RETO ESTADISTICA SALARIOS

María del Carmen Vargas Villarreal A00828570

2022-08-22

1. EXPLORACIÓN DE LA BASE DE DATOS

A) Accede a la base de datos de Data Science Jobs Salaries

```
datos_salarios <- read.csv("ds_salaries.csv", header=TRUE)
#datos_salarios
```

B) Explora las variables y familiarizate con su significado

1) Identifica la cantidad de datos y variables presentes.

```
# Dimesión del data frame (renglones = cantidad de datos, columnas = cantidad de variables)
dim(datos_salarios)
```

```
## [1] 607  12
```

2) Clasifica las variables de acuerdo a su tipo y escala de medición.

```
# Variables categóricas.
# Debajo de cada variable se puede observar el tipo de variable, ya sea tipo <int> o tipo <char>

categoricas = datos_salarios[, c( 'experience_level', 'employment_type', 'job_title', 'salary_currency'

head(categoricas)
```

```
##   experience_level employment_type                 job_title salary_currency
## 1               MI              FT              Data Scientist             EUR
## 2               SE              FT Machine Learning Scientist             USD
## 3               SE              FT           Big Data Engineer             GBP
## 4               MI              FT       Product Data Analyst             USD
## 5               SE              FT  Machine Learning Engineer             USD
## 6               EN              FT               Data Analyst             USD
##   employee_residence company_location company_size
## 1                 DE               DE            L
## 2                 JP               JP            S
## 3                 GB               GB            M
## 4                 HN               HN            S
## 5                 US               US            L
## 6                 US               US            L
```

```r
# Variables números
# Debajo de cada variable se puede observar el tipo de variable, de tipo <int>

numericas = datos_salarios[, c('salary', 'salary_in_usd', 'work_year', 'remote_ratio')]

head(numericas)
```

```
##   salary salary_in_usd work_year remote_ratio
## 1  70000         79833      2020            0
## 2 260000        260000      2020            0
## 3  85000        109024      2020           50
## 4  20000         20000      2020            0
## 5 150000        150000      2020           50
## 6  72000         72000      2020          100
```

C) Exploración de la base de datos

1) Cálculo de medidas estadísticas:

Variables cuantitativas

Medidas de tendencia central: promedio, media, mediana y moda de los datos.

```r
# Medidas de tendencia central: Min, Max, Promedio, media, mediana, 1st Quartile, 3rd Quartile
summary(numericas)
```

```
##      salary         salary_in_usd      work_year     remote_ratio
##  Min.   :    4000   Min.   :  2859   Min.   :2020   Min.   :  0.00
##  1st Qu.:   70000   1st Qu.: 62726   1st Qu.:2021   1st Qu.: 50.00
##  Median :  115000   Median :101570   Median :2022   Median :100.00
##  Mean   :  324000   Mean   :112298   Mean   :2021   Mean   : 70.92
##  3rd Qu.:  165000   3rd Qu.:150000   3rd Qu.:2022   3rd Qu.:100.00
##  Max.   :30400000   Max.   :600000   Max.   :2022   Max.   :100.00
```

```r
# Medidas de dispersión solo para Salary

# Summary de variable salario
summary(datos_salarios$salary)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     4000    70000   115000   324000   165000 30400000
```

```r
# Medidas de dispersión solo para Salary in USD

summary(datos_salarios$salary_in_usd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2859   62726  101570  112298  150000  600000
```

```r
# Medidas de dispersión solo para Work Year

summary(datos_salarios$work_year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2020    2021    2022    2021    2022    2022
```

```
# Medidas de dispersión solo para Remote Ratio

summary(datos_salarios$remote_ratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   50.00  100.00   70.92  100.00  100.00
```

Medidas de dispersión: rango: máximo - mínimo, varianza, desviación estándar.

```
# Medidas de dispersión: Varianza y desviación estándar para Salary

var(datos_salarios$salary)
```

```
## [1] 2.38504e+12
```

```
sd(datos_salarios$salary)
```

```
## [1] 1544357
```

```
# Medidas de dispersión: Varianza y desviación estándar para Salary in USD

var(datos_salarios$salary_in_usd)
```

```
## [1] 5034932663
```

```
sd(datos_salarios$salary_in_usd)
```

```
## [1] 70957.26
```

```
# Medidas de dispersión: Varianza y desviación estándar para Work Year

var(datos_salarios$work_year)
```

```
## [1] 0.4790481
```

```
sd(datos_salarios$work_year)
```

```
## [1] 0.692133
```

```
# Medidas de dispersión: Varianza y desviación estándar para Remote Ratio

var(datos_salarios$remote_ratio)
```

```
## [1] 1657.233
```

```
sd(datos_salarios$remote_ratio)
```

```
## [1] 40.70913
```

Variables cualitativas

Tabla de distribución de frecuencia

Moda

```
# options(scipen = 500)
```

Experience Level

```r
# Para ver la cantidad de datos por cada categoria de Experience Level (se observa en la consola)
print("Experience Level")
```

```
## [1] "Experience Level"
```

```r
table(datos_salarios$experience_level)
```

```
##
##  EN  EX  MI  SE
##  88  26 213 280
```

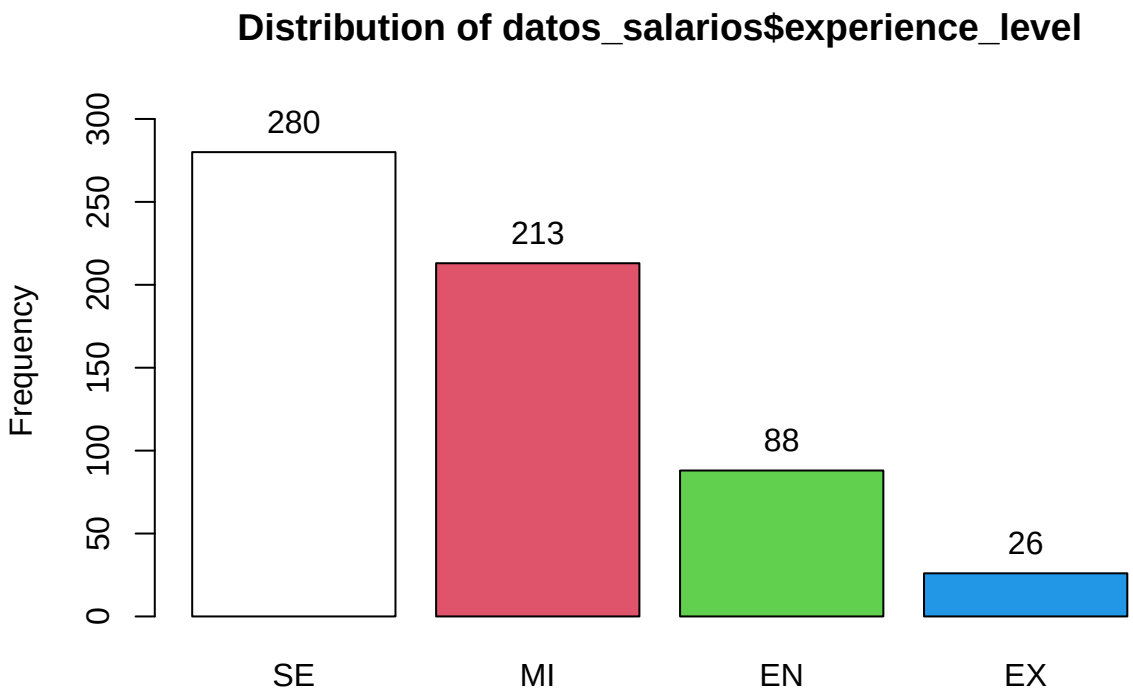
```r
# Paquete que te da atuomáticamente la tabla de distribución de frecuencias (incluyendo la cantidad en
install.packages('epiDisplay')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```r
library(epiDisplay)
```

```
## Loading required package: foreign
## Loading required package: survival
## Loading required package: MASS
## Loading required package: nnet
```

```r
tab1(datos_salarios$experience_level, sort.group = "decreasing", cum.percent = TRUE)
```




```
## datos_salarios$experience_level :
##         Frequency Percent Cum. percent
## SE            280    46.1         46.1
## MI            213    35.1         81.2
## EN             88    14.5         95.7
## EX             26     4.3        100.0
##   Total       607   100.0        100.0
```

Employment Type

```
# Para ver la cantidad de datos por cada categoria de Employment type (se observa en la consola)
print("Employment type")
```

```
## [1] "Employment type"
```

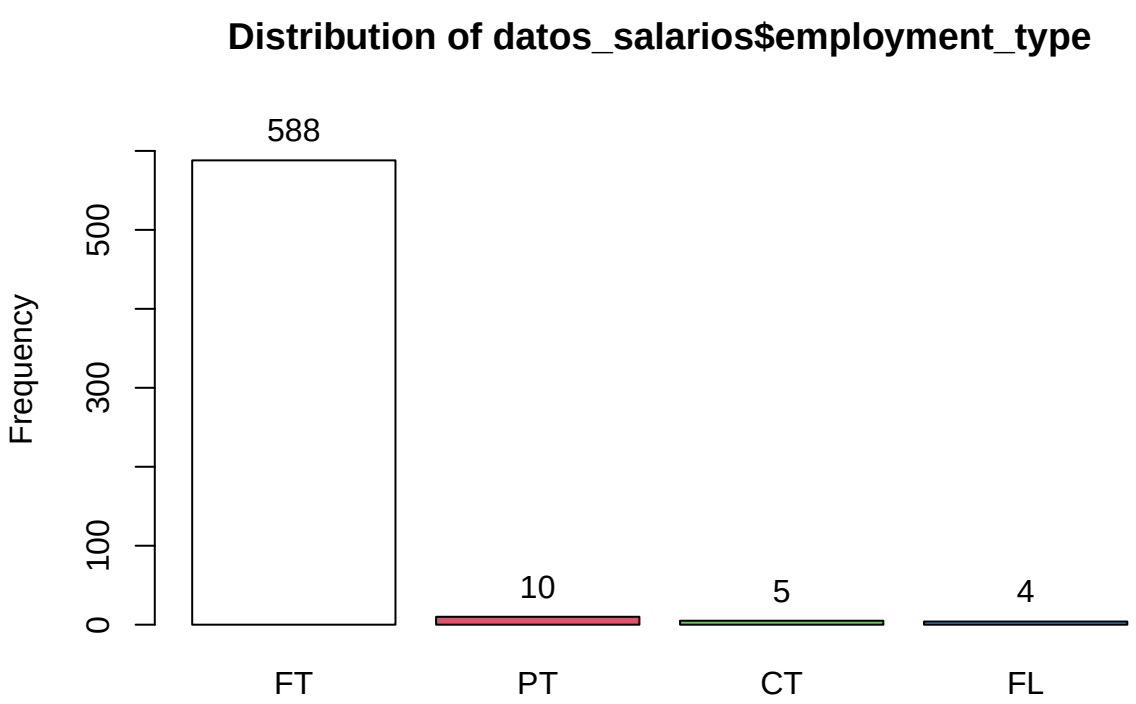
```
table(datos_salarios$employment_type)
```

```
##
##  CT  FL  FT  PT
##   5   4 588  10
```

```
# Paquete que te da atuomáticamente la distribución de frecuencias (incluyendo la cantidad en cada barr
install.packages('epiDisplay')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(epiDisplay)
tab1(datos_salarios$employment_type, sort.group = "decreasing", cum.percent = TRUE)
```




```
## datos_salarios$employment_type :
##         Frequency Percent Cum. percent
## FT            588    96.9         96.9
## PT             10     1.6         98.5
## CT              5     0.8         99.3
## FL              4     0.7        100.0
##   Total       607   100.0        100.0
```

Job Title

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:epiDisplay':
##
##     alpha
```

```
# Para ver la cantidad de datos por cada categoria de Job title (se observa en la consola)
print("Job title")
```

```
## [1] "Job title"
```

```
table(datos_salarios$job_title)
```

```
##
##            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
```

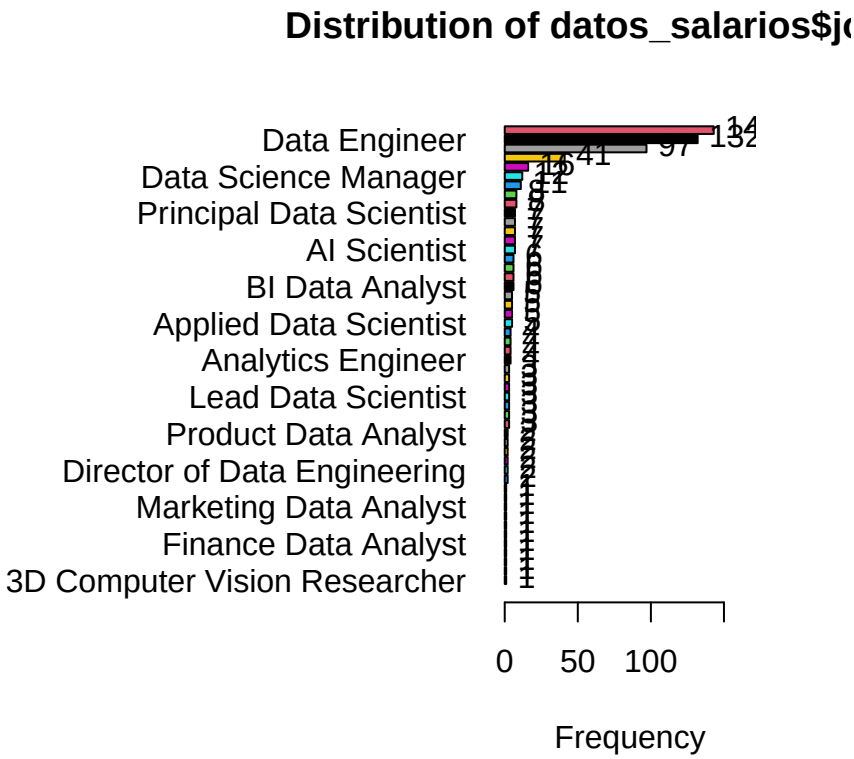
```
##                                     12
##                           Data Scientist
##                                    143
##                          Data Specialist
##                                      1
##              Director of Data Engineering
##                                      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
```

```
# Paquete que te da atuomáticamente la tabla de distribución de frecuencias (incluyendo la cantidad en
install.packages('epiDisplay')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(epiDisplay)
tab1(datos_salarios$job_title, sort.group = "decreasing", cum.percent = TRUE)
```




```
## datos_salarios$job_title :
##                                      Frequency Percent Cum. percent
## Data Scientist                             143    23.6         23.6
## Data Engineer                              132    21.7         45.3
## Data Analyst                                97    16.0         61.3
## Machine Learning Engineer                   41     6.8         68.0
## Research Scientist                          16     2.6         70.7
## Data Science Manager                        12     2.0         72.7
## Data Architect                              11     1.8         74.5
## Machine Learning Scientist                   8     1.3         75.8
## Big Data Engineer                            8     1.3         77.1
## Principal Data Scientist                     7     1.2         78.3
## Director of Data Science                     7     1.2         79.4
## Data Science Consultant                      7     1.2         80.6
## Data Analytics Manager                       7     1.2         81.7
## AI Scientist                                 7     1.2         82.9
## ML Engineer                                  6     1.0         83.9
## Lead Data Engineer                           6     1.0         84.8
```

