

# Assignment 3

## Collaborative Development of Data Explorer Web App

---

Group 11

9th November 2023

Vishal Raj (14227627)

Rohit Sharma (24590960)

Ronik Jayakumar (24680264)

Zhicheng William Dai (13628019)

GitHub: [dsp\\_at3\\_group\\_11](#)

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

## Table of Contents

<b>1. Executive Summary</b>	<b>2</b>
Project Overview:	2
Problem Statement and Context:	2
Significance of the Project:	2
Achieved Outcomes:	2
<b>2. Introduction</b>	<b>3</b>
Project Objectives and Goals:	3
Stakeholders and Requirements:	3
Addressing Stakeholder Requirements:	4
Relevance and Context of the Project:	4
Achieving Project Outcomes:	4
<b>3. Web App Presentation</b>	<b>5</b>
<b>4. Reflecting On Building Data Product</b>	<b>6</b>
a. Importance of developing data products	6
b. Key skills and technologies required for developing data products	7
c. Data Product Use Cases	7
d. Reflection on the current trend of AI advancement	8
<b>5. Collaboration</b>	<b>9</b>
a. Individual Contributions	9
b. Group Dynamic	11
c. Ways of Working Together	12
d. Issues Faced	12
<b>6. Conclusion</b>	<b>13</b>
<b>7. References</b>	<b>14</b>

# 1. Executive Summary

## Project Overview:

The “CSV Explorer app” project of our group was initiated to deliver an innovative data analysis solution that democratises the exploration of complex datasets. The project's core was the development of a web-based application tailored to simplify the process of data interaction for users with varying levels of expertise in data handling. Recognising the challenges organisations face in leveraging their data due to technical complexities, our solution catalyses insight generation by providing a streamlined interface for data analysis.

Our web application capitalises on user-friendly design principles to offer an accessible platform for dissecting CSV datasets. By facilitating a seamless transition from data upload to insightful visualisations, the application is an exemplary tool in decision support systems. This initiative reflects the intersection of cutting-edge technology with practical, user-centric applications, underscoring the potential for advanced analytics in everyday business operations.

## Problem Statement and Context:


In the age of data abundance, the ability to swiftly analyse and draw conclusions from available information is crucial. However, technical barriers often preclude non-technical users from engaging with data analytics tools, resulting in underutilised data and missed opportunities for data-driven decision-making. The project was conceived against this backdrop, aiming to empower users by simplifying data analysis workflows and eliminating the dependency on specialised knowledge or software.

## Significance of the Project:

The project's significance transcends mere data analysis, offering significant value in educational settings for teaching data literacy and business environments where quick data appraisals are needed. By equipping users with tools to perform preliminary data analysis, the application promotes a data-informed culture, fosters curiosity, and enhances the overall analytical capabilities of its users.

## Achieved Outcomes:

The CSV Explorer was designed with four distinct tabs, each dedicated to a different data type: DataFrame, Numeric, Text, and Datetime. This modular approach ensures that users can focus on specific aspects of their data without being overwhelmed by unnecessary complexity. The outcome is a versatile tool that not only provides users with detailed analysis but also equips



them with the means to export their findings for further use or reporting. Our application should be an invaluable asset in educational workshops and small-to-medium enterprise (SME) data reviews and a potent tool for researchers requiring quick access to data insights.

Furthermore, by fostering an iterative development process, the team has laid a robust foundation for continuous improvement, positioning the CSV Explorer app at the forefront of accessible data analysis tools.

In summary, our project delivers on its promise of a scalable, user-friendly application that provides robust data analysis capabilities. The realisation of this project marks a significant step forward in making data analysis more approachable and actionable for a broader audience, setting a new benchmark for accessible, data-driven tools in the educational and professional spheres.



## 2. Introduction

### Project Objectives and Goals:

The primary objective of our project was to design and deploy a web-based application that simplifies data analysis tasks for users. The overarching goal was to create a platform that would serve as a bridge between complex data manipulations and the end-user, effectively enabling users to gain insights from their datasets without the prerequisite of technical expertise in data science or programming. This initiative aimed to address the gap in data analytics by providing a tool that is both comprehensive in functionality and intuitive in use.

