Momento de Retroalimentación Módulo 1

Técnicas de procesamiento de datos para el análisis estadístico y para la construcción de modelos

Ariadna Jocelyn Guzmán Jiménez - A01749373

2022-08-28

Lectura de datos

```
db = read.csv("ds_salaries.csv")
salary = db$salary_in_usd
modality = db$remote_ratio
experience = db$experience_level
country = db$company_location
jobType = db$job_title
```

Medidas estadísticas

Variables cuantitativas: Medidas de tendencia central y dispersión

- Salario
- Remote ratio (modalidad de trabajo)

```
library(modeest)

# Medidas de tendencia central

n = length(db$X) #N
sprintf("=========== Medidas de tendencia central ===========")
```

```
## [1] "=========== Medidas de tendencia central ==========="
```

```
sprintf("Numero de datos: %s", n)
```

```
## [1] "Numero de datos: 607"
```

```
meanSalary = mean(salary) #Promedio de salario
sprintf("Promedio de salarios: %s", meanSalary)
```

```
## [1] "Promedio de salarios: 112297.86985173"
```

```r
medianSalary = median(salary) #Mediana de salario
sprintf("Mediana de salarios; %s", medianSalary)
```

```
## [1] "Mediana de salarios; 101570"
```

```r
modeSalary = mlv(salary, method = "mfv")[1] #Moda de salario
sprintf("Moda de salarios: %s", modeSalary)
```

```
## [1] "Moda de salarios: 100000"
```

```r
meanModality = mean(modality) #Promedio de modalidad
sprintf("Promedio de modalidad de trabajo (proporción); %s", meanModality)
```

```
## [1] "Promedio de modalidad de trabajo (proporción); 70.9225700164745"
```

```r
medianModality = median(modality) #Mediana de modalidad
sprintf("Mediana de modalidad; %s", medianModality)
```

```
## [1] "Mediana de modalidad; 100"
```

```r
modeModality = mlv(modality, method = "mfv")[1] #Moda de modalidad
sprintf("Moda de modalidad: %s", modeModality)
```

```
## [1] "Moda de modalidad: 100"
```

```r
# Medidas de dispersion
sprintf("=========== Medidas de dispersión ===========")
```

```
## [1] "=========== Medidas de dispersión ==========="
```

```r
maxSalary = max(salary) # Maximo valor de salario
sprintf("Salario máximo: %s", maxSalary)
```

```
## [1] "Salario máximo: 600000"
```

```r
minSalary = min(salary) # Minimo valor de salario
sprintf("Salario mínimo: %s", minSalary)
```

```
## [1] "Salario mínimo: 2859"
```

```r
deSalary = sd(salary) # Desviacion estandar salario
sprintf("Desviacion estandar del salario %s", deSalary)
```

```
## [1] "Desviacion estandar del salario 70957.2594113957"
```

```r
sSalary = var(salary) # Varianza de salario
sprintf("Varianza de salario: %s", sSalary)
```

```
## [1] "Varianza de salario: 5034932663.1761"
```

```r
maxModality = max(modality) # Maximo valor de modalidad
sprintf("Porcentaje de modalidad máximo: %s", maxModality)
```

```
## [1] "Porcentaje de modalidad máximo: 100"
```

```r
minModality = min(modality) # Minimo valor de modalidad
sprintf("Porcentaje de modalidad mínimo: %s", minModality)
```

```
## [1] "Porcentaje de modalidad mínimo: 0"
```

```r
deModality = sd(modality) # Desviacion estandar modalidad
sprintf("Desviacion estandar del porcentaje de modalidad: %s", deModality)
```

```
## [1] "Desviacion estandar del porcentaje de modalidad: 40.7091300402212"
```

```r
sModality = var(modality) # Varianza de modalidad
sprintf("Varianza de modalidad: %s", sModality)
```

```
## [1] "Varianza de modalidad: 1657.23326863164"
```

Variables cualitativas: tablas de distribución de frecuencia

- Experience level
- Company location
- Job title

```r
experienceTable = table(experience)
cat("Tabla de frecuencia de niveles de experiencia en Data Science\n\n")
```

```
## Tabla de frecuencia de niveles de experiencia en Data Science
```

```r
experienceTable
```

