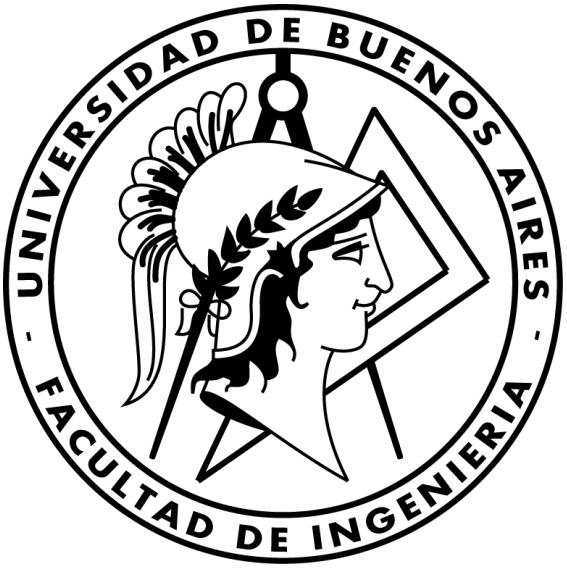
Trabajo Práctico: Procesamiento y Carga de Datos en una Base de Datos

Materia: Base de Datos

Grupo: 6



Alumno/a	*Padrón*
Victoria Avalos	108434
Gonzalo Manuel Calderón	107143
Mateo Liberini	104867
Franco Agustin Rodriguez	108799
Sebastián Brizuela	105288
Urbano Sol Guadalupe	109525

Índice

Introducción

Este Trabajo Práctico se centra en el análisis de un Conjunto de Datos seleccionado libremente, con el objetivo de aplicar y profundizar en los conceptos y técnicas discutidos durante el curso. Para ello, implementaremos un proceso de ETL (Extracción, Transformación y Carga). Este proceso se ha convertido en una piedra angular en el ámbito de la integración de datos, permite a las organizaciones recopilar datos de diversas fuentes, limpiarlos, transformarlos y cargarlos en un almacén de datos centralizado o data mart.

Iniciaremos extrayendo una base de datos a elección, para luego transformarla adecuadamente mediante un proceso batch. Una vez que los datos estén limpios y preparados, los exportamos a una base de datos SQLite. Esto nos permitirá acceder fácilmente a los datos preprocesados y realizar consultas SQL para análisis avanzados. Este enfoque no solo nos permitirá aplicar prácticamente los conocimientos adquiridos, sino que también nos proporcionará una plataforma robusta para realizar análisis significativos y extraer insights valiosos que respalden la toma de decisiones.

En cuanto a la estructura del trabajo, este está organizado en secciones que corresponden a cada paso del proceso de análisis y preprocesamiento de datos. Cada sección aborda un aspecto específico del trabajo, desde la carga inicial de datos hasta la exportación final a la base de datos.

