# Contexto del problema de negocio

## 1.1 Descripción del sector

La empresa se encuentra operando en la industria bancaria, lo cual representa una parte fundamental en la economía de cada país. Donde la dinámica del negocio consiste en captar recursos, mediante créditos para posteriormente realizar inversiones a corto o largo plazo. Actuando como punto de encuentro entre el ahorro y la inversión. La base de datos del presente proyecto se encuentra actualizada como fecha última en Abril del 2024.

## 1.2 Problema

En la actualidad la empresa presenta diferentes retos para conocer el comportamiento de compra de las tarjetas de crédito que ofrece. Desconociendo en muchas cosas las preferencias que tienden a optar un cliente de un producto, los cuáles existe limitación en estrategias agresivas para atender esa demanda que viene estancada durante los últimos meses. El problema principal es analizar con exactitud los requerimientos de los clientes para evaluar si existe alguna estacionalidad o tendencia durante el tiempo.

Es importante definir estrategias para que los clientes prefieran los productos que se cuenta, de acuerdo con la edad, ingreso, categoría, frecuencia de compra, entre otros.

# Descripción del conjunto de datos

En el presente proyecto se utilizó un conjunto de datos de Kaggle, una plataforma en línea de ciencia de datos que ofrece acceso a una amplia variedad de conjuntos públicos para investigación y desarrollo.

Aquí la ruta de la base de datos pública:

[**Customer Purchases Behaviour Dataset (kaggle.com)**](https://www.kaggle.com/datasets/sanyamgoyal401/customer-purchases-behaviour-dataset/data)

## 2.1 Resúmenes de Campos

* + 1. **ID**

Identificador único otorgado a cada cliente, el tipo de dato es String.

* + 1. **AGE**

Edad los clientes compradores, el tipo de dato es Integer.

* + 1. **GENDER**

Género de los clientes compradores, el tipo de dato es String.

* + 1. **INCOME**

Ingreso correspondiente a cada cliente, el tipo de dato es Integer.

* + 1. **EDUCATION**

Nivel de educación de los clientes, el tipo de dato es String.

* + 1. **REGIÓN**

Ubicación o territorio de los clientes, el tipo de dato es String.

* + 1. **LOYALTY STATUS**

Estado de lealtad de compra de los clientes, el tipo de dato es String.

* + 1. **PURCHASE FREQUENCY**

Frecuencia de compra de los clientes, el tipo de dato es String.

* + 1. **PURCHASE AMOUNT**

Monto de compra acumulada de los clientes, el tipo de dato es Integer.

* + 1. **PRODUCT CATEGORY**

Categoría del producto comprado por los clientes,el tipo de dato es String.

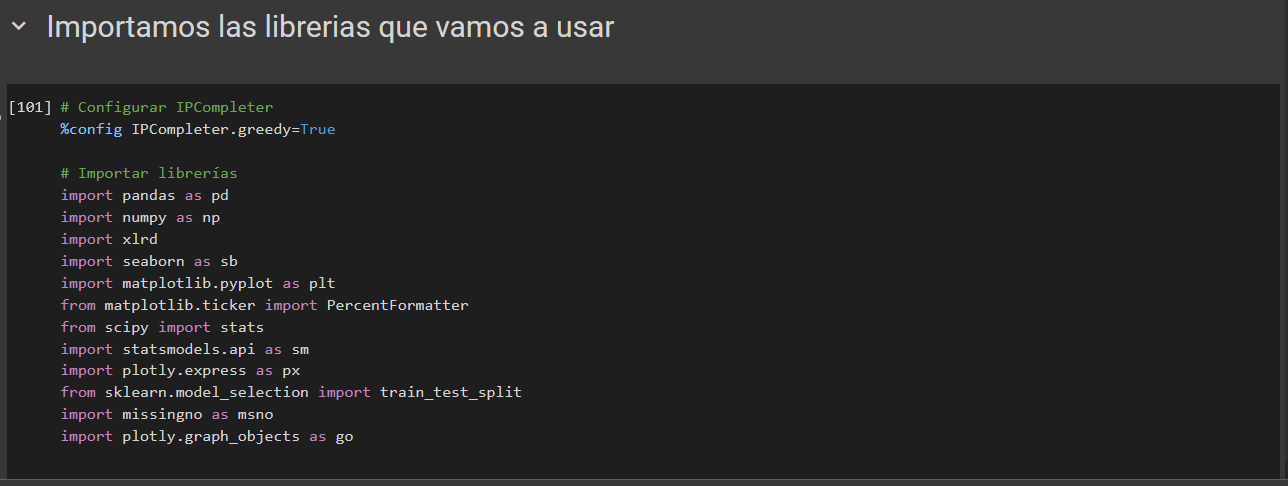
* + 1. **PROMOTION\_USAGE**

Identificación de uso de promociones de los clientes, el tipo de datos es Integer.

* + 1. **SATISFACTION SCORE**

Puntaje de satisfacción de los clientes, el tipo de compra es Integer.

