CMDA-3654

Homework 3

Eduardo Salvador

Due as a .pdf upload

Instructions:

Delete this Instructions section from your write-up!!

I have given you this assignment as an .Rmd (R Markdown) file.

- Change the name of the file to: `Lastname_Firstname_CMDA_3654_HW3.Rmd` , and your output should therefore match but with a `.pdf` extension.
- You need to edit the R Markdown file by adding chunks and filling them in appropriately with your code. Output will be generated automatically when you compile the document.
- You also need to add your own text before and after the chunks to explain what you are doing or to interpret the output.

Required: The final product that you turn in must be a .pdf file.

- You MUST Knit this document directly to a PDF, you are not allowed to knit to any other file type and then convert.

This assignment is to be done using Base R methods only!

The next assignment is devoted completely to plotting using ggplot2, so the use of ggplot2 is not allowed here.

Problem 1: (30 pts) Basic Summaries and Plotting with `Base R`

Install and load the `MASS` package for this problem, and load the `birthwt` data set that comes installed with `MASS` . This data set contains information on infant birth weight as well as observed risk factors. To find out more about this data set, see the help page `?birthwt` . In the following exercises, be sure to create an appropriate legend when neccesary, and label all axes and plots accordingly. ?birthwt a. Provide univariate summaries for the variables in this data set.

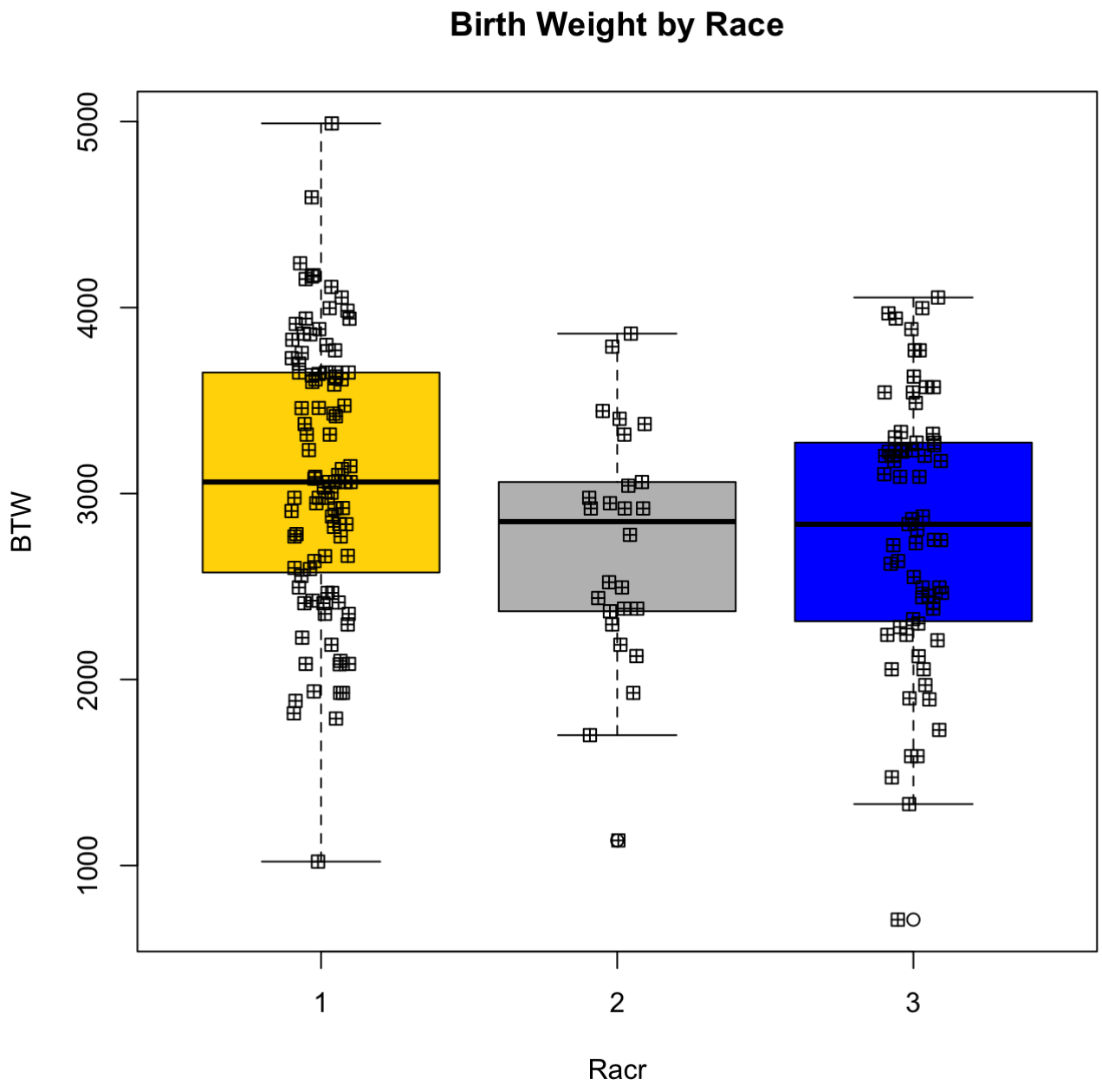
```
library(MASS)
#Univariate summary
for (i in 1:ncol(birthwt)){
  cat("variable:",colnames(birthwt[i]),"\n")
  print(summary(birthwt[,i]))}
```

```
variable: low
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.3122  1.0000  1.0000 
variable: age
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  14.00   19.00   23.00   23.24   26.00   45.00 
variable: lwt
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   80.0   110.0   121.0   129.8   140.0   250.0 
variable: race
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   1.000   1.000   1.847   3.000   3.000 
variable: smoke
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.3915  1.0000  1.0000 
variable: ptl
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.1958  0.0000  3.0000 
variable: ht
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.00000 0.00000 0.00000 0.06349 0.00000 1.00000 
variable: ui
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.1481  0.0000  1.0000 
variable: ftv
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.7937  1.0000  6.0000 
variable: bwt
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    709    2414    2977    2945    3487    4990 
```

```
  cat("\n")
```

