**Appendix 2**

**R code for generating salary\_data**

# Install necessary packages

#install.packages("dplyr")

# Load necessary libraries

library(dplyr)

# Set seed for reproducibility

set.seed(123)

# Number of individuals

n <- 2500

# Generate age, sex, university degree status, and field of work

age <- round(runif(n, min = 20, max = 60)) # Uniform distribution for age (20 to 60)

sex <- sample(c("Male", "Female"), n, replace = TRUE, prob = c(0.5, 0.5)) # Randomly sample Male and Female

university\_degree <- sample(c("Yes", "No"), n, replace = TRUE, prob = c(0.7, 0.3)) # 70% Yes, 30% No

field\_of\_work <- sample(c("Healthcare", "Engineering", "Law", "Business", "Arts", "Finance"), n, replace = TRUE)

# Function to generate annual salary based on inputs

generate\_annual\_salary <- function(age, sex, university\_degree, field\_of\_work) {

base\_salary <- 20 # Base salary in thousands

# Age effect: Piecewise approximation capturing rapid increase, slower growth, and eventual plateau or small decrease

age\_effect <- ifelse(age <= 25, 8 \* (age - 20),

ifelse(age <= 40, 50 + 2 \* (age - 25),

90 - 1 \* (age - 40)))

# Sex effect: Adjust for gender pay gap (not real-world accurate, just for example)

sex\_effect <- ifelse(sex == "Male", 10, 0)

# University degree effect: Individuals with a degree earn more

degree\_effect <- ifelse(university\_degree == "Yes", 15, 0)

# Field of work effect: Different fields have different average salaries

work\_effect <- case\_when(

field\_of\_work == "Healthcare" ~ min(80, 30 + rnorm(1, sd = 15)),

field\_of\_work == "Engineering" ~ min(800, 60 + rnorm(1, sd = 30)),

field\_of\_work == "Law" ~ min(800, 50 + rnorm(1, sd = 20)),

field\_of\_work == "Business" ~ min(80, 40 + rnorm(1, sd = 10)),

field\_of\_work == "Arts" ~ min(80, 25 + rnorm(1, sd = 10)),

field\_of\_work == "Finance" ~ min(800, 100 + rnorm(1, sd = 40))

)

# Introduce moderate variability

variability <- rnorm(n, mean = 0, sd = 12)

# Calculate final annual salary

annual\_salary <- base\_salary + age\_effect + sex\_effect + degree\_effect + work\_effect + variability

return(annual\_salary)

}

# Generate annual salary using the function

annual\_salary <- generate\_annual\_salary(age, sex, university\_degree, field\_of\_work)

# Create data frame

salary\_data <- data.frame(age = age, sex = sex, university\_degree = university\_degree, field\_of\_work = field\_of\_work, annual\_salary = annual\_salary)

# Print first few rows of the data frame

head(salary\_data)