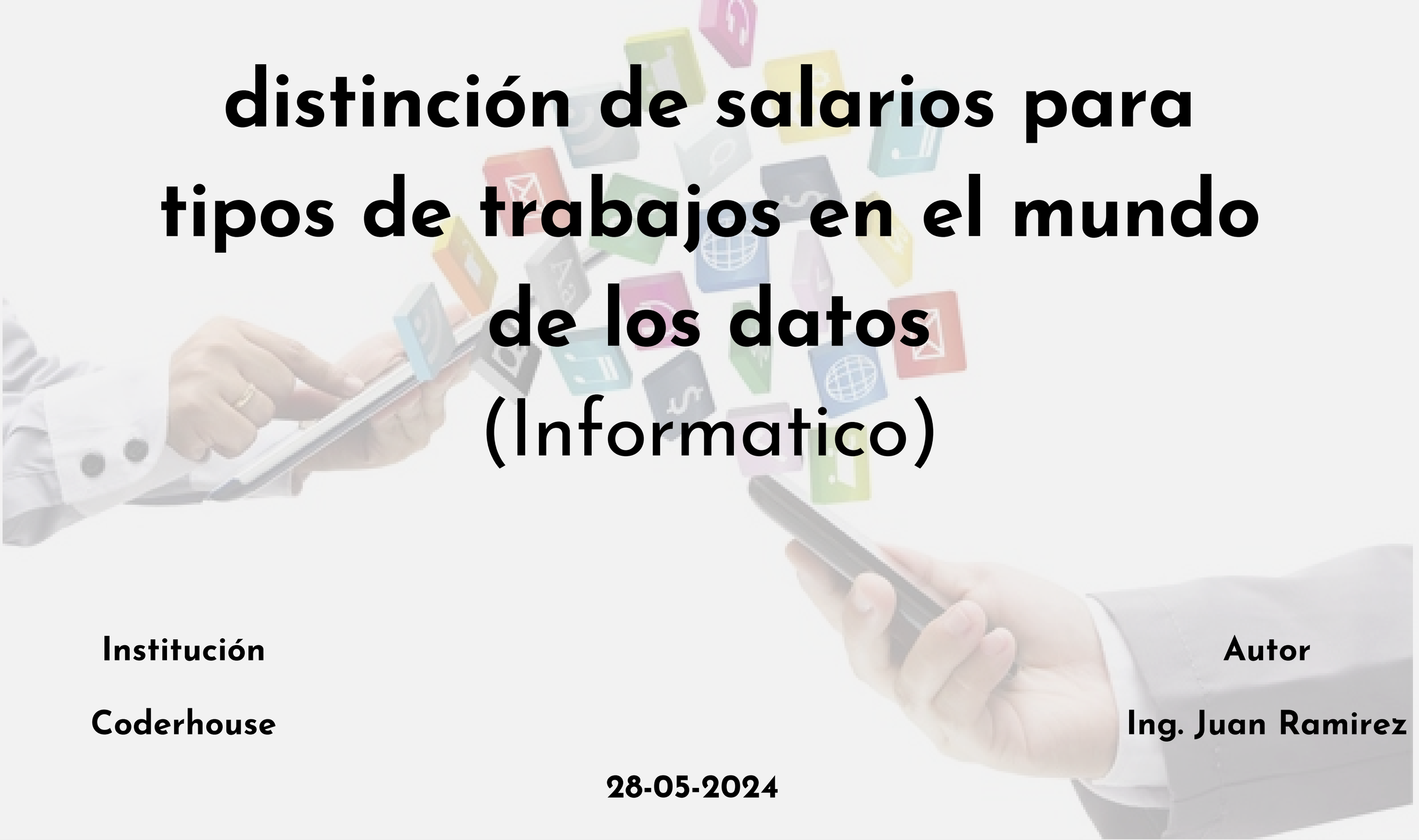


# distinción de salarios para tipos de trabajos en el mundo de los datos (Informatico)


**Institución**

**Coderhouse**

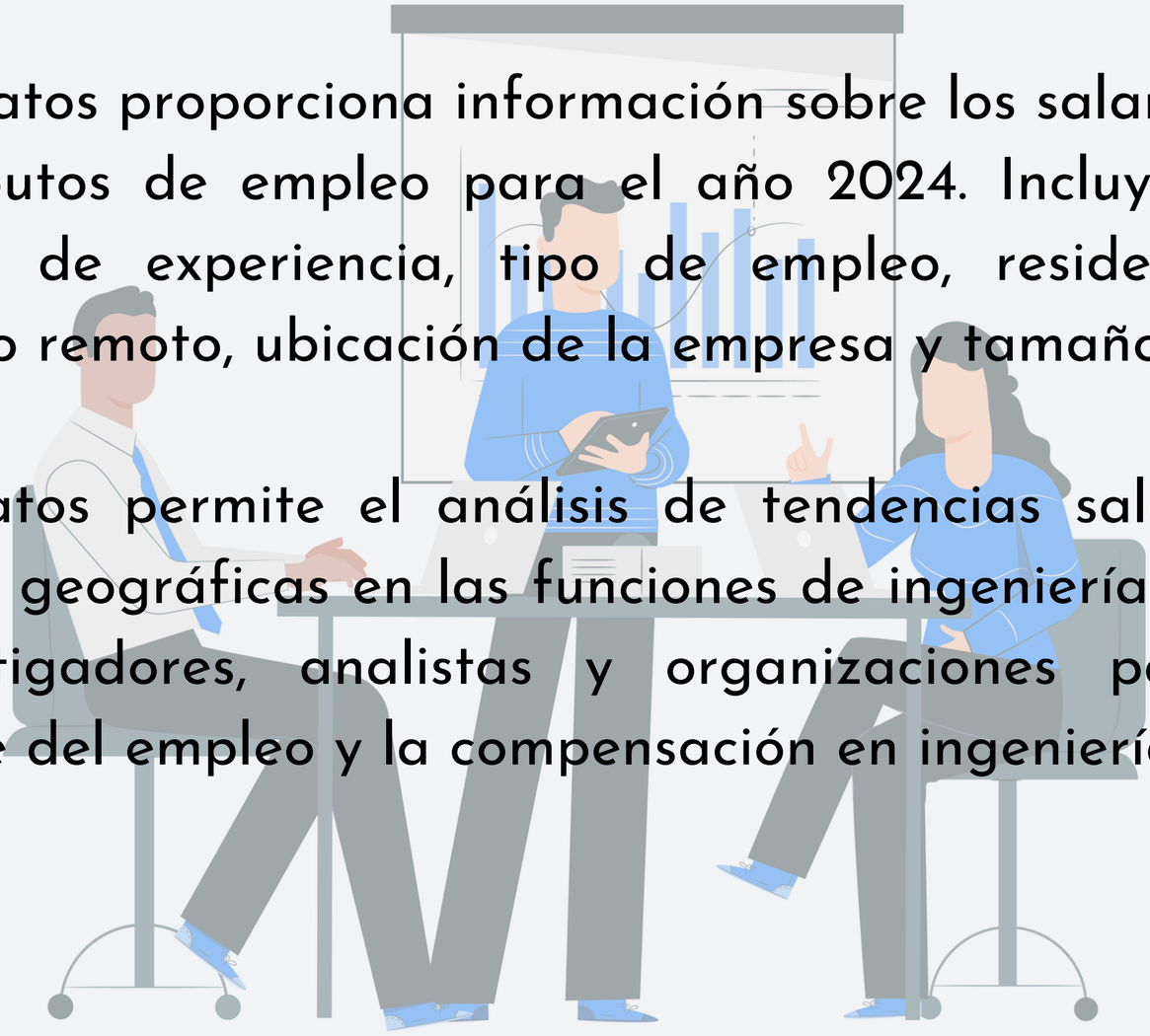
**Autor**

**Ing. Juan Ramirez**


**28-05-2024**

# Indice

# 1. Introducción



Este conjunto de datos proporciona información sobre los salarios de los ingenieros de datos y los atributos de empleo para el año 2024. Incluye información como salario, cargo, nivel de experiencia, tipo de empleo, residencia del empleado, proporción de trabajo remoto, ubicación de la empresa y tamaño de la empresa.

