Homework 2

YOUR NAME HERE

2026-01-05

Introduction

In this assignment you will practice summarizing variables, creating new variables, handling missing data, and managing factor variables. You will work with two data sets from the `openintro` package

```
library(forcats)
library(gtsummary)
ncbirths <- openintro::ncbirths
smoking <- openintro::smoking
```

Part I: Working with Data Frames (NC Births)

A. Summarizing variables

1. Calculate the `mean` and `median` of the father's age (`fage`), removing missing values as needed.

2. Pregnancies typically last about 38 weeks. *Update* the existing variable `weeks` so that any value greater than 38 is set to 38. That is, for all record where `weeks>38`, change the value of `weeks` to `<- 38`. Display the `summary` of `weeks` to confirm that the maximum value is now 38.

3. Create a new logical variable called `missing_gained` that indicates whether the variable `gained` is missing. Show a table of your result.

4. Calculate the **proportion** of records with missing values in `gained` using **two different methods**.

5. Use the `ifelse` function to create a new variable called `term_status` where pregnancies with `weeks >= 37` are labeled `"term"` and pregnancies with `weeks < 37` are labeled `"preterm"`. Create a frequency table to verify your result.

Part II: Working with Factors (Smoking Data)

1. Use `fct_count()` to examine the distribution of the variable `ethnicity`.
2. Create a new factor variable `ethnicity_collapsed` by modifying the variable `ethnicity` such that:

- `"Refused"` and `"Unknown"` are dropped
- `"Asian"` and `"Chinese"` are combined into `"Asian"`
- all other levels remain unchanged

Verify your recode using a two-way table comparing the old `ethnicity`and new variables.

3. Using `ethnicity_collapsed`, create a new variable called `ethnicity_code` with the following labels: `"A"` for Asian, `"B"` for Black, `"M"` for Mixed, `"W"` for White. Display a frequency table using `tbl_summary` of the new variable.
4. Using the frequencies from the table above, reorder the levels of `ethnicity_collapsed` from *least frequent to most frequent.* Print a table of the reordered factor to confirm the new order.
5. Create a new factor variable called `nationality_lumped` from `nationality` that keeps the four most frequent nationalities and combines all remaining levels into a single category called “Other”. Display a table of the new variable that includes both the frequency and percent (n%)