Project 1

Siddhesh Mahadeshwar

Due: Thu 01/31 @ 11:59pm

Contents

```
1 library(modeest)
```

```
## Warning: package 'modeest' was built under R version 4.0.3
```

```
1 library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
1 library(stats)
2 library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Q1

Q1(a)

- **age**: The variable age is the age of a given individual in the data set; the unit of measurement is in years (Numeric - ratio).
- **workclass**: The variable workclass represents the employment status of an individual; the options are a set list of possible choices (Categorical - nominal).
- **fnlwgt**: This variable represents the noninstitutional population that the census can assume the data might represent; the final weight is calculated via a set of 3 controls (Numeric - continuous).
- **education**: The education variable marks the highest education level of the individual; there are a set number of possible options (Categorical - nominal).
- **education-num**: This variable represents the numerical representation of the highest level of education achieved by the individual; the data in this column is numeric - continuous.
- **marital-status**: This variable represents the marital status of the individual at the time of data collection; there are a set of possible options (Categorical - nominal)
- **occupation**: This variable represents the job/occupation type of the individual at the time of data collection; there are a set number of possible options (Categorical - nominal).
- **relationship**: This variable represents what the surveyed individual is as a relation to others; there are a set of possible options (Categorical - nominal).
- **race**: This variable represents the race of the individual; the possible options include some races (Categorical - nominal)
- **sex**: This variable represents the sex of the individual; there are two possible options (Categorical - nominal)
- **capital-gain**: This variable holds the capital gains (in dollars) of each surveyed individual (Numerical - continuous).
- **capital-loss**: This variable holds the capital losses (in dollars) of each surveyed individual (Numerical - continuous).
- **hours-per-week**: This variable holds the number of hours worked by the individual per week (units in hours); the data is numerical and continuous.
- **native-country**: This variable holds a collection of some of possible countries that the surveyed individuals come from (Categorical - nominal).
- **predicted-income**: This variable represents two categories that serve as predictions on whether that particular individual makes more than 50k dollars or less than or equal to 50k dollars (Dollars, categorical - nominal).

Q1(b)

(i)

"?" was the symbol used in the original adult.data.txt file to represent missing values in the data set.

(ii)

```
adult <- read.csv("adult.data.txt")

names(adult) = c("age", "workclass", "fnlgwt", "education", "education-num",
                 "marital-status", "occupation", "relationship", "race", "sex",
                 "capital-gain", "capital-loss", "hours-per-week", "native-country",
                 "predicted-income")
# The income variable represents the two unnamed categories from the original dataset:
# >50K & <=50K.


summary(adult == " ?") # this line can be used to display all variables at once
```

```
##     age          workclass         fnlgwt         education
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:32560     FALSE:30724     FALSE:32560     FALSE:32560
##                  TRUE :1836
##  education-num   marital-status  occupation      relationship
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:32560     FALSE:32560     FALSE:30717     FALSE:32560
##                                  TRUE :1843
##     race            sex          capital-gain    capital-loss
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:32560     FALSE:32560     FALSE:32560     FALSE:32560
##
##  hours-per-week  native-country  predicted-income
##  Mode :logical   Mode :logical   Mode :logical
##  FALSE:32560     FALSE:31977     FALSE:32560
##                  TRUE :583
```

```
# "TRUE" values represent the missing data counted

### The calculated percentages below represent the missing data from each variable
# age = 0%
# workclass = 1836/30724 * 100 = 5.98%
# fnlgwt = 0%
# education = 0%
# education-num = 0%
# marital-status = 0%
# occupation = 1843/30717 * 100 = 6%
# relationship = 0%
```

```
# race = 0%
# sex = 0%
# capital-gain = 0%
# capital-loss = 0%
# hours-per-week = 0%
# native-country = 583/31977 * 100 = 1.82%
# predicted-income = 0%


