PROJECT REPORT ON

GLOBAL SUPERSTORE DATA

Course: BDA100NAA
Instructor: Omid Aghababaei Tafreshi
Prepared by: Rishi Raj Das, Luka Lazarev, Duy Bao Khang Nguyen
Student ID: 165212234, 143871242, 185000239

Descriptive Statistics

1.1 Dataset

1.1.1 Introduction

This project utilizes a retail dataset from a global superstore, covering four years of transaction data. The dataset includes detailed records of customer purchases, shipping details, and product information. Through this data, the project aims to uncover key business insights related to customer behavior, product performance, and regional sales trends.

To achieve this, three core data analysis techniques were applied:

1. Descriptive Statistics to summarize and visualize data distributions.
2. Linear Regression to model the relationship between sales and profit.
3. Time Series Analysis to forecast future sales trends.

The goal is to support decision-making in marketing, inventory planning, and sales strategies, while also working to improve profit margins through data-driven insights.

1.1.2 Dataset Characteristics

1. Number of instances (rows): 9,800
2. Number of variables (columns): 18

- Numerical variables: Row ID, Postal Code, Sales, Profit.
- Categorical variables: Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name, Segment, Country, City, State, Region, Product ID, Category, Sub-Category, Product Name.

1.2 Categorical Variables

1.2.1 Frequency Tables

We analyzed categorical variables from the retail dataset by constructing frequency tables for key fields. These tables include frequency, relative frequency, cumulative frequency, and cumulative relative frequency.

Below are the results and interpretations for two categorical variables: Category and Region.

Category	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
Office Supplies	5909	0.602959184	5909	0.602959184
Furniture	2078	0.212040816	7987	0.815
Technology	1813	0.185	9800	1

Category	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
West	3140	0.320408163	3140	0.320408163
East	2785	0.284183673	5925	0.604591837
Central	2277	0.232346939	8202	0.836938776
South	1598	0.163061224	9800	1

Product Category Sales

1. Office Supplies account for most sales, contributing approximately 60.3% of total transactions.
2. Furniture represents the second largest category, with 21.2% of total sales.
3. Technology holds the smallest share, contributing 18.5% of sales. Combined, Office Supplies and Furniture account for over 80% of all product category sales, highlighting their significance in the company's revenue stream.

Regional Sales Distribution

1. The West region leads in sales performance, making up 32% of all recorded sales.
2. The East region follows closely, contributing 28.4% of sales.
3. The Central region accounts for 23.2%, indicating moderate performance.
4. The South region contributes the least, with 16.3% of total sales.
5. Together, the West and East regions contribute more than 60%, making them crucial markets for focused business strategy and resource allocation.

1.2.2 Pivot Tables

This Pivot Table contains the top cities by their sales volume:

City	Count of City
New York City	891
Los Angeles	728
Philadelphia	532
San Francisco	500
Seattle	426
Houston	374
Chicago	308
Columbus	221
San Diego	170
Springfield	161
Dallas	156
Jacksonville	125
Detroit	115
Newark	92
Jackson	82
Columbia	81
Richmond	81
Aurora	68
Phoenix	63
Arlington	60
San Antonio	59

Key Highlights:

- Over the 4-year period, New York City had the highest number of orders (891), followed by:
  1. Los Angeles (728)
  2. Philadelphia (532)
  3. San Francisco (500)
  4. Seattle (426)

- These top 5 cities alone account for a significant share of total orders.
Urban hubs dominate the sales volume, reflecting high demand in major metropolitan areas.

- The top 20 cities collectively represent a strong concentration of customer activity
There are hundreds of other cities with lower order volumes, but the focus here is on the highest-performing locations for clearer insights.

Implications:

1. Marketing efforts and stock planning can be prioritized in top-performing cities.
2. High-order cities may benefit from regional promotions, faster delivery options, and localized inventory management.

The next Pivot Table is a profit breakdown by region & shipping mode:

Sum of Profit / Ship Mode	East	West	Central	South	Grand Total
Standard Class	82796.66	57499.05	68650.50	54081.61	263027.82
Second Class	18118.05	26396.64	16320.07	19811.02	80645.79
First Class	22176.61	29192.96	8922.21	7037.39	67329.17
Same Day	7525.95	6133.51	2658.61	5674.24	21992.31
Grand Total	**130617.27**	**119222.17**	**96551.39**	**86604.27**	**432995.10**
	30.17%	27.53%	22.30%	20.00%	100.00%

Key Insights:

East Region leads in profitability, contributing over 30% of total profit (≈$130K), followed by:
West: 27.5%

Central: 22.3%
South: 20%
The most profitable shipping mode across all regions is Standard Class — generating over $263K, which is more than 60% of total profit.
East & West benefit significantly from First Class and Second Class, suggesting premium shipping is more profitable in these areas.

