**Enterprise Business Analytics: Practical Assignment**

**Programming Language used: Python**

**GitHub Link:**

**Task 1**

1. **Introduction**

**The objective of the Assignment**

The objective of this assignment, in line with Enterprise Business Analytics, was to create a detailed dataset that captures comprehensive information on bookstore data. This dataset aims to provide actionable insights into sales performance, customer preferences, and market trends within the book industry. By analyzing this created dataset, businesses can make data-driven decisions to enhance inventory management, marketing strategies, and overall operational efficiency.

**Description of the dataset**

The created dataset focuses on book sales, encompassing a variety of attributes that offer a detailed view of the publishing industry's landscape. It includes columns such as Book Title, Author, Publisher, Genre, Publication Date, Pages, Sales, Customer Rating, Customer Reviews, Borrowed Count, and Social Media Mentions. This data is collected from diverse sources including publishers, bookstores, online platforms, libraries, and social media, ensuring a comprehensive and multi-faceted perspective on book performance and customer engagement.

Each entry in the dataset represents a specific book, providing key details like the title, author, and publisher, which are crucial for identifying and categorizing the books. The genre and publication date columns offer insights into the types of books being published and their release timelines, which are critical for trend analysis. The number of pages is also included, potentially serving as a variable in understanding reader preferences and sales performance based on book length.

Sales data and customer feedback are pivotal components of this dataset. The Sales column reflects the commercial success of each book, while the Customer Rating and Customer Reviews columns provide qualitative insights into reader satisfaction and preferences. This combination of quantitative and qualitative data allows for a more nuanced analysis of book performance, helping publishers and retailers to make informed decisions regarding inventory management, marketing strategies, and future publications.

Additionally, the dataset includes Borrowed Count and Social Media Mentions, which broaden the scope of analysis beyond direct sales. Borrowed Count captures the popularity of books within library systems, highlighting reader interest that might not be reflected in sales alone. Social Media Mentions offer a real-time gauge of public interest and discussion, providing a pulse on current trends and potential viral hits. These elements make the dataset a valuable tool for comprehensive market analysis and strategic planning in the book industry.

1. **Data Collection**

**Data Sources**

* Publisher Sales Data: Data collected from various book publishers.
* Retail Sales Data: Data from bookstores and online retailers.
* Library Data: Borrowing statistics from libraries.
* Customer Reviews: Reviews from online platforms like Goodreads and Amazon.
* Market Reports: Industry reports on book sales and trends.
* Social media: Mentions and discussions about books on social media platforms.

**Data Collection Methods**

* Publisher Sales Data: Monthly sales reports from publishers.
* Retail Sales Data: Sales data collected from point-of-sale systems in bookstores and online sales reports.
* Library Data: Borrowing statistics from library databases.
* Customer Reviews: Scraping reviews and ratings from online platforms.
* Market Reports: Purchasing and analyzing industry market reports.
* Social media: Using social media listening tools to gather data on book discussions and mentions.

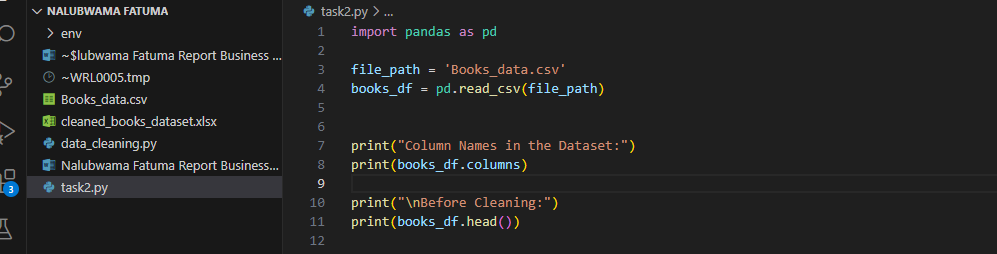
**Relevance of the Dataset**

This dataset is relevant for business analytics as it provides insights into:

* Sales Performance: Understanding which books are performing well in different markets.
* Customer Preferences: Analyzing customer feedback to understand preferences and trends.
* Market Trends: Identifying emerging trends and shifts in the book industry.
* Inventory Management: Optimizing stock levels based on sales and borrowing trends.
* Marketing Strategies: Develop targeted marketing campaigns based on customer demographics and preferences.

**Task 2**

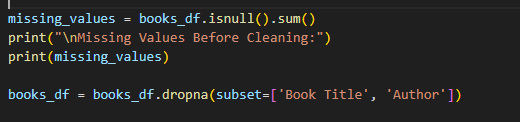
* 1. **Data cleaning steps**
     1. Loading dataset



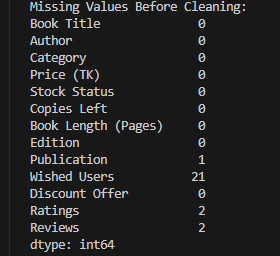
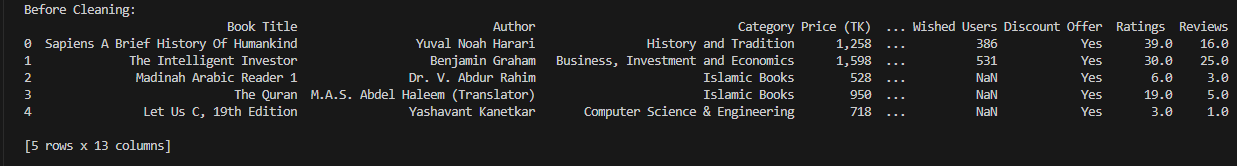
The dataset is loaded from a CSV file into a panda Frame using pandas.

