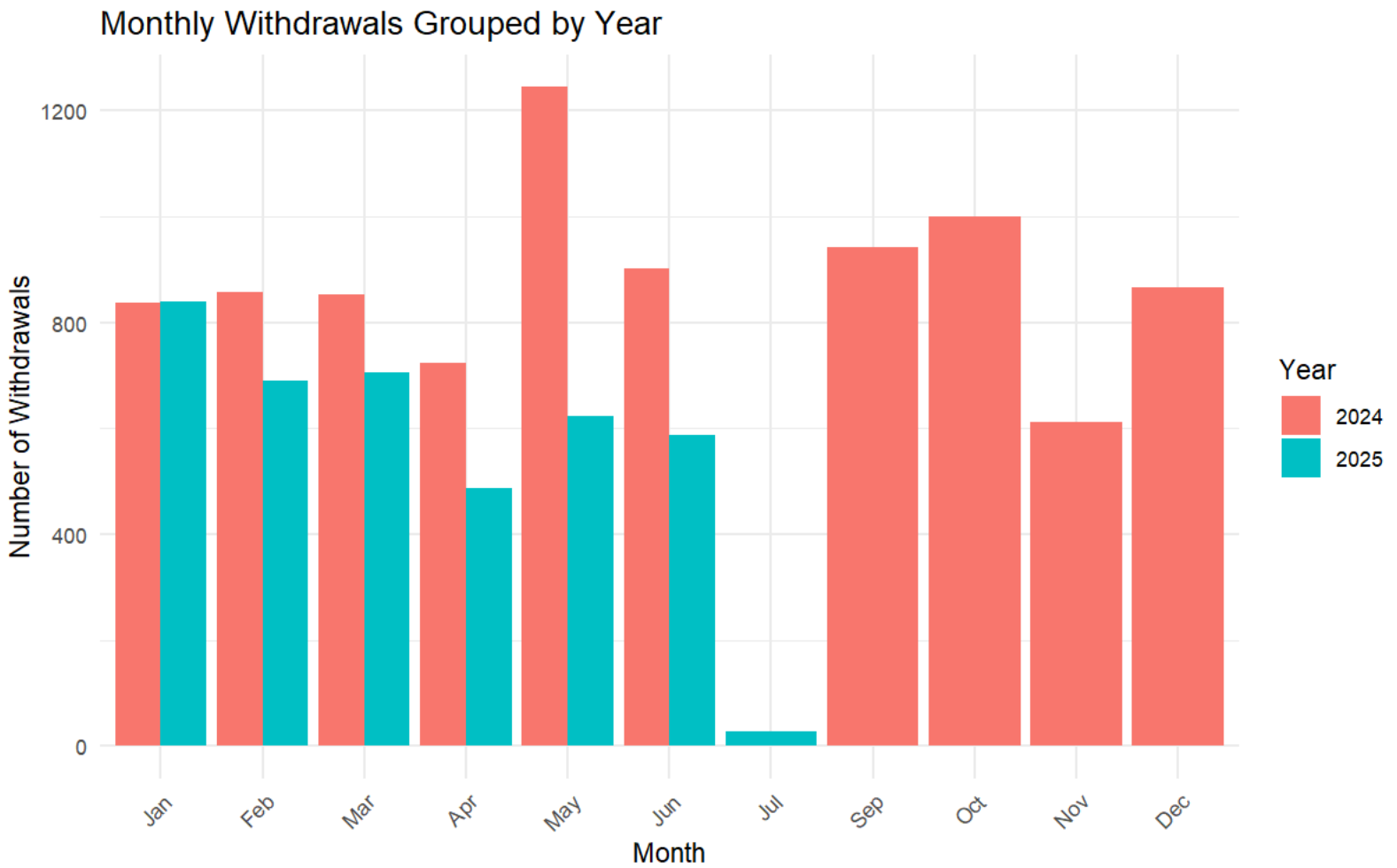

Figure: Figure: Skill Enrollment by Quarter.