```
## Computer Vision Engineer                          6     1.0         85.8
## BI Data Analyst                                   6     1.0         86.8
## Head of Data                                      5     0.8         87.6
## Data Engineering Manager                          5     0.8         88.5
## Business Data Analyst                             5     0.8         89.3
## Applied Data Scientist                            5     0.8         90.1
## Head of Data Science                              4     0.7         90.8
## Data Analytics Engineer                           4     0.7         91.4
## Applied Machine Learning Scientist                4     0.7         92.1
## Analytics Engineer                                4     0.7         92.8
## Principal Data Engineer                           3     0.5         93.2
## Machine Learning Infrastructure Engineer          3     0.5         93.7
## Machine Learning Developer                        3     0.5         94.2
## Lead Data Scientist                               3     0.5         94.7
## Lead Data Analyst                                 3     0.5         95.2
## Data Science Engineer                             3     0.5         95.7
## Computer Vision Software Engineer                 3     0.5         96.2
## Product Data Analyst                              2     0.3         96.5
## Principal Data Analyst                            2     0.3         96.9
## Financial Data Analyst                            2     0.3         97.2
## ETL Developer                                     2     0.3         97.5
## Director of Data Engineering                      2     0.3         97.9
## Cloud Data Engineer                               2     0.3         98.2
## Staff Data Scientist                              1     0.2         98.4
## NLP Engineer                                      1     0.2         98.5
## Marketing Data Analyst                            1     0.2         98.7
## Machine Learning Manager                          1     0.2         98.8
## Lead Machine Learning Engineer                    1     0.2         99.0
## Head of Machine Learning                          1     0.2         99.2
## Finance Data Analyst                              1     0.2         99.3
## Data Specialist                                   1     0.2         99.5
## Data Analytics Lead                               1     0.2         99.7
## Big Data Architect                                1     0.2         99.8
## 3D Computer Vision Researcher                     1     0.2        100.0
##   Total                                         607   100.0        100.0
```

Salary Currency

```
# Para ver la cantidad de datos por cada categoria de Salary Currency (se observa en la consola)
print("Salary Currency")
```

```
## [1] "Salary Currency"
```

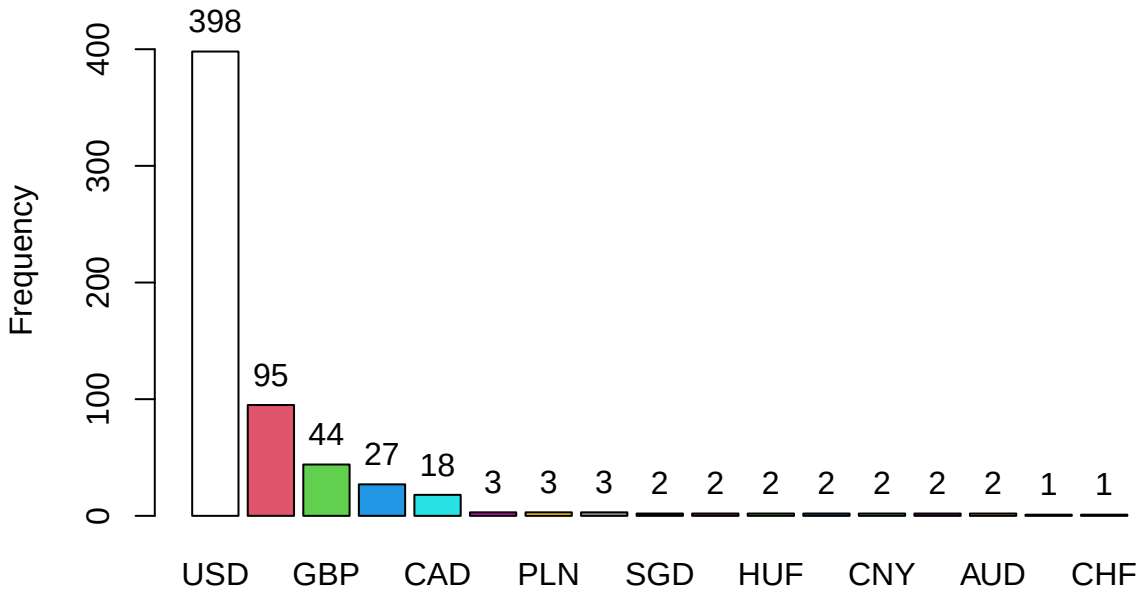
```
table(datos_salarios$salary_currency)
```

```
##
## 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
```

```
# Paquete que te da atuomáticamente la distribución de frecuencias (incluyendo la cantidad en cada barr
install.packages('epiDisplay')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(epiDisplay)
tab1(datos_salarios$salary_currency, sort.group = "decreasing", cum.percent = TRUE, width=0.9)
```




```
## datos_salarios$salary_currency :
##         Frequency Percent Cum. percent
## USD           398    65.6         65.6
## EUR            95    15.7         81.2
## GBP            44     7.2         88.5
## INR            27     4.4         92.9
## CAD            18     3.0         95.9
## TRY             3     0.5         96.4
## PLN             3     0.5         96.9
## JPY             3     0.5         97.4
## SGD             2     0.3         97.7
## MXN             2     0.3         98.0
## HUF             2     0.3         98.4
## DKK             2     0.3         98.7
## CNY             2     0.3         99.0
## BRL             2     0.3         99.3
## AUD             2     0.3         99.7
## CLP             1     0.2         99.8
## CHF             1     0.2        100.0
##   Total       607   100.0        100.0
```

Employee Residence

```
# Para ver la cantidad de datos por cada categoria de Employee Residence (se observa en la consola)
print("Employee residence")
```

```
## [1] "Employee residence"
```

```
table(datos_salarios$employee_residence)
```

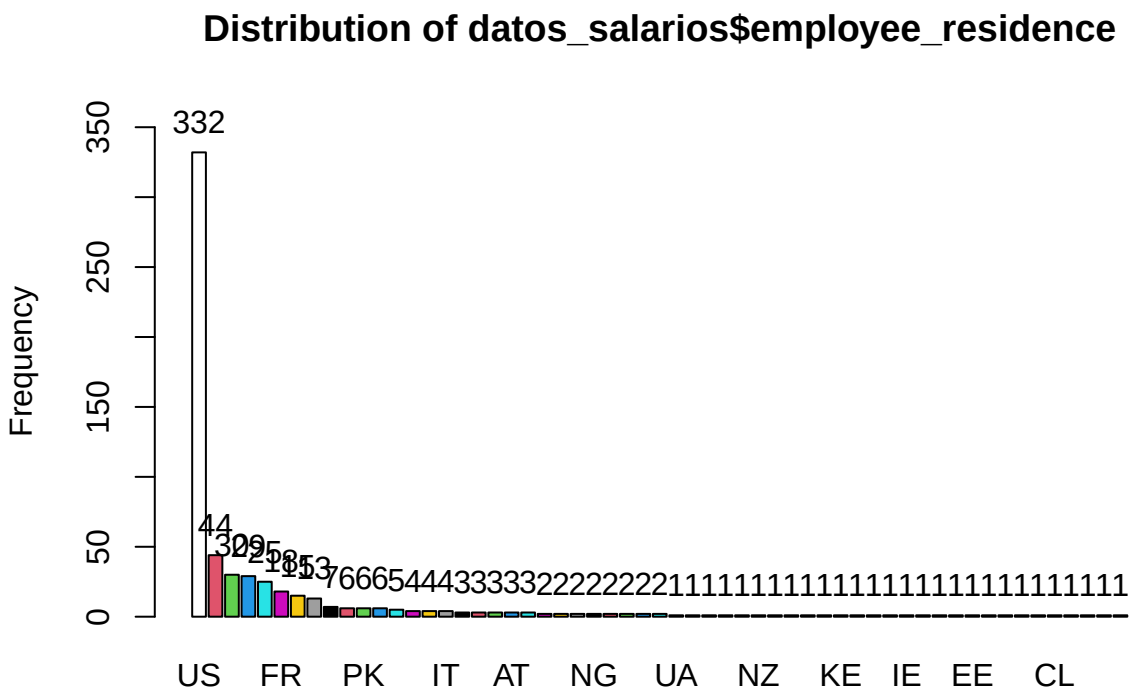
```
##
##  AE  AR  AT  AU  BE  BG  BO  BR  CA  CH  CL  CN  CO  CZ  DE  DK  DZ  EE  ES  FR
##   3   1   3   3   2   1   1   6  29   1   1   1   1   1  25   2   1   1  15  18
##  GB  GR  HK  HN  HR  HU  IE  IN  IQ  IR  IT  JE  JP  KE  LU  MD  MT  MX  MY  NG
```

```
##  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
```

```
# Paquete que te da atuomáticamente la distribución de frecuencias (incluyendo la cantidad en cada barr
install.packages('epiDisplay')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(epiDisplay)
tab1(datos_salarios$employee_residence, sort.group = "decreasing", cum.percent = TRUE)
```




```
## datos_salarios$employee_residence :
##         Frequency Percent Cum. percent
## US            332    54.7        54.7
## GB             44     7.2        61.9
## IN             30     4.9        66.9
## CA             29     4.8        71.7
## DE             25     4.1        75.8
## FR             18     3.0        78.7
## ES             15     2.5        81.2
## GR             13     2.1        83.4
## JP              7     1.2        84.5
## PT              6     1.0        85.5
## PK              6     1.0        86.5
## BR              6     1.0        87.5
## NL              5     0.8        88.3
## RU              4     0.7        89.0
## PL              4     0.7        89.6
## IT              4     0.7        90.3
## VN              3     0.5        90.8
## TR              3     0.5        91.3
## AU              3     0.5        91.8
```

```
## AT              3     0.5         92.3
## AE              3     0.5         92.8
## SI              2     0.3         93.1
## SG              2     0.3         93.4
## RO              2     0.3         93.7
## NG              2     0.3         94.1
## MX              2     0.3         94.4
## HU              2     0.3         94.7
## DK              2     0.3         95.1
## BE              2     0.3         95.4
## UA              1     0.2         95.6
## TN              1     0.2         95.7
## RS              1     0.2         95.9
## PR              1     0.2         96.0
## PH              1     0.2         96.2
## NZ              1     0.2         96.4
## MY              1     0.2         96.5
## MT              1     0.2         96.7
## MD              1     0.2         96.9
## LU              1     0.2         97.0
## KE              1     0.2         97.2
## JE              1     0.2         97.4
## IR              1     0.2         97.5
## IQ              1     0.2         97.7
## IE              1     0.2         97.9
## HR              1     0.2         98.0
## HN              1     0.2         98.2
## HK              1     0.2         98.4
## EE              1     0.2         98.5
## DZ              1     0.2         98.7
## CZ              1     0.2         98.8
## CO              1     0.2         99.0
## CN              1     0.2         99.2
## CL              1     0.2         99.3
## CH              1     0.2         99.5
## BO              1     0.2         99.7
## BG              1     0.2         99.8
## AR              1     0.2        100.0
##   Total       607   100.0        100.0
```

Company Location

```
# Para ver la cantidad de datos por cada categoria de Company Location (se observa en la consola)
print("Company Location")
```

```
## [1] "Company Location"
```

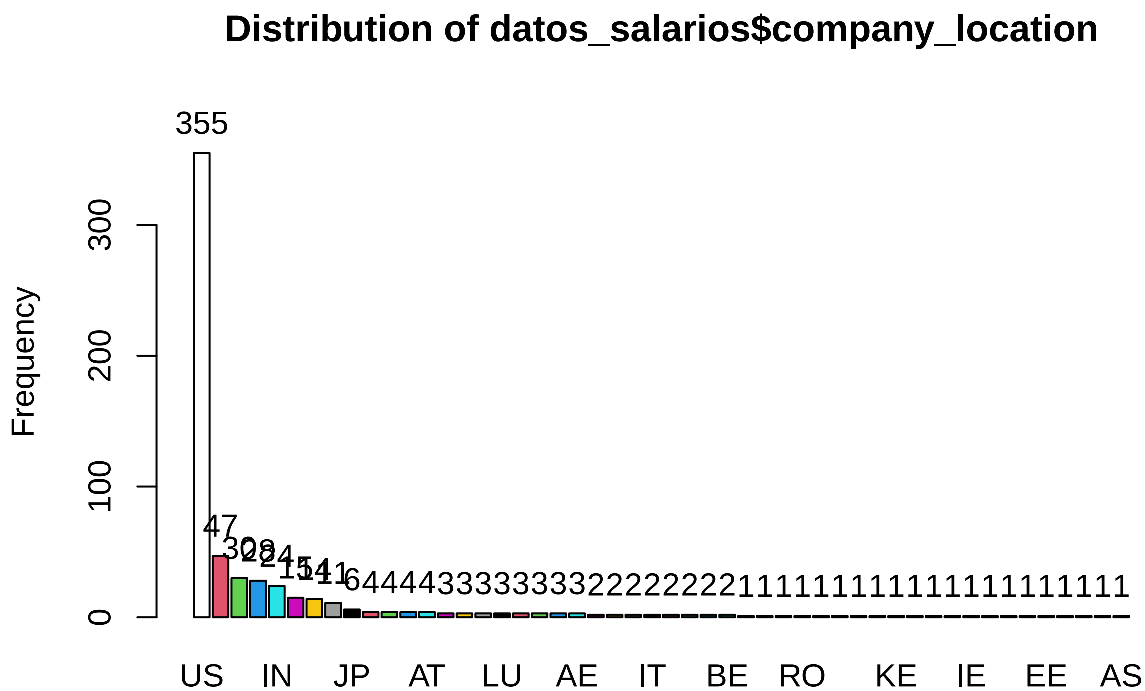
```
table(datos_salarios$company_location)
```

```
##
##  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
```

```r
# Paquete que te da atuomáticamente la distribución de frecuencias (incluyendo la cantidad en cada barr
install.packages('epiDisplay')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```r
library(epiDisplay)
tab1(datos_salarios$company_location, sort.group = "decreasing", cum.percent = TRUE)
```




```
## datos_salarios$company_location :
##         Frequency Percent Cum. percent
## US            355    58.5         58.5
## GB             47     7.7         66.2
## CA             30     4.9         71.2
## DE             28     4.6         75.8
## IN             24     4.0         79.7
## FR             15     2.5         82.2
## ES             14     2.3         84.5
## GR             11     1.8         86.3
## JP              6     1.0         87.3
## PT              4     0.7         88.0
## PL              4     0.7         88.6
## NL              4     0.7         89.3
## AT              4     0.7         90.0
## TR              3     0.5         90.4
## PK              3     0.5         90.9
## MX              3     0.5         91.4
## LU              3     0.5         91.9
## DK              3     0.5         92.4
## BR              3     0.5         92.9
## AU              3     0.5         93.4
## AE              3     0.5         93.9
## SI              2     0.3         94.2
```

```
## RU              2     0.3         94.6
## NG              2     0.3         94.9
## IT              2     0.3         95.2
## CZ              2     0.3         95.6
## CN              2     0.3         95.9
## CH              2     0.3         96.2
## BE              2     0.3         96.5
## VN              1     0.2         96.7
## UA              1     0.2         96.9
## SG              1     0.2         97.0
## RO              1     0.2         97.2
## NZ              1     0.2         97.4
## MY              1     0.2         97.5
## MT              1     0.2         97.7
## MD              1     0.2         97.9
## KE              1     0.2         98.0
## IR              1     0.2         98.2
## IQ              1     0.2         98.4
## IL              1     0.2         98.5
## IE              1     0.2         98.7
## HU              1     0.2         98.8
## HR              1     0.2         99.0
## HN              1     0.2         99.2
## EE              1     0.2         99.3
## DZ              1     0.2         99.5
## CO              1     0.2         99.7
## CL              1     0.2         99.8
## AS              1     0.2        100.0
##   Total       607   100.0        100.0
```

Company Size

```
# Para ver la cantidad de datos por cada categoria de Company Size
print("Company Size")
```

