

Assignment 3

Collaborative Development of Data Explorer Web App

Group 8

2024-11-08

Fatemeh Elyasifar (25589351)

Krishnan Unni Prasad (25225362)

Prisa Senduangdeth (25402088)

https://github.com/KrishUnni-Z/dsp_at3_group8.git

94692 - Data Science Practise
Master of Data Science and Innovation
University of Technology of Sydney

Table of Contents

1.	Executive Summary	2
a.	Problem Statement	2
b.	Outcome	3
2.	Introduction	4
3.	Web App Presentation	5
a.	Purpose and Main Functionalities	5
b.	Set up and Launch Instructions	10
c.	Potential Users and Use Case	11
d.	Potential Commercialisation	12
e.	Limitations and Potential Improvements	12
4.	Reflecting On Building Data Product	13
a.	Importance for Data Scientists to Develop Data Products	13
b.	Essential Skills and Technologies for Developing Data Products	13
c.	Other Types of Data Products Developed by Data Scientists	13
d.	Reflection on AI Advancements and Innovation in Data Products	14
5.	Collaboration	15
a.	Individual Contributions	15
b.	Group Dynamic	15
c.	Ways of Working Together	15
d.	Issues Faced	16
■ ■ ■		16
6.	Conclusion	17
a.	Key Findings and Insights	17
b.	Project Success and Stakeholder Requirements	17
c.	Recommendations and Future Work	17
7.	References	19
■ ■ ■		

1. Executive Summary

This project involved developing a comprehensive web application for exploring and analysing CSV file datasets through an interactive interface. The application is designed with four main tabs to aid users in viewing and interpreting their dataset:

- **Overall Information Tab:** Provides a summary of the dataset, including row and column counts, duplicates, and missing values.
- **Numeric Column Tab:** Allows users to select 'numeric' columns and displays detailed statistics such as unique values, missing values, and distribution charts.
- **Text Column Tab:** Enables users to select 'text' columns and view detailed analyses, including unique value counts, empty strings, and character-type distributions.
- **DateTime Column Tab:** Facilitates the exploration of 'datetime' columns with information on date ranges, counts of specific dates, and visualisations of data distributions.

The primary aim of this web application is to help users streamline the process of exploratory data analysis (EDA) by allowing them to quickly upload their datasets in CSV files and gain insights into the data through an intuitive interface.

a. Problem Statement

One line Statement: The application aims to address the need for a user-friendly EDA tool that removes coding requirements, offering value to data analysts, organizational departments, business professionals, and academic users who seek fast, accessible data insights without complex tools.

The challenge addressed by this application is the need for an efficient and user-friendly tool that allows users to perform initial EDA quickly without extensive coding or complex data tools. The application is particularly valuable for:

- **Data analysts** who need to perform initial data checks and summary analyses.
- **Departments across an organisation** (e.g., marketing, sales, HR) that require quick data insights for decision-making.
- **Entrepreneurs and business professionals** who want to understand their data without delving deep into programming or data science tools.
- **Academic professionals and students** who want quick and efficient access to data insights to support their research and studies.

The ability to upload a CSV file and immediately visualise and explore the data helps save time and enhances productivity, providing users with an accessible approach to data analysis.



b. Outcome

The project resulted in a fully functioning web application that meets the outlined requirements. Users can easily navigate through the tabs to perform various types of analyses, from simple overviews to detailed breakdowns of numeric, text, and date-time columns. Additionally, the project showcased effective teamwork using GitHub for version control, allowing for collaborative coding. This helped create a cohesive workflow, reinforcing best practices in software development and team management.





2. Introduction

This project's main objective was to develop an interactive web application that facilitates quick and easy exploratory data analysis (EDA) for CSV files. The desired outcome was to create a tool that simplifies data visualisation and understanding through intuitive features, allowing users to:

- View overall dataset statistics.
- Analyse specific numeric, text, and datetime columns with ease.
- Interact with the data through visualisations and summary tables.

Stakeholders and the requirements

- **Data Analysts:** Need an easy-to-use tool for quick EDA to support more profound analysis.
- **Department Teams** (e.g., marketing, sales, HR): Require initial data insights to inform strategies and decision-making without extensive data manipulation.
- **Entrepreneurs and Business Professionals:** Benefit from simplified tools for data exploration to support data-driven business insights.
- **Academic Professionals and Students:** Need a tool to help them speed up their initial data exploratory analysis to gain access to insights for their research and studies.

