Entregable I

Proyecto Data Visualization con Power BI

Proyecto Final Borrador

Integrantes Grupo 1:

Anzules Fuentes Abraham Joel

Baño Cordero Christell Nicole

Cevallos Cobos Luis Fernando

Morante Murillo Madelayne Betsabeth

Tulcan Alvarez José David

Vergara Villafuerte Nagelly Dayanna

Data Visualization

MSc. Martha Tomalá

Data Foundations Program

MINTEL - ESPOL

Curso CBMP4

21 de ene. de 26

Proyecto: Data Visualization with Power BI

Entregable I – Exploración y Preprocesamiento de Datos (Python)

Objetivo:
El objetivo de este entregable es comprender, explorar y preparar un dataset de marketing para su posterior análisis y visualización en Power BI. El estudiante deberá aplicar criterios de análisis exploratorio, limpieza y transformación de datos utilizando Python.

1. Diccionario de Datos (Columnas Originales)

A continuación, se presenta el diccionario de datos con las columnas tal como se encuentran en el archivo original entregado:

Columna	Tipo esperado	Descripción
ID	Numérico	Identificador único del cliente
Income	Texto	Ingreso anual del cliente, contiene símbolo $
Age	Numérico	Edad del cliente
Customer_Days	Numérico	Antigüedad del cliente en días
Kidhome	Numérico	Número de niños en el hogar
Teenhome	Numérico	Número de adolescentes en el hogar
MntWines	Texto	Monto gastado en vinos
MntFruits	Texto	Monto gastado en frutas
MntMeatProducts	Texto	Monto gastado en productos cárnicos
MntFishProducts	Texto	Monto gastado en pescados
MntSweetProducts	Texto	Monto gastado en dulces
MntGoldProds	Texto	Monto gastado en productos premium
MntTotal	Texto	Gasto total del cliente

NumWebPurchases	Numérico	Número de compras realizadas por web
NumCatalogPurchases	Numérico	Número de compras por catálogo
NumStorePurchases	Numérico	Número de compras en tienda
NumDealsPurchases	Numérico	Compras realizadas con descuento
NumWebVisitsMonth	Numérico	Visitas mensuales al sitio web
marital_Single	Binaria	Cliente soltero
marital_Married	Binaria	Cliente casado
marital_Together	Binaria	Cliente en unión
marital_Divorced	Binaria	Cliente divorciado
marital_Widow	Binaria	Cliente viudo
education_Basic	Binaria	Nivel educativo básico
education_Graduation	Binaria	Nivel educativo universitario
education_Master	Binaria	Nivel educativo maestría
education_PhD	Binaria	Nivel educativo doctorado
AcceptedCmpOverall	Binaria	Aceptó al menos una campaña
Response	Binaria	Respuesta positiva a la última campaña

Actividades del Entregable

PARTE 1: Notebook

A. Análisis Exploratorio de Datos (EDA)

- Explorar la estructura del dataset.
- Identificar outliers evidentes y posibles problemas de calidad.
- Presentar observaciones relevantes.

B. Preprocesamiento de Datos

- En base a las observaciones anteriores, realice el preprocesamiento de datos para mejorar la calidad del mismo.

C. Reconstrucción de variables categóricas:

- Crear la columna Marital_Status a partir de marital_*.
- Crear la columna Education a partir de education_*.
- Eliminar las columnas originales utilizadas.

D. Tratamiento de valores inválidos:

- Identificar edades con valor 99999.
- Proponer y justificar una estrategia de tratamiento.

E. Validaciones finales:

- Confirmar los tipos de datos correctos.
- Verificar coherencia general del dataset.

PARTE 2: Informe analítico

- Los grupos están seleccionados a uno de los siguientes enfoques:
  - Perfil del cliente
  - Comportamiento de compra

