603 BIG DATA PROJECT

Book recommendation system

By:
Group 6

Soumitra Joshi
Shanmukha Surya Sriteja
Soniya koduru

Introduction and Course relevance

Introduction :

- A recommendation engine is a class of machine learning which offers relevant suggestions to the customer. Before the recommendation system, the major tendency to buy was to take a suggestion from friends. But Now Google knows what news you will read, Youtube knows what type of videos you will watch based on your search history, watch history, or purchase history.
- A recommendation system enhances customer loyalty by suggesting products based on their preferences and trends, even for new visitors, helping businesses increase sales and profits.

Course Relevance:

In our project, we aim to utilize big data techniques to recommend books to users based on their preferences, using the Amazon Books Review dataset from Kaggle. This data is a bigdata which is of around 3GB size and the tools used will be Hadoop , Apache Spark , and other visualization tools.

Objective

Develop a book recommendation system which suggests the nearest 5 books to users based on their preferences and reviews.

Specific Problems/Questions:

- How can we effectively recommend books to users based on their preferences and reviews?
- What is the best way to implement collaborative filtering with k-nearest neighbors for book recommendations?
- How can we evaluate the performance of the recommendation system and ensure its accuracy?
- What challenges are involved in processing and analyzing the Amazon Books Review dataset, and how can they be overcome?

Data Sources and collection

Data Sources:

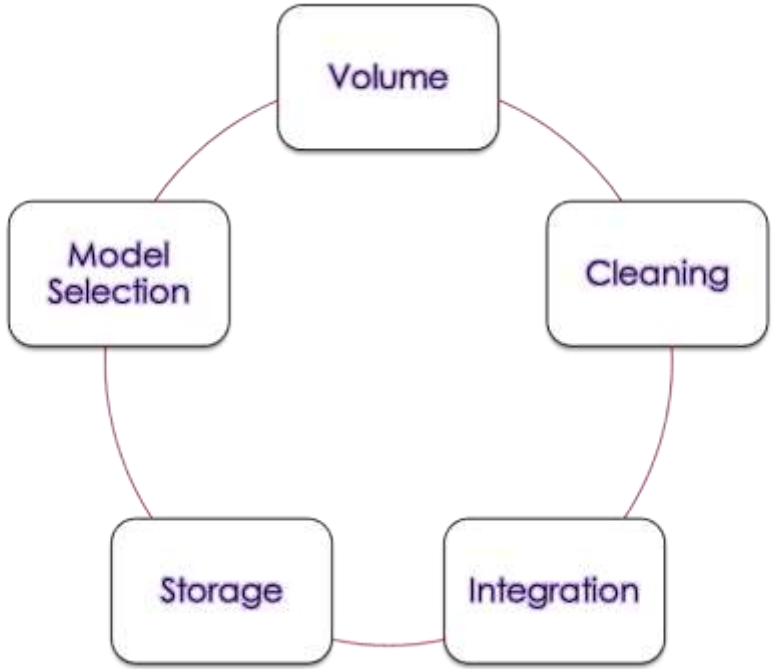
- Our primary source of data was Kaggle, a renowned platform for datasets spanning various domains.
- Specifically, we accessed the "Amazon book reviews" dataset hosted on Kaggle.
- This dataset offered a wealth of information about books, encompassing details such as titles, authors, categories, user ratings, prices, and more, making it an invaluable resource for our project.

Collection Process:

- We obtained the dataset by downloading it directly from Kaggle's website, leveraging their user-friendly interface.
- The dataset, totaling around 3 GB in size, comprised two CSV files, each containing structured data essential for our analysis.
- Accessing and acquiring the dataset from Kaggle was a straightforward process, facilitated by the platform's accessibility and comprehensive documentation.

Challenges

- Managing the substantial size of the dataset presented challenges in terms of storage, processing, and computational resources.
- Ensuring data quality and consistency across the multiple CSV files demanded meticulous attention to detail during the collection phase.
- Despite these challenges, we navigated through the collection process adeptly, ultimately securing the necessary data for our analysis and modeling endeavors.



