

Ecommerce de Indumentaria en India: desafíos y oportunidades

Trabajo final

Gerson Molina Vaca
gersonmolinavaca@gmail.com



El mercado de mayor potencial a nivel mundial

Introducción

En las últimas dos décadas, la industria del comercio electrónico ha experimentado una transformación radical. Hacia el año 2010, los marketplaces digitales como Amazon comenzaron a consolidar una masividad considerable. En el camino hacia 2015, las redes sociales se afirmaron como canales de referencia y moldeadores del comportamiento del usuario. En los años subsiguientes, empresas de todas las industrias comenzaron, en diferentes escalas, a transformar digitalmente sus modelos de servicio.

Sin embargo, la gran aceleración se produjo hacia el 2020, cuando la pandemia de COVID-19 forzó a los consumidores rezagados de todo el mundo a adaptar su comportamiento y, en última instancia, a volcarse inevitablemente hacia el comercio electrónico.

Si bien todo este movimiento hacia las compras en línea siempre fue evidente en países occidentales, la velocidad de adopción en naciones más allá de la cortina de Occidente no fue tan clara ni consistente.

Con esta interrogante en mente, me propuse examinar de cerca el impacto del comercio electrónico en la India, una de las grandes promesas de la próxima década. Este caso de estudio se centra en las ventas de Amazon, como líder mundial del comercio electrónico; tomando una partición de datos del 2022: los suficientemente actual para retratar un escenario post pandemia. A través de este análisis, aspiro a arrojar luz sobre el fenómeno global del comercio electrónico y cómo ha influido en las economías y sociedades de regiones anteriormente menos exploradas en este contexto.

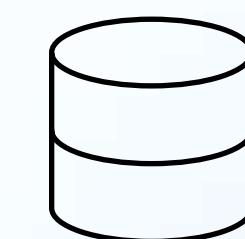
Índice

Intro	02
Presentación del Data Set.....	03
Análisis y toma de decisiones ...	05
Variable Target	06
Métricas y visualizaciones	07
Definiciones para el Modelo	12
Conclusión	15

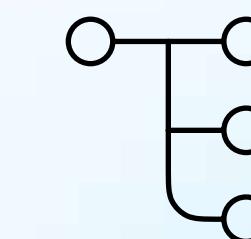
Presentación del Data Set: Ventas de Amazon India

En la categoría Indumentaria para el año 2022

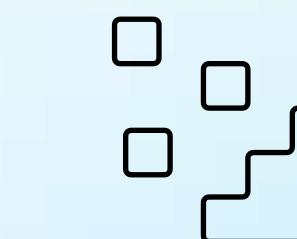
Prelimpieza:

**116.000**

Registros

**23**

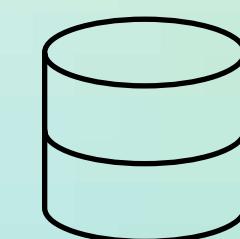
Variables totales

**2.668.000**

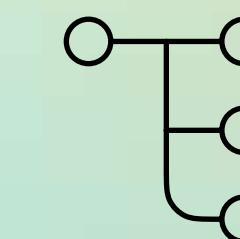
Datos



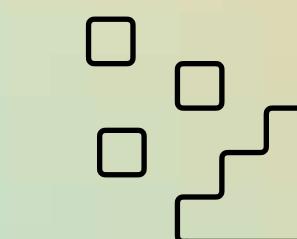
Postlimpieza:

