1. Overview

This project analyzes the *"Supermarket Sale"* dataset, which contains sales transaction records from three supermarket branches located in **New York, Los Angeles, and Chicago (USA)**. It includes details about products, customers, and sales information, enabling insights into **customer behavior** and **branch performance**.

- **Dataset Size:** 253 rows × 8 columns
- **Final Size (after cleaning):** 239 rows × 8 columns
- **Missing Values:** 12 missing values (1 out of 11 row has 2 missing values) removed due to incomplete data

Columns:

- sale_id: Unique sale transaction ID
- branch: Supermarket branch (A, B)
- city: City location (New York, Los Angeles, Chicago)
- customer_type: Member or Normal customer
- product_name: Name of product
- product _category: Product category (e.g., Fruits, Stationery, Beverages)
- quantity: Number of units sold
- total_price: Total sales value (USD)

Data Sources

Although the original source is unspecified, the dataset likely represents **sample supermarket transaction data** used for educational purposes. It reflects common structures found in real retail sales systems and public learning resources. Comparable data is often adapted from the following open platforms:

- **Kaggle**: Retail and supermarket sales datasets for data analytics practice.
- **UCI Machine Learning Repository**: Business and marketing datasets for academic study.

- **Open educational resources:** Sample data provided for teaching statistics and business analytics.

2. **Data cleaning**

The dataset was reviewed and cleaned to ensure accuracy and consistency before analysis. The following steps were taken during the data cleaning process:

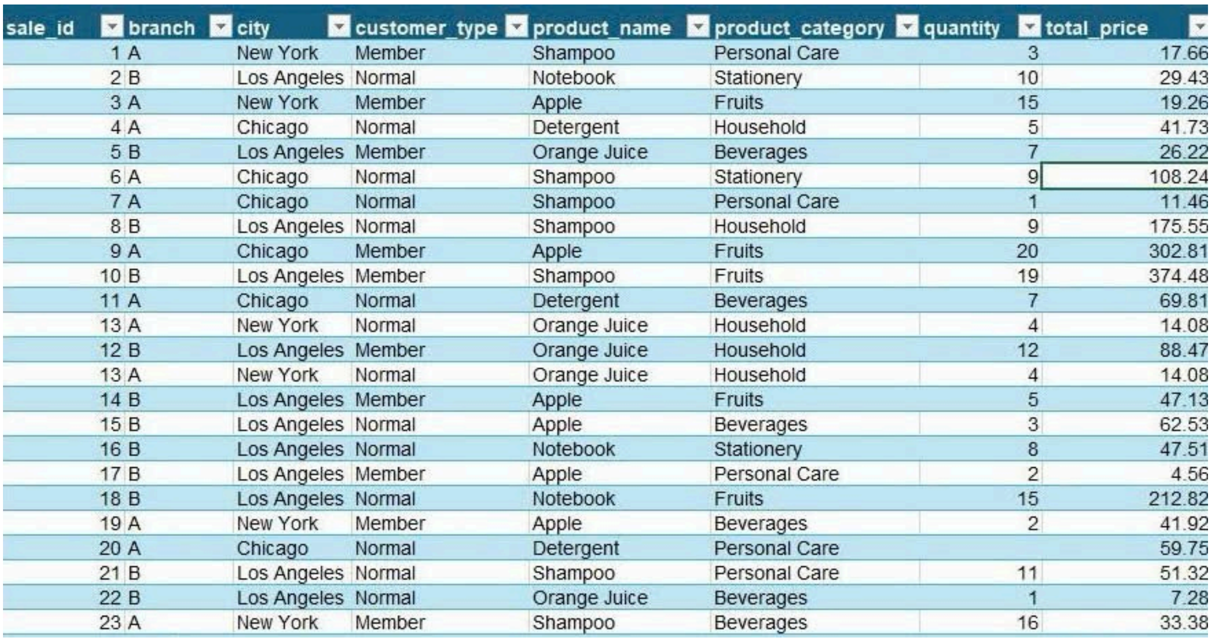
2.1. Data Formatting

All columns were formatted with appropriate data types to ensure data consistency and avoid calculation errors. Specifically, *Sale_ID*, *Branch*, *City*, *Customer_Type*, *Product_Name*, and *Product_Category* were set to **Text format**, as they represent categorical data. The *Quantity* column was formatted as **Number (0 decimal places)**, and *Total_Price* was formatted as **Currency (2 decimal places)**. Although *Sale_ID* contains numeric characters, it was formatted as Text because it acts as a unique identifier rather than a numerical value.

