R Notebook

```
knitr::opts_chunk$set(fig.width=6, fig.height=6)
```

Assignment 1

Pre-requisite

install package

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

load package

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

library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.1.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3

##
## 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
```

Load dataset

```r
# 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
```

Exploratory Data Analysis

View the summary of each numeric column

```r
summary(dataset)
```

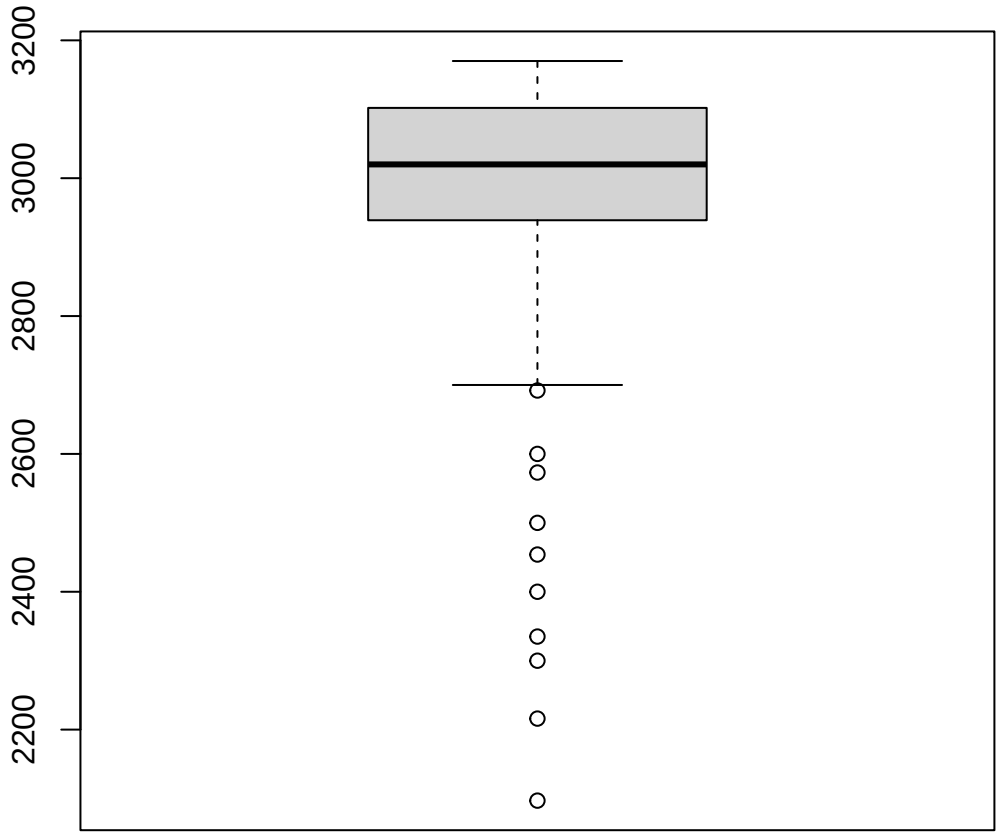
```
##     Salaries    Years of Experience      Age           Gender
##  Min.   :2097   Min.   : 4.00       Min.   :26.00   Length:63
##  1st Qu.:2939   1st Qu.:10.65       1st Qu.:29.00   Class :character
##  Median :3020   Median :12.20       Median :31.00   Mode  :character
##  Mean   :2939   Mean   :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
```

Salaries

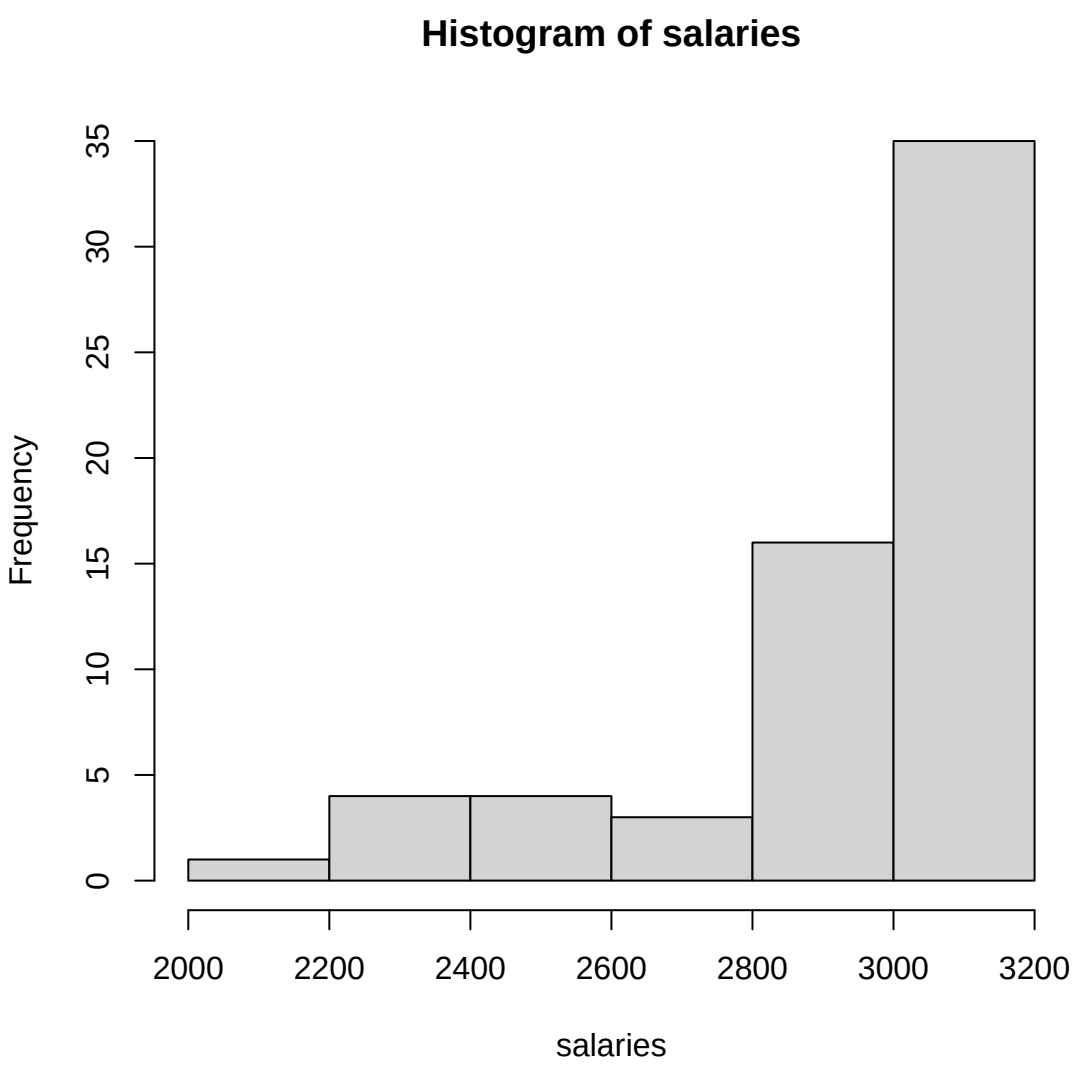
```r
summary(salaries)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2097    2939    3020    2939    3102    3170
```

