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

Módulo 1: Estadística para ciencia de datos

Inteligencia artificial avanzada para la ciencia de datos I

Grupo 101

Resumen

La problemática por resolver es el obtener información sobre la relación entre el tamaño de la compañía y el salario; el salario al que puede aspirar un analista de datos; encontrar los empleos más populares en Estados Unidos y la modalidad con mayor salario. Los métodos y técnicas estadísticas utilizadas son cálculos de tendencia central, dispersión, tablas de distribución de frecuencia, medidas de posición y análisis de distribución de los datos (histogramas y diagramas). Después de todo el proceso de exploración y análisis se obtuvo que no tiene tanta importancia el tamaño de la compañía en relación con los salarios; el promedio de salario de un analista de datos es de $92893.06; los empleos más populares en Estados Unidos son manager de ciencia de datos, manager de analítica de datos y analista de datos BI; y la modalidad que tiene mejor salario es la que es 100% remota.

Introducción

En este portafolio se utilizará la base de datos de "Data Science Job Salaries" para explorar y analizar los datos y así poder resolver una problemática. El dataset contiene los siguientes atributos:

- work_year: el año que se pagó el salario
- experience_level: nivel de experiencia en el trabajo EN (nivel inicial), MI (junior, nivel medio), SE (nivel senior) y EX (nivel ejecutivo)
- employment_type: tipo de empleo PT (medio tiempo), FT (tiempo completo), CT (contrato), FL (freelance)
- job_title: nombre del trabajo
- salary: salario
- salary_currency: moneda del salario pagado
- salaryinusd: salario en dólares
- employee_residence: país de origen del empleado
- remote_ratio: cantidad de trabajo realizado a distancia
- company_location: ubicación de la compañía

- company_size: tamaño de la compañía S (menos de 50 empleados), M (50-250 empleados) y L (más de 250 empleados)

La problemática consiste en contestar las siguientes preguntas: ¿influye el tamaño de la compañía en el salario que puede ofrecer a un analista de datos? ¿cuál es el salario al que pueda aspirar un analista de datos? ¿cuáles son los empleos más populares en Estado Unidos? ¿cuál es la modalidad con mayor salario?

Por lo tanto, se utilizarán los conocimientos adquiridos en el módulo 1 para poder realizar la exploración de la base de datos que incluye el cálculo de medidas estadísticas, el uso de herramientas de visualización e identificación de problemas de calidad de datos al igual que el análisis de datos para contestar las preguntas anteriores. De esta manera, se realizarán cálculos de tendencia central, dispersión, tablas de distribución de frecuencia, medidas de posición y análisis de distribución de los datos (histogramas y diagramas) para poder adentrarnos en los datos. Realizando este proceso no solo se logrará resolver las preguntas planteadas, sino que se podrá tener un acercamiento y un análisis de valor acerca de los trabajos relacionados con ciencia de datos en el mundo.

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.

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

```
## [1] "Número de datos: 607"
```

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

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

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

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

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

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

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

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

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

```
## [1] "Salario mínimo: 2859"
```

```
## [1] "Modalidad mínima: 0"
```

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

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

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

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

Variables cualitativas y Variables categóricas

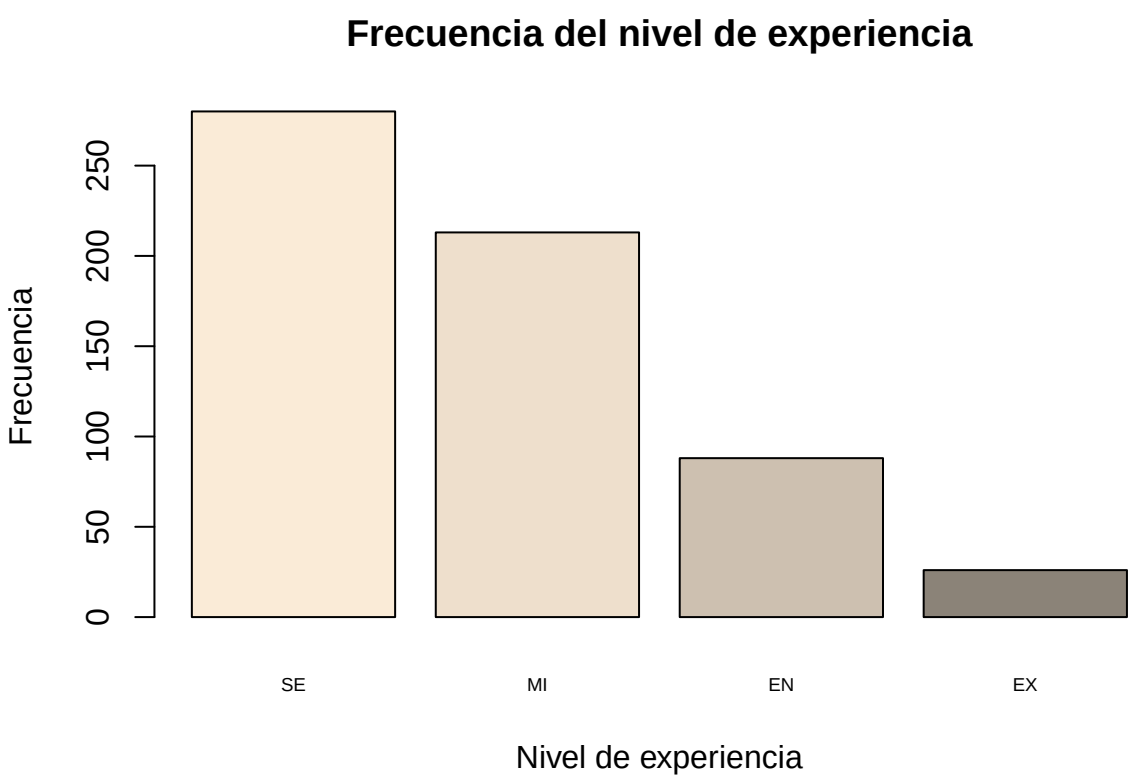
Las variables cualitativas y categóricas que se utilizarán son el nivel de experiencia, el nombre de empleo, el tipo de empleo, la ubicación de la compañía y el tamaño de la compañía. Únicamente se excluyó el tipo de cambio ya que es muy similar a los datos de ubicación de la compañía.

