MA 615 HW2

Problem 1

Load the gapminder data from the gapminder package.

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1704 obs. of  6 variables:
##  $ country  : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ year     : int  1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...
##  $ lifeExp  : num  28.8 30.3 32 34 36.1 ...
##  $ pop      : int  8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 16317921 22
##  $ gdpPercap: num  779 821 853 836 740 ...
```

How many continents are included in the data set? There are 5 continents in this data set.

```
## [1] "Africa"   "Americas" "Asia"     "Europe"   "Oceania"
```

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

```
attach(dat.gap)
num.country<-unique(country)
num.of.num.country<-length(num.country)
num.of.num.country
```

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

```
dat.gap52<-subset(dat.gap, year == 1952)
dat.gap52 %>% group_by(dat.gap52$continent) %>%tally()
```

```
## # A tibble: 5 x 2
##   `dat.gap52$continent`     n
##   <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.

```
attach(dat.gap)
```

```
## The following objects are masked from dat.gap (pos = 3):
##
##     continent, country, gdpPercap, lifeExp, pop, year
```

```
tab_1<-dat.gap %>%
  group_by(continent) %>%
  summarise(pop.mean = mean(pop), gdpPercap.mean = mean(gdpPercap))
kable(tab_1, format = "latex", booktabs=TRUE, digits = 2,
      col.names = c("continent", "total population", "gdpPercap.mean"),
      caption = "Total population and GDP per capita by Continent")
```

```
knitr::kable(tab_1)
```

Table 1: Total population and GDP per capita by Continent

continent	total population	gdpPercap.mean
Africa	9916003	2193.75
Americas	24504795	7136.11
Asia	77038722	7902.15
Europe	17169765	14469.48
Oceania	8874672	18621.61

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

continent	gdpPercap.mean in 52
Africa	1252.57
Americas	4079.06
Asia	5195.48
Europe	5661.06
Oceania	10298.09

continent	pop.mean	gdpPercap.mean
Africa	9916003	2193.755
Americas	24504795	7136.110
Asia	77038722	7902.150
Europe	17169765	14469.476
Oceania	8874672	18621.609

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

```
tab_2<-dat.gap52 %>%
  group_by(continent) %>%
  summarise(gdpPercap.mean52 = mean(gdpPercap))
kable(tab_2, format = "latex", booktabs=TRUE, digits = 2,
      col.names = c("continent", "gdpPercap.mean in 52"),
      caption = "GDP per capita for the countries in each continent")
```

```
knitr::kable(tab_2)
```

continent	gdpPercap.mean52
Africa	1252.572
Americas	4079.063
Asia	5195.484
Europe	5661.057
Oceania	10298.086

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
dat.gap07<-filter(dat.gap, year == 2007)
tab_3<-dat.gap07 %>%
  group_by(continent) %>%
  summarise(gdpPercap.mean07 = mean(gdpPercap))
kable(tab_3, format = "latex", booktabs=TRUE, digits = 2,
      col.names = c("continent", "gdpPercap.mean in 07"),
      caption = "GDP per capita for the countries in each continent")
```

Table 3: GDP per capita for the countries in each continent

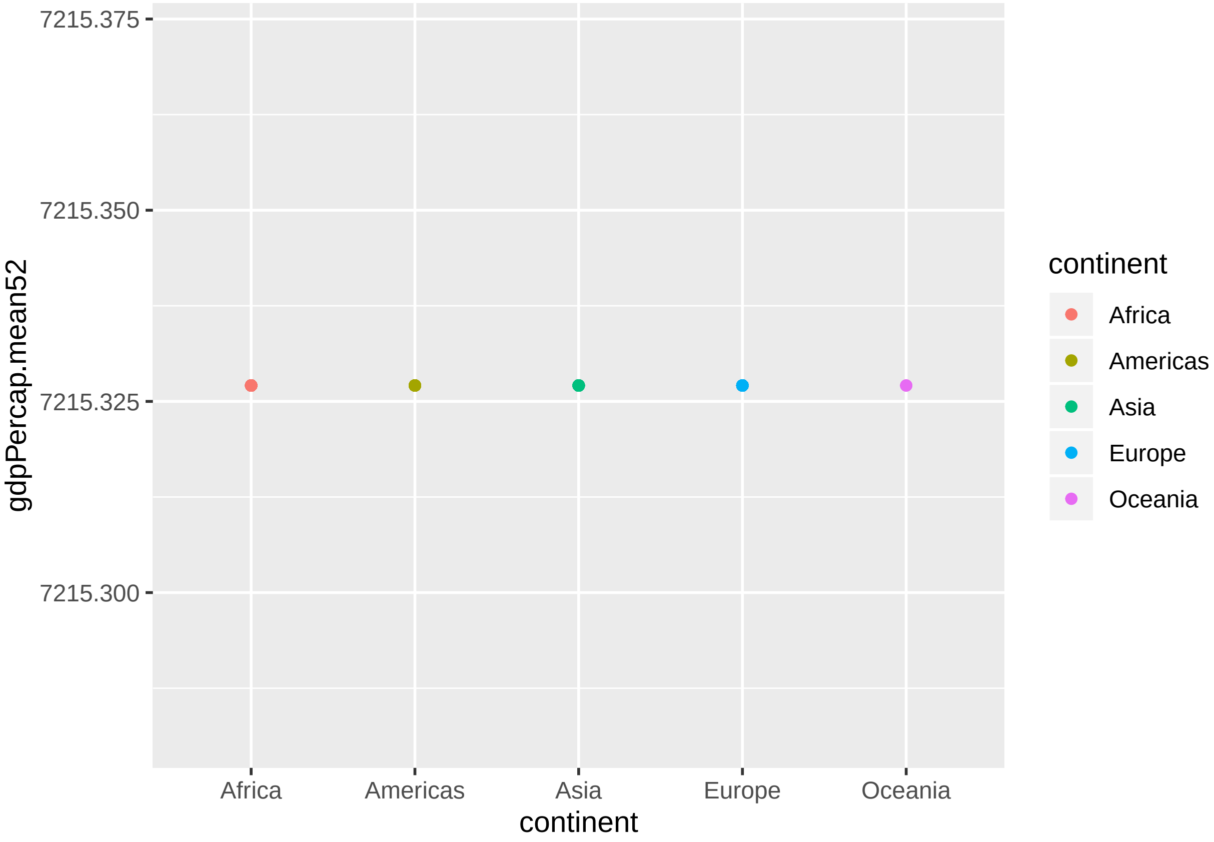
continent	gdpPercap.mean in 07
Africa	3089.03
Americas	11003.03
Asia	12473.03
Europe	25054.48
Oceania	29810.19

