

# Tidyverse Problem Set

*Dae Hyun Lee*

*October 4, 2019*

The purpose of this problem set is to provide data contexts in which to exercise the capabilities of the tidyverse. While some questions require specific answers, other parts of the problems have been written to be purposely ambiguous, requiring you to think through the presentation details of your answer.

## HOLD THE PRESSES!

As I was preparing to post these problems yesterday, I noticed that tidyr had been updated in the last few weeks. I was looking for more exercises on `gather()` and `spread()` – which are always difficult to master. And I found that they have been superseded!! Why do I love working with R as the tidyverse is on a path of continuous improvement? Because the improvements come from developers who write things like this:

*For some time, it's been obvious that there is something fundamentally wrong with the design of `spread()` and `gather()`. Many people don't find the names intuitive and find it hard to remember which direction corresponds to spreading and which to gathering. It also seems surprisingly hard to remember the arguments to these functions, meaning that many people (including me!) have to consult the documentation every time. [Hadley Wickham, Pivot Vignette](#)*

So... before you do anymore tidyverse exercises, Read this [tidyr 1.0.0](#).

Then go to the [tidyr cran page](#) and to the examples and exercises in the new vignettes.

In your solutions to the problems below, if you need to use table reshaping functions from TidyR, be sure that you use `pivot_longer()`, and `pivot_wider()`.

## Problem 1

Load the gapminder data from the gapminder package.

```
library(gapminder)
data(gapminder)
```

How many continents are included in the data set?

```
levels(gapminder$continent)
```

```
## [1] "Africa" "Americas" "Asia" "Europe" "Oceania"
```

*There are total 5 continents in the data set.*

How many countrys are included? How many countries per continent?

```
head(fct_unique(gapminder$country))
```

```
## [1] Afghanistan Albania Algeria Angola Argentina Australia
## 142 Levels: Afghanistan Albania Algeria Angola Argentina ... Zimbabwe
```

*There are total 142 countrys are included in the data set.*

Using the gapminder data, produce a report showing the continents in the dataset, total population per continent, and GDP per capita. Be sure that the table is properly labeled and suitable for inclusion in a printed report.

```
levels(gapminder$continent)
```

```
## [1] "Africa" "Americas" "Asia" "Europe" "Oceania"
```

*saf*

Produce a well-labeled table that summarizes GDP per capita for the countries in each continent, contrasting the years 1952 and 2007. *wdw*

Product a plot that summarizes the same data as the table. There should be two plots per continent. *wdwd*

Which countries in the dataset have had periods of negative population growth? *oo*

Illustrate your answer with a table or plot. *oo*

Which countries in the dataset have had the highest rate of growth in per capita GDP? *k*

Illustrate your answer with a table or plot. *j*

## Problem 2

The data for Problem 2 is the Fertility data in the AER package. This data is from the 1980 US Census and is comprised of data on married women aged 21-35 with two or more children. The data report the gender of each woman's first and second child, the woman's race, age, number of weeks worked in 1979, and whether the woman had more than two children.

```
library(AER)

## Loading required package: car
## Loading required package: carData
##
## Attaching package: 'car'
## The following object is masked from 'package:dplyr':
##
##     recode
## The following object is masked from 'package:purrr':
##
##     some
## Loading required package: lmtest
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
## Loading required package: sandwich
## Loading required package: survival
data("Fertility")
data(Fertility)
##glimpse(Fertility)
##Fertility %>% slice(30:40)
Fertility <- as_tibble(Fertility)
##Fertility
```

There are four possible gender combinations for the first two Children. Produce a plot the contrasts the frequency of these four combinations. Are the frequencies different for women in their 20s and women who are older than 29?

Produce a plot that contrasts the frequency of having more than two children by race and ethnicity.

## Problem 3

Use the mtcars and mpg datasets.

How many times does the letter “e” occur in mtcars rownames?

How many cars in mtcars have the brand Merc?

How many cars in mpg have the brand(“manufacturer” in mpg) Merc?

Contrast the mileage data for Merc cars as reported in mtcars and mpg. Use tables, plots, and a short explanation.

## Problem 4

Install the babynames package.

```
library(babynames)
data(babynames)
babyn <- as_tibble(babynames)
babyn
```

```
## # A tibble: 1,924,665 x 5
##   year sex  name      n  prop
##   <dbl> <chr> <chr>   <int> <dbl>
## 1  1880 F    Mary     7065 0.0724
## 2  1880 F    Anna     2604 0.0267
## 3  1880 F    Emma     2003 0.0205
## 4  1880 F  Elizabeth  1939 0.0199
## 5  1880 F   Minnie   1746 0.0179
## 6  1880 F  Margaret  1578 0.0162
## 7  1880 F    Ida     1472 0.0151
## 8  1880 F   Alice   1414 0.0145
## 9  1880 F   Bertha   1320 0.0135
##10  1880 F    Sarah   1288 0.0132
## # ... with 1,924,655 more rows
```

Draw a sample of 500,000 rows from the babynames data

```
##hist(babyn$year)
s <- sample(x = 1:1924665, size = 500000, replace = FALSE)
babyn_500000 <- babyn %>% slice(s)
babyn_500000
```

```
## # A tibble: 500,000 x 5
##   year sex  name      n  prop
##   <dbl> <chr> <chr>   <int> <dbl>
## 1  1937 F  Marillyn    10 0.00000908
## 2  1909 F   Jerry     16 0.0000435
## 3  1968 F  Loredana    5 0.00000292
## 4  1995 M   Latique    6 0.00000298
## 5  1958 F  Calleen    5 0.00000242
## 6  1985 F   Jaren     10 0.00000542
## 7  1994 F  Migdalia   22 0.0000113
## 8  1982 F   Lirio      6 0.00000331
## 9  1988 F   Xylina     21 0.0000109
##10  1966 F   Moria     10 0.00000570
## # ... with 499,990 more rows
```

Produce a tabble that displays the five most popular boy names and girl names in the years 1880,1920, 1960, 2000.

What names overlap boys and girls?

What names were used in the 19th century but have not been used in the 21st century?

Produce a chart that shows the relative frequency of the names “Donald”, “Hilary”, “Hillary”, “Joe”, “Barrack”, over the years 1880 through 2017.