Tidyverse Problem Set

Runxin Yu

October 4, 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 And I found that they have been superceded!!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)
data(gapminder)
```

How many continents are included in the data set?

```
fct_unique(gapminder$continent)
```

```
## [1] Africa   Americas Asia     Europe   Oceania
## Levels: Africa Americas Asia Europe Oceania
```

```
# It can be shown from the data above that there are 5 continents included in the data set.
```

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

```
fct_unique(gapminder$country)
```

```
##   [1] Afghanistan              Albania
##   [3] Algeria                  Angola
##   [5] Argentina                Australia
##   [7] Austria                  Bahrain
##   [9] Bangladesh               Belgium
##  [11] Benin                    Bolivia
##  [13] Bosnia and Herzegovina   Botswana
##  [15] Brazil                   Bulgaria
##  [17] Burkina Faso             Burundi
##  [19] Cambodia                 Cameroon
```

```
##  [21] Canada                  Central African Republic
##  [23] Chad                    Chile
##  [25] China                   Colombia
##  [27] Comoros                 Congo, Dem. Rep.
##  [29] Congo, Rep.             Costa Rica
##  [31] Cote d'Ivoire           Croatia
##  [33] Cuba                    Czech Republic
##  [35] Denmark                 Djibouti
##  [37] Dominican Republic      Ecuador
##  [39] Egypt                   El Salvador
##  [41] Equatorial Guinea       Eritrea
##  [43] Ethiopia                Finland
##  [45] France                  Gabon
##  [47] Gambia                  Germany
##  [49] Ghana                   Greece
##  [51] Guatemala               Guinea
##  [53] Guinea-Bissau           Haiti
##  [55] Honduras                Hong Kong, China
##  [57] Hungary                 Iceland
##  [59] India                   Indonesia
##  [61] Iran                    Iraq
##  [63] Ireland                 Israel
##  [65] Italy                   Jamaica
##  [67] Japan                   Jordan
##  [69] Kenya                   Korea, Dem. Rep.
##  [71] Korea, Rep.             Kuwait
##  [73] Lebanon                 Lesotho
##  [75] Liberia                 Libya
##  [77] Madagascar              Malawi
##  [79] Malaysia                Mali
##  [81] Mauritania              Mauritius
##  [83] Mexico                  Mongolia
##  [85] Montenegro              Morocco
##  [87] Mozambique              Myanmar
##  [89] Namibia                 Nepal
##  [91] Netherlands             New Zealand
##  [93] Nicaragua               Niger
##  [95] Nigeria                 Norway
##  [97] Oman                    Pakistan
##  [99] Panama                  Paraguay
## [101] Peru                    Philippines
## [103] Poland                  Portugal
## [105] Puerto Rico             Reunion
## [107] Romania                 Rwanda
## [109] Sao Tome and Principe   Saudi Arabia
## [111] Senegal                 Serbia
## [113] Sierra Leone            Singapore
## [115] Slovak Republic         Slovenia
## [117] Somalia                 South Africa
## [119] Spain                   Sri Lanka
## [121] Sudan                   Swaziland
## [123] Sweden                  Switzerland
## [125] Syria                   Taiwan
## [127] Tanzania                Thailand
```

```
## [129] Togo                     Trinidad and Tobago
## [131] Tunisia                  Turkey
## [133] Uganda                   United Kingdom
## [135] United States            Uruguay
## [137] Venezuela                Vietnam
## [139] West Bank and Gaza       Yemen, Rep.
## [141] Zambia                   Zimbabwe
## 142 Levels: Afghanistan Albania Algeria Angola Argentina ... Zimbabwe
```

```
# It can be shown from the data above that there are 142 countries included in the data set.
gapminder %>% group_by(continent) %>% summarize(num_obs = n(), num_countries = n_distinct(country))
```

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

```
# From the data above, there are 52 contries in Africa, 25 countries in Americas, 33 countries in Asia,
# 30 countires in Europe and 2 countries in Oceania.
```

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.

