Exploratory Data Analysis of Flipkart's 2014-2018 Sales Data

August 2025

1 Introduction

This report presents an exploratory data analysis (EDA) of Flipkart's sales data from 2014 to 2018, aimed at identifying key revenue and profit drivers across products, channels, and regions, uncovering seasonal trends, detecting outliers, and aligning performance with 2017 budgets. The analysis leverages Python libraries (Pandas, NumPy, Matplotlib, Seaborn) to process and visualize data from the ”Regional Sales Dataset.xlsx” file, which includes sales orders, customer details, product information, regional data, state-region mappings, and 2017 budgets. The insights derived inform strategies for pricing, promotions, and market expansion to ensure sustainable growth and reduced concentration risk.

2 Data Preparation and Cleaning

The dataset comprises six Excel sheets: Sales Orders, Customers, Products, Regions, State Regions, and 2017 Budgets. These were loaded into Pandas DataFrames and merged to create a comprehensive dataset for analysis. The Sales Orders DataFrame, containing order details such as OrderNumber, OrderDate, Customer Name Index, Channel, and financial metrics (Unit Price, Line Total, Total Unit Cost), was merged with Customers, Products, Regions, and Budgets DataFrames using appropriate keys (e.g., Customer Name Index, Product Description Index, Delivery Region Index, and Product Name). The State Regions DataFrame required preprocessing to set the first row as column headers and remove it from the data.

Null value checks revealed no missing data across all DataFrames, ensuring data integrity. A Profit column was calculated as the difference between Line Total and Total Unit Cost, enabling profitability analysis. The merged dataset facilitated a multidimensional analysis of sales performance.

3 Exploratory Data Analysis

3.1 Distribution and Outlier Analysis

Histograms and box plots were used to examine the distributions of Line Total and Total Unit Cost. The histograms showed right-skewed distributions, indicating a higher frequency of lower-value transactions with some high-value outliers. Box plots confirmed significant outliers in both metrics, suggesting potential large transactions or data anomalies requiring further investigation.

3.2 Relationship Analysis

Scatter plots explored relationships between Order Quantity and Line Total, and Unit Price and Line Total. A positive correlation was observed, with higher quantities and unit prices generally leading to higher Line Totals, though variability suggested other influencing factors.

3.3 Revenue and Profit Analysis

The data was grouped by Product Name, Channel, and Region to compute total revenue (sum of Line Total) and total profit (sum of Profit). The top 10 products by revenue and profit highlighted key performers, with some products showing higher profitability despite lower revenue rankings. Channel performance revealed that Wholesale and Distributor channels contributed significantly to revenue, while Export channels showed varied profitability. Regionally, certain regions dominated revenue and profit, indicating market concentration that may pose risks.

3.4 Seasonal Trends

Time-based analysis involved extracting year, month, and day-of-week from OrderDate. Monthly sales and profit trends, visualized via line plots, showed seasonal patterns with peaks in specific months, likely tied to festive seasons or promotions. Weekly analysis indicated consistent sales across days, with slight variations by year. These trends suggest opportunities for targeted promotions during high-demand periods.

3.5 Outlier Detection by Dimension

Box plots by Channel, Region, and top 10 products by revenue revealed outliers in Line Total and Profit. Certain channels and regions exhibited greater variability, potentially due to high-value transactions or regional market dynamics. Product-specific outliers suggested varying cost structures or pricing strategies.

3.6 Correlation Analysis

A correlation matrix for numerical variables (Order Quantity, Unit Price, Line Total, Total Unit Cost, Profit) was visualized using a heatmap. Strong positive correlations were observed between Line Total and both Order Quantity and Unit Price, confirming their role as revenue drivers. Profit showed moderate correlation with Line Total, indicating cost impacts on profitability.

3.7 Sales Distribution

A pie chart illustrated the distribution of sales across channels, with Wholesale and Distributor channels dominating, highlighting channel concentration. A time series plot of daily Line Total showed spikes on specific dates, potentially linked to promotional events or bulk orders.

4 Insights and Recommendations

The analysis uncovered several actionable insights:

- Key Drivers: High-revenue products and channels (Wholesale, Distributor) are critical, but profitability varies, suggesting a need for cost optimization in low-margin products.
- Seasonal Trends: Seasonal peaks indicate opportunities for targeted promotions during high-demand months to boost revenue.
- Outliers: Significant outliers in Line Total and Profit across channels and regions warrant investigation to distinguish between legitimate transactions and potential data errors.
- Concentration Risk: Dominance of specific channels and regions suggests a need for market diversification to mitigate risks.
- Pricing and Promotions: Strong correlations between Order Quantity, Unit Price, and Line Total support dynamic pricing strategies to maximize revenue.

Recommendations include optimizing pricing for high-margin products, timing promotions with seasonal peaks, expanding into underrepresented regions, and validating outliers to ensure data accuracy. These strategies can enhance profitability and support sustainable growth.

5 Conclusion

This EDA of Flipkart's 2014-2018 sales data provides a comprehensive understanding of revenue and profit dynamics, seasonal trends, and outliers. By leveraging these insights, Flipkart can refine pricing, promotions, and market expansion strategies to achieve sustainable growth and reduce concentration risk. Future analyses could incorporate predictive modeling to forecast sales and refine budget alignments.