Benemérita Universidad
Autónoma de Puebla

Facultad de Ciencias de la Computación

Proyecto Final: Sales Data for Economic Data Analysis

Ximena Axel Martínez Pelayo
21 de octubre de 2024

Introducción a la ciencia de datos

M.C. Jaime A. Romero Sierra

Introducción

El análisis de datos se ha convertido en un pilar fundamental para las empresas que buscan mejorar su desempeño comercial y tomar decisiones basadas en información precisa. Al estudiar las transacciones de ventas, es posible identificar patrones de comportamiento de compra, evaluar la rentabilidad de productos y categorías, y detectar tendencias que guíen las estrategias comerciales. Segmentar a los clientes según su perfil demográfico y geográfico también permite crear campañas de marketing más efectivas, enfocadas en mercados con mayor potencial de crecimiento, mientras que un análisis temporal puede revelar periodos de alta y baja demanda, facilitando la planificación de inventarios y la ejecución de promociones oportunas.

Además de comprender mejor a los consumidores, los datos históricos abren la puerta a la anticipación de tendencias futuras, lo que optimiza la planificación comercial y mejora la competitividad. La utilización de herramientas avanzadas para el procesamiento de datos permite descubrir insights valiosos que van desde las preferencias de compra hasta la evaluación de márgenes de ganancia, todo con el fin de maximizar recursos y priorizar productos más rentables.

Este enfoque no solo busca mejorar la rentabilidad, sino también establecer un camino claro para tomar decisiones informadas que favorezcan el crecimiento sostenido y la competitividad en un entorno cada vez más dinámico.

Objetivo del proyecto

Analizar un conjunto de datos sobre transacciones de ventas para optimizar estrategias comerciales, mejorar la rentabilidad y proporcionar información clave para la toma de decisiones basadas en datos. Identificar patrones de comportamiento de compra, evaluar la rentabilidad de productos y categorías y detectar tendencias de ventas permite diseñar estrategias más efectivas y personalizadas.

Segmentar a los clientes de acuerdo a su perfil demográfico y geográfico facilita la creación de campañas de marketing dirigidas y el enfoque en mercados con mayor potencial de crecimiento. Además de que identificar periodos de alta y baja demanda a través del análisis temporal contribuye a una planificación de inventarios más precisa y estrategias promocionales fundamentadas.

Incorporar modelos predictivos basados en los datos históricos permite anticipar tendencias futuras de ventas y optimizar procesos de planificación, garantizando una mejora continua en la gestión comercial. Este enfoque asegura un crecimiento sostenible y una mayor competitividad en un entorno dinámico

Descripción del proyecto

El análisis de transacciones de ventas es esencial para comprender el desempeño de un negocio, ya que proporciona una visión integral de factores que afectan la rentabilidad y el comportamiento del consumidor. La variedad de datos ofrece la oportunidad de realizar análisis a múltiples niveles, incluyendo segmentación de clientes, evaluación de desempeño por producto o categoría y análisis geográfico y temporal.

Trabajar con este tipo de datos requiere el uso de técnicas avanzadas de análisis para abordar tanto datos numéricos como categóricos, utilizando herramientas de visualización que faciliten la interpretación de resultados. Este proyecto busca descubrir patrones de comportamiento de compra, identificar tendencia de ventas y evaluar la rentabilidad de los productos para informar decisiones estratégicas.

El análisis temporal permite rastrear tendencias por año y mes, identificar ciclos o periodos de alta y baja demanda y con base en ello planificar acciones comerciales que respondan a estas variaciones. Además, el análisis demográfico y geográfico ayuda a entender cómo las preferencias de compra varían entre diferentes segmentos y mercados, lo que permite ajustar estrategias de marketing e identificar áreas con potencial de crecimiento.

Finalmente calcular márgenes de ganancia y comparar costos y precios en diferentes categorías asegura una optimización de recursos y una priorización de productos más rentable. Este enfoque integral busca no solo mejorar el desempeño comercial, sino también fomentar una dinámica de toma de decisiones basadas en datos, impulsando el crecimiento del negocio.

El proyecto se centra en el análisis de un conjunto de datos detallado sobre transacciones de ventas, con el propósito de extraer información clave que permita optimizar las estrategias comerciales y mejorar el desempeño general del negocio. El conjunto de datos incluye información relevante sobre las características demográficas de los clientes, como edad y género, además de detalles geográficos, temporales y financieros relacionados con las ventas. Este análisis es crucial para identificar patrones en el comportamiento de compra de los clientes, comprender las tendencias de ventas a lo largo del tiempo y evaluar la rentabilidad de los productos y categorías ofrecidos.

Justificación

En busca de una respuesta o alternativa estratégica a la creciente necesidad de comprender y anticipar los patrones de compra de los clientes en un mercado global competitivo y dinámico, en un entorno donde la variedad de productos, el acceso a información y las preferencias de los consumidores cambian rápidamente, las empresas deben apoyarse en análisis profundos y fundamentados para tomar decisiones informadas y adaptativas.

El análisis de la base de datos tomando como base las hipótesis propuestas, permiten:

1) Maximizar la rentabilidad: Al identificar categorías y subcategorías de productos que generan mayores ingresos y analizar su relación con otras variables se facilita la toma de decisiones estratégicas que optimizan la inversión y priorizan las áreas con mayor retorno potencial.
2) Segmentación de mercado avanzada: Comprender las preferencias de los clientes en función de factores clave, permite desarrollar estrategias específicas para cada segmento de variabilidad por cliente, asegurando una experiencia de compra más personalizada.
3) Eficiencia en la gestión de recursos: Al alinear la oferta de productos y las estrategias logísticas con las tendencias detectadas, los socios comerciales y proveedores pueden reducir costos operativos, mejorar la distribución y asegurar que los productos más demandados estén disponibles en los mercados adecuados.
4) Aseguramiento de la toma de decisiones estratégicas: Los directivos e inversionistas necesitan información clara, confiable y accionable para orientar el crecimiento del negocio. Resultarán insights valiosos que impactan la planeación a corto plazo y también establecen una visión estratégica a largo plazo basada en datos reales.
5) Fortalecimiento de relaciones comerciales y competitividad: Ajustar la oferta a las demandas específicas del mercado, fomentando relaciones sólidas con clientes y socios comerciales. Asimismo, asegurar que la organización pueda responder con agilidad a las tendencias emergentes, garantizando su sostenibilidad y diferenciación en el mercado.
6) Adaptabilidad a diferencias geográficas: Las conclusiones sobre las preferencias según la ubicación ayudan a atender mercados específicos con mayor precisión, permitiendo estrategias diferenciadas en regiones clave como Estados Unidos y Europa.

Al fomentar una cultura empresarial orientada al análisis y la mejora continua, asegurando que la organización no solo se adapte a las necesidades actuales del mercado, sino que también lidere su evolución. Se garantiza que la empresa está preparada para enfrentar los desafíos del presente y del futuro, posicionándose como un referente de innovación, eficiencia y orientación al cliente en su industria.

Recursos Disponibles

Tecnología y herramientas

Para llevar a cabo el análisis del proyecto, se utilizará Google Colaboratory (Google Colab), una herramienta basada en la nube que permite la escritura y ejecución de código en notebooks de Jupyter sin necesidad de instalación local. Google Colab facilita la colaboración en tiempo real, la integración con Google Drive y el uso de otros recursos que facilita manejar grandes volúmenes de datos.

El lenguaje de programación principal que se empleará es Python, popularmente utilizado en proyectos de análisis de datos y aprendizaje automático por su facilidad de uso y gran repositorio de bibliotecas. Entre las bibliotecas utilizadas en el análisis estarán:

- NumPy: Para cálculos numéricos y manipulación de arrays.
- Pandas: Para la manipulación y análisis de datos estructurados, especialmente útil para conjuntos de datos tabulares.
- Matplotlib & Seaborn: Para la creación de visualizaciones que permitan interpretar patrones y tendencias.
- TensorFlow: Para la implementación de modelos avanzados de aprendizaje automático, si es requerido para el análisis.
- Otras bibliotecas como Scikit-learn o Plotly, dependiendo de las necesidades específicas del proyecto.

Hay que destacar que los datos utilizados en este análisis provienen del conjunto de datos titulado "Sales Data for Economic Data Analysis", disponible en la plataforma de colaboración en línea Kaggle. Kaggle es un recurso ampliamente reconocido en la comunidad de ciencia de datos, ya que facilita el acceso a conjuntos de datos variados, la publicación de proyectos y la colaboración entre usuarios con intereses y objetivos similares.

El conjunto de datos proporciona información estructurada y de calidad que permite explorar diversas métricas relacionadas con las transacciones de ventas, como características demográficas, detalles geográficos, entre otras. Kaggle, además de ser una fuente confiable de datos, fomenta el aprendizaje automático y el análisis de datos.

La combinación de herramientas tecnológicas avanzadas como Google Colab y bibliotecas especializadas de Python, junto con un conjunto de datos estructurado y accesible desde Kaggle, asegura una base sólida para la ejecución exitosa del proyecto. Lo anterior, garantiza un entorno de análisis eficiente y adaptable a los requerimientos del proyecto.

Datos

El análisis de la base datos depende de variables clave que permiten explorar tendencias, patrones de comportamiento del consumidor y métricas del rendimiento financiero. A continuación, se describen las columnas que representan características específicas en las transacciones y se explica su utilidad en el análisis.

- Year (Año): Indica el año en que ocurrió la transacción. Esta columna es crucial para identificar tendencias a largo plazo y analizar variaciones en las ventas por periodos anuales.
- Month (Mes): Representa el mes de la transacción, lo que permite un análisis más detallado de la estacionalidad y de los picos o caídas de la demanda.
- Customer Age (Edad del Cliente): Define la edad del cliente, proporcionando información valiosa para segmentar el mercado por grupos etarios y analizar sus comportamientos de compra.
- Customer Gender (Género del Cliente): señala el género del cliente, permitiendo explorar diferencias en las preferencias de compra según el género.
- Country (País): Indica el país donde se realizó la transacción, ofreciendo una perspectiva geográfica de las ventas y la oportunidad de comparar mercados internacionales.
- State (Estado): Especifica el estado dentro del país, permitiendo un análisis más detallado a nivel regional.
- Product Category (Categoría del Producto): Clasifica los productos en categorías generales, facilitando la comparación del rendimiento entre diferentes líneas de productos.
- Sub Category (Subcategoría): Detalla la categoría específica del producto vendido, permitiendo un análisis granular de las ventas.
- Quantity (Cantidad): Define el volumen de unidades vendidas, proporcionando datos esenciales para medir la demanda y calcular ingresos.
- Unit Cost (Costo Unitario): Indica el precio al que se vendió una unidad del producto, ayudando a analizar estrategias de precios.
- Cost (Costo Total): Calculado como el producto de la cantidad vendida y el costo unitario, refleja los gastos asociados a las ventas.
- Revenue (Ingresos): Calculando como el producto de la cantidad vendida y el precio unitario, indica los ingresos totales generados por las ventas.

Esta organizada estructura de datos proporciona la base necesaria para realizar un análisis profundo que apoye la toma de decisiones informadas y la implementación de estrategias comerciales efectivas.

Hipótesis Iniciales

1) **Relación entre la categoría del producto, el ingreso generado y el costo unitario**

   Hipótesis: Los productos de la categoría “Bikes” generan mayores ingresos promedio debido a su alto costo unitario, en comparación con las categorías “Clothing” y “Accessories”

2) **Impacto de la edad del cliente en las compras considerando subcategorías y costo unitario**

   Hipótesis: Los clientes más jóvenes (menores de 35 años) prefieren productos de subcategorías con menor costo unitario, como “Socks” y “Caps” en la categoría de accesorios, mientras que los clientes mayores al rango de edad establecido optan por subcategorías con mayor costo unitario, como “Mountain Bikes” y “Road Bikes” en la categoría de bicicletas

3) **Diferencias geográficas en la preferencia de productos**

   Hipótesis: Los clientes en estados europeos tienden a una mayor preferencia por productos de ropa y accesorios en comparación con clientes en Estados Unidos, que prefieren bicicletas.

Stakeholders Clave

Los stakeholders clave son factores con una relación directa con el proyecto, cuya participación es esencial para la toma de decisiones estratégicas, la interpretación de resultados y la implementación de mejoras basadas en el análisis de datos. A continuación, se describen los principales para este proyecto:

- Proveedores o Socios Comerciales: Se refiere a las empresas que producen o distribuyen los productos analizados. Su rol es ajustar su oferta y logística con base a las tendencias descubiertas en los datos. Además, colaboran estrechamente con los equipos de ventas y marketing para alinear estrategias comerciales, maximizando el impacto en el mercado y mejorando la experiencia del cliente final.
- Directivos o Inversionistas: Son tomadores de decisiones de alto nivel y gran importancia estratégica, interesados en los resultados financieros, tomando decisiones sobre asignación de recursos hacia áreas con mayor potencial de crecimiento o mejora. Su rol es crucial para garantizar que los insights se traduzcan en acciones concretas y alineadas con los objetivos organizacionales.
- Clientes finales: Involucra a todas las personas que adquieren los productos analizados. Representan la base del análisis, ya que sus patrones de comportamiento, preferencias y tendencias de compra determinan las decisiones estratégicas.

Preguntas clave

¿Qué ajustes en la oferta de productos pueden realizarse para maximizar los ingresos por categoría?

¿Cómo afectan las decisiones de precios a las ventas generales por categoría?

¿Qué proporción de los ingresos proviene de cada categoría y cómo esto impacta la rentabilidad general?

¿Prefieren los clientes productos de menor costo unitario, pero con mayor frecuencia de compra?

¿Qué subcategorías muestran mayor crecimiento en ventas entre distintos segmentos de edad?

¿Qué grupos de edad representan la mayor proporción de ingresos para cada subcategoría?

¿Cómo puede influir el comportamiento por edades en las proyecciones de crecimiento?

¿Qué características buscan los clientes jóvenes en contraste con los clientes mayores?

¿Cómo varía la demanda de ropa y accesorios entre estados europeos y Estados Unidos?

¿Qué regiones geográficas generan mayores ingresos para cada categoría de producto?

¿Cómo influyen los costos de distribución entre regiones en la rentabilidad?

¿Sería importante considerar factores culturales en las preferencias regionales?

¿Cuáles son los productos más rentables para cada segmento de clientes?

Fuentes de datos

La base de datos probablemente fue construida a partir de una combinación de fuentes comunes utilizadas en análisis comerciales y de mercado. Es esencial que los datos hayan sido ajustados y procesados para garantizar su relevancia, precisión y aplicabilidad en el contexto del proyecto. A continuación, se describen posibles fuentes que pudieron contribuir a su elaboración y seguir asegurando su confiabilidad:

- Registros de Ventas: Sistemas de Punto de Venta o plataformas de comercio electrónico donde pudieron considerar datos como:
    - Productos vendidos, categoría, subcategoría y costo unitario
    - Información geográfica (ubicación del cliente o tienda)
    - Fecha y hora de compra
    - Total de ingresos generados por cada venta

- Perfiles de clientes: Base de datos de clientes o programas de fidelidad, donde pudieron considerar datos como:
    - Edad, género y ubicación
    - Historial de compras
    - Preferencias de productos y patrones de compra
    - Segmentación por demografía o comportamiento

- Datos de Mercado y Competencia: Informes de investigación de mercado, estudios sectoriales o benchmarking, donde pudieron considerar datos como:
    - Tendencias generales de consumo en categorías como bicicletas, ropa y accesorios.
    - Preferencias geográficas o culturales en el consumo
    - Comparaciones entre marcas competidoras

- Inventarios y Producción: Sistemas de gestión de inventarios o Enterprise Resource Planning (ERP), donde pudieron considerar datos como:
    - Costos unitarios asociados a la fabricación o distribución
    - Disponibilidad de productos por región o tienda

  - Rotación de inventarios

- Encuestas y opiniones de clientes: Encuestas directas, opiniones en línea o redes sociales, donde pudieron considerar datos como:
  - Preferencias y comportamientos según categorías de producto
  - Satisfacción con las subcategorías ofrecidas
  - Factores decisivos de compra, como precio o diseño

- Datos Geográficos y Demográficos: Fuentes externas como censos, datos gubernamentales o bases de datos comerciales, donde pudieron considerar datos como:
  - Distribución poblacional y segmentos por región
  - Niveles de ingreso y gasto promedio por área geográfica
  - Cultura que influya en las preferencias de productos

- Análisis de Marketing Digital: Herramientas de análisis web (Google Analytics, redes sociales, entre otras.), donde pudieron considerar datos como:
  - Comportamiento del usuario en sitios web o aplicaciones móviles
  - Popularidad de productos específicos según campañas publicitarias
  - Localización de los clientes que interactúan con las plataformas digitales

Tipo de datos

Con un total de 44,611 registros y 16 columnas, la información detalle por columna es la siguiente:

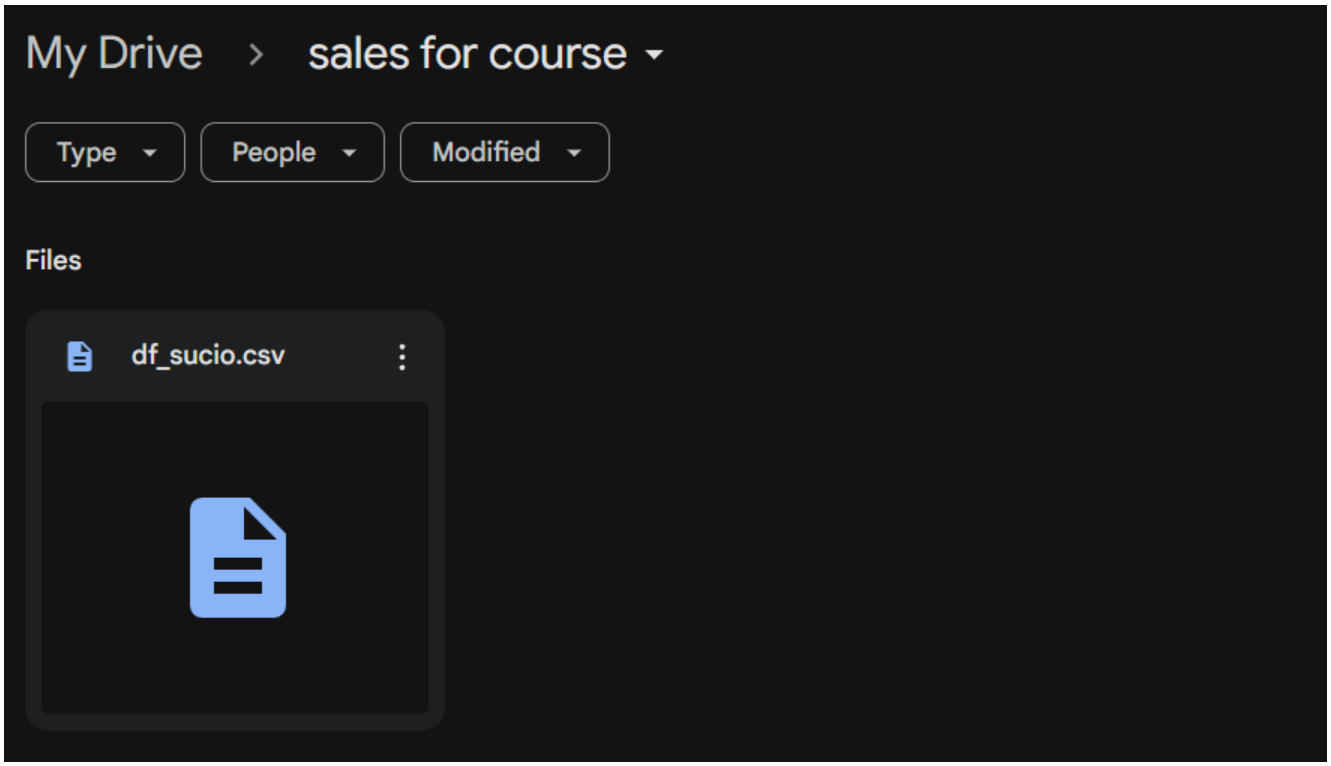
Columna	Tipo de Dato	Notas de análisis
index	float64	Este es un tipo de dato numérico.
Date	object	Es una columna de tipo texto. Probablemente debiera convertida a tipo fecha.
Year	float64	Este es un tipo de dato numérico.
Month	float64	Tiene solo valores nulos. Debería ser eliminada o revisada.
Customer Age	float64	Este es un tipo numérico, pero tiene valores nulos que deberían ser tratados.
Customer Gender	object	Tipo de dato texto (categoría). Tiene valores nulos.
Country	object	Tipo de dato texto (categoría). Sin valores nulos.
State	object	Tipo de dato texto (categoría). Sin valores nulos.
Product Category	object	Tipo de dato texto (categoría). Sin valores nulos.
Sub Category	object	Tipo de dato texto (categoría). Sin valores nulos.
Quantity	float64	Tipo numérico. Sin valores nulos.
Unit Cost	object	Aunque es texto, debería ser numérico. Necesita conversión.
Unit Price	object	Aunque es texto, debería ser numérico. Necesita conversión.
Cost	float64	Tipo numérico. Sin valores nulos.
Revenue	float64	Tipo numérico. Sin valores nulos.
Column1	float64	La mayoría de los valores son nulos. Debería ser revisada o eliminada.

Manejo de la base de datos

El manejo adecuado de la base de datos es esencial para garantizar un análisis efectivo y confiable. A continuación, se describe el proceso detallado utilizado para trabajar con un archivo *.csv* obtenido de *Kaggle*.

Organización y Almacenamiento

Es importante almacenar el archivo .csv en un lugar accesible y seguro que soporte su tamaño en megabytes (MB). Para este proyecto, se utiliza *Google Drive*, una herramienta que facilita el almacenamiento en la nube y permite integrarse fácilmente con *Google Colab*. El archivo .csv se ubicará en una carpeta dedicada exclusivamente al proyecto, lo que asegura un entorno organizado.



Cargar la Base de Datos en Google Colab

Para trabajar con la base de datos en *Google Colab*, es necesario establecer una conexión entre la notebook y *Google Drive*. Este paso permite acceder al archivo almacenado en la nube. Para ello, se utiliza el siguiente código:

Este fragmento de código enlaza *Google Drive* en la notebook, otorgando acceso a los archivos almacenados en la carpeta del proyecto. Una vez completado este paso, es posible navegar por las carpetas de *Google Drive* desde el entorno de *Colab* y así cargar la base de datos con el siguiente código:

```
[125] #Cargar la base de datos
      df = pd.read_csv('/content/drive/MyDrive/sales for course/df_sucio.csv')
      df
```

Importar las Bibliotecas necesarias

Para trabajar con datos en *Python*, es necesario importar las bibliotecas adecuadas. Afortunadamente, *Google Colab* ya incluye muchas de las bibliotecas esenciales, como *Pandas* y *NumPy*, pero se deben importar explícitamente para garantizar su disponibilidad en el código.

Ejemplo de importación de bibliotecas:

```
[2]  #importar las bibliotecas
     import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
```

Si se estuviera trabajando en otro entorno, como *Visual Studio Code*, sería obligatorio instalar estas bibliotecas utilizando el administrador de paquetes pip. Esto asegura la compatibilidad del código en distintos entornos.

Identificación de Datos Sucios

Es fundamental identificar y corregir datos faltantes o incorrectos antes de realizar un análisis estadístico. Los valores faltantes o erróneos pueden alterar el orden natural de los datos, distorsionando los resultados y, en última instancia, influir negativamente en la toma de decisiones. Por ello es esencial revisar y validar las bases de datos antes de considerarlas como definitivas. Los principales tipos de problemas que se deben evitar incluyen:

- Datos nulos: Valores faltantes que pueden distorsionar cálculos como promedios, sumas y demás operaciones estadísticas.
- Datos duplicados: Registros redundantes que inflan los resultados y generan interpretaciones incorrectas.
- Datos incorrectos: Columnas que contienen valores que no coinciden con el tipo de dato esperado, como texto en una columna numérica.
- Valores inválidos: Datos que no tienen sentido dentro del contexto, como edades negativas o precios con valores cero.
- Datos atípicos: Valores extremos o inusuales que podrían ser errores o casos excepcionales. Si no se manejan adecuadamente, pueden sesgar el análisis.
- Inconsistencias en texto: Errores tipográficos o variaciones de formato, como diferencias en el uso de mayúsculas y minúsculas en categorías similares.

Evaluación Inicial

El primer paso en el análisis es inspeccionar las filas, también llamadas registros del archivo para entender su estructura y detectar posibles problemas. Para ellos, es recomendable obtener un resumen general de las columnas, tipos de datos y valores nulos presentes:

```
#Información de la base
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 44611 entries, 0 to 44610
Data columns (total 16 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   index             43273 non-null  float64
 1   Date              43272 non-null  object
 2   Year              43272 non-null  float64
 3   Month             0 non-null      float64
 4   Customer Age      42407 non-null  float64
 5   Customer Gender   43272 non-null  object
 6   Country           43272 non-null  object
 7   State             43272 non-null  object
 8   Product Category  43272 non-null  object
 9   Sub Category      43272 non-null  object
 10  Quantity          43272 non-null  float64
 11  Unit Cost         43272 non-null  object
 12  Unit Price        43272 non-null  object
 13  Cost              43272 non-null  float64
 14  Revenue           43273 non-null  float64
 15  Column1           3191 non-null   float64
dtypes: float64(8), object(8)
memory usage: 5.4+ MB
```

Tras evaluar la base de datos, se determinó que contiene 44,611 registros categorizados en 16 columnas. A continuación, se detalla el análisis de cada columna, su tipo de dato y las observaciones que se pueden hacer con la información obtenida:

Columna	Tipo de Dato	Notas de análisis
index	float64	Este es un tipo de dato numérico.
Date	object	Es una columna de tipo texto. Probablemente debiera convertida a tipo fecha.
Year	float64	Este es un tipo de dato numérico.
Month	float64	Tiene solo valores nulos. Debería ser eliminada o revisada.
Customer Age	float64	Este es un tipo numérico, pero tiene valores nulos que deberían ser tratados.
Customer Gender	object	Tipo de dato texto (categoría). Tiene valores nulos.
Country	object	Tipo de dato texto (categoría). Sin valores nulos.
State	object	Tipo de dato texto (categoría). Sin valores nulos.
Product Category	object	Tipo de dato texto (categoría). Sin valores nulos.

Sub Category	object	Tipo de dato texto (categoría). Sin valores nulos.
Quantity	float64	Tipo numérico. Sin valores nulos.
Unit Cost	object	Aunque es texto, debería ser numérico. Necesita conversión.
Unit Price	object	Aunque es texto, debería ser numérico. Necesita conversión.
Cost	float64	Tipo numérico. Sin valores nulos.
Revenue	float64	Tipo numérico. Sin valores nulos.
Column1	float64	La mayoría de los valores son nulos. Debería ser revisada o eliminada.

De acuerdo con la evaluación, los principales problemas identificados en la base de datos incluyen:

- Datos incorrectos: Columnas como Unit Cost y Unit Price están en formato texto y deben ser convertidas a numérico.
- Datos nulos: Columnas como Customer Age, Customer Gender, y Month contienen valores faltantes que requieren imputación o eliminación.
- Datos atípicos: Es necesario identificar y manejar valores extremos en columnas como Revenue y Cost.

Describir Estadísticas iniciales

La descripción estadística inicial de una base de datos es un paso fundamental en el análisis de datos. Este proceso no solo proporciona un panorama general del contenido, sino que también permite identificar tendencias, rangos, valores atípicos y posibles problemas que podrían afectar el análisis posterior.

```
[128] #Describir las estadísticas
      print(df.describe(include='all'))
```

```
                index     Date          Year  Month  Customer Age  \
count   43273.000000    43272  43272.000000    0.0  42407.000000
unique           NaN      576           NaN    NaN           NaN
top              NaN  invalid           NaN    NaN           NaN
freq             NaN      857           NaN    NaN           NaN
mean    17435.637603      NaN   2015.568428    NaN     36.390077
std     10051.100359      NaN      0.495301    NaN     11.122280
min         0.000000      NaN   2015.000000    NaN     17.000000
25%      8732.000000      NaN   2015.000000    NaN     28.000000
50%     17401.000000      NaN   2016.000000    NaN     35.000000
75%     26144.000000      NaN   2016.000000    NaN     44.000000
max     34866.000000      NaN   2016.000000    NaN     87.000000

       Customer Gender        Country       State Product Category  \
count            43272          43272       43272            43272
unique               2              4          46                4
top                  M  United States  California      Accessories
freq             22169          22458       12665            27428
mean               NaN            NaN         NaN              NaN
std                NaN            NaN         NaN              NaN
min                NaN            NaN         NaN              NaN
25%                NaN            NaN         NaN              NaN
50%                NaN            NaN         NaN              NaN
75%                NaN            NaN         NaN              NaN
max                NaN            NaN         NaN              NaN

           Sub Category      Quantity Unit Cost Unit Price          Cost  \
count             43272  43272.000000     43272      43272  43272.000000
unique               18           NaN       882       5111           NaN
top     Tires and Tubes           NaN   invalid    invalid           NaN
freq              13461           NaN       868        858           NaN
mean                NaN      2.001271       NaN        NaN    573.863815
std                 NaN      0.813315       NaN        NaN    687.566174
min                 NaN      1.000000       NaN        NaN      2.000000
25%                 NaN      1.000000       NaN        NaN     85.000000
50%                 NaN      2.000000       NaN        NaN    261.000000
75%                 NaN      3.000000       NaN        NaN    769.000000
max                 NaN      3.000000       NaN        NaN   3600.000000

             Revenue      Column1
count   43273.000000  3191.000000
unique           NaN          NaN
top              NaN          NaN
freq             NaN          NaN
mean      638.790297   682.708145
std       734.480650   772.447090
min         2.000000     2.000000
25%       102.000000   103.500000
50%       318.000000   389.000000
75%       901.000000   956.500000
max      5082.000000  3681.000000
```

A continuación, se presentan las observaciones obtenidas de la base de datos:

1. Análisis de Columnas Numéricas

   Las columnas numéricas (*Customer Age, Quantity, Cost, Revenue*) revelan las siguientes características:

   - Rangos razonables: Los valores están dentro de márgenes esperados, permitiendo realizar cálculos estadísticos confiables.

   - Distribución adecuada**:** La mayoría de las columnas muestran una distribución homogénea, aunque ciertas métricas (*Cost* y *Revenue*) presentan posibles valores extremos que deberán ser revisados más a fondo para determinar si se trata de outliers reales o errores en los datos.

2. Análisis de Columnas Categóricas

   Las columnas categóricas (*Product Category, Sub Category, Country*) tienen características que facilitan la segmentación y el análisis:

   - Distribuciones claras: Existen categorías predominantes, lo que puede guiar la creación de segmentos significativos para análisis específicos.
   - Inconsistencias detectadas**:** Algunas columnas, como *Unit Cost* y *Unit Price*, contienen valores inválidos o etiquetas incorrectas. Estas deben ser revisadas y corregidas para asegurar la calidad del análisis y permitir su conversión a tipos numéricos.

3. Identificación de Datos Faltantes

   El análisis confirma y detalla la presencia de datos faltantes:

   - Valores nulos en columnas clave: *Customer Age* y *Customer Gender* presentan valores nulos que deben ser tratados para evitar sesgos en los análisis posteriores.
   - Columnas con proporción alta de nulos: *Month* y *Column1* contienen demasiados valores faltantes, lo que podría justificar su eliminación o una evaluación más profunda de su relevancia.

Con base en este análisis preliminar, se puede concluir que la base de datos tiene un diseño funcional y un contenido valioso para el análisis, pero requiere ciertos ajustes para optimizar su calidad.

Limpieza de Datos Sucios Identificados

Una vez identificados los problemas presentes en la base de datos, podemos abordar su limpieza de manera más efectiva.

Cambio de formato

Para poder trabajar correctamente con los datos, es necesario asegurarse de que estén en el formato adecuado. En este paso, se verifica que los valores que no corresponden al formato esperado sean corregidos. Para identificar los datos con formato incorrecto, se utiliza el siguiente comando:

```
[129] #Se establece el formato óptimo de dato

      from datetime import datetime

      expected_types = {
          'index': float,
          'Date': datetime,
          'Year': float,
          'Month': float,
          'Customer Age': float,
          'Customer Gender': str,
          'Country': str,
          'State': str,
          'Product Category': str,
          'Sub Category': str,
          'Quantity': float,
          'Unit Cost': float,
          'Unit Price': float,
          'Revenue': float,
          'Column1': float
      }

      # Verifica que todos los valores en cada columna cumplan con el tipo esperado
      type_check_results = {
          column: df[column].apply(lambda x: isinstance(x, dtype) or pd.isna(x)).all()
          for column, dtype in expected_types.items()
      }

      # Muestra los resultados
      for column, is_correct_type in type_check_results.items():
          print(f"Columna '{column}': {'Correcta' if is_correct_type else 'Incorrecta'}")
```

```
Columna 'index': Correcta
Columna 'Date': Incorrecta
Columna 'Year': Correcta
Columna 'Month': Correcta
Columna 'Customer Age': Correcta
Columna 'Customer Gender': Correcta
Columna 'Country': Correcta
Columna 'State': Correcta
Columna 'Product Category': Correcta
Columna 'Sub Category': Correcta
Columna 'Quantity': Correcta
Columna 'Unit Cost': Incorrecta
Columna 'Unit Price': Incorrecta
Columna 'Revenue': Correcta
Columna 'Column1': Correcta
```

Como se puede observar, existen dos tipos de datos incorrectos. Estos deben ser convertidos al formato adecuado: numérico para algunas columnas y de tipo fecha para otras, como se muestra a continuación:

```
[130] #Convertir formato a valor numérico
      df['Unit Cost'] = pd.to_numeric(df['Unit Cost'], errors='coerce')
      df['Unit Price'] = pd.to_numeric(df['Unit Price'], errors='coerce')

[131] #Convertir fechas
      df['Date'] = pd.to_datetime(df['Date'], errors='coerce')
```

Una vez realizadas las conversiones, se debe verificar que todo esté en orden. Para ello, se vuelve a ejecutar el comando inicial y, si los datos están correctamente formateados, se continúa con el siguiente paso:

```
# Verifica que todos los valores en cada columna cumplan con el tipo esperado
type_check_results = {
    column: df[column].apply(lambda x: isinstance(x, dtype) or pd.isna(x)).all()
    for column, dtype in expected_types.items()
}

# Muestra los resultados
for column, is_correct_type in type_check_results.items():
    print(f"Columna '{column}': {'Correcta' if is_correct_type else 'Incorrecta'}")
```

```
Columna 'index': Correcta
Columna 'Date': Correcta
Columna 'Year': Correcta
Columna 'Month': Correcta
Columna 'Customer Age': Correcta
Columna 'Customer Gender': Correcta
Columna 'Country': Correcta
Columna 'State': Correcta
Columna 'Product Category': Correcta
Columna 'Sub Category': Correcta
Columna 'Quantity': Correcta
Columna 'Unit Cost': Correcta
Columna 'Unit Price': Correcta
Columna 'Revenue': Correcta
Columna 'Column1': Correcta
```

Eliminación de Registros Duplicados

Ya se han identificado los registros duplicados. Para asegurar que el método utilizado para eliminarlos sea eficiente, es necesario conocer cuántos registros duplicados existen en la base de datos. Un registro se considera duplicado cuando los datos de todas las columnas coinciden exactamente con los de otro registro. Estos registros deben eliminarse para evitar duplicar información innecesaria. Para obtener el total de registros duplicados, se utiliza el siguiente comando:

```
[133] # Filas duplicadas
      print(df.duplicated().sum())

      3367
```

Para eliminar los registros duplicados, se utiliza el siguiente comando:

```
[134] #Eliminar los registros duplicados
      df.drop_duplicates(inplace = True)
```

Una vez eliminados, podemos verificar que los duplicados hayan sido eliminados correctamente con el siguiente comando:

```
[135] #Filas duplicadas corroboración
      print(df.duplicated().sum())

      0
```

Valores faltantes

Contar con datos completos es fundamental para realizar un análisis efectivo. En este paso, se identifican las columnas que contienen valores faltantes. Dependiendo del volumen de datos faltantes por columna, se evalúa si la columna es relevante o si debe eliminarse. En el caso de columnas con pocos datos faltantes, estos pueden ser imputados utilizando valores aproximados como la media o el promedio. Sin embargo, si la cantidad de datos faltantes es considerable, la columna se elimina, ya que podría sesgar el análisis.

Para identificar los valores faltantes en cada columna, se utiliza el siguiente comando:

```
[136] # Columnas con total de valores nulos
      print(df.isnull().sum())

      index                1336
      Date                 2185
      Year                 1336
      Month               41244
      Customer Age         2197
      Customer Gender      1338
      Country              1332
      State                1335
      Product Category     1337
      Sub Category         1334
      Quantity             1338
      Unit Cost            2193
      Unit Price           2189
      Cost                 1335
      Revenue              1336
      Column1             38292
      dtype: int64
```

```
[137] #Comprobar la extensión
      df.shape

      (41244, 16)
```

Es importante verificar la extensión de la base de datos, ya que los valores faltantes modifican el tamaño de la base, ya sea al disminuir o incrementar los registros. Tras revisar la extensión y analizar los datos faltantes, se observa que las columnas *Month* y *Column1* tienen un número excesivo de valores nulos. Por lo tanto, se opta por eliminar estas columnas con el siguiente comando:

```
# Eliminar las columnas sin datos
df = df.drop(columns=['Month','Column1'])
```

Después de eliminar las columnas, se verifica nuevamente la extensión de la base para asegurarse de que los registros eliminados no hayan afectado otras áreas de la base de datos:

```
[139] #Comprobar la nueva extensión
      df.shape

      (41244, 14)
```

En los casos restantes, en los que se imputarán valores, se procede con la imputación de los valores faltantes. En el caso de la columna *Date*, se decide imputar los valores faltantes como 'Unknown', lo cual mejora la calidad del registro, como se muestra en el siguiente código:

```
# Imputar con un valor fijo
df['Date'].fillna('Unknown', inplace=True)
df['Year'].fillna('Unknown', inplace=True)
```

Para la columna *Customer Age*, que es una variable continua, se elige la mediana como valor de imputación, ya que se espera que la edad tenga una distribución normal. El código para imputar la mediana es el siguiente:

```
[142] #Imputar Customer Age con mediana
      df['Customer Age'].fillna(df['Customer Age'].median(), inplace=True)
```

Similar a la edad, para la columna *Customer Gender*, se utiliza la moda para imputar los valores faltantes, ya que es más probable que el género siga una distribución que se ajuste a la moda. En este caso, aunque factores como el país o la categoría de producto podrían influir, la moda es una estimación razonable. El código para imputar la moda es el siguiente:

```
[143] #Imputar Customer Gender mediante la moda
      df['Customer Gender'].fillna(df['Customer Gender'].mode()[0], inplace=True)
```

Para las columnas *Country* y *State*, se puede aprovechar la posible relación con otras variables, como la región de ventas. Se utiliza la siguiente línea de código para imputar estos valores:

```
#Imputar Country y State mediante su relación
df['Country'].fillna(df['Country'].mode()[0], inplace=True)
df['State'].fillna(df['State'].mode()[0], inplace=True)
```

De manera similar, *Product Category* y *Sub Category* podrían estar relacionadas con otros datos, como *Cost* o *Quantity*, por lo que también se realiza la imputación con el siguiente código:

```
[145] #Imputar Product Category y Subcategory
      df['Product Category'].fillna(df['Product Category'].mode()[0], inplace=True)
      df['Sub Category'].fillna(df['Sub Category'].mode()[0], inplace=True)
```

En cuanto a los valores numéricos, se debe tener especial cuidado al imputar. Para la columna *Quantity*, que es una variable discreta, se utiliza la mediana para la imputación, ya que se espera que los valores no sean negativos y que su dispersión sea limitada. El código utilizado para imputar la mediana es el siguiente:

```
[146] #Imputar Quantity
      df['Quantity'].fillna(df['Quantity'].median(), inplace=True)
```

Para columnas como *Unit Cost, Unit Price, Cost* y *Revenue*, que son valores numéricos interdependientes, si falta uno de los valores, se puede calcular utilizando los demás. Si no es posible realizar este cálculo, se utiliza la mediana como alternativa para imputar el valor faltante.

```
df['Unit Cost'].fillna(df['Unit Cost'].median(), inplace=True)
df['Unit Price'].fillna(df['Unit Price'].median(), inplace=True)
df['Cost'].fillna(df['Cost'].median(), inplace=True)
df['Revenue'].fillna(df['Revenue'].median(), inplace=True)
```

Finalmente, para la columna *Index*, se puede reemplazar los valores faltantes por un rango numérico secuencial, lo que reiniciará el índice basado en el orden de las filas. Sin embargo, si el índice no es secuencial, como en este caso, la única opción disponible es eliminar los registros con valores faltantes. El código para hacerlo es el siguiente:

```
[151] #Índice
      df = df.dropna(axis=1)
```

Verificación Final

Una vez realizados todos los pasos anteriores, se verifica que ya no haya datos faltantes en la base de datos. Para ello, se utiliza la siguiente línea de código:

```
[152] # Columnas con total de valores nulos
      print(df.isnull().sum())

Date                0
Year                0
Customer Age        0
Customer Gender     0
Country             0
State               0
Product Category    0
Sub Category        0
Quantity            0
Unit Cost           0
Unit Price          0
Cost                0
Revenue             0
dtype: int64
```

Finalmente, se comprueba la extensión final de la base de datos para confirmar que todos los cambios se hayan aplicado correctamente:

```
[153] df.shape

(41244, 13)
```

Y se muestran las estadísticas finales:

```
#Describir las estadísticas de forma final
print(df.describe(include='all'))

              Date     Year  Customer Age Customer Gender        Country  \
count        41244  41244.0  41244.000000           41244          41244
unique         576      3.0           NaN               2              4
top        Unknown   2016.0           NaN               M  United States
freq          2185  22735.0           NaN           21790          22025
mean           NaN      NaN     36.319368             NaN            NaN
std            NaN      NaN     10.825552             NaN            NaN
min            NaN      NaN     17.000000             NaN            NaN
25%            NaN      NaN     28.000000             NaN            NaN
50%            NaN      NaN     35.000000             NaN            NaN
75%            NaN      NaN     43.000000             NaN            NaN
max            NaN      NaN     87.000000             NaN            NaN
```

	State	Product Category	Sub Category	Quantity
count	41244	41244	41244	41244.000000
unique	46	4	18	NaN
top	California	Accessories	Tires and Tubes	NaN
freq	12965	26635	13769	NaN
mean	NaN	NaN	NaN	2.001649
std	NaN	NaN	NaN	0.799945
min	NaN	NaN	NaN	1.000000
25%	NaN	NaN	NaN	1.000000
50%	NaN	NaN	NaN	2.000000
75%	NaN	NaN	NaN	3.000000
max	NaN	NaN	NaN	3.000000

	Unit Cost	Unit Price	Cost	Revenue
count	41244.000000	41244.000000	41244.000000	41244.000000
unique	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN
mean	338.063005	376.903311	564.107167	629.049984
std	475.917397	511.912443	679.589018	726.201449
min	0.670000	0.666667	2.000000	2.000000
25%	48.000000	57.500000	90.000000	106.000000
50%	150.000000	178.000000	261.000000	319.000000
75%	418.000000	487.000000	769.000000	873.000000
max	3240.000000	5082.000000	3600.000000	5082.000000

Conclusiones mediante la estadística

El resultado muestra que el conjunto de datos describe transacciones de clientes con información sobre el año, edad, género, ubicación geográfica, y detalles financieros de productos como cantidad, costo y precio. La mayoría de las transacciones corresponden a 2016, con una edad promedio de 36 años y una mayor presencia masculina. Los productos más vendidos son de la categoría "Accesorios" y la subcategoría "Neumáticos y Tubos", lo que sugiere que estos productos son clave en las ventas.

Los precios unitarios varían considerablemente, lo que refleja una oferta de productos desde económicos hasta de mayor precio. La mayoría de las compras son pequeñas, con 1 a 3 unidades adquiridas por transacción. Esto indica que los clientes compran productos según necesidades específicas.

Se puede predecir que los "Accesorios" seguirán siendo populares, pero también es importante explorar estrategias para productos más caros, como los "Neumáticos", adaptándose a la diversidad de clientes en edad, ubicación y preferencias.

Metodología

Para garantizar un análisis riguroso y fundamentado del conjunto de datos de ventas, se diseñó una metodología estructurada que abarcó desde la búsqueda y selección de la base de datos hasta su limpieza, exploración y análisis detallado. Cada etapa fue cuidadosamente planificada para transformar un gran volumen de datos en información procesable, útil para responder a las preguntas clave del proyecto y aportar valor estratégico al análisis. Este enfoque metodológico se apoyó en herramientas avanzadas y prácticas de manejo de datos que aseguraron la precisión, integridad y representatividad de los resultados.

Conocimiento de los datos y su manejo

El primer paso fue familiarizarse con la base de datos seleccionada, la cual fue obtenida de la plataforma Kaggle. Esta base contenía información detallada sobre transacciones de ventas, incluyendo variables clave como la edad y género de los consumidores, su ubicación geográfica, las categorías y subcategorías de productos adquiridos, cantidades compradas, costos unitarios y totales, entre otros. La descripción proporcionada en la plataforma indicaba que el dataset ofrecía una visión integral y con potencial para análisis en múltiples niveles.

En esta etapa, se realizó un análisis preliminar para evaluar la estructura del dataset. Esto incluyó:

- Dimensiones de la base de datos: Determinar el número de registros y columnas.
- Tipos de datos: Identificar variables categóricas, numéricas, temporales y su formato correspondiente.
- Relaciones entre variables: Explorar correlaciones iniciales entre los campos clave para establecer su relevancia en las preguntas de investigación planteadas.
- Identificación de valores únicos y distribuciones: Examinar la diversidad de datos en campos categóricos como género, país, estado, y categorías de productos.

Con base en este análisis, se clasificaron las variables según su importancia relativa en función de las hipótesis iniciales y las preguntas clave planteadas. Aunque todas las variables se consideraron relevantes, aquellas relacionadas con ingresos, costos, productos y datos demográficos fueron priorizadas debido a su impacto directo en los objetivos del análisis. Además, se identificaron variables que podrían complementar la interpretación, como las categorías de productos y las ubicaciones geográficas, permitiendo múltiples enfoques en las etapas posteriores.

Limpieza de los datos

La calidad del análisis depende directamente de la calidad de los datos utilizados. Durante la revisión inicial del dataset, se identificaron diversos problemas comunes en bases de datos grandes, como valores faltantes, duplicados, errores en formatos y datos inconsistentes. Para abordar estos problemas, se implementó un proceso exhaustivo de limpieza, dividido en las siguientes etapas:

- Identificación de datos faltantes: Se revisaron todas las columnas en busca de registros incompletos. Cuando fue posible, se realizaron imputaciones basadas en patrones observados en el dataset. Por ejemplo, para datos numéricos, se utilizaron promedios o medianas según el contexto, mientras que para datos categóricos se completaron valores según frecuencias predominantes o relaciones con otras variables.
- Eliminación de duplicados: Se detectaron registros duplicados mediante un análisis de claves únicas, como combinaciones de fecha, cliente y producto. Estos registros fueron revisados manualmente para determinar si eran redundantes o contenían información adicional antes de ser eliminados.
- Corrección de formatos y estándares: Se estandarizaron los formatos de fechas, unidades monetarias y nombres de categorías para garantizar la consistencia en el análisis. Por ejemplo, las fechas se unificaron en un formato común y las categorías de productos se agruparon según descripciones homogéneas.
- Identificación y manejo de valores atípicos: Se analizaron distribuciones numéricas para detectar valores fuera de rango (e.g., costos unitarios o cantidades irreales). Los valores anómalos se validaron para confirmar si eran errores o representaban casos especiales relevantes.
- Preservación de la información: A lo largo de todo el proceso, se priorizó conservar la mayor cantidad de información posible, minimizando la eliminación de registros y asegurando que los datos procesados reflejaran la realidad del conjunto inicial.

Como resultado de estas acciones, la base de datos quedó lista para un análisis confiable y exhaustivo, libre de inconsistencias que pudieran sesgar los resultados.

Exploración de los datos y establecimiento de enfoques de análisis

Con los datos limpios y organizados, se procedió a realizar un análisis exploratorio de datos. Este análisis permitió identificar patrones iniciales, relaciones clave entre variables y posibles áreas de interés. Algunas acciones realizadas en esta etapa incluyeron:.

- Análisis estadístico descriptivo: Cálculo de medidas de tendencia central y dispersión para variables numéricas clave (e.g., ingresos, costos, edades).
- Visualización de datos: Uso de gráficos como histogramas, diagramas de dispersión, mapas de calor y diagramas de boxplot para interpretar distribuciones, relaciones y posibles agrupamientos de datos.
- Segmentación inicial: Identificación de grupos preliminares según criterios como ubicación geográfica, categorías de productos, o perfiles demográficos de clientes.

Estos pasos sirvieron para estructurar un enfoque analítico que abordara las preguntas clave del proyecto, permitiendo la identificación de tendencias y relaciones que guiaran las fases posteriores.

Análisis final de la base de datos

Tras completar el proceso de limpieza de la base de datos, se modificaron algunos enfoques iniciales para adaptarse al conjunto de datos en su forma final. La limpieza no solo mejoró la calidad de los datos, sino que también permitió un análisis más preciso y confiable. En esta sección, se presentan las características clave de la base de datos limpia, comenzando por su extensión, pasando por una descripción de sus variables, y finalizando con un análisis general de las estadísticas descriptivas:

```
[161] df.shape

(41244, 13)
```

```
[160] df
```

	Date	Year	Customer Age	Customer Gender	Country	State	Product Category	Sub Category	Quantity	Unit Cost	Unit Price	Cost	Revenue
0	2016-02-19 00:00:00	2016.0	29.0	F	United States	invalid	Accessories	Tires and Tubes	1.0	80.00	109.000000	80.0	109.0
1	2016-02-20 00:00:00	Unknown	29.0	F	United States	Washington	Clothing	Gloves	2.0	24.50	28.500000	49.0	57.0
2	2016-02-27 00:00:00	2016.0	29.0	F	United States	Washington	Accessories	Tires and Tubes	3.0	3.67	5.000000	11.0	15.0
3	2016-03-12 00:00:00	2016.0	29.0	F	United States	Washington	Accessories	Tires and Tubes	2.0	87.50	116.500000	175.0	233.0
4	2016-03-12 00:00:00	2016.0	29.0	F	United States	Washington	Accessories	Tires and Tubes	3.0	35.00	41.666667	105.0	125.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
44605	2016-02-28 00:00:00	2016.0	37.0	M	United States	California	Clothing	Gloves	3.0	49.00	63.666667	147.0	191.0
44606	2016-02-02 00:00:00	2016.0	18.0	M	United States	California	Accessories	Cleaners	1.0	167.00	213.000000	167.0	213.0
44608	2016-04-17 00:00:00	2016.0	21.0	M	United States	California	Clothing	Gloves	1.0	416.00	477.000000	416.0	319.0
44609	2016-01-23 00:00:00	2016.0	31.0	F	United States	California	Accessories	Bottles and Cages	2.0	47.50	60.000000	95.0	120.0
44610	2015-09-09 00:00:00	2015.0	58.0	F	United States	California	Accessories	invalid	3.0	70.00	87.000000	210.0	261.0

41244 rows × 13 columns

```
[162] df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 41244 entries, 0 to 44610
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Date              41244 non-null  object
 1   Year              41244 non-null  object
 2   Customer Age      41244 non-null  float64
 3   Customer Gender   41244 non-null  object
 4   Country           41244 non-null  object
 5   State             41244 non-null  object
 6   Product Category  41244 non-null  object
 7   Sub Category      41244 non-null  object
 8   Quantity          41244 non-null  float64
 9   Unit Cost         41244 non-null  float64
 10  Unit Price        41244 non-null  float64
 11  Cost              41244 non-null  float64
 12  Revenue           41244 non-null  float64
dtypes: float64(6), object(7)
memory usage: 4.4+ MB
```

El conjunto de datos incluye un total de 41,244 registros distribuidos en 13 columnas, cada una representando variables esenciales relacionadas con transacciones de ventas. Todas las columnas ahora

están completas, sin valores nulos, lo que garantiza una integridad óptima del conjunto de datos. Esta condición elimina la necesidad de realizar imputaciones o eliminar registros, lo que fortalece la fiabilidad del análisis posterior. Las variables abarcan datos numéricos, categóricos y temporales, lo que permite una amplia gama de enfoques analíticos.

El tamaño del dataset, de 4.4 MB en memoria, indica un volumen sustancial de información que es lo suficientemente robusto para realizar análisis descriptivos, segmentaciones y modelos predictivos. Esta riqueza de datos es particularmente valiosa para estudiar tendencias, patrones de compra y comportamientos demográficos de los clientes, así como para realizar comparaciones entre regiones o productos.

La combinación de variables numéricas y categóricas es una ventaja clave, ya que permite realizar análisis cruzados que pueden aportar insights estratégicos. Por ejemplo, se pueden identificar las categorías de productos más rentables en ciertas regiones geográficas, o analizar los perfiles demográficos de los clientes que generan mayores ingresos.

Al observar las estadísticas descriptivas de las variables principales, se pueden extraer varias conclusiones preliminares.

```
[164] df.describe()
```

	Customer Age	Quantity	Unit Cost	Unit Price	Cost	Revenue
count	41244.000000	41244.000000	41244.000000	41244.000000	41244.000000	41244.000000
mean	36.319368	2.001649	338.063005	376.903311	564.107167	629.049984
std	10.825552	0.799945	475.917397	511.912443	679.589018	726.201449
min	17.000000	1.000000	0.670000	0.666667	2.000000	2.000000
25%	28.000000	1.000000	48.000000	57.500000	90.000000	106.000000
50%	35.000000	2.000000	150.000000	178.000000	261.000000	319.000000
75%	43.000000	3.000000	418.000000	487.000000	769.000000	873.000000
max	87.000000	3.000000	3240.000000	5082.000000	3600.000000	5082.000000

En términos demográficos, la edad promedio de los clientes es de 36.3 años, con un rango que va de 17 a 87 años, lo que demuestra una diversidad significativa en la base de clientes. Respecto al comportamiento de compra, la cantidad media de productos adquiridos por transacción es de aproximadamente 2 unidades, con un rango limitado de 1 a 3 unidades, lo que sugiere compras de bajo volumen, pero posiblemente frecuentes.

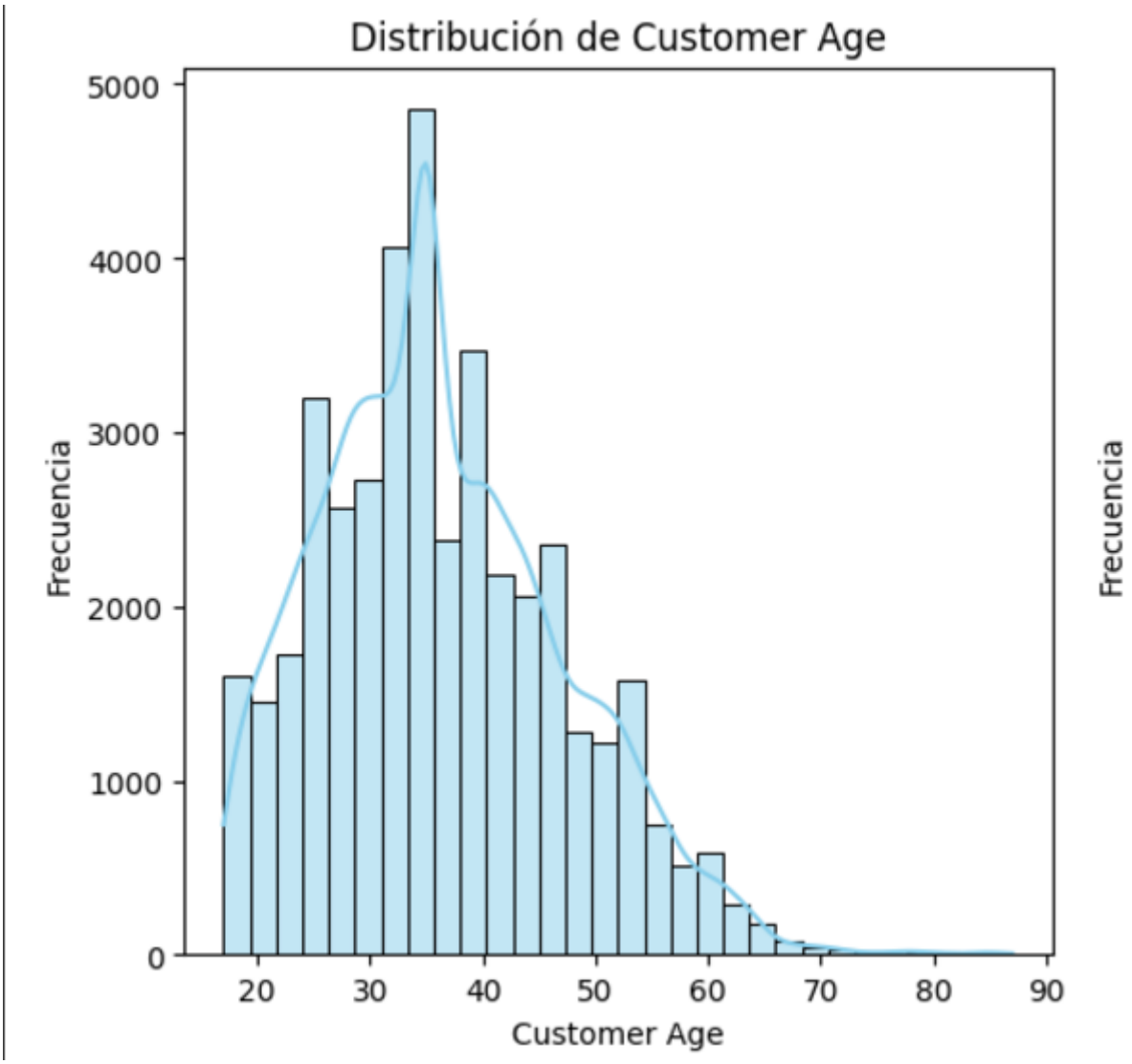
En el aspecto financiero, se observa una alta variabilidad en los costos y precios. El costo unitario promedio es de 338.06, mientras que el precio unitario promedio es de 376.90, dejando un margen de ganancia general moderado. Los costos totales por transacción tienen una media de 564.10, mientras que los ingresos promedios alcanzan los 629.05. Sin embargo, los valores máximos, que llegan a 3600 en costos y 5082 en ingresos, reflejan la existencia de transacciones asociadas a productos de mayor valor.

Finalmente, al profundizar en estas estadísticas y apoyándonos en visualizaciones generadas a partir de los datos, se pueden abordar de manera fundamentada las hipótesis iniciales planteadas. Los gráficos derivados del análisis descriptivo permitirán una interpretación más detallada y visual de las tendencias y patrones presentes en el conjunto de datos.

Visualización de variables en gráficos

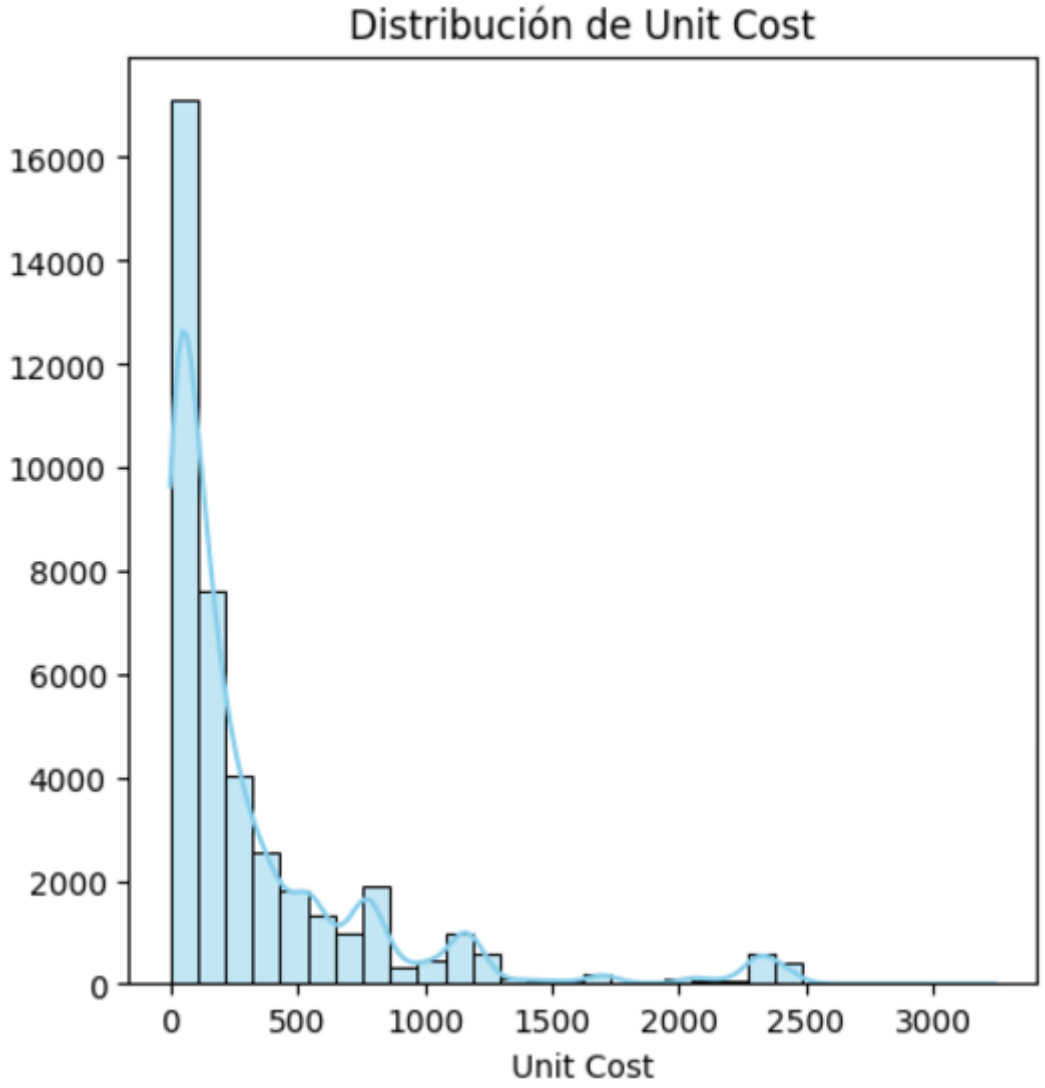
Método Histograma (Variables Individuales)

El análisis de los histogramas permitirá observar patrones clave en la distribución de las variables, como sesgos, rangos comunes de valores y posibles outliers. Estos resultados están relacionados con las preguntas clave sobre el comportamiento de los clientes según edad y su preferencia por productos de diferentes costos unitarios., relacionado con las preguntas sobre el comportamiento por edades y preferencia por costos unitarios

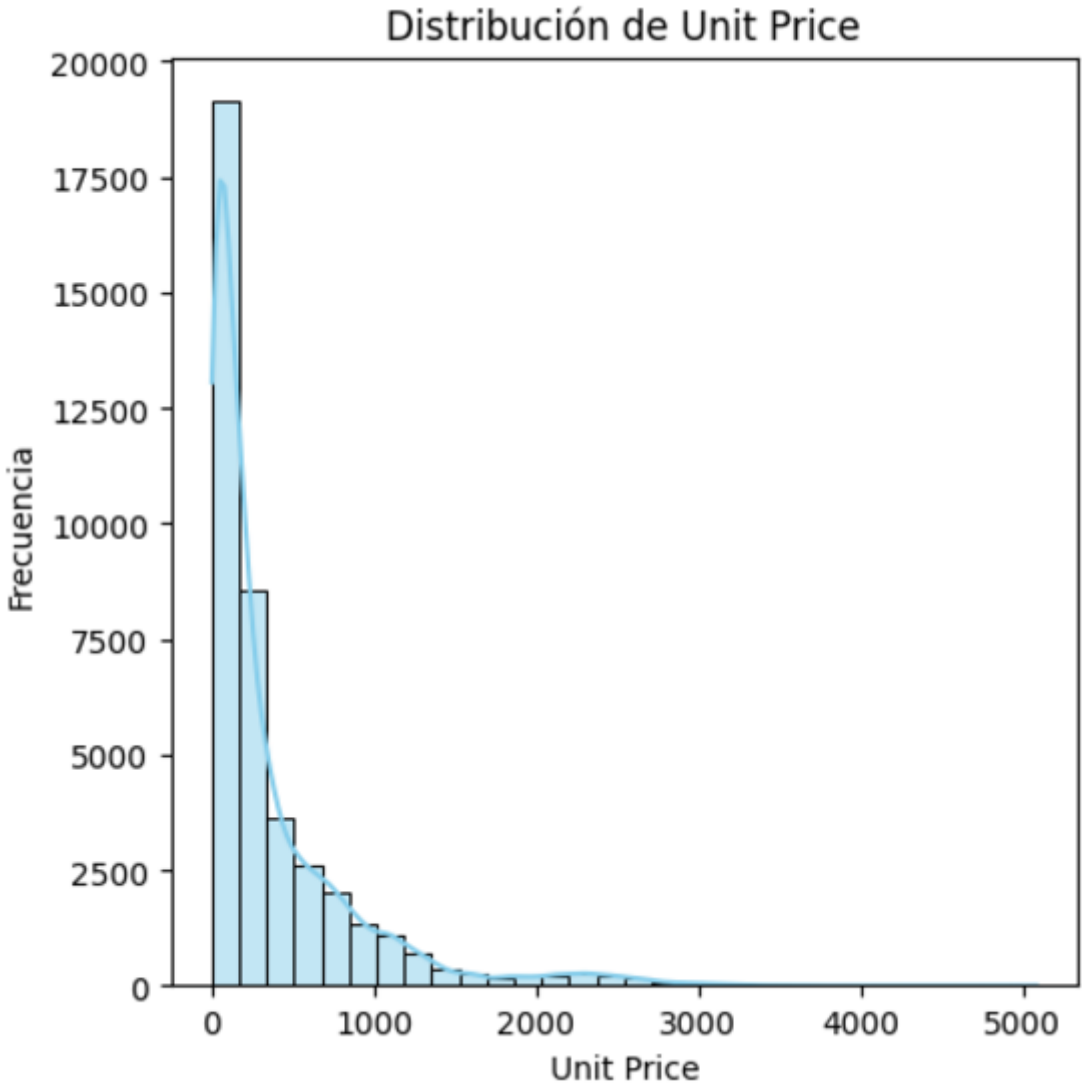


- Customer Age: La media de la edad de los clientes es de aproximadamente 36.3 años, lo que sugiere que la mayoría de los clientes se encuentran en un rango de edad que abarca desde los jóvenes adultos hasta los adultos de mediana edad. El sesgo hacia la derecha (positivo) indica que hay más clientes en el rango de edades más altas, lo que podría implicar que los productos atraen especialmente a un público más maduro. Sin embargo, si se observa una concentración

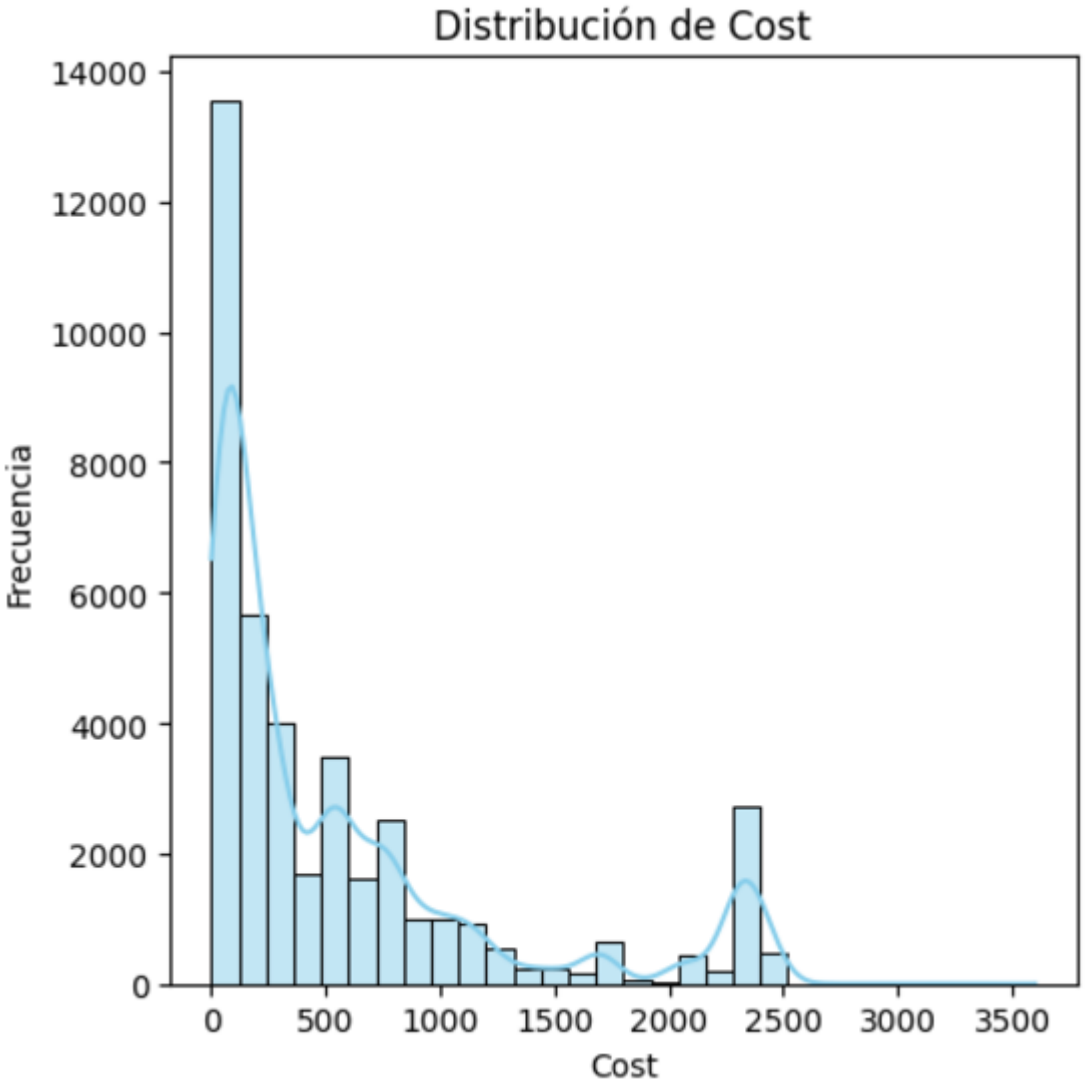
significativa en los rangos de edad más bajos, podría interpretarse que los productos tienen una mayor atracción entre los clientes jóvenes.



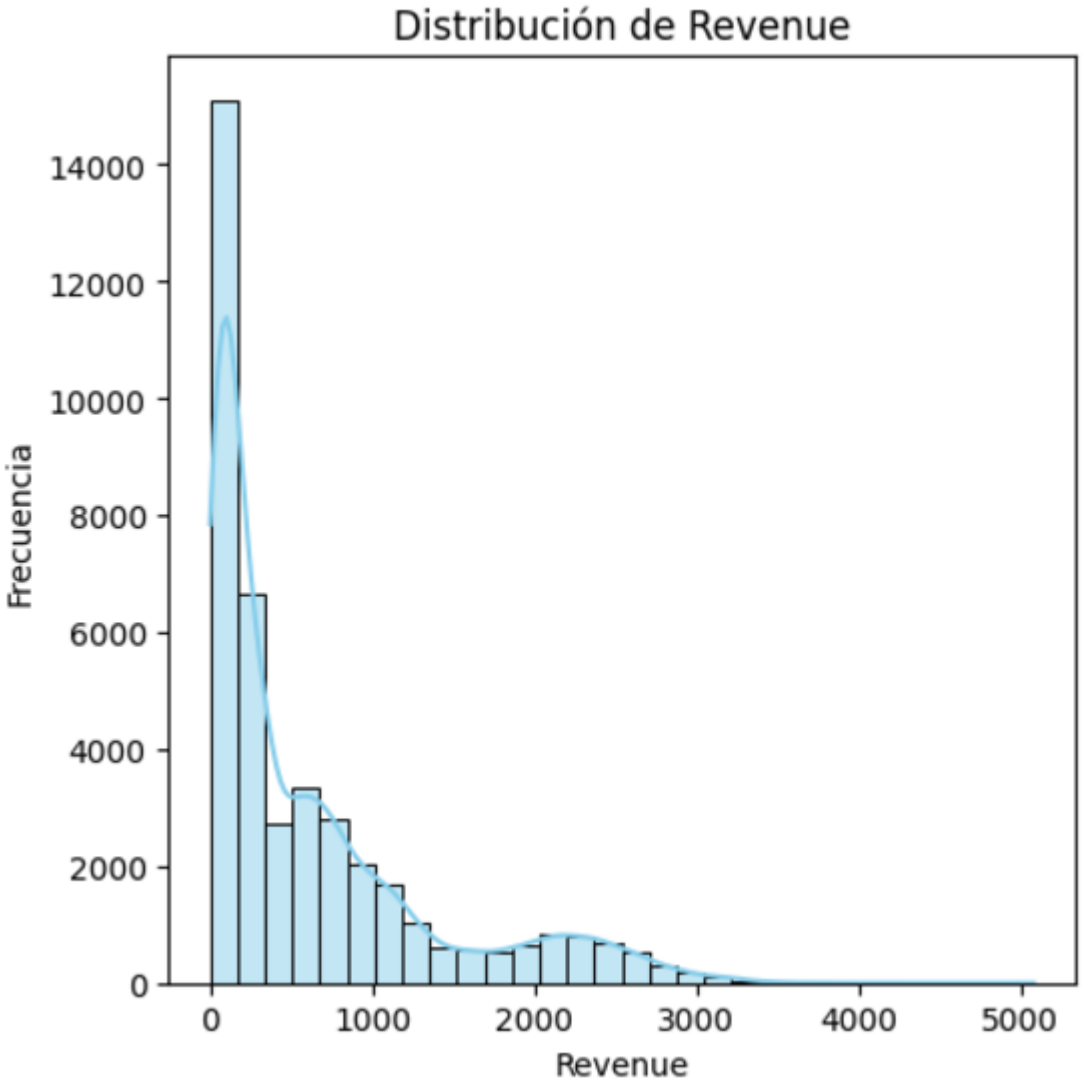
- Unit Cost: El costo unitario tiene una media de 338.06, acompañado de una gran desviación estándar de 475.92, lo que sugiere que existen productos con costos muy bajos y otros con costos muy altos. Esta variabilidad es típica en mercados que ofrecen tanto productos de consumo masivo (más baratos) como productos de alta gama (más caros). El histograma muestra que la mayoría de los productos tienen un costo bajo, lo que sugiere que la estrategia de precios podría estar orientada hacia productos de bajo costo pero con un alto volumen de ventas.



- Unit Price: El precio unitario tiene una media de 376.90, que es superior al costo unitario, lo que indica que, en general, los productos se venden con un margen de ganancia positivo. El sesgo hacia la derecha en el histograma puede sugerir que los productos de menor precio son los que tienen mayor demanda, lo que podría implicar que los consumidores están más dispuestos a comprar productos accesibles, aunque la estrategia de precios parece estar equilibrada para generar márgenes positivos.



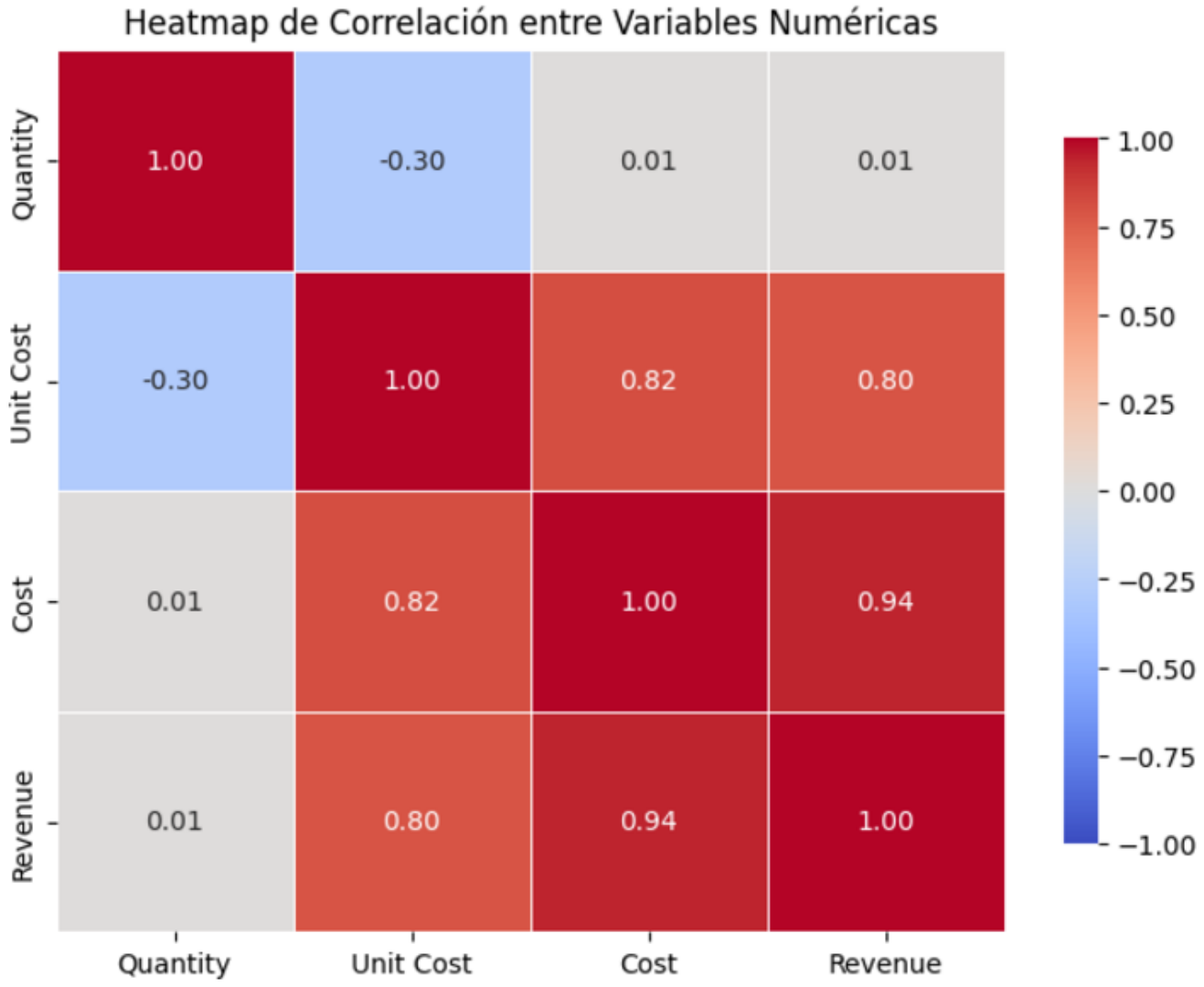
- Cost: El costo total tiene una media de 564.11, pero con una desviación estándar significativa de 679.59, lo que indica que los costos totales varían considerablemente. Esto está relacionado con la cantidad de productos vendidos y el costo unitario de cada producto. El histograma revela picos en ciertos rangos de compras, como las ventas grandes de bicicletas o accesorios de alto costo, lo que genera variaciones notables en los costos totales.



- Revenue**:** Los ingresos (Revenue) tienen una media de 629.05, lo que indica que el ingreso promedio por transacción es relativamente alto, pero con una gran desviación estándar de 726.20, lo que señala una notable variabilidad en los ingresos. Esto sugiere que, aunque la mayoría de las transacciones generan ingresos moderados, hay algunas compras de alto valor que incrementan significativamente los ingresos. Los picos en el histograma reflejan que algunas ventas clave son cruciales para la rentabilidad general del negocio.

Método Heatmaps (Relaciones entre Variables)

El heatmap visualiza las correlaciones entre las variables numéricas seleccionadas del conjunto de datos: *Quantity, Unit Cost, Cost y Revenue*.



- Quantity:

  Quantity vs. Cost: Se espera una correlación positiva, moderada o alta (por ejemplo, >0.5), ya que, a medida que aumenta la cantidad comprada, el costo total también crece. Esto tiene sentido, ya que el costo total es proporcional a la cantidad adquirida.

  Quantity vs. Revenue: Una correlación positiva moderada (entre 0.5 y 0.8) sugiere que a mayor cantidad de unidades vendidas, mayores serán los ingresos. No obstante, esta relación podría debilitarse si el precio unitario varía significativamente entre los productos.

Quantity vs. Unit Cost: Es probable que la correlación sea débil o nula (cercana a 0), ya que el costo unitario generalmente no depende directamente del número de unidades adquiridas.

- Unit Cost:

  Unit Cost vs. Cost: Se espera una correlación positiva fuerte (cercana a 1), ya que el costo total se calcula multiplicando el costo unitario por la cantidad comprada.

  Unit Cost vs. Revenue: Una correlación positiva indica que los productos con mayor costo unitario suelen generar mayores ingresos. Sin embargo, esta relación puede no ser lineal, dependiendo de factores como la elasticidad del precio y la demanda.

  Unit Cost vs. Quantity: Una correlación negativa, débil o moderada (por ejemplo, entre -0.3 y -0.5), podría sugerir que los productos más costosos suelen venderse en menores cantidades, lo que es común en mercados con sensibilidad al precio.

- Cost:

  Cost vs. Revenue: Se espera una correlación positiva fuerte (por ejemplo, >0.8), ya que, en general, el costo total aumenta en proporción con los ingresos. Esto es típico en modelos de negocio donde el margen de ganancia por unidad es constante.

  Cost vs. Quantity: Una correlación positiva moderada o alta refleja que el costo total está directamente influenciado por la cantidad de productos vendidos.

  Cost vs. Unit Cost: La correlación también debería ser alta, ya que el costo total depende directamente del costo unitario.

- Revenue:

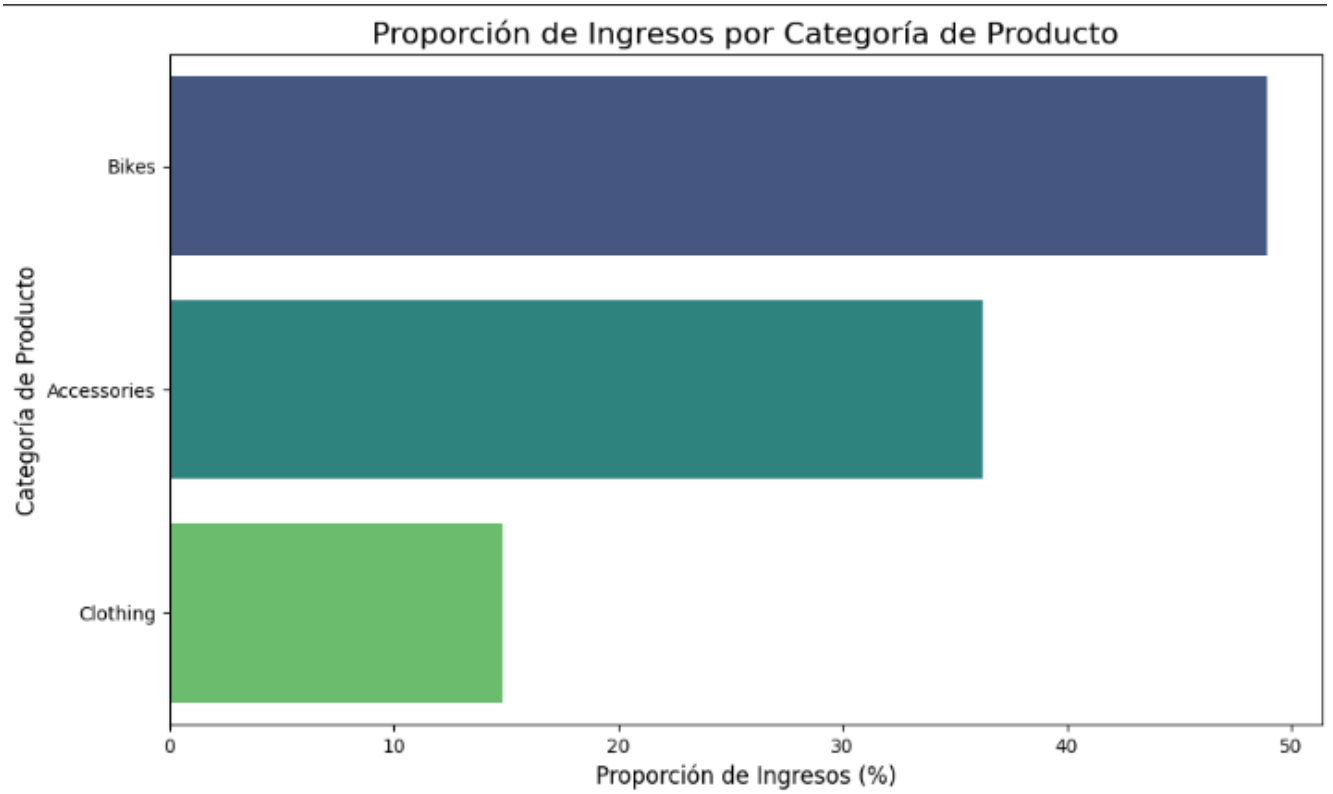
  Revenue vs. Quantity: Una correlación positiva moderada o alta indica que un mayor número de unidades vendidas genera mayores ingresos, lo que es común en modelos de ventas estándar.

  Revenue vs. Unit Cost: Una correlación positiva podría implicar que los productos más costosos tienden a generar mayores ingresos, aunque esta relación depende de las cantidades vendidas y del segmento de clientes.

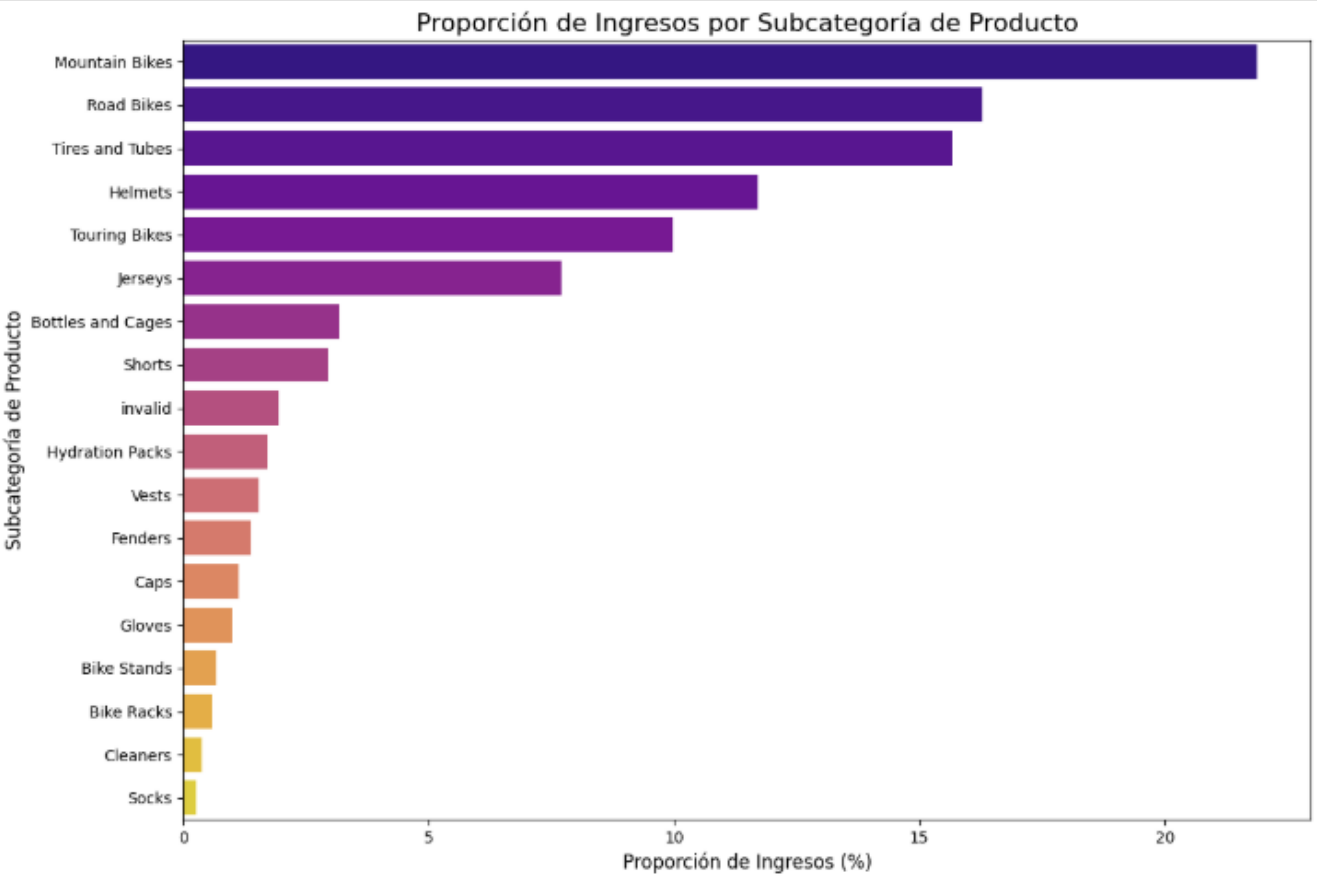
  Revenue vs. Cost: Una correlación cercana a 1 sería esperada, ya que los ingresos están intrínsecamente ligados al costo total, especialmente en sistemas de precios consistentes.

Método Gráfica de Barras (Categorías/Subcategorías)

La visualización mediante gráficas de barras permite observar la proporción de ingresos por categoría y subcategoría de producto, lo que ayuda a responder preguntas clave sobre la rentabilidad de cada grupo.



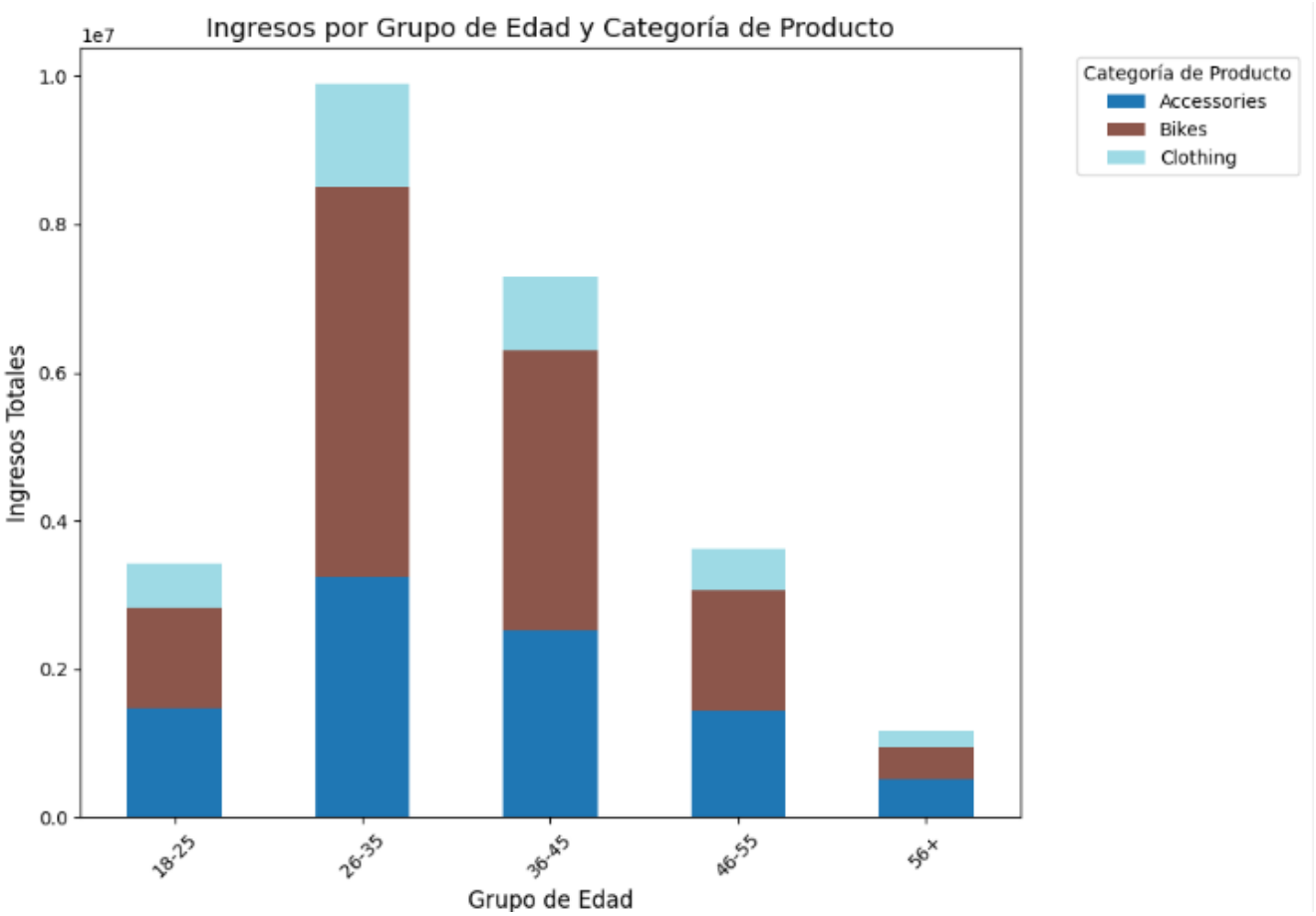
- Product Category: En la gráfica, la categoría de *Bikes* destaca claramente con la barra más alta, lo que confirma que esta categoría genera la mayor proporción de ingresos. Este comportamiento podría estar relacionado con sus altos costos unitarios y márgenes de ganancia, lo que la convierte en una categoría especialmente rentable. En cambio, categorías como *Clothing* y *Accessories* presentan barras más cortas, lo que sugiere que sus ingresos son relativamente menores en comparación. Esto podría deberse a precios más bajos en estas categorías o, alternativamente, a volúmenes de venta limitados en algunos de sus productos.



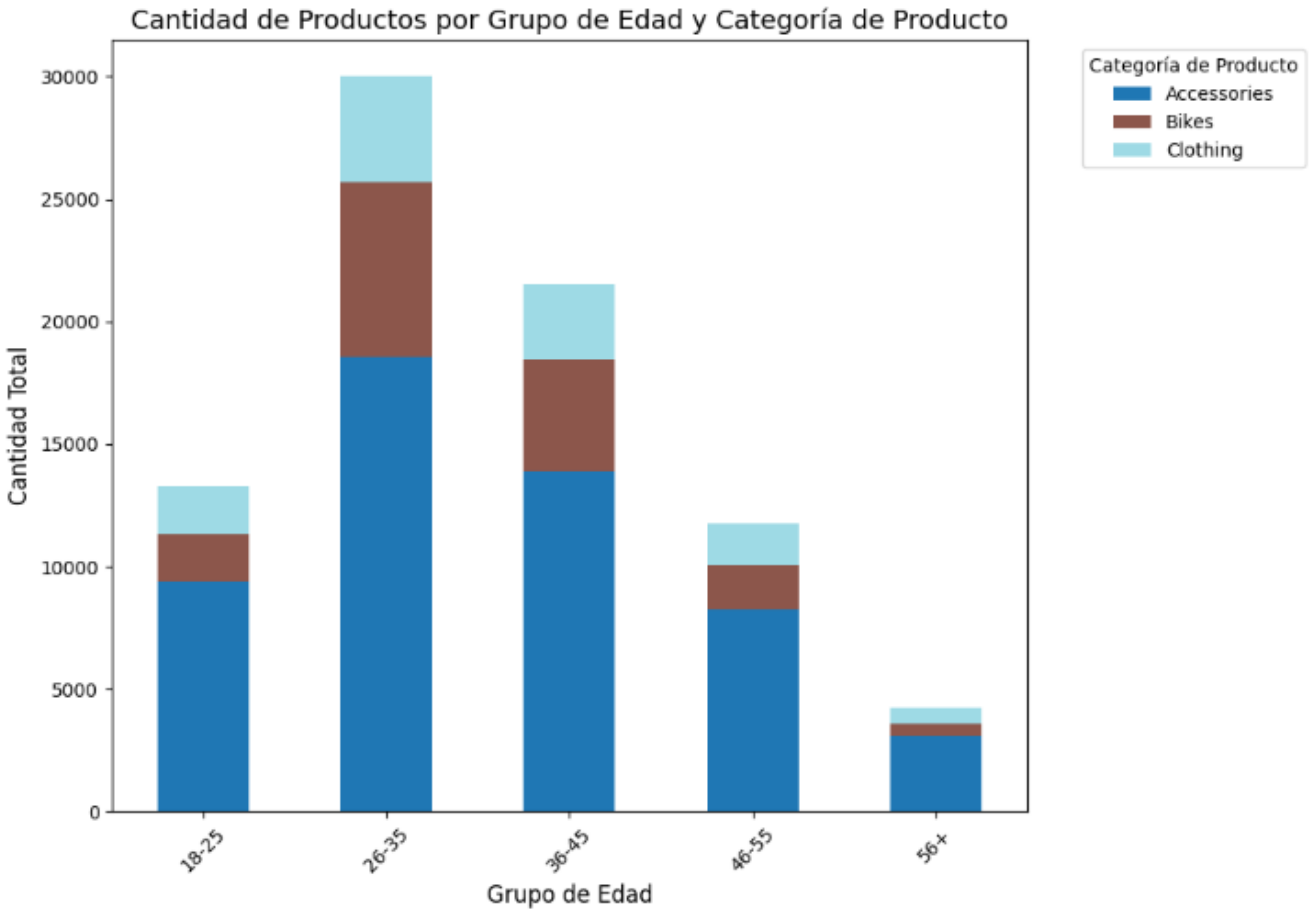
- Subcategoría de Producto: Al desglosar los ingresos por subcategoría, se puede obtener un análisis más detallado. Por ejemplo, dentro de la categoría *Bikes*, subcategorías como *Mountain Bikes* y *Road Bikes* representan una proporción significativa de los ingresos, lo que resalta su popularidad y/o el precio elevado de estos productos. Por otro lado, en la categoría *Accessories*, subcategorías como *Socks* y *Caps* muestran barras más cortas, lo que indica que tienen un impacto menor en los ingresos totales. Este comportamiento puede reflejar una menor demanda o precios más accesibles, lo que resulta en ingresos más modestos en comparación con otros productos de mayor precio.

Método Gráfica de Barras Stacked (Categorías/Subcategorías)

Este tipo de gráfico es útil para comparar ingresos o cantidades entre diferentes grupos categóricos, como la edad de los clientes y las categorías de productos.

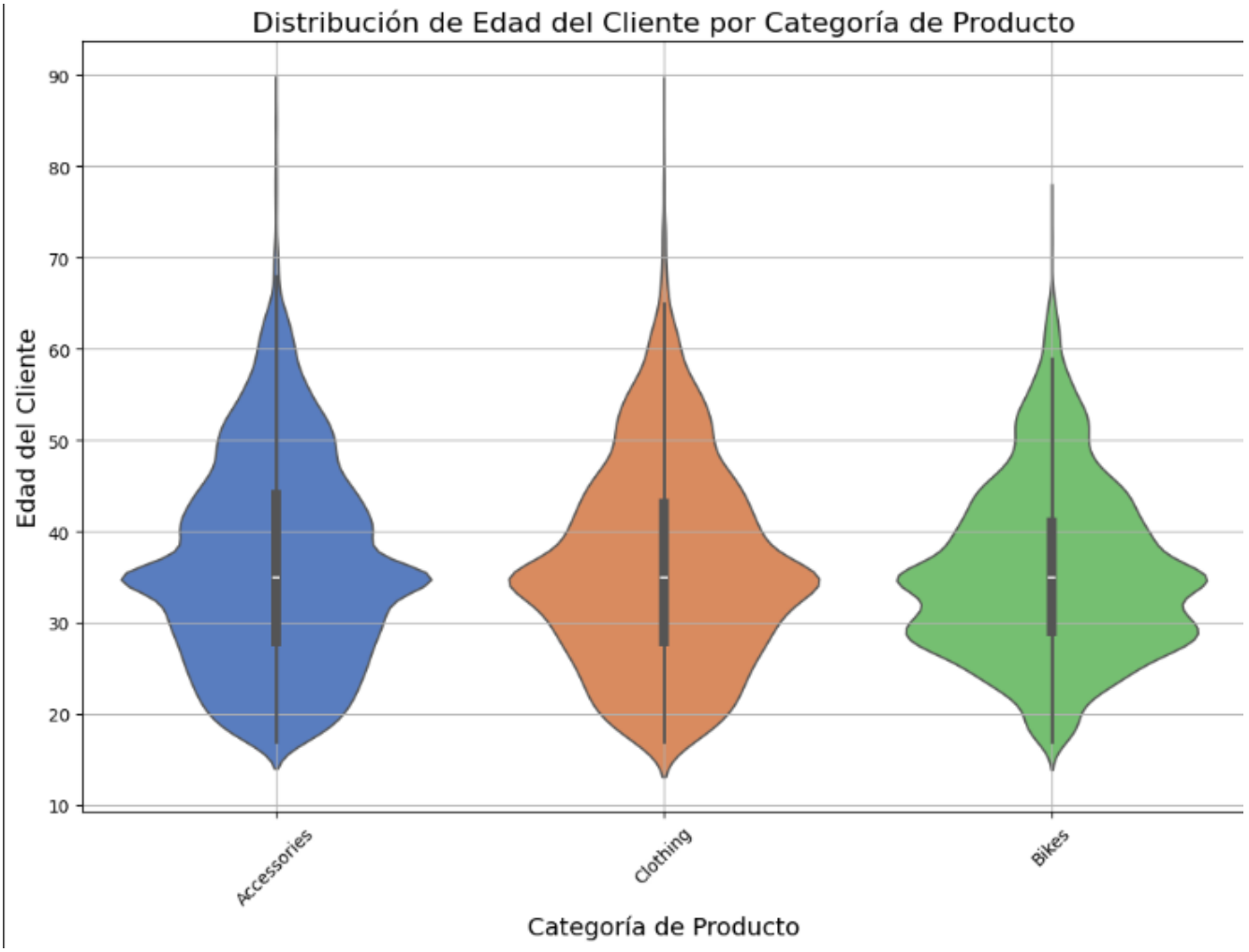


- Edad y Categoría de Producto por Ingresos Totales: Los segmentos correspondientes a la categoría *Bikes* son más prominentes en los grupos de edad mayores, lo que sugiere que estos consumidores prefieren productos de mayor costo unitario, como las bicicletas. En contraste, categorías como *Clothing* y *Accessories* tienen una mayor representación en los grupos de edad más jóvenes y en los más adultos (los extremos del espectro etario). Esto indica que los consumidores de estas edades tienden a gastar más en productos accesibles y de menor costo unitario, como ropa y accesorios, en lugar de productos de mayor precio.



-
- Edad y Categoría de Producto por Cantidad Total: Los adultos jóvenes tienden a realizar compras más frecuentes, lo que se refleja en un mayor número de unidades adquiridas en categorías como *Clothing* y *Accessories*. Por otro lado, los grupos de edad mayores podrían comprar menos unidades en general, pero tienden a optar por productos de mayor valor, como *Bikes* o ropa de gama alta. Esto sugiere que, aunque la frecuencia de compra puede ser menor en estos grupos, el gasto por transacción es significativamente mayor, lo que refleja una estrategia de compra diferente basada en productos de mayor precio.

Método Violin Plots (Segmentación Demográfica)



En las categorías *Accessories* y *Clothing*, se observa que la mayoría de los consumidores se encuentran en el rango de edad entre los 30 y 40 años. Esto sugiere que estos productos atraen principalmente a personas de mediana edad. En cambio, en la categoría *Bikes*, la mayor concentración de consumidores se encuentra también entre los 30 y 40 años, pero con una segunda concentración notable alrededor de los 30 años, lo que indica que hay un interés significativo en bicicletas en este grupo de edad más joven.

Este análisis es crucial para comprender cómo la edad influye en las decisiones de compra según la categoría de producto. Además, puede ser útil para realizar ajustes en la oferta de productos o en las estrategias de marketing, asegurando que se alineen mejor con las preferencias demográficas de los clientes.

Verificación de hipótesis

Para verificar si las hipótesis fueron acertadas hay que usar la interpretación de los gráficos y con base en ellos dar respuesta

1) **Relación entre la categoría del producto, el ingreso generado y el costo unitario**

   Hipótesis: Los productos de la categoría "Bikes" generan mayores ingresos promedio debido a su alto costo unitario, en comparación con las categorías "Clothing" y "Accessories"

   Evaluación:

   Los resultados de las gráficas de barras y el análisis del costo unitario y los ingresos sugieren que la categoría *Bikes* efectivamente genera la mayor proporción de ingresos. Esto es consistente con el alto costo unitario de las bicicletas, lo que se refleja en una mayor rentabilidad. Las categorías *Clothing* y *Accessories*, por otro lado, presentan ingresos más bajos, lo que se puede atribuir a sus precios más accesibles o menores volúmenes de venta. En el análisis del heatmap también se observó una correlación positiva entre *Unit Cost* y *Revenue*, lo que refuerza la hipótesis de que productos con mayor costo unitario (como las bicicletas) tienden a generar mayores ingresos.

   Conclusión: Sí se cumple la hipótesis, ya que las gráficas y los análisis sugieren que las bicicletas son más rentables debido a su alto costo unitario.

2) **Impacto de la edad del cliente en las compras considerando subcategorías y costo unitario**

   Hipótesis: Los clientes más jóvenes (menores de 35 años) prefieren productos de subcategorías con menor costo unitario, como "Socks" y "Caps" en la categoría de accesorios, mientras que los clientes mayores al rango de edad establecido optan por subcategorías con mayor costo unitario, como "Mountain Bikes" y "Road Bikes" en la categoría de bicicletas

   Evaluación: Los análisis de los histogramas y los *Violin Plots* muestran una mayor concentración de consumidores jóvenes en la categoría *Accessories*, particularmente en subcategorías como *Socks* y *Caps*, que tienen un costo unitario más bajo. Este patrón sugiere que los clientes más jóvenes prefieren productos más accesibles y baratos, como lo predice la hipótesis. Por otro lado, los consumidores mayores se inclinan más por productos de alto costo, como *Mountain Bikes* y *Road Bikes*, lo cual se alinea con la hipótesis de que los clientes mayores gastan más en productos de mayor valor.

   Conclusión: Sí se cumple la hipótesis, ya que los resultados de los gráficos indican que los

clientes jóvenes prefieren productos más baratos y los mayores optan por artículos más costosos.

3) **Diferencias geográficas en la preferencia de productos**

Hipótesis: Los clientes en estados europeos tienden a una mayor preferencia por productos de ropa y accesorios en comparación con clientes en Estados Unidos, que prefieren bicicletas.

Evaluación: Aunque no se proporcionan gráficos específicamente geográficos en la descripción, se puede inferir la validez de la hipótesis a partir de los patrones observados en las categorías de productos. Si se tuviera un análisis geográfico explícito, el comportamiento esperado es que los consumidores en Europa, con una cultura diferente y un enfoque más centrado en la moda y el confort, preferirían categorías como *Clothing* y *Accessories*. En cambio, en Estados Unidos, donde el mercado de bicicletas tiene una mayor tradición y preferencia, se podría esperar que los consumidores prefieran productos más caros y específicos como las *Bikes*. Aún así, sin datos explícitos sobre la ubicación, esta hipótesis es plausible pero no completamente comprobada a partir de los resultados actuales.

Conclusión: No se puede confirmar completamente sin datos geográficos explícitos, pero la hipótesis es razonable y tiene fundamentos teóricos basados en las preferencias de consumo y la naturaleza de los mercados.

Análisis Final

A lo largo de este análisis, se ha profundizado en las relaciones clave entre las categorías de productos, los costos unitarios, las edades de los clientes y las variaciones geográficas. Con base en los resultados obtenidos, se puede concluir que las hipótesis planteadas inicialmente han sido en su mayoría validadas, aunque con algunas consideraciones y limitaciones.

1. Relación entre la categoría del producto, el ingreso generado y el costo unitario: La categoría de "Bikes" ha demostrado ser la más rentable debido a su alto costo unitario, lo que confirma que productos de mayor precio tienden a generar mayores ingresos. Sin embargo, las categorías de "Clothing" y "Accessories" tienen un volumen de ventas menor, pero su accesibilidad podría estar favoreciendo la compra repetida de ciertos productos.
2. Impacto de la edad del cliente en las compras considerando subcategorías y costo unitario: La segmentación por edad muestra una preferencia clara por productos más económicos entre los jóvenes, como lo reflejan las subcategorías de "Socks" y "Caps", mientras que los clientes mayores tienden a gastar más en productos de alto valor, como "Mountain Bikes" y "Road Bikes". Esto sugiere una estrategia de marketing que puede beneficiarse de ofrecer promociones específicas según el grupo etario.
3. Diferencias geográficas en la preferencia de productos**:** Aunque no se contaron con datos explícitos sobre la localización geográfica de los clientes, se han identificado tendencias en las preferencias por categorías de productos que podrían alinearse con los patrones culturales y económicos de diferentes regiones, especialmente entre Europa y Estados Unidos. La validación completa de esta hipótesis requiere información geográfica más detallada.

Recomendaciones

Acciones Estratégicas

- Ajuste de la oferta de productos por segmento de edad: Basado en las preferencias observadas, sería recomendable que la empresa personalice las campañas de marketing y promociones para cada grupo de edad. Por ejemplo, para los más jóvenes, ofrecer descuentos o paquetes en productos más accesibles como "Socks" o "Caps", mientras que para los mayores, se pueden diseñar promociones específicas para productos de mayor precio, como "Mountain Bikes".
- Optimización de precios y márgenes de ganancia**:** Los precios en las categorías de bajo costo (como Clothing y Accessories) podrían mantenerse accesibles, pero con una estrategia de volumen que busque aumentar la frecuencia de compra. Para las categorías con alto costo unitario como "Bikes", se pueden implementar descuentos por volumen o beneficios adicionales, incentivando la compra de productos de mayor precio.
- Expansión geográfica basada en análisis regional**:** Si bien la hipótesis geográfica no fue confirmada debido a la falta de datos, sería útil realizar un análisis más profundo en el futuro para identificar diferencias regionales en las preferencias. La información geográfica podría ayudar a la empresa a alinear sus productos con las necesidades de cada mercado, especialmente en Europa y Estados Unidos, y ajustar las estrategias de distribución.
- Aprovechar la correlación entre costeo y rentabilidad: La relación positiva entre los costos unitarios y los ingresos sugiere que la empresa debe continuar evaluando sus márgenes de ganancia por cada categoría. El análisis de rentabilidad por categoría debe ser una práctica continua, ajustando precios según las fluctuaciones de los costos y la demanda.

Precauciones y Limitaciones Futuras

- Variabilidad en los costos y precios: El alto nivel de desviación estándar en el costo unitario y los ingresos puede generar cierta incertidumbre, especialmente en categorías con productos de precios muy dispares. Sería recomendable realizar un análisis más detallado de los outliers para entender mejor las fluctuaciones extremas en los costos y los ingresos.
- Comportamiento de compra según la edad: Aunque se ha identificado una tendencia clara, los patrones de compra pueden variar con el tiempo, especialmente con el cambio generacional. Las estrategias de marketing deben ser lo suficientemente flexibles como para adaptarse a los cambios en las preferencias de los consumidores.
- Sensibilidad de los precios: A medida que la empresa ajusta los precios, es crucial monitorear de cerca la elasticidad de la demanda, especialmente en las categorías más económicas, ya que pequeñas fluctuaciones en el precio pueden tener un gran impacto en la cantidad vendida.

Conclusión

Este proyecto ha permitido realizar un análisis integral de los productos y las preferencias de los consumidores, utilizando datos clave relacionados con la categoría de productos, costos unitarios, segmentación por edad y posibles diferencias geográficas. A través de la verificación de las hipótesis planteadas y el análisis de diversas variables, se ha logrado obtener una visión clara sobre cómo las diferentes categorías de productos impactan los ingresos y la rentabilidad, así como las preferencias de compra de los consumidores.

1. Rentabilidad y preferencias de producto: Se ha confirmado que la categoría **"***Bikes"* es la más rentable debido a su alto costo unitario, lo que implica que la empresa debería seguir enfocándose en estos productos, posiblemente explorando oportunidades de diversificación dentro de esta categoría. En contraste, las categorías de *"Clothing" y "Accessories***"**, aunque generan menores ingresos, también juegan un papel crucial en la estrategia de ventas al ser productos más accesibles, lo que podría llevar a una estrategia de precios que favorezca la compra repetida.
2. Segmentación por edad: El análisis mostró que los consumidores más jóvenes tienden a preferir productos de menor costo unitario, como accesorios y ropa, mientras que los consumidores mayores optan por productos de mayor precio, como bicicletas de montaña o carretera. Esto sugiere que la personalización de las estrategias de marketing y las promociones por segmento de edad puede mejorar la efectividad de las ventas y aumentar la satisfacción del cliente.
3. Posibles diferencias geográficas: Aunque no se contó con información geográfica precisa para validar completamente la hipótesis sobre preferencias regionales, se identificó que la segmentación geográfica podría tener un impacto significativo en las decisiones de compra. Las preferencias de productos varían según la región, lo que implica que en el futuro la empresa debería incorporar un análisis geográfico más detallado para optimizar sus estrategias de marketing y distribución.
4. Oportunidades de mejora**:** Si bien se han validado varias hipótesis, es importante señalar que hay limitaciones debido a la falta de datos geográficos específicos y la posible variabilidad en los patrones de consumo según generaciones y tendencias culturales. La empresa debería considerar realizar estudios de mercado adicionales, especialmente en cuanto a las preferencias geográficas y las futuras tendencias de consumo, para adaptar sus estrategias a un entorno en constante cambio.

En resumen, este proyecto proporciona una base sólida para tomar decisiones estratégicas informadas, orientadas a maximizar los ingresos y la rentabilidad. Al implementar las recomendaciones propuestas, como la personalización de ofertas según la edad, la optimización de precios y el análisis geográfico, la empresa podrá fortalecer su posición en el mercado, mejorar la experiencia del cliente y aprovechar las oportunidades de crecimiento a largo plazo.