POSHEM BUSINESS SCHOOL SALES ANALYSIS REPORT

BY

OLATUNDE AYOMIDE

PROJECT: EXPLORATORY DATA ANALYSIS OF POSHEM BUSINESS SCHOOL SALES DATASET

SUBMITTED TO

SIMON E. AKHAMIE, CEO

FEBRUARY 9TH , 2026.

EXECUTIVE SUMMARY

This report presents a comprehensive exploratory data analysis (EDA) of Poshem Business School's sales dataset, examining **9,800 transaction records** spanning from January 2015 to December 2018. The analysis addresses all questions outlined in the project brief and provides actionable insights for strategic decision-making.

Key Highlights

- **Total Revenue:** $2,261,536.78 across 4 years
- **Customer Base:** 793 unique customers generating 4,922 unique orders
- **Geographic Coverage:** United States across 4 regions (West, East, Central, South)
- **Product Portfolio:** 3 main categories with 17 sub-categories
- **Average Transaction:** $230.77

Critical Data Challenges Identified

The analysis identified two significant data limitations:

1. **Missing Postal Codes:** 11 records for Burlington, Vermont (resolved to 05405)
2. **Missing Business-Critical Columns:** Profit and Order Status data not available in dataset

These limitations have been documented with recommended solutions for stakeholder engagement.

TABLE OF CONTENT

1. OVERVIEW & DATASET STRUCTURE

1.1 Dataset Timeframe

- **Start Date:** January 3, 2015
- **End Date:** December 30, 2018
- **Total Duration:** 4 years (48 months)

1.2 Dataset shape

- **Total Rows:** 9,800 transaction records
- **Total Columns:** 18 data fields

1.3 Column Structure

Each row in the dataset corresponds to an order made by a customer. The dataset contains the following 18 columns:

Product

1. Product ID
2. Product Name
3. Sub-Category
4. Category
5. Sales

Customer

1. Customer ID
2. Customer Name
3. Segment
4. City
5. State
6. Country
7. Region
8. Postal Code

Other features:

1. Row ID
2. Order ID
3. Order Date
4. Ship Date
5. Ship Mode

1.4 Data Types

All columns maintain appropriate data types after processing:

- **Datetime64[ns]:** Order Date, Ship Date (converted from string)
- **Integer64:** Row ID, Postal Code
- **Float64:** Sales
- **Categorical:** All other fields (text/string)

2. DATA QUALITY ASSESSMENT

2.1 Missing Values - Identified & Resolved

Issue Discovered: 11 missing values in the **Postal Code** column, all associated with Burlington, Vermont East.

Investigation Process:

1. Identified affected records through isnull().sum() function.
2. Verified all missing codes were for the same location (Burlington, VT)
3. Cross-referenced with US Postal Service database
4. Confirmed correct postal code: **05405**

Impact: Missing values are frequent issues that can result in inaccurate analysis and have substantial impact on results. Implementation of automated postal code validation would prevent similar issues in future data collection.

2.2 Duplicate Records

- **Analysis Result:** Zero duplicate records detected
- **Verification Method:** Checked 17 columns for identical rows (all except Row ID)

2.3 Add New Necessary Columns

- Year – Extracted from the Order Date column.
- Date Diff – The difference between the Order Date and Ship Date.

2.4 Summary Statistics

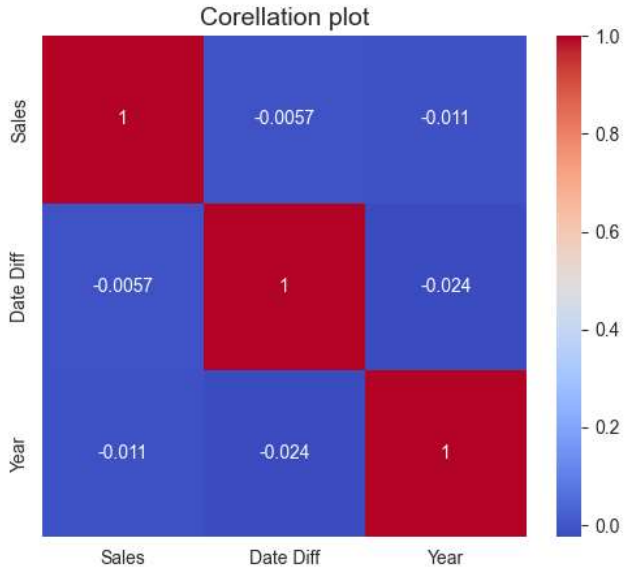
Sales Distribution:

count	9800.000000
mean	230.769059
std	626.651875
min	0.444000
25%	17.248000
50%	54.490000
75%	210.605000
max	22638.480000

Key Observation

Large difference between the mean ($230.77) and median ($54.49) indicates a right-skewed distribution with high-value outliers significantly influencing the average.