```
knitr::kable(tab_3)
```

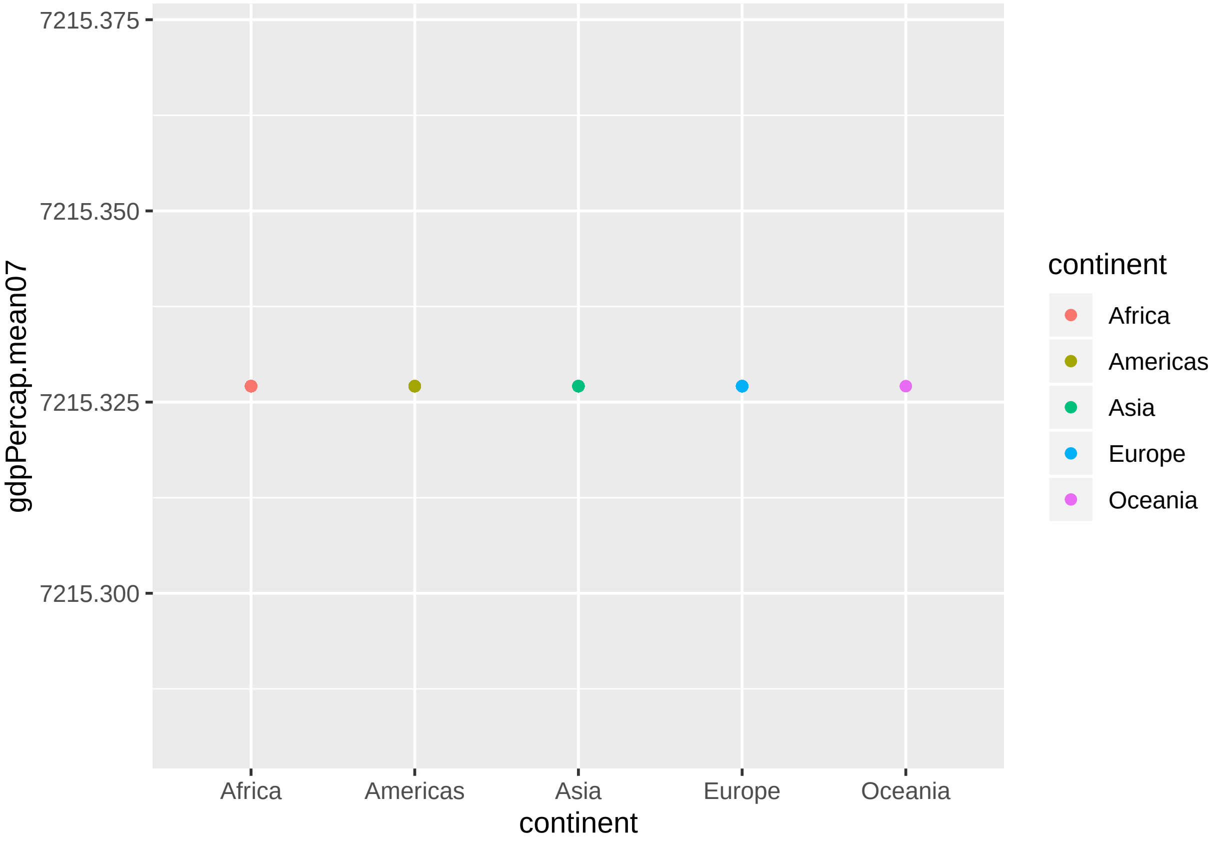
continent	gdpPercap.mean07
Africa	3089.033
Americas	11003.032
Asia	12473.027
Europe	25054.482
Oceania	29810.188

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

```
par(mfrow=c(2,2))
dat.gap52$gdpPercap.mean52=mean(gdpPercap)
ggplot(data = dat.gap52) +
    geom_point(mapping = aes(x = continent, y = gdpPercap.mean52,color=continent))
```



```
dat.gap07$gdpPercap.mean07=mean(gdpPercap)
ggplot(data = dat.gap07) +
  geom_point(mapping = aes(x = continent, y = gdpPercap.mean07,color=continent))
```



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

```
nrow(dat.gap)
```

```
## [1] 1704
```

```
for(i in 1:nrow(dat.gap)){
  a=dat.gap$pop[i]
  b=dat.gap$pop[i+1]
  c <- a/b
  dat.gap$negative_check[i]<-c
}
country.unique<-unique(dat.gap$country)
for(p in 1:length(country.unique)){
  if(dat.gap$country[p]!=dat.gap$country[p+1]){
    dat.gap$negative_check[p]<-0
  }
}

getcountry<-rep(0,length(dat.gap$negative_check)-1)
for(g in 1:length(getcountry)){
  if(dat.gap$negative_check[g]>1){
    getcountry<-dat.gap$country[g]
  }
}
getcountry
```

```
## [1] Zambia
## 142 Levels: Afghanistan Albania Algeria Angola Argentina ... Zimbabwe
```

```
getyear<-rep(0,length(dat.gap$negative_check)-1)
for(g in 1:length(getyear)){
  if(dat.gap$negative_check[g]>1){
    getyear<-dat.gap$year[g]
  }
}
getyear
```

```
## [1] 2007
```

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

```
for(r in 1:nrow(dat.gap)){
  num=(dat.gap$gdpPercap[r+1]-dat.gap$gdpPercap[r])
  den=dat.gap$gdpPercap[r]
  rate <- num/den
  dat.gap$rate.of.growth[r]<-rate
}
MAX.RATE=max(dat.gap$rate.of.growth,na.rm = TRUE)
MAX.RATE
```

```
## [1] 8.49069
```

```
get.max.country<-rep(0,length(dat.gap$rate.of.growth)-1)
for(m in 1:length(get.max.country)){
  if(dat.gap$rate.of.growth[m]==MAX.RATE){
    get.max.country<-dat.gap$country[m]
  }
}
get.max.country
```

```
## [1] Gambia
## 142 Levels: Afghanistan Albania Algeria Angola Argentina ... Zimbabwe
```

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 difference from women in 20s and women older than 29.

