Exploratory Data Analysis (EDA) with Pandas in Retail Store

The purpose of this project is to explore and analyse the Retail Store dataset using the Pandas framework to derive insights into customer behaviour, product trends, and sales performance.

The goal of this analysis is to understand the structure, quality, and trends within the retail e-commerce data to make informed business decisions. The data contains information about products, customers, sales, and transactions.

Goals of the Project:

- Explore the Retail store 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-making.

Materials and Methods

The data for this project is from a Retail store platform, containing information about orders, products, Payment MEthod, and Order quantity. This dataset includes sales data, product categories, order dates, Order quantity, and more. The analysis aims to understand sales performance, customer behaviour, product trends, and preferred shopping mode.

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 **Retail store Dataset** to gain meaningful insights into **sales performance, customer behavior, and Payment Methods**. Below are the key outcomes:

1. Sales Performance

- **Sales by Product Category**: The project groups sales based on different product category to identify the most popular products in the retail store.

- **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 Behaviour 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. Profitability & Business Growth

- **Top Performing Categories**: Helps understand **Top performing Categories** and identify areas for improving Product Availability and quality
- **Year-over-Year Sales Growth**: Tracks annual sales growth percentages, enabling better financial planning.

Feature Engineering:

Created new columns such as:

- **order_year, order_month, order_weekday** (Extracted from order_date).
- **returning_customer** (Boolean flag indicating repeated customers).

Key Questions and Insights to be Addressed:

- What is the total sales by Category?

```
sales_by_category =
df.groupby('Category')['Quantity'].sum().sort_values(ascending=False)
print("\nSales by Category:\n", sales_by_category)
```

- Which product categories have the highest sales?

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

Answer : Sales by Category:

```
 Sales by Category:

 Category

Food                                 7449.042839

Milk Products                        7402.641686

Beverages                            7383.509890

Furniture                            7368.443992

Butchers                             7355.042839

Computers and electric accessories   7233.641686

Electric household essentials        7179.108737

Patisserie                           7139.504125
```

```
Name: Quantity, dtype: float64


Money Collected by a product Categorys:

 Category

Butchers                              186374.684972

Electric household essentials         175502.836100

Beverages                             175168.359408

Food                                  172468.684972

Furniture                             170208.208279

Computers and electric accessories    165179.161664

Patisserie                            164055.242869

Milk Products                         158660.661664
```

- How does the sales trend change over time?

  ```
  monthly_sales = df.groupby(['order_year',
  'order_month'])['Quantity'].sum().reset_index()

  print("Monthly Sales Trend:\n", monthly_sales)
  ```

  Answer:

  ```
   . Year-over-Year Sales Growth:

   order_year

  2022          NaN

  2023   -0.030834

  2024    0.085171

  2025   -0.950822
  ```

```
Name: Total Spent, dtype: float64
Monthly Sales Trend:
     order_year order_month     Quantity
0          2022     2022-01  1896.191929
1          2022     2022-02  1582.461286
2          2022     2022-03  1591.862439
3          2022     2022-04  1520.263592
4          2022     2022-05  1469.461286
5          2022     2022-06  1577.928337
6          2022     2022-07  1699.191929
7          2022     2022-08  1693.395388
8          2022     2022-09  1543.263592
9          2022     2022-10  1482.461286
10         2022     2022-11  1628.191929
11         2022     2022-12  1504.658980
12         2023     2023-01  1870.796541
13         2023     2023-02  1443.395388
14         2023     2023-03  1521.994235
15         2023     2023-04  1424.461286
16         2023     2023-05  1607.994235
17         2023     2023-06  1642.928337
18         2023     2023-07  1687.527184
19         2023     2023-08  1345.796541
20         2023     2023-09  1499.395388
21         2023     2023-10  1492.928337
```

```
22        2023     2023-11  1372.862439
23        2023     2023-12  1590.060133
24        2024     2024-01  1835.658980
25        2024     2024-02  1472.928337
26        2024     2024-03  1619.593082
27        2024     2024-04  1682.461286
28        2024     2024-05  1698.593082
29        2024     2024-06  1658.461286
30        2024     2024-07  1602.197694
31        2024     2024-08  1695.527184
32        2024     2024-09  1490.329490
33        2024     2024-10  1580.395388
34        2024     2024-11  1603.395388
35        2024     2024-12  1901.126031
36        2025     2025-01   980.796541
```

