Exploratory Data Analysis Report

1. Introduction

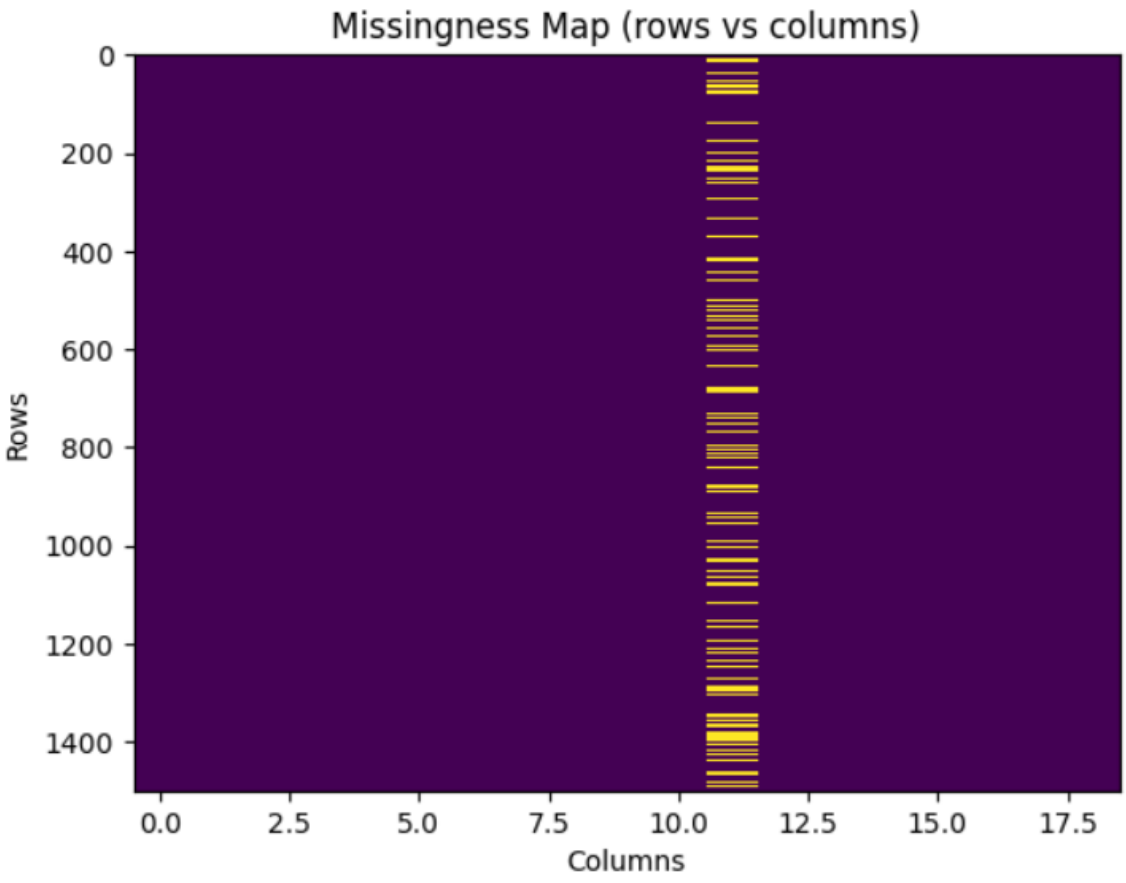
This report presents the findings of Exploratory Data Analysis (EDA) performed on the Product Retail Sales dataset. The objective of the analysis was to understand data distributions, detect anomalies, analyze correlations, and extract meaningful business insights through statistical summaries and visualizations.

2. Dataset Overview

- Number of rows: 1500
- Number of columns: 19
- Columns: Date, Region, Product, Quantity, UnitPrice, StoreLocation, CustomerType, Discount...

