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

2022-09-18

```
knitr::opts_chunk$set(echo = FALSE)
```

Módulo 1: Estadística para ciencia de datos

Inteligencia artificial avanzada para la ciencia de datos I

Grupo 101

Resumen

Mediante este trabajo, se toma en cuenta un dataset de Kaggle llamado "Data Science Jobs Salaries", en donde, como su nombre lo menciona, nos proporciona datos indicadores para poder determinar el salario de una persona con este puesto. Para ello, se muestran a continuación técnicas de procesamiento que nos ayudan al entendimiento de datos para poder saber cuáles son las variables más y menos importantes y así, posteriormente, poder implementar un modelo que nos ayude para una correcta predicción.

Introducción

En los últimos años, los trabajos como Data Scientist han ido en alza, ya que nos ayuda para predecir pronósticos, resultados y hasta poder implementarlos en sistemas inteligentes. Dado esto, resulta de suma importancia poder conocer las condiciones que see necesitar para aspirar y tener un mejor sueldo, por lo qur la base de datos "Data Science Jobs Salaries", nos ayuda a esto. Dentro de este set de datos, contamos con atributos a considerar como:

- **experience_level** : Nivel de experiencia en el trabajo.
- **employment-type** : Tipo de trabajo (Part-time, Full-time, Contract o Freelance)
- **job_title** : Rol trabajado durante el año.
- **salary** : Total de salario bruto pagado.
- **salary_currency** : Salario pagado como código de moneda ISO 4217.
- **salary_in_usd** : Salario en USD.
- **employee_residence** : País de residencia principal del empleado.
- **remote_ratio** : Modalidad de trabajo (remoto, parcialmente remoto, presencial).
- **company_location** : País de oficina principal o sucrusal contratante del empleador.
- **company_size** : Número promedio de personas que trabajaron para la empresa durante el año.

Dadas las descripciones de cada variable, nos surgen interesantes las siguientes preguntas para poder resolver y poder ir sobre ellas para poder hacer predicciones futuras, las cuales son:

- *¿Cuánto puede esperar de salario un Data Analyst?*
- *¿Cuál es el top 5 de países con mejores salarios?*
- *¿Cuál es la modalidad mejor pagada?*

Análisis de resultados

Lectura de datos

Medidas estadísticas

Es importante obtener medidas estadísticas de nuestras variables para poder conocer datos generalizados de nuestro problema.

Variables cuantitativas: Medidas de tendencia central y dispersión

- Salario
- Remote ratio (modalidad de trabajo)

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

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

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

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

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

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

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

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

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

## [1] "Salario máximo: 600000"

## [1] "Salario mínimo: 2859"

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

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

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

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

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

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

Variables cualitativas: tablas de distribución de frecuencia

Las tablas de frecuencia, nos ayudan a conocer cuáles son los datos más y menos demandados respecto a un atributo y de esta manera poder hacer análisis para la resolución del problema. Además, la moda nos ayuda con un resumen sobre dichas tablas de frecuencia al desplegar cuál es el dato que más se repite.

- Experience level
- Company location
- Job title

Tablas de frecuencia

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

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

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

```
## 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
```

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

```
## 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

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