**113.600**

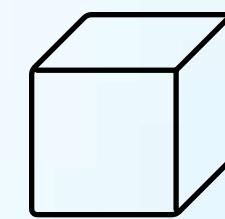
Registros

**28**

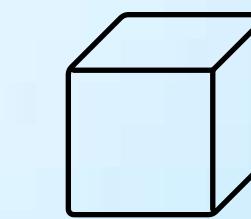
Variables totales

**3.180.800**

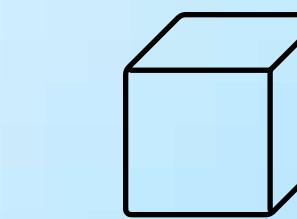
Datos

**2**

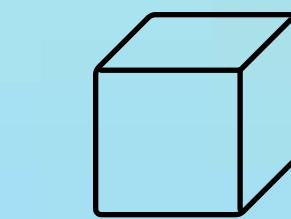
Variables Float64

**3**

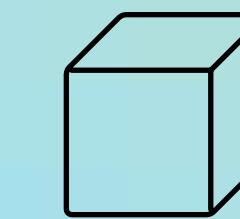
Variables Int64

**18**

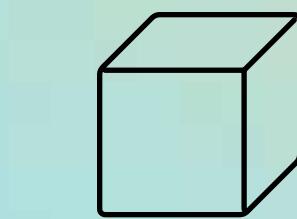
Variables Object

**1**

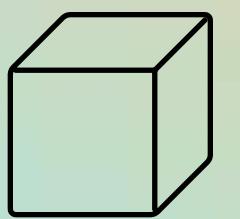
Variable Float64

**6**

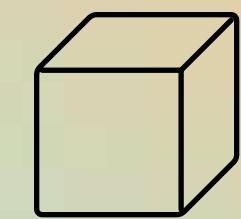
Variables Int64

**18**

Variables Object

**1**

V. Datetime 64

**3**

Variables nt64

Estructura del Data Set

index		Order ID	Date	Status	Fulfilment	ship-service-level	Style	SKU	ID Category	Category	...	Qty	currency	Amount	ship-city	ship-state	ship-postal-code	ship-country	promotion-ids	B2B	fulfilled-by	
0	1	171-9198151-1101146	04-30-22	Shipped	Merchant	Standard	JNE3781	JNE3781-KR-XXXL	5	kurta	...	1	INR	406.0	BENGALURU	KARNATAKA	560085.0	IN	Amazon PLCC Free Financing Universal Merchant ...		False	Easy Ship
1	2	404-0687676-7273146	04-30-22	Shipped	Amazon	Expedited	JNE3371	JNE3371-KR-XL	5	kurta	...	1	INR	329.0	NAVI MUMBAI	MAHARASHTRA	410210.0	IN	IN Core Free Shipping 2015/04/08 23-48-5-108		True	NaN
2	4	407-1069790-7240320	04-30-22	Shipped	Amazon	Expedited	JNE3671	JNE3671-TU-XXXL	8	Top	...	1	INR	574.0	CHENNAI	TAMIL NADU	600073.0	IN	NaN		False	NaN
3	5	404-1490984-4578765	04-30-22	Shipped	Amazon	Expedited	SET264	SET264-KR-NP-XL	7	Set	...	1	INR	824.0	GHAZIABAD	UTTAR PRADESH	201102.0	IN	IN Core Free Shipping 2015/04/08 23-48-5-108		False	NaN
4	6	408-5748499-6859555	04-30-22	Shipped	Amazon	Expedited	J0095	J0095-SET-L	7	Set	...	1	INR	653.0	CHANDIGARH	CHANDIGARH	160036.0	IN	IN Core Free Shipping 2015/04/08 23-48-5-108		False	NaN

Cabecera de los datos

Descripción de Variables

- 1. **index:** Un número de identificación único para cada registro en el conjunto de datos.
- 2. **Order ID:** El identificador único de cada orden realizada.
- 3. **Date:** La fecha en que se realizó la orden.
- 4. **Status:** El estado actual de la orden.
- 5. **Fulfilment:** El estado de cumplimiento de la orden.
- 6. **ship-service-level:** El nivel de servicio de envío utilizado para la orden.
- 7. **Style:** El estilo del producto ordenado.
- 8. **SKU:** El código de unidad de stock del producto.
- 9. **Category:** La categoría a la que pertenece el producto.
- 10. **Size:** El tamaño del producto, si es aplicable.
- 11. **ASIN:** El número de identificación único de Amazon Standard Identification Number para el producto.
- 12. **Courier Status:** El estado del envío por parte del servicio de mensajería.
- 13. **Qty:** La cantidad de productos ordenados.
- 14. **currency:** La moneda utilizada para la transacción.
- 15. **Amount:** El monto total de la orden.
- 16. **ship-city:** La ciudad de destino para el envío.
- 17. **ship-state:** El estado o provincia de destino para el envío.
- 18. **ship-postal-code:** El código postal de destino para el envío.
- 19. **ship-country:** El país de destino para el envío.
- 20. **promotion-ids:** Los identificadores de promoción aplicados a la orden, si los hay.
- 21. **B2B:** Indicador de si la transacción fue realizada entre empresas (B2B).
- 22. **fulfilled-by:** Quién cumplió la orden, si fue cumplida por una entidad diferente a la que la recibió originalmente.

Análisis y toma de decisiones

Luego de la exploración de las variables, procedí a su tratamiento y limpieza. Las siguientes son 4 decisiones destacadas:

1

Ánalisis: las variables "Courier status", "Ship-city", "Ship-state", "Ship-postal-code", "Ship-country" y "Promotion-ids" **contenían valores vacíos.**

Decisión: **Asigné** la sintaxis "No Especificados" a cada valor vacío, para mantener la consistencia con el resto de las variables en esa fila.

2

Ánalisis: las variables "Currency" y "Amount" **contenían filas con valores nulos**, en última instancia irrelevantes para el Modelo (dado que "Amount" es la variable target).

Decisión: **Eliminé** las filas con valores nulos dentro de estas variables.

3

Ánalisis: las variables "B2B" y "fulfilled-by" **contenían filas sin valores.**

Decisión: **Asigné** a los valores faltantes la imputación de la media.

4

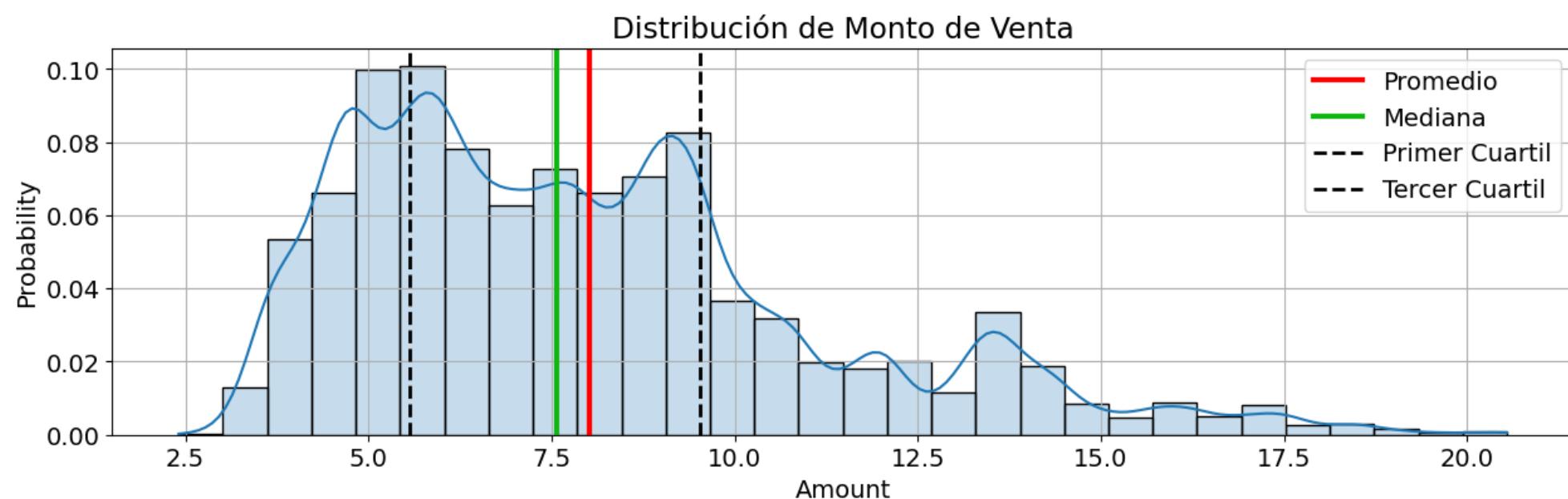
Ánalisis: la variable "Date" **contenía valores de "object"**, dificultando su procesamiento al momento de analizar por periodo.

