Momento Retroalimentación: Módulo 1 Técnicas de procesamiento de datos para el análisis estadístico y para la construcción de modelos

Amy Murakami Tsutsumi - A01750185

2022-08-22

Contestar dos preguntas e inventar una

Leer datos

```
db=read.csv("ds_salaries.csv") #leer la base de datos
salary = db$salary_in_usd   #para llamar una variable
remoteRatio = db$remote_ratio   #para llamar una variable
```

Exploración de la base de datos

1. Calcula medidas estadísticas

Variables cuantitativas

Las variables cuantitativas que se utilizarán son el salario en dólares y el porcentaje de trabajo remoto ya que son las variables más significativas. Se excluyeron las variables del año en que se paga el salario y el salario en su respectivo tipo de cambio ya que son valores que no aportan datos necesarios para las preguntas objetivo.

```
n = length(db$X) #N
sprintf("Número de datos: %s", n)
```

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

```
## [1] "Número de datos: 607"
```

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

```
## [1] "Promedio salario en dólares: 112297.86985173"
```

```r
meanRemoteRatio = mean(remoteRatio) #Promedio de modalidad
sprintf("Promedio de modalidad: %s", meanRemoteRatio)
```

```
## [1] "Promedio de modalidad: 70.9225700164745"
```

```r
medianSalary = median(salary) #Mediana de salario
sprintf("Mediana del salario en dólares: %s", medianSalary)
```

```
## [1] "Mediana del salario en dólares: 101570"
```

```r
medianRemoteRatio = median(remoteRatio) #Mediana de modalidad
sprintf("Mediana de modalidad: %s", medianRemoteRatio)
```

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

```r
library(modeest)
modeSalary = mlv(salary, method = "mfv")[1] #Moda de salario
sprintf("Moda del salario en dólares: %s", modeSalary)
```

```
## [1] "Moda del salario en dólares: 100000"
```

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

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

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

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

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

```r
maxRemoteRatio = max(remoteRatio) # Maximo valor de modalidad
sprintf("Modalidad máxima: %s", maxRemoteRatio)
```

```
## [1] "Modalidad máxima: 100"
```

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

```
## [1] "Salario mínimo: 2859"
```

```r
minRemoteRatio = min(remoteRatio) # Minimo valor de modalidad
sprintf("Modalidad mínima: %s", minRemoteRatio)
```

```
## [1] "Modalidad mínima: 0"
```

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

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

```r
varRemoteRatio = var(remoteRatio) # Varianza de modalidad
sprintf("Varianza de la modalidad: %s", varRemoteRatio)
```

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

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

```
## [1] "Desviación estándar del salario: 70957.2594113957"
```

```r
deRemoteRatio = sd(remoteRatio) # Desviacion estandar modalidad
sprintf("Desviación estándar de modalidad: %s", deRemoteRatio)
```

```
## [1] "Desviación estándar de modalidad: 40.7091300402212"
```

Variables cualitativas y Variables categóricas

Las variables cualitativas y categóricas que se utilizarán son el nivel de experiencia, el nombre de empleo, el tipo de empleo, la ubicación de la compañía y el tamaño de la compañía. Únicamente se excluyó el tipo de cambio ya que es muy similar a los datos de ubicación de la compañía.

```r
print("Tabla de distribución de frecuencia del nivel de experiencia:")
```

* Tabla de distribución de frecuencia, moda y distribución de los datos (diagramas de barras, diagramas de pastel)

```
## [1] "Tabla de distribución de frecuencia del nivel de experiencia:"
```

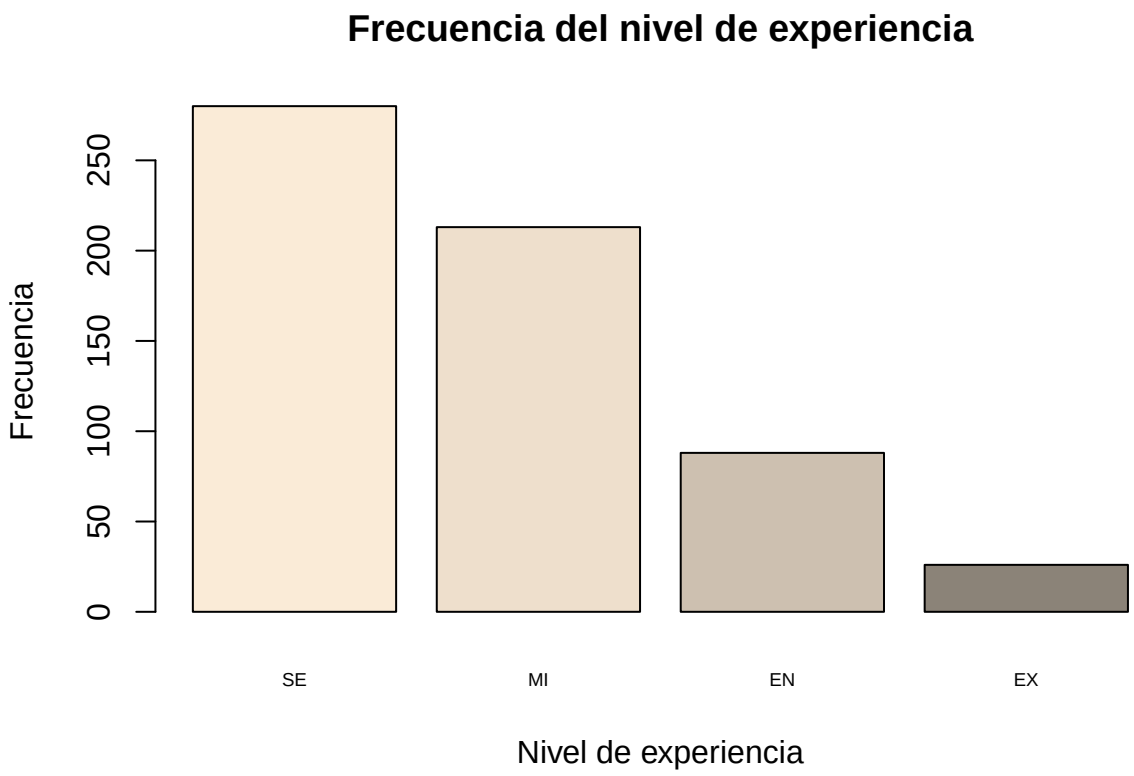
```r
experience = db$experience_level
experience_table = table(experience)
print(experience_table)
```

