# Momento de Retroalimentación (Portafolio Análisis)

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

Jorge Chávez Badillo A01749448

2022-08-28

## Preguntas Guía

1. ¿Cuál es el salario al que pueda aspirar un analista de datos?
2. ¿En qué países se ofrecen mejores salarios?
3. ¿Se han incrementado los salarios a lo largo del tiempo?
4. ¿Influye el nivel de experiencia en el salario?
5. ¿Influye el tamaño de la compañía en el salario que puede ofrecer a un analista de datos?
6. ¿Qué tipo de contrato (parcial, tiempo completo, etc) ofrece mejores salarios? ¿Qué tipo de contrato será el más conveniente?
7. Otras más que creas que se pueden contestar a partir de la base de datos.

## Exploración de la Base de Datos

### Acceder al la base de datos Data Science Job Salaries

```
# Lectura de la base de datos
db_salaries = read.csv("ds_salaries.csv")
```

### Exploración de variables

```
n_variables = length(db_salaries)
n_rows = nrow(db_salaries)

salaries_usd = db_salaries$salary_in_usd
job_title = db_salaries$job_title
company_size = db_salaries$company_size
experience_level = db_salaries$experience_level
job_modality = db_salaries$remote_ratio
country = db_salaries$company_location

sprintf("Número de Variables: %s", n_variables)
```

```
## [1] "Número de Variables: 12"
```

```r
sprintf("Número de Registros: %s", n_rows)
```

```
## [1] "Número de Registros: 607"
```

### Exploración de la base de datos

### 1. Medidas Estadísticas

a. Variables Cuantitativas:

- Medidas de Tendencia Central

```r
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Salarios
mean_salaries_usd = mean(salaries_usd)
median_salaries_usd = median(salaries_usd)
mode_salaries_usd = getmode(salaries_usd)
sprintf("Promedio Salarios: %s", mean_salaries_usd)
```

```
## [1] "Promedio Salarios: 112297.86985173"
```

```r
sprintf("Mediana Salarios: %s", median_salaries_usd)
```

```
## [1] "Mediana Salarios: 101570"
```

```r
sprintf("Moda Salarios: %s", mode_salaries_usd)
```

```
## [1] "Moda Salarios: 100000"
```

```r
# Modalidad de Trabajo
mean_job_modality = mean(job_modality)
median_job_modality = median(job_modality)
mode_job_modality = getmode(job_modality)
sprintf("Promedio Modalidad: %s", mean_job_modality)
```

```
## [1] "Promedio Modalidad: 70.9225700164745"
```

```r
sprintf("Mediana Modalidad: %s", median_job_modality)
```

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

```r
sprintf("Moda Modalidad: %s", mode_job_modality)
```

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

- Medidas de Dispersión

```r
# Salarios
max_salaries_usd = max(salaries_usd)
min_salaries_usd = min(salaries_usd)
sd_salaries_usd = sd(salaries_usd)
var_salaries_usd = var(salaries_usd)
sprintf("Máximo Salarios: %s", max_salaries_usd)
```

```
## [1] "Máximo Salarios: 600000"
```

```r
sprintf("Mínimo Salarios: %s", min_salaries_usd)
```

```
## [1] "Mínimo Salarios: 2859"
```

```r
sprintf("Desviación Estándar Salarios: %s", sd_salaries_usd)
```

```
## [1] "Desviación Estándar Salarios: 70957.2594113957"
```

```r
sprintf("Varianza Salarios: %s", var_salaries_usd)
```

```
## [1] "Varianza Salarios: 5034932663.1761"
```

```r
# Modalidad de Trabajo
max_job_modality = max(job_modality)
min_job_modality = min(job_modality)
sd_job_modality = sd(job_modality)
var_job_modality = var(job_modality)
sprintf("Máximo Modalidad: %s", max_job_modality)
```

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

```r
sprintf("Mínimo Modalidad: %s", min_job_modality)
```

```
## [1] "Mínimo Modalidad: 0"
```

```r
sprintf("Desviación Estándar Modalidad: %s", sd_job_modality)
```

```
## [1] "Desviación Estándar Modalidad: 40.7091300402212"
```

```r
sprintf("Varianza Modalidad: %s", var_job_modality)
```

```
## [1] "Varianza Modalidad: 1657.23326863164"
```

b. Variables Cualitativas:

```r
job_title_table = table(job_title)
print("Tabla de Distribución de Frecuencia de los Puestos de Trabajo: ")
```

