

# Product Sales Data Analysis Report

## Introduction

In the dynamic landscape of modern business, data-driven decision-making stands as a cornerstone for success. The ability to harness the vast volumes of data generated by businesses and extract actionable insights is paramount in gaining a competitive edge. The project, "Product Sales Data Analysis," delves into this realm, aiming to unravel the intricacies of sales data to drive informed decision-making and strategic planning.

In this report, we present a comprehensive analysis of sales data from a business context. The objectives encompass understanding product performance, identifying regional sales trends, and discerning overarching trends shaping sales dynamics. Leveraging the versatile capabilities of Python as a primary analytical tool, we navigate through the complexities of sales data to uncover valuable insights.

## Project Objectives:

The project is driven by the following key objectives:

- **Identifying Top-Performing Products:** We aim to analyze sales data to identify products that are driving revenue and exhibiting exceptional performance. By understanding the factors contributing to their success, we can inform future product strategies and marketing initiatives.
- **Uncovering Regional Sales Trends:** Geographical analysis of sales data enables us to identify regions or markets with the highest sales volumes. By uncovering regional sales trends, businesses can allocate resources effectively, tailor marketing efforts, and capitalize on emerging market opportunities.
- **Analyzing Sales Trends:** Historical sales data provides insights into patterns, fluctuations, and trends affecting overall sales performance. By analyzing sales trends, businesses can anticipate market changes, adapt strategies, and maintain a competitive advantage.

## Methodology

The approach to Product Sales Data Analysis encompasses the following steps:

- **Data Collection and Preprocessing:** We gather and preprocess sales data to ensure its readiness for analysis. This involves cleaning, transforming, and structuring the data for analysis.
- **Exploratory Data Analysis (EDA):** Through exploratory data analysis, we gain insights into the characteristics and distributions of sales data. EDA helps us identify patterns, outliers, and relationships within the data.
- **Statistical Analysis:** We employ statistical techniques to quantify relationships between variables, identify correlations, and assess significance.
- **Visualization and Reporting:** Utilizing Python's visualization libraries, we create informative charts, graphs, and interactive dashboards to visualize sales data and communicate insights effectively.

## Analysis

Based on the appendices, several insights into sales performance, regional distribution, order prioritization, and monthly sales trends can be analyzed by listed below:

- Appendix 1 shows the top 10 performing products by sales, with Apple Smart Phone, Full Size being the highest selling product, generating sales of \$86,936 and a profit of \$5,921,578. Smartphones from major brands like Apple, Cisco, Motorola, and Nokia dominate the top sales, indicating a high demand for these products. Office furniture like executive leather armchairs also feature in the top sellers.
- Appendix 2 reveals that the Central region generated the highest sales of \$2,822,612, followed by the South (\$1,600,960) and North (\$1,248,727) regions. Oceania and Southeast Asia had comparatively lower sales figures.
- Appendix 3 highlights the United States as the country with the highest sales of \$2,297,567, trailing Australia (\$925,257) and France (\$858,930). China, Germany, and Mexico also feature among the top sales generators.
- Appendix 4, a bar chart, illustrates the total sales by order priority. Medium priority orders account for the highest sales, followed by High and Low priority orders. Critical orders have the lowest sales volume.
- Appendix 5 presents the monthly sales trend over a period of several years. The graph shows a cyclical pattern with peaks and troughs, indicating seasonal variations in sales. There is an overall upward trend, with sales reaching their highest point in the last data point (2015-01).
- Appendix 6 shows the total sales segmented by customer type - Consumer, Corporate, and Home Office. The Consumer segment has the highest sales, followed by Corporate and then Home Office.
- Appendix 7 breaks down the total sales by geographic region. The East region has the highest sales, followed by North, Southeast Asia, and West. Africa has the lowest sales among the regions shown.
- Appendix 8 presents the distribution of sales by shipping mode. The majority (59.9%) of sales were shipped via Standard Class, followed by Second Class (20.3%), First Class (14.5%), and Same Day (5.3%).
- Appendix 9 displays the distribution of profit margins as a histogram. The data is heavily right-skewed, with most observations having negative profit margins between -4 and 0, and a high frequency around the 0 profit margin bin.
- Appendix 10 shows the yearly sales growth from 2011 to 2014. Sales grew at an increasing rate each year, starting at around 2.5 in 2011 and reaching over 4.25 by 2014, indicating strong year-over-year sales growth.
- Appendix 11, the proportion of returned products is relatively low at 6.1% compared to total sales at 93.9%. This suggests that the majority of sales are successful transactions without significant product returns.
- Appendix 12, the scatterplot shows a positive correlation between sales and profit, indicating that higher sales generally lead to higher profits. However, there is a wide variation in profits at different sales levels, implying that factors other than sales volume also influence profitability.

