PRA2 - Tipología y ciclo de vida de los datos: Limpieza y análisis de datos

Sergio Funes, David Ortiz

Junio de 2021

Contents

1. Descripción del dataset

2. Integración y selección de los datos de interés

```
data <- read.csv("recruitment_decision_tree.csv", header = TRUE, stringsAsFactors = TRUE)
head(data)
```

```
##   Serial_no Gender Python_exp Experience_Years    Education Internship Score
## 1         1   Male        Yes                0     Graduate         No  5139
## 2         2   Male         No                1     Graduate         No  4583
## 3         3   Male         No                0     Graduate        Yes  3000
## 4         4   Male         No                0 Not Graduate         No  2583
## 5         5   Male        Yes                0     Graduate         No  6000
## 6         6   Male         No                2     Graduate        Yes  5417
##   Salary...10E4 Offer_History Location Recruitment_Status
## 1             0             1    Urban                  Y
## 2           128             1    Rural                  N
## 3            66             1    Urban                  Y
## 4           120             1    Urban                  Y
```

```
## 5           141             1    Urban                  Y
## 6           267             1    Urban                  Y
```

```
sapply(data, function(x) class(x))
```

```
##          Serial_no             Gender         Python_exp   Experience_Years
##          "integer"           "factor"           "factor"          "integer"
##          Education         Internship              Score      Salary...10E4
##           "factor"           "factor"          "integer"          "integer"
##      Offer_History           Location Recruitment_Status
##          "integer"           "factor"           "factor"
```

```
data <- data[, -(1)]
```

3. Limpieza de los datos

3.1. Ceros y elementos vacíos

```
sapply(data, function(x) sum(is.na(x)))
```

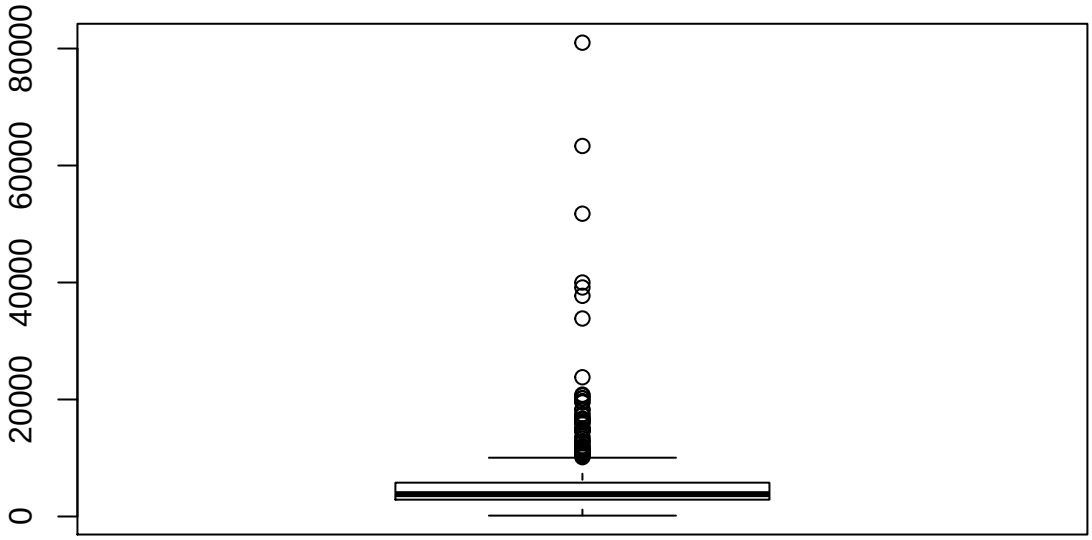
```
##             Gender         Python_exp   Experience_Years          Education
##                  0                  0                 15                  0
##         Internship              Score      Salary...10E4      Offer_History
##                  0                  0                 21                 50
##           Location Recruitment_Status
##                  0                  0
```

```
suppressWarnings(suppressMessages(library(VIM)))

data$Experience_Years <- kNN(data)$Experience_Years
data$Salary...10E4 <- kNN(data)$Salary...10E4
data$Offer_History <- kNN(data)$Offer_History

sapply(data, function(x) sum(is.na(x)))
```

```
##             Gender         Python_exp   Experience_Years          Education
##                  0                  0                  0                  0
##         Internship              Score      Salary...10E4      Offer_History
##                  0                  0                  0                  0
##           Location Recruitment_Status
##                  0                  0
```

