Los Salarios

Módulo 1: Estadística para ciencia de datos
Inteligencia Artificial Avanzada para la Ciencia de Datos
Angel Corrales Sotelo
A01562052
a01562052@tec.mx
Grupo 102

Repositorio con todos los procedimientos:
https://github.com/AngelCorso/Los-salarios

Los_salarios

September 18, 2022

1 Los salarios

2 Angel Corrales Sotelo

3 A01562052

4 B. Explora las variables y familiarízate con su significado.

```
     work_year experience_level employment_type      job_title  salary  \
602       2022               SE              FT  Data Engineer  154000
603       2022               SE              FT  Data Engineer  126000
604       2022               SE              FT   Data Analyst  129000
605       2022               SE              FT   Data Analyst  150000
606       2022               MI              FT   AI Scientist  200000

    salary_currency  salary_in_usd employee_residence  remote_ratio  \
602             USD         154000                 US           100
603             USD         126000                 US           100
604             USD         129000                 US             0
605             USD         150000                 US           100
606             USD         200000                 IN           100

    company_location company_size
602               US            M
603               US            M
604               US            M
605               US            M
606               US            L
```

4.1 1. Identifica la cantidad de datos y variables presentes.

```
Index(['work_year', 'experience_level', 'employment_type', 'job_title',
       'salary', 'salary_currency', 'salary_in_usd', 'employee_residence',
       'remote_ratio', 'company_location', 'company_size'],
      dtype='object')
```

```
work_year             607
experience_level      607
employment_type       607
job_title             607
salary                607
salary_currency       607
salary_in_usd         607
employee_residence    607
remote_ratio          607
company_location      607
company_size          607
dtype: int64
```

4.2 2. Clasifica las variables de acuerdo a su tipo y escala de medición.

4.2.1 Datos categóricos/cualitativos

Escala nominal: work_year, employment_type, job_title, salary_currency, employee_residence, company_location.

Escala ordinal: experience_level, remote_ratio, company_size.

4.2.2 Datos numéricos/cuantitativos

Escala de razón: salary, salary_in_usd.

5 C. Exploración de la base de datos

5.1 1. Calcula medidas estadísticas

5.1.1 Variables cuantitativas

- Medidas de tendencia central: promedio, media, mediana y moda de los datos.
- Medidas de dispersión: rango: máximo - mínimo, varianza, desviación estándar.

```
             salary  salary_in_usd
count  6.070000e+02     607.000000
mean   3.240001e+05  112297.869852
std    1.544357e+06   70957.259411
min    4.000000e+03    2859.000000
25%    7.000000e+04   62726.000000
50%    1.150000e+05  101570.000000
75%    1.650000e+05  150000.000000
max    3.040000e+07  600000.000000
```

```
   salary  salary_in_usd
0   80000       100000.0
1  100000            NaN
```

```
salary           2.385040e+12
salary_in_usd    5.034933e+09
dtype: float64
```

5.1.2 Variables cualitativas

- Tabla de distribución de frecuencia
- Moda

```
col_0      Freq
work_year
2020         72
2021        217
2022        318
```

```
col_0            Freq
employment_type
CT                  5
FL                  4
FT                588
PT                 10
```

```
col_0                                Freq
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
ML Engineer	6
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
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

col_0	Freq
salary_currency	
AUD	2
BRL	2
CAD	18
CHF	1
CLP	1
CNY	2
DKK	2
EUR	95
GBP	44
HUF	2
INR	27
JPY	3
MXN	2
PLN	3
SGD	2
TRY	3
USD	398

col_0	Freq
employee_residence	
AE	3
AR	1
AT	3
AU	3
BE	2
BG	1
BO	1
BR	6
CA	29
CH	1
CL	1
CN	1
CO	1
CZ	1
DE	25
DK	2
DZ	1
EE	1
ES	15
FR	18
GB	44
GR	13
HK	1
HN	1
HR	1
HU	2
IE	1
IN	30
IQ	1
IR	1
IT	4
JE	1
JP	7
KE	1
LU	1
MD	1
MT	1
MX	2
MY	1
NG	2
NL	5
NZ	1
PH	1
PK	6
PL	4

PR	1
PT	6
RO	2
RS	1
RU	4
SG	2
SI	2
TN	1
TR	3
UA	1
US	332
VN	3

col_0	Freq
company_location	
AE	3
AS	1
AT	4
AU	3
BE	2
BR	3
CA	30
CH	2
CL	1
CN	2
CO	1
CZ	2
DE	28
DK	3
DZ	1
EE	1
ES	14
FR	15
GB	47
GR	11
HN	1
HR	1
HU	1
IE	1
IL	1
IN	24
IQ	1
IR	1
IT	2
JP	6
KE	1
LU	3

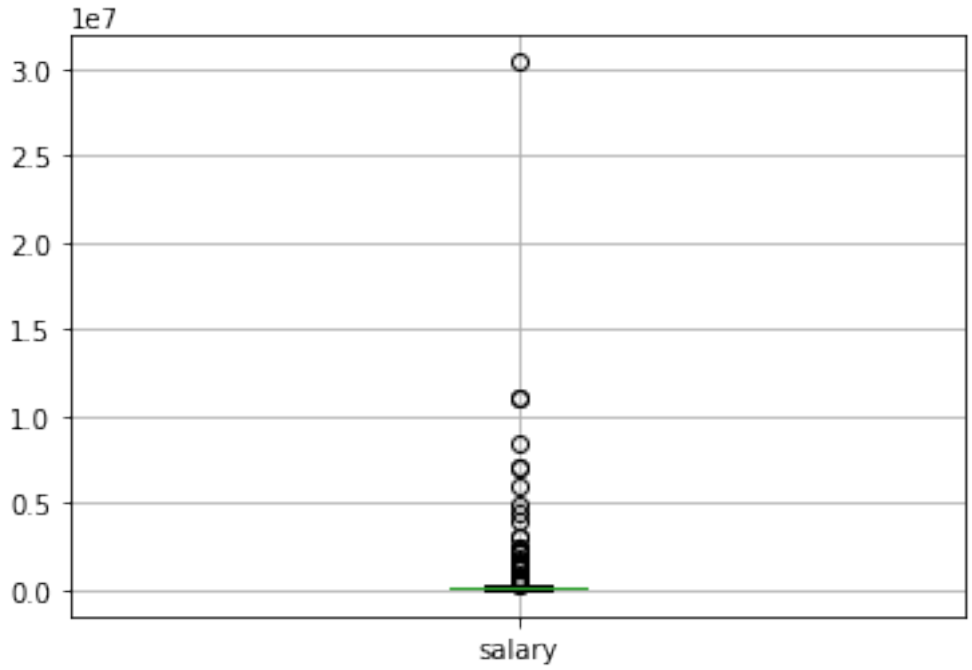
```
MD                   1
MT                   1
MX                   3
MY                   1
NG                   2
NL                   4
NZ                   1
PK                   3
PL                   4
PT                   4
RO                   1
RU                   2
SG                   1
SI                   2
TR                   3
UA                   1
US                 355
VN                   1

col_0             Freq
experience_level
EN                  88
EX                  26
MI                 213
SE                 280

col_0         Freq
remote_ratio
0              127
50              99
100            381

col_0         Freq
company_size
L              198
M              326
S               83

   work_year experience_level employment_type       job_title salary_currency  \
0       2022               SE              FT  Data Scientist             USD

  employee_residence  remote_ratio company_location company_size
0                 US           100               US            M
```

5.2 2. Explora los datos usando herramientas de visualización

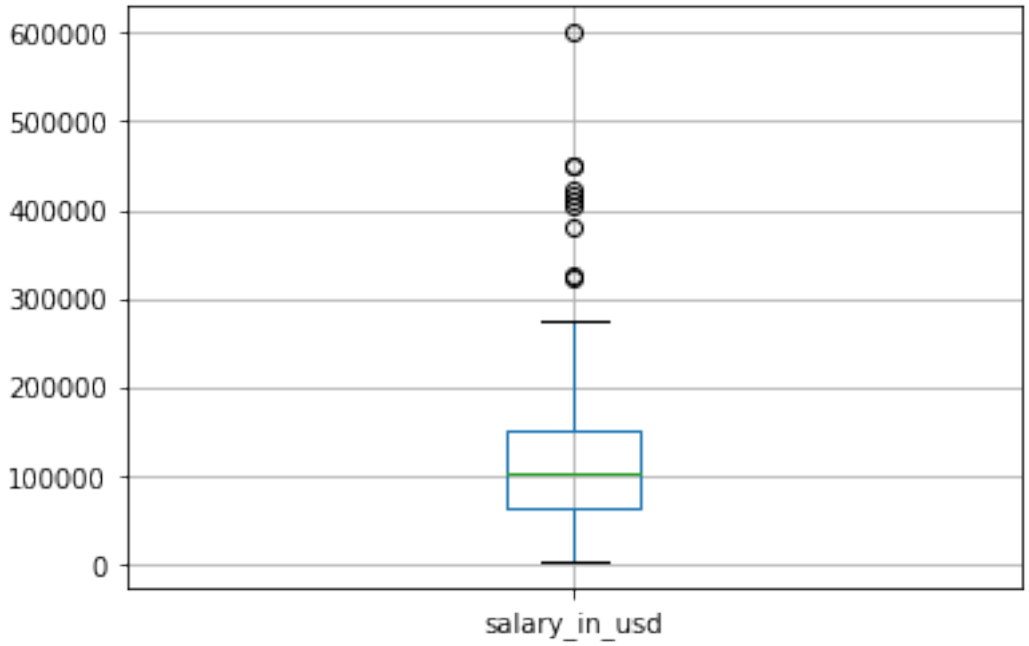
5.2.1 Variables cuantitativas:

- Medidas de posición: cuartiles, outlier (valores atípicos), boxplots
- Análisis de distribución de los datos (Histogramas). Identificar si tiene forma simétrica o asimétrica.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fdd15ba0790>
```



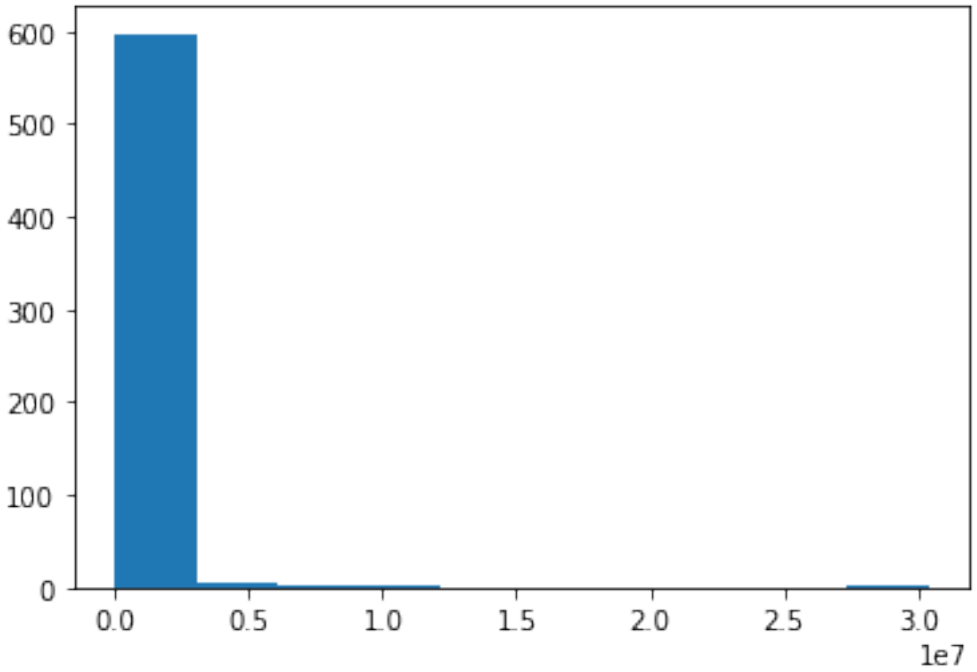

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fdd151b2850>
```

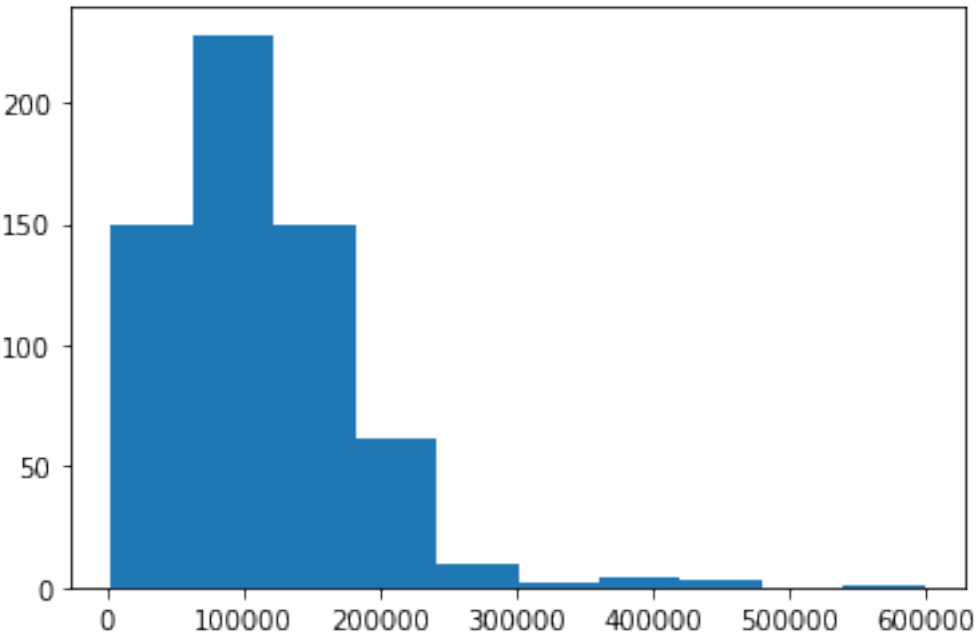



