**BENEMÉRITA UNIVERSIDAD**

AUTÓNOMA DE PUEBLA



**Práctica de Laboratorio - Unidad 2: Introducción a la Limpieza de Datos**

**Reporte de Limpieza de una Base de Datos Ensuciada**

Asignatura: INTRODUCCION A LA CIENCIA DE DATOS

Carrera: INGENIERIA EN CIENCIA DE DATOS

Nombre del alumno: JOANA MONTSERRAT CARMONA NOGUEZ

Nombre del profesor: JAIME A. ROMERO SIERRA

Horario de la materia: LUNES, MARTES Y JUEVES

## No. De cuenta: 202466401

## Fecha de entrega: 22 DE OCTUBRE DE 2024

Práctica de Laboratorio - Unidad 2: **Introducción a la Limpieza de Datos**

Materia: **Introducción a la Ciencia de Datos**

Unidad 2: **Procesamiento y Limpieza de Datos**

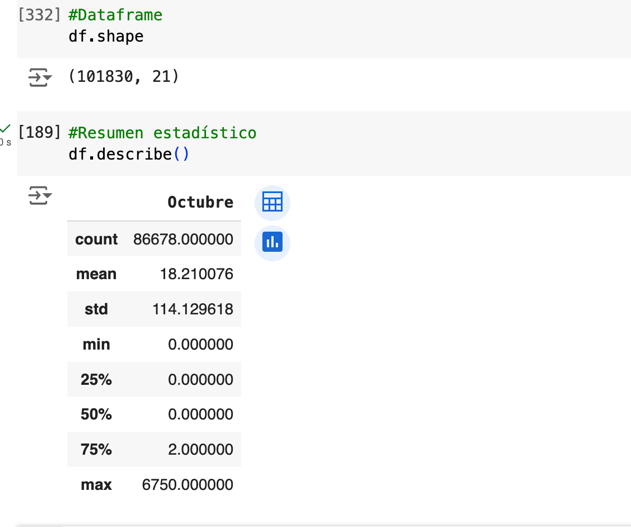
Título de la Práctica: **Limpieza de una Base de Datos Ensuciada**

Nombre: **Joana Montserrat Carmona Noguez**

Día y horario de la materia: **Lunes, Martes, Jueves**

Análisis inicial:

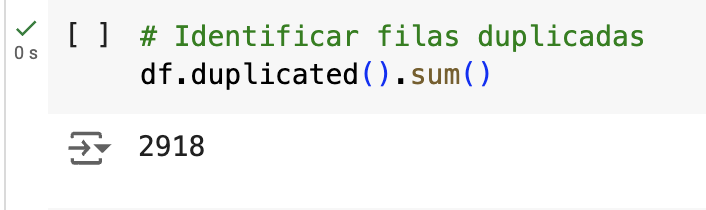
▪ Resumen estadístico de la base de datos antes de la limpieza.



▪ Tabla que muestre el porcentaje de valores faltantes por columna.



▪ Total de filas duplicadas encontradas.



▪ Descripción de los tipos de datos originales y los problemas encontrados.



Proceso de limpieza:

▪ Describir qué métodos utilizaron para limpiar la base de datos (eliminación, imputación, etc.).

Se aplicaron varias técnicas para limpiar la base de datos. Primero, los valores faltantes (NaN) en columnas clave como Entidad, Bien jurídico afectado, Tipo de delito y Subtipo de delito fueron reemplazados por "Desconocido", lo que permitió mantener la integridad del conjunto de datos y facilitar la identificación de información incompleta sin perder registros importantes. En los casos donde los datos faltantes eran irrecuperables o irrelevantes, se optó por eliminar las filas afectadas, garantizando que la base de datos permaneciera consistente.

También se corrigieron las inconsistencias en los tipos de datos, por ejemplo, la columna Año fue convertida a formato numérico para asegurar que pudiera utilizarse en análisis cronológicos sin errores. Además, se eliminaron los duplicados, evitando que los eventos delictivos se contaran más de una vez, lo que podría sesgar los resultados. Por último, se corrigieron valores mal escritos o inválidos en algunas columnas, reemplazándolos con valores correctos para asegurar la coherencia de los datos.

▪ Mostrar antes y después de cada paso clave (por ejemplo, antes y

después de eliminar duplicados).

