Exploratory Data Analysis Report
large_sales_dataset.csv

Rows: 500 Columns: 7

Numeric columns: 3 Categorical columns: 4

This report includes: dataset head, missing value summary, descriptive statistics, correlation heatmap, histograms, boxplots for numeric features, and sample scatterplots for pairs of numeric variables. Observations are provided after visuals.

Top missing columns (by %):
1. Order_ID: 0.0%
2. Order_Date: 0.0%
3. Category: 0.0%
4. Product: 0.0%
5. Region: 0.0%
6. Sales: 0.0%
7. Profit: 0.0%

First 10 rows (dataset head)

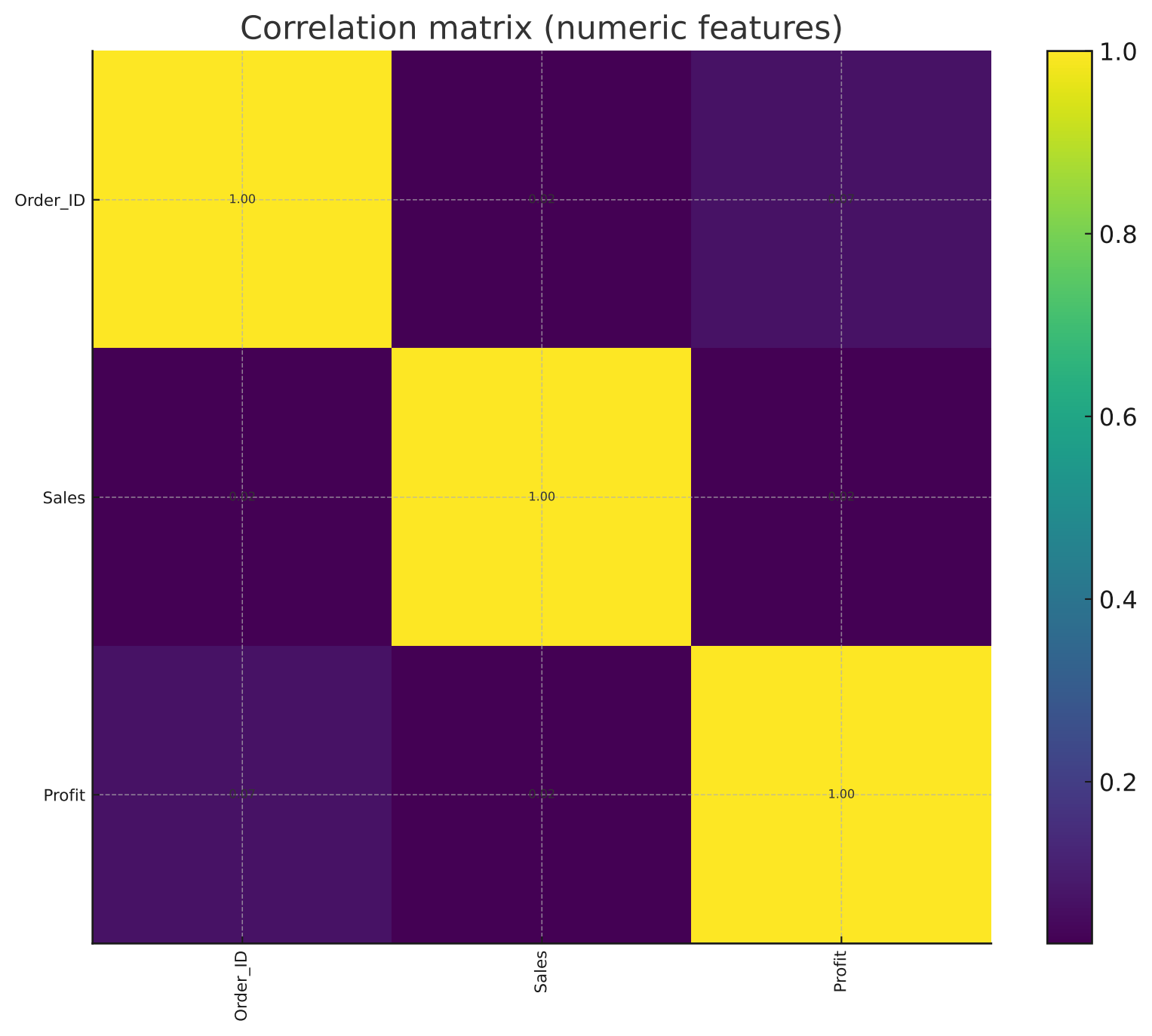
Order_ID	Order_Date	Category	Product	Region	Sales	Profit
2001	2023-01-01	Office Supplies	Keyboard	East	1858	294
2002	2023-01-02	Furniture	Whiteboard	South	31	-37
2003	2023-01-03	Office Supplies	Chair	South	1361	-29
2004	2023-01-04	Office Supplies	Tablet	East	1635	173
2005	2023-01-05	Furniture	Markers	East	235	219
2006	2023-01-06	Furniture	Table	East	358	237
2007	2023-01-07	Office Supplies	Sofa	West	1819	265
2008	2023-01-08	Technology	Sofa	South	680	394
2009	2023-01-09	Office Supplies	Laptop	West	1124	65
2010	2023-01-10	Office Supplies	Router	West	1770	443