```
library("AER")
```

```
## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## Loading required package: lmtest

## Loading required package: zoo
```

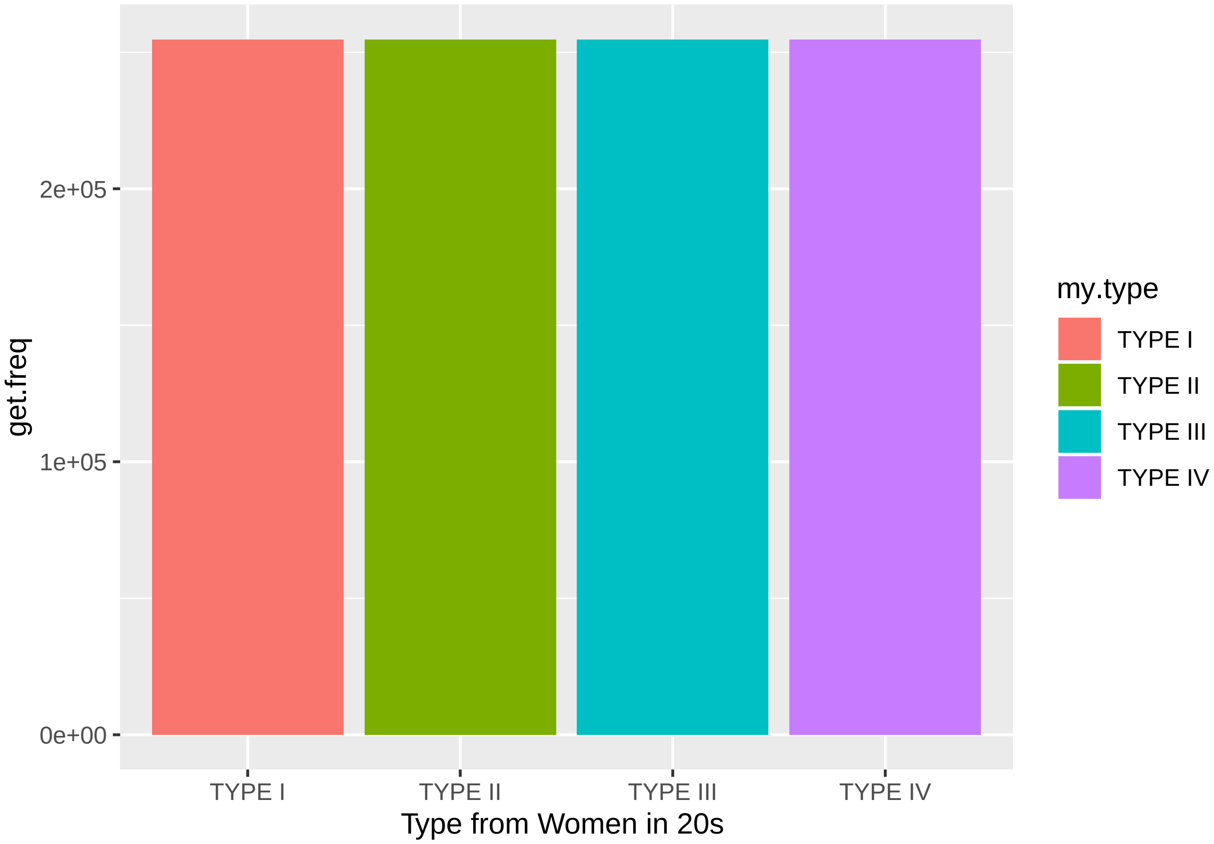
```
##
## Attaching package: 'zoo'

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

## Loading required package: sandwich

## Loading required package: survival
```

```
data(Fertility)
data.Fert<-Fertility
data.Fert$cat<-ifelse(data.Fert$age<30, data.Fert$cat<-"20s", data.Fert$cat<-"30+")
ty.1.20<-data.Fert[data.Fert$gender1=="male"&&data.Fert$gender2=="male"&&data.Fert$cat=="20s"]
ty.2.20<-data.Fert[data.Fert$gender1=="female"&&data.Fert$gender2=="male"&&data.Fert$cat=="20s"]
ty.3.20<-data.Fert[data.Fert$gender1=="male"&&data.Fert$gender2=="female"&&data.Fert$cat=="20s"]
ty.4.20<-data.Fert[data.Fert$gender1=="female"&&data.Fert$gender2=="female"&&data.Fert$cat=="20s"]
get.freq<-c(nrow(ty.1.20),nrow(ty.2.20),nrow(ty.3.20),nrow(ty.4.20))
my.type<-c("TYPE I","TYPE II","TYPE III","TYPE IV")
