CS108 STATISTIS

Summer 2019

Assignment 1

Gurgen Hayrapetyan

Load “Trucking_jobs.csv” datain your R environment.

```
jobs <- read.csv("jobs.csv")
```

How many observations are in the data set?

We can find the number of observation by looking at length of one of the variables

```
observations <- length(jobs$sex)
print(observations)
```

```
## [1] 129
```

What are the variables/features and their data types? Indicate if they are categorical or quantitative variables.

To answer this question we can use `str` function

```
str(jobs)
```

```
## 'data.frame':    129 obs. of  5 variables:
##  $ sex       : Factor w/ 2 levels "F","M": 2 1 1 2 2 2 2 1 2 2 ...
##  $ earnings  : int  35000 36800 25000 45000 30000 60000 40000 30000 25000 30000 ...
##  $ age       : int  25 62 34 44 34 46 30 26 43 37 ...
##  $ title     : Factor w/ 32 levels "AGENT REP","AGENT/TRAINING REP",..: 24 4 27 4 27 4 24 12 27 27 .
##  $ hiredyears: int  0 5 1 0 3 0 5 3 1 3 ...
```

From this we can see that `sex` and `title` variables are categorical while `earnings`, `age`, and `hiredyears` are quantative

What is the mean, standard deviation, median, and range of earnings.

we can access earning from our dataset by `jobs$earnings`.

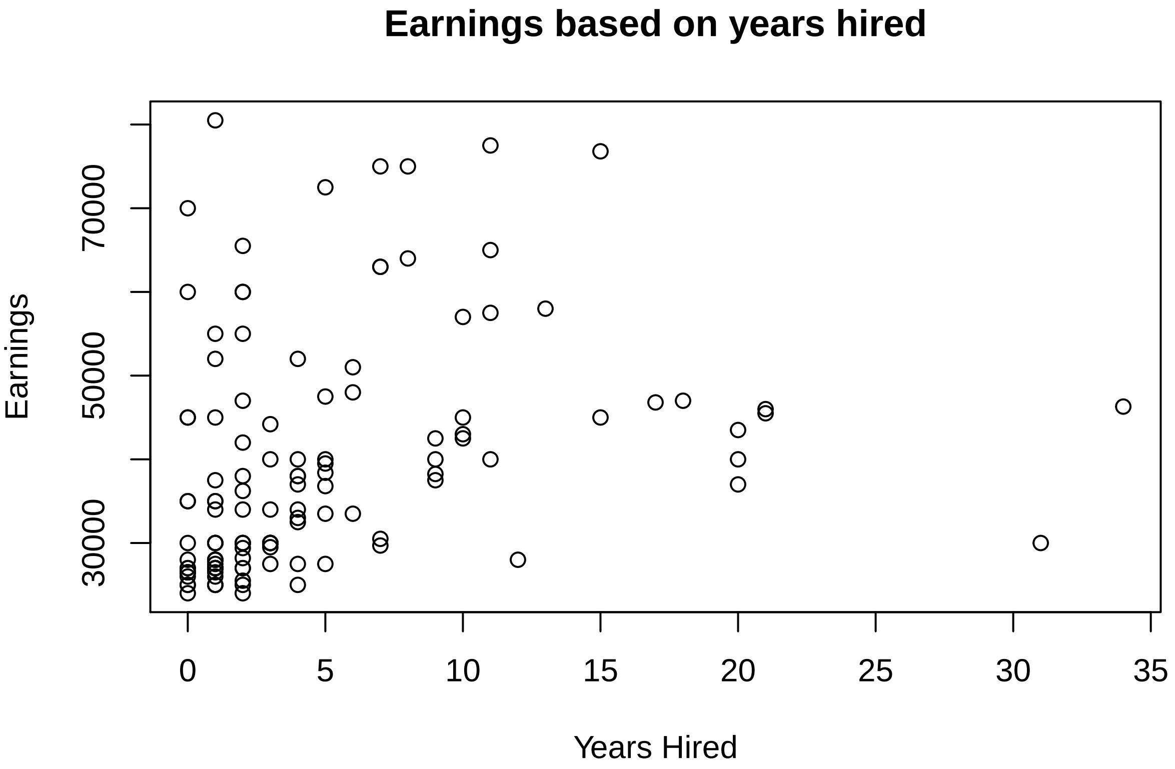
```
earnings <- jobs$earnings
earnings_mean <- mean(earnings)
earnings_sd <- sqrt(var(earnings))
earnings_median <- median(earnings)
earnings_range <- range(earnings)[2] - range(earnings)[1]

# code for visualizing the values

data <- c(earnings_mean, earnings_sd, earnings_median, earnings_range)
names(data) <- c(" Mean", "Standard Deviation", "Median", "Range")

labelAll <- paste(names(data), ": ", data, "\n", sep="")
cat(labelAll)
```

```
##  Mean: 38768.1007751938
##  Standard Deviation: 13805.056240743
##  Median: 34000
##  Range: 56500
```