```
df1 <- gapminder[,c("continent","pop","gdpPercap")]
report <- aggregate(df1[,2:3],by=list(df1$continent),sum)
kable(report, digits = 2,
      col.names = c("Continent", "Total population per continent", "GDP per capita"),
      caption = "Total Population and GDP per capita by continents" ,align = 'c') %>%
  kable_styling(latex_options = 'hold_position',font_size = 15.5,full_width = F)%>%
  column_spec(1,bold = T)
```

Table 1: Total Population and GDP per capita by continents

Continent	Total population per continent	GDP per capita
Africa	6187585961	1368902.9
Americas	7351438499	2140833.1
Asia	30507333901	3129251.6
Europe	6181115304	5209011.2
Oceania	212992136	446918.6

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

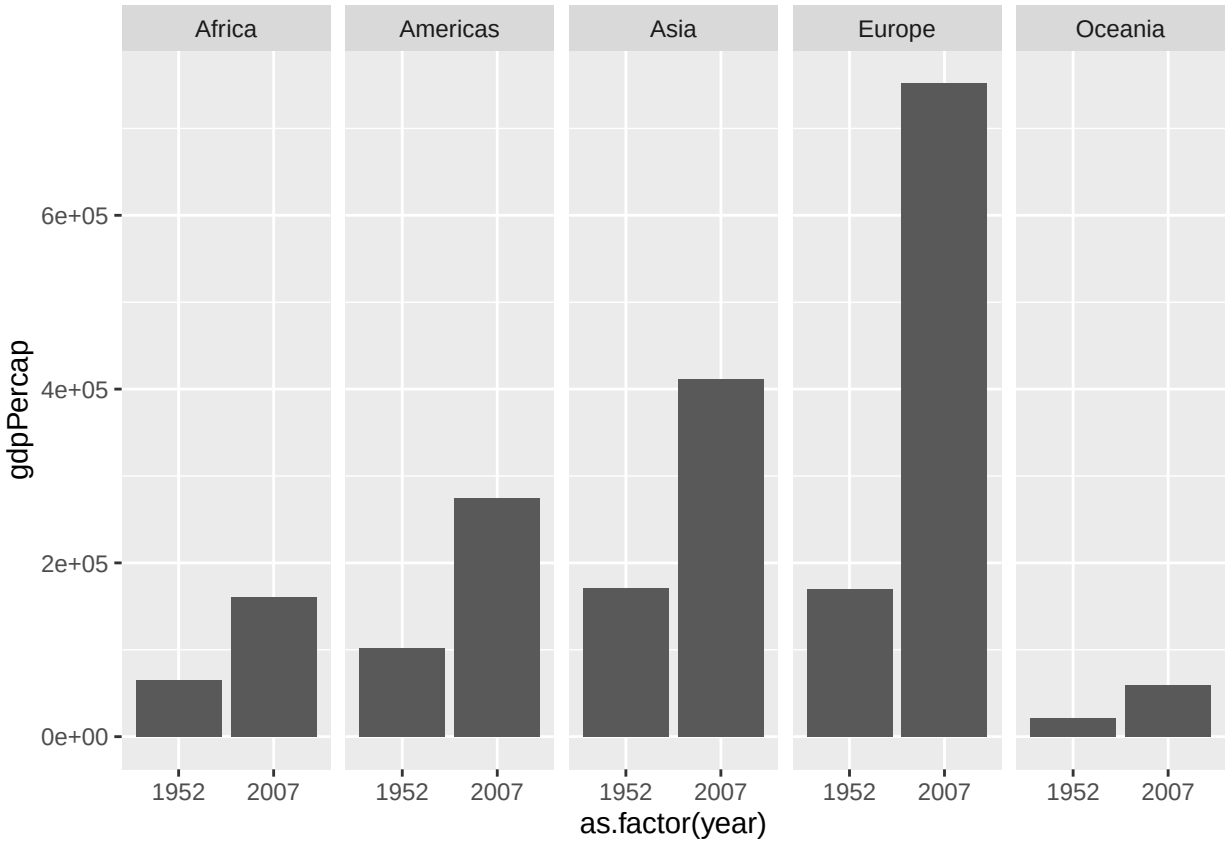
```
options(digits = 3)
gdp_52 <- gapminder %>% select(country,continent,year,gdpPercap) %>% filter(year == 1952)
gdp_07 <- gapminder %>% select(country,continent,year,gdpPercap) %>% filter(year == 2007)

gdp_52$year <- NULL
gdp_07$year <- NULL
```

```
table_top <- cbind(gdp_52, gdp_07)
colnames(table_top) <- c("Country", "Continent", "GDP",
                         "Country", "Continent", "GDP")
kable(table_top, digits = 3, booktabs=TRUE, caption = "GDP per capita for the countries in each continer
  kable_styling(latex_options = 'hold_position',font_size = 12) %>%
  column_spec(c(1,4),bold = T) %>%
  add_header_above(c("GDP 1952"=3,"GDP 2007"=3))
```

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)) %>%
  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.

