EDA Report
Milestone 1

INTRODUCTION & OBJECTIVES

This report details the exploratory data analysis (EDA) and preprocessing steps performed on the "Original Data.csv" dataset. The primary objective of this milestone was to clean, analyze, and enrich the data to uncover key business insights and prepare a robust dataset for building sales and demand forecasting models. This process involved handling missing values, managing outliers, and extensive feature engineering.

DATA OVERVIEW & PREPROCESSING

The dataset was loaded and systematically cleaned to ensure data quality and readiness for analysis and modeling.

- Original Dataset: 51,931 rows and 24 columns.

Initial Cleaning

- Dropped 5 non-essential columns: Row ID, Order ID, Customer ID, Postal Code, and Product ID.
- Removed 173 duplicate rows from the dataset.
- Removed 468 rows containing 4 or more null values, as they were unsuitable for imputation.

Missing Values Handled

- Profit: 86 missing values were imputed using the column mean.
- Order Priority: 830 missing values were filled using the column mode ("Medium").
- Market: 672 missing values were imputed using the most frequent (mode) market value from their corresponding Region.

Outlier Management

- Boxplots identified a significant number of outliers in Sales, Discount, Profit, and Shipping Cost.
- These extreme values were managed by replacing them with the median (for Profit and Discount) or mean (for Sales and Shipping Cost) to normalize the data for forecasting.

Final Dataset:

51,290 rows (after cleaning) and 31 columns (after feature engineering and preprocessing)

FEATURE ENGINEERING

Feature	What It Is & Why We Added It
Total Revenue (Total_sales)	Calculated the *true* order value (Price x Quantity). The original Sales column was just the price of one item.
Time Features	Broke down the order date into useful parts like Month, Year, Day_of_Week, and a weekend flag (Is_Weekend).
Smart Season (Season)	Correctly identified the Season (e.g., Summer/Winter) for all 147 countries, since "July" is Summer in the US but Winter in Australia.
Holiday Flag (Is_Holiday)	Added a flag (1 or 0) to show if an order was placed on a public holiday in that specific country.
Promotion Flags	Added simple yes/no flags (1 or 0) to show if an item had *any* discount or an *above-average* discount.
Lag Features (for Forecasting)	Added the sales data from 1 day, 7 days, and 30 days ago. This helps the model understand recent trends.

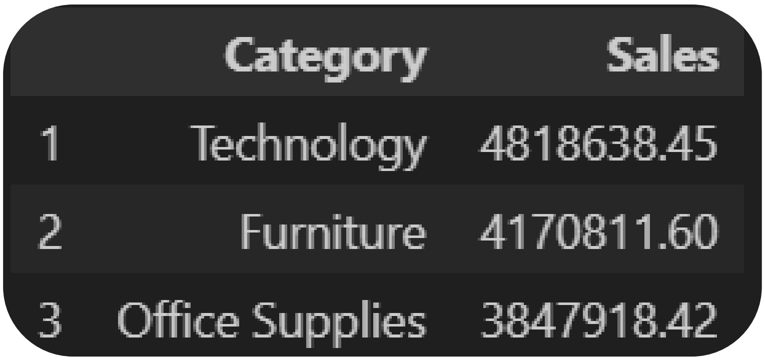
DATA EXPLORATION INSIGHTS

Total Sales
$12.84 Million

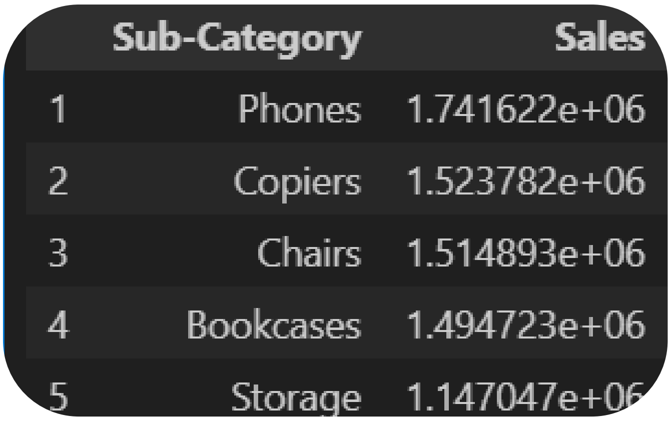
Total Profit
$1.49 Million

Total discounts
$7375.12

Average Sales per Order
$247.20



	Category	Sales
1	Technology	4818638.45
2	Furniture	4170811.60
3	Office Supplies	3847918.42



	Sub-Category	Sales
1	Phones	1.741622e+06
2	Copiers	1.523782e+06
3	Chairs	1.514893e+06
4	Bookcases	1.494723e+06
5	Storage	1.147047e+06

Financial & Discount Analysis

The Discount-Profit Crisis: This is the most critical financial insight. The data shows a direct and severe negative correlation between discounts and profit.

