Customer Shopping Behavior Analysis

By Agrim Sharma

Dhirubhai Ambani University

Data Analysis Project Report *(for overview purposes)*

CSBA- GitHub

Contact Info

College ID: 202301087

Reach at: https://www.linkedin.com/in/agrimtdk/

Email at: agrimsharma20092005@gmail.com

GitHub Profile: https://github.com/agrimtdk

Abstract

This report analyzes customer shopping behavior using a structured dataset integrated with Python-based exploratory analysis and SQL-based data persistence. The objective is to identify patterns in purchasing behavior across demographic segments, product

categories, and transactional characteristics, with an emphasis on data quality, feature engineering, and interpretable insights. The analysis is descriptive in nature and does not attempt causal inference.

1. Problem Definition and Data Overview

The purpose of this analysis is to understand how customer demographics and purchasing attributes relate to spending behavior and review ratings. The dataset contains individual-level customer transaction records, including age, gender, product category, purchase amount, review rating, discount usage, and purchase frequency indicators. The data was sourced from a flat CSV file and subsequently processed using Python before being loaded into a relational database for structured querying.

Initial inspection showed a moderate-sized dataset with mixed data types and limited missingness. Missing values were primarily observed in the review rating field and were treated systematically to preserve analytical validity.

2. Data Cleaning and Feature Engineering

Data preprocessing was conducted using pandas. Column names were standardized to lowercase to ensure consistency across Python and SQL environments. Missing review ratings were imputed using the median rating within each product category, a conservative strategy that preserves category-level distributional properties while avoiding mean distortion.

Additional derived features were introduced to enhance analytical resolution. Customers were segmented into age groups using quartile-based binning, allowing comparison across relative age cohorts rather than absolute ages. Purchase frequency, originally stored as categorical text, was mapped to an ordinal numeric representation reflecting increasing purchasing regularity. A logical consistency check confirmed that discount application and promotional code usage were perfectly aligned, allowing the redundant promotional code column to be removed without information loss.

3. Exploratory Analysis

Descriptive statistics were computed to summarize central tendencies and dispersion across numerical variables, including age and purchase amount. The analysis indicated variability in spending behavior across both age groups and product categories. Category-level differences were particularly evident in review ratings and purchase amounts, suggesting heterogeneous customer expectations and price sensitivity.

Age group segmentation revealed differences in purchasing frequency and average spend, with middle quartiles generally exhibiting higher transaction regularity. However, these patterns are correlational and should not be interpreted as causal without further modeling.

4. Database Integration and SQL Layer

Following preprocessing, the cleaned dataset was loaded into a PostgreSQL database using SQLAlchemy. This step ensures reproducibility, scalability, and compatibility with downstream analytics workflows. Basic SQL validation queries were executed to confirm successful ingestion, table availability, and database connectivity.

The SQL layer is designed to support aggregation and filtering operations consistent with the exploratory analysis performed in Python. While advanced analytical queries were not the focus of this project, the database structure enables extension to more complex business intelligence use cases.

5. Limitations

This analysis is subject to several limitations. First, the dataset represents a snapshot in time and lacks temporal depth, preventing trend or cohort analysis. Second, imputation of missing review ratings, while methodologically reasonable, introduces assumptions that may not hold uniformly across customers. Third, the absence of behavioral context such as income, location, or marketing exposure constrains interpretability.

6. Conclusion

The project demonstrates a complete data analysis pipeline, from raw data ingestion and cleaning to exploratory analysis and database integration. The findings highlight meaningful variation in customer shopping behavior across demographic and product dimensions, while maintaining analytical transparency and methodological discipline. Future work could extend this analysis through predictive modeling, time-series data, or experimental validation.