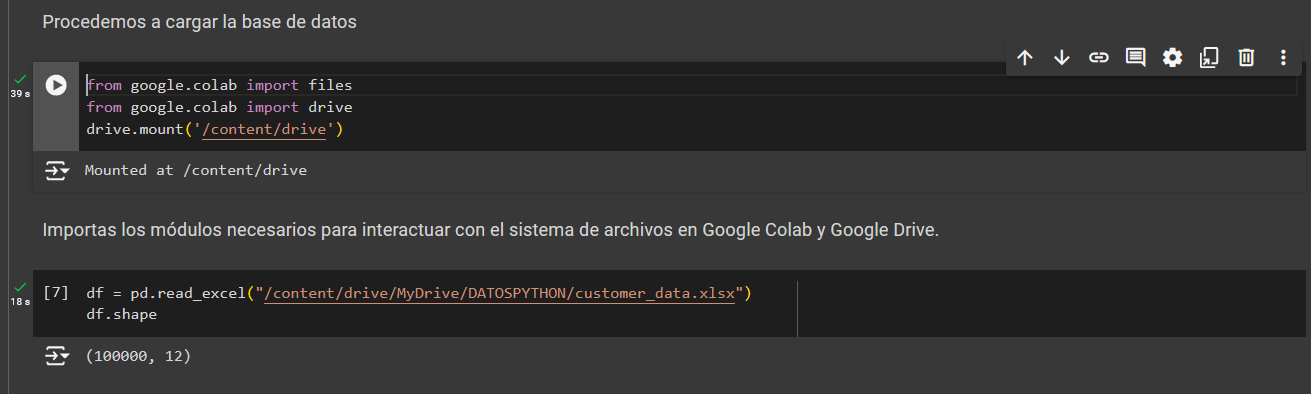
# 3 Diagnóstico e Insights

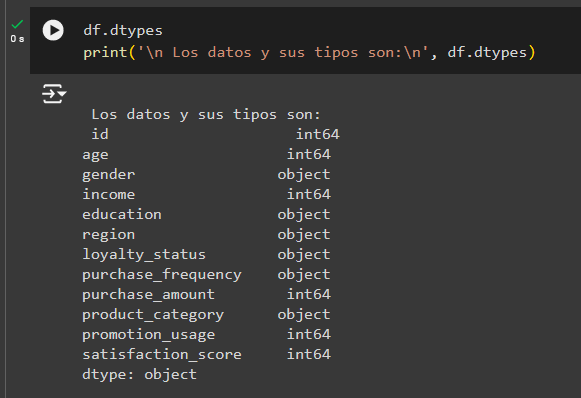


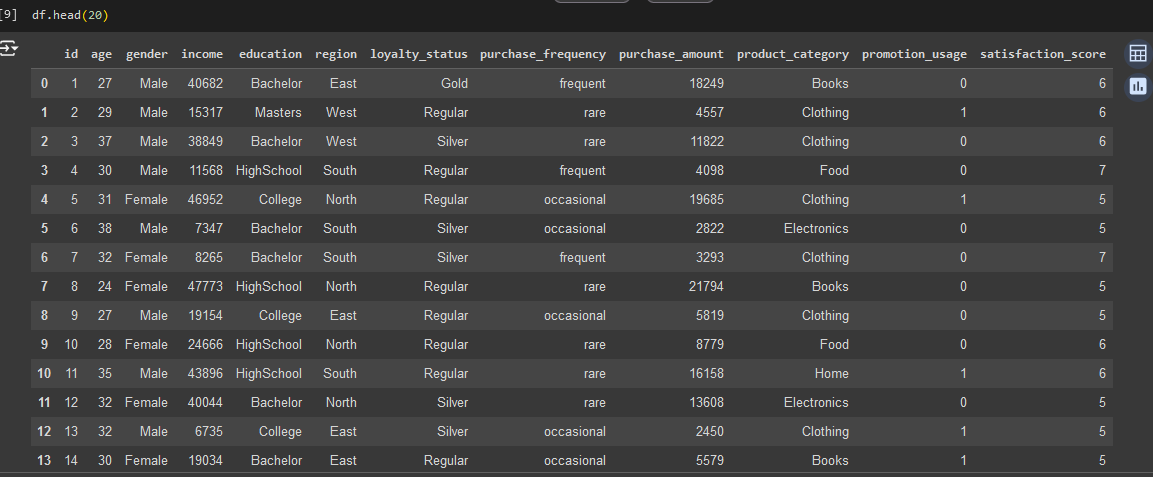
* **pandas** es esencial para la manipulación y análisis de datos tabulares, permitiendo la carga y visualización de datos desde un archivo Excel
* **numpy** proporciona soporte para arreglos y matrices de gran tamaño, facilitando las operaciones matemáticas y lógicas.
* **xlrd** se utiliza para leer datos de archivos Excel, funcionando como backend para pandas
* Para la visualización de datos, **seaborn y matplotlib** son herramientas cruciales; seaborn crea gráficos estadísticos atractivos y matplotlib permite crear una amplia variedad de gráficos en 2D
* **PercentFormatter de matplotlib** ayuda a formatear los ejes en porcentajes
* **scipy** ofrece una colección de algoritmos y funciones matemáticas y estadísticas
* **statsmodels** facilita la estimación de modelos estadísticos y pruebas.
* Para gráficos interactivos, **plotly.express** es ideal, permitiendo visualizaciones dinámicas y exploratorias.
* Para realizar modelado predictivo, **scikit-lear**n es esencial; esta librería proporciona herramientas para dividir los datos en conjuntos de entrenamiento y prueba y realizar tareas de machine learning
* **missingno** es útil para identificar y entender patrones en los datos faltantes mediante visualizaciones específicas

La ruta proporcionada ([/content/drive/MyDrive/DATOSPYTHON/customer\_data.xlsx](https://colab.research.google.com/drive/1Rhk0QRXdP2pdtFuULMqegnM-LiOBQ2r5)) es la ubicación del archivo Excel en Google Drive. Una vez leído, los datos se almacenan en la **variable df** como un DataFrame de pandas, que permite su manipulación y análisis.

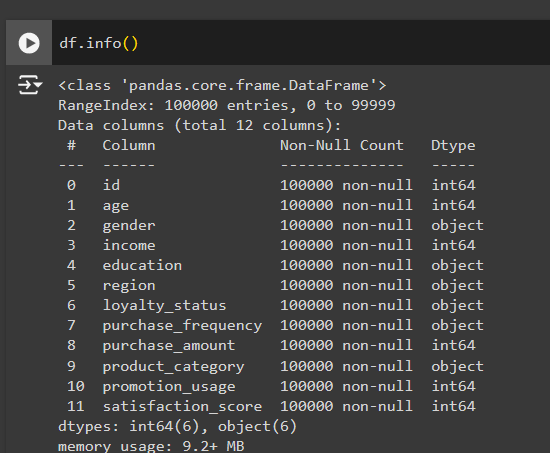
df.shape -> este atributo devuelve una tupla que indica el número de filas y columnas en el DataFrame ((número\_de\_filas, número\_de\_columnas))

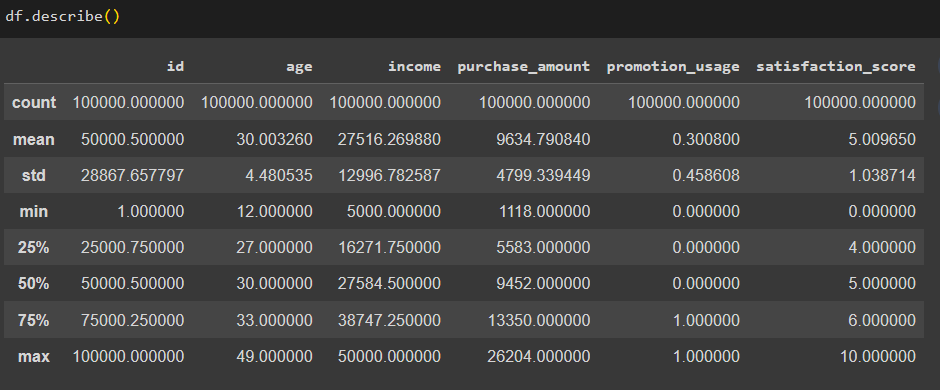


Utilizamos df.dtypes para obtener los tipos de datos de cada columna y luego imprimimos esta información junto con un mensaje explicativo. También utilizamos la función head() para verificar que los datos pueden ser visualizados.