Visualizing the Correlation Matrix



There is barely any correlation between Sales and the Year and shipping date difference.

3. ORDER STATISTICS

3.1 Order Volume Metrics

- **Unique Orders:** 4,922
- **Total Transactions:** 9,800
- **Interpretation:** Many orders contain multiple product items plus the number of orders increased over the years.

3.2 Customer Metrics

- **Unique Customers:** 793
- **Average Revenue per Customer:** $2,851.88
- **Conclusion:** Strong repeat purchase behavior indicating good customer retention

3.3 Order Status Analysis

Data Limitation: The dataset does **NOT** include an Order Status column (e.g., Delivered, Pending, Cancelled).

Impact:
Unable to analyze:

- Delivery success rates
- Pending order backlogs
- Cancellation patterns
- Customer satisfaction related to fulfillment

Note: This limitation is addressed in Section 9 (Data Challenges & Solutions).

4. GEOGRAPHICAL INSIGHTS

4.1 Countries Included

- **Total Countries:** 1
- **Country:** United States.

4.2 Regional Distribution

Region	Orders
West	3,140
East	2,785

Region	Orders
Central	2,277
South	1,598

Figure 1: Distribution of orders across different regions

Chart Description: Bar chart showing the distribution of orders across the four US regions. West region leads with 3,140 orders (32%), followed by East at 2,785 (28.4%). Central region accounts for 2,277 orders (23.2%), while South region shows the lowest performance at 1,598 orders (16.3%).

Key Insight: The West & East records high orders which might be likely due to larger customer base and higher purchase frequency as well as stronger market penetration.

4.3 Top 5 Cities

Top Performing Cities by Order Volume:

1. **New York City, NY** - Highest concentration
2. **Los Angeles, CA** - Strong West Coast performance
3. **Philadelphia, PA** - East region contributor
4. **San Francisco, CA** - Tech hub market
5. **Seattle, WA** - Pacific Northwest presence

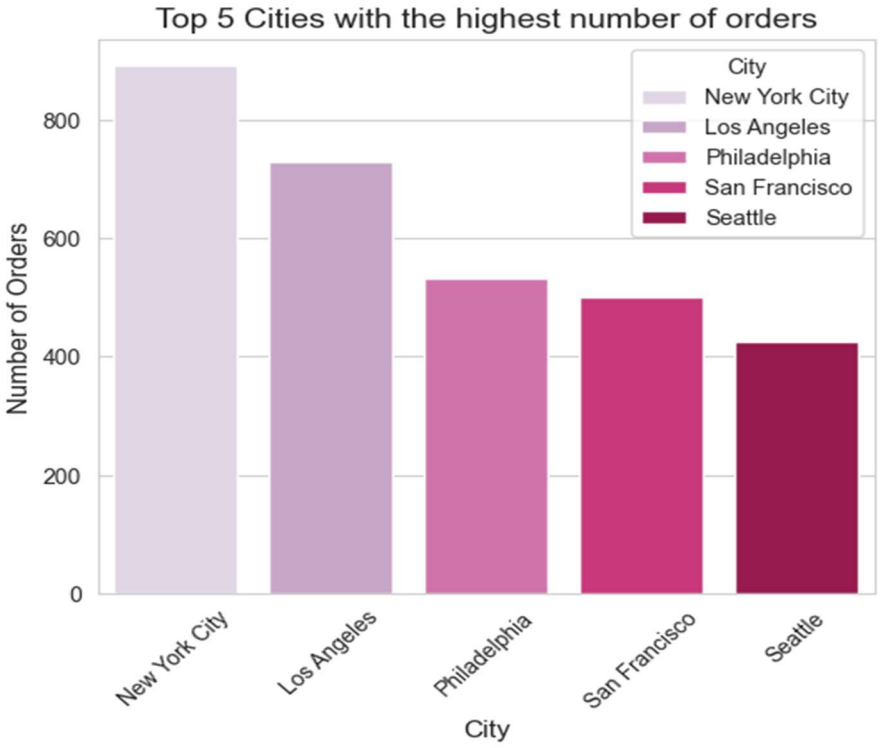


Figure 2: Top 5 Cities Bar Chart

Chart Description: Horizontal bar chart displaying the five cities with the highest number of orders. Chart uses a gradient color scheme to emphasize ranking hierarchy.