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

```r
modeExperience = mlv(experience, method = "mfv")[1] #Moda
sprintf("Moda del nivel de experiencia: %s", modeExperience)
```

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

```r
sorted_table = sort(experience_table, decreasing = TRUE)[1:4]
barplot(sorted_table, width = 1, cex.names = 0.6, xlab="Nivel de experiencia", ylab="Frecuencia", col =
```




```r
print("Tabla de distribución de frecuencia del tipo de empleo:")
```

```
## [1] "Tabla de distribución de frecuencia del tipo de empleo:"
```

```r
employment_type = db$employment_type
employment_type_table = table(employment_type)
print(employment_type_table)
```

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

```r
modeEmployment = mlv(employment_type, method = "mfv")[1] #Moda
sprintf("Moda del tipo de empleo: %s", modeEmployment)
```

```
## [1] "Moda del tipo de empleo: FT"
```

```
sorted_table = sort(employment_type_table, decreasing = TRUE)[1:4]
barplot(sorted_table, width = 1, cex.names = 0.6, xlab="Tipo de empleo", ylab="Frecuencia",col = c("aqua
```


```
print("Tabla de distribución de frecuencia del empleo:")
```

```
## [1] "Tabla de distribución de frecuencia del empleo:"
```

```
job_title = db$job_title
job_title_table = table(job_title)
print(job_title_table)
```

```
## job_title
##            3D Computer Vision Researcher
##                                        1
##                              AI Scientist
##                                        7
##                         Analytics Engineer
##                                        4
##                     Applied Data Scientist
##                                        5
##       Applied Machine Learning Scientist
##                                        4
```

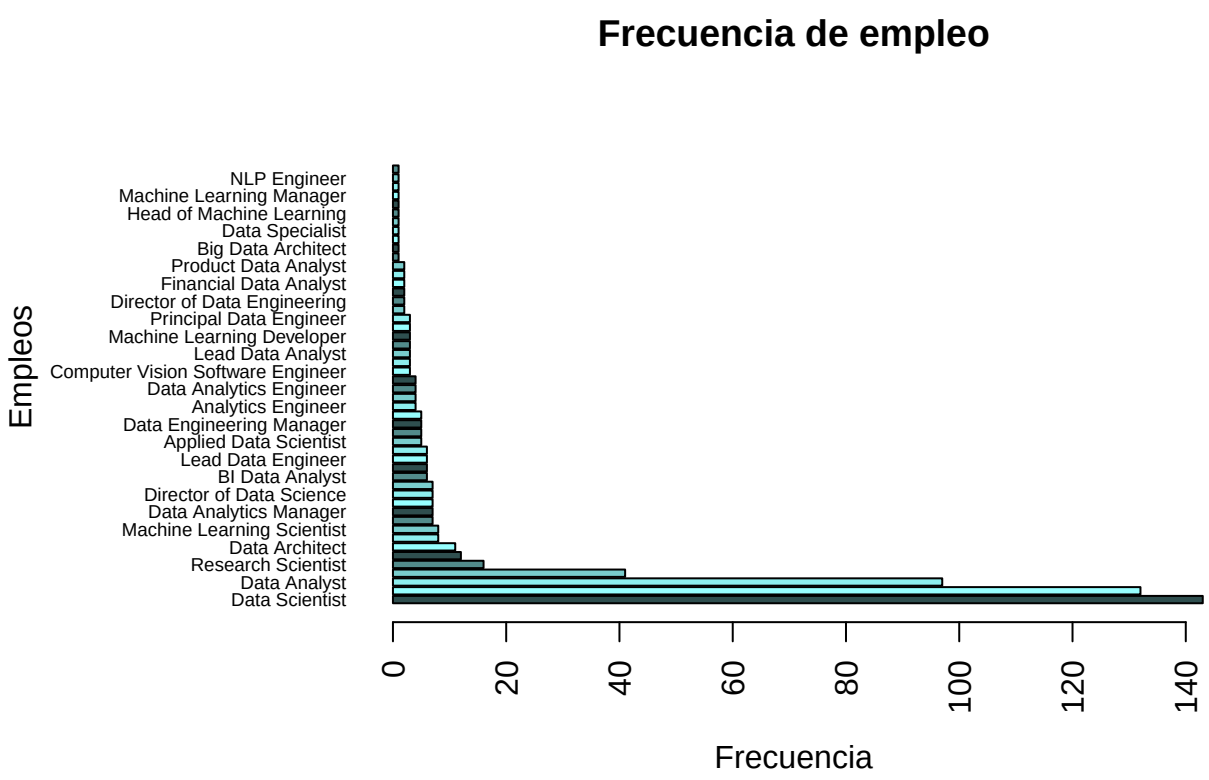
```
##                         BI Data Analyst
##                                       6
##                      Big Data Architect
##                                       1
##                       Big Data Engineer
##                                       8
##                    Business Data Analyst
##                                       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
```

```
modeJobTitle = mlv(job_title, method = "mfv")[1] #Moda
sprintf("Moda del empleo: %s", modeJobTitle)
```

```
## [1] "Moda del empleo: Data Scientist"
```

```
sorted_table = sort(job_title_table, decreasing = TRUE)[1:54]
par(mar=c(5,10,4,1)+.1)
barplot(sorted_table, width = 1, cex.names = 0.6, col = c("darkslategray", "darkslategray1", "darkslateg
title(xlab = "Frecuencia", line = 3)
title(ylab = "Empleos", line = 9)
```




```
print("Tabla de distribución de frecuencia de la ubicación de la compañía:")
```

```
## [1] "Tabla de distribución de frecuencia de la ubicación de la compañía:"
```

