exercises chapter 1 LMR

waheeb Algabri

1.The dataset teengamb concerns a study of teenage gambling in Britain. Make a numerical and graphical summary of the data, commenting on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data.

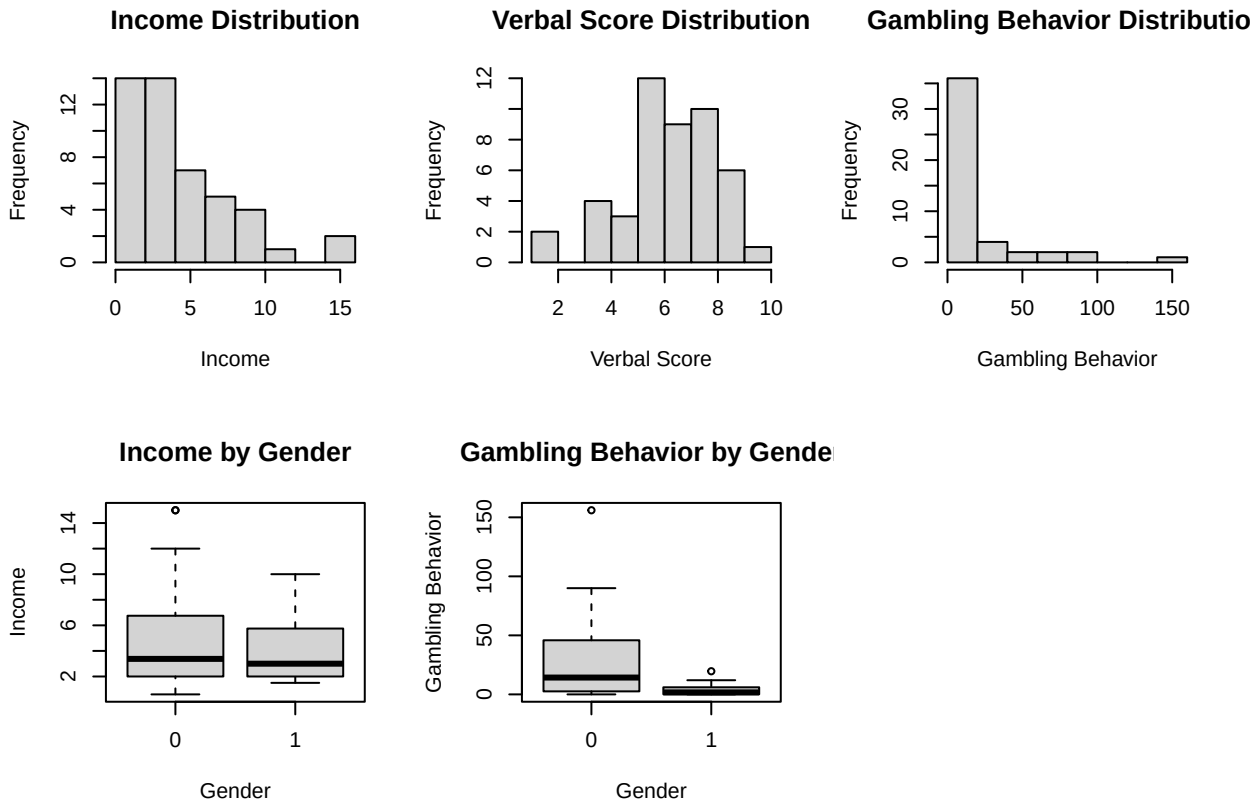
```
library(faraway)
data(teengamb)
head(teengamb)
```

```
##   sex status income verbal gamble
## 1   1     51   2.00      8    0.0
## 2   1     28   2.50      8    0.0
## 3   1     37   2.00      6    0.0
## 4   1     28   7.00      4    7.3
## 5   1     65   2.00      8   19.6
## 6   1     61   3.47      6    0.1
```

```
summary(teengamb)
```

```
##       sex            status          income           verbal
##  Min.   :0.0000   Min.   :18.00   Min.   : 0.600   Min.   : 1.00
##  1st Qu.:0.0000   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00
##  Median :0.0000   Median :43.00   Median : 3.250   Median : 7.00
##  Mean   :0.4043   Mean   :45.23   Mean   : 4.642   Mean   : 6.66
##  3rd Qu.:1.0000   3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00
##  Max.   :1.0000   Max.   :75.00   Max.   :15.000   Max.   :10.00
##      gamble
##  Min.   :  0.0
##  1st Qu.:  1.1
##  Median :  6.0
##  Mean   : 19.3
##  3rd Qu.: 19.4
##  Max.   :156.0
```

```
# Graphical summary
par(mfrow=c(2, 3))
hist(teengamb$income, main="Income Distribution", xlab="Income")
hist(teengamb$verbal, main="Verbal Score Distribution", xlab="Verbal Score")
hist(teengamb$gamble, main="Gambling Behavior Distribution", xlab="Gambling Behavior")
boxplot(income ~ sex, data=teengamb, main="Income by Gender", xlab="Gender", ylab="Income")
boxplot(gamble ~ sex, data=teengamb, main="Gambling Behavior by Gender", xlab="Gender", ylab="Gambling
```



2.The dataset uswages is drawn as a sample from the Current Population Survey in 1988. Make a numerical and graphical summary of the data as in the previous question

```
data("uswages")
head(uswages)
```

```
##         wage educ exper race smsa ne mw so we pt
## 6085  771.60   18    18    0    1  1  0  0  0  0
## 23701 617.28   15    20    0    1  0  0  0  1  0
## 16208 957.83   16     9    0    1  0  0  1  0  0
## 2720  617.28   12    24    0    1  1  0  0  0  0
## 9723  902.18   14    12    0    1  0  1  0  0  0
## 22239 299.15   12    33    0    1  0  0  0  1  0
```

```
summary(uswages)
```

```
##       wage              educ           exper            race
##  Min.   :  50.39   Min.   : 0.00   Min.   :-2.00   Min.   :0.000
##  1st Qu.: 308.64   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:0.000
##  Median : 522.32   Median :12.00   Median :15.00   Median :0.000
##  Mean   : 608.12   Mean   :13.11   Mean   :18.41   Mean   :0.078
##  3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00   3rd Qu.:0.000
##  Max.   :7716.05   Max.   :18.00   Max.   :59.00   Max.   :1.000
##       smsa             ne              mw              so
##  Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
```

```
##  1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :0.000   Median :0.0000   Median :0.0000
##  Mean   :0.756   Mean   :0.229   Mean   :0.2485   Mean   :0.3125
##  3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##        we              pt
##  Min.   :0.00   Min.   :0.0000
##  1st Qu.:0.00   1st Qu.:0.0000
##  Median :0.00   Median :0.0000
##  Mean   :0.21   Mean   :0.0925
##  3rd Qu.:0.00   3rd Qu.:0.0000
##  Max.   :1.00   Max.   :1.0000
```

2-The dataset uswages is drawn as a sample from the Current Population Survey in 1988. Make a numerical and graphical summary of the data as in the previous question.