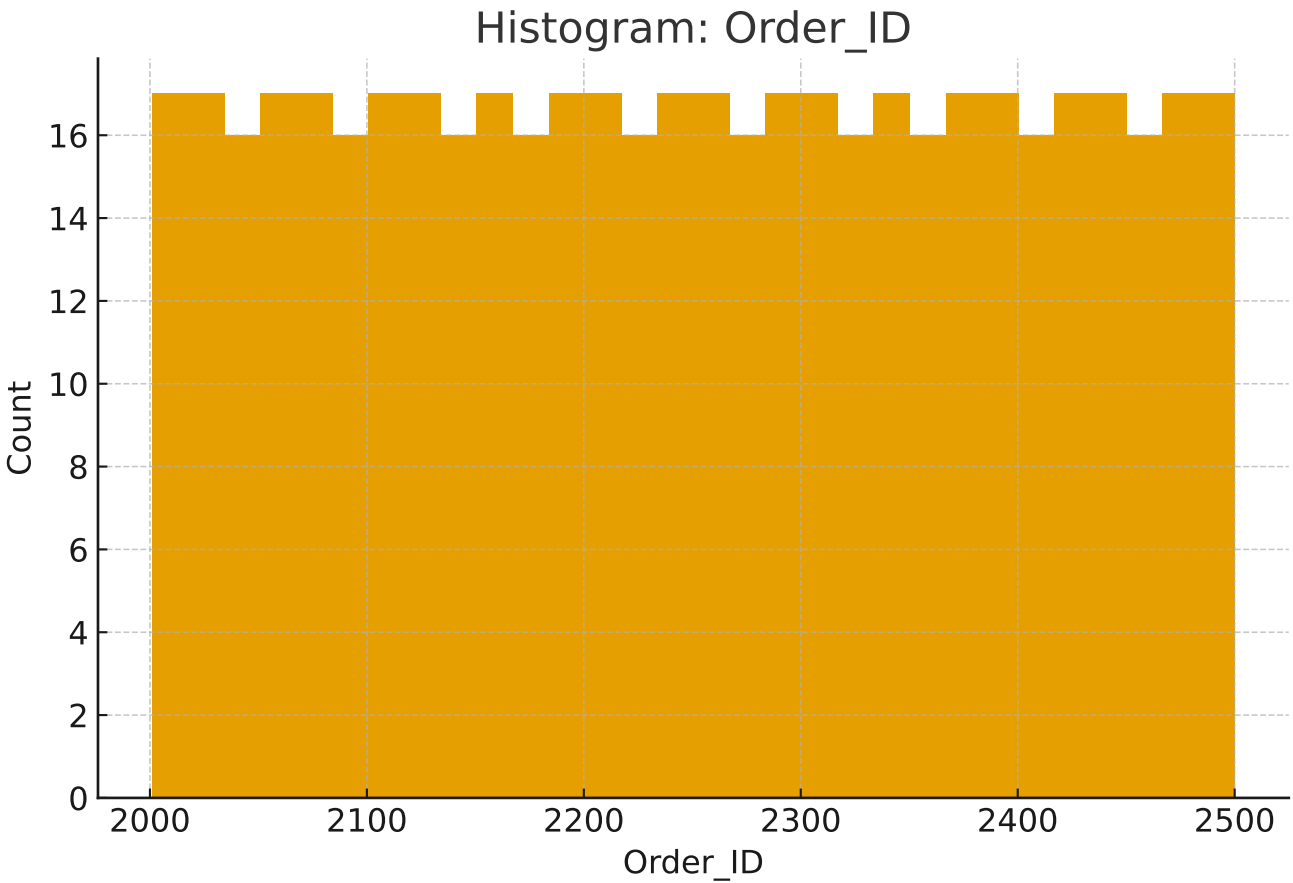
Missing Values Summary (top 20)