```r
boxplot(salaries)
```




```
hist(salaries)
```

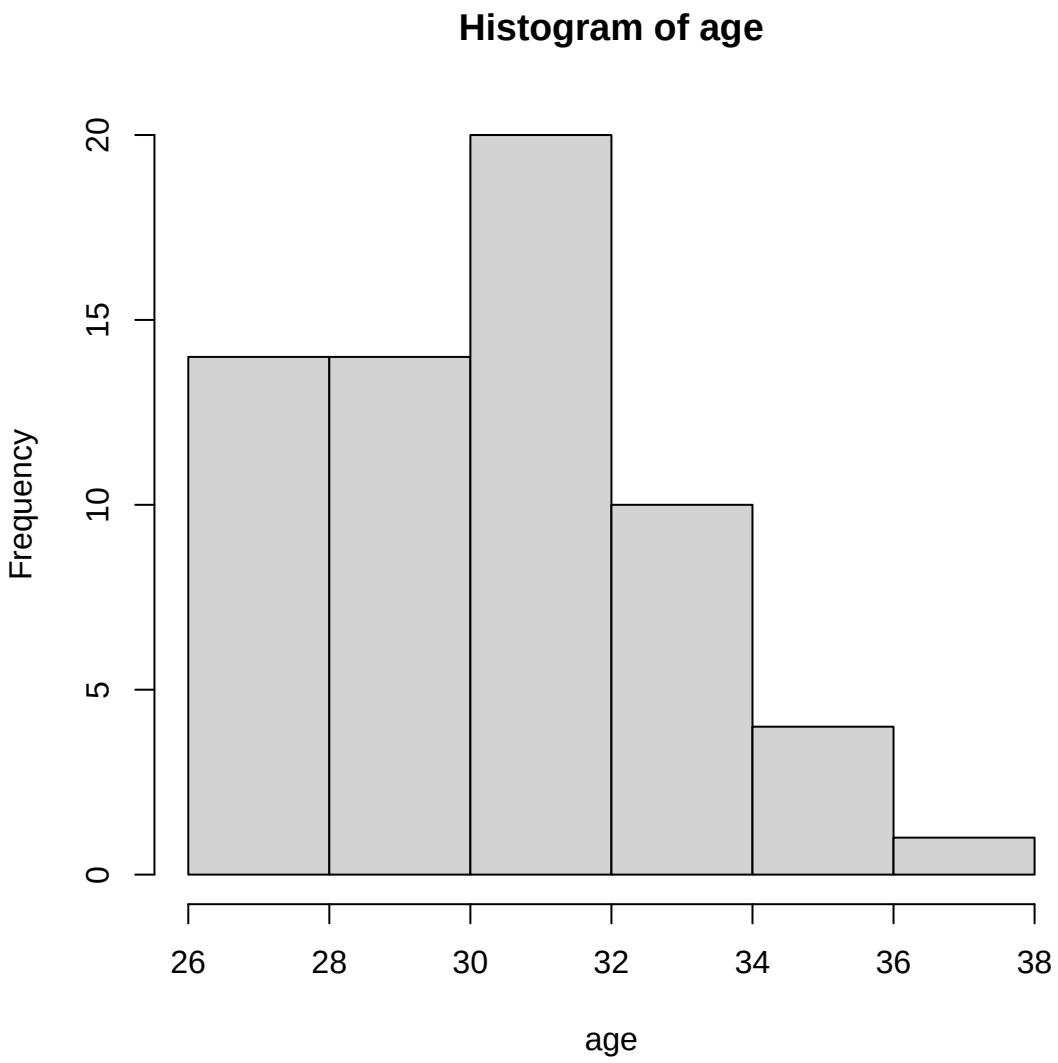


From the histogram and the box plots, the data show the following:

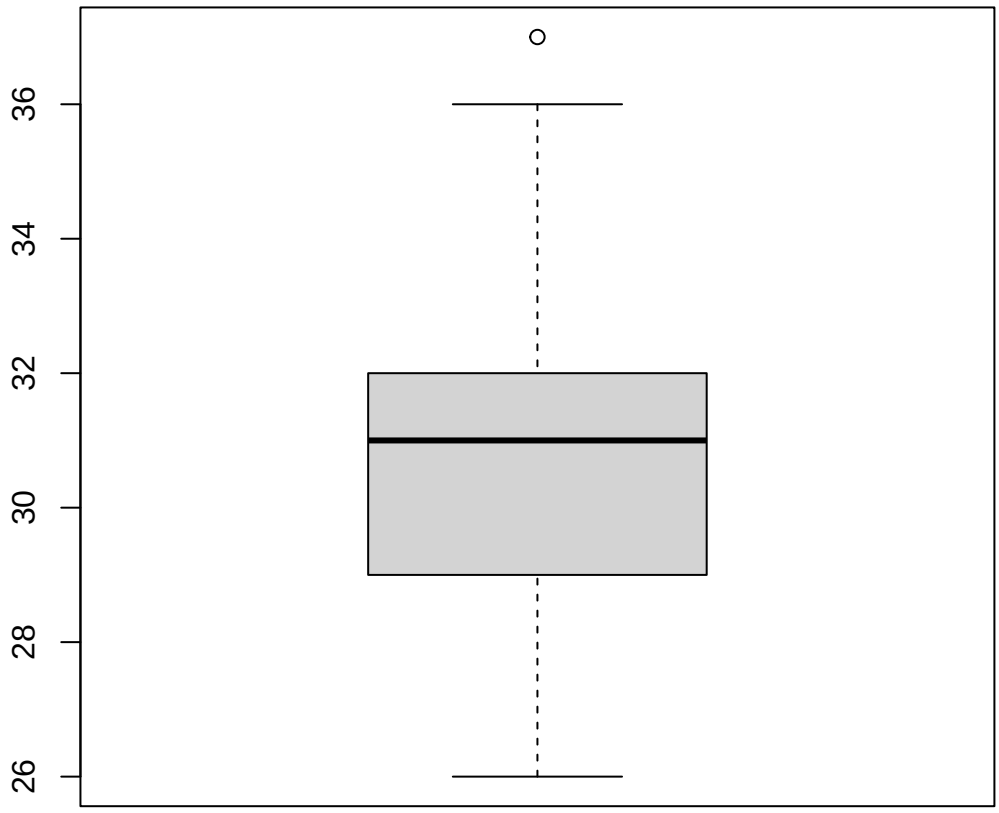
1. The salaries data is continuous
2. A lot of people are earning 3000 to 3200.
3. Salaries is negative skewed because the mean is less than the median

Age