2.2. Checked for missing values

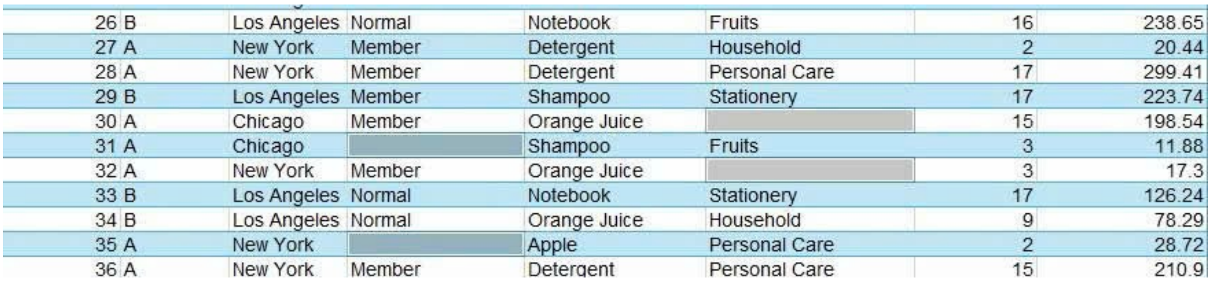
Step 1: Select all data → go to **Home → Find & Select → Go To Special → Blanks**.

Excel highlights all blank cells.



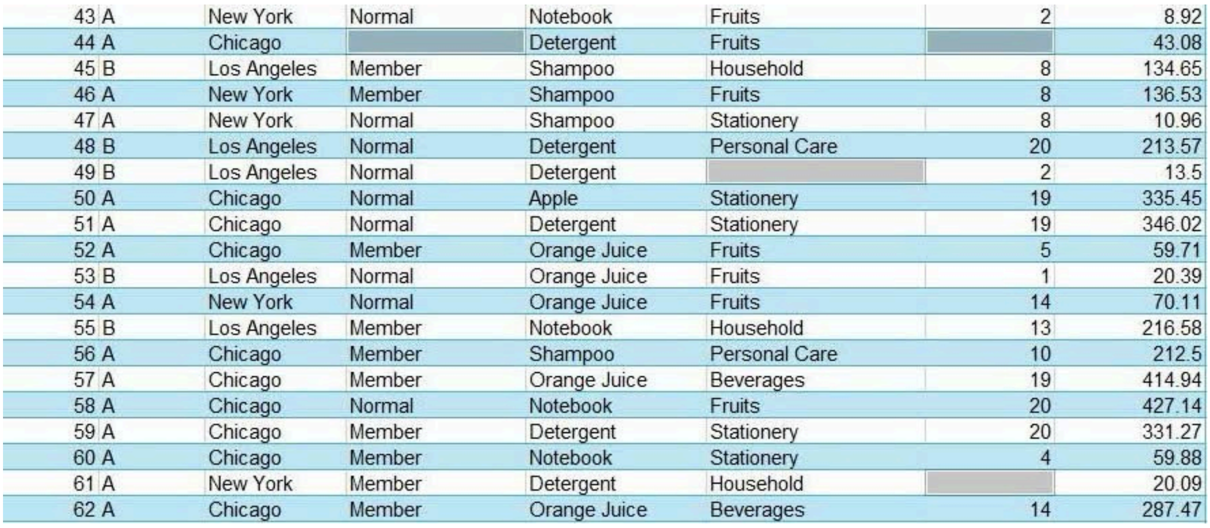
sale_id	branch	city	customer_type	product_name	product_category	quantity	total_price
1	A	New York	Member	Shampoo	Personal Care	3	17.66
2	B	Los Angeles	Normal	Notebook	Stationery	10	29.43
3	A	New York	Member	Apple	Fruits	15	19.26
4	A	Chicago	Normal	Detergent	Household	5	41.73
5	B	Los Angeles	Member	Orange Juice	Beverages	7	26.22
6	A	Chicago	Normal	Shampoo	Stationery	9	108.24
7	A	Chicago	Normal	Shampoo	Personal Care	1	11.46
8	B	Los Angeles	Normal	Shampoo	Household	9	175.55
9	A	Chicago	Member	Apple	Fruits	20	302.81
10	B	Los Angeles	Member	Shampoo	Fruits	19	374.48
11	A	Chicago	Normal	Detergent	Beverages	7	69.81
13	A	New York	Normal	Orange Juice	Household	4	14.08
12	B	Los Angeles	Member	Orange Juice	Household	12	88.47
13	A	New York	Normal	Orange Juice	Household	4	14.08
14	B	Los Angeles	Member	Apple	Fruits	5	47.13
15	B	Los Angeles	Normal	Apple	Beverages	3	62.53
16	B	Los Angeles	Normal	Notebook	Stationery	8	47.51
17	B	Los Angeles	Member	Apple	Personal Care	2	4.56
18	B	Los Angeles	Normal	Notebook	Fruits	15	212.82
19	A	New York	Member	Apple	Beverages	2	41.92
20	A	Chicago	Normal	Detergent	Personal Care		59.75
21	B	Los Angeles	Normal	Shampoo	Personal Care	11	51.32
22	B	Los Angeles	Normal	Orange Juice	Beverages	1	7.28
23	A	New York	Member	Shampoo	Beverages	16	33.38