# create a new data frame with missing values removed from the data
n_adult <- adult[adult$workclass != " ?" | adult$occupation != " ?" |
                 adult$`native-country` != " ?",]
```

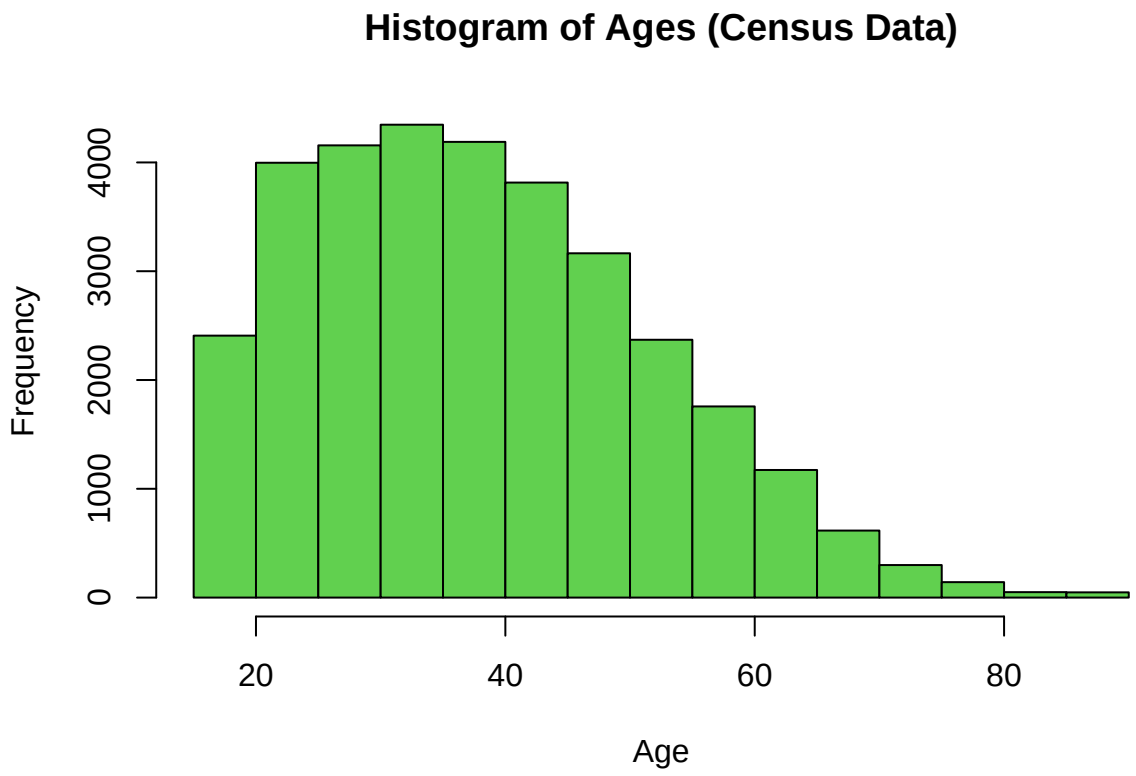
Q1(c)

- Categorical variables: workclass, education, marital-status, occupation, relationship, race, sex, native-country, predicted-income
- Numeric variables: age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week

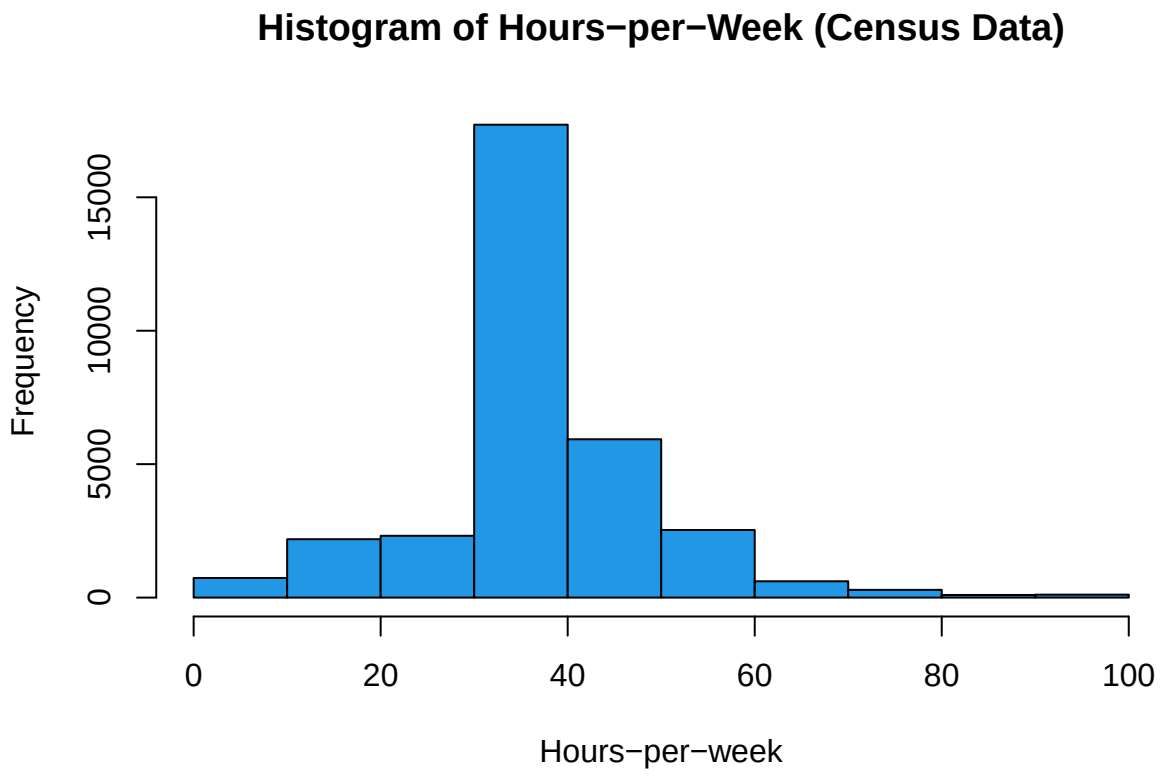
Q1(d)

(i)

```
hist(n_adult$age, breaks = 20,
     xlab = "Age",
     ylab = "Frequency",
     main = "Histogram of Ages (Census Data)",
     col = 3)
```

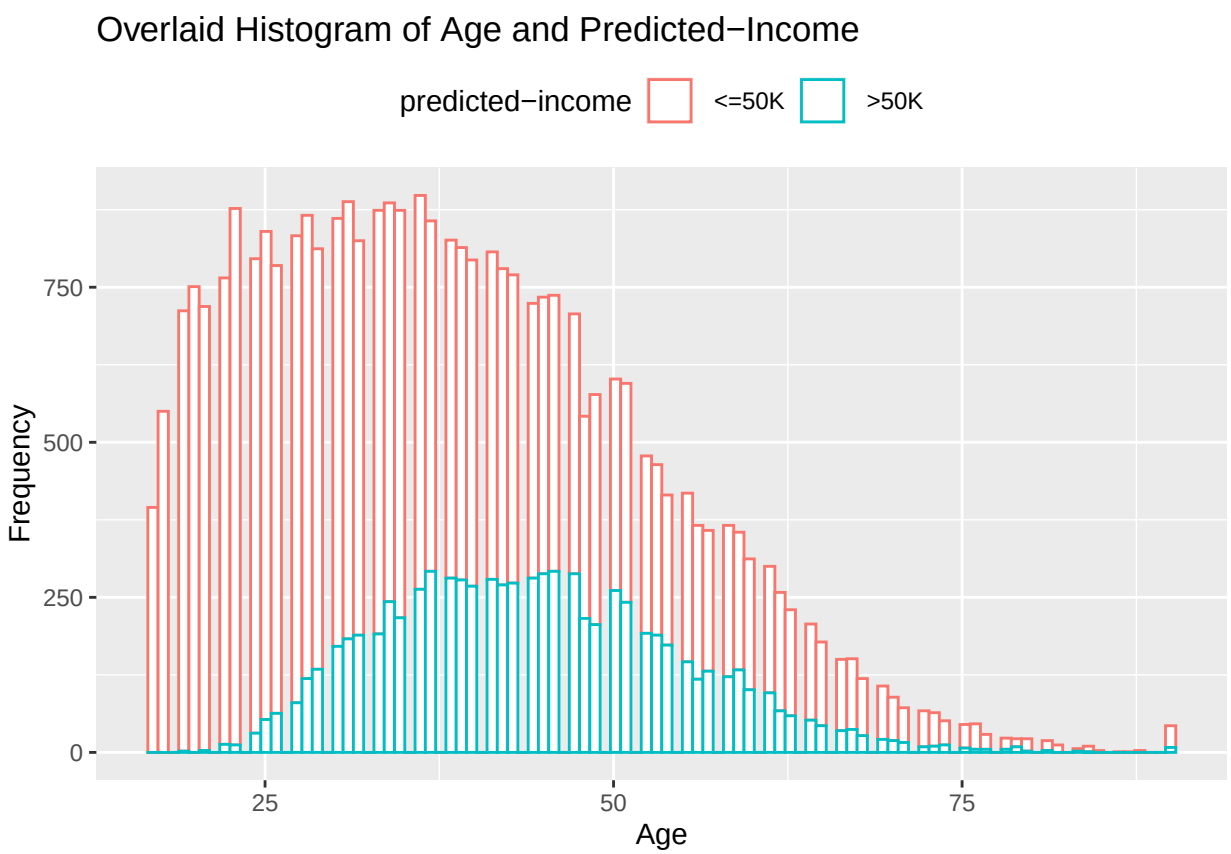