Tools and Technologies

Big Data Platforms: Apache Hadoop, Apache Spark, Plotly and Matplotlib for visualization

- **Apache Hadoop:**
  - Apache Hadoop serves as the backbone of our big data infrastructure, providing distributed storage and processing capabilities.
  - Its distributed file system (HDFS) allows us to store vast amounts of data across multiple nodes, while its MapReduce framework facilitates parallel processing, enabling efficient analysis of large datasets.
- **Apache Spark:**
  - Apache Spark is a powerful data processing engine that utilizes an in-memory computation model, making it highly suitable for processing and analyzing big data.
  - Its ability to perform computations in memory significantly accelerates processing speeds, leading to faster analytics and insights.
  - Spark's versatile APIs support a wide range of data processing tasks, including batch processing, streaming analytics, machine learning, and graph processing, making it a versatile tool for various use cases.

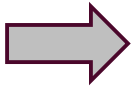
Tools and Technologies

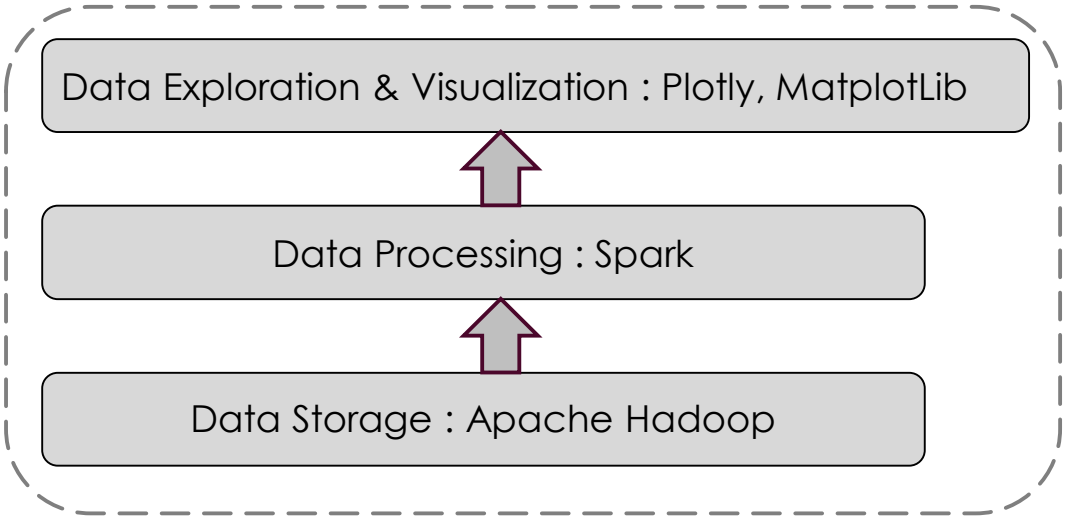
Plotly or Matplotlib :

- Plotly and Matplotlib are both powerful data visualization libraries in Python, offering a wide range of chart types and customization options.
- Plotly excels in interactive visualizations, allowing users to create dynamic plots and dashboards that can be explored and shared.
- Matplotlib is known for its flexibility and ability to create static, animated, and interactive visualizations with detailed customization options.
- Both libraries are well-suited for handling large datasets and can be integrated seamlessly into Python data analysis workflows.
- The choice between Plotly and Matplotlib often depends on the specific requirements of the visualization task and the user's preferences for interactivity and customization.