Decisión: **Convertí** los datos al formato "datetime" y "datatype".

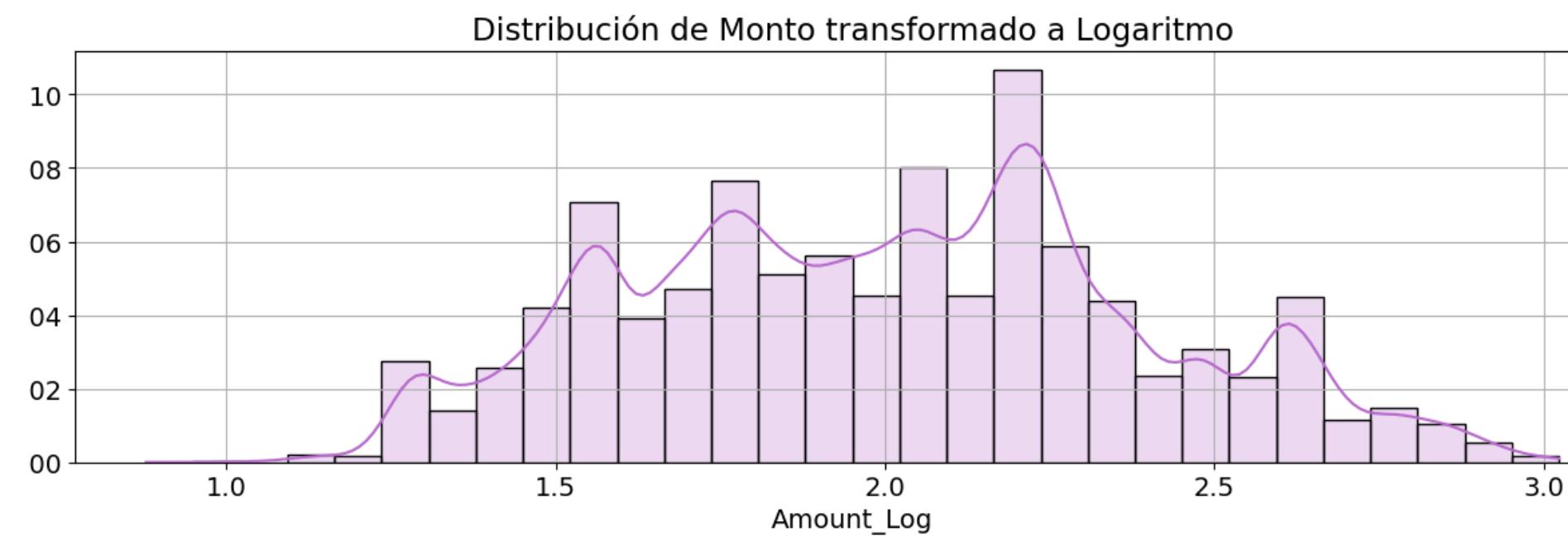
Eligiendo la variable target: Amount

La variable Amount reporta el monto total de cada venta. Después de realizar el análisis exploratorio definí que como variable Target sería la más indicada para la predicción del modelo.

1. Usando la fórmula: “`CalculaMetricas(df['Amount'])`” validé que la distribución sea simétrica.



2. A continuación, transformé la variable a Logaritmo para normalizar los datos.



Métricas y visualizaciones

Escogí las siguientes 5 visualizaciones porque ofrecen un abanico diverso de posibilidades con mi Modelo.