```r
hist(n_adult$`hours-per-week`, breaks = 10,
     xlab = "Hours-per-week",
     ylab = "Frequency",
     main = "Histogram of Hours-per-Week (Census Data)",
     col = 4)
```



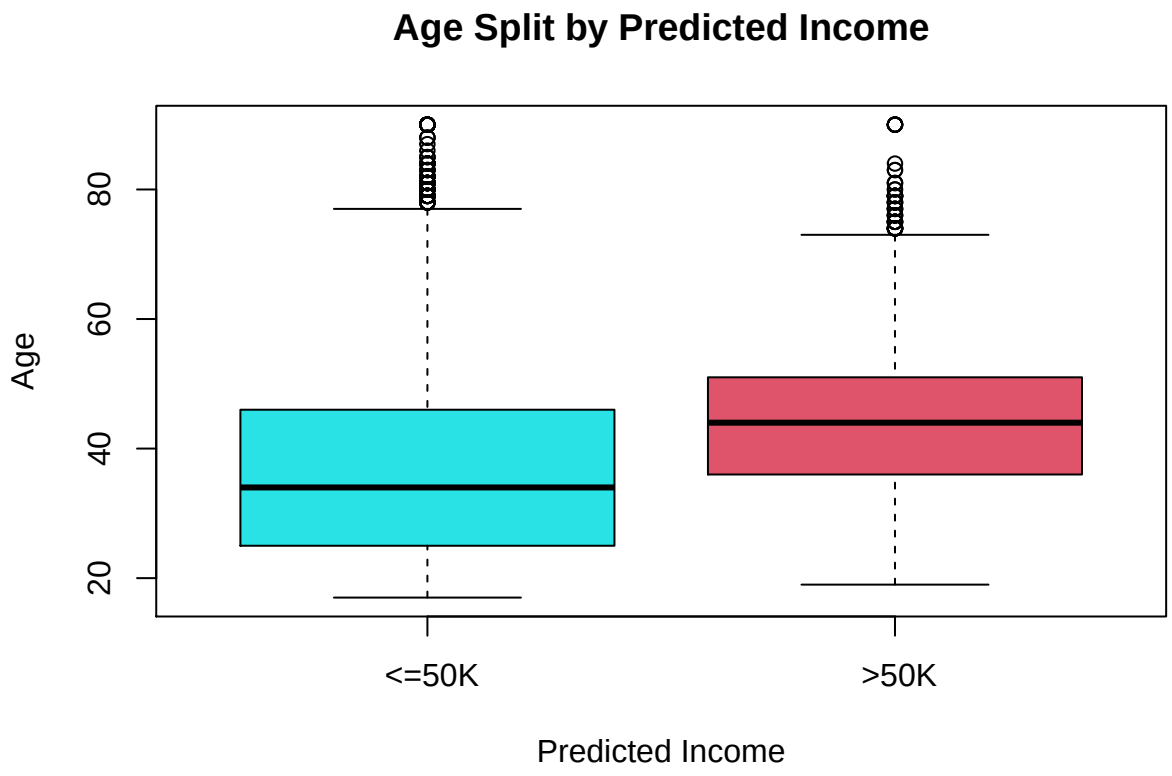
(ii)

