Momento de Retroalimentacion Estadistica

Facundo Vecchi A01283666 - Ricardo Arriaga A01570553

Importacion de datos

```
library(dplyr)
```

```
##
## 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
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(modeest)

salaries <- read.csv("ds_salaries.csv")
salaries <- subset(salaries, select = -X)
```

Medidas de tendencia central y dispersión

```
for (col in names(salaries)) {
  if (class(salaries[, col]) == "integer") {
    c <- nchar(col)
    cl <- 0
    cr <- 0
    if (c %% 2 != 0) {
      cl <- c / 2
      cr <- c / 2
    } else {
      cl <- c / 2
      cr <- c / 2 + 1
    }

    cat(strrep('-', 30 - cl), col, strrep('-', 30 - cr), "\n")
```

```
    cat("Promedio: ", mean(salaries[, col]), " ",
        "Mediana: ", median(salaries[, col]), " ",
        "Moda: ", mfv(salaries[, col]), "\n")
    cat("Desviacion estandar: ", sd(salaries[, col]),
        " ", "Varianza: ", var(salaries[, col]), "\n")
    cat("Minimo: ", min(salaries[, col]), " ",
        "Maximo: ", max(salaries[, col]), "\n")
    cat("\n")
  }
}
```

```
## ------------------------ work_year ------------------------
## Promedio:  2021.405   Mediana:  2022   Moda:  2022
## Desviacion estandar:  0.692133   Varianza:  0.4790481
## Minimo:  2020   Maximo:  2022
##
## ------------------------ salary ------------------------
## Promedio:  324000.1   Mediana:  115000   Moda:  80000 100000
## Desviacion estandar:  1544357   Varianza:  2.38504e+12
## Minimo:  4000   Maximo:  30400000
##
## ---------------------- salary_in_usd ----------------------
## Promedio:  112297.9   Mediana:  101570   Moda:  100000
## Desviacion estandar:  70957.26   Varianza:  5034932663
## Minimo:  2859   Maximo:  600000
##
## ---------------------- remote_ratio ----------------------
## Promedio:  70.92257   Mediana:  100   Moda:  100
## Desviacion estandar:  40.70913   Varianza:  1657.233
## Minimo:  0   Maximo:  100
```

```
for (col in names(salaries)) {
  if (class(salaries[, col]) == "character") {
    c <- nchar(col)
    cl <- 0
    cr <- 0
    if (c %% 2 != 0) {
      cl <- c / 2
      cr <- c / 2
    } else {
      cl <- c / 2
      cr <- c / 2 + 1
    }

    cat(strrep('-', 30 - cl), col, strrep('-', 30 - cr), "\n")
    cat("Moda: ", mfv(salaries[, col]), "\n")
    cat("\n")
    print(table(salaries[, col]))
    cat("\n")
  }
}
```

```
## ---------------------- experience_level ----------------------
## Moda:  SE
##
```

```
##
##  EN  EX  MI  SE
##  88  26 213 280
##
## ---------------------- employment_type ----------------------
## Moda:  FT
##
##
##  CT  FL  FT  PT
##   5   4 588  10
##
## ------------------------- job_title -------------------------
## Moda:  Data Scientist
##
##
##            3D Computer Vision Researcher
##                                        1
##                               AI Scientist
##                                        7
##                         Analytics Engineer
##                                        4
##                    Applied Data Scientist
##                                        5
##       Applied Machine Learning Scientist
##                                        4
##                             BI Data Analyst
##                                        6
##                          Big Data Architect
##                                        1
##                           Big Data Engineer
##                                        8
##                      Business Data Analyst
##                                        5
##                        Cloud Data Engineer
##                                        2
##                  Computer Vision Engineer
##                                        6
##        Computer Vision Software Engineer
##                                        3
##                                Data Analyst
##                                       97
##                   Data Analytics Engineer
##                                        4
##                       Data Analytics Lead
##                                        1
##                    Data Analytics Manager
##                                        7
##                              Data Architect
##                                       11
##                               Data Engineer
##                                      132
##                  Data Engineering Manager
##                                        5
##                   Data Science Consultant
```

```
##                                     7
##                     Data Science Engineer
##                                     3
##                      Data Science Manager
##                                    12
##                           Data Scientist
##                                   143
##                          Data Specialist
##                                     1
##              Director of Data Engineering
##                                     2
##                  Director of Data Science
##                                     7
##                             ETL Developer
##                                     2
##                     Finance Data Analyst
##                                     1
##                   Financial Data Analyst
##                                     2
##                               Head of Data
##                                     5
##                     Head of Data Science
##                                     4
##                  Head of Machine Learning
##                                     1
##                          Lead Data Analyst
##                                     3
##                         Lead Data Engineer
##                                     6
##                       Lead Data Scientist
##                                     3
##            Lead Machine Learning Engineer
##                                     1
##                Machine Learning Developer
##                                     3
##                 Machine Learning Engineer
##                                    41
## Machine Learning Infrastructure Engineer
##                                     3
##                  Machine Learning Manager
##                                     1
##                Machine Learning Scientist
##                                     8
##                    Marketing Data Analyst
##                                     1
##                               ML Engineer
##                                     6
##                              NLP Engineer
##                                     1
##                    Principal Data Analyst
##                                     2
##                   Principal Data Engineer
##                                     3
##                  Principal Data Scientist
```

```
##                                     7
##                       Product Data Analyst
##                                     2
##                         Research Scientist
##                                    16
##                       Staff Data Scientist
##                                     1
##
## ---------------------- salary_currency ----------------------
## Moda:  USD
##
##
## AUD BRL CAD CHF CLP CNY DKK EUR GBP HUF INR JPY MXN PLN SGD TRY USD
##   2   2  18   1   1   2   2  95  44   2  27   3   2   3   2   3 398
##
## --------------------- employee_residence ---------------------
## Moda:  US
##
##
##  AE  AR  AT  AU  BE  BG  BO  BR  CA  CH  CL  CN  CO  CZ  DE  DK  DZ  EE  ES  FR
##   3   1   3   3   2   1   1   6  29   1   1   1   1   1  25   2   1   1  15  18
##  GB  GR  HK  HN  HR  HU  IE  IN  IQ  IR  IT  JE  JP  KE  LU  MD  MT  MX  MY  NG
##  44  13   1   1   1   2   1  30   1   1   4   1   7   1   1   1   1   2   1   2
##  NL  NZ  PH  PK  PL  PR  PT  RO  RS  RU  SG  SI  TN  TR  UA  US  VN
##   5   1   1   6   4   1   6   2   1   4   2   2   1   3   1 332   3
##
## ---------------------- company_location ----------------------
## Moda:  US
##
##
##  AE  AS  AT  AU  BE  BR  CA  CH  CL  CN  CO  CZ  DE  DK  DZ  EE  ES  FR  GB  GR
##   3   1   4   3   2   3  30   2   1   2   1   2  28   3   1   1  14  15  47  11
##  HN  HR  HU  IE  IL  IN  IQ  IR  IT  JP  KE  LU  MD  MT  MX  MY  NG  NL  NZ  PK
##   1   1   1   1   1  24   1   1   2   6   1   3   1   1   3   1   2   4   1   3
##  PL  PT  RO  RU  SG  SI  TR  UA  US  VN
##   4   4   1   2   1   2   3   1 355   1
##
## ------------------------ company_size ------------------------
## Moda:  M
##
##
##   L   M   S
## 198 326  83
```