1. Relación entre “Cantidad vendida” y “Valor de ventas”

2. Costo promedio por categorías

3. Ventas por cantidad

4. Meses de mayor venta

5. Ventas por periodo

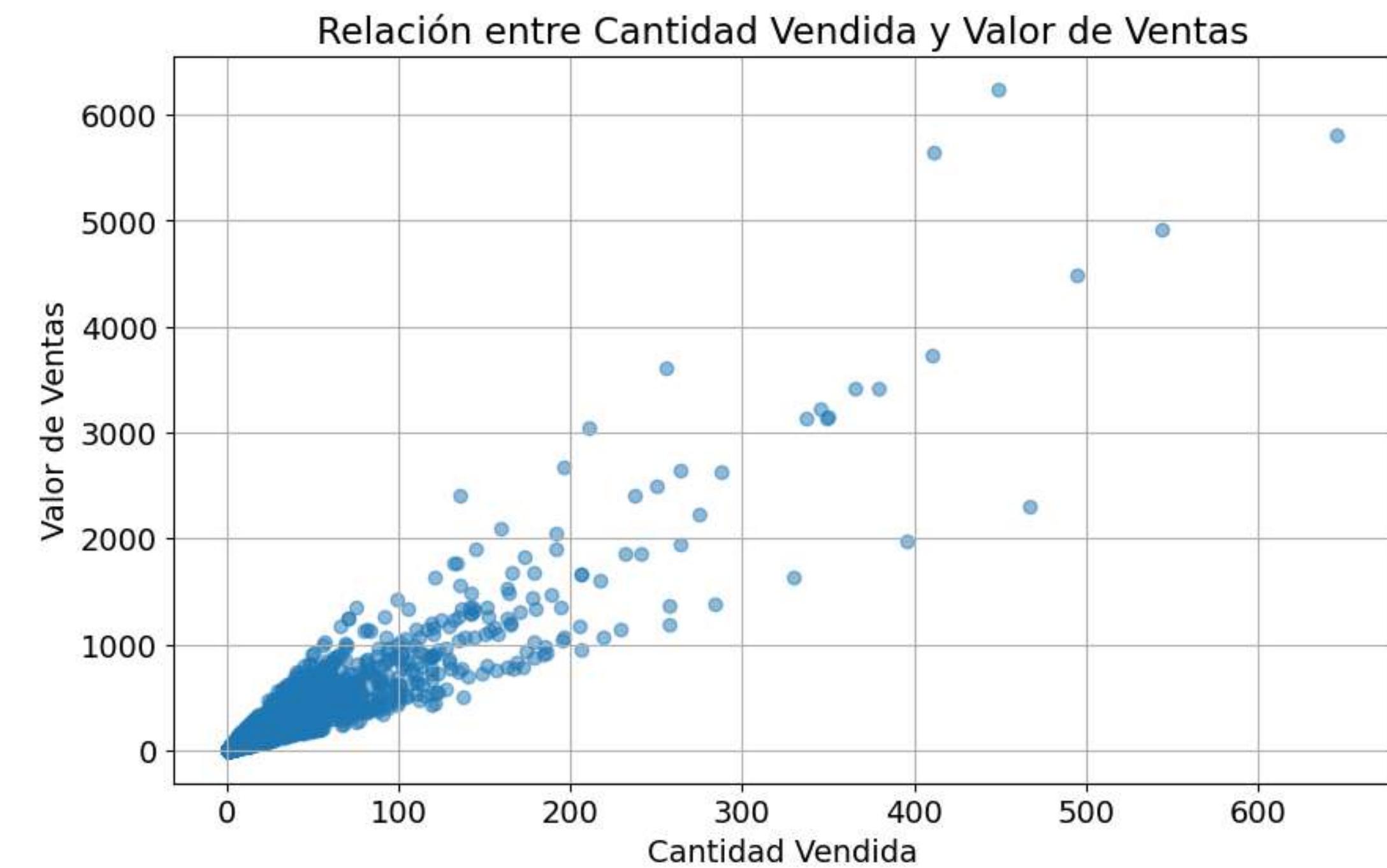


Métricas y visualizaciones

1. Relación entre “Cantidad vendida” y “Valor de ventas”

Se puede determinar en primera instancia una fuerte relación positiva entre la cantidad de ventas y el valor de las mismas, lo que podría significar que a medida que la cantidad aumenta, el valor de las ventas tienden a aumentar positivamente.

Concluyendo qué, mientras mas unidades venda genero mas ingresos por esas ventas.

