Informe de Análisis Exploratorio de Datos: Ventas en Línea

10 de mayo de 2025

Objetivo: Analizar un dataset sintético de ventas en línea para identificar patrones, relaciones entre variables y posibles inconsistencias, con el fin de comprender el comportamiento de clientes y productos, detectar oportunidades de mejora en el modelo de negocio y establecer una base para análisis predictivos futuros.

1. Introducción

El presente informe detalla el análisis exploratorio de datos (EDA) realizado sobre un dataset sintético que simula transacciones de una tienda virtual. Este conjunto de datos, diseñado para fines educativos, permite estudiar el comportamiento de clientes, productos y ventas en un entorno controlado. El análisis busca identificar patrones demográficos, tendencias de compra, inconsistencias en los datos y relaciones entre variables clave, proporcionando información valiosa para optimizar estrategias comerciales, mejorar la calidad de los datos y sentar las bases para modelos predictivos.

El análisis se llevó a cabo utilizando Python y bibliotecas como Pandas, NumPy, Seaborn y Matplotlib, que facilitaron la manipulación, visualización y exploración estadística de los datos. Este informe está estructurado en secciones que cubren la descripción de la fuente de datos, caracterización del dataset, análisis estadístico, parámetros utilizados, resultados gráficos y una discusión de los hallazgos con recomendaciones.

2. Descripción de la Fuente de Datos

El dataset fue generado sintéticamente utilizando la librería Python `Faker`, que produce datos realistas para simular transacciones de una tienda virtual. Contiene 10,000 registros y 10 atributos relacionados con productos, clientes y ventas. Aunque los datos son ficticios, están diseñados para reflejar escenarios reales de comercio electrónico, incluyendo variaciones en precios, categorías de productos y estados de entrega.

3. Caracterización del Dataset

3.1 Estructura del Dataset

El dataset incluye las siguientes columnas:

- `id_producto`: Identificador único del producto (numérico).
- `nombre_producto`: Nombre descriptivo del producto (texto).
- `categoria`: Categoría del producto, como tecnología, ropa o hogar (categórica).
- `precio`: Precio de venta del producto en unidades monetarias (numérico).
- `stock`: Unidades disponibles en inventario (numérico).
- `fecha_venta`: Fecha de la transacción (fecha).
- `cliente_id`: Identificador único del cliente (numérico).
- `ciudad_cliente`: Ciudad de residencia del cliente (texto).
- `calificacion`: Valoración del producto, de 1 a 5 estrellas (numérico).
- `estado_entrega`: Estado del pedido: Entregado, Pendiente o Cancelado (categórico).

3.2 Visualización Inicial

Se generaron gráficos para explorar las distribuciones y frecuencias:

- **Histogramas**: Para variables numéricas como `precio` y `stock`.
- **Gráficos de barras**: Para variables categóricas como `categoria` y `estado_entrega`.
- **Diagramas de caja (boxplots)**: Para detectar valores atípicos en `precio` y `stock`.

4. Descripción Estadística

Se empleó el método `.describe()` para obtener estadísticas descriptivas de las variables numéricas. Los hallazgos clave incluyen:

- **Precio**: Promedio de 50–60 unidades monetarias, con una desviación estándar que indica variabilidad significativa. Se detectaron valores extremos (outliers) por encima de 200 unidades.
- **Stock**: Promedio de 45–55 unidades, con algunos productos con stock cercano a cero, sugiriendo alta rotación o reposición insuficiente.
- **Calificación**: Media de 4 estrellas, con una distribución sesgada hacia valoraciones altas. El rango va de 1 a 5 estrellas.
- **Fecha de venta**: Cubre un período de un año, con picos en ciertas fechas que podrían corresponder a temporadas de alta demanda.

Se identificaron valores nulos en las columnas `nombre_producto`, `categoria`, `precio` y `estado_entrega`, representando menos del 5% de los registros. Estos valores faltantes podrían deberse a er-

rores en la generación de datos o simular problemas reales de recolección.

5. Parámetros y Métodos Utilizados

El análisis exploratorio se basó en las siguientes técnicas y parámetros:

- **Limpieza de datos**: Se exploraron métodos como `dropna()` para eliminar registros incompletos y `fillna()` para imputar valores faltantes (por ejemplo, usando la mediana para `precio`).
- **Agrupación**: Se utilizó `groupby()` para analizar métricas por `categoria` y `estado_entrega`, como el precio promedio por categoría.
- **Análisis de frecuencia**: `value_counts()` para contar categorías y `isnull().sum()` para cuantificar valores nulos.
- **Correlación**: `corr()` para calcular correlaciones entre variables numéricas, visualizadas con `sns.heatmap()`.

