Exploratory Data Analysis (EDA) with Pandas in E-commerce

The purpose of this project is to explore and analyze an e-commerce dataset using the Pandas framework to derive insights into customer behavior, product trends, and sales performance.

Goals of the Project:

- Explore the e-commerce dataset using Pandas.
- Perform feature engineering to derive useful insights.
- Visualize data distributions and trends with various plot types.
- Summarize key findings that can aid in business decision-makin

Materials and Methods

The data for this project is from a simulated e-commerce platform, containing information about orders, products, customer regions, and shipping details. This dataset includes sales data, product categories, order dates, profit margins, and more. The analysis aims to understand sales performance, customer behavior, product trends, and shipping delays.

General Part

- **Libraries Import**: Pandas, NumPy, Seaborn, Matplotlib
- **Dataset Exploration**: Initial exploration of the dataset, checking for missing values, duplicates, and generating summary statistics.
- **Feature Engineering**: Transformation of date columns and creation of new features like shipping delay and profit margin.
- **Visualization in Pandas**: Distribution analysis, relationships between variables, and time-based trends.

Project Outcome & Insights

The project performs **Exploratory Data Analysis (EDA)** on an **e-commerce dataset** to gain meaningful insights into **sales performance, customer behavior, and shipping efficiency**. Below are the key outcomes:

1. Sales Performance

- **Customer Segment Wise Top Sales**: The project groups sales based on different customer segments to identify the most profitable segments.
- **Time Series Analysis**: It shows **sales trends over time**, helping businesses identify seasonal fluctuations and peak sales periods.
- **Top Performing Categories**: Identifies the product categories with the highest sales and revenue.

2. Customer Behavior Analysis

- **Returning Customers**: The analysis helps in understanding customer retention by identifying customers who have made multiple purchases.
- **Top 10 High-Spending Customers**: Helps businesses recognize their most valuable customers and plan targeted marketing strategies.

3. Shipping Performance

- **Shipping Delay Categorization**: The `shipping_category` feature segments shipments into **Same Day, Fast, Moderate, and Delayed**, allowing better logistics management.
- **Profitability by Shipping Type**: Analyzes which shipping methods are most profitable and efficient.

4. Profitability & Business Growth

- **Profit Margin Analysis**: Helps understand **profitability per order** and identify areas for improving profit margins.
- **Year-over-Year Sales Growth**: Tracks annual sales growth percentages, enabling better financial planning.

Feature Engineering:

Created new columns such as:

- **actual_shipping_delay** (Days between order date and ship date).
- **profit_margin** (Profit per order / Sales per order).
- **order_year, order_month, order_weekday** (Extracted from order_date).
- **returning_customer** (Boolean flag indicating repeated customers).
- **shipping_category** (Binned shipping delays into categories: Same Day, Fast, Moderate, Delayed).

Key Questions and Insights to be Addressed:

- What is the total sales by region?

  ```
  sales_by_region = df.groupby('customer_region')['sales_per_order'].sum().sort_values(ascending=False)

  print(sales_by_region)
  ```

  Answer: customer_region

  ```
  West       2.207444e+06

  East       1.970007e+06

  Central    1.621669e+06

  South      1.076112e+06
  ```

- Which product categories have the highest sales?

  ```
  sales_by_category = df.groupby('category_name')['sales_per_order'].sum().sort_values(ascending=False)
  ```

  Answer : Sales by Category:

  ```
   category_name
  Office Supplies    4.115134e+06
  Furniture          1.485964e+06
  Technology         1.274136e+06
  ```

- What is the relationship between sales and profit margins?

```
correlation
=df[['sales_per_order','profit_per_order']].corr()

print(correlation)
```

Answer:

Correlation between sales and profit:

```
                 sales_per_order  profit_per_order

sales_per_order         1.000000         0.130008

profit_per_order        0.130008         1.000000
```

- How does the sales trend change over time?

```
monthly_sales=df.groupby(['order_year','order_month'])['sales
_per_order'].sum()

print(monthly_sales)
```

