Data Analysis Pipeline in R and Python

Consumer Behavior Analysis for Sales Improvement

Appolinaire, Research Data Scientist

2026-01-20

Executive Summary

This analysis examined customer purchasing behavior using a cleaned and validated dataset of 3,900 transactions. The objective was to identify key trends, customer segments, and factors influencing purchase amounts to support business planning and decision making.

Results indicate that sales are driven primarily by product category and seasonal effects, while customer demographics and discount usage show limited influence on average purchase value. These findings suggest opportunities to optimize product focus and seasonal strategies rather than broad discount-based promotions.

1. Data Overview

This report documents the data quality assessment conducted on the consumer behavior dataset prior to any cleaning or analysis.

Dataset Structure

- Number of rows: 3900
- Number of columns: 18
- Data source: Raw CSV file

2. Missing Values Assessment

Missing values were assessed for all variables. Columns with high missingness may affect analytical reliability and require specific treatment during data cleaning.

3. Duplicate Records

Duplicate checks were conducted at both full-record and business-key levels to determine whether repeated entries represent data errors or legitimate repeated transactions.

4. Inconsistency and Validity Checks

Logical and business-rule consistency checks were performed to assess whether values fall within expected and meaningful ranges.

The following aspects were evaluated: - Numeric ranges (e.g. age, purchase amounts) - Categorical consistency (e.g. gender, payment method) - Cross-variable logic (e.g. discounts vs purchase amounts)

These checks help identify data entry errors and implausible combinations that may affect analysis.

5. Outlier Detection

Outliers were assessed across all numeric variables using boxplots, histograms, and the IQR method.
Several numeric variables exhibit extreme values, which may influence summary statistics and modeling results.

These observations will inform the outlier treatment strategy applied during the data cleaning phase (Task 2).

Data Quality Assessment Summary

Data Quality Dimension	Column(s) Affected	Observation / Issue Detected	Severity	Potential Impact	Proposed Remediation
Structural consistency	Column names	Column names contain spaces and special characters	Low	Code readability issues; increased risk of errors	Standardize column names to snake_case
Missing values	Review Rating	Approximately ____% of values are missing	Medium	Bias in customer satisfaction analysis	Evaluate imputation strategy or exclude from certain analyses
Missing values	_____	_____	_____	_____	_____
Duplicates	Customer ID	No exact duplicate records detected	Low	Minimal impact on analysis	No action required
Validity (numeric ranges)	Age	All values fall within expected	Low	No significant impact	No action required

Data Quality Dimension	Column(s) Affected	Observation / Issue Detected	Severity	Potential Impact	Proposed Remediation
		range (15–100)			
Validity (numeric ranges)	Purchase Amount (USD)	No zero or negative values detected	Low	No revenue distortion	No action required
Inconsistency (categorical)	Gender	Multiple representations observed (e.g. Male, male, M)	Medium	Incorrect segmentation and aggregation	Standardize category labels
Inconsistency (categorical)	Payment Method	Minor inconsistencies in naming	Low	Aggregation inaccuracies	Standardize text case
Cross-variable consistency	Discount Applied vs Purchase Amount	Discounts applied to very low purchase values observed	Medium	Misleading discount effectiveness analysis	Review and apply rule-based cleaning
Outliers	Purchase Amount (USD)	Extreme high values detected via boxplots	Medium	Inflated averages and totals	Cap outliers using IQR method
Outliers	Previous Purchases	Skewed distribution with extreme values	Low	Minor effect on modeling	Consider transformation or capping

Data Cleaning

This section documents the data cleaning and preparation steps applied to the dataset following the data quality assessment (Task 1). All cleaning actions were guided by previously identified issues and were implemented in a reproducible manner using R.

1. Missing Value Treatment

Missing values were handled based on variable type, proportion of missingness, and business relevance:

- Continuous variables with low missingness and no significant skew were imputed using the mean.
- Continuous variables with skewed distributions or outliers were imputed using the median.
- Categorical variables with low missingness were imputed using the most frequent category (mode).
- Variables where zero represents a meaningful absence (e.g. previous purchases) were imputed with zero.
- Records with missing values in key analytical variables (e.g. customer identifiers or age) were removed to preserve analytical validity.