*** 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:"
```

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

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

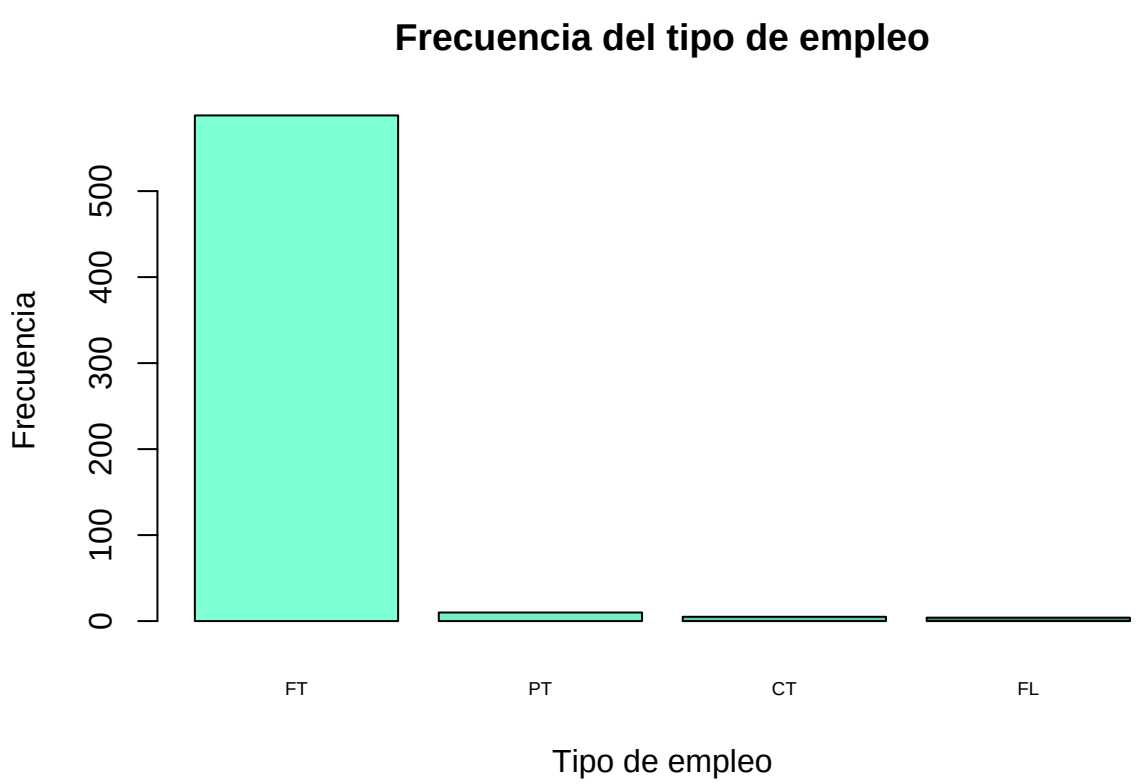


Se puede observar que el nivel de experiencia más popular es el Intermediate SE Senior-level y el menos popular es Expert EX Executive-level.

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

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

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



En el caso del tipo de empleo, el más popular es FT que significa de tiempo completo; mientras que los otros (tiempo completo, contrato y freelance) tienen menos de 11 empleados.

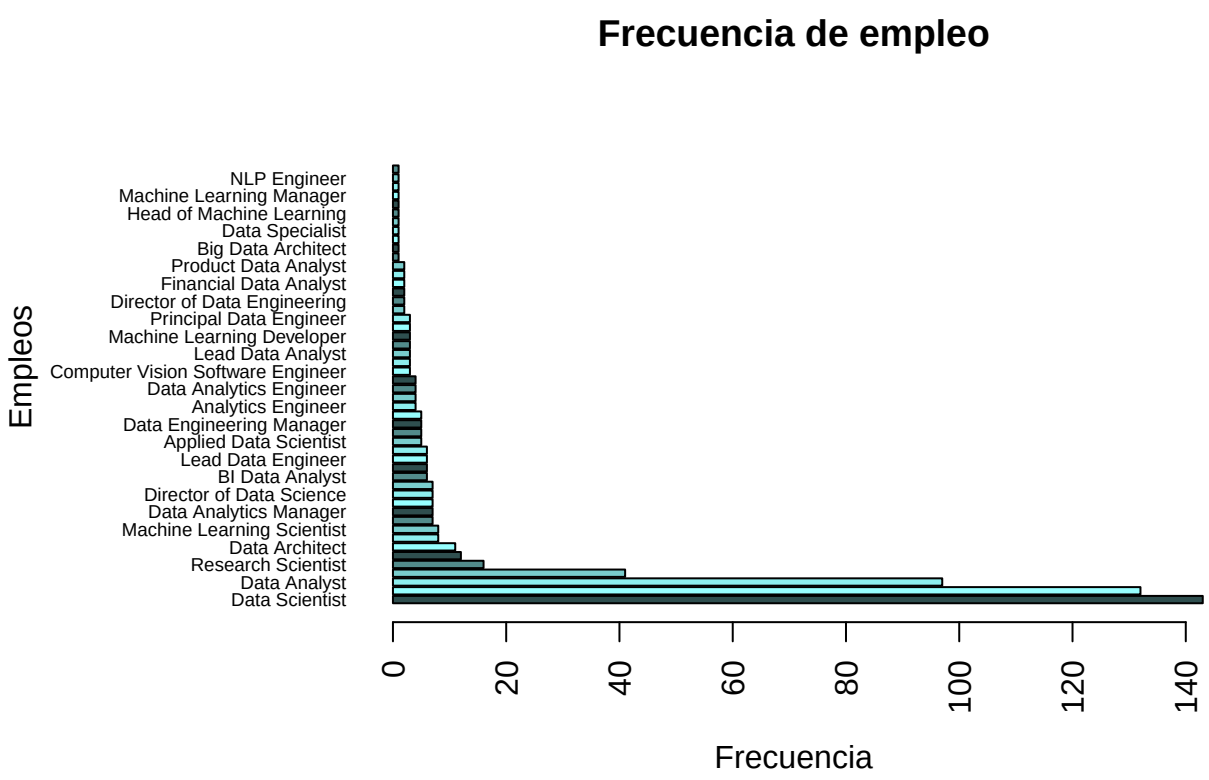
```
## [1] "Tabla de distribución de frecuencia del empleo:"
```

