



# Análisis salarial de científicos de datos.

Introducción a la ciencia de datos

Ricardo Xochitiotzi Flores

27/11/2024

M.C. Jaime Alejandro Romero Sierra



## 2 introduccion

### **Objetivo.**

El objetivo principal de este proyecto es analizar las relaciones entre diversas características laborales, como el tipo de empleo, nivel de experiencia, tamaño de la empresa, modalidad de trabajo (remoto o presencial), y los salarios en dólares (USD) de científicos de datos. Asimismo, se busca desarrollar un modelo de machine learning que permita predecir el salario en función de las características laborales disponibles.

### **Justificación del problema.**

El sector de la ciencia de datos se encuentra en constante crecimiento y transformación debido al impacto de la digitalización y la tecnología. Este análisis resulta relevante porque permite identificar patrones salariales y factores determinantes en los salarios, ayudando a los profesionales del área a tomar decisiones informadas sobre su carrera, y a las empresas a estructurar mejores políticas de compensación. Estudiar esta problemática también aporta información valiosa sobre cómo la modalidad de trabajo remoto y el tamaño de la empresa influyen en las remuneraciones, temas críticos en el contexto laboral actual.

### **Fuentes de datos.**

Los datos fueron recabados de la página para científicos de datos “Kaggle”, la cual era una base de datos obtenidos de empresas internacionales y recabados todos los datos en una sola base de datos, originalmente contaba con 47,579 datos y 11 columnas, después de la limpieza de datos cuenta con 22,124 datos y 9 columnas, notamos categorías que nos ayudaran a filtrar los datos para tener un mejor control y visualización de los datos como lo son: Nivel de experiencia(experience\_level) y tipo de empleo(employment\_type).



### 3.- Metodología

Procesamiento de limpieza de datos

## Análisis inicial

### Resumen estadístico de los datos

Nuestro dataframe cuenta con un total de 47,579 filas y 11 columnas

```
#Visualizamos la cantidad total de filas y columnas (en ese orden)
df.shape
```

[4] ✓ 0.0s

.. (47579, 11)

**Porcentaje de valores faltantes** por columna(En todos los casos el porcentaje de NaN son los mismos)

```
#Calculando el porcentaje de valores faltantes por columna
df.isnull().mean()*100
```

✓ 0.0s

work_year	3.999664
experience_level	3.999664
employment_type	3.999664
job_title	3.999664
salary	3.999664
salary_currency	3.999664
salary_in_usd	3.999664
employee_residence	3.999664
remote_ratio	3.999664
company_location	3.999664
company_size	3.999664
dtype:	float64



## Total de filas duplicadas

```
#Suma total de los datos duplicados
df.duplicated().sum()
```

✓ 0.0s

np.int64(14793)

En la descripción de **tipos de datos** encontramos ciertos problemas; Como primera observación tenemos que la columna llamada “salary\_in\_usd” esta como tipo objeto y no como tipo numérico(int o float), lo mismo pasa con la fecha que para mejor uso debería estar en tipo “date”

```
#Tipo de valor por columna
df.info()
```

[6] ✓ 0.0s

```
... <class 'pandas.core.frame.DataFrame'>
RangeIndex: 47579 entries, 0 to 47578
Data columns (total 11 columns):
#   Column                Non-Null Count  Dtype
---  -
0   work_year              45676 non-null  object
1   experience_level        45676 non-null  object
2   employment_type        45676 non-null  object
3   job_title              45676 non-null  object
4   salary                 45676 non-null  float64
5   salary_currency        45676 non-null  object
6   salary_in_usd          45676 non-null  object
7   employee_residence     45676 non-null  object
8   remote_ratio           45676 non-null  float64
9   company_location       45676 non-null  object
10  company_size           45676 non-null  object
dtypes: float64(2), object(9)
memory usage: 4.0+ MB
```



## Proceso de limpieza de datos

### Eliminación o imputación de datos faltantes

Eliminando las columnas con NaN de salary\_currency y salary. Hacemos esto ya que tenemos otra columna que nos facilita la conversión de los salarios a dólares. Esto debido a que no ocuparemos el salario inicial y la moneda en la que se dio dicho salario. Entonces esto quiere decir que no afectará a nuestros modelos de predicción ya que no va relacionado con el objetivo

```
#Eliminando las columnas con NaN de salary_currency y salary
#Hacemos esto ya que tenemos otra columna que nos facilita la conversión de los salarios a dolares.
#Esto debido a que no ocuparemos el salario inicial y la moneda en la que se dio dicho salario.
#Entonces esto quiere decir que no afectará a nuestros modelos de predicción ya que no va relacionado con el objetivo
df2=df.drop(columns=['salary'])
df2=df2.drop(columns=['salary_currency'])
df2
```

188] ✓ 0.0s Python

	work_year	experience_level	employment_type	job_title	salary_in_usd	employee_residence	remote_ratio	company_location	company
0	2020	EN	FT	Azure Data Engineer	100000	MU	0.0	MU	
1	2020	EN	CT	Staff Data Analyst	44753	CA	NaN	CA	
2	2020	SE	FT	Staff Data Scientist	164000	US	50.0	US	
3	2020	EN	FT	Data Analyst	47899	DE	0.0	DE	
4	2020	EX	FT	Data Scientist	300000	US	100.0	US	

En la siguiente imagen se muestra como decidí imputar las demás columnas



```
#TIPO DE IMPUTACION DE DATOS PARA CADA COLUMNA
#Año -- moda
#experience -- nueva cat
#tipoempleo -- moda
#titulo -- nueva cat
#salario usd -- media
#residence -- moda
#remote_ratio -- moda
#ubi comp -- nueva cat
#tamaño comp -- moda
```

- Moda: Simple de implementar y efectivo si hay una categoría dominante en la columna.
- Nueva categoría: No introduce sesgo en los datos, y mantiene la información de que originalmente había valores faltantes, esto ayuda a no alterar demasiado los datos.
- Media: Para datos numéricos, si el dataframe no tiene muchos outliers (valores extremadamente altos o bajos), puedes utilizar la media como referencia para la imputación.

```
#Imputando los datos para no alterar demasiado los analisis de predicción
df3['work_year'] = df3['work_year'].fillna(df3['work_year'].mode()[0])
df3['experience_level'] = df3['experience_level'].fillna('Unknown')
df3['employment_type'] = df3['employment_type'].fillna(df3['employment_type'].mode()[0])
df3['job_title'] = df3['job_title'].fillna('Unknown')
df3['salary_in_usd'] = df3['salary_in_usd'].fillna(df3['salary_in_usd'].mean())
df3['employee_residence'] = df3['employee_residence'].fillna(df3['employee_residence'].mode()[0])
df3['remote_ratio'] = df3['remote_ratio'].fillna(df3['remote_ratio'].mode()[0])
df3['company_location'] = df3['company_location'].fillna('Unknown')
df3['company_size'] = df3['company_size'].fillna(df3['company_size'].mode()[0])
```

314] ✓ 0.0s

+ Código + Markdown



```

df3.isnull().sum()
16] ✓ 0.0s

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

```

## Eliminación de duplicados

Visualizando los datos duplicados

```

#Visualizando los datos duplicados
df4[df4.duplicated()]
✓ 0.0s

```

	work_year	experience_level	employment_type	job_title	salary_in_usd	employee_residence	remote_ratio	company_location	com
137	2021	MI	FT	Data Scientist	150000.0	US	100.0	US	
179	2021	MI	FT	Data Engineer	200000.0	US	100.0	US	
241	2021	EN	FT	Data Scientist	90000.0	US	100.0	US	
265	2021	MI	FT	Data Scientist	90734.0	DE	50.0	DE	
350	2022	SE	FT	Data Engineer	175000.0	US	0.0	US	
...	...	...	...	...	...	...	...	...	...
47573	2024	SE	FT	Machine Learning Engineer	180000.0	US	0.0	US	
47574	2024	SE	FT	Engineer	246000.0	US	0.0	US	
47575	2024	MI	FT	Machine Learning Engineer	171000.0	US	0.0	US	



## Eliminando los datos duplicados

```
#Eliminando datos duplicados
df5=df4.drop_duplicates()
df5
```

Python

	work_year	experience_level	employment_type	job_title	salary_in_usd	employee_residence	remote_ratio	company_location	com
0	2020	EN	FT	Azure Data Engineer	100000.0	MU	0.0	MU	
1	2020	EN	CT	Staff Data Analyst	44753.0	CA	0.0	CA	
2	2020	SE	FT	Staff Data Scientist	164000.0	US	50.0	US	
3	2020	EN	FT	Data Analyst	47899.0	DE	0.0	DE	
4	2020	EX	FT	Data Scientist	300000.0	US	100.0	US	
...	...	...	...	...	...	...	...	...	...
47516	2021	SE	FT	Software	140350.0	US	0.0	US	

## Corrección de los datos

```
#Convirtiendo a entero y fecha los datos que lo necesitan
df3['salary_in_usd']=pd.to_numeric(df3['salary_in_usd'],errors='coerce')
df3['salary_in_usd']=df3['salary_in_usd'].astype(float)

df3['work_year']=pd.to_numeric(df3['work_year'],errors='coerce')
df3['work_year']=df3['work_year'].astype(float)

df3.info()
```

Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 47579 entries, 0 to 47578
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   work_year              44752 non-null  float64
1   experience_level       45676 non-null  object
2   employment_type        45676 non-null  object
3   job_title              45676 non-null  object
4   salary_in_usd          44765 non-null  float64
5   employee_residence     45676 non-null  object
6   remote_ratio           45676 non-null  float64
7   company_location       45676 non-null  object
8   company_size           45676 non-null  object
dtypes: float64(3), object(6)
memory usage: 3.3+ MB
```





## Corrección de valores invalid

```
#Mostrando los invalid values

for i in (df3.columns) :
    print(f"En la columna {i} los invalid_value son: {df3[df3[i] == 'bbb'].shape[0]}")
```

✓ 0.0s

En la columna work\_year los invalid\_value son: 0  
 En la columna experience\_level los invalid\_value son: 912  
 En la columna employment\_type los invalid\_value son: 0  
 En la columna job\_title los invalid\_value son: 0  
 En la columna salary\_in\_usd los invalid\_value son: 0  
 En la columna employee\_residence los invalid\_value son: 913  
 En la columna remote\_ratio los invalid\_value son: 0  
 En la columna company\_location los invalid\_value son: 0  
 En la columna company\_size los invalid\_value son: 0

## Eliminando los invalid

```
#Eliminamos los invalid_values('bbb')
df4=df3[df3['experience_level']!='bbb']
df4=df4[df4['employee_residence']!='bbb']
df4=df4[df4['work_year']!='bbb']

for i in (df4.columns) :
    print(f"En la columna {i} los invalid_value son: {df4[df4[i] == 'bbb'].shape[0]}")
```


19] ✓ 0.0s

En la columna work\_year los invalid\_value son: 0  
 En la columna experience\_level los invalid\_value son: 0  
 En la columna employment\_type los invalid\_value son: 0  
 En la columna job\_title los invalid\_value son: 0  
 En la columna salary\_in\_usd los invalid\_value son: 0  
 En la columna employee\_residence los invalid\_value son: 0  
 En la columna remote\_ratio los invalid\_value son: 0  
 En la columna company\_location los invalid\_value son: 0  
 En la columna company\_size los invalid\_value son: 0



## Resultados

La base de datos nos queda con las siguientes características:

 <pre>df5.shape</pre> <p>✓ 0.0s</p> <p>(22124, 9)</p>	<p>22,124 filas</p> <p>9 columnas</p> <p>3 datos de tipo numérico (año, salario, trabajo remoto)</p>
--	--

Comprobaciones de que no hay 'Nan' y 'bbb'

```
df5.duplicated().sum()
np.int64(0)
```

```
for i in (df5.columns) :
    print(f"En la columna {i} los invalid_value son: {df5[df5[i] == 'bbb'].shape[0]}")
```

En la columna work\_year los invalid\_value son: 0  
 En la columna experience\_level los invalid\_value son: 0  
 En la columna employment\_type los invalid\_value son: 0  
 En la columna job\_title los invalid\_value son: 0  
 En la columna salary\_in\_usd los invalid\_value son: 0  
 En la columna employee\_residence los invalid\_value son: 0  
 En la columna remote\_ratio los invalid\_value son: 0  
 En la columna company\_location los invalid\_value son: 0  
 En la columna company\_size los invalid\_value son: 0

Tabla que muestre el porcentaje de valores faltantes final por columna



```

df6=df5.isnull().mean() * 100
df6
2] ✓ 0.0s

work_year      0.0
experience_level 0.0
employment_type 0.0
job_title       0.0
salary_in_usd   0.0
employee_residence 0.0
remote_ratio    0.0
company_location 0.0
company_size    0.0
dtype: float64

```

### 3.1.- Descripción general de los datos

Notamos que después de terminar con la limpieza general de los datos nos queda lo siguiente:

```

df5.shape
24] ✓ 0.0s

(22124, 9)

```

22,124 filas

9 columnas

- Categóricas: job\_title, experience\_level, employment\_type, company\_location, experience\_level, company\_size.
- Numéricas: salary\_in\_usd, remote\_ratio.