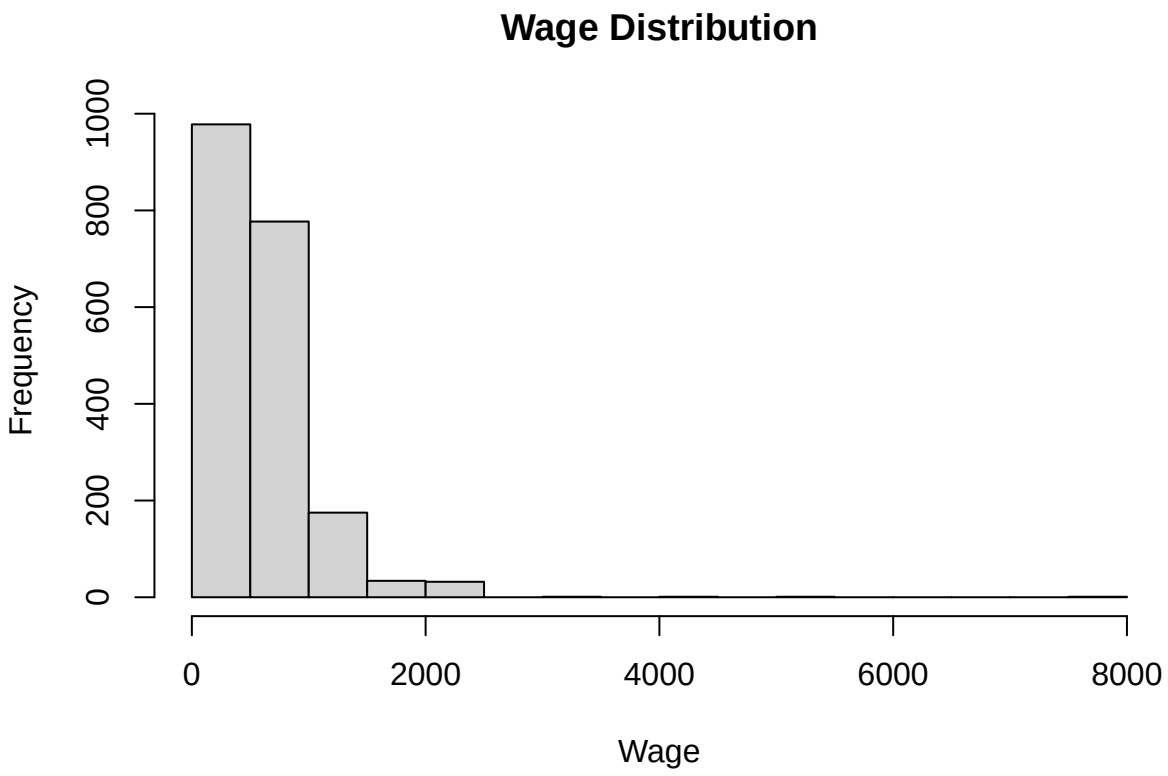
```
head(uswages)
```

```
##         wage educ exper race smsa ne mw so we pt
## 6085  771.60   18    18    0    1  1  0  0  0  0
## 23701 617.28   15    20    0    1  0  0  0  1  0
## 16208 957.83   16     9    0    1  0  0  1  0  0
## 2720  617.28   12    24    0    1  1  0  0  0  0
## 9723  902.18   14    12    0    1  0  1  0  0  0
## 22239 299.15   12    33    0    1  0  0  0  1  0
```

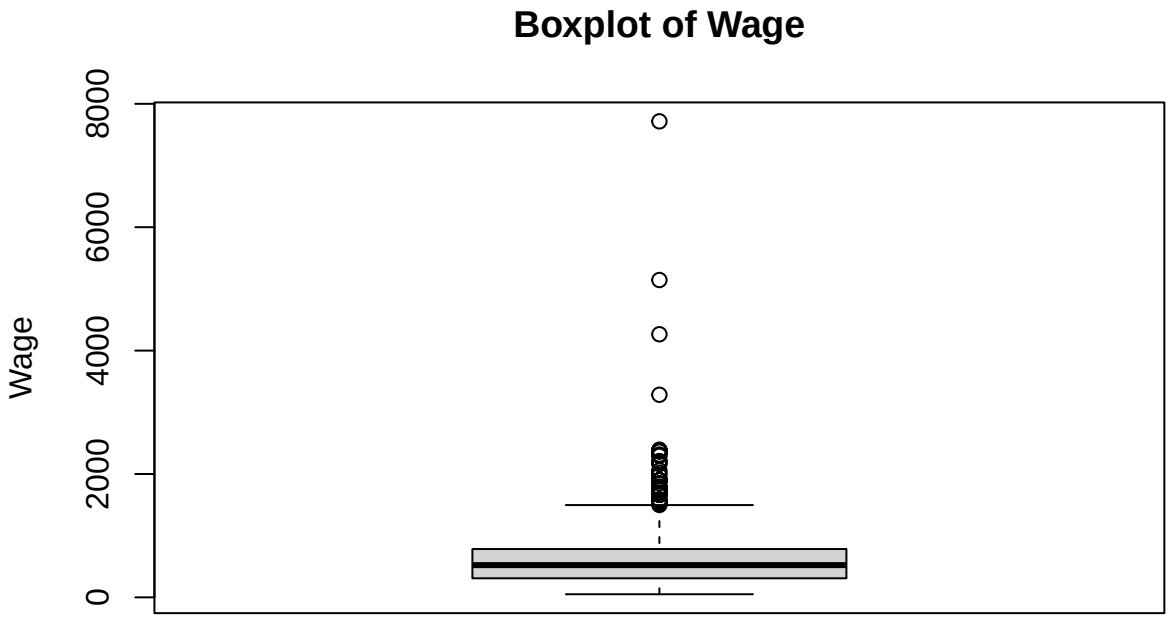
```
summary(uswages)
```

```
##       wage              educ           exper            race
##  Min.   :  50.39   Min.   : 0.00   Min.   :-2.00   Min.   :0.000
##  1st Qu.: 308.64   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:0.000
##  Median : 522.32   Median :12.00   Median :15.00   Median :0.000
##  Mean   : 608.12   Mean   :13.11   Mean   :18.41   Mean   :0.078
##  3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00   3rd Qu.:0.000
##  Max.   :7716.05   Max.   :18.00   Max.   :59.00   Max.   :1.000
##       smsa             ne              mw              so
##  Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :0.000   Median :0.0000   Median :0.0000
##  Mean   :0.756   Mean   :0.229   Mean   :0.2485   Mean   :0.3125
##  3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##        we              pt
##  Min.   :0.00   Min.   :0.0000
##  1st Qu.:0.00   1st Qu.:0.0000
##  Median :0.00   Median :0.0000
##  Mean   :0.21   Mean   :0.0925
##  3rd Qu.:0.00   3rd Qu.:0.0000
##  Max.   :1.00   Max.   :1.0000
```

```
hist(uswages$wage, main="Wage Distribution", xlab="Wage")
```

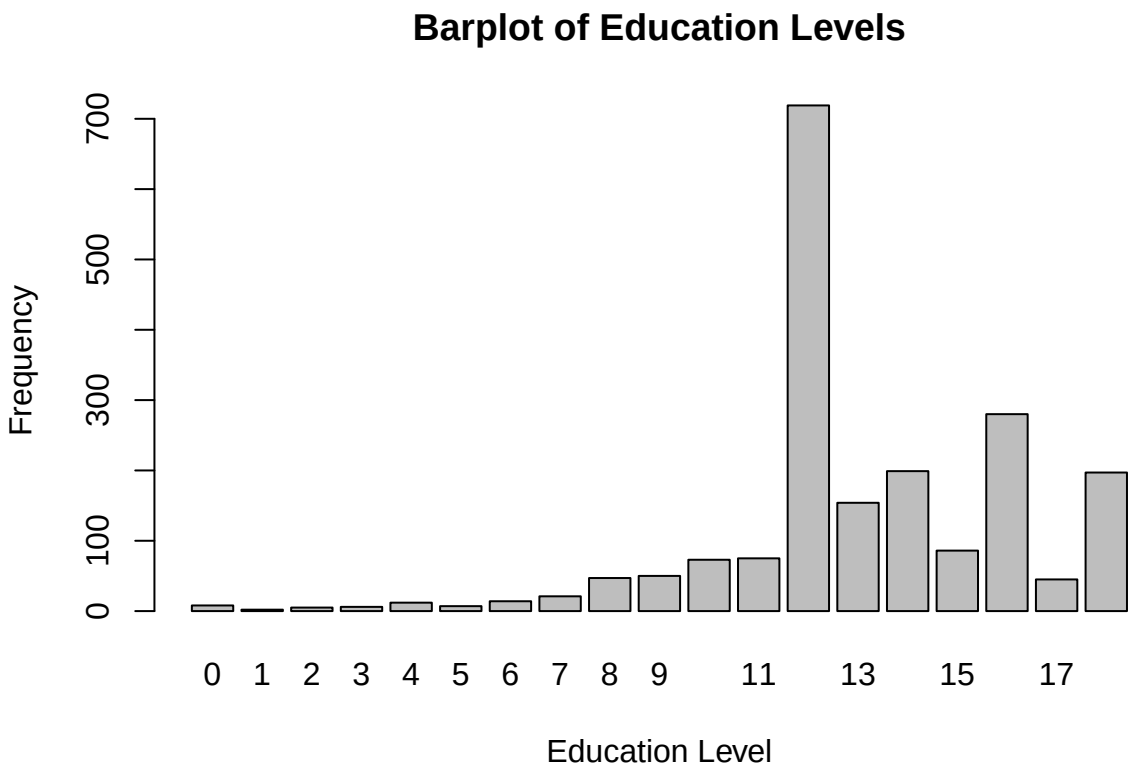



```
boxplot(uswages$wage, main="Boxplot of Wage", ylab="Wage")
```




```
plot(uswages$exper, uswages$wage, main="Scatterplot of Wage vs Experience", xlab="Experience", ylab="Wag
```


```
barplot(table(uswages$educ), main="Barplot of Education Levels", xlab="Education Level", ylab="Frequency
```




```
ggplot(uswages, aes(x = exper, y = wage)) +
  geom_point() +
  labs(title = "Scatterplot of Wage vs Experience", x = "Experience", y = "Wage")
```

