Tidyverse Problem Set

MA615 Xiaofan Xia

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 And I found that they have been superceded!!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)
head(gapminder)
```

```
## # A tibble: 6 x 6
##   country     continent  year lifeExp      pop gdpPercap
##   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
## 1 Afghanistan Asia       1952    28.8  8425333      779.
## 2 Afghanistan Asia       1957    30.3  9240934      821.
## 3 Afghanistan Asia       1962    32.0 10267083      853.
## 4 Afghanistan Asia       1967    34.0 11537966      836.
## 5 Afghanistan Asia       1972    36.1 13079460      740.
## 6 Afghanistan Asia       1977    38.4 14880372      786.
```

How many continents are included in the data set?

```
str(gapminder$continent) #There are 5 continents included in the dataset.
```

```
##  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(gapminder$country) #There are 142 countries included in the dataset.
```

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

```
gapminder.country <- gapminder %>%
  group_by(continent) %>%
  summarize(n = n(),
            n_countries = n_distinct(country))

gapminder.country #number of countries per continent
```

```
## # 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.

```
gapminder.pop <-
  gapminder%>%
  group_by(continent) %>%
  summarize(population=sum(as.numeric(pop)),GDP=sum(gdpPercap))

kable(gapminder.pop) #total population and total GDP for each continent in table
```

continent	population	GDP
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.

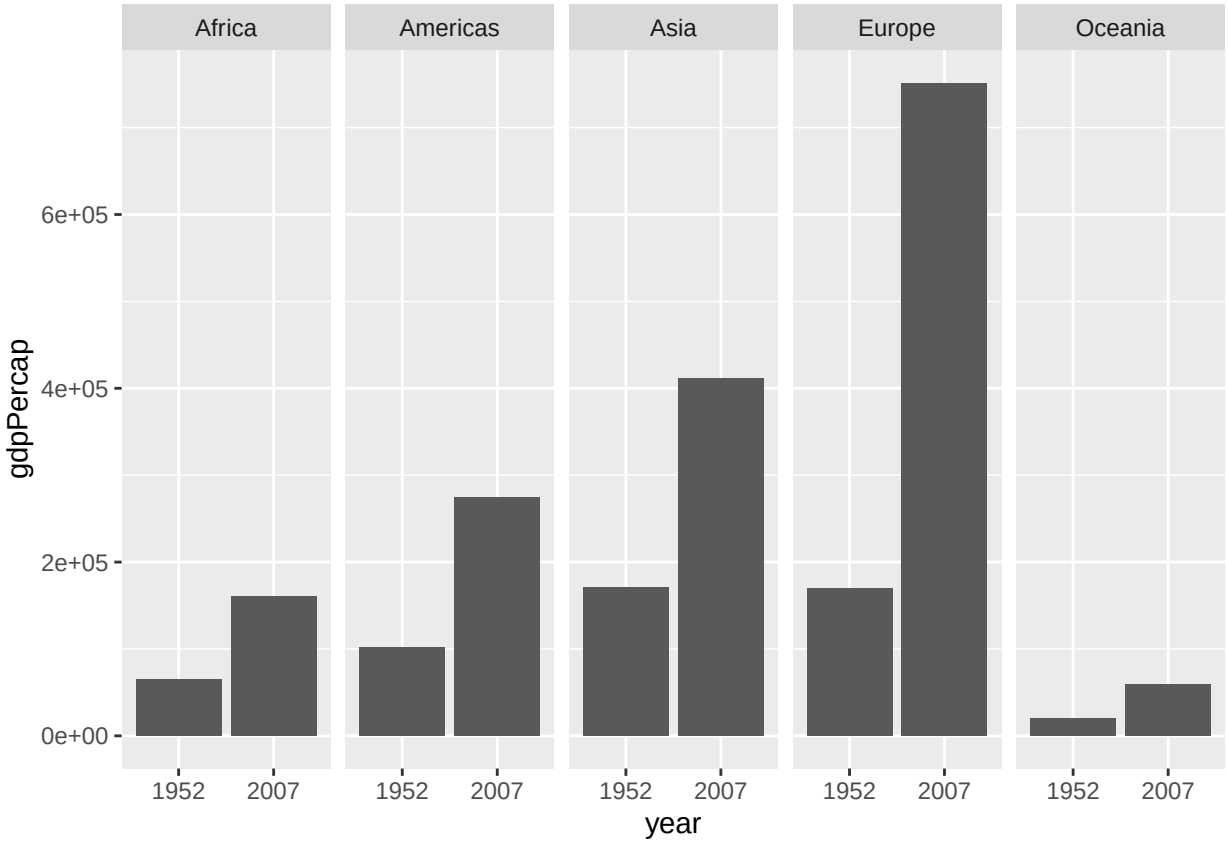
```
gapminder.1952 <-
  gapminder%>%
  filter(year==1952)%>%
  group_by(continent)%>%
  summarize(GDP=sum(gdpPercap)) #data of 1952

year.1952 <- c(rep(1952,5))

gapminder.1952.new <- cbind(year.1952,gapminder.1952)

gapminder.2007 <-
  gapminder%>%
  filter(year==2007)%>%
  group_by(continent) %>%
  summarize(GDP=sum(gdpPercap)) #data of 2007

year.2007 <- c(rep(2007,5))
```

```
gapminder.2007.new <- cbind(year.2007,gapminder.2007)

combined.gapminder.year <- cbind(gapminder.1952.new,gapminder.2007.new)

kable(combined.gapminder.year) #a table that summarizes GDP per capita for the countries in each contin
```