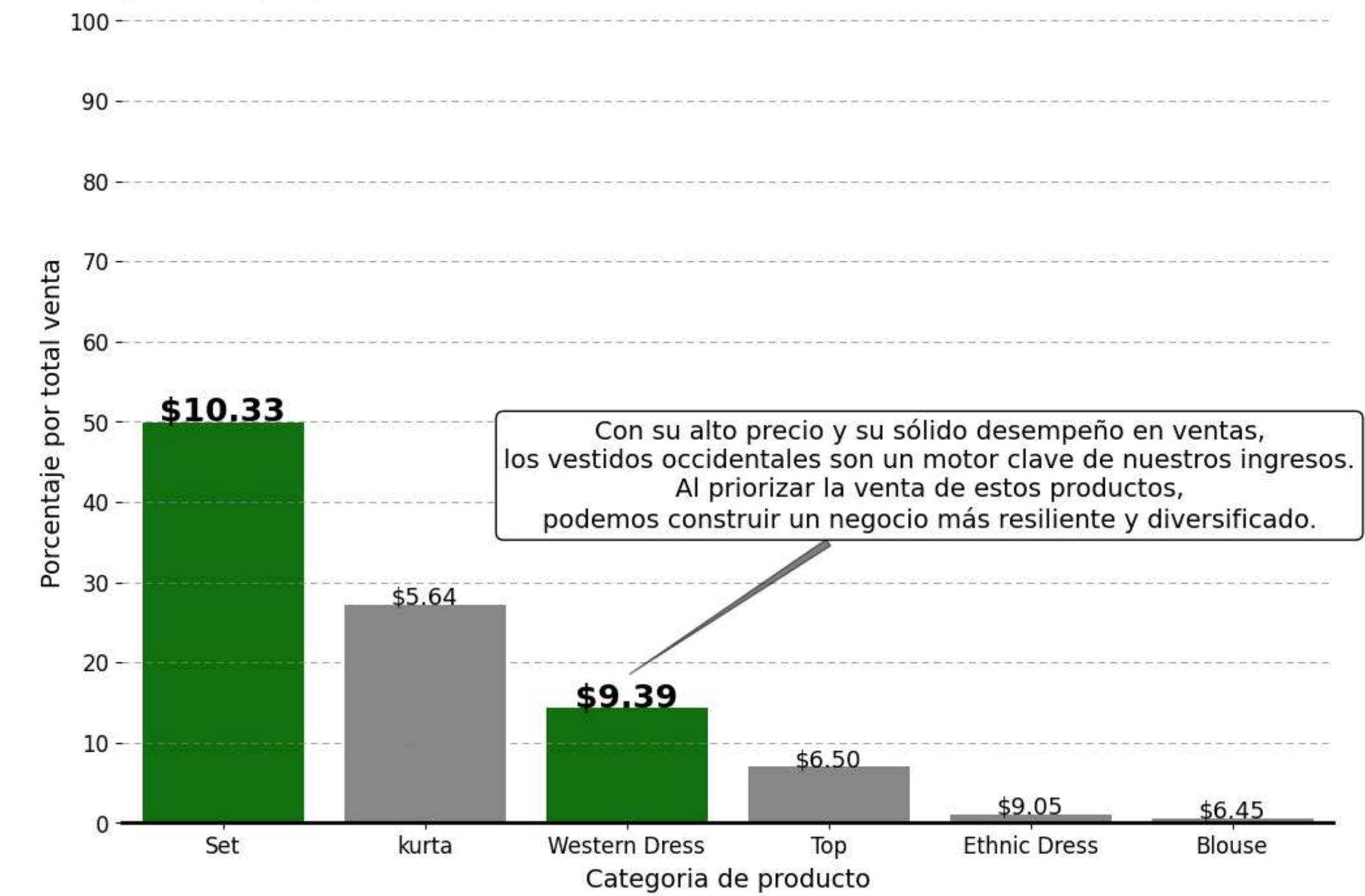


Métricas y visualizaciones

2. Costo promedio por categorías

Los vestidos occidentales representan un activo importante en nuestro portafolio de productos, con su combinación de alto precio y sólido desempeño en ventas. Al centrarnos en este segmento, podemos fortalecer nuestra posición en el mercado y construir un negocio más resiliente y diversificado. Al priorizar la venta de estos vestidos, podemos maximizar nuestros ingresos y establecer una base sólida para el crecimiento futuro.

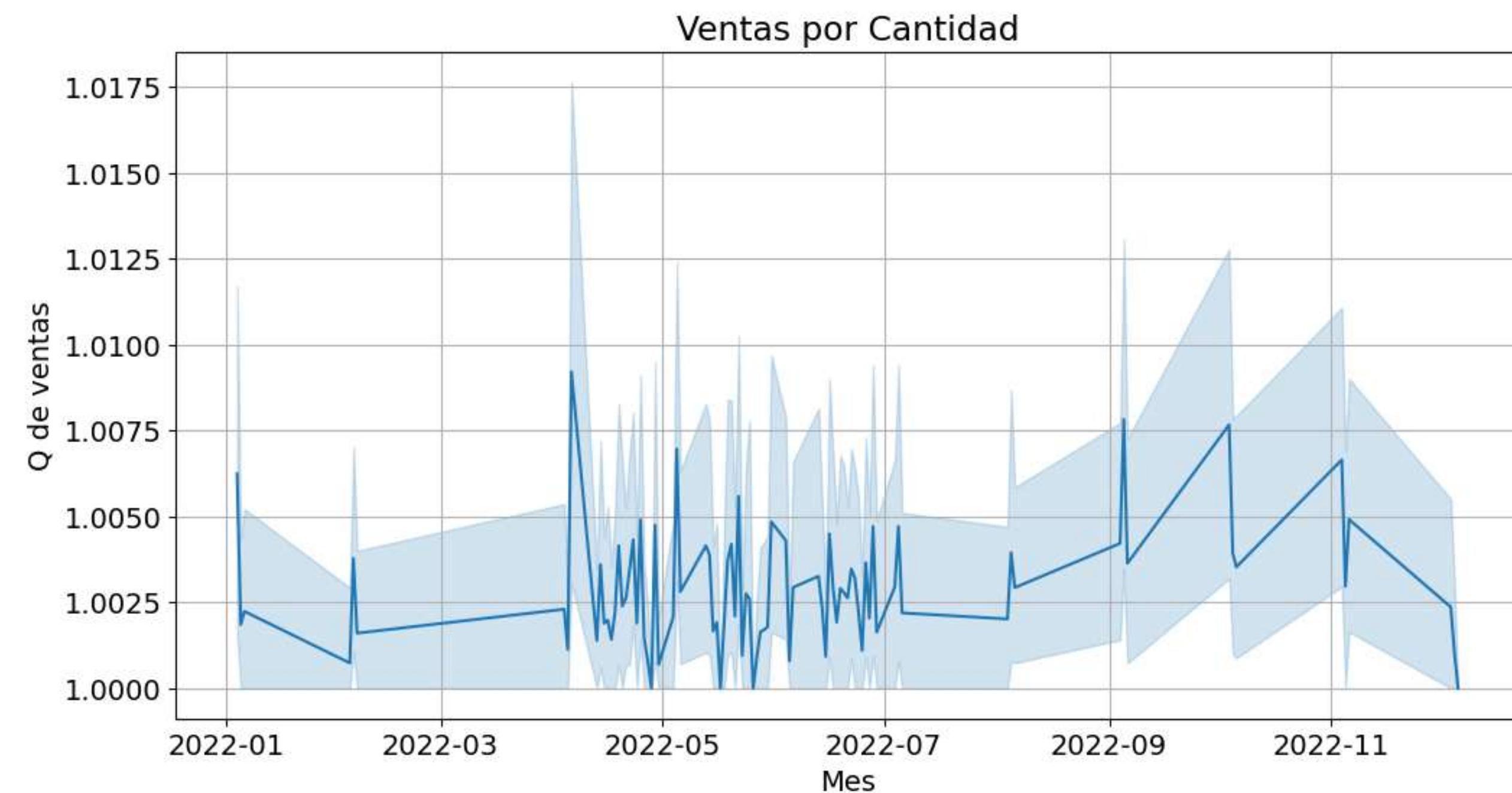
Porcentaje de categoría de producto para ingresos netos
Costo promedio por producto mostrado