Monthly Withdrawals: 2024 vs 2025

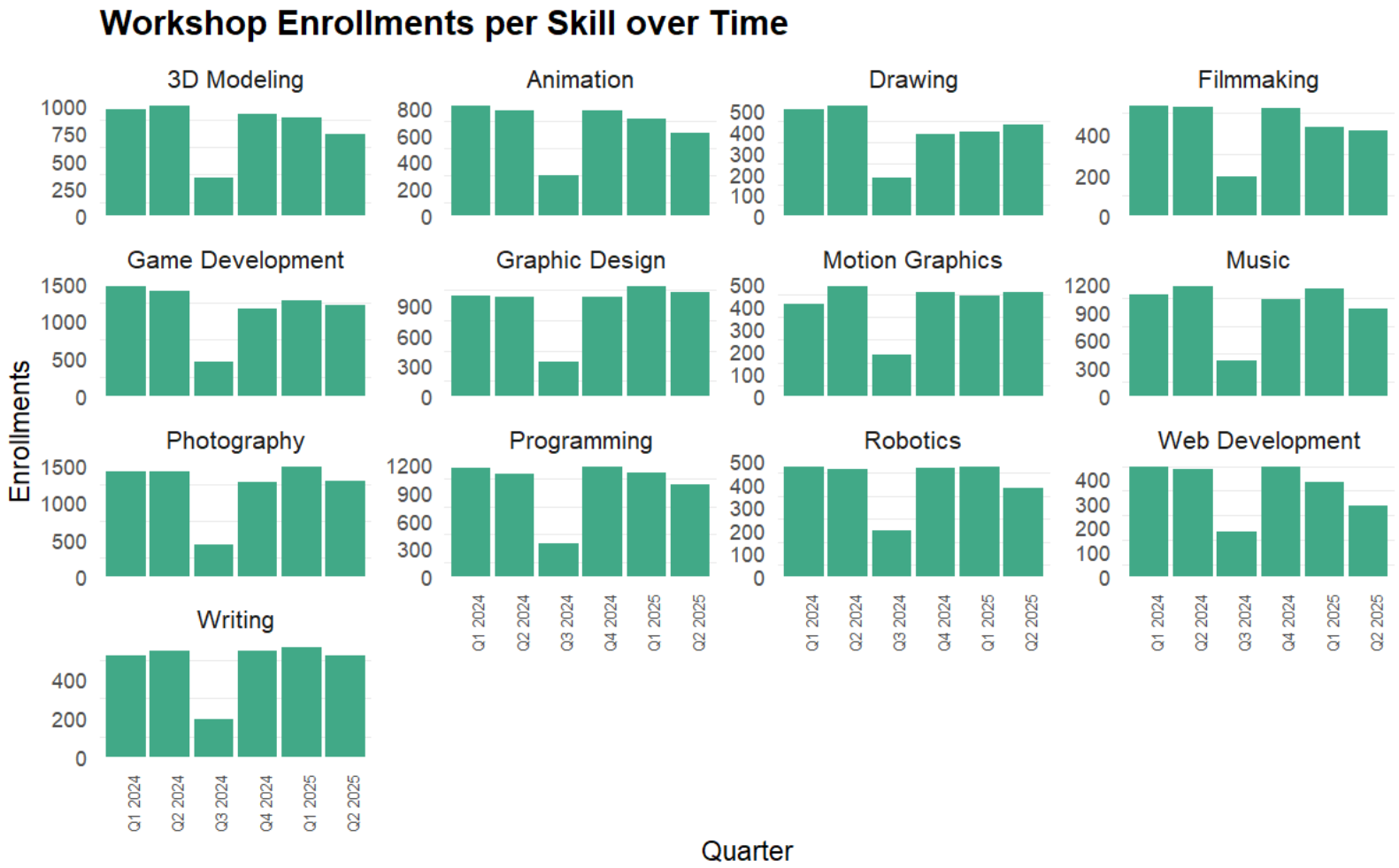

Figure: Figure: Skill Enrollment by Quarter.

Workshop Outcome Types

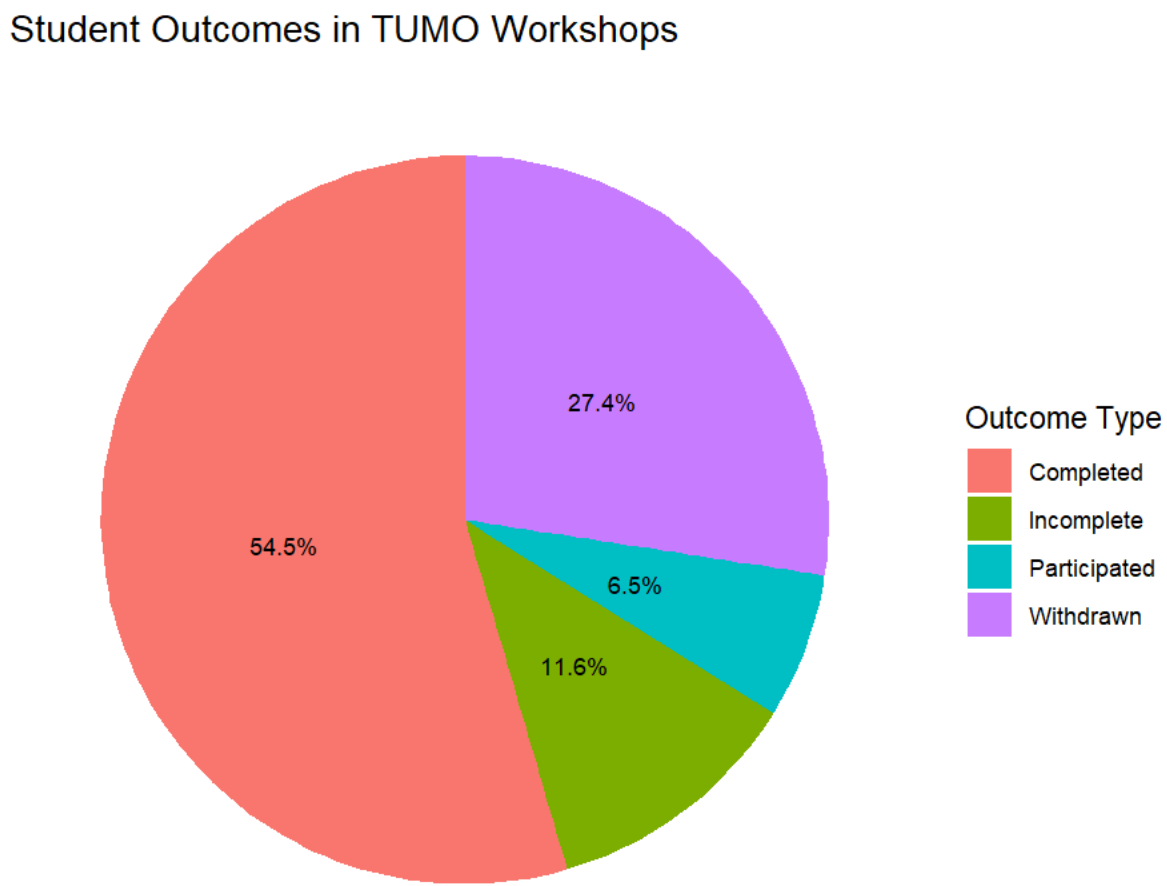

Figure: Figure: Workshop Outcome Types.

