Assignment 1 Linear Models

Getrude Gichuhi

28/03/2022

Use Salaries as your response variable. Ensure you convert gender and region to factor variables

1) Use R markdown and submit a PDF document
2) Explore the data by doing relevant histograms, boxplots, scatter plots etc and interpret your results
3) Fit a linear regression with salary as response variable and all the rest at exploratory variables
4) Interpret all your output

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

load the libraries to use

```
library("readxl")
```

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

loading the dataset

```
df <- read_excel("Data1.xlsx")
print(head(df))
```

```
## # A tibble: 6 x 5
##   Salaries `Years of Experience`   Age Gender Region
##      <dbl>                 <dbl> <dbl> <chr>   <dbl>
## 1     2097                   4      26 Male        1
## 2     2216                   4      27 Female      2
## 3     2300                   5      27 Male        3
## 4     2335                   6      28 Female      1
## 5     2400                   6.5    28 Male        2
## 6     2454                   7.2    29 Female      3
```

```
summary(df)
```

```
##     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
```

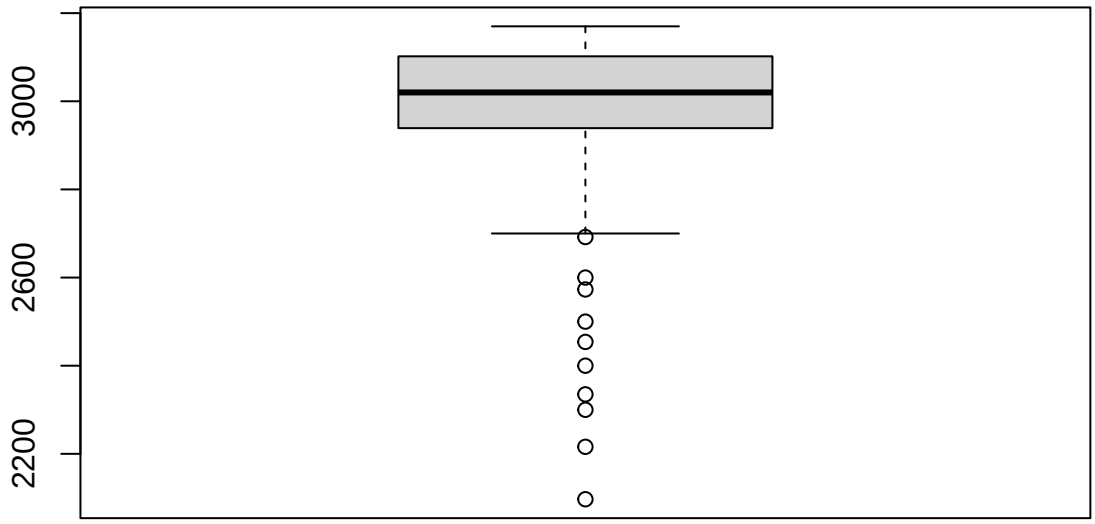
Scatter Plots

```
#plot(df$`Years of Experience`, df$Salaries)

scatter.smooth(x=df$`Years of Experience`, y=df$Salaries, main="Salaries ~ Experience")
```

Box Plots

```
boxplot(df$Salaries)
```




```
boxplot(df$Salaries,
main = "Salaries and Years of experience",
xlab = "`Years of Experience`",
ylab = "Salaries",
col = "orange",
border = "brown",
horizontal = TRUE,
notch = TRUE
)
```

Histogram

```
Salaries <- df$Salaries
hist(Salaries)
```


```
# histogram with added parameters
hist(Salaries,
main="Maximum Salaries",
xlab="`Years of Experience`",
ylab= "Salaries",
col="darkmagenta",
freq=FALSE
)
```

Salaries increases with the number of years of experience. The more experienced, the more salary is earned.

```
#plot(df$Age, df$Salaries)

scatter.smooth(x=df$Age, y=df$Salaries, main="Salaries ~ Age")
```


```
boxplot(df$Salaries,
main = "Salaries and Years of experience",
xlab = "Age",
ylab = "Salaries",
col = "black",
border = "blue",
horizontal = TRUE,
notch = TRUE
)
```


```
hist(Salaries,
main="Maximum Salaries",
xlab="Age",
ylab= "Salaries",
col="darkgreen",
freq=FALSE
)
```

From ages 28 to around 33 more people seems to earn more from around 3000 - 3200. The data shows that the more energetic one is and young, the salaries are more, while only a handful of them still earn more as the ages progresses.
Box Plots

```
library(caret)
```

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

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(tidyverse)
```

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

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4
```

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

## Warning: package 'forcats' was built under R version 4.1.3

## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
```

```
library(glmnet)
```

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

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-3
```

```
library(dplyr)
```

```
df <- read_excel("Data1.xlsx")
view(df)
```

```
linearMod <- lm(Salaries ~`Years of Experience` +  Age + Region, data=df)  # build linear regression mo
print(linearMod)
```

```
##
## Call:
## lm(formula = Salaries ~ `Years of Experience` + Age + Region,
##     data = df)
##
## Coefficients:
##           (Intercept)  `Years of Experience`                    Age
##              1711.423                92.792                  5.289
##                Region
##                -8.115
```

Based on the outcome it shous that the Intercept of the salaries is = 1711.423 which is Beta zero. while Beta one is 92.792 * Years of Experience, Beta 2 * Age and Beta 3 * Region.