Building Insights from Web Data:

Web Scraping, Feature Selection, and Classification for Best-Sellers

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

### **Bachelor of Technology In**

### **Computer Science and Engineering**

### **School of Engineering and Sciences**

Submitted by

### **Geethika Sruthi Rudrapati - AP21110010285**

### **Kavya Gudiseva - AP21110010294**

### **Naga Sai Sandeep Pipalla - AP21110010317**

### **Harika Kommu - AP21110010132**

### **Sravanthi Murukonda - AP21110010144**

### **Sailesh Kollipara - AP21110010152**



Under the Guidance of

Dr.Sabyasachi Dutta

**Certificate**

Date: 09/05/2024

This is to certify that the work present in this Project entitled **“**Building Insights from Web Data: Web Scraping, Feature Selection, and Classification for Best-Sellers**”** has been carried out by **Sruthi R, Kavya G, Sandeep P, Harika K, Sravanthi M and Sailesh K** under our supervision. The work is genuine, original, and suitable for submission to the **SRM University – AP** for the award of Bachelor of Technology in **School of Engineering and Sciences.**

### **Supervisor**

(Signature)

Dr Sabyasachi Dutta

Assistant Professor, CSE Department SRM UNIVERSITY, AP.

**Acknowledgements**

We would like to acknowledge all those without whom this project would not have been successful. Firstly, we would like to thank our Professor Dr Sabyasachi Dutta who guided us throughout the project and gave his immense support. He made us understand how to successfully complete this project and without him, the project would not have been complete.

This project has been a source to learn and bring our theoretical knowledge to the real-life world. So, we would really acknowledge his help and guidance for this project.

**Abstract**

This project delves into the world of best-selling books, aiming to uncover the characteristics that contribute to their success. We employ a data-driven approach that leverages web scraping, feature engineering, and machine learning techniques.

The initial stage involves web scraping data from a relevant source, likely the Wikipedia list of best-selling books. To identify patterns and potential clusters within the data, we plan to explore both classification and clustering algorithms. Classification, assuming this is the primary task, may involve feature selection techniques like filter methods and wrapper methods to choose the most informative features for building accurate models. Clustering algorithms could shed light on groupings of best-sellers based on shared characteristics.

**Contents**

1. Introduction
2. System Requirements
3. Methodology
4. Implementation and Results
5. Conclusion

**1. Introduction**

We embark on this journey by ethically scraping data from reputable online sources (likely the Wikipedia list of best-selling books). This raw data serves as the foundation for our exploration. In the feature engineering stage, we clean and refine the data, extracting meaningful insights.This may involve creating new features based on genre, publication date, keywords,etc.

Two key approaches will be explored: classification and clustering. Classification, assuming it's our primary focus, aims to categorize books based on their characteristics. This might involve feature selection techniques like filter methods to choose the most informative features for building accurate models. On the other hand, clustering algorithms can reveal groups of best-sellers that share similar characteristics, offering a different lens for understanding literary success.

Wrapper methods, on the other hand, take a more collaborative approach. They essentially "wrap" around the classification model and iteratively evaluate different feature subsets by training the model with each subset and measuring its performance. The wrapper method ultimately selects the feature subset that yields the best performance on the chosen evaluation metric. While wrapper methods can potentially lead to superior feature selection, they come at the cost of higher computational demands due to the repeated model training involved. We'll carefully consider these trade-offs when selecting the most appropriate feature selection technique for our classification task.

**2. System Requirements**

Software Requirements:

1. Operating System:

Recommended: Windows 10, macOS, or a Linux distribution (e.g., Ubuntu, CentOS)

Ensure compatibility with required software libraries and tools

2. Python:

Version: Python 3.x (e.g., Python 3.7, Python 3.8)

3. Integrated Development Environment (IDE):

Recommended: Jupyter Notebook, JupyterLab, or Visual Studio Code with Python extensions

Hardware Requirements:

1. Processor:

Minimum: Dual-core processor (e.g., Intel Core i3 or equivalent)

Recommended: Quad-core processor or higher (e.g., Intel Core i5/i7 or equivalent)

2. Memory (RAM):

Minimum: 8 GB RAM

Recommended: 16 GB RAM or higher for handling large datasets and complex computations

3. Storage:

Sufficient free disk space for storing datasets, Python scripts, and generated outputs

SSD recommended for faster data access and computation speed

**3. Methodology**

**3.1 Description about the dataset**

The dataset likely holds information about best-selling books, scraped from a Wikipedia page. Each entry represents a book with details like title, author, original language, publication year, estimated sales , and genre . The data may require further processing before analysis. This dataset could be valuable for tasks like classifying books by genre or era, or uncovering clusters of best-sellers with shared characteristics.

**3.2 Programming Language**

The entire project was executed using the Python programming language, known for its simplicity and versatility. Python's extensive ecosystem of libraries facilitated various project phases, from data processing to analysis and visualization. Its intuitive syntax allowed for rapid development and easy debugging, streamlining the implementation process.

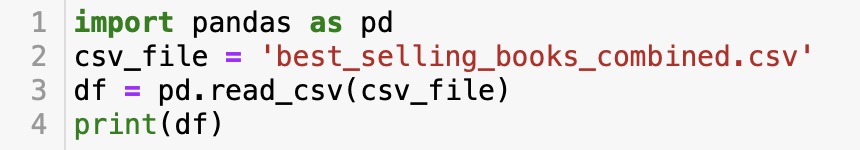
**3.3 Libraries Used**

* **pandas:** Used for data manipulation and creating the DataFrame.
* **requests:** Used to send HTTP GET requests to the target URL.
* **BeautifulSoup:** Used to parse the HTML content of the webpage.
* **seaborn (sns):** This library provides a high-level interface for creating statistical graphics based on Matplotlib.
* **matplotlib.pyplot (plt):** It offers various functionalities for creating different plot types.
* **sklearn.feature\_selection:** provides functionalities for selecting informative features from datasets.

