# Los_salarios

September 18, 2022

## 0.1 Data processing techniques for statistical analysis and model building

Student: Jesús David Núñez Rodríguez A01634928

### 0.1.1 problem description:

Identify the conditions that make a person specialized in data analysis have a better, salary according to the database provided by Kaggle, in a sample of people who are dedicated to data analysis in different parts of the world.

Questions:
What are the best paid jobs?
The salary has increase over time?

**Preview of the data:**

```
   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
```

```
[607 rows x 11 columns]
```

## 1 Exploratory phase of data base

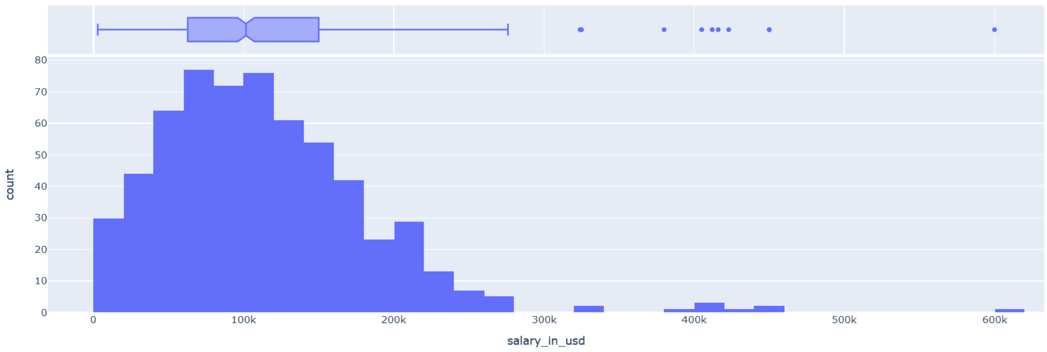
There is 607 rows wich will be consider n
#### **Types of atributes:**
Categorical (nominal):
Categorical (ordinal): work_year, experience_level, employment_type, job_title, salary_currency, employee_residence, company_location, company_size
Numeric: salary, salary_in_usd, remote_ratio

Describe function gives measures of central tendency and measures of dispersion

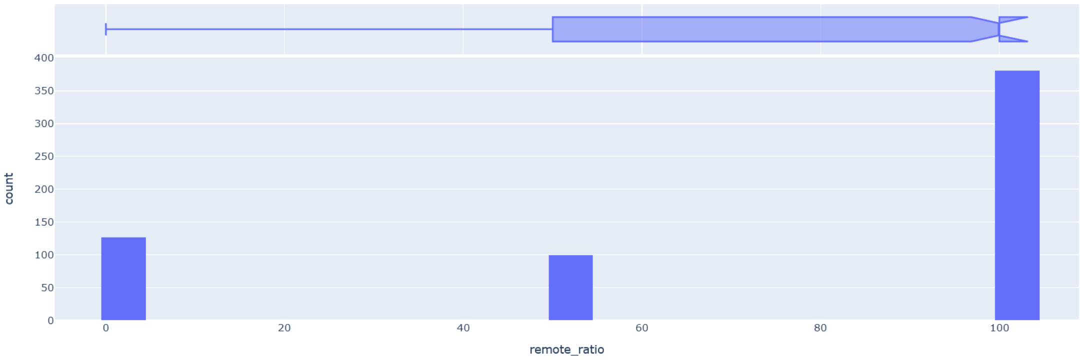
```
          work_year experience_level employment_type       job_title  \
count    607.000000              607             607             607
unique          NaN                4               4              50
top             NaN               SE              FT  Data Scientist
freq            NaN              280             588             143
mean    2021.405272              NaN             NaN             NaN
std        0.692133              NaN             NaN             NaN
min     2020.000000              NaN             NaN             NaN
25%     2021.000000              NaN             NaN             NaN
50%     2022.000000              NaN             NaN             NaN
75%     2022.000000              NaN             NaN             NaN
max     2022.000000              NaN             NaN             NaN

              salary salary_currency  salary_in_usd employee_residence  \
count   6.070000e+02             607     607.000000                607
unique           NaN              17            NaN                 57
top              NaN             USD            NaN                 US
freq             NaN             398            NaN                332
mean    3.240001e+05             NaN  112297.869852                NaN
std     1.544357e+06             NaN   70957.259411                NaN
min     4.000000e+03             NaN    2859.000000                NaN
25%     7.000000e+04             NaN   62726.000000                NaN
50%     1.150000e+05             NaN  101570.000000                NaN
75%     1.650000e+05             NaN  150000.000000                NaN
max     3.040000e+07             NaN  600000.000000                NaN

        remote_ratio company_location company_size
count      607.00000              607          607
unique           NaN               50            3
top              NaN               US            M
freq             NaN              355          326
mean        70.92257              NaN          NaN
std         40.70913              NaN          NaN
```

```
min          0.00000              NaN          NaN
25%         50.00000              NaN          NaN
50%        100.00000              NaN          NaN
75%        100.00000              NaN          NaN
max        100.00000              NaN          NaN
```