El conjunto de datos permite el análisis de tendencias salariales, patrones de empleo y variaciones geográficas en las funciones de ingeniería de datos. Puede ser utilizado por investigadores, analistas y organizaciones para comprender el panorama cambiante del empleo y la compensación en ingeniería de datos.

# 2. Situación

En 2024, la demanda de ingenieros de datos ha crecido notablemente, con variaciones salariales significativas según el nivel de experiencia. Los profesionales con experiencia avanzada y especializados en big data son los más solicitados y mejor remunerados. Las empresas tecnológicas en áreas urbanas ofrecen los salarios más altos, especialmente para roles senior. Además, muchas compañías están adoptando modalidades de trabajo remoto, lo que amplía las oportunidades para ingenieros de datos con distintos niveles de experiencia. Las grandes corporaciones suelen ofrecer paquetes más competitivos que las pequeñas y medianas empresas. En resumen, el mercado laboral favorece a los ingenieros de datos experimentados, reflejando su valor estratégico en la economía digital.

# 3. Objetivos y Alcances

**Realizar la limpieza correspondiente de los datos**, para poder aclarar la realidad actual de los salarios impuestos a los ingenieros de datos en rasgos generales.

**Identificar Patrones de Comportamiento** o tendencias en los salarios calculados en USD, para los distintos cargos indicados con diferentes niveles de experienca.

**Crear un modelo de Machine Learning** que logre identificar los distintos tipos de salarios, que deben asignarse a los cargos, dependiendo de su nivel de experiencia. Como primer punto se partira con un modelo basado meramente en el cargo de **Data Scientist**

# 4. Hipótesis

**Los salarios van acorde al nivel de experiencia del tipo de trabajo.**

Los empleados con mayor experiencia (por ejemplo, nivel "EX" - Experto) tienden a ganar salarios más altos en comparación con aquellos con menos experiencia como “SE” o “MI” (senior o midlevel).

**Los salarios definidos en niveles de experiencia de los distintos tipos de trabajo son muy similares entre si.**

Independientemente de ser  'Data Analyst', 'Machine Learning', 'Data Engineer' o 'Data Scientist' sus salarios no varian (son simlares entre si)

**Utilizando los datos historicos y bajo un modelo supervisado, podemos generar un modelo que prediga con precision los salarios que deden ser asignados a los distintos tipos de cargos**

# 6. Variables del Data set

| | | |
|---|---|---|
| | | |
| | | |
| | | |

**work_year:** El año en que se recogieron los datos (2024).

**experience_level:** El nivel de experiencia del empleado, categorizado como EX (ingeniero experto), SE (ingeniero sénior), MI (ingeniero de nivel medio) o EL (ingeniero de nivel inicial).