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

```r
countryTable = table(country)
cat("Tabla de frecuencia de países de origen de compañías de Data Science\n\n")
```

```
## Tabla de frecuencia de países de origen de compañías de Data Science
```

```
countryTable
```

```
## country
##  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
```

```
jobTypeTable = table(jobType)
cat("Tabla de frecuencia de tipos de trabajo en Data Science\n\n")
```

```
## Tabla de frecuencia de tipos de trabajo en Data Science
```

```
jobTypeTable
```

```
## jobType
##            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
```

Moda

```
modeExp = mlv(experience, method = "mfv")[1]
sprintf("Moda de experiencia: %s", modeExp)
```

```
## [1] "Moda de experiencia: SE"
```

```
modeCountry = mlv(country, method = "mfv")[1]
sprintf("Moda de país: %s", modeCountry)
```

```
## [1] "Moda de país: US"
```

```
modeJob = mlv(jobType, method = "mfv")[1]
sprintf("Moda de tipo de trabajo: %s", modeJob)
```

```
## [1] "Moda de tipo de trabajo: Data Scientist"
```

Herramientas de visualización

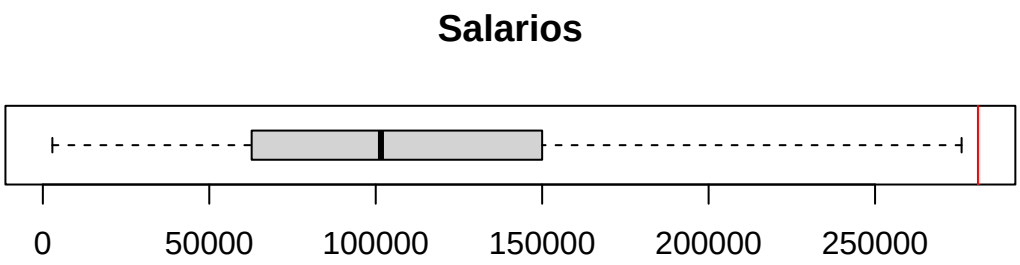
Variables cuantitativas: Medidas de posición y análisis de distribución

Distribución de datos de variables categóricas

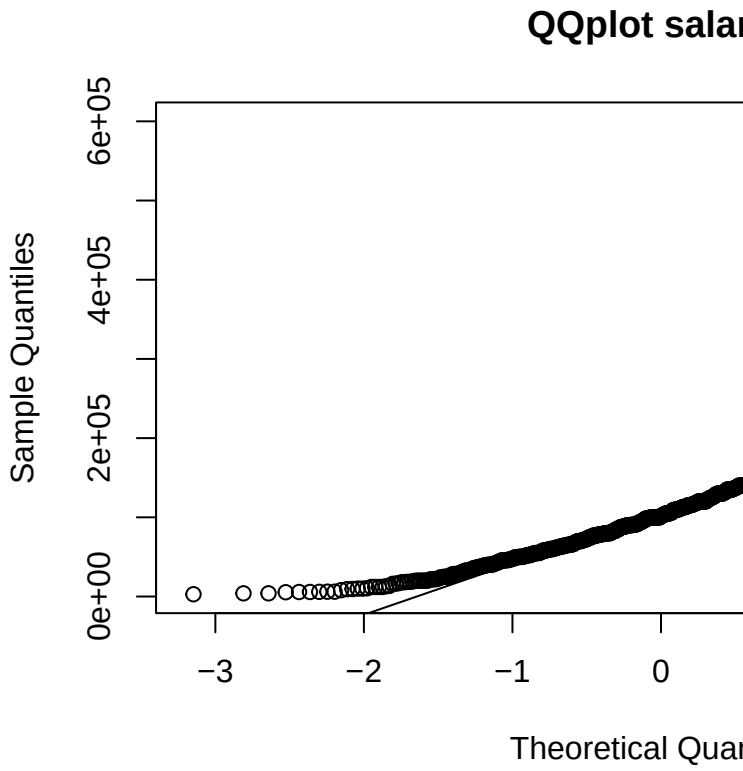
```
#Cuartiles
q1_s = quantile(salary, 0.25)
q3_s = quantile(salary, 0.75)
Rs = q3_s - q1_s
y2= q3_s + 1.5 * Rs

par(mfrow=c(2,1))  #Matriz de gráficos
boxplot(salary,horizontal=TRUE,ylim=c(0,y2) , main= "Salarios")
abline(v=y2,col="red")  #linea vertical en el límite de los datos atípicos
X= db[salary<y2,c("salary_in_usd")] #Quitar datos atípicos de la matriz db en la variable X
summary(X)
```