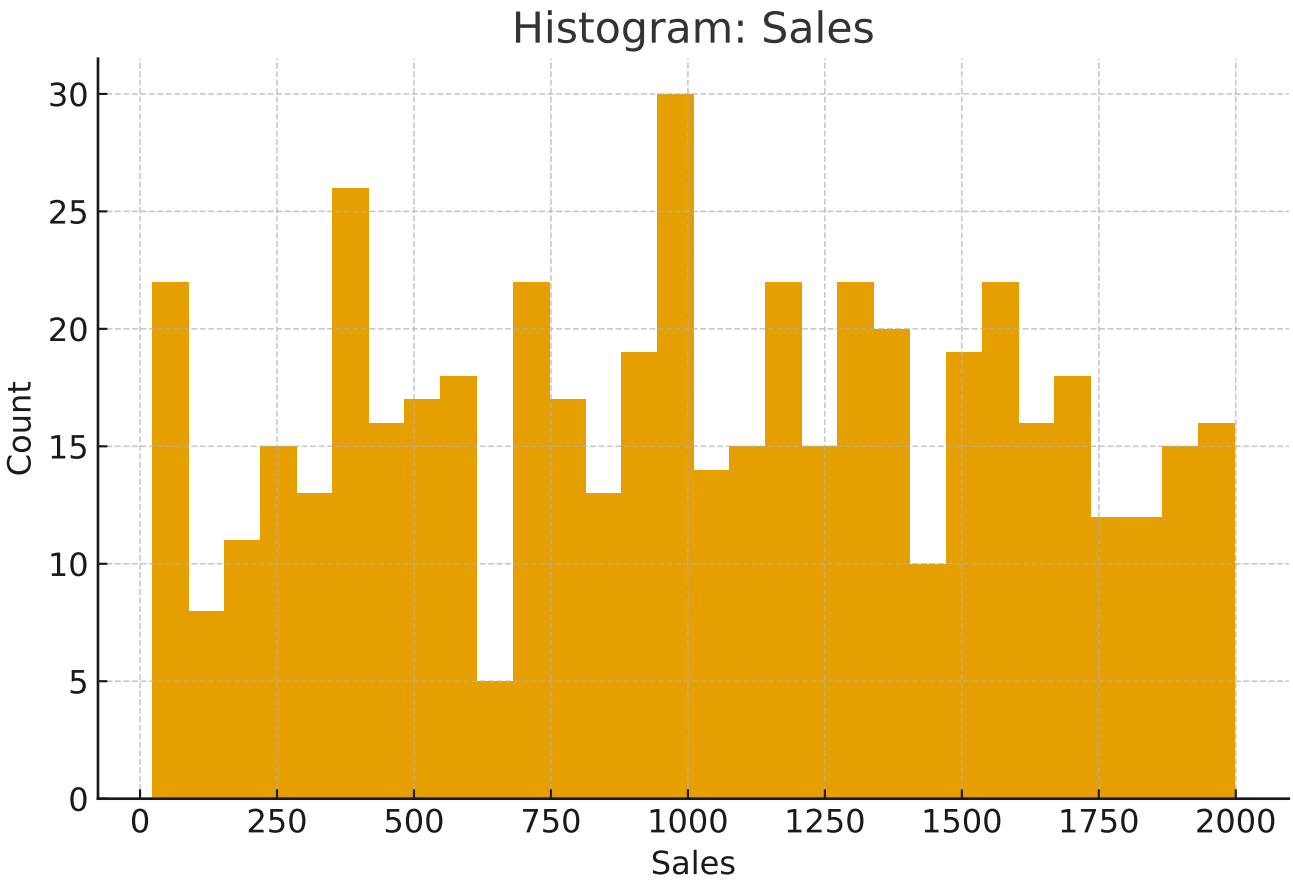
column	missing_count	missing_percent
Order_ID	0	0.0
Order_Date	0	0.0
Category	0	0.0
Product	0	0.0
Region	0	0.0
Sales	0	0.0
Profit	0	0.0

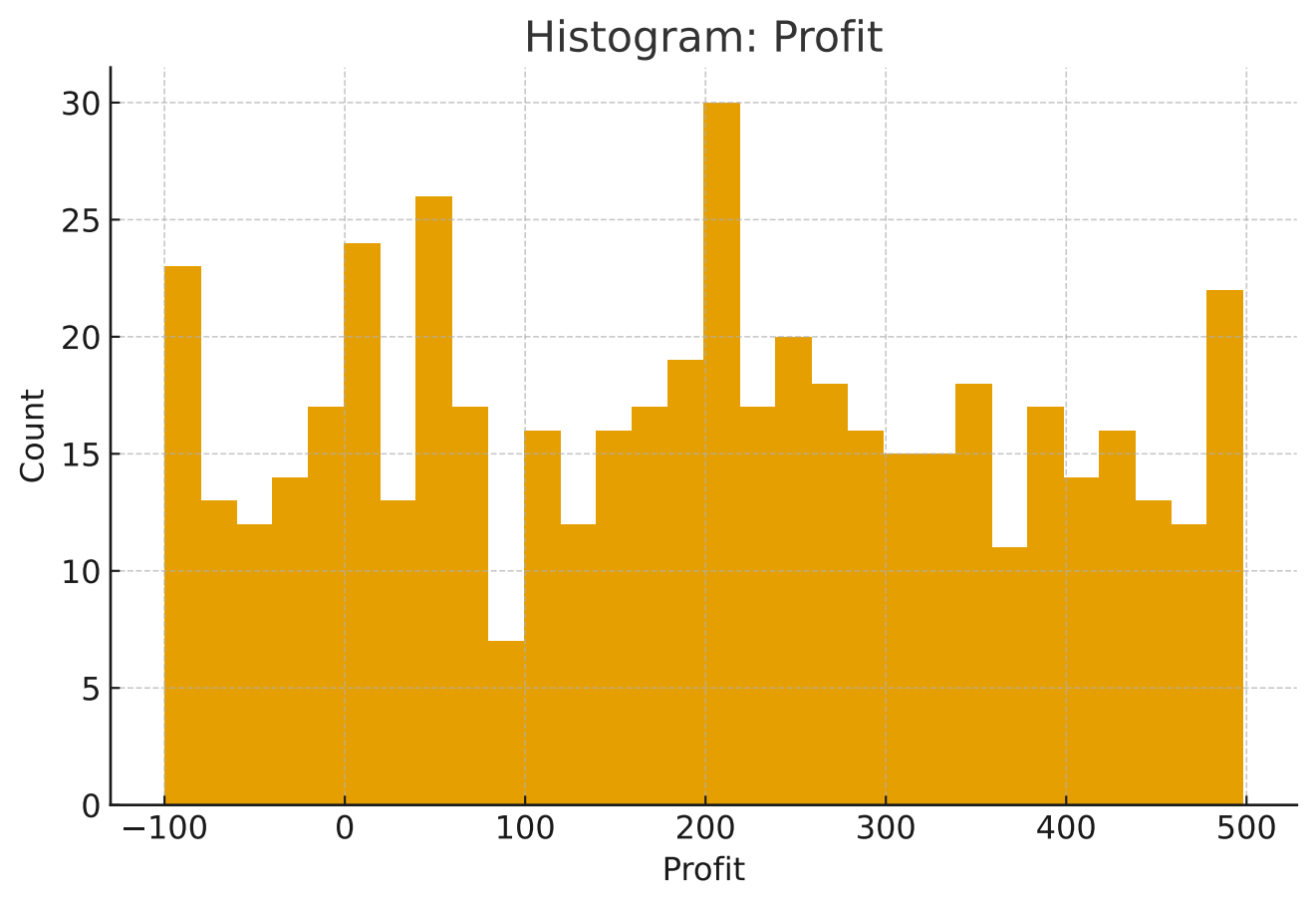
Descriptive Statistics (numeric + object -- top 20 rows)