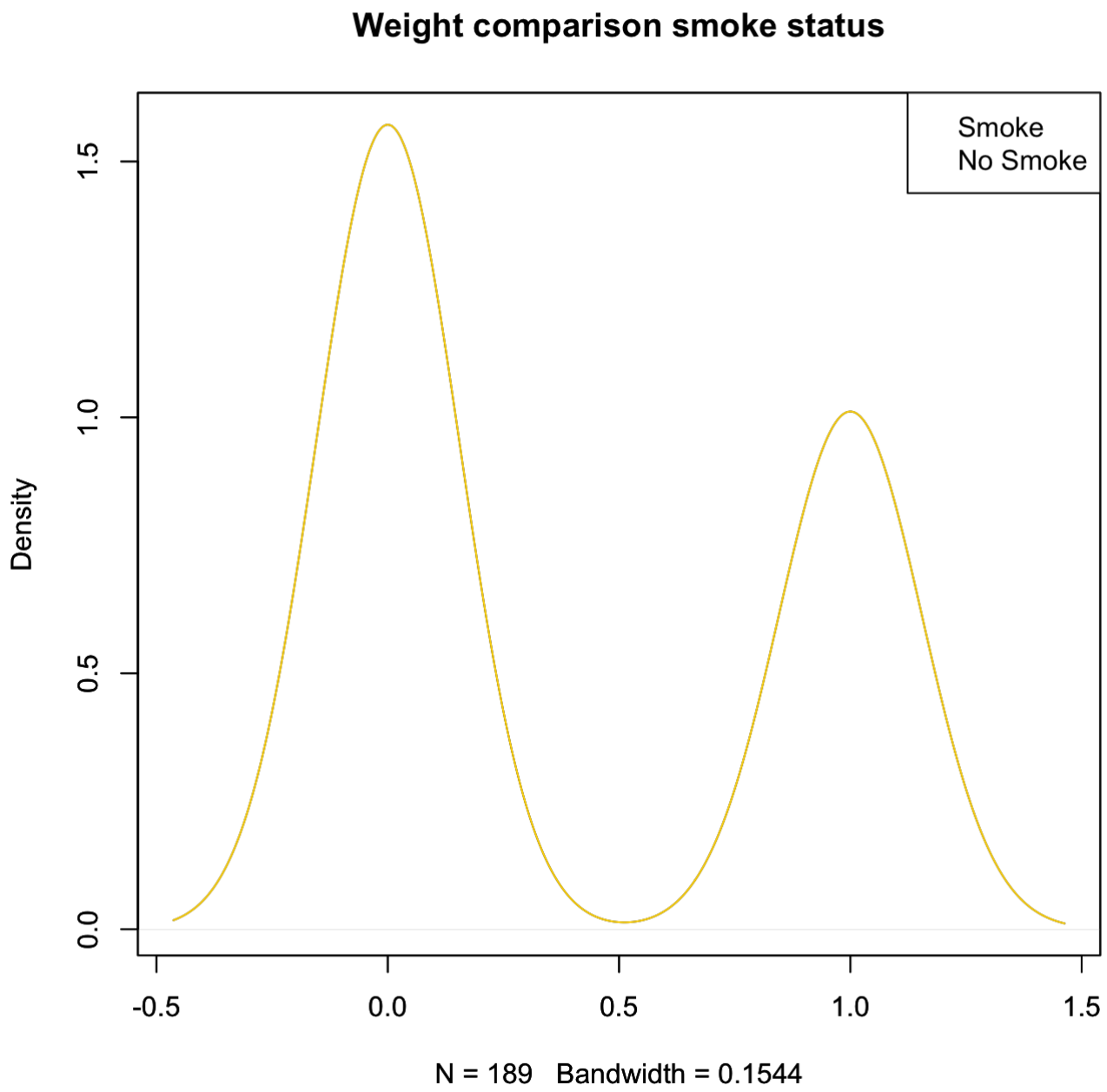
b. Create a boxplot of birth weight (`bwt`) by `race` . Notice that the variable race is numerically coded. Make sure to assign the proper factor names when creating your plot. You should use different colors for each boxplot. Overlay a jittered stripcharts.

```
#Create plot
boxplot(birthwt$bwt~birthwt$race,
        xlab="Racr",
        ylab="BTW",
        main="Birth Weight by Race",
        col=c("gold", "grey", "blue"),
        alpha=0.3)
stripchart(birthwt$bwt~birthwt$race, vertical=T, data=birthwt,method="jitter",pch=12,add
=T)
```



c. Create an overlayed density plot of birth weight given the smoking status of the mother, that is, make sure both densities are displayed onto the same plot. Use different colors and a legend.

```
#Creating density plot
plot(density(birthwt$smoke),
     main = "Weight comparison smoke status")
par(new=T)
plot(density(birthwt$smoke),
     main="",
     col="gold")
legend("topright",c("Smoke","No Smoke"),col=c("blue","purple"))
```

