Reporte final de Los salarios

September 18, 2022

Yoceline Aralí Mata A01562116

Inteligencia Artificial para la Ciencia de Datos

Grupo: 102

Módulo: Estadística para la Ciencia de Datos

1 EXPLORACIÓN DE LA BASE DE DATOS

Puede revisar el notebook y los datos en el siguiente link: https://github.com/AraliMata/Estadistica/tree/main/Los%20salarios

```
The rpy2.ipython extension is already loaded. To reload it, use:
  %reload_ext rpy2.ipython
```

1.1 B. Explora las variable y familiarízate con su significado

1. Identifica la cantidad de datos y variables presentes.
2. Clasifica las variables de acuerdo a su tipo y escala de medición.

```
   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 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

<class 'pandas.core.frame.DataFrame'>
Int64Index: 607 entries, 0 to 606
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   work_year           607 non-null    int64
 1   experience_level    607 non-null    object
 2   employment_type     607 non-null    object
 3   job_title           607 non-null    object
 4   salary              607 non-null    int64
 5   salary_currency     607 non-null    object
 6   salary_in_usd       607 non-null    int64
 7   employee_residence  607 non-null    object
 8   remote_ratio        607 non-null    int64
 9   company_location    607 non-null    object
 10  company_size        607 non-null    object
dtypes: int64(4), object(7)
memory usage: 56.9+ KB
```

Tipo de variables y escala de medición

- 0 work_year | cualitativa | escala nominal
- 1 experience_level | cualitativa | escala ordinal

- 2 employment_type | cualitativa | escala nominal
- 3 job_title | cualitativa | escala nominal
- 4 salary | cuantitativa | escala de razón
- 5 salary_currency | cualitativa | escala nominal
- 6 salary_in_usd | cuantitativa | escala de razón
- 7 employee_residence | cualitativa | escala nominal
- 8 remote_ratio | cualitativa | escala **nominal u ordinal**
- 9 company_location | cualitativa | escala nominal
- 10 company_size | cualitativa | escala ordinal

1.2 Exploración de la base de datos

1.2.1 1. Calcula medidas estadísticas

Variables cuantitativas Medidas de tendencia central

Medidas de dispersión

```
             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
```

Variables cualitativas

- Tabla de distribución de frecuencia

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

```
col_0             frequency
experience_level
EN                       88
EX                       26
MI                      213
SE                      280
```

col_0	frequency
employment_type	
CT	5
FL	4
FT	588
PT	10

col_0	frequency
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
```

salary_currency	frequency
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

employee_residence	frequency
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	frequency
remote_ratio	
0	127
50	99
100	381

col_0	frequency
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         frequency
company_size
L                   198
M                   326
S                    83
```

- Moda

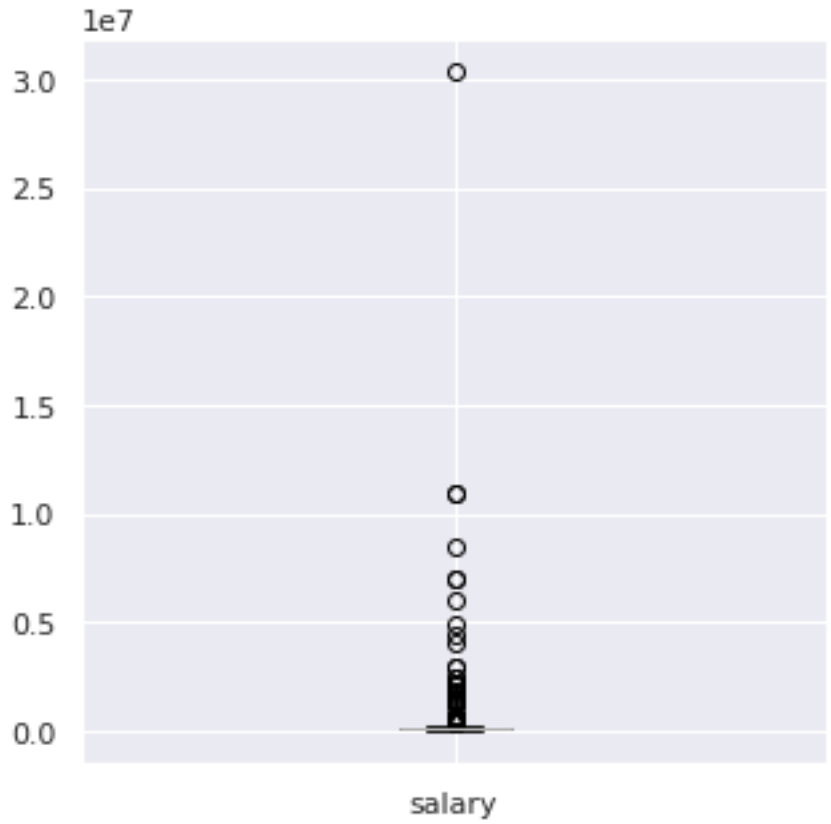
```
   work_year experience_level employment_type       job_title salary_currency  \
0       2022               SE              FT  Data Scientist             USD

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

1.2.2 2. Explora los datos usando herramientas de visualización

Variables cuantitativas

- Medidas de posición

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbbbac73910>
```




```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbbba6b9fd0>
```

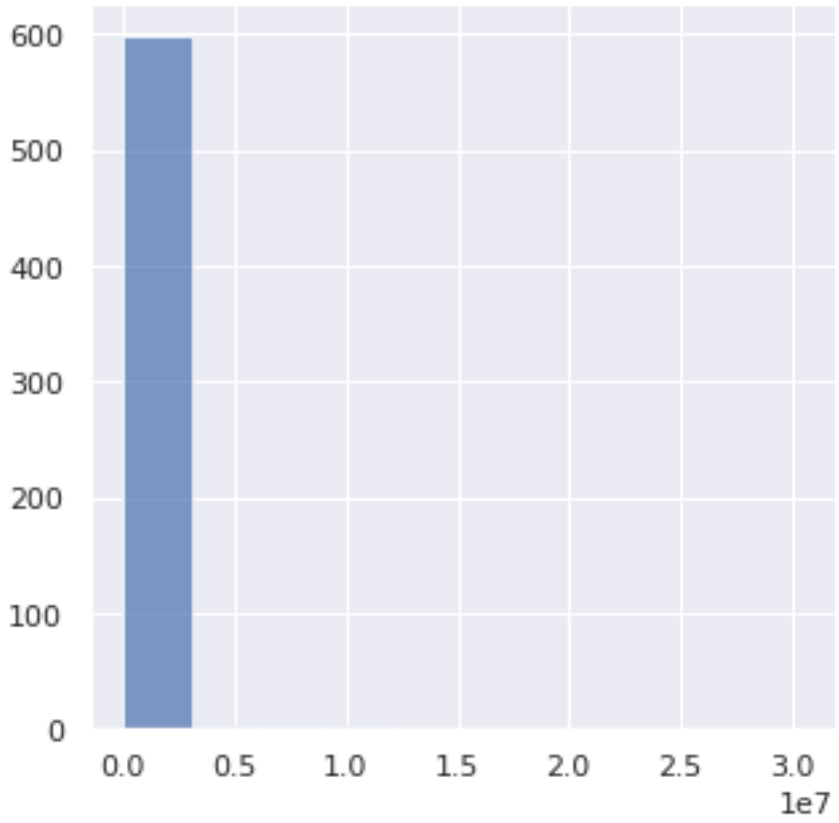
- Outliers

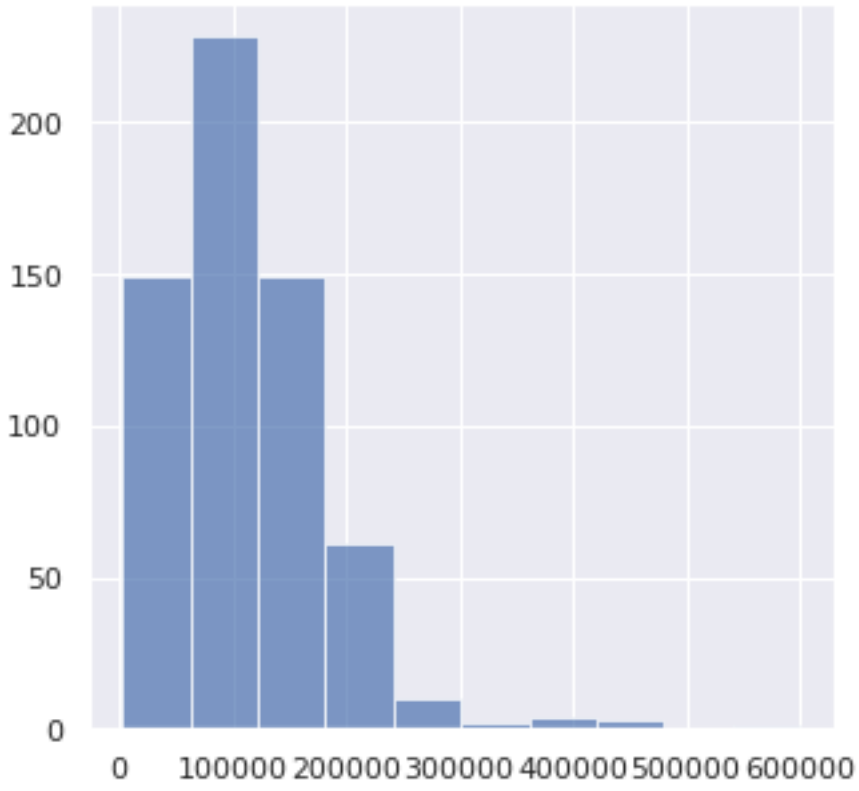
```
     work_year experience_level employment_type                 job_title  \
7         2020               MI              FT            Data Scientist
102       2021               MI              FT          BI Data Analyst
136       2021               MI              FT              ML Engineer
137       2021               MI              FT              ML Engineer
177       2021               MI              FT            Data Scientist
285       2021               SE              FT     Data Science Manager
384       2022               EX              FT  Head of Machine Learning

       salary salary_currency  salary_in_usd employee_residence  remote_ratio  \
7    11000000             HUF          35735                 HU            50
102  11000000             HUF          36259                 HU            50
136   7000000             JPY          63711                 JP            50
137   8500000             JPY          77364                 JP            50
177  30400000             CLP          40038                 CL           100
285   7000000             INR          94665                 IN            50
384   6000000             INR          79039                 IN            50

    company_location company_size
```

```
7                 HU            L
102               US            L
136               JP            S
137               JP            S
177               CL            L
285               IN            L
384               IN            L

work_year             7
experience_level      7
employment_type       7
job_title             7
salary                7
salary_currency       7
salary_in_usd         7
employee_residence    7
remote_ratio          7
company_location      7
company_size          7
dtype: int64

     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

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
```