```
## 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
```

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

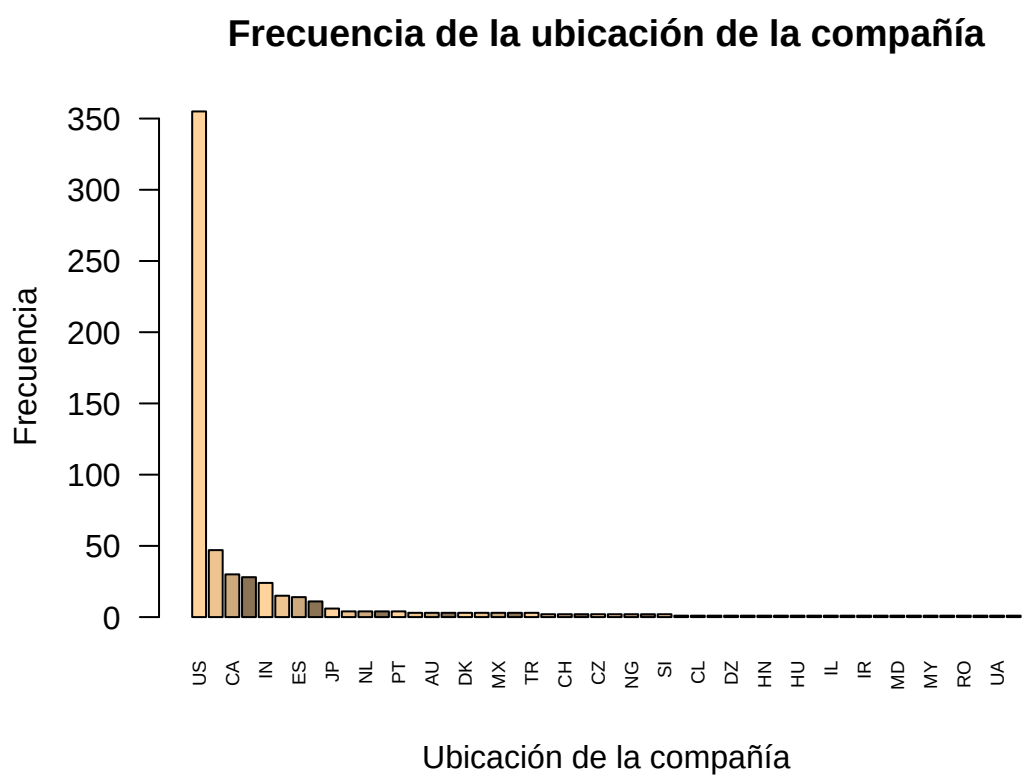


La gráfica anterior muestra que los empleos con mayor frecuencia son data scientist, magine learning engineer y data architect.

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

