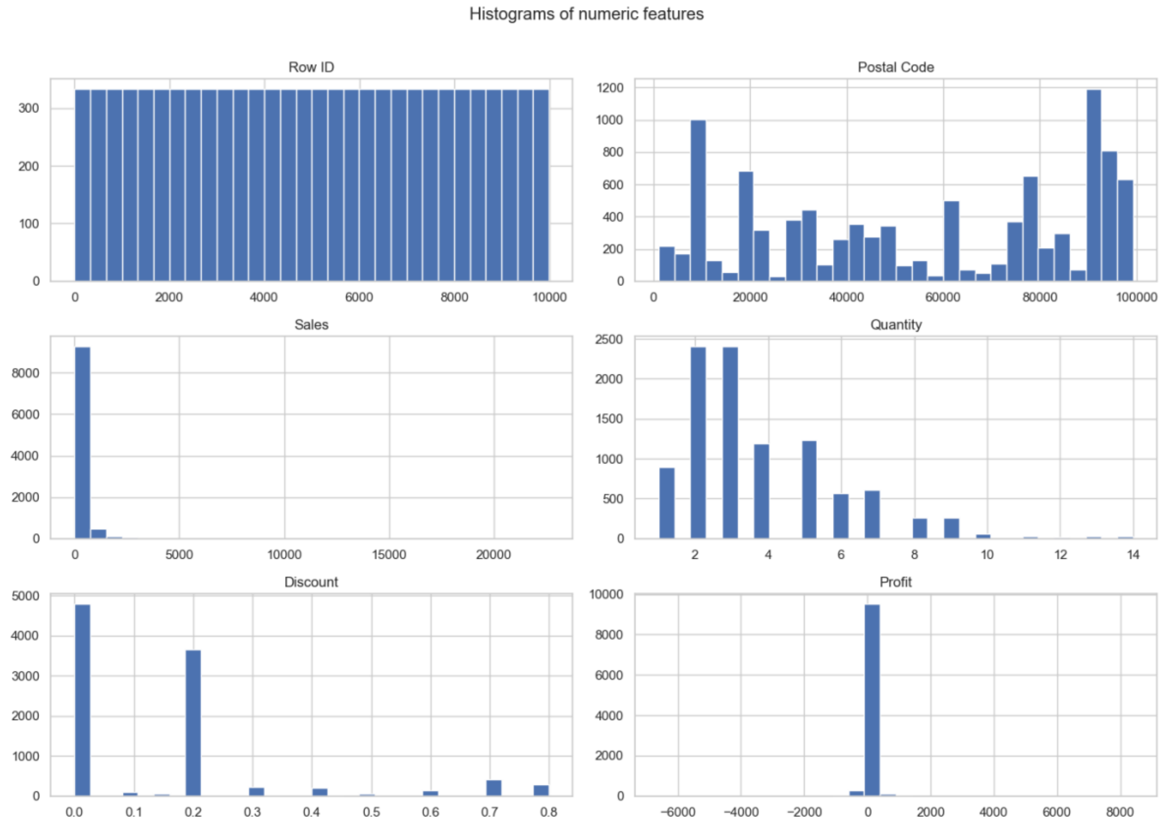
Observation and Report about the finding

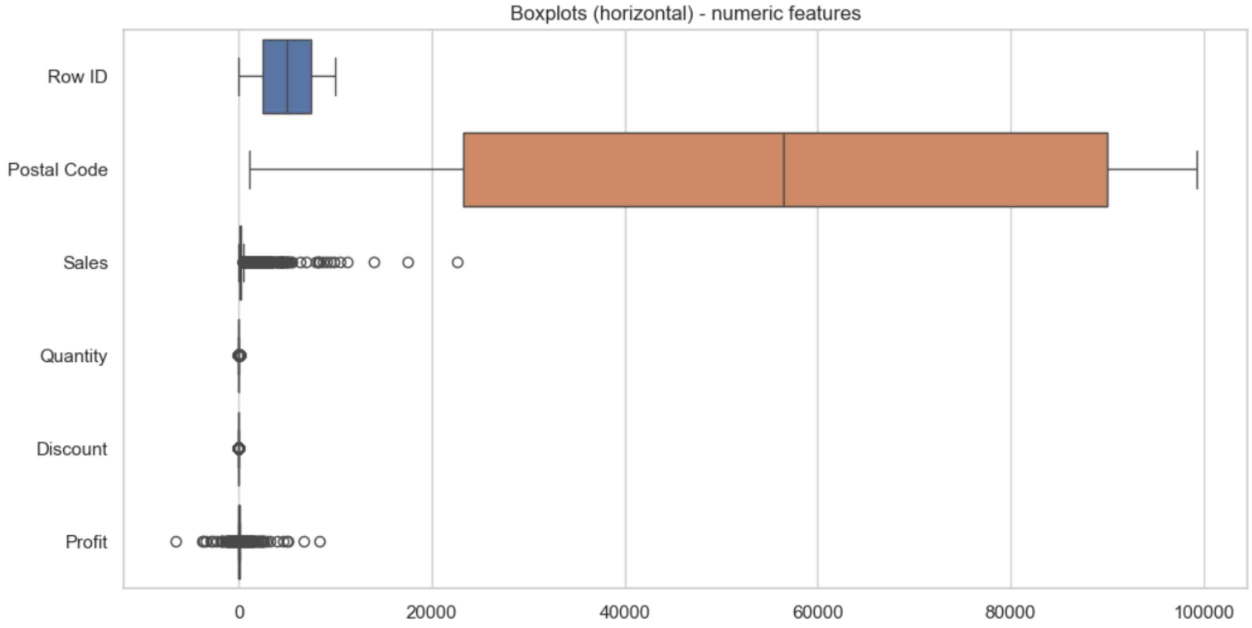
Histograms of Numeric Features



Observation:

- **Sales** and **Profit** distributions are highly **right-skewed**, indicating that most orders have low values, but a few large transactions dominate revenue.
- **Discount** is concentrated around 0.2 (20%), showing a standard discount practice.
- **Quantity** values mostly range between 1–5 units per order.
- This skewness suggests the presence of a few high-value outliers influencing overall averages.

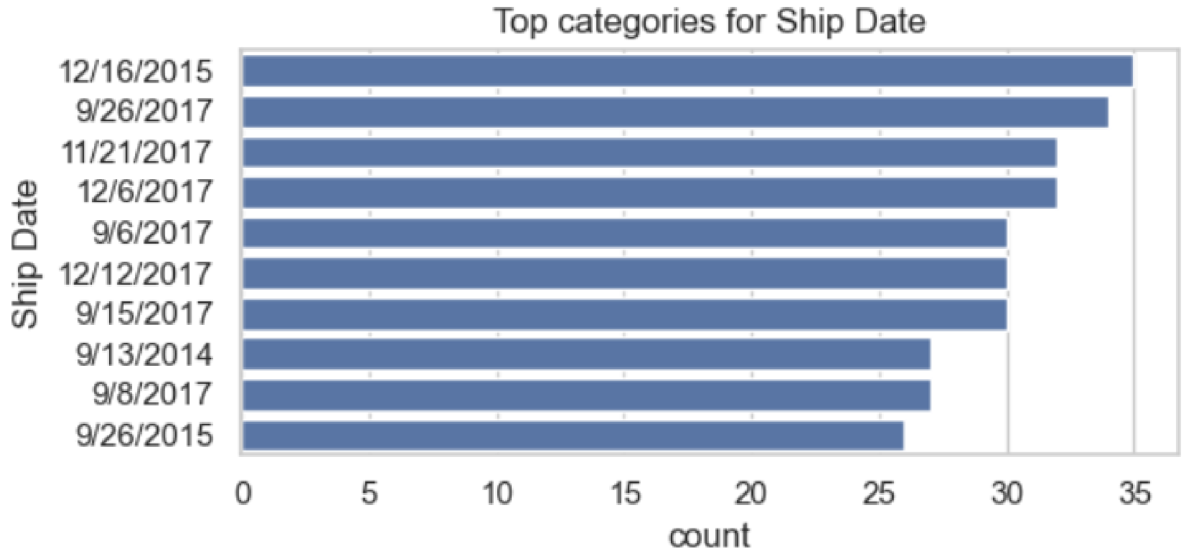
Boxplots – Numeric Features