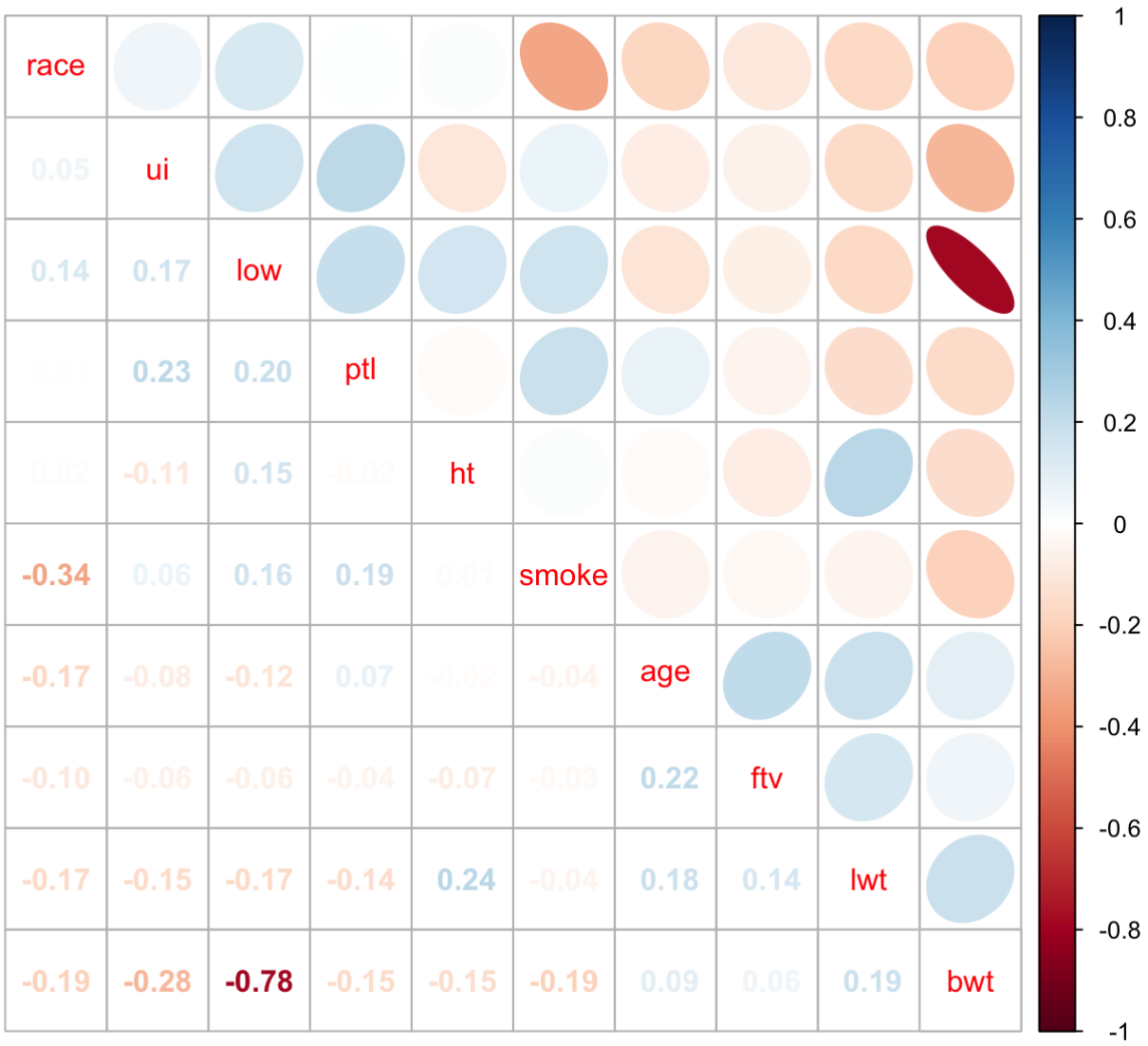


d. Create a correlogram for all quantitative variables and comment on what you observe.

```
library(corrplot)
round( M <- cor(birthwt[, c("low", "age", "lwt", "race","smoke","ptl","ht","ui","ftv","b
wt")]),3)
```

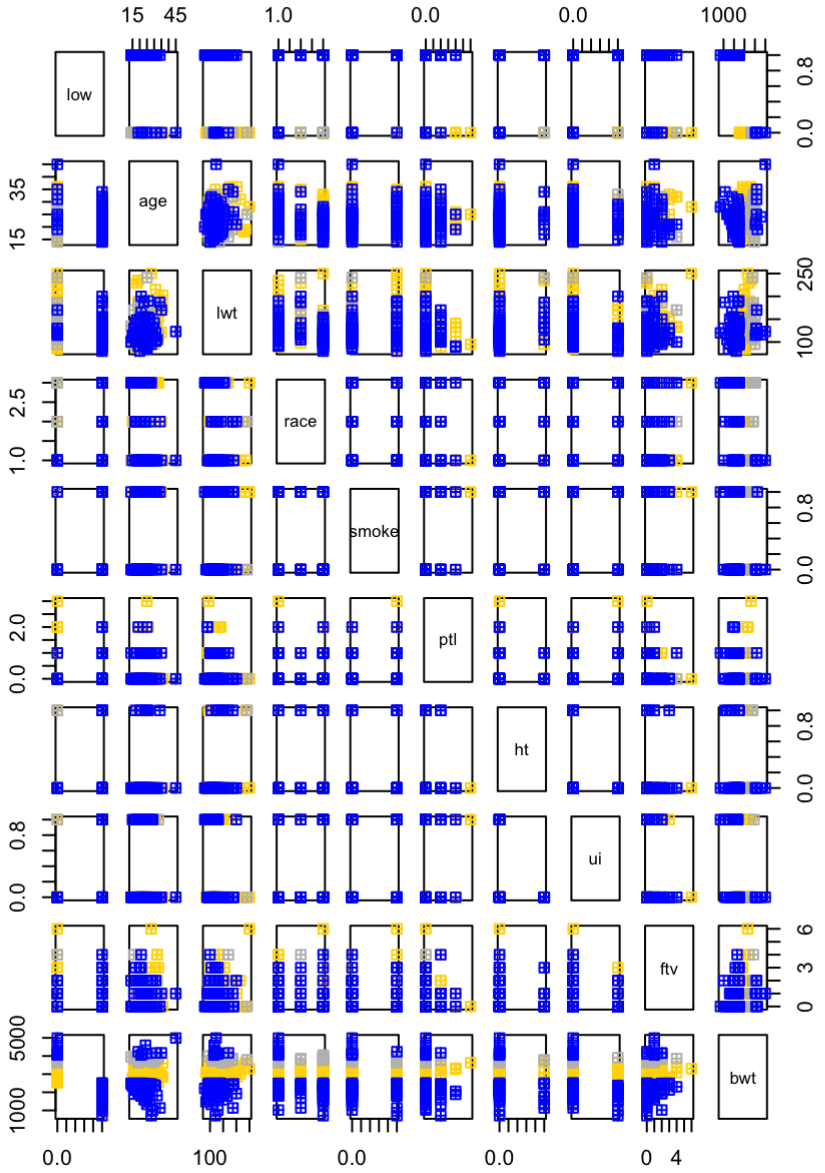
```
         low    age    lwt   race  smoke    ptl     ht     ui    ftv    bwt
low    1.000 -0.119 -0.170  0.138  0.161  0.196  0.152  0.169 -0.063 -0.785
age   -0.119  1.000  0.180 -0.173 -0.044  0.072 -0.016 -0.075  0.215  0.090
lwt   -0.170  0.180  1.000 -0.165 -0.044 -0.140  0.236 -0.153  0.141  0.186
race   0.138 -0.173 -0.165  1.000 -0.339  0.008  0.020  0.054 -0.098 -0.195
smoke  0.161 -0.044 -0.044 -0.339  1.000  0.188  0.013  0.062 -0.028 -0.190
ptl    0.196  0.072 -0.140  0.008  0.188  1.000 -0.015  0.228 -0.044 -0.155
ht     0.152 -0.016  0.236  0.020  0.013 -0.015  1.000 -0.109 -0.072 -0.146
ui     0.169 -0.075 -0.153  0.054  0.062  0.228 -0.109  1.000 -0.060 -0.284
ftv   -0.063  0.215  0.141 -0.098 -0.028 -0.044 -0.072 -0.060  1.000  0.058
bwt   -0.785  0.090  0.186 -0.195 -0.190 -0.155 -0.146 -0.284  0.058  1.000
```

```
corrplot.mixed(M,upper="ellipse",order="AOE")
```



