Lab 3 Homework: Data Cleaning and Manipulation

Background

Every two years, the CDC conducts national surveys in schools to monitor and assess the six largest contributors to youth morbidity and mortality. These contributors include not only health risks such as high body mass index, but also risky behaviors such as tobacco and alcohol use, as well as drunk driving and failure to use seatbelts. In 2013, 47 states participated in this school-based survey, yielding 13,583 respondents and 213 variables. Full survey and data documentation can be accessed on the CDC website. A subset of this dataset, which has no missing data for 16 selected variables, is provided in the file **cdc.csv**.

You are provided with a subset of the **cdc.csv** dataset, which includes the following variables:

Variable	**Description**
age	Q1: How old are you?
gender	Q2: What is your sex? (1 refers to Male and 0 refers to Female)
height_m	Calculated variable: height in meters
weight_kg	Calculated variable: weight in kilograms
bmi	Calculated variable: body mass index = weight_kg / $(\text{height\_m})^2$
BMIPCT	Calculated variable: BMI percentile for age and sex

Table 1: Variable Descriptions

Questions

1. **Import** `cdc.csv` and examine a summary of all variables included in the dataset.

2. The **age** variable is provided for each respondent, and the survey was conducted in **2013**, calculate the birth year of each respondent and create a new variable called **year_birth**.

3. The **BMI percentile (BMIPCT)** variable allows us to classify respondents based on their BMI status. Create a new variable called **BMI_category** classifying BMI status as follows:

   - Underweight: BMIPCT < 5th percentile
   - Healthy weight: 5th ≤ BMIPCT < 85th percentile
   - Overweight: 85th ≤ BMIPCT < 95th percentile
   - Obese: BMIPCT ≥ 95th percentile

   Check your new variable to ensure it was created correctly.

4. The **gender** variable is coded as 1 for Male and 0 for Female. Create a new variable called **gender_label** that classifies individuals as either "Male" or "Female", and verify the accuracy of this variable.

5. **Examine the relationship between BMI and age.**

   (a) Produce a scatter plot showing the relationship between BMI and age. Can you easily identify a trend? Why or why not?

   (b) Use a boxplot to examine the distribution of BMI by gender. Do you notice any patterns?

6. **Answer the following questions to check if your coding was successful.** (Round all answers to one decimal place, and be sure to use your new variables when answering.)

   (a) What is the average BMI for male respondents?

   (b) What is the percentage of respondents classified as "Obese" in the dataset?

   (c) How many individuals are classified as having a "Healthy weight"?