# *DATA SCIENCE SALARY TRENDS*

## *2023*

Verónica Arasanz

Comisión 46310

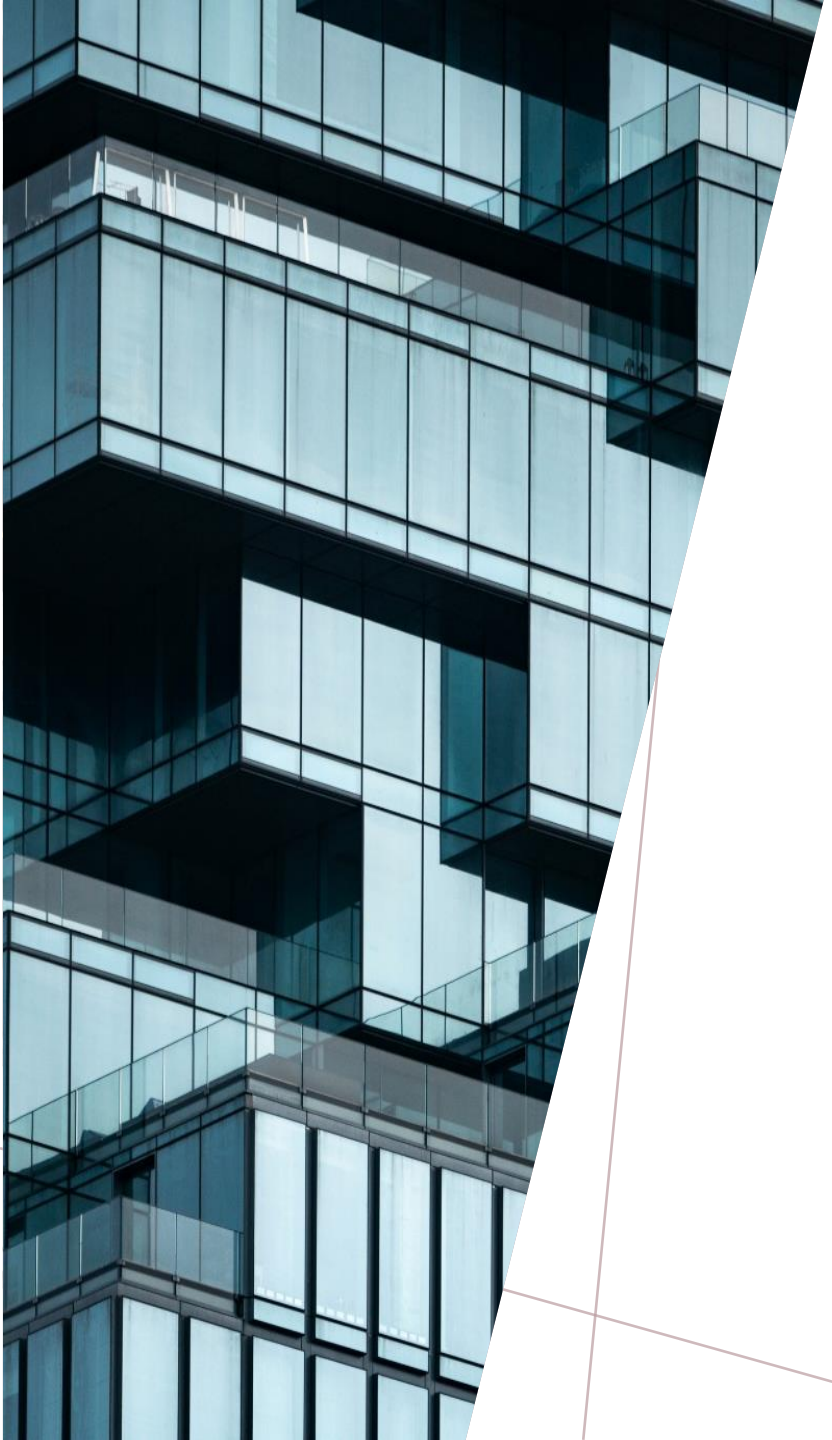
2023

# *ABSTRACT*

Este trabajo de ciencia de datos se centra en el análisis de los salarios de Data Science desde el 2020 al 2023.

El objetivo principal es comprender las tendencias, relaciones y factores que influyen en los salarios de Data Science en ese período de tiempo.

Se espera que este trabajo proporcione una visión completa de los salarios de Data Science durante estos años

