



Web Application

For Automated Data Cleaning






Web applications for automated data cleaning are becoming increasingly popular as organizations strive to make sense of the vast amounts of data they collect. These tools are designed to help users quickly and easily clean, organize, and prepare their data for analysis. In this report, we will explore the features and capabilities of some of the most popular web applications for automated data cleaning, and discuss their potential benefits and drawbacks. We will also examine the ways in which these tools can help organizations improve their data quality, increase efficiency, and make better use of their data. Overall, the goal of this report is to provide readers with a comprehensive understanding of the web-based tools available for automated data cleaning, and how they can be used to support data-driven decision making.

Project Summary

This Python-based web application performs Exploratory Data Analysis (EDA) and is a tool that allows users to easily explore and analyze large datasets through a web-based interface. This application typically utilizes the Python programming language and various data analysis libraries, such as Pandas, to provide users with a range of visualization and statistical tools for understanding their data. Some of the key features of these applications include the ability to import data from various sources, perform basic data cleaning and preprocessing tasks, and generate various types of visualizations (e.g., histograms, scatter plots, box plots) to help users identify patterns and trends in their data. Additionally, these applications may also provide users with the ability to perform statistical tests, calculate summary statistics, and interact with the data in real-time. Overall, a Python-based web application for EDA is a powerful tool that can help users quickly and easily gain insights from their data and make data-driven decisions.

Project Introduction

A Python-based automated data cleaning program can achieve several key objectives, including:

-  Removing or correcting errors and inconsistencies in the data: This can include removing duplicate records, correcting inaccuracies, and standardizing formatting.
-  Identifying and handling missing or null values: This can include filling in missing values with appropriate defaults or removing records with too many missing values.
-  Handling outliers and anomalies in the data: This can include identifying and removing extreme values that may be errors or removing records that do not conform to expected patterns.
-  Data Transformation: This can include converting data into the desired format or units, and join or aggregate the data from multiple sources.
-  Data Visualization: This can include creating visualizations of the cleaned data to help identify patterns and trends that may not be immediately apparent in the raw data.

Automating the data cleaning process: This can include creating scripts or programs that can be run periodically to keep the data clean and up-to-date.

By achieving these objectives, a Python-based automated data cleaning program can help improve the quality and reliability of the data used in a wide range of applications, including machine learning, business intelligence, and data analysis.

"Data cleaning is like house cleaning: everyone knows it needs to be done, but no one wants to do it." - Robert Grossman

Technology and Literature Review

A technology and literature review for a Streamlit application that cleans data would involve researching and evaluating existing technologies and literature related to data cleaning and Streamlit.

One technology that could be used in a Streamlit data cleaning application is the Python pandas library. Pandas provides a wide range of data cleaning and manipulation functions, such as removing duplicates, handling missing values, and standardizing data formats. It also integrates well with other popular data analysis libraries, such as NumPy and scikit-learn.

Another technology that could be used is the Dedupe library, which is a python library for de-duplicating and merging datasets. It uses machine learning to perform record linkage and can be integrated with pandas dataframe.

Streamlit, on the other hand, is a Python library for building interactive web-based data visualization and machine learning applications. It allows for easy creation of user interfaces and widgets, making it well-suited for data cleaning applications where users need to interact with the data and make decisions about how to clean it.

In literature review, one can find several research papers and articles that discuss the importance of data cleaning, the challenges of data cleaning, and the various techniques and methods used for data cleaning. Additionally, there are also research papers that discuss the use of machine learning and natural language processing techniques for data cleaning.

In conclusion, by combining the power of Pandas and Streamlit, one can create a powerful and user-friendly data cleaning application that allows users to easily identify and correct errors and inconsistencies in their data, and automate the cleaning process.

Survey using online data

For a data cleaning project, an online survey could be used to gather information on the quality and accuracy of the data that needs to be cleaned. Here are a few examples of how an online survey could be used in a data cleaning project:

"Online surveys are revolutionizing the way we gather data, offering a cost-effective, fast, and convenient way to reach a large and diverse sample of participants." - Anonymous

Data Quality Assessment: Creating an online survey to gather feedback from stakeholders or data users on the quality and accuracy of the data that needs to be cleaned. This survey could include questions on data completeness, accuracy, consistency, and usability.

