Assignment 1

79546 - Stephen K. Ng'etich

Contents

1 Pre-requisite **1**

1.1 Install package 1

1.2 Load package 1

1.3 Load dataset 2

2 Exploratory Data Analysis **2**

2.1 Salaries Distribution 2

2.2 Age Distribution 4

2.3 Experience Distribution 6

2.4 Age Distribution by Gender 7

2.5 Salary Distribution per Region 9

2.6 Correlation among the variables 10

3 Model Estimation **11**

1 Pre-requisite

1.1 Install package

```
#install.packages("readxl")
#install.packages('dplyr')
```

1.2 Load package

```
# Clear variables
rm(list=ls())

library(readxl)
library(dplyr)
```

1.3 Load dataset

```
# Load Dataset
dataset <- read_excel("dataset/Data1.xlsx")

salaries = dataset$Salaries
experience = dataset$`Years of Experience`
age = dataset$Age
gender = dataset$Gender
region = dataset$Region
```

2 Exploratory Data Analysis

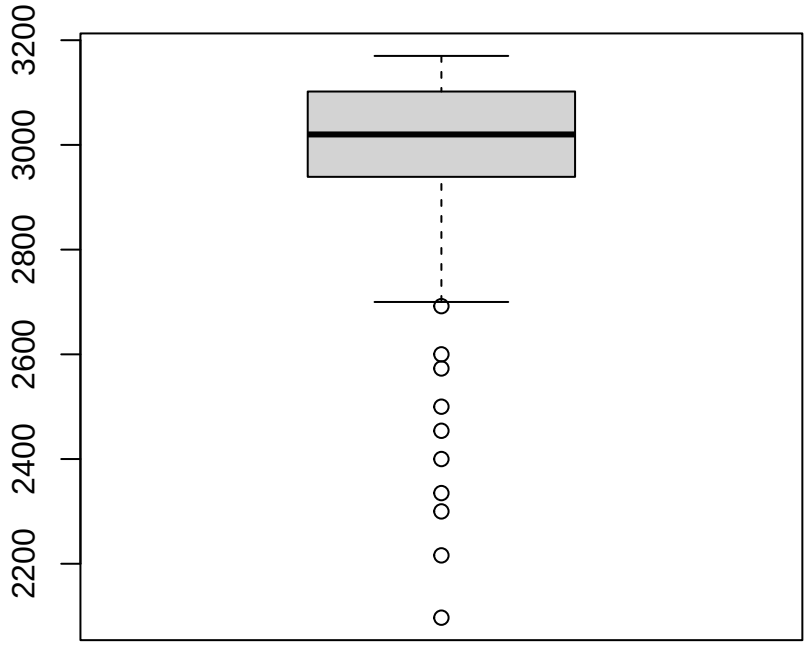
View the summary statistics of each numeric column

```
summary(dataset)
```