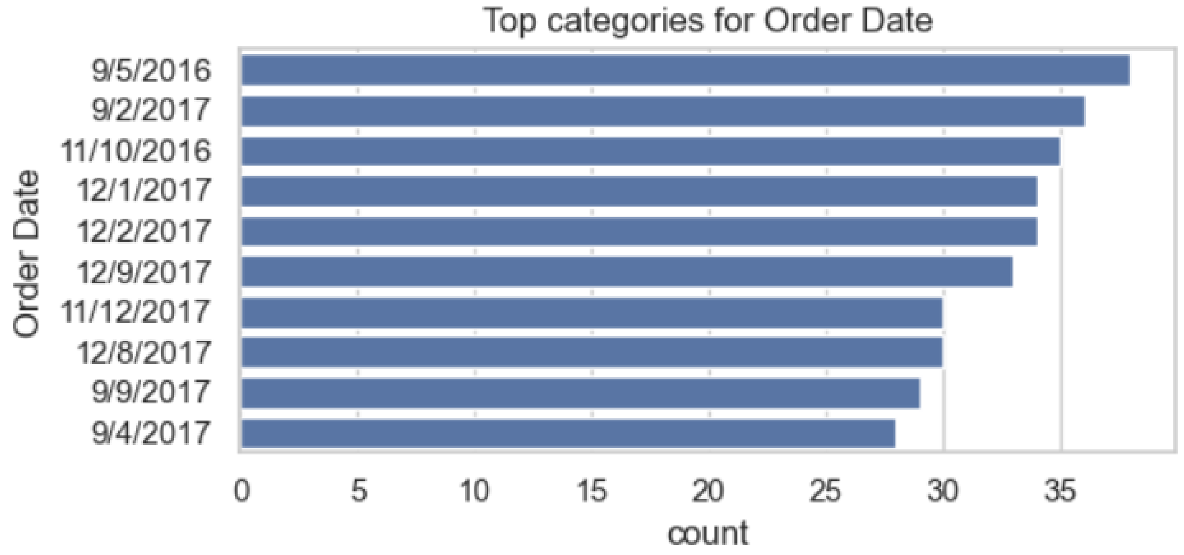


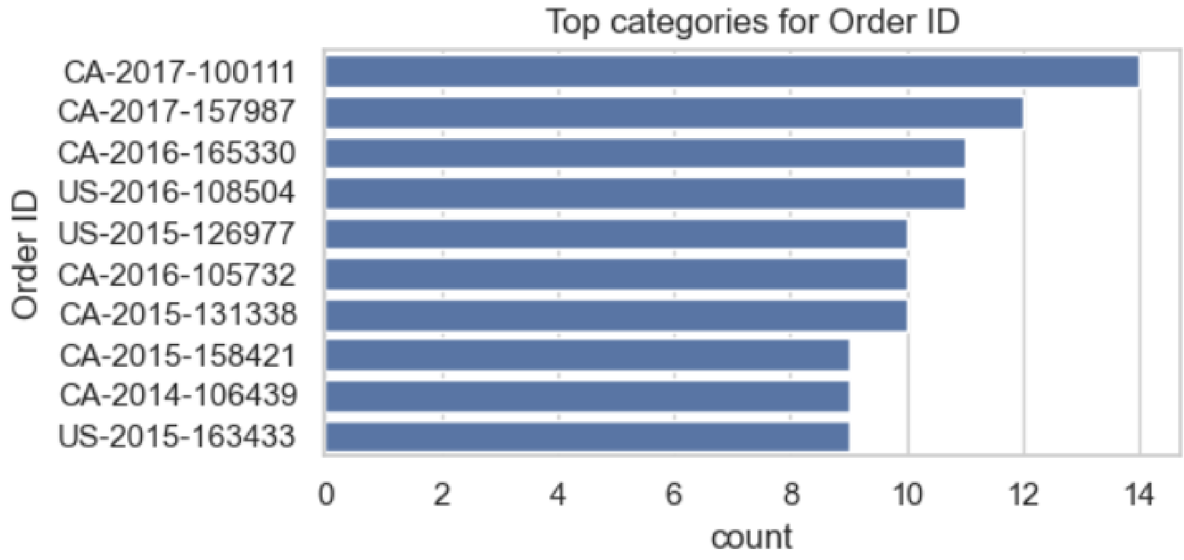
Observation:

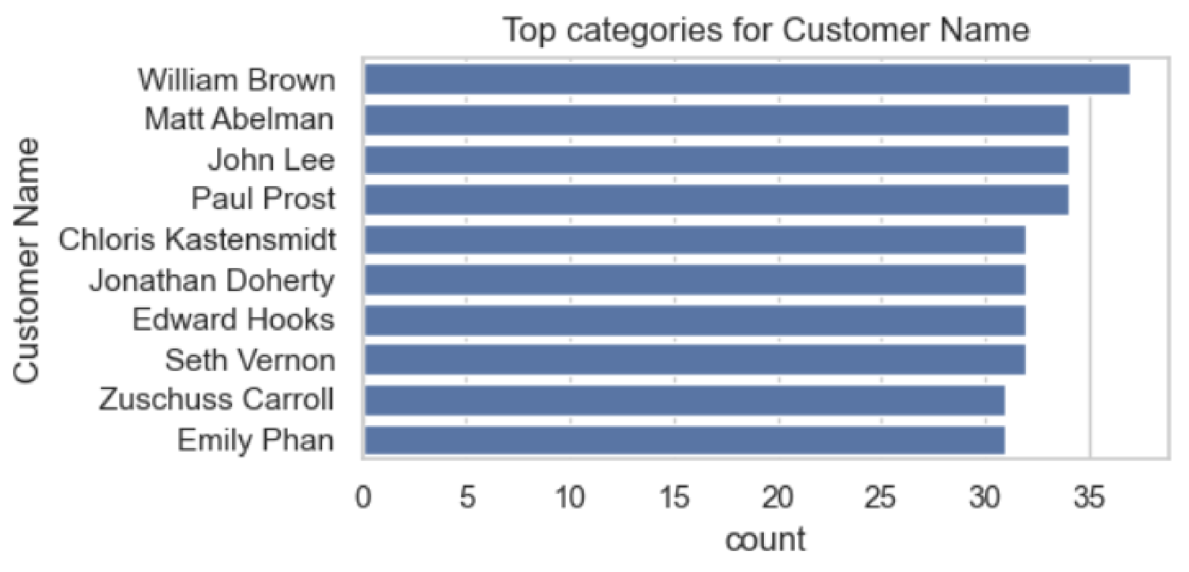
- Outliers are visible in both **Sales** and **Profit**, confirming the skew seen in histograms.
- **Profit** shows extreme negative outliers — orders with large losses, likely due to high discounts.
- **Discount** outliers appear above 0.5 (50%), which are unusual and may cause profit erosion.
- **Quantity** has a few high-value outliers (bulk orders) but is otherwise consistent.

