Momento de Retroalimentación (Portafolio Análisis): Técnicas de procesamiento de datos para el análisis estadístico y para la construcción de modelos

Inteligencia Artificial Avanzada para la Ciencia de Datos Módulo 1: Estadística para la Ciencia de Datos

Jorge Chávez Badillo A01749448 Grupo 101

2022-09-18

Salarios

Resumen

Para este problema se tiene la base de datos ds_salaries, donde se tienen diferentes atributos sobre los salarios de una persona especialista en el análisis de datos tenga un mejor sueldo en las diferentes partes del mundo. Para lo cual se decidió abordar el problema utilizando diferentes herraminetas estadísticas, en primera instancia tenemos el análisis estadístico general para el entendimiento de los datos y posteriormente se tienen unas gráficas de frecuencia, donde obtenemos los puestos de trabajo más populares, lo cual es un aooyo a poder resolver el problema. Algunos de los principales resultados obtenidos fueron que el top 5 de puestos de trabajo tenemos que estos son: Data Scientist, Data Engineer, Data Analyst, Machine Learning Engineer y Research Scientist. Por otro lado, se obtuvieron los niveles de experiencia del dataset, teniendo que la mayoría son Senior-level y Mid-level.

Introducción

Para la solución de este momento de retroalimentación sobre los salarios de trabajos en el dominio de Data Science, fue necesario hacer una exploración de los datos para familiarizarse con su significado, la identificación de las variables cuantitativas y cualitativas y también la implementación de herramientas de visualización para poder generar un mejor entendimiento.

El significado de cada uno de los atributos de la base de datos son los siguientes:

- *work_year* El año en el que fue pagado el salario.

- *experience_level* Nivel de experiencia del puesto de trabajo durante ese año.
- *employment_type* Tipo de empleo (tiempo completo, medio tiempo, freelance)
- *job_title* Nombre del puesto.
- *salary* Monto del salario.
- *salary_currency* Código de cambio.
- *salaryinusd* Monto del salario en dólares.
- *employee_residence* País.
- *remote_ratio* Porcentaje de modalidad en trabajo remoto.
- *company_location* Páis de la empresa.
- *company_size* Promedio de gente que trabajó para la compañia durante ese año.

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

```
knitr::opts_chunk$set(echo = FALSE)
# Lectura de la base de datos
db_salaries = read.csv("ds_salaries.csv")
```

Exploración de variables

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

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

Exploración de la base de datos

1. Medidas Estadísticas

a. Variables Cuantitativas:

- Medidas de Tendencia Central

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

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

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

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

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

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

- Medidas de Dispersión

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

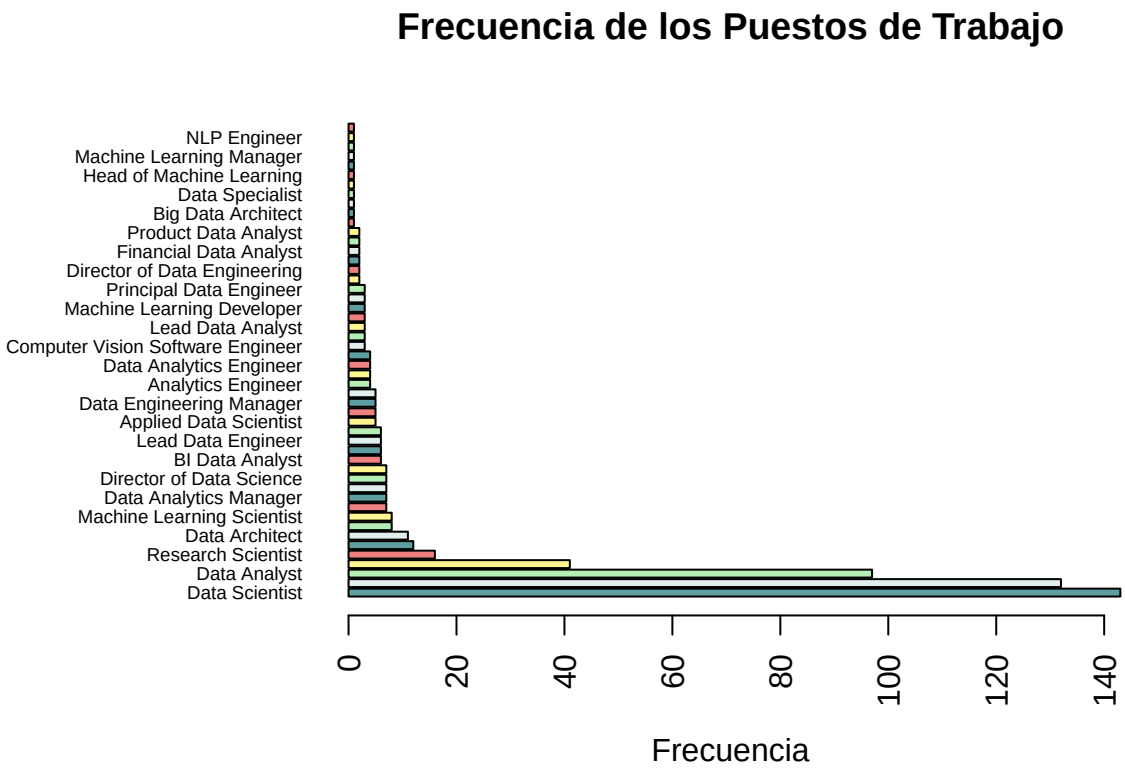
```
## [1] "Mínimo Salarios: 2859"

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

## [1] "Varianza Salarios: 5034932663.1761"

## [1] "Máximo Modalidad: 100"

## [1] "Mínimo Modalidad: 0"

## [1] "Desviación Estándar Modalidad: 40.7091300402214"

## [1] "Varianza Modalidad: 1657.23326863165"
```