Plot a scatterogram of earnings based on hire year

Our Y-axis will be the hireyears X-axis the earnings

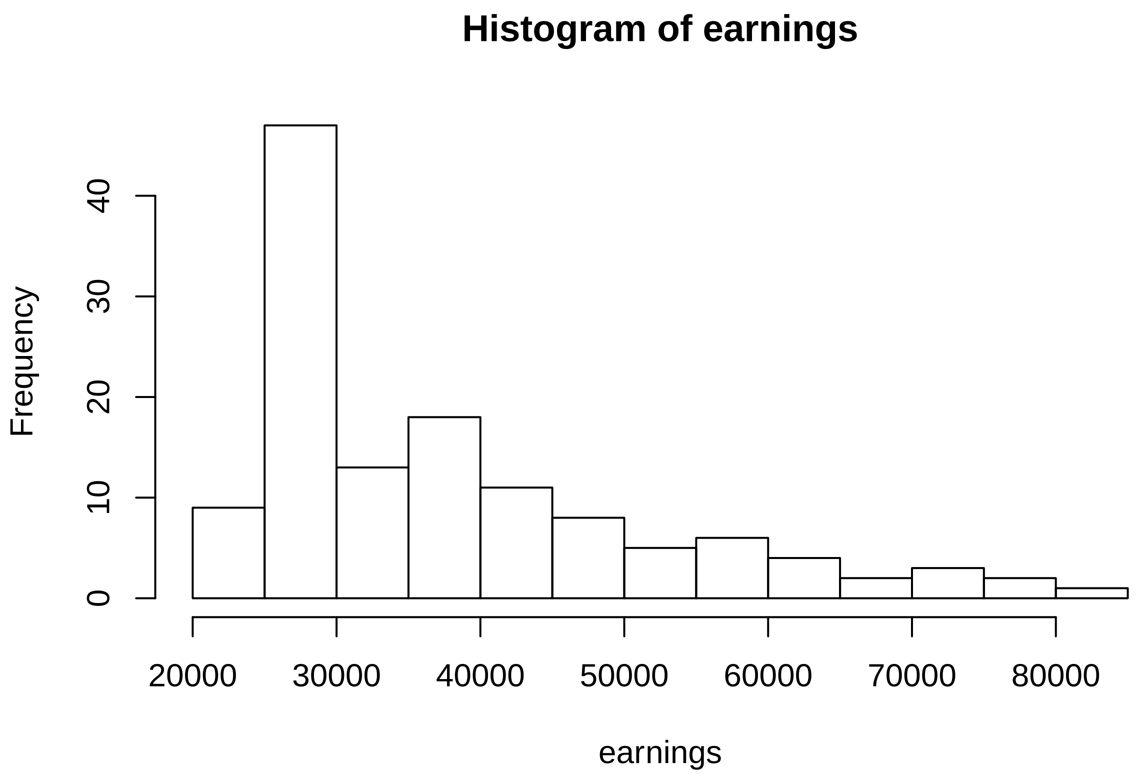
```
x <- jobs$hiredyears
y <- jobs$earnings
plot(x, y, main = "Earnings based on years hired", xlab = "Years Hired", ylab = "Earnings")
```



Plot the histogram of earnings by count and density

We can plot the histgram by just using the `hist` function

```
hist(earnings)
```



Find quartiles, 80th quantile(percentile), and IQR of the earnings using quantiles function. What is the median

```
quartiles <- quantile(earnings)
quantile_80 <- quantile(earnings, probs = c(0.8))
earnings_iqr <- IQR(earnings)
```

```
print(quartiles)
```

```
##    0%   25%   50%   75%  100% 
## 24000 28000 34000 45000 80500
```

```
print(quantile_80)
```