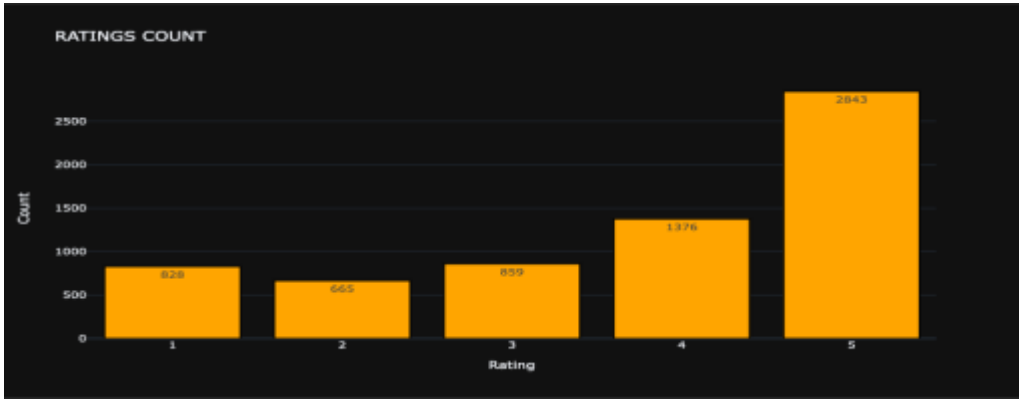
Data Stack Diagram

Data Source:
- Kaggle

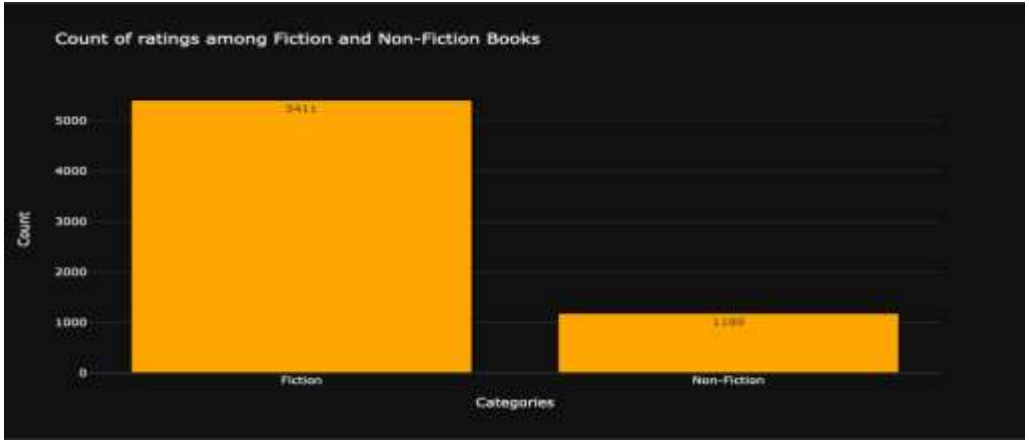


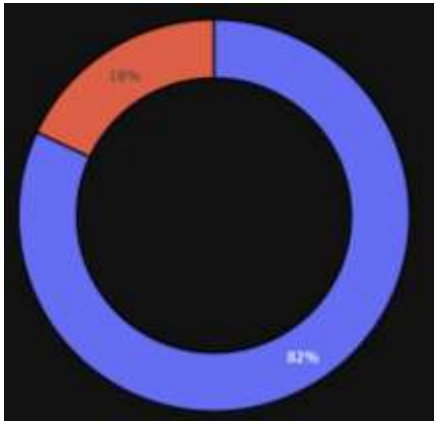


Reports and Insights Generation



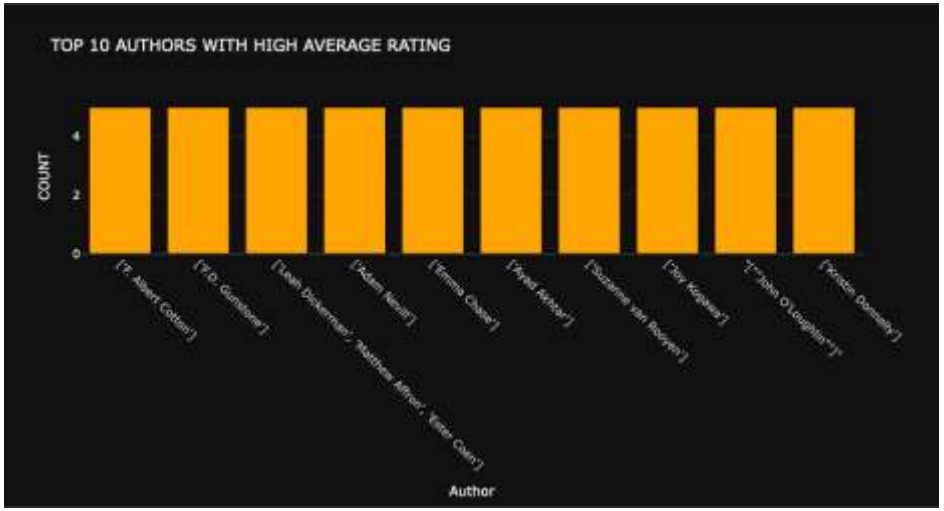
5
4
3
2
1

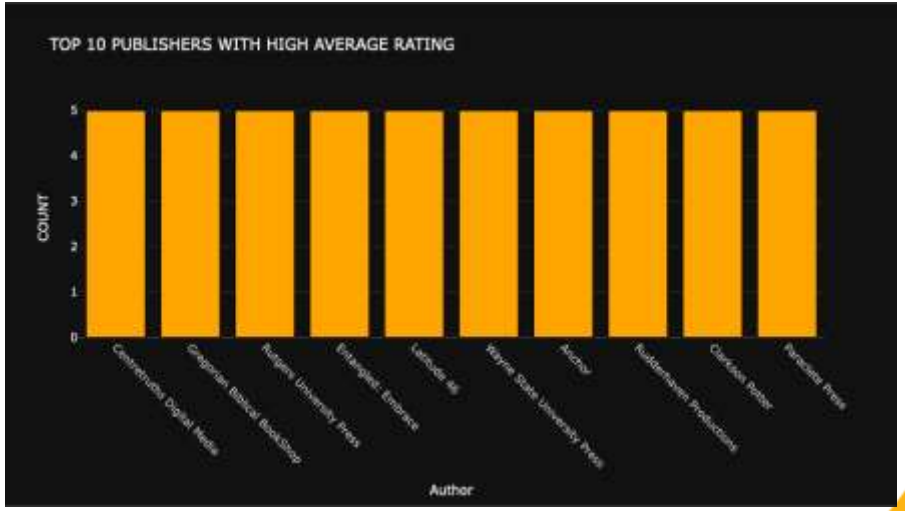




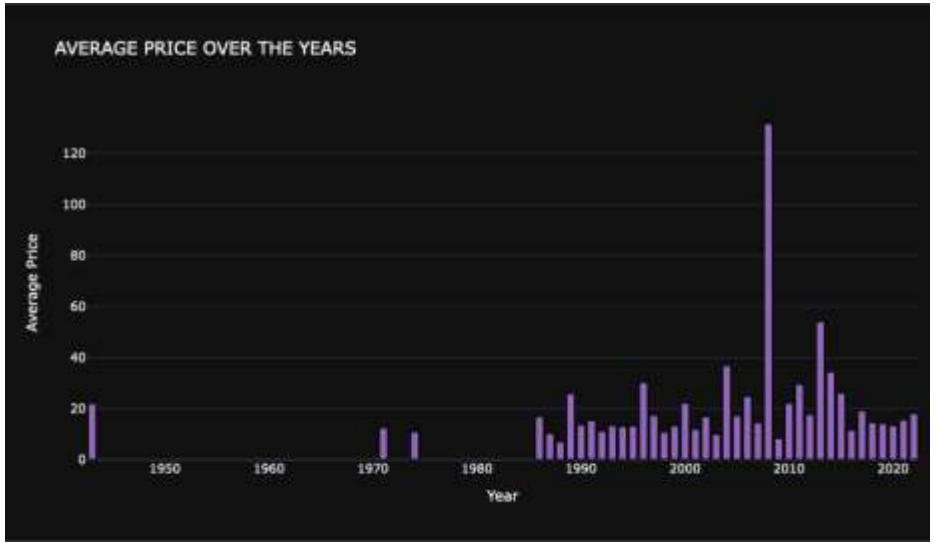
Fiction
Non-Fiction

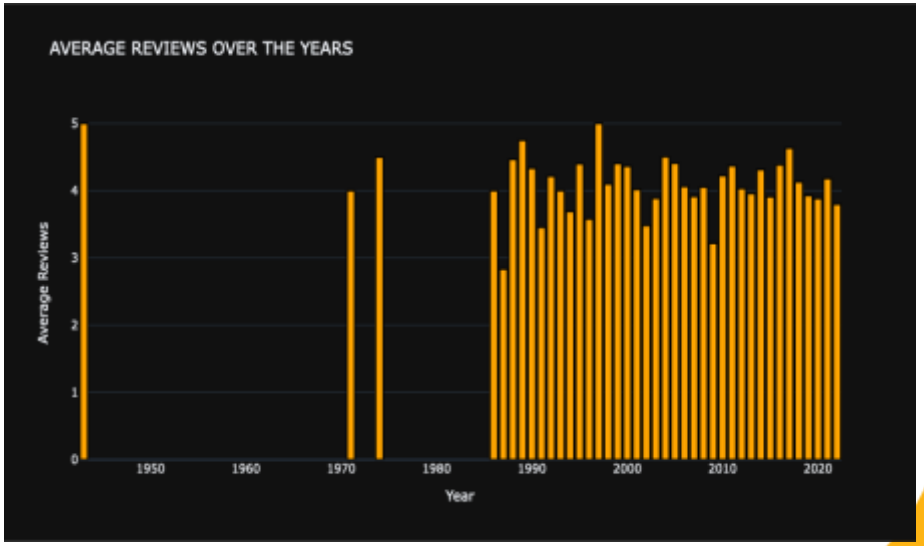