```
gapminder_drop <- drop_na(gapminder)
gapminder_drop %>% 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))
```

```
## # 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
```

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.

```
## 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
```

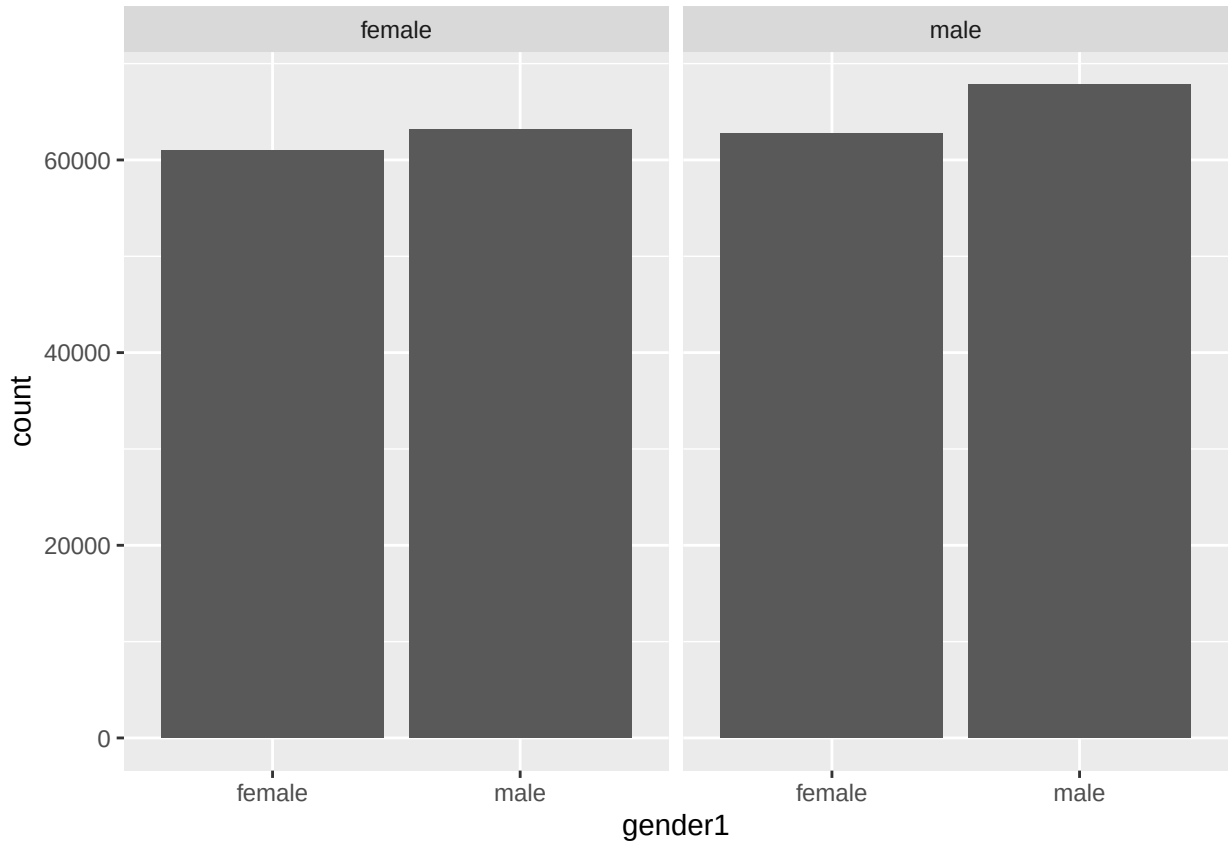
```
## Loading required package: sandwich
```

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

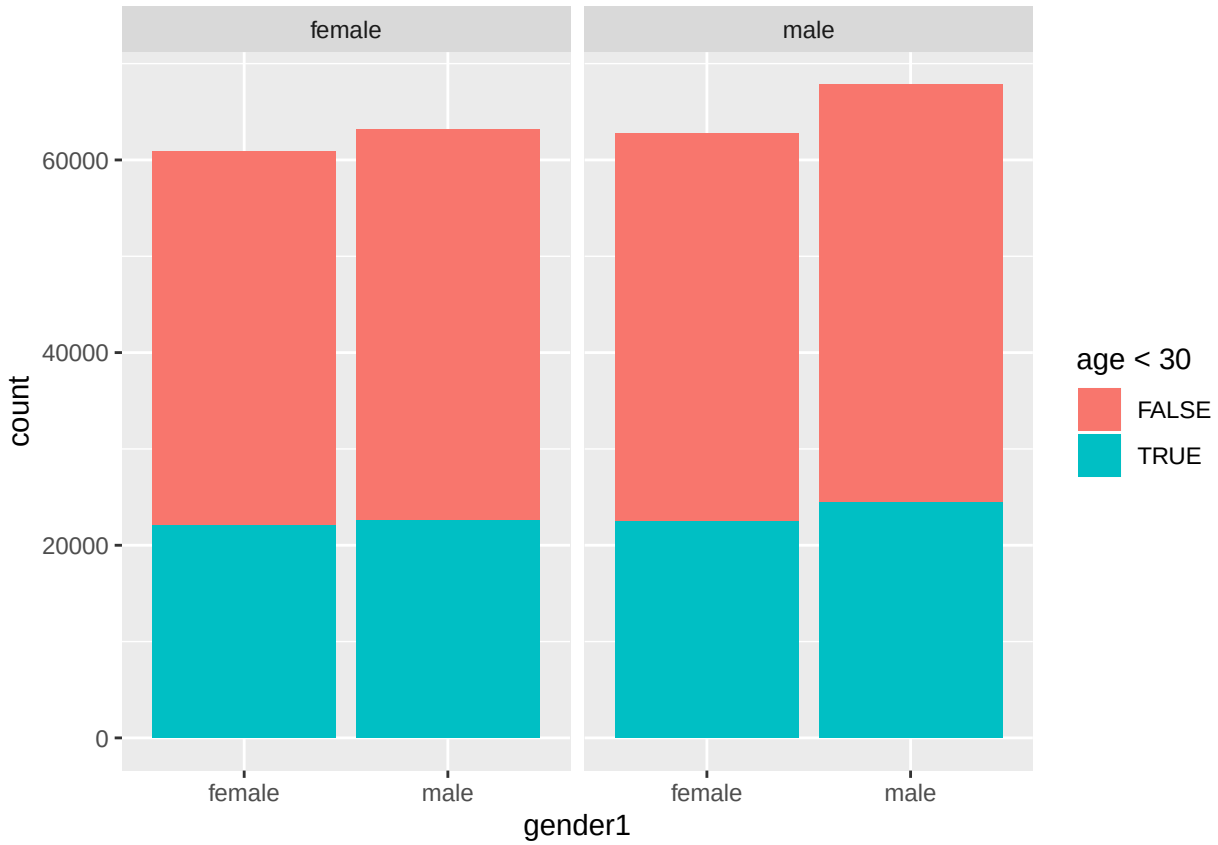
There are four possible gender combinations for the first two Children. Product a plot the contracts the frequency of these four combinations. Are the frequencies different for women in their 20s and wemen who are older than 29?

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