Medidas de distribución y medidas de posicion

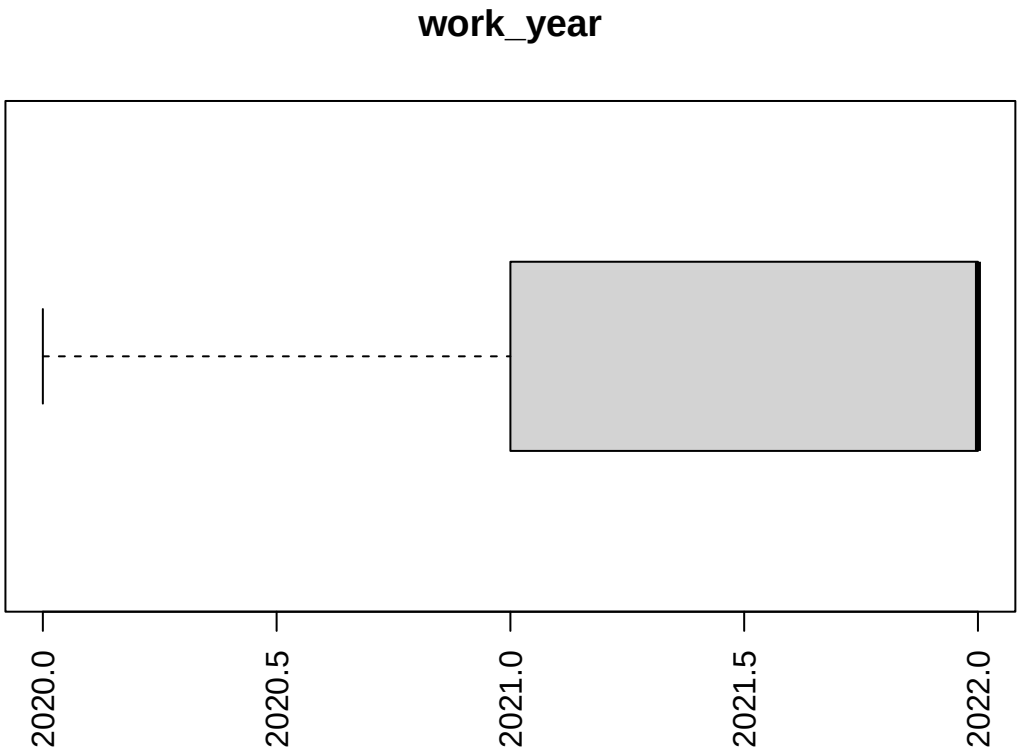
```r
for (col in names(salaries)) {
  if (class(salaries[, col]) == "integer") {
    x <- salaries[, col]
    q <- quantile(x, c(0.25, 0.75))
    ri <- q[2] - q[1]

    c <- nchar(col)
```

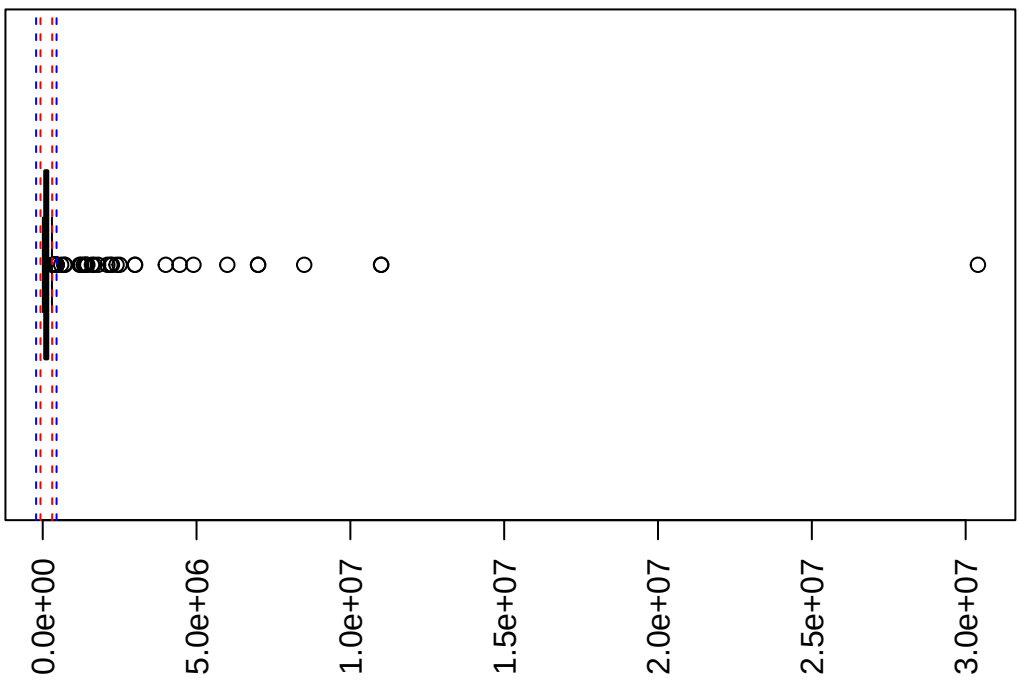
```r
    cl <- 0
    cr <- 0
    if (c %% 2 != 0) {
      cl <- c / 2
      cr <- c / 2
    } else {
      cl <- c / 2
      cr <- c / 2 + 1
    }

    cat(strrep('-', 30 - cl), col, strrep('-', 30 - cr), "\n")
    cat("Quartil 1: ", q[1], " ", "Quartil 3: ", q[2], "\n")
    boxplot(x, main = col, las = 2, xlab = "", ylab = "", horizontal = TRUE)
    abline(v = q[1] - 1.5 * ri, lty = 2, col = "red")
    abline(v = q[2] + 1.5 * ri, lty = 2, col = "red")
    abline(v = q[1] - 3 * ri, lty = 2, col = "blue")
    abline(v = q[2] + 3 * ri, lty = 2, col = "blue")
  }
}
```