```
       salary
7    11000000
102  11000000
136   7000000
137   8500000
177  30400000
263   4900000
285   7000000
384   6000000

salary    8
dtype: int64
```

```
     salary_in_usd
1           260000
25          325000
33          450000
37          250000
63          412000
74          235000
78          270000
93          276000
97          450000
115         225000
138         220000
```

```
141         240000
157         423000
160         230000
167         250000
173         235000
224         225000
225         416000
231         256000
252         600000
309         242000
321         220110
337         243900
342         224000
398         215300
416         260000
421         241000
444         215300
472         220000
477         220000
482         324000
483         216000
486         230000
519         380000
523         405000
532         214000
534         266400
535         213120
582         220110
592         230000

salary_in_usd    40
dtype: int64
```

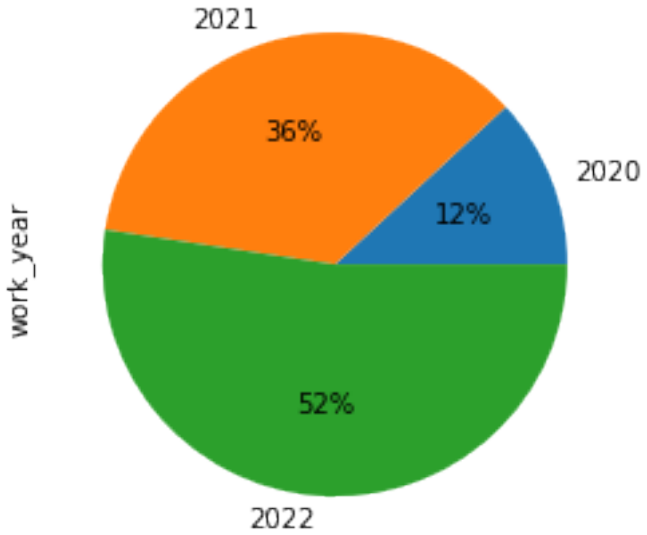


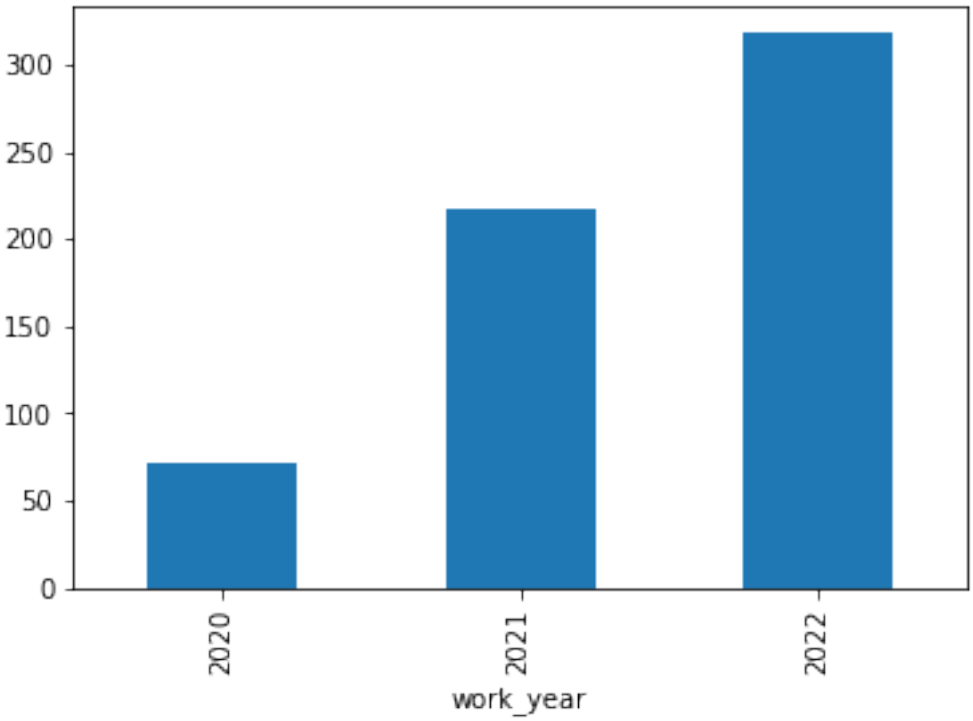


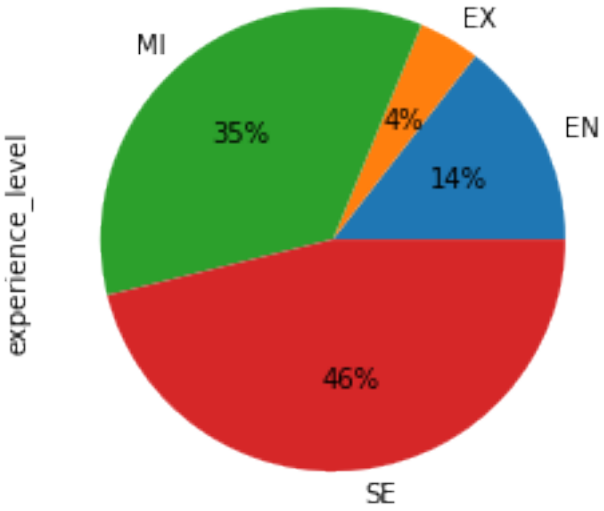
5.2.2 Variables categóricas:

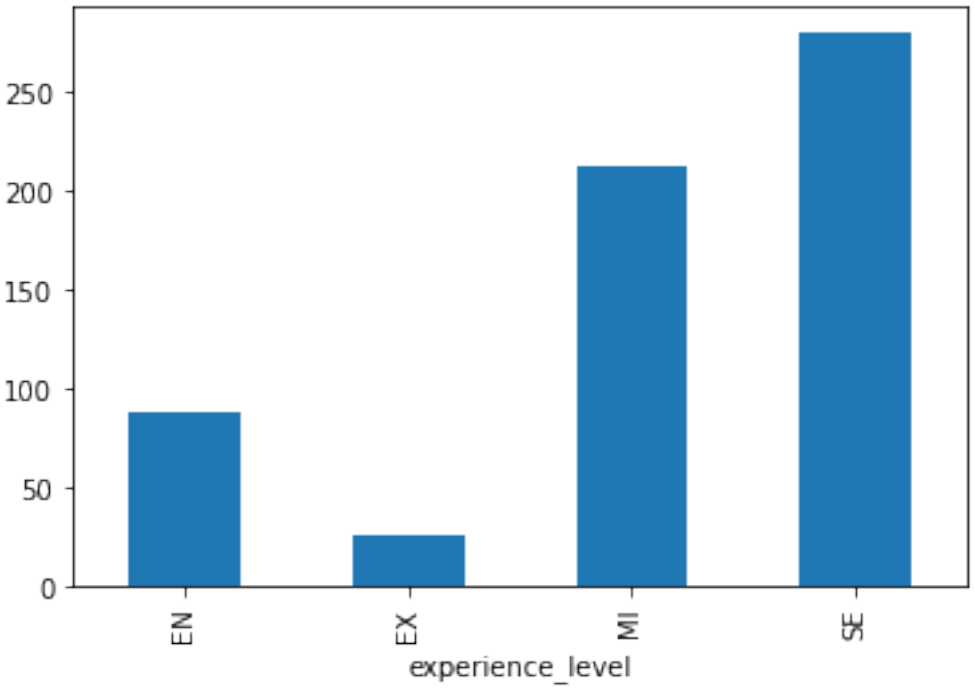
- Distribución de los datos (diagramas de barras, diagramas de pastel)

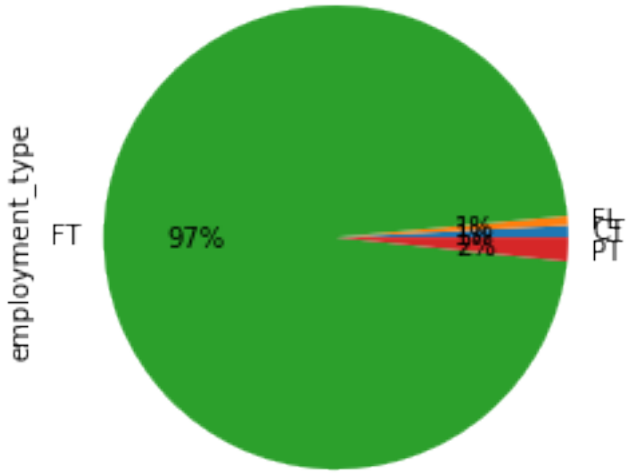
```
['work_year',
 'experience_level',
 'employment_type',
 'job_title',
 'salary_currency',
 'employee_residence',
 'remote_ratio',
 'company_location',
 'company_size']
```

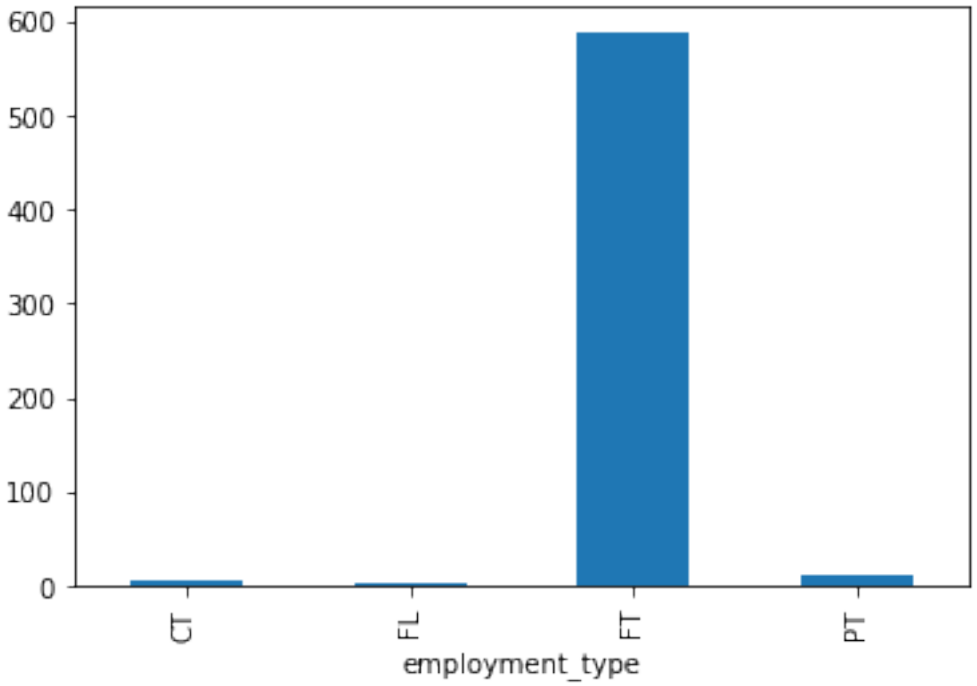


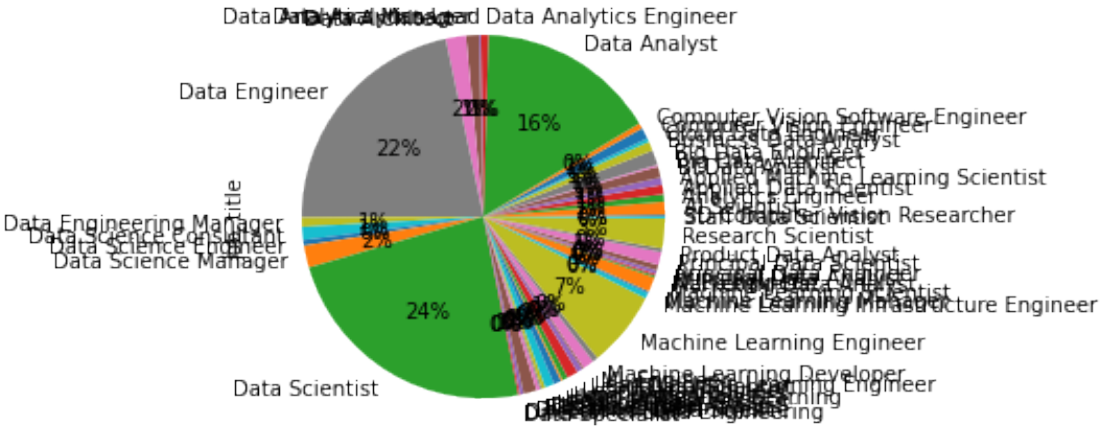

300
250
200
150
100
50
0
2020
2021
2022
work_year

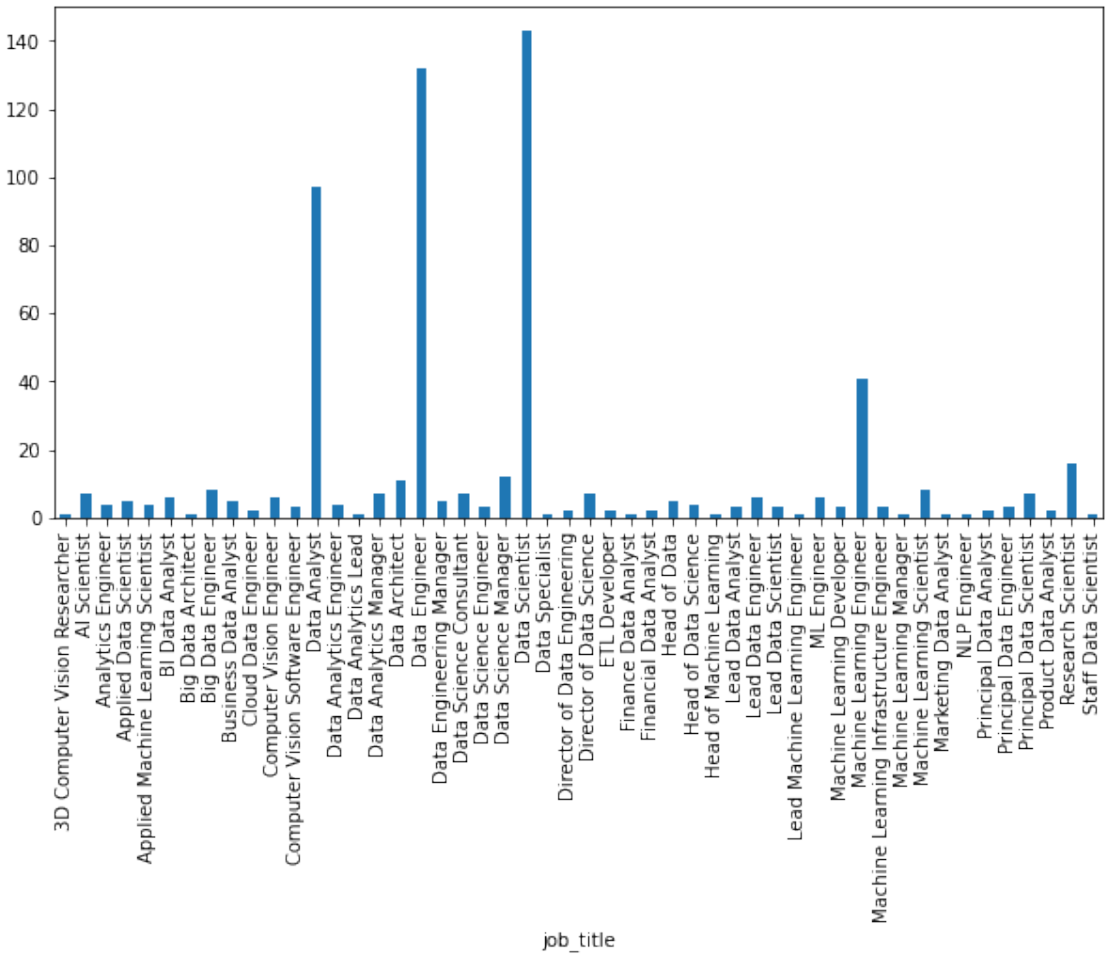

MI
EX
EN
35%
4%
14%
46%
SE
experience_level

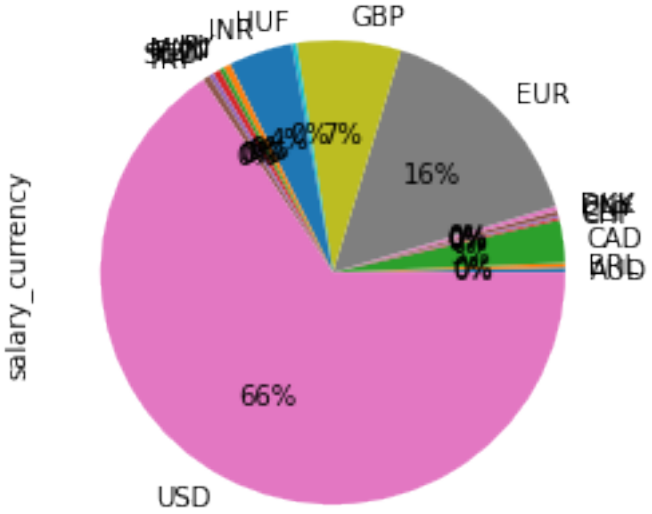

250
200
150
100
50
0
EN
EX
MI
SE
experience_level

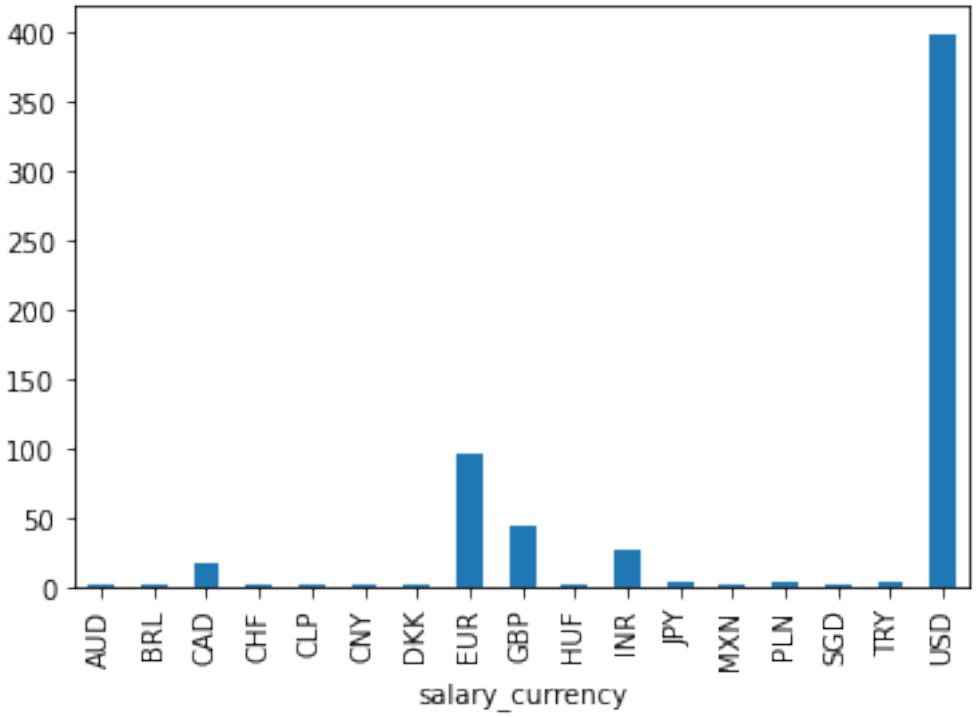

employment_type
FT
97%
3%
2%
FL
CT
PT

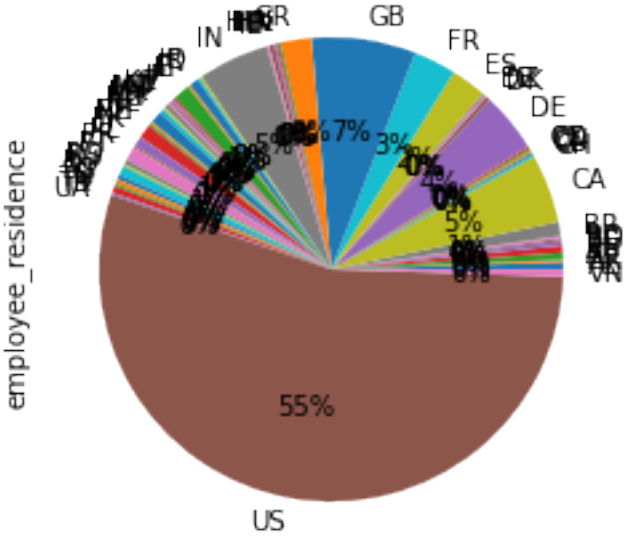

600
500
400
300
200
100
0
CT
FL
FT
PT
employment_type

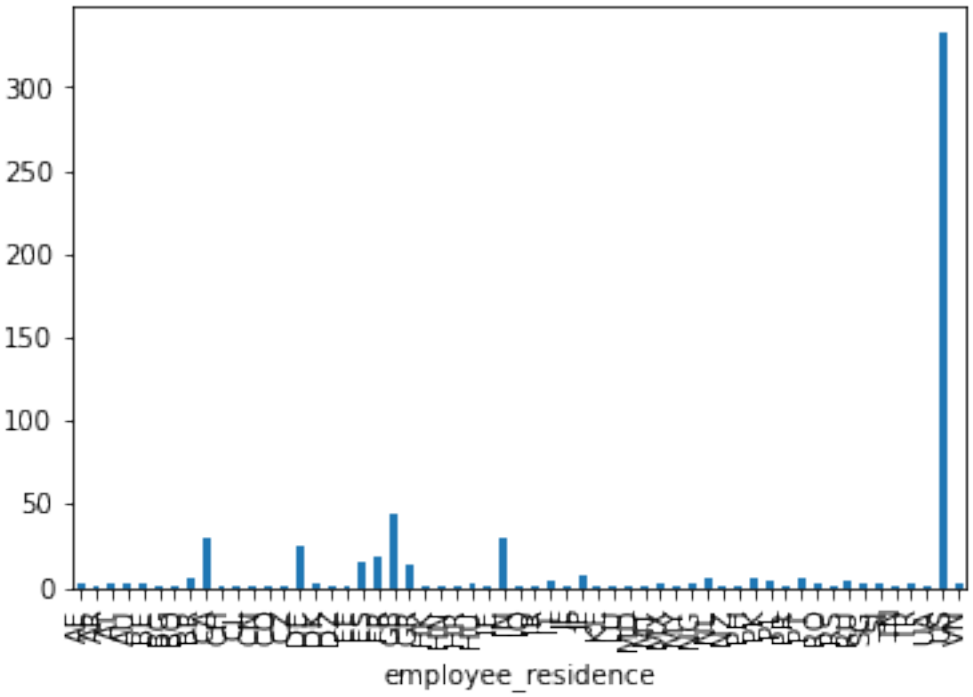

Data Analytics Engineer
Data Analyst
Data Engineer
22%
16%
Computer Vision Software Engineer
Applied Machine Learning Scientist
Applied Data Scientist
Research Scientist