```
## 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)
```



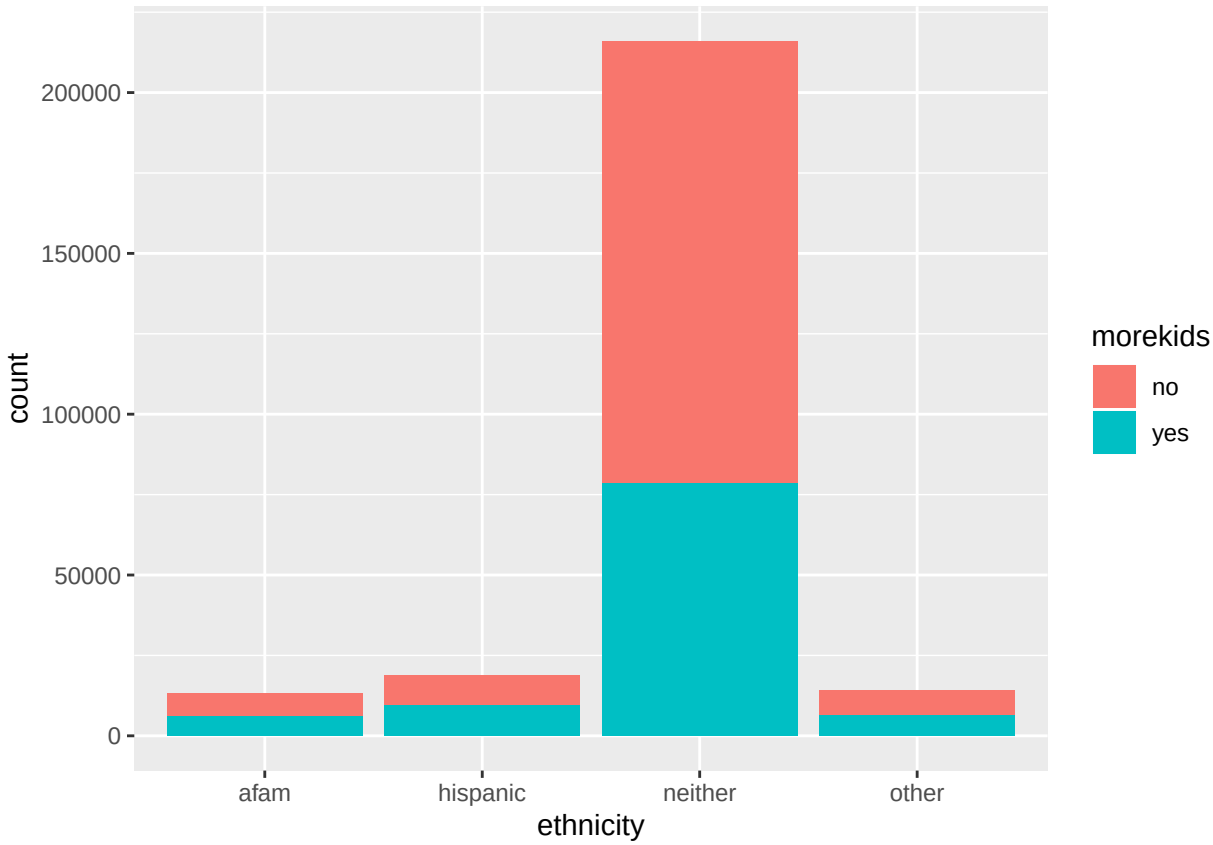