b. Variables Cualitativas:

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

## 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
```

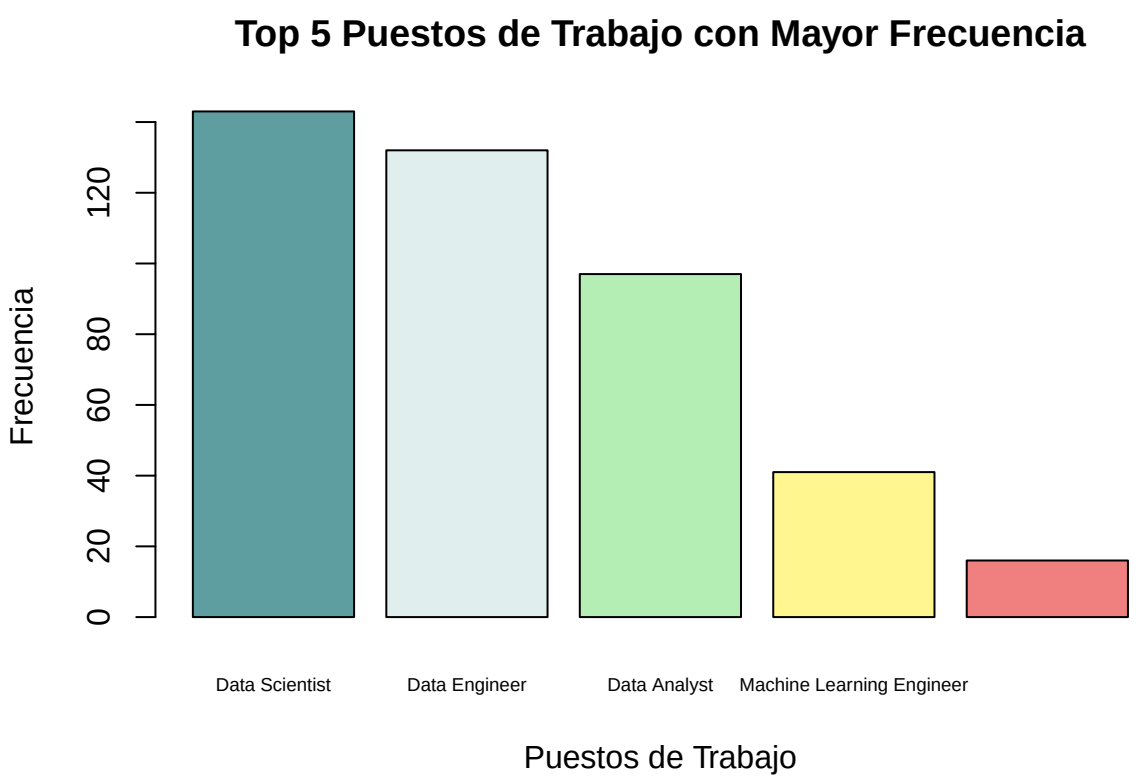
```
##                                 1
##                  Principal Data Analyst
##                                 2
##                 Principal Data Engineer
##                                 3
##                Principal Data Scientist
##                                 7
##                   Product Data Analyst
##                                 2
##                    Research Scientist
##                                16
##                   Staff Data Scientist
##                                 1
```



En el gráfico anterior se muestran los diferentes puestos de trabajo y su frecuencia en el dataset.

