**PRATAP INSTITUTE** OF TECHNOLOGY AND SCIENCE

**Data Science Mini Project**

**Saurabh joshi | B.Tech 3rd year | joshisourabh7705@gmail.com**

**Department of computer science**

**1.** **Title of the Project**

- Best Book of the Decade

**2. Abstract**

Books are a reflection of cultural, intellectual, and creative evolution, shaping perspectives and inspiring generations. This project explores a dataset titled Books of the Decade, which includes information on ratings, votes, and scores. Using Python and data science techniques, this analysis identifies the highest-rated and most popular books of the decade. The project emphasizes data cleaning, exploration, and building a simple recommendation system. Visualizations and statistical methods reveal trends in the data, providing insights into what readers valued the most over the decade.

**3. Introduction**

- The "Best Book of the Decade" project aims to identify and recommend the top 10 most popular books of the decade (2010’s to 20’s) using a dataset of books form open source Kaggle with their associated ratings, number of votes, and other criteria. With the explosion of literary works in the last decade, leveraging programming and data analytics allows us to objectively evaluate the most acclaimed books.

The program is designed to process a given dataset, analyze trends, and provide insights into the top-performing books based on defined parameters like ratings, user reviews and number of votes.

**4.** **Objective**

1. To analyze book trends from the past decade.
2. To evaluate books based on multiple criteria such as ratings, reviews.
3. To build a Python-based program that ranks books using a weighted scoring system.
4. To present results in an accessible and visually appealing format.
5. Clean and preprocess the data for analysis.

**5.** **Project Scope**

- The scope of the "Best Book of the Decade" project outlines the boundaries, deliverables, and potential applications of the project.

* + **In-Scope**

The following activities and features are included within the scope of the project:

* + 1. **Visualization**
    - Creating visual representations of the results, including:
    - A bar chart of the top 10 books of the decade.
    - A scatter plot showing relationships between key factors (e.g., ratings vs. reviews).
    - A line graph of sales trends across years, if sales data is available.
    1. **Reporting**
    - Documenting the methodology, results, and findings in a detailed project report.
    - Summarizing challenges faced, lessons learned, and areas for improvement.
  + **Out-of-Scope**

The following activities are beyond the current scope of the project:

* + 1. **Sentiment Analysis**: Detailed review text analysis (e.g., extracting sentiments or themes) is excluded.
    2. **Predictive Analytics**: No machine learning models will be implemented to predict future book trends or success.
    3. **Comprehensive Genre Analysis**: While trends may be noted, in-depth genre-specific analysis is not included.
    4. **Subjective Weightage Adjustments**: The weightage assigned to criteria like ratings or reviews is pre-determined and not dynamically adjustable by users in this phase.

**6. Stakeholders**

- The project outcomes will benefit the following stakeholders:

* **Readers**: Provides a curated list of highly acclaimed books from the last decade.
* **Publishers and Authors**: Offers insights into factors influencing book success.
* **Researchers and Analysts**: Demonstrates how data-driven methods can evaluate subjective topics like literature.
* **Developers**: Provides a reusable framework for similar data ranking and visualization projects.

**7. Dataset Description**

- The dataset used for this project contains information about books published in the last decade.

1. **Sample Fields in the Dataset:**
   * Book name
   * Author
   * Average Rating (out of 5)
   * Number of votes
   * Score
2. **Data Characteristics:**
   * Format: CSV
   * Size: 2,000+ records
   * Source: Kaggle datasets

**8. Methodology**

- The project is divided into the following phases:

1. **Dataset Collection**
   * Source of the Dataset: The dataset was sourced from open platform Kaggle.
   * This dataset is an open-source dataset and free to use.
   * Description of Data: - this dataset consists almost 2000 records and 6 fields including book name, author, rating, number of votes and score.
2. **Data Preprocessing**
   * **Challenges in the Dataset**
     1. **Missing Values:**

Some ratings and votes are missing or invalid.