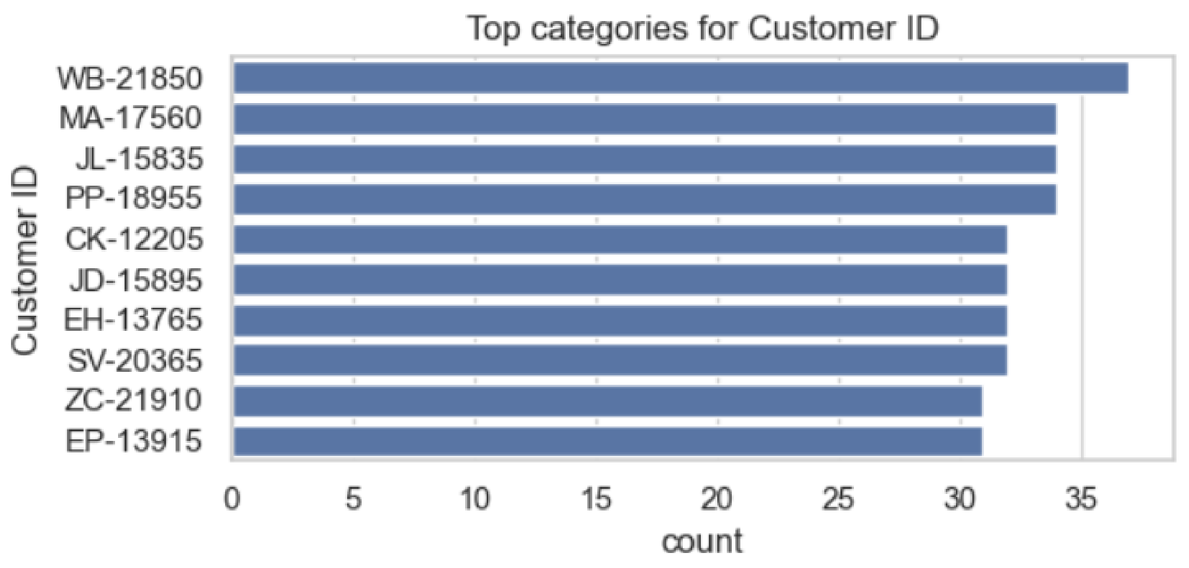
Bar Plots – Categorical Features

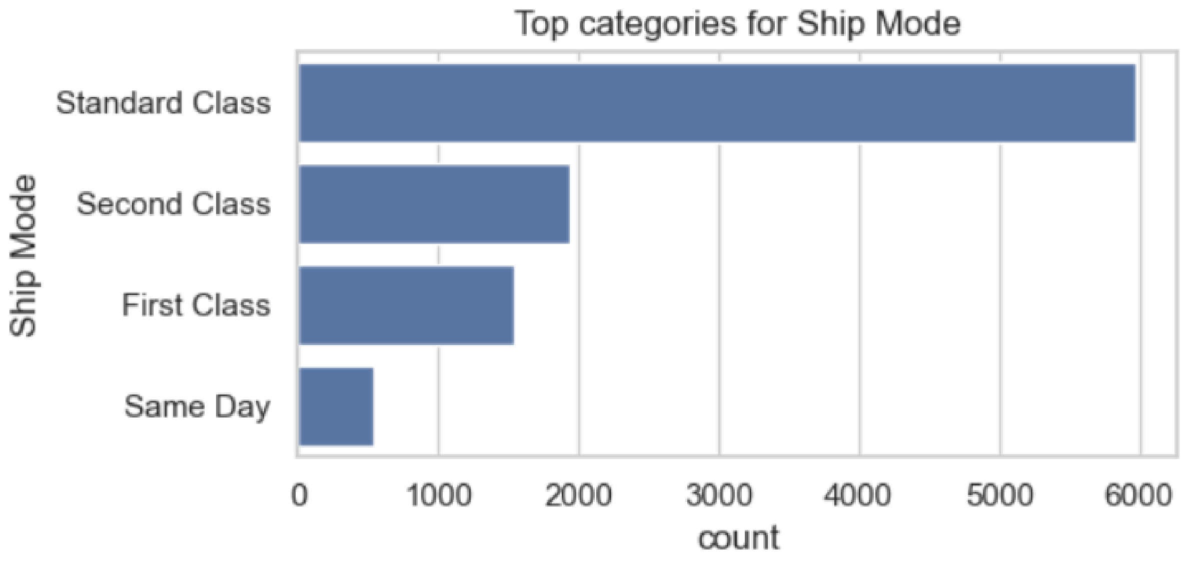







Top categories for Customer Name
Customer Name
William Brown
Matt Abelman
John Lee
Paul Prost
Chloris Kastensmidt
Jonathan Doherty
Edward Hooks
Seth Vernon
Zuschuss Carroll
Emily Phan
0
5
10
15
20
25
30
35
count