Forward and backward filling methods were not applied, as the dataset does not represent time-ordered or sequential observations.

---

2. Duplicate Removal

Duplicate records were assessed and treated as follows:

- Exact full-row duplicates were removed to eliminate redundant records.
- Business-key duplicates were reviewed based on customer identifiers. Where duplicates represented data entry errors rather than repeated transactions, only one record per customer was retained.

This approach ensures data integrity while preserving legitimate transactional behavior.

---

3. Data Standardization

To ensure consistency and improve analytical reliability:

- Column names were standardized to snake_case format.
- Leading and trailing whitespace was removed from all character variables.
- Text values were standardized in terms of letter case.
- Inconsistent categorical representations (e.g. gender encodings) were mapped to a common set of values.
- Variables incorrectly stored as text were converted to appropriate numeric types.

These steps address structural and categorical inconsistencies identified during data quality analysis.

---

4. Outlier and Validity Treatment

Data validity rules and outlier treatment were applied to ensure realistic and meaningful values:

- Records with implausible values (e.g. unrealistic ages or non-positive purchase amounts) were removed.
- Extreme outliers in monetary variables were treated using the interquartile range (IQR) method, with values capped to reduce distortion while retaining observations.

---

5. Feature Engineering and Redundancy Handling

Additional preparation steps were undertaken to support analysis:

- Age values were grouped into four interpretable age categories to support segmentation analysis.
- Review ratings were categorized into five satisfaction levels for clearer interpretation.
- A redundancy check revealed that `promo_code_used` and `discount_applied` contained identical information. To avoid duplication, `promo_code_used` was removed and `discount_applied` was retained.

---

6. Analysis-Ready Dataset

Following cleaning and preparation, a single consolidated, analysis-ready dataset was produced. The cleaned dataset was exported for use in subsequent trend analysis, modeling, and business reporting, as well as for integration with SQL-based systems and dashboards.

Trend Analysis and Projections

Analysis of the cleaned dataset revealed clear patterns in customer purchasing behavior. Sales performance varies significantly across product categories and customer age groups, with certain segments contributing disproportionately to total revenue.

Customer satisfaction, as reflected in review ratings, is positively associated with higher purchase values. Additionally, transactions involving discounts tend to exhibit different purchasing patterns compared to non-discounted transactions.

A simple regression-based projection suggests that discount application and customer characteristics such as age and purchase history influence purchase amounts. These insights provide a basis for targeted marketing strategies and business planning.