**employment_type:** El tipo de empleo, como tiempo completo (FT), tiempo parcial (PT), contrato (C) o autónomo (F).

**job_title:** El título o función del empleado dentro de la empresa, por ejemplo, ingeniero de inteligencia artificial.

**salary:** El salario del empleado en la moneda local (por ejemplo, 202 730 USD).

**salary_currency:** La moneda en la que está denominado el salario (por ejemplo, USD).

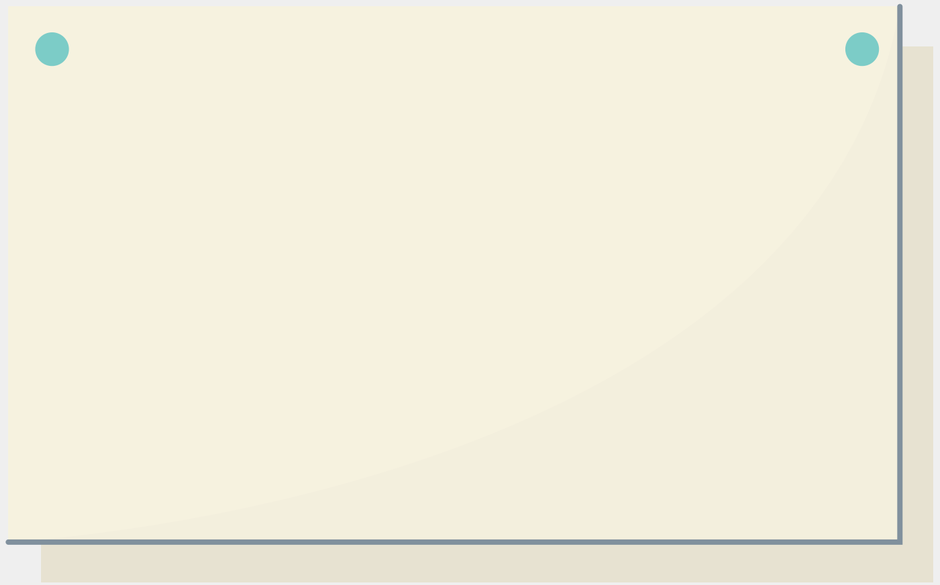
**salary_in_usd:**El salario convertido a dólares estadounidenses para fines de estandarización.

**employee_residence:** El país de residencia del empleado.

**remote_ratio:** La proporción que indica el alcance del trabajo remoto permitido en el puesto (0 para ningún trabajo remoto, 1 para completamente remoto).

**company_location:** La ubicación de la empresa donde trabaja el empleado.

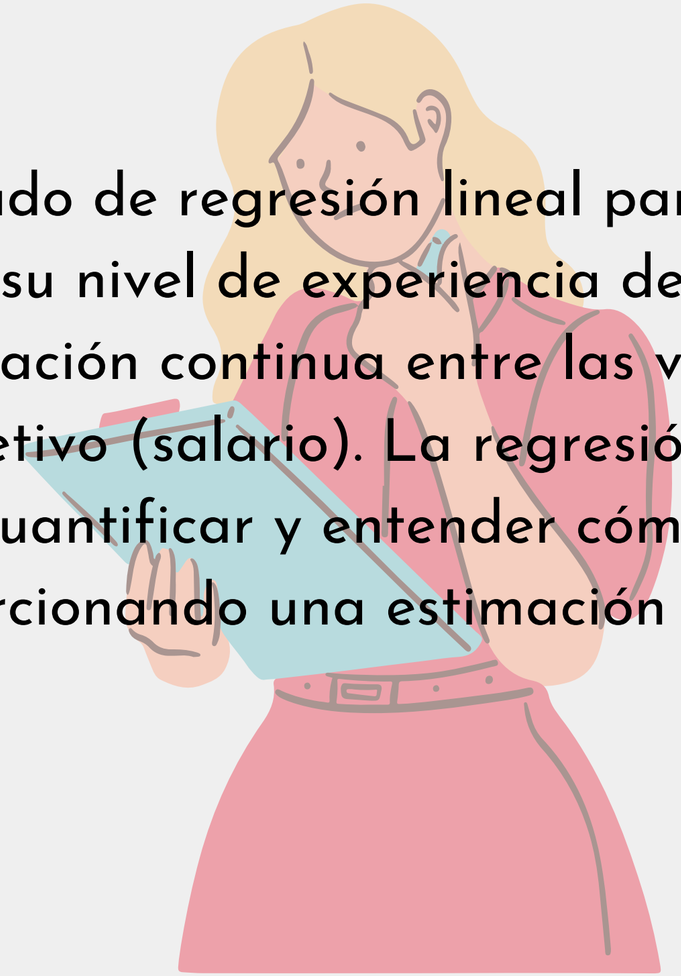
**company_size:**El tamaño de la empresa, a menudo categorizado por el número de empleados (S para pequeña, M para mediana, L para grande).