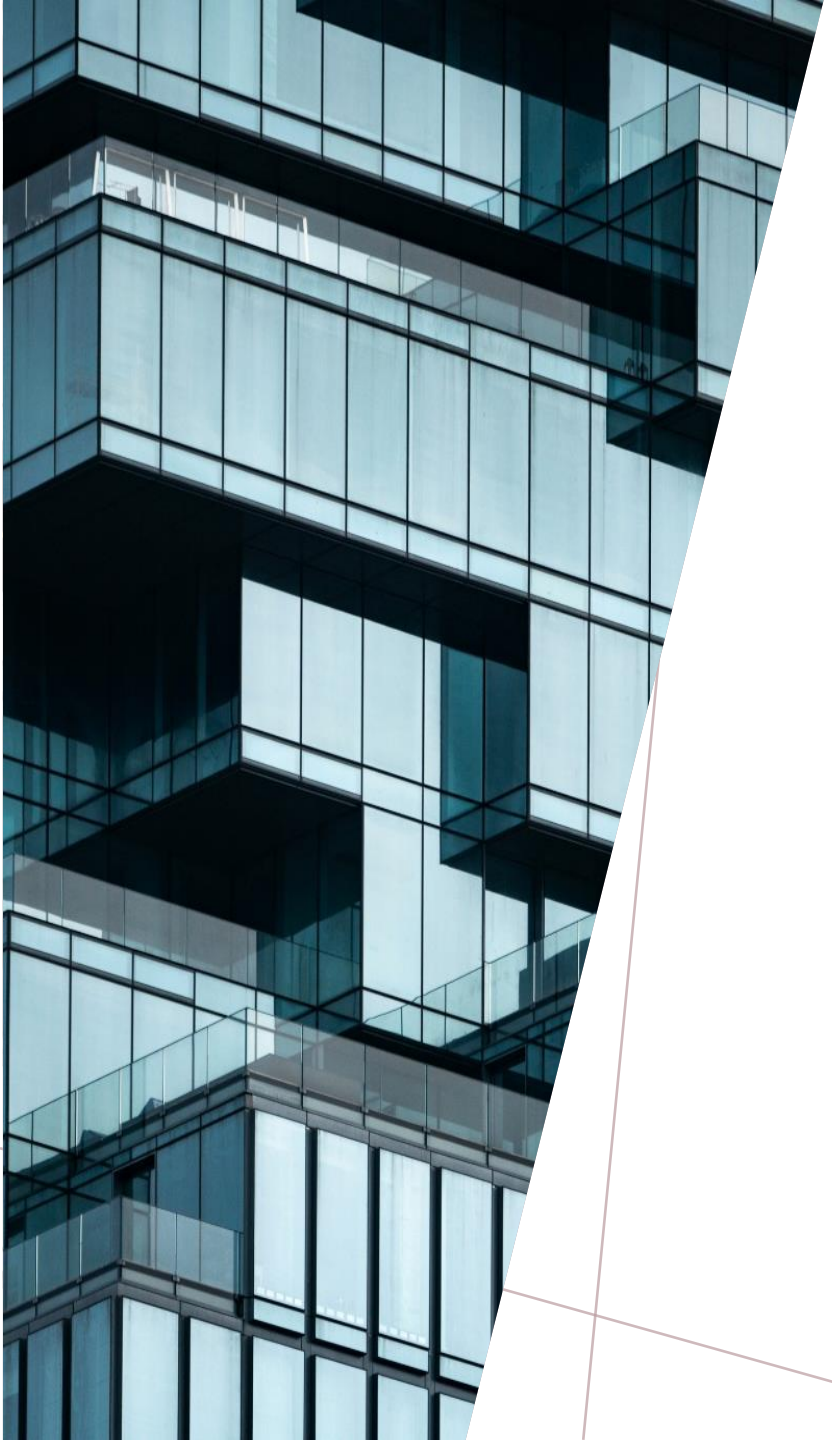


# CONTEXTO COMERCIAL

Una empresa multinacional, que está queriendo crear su área de Data Science, precisa analizar el mercadero laboral para decidir su estrategia de contratación del personal

# PREGUNTAS DESCRIPTIVAS

1) ¿Cuál es la tendencia general de los salarios de Data Science en los últimos cuatro años (2020-2023)?
2) ¿Existen diferencias significativas en los salarios de Data Science entre empresas de diferentes tamaños?
3) ¿Cómo varía el salario de Data Science según la experiencia laboral?

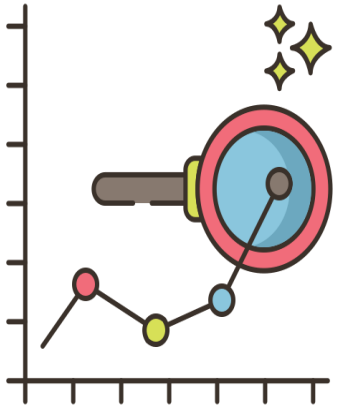


# *HIPÓTESIS*

1) Los salarios de Data Science han experimentado un aumento constante a lo largo de los años.
2) Existe una correlación positiva entre el tamaño de la empresa y los salarios de Data Science.
3) Los profesionales con más experiencia en Data Science ganan más que los que tienen menos experiencia.

# DEFINICIÓN DE OBJETIVO

Predecir el salario de un trabajador de data Science de acuerdo a su experiencia, tipo de trabajo y tamaño de empresa en que trabaja.



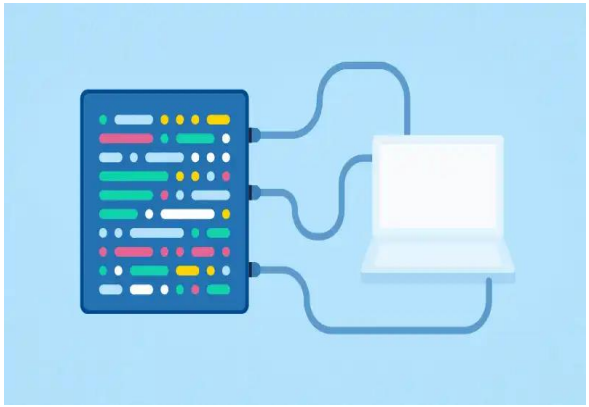
# ANÁLISIS PREDICTIVO

1) ¿Se puede desarrollar un modelo de regresión que prediga los salarios de Data Science en función de la experiencia laboral, el país y el tamaño de la empresa?

