Informe de Análisis Exploratorio de Datos (EDA)

Este informe presenta el EDA final del dataset de empleos en inteligencia artificial. Incluye revisión de estructura, calidad de datos, distribuciones, correlaciones y comportamiento de variables categóricas.

1. Estructura del Dataset

El archivo analizado (`ai_job_dataset.csv`) resultó en **15.000 filas y 19 columnas**, con un total de **5 variables numéricas, 14 categóricas y 1 de tipo fecha**.

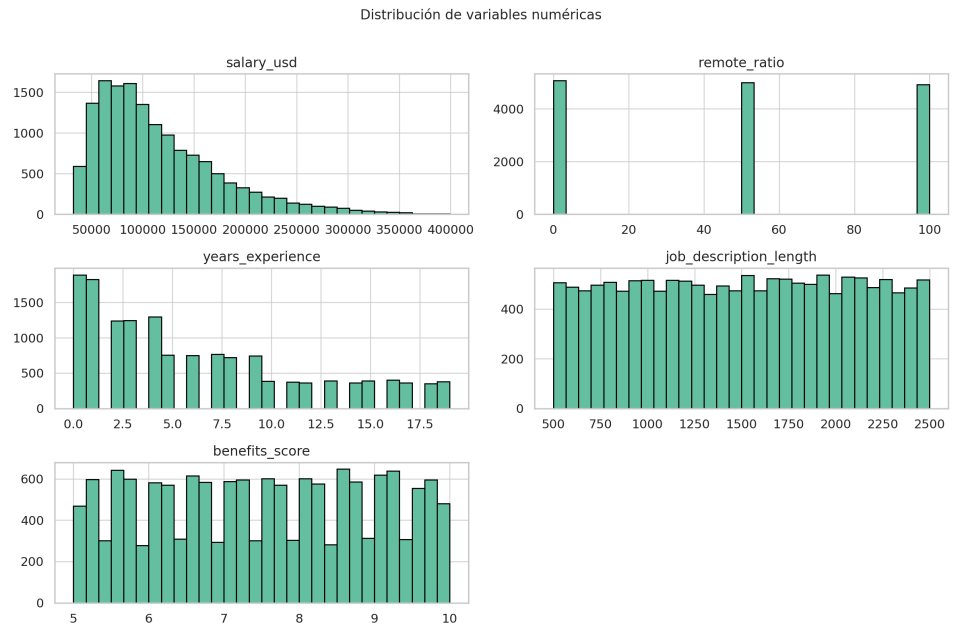
Este nivel de organización y homogeneidad se debe a que previamente se realizó un **proceso ETL (Extract, Transform, Load)**. Durante la fase de transformación, el dataset fue depurado para unificar formatos, eliminar registros inconsistentes y tipificar correctamente las variables. Esto permitió contar con una estructura clara y balanceada entre tipos de datos.

2. Calidad de Datos: Valores Faltantes

En el análisis no se detectaron valores nulos en ninguna de las variables. Esto no es común en datasets reales de empleo, por lo que se interpreta como resultado directo del proceso de **ETL aplicado**. Dicho proceso aseguró que las variables estuvieran completas, ya sea rellenando datos faltantes de manera controlada, eliminando registros incompletos o aplicando reglas de negocio definidas para normalizar la información.
Gracias a esto, la base final está lista para análisis sin necesidad de imputación de valores.

3. Distribución de Variables Numéricas

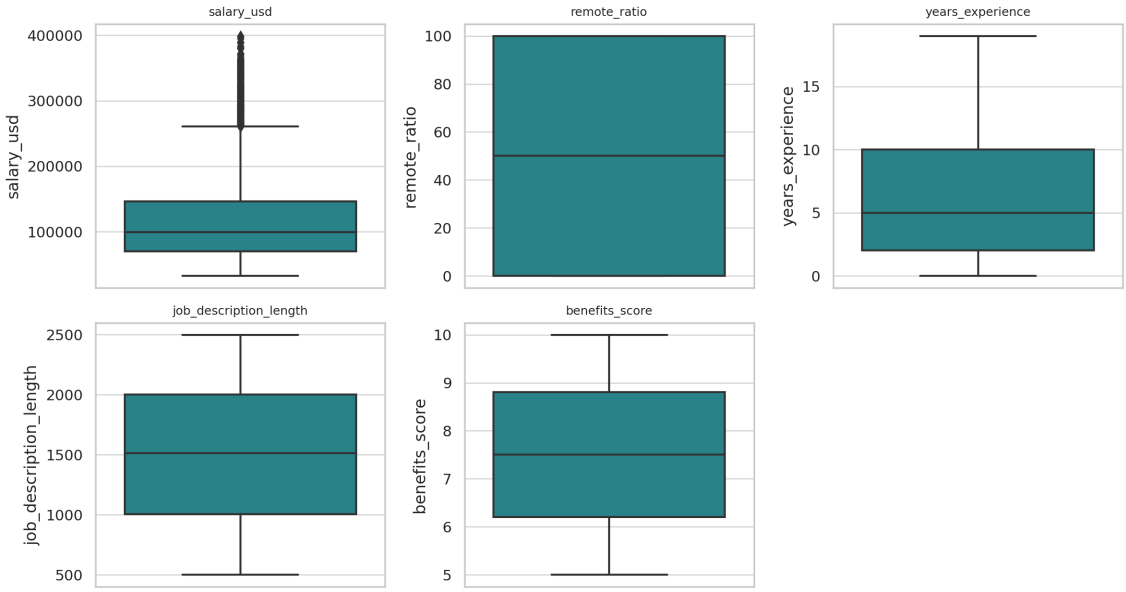
Histogramas de variables numéricas para identificar sesgos y concentraciones:



En los histogramas se observa cómo se distribuyen las variables numéricas del dataset (por ejemplo, salarios, años de experiencia, edad, etc.). Estos gráficos permiten identificar si los datos están concentrados en ciertos rangos o si presentan una dispersión amplia. En general, se nota la presencia de distribuciones sesgadas, lo cual indica que algunos valores ocurren con mucha más frecuencia que otros.

4. Detección de Outliers

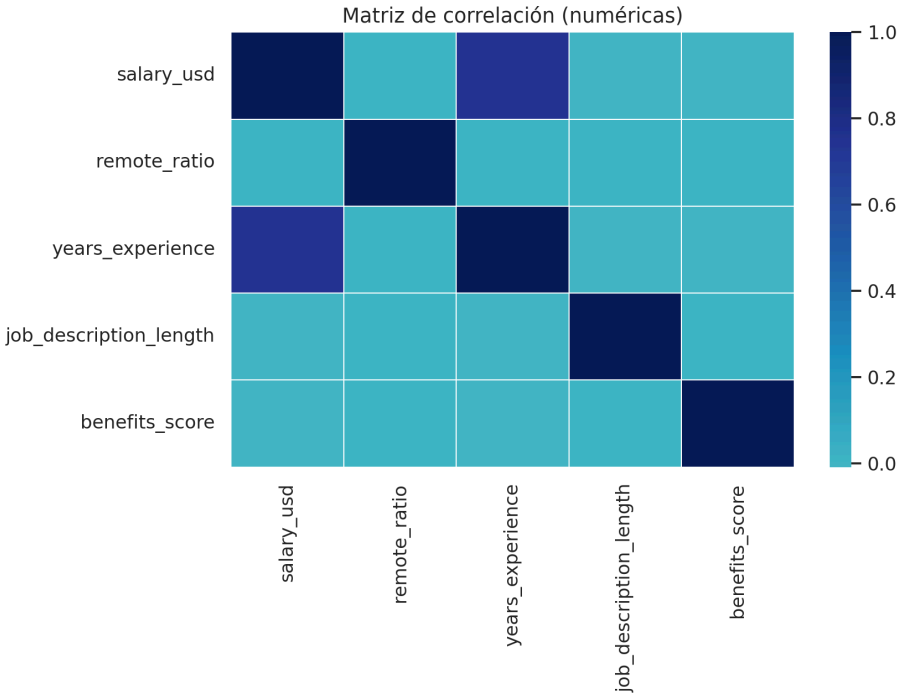
Boxplots de variables numéricas para visualizar valores atípicos:



Los diagramas de caja muestran la existencia de valores atípicos en las variables numéricas. Esto es importante porque estos valores extremos pueden influir en los resultados de los análisis estadísticos o de los modelos predictivos. Se observa que varias variables tienen puntos alejados del rango intercuartílico, lo que sugiere la necesidad de decidir si se corrigen, eliminan o mantienen estos datos según el objetivo del análisis.

5. Correlaciones

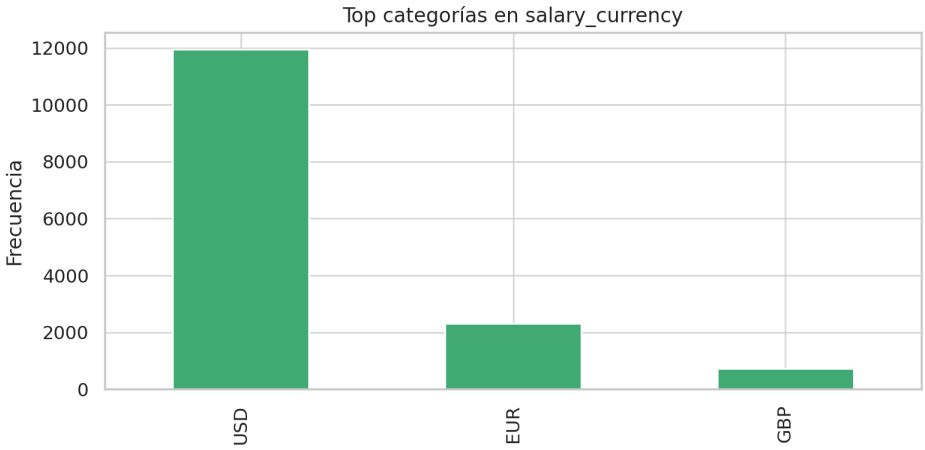
Matriz de correlación entre variables numéricas:

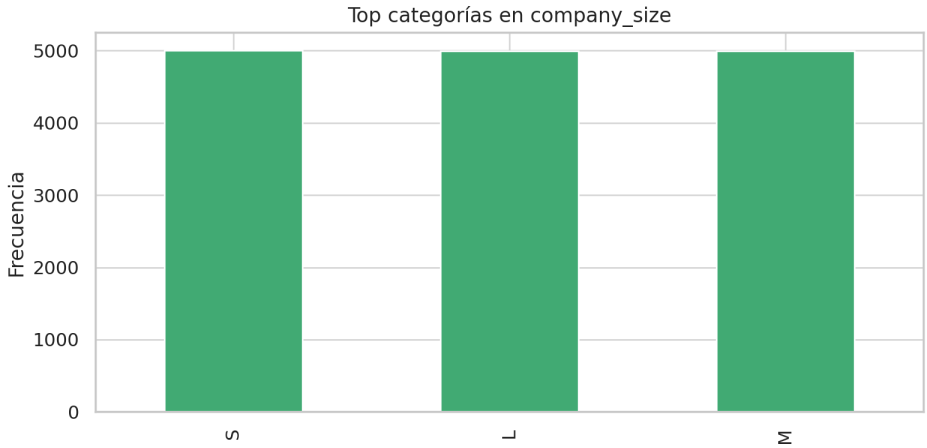


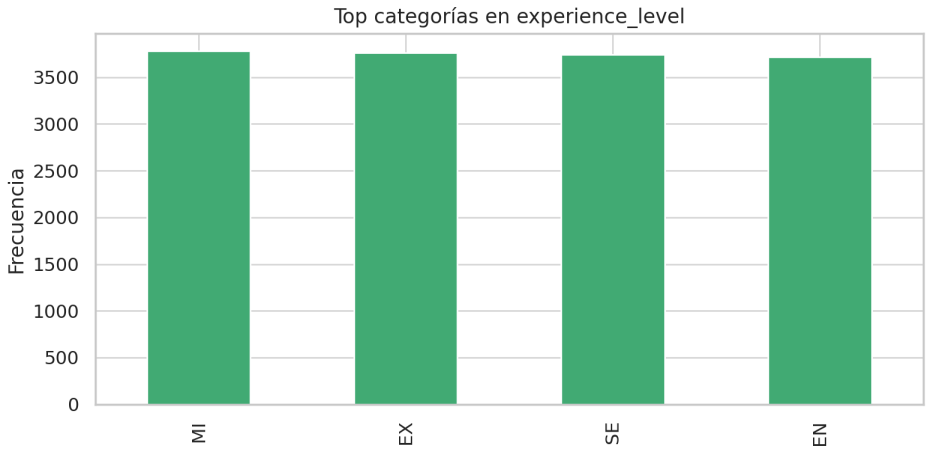
La matriz de correlación evidencia qué tan relacionadas están entre sí las variables numéricas. En este caso, algunas variables muestran correlaciones significativas, lo cual es útil para detectar redundancias de información o posibles dependencias. También ayuda a identificar qué variables podrían tener mayor peso en modelos predictivos relacionados con empleos en IA.

6. Variables Categóricas

Distribución de categorías más frecuentes en variables categóricas:







Los gráficos de barras muestran cuáles son las categorías más frecuentes en variables como puesto, país, tipo de contrato, nivel de experiencia, entre otras. Esto permite ver tendencias claras, como la concentración de oportunidades laborales en ciertos roles o regiones. También refleja la diversidad (o falta de ella) en las categorías registradas en el dataset.

7. Conclusiones del EDA

El conjunto de datos fue perfilado y visualizado para comprender su estructura y calidad. Los gráficos permiten comprender en detalle la estructura y comportamiento del dataset de empleos en inteligencia artificial. Se confirma que los datos están completos (sin nulos), aunque presentan distribuciones sesgadas y valores atípicos que deben ser gestionados. La matriz de correlación revela relaciones importantes entre variables numéricas, y las distribuciones categóricas muestran patrones de concentración en roles, países y niveles de experiencia. En conjunto, este análisis brinda un diagnóstico sólido para avanzar hacia la ingeniería de características y la construcción de modelos predictivos más precisos.