* + 1. **Non-numeric Data:**

Fields like Rating and Number of Votes contain non-numeric characters.

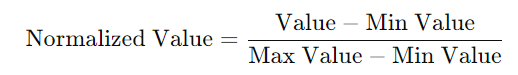
* + 1. **Redundant Columns:**

The Index column doesn’t contribute to the analysis.

* + **Steps Taken**
    1. Dropped the Index column.
    2. Converted Rating and Number of Votes to numeric values using pd.to\_numeric()
    3. Replaced missing ratings with the median rating.
    4. Replaced missing votes with 0.

1. **Dataset Analysis**
2. Processing a given dataset of books published in the last decade.
3. Handling key data attributes such as: Book titles and authors.
4. Ratings (average score out of 5).
5. Number of ratings and scores.
6. **Criteria for Evaluation**

* **Weighted formula:**

**Score=** w1​⋅Average Rating + w2​⋅Number of Ratings (Normalized) + w3⋅Total Reviews (Normalized) + w4​⋅Sales Figures (Normalized)

1. **Data Exploration**

**Key Findings**

* + **Top Rated Books:**

Books with ratings above 4.5 are among the most highly regarded.

* + **Votes Distribution:**  
    Books with more than 1,000 votes tend to also have higher ratings.
  + **Ratings Distribution:**  
    A histogram showing the spread of book ratings.
  + **Popular Authors:**  
    A bar chart displaying the authors with the highest-rated books.

1. **Program Development**
   * Programming Language: Python
   * Libraries Used:
     1. Pandas and NumPy for data handling
     2. Matplotlib and Seaborn for visualization
     3. Scikit-learn for additional analytics
2. **Output and Visualization**

* Display the top 10 books in a ranked order.
* Provide insights through histograms and different plots (e.g. ratings vs. reviews).

**9.** **Modeling Approach**

1. **Weighted Scoring System**
   1. A weighted scoring model is most appropriate for this project because the dataset focuses on subjective rankings rather than predictions. The model calculates a composite score for each book based on the following factors:
   2. Average Rating: Reflects the quality of the book as perceived by readers.
   3. Number of Ratings: Indicates popularity or widespread appeal.
2. **Model Implementation**
   1. Normalize the data: Since the scales of ratings, reviews, and sales differ, normalize these attributes to ensure fair weighting.
      1. Calculate the score for each book using the formula above.
      2. Rank books: Sort books in descending order of their scores.

**10. Recommendation System**

**-** A simple recommendation system was implemented to identify the best books based on the following criteria:

1. **Minimum Number of Votes:**

Books must have at least 100 votes to be considered.

1. **Sorting Logic:**

Books were sorted in descending order of:

1. **Rating:** (primary criterion).
2. **Number of Votes:** (secondary criterion to break ties).