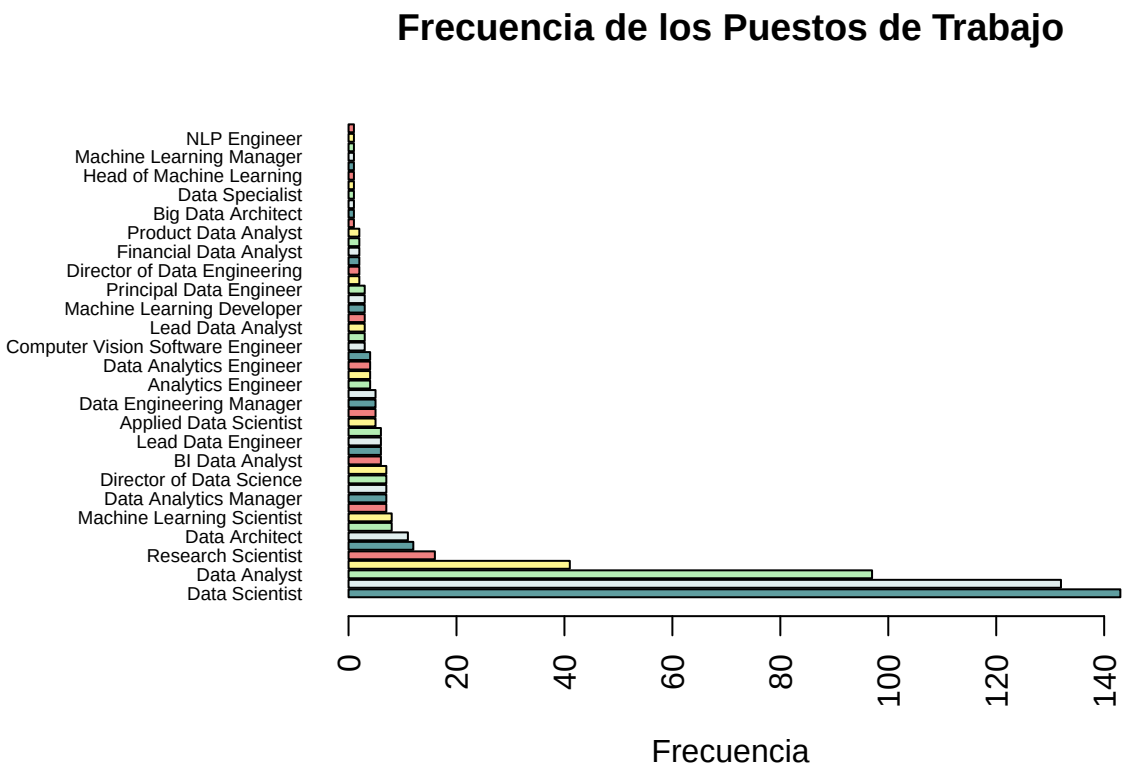
```
## [1] "Tabla de Distribución de Frecuencia de los Puestos de Trabajo: "
```

```
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
```

```
sorted_table = sort(job_title_table, decreasing = TRUE)
par(mar = c(5.1, 10.1, 4.1, 2.1))
barplot(sorted_table, width = 1, cex.names = 0.6, col = c("cadetblue", "azure2", "darkseagreen2", "khaki
```



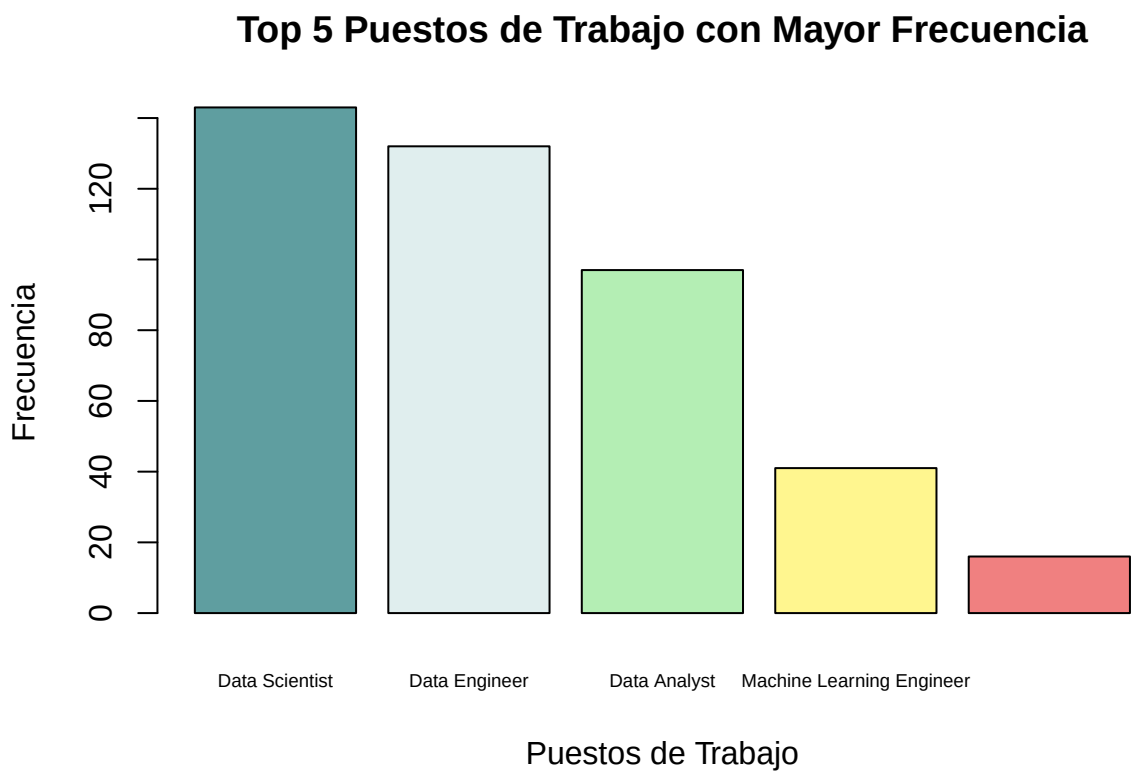

```
job_title_table = table(job_title)
print("Tabla de Distribución de Frecuencia de los 5 Puestos de Trabajo: ")
```

```
## [1] "Tabla de Distribución de Frecuencia de los 5 Puestos de Trabajo: "
```

```
sorted_table = sort(job_title_table, decreasing = TRUE)[1:5]
sorted_table
```

```
## job_title
##            Data Scientist            Data Engineer             Data Analyst
##                       143                      132                       97
## Machine Learning Engineer       Research Scientist
##                        41                       16
```

```
barplot(sorted_table, width = 1, cex.names = 0.6, col = c("cadetblue", "azure2", "darkseagreen2", "khak
```




```
mode_job_title = sort(job_title_table, decreasing = TRUE)[1:1]
print("Moda de los Puestos de Trabajo")
```

```
## [1] "Moda de los Puestos de Trabajo"
```

```
mode_job_title
```

```
## Data Scientist
##            143
```