```
## 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
```

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

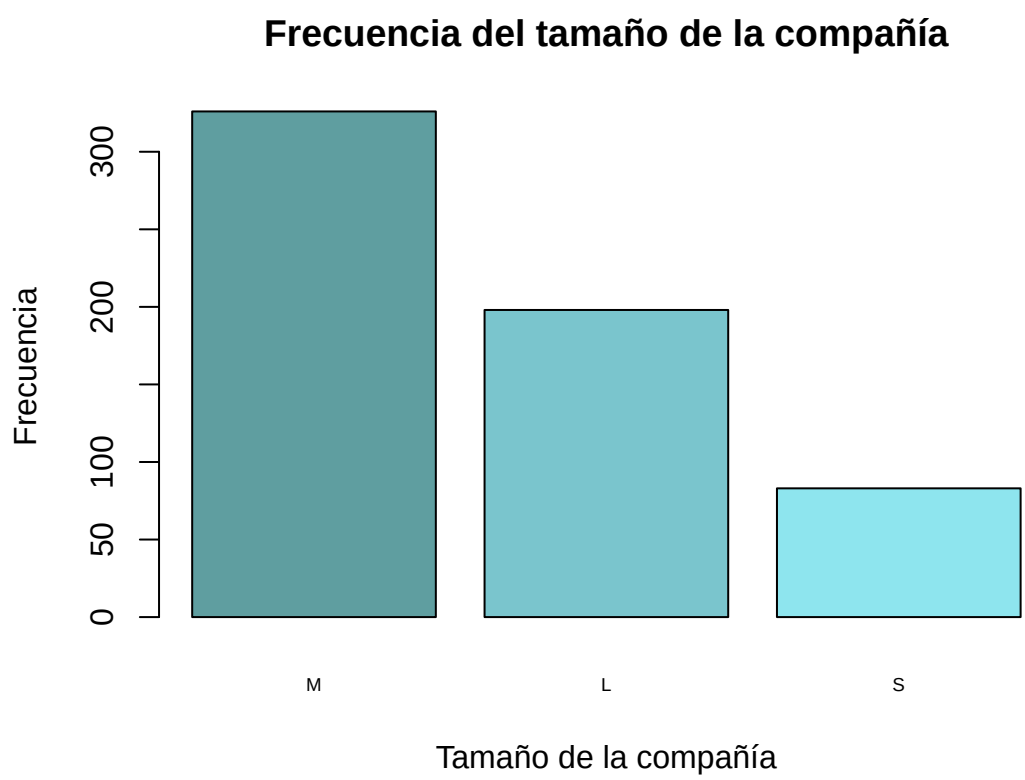


En cuanto a la ubicación de la compañía, Estados Unidos toma el primer lugar con 355 empleados.

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

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

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



El tamaño de compañía más común en trabajos relacionados con ciencia de datos es el mediano. Luego se tienen compañías grandes y por último compañías pequeñas.

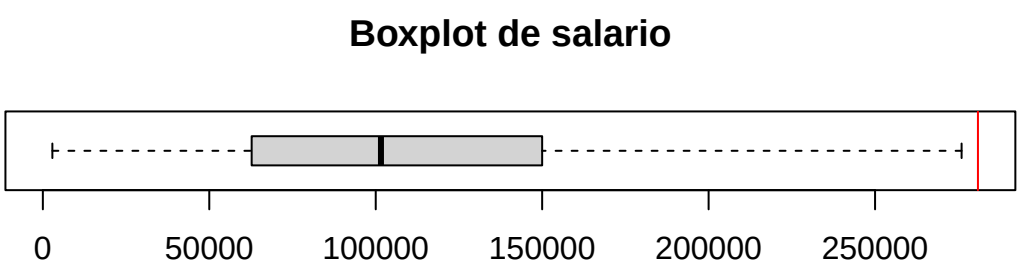
Explora los datos usando herramientas de visualización

Variables cuantitativas:

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

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

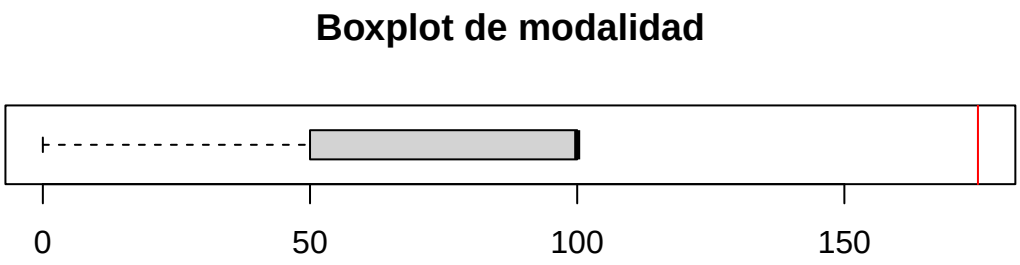
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2859   62649  100000  107169  148261  276000
```



En la gráfica anterior podemos observar que se tiene una distribución de sesgo a la derecha, ya que la mayoría de los datos se concentran en la parte izquierda de la distribución. Por lo tanto, es una distribución asimétrica.