```

'Resumen Estadístico Numérico':
work_year  salary_in_usd  remote_ratio
count  22124.000000  22124.000000  22124.000000
mean    2023.619463  156309.909332   26.701772
std      0.667846   74687.381612   43.911717
min     2020.000000   15000.000000    0.000000
25%     2023.000000   104800.000000    0.000000
50%     2024.000000   148000.000000    0.000000
75%     2024.000000   195050.000000   100.000000
max     2024.000000   800000.000000   100.000000,
'Frecuencia de Categorías': {'experience_level': experience_level
SE      11513
MI      6226
EN      2110
Unknown  1622
EX       653
Name: count, dtype: int64,
'employment_type': employment_type
...
'company_size': company_size
M      21035
L       884
S       205
Name: count, dtype: int64}}

```

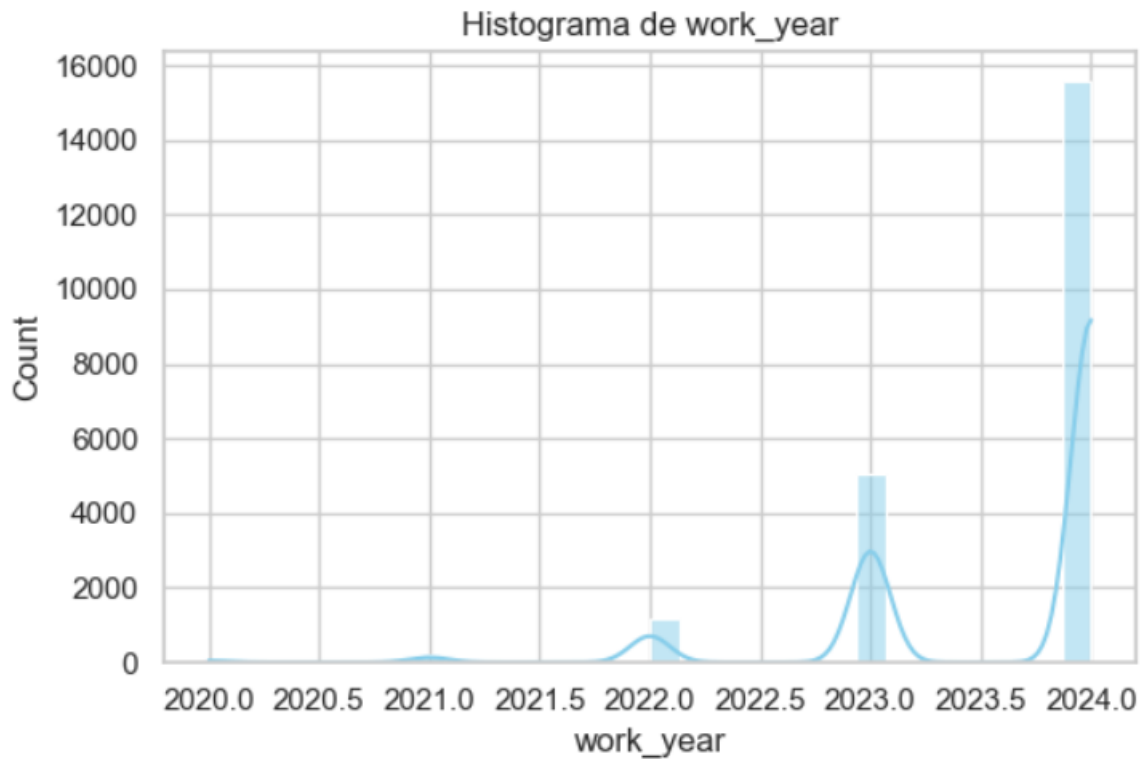
- **Tendencias temporales:** Los datos se centran en los años más recientes (2023-2024), probablemente reflejando cambios en las dinámicas laborales.
- **Salarios:** Aunque la mayoría de los salarios son altos (en promedio \$156,309), hay una gran dispersión, lo que sugiere la inclusión de trabajos con diferentes niveles de responsabilidad, experiencia y localización.
- **Trabajo remoto:** Aunque todavía prevalecen los empleos presenciales, el trabajo remoto está ganando terreno, especialmente en roles tecnológicos y científicos.



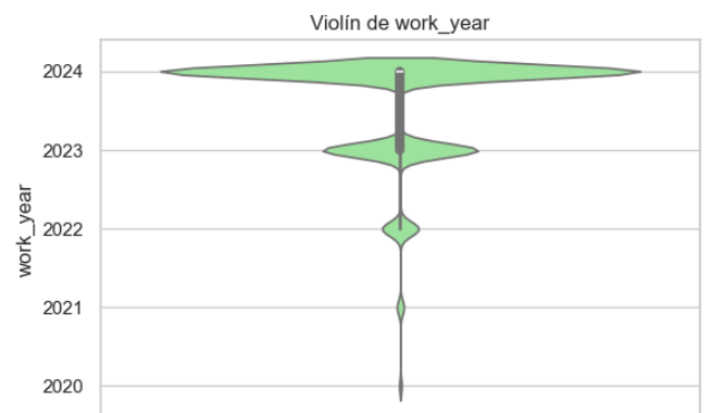
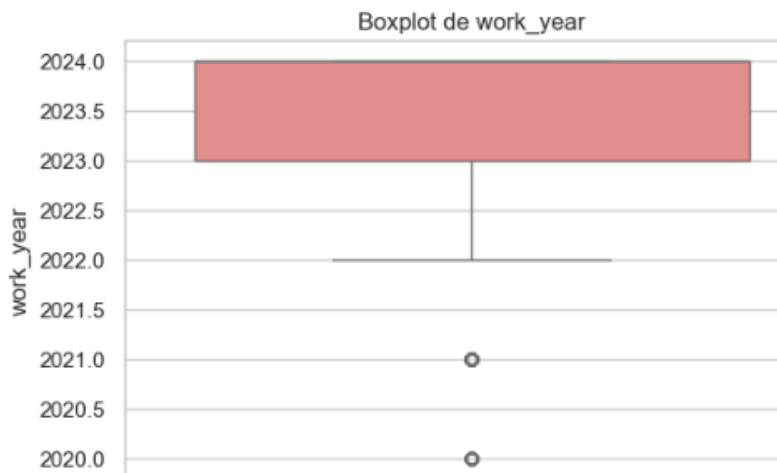
### 3.2.- Visualización y Distribución de Variables Individuales

- Observaciones en variables numéricas.

1.- Año de trabajo(work\_year en el que se registró)



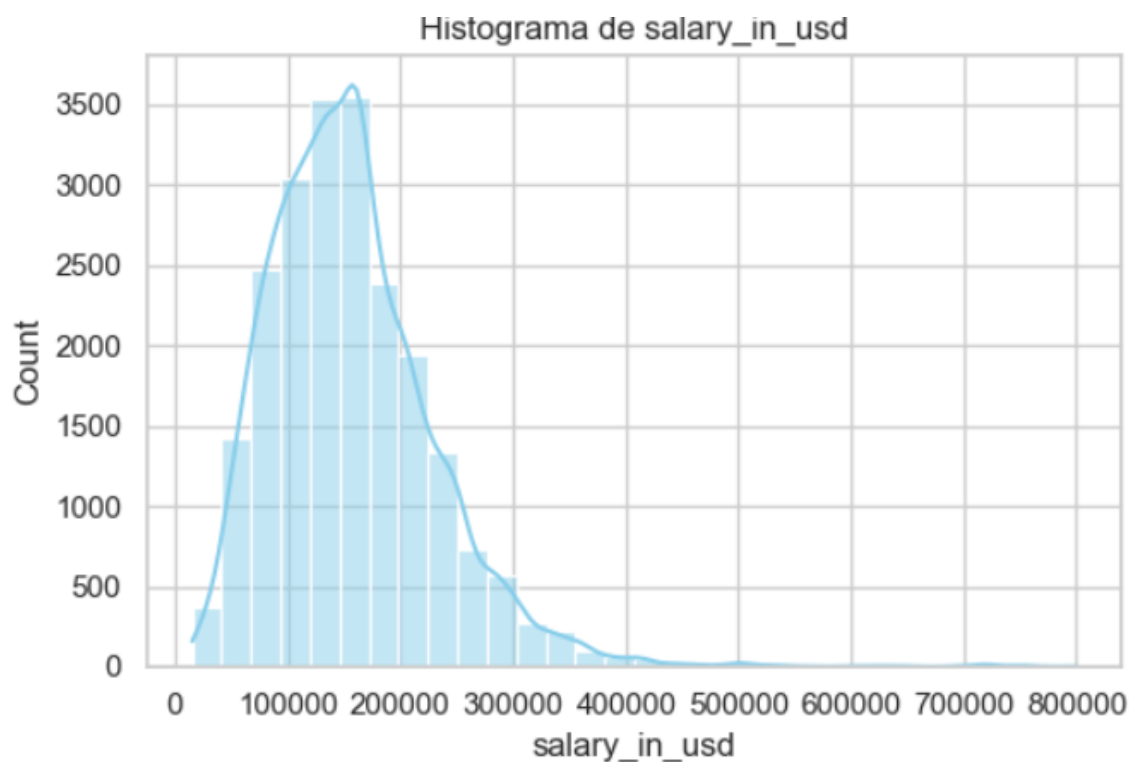
En el siguiente histograma podemos observar algo que nos puede servir a deducir que esta carrera se ha disparado por completo en el 2024 ya que la distribución está sesgada hacia 2024, lo que confirma que la mayoría de los registros pertenecen a este año.



En cuanto al boxplot y el gráfico de violín no se observan más cambios, ya que no se observan outliers significativos, ya que los años están limitados entre 2020 y 2024..

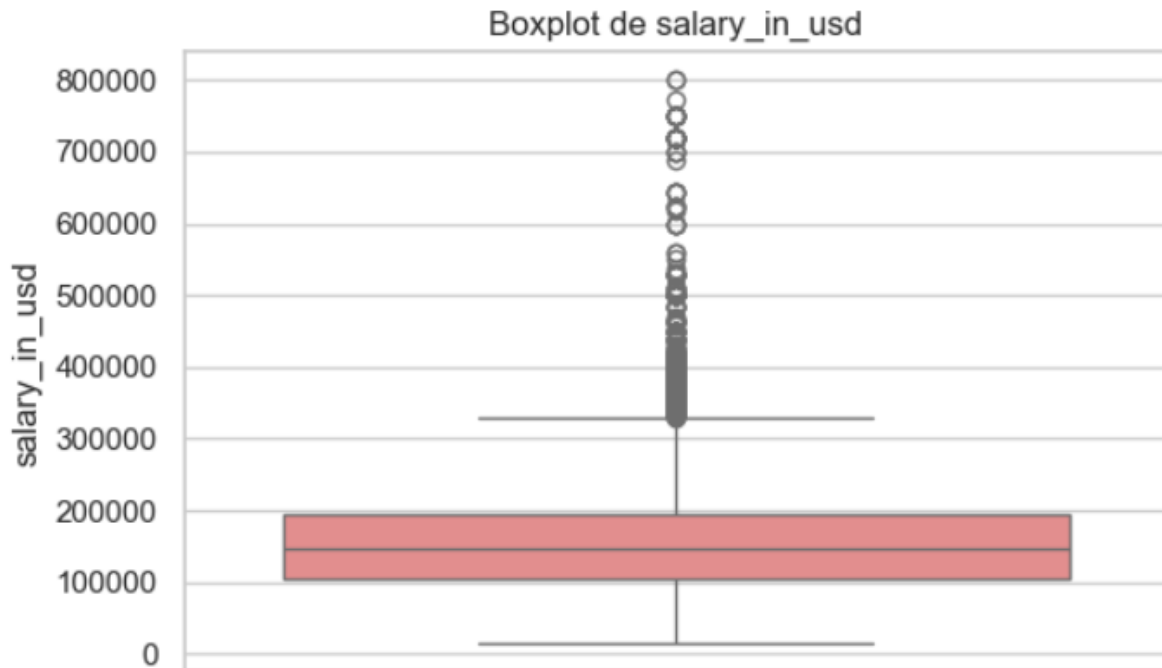
## 2.- Salario (salary\_in\_usd)

En cuanto al salario las gráficas nos brindan mas información importante de la cual nos estaremos apoyando

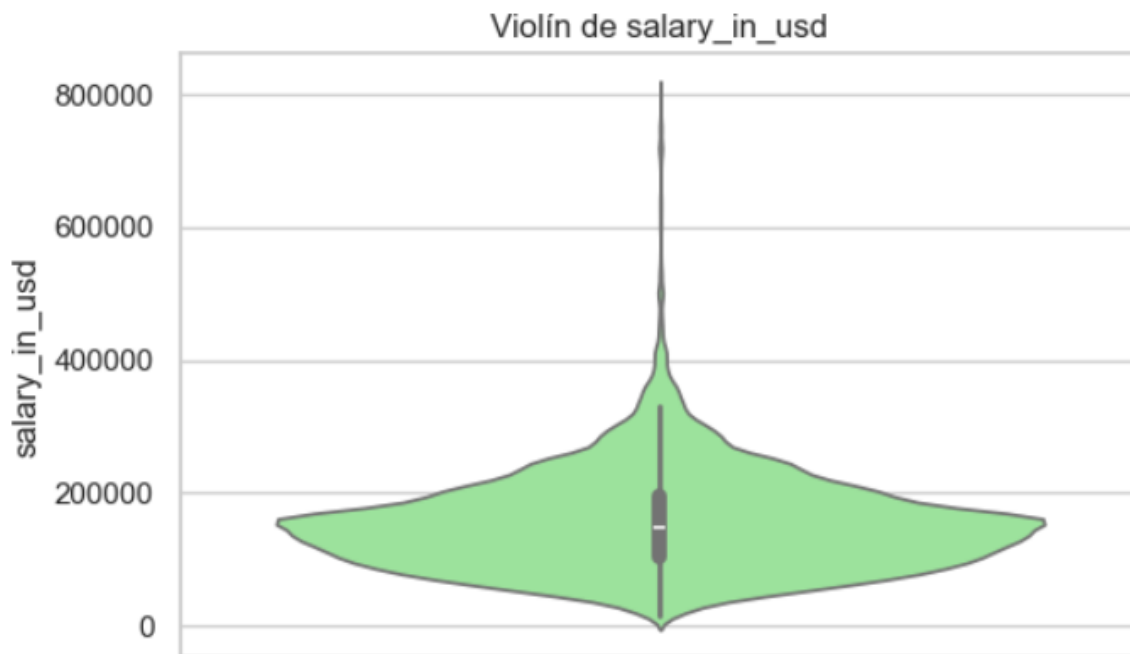


En la siguiente gráfica podemos observar como la distribución es asimétrica positiva, con una concentración alrededor de \$100,000-\$200,000 y algunos valores extremos por encima de \$500,000. Los valores que realmente nos importan van a ser los que se encuentran de nuestro rango \$100,000-\$200,000, ya que esto nos puede dar una gran pista mas adelante si es directamente proporcional a la experiencia y titulo laboral de cada científico de datos.

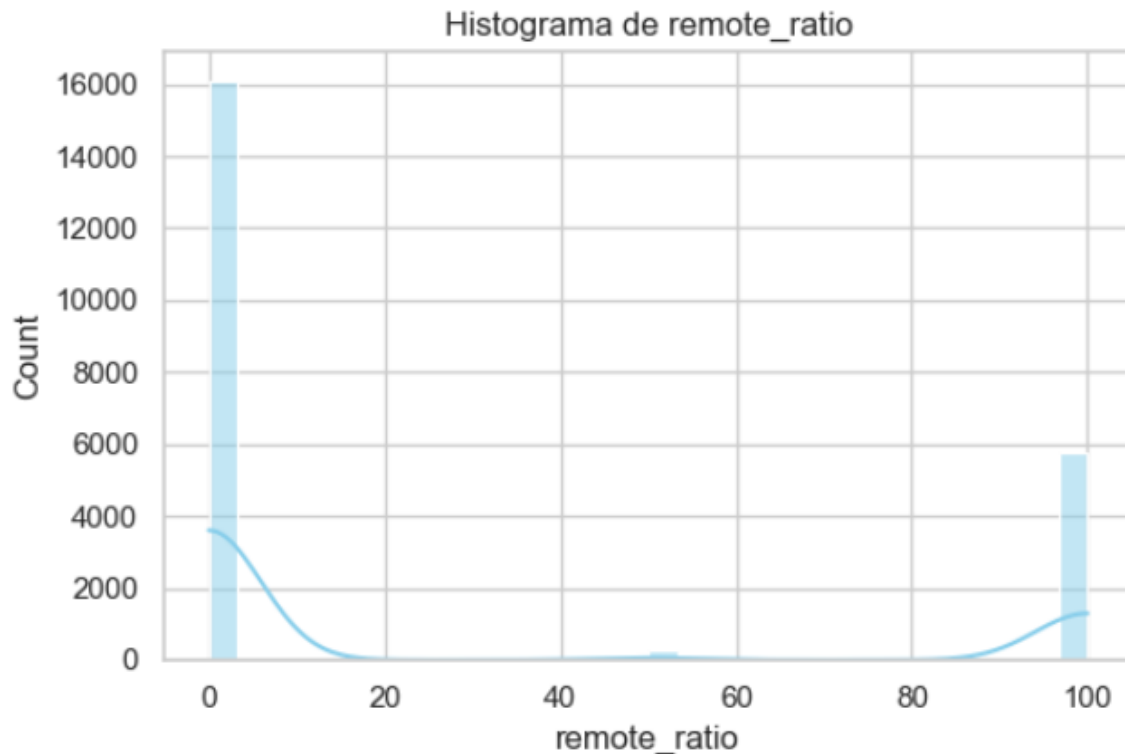




En la siguiente gráfica podemos notar más los outliers, identificamos diversos outliers en el rango más alto de salarios, como el salario máximo de \$800,000. Notamos que para no afectar los resultados estos serán los valores que tendremos que eliminar.



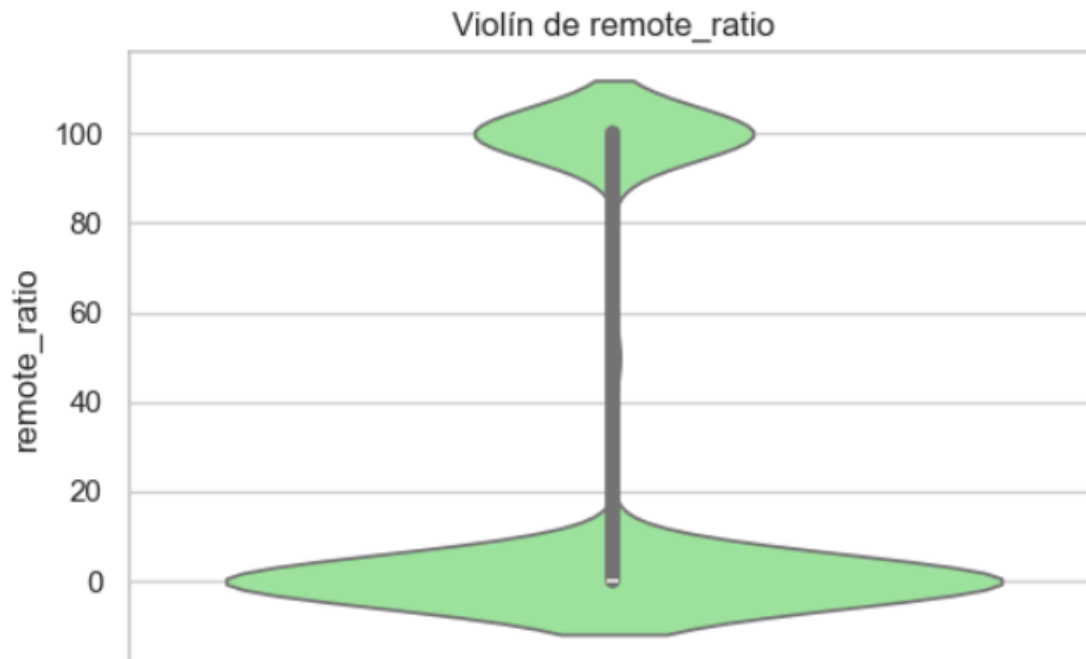
Aquí nuevamente nos confirma lo que ya hemos podido observar en las dos anteriores gráficas, nos muestra densidad en salarios intermedios (alrededor de (\$150,000), pero una dispersión significativa hacia valores altos.



La mayoría de los trabajos son presenciales (remote\_ratio = 0), con menos empleos híbridos (50%) y completamente remotos (100%), esto nos da una pista clave, y es que podemos empezar a darnos cuenta de que esto puede dar una mayor probabilidad de tener un mejor salario.



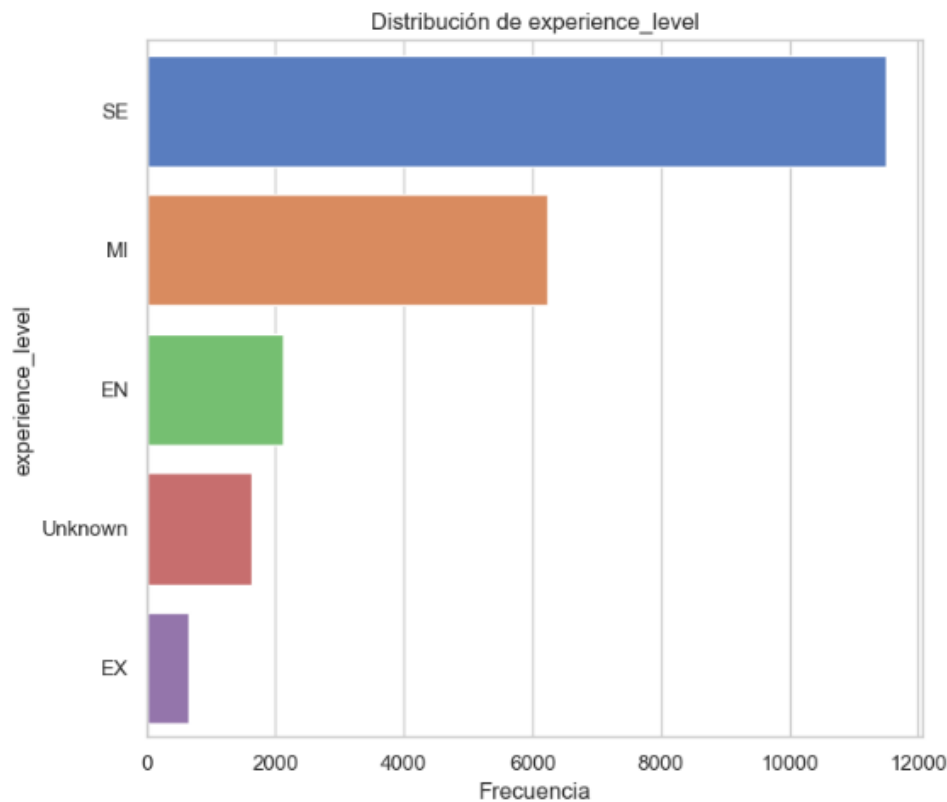




Reflejan la concentración en trabajos presenciales, pero también la presencia de empleos con mayor flexibilidad en donde del 80% – 100 % se concentran trabajos con una mejor flexibilidad.

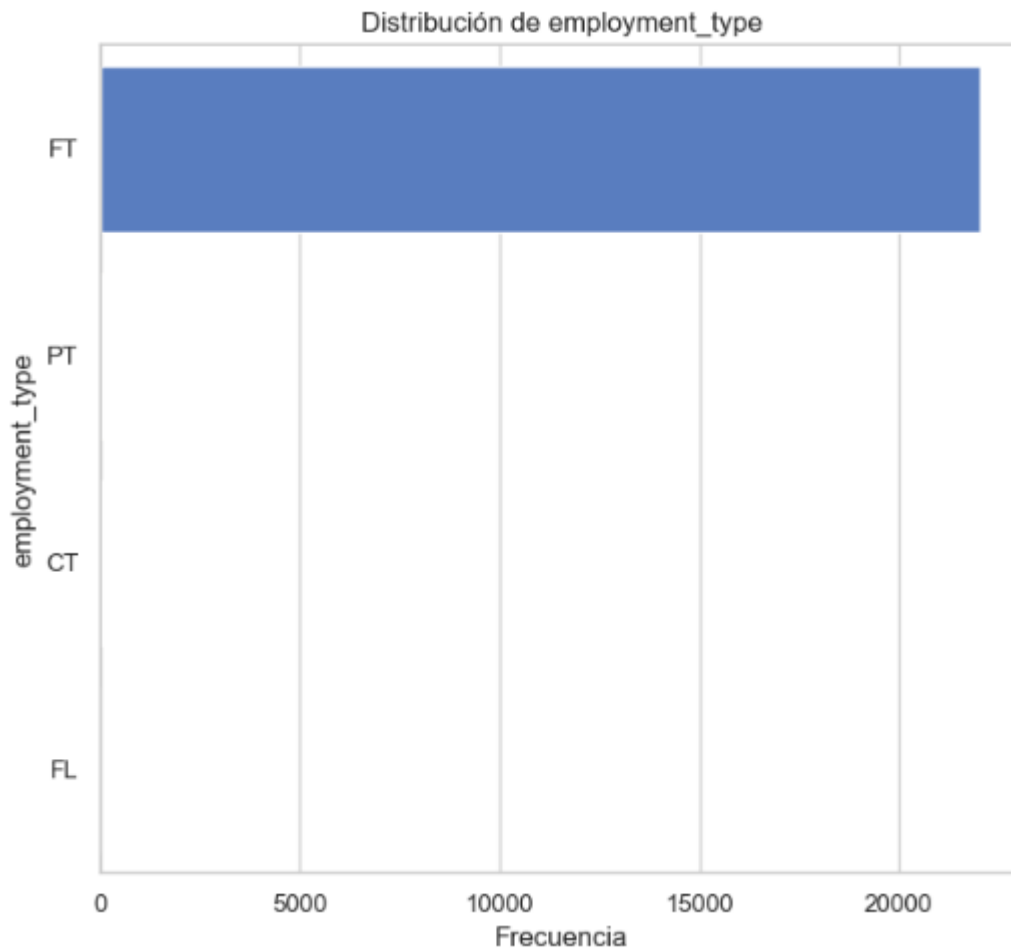


- Observación de variables categóricas



Notamos que tiene una distribución dominada por niveles "SE" (senior-level) y "MI" (mid-level), lo que indica que la mayoría de los roles requieren experiencia previa. Las categorías menores son: "EX" (executive-level) y "EN" (entry-level), además de algunos registros etiquetados como "Unknown" que son registros desconocidos.

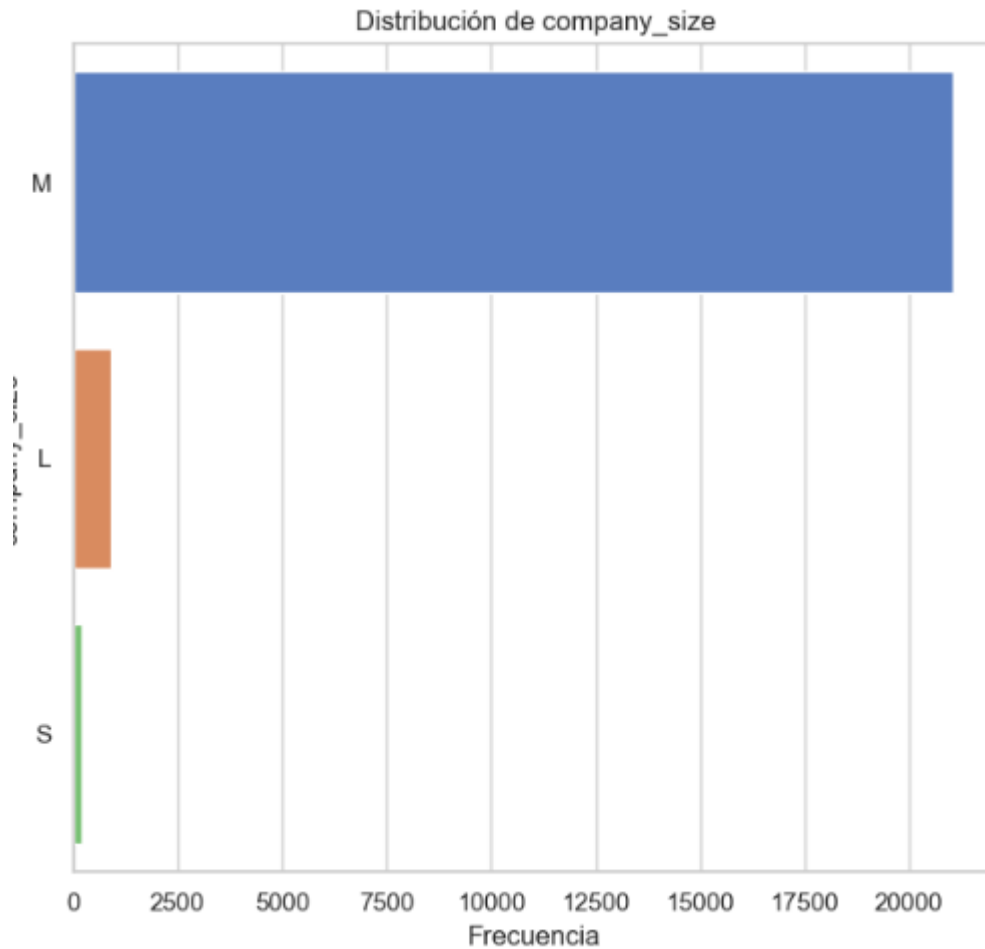




- **PT** Part-time
- **FT** Full-time
- **CT** Contract
- **FL** Freelance

Aquí si notamos un completo dominio por parte de la categoría FT, la categoría "FT" (full-time) domina completamente. Los tipos de empleo no tradicionales como "PT", "CT", y "FL" son raros.



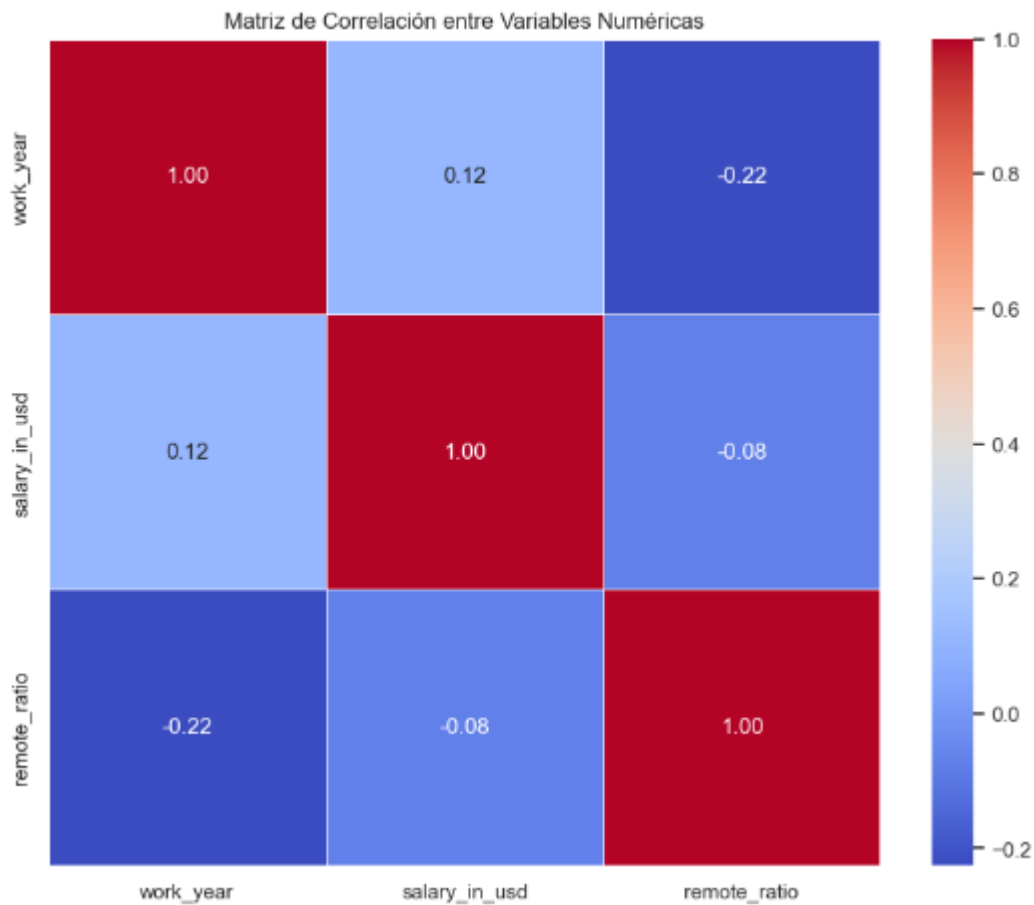