2) ¿Se podría utilizar la clasificación para predecir el tipo de empleo (por ejemplo, tiempo completo o medio tiempo) en función de otros atributos?

# *DATASET*

El conjunto de datos utilizado tiene como objetivo analizar las tendencias salariales en el campo de la Ciencia de Datos para los años 2020 a 2023.



# DATA FIELDS DEL DATASET

1)work_year: Representing the specific year of salary data collection.

2)Experience_level: The level of work experience of the employees, categorized as EN (Entry-Level), EX (Experienced), MI (Mid-Level), SE (Senior).

3)Employment_type: The type of employment, labelled as FT (Full-Time), CT (Contract), FL (Freelance), PT (Part-Time).

4)Job_title: The job titles of the employees, such as "Applied Scientist", "Data Quality Analyst" , etc.

5)Salary: The salary figures in their respective currency formats.

6)Salary_currency: The currency code representing the salary.

7)Salary_in_usd: The converted salary figures in USD for uniform comparison.

8)Company_location: The location of the companies, specified as country codes (e.g., "US" for the United States and "NG" for Nigeria).

9)Company_size: The size of the companies, classified as "L" (Large), "M" (Medium), and "S" (Small).

# *DATA WRANGLING*

- Se quitaron datos duplicados
- Se verificó que no existen datos nulos
- Se eliminaron las columnas de salario y moneda, por contar con la columna salario en dólares.
- Se agruparon los títulos de trabajo que representaba el mismo trabajo, para mejorar el análisis con menos categorías
- Identificamos outliers en el modelo y los eliminamos.

# *HEAD*


| | work_year | experience_level | employment_type | job_title | salary_in_usd | company_location | company_size |
|---|---|---|---|---|---|---|---|
| 0 | 2023 | Entry-Level | Full-time | Applied Scientist | 213660 | US | Large |
| 1 | 2023 | Entry-Level | Full-time | Applied Scientist | 130760 | US | Large |
| 2 | 2023 | Entry-Level | Full-time | Data Analyst | 100000 | NG | Large |
| 3 | 2023 | Entry-Level | Full-time | Data Analyst | 30000 | NG | Large |
| 4 | 2023 | Entry-Level | Full-time | Applied Scientist | 204620 | US | Large |

# *REPORTE*

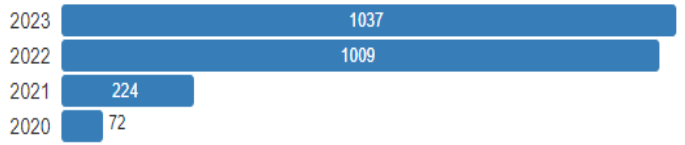
Conociendo nuestro DATA FRAME.
Análisis univariado

### work_year

Categorical

HIGH CORRELATION

| | |
|---|---|
| **Distinct** | 4 |
| **Distinct (%)** | 0.2% |
| **Missing** | 0 |
| **Missing (%)** | 0.0% |
| **Memory size** | 18.4 KiB |

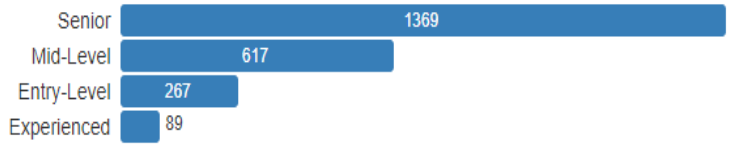



### experience_level

Categorical

HIGH CORRELATION

| | |
|---|---|
| **Distinct** | 4 |
| **Distinct (%)** | 0.2% |
| **Missing** | 0 |
| **Missing (%)** | 0.0% |
| **Memory size** | 18.4 KiB |



# REPORTE

## Análisis univariado

### employment_type

Categorical

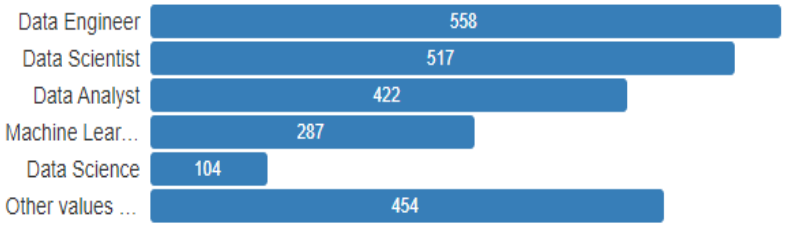
IMBALANCE

| | |
|---|---|
| **Distinct** | 4 |
| **Distinct (%)** | 0.2% |
| **Missing** | 0 |
| **Missing (%)** | 0.0% |
| **Memory size** | 18.4 KiB |


### job_title

Categorical

| | |
|---|---|
| **Distinct** | 41 |
| **Distinct (%)** | 1.8% |
| **Missing** | 0 |
| **Missing (%)** | 0.0% |
| **Memory size** | 18.4 KiB |