```
hist(age)
```

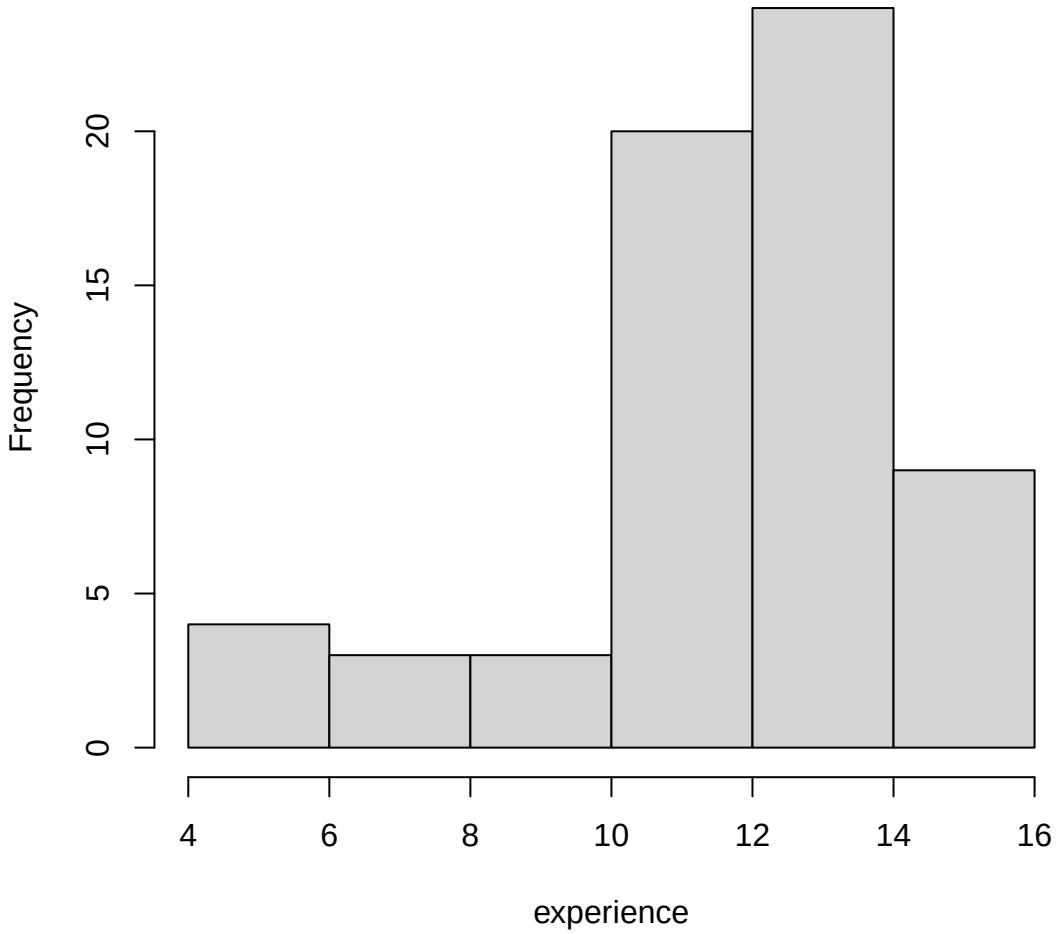



```
boxplot(age)
```



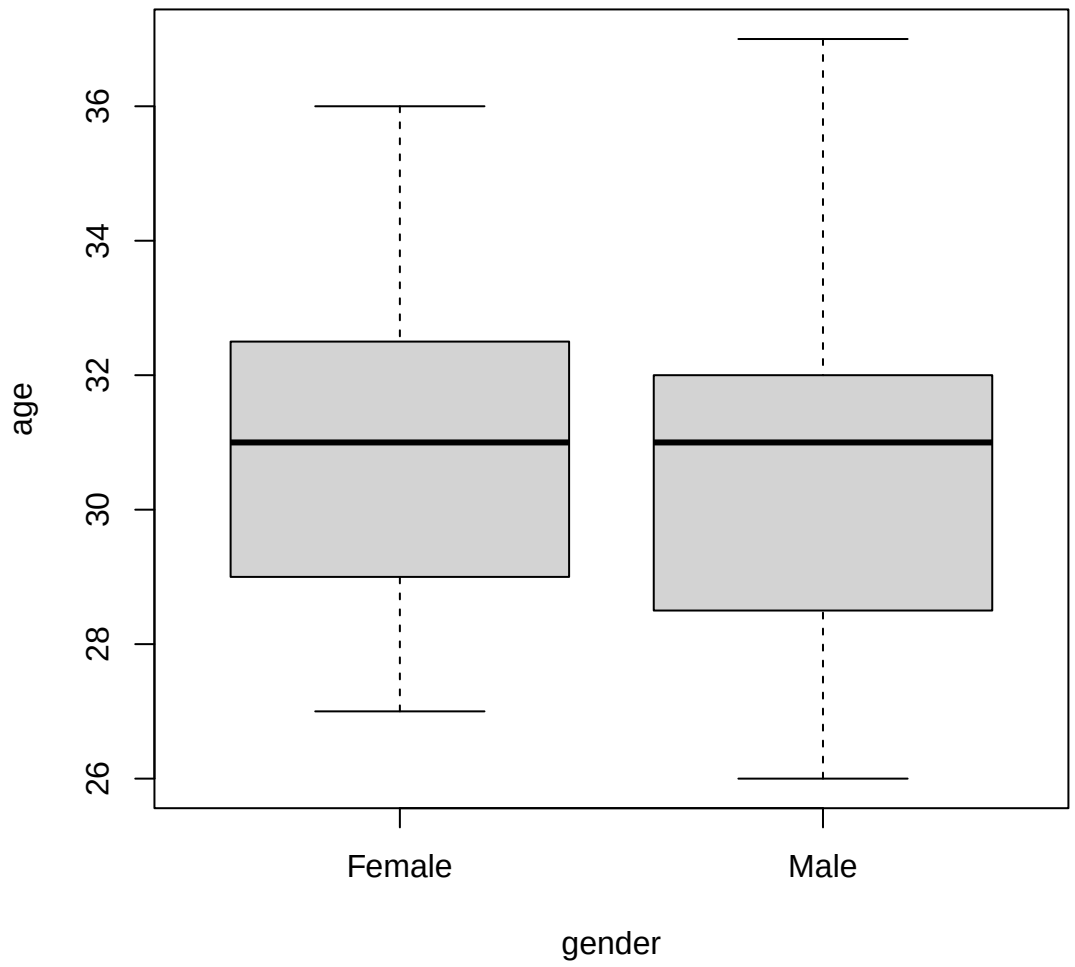
Age is negatively skewed with people 30 to 32 highly represented.

