Data Wrangling in R with the Tidyverse (Part 1) - Practice Solutions

Jessica Minnier, PhD & Meike Niederhausen, PhD
OCTRI Biostatistics, Epidemiology, Research & Design (BERD) Workshop

2019/04/18 (Part 1)
and again! 2019/05/16 (Part 1)

slides: bit.ly/berd_tidy1
pdf: bit.ly/berd_tidy1_pdf

Load the data and packages

```
# install.packages("tidyverse")
library(tidyverse)
library(lubridate)
demo_data <- read_csv("data/yrbss_demo.csv")
```

Practice 1

1. Import `demo_data.csv` in the `data` folder if you haven't already done so.
2. Filter `newdata` to only keep Asian or Native Hawaiian/other PI subjects that are in the 9th grade, and save again as `newdata`.
3. Filter `newdata` to remove subjects younger than 13, and save as `newdata`.
4. Remove the column `race4`, and save as `newdata`.
5. How many rows does the resulting `newdata` have? How many columns?

Practice 1 Solutions (1/2)

```
newdata <- demo_data %>%
  filter(race7 %in% c("Asian","Native Hawaiian/other PI"),
         grade == "9th",
         age != "12 years old or younger") %>%
  select(-race4)
newdata
```

```
# A tibble: 503 x 7
    record age          sex    grade race7                    bmi stweight
     <dbl> <chr>        <chr>  <chr> <chr>                  <dbl>    <dbl>
 1  924270 15 years old Male   9th   Asian                   30.7     81.6
 2 1310726 14 years old Female 9th   Asian                   30.7     81.6
 3  256154 14 years old Male   9th   Asian                   NA       NA
 4  930610 14 years old Female 9th   Native Hawaiian/other … 20.9     59.0
 5  256461 15 years old Male   9th   Asian                   NA       NA
 6  767725 14 years old Female 9th   Asian                   19.1     50.8
 7  769030 15 years old Female 9th   Native Hawaiian/other … 19.4     NA
 8  923983 15 years old Male   9th   Asian                   21.0     70.3
 9  931000 14 years old Female 9th   Asian                   18.9     45.4
10 1305660 15 years old Male   9th   Asian                   24.4     64.9
# … with 493 more rows
```

Practice 1 Solutions (2/2)

```
dim(newdata) # both nrow and ncol
```

```
[1] 503   7
```

```
nrow(newdata)
```

```
[1] 503
```

```
ncol(newdata)
```

```
[1] 7
```

Practice 2

Do the following data wrangling steps in order so that the output from the previous step is the input for the next step. Save the results in each step as `newdata`.

1. Import `demo_data.csv` in the `data` folder if you haven't already done so.
2. Create a variable called `grade_num` that has the numeric grade number (use `as.numeric`).
3. Filter the data to keep only students in grade 11 or higher.
4. Filter out rows when `bmi` is `NA`.
5. Create a binary variable called `bmi_normal` that is equal to 1 when `bmi` is between 18.5 to 24.9 and 0 when it is outside that range.
6. Arrange by `grade_num` from highest to lowest
7. Save all output to `newdata`.

Practice 2 Solutions (1/2)

```
newdata <- demo_data %>%
  separate(grade, c("grade_num"), sep = "th") %>%
  mutate(grade_num = as.numeric(grade_num)) %>%
  filter(grade_num >= 11,
         !is.na(bmi)) %>%
  mutate(
    bmi_normal = case_when(
      (18.5 <= bmi) & (bmi <= 24.9) ~ 1,
      bmi > 24.9 ~ 0,
      bmi < 18.5 ~ 0,
      )
    ) %>%
  arrange(desc(grade_num))
newdata
```

```
# A tibble: 6,630 x 9
    record age      sex    grade_num race4  race7    bmi stweight bmi_normal
     <dbl> <chr>    <chr>      <dbl> <chr>  <chr>  <dbl>    <dbl>      <dbl>
 1  333862 17 year… Fema…         12 White  White   20.2     57.2          1
 2 1309082 17 year… Male          12 White  White   19.3     59.0          1
 3  506337 18 year… Male          12 Hispa… Hispa…  33.1    123.           0
 4  938291 18 year… Fema…         12 White  White   21.7     64.9          1
 5 1316277 18 year… Fema…         12 White  White   21.6     49.9          1
```

Practice 2 Solutions - Alternative (2/2)

```
newdata <- demo_data %>%
  mutate(
    grade_num = str_replace(grade,"th",""),
    grade_num = as.numeric(grade_num),
    ) %>%
  filter(grade_num >= 11,
         !is.na(bmi)) %>%
  mutate(
    bmi_normal = case_when(
      (18.5 <= bmi) & (bmi <= 24.9) ~ 1,
      TRUE ~ 0, # TRUE is like "else", also changes NAs if any
      )
    ) %>%
  arrange(desc(grade_num))
newdata
```

```
# A tibble: 6,630 x 10
   record age   sex   grade race4 race7   bmi stweight grade_num bmi_normal
    <dbl> <chr> <chr> <chr> <chr> <chr> <dbl>    <dbl>     <dbl>      <dbl>
 1 3.34e5 17 y… Fema… 12th  White White  20.2     57.2         12          1
 2 1.31e6 17 y… Male  12th  White White  19.3     59.0         12          1
 3 5.06e5 18 y… Male  12th  Hisp… Hisp…  33.1    123.          12          0
 4 9.38e5 18 y… Fema… 12th  White White  21.7     64.9         12          1
```