Tidyverse Problem Set

Dae Hyun Lee

October 4, 2019

Problem 1

Load the gapminder data from the gapminder package.

How many continents are included in the data set?

```
## [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?

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

There are total 142 courtrys 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.

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

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

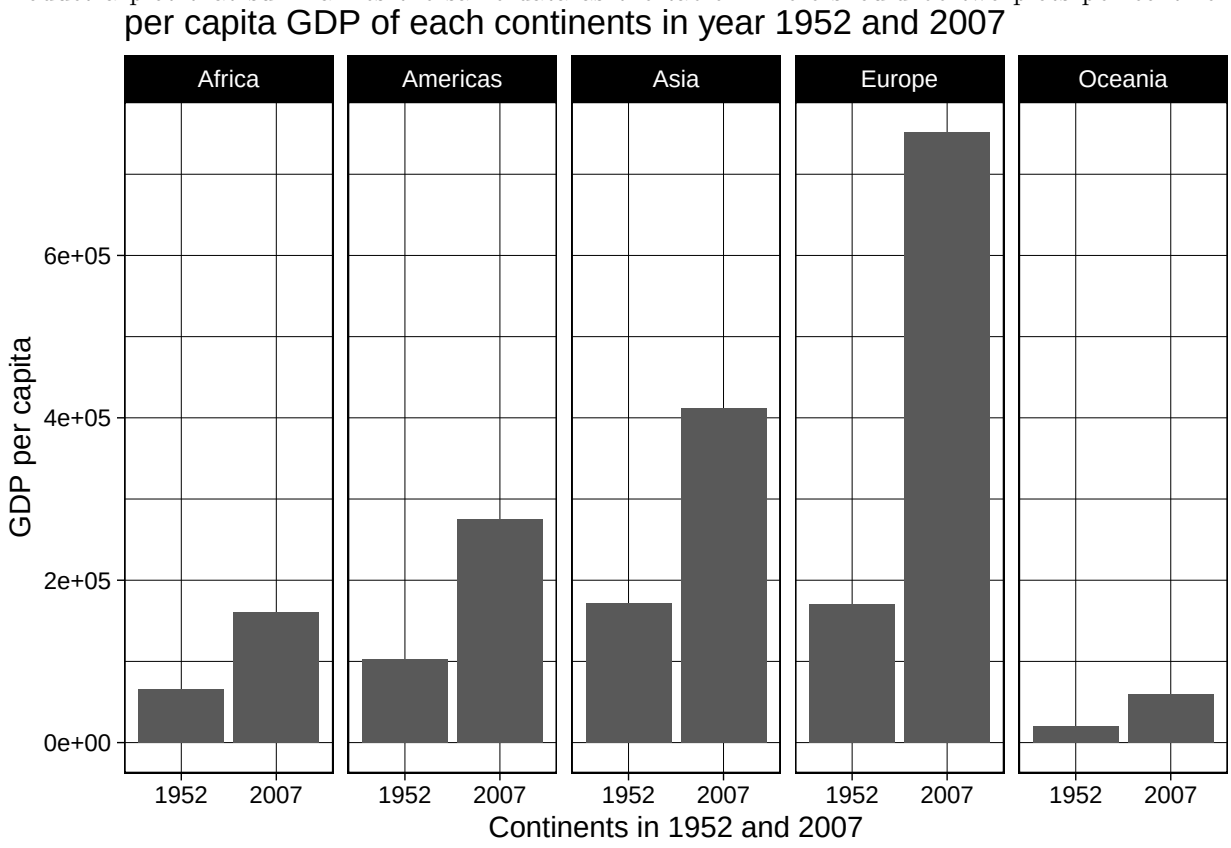


Table 1: Total population and per capita GDP of each continent in 2002

Continent	Total Population	GDP per Capita
Africa	659081517	2071.61
Americas	739274104	16566.92
Asia	3133292191	3234.40
Europe	558142797	18420.19
Oceania	20919651	22593.03

Table 2: GDP per capita by continents in 1952 and in 2007

Continent	GDP per capita in 1952	GDP per capita in 2007
Africa	1311.22	2560.93
Europe	6096.66	25244.05
Asia	806.36	5432.37
Americas	8528.04	21602.75
Oceania	10136.10	32884.56

Which countries in the dataset have had periods of negative population growth? Illustrate your answer with a table or plot.

