STT 301 Homework Assignment 3

Shawn Santo

October 23, 2018

Homework Assignment 3 is due Wednesday, October 31 at 12:40pm EST.

Instructions

This assignment is to be done in groups using R Markdown. Groups are posted on D2L. One Rmd file per group should be submitted to the dropbox folder by the above deadline with each individual's name listed in the "author" section.

Everyone in the group will earn the same grade.

Rubric

- **Total**: 10 points.
- **Correctness**: Point values for the question and their respective parts are listed. Partial credit is available. Hard-coded solutions will not receive full credit.
- **Knitting**: Deduction of 0.5 points if the Rmd file does not knit for any reason.
- **Style**: Use a third-level header to off-set each question in your solutions - as is done below. For questions with multiple parts (part a, part b, etc), use fourth-level headers to off-set the parts in your solutions - as is done below. Use code comments for subsubparts. Coding style is very important. You will receive a deduction of up to 1.0 point if you do not adhere to good coding style. What I am looking for in terms of style includes:
  - appropriate variable use and naming
  - appropriate function use
  - good code commenting
  - consistent code syntax
- **Code documentation**: Code should be well documented.
- **Late Submission**: Late homework will not be accepted.

Please do not include the above Rubric, Instructions, and homework deadline sections in your solutions.

Question 1 (6 points)

The `tb_cases.csv` file (available on D2L - Data Sets section) contains tuberculosis (TB) cases by country, year, age, gender, and diagnosis method. The data is from 1980 to 2013. A data dictionary is available at http://www.who.int/tb/country/data/download/en/ (http://www.who.int/tb/country/data/download/en/).

The objective in question 1 is to make the data tidy. Each of the subsequent parts will help you in the process of tidying the data. The resulting tibble is shown (by default only 10 rows) for most of the parts. You must use the functions in the `tidyr` package and `dplyr` package (both of which are loaded when you load the `tidyverse` package) along with the pipe operator (where applicable) to earn full credit. You will also need to load the `stringr` package.

You should think about why you are doing what you are doing and how it is a step in the process to tidy data. This may be a large component of your project, so it is imperative to have a good understanding of what is going on and why it is being done.

Part a (0.25 points)

Read in the `tb_cases.csv` file and save it as an object named `tb.cases`. Convert `tb.cases` to a tibble using `as_tibble` and save it as `tb.cases`. The result should be as below.

```
# A tibble: 7,240 x 60
   country iso2  iso3   year new_sp_m014 new_sp_m1524 new_sp_m2534
   <chr>   <chr> <chr> <int>       <int>        <int>        <int>
 1 Afghan~ AF    AFG    2013          NA           NA           NA
 2 Albania AL    ALB    2013          NA           NA           NA
 3 Algeria DZ    DZA    2013          NA           NA           NA
 4 Americ~ AS    ASM    2013          NA           NA           NA
 5 Andorra AD    AND    2013          NA           NA           NA
 6 Angola  AO    AGO    2013          NA           NA           NA
 7 Anguil~ AI    AIA    2013          NA           NA           NA
 8 Antigu~ AG    ATG    2013          NA           NA           NA
 9 Argent~ AR    ARG    2013          NA           NA           NA
10 Armenia AM    ARM    2013          NA           NA           NA
# ... with 7,230 more rows, and 53 more variables: new_sp_m3544 <int>,
#   new_sp_m4554 <int>, new_sp_m5564 <int>, new_sp_m65 <int>,
#   new_sp_f014 <int>, new_sp_f1524 <int>, new_sp_f2534 <int>,
#   new_sp_f3544 <int>, new_sp_f4554 <int>, new_sp_f5564 <int>,
#   new_sp_f65 <int>, new_sn_m014 <int>, new_sn_m1524 <int>,
#   new_sn_m2534 <int>, new_sn_m3544 <int>, new_sn_m4554 <int>,
#   new_sn_m5564 <int>, new_sn_m65 <int>, new_sn_f014 <int>,
#   new_sn_f1524 <int>, new_sn_f2534 <int>, new_sn_f3544 <int>,
#   new_sn_f4554 <int>, new_sn_f5564 <int>, new_sn_f65 <int>,
#   new_ep_m014 <int>, new_ep_m1524 <int>, new_ep_m2534 <int>,
#   new_ep_m3544 <int>, new_ep_m4554 <int>, new_ep_m5564 <int>,
#   new_ep_m65 <int>, new_ep_f014 <int>, new_ep_f1524 <int>,
#   new_ep_f2534 <int>, new_ep_f3544 <int>, new_ep_f4554 <int>,
#   new_ep_f5564 <int>, new_ep_f65 <int>, newrel_m014 <int>,
#   newrel_m1524 <int>, newrel_m2534 <int>, newrel_m3544 <int>,
#   newrel_m4554 <int>, newrel_m5564 <int>, newrel_m65 <int>,
#   newrel_f014 <int>, newrel_f1524 <int>, newrel_f2534 <int>,
#   newrel_f3544 <int>, newrel_f4554 <int>, newrel_f5564 <int>,
#   newrel_f65 <int>
```