# 7. Tipo de aprendizaje utilizado

Se utilizó un modelo supervisados ya que son eficaces cuando se cuenta con un conjunto de datos etiquetado, permitiendo al modelo aprender de las observaciones conocidas para hacer predicciones precisas sobre nuevos datos. Esto es crucial en contextos salariales, donde se busca entender y prever tendencias económicas y de compensación basadas en la experiencia laboral.

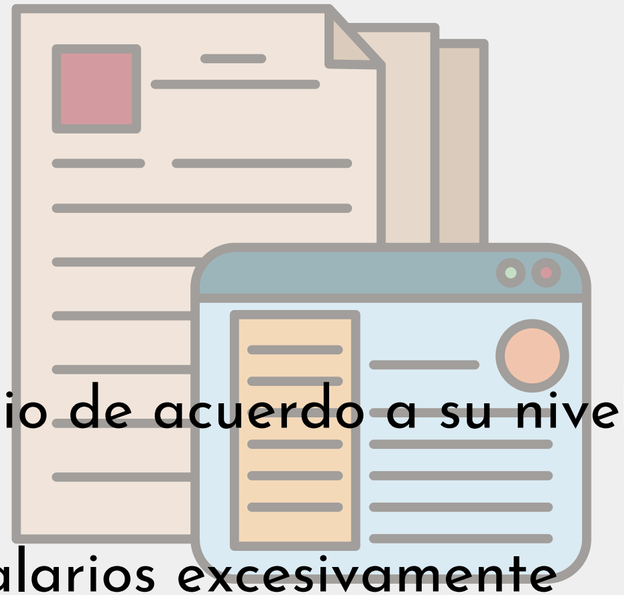
# 8. Modelo de resolución



Se utilizó un modelo supervisado de regresión lineal para predecir los salarios de los Data Scientists en función de su nivel de experiencia debido a la naturaleza del problema, que implica una relación continua entre las variables predictoras (nivel de experiencia) y la variable objetivo (salario). La regresión lineal es adecuada para este análisis porque permite cuantificar y entender cómo varía el salario con respecto al nivel de experiencia, proporcionando una estimación directa de esta relación.

# 9.Resumen del Proceso de Análisis de Salarios y modelo de ML para Data Scientists

1. Importación y Preparación de Datos:
   - Se importaron los datos desde un archivo CSV que incluía información sobre nivel de experiencia, tipo de empleo, título del trabajo, salario, residencia del empleado, proporción de trabajo remoto, ubicación de la empresa y tamaño de la empresa.
   - Se verificaron existencia de datos nulos o perjudiciales para el set de datos a utilizar
   - Se convirtieron las categorías del nivel de experiencia ('EN', 'EX', 'MI', 'SE') a valores numéricos ('1', '4', '2', '3') para facilitar el análisis.
   - Se agruparon algunos títulos de trabajo similares bajo una misma categoría, por ejemplo, 'Data Science' y 'Data Scientist'