This project addresses these needs by providing a web application that is easy to use but still provides comprehensive analysis. By meeting these objectives, the project enables stakeholders to gain quick, actionable insights, streamline their data exploration processes, and support data-driven decisions.



3. Web App Presentation

a. Purpose and Main Functionalities

The purpose of this web application is to provide users with an interactive and user-friendly tool for conducting exploratory data analysis (EDA) on their dataset by utilising CSV files. It is designed to simplify the process of understanding and visualising data in the initial phase by allowing users to upload their own datasets and access key insights across four main data categories. The application is structured with four main tabs, allowing users to easily navigate and analyse different types of data—overall information, numeric, text, and date-time—without needing advanced programming skills.

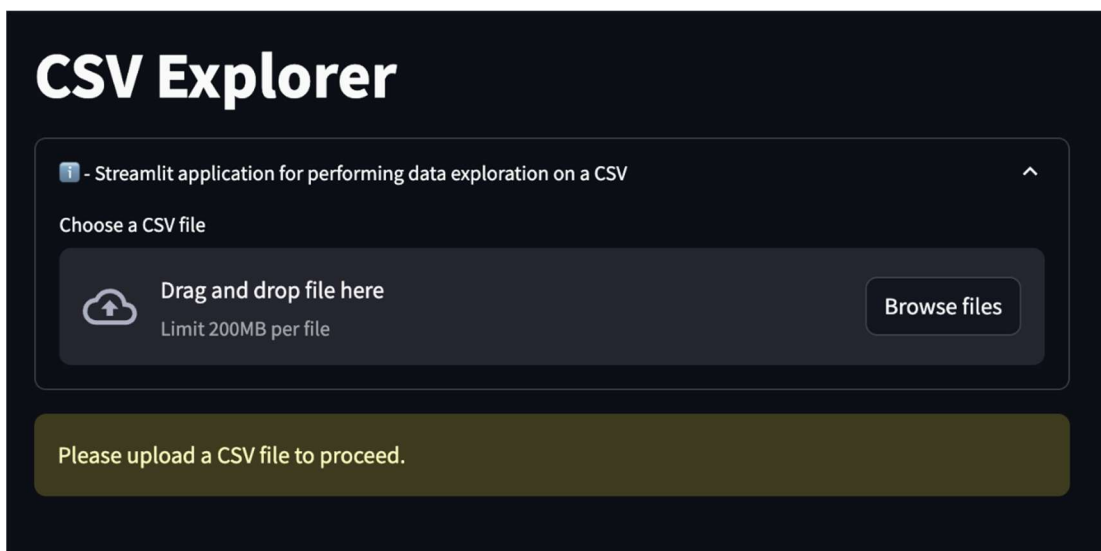


Figure 1. The main page where users choose to upload their CSV file.

1) Data Frame Overview

This tab offers an overview of the dataset, providing a summary of the number of rows, columns, duplicate rows, and missing values. Additionally, the application displays the column names along with their data types and memory usage. Users can also view the head, tail, or a sample of the dataset, selecting any number of rows from 5 to 50.