Aquí igualmente, notamos una gran dominación por parte de una categoría dominada por empresas medianas ("M"), con menos representación de empresas grandes ("L") y pequeñas ("S").

- **Variables Numéricas:** Las distribuciones revelan patrones interesantes, como la alta concentración en años recientes, los outliers en salarios, y la prevalencia de empleos presenciales.
- **Variables Categóricas:** Existen categorías dominantes en casi todas las variables, destacando la concentración en roles senior y empresas medianas, además de la ubicación en los Estados Unidos.



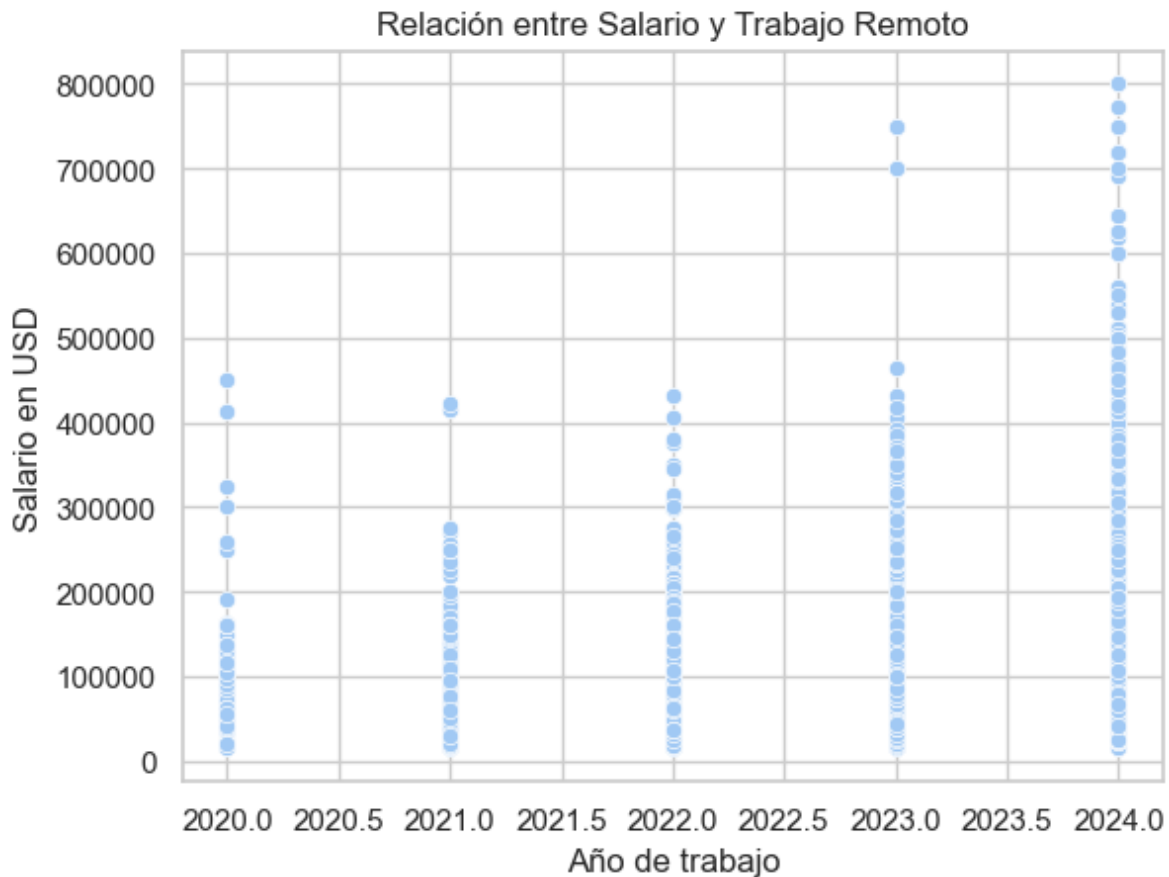
### 3.3.- Correlación entre variables



Al tener pocas variables numéricas y más categóricas, el heatmap es pequeño, pero aun así nos brinda información de correlaciones importantes.

La primera correlación que notamos es la del salario (salary\_in\_usd) con el año de trabajo (work\_year) y con esta información podemos llegar a una idea clara.

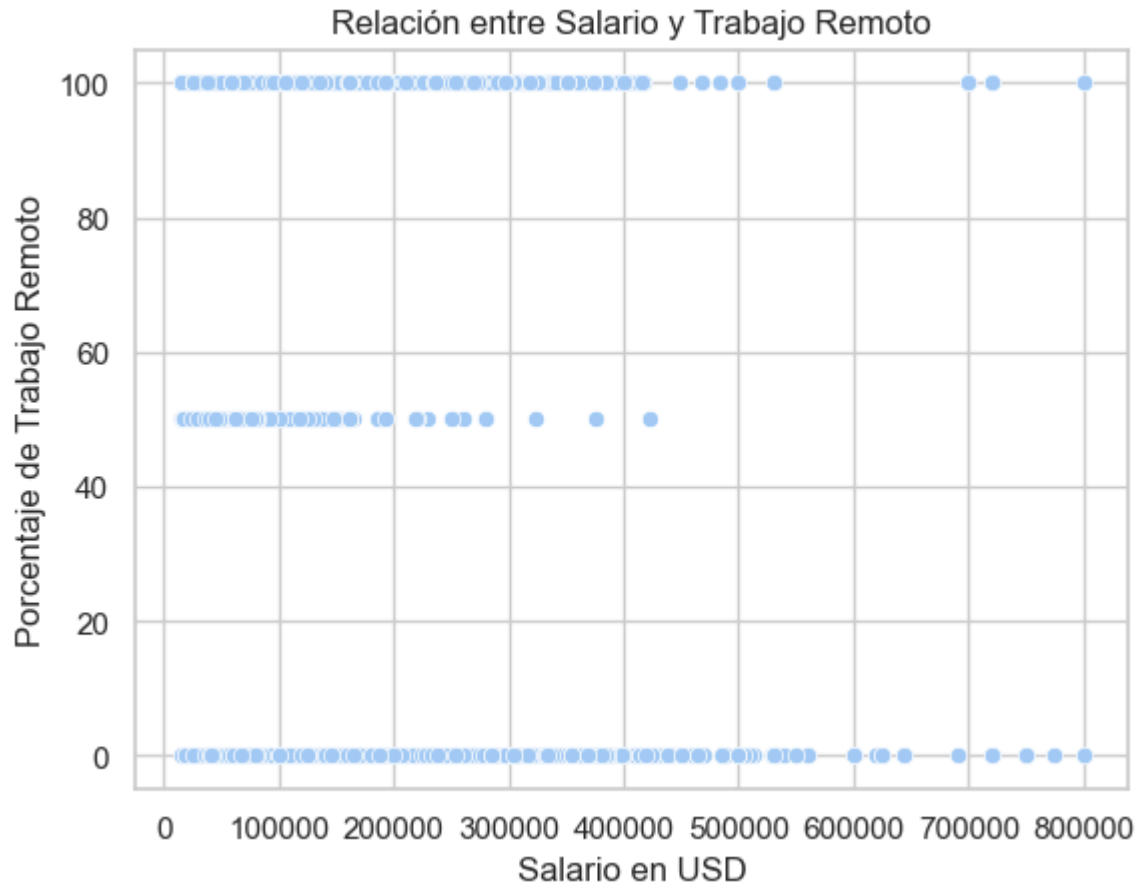




En esta gráfica de dispersión (scatter plot) podemos darnos cuenta de dos cosas, la primera es el aumento significativo de valor en cuanto a salario, y la segunda es el gran aumento y demanda de como va escalando, es decir, oferta y demanda que con el paso del tiempo se verá completamente superado a los números actuales.

Y la segunda correlación que podemos notar es la de el salario (salary\_in\_usd) y el trabajo remoto (remote\_ratio), esto es un punto importante ya que podemos ver que el salario depende de principalmente 2 porcentajes, 0% (sin trabajo remoto) y 100% (trabajo totalmente remoto).





Esto nos demuestra que, en el futuro análisis, podremos encontrar que estos dos datos si pueden afectar directamente al salario que ganes.

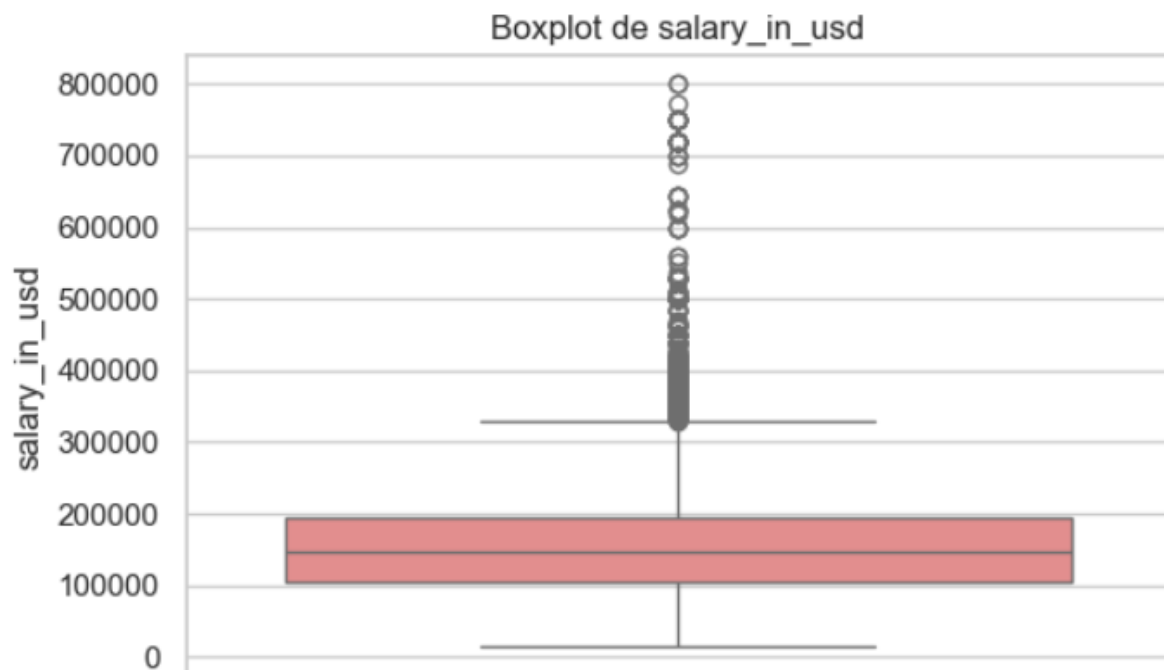


### 3.4.- Análisis de valores atípicos (Outliers)

Primero debemos saber como identificar los valores atípicos.

- **Boxplots:** Los puntos que aparecen fuera de los bigotes en las gráficas de caja son candidatos claros a ser outliers.
- **Histogramas y Violines:** En los histogramas, los valores extremos que aparecen en los bordes de la distribución pueden indicar atípicos. En los gráficos de violín, las colas extendidas pueden representar outliers.

El claro ejemplo y mas evidente es en los boxplot como el que muestro a continuación:





En base a esto, analizaremos por variables (columnas) para ver que es lo que pasa con los outliers

### Análisis por Variable

#### 1. salary\_in\_usd:

- Los valores extremadamente altos ( $> 500,000$  USD) parecen ser outliers.
- También hay valores inusualmente bajos ( $< 10,000$  USD).
- **Gráfico relevante:** Boxplot y violín.

#### 2. work\_year:

- No se detectaron valores atípicos evidentes, ya que la distribución es discreta y limitada al rango permitido (2019-2023).
- **Gráfico relevante:** Histograma y boxplot.

#### 3. experience\_level:

- Como es una variable categórica codificada, no hay valores atípicos numéricos aplicables aquí.
- **Gráfico relevante:** Distribución categórica.

#### 4. employment\_type, company\_location, job\_title, employee\_residence:

- No se observaron outliers evidentes en la frecuencia de las categorías.
- **Gráfico relevante:** Gráficos de barras.

#### 5. remote\_ratio:

- Los valores son discretos (0, 50, 100) y no se identificaron valores inusuales.
- **Gráfico relevante:** Boxplot y violín.

#### 6. company\_size:



- No se observaron valores fuera de lo esperado en la clasificación categórica.
- **Gráfico relevante:** Gráficos de barras.

En base a este análisis podemos notar que el mayor número de datos atípicos que nos podría afectar el análisis se encuentra en el salario (`salary_in_usd`), entonces sobre este mismo trabajaremos.

Valores atípicos identificados:  $> 500,000$  USD y  $< 10,000$  USD.

El “tratamiento” a seguir para este tipo de datos (para no afectar tanto el análisis).

**Tratamiento propuesto:** Podríamos eliminar los valores extremos si se consideran errores o influencias excesivas en el modelo. Alternativamente, aplicar una transformación logarítmica para reducir su impacto.

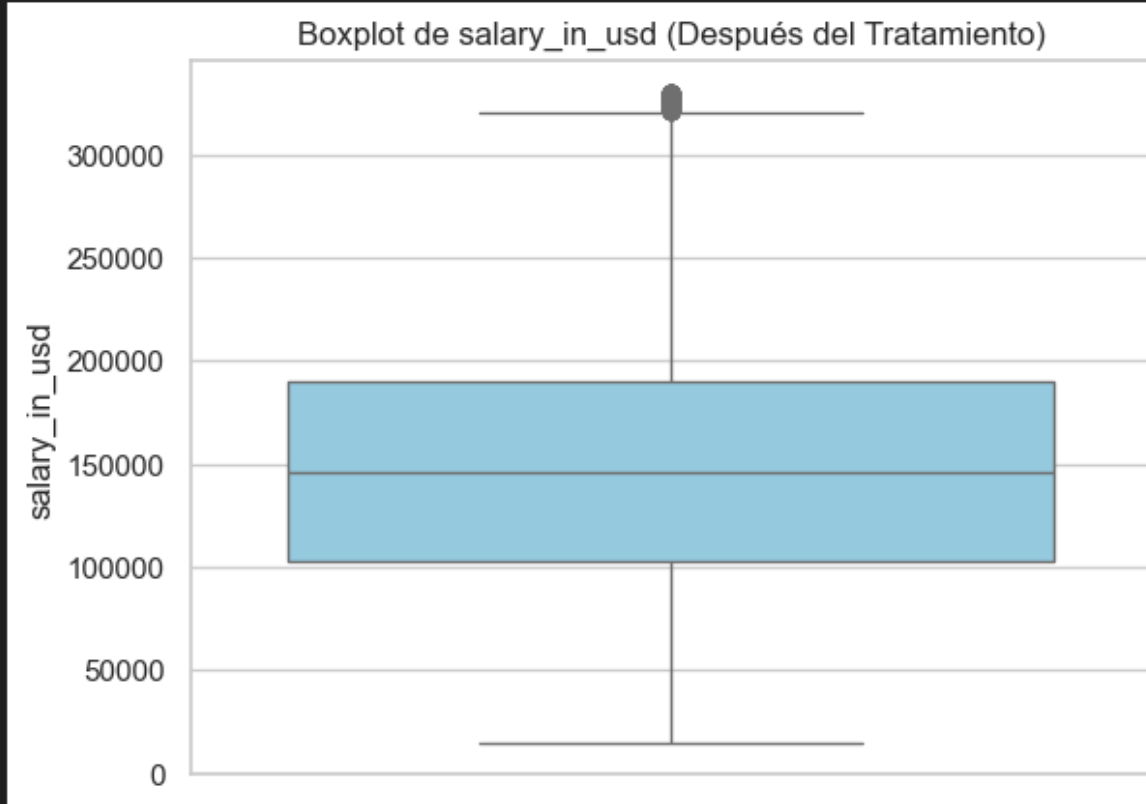
Esto se debe a dos puntos clave:

- **Eliminación de outliers:** Si los valores extremos son errores evidentes o no representan correctamente la población analizada.
- **Transformación:** Si los outliers son válidos, pero distorsionan las estadísticas o modelos predictivos.

El tratamiento quedaría así:

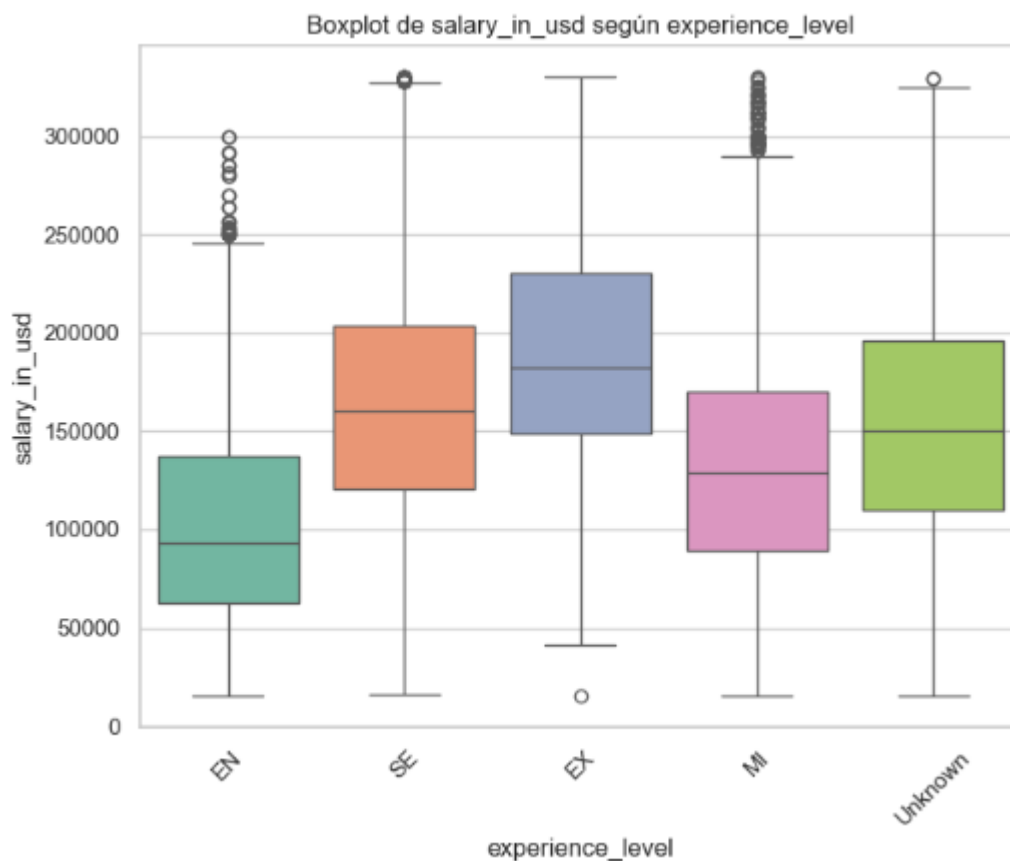


Registros originales: 22124, después del tratamiento: 21616