```
## [1] 109
```

There are total 109 countries which have had periods of negative population growth.

Which countries in the dataset have had the highest rate of growth in per capita GDP? Illustrate your answer with a table or plot.

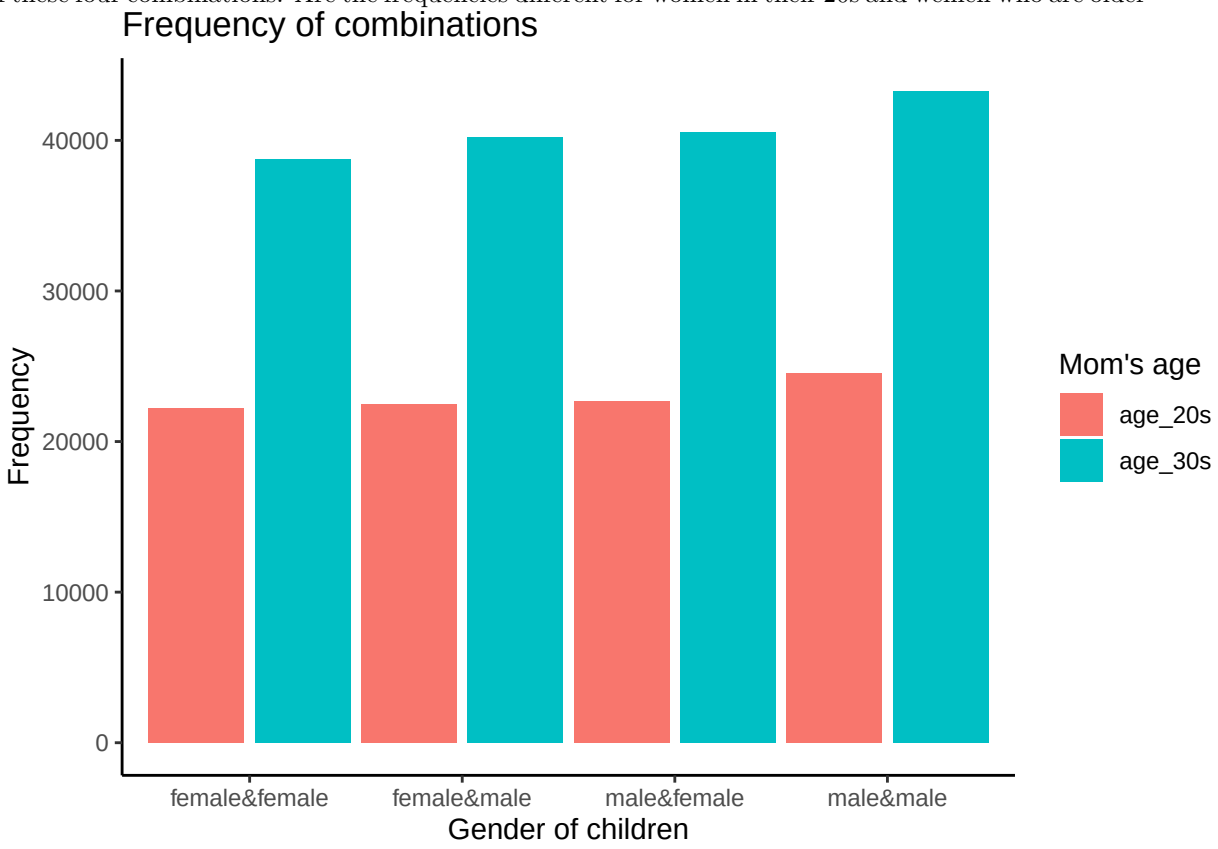
```
## # A tibble: 10 x 7
## # Groups:   country [7]
##    country      continent  year lifeExp      pop gdpPercap gdpChange
##    <fct>        <fct>     <int>   <dbl>    <int>     <dbl>     <dbl>
##  1 Kuwait       Asia       1977    69.3  1140357    59265.    50082.
##  2 Kuwait       Asia       1982    71.3  1497494    31354.    27911.
##  3 Kuwait       Asia       1962    60.5   358266    95458.    18065.
##  4 Kuwait       Asia       1967    64.6   575003    80895.    14563.
##  5 Saudi Arabia Asia       1987    66.3 14619745    21198.    12495.
##  6 Iraq         Asia       1992    59.5 17861905     3746.     7898.