```
## [1] "Company Size"
```

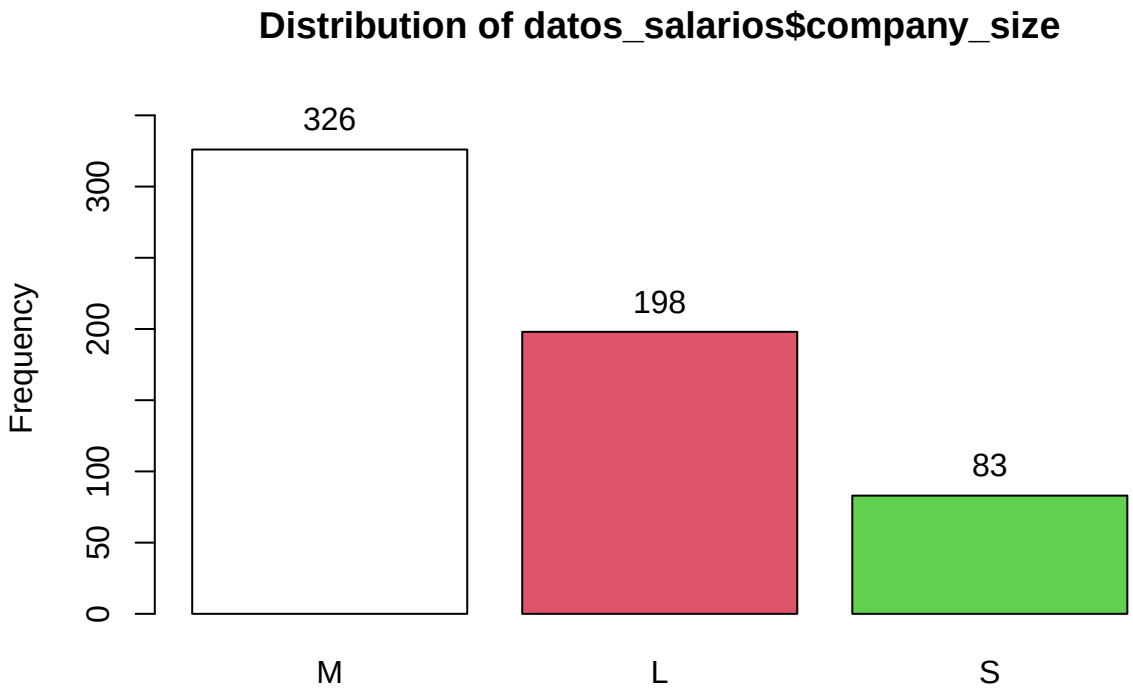
```
table(datos_salarios$company_size)
```

```
##
##   L   M   S
## 198 326  83
```

```
# Paquete que te da atuomáticamente la distribución de frecuencias (incluyendo la cantidad en cada barr
install.packages('epiDisplay')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(epiDisplay)
tab1(datos_salarios$company_size, sort.group = "decreasing", cum.percent = TRUE)
```




```
## datos_salarios$company_size :
##         Frequency Percent Cum. percent
## M             326    53.7         53.7
## L             198    32.6         86.3
## S              83    13.7        100.0
##   Total       607   100.0        100.0
```

2) Explora los datos usando herramientas de visualización:

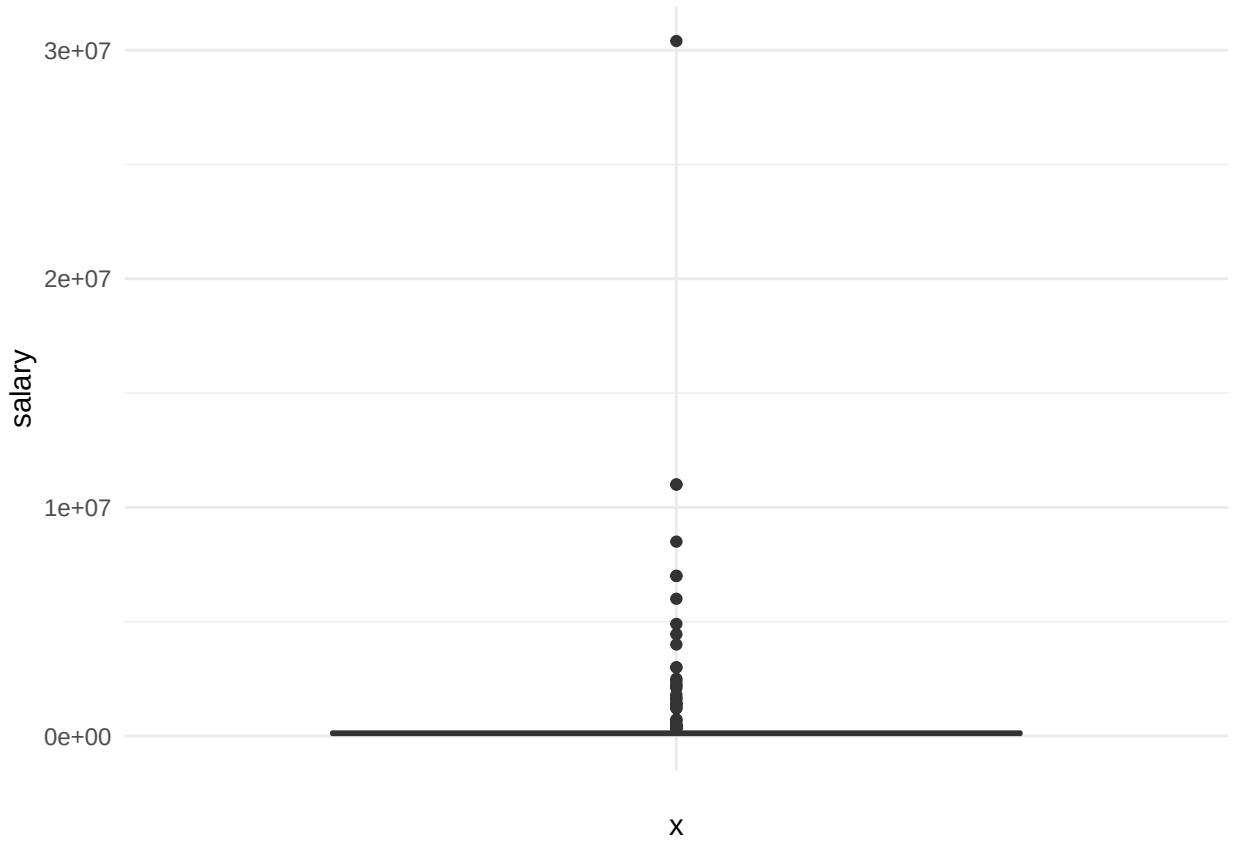
Variables cuantitativas:

Medidas de posición: cuartiles, outlier (valores atípicos), boxplots

```
# Boxplot para Salary

summary(datos_salarios$salary)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     4000    70000   115000   324000   165000 30400000
```

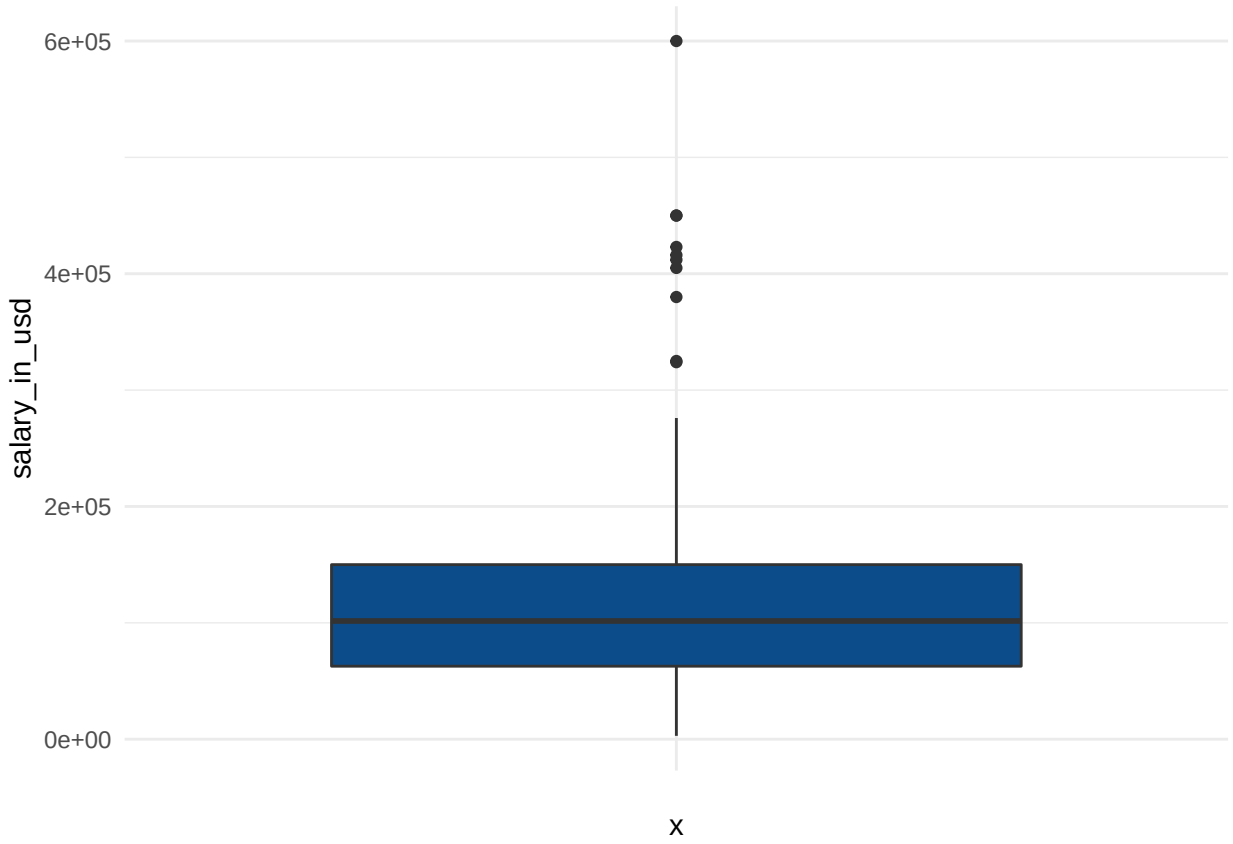
```
ggplot(datos_salarios) +
  aes(x = "", y = salary) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```




```
# Boxplot para Salary in USD

summary(datos_salarios$salary_in_usd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2859   62726  101570  112298  150000  600000
```

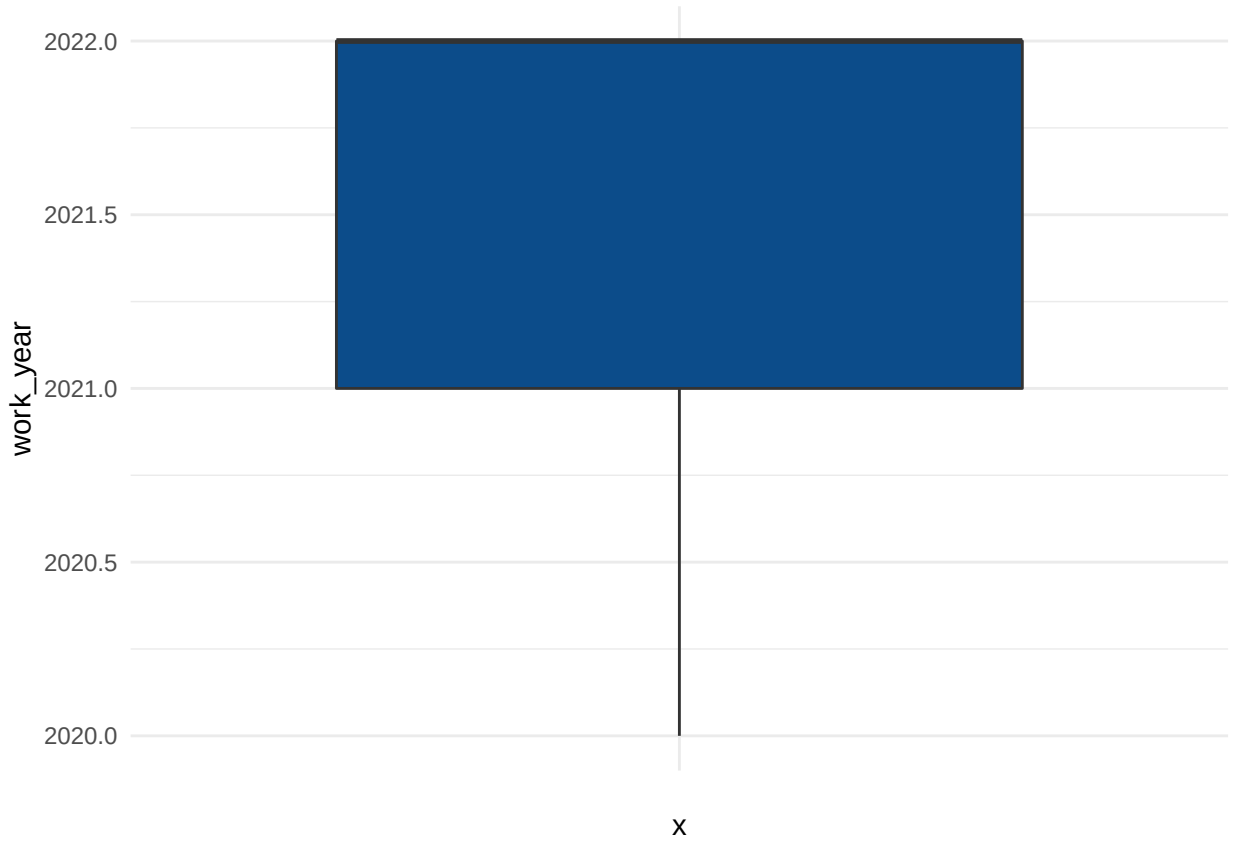
```
ggplot(datos_salarios) +
  aes(x = "", y = salary_in_usd) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```




