Marketing-Oriented Outlier Analysis

Identifying Anomalies in Retail Transaction Data

Prepared by: Junior Data Scientist

Date: July 2025

Problem Statement

In this project, I performed advanced outlier detection on online transactional data to ensure data quality for marketing analytics. Using techniques like IQR, MAD, and Isolation Forest, I filtered erroneous or extreme records that could skew customer insights. Then created an isolated dataset for the outliers to view later and understand how the different models are flagging outliers. This workflow enhances downstream marketing efforts such as segmentation, campaign targeting, price optimization, and lifetime value prediction by ensuring clean and interpretable input data.

Exploratory Data Analysis

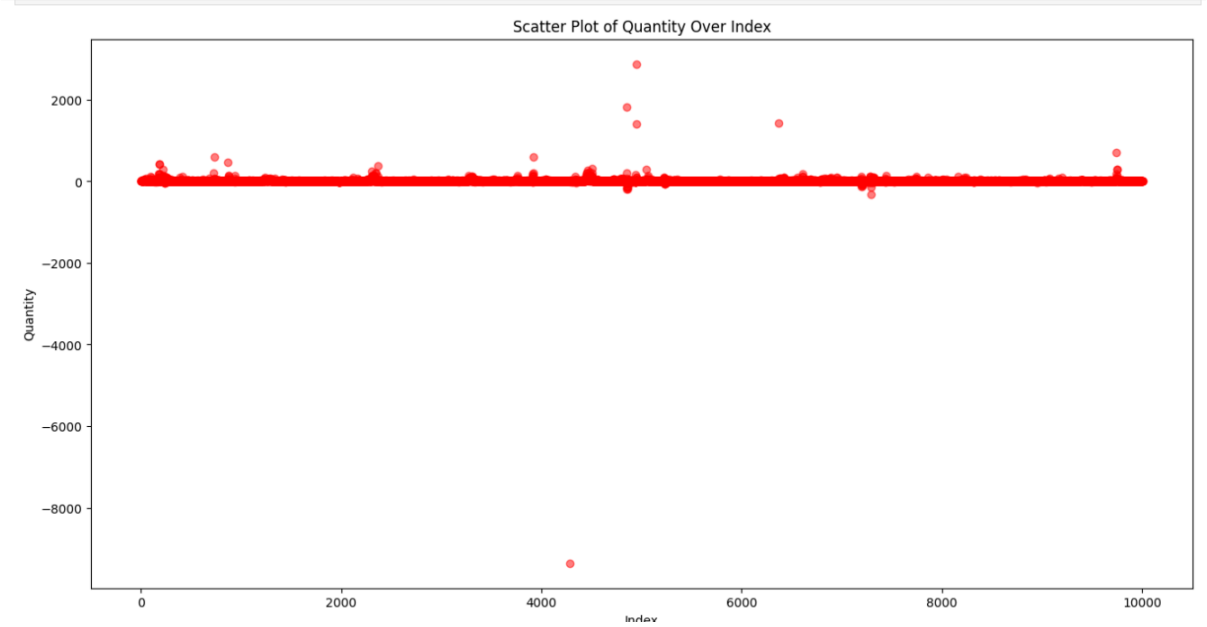
To begin uncovering hidden patterns and irregularities in the dataset, extensive exploratory data analysis (EDA) was performed on two key features: `UnitPrice` and `Quantity`, which directly impact revenue and marketing insights.