Figure 1: Missing values (row 20 - column: quantity)



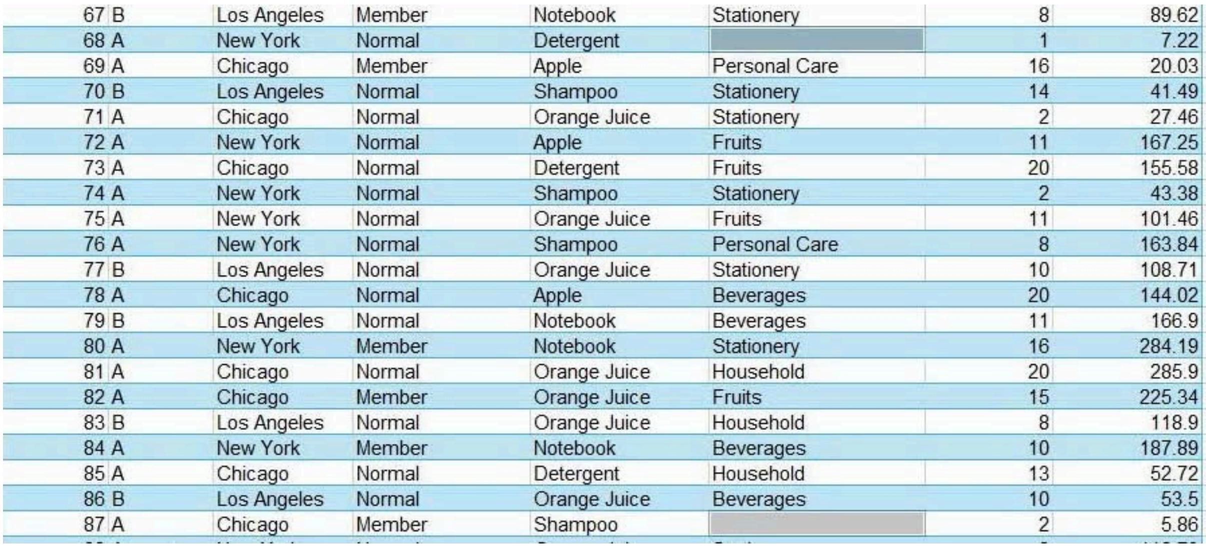
26	B	Los Angeles	Normal	Notebook	Fruits	16	238.65
27	A	New York	Member	Detergent	Household	2	20.44
28	A	New York	Member	Detergent	Personal Care	17	299.41
29	B	Los Angeles	Member	Shampoo	Stationery	17	223.74
30	A	Chicago	Member	Orange Juice		15	198.54
31	A	Chicago		Shampoo	Fruits	3	11.88
32	A	New York	Member	Orange Juice		3	17.3
33	B	Los Angeles	Normal	Notebook	Stationery	17	126.24
34	B	Los Angeles	Normal	Orange Juice	Household	9	78.29
35	A	New York		Apple	Personal Care	2	28.72
36	A	New York	Member	Detergent	Personal Care	15	210.9

Figure 2: Missing values (row 30, 31, 32, 35 - column customer_type & product_category)