```
## frequencies compariasion for women in their 20s and wemen who are older than 29
ggplot(data = Fertility)+
  geom_bar(mapping = aes(x=gender1,fill = age <30))+
  facet_grid(.~gender2)
```




```
## 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")
```

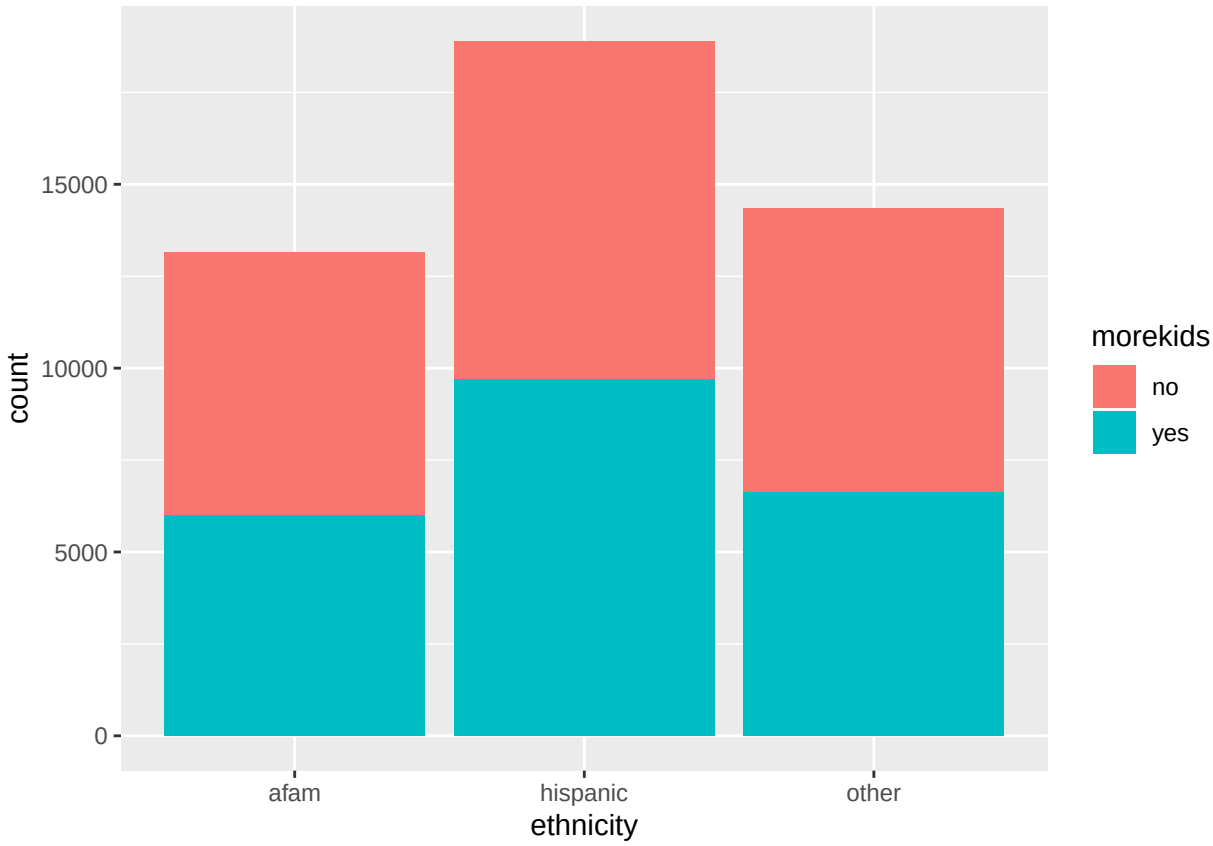
```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
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?

```
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
```

```
# The letter "e" in mtcars rownames occur 25 times.
```