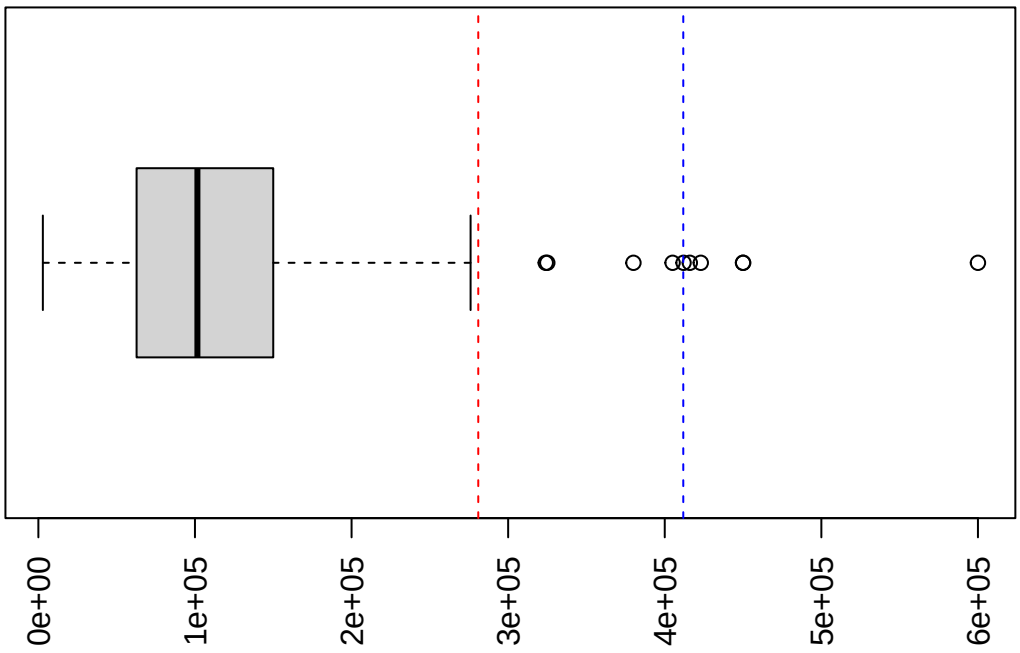
```
## ------------------------- work_year -------------------------
## Quartil 1:  2021   Quartil 3:  2022
```



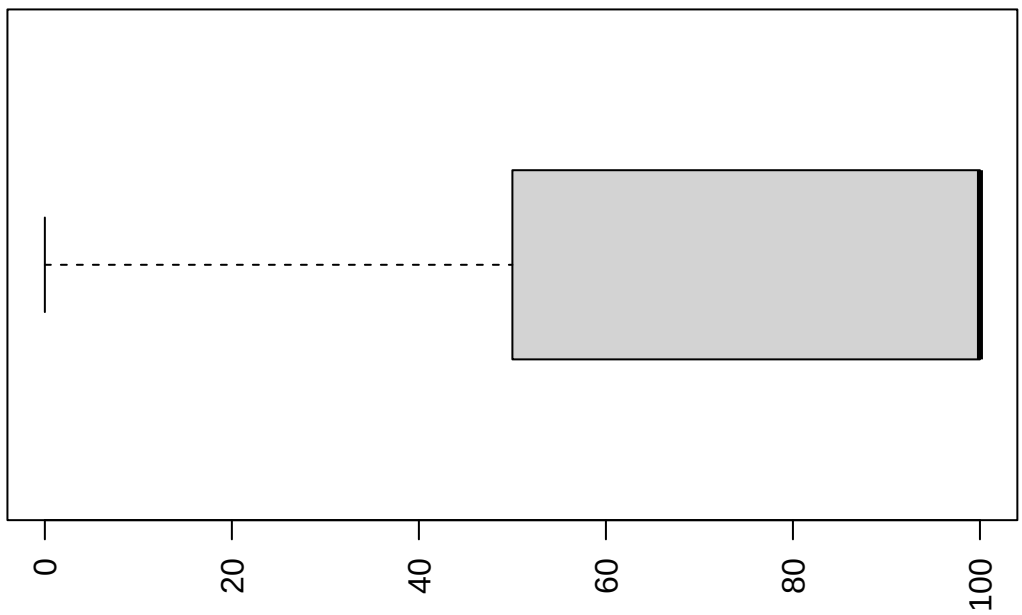

```
## -------------------------- salary --------------------------
## Quartil 1:  70000   Quartil 3:  165000
```



