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 And I found that they have been superceded!!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?

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

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

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

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

```
length(unique(gapminder$continent))
```

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

```
## 5 continents
length(unique(gapminder$country))
```

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

```
## 142 countries
length(unique(gapminder$country))
```

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

```
gm<-gapminder
gm %>% group_by( continent) %>% summarise(n_country = n_distinct(country))
```

```
## # A tibble: 5 x 2
##   continent n_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.

```
## delete all null rows
gm1 <- drop_na(gm)
## total_pop is the total population per continent
## gdp_per_capita is the GDP per capita
gm1 %>%
  group_by(continent) %>%
  summarise(total_pop = sum(as.numeric(pop)), gdp_per_capita = sum(as.numeric(pop)*gdpPercap)/total_pop
```

```
## # A tibble: 5 x 3
##   continent   total_pop gdp_per_capita
##   <fct>           <dbl>          <dbl>
## 1 Africa     6187585961          2108.
## 2 Americas   7351438499         15477.
## 3 Asia      30507333901          2950.
## 4 Europe     6181115304         15693.
## 5 Oceania     212992136         21205.
```

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

```
gm2 <- gm1 %>%
  select(continent, country, year, gdpPercap)%>%
  group_by(continent, country) %>%
  filter(year ==1952 | year == 2007)%>%
  spread(year, gdpPercap)%>%
  arrange(desc(continent))%>%
  group_by(continent)
gm2
```

```
## # A tibble: 142 x 4
## # Groups:   continent [5]
##    continent country                `1952` `2007`
##    <fct>     <fct>                   <dbl>  <dbl>
##  1 Oceania   Australia              10040. 34435.
##  2 Oceania   New Zealand            10557. 25185.
##  3 Europe    Albania                 1601.  5937.
```

```
##  4 Europe    Austria                 6137. 36126.
##  5 Europe    Belgium                 8343. 33693.
##  6 Europe    Bosnia and Herzegovina   974.  7446.
##  7 Europe    Bulgaria                2444. 10681.
##  8 Europe    Croatia                 3119. 14619.
##  9 Europe    Czech Republic          6876. 22833.
## 10 Europe    Denmark                 9692. 35278.
## # ... with 132 more rows
```

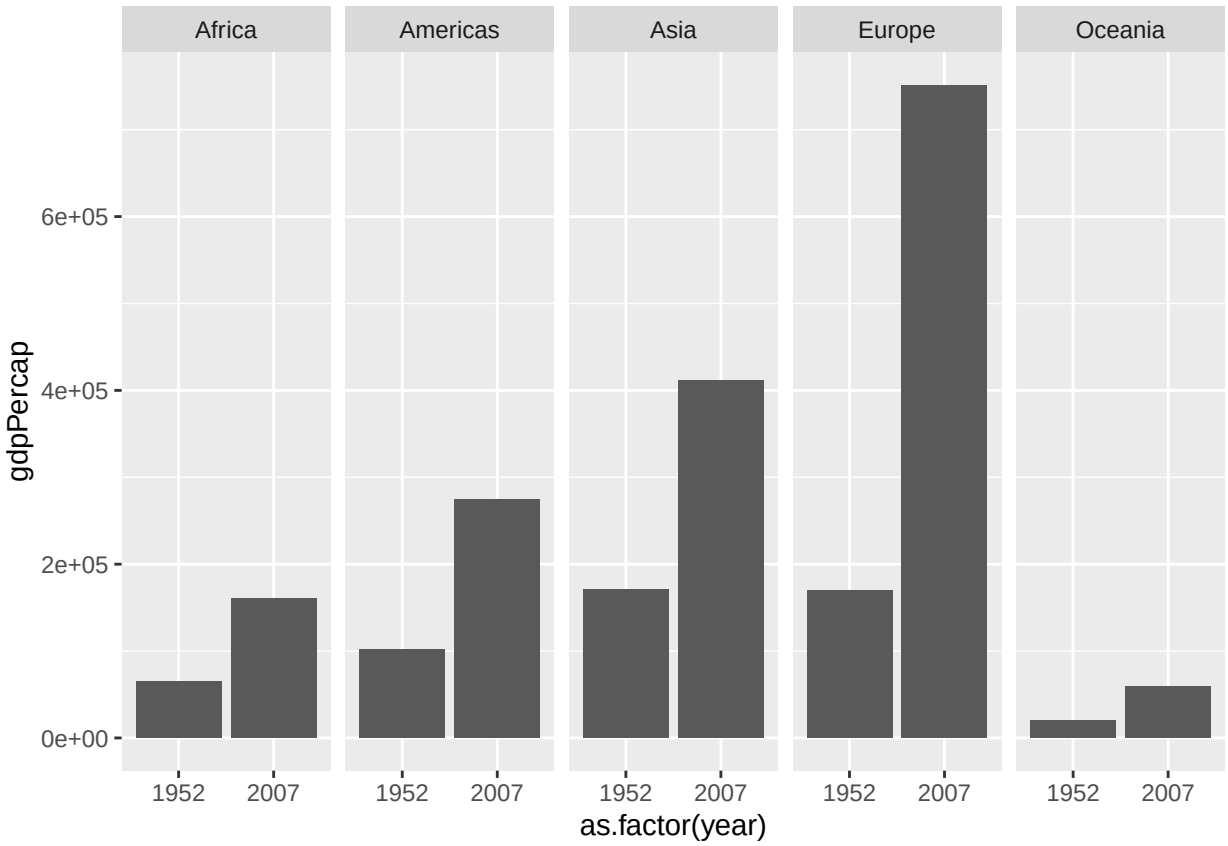
Product a plot that summarizes the same data as the table. There should be two plots per continent.

