

Documentación de Limpieza y Modelado de Datos

1. Extracción de datos

Se extrajo la base de datos original desde el Google Sheets público del Salary Survey de Ask A Manager. Los datos fueron cargados en Python (Google Colab) para su procesamiento y limpieza.

2. Renombramiento de variables

Se realizó un mapeo de nombres originales (preguntas del formulario) hacia nombres estandarizados en español para facilitar el análisis. Se utilizó un rename_map aplicado sobre el DataFrame.

3. Eliminación de registros nulos

Se eliminaron registros con valores nulos en variables críticas: nombre_trabajo, ciudad, industria, nivel_educacion, genero y etnia, asegurando consistencia analítica.

4. Limpieza de variables salariales

Se limpian las columnas salario_anual y compensacion_monetaria eliminando comas, caracteres especiales y textos. Posteriormente se convirtieron a formato numérico.

5. Estandarización de monedas

Se consolidaron las columnas moneda y otra_moneda en una sola variable moneda_total. Se estandarizaron nombres de monedas duplicadas o mal escritas.

6. Conversión de divisas

Se aplicaron tasas de conversión (fx_to_usd) para llevar todos los salarios a USD. Posteriormente se convirtieron a COP usando la TRM del día del análisis.

7. Cálculo de compensación total

Se generaron los campos salario_usd, compensacion_usd, salario_cop, compensacion_cop y total_compensacion_cop como base para el análisis.

8. Limpieza geográfica

Se estandarizaron países mediante un country_map. Se eliminaron registros que no correspondían a países reales.

9. Imputación de ciudades

Cuando la ciudad era nula, se imputó la capital del país correspondiente mediante capital_map.

10. Tratamiento de outliers

Se eliminaron valores extremos utilizando percentiles (5% inferior y 5% superior) para evitar distorsiones en el análisis.

11. Dataset final

Se consolidó un dataset modelado listo para visualización en Looker Studio y documentación analítica.