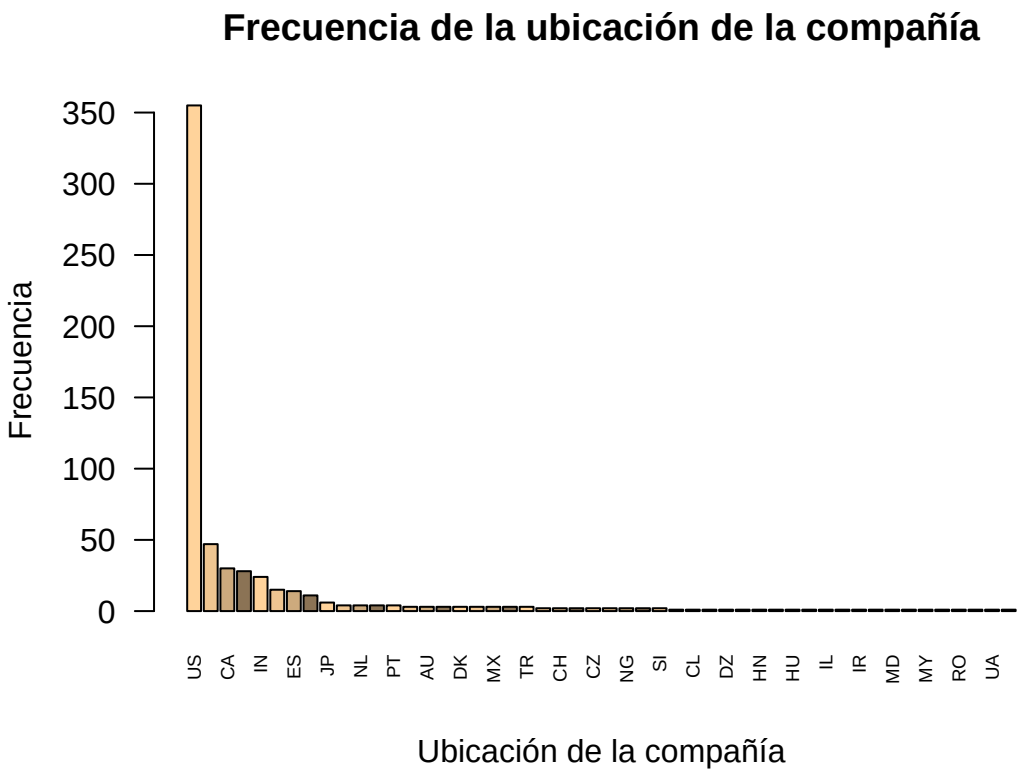
```
company_location = db$company_location
company_location_table = table(company_location)
print(company_location_table)
```

```
## 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
```

```
modeCompLoc = mlv(company_location, method = "mfv")[1] #Moda
sprintf("Moda de la ubicación del empleo: %s", modeCompLoc)
```

```
## [1] "Moda de la ubicación del empleo: US"
```

```
sorted_table = sort(company_location_table, decreasing = TRUE)[1:50]
par(mar=c(5,8,4,1)+.1)
barplot(sorted_table, width = 1, cex.names = 0.6, col = c("burlywood1", "burlywood2", "burlywood3", "bu
```




```
print("Tabla de distribución de frecuencia del tamaño de la compañía:")
```

```
## [1] "Tabla de distribución de frecuencia del tamaño de la compañía:"
```

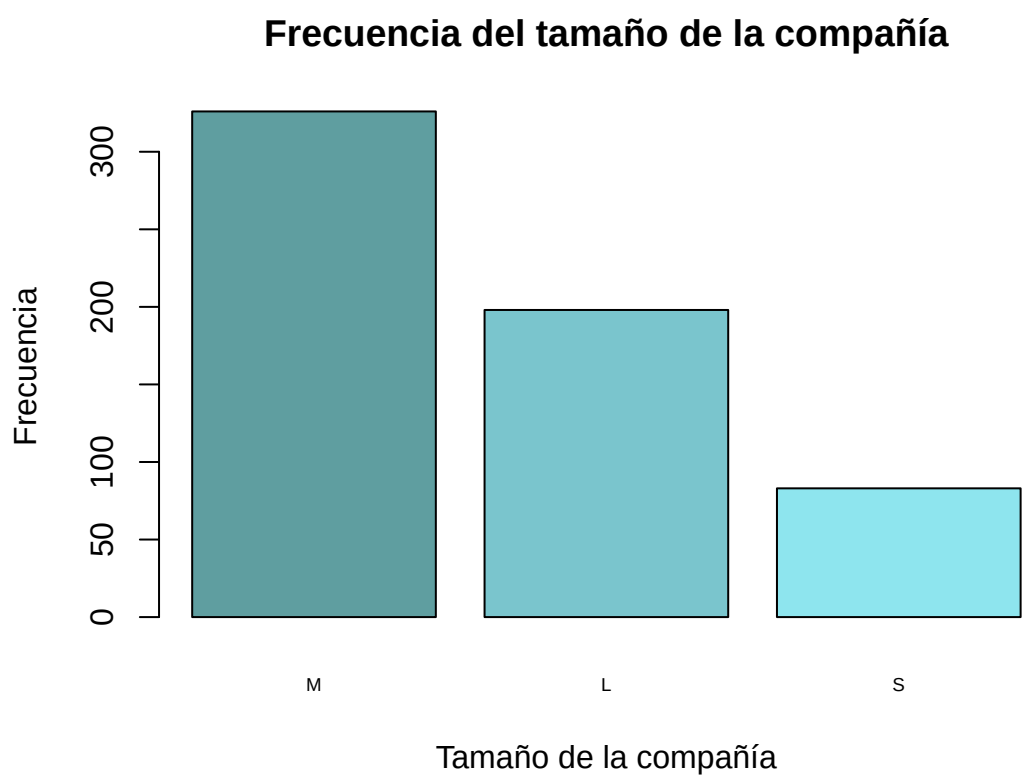
```
company_size = db$company_size
company_size_table = table(company_size)
print(company_size_table)
```

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

```
modeCompSize = mlv(company_size, method = "mfv")[1] #Moda
sprintf("Moda de la ubicación del empleo: %s", modeCompSize)
```

```
## [1] "Moda de la ubicación del empleo: M"
```

```
sorted_table = sort(company_size_table, decreasing = TRUE)[1:3]
par(mar=c(5,8,4,1)+.1)
barplot(sorted_table, width = 1, cex.names = 0.6, col = c("cadetblue", "cadetblue3", "cadetblue2"), mai
```