3-The dataset 'prostate' is derived from a study involving 97 men diagnosed with prostate cancer, all of whom were scheduled to undergo a radical prostatectomy. Provide a numerical and graphical summary of the data, similar to the previous question.

```
head(prostate)
```

```
##       lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
## 1 -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
## 2 -0.9942523  3.3196  58 -1.386294   0 -1.38629       6     0 -0.16252
## 3 -0.5108256  2.6912  74 -1.386294   0 -1.38629       7    20 -0.16252
## 4 -1.2039728  3.2828  58 -1.386294   0 -1.38629       6     0 -0.16252
## 5  0.7514161  3.4324  62 -1.386294   0 -1.38629       6     0  0.37156
## 6 -1.0498221  3.2288  50 -1.386294   0 -1.38629       6     0  0.76547
```

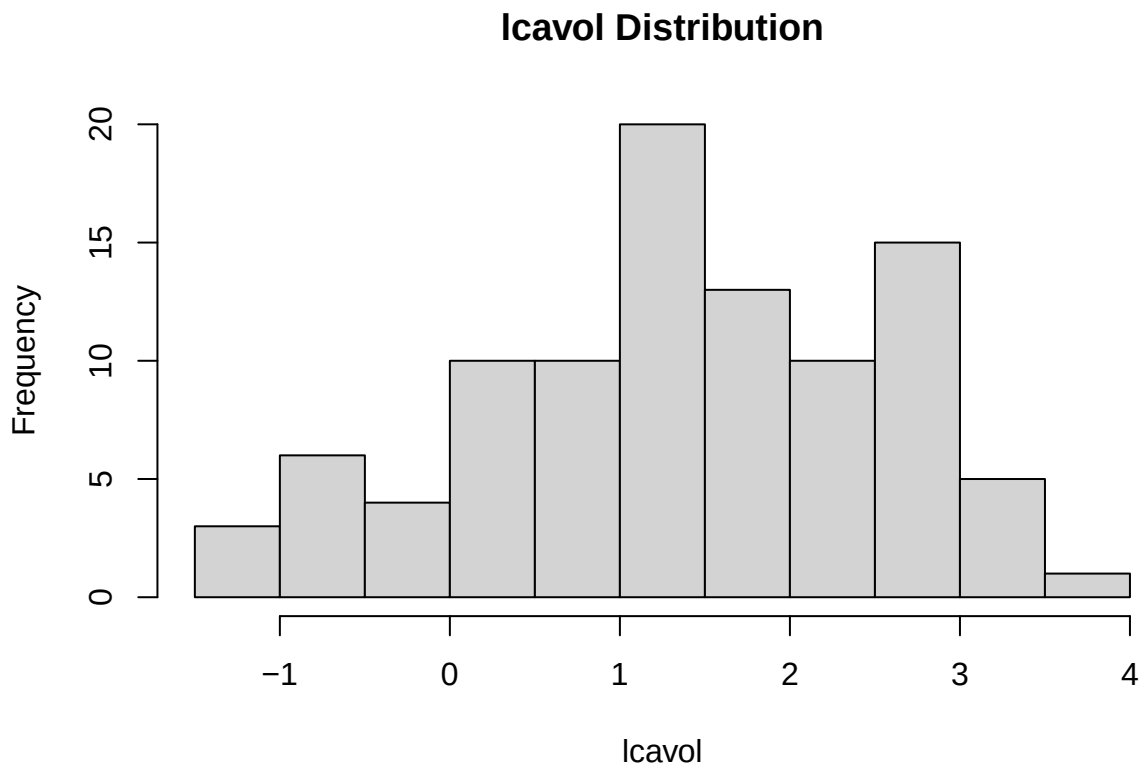
```
summary(prostate)
```

```
##      lcavol           lweight           age             lbph
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001
##  Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004
##  3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581
##  Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263
##       svi              lcp             gleason          pgg45
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00
```

```
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38
##  3rd Qu.:0.0000   3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00
##       lpsa
##  Min.   :-0.4308
##  1st Qu.: 1.7317
##  Median : 2.5915
##  Mean   : 2.4784
##  3rd Qu.: 3.0564
##  Max.   : 5.5829
```

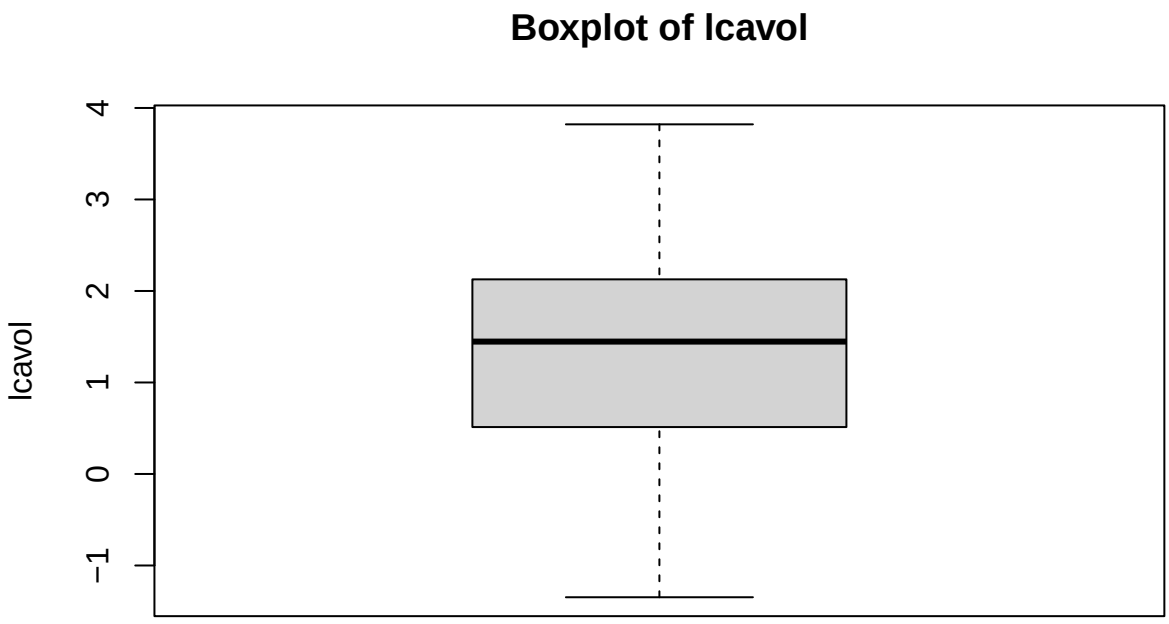
Histogram for lcavol:

```
hist(prostate$lcavol, main="lcavol Distribution", xlab="lcavol")
```



