Limpieza y Transformación de un Dataset

María José Hernández Rosales

Juan Guillermo González Gutiérrez

Universidad Digital de Antioquia

Proyecto Integrado III

Sharon Karin Camacho

24 de mayo de 2025

Descripción y necesidad de limpieza

1. Se crea una copia del dataframe para efectuar el análisis y la transformación de los datos.

Imagen 1

```
# Se crea una copia del dataframe
copy_combined_df = combined_df.copy()
```

2. De acuerdo a lo evidenciado en la actividad exploratoria del dataset, se identificaron variables cuyos valores nulos superaban el 70%. Al no ser datos relevantes se toma la decisión de eliminarlas (Standard error, Lower Confidence Interval, Upper Confidence Interval).

Imagen 2

```
# Eliminar las columnas con más del 70% de valores nulos
copy_combined_df.drop(columns=eliminar_columna_mayor_70, inplace=True)
```

3. En el dataset, varias columnas tienen valores nulos. Para llenar estos valores se implementa como estrategia trabajarlos como la moda (Happiness Score, Economy (GDP per Capita), Family, Health (Life Expectancy, Trust (Government Corruption), Generosity, Dystopia Residual, Social support) ya que son variables indispensables para su posterior análisis.

Imagen 3

```
[OK] 'Happiness Score' rellenada con la moda: 2.905
[OK] 'Economy (GDP per Capita)' rellenada con la moda: 0.0
[OK] 'Family' rellenada con la moda: 0.0
[OK] 'Health (Life Expectancy)' rellenada con la moda: 0.815
[OK] 'Trust (Government Corruption)' rellenada con la moda: 0.082
[OK] 'Generosity' rellenada con la moda: 0.175
[OK] 'Dystopia Residual' rellenada con la moda: 0.32858
[OK] 'Social support' rellenada con la moda: 1.125
```

4. La variable categórica “Region” presenta casi el 50% de valores nulos. Para dar solución, se crea un diccionario con las series país y región.

Imagen 4

	0
Country	0.000000
Region	49.760766
Overall rank	0.000000
Happiness Score	0.000000
Economy (GDP per Capita)	0.000000
Family	0.000000
Health (Life Expectancy)	0.000000
Freedom	0.000000
Trust (Government Corruption)	0.000000
Generosity	0.000000
Dystopia Residual	0.000000
Year	0.000000
Social support	0.000000

5. Todas las variables tienen el tipo indicado para el análisis que se requiere.

Validación

Completitud de datos: el dataset contiene 627 filas y 13 columnas. No hay valores nulos.

Relevancia de las variables: Las columnas representan valores comunes en estudios de felicidad por países.

Granularidad: Es la adecuada para el análisis comparativo anuales por país