Por últimos el utilizamos la funcion info() y el describe() ya que proporciona un resumen conciso del DataFrame, incluyendo el índice, el número de entradas, el tipo de datos de cada columna, y la cantidad de valores no nulos en cada columna. Es especialmente útil para obtener una visión general de la estructura del DataFrame, detectar datos faltantes y verificar los tipos de datos de las columnas





La lectura del df.describe() es la siguiente:  
count:

* Significado: Número de valores no nulos en cada columna.
* Interpretación: Cada columna tiene 100,000 entradas válidas, lo que indica que no hay valores nulos.

mean:

* Significado: Media aritmética de los valores en cada columna.
* Interpretación:
  + id: La media del identificador es 50,000.5, lo que tiene sentido en un rango de 1 a 100,000.
  + age: La edad media es 30.003.
  + income: El ingreso promedio es 27,516.27.
  + purchase\_amount: El monto promedio de compra es 9,634.79.
  + promotion\_usage: La media de uso de promociones es 0.301, indicando que aproximadamente el 30.1% de los registros usan promociones.
  + satisfaction\_score: La puntuación de satisfacción media es 5.01.

std (desviación estándar):

* Significado: Mide la dispersión de los valores con respecto a la media.
* Interpretación:
  + id: La desviación estándar de 28,867.66 indica una amplia dispersión en los valores del identificador.
  + age: La desviación estándar de 4.48 años muestra variabilidad en las edades.
  + income: La desviación estándar de 12,996.78 indica una gran variabilidad en los ingresos.
  + purchase\_amount: La desviación estándar de 4,799.34 muestra variabilidad en los montos de compra.
  + promotion\_usage: La desviación estándar de 0.459 indica variabilidad en el uso de promociones.
  + satisfaction\_score: La desviación estándar de 1.039 muestra cierta variabilidad en las puntuaciones de satisfacción.

min (valor mínimo):

* Significado: El valor mínimo en cada columna.
* Interpretación:
  + id: El valor mínimo del identificador es 1.
  + age: La edad mínima es 12 años.
  + income: El ingreso mínimo es 5,000.
  + purchase\_amount: El monto mínimo de compra es 1,118.
  + promotion\_usage: El valor mínimo es 0, indicando que algunos registros no usan promociones.
  + satisfaction\_score: La puntuación de satisfacción mínima es 0.

25% (percentil 25):

* Significado: El valor debajo del cual cae el 25% de los datos.
* Interpretación:
  + id: El 25% de los identificadores están por debajo de 25,000.75.
  + age: El 25% de las edades están por debajo de 27 años.
  + income: El 25% de los ingresos están por debajo de 16,271.75.
  + purchase\_amount: El 25% de los montos de compra están por debajo de 5,583.
  + promotion\_usage: El 25% de los registros no usan promociones.
  + satisfaction\_score: El 25% de las puntuaciones de satisfacción están por debajo de 4.

50% (percentil 50 o mediana):

* Significado: El valor debajo del cual cae el 50% de los datos (mediana).
* Interpretación:
  + id: El 50% de los identificadores están por debajo de 50,000.5.
  + age: El 50% de las edades están por debajo de 30 años.
  + income: El 50% de los ingresos están por debajo de 27,584.5.
  + purchase\_amount: El 50% de los montos de compra están por debajo de 9,452.
  + promotion\_usage: El 50% de los registros no usan promociones.
  + satisfaction\_score: El 50% de las puntuaciones de satisfacción están por debajo de 5.

75% (percentil 75):

* Significado: El valor debajo del cual cae el 75% de los datos.
* Interpretación:
  + id: El 75% de los identificadores están por debajo de 75,000.25.
  + age: El 75% de las edades están por debajo de 33 años.
  + income: El 75% de los ingresos están por debajo de 38,747.25.
  + purchase\_amount: El 75% de los montos de compra están por debajo de 13,350.
  + promotion\_usage: El 75% de los registros usan promociones.
  + satisfaction\_score: El 75% de las puntuaciones de satisfacción están por debajo de 6.

max (valor máximo):

* Significado: El valor máximo en cada columna.
* Interpretación:
  + id: El valor máximo del identificador es 100,000.
  + age: La edad máxima es 49 años.
  + income: El ingreso máximo es 50,000.
  + purchase\_amount: El monto máximo de compra es 26,204.
  + promotion\_usage: El valor máximo es 1, indicando que algunos registros usan promociones.
  + satisfaction\_score: La puntuación de satisfacción máxima es 10.

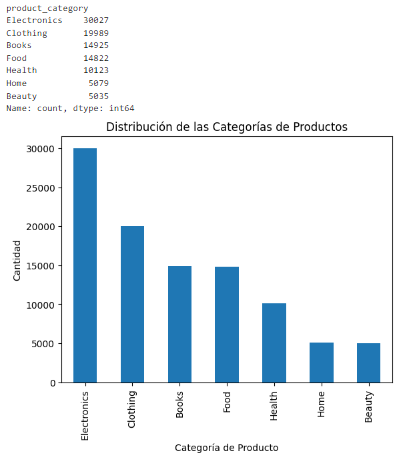
# 4. Manipulación de datos

**4.1 Identificando la distribución de la categoría de productos**

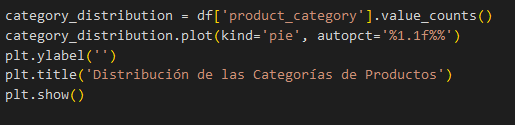
**Forma 1**

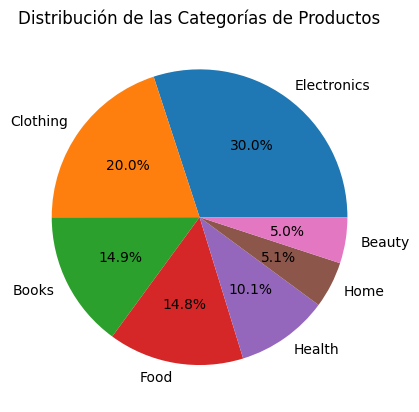
Se pudo reconocer que categorías de productos son las más preferidas en la compra de los clientes, siendo las de mayor porcentaje la categoría de electrodomésticos y ropa. Donde se tendrá que poner más énfasis en estos 2 últimos.