Answer:

```
order_year order_month  sales_per_order

0       2021     2021-01    278026.849371

1       2021     2021-02    359211.126956

2       2021     2021-03    428262.210236

3       2021     2021-04    370351.629310

4       2021     2021-05    365224.566823

5       2021     2021-06    402157.323752

6       2021     2021-07    350417.948708
```

```
7       2021    2021-08    295852.041774
8       2021    2021-09    257718.098723
9       2021    2021-10    109834.018045
10      2021    2021-11     39807.478701
11      2021    2021-12     24551.260457
12      2022    2022-01    427267.288370
13      2022    2022-02    425011.068305
14      2022    2022-03    409340.048349
15      2022    2022-04    388161.257242
16      2022    2022-05    414351.287871
17      2022    2022-06    381887.037021
18      2022    2022-07    372165.315867
19      2022    2022-08    310705.086154
20      2022    2022-09    276254.015330
21      2022    2022-10    115058.862121
22      2022    2022-11     58542.990361
23      2022    2022-12     15073.950001
```

- What is the average shipping delay, and how does it vary by shipping type?

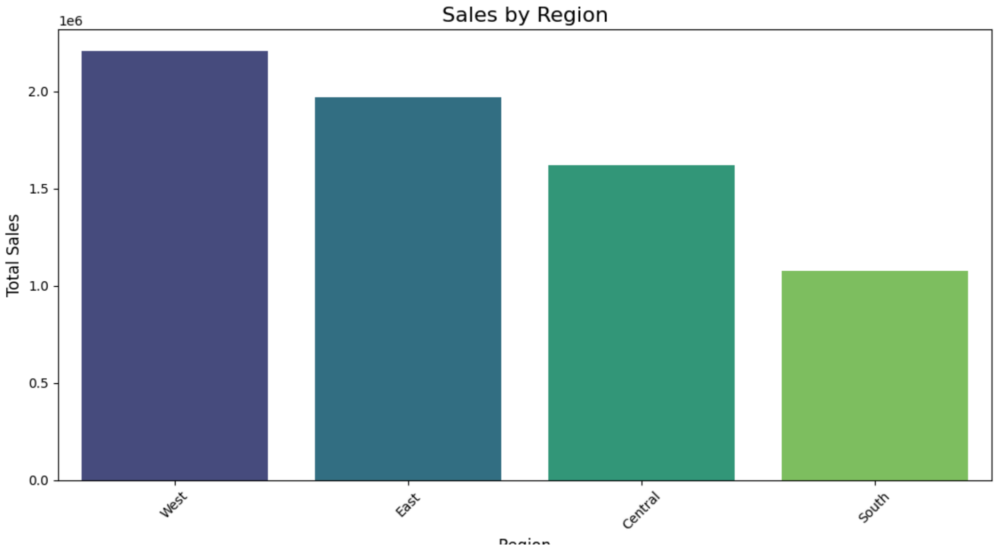
```
avg_shipping_delay = df.groupby('shipping_type')['actual_shipping_delay'].mean()

print(avg_shipping_delay)
```

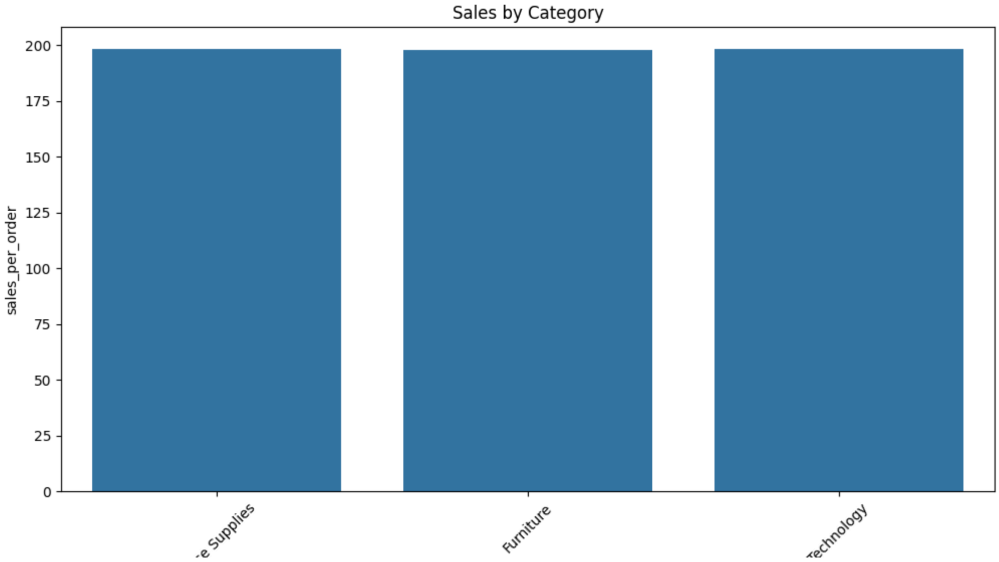
Visualization:

Several charts created to present inside including:

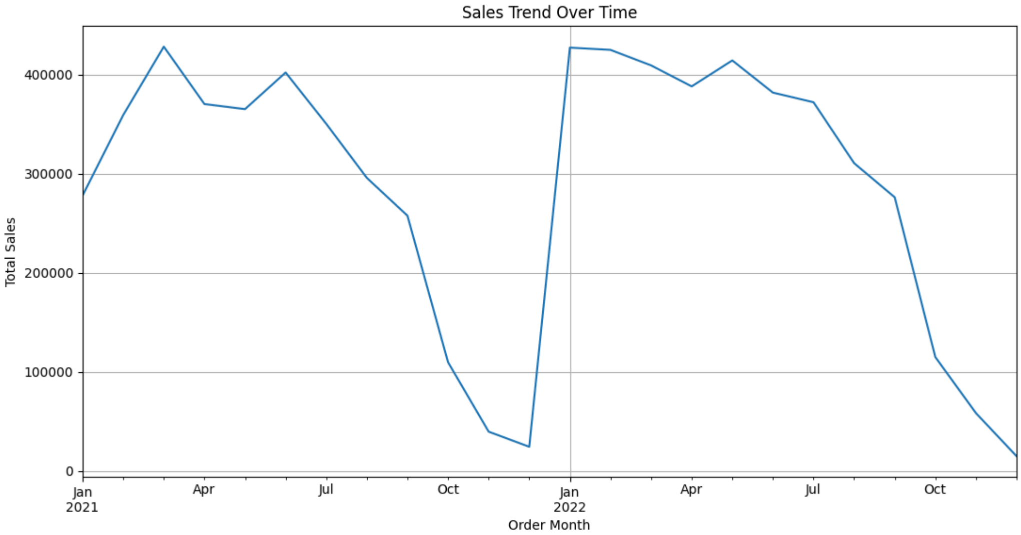
- Sales by region (Bar chart)



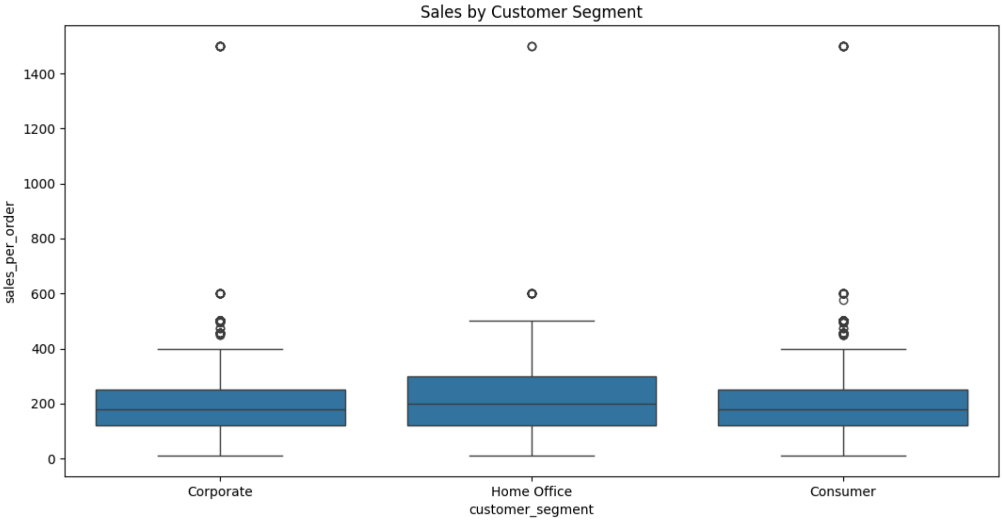
- Sales by category (Bar chart)



- Sales trends over time (Line chart)



- Boxplot showing sales distribution across customer segments


Sales by Customer Segment
sales_per_order
1400
1200
1000
800
600
400
200
0
Corporate
Home Office
Consumer
customer_segment