Answers with R codes

1. Data Inspection:

(a) (1pt) How many observations are there in the dataset? How many variables are there in the dataset?

Answer: 60 observations and 8 variables.

Code: data <- read.csv("C://Users//91626//Downloads//NLS2023.csv")

(b) (1.5pt) For each variable, how many missing entries are there? Also, give the line or lines of R code used to find this answer.

Answer: Urban = 1

Siblings = 0

White = 1

Christian = 0

FamilySize = 0

Height = 0

Weight = 1

Income = 2

Code: missing_counts <- colSums(is.na(data))

print(missing_counts)

(c) (0.5pt) Using the sort() function, find the largest number of siblings for an individual in the dataset. Also give the line or lines of R code to find this answer.

Answer: 8

Code: max_siblings <- max(sort(data$Siblings, decreasing = TRUE)[1])

print(max_siblings)

(d) (1pt) Using the order() function, order the data according to the FamilySize variable from largest to smallest. What is the largest family size in the dataset? Also, give the line or lines of R code used to find this answer.

Answer: 7

Code: largest_to_smallest <- order(-data$FamilySize)

largest_family_size <- data$FamilySize[largest_to_smallest[1]]

print(largest_family_size)

2. Data Cleaning:

(a) (1pt) Remove all individuals from the dataset with 0 income. Give the line or lines of R code for this step.

Code: data <- subset(data, Income != 0)

(b) (1pt) Afterward, replace all missing incomes with the average of the non-missing incomes. Give the line or lines of R code for this step.

Code: average_income <- mean(data$Income, na.rm = TRUE)

data$Income[is.na(data$Income)] <- average_income

(c) (1pt) Apply the log transformation (either base e or base 10) to the incomes and add this as a column to the NLS2023 data frame. Give the line or lines of R code to perform these steps.

Code: data$logIncome <- log(data$Income)

(d) (1pt) Apply the square root transformation to heights and add this as a column to the NLS data frame. Give the line or lines of R code to perform these steps.

Code: data$sqrtHeight <- sqrt(data$Height)

(e) (2pt) After performing steps (a), (b), (c), and (d), export this cleaned dataset as a .csv file called CleanNLS.csv. Give the line or lines of R code to perform this step. Also, upload this file onto Canvas.

Code: write.csv(data, file = "CleanNLS.csv", row.names = FALSE)