e. Make a scatterplot matrix using the `pairs()` function for all numeric variables. Color the points in the scatterplot matrix using different colors depending on race. A legend might be kinda tricky in this case, but not impossible. I'll settle for a description of which groups the colors represent.

```
#selecting colums to get numbers
SCol<-birthwt[,c("low", "age", "lwt", "race","smoke","ptl","ht","ui","ftv","bwt")]

#Coloring scatterplot points
Coloring<-vector()
Coloring[1:96]<-"gold"
Coloring[97:122]<-"grey"
Coloring[123:189]<-"blue"

#Creating graph with legend and to fit
par(xpd=T)
pairs(SCol, pch=12, col=Coloring,oma=c(4,5,6,19))
legend("topright",pch<-12,legend = c("Race 1","Race 2", "Race 3"), col<-c("gold","grey",
"blue"))
```



Problem 2: (30 pts) Census Data

Turn your attention to the `adult.csv` data set..

a. Provide univariate summaries for the variables in this data set.

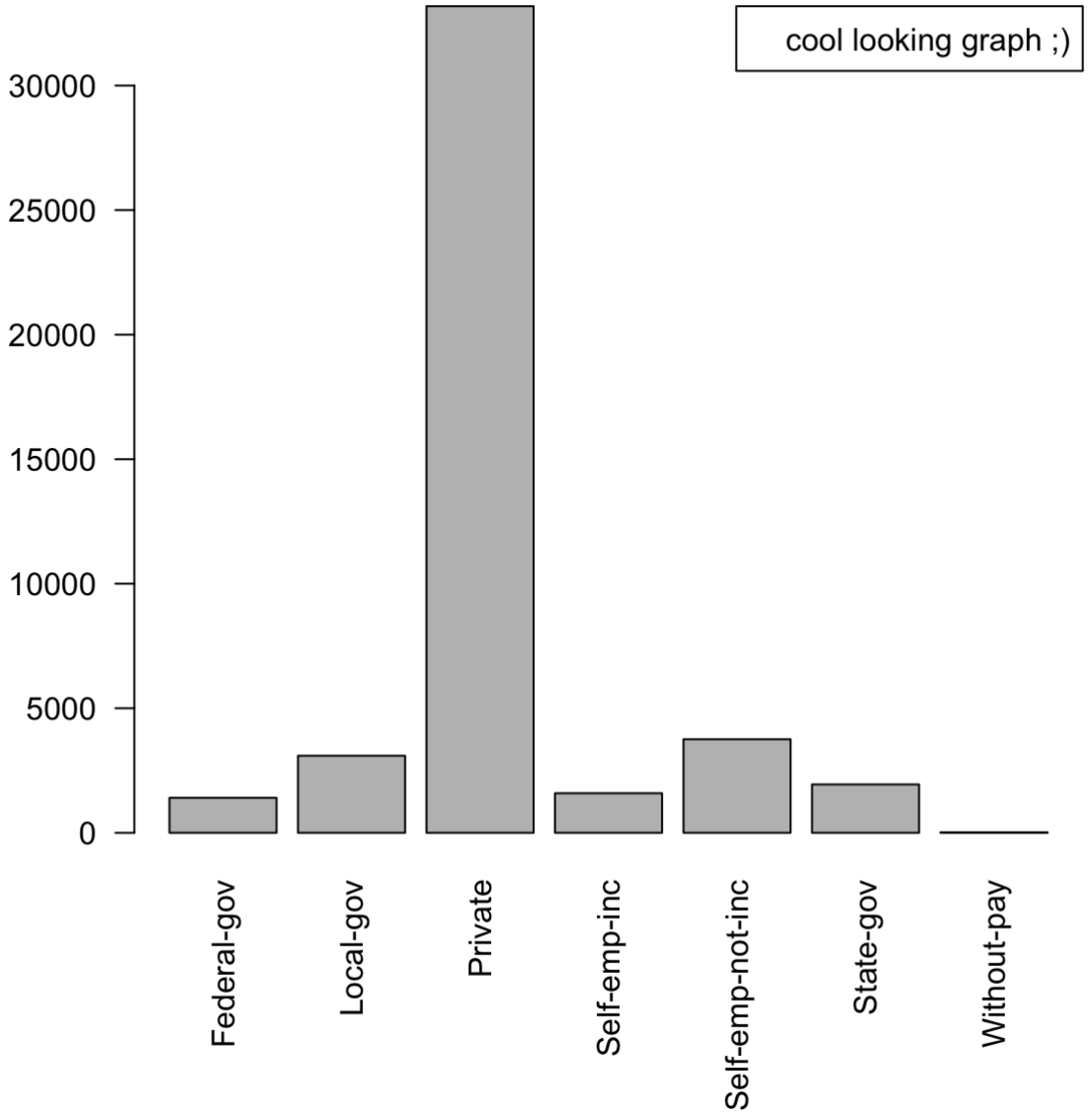
```
#Read
adult<-read.csv(file = "/Users/eduardosalvador/Desktop/FINAL\ Spring\ Semester\ 2021/CMD
A\ /Assignments/HW3/adult.csv")

#Univariate summary
for (i in 1:ncol(adult)){
  cat("variable:",colnames(adult[i]),"\n")
  print(summary(adult[,i]))
  cat("\n")
}
```

```
variable: age
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.00   28.00   37.00   38.51   47.00   90.00

variable: workclass
   Length     Class      Mode
    44993 character character

variable: fnlwgt
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13492  117392  178322  189757  237943 1490400

variable: education
   Length     Class      Mode
    44993 character character

variable: marital
   Length     Class      Mode
    44993 character character

variable: occupation
   Length     Class      Mode
    44993 character character

variable: relationship
   Length     Class      Mode
    44993 character character

variable: race
   Length     Class      Mode
    44993 character character

variable: sex
   Length     Class      Mode
    44993 character character

variable: capgain
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     0.0     0.0   598.1     0.0 41310.0

variable: caploss
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    0.00   89.05    0.00 4356.00

variable: hourspweek
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   40.00   40.00   40.89   45.00   99.00

variable: native
   Length     Class      Mode
    44993 character character
```

```
variable: pay
   Length     Class      Mode
    44993 character character
```