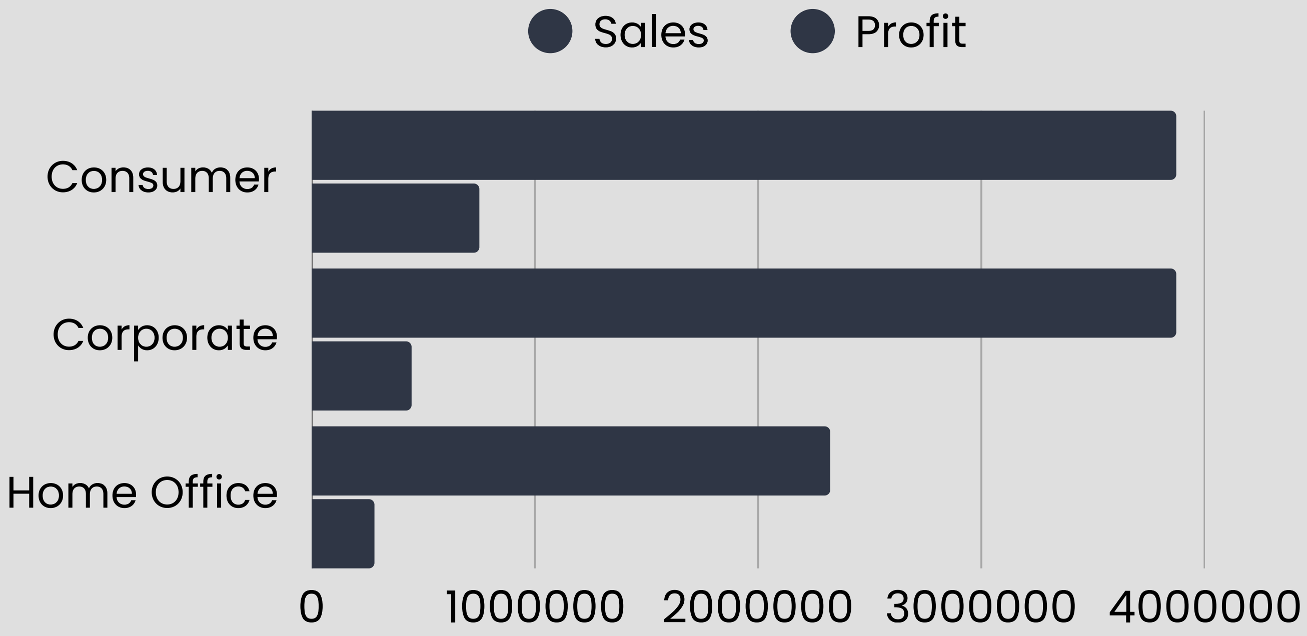
- Average Profit (No Discount): $61.19
- Average Profit (Any Discount): -$13.34 (a net loss)

