Report on Data Cleaning and Processing Task

Documenting the Example

Mahir Hashimov
12695120
Unit 4
Deciphering Big Data
University of Essex Online
21 June 2024

Contents

Data Cleaning and Transformation

Introduction

Data cleaning is a critical step in the data analysis process, ensuring that datasets are accurate, consistent, and ready for analysis (Gudivada et al. 2017 & Fan et al. 2021). This report details the process of cleaning and preprocessing a dataset obtained from UNICEF survey data, which involved handling mixed data types, renaming columns, identifying and removing duplicates, managing missing values, and saving the cleaned data to a SQLite database. The task utilized Python and its libraries pandas and sqlite3.

Objective

The primary objective of this task was to clean and preprocess the raw data contained in mn.csv using human-readable headers from mn_headers.csv. Specific goals included:

1. Importing and inspecting the raw data.
2. Renaming columns based on a mapping dictionary.
3. Identifying and removing duplicate rows.
4. Handling missing data.
5. Saving the cleaned dataset to a SQLite database

Methodology

The task was approached systematically, with a focus on modularity and reusability of the code. The key steps and the corresponding code functions are detailed below:

1. Loading Data

The raw data and headers were loaded into pandas DataFrames. This step involved reading the CSV files and ensuring that the data was correctly imported for further processing.

2. Creating a Headers Dictionary

A dictionary mapping the acronyms to their human-readable labels was created using pandas. This dictionary facilitated the renaming of columns in the main dataset, enhancing readability and interpretability.

3. Renaming Columns

The columns in the dataset were renamed using the headers dictionary. This step was essential for transforming cryptic headers into meaningful names, aiding subsequent data analysis.

4. Removing Duplicate Rows

Duplicate rows were identified and removed using pandas' drop_duplicates method. This ensured the integrity of the dataset by eliminating redundant information.

5. Checking for Missing Values

The dataset was checked for missing values, and their counts were obtained. This step was crucial for identifying columns that required further cleaning or imputation.

6. Saving the Cleaned Data

The cleaned dataset was saved to a SQLite database using the sqlite3 module. This involved creating a table structure suitable for storing the cleaned data and inserting the data into the table.

Challenges and Solutions

Several challenges were encountered during the data cleaning process:

1. Mixed Data Types:

The raw data contained columns with mixed data types, causing warnings during loading. This issue was managed by using the low_memory=False parameter in pandas, allowing for efficient data loading without type inference problems.

2. Duplicate Column Names:

Some columns had duplicate names after renaming. This was resolved by ensuring that the headers dictionary correctly mapped each acronym to a unique, human-readable label.

3. Handling Missing Values:

The dataset contained missing values across several columns. While this task focused on identifying these missing values, further steps could involve imputation or deletion based on the analysis requirements.

4. Data Integrity:

Ensuring the accuracy and consistency of the dataset was paramount. The process involved verifying that the renaming and duplicate removal steps did not alter the data structure or introduce errors.

5. Database Integrity:

Initially, the dataset library was used for database interactions. However, due to installation issues, the sqlite3 module was employed. This required modifying the script to ensure compatibility with SQLite, involving creating table structures and inserting data manually.

Critical Analysis and Technical Considerations

The data cleaning process necessitates a careful balance between automation and manual inspection (Hosseinzadeh et al., 2021). Automated scripts enhance efficiency and repeatability, but manual checks are essential to ensure the process's correctness and completeness (McKinney, 2022).

From a technical perspective, the use of pandas for data manipulation proved effective, given its robust handling of large datasets and versatile data manipulation capabilities. The switch to sqlite3 for database interactions highlighted the importance of flexibility in tool selection, ensuring that the process could be completed despite environmental constraints.

The modular approach adopted in the script enhanced its clarity and reusability. Each function was designed to handle a specific aspect of the cleaning process, making the script easy to understand, debug, and extend.

Conclusion

The data cleaning and preprocessing task was successfully completed, resulting in a cleaned dataset ready for further analysis. The systematic approach ensured that the data was accurately processed, with challenges effectively managed through appropriate technical solutions. This report underscores the importance of meticulous data cleaning practices in the broader context of data analysis, laying a strong foundation for reliable and insightful results.

The script developed can serve as a template for similar tasks in the future, demonstrating the effectiveness of combining automated data manipulation with critical human oversight. This approach ensures that datasets are not only clean and consistent but also well-documented and easily understandable, facilitating further analysis and decision-making.The script developed can serve as a template for similar tasks in the future, demonstrating the effectiveness of combining automated data manipulation with critical human oversight.

References

Gudivada, V.N., Apon, A. and Ding, J., 2017. Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. International Journal on Advances in Software, 10(1-2). Available at: http://www.iariajournals.org/software/ [Accessed 19 June 2024].

Fan, C., Chen, M., Wang, X., Wang, J. and Huang, B., 2021. A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. Frontiers in Energy Research, [online] 9. Available at: https://doi.org/10.3389/fenrg.2021.652801 [Accessed 19 June 2024].

Hosseinzadeh, M., Azhir, E., Ahmed, O.H., Ghafour, M.Y., Ahmed, S.H., Rahmani, A.M. and Vo, B., 2021. Data cleansing mechanisms and approaches for big data analytics: a systematic study. Journal of Ambient Intelligence and Humanized Computing, 14, pp.99-111. Available at: https://doi.org/10.1007/s12652-021-03590-2 [Accessed 20 June 2024].

McKinney, W., 2022. Python for Data Analysis. O'Reilly Media, Inc.