# Tidyverse Problem Set

*MA615*

*September 29, 2019*

The purpose of this problem set is to provide data contexts in which to exercise the capabilitiues of the tidyverse. While some questons 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 updata 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 And I found that they have been superceded!! Why do I love working with R as the tidyversie 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 Vingette

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 exercies 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.

How many continents are included in the data set?

```
## [1] 5
```

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

```
## [1] 142
```

```
## # A tibble: 5 x 2
##   continent `n_distinct(country)`
##   <fct>                     <int>
## 1 Africa                       52
## 2 Americas                     25
## 3 Asia                         33
## 4 Europe                       30
## 5 Oceania                       2
```

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.

Since simply add up all GDP per capita is meaningless, I first add an addition column "total GDP" to the table which mutiply gdpPercap and pop , then make the aggregate, and finally devided by total population.

| Continent | Total Population | total GDP | GDP per Capita |
|---|---|---|---|
| Africa | 929539692 | 2.380486e+12 | 2560.93 |

| Continent | Total Population | total GDP | GDP per Capita |
|---|---|---|---|
| Americas | 898871184 | 1.941809e+13 | 21602.75 |
| Asia | 3811953827 | 2.070795e+13 | 5432.37 |
| Europe | 586098529 | 1.479550e+13 | 25244.05 |
| Oceania | 24549947 | 8.073141e+11 | 32884.56 |

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

I selected some countries for the table since the original one is way too long, see gdpPercapTable fot the full information.

| Country | Continent | 1952 GDP | 2007 GDP |
|---|---|---|---|
| Algeria | Africa | 2449.0 | 6223.4 |
| Angola | Africa | 3520.6 | 4797.2 |
| Benin | Africa | 1062.8 | 1441.3 |
| Argentina | Americas | 5911.3 | 12779.4 |
| Bolivia | Americas | 2677.3 | 3822.1 |
| Brazil | Americas | 2108.9 | 9065.8 |
| Bahrain | Asia | 9867.1 | 29796.0 |
| Bangladesh | Asia | 684.2 | 1391.3 |
| Cambodia | Asia | 368.5 | 1713.8 |
| Albania | Europe | 1601.1 | 5937.0 |
| Austria | Europe | 6137.1 | 36126.5 |
| Belgium | Europe | 8343.1 | 33692.6 |
| Australia | Oceania | 10039.6 | 34435.4 |
| New Zealand | Oceania | 10556.6 | 25185.0 |

Product a plot that summarizes the same data as the table. There should be two plots per continent. long, see gdpPercapTable fot the full information.

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

```
## # A tibble: 27 x 2
##    country                `n_distinct(year)`
##    <fct>                               <int>
##  1 Afghanistan                             1
##  2 Bosnia and Herzegovina                  2
##  3 Bulgaria                                4
##  4 Cambodia                                1
##  5 Croatia                                 1
##  6 Czech Republic                          3
##  7 Equatorial Guinea                       1
##  8 Germany                                 2
##  9 Guinea-Bissau                           1
## 10 Hungary                                 5
## # ... with 17 more rows
```

Illustrate your answer with a table or plot.

| Country | Year | Decrease |
|---|---|---|
| Afghanistan | 1982 | -1998556 |
| Bosnia and Herzegovina | 1992 | -82964 |

| Country | Year | Decrease |
|---|---|---|
| Bosnia and Herzegovina | 1997 | -649013 |
| Bulgaria | 1992 | -313452 |
| Bulgaria | 1997 | -592449 |
| Bulgaria | 2002 | -404258 |
| Bulgaria | 2007 | -338941 |
| Cambodia | 1977 | -471999 |
| Croatia | 1997 | -49418 |
| Czech Republic | 1997 | -14995 |
| Czech Republic | 2002 | -44412 |
| Czech Republic | 2007 | -27551 |
| Equatorial Guinea | 1977 | -84928 |
| Germany | 1977 | -556315 |
| Germany | 1987 | -616968 |
| Guinea-Bissau | 1967 | -26533 |
| Hungary | 1987 | -92795 |
| Hungary | 1992 | -264056 |
| Hungary | 1997 | -104000 |
| Hungary | 2002 | -161371 |
| Hungary | 2007 | -127205 |
| Ireland | 1957 | -73936 |
| Ireland | 1962 | -48220 |
| Kuwait | 1992 | -473392 |
| Lebanon | 1982 | -28911 |
| Lesotho | 2007 | -34123 |
| Liberia | 1992 | -356440 |
| Montenegro | 2007 | -35494 |
| Poland | 2002 | -28981 |
| Poland | 2007 | -107735 |
| Portugal | 1972 | -132550 |
| Romania | 1997 | -234569 |
| Romania | 2002 | -158121 |
| Romania | 2007 | -128281 |
| Rwanda | 1997 | -77620 |
| Serbia | 2002 | -225035 |
| Slovenia | 2002 | -115 |
| Slovenia | 2007 | -2252 |
| Somalia | 1992 | -822059 |
| South Africa | 2007 | -435794 |
| Switzerland | 1977 | -84976 |
| Trinidad and Tobago | 1992 | -7667 |
| Trinidad and Tobago | 1997 | -45568 |
| Trinidad and Tobago | 2002 | -36269 |
| Trinidad and Tobago | 2007 | -45224 |
| West Bank and Gaza | 1972 | -53064 |