b. Create a bar chart displaying the counts of working class for all United States citizens.

```
#Fitting the name of each working class
par(mar=c(8,7,5,3))
#las=2 sets the bar horizontally
barplot(table(adult$workclass),las=2)
#Fitting back the values
par(mar=c(5.1,4.1,4.1,2.1))
#Setting unnecessary label
legend("topright",c("cool looking graph ;)"))
```



c. Make a bivariate frequency table for the `workclass` variable as the rows and `race` as the columns. Show this table. In a second table, show the same table but with the marginal frequencies added.

```
#Creating table
bivariantet<-table(adult$workclass,adult$race)
bivariantet
```

```
                 Amer-Indian-Eskimo Asian-Pac-Islander Black Other White
  Federal-gov                    33                 56   254    10  1051
  Local-gov                      65                 57   427    12  2533
  Private                       278                950  3141   298 28520
  Self-emp-inc                    2                 54    37     4  1492
  Self-emp-not-inc               34                 89   126    13  3494
  State-gov                      23                 85   233    13  1588
  Without-pay                     0                  1     1     0    19
```

```
#Adding margins
addmargins(bivariantet)
```

```
                 Amer-Indian-Eskimo Asian-Pac-Islander Black Other White
  Federal-gov                    33                 56   254    10  1051
  Local-gov                      65                 57   427    12  2533
  Private                       278                950  3141   298 28520
  Self-emp-inc                    2                 54    37     4  1492
  Self-emp-not-inc               34                 89   126    13  3494
  State-gov                      23                 85   233    13  1588
  Without-pay                     0                  1     1     0    19
  Sum                           435               1292  4219   350 38697

                    Sum
  Federal-gov      1404
  Local-gov        3094
  Private         33187
  Self-emp-inc     1589
  Self-emp-not-inc 3756
  State-gov        1942
  Without-pay        21
  Sum             44993
```

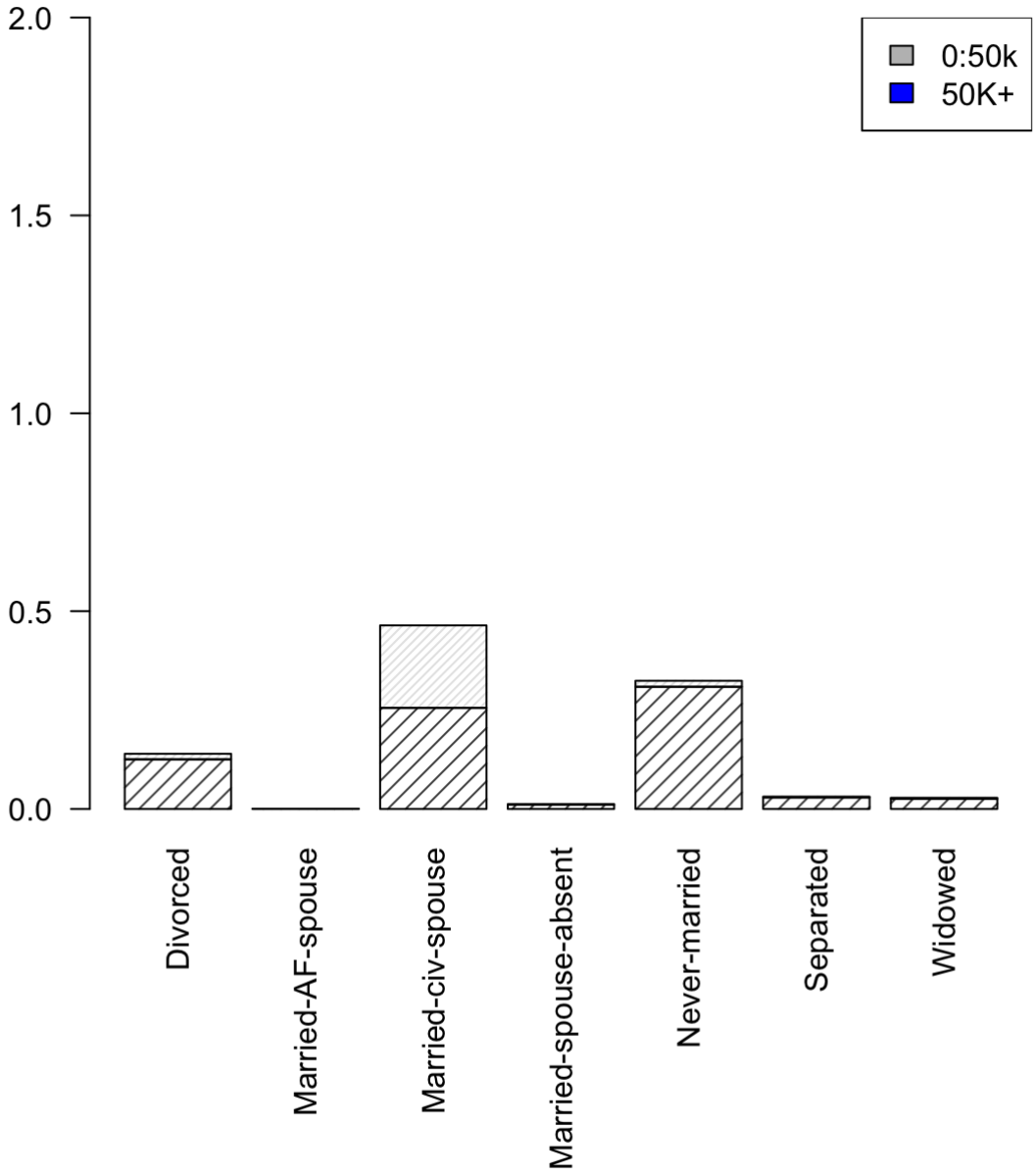
d. Make a three-way frequency table using the `xtabs()` function for the `workclass`, `race`, and `sex` variable (have sex be the 3rd dimension). Then use `ftable()` to flatten the 3-D table.

```
#three-way table
three_way<-xtabs(~race+workclass+sex,data=adult)
ftable(three_way)
```

```
                                  sex Female  Male
race                 workclass
Amer-Indian-Eskimo   Federal-gov              19    14
                     Local-gov                31    34
                     Private                 100   178
                     Self-emp-inc              1     1
                     Self-emp-not-inc          6    28
                     State-gov                 9    14
                     Without-pay               0     0
Asian-Pac-Islander   Federal-gov              15    41
                     Local-gov                19    38
                     Private                 344   606
                     Self-emp-inc             12    42
                     Self-emp-not-inc         15    74
                     State-gov                30    55
                     Without-pay               0     1
Black                Federal-gov             120   134
                     Local-gov               223   204
                     Private                1555  1586
                     Self-emp-inc              6    31
                     Self-emp-not-inc         37    89
                     State-gov               138    95
                     Without-pay               0     1
Other                Federal-gov               5     5
                     Local-gov                 3     9
                     Private                 109   189
                     Self-emp-inc              0     4
                     Self-emp-not-inc          1    12
                     State-gov                 7     6
                     Without-pay               0     0
White                Federal-gov             285   766
                     Local-gov               966  1567
                     Private                9299 19221
                     Self-emp-inc            180  1312
                     Self-emp-not-inc        550  2944
                     State-gov               570  1018
                     Without-pay               7    12
```