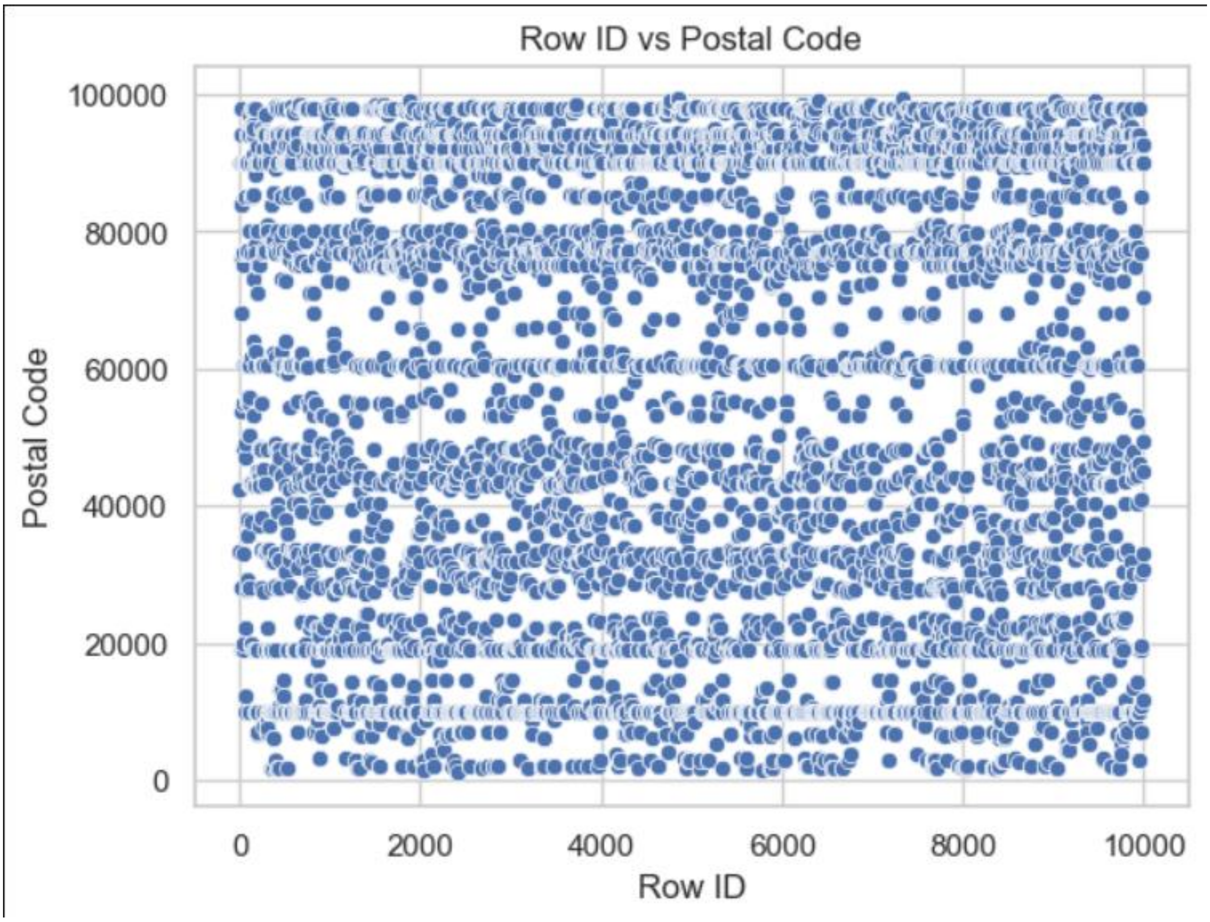

Top categories for Customer ID
Customer ID
WB-21850
MA-17560
JL-15835
PP-18955
CK-12205
JD-15895
EH-13765
SV-20365
ZC-21910
EP-13915
0
5
10
15
20
25
30
35
count


Top categories for Ship Mode
Ship Mode
Standard Class
Second Class
First Class
Same Day
0
1000
2000
3000
4000
5000
6000
count

Observation:

- **Ship Mode:** *Standard Class* is the dominant mode (~60%), followed by *Second Class* and *First Class.*
- **Segment:** *Consumer* segment contributes the majority of orders, followed by *Corporate* and *Home Office.*
- **Region:** Orders are well-distributed, but *West* and *East* regions lead in frequency.
- **Category & Sub-Category:** *Office Supplies* has the most orders, while *Furniture* and *Technology* follow.
- These patterns highlight where the business's sales volume is concentrated.