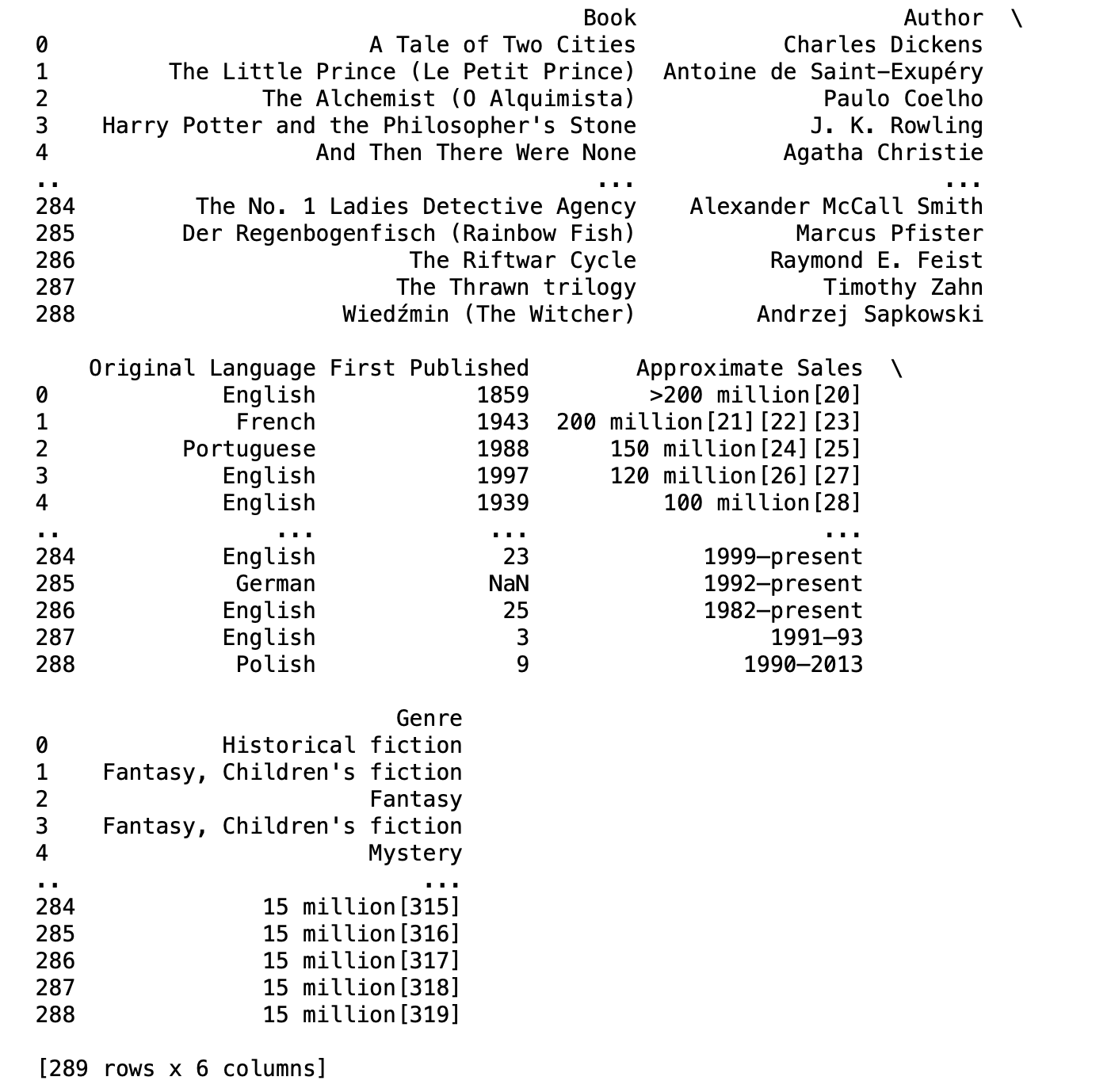
**4. Implementation And Results**

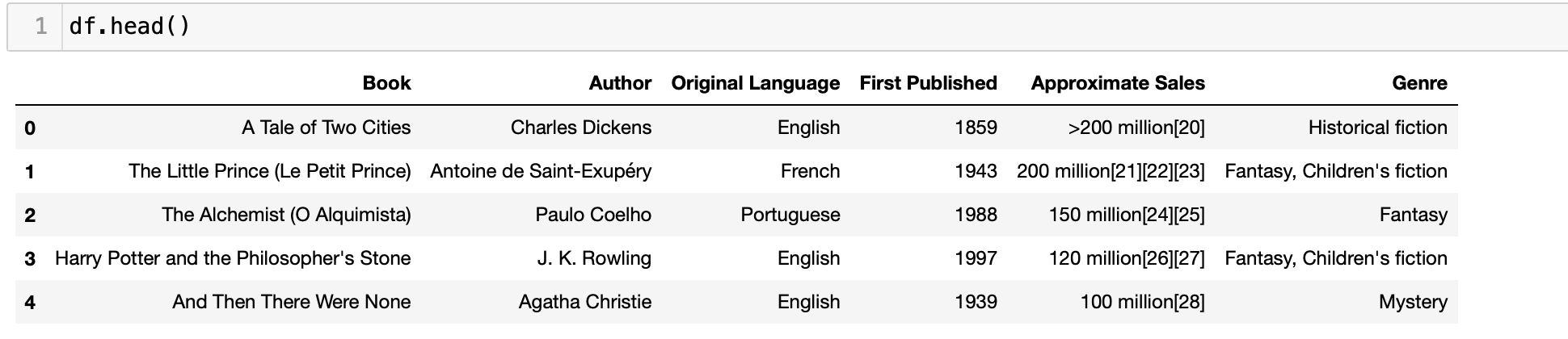