Registered vs Incomplete Students Over Time

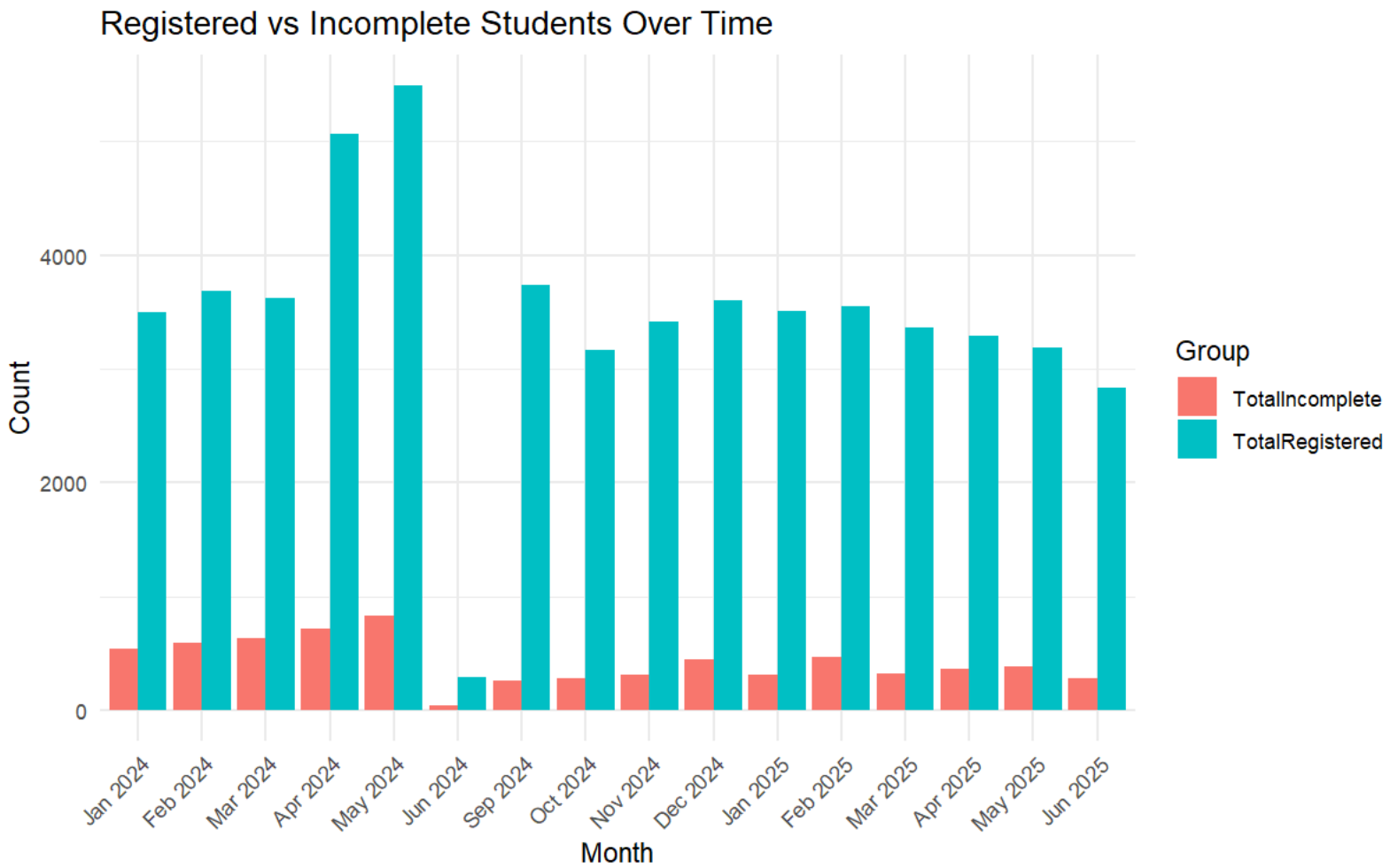

Figure: Figure: Registration vs Incompletion.