my.frame<-data.frame(get.freq,my.type)
```

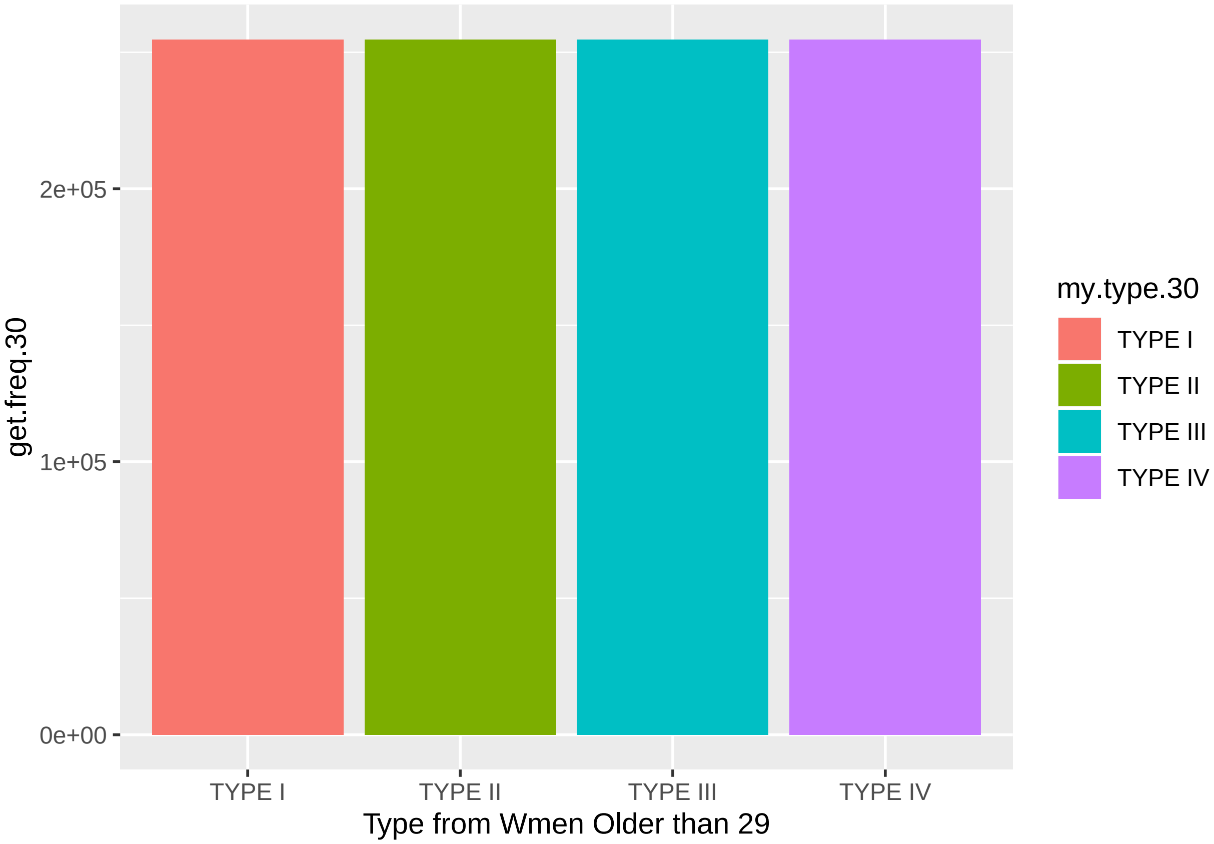
```
ty.1.30<-data.Fert[data.Fert$gender1=="male"&&data.Fert$gender2=="male"&&data.Fert$cat=="30+"]
ty.2.30<-data.Fert[data.Fert$gender1=="female"&&data.Fert$gender2=="male"&&data.Fert$cat=="30+"]
ty.3.30<-data.Fert[data.Fert$gender1=="male"&&data.Fert$gender2=="female"&&data.Fert$cat=="30+"]
ty.4.30<-data.Fert[data.Fert$gender1=="female"&&data.Fert$gender2=="female"&&data.Fert$cat=="30+"]
get.freq.30<-c(nrow(ty.1.30),nrow(ty.2.30),nrow(ty.3.30),nrow(ty.4.30))
my.type.30<-c("TYPE I","TYPE II","TYPE III","TYPE IV")
my.frame.30<-data.frame(get.freq.30,my.type.30)
```

Product a plot the contracts the frequency of these four combinations. Are the frequencies difference from women in 20s and women older than 29.

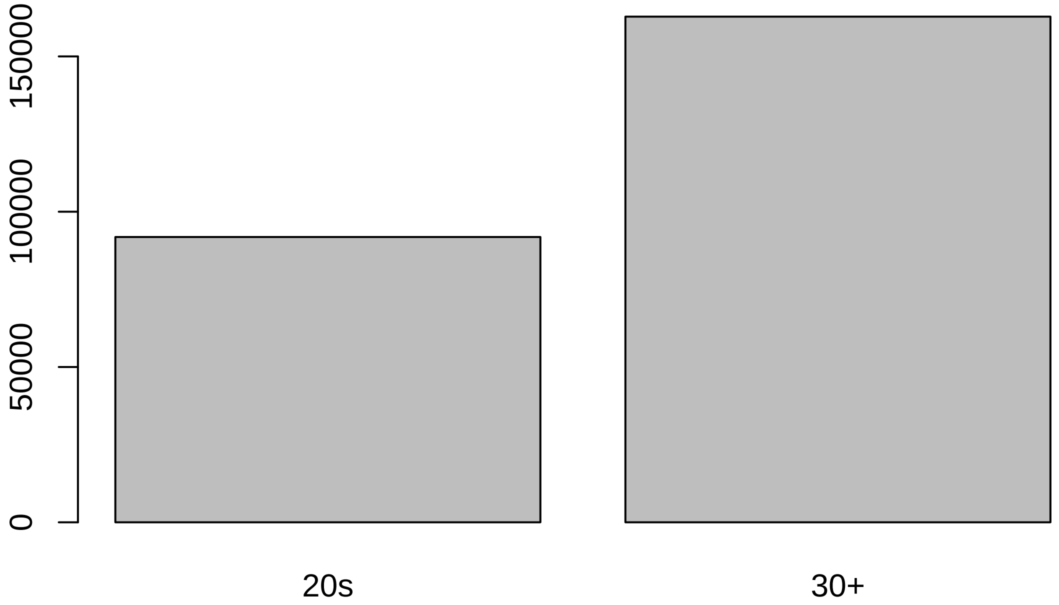
```
ggplot(data = my.frame) +
          geom_bar(
            mapping = aes(x = my.type, y = get.freq,fill=my.type), stat = "identity"
          )+xlab("Type from Women in 20s ")
```