**Web Scrapping**

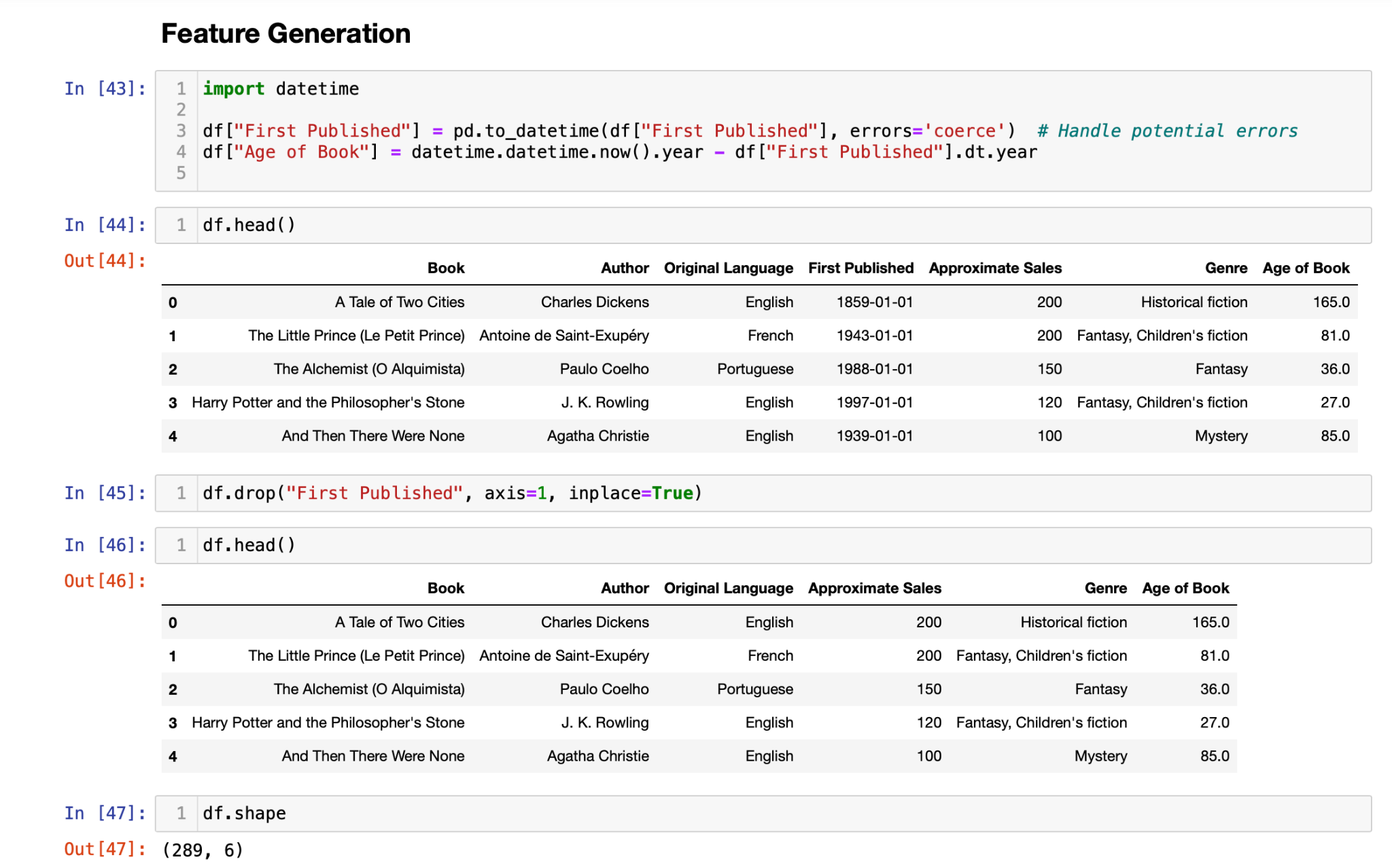




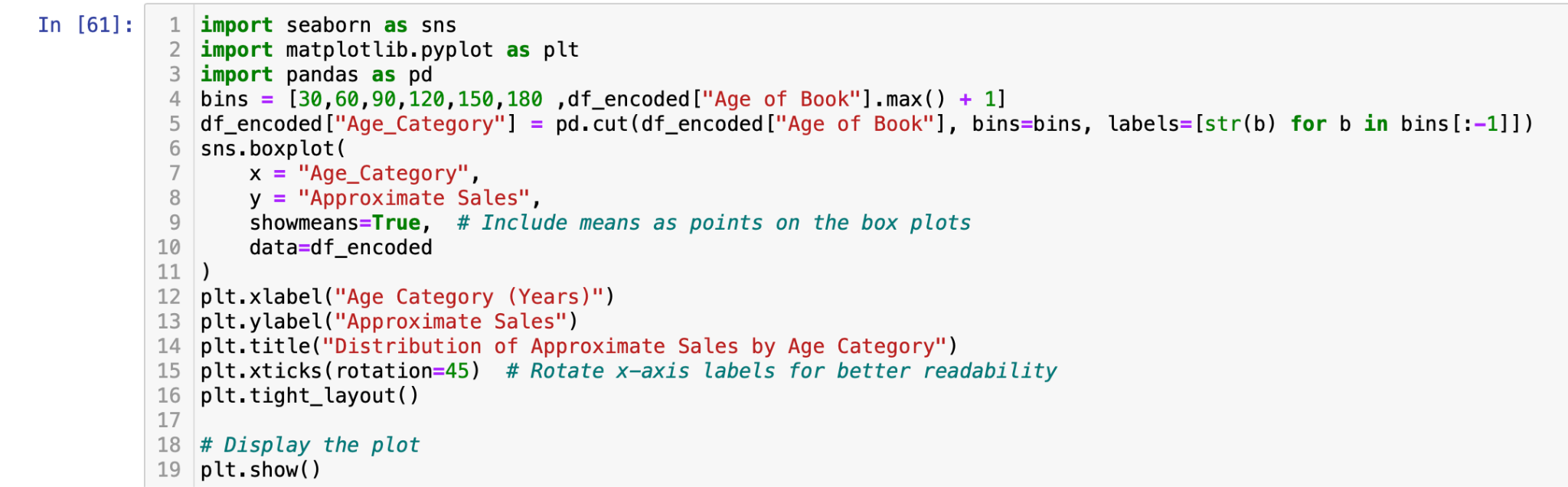
**Output**