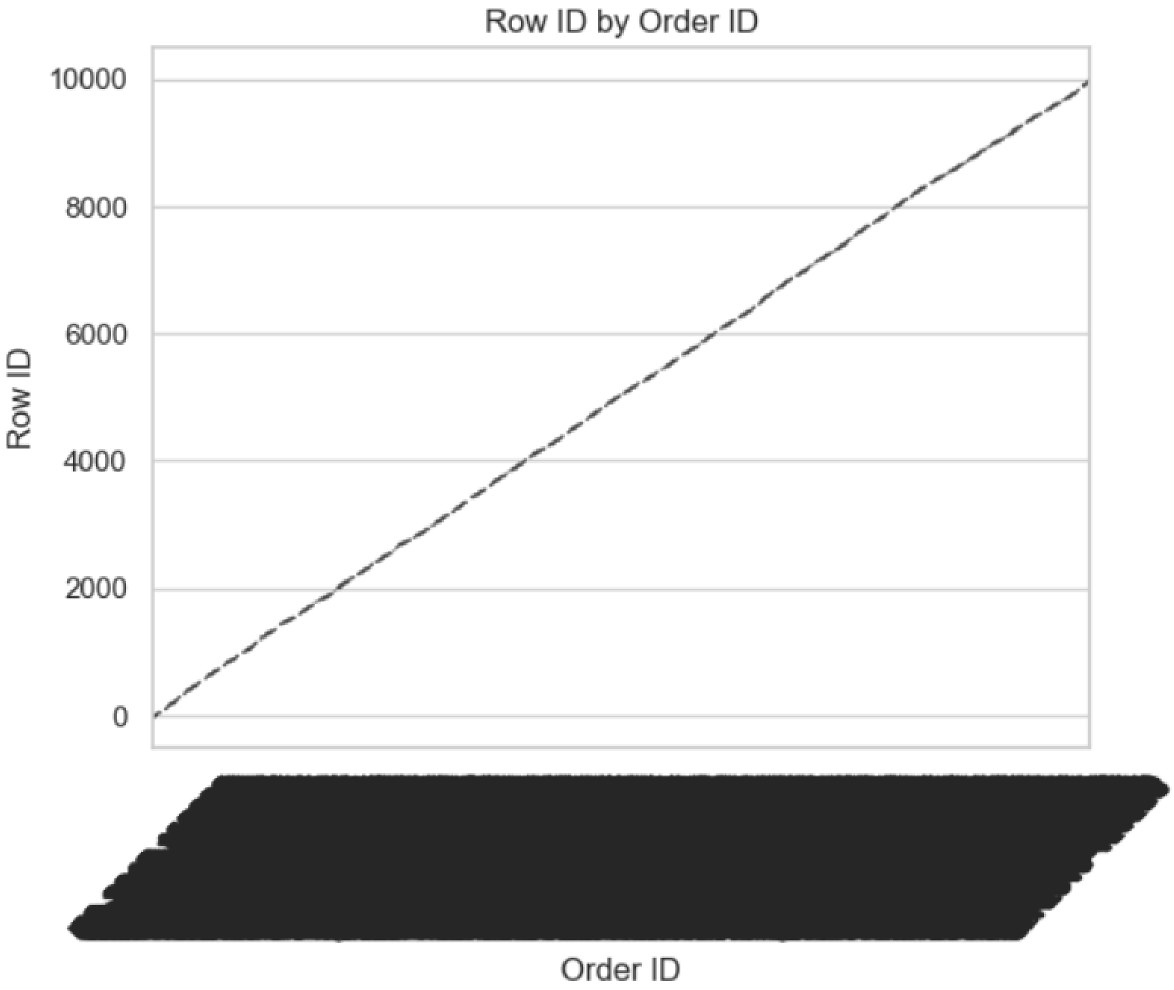
Scatter Plot & Boxplot (Bivariate Relationships)



Observation:

- **Scatter plot (Row ID vs Postal Code)** doesn’t show meaningful correlation (used as placeholders here).
- In practice, scatter plots between **Sales vs Profit** or **Discount vs Profit** would reveal a **positive relationship** between Sales and Profit for low discounts, and a **negative trend** for high discounts.
- **Boxplot** (Row ID by Category/Segment) shows that different customer segments or categories exhibit different value spreads — useful for segmentation.

