Tidyverse Problem Set

MA615 Xinci Chen

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 And I found that they have been superceded!!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.

```
library(gapminder)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------ -
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
data("gapminder")
gapminder1 <- gapminder
```

How many continents are included in the data set?

```
str(gapminder1$continent)
```

```
##  Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
```

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

```
str(gapminder1$country)
```

```
##  Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
gapminder1 %>%
  group_by(continent) %>%
  summarize(n = n(),
            n_countries = n_distinct(country))
```

```
## # A tibble: 5 x 3
##   continent     n n_countries
##   <fct>     <int>       <int>
## 1 Africa      624          52
## 2 Americas    300          25
## 3 Asia        396          33
## 4 Europe      360          30
## 5 Oceania      24           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.

```
gapminder1 %>%
  group_by(continent) %>%
  summarize(total_pop=sum(as.numeric(pop)),total_gdp=sum(gdpPercap))
```

```
## # A tibble: 5 x 3
##   continent   total_pop total_gdp
##   <fct>           <dbl>     <dbl>
## 1 Africa     6187585961  1368903.
## 2 Americas   7351438499  2140833.
## 3 Asia      30507333901  3129252.
## 4 Europe     6181115304  5209011.
## 5 Oceania     212992136   446919.
```

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

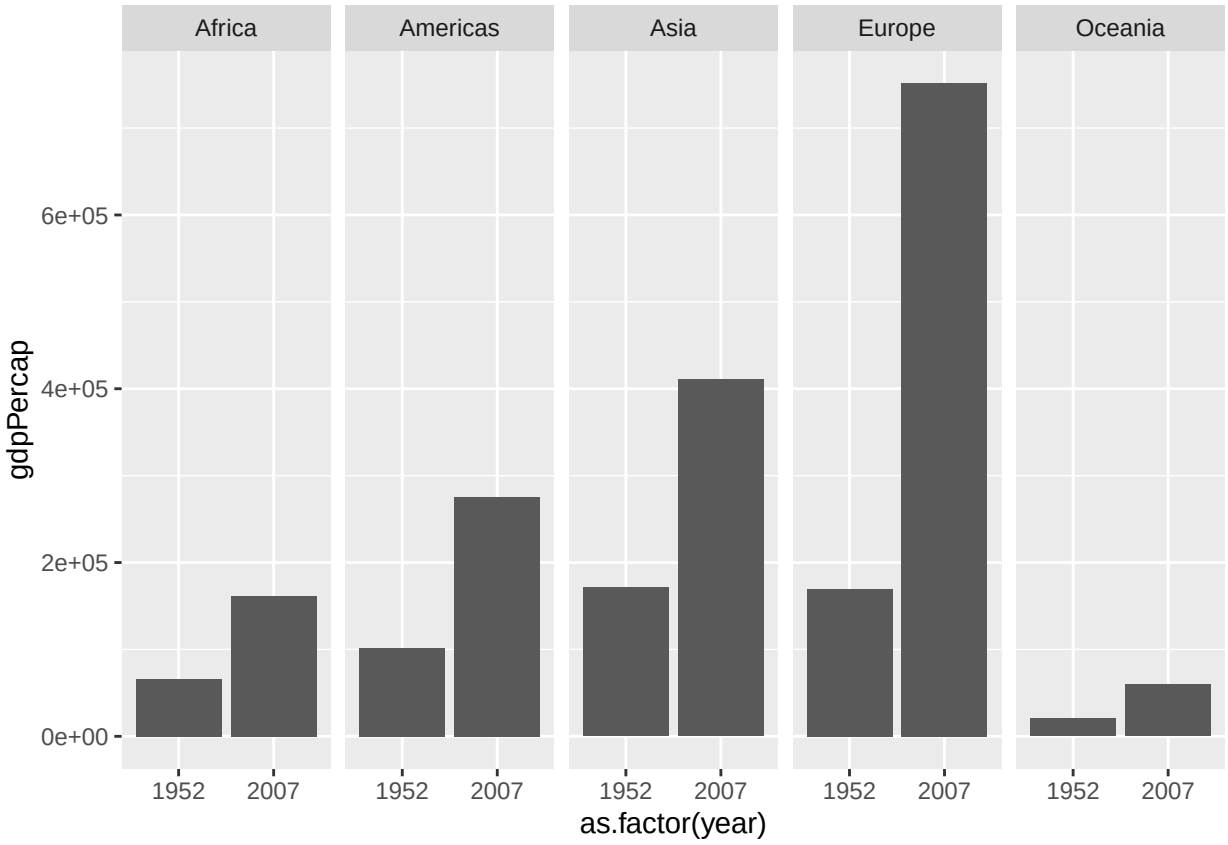
```
gapminder1 %>%
  filter(year %in% c(1952, 2007)) %>%
  group_by(continent,year) %>%
  summarize(total_gdp=sum(gdpPercap))
```