```
# Boxplot para Work Year
summary(datos_salarios$work_year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2020    2021    2022    2021    2022    2022
```

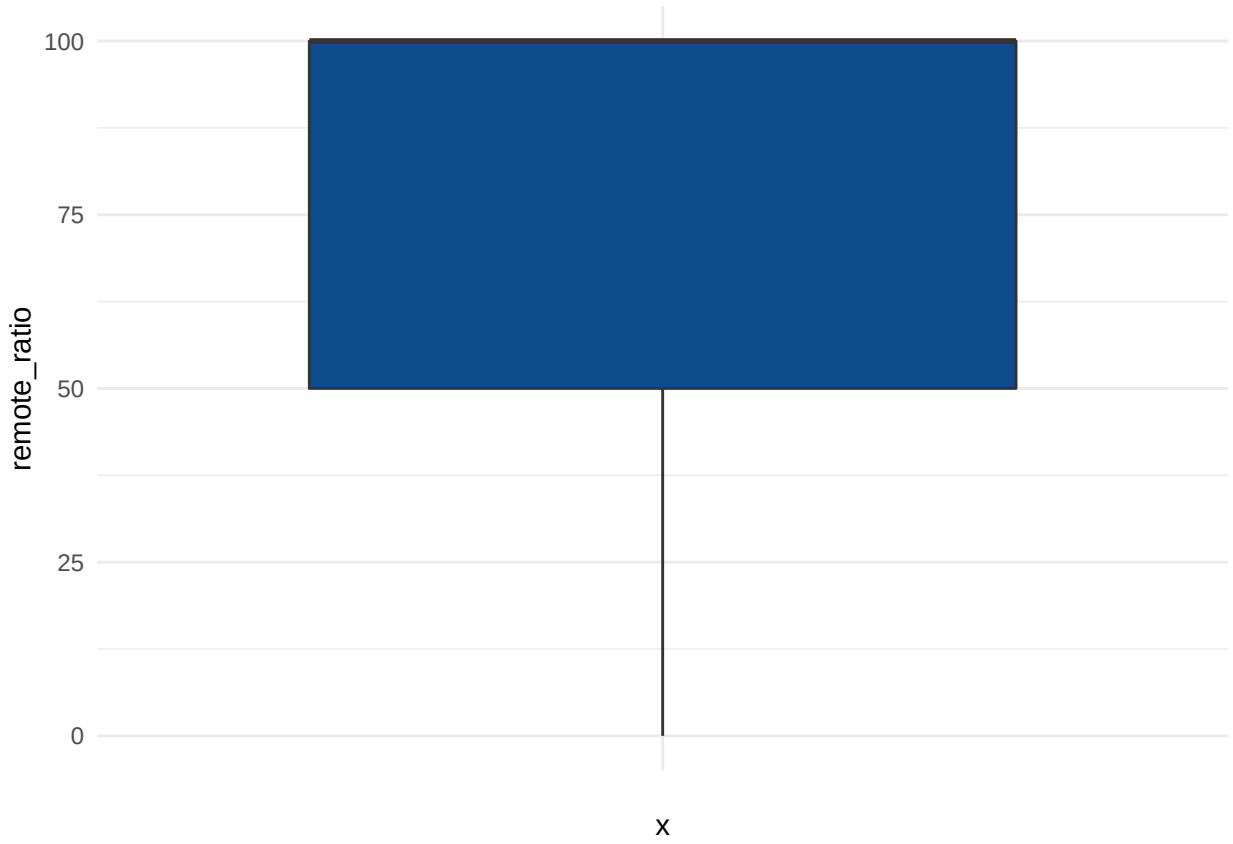
```
ggplot(datos_salarios) +
  aes(x = "", y = work_year) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```




```
# Boxplot para Remote Ratio

summary(datos_salarios$remote_ratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   50.00  100.00   70.92  100.00  100.00
```

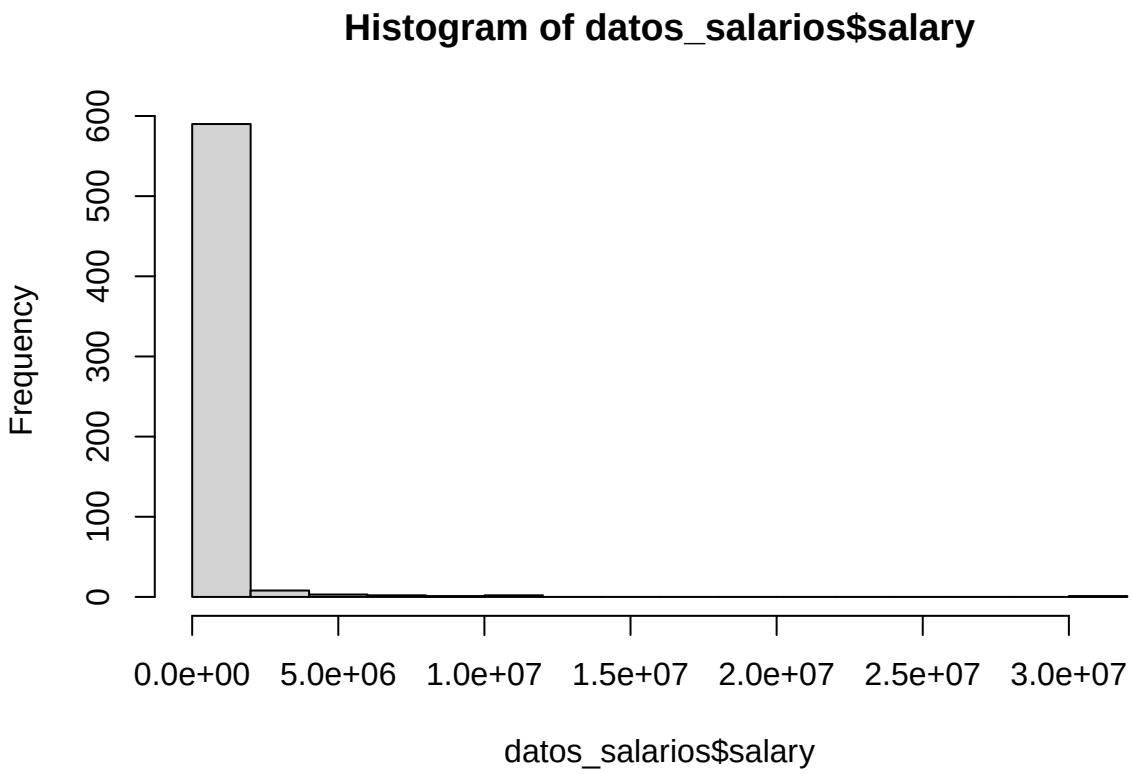
```
ggplot(datos_salarios) +
  aes(x = "", y = remote_ratio) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



Análisis de distribución de los datos (Histogramas). Identificar si tiene forma simétrica o asimétrica

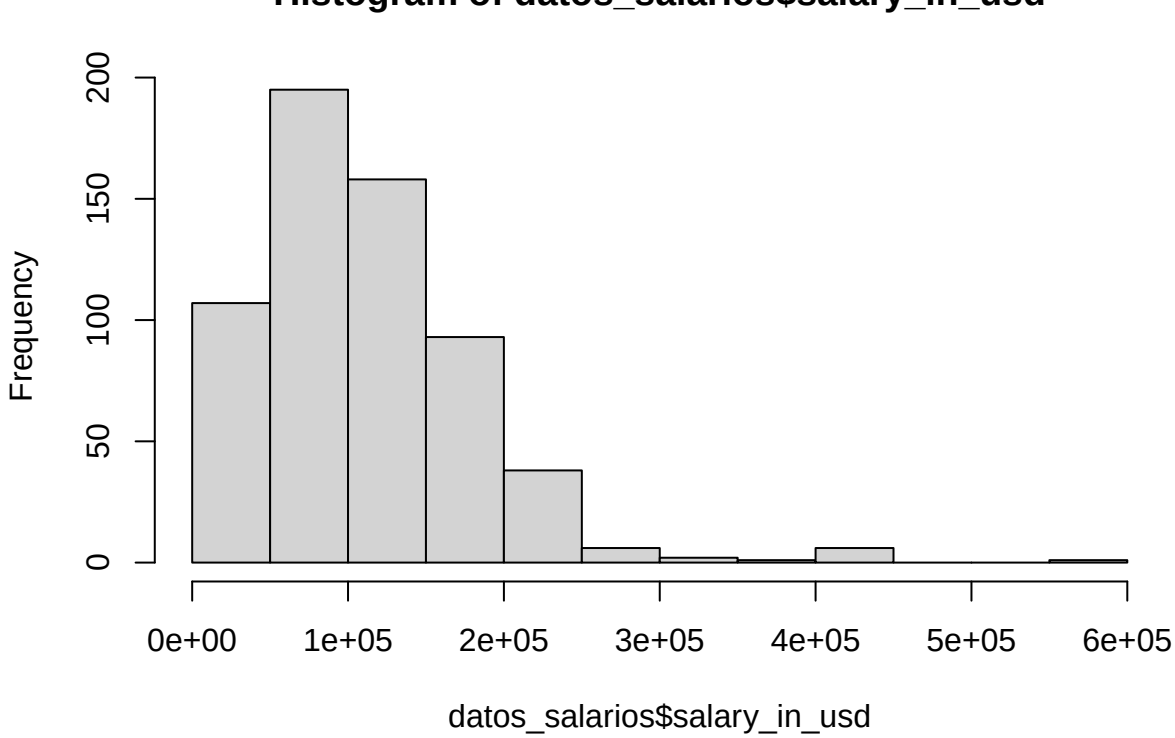
Histograma de Salary

```
hist(datos_salarios$salary)
```



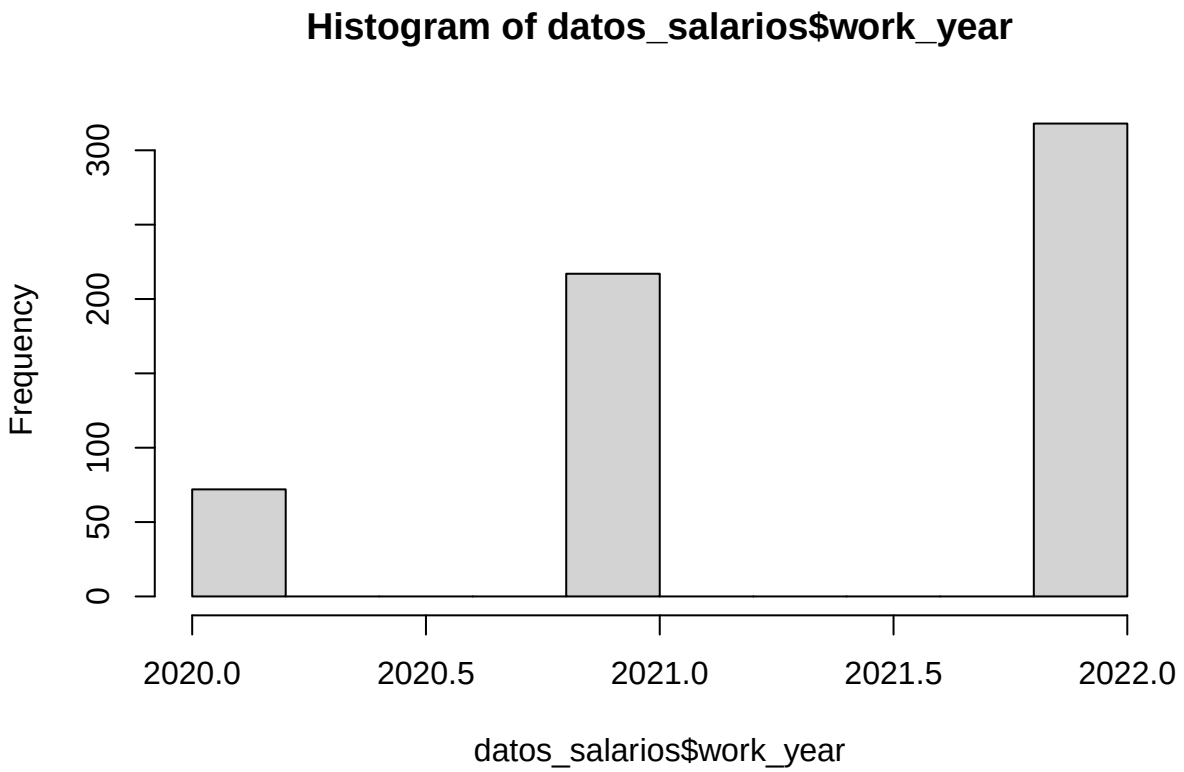
Histograma de Salary in usd

```
hist(datos_salarios$salary_in_usd)
```



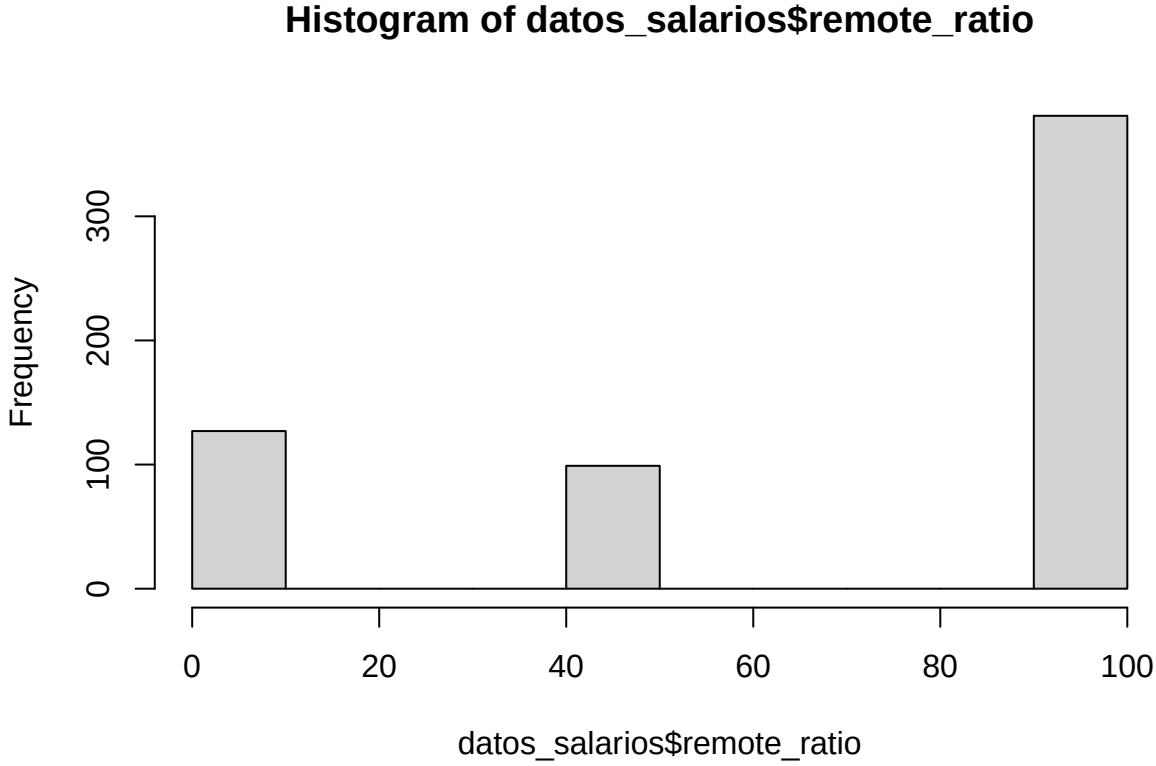
Histograma de Work Year

