

Student Engagement and Performance Segmentation

```
student_performance <- read.csv("../Data/TUMO Yerevan_Students Performance_Table - Sheet1.csv")
```

```
str(student_performance)
```

```
## 'data.frame': 10139 obs. of 14 variables:
## $ TumoID : num 2.41e+11 2.41e+11 2.31e+11 2.30e+11 2.50e+11 ...
## $ Age : int 14 13 14 14 13 13 13 15 20 14 ...
## $ Classification : chr "T" "T" "M" "M" ...
## $ Schedule : chr "Sunday 13:30" "Monday 17:30" "Sunday 13:30" "Tuesday 19:30" ...
## $ Status : chr "Active" "Preclosed" "Active" "Active" ...
## $ RetentionGrouped : chr "0.5 - 1 Year" "0.1 - 0.5 Year" "1.5 - 2 Year" "2 - 2.5 Year" ...
## $ Awarded : int 38 8 36 35 12 13 19 20 22 52 ...
## $ Rejected : int 6 0 4 2 4 2 1 3 0 1 ...
## $ Completed : int 1 0 5 4 0 0 1 2 0 6 ...
## $ Incomplete : int 0 0 0 2 0 0 0 0 0 1 ...
## $ Participated : int 0 0 0 0 0 1 2 0 0 0 ...
## $ Withdrawn : int 0 0 0 2 0 0 0 0 1 1 ...
## $ LearningLabs.Completed: int 0 0 1 2 0 0 0 0 0 0 ...
## $ AttendingSince : chr "10/10/2024" "2/10/2025" "12/7/2023" "6/6/2023" ...
```

```
student_info <- read.csv("../Data/TUMO Yerevan Center Report_Students List_Table - Sheet1.csv", colClasses = "character")
str(student_info) #
```

```
## 'data.frame': 10428 obs. of 11 variables:
## $ TumoID : chr "2.30326E+11" "2.30113E+11" "2.40401E+11" "2.50414E+11" ...
## $ BirthDate : chr "11-Jul-10" "9-Nov-10" "20-Mar-12" "29-Jan-13" ...
## $ Classification : chr "M" "U" "T" "T" ...
## $ Status : chr "Active" "Active" "Active" "Active" ...
## $ StudentSchedule : chr "Sunday 13:30" "Friday 17:30" "Wednesday 15:30" "Wednesday 19:30" ...
## $ AttendingSince : chr "14-Sep-23" "12-Apr-23" "5-Jun-24" "7-May-25" ...
## $ RetentionByMonths: int 22 27 13 2 27 21 18 6 6 8 ...
## $ RetentionGrouped : chr "1.5 - 2 Year" "2 - 2.5 Year" "1 - 1.5 Year" "0.1 - 0.5 Year" ...
## $ Age : int 15 14 13 12 14 15 14 14 12 12 ...
## $ Present : chr "109" "120" "64" "13" ...
## $ PresenceRatio : chr "85" "90" "86" "81" ...
```

```
student_info$TumoID <- as.numeric(student_info$TumoID)
options(scipen = 999)
```

```
student_performance$task_rating <- round(student_performance$Awarded /
  (student_performance$Awarded + student_performance$Rejected), 2)

student_performance$training_rating <- round(student_performance$Completed /
  (student_performance$Incomplete + student_performance$Participated +
    student_performance$Withdrawn + student_performance$Completed), 2)
```

```
str(student_performance)
```

```
## 'data.frame':    10139 obs. of  16 variables:
## $ TumoID          : num  240712000018 240924000012 230619000016 230121000032 250415000009 ...
## $ Age             : int   14 13 14 14 13 13 13 15 20 14 ...
## $ Classification   : chr   "T" "T" "M" "M" ...
## $ Schedule         : chr   "Sunday 13:30" "Monday 17:30" "Sunday 13:30" "Tuesday 19:30" ...
## $ Status           : chr   "Active" "Preclosed" "Active" "Active" ...
## $ RetentionGrouped : chr   "0.5 - 1 Year" "0.1 - 0.5 Year" "1.5 - 2 Year" "2 - 2.5 Year" ...
## $ Awarded          : int   38 8 36 35 12 13 19 20 22 52 ...
## $ Rejected         : int    6 0 4 2 4 2 1 3 0 1 ...
## $ Completed        : int    1 0 5 4 0 0 1 2 0 6 ...
## $ Incomplete       : int    0 0 0 2 0 0 0 0 0 1 ...
## $ Participated     : int    0 0 0 0 0 1 2 0 0 0 ...
## $ Withdrawn        : int    0 0 0 2 0 0 0 0 1 1 ...
## $ LearningLabs.Completed: int  0 0 1 2 0 0 0 0 0 0 ...
## $ AttendingSince   : chr   "10/10/2024" "2/10/2025" "12/7/2023" "6/6/2023" ...
## $ task_rating      : num   0.86 1 0.9 0.95 0.75 0.87 0.95 0.87 1 0.98 ...
## $ training_rating  : num    1 NaN 1 0.5 NaN 0 0.33 1 0 0.75 ...
```

```
table(student_performance$task_rating)
```

```
##
## 0.36 0.5 0.52 0.53 0.54 0.57 0.58 0.59 0.6 0.61 0.62 0.63 0.64 0.65 0.67 0.68
## 1 1 1 1 2 4 4 2 4 1 6 4 6 17 14 14
## 0.69 0.7 0.71 0.72 0.73 0.74 0.75 0.76 0.77 0.78 0.79 0.8 0.81 0.82 0.83 0.84
## 23 21 30 23 28 38 53 54 57 77 96 89 108 128 144 113
## 0.85 0.86 0.87 0.88 0.89 0.9 0.91 0.92 0.93 0.94 0.95 0.96 0.97 0.98 0.99 1
## 182 236 167 352 331 392 384 472 534 631 727 776 793 682 150 2166
```

```
table(student_performance$training_rating)
```

```
##
## 0 0.08 0.09 0.1 0.11 0.12 0.13 0.14 0.17 0.18 0.2 0.21 0.22 0.23 0.25 0.27
## 946 6 5 4 8 17 1 27 38 9 75 3 17 6 190 7
## 0.28 0.29 0.3 0.31 0.32 0.33 0.35 0.36 0.37 0.38 0.39 0.4 0.41 0.42 0.43 0.44
## 1 61 26 4 2 365 2 20 1 54 2 181 2 11 68 42
## 0.45 0.46 0.47 0.48 0.5 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59 0.6 0.62 0.64
## 20 9 7 1 1012 3 6 11 27 52 114 25 2 252 117 32
## 0.65 0.67 0.68 0.69 0.7 0.71 0.72 0.73 0.74 0.75 0.76 0.77 0.78 0.79 0.8 0.81
## 3 754 2 14 55 156 2 35 1 483 6 15 49 13 320 4
## 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89 0.9 0.91 0.92 0.93 0.94 0.95 0.96 1
## 23 233 1 9 143 6 108 56 38 17 13 8 4 1 1 2520
```