```
ggplot(data = my.frame.30) +
          geom_bar(
            mapping = aes(x = my.type.30, y = get.freq.30,fill=my.type.30), stat = "identity"
          )+xlab("Type from Wmen Older than 29")
```

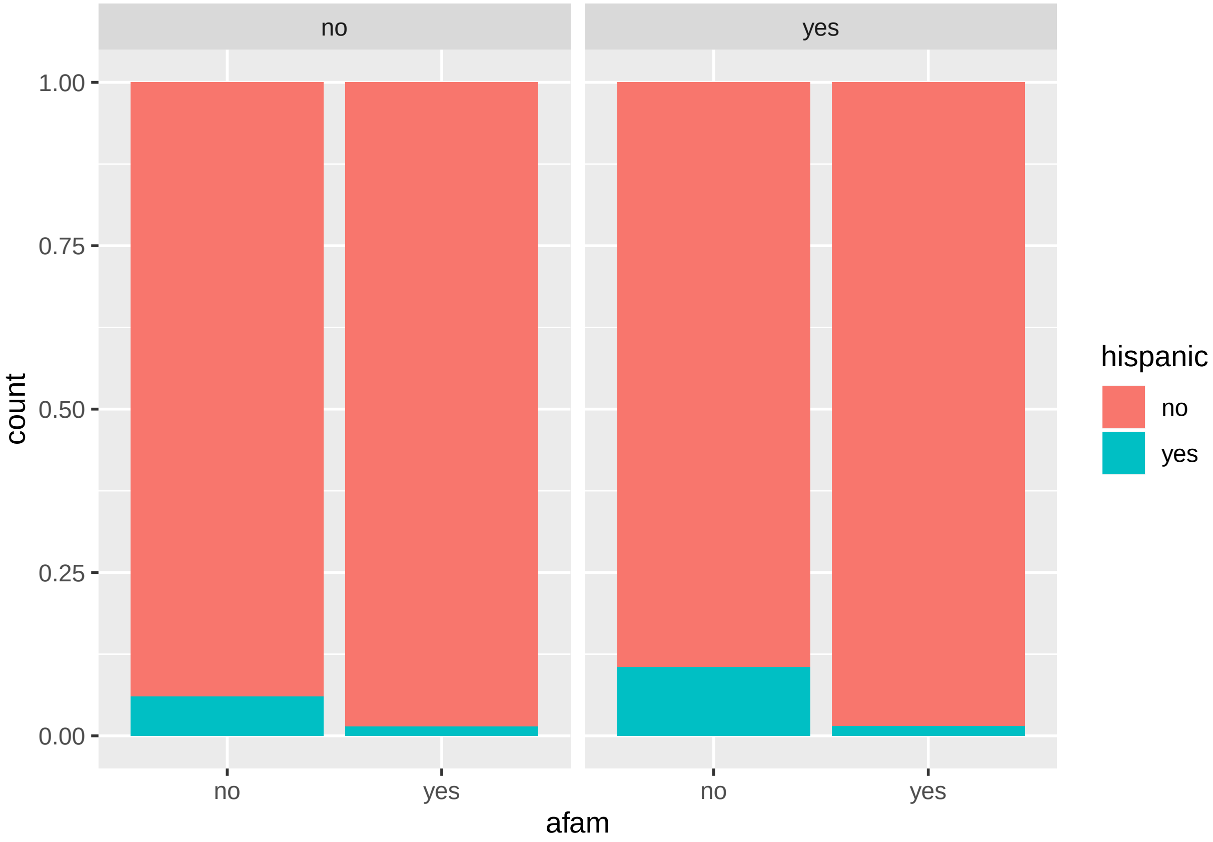


```
counts.age<-table(data.Fert$cat)
barplot(counts.age)
```



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

```
ggplot(data.Fert)+aes(x=afam,fill=hispanic)+geom_bar(position = "fill")+facet_grid(.~morekids)
```



##Problem 3 Use the mtcars and mpg datasets.

```
library("stringr")
data(mpg)
dat.mpg<-mpg
data(mtcars)
dat.mtcar<-mtcars
```

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

```
dat.mtcar <- cbind(dat.mtcar,names=row.names(dat.mtcar))
e<-str_count(dat.mtcar$names, "e")
e<-sum(e)
e
```

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

How many cars in mtcars have the brand Merc?

```
str_count(dat.mtcar$names, "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
```

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

```
dat.mpg$coutM<-ifelse(dat.mpg$manufacturer=="mercury", dat.mpg$coutM<-"T", dat.mpg$coutM<-"F")
table(dat.mpg$coutM)
```