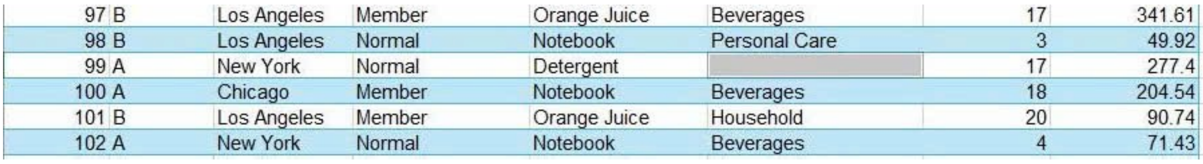
43	A	New York	Normal	Notebook	Fruits	2	8.92
44	A	Chicago		Detergent	Fruits		43.08
45	B	Los Angeles	Member	Shampoo	Household	8	134.65
46	A	New York	Member	Shampoo	Fruits	8	136.53
47	A	New York	Normal	Shampoo	Stationery	8	10.96
48	B	Los Angeles	Normal	Detergent	Personal Care	20	213.57
49	B	Los Angeles	Normal	Detergent		2	13.5
50	A	Chicago	Normal	Apple	Stationery	19	335.45
51	A	Chicago	Normal	Detergent	Stationery	19	346.02
52	A	Chicago	Member	Orange Juice	Fruits	5	59.71
53	B	Los Angeles	Normal	Orange Juice	Fruits	1	20.39
54	A	New York	Normal	Orange Juice	Fruits	14	70.11
55	B	Los Angeles	Member	Notebook	Household	13	216.58
56	A	Chicago	Member	Shampoo	Personal Care	10	212.5
57	A	Chicago	Member	Orange Juice	Beverages	19	414.94
58	A	Chicago	Normal	Notebook	Fruits	20	427.14
59	A	Chicago	Member	Detergent	Stationery	20	331.27
60	A	Chicago	Member	Notebook	Stationery	4	59.88
61	A	New York	Member	Detergent	Household		20.09
62	A	Chicago	Member	Orange Juice	Beverages	14	287.47

Figure 3: Missing values (row 44, 49, 61 column customer_type, product_category & qiantity)



67	B	Los Angeles	Member	Notebook	Stationery	8	89.62
68	A	New York	Normal	Detergent		1	7.22
69	A	Chicago	Member	Apple	Personal Care	16	20.03
70	B	Los Angeles	Normal	Shampoo	Stationery	14	41.49
71	A	Chicago	Normal	Orange Juice	Stationery	2	27.46
72	A	New York	Normal	Apple	Fruits	11	167.25
73	A	Chicago	Normal	Detergent	Fruits	20	155.58
74	A	New York	Normal	Shampoo	Stationery	2	43.38
75	A	New York	Normal	Orange Juice	Fruits	11	101.46
76	A	New York	Normal	Shampoo	Personal Care	8	163.84
77	B	Los Angeles	Normal	Orange Juice	Stationery	10	108.71
78	A	Chicago	Normal	Apple	Beverages	20	144.02
79	B	Los Angeles	Normal	Notebook	Beverages	11	166.9
80	A	New York	Member	Notebook	Stationery	16	284.19
81	A	Chicago	Normal	Orange Juice	Household	20	285.9
82	A	Chicago	Member	Orange Juice	Fruits	15	225.34
83	B	Los Angeles	Normal	Orange Juice	Household	8	118.9
84	A	New York	Member	Notebook	Beverages	10	187.89
85	A	Chicago	Normal	Detergent	Household	13	52.72
86	B	Los Angeles	Normal	Orange Juice	Beverages	10	53.5
87	A	Chicago	Member	Shampoo		2	5.86

Figure 4: Missing values (row 68, 87 - column product_category)



97	B	Los Angeles	Member	Orange Juice	Beverages	17	341.61
98	B	Los Angeles	Normal	Notebook	Personal Care	3	49.92
99	A	New York	Normal	Detergent		17	277.4
100	A	Chicago	Member	Notebook	Beverages	18	204.54
101	B	Los Angeles	Member	Orange Juice	Household	20	90.74
102	A	New York	Normal	Notebook	Beverages	4	71.43

Figure 5: Missing value (row 99 - product_category)

A total of **12 missing values** were identified in the dataset across three columns. Missing entries were found in:

- **quantity** column: rows **20, 44, 61**
- **customer_type** column: rows **31, 35, 44**
- **product_category** column: rows **30, 32, 49, 68, 87, 99**

These missing values represented incomplete sales information, such as unknown quantities, customer types, or product categories. Such gaps could lead to inaccurate descriptive statistics and biased insights.

Step 2: Delete rows that contain missing data (Right-click → Delete → Entire Row or Ctrl -).