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

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




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

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

Para resolver las preguntas base, y para poder brindar mayor información sobre el dataset fue necesario obtener el top 5 de puestos de una manera visual.

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

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

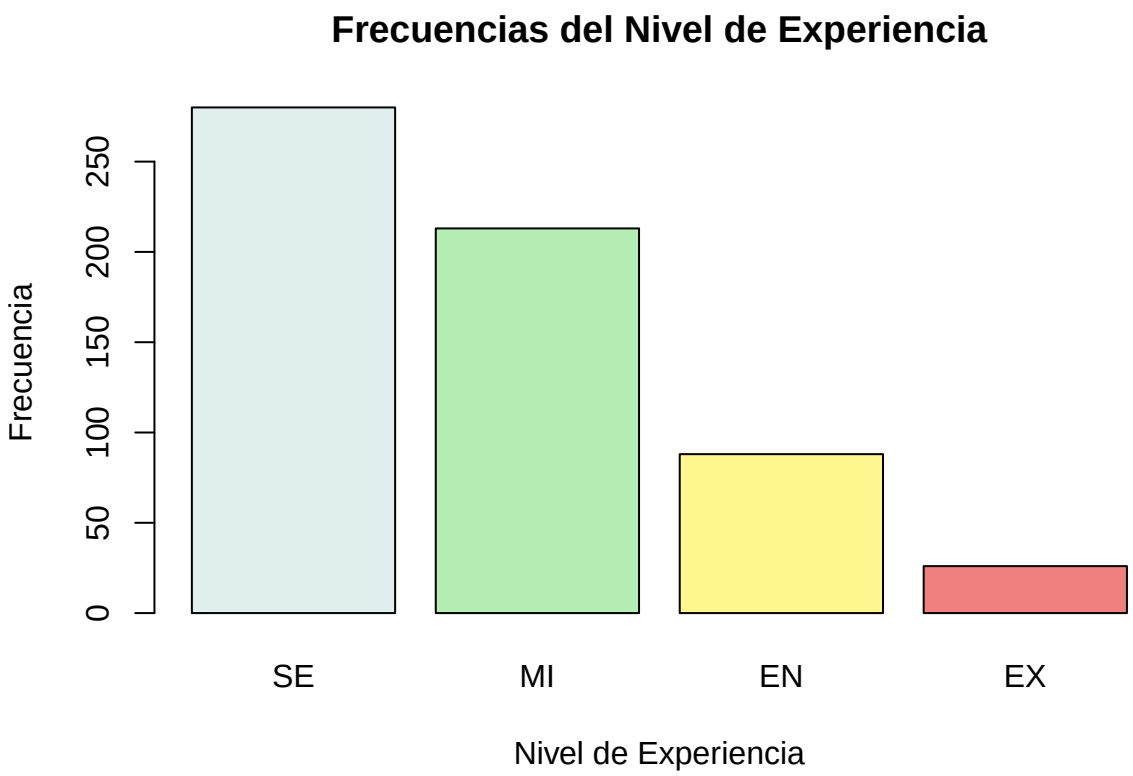
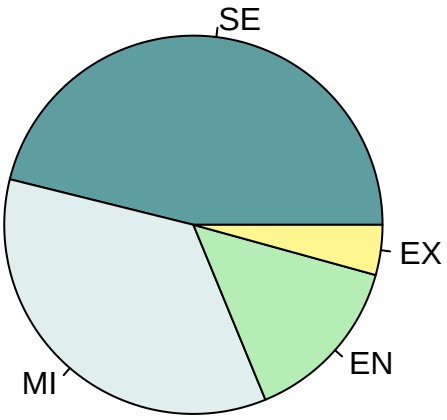

Frecuencias del Nivel de Experiencia
Frecuencia
0
50
100
150
200
250
SE
MI
EN
EX
Nivel de Experiencia

Gráfico de Pastel sobre el Nivel de Experiencia

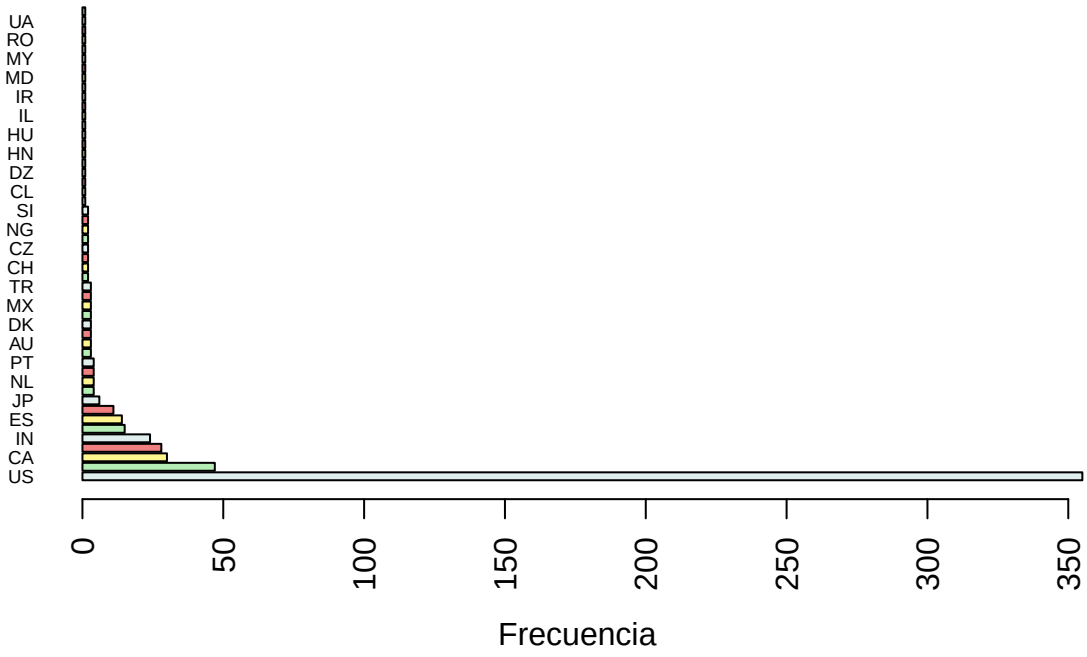



