**UNIVERSIDAD DE LA SABANA**

**FACULTAD DE INGENIERÍA**

**PROGRAMACIÓN APLICADA**

Importación y Limpieza de Datos

Objetivo: Importar los datos y realizar una limpieza preliminar.

* Limpieza de Datos Inicial: Identificación y errores.

**Identificación:**

Principalmente se utilizaran dos variables relacionadas: desaparecidos (7 bases de datos), Homicidios (8 bases de datos).

* **Descripción de base de datos Desaparecidos:**
  + Colombia, Solo departamento, Solo municipio, Bogotá: Cantidad hombre y mujer, años, escolaridad, estado conyugal, factor de vulnerabilidad, ancestro racial, estado de la desaparición, entidad receptora del caso, zona del hecho, mes del hecho, día del hecho, en el extranjero hombre y mujer, el total de desaparecidos.
* Departamento con su municipio: Cantidad hombre y mujer, años, escolaridad, estado conyugal, factor de vulnerabilidad, ancestro racial, estado de la desaparición, entidad receptora del caso, zona del hecho, mes del hecho, día del hecho, en el extranjero hombre y mujer, el total de desaparecidos, tasa por cada 100.000 habitantes código divipola de cada departamento y municipio.
* **Descripciones de base de datos Homicidios:**
  + Colombia, Solo departamento, Solo municipio, Bogotá: Cantidad hombre y mujer, años, escolaridad, estado conyugal, factor de vulnerabilidad, ancestro racial, presunto agresor agrupado y al detalle, circunstancia del hecho, actividad durante el hecho, mecanismo causal, diagnóstico topográfico de la lesión, zona del hecho, escenario del hecho, mes del hecho, día del hecho, hora del hecho.
  + Departamento con su municipio: Cantidad hombre y mujer, años, escolaridad, estado conyugal, factor de vulnerabilidad, ancestro racial, presunto agresor agrupado y al detalle, circunstancia del hecho, actividad durante el hecho, mecanismo causal, diagnóstico topográfico de la lesión, zona del hecho, escenario del hecho, mes del hecho, día del hecho, hora del hecho, tasa por cada 100 habitantes código divipola de cada departamento y municipio.

**Errores.**

Las bases de datos presentan varías condiciones diferentes que impide globalizar errores comunes.

Las bases de datos de “Homicidios” no presentaron errores al momento de hacer todo el proceso de limpieza de los datos; cabe recalcar que la base de datos de Homicidios 2023, aunque no presentó errores finales, si se manejo cuidadosamente pues los datos compartían un solo documentos con lesiones en general (Accidentes no fatales, suicidios, etc),

Las bases de datos de “Desaparecidos” si presenta muchos más errores en la tabla final que se presenta después de todo el proceso, en algunas bases el error va desde 500 hasta 700 celdas de error, error que seguimos cuidadosamente y determinamos que probablemente es por la presentación de celdas “vacías”.

* Documentación de la Limpieza de Datos: Descripción de los pasos tomados para limpiar los datos y las decisiones justificadas.

**Descripción de los pasos tomados para limpiar los datos.**

para nuestro análisis eran "Departamento de Hecho" y "Municipio de Hecho". Estos campos eran fundamentales para profundizar en el análisis, por lo que decidimos enfocarnos en estos.

A continuación, trans pusimos las tablas en ambas bases de datos y eliminamos todos los datos que representaban ruido, como totales y variables irrelevantes para nuestro estudio. Para mantener la coherencia, renombramos las variables de manera clara. Posteriormente, utilizando una tabla de referencia y un rango específico para organizar la información. Esto nos permitió construir una base de datos final más limpia, y Finalmente, unificamos todas las variables relevantes para facilitar la búsqueda de datos específicos y garantizar que el análisis posterior sea más eficiente.

**Las decisiones justificadas.**

No utilizamos datos como "General Colombia", "Departamento", "Municipio", ni "Bogotá DC" porque estos no proporcionaron la profundidad de información necesaria para nuestro análisis. Aunque estos campos ofrecían un panorama general, carecían de la especificidad requerida para profundizar

Los datos que seleccionamos, como "Departamento de Hecho" y "Municipio de Hecho", no solo incluían la información presentada en los campos más generales, sino que además estaban más desglosados ​​y detallados. Este nivel de detalle nos permitió realizar un análisis más completo y preciso, ya que contábamos con un mayor volumen de datos y un desglose más exhaustivo, lo que es crucial para identificar patrones y tendencias con mayor exactitud.