Customer Shopping Behavior Analysis

1. Project Overview

This dataset contains 504 retail transactions from shopping malls in Istanbul, capturing customer demographics, product categories, quantities, prices, payment methods, and purchase dates to analyze sales performance and shopping behavior.

2. Dataset Summary

- Rows: 504

- Columns: 10

- Key Features:

  - Customer demographics: Gender, Age
  - Transaction details: Invoice Number, Customer ID, Category, Quantity, Price, Invoice Date
  - Shopping context: Shopping Mall, Payment Method

- Missing Data: 1 missing value in each field except Invoice Number (total 9 missing values)

3. Exploratory Data Analysis using Python

We began with data preparation and cleaning in Python:

- **Data Loading:** Imported the dataset using pandas.

- **Initial Exploration:** Used df.info()to check structure and .describe()for summary statistics.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 504 entries, 0 to 503
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Invoice Number  504 non-null    object 
 1   Customer ID     503 non-null    object 
 2   Gender          503 non-null    object 
 3   Age             503 non-null    float64
 4   Category        503 non-null    object 
 5   Quantity        503 non-null    float64
 6   Price           503 non-null    float64
 7   Payment Method  503 non-null    object 
 8   Invoice Date    503 non-null    object 
 9   Shopping Mall   503 non-null    object 
dtypes: float64(3), object(7)
memory usage: 39.5+ KB
```

	count	mean	std	min	25%	50%	75%	max
Age	503.0	40.862823	13.424982	18.00	30.0	39.00	51.00	72.0
Quantity	503.0	2.697813	1.321043	1.00	2.0	2.00	4.00	5.0
Price	503.0	1206.421948	3064.507191	5.23	41.9	143.36	900.24	26250.0

- **Missing Data Handling**: Checked for null values and removed the single record containing missing values to keep the dataset clean.

