CUSTOMER DATA ANALYSIS

A Project Report

Certified Data Analysis

By

PTID-CDA-NOV-25-864

Kattamreddy Jahnavi

Eragaraju Manoj Kumar

Bakka Vijay Raju

Ponguru Venkata Sai Sudha

CUSTOMER DATA ANALYSIS

Introduction

Customer behaviour has become one of the most valuable sources of insight for modern businesses. Understanding how customers shop, what products they prefer, how much they spend, and which payment methods they use helps companies improve sales strategies, product placement, and overall customer satisfaction.

This project focuses on analysing customer shopping data collected from 10 different shopping malls in Istanbul between 2021 and 2023. The dataset contains detailed information about customer demographics, purchasing patterns, product categories, payment methods, and transaction timestamps. By examining these factors, the project aims to uncover meaningful trends and patterns that can support data-driven decision-making.

the data is taken from a customer dataset and analyzed using different methods like charts, tables, and basic statistics. The main aim is to find important patterns and trends that will help businesses make better decisions and the tools are covered are SQL, Excel, orange, Power BI, etc. Overall, this project shows how data analysis can be used to understand customers and make smarter business choices.

Objective of the project

The primary objective of this project is to analyze customer shopping behavior across multiple shopping malls using the provided dataset. The aim is to derive meaningful insights that can help the company understand trends, improve decision-making, and enhance business strategies. Specifically, the project focuses on we covered a database samples, according to that we have defined the insights and interpretation about the projects.

Understanding shopping distribution across gender and age groups to identify the major customer segments.

Determining which gender and age categories purchase more products and generate higher revenue.

Analysing product category performance in relation to demographic and transactional attributes.

Studying the relationship between payment methods and other customer attributes, such as gender, age, and shopping malls.
Identifying spending patterns using quantity, price, and purchase frequency.
Visualizing the dataset in Power BI/Tableau to present insights clearly and support decision-making.
Providing data-driven recommendations to help shopping mall management and retailers enhance marketing, product placement, and customer engagement strategies

Dataset Description;

The dataset used in this project contains detailed customer transaction records collected from 10 shopping malls in Istanbul between 2021 and 2023. It provides comprehensive information on customer demographics, purchase behavior, payment preferences, and revenue contribution. The dataset is structured and stored in a MySQL schema named customer sales analysis.

1. Invoice no

- A unique transaction ID generated for every purchase.
- Helps identify and track individual customer purchases.

2. Customer_id

- A unique identifier assigned to each customer.
- Used for grouping purchases made by the same person over time.

3. Gender

- Specifies the gender of the customer: *Male* or *Female*.
- Used for analyzing gender-based shopping trends.

4. Age

- Represents the age of the customer at the time of purchase.
- Useful for age-wise segmentation, revenue analysis, and understanding buying behavior.

5. Category

- Product category purchased by the customer, such as:
  - Clothing

  - Cosmetics
  - Books
  - Technology
  - Food & Beverages
- Helps in identifying top-selling and least-performing categories.

6. Quantity

- Number of units/products purchased in a single transaction. Used for analysing total product sales and customer demand.

7. Price

- Unit price of the individual product purchased.
- When multiplied with quantity, determines the revenue for the transaction.

8. Payment_method

- Mode of payment used by the customer:
  - Cash
  - Credit Card
  - Debit Card
  - UPI
- Helps in understanding preferred payment channels.

9. Shopping_mall

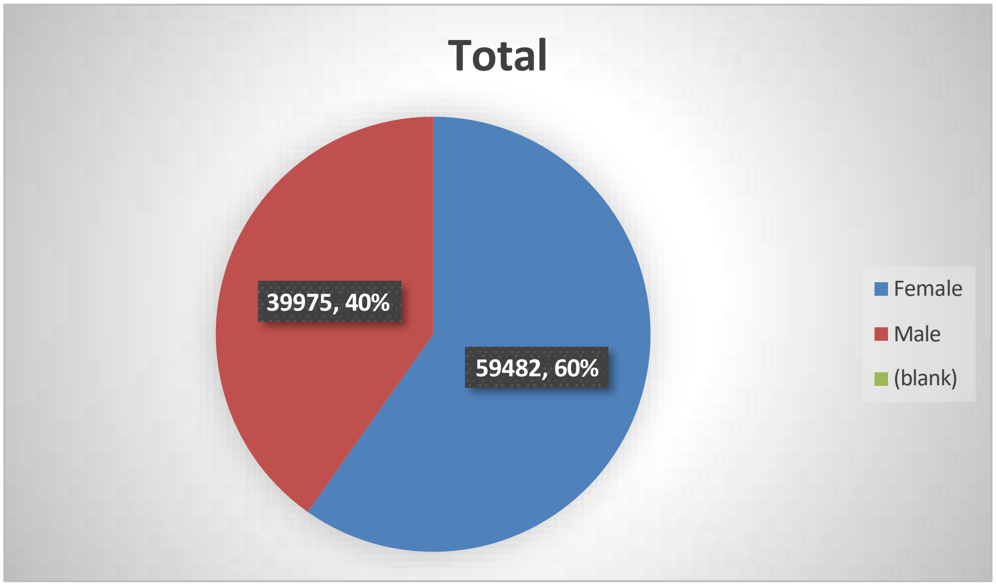
- Name of the mall where the purchase was made.
- Used to compare performance among 10 malls in Istanbul.

10. Revenue

- Total amount collected in a transaction.
- Calculated as:
  Revenue = Quantity × Price
- Primary metric used for evaluating sales performance.