3. Missing Values

The dataset contains missing values across a column. Below is the Graph of missingness



4. Descriptive Statistics

- **Numeric columns summary:**

[2]:

	count	mean	std	min	25%	50%	75%	max
Quantity	1500.0	10.410667	5.735732	1.00	5.00000	11.000	15.0000	20.00
UnitPrice	1500.0	298.826947	169.100075	5.52	151.02000	294.740	446.7025	599.72
Discount	1500.0	0.073133	0.055979	0.00	0.00000	0.050	0.1000	0.15
TotalPrice	1500.0	2919.994952	2522.789977	6.97	867.13625	2174.724	4414.7235	11077.00
Returned	1500.0	0.248000	0.431996	0.00	0.00000	0.000	0.0000	1.00
ShippingCost	1500.0	27.507293	13.093453	5.01	16.70000	27.100	38.6350	49.98

- **Categorical columns summary:**

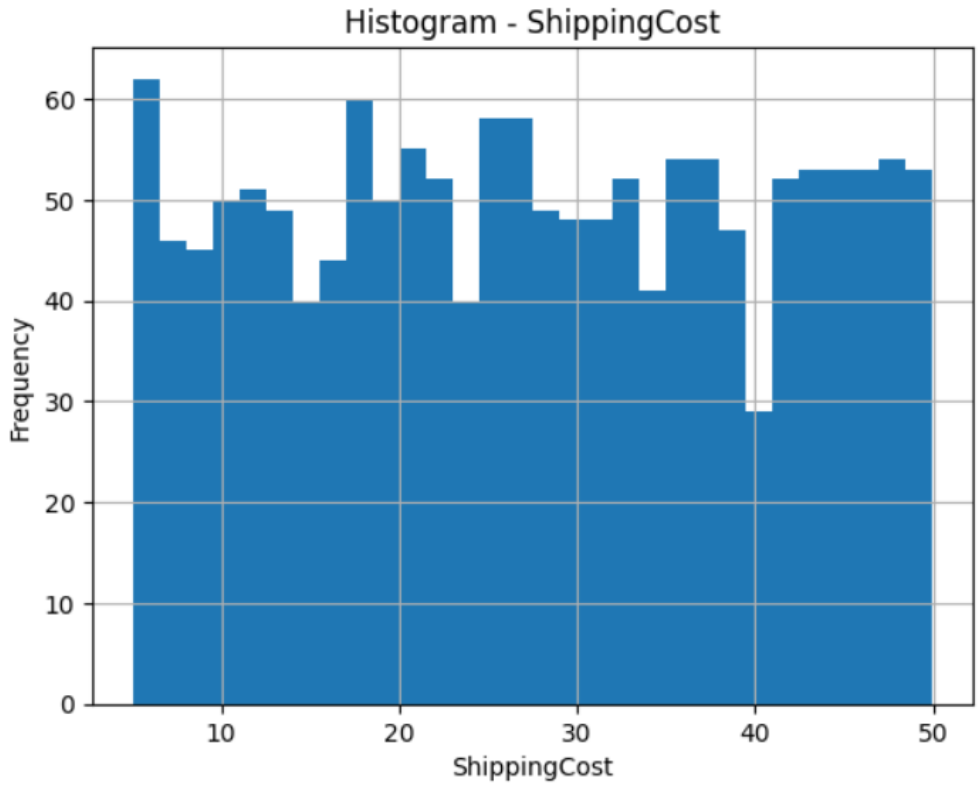
[3]:

	count	unique	top	freq
Date	1500	747	2023-08-06	7
Region	1500	5	East	311
Product	1500	7	Tablet	240
StoreLocation	1500	4	Store D	400
CustomerType	1500	2	Wholesale	756
Salesperson	1500	6	Carlos	270
PaymentMethod	1500	5	Online	323
Promotion	1130	3	FREESHIP	419
OrderID	1500	1500	REG100000	1
CustomerName	1500	1371	Cust 8150	4
OrderDate	1500	747	2023-08-06	7
DeliveryDate	1500	733	2025-05-21	6
RegionManager	1500	5	Eric	311