Attendance by Gender

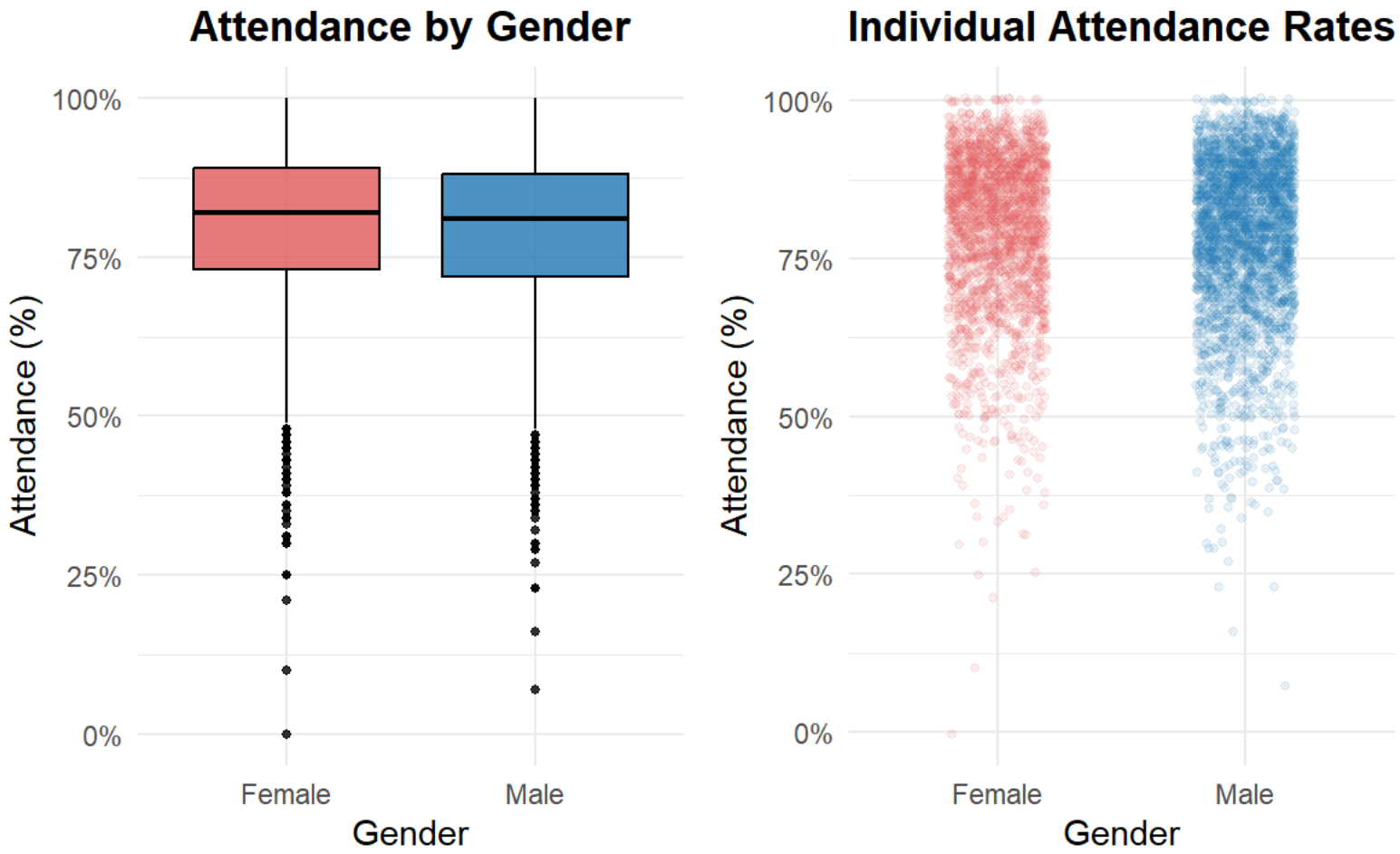

Figure: Figure: Attendance by Gender.

