Optimizing E-Commerce Profitability through Data Analytics

Peter Cheng

2025-11-05

Dataset credit: *E-Commerce Sales Dataset — "Unlock Profits with E-Commerce Sales Data"* on Kaggle, authored by **ANil**.

Background

E-commerce retailers operate across multiple platforms (Amazon, Flipkart, Myntra, Ajio, etc.) with heterogeneous pricing, fulfillment, and fee structures. This project analyzes public, SKU-level sales and pricing data to identify profit drivers and deliver actionable recommendations.

Project Objective

Analyze and predict **e-commerce profitability** across platforms, fulfillment methods, and international markets using transactional, pricing, and inventory data.
Specifically, the project aims to:

- **Identify the most profitable product categories** and fulfillment strategies (e.g., Amazon FBA vs. Merchant Fulfilled, Shiprocket vs. INCREFF).
- **Compare platform and market performance** in terms of margin, sales volume, and operational reliability.
- **Quantify key profitability drivers** — including price ratios, catalog identity, platform alignment, and fulfillment efficiency.
- **Provide actionable recommendations** for dynamic pricing, inventory allocation, and cross-market fulfillment planning.

Business Question

Which **product categories**, **sales platforms**, and **fulfillment methods** generate the highest profit margins,
and how do factors such as **pricing structure**, **catalog mix**, and **operational performance** influence e-commerce profitability across domestic and international markets?

Data Source

Primary dataset: *Unlock Profits with E-Commerce Sales Data* (Kaggle), **author: ANil**.

Methodology

1. **Prepare** → Import & merge CSVs; harmonize schemas; standardize currency; parse dates.
2. **Process** → Remove duplicates; handle missing; derive features (`profit`, `profit_margin`, `month`, etc.).
3. **Analyze** → Descriptive trends; regression for drivers; clustering SKUs; time-series.
4. **Share** → ggplot2 visuals + Tableau Public dashboard.
5. **Act** → Prioritized actions (pricing, inventory, fulfillment).

Tools & Libraries

All analysis is conducted in **RStudio**, with two distinct environments for different phases of the workflow. The **Data Preparation** stage focuses on cleaning and structuring datasets, while the **Analysis & Visualization** stage expands into modeling, clustering, and advanced reporting.

Data Preparation Environment

Used in the *E-Commerce Preparation.Rmd* file for cleaning, validation, and schema consistency checks.

- **tidyverse** — data wrangling and manipulation (`dplyr`, `tidyr`, `stringr`, `readr`)
- **readr** — efficient import of large CSV files
- **janitor** — cleaning and standardizing variable names and tabular data
- **lubridate** — handling and transforming date-time variables
- **skimr** — quick exploratory summaries and descriptive statistics

These packages ensure a clean, well-structured dataset ready for downstream modeling and visualization.

Analysis & Visualization Environment

Used in the *E-Commerce Analysis & Visualization.Rmd* file for modeling, feature evaluation, and visualization.

Core wrangling & plotting

- **tidyverse** — core data wrangling and plotting functions (`dplyr`, `tidyr`, `ggplot2`, `forcats`)
- **lubridate** — handling time-related fields

Modeling & evaluation

- **broom** — tidy model outputs and coefficient summaries
- **caret** — unified framework for regression/classification workflows
- **randomForest** — tree-based baseline models
- **glmnet** — regularized GLMs (lasso/ridge)
- **rpart** – recursive partitioning for interpretable decision trees
- **rpart.plot** – visualization of decision tree structures

Diagnostics & Visualization

- **ggcorrplot** — correlation heatmaps for feature diagnostics

Reporting

- **gt** — publication-quality summary tables

  Compared with the preparation phase, this environment adds modeling (`caret`, `glmnet`, `randomForest`), diagnostics (`factoextra`, `ggcorrplot`), and reporting (`gt`) to support hypothesis testing and insight generation.

All package installation and environment setup are handled programmatically at the start of each `.Rmd` file to ensure **reproducibility** and **runtime efficiency** throughout the workflow.

Analysis Plan

The **comprehensive data integration and cleaning** will be conducted during the formal Exploratory Data Analysis (EDA) phase to ensure data quality and structure are fully understood before modeling. This proposal focuses on the analytical design rather than code execution.

Data Preparation

- Import and inspect all relevant CSV files (Amazon, P&L reports, Warehouse Comparison, Inventory, etc.).
- Harmonize identifiers such as SKU and product category across datasets.
- Merge compatible tables to create a unified master dataset for profitability analysis.
- Handle missing values, duplicates, and inconsistent data types.
- Derive analytical variables including `profit`, `profit_margin`, `platform`, `category`, and `fulfillment_type`.
- Conduct preliminary descriptive summaries and outlier detection.

Modeling & Analytical Approach

- **Descriptive Analysis:**
  Compute summary statistics and visualize revenue, profit, and margin trends by platform, category, and fulfillment method.
- **Comparative Analysis:**
  Use grouped metrics and boxplots to compare profitability across e-commerce platforms (Amazon, Flipkart, Ajio, etc.).
- **Regression Modeling:**
  Build multiple linear regression or generalized linear models to identify which factors (e.g., platform, category, weight, fulfillment type) significantly influence profit margin.

- **Clustering:**
  Apply K-means or hierarchical clustering to segment products based on sales volume and profitability characteristics.
- **Predictive Forecasting:**
  If sufficient time-series data is available, explore revenue forecasting using ARIMA or Prophet models.

Visualization & Reporting

- Use **ggplot2** for visual storytelling (trend lines, bar charts, correlation heatmaps).
- Create an interactive **Tableau Public** dashboard to highlight profitability insights by category and platform.
- Summarize all findings in the final R Markdown report and executive summary.

Deliverables

- Cleaned master dataset
- R Markdown analysis report
- Presentation Slides

References

- ANil. *E-Commerce Sales Dataset — Unlock Profits with E-Commerce Sales Data.* Kaggle.