Task 1: Shopping Distribution according to the gender.

This task analyzes how many shopping transactions were made by male and female customers. Using PivotTables or BI tools, the dataset is grouped by gender and the number of invoices is counted. This helps identify which gender is more active in shopping.



Interpretation

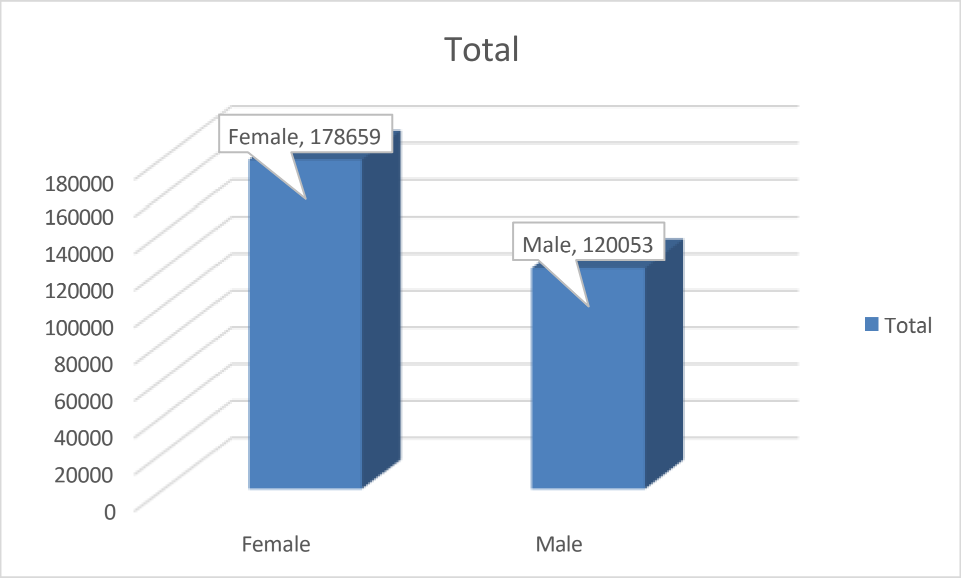
Here the Female customer made a transaction about 59482 and the male contributed about 39975, where majorly the female customers where making shopping distribution (60%) than the male customers (40%).

Row Labels	**Count of invoice_no**
Female	59482
Male	39975
(blank)	
Grand Total	**99457**

Task2: Which gender did we sell more products to?

Compare which Gender contribute more revenue to the company

Row Labels	Sum of quantity
Female	178659
Male	120053
Grand Total	**298712**



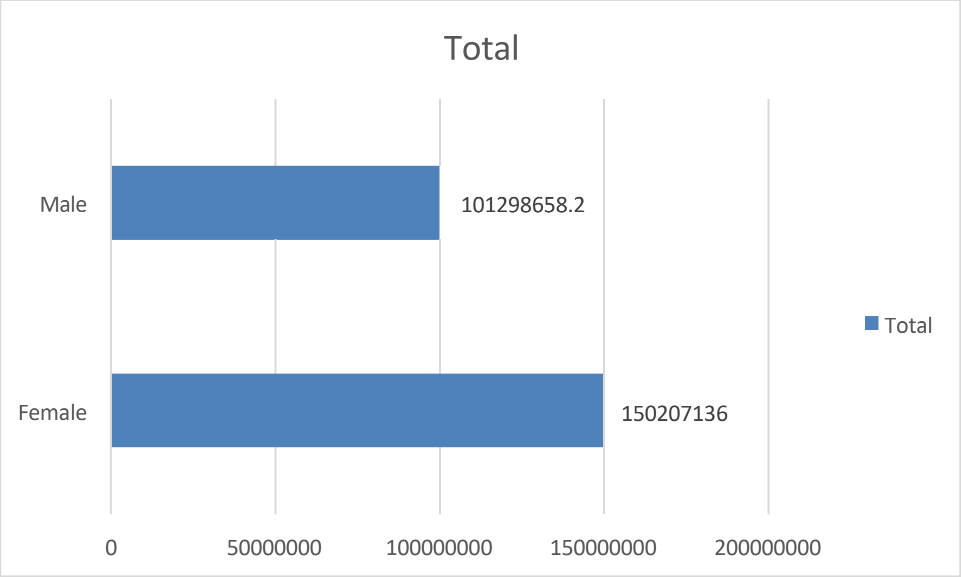
Here, the total quantity of products purchased is summed for each gender. This identifies which gender buys more items in total.

Interpretation

- Female customers generated more revenue, contributing about 60% of total sales. This suggests targeted campaigns towards female buyers could enhance profits further.
- The quantity-based analysis reveals which gender buys more items overall. If female customers purchased more units, it means they have higher demand frequency and shop in larger quantities. This group contributes significantly to unit sales and may respond well to quantity-based offers, bundle deals, and seasonal promotions.

Task 3 : Maximum revenue generated according to Gender

Calculate total revenue collected from each gender and average revenue per transaction to see which gender contributes more money.



Row Labels	Sum of Total_Revenue
Female	150207136
Male	101298658.2
Grand Total	**251505794.3**

For each gender, total revenue is calculated (Quantity × Price).

This shows who contributes more to total sales.

Interpretation:

The revenue analysis indicates which gender brings more monetary value to the business.