```
## [1] "Moda de país: US"
```

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

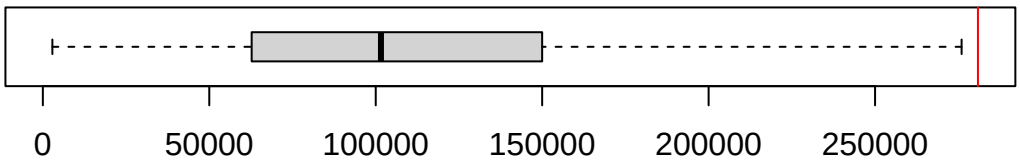
Herramientas de visualización

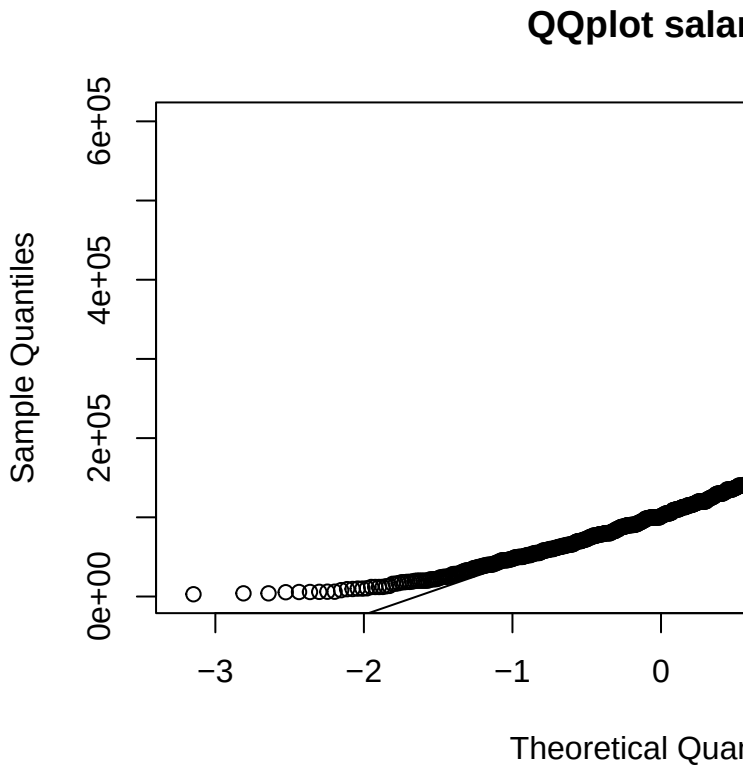
Variables cuantitativas: Medidas de posición y análisis de distribución

Distribución de datos de variables categóricas

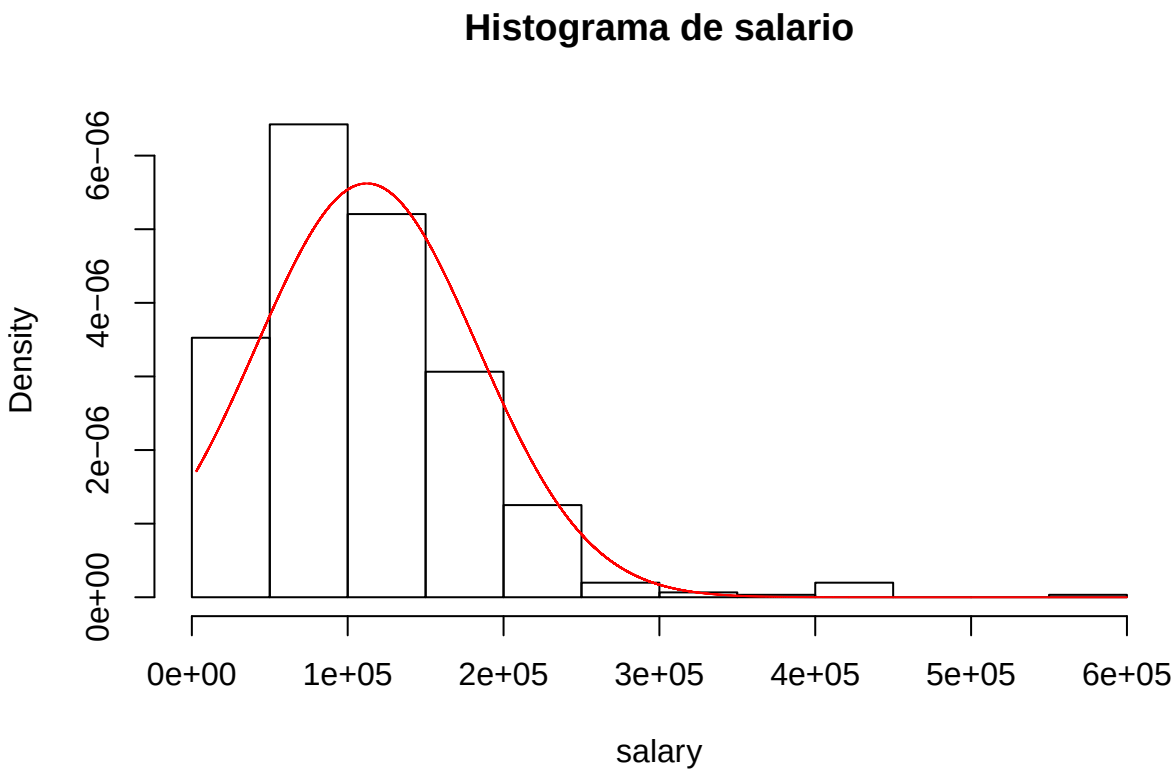
Eliminación de datos atípicos en salarios

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


Salarios
0
50000
100000
150000
200000
250000



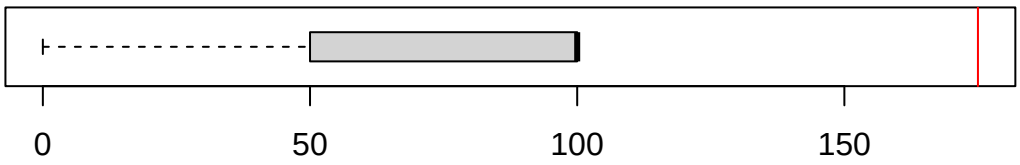
Exploración de la normalidad de la variable salarios

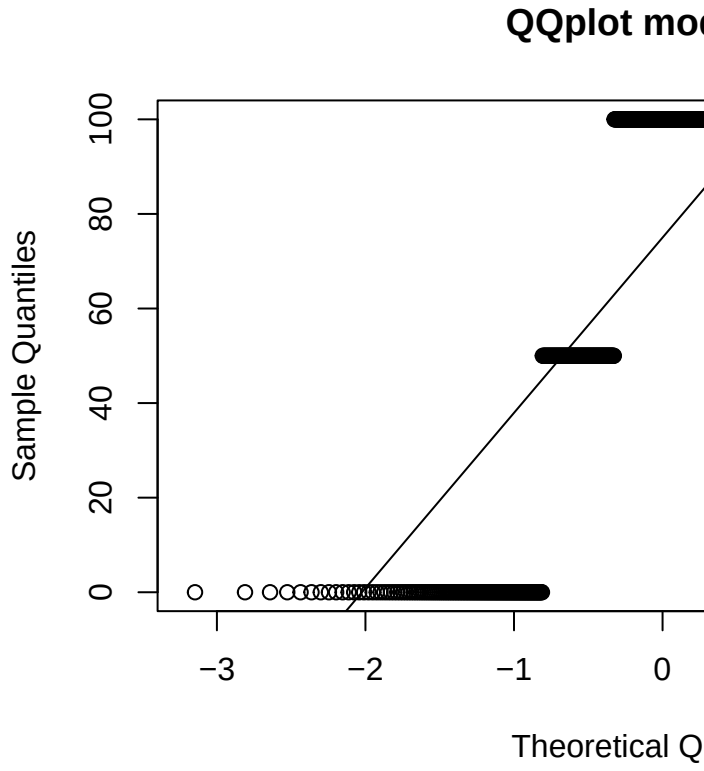


