November 9th 2023

ASSIGNMENT 3 COLLABORATIVE DEVELOPMENT OF DATA EXPLORER WEB APP

94692 - Data Science Practise

Master of Data Science and Innovation
University of Technology of Sydney

Group 16:

Aditi Vyas (24666152)

Siheng Mu (13475823)

Sinh Thanh Nguyen (25099704)

Xinhui Wang (14380428)

GitHub:
https://github.com/sinhthanhngds/dsp_at3_group16.git

1. Executive Summary

1.1 Overview and Objectives

The CSV Explorer Web App project aimed to revolutionise the exploratory data analysis (EDA) landscape by developing a Streamlit-based application tailored for rapid and comprehensive examination of CSV datasets. EDA is a cornerstone in data science, setting the groundwork for all subsequent analysis. This application emerges as an innovative solution, offering a streamlined and intuitive interface that enables data professionals to delve into datasets with unprecedented speed and clarity. Its significance is particularly notable in an era where the volume and complexity of data burgeon incessantly, demanding tools that not only keep pace but also enhance the analytical prowess of users in data-centric sectors.

1.2 Problem Statement and Context

Data scientists, data analysts, and data-driven decision maker often face challenges in quickly understanding and assessing new datasets due to the lack of accessible tools that can provide immediate and in-depth analysis. The problem is compounded when dealing with large datasets that may contain a mix of text, numeric, and datetime information. This project was undertaken in the context of providing a solution that addresses these challenges by offering a single interface for performing EDA across various data types within CSV files.

1.3 Achieved Outcomes and Results

This project successfully led to the creation of a versatile web application that engages users with four unique sections. Each section is designed to handle different types of data found in CSV file. The app gives a detailed look at the data, showing important details such as the number of rows and columns, any repeated data, and how much computer memory the data uses. What makes this app stand out is its ability to not only provide key statistics for numbers and text but also to visualise this information in charts and graphs that make the data's story clear. For date and time data, the app offers advanced analysis and points out any unusual data that might need a closer look. As the

result, it becomes a powerful tool that makes data analysis easier and faster, helping users quickly find important insights and make decisions based on data.

In summary, the CSV Explorer Web App represents a significant advancement in the tools available to data-related specialists for EDA. It was developed in response to the need for a more efficient way to understand complex datasets, by providing a comprehensive, interactive, and user-friendly platform. The application's ability to handle diverse data types and provide immediate visual and statistical feedback ensures that it is not only a time-saver but also a catalyst for more informed data analysis and decision-making.

2. Introduction

For both individuals and organisations, the capacity to efficiently analyse and make data-driven choices has become critical in the age of big data and abundant information. Data analysis is becoming a vital ability for workers in many other fields, not only data scientists and statisticians. We propose "ExploroData," a web application that helps users explore and analyse data in an easy-to-use, interactive manner. It fills a gap left by the complexity of previous tools and the rising demand for data analysis.

2.1 Project Stakeholders

Stakeholders for our “ExploroData” web app can vary depending on the context and purpose of the application. Possible stakeholders will be list below:

- Data Scientists: Before exploring more complex modelling and analytics, data scientists may find that the EDA web app expedites the preliminary phases of data exploration. The app's data processing and visualisation features can be advantageous to them.
- Developers and IT Teams: One of the important stakeholders is the development team, which oversees creating and managing the web application. This comprises database managers, front-end and back-end developers, and IT support staff who guarantee the dependability, security, and functionality of the application.
- Business Owners and Managers: Since the EDA web app may impact operational efficiency, market insights, and strategic choices, business owners and managers may have an interest in it in a corporate setting.
- User Interface (UI) and User Experience (UX) Designers: These experts are responsible for designing an interface that is simple to use and intuitive. To make the app interesting and simple to use, their feedback is essential.

2.2 Project aim

To satisfy stakeholders’ requirements, out web application offers the following key functionalities:

- Data Preview: The application enables users to quickly view the content of uploaded CSV files. The previews help users ensure they have access to the correct data set.
- Data Exploration: Users can navigate and explore contents of the CSV file, including data overview, data structure, for example, number of rows, columns, missing values, data types, memory usage. The application may also calculate and present summary statistics, such as min, max, mean, median, mode, standard deviation, etc, to help users understand data's behaviours.
- Data Visualization: Users can gain an overview of data distribution of each selected column by providing interactive histograms on the application.