Eliminación de datos atípicos en salarios

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2859   62649  100000  107169  148261  276000
```

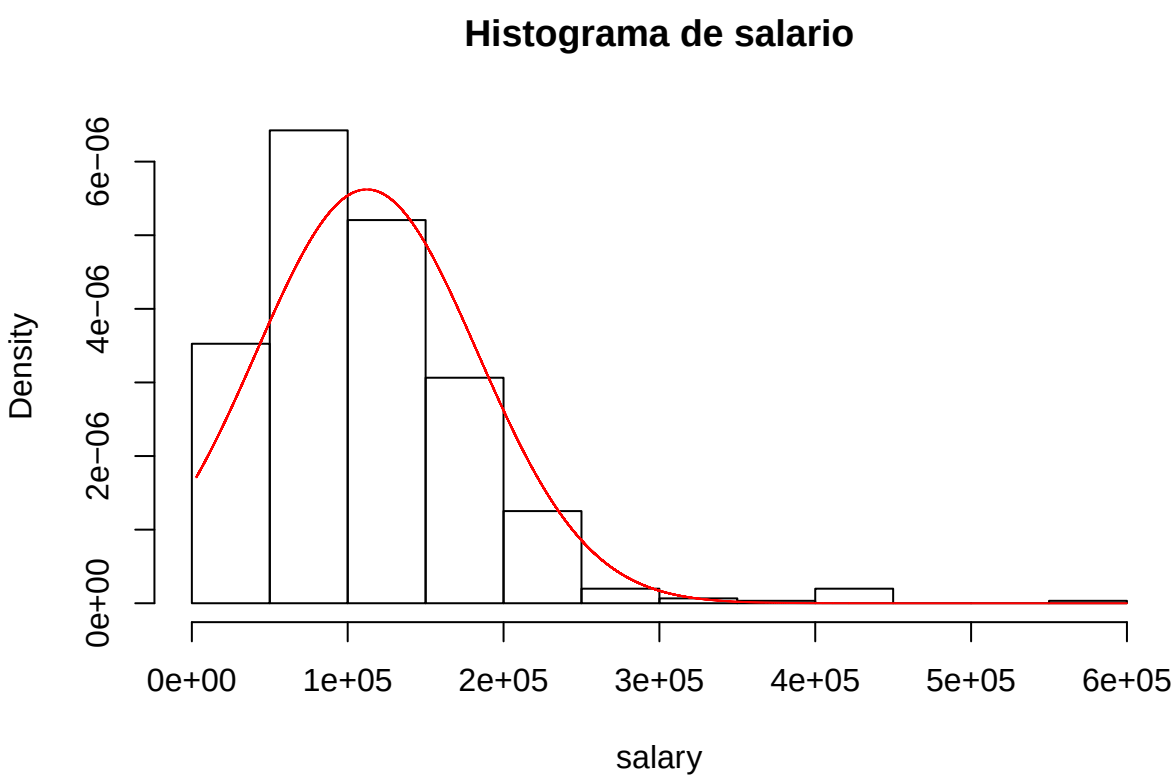



