ANALYZING THE RELATIONSHIP BETWEEN GENDER AND ACADEMIC PERFORMANCE IN ONLINE COURSES USING CHI-SQUARE TEST IN R

1. Project Background and Objective

An online education provider in the UK wanted to understand if **student gender is associated with academic outcomes** in their asynchronous course environment. Concerns were raised by instructional designers about possible group-specific performance trends that could inform **equity-based learning interventions**.

Research Objective: Determine whether **student gender and performance bands** are statistically associated using a chi-square test of independence.

2. Hypotheses and Test Specification

- **Null Hypothesis (H_0):** Gender and academic performance are independent (no association).

$$H_0: \text{Gender} \perp \text{Performance}$$

- **Alternative Hypothesis (H_1):** Gender and academic performance are associated.

$$H_1: \text{Gender} / \perp \text{Performance}$$

Statistical Test Used: Chi-square test of independence using **R (base + tidyverse)**

3. Dataset Overview

Variable	Type	Description
Student_ID	Factor	Unique identifier
Gender	Factor	Male / Female
Performance	Factor	Categorical: High, Medium, Low

Sample Size: 600 students **Source:** LMS export (CSV) from two intro-level courses across a 10-week period

Performance Bands:

- High: Avg. score ≥ 80%

- Medium: 60%–79%
- Low: < 60%

4. Data Analysis in R

```
# Load packages

library(tidyverse)

# Import and prepare data

data <- read.csv("student_performance.csv")

table_data <- table(data$Gender, data$Performance)

# Chi-square test

chisq_result <- chisq.test(table_data)

# Output

chisq_result
```

5. Output Summary

Gender	High	Medium	Low	Total
Male	102	136	62	300
Female	138	122	40	300

- **Chi-square Statistic:** 9.47
- **Degrees of Freedom (df):** 2
- **p-Value: 0.0088**
- **Expected Counts:** All ≥ 5 → assumption met

6. Interpretation of Results

- Since **p < 0.05**, we reject the null hypothesis.
- There is a statistically significant association between **gender** and **academic performance**.

- **Female students** had a greater proportion of high performers compared to males in this sample.

7. Visualization

- **Stacked Bar Chart:** Gender vs. performance proportion
- **Mosaic Plot:** R's base graphics mosaic plot to show residuals
- **Standardized Residuals Plot:** Contribution of each cell to overall chi-square value

```
# Bar plot
ggplot(data, aes(x = Gender, fill = Performance)) +
  geom_bar(position = "fill") +
  ylab("Proportion") +
  ggtitle("Performance Distribution by Gender")
```

8. Deliverables

- Cleaned dataset (.csv) and reproducible R script (.R)
- 5-page analysis report including statistical assumptions, visuals, and recommendations
- PowerPoint slide deck for internal presentation
- Markdown-based executive summary hosted on company Confluence

9. Strategic Implications for the Client

- Instructional designers adjusted the course pacing and module assessments to address variation in performance
- Suggested targeted student outreach strategies, particularly for **low-performing male learners**
- Result fed into internal diversity and inclusion dashboard for ongoing monitoring

10. Relevance to Academia and Industry

- **Educational Sector:** Useful for LMS providers, online universities, or corporate L\&D teams monitoring learner success

- **Academic Use:** Teaches hypothesis testing for categorical variables using R; ideal for stats and education analytics classes