```
hist(experience)
```

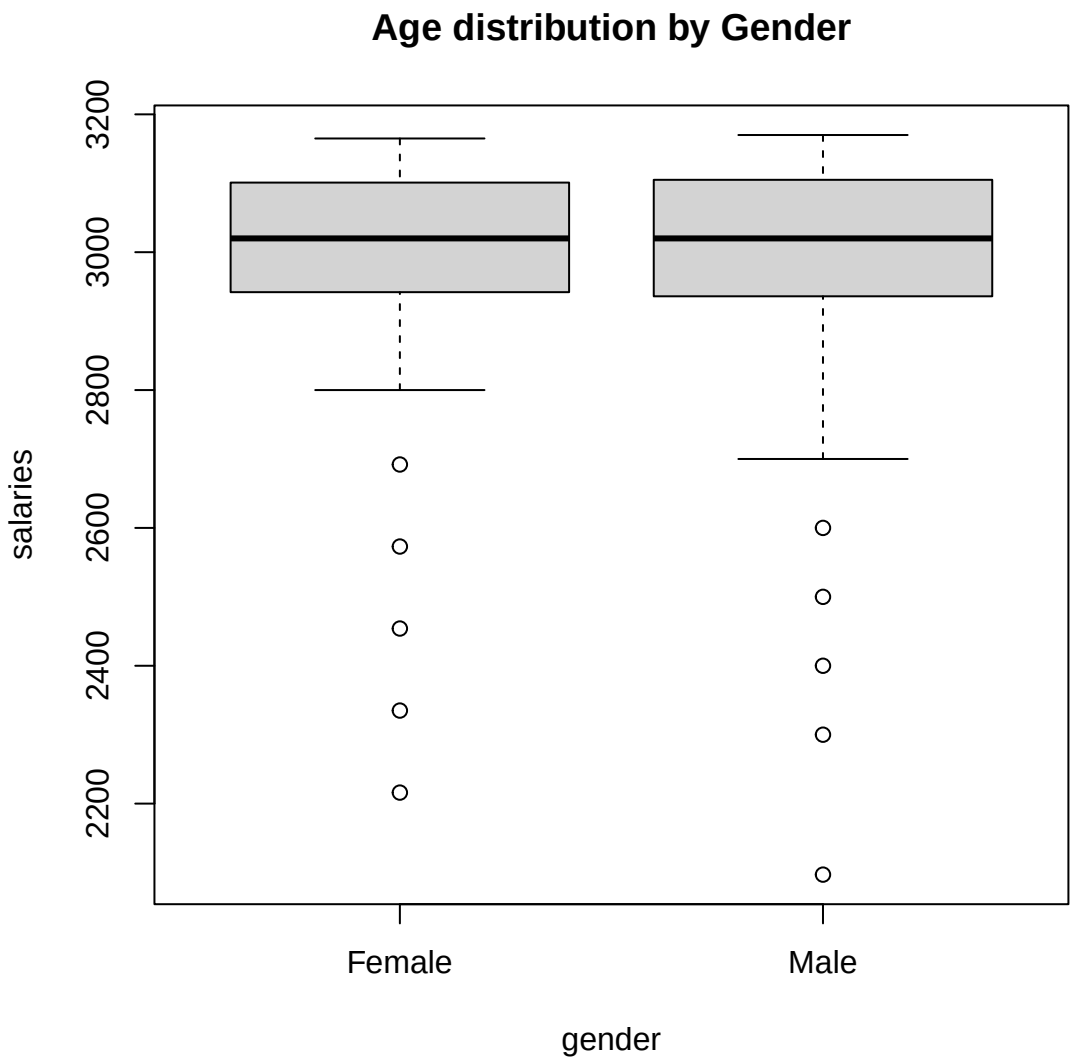


This data represent people have more than 10 years experience