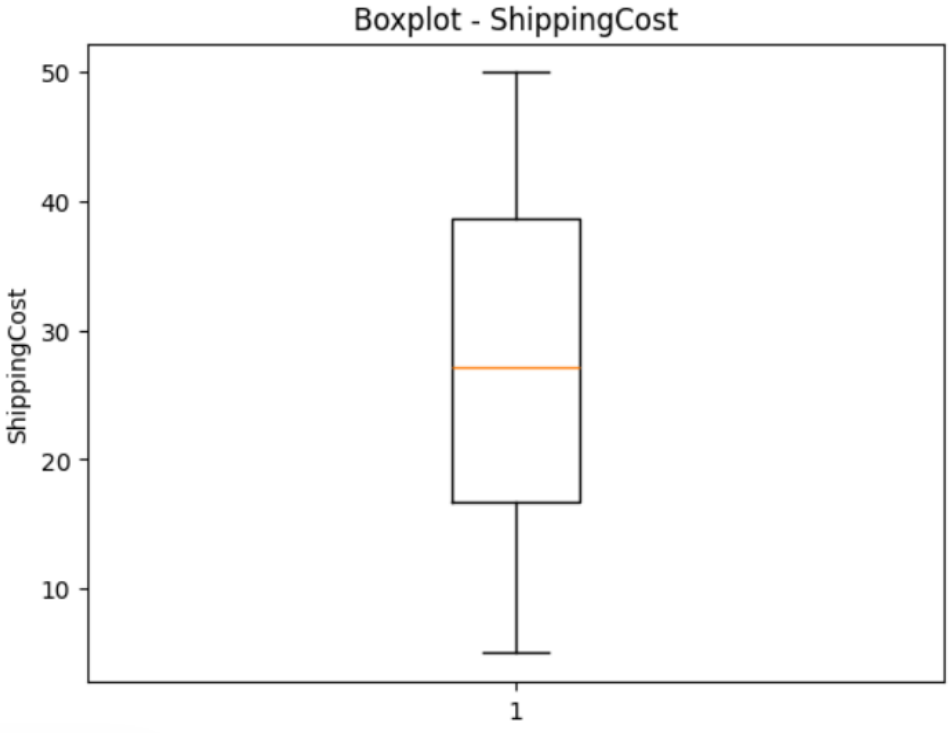
5. Key Charts

Below are some of the key visualizations generated during the analysis.