Data Validation: Creating an online survey to gather validation information from external sources to validate the data that needs to be cleaned. This survey could include questions that are specific to the data, such as asking for specific values or ranges of values to confirm the accuracy of the data.

Error Identification: Creating an online survey to gather information on the types and frequencies of errors present in the data. This survey could include questions on the type of errors, such as missing values, incorrect values, or duplicate records.

Data Annotation: Creating an online survey to gather information that can be used to annotate the data. This survey could include questions on the meaning, context or the format of the data.

Human in the loop: Creating an online survey to gather feedback on the results of the data cleaning process. This survey could include questions on the effectiveness of the cleaning process, the completeness of the cleaned data, and the usefulness of the cleaned data for specific tasks.

It's important to keep in mind that when conducting a survey using online data for this project the members involved were mostly a party in this project.

Project development approach

The project development approach for a Streamlit application that cleans data should involve the following steps:

- ✚ Define the problem and objectives: Clearly define the problem that the data cleaning application is trying to solve and the specific objectives that the application needs to achieve. This will help guide the design and development of the application.
- ✚ Data Exploration: Perform an initial exploration of the data to understand its structure, quality, and characteristics. This will help identify any potential issues that need to be addressed in the cleaning process.
- ✚ Data Cleaning: Develop a plan for the data cleaning process, including the specific cleaning tasks that need to be performed, such as removing duplicates, handling missing values, and standardizing data formats.
- ✚ Build the User Interface: Build the user interface using Streamlit, creating widgets and interactive elements that allow users to interact with the data and make decisions about how to clean it.
- ✚ Automation: Automate the data cleaning process by creating scripts or programs that can be run periodically to keep the data clean and up-to-date.
- ✚ Test and Validate: Test the application and validate the cleaning process using a sample of the data. This will help identify any issues or bugs that need to be addressed and ensure that the application is meeting its objectives.
- ✚ Deployment: Deploy the application on a suitable platform, such as a web server or cloud service, to make it accessible to users.
- ✚ Monitor and improve: Monitor the usage of the application and gather feedback from users, this will help to improve the application and fix any issues that arise.

By following this development approach, a Streamlit data cleaning application can be developed that is easy to use, efficient, and effective at cleaning data.

Agile framework

Building an Agile framework for data cleaning can involve the following steps:

- ✚ Define the scope: Clearly define the scope of the data cleaning project, including the specific data sets that need to be cleaned, the cleaning tasks that need to be performed, and the objectives of the project.
- ✚ Form a cross-functional team: Assemble a cross-functional team with members from different departments such as data analysts, developers, and quality assurance, to work together to achieve the project objectives.
- ✚ Prioritize the backlog: Prioritize the backlog of data cleaning tasks based on their importance, dependencies, and risks.

- 📌 Plan sprints: Plan sprints, which are short cycles of development, typically lasting 2 weeks, during which a specific set of tasks will be completed.
- 📌 Hold daily stand-up meetings: Hold daily stand-up meetings to discuss progress, identify and resolve any issues, and plan for the next day's tasks.
- 📌 Implement Continuous Integration and Continuous Deployment: Implement continuous integration and continuous deployment (CI/CD) practices to automate the data cleaning process, allowing for faster and more efficient data cleaning.
- 📌 Conduct regular reviews and retrospectives: Conduct regular reviews and retrospectives to evaluate the progress of the project, gather feedback from team members, and identify areas for improvement.
- 📌 Continuously monitor and improve: Continuously monitor and improve the data cleaning process by implementing new techniques, tools, and best practices as they become available.

By following these steps, an Agile framework for data cleaning can be developed that allows for flexible and efficient data cleaning, as well as a continuous improvement of the process. Agile development process is known for its flexibility, collaboration and rapid delivery, this is why it is a great fit for data cleaning projects.