The last Pivot Table shows us the profitability by product sub-category:

Sub-Catego	Sum of Profit	Sum of Sales	Profit to Sales Ra
Machines	78647.74	189238.631	41.56%
Copiers	64683.12	146248.094	44.23%
Chairs	51190.32	322822.731	15.86%
Phones	48833.34	327782.448	14.90%
Binders	46725.58	200028.785	23.36%
Tables	34235.73	202810.628	16.88%
Storage	27492.34	219343.392	12.53%
Accessories	19893.99	164186.7	12.12%
Bookcases	19354.58	113813.1987	17.01%
Supplies	14544.83	46420.308	31.33%
Appliances	13935.66	104618.403	13.32%
Furnishings	6749.47	89212.018	7.57%
Paper	3960.24	76828.304	5.15%
Art	1167.52	26705.41	4.37%
Envelopes	878.18	16128.046	5.45%
Labels	648.70	12347.726	5.25%
Fasteners	53.76	3001.96	1.79%
Grand Total	**432995.10**	**2261536.783**	

Top Performers – High Profit & Efficiency:

1. Copiers (44.23% profit margin) and Machines (41.56% profit margin) lead in both profit and profit-to-sales ratio.
2. Though not top in sales, these items yield exceptional returns and are key drivers of overall profitability.

High Sales, Moderate Profitability:

1. Phones, Chairs, and Binders generate significant sales volume but show lower profit margins:
   - Phones: 14.9%
   - Chairs: 15.86%
   - Binders: 23.36%
2. These are core products, but margins can be optimized.

Balanced but Underperforming:

Storage, Accessories, and Furnishings have moderate sales but below-average profit ratios (7–13%). These may benefit from pricing reviews or promotional repositioning.

1.3 Numerical Variables

1.3.1 Sales Data Summary

Min	2.025
Max	28106.716
Average	1838.647791
Q1	378.519
Q2	1058.397

Q3	2382.217
Q4	28106.716

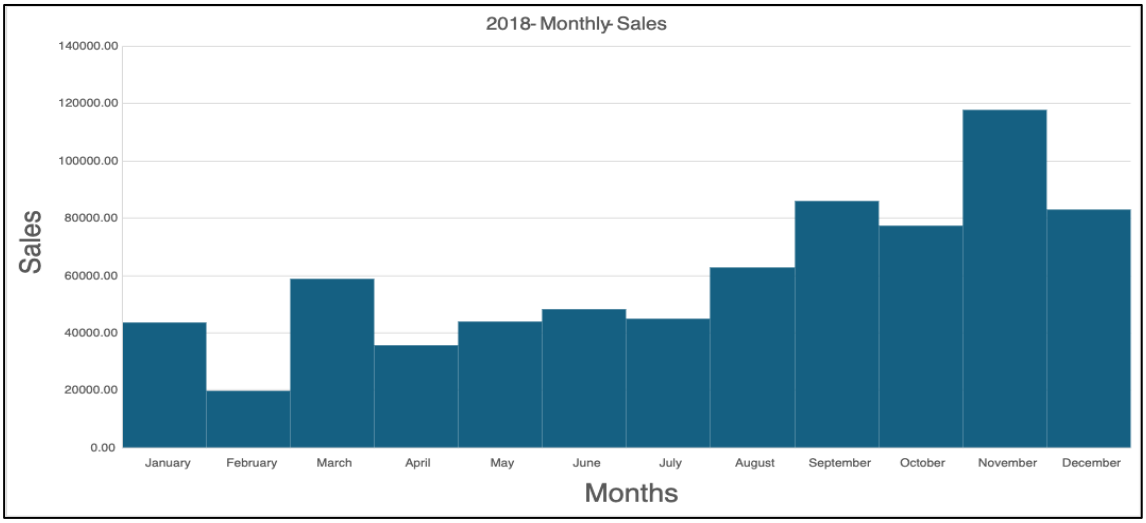
The quartile values provide a summary of the distribution of sales across all transactions in the dataset:

Q1 (First Quartile = 378.52):
This means that 25% of all sales transactions are less than or equal to $378.52. These represent low-value sales and could correspond to smaller or discounted items.
Q2 (Median = 1058.40):
The median sales value is $1,058.40, indicating that half of all sales fall below this amount, while the other half exceed it. This serves as a central reference point for typical transaction sizes.
Q3 (Third Quartile = 2382.22):
75% of sales transactions are less than or equal to $2,382.22. Only the top 25% of sales exceed this value, showing the threshold for high-value purchases.