Product Data Analyst
Machine Learning Infrastructure Engineer
Data Engineering Manager
Data Science Consultant
Data Science Engineer
Data Science Manager
Title
24%
7%
Machine Learning Engineer
Machine Learning Developer
Data Scientist

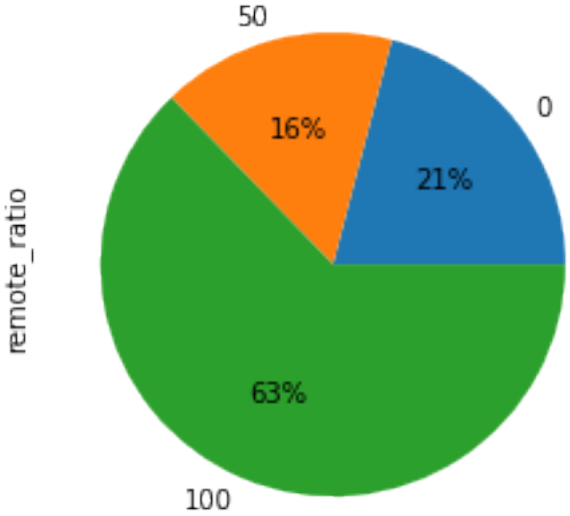

140
120
100
80
60
40
20
0
3D Computer Vision Researcher
AI Scientist
Analytics Engineer
Applied Data Scientist
Applied Machine Learning Scientist
BI Data Analyst
Big Data Architect
Big Data Engineer
Business Data Analyst
Cloud Data Engineer
Computer Vision Engineer
Computer Vision Software Engineer
Data Analyst
Data Analytics Engineer
Data Analytics Lead
Data Analytics Manager
Data Architect
Data Engineer
Data Engineering Manager
Data Science Consultant
Data Science Engineer
Data Science Manager
Data Scientist
Data Specialist
Director of Data Engineering
Director of Data Science
ETL Developer
Finance Data Analyst
Financial Data Analyst
Head of Data
Head of Data Science
Head of Machine Learning
Lead Data Analyst
Lead Data Engineer
Lead Data Scientist
Lead Machine Learning Engineer
ML Engineer
Machine Learning Developer
Machine Learning Engineer
Machine Learning Infrastructure Engineer
Machine Learning Manager
Machine Learning Scientist
Marketing Data Analyst
NLP Engineer
Principal Data Analyst
Principal Data Engineer
Principal Data Scientist
Product Data Analyst
Research Scientist
Staff Data Scientist
job_title

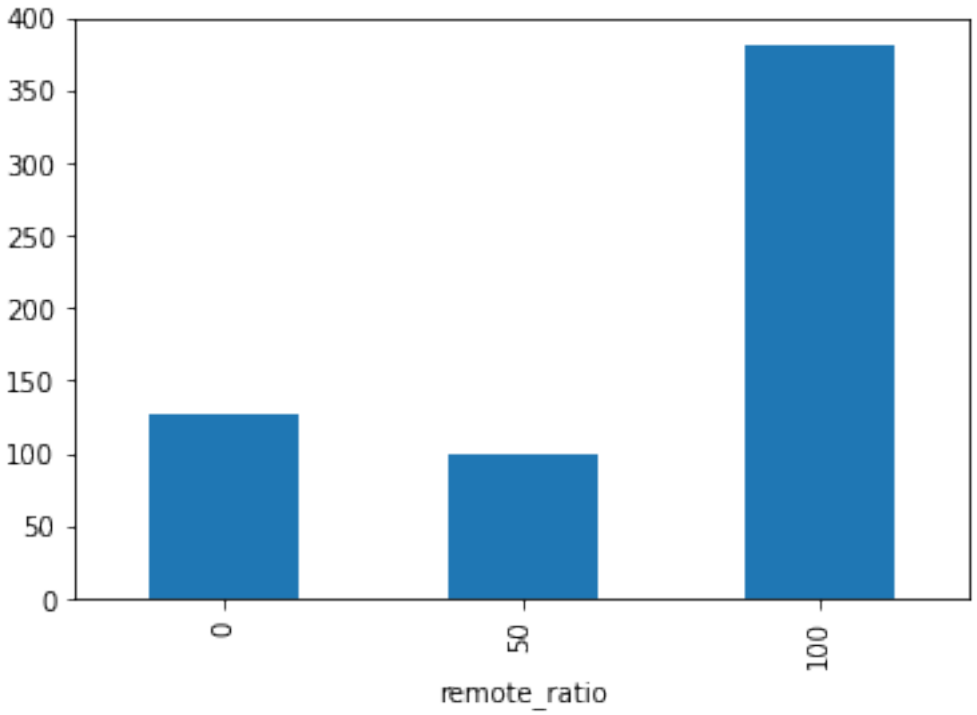

HUF
GBP
EUR
16%
CAD
66%
USD
salary_currency

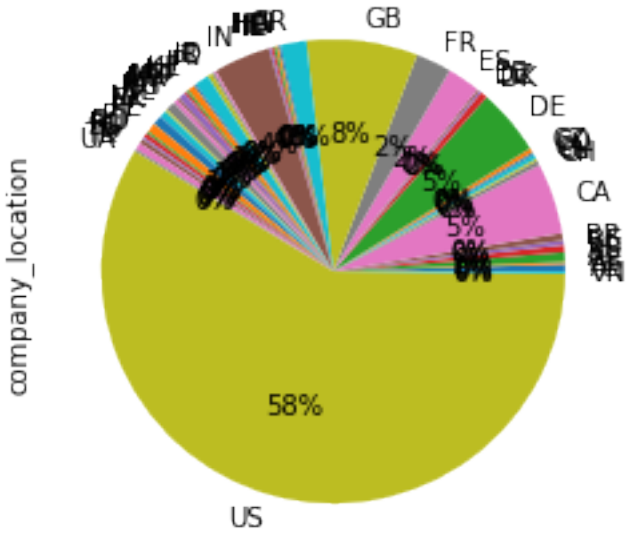

400
350
300
250
200
150
100
50
0
AUD
BRL
CAD
CHF
CLP
CNY
DKK
EUR
GBP
HUF
INR
JPY
MXN
PLN
SGD
TRY
USD
salary_currency

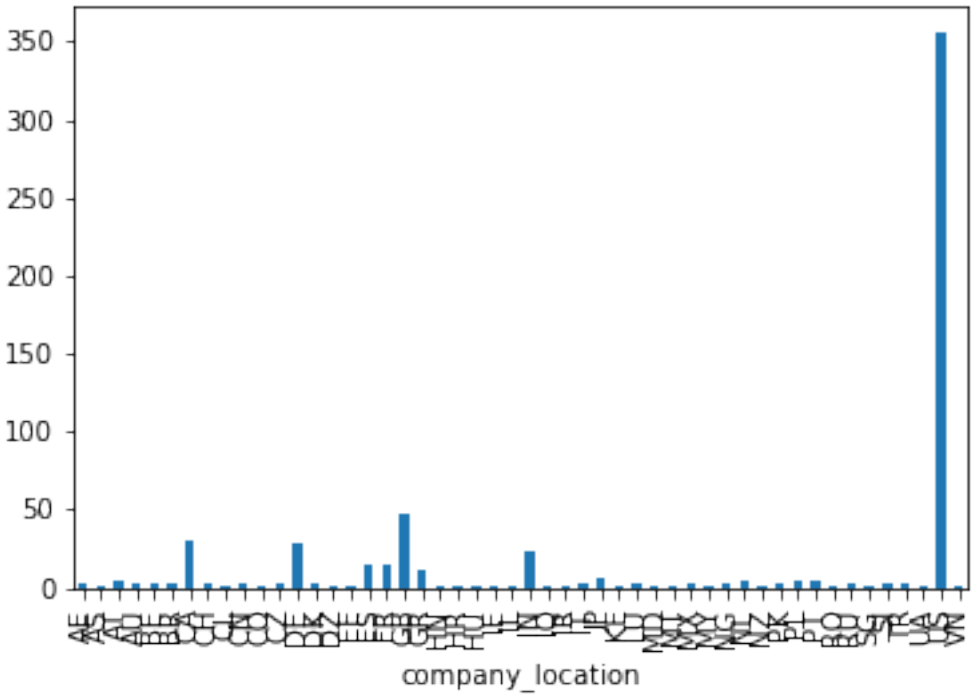

employee_residence
GB
FR
DE
CA
IN
7%
5%
55%
US

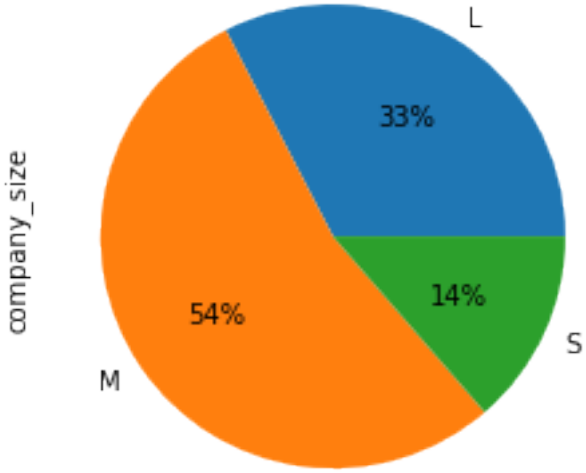

300
250
200
150
100
50
0
employee_residence

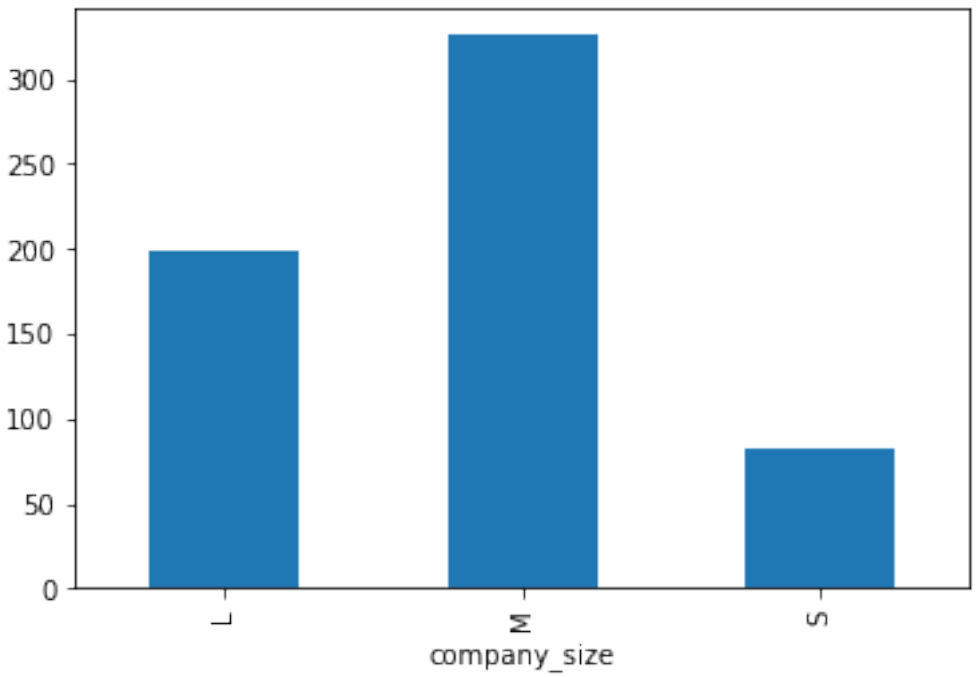

50
0
16%
21%
63%
100
remote_ratio


400
350
300
250
200
150
100
50
0
0
50
100
remote_ratio


company_location
GB
FR
DE
CA
US
58%
8%
5%
5%


350
300
250
200
150
100
50
0
company_location


company_size
L
33%
14%
S
54%
M



5.3 3. Identifica problemas de calidad de datos (registros duplicados, valores faltantes, outliers, etc).

