

15 December 2023

PLAKSHA

LIBRARY RECOMMENDATION
SYSTEM

Nainika Gupta, Nandana N, Niranjani A

Machine Learning &
Pattern Recognition



Library Dataset

Plaksha Library Recommendation System

AI3011: Machine Learning and Pattern Recognition



**Data pre-processing
+
Feature engineering**

**Hybrid Model
Content Based Filtering
+
Collaborative Filtering**

Performance metrics

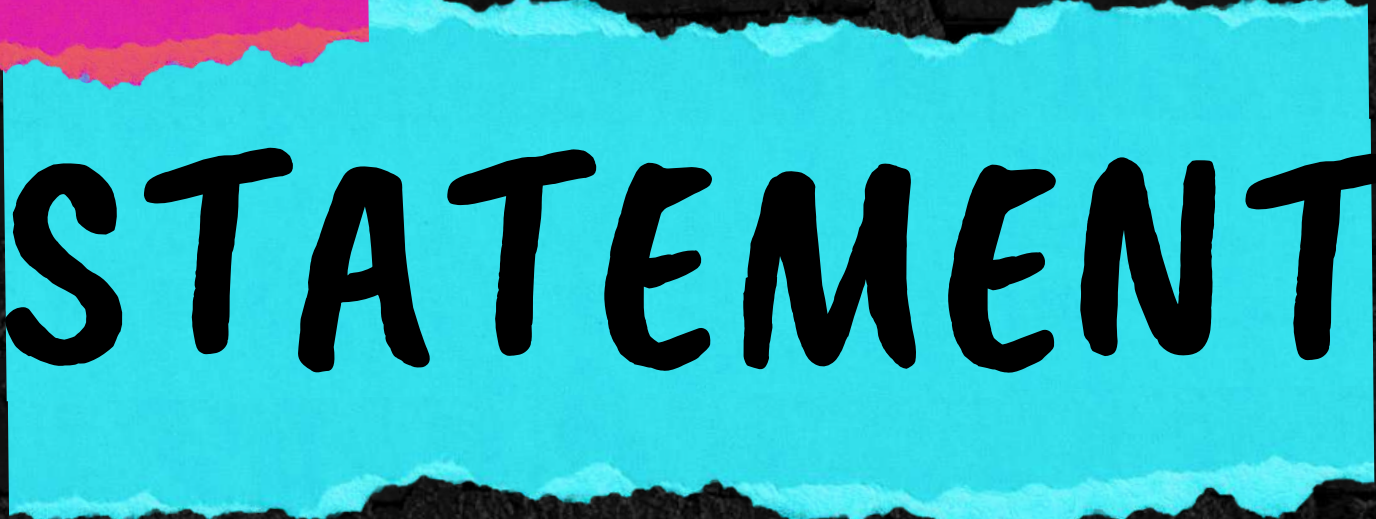


Book Recommendation

Nainika Gupta, Nandana. N, Niranjani. A



PROBLEM



STATEMENT



This project aims to develop a library recommendation system to enhance students' academic reading. It will utilize data analysis over time to suggest book purchases based on semester, subject, and individual interests, aiming to optimize the library's collection and boost students' academic experience.

Why?

- Importance of academic reading
- Purchasing new books based on recommendations and interests (library)
- Enhance student engagement
- Recommending books to students based on academic needs and individual interests

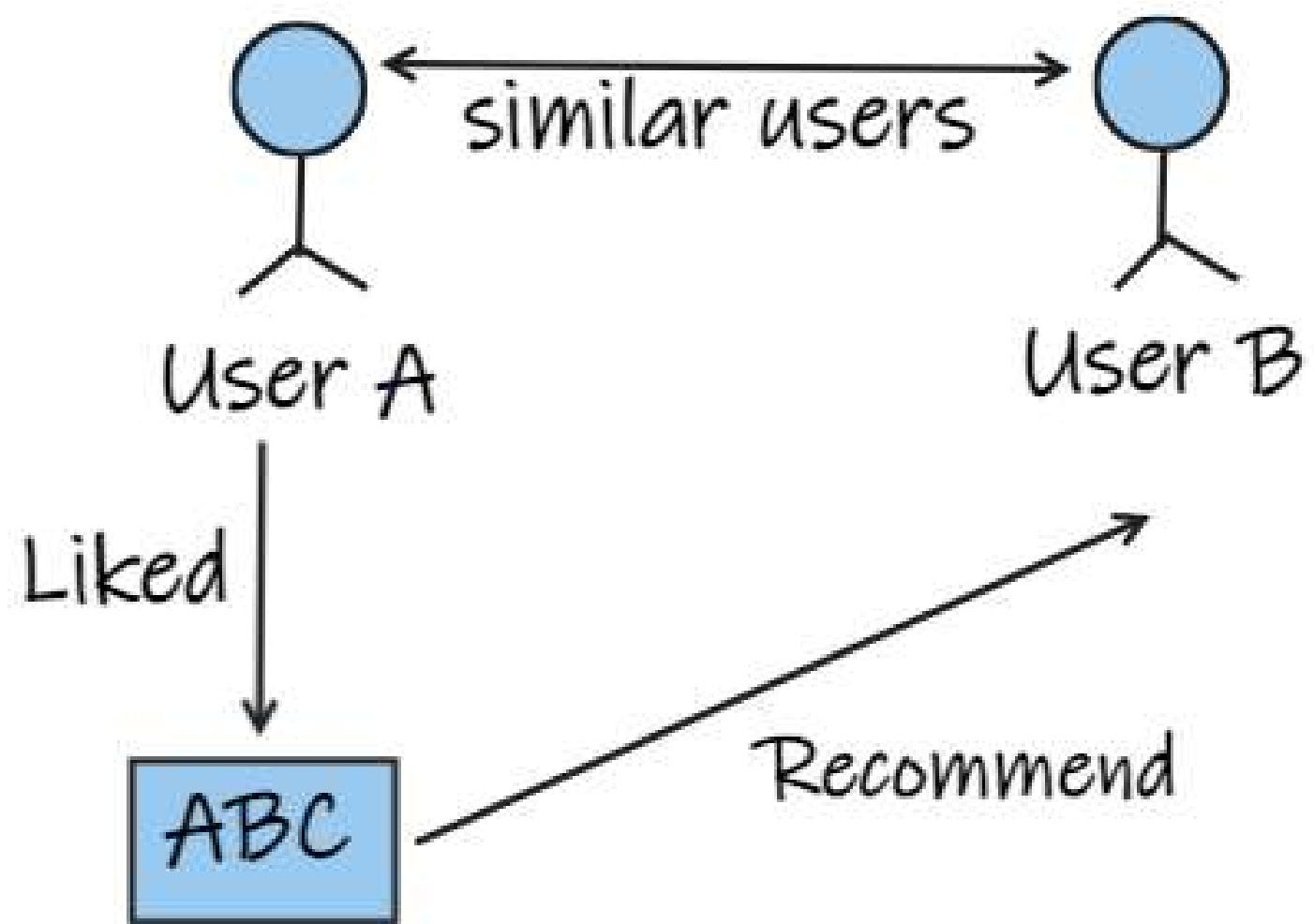
Potential impact: Facilitating informed book recommendations based on previous data, and reading preferences of others to positively impact academic performances



