

1) Describe the Data

The dataset contains 1,000 simulated sales transactions from 2023 (with a small spill into 2024). Key fields include product identifiers, dates, sales representatives, regions, monetary amounts, quantities, product categories, unit cost/price, customer type, discounts, payment methods, sales channels, a combined region–rep key, and derived fields for profit and discounted price. Variable types are shown in the Table below :

Product_ID	num	1052 1093 1015 1072 1061 ...
Sale_Date	Dat	, format: "2023-02-03" "2023-04-21" "2023-09-21"
Sales_Rep	chr	"Bob" "Bob" "David" "Bob" ...
Region	chr	"North" "West" "South" "South" ...
Sales_Amount	num	5054 4384 4631 2168 3750 ...
Quantity_Sold	num	18 17 30 39 13 32 29 46 30 18 ...
Product_Category	chr	"Furniture" "Furniture" "Food" "Clothing" ...
Unit_Cost	num	153 3816 262 4330 637 ...
Unit_Price	num	267 4209 371 4468 693 ...
Customer_Type	chr	"Returning" "Returning" "Returning" "New" ...
Discount	num	0.09 0.11 0.2 0.02 0.08 0.21 0.14 0.12 0.05 0.13
Payment_Method	chr	"Cash" "Cash" "Bank Transfer" "Credit Card" ...
Sales_Channel	chr	"Online" "Retail" "Retail" "Retail" ...
Region_and_Sales_Rep	chr	"North-Bob" "West-Bob" "South-David" "South-Bob"
Profit	num	114.5 393 109.8 137.7 55.3 ...
discountPrice	num	243 3746 297 4378 637 ...

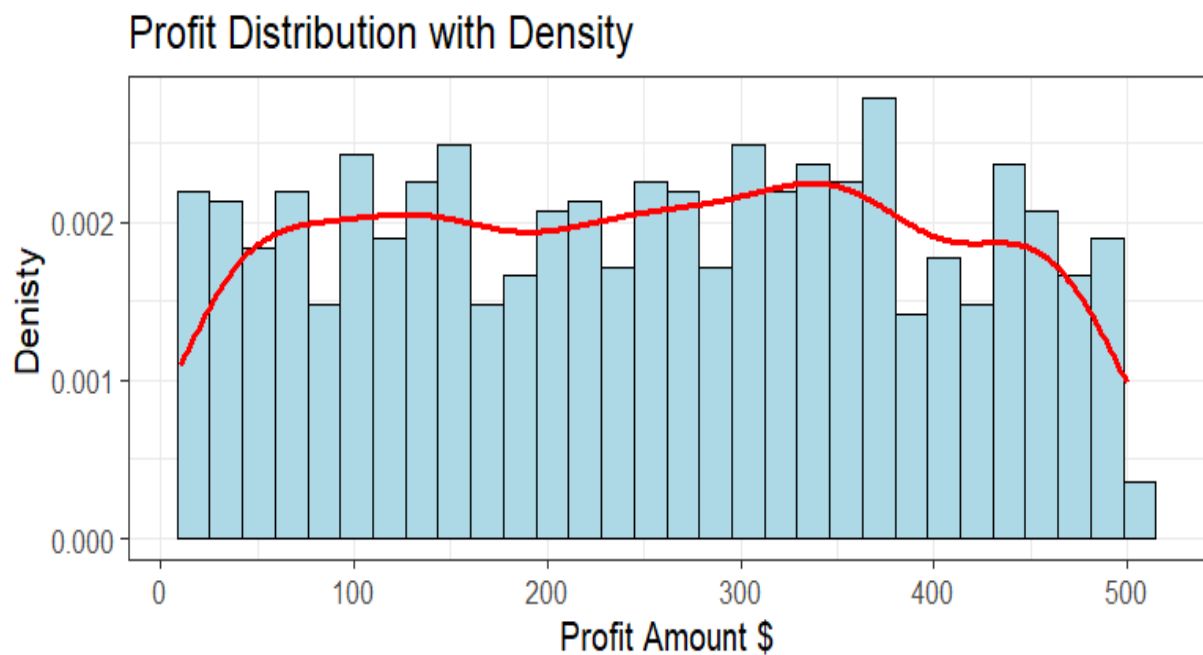
2) Summary Statistics (Key Numeric Variables)

Sales_Amount	Values	Quantity_Sold	Values	Unit_Cost	Values
Min.	100.1	Min.	1.00	Min	60.28
1st Qu	2550.3	1st Qu.	13.00	1st Qu	1238.38
Median	5019.3	Median	25.00	Median	2467.24
Mean	5019.3	Mean	25.36	Mean	2475.30
3rd Qu	7507.4	3rd Qu.	38.00	3rd Qu	3702.86
Max.	9989.0	Max.	49.00	Max	4995.30

Unit_Price	Values	Discount	Values	Profit	Values	discountPrice	Values
Min	167.1	Min	0.0000	Min	10.31	Min	127.1
1st Qu	1509.1	1st Qu	0.0800	1st Qu	133.65	1st Qu	1275.2
Median	2696.4	Median	0.1500	Median	256.52	Median	2303.3
Mean	2728.4	Mean	0.1524	Mean	253.14	Mean	2314.8
3rd Qu	3958.0	3rd Qu	0.2300	3rd Qu	367.61	3rd Qu	3313.8
Max	5442.1	Max	0.3000	Max	499.93	Max	5402.3

3) Profit Distribution with Density

What we see: The histogram + density curve for **Profit** shows a unimodal shape centered ~250 with limited skew and light tails (consistent with the quartile summary). Business-wise, this indicates stable unit profitability across transactions.

