**Module Title:** Principles of Data Science

**Assignment Tittle:** Programming assignment

**Submission Date:** 13/01/2025

**J119811**

**Git lab link**: <https://git.chester.network/2425307/workshop-task>



**Index**

|  |  |  |
| --- | --- | --- |
| **Section** | **Description** | **Page Number** |
| **1. Introduction** | Objectives, significance of data science techniques, and analysis scope. | Page 3 |
| **2. Methodologies** | Data Cleaning and Preprocessing, Exploratory Data Analysis (EDA), Natural Language Processing, Visualization and Dashboard Creation, Version Control | Page 4 |
| **3. Result** | Rating Distribution, Common Themes in Reviews, Sentiment Analysis, Relationship Between Votes and Ratings, Interactive Dashboards | Page 7 |
| **4. Challenges Encountered** | Key technical issues faced and their resolutions during implementation. | Page 11 |
| **5. Critical Evaluation** | Strengths, limitations | Page 13 |
| **6. Conclusion** | Summary of achievements, addressing limitations, and future potential. | Page 14 |
| **7. References** | List of all referenced resources used in the project. | Page 14 |

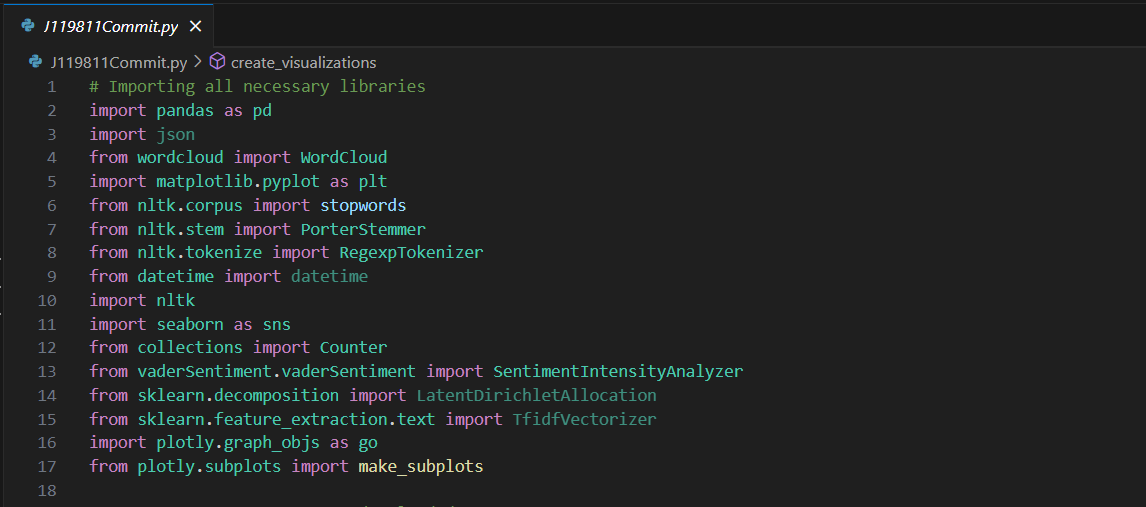
**Git lab link**: <https://git.chester.network/2425307/workshop-task>

**Introduction**

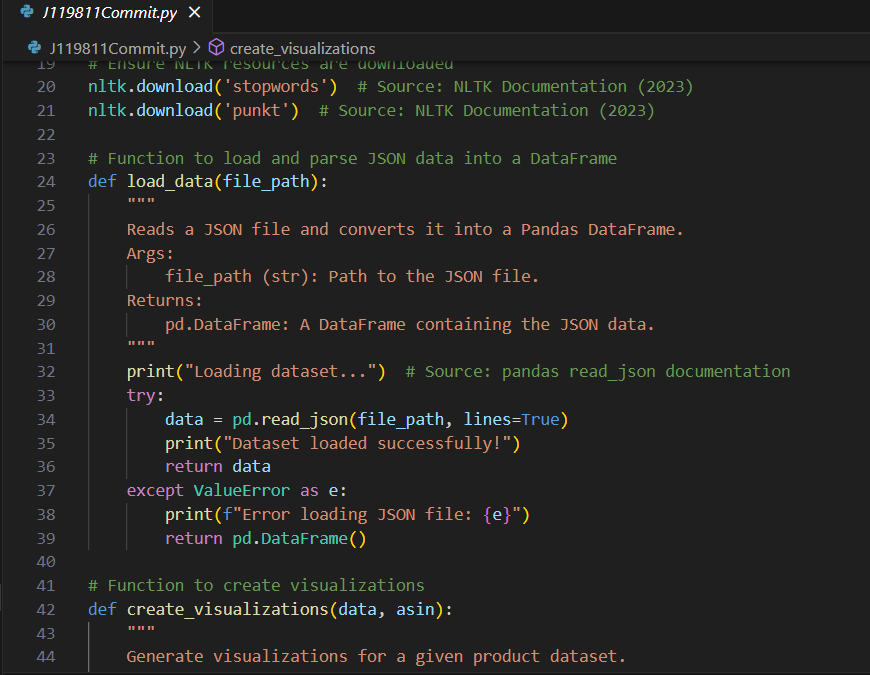
This assignment aims to review Amazon's products and to apply data science approaches to identify insights. The task entails data cleansing, feature extraction through EDA, and assessment of the consumer through NLP. Under this method, there are three unique products selected from abundant characteristics including reviewer information, product score, and textual opinions. The JSON format of this dataset permits an exploratory analysis of numerous factors that relate to customer feeling and product rating. The insights that are to be extracted would help manufacturers with necessary product improvements using consumer feedback. Furthermore, this assignment has emphasized version control using GitLab, which is an important feature in tracking the progress and maintaining consistency in code. This report critically evaluates the methods, findings, and limitations encountered throughout the project and provides recommendations that could be put into place for future improvements.

