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 And I found that they have been superceded!!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

1.1 Load the gapminder data from the gapminder package.

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

1.2 How many continents are included in the data set?

```
continents_num <- length(unique(gapminder$continent))
paste("Total",continents_num,"continents are included")
```

```
## [1] "Total 5 continents are included"
```

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

```
## count num of countries
country_num <- length(unique(gapminder$country))
paste("Total",country_num,"countries are included")
```

```
## [1] "Total 142 countries are included"
```

```
## summarize num of counries in each countinent
kable(gapminder %>% group_by(continent) %>% summarize(num_countries = n_distinct(country)),align ="c",b
kable_styling(latex_options = 'hold_position',font_size = 12,full_width = F,position = "center")
```

```
paste("Africa : 52, Americas: 25, Asia:33, Oceania: 2")
```

```
## [1] "Africa : 52, Americas: 25, Asia:33, Oceania: 2"
```

continent	num_countries
Africa	52
Americas	25
Asia	33
Europe	30
Oceania	2

1.4 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.

```
report <- gapminder %>% group_by(continent)%>% summarize(TTL_Avg_pop = sum(as.numeric(pop))/length(uniqu

kable(report,digits = 2,booktabs=TRUE,align = 'c') %>%
  kable_styling(latex_options = 'hold_position',font_size = 12,full_width = F)
```

continent	TTL_Avg_pop	TTL_gdpPercap
Africa	515632163	2108.19
Americas	612619875	15477.07
Asia	2542277825	2949.60
Europe	515092942	15692.82
Oceania	17749345	21204.96

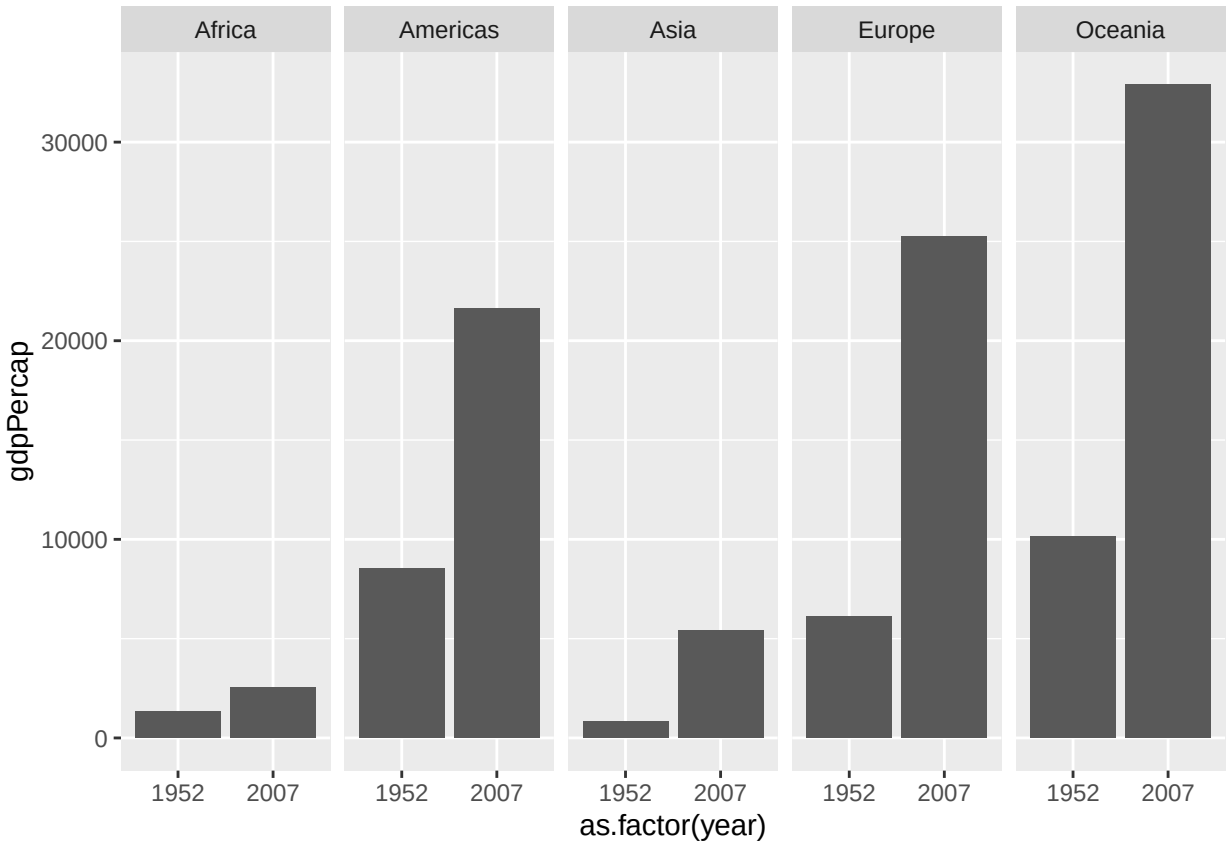
1.5 Produce a well-labeled table that summarizes GDP per capita for the countries in each continent, contrasting the years 1952 and 2007.

```
df1 <- gapminder %>% select(continent,country,year,gdpPercap) %>% filter (year %in% c(1952,2007)) %>% sp
colnames(df1) <- c("continent","country","year_1952","year_2007")
df1 <- df1 %>% arrange(continent,desc(year_2007))
kable(df1,digits =2,booktabs=TRUE,caption="GDP per capita for the countries of each continent in year 19
  kable_styling(latex_options = 'hold_position',font_size =12,full_width = F)
```

```
#kable(df1,digits =2,booktabs=TRUE,caption="GDP per capita for the countries of each continent in year
#kable_styling(latex_options = 'hold_position',font_size =12,full_width = F)
```