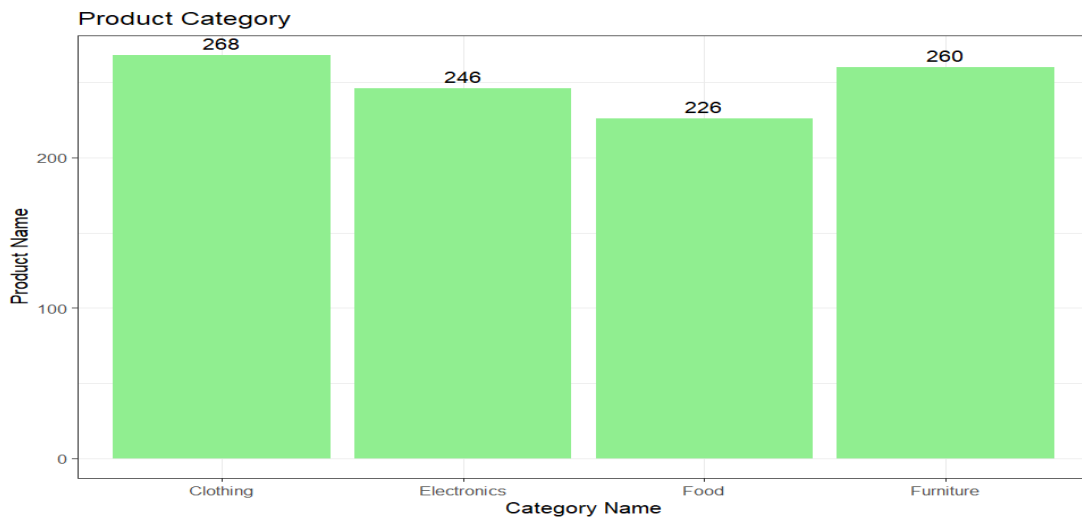


4) Categorical Profiles (Figures)

Product Category

Categories present: **Furniture, Food, Clothing, Electronics.**

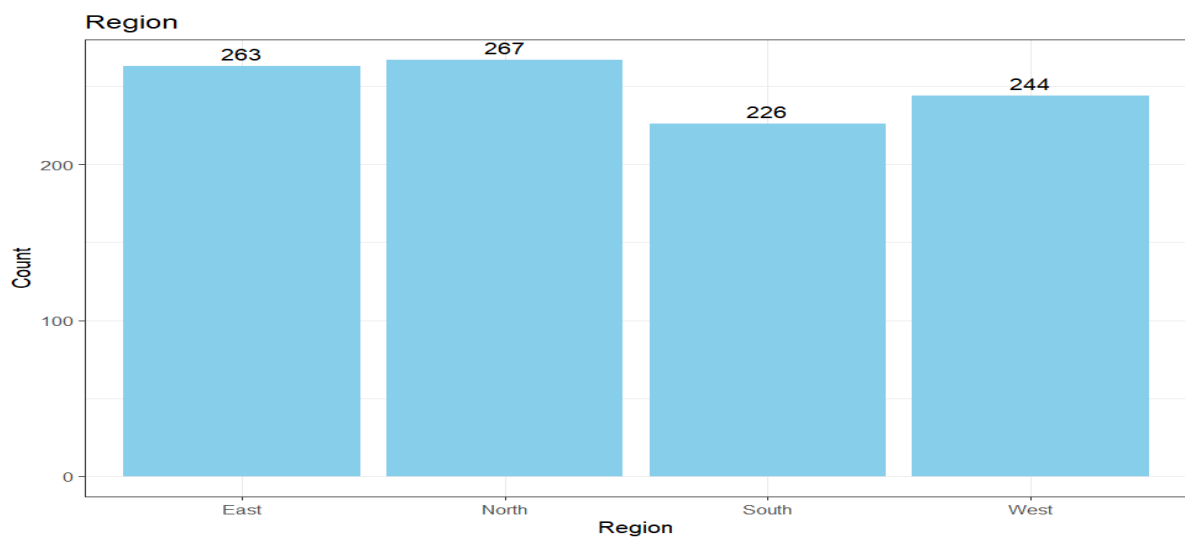
Reading: These are the four major lines examined in the subsequent profitability section.