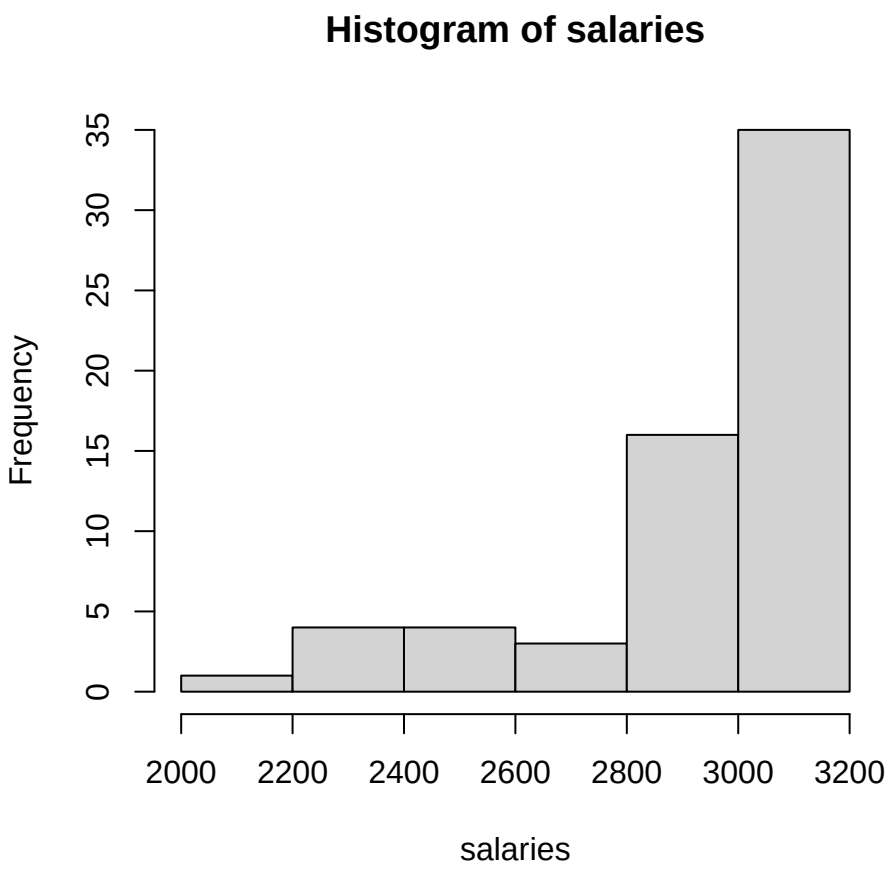
```
##     Salaries    Years of Experience      Age           Gender
##  Min.   :2097   Min.   : 4.00       Min.   :26.00   Length:63
##  1st Qu.:2939   1st Qu.:10.65       1st Qu.:29.00   Class :character
##  Median :3020   Median :12.20       Median :31.00   Mode  :character
##  Mean   :2939   Mean   :11.67       Mean   :30.79
##  3rd Qu.:3102   3rd Qu.:13.00       3rd Qu.:32.00
##  Max.   :3170   Max.   :15.80       Max.   :37.00
##      Region
##  Min.   : 1.0
##  1st Qu.: 1.0
##  Median : 2.0
##  Mean   : 2.2
##  3rd Qu.: 3.0
##  Max.   :15.6
```