```
ggplot(n_adult, aes(x=age, color = `predicted-income`)) +
       geom_histogram(fill="white", bins = 100) +
       xlab("Age") + ylab("Frequency") + theme(legend.position = "top") +
       labs(title = "Overlaid Histogram of Age and Predicted-Income",
            x = "Age", y = "Frequency")
```

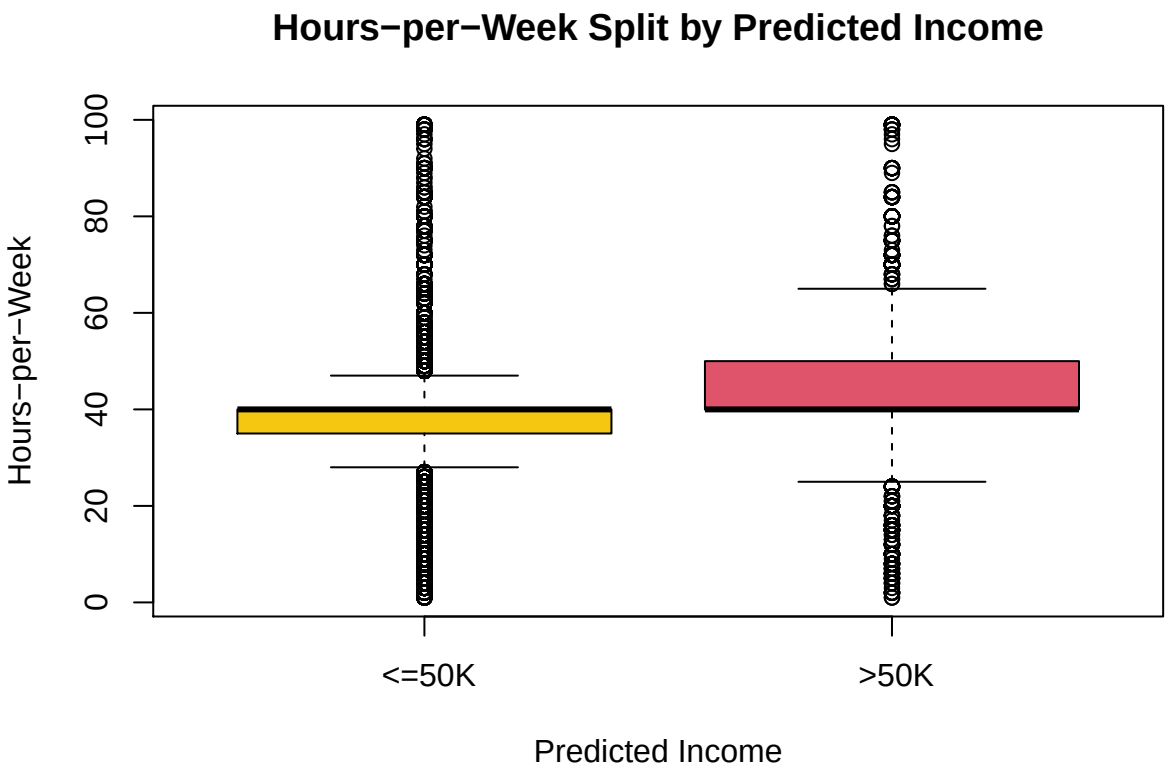


(iii)

```
age_income <- split(n_adult$age, n_adult$`predicted-income`)
hours_income <- split(n_adult$`hours-per-week`, n_adult$`predicted-income`)

boxplot(age_income,
        main = "Age Split by Predicted Income",
        xlab = "Predicted Income",
        ylab = "Age",
        col = c(5,10))
```




```r
boxplot(hours_income,
        main = "Hours-per-Week Split by Predicted Income",
        xlab = "Predicted Income",
        ylab = "Hours-per-Week",
        col = c(7,18))
```



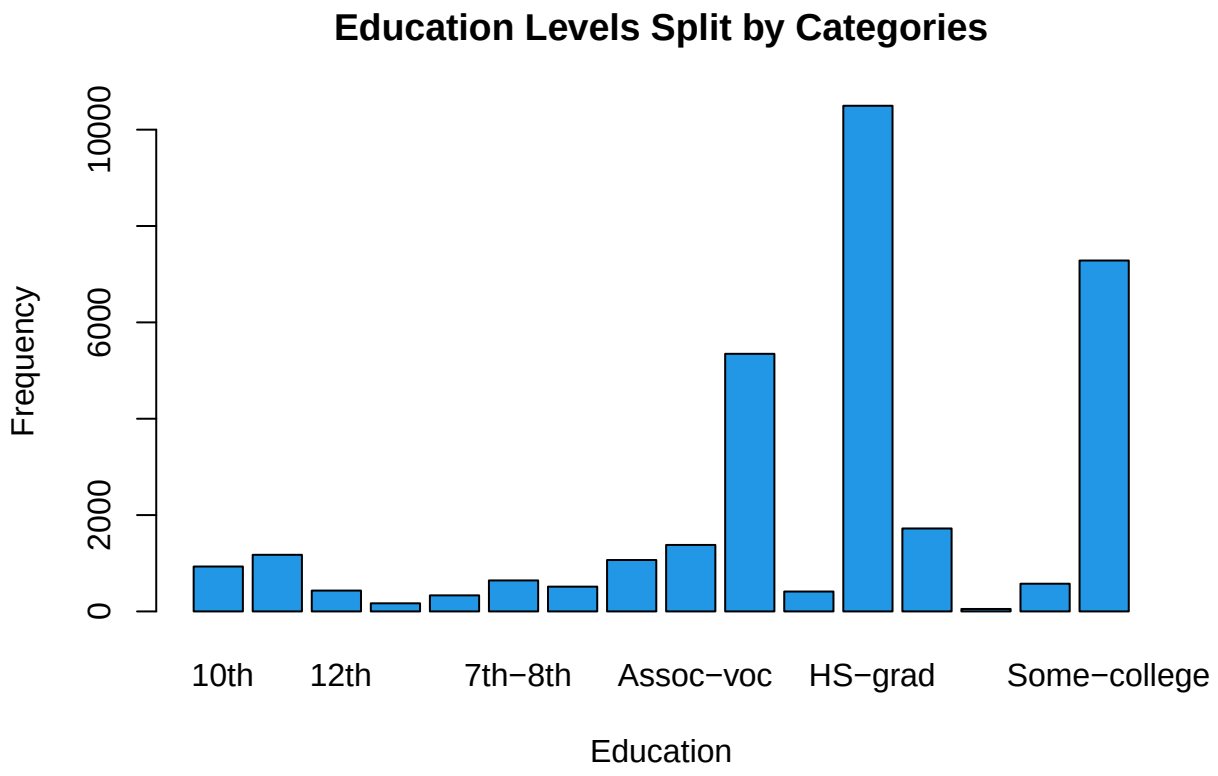
(iv)

Based on all the produced graphics above, it is quite apparent that there is a connection between the age of the individuals, the hours they work per week, and their predicted income. The general observed trend is that older people as well as people who tend to work more than the 35-40 hour average also tend to earn more, namely, more than 50k. It has also been revealed that the frequency of individuals earning more than 50k is significantly less than those who earn less than 50k (based on the overlaid histogram). Lastly, it can also be observed in the parallel boxplots that the individuals who had a predicted income higher than 50k typically worked between 40-50 hours while those who had a predicted income less than 50k worked between 35-40 hours.

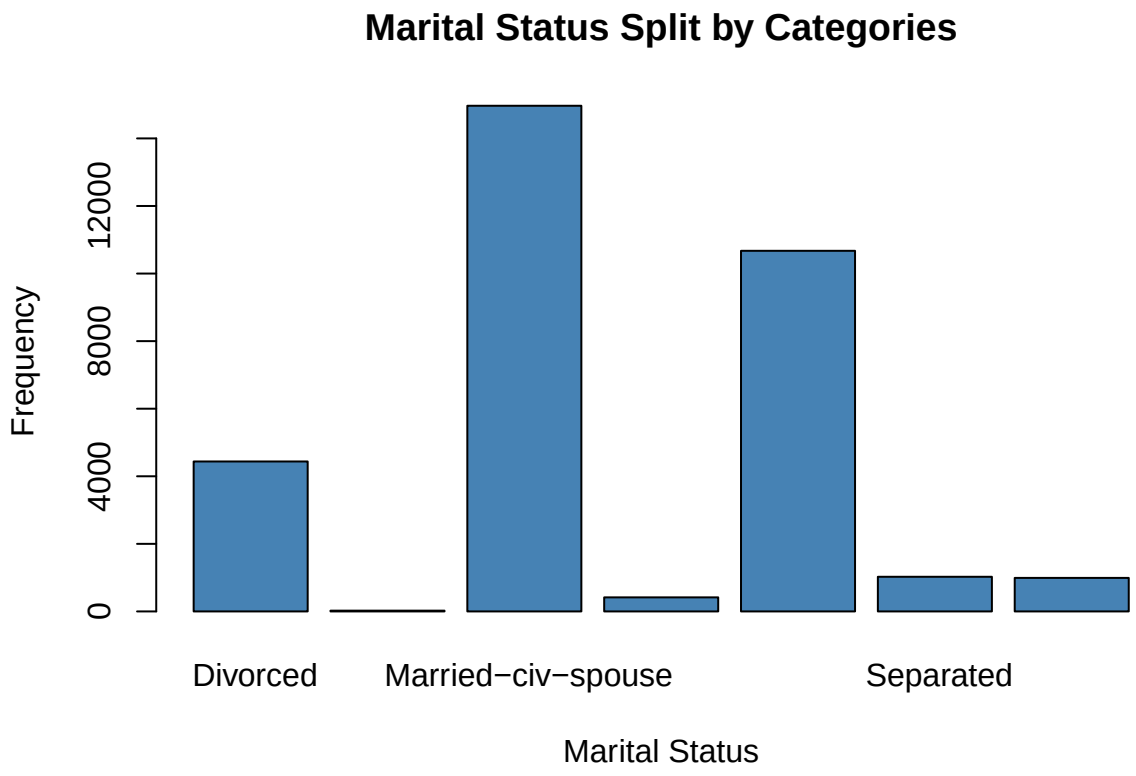
Q1(e)

(i)