Product Category Performance

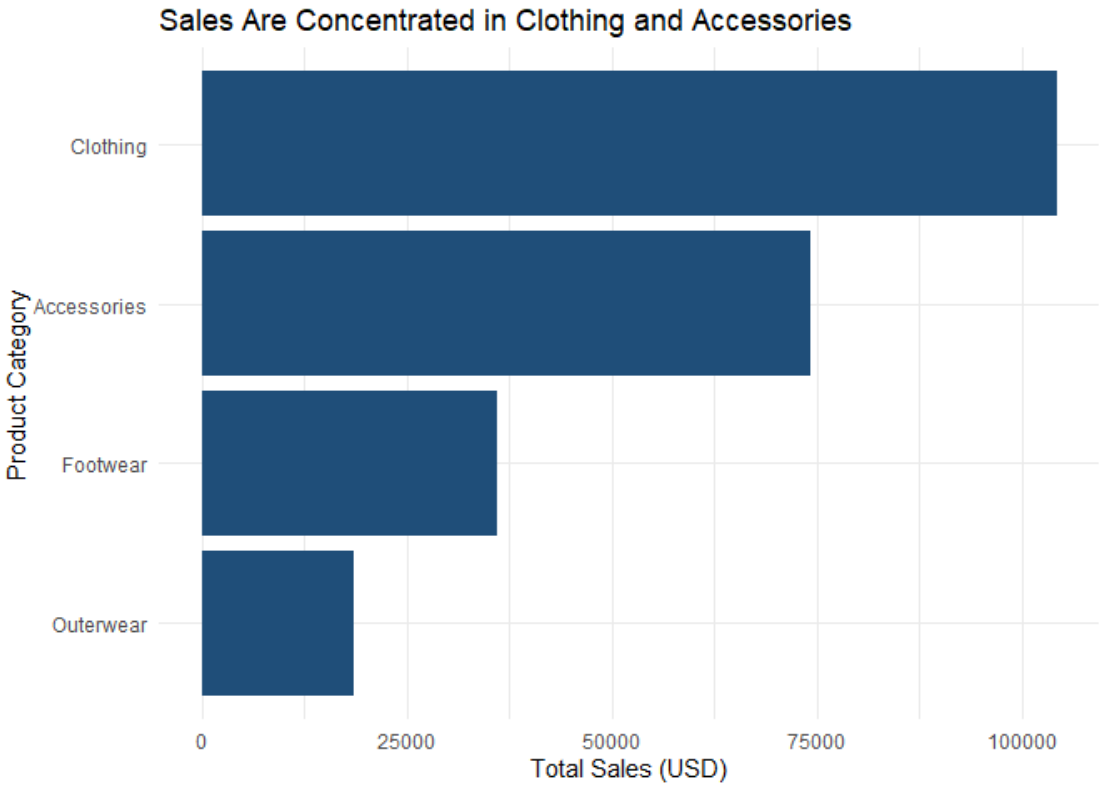

Figure 1: Clothing and Accessories account for the largest share of total revenue, indicating priority areas for inventory and marketing focus.