Hypothesis 1: Attendance correlates with student performance

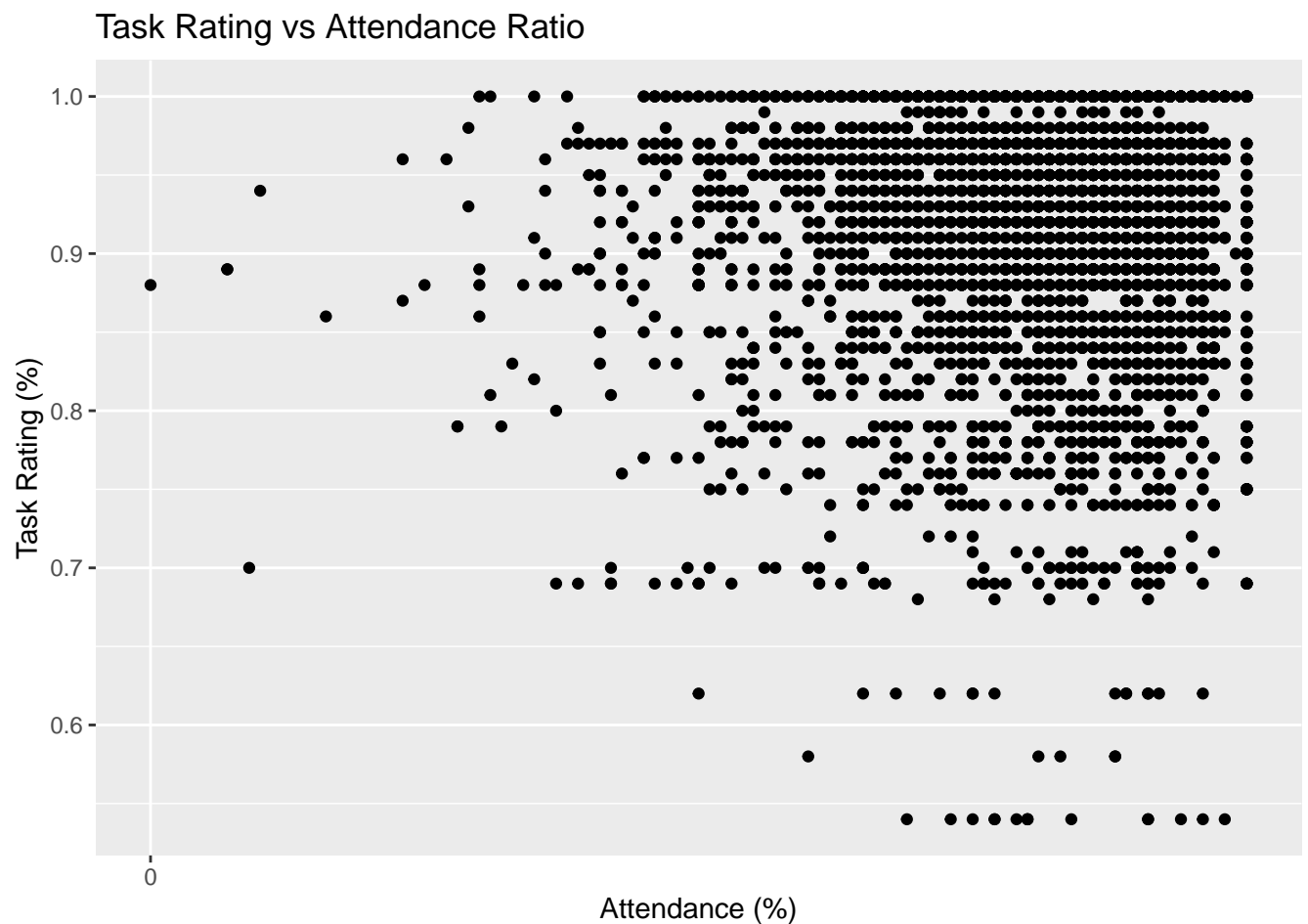
Assumption: Students with higher attendance are more likely to perform better (complete courses) compared to students with low attendance.