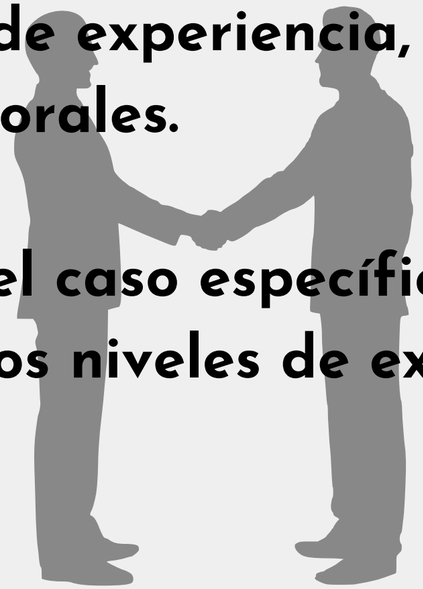


2. **Análisis Exploratorio de Datos:**
   - Se visualizaron los tipos de trabajos y sus salarios promedio de acuerdo a su nivel de experiancia.
   - Se identificaron y eliminaron outliers, como por ejemplo salarios excesivamente altos para 'Data Analyst'.
3. **Correlación:**
   - Se calculó el coeficiente de correlación entre el nivel de experiencia y el salario para los Data Scientists, resultando en un valor de 0.3947, indicando una relación positiva moderada.

# 10. Conclusión

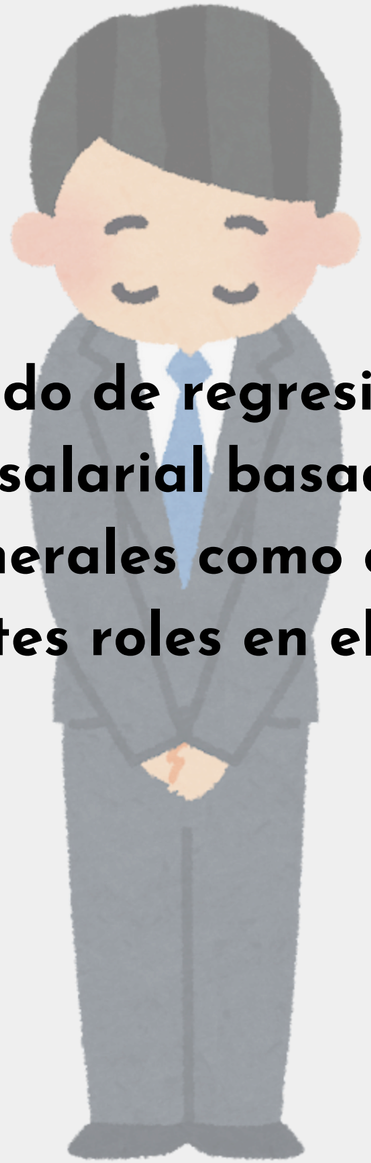
En el análisis realizado, se encontró que el uso de un modelo supervisado de regresión lineal fue viable para predecir los salarios de los Data Scientists en función de su nivel de experiencia. El modelo mostró una relación positiva moderada entre el nivel de experiencia y el salario, reflejada en un coeficiente de correlación de 0.3947. Esta relación indica que, en general, los salarios tienden a aumentar con el nivel de experiencia, lo cual es un resultado esperado y coherente con las prácticas laborales.

Sin embargo, se observó que en el caso específico de los 'Data Analysts', los salarios no están alineados con los niveles de experiencia de manera consistente.



Específicamente, se encontró que los 'Data Analysts' con nivel de experiencia senior ganan solo un poco más en promedio que aquellos con nivel de experiencia experto, lo cual es una anomalía en comparación con otras ocupaciones. Además, se identificó que los salarios para 'Data Analysts' están significativamente por debajo de los de otros cargos de trabajo, independientemente del nivel de experiencia.

En contraste, los salarios definidos para niveles de experiencia en los distintos tipos de trabajos analizados (como Data Engineer, Machine Learning Engineer, y Data Scientist) son muy similares entre sí y siguen una tendencia esperada de aumento con la experiencia. Este hallazgo resalta una consistencia salarial en la mayoría de los roles, exceptuando el caso de los 'Data Analysts'.



En resumen, el modelo supervisado de regresión lineal ha demostrado ser una herramienta útil para el análisis salarial basado en la experiencia, permitiendo identificar tanto tendencias generales como excepciones significativas en los patrones salariales de diferentes roles en el campo de la ciencia de datos.