Region

Regions present: **North, West, South, East.**

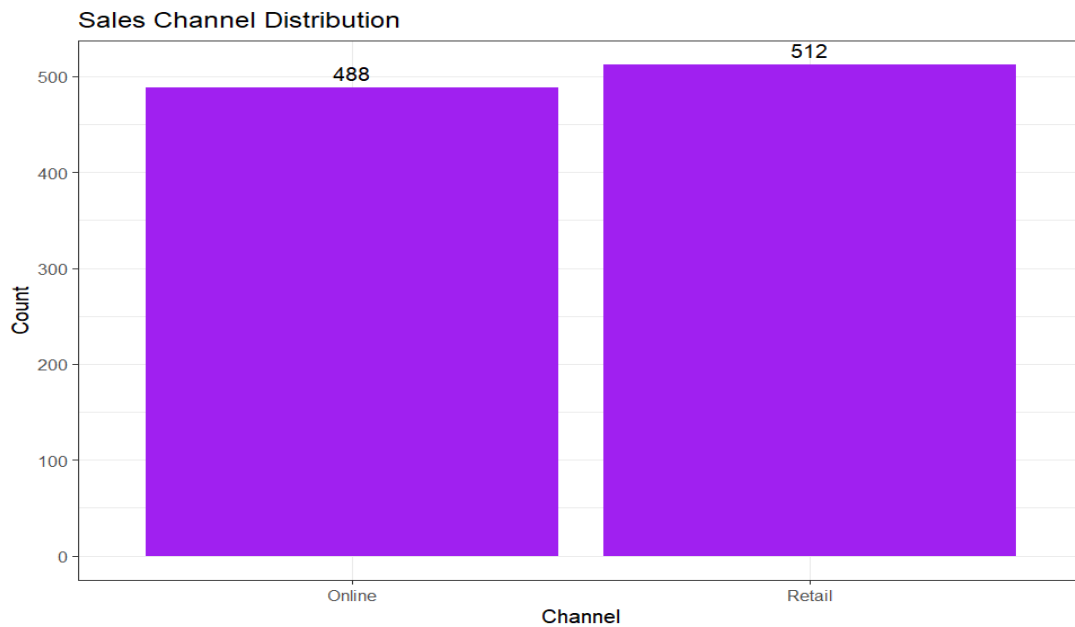
Reading: Coverage across four regions allows comparing geographic performance.



Sales Channel

Channels present: **Online, Retail.**

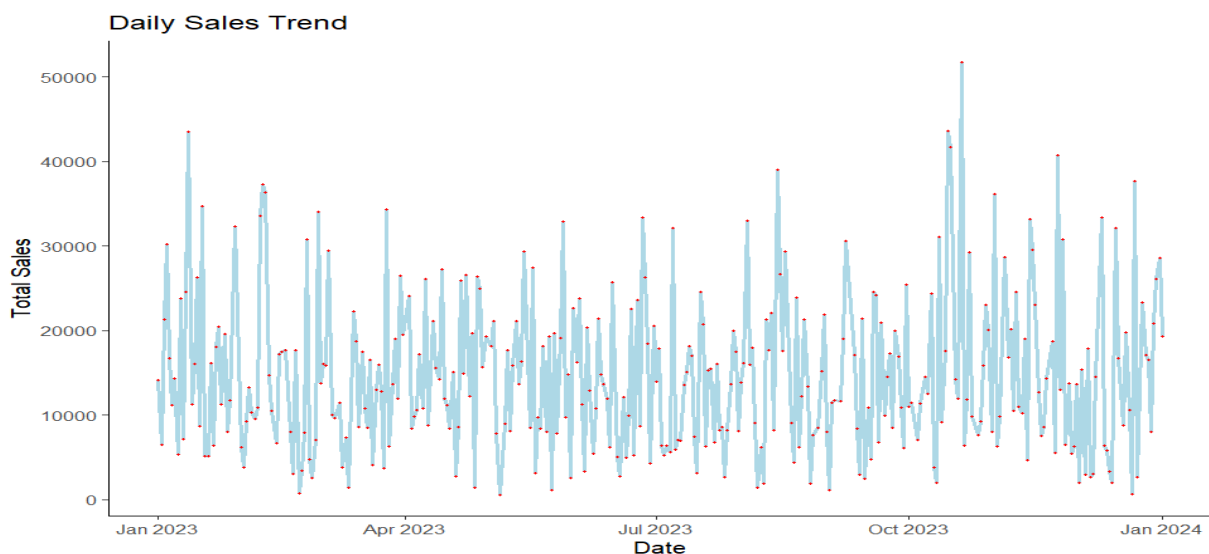
Reading: Enables channel-level comparisons for behavior, pricing, and promotions