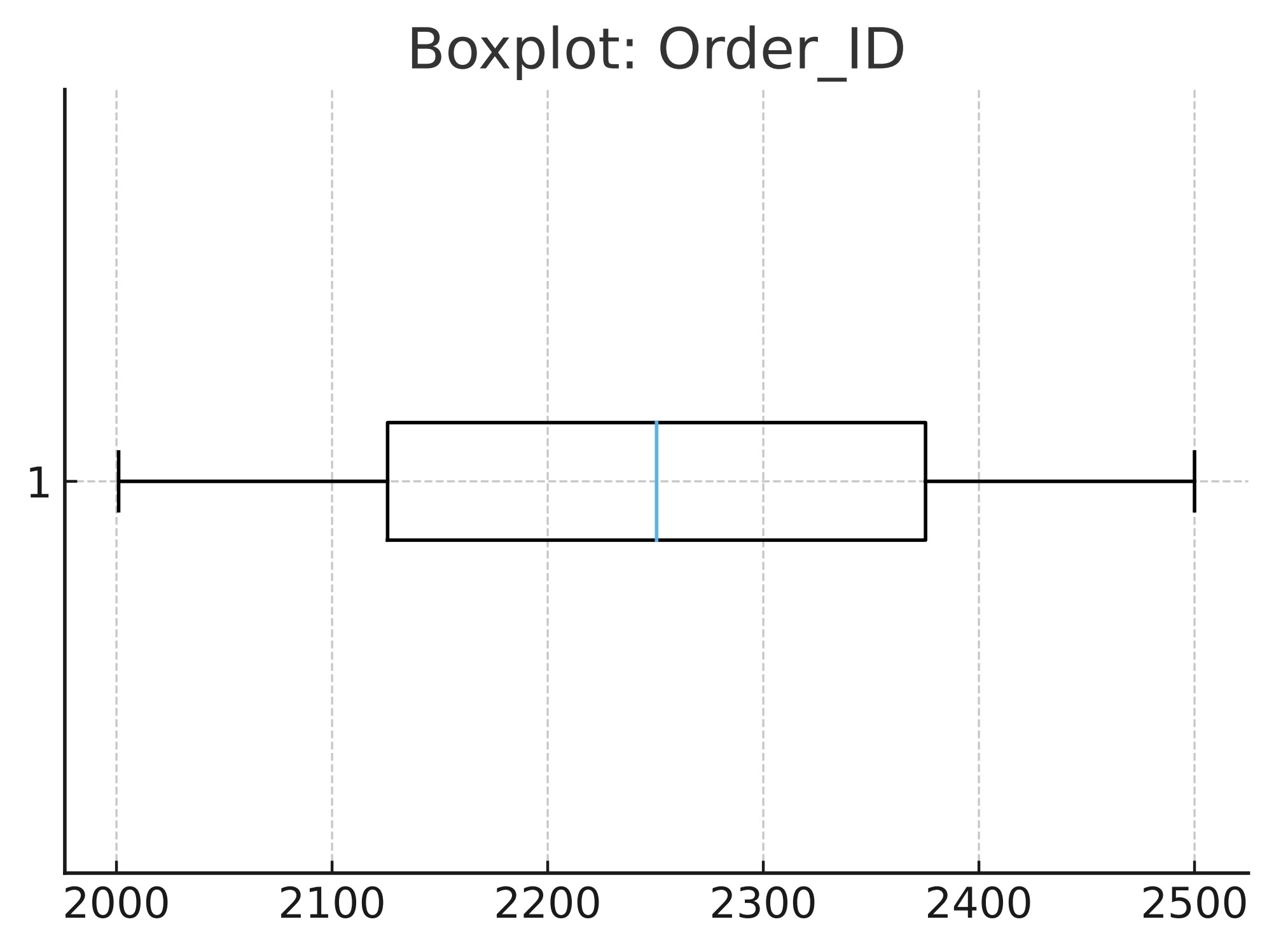
feature	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Order_ID	500.0	nan	nan	nan	2250.5	144.51518327679989	2001.0	2125.75	2250.5	2375.25	2500.0
Order_Date	500	500	2023-01-01	1	nan	nan	nan	nan	nan	nan	nan
Category	500	3	Furniture	168	nan	nan	nan	nan	nan	nan	nan
Product	500	20	Tablet	34	nan	nan	nan	nan	nan	nan	nan
Region	500	4	East	132	nan	nan	nan	nan	nan	nan	nan
Sales	500.0	nan	nan	nan	1056.814	556.7462087790665	22.0	540.0	1008.0	1483.0	1998.0
Profit	500.0	nan	nan	nan	194.548	194.5483611094157	-100.0	45.75	200.5	332.0	498.0

