Assignment 3
Collaborative Development of Data Explorer Web App

Date: 10/11/2023

Students of Group 14:

Tarun Krishnan (25697928)

Phillip White-Wall (10826612)

Muhsin Dawood (24897124)

Able Varghese (24712198)

https://github.com/the-tunnster/dsp-at3-group14

94692 - Data Science Practice

Master of Data Science and Innovation

University of Technology Sydney

Table of Content

Executive Summary and Introduction

The project aims to create a development environment and replicate the experience of working as a team. The use of continuous integration via version control systems allows the team to experiment with progressing changes and rolling back issues, as well as merging and pushing changes upstream and dealing with conflicts in such situations.

The primary goal of this project is to develop a web application that will be based on the Streamlit platform, along with the use of python libraries to populate various tables of information and data. The use of sessions and session states, along with python classes and object definitions is another aspect of the project.

The outcome of this project is a multifaceted web application that allows users to upload a CSV file, and obtain information about the different aspects of the data frame thus obtained. The project contains four tabs, each of which deal with the various data types and the columns associated with them.

Web App Presentation

Purpose and Functionalities

The web application serves as a tool for exploring different data type columns in datasets. Users can select columns they wish to explore further in depth, convert them to the specified format, and visualize various statistical analyses conducted. The application provides an intuitive interface, allowing users to seamlessly navigate through the features. Additionally, it offers export options for the generated visualizations, enhancing the usability for further analysis. The application is split into 4 tabs designed to explore different aspects and datatypes; Dataframe, Numeric, Text and Datetime. The specific content provided in each table can be seen below:

Figure 1: Functionalities of the features available in the application

Tabs	**Function**
Dataframe	Summarized information of the dataset, exploring the descriptive statistics allowing for modification on the number of rows and columns wished to be explored
Numeric	Explores the numerical columns, outlining the descriptive statistics, further allowing users to visualize columns in a histogram.
Text	Explores the text-based columns, outlining the key stats for missing characters, further allowing users to visualize columns in a bar chart.
Datetime	Detects the datetime specific columns, allowing to explore for ranges and visualizing in a bar chart for a time-series analysis.

Instructions to launch the Application

The application is contained within a folder with our group name. Within the folder, the backend logic and front-end display code can be seen in the four different tables, including an app.py script to run the application. Supporting the use of the application is the README.md file that meticulously explains how to use the application.

The application set up:

1. ensure you have python installed and running. You also need to install a package manager, and we have used pip3 for this project. The version for the python runtime is Python 3.9.18
2. Using the command lines found in the README.md file, you will create and activate a new virtual environment
3. We must then install the required external packages for the project to work, namely Streamlit, Pandas, Altair.

Launching the application:

The fastest way to run this project is to run the following command lines found in the README.md file. It is important to note that the application will run as per the file directory, therefore you must include the correct file directory after the “run” command. For example, in our case it would be as follows:

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

Potential Market

The python language remains the highest in-demand for the data industry, with data science professional's primarily using the several packages to produce descriptive analytics for the purpose of business intelligence.

Our EDA application demonstrates products that are similar to existing competition such as Y-data Profiling, an open-source tool leveraging Pandas Dataframes, is a tool that can simplify and accelerate such tasks, performing an in-depth EDA analysis on the specified dataset.

Drawing from the user feedback of Y-data profiling and similarity in function of ou product, we can assume that the potential user market of our product will exhibit a similar market share as follow:

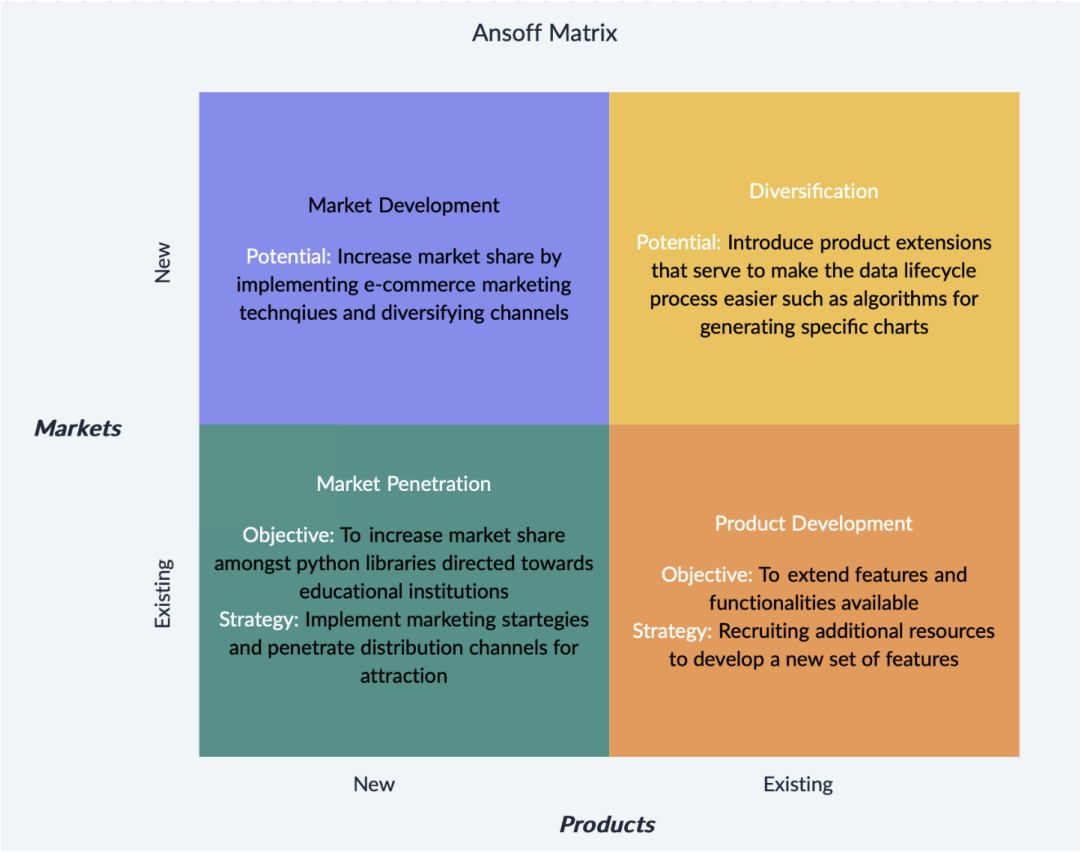
1. **Data Scientists and Analysts:** These users will use the application for the purpose of insight exploration for specific data types. Moreover, it will be utilized for identifying data noise by illustration of missing values and data cleaning potential.
2. **Data Engineers:** These user ideally use the application of validation purposes to check the quality and integrity of datasets before loading them into databases or data warehouses
3. **Students of Data Science and Computer Science:** Provides students with educational specific reporting with appealing visuals.
4. **PHD Research Students:** Serves research students as a quick and efficient tool for exploratory data analysis in research projects.
5. **Consultants in Data services:** Often dealing with clients, the tool will aim to provide clients with a comprehensive overview of their datasets, helping them make informed decisions and integrate solutions of insights gained.

Commercialization

Exploratory data analysis is a critical component of the data lifecycle. This data specific schema informs the data users with a potential approach to producing predictive modeling and capabilities that is both differentiated and valued. Moreover, the process of EDA is the catalyst of the reaction of decisions to be made such as feature selection and engineering.

While currently an internal tool, the application's potential commercialization lies in offering it as part of data analysis suites or services. The application is particularly useful in commercialization with educational institutions as the primary market is students, educational professionals and researchers. The commercial capabilities of the product can be visualized in the matrix below:

Figure 2: Ansoff Matrix for Commercialization



The Ansoff Matrix is a strategic planning tool that helps businesses decide their product and market growth strategy. It consists of four growth strategies: Market Penetration, Market Development, Product Development, and Diversification. The Existing section outlines the potential objectives and strategies to deploy for commercializing the tool. While the New changes are far along the product life cycle, demonstrating its potential expansion and use cases.