- Appendix 13, the monthly sales trend reveals a cyclical pattern, with peaks and troughs occurring approximately every six months. This could be due to seasonal factors or marketing campaigns. Overall, there is an upward trend in sales from 2011 to 2015, suggesting a growing business.

### **Machine Learning approach**

In order to prepare for the Machine Learning (ML) prediction approach for sales, the goal is to determine the approach and in this case the process will be regression with 5 different algorithms: Linear Regression, Ridge, Lasso, Decision Tree and Random Forest with additional ensemble learning method. The goal is to find the best prediction method and determine what factor drives high sales based on a given variable.

Some data cleaning by IQR method with preprocessing method includes calculating the difference in order date and shipped date, removing unneeded variable (such as 'order\_id', 'order\_date', 'ship\_date', 'customer\_name', 'country', 'market', 'state', 'region', 'product\_id', 'product\_name', 'sub\_category'). Next is to detect collinearity in appendix 14 and as result there are 2 colinearity, sales to shipping cost, which makes sense because more sales = more shipping. Second is profit per quantity to profit and profit margin, this we only keep profit for overall insights. Moreover, knowing the distribution of target data is critical to make a decision to include an extra step of transformation or scaling method and from appendix 15, the target follows the Poisson distribution, therefore a decision for scaling method is needed.

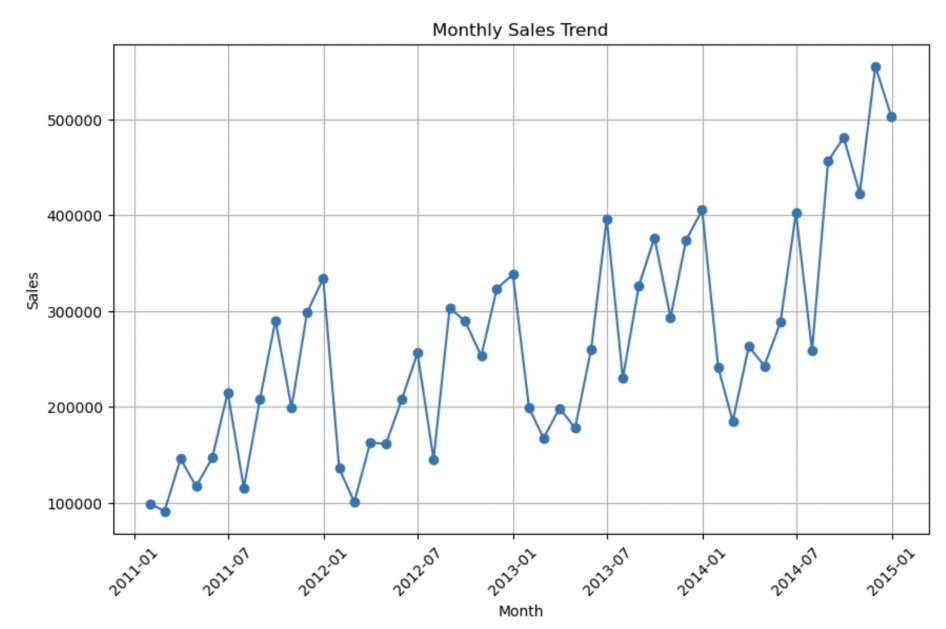
In the 5 baseline model in appendix 15, there is a positive result and the best performing model is XGBoost, regarding the high Mean Squared Error (MSE) rate. However, the Linear Regression, Ridge and Lasso can be improved since the data is not yet scaled. This gives another result in appendix 16, even with the scaling method applied, it only shows a minimal increase in performance efficiency. Therefore, the algorithm needs to be applied in this case is XGBoost since it yields the lowest error and highest R2 score