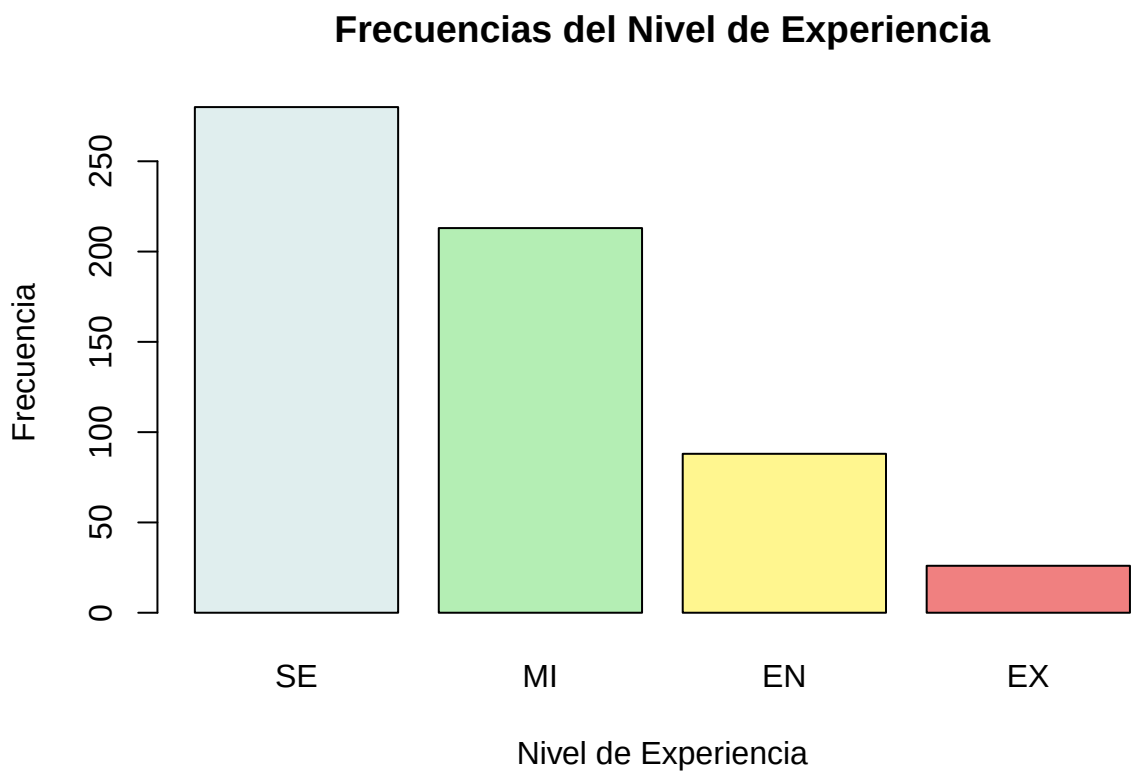
```
experience_level_table = table(experience_level)
print("Tabla de Distribución de Frecuencia del Nivel de Experiencia: ")
```

```
## [1] "Tabla de Distribución de Frecuencia del Nivel de Experiencia: "
```

```
experience_level_table
```

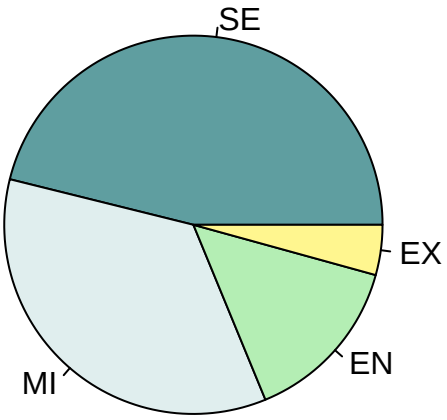
```
## experience_level
##  EN  EX  MI  SE
##  88  26 213 280
```

```r
sorted_table = sort(experience_level_table, decreasing = TRUE)
barplot(sorted_table, width = 1, cex.names = 1, col = c("azure2", "darkseagreen2", "khaki1", "lightcoral
```




```r
pie(sorted_table, main = "Gráfico de Pastel sobre el Nivel de Experiencia", col = c("cadetblue", "azure1
```