### Stakeholders and Requirements:

The stakeholders of this project include a broad spectrum of users ranging from academic professionals and students to business analysts and data enthusiasts. Educators require tools that can be seamlessly integrated into curriculums to facilitate hands-on learning experiences in data literacy. Business analysts and SMEs must quickly interpret data for timely decision-making. Thus, the requirements are multifaceted, entailing ease of use, flexibility in data handling, robust analytics capabilities, and reliability.



## Addressing Stakeholder Requirements:

To meet these diverse requirements, the CSV Explorer was designed with a user-centric approach. The application's modular design allows for targeted data analysis, with separate tabs addressing DataFrame overview, numerical data analysis, text data insights, and date-time series evaluations. We developed these tabs to deliver specific analytical functionalities, from basic descriptive statistics to advanced data visualisations and trend analysis, catering to the nuanced needs of our stakeholders.

## Relevance and Context of the Project:

In the current data-driven landscape, the ability to rapidly convert raw data into actionable intelligence is a critical competency across all sectors. However, the prevalent data analysis tools often present a steep learning curve and are not tailored to the needs of users who lack specialised training in data analytics. This gap impedes the broader adoption of data-driven practices and stifles the potential benefits of informed decision-making. The CSV Explorer initiative, therefore, was strategically positioned to lower these barriers and facilitate a broader adoption of data analytics by providing a more accessible entry point to data analysis.

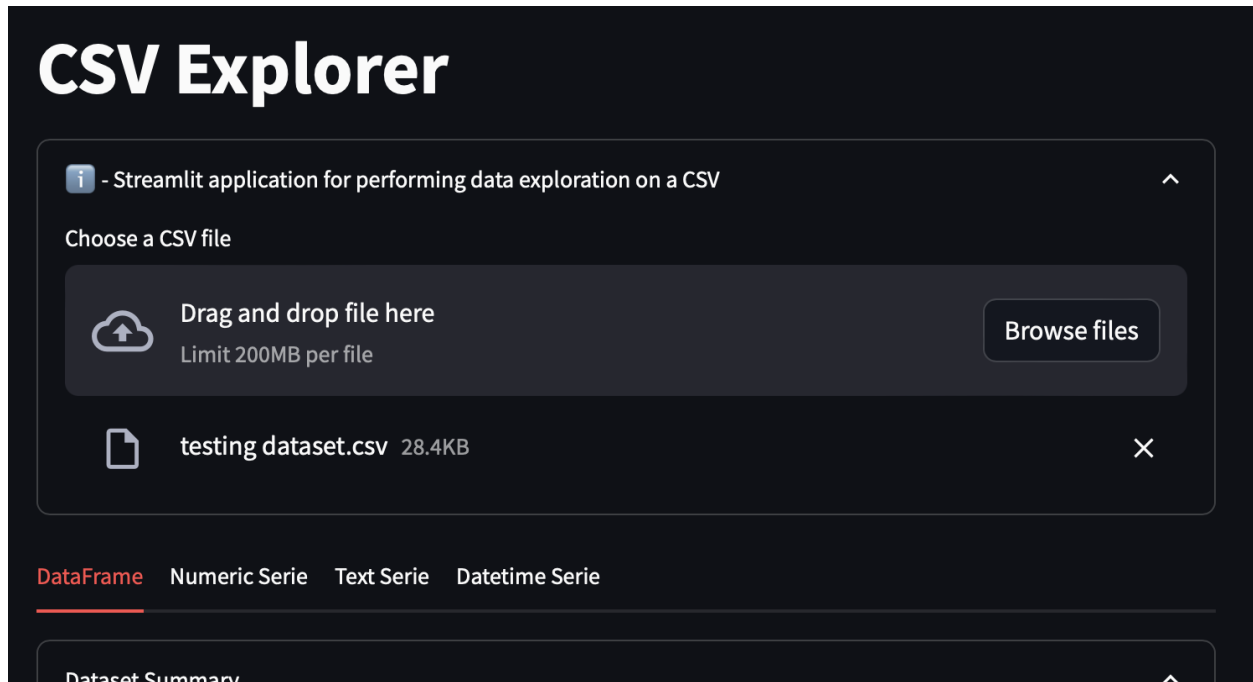
## Achieving Project Outcomes:

Through diligent project management and collaborative development efforts, the team successfully translated the requirements into a functional and practical product. We crafted the application's architecture to accommodate future expansions and integration of new features. The successful deployment and user adoption of the CSV Explorer within its first iteration signifies a commendable achievement in providing a solution that aligns with the vision of making data analytics a commonplace skill.

In essence, introducing the CSV Explorer app to the market will signify a paradigm shift in data analytics, empowering users to leverage the power of their data and fostering an inclusive environment where data literacy is not a privilege but a norm. The tool's reception and usage statistics reflect its efficacy and the value it adds to the stakeholders' data analysis endeavours.



### 3. Web App Presentation



App Screenshot

#### Purpose and Functionality:

Our development team at DSP AT3 Group 11 has crafted the CSV Explorer, a tailored web application engineered to streamline the data analysis workflow. The application is a testament to simplicity and efficiency, enabling users to delve into CSV datasets with unparalleled ease. With its robust suite of analysis tools, users can engage in comprehensive data examination through summarisations, intricate numeric and textual data analysis, and detailed date-time series review. This integration of multiple analysis features into one cohesive unit allows for a multifaceted approach to data interrogation within a singular, accessible environment.

#### Setup and Launch Instructions:

The CSV Explorer is designed for ease of use, allowing users to interact with their data using a local setup. To engage with the application, users must have Python and Streamlit installed on their local machine. Once the setup prerequisites are met, launching the application is as simple as navigating to the application's directory in the terminal and executing the 'streamlit run streamlit\_app.py' command. This action initiates the app, which then runs locally and is accessible through the default web browser, offering a private and secure way to analyse sensitive data without requiring an internet-hosted service.

## Potential Users and Use Cases:

The CSV Explorer caters to a diverse audience ranging from data analysts and business strategists to academic researchers and students. Its capacity to simplify complex data manipulation makes it an essential tool for anyone needing to make informed decisions based on data. For instance, business professionals might use it to quickly evaluate performance metrics, while educators can incorporate it into their curriculum to offer hands-on experience in data handling and visualisation. Journalists may also find the app beneficial for sifting through large datasets to pinpoint trends and stories.

## Commercialisation and Benefits:

The commercial viability of the CSV Explorer lies in its ability to democratise data analysis. By eliminating the need for advanced technical skills or costly software, it presents an attractive solution for businesses and educational institutions alike. It enhances productivity by saving time, reducing the likelihood of human error, and allowing users to concentrate on extracting valuable insights from their data.

## Limitations and Improvements:

Current limitations include the exclusive support for CSV file formats and the app's dependency on local machine resources, which may restrict its performance with voluminous datasets. Future iterations could introduce compatibility with additional file types and cloud-based processing to manage larger datasets more effectively. The integration of predictive analytics through machine learning could also extend its functionality, transitioning from a descriptive tool to one that can anticipate trends and outcomes.



## 4. Reflecting On Building Data Product

### a. Importance of developing data products

In the current era of vast petabyte data lakes, the role of data professionals has become increasingly crucial. Data Engineers and Scientists are the architects who build hyper scalable data pipelines to turn vast oceans of raw data into practical and actionable insights that organisations can leverage to build and develop use cases for cutting edge data products.

3 top benefits of developing data products are:

1. **Bringing Insights to Action:** Data Scientists with valuable and deep domain knowledge are able to generate valuable insights from data analysis which in turn allows these insights to be converted into tangible solutions, tools, or services that can benefit society as a whole.
2. **Enhancing User Experience:** Data products often cater to end-users, can often offer an one stop shop of products suites which allow for a centralised and unique experience
3. **Innovation and Competitive Edge:** Developing data products often involves developing new and groundbreaking new technologies and methodologies, which can spark a new technological era, one such example would be the development of transformers, which led to the development of Large Language Models.