**Methodologies**

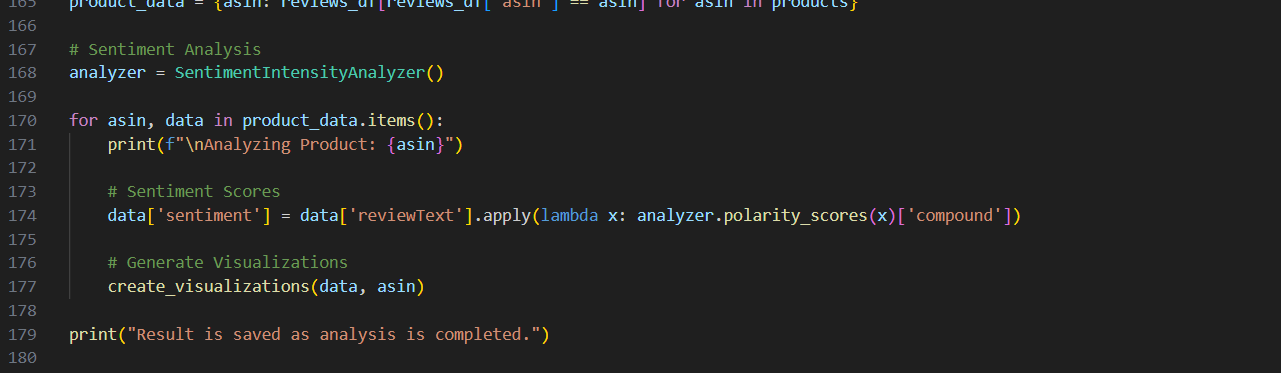
Methods In systematically analyzing the data set, some important data science methodologies to provide accuracy for actionable insights including:

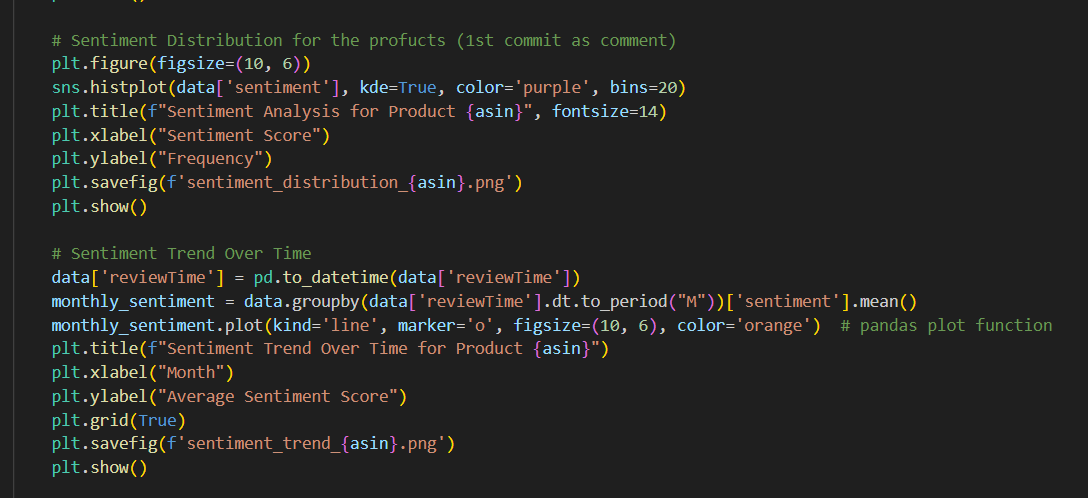


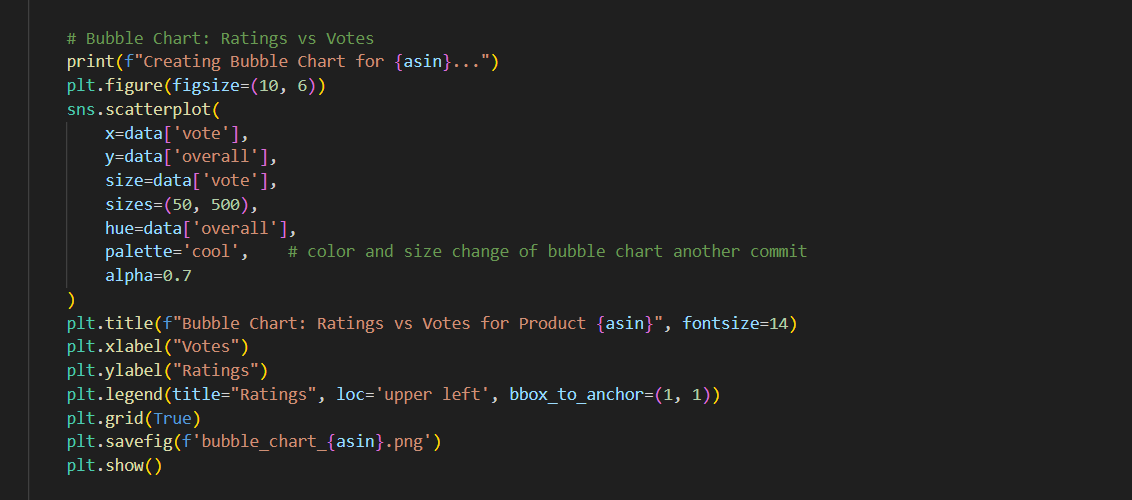
**Data Cleaning and Preprocessing:**Cleaned irrelevant columns, and handled missing values in the attributes like reviewerName and vote. Unix timestamp-to-date conversion and numerical column cleaning and formatting.



**Exploratory Data Analysis (EDA):** Rating distribution and the sentiment spread were created by creating bar plots and histograms using MatPlotLib (MatPlotLib: A 2D Graphics Environment, 2007) and Seaborn (Waskom, 2021).