## Gráfico de Pastel sobre el Nivel de Experiencia




```
mode_experience_level = sort(experience_level, decreasing = TRUE)[1:1]
print("Moda del Nivel de Experiencia")
```

```
## [1] "Moda del Nivel de Experiencia"
```

```
mode_experience_level
```

```
## [1] "SE"
```

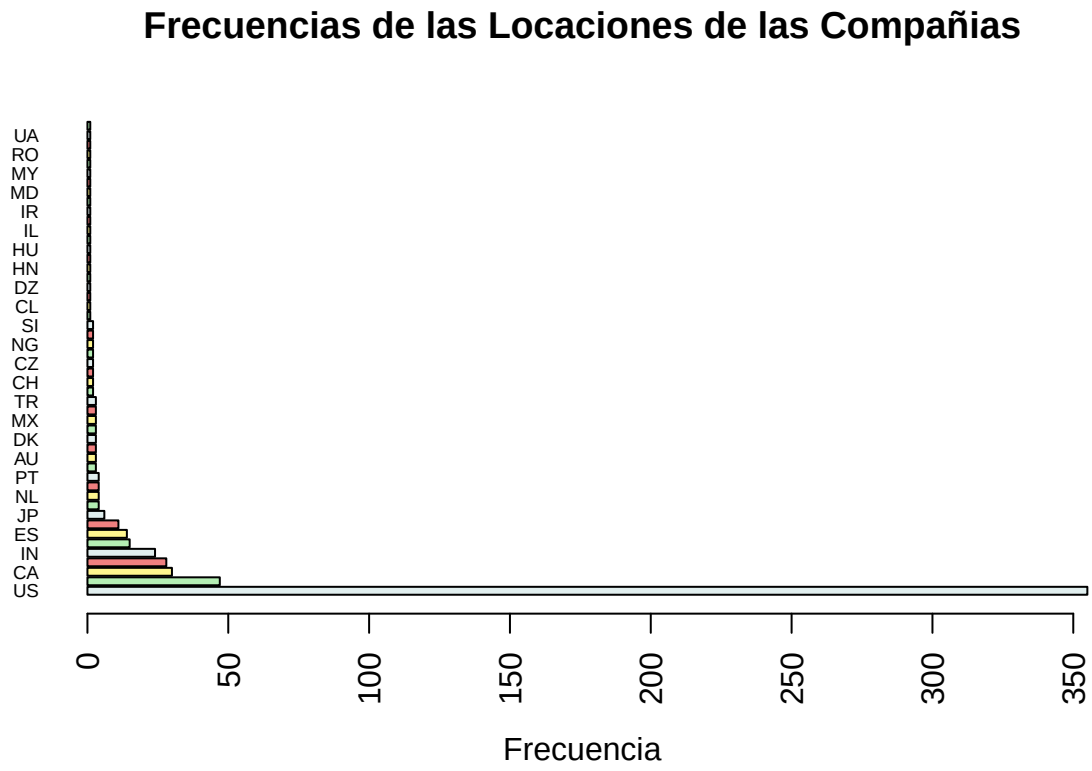
```
country_table = table(country)
print("Tabla de Distribución de Frecuencia del Paises: ")
```

```
## [1] "Tabla de Distribución de Frecuencia del Paises: "
```

```
country_table
```

```
## 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
```

```r
sorted_table = sort(country_table, decreasing = TRUE)
barplot(sorted_table, width = 1, cex.names = 0.6, col = c("azure2", "darkseagreen2", "khaki1", "lightco
```




```r
mode_country = sort(country_table, decreasing = TRUE)[1:1]
print("Moda del Nivel de Experiencia")
```

```
## [1] "Moda del Nivel de Experiencia"
```

```r
mode_country
```

```
##  US
## 355
```

**2. Exploración de Datos Usando Herramientas de Visualización**

a. Variables Cuantitativas:

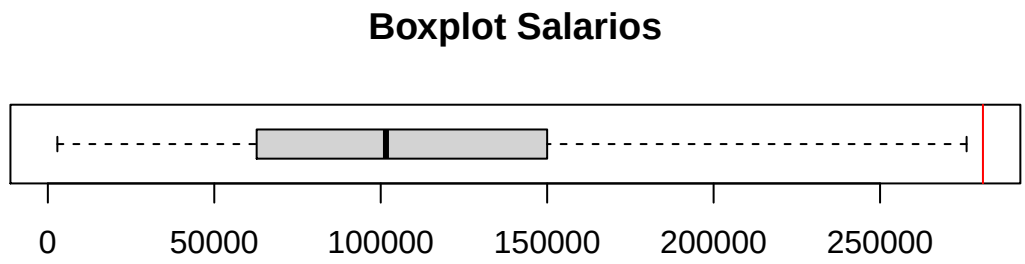
- Medidas de Posición

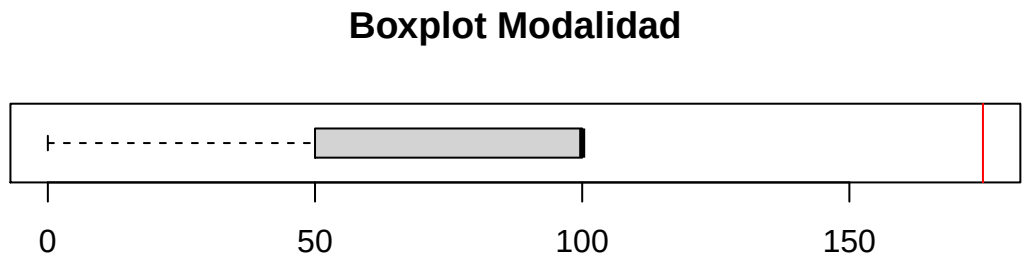
```r
# Cuartiles Para Salarios
q1_s = quantile(salaries_usd, 0.25)
q3_s = quantile(salaries_usd, 0.75)
```

```r
rc_s = q3_s - q1_s # Rango intercuartílico
y2_s = q3_s + 1.5 * rc_s
# IQR(salaries_usd)

# Cuartiles Para Modalidad
q1_m = quantile(job_modality, 0.25)
q3_m = quantile(job_modality, 0.75)
rc_m = q3_m - q1_m # Rango intercuartílico
y2_m = q3_m + 1.5 * rc_m
# IQR(job_modality)

par(mfrow = c(2, 1))

boxplot(salaries_usd, main = "Boxplot Salarios", horizontal = TRUE, ylim = c(0, y2_s))
abline(v = y2_s, col="red")
X_s = db_salaries[salaries_usd < y2_s, c("salary_in_usd")]
summary(X_s)
```

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

```r
boxplot(job_modality, main = "Boxplot Modalidad", horizontal = TRUE, ylim = c(0, y2_m))
abline(v = y2_m, col="red")
```



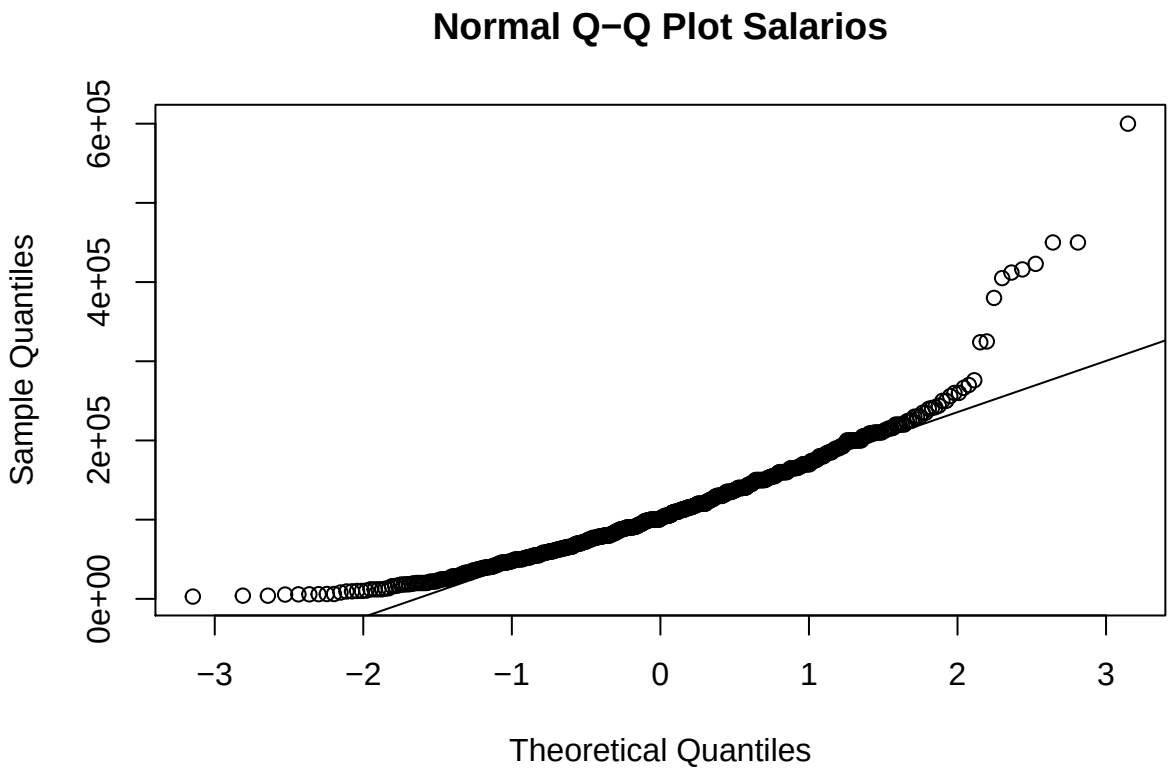



```r
X_m = db_salaries[job_modality < y2_m, c("remote_ratio")]
summary(X_m)
```