- Histogramas


600
500
400
300
200
100
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1e7



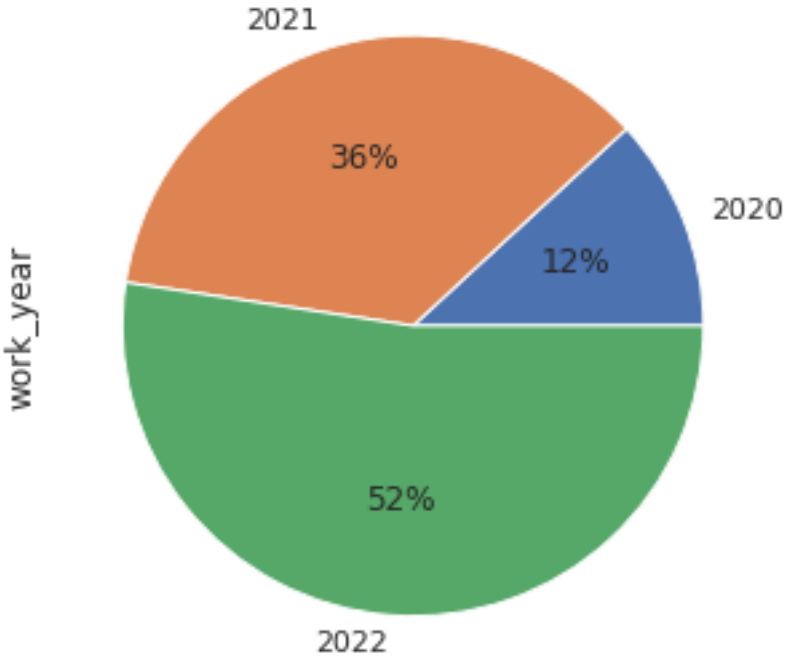
Observaciones

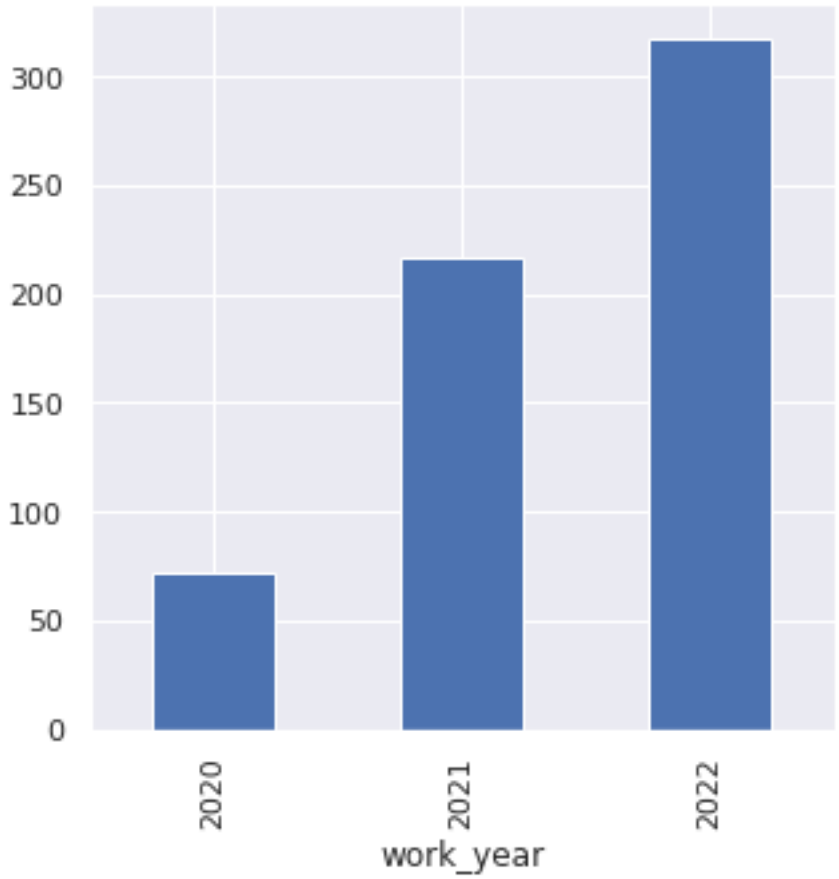
- Para ninguna de las variables de salario los histogramas tienen forma simétrica, sino que las frecuencias se encuentran hacía la izquierda.