By a quick review of the metrics, it can be said that most of the jobs in df are from people working in US companies, because the most common company location (up to 58%) is in US. Furthermore the **mean salary is $112,297.00**



The right tail is considerably longer, also is a right-skewed (asymmetric) distribution due to the outliers that pull the mean to the right.



Its important to notice that remote ratio is a numeric atribute but behave like a categorical, would be useful validate with the source of the data the nature of it.

check for NaN values in df to avoid inconsistences in further analisis

```
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
```

Frecuency of categorical variables

```
SE    280
MI    213
EN     88
EX     26
Name: experience_level, dtype: int64
Data Scientist                              143
Data Engineer                               132
Data Analyst                                 97
Machine Learning Engineer                    41
Research Scientist                           16
Data Science Manager                         12
Data Architect                               11
Big Data Engineer                             8
Machine Learning Scientist                    8
Principal Data Scientist                      7
AI Scientist                                  7
Data Science Consultant                       7
Director of Data Science                      7
Data Analytics Manager                        7
ML Engineer                                   6
Computer Vision Engineer                      6
BI Data Analyst                               6
Lead Data Engineer                            6
Data Engineering Manager                      5
Business Data Analyst                         5
Head of Data                                  5
Applied Data Scientist                        5
Applied Machine Learning Scientist            4
Head of Data Science                          4
Analytics Engineer                            4
Data Analytics Engineer                       4
Machine Learning Developer                    3
Machine Learning Infrastructure Engineer      3
Lead Data Scientist                           3
Computer Vision Software Engineer             3
Lead Data Analyst                             3
Data Science Engineer                         3
Principal Data Engineer                       3
Principal Data Analyst                        2
ETL Developer                                 2
```

```
Product Data Analyst                             2
Director of Data Engineering                     2
Financial Data Analyst                           2
Cloud Data Engineer                              2
Lead Machine Learning Engineer                   1
NLP Engineer                                     1
Head of Machine Learning                         1
3D Computer Vision Researcher                    1
Data Specialist                                  1
Staff Data Scientist                             1
Big Data Architect                               1
Finance Data Analyst                             1
Marketing Data Analyst                           1
Machine Learning Manager                         1
Data Analytics Lead                              1
Name: job_title, dtype: int64
US    355
GB     47
CA     30
DE     28
IN     24
FR     15
ES     14
GR     11
JP      6
NL      4
AT      4
PT      4
PL      4
LU      3
PK      3
BR      3
AE      3
MX      3
AU      3
TR      3
DK      3
IT      2
CZ      2
SI      2
RU      2
CH      2
NG      2
CN      2
BE      2
VN      1
EE      1
AS      1
```

```
DZ      1
MY      1
MD      1
KE      1
SG      1
CO      1
IR      1
CL      1
MT      1
IL      1
UA      1
IQ      1
RO      1
HR      1
NZ      1
HU      1
HN      1
IE      1
Name: company_location, dtype: int64
```

# 2 data preprocesing

Will be remove outliers based in salary. This help to make a better analisis. however the df still shows bias to the right but with shorter tails.

```
The new size of df(after drop outliers) is: 598
```

The atributes to analize will be: As inpedendent: experience_level, job_title, company_location As dependent: salary_in_usd

Salary and salary currency will be drop due that salary_in_usd standarize the income in a unique scala. Remote ratio will be drop for its unsure nature