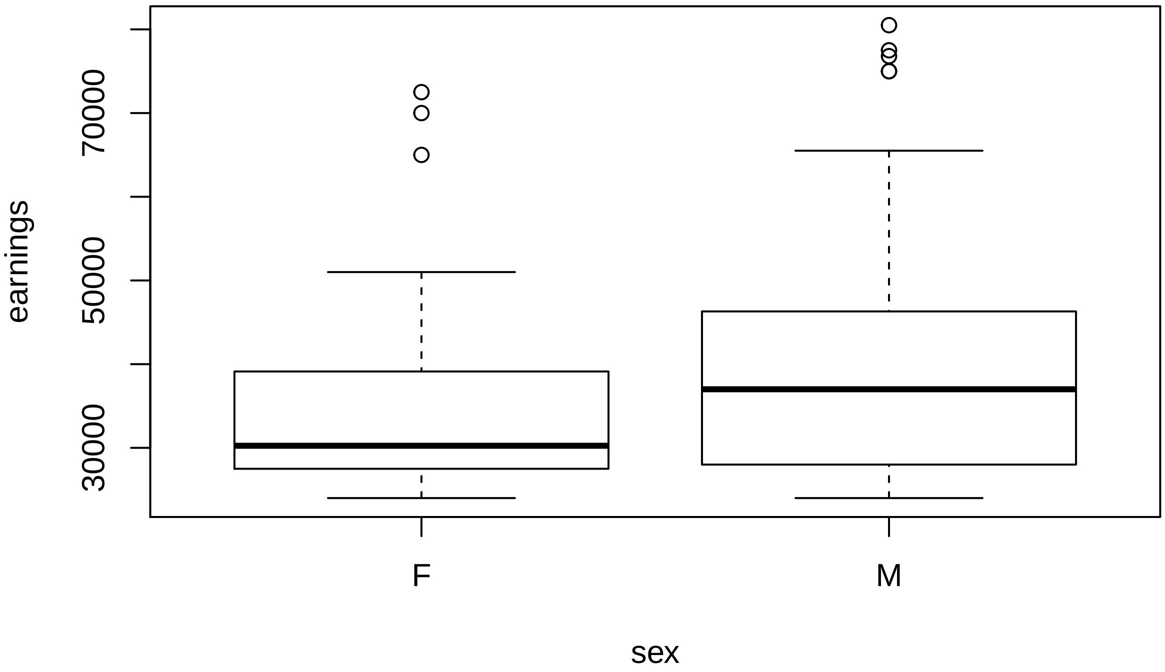
```
##   80% 
## 47000
```

```
print(earnings_iqr)
```

```
## [1] 17000
```

Plot the boxplot of earnings and separate by sex. Find values of outliers.

```
earnings <- jobs$earnings
sex <- jobs$sex

values <- boxplot(earnings ~ sex)
```



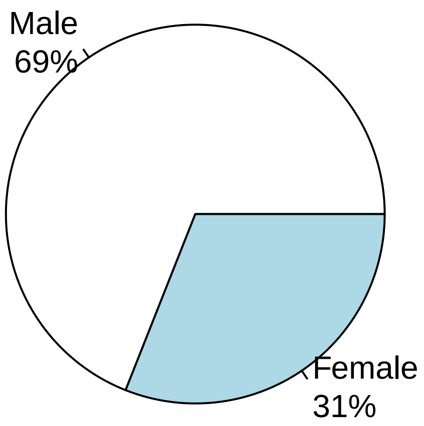
```
outliers <- values$out
print(outliers)
```

```
## [1] 72500 70000 65000 76800 75000 80500 75000 77500
```

Plot pie chart for percent male and female in the data set.

```
# seperating our data by gender
all_males <- subset(jobs, sex == "M")
all_females <- subset(jobs, sex == "F")

# getting number of observations in each sample
all_observations <- length(jobs$sex)
males <- round(length(all_males$sex) * 100 / all_observations)
females <- round(length(all_females$sex) * 100 / all_observations)
data <- c(males, females)

# generating labels
names(data) <- c("Male", "Female")
labelAll <- paste(names(data), "\n", data, "%", sep="")

# plotting the chart
pie(data, main="Male/Female Ratio of Jobs", labels = labelAll)
```

Male/Female Ratio of Jobs


Male
Female