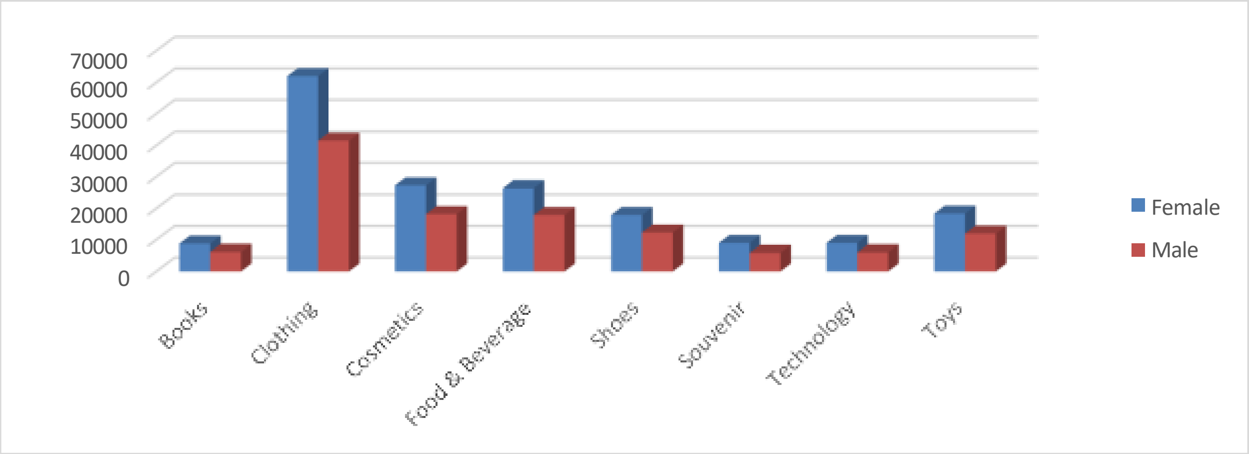
Even if both genders make similar numbers of transactions, the group with higher revenue has a stronger purchasing capacity and higher-value shopping patterns.

This gender becomes the primary revenue-driving segment, and targeted campaigns for them could further increase overall sales.

Task4: Category which are related to customer age ,gender, payment method, revenue

This task compares product categories with factors such as age, gender, payment method, and quantity. PivotTables cross-tab categories with other fields.

Category vs Gender

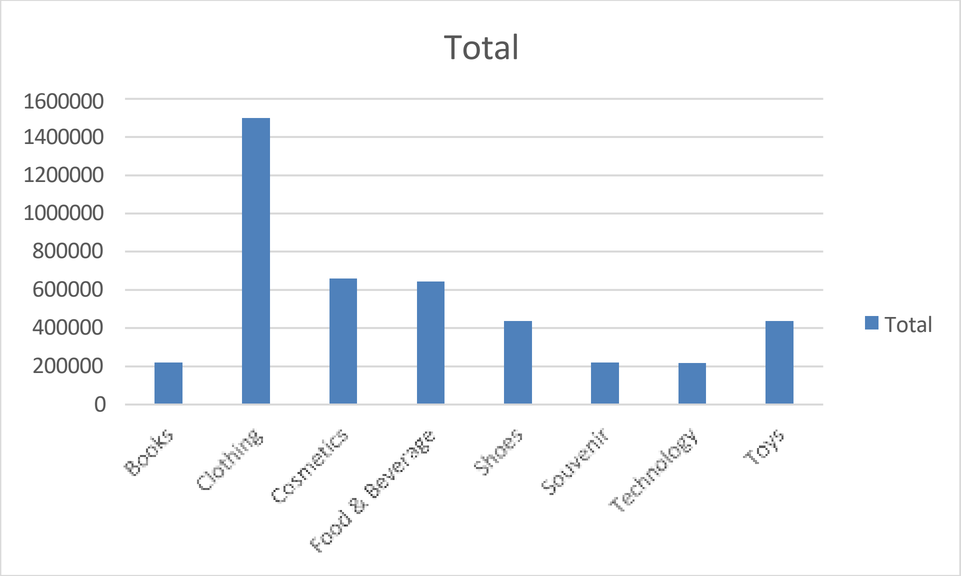


Sum of quantity	Column Labels		
Row Labels	**Female**	**Male**	**Grand Total**
Books	8776	6206	14982
Clothing	62039	41519	103558
Cosmetics	27261	18204	45465
Food & Beverage	26362	17915	44277
Shoes	17906	12311	30217
Souvenir	8976	5895	14871
Technology	8977	6044	15021
Toys	18362	11959	30321
Grand Total	**178659**	**120053**	**298712**

Interpretation

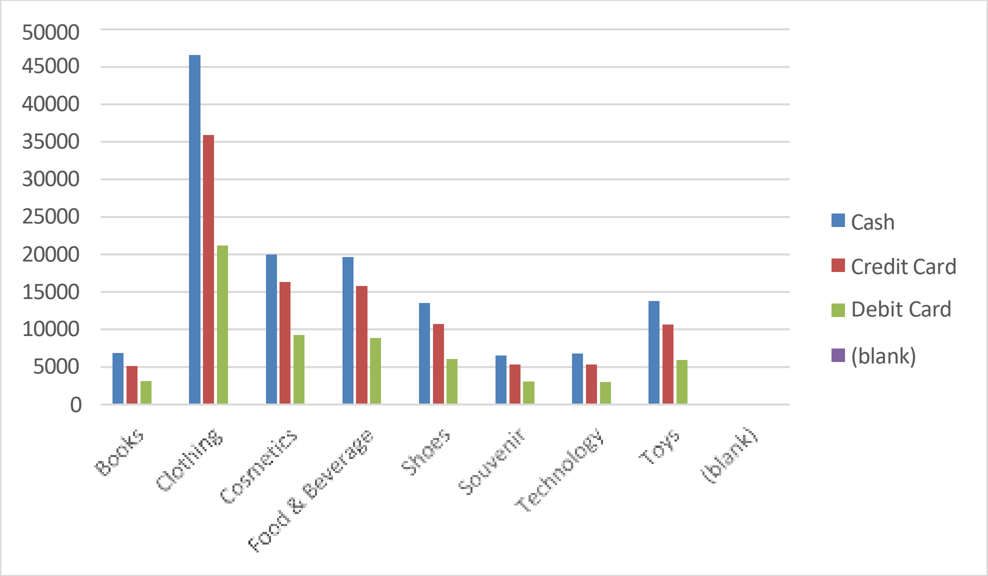
- Most of the female Buyers are focused to buy clothing
- More females from the age of 18 to 24 generates more revenue compared to the males.
- The cash payment method is frequently used by the buyers over 70%.