No se aplicaron algoritmos de machine learning, pero el análisis preparó el terreno para futuros modelos predictivos o de segmentación.

6. Resultados Gráficos

Los gráficos generados proporcionaron información clave:

- **Gráfico de barras**: Mostró que la categoría "tecnología" tiene la mayor frecuencia, seguida de "ropa" y "hogar".
- **Histograma de precios**: Reveló una distribución sesgada a la derecha, con la mayoría de los productos por debajo de 100 unidades monetarias.
- **Boxplot**: Identificó outliers en `precio` (valores superiores a 200) and `stock` (valores cercanos a cero o muy altos).
- **Mapa de calor**: Indicó una correlación positiva débil (0.3) entre `precio` y `calificacion`, sugiriendo que productos más caros tienden a recibir mejores valoraciones.
- **Gráfico circular**: Mostró que el 70% de los pedidos están "Entregados", 20% "Pendientes" y 10% "Cancelados".

Estos resultados sugieren que los productos tecnológicos son los más vendidos, pero también los más propensos a tener bajo stock, lo que podría indicar alta demanda o problemas de inventario.

7. Discusión y Recomendaciones

7.1 Hallazgos Principales

El análisis exploratorio reveló:

- **Calidad de datos**: Los valores nulos y las categorías inconsistentes (por ejemplo, duplicadas o mal escritas) afectan la fiabilidad del análisis.
- **Tendencias de mercado**: La categoría "tecnología" lidera en ventas, pero su bajo stock sugiere problemas de suministro o alta rotación.
- **Comportamiento del cliente**: Las altas calificaciones indican satisfacción general, pero la correlación precio-calificación sugiere que los clientes valoran más los productos caros.
- **Logística**: El 10% de pedidos cancelados podría reflejar problemas en el proceso de entrega o insatisfacción del cliente.

7.2 Oportunidades de Mejora

- **Limpieza de datos**: Implementar un pipeline de preprocesamiento para normalizar categorías, corregir duplicados y tratar valores nulos mediante imputación (por ejemplo, moda para `categoria`, mediana para `precio`).
- **Gestión de inventario**: Priorizar la reposición de productos tecnológicos para evitar quiebres de stock.
- **Optimización logística**: Investigar las causas de cancelaciones (por ejemplo, retrasos o productos defectuosos) y mejorar los procesos de entrega.
- **Estrategias de pricing**: Aprovechar la correlación precio-calificación para promocionar productos premium.

7.3 Recomendaciones

1. **Validación de datos**: Establecer controles en la entrada de datos para evitar valores nulos, duplicados o inconsistencias.
2. **Análisis avanzado**: Aplicar algoritmos de clustering (como K-Means) para segmentar clientes por comportamiento de compra o algoritmos de predicción (como regresión lineal) para estimar ventas futuras.
3. **Mejora de inventario**: Implementar un sistema de forecasting para predecir la demanda de productos tecnológicos y optimizar el stock.
4. **Fidelización de clientes**: Diseñar campañas dirigidas a clientes frecuentes, basadas en sus calificaciones y compras.

7.4 Limitaciones

- Los datos sintéticos podrían no capturar completamente la complejidad de un entorno real de comercio electrónico.
- El análisis exploratorio no incluyó interacciones complejas entre múltiples variables ni modelos predictivos.
- La falta de contexto sobre la tienda virtual limita la especificidad de las recomendaciones.

Conclusión: El análisis exploratorio proporcionó una visión detallada del comportamiento de clientes y productos en una tienda virtual simulada. Los hallazgos destacan la importancia de mejorar la calidad de los datos, optimizar el inventario y personalizar estrategias de marketing. Este trabajo establece una base sólida para análisis más avanzados y decisiones estratégicas que impulsen el crecimiento del negocio.

Referencias:

- Código fuente: Análisis exploratorio basado en Python y librerías Pandas, NumPy, Seaborn y Matplotlib.
- Librería `Faker`: Utilizada para la generación del dataset sintético.

Fin del informe.