Student Dynamics & Learning Patterns

2025-08-02

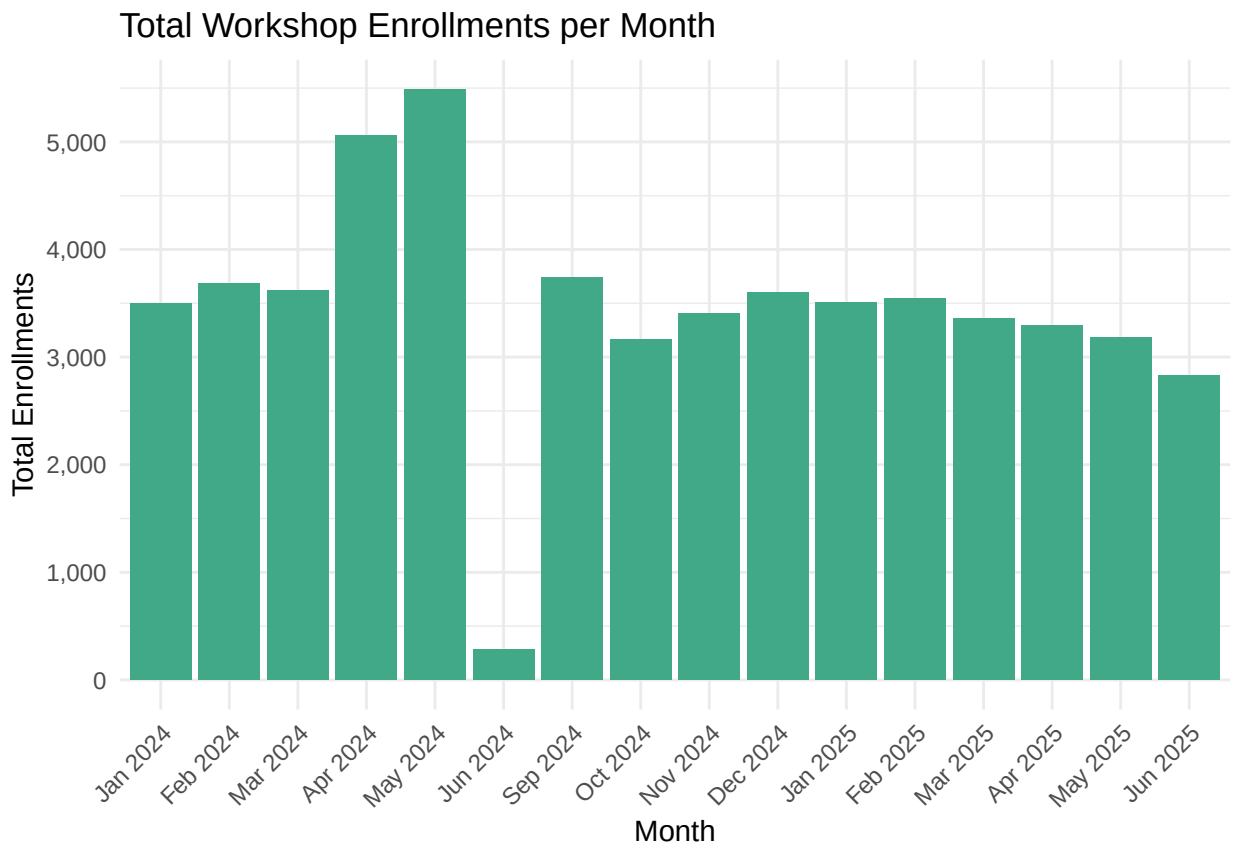
Research Goals:

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).

```
merged_df %>%
  filter(!is.na(StartDate)) %>%
  mutate(
    StartDate = as.Date(StartDate, format = "%d-%b-%y"),
    MonthYear = format(StartDate, "%b %Y"),
    MonthYear = factor(MonthYear, levels = unique(format(sort(StartDate), "%b %Y")))
  ) %>%
  group_by(MonthYear) %>%
  summarise(TotalEnrollments = n(), .groups = "drop") %>%
  ggplot(aes(x = MonthYear, y = TotalEnrollments)) +
  geom_col(fill = "#41a987") +
  scale_y_continuous(
    breaks = seq(0, 6000, by = 1000),
    labels = scales::comma_format()
  ) +
  labs(
    title = "Total Workshop Enrollments per Month",
    x = "Month",
    y = "Total Enrollments"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Chi-square test: Skill vs MonthYear
skill_month_table <- merged_df %>%
  filter(!is.na(StartDate), !is.na(Skill)) %>%
  mutate(
    StartDate = as.Date(StartDate, format = "%d-%b-%y"),
    MonthYear = format(StartDate, "%b %Y")
  ) %>%
  count(Skill, MonthYear) %>%
  pivot_wider(names_from = MonthYear, values_from = n, values_fill = 0) %>%
  column_to_rownames("Skill") %>%
  as.matrix()

chisq_result_H1 <- chisq.test(skill_month_table)
chisq_result_H1
```

```
##
##  Pearson's Chi-squared test
##
## data:  skill_month_table
## X-squared = 912.4, df = 180, p-value < 2.2e-16
```

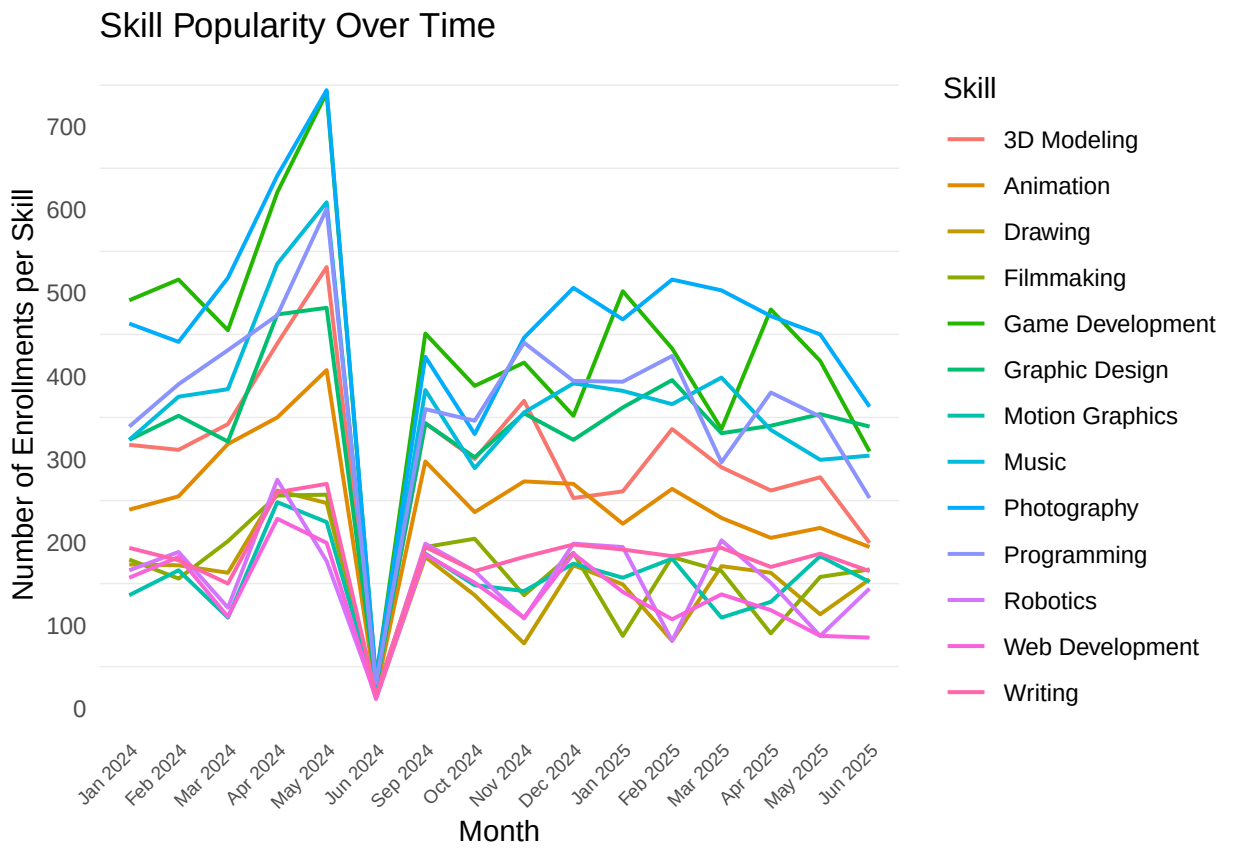
```r
if (chisq_result_H1$p.value < 0.05) {
  cat(" H  is supported: Skill popularity changes significantly over time.\n")
} else {
  cat(" Fail to reject H : No significant difference in skill enrollments over time.\n")
```

```
}
```