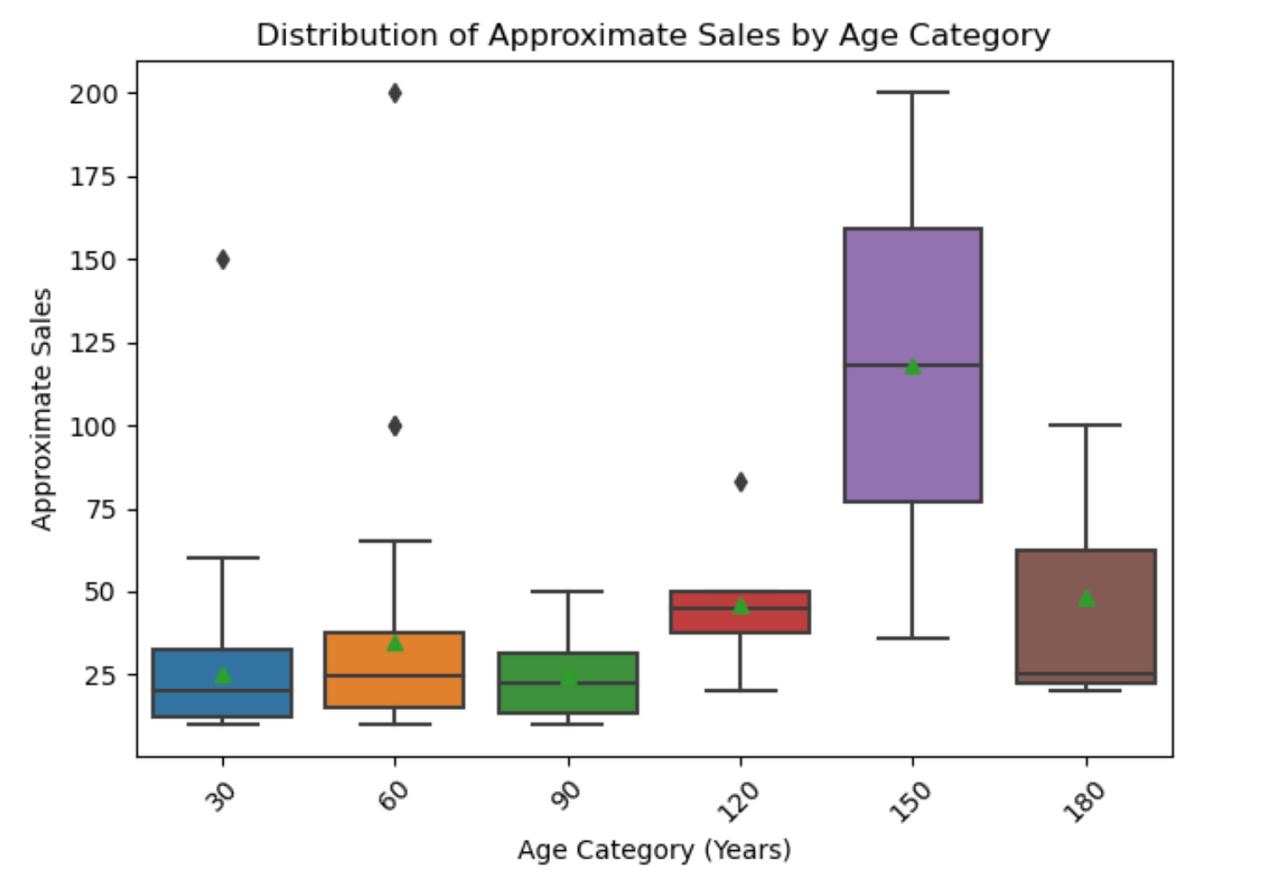




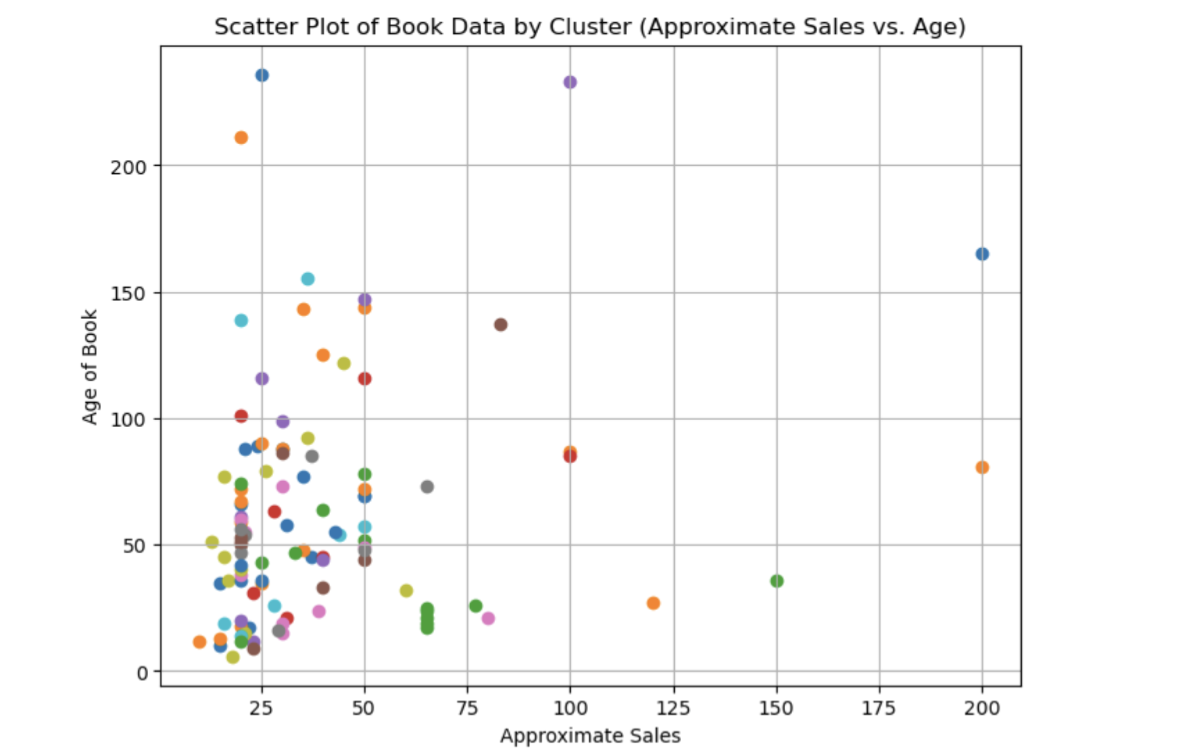
**Box plot : Approximate sales by age category**

