# SUPERSTORE SALES REPORT

## PYTHON PROJECT ON DATA ANALYSIS

*jasikagupta04*

*Jasika Gupta*

*jasikagupta04@gmail.com*

# PROBLEM STATEMENT

The dataset contains information about customer orders, including order dates, shipping dates, product categories, sub-categories, sales amounts, and regional distribution. The goal is to analyze the data to uncover insights and improve business performance.

Specifically, the analysis aims to identify the regions, categories, and sub-categories contributing the sales, examine trends in sales performance over time, and assess the impact of shipping modes on delivery efficiency.  Based on these insights, the goal is to recommend strategies to boost sales in underperforming areas, optimize shipping processes to improve customer satisfaction, and enhance overall operational efficiency.

# OBJECTIVE

To use exploratory data analysis (EDA) and data visualization to identify key trends, patterns, and anomalies in the dataset, leading to actionable insights that enhance overall business performance.

# DATA GATHERING

The dataset represents sales data collected from a superstore chain operating across various states and regions in the United States. Each transaction is recorded with details like the order date, shipping date, sales amount, product category, sub-category, and customer location, including city, state, and region.

For instance, the dataset categorizes regions as East, West, Central, and South, while also providing granular insights at the city level. Online and in-store sales data are consolidated into a centralized database, ensuring a comprehensive view of operations. Data preparation involves standardizing date formats, handling missing or incomplete fields, and aggregating data for analysis at different levels, such as region or product category. This data collection process enables the superstore to track sales trends, evaluate the efficiency of shipping modes, and understand customer behavior across regions, ultimately helping to optimize business strategies and improve overall performance.

# DATA PREPARATATION

- Data Cleaning:
  - Null values or duplicate data are identified and handled appropriately.
  - For Example, ‘Postal Code’ had several null values, which I replaced with ‘Unknown’.
  - Unnecessary columns like 'Order ID', 'Customer ID', 'Customer Name', 'Postal Code', 'Product ID' & 'Product Name' were removed.
- Data Transformation:
  - Date columns ‘Order Date’ and ‘Ship Date’ are changed into date format from object format.
  - Day, month and year is being extracted from the ‘Order Date’ for further analysis.

# DESCRIPTIVE ANALYSIS

| Region | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Central | 2277.0 | 216.357889 | 636.040148 | 0.444 | 14.560 | 45.920 | 201.9600 | 17499.950 |
| East | 2785.0 | 240.401697 | 626.366105 | 0.852 | 17.712 | 54.960 | 211.9600 | 11199.968 |
| South | 1598.0 | 243.524067 | 779.850548 | 1.167 | 17.088 | 54.114 | 209.9475 | 22638.480 |
| West | 3140.0 | 226.184613 | 524.240789 | 0.990 | 19.440 | 61.002 | 215.6065 | 13999.960 |

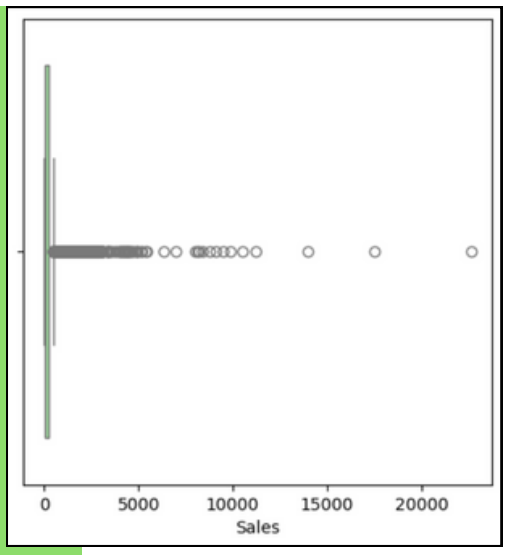
## Sales by Region

- West has the highest number of sales transactions (3140), while South has the lowest (1598).
- The South region has the highest average sales (243.52) per transaction while the Central region has the lowest with 216.36.
- The South region has the highest sales variation (779.85), meaning sales amounts fluctuate significantly.
- The percentiles (25%, 50%, 75%) show the distribution of sales values.
  - The West region consistently has higher values, meaning sales transactions tend to be larger.
  - The Central region has the lowest 25%, 50%, and 75% values, indicating lower sales transactions.