Category vs Age



Row Labels	Sum of age
Books	216882
Clothing	1497054
Cosmetics	657937
Food & Beverage	640605
Shoes	436027
Souvenir	216922
Technology	216669
Toys	437032
Grand Total	**4319128**

Category vs Payment Method



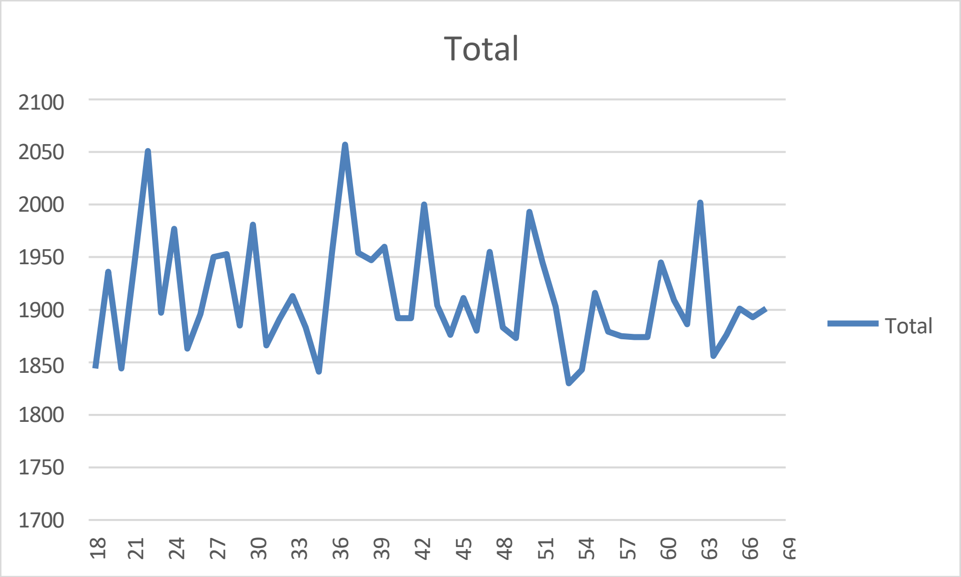
Sum of quantity	Column Labels				
Row Labels	Cash	Credit Card	Debit Card	(blank)	Grand Total
Books	6831	5062	3089		14982
Clothing	46542	35877	21139		103558
Cosmetics	19931	16283	9251		45465
Food & Beverage	19623	15792	8862		44277
Shoes	13492	10719	6006		30217
Souvenir	6486	5318	3067		14871
Technology	6720	5320	2981		15021
Toys	13745	10674	5902		30321
(blank)					
Grand Total	**133370**	**105045**	**60297**		**298712**

Interpretation:

Younger age groups may prefer categories like technology and clothing, while older groups may lean toward home essentials or books. Gender differences may also appear—for example, females often purchase clothing or cosmetics more frequently. Payment methods may vary by category, showing convenience-based behavior. These insights help companies decide what to stock more, which categories to promote, and how to personalize offers.

Task 5: Shopping Distribution According to the age

Age-wise transaction analysis by counting invoices for each age. This shows which age groups shop more frequently.



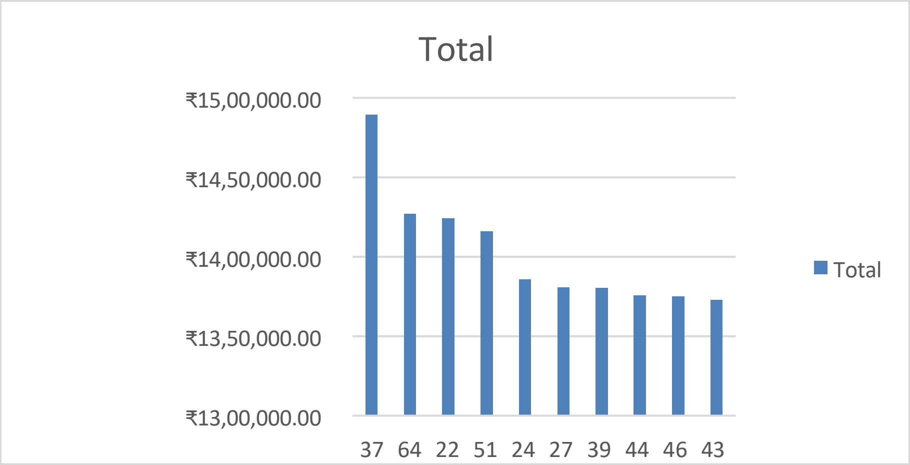
Interpretation

- Age-wise shopping distribution highlights the most active age groups.
- Typically, the 25–44 age bracket drives a large portion of total transactions because they have strong purchasing power and financial independence.
- Younger and older groups may shop less frequently. This helps businesses tailor loyalty programs and promotions based on age segments.