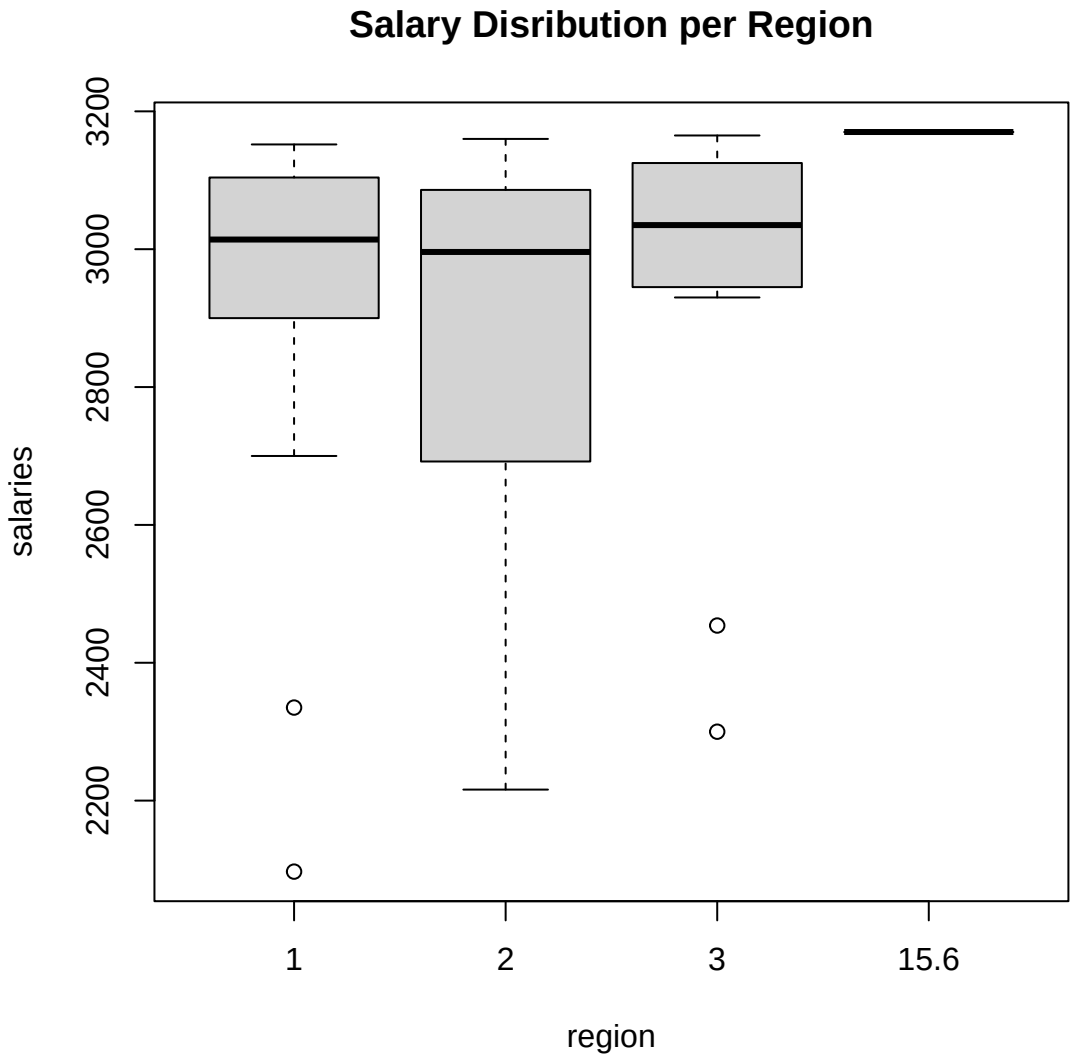
```
boxplot(age~gender,
        main="Age distribution by Gender")