```
library(ggplot2)

gm3 <- gm1 %>%
  select(continent, year, gdpPercap)%>%
  group_by(continent) %>%
  filter(year ==1952 | year == 2007)
gm3
```

```
## # A tibble: 284 x 3
## # Groups:   continent [5]
##    continent  year gdpPercap
##    <fct>     <int>     <dbl>
##  1 Asia       1952      779.
##  2 Asia       2007      975.
##  3 Europe     1952     1601.
##  4 Europe     2007     5937.
##  5 Africa     1952     2449.
##  6 Africa     2007     6223.
##  7 Africa     1952     3521.
##  8 Africa     2007     4797.
##  9 Americas   1952     5911.
## 10 Americas   2007    12779.
## # ... with 274 more rows
```

```
ggplot(data = gm3) +
  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?

```
dt_2 <- gapminder%>%
  select(country,year,pop)%>%
  spread(year,pop)%>%
  transmute(country,`1957`=`1957`-`1952`, `1962`=`1962`-`1957`,`1967`=`1967`-`1962`,`1972`=`1972`-`1967

negative_growth_1957 <- arrange(dt_2, `1957`)
negative_growth_1962 <- arrange(dt_2, `1962`)
negative_growth_1967 <- arrange(dt_2, `1967`)

dt_2
```

```
## # A tibble: 142 x 12
##    country `1957` `1962` `1967` `1972` `1977` `1982` `1987` `1992` `1997`
##    <fct>    <int>  <int>  <int>  <int>  <int>  <int>  <int>  <int>  <int>
##  1 Afghan~ 8.16e5 9.45e6 2.09e6 1.10e7 3.89e6 1.08e7 3.07e6 1.32e7 8.98e6
##  2 Albania 1.94e5 1.53e6 4.50e5 1.81e6 6.95e5 2.33e6 7.45e5 2.58e6 8.46e5
##  3 Algeria 9.91e5 1.00e7 2.75e6 1.20e7 5.14e6 1.73e7 5.97e6 2.03e7 8.75e6
##  4 Angola  3.29e5 4.50e6 7.51e5 5.14e6 1.02e6 6.27e6 1.61e6 7.13e6 2.75e6
##  5 Argent~ 1.73e6 1.96e7 3.38e6 2.14e7 5.59e6 2.60e7 5.66e6 2.83e7 7.91e6
##  6 Austra~ 1.02e6 9.77e6 2.10e6 1.11e7 3.00e6 1.31e7 3.17e6 1.43e7 4.25e6
##  7 Austria 3.81e4 7.09e6 2.85e5 7.26e6 3.09e5 7.29e6 2.90e5 7.63e6 4.44e5
##  8 Bahrain 1.82e4 1.54e5 4.85e4 1.82e5 1.15e5 3.29e5 1.25e5 4.04e5 1.94e5
##  9 Bangla~ 4.48e6 5.24e7 1.05e7 6.03e7 2.01e7 8.26e7 2.12e7 9.26e7 3.08e7
## 10 Belgium 2.59e5 8.96e6 5.97e5 9.11e6 7.10e5 9.26e6 6.11e5 9.43e6 7.65e5
## # ... with 132 more rows, and 2 more variables: `2002` <int>, `2007` <int>
```

```r
library(knitr)
library(esquisse)
knitr::opts_chunk$set(fig.pos = 'H')
library(kableExtra)
```