```
qqnorm(salary,  main= " QQplot salarios")
qqline(salary)
```



Exploración de la normalidad de la variable salarios

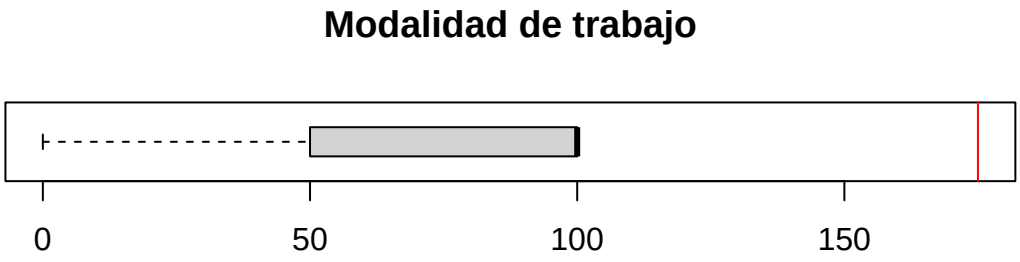
```
hist(salary,prob=TRUE,col=0,  main= "Histograma de salario")
x=seq(min(salary),max(salary),0.1)
y=dnorm(x,mean(salary),sd(salary))
lines(x,y,col="red")
```



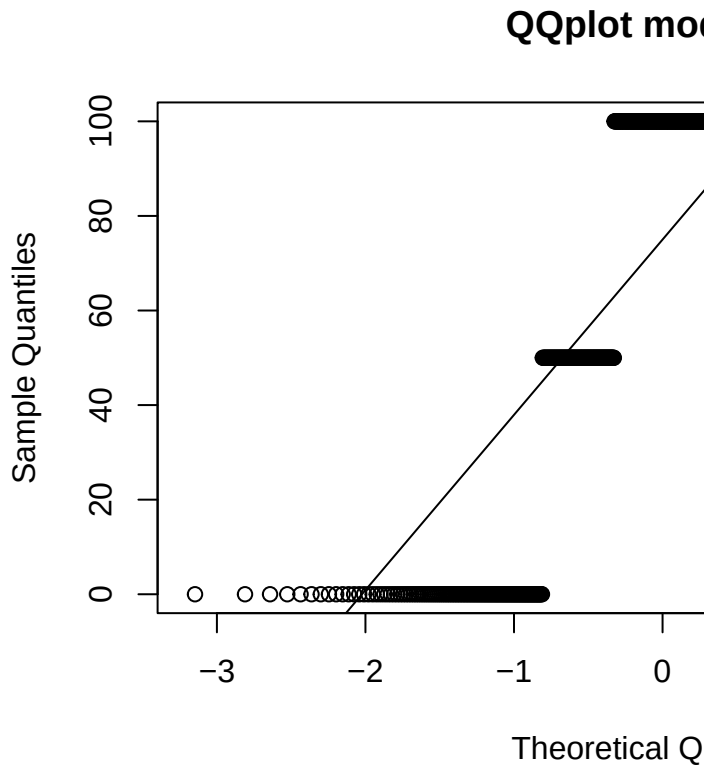

```
#Cuartiles
q1_m = quantile(modality, 0.25)
q3_m = quantile(modality, 0.75)
Rm = q3_m - q1_m
y2= q3_m + 1.5 * Rm