```
##
##   F   T
## 230   4
```

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

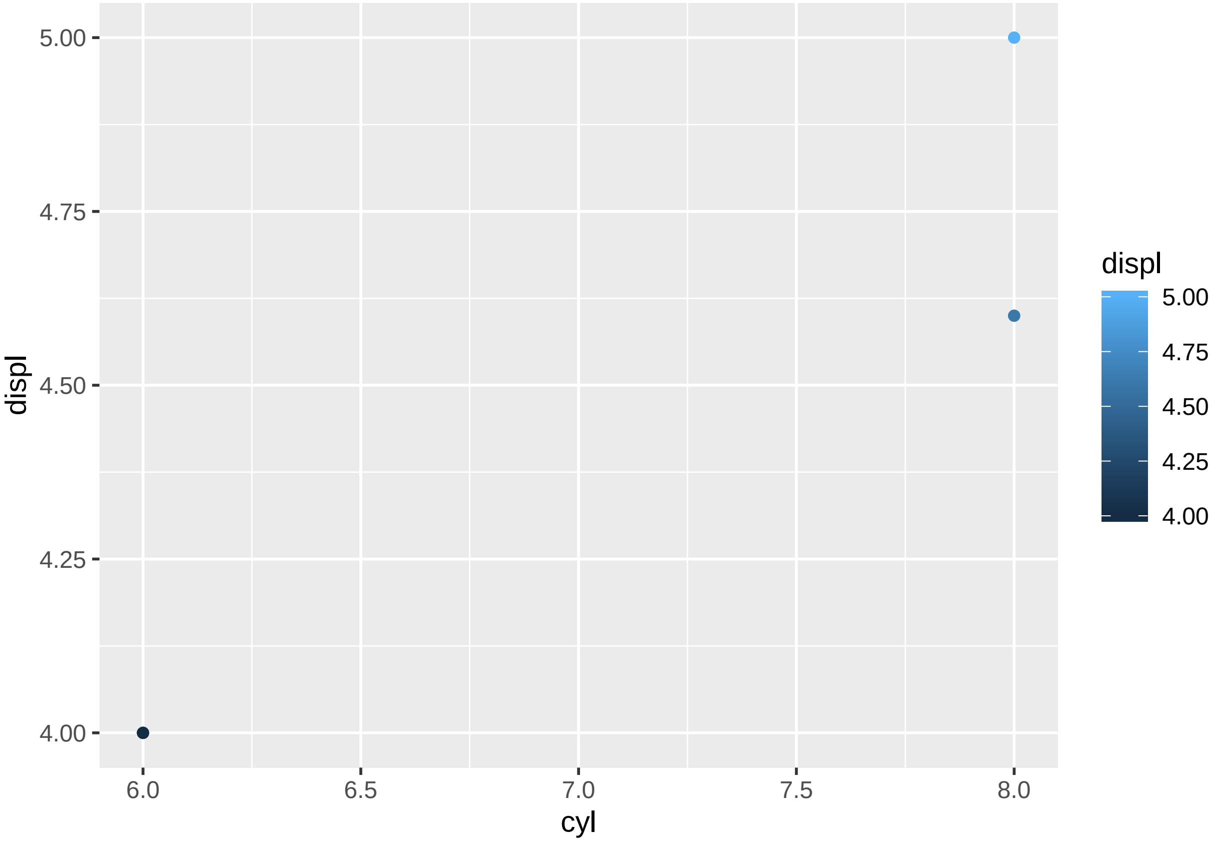
```
mpg.Merc<-dat.mpg[which(str_count(dat.mpg$manufacturer,"mercury")%in%c(1)),]
mtcars.Merc<-dat.mtcar[which(str_count(dat.mtcar$names,"Merc")%in%c(1)),]
knitr::kable(mpg.Merc)
```

manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class	coutM
mercury	mountaineer 4wd	4.0	1999	6	auto(l5)	4	14	17	r	suv	T
mercury	mountaineer 4wd	4.0	2008	6	auto(l5)	4	13	19	r	suv	T
mercury	mountaineer 4wd	4.6	2008	8	auto(l6)	4	13	19	r	suv	T
mercury	mountaineer 4wd	5.0	1999	8	auto(l4)	4	13	17	r	suv	T

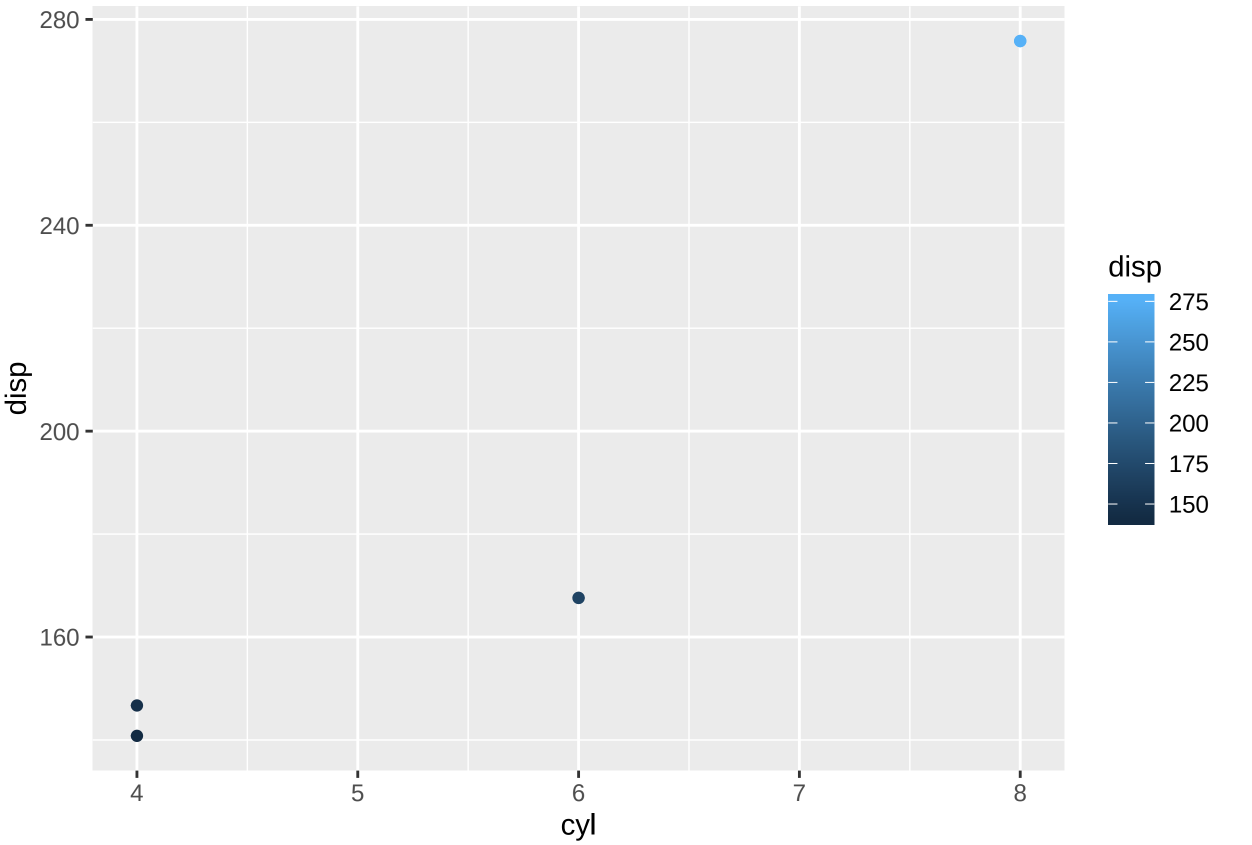
```
knitr::kable(mtcars.Merc)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb	names
Merc 240D	24.4	4	146.7	62	3.69	3.19	20.0	1	0	4	2	Merc 240D
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2	Merc 230
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4	Merc 280
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4	Merc 280C
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3	Merc 450SE
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3	Merc 450SL
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18.0	0	0	3	3	Merc 450SLC

```
par(mfrow=c(2,2))
ggplot(data = mpg.Merc) +
      geom_point(mapping = aes(x = cyl, y = displ,color=displ))
```



```
ggplot(data = mtcars.Merc) +
    geom_point(mapping = aes(x = cyl, y = disp,color=disp))
```



