Entregable I – Exploración y Preprocesamiento de Datos

Enfoque:

Comportamiento de compra

Grupo No. 6:

Altamirano Ortuño Ashley

Falconi Ortiz Valentina

Figueroa Benitez Frederick

Reyes Holguin Roberto

Salvatierra Samaniego Jairo

Módulo:

Data Visualization

Profesor/a:

Ing. Martha Tomalá

Enero 2026

Estado actual del dataset y las mejoras implementadas

Problemas detectados:

1. **Ausencia de identificador único:** No existía columna ID para identificar clientes de manera individual pese a que en el documento se mencionaba que existía
2. **Formato incorrecto en variables monetarias:** Columnas estaban en formato texto con símbolos "$" y comas
3. **Variables categóricas fragmentadas:** Las variables Marital_Status y Education ocupaban múltiples columnas
4. **Presencia de outliers:** Se identificaron valores atípicos en 4 variables de comportamiento de compra

Mejoras Realizadas en el Preprocesamiento

1. **Creación de identificador único**
   - Se generó columna ID como primera columna del dataset lo que resuelve la ausencia de clave primaria en el dataset original
2. **Conversión de variables monetarias**
   - Limpieza de símbolos "$" y comas en 8 columnas monetarias
   - Conversión de tipo texto (object) a tipo numérico (float64)
   - Variables procesadas: Income, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, MntTotal
3. **Reconstrucción de variables categóricas**
   - Marital_Status: Consolidación de 5 columnas (marital_Divorced, marital_Married, marital_Single, marital_Together, marital_Widow) en una sola variable categórica
   - Education: Consolidación de 5 columnas (education_2n Cycle, education_Basic, education_Graduation, education_Master, education_PhD) en una sola variable categórica
   - Eliminación de 10 columnas originales y creación de 2 nuevas
4. **Análisis de outliers**
   - Se analizaron 10 variables numéricas mediante método IQR (Rango Intercuartílico)
   - Se detectaron outliers en 4 variables:
     - **NumDealsPurchases:** 82 outliers detectados

    - **NumWebPurchases:** 3 outliers detectados
    - **NumCatalogPurchases:** 20 outliers detectados
    - **NumWebVisitsMonth:** 8 outliers detectados
- **Decisión:** Los outliers se mantuvieron por representar comportamientos extremos válidos porque pueden ser clientes de alto valor o muy activos

Validación de calidad

- Sin valores nulos
- Sin valores inválidos en Age, es decir que 0 registros con valor 99999 como se mencionaba
- Sin duplicados
- Tipos de datos correctos validados
- Dataset exportado como marketing_procesado.csv

Resultado Final

Dataset limpio con 2,205 registros y 32 columnas listo para un análisis

Preguntas de negocio

Pregunta 1: ¿Qué canal genera mayor volumen de compras?

Justificación: Identificar si los clientes prefieren comprar por web, tienda física, catálogo o con descuentos permite enfocar recursos en el canal más usado.

Pregunta 2: ¿Qué categoría de productos concentra el mayor gasto?

Justificación: Determinar si los clientes gastan más en vinos, carnes, pescados, frutas, dulces o productos premium permite priorizar inventario y promociones.

Pregunta 3: ¿Cuál es el ticket promedio de compra de los clientes?

Justificación: Conocer el gasto promedio por cliente permite diseñar estrategias para incrementar el valor de cada transacción.

Pregunta 4: ¿Existe relación entre visitas al sitio web y compras realizadas por ese canal?

Justificación: Analizar si más visitas generan más compras web permite evaluar la efectividad del sitio y detectar oportunidades de mejora.

Pregunta 5: ¿Con qué frecuencia compran los clientes?

Justificación: Medir la frecuencia de compra permite identificar clientes activos vs inactivos y diseñar campañas de reactivación.

KPIS

KPI 1: Volumen Total de Compras por Canal

KPI 2: Gasto Total por Categoría de Producto

KPI 3: Ticket Promedio (Gasto Promedio por Cliente)

KPI 4: Ratio Compras Web / Visitas Web

KPI 5: Frecuencia Promedio de Compra Total