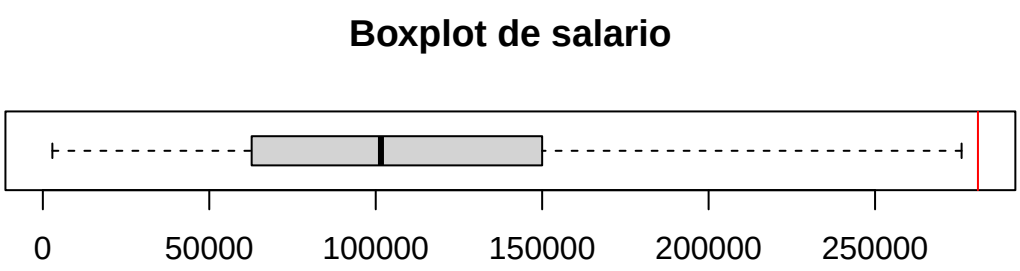
Explora los datos usando herramientas de visualización ## Variables cuantitativas: ### Medidas de posición: cuartiles, outlier (valores atípicos), boxplots

```
print("Cuartiles de salario")
```

```
## [1] "Cuartiles de salario"
```

```
q1_c=quantile(salary,0.25)  #Cuantil 1
q3_c = quantile(salary, 0.75) #Cuantil 3
ri_c= IQR(salary) #Rango intercuartílico
y2 = q3_c+1.5*ri_c
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(salary,horizontal=TRUE,ylim=c(0,y2),main="Boxplot de salario")
abline(v=q3_c+1.5*ri_c,col="red")  #linea vertical en el límite de los datos atípicos
X = db[salary<q3_c+1.5*ri_c,c("salary_in_usd")] #Quitar datos atípicos de la matriz M en la variable X
summary(X)
```

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

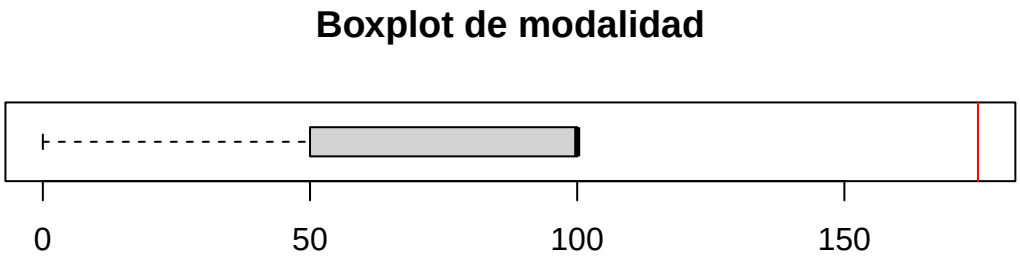



```
print("Cuartiles de modalidad")
```

```
## [1] "Cuartiles de modalidad"
```

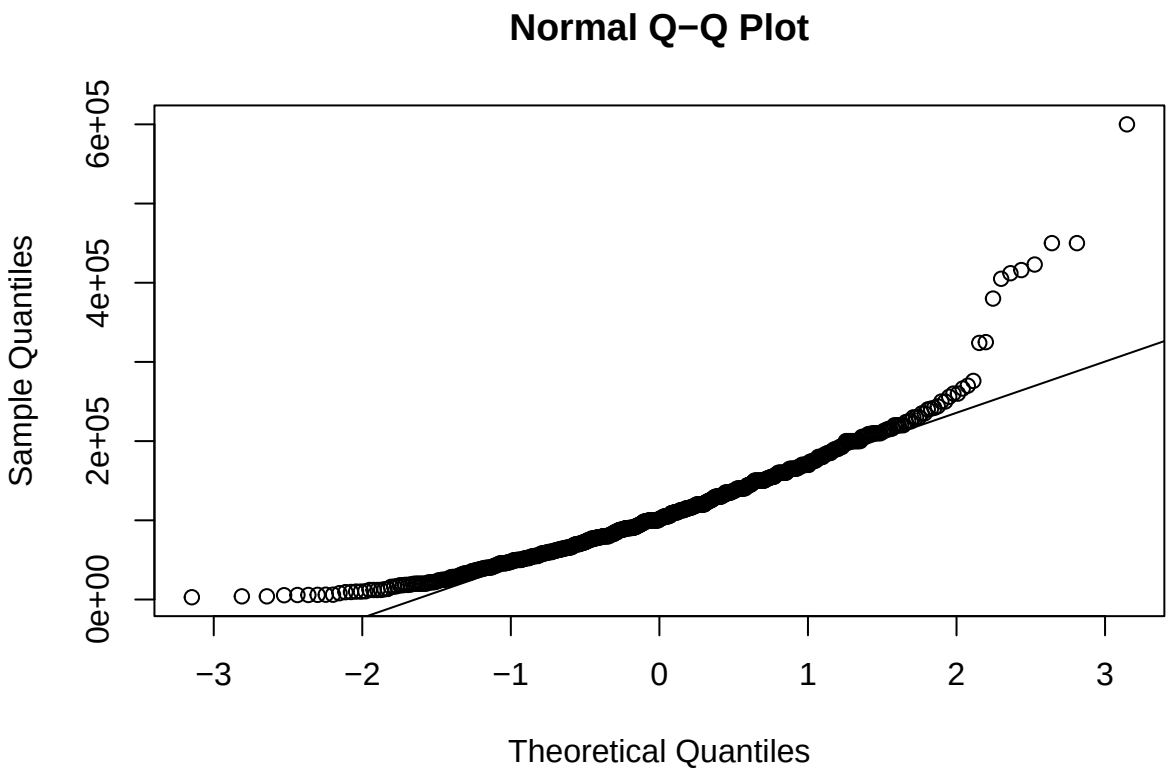
```
q1_c=quantile(remoteRatio,0.25)  #Cuantil 1
q3_c = quantile(remoteRatio, 0.75) #Cuantil 3
ri_c= IQR(remoteRatio) #Rango intercuartílico
y2 = q3_c+1.5*ri_c
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(remoteRatio,horizontal=TRUE,ylim=c(0,y2), main="Boxplot de modalidad")
abline(v=q3_c+1.5*ri_c,col="red")  #linea vertical en el límite de los datos atípicos
X = db[remoteRatio<q3_c+1.5*ri_c,c("remote_ratio")] #Quitar datos atípicos de la matriz M en la variabl
summary(X)
```