Failure Rates by Age & Gender

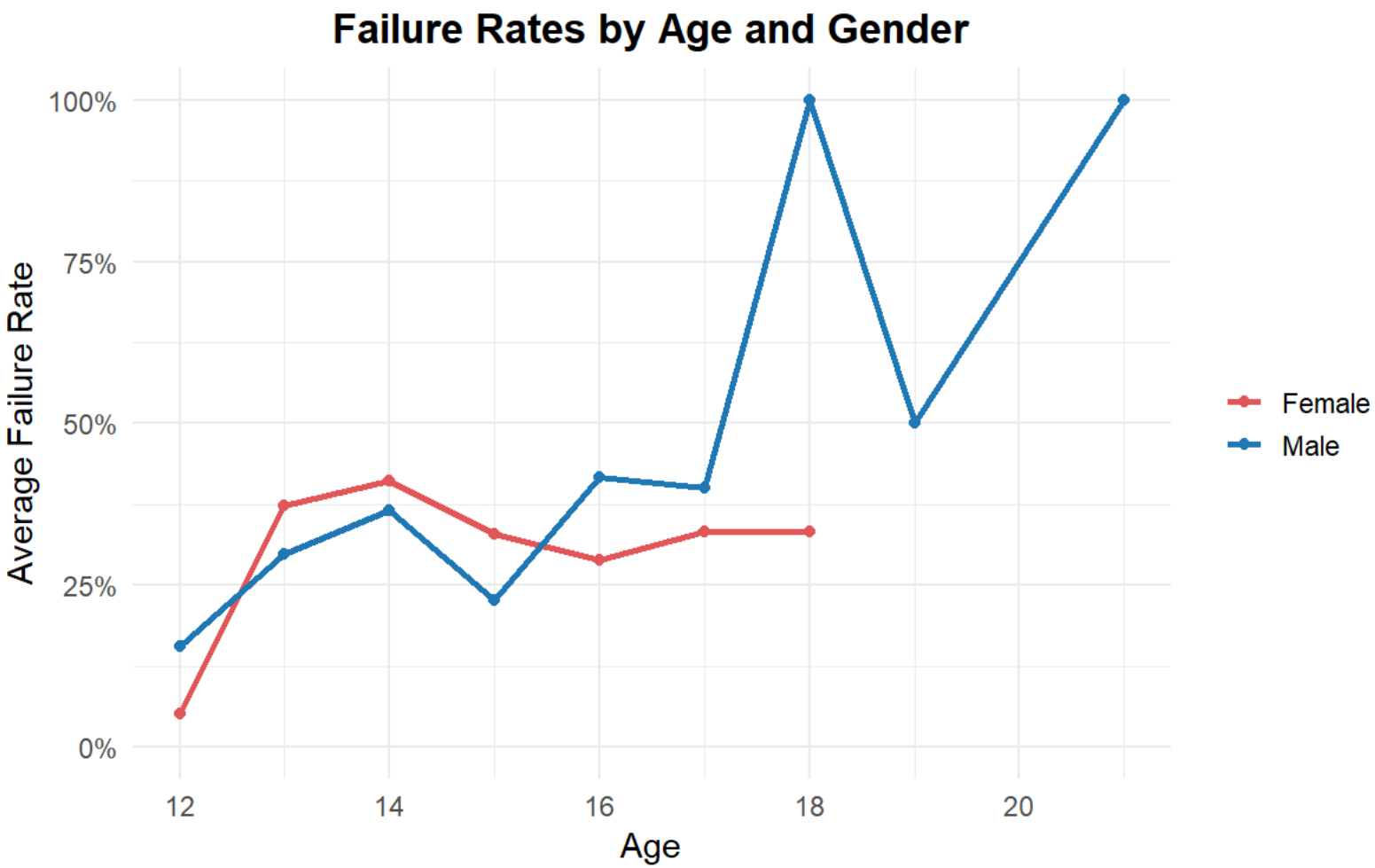

Figure: Figure: Failure Rates by Age and Gender.

Monotonicity: Presence Rate vs Task Rating

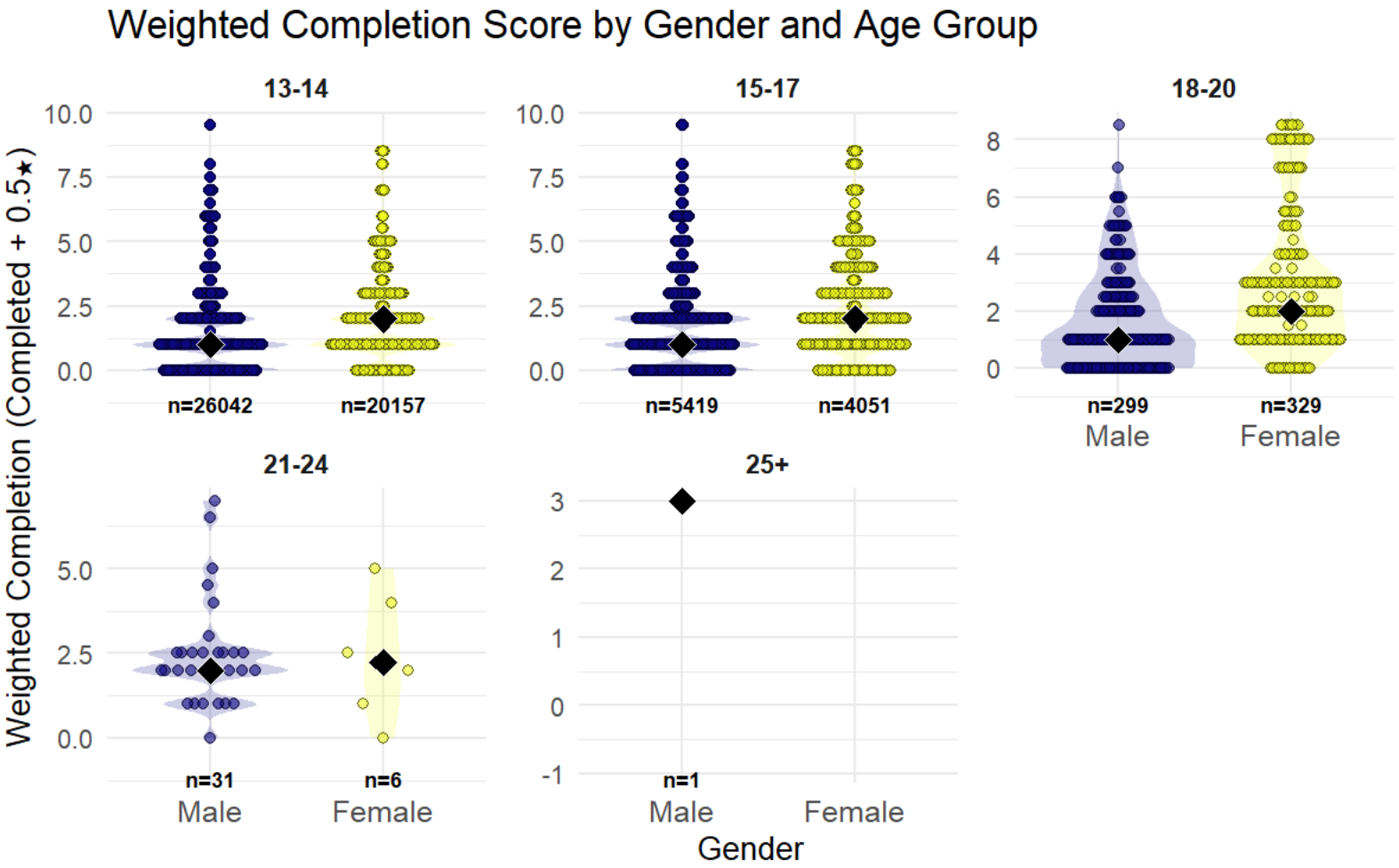

Figure: Figure: Weighted Completion Score by Gender and Age Group.