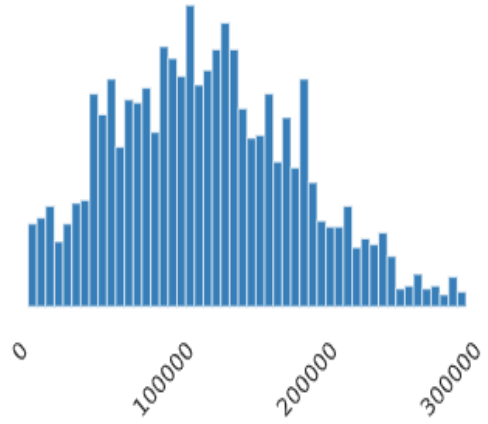
# *REPORTE*

## Análisis univariado

### salary_in_usd

Real number (ℝ)

| | | | |
|---|---|---|---|
| **Distinct** | 1009 | **Minimum** | 5132 |
| **Distinct (%)** | 43.1% | **Maximum** | 310000 |
| **Missing** | 0 | **Zeros** | 0 |
| **Missing (%)** | 0.0% | **Zeros (%)** | 0.0% |
| **Infinite** | 0 | **Negative** | 0 |
| **Infinite (%)** | 0.0% | **Negative (%)** | 0.0% |
| **Mean** | 129651.46 | **Memory size** | 18.4 KiB |

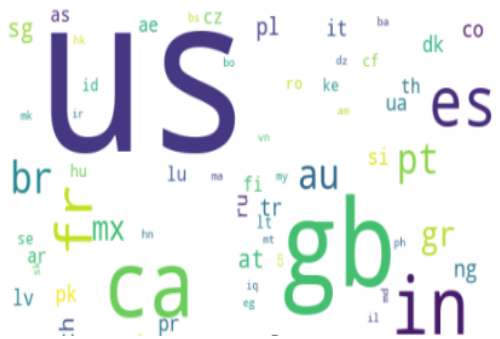


# *REPORTE*

## Análisis univariado

company_location
Text

| | |
|---|---|
| **Distinct** | 72 |
| **Distinct (%)** | 3.1% |
| **Missing** | 0 |
| **Missing (%)** | 0.0% |
| **Memory size** | 18.4 KiB |

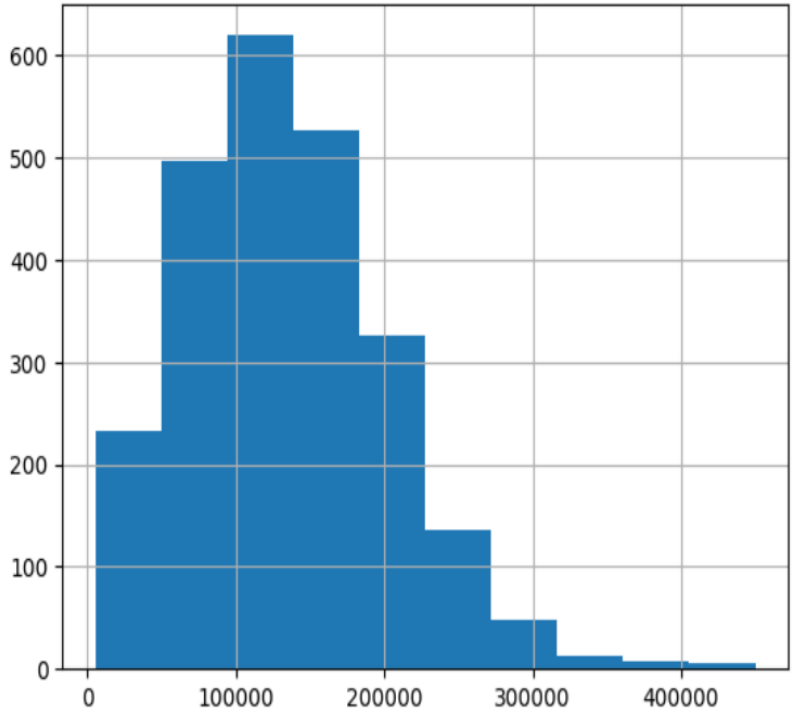



company_size
Categorical

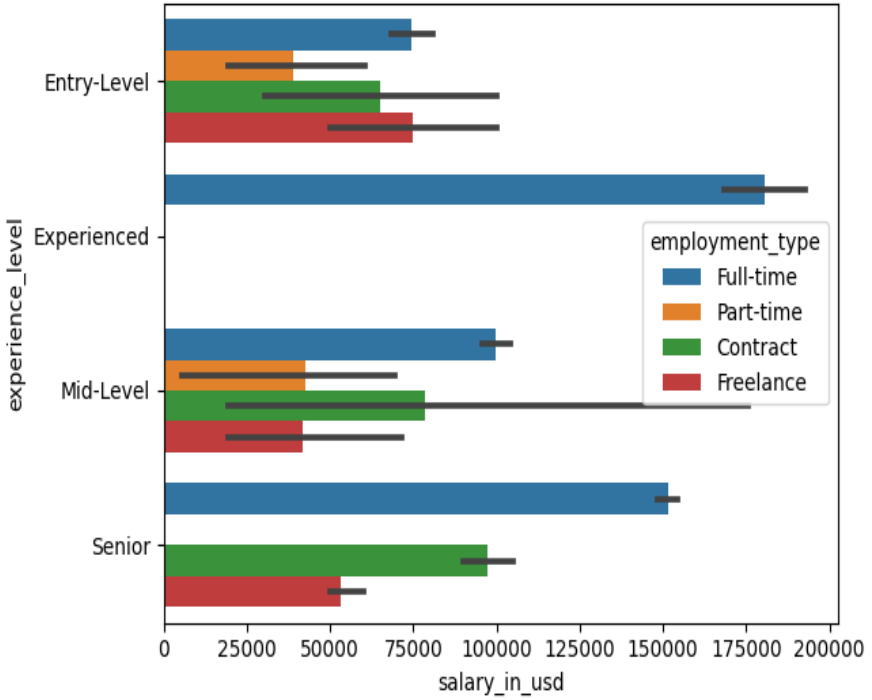
| | |
|---|---|
| **Distinct** | 3 |
| **Distinct (%)** | 0.1% |
| **Missing** | 0 |
| **Missing (%)** | 0.0% |
| **Memory size** | 18.4 KiB |