Rationale: Students who are actively attending classes may engage more with the material, leading to better performance.

```
merged_df <- inner_join(student_info, student_performance, by = "TumoID")
```

```
merged_df$PresenceRatio <- round(as.integer(merged_df$PresenceRatio) / 100, 2)
```

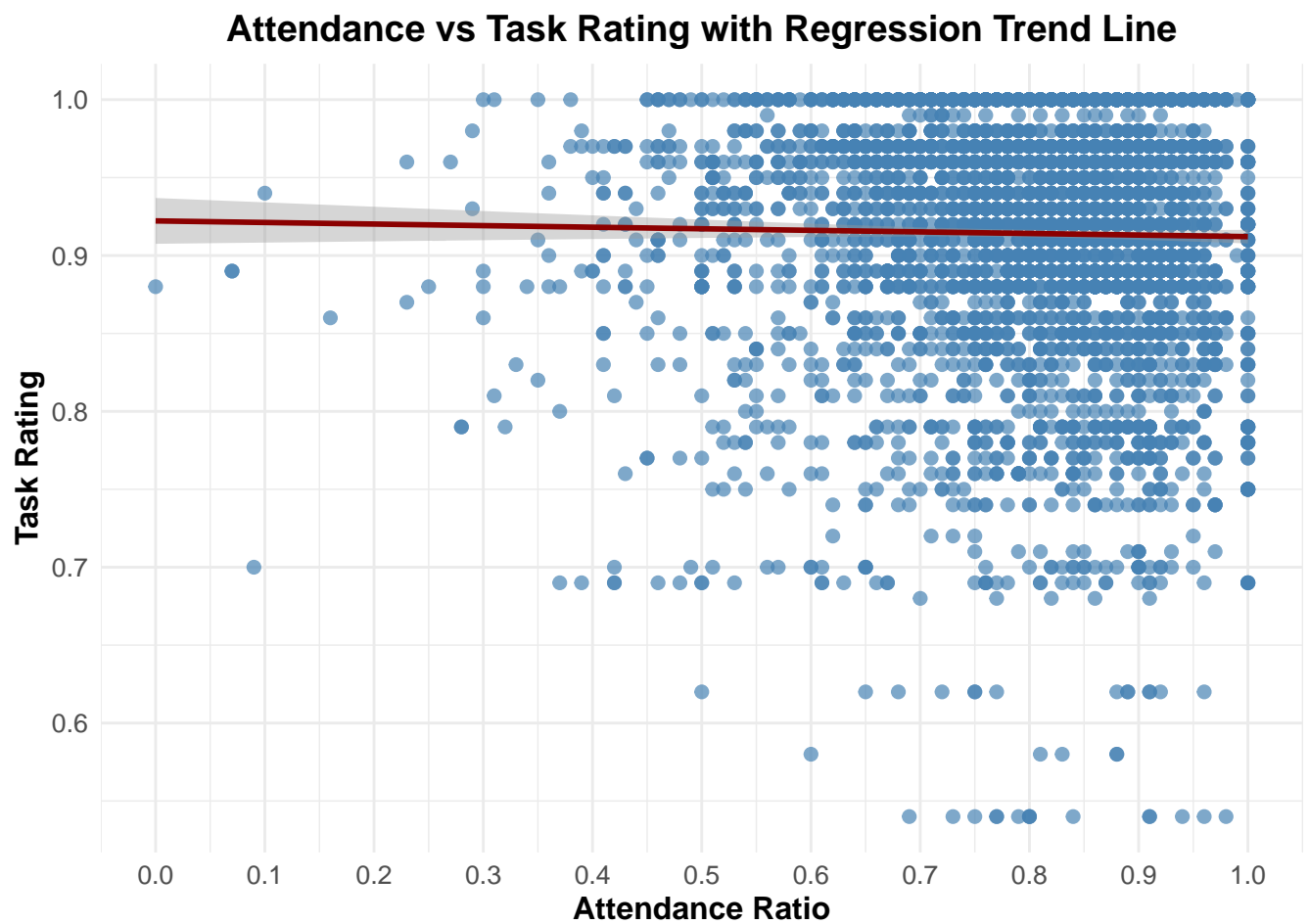
```
ggplot(merged_df, aes(x = PresenceRatio, y = task_rating)) +  
  geom_point() +  
  labs(  
    title = "Task Rating vs Attendance Ratio",  
    x = "Attendance (%)",  
    y = "Task Rating (%)"  
  ) +  
  scale_x_continuous(breaks = seq(0, 100, 10))
```