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

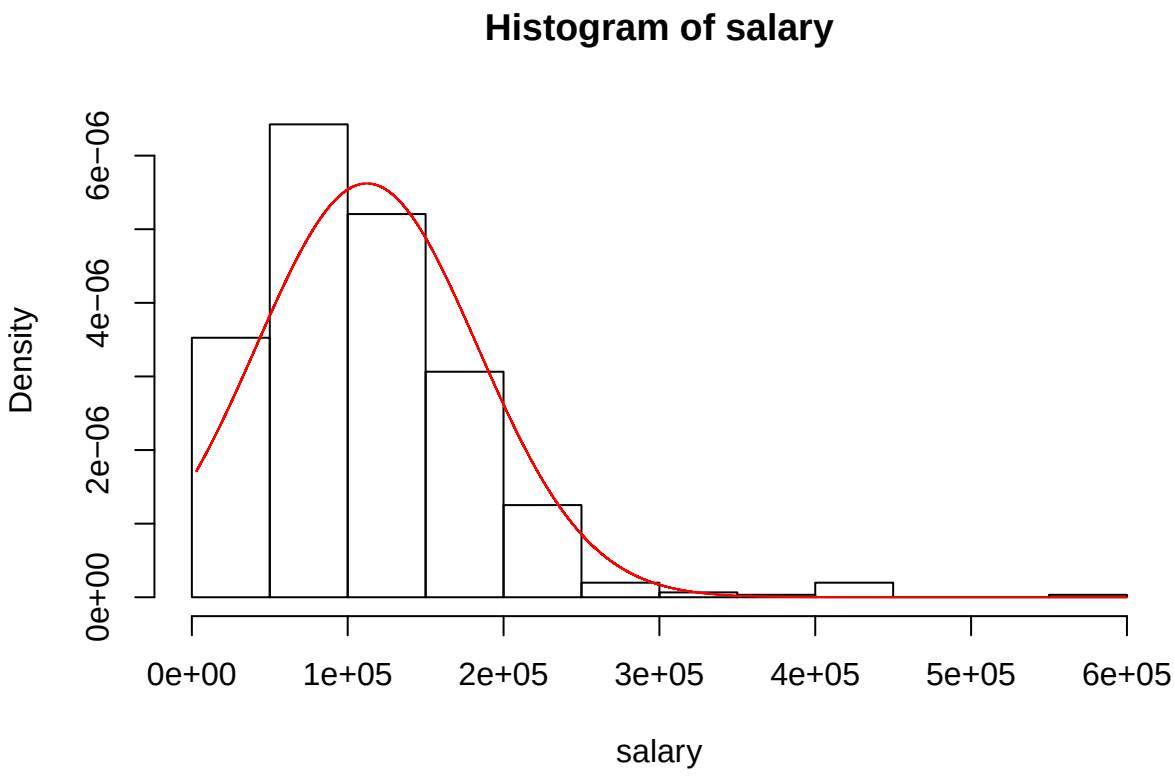


Análisis de distribución de los datos (Histogramas). Identificar si tiene forma simétrica o asimétrica

```
qqnorm(salary)
qqline(salary)
```




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




```
library(moments)
```

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

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

```
skewness(salary)
```

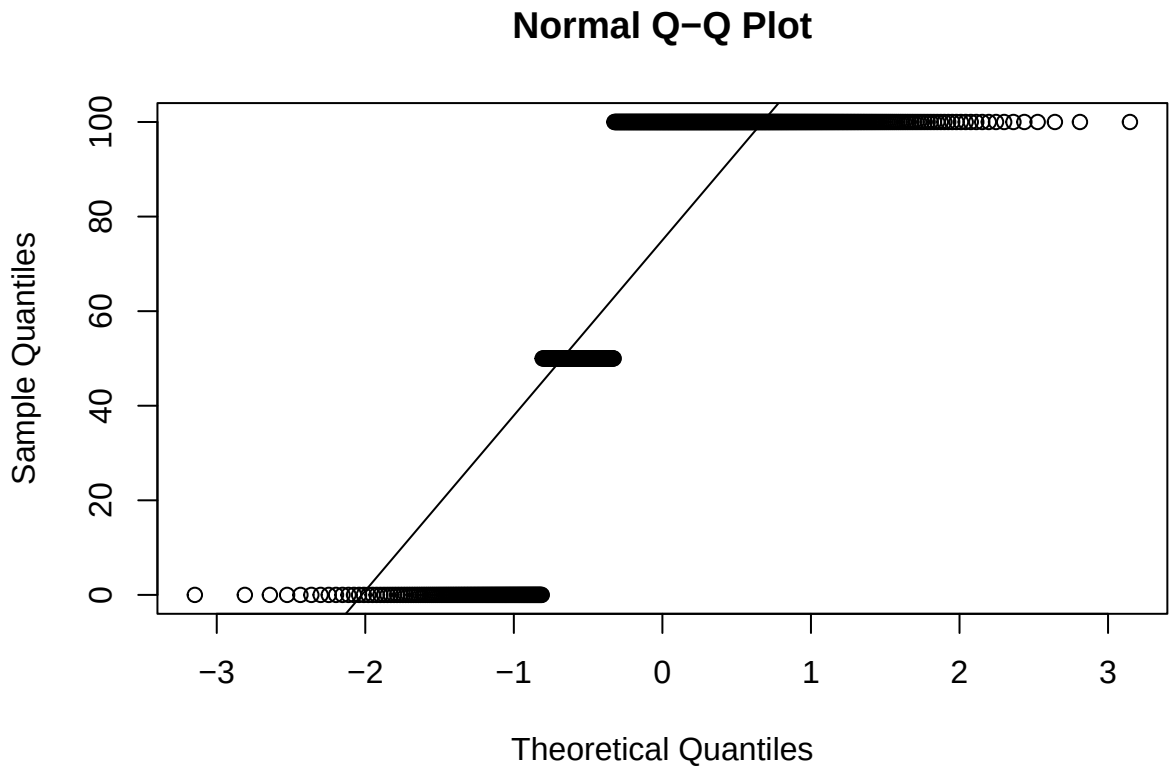
```
## [1] 1.663421
```

```
kurtosis(salary)
```