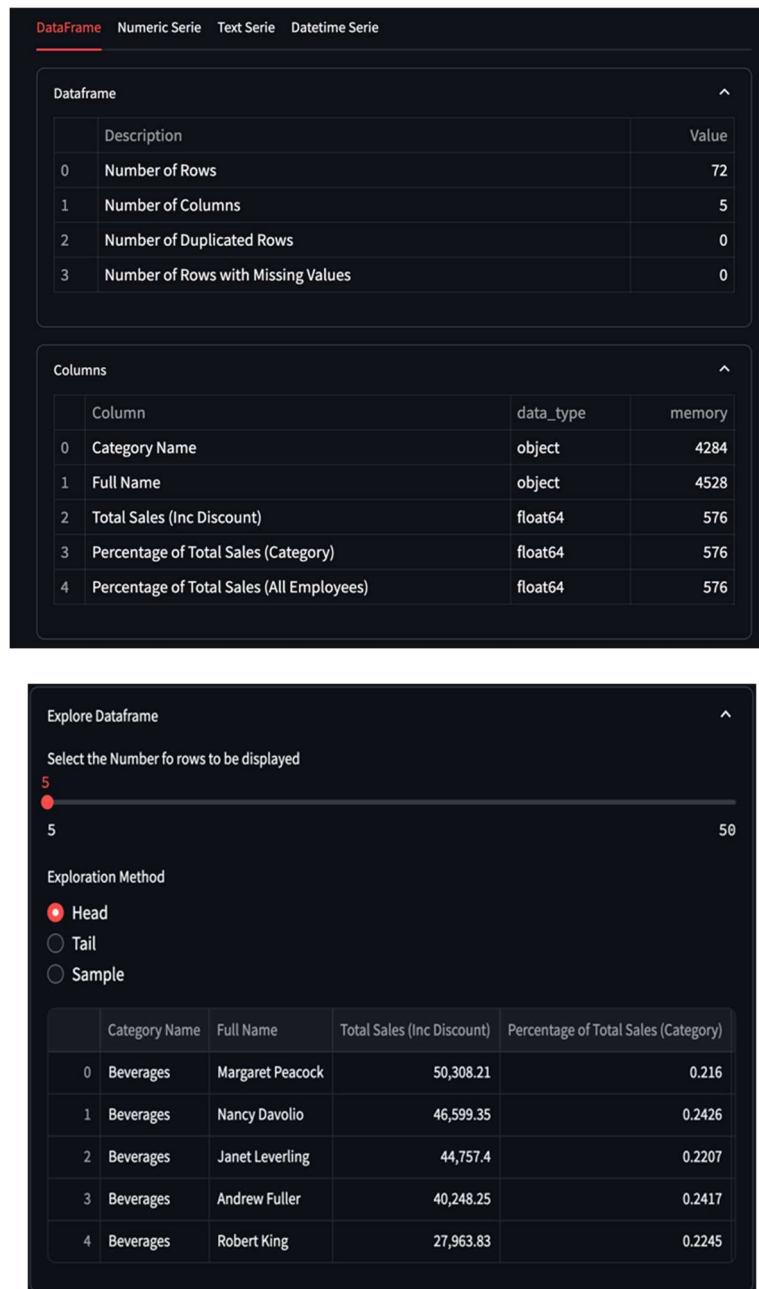


Figure 2. DataFrame tab, showing overall information of the dataset.

2) Numeric Serie

This tab allows users to explore each numeric column in the dataset in detail, as shown in Figure 3. By selecting a numeric column from a dropdown menu, users can view essential statistics and a column summary, including the number of unique values, missing values, occurrences of zero, and negative values, along with descriptive statistics such as the mean, standard deviation, minimum, maximum, and median, as shown in Figure 4. Moreover, an interactive histogram visually displays the distribution of values within the column, useful for examining distributions such as the number of items per shopping cart. Additionally, users can view a table listing the top 20 most frequent values, along with their occurrences and percentages, providing a deeper understanding of the numeric data.

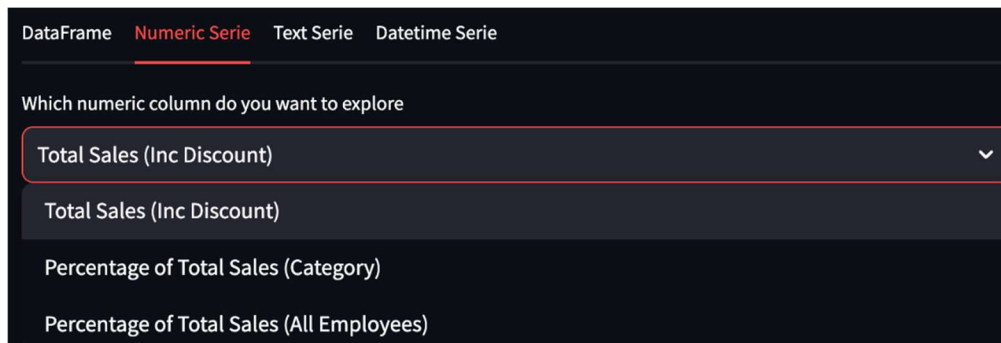


Figure 3. Users can select numeric columns to view the insights.

