Lab 4

Enter the names of the group members here: Austine Do, Graceanne Becker, Catherine Zhong

This assignment is due by the end of the lab. Only one student in the group submits a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

In this lab, you will explore the `who2` dataset which comes with `tidyr`. Let's first load the packages we will need to complete this lab (`tidyr`, `dplyr` and `ggplot2`, all contained in`tidyverse`):

```
# Load the package
library(tidyverse)
```

Take a quick look at the dataset:

```
# Take a quick look
head(who2)
```

```
## # A tibble: 6 x 58
##   country      year sp_m_014 sp_m_1524 sp_m_2534 sp_m_3544 sp_m_4554 sp_m_5564
##   <chr>       <dbl>    <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 Afghanistan  1980       NA        NA        NA        NA        NA        NA
## 2 Afghanistan  1981       NA        NA        NA        NA        NA        NA
## 3 Afghanistan  1982       NA        NA        NA        NA        NA        NA
## 4 Afghanistan  1983       NA        NA        NA        NA        NA        NA
## 5 Afghanistan  1984       NA        NA        NA        NA        NA        NA
## 6 Afghanistan  1985       NA        NA        NA        NA        NA        NA
## # i 50 more variables: sp_m_65 <dbl>, sp_f_014 <dbl>, sp_f_1524 <dbl>,
## #   sp_f_2534 <dbl>, sp_f_3544 <dbl>, sp_f_4554 <dbl>, sp_f_5564 <dbl>,
## #   sp_f_65 <dbl>, sn_m_014 <dbl>, sn_m_1524 <dbl>, sn_m_2534 <dbl>,
## #   sn_m_3544 <dbl>, sn_m_4554 <dbl>, sn_m_5564 <dbl>, sn_m_65 <dbl>,
## #   sn_f_014 <dbl>, sn_f_1524 <dbl>, sn_f_2534 <dbl>, sn_f_3544 <dbl>,
## #   sn_f_4554 <dbl>, sn_f_5564 <dbl>, sn_f_65 <dbl>, ep_m_014 <dbl>,
## #   ep_m_1524 <dbl>, ep_m_2534 <dbl>, ep_m_3544 <dbl>, ep_m_4554 <dbl>, ...
```

The `who2` dataset contains information about tuberculosis (TB) cases per country over the years. The TB cases are reported in the columns `sp_m_014:rel_f_65` following these conventions:

1. All columns denote `new` cases.
2. The first two/three letters describe the method of diagnosis: `rel` = relapse, `sn` = negative pulmonary smear, `sp` = positive pulmonary smear, `ep` = extra pulmonary.
3. The next letter indicates the gender category: females `f` or males `m`.
4. The remaining numbers gives the age group (for example, `014` means 0-14 years old).

The goal of the lab is to compare tuberculosis (TB) cases across countries and over time, comparing number of cases per age group or per gender category.

Question 1: (2 pts)

Is the `who2` dataset tidy for comparing tuberculosis (TB) cases across countries and over time? Why/ Why not?

The data set is not tidy because each of the variables in the data set does not have its own column.

Question 2: (4 pts)

Using a `tidyr` function, put all of the column names with format *diagnosis_gender_age* into a single column (call it `diagnosis_gender_age`) and all of their cell values into another single column (call it "cases"). Call the resulting dataset `long_who`. How many rows does the `long_who` dataset have?

```
# pivoting the dataset longer and calling it 'long_who'
who2_col_names <- colnames(who2)

long_who <- pivot_longer(who2,
                         cols = who2_col_names[3:58],
                         names_to = 'diagnosis_gender_age',
                         values_to = 'cases')
long_who
```

```
## # A tibble: 405,440 x 4
##    country      year diagnosis_gender_age cases
##    <chr>       <dbl> <chr>                <dbl>
##  1 Afghanistan  1980 sp_m_014                NA
##  2 Afghanistan  1980 sp_m_1524               NA
##  3 Afghanistan  1980 sp_m_2534               NA
##  4 Afghanistan  1980 sp_m_3544               NA
##  5 Afghanistan  1980 sp_m_4554               NA
##  6 Afghanistan  1980 sp_m_5564               NA
##  7 Afghanistan  1980 sp_m_65                 NA
##  8 Afghanistan  1980 sp_f_014                NA
##  9 Afghanistan  1980 sp_f_1524               NA
## 10 Afghanistan  1980 sp_f_2534               NA
## # i 405,430 more rows
```

```
glimpse(long_who)
```

```
## Rows: 405,440
## Columns: 4
## $ country              <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afg~
## $ year                 <dbl> 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1~
## $ diagnosis_gender_age <chr> "sp_m_014", "sp_m_1524", "sp_m_2534", "sp_m_3544"~
## $ cases                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

long_who has 405,440 rows in the data set.

Question 3: (4 pts)

Next, separate the `diagnosis_gender_age` variable into `diagnosis`, `gender`, and `age`. Call the resulting dataset `tidy_who`. Is that data tidy?