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

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

```r
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
library(tidyverse)
library(dplyr)
library(expss)
```

```
##
## Use 'expss_output_viewer()' to display tables in the RStudio Viewer.
##  To return to the console output, use 'expss_output_default()'.
```

```
##
## Attaching package: 'expss'
```

```
## The following objects are masked from 'package:magrittr':
##
##     and, equals, or
```

```
## The following objects are masked from 'package:stringr':
##
##     fixed, regex
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, compute, contains, first, last, na_if, recode, vars
```

```
## The following objects are masked from 'package:purrr':
##
##     keep, modify, modify_if, transpose
```

```
## The following objects are masked from 'package:tidyr':
##
##     contains, nest
```

```
## The following object is masked from 'package:ggplot2':
##
##     vars
```

```
library(tidyr)
  options(tinytex.verbose = TRUE)
  opts_chunk$set(echo = TRUE)


neg_inc = gapminder %>%
  group_by(country) %>%
  summarise(t = sum(diff(pop) > 0), l = length(pop), n = 11 - t) %>%
              filter(t < 11) %>%
              arrange(n)
colnames(neg_inc) = c("Country", "", "", "# of year of negative pop growth")
neg_inc = cbind(neg_inc[1:9, ], neg_inc[10:18, ], neg_inc[19:27, ])
kable(neg_inc[, c(1, 4, 5, 8, 9, 12)], caption = "Countries had periods of negative poppulation growth"
  kable_styling(latex_options = c("HOLD_position")) %>%
  column_spec(c(1, 2, 3, 4, 5, 6), width = "7em")
```

Table 1: Countries had periods of negative poppulation growth

Country	# of year of negative pop growth	Country.1	# of year of negative pop growth.1	Country.2	# of year of negative pop growth.2
Afghanistan	1	Montenegro	1	Germany	2
Cambodia	1	Portugal	1	Ireland	2
Croatia	1	Rwanda	1	Poland	2
Equatorial Guinea	1	Serbia	1	Slovenia	2
Guinea-Bissau	1	Somalia	1	Czech Republic	3
Kuwait	1	South Africa	1	Romania	3
Lebanon	1	Switzerland	1	Bulgaria	4
Lesotho	1	West Bank and Gaza	1	Trinidad and Tobago	4
Liberia	1	Bosnia and Herzegovina	2	Hungary	5

```
neg_inc
```

```
##             Country Var.2 Var.3 # of year of negative pop growth
## 1       Afghanistan    10    12                                1
## 2          Cambodia    10    12                                1
## 3           Croatia    10    12                                1
## 4 Equatorial Guinea    10    12                                1
## 5     Guinea-Bissau    10    12                                1
## 6            Kuwait    10    12                                1
## 7           Lebanon    10    12                                1
## 8           Lesotho    10    12                                1
## 9           Liberia    10    12                                1
##                 Country Var.6 Var.7 # of year of negative pop growth
## 1             Montenegro    10    12                                1
## 2               Portugal    10    12                                1
## 3                 Rwanda    10    12                                1
## 4                 Serbia    10    12                                1
## 5                Somalia    10    12                                1
```

```
## 6           South Africa    10    12                                  1
## 7            Switzerland    10    12                                  1
## 8     West Bank and Gaza    10    12                                  1
## 9 Bosnia and Herzegovina     9    12                                  2
##               Country Var.10 Var.11 # of year of negative pop growth
## 1               Germany      9     12                                2
## 2               Ireland      9     12                                2
## 3                Poland      9     12                                2
## 4              Slovenia      9     12                                2
## 5      Czech Republic      8     12                                3
## 6               Romania      8     12                                3
## 7              Bulgaria      7     12                                4
## 8 Trinidad and Tobago      7     12                                4
## 9               Hungary      6     12                                5
```

```
##....etc
```