par(mfrow=c(2,1))  #Matriz de gráficos
boxplot(modality,horizontal=TRUE,ylim=c(0,y2) , main= "Modalidad de trabajo")
abline(v=y2,col="red")  #linea vertical en el límite de los datos atípicos
X= db[modality<y2,c("remote_ratio")] #Quitar datos atípicos de la matriz db en la variable X
summary(X)
```

Eliminación de datos atípicos en modalidad

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

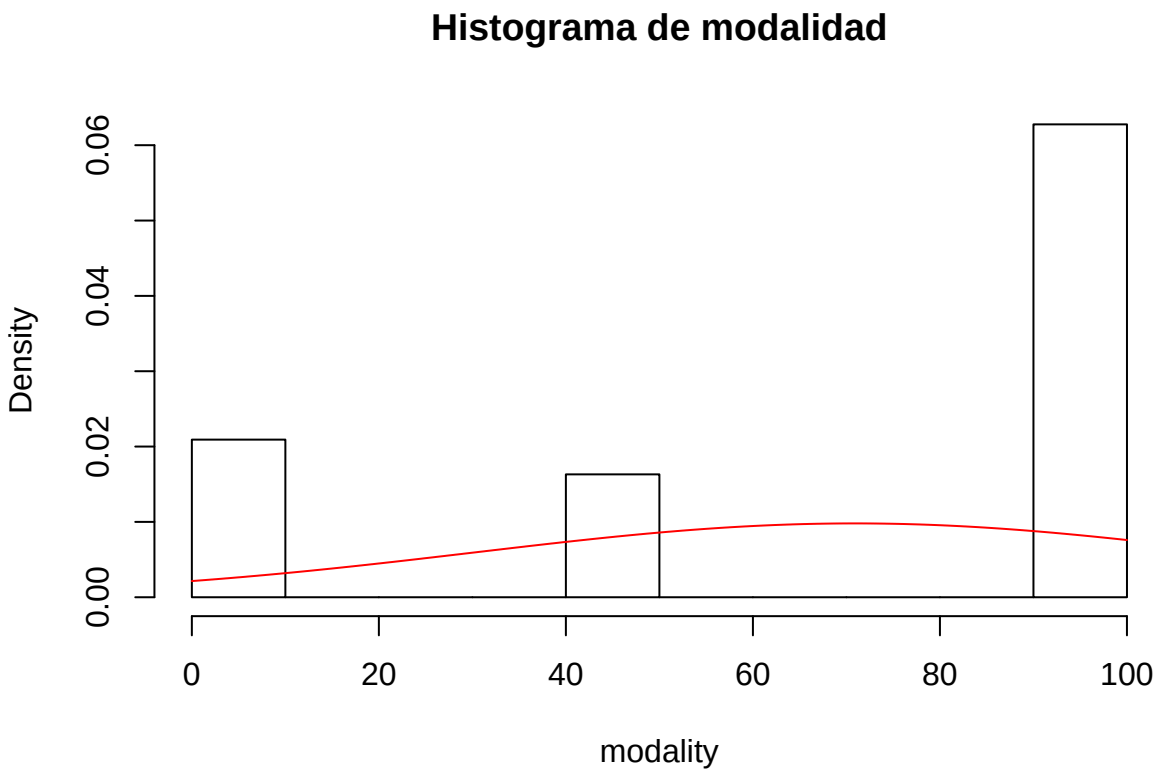



```
qqnorm(modality,  main= " QQplot modalidad")
qqline(modality)
```



Exploración de la normalidad de la variable modalidad

```
hist(modality,prob=TRUE,col=0,  main= "Histograma de modalidad")
x=seq(min(modality),max(modality),0.1)
y=dnorm(x,mean(modality),sd(modality))
lines(x,y,col="red")
```




```
library(moments)
```

Curtosis y sesgo de salarios

```
##
## Attaching package: 'moments'