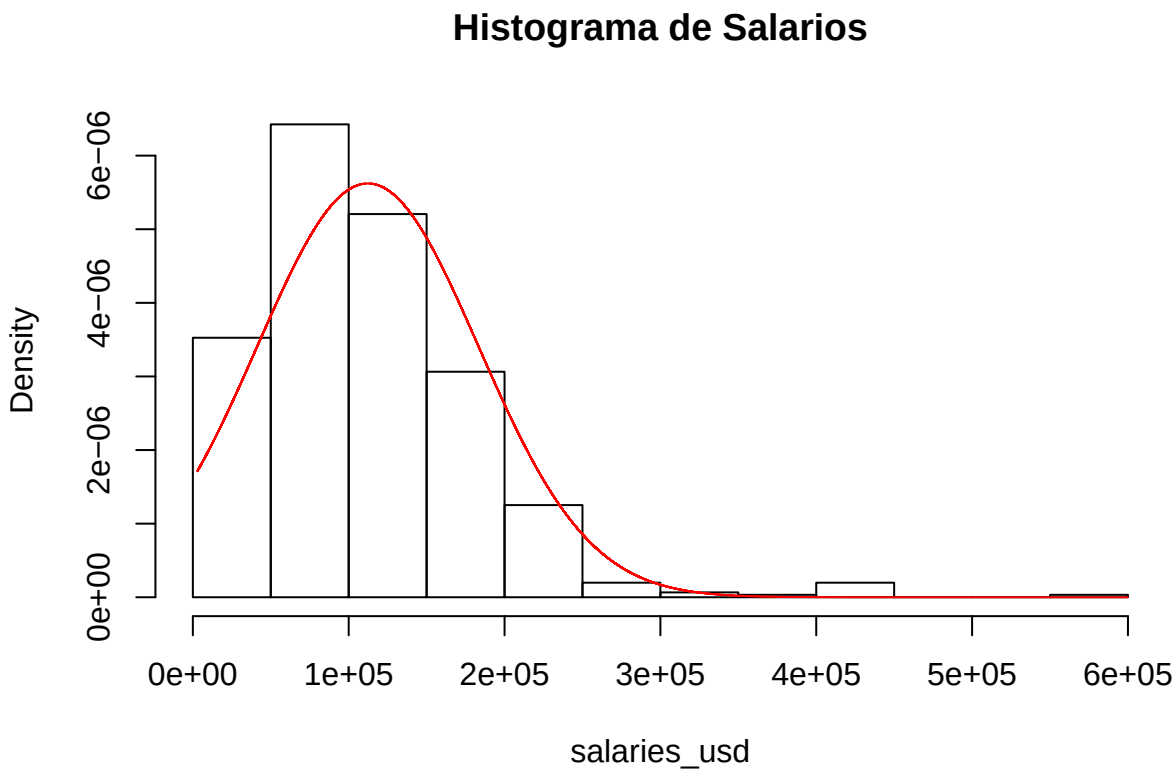
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   50.00  100.00   70.92  100.00  100.00
```

- Análisis de Distribución de los Datos