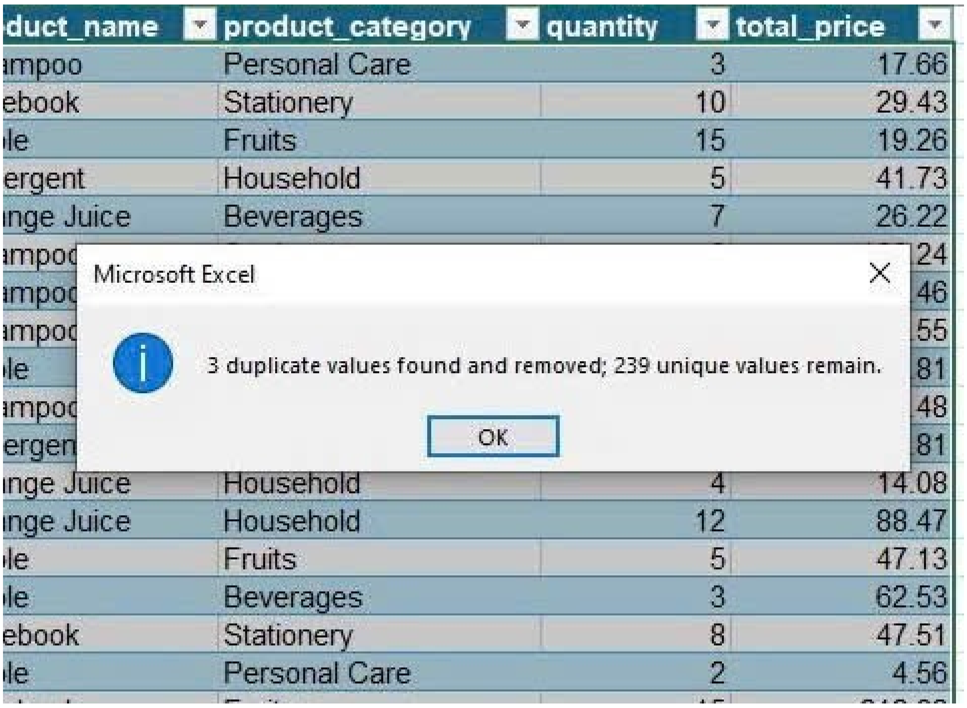
A total of **12 missing values** were detected in key columns such as quantity, customer_type, and product_category. Because these fields contain essential information for calculating sales and understanding customer behavior, filling them with estimated values could distort the analysis. Therefore, the missing records were **removed** to maintain data accuracy and reliability for further descriptive analysis.

2.3. Checked for duplicate records

Step 1: Select the whole table (ctrl A)

Step 2: Go to **Data → Remove Duplicates**.

Step 3: Tick all columns → OK.



duct_name	product_category	quantity	total_price
mpoo	Personal Care	3	17.66
ebook	Stationery	10	29.43
le	Fruits	15	19.26
ergent	Household	5	41.73
nge Juice	Beverages	7	26.22
mpoc			24
mpoc			46
mpoc			55
le			81
mpoc			48
ergen			81
nge Juice	Household	4	14.08
nge Juice	Household	12	88.47
le	Fruits	5	47.13
le	Beverages	3	62.53
ebook	Stationery	8	47.51
le	Personal Care	2	4.56

Microsoft Excel

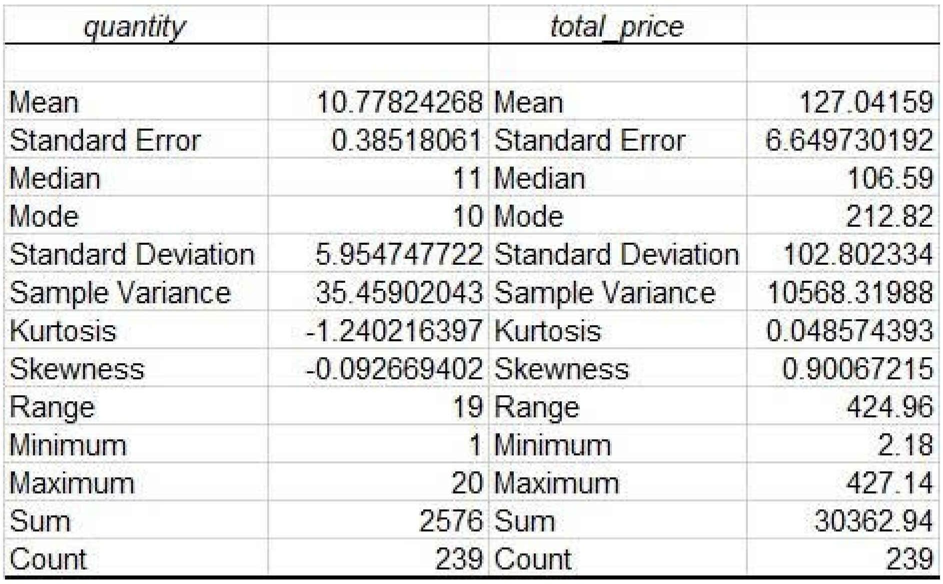
3 duplicate values found and removed; 239 unique values remain.

Figure 5: Duplicate values

After handling missing values, a duplicate check was performed using Excel's *Remove Duplicates* function. Three identical rows were found and deleted, resulting in **239 unique records**. The cleaned dataset now contains **8 columns** with no missing or duplicate values, ensuring consistency and reliability for descriptive analysis and visualization.

After cleaning, the dataset contains 239 valid records, with no missing or duplicate values, ready for accurate analysis.