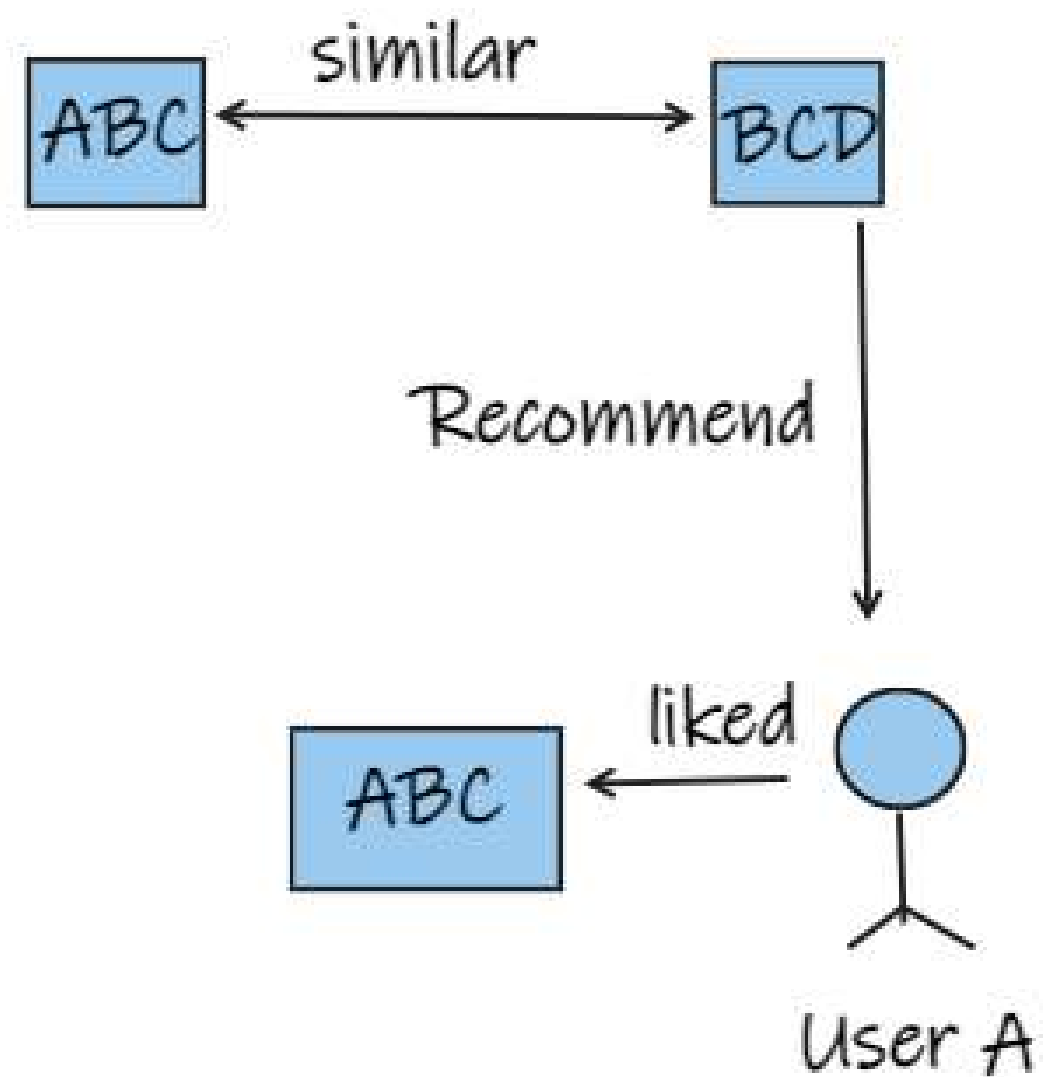
Literature Survey

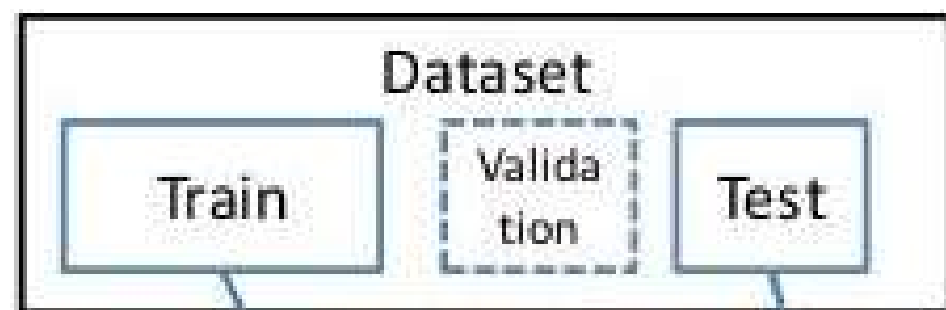


Collaborative Filtering



Content Based

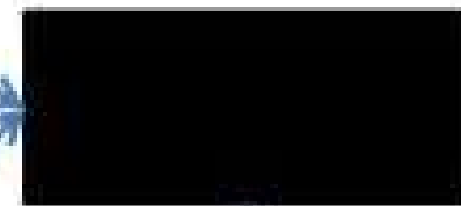




generates

a ranking
(for a user)

a prediction for a
given item (and user)



precision
error
coverage

...