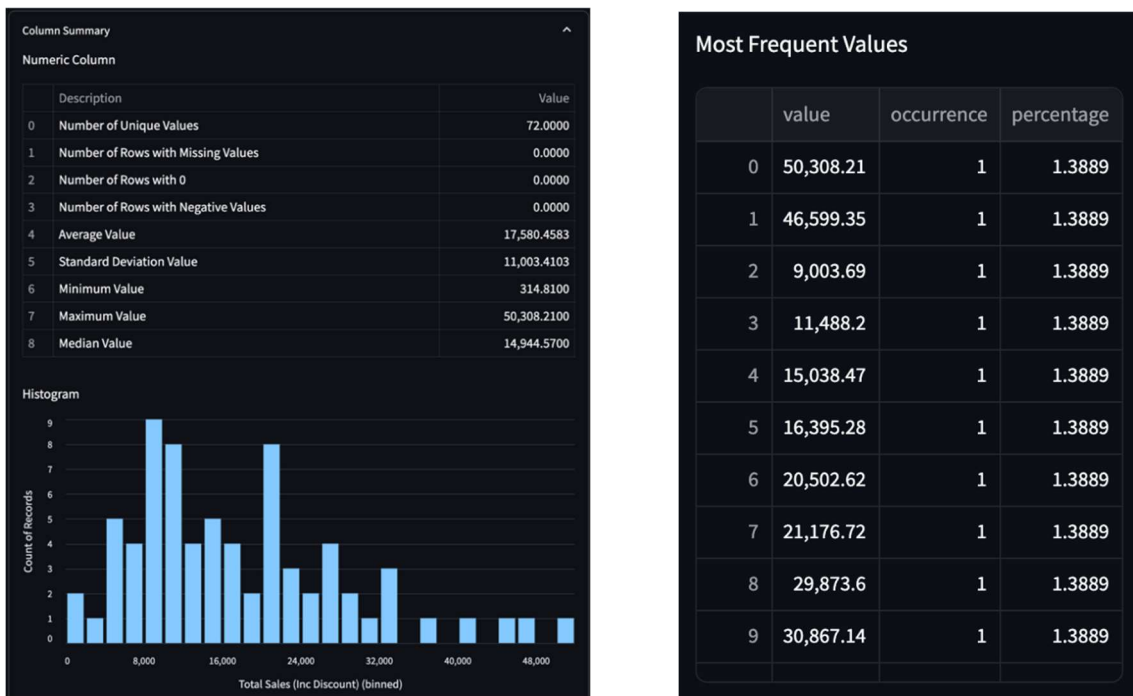


Figure 4. Numeric column insights.

3) Text Serie

This tab is designed to analyse text-based columns within the dataset. After selecting a text column, users are presented with key information, including the number of unique values, missing values, rows containing empty strings, rows with only whitespace, and counts of rows with exclusively lowercase or uppercase characters. The tab also identifies the most common value (mode) in the selected column. An interactive bar chart visually represents the frequency of each unique value in the column, and a table displays the top 20 most frequent values with their occurrence counts and percentages, providing users with a comprehensive analysis of text data characteristics.