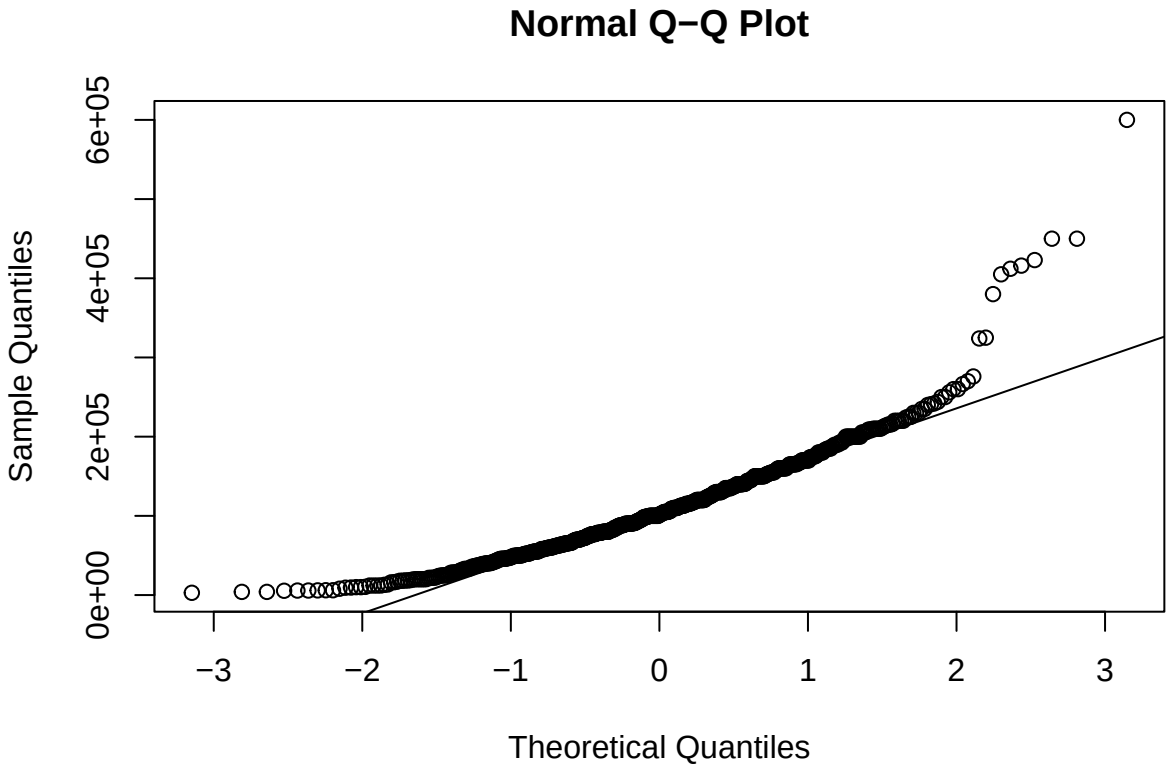
```
## [1] "Cuartiles de modalidad"
```

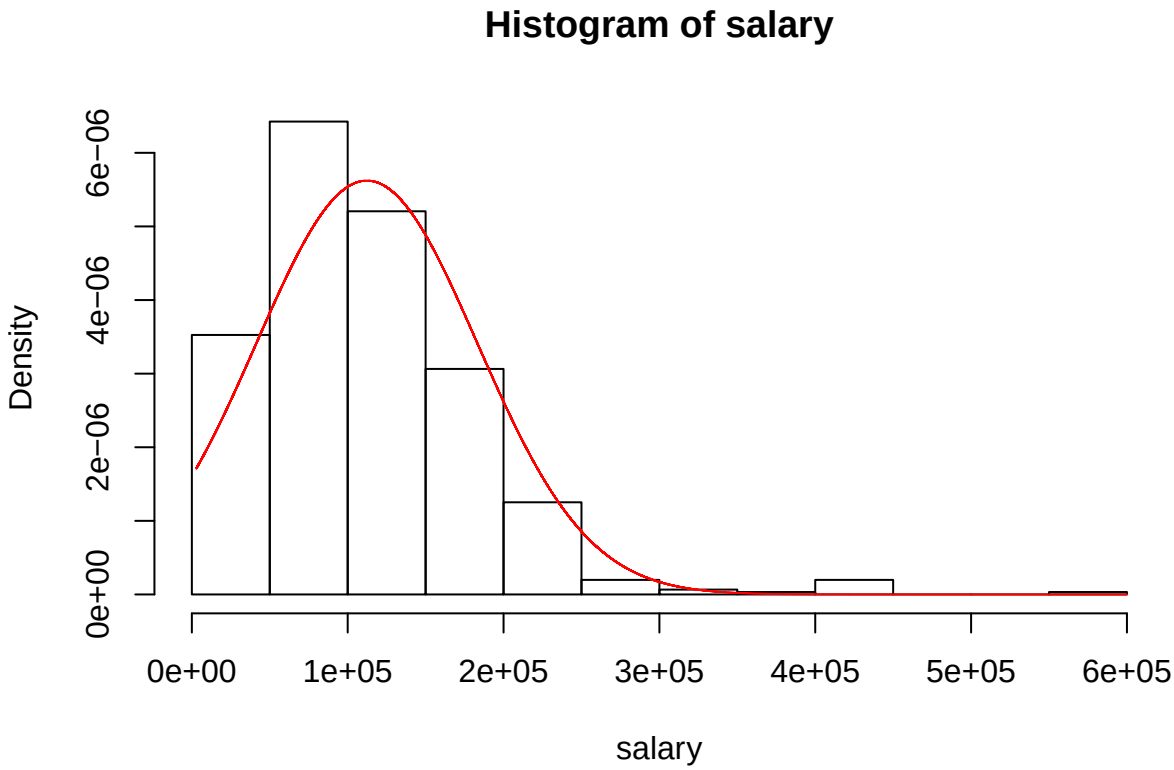
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   50.00  100.00   70.92  100.00  100.00
```



En la gráfica anterior podemos observar que se tiene una distribución de sesgo a la izquierda, ya que la mayoría de los datos se concentran en la parte derecha de la distribución. Por lo tanto, es una distribución asimétrica.