## b. Key skills and technologies required for developing data products


There are many skills and technologies required for developing data products, in this document I will focus on what I believe are the top 5.

1. **Data Analysis, Modeling and Domain Knowledge:** : Strong skills in data analysis, statistics, and machine learning algorithms and the ability to interpret analysis coupled with strong domain knowledge are often crucial to the success of any data product.
2. **Programming Languages:** Strong ability in languages such as Python, R and SQL is fundamental for data manipulation, analysis, and building models.
3. **Data Engineering:** Knowledge of data pipelines, data architecture, code versioning, and ETL processes are extremely useful for creating the backbone of any data products and often rely on tools such as Spark, Kafka, Airflow and Github.
4. **Data Visualization/ Storytelling/Narrative:** Ability to create effective visualisations and narratives to communicate insights are often critical to convey the success of a product, these often rely on Tableau or Power BI.
5. **Cloud Platforms:** Familiarity with cloud platforms such as AWS, Azure, or Google Cloud is increasingly important due to their extensive data storage and computing capabilities.

## c. Data Product Use Cases

There are many potential use cases for data scientists to develop innovative data products, some of the common and popular use cases are:



- 
1. **Predictive Models / Machine Learning Models:** These models often will use algorithms such as Regression, Boosting, Neural Networks or ARIMA to forecast/predict future trends and/or outcomes based on historical data. Examples include sales forecasts, predicting customer churn in businesses.
  2. **Recommendation Systems:** Data scientists can create recommendation engines that suggest products, content, or services based on user behaviour and preferences, commonly seen in social media and streaming platforms, such as Netflix.
  3. **Image and Video Analysis:** Products for image recognition, object detection, video analytics, or facial recognition systems, often used in security, retail, healthcare, and entertainment.
  4. **Natural Language Processing (NLP) and Large Language Model (LLM) Solutions:** Products such as chatbots, sentiment analysis tools, language translation services, and text summarization are some of the current and potential future use cases.

#### d. Reflection on the current trend of AI advancement

As society enters a golden era of AI advancement, numerous future use cases have been evaluated by Gartner's in their Hype Cycle for Artificial Intelligence. According to Gartner's, the landscape of cutting-edge innovation triggers in the next 3-5 years is on the cusp of several significant developments.

Firstly, is the rise of AI Simulation, utilising Artificial intelligence to simulate 'human-like' intelligence processes by machines, especially computer systems.

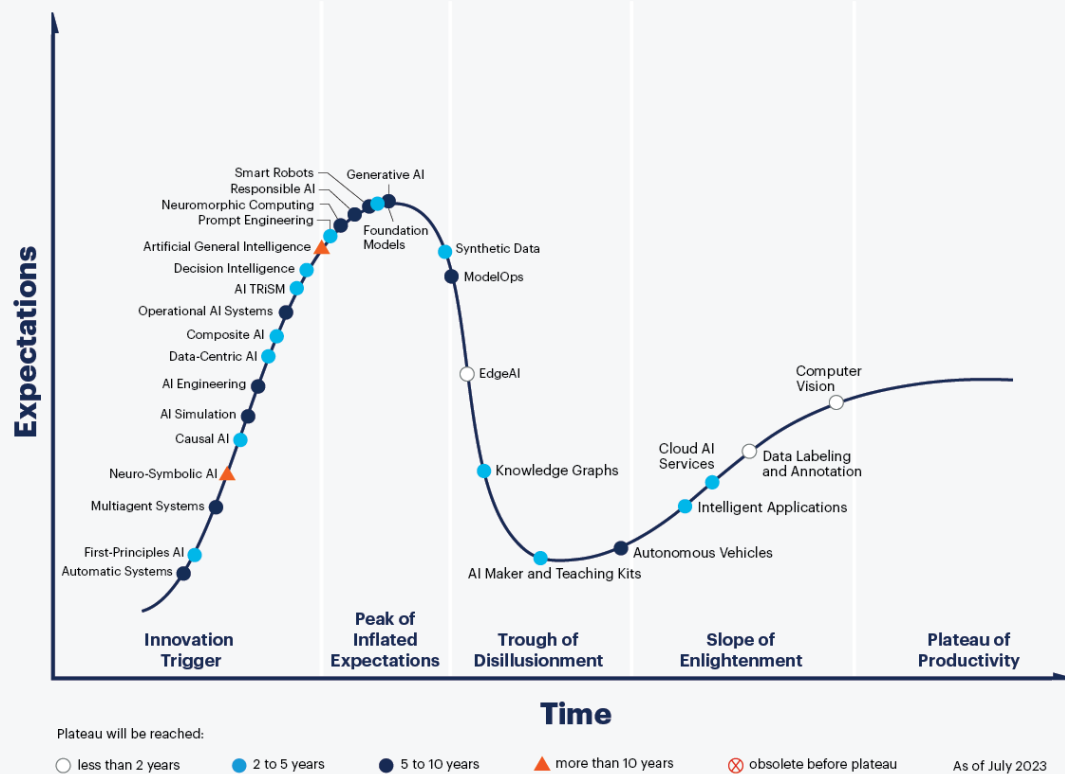
Secondly, Casual AI, which seeks to identify the degree of relatedness of variables in a system and identifies the root cause by understanding any variables that may have led to it.