Flow of Students

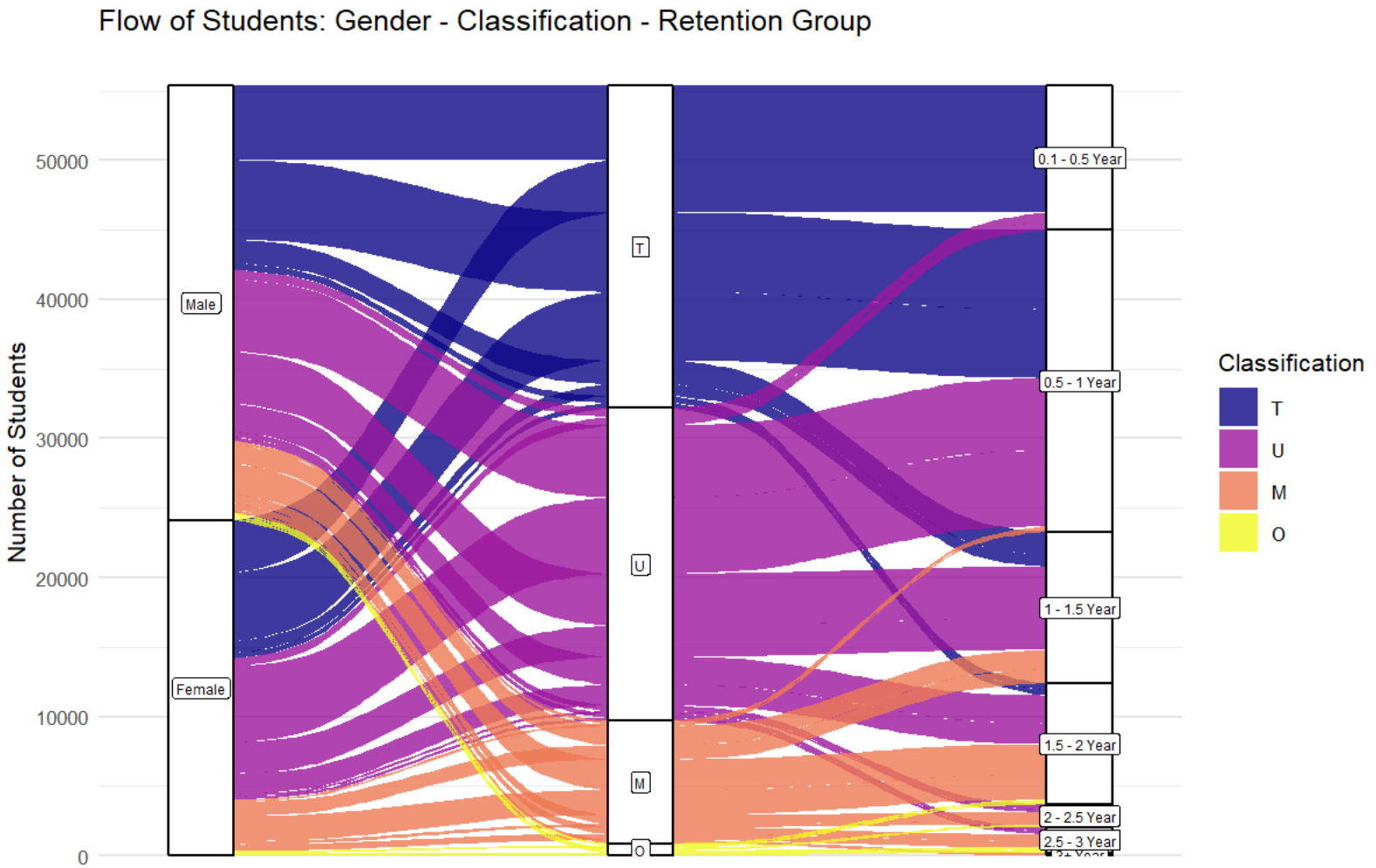

Figure: Figure: Alluvial Plot.