3.2. Valores extremos

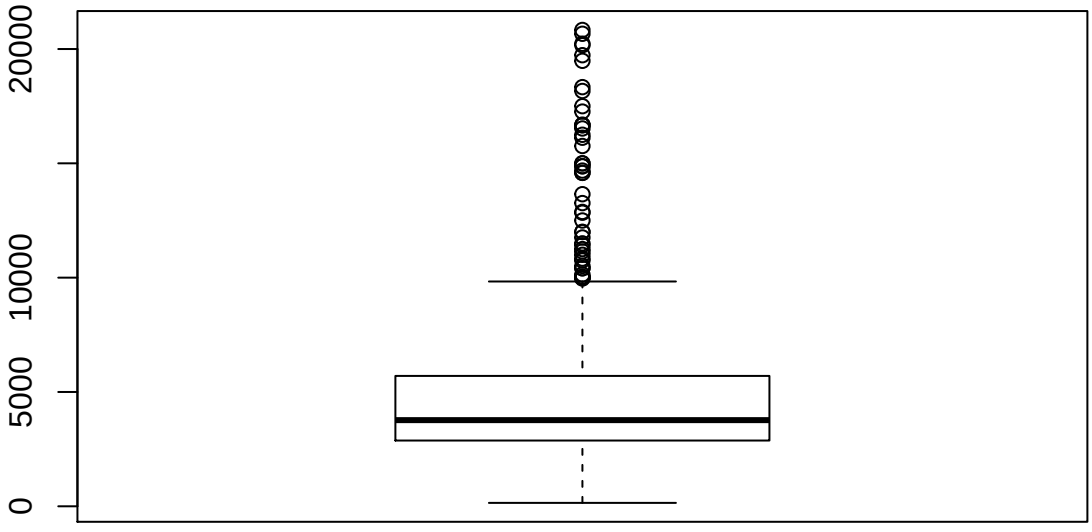
```
boxplot(data$Score)
```



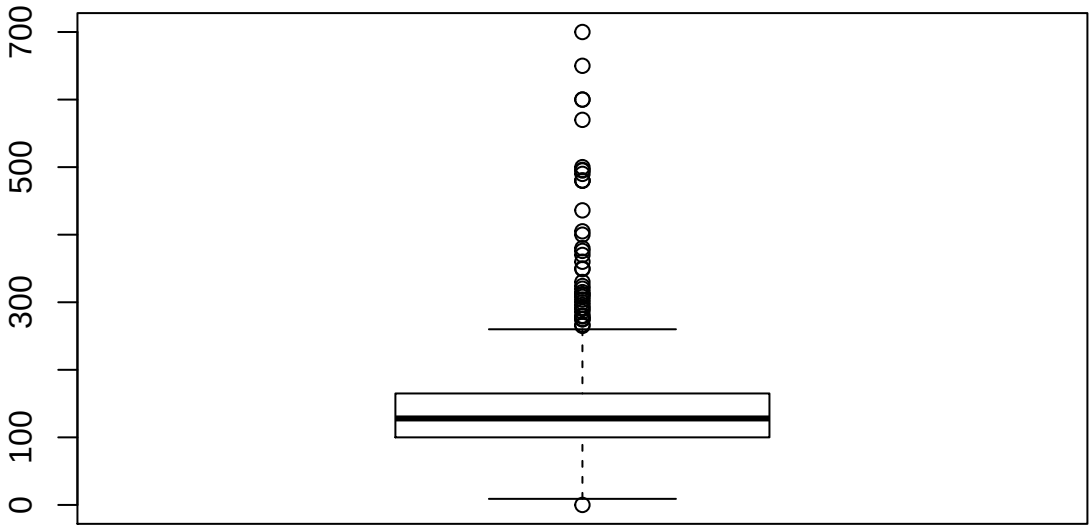

```
boxplot.stats(data$Score)$out
```

```
##  [1] 12841 12500 11500 10750 13650 11417 14583 10408 23803 10513 20166 14999
## [13] 11757 14866 39999 51763 33846 39147 12000 11000 16250 14683 11146 14583
## [25] 20667 20233 15000 63337 19730 15759 81000 14880 12876 10416 37719 16692
## [37] 16525 16667 10833 18333 17263 20833 13262 17500 11250 18165 10139 19484
## [49] 16666 16120 12000
```

```
data$Score[data$Score > 23000] <- NA

data$Score <- kNN(data)$Score

boxplot(data$Score)
```



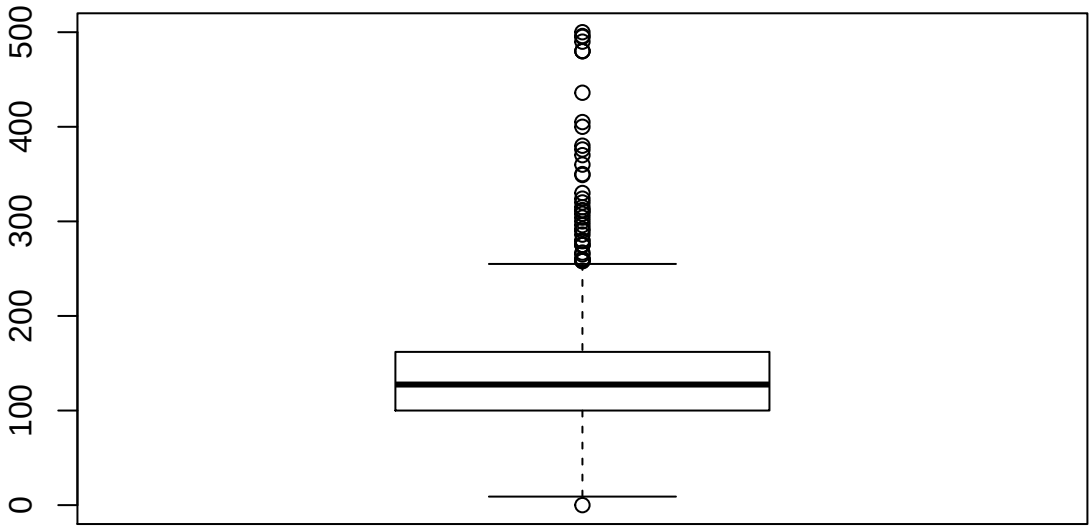

```
boxplot(data$Salary...10E4)
```




```
boxplot.stats(data$Salary...10E4)$out
```

```
##  [1]   0 267 349 315 320 286 312 265 370 650 290 600 275 700 495 280 279 304 330
## [20] 436 480 300 376 490 308 570 380 296 275 360 405 500 480 311 480 400 324 600
## [39] 275 292 350 496
```

```
data$Salary...10E4[data$Salary...10E4 > 500] <- NA

data$Salary...10E4 <- kNN(data)$Salary...10E4

boxplot(data$Salary...10E4)
```



3.3. Exportación de los datos preprocesados

```
write.csv(data, "recruitment_decision_tree_clean.csv")
```