DATASET &

**FEATURES
PREPROCESSING**

DATASET

**Plaksha
library
dataset**

WEBSCRAPING

**Genre &
Ratings**

**MISSING & NAN
VALUES**

**Feature
engineering &
pre-
processing**



Dataset

- Requested the Library Committee for the library data
- Drafted a data privacy statement signed by the team-members
- Forwarded it to the Office of Research and Library Committee
- Received data without any personal identifier (string manipulation)

SIGNED DOCUMENT

DATA PRIVACY
STATEMENT

TEAM

Possible Ethical and Privacy Issues

- User Identification: Even without names, certain combinations of data (e.g., borrowing history and department) might make it possible to identify specific individuals.
- Reading preference: Knowing a person's borrowing history can reveal sensitive information such as their personal interest, politics, etc.
- Data accessibility: Data breaches or unauthorized access can compromise user privacy.
- Privacy Statement: Drafted a data privacy statement signed by the team-members, approved by Library Committee and the Office of Research
- Data Security: We promise to keep the data confidential and only use it for the purpose of the project. We will not discuss the contents with anyone outside of the team.
- Data Retention Policy: We promise to erase the data from our personal devices after the access period.

We have access to a tabular dataset that has been diligently maintained by the library staff from December 2021 to September 2023. This dataset encompasses records of books issued and returned by all the batches of UG & TLP students, featuring a total of 7,292 data points and 8 unique features.

Date of Issue & Return

Batch

Title of the Book

Barcode

User id

Author's First Name

Transaction: Check in/Check out

Author's Last Name



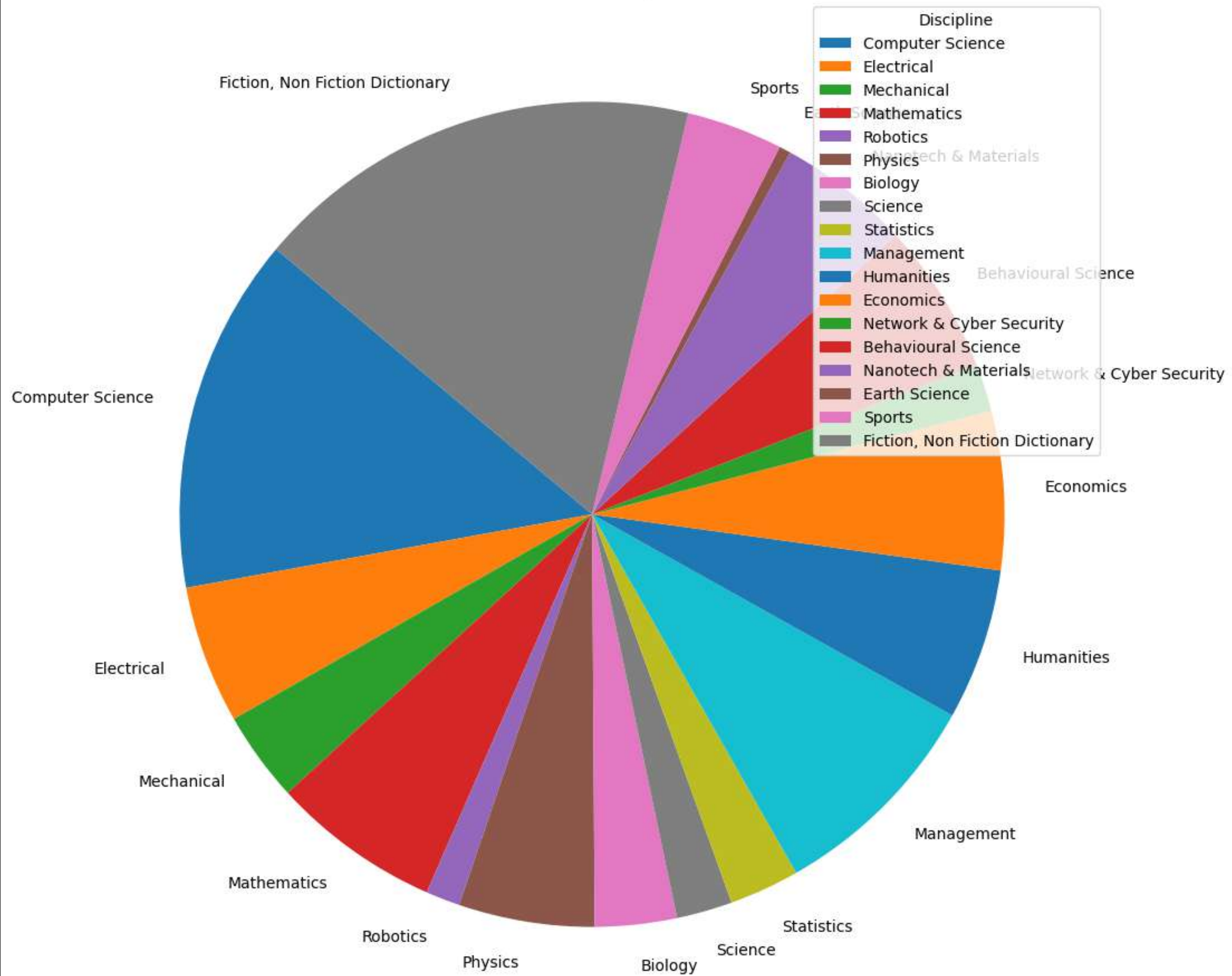
- Webscraping - openlibrary & google apis
- Missing ratings - Drawn from normal distribution
- Removed Nan values