# GRÁFICOS

**Salario en dólares:** me brinda información de la distribución de los salarios, el promedio de los mismos se ubica @ USD 133.000

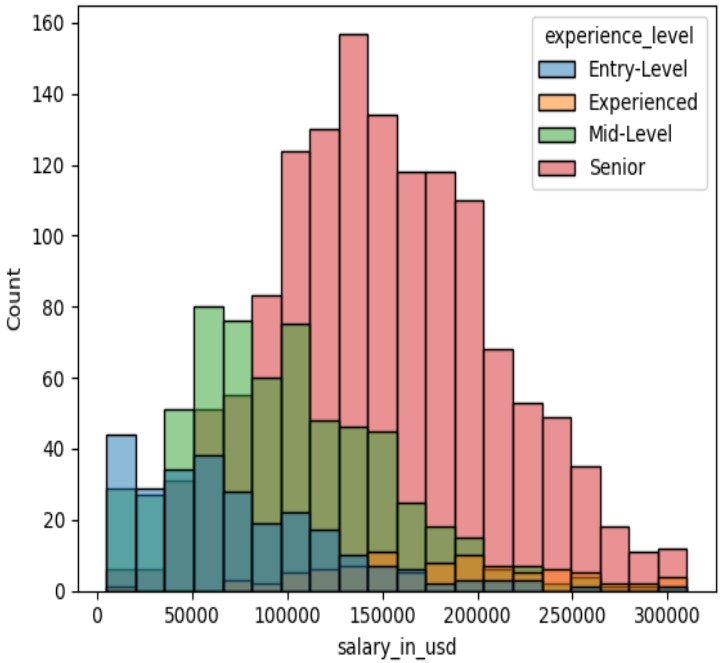



**Tipo de empleo**: Este gráfico bivariado me indica que a medida que va creciendo el nivel de experiencia y el salario, hay menos contratos Freelance y Part Time y más contratos contract y Full Time

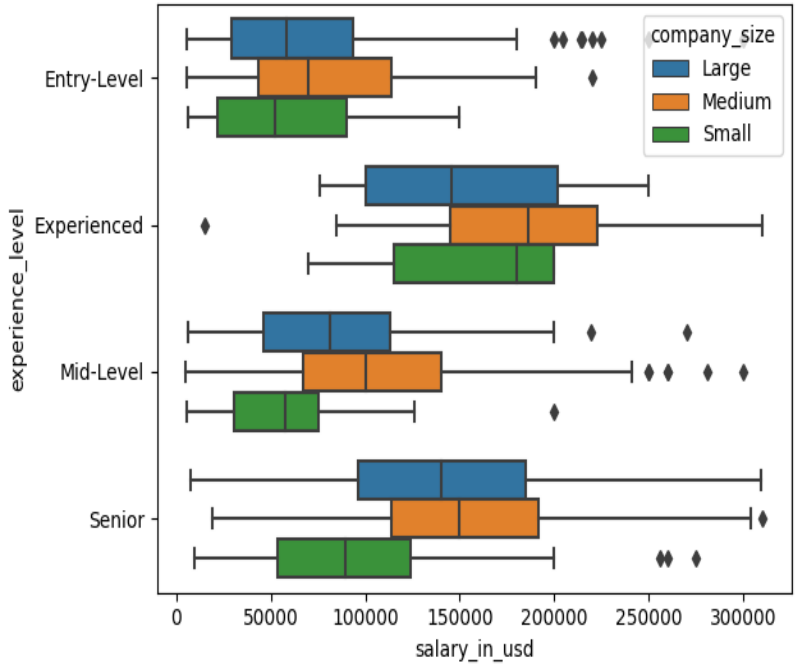


# GRÁFICOS

**Nivel de experiencia:** Agrupando por nivel de experiencia, podemos decir que los que tienen un nivel de experiencia senior ganan como media @ USD 151.000. Mientras que los que son Mid-level tienen un salario medio de @ USD 99.000. La media de los entry-level es de @ USD 73.000. Mientras que el de los experienced es usd 181.000

