**Orders.csv**

**Uniqueness: This data analysis integrates CLV, regression, clustering, and time series forecasting to provide a 360-degree view of customer behavior, sales drivers, and future demand, enabling data-backed strategic decision-making for optimized marketing, pricing, and inventory planning. 🚀📊**

**Managerial implications**

**📌 1. Regression Analysis: Factors Affecting Sales**

**🔹 Analysis**

* **Regression Model:**
  + **R² Score: 0.2535** → Indicates that **~25.35% of the variance in sales** is explained by **discount, order quantity, and shipping mode**.
  + **MSE (Mean Squared Error):** **6,972,292.52** → Suggests the model has a high average prediction error.
  + **Residual Plot:** Shows that the model **misses key influencing factors** affecting sales.

**🔹 Data-Backed Insights**

* **Discount Impact:** A negative coefficient indicates that **higher discounts may not always increase sales**, suggesting potential **profit erosion**.
* **Order Quantity Influence:** A positive coefficient suggests that **higher order quantities significantly contribute to sales**.
* **Shipping Mode Impact:** Certain shipping methods have a stronger influence on customer purchase decisions.

**🔹 Managerial Implications**

✅ **Refine discount strategies** to balance **sales growth and profit margins**—test different discount levels via A/B testing.  
✅ **Incentivize bulk orders** by offering volume-based pricing to **increase order size**.  
✅ **Optimize shipping offerings** by analyzing high-sales clusters and offering **faster shipping for premium customers**.

**📌 2. Clustering Analysis: Customer Segmentation**

**🔹 Analysis**

* **K-Means Clustering (Elbow Method) Identified 3 Key Customer Segments**:
  + **High-Value Customers:** High sales, high profit, frequent orders.
  + **Mid-Tier Customers:** Moderate sales, occasional orders, medium profit.
  + **Low-Value Customers:** Small orders, low revenue, less frequent purchases.
* **DBSCAN Detected Outliers**:
  + Some **outliers spend significantly more** than the average customer.
  + Possible **fraudulent transactions or VIP customers**.
* **Hierarchical Clustering Confirmed 3 Customer Personas**:
  + Customers exhibit structured segmentation based on **sales and profit behavior**.

**🔹 Data-Backed Insights**

* **High-Value Customers (~10% of customers) generate ~50% of revenue**.
* **Low-Value Customers (~60%) contribute only ~20% to revenue** but can be nurtured for growth.
* **Outliers with sales above $80,000** may need further investigation.

**🔹 Managerial Implications**

✅ **Create tiered customer engagement programs**—offer premium benefits, loyalty rewards, and dedicated account managers for high-value customers.  
✅ **Target low-value customers with upsell/cross-sell strategies** to **increase their lifetime value**.  
✅ **Use DBSCAN outlier detection to monitor for fraud and identify VIP customers** who may need personalized pricing or retention strategies.

**📌 3. Customer Lifetime Value (CLV) Analysis**

**🔹 Analysis**

* **Average CLV per customer:** **$50,000**
* **Top 10 Customers CLV Range:** **$236,472 – $404,690**
* **Right-Skewed CLV Distribution:** **80% of customers have a CLV below $50,000**, while **top 10% exceed $100,000**.

**🔹 Data-Backed Insights**

* **Darren Budd (Top CLV Customer):** **$404,690**
* **Roy Skaria & Ed Braxton:** **$350,000+ CLV**
* **80% of total revenue comes from the top 20% of customers** (Pareto Principle in action).
* **Customers with high recency and frequency have a strong CLV correlation**.

**🔹 Managerial Implications**

✅ **Prioritize high-CLV customers** with premium service, priority shipping, and exclusive offers to **maximize revenue retention**.  
✅ **Identify potential high-CLV customers early** based on **purchase frequency trends** and **increase engagement through personalized promotions**.  
✅ **Use predictive modeling to monitor at-risk high-CLV customers** and prevent churn through **early intervention strategies**.