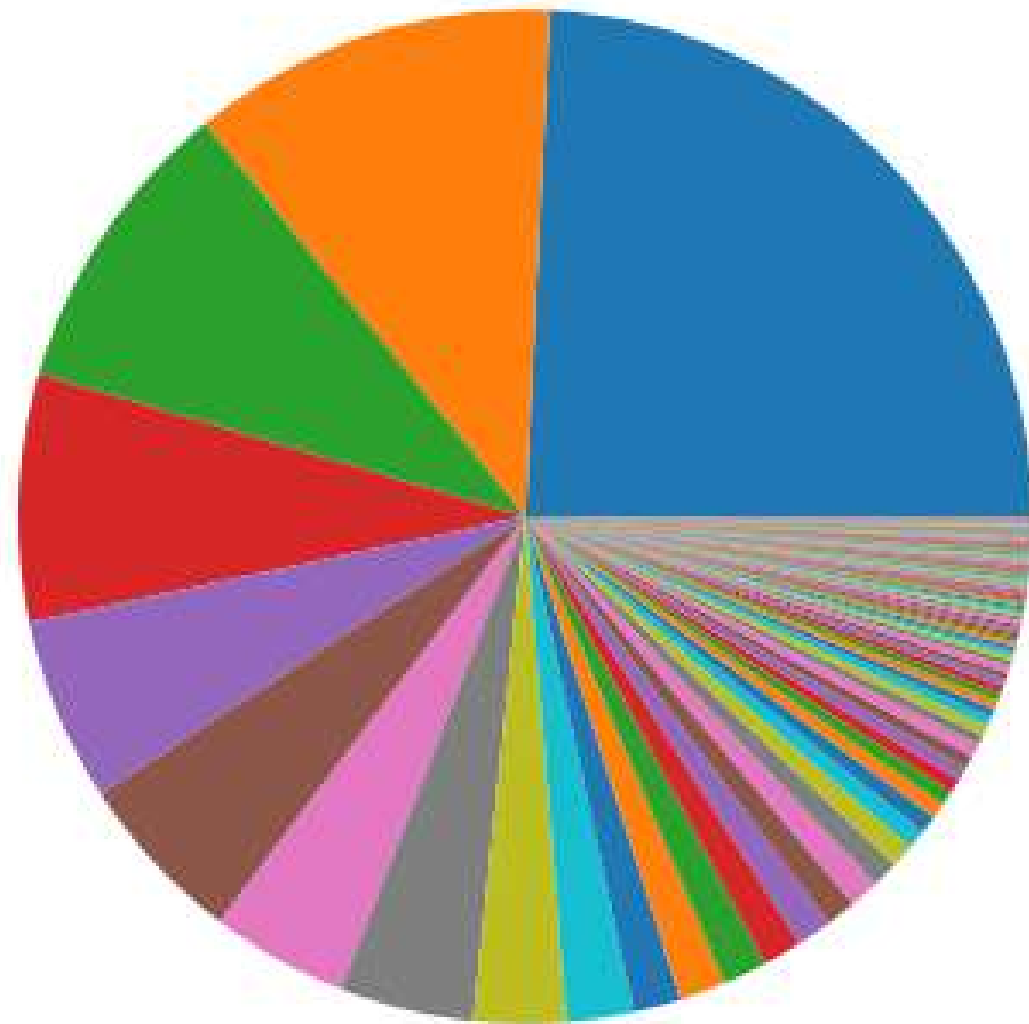
7293 -> 5229



Visualization

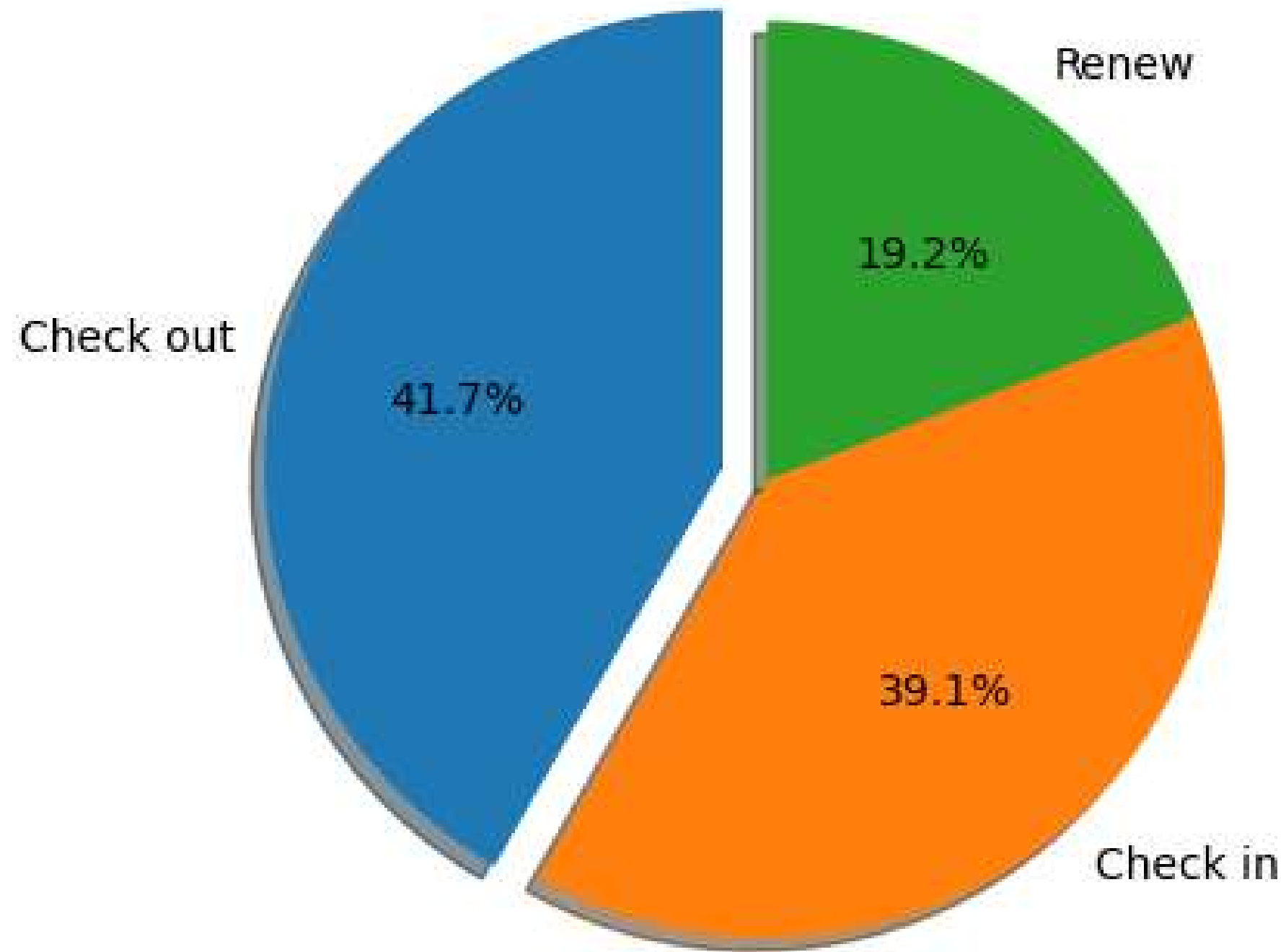
Pie Chart of Categories



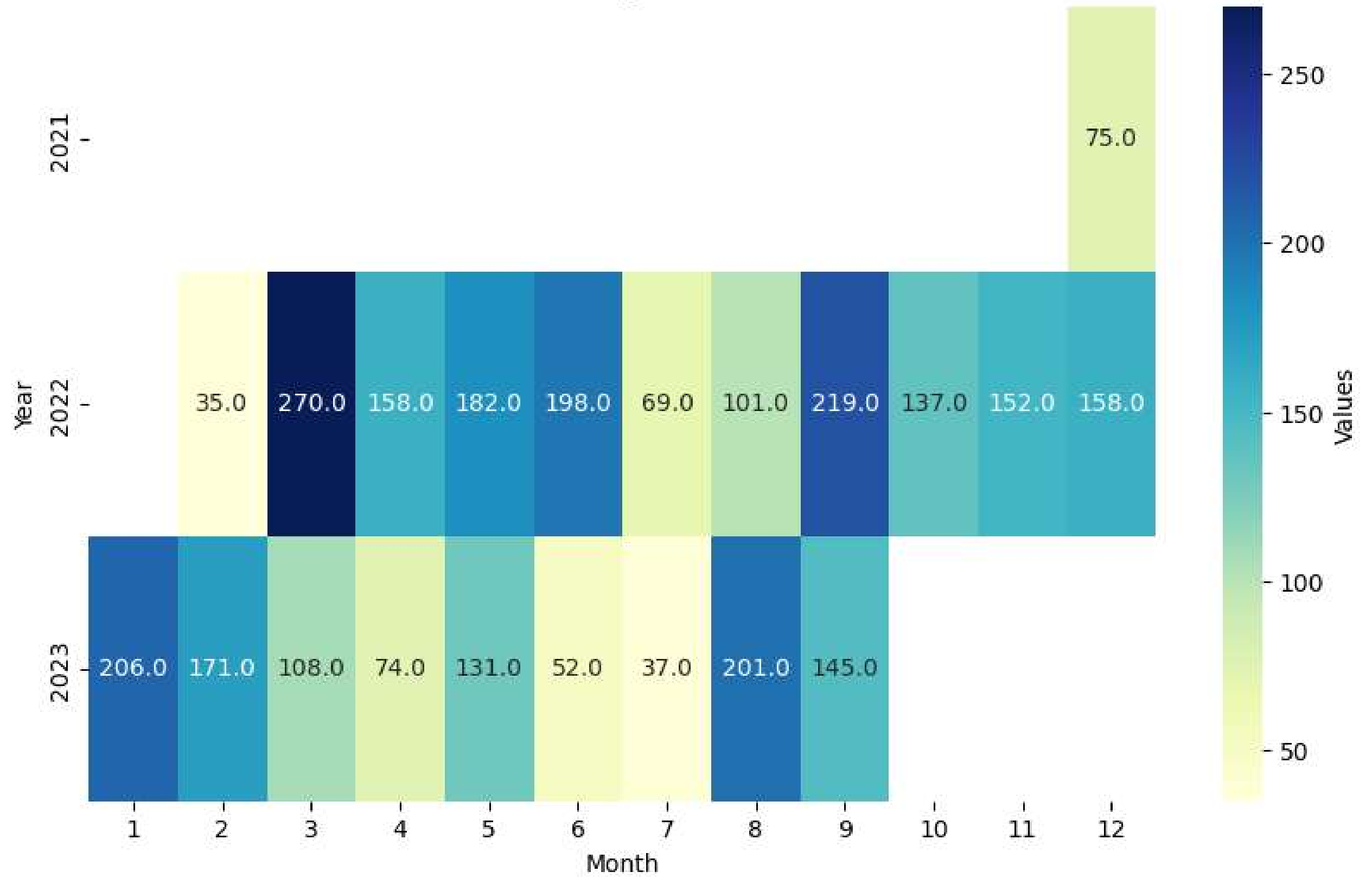


- Computer Science
- Mathematics
- Fiction
- Business & Economics
- Technology & Engineering
- Economics
- Self-Help
- Science
- Biology
- Biography & Autobiography
- Computers
- Computer algorithms
- Physics
- History
- Philosophy
- Psychology
- Social Science
- Political Science
- Medical
- Computer architecture
- Popular Science
- Body, Mind & Spirit