Analysis: Major metropolitan areas dominate order volume, with a particular concentration in coastal cities. This urban focus suggests that marketing efforts targeting large metropolitan areas yield the highest ROI.

5. PRODUCT CATEGORY ANALYSIS

5.1 Category Overview

Total Categories: 3 main categories

Total Sub-Categories: 17 product sub-categories

Category Breakdown:

Category	Orders	Percentage
Office Supplies	5,909	60.3%
Furniture	2,078	21.2%
Technology	1,813	18.5%

5.2 Sales by Category

Category	Total Sales
Technology	$836,154.03
Furniture	$741,999.80
Office Supplies	$683,383.10

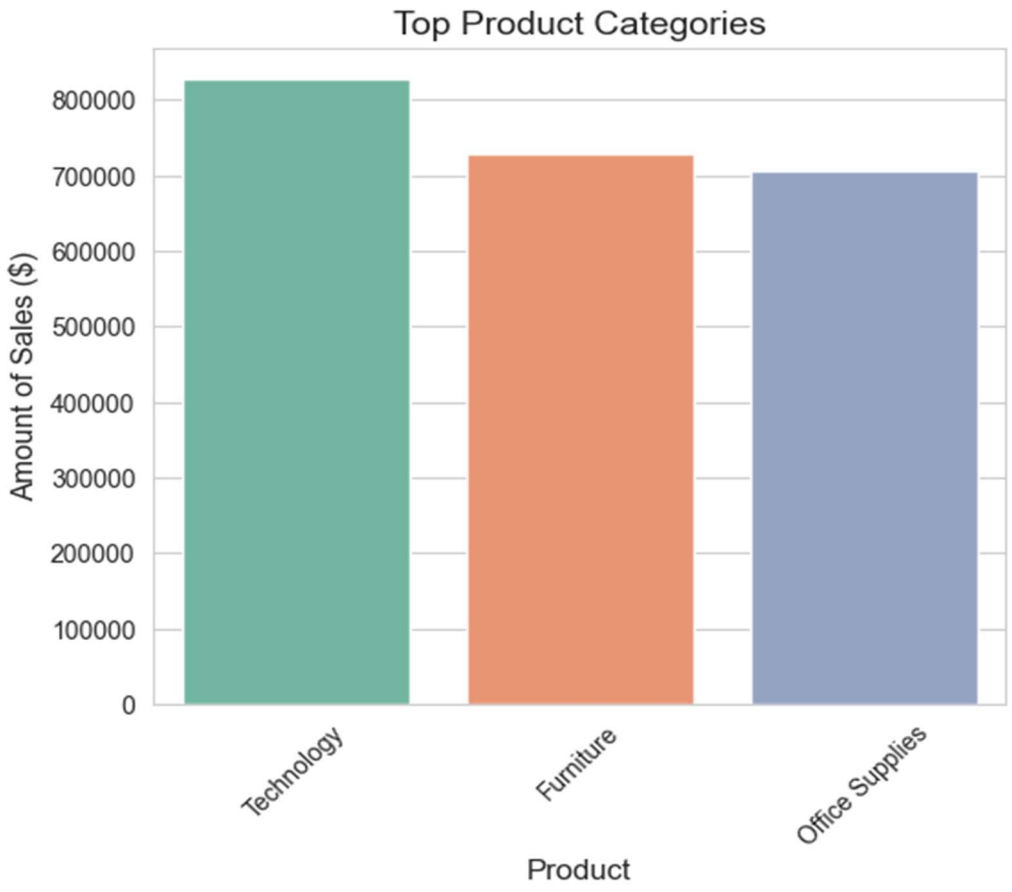


Figure 3: Category Sales Bar Chart

Chart Description: Bar chart showing total sales revenue by product category. Technology leads despite having fewer transactions, indicating higher per-transaction values.

Critical Finding: While Office Supplies accounts for 60% of **transaction volume**, Technology generates 37% of total **revenue** with only 18.5% of transactions. This indicates:

- Technology products have higher average sale values
- Office Supplies drives transaction frequency (consumables)
- Furniture provides balanced mid-range performance

5.3 Top 5 Selling Products

Highest Revenue Products

1. **Canon image CLASS 2200 Advanced Copier** - Premium office equipment

2. **Cisco TelePresence System EX90 Videoconferencing Unit** - High-value technology
3. **HON 5400 Series Task Chairs** - Bulk furniture purchases
4. **Hewlett Packard LaserJet 3310 Copier** - Essential office technology
5. **GBC DocuBind TL300 Electric Binding System** - Specialized equipment

Strategic Implication: Top revenue products are high-ticket items (copiers, telecommunications, office furniture) rather than high-volume consumables. This suggests that large individual transactions significantly impact revenue performance.

