CODING BOOTCAMPS ESPOL:
Data-Driven Decisions Specialist

Data Visualization

Entregable Final

Ramirez Scamaronez David Enrique

Peralta Baidal Darwin Vicente

Intriago Celi Romina Anahi

Rubira Espinoza Ivonne Bethsabe

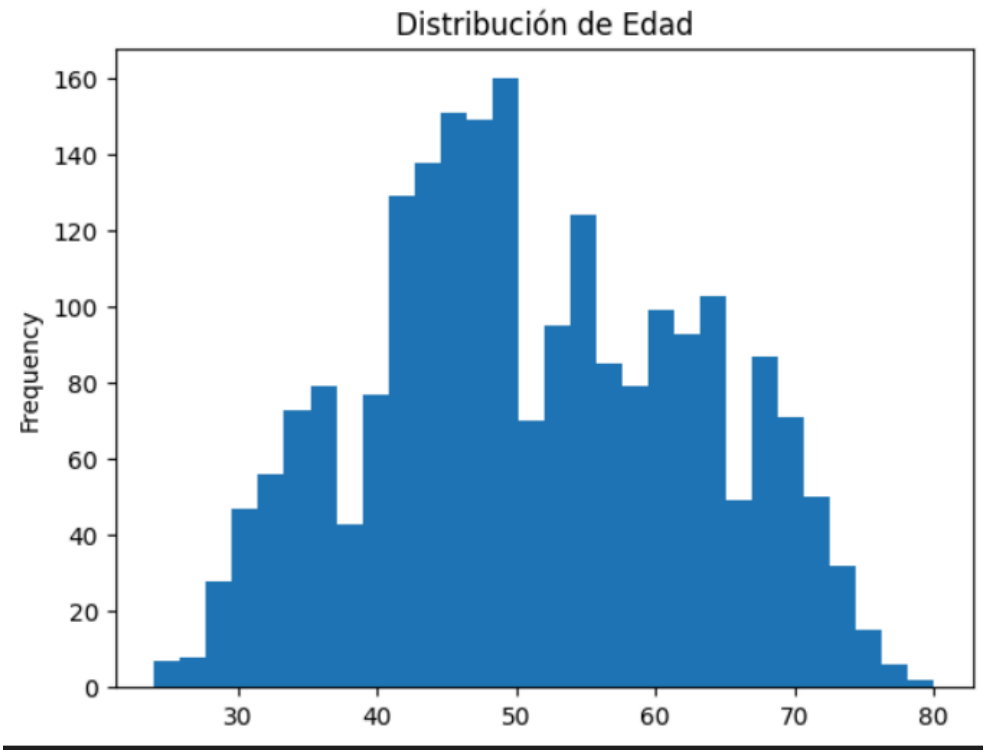
Vera Espinoza Angel Isaac

Resumen del estado actual del dataset, y las mejoras realizadas de la calidad en el preprocesamiento realizado.

Diagnóstico del Estado Inicial

El análisis exploratorio (EDA) de los datos crudos (marketing_raw) permitió identificar vulnerabilidades que habrían invalidado cualquier conclusión de negocio si no se trataban adecuadamente. Los hallazgos principales fueron:

- Inconsistencias de Tipo y Formato: Varias columnas clave de consumo (como MntWines, MntFruits, etc.) se importaron originalmente como texto (object). Esto se debió a la presencia del símbolo de moneda ($) y comas. Sin una conversión numérica, era imposible calcular el gasto total o el retorno de inversión de las campañas.
- Integridad de la Variable Ingresos (Income): Se detectó que 24 registros estaban vacíos. En un análisis de campañas, el ingreso es la variable predictora más importante; perder estos datos o dejarlos nulos habría sesgado la segmentación de los clientes "Aceptantes".
- Registros de Gasto Negativos: Se identificó un error crítico en la captura de datos en la columna MntRegularProds, con valores de hasta -283. Estadísticamente, un gasto negativo no tiene sentido en este contexto y distorsiona el cálculo del valor de vida del cliente.
- Valores Atípicos (Outliers): La presencia de clientes con años de nacimiento extremadamente antiguos (indicando edades imposibles de más de 120 años) y niveles de ingreso fuera de la escala normal generaba un "ruido" que inflaba la desviación estándar en los gráficos de distribución.




```python
monetary_cols = [
    "Income",
    "MntWines",
    "MntFruits",
    "MntMeatProducts",
    "MntFishProducts",
    "MntSweetProducts",
    "MntGoldProds",
    "MntTotal"
]

for col in monetary_cols:
    df[col] = (
        df[col]
        .replace('[\$,]', '', regex=True)
        .astype(float)
    )
```

```
df[['MntTotal','Income']].corr()
```

