DS 202 - Working in teams: Sample solution for the test case

Your Name

# Solution to the test case

The dataset brfss\_iowa.csv (linked from website) contains 6227 records from the Behavioral Risk Factor Surveillance System (BRFSS) for Iowans.

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation’s premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors.

It is conducted annually by the Center for Disease Control and Prevention (CDC).

Codebook with detailed explanations of variables is [available here](https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf).

## The Problem

1. Complete the following tasks and **write instructions / documentation** for a collaborator to be able to reproduce your findings.
2. Verify that there are 6227 cases (= number of interviews) in the data.
3. Verify that there are variables WEIGHT2 and HEIGHT3 in the data and read the description of the variables in the codebook.
4. How are height and weight related? Find correlations of weight and height by gender (SEX). How many values are the correlations based on for each gender?
5. Write a short report of your findings. Address potential problems in the data.

### Read the data and check dimensions

library(tidyverse)  
iowa <- readr::read\_csv("https://github.com/ds202-at-ISU/materials/raw/master/01\_collaborative-environment/data/brfss\_mn.csv")   
#iowa <- read\_csv("~/Documents/Data/BRFSS/brfss-iowa-2020.csv")  
# the read\_csv function from the readr package is faster than the standard read.csv  
# but the output is a tibble (we'll come back to that)  
  
dim(iowa)

## [1] 7307 330

# Yes, there are 6227 observations (rows) in the data set and a lot of columns/variables  
head(iowa)

## # A tibble: 6 × 330  
## X\_STATE FMONTH IDATE IMONTH IDAY IYEAR DISPC…¹ SEQNO X\_PSU CTELE…² PVTRE…³  
## <dbl> <dbl> <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 29 8 08082… 08 08 2015 1200 2.02e9 2.02e9 NA NA  
## 2 29 11 11172… 11 17 2015 1200 2.02e9 2.02e9 NA NA  
## 3 29 1 01152… 01 15 2015 1200 2.02e9 2.02e9 NA NA  
## 4 29 5 05202… 05 20 2015 1200 2.02e9 2.02e9 NA NA  
## 5 29 5 05212… 05 21 2015 1200 2.02e9 2.02e9 NA NA  
## 6 29 6 06052… 06 05 2015 1200 2.02e9 2.02e9 NA NA  
## # … with 319 more variables: COLGHOUS <lgl>, STATERES <dbl>, CELLFON3 <dbl>,  
## # LADULT <lgl>, NUMADULT <dbl>, NUMMEN <dbl>, NUMWOMEN <dbl>, CTELNUM1 <dbl>,  
## # CELLFON2 <dbl>, CADULT <dbl>, PVTRESD2 <dbl>, CCLGHOUS <lgl>, CSTATE <dbl>,  
## # LANDLINE <dbl>, HHADULT <dbl>, GENHLTH <dbl>, PHYSHLTH <dbl>,  
## # MENTHLTH <dbl>, POORHLTH <dbl>, HLTHPLN1 <dbl>, PERSDOC2 <dbl>,  
## # MEDCOST <dbl>, CHECKUP1 <dbl>, BPHIGH4 <dbl>, BPMEDS <dbl>, BLOODCHO <dbl>,  
## # CHOLCHK <dbl>, TOLDHI2 <dbl>, CVDINFR4 <dbl>, CVDCRHD4 <dbl>, …

### Variables HEIGHT3 and WEIGHT2

library(tidyverse)  
iowa %>% select(HEIGHT3, WEIGHT2) %>% head()

## # A tibble: 6 × 2  
## HEIGHT3 WEIGHT2  
## <dbl> <dbl>  
## 1 502 120  
## 2 509 165  
## 3 505 180  
## 4 504 130  
## 5 600 195  
## 6 503 200

Variables are there … but … some data values look odd.

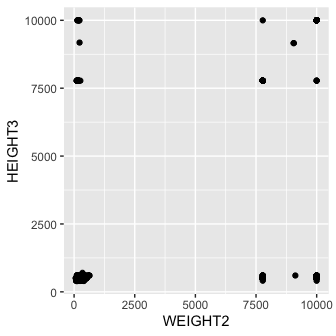
Let’s follow up by having a look at the Codebook.

From the [codebook](https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf): HEIGHT3 and WEIGHT2 are the originally recorded variables.

Let’s plot that …

Load the ggplot2 package to plot weight versus height:

library(ggplot2)  
library(tidyverse)  
iowa %>%   
 ggplot(aes(x = WEIGHT2, y = HEIGHT3)) + geom\_point()



Height and weight should show a somewhat positive correlation. These clusters are an indication of the data coding.

### Next steps

Obviously, we need to spend some time cleaning these variables before we can make use of them.