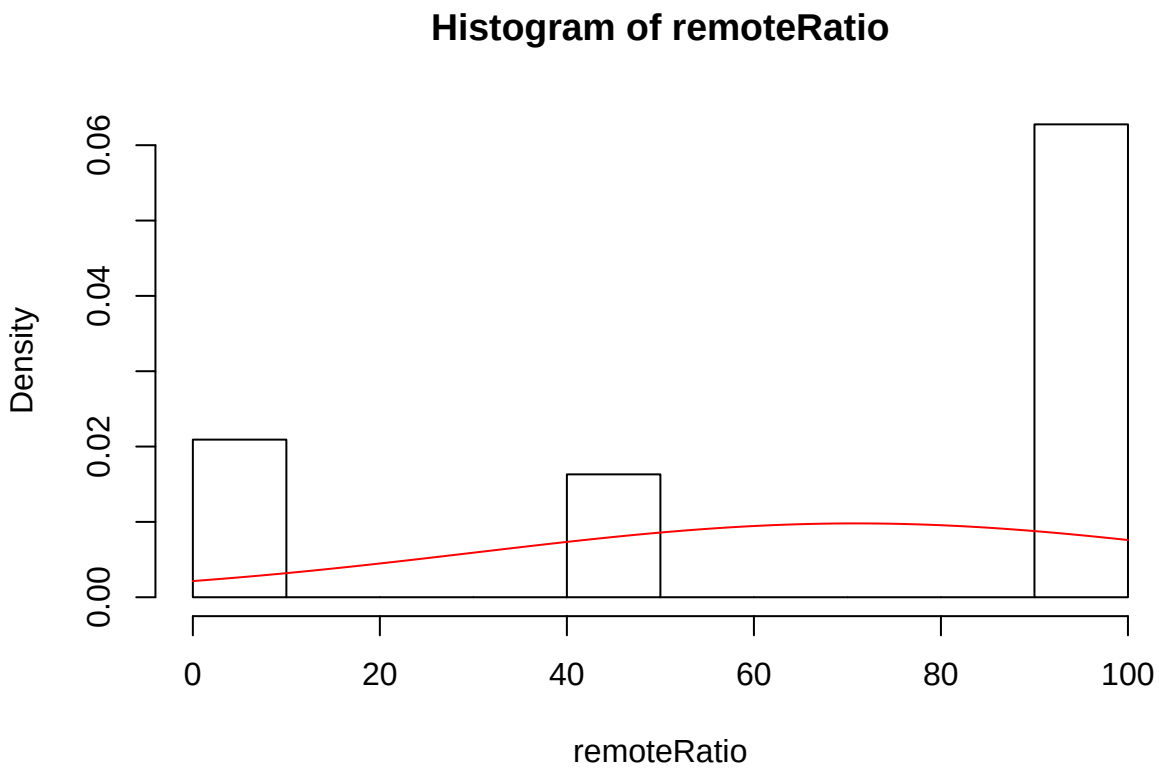
```
## [1] 9.291709
```

La gráfica anterior tiene una asimetría postiva, es decir, tiene un sesgo a la derecha. Además, el valor del coeficiente de sesgo al ser un valor mayor a uno significa que esta muy sesgada a la derecha. Incluso el valor de la curtosis, al ser un número mayor a 3, indica que es leptocúrtica.

```
qqnorm(remoteRatio)
qqline(remoteRatio)
```




```
hist(remoteRatio,prob=TRUE,col=0)
x=seq(min(remoteRatio),max(remoteRatio),0.1)
y=dnorm(x,mean(remoteRatio),sd(remoteRatio))
lines(x,y,col="red")
```




```
library(moments)
skewness(remoteRatio)
```

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

```
kurtosis(remoteRatio)
```

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

La gráfica anterior es asimétrica. Además, el valor del coeficiente de sesgo al ser un valor menor a uno significa que esta muy sesgada a la izquierda. Incluso el valor de la curtosis, al ser un número menor a 3, indica que es platicúrtica.

Analizar los datos y contestar las preguntas guía

¿Influye el tamaño de la compañía en el salario que puede ofrecer a un analista de datos?

```
data_analyst= db[db$job_title == "Data Analyst",]
data_small = data_analyst[data_analyst$company_size == "S",]
small = mean(data_small$salary_in_us)
sprintf("Promedio del salario en empresa pequeña: %s", small)
```

```
## [1] "Promedio del salario en empresa pequeña: 47759"
```