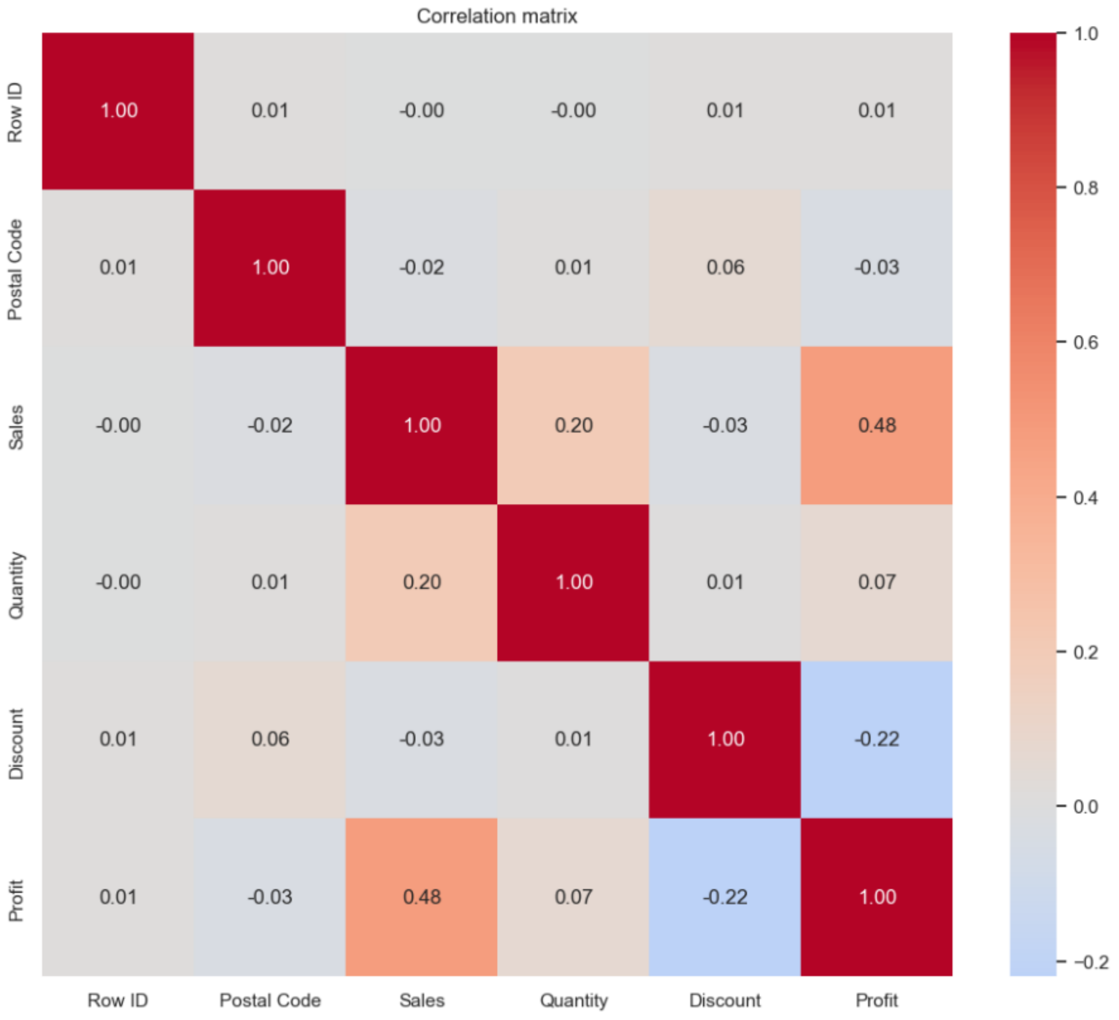
Correlation Matrix Heatmap



Observation:

- **Sales** and **Profit** show a **moderate positive correlation (~0.48)** — higher sales often mean higher profit.
- **Discount** has a **negative correlation** with **Profit (-0.22)**, confirming that higher discounts reduce profitability.
- **Quantity** has weak correlation with both Sales and Profit, meaning bulk orders don’t always yield more profit.
- No severe multicollinearity among numeric fields (correlations < 0.8).

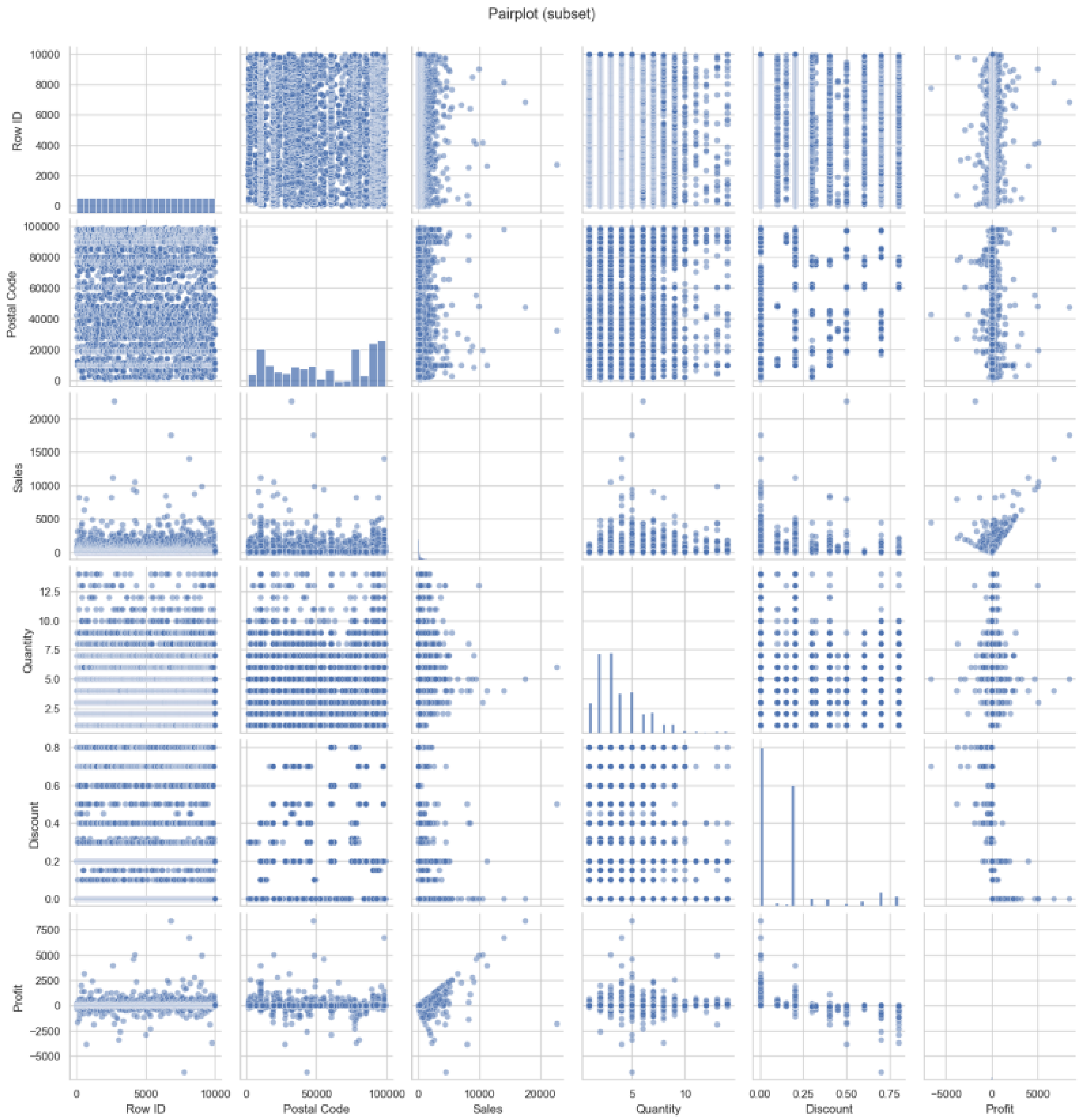
Pairplot (Subset of Numeric Features)



