CHANDIGARH UNIVERSITY

Discover. Learn. Empower.

CASE STUDY

ON

Data Cleaning in E-Commerce Customer Analytics

Submitted By

Name: Gurman Garg

UID- 24MCI10248

Class- MCA(AI-ML)

Section- 24MAM-2B

Submitted To

Dr. Arun kumar

CHANDIGARH UNIVERSITY

Discover. Learn. Empower.

Case Study: Data Cleaning in E-Commerce Customer Analytics

Problem statement

An e-commerce company, XYZ, aims to enhance its customer experience and improve sales through data-driven insights. The company collects vast amounts of data from customer transactions, website interactions, and feedback surveys. However, as the data grows, inconsistencies, duplicates, and errors have emerged, into leading to unreliable analytics.

Take dataset from any source, clean it by applying different data cleaning techniques whatever required that along with the explanation.

Rubrics for Relevant evaluation:

1. **dataset – 2 marks**
2. **Novel dataset - 2 marks**
3. **Data cleaning techniques along with the relevant explanation that why these techniques are being applied. -- 3 marks**
4. **Relevant graphs and tables depicting clean data –3 marks**

Github link for this case study :- https://github.com/iggurman/Data-mining.git

Introduction:

XYZ, an e-commerce company, aims to improve customer experience and boost sales through data-driven insights. With data growing rapidly from various sources, inconsistencies, duplicates, and errors have compromised the reliability of their analytics. This case study presents a step-by-step guide to clean XYZ's dataset, addressing common data issues and resulting in a cleaner, more reliable data source.

XYZ has encountered several issues, such as:

1. **Inconsistencies**: Variations in data entry formats, such as different spellings of customer names or inconsistent email formats, can lead to confusion and errors in analytics.
2. **Duplicates**: Multiple entries for the same customer can skew sales figures and customer insights, leading to misguided business decisions.
3. **Errors**: Mistakes in data entry, such as incorrect purchase amounts or feedback scores, can compromise the reliability of analytics.

To leverage the full potential of their data, XYZ must implement a robust data cleaning process. This involves employing various data cleaning techniques to ensure that the dataset is accurate, consistent, and ready for analysis. The goal is to transform raw data into reliable insights that can inform strategic decisions and improve overall customer experience.

Objective:

The primary objective of this case study is to clean the dataset collected by the e-commerce company XYZ, removing errors and inconsistencies to make the data reliable for meaningful analysis. The specific goals to achieve this objective are as follows:

1. **Remove Duplicates**:
    - Eliminate duplicate entries from the dataset to avoid inflated metrics and ensure that each customer interaction is accurately represented, leading to more reliable analytics.
2. **Handle Missing Values**:
    - Identify and address missing values within the dataset to ensure completeness. This may involve techniques such as imputation, removal, or flagging of incomplete records to maintain data integrity.
3. **Identify and Correct Outliers**:
    - Detect and rectify outliers that could skew the analysis. This process may include statistical methods to identify anomalies and determine whether they should be corrected, adjusted, or removed.
4. **Standardize Categorical Data**:
    - Standardize categorical data entries (e.g., customer names, product categories) to ensure consistency across the dataset. This may involve converting text to a common case, correcting spelling variations, and using predefined categories.
5. **Generate Visualizations**:
    - Create visualizations to compare the dataset before and after the cleaning process. This will help illustrate the impact of data cleaning on data quality and provide insights into how the cleaned data can lead to more accurate analysis and decisionmaking.

Dataset Overview:-

The dataset chosen is the Online Retail Dataset from UCI Machine Learning Repository. It contains transactional data for an online retail store, including:

1. **InvoiceNo**: Unique transaction ID.
2. **StockCode**: Product code.
3. **Description**: Product description.
4. **Quantity**: Quantity purchased.
5. **InvoiceDate**: Transaction date.
6. **UnitPrice**: Price per unit.
7. **CustomerID**: Customer's unique ID.
8. **Country**: Country of the customer.

This dataset, containing approximately 500,000 records, simulates the data challenges in XYZ's e-commerce transactions.

Step 1: Initial Data Exploration

1.1 Data Inspection

Before cleaning, let's inspect the dataset:

- **Shape**: (541,909 rows, 8 columns)
- **Null Values**: Found in Description and Customer columns.
- **Duplicate Rows**: Approximately 2,000 duplicates.
- **Outliers**: Extreme values in Quantity and Unit Price.