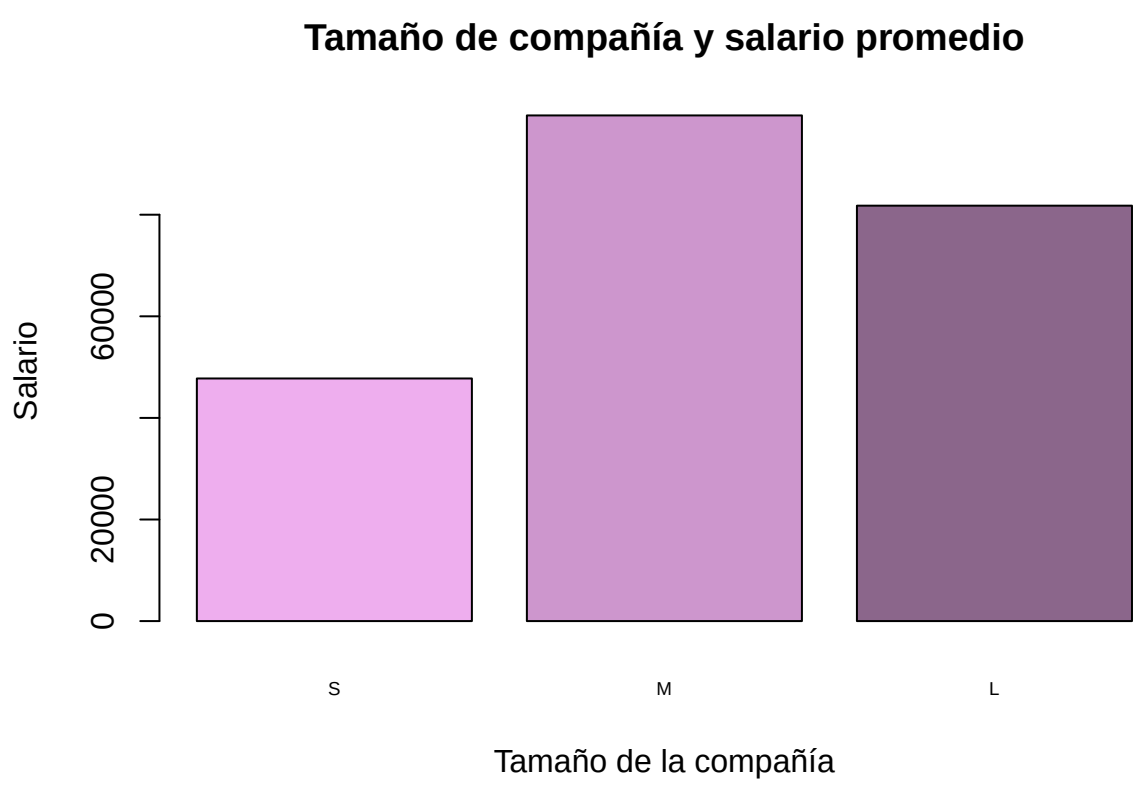
```
data_medium = data_analyst[data_analyst$company_size == "M",]
medium = mean(data_medium$salary_in_us)
sprintf("Promedio del salario en empresa mediana: %s", medium)
```

```
## [1] "Promedio del salario en empresa mediana: 99545.3421052632"
```

```
data_large = data_analyst[data_analyst$company_size == "L",]
large = mean(data_large$salary_in_us)
sprintf("Promedio del salario en empresa grande: %s", large)
```

```
## [1] "Promedio del salario en empresa grande: 81777.6153846154"
```

```
datos_tamaño = c("S"=small, "M"=medium, "L"=large)
barplot(datos_tamaño, width = 1, cex.names = 0.6, col = c("plum2", "plum3", "plum4"), main="Tamaño de c
```



Al observar los datos podemos notar que no es de tanta importancia el tamaño de la compañía al considerar los mejores salarios, ya que en promedio los mejores salarios se encuentran en empresas medianas.

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

```r
data_analyst_salary = db[db$job_title == "Data Analyst", ]
mean_salary_da = mean(data_analyst_salary$salary_in_usd)
sprintf("Promedio del salario de analista de datos: %s", mean_salary_da)
```

```
## [1] "Promedio del salario de analista de datos: 92893.0618556701"
```

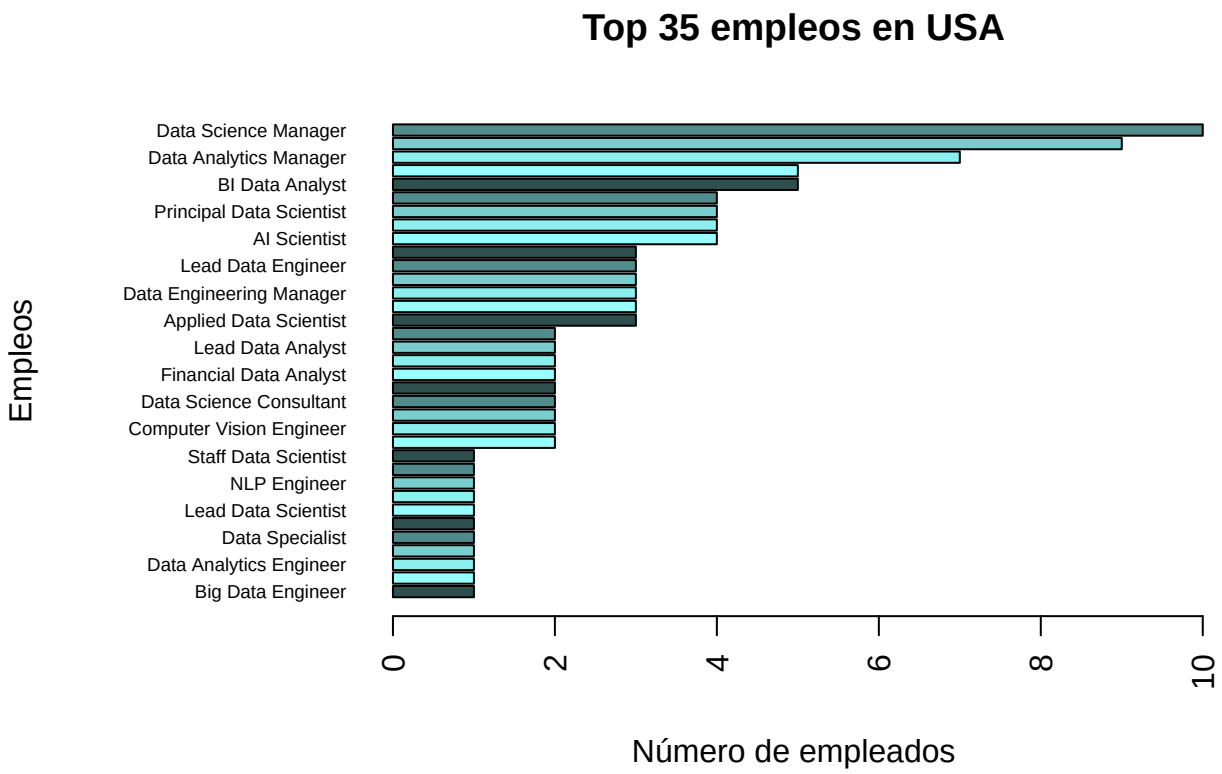
Top 35 empleos en USA

```r
data_usa = db[db$company_location == "US", ]
data_job_usa = data_usa$job_title
data_job_usa_table = table(data_job_usa)
print(data_job_usa_table)
```

```
## data_job_usa
##                              AI Scientist
##                                         4
##                       Analytics Engineer
##                                         4
##                    Applied Data Scientist
##                                         3
##       Applied Machine Learning Scientist
##                                         3
##                           BI Data Analyst
##                                         5
##                         Big Data Engineer
##                                         1
##                      Business Data Analyst
##                                         2
##                        Cloud Data Engineer
##                                         1
##                  Computer Vision Engineer
##                                         2
##        Computer Vision Software Engineer
##                                         2
##                               Data Analyst
##                                        71
##                    Data Analytics Engineer
##                                         1
##                        Data Analytics Lead
##                                         1
##                     Data Analytics Manager
##                                         7
##                             Data Architect
##                                         9
##                              Data Engineer
##                                        85
##                  Data Engineering Manager
##                                         3
##                   Data Science Consultant
##                                         2
```

```
##                     Data Science Manager
##                                       10
##                           Data Scientist
##                                       84
##                          Data Specialist
##                                        1
##              Director of Data Engineering
##                                        1
##                  Director of Data Science
##                                        2
##                    Financial Data Analyst
##                                        2
##                               Head of Data
##                                        2
##                     Head of Data Science
##                                        3
##                        Lead Data Analyst
##                                        2
##                       Lead Data Engineer
##                                        3
##                      Lead Data Scientist
##                                        1
##                  Machine Learning Engineer
##                                       16
## Machine Learning Infrastructure Engineer
##                                        1
##                 Machine Learning Scientist
##                                        5
##                                ML Engineer
##                                        2
##                               NLP Engineer
##                                        1
##                     Principal Data Analyst
##                                        1
##                    Principal Data Engineer
##                                        3
##                   Principal Data Scientist
##                                        4
##                         Research Scientist
##                                        4
##                     Staff Data Scientist
##                                        1
```

```
sorted_table = sort(data_job_usa_table, decreasing = FALSE)[1:35]
par(mar=c(5,10,4,1)+.1)
barplot(sorted_table, width = 1, cex.names = 0.6, col = c("darkslategray", "darkslategray1", "darkslateg
title(xlab = "Número de empleados", line = 3)
title(ylab = "Empleos", line = 9)
```