### **Conclusion**

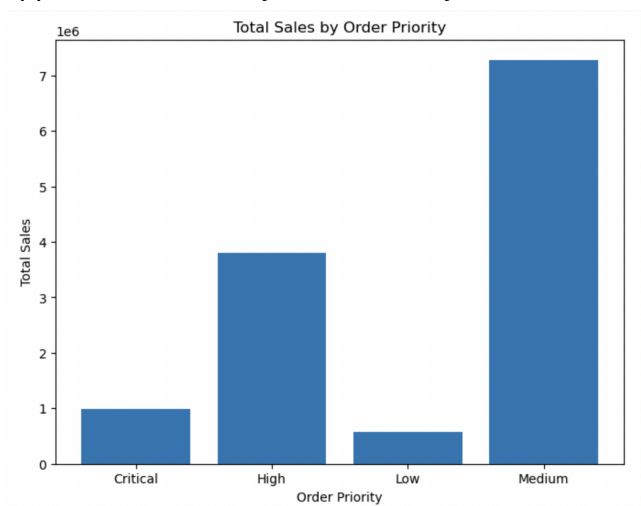
In conclusion, the "Product Sales Data Analysis" project demonstrated the power of data-driven decision-making in driving business success. By leveraging advanced analytics techniques, businesses can gain a deeper understanding of their sales dynamics, optimize strategies, and capitalize on market opportunities. Moving forward, continuous monitoring and analysis of sales data will be essential for adapting to evolving market trends and maintaining a competitive edge in the marketplace.

Appendix

Appendix 1: Monthly Sales Trend



Appendix 2: Sales by Order Priority



Appendix 3: Top 10 performing product by sales

Top 10 performing products by sales:

	product_name	sales	profit
310	Apple Smart Phone, Full Size	86936	5921.5786
970	Cisco Smart Phone, Full Size	76441	17238.5206
2415	Motorola Smart Phone, Full Size	73159	17027.1130
2501	Nokia Smart Phone, Full Size	71904	9938.1955
866	Canon imageCLASS 2200 Advanced Copier	61600	25199.9280
1837	Hon Executive Leather Armchair, Adjustable	58200	5997.2541
2631	Office Star Executive Leather Armchair, Adjustable	50667	4710.9840
1714	Harbour Creations Executive Leather Armchair, ...	50120	10427.3260
2988	Samsung Smart Phone, Cordless	48654	-198.0900
2502	Nokia Smart Phone, with Caller ID	47880	9465.3257

#### Appendix 4: Analysis Region of highest sales

Region with the highest sales:

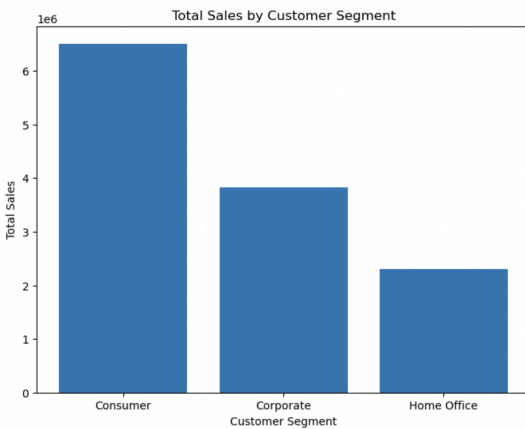
	region	sales
3	Central	2822612
10	South	1600960
7	North	1248727
9	Oceania	1100207
11	Southeast Asia	884438
8	North Asia	848349
5	EMEA	806184
0	Africa	783776
4	Central Asia	752839
12	West	725514