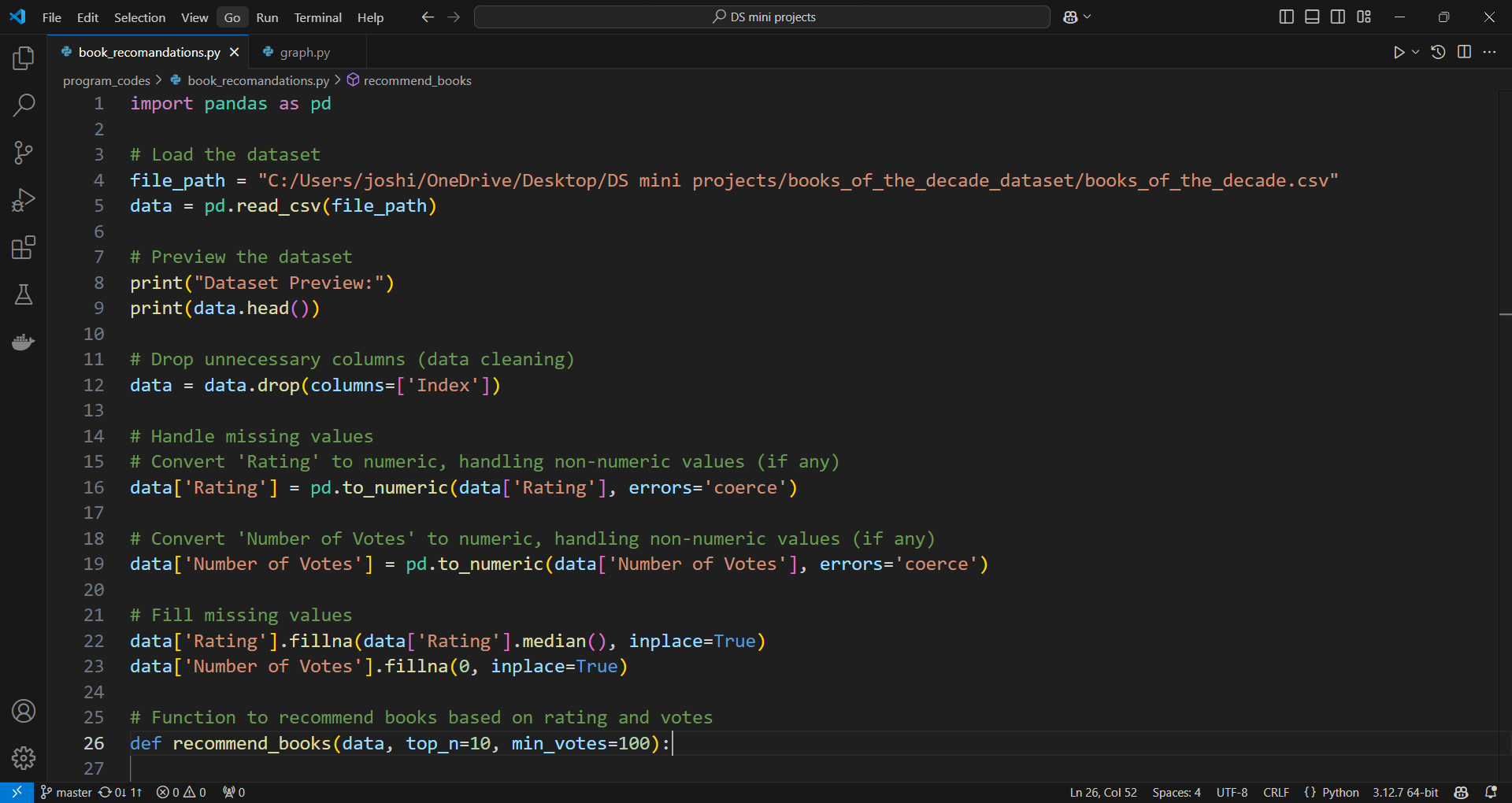