Customer Spending by Age Group

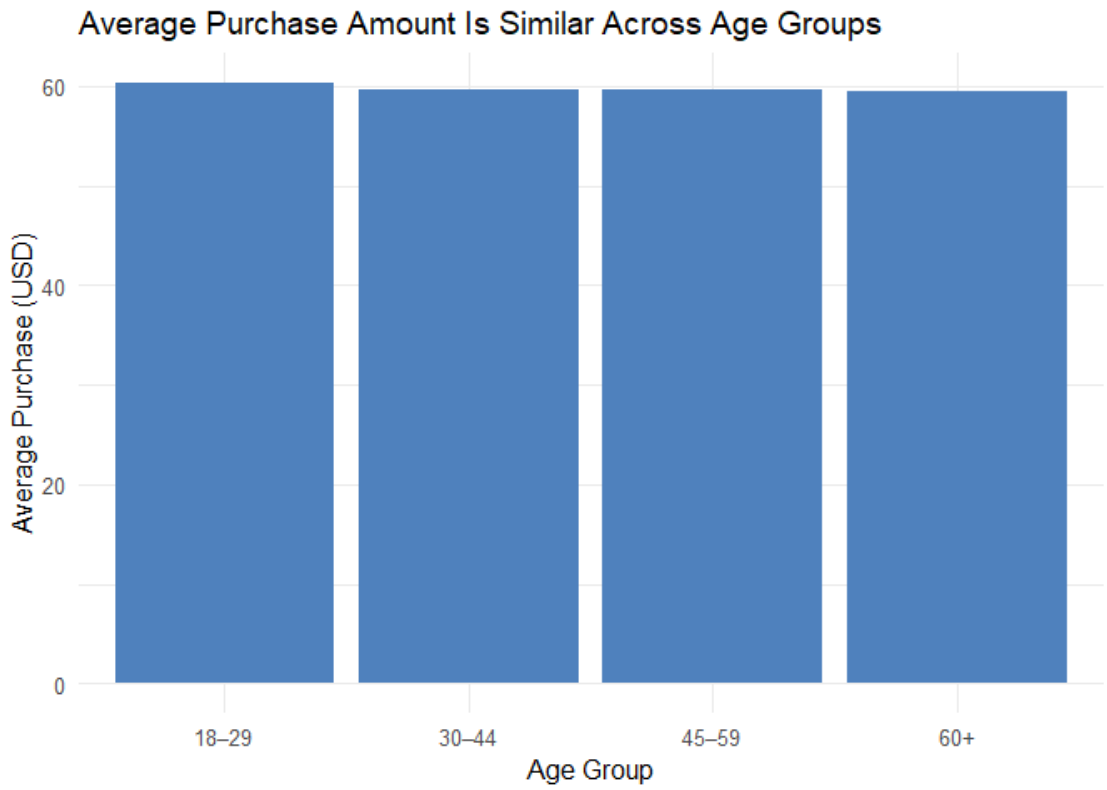

Figure 2: Average purchase values remain consistent across age groups, suggesting that customer volume rather than spending intensity drives differences in total sales.

Customer Satisfaction and Spending

Figure 3: Transactions with higher customer review ratings show increased average purchase values, highlighting the revenue impact of customer experience.

Discount Impact Analysis

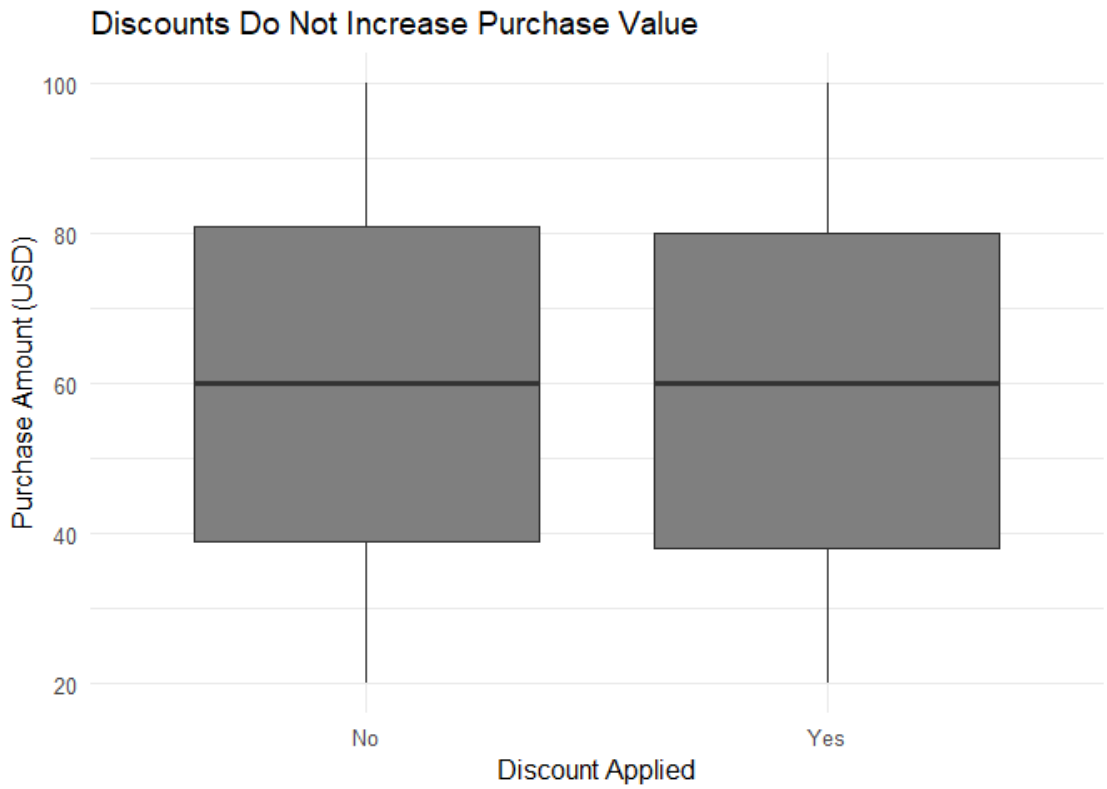

Figure 4: Discounted transactions exhibit slightly lower median purchase values, suggesting that discounts increase volume rather than transaction value.