```
## [1] "Moda del Nivel de Experiencia"
```

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

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

```
## 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
```




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

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

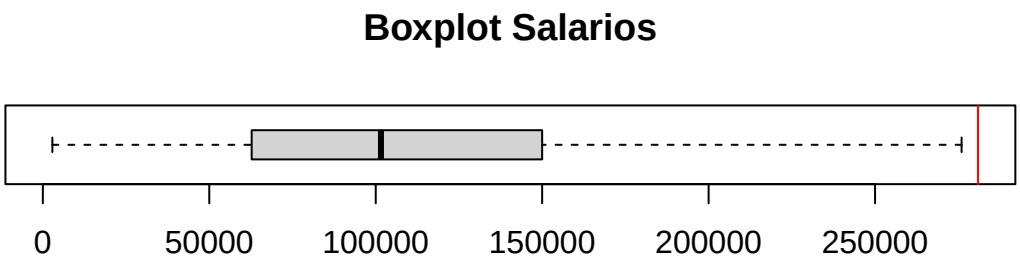
Gráfico sobre las locaciones de las compañias.

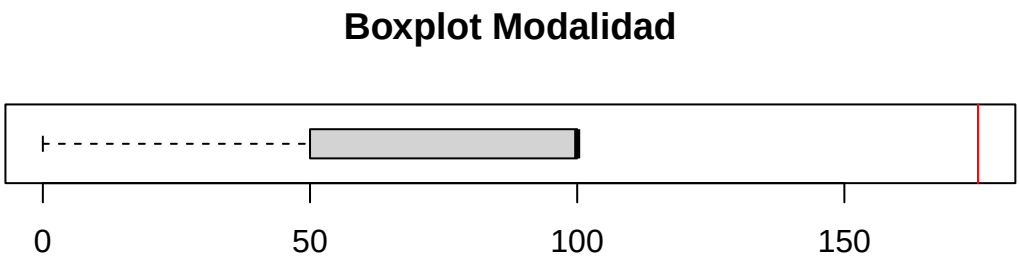
2. Exploración de Datos Usando Herramientas de Visualización

a. Variables Cuantitativas:

- Medidas de Posición

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

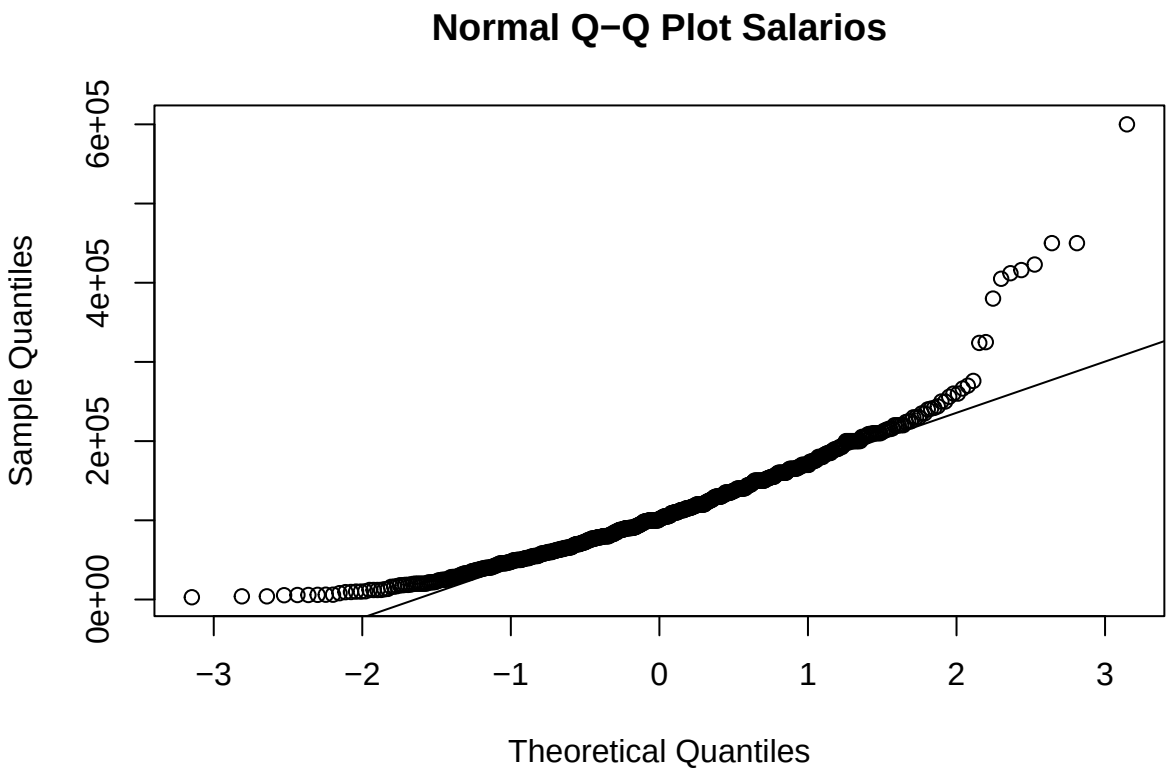


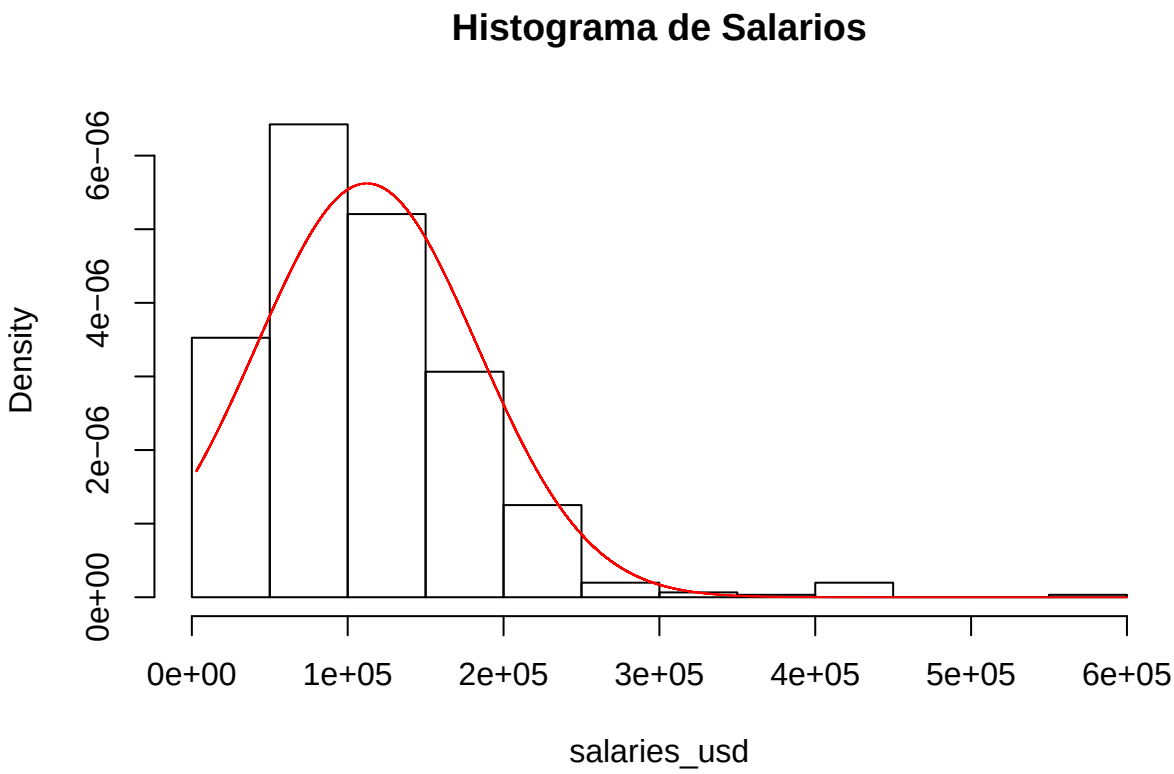



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   50.00  100.00   70.92  100.00  100.00
```