- **Duplicate & Type Cleaning**: Checked for fully duplicated rows and standardized data types by converting Invoice Date to datetime and casting Age and Quantity to integer.

- **Feature Engineering:**

    - Created an Age Group column by binning customer ages into four segments (Young Adult, Adult, Middle-aged, Senior) using quantile-based binning.
    - Created a Total Price column as Quantity × Price and rounded values to two decimals for revenue analysis.

- **Outlier Analysis**: Used the IQR method on the Total Price column to detect extreme purchase values for further inspection.

- **Database Integration**: Connected Python to SQL Server and loaded the cleaned dataset into the data_cleaned table, only inserting invoices that do not already exist in the database.

4. Data Analysis using SQL (Business Transactions)

We performed structured analysis in PostgreSQL to answer key business questions:

1. **Revenue by Gender**

	Gender	Revenue
1	Female	690682.049999999
2	Male	1141076.83

2. **Revenue by month and year**

	Year	Month	Revenue
1	2023	1	35734.04
2	2023	2	476438.9
3	2023	3	81732.82
4	2023	4	82719.81
5	2023	5	28970.64
6	2023	6	16120.79
7	2023	7	21988.62
8	2023	8	55201.08
9	2023	9	91814.58
10	2023	10	42265.08
11	2023	11	13494.59
12	2023	12	83087.73
13	2024	1	64839.53
14	2024	2	270527.25
15	2024	3	41876.87
16	2024	4	67152.46

3. **Revenue by day of the week**

	WeekdayName	Total_Orders
1	Sunday	79
2	Monday	62
3	Tuesday	74
4	Wednesday	73
5	Thursday	85
6	Friday	73
7	Saturday	57

4. **Revenue by age group**

	Age_Group	Revenue
1	Adult	482589.84
2	Middle-aged	301293.29
3	Senior	797035.47
4	Young Adult	250840.28

5. **Payment methods by order**

	Payment_Method	Revenue
1	Cash	333130.16
2	Credit Card	1263294.46
3	Debit Card	235334.26

6. **Top 3 category by revenue**

	Category	Revenue
1	Food & Beverage	3901.58