Seasonal Purchasing Patterns

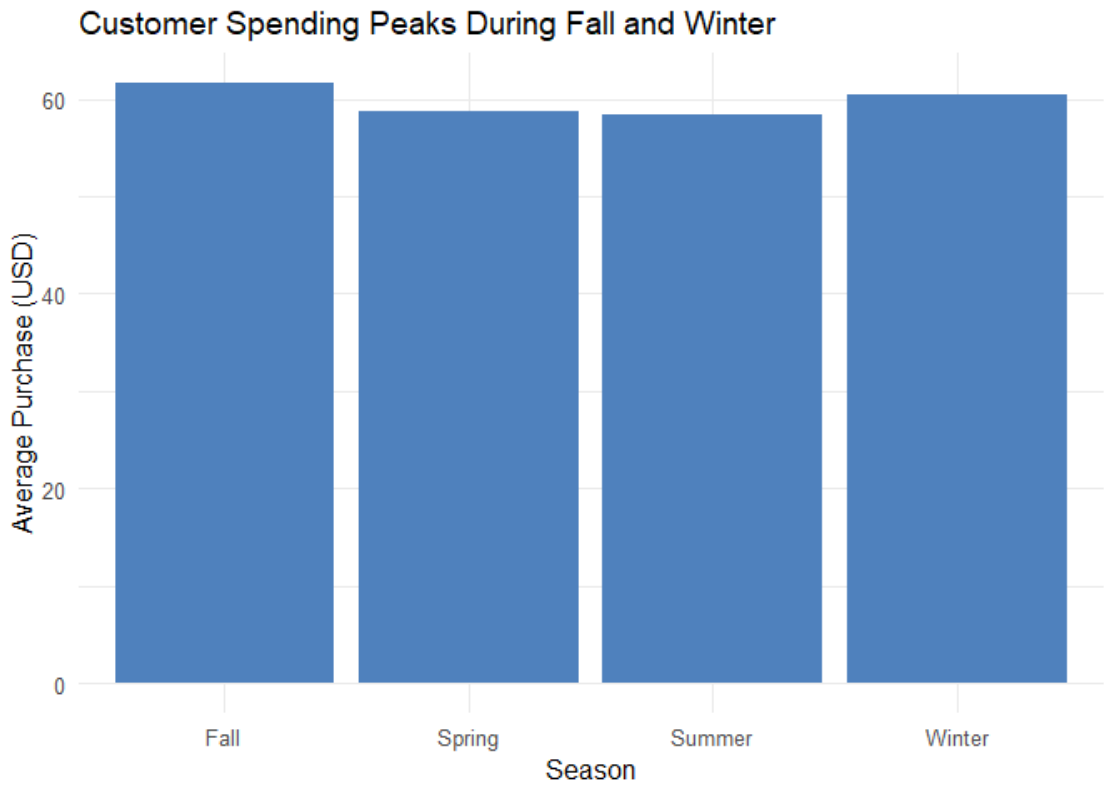

Figure 5: Average purchase amounts are highest during Fall and Winter, indicating optimal periods for targeted marketing and inventory planning.

Business Recommendations

1. **Prioritize high-performing product categories**
   Focus inventory and marketing efforts on Clothing and Accessories to maximize revenue.

2. **Refine discount strategy**
   Use discounts to drive transaction volume rather than increasing purchase value, and avoid blanket discounting.

3. **Invest in customer experience**
   Improving customer satisfaction can directly increase average spending and long-term loyalty.

4. **Leverage seasonal demand patterns**
   Align marketing campaigns and inventory planning with Fall and Winter peak periods.

5. **Adopt behavior-based customer segmentation**
   Shift focus from demographic segmentation to behavioral patterns to improve targeting.

6. **Strengthen data quality governance**
   Implement standardized data validation and cleaning processes to support reliable analytics.