* + 1. Handling Missing Values

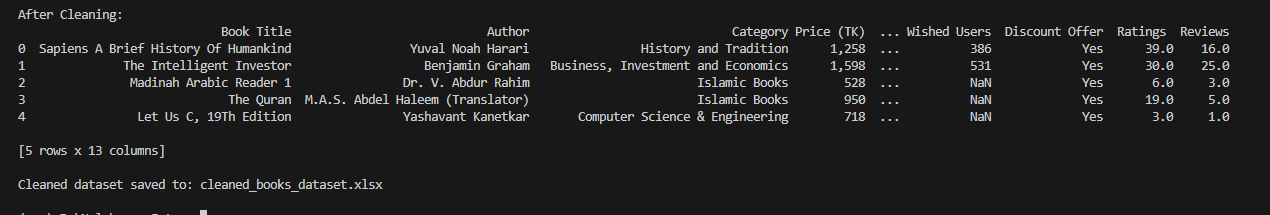
Checking for Missing Values and handling them by dropping them. For example by Title and Author.



* + 1. Removing duplicates and handling date formats
  1. **Before and after snapshots**

****

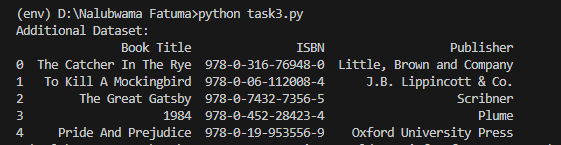
**After Data cleaning**

****

**Task 3**

**Process of Data Integration**

I created a smaller dataset containing 'Book Title', 'ISBN', and 'Publisher' as new columns.

****

Loaded both datasets into Pandas DataFrames.

The original dataset was cleaned by handling missing values, converting data types, and standardizing formats.

The datasets were then merged using the **pd.merge** function on the 'Book Title' column with a left join to retain all records from the original dataset and add matching records from the additional dataset.

**Challenges Faced**

Handling Text Case: Ensuring that the text case is consistent in both datasets to avoid mismatches during merging.

Handling Non-Numeric Values: The 'Copies Left' column contained non-numeric values like '50+' which needed to be converted to numeric.

Missing Data: Ensuring that missing values are properly handled in both datasets to prevent data loss during the merge.

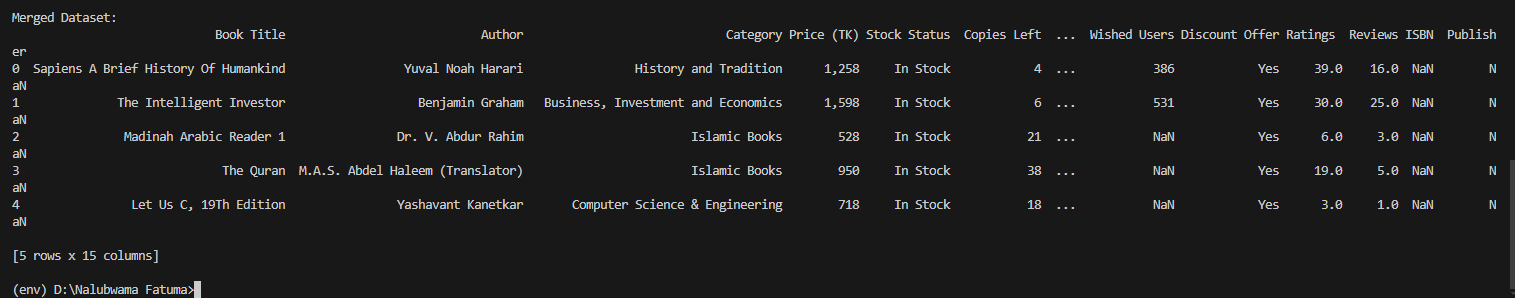
Ensuring Unique Keys: I had to make sure that the 'Book Title' column has consistent and unique values across both datasets for an accurate merge.

**Rationale for Integrating the Sources**

Integrating the dataset was valuable for the verification and validation of book information, and it was also necessary to enrich the dataset with more attributes to make it more useful.

**Data integration**

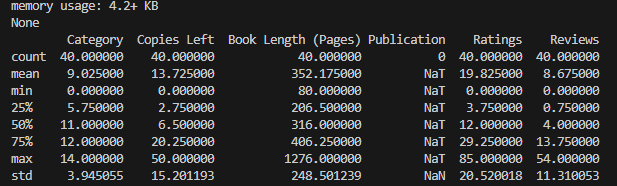
A small dataset was then integrated into the old dataset.