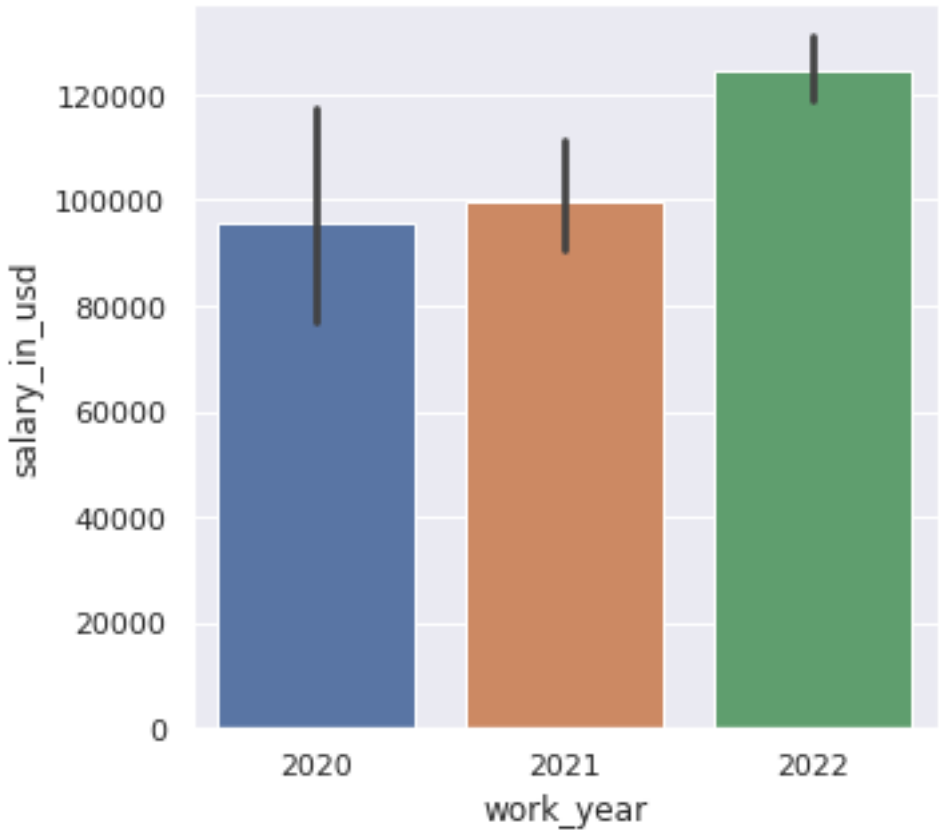
Variables categóricas

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

Work_year


2021
36%
2020
12%
work_year
52%
2022

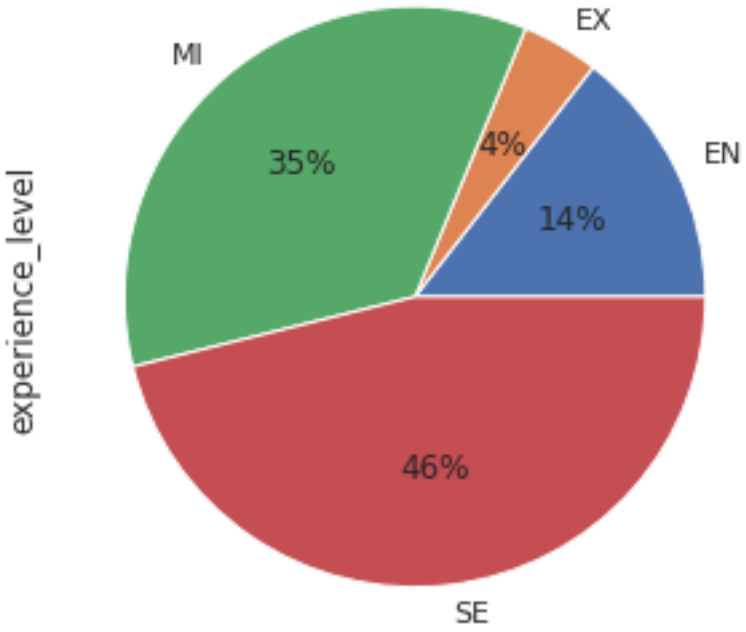

300
250
200
150
100
50
0
2020
2021
2022
work_year

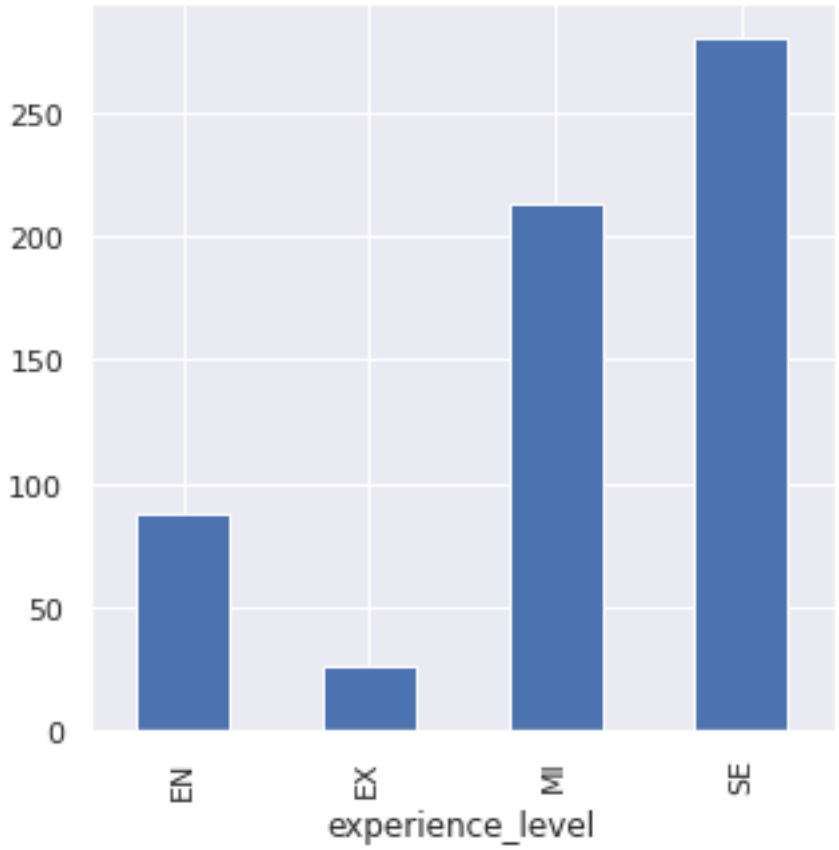


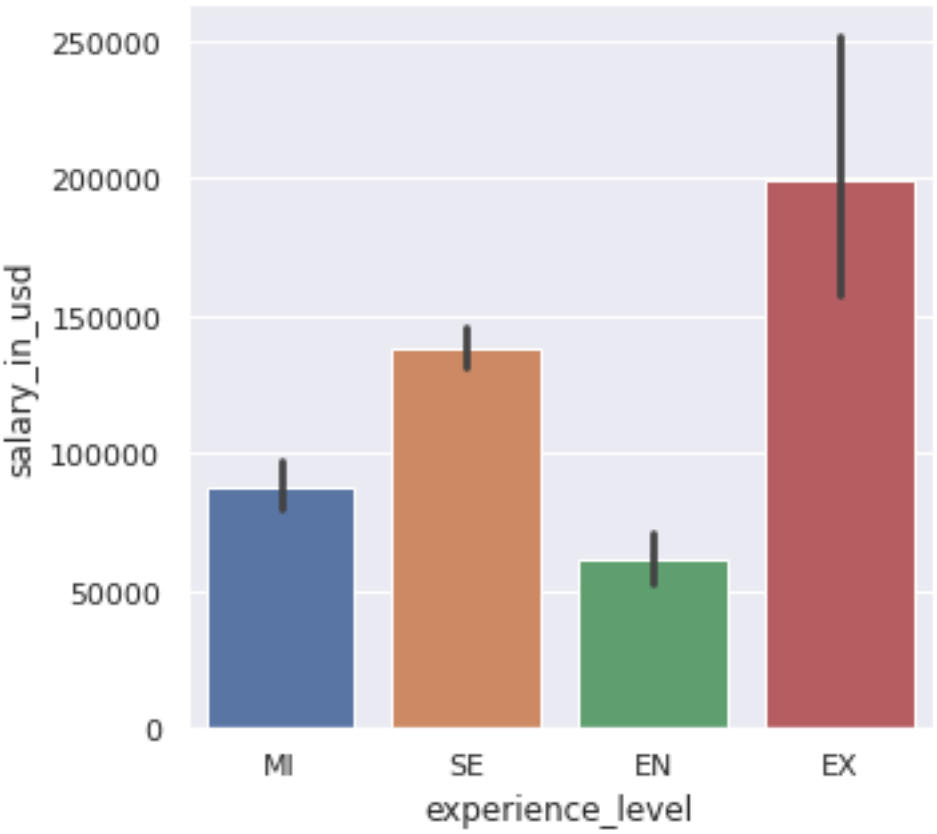
Observaciones

- La cantidad de datos de 2020 a 2022 incrementó, lo cual podría indicar que aumentó la cantidad de trabajos en el área de Ciencia de Datos.
- El promedio del salario aumentó a través de los años.

Experience_level


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

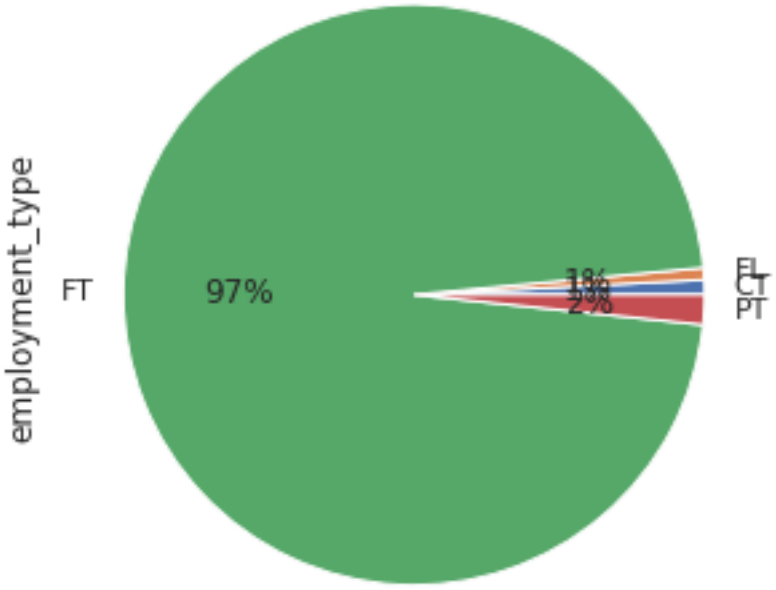

250
200
150
100
50
0
EN
EX
MI
SE
experience_level

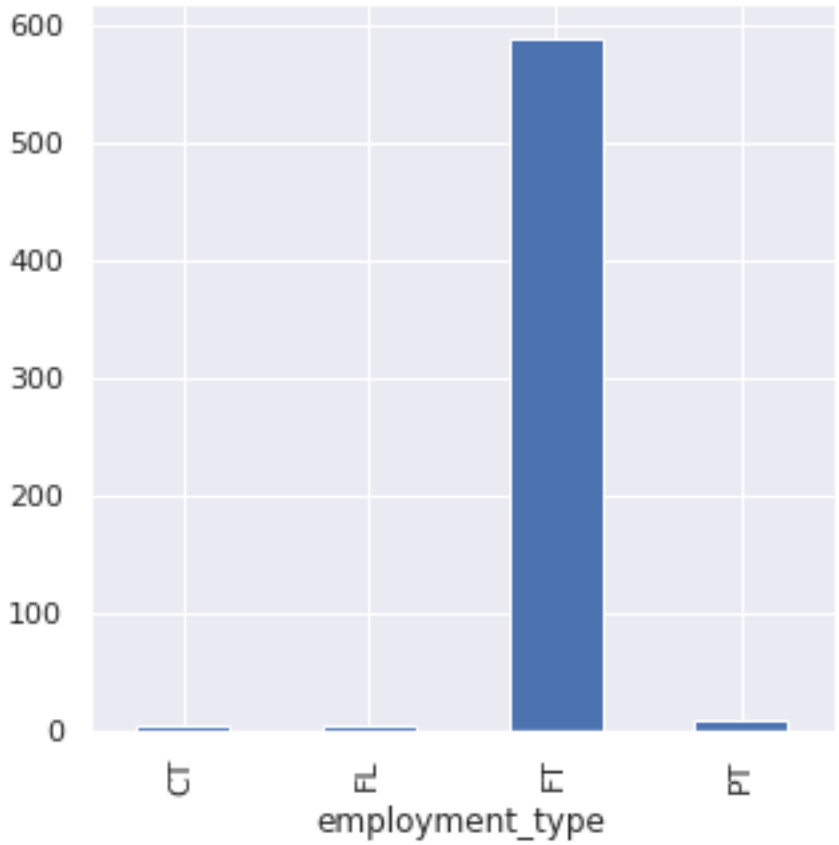


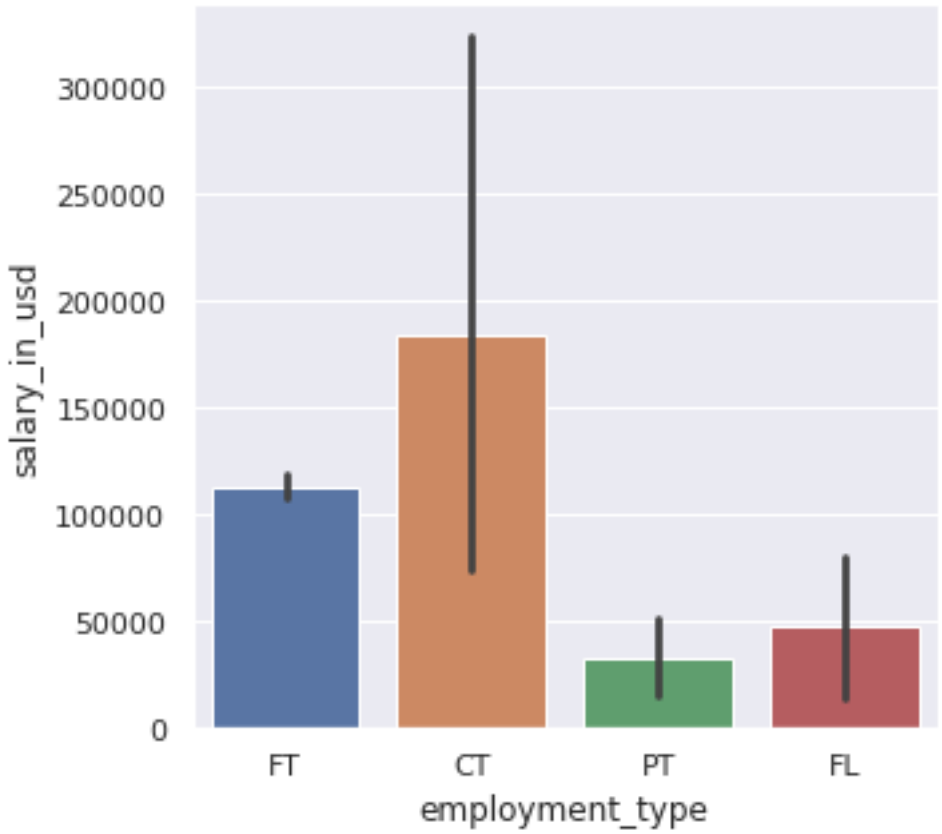
Observaciones

- Al comparar estas dos variables se observa que el salario promedio coincide con la experiencia, alguien experto gana más que un Senior que a su vez gana más que un Junior Mid-level que a su vez gana más que alguien en Entry-level.
- Por esta razón se decide considerar la variable en el entrenamiento del modelo, ya que, según los datos, el nivel de experiencia influye en el salario.