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

| country |
|---|
| Equatorial Guinea |
| Taiwan |
| Korea, Rep. |

| country |
|---|
| Singapore |
| Botswana |
| Hong Kong, China |
| China |
| Oman |
| Thailand |
| Japan |

Illustrate your answer with a table or plot.

| Country | GrowthRate in % |
|---|---|
| Equatorial Guinea | 3135.54 |
| Taiwan | 2279.41 |
| Korea, Rep. | 2165.51 |
| Singapore | 1936.30 |
| Botswana | 1376.65 |
| Hong Kong, China | 1200.57 |
| China | 1138.39 |
| Oman | 1120.64 |
| Thailand | 884.22 |
| Japan | 884.04 |

**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 date 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.

There are four possible gender combinations for the first two Children. Product a plot the contracts the frequency of these four combinations. Are the frequencies different for women in their 20s and wemen 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?

```
## [1] 25
```

How many cars in mtcars have the brand Merc?

```
## [1] 7
```

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

```
## [1] 4
```

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

| Model | MPG |
|---|---|
| Merc 240D | 24.4 |
| Merc 230 | 22.8 |
| Merc 280 | 19.2 |
| Merc 280C | 17.8 |
| Merc 450SE | 16.4 |
| Merc 450SL | 17.3 |
| Merc 450SLC | 15.2 |

| Manufacture | Model | year | city MPG | hyw MPG | avg MPG |
|---|---|---|---|---|---|
| mercury | mountaineer 4wd | 1999 | 14 | 17 | 15.5 |
| mercury | mountaineer 4wd | 2008 | 13 | 19 | 16.0 |
| mercury | mountaineer 4wd | 2008 | 13 | 19 | 16.0 |
| mercury | mountaineer 4wd | 1999 | 13 | 17 | 15.0 |

One question about this problem is that: Merc in mpg dataset represent "Mercury", while merc in mtcars dataset represent "Mercedes", so how do we contrast cars from two different manufacturer?

From the table we can see that the average MPG of Mercury mountaineer 4wd 1999 has the lowest mpg amoung all cars, while Mercedes 240D has the highest by 24.4

**Problem 4**

Install the babynames package.

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

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

| year | name | year | name | year | name | year | name |
|---|---|---|---|---|---|---|---|
| 1880 | John | 1920 | Mary | 1960 | David | 2000 | Jacob |
| 1880 | William | 1920 | John | 1960 | Michael | 2000 | Michael |
| 1880 | Mary | 1920 | William | 1960 | James | 2000 | Matthew |
| 1880 | James | 1920 | Robert | 1960 | John | 2000 | Joshua |
| 1880 | Charles | 1920 | James | 1960 | Robert | 2000 | Emily |

What names overlap boys and girls?

```
MaleName = filter(babynames1, sex=='M')
MaleName = unique(MaleName$name)
FemaleName = filter(babynames1, sex == 'F')
FemaleName = unique(FemaleName$name)
overlap = intersect(MaleName, FemaleName)
```

I have found 3072 overlap names in the search, for example: John, William, James, Charles. . . ..

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

```
baby19th = filter(babynames1, year<1900)
baby19th = unique(baby19th$name)
baby21th = filter(babynames3, year>=2000)
```

```
baby21th = unique(baby21th$name)
newNames = setdiff(baby19th, intersect(baby19th, baby21th))
```

Some examples are Myrtle, Nannie, Bertie....

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