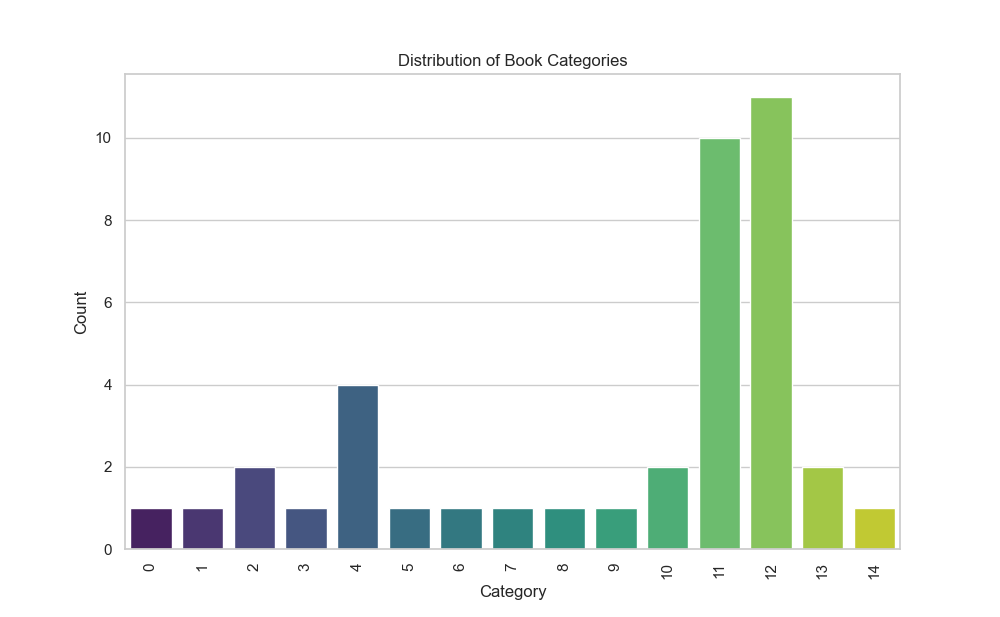
**Task4**

**Findings from the Analysis**

**Statistical analysis of the dataset**

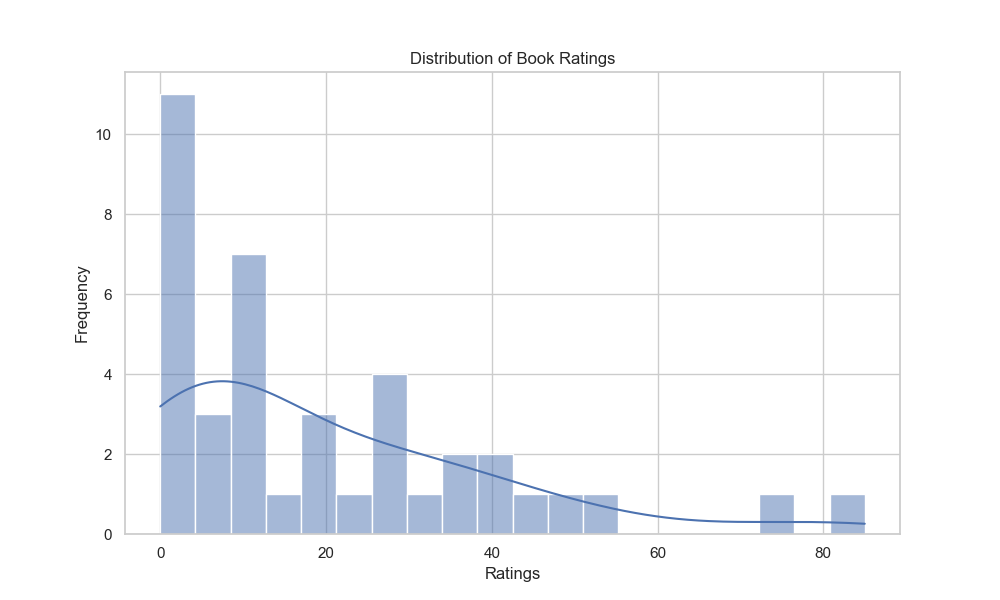
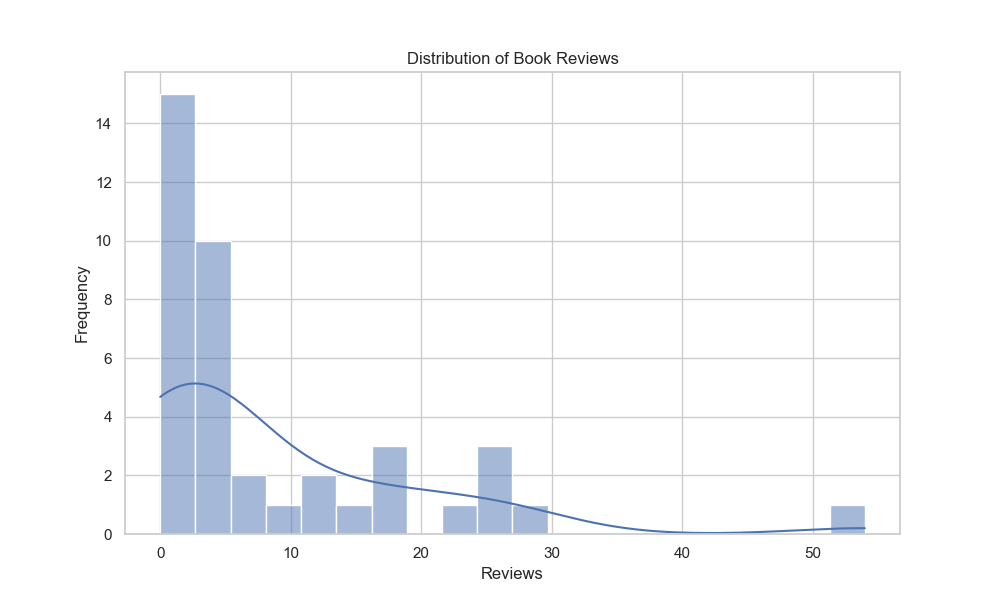
****

**EDA Insights**



The distribution of book categories shows the diversity in the dataset as shown above