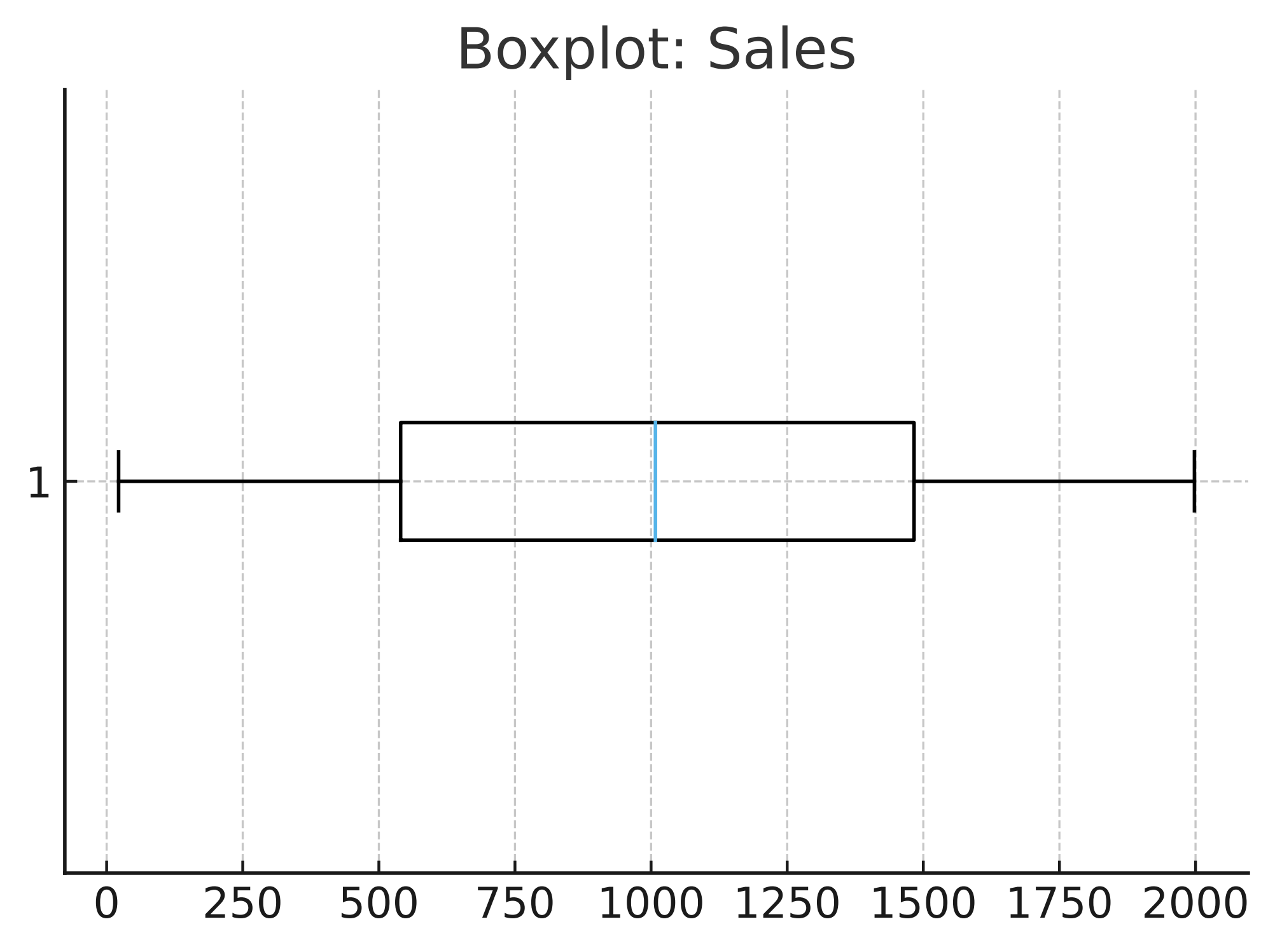

Correlation matrix (numeric features)
Order_ID
Sales
Profit
1.00
0.00
0.00
0.00
1.00
0.00
0.00
0.00
1.00
Order_ID
Sales
Profit
1.0
0.8
0.6
0.4
0.2