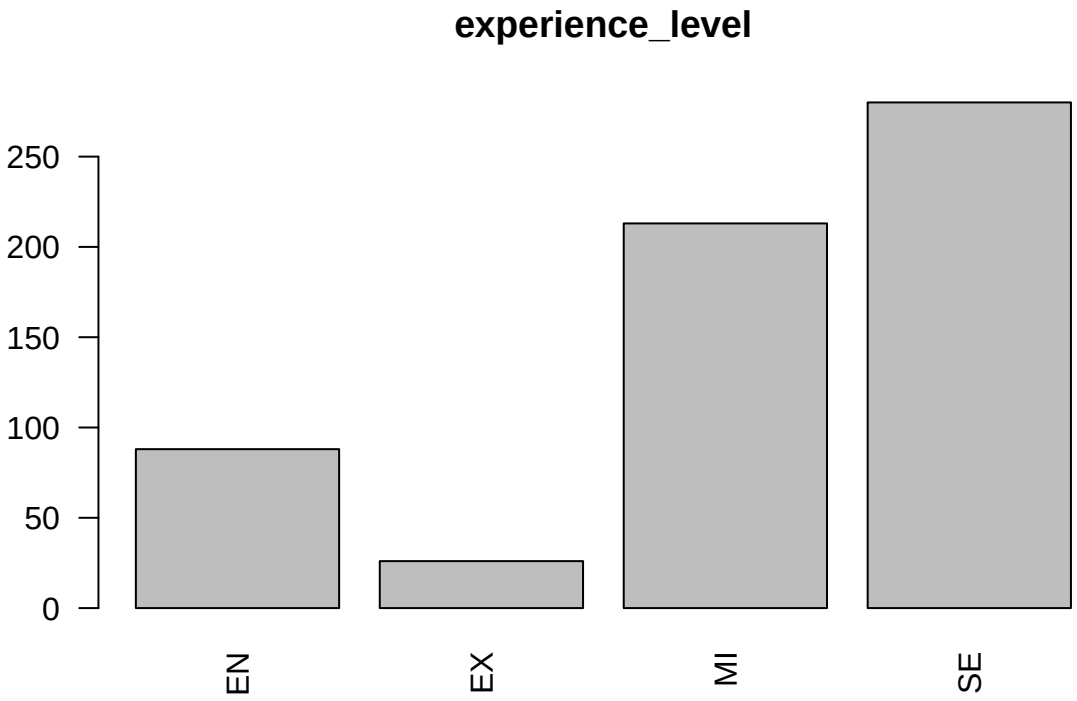

```
## ----------------------- salary_in_usd -----------------------
## Quartil 1:  62726   Quartil 3:  150000
```

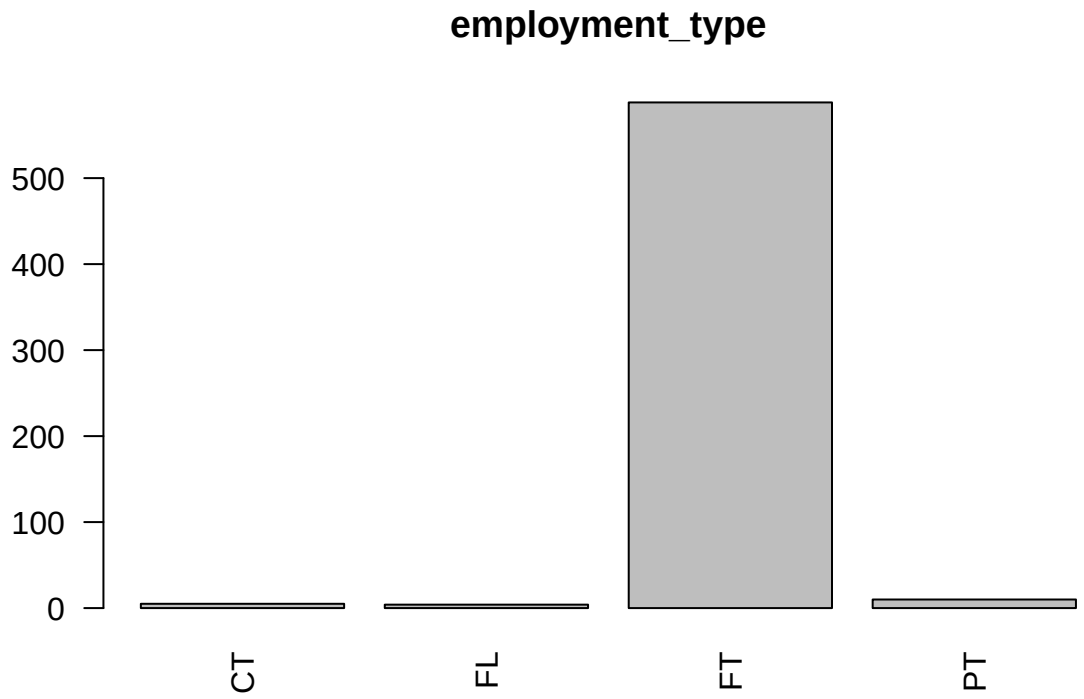


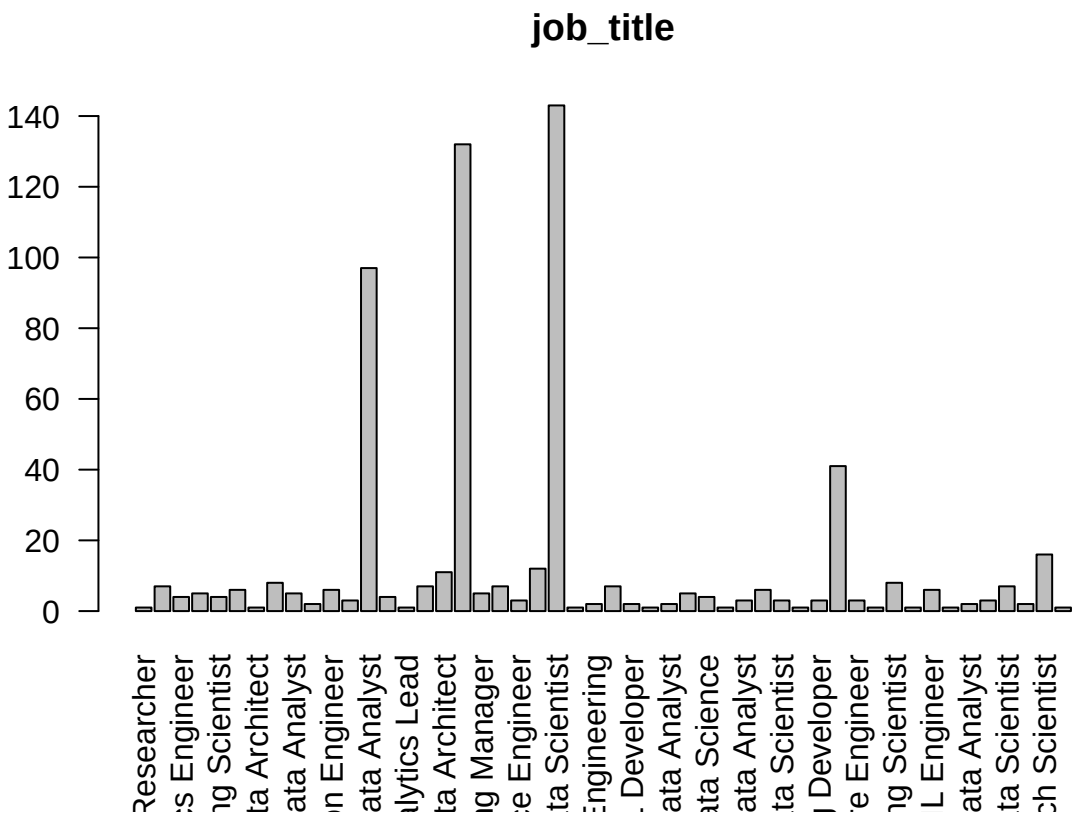

```
## ------------------------ remote_ratio ----------------------
## Quartil 1:  50   Quartil 3:  100
```