2.1 Salaries Distribution

```
boxplot(salaries)
```



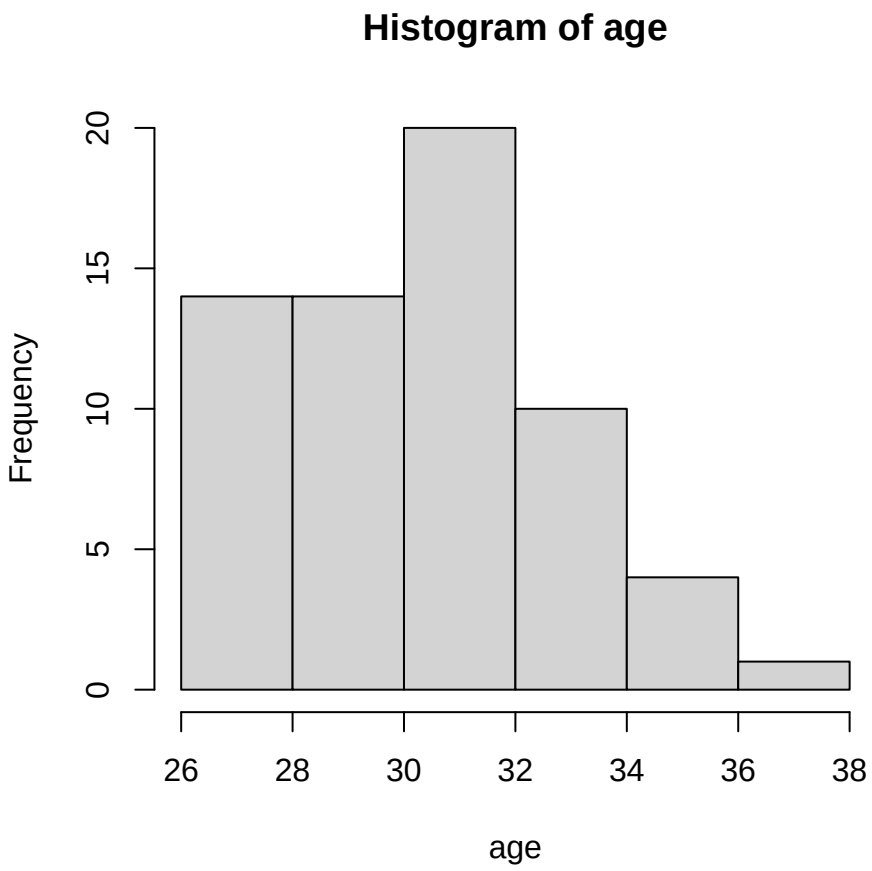

```
hist(salaries)
```



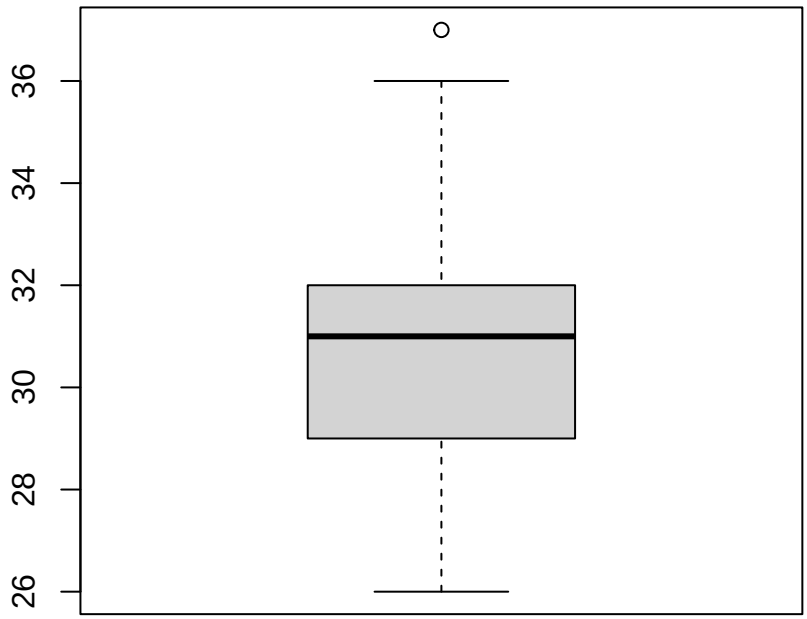
From the histogram and the box plots, the data show the following:
1. The salaries data is continuous
2. A lot of people are earning 3000 to 3200.
3. Salaries is negative skewed because the mean is less than the median

2.2 Age Distribution

```
hist(age)
```



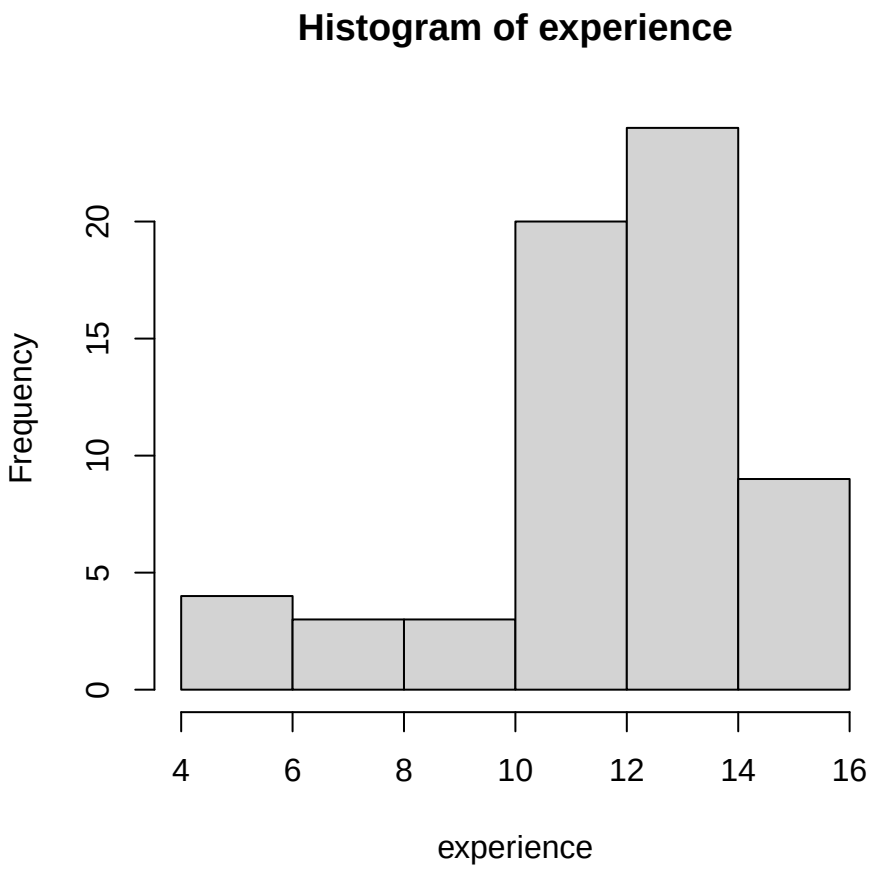

```
boxplot(age)
```



Age is negatively skewed with people 30 to 32 highly represented.