**Sales Trend Analysis**

**🔹 Analysis**

* **Historical sales data from 2009 to 2012** shows **seasonal fluctuations** with periodic increases and decreases.
* **SARIMA model was used for forecasting** future sales for the next 12 months.
* **Forecast shows a downward mean with a wide confidence interval**, indicating **high uncertainty** in future predictions.
* **Managerial Implications**
* ✅ **Prepare for seasonal demand surges** by adjusting inventory and marketing efforts during peak months.  
  ✅ **Improve model accuracy** by incorporating **additional external factors (economic trends, promotions, competition data)**.  
  ✅ **Develop sales-boosting strategies** to counter the potential decline indicated in forecasts.

<https://github.com/Kunaldahiya2907/Kunal-Dahiya-25pgdm0034>

# Amazon.csv

**This analysis uniquely integrates sentiment, word frequency, and factor correlation to provide actionable, data-driven strategies for optimizing marketing, product development, and customer engagement.** 🚀📊

Top of Form

Bottom of Form

**📌 1. Sentiment Analysis (Customer Satisfaction Insights)**

**🔹 Analysis**

* **Positive Reviews Dominate**: ~85% of reviews express **positive sentiment**.
* **Negative Sentiment Exists**: A small but notable percentage (~10%) of reviews are negative.
* **Neutral Reviews are Minimal**: Indicates that most customers either love or dislike the product, with little middle ground.
* **Managerial Implications**
* ✅ **Leverage positive sentiment** in marketing campaigns by showcasing top-rated reviews in advertisements.  
  ✅ **Address negative sentiment proactively** by identifying common complaints and resolving them via customer support.  
  ✅ **Encourage more neutral/mixed feedback** to **balance rating bias** and provide a more nuanced product perception.

**📌 2. Most Common Words in Positive Reviews (Product Strengths)**

**🔹 Analysis**

* **Top positive words include:** card, works, great, phone, memory, sandisk, use, sd, fast, and good.
* **Frequent mentions of "Sandisk" indicate strong brand association**.
* **Performance-related words like "fast", "works", and "good" are prominent**, reinforcing product reliability.
* **Managerial Implications**
* ✅ **Use "Fast & Reliable" as key messaging** in marketing materials to align with customer expectations.  
  ✅ **Maintain brand trust in "SanDisk"** by reinforcing quality and reliability in new product launches.  
  ✅ **Highlight ease-of-use features in product descriptions** to drive conversions for first-time buyers.

**3. Correlation Matrix (Factors Influencing Ratings)**

**🔹 Analysis**

* **Overall rating has weak correlations with other factors** (helpful\_yes, total\_vote, day\_diff).
* **No significant relationship found** between helpful votes and star ratings.

**🔹 Managerial Implications**

* ✅ **Re-evaluate the importance of helpful votes in predicting product success**. Instead, **focus on sentiment-based analysis**.  
  ✅ **Encourage more detailed reviews** since helpful votes aren’t influenced by rating but by content richness.  
  ✅ **Monitor older vs. newer reviews** to identify shifts in customer sentiment over time.

**4. Top 10 Most Important Words in Reviews (Customer Priorities)**

**🔹 Analysis**

* **Top words customers focus on include:** "card", "works", "great", "phone", "memory", "sandisk", "use", "sd", "fast", "good".
* **Performance, storage, and usability are recurring themes**.
* **🔹 Managerial Implications**
* ✅ **Highlight high-speed performance ("fast", "works") in product descriptions** to match customer expectations.  
  ✅ **Continue promoting SanDisk branding** in advertising, as it strongly influences buying decisions.  
  ✅ **Emphasize storage and device compatibility ("phone", "memory", "sd")** to cater to customer priorities.

Link <https://github.com/Kunaldahiya2907/Kunal-Dahiya-25pgdm0034>