Overall, the interquartile range (Q3 - Q1) of approximately $2,003.70 reflects the spread of the middle 50% of sales. This highlights the presence of a moderate variation in transaction values, with a small percentage of exceptionally large sales (seen in the max value of $28,106.72) pushing the average higher than the median.

1.3.2 Histograms

1. First histogram is about the monthly sales distribution:

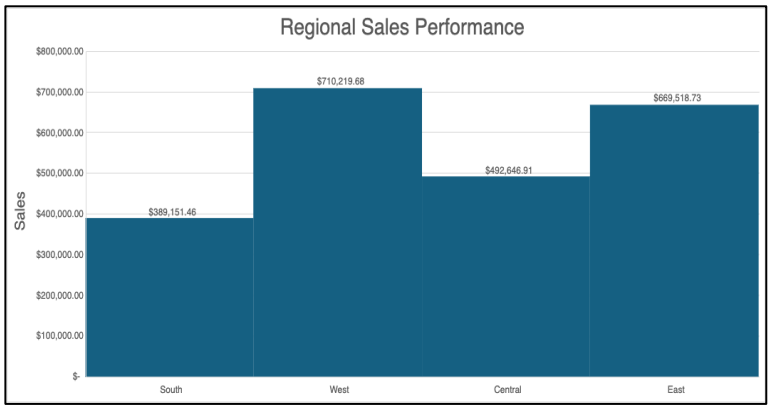


1. Distribution: Left-skewed – most sales in lower-to-mid range, peak towards year-end.

2. Peak Months: Major spike in November ($117K) and December ($83K) due to holiday season.

3. Lowest Sales: February ($19K) – shortest month, low activity.
4. Trend: Gradual rise from August onward, highlighting strong Q4 performance.
5. Insight: Focus marketing & inventory efforts in late-year months.

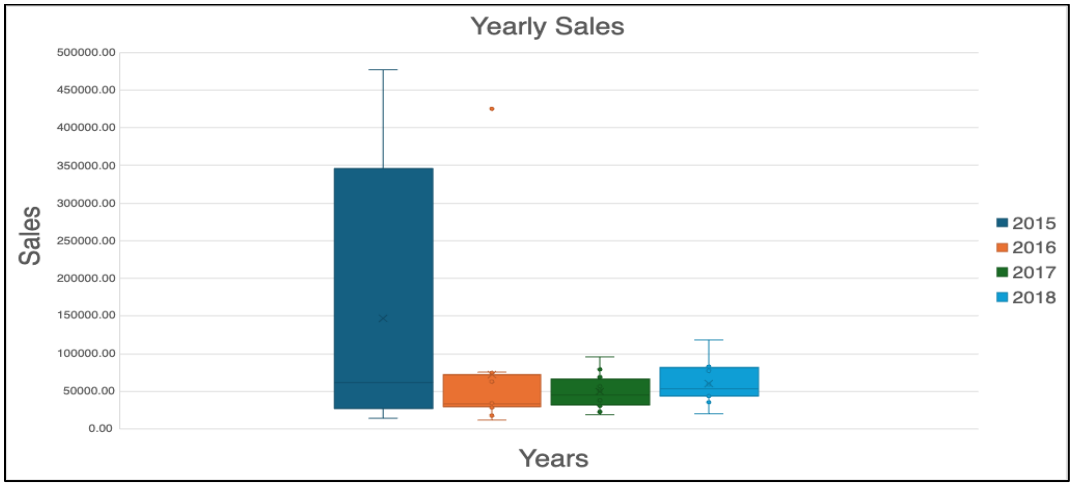
2. The next histogram is about the regional sales distribution:



1. West region leads with the highest sales at $7.1M & East follows closely with $6.7M.
2. Clear performance gap between top (West, East) and lower (Central, South) regions.
3. Suggests stronger sales strategies or market potential in West and East & highlights opportunities for growth in South and Central regions.

1.3.3 Box Plot

1. First is a box plot related to the yearly sales:

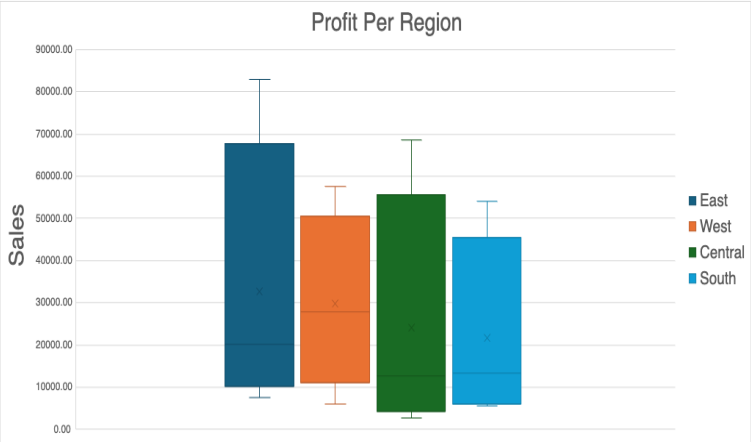


Variability:
2015 & 2016 show exceedingly high variability due to extreme sales spikes in Feb–July (outliers).
2017 & 2018 have more consistent sales, showing lower variability.