Illustrate your answer with a table or plot.

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

```
gm4 <- gm1 %>% 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))
gm4
```

```
## # A tibble: 9 x 4
##   country           `1952` `2007` growth_rate
##   <fct>              <dbl>  <dbl>       <dbl>
## 1 Equatorial Guinea   376. 12154.       31.4
## 2 Taiwan             1207. 28718.       22.8
## 3 Korea, Rep.        1031. 23348.       21.7
## 4 Singapore          2315. 47143.       19.4
## 5 Botswana            851. 12570.       13.8
## 6 Hong Kong, China   3054. 39725.       12.0
## 7 China               400.  4959.       11.4
## 8 Oman               1828. 22316.       11.2
## 9 Thailand            758.  7458.        8.84
```

Illustrate your answer with a table or plot.

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.

```
library(AER)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:expss':
##
##     recode
```

```
## 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'
```

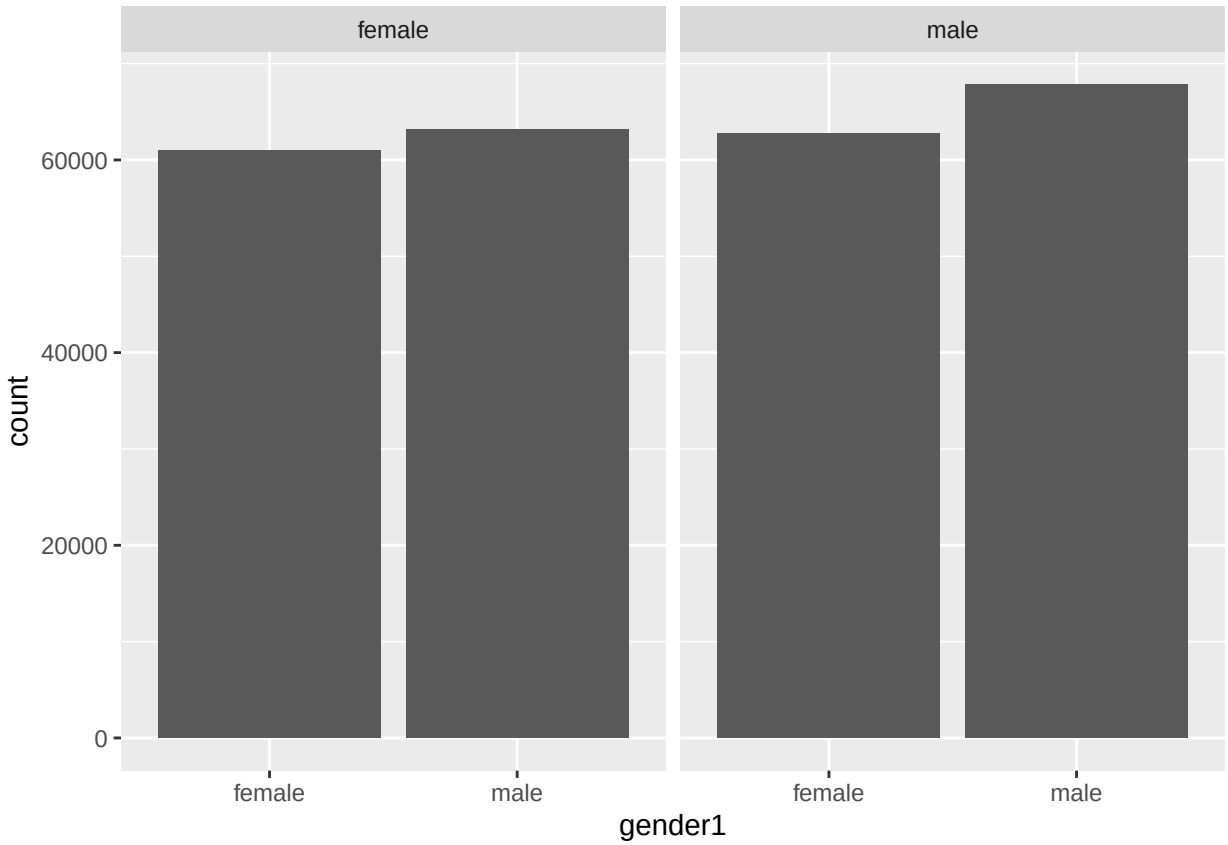
```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

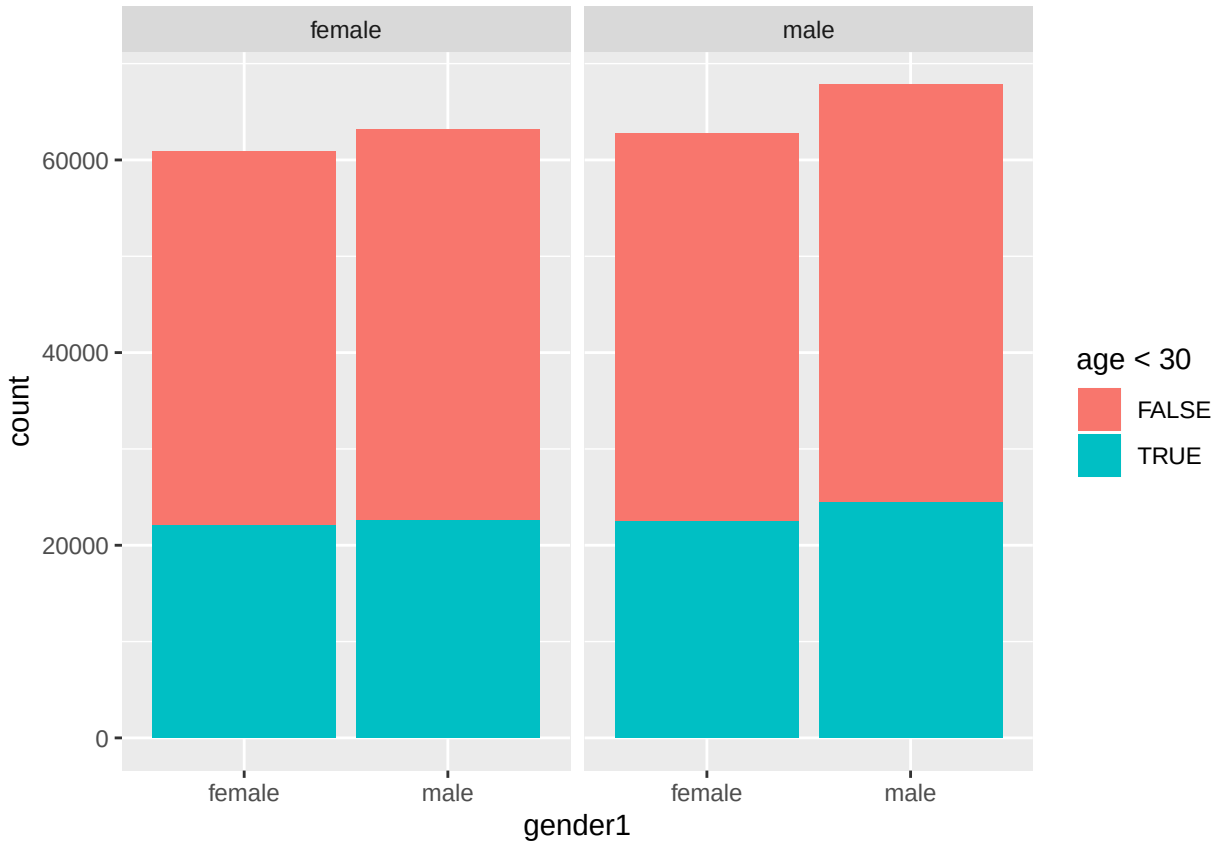
```
## Loading required package: survival
```