```
barplot(table(n_adult$education), col = 4,
        main = "Education Levels Split by Categories",
        xlab = "Education", ylab = "Frequency")
```



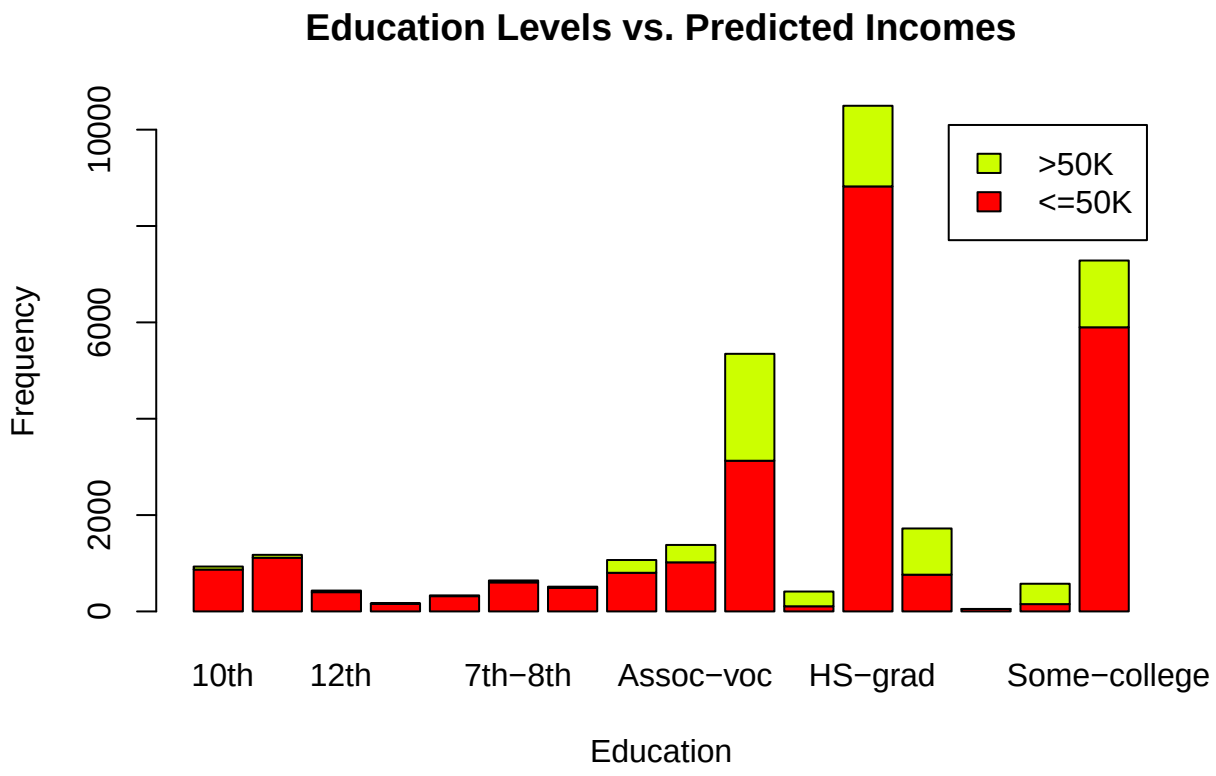