Distribution:
2017 & 2018 distributions are more balanced with tighter IQRs.
2015 & 2016 are skewed due to abnormally high sales months.

Outliers:
Large outliers in 2015 (Feb–July) and 2016 (June) inflate overall sales.
2018 has the least outliers, indicating more stable monthly performance.

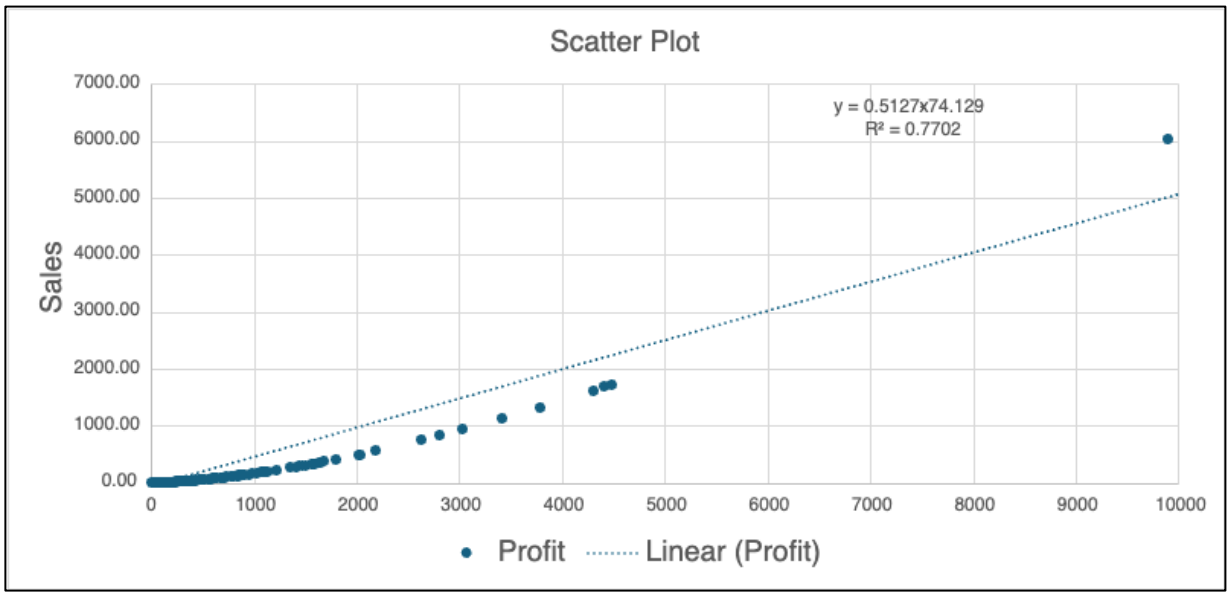
2. The next one is a box plot related to the regional:



1. **High variability** in profits, especially in the East and Central regions.
2. **Standard Class** profits in East and Central is outliers.
3. **South** shows the **least variability**.
4. Distribution is **right-skewed** due to high-profit values.

Linear Regression

2.1 Summary: Sales vs Profit



The scatter plot of Sales vs Profit reveals a strong positive linear relationship between the two variables. This is supported by the following key findings from the regression model:

Correlation Coefficient (r = 0.8776):

- This value is remarkably close to 1, indicating a strong and positive linear correlation. As sales increase, profits also tend to increase proportionally.

Regression Equation:

Profit = 0.5127 × Sales – 74.13

- The slope (b1 = 0.5127) means that, on average, every 1 dollar increase in sales is associated with a 0.5127 cent increase in profit.

- The intercept (b0 = -74.13) suggests that when sales are zero, the predicted profit would be approximately –74.13, which can be interpreted as a baseline or fixed cost in this context.

R^2 (R^2 = 0.7702):

- This means that 77.02% of the variability in profit is explained by the variation in sales, indicating that the model provides a good fit. The remaining 22.98% of variability is due to other factors not captured by the model.