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





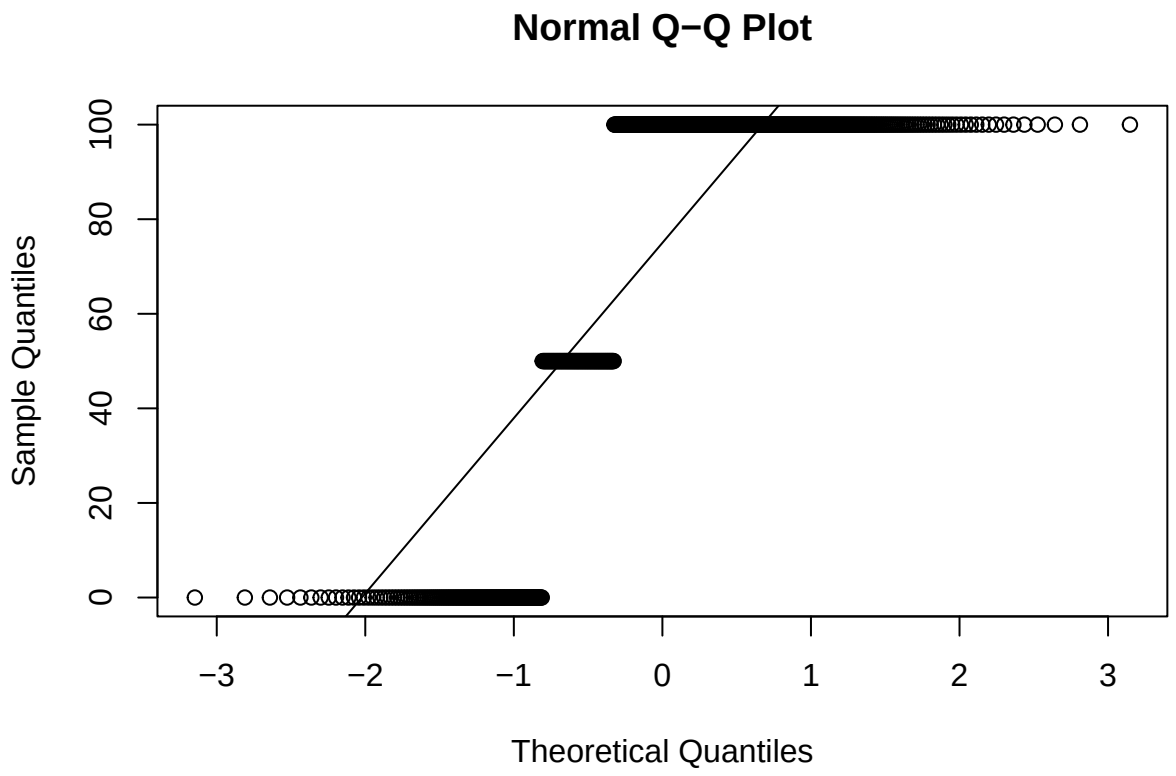
```
##
## Attaching package: 'moments'
```

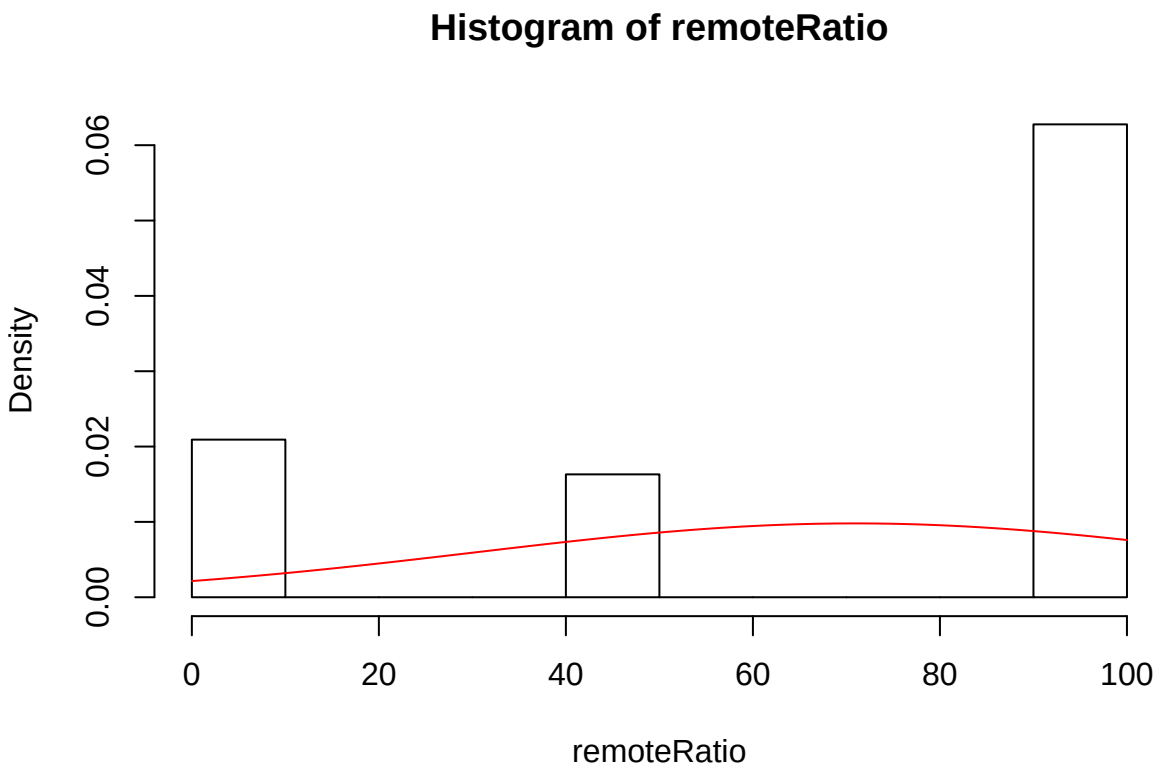
```
## The following object is masked from 'package:modeest':
##
##     skewness
```

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

```
## [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.