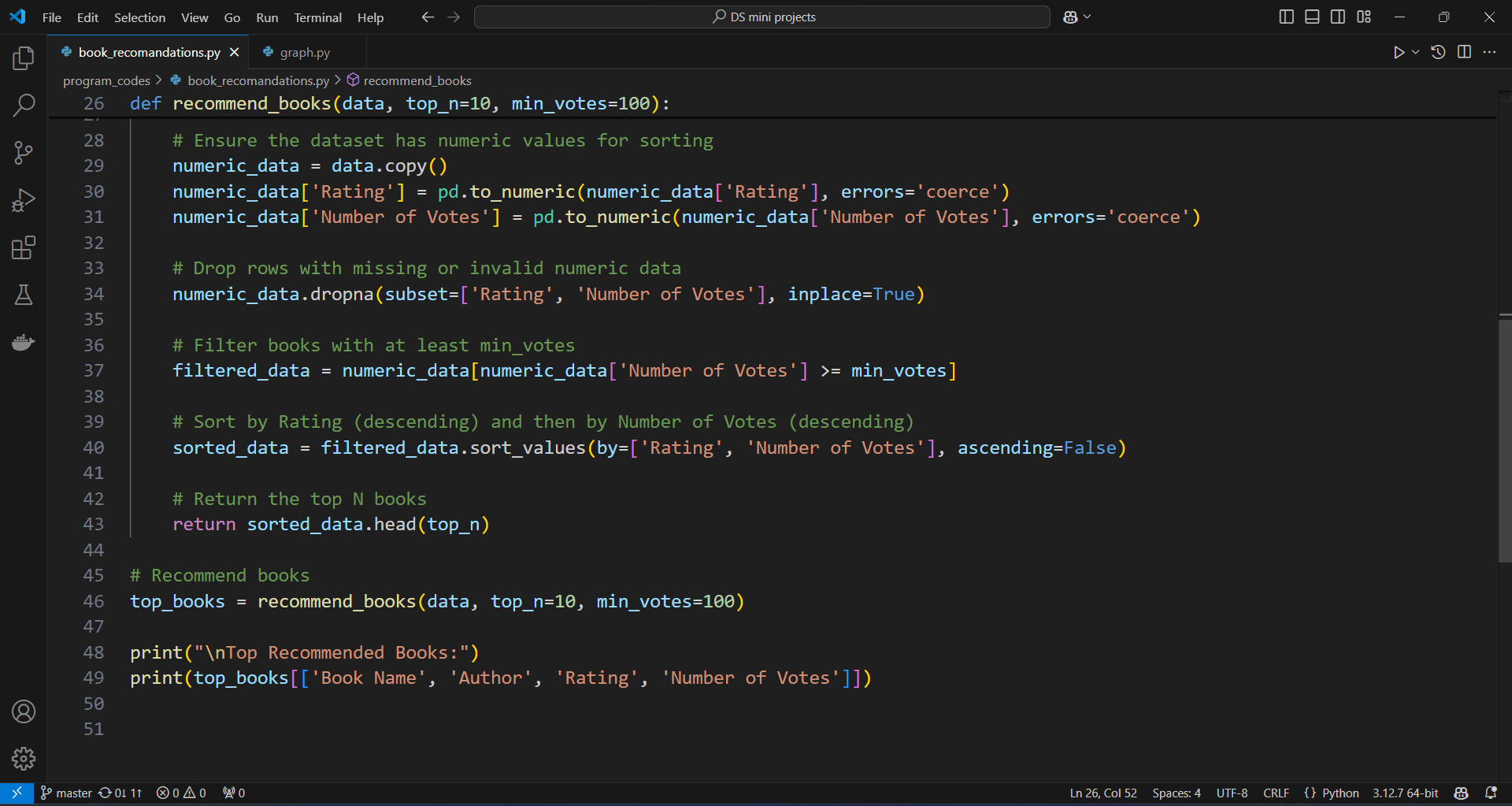
**11. Implementation**