```
## # A tibble: 10 x 3
## # Groups:   continent [5]
##    continent  year total_gdp
##    <fct>     <int>     <dbl>
##  1 Africa     1952    65134.
```

```
##  2 Africa     2007   160630.
##  3 Americas   1952   101977.
##  4 Americas   2007   275076.
##  5 Asia       1952   171451.
##  6 Asia       2007   411610.
##  7 Europe     1952   169832.
##  8 Europe     2007   751634.
##  9 Oceania    1952    20596.
## 10 Oceania    2007    59620.
```

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

```
gapminder1 %>%
  filter(year %in% c(1952, 2007)) %>%
  ggplot()+
  geom_bar(mapping=aes(x=as.factor(year),y=gdpPercap),stat="identity")+
  facet_grid(.~continent)
```



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

Illustrate your answer with a table or plot.

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

Illustrate your answer with a table or plot.

```
distinct(gapminder,year)
```

```
## # A tibble: 12 x 1
##     year
##    <int>
```

Table 1: Top 10 countries with the highest population growth rate from 1952 to 2007

country	1952	2007	growth_rate
Equatorial Guinea	375.6	12154.1	31.4
Taiwan	1206.9	28718.3	22.8
Korea, Rep.	1030.6	23348.1	21.7
Singapore	2315.1	47143.2	19.4
Botswana	851.2	12569.9	13.8
Hong Kong, China	3054.4	39725.0	12.0
China	400.4	4959.1	11.4
Oman	1828.2	22316.2	11.2
Thailand	757.8	7458.4	8.8

```
##  1  1952
##  2  1957
##  3  1962
##  4  1967
##  5  1972
##  6  1977
##  7  1982
##  8  1987
##  9  1992
## 10  1997
## 11  2002
## 12  2007
```

```
# the first year is 1952, the last year is 2007
p1 <- gapminder%>%
  select(country,year,gdpPercap)%>%
  filter(year %in% c(1952,2007)) %>%
  spread(year,gdpPercap)%>%
  mutate(growth_rate = `2007`/`1952`-1)%>%
  filter(rank(desc(growth_rate)) < 10)%>%
  arrange(desc(growth_rate))

kable(p1, digits = 1, caption = "Top 10 countries with the highest population growth rate from 1952 to 1
  kable_styling()
```

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(tidyr)
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
```

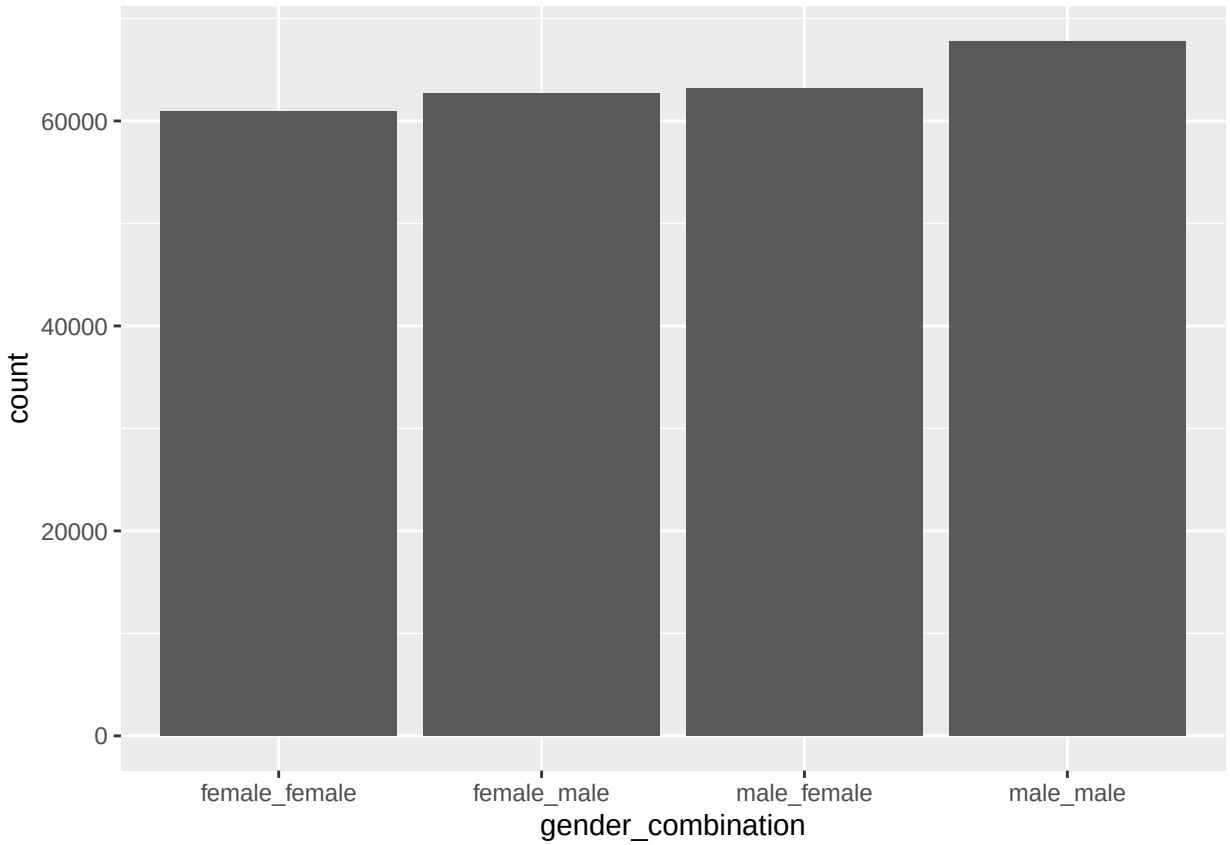
```
## Loading required package: survival
```

```
data(Fertility)
```