employment_type


employment_type
FT
97%
FL
CT
PT

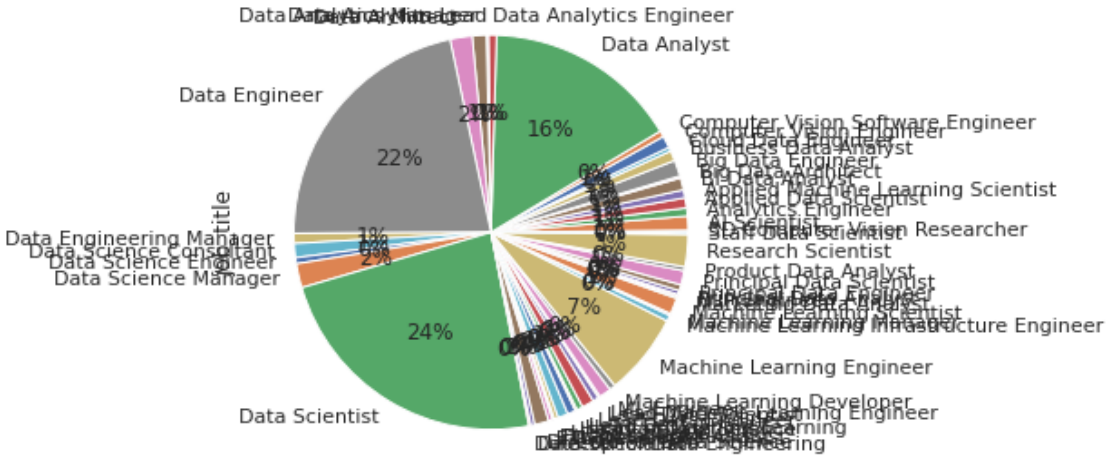

600
500
400
300
200
100
0
CT
FL
FT
PT
employment_type

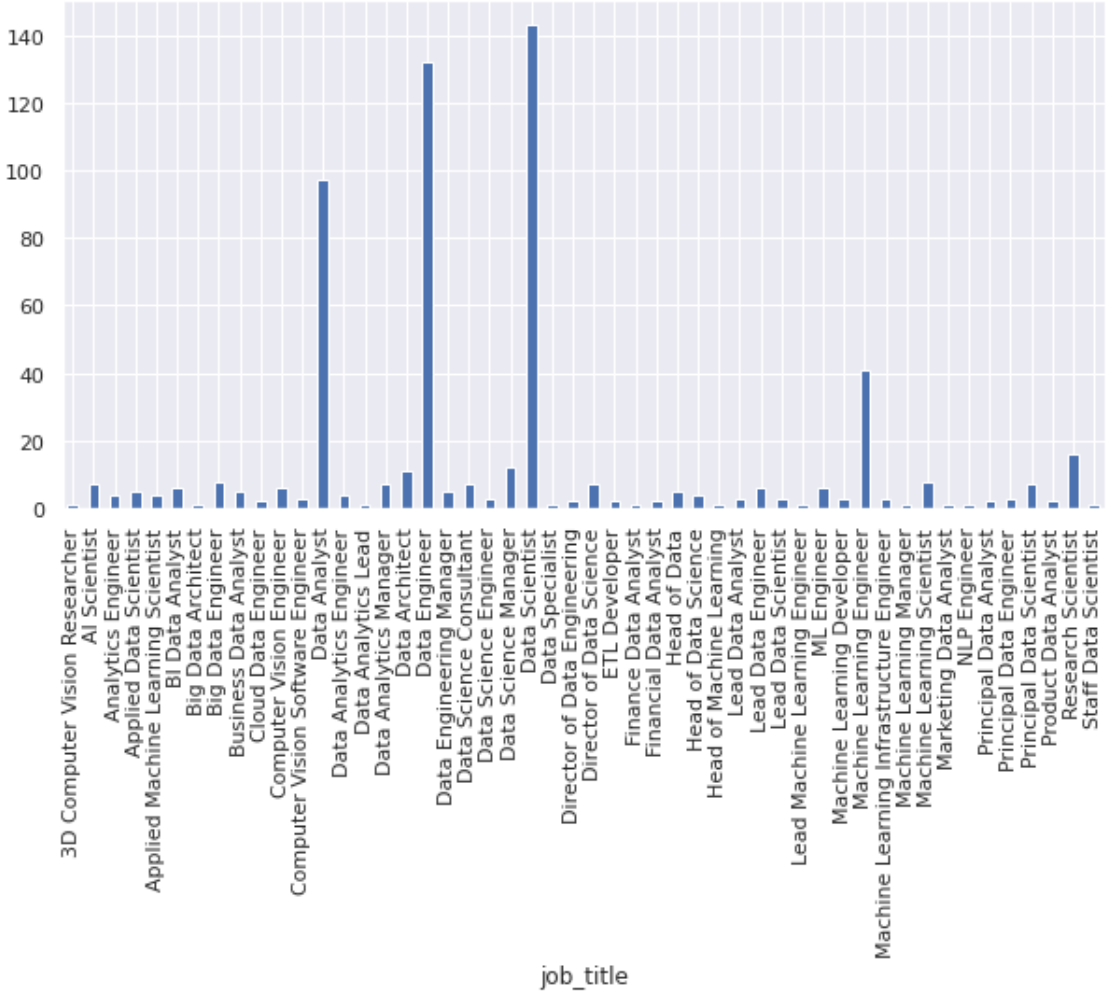


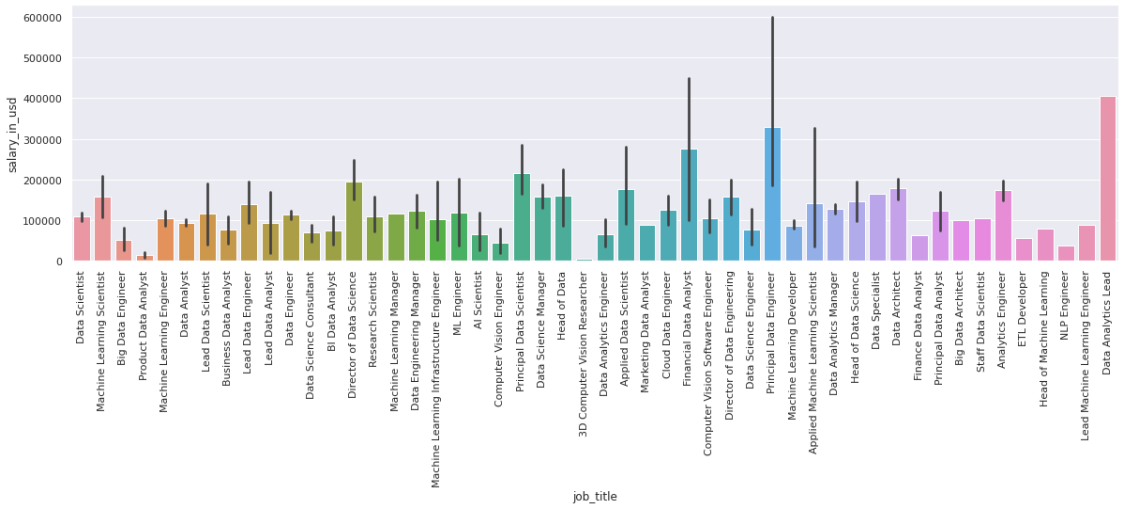
Observaciones

- El salario cambia de acuerdo con el tipo de empleo, alguién que tiene contrato o trabaja a tiempo completo gana más que alguien que trabaja por su cuenta o tiempo parcial.
- Se considera incluir la variable para el entrenamiento del modelo ya que parece estar correlacionada con el salario.

job_title


Data Analyst
Data Engineer
Computer Vision Software Engineer
Business Data Analyst
Big Data Engineer
Applied Machine Learning Scientist
Analytics Engineer
Research Scientist
Product Data Analyst
Machine Learning Engineer
Machine Learning Developer
Data Scientist
Data Engineering Manager
Data Science Consultant
Data Science Engineer
Data Science Manager
22%
16%
24%
7%
1%
2%