How many cars in mtcars have the brand Merc?

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

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

```
# There are 7 cars in mtcars have the brand Merc.
```

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

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

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

```
# There are 4 cars in mpg have the brand Merc.
```

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 = 1:1924665, size = 500000, replace = FALSE)
```

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

```
# For boys' names
popular_boy_1880 <- babyn %>% select(year,sex,name,n) %>%
    filter(sex == "M", year == 1880) %>% head(5)
popular_boy_1880$year <- NULL
popular_boy_1920 <- babyn %>% select(year,sex,name,n) %>%
    filter(sex == "M", year == 1920) %>% head(5)
popular_boy_1920$year <- NULL
popular_boy_1960 <- babyn %>% select(year,sex,name,n) %>%
    filter(sex == "M", year == 1960) %>% head(5)
popular_boy_1960$year <- NULL
popular_boy_2000 <- babyn %>% select(year,sex,name,n) %>%
    filter(sex == "M", year == 2000) %>% head(5)
popular_boy_2000$year <- NULL

table_boy <- cbind(popular_boy_1880,popular_boy_1920,popular_boy_1960,popular_boy_2000)
colnames(table_boy) <- c("Sex", "Name","Population",
                      "Sex", "Name", "Population",
                      "Sex", "Name", "Population",
                      "Sex", "Name", "Population"
                      )
kable(table_boy, digits = 3,booktabs=TRUE,caption = "Top 5 popular boys names: 1880,1920,1960,2000",ali
 add_header_above(c("Name(boy) 1880"=3,
                      "Name(boy) 1920"=3,
                      "Name(boy) 1960"=3,
                      "Name(boy) 2000"=3))
```

```
# For girls' names
popular_girl_1880 <- babyn %>% select(year,sex,name,n) %>%
    filter(sex == "F", year == 1880) %>% head(5)
popular_girl_1880$year <- NULL
popular_girl_1920 <- babyn %>% select(year,sex,name,n) %>%
    filter(sex == "F", year == 1920) %>% head(5)
popular_girl_1920$year <- NULL
popular_girl_1960 <- babyn %>% select(year,sex,name,n) %>%
    filter(sex == "F", year == 1960) %>% head(5)
popular_girl_1960$year <- NULL
popular_girl_2000 <- babyn %>% select(year,sex,name,n) %>%
    filter(sex == "F", year == 2000) %>% head(5)
popular_girl_2000$year <- NULL

table_girl <- cbind(popular_girl_1880,popular_girl_1920,popular_girl_1960,popular_girl_2000)
colnames(table_boy) <- c("Sex", "Name","Population",
```

```
                      "Sex", "Name", "Population",
                      "Sex", "Name", "Population",
                      "Sex", "Name", "Population"
                      )
kable(table_girl, digits = 3,booktabs=TRUE,caption = "Top 5 popular girls names: 1880,1920,1960,2000",a
 add_header_above(c("Name(girl) 1880"=3,
                      "Name(girl) 1920"=3,
                      "Name(girl) 1960"=3,
                      "Name(girl) 2000"=3))
```

What names overlap boys and girls?