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

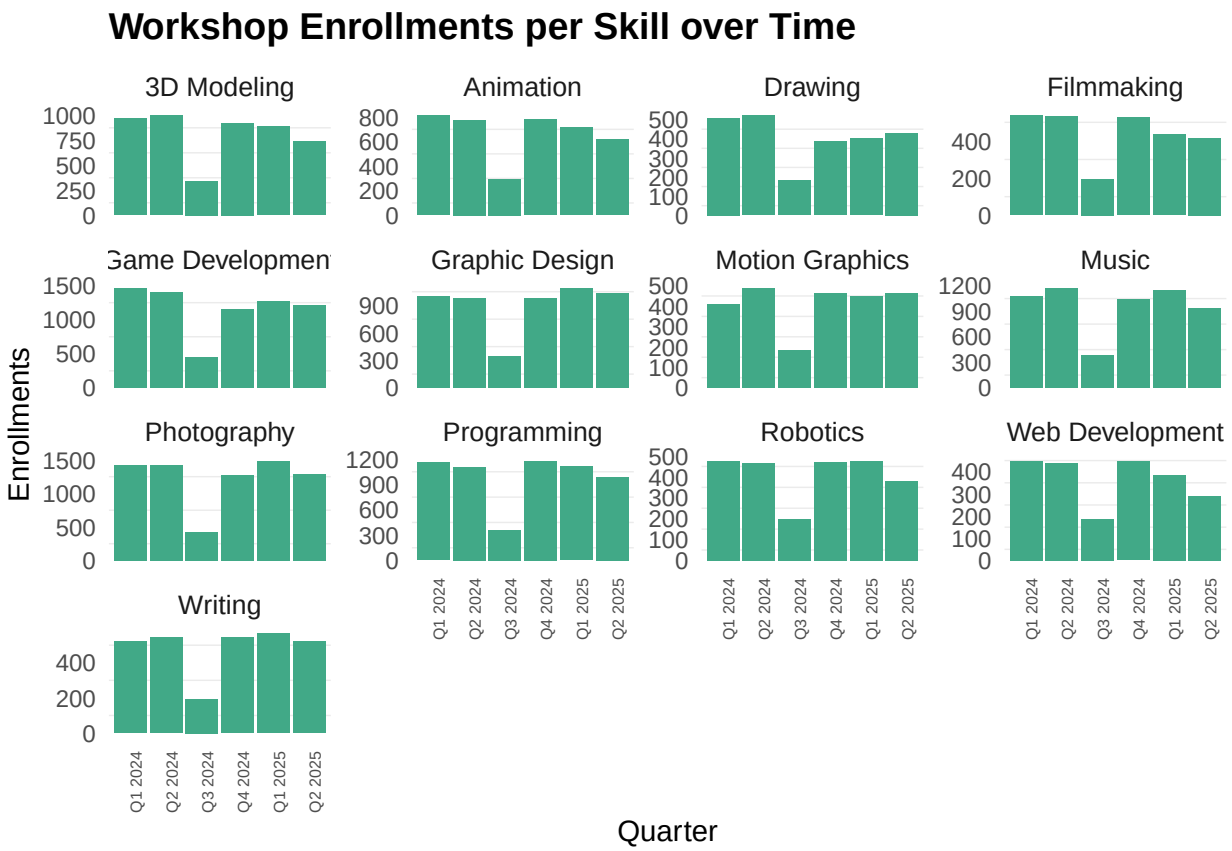
Skill popularity over Time

```
merged_df %>%
  filter(!is.na(StartDate), !is.na(Skill)) %>%
  mutate(
    StartDate = as.Date(StartDate, format = "%d-%b-%y"),
    MonthYear = format(StartDate, "%b %Y"),
    MonthYear = factor(MonthYear, levels = unique(format(sort(StartDate), "%b %Y")))
  ) %>%
  group_by(MonthYear, Skill) %>%
  summarise(Enrollments = n(), .groups = "drop") %>%
  ggplot(aes(x = MonthYear, y = Enrollments, color = Skill, group = Skill)) +
  geom_line(linewidth = 0.7) +
  scale_y_continuous(
    breaks = seq(0, 900, by = 100),
    labels = scales::comma_format()
  ) +
  labs(
    title = "Skill Popularity Over Time",
    x = "Month",
    y = "Number of Enrollments per Skill"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 7),
    legend.position = "right",
    panel.grid.major = element_blank()
  )
```