5) Aggregate Sales Over Time

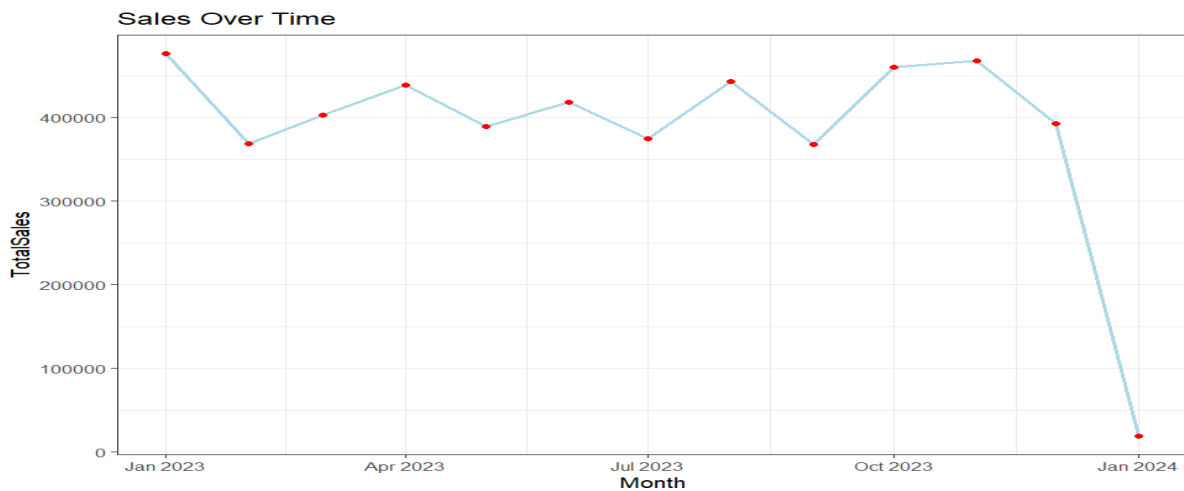
Daily Sales Trend (Figure)

Reading: Day-to-day totals fluctuate within a stable band; no extreme spikes after outlier removal checks.



Monthly Sales (Figure)

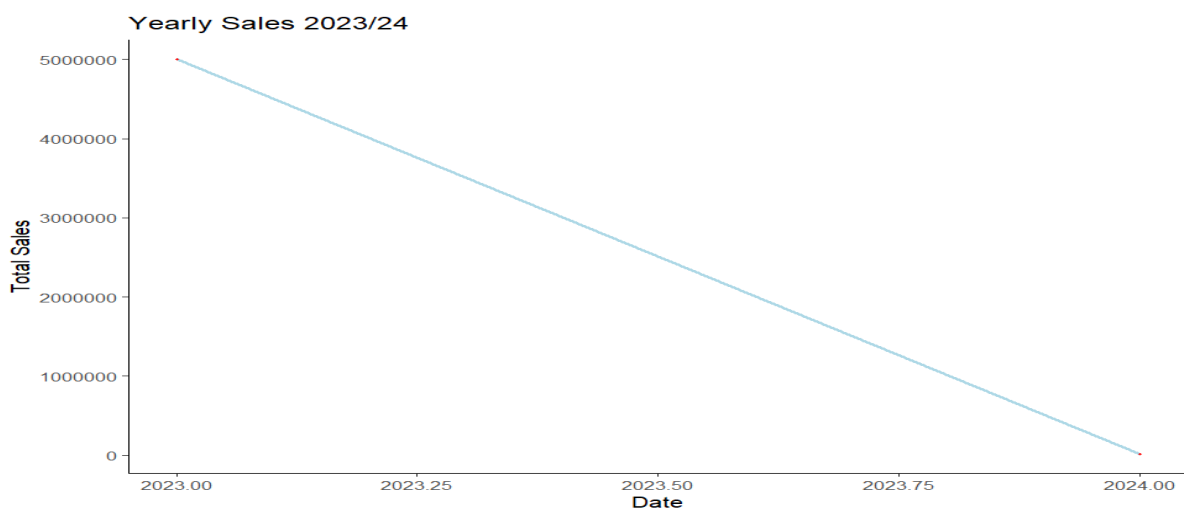
Reading: Monthly aggregation smooths noise and shows steady throughput consistent with the synthetic design.



Yearly Sales

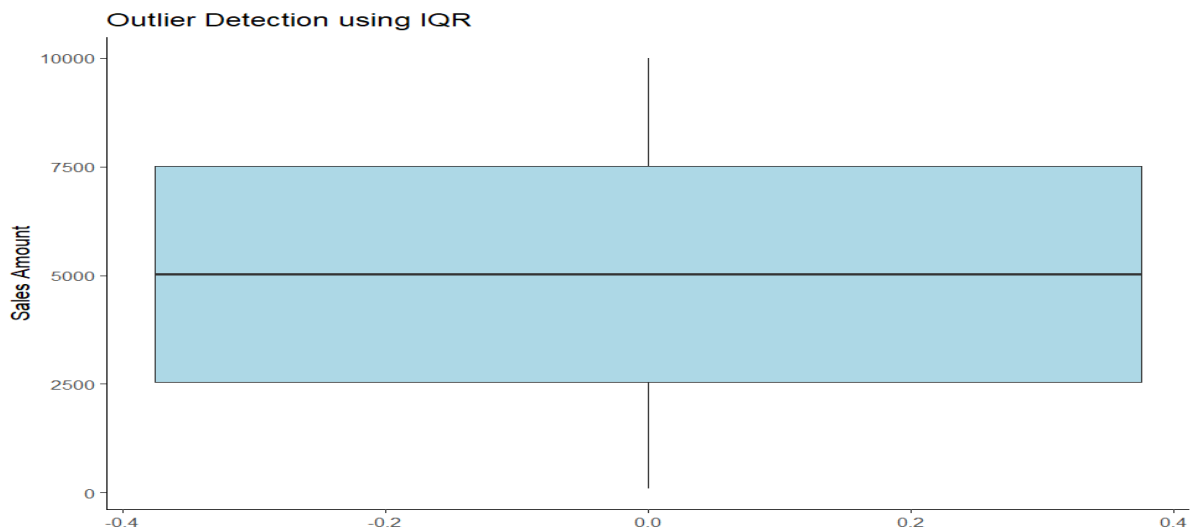
- **2023: 4,999,937**
- **2024: 19,328** (partial period)
- **Total: 5,019,265**