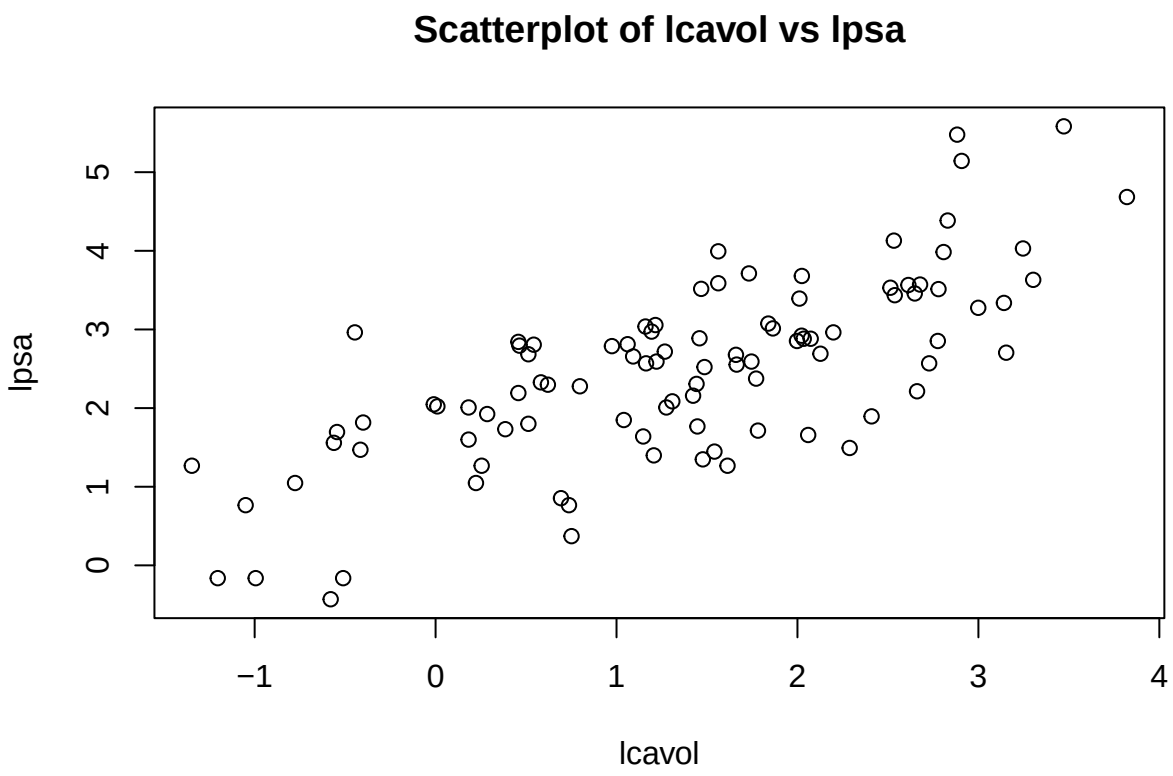
Boxplot for lcavol:

```
boxplot(prostate$lcavol, main="Boxplot of lcavol", ylab="lcavol")
```



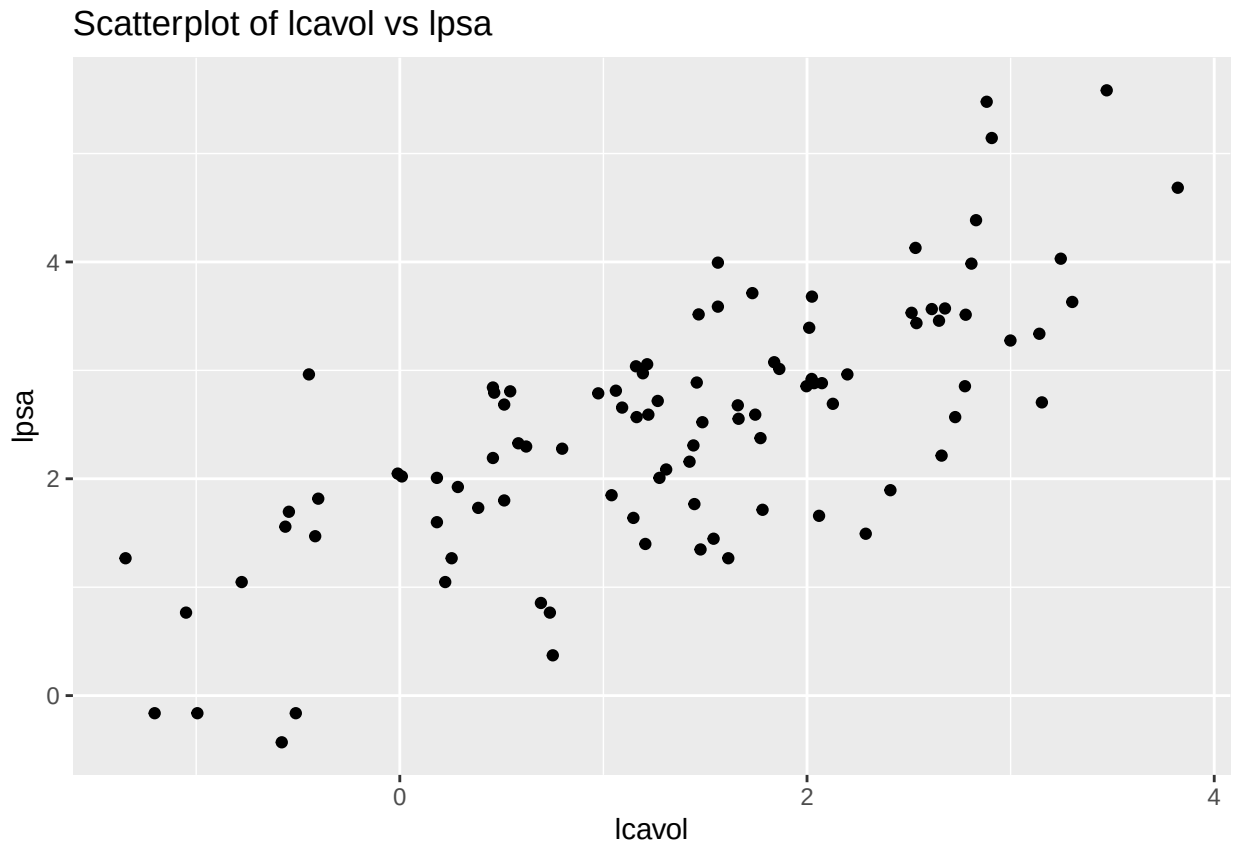
Scatterplot for lcavol vs. lpsa:

```
plot(prostate$lcavol, prostate$lpsa, main="Scatterplot of lcavol vs lpsa", xlab="lcavol", ylab="lpsa")
```




```
barplot(table(prostate$gleason), main="Barplot of Gleason Score", xlab="Gleason Score", ylab="Frequency
```


```
ggplot(prostate, aes(x = lcavol, y = lpsa)) +
  geom_point() +
  labs(title = "Scatterplot of lcavol vs lpsa", x = "lcavol", y = "lpsa")
```



4-The dataset sat comes from a study entitled "Getting What You Pay For: The Debate Over Equity in Public School Expenditures." Make a numerical and graphical summary of the data as in the first question.

```
head(sat)
```