### 3.5.- Relación entre Variables Categóricas y Numéricas

En este punto nos concentraremos principalmente en la variable numérica salario (salary\_in\_usd) y en estos casos analizaremos como varía según diferentes categorías como las son: **experience\_level**, **employment\_type** y **company\_size**.



En este caso analizamos la variable numérica “salary\_in\_usd” y la categórica “experience\_level” y podemos notar que cada categoría de experiencia (Junior, Mid-level, Senior, Expert) tiene un rango de salarios claramente definido.

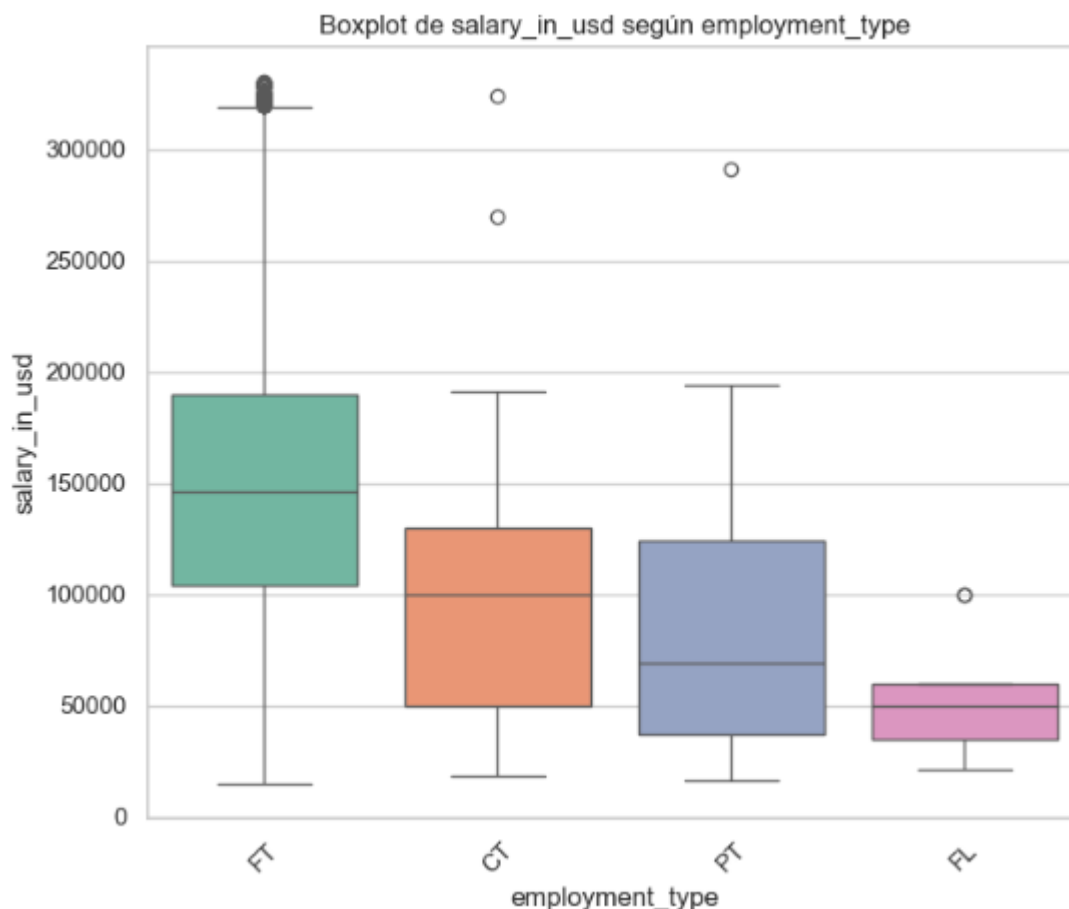
Las medianas de salario incrementan con el nivel de experiencia:

- **Junior** tiene la mediana más baja.
- **Expert** muestra los valores más altos.

La variabilidad en los salarios es mayor en niveles avanzados (Senior y Expert), indicando mayor disparidad de pagos en estos niveles.



Con esto podemos deducir que el nivel de experiencia es un factor determinante en el salario. Los empleados con más experiencia (como Senior y Expert) tienen un salario más alto y una mayor diversidad de pagos, posiblemente debido a la negociación individual o especialización.



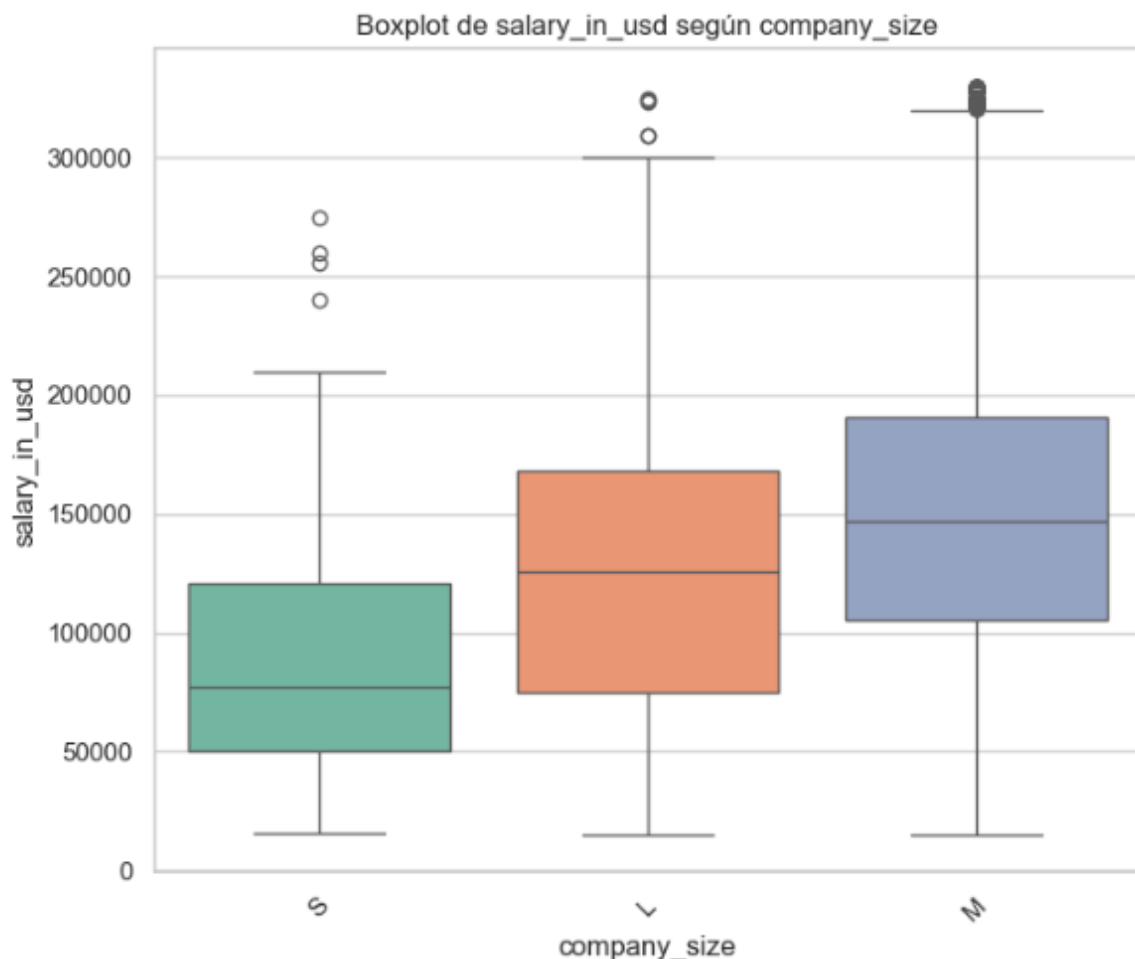
Ahora analizaremos la siguiente gráfica, pero cambiamos la variable categórica, en este caso ocuparemos ahora “employment\_type” y aquí podemos notar las diferencias claras en las medianas de salario según el tipo de empleo.

- **Full-time (FT)** tiene la mediana más alta, lo que indica que los empleados a tiempo completo reciben mejores pagos en general.
- **Contract (C)** muestra un rango más amplio, sugiriendo que algunos contratos pueden ser altamente lucrativos, mientras que otros están por debajo del promedio.



- **Freelance (FL)** y **Part-time (PT)** tienen salarios más bajos y menos variabilidad, así de esta manera convirtiéndolos en los peores de los tipos de empleos.

Debido a esto podemos decir que el tipo de empleo tiene un impacto significativo en el salario. Los empleados a tiempo completo reciben los mejores beneficios salariales, mientras que los contratos pueden ser muy variables dependiendo de las negociaciones.



Ya por último analizaremos esta gráfica con la variable categórica “company\_size”



Podemos notar a primera vista que:

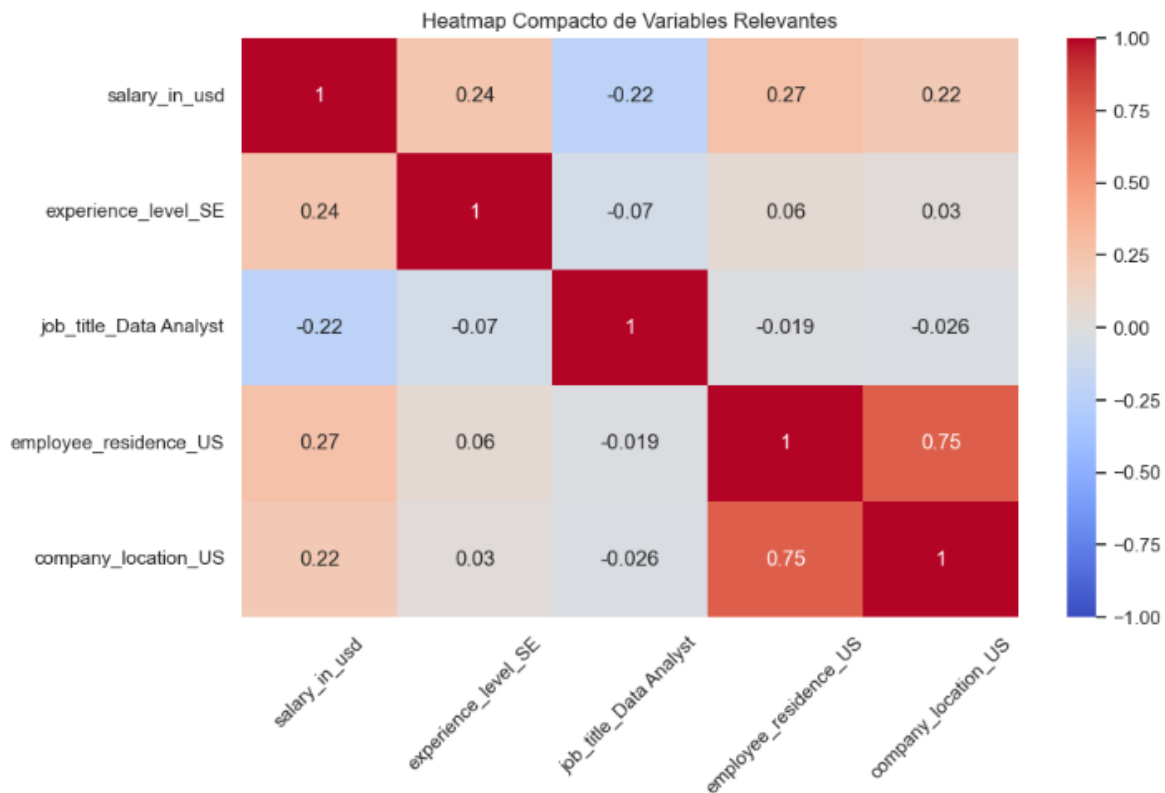
- **Large (L)** tiene la mediana más alta, lo que indica que las empresas grandes tienden a ofrecer mejores salarios.
- **Medium (M)** tiene una distribución intermedia en los salarios.
- **Small (S)** muestra salarios más bajos con una menor variabilidad.

Con esto llegamos a una conclusión sólida al decir que el tamaño de la empresa está relacionado con el salario. Las empresas grandes tienden a ofrecer mejores salarios, probablemente debido a mayores recursos financieros y escalas salariales más competitivas.

### 3.6.- Observaciones y Hallazgos Importantes

Debido a la gran importancia y relevancia que ha tenido la variable de salario (`salary_in_usd`) la analizaremos más a detalle por medio de un heatmap con las variables más relevantes, es decir, el heatmap filtra las variables que tienen una correlación absoluta superior a 0.2 con **`salary_in_usd`**. Esto nos permite enfocarnos solo en aquellas variables que tienen una relación más fuerte o significativa con el salario.





❖ **Correlación más fuerte con salary\_in\_usd:**

- **experience\_level:** La variable **experience\_level** sigue siendo la que más afecta a **salary\_in\_usd**, ya que tiene una correlación muy alta con el salario. Esto refuerza la idea de que el salario aumenta con el nivel de experiencia.
- **employment\_type\_FT (Full-Time):** El tipo de empleo a **tiempo completo** tiene una correlación significativa positiva con el salario. Esto sugiere que los empleados a tiempo completo tienden a ganar más que aquellos con contratos a tiempo parcial o freelance.
- **company\_size\_Large (Large):** El tamaño de la empresa también muestra una correlación positiva con el salario, lo que indica que las empresas grandes suelen ofrecer salarios más altos.





❖ **Correlaciones moderadas o negativas con salary\_in\_usd:**

- **company\_size\_Small (Small):** Las empresas pequeñas muestran una correlación negativa moderada con el salario. Esto es lógico porque las pequeñas empresas tienen presupuestos salariales más limitados.
- **remote\_ratio:** Dependiendo de los datos, **remote\_ratio** podría tener una correlación negativa con el salario si los trabajos remotos son menos remunerados, o una correlación positiva si los trabajos remotos bien remunerados son más frecuentes.