2.3 Experience Distribution

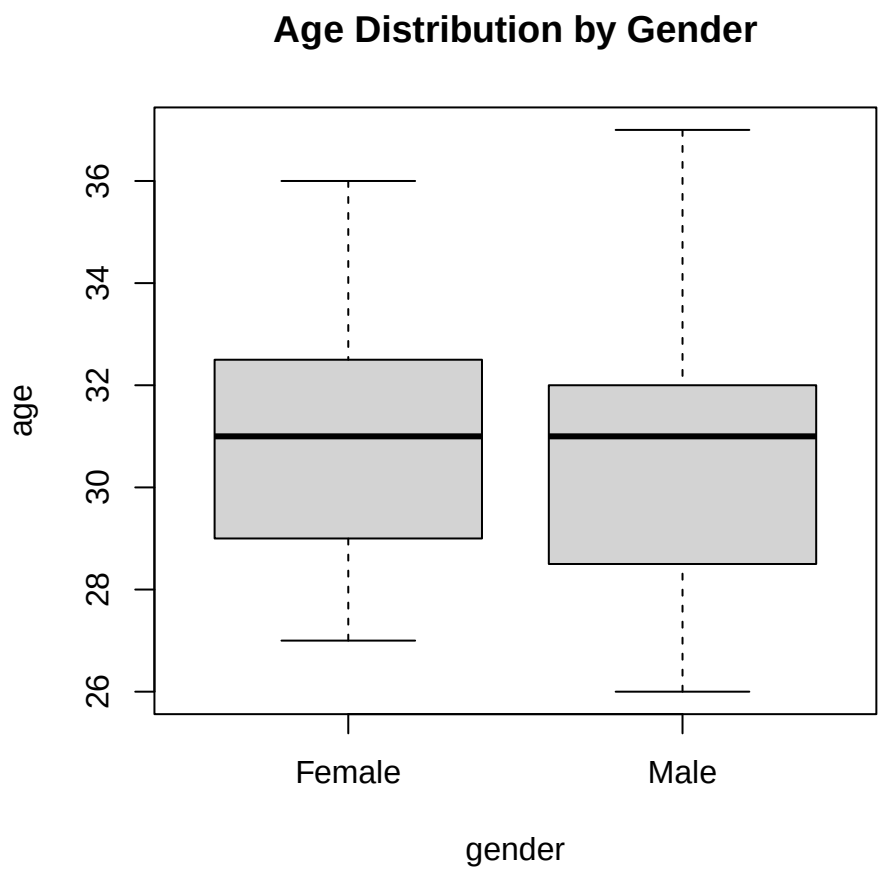
```
hist(experience)
```



This data represent people have more than `10 years experience`

2.4 Age Distribution by Gender

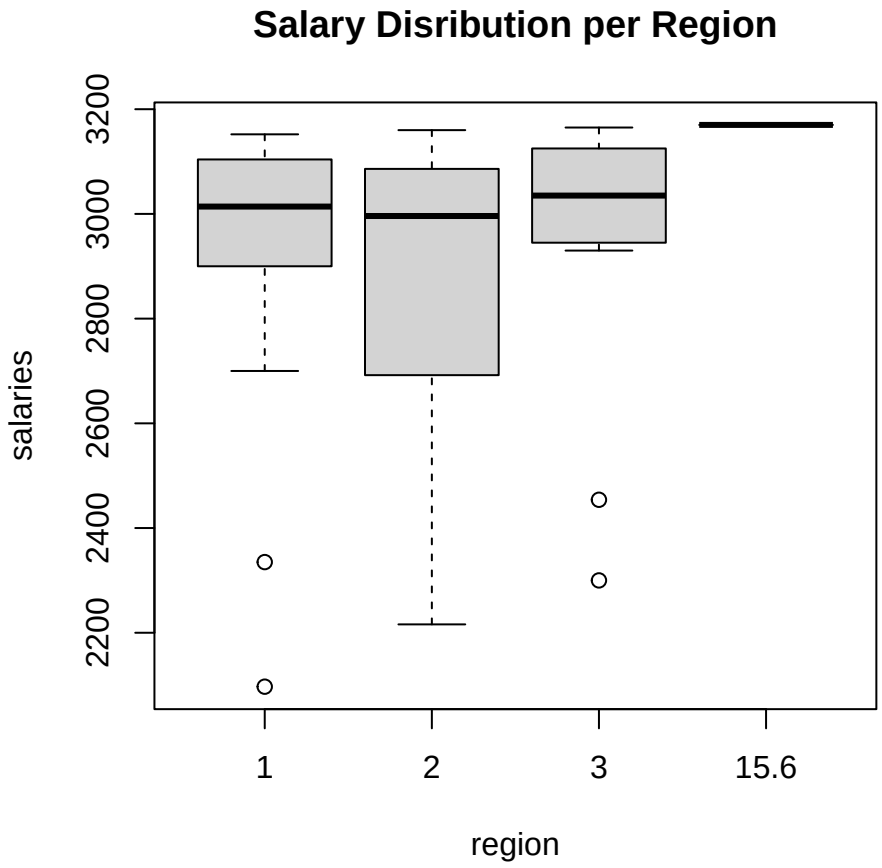
```
boxplot(age~gender,
        main="Age Distribution by Gender")
```




```
boxplot(salaries~gender,
        main="Salary distribution by Gender")
```