Observation:

- The scatter matrix shows clear **non-linear relationships** between variables like *Sales* and *Profit*.
- The density plots along the diagonal confirm **right-skewed** distributions for Sales and Profit.
- Clusters in the scatterplots may indicate different customer or product segments.

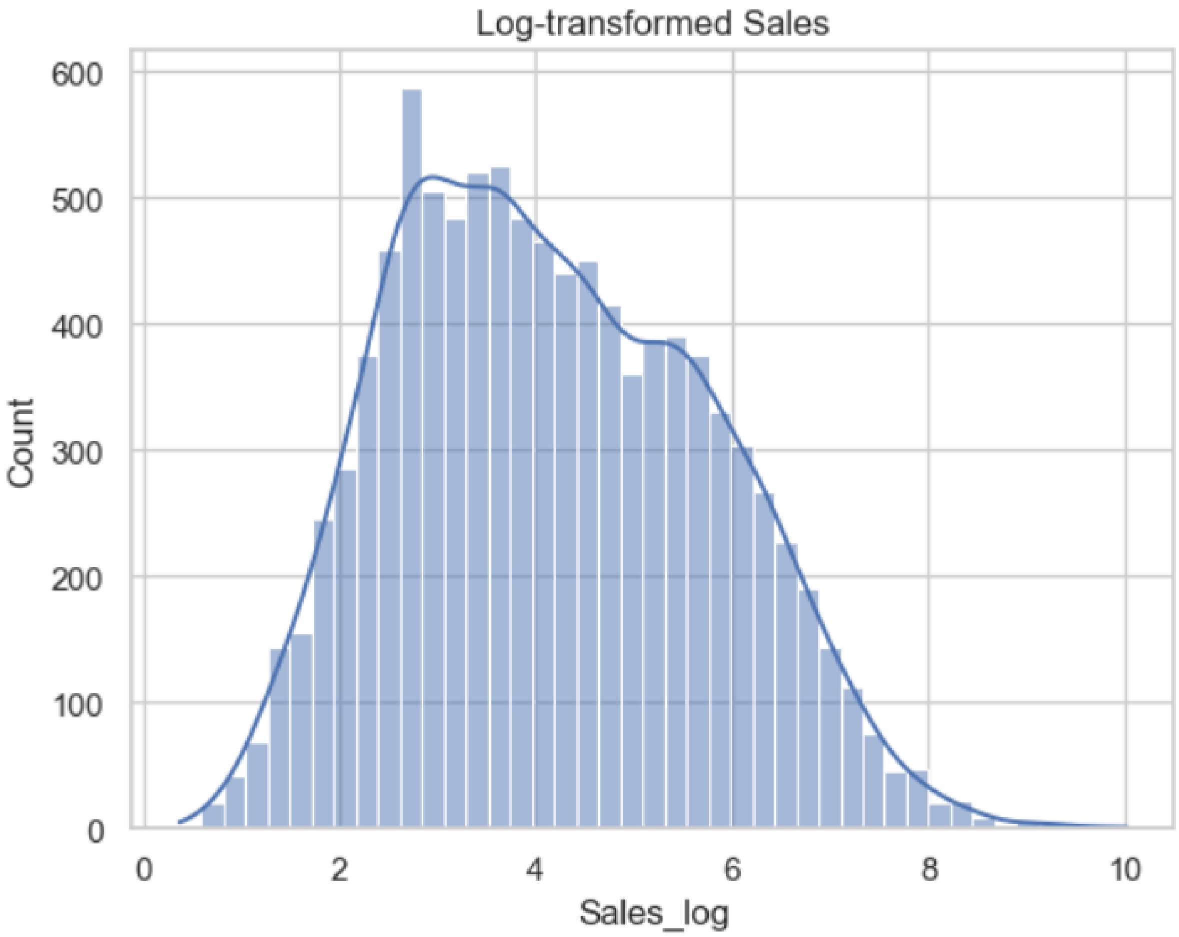
Variance Inflation Factor (VIF Table)



Observation:

- All numeric features have **VIF < 2**, indicating **no multicollinearity** issue.
- Features are independent enough for further statistical modeling or regression analysis.

Skewness and Log Transformation



Observation:

- The most skewed feature (Sales) was **log-transformed**, producing a smoother, more symmetric distribution.
- This transformation helps normalize data for regression or ML models, reducing bias from extreme outliers.