Scatterplot: Presence Rate & Task Rating Trend

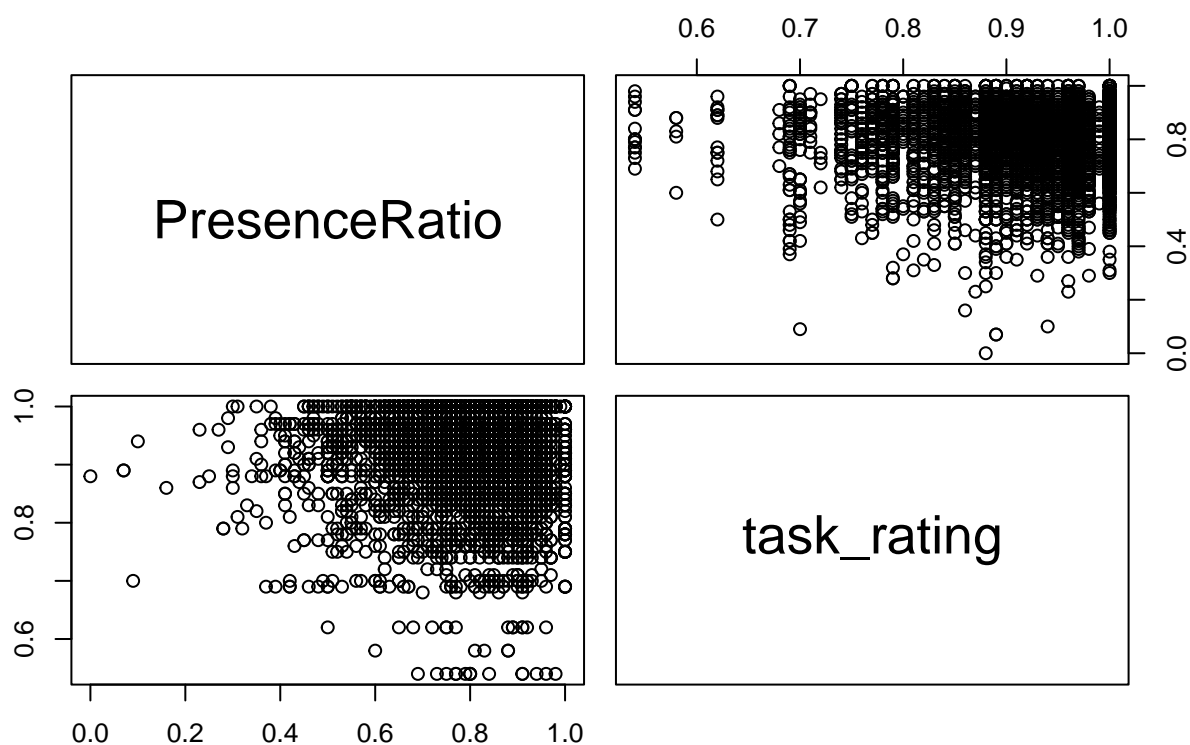
```
ggplot(merged_df, aes(x = PresenceRatio, y = task_rating)) +  
  geom_point(color = "steelblue", size = 2, alpha = 0.7) + # points  
  geom_smooth(method = "lm", se = TRUE, color = "darkred", linewidth = 1) + # trend line with CI
```

```
labs(
  title = "Attendance vs Task Rating with Regression Trend Line",
  x = "Attendance Ratio",
  y = "Task Rating"
) +
scale_x_continuous(breaks = seq(0, 1, 0.1)) + # ticks every 0.1
theme_minimal() +
theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
  axis.title = element_text(face = "bold", size = 12),
  axis.text = element_text(size = 10)
)
```



```
pairs(merged_df[, c("PresenceRatio", "task_rating")],
  main = "Scatterplot Matrix")
```

Scatterplot Matrix



Linear Relationship: Presence Rate VS Task Rating

```
spearman_cor_test_result <- cor.test(merged_df$PresenceRatio, merged_df$task_rating,
  method = "spearman", exact = FALSE, use = "complete.obs")
spearman_cor_test_result
```

```
##
## Spearman's rank correlation rho
##
## data: merged_df$PresenceRatio and merged_df$task_rating
## 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_cor_test_result <- cor.test(merged_df$PresenceRatio, merged_df$task_rating,
  method = "pearson", exact = FALSE, use = "complete.obs")
pearson_cor_test_result
```