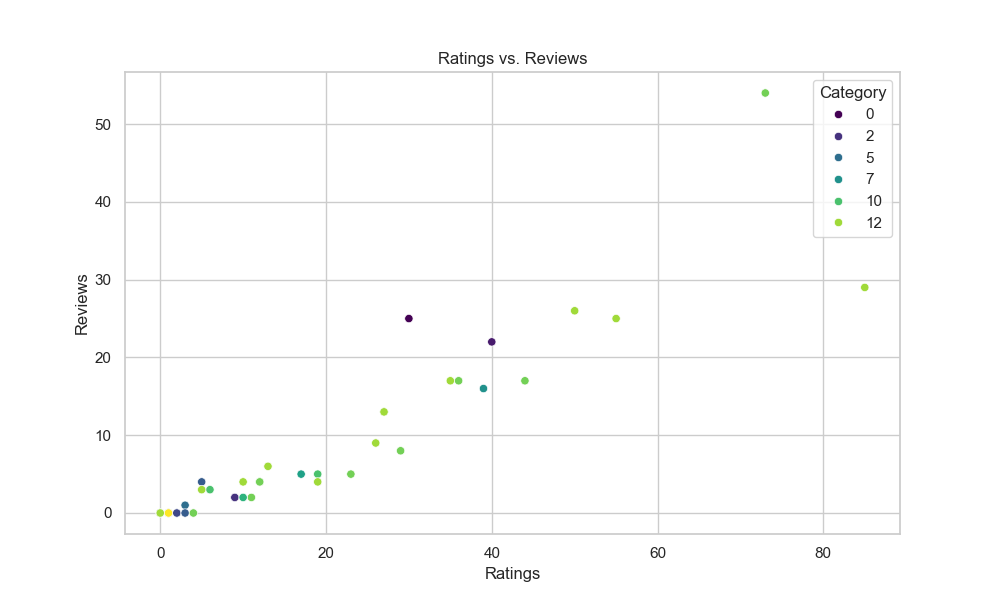
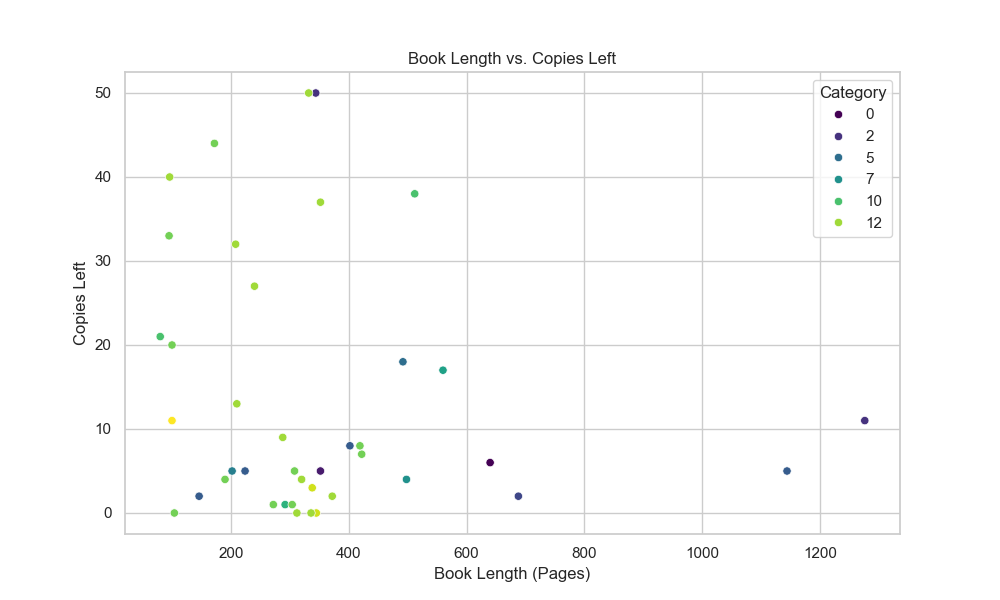
The distribution of book ratings and reviews indicates that most books have moderate ratings and reviews as depicted in the figures below.



The scatter plots reveal relationships between different features, such as ratings and reviews, book-length, and copies left.

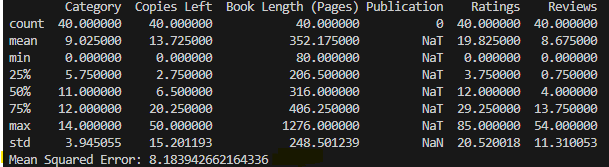
**Clustering**

KMeans clustering grouped the books into three clusters based on ratings, reviews, book-length, and copies left. The scatter plot visualizes these clusters.

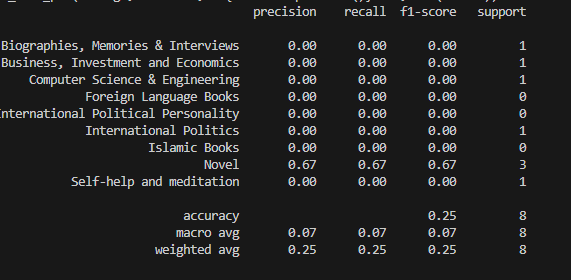


**Regression Analysis**

Linear regression was used to predict the number of reviews based on ratings, book-length, and copies left. The mean squared error indicates the accuracy of the model.