Reading: Nearly all sales occur in 2023, with a small carryover into early 2024, matching the dataset's date range



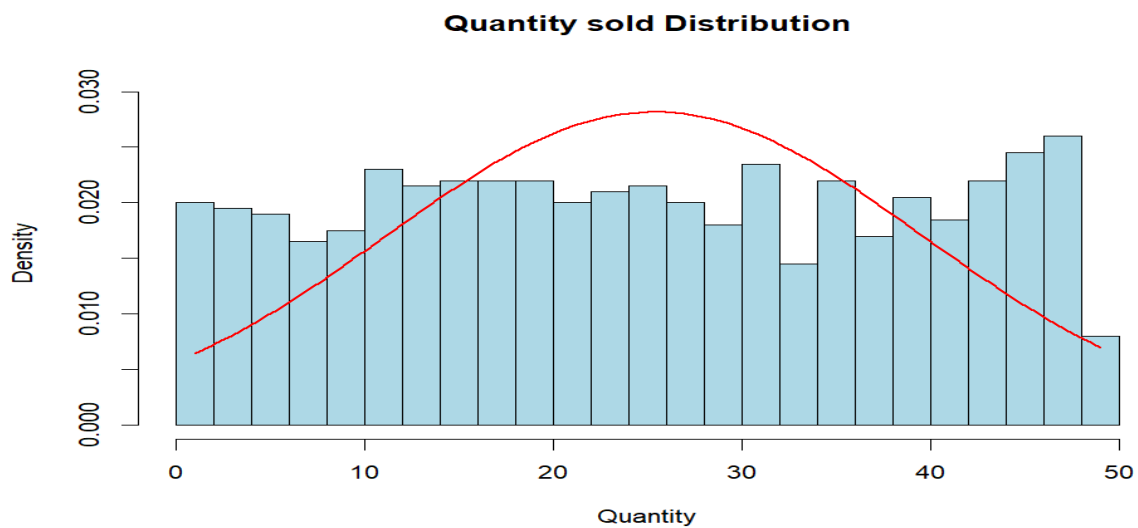
6) Outlier Detection (IQR Method)

Using the Tukey $1.5 \times \text{IQR}$ rule on transaction-level **Sales_Amount**, no values fell beyond the lower or upper fences; the distribution's quartiles produce wide enough bounds that cover all observations (consistent with the summary). **Result: no IQR outliers detected.**

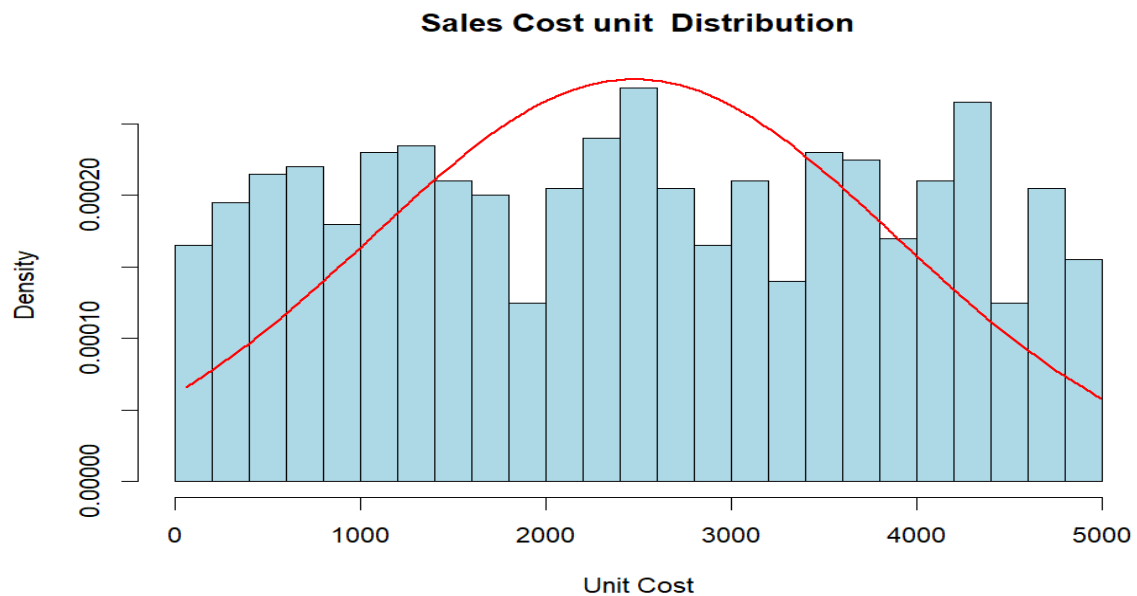


7) Distributions of Other Numeric Variables.

