



Final EDA Report for Shopify Sales Data



Dataset Summary

- **Total Records:** 7,431
 - **Total Features:** 19 (original), later expanded after feature engineering
 - **Data Source:** Shopify Sales Excel file
-



Step 1: Data Overview

- **No duplicate rows** found.
 - **Missing values:**
 - **Product Id:** 11 missing
 - **Variant Id:** 4 missing
 - **All key numerical and categorical columns properly typed**
 - **Datetime column Invoice Date converted** to proper datetime format
-



Step 2: Data Cleaning

- Combined **Billing Address First Name** and **Last Name** into **Full Name**
 - Normalized city names using **.capitalize()**
 - Removed missing values or imputed if necessary
 - Verified and dropped 0 zero-value rows in Quantity/Price/Tax
-



Step 3: Univariate Analysis

- **Numerical Summary:**
 - Most orders have **Quantity = 1**
 - Prices (Subtotal/Total) are **right-skewed** — few high-value outliers
- **Boxplots & Histograms:**
 - Outliers confirmed in price and tax columns
 - Concentration in low price ranges
- **Categorical Summary:**
 - Most sales from **United States**
 - **USD** is the dominant currency
 - **Shopify Payments** is the most used gateway
 - Few dominant product types

Step 4: Bivariate Analysis

- **Correlation Matrix:**
 - **Subtotal Price**, **Total Price Usd**, and **Total Tax** are perfectly correlated ($r = 1.00$)
 - **Quantity** has a moderate positive correlation with other price-based fields
 - **Scatter Plots:**
 - Showed strong linearity between price-related columns
 - Quantity has a **non-linear** relationship with price
 - **Boxplots (Categorical vs Numerical):**
 - Certain product types have consistently higher values
-

Key Business Insights






1. ☒ Most orders are **small, single-item** purchases
 2. ☒ **High-value transactions** are rare but significantly impact revenue
 3. ☒ Business is **geographically concentrated** (mainly U.S. customers)
 4. ☒ Revenue is **driven by few product types** and gateways
 5. ☒ **Tax and Total Price** are strongly tied — tax likely a fixed percentage
-

Step 5: Handling Missing & Duplicate Data

- 11 missing values in **Product Id**, 4 in **Variant Id** — handled
 - **0 duplicate rows**
 - No zero-values in **Quantity**, **Price**, or **Tax** columns
-

Step 6: Feature Engineering

New features created:

-  **Year, Month, Weekday, Hour** — extracted from **Invoice Date**
 -  **Revenue per Unit** = **Total Price** / **Quantity**
 -  **High Tax Order** = if tax > 95th percentile
 -  **Revenue Category** = segmented into bins: Very Low → Very High
 -  **Country_Product** = combined location + product type
-

Step 7: Encoding & Transformation

- **Label Encoding** applied to binary categories
- **One-Hot Encoding** used for multi-class variables (e.g., product type, city)
- **Min-Max Scaling** done for numerical features
- Final dataset shape: **(7431, 27211 columns)** after one-hot encoding

Step 8: Correlation & Feature Selection

- Identified highly correlated pairs:
 - Total Price ~ Subtotal Price ~ Total Tax
 - Product Id ~ Variant Id
- Final cleaned dataset exported to Cleaned_Shopify_Sales.csv

Final Output Ready For:

☒ Machine Learning ☒ Dashboards (Power BI/Tableau) ☒ Reporting & Decision Making ☒ Portfolio Projects