Métricas y visualizaciones

3. Ventas por cantidad

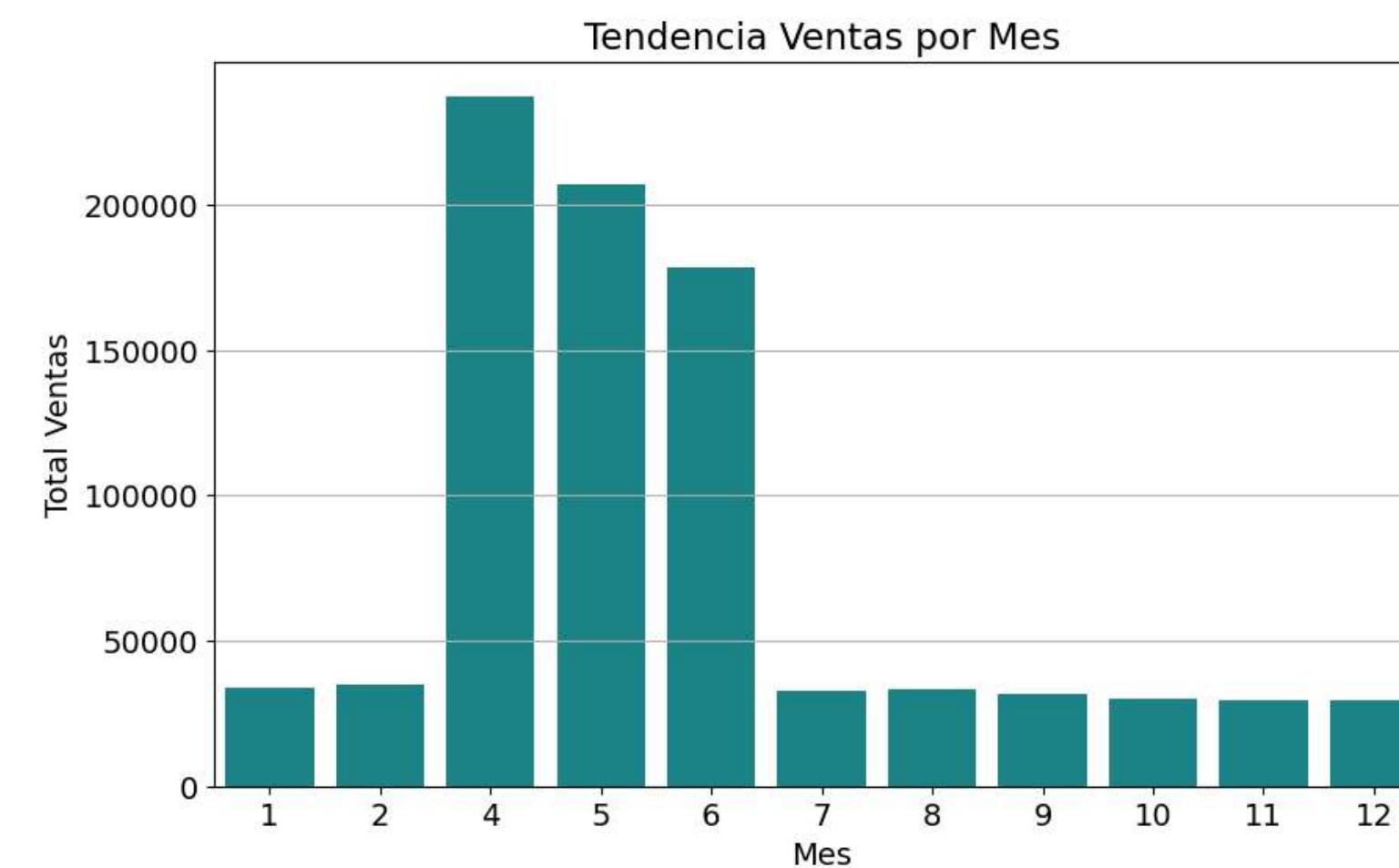
El análisis de las ventas mensuales en unidades resalta la presencia de un patrón estacional en nuestra actividad comercial, con un crecimiento significativo durante los meses de abril a junio. Al ajustar nuestras estrategias comerciales para capitalizar este aumento estacional en la demanda y al mismo tiempo diversificar nuestra oferta de productos, podemos maximizar nuestras oportunidades de crecimiento y mantener una base de clientes sólida y comprometida a lo largo del año.



Métricas y visualizaciones

4. Meses de mayor venta

Las ventas se observan amesetadas a lo largo del año, excepto un pico durante los meses de verano (Abril, Mayo, Junio), que registran hasta 6 veces mas las ventas, lo cual puede deberse a estacionalidad. Siendo Abril el mes con mayor cantidad de ventas.



```
import locale

# Establecer la configuración local para el formato de moneda
locale.setlocale(locale.LC_ALL, 'en_US.UTF-8')

# Formatear la columna 'Total Ventas' con el signo de dólar y puntos de miles
df_mes['Total Ventas'] = df_mes['Total Ventas'].map('${:,.2f}'.format)

print(df_mes)
```