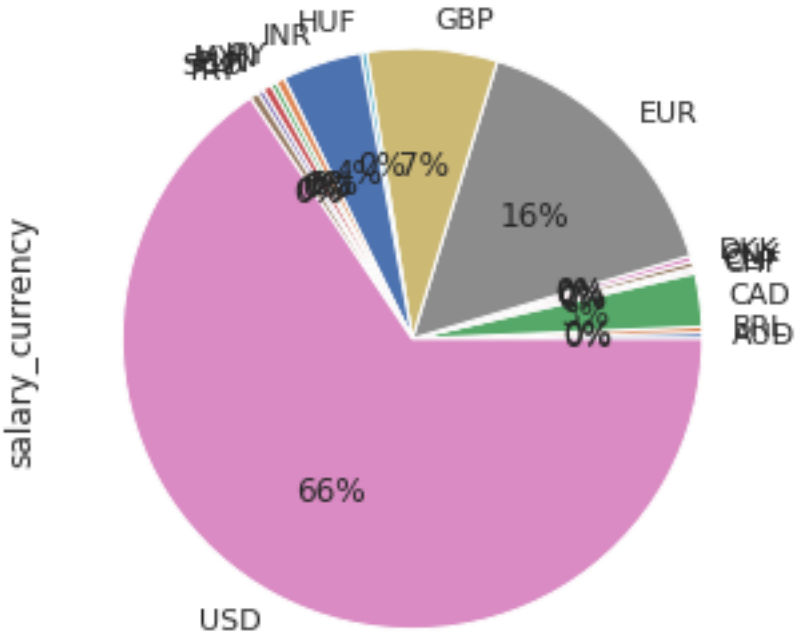

140
120
100
80
60
40
20
0
3D Computer Vision Researcher
AI Scientist
Analytics Engineer
Applied Data Scientist
Applied Machine Learning Scientist
BI Data Analyst
Big Data Architect
Big Data Engineer
Business Data Analyst
Cloud Data Engineer
Computer Vision Engineer
Computer Vision Software Engineer
Data 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

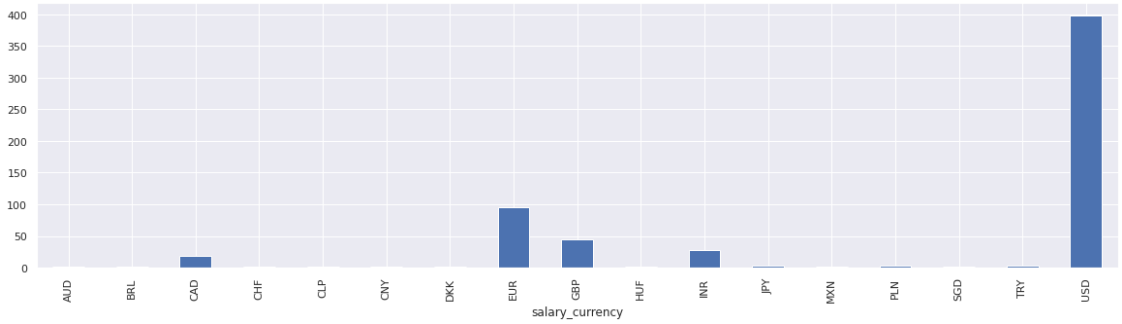


Observaciones

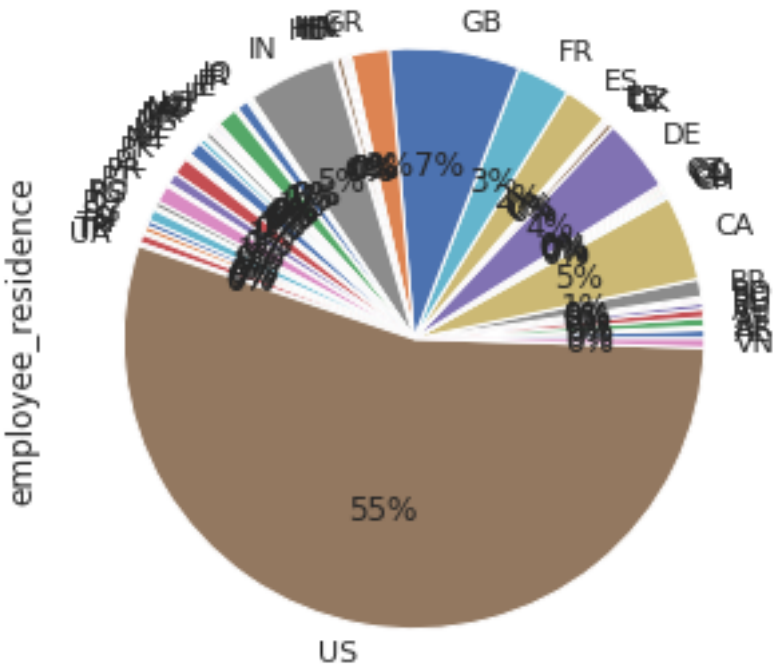
- La mayoría de los trabajos en ciencia de datos pertenecen a uno de estos grupos: analista de datos, ingeniero de datos y científico de datos.
- Existen ciertos puestos en los que se gana más que en la mayoría como Data Analytics Lead, Principal Data Engineer o Financial Data Analyst. Anque la mayoría gana entre 100000 y 200000.

salary_currency

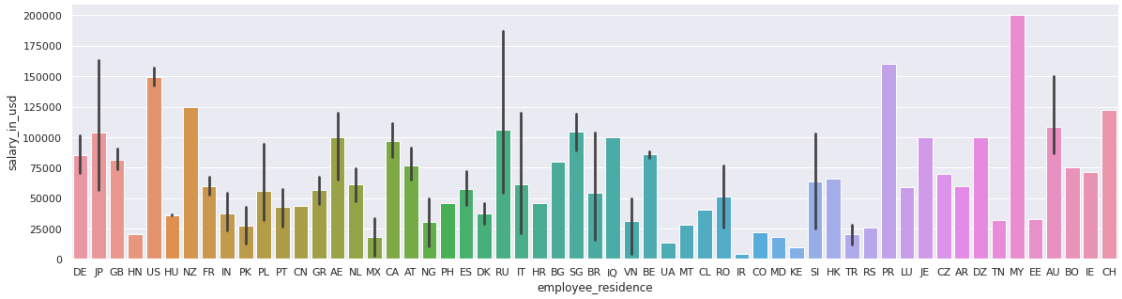




employee_residence


employee_residence
GB
FR
ES
DE
CA
IN
US
55%
5%
4%
3%
7%

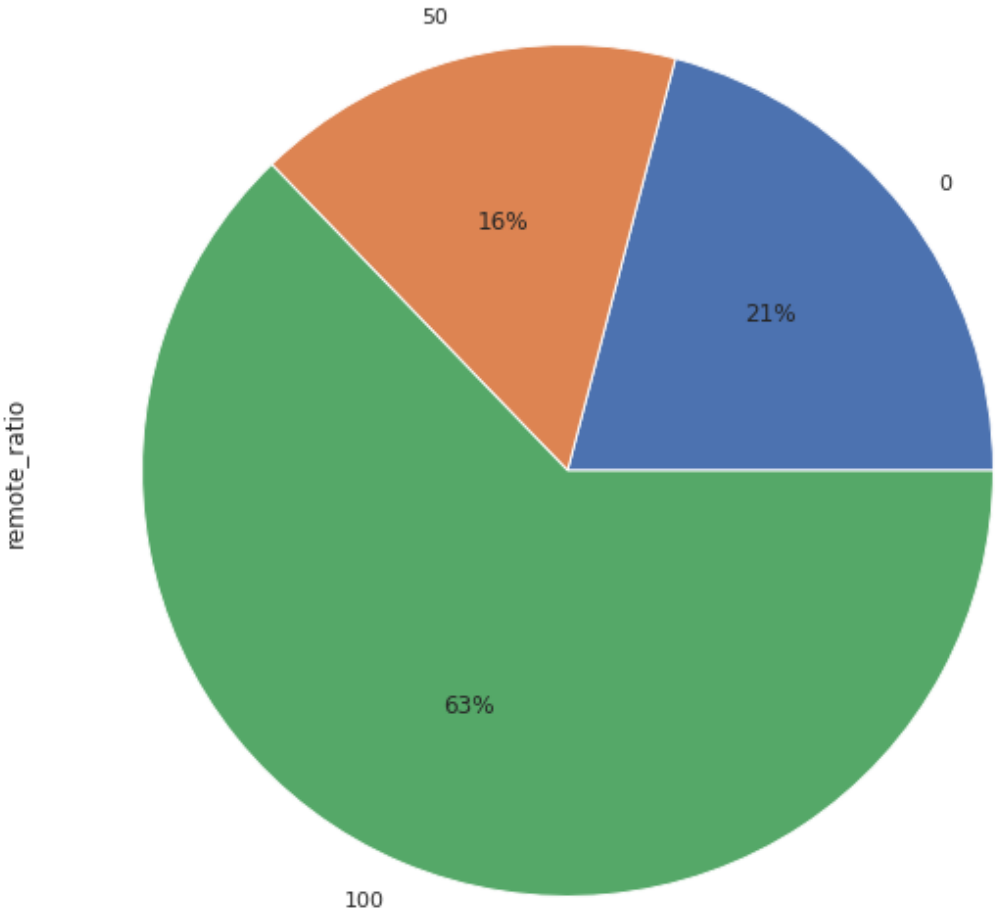

300
250
200
150
100
50
0
AE AR AT AU BE BG BO BR CA CH CL CN CO CZ DE DK DZ EE ES FR GB GR HK HN HR HU IE IN IQ IR IT JE JP KE LU MD MT MX MY NG NL NZ PH PK PL PR PT RO RS RU SG SI TN TR UA US VN
employee_residence