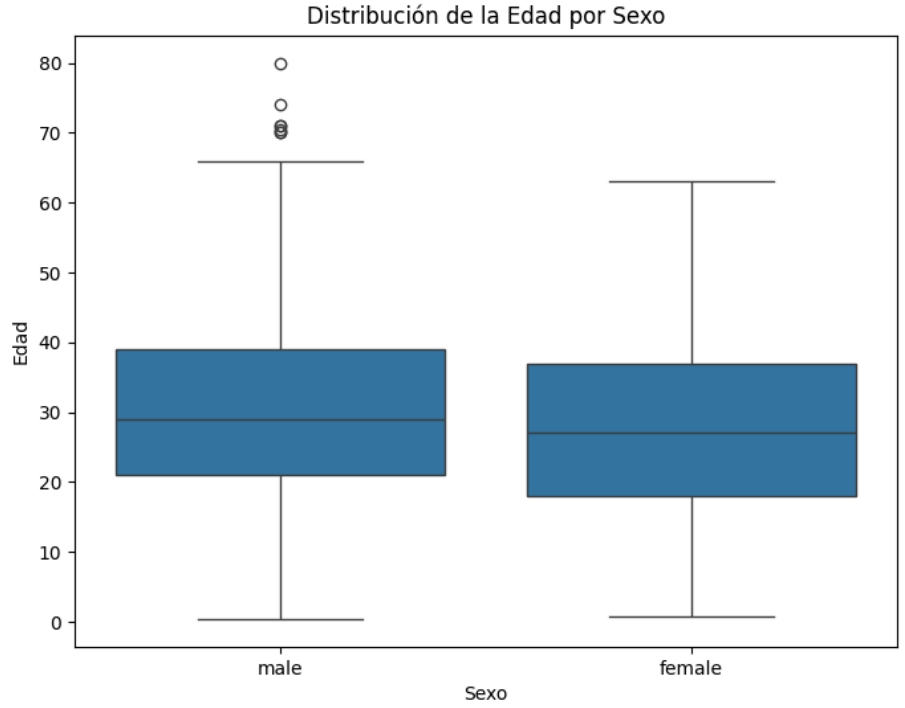
Análisis exploratorio

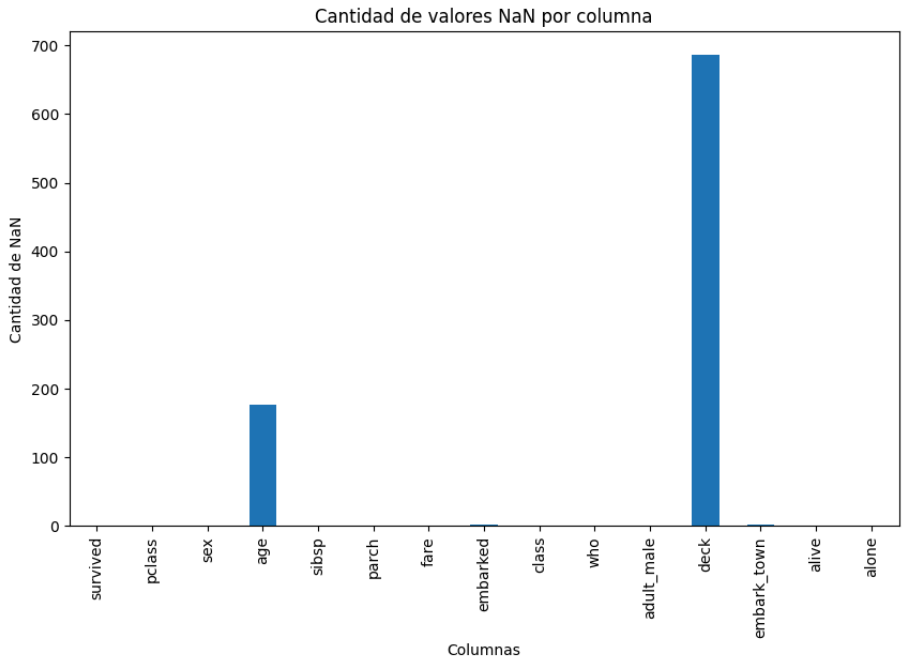
Hemos decidido utilizar el Dataset de Titanic brindado por la plataforma Kaggle, este es muy utilizado en el ámbito de la ciencia de datos. Este consta de la información de los pasajeros, desde sí sobrevivieron, hasta edad y acompañantes. Este fue propuesto, ya que su propósito es el de utilizarlo para hacer un análisis y preprocesamiento de datos para un uso apropiado de los mismos. Lo cual resulta ideal para nuestro objetivo.

Decidimos utilizar el framework de Pandas para realizar tanto el análisis exploratorio como el preprocesamiento, debido al tamaño moderado del dataset que elegimos. Comenzamos obteniendo información básica sobre los datos, como la cantidad de filas, tipo de dato que contiene cada columna y cantidad de elementos nulos en cada columna. Luego, con el valor de la cantidad de filas (889), obtuvimos el tamaño que debería tener una muestra aleatoria para garantizar que sea representativa con un 5% de margen de error y un 95% de confianza. Entonces, generamos una muestra de dicho valor, el cual resultó ser 269, para poder apreciar la composición de los datos.

También obtuvimos datos estadísticos sobre las columnas numéricas, con su cantidad de elementos no nulos, promedios, máximos, mínimos y percentiles.

Finalmente realizamos dos visualizaciones: un boxplot con la distribución de la edad por sexo y un barplot con la cantidad de valores *NaN* por columna.





Preprocesamiento

En esta etapa, se tomaron varias decisiones importantes que incluyen la eliminación de varias columnas que se consideraron redundantes o que contenían una cantidad significativa de datos nulos.

La primera columna eliminada fue "***male***", ya que la información sobre si un individuo era adulto y masculino o no, se podía inferir de manera adecuada a través de la columna "***who***". Dado que esta última columna proporcionaba una clasificación más detallada que incluía género y edad, se consideró más relevante y suficiente para el análisis.