3. Web App Presentation

Exploratory Data Analyse (EDA) web app built in this project makes data exploration and analysis easier for those who don't have a lot of experience with data science by combining interactive dashboards, statistical tools, and user-friendly data visualisation. Easy data input, manipulation, and visualisation are all possible for users in a web-based environment, which does not require a complicated program installed or in-depth training.

3.1 Functionalities

The functionalities of our EDA web app are organized into two main categories: dataset description and dataset feature analysis.

- Dataset Description:
    - Data Overview: Users can import their dataset and gain insights into its essential features, including data types, a preview of a sample of data points, and a summary of key dataset characteristics.
- Dataset Feature Analysis:
    - Numeric Data Analysis: This section offers an in-depth analysis of numerical data within the dataset.
    - Text Data Analysis: Users can explore and analyse text-based data.
    - Datetime Data Analysis: This section is dedicated to the analysis of date and time data.

For each of the data analysis sections, users can select a target column for analysis. Based on the chosen column, the web app provides the following insights:

- Feature Insights: An overview of the selected feature.
- Frequency Table: A table displaying the distribution of unique values.
- Frequency Histogram: A graphical representation of the distribution of unique values.

In a world where creativity and decision-making are encouraged by data-driven insights, this EDA web app shows itself as a potent tool suitable for novice and expert data analysts alike. Its user-friendly design and extensive feature set turn the sometimes-difficult process of data analysis into a fun and useful activity. To shed light on how this online application contributes to data analysis, this project report will examine the development, usage, and function of this EDA web app.

3.2 Application setup

Our web application was supported by the following environment and packages with specific versions, make sure they have been installed properly:

- Python: 3.9.13
- Streamlit: 1.13.0
- Altair: 4.2.0
- Pandas: 2.0.3

Application installation would take the following steps:

1. Open Terminal (on MacOS) or Commands Prompt (on Windows).
2. Direct to the directory that the application would in installed by:

cd [installation_directory]

3. Copy, paste and execute the following:

git clone git@github.com:sinhthanhngds/dsp_at3_group16.git

The Terminal / Command Prompt would show whether the installation is successful.

3. Application launch

To launch our application:

1. Open Terminal / Command Prompt.
2. Direct to the directory storing the installation, where contains 'app' folder by:

cd [your_installation_directory]/dsp_at3_group16/app

3. Copy, paste and execute the following command:

streamlit run streamlit_app.py

The application is successfully launched. Recently, users can start exploring the data by uploading the CSV file.

3.4 Potential Users and Use Cases

The potential users of this application could extend to Data Analysts, Data Scientists, Business Professionals, and even small business owners with the following use cases:

- Data Scientists: performing data cleaning, data visualization for further data pre-processing, including data transformation and scaling. They may use the application to easily import and visualize the data, apply filters, and generate insights.
    - Benefits: Streamlined data analysis, improved data visualization, and quicker data manipulation.
- Business Professionals: quickly opening and viewing the CSV file, enabling them to make instant data-driven decisions, create reports and share the findings with colleagues.
    - Benefits: Improved decision-making, time-saving, and enhanced data-sharing capabilities.
- Business owners and managers: Managing inventory, sales data and customer records. They can also utilize the application to analyse sales trends and make informed decisions.
    - Benefits: Better inventory management, improved sales insights, and informed business decisions.

3.5 Commercial potential

Here are some aspects to consider for the commercialization of such an application in the future:

- Freemium Model: Offering a basic version of the application for free, allowing users to perform essential tasks with limitations. Then, provide a premium or paid

version with advanced features, additional uploading size, or enhanced support. This model can attract a wide user base and convert some of them into paying customers.

- API Integration: Develop API (Application Programming Interfaces) to allow integration with other software tools and platforms. This can expand the application's reach and make it an essential part of various business processes.
- Customer Support and Training: Offer premium customer support, training and consulting services and address any issues or challenges they may face.