```
work_year             0
experience_level      0
employment_type       0
job_title             0
salary                0
salary_currency       0
salary_in_usd         0
employee_residence    0
remote_ratio          0
company_location      0
company_size          0
dtype: int64
```

No hay ningún valor faltante

```
     work_year experience_level employment_type                 job_title  \
217       2021               MI              FT            Data Scientist
256       2021               MI              FT             Data Engineer
331       2022               SE              FT              Data Analyst
332       2022               SE              FT              Data Analyst
333       2022               SE              FT              Data Analyst
```

353	2022	SE	FT	Data Scientist
362	2022	SE	FT	Data Analyst
363	2022	SE	FT	Data Analyst
370	2022	SE	FT	Data Scientist
374	2022	MI	FT	ETL Developer
377	2022	SE	FT	Data Engineer
385	2022	SE	FT	Data Engineer
392	2022	SE	FT	Data Analyst
393	2022	SE	FT	Data Analyst
406	2022	MI	FT	Data Analyst
438	2022	SE	FT	Machine Learning Engineer
439	2022	SE	FT	Machine Learning Engineer
443	2022	MI	FT	Data Engineer
446	2022	SE	FT	Data Engineer
447	2022	SE	FT	Data Engineer
473	2022	SE	FT	Data Scientist
527	2022	SE	FT	Data Analyst
529	2022	SE	FT	Data Analyst
536	2022	SE	FT	Data Analyst
537	2022	SE	FT	Data Engineer
545	2022	SE	FT	Data Engineer
547	2022	SE	FT	Data Engineer
551	2022	SE	FT	Data Scientist
555	2022	SE	FT	Data Engineer
566	2022	SE	FT	Data Analyst
569	2022	SE	FT	Data Scientist
571	2022	SE	FT	Data Scientist
572	2022	SE	FT	Data Analyst
574	2022	SE	FT	Data Scientist
575	2022	SE	FT	Data Scientist
576	2022	SE	FT	Data Scientist
578	2022	SE	FT	Data Engineer
587	2022	SE	FT	Data Scientist
588	2022	SE	FT	Data Analyst
592	2022	SE	FT	Data Scientist
596	2022	SE	FT	Data Scientist
597	2022	SE	FT	Data Analyst

	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	\
217	76760	EUR	90734	DE	50	
256	200000	USD	200000	US	100	
331	90320	USD	90320	US	100	
332	112900	USD	112900	US	100	
333	90320	USD	90320	US	100	
353	123000	USD	123000	US	100	
362	130000	USD	130000	CA	100	
363	61300	USD	61300	CA	100	

370	123000	USD	123000	US	100
374	50000	EUR	54957	GR	0
377	165400	USD	165400	US	100
385	132320	USD	132320	US	100
392	112900	USD	112900	US	100
393	90320	USD	90320	US	100
406	58000	USD	58000	US	0
438	189650	USD	189650	US	0
439	164996	USD	164996	US	0
443	60000	GBP	78526	GB	100
446	209100	USD	209100	US	100
447	154600	USD	154600	US	100
473	140000	USD	140000	US	100
527	135000	USD	135000	US	100
529	90320	USD	90320	US	100
536	112900	USD	112900	US	100
537	155000	USD	155000	US	100
545	115000	USD	115000	US	100
547	130000	USD	130000	US	100
551	140400	USD	140400	US	0
555	160000	USD	160000	US	100
566	170000	USD	170000	US	100
569	140000	USD	140000	US	100
571	140000	USD	140000	US	100
572	100000	USD	100000	US	100
574	210000	USD	210000	US	100
575	140000	USD	140000	US	100
576	210000	USD	210000	US	100
578	100000	USD	100000	US	100
587	140000	USD	140000	US	100
588	99000	USD	99000	US	0
592	230000	USD	230000	US	100
596	210000	USD	210000	US	100
597	170000	USD	170000	US	100

	company_location	company_size
217	DE	L
256	US	L
331	US	M
332	US	M
333	US	M
353	US	M
362	CA	M
363	CA	M
370	US	M
374	GR	M
377	US	M

385	US	M
392	US	M
393	US	M
406	US	S
438	US	M
439	US	M
443	GB	M
446	US	L
447	US	L
473	US	M
527	US	M
529	US	M
536	US	M
537	US	M
545	US	M
547	US	M
551	US	L
555	US	M
566	US	M
569	US	M
571	US	M
572	US	M
574	US	M
575	US	M
576	US	M
578	US	M
587	US	M
588	US	M
592	US	M
596	US	M
597	US	M

Registros duplicados.