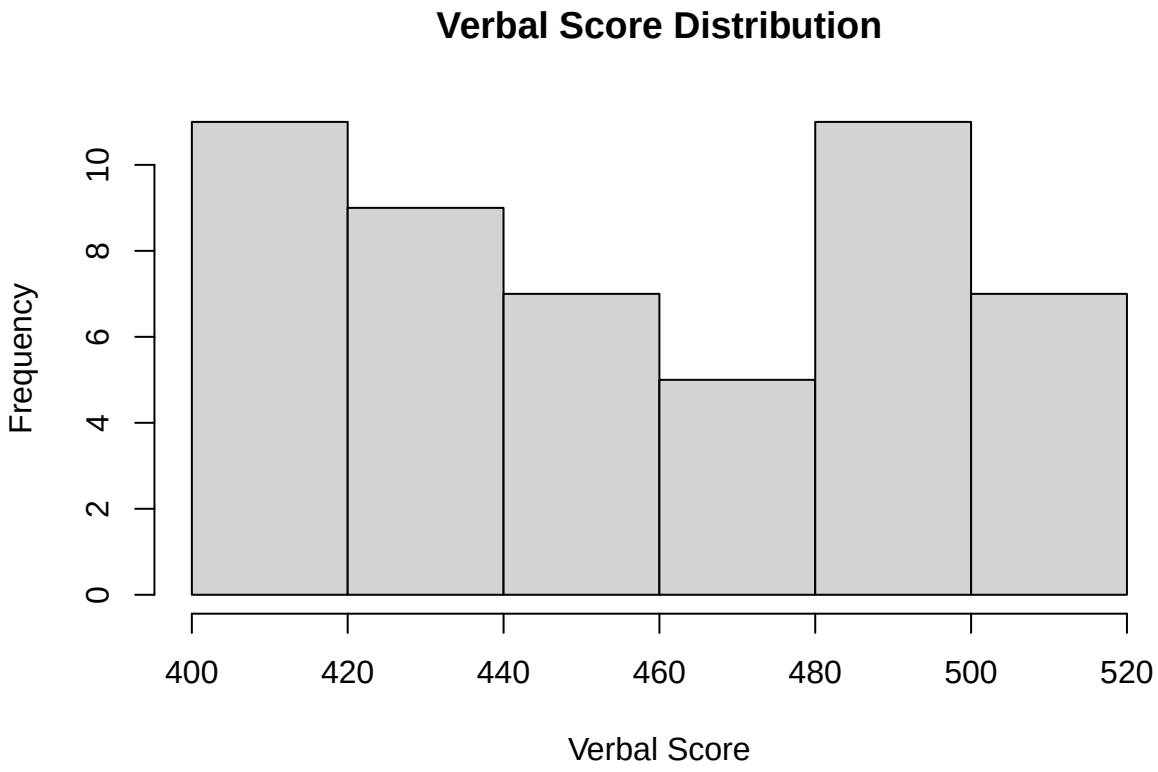
```
##            expend ratio salary takers verbal math total
## Alabama     4.405  17.2 31.144      8    491  538  1029
## Alaska      8.963  17.6 47.951     47    445  489   934
## Arizona     4.778  19.3 32.175     27    448  496   944
## Arkansas    4.459  17.1 28.934      6    482  523  1005
## California  4.992  24.0 41.078     45    417  485   902
## Colorado    5.443  18.4 34.571     29    462  518   980
```

```
summary(sat)
```

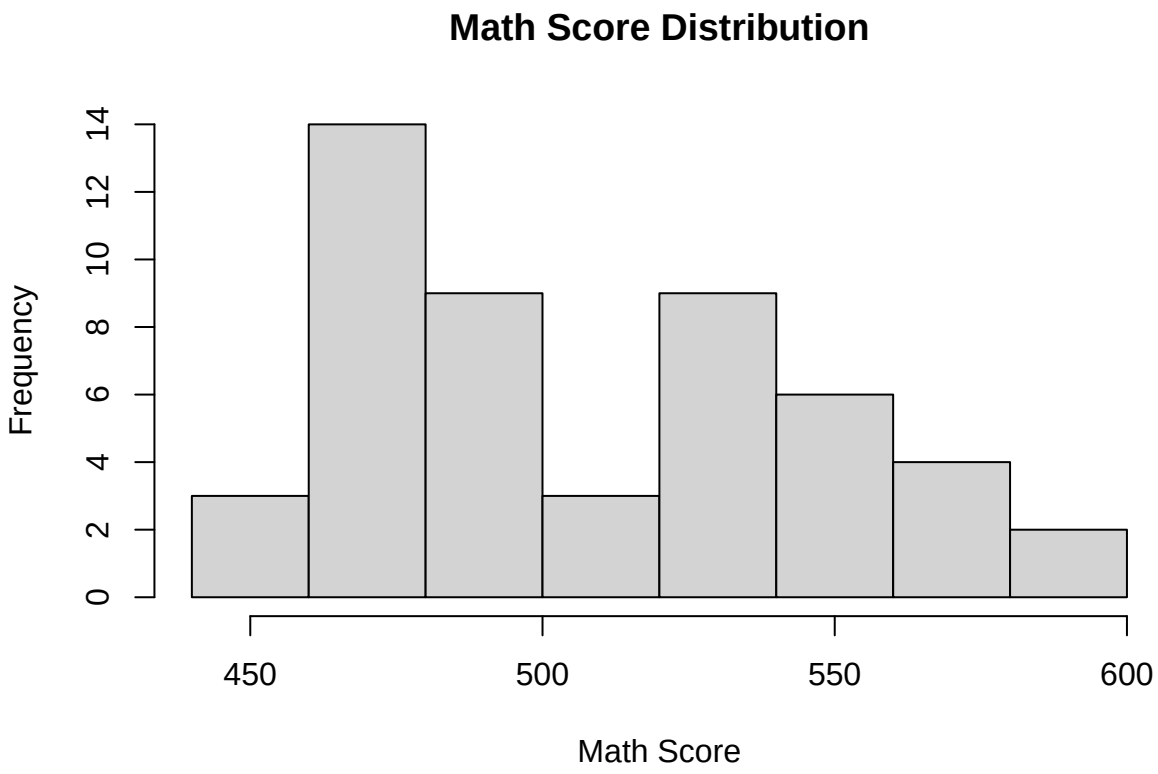
```
##      expend          ratio           salary          takers
##  Min.   :3.656   Min.   :13.80   Min.   :25.99   Min.   : 4.00
##  1st Qu.:4.882   1st Qu.:15.22   1st Qu.:30.98   1st Qu.: 9.00
##  Median :5.768   Median :16.60   Median :33.29   Median :28.00
##  Mean   :5.905   Mean   :16.86   Mean   :34.83   Mean   :35.24
##  3rd Qu.:6.434   3rd Qu.:17.57   3rd Qu.:38.55   3rd Qu.:63.00
##  Max.   :9.774   Max.   :24.30   Max.   :50.05   Max.   :81.00
##      verbal           math           total
##  Min.   :401.0   Min.   :443.0   Min.   : 844.0
```

```
##  1st Qu.:427.2   1st Qu.:474.8   1st Qu.: 897.2
##  Median :448.0   Median :497.5   Median : 945.5
##  Mean   :457.1   Mean   :508.8   Mean   : 965.9
##  3rd Qu.:490.2   3rd Qu.:539.5   3rd Qu.:1032.0
##  Max.   :516.0   Max.   :592.0   Max.   :1107.0
```