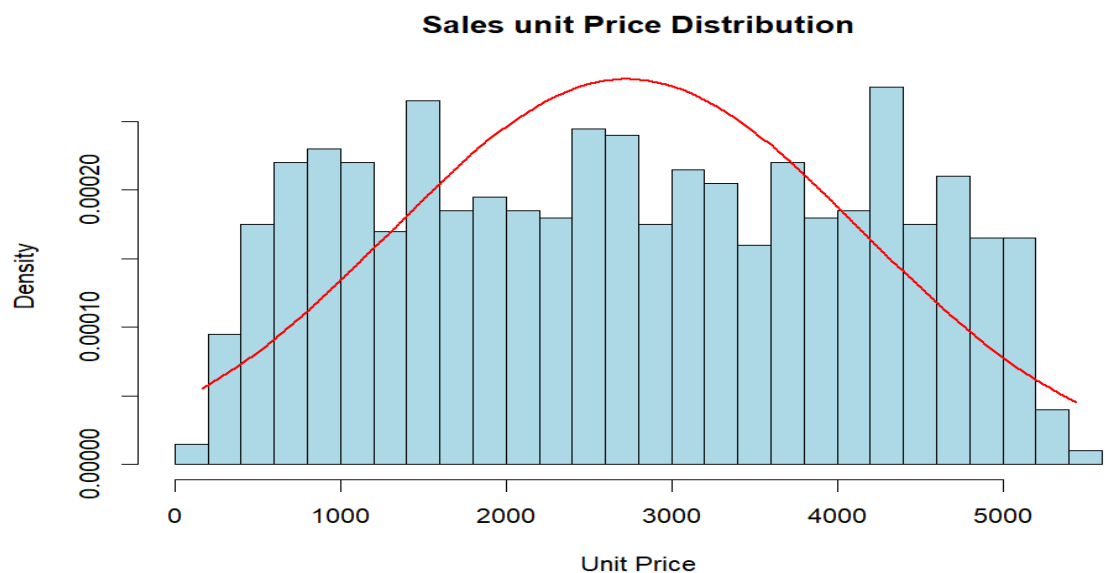
Quantity_Sold: Centered around ~25 with a broad but regular spread.



Unit_Cost / Unit_Price: Both exhibit wide ranges with prices reliably above costs as expected. **Reading:** These distributions confirm the dataset's internal consistency for downstream profitability analysis.



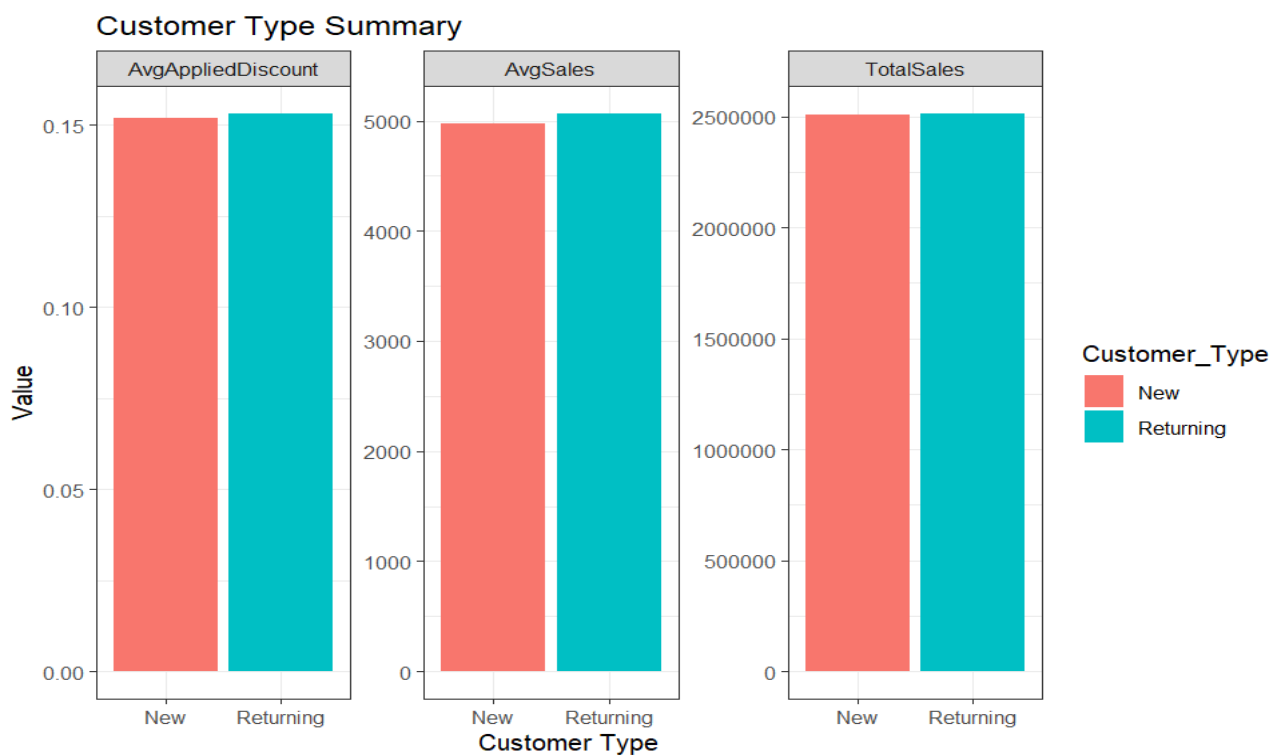
- Distribution of the sales unit Price



8) Customer Type Summary (Table & Figure).