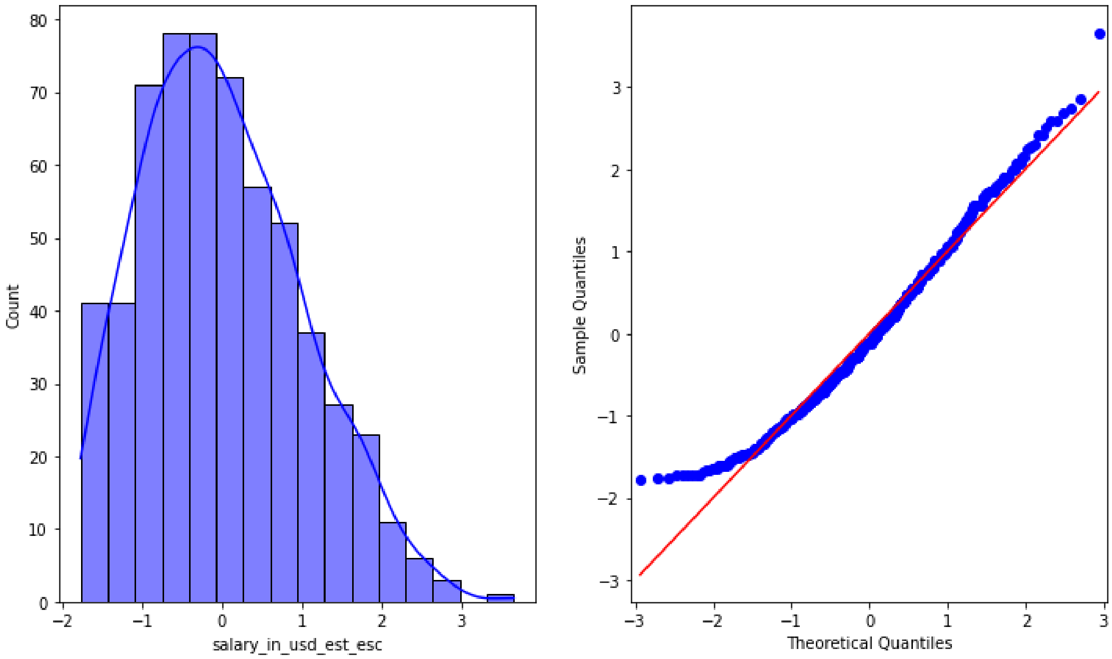
# 3 Does the level of experience influence the salary?

### 3.0.1 H0: The experience groups have equal mean

### 3.0.2 H1: At least one group introduce significance to displace the mean

dict for experience map: 0 = 'EN', 1 = 'EX', 2 = 'MI', 3 = 'SE'

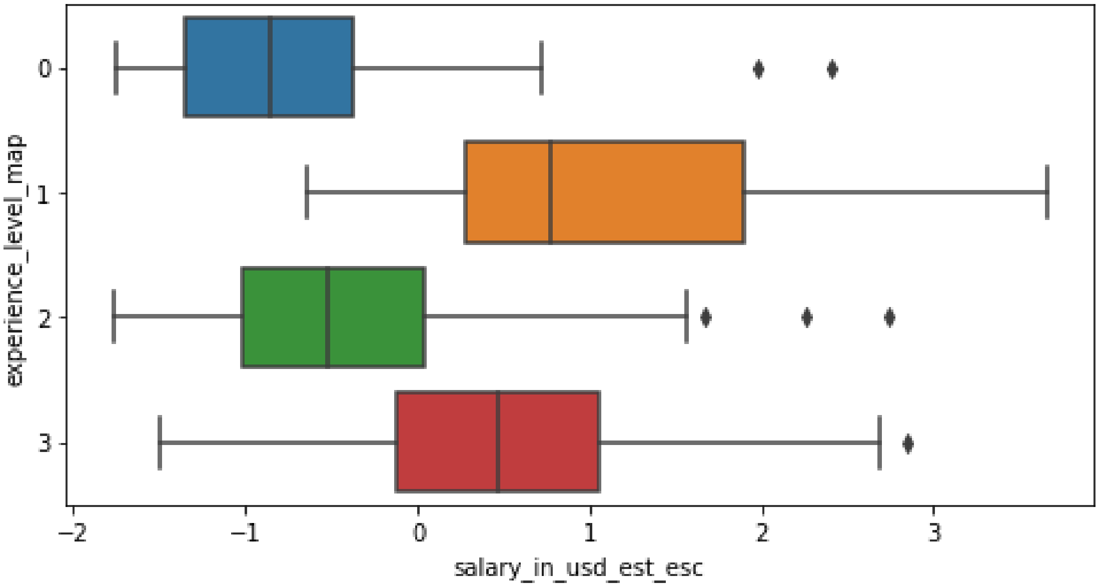
```
   work_year experience_level employment_type                   job_title  \
0       2020               MI              FT              Data Scientist   
1       2020               SE              FT  Machine Learning Scientist   
2       2020               SE              FT           Big Data Engineer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

   salary_in_usd employee_residence company_location company_size  \
0        79833                 DE                DE            L
1       260000                 JP                JP            S
2       109024                 GB                GB            M
3        20000                 HN                HN            S
4       150000                 US                US            L

   experience_level_map  job_title_map  company_location_map
0                     2             21                    12
1                     3             40                    29
2                     3              7                    18
3                     2             46                    20
4                     3             37                    48
```