```




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



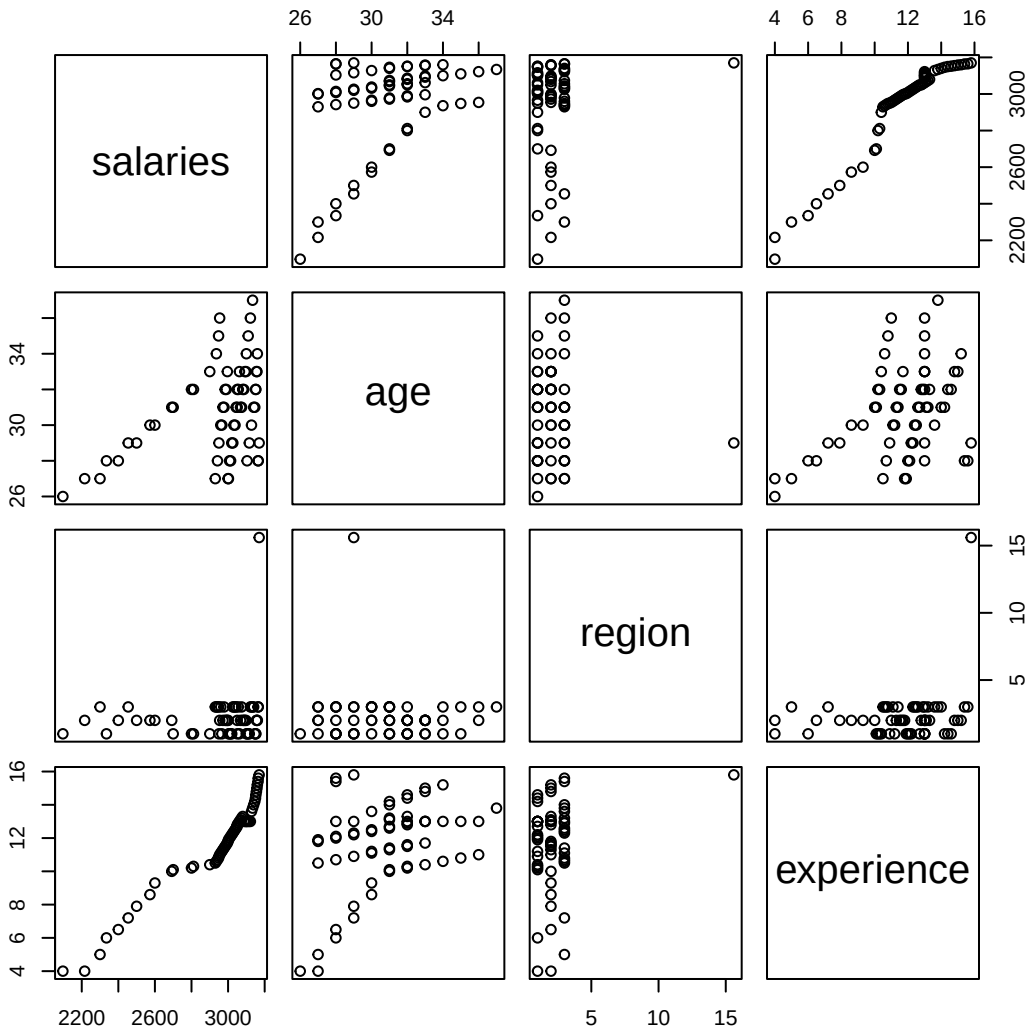