year.1952	continent	GDP	year.2007	continent	GDP
1952	Africa	65133.77	2007	Africa	160629.70
1952	Americas	101976.56	2007	Americas	275075.79
1952	Asia	171450.97	2007	Asia	411609.89
1952	Europe	169831.72	2007	Europe	751634.45
1952	Oceania	20596.17	2007	Oceania	59620.38

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

```
gapminder.gdp <-
  gapminder%>%
  filter(year==c(1952, 2007)) #data of 1952 and 2007

ggplot(gapminder.gdp,aes(year,gdpPercap))+
  geom_bar(mapping=aes(x=as.factor(year),y=gdpPercap),stat="identity")+
  facet_grid(.~continent) #a plot that summarizes GDP per capita for the countries in each continent, c
```



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.

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
```

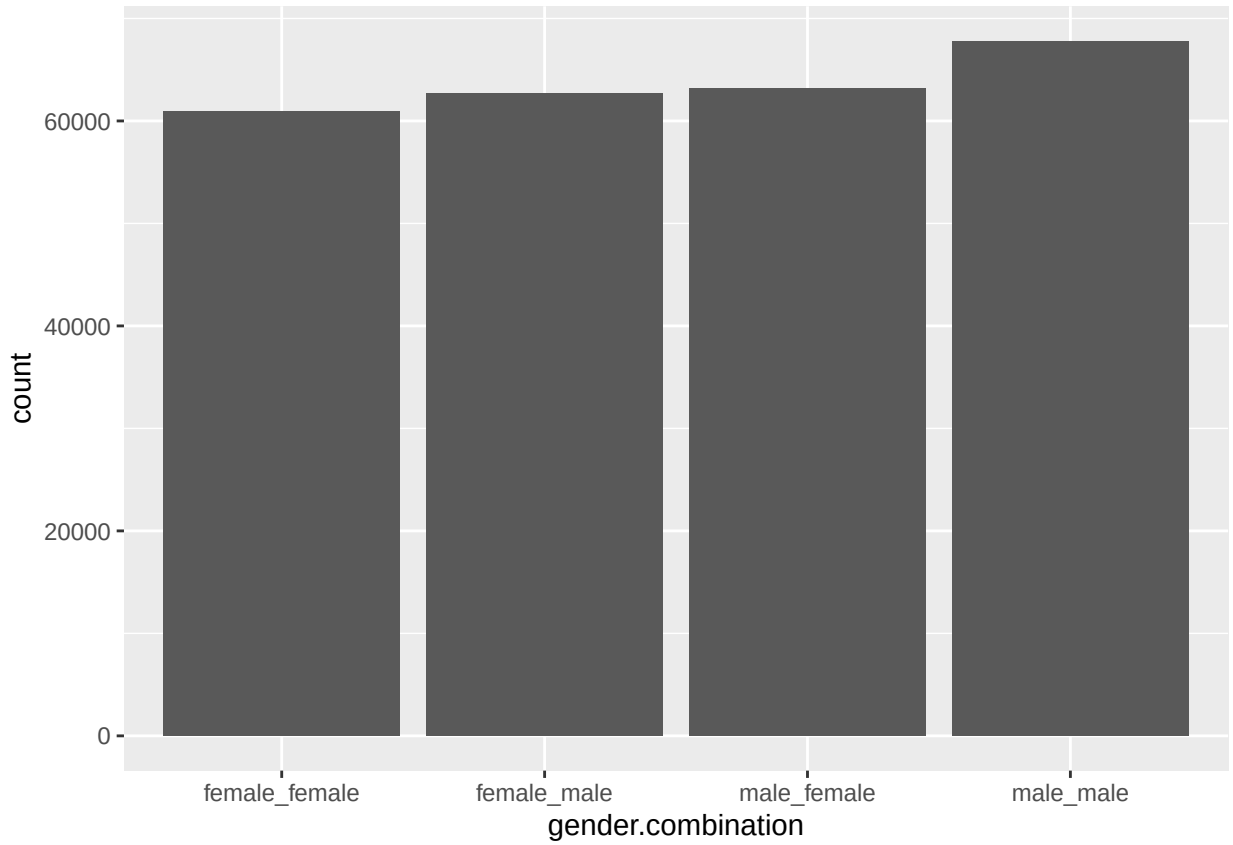
```
## Loading required package: survival
```