Work breakdown structure

Following an Agile Development approach, we were able to break the entire project into “three” separate parts, namely **Web Application Development (UI Development)**, **Automated Data Cleaning Process** and **Dynamic EDA Development**

We followed the following steps and developed the parts in synergy

- 📌 Project Initiation: Define the scope of the project, assemble the team, address any doubts, and establish project goals and objectives.
- 📌 Data Exploration: Perform an initial exploration of the data to understand its structure, quality, and characteristics.
- 📌 Data Cleaning: Develop a plan for the data cleaning process, including the specific cleaning tasks that need to be performed, such as removing duplicates, handling missing values, and standardizing data formats.
- 📌 User Interface Design: Design the user interface using Streamlit, creating widgets and interactive elements that allow users to interact with the data and make decisions about how to clean it.
- 📌 Automation: Automate the data cleaning process by creating scripts or programs that can be run periodically to keep the data clean and up-to-date.
- 📌 Testing and Validation: Test the application and validate the cleaning process using a sample of the data.
- 📌 Deployment: Deploy the application on a suitable platform, such as a web server or cloud service, to make it accessible to users.

- 📌 Monitoring and maintenance: Monitor the usage of the application and gather feedback from users, this will help to improve the application and fix any issues that arise.

Risk Management

We observed the following steps to ensure the Safe and Sound application of our web application and successive user

- 📌 Test the program: Test the program thoroughly with a variety of test data sets, including edge cases and scenarios that may cause errors. This will help identify and correct any bugs or issues that may cause the program to crash.
- 📌 Use exception handling: Use exception handling to catch and handle any errors that may occur during the execution of the program. This will help prevent the program from crashing and allow it to continue running.
- 📌 Use version control: Use version control, such as Git, to track and manage changes to the code, allowing you to easily revert to a previous version of the code if the program crashes or an error occurs.
- 📌 Monitor the performance: Monitor the performance of the program to detect any potential issues or errors that may cause it to crash. This can be done using performance monitoring tools, such as Python profilers, that can identify bottlenecks and potential issues in the code.
- 📌 Regular updates: Keep the program updated with the latest libraries and modules, as well as the updates for the operating system and the dependencies.
- 📌 Backup the data: Backup the data before running the program, this will help in case of data loss or corruption.
- 📌 Validate input data: Validate the input data before running the cleaning process, to ensure that it meets the required format and data types. This can help prevent errors that may cause the program to crash.
- 📌 Use robust libraries: Use robust and well-established libraries for data cleaning, such as pandas and numpy, which have been extensively tested and have a large user base, this can help prevent issues that may cause the program to crash.

Training module

The Web Application that we have developed is Highly Automated and tailored with the principals of Bespoke Tailoring in mind to address a particular type of data set pertaining to a certain domain.

However, Literature could be forwarded or a live demonstration to highlight the scope of our application.

User module

A user module for a data cleaning application could include a variety of features and functionality that allow users to interact with the application and clean their data. Here are examples of what our user module for a data cleaning application offers include:

- ✚ Data upload: A feature that allows users to upload their data to the application for cleaning. This could include support for various file types, such as CSV, Excel, and JSON.
- ✚ Data visualization: A feature that allows users to view their data in various formats, such as tables, charts, and graphs, to help them better understand the data and identify any issues that need to be addressed.
- ✚ Data cleaning options: A feature that allows users to select from a variety of data cleaning options, such as removing duplicates, handling missing values, and standardizing data formats.
- ✚ Data validation: A feature that allows users to validate the cleaned data, to ensure that it meets the required standards and that all the necessary cleaning steps have been performed.
- ✚ Data export: A feature that allows users to export their cleaned data in various file formats, such as CSV, Excel, and JSON.
- ✚ User settings: A feature that allows users to customize their settings and preferences, such as setting default cleaning options and selecting the preferred file format for data export.
- ✚ User feedback: A feature that allows users to give feedback and suggest improvements to the application.
- ✚ Help and support: A feature that provides users with access to help and support resources, such as documentation, tutorials, and FAQs.
- ✚ Reporting: A feature that allows users to generate reports, such as statistics about the data, the cleaning process, and the performance of the application.
- ✚ A user module is a crucial part of the data cleaning application, as it is the interface that users interact with, it should be designed in a user-friendly way, easy to navigate and understand.

Admin module

An admin module is a set of features and functionality within a data cleaning application that is intended for use by an administrator or a group of users with administrative privileges. The main purpose of an admin module is to allow these users to manage and maintain the application, as well as to perform tasks that are not typically available to regular users. Here are a few examples of what an admin module for our data cleaning application could include:

- ✚ User management: A feature that allows administrators to manage and maintain user accounts, such as creating new accounts, editing existing accounts, and disabling or deleting accounts.