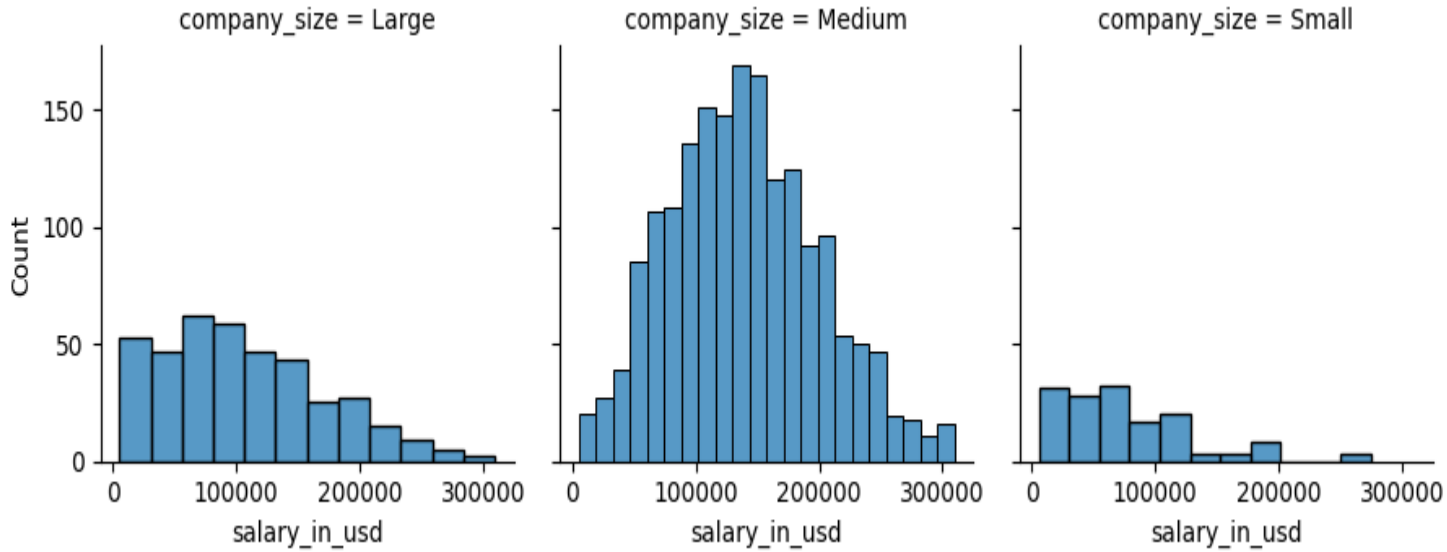



**Nivel de experiencia**: Este gráfico bivariado, me muestra que todos los niveles de salario son mejores en las empresas medianas y peores en las empresas pequeñas.



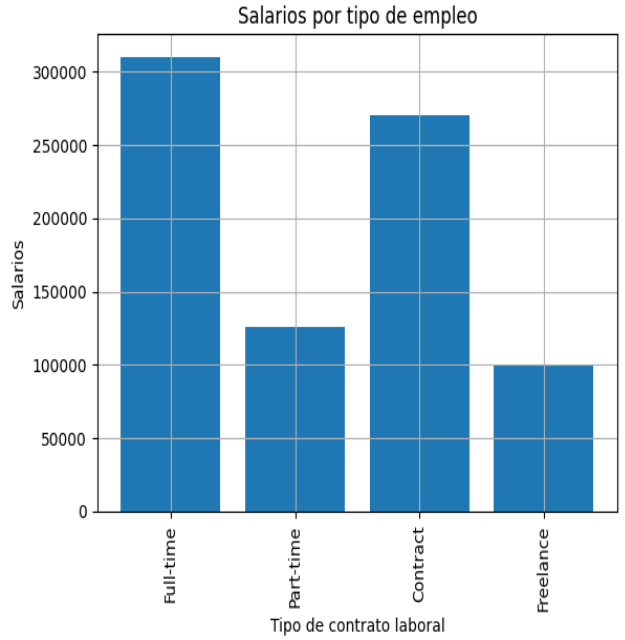
# GRÁFICOS

**Tamaño de la empresa:** La mayoría de los trabajadores corresponden a empresas medianas, donde el promedio de los salarios se ubica en USD 139.000, en las empresas grandes el promedio es un poco más bajo @ USD 106.000 con menos trabajadores. Finalmente, en las empresas chicas tiende a ser USD 75.000 con muy pocos trabajadores

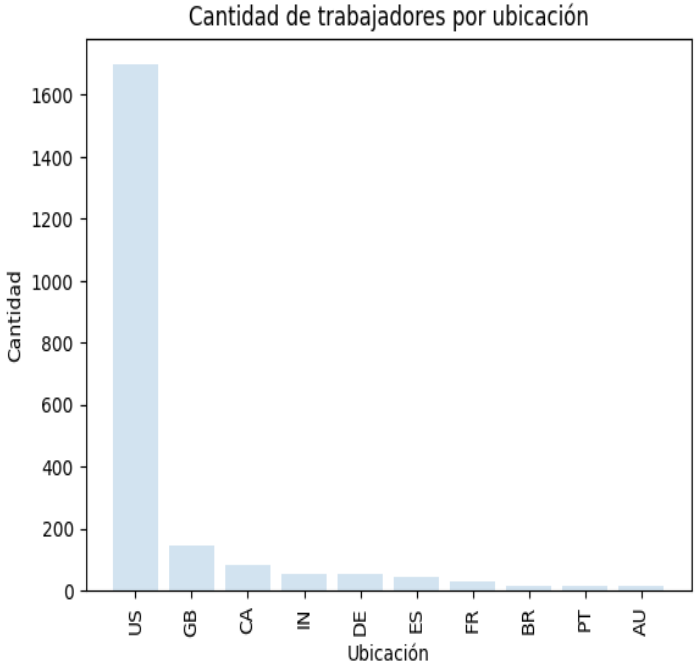


# GRÁFICOS

**Salario por tipo de contrato laboral:** mediante este análisis bivariado, podemos decir que con el tipo de contrato laboral que más dinero ganan es con el de Full Time, luego con el de Contract. Lejos figuran los contratos de Part time y por último Freelance.