Highest Order Products

1. Staple envelope
2. Staples
3. Easy-staple paper
4. Avery Non-Stick Binders
5. Staple remover

Chart Description: Bar chart displaying the five individual products generating the highest number of orders over the analysis period.

Strategic Implication: Customers probably prefer these products cause they are cheap or cause of its usefulness.

6. CUSTOMER SEGMENTATION

6.1 Segment Distribution

Segment	Orders	Percentage
Consumer	5,101	52.1%
Corporate	2,953	30.1%
Home Office	1,746	17.8%

Analysis: Consumer segment dominates order volume, representing over half of all transactions. However, volume alone doesn't indicate value segment profitability requires average order value analysis.

6.2 Average Order Value by Segment

Segment	Average Order Value
Home Office	$243.40
Corporate	$233.15
Consumer	$225.07

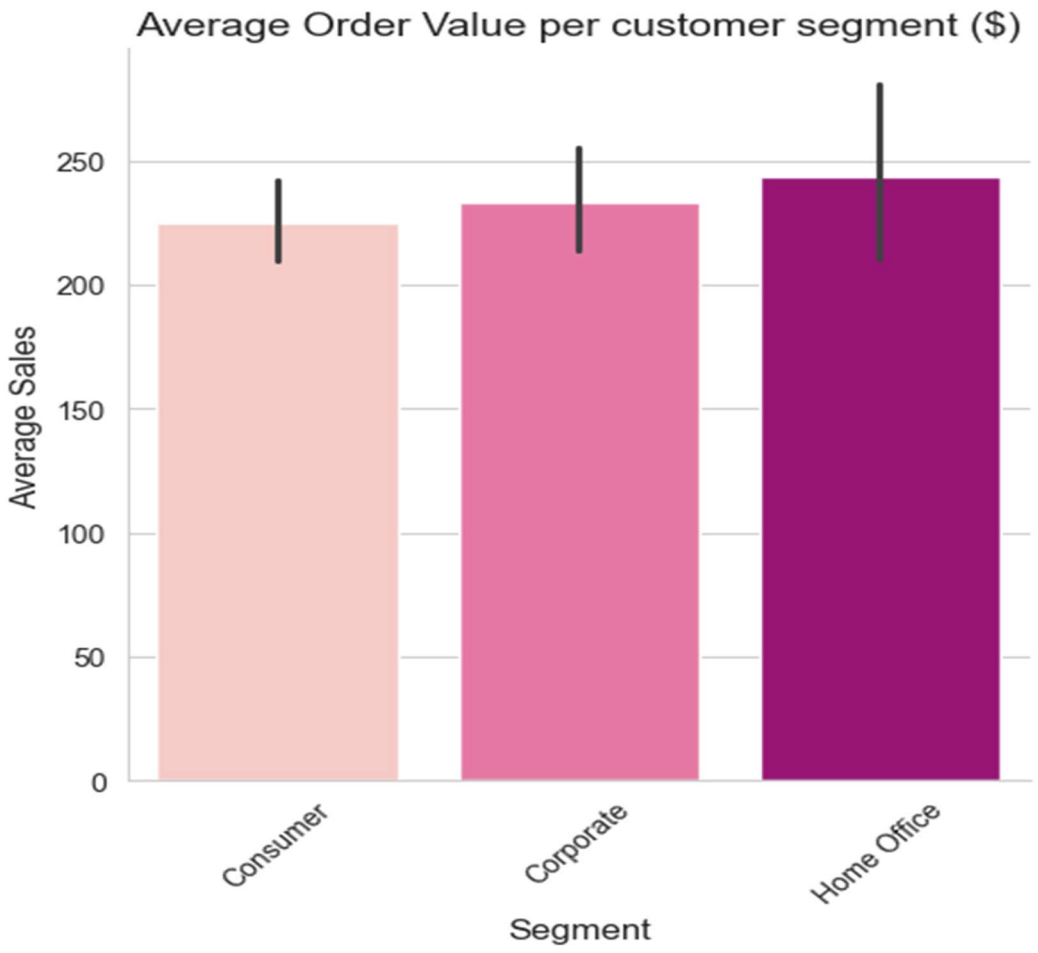


Figure 6: Average Order Value by Segment Bar Chart

Chart Description: Bar chart comparing average order values across the three customer segments, demonstrating that Hone Office customers generate the highest per-transaction revenue despite lower overall order volume.