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

```
## [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.

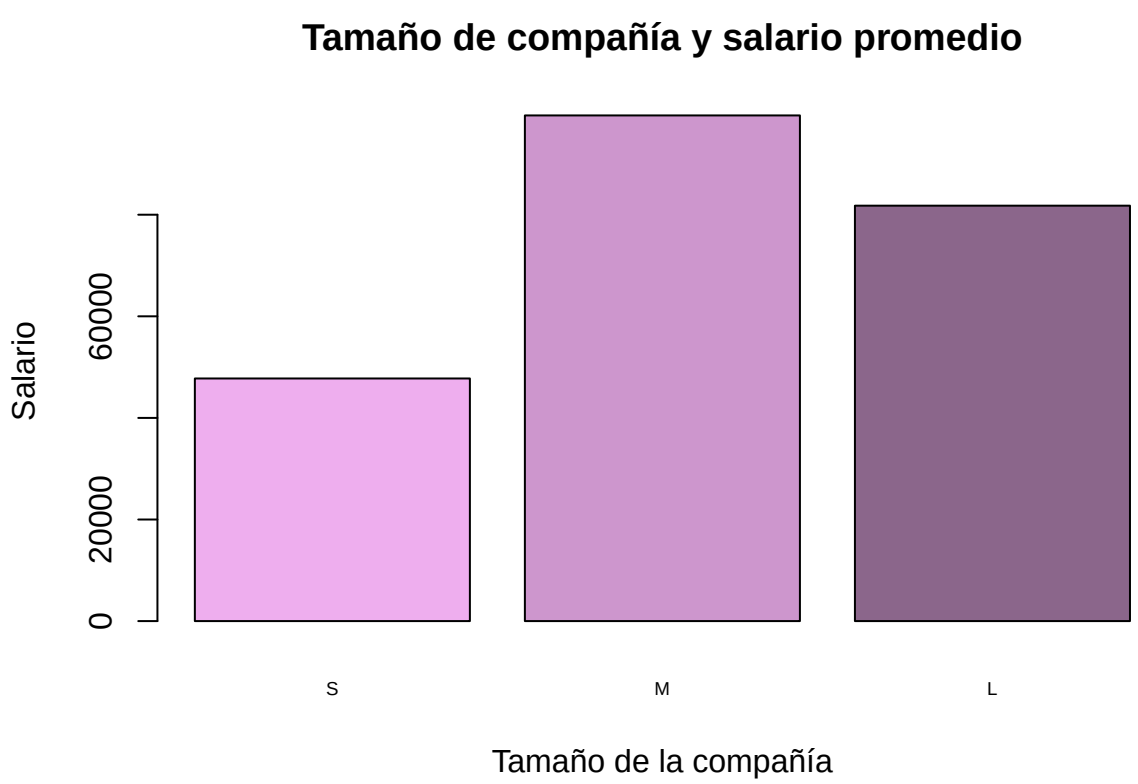
Analizar los datos y contestar las preguntas guía

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

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

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

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



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?

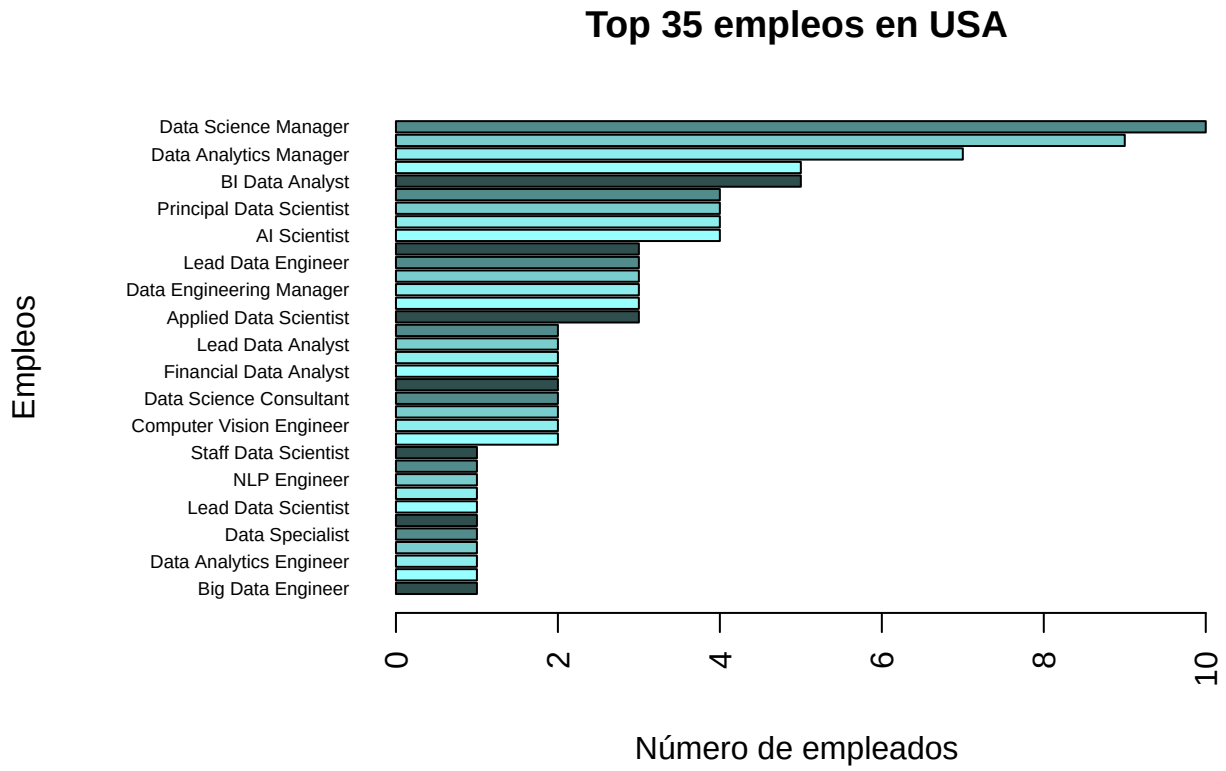
```
## [1] "Promedio del salario de analista de datos: 92893.0618556701"
```

Top 35 empleos en USA