- Campañas y respuesta

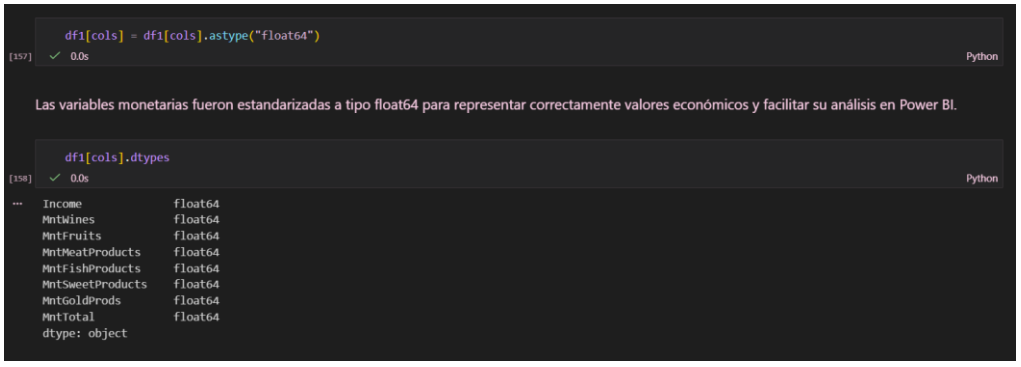
- Para el enfoque seleccionado (Perfil del cliente) debe:
  - Proponer al menos 5 preguntas de negocio. *(En Anexos hay ejemplos. puede utilizar hasta 3 preguntas de las recomendadas, y debe proponer al menos otras 2)*
  - Definir KPIs que permitan responder dichas preguntas.

Resumen del estado actual del dataset y mejoras realizadas

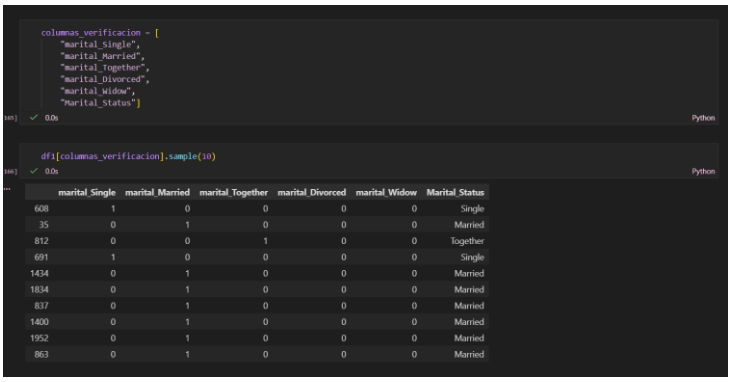
El dataset analizado corresponde a información de clientes de una empresa de marketing, incluyendo variables demográficas, comportamiento de compra y respuestas a campañas. En su estado original, el conjunto de datos presentaba inconsistencias en los tipos de datos, especialmente en las variables monetarias, así como una estructura categórica fragmentada en múltiples columnas binarias.

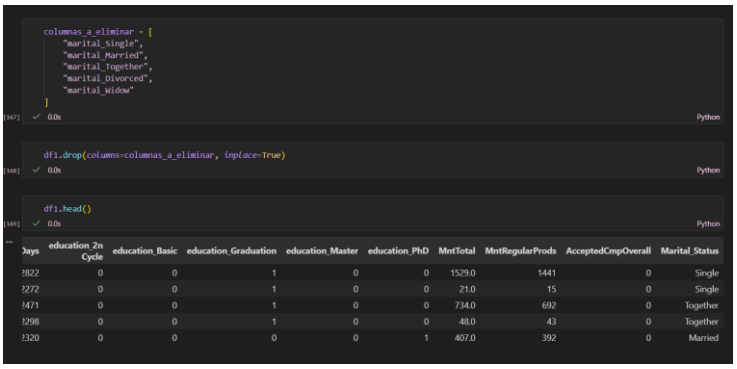
Durante la fase de exploración y preprocesamiento, se realizaron las siguientes mejoras de calidad:

- Limpieza y estandarización de variables monetarias (Income y Mnt*), eliminando símbolos de moneda y separadores de miles, y convirtiéndolas a formato numérico.



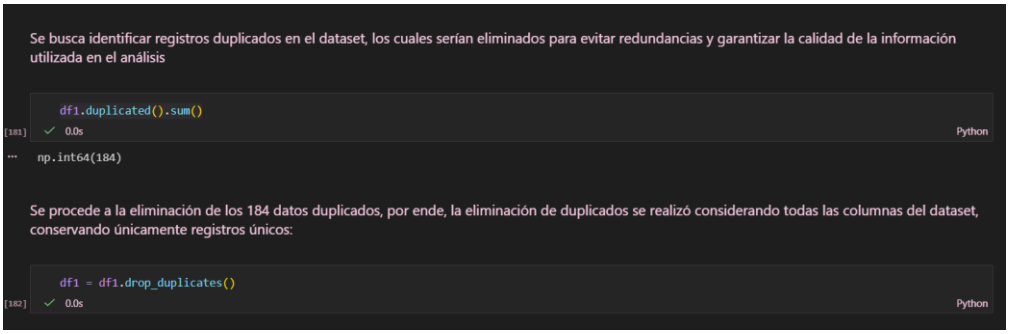
- Reconstrucción de variables categóricas, consolidando las columnas binarias de estado civil y nivel educativo en las variables Marital_Status y Education, respectivamente.