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

```
gapminder %>%
  filter(year %in% c(1952, 2007)) %>% group_by(continent,year)%>% summarize(gdpPercap = sum(gdpPercap*as
  ggplot()+
  geom_bar(mapping=aes(x=as.factor(year),y=gdpPercap),stat="identity") +
  facet_grid(.~continent)
```



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

```
df2 <- gapminder %>% select(continent,country,year,pop) %>% filter (year %in% c(1952,2007)) %>% spread(y
colnames(df2) <- c("continent","country","year_1952","year_2007")
filter (df2,year_2007<year_1952)
```

```
## # A tibble: 0 x 4
## # ... with 4 variables: continent <fct>, country <fct>, year_1952 <int>,
## #   year_2007 <int>
```

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

```
df3 <- df1 %>% mutate(growth_rate = (year_2007-year_1952)/year_1952) %>% arrange(desc(growth_rate))
head(df3,1)
```

```
## # A tibble: 1 x 5
##   continent country           year_1952 year_2007 growth_rate
##   <fct>     <fct>                 <dbl>     <dbl>       <dbl>
## 1 Africa    Equatorial Guinea      376.    12154.        31.4
```

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.

```
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
```

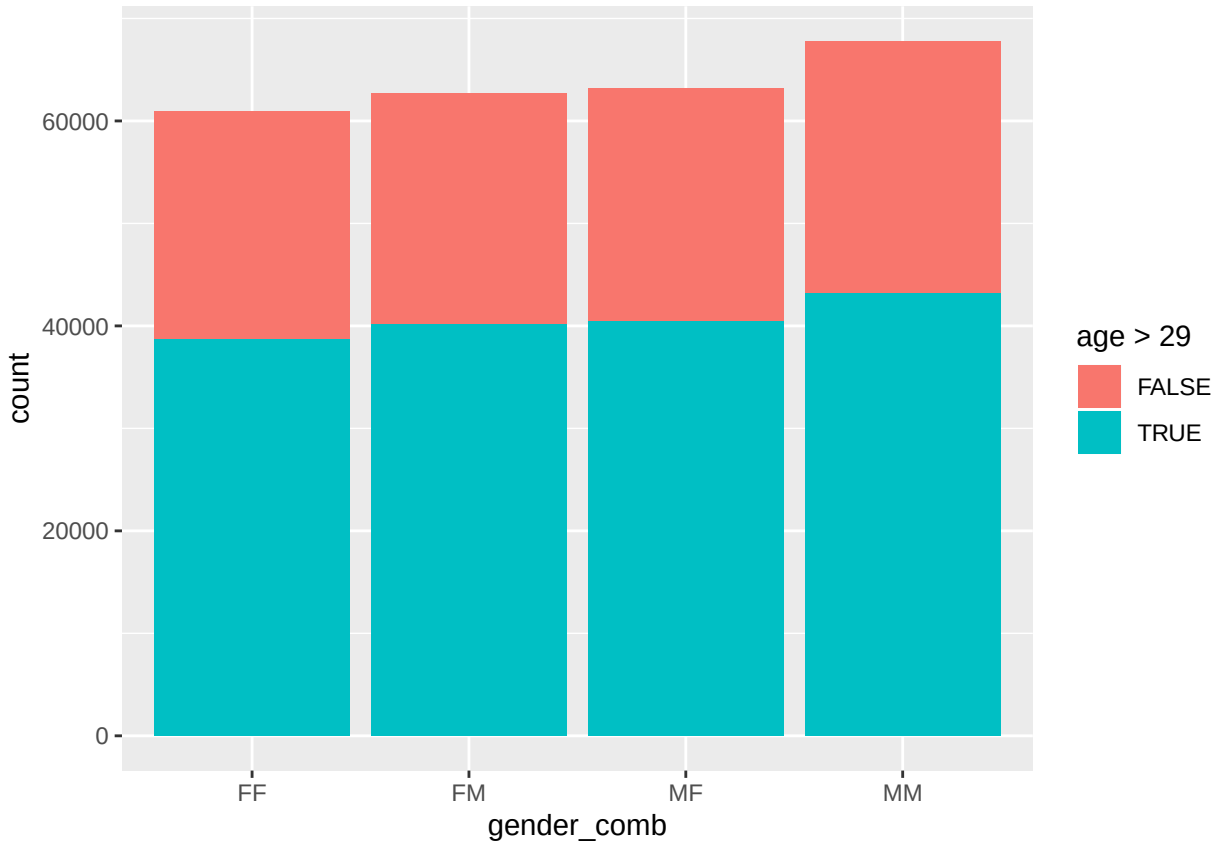
```
## Loading required package: survival
```

```
data(Fertility)
```