##  7 Gabon        Africa     1982    56.6   753874    15113.     6632.
##  8 Serbia       Europe     1992    71.7  9826397     9325.     6546.
##  9 Libya        Africa     1987    66.2  3799845    11771.     5594.
## 10 Croatia      Europe     1992    72.5  4494013     8448.     5375.
```

Kuwait have had the highest rate of growth in per capita GDP in 1977.

Problem 2

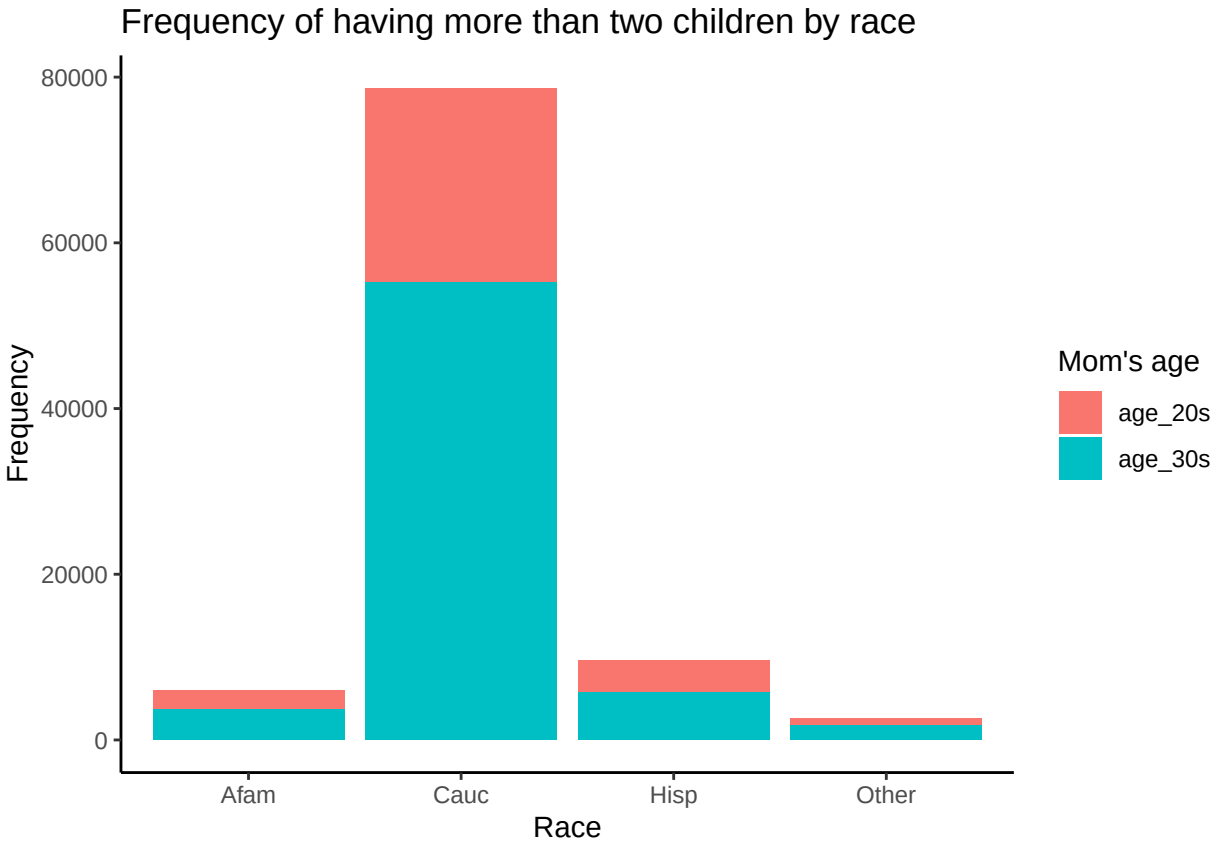
The data for Problem 2 is the Fertility data in the AER package. This data is from the 1980 US Census and is comprised of date on married women aged 21-35 with two or more children. The data report the gender of each woman's first and second child, the woman's race, age, number of weeks worked in 1979,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



It is clear that there are differences in frequency between two age groups.

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
```