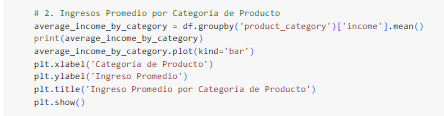
**Forma 2**

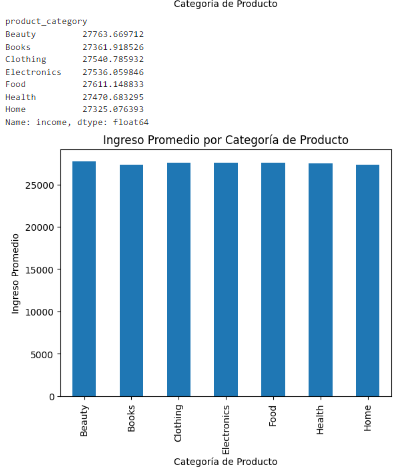
****

****

**4.2 Identificando los ingresos promedio por categoría del producto**

Con relación a los ingresos promedio, no existe una diferencia marcada respecto a la categoría de productos. La categoría de “belleza” ligeramente está por encima de las demás.

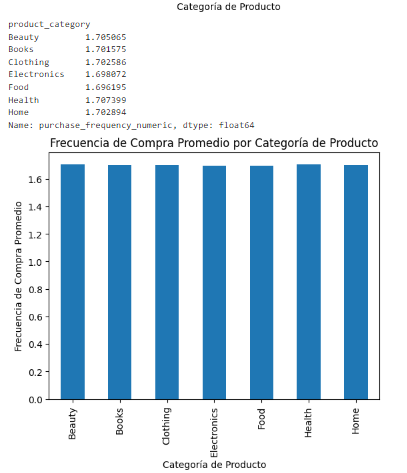




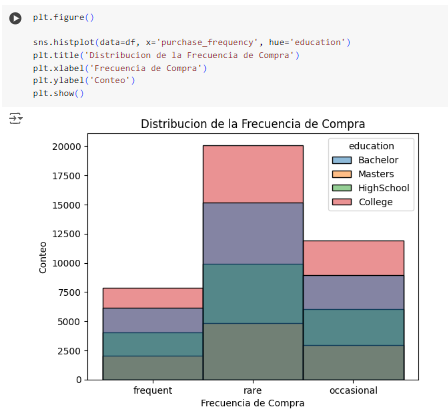
**4.3 Identificando la frecuencia de compra**

Se pudo realizar un análisis completo de la frecuencia de compra de los clientes. Aprovechando la máxima información con la que se dispone de la data:

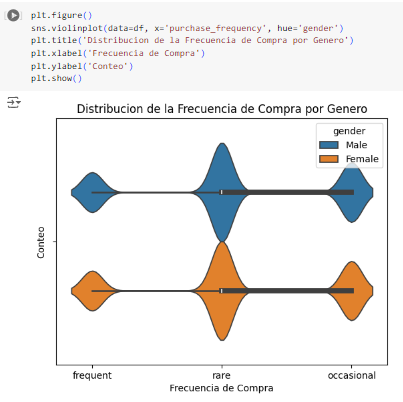




Se pudo identificar que la frecuencia de compra “raro” contempla un monto alto, con una mayor representación los clientes con nivel de educación de “Maestros”, seguidos por los clientes con un nivel de educación “Bachiller”. Dicha información resulta valiosa a la hora de tomar alguna decisión, y entender el porcentaje del nivel de educación de los clientes que compran frecuentemente.

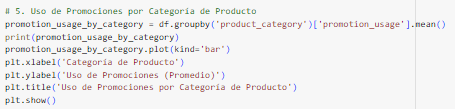


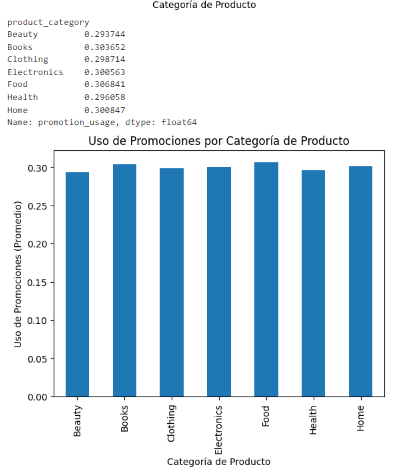
Seguidamente se pudo identificar la distribución en la frecuencia de compra de los clientes por género, esto resulta importante porque te da un panorama claro de los clientes más activos y realizan compras con una mayor frecuencia.



**4.4 Identificando la cantidad de promociones por categoría de producto**

Se identificó a su vez las promociones utilizadas por categoría de producto, para así comprender en qué productos los clientes utilizan una mayor promoción a la hora de tomar una decisión. Donde podemos notar que la categoría de “Libros” es la que mayor cantidad de promociones se utilizan.