```
42
```

Existen 42 registros duplicados

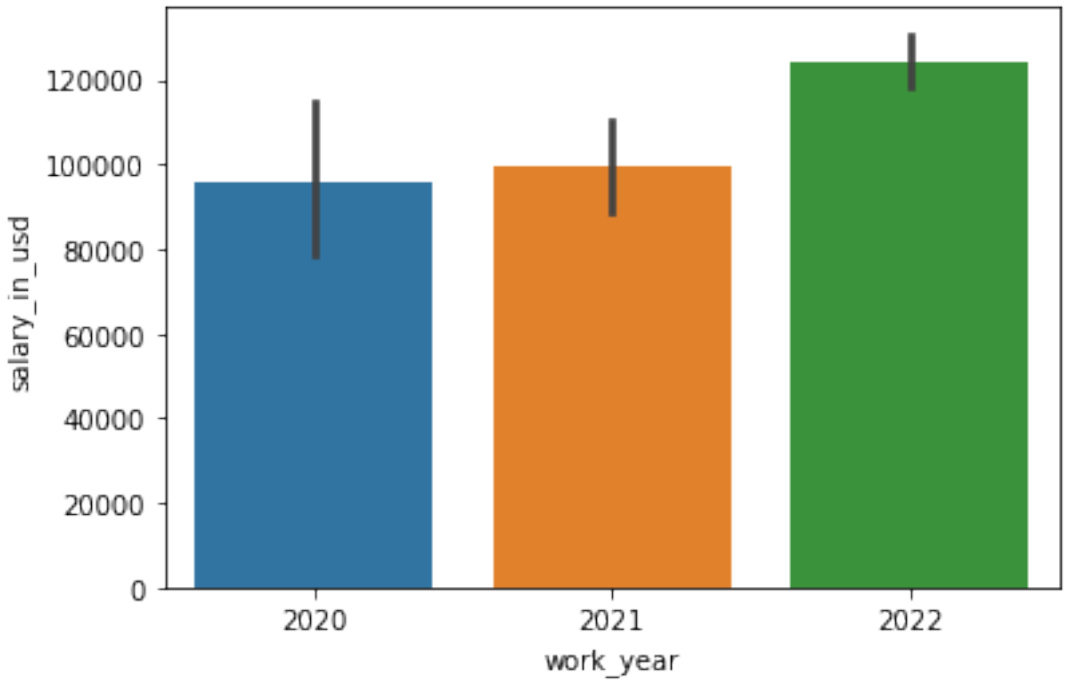
6 D. Preparación de los datos:

6.1 1. Selecciona el conjunto de datos a utilizar.

- Decide qué conjunto de datos se utilizará. Identifica variables objetivo. En caso necesario, explica por qué se incluyeron o excluyeron variables.
- Maneja datos categóricos: transforma a datos numéricos si es necesario.
- En caso de necesidad de recorte de datos (atípicos, faltantes, duplicados, etc), explica el motivo de la reducción.

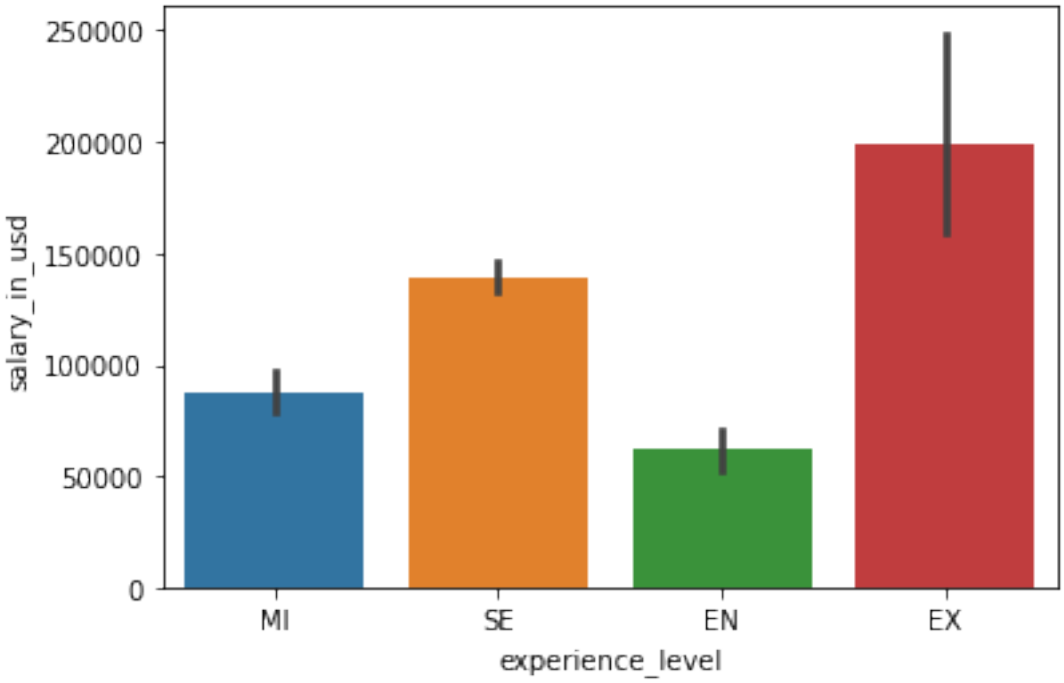
- Maneja apropiadamente datos atípicos.

```
   work_year experience_level employment_type                 job_title  \
0       2020               MI              FT              Data Scientist
1       2020               SE              FT  Machine Learning Scientist
2       2020               SE              FT           Big Data Engineer
3       2020               MI              FT        Product Data Analyst
4       2020               SE              FT   Machine Learning Engineer

   salary salary_currency  salary_in_usd employee_residence  remote_ratio  \
0   70000             EUR          79833                 DE             0
1  260000             USD         260000                 JP             0
2   85000             GBP         109024                 GB            50
3   20000             USD          20000                 HN             0
4  150000             USD         150000                 US            50