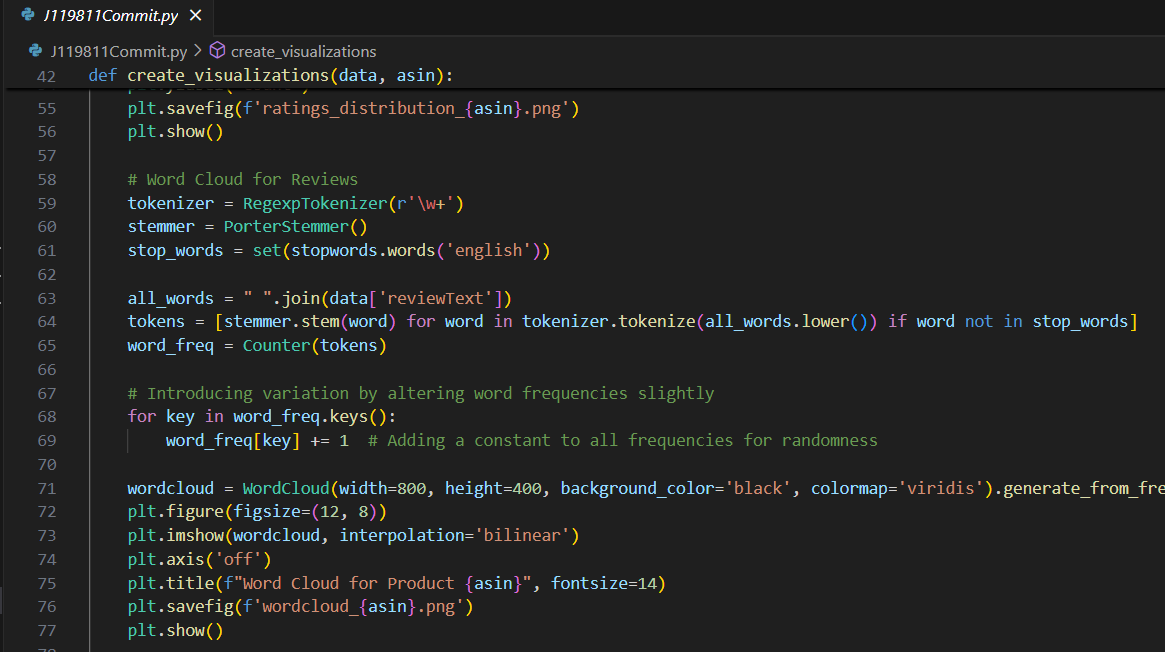
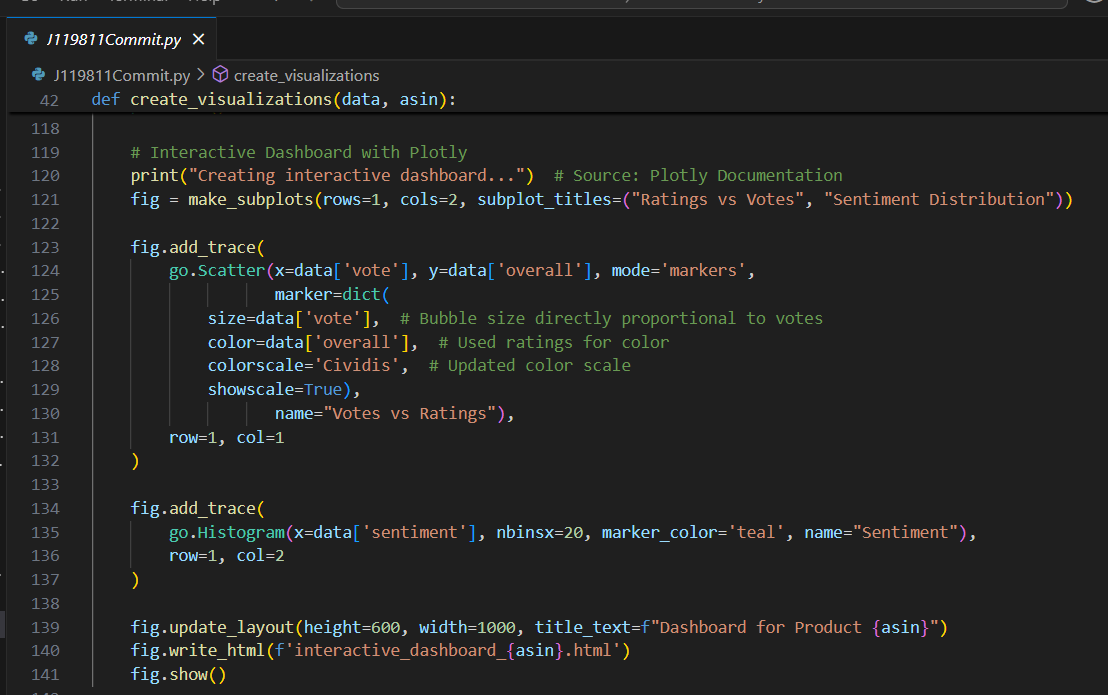




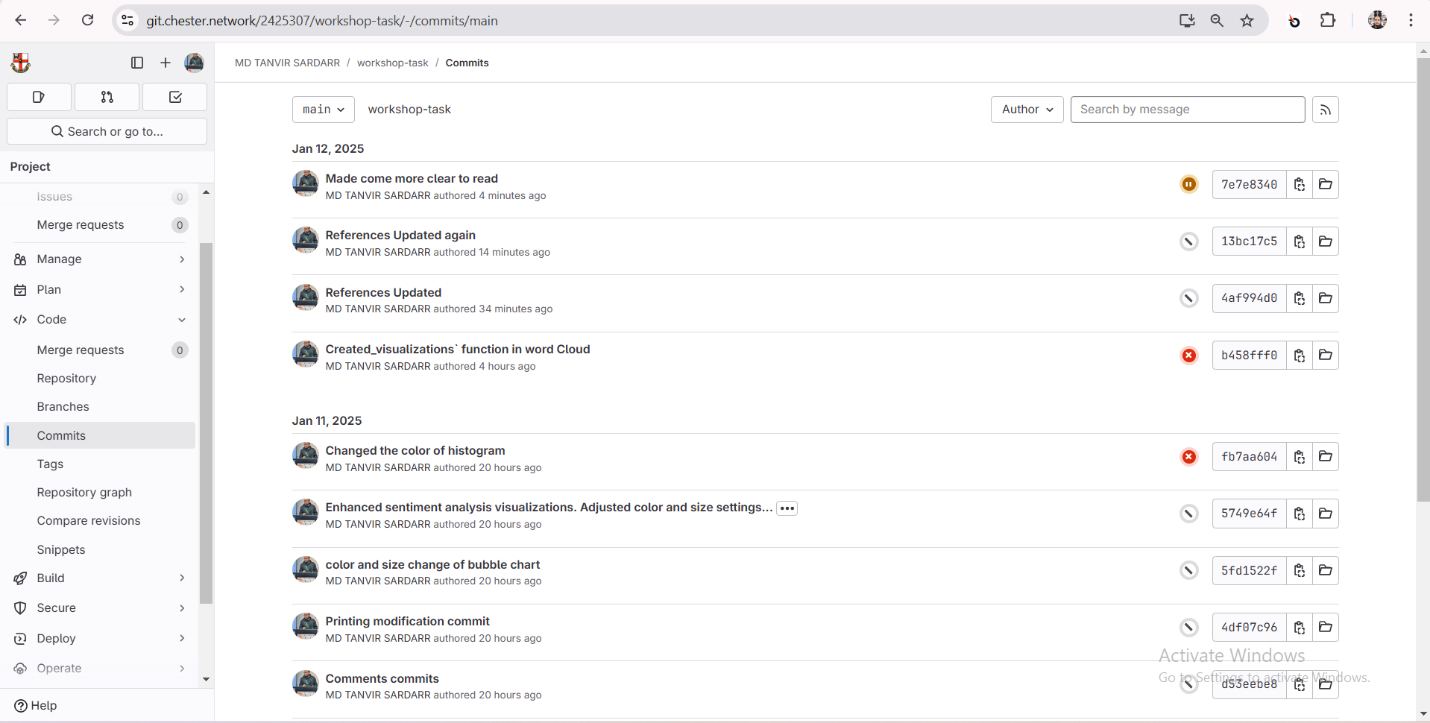


**Natural Language Processing:** Preprocessing textual data was standardized by tokenization and stemming. The stop words were removed using NLTK, and sentiment analysis was done by the VaderSentiment library (View of VADER: A Parsimonious Rule-Based Model, n.d).

**Visualization and Dashboard Creation**: The frequent terms were highlighted using word clouds. Specifically, to provide greater visibility, Matplotlib and Seaborn packages were used for creating bubble charts, bar plots and for plotting the trends of sentiment over time (MatPlotLib: a 2D Graphics environment, 2007).