```
data("Fertility")
head(Fertility)
```

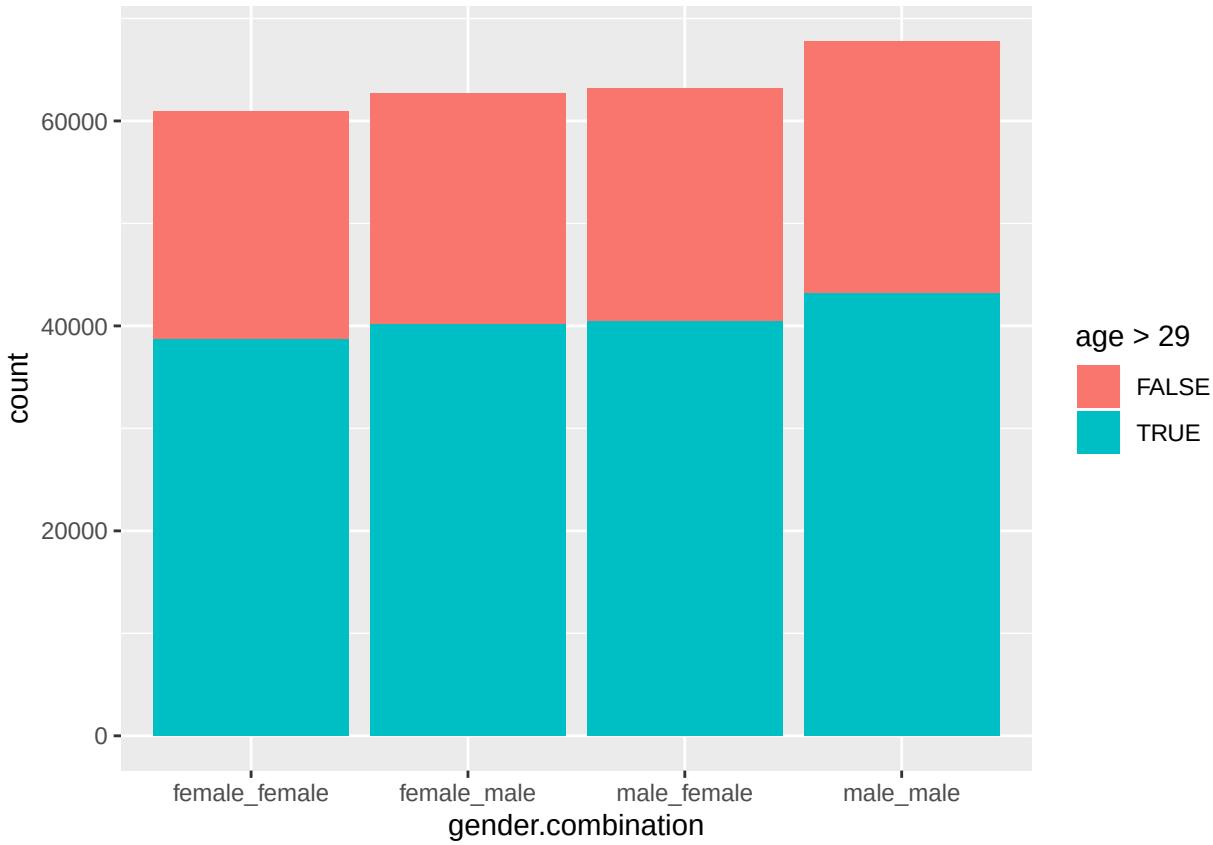
```
##   morekids gender1 gender2 age afam hispanic other work
## 1       no    male  female  27   no       no    no    0
## 2       no  female    male  30   no       no    no   30
## 3       no    male  female  27   no       no    no    0
## 4       no    male  female  35  yes       no    no    0
## 5       no  female  female  30   no       no    no   22
## 6       no    male  female  26   no       no    no   40
```

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.gender <-
  Fertility %>%
  unite(gender.combination, gender1, gender2) %>%
  select(gender.combination,age) %>%
  arrange(gender.combination)

ggplot(fertility.gender, aes(x=gender.combination)) + geom_bar() #plot the contracts the frequency of t
```