Workshop Enrollments per Skill over Time

```
merged_df %>%
  filter(!is.na(StartDate), !is.na(Skill)) %>%
  mutate(
    StartDate = as.Date(StartDate, format = "%d-%b-%y"),
    QuarterYear = paste0("Q", quarter(StartDate), " ", year(StartDate)),
    QuarterYear = factor(QuarterYear, levels = unique(paste0("Q", quarter(sort(StartDate)), " ", year(s
  ) %>%
  group_by(QuarterYear, Skill) %>%
  summarise(TotalEnrollments = n(), .groups = "drop") %>%
  ggplot(aes(x = QuarterYear, y = TotalEnrollments)) +
  geom_bar(stat = "identity", fill = "#41a987") +
  facet_wrap(~ Skill, scales = "free_y") +
  scale_x_discrete(drop = FALSE) +  # Keep x breaks even if not all quarters are in every skill
  labs(
    title = "Workshop Enrollments per Skill over Time",
    x = "Quarter",
    y = "Enrollments"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1, size = 6),
    strip.text = element_text(size = 10),
    plot.title = element_text(size = 14, face = "bold"),
    panel.grid.major = element_blank()
  )
```



H2. Withdrawal Trends Over Time Hypothesis: Withdrawals are more frequent in certain months (e.g. May, September).

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.

```
withdrawal_test <- merged_df %>%
  filter(PassStatus == "Withdrawn", !is.na(WithdrawDate)) %>%
  mutate(
    WithdrawDate = parse_date_time(WithdrawDate, orders = c("d-b-y HMS", "B d, Y, I:M:S p")),
    Month = month(WithdrawDate, label = TRUE)
  ) %>%
  count(Month)

chisq_result_H2 <- chisq.test(withdrawal_test$n)
chisq_result_H2
```