3.6 Limitations

Our application works flawlessly with ordinary CSV files, however, there could be some limitations when faced with complex data, for example:

- File size limitation: CSV files can vary greatly in size, while our application has an maximum file upload limit of only 200MB.
- Lack of Advanced Analysis Tools: While the application can help with basic data manipulation, it does not offer advanced statistical analysis or machine learning capabilities, which might be required for in-depth data analysis.
- Data Import Compatibility: the application may be unable to handle all CSV variations, and issues with delimiters, encoding, or file formats may arise during importing operations.

3.7 Improvements

Although the application is robust and informative, there should be more advanced functionalities developed on it for further superiority. The following improvements could be considered.

- Advanced Data Visualization: Incorporate advanced charting and graphing capabilities to enable users to visualize data trends and patterns more effectively. Support a wide range of chart types (e.g., bar charts, scatter plots, heatmaps) to cater to diverse data needs.

- Data Transformation Tools: Enhance the application with powerful data transformation features, including the ability to merge, split, and pivot columns, apply mathematical operations, and create calculated fields.
- Machine Learning Integration: Incorporate machine learning and predictive analytics capabilities for advanced data analysis and pattern recognition.
- User Feedback Mechanism: Implement a feedback system that allows users to suggest improvements and report issues directly from within the application.

4. Reflecting on Building Data Product

The Data Explorer Web App this report introduced is one of the examples of the data products that data scientists build to perform exploratory data analysis (EDA) to understand the input datasets, identify issues and limitations thereby can have comprehensive analysis of the data and extract the insights and patterns behind the collected datasets.

Building data products can automate tasks, therefore enhancing work efficiency and Improving data analysis accuracy in daily data-related projects as data products can follow pre-programmed processes to perform tasks consistently without human errors. Data products can handle repetitive tasks faster and with greater accuracy than manual processes. This will free up data scientists to focus on more complex projects. It is not only beneficial for individual data scientists but also enhances the overall competitive advantage of the organisation in the industry.

Nowadays data analysis is widely applied across various industries. For instance, with well-functioned models, a supermarket can set better deals for its products, leading to increased sales and profits. A marketing company will be able to target its audiences based on the analysis of their preferences and characteristics. This kind of data product has become one of the most important revenue generators for IT/data analysis companies and data-driven consulting companies. A good data product hinges on the expertise of data scientists for its development, ongoing maintenance, and continuous improvements. Therefore, the developing data products skills are essential for data scientists in terms of career growth and individual development.

Data scientists require a combination of skills and technologies to develop data products. Some of the key skills are:

- Data analysis: data analysis includes identifying the business question, collecting the raw data sets, cleaning the data, analysing the data and interpreting the result.
- Statistical analysis: statistical analysis is one of the most important data analysis methods, it includes the process of collecting, organizing, interpreting, and drawing conclusions from data.
- Programming: Python and R are widely used for data analysis, while languages like Java, and C++ may be used for building applications and interfaces.

One potential use case for developing innovative data products our team thinks for commercial applications is a loan default production model. Currently, most banks still rely on companies' financial reports and collateral valuation results to analyse the likelihood of potential default. The disadvantages include financial reports and valuations being lagging Indicators which may not be able to provide real-time insights and financial reports can be subject to manipulation which may affect the accuracy. However, the transactions that happen on companies' or individuals' bank accounts are usually more accurate and real-time. Therefore, since it is easy for banks to collect customers' transaction history, our team believes a data product can be developed to analyse customers' financial situation and predict the default rate.

5. Collaboration

5.1 Individual Contributions

5.1.1 Siheng Mu

Siheng is tasked with handling the numerical data analysis tab. This involves filtering the dataset, processing a user-selected numerical column, and extracting meaningful information into tables and a data visualisation graph. Siheng regularly seeks out teammate opinions and engages in self-reflection to identify areas for development in his own work. Siheng also actively participates in brainstorming sessions and feedback sessions with other team members to improve the EDA online application.