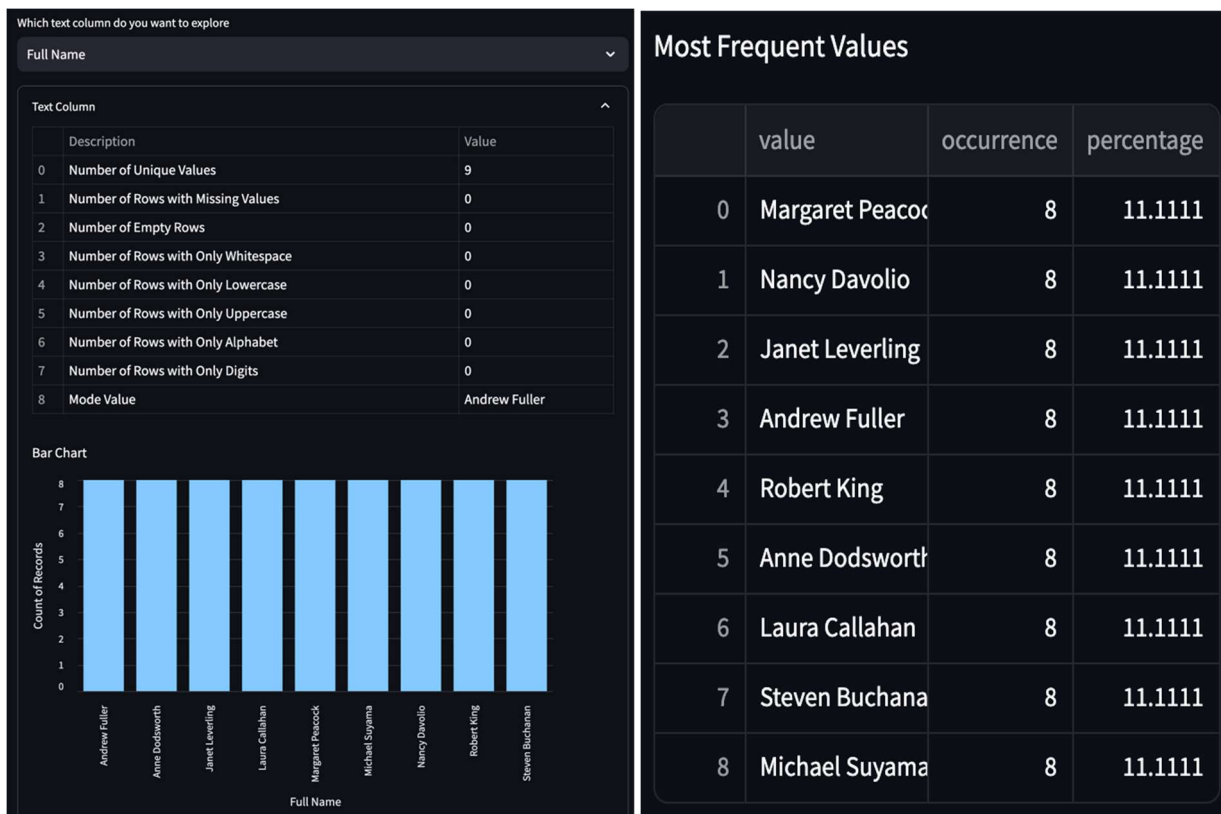


Figure 5. Text column insights

4) DateTime Serie

This tab enables users to analyse datetime columns. If no datetime columns are detected, the application will attempt to convert text columns into datetime format to identify any dates stored as text. For each datetime column, the tab provides essential information, such as the number of unique dates, missing values, minimum and maximum dates, and counts of dates that fall on weekdays, weekends, or in the future. Additionally, it highlights occurrences of specific dates like 1900-01-01 and 1970-01-01. An interactive histogram allows users to explore the distribution of dates within the column, and a table lists the top 20 most frequent dates, along with their percentages, offering valuable insights into time-related data trends.

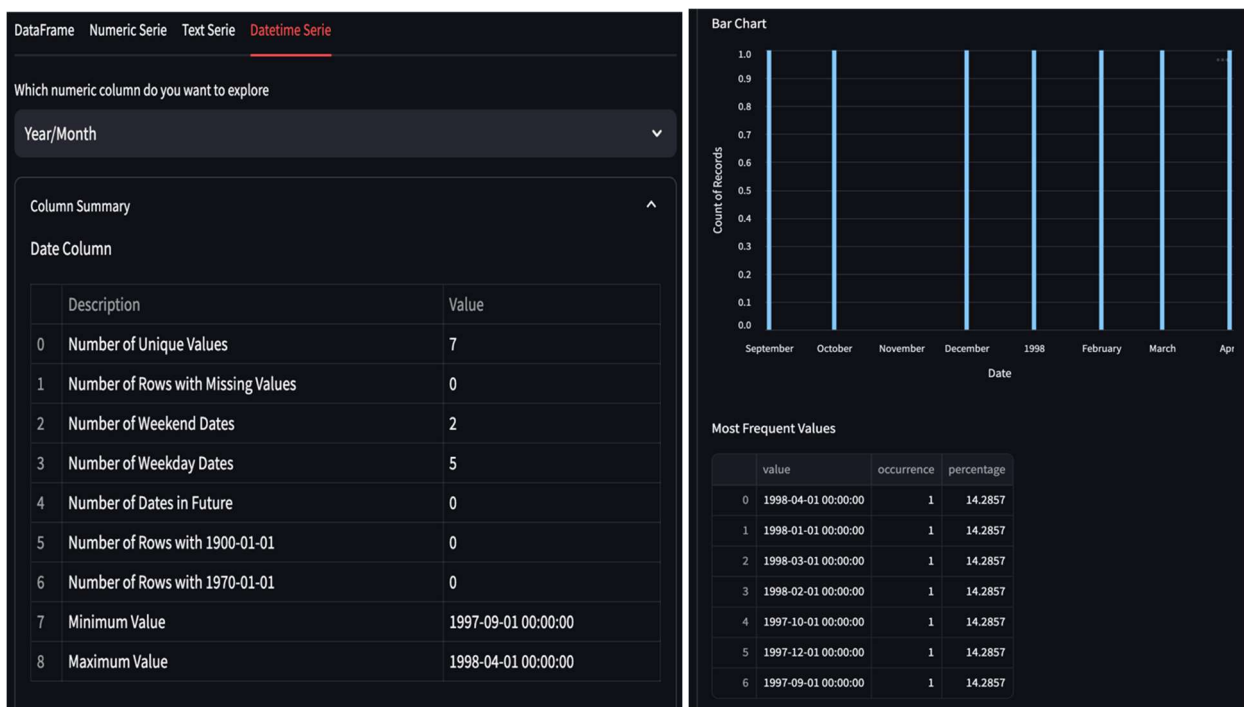


Figure 6. DateTime column insights

b. Set up and Launch Instructions

The development of this web application used Python version 3.12.4 and has a prerequisite of the following libraries:

- streamlit: version 1.13.0
- pandas: version 2.0.3
- altair: version 4.2.0
- Other Python standard libraries, including sys, os, and pathlib.