```
hist(datos_salarios$work_year)
```



Histograma de Remote Ratio

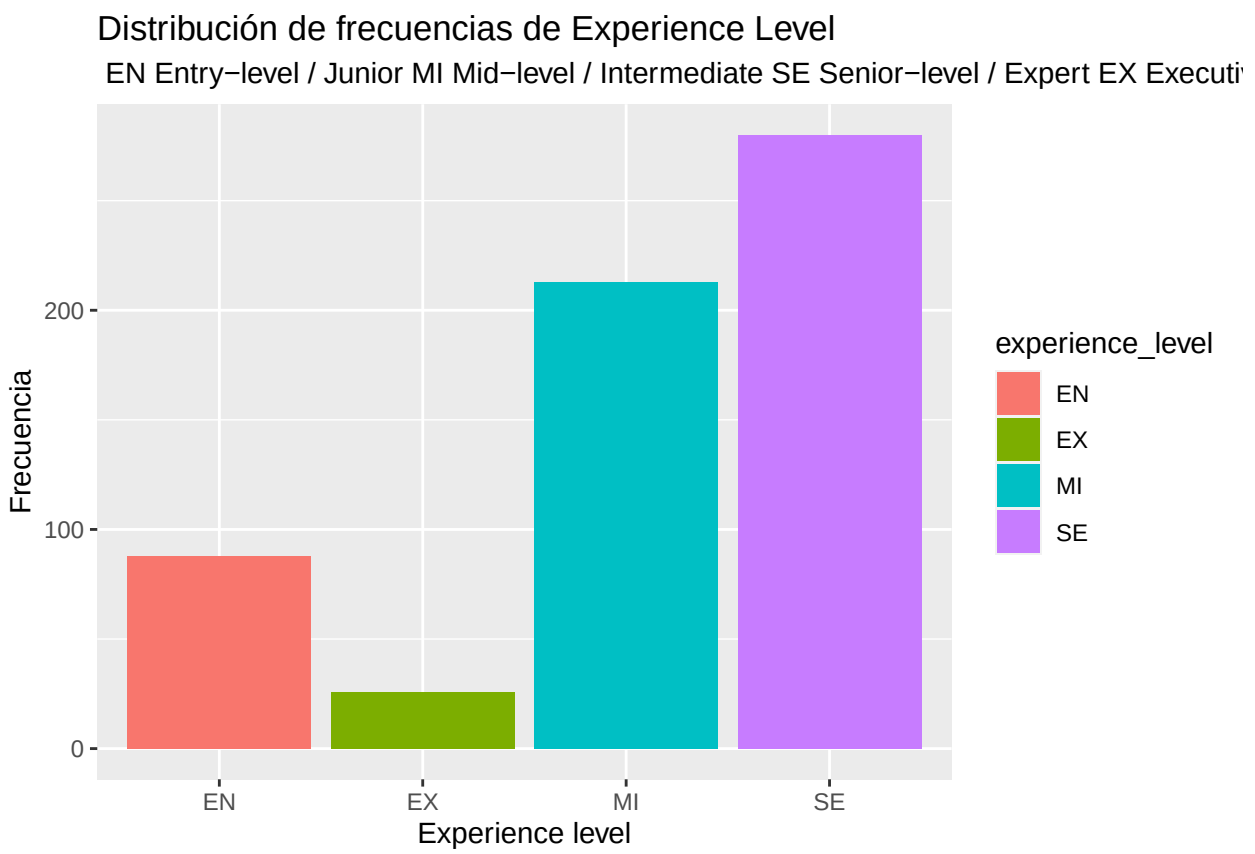
```
hist(datos_salarios$remote_ratio)
```



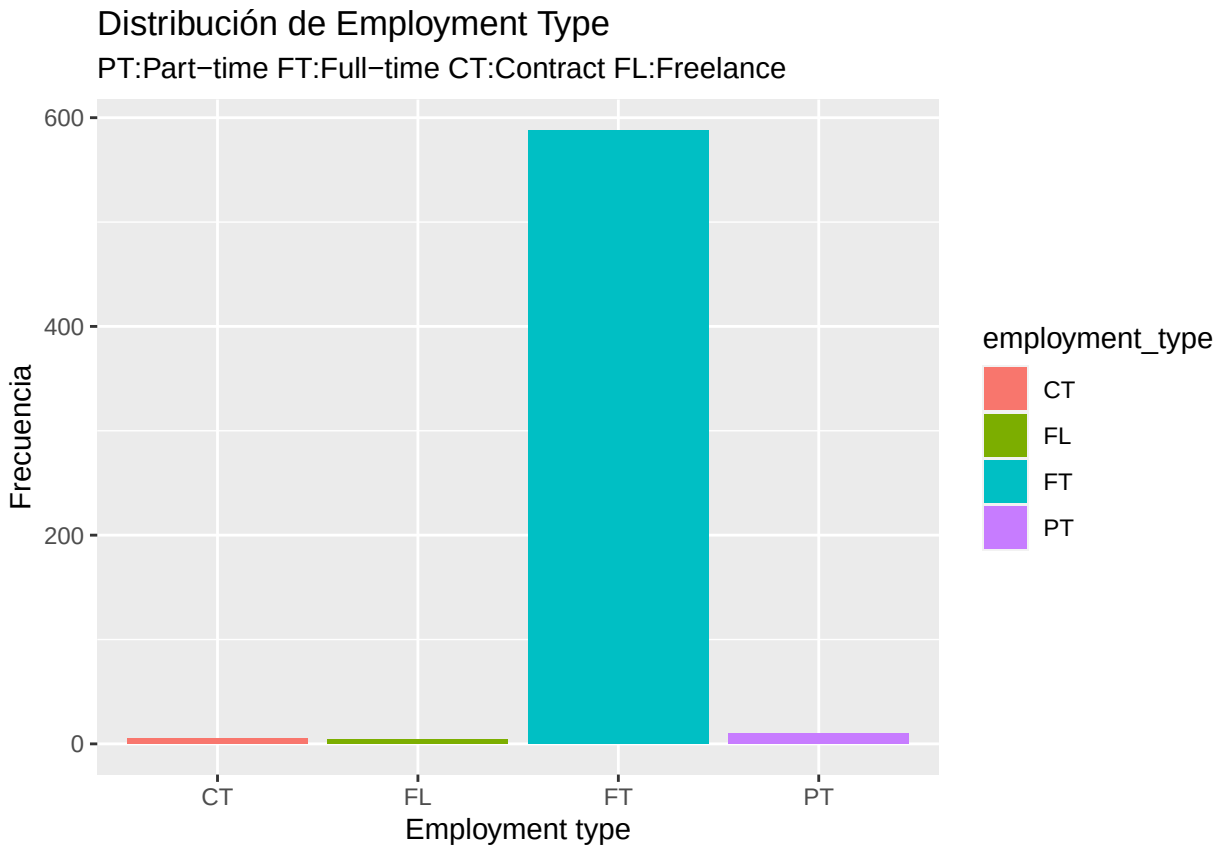
Variables categóricas

Distribución de los datos (diagramas de barras, diagramas de pastel)

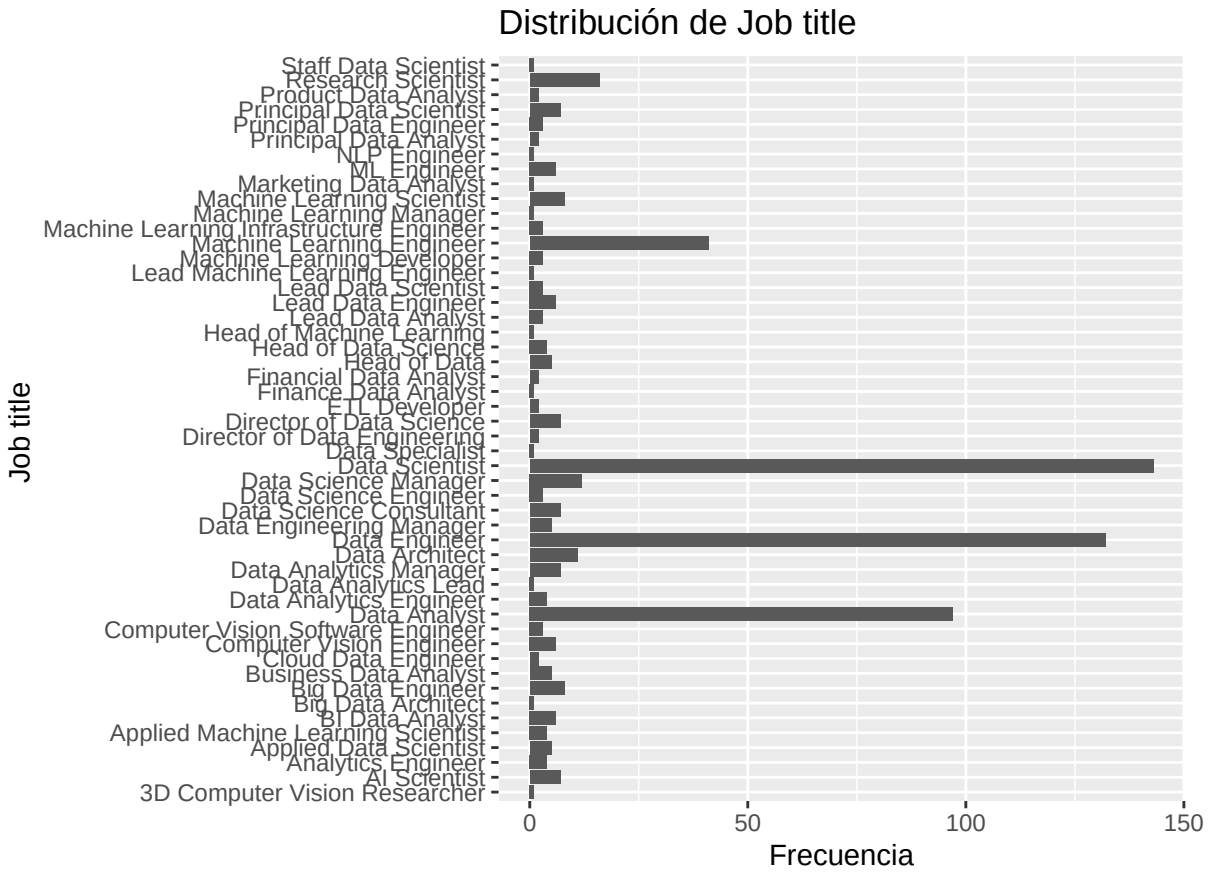
```
# Para ver distribución por colores de Employment Level

ggplot(data=datos_salarios) +
geom_bar(aes(x = experience_level,  fill = experience_level))  + labs(title="Distribución de frecuencias
```



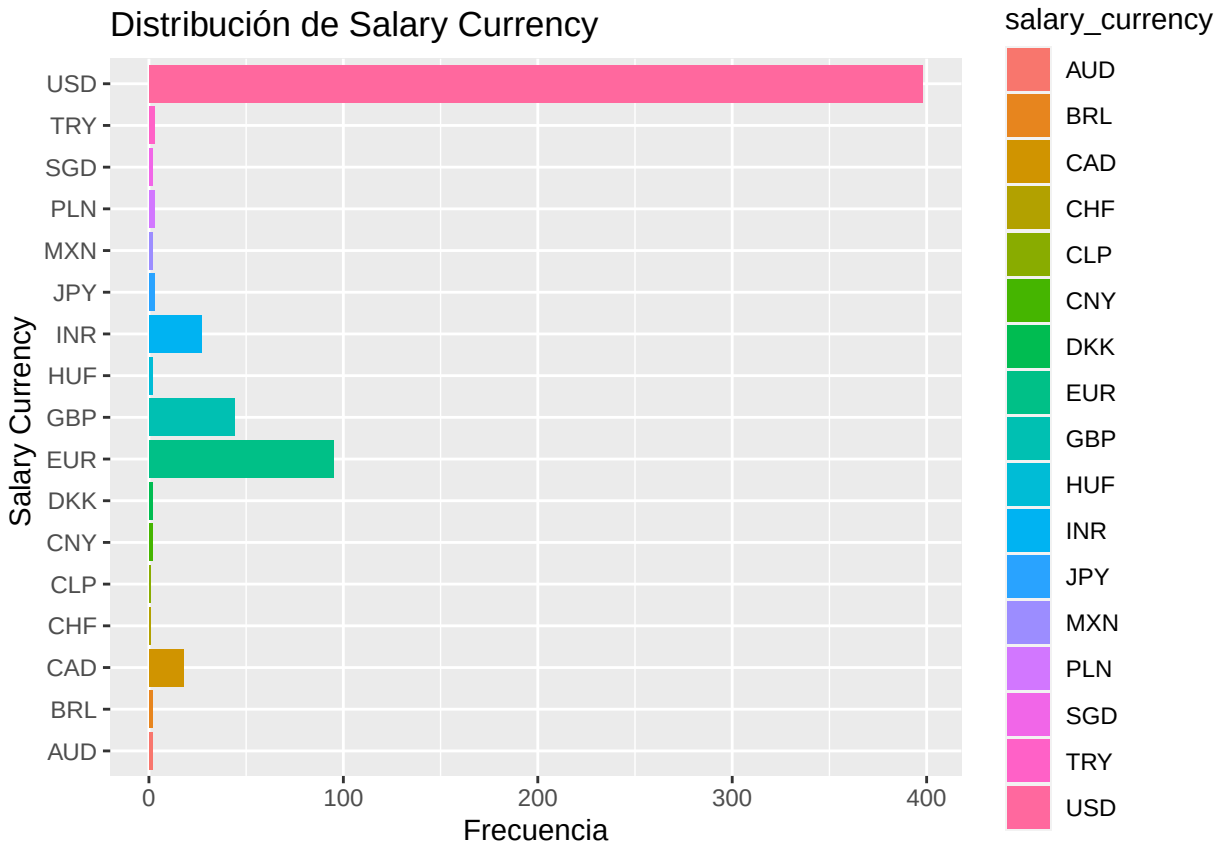

```
# Para ver distribución por colores de Employment Type

ggplot(data=datos_salarios) +
geom_bar(aes(x = employment_type,  fill = employment_type))  + labs(title="Distribución de Employment Ty
```



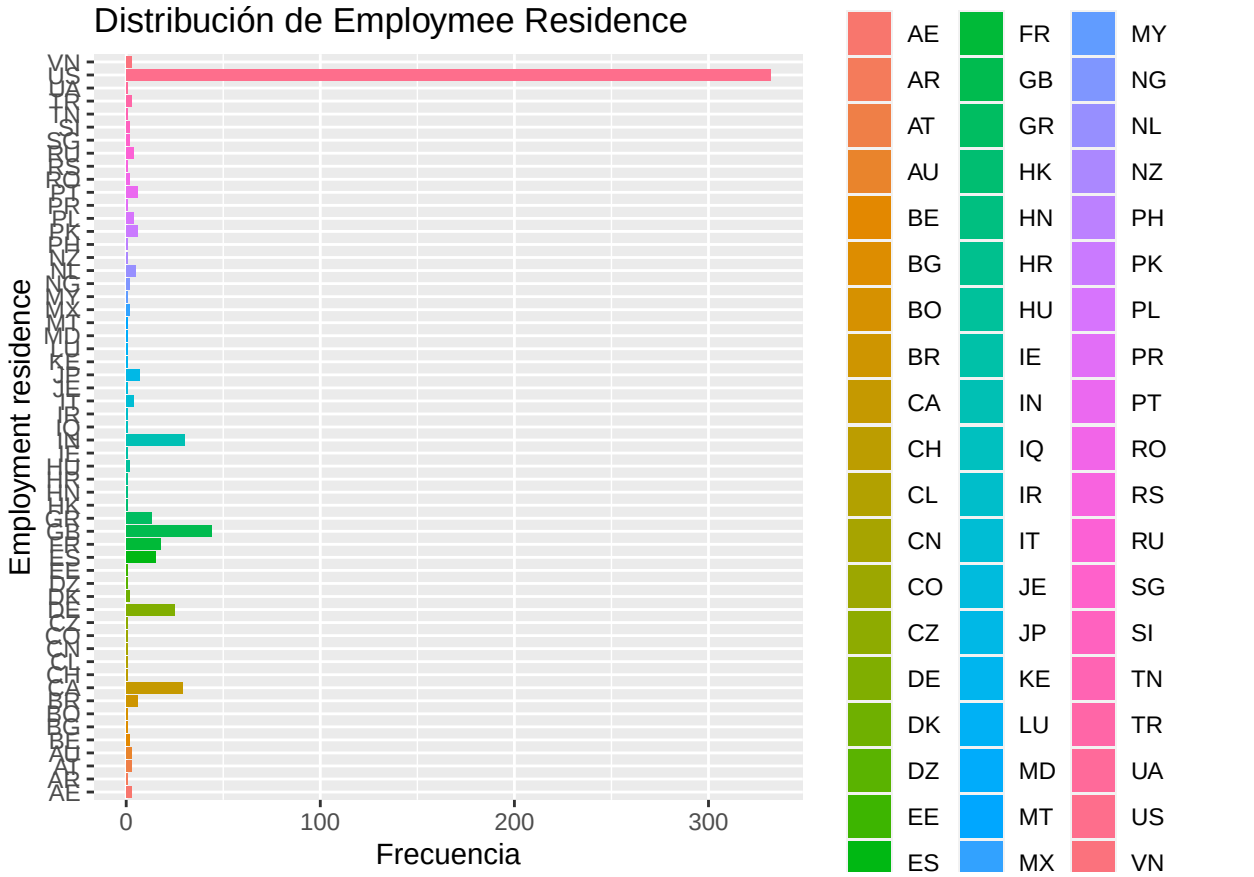