```r
barplot(table(n_adult$`marital-status`), col = "steelblue",
        main = "Marital Status Split by Categories",
        xlab = "Marital Status", ylab = "Frequency")
```



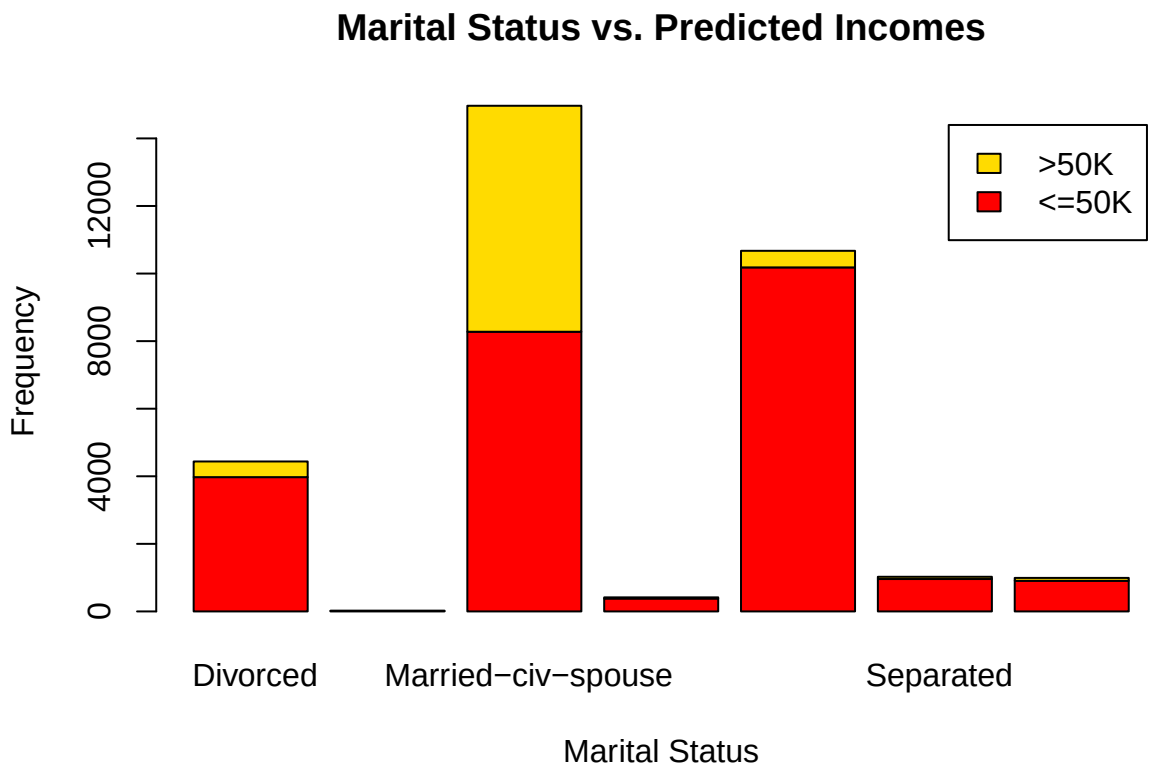

```
# las = 2 can be added to the above barplots to make the labels vertical and
# prevent cutting off. However, they then overlap the x-axis label. For presentation
# sake, this view was chosen.
```

(ii)

```
edu_inc <- table(n_adult$education, n_adult$`predicted-income`)

barplot(t(edu_inc), legend.text = TRUE, col = rainbow(5),
        main = "Education Levels vs. Predicted Incomes",
        xlab = "Education", ylab = "Frequency")
```




```r
marital_inc <- table(n_adult$`marital-status`, n_adult$`predicted-income`)

barplot(t(marital_inc), legend.text = TRUE, col = rainbow(7),
        main = "Marital Status vs. Predicted Incomes",
        xlab = "Marital Status", ylab = "Frequency")
```




```
# The order of the stacks could not be manipulated to put <=50k as the top stack.
# Instead, a legend is provided for clarity.
```

(iii)

Looking at the stacked barplots of education levels vs. predicted incomes, it is apparent that the higher levels of education yield more individuals who earn more than 50k. Likewise, lower education level individuals have lower than 50k incomes. When looking at the relationship between marital status and predicted income, it seems that almost all except Divorced, Married-civ-spouse, and Never-married are individuals earning below 50k. Married-civ-spouse is the marital status with the greatest number of individuals with a predicted income higher than 50k.

Q2

Q2(a)

```
sports <- read.csv("nfl-20-running-stats.csv")

# New data frame including only those players with the requirements met
my.sport <- sports[(sports$Pos == "RB" | sports$Pos == "FB" | sports$Pos == "rb"
                    | sports$Pos == "fb") & (sports$G >= 6),]
```

```
length(my.sport$Player)
```

```
## [1] 80
```

After cleaning the data to include only those players needed for further analysis, we have a pool of 80 players. Note that this includes part-time starters and primary starters.

Q2(b)

```
# all calculations for the variable TD
mean(my.sport$TD)     # mean
```

```
## [1] 4.1125
```

```
median(my.sport$TD)   # median
```

```
## [1] 3
```

```
mfv(my.sport$TD)      # mode
```

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

```
# all calculations for the variable Fmb
mean(my.sport$Fmb)    # mean
```

```
## [1] 1.1875
```

```
median(my.sport$Fmb)  # median
```

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