Luego, se removió la columna "***embarked***", ya que la información sobre el puerto de embarque de los pasajeros estaba contenida de manera más explícita y clara en la columna "***embark_town***". Esta decisión simplificó el conjunto de datos al eliminar la redundancia de información.

La columna "***class***" también fue excluida, ya que la información sobre la clase que le corresponde a pasajero se podía deducir de la columna "***pclass***", que representa el número de clase. Al eliminar "class", se redujo la duplicación de información y se mantuvo la integridad de los datos.

Una columna más que se quitó fue "***deck***", debido a la gran cantidad de valores nulos que contenía. Dado que la falta de datos era demasiado significativa, no era factible utilizar esta columna de manera efectiva en el análisis. Por lo tanto, se optó por eliminarla para mantener la calidad y la coherencia de los datos.

Finalmente, la columna "***alone***" también se consideró redundante y se descarto. Esta decisión se basó en la observación de que la información sobre si un pasajero viajaba solo o acompañado podía deducirse de manera precisa examinando las columnas "***sibsp***" (número de hermanos/cónyuges a bordo) y "***parch***" (número de padres/hijos a bordo).

Luego, analizamos qué hacer con las filas duplicadas. Resulta que había pasajeros que compartían las mismas características que otros, por lo que había una considerable cantidad de filas repetidas. Ya que no se puede perder la información de numerosos pasajeros pero la tabla no está de forma óptima, decidimos mantener el primer elemento de las filas repetidas y agregar una columna extra con el número de los pasajeros que comparten las características de su fila. De ésta forma no se pierde información y la tabla resulta más eficiente.

Respecto de las imputaciones, habían dos pasajeros los cuales tenían "***embark_town***" como NaN. Decidimos eliminar dichas filas ya que es imposible inferir de los demás datos una posible ciudad. Luego, habían muchas edades faltantes, por lo que decidimos imputarlas con la edad promedio de su respectivo sexo. En otras palabras, calculamos la edad promedio de las mujeres y de los hombres a bordo, y reemplazamos las edades faltantes en base a si eran hombres o mujeres.

Finalmente, decidimos eliminar la columna "***who***" (que determina si era hombre, mujer o infante) ya que sus datos pueden verse con las columnas "sex" y "age"; no la eliminamos antes ya que lo utilizamos primero para imputar las edades faltantes.

En resumen, estas acciones de preprocesamiento simplificaron y mejoraron la calidad del conjunto de datos, eliminando la redundancia de información y abordando los valores nulos de manera efectiva. El conjunto de datos preprocesado está ahora listo para ser analizado en profundidad, revelando patrones y tendencias que pueden ofrecer valiosos insights para futuras investigaciones.

Conclusiones

La integración, transformación y carga eficientes de datos no solo mejoran la calidad de la información, sino que también habilitan análisis avanzados y toma de decisiones informadas. Al automatizar estos procesos, se puede ahorrar tiempo, reducir errores y enfocarse en generar valor a partir de sus datos. En resumen, una aplicación ETL es una inversión estratégica que puede proporcionar una ventaja competitiva significativa al permitir una gestión y análisis de datos más eficientes y efectivos.

Después de completar el proceso de preprocesamiento y carga de datos en una base de datos SQLite pudimos sacar varias conclusiones.

En primer lugar, se identificaron varios problemas comunes en los datos, como valores nulos, datos fuera de rango, inconsistencias en el formato y duplicados. Se propusieron soluciones para abordar cada uno de estos problemas, incluyendo la eliminación o imputación de valores nulos, la normalización de datos, la conversión de tipos de datos y la eliminación de duplicados. A su vez, pudimos notar bastante redundancia entre las distintas columnas.

Para todo el proceso de transformación de datos pudimos documentar el paso a paso en el notebook justificando cada una de nuestras decisiones.

Finalmente pudimos exportar el conjunto de datos trabajado a una base de datos SQLite y verificar que los datos fueron almacenados de manera consistente realizando las correspondientes consultas.

Con lo realizado a lo largo de este trabajo, pudimos ver la importancia que tiene un procesamiento ETL para el manejo de grandes volúmenes de datos y la obtención de información precisa y útil de ellos.

Anexos

- Google Colab
- Dataset Titanic
- Calculadora de muestra