- ✚ Data management: A feature that allows administrators to manage and maintain the data stored in the application, such as importing, exporting, and backing up data.
- ✚ Application settings: A feature that allows administrators to configure and manage the settings of the application, such as setting up data validation rules, setting up data cleaning options, and setting up data export options.
- ✚ Performance monitoring: A feature that allows administrators to monitor the performance of the application, such as tracking the number of users, the amount of data processed, and the resource usage.
- ✚ Logging and auditing: A feature that allows administrators to view logs and audit trails of the application, such as user activity logs and system logs.
- ✚ Maintenance: A feature that allows administrators to perform maintenance tasks on the application, such as updating software, backing up data and monitoring the system.
- ✚ Security: A feature that allows administrators to manage security-related tasks, such as setting up user roles and permissions, and monitoring the system for security breaches.
- ✚ Feedback and support: A feature that allows administrators to manage feedback and support requests from users, such as tracking and responding to support tickets.

Testing

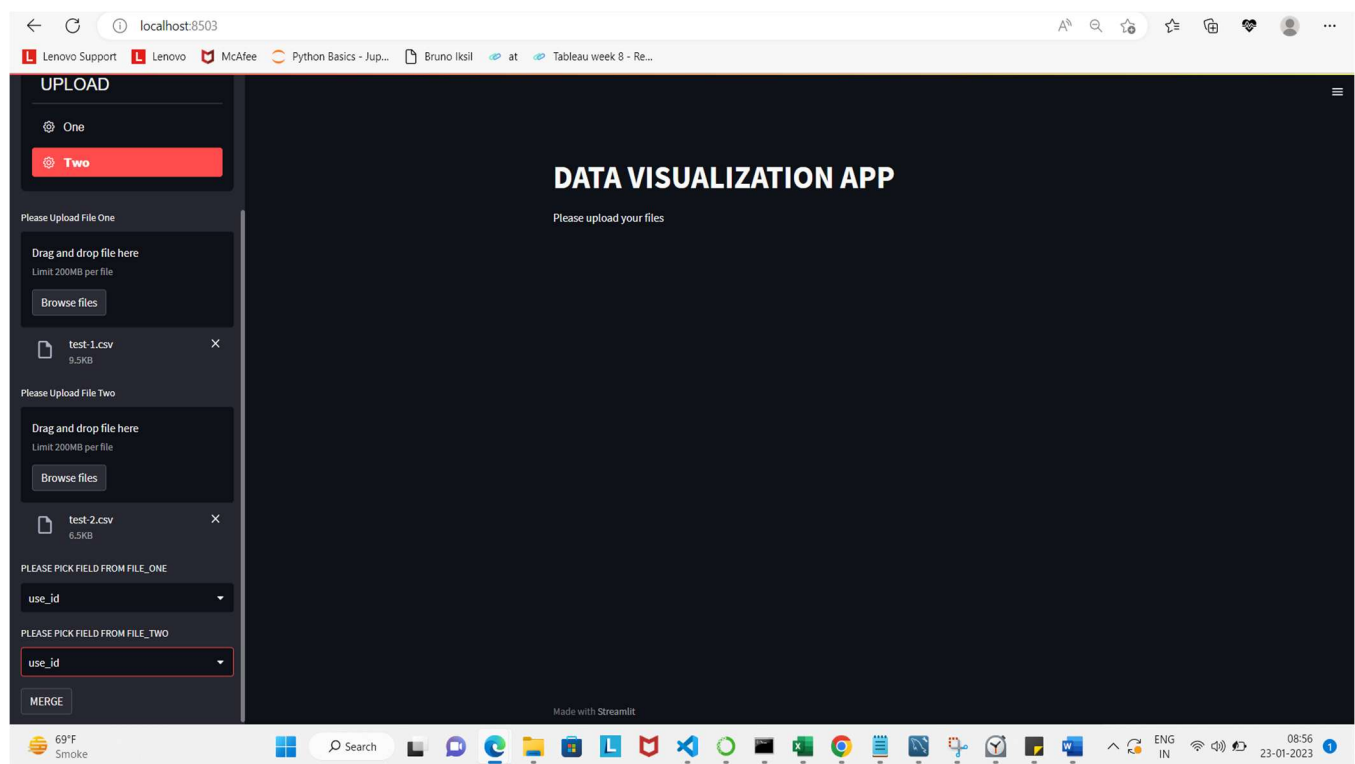
Testing is an important step in the development of a data cleaning application, as it helps to ensure that the application is functioning correctly and meets the requirements of the users. Here are a few examples of the types of testing that could be performed for a data cleaning application:

- ✚ Functional testing: Testing that verifies that the application functions as intended and meets the requirements of the users. This could include testing the data upload, cleaning, validation and export features of the application.
- ✚ Performance testing: Testing that verifies that the application can handle the expected load and usage patterns, such as the amount of data, number of users and the complexity of the data.
- ✚ Security testing: Testing that verifies that the application is secure and that the data is protected from unauthorized access. This could include testing for vulnerabilities such as SQL injection, cross-site scripting, and data breaches.
- ✚ User acceptance testing: Testing that verifies that the application meets the needs of the users and is easy to use. This testing is done by actual users of the application, who can provide feedback and suggest improvements.
- ✚ Regression testing: Testing that verifies that the application continues to function correctly after changes or updates have been made. This testing is done after any modification to the application to ensure that the changes don't have negative impact on the existing functionalities.
- ✚ Integration testing: Testing that verifies that the application integrates correctly with other systems or applications, such as databases, external APIs, and other software.
- ✚ Automated testing: Automating the testing process, this can help to increase the efficiency, reduce the time of testing and lower the costs of the testing process.

System Design

A user interface is like a joke. If you have to explain it, it's not that good". — Martin Leblanc

While designing the user interface for this application this is what we had in mind, hoping that we have done a decent enough job to let our User Interface speak for itself.



Graphs and visuals

We have relied on Pandas Profiling is a Python library that generates a comprehensive report on the characteristics of a data set. It provides a quick and easy way to create a summary of a large data set, which can be useful for data exploration and understanding the structure of a data set. Here are a few advantages of using Pandas Profiling:

- ✚ Saves Time: Generates a report with a single line of code, it saves a lot of time as compared to manual exploration of the data.
- ✚ Comprehensive: Generates a comprehensive report that includes information about the data types, missing values, unique values, and distribution of variables.
- ✚ Visualizations: Includes a variety of visualizations, such as histograms, bar charts, and scatter plots, that can help to quickly identify patterns and trends in the data.
- ✚ Memory-efficient: It is memory-efficient and can handle large data sets, it can handle large data sets without consuming too much memory, making it suitable for large datasets.
- ✚ Customizable: It is customizable, it allows to customize the output and include or exclude certain features.

NOTE: We also have provision to provide interactive graphs

Conclusion

A company (or an individual for that matter) can observe several major benefits if they opt for automated data cleaning:

- ✚ Increased efficiency: Automated data cleaning can greatly increase the efficiency of the data cleaning process, as it can quickly and accurately clean large data sets, reducing the time and resources required for manual cleaning.
- ✚ Improved data quality: Automated data cleaning can help to improve the quality of the data by removing errors and inconsistencies, and by standardizing the data format. This can help to ensure that the data is accurate and reliable, which is important for making informed business decisions.
- ✚ Reduced costs: Automated data cleaning can help to reduce the costs associated with data cleaning, as it can significantly reduce the time and resources required for manual cleaning.
- ✚ Better decision making: Automated data cleaning can help to improve the quality of the data, allowing the company to make more informed business decisions.
- ✚ Scalability: Automated data cleaning allows a company to scale the data cleaning process as their data grows, this can help to ensure that the data cleaning process can keep up with the company's data needs.
- ✚ Compliance: Automated data cleaning can help a company to comply with regulations related to data protection and security, by ensuring that the data is accurate, complete, and free from errors.
- ✚ Flexibility: Automated data cleaning can be applied to various types of data, from structured data to unstructured data. This flexibility can help a company to improve their data handling across different departments, systems and data types.
- ✚ Overall, automated data cleaning can help a company to improve the efficiency and accuracy of their data cleaning process, while also reducing costs and improving the quality