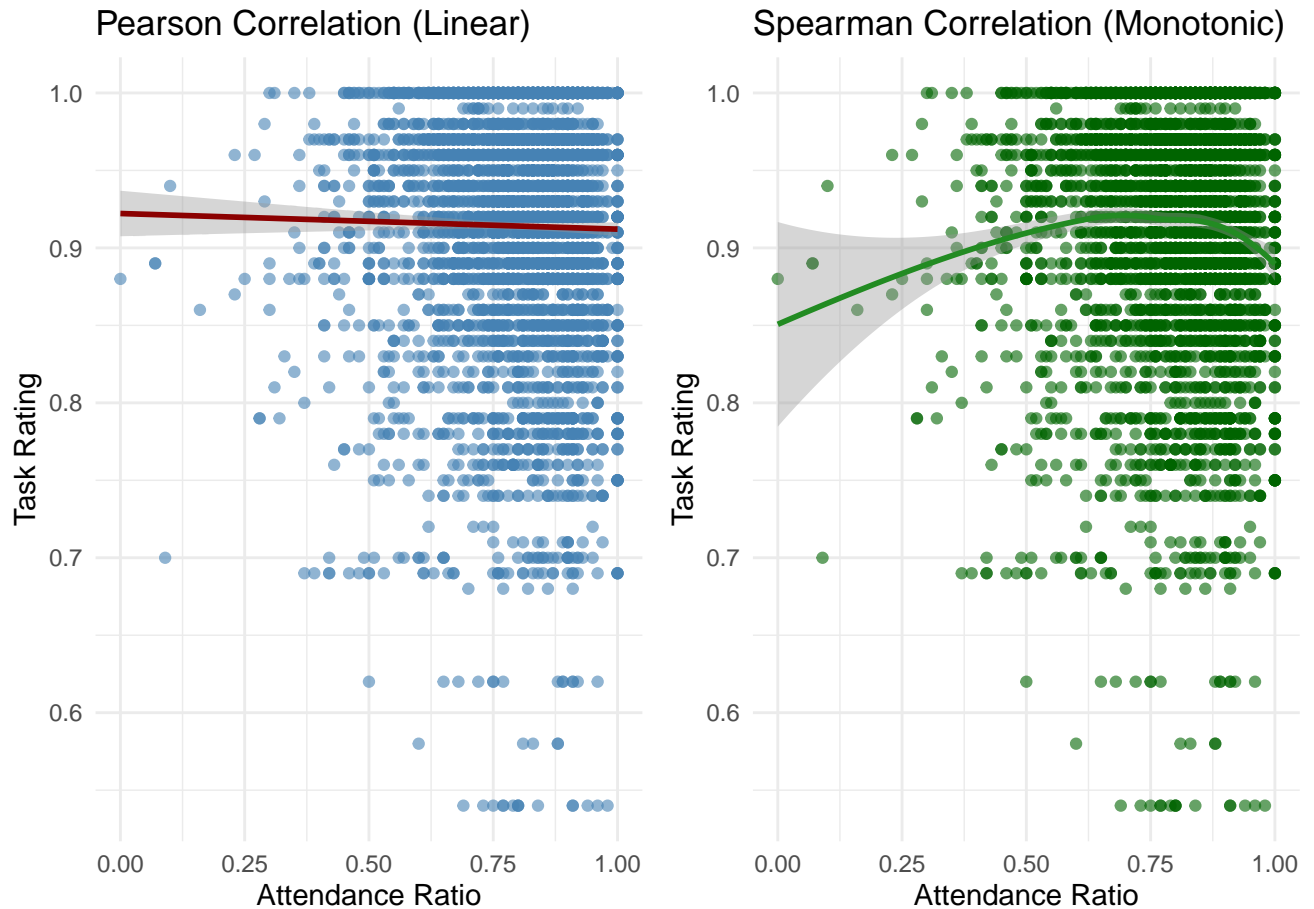
```
##
## Pearson's product-moment correlation
##
## data: merged_df$PresenceRatio and merged_df$task_rating
## t = -1.0881, df = 4216, p-value = 0.2766
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.04691147 0.01343134
## sample estimates:
## cor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

Linearity and Monotonicity Plots

```
p1 <- ggplot(merged_df, aes(x = PresenceRatio, y = task_rating)) +
  geom_point(alpha = 0.6, color = "steelblue") +
  geom_smooth(method = "lm", se = TRUE, color = "darkred") +
  labs(title = "Pearson Correlation (Linear)",
    x = "Attendance Ratio",
    y = "Task Rating") +
  theme_minimal()

# Spearman scatterplot with loess smooth (captures monotonic relationship)
p2 <- ggplot(merged_df, aes(x = PresenceRatio, y = task_rating)) +
  geom_point(alpha = 0.6, color = "darkgreen") +
  geom_smooth(method = "loess", se = TRUE, color = "forestgreen") +
  labs(title = "Spearman Correlation (Monotonic)",
    x = "Attendance Ratio",
    y = "Task Rating") +
  theme_minimal()