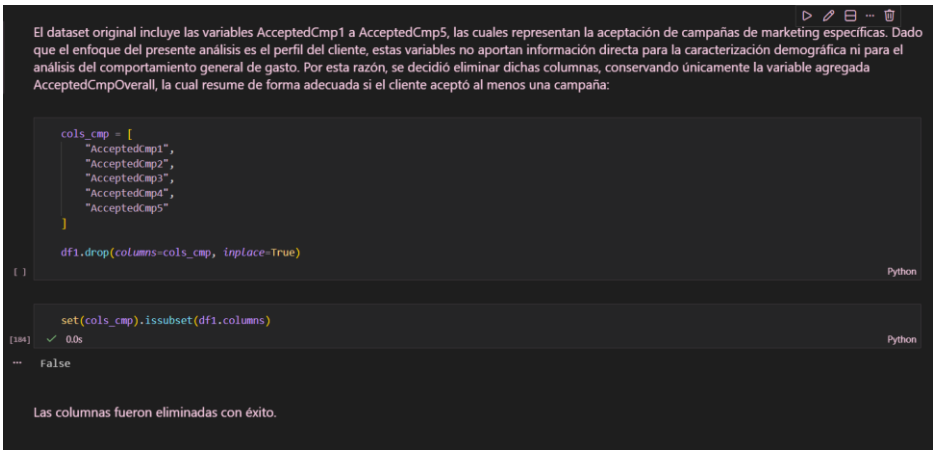


El mismo procedimiento se realizó con las columnas de Education.

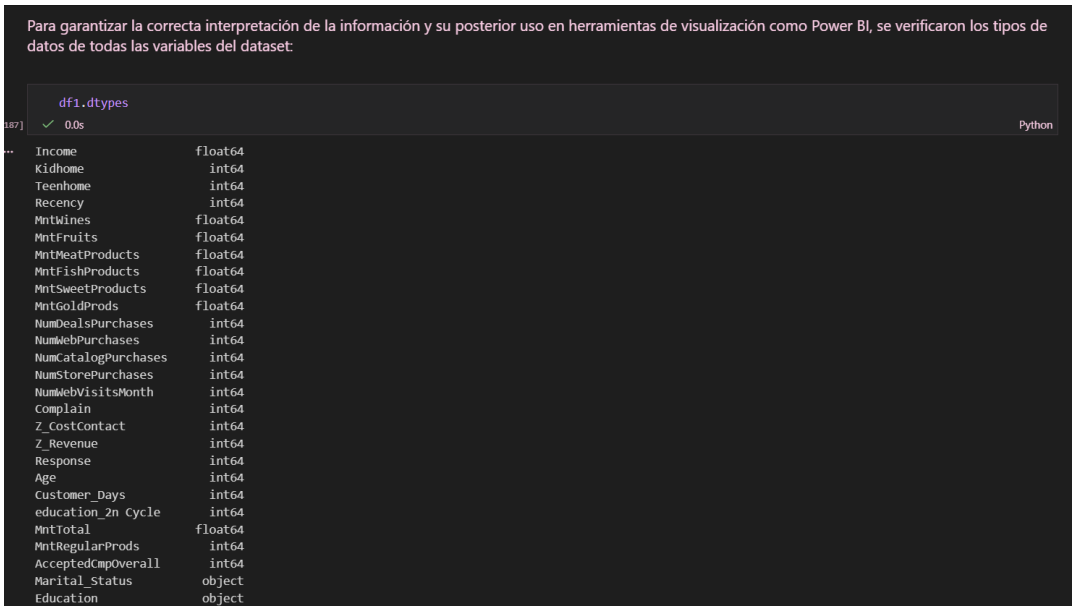
- Eliminación de columnas redundantes utilizadas en la reconstrucción categórica, mejorando la legibilidad y simplicidad del dataset.
- Identificación y eliminación de registros duplicados para evitar redundancias y garantizar la calidad de la información.