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

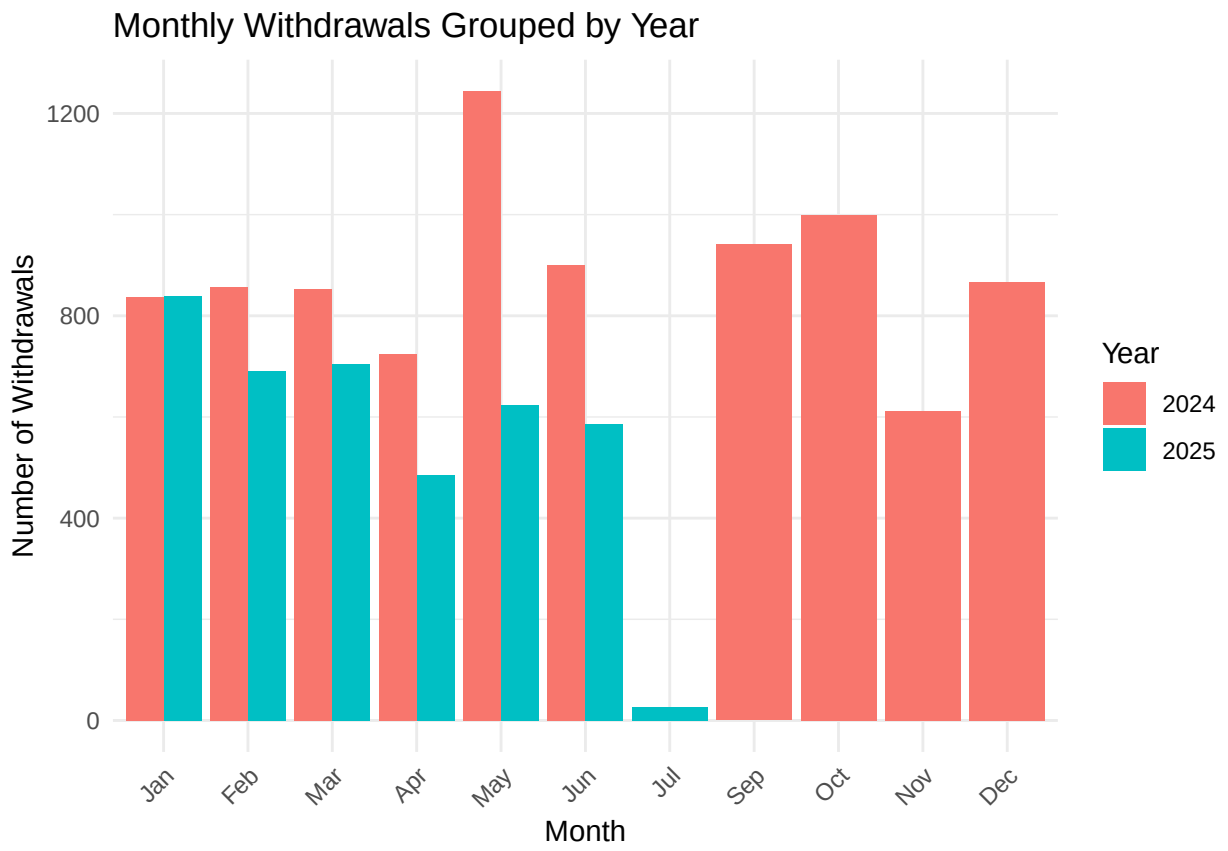
```
if (chisq_result_H2$p.value < 0.05) {
  cat("  H  is supported: Withdrawals are not evenly distributed across months.\n")
```

```r
} else {
  cat("  Fail to reject H : Withdrawals appear evenly distributed by month.\n")
}
```

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

Monthy Withdrawals: 2024 VS 2025

```r
merged_df %>%
  filter(PassStatus == "Withdrawn", !is.na(WithdrawDate)) %>%
  mutate(
    WithdrawDate = parse_date_time(WithdrawDate, orders = c("d-b-y HMS", "B d, Y, I:M:S p")),
    Month = factor(month(WithdrawDate, label = TRUE), levels = month.abb),
    Year = as.factor(year(WithdrawDate))
  ) %>%
  count(Month, Year) %>%
  ggplot(aes(x = Month, y = n, fill = Year)) +
  geom_col(position = "dodge") +
  labs(
    title = "Monthly Withdrawals Grouped by Year",
    x = "Month",
    y = "Number of Withdrawals"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
withdraw_counts <- merged_df %>%
  filter(PassStatus == "Withdrawn", !is.na(WithdrawDate)) %>%
  mutate(WithdrawDate = parse_date_time(WithdrawDate, orders = c("d-b-y HMS", "B d, Y, I:M:S p")),
         Month = month(WithdrawDate, label = TRUE)) %>%
  count(Month) %>%
  arrange(desc(n))
print(withdraw_counts)
```

```
## # A tibble: 11 x 2
##    Month     n
##    <ord> <int>
##  1 May    1868
##  2 Jan    1677
##  3 Mar    1558
##  4 Feb    1546
##  5 Jun    1488
##  6 Apr    1210
##  7 Oct     999
##  8 Sep     941
##  9 Dec     867
## 10 Nov     612
## 11 Jul      27
```

H3. Failure Trends Over Time Hypothesis: The frequency of unsuccessful outcomes (Incomplete, Complete) varies seasonally.