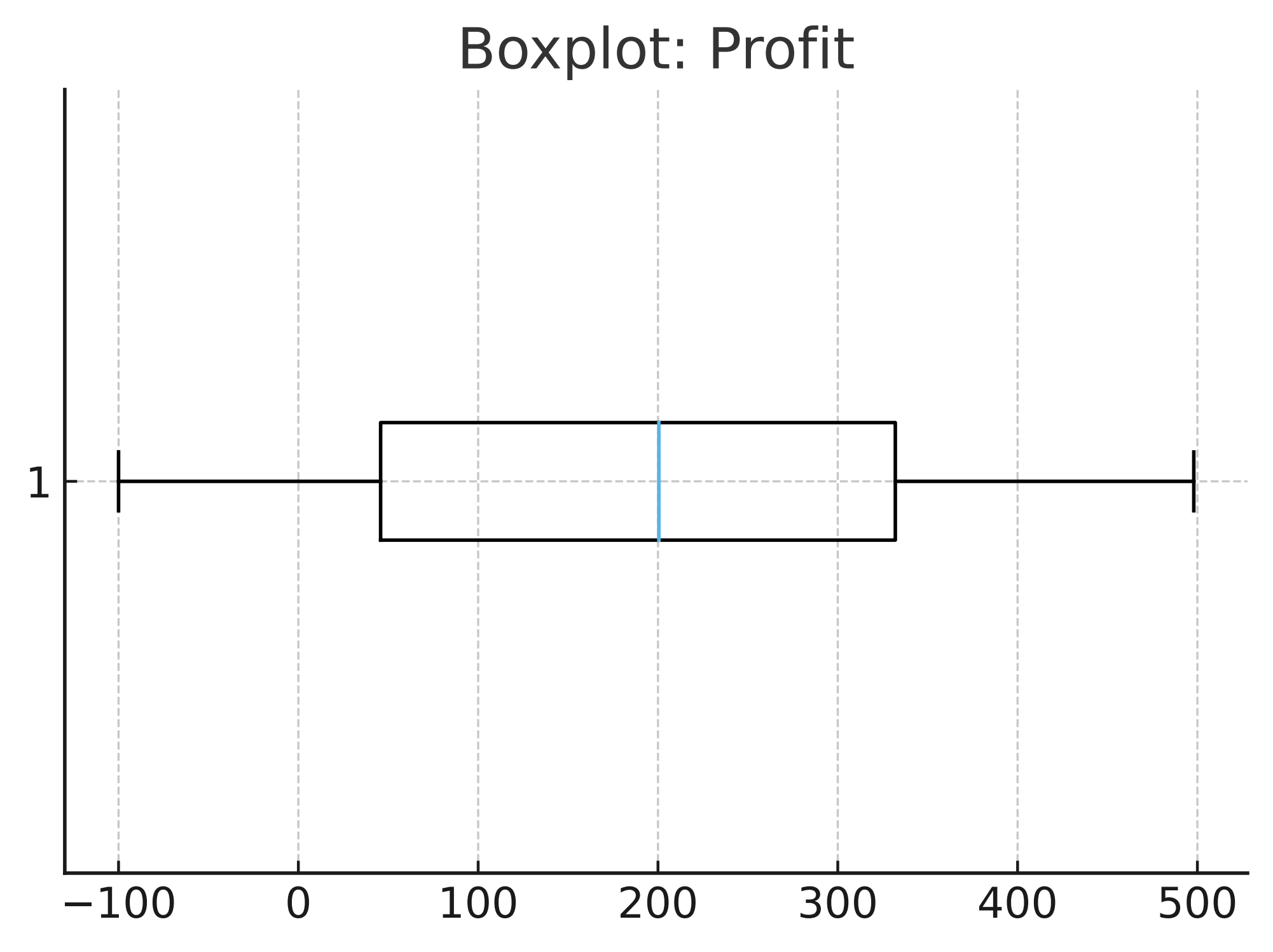

Histogram: Order_ID
Count
0
2
4
6
8
10
12
14
16
2000
2100
2200
2300
2400
2500
Order_ID