Para los salarios se cuenta con una distribución hacia la izquiera con respecto a su media y mediana, lo que indica una asimetría positiva con sesgo a la derecha.

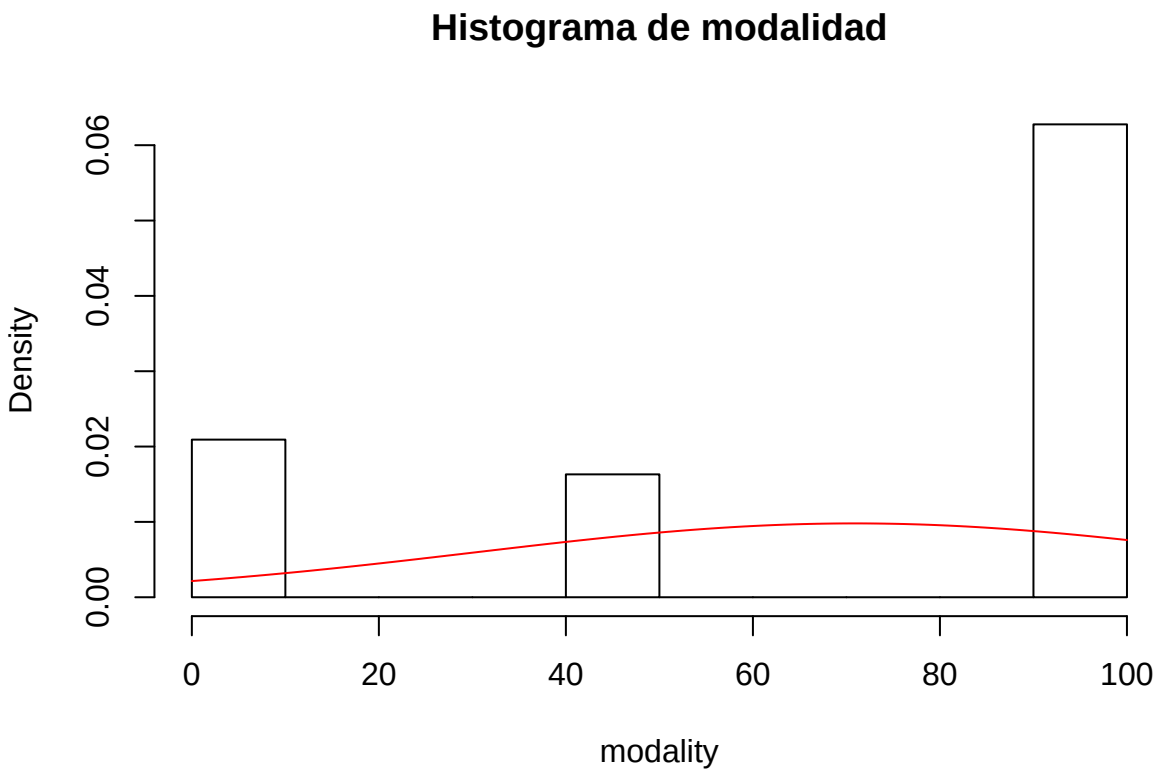
Eliminación de datos atípicos en modalidad

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


Modalidad de trabajo
0
50
100
150



Exploración de la normalidad de la variable modalidad



Cuenta con un sesgo a la izquierda con su distribución centrada a la derecha con respecto a su media y mediana.

Curtosis y sesgo de salarios

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

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

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

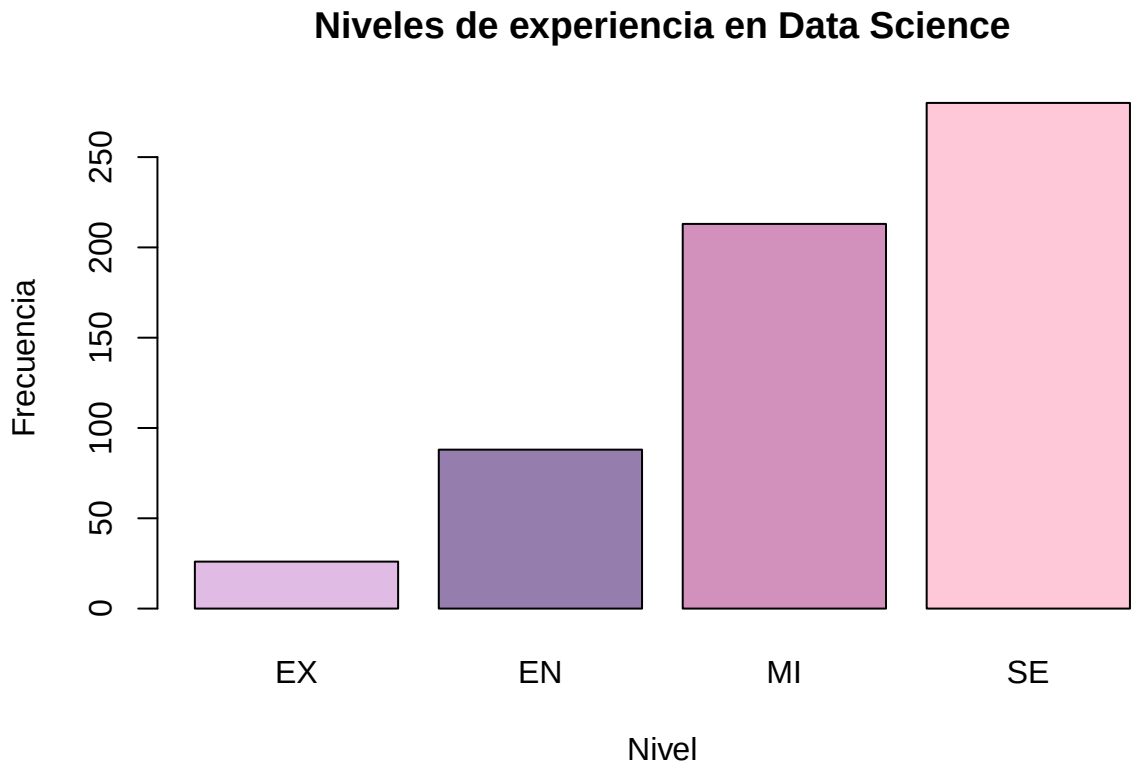
```