```
## 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
```



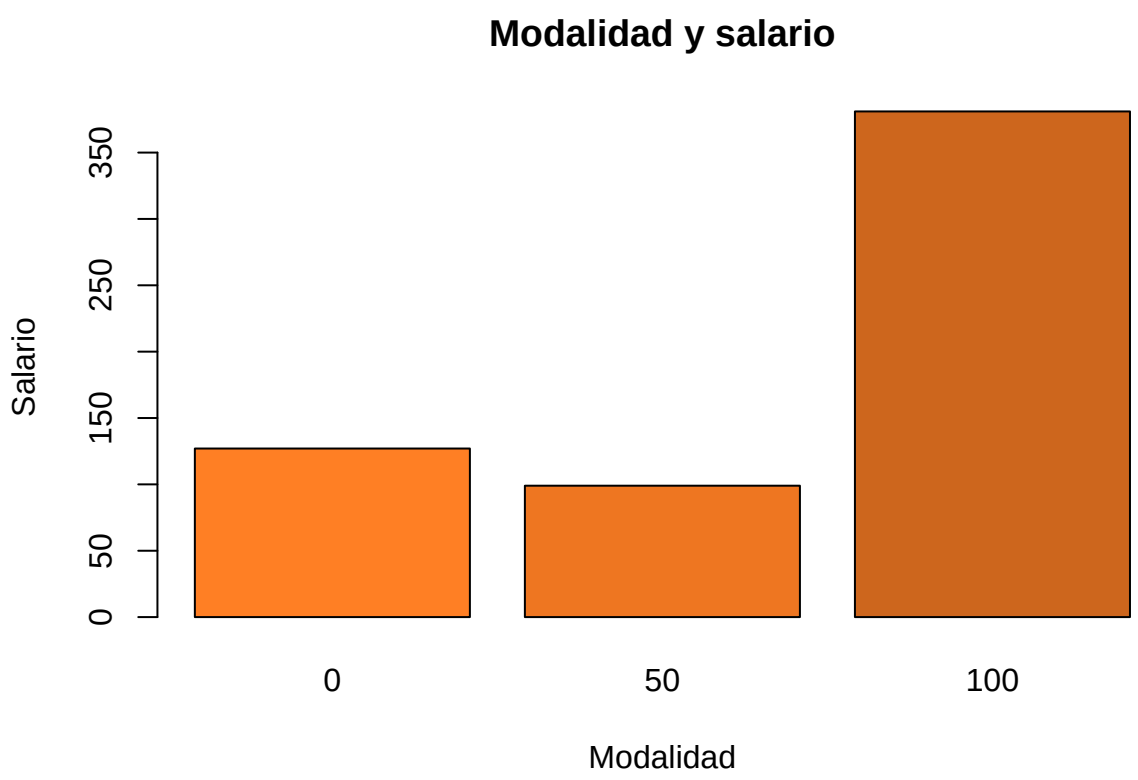
Podemos notar que los empleos más populares en Estados Unidos son: Data Science Manager, Data Analytics Manager y BI Data Analyst.

Modalidad con mayor salario

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

```
## [1] "Modalidad en línea"
```

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



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

Conclusión

En conclusión, las herramientas estadísticas utilizadas fueron bastante útiles para poder responder las preguntas planteadas en la problemática. Considero que es esencial el realizar el proceso de exploración de datos para poder analizar cada una de las variables cuantitativas y cualitativas para conocer su importancia y la forma en la que se pueden relacionar con otras variables. Incluso, los diagramas, histogramas y boxplots nos proporcionaron información valiosa de la distribución de datos y los valores atípicos que existen. El realizar todo este proceso de exploración nos permitió realizar un análisis más completo y proporcionar las respuestas indicadas a cada pregunta planteada. La creación de este portafolio fue de gran ayuda para poder entender los datos que se manejan y generar una solución de calidad.

Anexos

Liga de repositorio: https://github.com/A01750185/RetroM1-AnalisisEstadistico.git