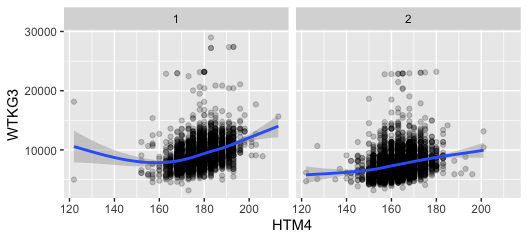
We have a couple of choices:

1. We can buckle down and do that clean-up and then move on to the calculation.
2. We can take another look at the variables and investigate what HTIN4, HTM4, and WTKG3 are …

## Route 2

Let’s take the second route first and draw a picture of height and weight in metric units:

iowa %>%   
 ggplot(aes(x = HTM4, y = WTKG3)) + geom\_point(alpha = 0.2) +   
 facet\_grid(.~SEX) +  
 geom\_smooth()



These charts look much better! We see a general increase in weight as height increases. The variability in weight is huge, though. On average, women (SEX = 2) are shorter and lighter.

### Correlations of height and weight by gender (Route 2)

Based on variables HTM4 and WTKG3

iowa %>%   
 group\_by(SEX) %>%  
 summarize(  
 n = n(),  
 cor = cor(HTM4, WTKG3, use="pairwise.complete"),  
 non\_missing = sum(!is.na(HTM4) & !is.na(WTKG3)),  
 non\_missing\_perc = non\_missing/n\*100  
 )

## # A tibble: 2 × 5  
## SEX n cor non\_missing non\_missing\_perc  
## <dbl> <int> <dbl> <int> <dbl>  
## 1 1 3055 0.333 2952 96.6  
## 2 2 4252 0.252 3845 90.4

## Route 1

This is a bit more work.

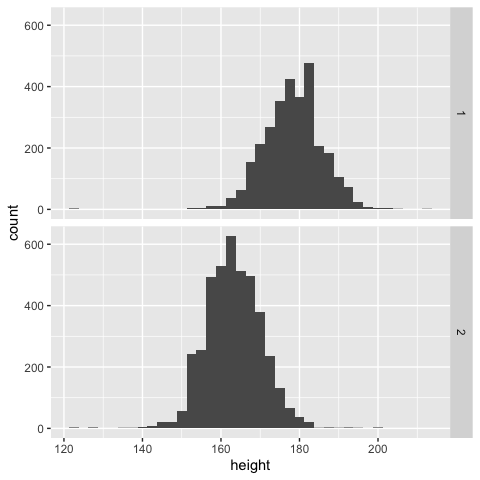
# first of all, replace all missing values by NAs  
iowa <- iowa %>% mutate(  
 HEIGHT3 = replace(HEIGHT3, HEIGHT3 %in% c(7777,9999), NA),  
 WEIGHT2 = replace(WEIGHT2, WEIGHT2 %in% c(7777,9999), NA)  
)

Let’s take on height first:

iowa <- iowa %>% mutate(  
 feet = HEIGHT3 %/% 100, # feet and inch only make sense for standard values,  
 inch = HEIGHT3 %% 100, # but we can calculated them for all  
 height = ifelse(HEIGHT3 >= 9000,   
 HEIGHT3 - 9000, # transform all metric values by subtracting 9000  
 feet \*30.48 + 2.54\*inch # add feet and inch converted to centimeter  
 )  
)

### Now plot!

iowa %>% filter(!is.na(height)) %>%  
 ggplot(aes(x = height)) + geom\_histogram(binwidth = 2.5) +  
 facet\_grid(SEX~.)



We get similar findings - for height at least: women are on average shorter than men.

Now we do the same thing for weight as we did for height before: 1 pound is 0.454 kg

iowa <- iowa %>% mutate(  
 weight = ifelse(WEIGHT2 >= 9000,   
 WEIGHT2 - 9000, # transform all metric values by subtracting 9000  
 WEIGHT2\*.454 # 1 lbs = 0.454 kg  
 )  
)

### Correlations of weight and height by gender (Route 1)

iowa %>%   
 group\_by(SEX) %>%  
 summarize(  
 n = n(),  
 cor = cor(height, weight, use="pairwise.complete"),  
 non\_missing = sum(!is.na(height) & !is.na(weight)),  
 non\_missing\_perc = non\_missing/n\*100  
 )

## # A tibble: 2 × 5  
## SEX n cor non\_missing non\_missing\_perc  
## <dbl> <int> <dbl> <int> <dbl>  
## 1 1 3055 0.333 2952 96.6  
## 2 2 4252 0.251 3845 90.4

## Recap 1

* Route 1 and 2 give identical solutions
* Read the codebook, visualize the data, even if the question does not ask for it.

## Recap 2

* the solution is disseminated from an R Markdown document, located at [a github repo](https://github.com/Stat585-at-ISU/materials-2019/tree/master/01_collaborative-environment).
* code and text/documentation are interwoven: reproducible and self-documenting.
* extend or refine analyses by copying and modifying code blocks.
* disseminate your work by sharing the RMarkdown file