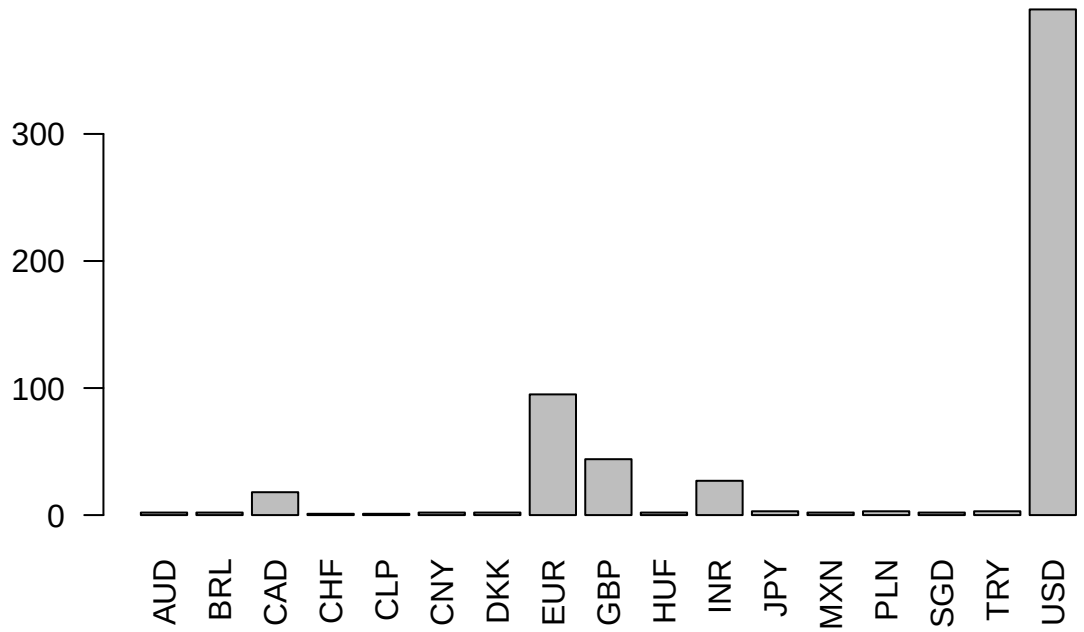


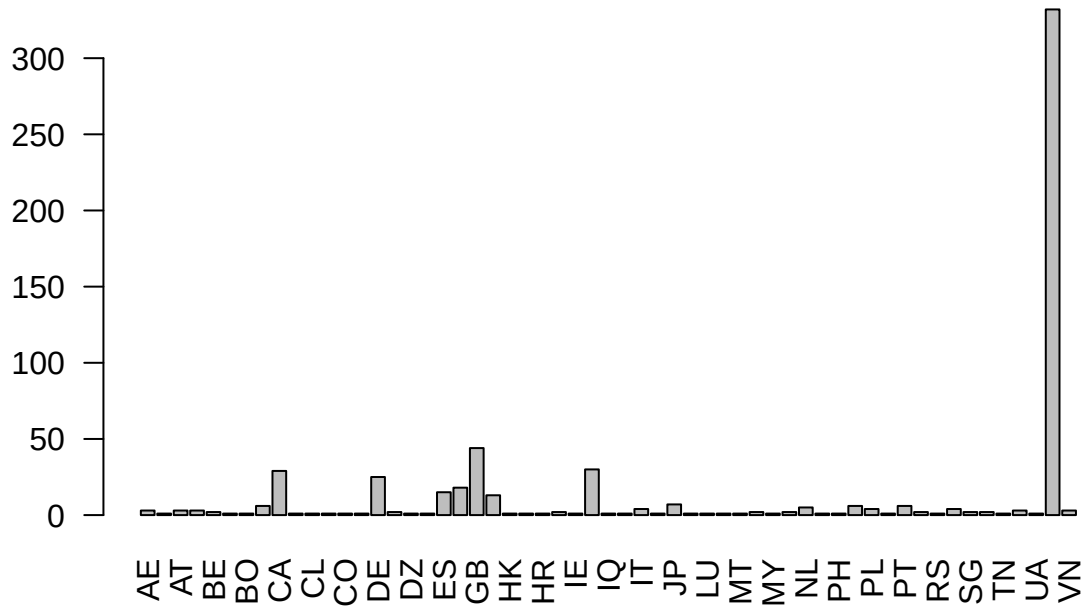

```
for (col in names(salaries)) {
  if (class(salaries[, col]) == "character") {
    x <- table(salaries[, col])
    barplot(x, main = col, las = 2, xlab = "", ylab = "")
  }
}
```