#### Appendix 5: Analysis of country with highest sales

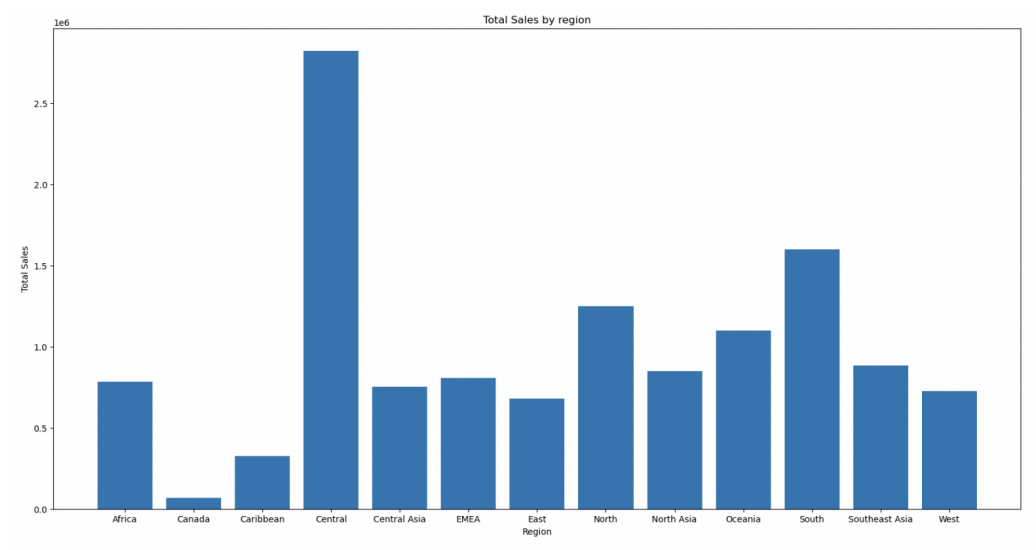
country with the highest sales:

	country	sales
139	United States	2297567
6	Australia	925257
44	France	858930
26	China	700591
47	Germany	628857
81	Mexico	623155
57	India	589664
138	United Kingdom	528570
58	Indonesia	404887
17	Brazil	361098

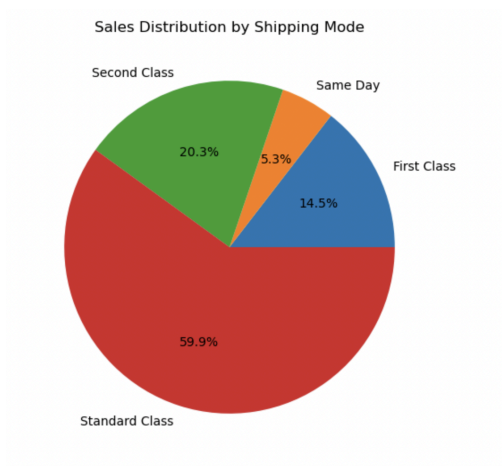
#### Appendix 6: Sales by Customer Segments