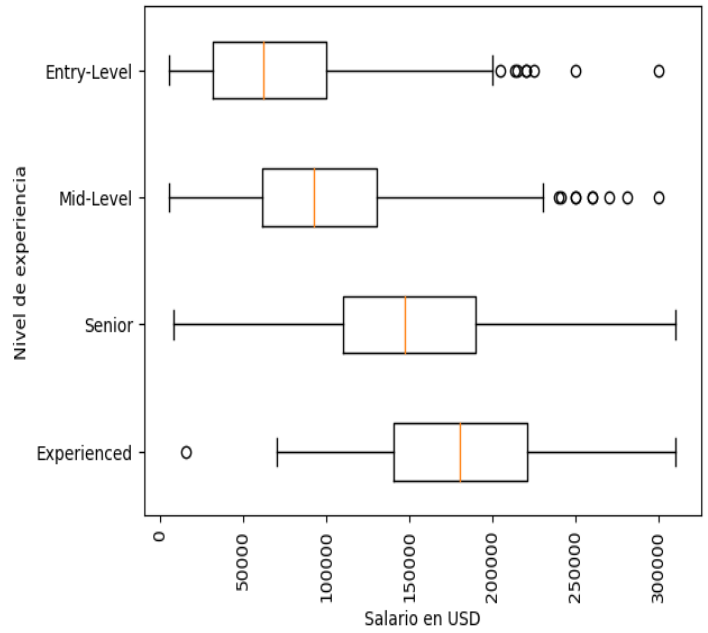



**Cantidad de trabajadores por ubicación**: En el país dónde más trabajadores hay de Data Science es lejos Estados Unidos.

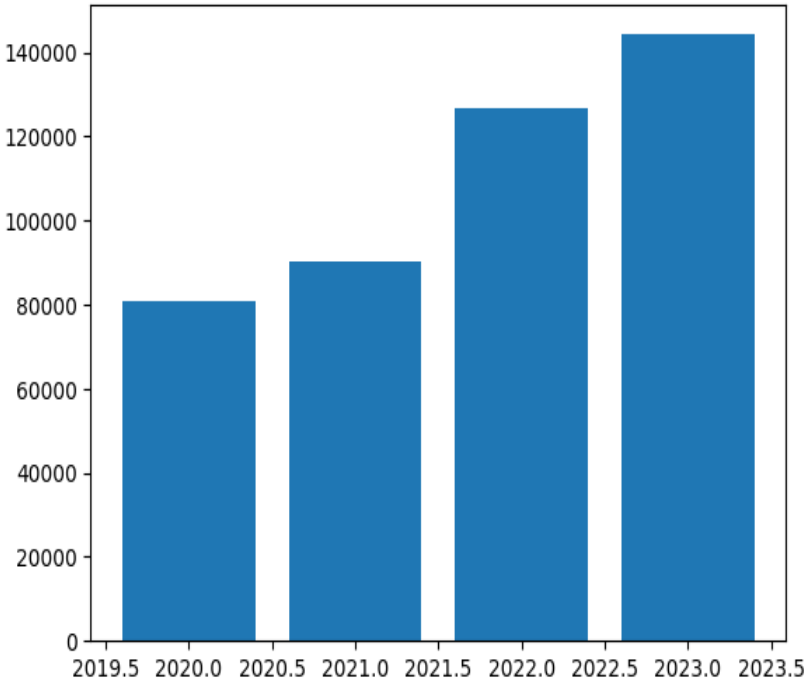


# GRÁFICOS

**Nivel de experiencia:** Según este gráfico, los que son Experienced tienen el mejor rango de sueldos, salvo por algunos trabajadores aislados Mid level y Entry Level.