## The following object is masked from 'package:modeest':
##
##     skewness
```

```
skewness(salary)
```

```
## [1] 1.663421
```

```
kurtosis(salary)
```

```
## [1] 9.291709
```

```
skewness(modality)
```

Curtosis y sesgo de modalidad

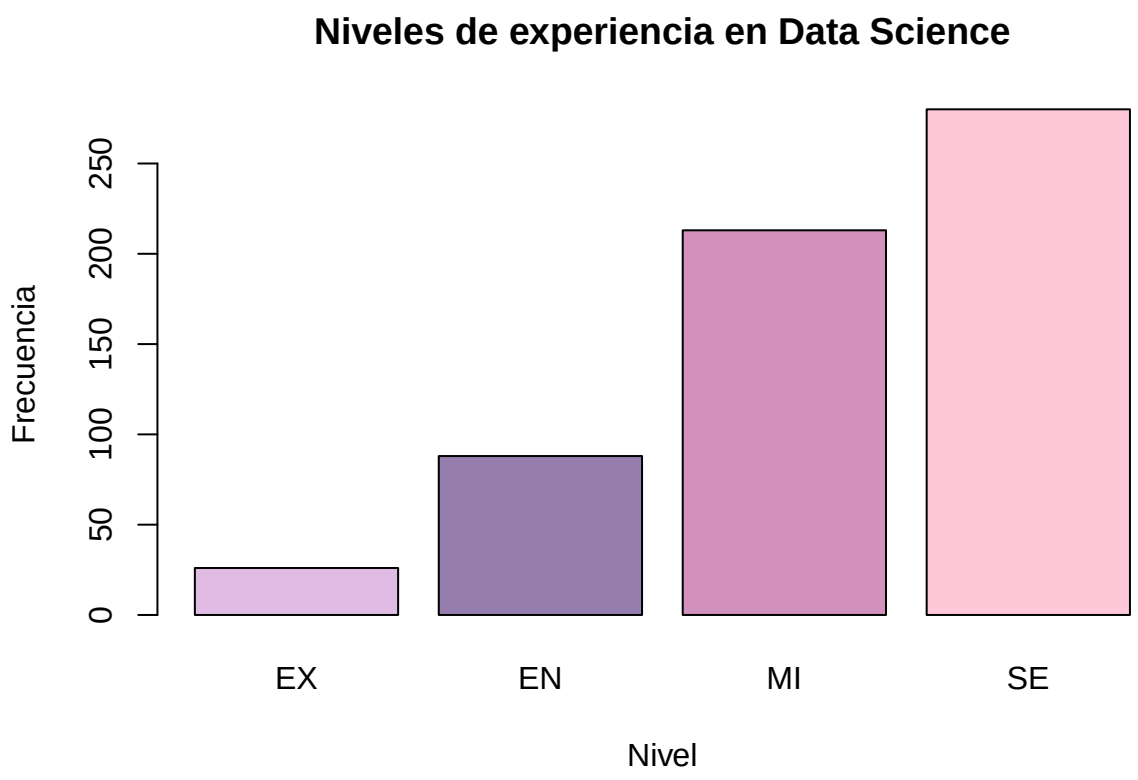
```
## [1] -0.9019881
```