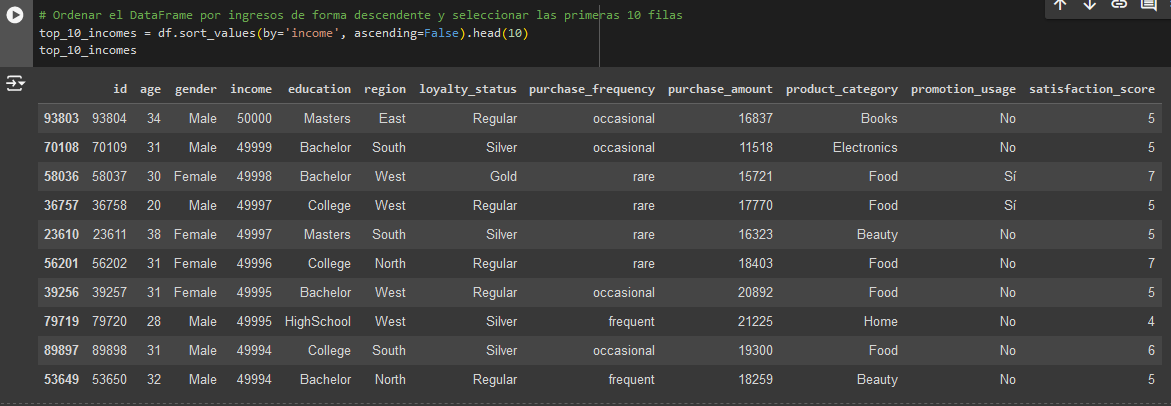




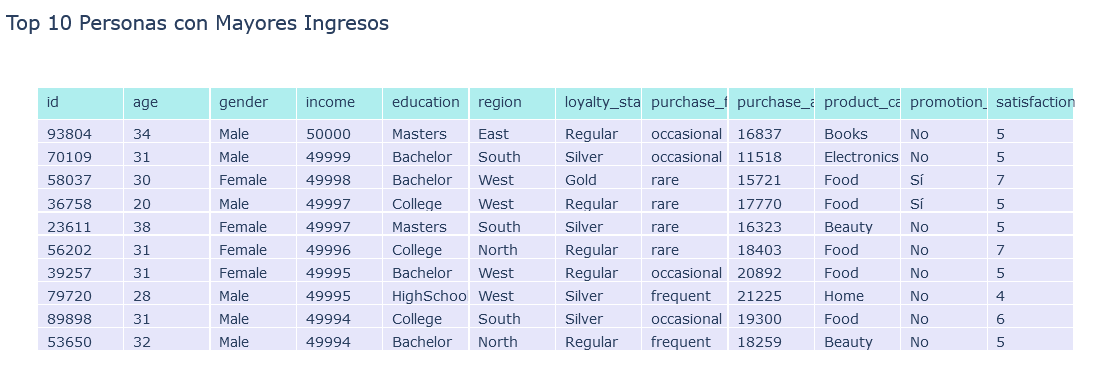
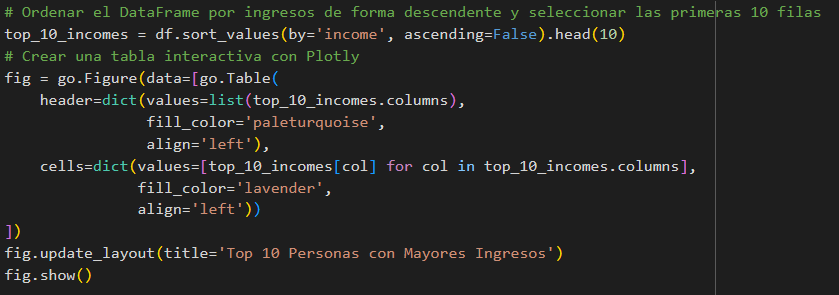
**4.5 Identificar a los 10 clientes con mayores ingresos**

Identificar a los clientes con los mayores ingresos puede ayudar a comprender mejor a este segmento de clientes. Conocer sus características demográficas y patrones de compra permite diseñar estrategias de fidelización específicas, ofertas personalizadas y campañas de marketing dirigidas, maximizando así las oportunidades de ventas y la satisfacción del cliente.

**Forma 1**

****

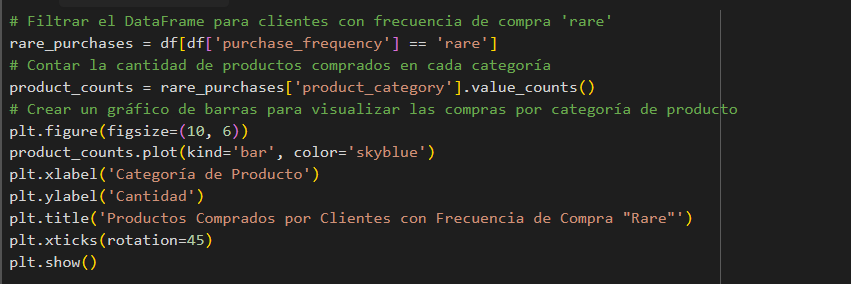
**Forma 2**

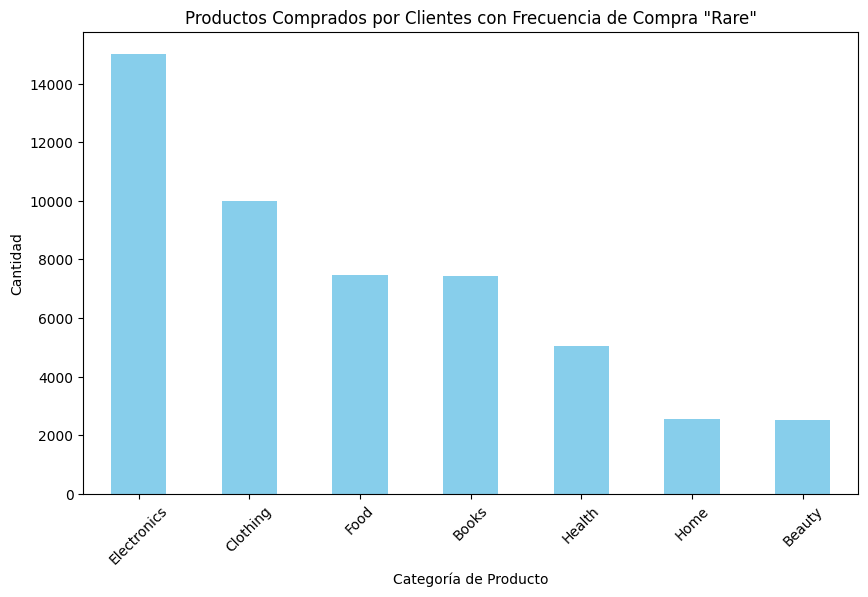
****

**4.6 Productos Comprados por Clientes con Frecuencia de Compra "Rare"**

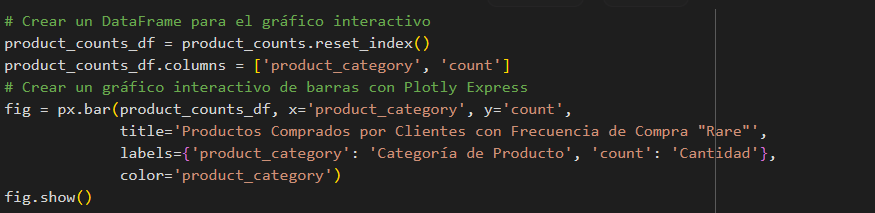
Entender qué productos compran los clientes con una frecuencia de compra 'rare' permite ajustar el inventario y las promociones específicamente para estos productos. Esto puede mejorar la eficiencia del inventario y aumentar las ventas al proporcionar incentivos dirigidos a los productos que estos clientes compran esporádicamente.

