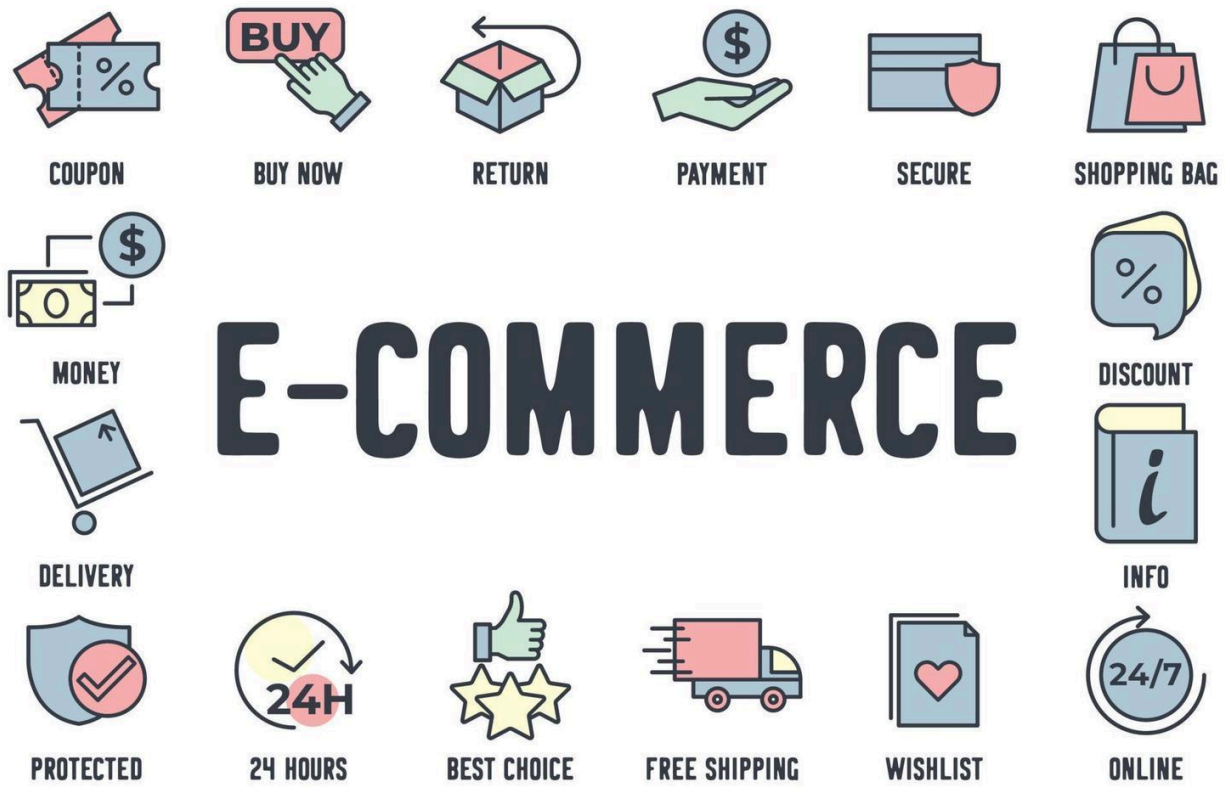


E-Commerce Data Analytics Project

Uncovering Insights for Business Growth

28-01-2025

Aniket kumar
Patna(Bihar)

Google Colab

link:- E-commerce Project.ipynb

1. Project Overview

Introduction

This project analyzes e-commerce data to uncover patterns and trends in sales, profits, and customer behavior. By leveraging data analytics tools and techniques, the project aims to assist businesses in making informed decisions to optimize operations and increase profitability.

Objectives

- To identify key performance metrics such as top-performing products, regions, and customer segments.
- To analyze shipping methods, delivery status, and their impact on profitability.
- To provide actionable insights to improve business strategies.

Importance

Analyzing e-commerce data helps businesses understand their operations better, optimize their supply chain, and create targeted marketing strategies that drive growth.