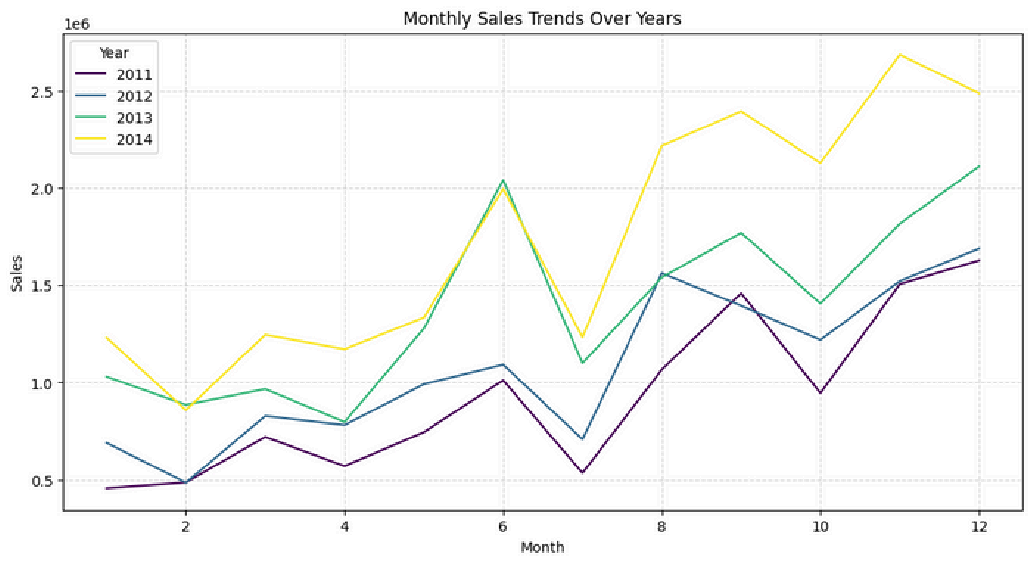
The "Furniture Problem": High sales do not equal high profit.

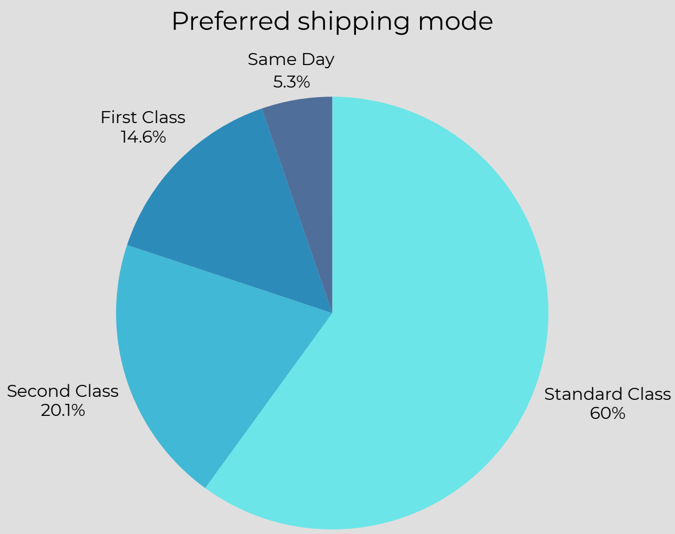
- Furniture is the #2 highest-selling category (based on unit price) but is the #1 least-profitable category, often generating significant losses.
- This is driven by sub-categories like Tables and Bookcases, which are among the top 5 least profitable products.

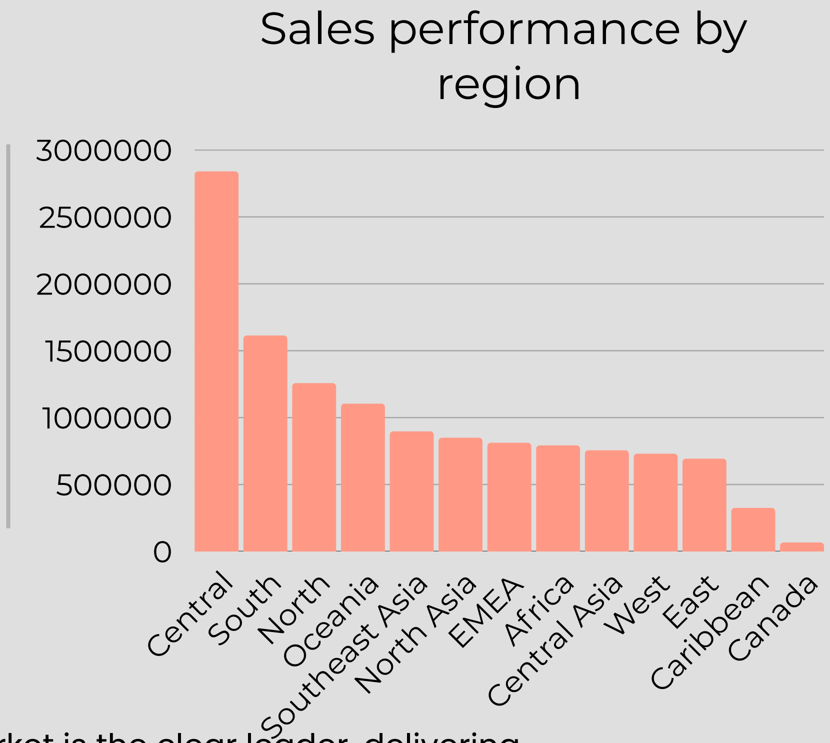
The Profit Winner: Technology is the most profitable category, driven by high-profit items like Copiers and Phones.