- What is the preferred Shoping mode?

```
 sales_by_location =
df.groupby('Location')['Quantity'].sum().sort_values(ascending=False)

print("\nSales by location:\n", sales_by_location)
```

```
 Answer:
```

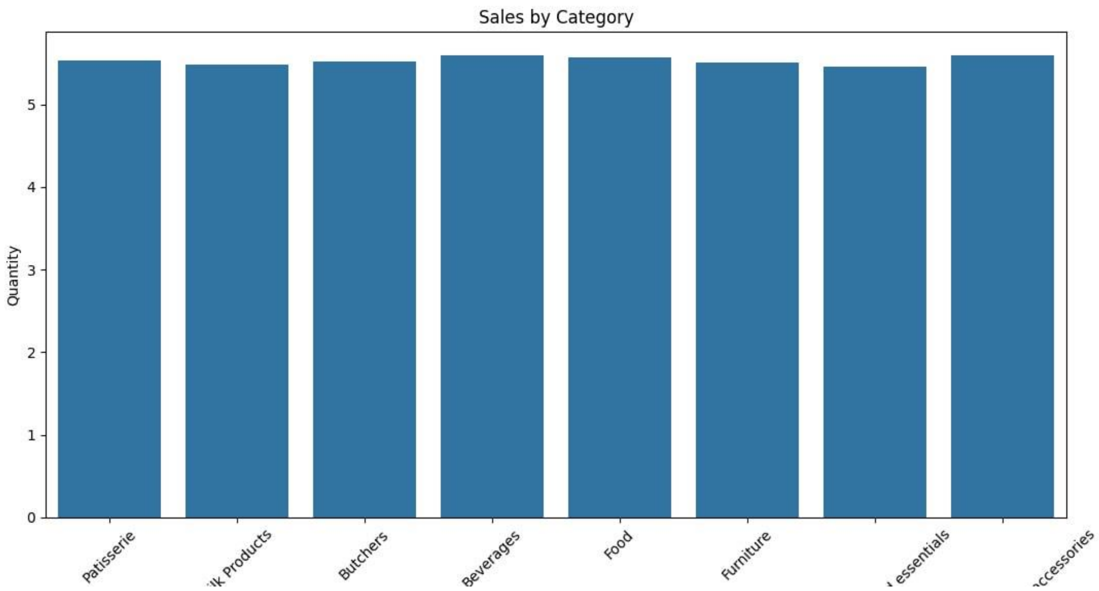
```
Sales by location:

Location
```

In-store	34304.477786
Online	24206.458006

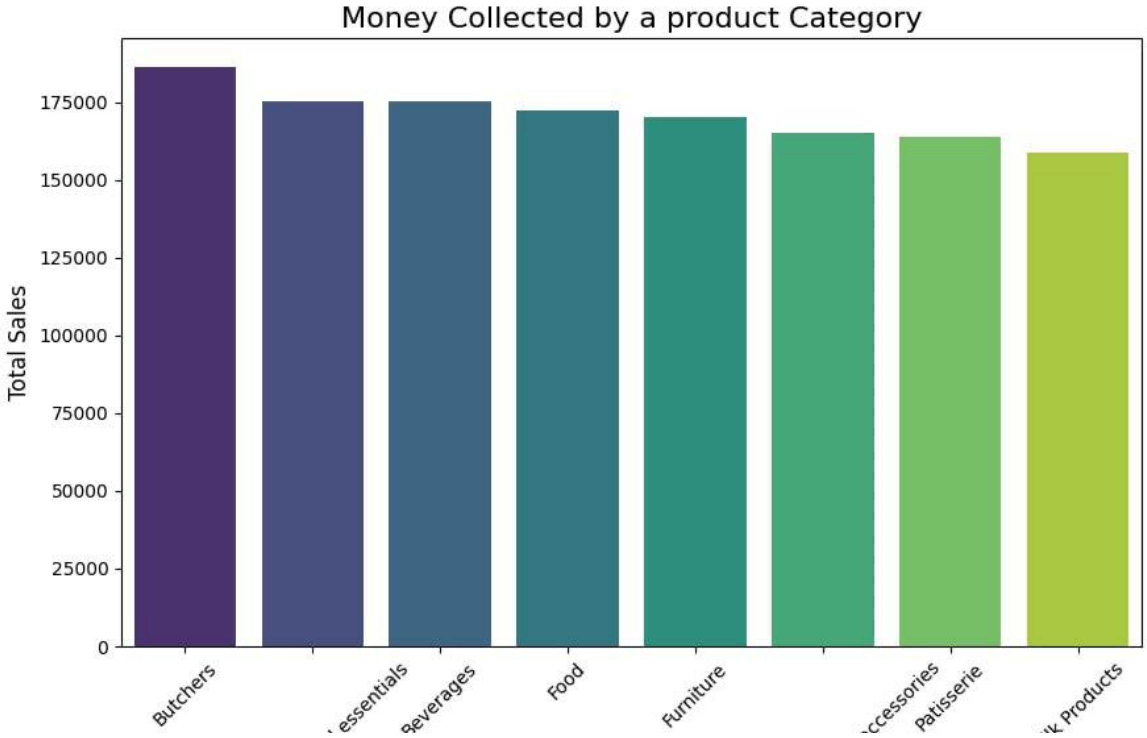