```
# Para ver distribución de Job Title

ggplot(data = datos_salarios, aes(x = job_title), size(0.30)) +
    geom_bar(position = "dodge",  width=0.9)+coord_flip() + theme(plot.margin = margin(0.0001,.8,0.02,.
```



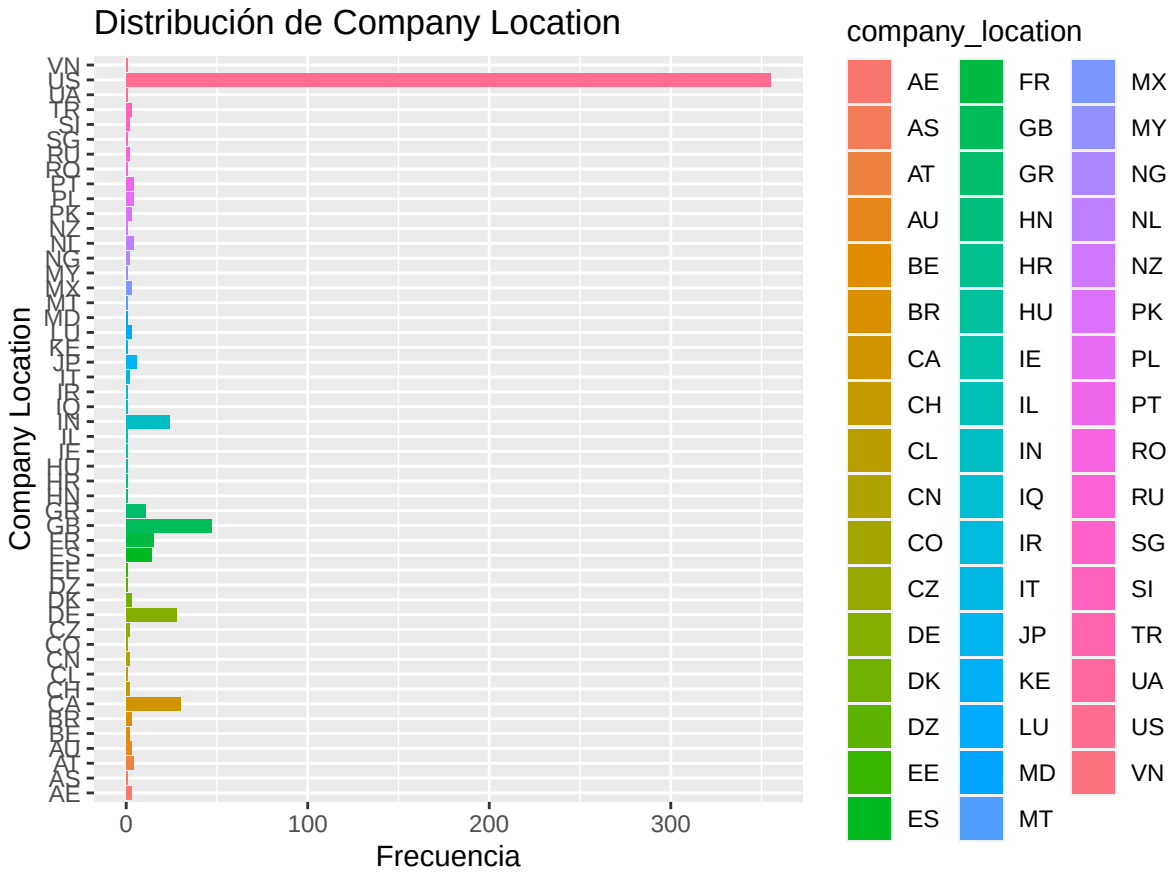

```
# Para ver la tabla de frecuencias por colores de Salary Currency

ggplot(data=datos_salarios) +
geom_bar(aes(x = salary_currency,  fill = salary_currency))  + labs(title="Distribución de Salary Curre
```



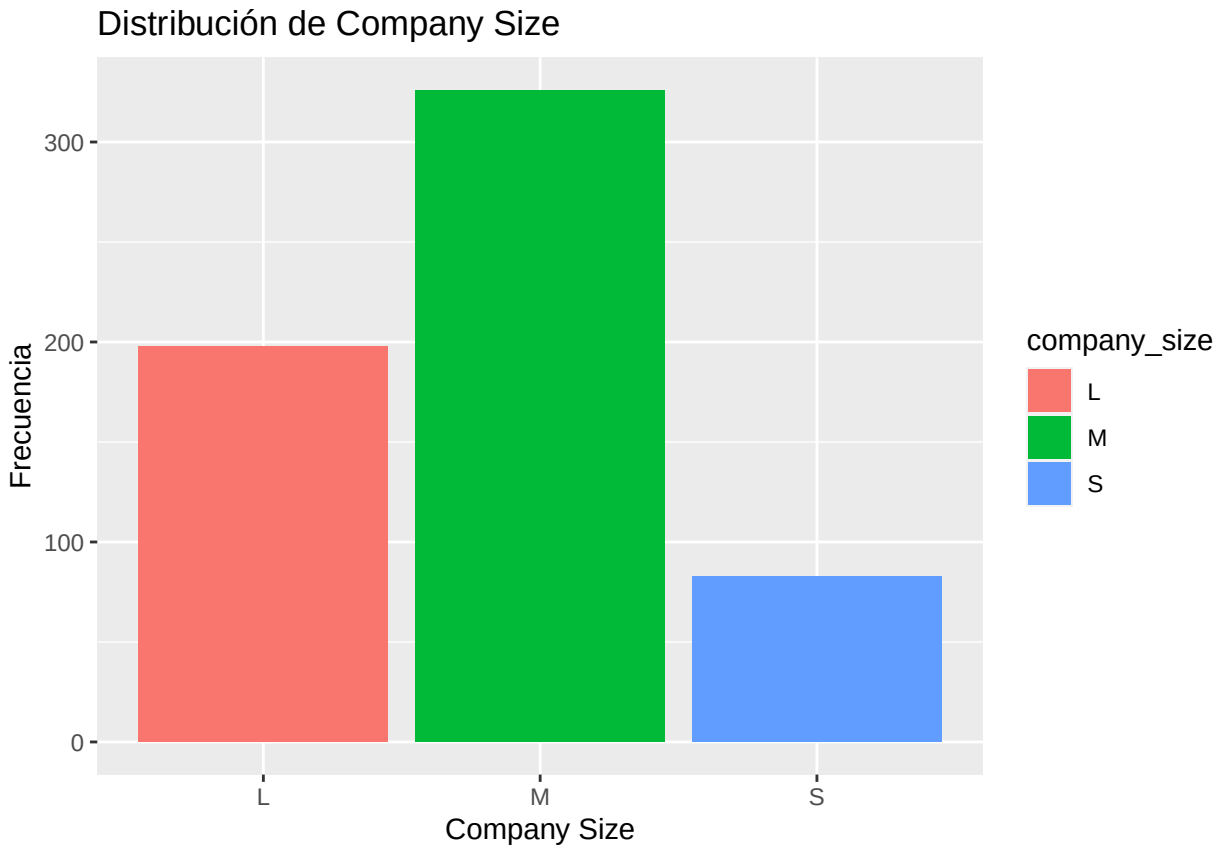

```
# Para ver distribución por colores de Employee Residence

ggplot(data=datos_salarios) +
geom_bar(aes(x = employee_residence,  fill = employee_residence))  + labs(title="Distribución de Employ
```




```
# Para ver distribución  por colores de Company Location
ggplot(data=datos_salarios) +
geom_bar(aes(x = company_location,  fill = company_location))  + labs(title="Distribución de Company Lo
```




```
# Para ver distribución  por colores de Company Size
ggplot(data=datos_salarios) +
geom_bar(aes(x = company_size,  fill = company_size))  + labs(title="Distribución de Company Size", x =
```



D) Preparación de los datos

1) Selecciona el conjunto de datos a utilizar

Tomando en consideración las preguntas objetivo a contestar: 2. ¿En qué países se ofrecen mejores salarios? El conjunto de datos necesarios son: - company_location y salary_in_usd

Para este caso, se incluyó una nueva base de datos llamada iso3166, y a partir de ella se pudo graficar un mapa mundial asociado a los paises incluidos en "company_location" del data set original siendo Salaries.csv.

También se incluyó una nueva variable para transformar la columna de "company_location" en 0 y 1 para facilitar la visualización. Al ser un primer acercamiento, faltan cosas por mejorar en dicha gráfica del mundo como asociar por colores el salario promedio, pero por el momento funciona para ver con qué paises se está trabajando.

Así mismo, se dio con la respuesta con una gráfica de barras. (Procedimiento se encontrará más adelante)

3. ¿Se han incrementado los salarios a lo largo del tiempo?

- salary_in_usd y work_year

En este caso no se requirió de transformar, discretizar, agregar variables y demás. Se obtuvo la respuesta de manera directa al graficar el promedio de Salary in USD vs Work Year. (Procedimiento se encontrará más adelante)

6. ¿Qué tipo de contrato (parcial, tiempo completo, etc) ofrece mejores salarios? ¿Qué tipo de contrato será el más conveniente? El conjunto de datos necesarios son:

- experience_level y salary_in_usd

En este caso no se requirió de transformar, discretizar, agregar variables y demás. Se obtuvo la respuesta de manera directa al graficar el promedio de Salary in USD vs Employment Type. (Procedimiento se encontrará más adelante)

4. ¿Influye el nivel de experiencia en el salario? El conjunto de datos necesarios son:

- experience_level y salary_in_usd

En este caso no se requirió de transformar, discretizar, agregar variables y demás. Se obtuvo la respuesta de manera directa al graficar el promedio de Salary in USD vs Experience level. (Procedimiento se encontrará más adelante)

Procedimiento para responder a "2. ¿En qué países se ofrecen mejores salarios?"

```
# Cargamos las librerias a utilizar
library(countrycode)
library(highcharter)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(maps)
```

```
# Para ver las abreviaturas con las que se están trabajando en Salaries.csv
#datos_salarios$company_location
```

```
# Cargamos dataset iso3166, ya que es el tipo de "country code" que se utiliza en los datos originales

dat <- iso3166
head(dat)
```

```
##   a2  a3       ISOname               mapname sovereignty
## 1 AW ABW         Aruba                 Aruba Netherlands
## 2 AF AFG   Afghanistan           Afghanistan Afghanistan
## 3 AO AGO        Angola                Angola      Angola
## 4 AI AIA      Anguilla              Anguilla    Anguilla
## 5 AX ALA Aland Islands Finland:Aland Islands     Finland
## 6 AL ALB       Albania               Albania     Albania
```

```
# Renombramos la variable a3 a iso-a3, con el fin de identificar más fácilmente esta variable al moment
```

```
dat <- rename(dat, "iso-a3" = a3)
head(dat)
```

```
##   a2 iso-a3       ISOname               mapname sovereignty
## 1 AW    ABW         Aruba                 Aruba Netherlands
## 2 AF    AFG   Afghanistan           Afghanistan Afghanistan
## 3 AO    AGO        Angola                Angola      Angola
## 4 AI    AIA      Anguilla              Anguilla    Anguilla
## 5 AX    ALA Aland Islands Finland:Aland Islands     Finland
## 6 AL    ALB       Albania               Albania     Albania
```

```
# Se imprimen los códigos de los países que se encuentran en "Salaries.csv", con el fin de comparar qué

countries_of_company_location <- datos_salarios$company_location
countries_of_company_location
```