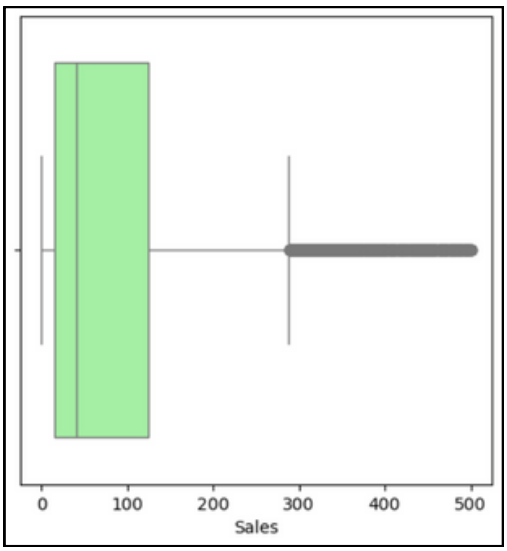
| Category | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Furniture | 2078.0 | 350.653790 | 501.489219 | 1.892 | 47.19 | 182.610 | 435.237 | 4416.174 |
| Office Supplies | 5909.0 | 119.381001 | 383.761427 | 0.444 | 11.76 | 27.360 | 79.470 | 9892.740 |
| Technology | 1813.0 | 456.401474 | 1116.818701 | 0.990 | 67.98 | 167.944 | 453.576 | 22638.480 |

## Sales by Category

- The Office Supplies category has the highest number of transactions (5909) but the lowest average sales (119.38).
- The Technology category has the highest average sales per transaction (456.40) and the highest maximum sale (22,638.48). It also have the highest variation (1116.81), meaning there are frequent high-value sales.
- Furniture has an average sale of 350.65, higher than Office Supplies but lower than Technology. It has a median (50%) sales value of 182.61, meaning half the sales are below this amount.