Scatterplot: Presence Rate & Task Rating Trend

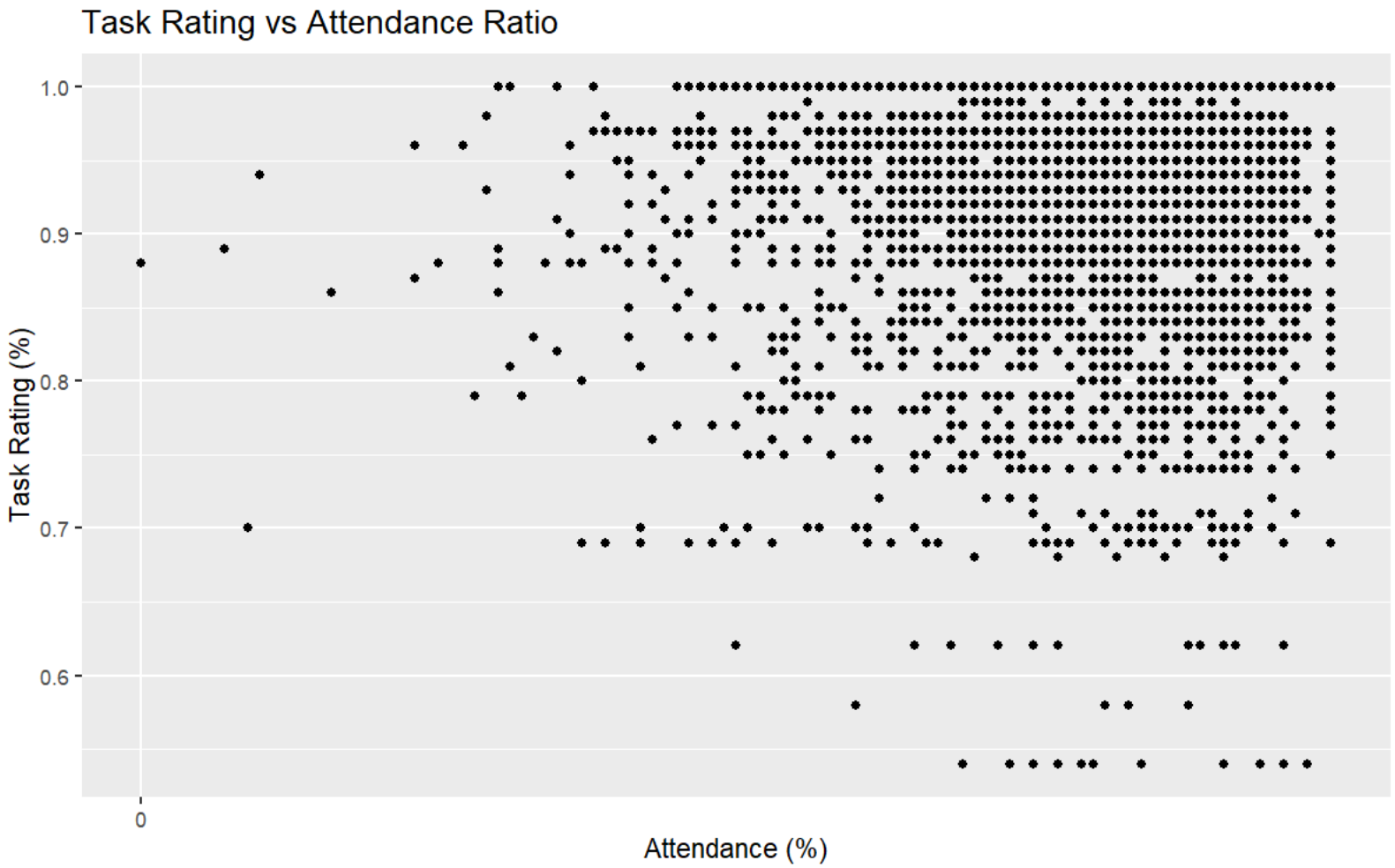

Figure: Figure: Presence Rate & Task Rating Trend.