5.1.2 Xinhui Wang

Xinhui was tasked with coding with the data frame tab that is used for quick data exploration. It allows users to upload a CSV and extract information on the count of rows and columns. It also provides basic data quality checks like the count of missing rows and duplicates.

The second session of the tab allows users to view a snapshot of the data. Users have the choice to view between 5 to 50 rows from the top, bottom or a random sample of the data provided.

Xinhui also reviewed the pull requests from other members, and suggested the codes where appropriate to ensure the web app runs cohesively end to end.

5.1.3 Aditi Vyas

Aditi Vyas played a key role in developing the Text Data Analysis tab for the CSV Explorer Web App. Her work focused on enhancing the tab's functionality, making it easier for users to conduct in-depth analyses of textual data. By streamlining the user interface and the backend processes, she ensured that the app could handle complex analyses with ease. Her collaborative efforts in regularly refining the tool, based on team input, contributed significantly to the project's successful outcome.

5.1.4 Sinh Thanh Nguyen

Sinh Thanh's task was developing the DateTime tab for the CSV Explorer Web App. He also initialized the repository on GitHub, tested outcomes from development branches, and reviewed and performed modifications if there were any errors or conflicts when

merging development branches to “main”. Besides, Sinh Thanh took part in synchronizing tabs’ appearance and standardizing users’ experience.

5.2 Group Dynamic

Our team created a WhatsApp group with four team members in it on 13 Oct 2023 for daily communication and quick chats. A GitHub repository was also created on the same date for us to work on the coding.

Our first meeting was on 16 Oct 2023 via Google meeting, we divided the coding part for each member as follows:

1. tab_df: Xinhui
2. tab_num: Siheng
3. tab_text: Aditi
4. tab_date: Sinh Thanh

The agreed due date for the coding part was 31 Oct 2023 and the second meeting was scheduled on 1 Nov 2023.

The coding part was finished by 31 Oct 2023 as planned, Henry and Sinh also made a couple of amendments to ensure the codes were working well.

We had our second meeting on 1 Nov 2023 to discuss the final report. We aimed at finishing the first 4 sections by 5 Nov 2023. The work was divided as follows:

1. Executive Summary: Aditi
2. Introduction: Siheng
3. Web App Presentation: Sinh Thanh
4. Reflecting on Building Data Product: Xinhui

A Google Share document was created for us to work on the report.

The third meeting was scheduled for 5 Nov 2023. We did some proofreading on each other’s part and shared some suggestions and opinions on how to improve the report. And divided the rest of the work as follows:

1. Individual Contributions and issues faced – individual work

2. README.md: Sinh
3. Group Dynamic: Ellie
4. Ways of Working Together: Siheng
5. Conclusion: Aditi

The last meeting happened on 9 Nov 2023. We did some grammar checks for the report and finally submitted the assignment.

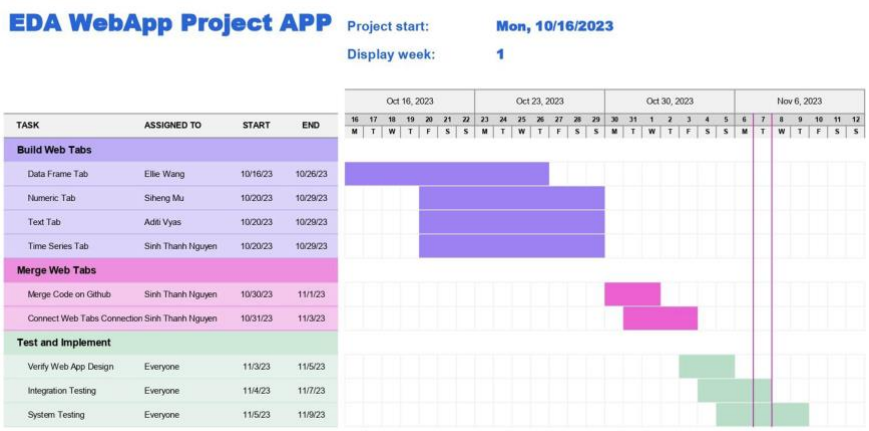
In general, our team cooperation is extremely smooth. Everyone can complete their assigned tasks on schedule, and the atmosphere is very harmonious and friendly. We have also developed friendships with each other during the collaborative process.