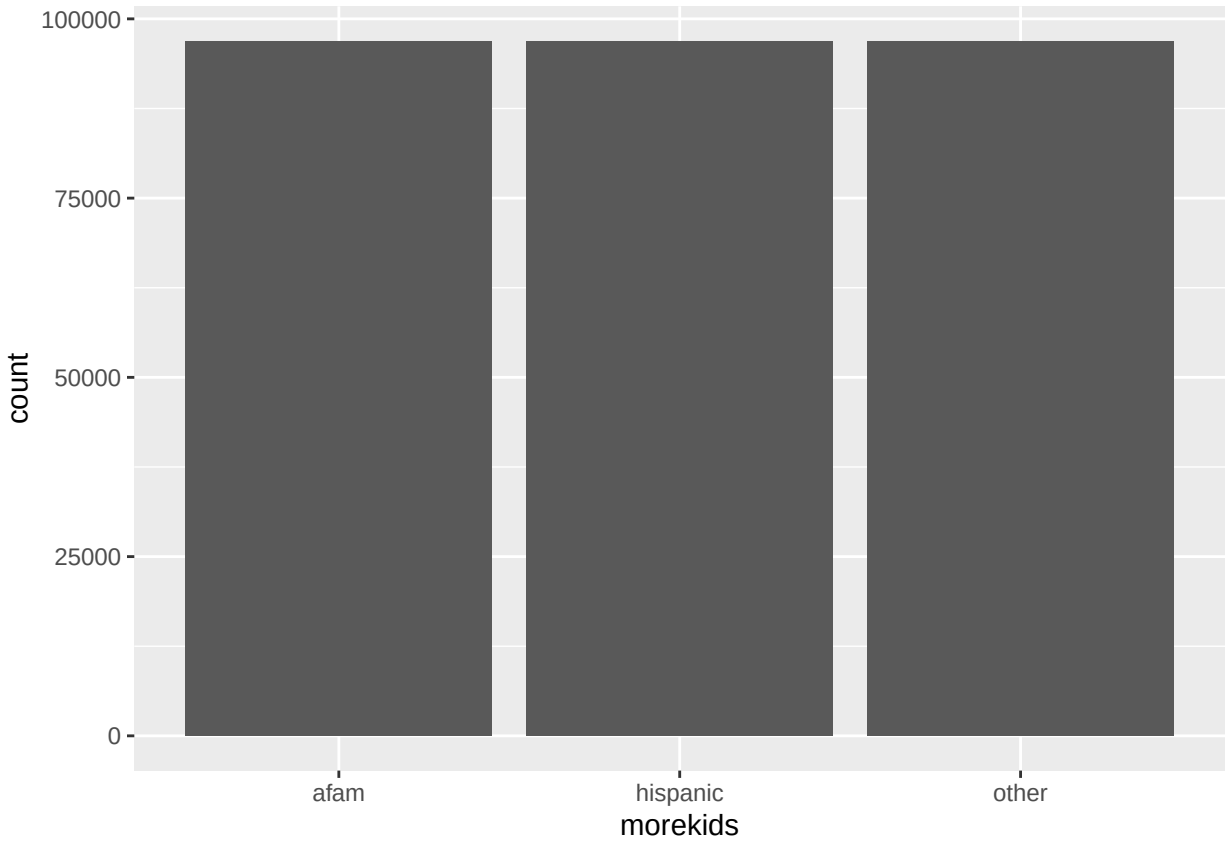



```
ggplot(fertility.gender, aes(x=gender.combination, fill=age>29)) + geom_bar() #plot the contracts the f
```



