````
---
title: "THI CUỐI HỌC KÌ 2 - XỬ LÝ SỐ LIỆU THỐNG KÊ"
author: "Trần Thị Như Phụng (B2203776)"
date: "`r Sys.Date()`"
output:
  html_document: default
  pdf_document: default
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
set.seed(12345)
# Gán cố định hai số cuối MSSV
xy <- 76  # Hai số cuối MSSV là 76

# Cài đặt và load các package cần thiết
pkgs <- c("dplyr","tidyr","knitr","ggplot2","gridExtra","corrplot","pROC")
install.packages(setdiff(pkgs, rownames(installed.packages())), repos = "https://cloud.r-
project.org")
lapply(pkgs, require, character.only = TRUE)
```

# BÀI 1 (4 điểm)

```{r load-data1}
# 1. Tải và in dữ liệu
data76 <- read.csv("https://drive.google.com/uc?
export=download&id=1FSrVQRyUAVH6ITMAoItO0PnwiAIAcjaB")
head(data76)
```

```{r dim-data1}
# 2. Số hàng, cột và tên biến
dim(data76)
names(data76)
```

```{r remove-cols1}
# 3. Xóa cột không cần thiết
data76 <- data76[, !(names(data76) %in% c("Student_ID","First_Name","Last_Name","Email"))]
names(data76)
```

```{r rename-vars1}
# 4. Đổi tên biến
names(data76)[names(data76) == "Attendance...."]       <- "Attendance"
names(data76)[names(data76) == "Stress_Level..1.10."] <- "Stress_Level"
names(data76)
```

```{r missing1}
# 5. Kiểm tra và điền giá trị thiếu
sapply(data76, function(x) sum(is.na(x)))
# Điền Attendance bằng trung bình
data76$Attendance[is.na(data76$Attendance)] <- mean(data76$Attendance, na.rm = TRUE)
# Điền Assignments_Avg bằng trung vị
data76$Assignments_Avg[is.na(data76$Assignments_Avg)] <- median(data76$Assignments_Avg,
na.rm = TRUE)
# Điền Parent_Education_Level
data76$Parent_Education_Level[is.na(data76$Parent_Education_Level)] <- "Not Reported"
head(data76, 10)
```

```{r stats1}
````
````

```r
# 6. Thống kê cơ bản
library(dplyr)
library(tidyr)
library(knitr)
vars <- c("Total_Score","Attendance","Study_Hours_per_Week","Sleep_Hours_per_Night")
stat_summary <- data76 %>%
  summarise(across(all_of(vars), list(
    mean   = ~mean(.),
    median = ~median(.),
    min    = ~min(.),
    max    = ~max(.),
    sd     = ~sd(.)
  ))) %>%
  pivot_longer(everything(), names_to = c("Variable","Statistic"), names_sep = "_") %>%
  pivot_wider(names_from = Statistic, values_from = value)

kable(stat_summary)
```

```{r hist1}
# Histogram
par(mfrow = c(2,2))
for(v in vars) {
  hist(data76[[v]], main = paste("Histogram of", v), xlab = v)
}
par(mfrow = c(1,1))
```

```{r normality1}
# 7. Q-Q plot và Shapiro-Wilk
gg <- data76$Total_Score
par(mfrow = c(1,2))
qqnorm(gg); qqline(gg)
shapiro.test(gg)
par(mfrow = c(1,1))
```

```{r compare-gender1}
# 8. So sánh theo Gender và boxplot
library(ggplot2)
agg_gender <- data76 %>% group_by(Gender) %>% summarise(mean_Total_Score =
mean(Total_Score))
knitr::kable(agg_gender)

ggplot(data76, aes(x = Gender, y = Total_Score, fill = Gender)) +
  geom_boxplot() +
  theme_minimal()
```

```{r multi-plots}
# 9. 4 plots gom vào 1
library(ggplot2)
library(gridExtra)

p1 <- ggplot(data76, aes(x = Midterm_Score, y = Final_Score)) +
  geom_point() + geom_smooth(method = "lm") + theme_minimal()

p2 <- ggplot(data76, aes(x = Study_Hours_per_Week, y = Total_Score, color = Gender)) +
  geom_point() + theme_minimal()

p3 <- ggplot(data76, aes(x = Study_Hours_per_Week, y = Total_Score, color = Grade)) +
  geom_point() + theme_minimal()

p4 <- ggplot(data76, aes(x = "", fill = Grade)) +
  geom_bar(width = 1) + coord_polar("y") + theme_void()
```

```
grid.arrange(p1, p2, p3, p4, nrow = 2, ncol = 2)
```

```{r corr1}
# 10. Ma trận tương quan và heatmap
library(corrplot)
num_vars <- c("Attendance","Midterm_Score","Final_Score","Quizzes_Avg",
              "Participation_Score","Projects_Score","Total_Score",
              "Study_Hours_per_Week","Stress_Level","Sleep_Hours_per_Night")
corr_mat <- cor(data76[, num_vars], use = "complete.obs")
corrplot(corr_mat, method = "color", addCoef.col = "black", number.cex = 0.7)

# Biến tương quan cao nhất với Final_Score:
corr_mat["Final_Score", ][order(-abs(corr_mat["Final_Score", ]))]
```

```{r chi1}
# 11. Kiểm định chi-squared giữa Grade và Stress_Level
tbl <- table(data76$Grade, data76$Stress_Level)
chisq.test(tbl)
```

```{r anova1}
# 12. ANOVA và post-hoc
anova_res <- aov(Total_Score ~ Department, data = data76)
summary(anova_res)
TukeyHSD(anova_res)
```

BÀI 2 (2 điểm)

```{r load-data2}
# 1. Tải và in dữ liệu
 data2_76 <- read.csv("https://drive.google.com/uc?
export=download&id=1VP25ESfeG8bpJKOttNe95CN9H0YjHqqo")
head(data2_76)
dim(data2_76)
```

```{r stats2}
# 2. Thống kê cơ bản của gre và gpa theo admit
data2_76 %>%
  group_by(admit) %>%
  summarise(
    mean_gre   = mean(gre), median_gre = median(gre), min_gre = min(gre), max_gre =
max(gre), sd_gre = sd(gre),
    mean_gpa   = mean(gpa), median_gpa = median(gpa), min_gpa = min(gpa), max_gpa =
max(gpa), sd_gpa = sd(gpa)
  ) %>%
  knitr::kable()
```

```{r factor2}
# 3. Chuyển rank thành factor và bảng chéo
 data2_76$rank <- factor(data2_76$rank, levels = 1:4)
table(data2_76$admit, data2_76$rank)
```

```{r logistic2}
# 4. Mô hình logistic
 model2 <- glm(admit ~ gre + gpa + rank, data = data2_76, family = binomial)
 summary(model2)
```

```{r odds2}
# 6. Odds ratios (làm tròn 2 chữ số)
exp_coef <- exp(coef(model2))
round(exp_coef, 2)
```

```{r roc2}
# 7. ROC và AUC
library(pROC)
roc_res <- roc(data2_76$admit, fitted(model2))
plot(roc_res, main = paste("ROC curve (AUC =", round(auc(roc_res), 2), ")"))
```

```{r importance2}
# 8. Biến quan trọng theo Odds Ratio
imp <- data.frame(
  Variable    = names(coef(model2))[-1],
  Odds_Ratio = round(exp(coef(model2))[-1], 2)
)
knitr::kable(imp)
```