Linearity and Monotonicity

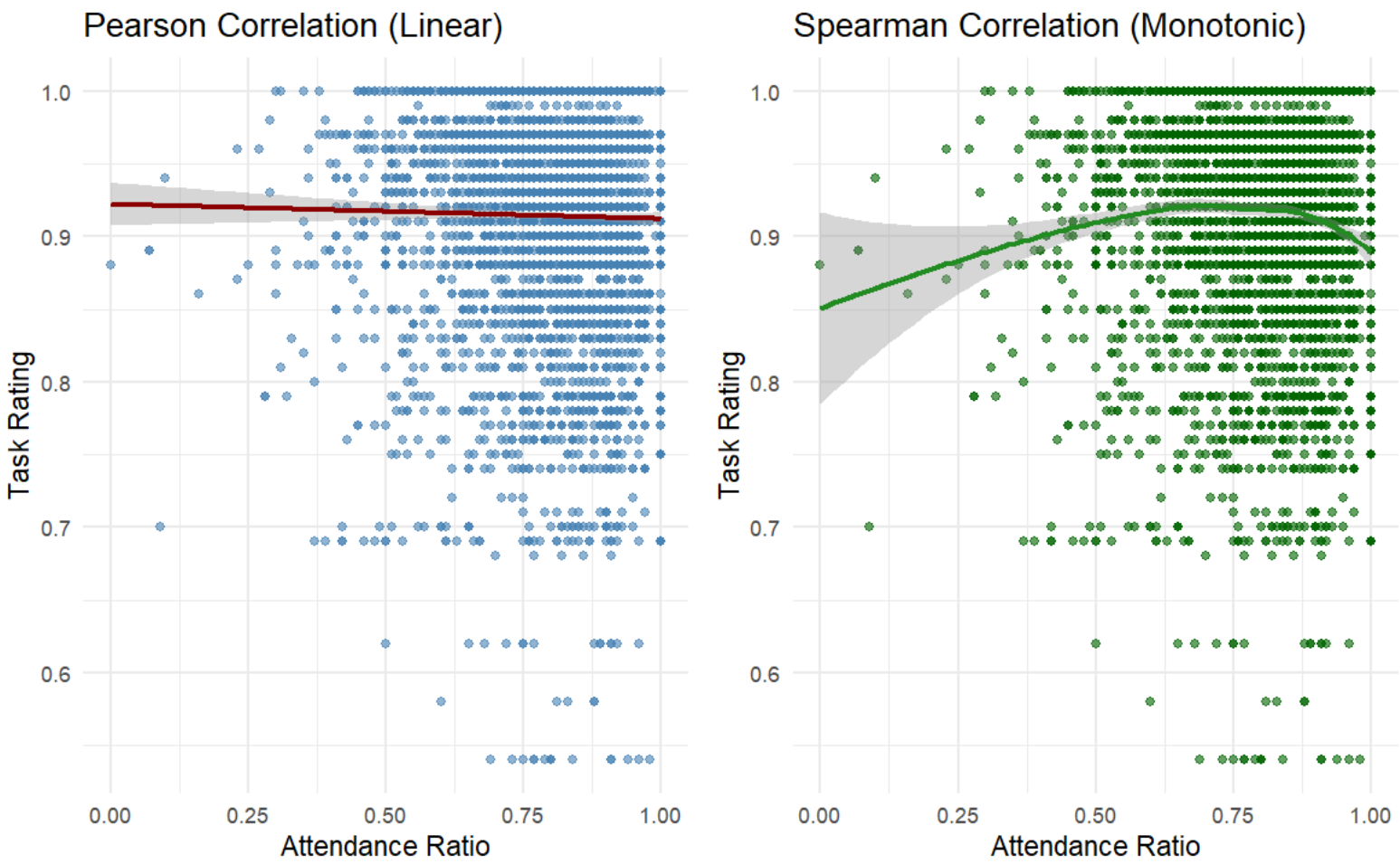

Figure: Figure: Attendance by Gender.

PCA Methodology

- Using `prcomp()` with centering and scaling
- Purpose: dimensionality reduction, pattern discovery

PCA Plot: First Two Principal Components

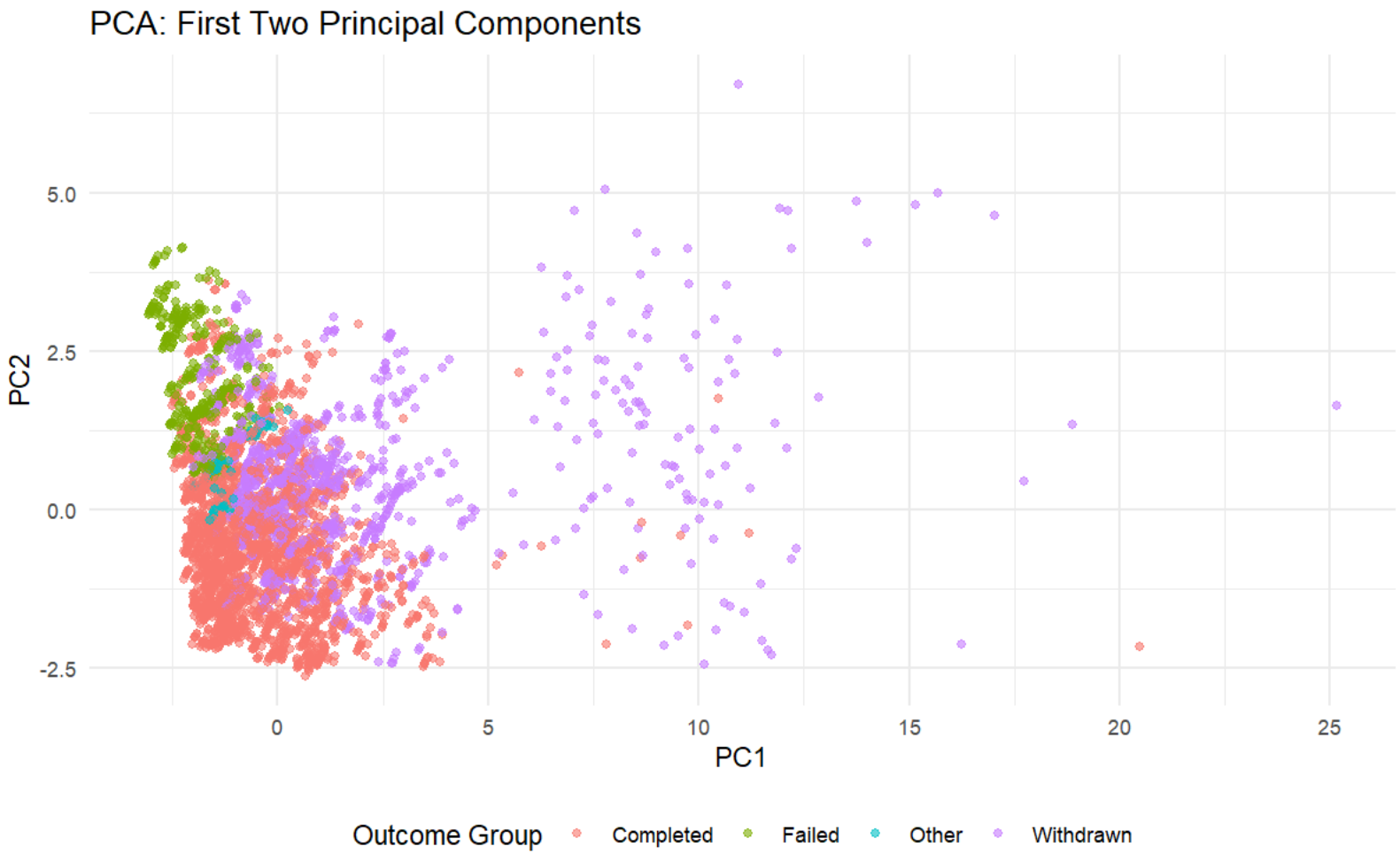

Figure: Figure: PCA .

Conclusions

- THANK YOU!