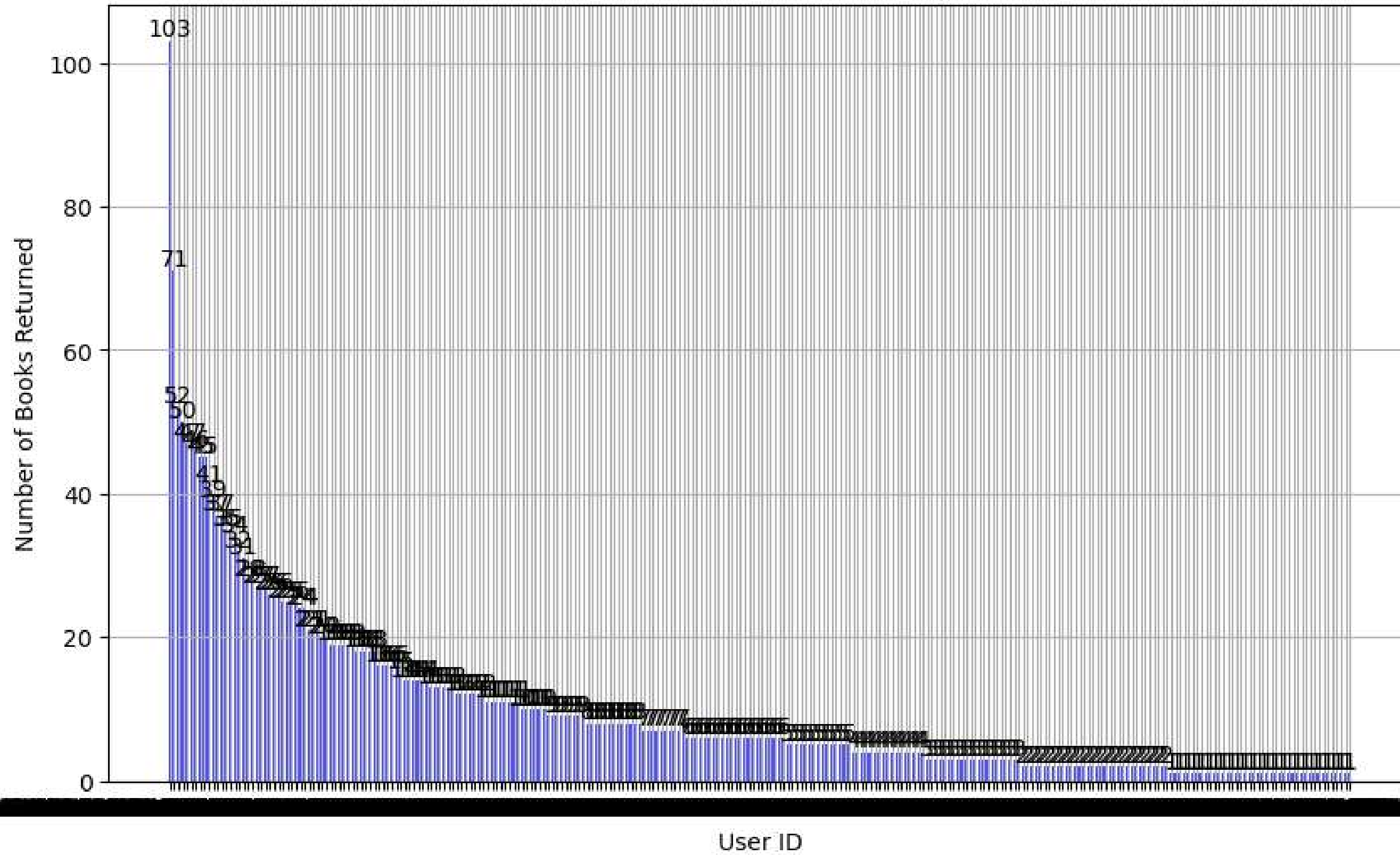
Propotion of transaction type



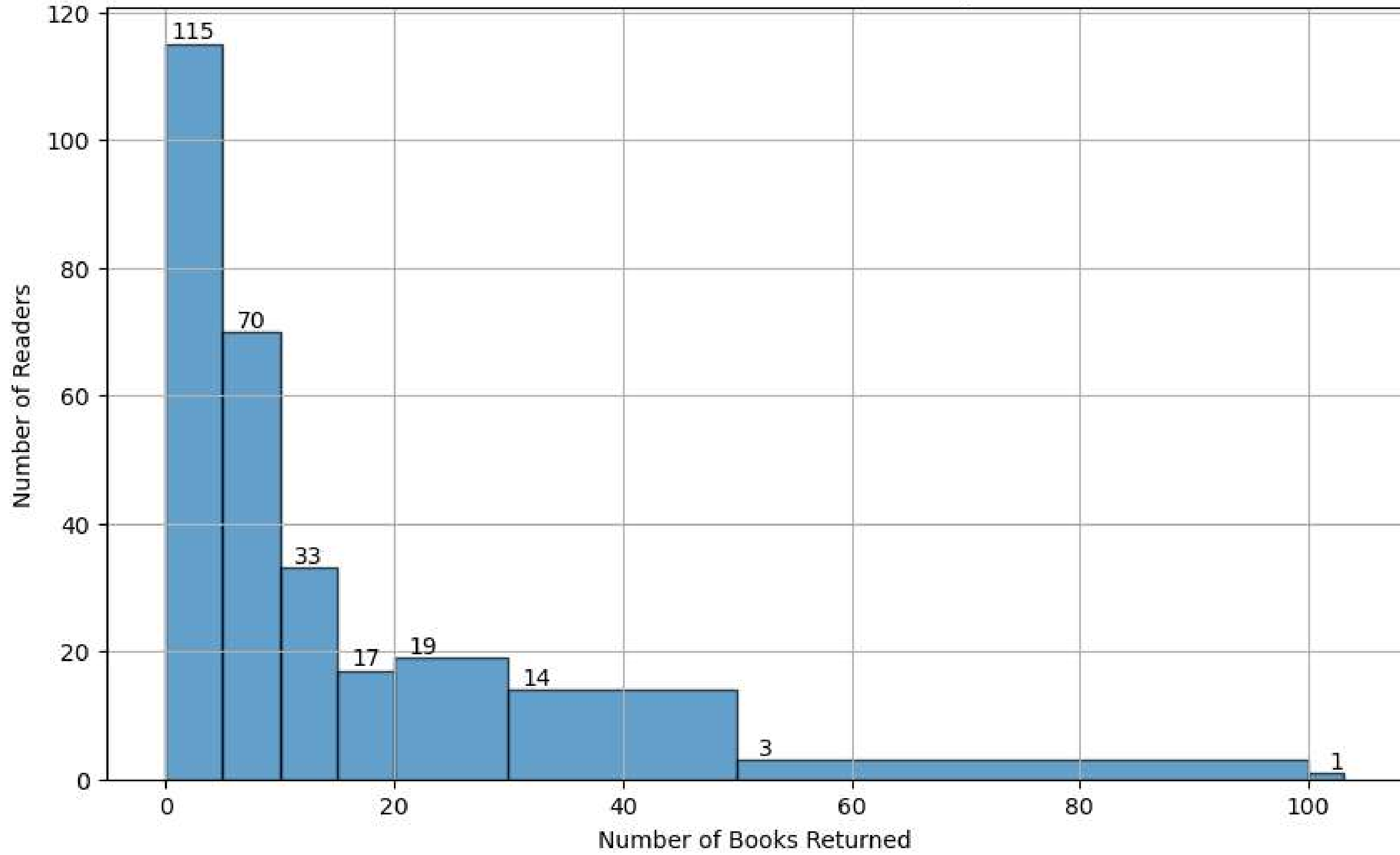
Books borrowed by Month and Year



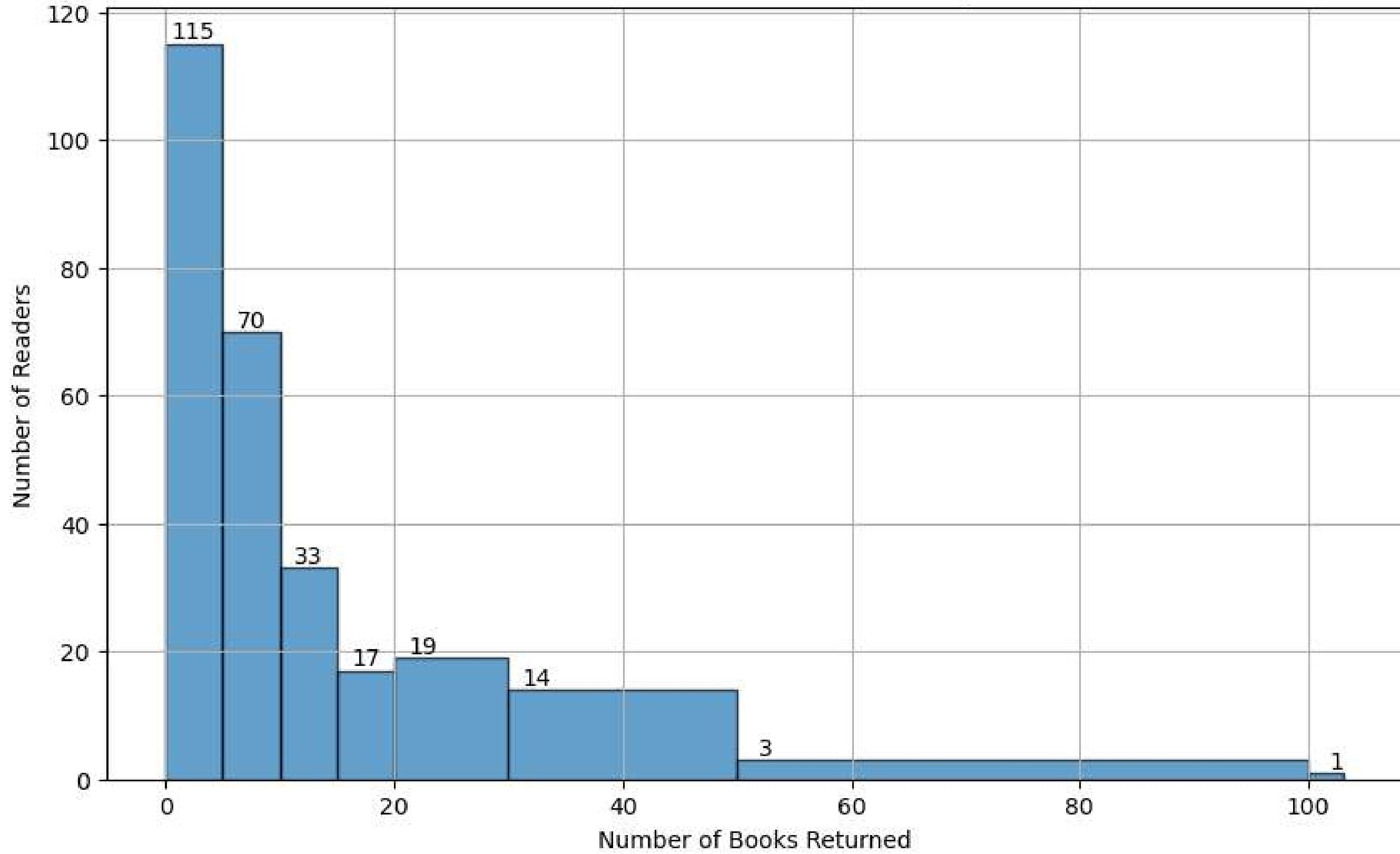
Distribution of User Activity



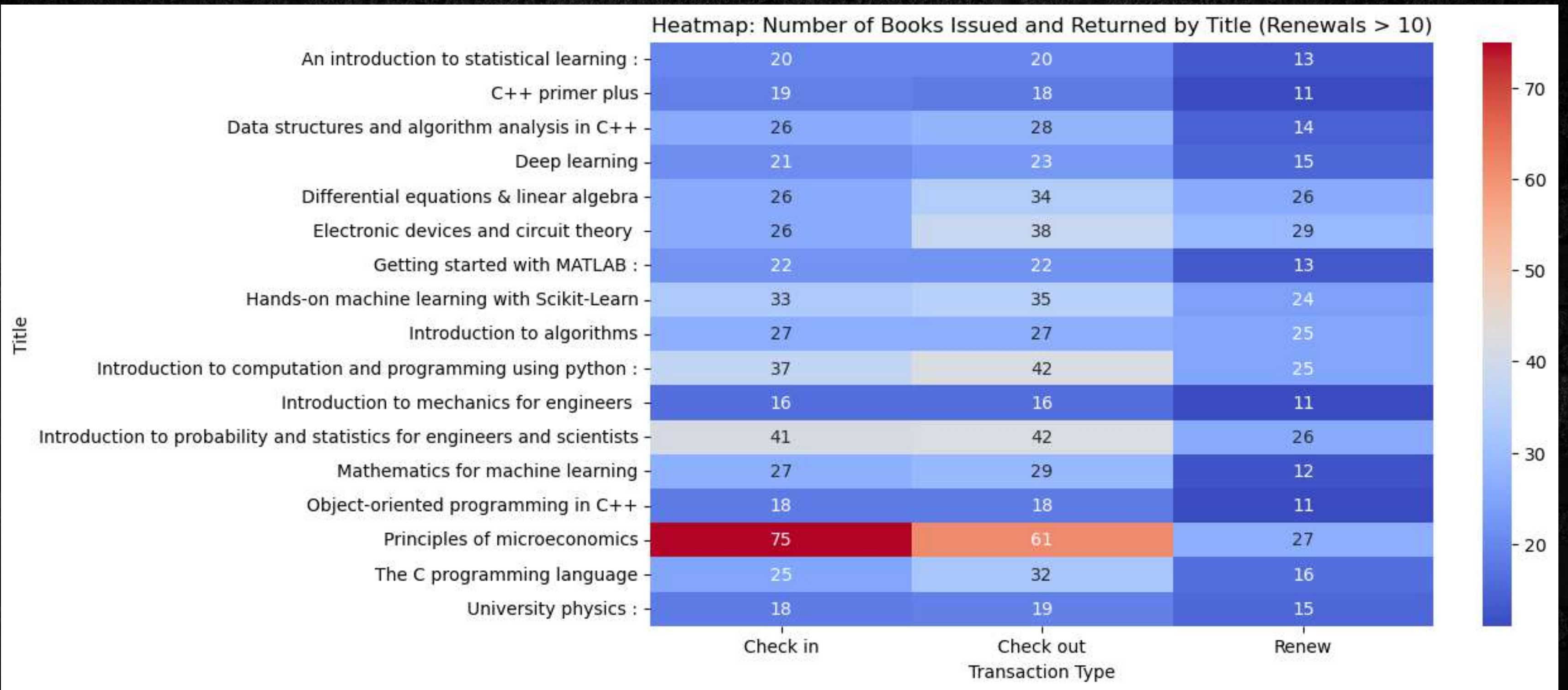
Distribution of Check in activity



Distribution of Check in activity



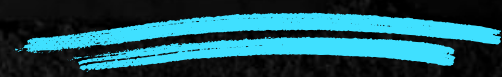
Heatmap of the most renewed books





ML

METHODOLOGY



Simple Recommendation

Weighted rating

$$\text{Weighted Rating (WR)} = \left(\frac{v}{v+m} \cdot R\right) + \left(\frac{m}{v+m} \cdot C\right)$$

where

- v is the number of votes for the movie
- m is the minimum votes required to be listed in the chart
- R is the average rating of the movie
- C is the mean vote across the whole report