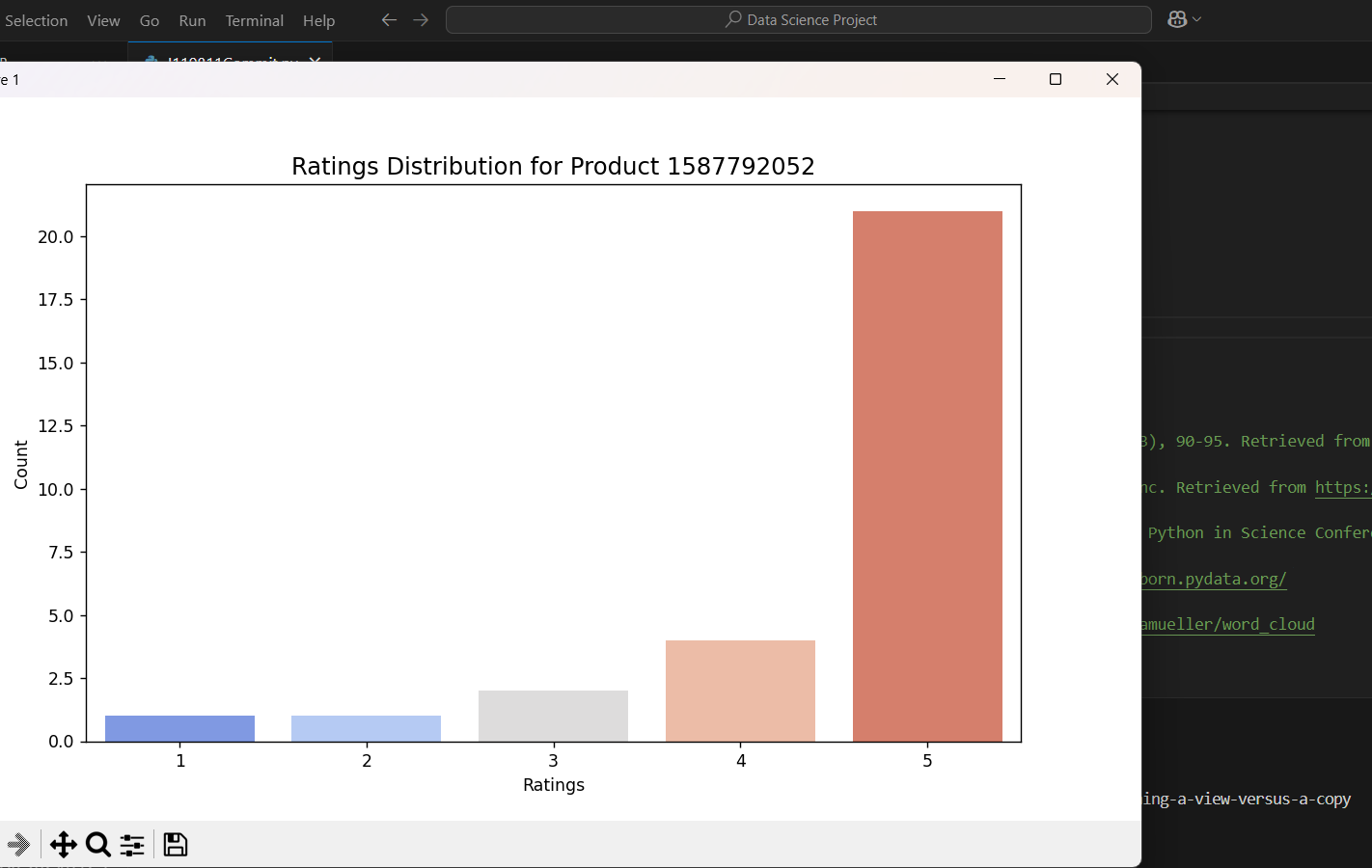


**Version Control:** GitLab was used as the versioning tool that enabled frequent committing to the repositories and enhancing coordination during parallel work.