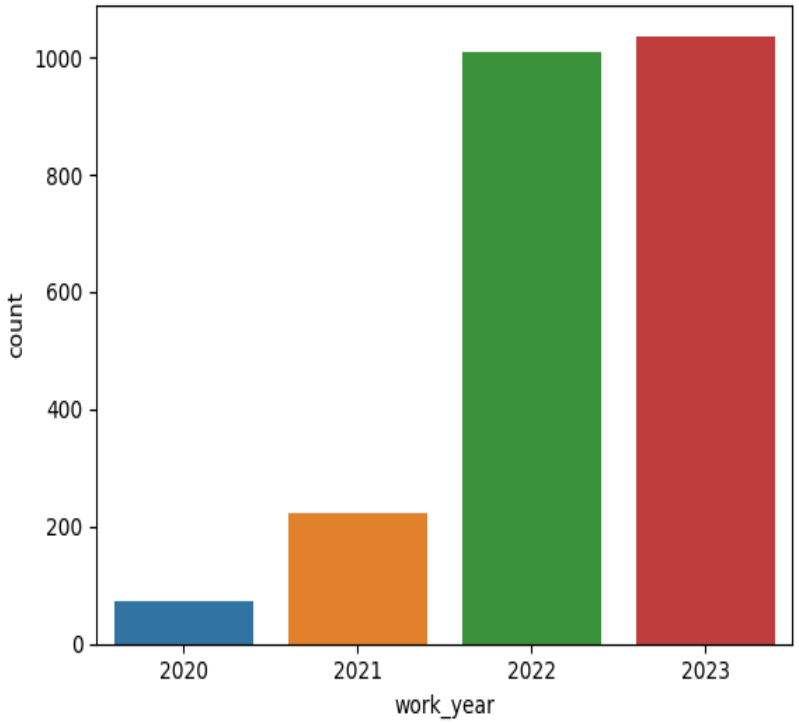



**Salario promedio por año:** El promedio de salarios se fue incrementando con los años, dando un salto importante del 2021 al 2022

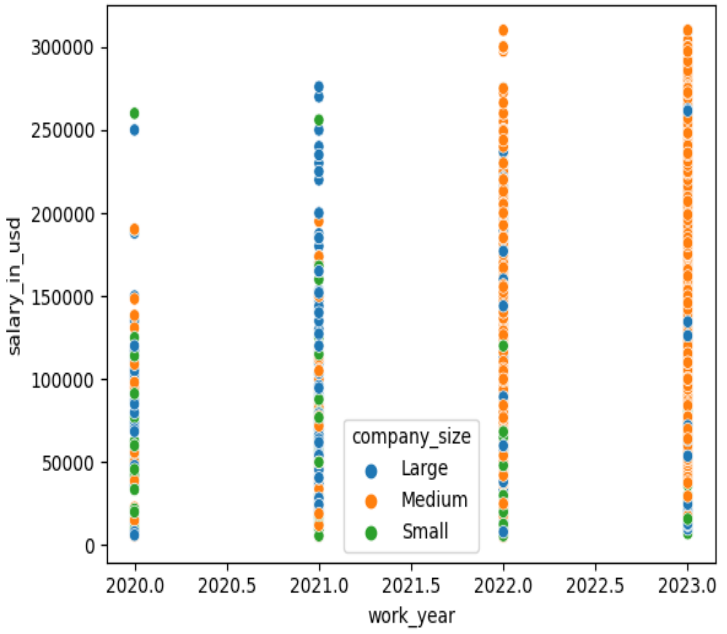


# GRÁFICOS

**Cantidad de trabajadores por año:** La cantidad de trabajadores empleados se incrementó considerablemente del 2021 al 2022




**Salario promedio por año por tipo de empresa:** El promedio de salarios fue creciendo durante los años. En el 2021 se puede decir que los salarios más altos y la mayor cantidad de trabajadores se ubicaban en las empresas grandes. En el 2022 y 2023, la mayor cantidad de trabajadores y los salarios más altos se ubicaban en empresas medianas

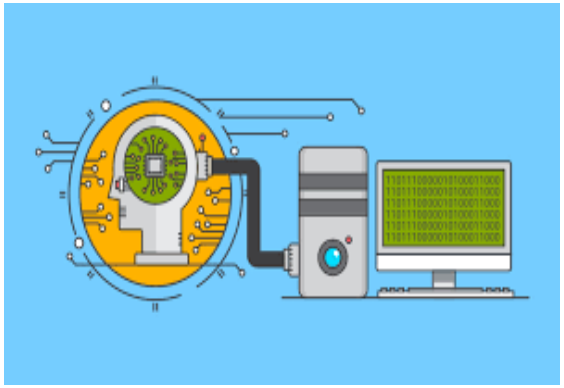


# *INSIGHTS*

- ✓ Luego de analizar todos los gráficos podríamos decir que el sueldo en promedio es mejor en empresas medianas y a su vez más trabajadores se sitúan en este tipo de empresas.
- ✓ Estados Unidos es lejos el país que más trabajadores de Data Science tiene.
- ✓ Un junior puede acceder a tener un contrato freelance o part time, luego a medida que aumenta la experiencia y el sueldo es más probable que requiera un contrato por tiempo determinado o full time.
- ✓ Podemos considerar también que los salarios van a ir en promedio de USD 72.000 a usd 181.000 dependiente el nivel de experiencia adquirido.
- ✓ El salario promedio fue aumentando a lo largo de los años con un salto importante en 2022.

# MACHINE LEARNING

1. Verificamos que utilizando todas las variables el R2 no era bueno (0.011). El modelo estaba overfiteando.
2. Hicimos el R2 de cada variable contra la variable sueldo en dólares, pero el mismo seguía sin ser el esperado y peor que con todas las variables.
3. Dividimos el data set en test y entrenamiento, utilizando todas las variables el R2 no es bueno (0.07)
4. Corro el modelo con las 30 mejores variables, mejoró el R2 (0.47)
5. Corro el modelo con las 15 mejores variables, el R2 es similar (0.46)




DANKSCHEEN
YAQHANYELAY
TASHAKKUR ATU
TINGKI
BÏYAN
SHUKRIA
GRACIAS
THANK
SUKSAMA
EKHMET
ARIGATO
SHUKURIA
MAAKE
GRAZIE
MEHRBANI
PALDIES
KOMAPSUMNIDA
GOZAIMASHITA
EFCHARISTO
YOU
JUSPAXAR
BOLZÏN
MERCI