```
##   [1] "DE" "JP" "GB" "HN" "US" "US" "US" "HU" "US" "NZ" "FR" "IN" "FR" "US" "US"
##  [16] "PK" "JP" "GB" "IN" "US" "CN" "IN" "GR" "US" "AE" "US" "NL" "MX" "US" "CA"
##  [31] "DE" "US" "US" "US" "FR" "AT" "US" "US" "NG" "US" "US" "ES" "PT" "US" "GB"
##  [46] "DE" "GB" "US" "US" "FR" "IN" "US" "DK" "DE" "US" "DE" "ES" "US" "US" "US"
##  [61] "US" "US" "IT" "US" "HR" "DE" "DE" "US" "US" "AT" "LU" "FR" "GB" "US" "US"
##  [76] "FR" "US" "IN" "US" "US" "DE" "US" "CA" "ES" "PL" "PL" "FR" "US" "US" "US"
##  [91] "DK" "DE" "IN" "US" "IN" "SG" "US" "US" "US" "US" "US" "US" "US" "US" "US"
## [106] "GB" "CA" "US" "US" "IN" "DE" "GB" "GB" "US" "NL" "US" "NG" "GR" "US" "US"
## [121] "RO" "US" "US" "GB" "ES" "US" "US" "IN" "IN" "IN" "IQ" "FR" "US" "BR" "US"
## [136] "US" "JP" "JP" "US" "US" "US" "US" "US" "US" "US" "BE" "FR" "US" "US" "US"
## [151] "JP" "US" "CA" "UA" "US" "CA" "IL" "US" "US" "US" "RU" "RU" "MT" "DE" "DE"
## [166] "US" "US" "US" "US" "US" "US" "GB" "US" "US" "PT" "US" "MX" "CL" "US" "US"
## [181] "IN" "DE" "US" "GB" "PK" "IR" "GB" "ES" "GB" "JP" "FR" "CO" "MD" "US" "CA"
## [196] "US" "KE" "IN" "US" "US" "AT" "US" "ES" "US" "US" "BR" "US" "US" "US" "US"
## [211] "SI" "FR" "GB" "CH" "DE" "US" "DK" "DE" "BE" "US" "PL" "GB" "IN" "GB" "CA"
## [226] "US" "CA" "DE" "US" "CA" "IN" "US" "US" "US" "US" "US" "CA" "ES" "VN" "IN"
## [241] "CA" "US" "US" "US" "AS" "GB" "FR" "TR" "GB" "US" "US" "US" "US" "IN" "US"
## [256] "CA" "US" "DE" "US" "DE" "US" "DE" "IN" "IN" "DE" "US" "US" "US" "TR" "DE"
## [271] "US" "BR" "DE" "DE" "FR" "US" "US" "ES" "TR" "LU" "US" "CN" "NL" "US" "CZ"
## [286] "IN" "SI" "US" "IT" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US"
## [301] "GB" "GB" "US" "US" "GB" "US" "US" "US" "US" "US" "US" "GB" "GB" "GB" "GB"
## [316] "US" "GB" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US"
## [331] "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US"
## [346] "US" "US" "US" "US" "US" "US" "US" "US" "US" "GB" "GB" "US" "US" "CA" "CA"
## [361] "CA" "CA" "CA" "CA" "US" "US" "US" "US" "US" "US" "US" "US" "US" "GR" "GR"
## [376] "CA" "US" "US" "US" "US" "US" "US" "US" "US" "US" "IN" "US" "GB" "US" "US" "GB"
## [391] "GB" "US" "US" "US" "US" "US" "DE" "GB" "US" "US" "US" "US" "US" "US" "US"
## [406] "GB" "US" "US" "GB" "US" "GB" "GB" "GR" "GR" "GB" "GB" "US" "MX" "US" "US"
## [421] "US" "US" "US" "US" "US" "US" "US" "ES" "US" "GB" "ES" "ES" "ES" "ES" "GB"
## [436] "GB" "ES" "GR" "US" "US" "GR" "GR" "GB" "GB" "US" "GR" "US" "US" "US" "PT"
## [451] "US" "CA" "CA" "US" "US" "US" "US" "DE" "IN" "IN" "PT" "US" "DE" "IN" "DE"
## [466] "US" "US" "US" "US" "US" "US" "US" "US" "US" "GB" "GB" "US" "US" "US" "US"
## [481] "AE" "AE" "US" "US" "US" "US" "US" "DZ" "US" "CZ" "US" "CA" "PL" "CA" "US"
## [496] "US" "DE" "US" "FR" "CA" "NL" "EE" "MY" "AU" "US" "AU" "US" "AT" "US" "US"
## [511] "AU" "CA" "US" "IE" "PK" "US" "US" "FR" "CH" "US" "CA" "LU" "GR" "US" "US"
## [526] "US" "US" "US" "US" "US" "CA" "CA" "US" "US" "US" "US" "US" "US" "US" "US"
## [541] "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US"
## [556] "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "GB" "US" "US"
## [571] "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US"
```

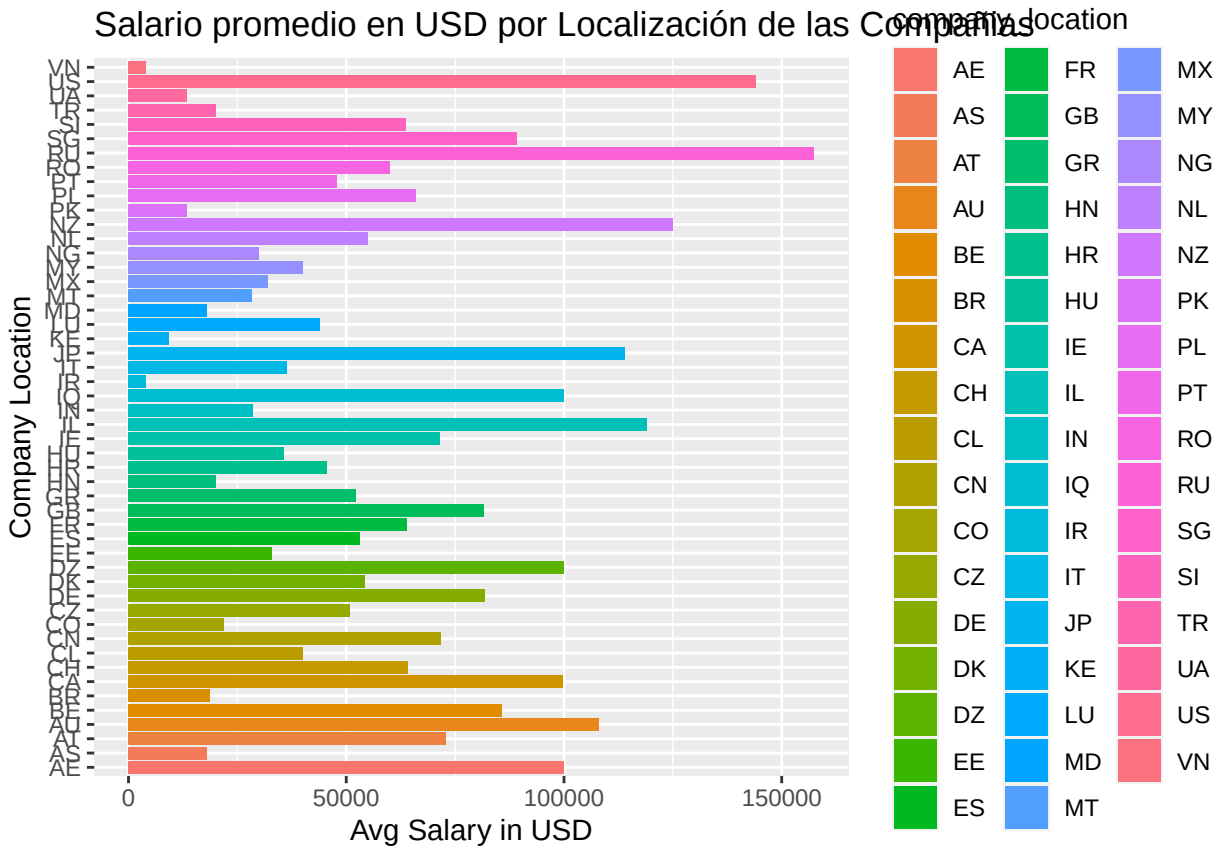
```
## [586] "US" "GB" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US" "US"
## [601] "CA" "CA" "US" "US" "US" "US" "US"
```

```
# La columna "a2" del data set cargado coincide con el mismo tipo de dato y valor de la columna "compan

dat$countries_of_company_location <- ifelse(dat$`a2` %in% countries_of_company_location, 1, 0)
head(dat, 10) # Se encuentra en la última columna esta nueva variable
```

```
##    a2 iso-a3              ISOname               mapname          sovereignty
## 1  AW    ABW                Aruba                 Aruba          Netherlands
## 2  AF    AFG          Afghanistan           Afghanistan          Afghanistan
## 3  AO    AGO               Angola                Angola               Angola
## 4  AI    AIA             Anguilla              Anguilla             Anguilla
## 5  AX    ALA        Aland Islands Finland:Aland Islands              Finland
## 6  AL    ALB              Albania               Albania              Albania
## 7  AD    AND              Andorra               Andorra              Andorra
## 8  AE    ARE United Arab Emirates  United Arab Emirates United Arab Emirates
## 9  AR    ARG            Argentina             Argentina            Argentina
## 10 AM    ARM              Armenia               Armenia              Armenia
##    countries_of_company_location
## 1                              0
## 2                              0
## 3                              0
## 4                              0
## 5                              0
## 6                              0
## 7                              0
## 8                              1
## 9                              0
## 10                             0
```

```
# Aqui se genera el mapa, marcando en azul los paises que corresponden a "company_location" del dataset

hcmap( # hcmap viene del paquete {highcharter}
  map = "custom/world-highres3", # Este parámetro muestra un mapa en alta resolución
  data = dat, # name of dataset
  joinBy = "iso-a3",
  value = "countries_of_company_location",
  showInLegend = FALSE, # hide legend
  nullColor = "#DADADA",
  download_map_data = TRUE
) %>%
  hc_mapNavigation(enabled = FALSE) %>%
  hc_legend("none") %>%
  hc_title(text = "Paises que se encuentran en la columna Location Company")
```

```
# Ahora se grafica un diagrama de barras, representando el Salary in USD vs company Location
ggplot(data=datos_salarios) +
geom_bar(aes(x = company_location, y = salary_in_usd, fill = company_location), stat = "summary", fun =
  labs(title="Salario promedio en USD por Localización de las Compañias", x = "Company Location", y = "
```




```
# Al ser 50 paises, es difícil observar cuáles realmente ofrecen un mejor salario en promedio. Por lo q


groupb_by_company_location_salary_in_usd = datos_salarios %>% group_by(company_location)  %>%
                    summarise(salary_in_usd = mean(salary_in_usd),
                              .groups = 'drop')
groupb_by_company_location_salary_in_usd
```

```
## # A tibble: 50 x 2
##    company_location salary_in_usd
##    <chr>                    <dbl>
##  1 AE                     100000
##  2 AS                      18053
##  3 AT                      72921.
##  4 AU                     108043.
##  5 BE                      85699
##  6 BR                      18603.
##  7 CA                      99824.
##  8 CH                      64114
##  9 CL                      40038
## 10 CN                      71666.
## # ... with 40 more rows
```

```
# Ahora lo que se realiza es ordenar de mayor a menor los datos, observando directamente el top 10 pais

groupb_by_company_location_salary_in_usd_decreasing <- groupb_by_company_location_salary_in_usd[order(g
groupb_by_company_location_salary_in_usd_decreasing
```

```
## # A tibble: 50 x 2
##    company_location salary_in_usd
##    <chr>                    <dbl>
##  1 RU                     157500
##  2 US                     144055.
##  3 NZ                     125000
##  4 IL                     119059
##  5 JP                     114127.
##  6 AU                     108043.
##  7 AE                     100000
##  8 DZ                     100000
##  9 IQ                     100000
## 10 CA                      99824.
## # ... with 40 more rows
```

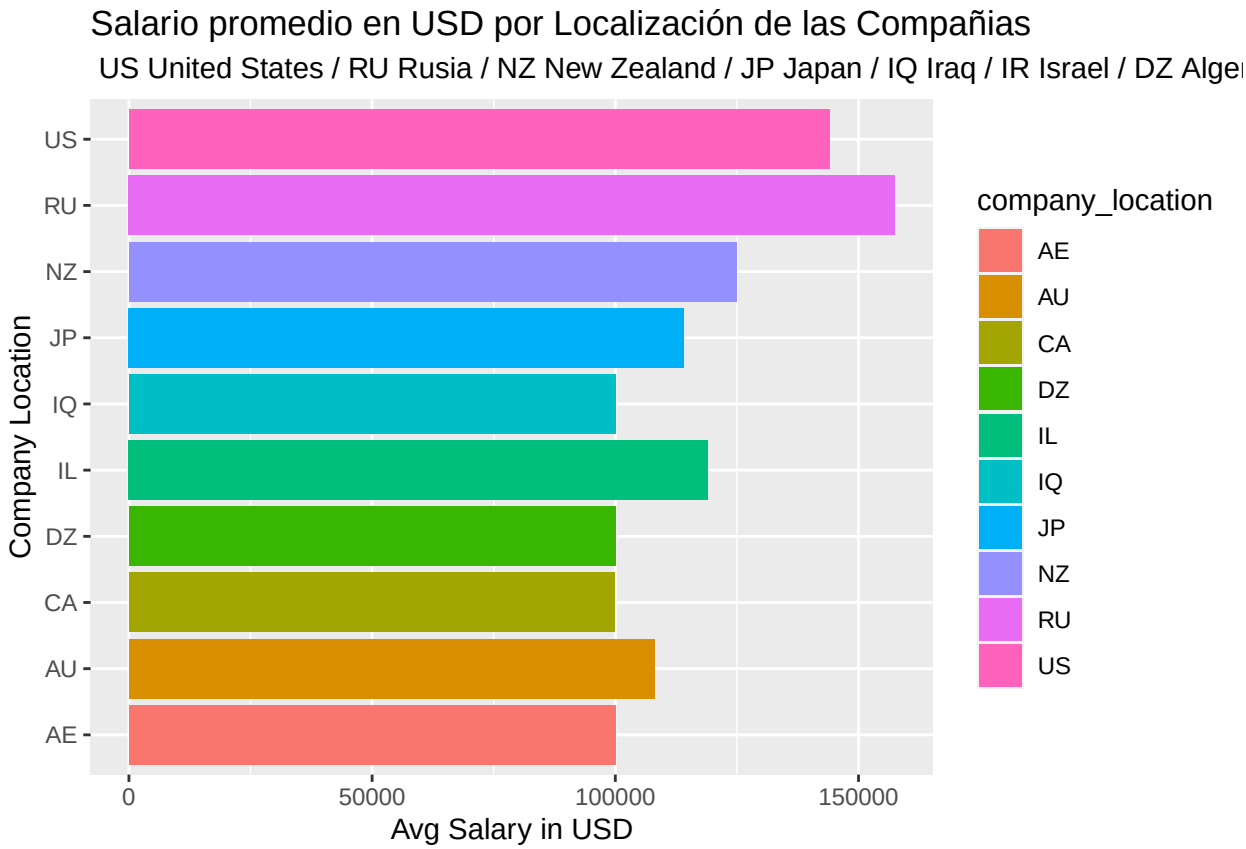
```
# Ahora se grafican los TOP 10 países con promedio de salario más alto para visualizarlo gráficamente
groupb_by_company_location_salary_in_usd_decreasing_TOP10 = groupb_by_company_location_salary_in_usd_de

groupb_by_company_location_salary_in_usd_decreasing_TOP10
```

```
## # A tibble: 10 x 2
##    company_location salary_in_usd
##    <chr>                    <dbl>
##  1 RU                     157500
##  2 US                     144055.
##  3 NZ                     125000
##  4 IL                     119059
##  5 JP                     114127.
##  6 AU                     108043.
##  7 AE                     100000
##  8 DZ                     100000
##  9 IQ                     100000
## 10 CA                      99824.
```

```
print("US United States / RU Rusia / NZ New Zealand / JP Japan / IQ Iraq / IR Israel / DZ Algeria / CA
```

```
## [1] "US United States / RU Rusia / NZ New Zealand / JP Japan / IQ Iraq / IR Israel / DZ Algeria / CA
```

```
ggplot(data=groupb_by_company_location_salary_in_usd_decreasing_TOP10) +
geom_bar(aes(x = company_location, y = salary_in_usd, fill = company_location), stat = "identity") +
  labs(title="Salario promedio en USD por Localización de las Compañias", x = "Company Location", y = "
```

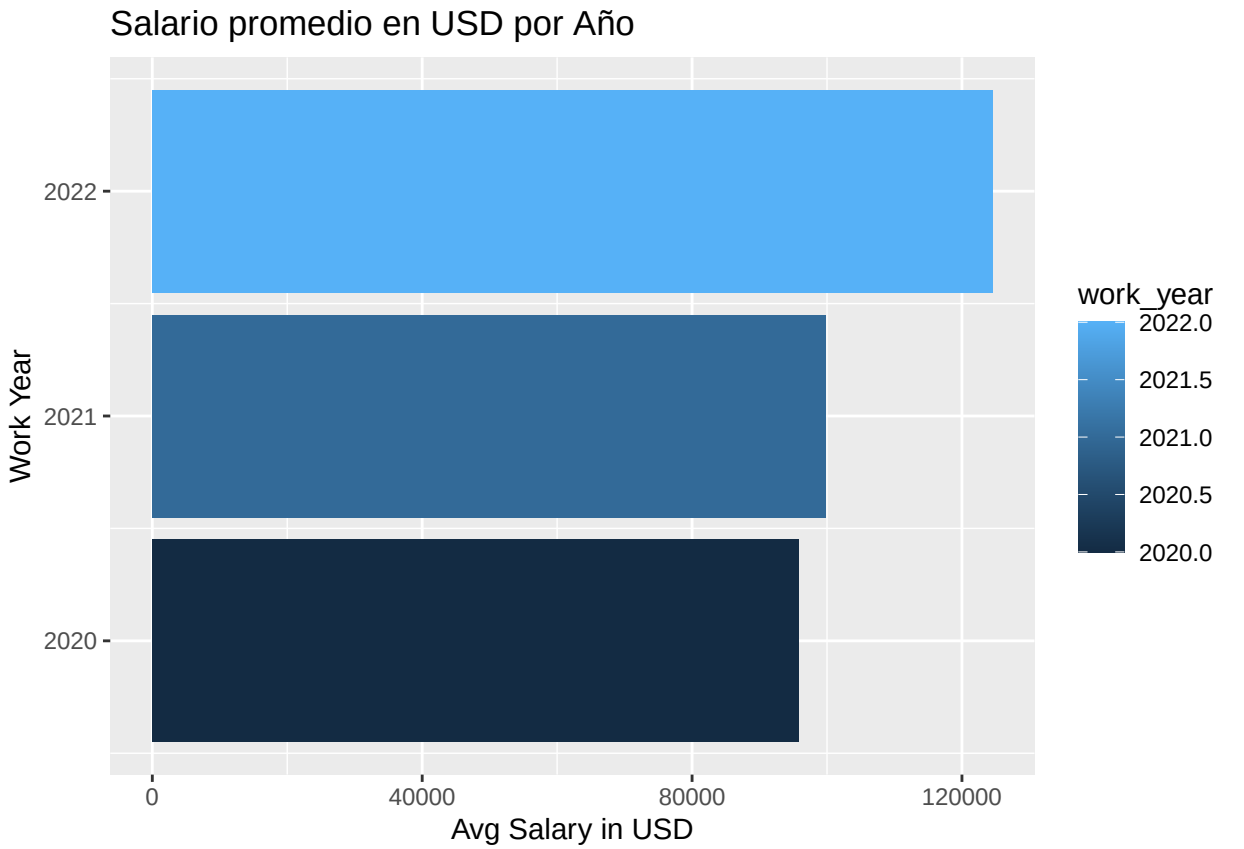