2. Project Goals

Primary Goals

- Analyze sales, profits, and order quantities to determine trends and patterns.
- Identify top-performing products and regions.
- Understand customer behavior and preferences.

Technical Goals

- Clean and preprocess the dataset to handle missing values and outliers.
- Create insightful visualizations to communicate findings effectively.

Specifications

Dataset Details

- **Columns:** order_date, sales_per_order, profit_per_order, shipping_type, customer_segment, etc.
- **Structure:** Contains rows representing individual transactions with key metrics for analysis.

Technologies Used

- Python: Pandas, NumPy, Matplotlib, Seaborn, Plotly.

Methodology

1. **Data Cleaning**: Remove missing values, handle duplicates, and standardize formats.
2. **Exploratory Analysis**: Use descriptive statistics and correlation analysis.
3. **Visualization**: Create meaningful charts to uncover trends and insights.

Analysis and insights

- This analysis is useful for determining which products generate the most revenue.
- It helps businesses:
    - Focus on best-sellers.
    - Optimize inventory management.
    - Plan targeted promotions or marketing efforts.

```
top_products = data.groupby('product_name')['sales_per_order'].sum().sort_values(ascending=False).head(10)
top_products_df = pd.DataFrame({
    'Product Name': top_products.index,
    'Total Sales': top_products.values})
top_products_df
```

	Product Name	Total Sales
0	Staples	71130.205243
1	Staple envelope	68870.791326
2	Easy-staple paper	61768.255159
3	KI Adjustable-Height Table	28939.058478
4	Staples in misc. colors	26292.450386
5	Avery Non-Stick Binders	25682.898499
6	Staple remover	22711.450373
7	Storex Dura Pro Binders	22395.180410
8	Staple-based wall hangings	21721.308369
9	Situations Contoured Folding Chairs, 4/Set	20660.038355

This analysis is helpful for identifying the most valuable customers who contribute the most to the business's sales. It can be used to:

- Target loyal customers with special promotions.
- Understand customer behavior and purchasing trends.
- Focus marketing efforts on high-value customers.

```
top_customers = data.groupby('customer_id')['sales_per_order'].sum().sort_values(ascending=False).head(10)
top_customers_df = pd.DataFrame({
    'Customer ID': top_customers.index,
    'Total Sales': top_customers.values})
top_customers_df
```

	Customer ID	Total Sales
0	C_ID_54153	3489.950012
1	C_ID_52997	3377.750000
2	C_ID_47998	3199.870025
3	C_ID_41932	3000.000000
4	C_ID_57604	2814.540018
5	C_ID_28784	2781.990036
6	C_ID_42758	2779.880035
7	C_ID_51208	2739.920006
8	C_ID_59491	2707.030014

This analysis is valuable for understanding which product categories are the most profitable. Businesses can use this insight to:

- Focus on expanding or marketing high-profit categories.

- Identify underperforming categories and optimize their strategies.
- Prioritize inventory and resources toward profitable categories.

```
top_categories = data.groupby('category_name')['profit_per_order'].sum().sort_values(ascending=False)
top_categories_df = pd.DataFrame({
    'Category Name': top_categories.index,
    'Total Profit': top_categories.values})
top_categories_df
```

	Category Name	Total Profit
0	Office Supplies	968710.521197
1	Furniture	345606.710524
2	Technology	298662.059737
3	Office Su	0.000000

This analysis is helpful for identifying which regions contribute the most to total sales. Businesses can use this information to:

- **Prioritize marketing campaigns** in high-performing regions.
- **Allocate resources effectively** based on sales potential in each region.
- **Understand regional trends** and target areas for growth or improvement.

```
top_regions = data.groupby('customer_region')['sales_per_order'].sum().sort_values(ascending=False)
top_regions_df = pd.DataFrame({
    'Region': top_regions.index,
    'Total Sales': top_regions.values})
top_regions_df
```