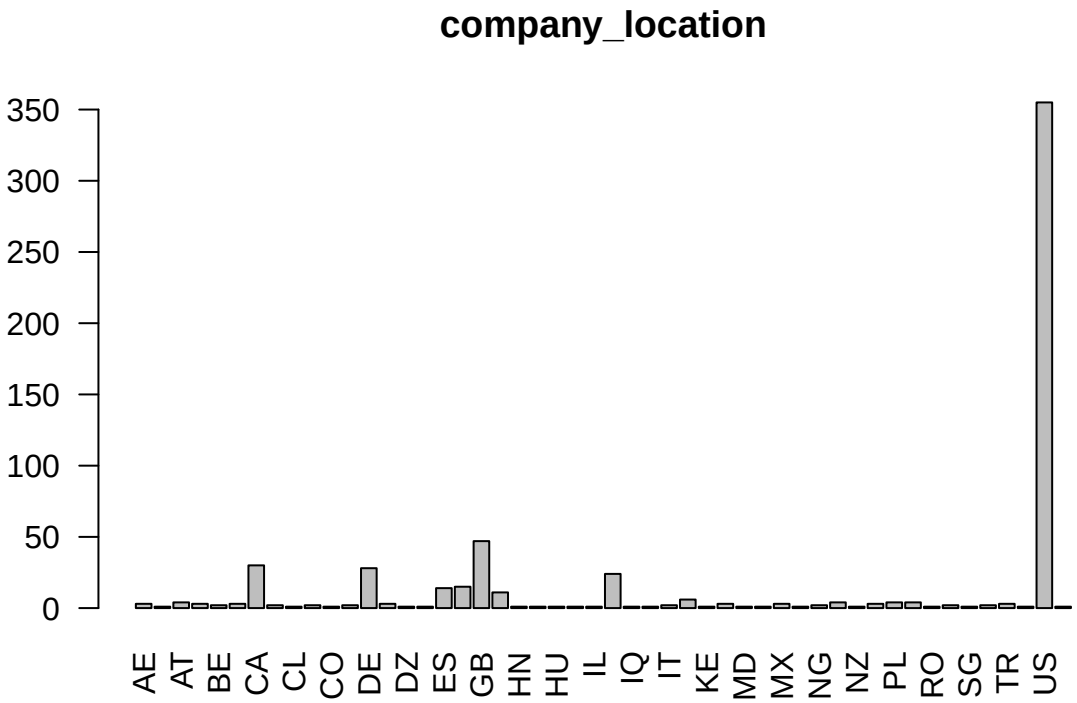

experience_level
250
200
150
100
50
0
EN
EX
MI
SE

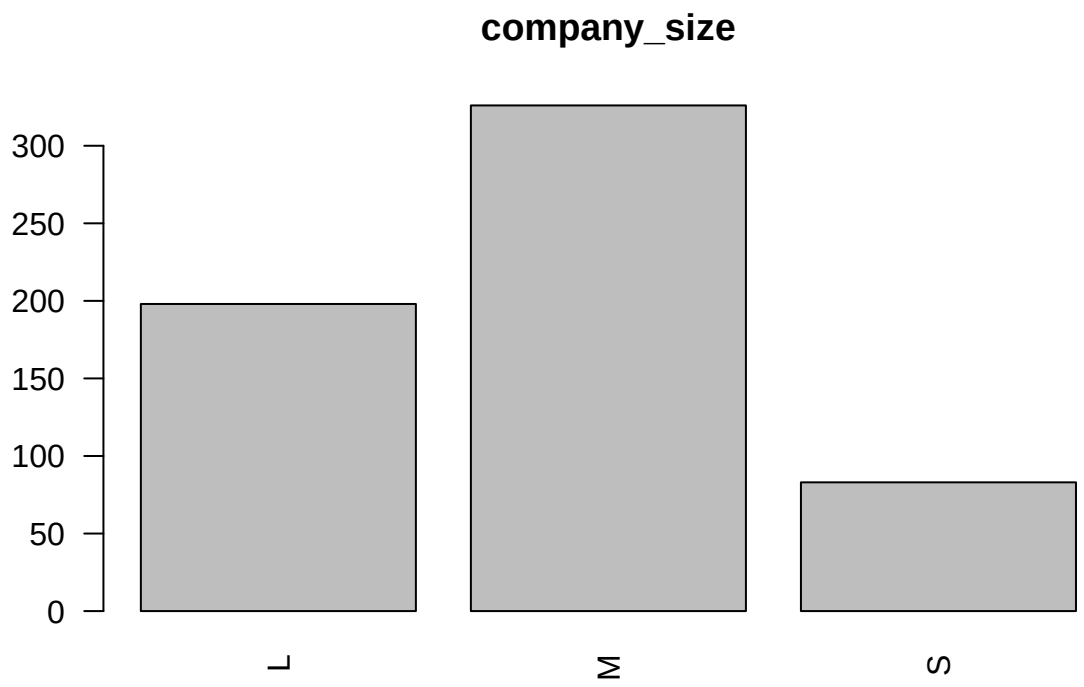

employment_type
500
400
300
200
100
0
CT
FL
FT
PT


job_title
140
120
100
80
60
40
20
0
Researcher
s Engineer
g Scientist
ta Architect
ata Analyst
n Engineer
ata Analyst
alytics Lead
ta Architect
g Manager
e Engineer
ta Scientist
Engineering
. Developer
ata Analyst
ata Science
ata Analyst
ta Scientist
Developer
e Engineer
g Scientist
L Engineer
ata Analyst
ta Scientist
ch Scientist


salary_currency
300
200
100
0
AUD
BRL
CAD
CHF
CLP
CNY
DKK
EUR
GBP
HUF
INR
JPY
MXN
PLN
SGD
TRY
USD


employee_residence
300
250
200
150
100
50
0
AE
AT
BE
BO
CA
CL
CO
DE
DZ
ES
GB
HK
HR
IE
IQ
IT
JP
LU
MT
MY
NL
PH
PL
PT
RS
SG
TN
UA
VN


company_location
350
300
250
200
150
100
50
0
AE
AT
BE
CA
CL
CO
DE
DZ
ES
GB
HN
HU
IL
IQ
IT
KE
MD
MX
NG
NZ
PL
RO
SG
TR
US