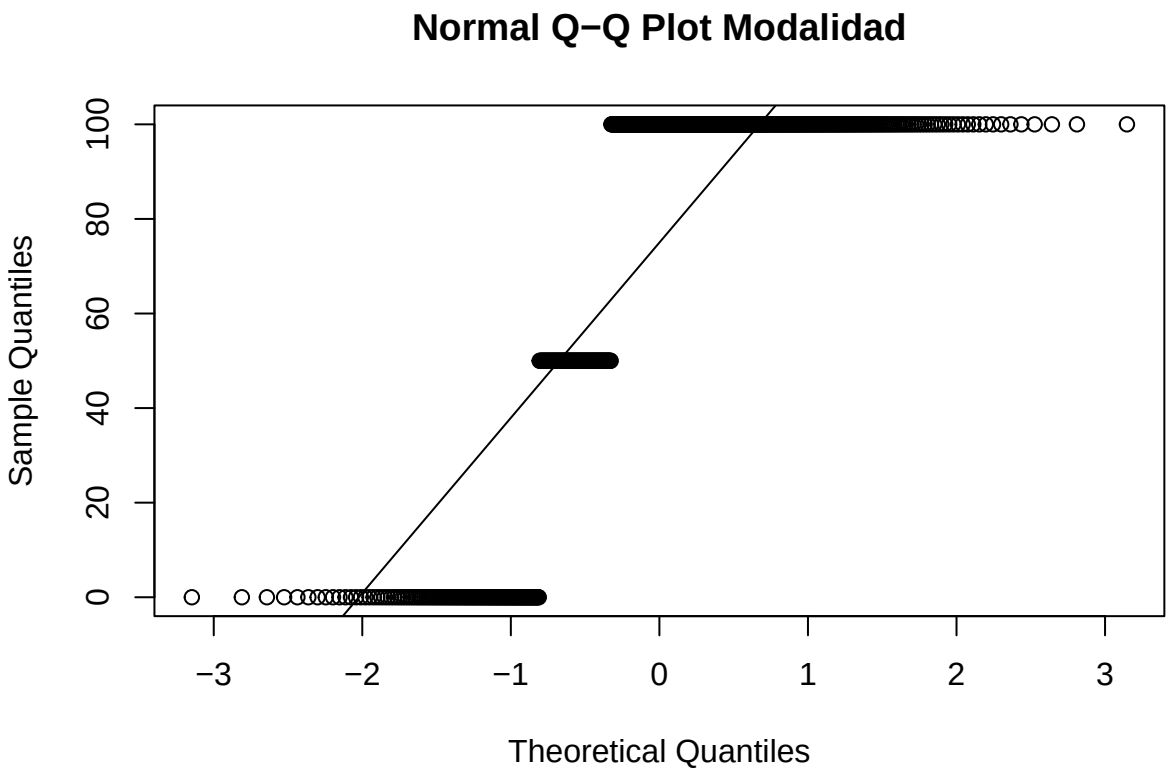
De acuerdo con los boxplots podemos observar que no existe una cantidad relevante de valores atípicos dentro de las variables utilizadas, pero si es posible identificar que se tienen sesgos muy aproximados a la simetría para salarios y para la modalidad se observa un sesgo hacia la izquierda.