Reports and Insights Generation



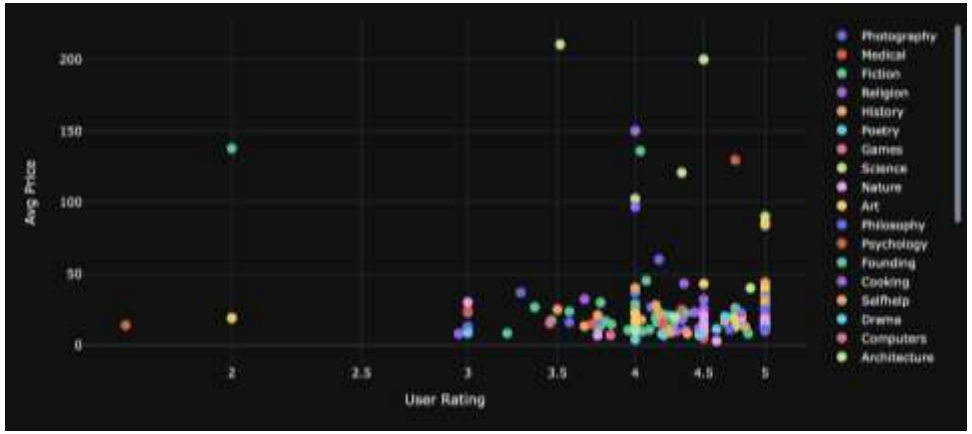


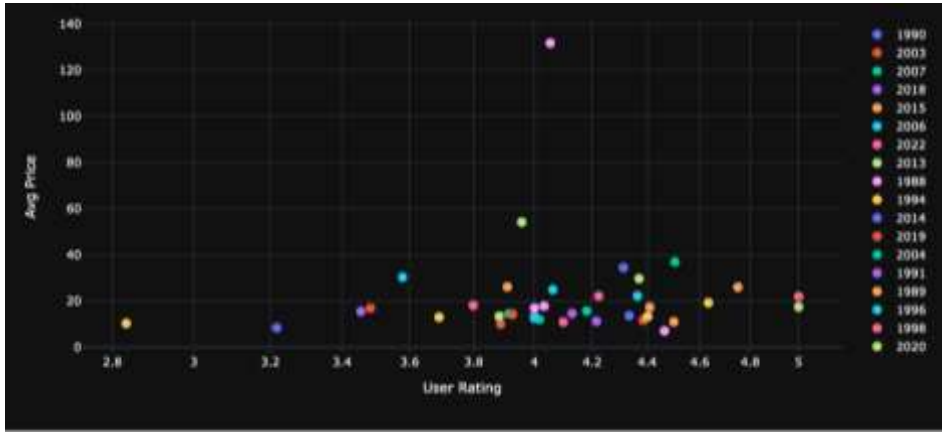
Reports and Insights Generation





User Rating & Avg price for different categories





User Rating & Avg price for different Years

Recommendation system suggesting nearest 5 books

- ✓ We used knn (nearest neighbors) model to develop the recommendation system using collaborative filtering
- ✓ Bucketed Random projection (RP) Locality Sensitive Hashing (LSH): For dimensionality reduction of the data and grouping similar points into the same buckets using hash functions.

Input

Enter the title of the book:

```
# # Provide input book title from the user
user_input_title = input("Enter the title of the book: ")

# Capitalize each word in the input title
user_input_title = user_input_title.title()
```