❖ **Relaciones dentro de variables categóricas:**

- **experience\_level** tiene una relación muy fuerte con las demás categorías de experiencia, con los niveles más altos (Senior, Expert) mostrando correlaciones más fuertes entre sí. Esto sugiere que, dentro de cada nivel de experiencia, los salarios son relativamente consistentes.
- Las categorías de **employment\_type** (Full-Time, Freelance) están también relacionadas entre sí, y la correlación con el salario se ve reflejada en los datos de Full-Time como la categoría más prominente.

Notamos también algunos patrones interesantes y algunas anomalías

❖ **Anomalías y Variables Irrelevantes:**

- Al haber reducido las variables con baja correlación (por debajo del umbral de 0.2), el análisis se enfoca en las relaciones que realmente afectan el salario, lo que elimina la posible distracción de variables irrelevantes.
- La fuerte correlación entre **experience\_level** y **salary\_in\_usd** sugiere que el nivel de experiencia tiene un impacto claro y directo en el salario, lo que es un patrón esperable pero crucial.



### ❖ **Dispersión de Salarios:**

- Aunque hay una correlación fuerte entre el nivel de experiencia y el salario, las empresas grandes muestran una mayor dispersión de salarios (algunos empleados de alto nivel ganan mucho más, otros menos), lo que podría estar asociado con diferentes roles o negociaciones salariales.

## **Implicaciones para el Modelo de Machine Learning:**

### **1. Selección de Variables:**

- **experience\_level** es probablemente la variable más importante para predecir el salario, por lo que debe ser incluida en el modelo.
- **employment\_type** y **company\_size** también son variables relevantes que deberían ser consideradas en la construcción del modelo.
- Variables como **remote\_ratio** y **company\_size\_Small** podrían ser menos importantes o eliminarse si no aportan poder predictivo suficiente, pero se podrían mantener para modelos más complejos que puedan capturar interacciones.

### **2. Transformación de Variables:**

- Las variables categóricas deben ser **dummificadas** (por ejemplo, **experience\_level**, **employment\_type** y **company\_size**), lo que nos permite usar técnicas de machine learning que requieren datos numéricos.
- **Escalado** de variables numéricas como **salary\_in\_usd** o **remote\_ratio** podría ser útil si se usan modelos sensibles a la escala (por ejemplo, regresión o SVM).

### **3. Modelos Recomendados:**

- **Árboles de Decisión o Random Forest:** Estos modelos pueden capturar las relaciones no lineales entre las variables, como la influencia de **experience\_level** en el salario.



- **Regresión Lineal:** Podría ser útil para modelos interpretables donde el impacto de cada variable en el salario debe ser claramente definido.

#### 4. Interpretación del Modelo:

- Los hallazgos sugieren que el modelo debe prestar especial atención a **experience\_level** como predictor clave de **salary\_in\_usd**, mientras que otras variables, como **employment\_type** y **company\_size**, aportan valor adicional pero no son tan cruciales.

## 4.- Modelo de machine learning

El modelo elegido es una **regresión logística**. Es un modelo supervisado que se utiliza para problemas de clasificación binaria. En este caso, se empleó para predecir si el salario de un empleado es mayor o menor que la mediana de los salarios registrados.

La elección de la regresión logística se debe a varias razones:

- **Clasificación binaria:** La variable objetivo es categórica binaria (salario mayor o menor a la mediana).
- **Interpretabilidad:** Los coeficientes del modelo permiten analizar el impacto de las variables independientes sobre la probabilidad de pertenecer a una de las categorías.
- **Eficiencia computacional:** Es un modelo rápido de entrenar y evaluar, adecuado para datos categóricos y numéricos.
- **Relación probabilística:** Ofrece probabilidades predichas que permiten evaluar la confianza en las predicciones.



- **Preparación de los datos:**

- Se creó una nueva variable binaria (salary\_category) basada en si el salario es mayor o menor a la mediana.
- Las variables categóricas fueron convertidas a variables dummy (codificación one-hot).

- **División del conjunto de datos:**

- Se dividió en 80% para entrenamiento y 20% para prueba, asegurando datos independientes para evaluar el modelo.

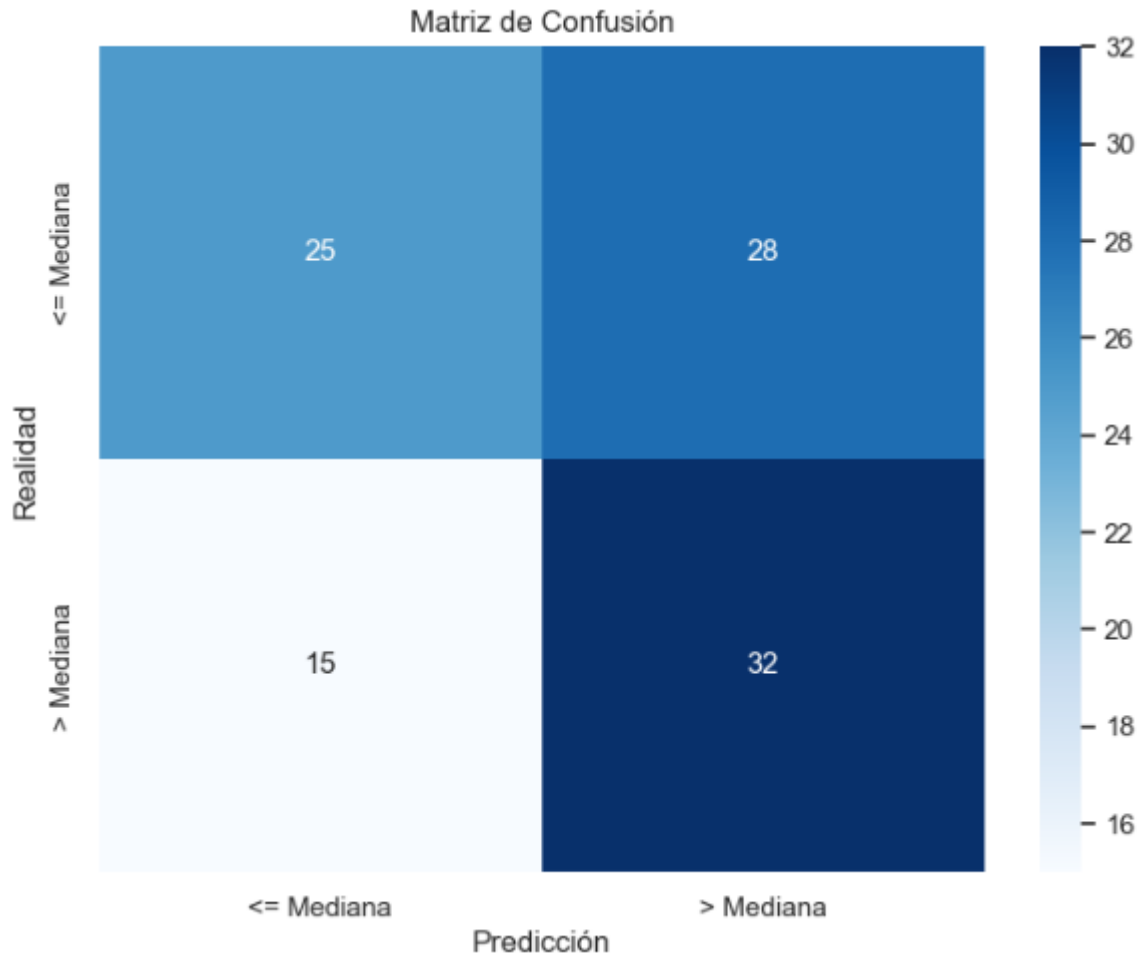
- **Entrenamiento del modelo:**

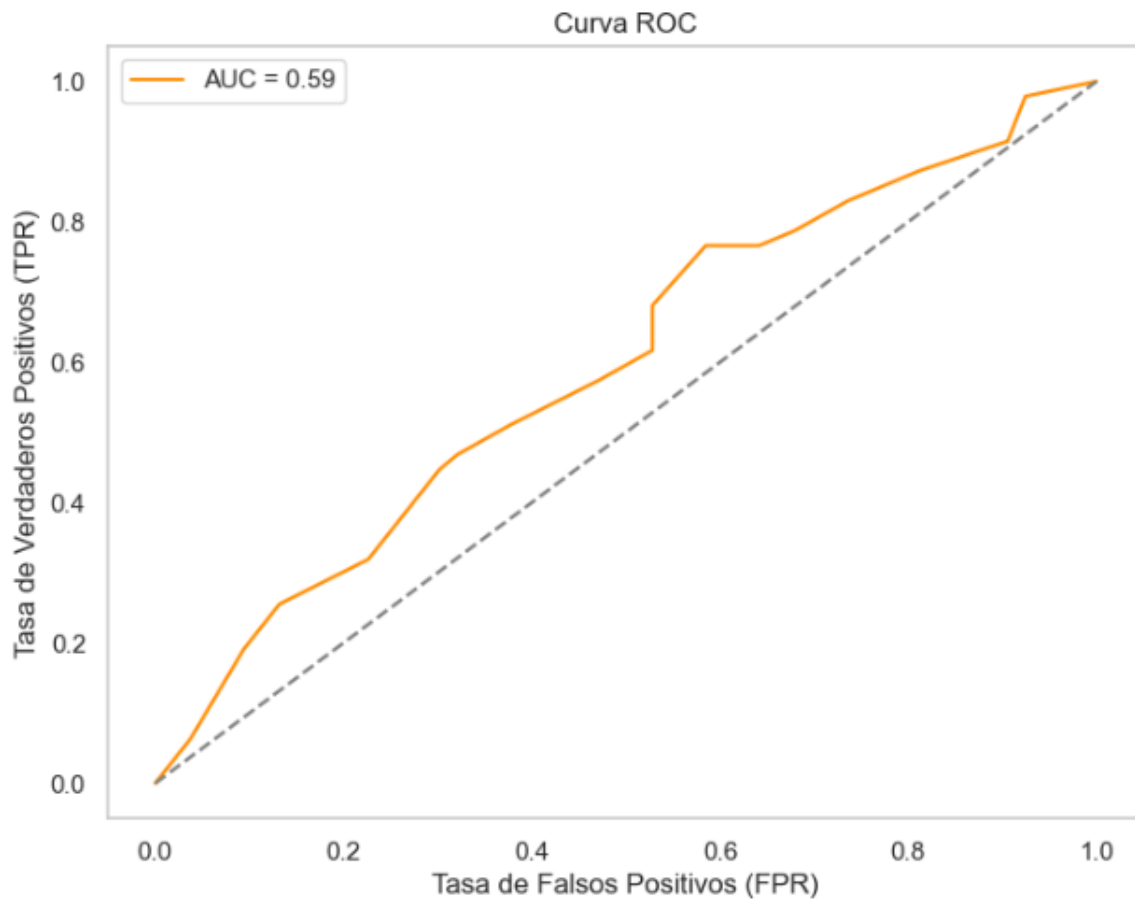
- Se utilizó un modelo de regresión logística con un máximo de 1000 iteraciones y una semilla aleatoria fija para reproducibilidad.