```
# This separates the diagonsis_gender_age into appropriate column variables
tidy_who <- separate(long_who, diagnosis_gender_age, into = c('diagonsis', 'gender', 'age'), sep = '_')
tidy_who
```

```
## # A tibble: 405,440 x 6
##    country      year diagonsis gender age   cases
##    <chr>       <dbl> <chr>     <chr>  <chr> <dbl>
##  1 Afghanistan  1980 sp        m      014      NA
##  2 Afghanistan  1980 sp        m      1524     NA
##  3 Afghanistan  1980 sp        m      2534     NA
##  4 Afghanistan  1980 sp        m      3544     NA
##  5 Afghanistan  1980 sp        m      4554     NA
##  6 Afghanistan  1980 sp        m      5564     NA
##  7 Afghanistan  1980 sp        m      65       NA
##  8 Afghanistan  1980 sp        f      014      NA
##  9 Afghanistan  1980 sp        f      1524     NA
## 10 Afghanistan  1980 sp        f      2534     NA
## # i 405,430 more rows
```

The data set tidy_who is now tidy since country, year, diagnosis, gender, and age now have their own column and each row/observation is distinct.

Question 4: (3 pts)

Let's take a look at missing values in `tidy_who`. There are some missing values for `cases`. But does a missing value mean that there was no case of TB for a specific country/year or does it mean that the WHO did not report the number of TB cases for a specific country/year? *Hint: Are there any zeros in our* ***`tidy_who`*** *dataset?*

```
# Filtering tidy_who to see how many rows has 0 and NA values in the `cases` column
tidy_who |>
    filter(cases == 0)
```

```
## # A tibble: 11,080 x 6
##    country      year diagonsis gender age   cases
##    <chr>       <dbl> <chr>     <chr>  <chr> <dbl>
##  1 Afghanistan  1997 sp        m      014       0
##  2 Afghanistan  1997 sp        m      65        0
##  3 Afghanistan  1997 sp        f      5564      0
##  4 Afghanistan  2007 sn        m      014       0
```

```
##  5 Afghanistan  2007 sn        m      1524      0
##  6 Afghanistan  2007 sn        m      2534      0
##  7 Afghanistan  2007 sn        m      3544      0
##  8 Afghanistan  2007 sn        m      4554      0
##  9 Afghanistan  2007 sn        m      5564      0
## 10 Afghanistan  2007 sn        m      65        0
## # i 11,070 more rows
```