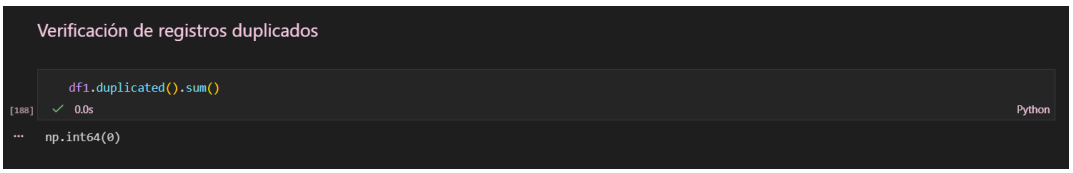


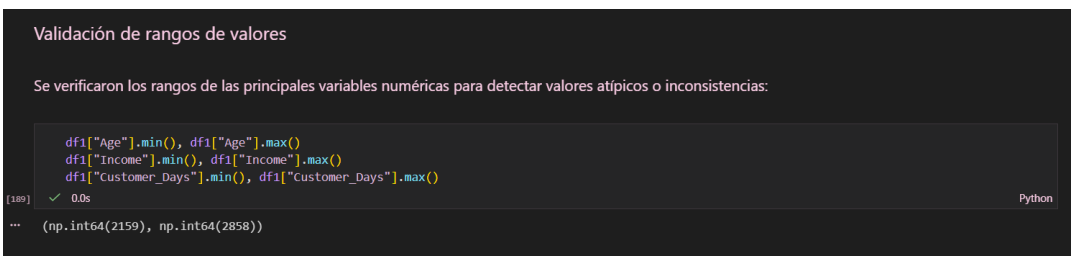
- Eliminación justificada de variables relacionadas con campañas individuales (AcceptedCmp1 a AcceptedCmp5), dado que el enfoque del análisis se centra en el perfil del cliente y no en la evaluación detallada de campañas.



- Validación final de rangos, valores nulos y tipos de datos, confirmando la coherencia general del dataset.



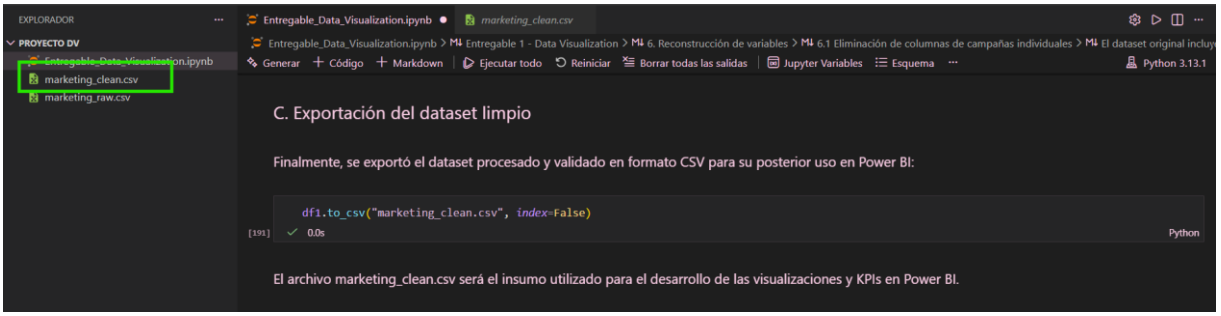




El resultado de la validación de rangos de valores representa antigüedad de clientes (≈ 6 a 8

años).

- Como resultado, se obtuvo un dataset limpio y estructurado, exportado como marketing_clean.csv, listo para su análisis y visualización en Power BI.



Enfoque analítico: Perfil del Cliente

El enfoque de *Perfil del Cliente* tiene como propósito analizar cómo las características demográficas y sociales influyen en el comportamiento de gasto de los clientes. Este análisis permite identificar segmentos de mayor valor y apoyar la toma de decisiones estratégicas relacionadas con marketing y fidelización.

Preguntas de negocio:

A partir del enfoque seleccionado, se plantearon las siguientes preguntas de negocio:

- ¿Qué nivel educativo presenta el mayor gasto promedio por cliente?
- ¿Qué estado civil concentra a los clientes con mayor gasto promedio?
- ¿Los clientes con hijos gastan más o menos que los clientes sin hijos?
- ¿Cómo varía el gasto promedio según los rangos de edad de los clientes?
- ¿Qué proporción de clientes corresponde a clientes de alto valor, definidos como aquellos que pertenecen al top 10% de gasto total?

KPIs definidos:

Para responder a las preguntas de negocio planteadas, se definieron los siguientes indicadores clave de desempeño (KPIs):

- *Gasto total:* Suma del gasto total (MntTotal) de todos los clientes.