```
kurtosis(modality)
```

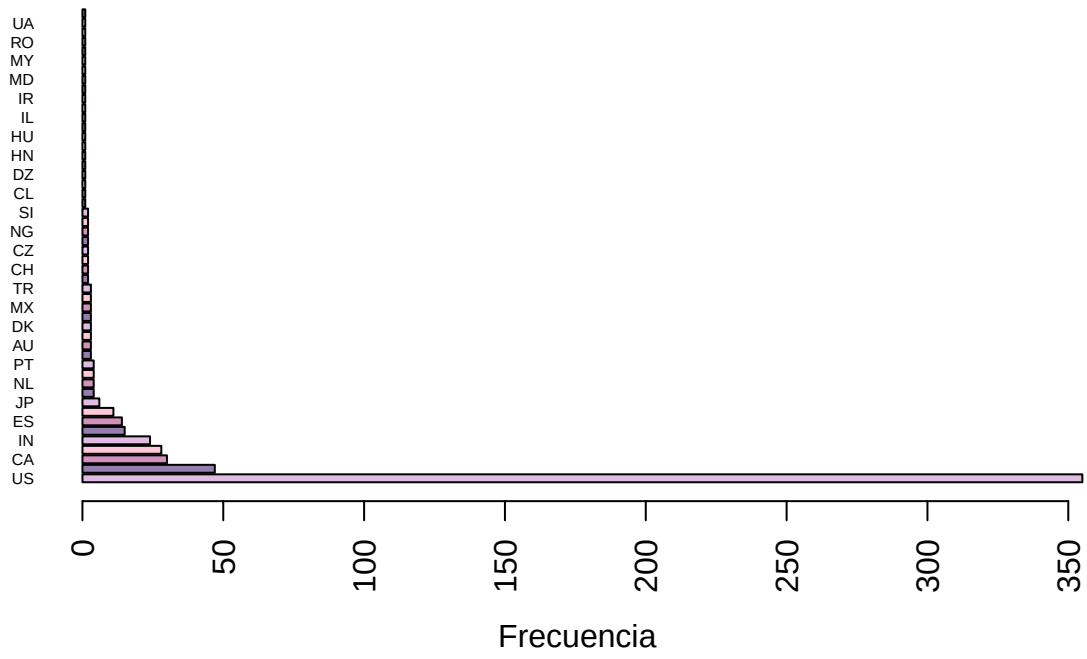
```
## [1] 2.109162
```

Variables categóricas: Distribución de los datos

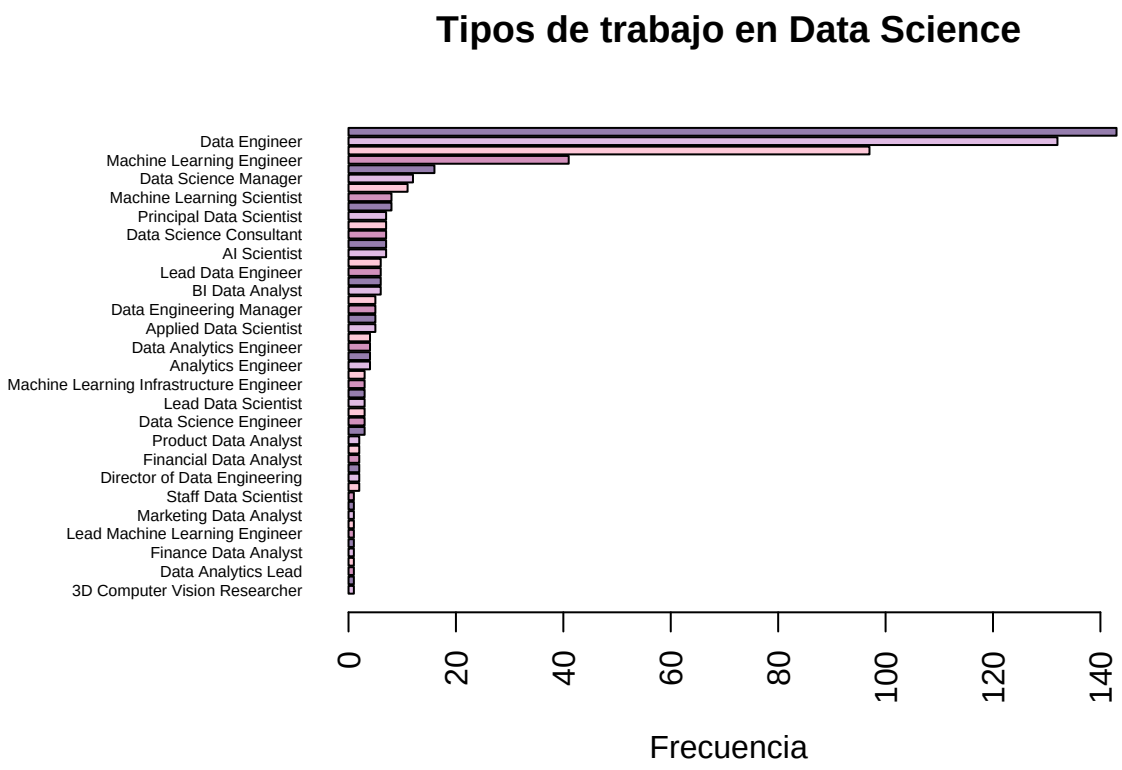
```
sorted_tableEx = sort(experienceTable)
barplot(sorted_tableEx, col= c("#E0BBE4", "#957DAD" , "#D291BC" , "#FEC8D8" ), main = "Niveles de experi
```




```
sorted_tableC = sort(countryTable, decreasing = TRUE)
barplot(sorted_tableC, col= c("#E0BBE4", "#957DAD" , "#D291BC" , "#FEC8D8" ), main = "Paises de origen
```




```
sorted_tableJT = sort(jobTypeTable)
par(mar=c(5.1,10,4.1,2.1)+.1)
barplot(sorted_tableJT, col= c("#E0BBE4", "#957DAD" , "#D291BC" , "#FEC8D8" ), main = "Tipos de trabajo
```



Preguntas

```
data_analyst_salary = db[db$job_title == "Data Analyst", ]
mean_salary_da = mean
sprintf("Promedio esperado de salario de un Data Analyst")
```

¿Cuánto puede esperar de salario un Data Analyst?