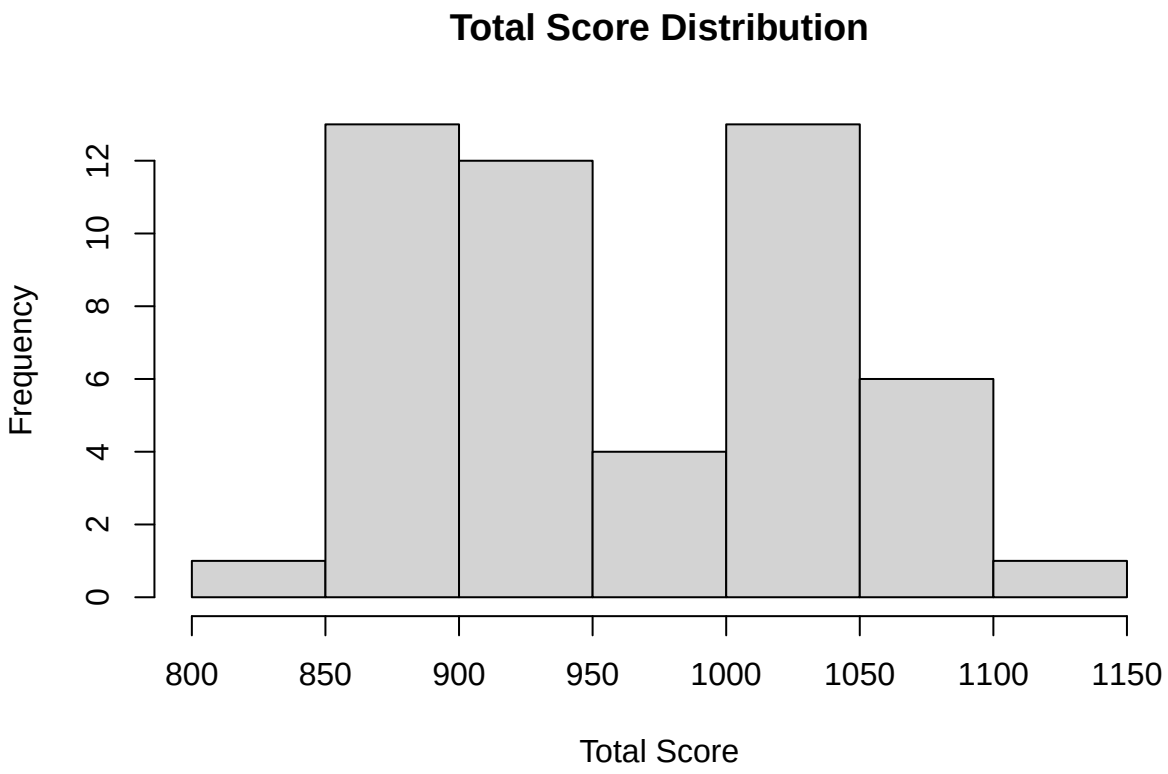
```
hist(sat$verbal, main="Verbal Score Distribution", xlab="Verbal Score")
```



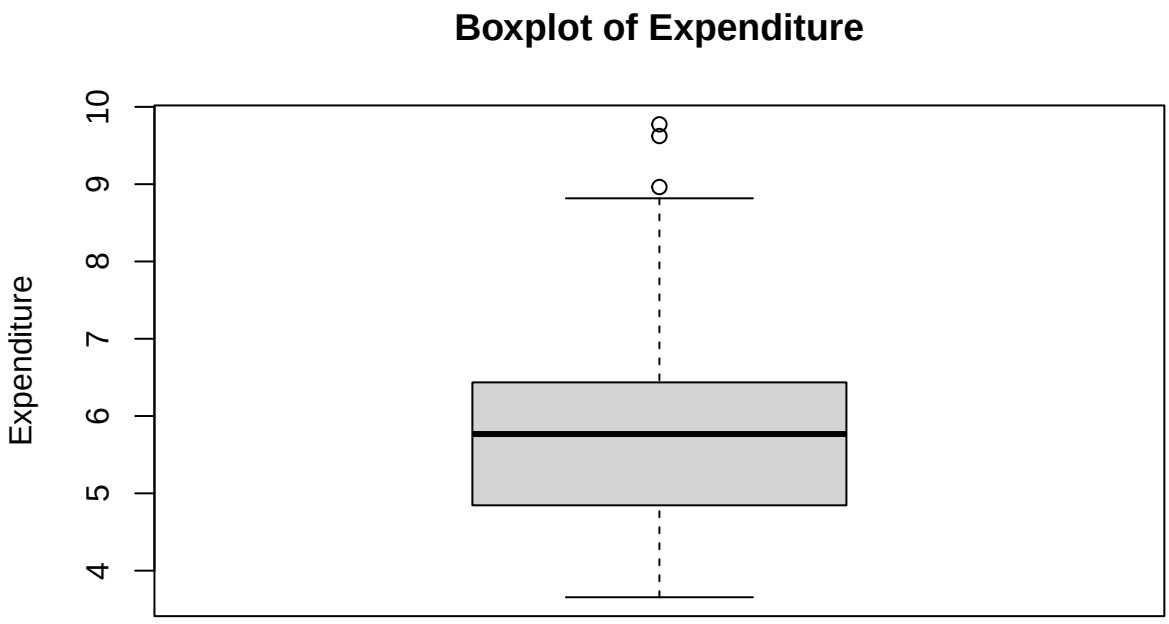

```
hist(sat$math, main="Math Score Distribution", xlab="Math Score")
```




```r
hist(sat$total, main="Total Score Distribution", xlab="Total Score")
```

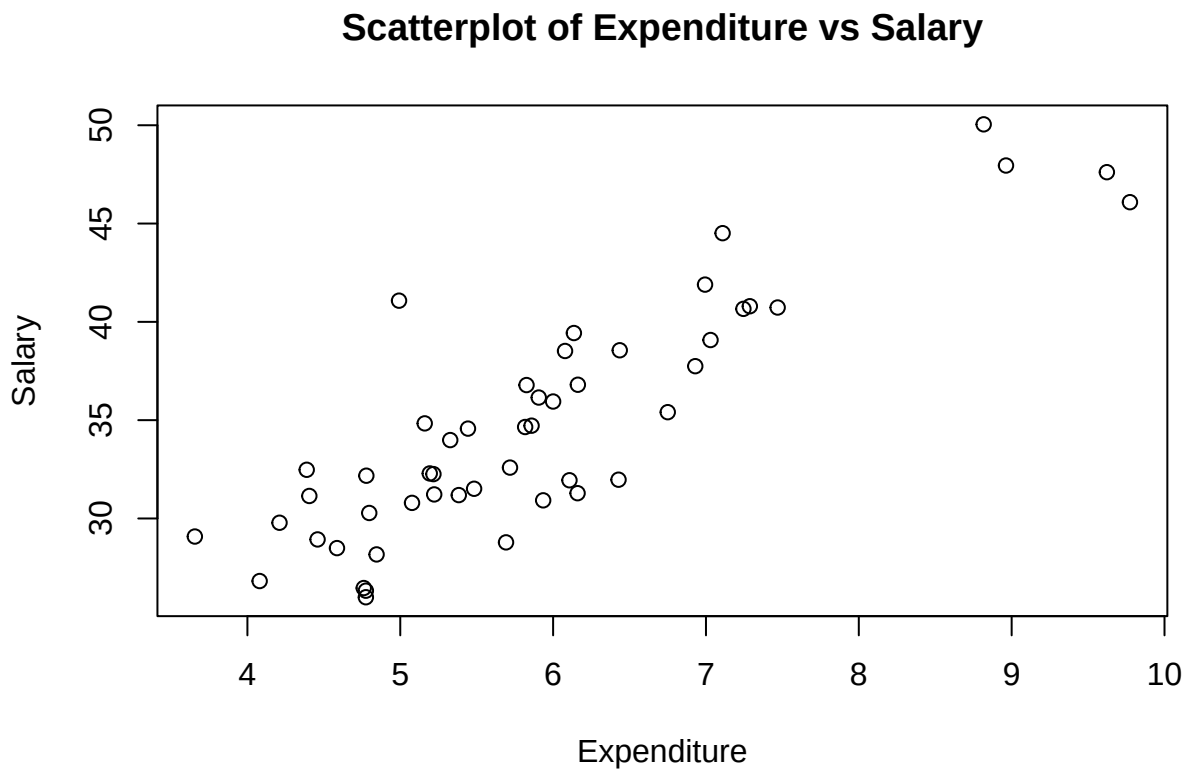



```
boxplot(sat$expend, main="Boxplot of Expenditure", ylab="Expenditure")
```