2	Souvenir	7856.25
3	Books	10620.15

7. **Bottom 3 category by revenue**

	Category	Revenue
1	Technology	1397550
2	Clothing	214557.2
3	Shoes	158142.16

8. **Top 3 shopping mall by revenue**

	Shopping_Mall	Revenue
1	Metrocity	638736.11
2	Forum Istanbul	273234.97
3	Cevahir AVM	270973.46

9. **Bottom 3 shopping mall by revenue**

	Shopping_Mall	Revenue
1	Kanyon	41097.08
2	Zorlu Center	76642.96
3	Akasya	101129.09

10. **Top 3 Category by Quantity Sold**

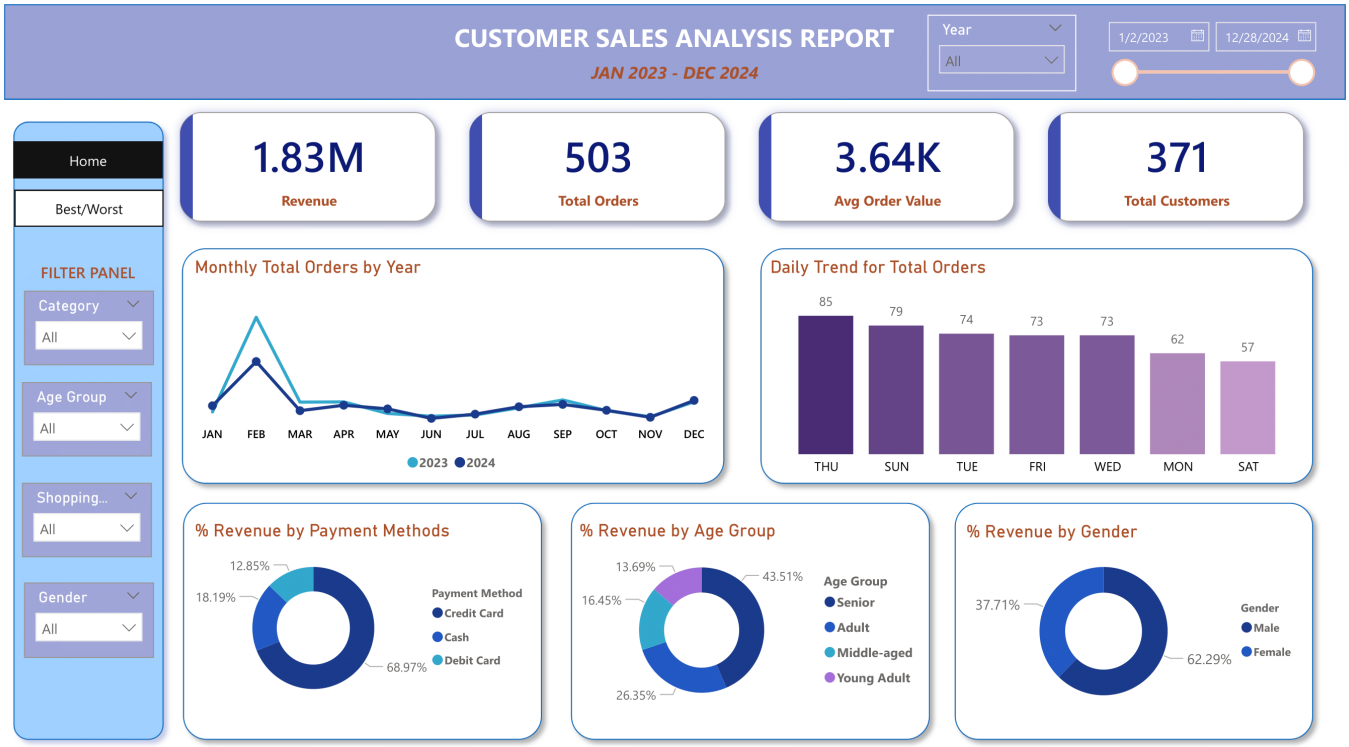
	Category	Total_Quantity
1	Clothing	215
2	Food & Beverage	206
3	Books	187

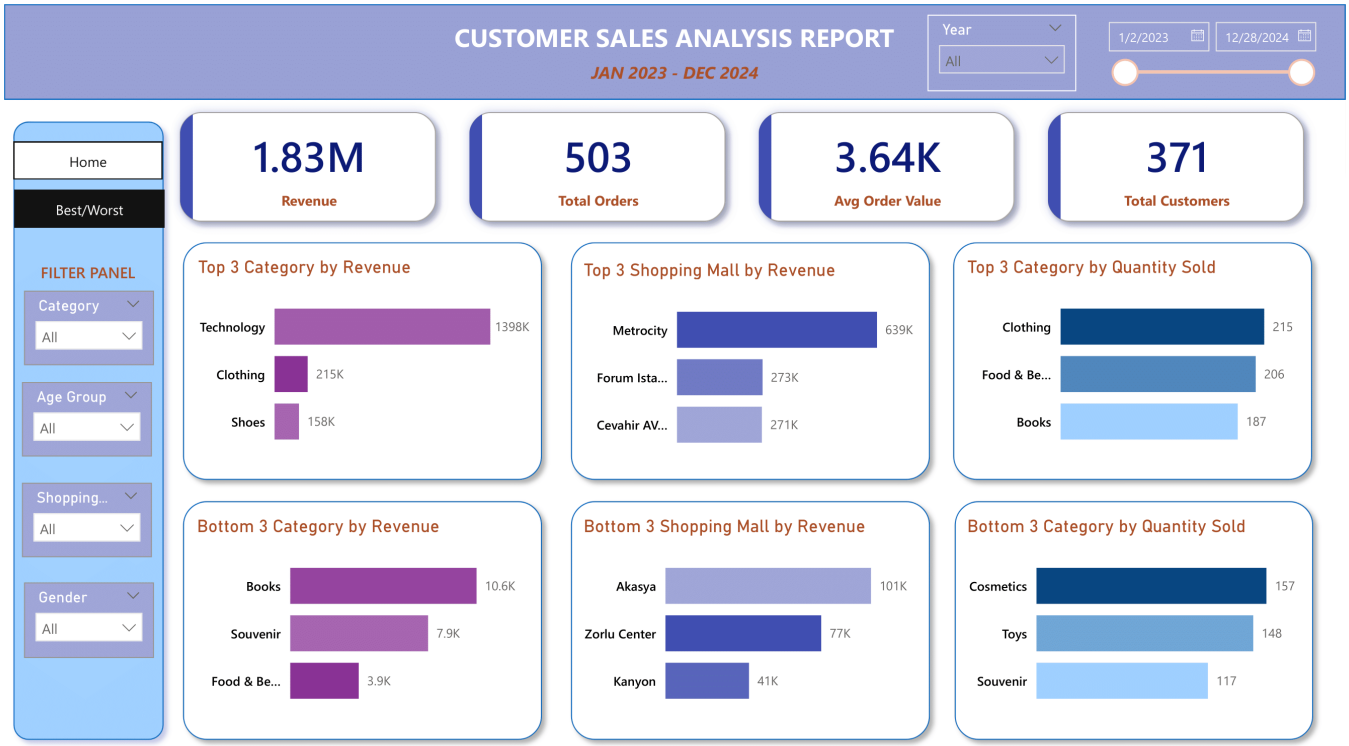
11. **Bottom 3 Category by Quantity Sold**

	Category	Total_Quantity
1	Souvenir	117
2	Toys	148
3	Cosmetics	157

5. Dashboard in Power BI

Finally, we built an interactive dashboard in **Power BI** to present insights visually.





6. Business Recommendations

- **Push Top Categories (Technology, Clothing, Shoes)** – Use **Technology, Clothing, and Shoes** (top 3 categories by revenue) as hero products in campaigns, bundle deals, and homepage highlights to maximize sales.
- **Fix Weak Categories (Food & Beverage, Souvenir, Books)** – These are the **bottom 3 categories by revenue**; test short-term discounts, bundles with top categories, or consider reducing shelf space if performance doesn't improve.
- **Focus on High-Value Age Groups** – Prioritize marketing budget and tailored offers for the age groups that show the highest spending in the analysis (e.g., targeted vouchers, member days).
- **Segment Malls by Performance** – Treat high-revenue malls as flagship locations, while using local promotions and events to support low-revenue malls.
- **Promote Preferred Payment Methods** – Ensure a smooth checkout experience for the most-used payment methods (especially Credit Card) and offer small incentives for digital payments.