****

**Output**

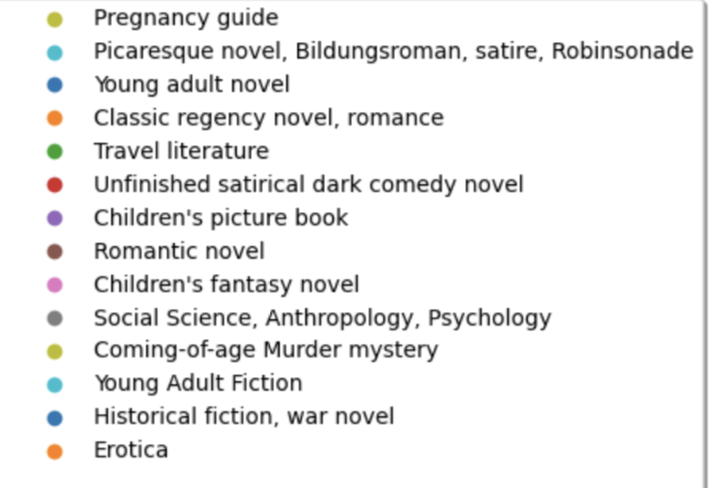
****

****

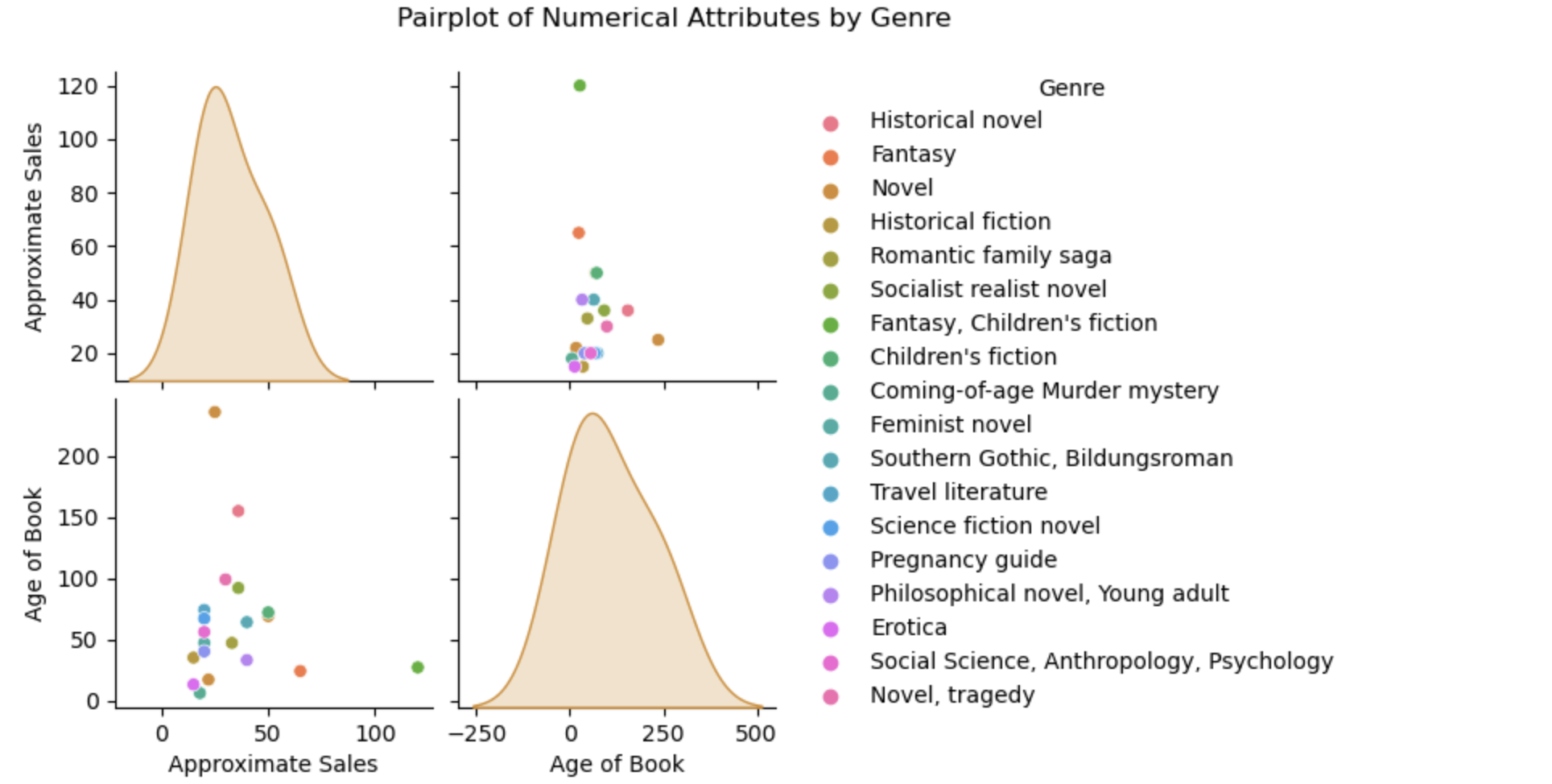
****

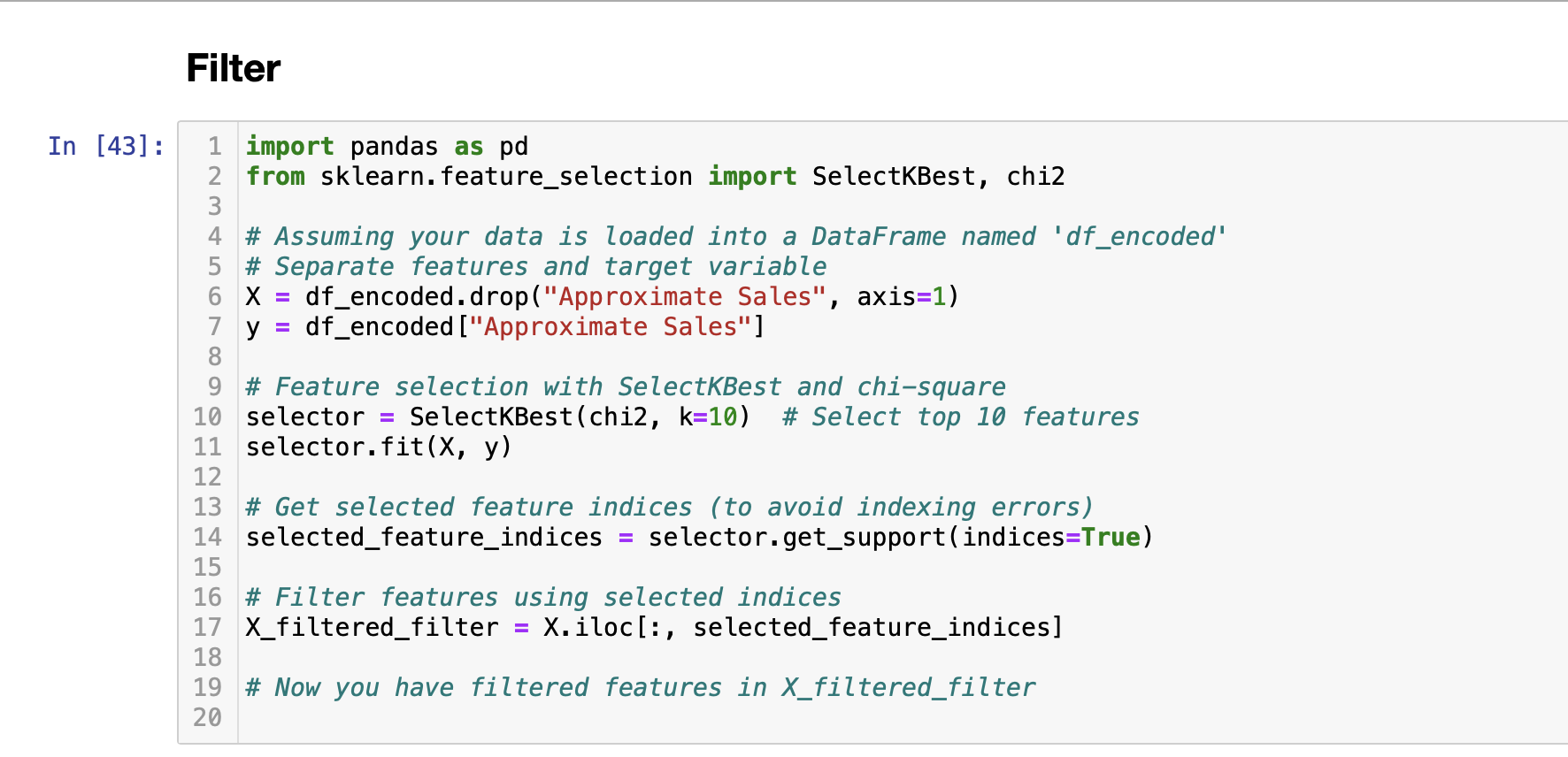
****

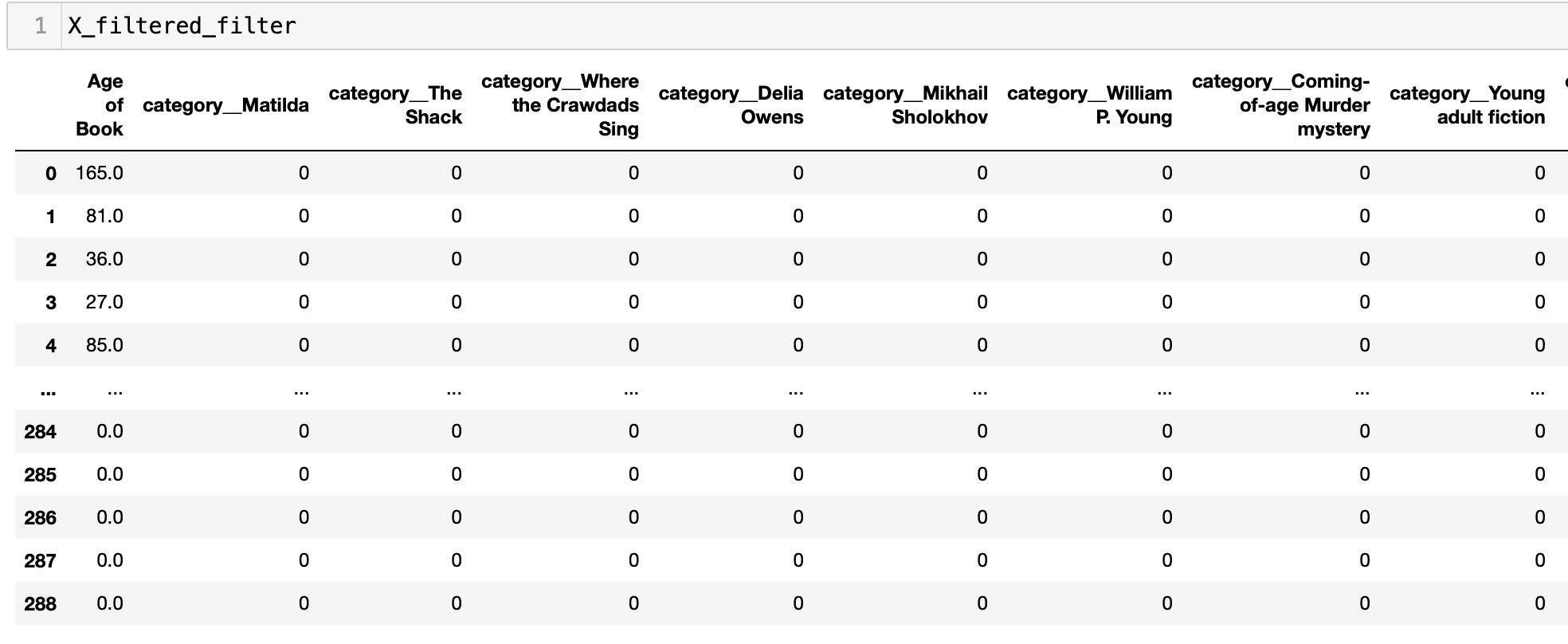
****

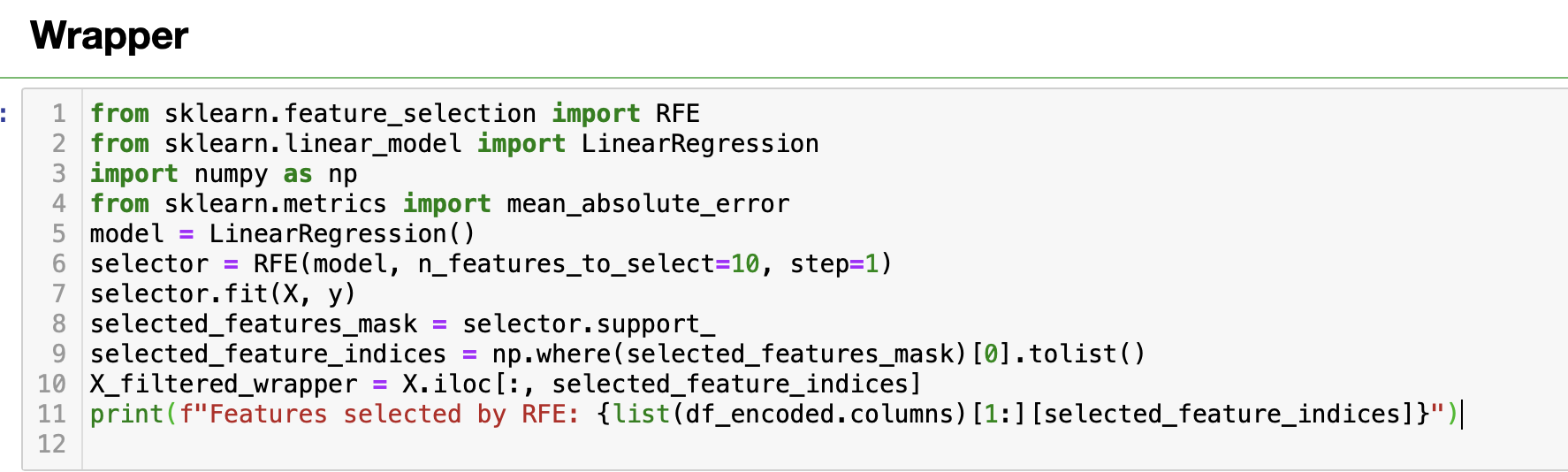
****

****

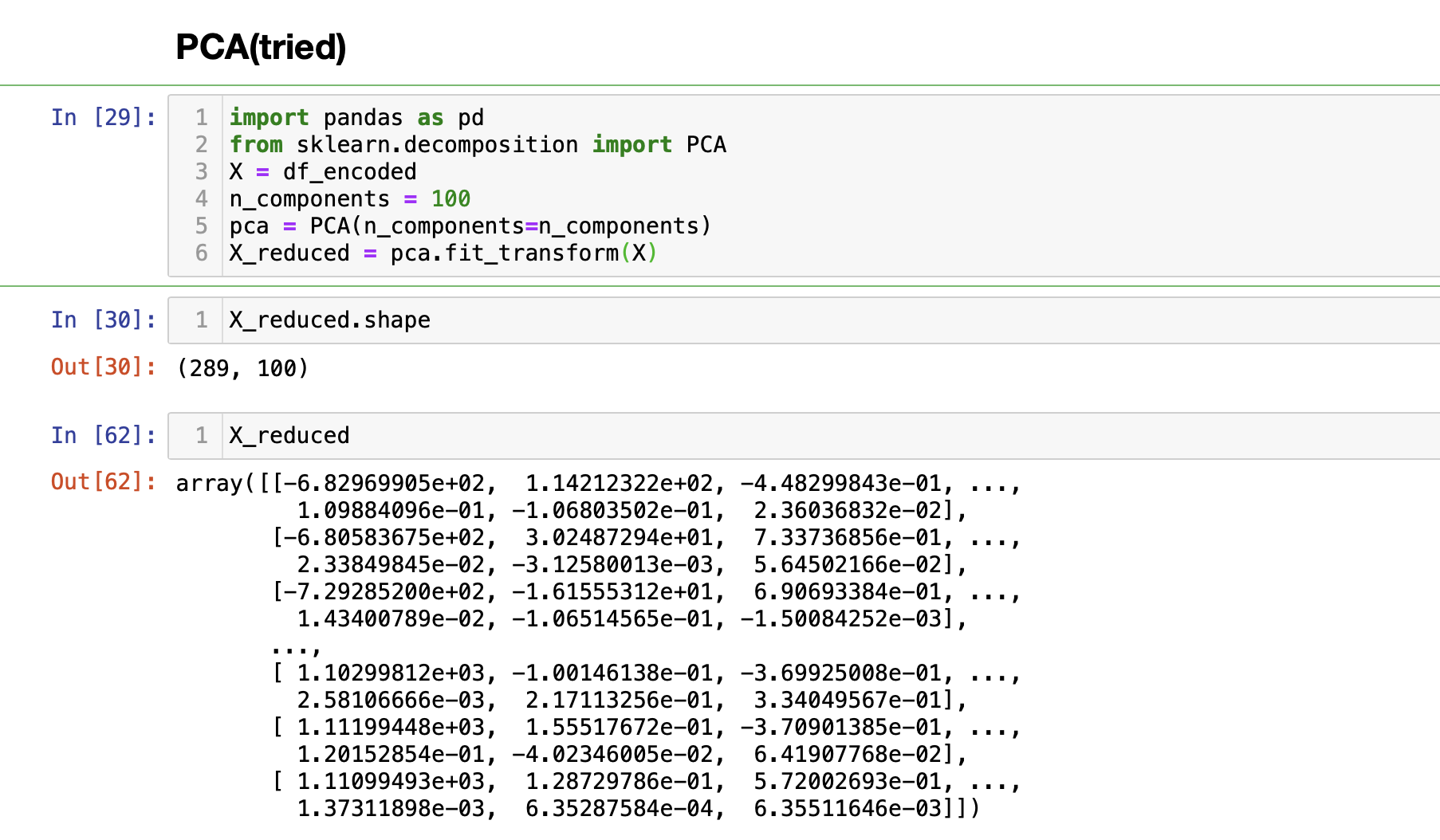
****

****

****

****

****

****

Wrapper can’t be implemented because there are too few numerical values, and when one-hot encoding is done it got increased to over 700 columns, so we have used PCA as for dimensionality reduction, it came down to around 100 columns. But as the output shows, the values were converted into numpy arrays and to implement wrapper we need a target variable which isn’t there.

These are our insights of our project.

**5. Conclusion**

The knowledge gleaned from this project has the potential to empower various stakeholders within the literary world. Publishers could utilize these insights to make informed decisions about book acquisition and marketing strategies. Marketers, armed with this knowledge, could tailor their approaches to reach the right audience for specific genres or themes. Even readers themselves can benefit from this project's findings. By understanding the characteristics of best-sellers, readers can explore new genres or authors that align with their preferences, enriching their literary journey.

In conclusion, this project has taken a significant step towards unlocking the secrets of best-selling books. The knowledge gleaned paves the way for further exploration and innovation within the literary landscape. By fostering collaboration between data science and the world of books, we can gain a deeper understanding of reader preferences and the intricate factors that contribute to a book's success.