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

```
fertility.race <-
  Fertility %>%
  select(morekids,afam,hispanic,other) %>%
  filter(morekids=="yes") %>%
  gather(afam, hispanic, other, key="morekids", value = "No")

ggplot(fertility.race, aes(x=morekids)) + geom_bar()
```



Problem 3

Use the mtcars and mpg datasets.

```
df.mtcars <- data.frame(mtcars)
df.mtcars <- tibble::rownames_to_column(df.mtcars, "car") #transfer the rownames in mtcars into first c
head(df.mtcars)
```

```
##                 car  mpg cyl disp  hp drat    wt  qsec vs am gear carb
## 1         Mazda RX4 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## 2     Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## 3        Datsun 710 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## 4    Hornet 4 Drive 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## 5 Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## 6           Valiant 18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
head(mpg)
```

```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans  drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>  <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(~ f        18    29 p     comp~
## 2 audi         a4      1.8  1999     4 manua~ f        21    29 p     comp~
## 3 audi         a4      2    2008     4 manua~ f        20    31 p     comp~
## 4 audi         a4      2    2008     4 auto(~ f        21    30 p     comp~
## 5 audi         a4      2.8  1999     6 auto(~ f        16    26 p     comp~
## 6 audi         a4      2.8  1999     6 manua~ f        18    26 p     comp~
```

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

```
number_e <- str_count(df.mtcars$car,"e") #Count numbers of letter "e" occurred in each mtcars car name(
number_e
```

```
##  [1] 0 0 0 2 1 0 1 1 1 1 1 1 1 1 2 1 2 0 0 0 0 3 1 0 1 0 1 0 1 1 1 0
```

```
sum(number_e) #The numnber of occurances of letter "e" in total is 25.
```

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

How many cars in mtcars have the brand Merc?

```
number_Merc <- str_count(df.mtcars$car,"Merc") #Count numbers of "Merc" occurred in each mtcars car nam
number_Merc
```

```
##  [1] 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
sum(number_Merc) #The numnber of occurances of "Merc" in total is 7.
```

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

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

```
number_Merc_mpg <- str_count(mpg$manufacturer,"merc") #Count numbers of "merc" occurred in each row of 
number_Merc_mpg
```

```
##   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [36] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##  [71] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [106] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1
## [141] 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [176] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [211] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
sum(number_Merc_mpg) #The numnber of occurances of "Merc" in total is 4.
```

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

```
#Not sure about if mercury is "Merc", but I am assuming mercury is denoted by "Merc".
```

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

```
df.mtcars.new <- df.mtcars %>%
  separate(car,into = c("manufacturer","model"),sep = " ") %>%
  select(manufacturer,mpg) %>%
  filter(manufacturer=="Merc") %>%
  mutate(manufacturer,mpg)
```