Critical Finding: While Consumer segment leads in order **volume** (52%), Hone Office segment leads in order **value** ($234.22 avg). This indicates different purchasing patterns:

- **Corporate:** Mid-range purchasing behavior
- **Consumer:** Individual purchases, mix of product types, cheaper items
- **Home Office:** Bulk purchases, higher-value items

Strategic Implication:

Different segments require tailored approaches:

- Consumer → Drives Volume, flash sales, discounts, free-shipping are effective to increase AOV.
- Corporate → Bulk pricing, subscription
- Home Office → Premium products drives revenue, emphasize productivity and quality.

7. SHIPPING OPERATIONS ANALYSIS

7.1 Available Shipping Modes

Shipping Mode	Orders	Avg Delivery Days
Standard Class	5,859	5.0 days
Second Class	1,903	3.0 days
First Class	1,506	2.0 days
Same Day	532	0.0 days

Analysis: Standard Class shipping represents the high majority of orders (59.8%). Premium shipping options (First Class + Same Day) accounts for only 20.8% of shipments, suggesting either:

- Cost-conscious customer base
- Lack of urgency in most orders
- Potential underutilization of premium options in different regions

7.2 Shipping Time Analysis

Figure 7: Shipping Time Box Plot by Mode

Chart Description: Box plot visualization displaying the distribution of shipping times (in days) for each shipping mode. Shows median, quartiles, and outliers. Standard Class exhibits the widest variability with some deliveries extending to 7days, while premium options show tighter, more

consistent delivery windows. However, there is an outlier for the Same day option, this may be due to a public holiday, malfunction or lack of staff.

Key Findings:

- **Standard Class:** Average 5 days, but high variability (3-7 day range)
- **Second Class:** Consistent 3-day delivery
- **First Class:** Reliable 2-day delivery
- **Same Day:** True same-day performance (0 days)

Operational Insight: Standard Class shipping time (5 days average) exceeds typical industry standards (3-4 days), presenting an opportunity for operational improvement.

7.3 Regional Shipping Patterns

Figure 8: Shipping Mode by Region Grouped Bar Chart

Chart Description: Grouped bar chart showing the distribution of shipping modes across the four regions. Each region displays four bars representing the four shipping modes, enabling direct comparison of shipping preferences across regions.

Finding: No significant regional variations in shipping mode preference detected. Standard Class shipping dominates uniformly across all four regions, suggesting:

- Standardized company shipping policies nationwide
- Consistent customer preferences regardless of geography

Opportunity:
The uniformity presents an opportunity to test region-specific shipping promotions to determine if regional variations in shipping preferences exist when cost barriers are removed.

8. SALES DISTRIBUTION ANALYSIS

8.1 Overall Sales Distribution

Statistical Summary:

Metric	**Value**
Mean	$230.77
Median	$54.49
Mode	~$10-20 range
Standard Deviation	$626.65
Skewness	Positive (right-skewed)
Range	$0.44 - $22,638.48