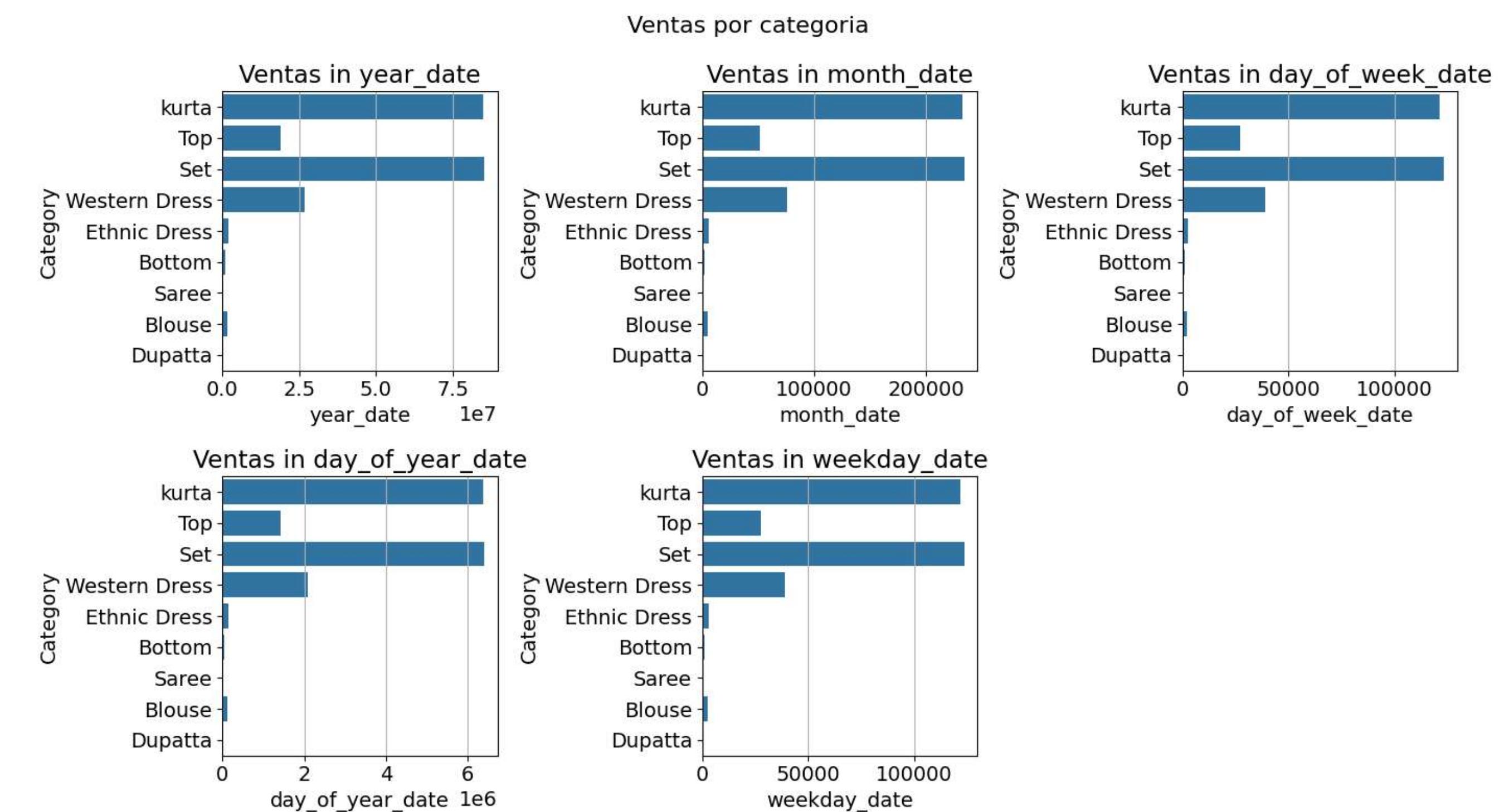
✓ 0.0s

Mes	Total Ventas
0	\$33,817.05
1	\$34,928.22
2	\$237,375.98
3	\$206,814.43
4	\$178,647.19
5	\$32,636.92
6	\$33,306.88
7	\$31,624.47
8	\$30,289.72
9	\$29,429.67
10	\$29,566.71
11	
12	

Métricas y visualizaciones

5. Ventas por periodo

Durante nuestro análisis de ventas, hemos observado patrones distintivos en la distribución de ventas por categoría a lo largo del año. Estos hallazgos ofrecen una visión valiosa sobre cómo varían las preferencias de los clientes en diferentes momentos estacionales, lo que nos permite ajustar nuestras estrategias comerciales de manera más efectiva para capitalizar estas tendencias estacionales.



Definiciones para el modelo

Creé nuevo dataframe llamado df_nuevo con las variables seleccionadas para la muestra y testeo de nuestro modelo, se separa la variable **Amount** debido a que es la variable predictora.

	MSE	R-squared
Linear Regression	0.528369	0.947268
Decision Tree	0.010741	0.998928
Random Forest	0.010199	0.998982



Al evaluar los 3 modelos disponibles, en primera instancia descarté al modelo de Regresión Lineal porque si bien ofrecía un buen rendimiento, sus limitaciones inherentes (debido a la suposición de linealidad) lo hacía menos atractivo.

Siguiente, entre Decision Tree y Random Forest concluí que Random Forest ofrecía una mejor predicción del modelo. **Random Forest** mostró un rendimiento notablemente sólido en términos de sus métricas, con un error cuadrático medio (MSE) muy bajo y un coeficiente de determinación (R-cuadrado) casi perfecto. Estos resultados sugieren que el modelo es altamente efectivo para explicar y predecir la variabilidad en nuestros datos de interés.