3. Descriptive Statistics



quantity		*total_price*	
Mean	10.77824268	Mean	127.04159
Standard Error	0.38518061	Standard Error	6.649730192
Median	11	Median	106.59
Mode	10	Mode	212.82
Standard Deviation	5.954747722	Standard Deviation	102.802334
Sample Variance	35.45902043	Sample Variance	10568.31988
Kurtosis	-1.240216397	Kurtosis	0.048574393
Skewness	-0.092669402	Skewness	0.90067215
Range	19	Range	424.96
Minimum	1	Minimum	2.18
Maximum	20	Maximum	427.14
Sum	2576	Sum	30362.94
Count	239	Count	239

Figure 6: Summary statistics

Quantity

Customers usually buy around 10 items per transaction, showing a stable and moderate shopping pattern. This consistency helps the supermarket predict product demand and manage inventory more effectively.

Total Price

The average spending per transaction is about USD 127, though some customers spend much higher amounts. This indicates the presence of *high-value customers* who contribute significantly to total revenue,a key group for targeted promotions or loyalty programs.

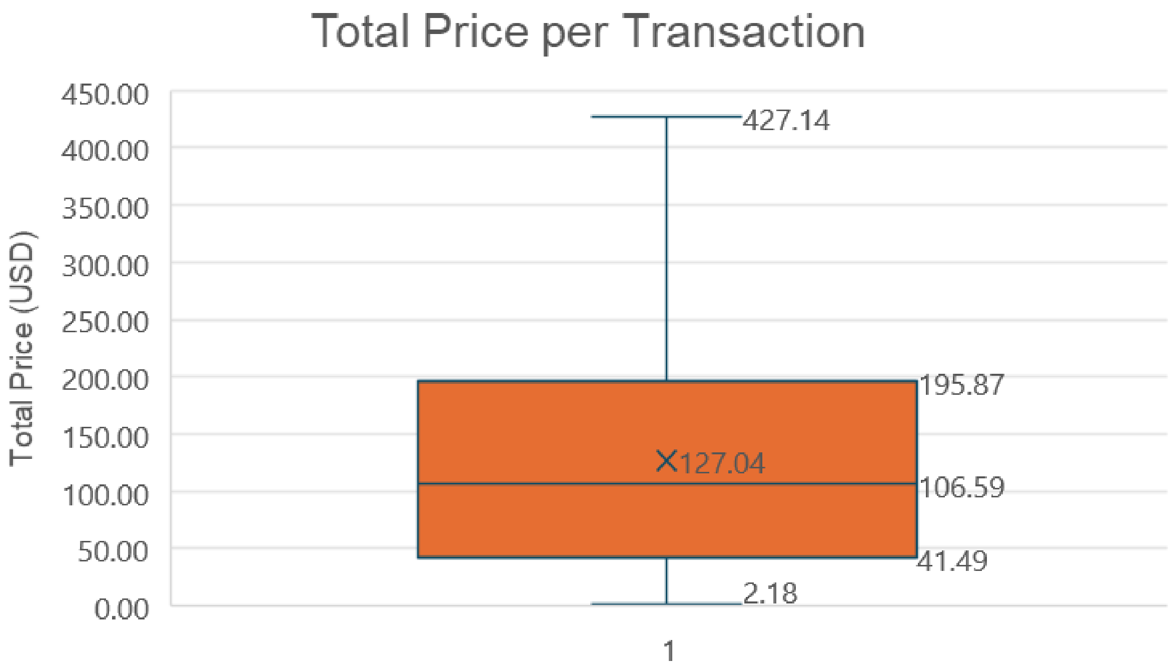


Figure 7: Box Plot of Total Price per Transaction:

The box plot illustrates the distribution of total spending per transaction. The median value is approximately **USD 106.59**, while the mean is slightly higher at **USD 127.04**, suggesting a **right-skewed distribution**. Most transactions fall within the range of **USD 40 to 200**, as indicated by the interquartile range (IQR), while the minimum and maximum values are **USD 2.18** and **USD 427.14**, respectively. This implies that most customers make moderate purchases, but a few high-value transactions significantly raise the overall average.

Overall, this distribution supports the findings from descriptive statistics and highlights variability in customer spending behavior, an important aspect to consider before moving on to deeper business insights.