Part b (1 point)

Modify `tb.cases` to get the result you see below. Save the new tibble as `tb.cases1`. If you get a tibble with 405,440 rows it is because you did not remove the `NA` values. The below tibble is enough to note all the changes that were made.

```
# A tibble: 76,046 x 6
   country             iso2  iso3   year diag        cases
 * <chr>               <chr> <chr> <int> <chr>       <int>
 1 Afghanistan         AF    AFG    2012 new_sp_m014   188
 2 Albania             AL    ALB    2012 new_sp_m014     0
 3 Algeria             DZ    DZA    2012 new_sp_m014    29
 4 Andorra             AD    AND    2012 new_sp_m014     0
 5 Angola              AO    AGO    2012 new_sp_m014   390
 6 Anguilla            AI    AIA    2012 new_sp_m014     0
 7 Antigua and Barbuda AG    ATG    2012 new_sp_m014     0
 8 Argentina           AR    ARG    2012 new_sp_m014    59
 9 Armenia             AM    ARM    2012 new_sp_m014     1
10 Australia           AU    AUS    2012 new_sp_m014     3
# ... with 76,036 more rows
```

Part c (1 point)

A note on the `diag` variable from `tb.cases1`.

1. The first three letters denote whether it is a new or old case of TB. In this data set all are new cases of TB.
2. The next two letters after `new` describe the type of TB.
   - `rel` stands for relapse cases
   - `ep` stands for extrapulmonary TB cases
   - `sn` stands for pulmonary TB cases that could not be diagnosed by a pulmonary smear
   - `sp` stands for pulmonary TB cases that could be diagnosed by a pulmonary smear
3. The subsequent letter gives the gender (`m` or `f`).
4. The numbers that conclude the string signify an age group.
   - 014 = 0-14 years old
   - 1524 = 15-24 years old
   - 2534 = 25-34 years old
   - 3544 = 35-44 years old
   - 4554 = 45-54 years old
   - 5564 = 55-64 years old
   - 65 = 65 or older

Look at a table of the `diag` variable from `tb.cases1`. You will notice that we have an inconsistency with regards to the values of this variable: `newrel` instead of `new_rel`. We will fix this so all values of the variable start with `new_rel`. To do this, use the `mutate` function (over-write the `diag` variable) along with `str_replace(diag, "newrel", "new_rel")`. Modify `tb.cases1` to get the result you see below. Save the new tibble as `tb.cases2`. You can look at a table of the `diag` variable from your new tibble to see if the change was made correctly.

```
# A tibble: 76,046 x 6
   country             iso2  iso3   year diag        cases
   <chr>               <chr> <chr> <int> <chr>       <int>
 1 Afghanistan         AF    AFG    2012 new_sp_m014   188
 2 Albania             AL    ALB    2012 new_sp_m014     0
 3 Algeria             DZ    DZA    2012 new_sp_m014    29
 4 Andorra             AD    AND    2012 new_sp_m014     0
 5 Angola              AO    AGO    2012 new_sp_m014   390
 6 Anguilla            AI    AIA    2012 new_sp_m014     0
 7 Antigua and Barbuda AG    ATG    2012 new_sp_m014     0
 8 Argentina           AR    ARG    2012 new_sp_m014    59
 9 Armenia             AM    ARM    2012 new_sp_m014     1
10 Australia           AU    AUS    2012 new_sp_m014     3
# ... with 76,036 more rows
```

Part d (1 point)

Modify `tb.cases2` to get the result you see below. Save the new tibble as `tb.cases3`. The below tibble is enough to note all the changes that were made.

```
# A tibble: 76,046 x 8
   country             iso2  iso3   year new   type  sex.age cases
   <chr>               <chr> <chr> <int> <chr> <chr> <chr>   <int>
 1 Afghanistan         AF    AFG    2012 new   sp    m014      188
 2 Albania             AL    ALB    2012 new   sp    m014        0
 3 Algeria             DZ    DZA    2012 new   sp    m014       29
 4 Andorra             AD    AND    2012 new   sp    m014        0
 5 Angola              AO    AGO    2012 new   sp    m014      390
 6 Anguilla            AI    AIA    2012 new   sp    m014        0
 7 Antigua and Barbuda AG    ATG    2012 new   sp    m014        0
 8 Argentina           AR    ARG    2012 new   sp    m014       59
 9 Armenia             AM    ARM    2012 new   sp    m014        1
10 Australia           AU    AUS    2012 new   sp    m014        3
# ... with 76,036 more rows
```

Part e (1 point)

Modify `tb.cases3` to get the result you see below. Save the new tibble as `tb.cases4`. The below tibble is enough to note all the changes that were made.

```
# A tibble: 76,046 x 5
   country              year type  sex.age cases
   <chr>               <int> <chr> <chr>   <int>
 1 Afghanistan          2012 sp    m014      188
 2 Albania              2012 sp    m014        0
 3 Algeria              2012 sp    m014       29
 4 Andorra              2012 sp    m014        0
 5 Angola               2012 sp    m014      390
 6 Anguilla             2012 sp    m014        0
 7 Antigua and Barbuda  2012 sp    m014        0
 8 Argentina            2012 sp    m014       59
 9 Armenia              2012 sp    m014        1
10 Australia            2012 sp    m014        3
# ... with 76,036 more rows
```