```
mfv(my.sport$Fmb)     # mode
```

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

Q2(c)

```
quantile(my.sport$X1D, c(.25, .37, .75))
```

```
##   25%   37%   75%
## 11.75 19.23 40.00
```

```
quantile(my.sport$Yds, c(.25, .37, .75))
```

```
##    25%    37%    75%
## 221.50 367.23 698.50
```

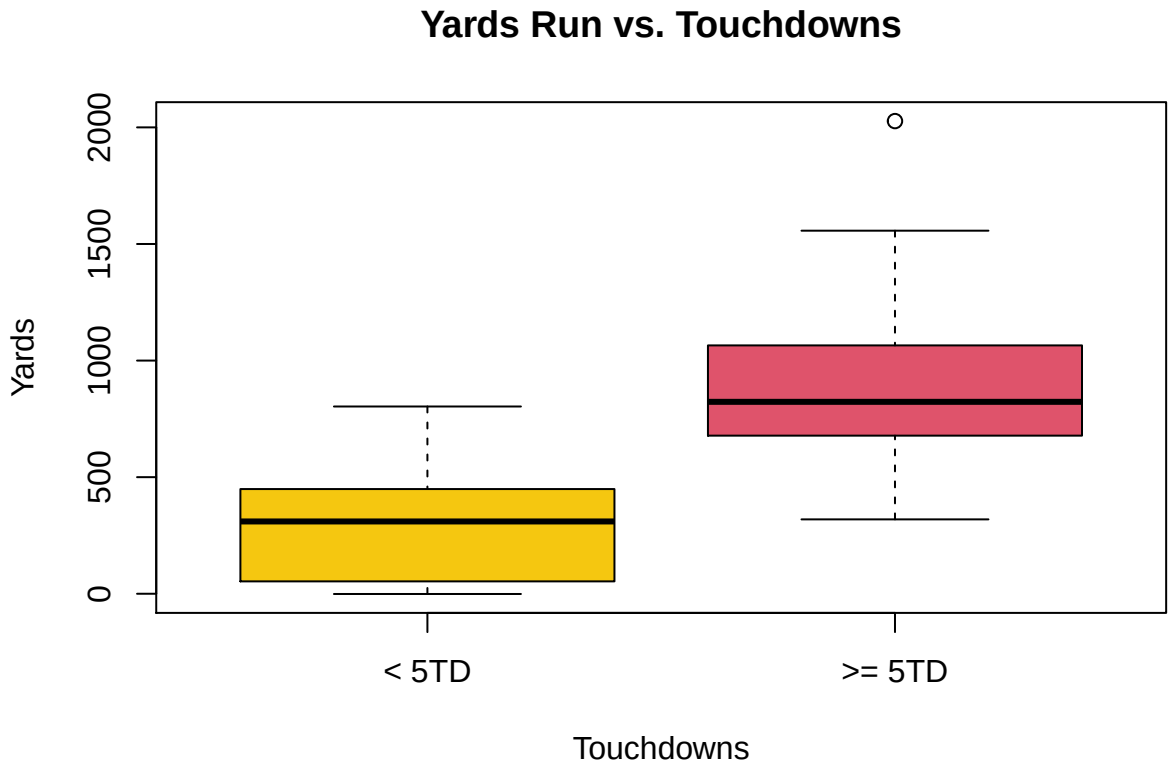
Q2(d)

```
summary(my.sport[, c("Y.G", "Lng")])
```

```
##       Y.G              Lng
##  Min.   : -0.10   Min.   :-1.00
##  1st Qu.: 18.30   1st Qu.:16.50
##  Median : 38.95   Median :34.00
##  Mean   : 38.46   Mean   :36.52
##  3rd Qu.: 57.00   3rd Qu.:51.75
##  Max.   :126.70   Max.   :98.00
```

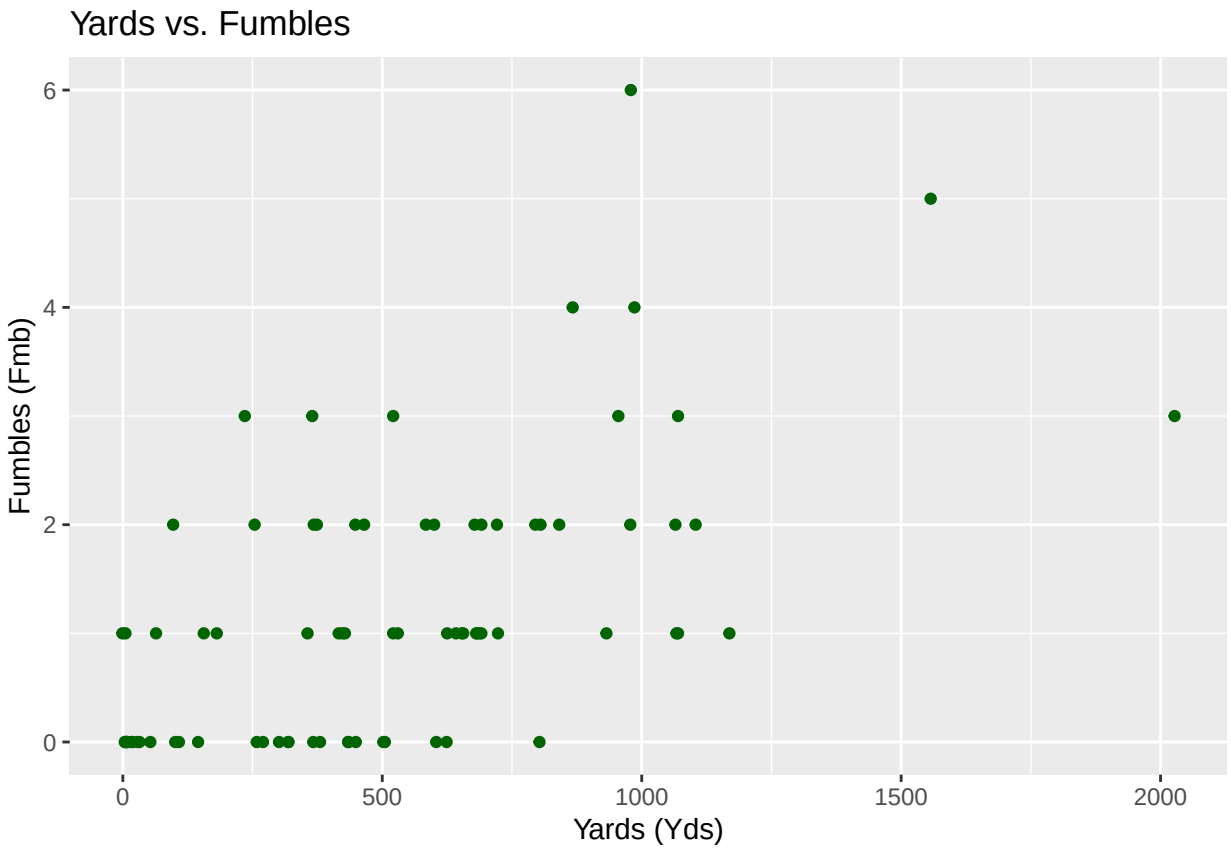
Q2(e)