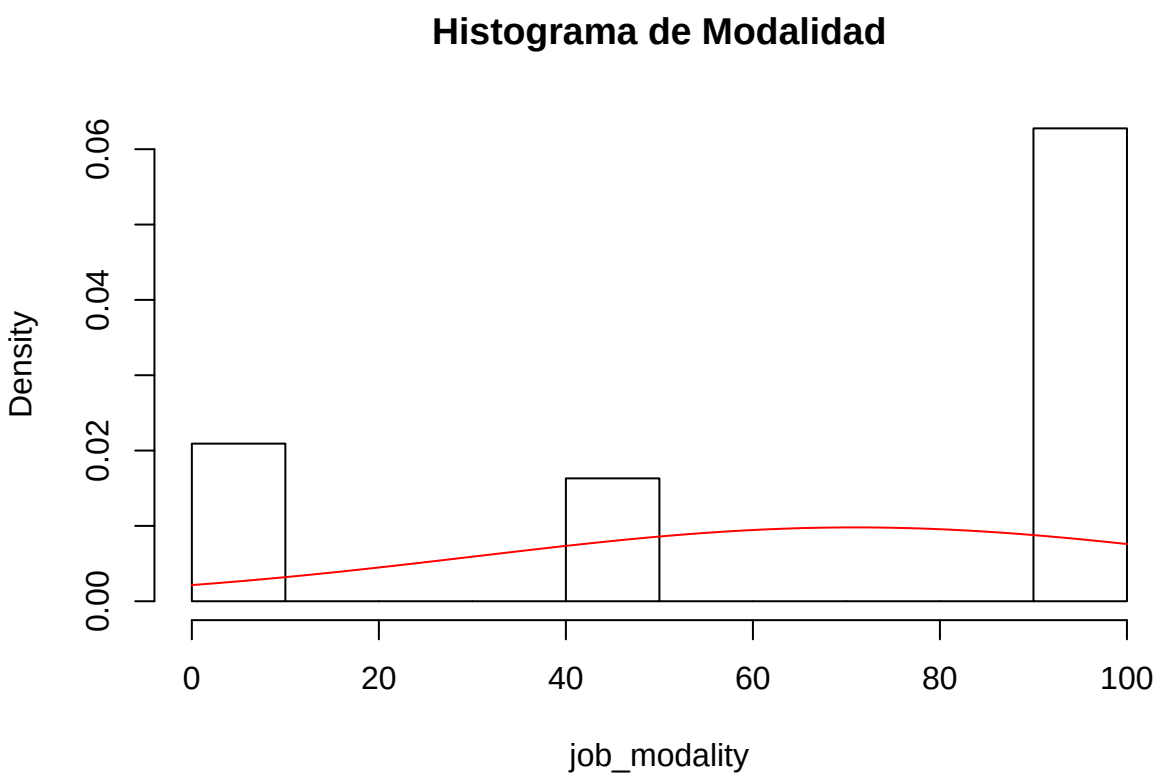
- Análisis de Distribución de los Datos


Normal Q−Q Plot Salarios
Sample Quantiles
0e+00
2e+05
4e+05
6e+05
−3
−2
−1
0
1
2
3
Theoretical Quantiles



De acuerdo con la gráfica de normalidad podemos observar que esta tiene un comportamiento aproximádamente ideal en su simetría.


Normal Q−Q Plot Modalidad
Sample Quantiles
0
20
40
60
80
100
−3
−2
−1
0
1
2
3
Theoretical Quantiles



Gracias a las gráficas podemos observar que la gráfica de normalidad cuenta con un sesgo hacia la izquierda ya que los datos se encuentran ligeramente recargados hacia la derecha.

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

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

```
## [1] "Sesgo de Modalidad: -0.901988052316292"
```

```
## [1] "Curtosis de Modalidad: 2.10916248872207"
```

De acuerdo con los datos calculados sobre la curtosis y sesgo, podemos observar que para los salarios se tiene un valor muy sesgado a la derecha ya que este es mayor a 0.5 y en la parte de la curtosis, tenemos que esta es leptocúrtica ya que esta es mayor a 0.5.

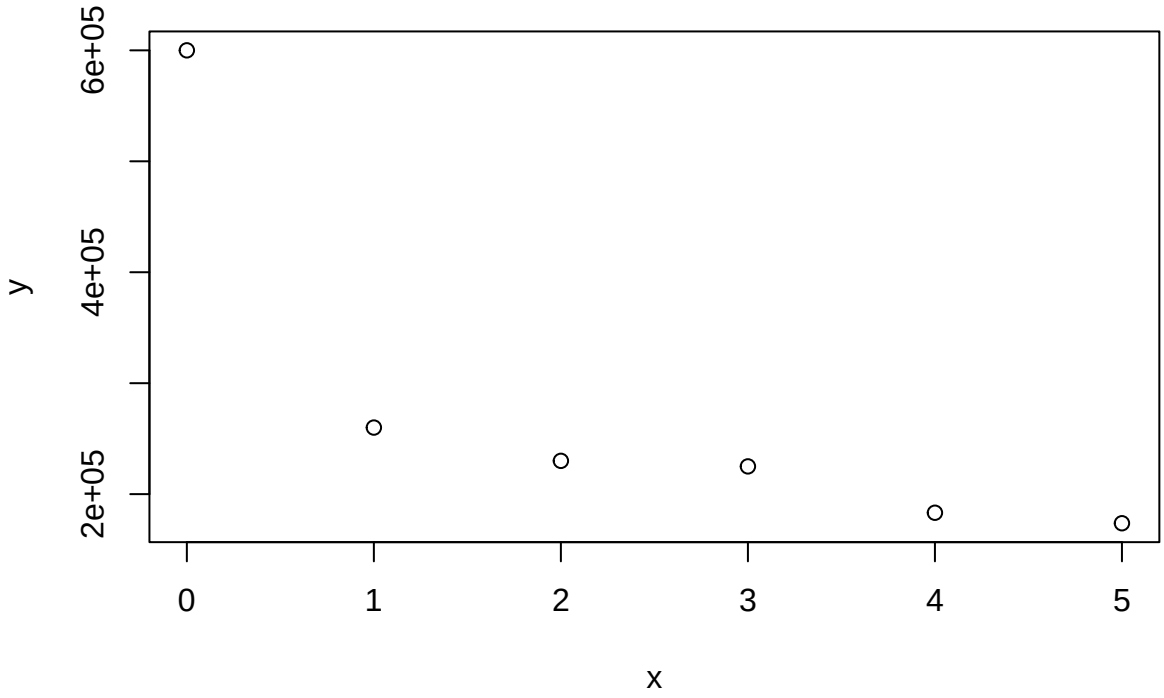
Por otro lado, para la parte de modalidades de trabajo, tenemos que el sesgo se encuentra moderadamente sesgada a la izquierda y la curtosis es leptocúrtica ya que se tiene un pico más pronunciado y colas más pesadas.

