Identifying Inconsistencies and Cleaning Dataset

by

Juliet Efemena Micheal

Learner ID 152151

Module 2 Assignment

BAN640: Data Modeling and Mining

October 11, 2024.

Part 1: Identify All Inconsistencies in the Dataset

On the process of cleaning the data, all the inconsistencies in the dataset are highlighted below:

1. Missing or Empty Values:

These are data fields that is not available or missing randomly in the dataset:

In the data field custName: Review shows the missing values in the customer's name.

Age: There are 7 missing values at random in the age column field

As well as DatePurchased: Missing 7 records.

Price: Four products have no price information on them.

RatingOfProduct: Analysis shows missing ratings of 7 entries.

And AdvertisingAgency: Has only one missing entry.

2. Outliers

To check whether any values in either of Age or Price that appear not to be normally low or high that looks not to be real.

From my observation, there are a few outliers 99.9 which is in the Price, while another one shows 12.99. Of which the insight of my analysis will be affected.

3. Inconsistent Date Formats

See if there are any inconsistencies in the date format of fields? E.g., the DatePurchased field is in a format that is "12/01/2023" versus "January 12, 2023"?

4. Typographical Errors

Check out the fields like custName, Product or AdvertisingAgency typist's errors or naming convention inconsistencies like "CocaCola" and "Coca Cola".

5. Inconsistent Data Types

Fields such as Age or Price should be numeric. Though in situations where they may contain text, symbols, or may be categorical.

Though the Price column should be numeric data, so we can check

for any doable formatting issues e.g., unexpected symbols or text.

Age is in float format, while this should be an integer, as age is usually a whole

number.

6. Potential Invalid Values:

The RatingOfProduct column might have values out of the range of what normally could be rated-e.g., 1 to 5. It' is worth checking what range of values are here.

7. Outdated or Incorrect Data

To check if the DatePurchased is valid or whether the Price is set to irrational values Showing a negative price. Though, I feel some of these prices are not correct because, looking back to 2013, there were one pair of Shoes that go for 79.99, and yet we have Shoes today in the year 2022 that goes for the same price, and I feel those prices should have skyrocketed.

Missing values for the custName, Age, and DatePurchased key fields can make the customers' data to be incomplete, and then affecting the overall integrity of the whole data set.

8. Redundant Entries:

Let it not be directly obvious but very necessary to verify for duplicate custID or multiples of a single customer record or information and even their presence would compromise the uniqueness of their records.

Part 2: Five Techniques/Methods to Solve Inconsistencies

Here are five techniques to address the issues identified and ensure high-quality, consistent data:

1. Handling Missing Data

❖ Imputation: For missing numeric values like Age or RatingOfProduct, imputation can be useful. For instance, you can fill missing Age values with the mean or median age of the available records.

❖ For categorical columns like custName, missing entries could either be filled with a placeholder (like "Unknown") or inferred based on other related fields, depending on the context.

❖ Drop rows or columns: If the number of missing entries is too high for certain rows or columns, consider dropping them if they do not significantly affect the analysis.

2. Ensuring Data Type Consistency

❖ Types of data cleaning: That date field Age is converted from float to int so that we can avoid any issues with floating-point numbers. The data could also be scrubbed by checking the Price field for any non-numeric characters or text (e.g., symbols) and clean those entries that if found.

❖ Implement consistency within the dataset using type casting-for instance, maintain the Price column as a float to two decimal places.

3. Data Validation and Cleaning

❖ Range Validating: The RatingOfProduct column should contain valid entries, e.g., between 1 and 5. Any outliers or invalid values should either be corrected or flagged so that it would be reviewed later.

❖ Regex and validation rules: In order to implement regular expressions or rules for fields like custName so that we would ensure valid names are captured.

4. Handling Duplicates

- Duplicate Customer IDs-custID: Identify these, and when found,

aggregate the duplicate records or keep only the newest if other fields, for instance, are DatePurchased.

- Duplicate rows: Whether there are rows that seem to be identical showing the exact same details, then such rows would be deleted since they would distort any evaluation.

5. Improve Data Collection Process

- **Mandatory fields:** At the point of collecting data, ensure all key fields like custName, Age, Product, and DatePurchased to be filled in. This would help in reducing the records that are not complete.

- **Input point field validation:** Implement validation of data while adding the information with restrictions on proper formatting of the fields, e.g., Price should be numeric or DatePurchased should be in a valid date format. Predefined selection options for specific fields, such as Advertising Agency, may reduce errors in entry, allowing users to select from a preselected list rather than keying free text that may be inconsistent or Misspelled.