New and Returning customers contribute **almost equally** to revenue (~2.51M each). Returning customers have a slightly higher **average ticket** and marginally higher average discount, suggesting loyalty-type incentives or higher-value purchases.

Customer_Type	TotalSales	AvgSales	AvgAppliedDiscount
1 New	2506258	4973	0.152
2 Returning	2513007	<u>5067</u>	0.153

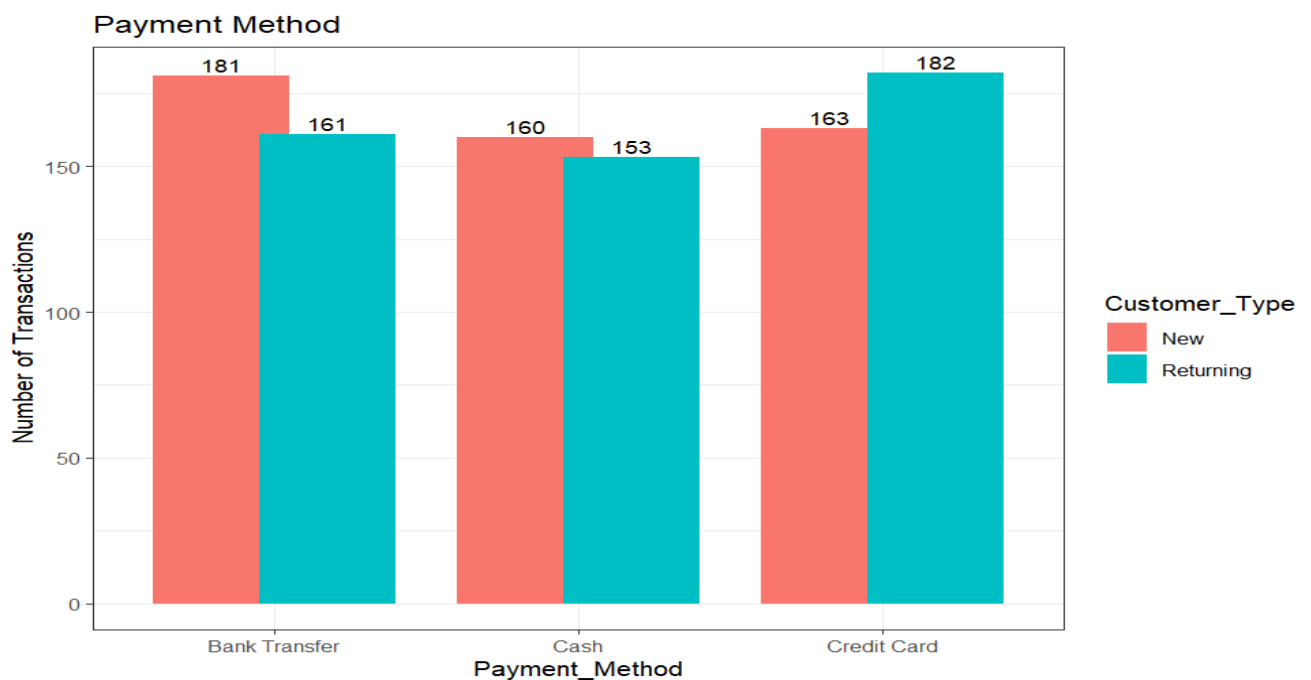


9) Payment Method Preferences by Customer Type (Table & Figure).

Preferences are distributed across methods with **Credit Card** slightly favored among **Returning** customers (18.2%). **New** customers lean a touch more toward **Bank Transfer** (18.1%). Overall shares sum to 100%, indicating balanced payment behavior across customer types.