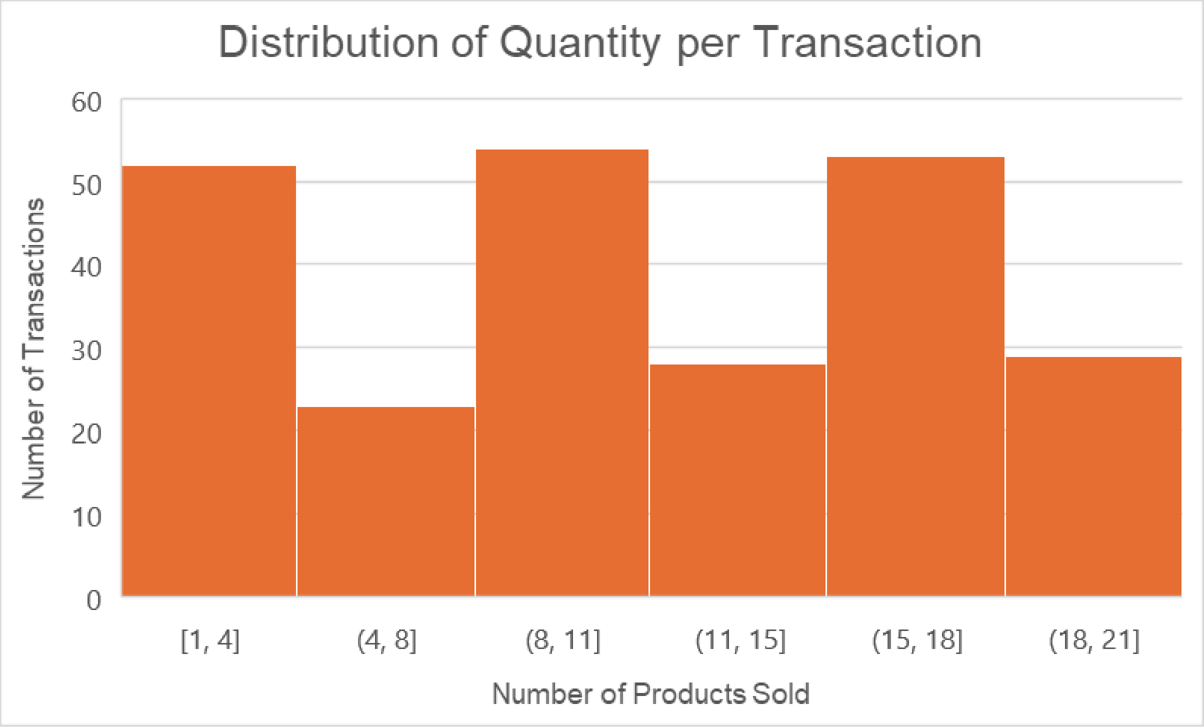


Figure 8: Distribution of quantity per transaction

The histogram above shows how many products customers usually buy in one transaction. Most customers purchase **between 8 and 11 items**, which is the largest group with about **55 transactions**. Buying **1–4 items** and **15–18 items** is also quite common (around **50–55 transactions**), while only a few customers buy **more than 18 items** (about **30 transactions**). This means most customers make **medium-sized purchases** instead of very small or very large ones. Overall, the chart shows that the number of products per purchase is **fairly balanced**, so customer buying behavior is **quite consistent**.

After understanding the overall distribution, the next step focuses on identifying key business insights