	Region	Total Sales
0	West	4.402020e+06
1	East	3.943938e+06
2	Central	3.197490e+06
3	South	2.232652e+06

This analysis helps businesses understand the performance of different shipping methods. It can be used to:

- **Optimize shipping options** by identifying which types drive the most sales, profit, or order quantities.
- **Evaluate profitability** of different shipping strategies.
- **Improve decision-making** for logistics and customer satisfaction.

```
shipping_performance = data.groupby('shipping_type').agg({
    'sales_per_order': 'sum',
    'profit_per_order': 'sum',
    'order_quantity': 'sum'
}).sort_values(by='sales_per_order', ascending=False)
shipping_performance_df = pd.DataFrame({
    'Shipping Type': shipping_performance.index,
    'Total Sales': shipping_performance['sales_per_order'].values,
    'Total Profit': shipping_performance['profit_per_order'].values,
    'Total Quantity': shipping_performance['order_quantity'].values})
shipping_performance_df
```

	Shipping Type	Total Sales	Total Profit	Total Quantity
0	Standard Class	9.213895e+06	1.070284e+06	92897.0
1	Second Class	2.288718e+06	2.659533e+05	21867.0
2	First Class	1.340832e+06	1.752297e+05	12596.0
3	Same Day	9.326546e+05	1.015126e+05	9409.0

This analysis provides insights into how different customer segments contribute to the business. It can be used to:

1. **Identify top-performing segments** that drive the most sales and profit.
2. **Understand segment-specific needs** to improve customer satisfaction.
3. **Spot underperforming segments** and plan improvements.

```
segment_performance = data.groupby('customer_segment').agg({
    'sales_per_order': 'sum',
    'profit_per_order': 'sum'
}).sort_values(by='sales_per_order', ascending=False)
segment_performance_df = pd.DataFrame({
    'Customer Segment': segment_performance.index,
    'Total Sales': segment_performance['sales_per_order'].values,
    'Total Profit': segment_performance['profit_per_order'].values
})
segment_performance_df
```

	Customer Segment	Total Sales	Total Profit
0	Consumer	7.121963e+06	853756.681177
1	Corporate	4.193593e+06	483525.619858
2	Home Office	2.460544e+06	275696.990423

This analysis helps businesses understand which cities contribute the most to total sales. It can be used to:

1. **Focus marketing and sales efforts** on high-performing cities.
2. **Analyze geographic trends** to identify growth opportunities.

```
top_cities = data.groupby('customer_city')['sales_per_order'].sum().sort_values(ascending=False).head(10)
top_cities_df = pd.DataFrame({
    'City': top_cities.index,
    'Total Sales': top_cities.values})
top_cities_df
```

	City	Total Sales
0	New York City	1.244157e+06
1	Los Angeles	1.053769e+06
2	Philadelphia	7.643363e+05
3	San Francisco	6.987824e+05
4	Seattle	5.829258e+05

This analysis provides insights into the distribution of delivery statuses. It can be used to:

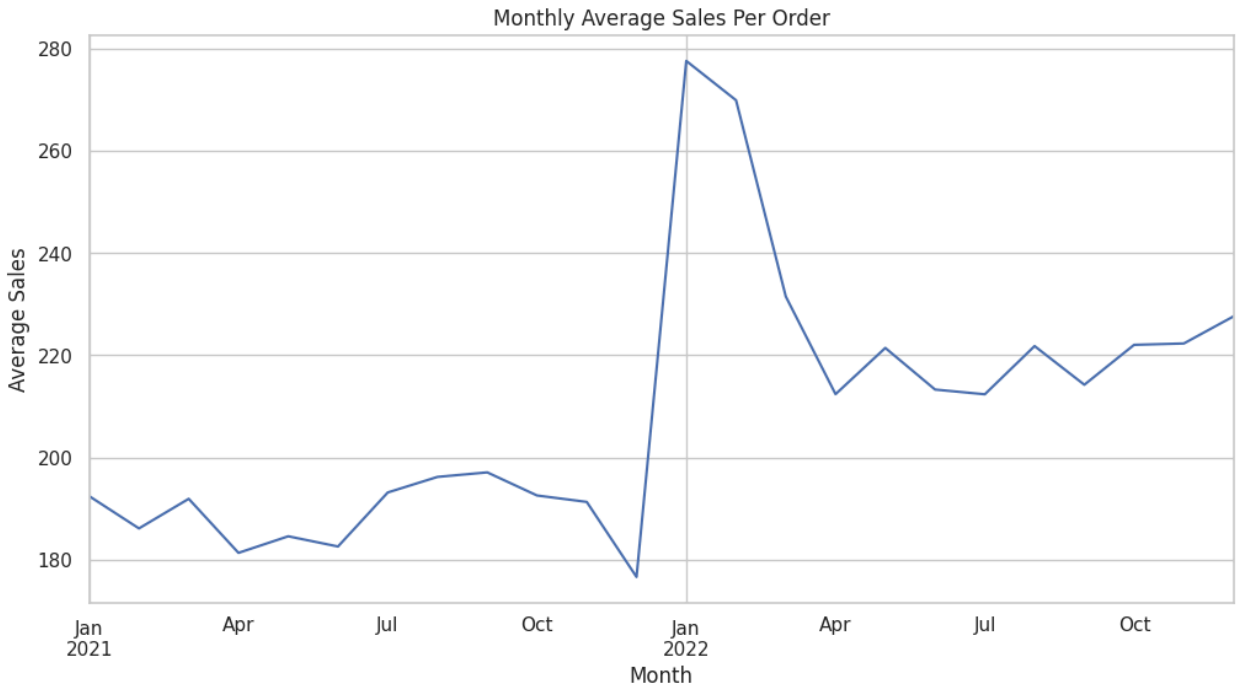
1. **Identify potential issues** (e.g., high "Pending" or "Canceled" counts).
2. **Improve logistics and customer service** by tracking the status of deliveries.

```
delivery_status_count = data['delivery_status'].value_counts()
delivery_status_df = pd.DataFrame({
    'Delivery Status': delivery_status_count.index,
    'Count': delivery_status_count.values})
delivery_status_df
```

	Delivery Status	Count
0	Advance shipping	21697
1	Late delivery	20046
2	Shipping on time	18872
3	Shipping canceled	4404

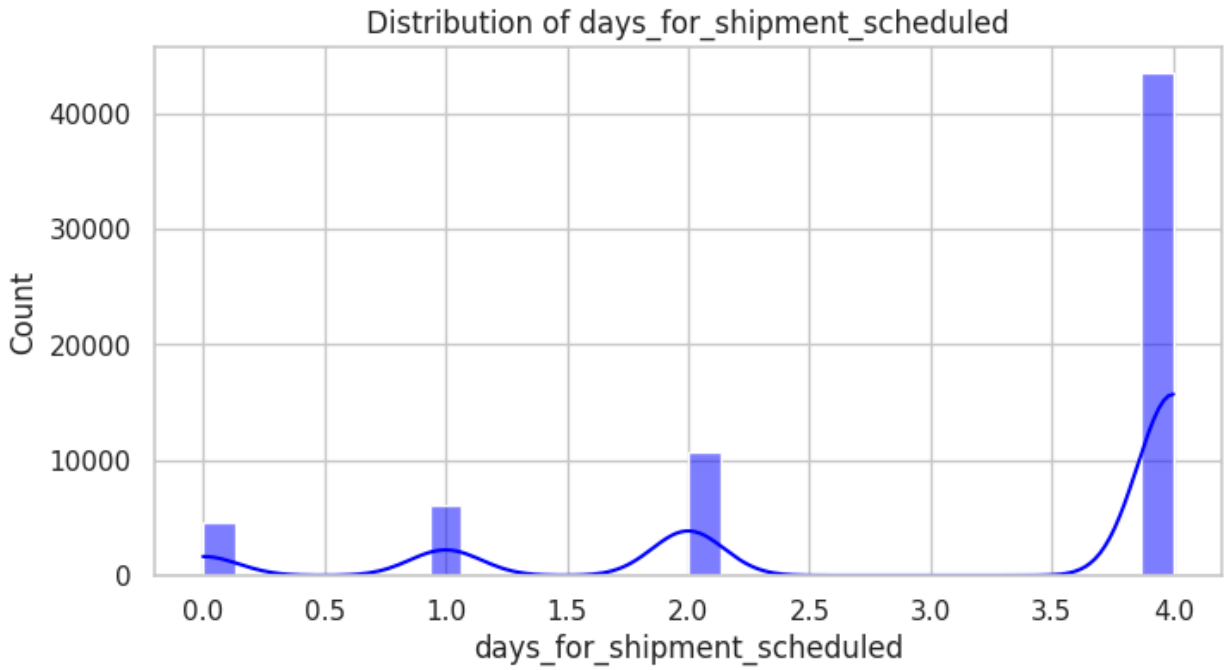
This code analyzes the **monthly sales trend** by calculating the average sales per order for each month and visualizing it with a line plot.

```
data['order_date'] = pd.to_datetime(data['order_date'], errors='coerce')  # Ensure date parsing
sales_trend = data.groupby(data['order_date'].dt.to_period('M'))['sales_per_order'].mean()

plt.figure(figsize=(12, 6))
sales_trend.plot()
plt.title("Monthly Average Sales Per Order")
plt.xlabel("Month")
plt.ylabel("Average Sales")
plt.grid(True)
plt.show()
```