Calidad de datos

```
for (col in names(salaries)) {
  c <- nchar(col)
  cl <- 0
  cr <- 0
  if (c %% 2 != 0) {
    cl <- c / 2
    cr <- c / 2
  } else {
    cl <- c / 2
    cr <- c / 2 + 1
  }
  cat(strrep('-', 30 - cl), col, strrep('-', 30 - cr), "\n")

  cat("NAs: ", sum(is.na(salaries[, col])), "\n")
}
```

```
## ------------------------- work_year -------------------------
## NAs:  0
## ---------------------- experience_level ----------------------
## NAs:  0
## ---------------------- employment_type -----------------------
## NAs:  0
## ------------------------- job_title -------------------------
```

```
## NAs:  0
## -------------------------- salary --------------------------
## NAs:  0
## ---------------------- salary_currency ----------------------
## NAs:  0
## ------------------------ salary_in_usd ------------------------
## NAs:  0
## --------------------- employee_residence ---------------------
## NAs:  0
## ------------------------ remote_ratio ------------------------
## NAs:  0
## ---------------------- company_location ----------------------
## NAs:  0
## ------------------------ company_size ------------------------
## NAs:  0
```

Preparacion de datos

Para nuestras predicciones hemos decidido eliminar las columnas de salario y tipo de moneda, ya que buscamos mantener un estandar en cuanto a los salarios.

```
salaries_clean <- subset(salaries, select = -c(salary, salary_currency))
```

Pregunta 1

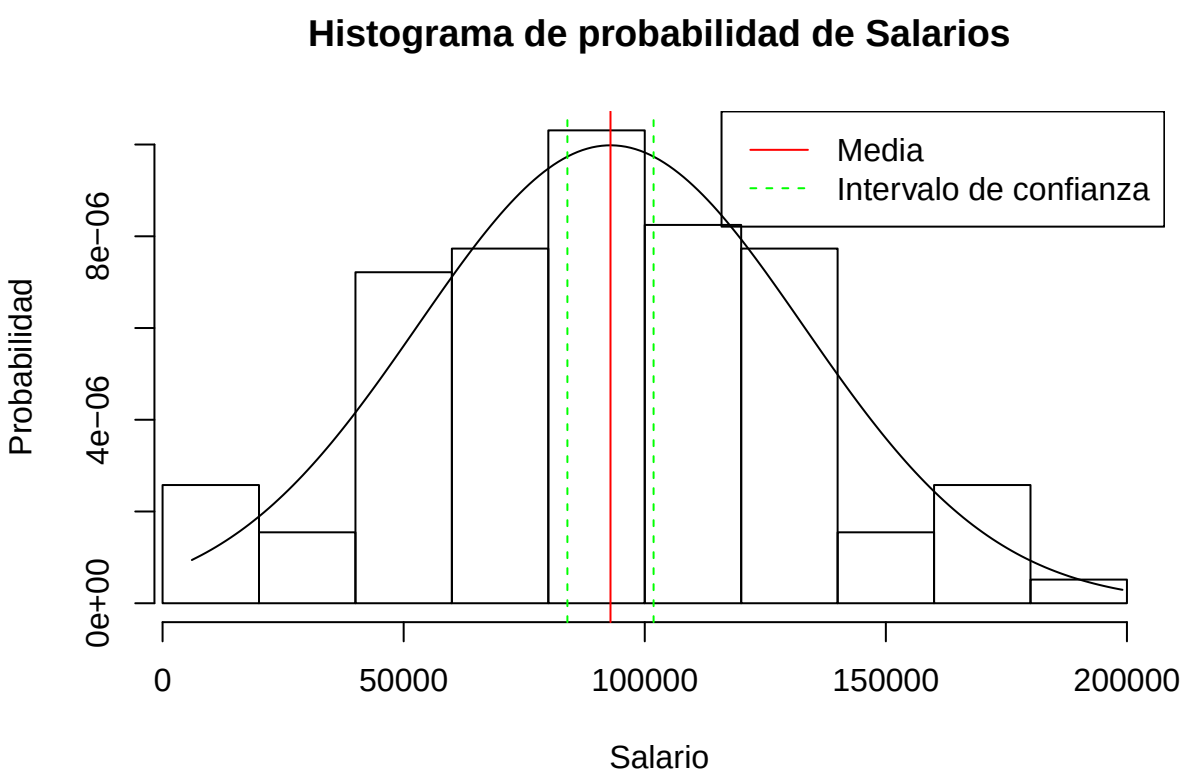
¿Cuál es el salario al que pueda aspirar un analista de datos?

```
salaries_da_x <- salaries_clean[salaries_clean$job_title == "Data Analyst", "salary_in_usd"]

salary_mean <- mean(salaries_da_x)
salary_sd <- sd(salaries_da_x)
salary_inters <- t.test(x = salaries_da_x, conf.level = 0.97)$conf.int

salaries_da_x_ <- seq(min(salaries_da_x), max(salaries_da_x), 1000)
salaries_da_y <- dnorm(salaries_da_x_, salary_mean, salary_sd)

hist(salaries_da_x, prob = TRUE, main = "Histograma de probabilidad de Salarios",
     xlab = "Salario",
     ylab = "Probabilidad",
     col = 0,
     xlim = c(min(salaries_da_x), max(salaries_da_x)))
lines(salaries_da_x_, salaries_da_y, col = "black")
abline(v = salary_mean, lty = 1, col = "red")
abline(v = salary_inters, lty = 2, col = "green")
legend("topright",
       legend = c("Media", "Intervalo de confianza"),
       col = c("red", "green"),
       lty = c(1, 2))
```




```
cat("En general, el salario al que puede aspirar un analista de datos es:",
    salary_mean, "USD\n",
    "La confianza en el salario al que puede aspirar un analista de datos es: \n",
    salary_inters[1], "USD -", salary_inters[2], "USD\n")
```

```
## En general, el salario al que puede aspirar un analista de datos es: 92893.06 USD
##  La confianza en el salario al que puede aspirar un analista de datos es:
##  83955.25 USD - 101830.9 USD
```