Users should follow the following steps to install and launch this web application:

1. Clone the GitHub repository to the local machine using the following command in the terminal:

```
git clone https://github.com/KrishUnni-Z/dsp_at3_group8.git
```

- Alternatively, users can also download the web application file by going to the given website.
- Users then navigate to the cloned or downloaded project.

```
cd dsp_at3_group8
```

2. Install the necessary libraries to launch the web application by using the following code.

```
pip install -r requirements.txt
```

3. After the installation has finished, the next step is to launch the web application by using the following code.

```
streamlit run app/streamlit_app.py
```

- By running the code, a URL will appear on the terminal, which users can use to enter the web application.

By following these three steps, users can now access the web application through their browsers to conduct initial exploratory data analysis on their datasets.

Remark: The application currently supports CSV files only. Users should ensure the file uploaded is in CSV format and has clean data where possible to avoid unexpected issues.

c. Potential Users and Use Case

The application's main purpose is to make exploratory data analysis accessible and efficient, providing value to a wide range of users across various fields. Below are the potential users and their specific use cases:

Data Analysts

- **Use case:** Data analysts can use this application for initial data exploration to understand the nature of the data before conducting more complex analyses. They can use the application to obtain insights into the dataset structure, missing values, and distributions, allowing analysts to make preliminary assessments without writing code.
- **Benefits:** Helps save time on basic exploration that can be initially reported to other non-technical stakeholders, allowing analysts to focus on more advanced analysis and model building.

Organisations with different departments

- **Use case:** The application can also be used by teams across various departments, such as marketing, sales, and HR, to gain insights into their data without requiring technical expertise. For instance, the Sales team can analyse transaction data, while the HR team can examine employee demographics or performance data. This feature is particularly helpful when preparing presentations that need quick, data-driven insights.
- **Benefits:** Enabling non-technical users to quickly gain insights into their data, facilitating data-driven decision-making and enhancing collaboration with data teams.

Entrepreneurs and Small Business Owners

- **Use case:** Business owners can upload customer or sales data to quickly identify trends, demographics, and key insights relevant to their operations. For example, they can analyze top-selling products, peak sales periods, or customer demographics to make data-driven decisions, without spending extensive time or effort.
- **Benefits:** Provides an easy-to-use solution for data exploration, allowing business owners to make informed decisions without hiring data specialists or investing in complex analytics software.

Academic Professionals and Students

- **Use case:** Academic professionals and students often work with diverse datasets from various sources during research. This application helps them efficiently analyze and extract insights, allowing them to quickly assess the data's suitability for their studies and make sense of the underlying patterns.
- **Benefit:** Saves time in understanding the data, enabling users to evaluate multiple data sources more quickly and effectively.



d. Potential Commercialisation

This web application has commercial potential due to being a user-friendly exploratory data analysis (EDA) tool aimed particularly towards non-technical users. It could be offered as a standalone product for small to medium-sized businesses or an add-on for existing business intelligence platforms. Moreover, with updates providing more features for complex data insights, this application has the potential to be a cloud-based SaaS solution for larger enterprises. Additionally, it could serve well as an educational tool for data literacy programs in universities and training sessions or as a customisable platform for consulting firms to provide quick data insights to clients. The application's simplicity and accessibility make it a valuable tool for organisations and individuals seeking data-driven insights without requiring advanced technical skills.

e. Limitations and Potential Improvements

As with any initial version, the current release of the CSV Explorer web application has certain limitations. Addressing these through future updates and enhancements will improve its usability, scalability, and functionality, making it more valuable for a wider range of users and increasing its potential for commercialization. Below are some identified limitations and suggested enhancements for future development.

Limitations

- Only supports CSV file format, limiting data input flexibility.
- May experience performance issues with huge datasets.
- Limited customisation options for visualisations and data analysis.

Potential Improvements:

- Add support for additional file types (e.g., Excel, JSON, databases).
- Optimise performance for handling large datasets more efficiently.
- Expand visualisation options for deeper and more complex insights.