Task 6. Which age cat did we sell more products to?

In this study states Comparing which age category has bought more products

AGE	Sum of Total_product_sale
37	₹ 14,89,242.75
64	₹ 14,27,031.58
22	₹ 14,24,246.25
51	₹ 14,15,905.64
24	₹ 13,85,679.04
27	₹ 13,80,902.40
39	₹ 13,80,451.45
44	₹ 13,75,697.54
46	₹ 13,75,062.98
43	₹ 13,73,004.24
Grand Total	**₹ 1,40,27,223.87**



Interpretation

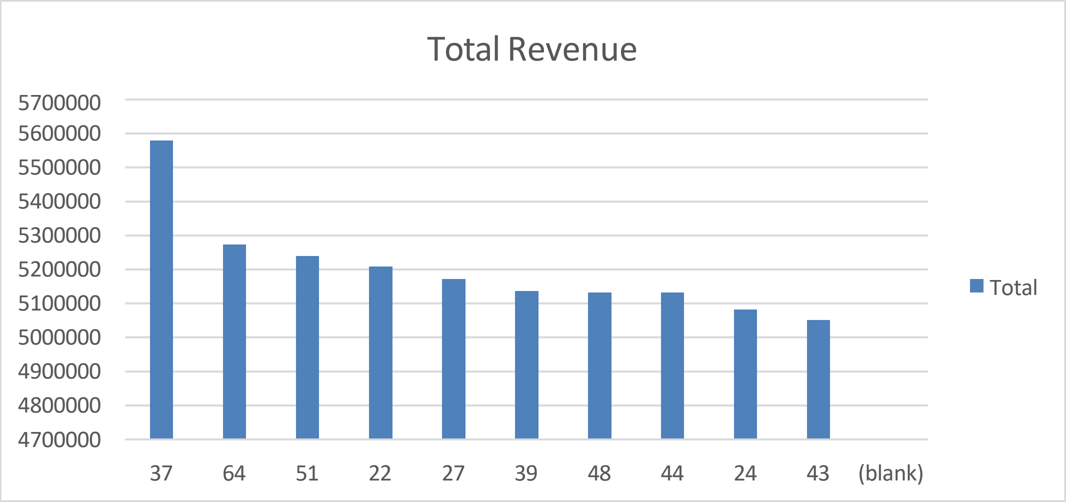
- The 35–44 age group purchased the highest number of products, leading total product sales.
- The lowest sales came from the 0–17 and 65+ age groups, showing minimal buying activity.

Task 7 : Total Revenue For According to Age

Compare which Age people contribute more revenue

Age	Total Revenue
37	5578540
64	5272475
51	5238725
22	5208841
27	5171859
39	5135673
48	5131748
44	5131687
24	5082410
43	5050324

Total Revenue For According to Age



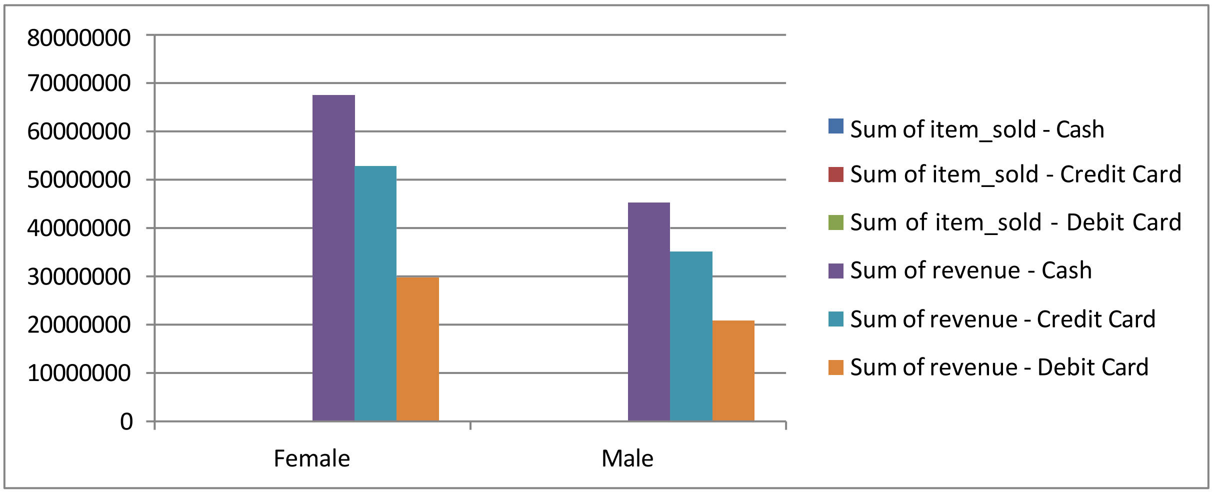
Interpretation

- People in higher age groups are earning more revenue compared to younger customers
- As age increases, revenue also shows an upward trend.