Pregunta 2

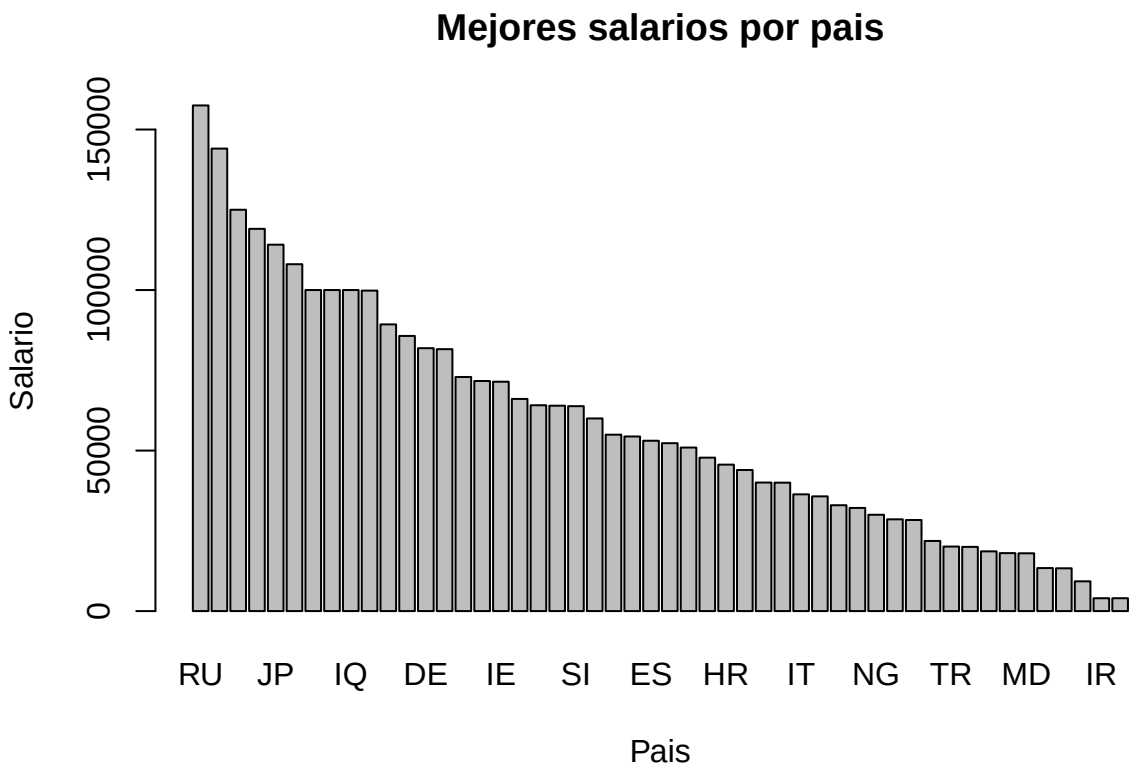
¿En qué países se ofrecen mejores salarios?

```
ISO_3166 <- fread("ISO-3166.csv", select = c("alpha-2", "name"))

salaries_country <- salaries_clean %>%
  group_by(company_location) %>%
  summarise(title_salary_mean = mean(salary_in_usd))

salaries_country <- salaries_country[order(-salaries_country$title_salary_mean),]

barplot(salaries_country$title_salary_mean,
        names.arg = salaries_country$company_location,
        main = "Mejores salarios por pais",
        xlab = "Pais",
        ylab = "Salario")
```




```
final_salaries <- data.table("Alias" = character(), "Pais" = character(), "Salario" = double())
for (i in 1:5) {
  name <- ISO_3166[ISO_3166$`alpha-2` == salaries_country$company_location[i], "name"]
  new_row <- data.table("Alias" = salaries_country$company_location[i],
                        "Pais" = name$name,
                        "Salario" = salaries_country$title_salary_mean[i])
  final_salaries <- rbind(final_salaries, new_row)
}
cat("\n")
```

```
cat("En general, los mejores salarios son en:", "\n")
```

```
## En general, los mejores salarios son en:
```

```
print(final_salaries)
```

```
##    Alias                     Pais  Salario
## 1:    RU       Russian Federation 157500.0
## 2:    US United States of America 144055.3
## 3:    NZ              New Zealand 125000.0
## 4:    IL                   Israel 119059.0
## 5:    JP                    Japan 114127.3
```

Pregunta 3

¿Que tanto afecta los salarios dependiendo del radio de trabajo remoto?

```r
titles <- unique(salaries_clean$job_title)

title_salary_remote <- salaries_clean %>%
  group_by(job_title, remote_ratio) %>%
  summarise(title_salary_mean = mean(salary_in_usd), .groups = 'drop')

title_salary_remote_0 <- title_salary_remote[title_salary_remote$remote_ratio == 0,]
title_salary_remote_50 <- title_salary_remote[title_salary_remote$remote_ratio == 50,]
title_salary_remote_100 <- title_salary_remote[title_salary_remote$remote_ratio == 100,]

acum_0_50 <- c()
acum_50_100 <- c()
acum_0_100 <- c()

for (title in titles) {
  title_salary_remote_0_ <- title_salary_remote_0[title_salary_remote_0$job_title == title,]
  title_salary_remote_50_ <- title_salary_remote_50[title_salary_remote_50$job_title == title,]
  title_salary_remote_100_ <- title_salary_remote_100[title_salary_remote_100$job_title == title,]


  if (nrow(title_salary_remote_100_) != 0 & nrow(title_salary_remote_0_) != 0) {
    acum_0_100 <- append(acum_0_100,
                         title_salary_remote_100_$title_salary_mean[1] /
                                title_salary_remote_0_$title_salary_mean[1] - 1)
  }
  if (nrow(title_salary_remote_50_) != 0 & nrow(title_salary_remote_0_) != 0) {
    acum_0_50 <- append(acum_0_50,
                        title_salary_remote_50_$title_salary_mean[1] /
                               title_salary_remote_0_$title_salary_mean[1] - 1)
  }
  if (nrow(title_salary_remote_100_) != 0 & nrow(title_salary_remote_50_) != 0) {
    acum_50_100 <- append(acum_50_100,
                          title_salary_remote_100_$title_salary_mean[1] /
                                 title_salary_remote_50_$title_salary_mean[1] - 1)
  }
}

cat("\n")
```

```r
cat("En general, el afecto de los salarios dependiendo del radio de trabajo remoto es:",
    "\n",
    "0% - 50% remoto:", mean(acum_0_50) * 100, "%\n",
    "50% - 100% remoto:", mean(acum_50_100) * 100, "%\n",
    "0% - 100% remoto:", mean(acum_0_100) * 100, "%\n")
```

```
## En general, el afecto de los salarios dependiendo del radio de trabajo remoto es:
##  0% - 50% remoto: -18.20465 %
##  50% - 100% remoto: 82.82617 %
##  0% - 100% remoto: 36.70695 %
```