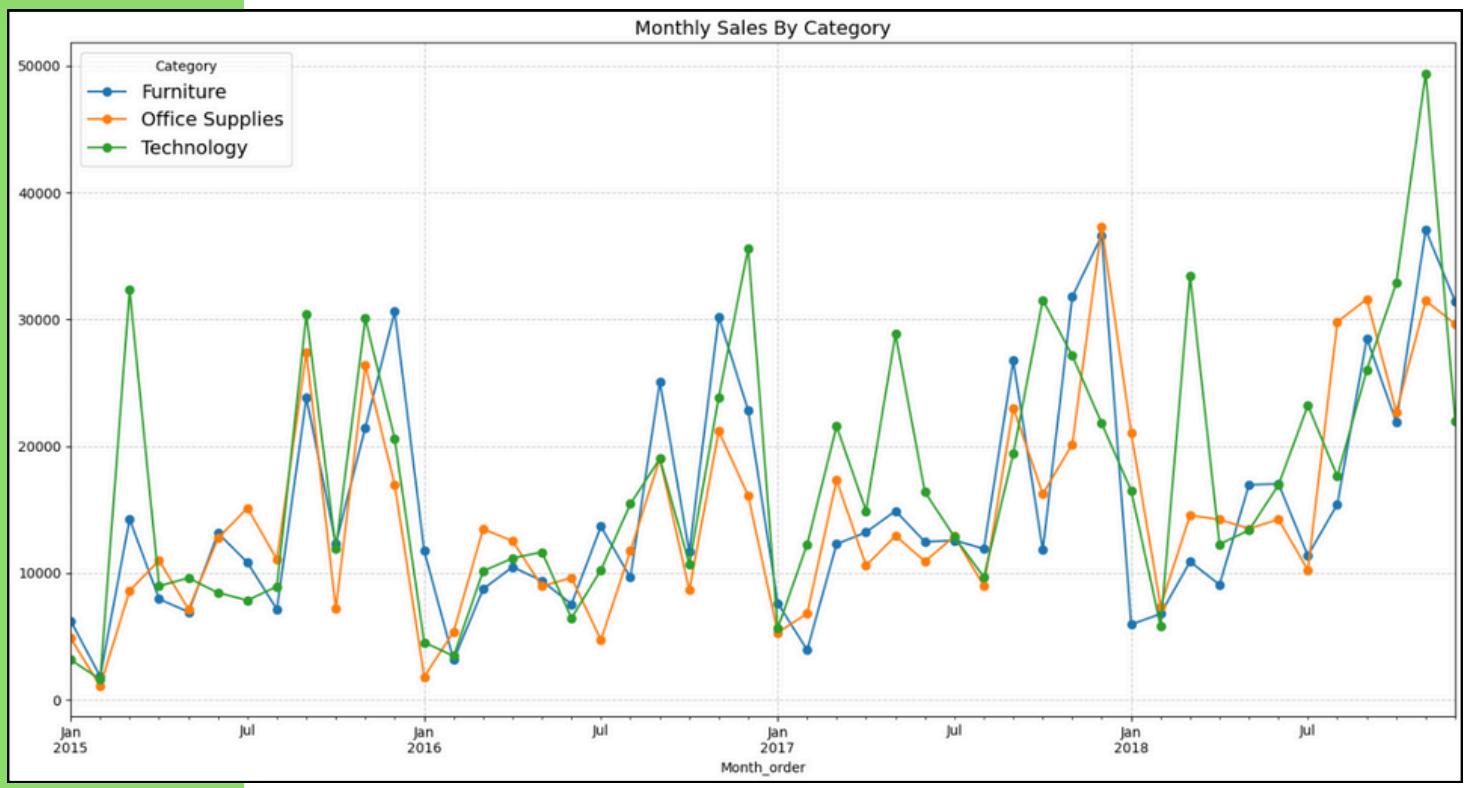

**Boxplot of Sales**




**Boxplot of Sales after removing the outliers**

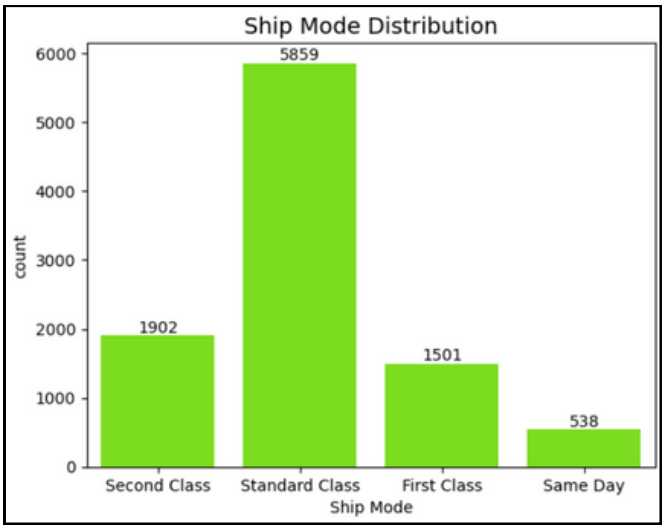
The sales data includes extreme outliers that significantly deviate from the bulk of the data points. The majority of the data is concentrated within a small range near the lower end, but there are several values extending up to and beyond 20,000. The presence of extreme outliers stretches the scale of the plot, making it difficult to analyze the distribution of the main body of data effectively.

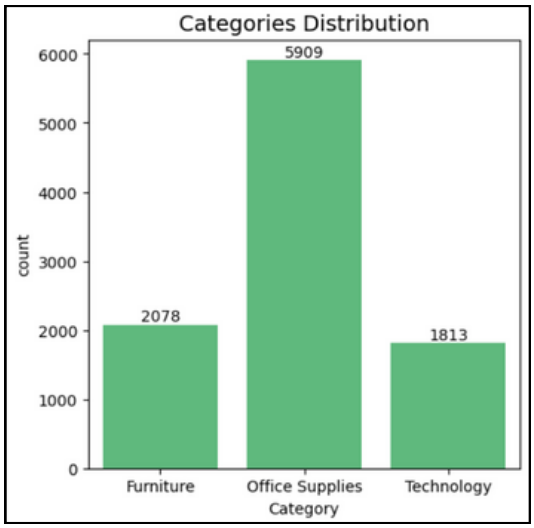
After removing the outliers, the boxplot provides a clearer representation of the central distribution. The interquartile range (IQR) is now visible, and the bulk of the sales data is concentrated between approximately 0 and 100.