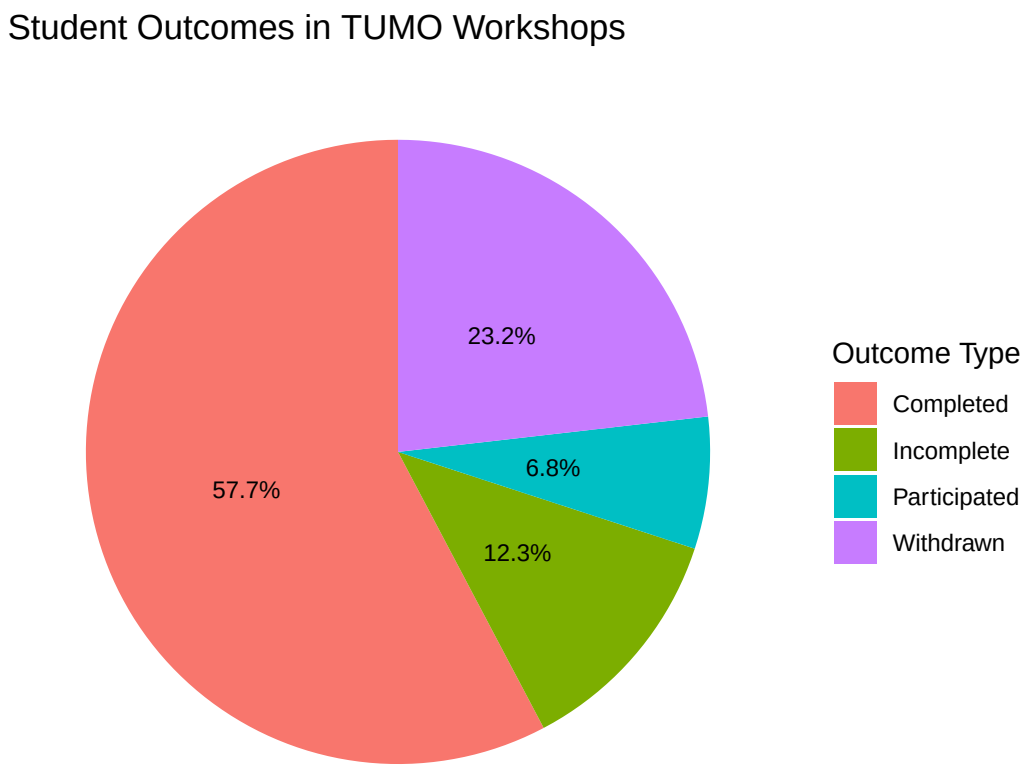
H (Null): The distribution of unsuccessful outcomes (Incomplete, Complete) is the same across all seasons (or months). H (Alternative): At least one season/month has a significantly higher failure rate (i.e., Incomplete or Participated).

Wprkshop Outcome Types

```
status_summary <- merged_df %>%
  filter(!is.na(PassStatus)) %>%
  mutate(
    OutcomeGroup = case_when(
      PassStatus == "Completed" ~ "Completed",
      PassStatus == "Participated" ~ "Participated",
      PassStatus == "Incomplete" ~ "Incomplete",
      TRUE ~ "Withdrawn"
    )
  ) %>%
  count(OutcomeGroup) %>%
  mutate(
    Percentage = round(100 * n / sum(n), 1),
    Label = paste0(OutcomeGroup, " (", Percentage, "%)")
  )

status_summary <- status_summary %>%
  mutate(Label = paste0(Percentage, "%"))

ggplot(status_summary, aes(x = "", y = n, fill = OutcomeGroup)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  geom_text(aes(label = Label), position = position_stack(vjust = 0.5), size = 3) +
  labs(
    title = "Student Outcomes in TUMO Workshops",
    x = NULL, y = NULL,
    fill = 'Outcome Type'
  ) +
  theme_minimal() +
  theme(
    axis.text = element_blank(),
    axis.ticks = element_blank(),
    panel.grid = element_blank()
  )
```



Chi-square test: Failure types vs Season

```
failure_season_table <- merged_df %>%
  filter(PassStatus %in% c("Incomplete", "Participated"), !is.na(StartDate)) %>%
  mutate(
    StartDate = as.Date(StartDate, format = "%d-%b-%y"),
    month = month(StartDate),
    Season = case_when(
      month %in% c(12, 1, 2) ~ "Winter",
      month %in% c(3, 4, 5) ~ "Spring",
      month %in% c(6, 7, 8) ~ "Summer",
      TRUE ~ "Fall"
    )
  ) %>%
  count(Season, PassStatus) %>%
  pivot_wider(names_from = PassStatus, values_from = n, values_fill = 0) %>%
  column_to_rownames("Season") %>%
  as.matrix()

chisq_result_H3 <- chisq.test(failure_season_table)
chisq_result_H3
```