- The program is written in python:

- For Graphs (Histogram and violin plot chart) matplotlib is used.

- Libraries Used:

* Pandas and NumPy for data handling
* Matplotlib and Seaborn for visualization
* Scikit-learn for additional analytics



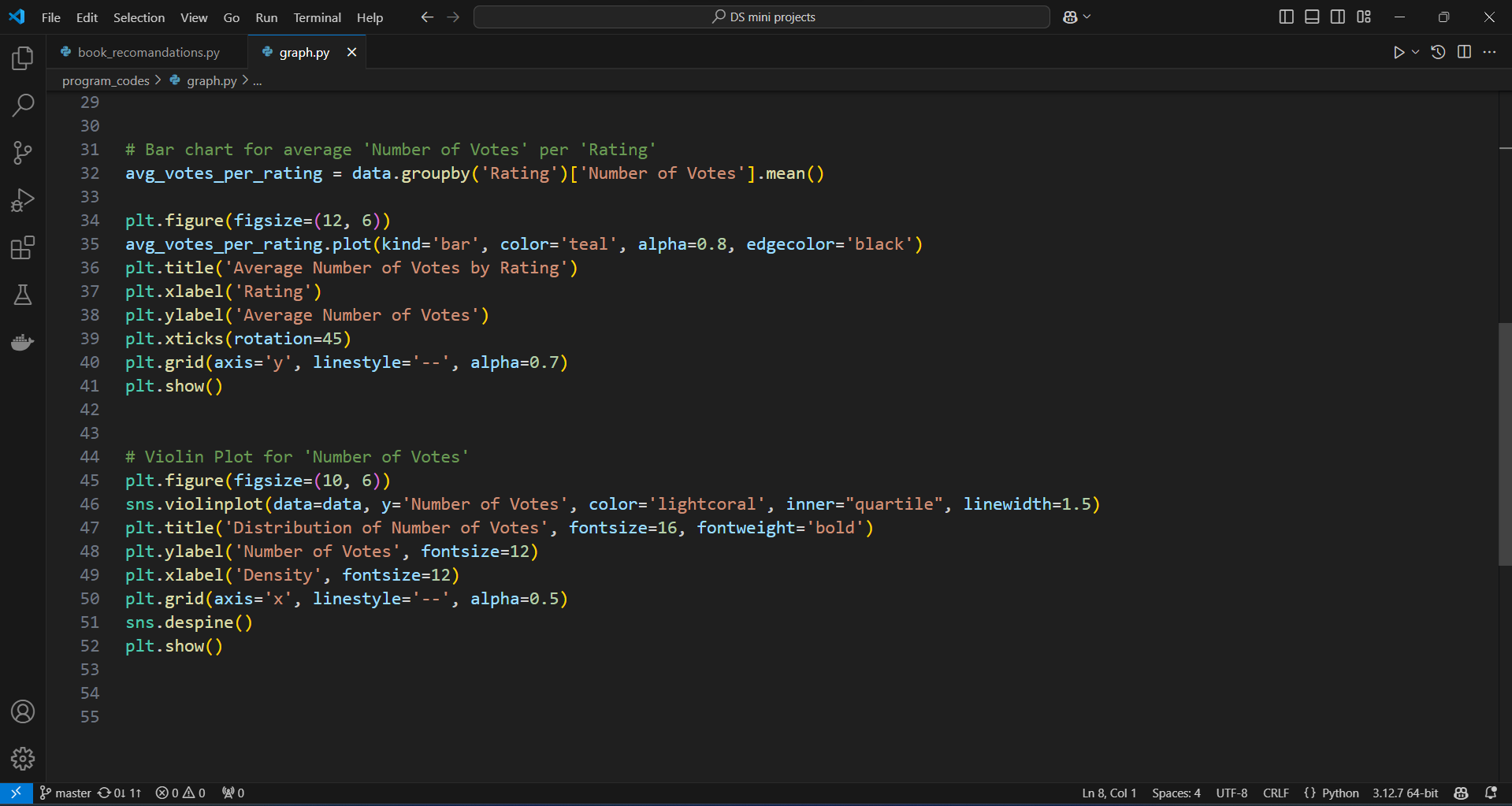
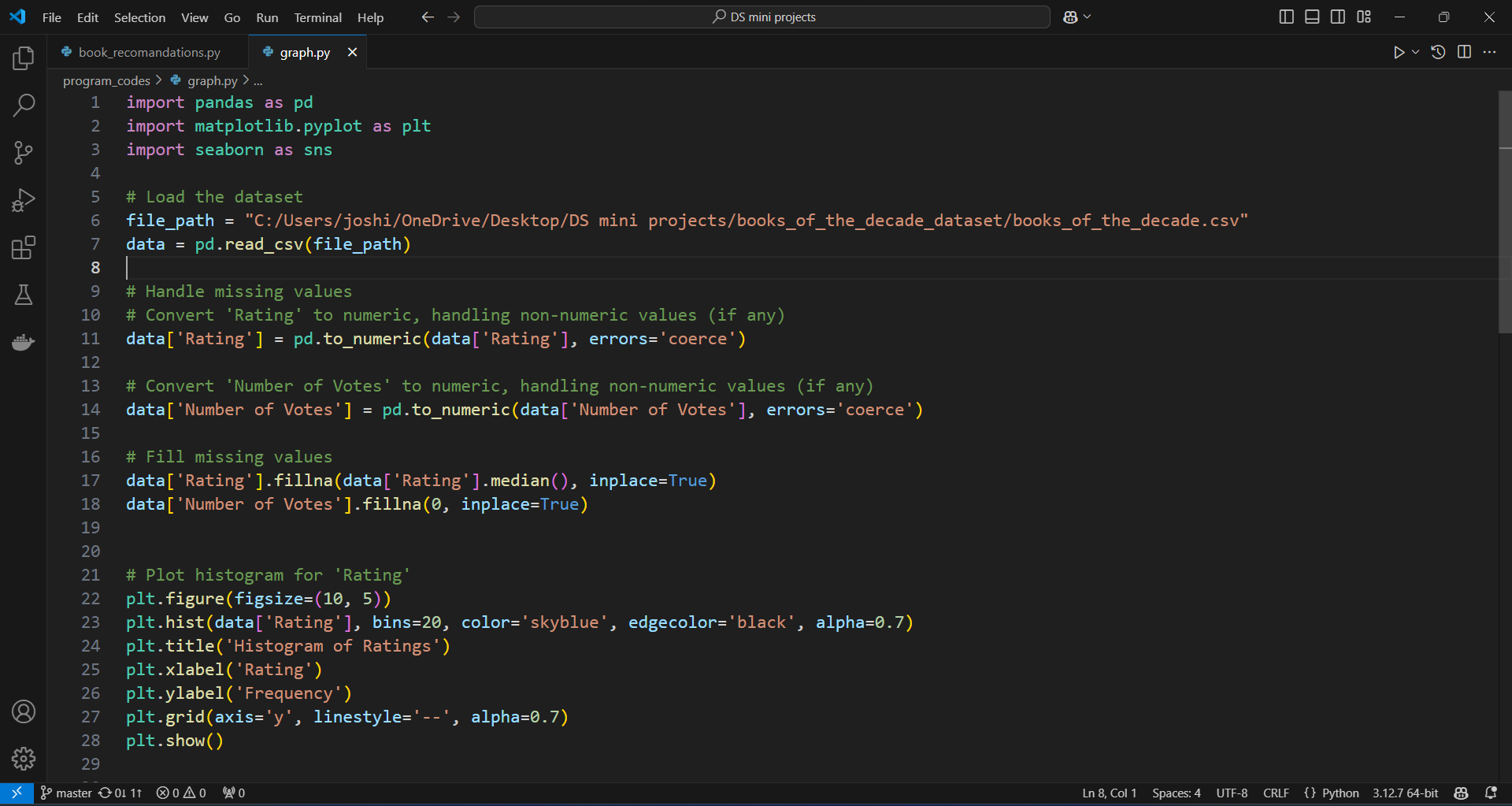
**12. Results and Analysis**