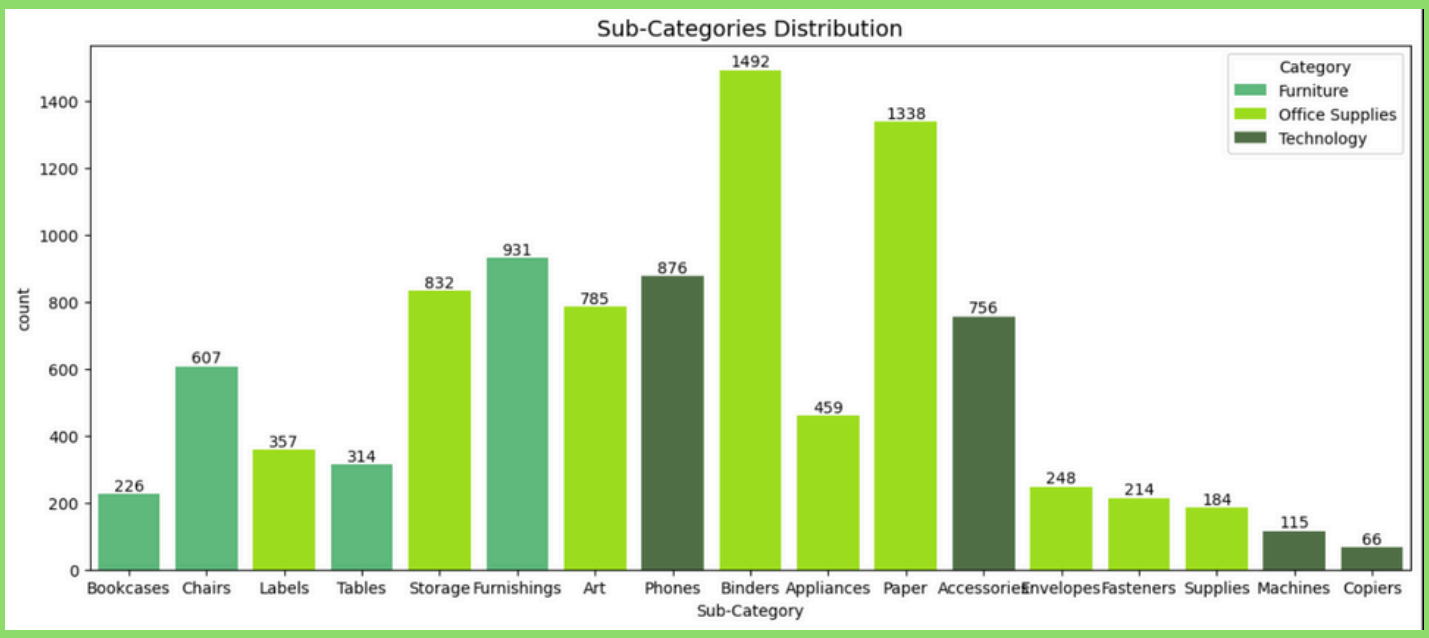


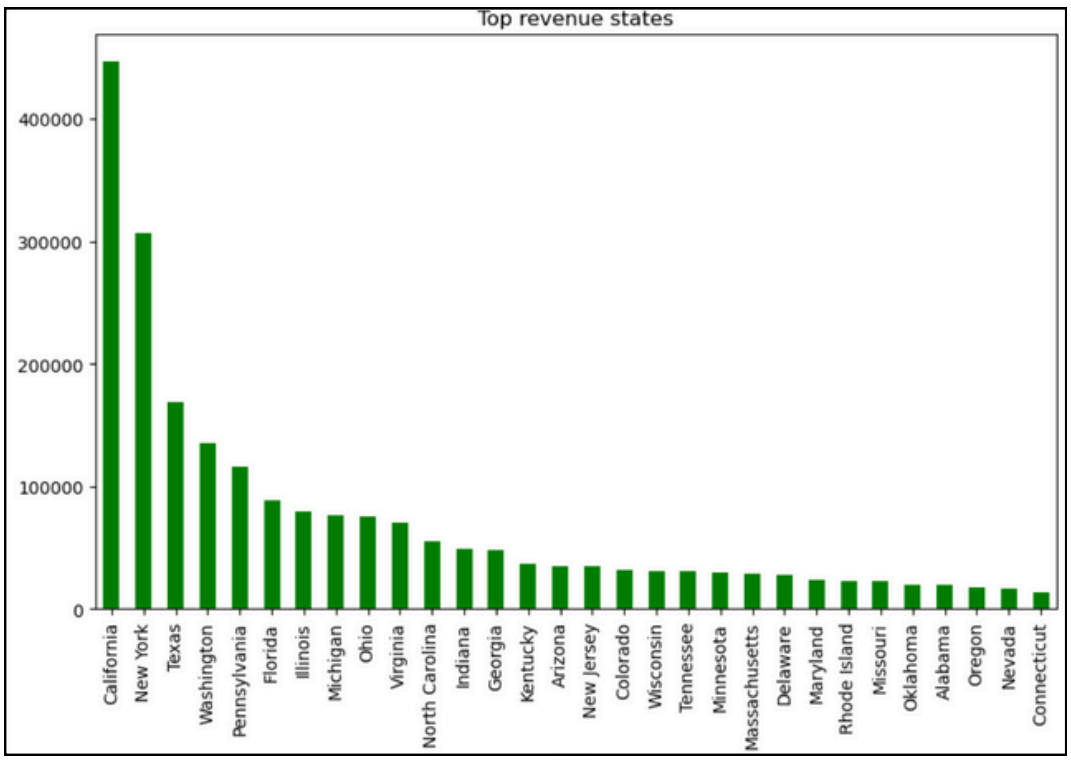

**Monthly Sales Distribution based on Product Category**

# DISTRIBUTION GRAPHS



# HEATMAP VISUALIZATION

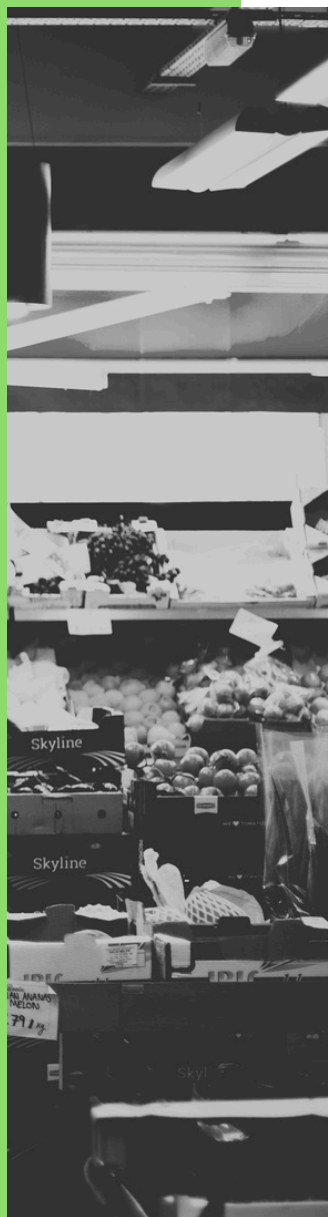

From the heatmap, it is evident that there are no strong correlations among the variables. The Sales variable exhibits negligible correlation with Order Day, Order Month, and Order Year, implying that sales figures do not significantly fluctuate based on the date of order. Similarly, Row ID, which is likely a unique identifier, shows no meaningful correlation with any other variable, as expected.

The negative correlation between Order Month and Order Day (-0.032) and Order Year (-0.018) is extremely weak, suggesting no clear trend in how sales behave over time.

Overall, the weak correlations indicate that none of the recorded variables strongly influence each other, emphasizing the need for further exploratory data analysis (EDA) to uncover potential patterns through other categorical or derived insights.

# LINEAR REGRESSION MODEL

```
x = data[['Days to Ship', 'Order Day', 'Order Month', 'Order Year']]
y = data[['Sales']]

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.25, random_state = 42)

model = LinearRegression()

model.fit(x_train, y_train)

LinearRegression
LinearRegression()

y_pred = model.predict(x_test)

mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f'Mean Squared Error: {mse}')
print(f'R-squared: {r2}')

Mean Squared Error: 651271.7595041806
R-squared: -0.004505352140149466
```




A linear regression model was trained to predict sales based on the features Days to Ship, Order Day, Order Month, and Order Year. The dataset was split into training and testing sets, with 25% of the data used for testing.

