Data Analysis Report: Sales Dataset

Dataset Info

- Source: Kaggle - Sales Data (https://www.kaggle.com/datasets/kyanyoga/sample-sales-data)
- File Used: sales_data_sample.csv
- Rows & Columns: 2823 rows × 25 columns
- Objective: Clean, analyze, and visualize sales data to derive meaningful insights.

Step 1: Data Loading

```
df = pd.read_csv("sales_data_sample.csv", encoding='latin1')
```

Step 2: Data Cleaning

- Dropped rows with missing STATE values.
- Filled missing POSTALCODE values with median (after converting to numeric).

```
df = df.dropna(subset=['STATE'])
df['POSTALCODE'] = pd.to_numeric(df['POSTALCODE'], errors='coerce')
postal_median = df['POSTALCODE'].median()
df['POSTALCODE'] = df['POSTALCODE'].fillna(postal_median)
```

Step 3: Basic Analysis

- Basic description: df.describe()
- Total sales: df['SALES'].sum()
- Year-wise sales: df.groupby("YEAR_ID")["SALES"].sum()

Data Analysis Report: Sales Dataset

Step 4: Visualizations

1. Monthly Sales Distribution:

```
sns.boxplot(x='MONTH_ID', y='SALES', data=df)
```

2. Product Line Sales:

```
df.groupby('PRODUCTLINE')['SALES'].sum().sort_values(ascending=False).plot(kind='barh')
```

3. Heatmap of Correlations:

```
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm')
```

Bonus Operations

A. Profit Column (fixed cost per unit = 70):

```
df['PROFIT'] = df['SALES'] - (df['QUANTITYORDERED'] * 70)
```

B. Top 5 Customers by Profit:

```
df.groupby('CUSTOMERNAME')['PROFIT'].sum().sort_values(ascending=False).head(5)
```

C. Pairplot:

```
sns.pairplot(df[['QUANTITYORDERED', 'PRICEEACH', 'SALES', 'MSRP', 'PROFIT']])
```

Final Output

- Cleaned CSV: cleaned_sales_data.csv
- All visuals and analysis performed using Python (Pandas, NumPy, Matplotlib, Seaborn).