EDA Report

1. Dataset Overview

- **Dataset Name**: `stores_sales_forecasting.csv`

2. Data Exploration

2.1 Missing Values

- **Missing Values Check**:
  - No missing values were found in the dataset.
  - All columns are complete.

2.2 Duplicates

- **Duplicates Check**:
  - No duplicate rows were found in the dataset.

2.3 Data Types

- **Numerical Columns**:
  - `Sales`, `Quantity`, `Discount`, `Profit`
- **Categorical Columns**:
  - `Ship Mode`, `Segment`, `Country`, `City`, `State`, `Region`, `Category`, `Sub-Category`, `Product Name`

3. Summary Statistics

3.1 Numerical Columns

Column	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Sales	9,994	229.86	623.25	0.44	17.28	54.49	209.94	22,638.48
Profit	9,994	28.66	234.26	-6,599.98	0.00	8.67	34.91	8,399.98
Quantity	9,994	3.79	2.23	1.00	2.00	3.00	5.00	14.00
Discount	9,994	0.16	0.21	0.00	0.00	0.00	0.20	0.80

3.2 Key Insights

4. Outlier Detection

4.1 Boxplots for Numerical Columns

- **Sales**:
  - Outliers are present in the upper range (very high sales values).
- **Profit**:
  - Outliers are present in both the lower (negative profits) and upper (very high profits) ranges.
- **Quantity**:
  - No significant outliers.
- **Discount**:
  - No significant outliers.

4.2 Outliers Removed

- Outliers were removed using the **IQR method**:
  - **Number of rows before removing outliers**: 9,994
  - **Number of rows after removing outliers**: 9,794
  - **Rows removed**: 200

5. Feature Engineering

5.1 New Features Created

- **Order Month**: Extracted from `Order Date` to analyze monthly trends.
- **Order Year**: Extracted from `Order Date` to analyze yearly trends.
- **Profit Margin**: Calculated as `Profit / Sales` to measure profitability.

5.2 Irrelevant Columns Dropped

- Columns like `Row ID`, `Order ID`, `Customer ID`, and `Product ID` were dropped as they are not useful for analysis.

6. Exploratory Data Analysis (EDA)

6.1 Univariate Analysis

- **Sales Distribution**:
  - Most sales are concentrated in the lower range (below $500).

    - The distribution is right-skewed, with a few high-value sales.
- **Profit Distribution**:
    - Most profits are concentrated around $0 to 0 to 50$.
    - The distribution is right-skewed, with a few high-profit outliers.
- **Quantity Distribution**:
    - Most orders have between 2 and 5 items.
    - The distribution is slightly right-skewed.

6.2 Bivariate Analysis

- **Sales vs Profit**:
    - There is a positive correlation between sales and profit.
    - High sales generally lead to higher profits, but some high sales result in losses (negative profit).
- **Sales by Region**:
    - The **West** region has the highest average sales, followed by the **East** and **Central** regions.
    - The **South** region has the lowest average sales.

6.3 Multivariate Analysis

- **Correlation Heatmap**:
    - `Sales` and `Profit` have a moderate positive correlation.
    - `Discount` has a weak negative correlation with `Profit`.

6.4 Time Series Analysis

- **Monthly Sales Trend**:
    - Sales peak in **November** and **December**, likely due to holiday shopping.
    - Sales are lowest in **January** and **February**.

7. Key Insights

1. **Sales and Profit**:
    a. High sales generally lead to higher profits, but some high sales result in losses due to discounts or other factors.
    b. Discounts have a weak negative impact on profit.
2. **Regional Performance**:
    a. The **West** region performs the best in terms of sales and profit.
    b. The **South** region underperforms compared to other regions.

3. **Seasonality**:
    a. Sales are highly seasonal, with peaks during the holiday season (November and December).
4. **Product Categories**:
    a. **Furniture** has the highest sales but the lowest profit margin.
    b. **Technology** has the highest profit margin.

8. Preprocessing Decisions

1. **Handling Missing Values**:
    a. No missing values were found, so no action was taken.
2. **Handling Outliers**:
    a. Outliers were removed using the IQR method to ensure data consistency.
3. **Feature Engineering**:
    a. New features like `Order Month`, `Order Year`, and `Profit Margin` were created to enhance analysis.
4. **Dropping Irrelevant Columns**:
    a. Columns like `Row ID`, `Order ID`, `Customer ID`, and `Product ID` were dropped as they are not useful for analysis.

9. Recommendations

1. **Focus on High-Profit Products**:
    a. Prioritize products with high profit margins, especially in the **Technology** category.
2. **Optimize Discounts**:
    a. Avoid excessive discounts, as they negatively impact profitability.
3. **Regional Strategies**:
    a. Invest more in the **West** region, which has the highest sales and profit.
    b. Investigate why the **South** region underperforms and implement targeted marketing strategies.
4. **Seasonal Promotions**:
    a. Capitalize on the holiday season (November and December) by running promotions and increasing inventory.

10. Visualizations

10.1 Sales Distribution

Most sales are concentrated in the lower range, with a few high-value outliers.

10.2 Profit Distribution

Most profits are around 0to0to50, with some high-profit outliers.

10.3 Sales vs Profit Scatterplot

Positive correlation between sales and profit, with some high sales resulting in losses.

10.4 Monthly Sales Trend

Sales peak in November and December due to holiday shopping.

11. Conclusion

This EDA provides a comprehensive understanding of the dataset, highlighting key trends, patterns, and areas for improvement. By leveraging these insights, the business can optimize its operations, improve profitability, and drive growth.