In parallel, AI Engineering is set to enhance the reliability and scalability of AI systems, focusing on robustness, efficient deployment and can help organisations increase efficiency, cut costs, increase profits, and make better business decisions.

Moreover, the pursuit of Artificial General Intelligence (AGI) remains as the holy grail and also a significant goal in AI research, aspiring to create intelligent systems that can perform a variety of tasks akin to 'human-like' capabilities.

On the other hand, Generative AI, while having reached high expectations, is anticipated to plateau in its innovation trajectory, likely facing a phase of reduced advancements in the coming years.

# Hype Cycle for Artificial Intelligence, 2023



[gartner.com](https://www.gartner.com)

Source: Gartner  
© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. 2079794

**Gartner**

Gartner. (2023). Hype Cycle for Artificial Intelligence 2023.

## 5. Collaboration

### a. Individual Contributions

- **Vishal Raj (Student A)** played a crucial role in developing the `DataFrame` tab, facilitating data upload, and creating a structured GitHub repository for our Streamlit app—his work on the CSV upload feature allowed for easy data entry, improving our app’s functionality. Vishal not only developed but also provided a comprehensive testing dataset, which was

integral for validating the application's various features, along with updating the Readme file.

In managing the GitHub repository, Vishal ensured that updates were systematically implemented and kept the main branch stable by carefully merging tested features. This attention to detail extended to his report writing, where he crafted the executive summary and introduction, outlining the project's scope and objectives.

Vishal's balanced approach to development and oversight helped maintain our project's momentum, leading to timely and cohesive final submissions. His efforts reflect the synergy between technical skills and project management needed for creating impactful data products.

- **Rohit Sharma (Student B)** played an instrumental role in the development of the numeric data analysis functionality within our Streamlit application, demonstrating a strong proficiency in data handling and visualisation. Rohit was responsible for the implementation and optimization of the `NumericColumn` class in `logics.py`, which serves as the backbone for numerical data processing in the application.

His work included developing robust methods to identify numerical columns, handle missing and negative values, and compute key statistics such as mean, median, and standard deviation. Additionally, Rohit's implementation of data visualisation using Altair charts for histogram representation significantly enhanced the user's ability to comprehend data distributions.

Rohit's meticulous approach to coding also extended to error handling and edge case considerations, ensuring that the application operates reliably with various datasets. His proactive testing and debugging led to high-quality code that consistently performs as intended.

- **Ronik Jayakumar (Student C)** developed the `text` tab which is tasked with reading textual data available in CSV files. Once the CSV file is loaded onto the webapp, the textual data is read and filtered through to make available for the viewer to select. Once a text column has been selected, the application gives the user an initial Exploratory Data Analysis. The features of the text tab are as follows:
  - An initial Exploratory Data Analysis with rows representing data features like - unique value count, missing values, alphabetical values etc.
  - A bar chart showing a visual representation of the frequency of values

- A frequency chart depicting the top 20 frequently occurring values and the percentage of their frequency.

