FLUJO DE TRABAJO DEL PROYECTO M3: ANÁLISIS DE DATOS DE E-COMMERCE

El proyecto se encuentra estructurado en 6 módulos que se describen a continuación:

PARTE 1: Creación y descarga de base de datos

En esta etapa inicial, se generó un dataset ficticio para simular un entorno real de ventas. Las principales actividades son:

- Generación Aleatoria: Se utilizó numpy para crear 100 registros con IDs de transacciones, fechas del año 2025 y nombres de clientes ficticios. Se organiza la información usando arrays.
- Definición de Variables: Se crearon columnas para regiones (RM, BioBio, Valparaíso, etc.), edades de clientes (18-75 años) y un catálogo de 20 productos con categorías y precios específicos.
- Cálculo Automático: El monto total (total_amount) se derivó de la multiplicación del precio por la cantidad.
- Exportación de datos: Se guardan datos en listas y se crea en un archivo .csv y se descarga.

PARTE 2: DataFrame y revisión de datos

Usando la librería Pandas, se procedió a realizar una exploración inicial (EDA) para entender la estructura de la información.

- Exploración: Uso de head(), tail() y describe() para obtener estadísticas descriptivas (media, desviación estándar, valores mínimos y máximos).
- Filtrado: Se aplicaron filtros para aislar datos específicos, como ventas exclusivas en la Región Metropolitana (RM), por dos regiones específicas (Coquimbo y Maule), por edad (> 40 años) y condicionales combinados como por región (RM) y valor de la compra (>100usd).

PARTE 3: Obtención de datos desde múltiples fuentes

El análisis se enriqueció integrando información de distintas fuentes externas. Se desarrollaron las siguientes actividades:

- Carga Multi-formato: Se importaron datos desde archivos .csv y .xlsx.
- Web Scraping: Se extrajo información en tiempo real desde Wikipedia sobre la población de las regiones de Chile usando read_html.
- Organización de la información extraída de la web: Se extrajo solo la tabla y las columnas útiles para nuestro análisis.
- Homologación de Datos: Se estandarizaron los nombres (ej. Regiones y clientes) para crear claves en común entre fuentes de datos.

- Unificar fuentes de datos: se creó un nuevo DataFrame a través de la unión de las tablas mediante la función merge().

PARTE 4: Manejo de valores perdidos y outliers

Para asegurar la calidad del análisis, se realizó una limpieza profunda del dataset a través de las siguientes actividades:

- Identificación y gestión de valores nulos: Se imputaron valores faltantes de price y total_amount mediante lógica matemática (Precio = Total / Cantidad y Total= precio X cantidad). Las calorías faltantes se completaron con la media del grupo. Se eliminaron registros no válidos.
- Tratamiento de outliers: Se identificó las variables (columnas) que podrían contar con valores atípicos (precio, monto, calorías y edad). Se aplicó uno de los métodos más robustos para la identificación de outliers que es el método del Rango Intercuartílico (IQR) para garantizar que los datos extremos no sesguen los resultados.

PARTE 5: Data Wrangling (Transformación)

En esta fase se prepararon los datos para el análisis estadístico avanzado, a través de las siguientes acciones:

- Eliminación de registros duplicados: Se eliminaron filas con datos iguales, manteniendo un valor único.
- Transformación de tipo de datos: Se clasificó las variables de acuerdo a su naturaleza, asegurando que los formatos sean correctos para fechas, enteros y flotantes.
- Creación de nuevas columnas calculadas: se crea una nueva variable con las calorías de todos los productos de la venta (calorías totales = calorías X cantidad).
- Aplicación de funciones personalizadas: Creación de rangos etarios (Joven, Adulto, Adulto Mayor) mediante funciones lambda. Mapeo de categorías usando la función map.
- Normalización y discretización: Se normalizó el valor total de las ventas (escala 0 a 1) para comparaciones uniformes entre gastos de clientes. Discretización de variables clasificando el monto total en un nivel de gasto "Bajo", "Medio" y "Alto" usando qcut.

PARTE 6: Agrupación y pivoteo de datos

Finalmente, en la parte 6 del proyecto se usaron técnicas para generaron las métricas clave para el negocio. Se desarrolló las siguientes acciones:

- Agrupación de datos: Se calculó la suma y el promedio de ventas agrupados por región y rango etario usando la función groupby.
- Tablas Dinámicas: Se utilizó pivot_table para reestructurar los datos, permitiendo visualizar el promedio de gasto de cada grupo de edad por región de forma horizontal. Se usó la función pivot.melt para visualizar nuevamente los datos en formato largo.
- Exportación: El flujo concluyó con la generación de los archivos finales en formato .csv y .xlsx listos para su presentación o uso en otros softwares como Excel o PowerBI.