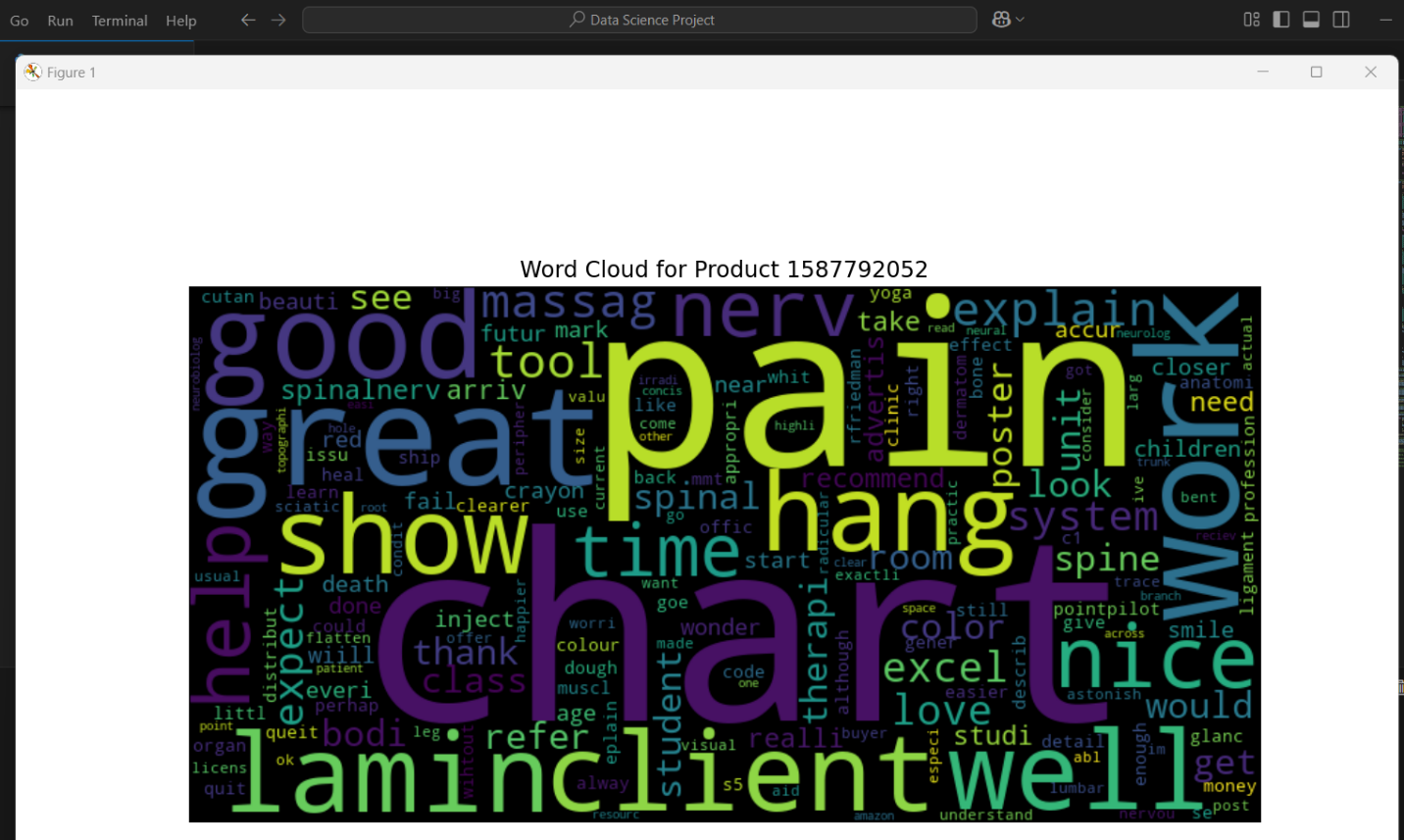


**Results**

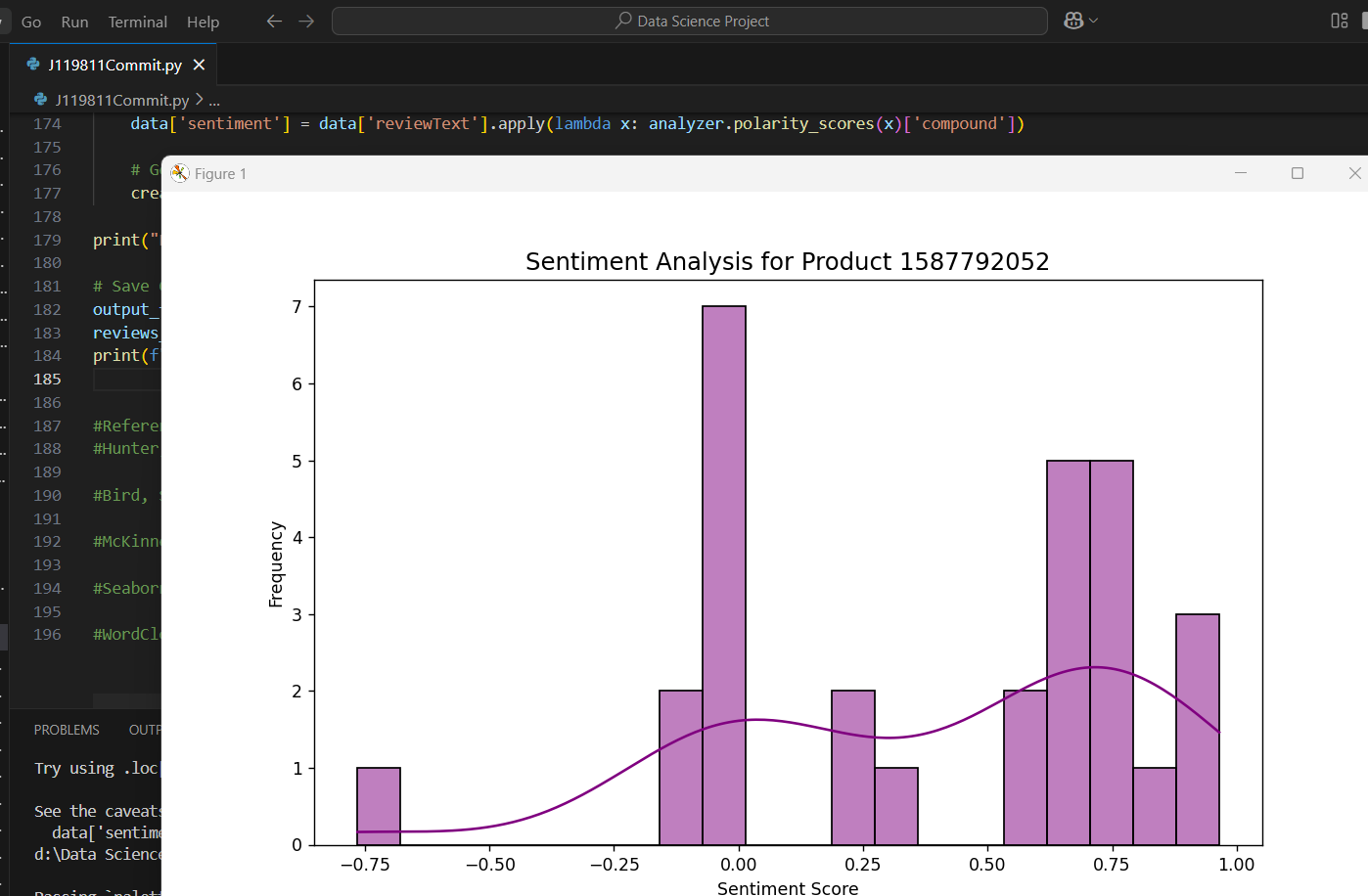
The examination of the dataset was useful for understanding specific aspects of customer feedback and outcomes of the products. Below are the key findings derived from the three selected products:

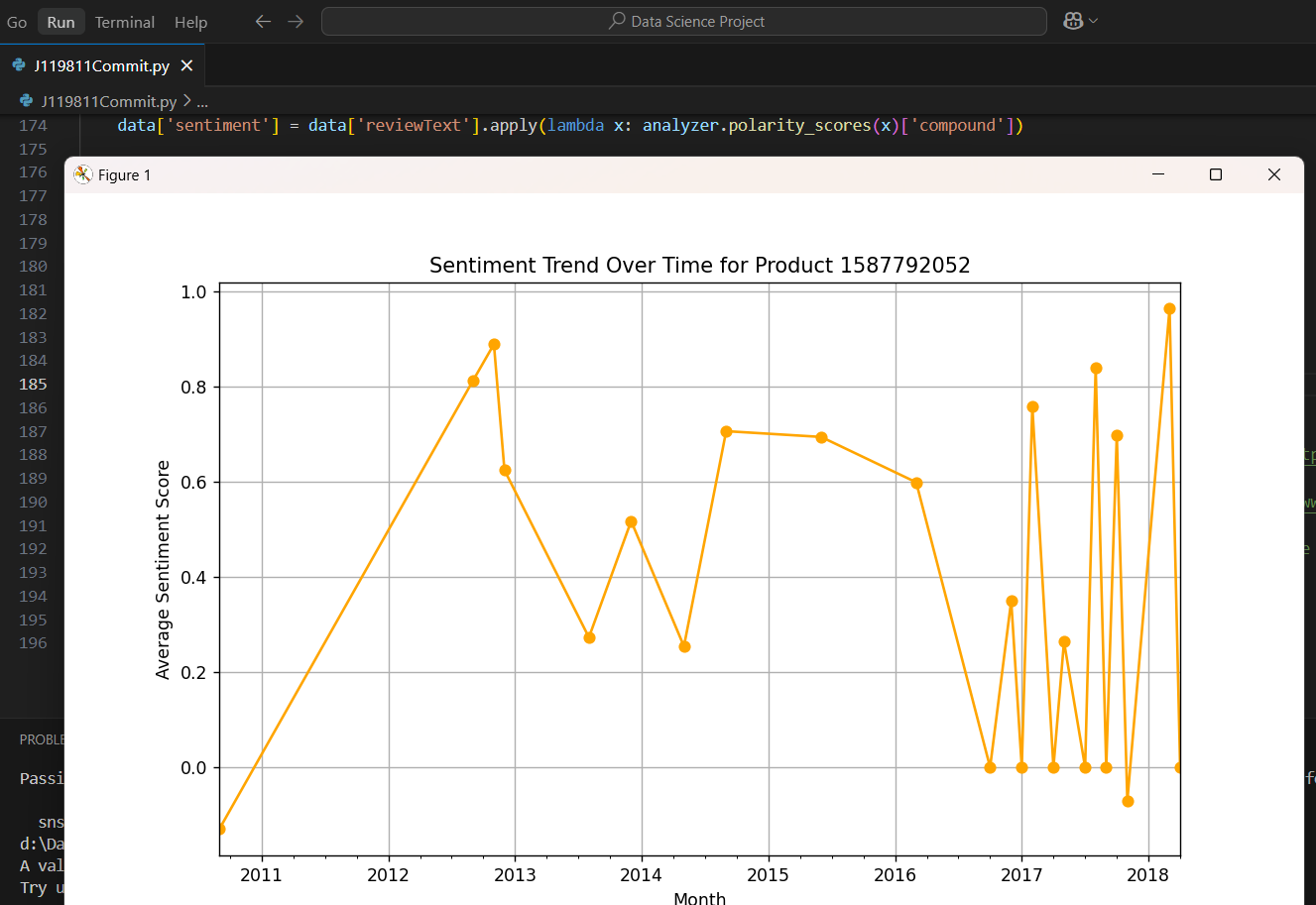
**Rating Distribution:** Through bar plots, the number of each type of rating was shown and this indicated the general perception of the customers.

**Common Themes in Reviews:** The word cloud for Product 1587792052 highlights key terms like "pain," "chart," "tool," "nice," and "help," reflecting its relevance in educational or therapeutic contexts. Positive words such as "great" and "excellent" indicate customer satisfaction, while terms like "spinal nerve" and "system" emphasize its specialized usage for anatomical or medical purposes. (Gallery of Examples — wordcloud, n.d.).