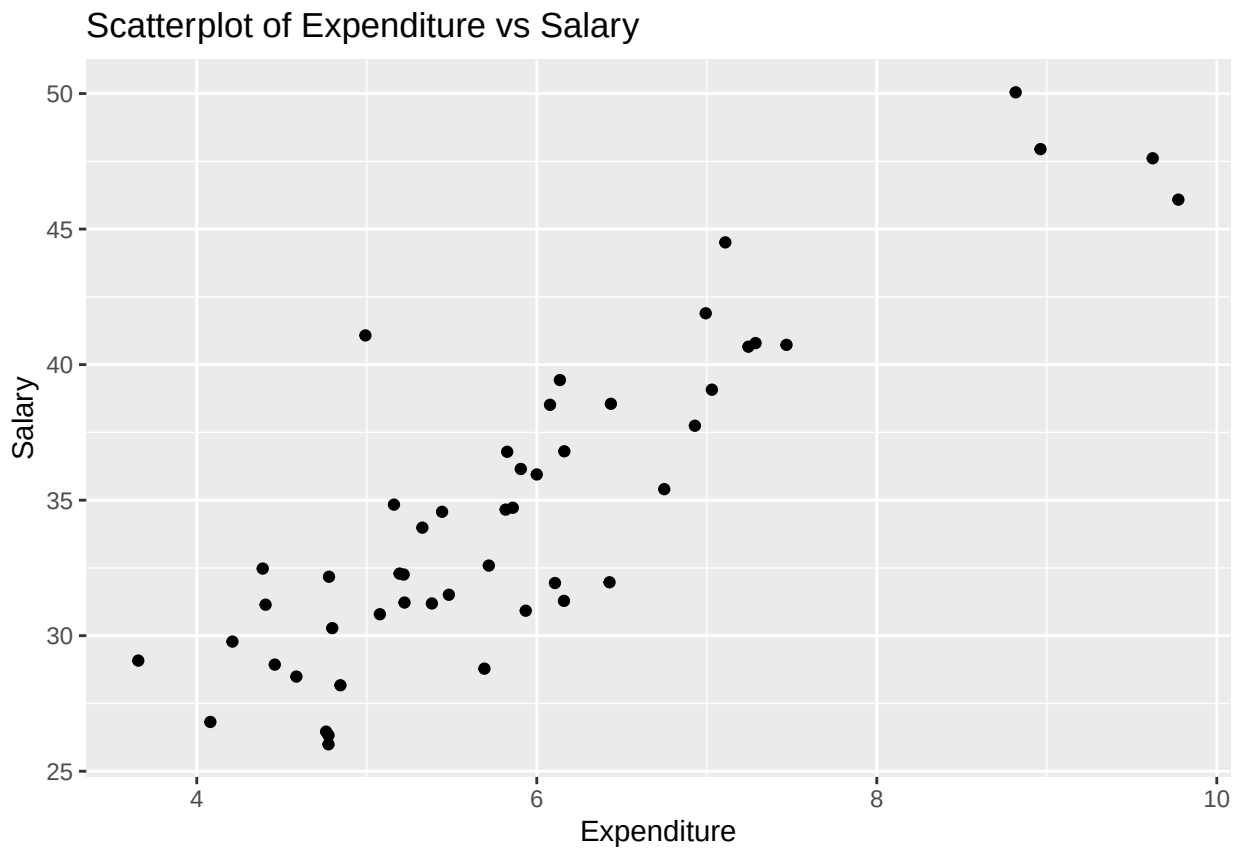


Scatterplot of Expenditure vs Salary

```
plot(sat$expend, sat$salary, main="Scatterplot of Expenditure vs Salary", xlab="Expenditure", ylab="Sala
```




```
ggplot(sat, aes(x = expend, y = salary)) +
  geom_point() +
  labs(title = "Scatterplot of Expenditure vs Salary", x = "Expenditure", y = "Salary")
```



5-The dataset divusa contains data on divorces in the United States from 1920 to 1996. Make a numerical and graphical summary of the data as in the first question.

```
head(divusa)
```

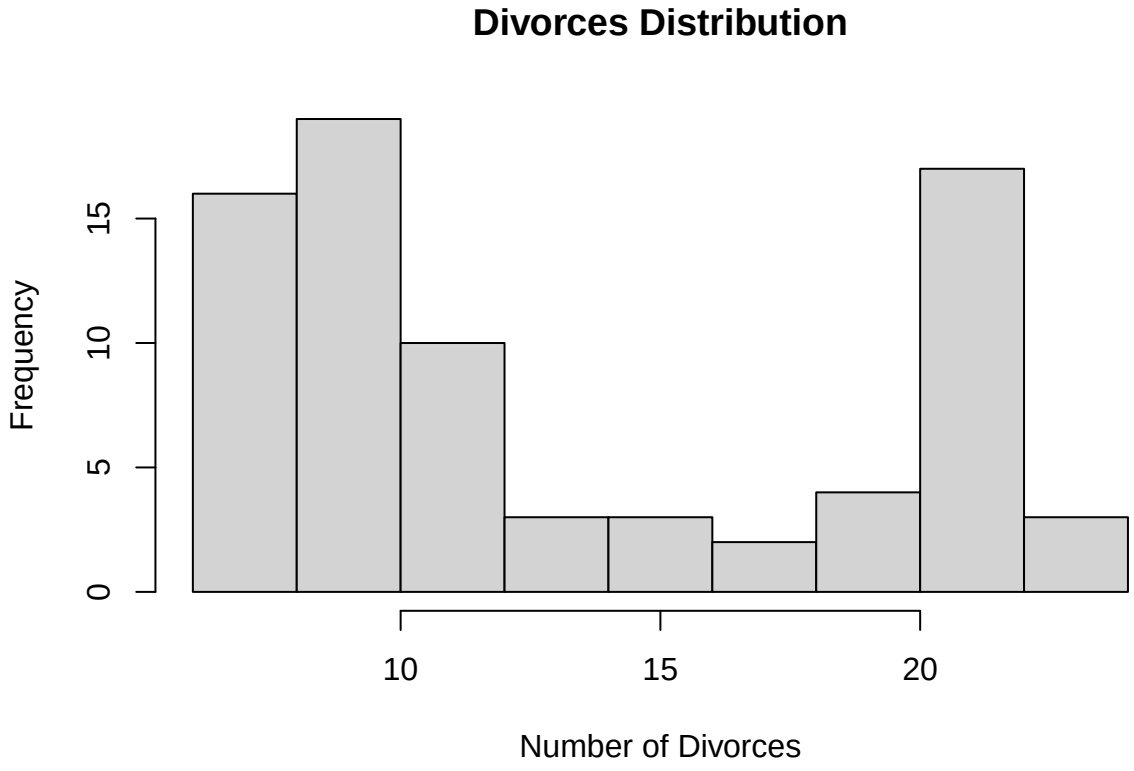
```
##   year divorce unemployed femlab marriage birth military
## 1 1920     8.0        5.2  22.70     92.0 117.9   3.2247
## 2 1921     7.2       11.7  22.79     83.0 119.8   3.5614
## 3 1922     6.6        6.7  22.88     79.7 111.2   2.4553
## 4 1923     7.1        2.4  22.97     85.2 110.5   2.2065
## 5 1924     7.2        5.0  23.06     80.3 110.9   2.2889
## 6 1925     7.2        3.2  23.15     79.2 106.6   2.1735
```

```
summary(divusa)
```