Visualization:

Several charts created to present inside including:

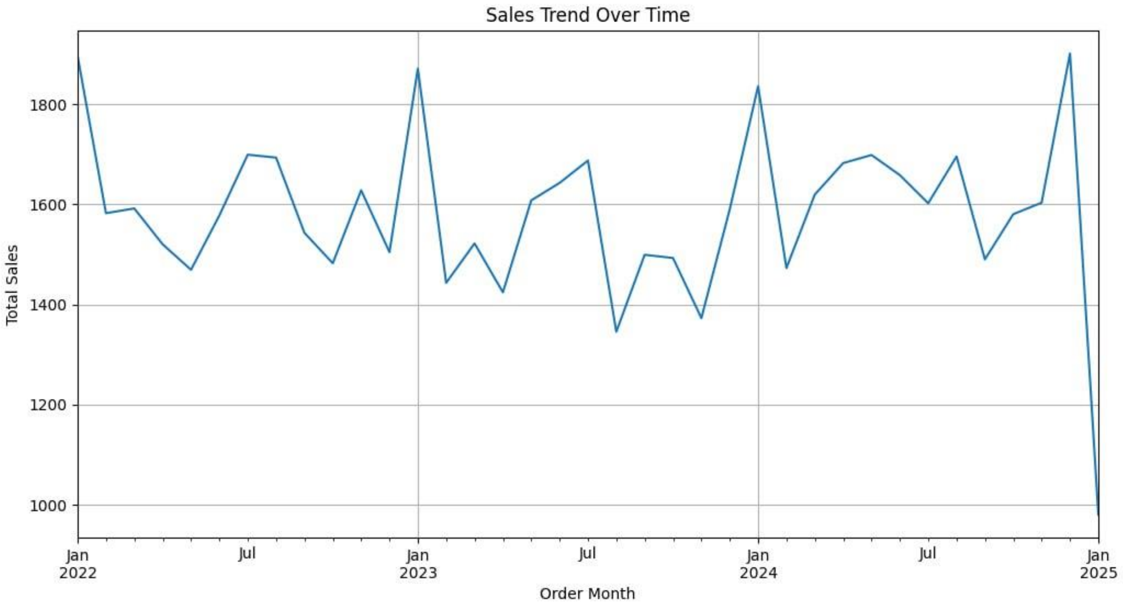
- Sales by Categoryn (Bar chart)



- Money collecetby product category (Bar chart)



- Sales trends over time (Line chart)


Sales Trend Over Time
Total Sales
1000
1200
1400
1600
1800
Order Month
Jan 2022
Jul
Jan 2023
Jul
Jan 2024
Jul
Jan 2025