Volume 10, Issue 3, March – 2025 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1342

Retail Refine: Enhancing Retail Transaction Data for Advanced Analytics

Samir Pandey[1]; Ami Shah[2]

[2]Assistant Professor,
[1,2]Department of Computer Science and Engineering, Parul University Vadodara, Gujarat, India

Publication Date: 2025/04/01

Abstract: In the era of big data, high-quality data is essential for accurate analysis and decision-making. This paper explores the process of data cleaning and preparation for advanced analytics, focusing on techniques such as handling missing values, outlier detection, data transformation, and feature engineering. A case study is presented using a dataset to perform time series analysis, cohort segmentation, churn analysis, and customer segmentation. The goal is to enhance data reliability and usability for machine learning and predictive modeling.

Keywords: *Data Cleaning, Data Preparation, Time Series Analysis, Cohort Segmentation, Churn Analysis, Outlier Detection, Feature Engineering.*

How to Cite: Samir Pandey; Ami Shah. (2025). Retail Refine: Enhancing Retail Transaction Data for Advanced Analytics. *International Journal of Innovative Science and Research Technology*, 10(3), 1668-1669. https://doi.org/10.38124/ijisrt/25mar1342.

I. INTRODUCTION

Data quality is a critical factor in the success of data-driven applications. Raw datasets often contain missing values, inconsistencies, duplicate records, and irrelevant information, which can negatively impact analytical models and business decisions. Poor data quality can lead to inaccurate insights, flawed forecasting, and inefficient resource allocation. Therefore, it is essential to establish a robust data cleaning and preparation pipeline to enhance the quality and usability of data before it is used for analytics or machine learning models.

This paper discusses systematic approaches to data cleaning and preparation, ensuring high-quality inputs for advanced analytics tasks such as time series forecasting and customer behavior analysis. By leveraging automated and scalable data preprocessing techniques, businesses can unlock valuable insights, improve decision-making, and optimize operational efficiency. The methods explored in this study are particularly relevant for industries dealing with large-scale transactional data, such as retail, finance, and e-commerce, where data accuracy directly impacts revenue and customer satisfaction.

II. RELATED WORK

Numerous studies emphasize the importance of data cleaning in data science. Techniques like mean imputation, data interpolation, and anomaly detection have been widely used. Existing solutions, such as OpenRefine and Pandas in Python, provide functionalities for structured data cleaning, but challenges remain in automating and optimizing the process for large-scale datasets.

III. DATA CLEANING TECHNIQUES

A. Handling Missing Values

- Imputation using mean, median, or mode.
- Forward/Backward filling for time series data
- Removing rows or columns with excessive missing values

B. Outlier Detection

- Z-score and IQR-based outlier removal.
- Winsorization for extreme values
- Anomaly detection using machine learning models

C. Data Transformation

- Normalization and standardization
- Encoding categorical variables
- Log transformation for skewed data

D. Feature Engineering

- Creating new variables from existing data
- Extracting time-based features (e.g., day of the week, seasonality)
- Dimensionality reduction techniques (PCA, LDA)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25mar1342

IV. ADVANCED ANALYTICS APPLICATIONS

- **Time Series Analysis:** Trend and seasonality decomposition. Moving averages and exponential smoothing. Forecasting with ARIMA and Prophet models.
- **Cohort Segmentation**: Grouping users based on sign-up behavior. Analyzing retention trends and customer lifetime value (CLV).
- **Churn Analysis**: Identifying factors contributing to customer churn. Predictive modeling for churn probability. Retention strategy recommendations
- **Top Customer Analysis**: Analyzing high-value customers based on purchase behavior. RFM (Recency, Frequency, Monetary) segmentation. Targeted marketing strategies.

➢ *Performance and Scalability*

The data cleaning pipeline was optimized using vectorized operations in Pandas, ensuring efficient processing of large datasets. Parallel computing and batch processing techniques were implemented to handle scalability issues. Data preprocessing steps were benchmarked to measure improvements in model accuracy and execution time.

➢ *Future Enhancements*

- Automated Data Cleaning Pipelines using AI/ML models.
- Real-Time Data Preprocessing for streaming applications.
- Integration with Cloud Services for scalable data preparation.

V. CONCLUSION

Effective data cleaning and preparation are crucial for advanced analytics. This paper demonstrated various techniques to improve data quality and optimize analytical processes. Implementing robust preprocessing techniques ensures better model performance, leading to more reliable insights in business and research applications.

By utilizing automated and scalable data cleaning methods, organizations can significantly reduce the risk of inaccurate decision-making caused by poor-quality data. The techniques discussed in this paper provide a strong foundation for improving data integrity, making data more actionable and reliable. Furthermore, the integration of machine learning models for anomaly detection and imputation can enhance the automation of data preparation processes.

As industries continue to rely heavily on data-driven decision-making, establishing effective data-cleaning frameworks will remain a priority. Future advancements in AI-driven data cleaning and real-time preprocessing will further enhance the capabilities of analytics pipelines. Ensuring high-quality data not only improves model accuracy but also drives better business intelligence and operational efficiency across various domains, particularly in retail, finance, and e-commerce sectors.

The methodologies outlined in this study pave the way for more sophisticated data preparation approaches, bridging the gap between raw data and actionable insights. With continued research and innovation in this space, organizations can leverage clean, structured, and well-prepared data to gain a competitive edge in the rapidly evolving digital landscape.

REFERENCES

[1]. Wes McKinney, "Python for Data Analysis," O'Reilly Media, 2017.
[2]. Hastie, T., Tibshirani, R., & Friedman, J., "The Elements of Statistical Learning," Springer, 2009.
[3]. J. Han, M. Kamber, & J. Pei, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2011.
[4]. Kaggle Datasets, https://www.kaggle.com/
[5]. Prophet Forecasting Model, https://facebook.github.io/prophet/