```
data(Fertility)

## the contracts the frequency of these four combinations
f_in20s<-Fertility %>% filter(age <30)
f_out20s<-Fertility %>% filter(age >=30)
ggplot(data = Fertility)+
  geom_bar(mapping = aes(x=gender1))+
  facet_grid(.~gender2)
```



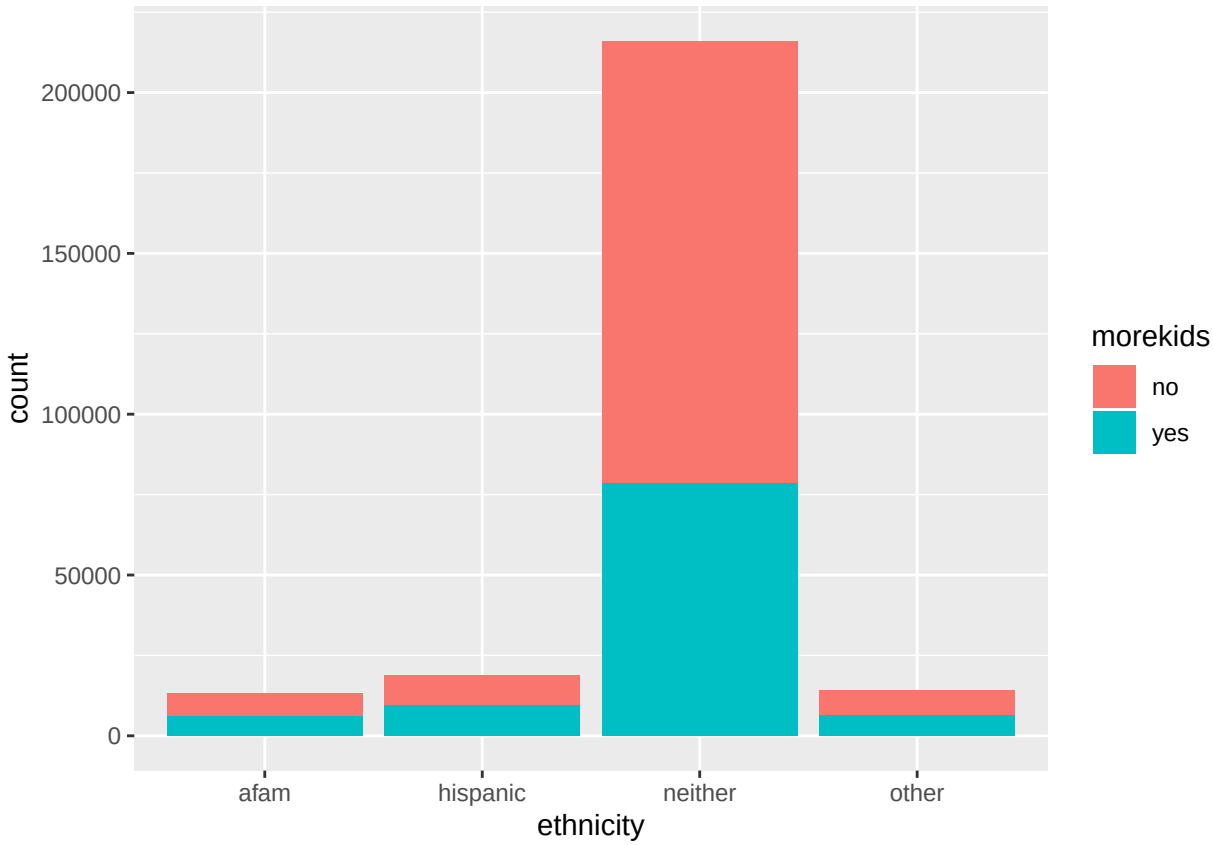