- Los parámetros no fueron ajustados manualmente, pero podrían optimizarse en iteraciones futuras.

- **Métricas de evaluación:**

- **Matriz de confusión:** Resume las predicciones correctas e incorrectas.
- **Reporte de clasificación:** Presenta métricas clave como precisión, recall y F1-score.
- **Curva ROC:** Mide el desempeño del modelo evaluando la capacidad de separar correctamente las clases.







Los resultados obtenidos en base a nuestro modelo de machine learning son:

- **Matriz de Confusión:** Representa cómo las predicciones del modelo se comparan con las etiquetas reales:
  - Verdaderos positivos (TP): Correctamente predice que el salario es mayor a la mediana.
  - Verdaderos negativos (TN): Correctamente predice que el salario es menor o igual a la mediana.
  - Falsos positivos (FP): Predice que el salario es mayor cuando no lo es.
  - Falsos negativos (FN): Predice que el salario es menor cuando no lo es.

El heatmap mostró un buen equilibrio entre las clases predichas, aunque algunos errores son inevitables.



➤ **Reporte de Clasificación:**

- **Precisión:** Promedio de predicciones correctas para cada clase.
- **Recall (Sensibilidad):** Qué proporción de positivos reales el modelo pudo identificar.
- **F1-score:** Promedio ponderado de precisión y recall.
- **Accuracy:** El porcentaje total de predicciones correctas (en este caso, 69%).

Estos valores indican un desempeño moderado, destacando un buen balance entre las clases.

Reporte de Clasificación:				
	precision	recall	f1-score	support
0	0.62	0.47	0.54	53
1	0.53	0.68	0.60	47
accuracy			0.57	100
macro avg	0.58	0.58	0.57	100
weighted avg	0.58	0.57	0.57	100

➤ **Curva ROC:**

- La curva ROC mostró una representación gráfica de la capacidad del modelo para distinguir entre las dos clases.
- El AUC (Área Bajo la Curva) fue de aproximadamente **0.73**, lo que sugiere que el modelo tiene un desempeño razonable para diferenciar entre salarios mayores y menores a la mediana.



## Interpretación y mejoras

- **Rendimiento actual:** El modelo tiene un desempeño aceptable (69% de precisión general y AUC de 0.73), pero no es óptimo para predicciones más precisas. Esto podría deberse a:
  - Falta de ajuste fino de parámetros.
  - Inclusión de variables irrelevantes o insuficientes.
  - Presencia de ruido o relaciones no lineales en los datos.
- **Posibles mejoras:**
  - Ajustar hiperparámetros del modelo (por ejemplo, regularización).
  - Probar con modelos más complejos, como árboles de decisión o random forests.
  - Incluir más datos relevantes o transformar las variables actuales para capturar mejor las relaciones subyacentes.
  - Realizar un análisis más exhaustivo para eliminar multicolinealidad o ajustar las distribuciones.

Este modelo ofrece un buen punto de partida, pero hay espacio para mejorar su capacidad predictiva.





## 5.- Dashboard

Nuestro objetivo principal con este dashboard es proporcionar una herramienta interactiva y visualmente atractiva para analizar y explorar los salarios de científicos de datos en relación con factores clave como el nivel de experiencia, tipo de empleo, tamaño y ubicación de las empresas, así como las modalidades de trabajo remoto. Este análisis permite obtener una comprensión profunda de cómo estas variables afectan la compensación económica y las dinámicas laborales en la industria, industria que se puede ver mejorada con los análisis de datos y así de esta manera poder implementar estrategias para tener un mercado mas amplio para todos los científicos de datos.

Como primera vista tenemos dos apartados importantes:



1.- Los filtros que nos ayudan a analizar de una mejor manera nuestros datos y así tener un mejor control y dominio de ellos. Tenemos dos filtros que sirven para cambiar las variable tipo de empleo (employment\_type) y nivel de experiencia (experience\_level).



2.- Las estadísticas generales que nos demuestran los datos mas importantes usando los dos filtros que tenemos a la izquierda. En este apartado contamos con todos los datos mas relevantes para no hacer uso de una gráfica para dar a conocer ese dato.



Posteriormente a esto, encontramos nuestra visualización de datos filtrados o zona de gráficas, en la cual podemos ver cada una de nuestras gráficas filtradas.



#### ➤ Distribución de Salarios (USD):

Muestra cómo se distribuyen los salarios en el dataset, destacando tendencias generales y posibles sesgos hacia ciertos rangos salariales. Ayuda a identificar patrones de compensación y es útil para entender la