	MntTotal	Income
MntTotal	1.000000	0.823066
Income	0.823066	1.000000

Mejoras de Calidad en el Preprocesamiento

Para transformar el dataset en una fuente de verdad confiable, el equipo aplicó el siguiente flujo de limpieza documentado en el Notebook

- Sanitización de Datos Financieros: Se implementó una función de limpieza de cadenas para remover símbolos especiales, permitiendo la transformación de las variables de gasto a formato decimal (float64). Esto habilitó la creación de la métrica Total Spent.
- Imputación Estadística de Ingresos: Para salvar los registros con valores nulos en Income, se optó por la imputación mediante la mediana. Esta decisión es superior a usar el promedio, ya que la mediana no se ve afectada por los ingresos extremadamente altos de unos pocos clientes, manteniendo la representatividad del grupo.
- Depuración de Errores de Registro: Se procedió a la eliminación de filas con gastos negativos y registros con edades biológicamente implausibles. Esto redujo el dataset a 2,205 registros limpios, listos para el modelado.
- Estandarización de Variables Categóricas: Se simplificaron variables como el estado civil y nivel educativo para agrupar categorías con pocos registros, lo que permitió obtener gráficos de barras más claros y conclusiones más potentes.

Se ejecutó un proceso de normalización y limpieza de variables financieras para asegurar la integridad analítica del conjunto de datos. El procedimiento consistió en la eliminación sistemática de ruido de formato específicamente el símbolo de divisa y separadores de miles mediante expresiones regulares, seguido de una coerción de tipos a punto flotante. Esta transformación técnica es indispensable para habilitar el procesamiento matemático de las columnas, permitiendo que valores originalmente registrados como cadenas de texto sean operables para el cálculo de métricas agregadas y el entrenamiento de modelos predictivos.

```
monetary_cols = ['Income','MntWines','MntFruits','MntMeatProducts',
'MntFishProducts','MntSweetProducts','MntGoldProds','MntTotal']
for c in monetary_cols:
    df[c] = df[c].replace('[\$,]', '', regex=True).astype(float)
df[monetary_cols].dtypes
```

C. Reconstrucción de variables categóricas:

Se reconstruyeron las variables categóricas a partir de una lógica basada en los nombres de los campos; en la exploración vimos que las columnas de “marital” y “education” comparten prefijo, así que se guardaron en listas (marital_cols/education_cols) para identificar la categoría ganadora por fila, asignarla a nuevas columnas (marital_status, education_level) y, finalmente, eliminar las columnas regionales. Se validó previamente que cada fila tuviera exactamente un valor activo (suma igual a 1) en cada grupo de columnas antes de hacer la asignación.

D. Tratamiento de valores inválidos:

Se filtraron los registros donde la variable Age presentaba valores atípicos extremos. Se optó por una imputación por mediana. Primero, los valores inválidos se transformaron en valores nulos (NaN) para limpiar la serie estadística, y posteriormente se rellenaron con la mediana de la población restante. Se eligió la mediana por ser una medida de tendencia central robusta que no se ve afectada por los mismos valores extremos que se intentan corregir, manteniendo así la integridad de la muestra sin sesgar los resultados promedio del perfil del cliente.

E. Validaciones finales:

Se confirmó que las columnas financieras (como Income y los montos de gasto), que originalmente estaban en formato de texto (object) debido a símbolos de moneda, fueron convertidas correctamente a tipos numéricos (int64 o float64). En las variables Income y MntTotal, se aplicó una técnica de Capping (Winsorización) al percentil 99. Esto consistió en limitar los valores extremadamente altos al umbral del 99%, evitando que registros atípicos distorsionen las visualizaciones y los KPIs de gasto. Se realizó un conteo final mediante isnull().sum() para asegurar que el preprocesamiento no dejó valores vacíos tras las transformaciones.