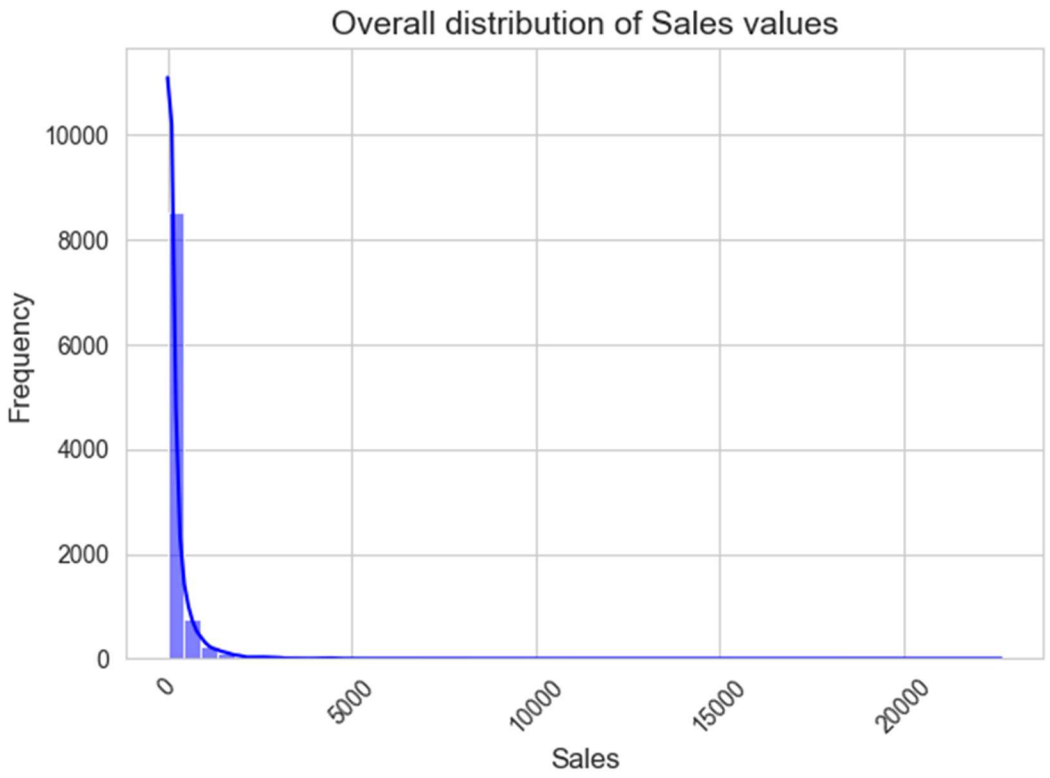


Figure 9: Sales Distribution by Histogram Plot

Interpretation: Highly right-skewed distribution with a long tail of high-value transactions. The median ($54.49) being significantly lower than the mean ($230.77) indicates most transactions cluster at lower values, with occasional large orders driving the average upward.

8.2 Outlier Analysis

Method: Interquartile Range (IQR) Analysis

Calculations:

- Q1 (25th percentile): $17.25
- Q3 (75th percentile): $210.61
- IQR: $193.36
- Lower Bound: $17.25 - (1.5 × $193.36) = -$272.79 (no lower outliers possible)
- Upper Bound: $210.61 + (1.5 × $193.36) = $500.64

Results:

- **Outliers Identified:** 1,145 transactions (11.7% of dataset)
- **Outlier Threshold:** Sales > $500.64
- **Highest Outlier:** $22,638.48

Outlier Characteristics:

- Predominantly **Technology** products (copiers, servers, telecommunications)
- Significant **Furniture** items (complete office setups, executive furniture)
- Almost exclusively **Corporate** and **Home Office** segments

Conclusion: Outliers represent important business opportunities rather than data anomalies. These high-value transactions account for a disproportionate share of total revenue and warrant dedicated account management strategies.