4. Reflecting On Building Data Product

a. Importance for Data Scientists to Develop Data Products


For data scientists, learning data product development is crucial, as it facilitates the transformation of raw data into valuable insights and practical tools that address real-world problems. From a career perspective, it expands their skill set beyond data analysis to include building functional products that can automate decision-making processes. This capability is highly sought after by businesses, as it helps streamline workflows, provide real-time insights, and support data-driven decisions across teams. Additionally, it fosters better collaboration between data scientists and stakeholders, ensuring that data science projects are aligned with business objectives.

b. Essential Skills and Technologies for Developing Data Products

1. **Programming:** Developing a data product requires proficiency in Python, R, and JavaScript for data manipulation, algorithm development, and web interface creation.
2. **Machine Learning:** Machine learning is the critical characteristic of understanding machine learning algorithms and frameworks, e.g., TensorFlow and Scikit-Learn, which would build predictive models.
3. **Web Development:** Familiarity with frameworks like Flask, Django, and Streamlit is critical for developing user-facing data products.
4. **Data Engineering:** Experience with SQL, NoSQL databases, and cloud platforms (e.g., AWS, Google Cloud) is vital for managing and processing large datasets.
5. **Data Visualization:** Mastery of visualisation libraries (e.g., Matplotlib, Plotly) and tools (e.g., Tableau) helps create interactive dashboards and reports, which is necessary when communicating with other stakeholders.

c. Other Types of Data Products Developed by Data Scientists

- **Recommendation systems:** Used in e-commerce, media platforms, and social networks to recommend products or content based on user preferences.
- **Predictive analytics tools:** Employed for forecasting demand, sales, and other key business metrics, helping organizations make informed decisions.
- **Chatbots and virtual assistants:** AI-powered systems that automate customer service and support, enhancing user experience and operational efficiency.
- **Data dashboards:** Interactive platforms that display real-time business metrics and KPIs, enabling stakeholders to monitor performance and make data-driven decisions.

- 
- **Automated report generation tools:** Generate insights or reports from the data without human intervention, reducing the need for manual analysis and saving time.

d. Reflection on AI Advancements and Innovation in Data Products

AI is proactively leading the data revolution in best practices, bringing forth intelligent and self-learning systems for responding, realistically interpreting, and adapting to trends in customer behaviour or trend data evolution. Techniques such as deep learning, natural language processing, and reinforcement learning enable a product to be personalised, autonomous, and perform effectively. By quickly processing large volumes of data, AI enhances the capabilities of existing data products, maximizing their utility for businesses and consumers alike. To stay competitive, data scientists must integrate AI features into their products, enabling greater accuracy, automation, and scalability in their solutions.



5. Collaboration

a. Individual Contributions

- **Fatemeh Elyasifar (25589351)**: Fatemeh worked on the project's **Date Tab** and **Numeric Tab**, focusing on data preprocessing and transformation tasks related to these features. She was also involved in troubleshooting and fixing minor code issues to ensure the smooth functionality of these tabs. Additionally, she contributed to reviewing and verifying the project report, helping ensure accuracy and completeness.
- **Krishnan Unni Prasad (25225362)**: Unni focused on the **Text Tab**, contributing to text data processing and feature engineering. He also played a key role in writing the **Readme** and **Report**, ensuring that all documentation was clear and complete.
- **Prisa Senduangdeth (25402088)**: Jack worked on the **Df Tab**, performing the necessary transformations and analyses on the dataset. He was also mainly involved in drafting the **Report** and addressing any last-minute revisions to the code to minimise errors and project documentation.

Combined teamwork allowed all members to correct little issues and check the Report for uniformity and accuracy in all segments.

b. Group Dynamic

From the beginning to the end of the project, the group dynamics were highly supportive and collaborative. Team members maintained active communication through WhatsApp and Microsoft Teams, ensuring timely updates and responses to any questions. Regular information sharing kept everyone aligned with project goals, deadlines, and deliverables. To further enhance collaboration, we used GitHub to dynamically work on and build each tab of the project simultaneously, allowing us to efficiently coordinate and integrate our individual contributions. This approach helped us address challenges proactively, maintain focus on deadlines, and ensure that all tasks were completed on time, leading to a successful project outcome.

c. Ways of Working Together

To manage the project effectively, we ensured that all task assignments and follow-ups were clearly defined. The project report was drafted using MS Word, while GitHub was used for code versioning and to facilitate collaboration on the development of different project tabs. We kept the entire team informed and up to date through a dedicated communication channel. Team meetings were held regularly, where progress was reviewed, and any necessary decisions were made collectively. Everyone contributed to the decision-making process, offering input and assistance as needed, ensuring that tasks were completed on time and the project moved forward smoothly. This approach, combined with the use of tools like GitHub and MS Word, helped streamline the management and execution of the project.



d. Issues Faced

A few challenges were encountered during the project, primarily related to code tuning and initial communication. At the beginning, there were difficulties in clarifying expectations and aligning everyone on the scope of work. This was addressed by improving communication through frequent updates and establishing clearer guidelines for task division. The team also faced challenges in fine-tuning the code to meet performance requirements. These issues were resolved by collaborating on debugging and optimizing the code to ensure smooth functionality.