## [1] 9.291709
```

Curtosis y sesgo de modalidad

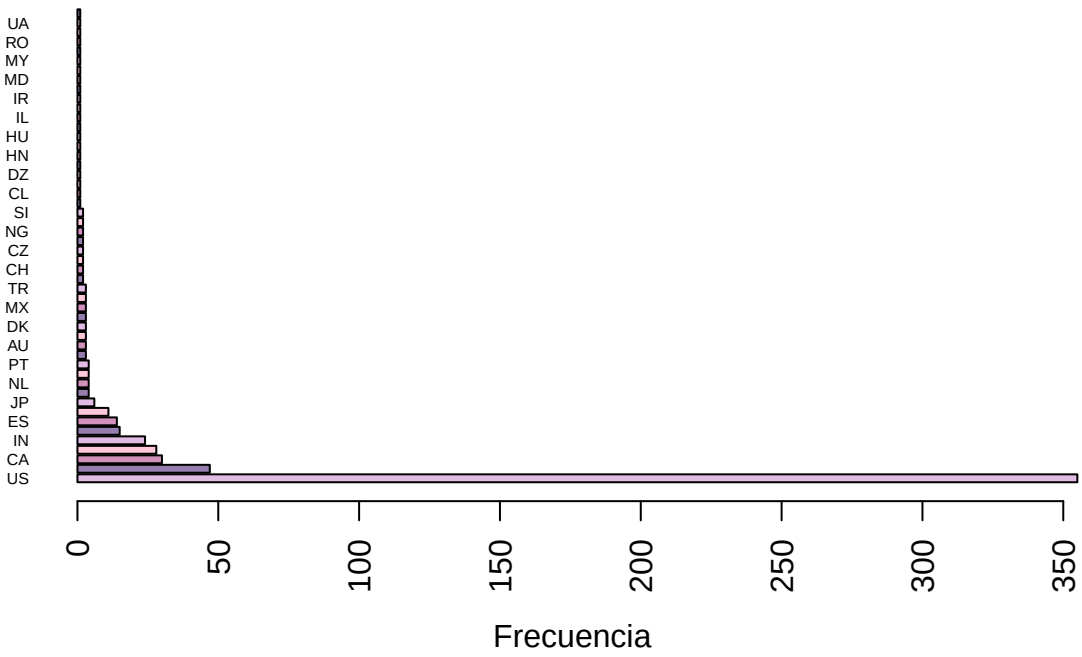
```
## [1] -0.9019881
```

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

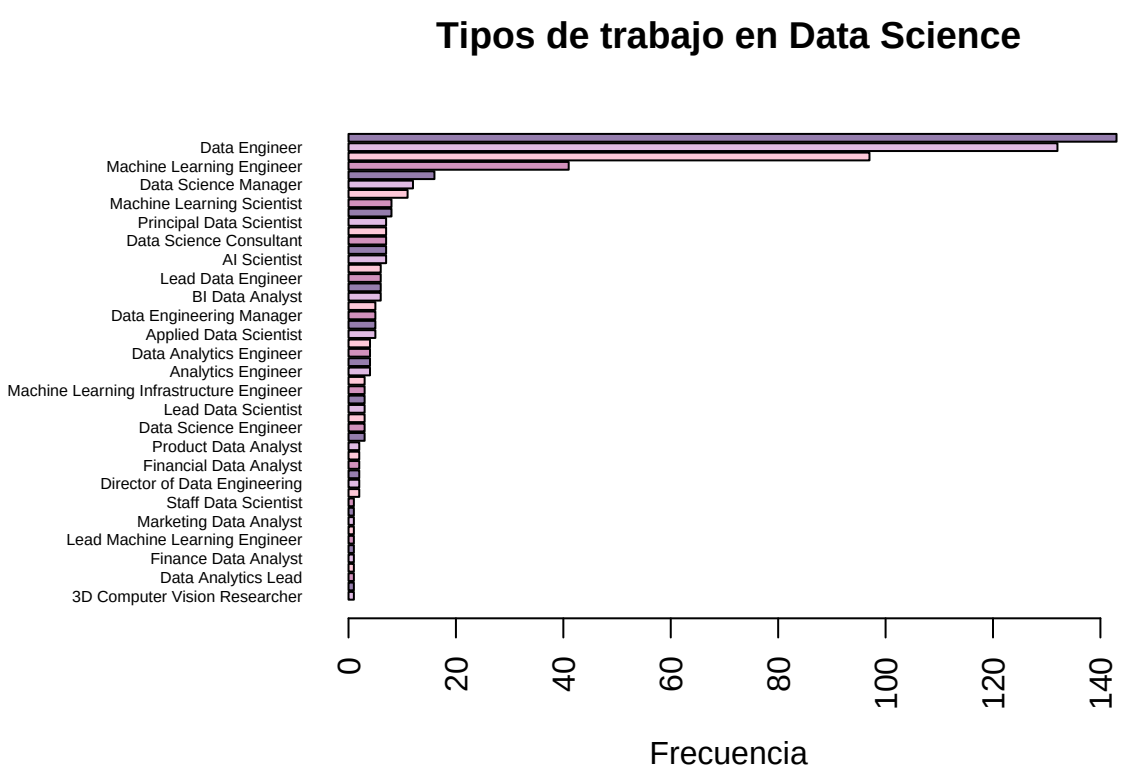
Variables categóricas: Distribución de los datos



De acuerdo a esto, vemos que la experiencia "SE (Senior-level)" es la más frecuente en nuestro set de datos.



Estados Unidos es el país que más lidera por mucho como país de origen de cada uno de los trabajadores de Data Scientist.



Data Engineer es el tipo que más frecuencia tiene dentro de Data Scientist, seguido de DataScience Manager con muy poca diferencia entre ellos.

Preguntas

¿Cuánto puede esperar de salario un Data Analyst?