```
overlap<-babyn %>% group_by(year,name) %>% summarise(count = length(sex)) %>% arrange(desc(count)) %>%
  filter(count>1)
unique(overlap$name)[1:10]
```

```
##  [1] "Addie"   "Allie"   "Alma"    "Alpha"   "Alva"    "Anna"    "Annie"
##  [8] "Arthur"  "Artie"   "Augusta"
```

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

```
nineteen <- babyn %>% filter(year > 1999)
nineteen <- unique(nineteen$name)
twenty <- babyn %>% filter(year < 1900)
twenty <- unique(twenty$name)
Int <- intersect(nineteen, twenty)
Int[1:10]
```

```
##  [1] "Emily"     "Hannah"    "Madison"   "Ashley"    "Sarah"
##  [6] "Alexis"    "Samantha"  "Jessica"   "Elizabeth" "Taylor"
```

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())
babyn %>% 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()
```

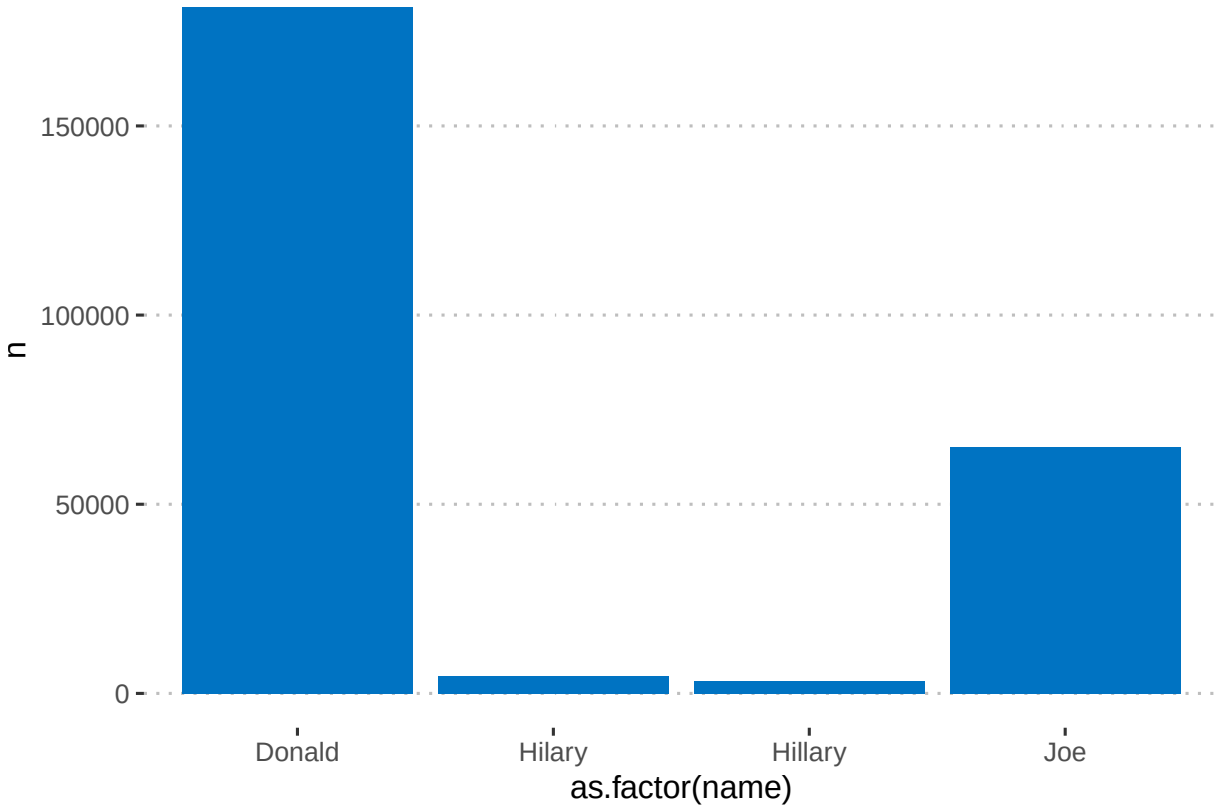

150000
100000
50000
0
n
Donald
Hilary
Hillary
Joe
as.factor(name)