```
tidy_who |>
    filter(is.na(cases))
```

```
## # A tibble: 329,394 x 6
##    country      year diagonsis gender age   cases
##    <chr>       <dbl> <chr>     <chr>  <chr> <dbl>
##  1 Afghanistan  1980 sp        m      014      NA
##  2 Afghanistan  1980 sp        m      1524     NA
##  3 Afghanistan  1980 sp        m      2534     NA
##  4 Afghanistan  1980 sp        m      3544     NA
##  5 Afghanistan  1980 sp        m      4554     NA
##  6 Afghanistan  1980 sp        m      5564     NA
##  7 Afghanistan  1980 sp        m      65       NA
##  8 Afghanistan  1980 sp        f      014      NA
##  9 Afghanistan  1980 sp        f      1524     NA
## 10 Afghanistan  1980 sp        f      2534     NA
## # i 329,384 more rows
```

There are 11,080 rows/observations that contain 0 for cases so that does not means that the rows/observations that contains missing values for cases had 0 cases, they just were not reported in the data set for that year, diagonsis, age group, gender, and country.

Question 5: (4 pts)

What about missing years for some countries? These missing years would not appear explicitly in the dataset, they just would not be there... Using `group_by()` and `summarize`, find the total number of distinct years for each country in `tidy_who`. Also report the minimum and maximum year contained in the dataset for each country. Which countries had less than the expected 34 years (1980 to 2013)? Why do you think these years are missing? *Hint: To understand why we have missing years, look at* ***`Serbia & Montenegro`****. What happened to this country in 2005?*

```
# Finding the number of distinct years by country and then finding the countries with missing years in
tidy_who |>
    group_by(country) |>
    summarize(num_of_distinct_year = n_distinct(year), max_year = max(year), min_year = min(year))
```

```
## # A tibble: 219 x 4
##    country             num_of_distinct_year max_year min_year
##    <chr>                              <int>    <dbl>    <dbl>
##  1 Afghanistan                           34     2013     1980
##  2 Albania                               34     2013     1980
##  3 Algeria                               34     2013     1980
```

```
##  4 American Samoa                      34     2013     1980
##  5 Andorra                             34     2013     1980
##  6 Angola                              34     2013     1980
##  7 Anguilla                            34     2013     1980
##  8 Antigua and Barbuda                 34     2013     1980
##  9 Argentina                           34     2013     1980
## 10 Armenia                             34     2013     1980
## # i 209 more rows
```

```
tidy_who |>
    group_by(country) |>
    summarize(num_of_distinct_year = n_distinct(year)) |>
    filter(num_of_distinct_year < 34)
```