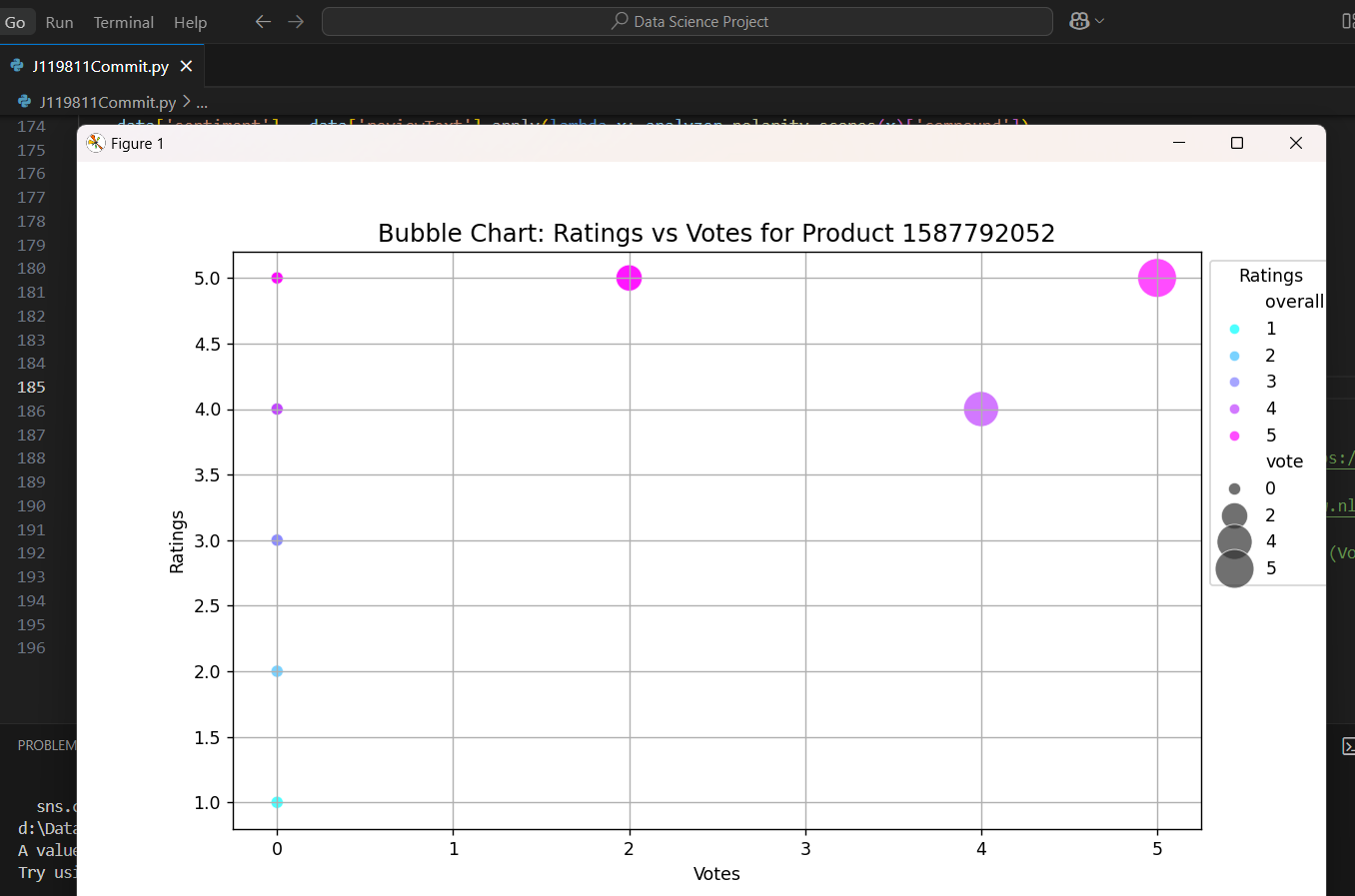


**Sentiment Analysis:** Time-series trends from sentiments are observed seasonally for better reviews.

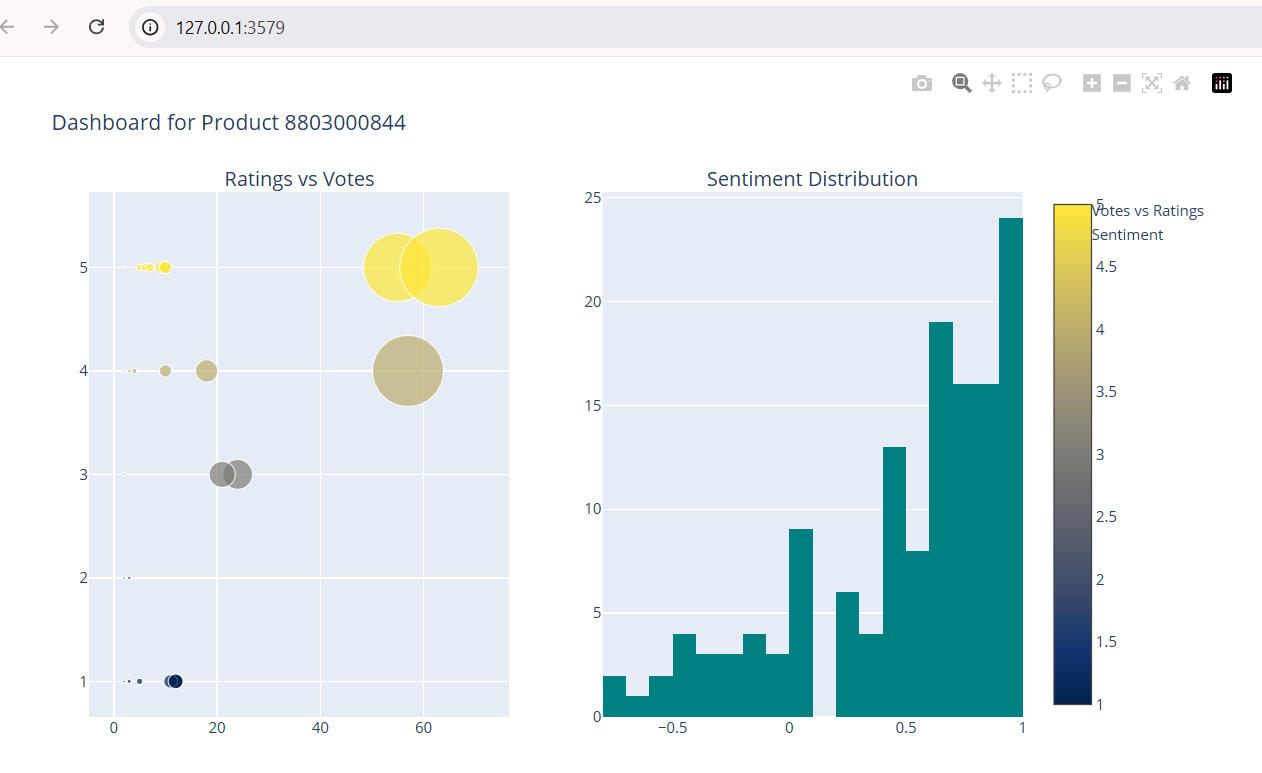




**Relationship Between Votes and Ratings:** An example was the use of bubble charts to demonstrate a connection between the number of votes a review had and the rating. (Waskom, 2021).



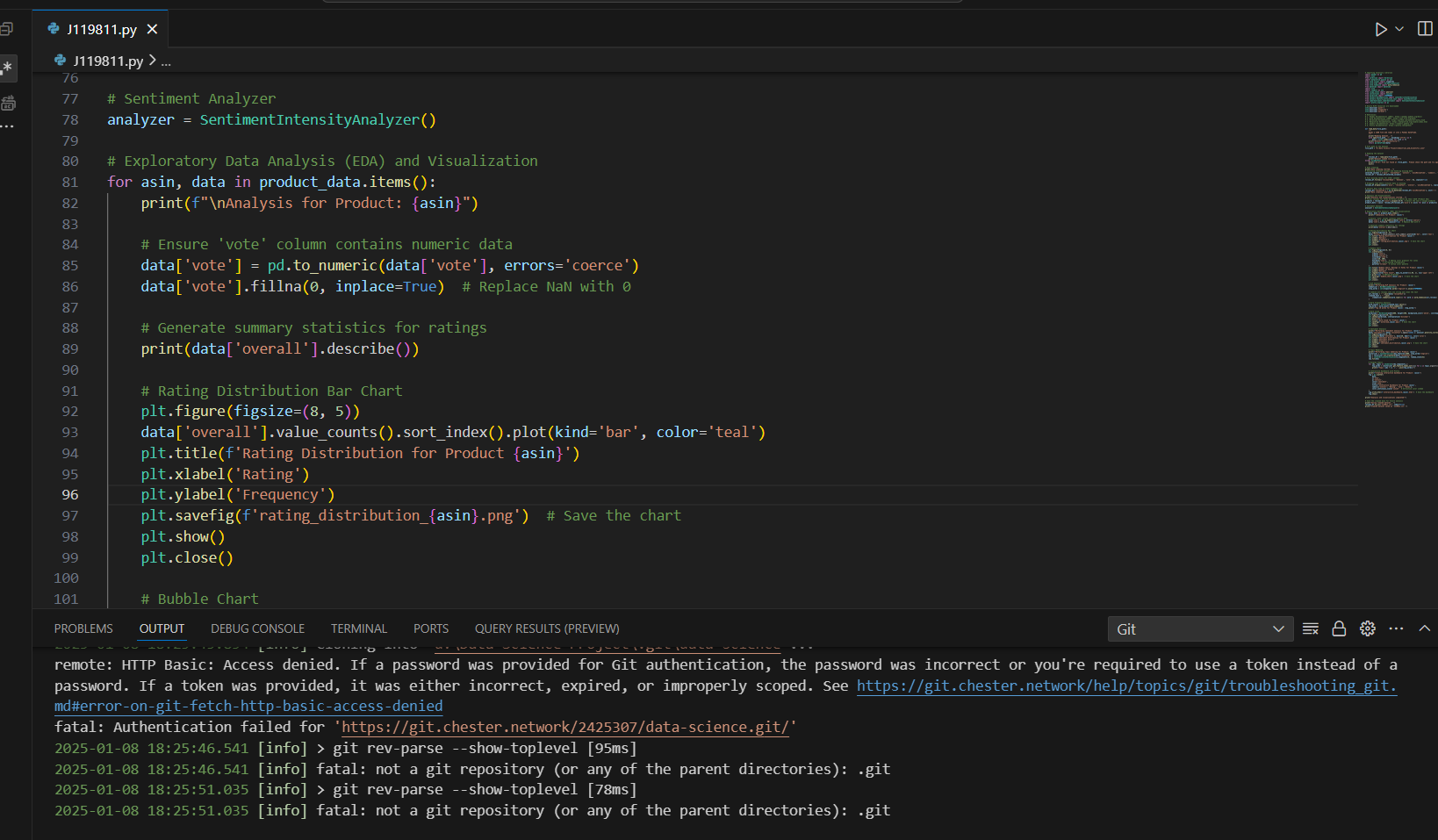
**Interactive Dashboards**: It offeres easy-to-use controls, an instant update of the graph for assessing sentiments and ratings in several dimensions (Plotly, n.d.).