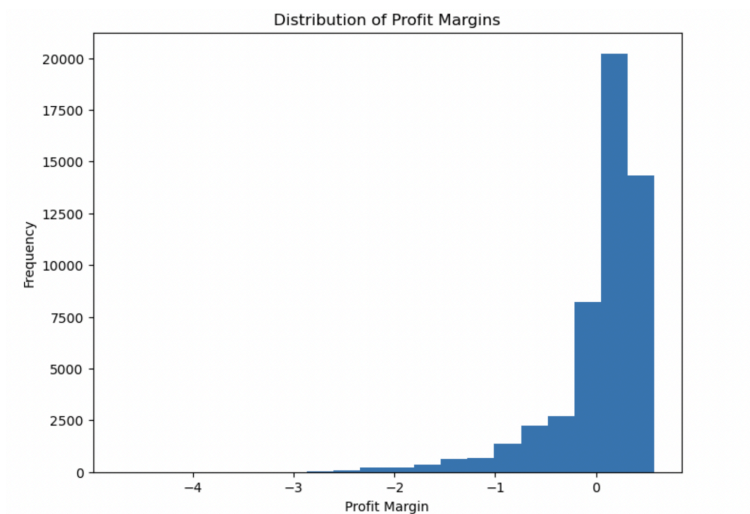
Appendix 7: Sales by Region



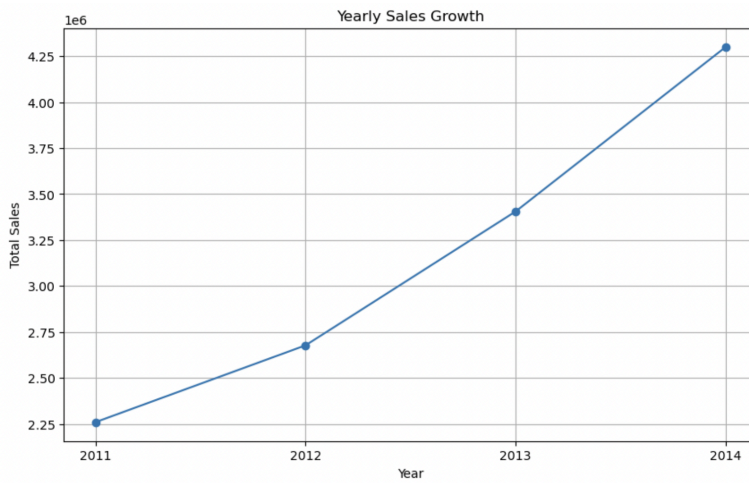
Appendix 8: Sales by shipping mode



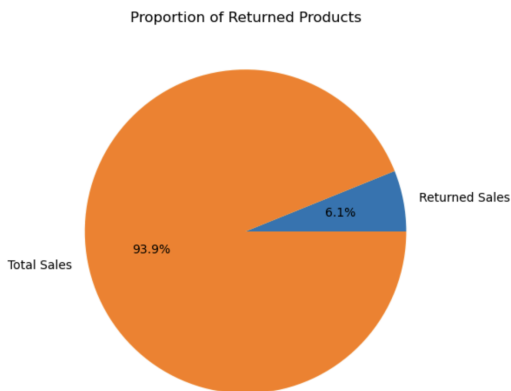
Appendix 9: Distribution of Profit Margins



### Appendix 10: Yearly Sales Growth



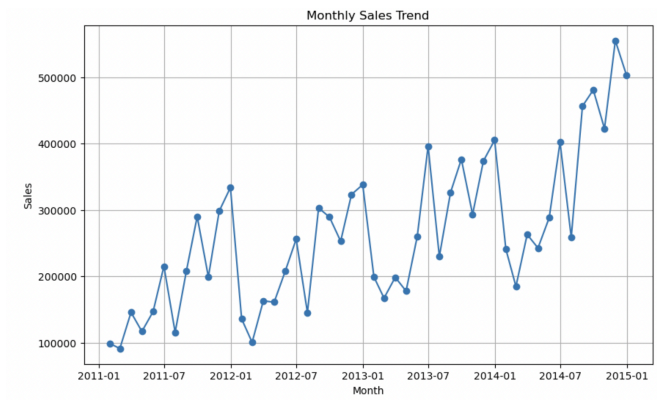
### Appendix 11: Proportion of Returned Product