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

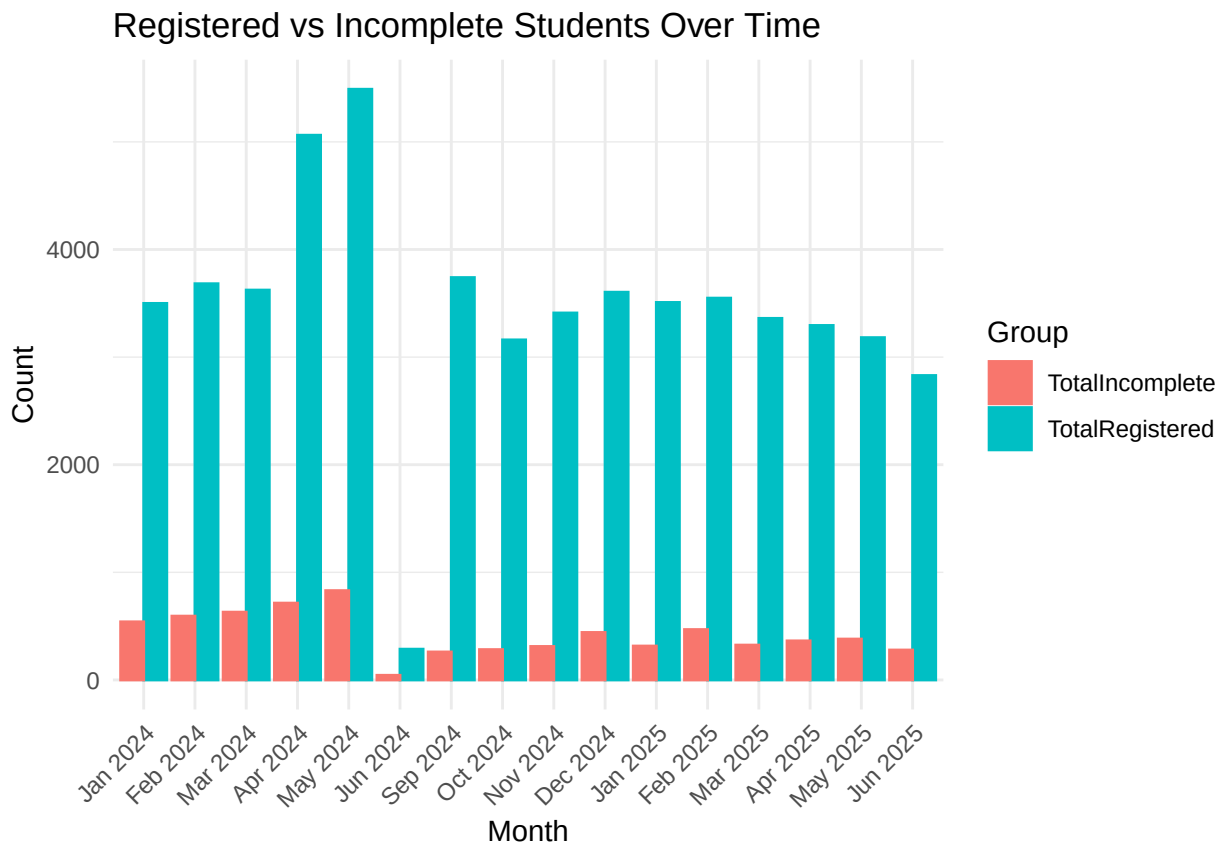
```r
if (chisq_result_H3$p.value < 0.05) {
  cat("  H  is supported: Failure outcomes vary significantly by season.\n")
} else {
  cat("  Fail to reject H : No significant seasonal trend in failure types.\n")
}
```