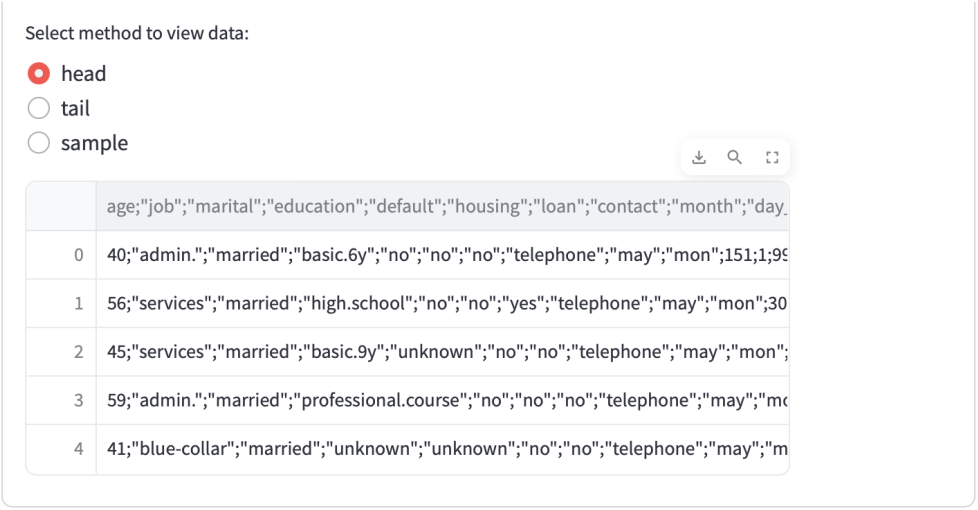
The commercialization of our product aims to target educational institutions with a shared ownership structure with the ministry of education. By implementing a subscriber based structure, the product is able to generate revenue stored as capital for investment and attract users through targeting marketing campaigns. The brief analysis showcases that the demand of the product amongst data users is quite vast which contributes to the success and potential commercialization of it.

Limitations and Improvements

The web application is a resounding success considering the initial aim of the project. While it does work, there are without doubt certain limitations of the applications. We will explore those limitations and demonstrate the future improvement required to develop for deployment.

1. The application is dependent on the quality of the dataset provided. For example, if the CSV file is not delimited correctly, it will showcase errors as seen from the below image:
2. The application is limited to 200MB for the CSV files, hence any EDA needed to be performed for big data analysis purposes will not be supported
3. The EDA feature list is limited to the 4 tabs, but moreover there is no range of options in viewing different types of charts to fulfill different purposes. This was thoroughly expressed in the Ansoff matrix as potential product development.

Figure 3: CSV file delimiter issue



Future improvements could include refining the detection algorithm for better accuracy. Additionally, incorporating user feedback mechanisms within the application can contribute to ongoing enhancements. Moreover, extending the option top perform dedicated tasks such as CSV delimiting, data cleaning and potential processing could improve the user feedback and build brand loyalty. Finally, an aspect we continuously discussed is integrating it into popular data science platforms or providing it as a standalone tool. The user-friendly design and interactive features position it well for broader adoption.

Reflecting on Building Data Product

Importance of Data Products

Data scientists spend a disproportionate amount of time on data wrangling and janitorial work. These are necessary but often repetitive tasks in the project lifecycle. The purpose of the Data Explorer Web App addresses the initial data exploration phase. Other products could be developed by data scientists to address other phases. When preparing data for model creation, products to address missing/null values, outliers, encoding, etc. would give way to time spent on model creation/evaluation instead. It's worthwhile building reusable data products to free up time for these more value-add activities.

Key Skills and Technologies

It is crucial for data scientists to possess a diverse range of skills, enabling them to effectively translate data insights into action. Soft skills around data visualization, user interface design, and domain-specific knowledge contribute to successful adoption of your data insights. Also adopting the appropriate enterprise tech stack makes for better integration within the business. Developing data products stakeholders can understand and engage with builds trust and encourages collaboration.

Trends in AI Advancement

AI assisted insight generation is already a reality for a lot of business enterprise systems. These features interpret sets of data often employing large language models to interpret user's questions. A level of skepticism exists for the outputs with businesses often relying on data analysts and scientists to verify.

Innovative products that leverage LLMs and AI are becoming more accessible for data scientists to develop more bespoke solutions. AI that are trained on company specific data will provide richer insights. These targeted products will experience better success than the generic models available today.

Collaboration

1. Individual Contributions

Tarun Krishnan :

- Created the github repository and added the team members.
- Created the individual branches and defined the version control strategy.
- Approved various pull requests and fixed issues by reverting commits.
- Created the README.md file and populated the google drive folder for the team to use.
- Developing the backend logic for the Dataframe Tab

Muhsin Dawood :

- Backend development of the Numerical Tab, implementing the key functions for the front-end display.
- Consistently testing and debugging for the seamless integration of the “Numeric Serie”
- Responsible for complete testing and relay of potential errors to be amended and changes to be implemented
- Contribution to the final draft, enhancement and proof-reading the report

Phillip White-Wall :

- Development of the Text Tab: This involved implementing the backend logic and presentation of text columns, as well as testing and debugging for integrating this feature.
- Contribution to report writing: authoring and contributing to overall voice of report.