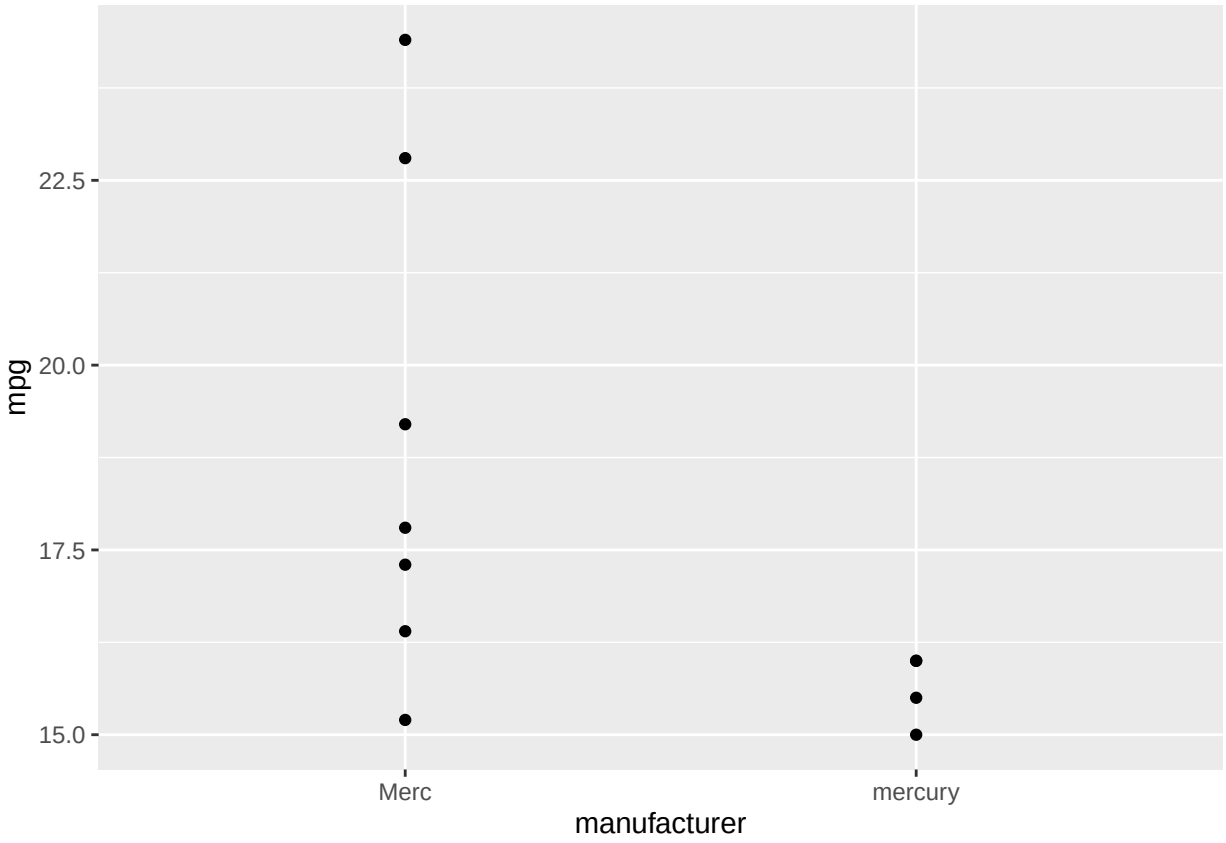
```
## Warning: Expected 2 pieces. Additional pieces discarded in 3 rows [2, 4,
## 29].
```

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

```
mpg.new <- mpg %>%
  select(manufacturer,cty,hwy) %>%
  filter(manufacturer=="mercury") %>%
  transmute(manufacturer,mpg=(cty+hwy)/2)

combined.mpg <- merge(df.mtcars.new,mpg.new,all=T)

kable(combined.mpg) #Contrast the mileage data for Merc cars in two different datasets by a table (diff
```

manufacturer	mpg
Merc	15.2
Merc	16.4
Merc	17.3
Merc	17.8
Merc	19.2
Merc	22.8
Merc	24.4
mercury	15.0
mercury	15.5
mercury	16.0
mercury	16.0

```
ggplot(data=combined.mpg,aes(manufacturer,mpg))+ geom_point() #Contrast the mileage data for Merc cars
```




```
#The mpg dataset has less samples than the mtcars dataset does. The data of mileage of Merc in mtcars h
```

Problem 4

Install the babynames package.

```
library(babynames)
head(babynames)
```

```
## # A tibble: 6 x 5
##    year sex   name          n   prop
##   <dbl> <chr> <chr>     <int>  <dbl>
```

```
## 1  1880 F     Mary       7065 0.0724
## 2  1880 F     Anna       2604 0.0267
## 3  1880 F     Emma       2003 0.0205
## 4  1880 F     Elizabeth  1939 0.0199
## 5  1880 F     Minnie     1746 0.0179
## 6  1880 F     Margaret   1578 0.0162
```

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

```
babynames.sample<-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.

```
babynames.1880.F <-
  babynames %>%
  filter(year==1880,sex=="F")
F.1880 <-
  babynames.1880.F %>%
  group_by(name) %>%
  summarise(sum(n))
F.1880 <- F.1880[order(-F.1880$`sum(n)`),]
F.1880.top5 <- F.1880[c(1:5),]
year.1 <- rep(1880,5)
baby.1880.F <- cbind(year.1,F.1880.top5)

babynames.1880.M <-
  babynames %>%
  filter(year==1880,sex=="M")
M.1880 <-
  babynames.1880.M %>%
  group_by(name) %>%
  summarise(sum(n))
M.1880 <- M.1880[order(-M.1880$`sum(n)`),]
M.1880.top5 <- M.1880[c(1:5),]
baby.1880.M <- cbind(year.1,M.1880.top5)