Table 2: GDP per capita for the countries in each continent: 1952, 2007

GDP 1952			GDP 2007		
Country	Continent	GDP	**Country**	Continent	GDP
Afghanistan	Asia	779	**Afghanistan**	Asia	975
Albania	Europe	1601	**Albania**	Europe	5937
Algeria	Africa	2449	**Algeria**	Africa	6223
Angola	Africa	3521	**Angola**	Africa	4797
Argentina	Americas	5911	**Argentina**	Americas	12779
Australia	Oceania	10040	**Australia**	Oceania	34435
Austria	Europe	6137	**Austria**	Europe	36126
Bahrain	Asia	9867	**Bahrain**	Asia	29796
Bangladesh	Asia	684	**Bangladesh**	Asia	1391
Belgium	Europe	8343	**Belgium**	Europe	33693
Benin	Africa	1063	**Benin**	Africa	1441
Bolivia	Americas	2677	**Bolivia**	Americas	3822
Bosnia and Herzegovina	Europe	974	**Bosnia and Herzegovina**	Europe	7446
Botswana	Africa	851	**Botswana**	Africa	12570
Brazil	Americas	2109	**Brazil**	Americas	9066
Bulgaria	Europe	2444	**Bulgaria**	Europe	10681
Burkina Faso	Africa	543	**Burkina Faso**	Africa	1217
Burundi	Africa	339	**Burundi**	Africa	430
Cambodia	Asia	368	**Cambodia**	Asia	1714
Cameroon	Africa	1173	**Cameroon**	Africa	2042
Canada	Americas	11367	**Canada**	Americas	36319
Central African Republic	Africa	1071	**Central African Republic**	Africa	706
Chad	Africa	1179	**Chad**	Africa	1704
Chile	Americas	3940	**Chile**	Americas	13172
China	Asia	400	**China**	Asia	4959
Colombia	Americas	2144	**Colombia**	Americas	7007
Comoros	Africa	1103	**Comoros**	Africa	986
Congo, Dem. Rep.	Africa	781	**Congo, Dem. Rep.**	Africa	278
Congo, Rep.	Africa	2126	**Congo, Rep.**	Africa	3633
Costa Rica	Americas	2627	**Costa Rica**	Americas	9645
Cote d'Ivoire	Africa	1389	**Cote d'Ivoire**	Africa	1545
Croatia	Europe	3119	**Croatia**	Europe	14619
Cuba	Americas	5587	**Cuba**	Americas	8948
Czech Republic	Europe	6876	**Czech Republic**	Europe	22833
Denmark	Europe	9692	**Denmark**	Europe	35278
Djibouti	Africa	2670	**Djibouti**	Africa	2082
Dominican Republic	Americas	1398	**Dominican Republic**	Americas	6025
Ecuador	Americas	3522	**Ecuador**	Americas	6873
Egypt	Africa	1419	**Egypt**	Africa	5581
El Salvador	Americas	3048	**El Salvador**	Americas	5728
Equatorial Guinea	Africa	376	**Equatorial Guinea**	Africa	12154
Eritrea	Africa	329	**Eritrea**	Africa	641
Ethiopia	Africa	362	**Ethiopia**	Africa	691
Finland	Europe	6425	**Finland**	Europe	33207
France	Europe	7030	**France**	Europe	30470

Table 3: Top 5 popular boys names: 1880,1920,1960,2000

Name(boy) 1880			Name(boy) 1920			Name(boy) 1960			Name(boy	
Sex	Name	Population	Sex	Name	Population	Sex	Name	Population	Sex	Name
M	John	9655	M	John	56913	M	David	85928	M	Jacob
M	William	9532	M	William	50147	M	Michael	84183	M	Michael
M	James	5927	M	Robert	48678	M	James	76842	M	Matthew
M	Charles	5348	M	James	47909	M	John	76096	M	Joshua
M	George	5126	M	Charles	28308	M	Robert	72369	M	Christophe

Table 4: Top 5 popular girls names: 1880,1920,1960,2000

Name(girl) 1880			Name(girl) 1920			Name(girl) 1960			Name(girl) 2000		
sex	name	n	sex	name	n	sex	name	n	sex	name	n
F	Mary	7065	F	Mary	70980	F	Mary	51474	F	Emily	25953
F	Anna	2604	F	Dorothy	36643	F	Susan	39200	F	Hannah	23080
F	Emma	2003	F	Helen	35097	F	Linda	37314	F	Madison	19967
F	Elizabeth	1939	F	Margaret	27997	F	Karen	36376	F	Ashley	17997
F	Minnie	1746	F	Ruth	26101	F	Donna	34133	F	Sarah	17697