Part f (1 point)

Modify `tb.cases4` to get the result you see below. Save the new tibble as `tb.cases5` . The below tibble is enough to note all the changes that were made.

```
# A tibble: 76,046 x 6
   country              year type  sex   age   cases
   <chr>               <int> <chr> <chr> <chr> <int>
 1 Afghanistan          2012 sp    m     014     188
 2 Albania              2012 sp    m     014       0
 3 Algeria              2012 sp    m     014      29
 4 Andorra              2012 sp    m     014       0
 5 Angola               2012 sp    m     014     390
 6 Anguilla             2012 sp    m     014       0
 7 Antigua and Barbuda  2012 sp    m     014       0
 8 Argentina            2012 sp    m     014      59
 9 Armenia              2012 sp    m     014       1
10 Australia            2012 sp    m     014       3
# ... with 76,036 more rows
```

Part g (0.25 points)

Modify `tb.cases5` to get the result you see below. Save the new tibble as `tb.cases6` . The below tibble is enough to note all the changes that were made.

```
# A tibble: 76,046 x 6
   country              year age   sex   type  cases
   <chr>               <int> <chr> <chr> <chr> <int>
 1 Afghanistan          2012 014   m     sp      188
 2 Albania              2012 014   m     sp        0
 3 Algeria              2012 014   m     sp       29
 4 Andorra              2012 014   m     sp        0
 5 Angola               2012 014   m     sp      390
 6 Anguilla             2012 014   m     sp        0
 7 Antigua and Barbuda  2012 014   m     sp        0
 8 Argentina            2012 014   m     sp       59
 9 Armenia              2012 014   m     sp        1
10 Australia            2012 014   m     sp        3
# ... with 76,036 more rows
```

Part h (0.50 points)

Use one of the apply family of functions to change each of the character variables to a factor. The tibble should remain named `tb.cases6` . The result is below.

```
# A tibble: 76,046 x 6
   country              year age   sex   type  cases
   <fct>               <int> <fct> <fct> <fct> <int>
 1 Afghanistan          2012 014   m     sp      188
 2 Albania              2012 014   m     sp        0
 3 Algeria              2012 014   m     sp       29
 4 Andorra              2012 014   m     sp        0
 5 Angola               2012 014   m     sp      390
 6 Anguilla             2012 014   m     sp        0
 7 Antigua and Barbuda  2012 014   m     sp        0
 8 Argentina            2012 014   m     sp       59
 9 Armenia              2012 014   m     sp        1
10 Australia            2012 014   m     sp        3
# ... with 76,036 more rows
```

Question 2 (2 points)

Use the functions in the `dplyr` package along with the tibble `tb.cases6` for the following parts.

Part a (0.50 points)

Create a tibble showing the total number of cases of TB for each age group and gender.

Part b (0.50 points)

In what country and year were TB cases highest?

Part c (0.50 points)

Create a tibble that shows the total number of TB cases for each year and diagnosis type, but only for years after 2009.

Part d (0.50 points)

Give code to produce the tibble you see below.

```
# A tibble: 29 x 4
# Groups:   country [5]
   country                   year total.cases avg.per.month
   <fct>                    <int>       <int>         <dbl>
 1 Brazil                    2008       70484         5874.
 2 Brazil                    2009       71572         5964.
 3 Brazil                    2010       70848         5904
 4 Brazil                    2011       71202         5934.
 5 Brazil                    2012       71072         5923.
 6 Brazil                    2013       75996         6333
 7 China                     2008      462596        38550.
 8 China                     2009      884477        73706.
 9 China                     2010      869092        72424.
10 China                     2011      865059        72088.
11 China                     2012      858861        71572.
12 China                     2013      847176        70598
13 India                     2008      615492        51291
14 India                     2009      624617        52051.
15 India                     2010      630164        52514.
16 India                     2011      642311        53526.
17 India                     2012      629589        52466.
18 Indonesia                 2008      292899        24408.
19 Indonesia                 2009      289044        24087
20 Indonesia                 2010      296272        24689.
21 Indonesia                 2011      313601        26133.
22 Indonesia                 2012      322882        26907.
23 Indonesia                 2013      325582        27132.
24 United States of America  2008       12893         1074.
25 United States of America  2009       11370          948.
26 United States of America  2010       10305          859.
27 United States of America  2011       10319          860.
28 United States of America  2012        9918          826.
29 United States of America  2013        9106          759.
```

Question 3 (2 points)

Use `ggplot` , or similar packages in the grammar of graphics, to create at least three distinct and informative visualizations of the `tb.cases6` tibble. The plots should be descriptive and well labeled as if you were using these in a presentation or paper.

Furthermore, you should pose at least one question you would be interested in further investigating based off the plots and your data analysis. The questions need not be answered. It is also okay if the questions you raise are unable to be answered based off the scope of the data. For example: “Does less government spending result in more relapse TB cases?”.