Tarea 4 de agosto: Revisión de variables de la base de datos

Samuel Blanco Castellanos

4 de agosto

1. Introducción

Este documento presenta un análisis automático de la base de datos cargada desde el archivo `Datos_2023_348.xlsx` (hoja: `Datos_2023_348`). La tabla contiene **72558** filas y **13** columnas.

2. Tipo de base y estructura

Tipo de base de datos (heurístico): Panel (longitudinal por ID y tiempo). *Clave sugerida*: `consecutive_origen`, `FEC_NOT`.

3. Diccionario de variables

Para cada variable se identifica su rol analítico, tipo de dato, número de valores únicos, porcentaje de nulos y un ejemplo.

Cuadro 1: Diccionario de variables con tipología y calidad.

Variable	Rol	Tipo	Unicos	Pct$_N ulos$	Ejemplo
FEC_NOT	Temporal	Fecha/Hora	395	0.0	2023-10-01
SEMANA	Métrica	Numérica (entera)	52	0.0	39
ANO	Categoría	Numérica (entera)	1	0.0	2023
FEC_CON	Temporal	Fecha/Hora	396	0.0	2023-09-30
confirmados	Categoría	Numérica (entera)	2	0.0	1
consecutive_origen	Identificador (ID)	Numérica (entera)	72558	0.0	40738
va_sispro	Categoría	Numérica (entera)	1	0.0	1
Estado_final_de_caso	Categoría	Numérica (entera)	3	0.0	3
nom_est_f_caso	Categoría	Texto	3	0.0	Confirmado por laboratorio
Departamento_ocurrencia	Texto libre	Texto	34	0.0	BOGOTA
Municipio_residencia	Texto libre	Texto	906	0.0	BOGOTA
Departamento_Notificacion	Texto libre	Texto	30	0.0	BOGOTA
Municipio_notificacion	Texto libre	Texto	170	0.0	BOGOTA

4. Calidad de datos

- Filas duplicadas detectadas: **0**.
- Columnas constantes (sin variación): 2.
- Columnas con más del 70 % de nulos: 0.
- Columnas numéricas con varianza cero: 2.

5. Relaciones entre datos

Se calcularon correlaciones entre variables numéricas para identificar relaciones lineales fuertes (tabla 2). Para variables categóricas, se recomienda evaluar asociaciones (p.ej., V de Cramér) y pruebas *chi-cuadrado* por pares de categorías de baja cardinalidad.

Cuadro 2: Top 10 correlaciones absolutas entre variables numéricas.

Var_1	Var_2	AbsCorr
SEMANA	consecutive_origen	0.0035588623827884094
SEMANA	Estado_final_de_caso	0.008874976907766114
SEMANA	confirmados	0.0005619849687798687
confirmados	consecutive_origen	0.026248788142751174
confirmados	Estado_final_de_caso	0.852486984983822
consecutive_origen	Estado_final_de_caso	0.02596438933042489

6. Recomendaciones

Qué sirve (priorizar para análisis/ETL):

- FEC_NOT
- SEMANA
- FEC_CON
- confirmados
- consecutive_origen (para joins / claves)
- Estado_final_de_caso
- nom_est_f_caso
- Departamento_ocurrencia
- Municipio_residencia
- Departamento_Notificacion
- Municipio_notificacion

Qué no sirve (eliminar o transformar):

- ANO
- va_sispro

7. Anexo: Supuestos y criterios

Criterios usados (resumen):

- **Rol**: Identificador si la cardinalidad es igual al número de filas y hay pocos nulos; Categoría si hay baja cardinalidad (≤ 20 únicos aprox.) o tipo booleano; Métrica si es numérica continua; Temporal si es fecha u hora; Texto libre si es cadena de alta cardinalidad.
- **Qué no sirve**: columnas constantes, con $> 70\,\%$ de nulos, duplicadas de otras columnas o con varianza cero.
- **Tipo de base**: heurístico basado en la presencia de columnas temporales y/o claves de entidad (ID) repetidas en el tiempo (panel).