3.1. Insight 1: Total Sales Revenue by Branch

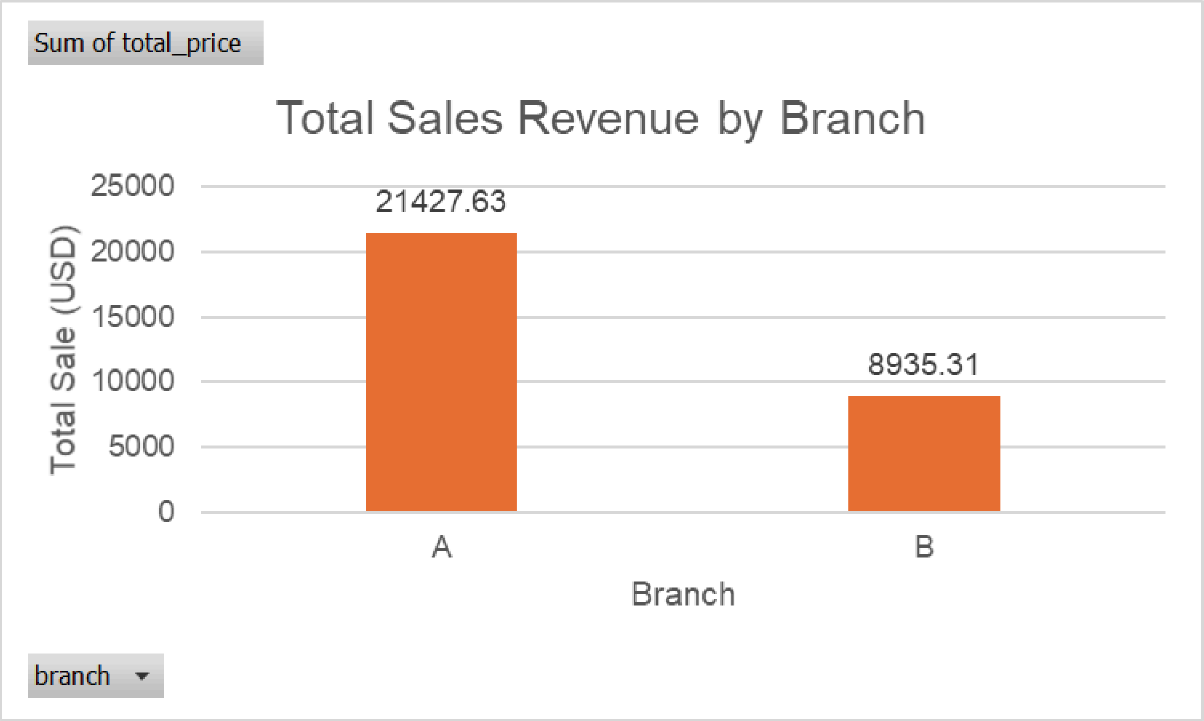


Figure 9: Total Sales Revenue by Branch

Branch A generated nearly double the revenue of Branch B (USD 21,427 vs. USD 8,935). This suggests stronger sales performance, possibly due to better location, a larger customer base, or more effective marketing. The company should consider expanding operations or investing more in Branch A's area to maximize overall business growth.

3.2. Insight 2: Average Spending by Customer Type

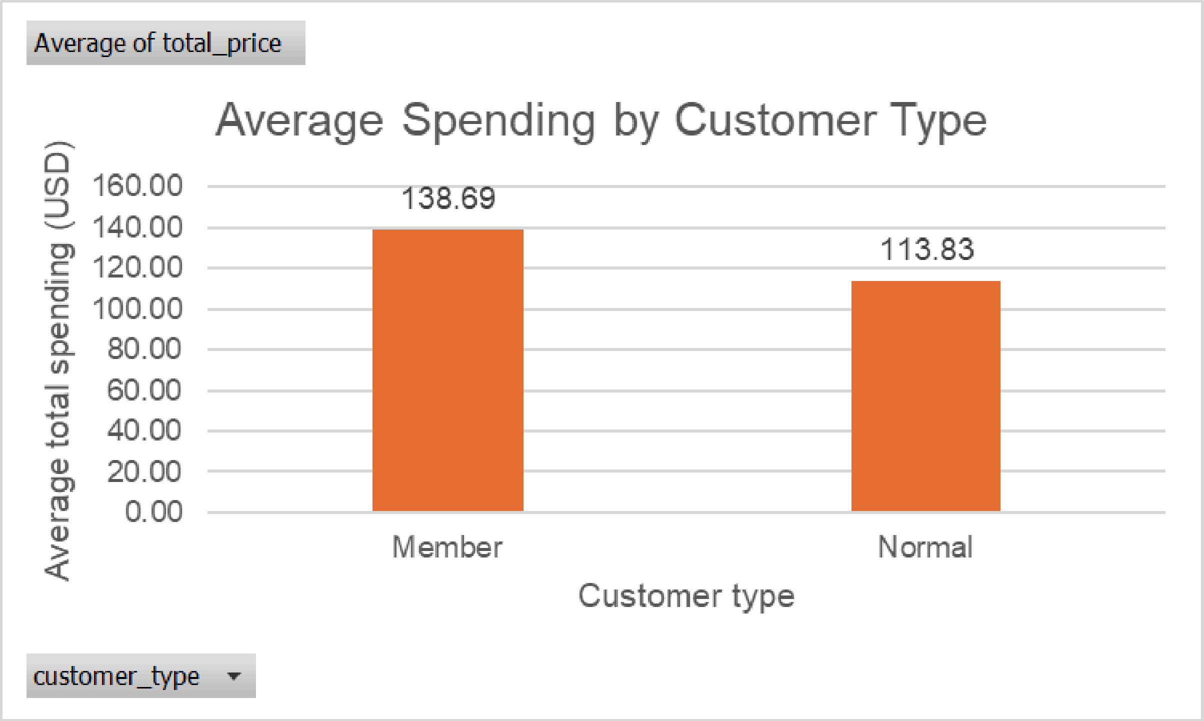


Figure 10: Average Spending by Customer Type

Member customers spend about 22% more than normal customers on average (USD 138.69 vs. USD 113.83). This shows that the membership program successfully encourages higher spending and customer loyalty. Maintaining and expanding this program could further increase total revenue and long-term customer retention.