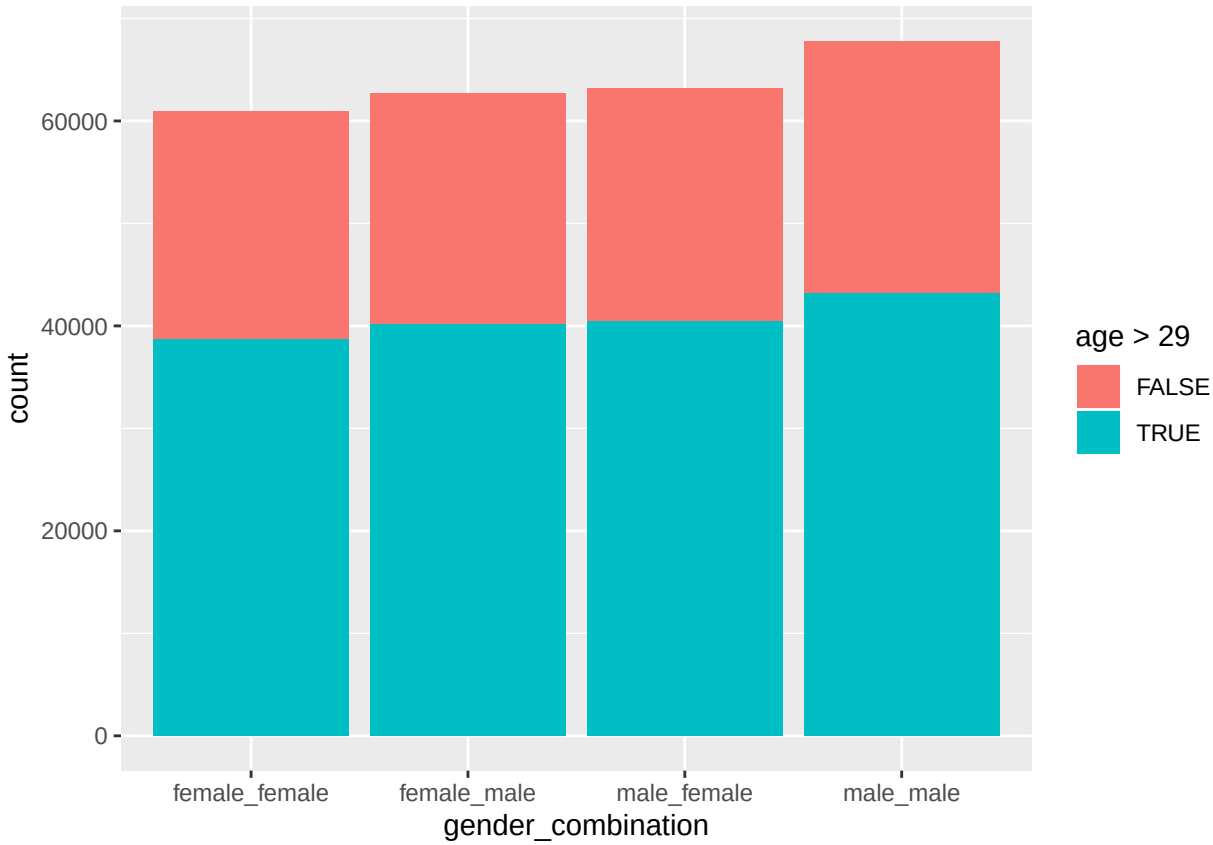
There are four possible gender combinations for the first two Children. Product a plot the contracts the frequency of these four combinations. Are the frequencies different for women in their 20s and wemen who are older than 29?

```
Fertility1 <- Fertility %>%
  unite(gender_combination, gender1, gender2) %>%
  select(gender_combination,age) %>%
  arrange(gender_combination)

#Plot that contracts the frequency of 4 combinations:
ggplot(data=Fertility1, aes(x=gender_combination)) +
  geom_bar()
```




```
#Plot that contracts the frequency of 4 combinations with difference age period:
ggplot(data=Fertility1, aes(x=gender_combination, fill=age>29)) +
  geom_bar()
```