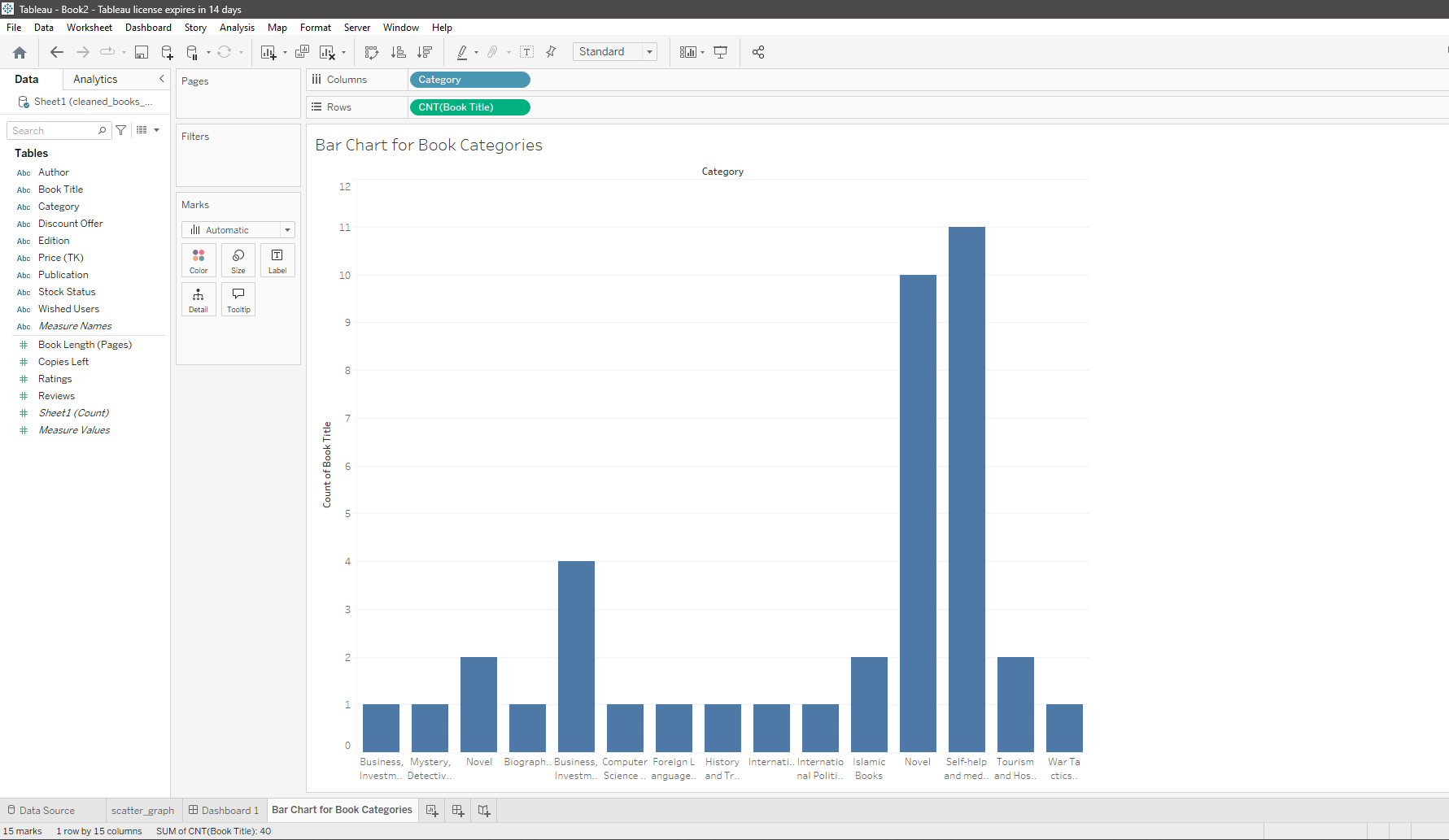
**Classification Analysis**

A linear regression model was used for the classification of book categories. The classification report provides detailed performance metrics for each category.