e. Create a **relative frequency stacked barchart** displaying the counts of `pay` categories with respect to the `marital` category. .

```
#Create table
btable=table(adult$pay, adult$marital)

rfreqtable<-prop.table(btable)

#Boundaries
par(mar=c(12,8,2,2))
#Setting it horizontal and adding labels
barplot(rfreqtable,las=2, density=c(15,30,45),ylim=c(0,2))
legend("topright",legend=c("0:50k","50K+"),fill=c("grey","blue"))
```



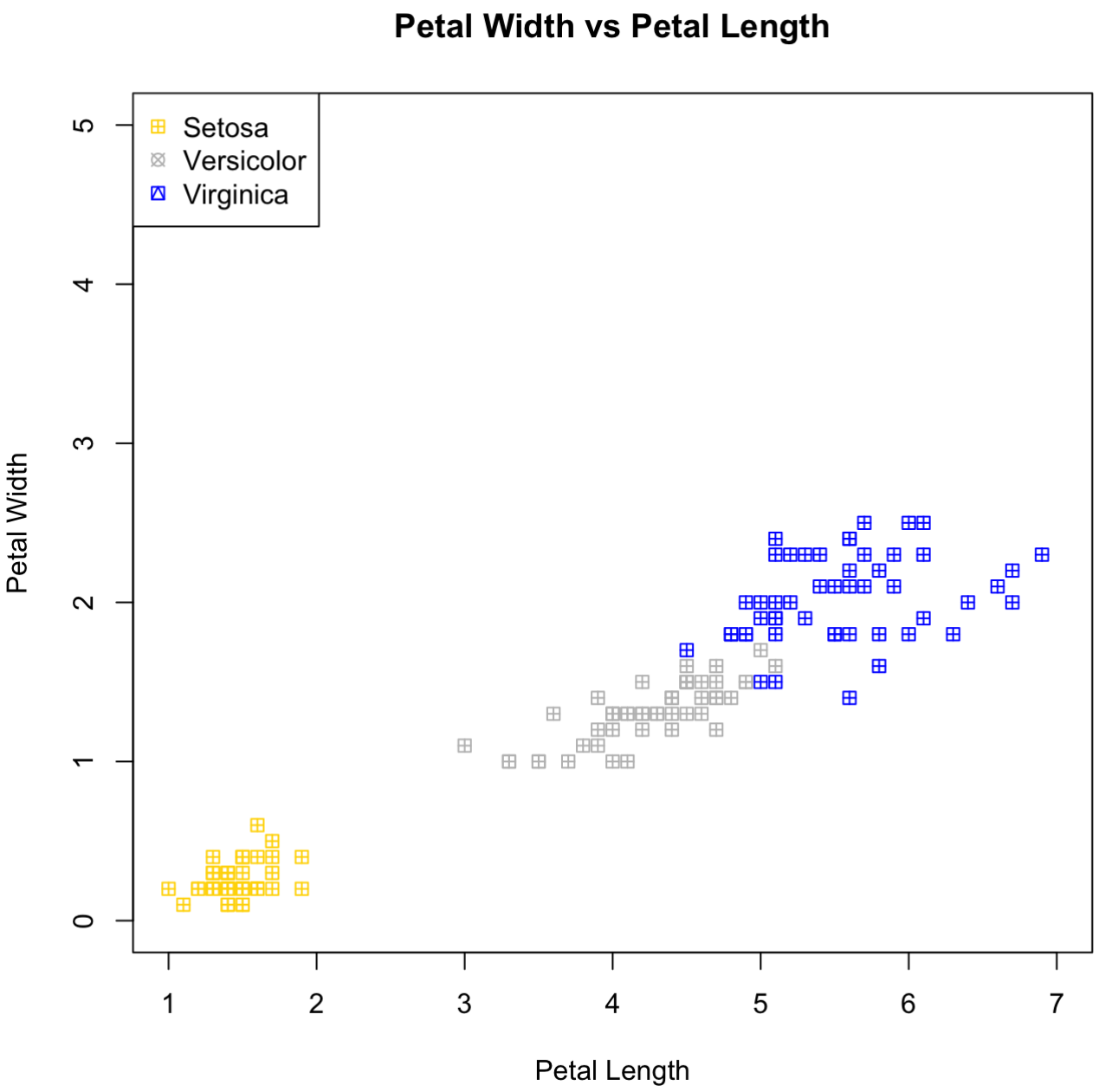

```
#Fitting back values
par(mar=c(5.1,4.1,4.1,2.1))
```

Problem 3: (20 pts) The `iris` dataset

(**Note:** When we say plot **"a" vs "b"**, by default "a" is on the y-axis, and "b" is on the x-axis.)