Soft tails distribution, high curtosis, leptocurtic distribution.

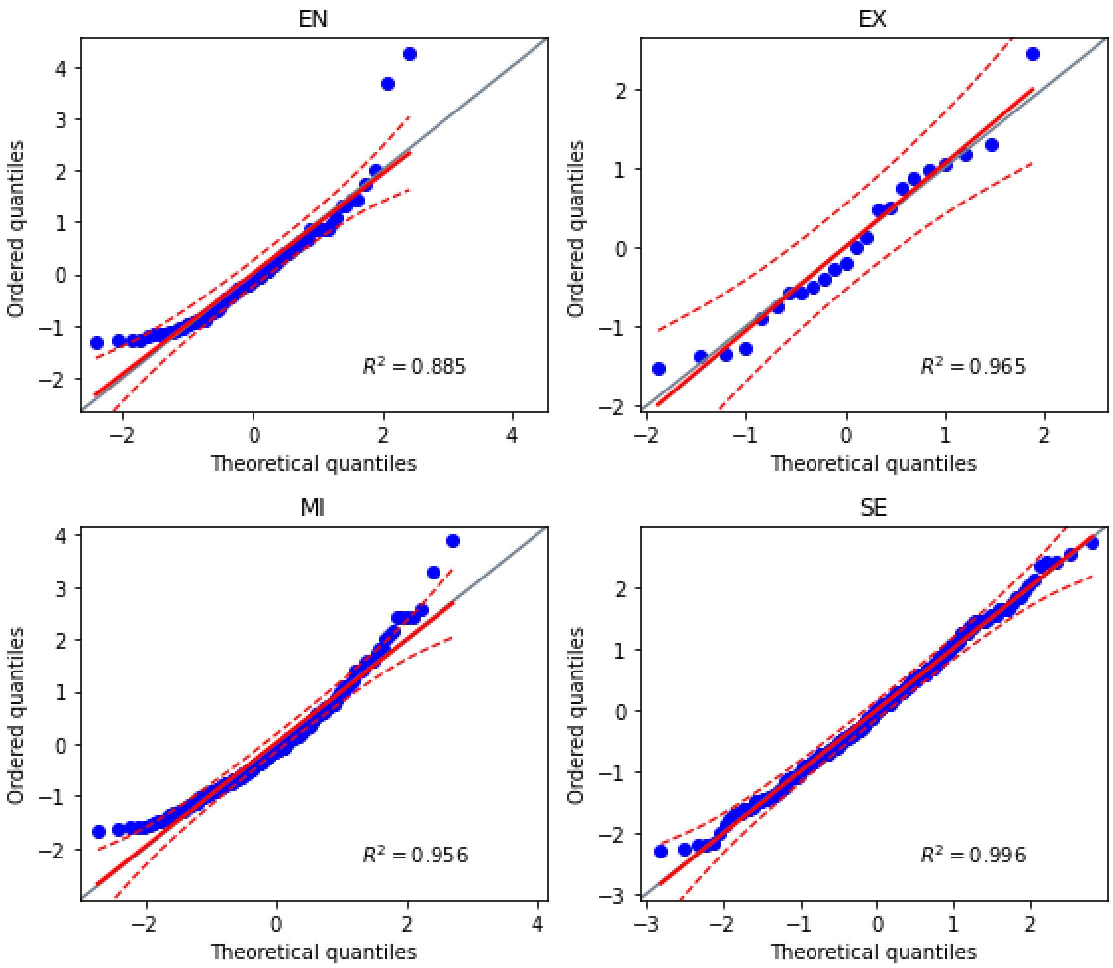
```
                        work_year              salary_in_usd                \
                             mean        std            mean           std
experience_level_map
0                     2021.011364   0.686392    61643.318182  44395.541126
1                     2021.521739   0.593109   167095.347826  65874.574937
2                     2021.285714   0.708314    82953.142857  48222.337602
3                     2021.628159   0.598005   135797.263538  51162.122770

                     job_title_map             company_location_map            \
                              mean        std                  mean        std
experience_level_map
```

```
0                        19.375000  11.439308            29.511364  16.312321
1                        19.869565   8.976253            36.652174  16.305628
2                        20.300000  10.183038            31.395238  16.128698
3                        20.350181   9.812930            40.862816  14.292593

                     salary_in_usd_est_esc
                                      mean       std
experience_level_map
0                                -0.775471  0.750247
1                                 1.006580  1.113225
2                                -0.415353  0.814917
3                                 0.477668  0.864597
```



