###############################################################

# Project: Problematic Internet Use and Common Mental Disorder

# Script: 02\_explore\_describe.R

# Purpose: Explore and describe the cleaned dataset

# Author: [Your Name]

# Date: [Today's Date]

###############################################################

# -------------------------------------------------------------

# 1. SETUP

# -------------------------------------------------------------

# Clear the environment (removes old variables from memory)

rm(list = ls())

# Set the working directory

# This should be the folder where your project files live.

# For example:

# setwd("C:/Users/YourName/Documents/psychology-data-analysis-R/01-students-internet-addiction")

# (If you opened this project using the .Rproj file, RStudio sets it automatically.)

# Load required packages

# If you don’t have any of these packages, install them first by running:

# install.packages(c("dplyr", "ggplot2", "psych", "skimr", "readr"))

library(dplyr) # For data manipulation

library(ggplot2) # For data visualization

library(psych) # For descriptive statistics

library(skimr) # For compact summaries

library(readr) # For reading data

# -------------------------------------------------------------

# 2. LOAD CLEAN DATA

# -------------------------------------------------------------

# Read your cleaned dataset

# (Replace with the correct path if needed)

data <- read\_csv("data/clean\_data.csv")

# Have a quick look at the data

View(data) # Opens spreadsheet view in RStudio

head(data) # Shows the first 6 rows

str(data) # Shows the structure (variable types)

summary(data) # Basic summaries for numeric variables

# -------------------------------------------------------------

# 3. UNDERSTAND VARIABLE TYPES

# -------------------------------------------------------------

# The structure tells you which variables are numeric, character, or factors.

# For example:

# - IAT\_total: numeric (Internet Addiction Test total score)

# - SRQ20\_total: numeric (Self-Reporting Questionnaire)

# - gender: factor (Male/Female)

# - age\_band: factor (e.g., 18-22, 23-27, etc.)

# - degree\_level: factor (Undergraduate/Postgraduate)

# - discipline: factor (Arts/Science/Other)

# Let’s confirm that categorical variables are set as factors:

data <- data %>%

mutate(

gender = as.factor(gender),

age\_band = as.factor(age\_band),

degree\_level = as.factor(degree\_level),

discipline = as.factor(discipline)

)

# -------------------------------------------------------------

# 4. DESCRIPTIVE STATISTICS FOR NUMERIC VARIABLES

# -------------------------------------------------------------

# The psych::describe() function gives mean, SD, min, max, skewness, etc.

numeric\_summary <- describe(select(data, IAT\_total, SRQ20\_total))

print(numeric\_summary)

# Save to CSV for your outputs folder

write.csv(numeric\_summary, "outputs/EDA/numeric\_summary.csv", row.names = TRUE)

# -------------------------------------------------------------

# 5. FREQUENCY TABLES FOR CATEGORICAL VARIABLES

# -------------------------------------------------------------

# We can use dplyr::count() or base R table()

# Gender

gender\_table <- data %>% count(gender)

print(gender\_table)

# Age band

age\_table <- data %>% count(age\_band)

print(age\_table)

# Degree level

degree\_table <- data %>% count(degree\_level)

print(degree\_table)

# Discipline

discipline\_table <- data %>% count(discipline)

print(discipline\_table)

# Combine them into one file for convenience

freq\_tables <- list(

gender = gender\_table,

age\_band = age\_table,

degree\_level = degree\_table,

discipline = discipline\_table

)

# Optional: save as separate sheets or write one by one

write.csv(gender\_table, "outputs/EDA/freq\_gender.csv", row.names = FALSE)

write.csv(age\_table, "outputs/EDA/freq\_age.csv", row.names = FALSE)

write.csv(degree\_table, "outputs/EDA/freq\_degree.csv", row.names = FALSE)

write.csv(discipline\_table, "outputs/EDA/freq\_discipline.csv", row.names = FALSE)