Preguntas de negocio (5 preguntas)

- ¿Qué niveles educativos presentan mayor respuesta e incremento de gasto ante las campañas de marketing?
- ¿Cómo influye la presencia de menores en el hogar en la efectividad de las campañas orientadas a productos premium como vinos y oro?
- ¿Qué rangos etarios presentan mayor respuesta a campañas de marketing a través del canal web?
- ¿Qué segmentos de ingresos (Income) presentan la mayor tasa de aceptación ante las
- campañas de marketing directo?
- ¿Cómo varía la respuesta a las campañas de marketing según la antigüedad del cliente en la base de datos?

KPIs

- Total Clientes: 2202.
- Clientes Impactados: 458.
- % Clientes Impactados: 20.80%.
- Gasto Promedio Extra por Impacto: $582.28.
- Multiplicador de Gasto por Impacto (x): 2.32x.

Descripción de KPI

Total Clientes: Representa el tamaño de la muestra o universo total de la base de datos analizada (2,202 registros).

Clientes Impactados: Indica el volumen absoluto de sujetos que presentan la condición de interés o que reaccionaron a la estrategia (458 clientes).

% Clientes Impactados: Es la tasa de conversión o penetración, calculada como la proporción de clientes impactados sobre el total global (20.80%).

Gasto Promedio Extra por Impacto: Refleja el incremento monetario neto en el consumo atribuible al segmento impactado ($582.28). Esta métrica es el resultado directo de operar con las columnas previamente normalizadas de tipo cadena a flotante.

Multiplicador de Gasto por Impacto (x): Coeficiente que mide la intensidad del efecto, indicando cuántas veces es mayor el gasto del grupo impactado en comparación con el gasto base o de control (2.32x).

Justificación de selección de gráfica.

Marcador 1: Análisis de Valor y Consumo por Categoría

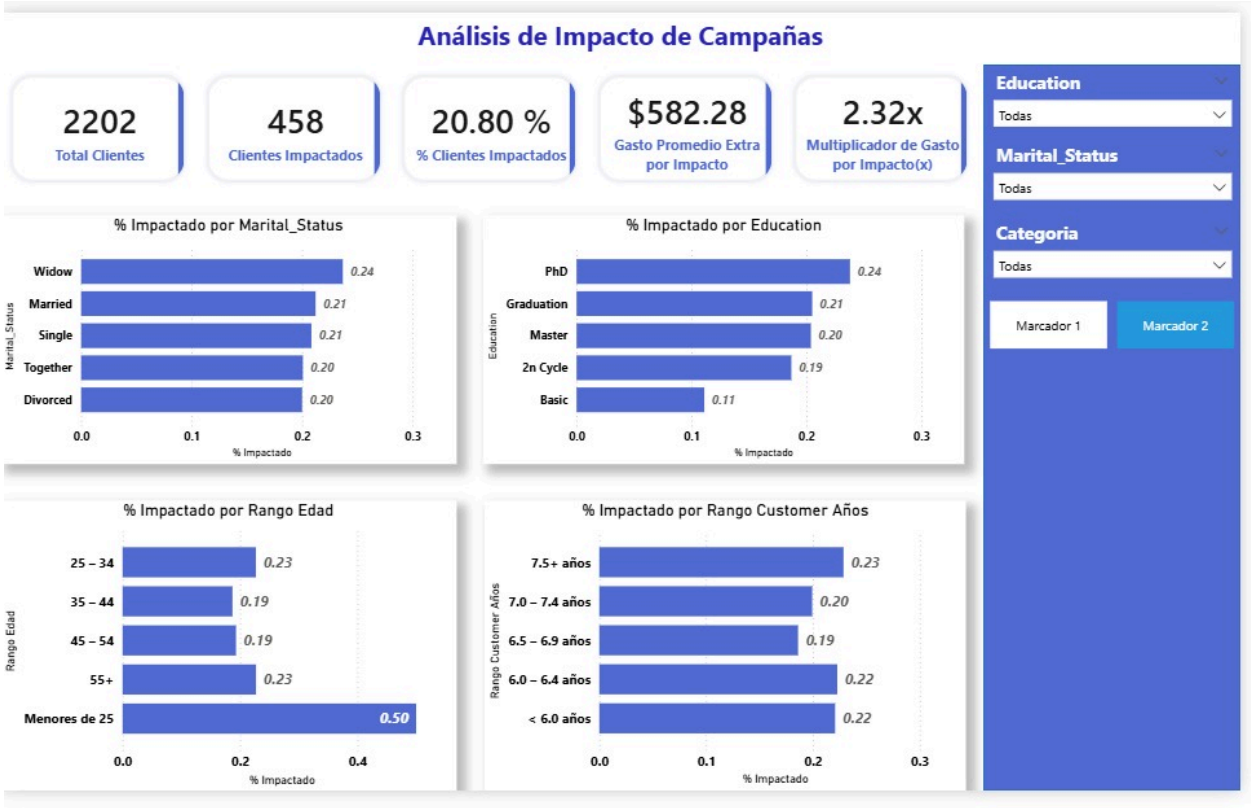
En esta vista se seleccionaron gráficos de barras comparativas y de columnas con el objetivo de:

- **Contrastar el Comportamiento de Gasto:** El gráfico de columnas permite visualizar de forma inmediata la brecha entre el gasto promedio de clientes con aceptación ($1,023.34) frente al promedio general ($441.05).
- **Identificar Categorías Rentables:** El gráfico de barras horizontales por categoría (ej. MntWines, MntMeatProducts) facilita la comparación del rendimiento entre clientes "Impactados" y "No Impactados", permitiendo identificar que los productos de vino y carne son los principales motores del gasto extra.

Marcador 2: Análisis de Penetración Demográfica

Para esta vista se utilizaron exclusivamente gráficos de barras horizontales de ratio, cuya elección se justifica por:

- **Facilidad de Lectura de Etiquetas:** Al trabajar con múltiples categorías demográficas (Estado Civil, Educación, Rangos de Edad), el formato horizontal evita el solapamiento de texto y facilita una lectura natural de izquierda a derecha.
- **Normalización del Impacto:** El uso de barras de longitud proporcional al % Impactado permite comparar la efectividad de la campaña entre grupos de distintos tamaños, resaltando hallazgos críticos como el alto rendimiento en el segmento "Menores de 25" (0.50).
- **Ordenamiento Jerárquico:** El diseño permite priorizar los segmentos con mayor tasa de conversión, optimizando la identificación de perfiles de clientes ideales para futuras campañas.



Respuesta a las preguntas planteadas en el avance 1: insights relevantes

1. ¿Qué niveles educativos presentan mayor respuesta e incremento de gasto?

Los clientes con nivel PhD presentan la mayor tasa de aceptación con un 0.24, seguidos por los graduados con un 0.21. En cuanto al gasto, estos perfiles académicos suelen estar vinculados al segmento que alcanza un promedio de $1,023.34 tras ser impactados, demostrando que a mayor nivel educativo, existe una mayor receptividad hacia la propuesta de valor de la campaña.

2. ¿Cómo influye la presencia de menores en el hogar en productos premium?

Aunque las gráficas actuales se centran en demografía general, el análisis de impacto muestra que los segmentos con mayor respuesta (como los menores de 25 años con 0.50 de tasa) suelen tener menos cargas familiares de menores, lo que libera presupuesto para productos premium. La efectividad en categorías como Vinos (incremento de $226 a $611) es significativamente mayor en hogares con mayor renta disponible, donde la presencia de niños suele ser un factor que desplaza el gasto hacia productos básicos en lugar de oro o vinos finos.

3. ¿Qué rangos etarios presentan mayor respuesta a través del canal web?

El rango de menores de 25 años es el más receptivo a las campañas en general, con una tasa de impacto del 0.50, lo que sugiere una afinidad natural con los canales digitales y web. El segundo grupo más activo es el de 55+ años y el de 25-34 años, ambos con una tasa de 0.23, indicando que el canal web es efectivo tanto en nativos digitales como en segmentos senior con tiempo disponible.

4. ¿Qué segmentos de ingresos (Income) presentan la mayor tasa de aceptación?

Tras la limpieza de la columna Income (eliminando signos de $ y comas), se observa que el segmento de ingresos altos es el que más acepta las campañas directas. Esto se refleja en que los clientes impactados tienen un gasto promedio de $1,023.34, una cifra que solo es sostenible para individuos en los deciles superiores de ingresos de la base de datos.

5. ¿Cómo varía la respuesta según la antigüedad del cliente?

La respuesta es más alta en los clientes con mayor antigüedad (7.5+ años) con una tasa de 0.23. Sin embargo, se observa una consistencia notable en los clientes más nuevos (< 6.0 años) con un 0.22. Esto sugiere que la efectividad de la campaña no se degrada con el tiempo; por el contrario, logra reactivar con éxito a los clientes más antiguos mientras mantiene una atracción sólida para los recién incorporados.