Sales, and Profit by Segment









Geographic & Market Analysis

Top Market: The APAC (Asia-Pacific) market is the clear leader, delivering both the highest total sales and the highest total profit.

Top Region: The Central region is the single most valuable region, outperforming all others in both sales ($2.8M) and profit ($312K).

Top Cities: The top 5 cities by sales are New York City ($264k), Los Angeles ($179k), Manila ($122k), Seattle ($120k), and San Francisco ($114k).

Customer & Seasonal Trends

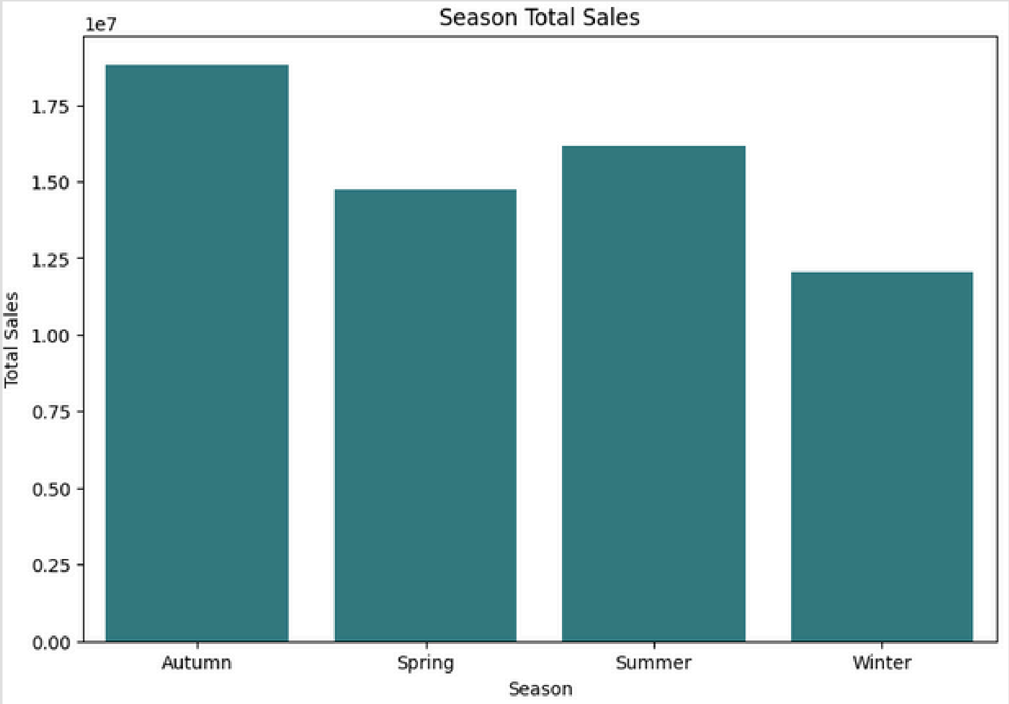
Top Customer Segment: The Consumer segment is the most important, driving the majority of revenue (over 51% of sales) and the most profit ($748K).

Overall Sales Trend: The business shows a strong and consistent upward trend in sales from 2011 to 2014.

Peak Season: Autumn is the highest-selling season.

Peak Months: Sales peak in December, August, and June.

Holiday/Weekend Effect: Analysis showed no significant negative impact on sales or profit during holidays or weekends; purchasing is consistent.



CONCLUSIONS & NEXT STEPS

Conclusion: The data has been successfully cleaned, processed, and enriched. The analysis reveals that while sales are growing, profitability is heavily undermined by discount strategies, particularly in the Furniture category.

Next Steps: The resulting dataset, Data After Milestone 1.csv, is now prepared for advanced modeling. The engineered features (especially Lag Features, Promotion Flags, and Seasonal tags) will be used to build and train sales and demand forecasting models.