# -------------------------------------------------------------

# 6. DISTRIBUTION PLOTS FOR NUMERIC VARIABLES

# -------------------------------------------------------------

# Histogram of IAT\_total (Internet Addiction)

ggplot(data, aes(x = IAT\_total)) +

geom\_histogram(binwidth = 5, fill = "skyblue", color = "black") +

labs(

title = "Distribution of Internet Addiction Test (IAT) Scores",

x = "IAT Total Score",

y = "Number of Students"

)

ggsave("outputs/EDA/hist\_IAT.png")

# Histogram of SRQ20\_total (Common Mental Disorder score)

ggplot(data, aes(x = SRQ20\_total)) +

geom\_histogram(binwidth = 2, fill = "salmon", color = "black") +

labs(

title = "Distribution of SRQ-20 Scores",

x = "SRQ-20 Total Score",

y = "Number of Students"

)

ggsave("outputs/EDA/hist\_SRQ20.png")

# Boxplot of IAT by CMD status (if CMD variable exists)

# Note: CMD is defined as SRQ20\_total >= 8

data <- data %>%

mutate(CMD = ifelse(SRQ20\_total >= 8, "Yes", "No"))

ggplot(data, aes(x = CMD, y = IAT\_total, fill = CMD)) +

geom\_boxplot() +

labs(

title = "Internet Addiction Scores by CMD Status",

x = "Probable CMD (SRQ-20 ≥ 8)",

y = "IAT Total Score"

)

ggsave("outputs/EDA/box\_IAT\_by\_CMD.png")

# -------------------------------------------------------------

# 7. RELATIONSHIP BETWEEN CATEGORICAL VARIABLES

# -------------------------------------------------------------

# Example: CMD prevalence by gender

cmd\_gender <- data %>%

group\_by(gender, CMD) %>%

summarise(n = n()) %>%

mutate(percent = round(100 \* n / sum(n), 1))

print(cmd\_gender)

# Visualize as stacked bar chart

ggplot(cmd\_gender, aes(x = gender, y = percent, fill = CMD)) +

geom\_bar(stat = "identity", position = "stack") +

labs(

title = "Prevalence of CMD by Gender",

x = "Gender",

y = "Percentage of Students"

)

ggsave("outputs/EDA/bar\_CMD\_by\_gender.png")

# -------------------------------------------------------------

# 8. CORRELATION BETWEEN CONTINUOUS VARIABLES

# -------------------------------------------------------------

# IAT\_total vs SRQ20\_total

correlation <- cor(data$IAT\_total, data$SRQ20\_total, use = "complete.obs")

print(paste("Correlation between IAT and SRQ-20:", round(correlation, 3)))

# Visualize with scatter plot

ggplot(data, aes(x = IAT\_total, y = SRQ20\_total)) +

geom\_point(color = "darkblue", alpha = 0.6) +

geom\_smooth(method = "lm", color = "red", se = TRUE) +

labs(

title = "Relationship between Internet Addiction and Mental Disorder Scores",

x = "Internet Addiction Test (IAT) Score",

y = "SRQ-20 Score"

)

ggsave("outputs/EDA/scatter\_IAT\_vs\_SRQ20.png")

# -------------------------------------------------------------

# 9. SUMMARIZE KEY FINDINGS (Optional Console Output)

# -------------------------------------------------------------

cat("\n====================== SUMMARY ======================\n")

cat("1. Sample size:", nrow(data), "students\n")

cat("2. Mean IAT score:", round(mean(data$IAT\_total, na.rm = TRUE), 2), "\n")

cat("3. Mean SRQ-20 score:", round(mean(data$SRQ20\_total, na.rm = TRUE), 2), "\n")

cat("4. CMD prevalence:", round(mean(data$CMD == 'Yes') \* 100, 1), "%\n")

cat("5. Correlation (IAT vs SRQ-20):", round(correlation, 3), "\n")

cat("=====================================================\n")

# End of Script

###############################################################