##Problem 4 Install the babynames package. Draw a sample of 500,000 rows from the babynames data

```
library("babynames")
data(babynames)
dat.baby<-babynames
dat.baby5<-dat.baby[sample(nrow(dat.baby),500000 ),]
```

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

```
top5.1880m<-filter(dat.baby5,dat.baby5$year=="1880")%>%
  group_by(sex,name)%>%summarise(total=sum(n))%>%arrange(desc(total))
 top5.1880m<-filter(top5.1880m, sex=="M")
```

```
top5.1880f<-filter(dat.baby5,dat.baby5$year=="1880")%>%
  group_by(sex,name)%>%summarise(total=sum(n))%>%arrange(desc(total))
 top5.1880f<-filter(top5.1880f, sex=="F")
```

```
top5.1920m<-filter(dat.baby5,dat.baby5$year=="1920")%>%
  group_by(sex,name)%>%summarise(total=sum(n))%>%arrange(desc(total))
 top5.1920m<-filter(top5.1920m, sex=="M")
```

```
top5.1880f<-filter(dat.baby5,dat.baby5$year=="1920")%>%
  group_by(sex,name)%>%summarise(total=sum(n))%>%arrange(desc(total))
 top5.1920f<-filter(top5.1880f, sex=="F")
```

```
top5.1960m<-filter(dat.baby5,dat.baby5$year=="1960")%>%
  group_by(sex,name)%>%summarise(total=sum(n))%>%arrange(desc(total))
 top5.1960m<-filter(top5.1960m, sex=="M")
```

```
top5.1960f<-filter(dat.baby5,dat.baby5$year=="1960")%>%
  group_by(sex,name)%>%summarise(total=sum(n))%>%arrange(desc(total))
 top5.1960f<-filter(top5.1960f, sex=="F")
```

```
top5.2000m<-filter(dat.baby5,dat.baby5$year=="2000")%>%
  group_by(sex,name)%>%summarise(total=sum(n))%>%arrange(desc(total))
 top5.2000m<-filter(top5.2000m, sex=="M")
```

```
top5.2000f<-filter(dat.baby5,dat.baby5$year=="2000")%>%
  group_by(sex,name)%>%summarise(total=sum(n))%>%arrange(desc(total))
 top5.2000f<-filter(top5.2000f, sex=="F")
```

```
m1880<-top5.1880m$total[1:5]
m1880n<-top5.1880m$name[1:5]
f1880<-top5.1880f$total[1:5]
f1880n<-top5.1880f$name[1:5]

m1920<-top5.1920m$total[1:5]
m1920n<-top5.1920m$name[1:5]
f1920<-top5.1920f$total[1:5]
f1920n<-top5.1920f$name[1:5]

m1960<-top5.1960m$total[1:5]
m1960n<-top5.1960m$name[1:5]
f1960<-top5.1960f$total[1:5]
f1960n<-top5.1960f$name[1:5]

m2000<-top5.2000m$total[1:5]
m2000n<-top5.2000m$name[1:5]
f2000<-top5.2000f$total[1:5]
f2000n<-top5.2000f$name[1:5]

top5.1800_20<-data.frame(cbind(m1880n,m1880,f1880n,f1880,m1920n,m1920,f1920n,f1920))
  colnames(top5.1800_20) <- c('1880 Male Name', '1880 Number',
                     "1880 Female Name", "1880 Number",
                     '1920 Male Name', '1920 Number',
                     "1920 Female Name", "1920 Number"
                     )
knitr::kable(top5.1800_20)
```

1880 Male Name	1880 Number	1880 Female Name	1880 Number	1920 Male Name	1920 Number	1920 Female
Frank	3242	James	47909	James	47909	Helen
Joseph	2632	Helen	35097	George	26893	Ruth
Robert	2415	George	26893	Edward	20095	Elizabeth
Arthur	1599	Ruth	26101	Frank	16432	Alice
Andrew	644	Edward	20095	Thomas	14938	Catherine

```
top5.1960_02<-data.frame(cbind(m1960n,m1960,f1960n,f1960,m2000n,m2000,f2000n,f2000))
colnames(top5.1960_02) <- c('1996 Male Name', '1996 Number',
                     "1996 Female Name", "1996 Number",
                     '2000 Male Name', '2000 Number',
                     "2000 Female Name", "2000 Number"
                     )
knitr::kable(top5.1960_02)
```

1996 Male Name	1996 Number	1996 Female Name	1996 Number	2000 Male Name	2000 Number	2000 Female
David	85928	Cynthia	26725	Jacob	34471	Ashley
Michael	84183	Barbara	24444	Christopher	24931	Alexis
Kevin	28388	Denise	15065	Daniel	22312	Samantha
Paul	25639	Cindy	14949	Ryan	20264	Brianna
Donald	22731	Kim	12474	Dylan	15401	Olivia