  company_location company_size
0               DE            L
1               JP            S
2               GB            M
3               HN            S
4               US            L
```



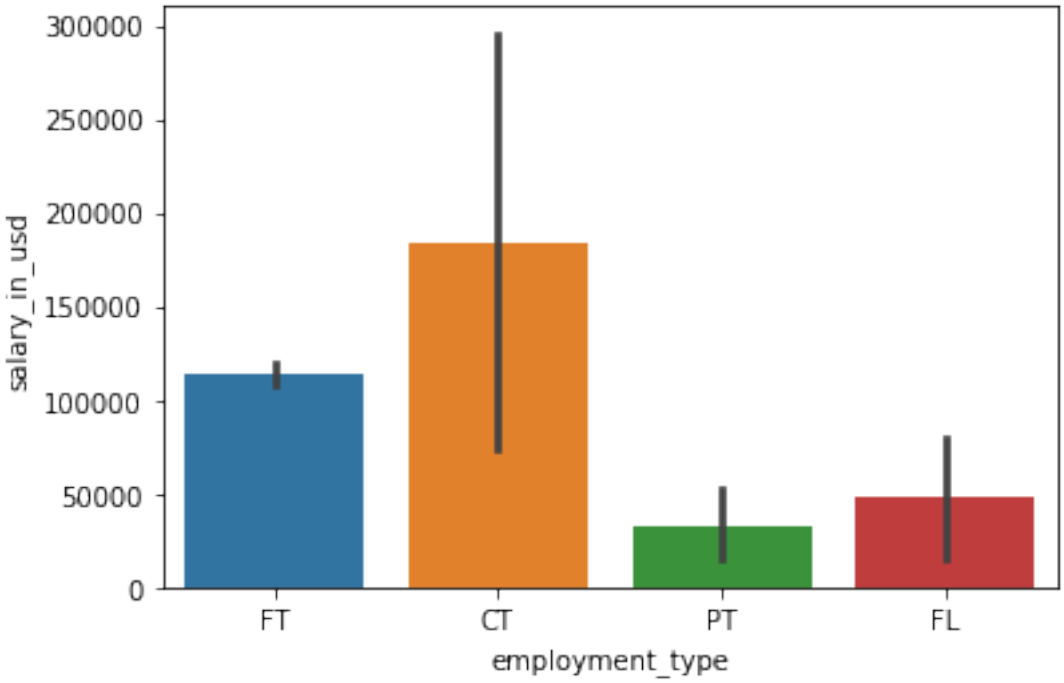
work_year. The year the salary was paid.

- El salario aumentó cada año.



Experience_level.

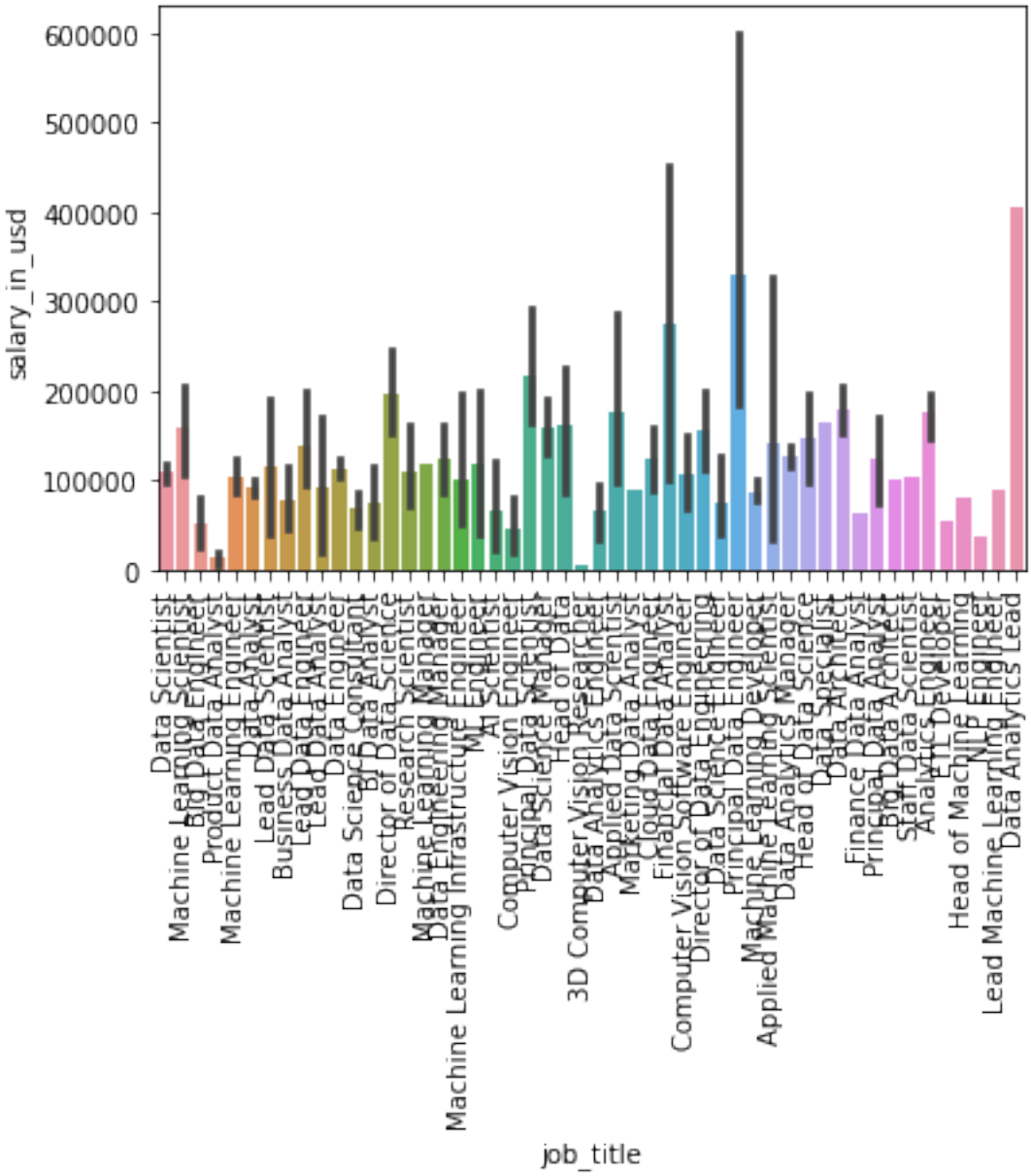
EN Entry-level / Junior MI Mid-level / Intermediate SE Senior-level / Expert EX Executive-level / Director - Parece haber una relación entre el nivel de experiencia y el salario, por lo que se decidió incluir esta variable al modelo. - Como podría esperarse, aparentemente el salario aumenta según el nivel de experiencia (a excepción del nivel Director del cual no se cuentan datos)



Employment_type.

PT Part-time FT Full-time CT Contract FL Freelance

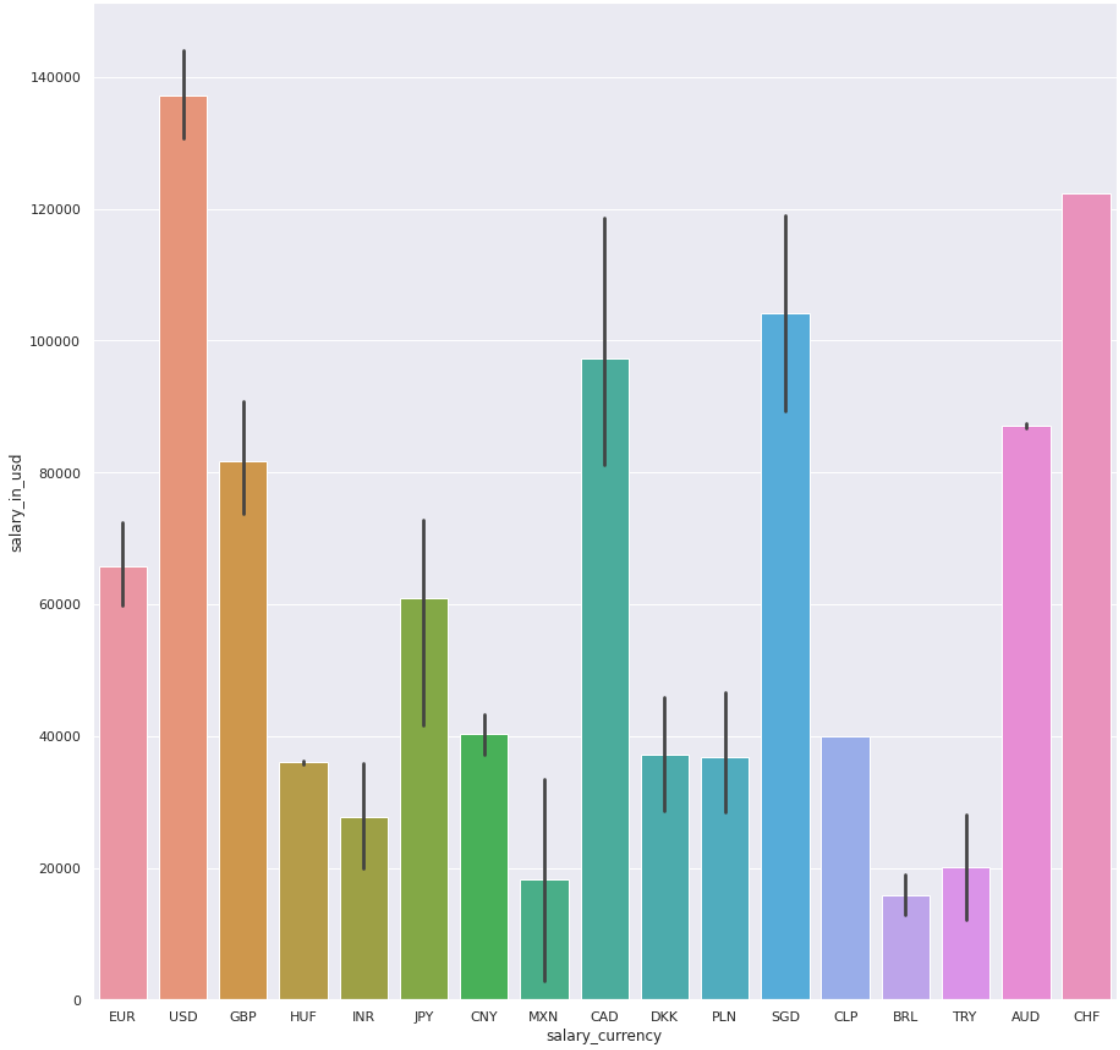
- Como podría esperarse, aparentemente el salario aumenta según el nivel de experiencia (a excepción del nivel Director del cual no se cuentan datos)



job_title.

The role worked in during the year.

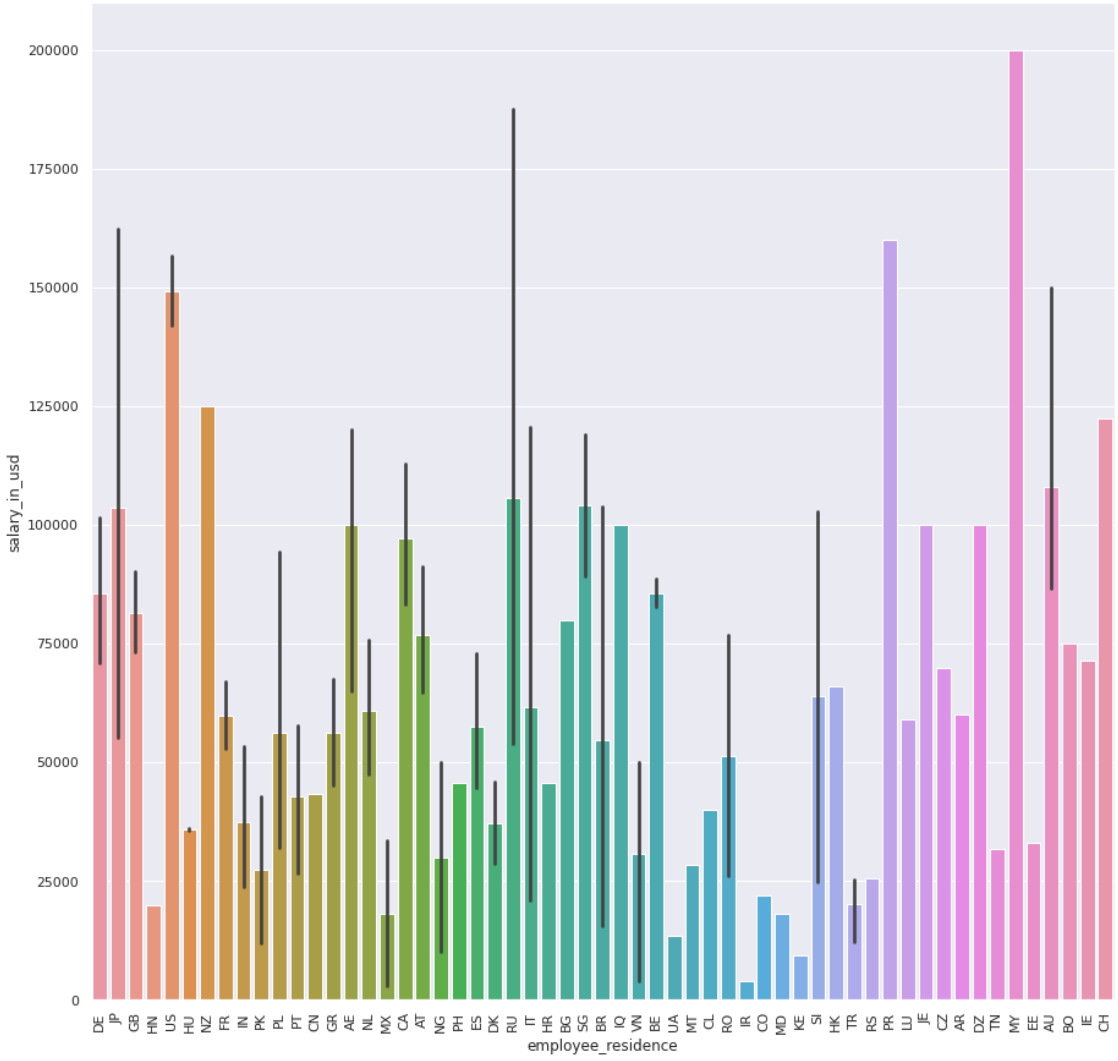
- Se observa diferencia de salarios entre los trabajos, por lo que podría ser una buena idea mantener la variable.



salary_currency.

The currency of the salary paid as an ISO 4217 currency code.

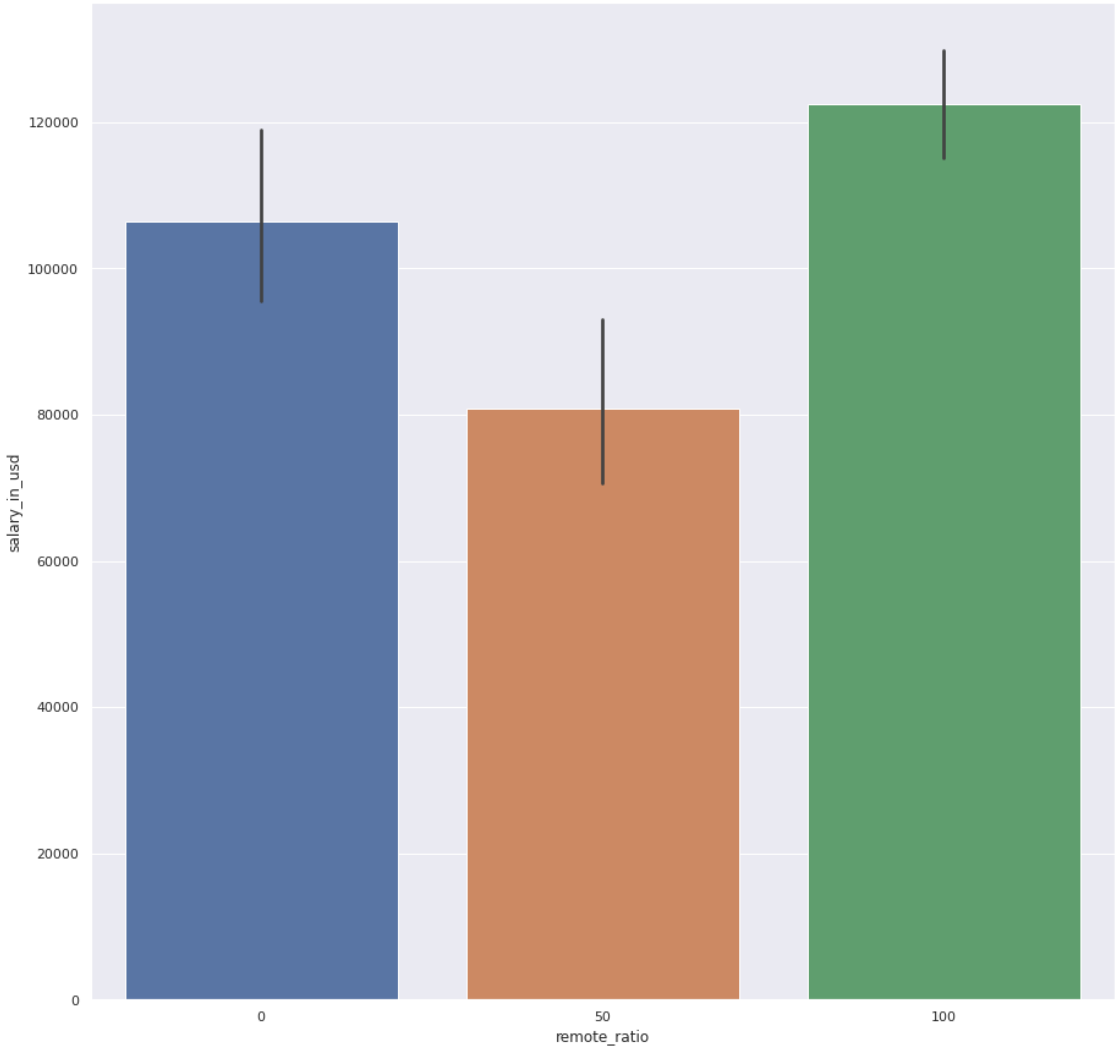
- Se observa la diferencia de salarios según el tipo de moneda, y por lo tanto la diferencia de salarios según la zona geográfica en la que se trabaje.



employee_residence.

Employee's primary country of residence in during the work year as an ISO 3166 country code.

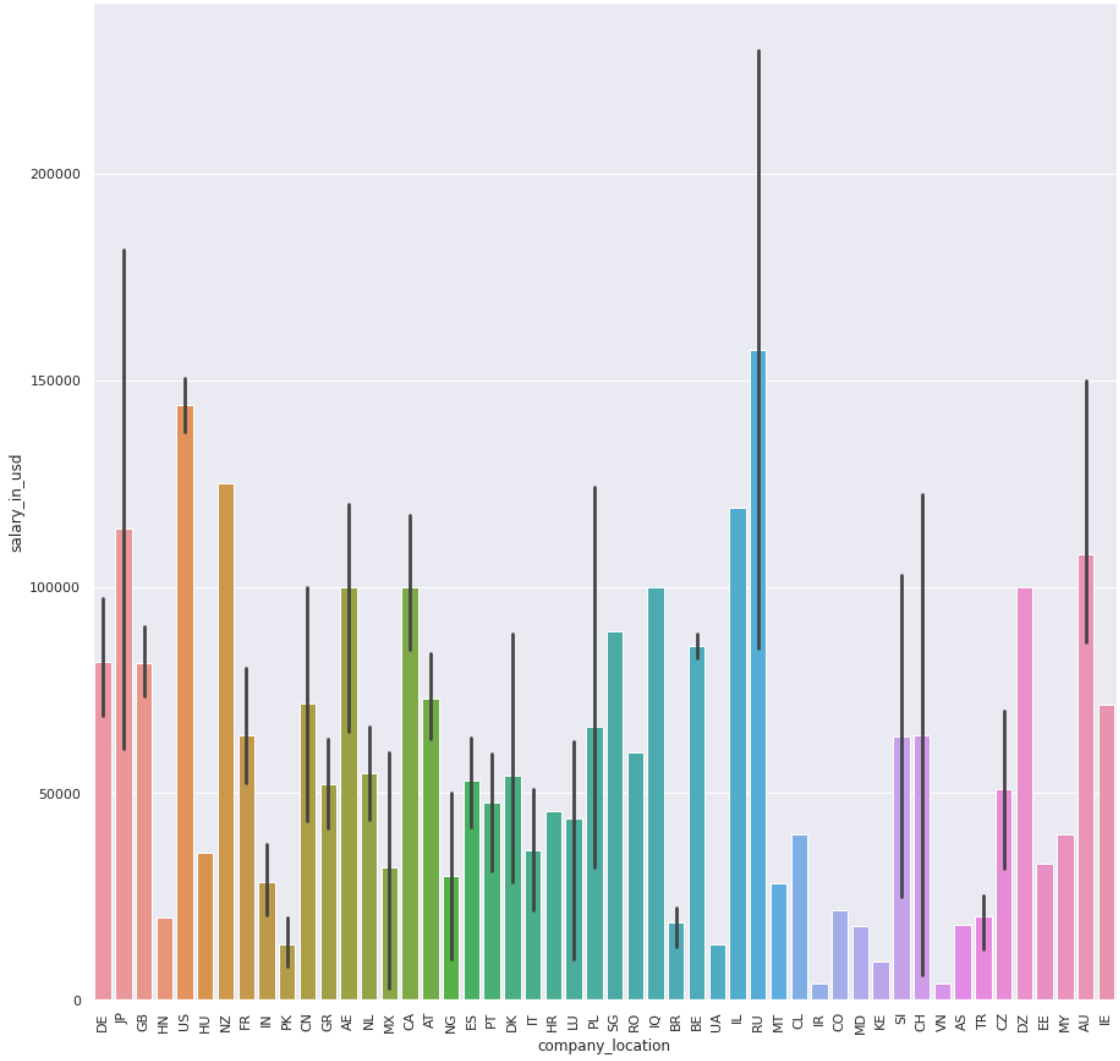
- Los intervalos tan grandes de error se pueden deber a la poca cantidad de datos que hay de esos lugares. Esta incertidumbre podría ser negativa para un modelo predictivo.



remote_ratio.

The overall amount of work done remotely, possible values are as follows: 0 No remote work (less than 20%) 50 Partially remote 100 Fully remote (more than 80%)

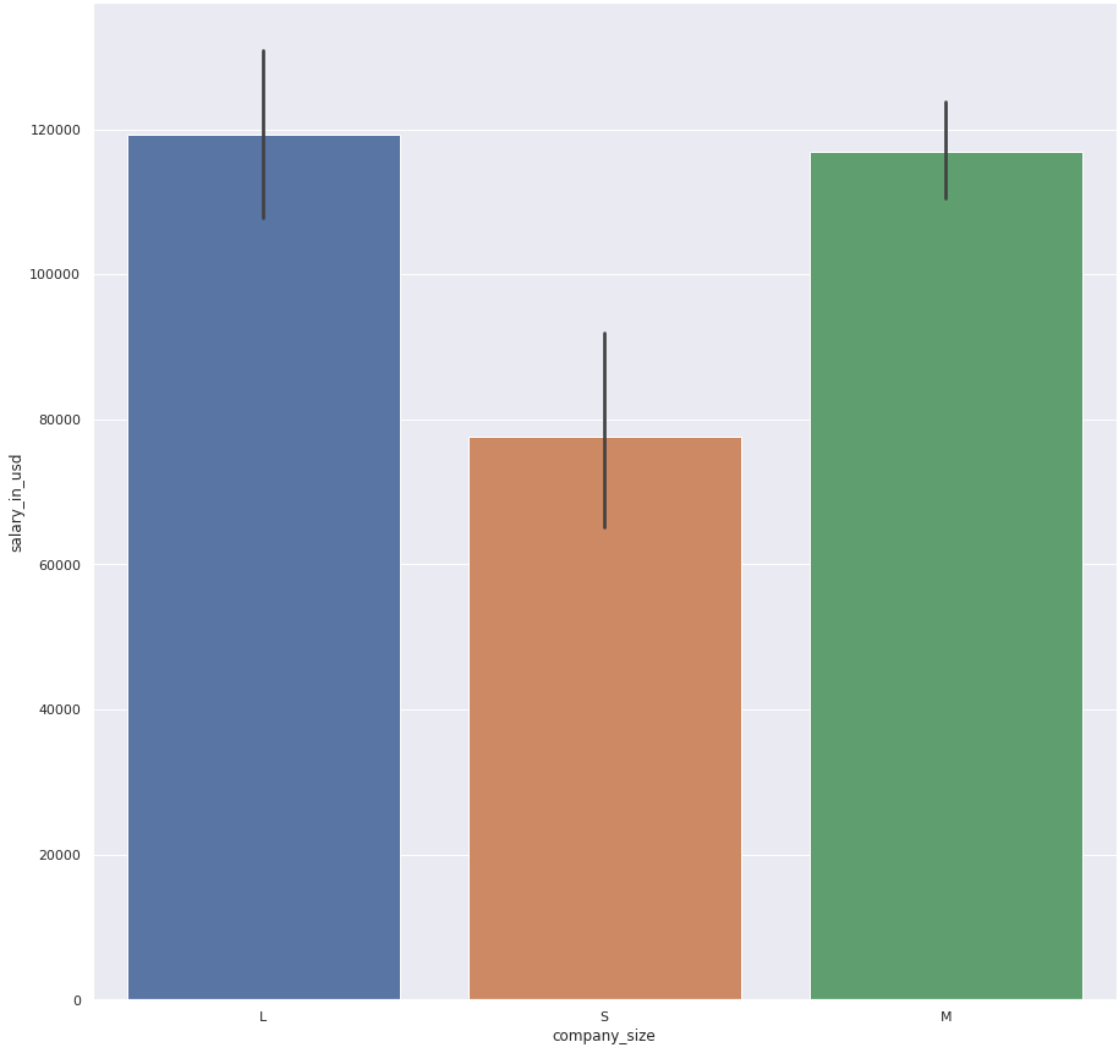
- No existe gran diferencia en los salarios dependiendo de la cantidad de trabajo realizado en remoto.
- Aparentemente los trabajos totalmente presenciales o remotos tienen mejores salrios que aquellos híbridos.



company_location.

The country of the employer's main office or contracting branch as an ISO 3166 country code.

- Existe gran diferencia entre los salarios dependiendo de la localización de la compañía, pero la confiabilidad no es tan grande debido a la gran longitud de los intervalos de error.



company_size.

The average number of people that worked for the company during the year: S less than 50 employees (small) M 50 to 250 employees (medium) L more than 250 employees (large)

- No existe diferencia significativa entre los salarios de compañías grandes o medianas.
- Se puede observar que el salario es menor en las compañías pequeñas.

Variables excluidas - salary - salary_currency

Debido a que ambos atributos denotan distintos tipos de medición, incluirlos en el modelo no es conveniente. Decidiendo únicamente dejar como atributo cuantitativo y objetivo ***salary_in_usd*** el cual representa los dos excluidos pero en una misma escala.

Variables incluidas - work_year, - experience_level, - employment_type, - job_title, - salary_currency, - employee_residence, - remote_ratio, - company_location, - company_size

Tras una comparación de las variables anteriores con salary_in_usd, todas las variables categóricas

demuestran distinción de salario en cada uno de sus niveles, por lo que se decidió utilizarlos.

Variable Objetivo - salary_in_usd

Se decidió utilizar salary_in_usd como variable objetivo con el objetivo de encontrar las condiciones que hacen que una persona especialista en analizar datos tenga un mejor suelo.

Outliers