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?

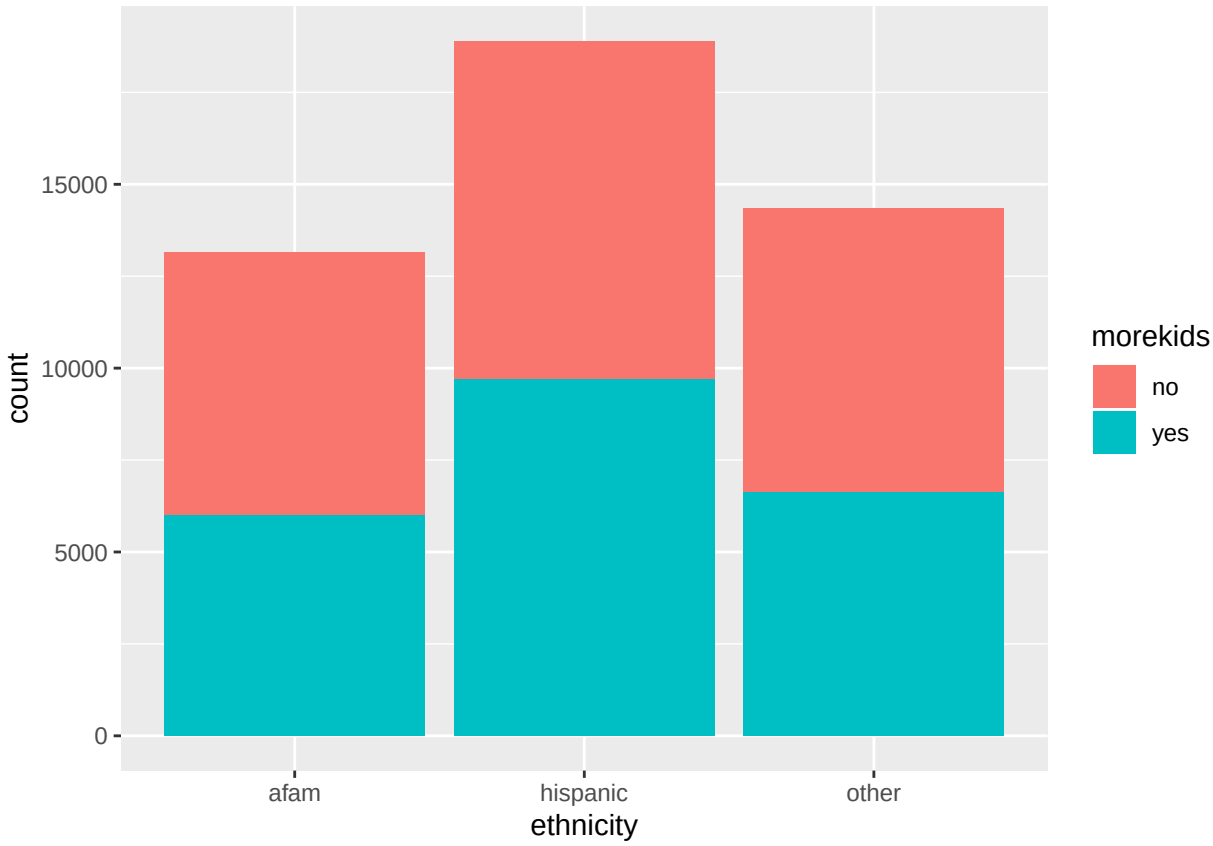
```
Fertility <- as_tibble(Fertility)
gender_comb <- Fertility %>% mutate(gender_comb = ifelse(gender1 == "male" & gender2 == "male", "MM", i

gender_comb %>%
    ggplot()+ geom_bar(mapping=aes(x=gender_comb,fill=age >29),stat="count")
```