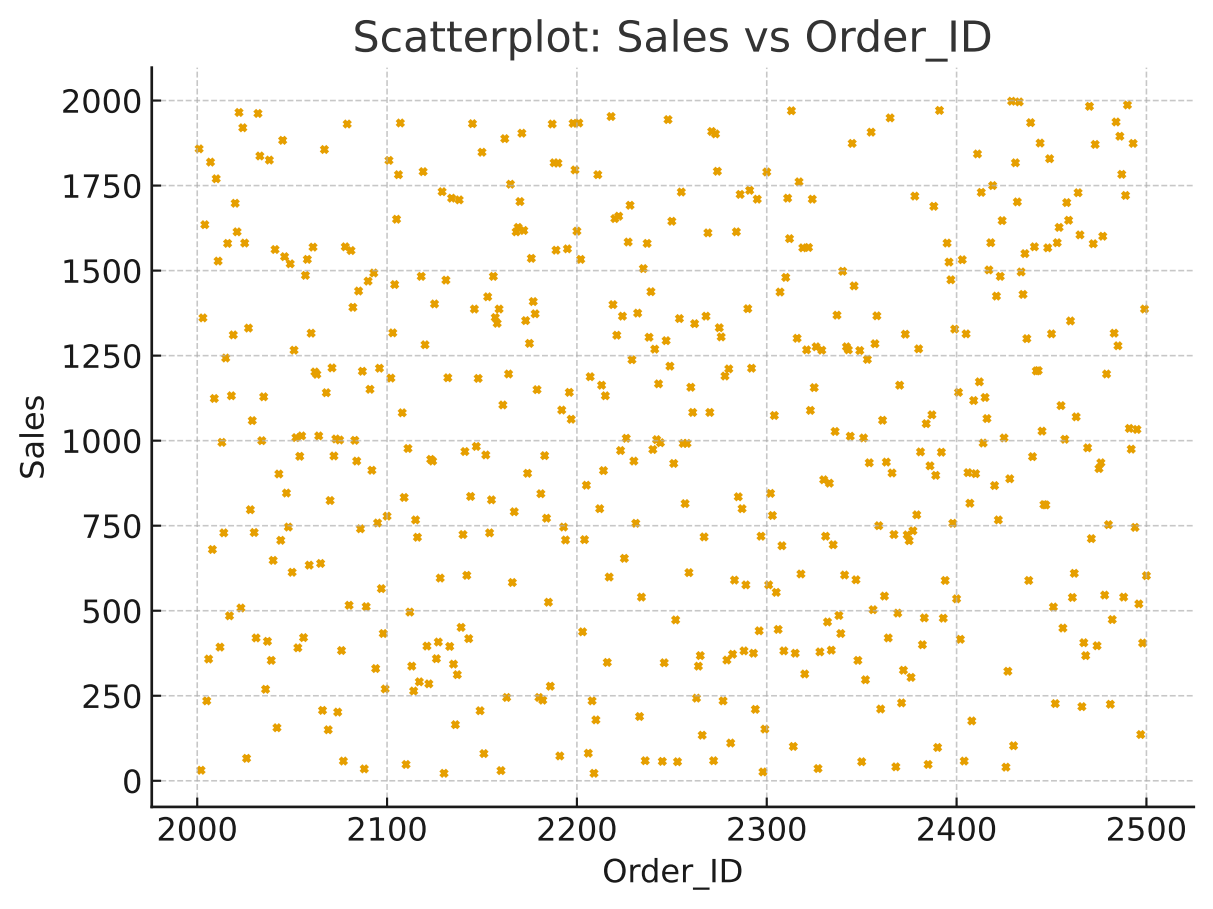

Histogram: Sales
Count
Sales
0
5
10
15
20
25
30
0
250
500
750
1000
1250
1500
1750
2000