**Results Analysis:**

- **Mean Squared Error (MSE): 651271.76**
  - A high MSE value suggests that the model's predictions deviate significantly from the actual sales values.
- **R-squared ($R^2$) Score: -0.004**
  - The $R^2$ value is close to 0, indicating that the independent variables used in the model do not explain the variance in sales. A negative $R^2$ score implies that the model performs worse than a simple mean-based prediction.

The model demonstrates poor predictive performance, likely due to the weak correlation between the selected independent variables and sales. This suggests that factors such as product category, discount, customer demographics, or promotional strategies may have a greater influence on sales than the chosen time-based features.

# KEY OBSERVATIONS:

*"Data-driven insights pave the way for smarter decisions, but the right features unlock true predictive power."*

The analysis of superstore sales data provided key insights into sales trends across different regions and product categories. Descriptive statistics revealed that sales performance varies significantly, with categories like Technology having the highest average sales and Office Supplies having the lowest. Regional sales distribution also highlighted variations, with the South region recording the highest maximum sales value.

To further explore sales patterns, a linear regression model was implemented using time-based features (Days to Ship, Order Day, Order Month, and Order Year) to predict sales. However, the model's performance was poor, as indicated by a high Mean Squared Error (MSE) and a negative R-squared ($R^2$) score. This suggests that the chosen variables do not sufficiently explain variations in sales, indicating the need for better feature selection.

# THANK YOU!

*jasikagupta04* *jasikagupta04@gmail.com* *Jasika Gupta*