Data Cleaning

Laura Albrecht

8/17/2020

1. Load in the Colorado COVID data set using the code below.
   Note: you will need to install the readr package first. read_csv is very similar to read.csv but is a faster when loading in large data sets.

```
library(readr)
colorado_covid <- read_csv("colorado_covid.csv")
```

2. Use head() and str() to see what is in the data set.

3. In the current format, each row represents one case. This isn't very helpful for understanding and visualizing the data. Try to reformat the data and save a new data frame to recreate the following:

```
## # A tibble: 19,919 x 5
## # Groups:   onset_dt, sex, age_group [3,912]
##    onset_dt   sex    age_group     ‘Race and ethnicity (combined)‘ cases
##    <date>     <chr>  <chr>         <chr>                           <int>
##  1 2020-03-01 Female 0 - 9 Years   Asian, Non-Hispanic                 1
##  2 2020-03-01 Female 0 - 9 Years   Black, Non-Hispanic                 1
##  3 2020-03-01 Female 0 - 9 Years   Hispanic/Latino                     1
##  4 2020-03-01 Female 0 - 9 Years   White, Non-Hispanic                 2
##  5 2020-03-01 Female 10 - 19 Years Unknown                             1
##  6 2020-03-01 Female 10 - 19 Years White, Non-Hispanic                 4
##  7 2020-03-01 Female 20 - 29 Years Asian, Non-Hispanic                 1
##  8 2020-03-01 Female 20 - 29 Years Black, Non-Hispanic                12
##  9 2020-03-01 Female 20 - 29 Years Hispanic/Latino                     5
## 10 2020-03-01 Female 20 - 29 Years Multiple/Other, Non-Hispanic        1
## # ... with 19,909 more rows
```

Hint: The function “n()” can be used to find a group size

4. The fourth column contains both race and ethnicity in one. Separate these variables into two columns called “race” and “ethnicity”.

5. In the age_group column, delete the word “Years” from every row.

Hint: There are many ways you could do this. Look at the separate function or gsub function help files to find two options.

6. Change the name of the first column to “date”.

Now your data should look like this:

```
## # A tibble: 19,919 x 6
## # Groups:   date, sex [533]
##    date       sex    age_group race            ethnicity    cases
##    <date>     <chr>  <chr>     <chr>           <chr>        <int>
##  1 2020-03-01 Female 0 - 9     Asian           Non-Hispanic     1
##  2 2020-03-01 Female 0 - 9     Black           Non-Hispanic     1
##  3 2020-03-01 Female 0 - 9     Hispanic/Latino <NA>             1
##  4 2020-03-01 Female 0 - 9     White           Non-Hispanic     2
##  5 2020-03-01 Female 10 - 19   Unknown         <NA>             1
##  6 2020-03-01 Female 10 - 19   White           Non-Hispanic     4
##  7 2020-03-01 Female 20 - 29   Asian           Non-Hispanic     1
##  8 2020-03-01 Female 20 - 29   Black           Non-Hispanic    12
##  9 2020-03-01 Female 20 - 29   Hispanic/Latino <NA>             5
## 10 2020-03-01 Female 20 - 29   Multiple/Other  Non-Hispanic     1
## # ... with 19,909 more rows
```

7. Use the ggplot2 package to visualize the data in a way you think is interesting. Assign variables to color, fill, or facets to display the data in different ways.