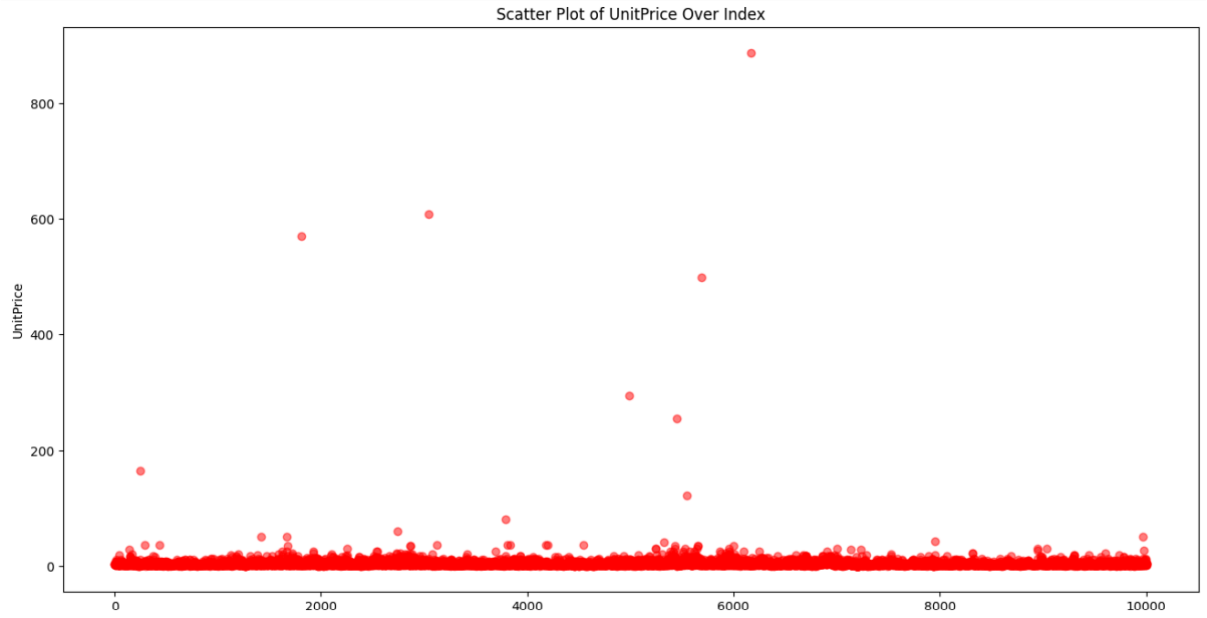
Histograms were used to assess the overall distribution shape of both variables. These visualizations revealed highly right-skewed distributions, with most values clustered near the lower end and a long tail of unusually high values. This suggested the presence of bulk purchases, pricing errors, or one-time promotional transactions — all of which could distort marketing models such as CLV or segmentation if left unaddressed.

Scatter plots were then plotted across transaction index ranges to detect unusual spikes. These plots exposed abrupt jumps and dense vertical bands, pointing to repetitive extreme values (e.g., same price or quantity repeated), possibly due to automated systems, catalog errors, or supplier-side batch orders.

Boxplots helped highlight the magnitude and frequency of these outliers. With long whiskers and numerous data points plotted far outside the interquartile range (IQR), the boxplots confirmed that both features had extreme variability not explained by regular business behavior.

Together, this exploratory phase provided a data-driven foundation for choosing outlier detection methods and justified the need for cleaning the dataset before applying any marketing analytics or modeling.





Outlier Detection and Validation through Regression

To ensure data integrity before applying marketing analytics, three outlier detection techniques were implemented. First, the Interquartile Range (IQR) method flagged extreme values based on the spread of the middle 50% of the data, effectively identifying simple univariate anomalies. Next, the Median Absolute Deviation (MAD) method was employed using the PyOD library, offering a robust alternative well-suited for skewed distributions where traditional measures like standard deviation fall short. Finally, a more advanced Isolation Forest, an unsupervised machine learning algorithm, was applied to detect multivariate outliers by isolating points that behave differently across several features simultaneously.

To assess how each technique impacted data quality, a linear regression model was trained to predict `TotalPrice` using relevant variables such as `Quantity`, `UnitPrice`, and country indicators. The models were evaluated using Root Mean Squared Error (RMSE) after filtering the dataset based on each outlier method. The best predictive performance was achieved when 25% of the data was excluded as outliers using Isolation Forest, suggesting these anomalies were significantly degrading model accuracy and obscuring real marketing patterns.

Final Insights

Dimensionality reduction using t-SNE was applied to transform high-dimensional retail transaction data into a 2D visual space. The resulting scatterplots, colored by features like `Quantity`, `UnitPrice`, and `TotalPrice`, revealed distinct transaction patterns and natural behavioral groupings that could support segmentation, pricing, and campaign design. However, the use of one-hot encoded `Country` variables introduced significant noise and distortion. Because t-SNE is sensitive to sparse and binary inputs, the encoded countries interfered with cluster integrity, limiting the interpretability of geographic patterns.

One of the more subtle but frustrating challenges in the project was working with these one-hot encoded variables. While encoding is required to make categorical data usable in machine learning models, it complicates downstream tasks like dimensionality reduction and interpretation. Once split into binary columns, the categorical meaning is lost — and reconstructing the original "Country" column after modeling is non-trivial, especially when some observations have been filtered out or when feature importance is assessed across multiple dummy variables. This made it difficult to communicate geographic insights clearly to a marketing audience.

From a marketing perspective, the project reinforces how clean, well-structured data underpins all reliable insights. Detecting and removing outliers is not just technical housekeeping — it's foundational for accurate customer segmentation, trustworthy revenue modeling, and effective targeting strategies. The measurable improvement in regression performance after outlier removal (as shown by lower RMSE) proved that anomaly filtering can have a direct, positive impact on business forecasting.

Ultimately, this project illustrates how thoughtful data preprocessing, interpretability-conscious modeling, and human-readable storytelling can bridge the gap between raw transaction data and actionable marketing intelligence. Whether the goal is to improve targeting, optimize pricing, or reduce churn, ensuring clean, high-quality data is the first non-negotiable step.