```
#testing1 <- dat[(dat$a2 == "AE"),]
#testing1
```

Respuesta: Con esta gráfica, se puede responder más fácilmente la pregunta "¿En qué países se ofrecen mejores salarios?". Observando la gráfica, los paises que ofrecen mejores salarios son: - US United States - RU Rusia - NZ New Zealand - JP Japan - IQ Iraq - IR Israel - DZ Algeria - CA Canada - AU Australia - AE United Arab Emirates

Procedimiento para responder a "3. ¿Se han incrementado los salarios a lo largo del tiempo?"

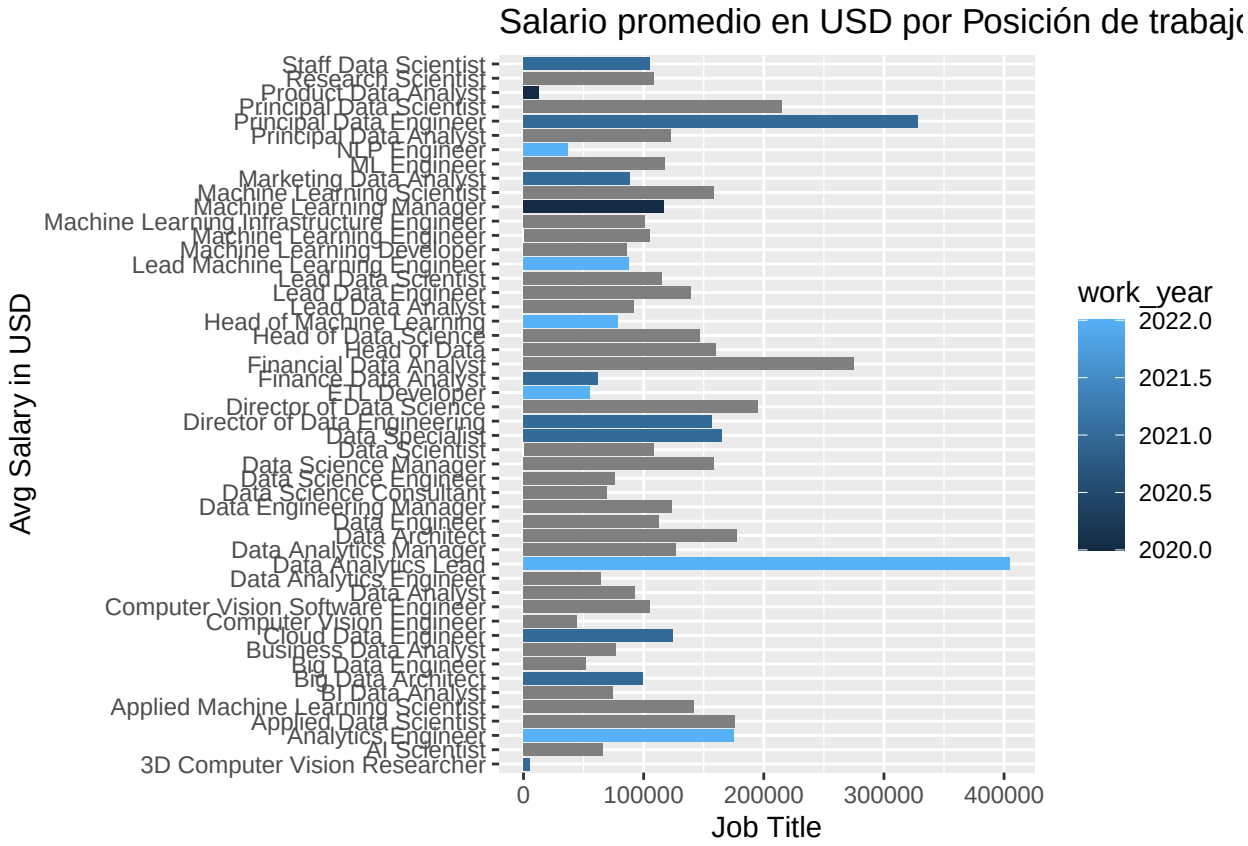
```
# Ahora se grafica un diagrama de barrar, representando el Salary in USD vs Work Year
ggplot(data=datos_salarios) +
geom_bar(aes(x = work_year, y = salary_in_usd, fill = work_year), stat = "summary", fun = "mean") +
  labs(title="Salario promedio en USD por Año", x = "Work Year", y = "Avg Salary in USD")+ coord_flip()
```



Respuesta: Como se observa la gráfica, obteniendo los promedios de Salary in USD por año, se puede afirmar que a medida que pasan los años incrementa considerablemente el salario promedio, llegando a más de 120,000 dólares.

Con la siguiente gráfica también se puede observar qué "Job title" tiene un mejor salario promedio a medida que pasa el tiempo:

```
options(scipen = 500)

ggplot(data=datos_salarios) +
geom_bar(aes(x = salary_in_usd, y = job_title, fill = work_year), stat = "summary", fun = "mean") +
      labs(title="Salario promedio en USD por Posición de trabajo y Años", x = "Job Title", y = "Avg Sa
```



Al parecer, viendo la gráfica, "Data Analytics Lead" tiene el mejor salario en la actualidad. Y para comprobarlo:

```
datos_salarios_testing <- datos_salarios[(datos_salarios$job_title == "Data Analytics Lead"),]
datos_salarios_testing
```

```
##       X work_year experience_level employment_type           job_title salary
## 524 523      2022               SE              FT Data Analytics Lead 405000
##     salary_currency salary_in_usd employee_residence remote_ratio
## 524             USD        405000                 US          100
##     company_location company_size
## 524               US            L
```

Con el siguiente filtro se puede corrobar la información observada de la gráfica, al igual que ver las siguientes posiciones con los salarios más altos de manera descendiente:

```
groupb_by_job_title_salary_in_usd = datos_salarios %>% group_by(job_title)  %>%
                    summarise(salary_in_usd = mean(salary_in_usd),
                              .groups = 'drop')
groupb_by_job_title_salary_in_usd
```

```
## # A tibble: 50 x 2
##    job_title                          salary_in_usd
##    <chr>                                      <dbl>
##  1 3D Computer Vision Researcher               5409
##  2 AI Scientist                               66136.
##  3 Analytics Engineer                        175000
##  4 Applied Data Scientist                    175655
##  5 Applied Machine Learning Scientist        142069.
```

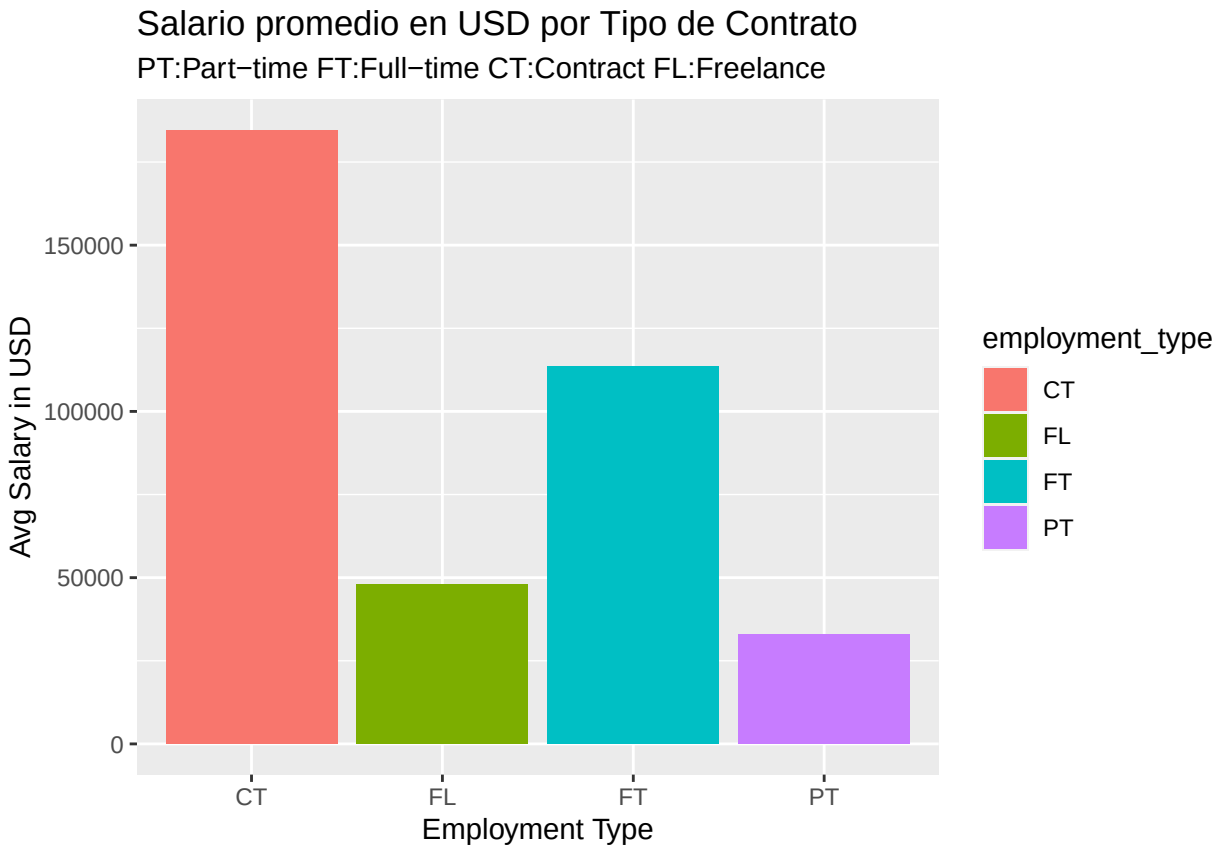
```
##  6 BI Data Analyst                          74755.
##  7 Big Data Architect                       99703
##  8 Big Data Engineer                        51974
##  9 Business Data Analyst                    76691.
## 10 Cloud Data Engineer                     124647
## # ... with 40 more rows
```

```
groupb_by_job_title_salary_in_usd_decreasing <- groupb_by_job_title_salary_in_usd[order(groupb_by_job_ti
groupb_by_job_title_salary_in_usd_decreasing
```

```
## # A tibble: 50 x 2
##    job_title                salary_in_usd
##    <chr>                            <dbl>
##  1 Data Analytics Lead             405000
##  2 Principal Data Engineer         328333.
##  3 Financial Data Analyst          275000
##  4 Principal Data Scientist        215242.
##  5 Director of Data Science        195074
##  6 Data Architect                  177874.
##  7 Applied Data Scientist          175655
##  8 Analytics Engineer              175000
##  9 Data Specialist                 165000
## 10 Head of Data                    160163.
## # ... with 40 more rows
```

Procedimiento para responder a "6. ¿Qué tipo de contrato (parcial, tiempo completo, etc) ofrece mejores salarios? ¿Qué tipo de contrato será el más conveniente?"

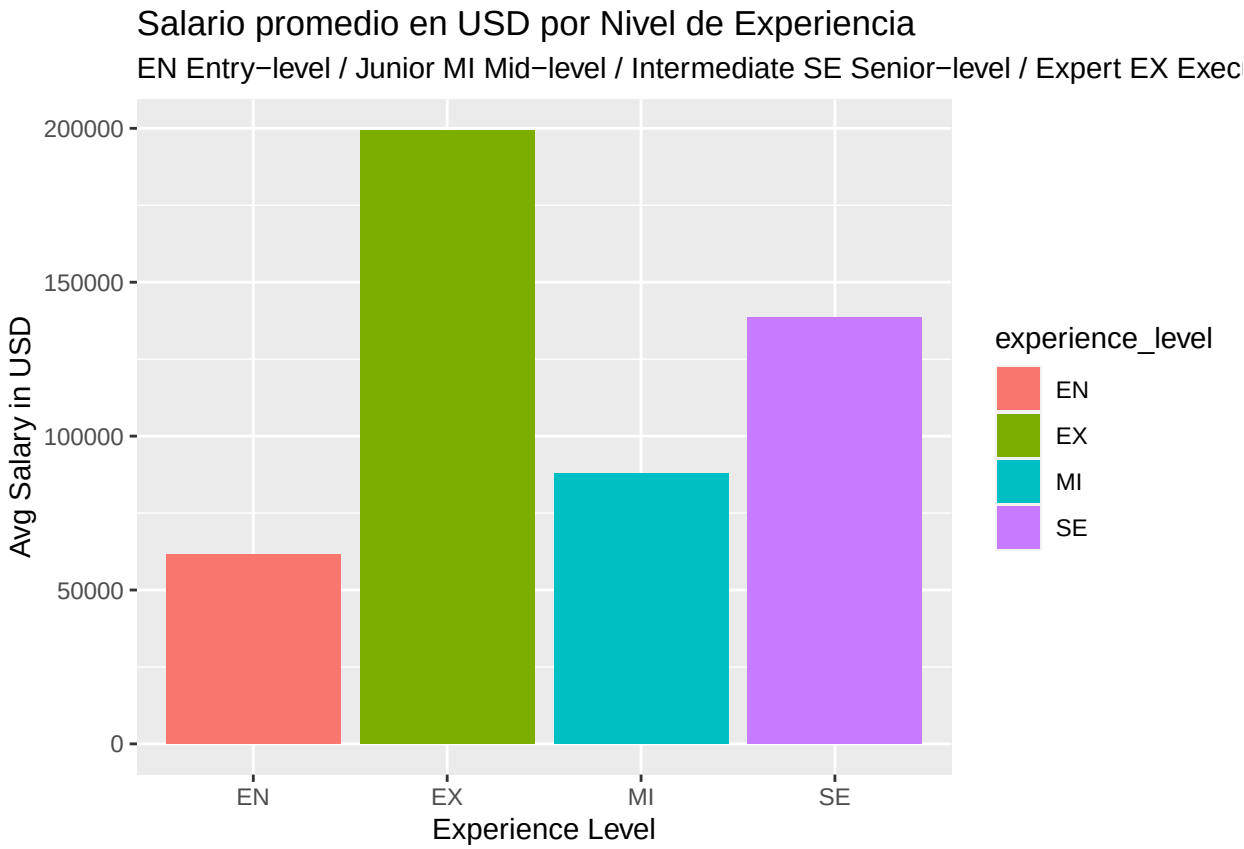
```
ggplot(data=datos_salarios) +
geom_bar(aes(x = employment_type, y = salary_in_usd, fill = employment_type), stat = "summary", fun = "
  labs(title="Salario promedio en USD por Tipo de Contrato", x = "Employment Type", y = "Avg Salary in
```



Respuesta: Como se observa, el tipo de contrato "CT: Contract" ofrece mejor salario, mientras que "PT: Part Time" el más bajo, lo cual tiene sentido debido a la cantidad de horas laborales que diferencian el tipo de contrato.

Procedimiento para responder a "6. ¿Qué tipo de contrato (parcial, tiempo completo, etc) ofrece mejores salarios? ¿Qué tipo de contrato será el más conveniente?"

```
ggplot(data=datos_salarios) +
geom_bar(aes(x = experience_level, y = salary_in_usd, fill = experience_level), stat = "summary", fun =
  labs(title="Salario promedio en USD por Nivel de Experiencia", x = "Experience Level", y = "Avg Salary
```



Respuesta: El nivel de experiencia si influye en el salario promedio que gana los especialistas, siendo Expert EX Executive-level / Director el más alto, superando los 150,000 dólares al año.