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?

```
data(mtcars)
data(mpg)
mtcars2 <- tibble::rownames_to_column(mtcars, "Car Name")
number_e <- str_count(mtcars2$`Car Name`,"e")
sum(number_e)
```

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

How many cars in mtcars have the brand Merc?

```
number_Merc <- str_count(mtcars2$`Car Name`,"Merc")
sum(number_Merc)
```

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

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

```
number_Merc_mpg <- str_count(mpg$manufacturer,"merc")
sum(number_Merc_mpg)
```

```
## [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.

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

```
library(babynames)
library(dplyr)
babynames5000<-sample_n(babynames,500000)
```

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

```
babynames1880<-filter(babynames,year==1880)
babynames1880count<-babynames1880%>%group_by(name)%>%summarise(sum(n))
babynames1880count<-babynames1880count[order(-babynames1880count$`sum(n)`),]
babynames1880top5<-babynames1880count[c(1:5),]
year<-rep(1880,5)
baby1880<-cbind(year,babynames1880top5)

babynames1920<-filter(babynames,year==1920)
babynames1920count<-babynames1920%>%group_by(name)%>%summarise(sum(n))
babynames1920count<-babynames1920count[order(-babynames1920count$`sum(n)`),]
babynames1920top5<-babynames1920count[c(1:5),]
year<-rep(1920,5)
baby1920<-cbind(year,babynames1920top5)

babynames1960<-filter(babynames,year==1960)
babynames1960count<-babynames1960%>%group_by(name)%>%summarise(sum(n))
babynames1960count<-babynames1960count[order(-babynames1960count$`sum(n)`),]
babynames1960top5<-babynames1960count[c(1:5),]
year<-rep(1960,5)
baby1960<-cbind(year,babynames1960top5)

babynames2000<-filter(babynames,year==2000)
babynames2000count<-babynames2000%>%group_by(name)%>%summarise(sum(n))
babynames2000count<-babynames2000count[order(-babynames2000count$`sum(n)`),]
babynames2000top5<-babynames2000count[c(1:5),]
year<-rep(2000,5)
baby2000<-cbind(year,babynames2000top5)

babynames_top5<-rbind(baby1880,baby1920,baby1960,baby2000)
```

What names overlap boys and girls?

```
boys<-filter(babynames,sex=='M')
girls<-filter(babynames,sex=='F')
overlap<-intersect(boys$name,girls$name)
#overlap
```

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

```
name19th<-filter(babynames,year>=1801 & year<=1900)
name21th<-filter(babynames,year>=1990 & year<=1999)
notin21st<-setdiff(name19th$name,name21th$name)
#notin21st
```

Produce a chart that shows the relative frequency of the names “Donald”, “Hilary”, “Hillary”, “Joe”,

"Barrack", over the years 1880 through 2017.

```
babynames1880and2017<-filter(babynames,year>=1880 & year<=2017)
n<-length(babynames$name)
babynames1880and2017<-filter(babynames1880and2017,name=="Donald"|name=="Hilary"|name=="Hillary"|name==
final<-babynames1880and2017%>%group_by(name)%>%summarise(sum(n)/length(babynames$name))
```