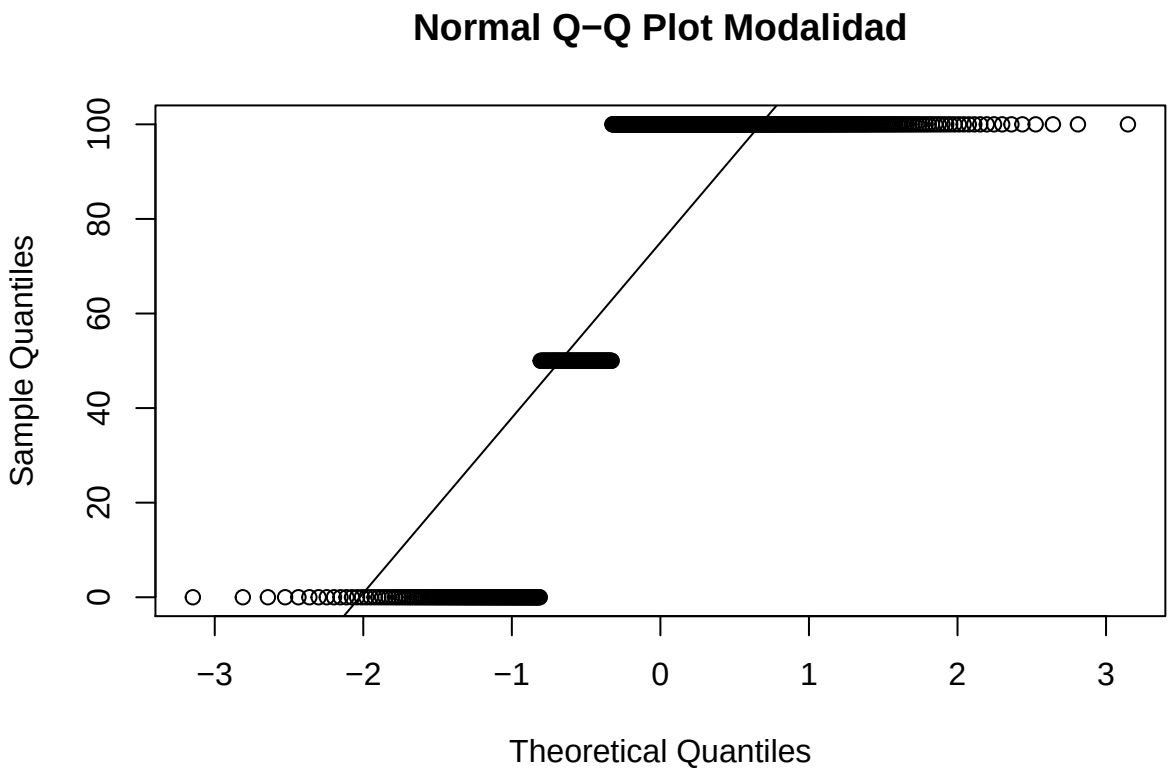
```r
# Salarios
qqnorm(salaries_usd, main = "Normal Q-Q Plot Salarios")
qqline(salaries_usd)
```



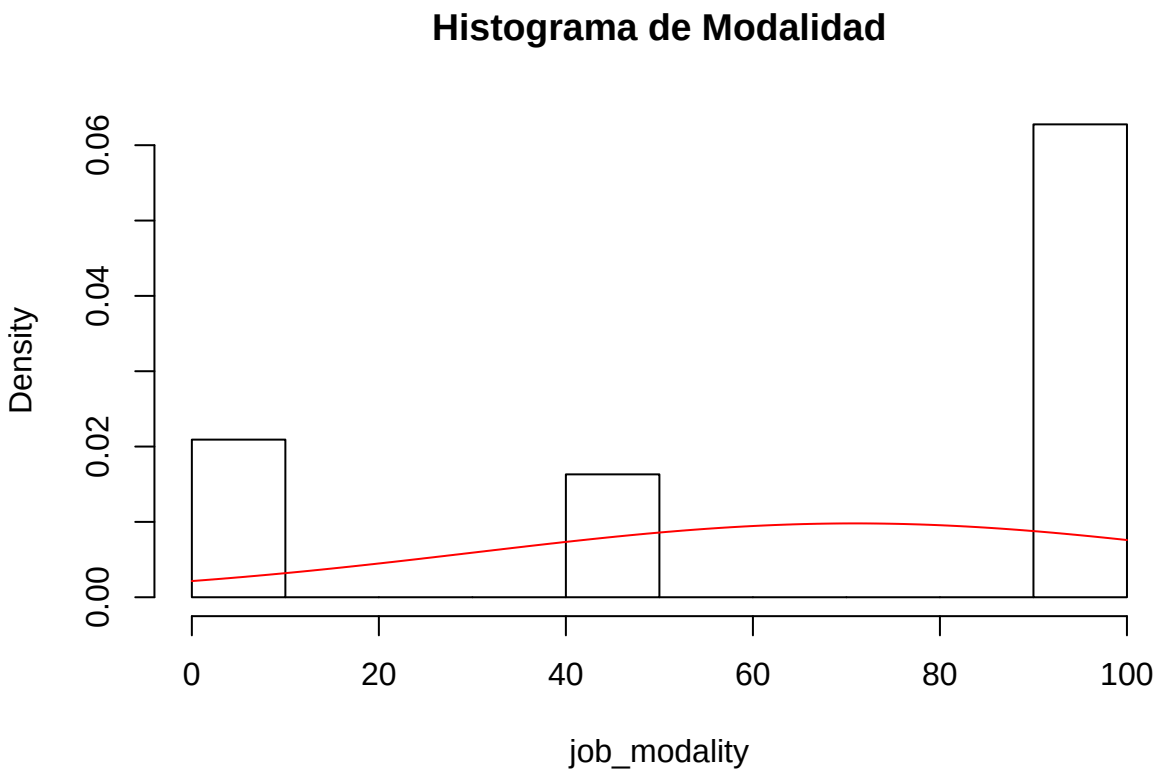