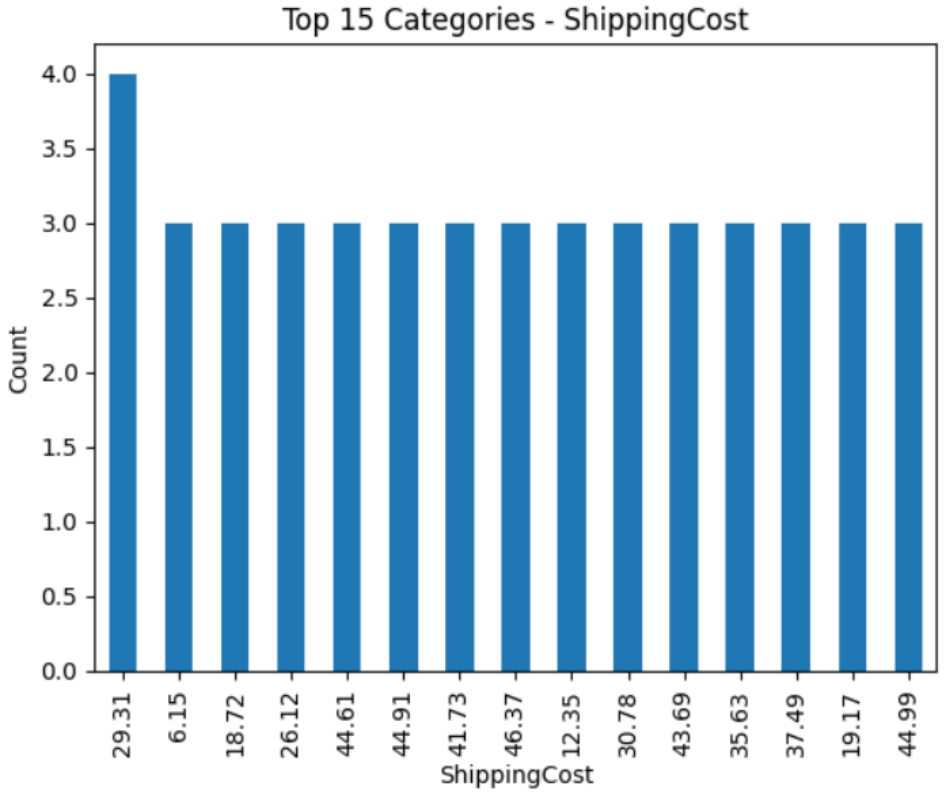
- **Histograms for numeric columns**



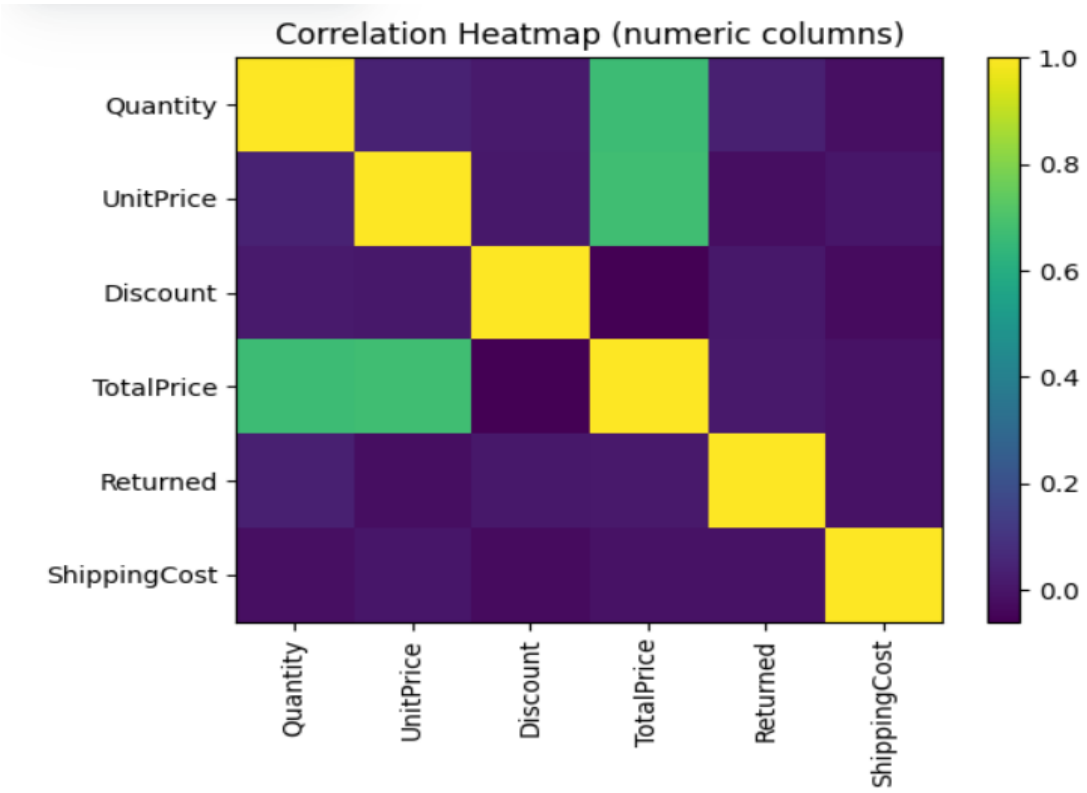
- **Boxplots for numeric columns**



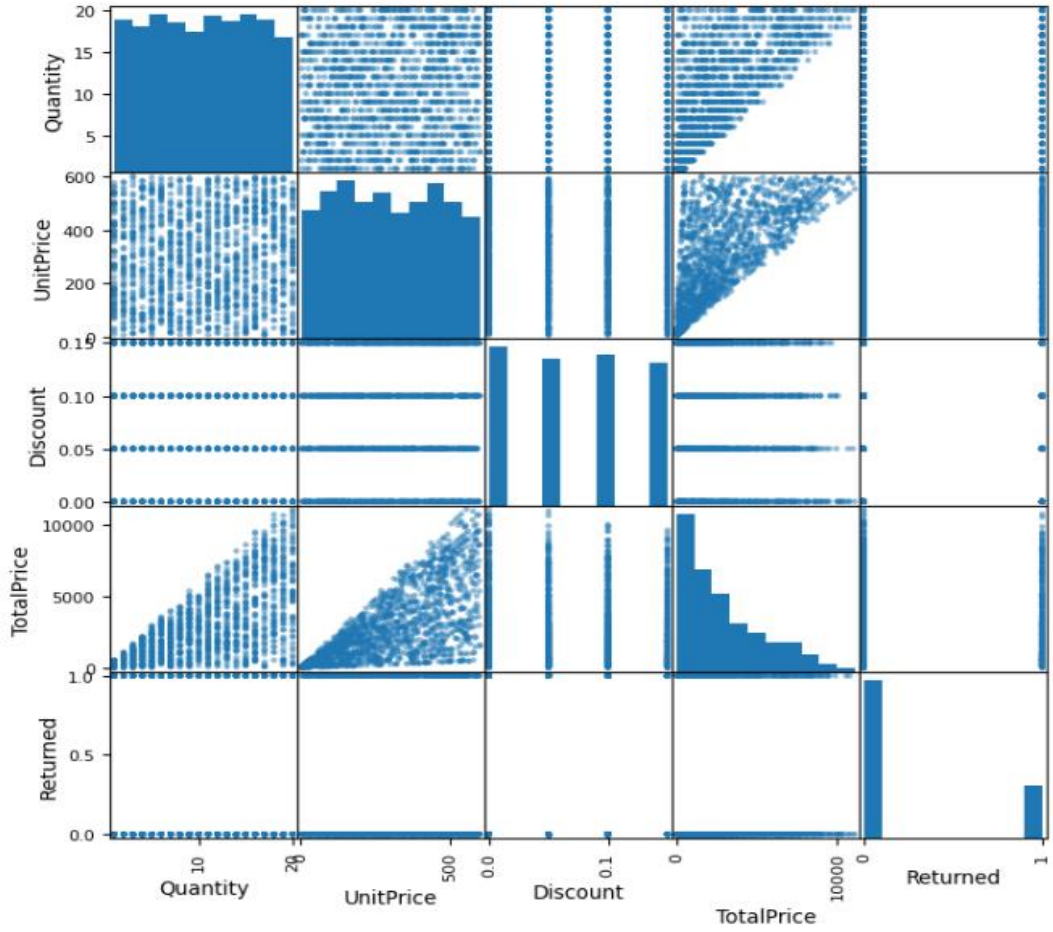
- **Bar plots for top categories**



- **Correlation heatmap**



- **Scatter matrix for first few numeric columns**



6. Insights

- The dataset contains both numeric and categorical variables, with some missing values requiring handling.
- Numeric distributions show skewness in some sales-related columns, suggesting the need for transformations.
- Outliers were detected in key numeric features, as observed in the boxplots.
- Correlation analysis revealed relationships between certain sales metrics, which can be used for predictive modeling.
- Category-level analysis showed significant variation in performance across different product groups.

7. Conclusion

The EDA provided valuable insights into the structure and characteristics of the Product Retail Sales dataset. The findings highlight the importance of data preprocessing (handling missing values and skewness) before modeling, and point towards potential features of interest (e.g., top-performing categories, correlated metrics) for deeper business analysis or predictive modeling.