```
## [1] "Promedio esperado de salario de un Data Analyst"
```

```
mean_salary_da
```

```
## function (x, ...)
## UseMethod("mean")
## <bytecode: 0x000001d41c954110>
## <environment: namespace:base>
```

```
sorted_db = db[order(db$salary_in_usd, decreasing = TRUE), ]
top_sorted_db = head(sorted_db, 100)
```

```
top_countries = top_sorted_db$company_location
top = head(unique(top_countries), 5)
df = data.frame(top_sorted_db$company_location, top_sorted_db$salary_in_usd)
print("Top 5 Países con Mejores Salarios: ")
```

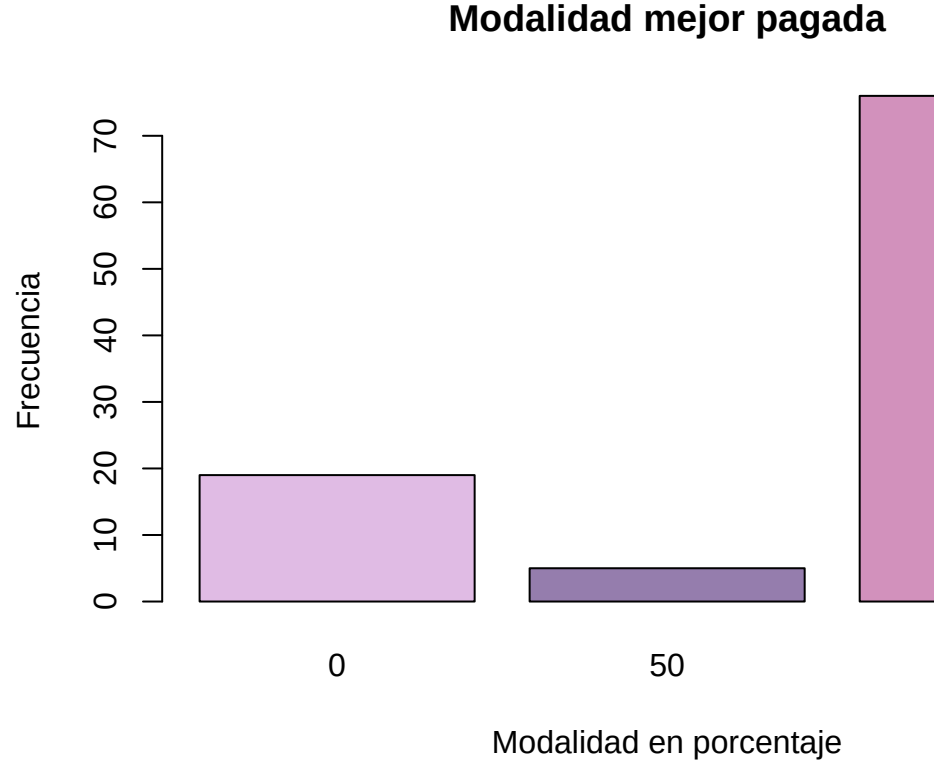
¿Cuál es el top 5 de países con mejores salarios?

```
## [1] "Top 5 Países con Mejores Salarios: "
```

```
top
```

```
## [1] "US" "JP" "RU" "CA" "GB"
```

```
top_modality = top_sorted_db$remote_ratio
top = head(unique(top_modality), 3)
barplot(table(top_modality), col= c("#E0BBE4", "#957DAD" , "#D291BC" , "#FEC8D8" ), xlab = "Modalidad e
```



¿Cuál es la modalidad mejor pagada?

```
sprintf("Modalidad de trabajo mejor pagada en porcentaje: %s", top[1])
```

```
## [1] "Modalidad de trabajo mejor pagada en porcentaje: 100"
```

```
cat("*0 = no hay trabajo remoto\n*50 = parcialmente remota\n*100 = totalmente remota")
```

```
## *0 = no hay trabajo remoto
## *50 = parcialmente remota
## *100 = totalmente remota
```