Task8: Category which are related to customer age ,gender, payment method, revenue

Comparing the Column where the category of purchasing item which influenced by the other factors like customer age, gender, payment method, and revenue.

item_sold	gender	age	payment_method	revenue
1802	Female	37	Cash	1790313.89
1748	Female	64	Cash	1562265.34
1709	Female	40	Cash	1490277.29
1536	Female	42	Cash	1471339.02
1544	Female	39	Cash	1457676.55
1552	Female	58	Cash	1443135.94



Interpretation

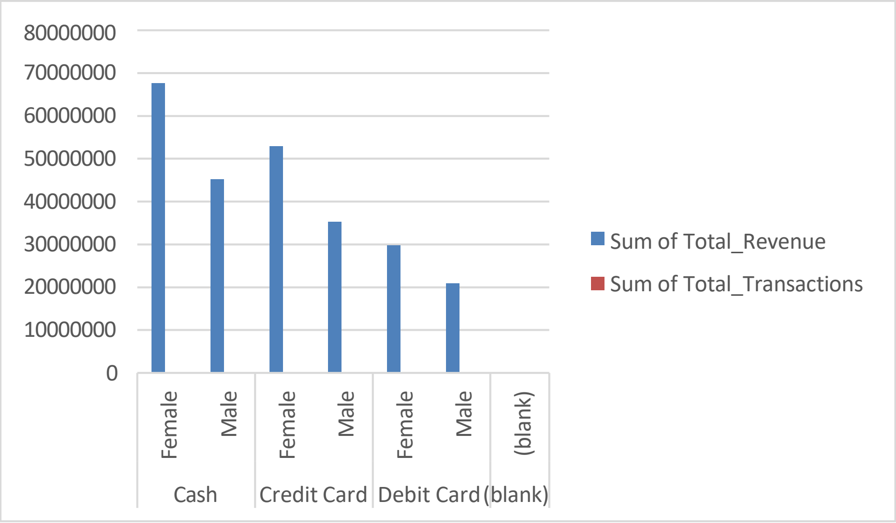
- Most of the female Buyers are focused to buy clothing .
- More females from the age of 18 to 24 generates more revenue compared to the males.
- The cash payment method is frequently used by the buyers over 70%.

Task 9 – Payment Method Analysis Report

Compare payment method for Customer GroupsBy Gender,Revenueand total transactions

Gender	Payment Method	Total Revenue	Total Transactions
Male	Debit Card	20838739.79	8117
Female	Debit Card	29757687.67	11962
Male	Credit Card	35201313.87	13920
Male	Cash	45258604.57	17938
Female	Credit Card	52875809.9	21011
Female	Cash	67573638.45	26509

Revenue by Payment Method



Interpretation

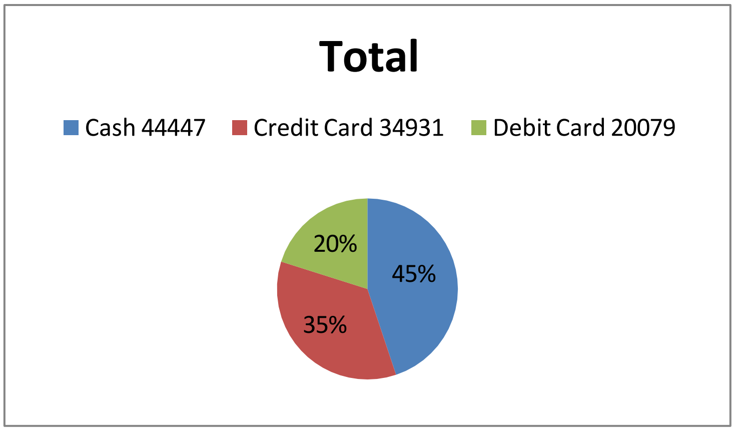
• Cash and Credit Card generate the highest revenue across customers.

• It helps identify the most preferred and profitable payment methods.

Task 10.How is the distribution of the payment method?

Here in this tasks states that where the Transaction Distribution undertaken by the each payment method.

Payment Type	Total_transcations
Cash	₹ 44,447.00
Credit Card	₹ 34,931.00
Debit Card	₹ 20,079.00
Grand Total	**₹ 99,457.00**