5.3 Ways of Working Together

In the field of software development, reaching effective code collaboration is crucial to project success, deadline compliance, and code quality maintenance. The Waterfall project management methodology and GitHub are two effective platforms that can aid in facilitating this kind of cooperation.

5.3.1 Waterfall Project Management

Software development is approached methodically and sequentially using the Waterfall project management technique. It consists of several phases, each of which builds on the one before it. Because of the incremental approach of the Waterfall technique, extensive planning and documentation are possible, resulting in well-defined project requirements and carefully managed adjustments.



5.3.2 GitHub

GitHub is essential for collaborative code development. By offering developers a centralised platform to work on their code concurrently, GitHub improves the Waterfall approach through version control and collaboration capabilities. Code quality and consistency are preserved throughout the project because teams can easily generate, review, and integrate code changes. The pull request, code review, and issue tracking capabilities of GitHub also help team members communicate clearly and take responsibility for each other's work, which is in line with the Waterfall method's emphasis on comprehensive documentation and defined procedures.

The advantages of GitHub that allow for version control, issue tracking, and code review are enhanced by Waterfall's distinct stages and documentation. This synergy promotes effective code collaboration, upholds code quality, and guarantees that projects go forward smoothly with clearly defined roles and responsibilities.

5.4 Issues Faced

The application must determine if any data has been imported for each dataset feature analysis section. If not, it must ensure that an error message is not displayed on the frontend display and that a blank analysis page is displayed instead. Nevertheless, there was some trouble ruling out an additional instance of a data import problem; this was fixed by some trial and error.

Inconsistent coding styles of team members were also considered as a barrier. Each team member has their way of implementing ideas on the app. Therefore, it took quite a long time to resolve the conflicts when merging their work.

Moreover, during the testing phase of the project, there were also conflicts between imported packages, which took a lot of effort to deal with.

6. Conclusion

The CSV Explorer Web App project concluded with several significant accomplishments:

6.1 Key outcomes

- Interactive tabs for each data type (text, numeric, datetime) allow for a tailored analytical experience.
- Users can now engage with data through an intuitive interface, with features such as dynamic visualisations and real-time statistical analyses.

6.2 Insights

- The app's design effectively reduces the complexity traditionally associated with EDA, making it more accessible to a wider range of users.
- The tool's capability to handle large datasets with efficiency has been a standout feature.

6.3 Achievements

- The application met all predefined objectives, providing a solution that aligns well with user needs and stakeholder requirements.
- Positive feedback from stakeholders and users alike indicates the application's success and its potential impact on the data analysis workflow.

Reflecting on the project's journey, the following points encapsulate its triumphs and roadmap for future enhancements:

6.4 Project Success

- The app has been successful in achieving its goal of simplifying EDA processes, as evidenced by user testimonials and stakeholder endorsements.
- Meeting the project's benchmarks, the tool has demonstrated its value by improving the efficiency and depth of data analysis.

6.5 Future work and Recommendations

- Incorporate advanced analytical features like predictive analytics through machine learning integration.
- Improve data cleaning processes, broadening the scope to handle a wider variety of data inconsistencies.
- Support for additional file formats could be explored to make the tool more versatile.
- Continued refinement based on user feedback will ensure the tool evolves to meet changing data analysis needs.

6.6 Next Steps

- Implement a feedback loop for continuous user-driven development.
- Plan for regular updates and maintenance to ensure the application remains at the forefront of EDA technology.
- Consider partnerships with academic and professional data science communities to keep the tool updated with the latest trends and methodologies in data analysis.

In conclusion, the CSV Explorer Web App has set a new standard in EDA tools, providing a combination of depth, flexibility, and user-friendliness. The ongoing commitment to enhancement and adaptability will ensure that the application continues to serve the needs of the data science community effectively.

7. References

Van Der Aalst, W., & van der Aalst, W. (2016). *Data science in action* (pp. 3-23). Springer Berlin Heidelberg.