

# Homework 2

Solutions

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: - `ncbirths` - `smoking`

```
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.

```
mean(ncbirths$fage, na.rm=TRUE)
```

[1] 30.25573

```
median(ncbirths$fage, na.rm=TRUE)
```

[1] 30

Characteristic	N = 1,000 <sup>1</sup>
missing_gained	27 (2.7%)
<sup>1</sup> n (%)	

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.

```
ncbirths$weeks[ncbirths$weeks > 38] <- 38
summary(ncbirths$weeks)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
20.00	37.00	38.00	37.17	38.00	38.00	2

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

```
ncbirths$missing_gained <- is.na(ncbirths$gained)
table(ncbirths$missing_gained)
```

FALSE	TRUE
973	27

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

```
table(ncbirths$missing_gained) |> prop.table()
```

FALSE	TRUE
0.973	0.027

```
tbl_summary(ncbirths, include = "missing_gained")
```

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.

```
ncbirths$term_status <- ifelse(ncbirths$weeks>=37, "term", "preterm")
table(ncbirths$term_status)
```

```
preterm      term
    152      846
```

## Part II: Working with Factors (Smoking Data)

1. Use `fct_count()` to examine the distribution of the variable `ethnicity`.

```
fct_count(smoking$ethnicity)
```

```
# A tibble: 7 x 2
  f      n
  <fct> <int>
1 Asian     41
2 Black     34
3 Chinese   27
4 Mixed     14
5 Refused   13
6 Unknown   2
7 White    1560
```

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.

```
smoking$ethnicity_collapsed <- fct_collapse(smoking$ethnicity,
                                              Asian = c("Asian", "Chinese"))

smoking$ethnicity_collapsed[smoking$ethnicity_collapsed %in% c("Refused", "Unknown")] <- NA
smoking$ethnicity_collapsed <- fct_drop(smoking$ethnicity_collapsed)
table(smoking$ethnicity, smoking$ethnicity_collapsed)
```

Characteristic	N = 1,691 <sup>1</sup>
recode_ethnicity	
A	68 (4.1%)
B	34 (2.0%)
M	14 (0.8%)
W	1,560 (93%)
Unknown	15
<sup>1</sup> n (%)	

	Asian	Black	Mixed	White
Asian	41	0	0	0
Black	0	34	0	0
Chinese	27	0	0	0
Mixed	0	0	14	0
Refused	0	0	0	0
Unknown	0	0	0	0
White	0	0	0	1560

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.

```
smoking$recode_ethnicity <- smoking$ethnicity_collapsed |>
  fct_recode("A"="Asian", "B"="Black", "M"="Mixed", "W"="White")

tbl_summary(smoking, include = recode_ethnicity)
```

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.

```
smoking$ethnicity_collapsed %>% fct_relevel("Mixed", "Black", "Asian", "White") %>% table()

.
.
.
Mixed Black Asian White
14     34     68   1560
```

Characteristic	N = 1,691 <sup>1</sup>
nationality_lumped	
British	538 (32%)
English	833 (49%)
Scottish	142 (8.4%)
Other	178 (11%)
<sup>1</sup> n (%)	

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%)

```
smoking$nationality_lumped <- fct_lump_n(smoking$nationality, n = 4)
tbl_summary(smoking, include = nationality_lumped)
```