This code generates histograms with kernel density estimation (KDE) for multiple numeric columns in the dataset to visualize their distributions.

```
numeric_columns = ['days_for_shipment_scheduled', 'days_for_shipment_real',
                   'order_item_discount', 'sales_per_order',
                   'order_quantity', 'profit_per_order']

for col in numeric_columns:
    plt.figure(figsize=(8, 4))
    sns.histplot(data[col], kde=True, bins=30, color='blue')
    plt.title(f"Distribution of {col}")
    plt.show()
```



This code calculates the **average profit per order** for each shipping type and visualizes the results as a bar chart.

```
avg_profit_by_shipping = data.groupby('shipping_type')['profit_per_order'].mean()

plt.figure(figsize=(8, 4))
avg_profit_by_shipping.plot(kind='bar', color='teal')
plt.title("Average Profit by Shipping Type")
plt.xlabel("Shipping Type")
plt.ylabel("Average Profit")
plt.xticks(rotation=45)
plt.show()
```

Milestones

- **Data Cleaning**: Removed rows with missing values.
- **EDA**: Analyzed dataset basics and missing data.
- **Key Insights**: Identified top products, order quantities, and revenue by customer state.
- **Correlation**: Visualized relationships between sales, profit, and order data.
- **Trends**: Analyzed monthly average sales.
- **Performance**: Analyzed sales/profit by product, customer, region, and shipping type.
- **Customer Segments**: Identified top customers and cities by sales.

- **Visualization**: Created Sunburst chart for sales by category, product, and region.

The analysis provides a clear view of sales, profit, and performance trends.

Conclusion

This e-commerce data analysis revealed key insights into sales performance, customer behavior, and operational efficiency.

Key Findings:

- **Sales & Profit:** Peak sales vary seasonally, with top products driving revenue.
- **Customer Insights:** A small group of high-value customers contributes significantly to sales.
- **Product Performance:** Some categories are highly profitable, while others underperform.
- **Shipping & Logistics:** Delivery delays highlight a need for supply chain improvements.
- **Customer Segmentation:** Different customer groups show distinct spending patterns.

Recommendations:

- Focus on high-performing products and optimize inventory.
- Implement loyalty programs for high-value customers.
- Improve logistics to reduce delays.
- Use regional sales data for targeted marketing.

This analysis provides valuable business insights, with future opportunities in predictive analytics for enhanced decision-making.