Sales Data Cleaning & Preprocessing Report

Dataset Overview

The dataset initially contained 185,950 rows and 11 columns, which captured customer purchase details such as order information, product details, pricing, and location.
The columns included:

- Order ID – Unique identifier for each order
- Product – Name of the product purchased
- Quantity Ordered – Number of units purchased
- Price Each – Price of a single unit
- Order Date – Date and time of purchase
- Purchase Address – Customer shipping address
- Month – Extracted month of purchase
- Sales – Total sales value per order (Quantity × Price)
- City – Extracted city from purchase address
- Hour – Extracted purchase hour from the order timestamp

We dropped the unnecessary column `Unnamed: 0`, which was only an index placeholder.

Descriptive Statistics & Distribution

Column	Mean	Std Dev	Min	25%	50%	75%	Max
Quantity Ordered	1.12	0.44	1	1	1	1	9
Price Each ($)	184.39	332.73	2.99	11.95	14.95	150	1700
Sales ($)	185.49	332.91	2.99	11.95	14.95	150	3400
Month	7.05	3.50	1	4	7	10	12
Hour (24hr format)	14.41	5.42	0	11	15	19	23

Key Findings

1. Product Quantity
    - Most purchases were single-unit orders (median = 1), with very few bulk orders (max = 9).
2. Pricing & Sales
    - The median price per item is $14.95, while some premium products go up to $1700.
    - The median sales per order is also $14.95, indicating that many orders consist of just one low-cost item.
    - A few high-value transactions significantly increase the average sales to $185.
3. Seasonality (Month-wise)
    - Data spans across 12 months (January–December).
    - Average purchase month = July, suggesting mid-year sales peaks.
4. Customer Buying Behavior (Hourly)
    - Purchases occur throughout the day (00:00–23:00).
    - The median purchase time is around 3 PM, with high activity between 11 AM – 7 PM.

Data Cleaning & Transformation

1. Handling Missing Values
    - Verified all key columns (Order ID, Product, Sales, etc.) had 0 missing values.
2. Removing Duplicates
    - Dropped duplicate entries based on a combination of Order ID, Product, Quantity Ordered, Order Date, and Purchase Address.
3. String Normalization
    - Stripped extra whitespaces from categorical columns: Product, Purchase Address, and City.
4. Feature Engineering
    - Extracted State from the `Purchase Address` column.
    - Derived Time and Day Name from the `Order Date` field.
    - Converted `Order Date` to proper datetime format for time-series analysis.
    - Added Calculated Sales (Quantity × Price) for validation against the given Sales column.
5. Final Dataset
    - After cleaning, the dataset contains 185,686 rows and 15 columns.
    - Exported cleaned dataset as `cleaned_data_sales.csv` for further analysis.

Conclusion

This cleaned dataset provides a reliable foundation for sales trend analysis, customer behavior insights, and predictive modeling.
The preprocessing ensures:

- No duplicates
- Standardized data formats
- Additional derived features (Month, Day Name, Hour, State) that enable deeper business intelligence and visualization.