What names overlap boys and girls?

```
overlapF<-subset(dat.baby5,dat.baby5$sex=="F",select = "name")
overlapM<-subset(dat.baby5,dat.baby5$sex=="M",select = "name")
overlap<-inner_join(overlapF, overlapM,by="name",copy=TRUE)
u.overlap<-unique(overlap)
u.overlap
```

```
## # A tibble: 7,420 x 1
##    name
##    <chr>
##  1 Justina
##  2 Ashlee
##  3 Sekai
##  4 Bettie
##  5 Alizae
##  6 Courtney
##  7 Rayne
##  8 Kevin
##  9 Li
## 10 Selma
## # ... with 7,410 more rows
```

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

```
dat.baby19<-subset(dat.baby,dat.baby$year==1880,select="name")
dat.baby21<-subset(dat.baby,dat.baby$year==2000,select="name")
overlapname<-inner_join(dat.baby19, dat.baby21,by="name",copy=TRUE)
notused<-dat.baby21 [! dat.baby19 %in% overlapname]
notused
```

```
## # A tibble: 29,769 x 1
##    name
##    <chr>
##  1 Emily
##  2 Hannah
##  3 Madison
##  4 Ashley
##  5 Sarah
##  6 Alexis
##  7 Samantha
##  8 Jessica
##  9 Elizabeth
## 10 Taylor
## # ... with 29,759 more rows
```

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

```
checknamesd<-filter(dat.baby,dat.baby$name=="Donald")
d<-length(checknamesd$name)
d
```

```
## [1] 226
```

```
checknamesh1<-filter(dat.baby,dat.baby$name=="Hilary")
h<-length(checknamesh1$name)
h
```

```
## [1] 193
```

```
checknamesh2<-filter(dat.baby,dat.baby$name=="Hillary")
h2<-length(checknamesh2$name)
h2
```

```
## [1] 174
```

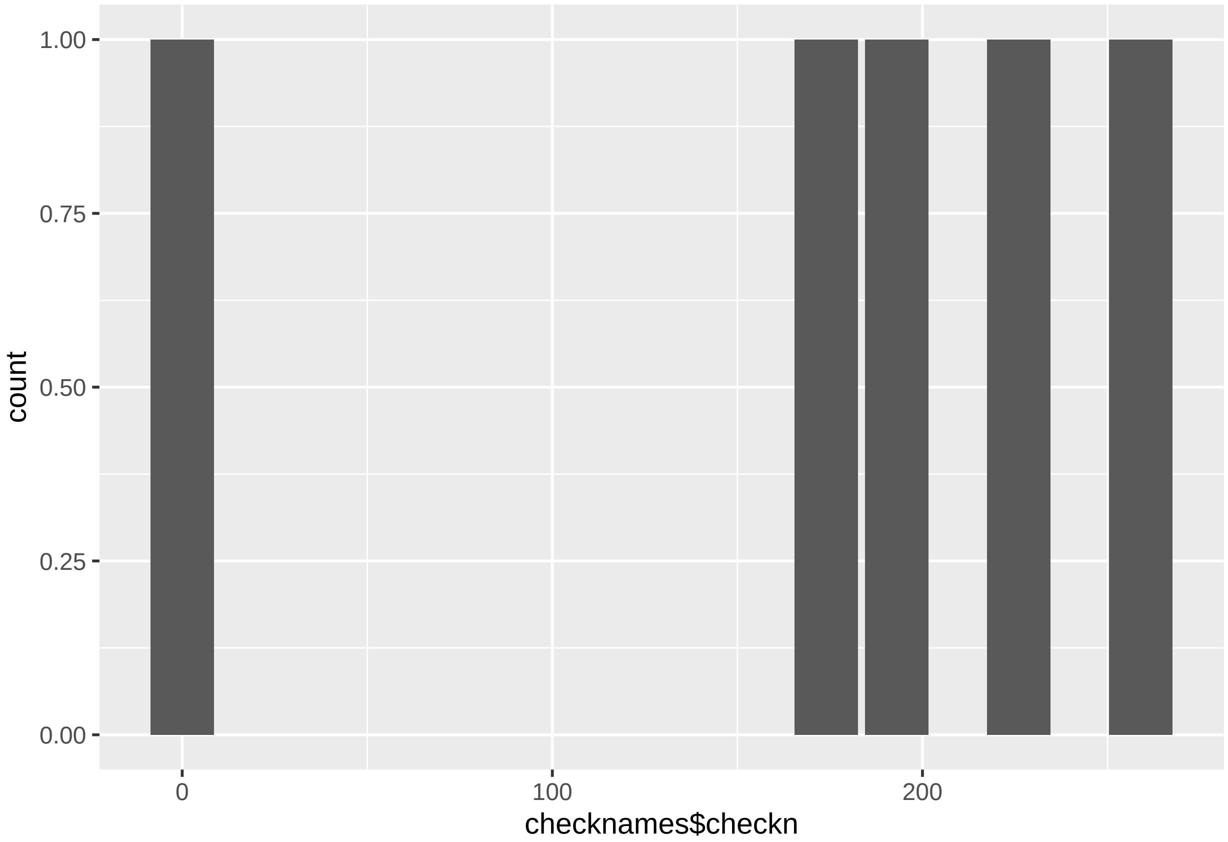
```
checknamesj<-filter(dat.baby,dat.baby$name=="Joe")
j<-length(checknamesj$name)
j
```

```
## [1] 259
```

```
checknamesb<-filter(dat.baby,dat.baby$name=="Barrack")
b<-length(checknamesb$name)
b
```

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

```
checkn<-c(d,h,h2,j,b)
checkname<-c("Donald", "Hilary", "Hillary", "Joe","Barrack")
checknames<-data.frame(checkn,checkname)
ggplot(checknames)+geom_bar(position = "fill")+aes(x=checknames$checkn)
```


1.00
0.75
0.50
0.25
0.00
count
0
100
200
checknames$checkn