8.3 Sales Distribution by Category

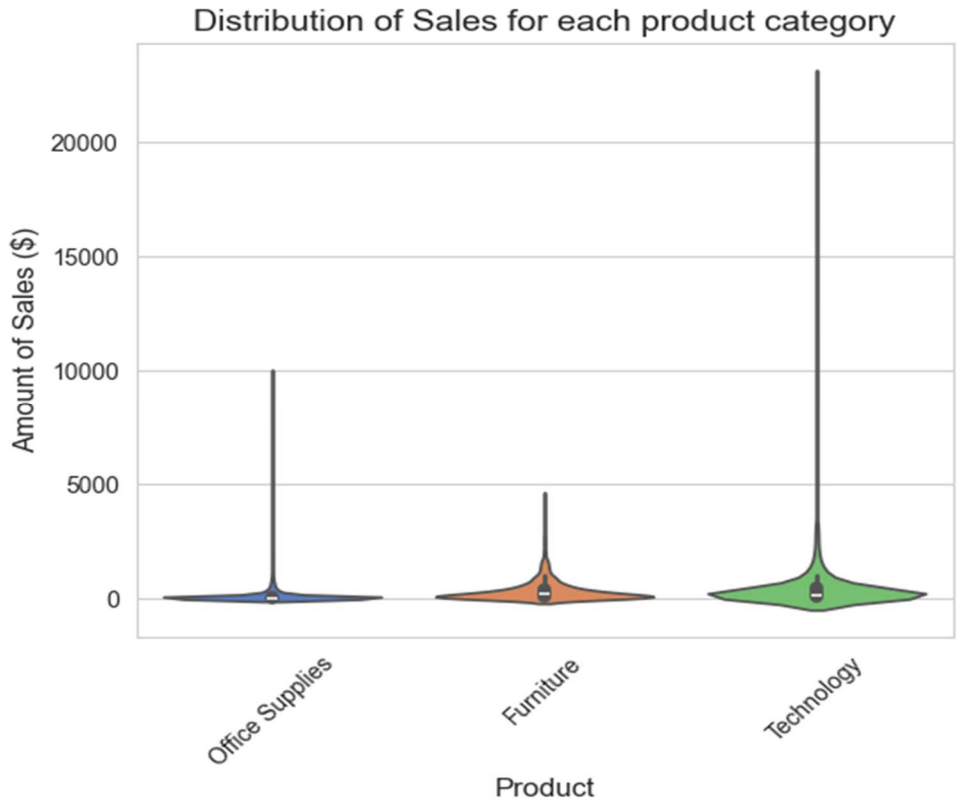


Figure 10: Sales Distribution by Category Violin Plot

Chart Description: Violin plot showing the full probability density distribution of sales values for each product category. Wider sections indicate higher concentration of transactions at that price point. Technology displays the widest distribution with numerous high-value outliers, Furniture shows moderate spread, and Office Supplies exhibits the tightest, most consistent pricing.

Category-Specific Insights:

Technology:

- Widest sales distribution ($0.44 - $22,638.48)
- Highest concentration of outliers
- Reflects mix of accessories ($10-50) and major equipment ($5,000+)
- Most variable category

Furniture:

- Moderate distribution range ($5 - $5,000 typical)
- Consistent mid-range pricing
- Few extreme outliers
- Predictable purchase patterns

Office Supplies:

- Tightest distribution ($1 - $200 typical)
- Fewest outliers
- Consumable nature drives consistent pricing
- Most predictable category

9. LIMITATIONS & SOLUTIONS

9.1 Overview of Data Limitations

During the exploratory data analysis, **two significant data gaps** were identified that limit the scope of strategic insights:

1. **RESOLVED:** Missing Postal Codes (11 records)
2. **UNRESOLVED:** Missing Business-Critical Columns (Profit, Order Status)

While the first issue was successfully addressed during analysis, the second represents a systemic data collection gap requiring stakeholder intervention.

9.2 CHALLENGE 1: Missing Profit Data

Issue Description

The project brief requested comprehensive **Profit Analysis** including:

- Calculate total profit for each order
- Determine average profit margin
- Identify top 5 orders with the highest profit

Critical Finding: The dataset contains **NO profit-related columns**:

- No "Profit" column
- No "Cost" column (would enable calculation: Profit = Sales - Cost)
- No "Profit Margin %" column (would enable calculation: Profit = Sales × Margin%)
- No ‘Discount’ column

Impact on Analysis

Cannot Perform:

- Profitability analysis by customer segment, product category or by geographic region
- ROI calculation for marketing initiatives
- Identification of high-profit vs. high-revenue customers
- Product mix optimization based on profit contribution

Current Limitation: All financial analysis focuses on **sales revenue** rather than **profitability**, which can lead to suboptimal strategic decisions. For example:

- A high-revenue customer may have low profit margins
- A low-revenue product may contribute higher profit margins
- Volume-driven strategies may not maximize profitability

Proposed Solutions

IMMEDIATE ACTION (Stakeholder Engagement Required):

Option 1: Retroactive Data Collection

- **Action:** Reach out to **Finance/Accounting Department**
- **Request:** Historical cost data of products sold (2015-2018)
- **Timeline:** 2-4 weeks for data extraction and validation
- **Result:** Calculate profit = Sales – Cost, as well as the discount that was applied for all transactions

Option 2: Product Cost Database

- **Action:** Contact **Inventory/Operations Management**
- **Request:** Standard cost database by Product ID

LONG-TERM SOLUTION (System Enhancement):

Action: Implement profit tracking in data collection systems

- **Owner:** IT Department / Data Engineering Team
- **Timeline:** Next system update cycle (Q2 2026)

1. Add "Unit Cost" field to order entry system
2. Add ‘Discount field’ to order entry system
3. Auto-calculate "Profit" field: (Sales - Cost)
4. Auto-calculate "Profit Margin %": ((Sales - Cost) / Sales × 100)
5. Make all three fields mandatory for order completion

- **Validation:** Implement data quality checks to prevent future gaps

9.3 CHALLENGE 2: Missing Order Status Data

Issue Description

The project brief requested analysis of:

- "What is the distribution of order statuses (e.g., delivered, pending)?"

