Limpieza y Documentación de Dataset de Freelancers Globales en Excel

Autor: Victor Manuel Saavedra Radilla

Fecha: 17/09/2025

Introducción

Este proyecto consiste en la limpieza y documentación de un dataset de 1000 freelancers globales. El dataset incluye información demográfica, profesional y financiera como género, edad, país, habilidades, años de experiencia, tarifa por hora, calificación y satisfacción del cliente. El objetivo es aplicar un proceso de limpieza en Excel, utilizando funciones, filtros, Power Query y tablas dinámicas, para preparar los datos para análisis posteriores.

Metodología

- Importar el dataset en Excel y usar Power Query para iniciar la limpieza.
- Eliminar espacios extra y normalizar mayúsculas/minúsculas.
- Estandarizar valores inconsistentes (ejemplo: 'f', 'FEMALE', 'female' → 'Female').
- Transformar formatos de columnas (tarifas sin símbolos, porcentajes a números).
- Tratar valores nulos mediante reemplazo con promedios o dejarlos vacíos según corresponda.
- Validar consistencia con filtros y tablas dinámicas.

Problemas encontrados

- Valores nulos en edad, experiencia, tarifa por hora, calificación y satisfacción del cliente.
- Inconsistencias en género (f, F, female, etc.) y estado activo (0, 1, N).
- Formatos mixtos en tarifa por hora (100, USD 100, $40).
- Porcentajes representados como texto (ej. '84%').
- Calificaciones con valor 0.0 interpretadas como 'no calificado'.

Proceso de Limpieza de Datos

Antes de realizar cualquier modificación en los datos, se recomienda crear una copia de la columna a trabajar. Todas las transformaciones deben aplicarse en esta copia, verificando los resultados antes de eliminar la columna original. Una vez validada la limpieza, la copia reemplaza a la columna inicial.

1. Columna name

La columna contenía nombres, apellidos y honoríficos en un mismo campo. El proceso seguido fue:

- Identificar registros con más de tres palabras, lo que permitía detectar posibles honoríficos.
- Los honoríficos encontrados fueron: Ms., Mrs., Mr., DDS, MD, Jr., Dr., III, PhD, DVM, Miss.- Se creó una columna condicional en Power Query para extraer el honorífico cuando estuviera presente.
- Posteriormente, se utilizaron funciones de reemplazo para eliminar estos honoríficos de lacolumna de nombres.
- Finalmente, se dividió la columna en ‘Nombre’ y ‘Apellido’, utilizando el espacio como delimitador.

2. Columna gender

Los valores presentaban variaciones como 'f', 'FEMALE' o 'female'. Para normalizarlos:

- Se aplicó la función de formato 'Capitalize Each Word'.
- Se reemplazó 'F' por 'Female' y 'M' por 'Male', usando la opción de coincidencia exacta.
- Se verificó en los filtros que únicamente quedaran 'Male' y 'Female'.

3. Columna age

- Se detectaron valores nulos.
- Se calculó el promedio con la función 'Average' en la pestaña 'Statistics'.
- El promedio redondeado se utilizó para reemplazar los valores nulos.

4. Columna years_of_experience

- Se siguió el mismo proceso que en la columna 'age': cálculo del promedio y reemplazo de valores nulos.

5. Columna hourly_rate (USD)

- Se identificaron formatos mixtos como '100', 'USD 100', '$40'.
- Se reemplazó 'USD ' y '$' por vacío.
- Se cambió el tipo de dato a número.

6. Columna rating

- Se eliminaron los valores iguales a 0.
- Los valores nulos se conservaron para no afectar el análisis.

7. Columna is_active

- Presentaba valores 0, 1, N, Y.
- Se aplicó 'Capitalize Each Word'.
- En Power Query se reemplazó: '0' y 'N' por 'Inactive'; '1' y 'Y' por 'Active'.- Finalmente, se transformó a tipo de dato booleano (TRUE/FALSE).