```
## frequencies compariasion for women in their 20s and wemen who are older than 29
ggplot(data = Fertility)+
  geom_bar(mapping = aes(x=gender1,fill = age <30))+
  facet_grid(.~gender2)
```




```
## contrasts the frequency of having more than two children
## by race and ethnicity for four groups of people:
## afam, hispanic, other, or neither or these
f3 <- Fertility %>%
  mutate(neither = (afam == "no" & hispanic == "no" & other == "no") )
f4 <- f3%>%
  within(neither[neither == TRUE]<- "yes")
f_race <-f4 %>% gather(`afam`,`hispanic`,`other`,`neither`, key = ethnicity, value = "yes")%>%
  filter(yes == "yes")
```

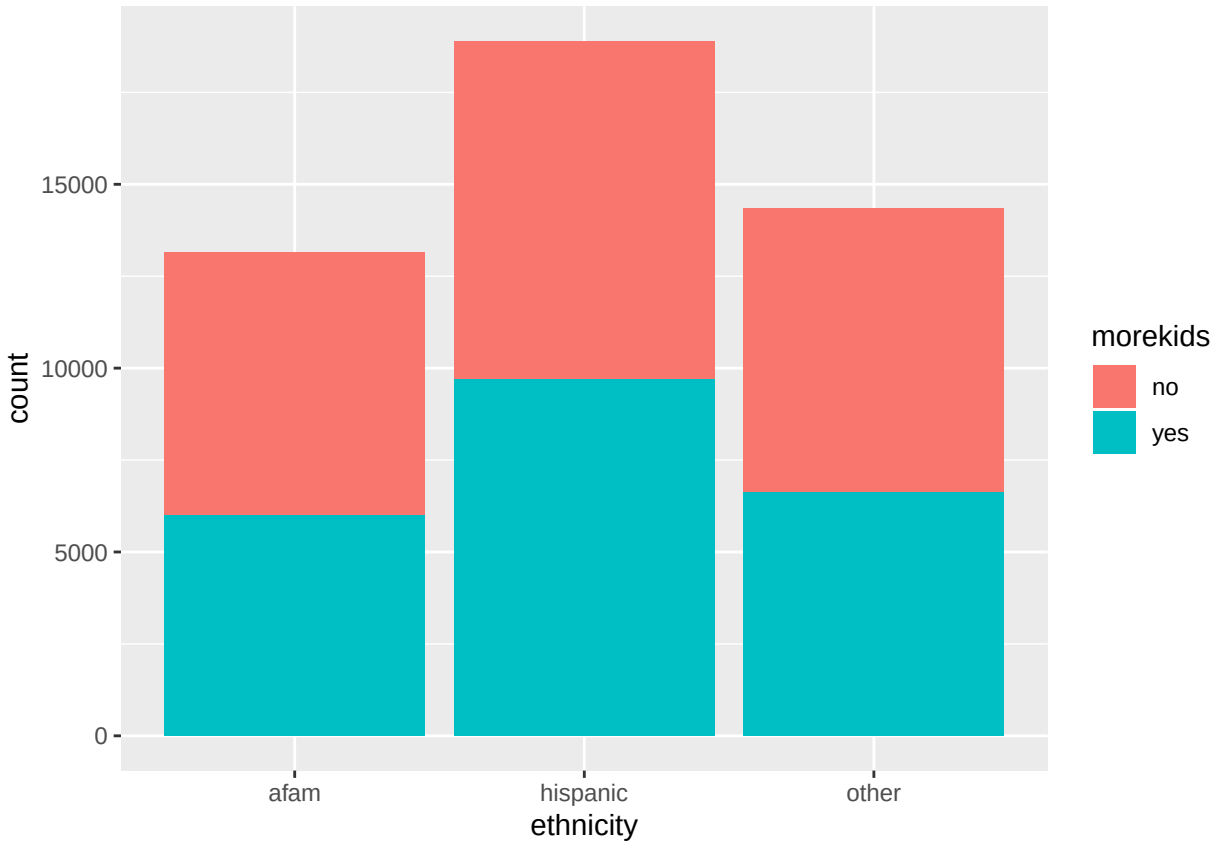
```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
ggplot(data = f_race)+
  geom_bar(mapping =aes(x=ethnicity,fill = morekids))
```