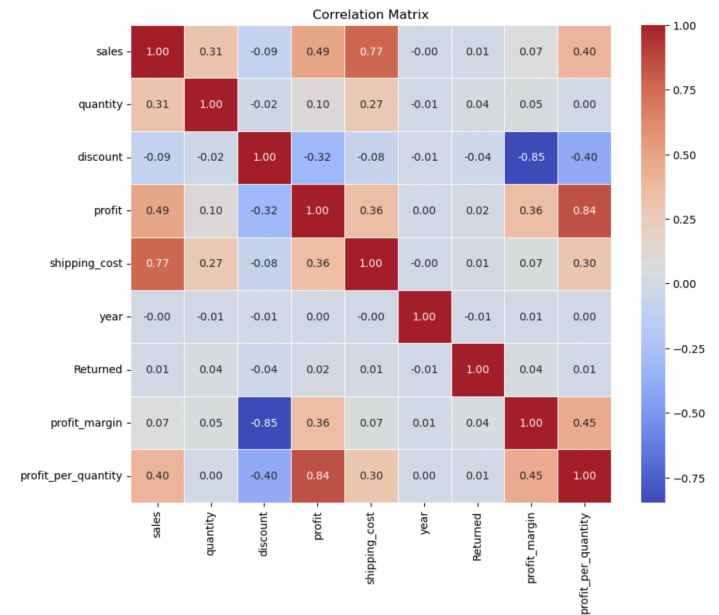
### Appendix 12: Sales vs Profit



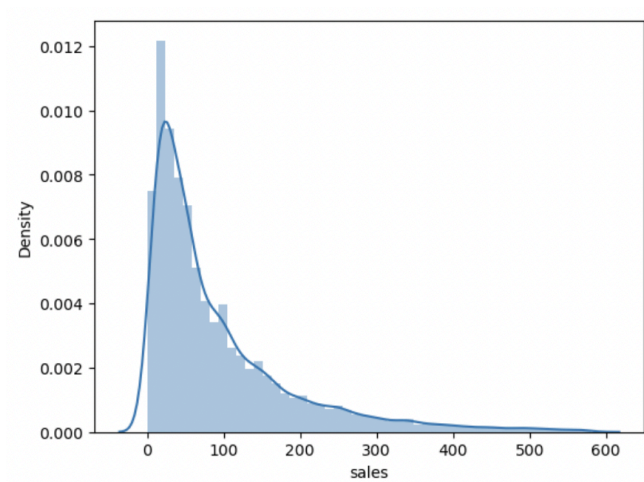
Appendix 13: Monthly Sales Trend



Appendix 14: Correlation Matrix



Appendix 15: Sales distribution





#### *Appendix 16: Baseline model result*

```
{ 'Linear Regression': { 'R2 Score': 0.3871069150353458,  
  'Mean Squared Error': 5732.120611624121},  
  'Ridge': { 'R2 Score': 0.3871120489054951,  
    'Mean Squared Error': 5732.072596784956},  
  'Lasso': { 'R2 Score': 0.3857024857354028,  
    'Mean Squared Error': 5745.255623806934},  
  'DecisionTreeRegressor': { 'R2 Score': 0.07286041677318289,  
    'Mean Squared Error': 8671.13048791132},  
  'RandomForestRegressor': { 'R2 Score': 0.5186225248614824,  
    'Mean Squared Error': 4502.112709221064},  
  'XGBosst': { 'R2 Score': 0.5338477148043013,  
    'Mean Squared Error': 4359.718175446616}}
```

#### *Appendix 17: Scaled data model*

```
{ 'Linear Regression': { 'R2 Score': 0.38636886749448696,  
  'Mean Squared Error': 5739.023247051244},  
  'Ridge': { 'R2 Score': 0.3863695681276035,  
    'Mean Squared Error': 5739.016694336542},  
  'Lasso': { 'R2 Score': 0.38700704056531277,  
    'Mean Squared Error': 5733.054693803046}}
```