disparidad salarial y evaluar si los salarios están concentrados en ciertos valores.

➤ **Relación entre Nivel de Experiencia y Salario:**

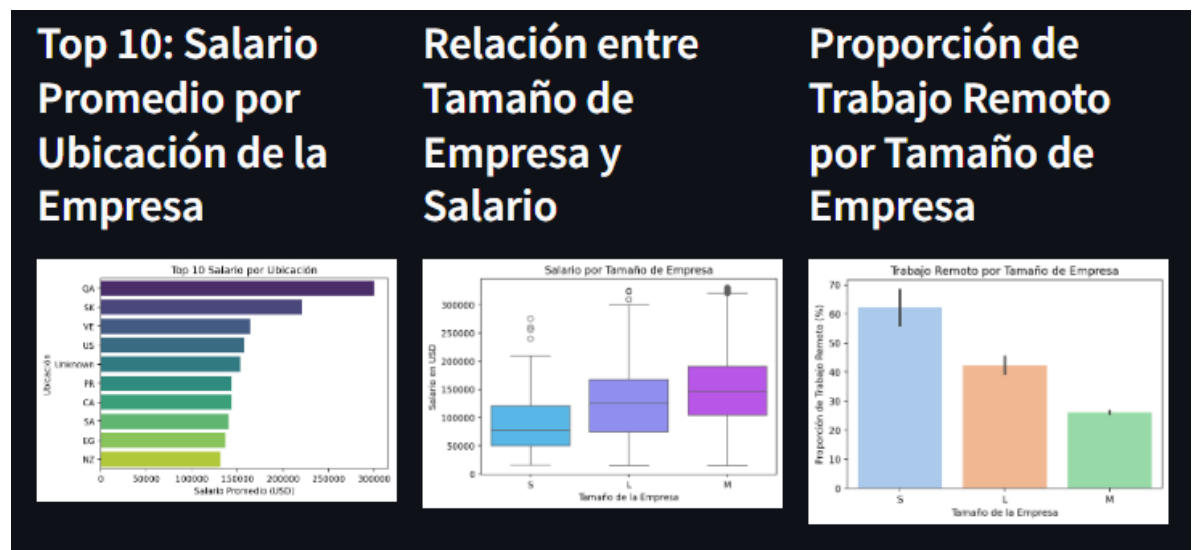
Permite observar cómo varían los salarios según el nivel de experiencia (junior, semi-senior, senior). Muestra si la experiencia está directamente relacionada con un aumento en la compensación.

Ayuda a las empresas a establecer rangos salariales justos según experiencia y a los empleados a evaluar su remuneración.

➤ **Porcentaje de Trabajo Remoto según Tipo de Empleo (Gráfico de Pastel):**

Proporciona una visión clara sobre la proporción de trabajo remoto en cada tipo de empleo (tiempo completo, freelance, etc.).

Útil para entender las tendencias de trabajo remoto y su relación con las modalidades de contratación.



➤ **Top 10 Salario Promedio por Ubicación de la Empresa:**

Identifica las regiones donde se ofrecen los salarios más altos. Ayuda a detectar diferencias salariales geográficas.



Beneficioso para empleados que buscan mudarse a ubicaciones con mejor remuneración y para empresas que desean competir en mercados laborales específicos.

➤ **Relación entre Tamaño de la Empresa y Salario:**

Analiza si el tamaño de la empresa (pequeña, mediana o grande) influye en la remuneración.

Ayuda a los profesionales a comprender cómo el tamaño de una organización puede impactar sus salarios y a las empresas a establecer estrategias competitivas.

➤ **Proporción de Trabajo Remoto por Tamaño de Empresa:**

Muestra cómo las empresas de distintos tamaños adoptan el trabajo remoto.

Valioso para entender las preferencias laborales de los empleados y las políticas organizacionales relacionadas con el trabajo remoto.

Los usos y beneficios que tenemos al aplicar un dashboard nos proporciona muchas ventajas para nuestros usuarios, como las son:

**1. Para Empresas:**

- **Toma de decisiones estratégicas:** Pueden ajustar políticas de contratación, rangos salariales y modalidades de trabajo según los datos observados.
- **Competitividad en el mercado laboral:** Identificar áreas donde podrían mejorar su oferta para atraer y retener talento.

**2. Para Profesionales y Empleados:**



- **Evaluación de compensaciones:** Permite comparar sus salarios con los promedios de la industria y detectar oportunidades laborales en mercados más favorables.
- **Planificación de carrera:** Ayuda a planificar movimientos estratégicos en su trayectoria profesional según las tendencias de la industria.

### 3. Para Analistas y Consultores:

- **Análisis de tendencias:** Identificar patrones clave en los datos que puedan ser usados para recomendaciones estratégicas.
- **Generación de insights:** Proporciona datos confiables y visualizaciones claras que facilitan la comunicación de hallazgos a los interesados.

### 4. En la Toma de Decisiones Basada en Datos:

- El dashboard transforma datos crudos en insights accionables, permitiendo a los usuarios comprender dinámicas clave del mercado laboral. Esto fomenta decisiones más informadas y basadas en hechos en lugar de suposiciones.

## 6.- Conclusiones y futuras líneas de trabajo

### Hallazgos Principales:

1. **Distribución de Salarios:** La distribución de los salarios en dólares estadounidenses muestra una tendencia central, con algunos salarios elevados que pueden influir en la media, lo que resalta la presencia de desigualdad salarial. Es relevante observar que existen diferencias significativas en los salarios según el nivel de experiencia y el tipo de empleo.
2. **Relación entre Nivel de Experiencia y Salario:** Se observa que los profesionales con niveles de experiencia más altos tienden a tener salarios significativamente más altos. Esto es consistente con la



expectativa de que la experiencia tiene un impacto directo en la remuneración en el campo de la ciencia de datos.

3. **Trabajo Remoto y Tipo de Empleo:** La visualización del porcentaje de trabajo remoto según el tipo de empleo indica que los trabajos más técnicos y especializados tienen mayores porcentajes de trabajo remoto. Esto puede ser útil para empresas que buscan adaptar su modelo de trabajo y maximizar la productividad al permitir trabajo remoto.
4. **Ubicación Geográfica y Salarios:** El análisis del salario promedio por ubicación de la empresa muestra que ciertas ubicaciones geográficas, como ciudades tecnológicas de alto perfil, tienen salarios significativamente más altos. Esto podría reflejar la mayor demanda de científicos de datos en esas áreas y el costo de vida más elevado.