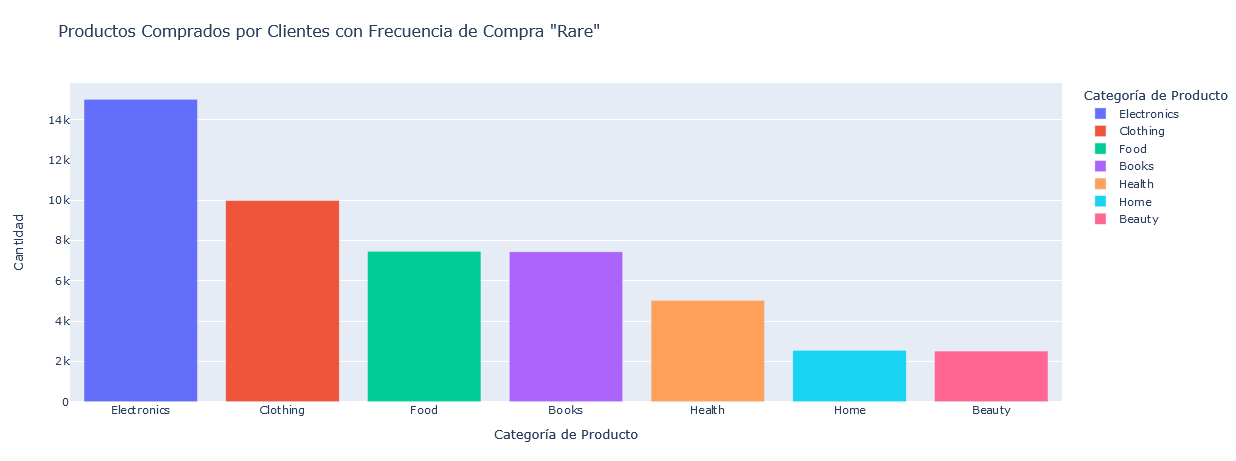
**Forma 1**

****

****

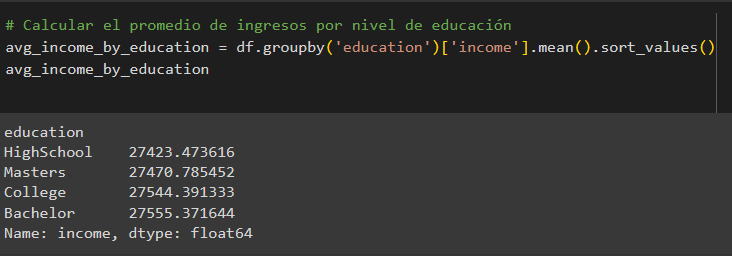
**Forma 2**

****

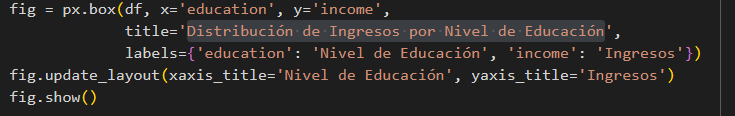
****

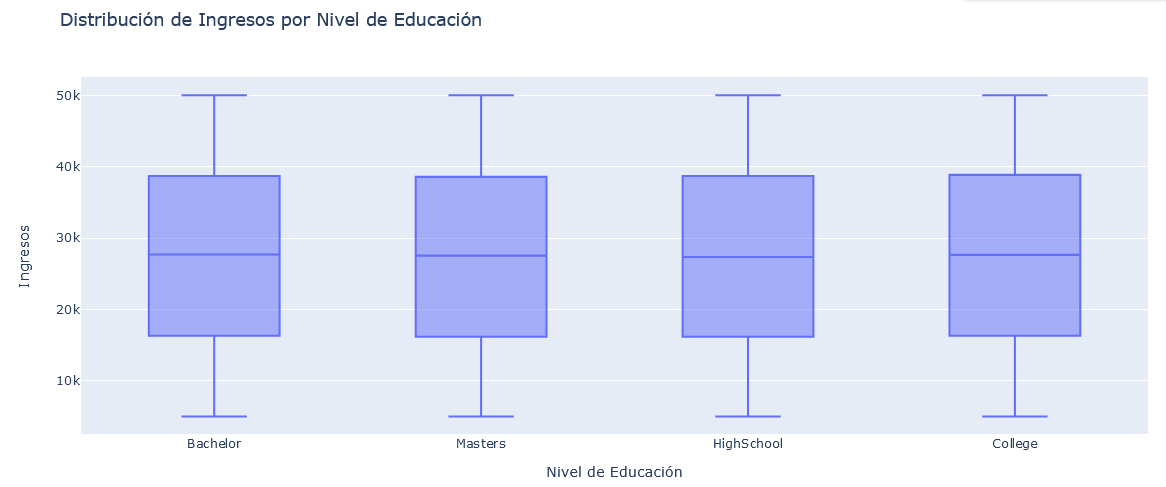
**4.7 Cálculo de promedio y distribución de Ingresos por Nivel de Educación**

Analizar la relación entre ingresos y nivel de educación ayuda a identificar los segmentos de clientes con mayor poder adquisitivo. Esto permite diseñar productos y estrategias de marketing más efectivas dirigidas a estos segmentos, mejorando así la rentabilidad y la eficiencia de las campañas.

****

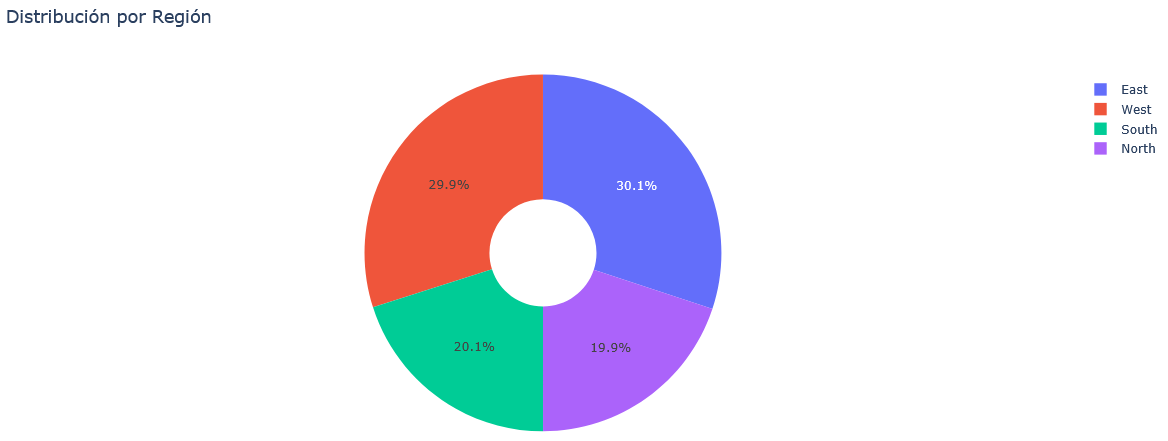
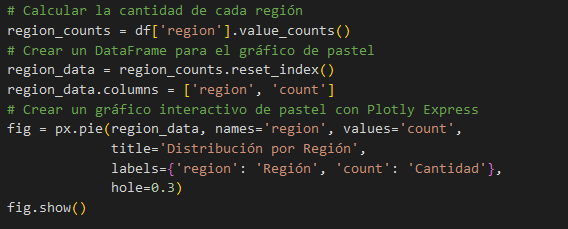
**Forma grafica**

****

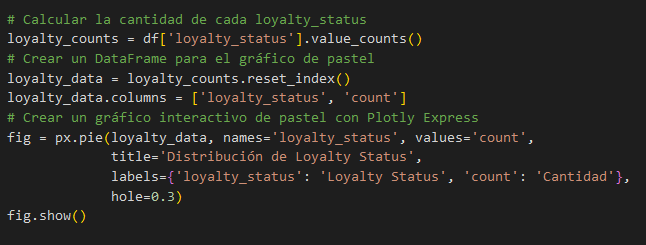
****

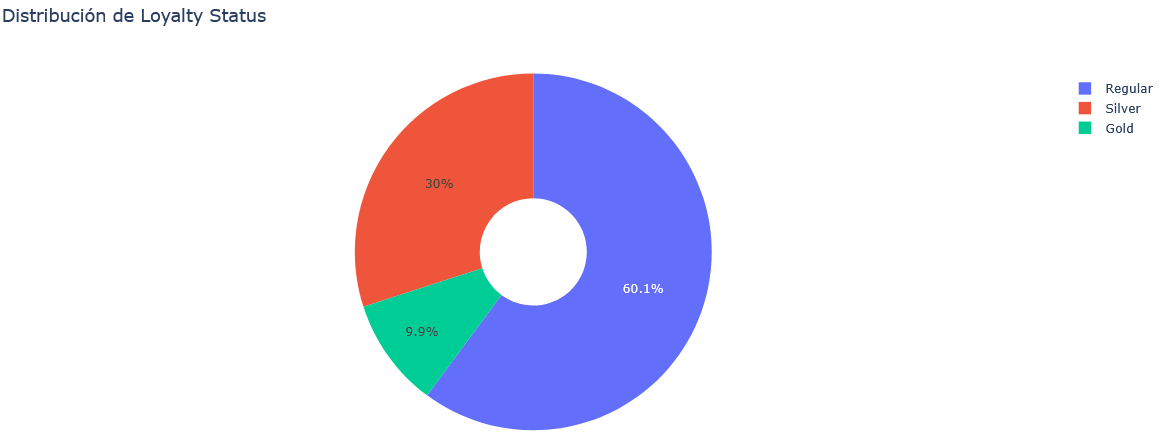
**4.8 Calculo del porcentaje de clientes por cada Región**

Conocer la distribución de clientes por región es crucial para la logística y la planificación de la expansión geográfica. Esto puede mejorar la eficiencia en la distribución de productos y permitir una mejor focalización en las campañas de marketing regionales.

****

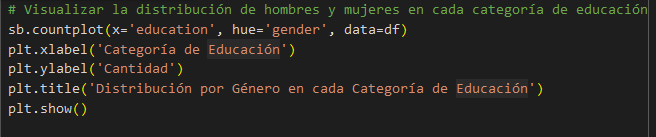
**4.9 Distribución de Loyalty Status**

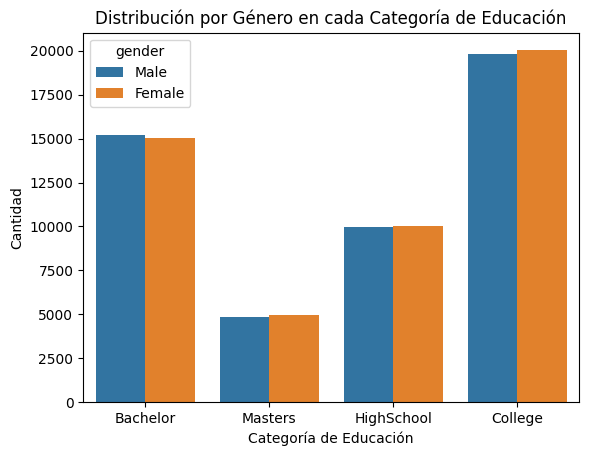
Analizar la distribución de loyalty status permite entender mejor la base de clientes y diseñar programas de lealtad más efectivos. Identificar la proporción de clientes en cada nivel de lealtad ayuda a enfocar esfuerzos en mejorar la retención y fidelización de cliente.****

****

**4.10 La distribución de hombres y mujeres en cada categoría de educación**

Conocer la distribución de género dentro de cada categoría de educación permite personalizar las estrategias de comunicación y marketing. Esto ayuda a crear campañas más inclusivas y efectivas, alineadas con la composición demográfica de los clientes.

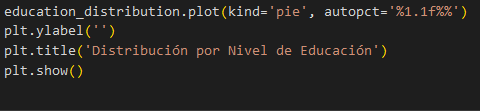
****

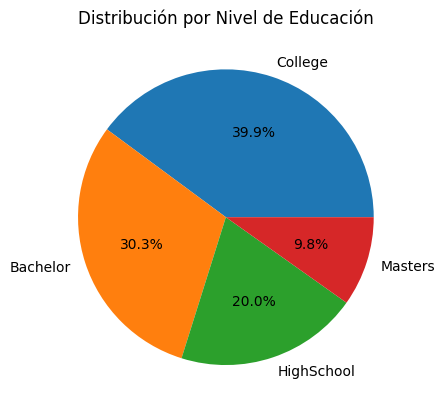
****

**4.11 Distribución por Nivel Educativo**

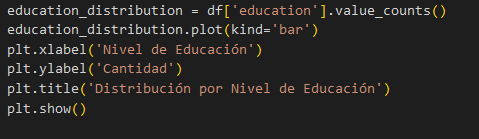
Entender la distribución de los niveles de educación entre los clientes permite diseñar productos y servicios adecuados a las necesidades y preferencias de cada grupo educativo. Esto mejora la segmentación del mercado y la precisión en las estrategias de marketing.

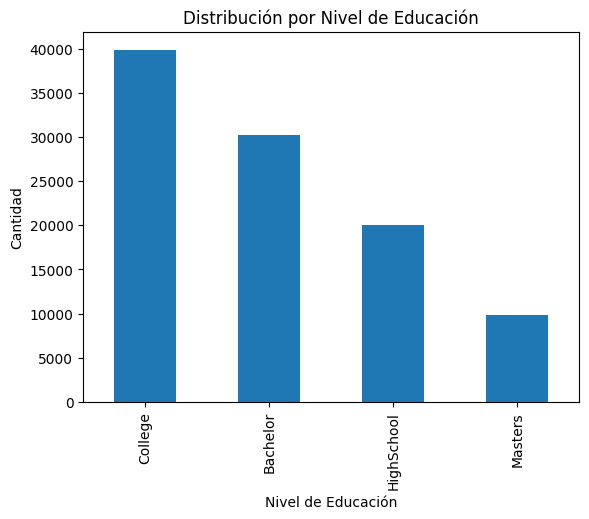
**Forma 1**

****

****

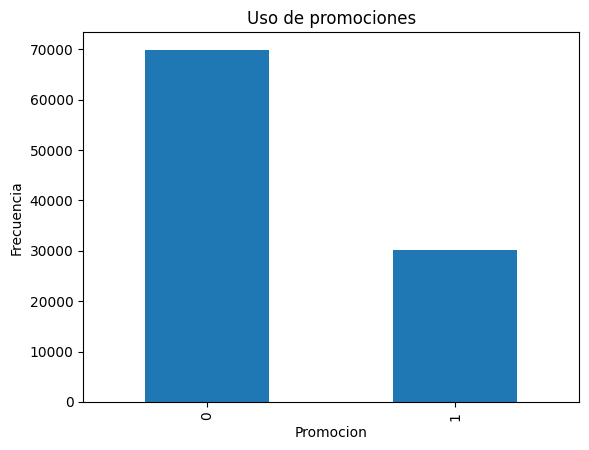
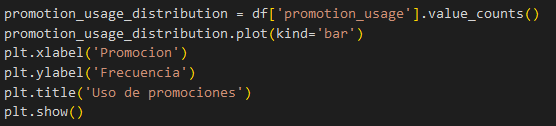
**Forma 2**

****

****

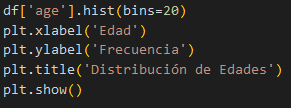
**4.12 Calcular la frecuencia de el uso de promociones**

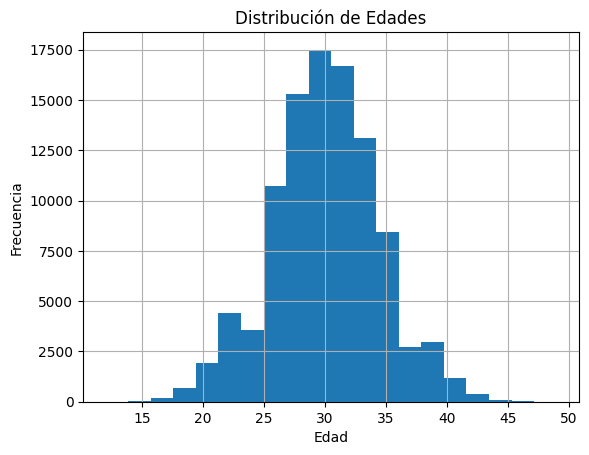
Analizar la frecuencia del uso de promociones en diferentes categorías de productos permite optimizar las campañas promocionales. Con esta información, se pueden diseñar promociones más efectivas y dirigidas, incrementando así las ventas y la satisfacción del cliente.

****

**4.13 Gráfica de la distribución de edad**

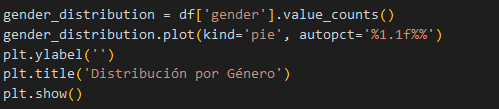
Conocer la distribución de la edad de los clientes ayuda a segmentar el mercado de manera efectiva. Esto permite personalizar las ofertas, los productos y las estrategias de marketing según las edades de los clientes, mejorando la relevancia y efectividad de las campañas.

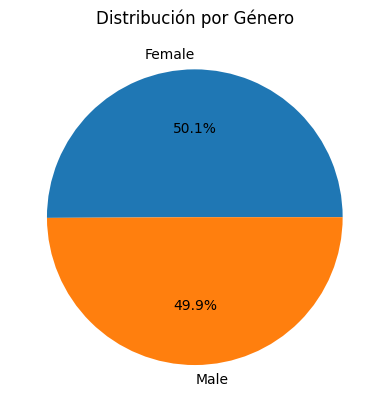
****

****

**4.14 Gráfica de distribución por Género**

Entender la proporción de género entre los clientes es crucial para diseñar estrategias de marketing y productos que resuenen con cada segmento. Esto permite una mejor personalización y adaptación de las campañas y ofertas a las necesidades y preferencias de cada grupo.

****

****

# 5. Conclusiones

El Análisis Exploratorio de Datos realizado sobre el conjunto de datos de ventas de la tienda nos ha permitido obtener conclusiones clave que pueden ser útiles para la toma de decisiones estratégicas y operativas:

* Existe una relación inversa entre el precio de los productos y el volumen de ventas. Esto sugiere que los consumidores tienden a comprar más productos cuando los precios son más bajos, alineándose con la ley de la demanda. Es importante considerar esta relación al establecer estrategias de precios y promociones.
* La matriz de correlación revela una correlación positiva significativa entre los descuentos aplicados y el volumen de ventas. Esto evidencia que las promociones y descuentos son efectivos para aumentar las ventas. Sin embargo, es necesario balancear los descuentos con los márgenes de beneficio para asegurar la rentabilidad.
* La mayor parte de los clientes prefieren productos de las categorías de electrodomésticos y ropa. Para aumentar las ventas y la satisfacción del cliente, es esencial centrar las estrategias de marketing y las promociones en estos dos sectores. Además, comprender las promociones más efectivas en cada categoría puede ayudar a diseñar campañas más dirigidas, como se observa en la alta utilización de promociones en la categoría de libros
* Los clientes con niveles de educación superiores, como Maestros y Bachilleres, tienden a comprar con menos frecuencia pero con montos de compra más altos. Este conocimiento permite crear estrategias específicas para incentivar compras más frecuentes entre estos grupos, tal vez mediante promociones exclusivas o programas de lealtad diseñados para altos ingresos y educación avanzada.
* La distribución de clientes por género, región y nivel educativo ofrece una base sólida para segmentar el mercado. Por ejemplo, la distribución de género dentro de cada nivel educativo puede ayudar a personalizar las estrategias de comunicación y marketing, asegurando que las campañas sean inclusivas y resuenen con las diversas demografías de clientes.
* Por último. Identificar y analizar los patrones de compra y características demográficas de los clientes con los mayores ingresos proporciona valiosos insights para diseñar estrategias de fidelización y ofertas personalizadas. Esto puede incluir programas VIP, promociones exclusivas y atención personalizada, lo que no solo aumenta las ventas, sino también la lealtad del cliente.