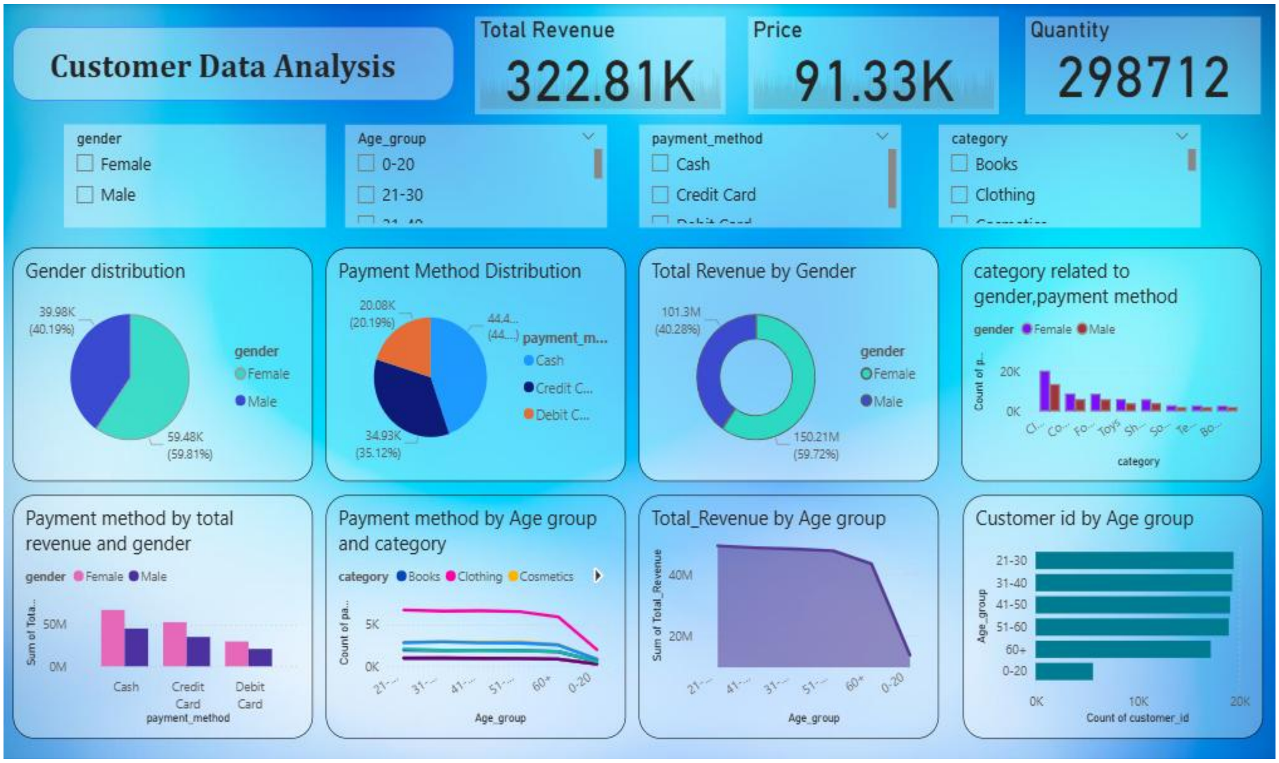


Interpretation

- Cash is the most preferred payment method, contributing 45% of total transactions, followed by Credit Cards at 35%.
- Debit Cards have the lowest share at 20%, indicating that customers use them significantly less compared to other payment options.

Visualization

Here we used the power BI tool to understand the dataset and insights of each factors by their purchasing behavior and revenue towards the company.



Summary of the report:

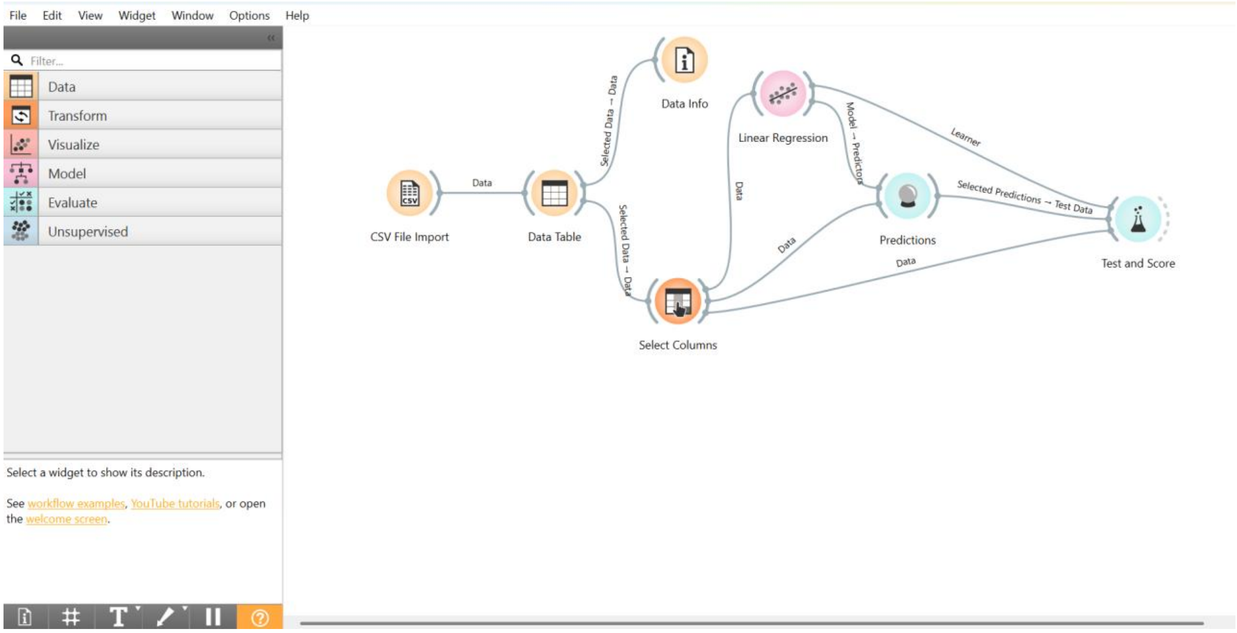
Total Revenue: ₹322.81K

Total Quantity Sold: 298712 units

1. Female customers purchased more quantity and generated higher revenue (about 60%) compared to male customers, making them the primary target audience for marketing and promotions.
2. Cash is the most preferred payment method, followed by Credit Card and Debit Card, indicating that customers mostly choose quick/instant payment options.
3. Clothing is the top-selling category, followed by Cosmetics and Food, while Books and Technology have the least sales in terms of quantity.
4. Mall of Istanbul generates the highest revenue, compared to other mall locations in the dataset.

Predictions

Here we predicted the R2 score of the databases using orange data mining software machine learning by linear regression model.