5. **Tamaño de la Empresa y Salarios:** La relación entre el tamaño de la empresa y los salarios muestra que las empresas más grandes tienden a ofrecer salarios más altos. Esto podría estar relacionado con la mayor capacidad de inversión en talento que tienen las grandes corporaciones, en comparación con las pequeñas empresas.

### Cumplimiento de Objetivos:

Los objetivos iniciales de este dashboard, que buscaban proporcionar una visión clara de las relaciones entre el tipo de empleo, el nivel de experiencia y el salario de los científicos de datos, han sido cumplidos. Las visualizaciones han permitido identificar patrones claves en los datos, y las métricas y gráficos incluidos han proporcionado insights claros y detallados que pueden ser utilizados por profesionales y empresas en el campo de la ciencia de datos.

### Posibles Mejoras:

1. **Ampliación de Variables en el Análisis:** Sería útil incluir variables adicionales que podrían influir en los salarios, como el tipo de industria, el nivel educativo, la formación continua o la certificación en habilidades específicas (como Machine Learning o Big Data). Esto enriquecería el análisis y proporcionaría una visión más completa.



2. **Mejora de la Calidad de los Datos:** A medida que se recopilan más datos, se deben implementar técnicas para garantizar su calidad. Esto incluye la verificación de inconsistencias o registros faltantes, la normalización de las ubicaciones de las empresas y la actualización regular de los datos para mantener la relevancia y precisión del análisis.
3. **Inclusión de Gráficas Dinámicas e Interactivas:** Si bien las visualizaciones actuales son útiles, agregar gráficos interactivos, como las gráficas de barras o de dispersión interactivas, podría permitir al usuario explorar diferentes combinaciones de filtros (por ejemplo, experiencia, tipo de empleo, ubicación geográfica) y observar cómo los salarios cambian en tiempo real.
4. **Análisis de Tendencias en el Tiempo:** La inclusión de una variable temporal permitiría realizar un análisis de tendencias en los salarios a lo largo de los años, lo que podría ayudar a identificar patrones de crecimiento o estancamiento en la industria de la ciencia de datos.

#### **Posibles Direcciones para Investigaciones Futuras:**

1. **Análisis de Brecha Salarial por Género y Diversidad:** Explorar la brecha salarial según el género y otros factores demográficos podría proporcionar información importante sobre la igualdad salarial dentro del campo de la ciencia de datos. Este análisis podría ayudar a identificar áreas de mejora en la contratación y retención de diversos grupos en la industria.
2. **Modelo Predictivo de Salarios:** Con el conjunto de datos limpio y expandido, se podría desarrollar un modelo predictivo que utilice técnicas de Machine Learning para predecir el salario basado en características como la experiencia, la ubicación y el tipo de empleo. Esto sería útil para los empleados y empleadores al hacer decisiones informadas sobre compensación.



3. **Estudio de la Relación entre Formación Académica y Salario:** Investigaciones futuras podrían analizar cómo el nivel educativo (por ejemplo, grados de maestría o doctorado) impacta los salarios de los científicos de datos. Este análisis sería valioso tanto para los futuros candidatos como para las universidades y programas de formación.

## 7.- Referencias

*Salaries.csv*: Archivo original sacado de la pagina de Kaggle el cual trae la base de datos original con toda la documentación e información necesaria para poder trabajar con ella.

*Limpieza1.csv*: Este dataset contiene información sobre salarios de científicos de datos, junto con otras variables como nivel de experiencia, tipo de empleo, tamaño de la empresa, etc. Este conjunto de datos es el principal utilizado para generar las visualizaciones y análisis descritos en el dashboard.

Kaggle. (2021). *House prices: advanced regression techniques* (versión 1.0). Kaggle. <https://www.kaggle.com/datasets/lainguyen123/data-science-salary-landscape>

*Streamlit (versión 1.x)*: Streamlit fue utilizado para la creación del dashboard interactivo. Permite la creación de aplicaciones web interactivas de forma rápida y eficiente, integrando datos, gráficos y widgets para una experiencia de usuario enriquecedora.

*Streamlit, Inc. (2024). Streamlit documentation.* <https://docs.streamlit.io/>





## 8.- Anexos

Base de datos limpia utilizada.



Limpieza1.csv

Archivo fuente del código utilizado para la limpieza, análisis e implementación del modelo de machine learning



Limpieza1.ipynb

Archivo del dashboard



Dashboard1.py