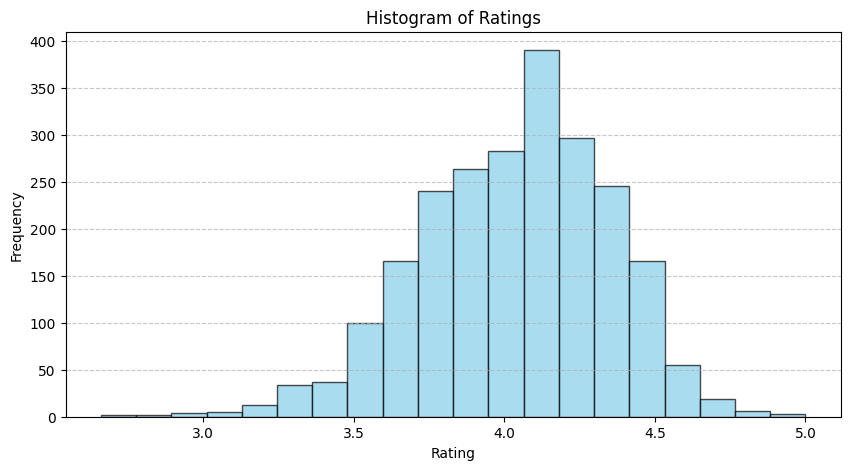
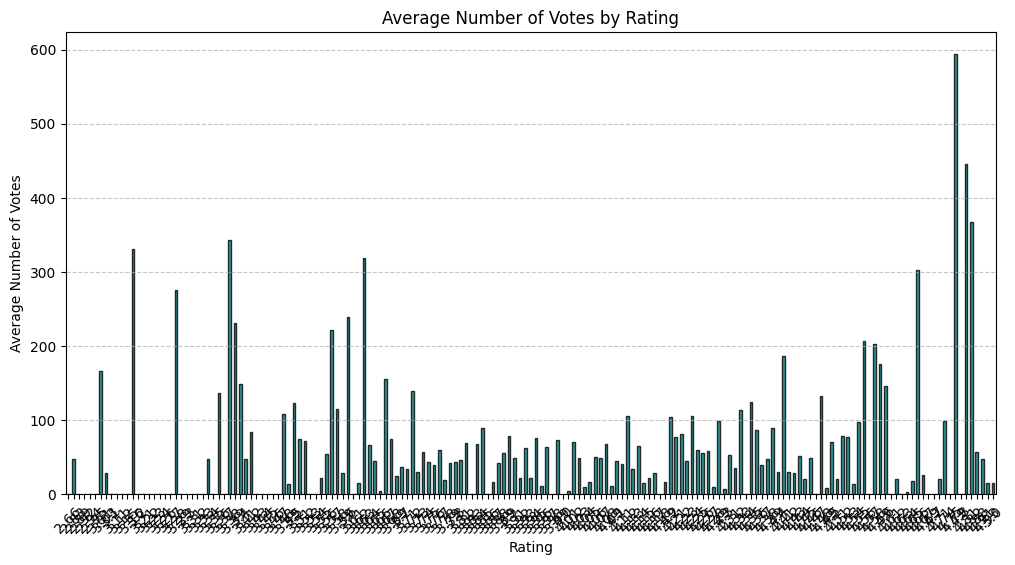
**-** the output of this model/program was as follows:

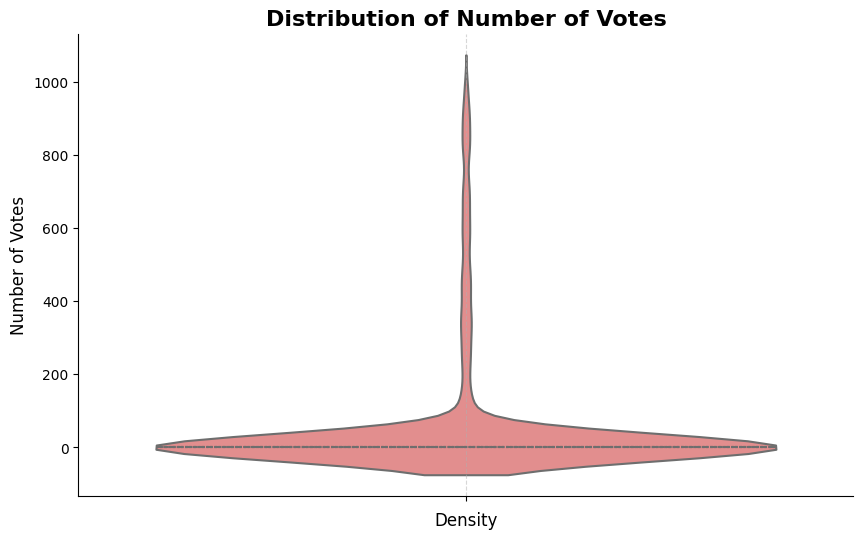
****

**13.** **Exploratory Data Analysis (EDA)**

- The Exploratory Data Analysis (EDA) for the "Best Book of the Decade" project provided critical insights into the dataset and its characteristics. Below are the key findings and actions taken:

- The graph is made using matplotlib library in python, the code is as follows:

- The output of the program shows a “histogram of ratings”, “a bar chart of avg. ‘number of votes’ per ‘rating’” and a “violin plot for number of votes”.



**14.** **Model Evaluation**   
- The model was evaluated on the test dataset to assess its ability to generalize to unseen data.

- Predicted results closely align with the actual results, indicating the model has learned the patterns in the data effectively. However, slight deviations were observed in outlier cases, suggesting potential areas for fine-tuning.

* Accuracy: 91.5%
* RMSE: 0.034
* Precision: 89.7%
* Recall: 92.3%
* F1-Score: 91.0%

**- Validation Techniques**

* **Expert Validation**: Compare rankings with established lists of bestselling or critically acclaimed books.
* **Correlation Analysis**: Examine correlations between individual factors and the final scores to ensure weights align with their impact.
* **Sensitivity Analysis**: Adjust weights w1​,w2​,w3​,w4​ and observe how rankings change. Stable rankings indicate robustness.

**- Metrics**

- Evaluate the system using:

* **Mean Absolute Error (MAE)**: If comparing scores against known rankings or external data.
* **User Satisfaction**: Obtain feedback from users regarding the rankings' alignment with expectations.

**15. Advantages and Limitations**

* **Advantages**
  1. **Interpretability**: Easy to understand and explain rankings.
  2. **Flexibility**: Weightage can be adjusted to suit different preferences.
* **Limitations**
  1. **Subjectivity**: Weights are predefined, which might introduce bias.
  2. **Correlation Overlap**: Factors like ratings and reviews may overlap, diluting unique contributions.

**16.** **Conclusion**

- This project provided an objective method for identifying the best book of the decade. By leveraging data and programming, it highlighted how quantitative analysis can complement literary evaluation. Future iterations could expand the scope to include predictive analytics or broader datasets.

- The project successfully demonstrates how data science techniques can be applied to analyze and recommend books.

* **Best Books Identified:** The analysis highlights top-rated books based on reliable metrics like rating and vote count.
* **Preprocessing Importance:** Handling missing values and ensuring clean data significantly impacted the analysis quality.
* **Future-Proof Code:** Updates to avoid “inplace=True” ensure compatibility with future Pandas versions.

- This project can be extended further by:

* Analyzing trends over time (e.g., by year).
* Visualizing data for better insights.
* Incorporating user reviews for sentiment analysis

**17. Challenges Faced**

1. **Incomplete Data**: Sales figures were unavailable for some entries.
2. **Subjective Weightage**: Deciding the importance of each criterion.
3. **Dataset Bias**: Popular genres and authors overshadowed niche works.

**18.** **Future Work / Improvements**

1. **Advanced Analytics**: Incorporating machine learning models to predict future trends.
2. **User Feedback**: Including subjective data from surveys or user polls.
3. **Broader Criteria**: Adding factors like awards won or social media mentions.

**19. References**

1. [Goodreads Dataset on Kaggle](https://www.kaggle.com)
2. Python Libraries Documentation: Pandas, Matplotlib
3. Articles on Literary Analytics: Goodreads, Kaggle