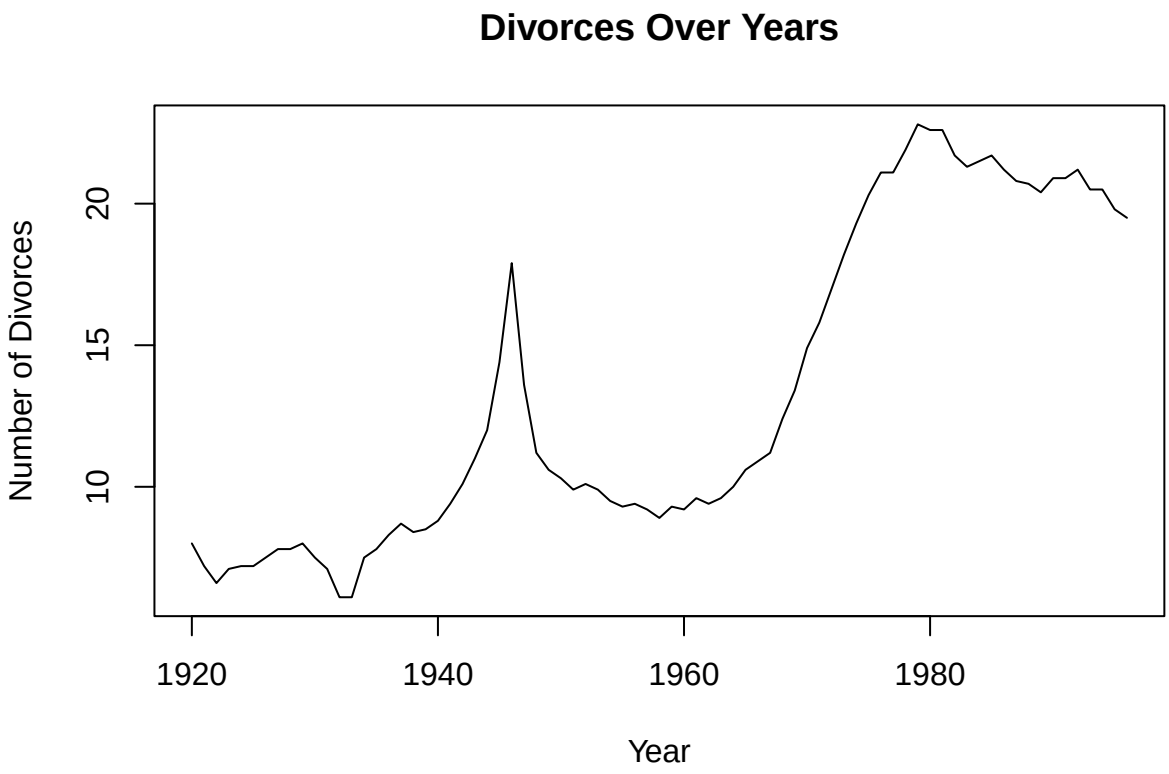
```
##       year         divorce        unemployed         femlab
##  Min.   :1920   Min.   : 6.10   Min.   : 1.200   Min.   :22.70
##  1st Qu.:1939   1st Qu.: 8.70   1st Qu.: 4.200   1st Qu.:27.47
##  Median :1958   Median :10.60   Median : 5.600   Median :37.10
##  Mean   :1958   Mean   :13.27   Mean   : 7.173   Mean   :38.58
##  3rd Qu.:1977   3rd Qu.:20.30   3rd Qu.: 7.500   3rd Qu.:47.80
##  Max.   :1996   Max.   :22.80   Max.   :24.900   Max.   :59.30
##     marriage          birth          military
##  Min.   : 49.70   Min.   : 65.30   Min.   : 1.940
##  1st Qu.: 61.90   1st Qu.: 68.90   1st Qu.: 3.469
```

```
##  Median : 74.10   Median : 85.90   Median : 9.102
##  Mean   : 72.97   Mean   : 88.89   Mean   :12.365
##  3rd Qu.: 80.00   3rd Qu.:107.30   3rd Qu.:14.266
##  Max.   :118.10   Max.   :122.90   Max.   :86.641
```

```
hist(divusa$divorce, main="Divorces Distribution", xlab="Number of Divorces")
```

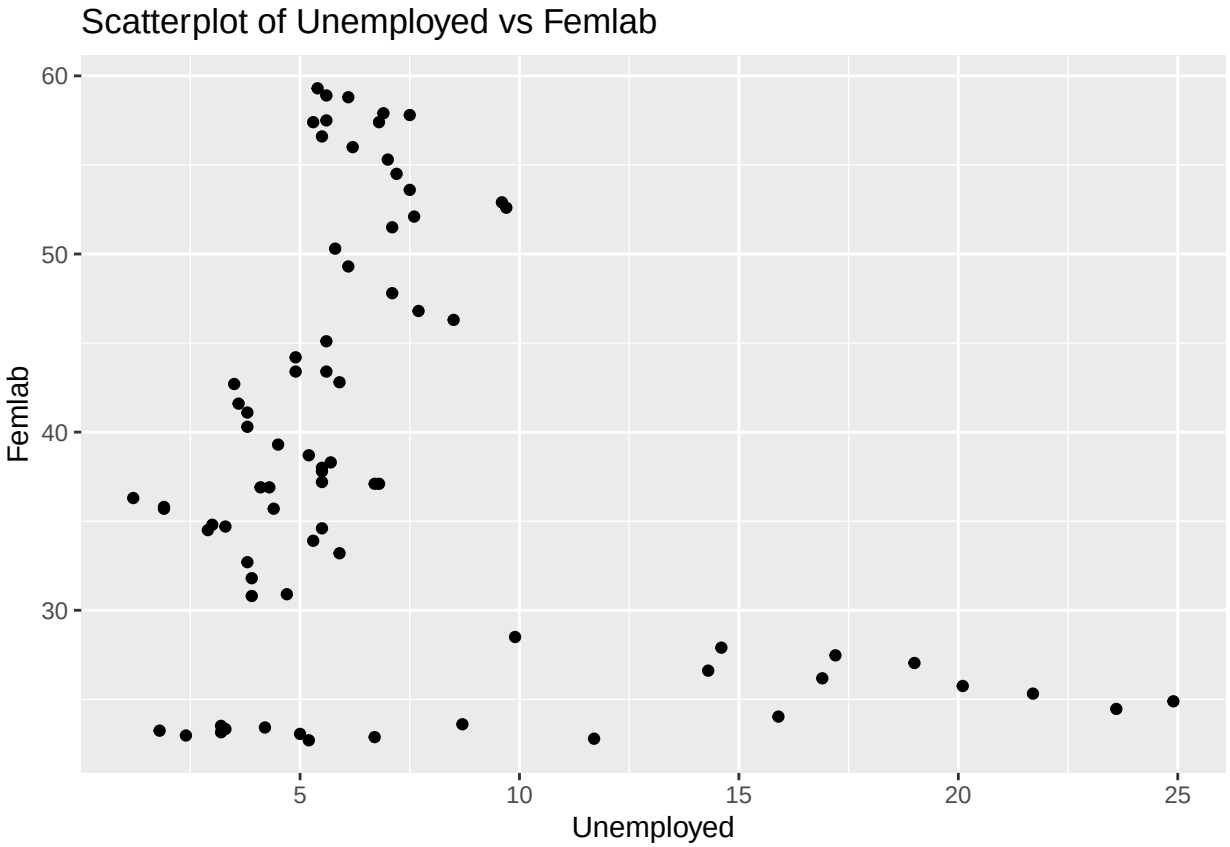



```
plot(divusa$year, divusa$divorce, type="l", main="Divorces Over Years", xlab="Year", ylab="Number of Di
```




```
plot(divusa$unemployed, divusa$femlab, main="Scatterplot of Unemployed vs Femlab", xlab="Unemployed", y
```


```
library(ggplot2)
ggplot(divusa, aes(x = unemployed, y = femlab)) +
  geom_point() +
  labs(title = "Scatterplot of Unemployed vs Femlab", x = "Unemployed", y = "Femlab")
```


Scatterplot of Unemployed vs Femlab
Femlab
60
50
40
30
5
10
15
20
25
Unemployed