Enter the title of the book: Bite

Output

List of 5 books

Jumpmetrics
Dreamspy
Scarab-4
Kunma
Sensation

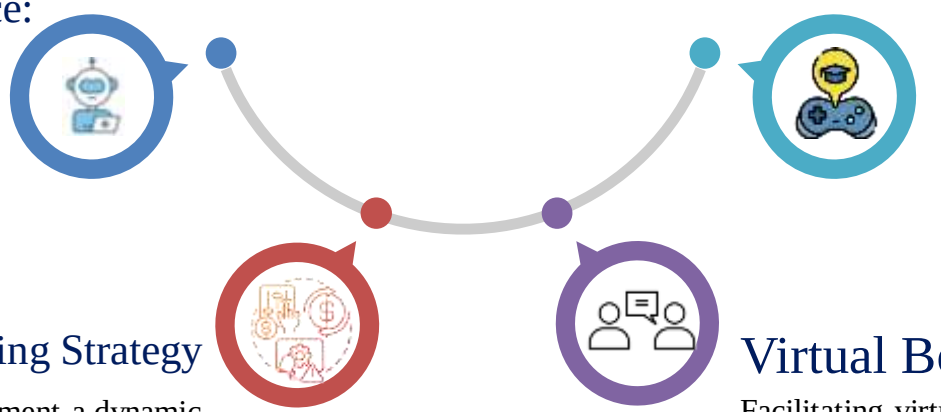
Reports and Insights Generation

Users	Online Retailers	Offline Retailers
➢ **Personalized Recommendations:** Users can receive recommendations for books similar to the ones they have rated highly, enhancing their reading experience. ➢ **Discover New Books:** Users can explore books in categories they might not have considered before, based on the recommendations. ➢ **Enhanced User Experience:** With relevant recommendations, users are more likely to find books that interest them, improving their overall satisfaction.	➢ **Customer Engagement:** Online retailers can use personalized recommendations to engage customers and increase the likelihood of purchase. ➢ **Inventory Management:** Retailers can optimize their inventory by stocking books that are frequently recommended or similar to popular books. ➢ **Targeted Marketing:** Retailers can use the data to create targeted marketing campaigns based on user preferences and book categories.	➢ **In-Store Promotions:** Offline retailers can promote recommended books in-store based on popular categories or similar books. ➢ **Customer Loyalty:** By offering personalized recommendations, offline retailers can enhance customer loyalty and increase repeat business. ➢ **Stocking Strategy:** Retailers can use the data to determine which books to stock based on user preferences and popular categories.

Unique Suggestions to Stakeholders

AI-Driven Customer Service:

Use AI-driven chatbots or virtual assistants to provide personalized customer service, assist with book recommendations, and handle inquiries, improving overall customer satisfaction.



Gamification

Introduce gamification elements to the platform, such as badges or rewards for exploring new genres, rating books, or participating in book clubs. This can increase user engagement and loyalty.

Dynamic Pricing Strategy

Utilize the data to implement a dynamic pricing strategy, where prices of books are adjusted based on demand, user ratings, and other relevant factors. This can help optimize revenue and maximize profitability.

Virtual Book Clubs

Facilitating virtual book clubs where readers can discuss books in real-time, interact with authors, and participate in exclusive events, creating a sense of community and engagement.

Future Opportunities and Challenges

- **Integration with Emerging Technologies:** Integrating data from multiple sources, such as social media, reading habits, and purchase history, and utilizing techniques like NLP and deep learning can provide richer insights for better recommendations.
- **Real-Time Recommendations:** Implementation of real-time recommendation systems with technologies such as Apache Flink or Kafka Streams for instant recommendations based on user interactions.
- **Augmented Reality (AR) and Virtual Reality (VR):** Using AR/VR to enable users to discover books in physical bookstores or libraries by pointing their smartphone camera at a book cover, which then displays reviews, ratings, and similar book recommendations.
- **Blockchain for Data Security:** Adoption of blockchain technology to enhance data security and transparency, ensuring trust in recommendation systems by securely storing user preferences.
- **Content Curation:** Continuously improving content curation algorithms to keep up with changing reader preferences and trends, while also promoting diverse and inclusive content.

References:

Data Source :

Mohamed Bekheet. (2022). Amazon Books Reviews. *Kaggle*. Amazon Books Reviews (kaggle.com)

Other References:

Raghav Agarwal (2021, June). Build Book Recommendation System using Unsupervised Learning [Blog post]. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2021/06/build-book-recommendation-system-unsupervised-learning-project/

Nandalal D (2021). Amazon Books EDA & Recommendation. *Kaggle*. https://www.kaggle.com/code/nandalald/amazon-books-eda-recommendation

Sohel Rana (2023, Dec). Enhancing Book Recommendations: Leveraging Deep Learning Algorithms for Personalized Suggestions. *Medium*. https://medium.com/@sohelrana.aiubPro/enhancing-book-recommendations-leveraging-deep-learning-algorithms-for-personalized-suggestions-a6f209849d36

Gupta, S., & Dave, M. (2019). A Recommendation System: Trends and Future. *International Journal of Engineering and Advanced Technology*. https://www.ijeat.org/wp-content/uploads/papers/v8i6S3/F12400986S319.pdf

Github Link: https://github.com/SoumitraJoshi7/DATA603_Project

Thankyou !!