Selección del modelo: Random Forest

```
# Definimos el modelo de Random Forest
random_forest = RandomForestRegressor(n_estimators=100, random_state=42)

# Entrenamos el modelo
random_forest.fit(X_train_scaled, y_train)

[31]: ✓ 3.6s
...
RandomForestRegressor
RandomForestRegressor(random_state=42)
```

Validación del modelo: positiva

```
# Predecimos en los datos de prueba
y_pred = random_forest.predict(X_test_scaled)

# Calculamos las métricas de evaluación
mse = mean_squared_error(y_test, y_pred)
r_squared = r2_score(y_test, y_pred)

print("Mean Squared Error:", mse)
print("R-squared:", r_squared)

✓ 0.2s

Mean Squared Error: 0.19789460614655854
R-squared: 0.9820936591101251
```

Resultado de la predicción

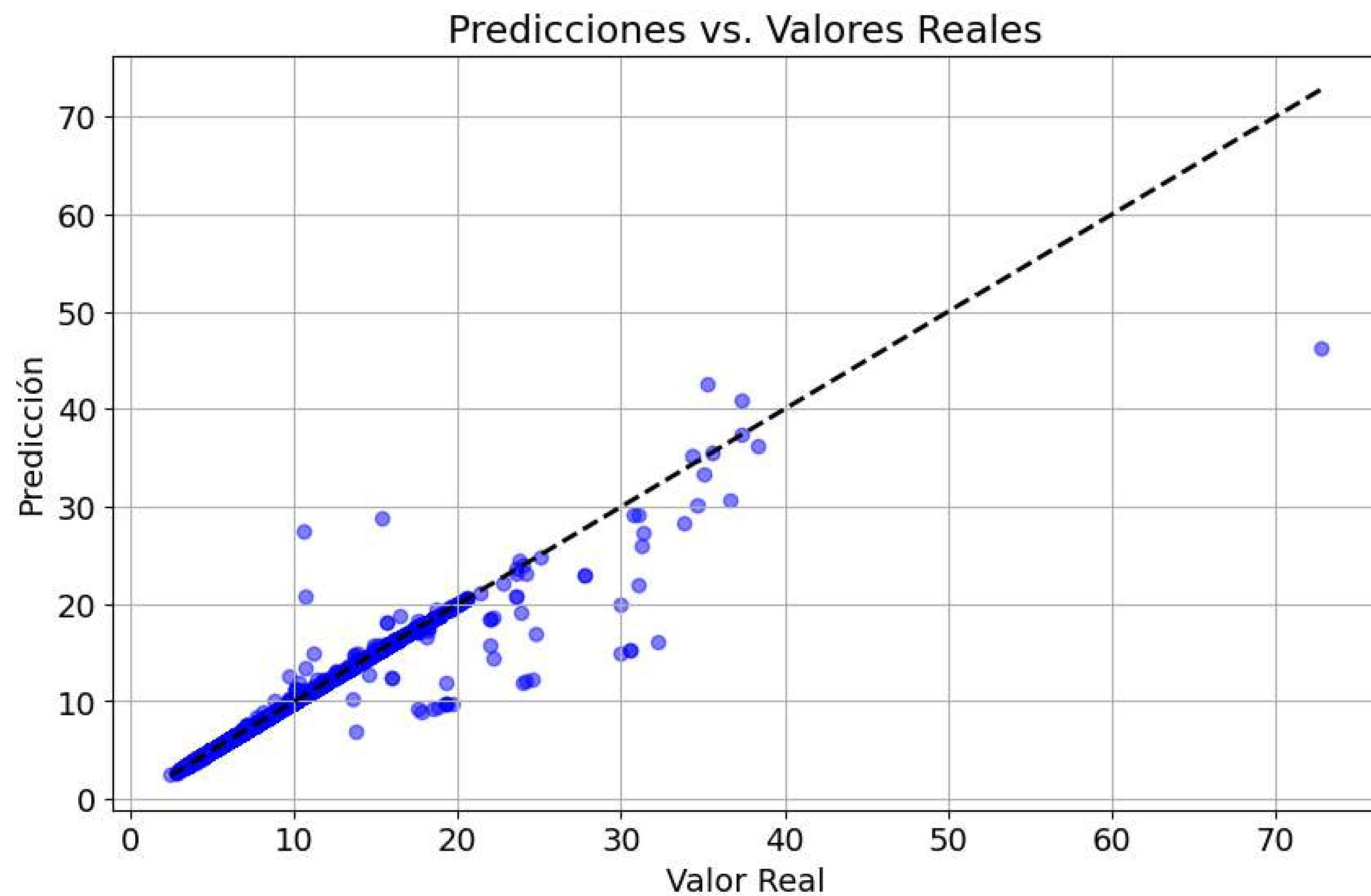
El resultado (imagen inferior) permite concluir que el modelo de Random Forest es el modelo de mejor potencial para este caso.

```
y_pred = random_forest.predict(X_test_scaled)

print("Predicciones:")
print(y_pred)
]
✓ 0.2s

Predicciones:
[5.8921156 6.2550796 7.0052052 ... 8.892618 5.8921156 3.5328496]
```

El gráfico de la derecha muestra que la mayor cantidad de resultados se concentra en una relación positiva al modelo.



Conclusiones

Tras el análisis exploratorio y la implementación de técnicas de limpieza aprendidas en la Clase, transformación y selección de variables, así como la evaluación de los 3 modelos, concluyo que el modelo **Random Forest** se destaca como la mejor para predecir las ventas en basado en este conjunto de datos. La transformación de la variable objetivo a su forma logarítmica y la normalización de los datos contribuyeron significativamente a mejorar la calidad de la predicción.

3 aprendizajes clave obtenidos con este modelo:

- Pude validar que existe un marcado incremento de ventas en torno a las fechas festivas. A esto lo pude comprobar desde el Modelo relacionando la variable target en función de los meses y la cantidad de ventas.

- Pude identificar que dentro del 80/20 de las categorías que más venden, el “20” que menor volumen de ventas ostenta totales similares o mayores al “80” de mayor volumen. En otras palabras, hay una gran ventana de oportunidad para maximizar ingresos.
- Pude validar que hay concentraciones interesantes para una estrategia comercial, como el top 3 de ciudades o el top 3 de categorías que se despegan de la media.

¡Muchas gracias!

Gerson Molina Vaca

gersonmolinavaca@gmail.com

Comisión: 46270