We imported our dataset into Orange in CSV format and connected it to the Data Table. After that, we created a new target column called Revenue, which is calculated using the formula:

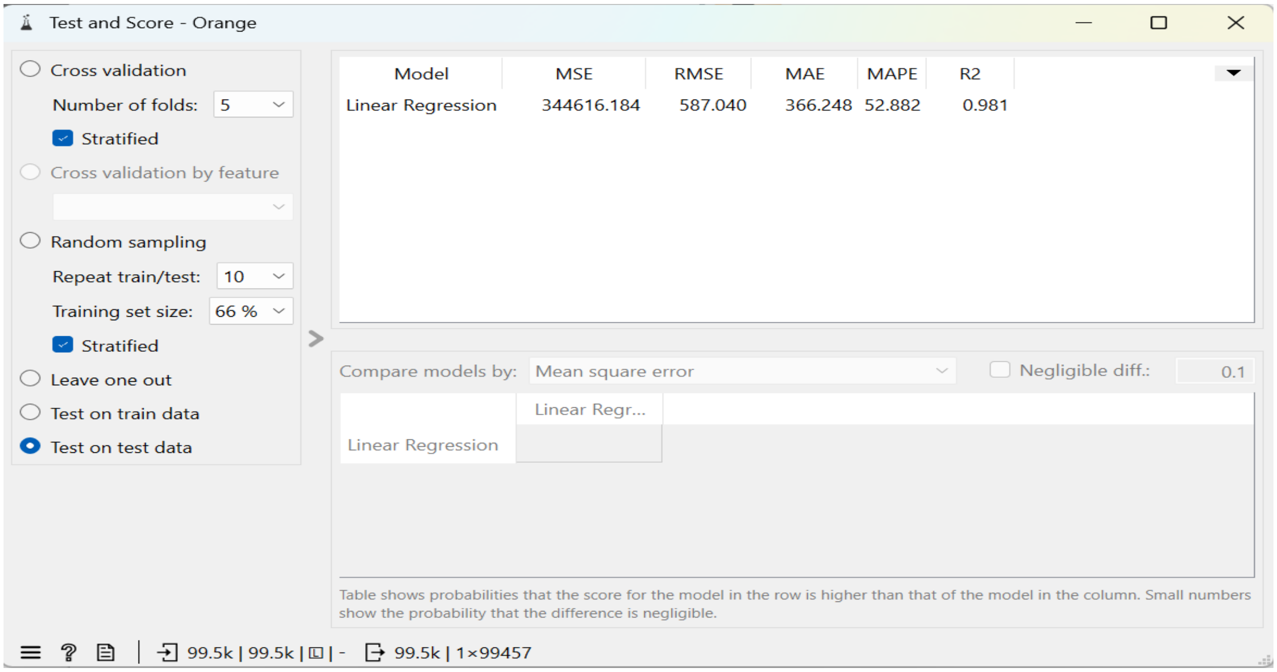
Revenue = Quantity × Price

This target value represents the income generated by the company based on products sold.

Next, we linked the data to the Linear Regression model to train the model and generate predictions for revenue. This helped us understand how the input values (price and quantity) affect the final revenue.

Finally, we used the Predictions and Test and Score widgets to evaluate the model's performance. The main goal of this testing is to check how accurately our model predicts revenue. For this, the R^2 score is used to measure how strongly revenue is related to the selected features. A higher R^2 value indicates that the model explains the variance well and fits the data strongly.

These test results help us understand how reliable the model is in estimating future revenue for the company.



Test and Score - Orange

○ Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

● Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

● Test on test data

Model	MSE	RMSE	MAE	MAPE	R2
Linear Regression	344616.184	587.040	366.248	52.882	0.981

Compare models by: Mean square error

□ Negligible diff.: 0.1

	Linear Regr...
Linear Regression	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

99.5k | 99.5k | | - 99.5k | 1×99457

R² Score: 0.981

MSE: 344616.184

The R² score (Coefficient of Determination) measures how well the model explains the changes in the target variable.

R² values range between 0 and 1:

- R² = 1 → the model explains 100% of the variation (perfect)
- R² = 0 → the model explains 0% of the variation
- Higher R² values → stronger model performance

In this project, the model achieved an R² value of 0.981, which means 98.1% of the variance in Revenue is explained by the model.
This indicates that our regression model is highly accurate, consistent, and reliable in predicting revenue based on the input variables.

The MSE score represents the average squared difference between the predicted values and the actual values. A lower MSE means better model performance.

The regression model achieved an R² score of 0.981, indicating that it explains 98.1% of the variance in the revenue, demonstrating a highly accurate and reliable predictive performance.

Tools

For this project we covered the tools are

1. **Excel** : Used to clean and organize the dataset, and to calculate the revenue column by multiplying Quantity and Price.
2. **SQL** : Used to run queries such as sorting, filtering, and extracting specific data from the dataset, and to export the cleaned data as a CSV file.
3. **Orange** : Used to build and train the Linear Regression model, generate prediction values, and evaluate the performance using metrics like R^2 score and MSE.
4. **Power BI** : Used to design an interactive dashboard and create visual reports, helping us understand patterns, trends, and insights from the dataset.

Conclusion

Key Insights

- This analysis of customer data from 10 major shopping malls in Istanbul between 2021 and 2023 provided valuable insights into purchasing trends and revenue performance.
- Female customers contributed the highest share, accounting for nearly 60% of total purchases and revenue, making them the primary focus group for future marketing strategies.
- Cash and credit card payments were the most preferred modes, indicating that customers favor instant or direct payment options. Clothing was the highest-selling category by volume, followed by cosmetics and food, while books and technology showed the lowest sales.
- Customers aged 35–44 and above generated the maximum revenue compared to younger age groups. Among all locations, the Mall of Istanbul emerged as the top-performing mall based on revenue.
- The predictive model built using Linear Regression achieved an R^2 score of 0.981, demonstrating that the model can accurately estimate revenue using quantity and price values.

Overall, this project shows how combining data analysis, visualization, and predictive modeling through tools like Excel, SQL, Power BI, and Orange can help businesses take informed decisions, understand customer behavior, and plan better strategies to improve sales performance