- *Gasto promedio por cliente:* Promedio del gasto total por cliente.
- *Número total de clientes:* Conteo distinto de clientes.
- *Gasto promedio por nivel educativo:* Promedio de MntTotal segmentado por Education.
- *Gasto promedio por estado civil:* Promedio de MntTotal segmentado por Marital_Status.
- *Gasto promedio con hijos vs sin hijos:* Comparación del gasto promedio según la presencia de hijos en el hogar.
- *Clientes de alto valor:* Número y proporción de clientes pertenecientes al top 10% de gasto total.

Estos KPIs fueron implementados en Power BI mediante medidas DAX y visualizaciones interactivas.

Desarrollo en Power BI alineado a las preguntas de negocio y KPIs.

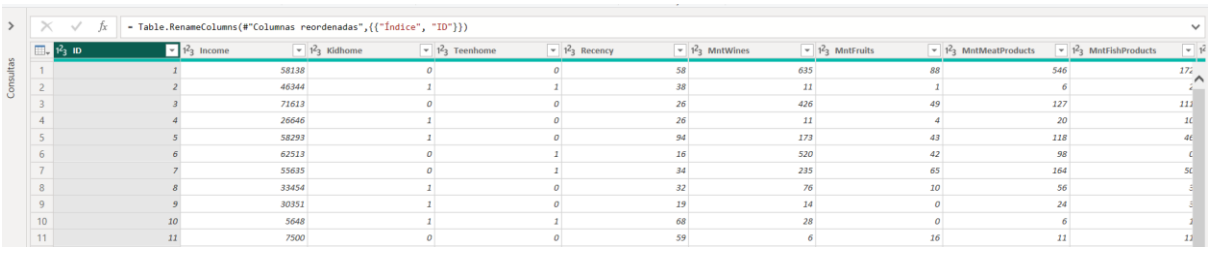
Formato visual de moneda

Se aplica formato *Moneda ($)* a *Income* y a todas las columnas *Mnt*;* aunque el dato se mantiene como float, se muestra al usuario en dólares para mejorar la interpretación y asegurar que el dashboard comunique valores económicos de forma inmediata.

Creación de una columna identificadora

Se crea un identificador único cuando el dataset no lo incluye explícitamente, para poder contar clientes sin duplicar registros y construir KPIs basados en clientes únicos.

Se lo creo yendo a Transformar datos → Selección de tabla → Agregar columna → Índice → Desde 1 → Renombrada a *ID.*



Creación de columnas derivadas para segmentación del perfil

El dataset contiene dos variables relacionadas con hijos (Kidhome y Teenhome), pero la pregunta de negocio requiere comparar dos segmentos claros: clientes con hijos vs sin hijos. Esta columna convierte el detalle numérico en una categoría de análisis, lo cual simplifica el filtrado y permite visualizaciones directas para comparar gasto promedio entre ambos grupos:

```
1 Has_Kids =
2 IF (
3     'marketing_clean'[Kidhome] + 'marketing_clean'[Teenhome] > 0,
4     "Con hijos",
5     "Sin hijos"
6 )
```

Analizar edad como valores individuales genera demasiadas categorías y dificulta la lectura.

Los rangos permiten resumir el comportamiento por grupos etarios, haciendo que las comparaciones sean más interpretables y útiles para definir perfiles predominantes y segmentos de valor:

```
1 Age_Group =
2 SWITCH(
3     TRUE(),
4     'marketing_clean'[Age] < 30, "18-29",
5     'marketing_clean'[Age] < 40, "30-39",
6     'marketing_clean'[Age] < 50, "40-49",
7     'marketing_clean'[Age] < 60, "50-59",
8     "60+"
9 )
```

Creación de medidas DAX

El KPI *Total de clientes* permite dimensionar el tamaño de la base de análisis y sirve como denominador para proporciones. Se usa DISTINCTCOUNT para evitar duplicidades y contar clientes únicos:

```
1 Total Customers =
2 DISTINCTCOUNT ( 'marketing_clean'[ID] )
```

La medida *Gasto Total* representa el volumen económico total generado por los clientes. Es esencial para comprender el valor general de la cartera:

```
1 Total Spend =
2 SUM ( 'marketing_clean'[MntTotal] )
```

La medida *Gasto promedio por cliente* permite comparar segmentos con independencia del tamaño del grupo. Es un KPI clave para evaluar qué perfiles tienden a gastar más:

```
1 Average Spend per Customer =
2 AVERAGE ( 'marketing_clean'[MntTotal] )
```