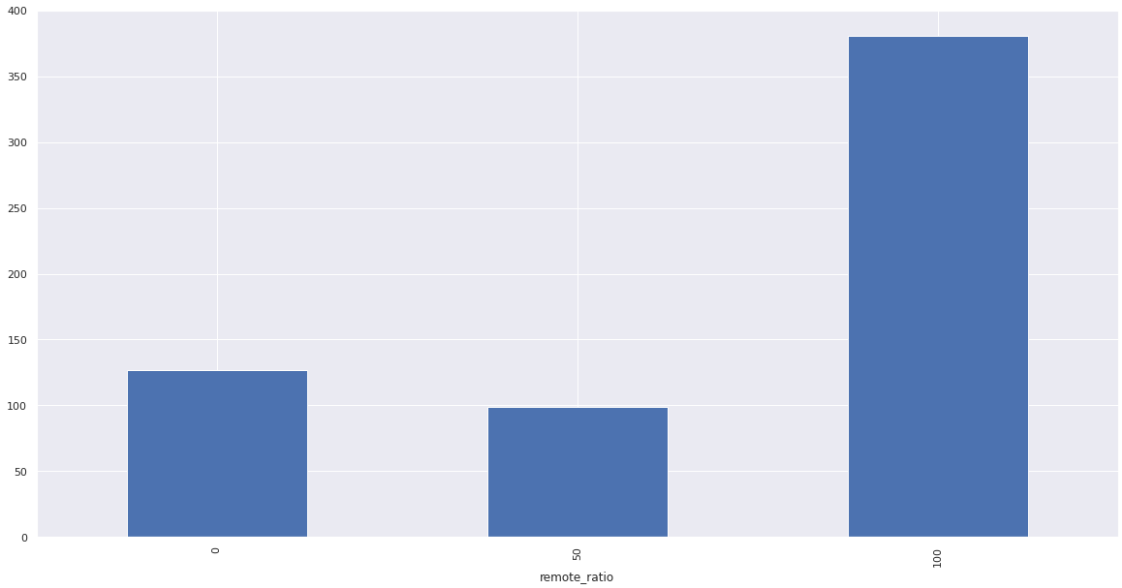

salary_in_usd
200000
175000
150000
125000
100000
75000
50000
25000
0
DE JP GB HN US HU NZ FR IN PK PL PT CN GR AE NL MX CA AT NG PH ES DK RU IT HR BG SG BR IQ VN BE UA MT CL RO IR CO MD KE SI HK TR RS PR LU JE CZ AR DZ TN MY EE AU BO IE CH
employee_residence

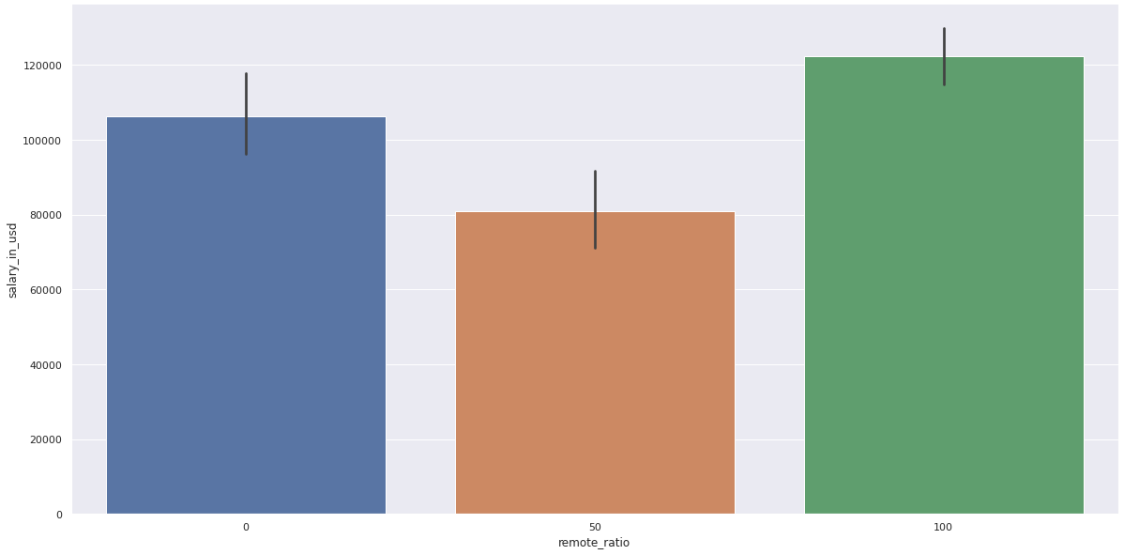
Observaciones

El salario cambia dependiendo del lugar de residencia de la persona, una persona en Malasia gana en promedio mucho más que alguien en México.