In conclusion, h0 is not rejected, that could be validate by the confidences intervals graph.

Validate normality in data



According to normality graph there ir normality in data, but is necesary to validate with complementary tests like Shapiro-Wilk and homocedasticity.

Normality test Shapiro-Wilk

```
          W      pval  normal
2  0.956378  0.000005   False
3  0.993931  0.331550    True
0  0.889819  0.000002   False
1  0.960074  0.464865    True
```

The result of Shapiro-Wilk test show that the data present inconsistencies in normality, in some.

**Homocedasticity test**

```
               W     pval  equal_var
levene  3.596672  0.01345      False
```

The second red flag is that the data do not present homocedasticity

### 3.0.3 One-way ANOVA test

```
                 Source          SS   DF         MS          F         p-unc  \
0  experience_level_map  175.653951    3  58.551317  82.543748  1.199071e-44
1                Within  421.346049  594   0.709337        NaN           NaN

        np2
0  0.294228
1       NaN
```

P-value is smaller than 0.05 wich is evidence to refuse the null hipotesis. Consequently, the experience level does affect the average income.

**Post-hoc Tukey test**

```
   A  B  mean(A)  mean(B)   diff     se       T  p-tukey  hedges
0  0  1   -0.775    1.007 -1.782  0.197  -9.035    0.000  -2.101
1  0  2   -0.775   -0.415 -0.360  0.107  -3.367    0.004  -0.426
2  0  3   -0.775    0.478 -1.253  0.103 -12.159    0.000  -1.485
3  1  2    1.007   -0.415  1.422  0.185   7.687    0.000   1.683
4  1  3    1.007    0.478  0.529  0.183   2.894    0.021   0.626
5  2  3   -0.415    0.478 -0.893  0.077 -11.588    0.000  -1.059
```

The groups 0,2 have a small diference in their means wich demostrate that these groups are similar,also 1,3

## 4 The salary has increase over time?

Description of data

```
           salary_in_usd               experience_level_map            \
                    mean           std                 mean       std
work_year
2020        82775.884058  53887.352872             1.652174  1.148222
2021        92860.436620  61531.282566             1.840376  1.108655
2022       122825.943038  54286.303186             2.430380  0.853505

          job_title_map            company_location_map            \
                   mean        std                 mean        std
work_year
2020          21.811594  10.936011            30.797101  15.995937
2021          22.070423  11.719188            32.887324  16.327326
2022          18.531646   8.437385            38.677215  15.363849

          salary_in_usd_est_esc
                          mean       std
work_year
2020                 -0.418348  0.910651
```

```
2021                     -0.247928  1.039827
2022                      0.258464  0.917393
```

There is no need in do a deeper analisis, from 2020 to 2021 it has increase the salary in a 12%, and a 48% from 2020 to 2022.

## 5 What are the best paid jobs?

```
job_title
Principal Data Engineer     192500.000000
Principal Data Scientist    181782.833333
Data Architect              177873.909091
Analytics Engineer          175000.000000
Director of Data Science    173419.666667
Data Specialist             165000.000000
Head of Data                160162.600000
Name: (salary_in_usd, mean), dtype: float64
```

The top 7 of better paid jobs are the list from above

Conclusion, the data is not reliable because it do not present normality or homoscedasticity in the case of the variable of experience.

Code link: https://github.com/a01634928/TC3006C_101_A01634928/tree/main/modulo_1/tecnicas_de%20procesamiento_de_datos_para_el_analisis%20_estadistico