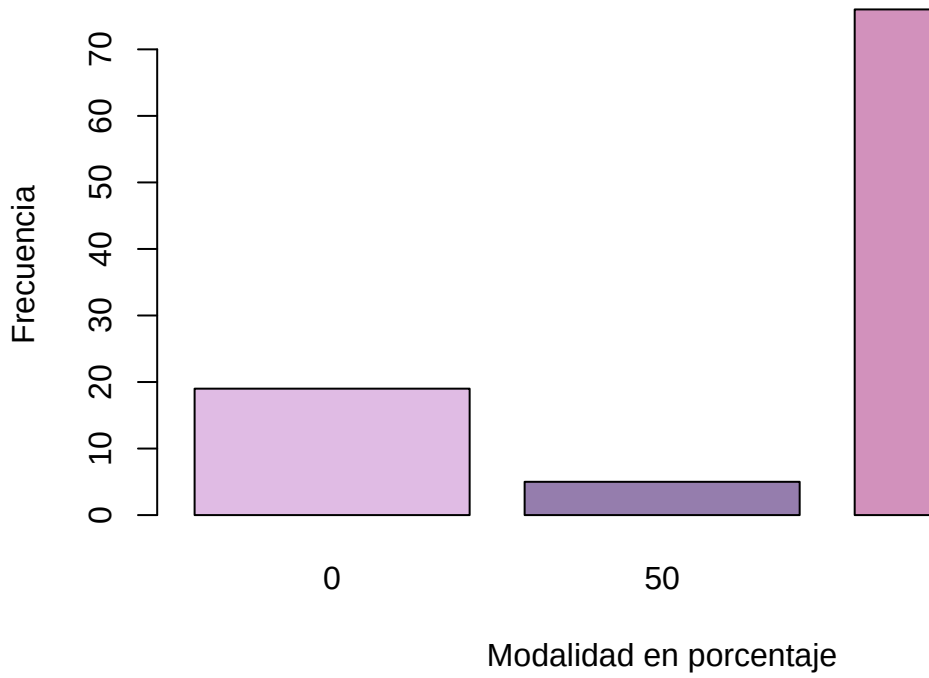
```
## [1] "Promedio esperado de salario de un Data Analyst"
```

```
## [1] 92893.06
```

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

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

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



¿Cuál es la modalidad mejor pagada?

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

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

Conclusión

De acuerdo con esta actividad, nos damos cuenta que la exploración de datos puede parecer interminable, ya que cada una de las condiciones ponen de su parte para llegar a el cálculo de un salario. En este caso, nos dedicamos a revisar cada una de ellas, las cualitativas y cuantitativas, viendo el impacto que tenían y a su vez, ayudandonos a resolver la preguntas base que planteamos desde la introducción. Para el análisis, nos basamos en la creación de filtros para determinar conclusiones concretas, en donde descubrimos que Estados Unidos es el país que más demanda tiene en origen y por ende, cuenta con mejores salarios, que un Data Analyst puede aspirar hasta ganar 92893.06 y que la modalidad mejor pagada, es presencial. Gracias a esto, nos damos una mejor idea sobre que datos indagar para posteriormente implementar un modelo estadístico base para la predicción de datos.

Anexos

Liga a github : https://github.com/A01749373/PortafolioAnalisisM1