Part 3: Clean the Data and Save it in Excel Format

This part requires confirm data cleaning that evolve a step-by-step guide to clean the dataset:

1. Import the Dataset:

Seeing that the dataset given in the case is in text format, it is important to ensure it is in the correct format. The data will then be imported into a spreadsheet - (Excel, Google Sheets) but this assignment was addressed using Python with the help of Pandas (pd.read_csv('file.txt')).

2. Handle Missing Data:

Approach: The fact that I used Python, Excel can also be used in finding missing value. Then, I used the VS code system with the aid of function. fillna() to fill missing values. Where I decided if to fill or drop rows with that same missing data.

3. Normalize Erroneous Formats:

Approach: The date columns were made in consistent date format;

Text to Columns can be used in Excel. Numeric columns for Age, and Price are cleaned to be string-free.

4. Remove Anomalies:

Approach: The ages that is showing outside the plausible limits are sorted using Excel formula=IF(AND(Age>18, Age<100), Age, "Outlier"). IQR or functions like .quantile() can be used in Python.

5. Replace Typographical Errors:

Standardization: Standardize both the names and product descriptions by using Excel's Find and Replace Python's fuzzy matching algorithms through the fuzzywuzzy library.

6. Remove Duplicates Data:

Strategy: In Excel, apply the feature Remove Duplicates; in Python, .drop_duplicates() removes repeated records based on custID or Product.

7. Save Cleaned Data

Save cleaned dataset into Excel format: e.g., Cleaned_Data.xlsx.

Part 4: Challenges and Opportunities

The five challenges and five opportunities a company could face when using this dataset are as follows:

Five Challenges:

The five potential challenges are as follows:

1. Data Quality Issues:

The challenges of lower data quality in the company with incomplete, inconsistent, or even outdated information can also result in incorrect reporting. E.g., if the sales data was captured wrongly or incorrectly, then the tendency to predict analysis will be difficult.

2. Perplexity in Data Clearing:

This is sometimes extremely hard and time wasting to manually clear large datasets because it is open to human mistake or error.

3. Data Privacy Concerns:

The challenge is to ensure data security and protection while maintaining data confidentiality and integrity where customer data such as custID and buy history must comply with privacy laws e.g., GDPR.

4. Handling Outliers:

Pointing out the consistent outliers versus data errors can be difficult to locate, leading to possibles skewed the report.

5. Inadequate tools:

Data mining and analytics requires resources and tools in any company. Then if any companies lack the proper data analytics tools that would be required for clearing of the data efficiently, the quality of dataset will be affected.

Five Opportunities:

Here are the five opportunities a company will experience while using the dataset**:**

1. Enhanced Decision-Making:

Clearing data gives a helpful insight into customer purchasing patterns that will help target audiences in making decisions.

2. One-on-One Marketing:

This helps in creating more personal, relevant, and engaging interaction. e.g., one- on-one interaction which allows the firm to develop targeted marketing campaigns and customer loyalty.

3. Enhance Customer Satisfaction:

Knowing the product ratings and buying history can help in promoting the products, customer services, and communication.

4. Efficient Resource Allocation:

Efficient Resource Allocation is one of the great opportunities from the dataset that reveals whether there is a valuable insight into the advertising agencies for more effective business that would help notify the marketing team about where exactly to channel the resources for more outgrowth.

5. Improve Customer Retention:

Evaluating buying patterns and trends, making companies identify or forecast customers that would have a tendency to stir up and take proactive steps in retaining them. However, proper customer insight would help improve better customer retention strategies.

References

1. Kowieski, Jon. “What Is Data Cleaning? Examples and How to Clean Your Data.” *ThoughtSpot*, 2 Dec. 2022, www.thoughtspot.com/data-trends/data-science/what-is-data-cleaning-and-how-to-keep-your-data-clean-in-7-steps.

2. van Eck, M.L., Lu, X., Leemans, S.J.J., & van der Aalst, W.M.P. (2015). Interactive data cleaning for process mining: A case study of an outpatient clinic's appointment system. Springer International Publishing. https://doi.org/10.1007/978-3-319-19069-3_19

3. Si, S., Xiong, W., & Che, X. (2023). Data quality analysis and improvement: A case study of a bus transportation system. Applied Sciences, 13(19), 11020. https://doi.org/10.3390/app131911020

4. Moore, D.S., & McCabe, G.P. (2023). Use data mining cleansing to prepare data for strategic decisions. IntechOpen. https://doi.org/10.5772/intechopen.96231