Abel Varghese :

- Implemented the backend logic for datetime column analysis in the Fourth tab. This involved parsing and interpreting datetime information from various data types.
- Conducted rigorous testing and debugging to ensure the reliability of the datetime analysis features. Collaborated with frontend developers to address integration issues and ensure seamless functionality.
- Contributed to the documentation of the Fourth tab's functionality, ensuring that users and future developers can understand the intricacies of datetime analysis and leverage the features effectively.

2. Group Dynamic

The group faced issues with respect to managing the version control strategy. However, these were overcome with a bit of communication and resource management. As a team each member decided on roles and responsibilities, and completed the tasks defined.

The group maintained effective communication through regular meetings and utilized collaborative tools for efficient project management. Each team member had clearly defined roles, fostering a positive group dynamic. Continuous communication channels, including Microsoft teams, facilitated prompt responses to queries and challenges.

3. Ways of Working Together

Primary communication was performed on a WhatsApp group chat, along with emails or pull requests and updates. Git was the version control system used, and GitHub was the repository service provider. Finally, Branches were created and managed to provide parallel collaboration among the team members, and frequent commits and updates provided a system of continuous integration.

Agile methodologies were employed, with weekly sprint meetings to track progress. Tools like Microsoft Teams facilitated communication and project management. Regular code reviews ensured the quality of the codebase. The group embraced an iterative development approach, allowing for flexibility in responding to emerging challenges.

4. Issues Faced

The issues faced include a combination of technical and collaborative issues. The summary of the issues faced can bee seen below, also outlining the approach to resolving the issues.

- There was an issue with pushing an incorrect branch update, but was dealt with using a git revert command.
- The Pycache files were being generated on every run and to avoid this issue, a gitignore file was added.
- Issues with communication and time management between members was an initial hurdle, but was resolved relatively quickly by setting up dedicated meetings and milestones to achieve the project within the deadline.
- Version control strategies were established, and there were a few issues with ensuring branch safety, which was enforced later.

Conclusion

The project aimed to develop based on the Streamlit platform, along with the use of python libraries to populate various tables of information and data. The testing of the web application on the team members previous assignments showcases that the project successfully delivered a valuable tool for exploratory data analysis. The testing showcased that while the web application provides crucial EDA analysis to perform further action such as data cleaning, modeling and feature engineering, it does present shortfalls in bypassing certain technical issues exhibited in the files imputed as mentioned in our limitation analysis.

Acknowledging the team's collaborative efforts, the web application reflects the significance of data products in modern data science. The tool not only addresses current data analysis needs but also sets the stage for future developments. The market study conducted showcases that the requirements of this product within the data industry is vast and distributed throughout different data roles, from students to paid professionals. As EDA is a critical component of the data lifecycle, this came to no surprise of the team members.

Future work involves refining the data type detection algorithm, incorporating advanced features, and exploring additional use cases based on user feedback. The application has built a solid foundation in providing users with the ability to perform the necessary EDA analysis for small projects. As seen from our potential commercialisation analysis, the room for improvement and deploying the application is possible as the market study suggests the adoption of a highly insightful EDA application could compete with existing and limited libraries such as Y-data profiling. The implied future steps in market and product development will not only allow to fit a gap but also penetrate a market and initiate growth in the already boosting data industry, improving and progressing the data lifecycle.

References

- Goyette, C. (2023, August 24). *How to supercharge data exploration with pandas profiling*. Domino Data Lab. https://domino.ai/blog/how-to-supercharge-data-exploration-with-pandas-profiling
- Vangala, S. (2023, September 24). *Pandas profiling: Uncovering insights from data with ease*. Medium. https://medium.com/@sankalpithav/pandas-profiling-uncovering-insights-from-data-with-ease-7b4224a5ef8
- *Pandas-profiling*. PyPI. (n.d.). https://pypi.org/project/pandas-profiling/
- Leppitsch, M. (2023, October 20). *Benefits of the data product approach*. Ascend.io. https://www.ascend.io/blog/benefits-of-the-data-product-approach/
- Marr, B. (2023, October 5). *The 10 most important AI trends for 2024 everyone must be ready for now*. Forbes. https://www.forbes.com/sites/bernardmarr/2023/09/18/the-10-most-important-ai-trends-for-2024-everyone-must-be-ready-for-now/?sh=7aa2380736bd
- *Ai trends outlook*. Deloitte United States. (n.d.). https://www2.deloitte.com/us/en/pages/consulting/articles/ai-trends.html