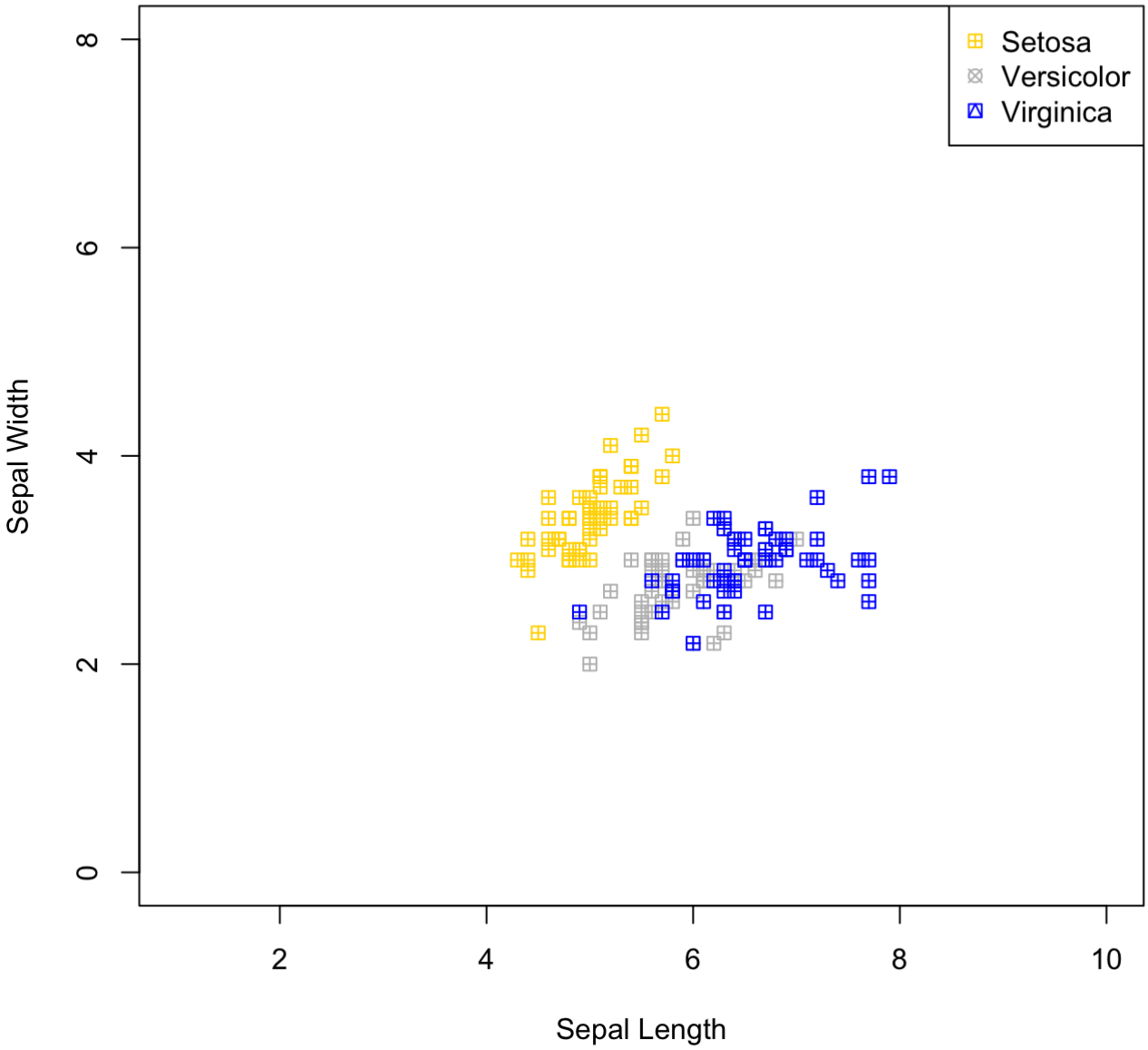
a. Plot the Petal Width vs Petal Length with different colors and plot characters for the different classes of plants. Be sure to add a legend.

```
library(datasets)
data("iris")

plot(iris$Petal.Length[1:50],iris$Petal.Width[1:50],
     main="Petal Width vs Petal Length",
     ylab="Petal Width",
     xlab="Petal Length",
     xlim=c(1,7),
     pch=12,
     col="gold",
     ylim=c(0,5))

points(iris$Petal.Length[51:100],iris$Petal.Width[51:100],
     main="Petal Width vs Petal Length",
     ylab="Petal Width",
     xlab="Petal Length",
     xlim=c(3,7),
     pch=12,
     col="grey",
     ylim=c(0,5))

points(iris$Petal.Length[101:150],iris$Petal.Width[101:150],
     main="Petal Width vs Petal Length",
     ylab="Petal Width",
     xlab="Petal Length",
     xlim=c(3,7),
     pch=12,
     col="blue",
     ylim=c(0,5))

legend("topleft",legend=c("Setosa","Versicolor","Virginica"),pch=c(12,13,14),col=c("gol
d","grey","blue"))
```



b. Plot the Sepal Width vs Sepal Length with different colors and plot characters for the different classes of plants. Be sure to add a legend.

```
plot(iris$Sepal.Length[1:50],iris$Sepal.Width[1:50],
     main="Sepal Width vs Sepal Length",
     ylab="Sepal Width",
     xlab="Sepal Length",
     xlim=c(1,10),
     pch=12,
     col="gold",
     ylim=c(0,8))

points(iris$Sepal.Length[51:100],iris$Sepal.Width[51:100],
     main="Sepal Width vs Sepal Length",
     ylab="Sepal Width",
     xlab="Sepal Length",
     xlim=c(1,10),
     pch=12,
     col="grey",
     ylim=c(0,8))

points(iris$Sepal.Length[101:150],iris$Sepal.Width[101:150],
     main="Sepal Width vs Sepal Length",
     ylab="Sepal Width",
     xlab="Sepal Length",
     xlim=c(1,10),
     pch=12,
     col="blue",
     ylim=c(0,8))

legend("topright",legend=c("Setosa","Versicolor","Virginica"),pch=c(12,13,14),col=c("gol
d","grey","blue"))
```



c. What proportion of flowers have a Petal Length greater than 4, Petal widths between 1 and 2, and Sepal Widths and Lengths within 0.5 units of their median values?

```
#Subsetting each variable
PetalLG4<-subset(iris,Petal.Length>4)
PetalsWB1_2<-subset(iris,Petal.Width>1 & Petal.Width<2)

#Outputting objects
cat("Amount of flowers with petal length greater than 4:",nrow(PetalLG4)/nrow(iris))
```

```
Amount of flowers with petal length greater than 4: 0.56
```

```
cat("Amount of flowers with petal widths between 1 and 2:", nrow(PetalsWB1_2)/nrow(iri
s))
```

```
Amount of flowers with petal widths between 1 and 2: 0.4266667
```

```
#Lengths within 0.5 units of their median value
median(iris$Sepal.Width)
```

```
[1] 3
```

```
median(iris$Sepal.Length)
```

```
[1] 5.8
```

d. Observing the plots in (a) and (b), if you had to distinguish between classes by using either petal dimensions or sepal dimensions, which one would you choose: petals or sepals, and why?

```
#IF I had to distinguish between classes of either petal or sepal, I would chose petal b
ecause the plot shows datapoints more spread out than sepal which ultimately makes it ea
sier to distinguish between classes
```

