Assignment - 1 : Tidy With Tidyr

Raj Kumar Seth (UMDS20006)

October 15, 2020

1 Library

```
library(tidyr)
library(tidyverse)
library(ggplot2)
```

2 Exercises

Q.1. Using prose, describe how the variables and observations are organized in each of the sample tables.

Ans:

```
table1

    A tibble: 6 x 4
Index country year cases population
int      chr     int     int     int
1 Afghanistan 1999 745 19987071
2 Afghanistan 2000 2666 20595360
3 Brazil 1999 37737 172006362
4 Brazil 2000 80488 174504898
5 China 1999 212258 1272915272
6 China 2000 213766 1280428583

    table2

    A tibble: 12 x 4
Index country year type count
int     chr     int     chr     int
1 Afghanistan 1999 cases 745
2 Afghanistan 1999 population 19987071
3 Afghanistan 2000 cases 2666
4 Afghanistan 2000 population 20595360
5 Brazil 1999 cases 37737
6 Brazil 1999 population 172006362
7 Brazil 2000 cases 80488
8 Brazil 2000 population 174504898
```

```
9 China 1999 cases 212258
10 China 1999 population 1272915272
11 China 2000 cases 213766
12 China 2000 population 1280428583
```

table3

A tibble: 6 x 3

```
Index country year rate
int   chr     int  chr
1 Afghanistan 1999 745/19987071
2 Afghanistan 2000 2666/20595360
3 Brazil 1999 37737/172006362
4 Brazil 2000 80488/174504898
5 China 1999 212258/1272915272
6 China 2000 213766/1280428583
```

table4a

A tibble: 3 x 3

```
Index country ‘1999‘ ‘2000‘
int   chr     int    int
1 Afghanistan 745 2666
2 Brazil 37737 80488
3 China 212258 213766
```

table4b

A tibble: 3 x 3

```
Index country ‘1999‘ ‘2000‘
int   chr      int    int
1 Afghanistan 19987071 20595360
2 Brazil 172006362 174504898
3 China 1272915272 1280428583
```

Q.2 Compute the rate for table2, and table4a + table4b. You will need to perform four operations:

a. Extract the number of TB cases per country per year.

b. Extract the matching population per country per year.

c. Divide cases by population, and multiply by 10000.

d. Store back in the appropriate place.

e.Which representation is easiest to work with? Which is hardest? Why?

Ans:

¡

Q.3 Re-create the plot showing change in cases over time using table2 instead of table1. What do you need to do first?

Ans:

Q.4. Why are gather() and spread() not perfectly symmetrical?
Carefully consider the following example:

```
stocks ¡- tibble(
year = c(2015, 2015, 2016, 2016),
half = c( 1, 2, 1, 2),
return = c(1.88, 0.59, 0.92, 0.17)
)
stocks spread(year, return) gather(”year”, ”return”, ‘2015‘:‘2016‘)
```

(Hint: look at the variable types and think about column names.) Both spread() and gather() have a convert argument. What does it do?

Ans:

Q.5. Why does this code fail?

```
table4a gather(1999, 2000, key = ”year”, value = ”cases”)
Error in eval(expr, envir, enclos):
Position must be between 0 and n
```

Ans:

Q.6. Why does spreading this tibble fail? How could you add a new column to fix the problem?

```
people ¡- tribble(
name, key, value,
————————————
”Phillip Woods”, ”age”, 45,
”Phillip Woods”, ”height”, 186,
”Phillip Woods”, ”age”, 50,
”Jessica Cordero”, ”age”, 37,
”Jessica Cordero”, ”height”, 156
)
```

Ans:

Q.7. Tidy this simple tibble. Do you need to spread or gather it? What are the variables?

```
preg ¡- tribble(
pregnant, male, female,
”yes”, NA, 10,
”no”, 20, 12
)
```

Ans:

Q.8. What do the extra and fill arguments do in separate()?
Experiment with the various options for the following two toy datasets:

```
tibble(x = c(”a,b,c”, ”d,e,f,g”, ”h,i,j”)) separate(x, c(”one”, ”two”, ”three”))
tibble(x = c(”a,b,c”, ”d,e”, ”f,g,i”)) separate(x, c(”one”, ”two”, ”three”))
```

Ans:

Q.9.. Both unite() and separate() have a remove argument. What does it do? Why would you set it to FALSE?

Ans:

Q.10. Compare and contrast separate() and extract(). Why arethere three variations of separation (by position, by separator,and with groups), but only one unite?

Ans:

Q.11. Compare and contrast the fill arguments to spread() and complete().

Ans:

Q.12. What does the direction argument to fill() do?

Ans:

Q.13 In this case study I set na.rm = TRUE just to make it easier tocheck that we had the correct values. Is this reasonable?
Think about how missing values are represented in this dataset. Are there implicit missing values? What's the difference between an NA and zero?

Ans:

Q.14. What happens if you neglect the mutate() step?

Ans:

Q.15. I claimed that iso2 and iso3 were redundant with country.Confirm this claim.

Ans:

Q.16.For each country, year, and sex compute the total number of cases of TB. Make an informative visualization of the data.

Ans: