VENTAS RETAIL

Análisis EDA

MARIA FERNANDA SALGADO FLORES

DESCRIPCIÓN DE DATA

El análisis se basa en un conjunto de datos sintéticos de Ventas Minoristas y Demografía de Clientes. Estos datos han sido diseñados meticulosamente para simular un entorno minorista dinámico y reflejar escenarios del mundo real.

🎯 Objetivo del Análisis Exploratorio de Datos (EDA)

El objetivo principal del Análisis Exploratorio de Datos (EDA) es desentrañar patrones de ventas complejos, identificar la influencia de factores demográficos como el género y la edad en el comportamiento de compra, y extraer información práctica que pueda guiar las estrategias comerciales.

🔗 Estructura y Enfoque

El conjunto de datos combina detalles de la transacción (ID, fecha, categoría de producto, cantidad, precio) con información demográfica del cliente (ID, sexo, edad), siendo el Importe Total (Total Amount) la métrica clave de rendimiento que exploramos.

DATA

	A	B	C	D	E	F	G
1	Transaction ID,Date,Customer ID,Gender,Age,Product Category,Quantity,Price per Unit,Total Amount						
2	1,2023-11-24,CUST001,Male,34,Beauty,3,50,150						
3	2,2023-02-27,CUST002,Female,26,Clothing,2,500,1000						
4	3,2023-01-13,CUST003,Male,50,Electronics,1,30,30						
5	4,2023-05-21,CUST004,Male,37,Clothing,1,500,500						
6	5,2023-05-06,CUST005,Male,30,Beauty,2,50,100						
7	6,2023-04-25,CUST006,Female,45,Beauty,1,30,30						
8	7,2023-03-13,CUST007,Male,46,Clothing,2,25,50						
9	8,2023-02-22,CUST008,Male,30,Electronics,4,25,100						
10	9,2023-12-13,CUST009,Male,63,Electronics,2,300,600						
11	10,2023-10-07,CUST010,Female,52,Clothing,4,50,200						
12	11,2023-02-14,CUST011,Male,23,Clothing,2,50,100						
13	12,2023-10-30,CUST012,Male,35,Beauty,3,25,75						
14	13,2023-08-05,CUST013,Male,22,Electronics,3,500,1500						
15	14,2023-01-17,CUST014,Male,64,Clothing,4,30,120						
16	15,2023-01-16,CUST015,Female,42,Electronics,4,500,2000						
17	16,2023-02-17,CUST016,Male,19,Clothing,3,500,1500						
18	17,2023-04-22,CUST017,Female,27,Clothing,4,25,100						
19	18,2023-04-30,CUST018,Female,47,Electronics,2,25,50						
20	19,2023-09-16,CUST019,Female,62,Clothing,2,25,50						
21	20,2023-11-05,CUST020,Male,22,Clothing,3,300,900						
22	21,2023-01-14,CUST021,Female,50,Beauty,1,500,500						
23	22,2023-10-15,CUST022,Male,18,Clothing,2,50,100						

Columna	Tipo de Dato Esperado
Transaction ID	Categórico/Texto
Date	Fecha y Hora
Customer ID	Categórico/Texto
Gender (Sexo)	Categórico
Age (Edad)	Numérico
Product Category	Categórico
Quantity (Cantidad)	Numérico
Unit Price (Precio Unitario)	Numérico
Total Amount (Importe Total)	Numérico

LIMPIEZA

1 Separar las columnas del archivo CSV

```
from google.colab import drive
drive.mount('/content/drive')
df=pd.read_csv('/content/drive/MyDrive/PYTHON/retail_sales_dataset.csv', sep=';')
display(df)
```

	Transaction ID	Date	Customer ID	Gender	Age	Product Category	Quantity	Price per Unit	Total Amount
0	1	2023-11-24	CUST001	Male	34	Beauty	3	50	150
1	2	2023-02-27	CUST002	Female	26	Clothing	2	500	1000
2	3	2023-01-13	CUST003	Male	50	Electronics	1	30	30
3	4	2023-05-21	CUST004	Male	37	Clothing	1	500	500
4	5	2023-05-06	CUST005	Male	30	Beauty	2	50	100
...	...	...	...	...	...	...	...	...	...

2 Identificar valores nulos para trabajarlos

	0
Transaction ID	0
Date	0
Customer ID	0
Gender	0
Age	0
Product Category	0
Quantity	0
Price per Unit	0
Total Amount	0

3

Identificar el tipo de datos y cambiarlo en caso sea necesario

	0
Transaction ID	int64
Date	object
Customer ID	object
Gender	object
Age	int64
Product Category	object
Quantity	int64
Price per Unit	int64
Total Amount	int64

El tipo de dato object se utiliza para representar información no numérica, es decir, valores que contienen texto, cadenas de caracteres o combinaciones alfanuméricas.
En el caso de que una columna contenga fechas, como “2025-10-28”, no debe permanecer como object. En estos casos, es recomendable convertirla al tipo datetime.

```
df['Date'] = pd.to_datetime(df['Date'], errors='coerce')
display(df.dtypes)
```

	0
Transaction ID	int64
Date	datetime64[ns]
Customer ID	object
Gender	object
Age	int64
Product Category	object
Quantity	int64
Price per Unit	int64
Total Amount	int64

ANÁLISIS DESCRIPTIVO

PRIMERAS FILAS

df.head()

	Transaction ID	Date	Customer ID	Gender	Age	Product Category	Quantity	Price per Unit	Total Amount
0	1	24/11/2023	CUST001	Male	34	Beauty	3	50	150
1	2	27/02/2023	CUST002	Female	26	Clothing	2	500	1000
2	3	13/01/2023	CUST003	Male	50	Electronics	1	30	30
3	4	21/05/2023	CUST004	Male	37	Clothing	1	500	500
4	5	6/05/2023	CUST005	Male	30	Beauty	2	50	100

ÚLTIMAS FILAS

df.tail()

	Transaction ID	Date	Customer ID	Gender	Age	Product Category	Quantity	Price per Unit	Total Amount
995	996	16/05/2023	CUST996	Male	62	Clothing	1	50	50
996	997	17/11/2023	CUST997	Male	52	Beauty	3	30	90
997	998	29/10/2023	CUST998	Female	23	Beauty	4	25	100
998	999	5/12/2023	CUST999	Female	36	Electronics	3	50	150
999	1000	12/04/2023	CUST1000	Male	47	Electronics	4	30	120

FILAS ALEATORIAS

df.sample(5)

	Transaction ID	Date	Customer ID	Gender	Age	Product Category	Quantity	Price per Unit	Total Amount
134	135	26/02/2023	CUST135	Male	20	Clothing	2	25	50
555	556	4/06/2023	CUST556	Female	18	Electronics	1	50	50
314	315	1/06/2023	CUST315	Male	47	Clothing	2	30	60
219	220	3/03/2023	CUST220	Male	64	Beauty	1	500	500
983	984	29/08/2023	CUST984	Male	56	Clothing	1	500	500

Descripción Numérica

df.describe()

	Transaction ID	Age	Quantity	Price per Unit	Total Amount
count	1000.000000	1000.00000	1000.000000	1000.000000	1000.000000
mean	500.500000	41.39200	2.514000	179.890000	456.000000
std	288.819436	13.68143	1.132734	189.681356	559.997632
min	1.000000	18.00000	1.000000	25.000000	25.000000
25%	250.750000	29.00000	1.000000	30.000000	60.000000
50%	500.500000	42.00000	3.000000	50.000000	135.000000
75%	750.250000	53.00000	4.000000	300.000000	900.000000
max	1000.000000	64.00000	4.000000	500.000000	2000.000000

Age (Edad): Las edades de los individuos en la muestra varían entre un mínimo de 18 años y un máximo de 64 años. La edad promedio (media) es de aproximadamente 41,39 años, y la edad mediana (50%) es de 42 años, lo que indica una distribución de edad bastante centrada. El 75% de las transacciones son realizadas por personas de 53 años o menos.

Quantity (Cantidad): La cantidad de unidades compradas por transacción es muy limitada, variando solo de 1 a 4 unidades. La media es de aproximadamente 2.51 unidades, y la desviación estándar es relativamente baja (1.13), lo que confirma que las cantidades por compra son consistentemente pequeñas.

Price per Unit (Precio por Unidad): Los precios unitarios tienen un rango considerable, yendo de un mínimo de 25.00 a un máximo de 500.00. Sin embargo, la mediana (50.00) es mucho menor que la media (179.89), y el 75% de las unidades se venden a 300.00 o menos. Esta gran diferencia entre la media y la mediana, junto con una alta desviación estándar (189.68), sugiere que la distribución del precio por unidad está sesgada a la derecha, lo que significa que la mayoría de las unidades se compran a precios bajos, pero existen algunas transacciones a precios muy altos que elevan el promedio.

Total Amount (Monto Total): Los montos totales de las transacciones oscilan entre un mínimo de 25.00 y un máximo de 2000.00. Al igual que el precio por unidad, la mediana (135.00) es significativamente menor que la media (456.00), y la desviación estándar es la más alta en términos absolutos (559.99). Esto refuerza la idea de un fuerte sesgo a la derecha, donde la mayoría de las transacciones son de montos pequeños, pero un número reducido de compras de alto valor contribuyen significativamente al monto total promedio.

AGE

```
age=df['Age'].mean()
max=df['Age'].max()
min=df['Age'].min()
print(f'La edad máxima de los clientes es {max}, la minima es {min}. Por lo tanto, el promedio es {age}')
```

```
La edad máxima de los clientes es 64, la minima es 18. Por lo tanto, el promedio es 41.392
```

GENDER

```
conteo=df['Gender'].value_counts()
mujer=conteo['Female']
hombre=conteo['Male']
print(f'Hay {mujer} clientes mujeres y {hombre} clientes hombres')
```

```
Hay 510 clientes mujeres y 490 clientes hombres
```

PRODUCT CATEGORY

```
producto = df['Product Category'].value_counts()
ropa = producto['Clothing']
electronica = producto['Electronics']
belleza = producto['Beauty']
print(f'Se venden unidades en Ropa, Electrodomésticos y Productos de Belleza, {ropa},{electronica} y {belleza} respectivamente. Hay {producto.sum()} unidades totales')
```

```
Se venden unidades en Ropa, Electrodomésticos y Productos de Belleza, 351,342 y 307 respectivamente. Hay 1000 unidades totales
```

EDA

Importar librerias:

```
import pandas as pd
import numpy as np
import matplotlib as plt
import matplotlib.pyplot as plt
import seaborn as sns
```

Para ir más allá de un simple análisis de ingresos y poder determinar la verdadera contribución de cada cliente y producto a la utilidad de la tienda, se requiere la inclusión del Costo Total y, por consiguiente, la columna de Margen (Ganancia Bruta). Esto nos permitirá un análisis de rentabilidad por segmento demográfico y categoría de producto.

```
df['Costo_Total'] = df['Total Amount']*0.60
df['Margen'] =df['Total Amount']-df['Costo_Total']
```

	Transaction ID	Date	Customer ID	Gender	Age	Product Category	Quantity	Price per Unit	Total Amount	Costo_Total	Margen
0	1	2023-11-24	CUST001	Male	34	Beauty	3	50	150	90.0	60.0
1	2	2023-02-27	CUST002	Female	26	Clothing	2	500	1000	600.0	400.0
2	3	2023-01-13	CUST003	Male	50	Electronics	1	30	30	18.0	12.0
3	4	2023-05-21	CUST004	Male	37	Clothing	1	500	500	300.0	200.0
4	5	2023-05-06	CUST005	Male	30	Beauty	2	50	100	60.0	40.0
...	...	...	...	...	...	...	...	...	...	...	...

Insights Clave

producto=df.groupby('Product Category')['Total Amount'].sum()

Product Category	Total Amount
Beauty	143515
Clothing	155580
Electronics	156905

La categoría de Electronics generó la mayor cantidad de ingresos (156,905), seguida muy de cerca por Clothing (155,580). Beauty es la que menos contribuye a los ingresos totales.

df.groupby('Gender')['Margen'].mean()

Gender	Margen
Female	182.619608
Male	182.171429

Aunque las clientes femeninas tienen un margen promedio ligeramente superior (aproximadamente $0.45 más alto), la diferencia es tan mínima que se puede concluir que ambos géneros compran artículos que generan una rentabilidad similar.

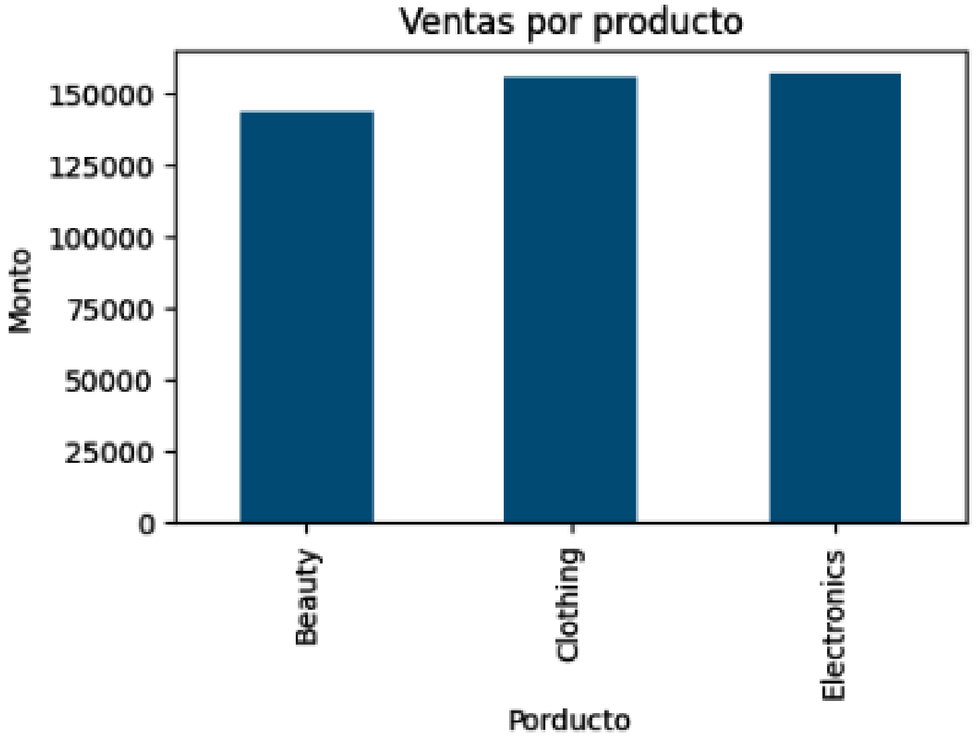
```
df['Mes'] = df['Date'].dt.to_period('M')
ventasmes = df.groupby('Mes')['Total Amount'].sum()
display(ventasmes)
```

Mes	Total Amount
2023-01	35450
2023-02	44060
2023-03	28990
2023-04	33870
2023-05	53150
2023-06	36715
2023-07	35465
2023-08	36960
2023-09	23620
2023-10	46580
2023-11	34920
2023-12	44690
2024-01	1530

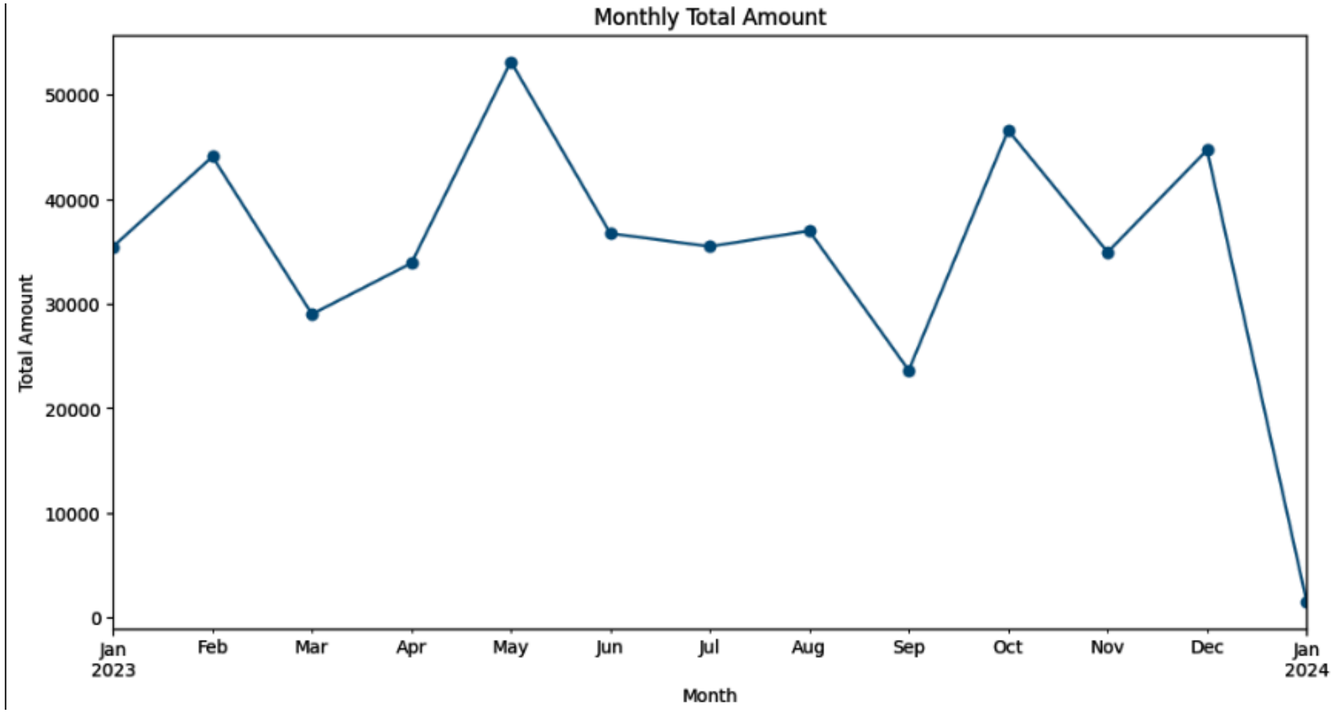
Se observa una fuerte estacionalidad con el pico de ventas ocurriendo en Mayo de 2023 ($53,150). Por otro lado, hay un notable declive, con Septiembre como el mes más bajo de 2023 ($23,620) y Enero de 2024 marcando el punto más bajo del período analizado ($1,530), lo cual es común después de la temporada navideña.

Gráficos

```
g1=df.groupby('Product Category')['Total Amount'].sum()
g1.plot(kind='bar',figsize=(5,3))
plt.title('Ventas por producto')
plt.xlabel('Porducto')
plt.ylabel('Monto')
```

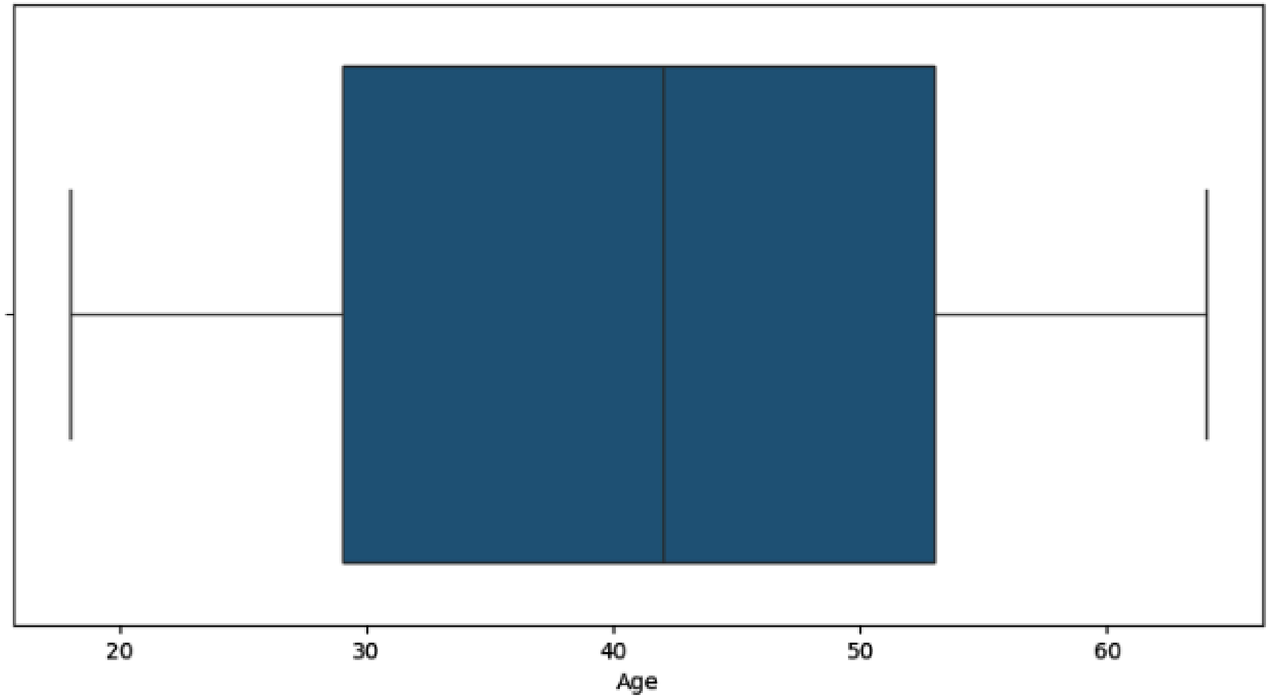



```
plt.figure(figsize=(12, 6))
ventasmes.plot(kind='line', marker='o')
plt.xlabel('Month')
plt.ylabel('Total Amount')
plt.title('Monthly Total Amount')
plt.show()
```



Box Plot por Edad

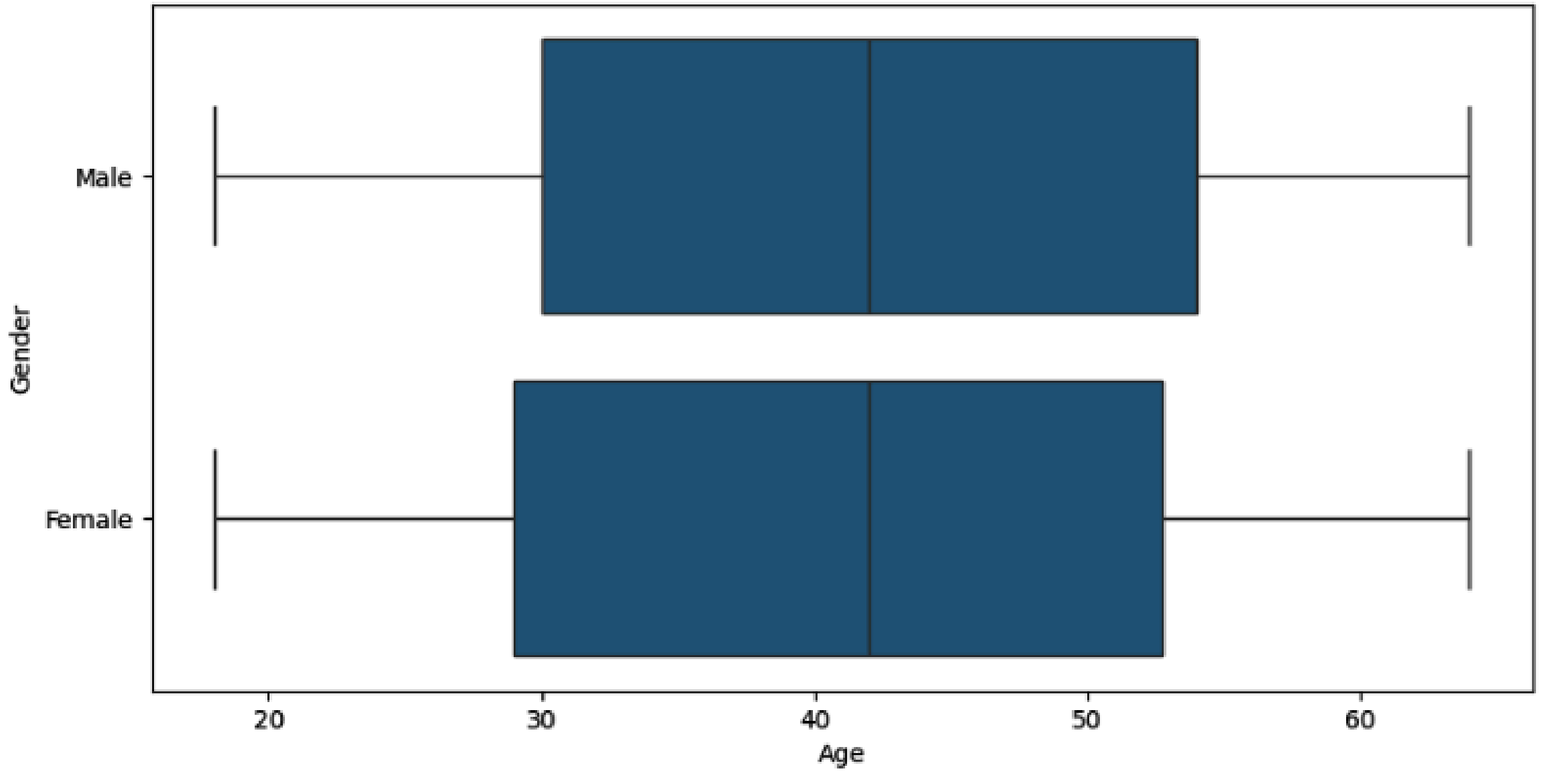
```
plt.figure(figsize=(10,5))
sns.boxplot(data=df, x='Age')
plt.show()
```



El rango de edad total de los clientes empieza desde los 19 hasta los 62 años aproximadamente.El rango intercuantil, donde se encuentran el 50% de los clientes, es entre 30 y 52 años. Dentro de este rango, se identificó que el promedio tiene 41 años aproximadamente. El primer cuartil tiene 30 años a menos. Mientras que el tercer cuartil tiene de 52 años a menos. Además, no hay valores atípicos.

Box Plot por Género

```
plt.figure(figsize=(10,5))
sns.boxplot(data=df, x='Age', y='Gender')
plt.show()
```

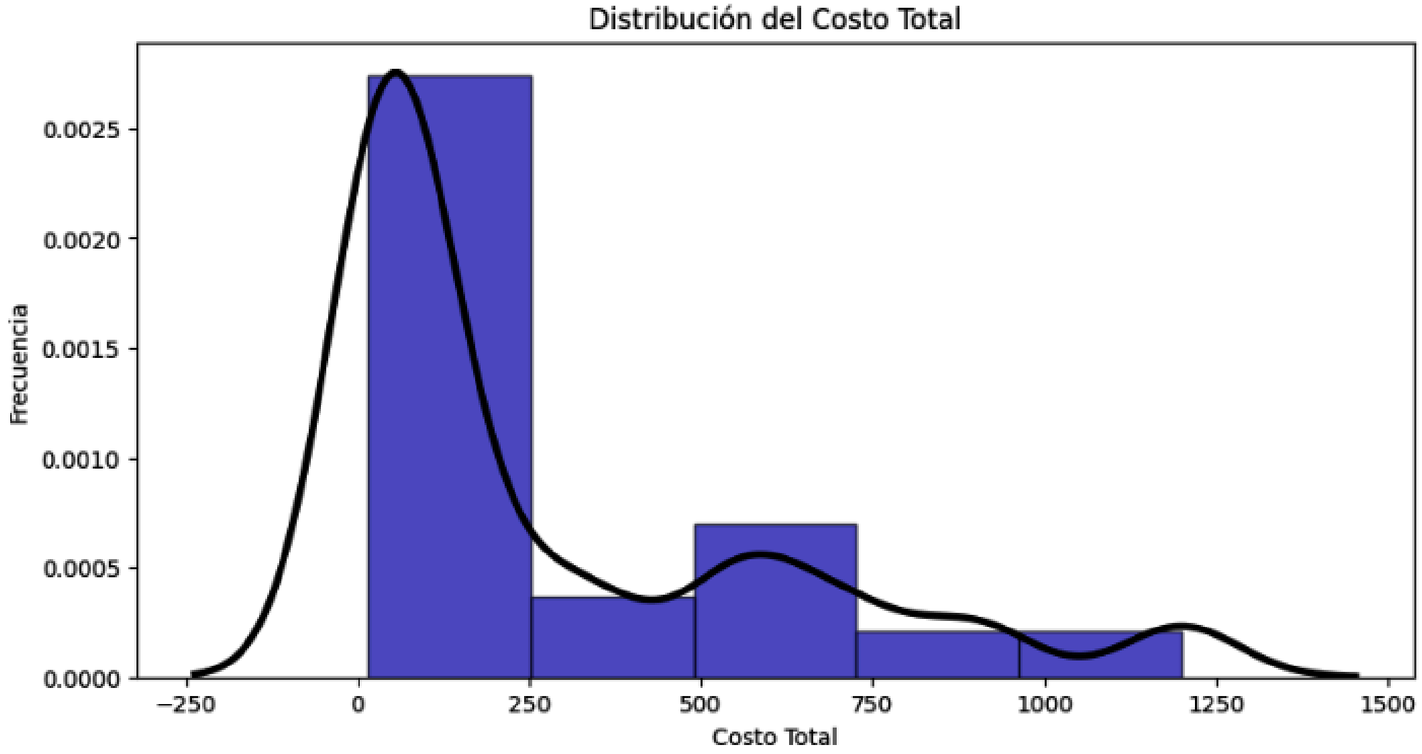


El rango completo en ambas géneros empiezas desde los 19 años hasta los 62 años.

- Male: El rango intercuantil, donde se concentra el 50% de los clientes, es entre 30 a 52 años. Dentro de este rango, se identificó que el promedio tiene 41 años. El primer cuartil tiene 30 a menos. Mientras que el tercer cuartil tiene 52 a menos. No presenta valores atípicos.
- Female: El rango intercuantil, donde se concentra el 50% de los clientes, es entre 29 a 51 años. Dentro de este rango, se identificó que el promedio tiene 41 años. El primer cuartil tiene 29 a menos. Mientras que el tercer cuartil tiene 51 a menos. No presenta valores atípicos.

Histograma

```
plt.figure(figsize=(10,5))
df['Costo_Total'].plot(kind='hist',bins=5,density=True,alpha=0.6,color='blue',edgecolor='black')
sns.kdeplot(df['Costo_Total'],color='black',linewidth=3)
plt.title('Distribución del Costo Total')
plt.xlabel('Costo Total')
plt.ylabel('Frecuencia')
```

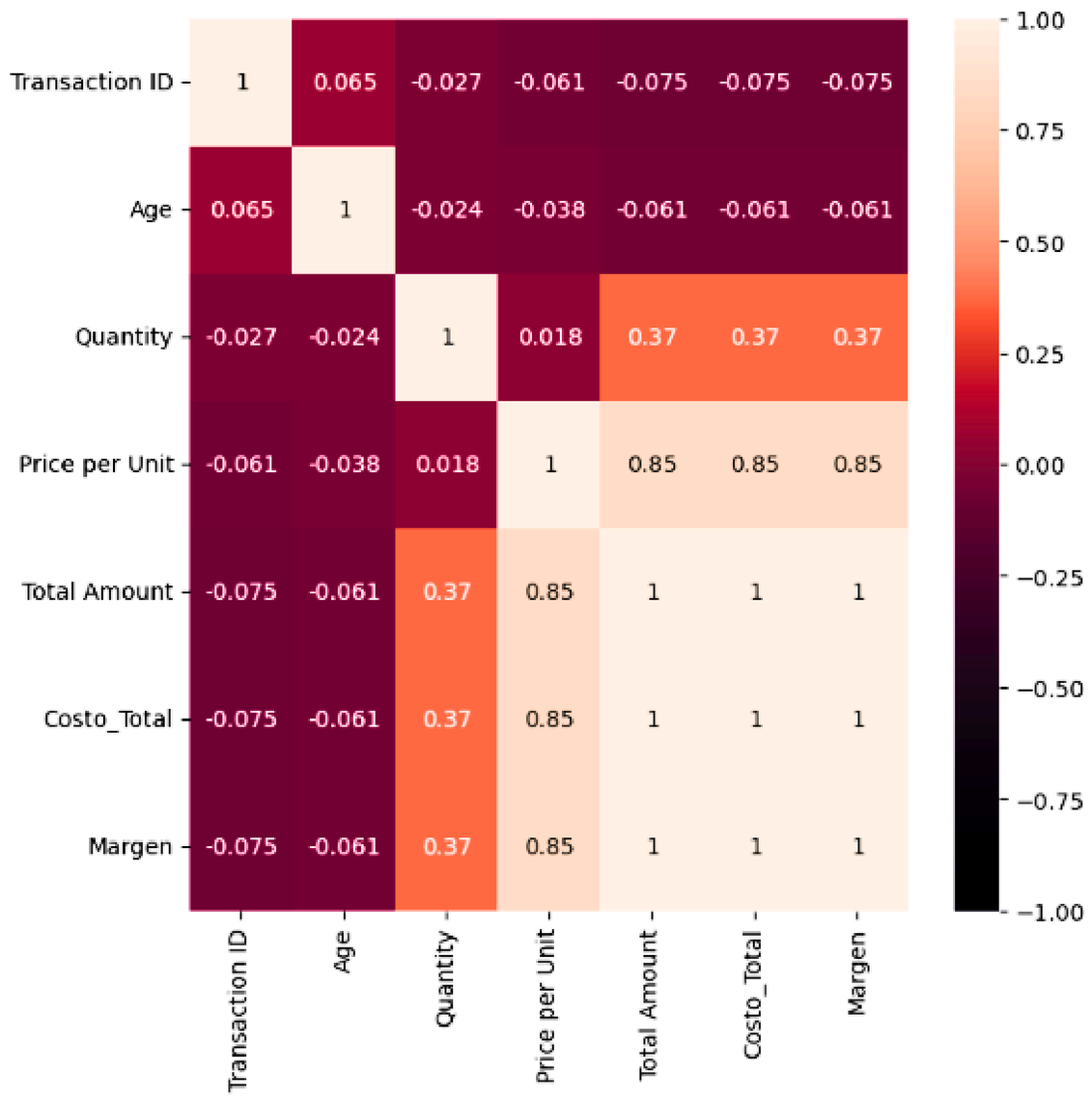


La gran mayoría de las transacciones (la moda) tienen un Costo Total bajo, concentrándose en el rango de 0 a aproximadamente 250 nuevos soles. A partir de ese punto, la frecuencia cae drásticamente, extendiéndose en una cola larga hacia costos más altos (500, 750 y más), lo que indica que las transacciones de Costo Total elevado son mucho menos comunes que las de bajo valor.

Mapa de Calor

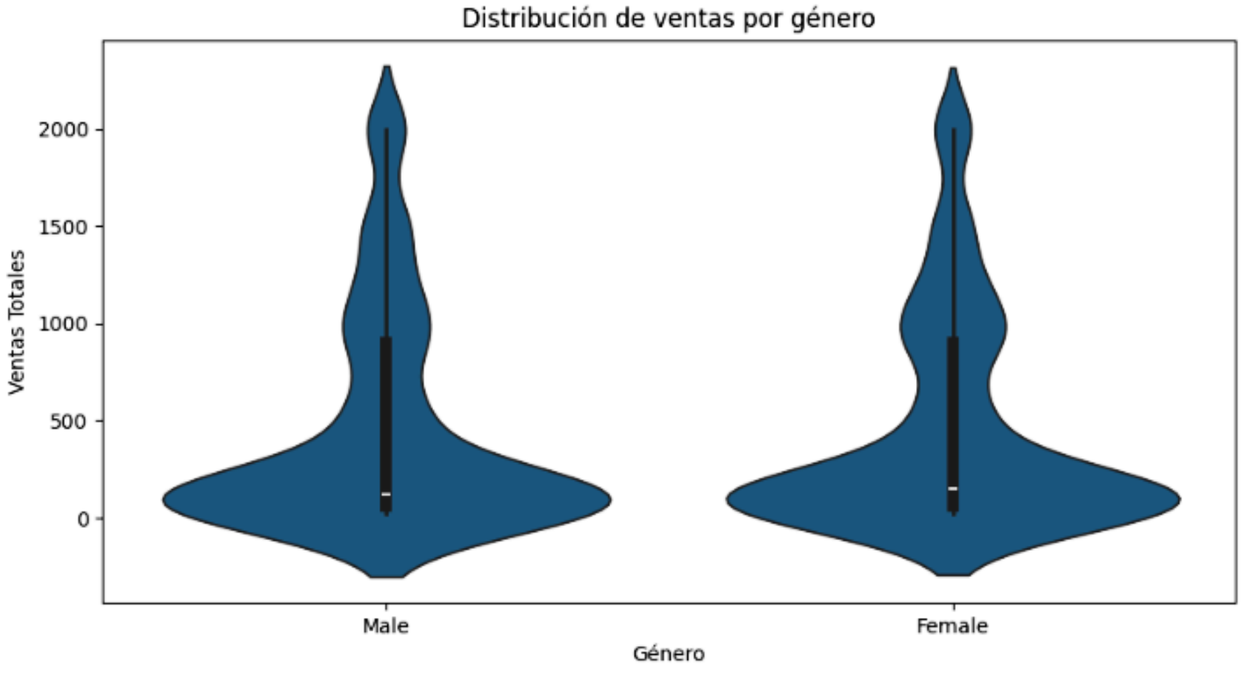
```
plt.figure(figsize=(7,7))
cor_matrix = df.select_dtypes(include=np.number).corr()
sns.heatmap(cor_matrix, annot=True,vmin=-1)
plt.show()
```

1. Price Per Unit & Total Amount: El valor de la correlación es cercano a la unidad, 0.85, es decir, una correlación perfecta. Esto indica que un aumento del precio aumenta el monto total.
2. Total Amount & Quantity: La correlación es positiva (0.37), pero muy pequeña. Los montos totales son impulsados más por el precio unitario que por la cantidad de artículos comprados.
3. Age & Otras variables: Muestra que la edad del cliente no tiene una relación lineal fuerte con ninguna otra métrica de la transacción (ni con la cantidad comprada, ni con el precio, ni con el monto total).



Violin Plot

```
plt.figure(figsize=(10,5))
sns.violinplot(data=df,x='Gender',y='Total Amount')
plt.title('Distribución de ventas por género')
plt.xlabel('Género')
plt.ylabel('Ventas Totales')
plt.show()
```



Se observa que la forma de ambos violines es similar, lo que indica que tanto hombres como mujeres presentan una distribución de ventas similares.En ambos géneros, se encuentra lo siguiente:

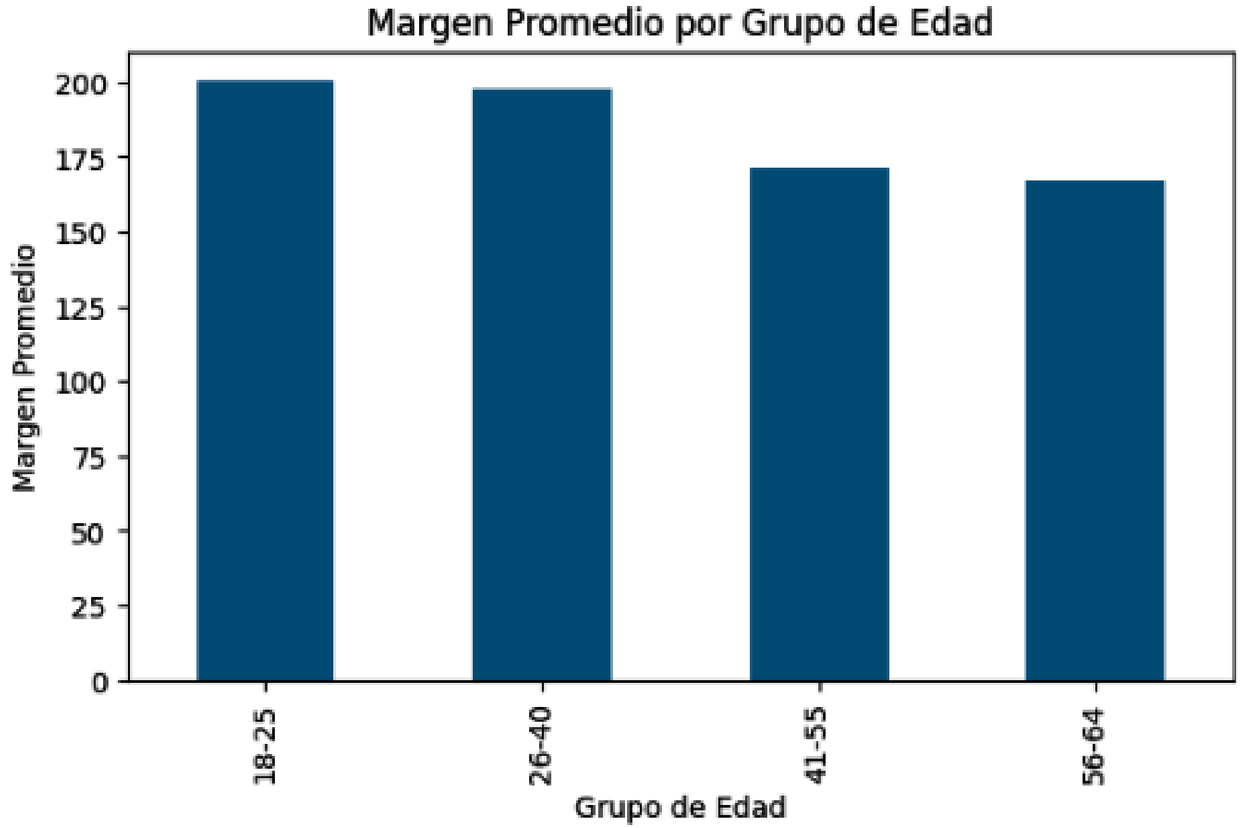
1. Las ventas totales empiezan desde los 0 a 2200 unidades monetarias.
2. La mayor densidad de los datos se concentra en montos bajos, entre 0 a 300.
3. La línea negra representa el rango intercuantil, donde se concentral el 50% central de los datos, que en el gráfico va desde aproximadamente desde 100 (Q1) a 900 (Q3) unidades monetarias.
4. La linea blanca dentro del rango intercuantil representa la mediana, es decir, divide la distribución en dos partes iguales, que se encuentran en 150 (Male) a 200 (Female)
5. A medida que el eje Y aumenta, los violines se estrechan, indicando que las ventas mayores a 1 000 son menos frecuentes. Sin embargo, existen algunos valores extremos que alcanzan hasta los 2 000 – 2 200, lo que evidencia la presencia de ventas atípicas o excepcionales en ambos géneros.

Barras por grupos de Edad

```
bins = [18, 25, 40, 55, 65]
labels = ['18-25', '26-40', '41-55', '56-64']
df['Age_Group'] = pd.cut(df['Age'], bins=bins, labels=labels, right=False)

g_age_margin = df.groupby('Age_Group')['Margen'].mean()
g_age_margin.plot(kind='bar', figsize=(7,4))
plt.title('Margen Promedio por Grupo de Edad')
plt.xlabel('Grupo de Edad')
plt.ylabel('Margen Promedio')
plt.show()
```

Aunque el análisis de correlación (en tu Mapa de Calor) indicó que la edad no tenía una relación lineal fuerte con el Monto Total, este gráfico demuestra que la edad sí tiene una influencia significativa en la rentabilidad de las compras.
Esto sugiere que los clientes más jóvenes (18-40 años) tienden a comprar artículos que, en promedio, tienen un precio más alto o un costo asumido más bajo, lo que se traduce en una mayor utilidad para el negocio por cada transacción.



CONCLUSIONES

1. Rentabilidad y Contribución por Categoría de Producto

- Utilidad Bruta: La categoría de Electronics demostró ser la principal contribuidora a la utilidad bruta de la tienda, al registrar el mayor Ingreso Total ($156,905) y, por ende, el mayor Margen Total. Le sigue de cerca Clothing, mientras que Beauty es la categoría con la menor contribución absoluta.
- Correlaciones: El Monto Total de las transacciones está fuertemente correlacionado con el Precio por Unidad (correlación de 0.85). En contraste, la correlación con la Cantidad de unidades compradas es débil (0.37). Esto sugiere que las ventas no se basan en el volumen de artículos por transacción, sino en el precio individual de los productos.

CONCLUSIONES

2. Influencia de Factores Demográficos

- Impacto de la Edad en el Margen: Se identificó una relación inversa entre la edad del cliente y la rentabilidad promedio por transacción. Los clientes más jóvenes, específicamente el grupo de 18-25 años, generan consistentemente el Margen Promedio más alto, mientras que los clientes del grupo 56-64 años generan el Margen Promedio más bajo. Esta conclusión es clave para enfocar estrategias de fidelización y marketing en el segmento etario de mayor valor.
- Género y Margen: No se encontró una diferencia estadísticamente significativa en el comportamiento de compra por género. Tanto el Margen Promedio por transacción como la distribución de las Ventas Totales son casi idénticos entre clientes masculinos y femeninos.

CONCLUSIONES

3. Estacionalidad y Comportamiento Temporal

- Pico Estacional: El mes de Mayo de 2023 registró el pico máximo de ingresos ($53,150). Se recomienda investigar qué eventos o estrategias comerciales impulsaron este resultado.
- Tendencia a la Baja: Se observan caídas significativas en la actividad comercial, siendo Septiembre de 2023 el mes con menor actividad dentro del año calendario ($23,620) y Enero de 2024 el punto más bajo del período analizado. Esta información es crucial para optimizar la gestión de inventario y planificar campañas de reactivación post-vacacionales.

GRACIAS