The project highlighted the importance of clear and early communication in avoiding misunderstandings and ensuring everyone is aligned. Regular progress updates and task assignments helped keep the project on track. In future collaborations, more structured planning in the initial phase would help mitigate challenges related to scope and communication. Additionally, ensuring all team members have access to the necessary resources early on is crucial to prevent delays.



6. Conclusion

This project has provided a robust and user-friendly exploratory data analysis (EDA) tool that meets the specific needs of data analysts, business teams, entrepreneurs, and academics. The web application enables users to upload and explore CSV datasets across four tailored tabs—general dataset overview, numeric column analysis, text column insights, and datetime analysis. Each tab offers a focused set of insights, allowing users to explore the data from multiple perspectives without requiring advanced technical skills.

a. Key Findings and Insights


1. **Overall Dataset Information:** The application effectively provided quick summaries on dataset size, missing values, and column details, offering immediate insight into data quality and structure. This feature proved essential for initial assessments, helping users assess if a dataset is ready for analysis.
2. **Numeric Columns Analysis:** Users could dive deep into numeric data distributions, variance, and identify outliers. This tab included key visualisations, such as histograms and box plots, which were instrumental in highlighting underlying patterns and potential data anomalies. These insights helped users understand central trends and detect skewed distributions that might impact downstream analyses.
3. **Text Column Exploration:** By offering word frequency counts and word clouds, the text analysis tab enabled users to capture dominant terms, recurring themes, and language patterns in text data. This insight was especially beneficial for users in marketing or customer service sectors, where understanding customer sentiment and feedback trends can provide actionable insights.
4. **Datetime Columns Analysis:** For datetime data, the application provided periodic patterns and seasonality insights, helping users understand trends over time. This was particularly valuable for business forecasting or trend analysis in industries like retail, where seasonality plays a significant role in decision-making.

b. Project Success and Stakeholder Requirements

The project successfully aligned with stakeholder needs by offering an accessible, comprehensive EDA solution tailored for various user roles. The choice to focus on CSV file support with in-depth, targeted analysis options was well-received, fulfilling the requirement for a streamlined tool that balances simplicity with analytical depth. The collaborative development approach on GitHub also facilitated efficient feature integration, ensuring all team contributions were harmonised.

c. Recommendations and Future Work

To further enhance the application's value, the following next steps are recommended:

- 
1. **Expand File Format Support:** Adding support for Excel, JSON, and other file types would make the application even more versatile, allowing users to analyse datasets from various sources.
 2. **Advanced Statistical Analyses and Transformation Options:** Including statistical tests (e.g., correlation tests, chi-square tests) and transformation tools (e.g., log transformation, normalisation) would add depth to the numeric analysis tab, benefiting users looking for a more sophisticated level of data preprocessing.
 3. **Automated Report Generation:** Implementing a feature to automatically generate summaries or detailed reports based on EDA findings would save users time and provide a convenient way to document and share insights.
 4. **Basic Predictive Modeling Capabilities:** Adding preliminary predictive analytics tools, such as trend forecasts or clustering for quick segmentation, would introduce users to predictive insights, bridging EDA with early-stage modelling.

Overall, the project has laid a strong foundation for a powerful EDA tool, with a potential for further enhancements that would deepen user insights and extend the application's utility across broader data analysis tasks.



7. References

- Streamlit. (2022). *Streamlit (Version 1.13.0)* [Computer software]. Streamlit, Inc. Available from <https://streamlit.io>
- McKinney, W. (2023). *pandas (Version 2.0.3)* [Computer software]. pandas development team. Available from <https://pandas.pydata.org>
- VanderPlas, J. (2022). *Altair: Declarative statistical visualization library (Version 4.2.0)* [Computer software]. Available from <https://altair-viz.github.io>
- Python Software Foundation. (2023). *Python Standard Library* [Computer software]. Available from <https://docs.python.org>