remote_ratio



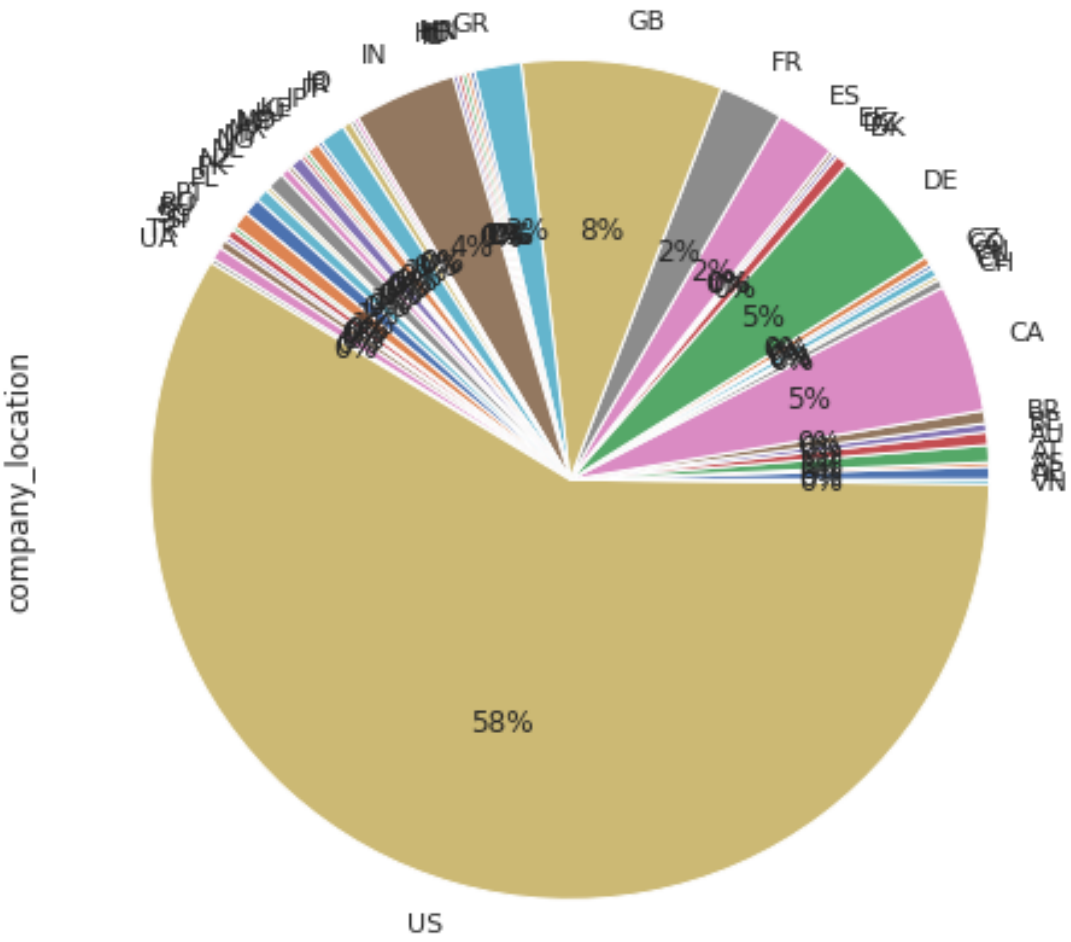


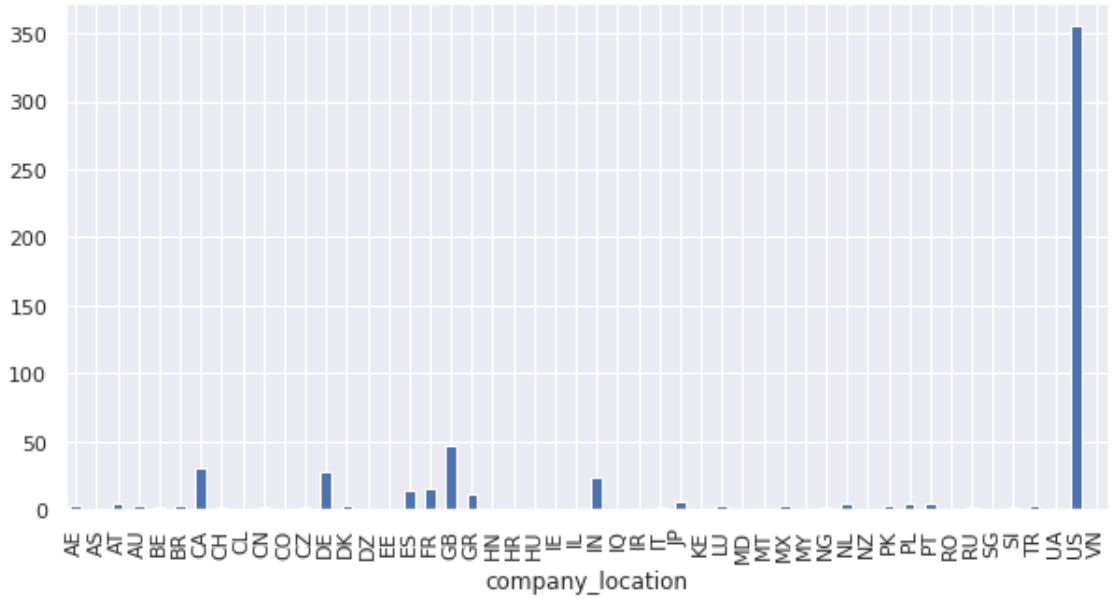


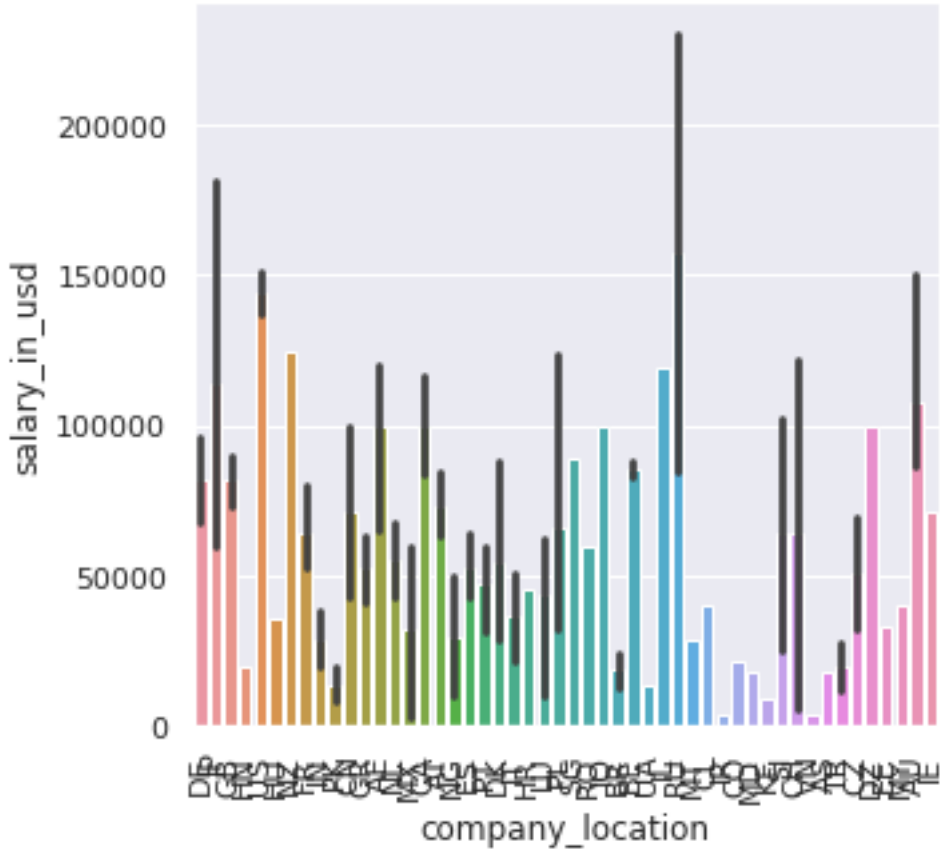
Observaciones

El promedio del salario varia si alguien trabaja presencial, remoto o híbrido. Es más alto el salario si alguien trabaja 100% remoto que alguien que trabaj híbrido. La diferencia de los promedios es de 40,000. Esta variable también podría estar relacionada con el salario.

company_location


company_location
US
58%
GB
8%
FR
2%
ES
DE
5%
CA
5%
IN
GR
UA

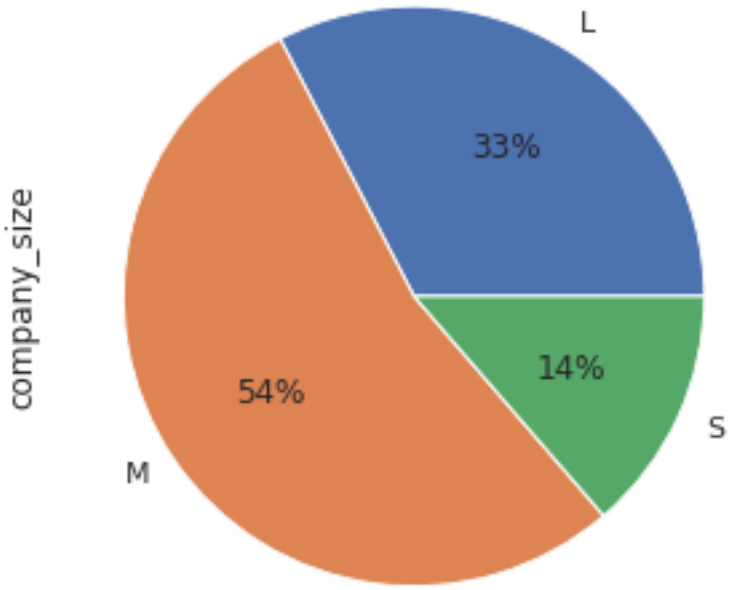

350
300
250
200
150
100
50
0
AE AS AT AU BE BR CA CH CL CN CO CZ DE DK DZ EE ES FR GB GR HN HR HU IE IL IN IQ IR IT JP KE LU MD MT MX MY NG NL NZ PK PL PT RO RU SG SI TR UA US VN
company_location

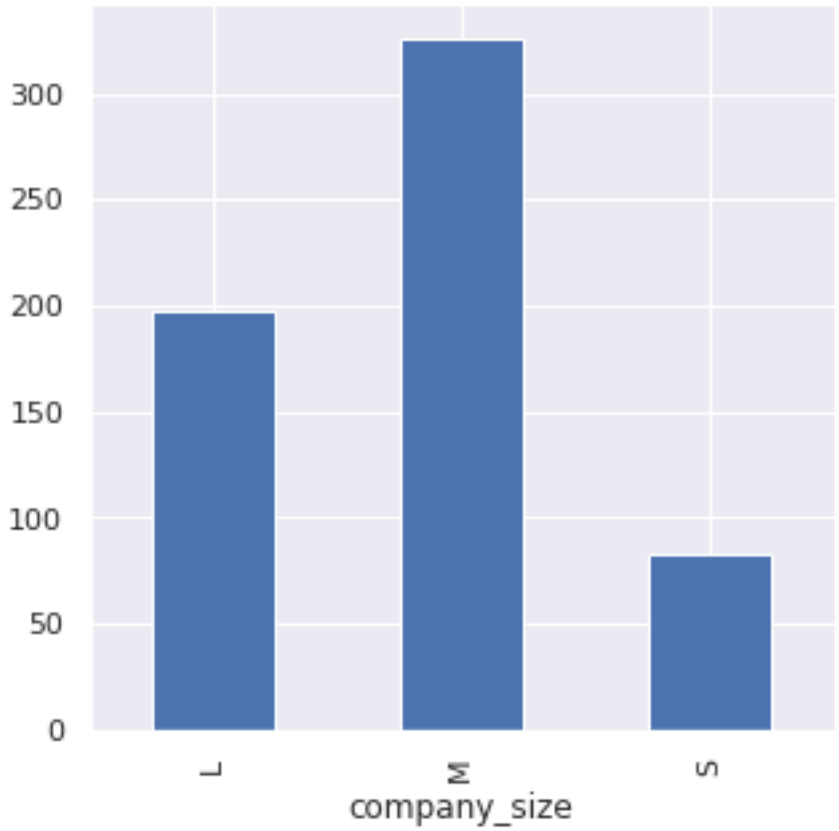

200000
150000
100000
50000
0
salary_in_usd
company_location

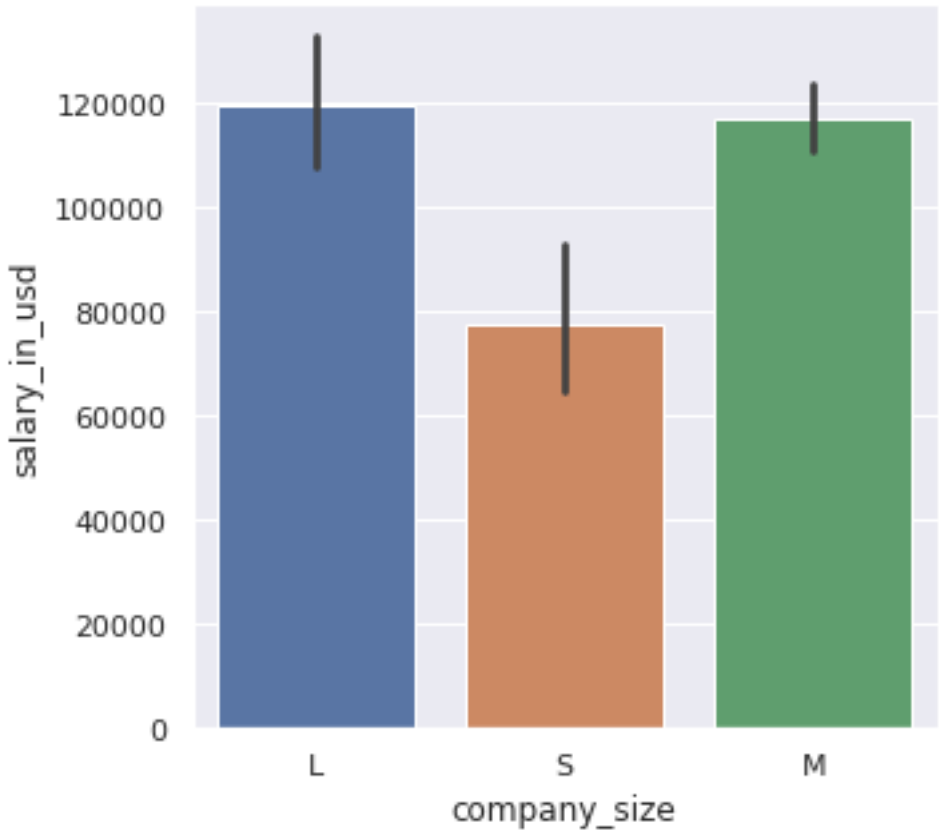
Observaciones

El salario también cambia dependiendo del lugar de la compañia.

company_size




300
250
200
150
100
50
0
L
M
S
company_size



1.2.3 3. Identifica problemas de calidad de datos

Variables con valores faltantes

```
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
```

Variables con registros duplicados

```
     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

42
```

No se encontraron valores faltantes y hay 42 registros duplicados. Se decidió que para el manejo de los registros duplicados, estos se van a eliminar.

1.3 D. Preparación de los datos

1.3.1 Selección de conjunto de datos

Variables no utilizadas: * Se decidió no utilizar las variables **salary** y **salary_currency** ya que es necesario que la variable objetivo (salary_in_usd) se encuentre en las mismas unidades y estas dos variables no proporcionan nueva información.

Variables utilizadas: * Se decidió utilizar las variables **work_year**,**experience_level**, **employment_type**, **job_title**, **employee_residence**, **remote_ratio**,**company_location**, **company_size** debido a que, como se observa en el punto anterior, tienen cierta influencia en el salario.