The letter "e" occurs 25 times in mtcars rownames

How many cars in mtcars have the brand Merc?

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

7 cars in mtcars have the brand Merc

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

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

There are no such cars that uses the brand Merc in mpg. Instead, there are 4 cars that have brand name mercury

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

Table 3: Mileage data for Merc in mpg

Manufacturer	Model	City miles per gallown	Highways miles per gallon
mercury	mountaineer 4wd	14	17
mercury	mountaineer 4wd	13	19
mercury	mountaineer 4wd	13	19
mercury	mountaineer 4wd	13	17

Table 4: Mileage data for Merc in mtcars

Model	Miles/(US)gallon
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

Problem 4

Install the babynames package.

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

```
## # A tibble: 500,000 x 5
##     year sex   name         n       prop
##    <dbl> <chr> <chr>    <int>      <dbl>
##  1  2000 F     Janissa     22 0.0000110
##  2  2017 M     Izriel       5 0.00000255
##  3  2014 F     Janielys     7 0.00000359
##  4  1983 F     Khadija     27 0.0000151
##  5  1890 M     Coy          8 0.0000668
##  6  1904 M     Edward    2334 0.0169
##  7  1999 F     Berkley     38 0.0000195
##  8  1980 F     Kalie       14 0.00000786
##  9  2004 M     Ashdon       8 0.00000379
## 10  1977 M     Romaldo      6 0.00000351
## # ... with 499,990 more rows
```

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

Table 5: The five most popular boy names and girl names

Rank	1880	1920	1960	2000
#1	John	Mary	David	Jacob
#2	William	John	Michael	Michael
#3	Mary	William	James	Matthew
#4	James	Robert	John	Joshua
#5	Charles	James	Robert	Emily

What names overlap boys and girls?

```
## [1] 10663
```

There are total 10663 overlapping names between boys and girls

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

```
## # A tibble: 6 x 6
##    year sex   name       n     prop only
##   <dbl> <chr> <chr>  <int>    <dbl> <lgl>
## 1  1880 F     Nannie   248 0.00254  TRUE
## 2  1880 F     Hulda     60 0.000615 TRUE
## 3  1880 F     Gussie    54 0.000553 TRUE
## 4  1880 F     Mittie    53 0.000543 TRUE
## 5  1880 F     Myrtie    45 0.000461 TRUE
## 6  1880 F     Virgie    41 0.000420 TRUE
```

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

```
## # A tibble: 4 x 4
##   Name          x   prop proportion
##   <chr>     <int>  <dbl>      <dbl>
## 1 Donald  181369 0.714       71.4
## 2 Hilary    4425 0.0174       1.74
## 3 Hillary   3356 0.0132       1.32
## 4 Joe      65027 0.256       25.6
```