Problem 4: (20 pts) The `babynames` dataset

Consider the `babynames` data from assignment 1 located within the R library package of the same name..

a. Create a subset of the data with female babies named “Mary” from 1880-2014. How many observations are in this subset?

```
library(babynames)
Name_Mary<-subset(babynames,year>=1880&year<=2014 & name=="Mary"&sex=="F")
cat("The amount of names corresponding to Mary is:",nrow(Name_Mary))
```

```
The amount of names corresponding to Mary is: 135
```

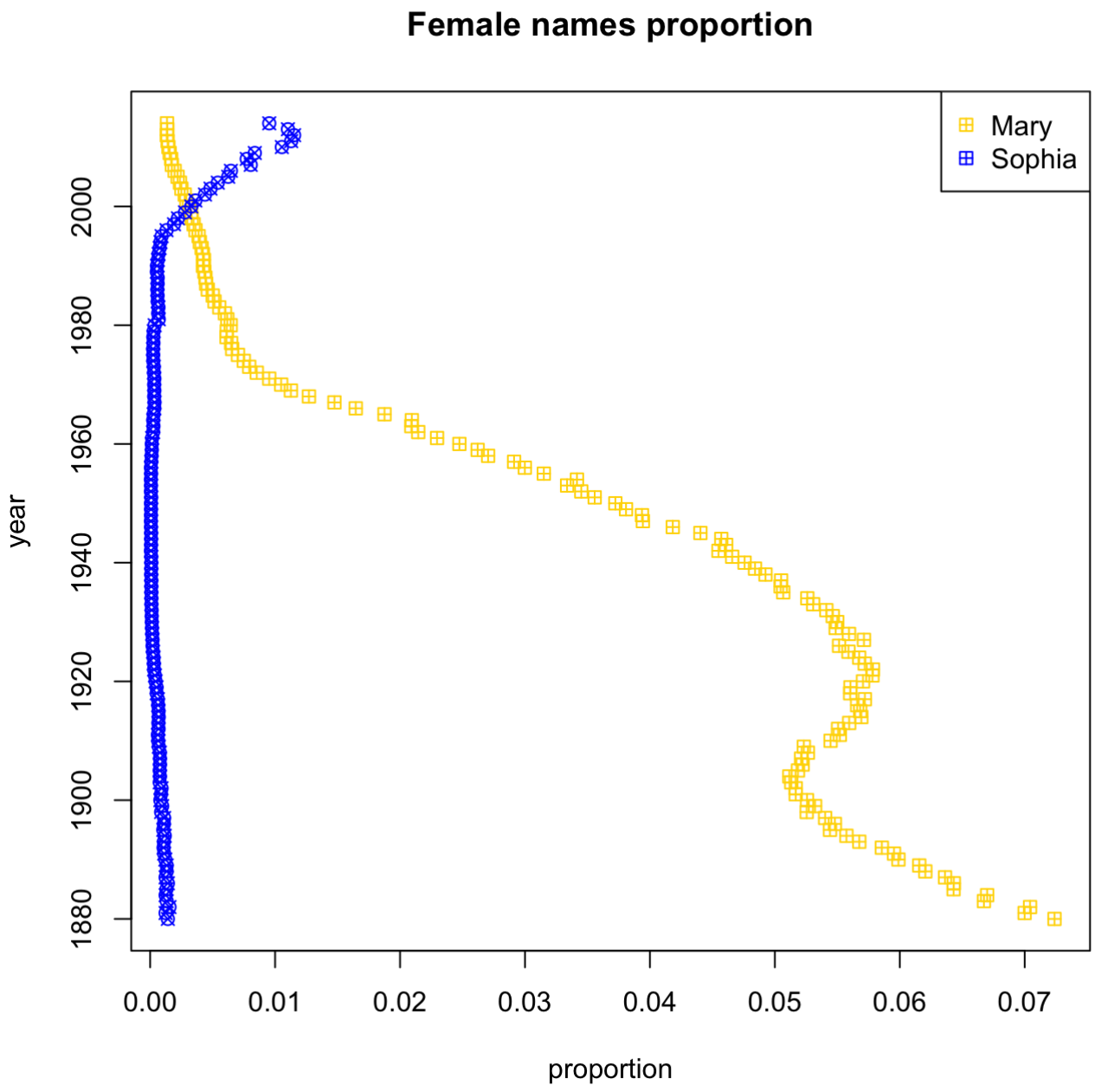
b. Create a subset of the data with female babies named “Sophia” from 1880-2014. How many observations are in this subset?

```
Name_Sophia<-subset(babynames,year>=1880&year<=2014 & name=="Sophia"&sex=="F")
cat("The amount of names corresponding to Sophia is:",nrow(Name_Sophia))
```

```
The amount of names corresponding to Sophia is: 135
```

c. Construct a plot of the proportion of female babies named “Mary” from 1880-2014. On the same plot, add/overlay a plot of the proportion of female babies named “Sophia” from 1880-2014. Use different colors for “Mary” vs “Sophia” and add a legend.

```
plot(Name_Mary$prop,Name_Mary$year, main = "Female names proportion",xlab="proportion",
 ylab ="year",pch=12,col="gold")
points(Name_Sophia$prop,Name_Sophia$year, main = "Female names proportion",xlab="proport
ion", ylab ="year",pch=13,col="blue")

legend("topright",legend=c("Mary","Sophia"),col = c("gold","blue"),pch=12)
```



d. Briefly describe your interpretation of the plot.

```
#My interpretation of the plot is that for the female name Mary, it seems that the name
 is being used less since 1880 and for the name Sophia, the plot shows an icrease in usa
ge since 1880 with a peak around 2012
```