```
boxplot(salaries~region, main="Salary Disribution per Region")
```



Region

2 has the highest salary variation

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



Salaries

and experience are correlated this means that experience is a reasonable predictor of size

```
simple.regression = lm(formula= salaries~experience)
summary(simple.regression)
```

```
##
## Call:
## lm(formula = salaries ~ experience)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -156.08  -50.08   14.22   50.31  100.24
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1846.491     38.583   47.86   <2e-16 ***
## experience    93.645      3.227   29.02   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.14 on 61 degrees of freedom
## Multiple R-squared:  0.9325, Adjusted R-squared:  0.9314
## F-statistic: 842.3 on 1 and 61 DF,  p-value: < 2.2e-16
```

```
lm1 =  lm(dataset,formula = Salaries~.)
summary(lm1)
```

```
##
## Call:
## lm(formula = Salaries ~ ., data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -127.39  -40.62   12.87   45.93  127.59
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1714.523    104.358  16.429   <2e-16 ***
## `Years of Experience`   92.780      3.559  26.072   <2e-16 ***
## Age                      5.253      3.667   1.433   0.1574
## GenderMale              -4.142     16.587  -0.250   0.8037
## Region                  -7.999      4.553  -1.757   0.0842 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.37 on 58 degrees of freedom
## Multiple R-squared:  0.9391, Adjusted R-squared:  0.9349
## F-statistic: 223.7 on 4 and 58 DF,  p-value: < 2.2e-16
```