```
## # A tibble: 9 x 2
##   country                          num_of_distinct_year
##   <chr>                                           <int>
## 1 Bonaire, Saint Eustatius and Saba                   4
## 2 Curacao                                             4
## 3 Montenegro                                          9
## 4 Netherlands Antilles                               30
## 5 Serbia                                              9
## 6 Serbia & Montenegro                                25
## 7 Sint Maarten (Dutch part)                           4
## 8 South Sudan                                         3
## 9 Timor-Leste                                        12
```

There are typically 34 distinct years per country. The minimum year is 1980 and the maximum year is 2013.The countries that had less than expected 34 years are `Bonaire, Saint Eustatius and Saba`, `Curacao`, `Montenegro`, `Netherlands Antilles`, `Serbia`, `Serbia & Montenegro`, `Sint Maarten, South Sudan`, `Timor-Leste`. We think these years are missing because of political conflict and various declarations of independence affecting the status of each country, therefore affecting accurate data reporting for those various countries.

Question 6: (6 pts)

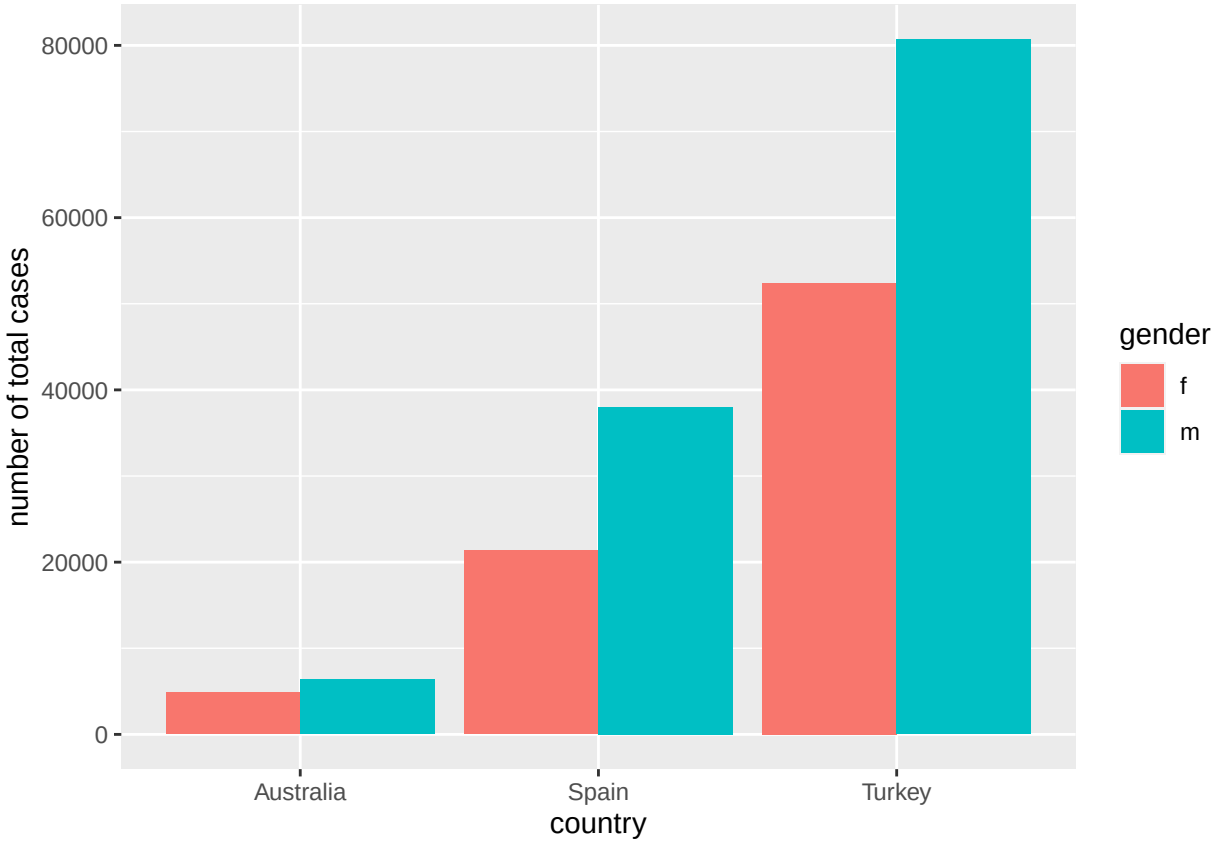
Investigate the total number of TB cases (adding up cases over all years and across all methods of diagnosis), in the countries of your choice (each group member picks a country), and comparing either age groups or gender categories. Write a research question that your investigation would answer. *For example, (create a question of your own, don't use this one!): How did the total number of TB cases differ between age groups in Belgium, France, and Germany?*

How did the total number of cases of all types of TB differ between genders across Turkey, Spain, and Australia?

Answer your research question using some `dplyr` functions and a `ggplot` visualization. Why should we be careful in interpreting what we see?

```
# Filtering the data set to only contain Turkey, Spain, and Australia, grouping
# by country and gender, then visualizing the total number of cases of all
# types of TB grouped by gender and country
```

```r
tidy_who |>
    filter(country %in% c('Turkey', 'Spain','Australia')) |>
    group_by(country, gender) |>
    summarize(num_of_case = sum(cases, na.rm = TRUE)) |>
    ggplot(aes(x = country, y = num_of_case, fill = gender)) +
    geom_bar(stat = 'identity', position = 'dodge') +
    labs(y = 'number of total cases')
```



We should be careful about the data visualization because there could be outliers and NA values causing variation in the data. Also, this is a small sample size of countries where there could be missing observations so the visualization could be misleading.

Question 7: (1 pt)

After investigating how the number of TB cases might change over time, did the data match your expectations or not? If the data differed from your expectation, provide a possible explanation for why the data differed from what you expected.

The data matches our expectation because male cases exceed females for each of the countries we explored in the data set. This shows that there might be a plausible relation between all types of TB cases across gender per country.

Formatting: (1 pt)

Make sure the names of all group members are included at the beginning of the document.

Knit your file! You can knit into pdf directly or into html. Once it knits in html, click on `Open in Browser` at the top left of the window pops out. Print your html file into pdf from your browser.

Any issue? Ask other classmates or TA!

Finally, remember to select pages for each question when submitting your pdf to Gradescope and to identify your group members.