Time Series Analysis

3.1 Data Refinement Process

Started with a raw dataset containing every individual transaction made over 4 years, each with its exact date.
Cleaned and filtered the data to focus only on relevant fields, Order Date and Sales.
Grouped transactions by year and month to calculate the total monthly sales for each year (2015–2018).

3.2 Forecasting Methods Used

Used the refined monthly sales data to perform Time Series Forecasting using 4 key methods:

Naive Method – Uses the most recent value as the forecast.
Average Method – Forecasts based on the average of all past sales.
Moving Average – Smooths fluctuations using the average of recent periods.
Exponential Smoothing – Applies greater weight to recent data for more responsive forecasting.

3.3 Time Series Patterns

Naïve Method(Most Recent Value):	
MAE:	72136.82089
MSE:	22238642421
MAPE:	130.4079866
Average of Historical Data Method:	
MAE:	90711.52308
MSE:	16199410354
MAPE:	201.4157133
Moving Average Forcasting:	
MAE:	68705.82941
MSE:	15262690090
MAPE:	136.7869507
Exponential Smoothing Method:	
MAE:	67785.07493
MSE:	14737648878
MAPE:	121.6183795

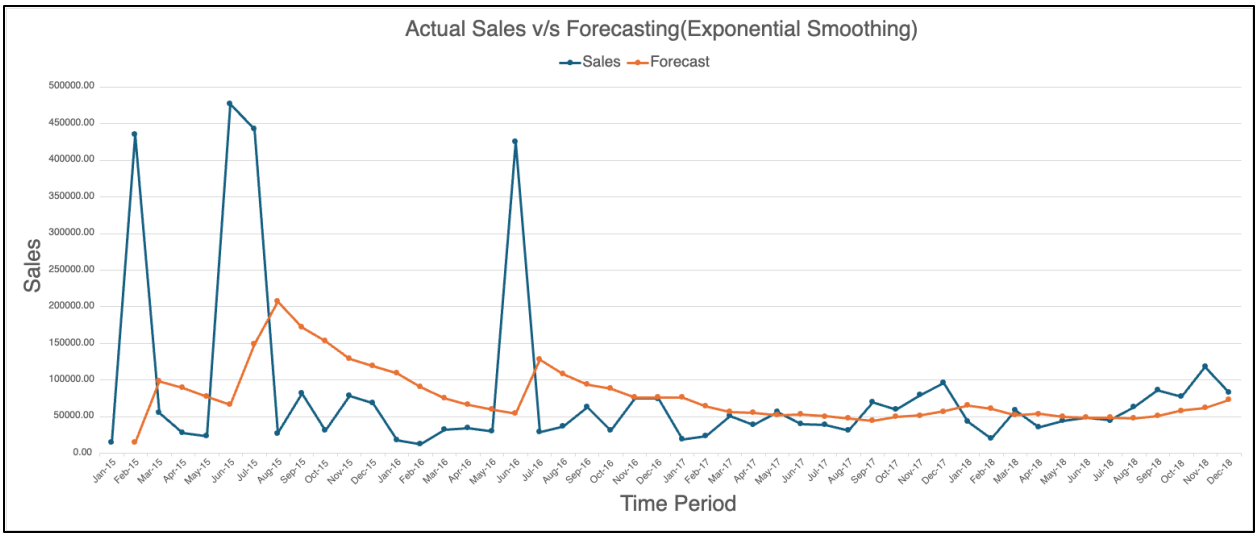
Among all four methods, Exponential Smoothing performs the best based on lowest error values.

Why Exponential Smoothing is Better:

1. Lowest MAE (Mean Absolute Error): Gives more accurate predictions.
2. Lowest MSE (Mean Squared Error): Reduces large forecasting errors.
3. Lowest MAPE (Mean Absolute Percentage Error): More consistent and reliable percentage-based accuracy.

It adapts quickly to recent changes in data, making it more responsive and ideal for dynamic sales trends.

Using Exponential Smoothing, the predicted sales for Jan 2019 is:
$ 75,053.96



Supplementary Information

Dataset Information:

Superstore Sales Dataset
link: https://www.kaggle.com/datasets/rohitsahoo/sales-forecasting/data

Project Documents:

Presentation: BDA100_Project_Final.pptx

Report: BDA100_Project_Report.docx

Excel Workbook: BDA100_Project_Final.xlsx

Course Details:
Course: Introduction to Data Science
Course Code: BDA100NAA
Instructor: Omid Aghababaei Tafreshi

Group 2:
Rishi Raj Das, 165212234
Luka Lazarev, 143871242
Duy Bao Khang Nguyen, 185000239

Note: access to the documents has been given to @seneca.ca emails.