Podemos notar que los empleos más populares en Estados Unidos son: Data Science Manager, Data Analytics Manager y BI Data Analyst.

Modalidad con mayor salario

```
sorted_db = db[order(db$salary_in_usd, decreasing = TRUE), ]
top_sorted_db = head(sorted_db, 1)
top_modality = top_sorted_db$remote_ratio
print("La Modalidad que Cuenta con un Mayor Salario es: ")
```

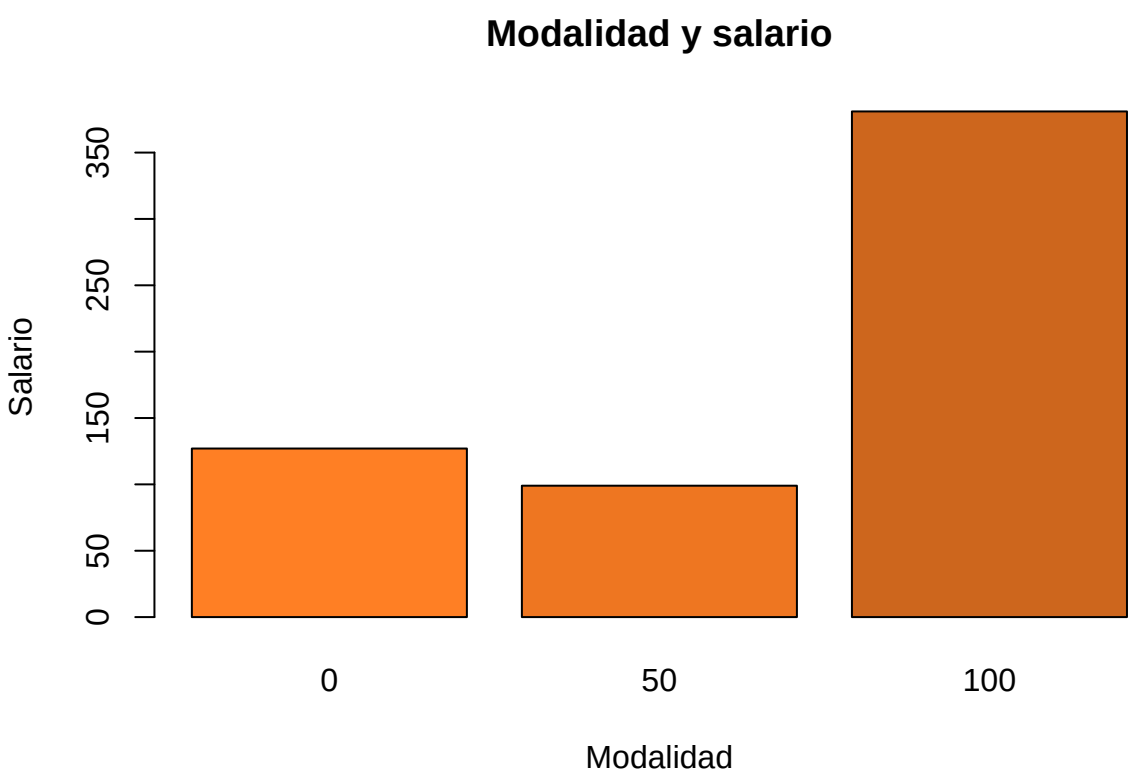
```
## [1] "La Modalidad que Cuenta con un Mayor Salario es: "
```

```
if (top_modality == 100) {
  print("Modalidad en línea")
  top_modality
}
```

```
## [1] "Modalidad en línea"
```

```
## [1] 100
```

```
barplot(table(sorted_db$remote_ratio), col = c("chocolate1", "chocolate2", "chocolate3"), main="Modalida
```



Podemos observar que la modalidad que tiene un mejor salario es la que es 100% remota.