```
     work_year experience_level employment_type  \
33        2020               MI              FT
63        2020               SE              FT
97        2021               MI              FT
157       2021               MI              FT
225       2021               EX              CT
252       2021               EX              FT
519       2022               SE              FT
523       2022               SE              FT

                            job_title  salary salary_currency  \
33                  Research Scientist  450000             USD
63                      Data Scientist  412000             USD
97               Financial Data Analyst  450000             USD
157  Applied Machine Learning Scientist  423000             USD
225            Principal Data Scientist  416000             USD
252             Principal Data Engineer  600000             USD
519              Applied Data Scientist  380000             USD
523                 Data Analytics Lead  405000             USD

     salary_in_usd employee_residence  remote_ratio company_location  \
33          450000                 US             0                US
63          412000                 US           100                US
97          450000                 US           100                US
157         423000                 US            50                US
225         416000                 US           100                US
252         600000                 US           100                US
519         380000                 US           100                US
523         405000                 US           100                US

    company_size
33             M
63             L
97             L
157            L
225            S
252            L
519            L
523            L

Cantidad de outliers:
```

```
work_year             8
experience_level      8
employment_type       8
job_title             8
salary                8
salary_currency       8
salary_in_usd         8
employee_residence    8
remote_ratio          8
company_location      8
company_size          8
dtype: int64
```

6.2 Transformación de datos

Con el propósito de preparar los datos para el entrenamiento de un modelo de inteligencia artificial, se decidió utilizar las técnicas de Ordinal Encoding y One Hot Encoding en las variables categóricas.

6.2.1 Ordinal Encoding

```
experience_level    607
company_size        607
remote_ratio        607
dtype: int64
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 607 entries, 0 to 606
Data columns (total 3 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   experience_level  607 non-null    float64
 1   company_size      607 non-null    float64
 2   remote_ratio      607 non-null    int64
dtypes: float64(2), int64(1)
memory usage: 14.4 KB
```

6.2.2 One Hot Encoding

Se decidió eliminar lo registros duplicados.

Se decidieron eliminar los valores atípicos tras identificar valores extraños con base en el contexto del problema.

6.2.3 Resultado Final

```
     salary_in_usd  experience_level  company_size  remote_ratio  year_2020  \
0            79833               2.0           0.0             0          1
1           260000               3.0           2.0             0          1
2           109024               3.0           1.0            50          1
```

```
3            20000               2.0           2.0             0          1
4           150000               3.0           0.0            50          1
..             …                  …             …              …          …
602         154000               3.0           1.0           100          0
603         126000               3.0           1.0           100          0
604         129000               3.0           1.0             0          0
605         150000               3.0           1.0           100          0
606         200000               2.0           0.0           100          0

     year_2021  year_2022  employment_CT  employment_FL  employment_FT  …  \
0            0          0              0              0              1  …
1            0          0              0              0              1  …
2            0          0              0              0              1  …
3            0          0              0              0              1  …
4            0          0              0              0              1  …
..         …          …              …              …              …  …
602          0          1              0              0              1  …
603          0          1              0              0              1  …
604          0          1              0              0              1  …
605          0          1              0              0              1  …
606          0          1              0              0              1  …

     company_location_PL  company_location_PT  company_location_RO  \
0                      0                    0                    0
1                      0                    0                    0
2                      0                    0                    0
3                      0                    0                    0
4                      0                    0                    0
..                   …                    …                    …
602                    0                    0                    0
603                    0                    0                    0
604                    0                    0                    0
605                    0                    0                    0
606                    0                    0                    0

     company_location_RU  company_location_SG  company_location_SI  \
0                      0                    0                    0
1                      0                    0                    0
2                      0                    0                    0
3                      0                    0                    0
4                      0                    0                    0
..                   …                    …                    …
602                    0                    0                    0
603                    0                    0                    0
604                    0                    0                    0
605                    0                    0                    0
606                    0                    0                    0
```