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

Registered VS Incomplete Students Over Time

```r
df_comp <- merged_df %>%
  filter(!is.na(PassStatus), !is.na(StartDate)) %>%
  mutate(
    StartDate = as.Date(StartDate, format = "%d-%b-%y"),
    MonthYear = format(StartDate, "%b %Y"),
    MonthYear = factor(MonthYear, levels = unique(format(sort(StartDate), "%b %Y")))
  ) %>%
  group_by(MonthYear) %>%
  summarise(
    TotalRegistered = n(),
    TotalIncomplete = sum(PassStatus == "Incomplete", na.rm = TRUE)
  )

df_comp %>%
  tidyr::pivot_longer(cols = c(TotalRegistered, TotalIncomplete),
                      names_to = "Group", values_to = "Count") %>%
  ggplot(aes(x = MonthYear, y = Count, color = Group, group = Group, fill = Group)) +
  geom_col(position = 'dodge') +
  labs(
    title = "Registered vs Incomplete Students Over Time",
    x = "Month",
    y = "Count"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Hypothesis Statement H (Null): The distribution of New vs Returning students is independent of time (month). H (Alternative): The distribution changes over time — i.e., some months have a higher/lower proportion of new students.

```
library(lubridate)
library(dplyr)
library(tidyr)

# each student's first seen month
student_first_seen <- merged_df %>%
  filter(!is.na(StartDate)) %>%
  mutate(StartDate = as.Date(StartDate, format = "%d-%b-%y")) %>%
  group_by(TumoID) %>%
  summarise(FirstMonth = format(min(StartDate), "%b %Y"), .groups = "drop")

# Tag new vs returning
student_type_tbl <- merged_df %>%
  filter(!is.na(StartDate)) %>%
  mutate(
    StartDate = as.Date(StartDate, format = "%d-%b-%y"),
    MonthYear = format(StartDate, "%b %Y")
  ) %>%
  left_join(student_first_seen, by = "TumoID") %>%
  mutate(StudentType = ifelse(MonthYear == FirstMonth, "New", "Returning")) %>%
  count(MonthYear, StudentType) %>%
  pivot_wider(names_from = StudentType, values_from = n, values_fill = 0) %>%
```

```
  column_to_rownames("MonthYear") %>%
  as.matrix()
chisq_result <- chisq.test(student_type_tbl)
chisq_result
```

```
##
##  Pearson's Chi-squared test
##
## data:  student_type_tbl
## X-squared = 16917, df = 15, p-value < 2.2e-16
```

```
if (chisq_result$p.value < 0.05) {
  cat("  H  is supported: Student type distribution (New vs Returning) varies significantly by month.\n"
} else {
  cat("  Fail to reject H : No significant difference in student type distribution across months.\n")
}
```

```
##   H  is supported: Student type distribution (New vs Returning) varies significantly by month.
```

New and Returning Students Over Time

```
student_first_seen <- merged_df %>%
  filter(!is.na(StartDate)) %>%
  mutate(StartDate = as.Date(StartDate, format = "%d-%b-%y")) %>%
  group_by(TumoID) %>%
  summarise(FirstMonth = format(min(StartDate), "%b %Y"), .groups = "drop")

first_vs_returning <- merged_df %>%
  filter(!is.na(StartDate), !is.na(TumoID)) %>%
  mutate(StartDate = as.Date(StartDate, format = "%d-%b-%y"),
         MonthYear = format(StartDate, "%b %Y")) %>%
  left_join(student_first_seen, by = "TumoID") %>%
  mutate(StudentType = ifelse(MonthYear == FirstMonth, "New", "Returning")) %>%
  count(MonthYear, StudentType) %>%
  pivot_wider(names_from = StudentType, values_from = n, values_fill = 0) %>%
  mutate(
    MonthYear = factor(MonthYear, levels = unique(format(sort(as.Date(paste0("01 ", MonthYear), format =
  ) %>%
  pivot_longer(cols = c("New", "Returning"), names_to = "Type", values_to = "Count")

ggplot(first_vs_returning, aes(x = MonthYear, y = Count, fill = Type)) +
  geom_col(position = "stack") +
  scale_y_continuous(
    breaks = seq(0, 6000, by = 1000),
    labels = scales::comma_format()
  ) +
  labs(
    title = "New vs Returning Students Over Time",
    x = "Month",
    y = "Number of Students",
```

```
    fill = "Student Type"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```