babynames.1920.F <-
  babynames %>%
  filter(year==1920,sex=="F")
F.1920 <-
  babynames.1920.F %>%
  group_by(name) %>%
  summarise(sum(n))
F.1920 <- F.1920[order(-F.1920$`sum(n)`),]
F.1920.top5 <- F.1920[c(1:5),]
year.2 <- rep(1920,5)
baby.1920.F <- cbind(year.2,F.1920.top5)

babynames.1920.M <-
  babynames %>%
  filter(year==1920,sex=="M")
M.1920 <-
  babynames.1920.M %>%
  group_by(name) %>%
  summarise(sum(n))
```

```r
M.1920 <- M.1920[order(-M.1920$`sum(n)`),]
M.1920.top5 <- M.1920[c(1:5),]
baby.1920.M <- cbind(year.2,M.1920.top5)

babynames.1960.F <-
  babynames %>%
  filter(year==1960, sex=="F")
F.1960 <-
  babynames.1960.F %>%
  group_by(name) %>%
  summarise(sum(n))
F.1960 <- F.1960[order(-F.1960$`sum(n)`),]
F.1960.top5 <- F.1960[c(1:5),]
year.3 <- rep(1960,5)
baby.1960.F <- cbind(year.3,F.1960.top5)

babynames.1960.M <-
  babynames %>%
  filter(year==1960, sex=="M")
M.1960 <-
  babynames.1960.M %>%
  group_by(name) %>%
  summarise(sum(n))
M.1960 <- M.1960[order(-M.1960$`sum(n)`),]
M.1960.top5 <- F.1960[c(1:5),]
baby.1960.M <- cbind(year.3,M.1960.top5)

babynames.2000.F <-
  babynames %>%
  filter(year==2000,sex=="F")
F.2000 <-
  babynames.2000.F %>%
  group_by(name) %>%
  summarise(sum(n))
F.2000<- F.2000[order(-F.2000$`sum(n)`),]
F.2000.top5 <- F.2000[c(1:5),]
year.4 <- rep(2000,5)
baby.2000.F <- cbind(year.4,F.2000.top5)

babynames.2000.M <-
  babynames %>%
  filter(year==2000,sex=="M")
M.2000 <-
  babynames.2000.M %>%
  group_by(name) %>%
  summarise(sum(n))
M.2000<- M.2000[order(-M.2000$`sum(n)`),]
M.2000.top5 <- M.2000[c(1:5),]
baby.2000.M <- cbind(year.4,M.2000.top5)

babytop5 <- cbind(baby.1880.F,baby.1880.M,baby.1920.F,baby.1920.M, baby.1960.F,baby.1960.M,baby.2000.F,

kable(babytop5)
```

year.1	name	sum(n)	year.1	name	sum(n)	year.2	name	sum(n)	year.2	name	sum
1880	Mary	7065	1880	John	9655	1920	Mary	70980	1920	John	56
1880	Anna	2604	1880	William	9532	1920	Dorothy	36643	1920	William	50
1880	Emma	2003	1880	James	5927	1920	Helen	35097	1920	Robert	48
1880	Elizabeth	1939	1880	Charles	5348	1920	Margaret	27997	1920	James	47
1880	Minnie	1746	1880	George	5126	1920	Ruth	26101	1920	Charles	28
What name	s overlap bo	ys and gi	rls?								

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

```
## [1] "John"    "William" "James"   "Charles" "George"  "Frank"
```

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)
notusedin21st <- setdiff(name19th$name,name21th$name)
head(notusedin21st)
```

```
## [1] "Bertie" "Nelle"  "Hulda"  "Mittie" "Myrtie" "Madge"
```

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

```
babynames.1880.2017 <- filter(babynames,year>=1880 & year<=2017)
n<-length(babynames$name)
babynames.1880.2017 <- filter(babynames.1880.2017,name=="Donald"|name=="Hilary"|name=="Hillary"|name==
ff<-babynames.1880.2017 %>%
  group_by(name) %>%
  summarise(sum(n)/length(babynames$name))

kable(ff)
```

name	sum(n)/length(babynames$name)
Donald	0.7360050
Hilary	0.0135655
Hillary	0.0154843
Joe	0.2400932