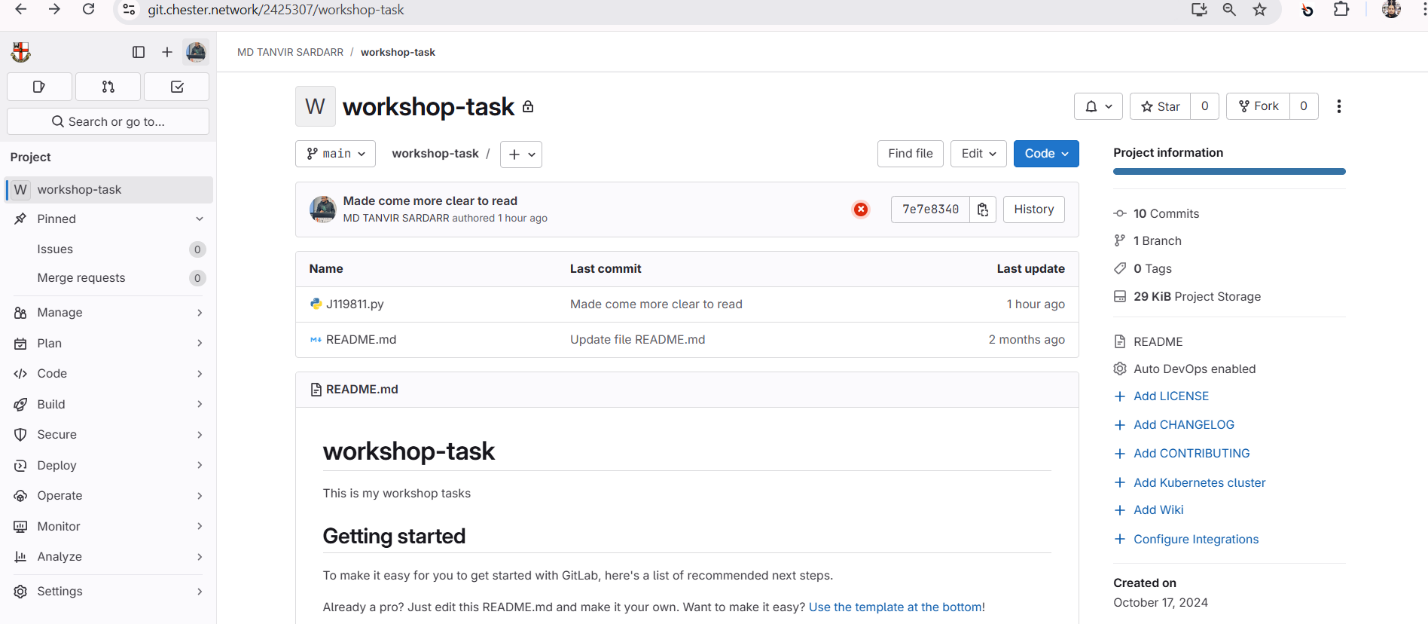


**Challenges Faced**

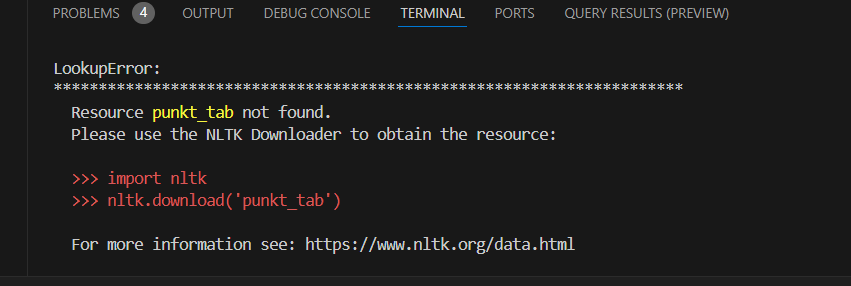
During the implementation of this project, a couple of challenges arose. The following is a summary of these challenges and their respective solutions:

**GitLab Authentication Challenge:** Manually attempted to connect to the university GitLab repository as ssh-keys and got a "Permission denied" error.



**Resolution**: This was done by using the university-authenticated GitLab web interface for authentication and repository management.

**NLTK Resource Error Challenge:** There was an error running the code saying that the resource punkt\_tab was missing, which was needed by the NLTK tokenizer.



**Solution:** The resource was downloaded using the command nltk.download('punkt'), and the tokenizer started working as expected.

**Critical Evaluation**

This project is successful in extracting meaningful insights from a very complex dataset while using a wide variety of data science techniques. This project has engaged a lot of data science techniques to extract important information from a very complicated data set. However, some areas can be improved upon:

**Strengths of the Tool:** The tool combined several libraries for data visualization, NLP, and data preprocessing; therefore, the analytical solution was complete. The interactive dashboard can be effective in easing the exploration of data, as well as presenting dynamic and rather visually appealing information to the users. Moreover, GitLab was utilized effectively to keep transparency and manage the accountability in the development process.

**Limitations:** The sentiment analysis is limited to English language reviews; therefore, all the valuable information from the non-English reviews, which may form a large portion of the data, is excluded. It also limits itself to a particular time window which means any long-term trend or shift in the customer sentiment was ignored. It also does not leverage advanced machine learning techniques that could offer better nuanced sentiment classification and maybe event trend predictions. It focused on only three products, hence making findings not generalize able across the whole dataset.

**Conclusion**

This project effectively showcased how data science techniques can analyze customer opinions and offer actionable recommendations for product enhancements. By utilizing a framework that encompassed data cleaning, exploratory data analysis, natural language processing, and advanced visualizations, it enabled a thorough analysis. The interactive dashboards increased accessibility, allowing stakeholders to spot trends and tackle issues. However, limitations such as focusing on only three products and depending on basic sentiment analysis methods limited its generalizability. To improve the tool’s impact, addressing these gaps by incorporating advanced machine learning and accommodating non-English reviews would be beneficial.

**References**

*Plotly*. (n.d.). <https://plotly.com/python/>

*View of VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text*. (n.d.). <https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399>

Waskom, M. (2021). seaborn: statistical data visualization. *The Journal of Open Source Software*, *6*(60), 3021. <https://doi.org/10.21105/joss.03021>

*Gallery of Examples — wordcloud 1.8.1 documentation*. (n.d.). <https://amueller.github.io/word_cloud/auto_examples/index.html>

*MatPlotLib: a 2D Graphics environment*. (2007, June 1). IEEE Journals & Magazine | IEEE Xplore. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4160265>