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

```
f_race_only_three <- Fertility %>% gather(`afam`,`hispanic`,`other`, key = ethnicity, value = yes )%>%
  filter(yes == "yes")
ggplot(data = f_race_only_three)+
  geom_bar(mapping =aes(x=ethnicity,fill = morekids))
```



Problem 3

Use the mtcars and mpg datasets.

```
data(mpg)
data(mtcars)
```

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

```
mtc <- as_tibble(rownames_to_column(mtcars, var = "Model"))
mtc$number.of.e <- str_count(mtc$Model, "e")
sum(mtc$number.of.e)
```

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

How many cars in mtcars have the brand Merc?

```
sum(str_count(mtc$Model,"Merc"))
```

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

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

```
sum(str_count(mpg$manufacturer,"mercury"))
```

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

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

Problem 4

Install the babynames package.

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

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

```
s <- sample(x = 1:nrow(babynames), size = 500000, replace = FALSE)
sample_babynames <- babynames[s,]
head(sample_babynames,10)
```

```
## # A tibble: 10 x 5
##     year sex   name        n       prop
##    <dbl> <chr> <chr>   <int>      <dbl>
##  1  1974 F     Prima      11 0.00000702
##  2  1990 F     Kelsye     10 0.00000487
##  3  1955 F     Cindra     33 0.0000165
##  4  2005 M     Carey      27 0.0000127
##  5  1986 M     Vannak      9 0.00000469
##  6  1910 F     Luda        5 0.0000119
##  7  1941 M     Gerson      8 0.00000638
##  8  1991 M     Neeraj      7 0.0000033
##  9  1935 M     Quinten    11 0.0000103
## 10  2013 M     Lochlin     7 0.00000347
```

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?

```
boys <- babyn %>% filter(sex == "M")
girls <- babyn %>% filter(sex == "F")
overlap <- intersect(boys$name,girls$name)
```

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

```
nineteenth <- babyn %>% filter(year >= 1800 & year <= 1899)
twentyth <- babyn %>% filter(year >= 2000 & year <= 2017)
#count(!(twentyth$name %in% nineteenth))
```

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

Table 1: GDP per capita for the countries of each continent in year 1952 vs 2007

continent	country	year_1952	year_2007
Africa	Gabon	4293.48	13206.48
Africa	Botswana	851.24	12569.85
Africa	Equatorial Guinea	375.64	12154.09
Africa	Libya	2387.55	12057.50
Africa	Mauritius	1967.96	10956.99
Africa	South Africa	4725.30	9269.66
Africa	Reunion	2718.89	7670.12
Africa	Tunisia	1468.48	7092.92
Africa	Algeria	2449.01	6223.37
Africa	Egypt	1418.82	5581.18
Africa	Namibia	2423.78	4811.06
Africa	Angola	3520.61	4797.23
Africa	Swaziland	1148.38	4513.48
Africa	Morocco	1688.20	3820.18
Africa	Congo, Rep.	2125.62	3632.56
Africa	Sudan	1615.99	2602.39
Africa	Djibouti	2669.53	2082.48
Africa	Cameroon	1172.67	2042.10
Africa	Nigeria	1077.28	2013.98
Africa	Mauritania	743.12	1803.15
Africa	Senegal	1450.36	1712.47
Africa	Chad	1178.67	1704.06
Africa	Sao Tome and Principe	879.58	1598.44
Africa	Lesotho	298.85	1569.33
Africa	Cote d'Ivoire	1388.59	1544.75
Africa	Kenya	853.54	1463.25
Africa	Benin	1062.75	1441.28
Africa	Ghana	911.30	1327.61
Africa	Zambia	1147.39	1271.21
Africa	Burkina Faso	543.26	1217.03
Africa	Tanzania	716.65	1107.48
Africa	Uganda	734.75	1056.38
Africa	Madagascar	1443.01	1044.77
Africa	Mali	452.34	1042.58
Africa	Comoros	1102.99	986.15
Africa	Guinea	510.20	942.65
Africa	Somalia	1135.75	926.14
Africa	Togo	859.81	882.97
Africa	Rwanda	493.32	863.09
Africa	Sierra Leone	879.79	862.54
Africa	Mozambique	468.53	823.69
Africa	Malawi	369.17	759.35
Africa	Gambia	485.23	752.75
Africa	Central African Republic	1071.31	706.02
Africa	Ethiopia	362.15	690.81