The overall code required multiple iterations and testing phases coordinated with team members to ensure accurate and efficient working. Future additions to this could include text preprocessing features such as punctuations removal, and natural language processing.

- **William Dai (Student D)** developed the `datetime` tab which is tasked with detecting and converting text to datetime time series data, some of the key features of that tab include:
  - An initial Exploratory Data Analysis on key features of the datetime data, such as weekends, weekdays, future dates, etc.
  - A time series histogram visualisation developed using Altair.
  - A streamlit dataframe table listing the occurrences and percentage of the top 20 most frequent values
- The overall code required multiple iterations and testing phases and strong coordination with team members to ensure successful integration of code, one such initiative we took early on was to ensure our development environments are aligned, this was done through the use of virtual environments and requirements.txt

## b. Group Dynamic

A comprehensive working dynamic was agreed upon in the first week of the project work.

**Communication** - To ensure a simple yet efficient means of communication, Whatsapp was chosen as the primary means of discussion. A group was made with all team members and responsibilities were assigned. Timelines were decided through team discussions and periodic progress checks were performed.

**Collaboration** - A github repository was set up for the team to push their code into. This enabled progress tracking, and efficient collaboration. A separate file was created for each student/tab with a temporary branch for testing and a main branch where the final code was pushed once all checks were passed.

**Documentation and Knowledge sharing** - Comprehensive documentation of requirements were made available on Github to ensure easy reference for team members. Knowledge sharing was also encouraged for team members to voice out their findings and insights.

**Reflect and Support** - The team was open and respectful to each others backgrounds and working styles. Each others achievements were celebrated and pain points were discussed.

**Incremental delivery** - An agile methodology was used to in feature delivery. We worked in incremental sprints taking one step at a time enabling us to quickly respond to change and incorporate team member feedback. This allowed for flexibility and continuous improvement.

### c. Ways of Working Together

- **Methodologies, Frameworks and Progress Tracking:**
  - An agile methodology was used to allow for incremental project building, easy adoption to change, and feedback incorporation.
  - Daily stand-ups were help via quick whatsapp chats to discuss team progress, issues faced, and new findings.
  - Weekly reviews and retrospectives were conducted where team members portrayed their work, received feedback, and discussed what went well and what can be improved.
- **Decision Making, Tools & Technologies:**
  - Informed discussions were held to ensure all inputs are taken to enable ease in prioritisation and feature handling.
  - Github was the core of our webapp building with all collaborative coding and testing being done using the website.
  - Google docs was used for collaborative report making where all team members had their assigned responsibilities.
  - Python Virtual environments and single source of truth requirements.txt to ensure our development environments are in sync.

### d. Issues Faced

- **Challenges:**
  - Scheduling conflicts - With varied work schedules, University assignments, and personal agendas, meeting scheduling was one of the biggest challenges faced.
  - Technical discrepancies - The team came with varied expertise in using the tools employed. Some technical difficulties were faced in the usage of some employed tools.
  - Communication gap - Due to schedule conflicts, team members would at times miss important communication.
- **Solution:**

- Scheduling conflicts - This was resolved by having frequent conversations and ensuring all discussed points were written down to ensure missing team members could revisit.
- Technical discrepancies - Open lines of communication enabled team members to voice out their difficulties. This enabled the others to voice out solutions.
- Communication gap - The communication means were used strictly for project related discussions to ensure no important messages are missed by missing team members.
- Lessons Learned:
  - Flexibility is key - The key to ensure inclusivity in a diversely scheduled environment is flexibility which would allow all team members to be a part.
  - Continuous skill building - Varied skill expertise helps for continuous skill building within the team.
  - Documentation - Maintaining clear, concise, and comprehensive documentation could help deal with conflict schedules and doubts clearing for future reference..