Análisis de Datos y Preguntas Guía Contestadas

1.

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

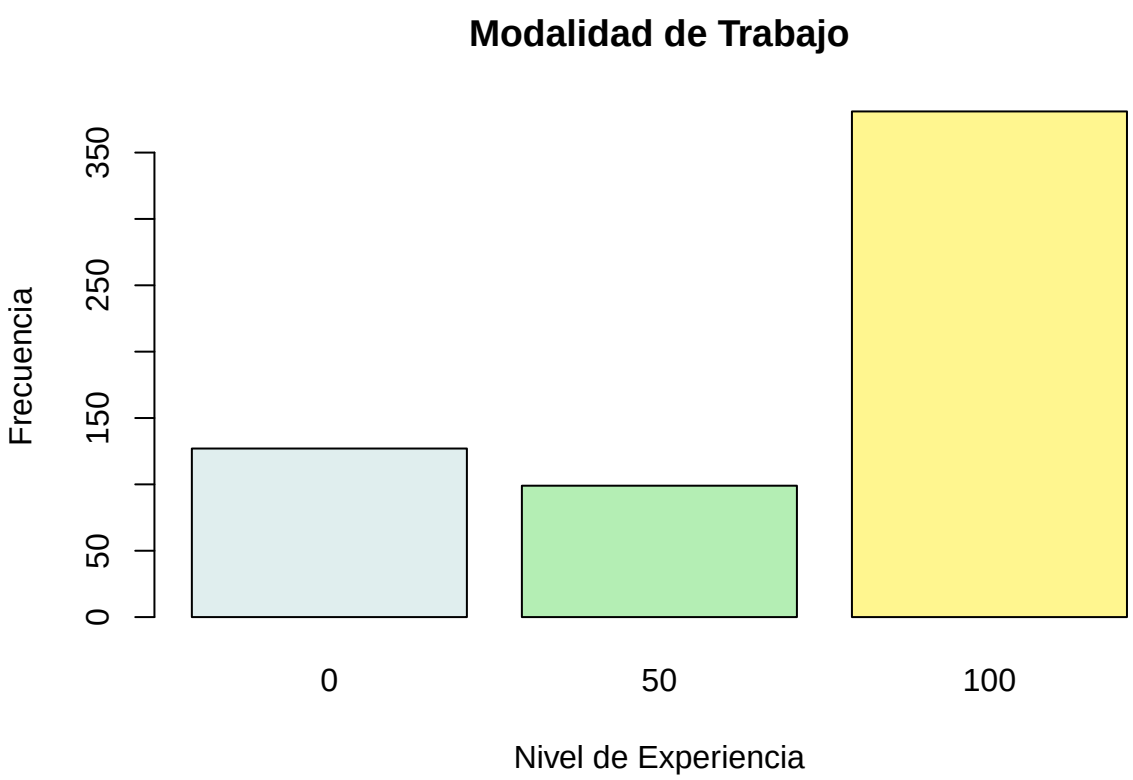
```
##
## Attaching package: 'dplyr'

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

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

##     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

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

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




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

```
## [1] "Modalidad en línea"
```

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



Conclusión

Al hacer un análisis de datos tan general, podemos hacer una gran variedad de procedimientos, es por ello que tener una serie de preguntas base nos ayuda mucho a poder encaminar el análisis y así realizar los procedimientos pertinentes para poder responder las preguntas, por ejemplo, en este caso se decidió responder cuál es el top 5 de países con los mejores salarios y se obtuvo que "US" "JP" "RU" "CA" "GB" son los que mejor salario tienen, lo que indica que es muy probable que al tener un trabajo en estos países, los ingresos del empleado serán mayores que si fuera en otros países, por otro lado, también tenemos que salario promedio al que puede aspirar un analista de datos: $ 92893.0618556701, lo cual es un salario bastante alto, lo que indica que las carreras relacionadas a data science reciben buenos salarios.

Anexos

Liga de Github: https://github.com/A01749448/momento-retroalimentacion-m1.git