```
     company_location_TR  company_location_UA  company_location_US  \
0                      0                    0                    0
1                      0                    0                    0
2                      0                    0                    0
3                      0                    0                    0
4                      0                    0                    1
..                   …                    …                    …
602                    0                    0                    1
603                    0                    0                    1
604                    0                    0                    1
605                    0                    0                    1
606                    0                    0                    1

     company_location_VN
0                      0
1                      0
2                      0
3                      0
4                      0
..                   …
602                    0
603                    0
604                    0
605                    0
606                    0

[557 rows x 168 columns]

salary_in_usd          557
experience_level       557
company_size           557
remote_ratio           557
year_2020              557
                      …
company_location_SI    557
company_location_TR    557
company_location_UA    557
company_location_US    557
company_location_VN    557
Length: 168, dtype: int64
```

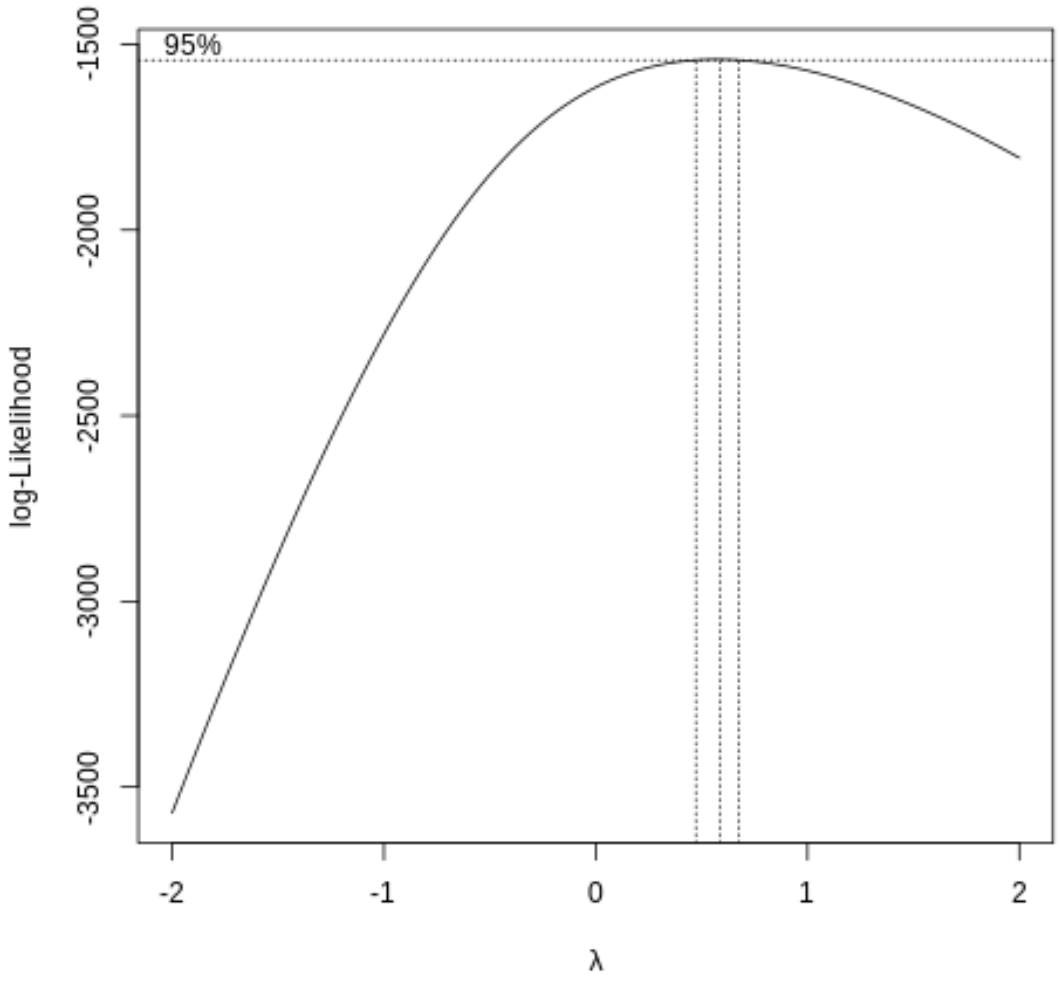
6.3 2. Transforma los datos en caso necesario.

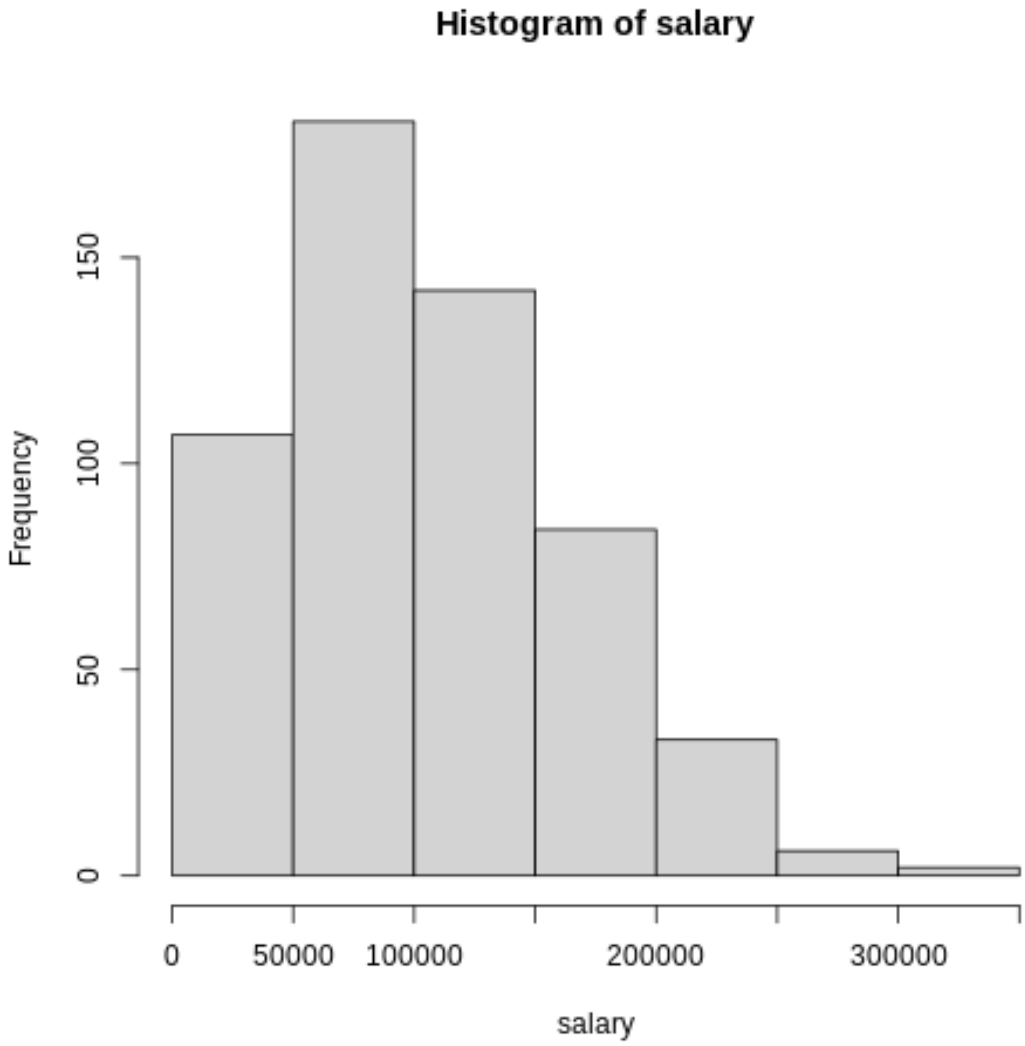
- Revisa si es necesario discretizar los datos
- Revisa si es necesario escalar y normalizar los datos
- Construye atributos si es conveniente

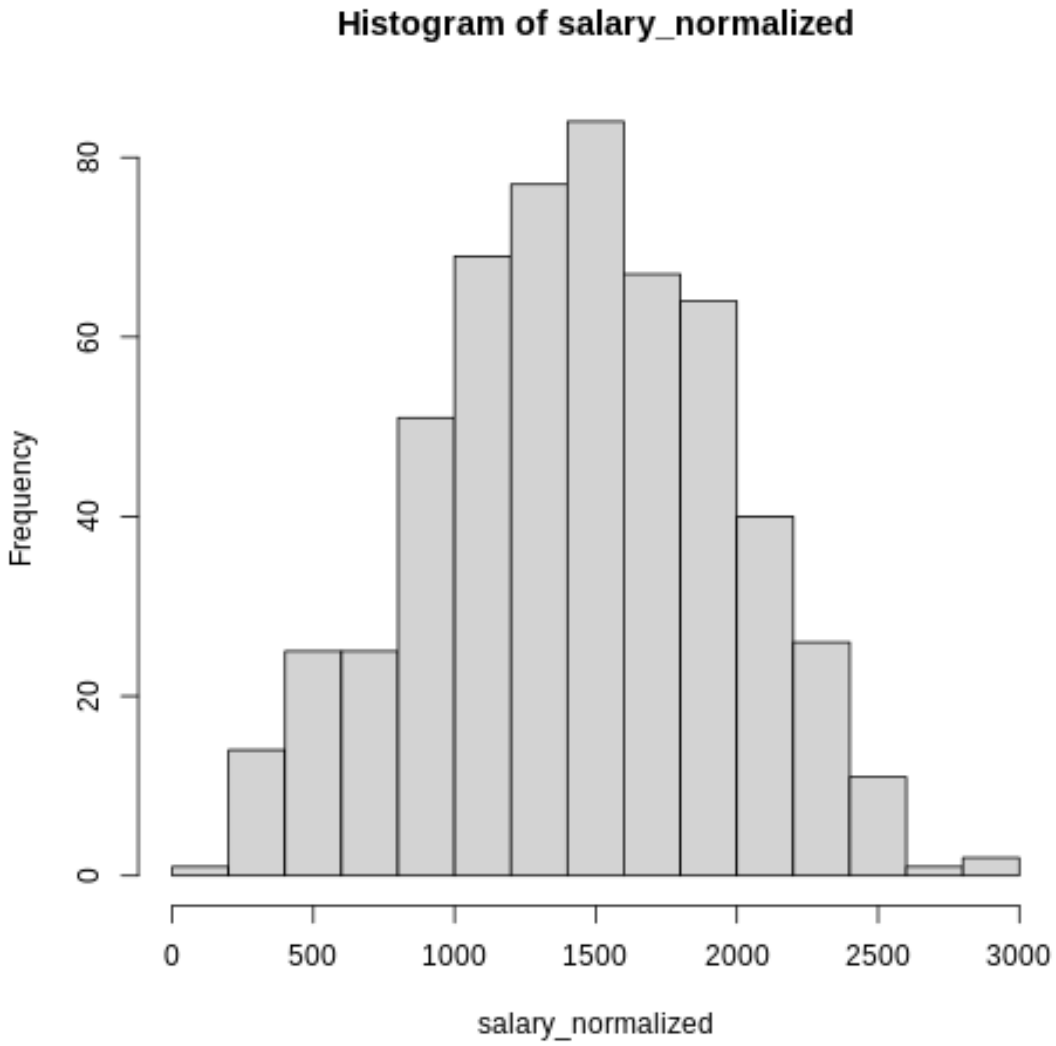
Discretizar No es necesario discretizar los datos debido a que solo contamos con una variable cuantitativa la cual es nuestra variable objetivo

Normalizar - Se decidió normalizar la variable objetivo salary_in_usd ya que no parece distribuirse de manera normal - Las demás variables al haber sido transformadas a variables dummies no será necesario normalizarlas.

```
[1] 0.5858586
```




Histogram of salary
Frequency
0
50
100
150
0
50000
100000
200000
300000
salary



Escalamiento

Se decidió escalar todas las variables mediante el escalamiento estándar

6.4 ANALIZA LOS DATOS Y PREGUNTAS GUÍA

- ¿En qué países se ofrecen mejores salarios?

```
                salary_in_usd
company_location
RU              157500.000000
US              144055.261972
NZ              125000.000000
```

Los mejores salarios en promedio se ofrecen en Rusia (RU), Estados Unidos (US) y Nueva Zelanda (NZ) con los valores mostrados en la tabla superior

- ¿Se han incrementado los salarios a lo largo del tiempo?

```
           salary_in_usd
work_year
2020        95813.000000
2021        99853.792627
2022       124522.006289
```

Aunque el incremento no fue mucho entre los años 2020 y 2021, se puede observar un claro incremento a través de los 3 años analizados con un cambio significativo entre el año 2021 y el año 2022

- ¿Influye el nivel de experiencia en el salario?

```
                  salary_in_usd
experience_level
EX                199392.038462
SE                138617.292857
MI                 87996.056338
EN                 61643.318182
```

Como se puede observar en los resultados y como es de esperar, el salario promedio incrementa dependiendo del nivel de experiencia.