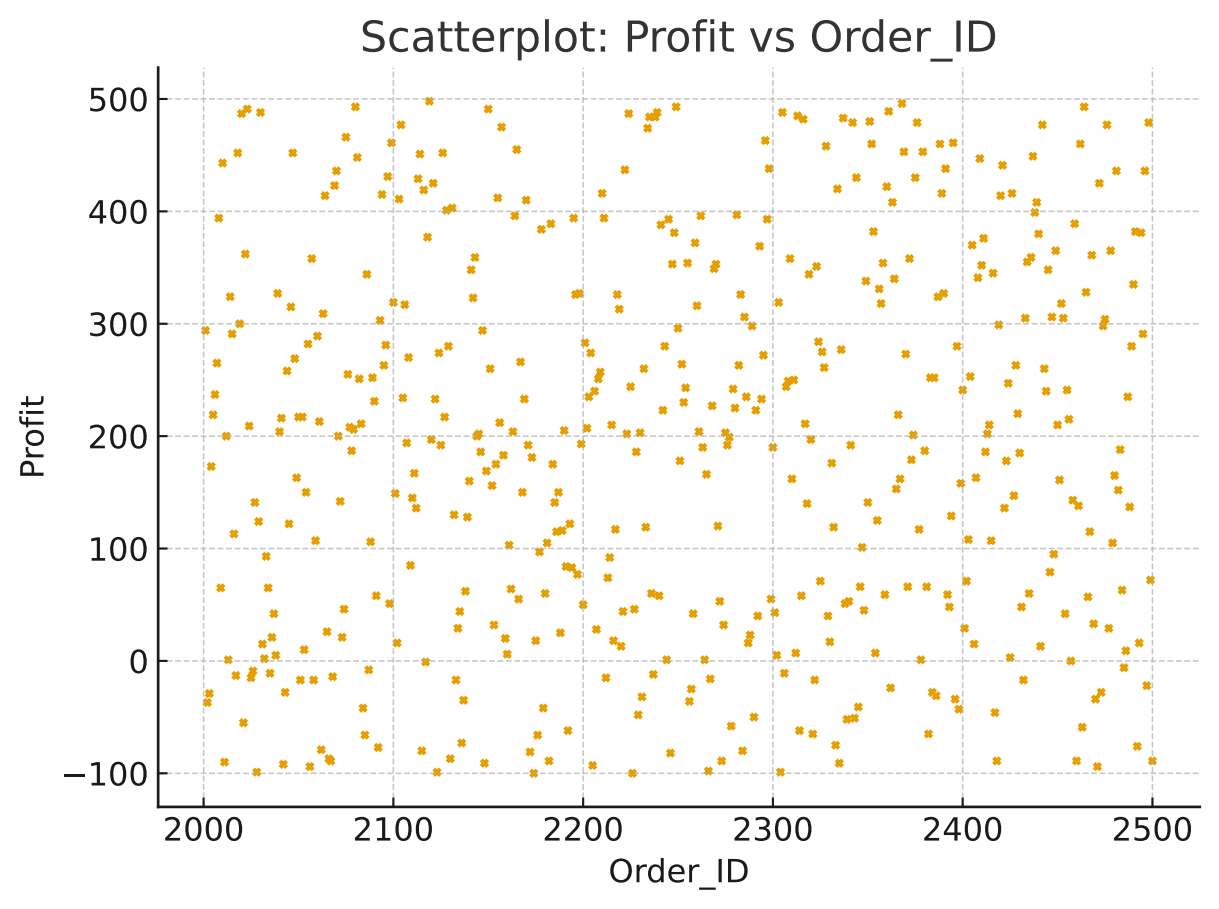

Histogram: Profit
Count
Profit
0
5
10
15
20
25
30
−100
0
100
200
300
400
500

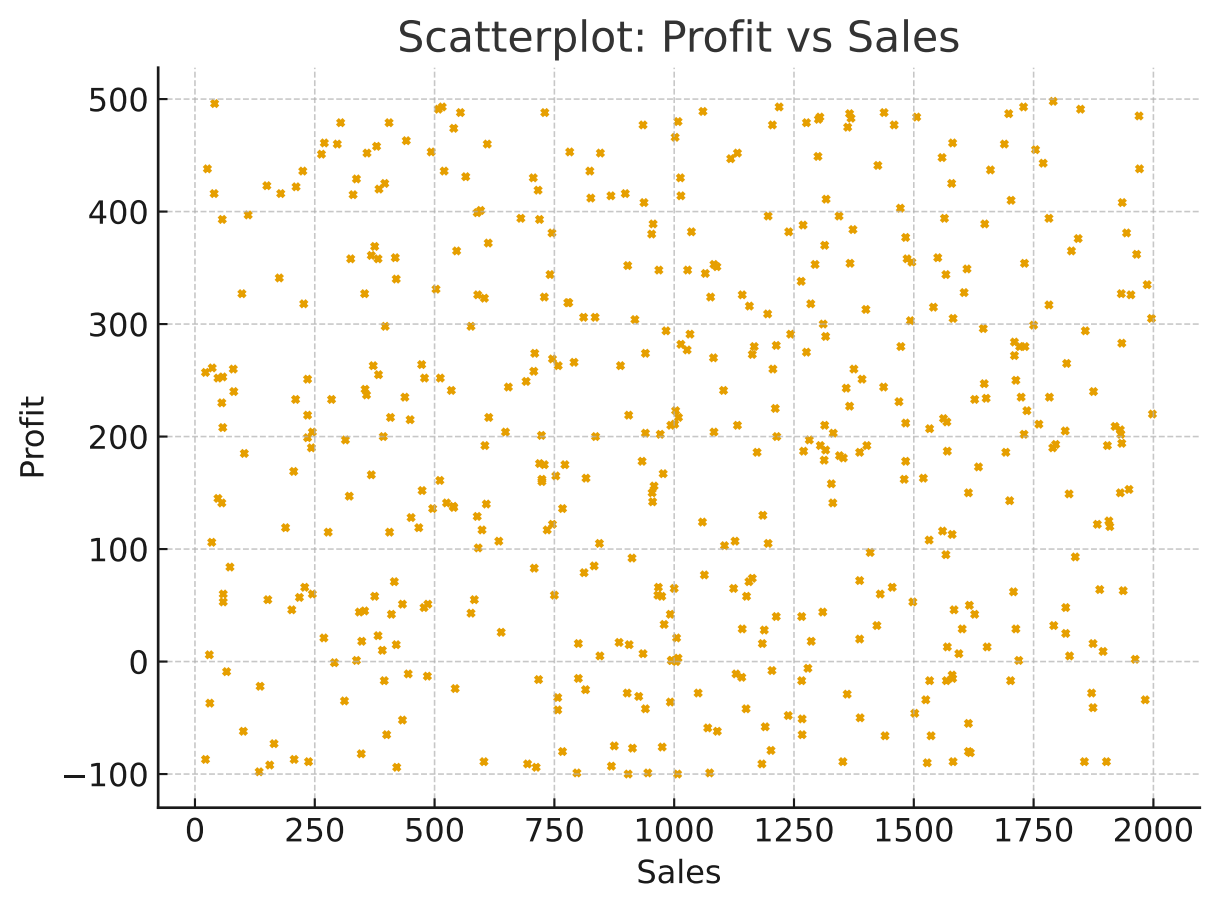

Boxplot: Order_ID
1
2000
2100
2200
2300
2400
2500


Boxplot: Sales
1
0
250
500
750
1000
1250
1500
1750
2000


Boxplot: Profit
1
−100
0
100
200
300
400
500


Scatterplot: Sales vs Order_ID
Sales
2000
1750
1500
1250
1000
750
500
250
0
2000
2100
2200
2300
2400
2500
Order_ID


Scatterplot: Profit vs Order_ID
Profit
500
400
300
200
100
0
−100
2000
2100
2200
2300
2400
2500
Order_ID


Scatterplot: Profit vs Sales
Profit
Sales
500
400
300
200
100
0
−100
0
250
500
750
1000
1250
1500
1750
2000

Automated Observations & Suggested Next Steps

Total rows: 500, total columns: 7
Top absolute correlations (pairs):
  Order_ID & Profit : 0.07
  Sales & Profit : 0.02
  Order_ID & Sales : 0.02
Suggested next steps: handle missing values (drop/impute), encode categorical variables, check and treat outliers, and build models for forecasting or classification depending on target variable.