Variable objetivo * Salario en dolares (salary_in_usd)

1.3.2 Transformación de datos categóricos.

- Para la transformación de las variables ordinales se va a utilizar la clase OrdinalEncoder(), ya que toma los valores de las variables y los convierte en numéricos.
- Para la transformación de las variables nominales se van a obtener las variables dummies de estas.

Variables ordinales

- Remote ratio
- Experience level
- Company size

```
     experience_level  company_size  remote_ratio
0                 2.0           0.0             0
1                 3.0           2.0             0
2                 3.0           1.0            50
3                 2.0           2.0             0
4                 3.0           0.0            50
..                …             …              …
602               3.0           1.0           100
603               3.0           1.0           100
604               3.0           1.0             0
605               3.0           1.0           100
606               2.0           0.0           100

[607 rows x 3 columns]

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

Variables nominales

- work_year
- employment_type
- job_title
- employee_residence
- company_location

```
     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

[607 rows x 168 columns]

salary_in_usd          607
experience_level       607
company_size           607
remote_ratio           607
year_2020              607
                      …
company_location_SI    607
company_location_TR    607
company_location_UA    607
company_location_US    607
company_location_VN    607
Length: 168, dtype: int64
```

1.3.3 Manejo de datos duplicados y atípicos.

Debido a que solo se encontraron 8 outliers y 42 duplicados, se decide eliminarlos.

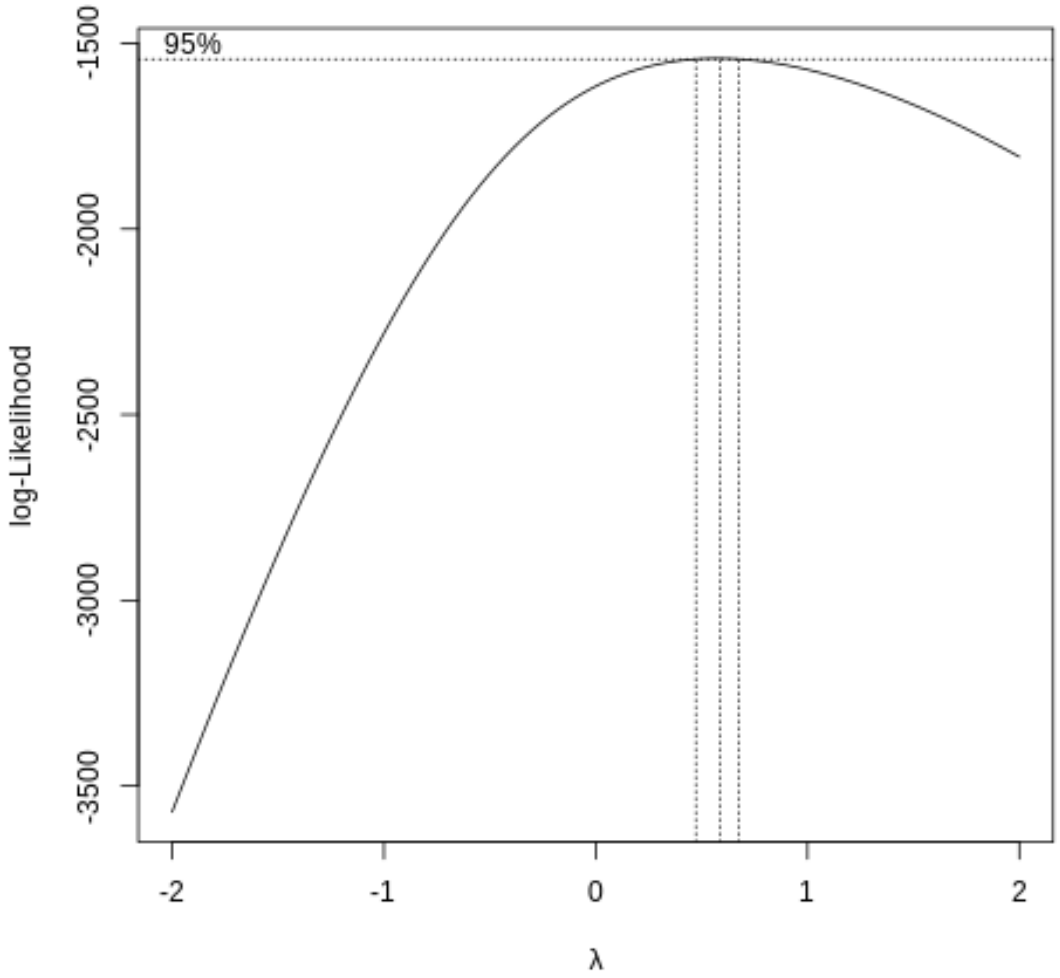
1.3.4 Escalamiento, normalización y discretización

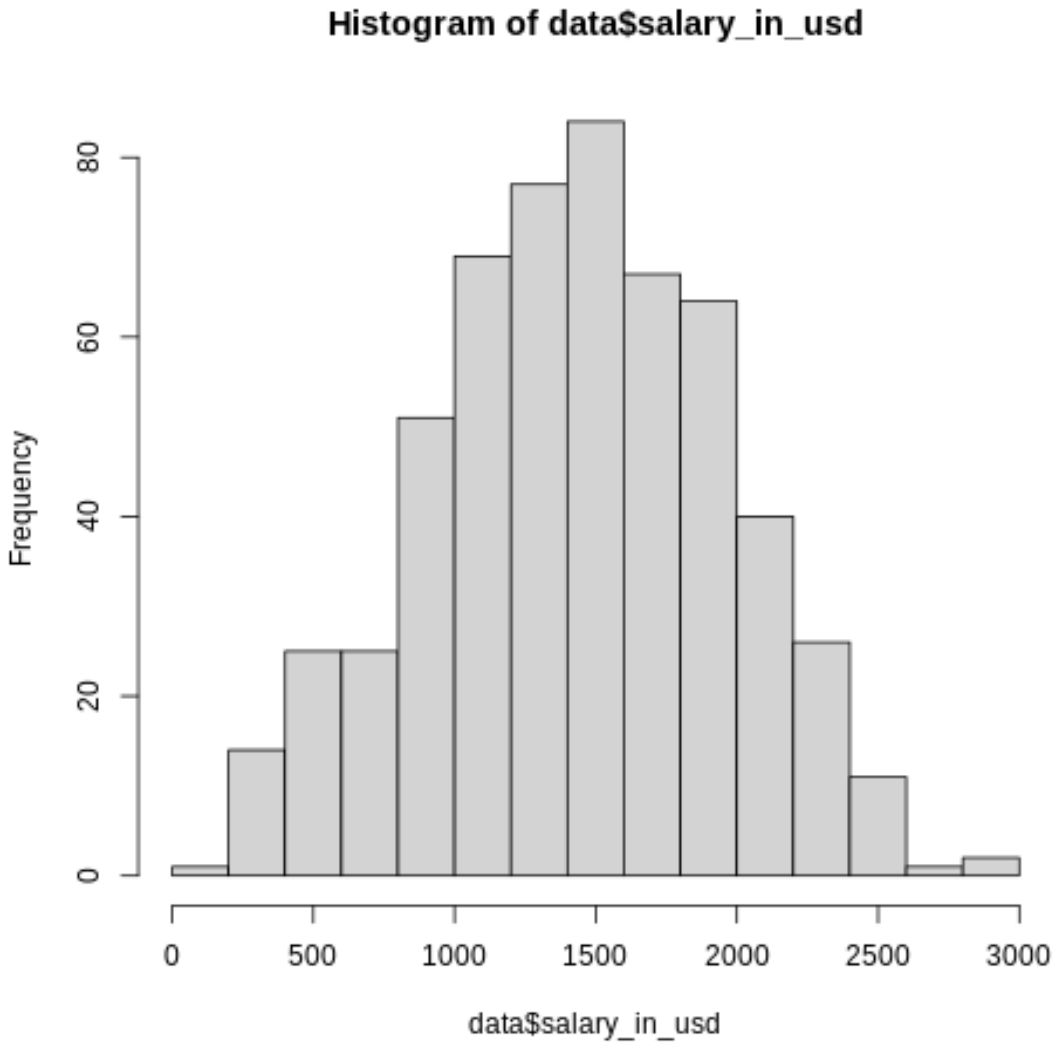
En este caso no es necesario **discretizar** los datos, ya que solo se tiene una variable numérica y se trata de la variable objetivo (salary_in_usd), la cual es discreta.

Debido a que la variable numérica salary_in_usd no tiene una forma normal, esta se va a **normalizar** para que sea apta para utilizarla en la propuesta de algún modelo. El resto de las variables se va a dejar sin normalizar ya que la mayoría se encuentra separada en variables dummies que solo toman valores de 0 y 1.

```
'%%R\ninstall.packages("MASS")'
```

```
[1] 0.5858586
```





También se realizará un **escalamiento** de los datos para que todas las variables estén en la misma escala, esto con un escalamiento estándar.

2 ANALIZA LOS DATOS Y CONTESTA TUS PREGUNTAS GUÍA

De acuerdo con lo encontrado en la fase de exploración fue posible contestar las siguientes preguntas de orientación.

¿Influye el nivel de experiencia en el salario?

Se encontró que el nivel de experiencia sí influye en el salario ya que al comparar la variable de experiencia con el salario se observó que en promedio alguien experto gana más que un Senior que

a su vez gana más que un Mid-level que a su vez gana más que alguien en Entry-level. Siendo los siguientes los salarios promedios por nivel se experiencia:

- Ex: 200000
- Se: 150000
- Mi: 90000
- En: 50000

¿Qué tipo de contrato (parcial, tiempo completo, etc) ofrece mejores salarios?

Se encontró que el tipo de contrato CT (Contract) es el que ofrece mejores salarios, ya que las personas con ese tipo de contrato ganan cerca de 200,000 en promedio, mientras que con otro tipo de contrato ganan menos de 120,000 en promedio.

¿En qué países se ofrecen mejores salarios?

Los países en los que se encuentran las compañías que ofrecen mejores salarios son Rusia, Estados Unidos y Nueva Zelanda.