2.5 Salary Distribution per Region

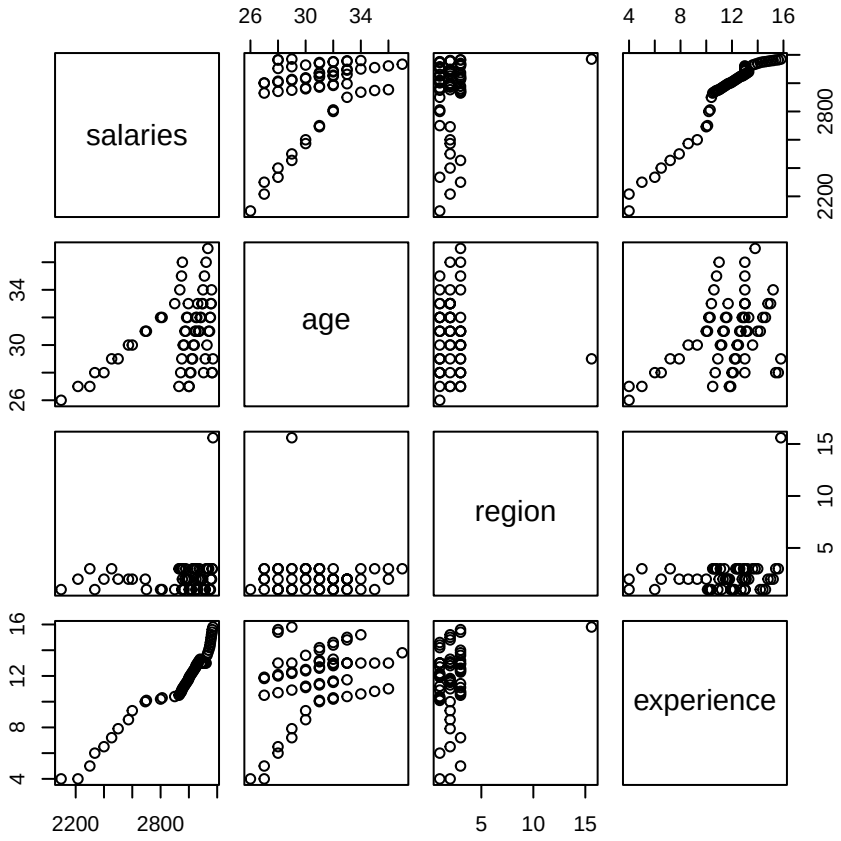
```
boxplot(salaries~region, main="Salary Disribution per Region")
```



Region 2 has the highest salary variation

2.6 Correlation among the variables

```
plot(data.frame(salaries,age,region,experience))
```



Salaries, age and experience are correlated this means that experience is a reasonable predictor of salaries

3 Model Estimation

```
lm1 =  lm(formula= salaries~experience+age)
summary(lm1)
```

```
##
## Call:
## lm(formula = salaries ~ experience + age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -134.49  -48.05   17.41   52.87  121.44
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1679.844    103.184  16.280   <2e-16 ***
## experience    91.176      3.478  26.212   <2e-16 ***
## age            6.347      3.654   1.737   0.0875 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.06 on 60 degrees of freedom
## Multiple R-squared:  0.9357, Adjusted R-squared:  0.9336
## F-statistic: 436.6 on 2 and 60 DF,  p-value: < 2.2e-16
```

Years of Experience is a good predictor of salaries since the p-value is less than 0.05

$$Salaries = 1679.855 , + , 91`Years of experience`$$