p1 + p2
```



Hypothesis 2: Number of courses started impacts performance outcomes

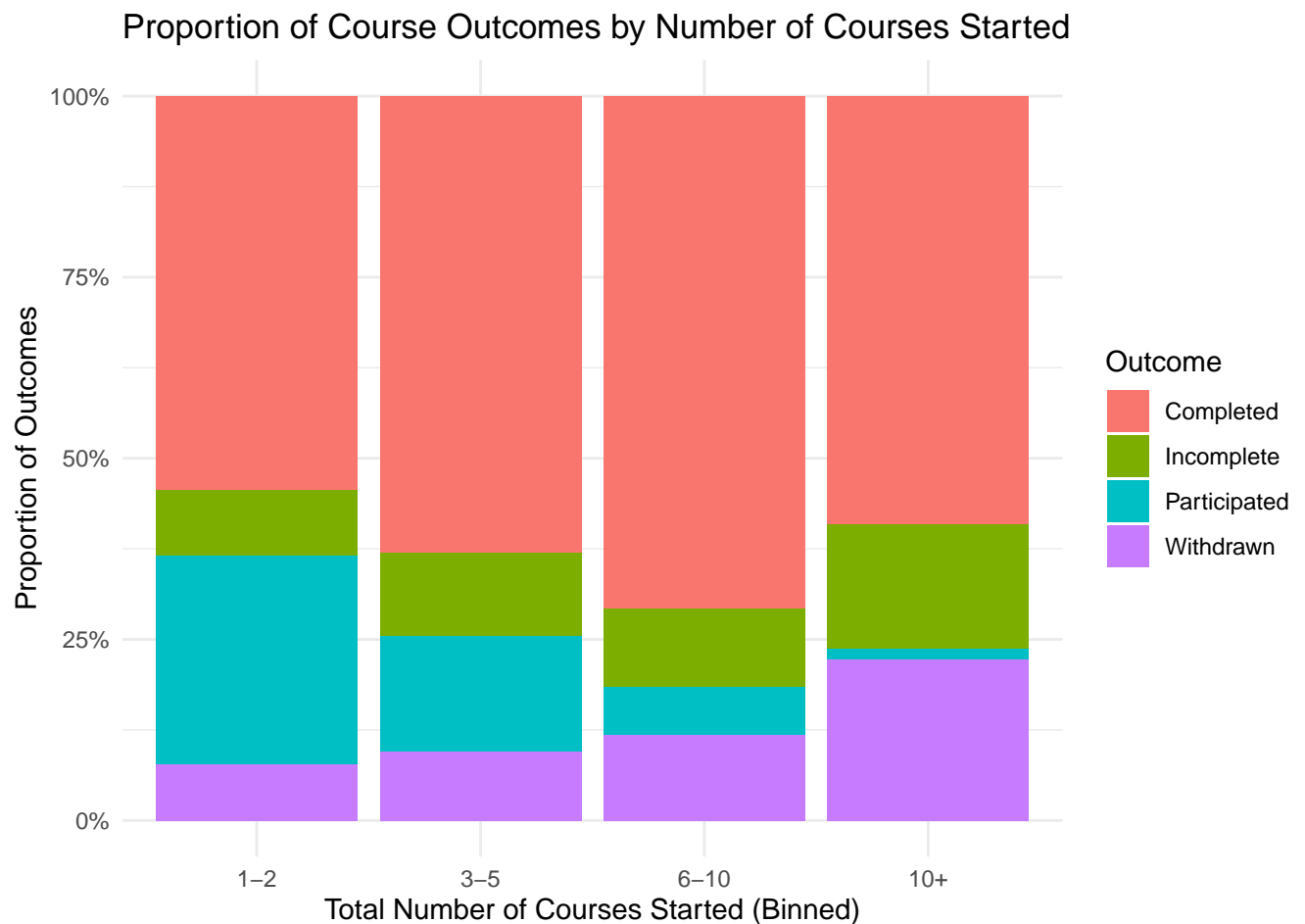
Assumption: Students who start more courses may either demonstrate strong engagement or be overwhelmed, which could lead to different performance outcomes (completed, failed, or withdrawn).

Rationale: Engaging with multiple courses may indicate either strong motivation or poor focus/time management skills, both of which could affect performance outcomes.

```
course_outcomes <- merged_df %>%
  mutate(total_courses = Completed + Incomplete + Participated + Withdrawn) %>%
  pivot_longer(
    cols = c(Completed, Incomplete, Participated, Withdrawn),
    names_to = "Outcome",
    values_to = "Count"
  )
```

Proportion of Course Outcomes by Number of Courses Started

```
course_outcomes <- course_outcomes %>%  
  mutate(total_bin = cut(total_courses, breaks = c(0, 2, 5, 10, Inf),  
    labels = c("1-2", "3-5", "6-10", "10+"), right = FALSE))  
  
ggplot(course_outcomes, aes(x = total_bin, y = Count, fill = Outcome)) +  
  geom_bar(stat = "identity", position = "fill") +  
  scale_y_continuous(labels = percent_format()) +  
  labs(  
    title = "Proportion of Course Outcomes by Number of Courses Started",  
    x = "Total Number of Courses Started (Binned)",  
    y = "Proportion of Outcomes",  
    fill = "Outcome"  
  ) +  
  theme_minimal()
```



```
df <- merged_df %>%  
  mutate(  
    total_courses = Completed + Participated + Rejected + Withdrawn,
```