```r
hist(salaries_usd, main = "Histograma de Salarios", prob = TRUE, col = 0)
x = seq(min(salaries_usd), max(salaries_usd), 0.1)
y = dnorm(x, mean(salaries_usd), sd(salaries_usd))
lines(x, y, col = "red")
```




```r
# Modalidad
qqnorm(job_modality, main = "Normal Q-Q Plot Modalidad")
qqline(job_modality)
```




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




```
library(moments)
# Salarios
sprintf("Sesgo de Salarios: %s", skewness(salaries_usd))
```

```
## [1] "Sesgo de Salarios: 1.66342133609776"
```

```
sprintf("Curtosis de Salarios: %s", kurtosis(salaries_usd))
```

```
## [1] "Curtosis de Salarios: 9.29170920802767"
```

```
# Modalidad
sprintf("Sesgo de Salarios: %s", skewness(job_modality))
```

```
## [1] "Sesgo de Salarios: -0.901988052316298"
```

```
sprintf("Curtosis de Salarios: %s", kurtosis(job_modality))
```

```
## [1] "Curtosis de Salarios: 2.10916248872211"
```

### Análisis de Datos y Preguntas Guía Contestadas

1.

```r
data_analyst_salary = db_salaries[db_salaries$job_title == "Data Analyst", ]
mean_salary_da = mean(data_analyst_salary$salary_in_usd)
sprintf("Salario Promedio al que Puede Aspirar un Analista de Datos: $ %s", mean_salary_da)
```

```
## [1] "Salario Promedio al que Puede Aspirar un Analista de Datos: $ 92893.0618556701"
```

```r
library(dplyr)
```

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

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

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

```r
sorted_db = db_salaries[order(db_salaries$salary_in_usd, decreasing = TRUE), ]
top_sorted_db = head(sorted_db, 100)
top_sorted_db = distinct(top_sorted_db, top_sorted_db$company_location, .keep_all = TRUE)
top_sorted_db
```

```
##     X work_year experience_level employment_type                  job_title
## 1 252      2021               EX              FT    Principal Data Engineer
## 2   1      2020               SE              FT Machine Learning Scientist
## 3 160      2021               EX              FT               Head of Data
## 4 224      2021               SE              FT Machine Learning Scientist
## 5 474      2022               MI              FT             Data Scientist
## 6 257      2021               SE              FT   Principal Data Scientist
##   salary salary_currency salary_in_usd employee_residence remote_ratio
## 1 600000             USD        600000                 US          100
## 2 260000             USD        260000                 JP            0
## 3 230000             USD        230000                 RU           50
## 4 225000             USD        225000                 US          100
## 5 140000             GBP        183228                 GB            0
## 6 147000             EUR        173762                 DE          100
##   company_location company_size top_sorted_db$company_location
## 1               US            L                             US
## 2               JP            S                             JP
## 3               RU            L                             RU
## 4               CA            L                             CA
## 5               GB            M                             GB
## 6               DE            M                             DE
```

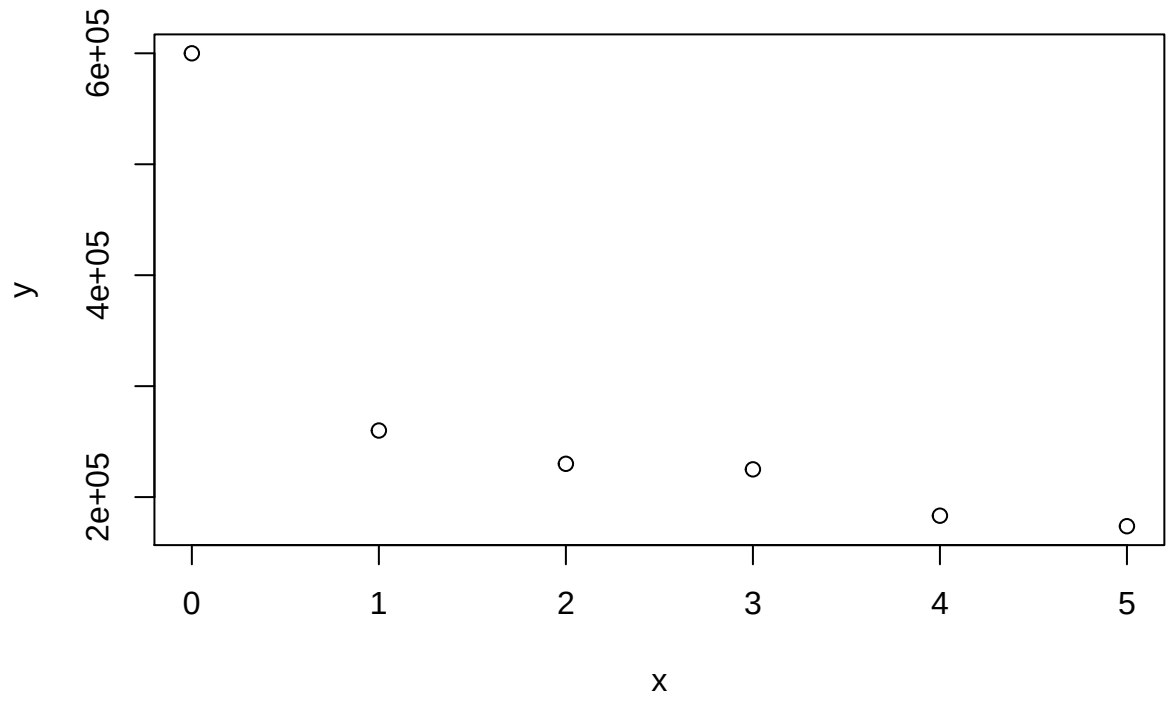
```r
top_countries = top_sorted_db$company_location
top = head(unique(top_countries), 5)
print("Top 5 Países con Mejores Salarios: ")
```

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

```r
top
```

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

```r
x = c(0, 1, 2, 3, 4, 5)#top_sorted_db$company_location
y = top_sorted_db$salary_in_usd
plot(x, y)
```




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

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

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

```
## [1] "Modalidad en línea"
```

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

```
barplot(table(sorted_db$remote_ratio), width = 1, cex.names = 1, col = c("azure2", "darkseagreen2", "kha
```