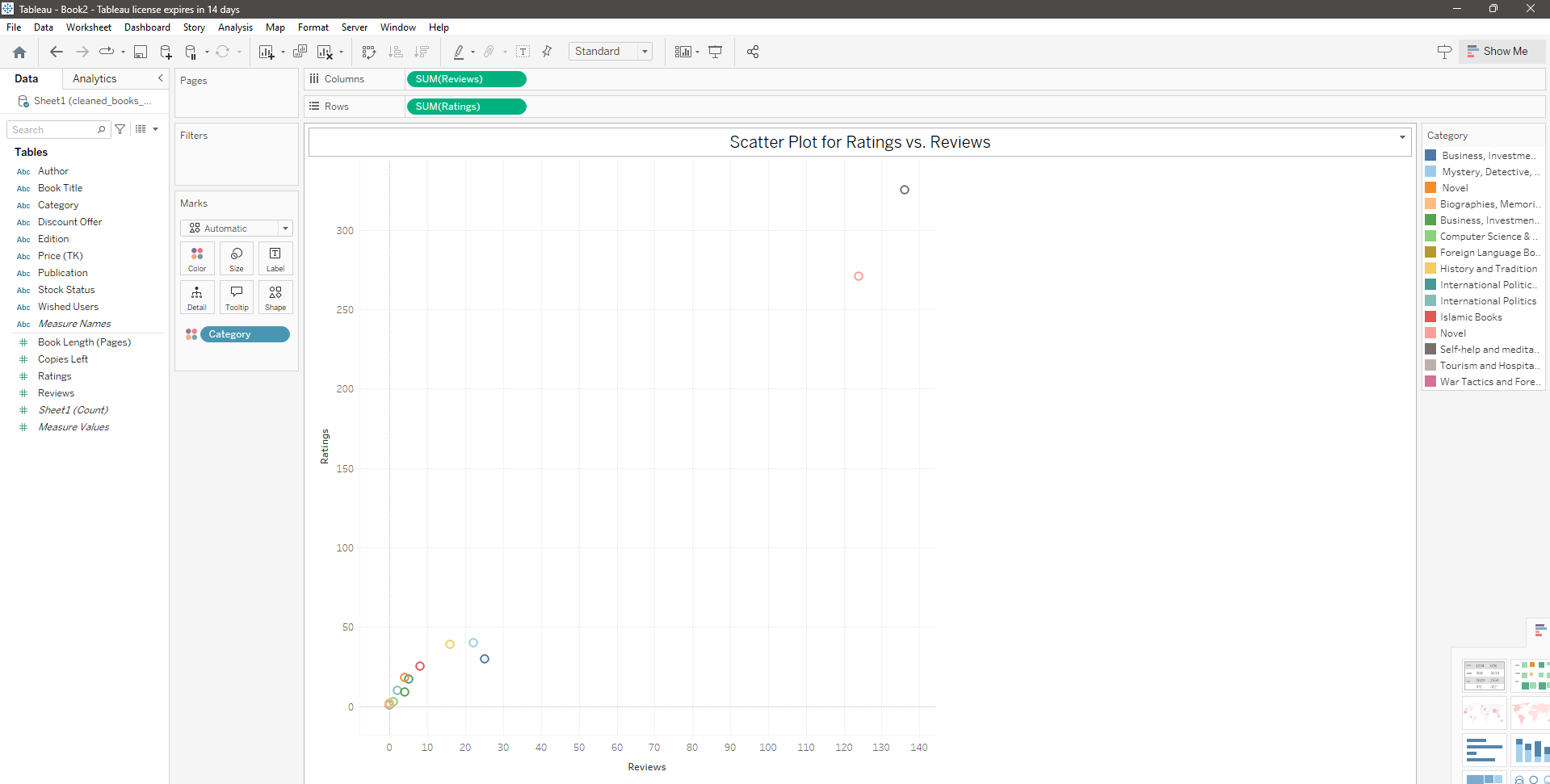
**Task 5**

**Insights from visualization.**

Bar Chart Insight. The bar chart reveals the distribution of books across various categories, which can inform decisions about which categories to focus on for new acquisitions or marketing.



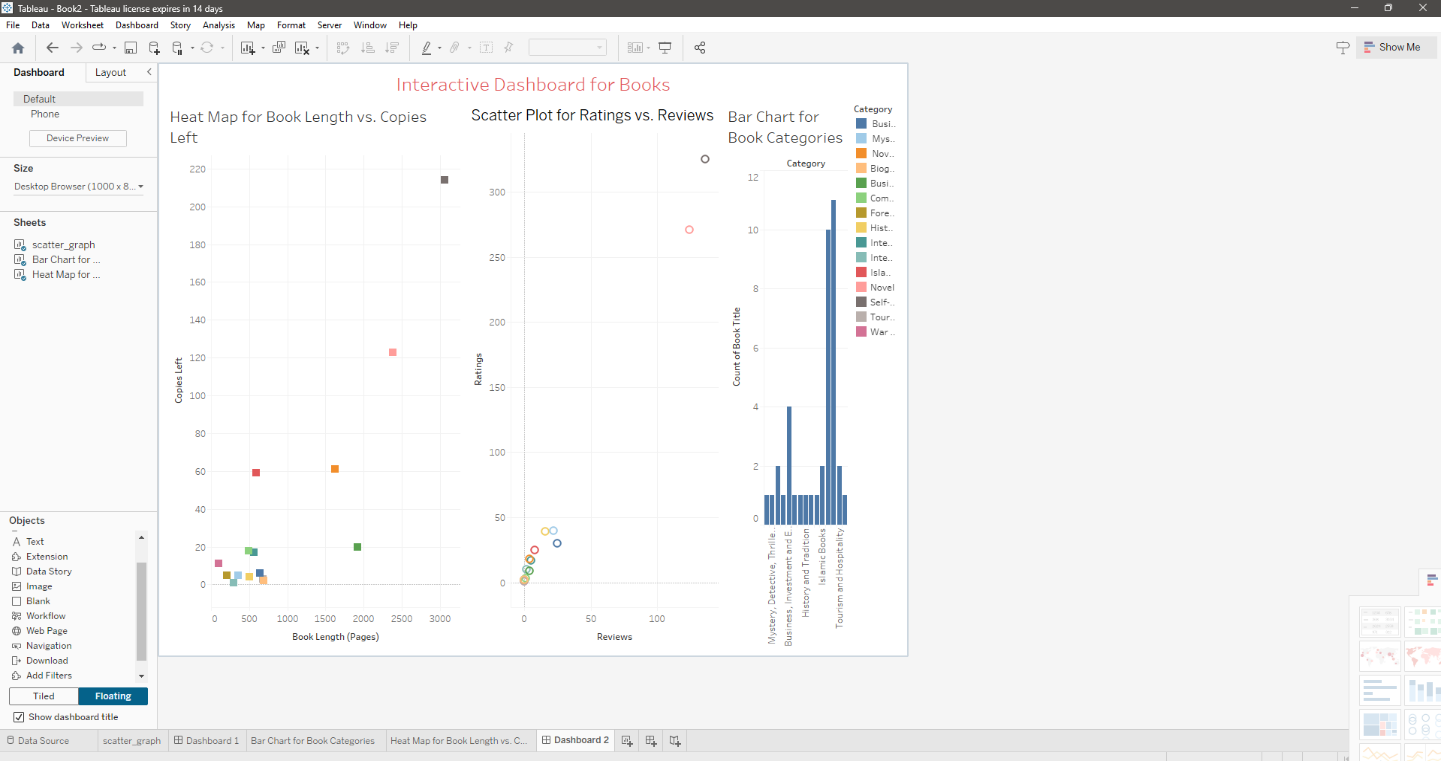
Scatter Plot Insight: The scatter plot highlights the relationship between ratings and reviews. Categories with high ratings but low reviews might need more visibility, while categories with low ratings and high reviews might need quality improvements.



Heat Map Insight: The heat map can identify patterns in book-length and copies left. For instance, if longer books have more copies left, it might indicate they are less popular, suggesting a need for adjusting the inventory strategy.