Analysing Payment Method

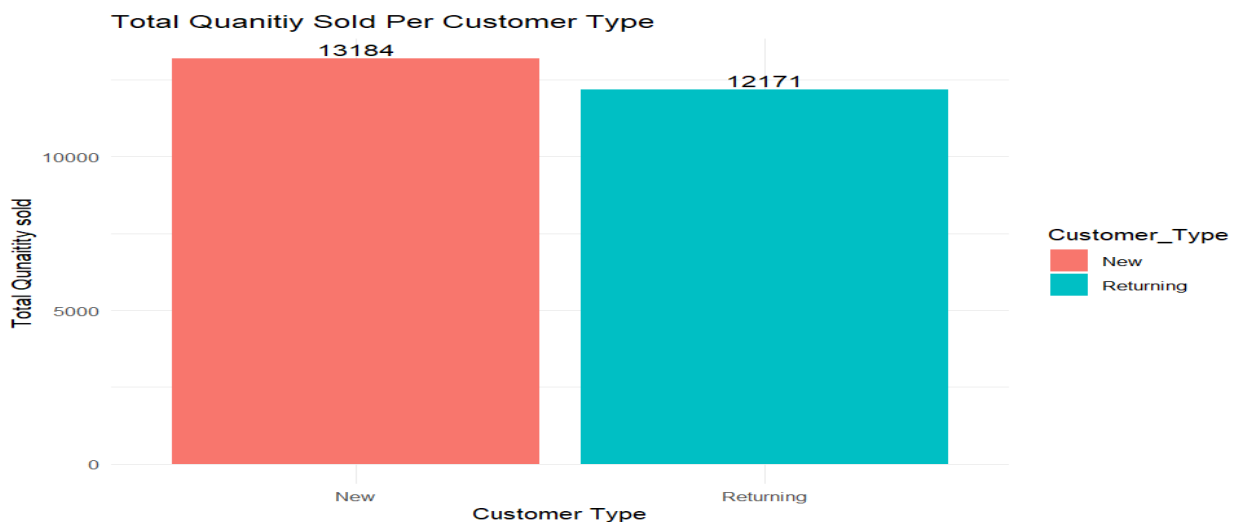
	Payment_Method	Customer_Type	count	Share
1	Bank Transfer	New	181	18.1
2	Bank Transfer	Returning	161	16.1
3	Cash	New	160	16
4	Cash	Returning	153	15.3
5	Credit Card	New	163	16.3
6	Credit Card	Returning	182	18.2



10) Total Quantity Sold per Customer Type (Table and Figure).

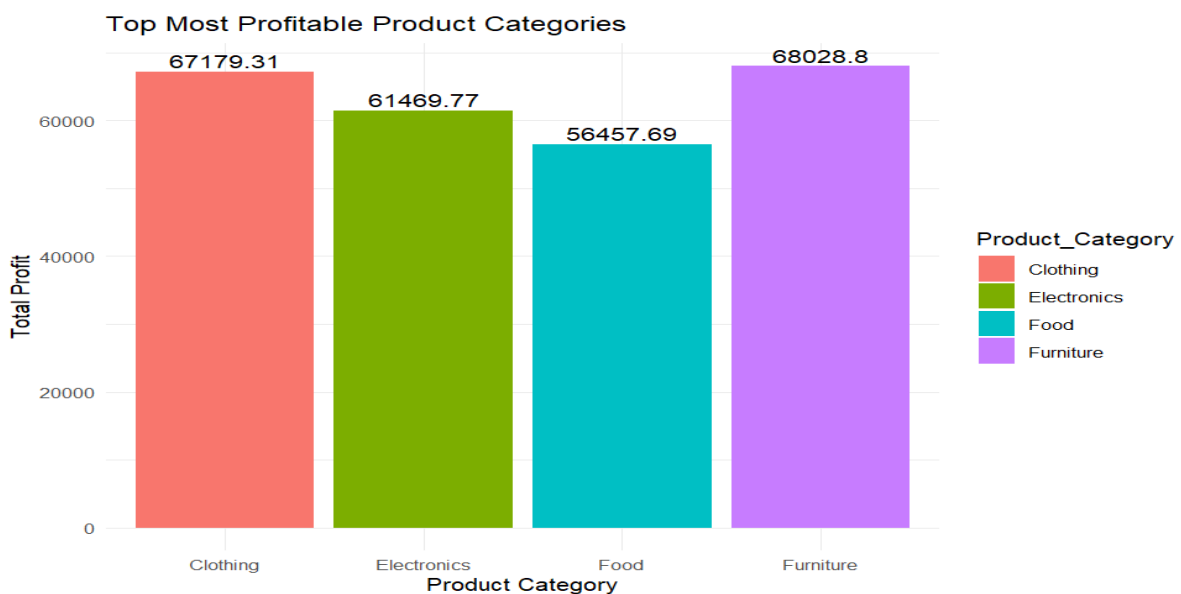
New customers purchased slightly more units in aggregate, aligning with the near-equal revenue split but possibly with smaller average prices per unit relative to Returning customers.

Customer_Type	TotalQuantity
New	13184
Returning	12171



11) Top Most Profitable Product Categories (Figure)

Furniture leads total profit, closely followed by Clothing; Electronics and Food also contribute materially. This ranking can guide category emphasis in promotions and inventory planning.



12) Data Quality, Source & Limitations

- **No missing values** reported in the working copy; distributions and summaries are internally consistent.
- **No IQR outliers** detected at the transaction level.
- **Source:** Synthetic practice dataset from Kaggle (for educational/testing use only). Conclusions are illustrative and **not** tied to real operations.

13) Conclusions & Practical Notes

- Revenue is split nearly 50/50 between New and Returning customers; Returning customers' slightly higher ticket suggests potential for **loyalty programs** or **upsell bundles**.
- Payment method preferences are broadly distributed, warranting **continued support of all three** options (Credit Card, Cash, Bank Transfer).
- Category profitability prioritizes **Furniture** and **Clothing**; targeted **promotions** and **stock depth** there can lift profit.
- No extreme outliers and stable distributions indicate a clean base for **time-series modeling** or **AB testing** on discounts.