El KPI de clientes de alto valor (Top 10% por gasto), *Umbral del top 10%,* sirve para definir un punto de corte para clasificar a los clientes de mayor valor con un criterio estadístico objetivo. El percentil 90 permite identificar el grupo superior de clientes según gasto:

```
1 High Value Threshold =
2 PERCENTILEX.INC(
3     ALL ( 'marketing_clean' ),
4     'marketing_clean'[MntTotal],
5     0.9
6 )
```

La medida *Número de clientes de alto valor* cuantifica cuántos clientes pertenecen al segmento de mayor gasto. Esto ayuda a dimensionar el grupo prioritario para estrategias de retención o fidelización:

```
1 High Value Customers =
2 CALCULATE(
3     DISTINCTCOUNT ( 'marketing_clean'[ID] ),
4     FILTER(
5         ALL ( 'marketing_clean' ),
6         'marketing_clean'[MntTotal] >= [High Value Threshold]
7     )
8 )
```

El indicador *Porcentaje de clientes de alto valor* expresa el tamaño relativo del segmento de alto valor y permite interpretarlo en términos proporcionales, no solo absolutos:

```
tructura  Formato
1 High Value Customer Percentage =
2 DIVIDE(
3     [High Value Customers],
4     [Total Customers],
5     0
6 )
```

Conclusiones

El desarrollo realizado permitió establecer una base sólida para el análisis del perfil del cliente, integrando de manera coherente las características demográficas y sociales con el comportamiento de gasto. A través del preprocesamiento de datos en Python y la posterior estructuración del modelo en Power BI, se logró transformar un conjunto de datos inicial con inconsistencias en un dataset limpio, organizado y analíticamente funcional.

La creación de variables derivadas, como los rangos de edad y la clasificación de clientes con hijos y sin hijos, facilitó la segmentación del conjunto de datos y permitió realizar comparaciones claras entre distintos perfiles de clientes. Estas transformaciones resultaron fundamentales para responder de forma directa a las preguntas de negocio planteadas, enfocadas en identificar qué características influyen en un mayor gasto promedio.

Asimismo, la definición e implementación de indicadores clave de desempeño (KPIs), tales como el gasto total, el gasto promedio por cliente, el número total de clientes y la identificación de clientes de alto valor pertenecientes al top 10% de gasto, proporcionó métricas objetivas y comparables para el análisis del comportamiento del cliente. Estos KPIs permiten evaluar el valor económico de los distintos segmentos y sirven como insumo para la toma de decisiones estratégicas relacionadas con marketing y fidelización.

En conjunto, el trabajo desarrollado sienta las bases para la realización futura de un dashboard interactivo en Power BI, el cual permitirá visualizar de forma dinámica los resultados obtenidos, explorar los datos mediante filtros y segmentadores, y profundizar en el análisis del perfil del cliente.

ANEXOS:

Enfoque: Perfil de cliente

Propósito: Analizar cómo las características demográficas influyen en el comportamiento de compra.

Preguntas guía (ejemplos)

1. ¿Qué nivel educativo presenta mayor gasto total?
2. ¿Los clientes con hijos gastan más o menos?
3. ¿La antigüedad del cliente impacta en el gasto?
4. ¿Qué perfil concentra los clientes de mayor valor?

KPIs sugeridos

- Gasto promedio por cliente
- Ticket promedio
- % de clientes por nivel educativo
- Gasto promedio por estado civil

Enfoque: Comportamiento de compra

Propósito: Entender cómo compran los clientes y a través de qué canales.

Preguntas guía (ejemplos)

1. ¿Qué canal genera mayor volumen de compras?
2. ¿Qué productos concentran mayor gasto?
3. ¿Los clientes que visitan más la web compran más?
4. ¿Qué canal está asociado a mayor ticket promedio?

KPIs sugeridos

- Gasto promedio por canal
- Distribución del gasto por categoría
- Frecuencia promedio de compra
- Ticket promedio por canal

Enfoque: Campañas y respuestas

Propósito: Evaluar la efectividad de campañas y el perfil del cliente que responde.

Preguntas guía (ejemplos)

1. ¿Qué porcentaje de clientes acepta campañas?
2. ¿Qué perfil responde mejor a campañas?
3. ¿Los clientes que aceptan campañas gastan más?
4. ¿Qué categorías aumentan tras campañas?

KPIs sugeridos

- Tasa de aceptación de campañas
- Gasto promedio de clientes que aceptan vs no aceptan
- % de clientes impactados
- Diferencia de gasto pre / post campaña (si aplica)