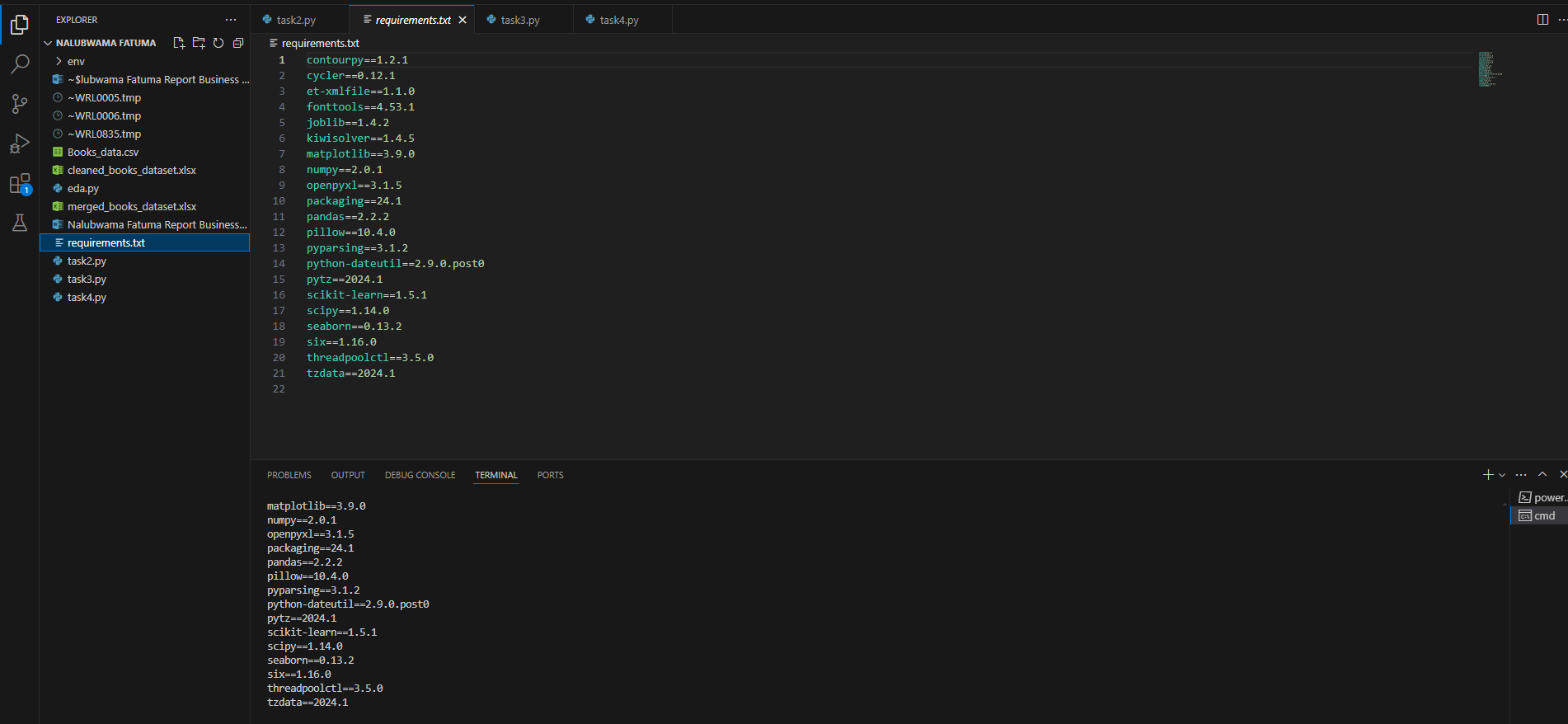
**Interactive Dashboard**

****

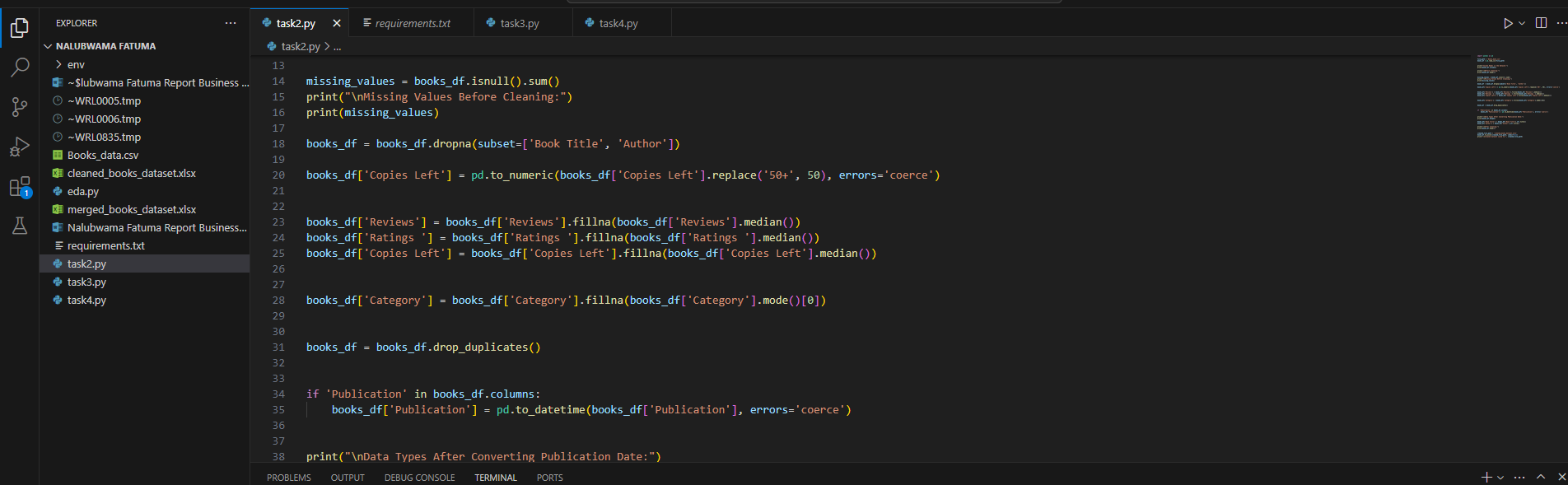
An interactive dashboard allows users to explore the data dynamically, offering more customized insights based on user queries.