Critical Finding: The dataset contains **NO order status column**:

- No "Order Status" field
- No "Delivery Status" field
- No "Fulfillment Status" field

Impact on Analysis

Cannot Analyze:

- Delivery success rates
- Pending order backlogs
- Cancellation patterns
- Order fulfillment efficiency
- Customer satisfaction related to fulfillment
- Late delivery frequency to know why they opt for a particular shipping mode
- Correlation between shipping mode and delivery success

Proposed Solutions

IMMEDIATE ACTION (Stakeholder Engagement Required):

Option 1: Shipping Carrier Data

- **Action:** Reach out to **Logistics/Shipping Department**
- **Request:** Delivery confirmation data from carriers

LONG-TERM SOLUTION:

Action: Add order status tracking to data warehouse

- **Additional Fields to Add:**
- Order Status (Pending/Shipped/Delivered/Cancelled)
- Delivery Date (actual)

- Delivery Delay (in days, if applicable)
- Return Status (if applicable)

9.4 CHALLENGE 3: Missing Shipping Cost Data

Impact on Analysis

Cannot Analyze:

- The correlation between shipping cost and Sales
- Shipping costs and Shipping mode
- Shipping cost & Profit

IMMEDIATE ACTION (Stakeholder Engagement Required):

Option 1: Shipping Carrier Data

- **Action:** Reach out to **Logistics/Shipping Department**
- **Request:** Delivery confirmation data from carriers

LONG-TERM SOLUTION:

Action: Add Shipping costs to data warehouse

10. KEY FINDINGS SUMMARY

10.1 Market Performance

- There has been increase in sales and orders in the company over the years.
- 6.2 orders per customer indicates good loyalty
- 60% of orders concentrated in West/East regions
- Almost Balanced revenue across the 3 product categories
- Top 5 cities drive disproportionate volume

10.2 Customer Insights

- Consumer segment accounts for 52% of orders
- Home Office segment has 3.5% higher average order value

10.3 Product Performance

- Technology category accounts for $836K, 37% of total
- Office Supplies 60% of orders
- Furniture (32.8% of revenue)
- Top 5 selling products are all premium equipment

10.4 Operational Performance

- Standard Class averages 5 days (above industry standard)
- Only 20.8% use First Class or Same Day
- No regional variations in shipping preferences

10.5 Data Quality Status

- Missing postal codes (11 records corrected to 05405)
- Zero duplicate records
- Profit data unavailable (requires stakeholder action)
- Discount unavailable
- Shipping cost unavailable
- Order status not tracked (limits fulfillment analysis)

11. ACTIONABLE RECOMMENDATIONS

11.1 PRIORITY 1: Address Data Gaps (Immediate)

Objective: Enable comprehensive profitability and operational analysis

Actions Required:

1. **Obtain Profit/Cost Data & Obtain Shipping Cost Data**

- Contact Finance Department for historical cost data
- Map product costs to transactions
- Calculate profit margins by segment, region, category
- Re-run analysis with profitability focus

2. **Integrate Order Status Data**

- Pull delivery status from fulfillment system
- Analyze delivery success rates
- Identify fulfillment bottlenecks
- Correlate status with customer segments

3. **Implement Long-Term Solution (Q2 2026)**

- Add profit tracking to order entry system
- Include order status in data warehouse

11.2 PRIORITY 2: Regional Expansion Strategy

- In regions and countries with low sales (South), the company should increase product marketing and advertising.

11.3 PRIORITY 3: Customer Segment Optimization

- Introduce Promotion strategies, discounts to Consumer segment to optimize profits
- Advertise with emphasis on durability, quality for Home Office customers.
- Corporate customers should be engaged with retention strategies.

11.4 PRIORITY 4: Shipping Efficiency

- Offer free shipping bonuses to customers from time to time.
- Tailor shipping modes by segment to balance cost e.g. prioritize fast shipping options for high value orders.

11.5 PRIORITY 5: Product Strategy

- Regular EDA on updated datasets
- Track KPIs: revenue growth, customer retention, regional balance
- Add order status field
- Track customer satisfaction scores
- Predict customer churn risk
- Price optimization models

12. CONCLUSION

The purpose of this report was to outline the procedures I used to analyze a Sales dataset. To begin, I performed some tasks like removing missing values from datasets, creating new columns, and changing the data type of columns (to_datetime). Then the article moves to the Exploration data Analysis where we explore questions at the business level related to our data, then give recommendations to make the data quality better and analyze better.