```
# Created 2 new columns to match the TD requirements
my.sport$new <- ifelse(my.sport$TD >= 5, ">= 5TD", "< 5TD")

new_split <- split(my.sport$Yds, my.sport$new)

boxplot(new_split,
        main = "Yards Run vs. Touchdowns",
        xlab = "Touchdowns",
        ylab = "Yards",
        col = c(7,18))
```



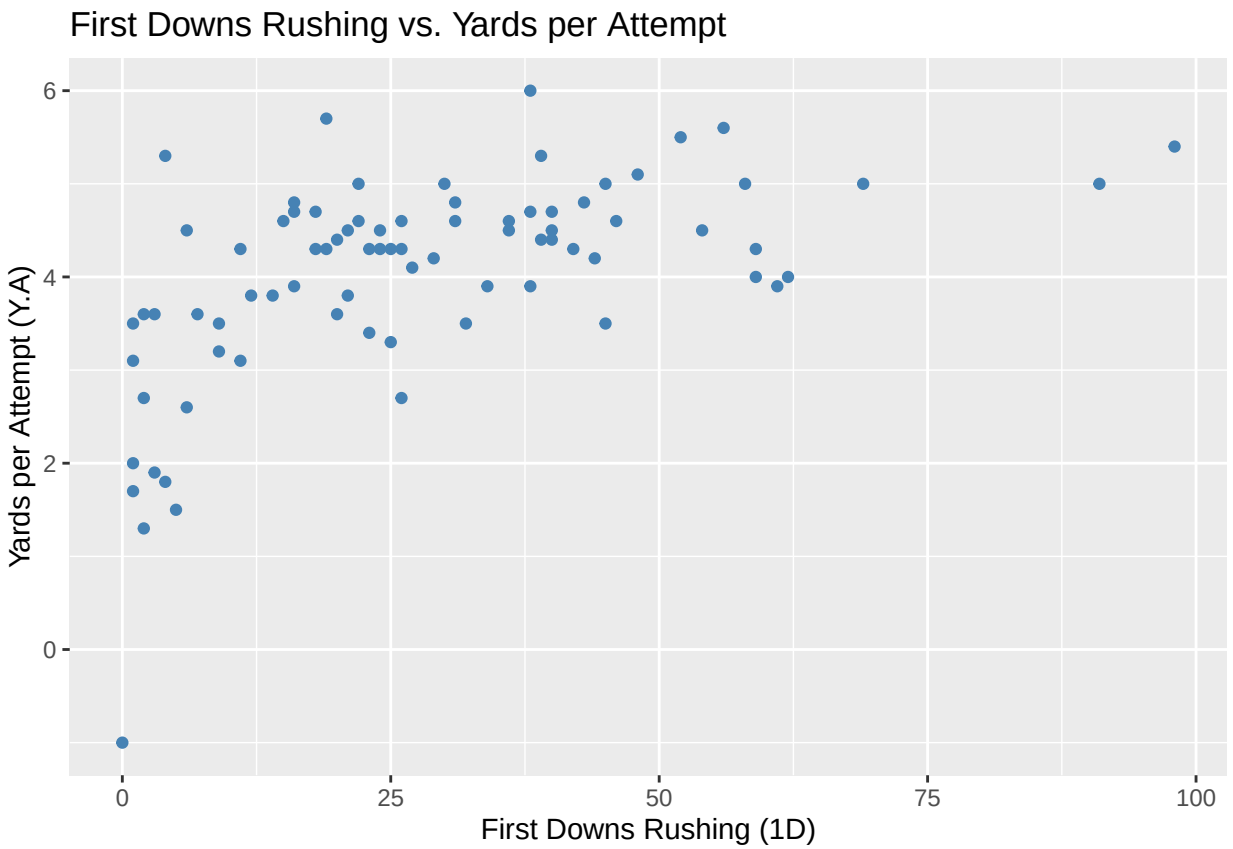
Q2(f)

```
ggplot(my.sport, aes(x=Yds, y=Fmb)) +
       geom_point(color = "darkgreen") +
       ylab("Fumbles (Fmb)") +
       xlab("Yards (Yds)") +
       ggtitle("Yards vs. Fumbles")
```



Q2(g)

```
ggplot(my.sport, aes(x=X1D, y=Y.A)) +
       geom_point(color = "steelblue") +
       ylab("Yards per Attempt (Y.A)") +
       xlab("First Downs Rushing (1D)") +
       ggtitle("First Downs Rushing vs. Yards per Attempt")
```


First Downs Rushing vs. Yards per Attempt
Yards per Attempt (Y.A)
6
4
2
0
0
25
50
75
100
First Downs Rushing (1D)