**Code Snippets**

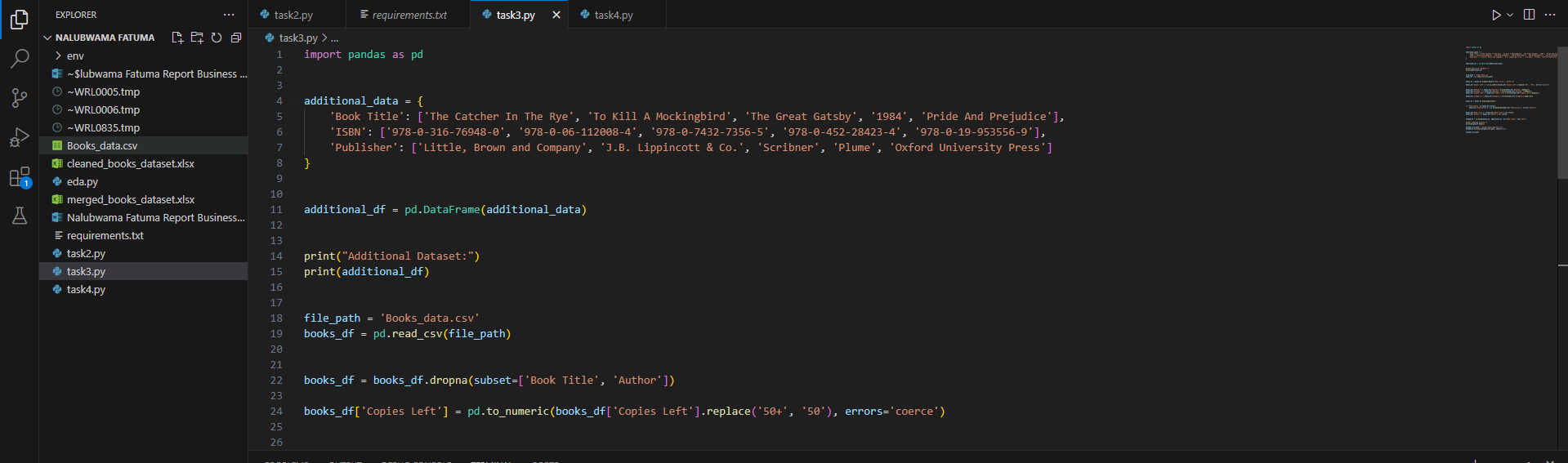
**Project Requirements**

****

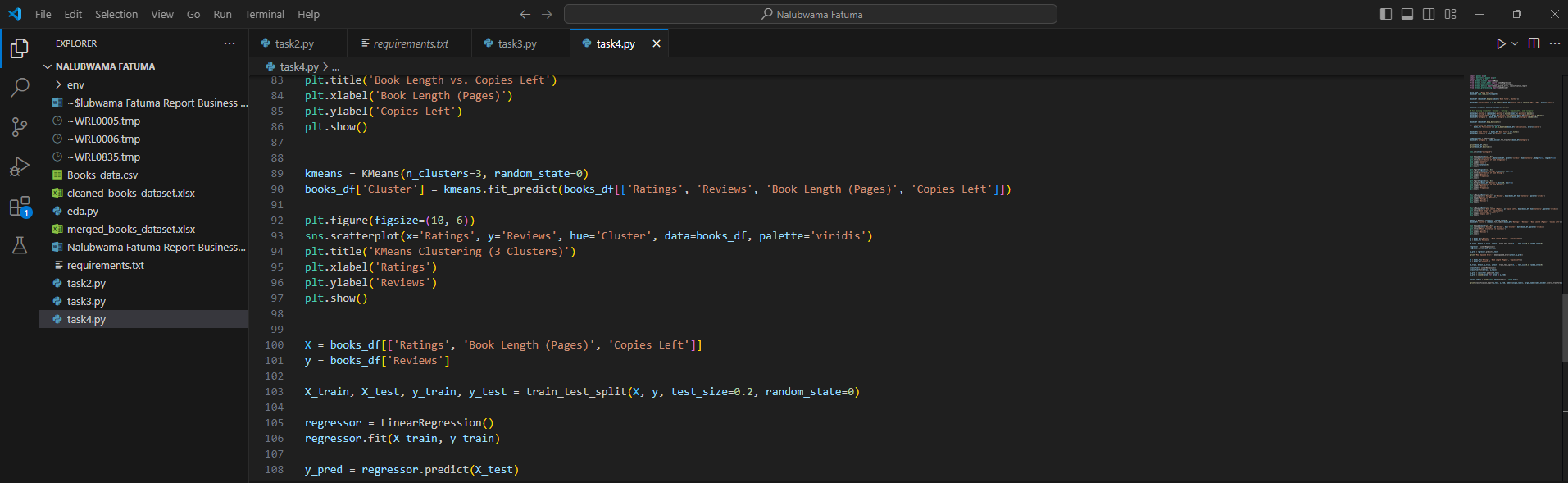
**Data cleaning**

****

**Creating additional dataset**

****

**Explanatory data analysis implementing the KMeans algorithm**

****