```
## Notice that there are some people have more than one ethnicity
f_test <- f3 %>%
  filter(afam=="yes" & hispanic == "yes")


## contrasts the frequency of having more than two children
## by race and ethnicity for three groups of people:
## afam, hispanic, other
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.
How many times does the letter "e" occur in mtcars rownames?

```
data(mtcars)
data(mpg)
## The letter "e" in mtcars rownames occur 25 times.
mtc <- as_tibble(rownames_to_column(mtcars, var = "Model"))
mtc$n_e<- str_count(mtc$Model, "e")
sum(mtc$n_e)
```

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

How many cars in mtcars have the brand Merc?

```
## 7 cars in mtcars have the brand Merc.
mtc$n_merc<- str_count(mtc$Model, "Merc")
sum(mtc$n_merc)
```

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

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

```
## 4 cars in mpg have the brand Merc.
mpg$n_merc<- str_count(mpg$manufacturer, "mercury")
sum(mpg$n_merc)
```

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

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

```
## creat table named "mtc_merc", with only Merc cars with data mpg.
mtc_merc <-mtc%>%
  separate(Model,sep = " ", into=c("brand","type"))%>%
  select(brand,mpg)%>%
  filter(brand == "Merc")
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 3 rows [2, 4,
## 29].
```

```
## Warning: Expected 2 pieces. Missing pieces filled with `NA` in 1 rows [6].
```

```
## creat table named "mpg_merc", with only Merc cars with data
## cty as "city miles per gallon" and hwy as "highway miles per gallon".
mpg_merc <- mpg %>%
  select(manufacturer, cty, hwy)%>%
  filter(manufacturer == "mercury")
```

Problem 4

Install the babynames package.

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

```
library(babynames)
data(babynames)
set.seed=2019
sample = sample(1924665, 500000)
bns<-babynames[sample,]
```

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

```
bn <-as.tibble(babynames)
```

```
## Warning: `as.tibble()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

```
## displays the five most popular boy names and girl names
#in the years 1880,1920, 1960, 2000.
bn_year <-bn %>%
  filter(year == "1880" |year == "1920" |year == "1960"|year == "2000") %>%
  group_by(year,sex) %>%
  filter(rank(desc(n))<=5)
```

What names overlap boys and girls?

```
# boys <- bn%>% filter(sex == "M")
# girls <- bn %>% filter(sex == "F")
# overlap <- intersect(boys$name,girls$name)
# overlap
# nrow(count(overlap))
# There are 10,663 names that overlap boys and girls.
```

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

```
# nineteenth <- bn %>% filter(year >= 1800 & year <= 1899)
# twentyth <- bn %>% filter(year >= 2000 & year <= 2017)
```

```
# count(!(twentyth$name %in% nineteenth))
# There are 591,925 names in the 19th century but have not been used in the 21sth century.
```

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

```
library(ggpubr)
```

```
##
## Attaching package: 'ggpubr'
```

```
## The following object is masked from 'package:expss':
##
##     compare_means
```

```
theme_set(theme_pubr())
bn %>% filter(name == c("Donald", "Hilary", "Hillary", "Joe", "Barrack"),
                year >= 1800 & year <= 2017) %>%
  ggplot()+
  geom_bar(mapping=aes(x = as.factor(name),y = n),stat="identity",fill = "#0073C2FF")+
  theme_pubclean()
```