## 6. Conclusion

The overall development of the webapp resulted in a versatile tool that can be used for initial Exploratory analysis. The application's architecture takes into account the important requirement for agility in business and research settings, giving quick and dependable access to data summaries, numeric and textual analysis, and date-time evaluations. It was created in response to a perceived need in tools that provide a comprehensive yet user-friendly way for non-technical individuals to explore data.

The overall project was successful in materialising the stakeholder requirement to simplify data exploration. It encompasses the 1st step of most data projects by allowing technical and non technical users to get a quick and easy insight.

The final project involved the following:

- Dataframe tab providing an overview of the table.
- Numeric tab providing information on all integer and float data.
- Text tab providing information on all textual data in the file.
- Datetime tab giving us all date and time related information existing in the file.

Future recommendations for the app would be the following:

- Addition of data cleaning techniques that perform both exploration and chosen data preprocessing. This could include removal of punctuations, dropping null values, basic mathematical operations and so on.
- Additional effort into the User Interface to increase ease of use and aesthetics.
- Custom visualisations which allows the user to select chart types based on their specific needs. This could involve word clouds, network graphs, etc.
- Machine learning integration to enable the app to train and test data.
- Compatibility with cloud based platforms like AWS, Google Cloud, and Azzure.
- Direct integration with various data sources for direct access.

In conclusion, the project has delivered a comprehensive application which can be used as a crucial analytical tool to gain valuable insights from data.



## 7. References

McKinney, W., Pandas Development Team. (2022). Pandas: Powerful Python data analysis toolkit. Retrieved from <https://pandas.pydata.org/pandas-docs/version/1.4.4/pandas.pdf>

Python Software Foundation. (2023). Python Language Reference. Retrieved from <https://www.python.org/>

Harris, C.R., Millman, K.J., van der Walt, S.J., et al. (2020). Array programming with NumPy. Nature, 585(7825), 357-362. <https://doi.org/10.1038/s41586-020-2649-2>

McKinney, W. (2017). Python for Data Analysis: Data wrangling with pandas, NumPy, and Ipython. O'Reilly Media, Inc.

Few, S. (2012). Show me the numbers: Designing tables and graphs to enlighten. Analytics Press.

Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis. Queue, 10(2), 30-55. <https://doi.acm.org/10.1145/2133416.2146416>

Wickham, H. (2014). Tidy data. The Journal of Statistical Software, 59(10). <https://www.jstatsoft.org/v59/i10>

Riche, N. H., Lee, B., Plaisant, C., & Fekete, J. D. (2010). Understanding interactive legends: a comparative evaluation with standard widgets. Computer Graphics Forum, 29(3), 1193-1202. <https://doi.org/10.1111/j.1467-8659.2009.01678.x>

Fry. (2008). Visualizing data. O' Reilly Media.

Streamlit. (2023). Streamlit — The fastest way to build custom ML tools. Retrieved from <https://www.streamlit.io/>

Altair Visualization Team. (2018). Altair: Declarative Visualisation in Python. Retrieved from <https://altair-viz.github.io/>

Gartner. (2023). Hype Cycle for Artificial Intelligence 2023. [Online image]. Available at: [https://emt.gartnerweb.com/ngw/globalassets/en/articles/images/hype-cycle-for-artificial-intelligence-2023.png?\\_gl=1\\*1fn2bfq\\*\\_ga\\*MjA5Nzg0NTY2Mi4xNjk5NTE0MjAw\\*\\_ga\\_R1W5CE5FEV\\*MTY5OTUxNDIwMC4xLjAuMTY5OTUxNDIwMS4lOS4wLjA](https://emt.gartnerweb.com/ngw/globalassets/en/articles/images/hype-cycle-for-artificial-intelligence-2023.png?_gl=1*1fn2bfq*_ga*MjA5Nzg0NTY2Mi4xNjk5NTE0MjAw*_ga_R1W5CE5FEV*MTY5OTUxNDIwMC4xLjAuMTY5OTUxNDIwMS4lOS4wLjA). Accessed on 09/11/2023

■ ■ ■