```

course_volume_group = cut(
  total_courses,
  breaks = c(-Inf, 2, 5, 10, Inf),
  labels = c("1-2", "3-5", "6-10", "10+")
)
)

outcome_table <- df %>%
  group_by(course_volume_group) %>%
  summarise(
    Completed = sum(Completed, na.rm = TRUE),
    Participated = sum(Participated, na.rm = TRUE),
    Failed = sum(Rejected, na.rm = TRUE),
    Withdrawn = sum(Withdrawn, na.rm = TRUE)
  )

```

```

outcome_matrix <- as.matrix(outcome_table[, -1])
rownames(outcome_matrix) <- outcome_table$course_volume_group
chisq_result <- chisq.test(outcome_matrix)

```

```
outcome_matrix
```

```

##      Completed Participated Failed Withdrawn
## 1-2         428          128    644         56
## 3-5        2355          557   2685        406
## 6-10        4782          746   5227        865
## 10+         2266           38   1568        631

```

Variable Independence Test

```
chisq_result
```

```

##
## Pearson's Chi-squared test
##
## data:  outcome_matrix
## X-squared = 762.73, df = 9, p-value < 0.00000000000000022

```

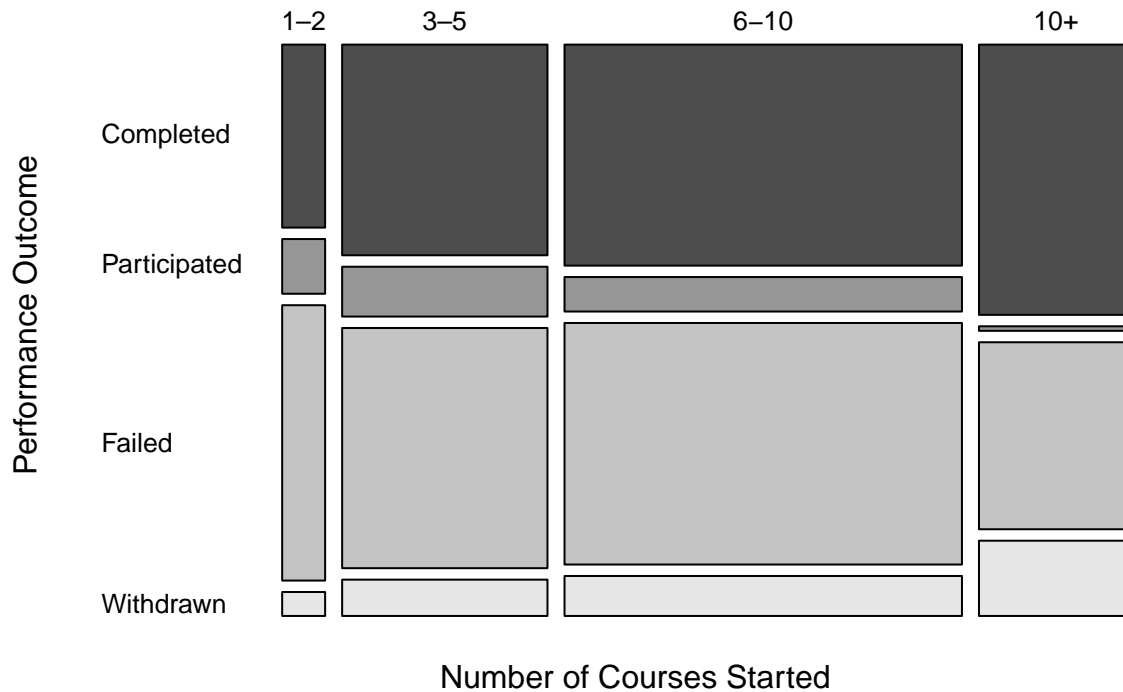
Course Volume VS Performance Outcome

```

mosaicplot(outcome_matrix,
  main = "Mosaic Plot: Course Volume vs Performance Outcome",
  color = TRUE,
  xlab = "Number of Courses Started",
  ylab = "Performance Outcome",
  las = 1,
  cex.axis = 0.8)

```

Mosaic Plot: Course Volume vs Performance Outcome



Hypothesis 3: 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.

```
merged_df$main_outcome_group <- case_when(
  merged_df$Withdrawn > 0 ~ "Withdrawn",
  merged_df$Completed > 0 ~ "Completed",
  merged_df$Rejected > 0 ~ "Failed",
  TRUE ~ "Other"
)
```

T-Test: Withdrawn vs Completed

```
t.test(PresenceRatio ~ main_outcome_group,
  data = merged_df %>% filter(main_outcome_group %in% c("Withdrawn", "Completed")))
```

```
##
## Welch Two Sample t-test
##
## data: PresenceRatio by main_outcome_group
## t = 6.3982, df = 2091.9, p-value = 0.0000000001934
## alternative hypothesis: true difference in means between group Completed and group Withdrawn is not equal to 0
## 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

```
t.test(PresenceRatio ~ main_outcome_group,
       data = merged_df %>% filter(main_outcome_group %in% c("Withdrawn", "Failed")))
```

```
##
## Welch Two Sample t-test
##
## data: PresenceRatio by main_outcome_group
## t = 5.8457, df = 1558.9, p-value = 0.0000000006131
## alternative hypothesis: true difference in means between group Failed and group Withdrawn is not equal to 0
## 95 percent confidence interval:
##  0.02520169 0.05065455
## sample estimates:
##      mean in group Failed mean in group Withdrawn
##           0.8106117           0.7726836
```