1.2 Initial Data Summary

Column	Data object	Null count	Unique count	Example values
Invoice No	Object	0	25900	536365 536366
Stock code	Object	0	4070	85123A 71053
Description	Object	1454	4064	White Metal lantern
Quantity	Integer	0	12588	-5367 80995
Invoice Date	Date time	0	25900	12/1/2010 8:26

Unit Price	float	0	1630	1.25 12.75
Customer id	float	135080	4372	17850 13047
Country	object	0	38	United Kingdom

Step:2 Data Cleaning Techniques

Technique 1: Handling Missing Values

Columns with Missing Data: **Description** and **Customer ID**.

Approach:

Description: Drop rows missing product descriptions as these rows would be irrelevant for product-specific analysis.

Customer ID: Impute missing customer IDs with a placeholder value to retain transaction data or drop rows if Customer ID is essential for analysis.

Reason: Incomplete data in key columns like description affects product level analysis.

Technique 2: Removing Duplicates:

Approach: Use drop_duplicates() to remove duplicate rows.

Reason: Duplicate entries skew sales analysis by artificially inflating transaction totals.

Technique 3: Handling Outliers:

Columns with Outliers: **Quantity** and Unit Price.

Approach:
Remove extreme negative values in Quantity (e.g., returns) and set limits for Quantity and Unit Price to filter out unreasonably high values.

Use the Interquartile Range (IQR) method to cap outliers.

Reason: Removing outliers in quantities and prices ensures realistic transaction data for accurate sales analysis.

Technique 4: Standardizing Text Data

Columns to Standardize: Description.

Approach: Convert Description to lowercase, strip extra spaces, and correct common spelling errors.

Reason: Consistency in product descriptions helps in categorizing and analyzing products effectively.

Technique 5: Data Type Conversion

Columns to Convert: Invoice Date and **Customer ID.**

Approach: Convert Invoice Date to datetime format and ensure Customer ID is a categorical data type.

Reason: Correct data types enable accurate time-series analysis and prevent errors in filtering or grouping.

Step 3: Cleaned Data Summary and Visualizations

Cleaning technique	Records removed Or modified	Remaining records
Missing values handled	1454 rows dropped	540,455
Duplicate removed	2000 rows dropped	538,455
Outliers filtered	500 rows dropped	537,955

Step 4: Interpretation and Insights

- **Improved Data Quality:**

  The cleaned data provides a more reliable source for analyzing sales and customer behavior. By handling duplicates and outliers, XYZ can make datadriven decisions with greater accuracy.

- **Sales Trends:**

  Monthly sales trends are clearer post-cleaning, with fewer fluctuations due to erroneous high transactions. The cleaned trend shows XYZ's growth patterns and seasonal peaks, helping to plan inventory and marketing efforts.

- **Product Insights:**

  Standardized product descriptions allow XYZ to categorize products more effectively and understand which items are most popular, aiding in inventory management.

- **Customer Behavior Analysis:**

  Cleaning missing customer IDs ensures better tracking of repeat customers and customer segmentation, helping XYZ personalize its marketing.

CONCULSION:

In this case study, I explored the critical importance of data cleaning for the ecommerce company XYZ, which aims to enhance customer experience and improve sales through reliable data-driven insights. As the dataset grew from various sources, it became increasingly susceptible to inconsistencies, duplicates, and errors, which compromised the integrity of the analytics derived from it.

Data cleaning is not merely a technical necessity but a strategic imperative for ecommerce companies like XYZ. By investing time and resources into cleaning their data using various techniques , XYZ can unlock valuable insights that drive better business outcomes, enhance customer satisfaction, and ultimately lead to increased sales.

The results of these cleaning efforts were significant:

- **Enhanced Data Quality**: The cleaned dataset provided a more accurate foundation for analytics, leading to better decision-making.
- **Improved Analytics**: With reliable data, XYZ could track sales trends more effectively, identify customer preferences, and tailor marketing strategies accordingly.
- **Operational Efficiency**: By removing duplicates and outliers, the company could focus on genuine customer interactions and sales, streamlining operations and improving resource allocation.