Principal Component Analysis (PCA)

Understanding 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.

```
df_numeric <- merged_df %>%
  select(where(is.numeric)) %>%
  na.omit()

groups <- merged_df %>%
  filter(complete.cases(select(., where(is.numeric)))) %>%
  pull(main_outcome_group)
```

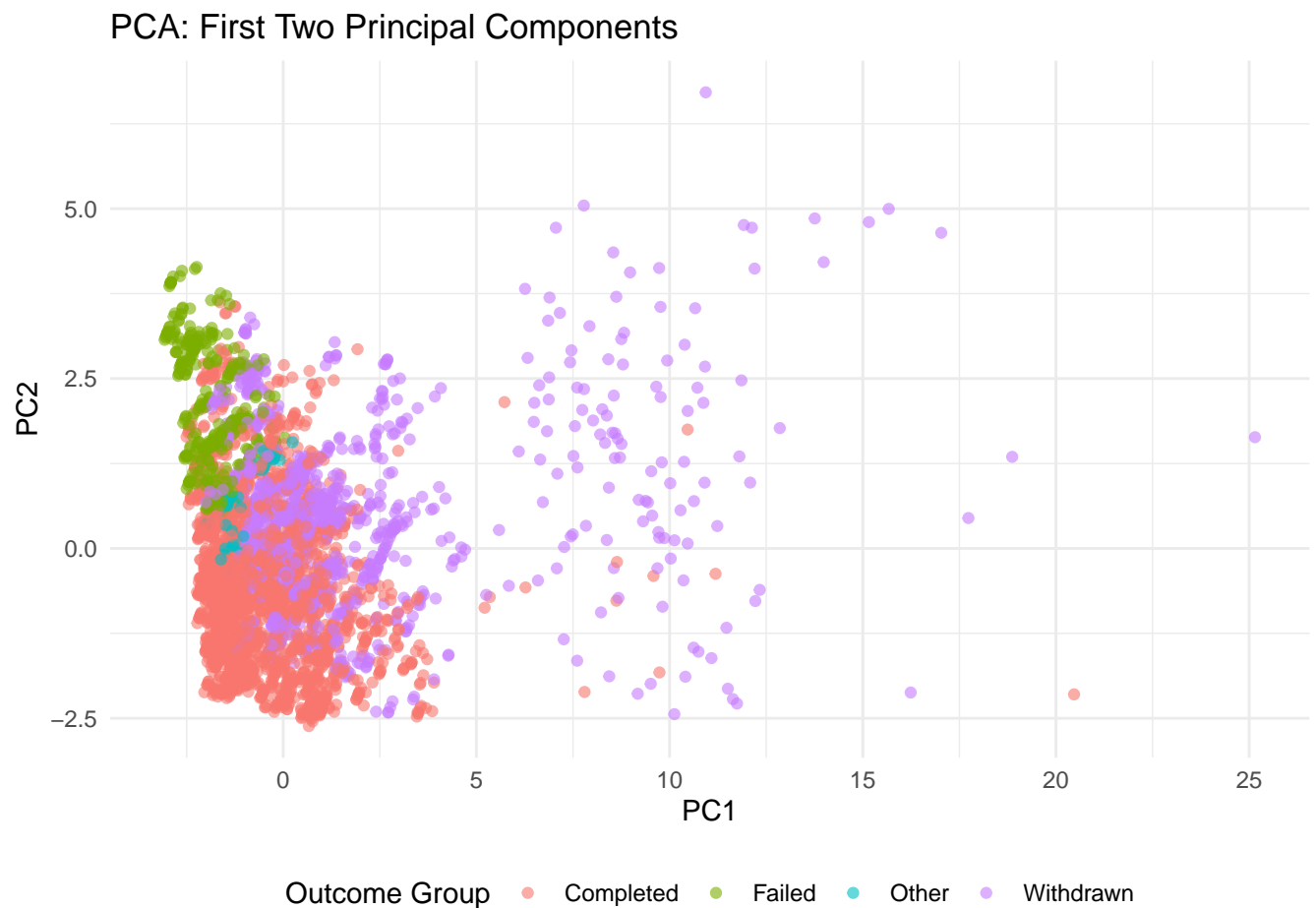
```
pca_result <- prcomp(df_numeric, center = TRUE, scale. = TRUE)

pca_df <- as.data.frame(pca_result$x)
pca_df$Group <- groups
```

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

PCA Plot: First Two Principal Components

```
ggplot(pca_df, aes(x = PC1, y = PC2, color = Group)) +  
  geom_point(alpha = 0.6, size = 1.5) +  
  labs(title = "PCA: First Two Principal Components",  
       color = "Outcome Group") +  
  theme_minimal() +  
  theme(legend.position = "bottom")
```



If Withdrawn points cluster separately from Completed or Failed, this supports the hypothesis: withdrawn students behave differently.