	User ID	Title	AuthorFullName	Ratings	Popularity	score
3533	urn:uuid:a28bb135-c584-49f8-8950-21483a74bc7a	Principles of microeconomics	N. Gregory Mankiw	4.969748	75	4.074924
4040	urn:uuid:e62f4428-4d7b-42c8-b030-e51bfafae8b2	Principles of microeconomics	N. Gregory Mankiw	4.922300	75	4.044246
1468	urn:uuid:0c4f0608-2ce4-42c7-83cc-e40bc9913bcf	Principles of microeconomics	N. Gregory Mankiw	4.906782	75	4.034213
432	urn:uuid:c6633c0e-918b-4fa6-8bab-21369d2c74b5	Principles of microeconomics	N. Gregory Mankiw	4.865084	75	4.007253
4592	urn:uuid:de785c04-cf61-4837-8c09-6315cf6feaf7	Principles of microeconomics	N. Gregory Mankiw	4.859929	75	4.003920

Content-based Recommendation

Cosine similarity

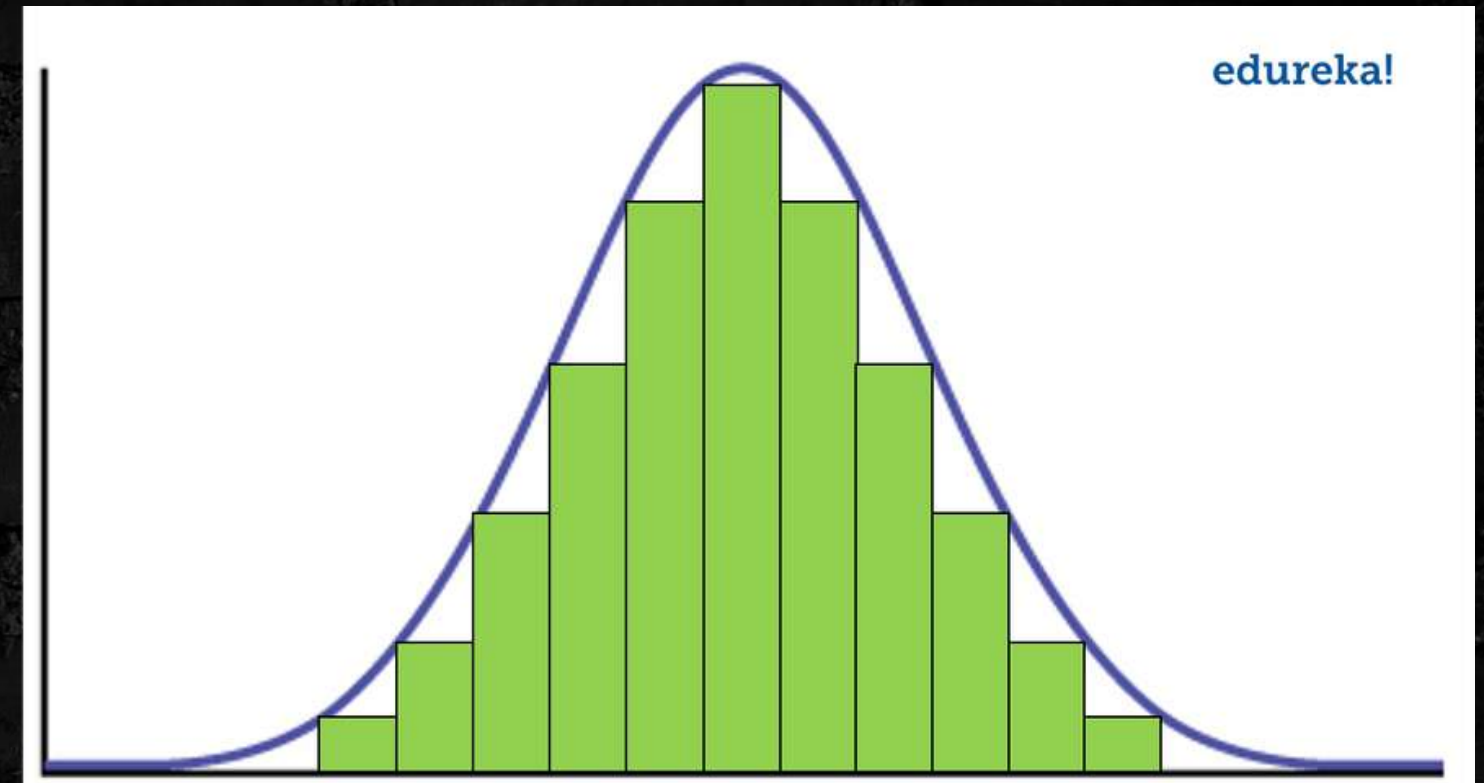
$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_i A_i B_i}{\sqrt{\sum_i A_i^2} \sqrt{\sum_i B_i^2}}$$

```
1 content_recommendation(rate, "Aunts aren't gentlemen")
```

	Title	AuthorFullName	Genre	Ratings	Popularity
2853	Piccadilly jim	P. G Wodehouse	Fiction	0.884125	1
2856	Piccadilly jim	P. G Wodehouse	Fiction	4.195725	1
3044	Uncles aunts and elephants	Ruskin Bond	Humor	1.641205	2
3047	Uncles aunts and elephants	Ruskin Bond	Humor	0.875421	2
3632	Uncles aunts and elephants	Ruskin Bond	Humor	3.992267	2

Collaborative filtering

Normal predictor - Assume ratings to be part of a normal distribution



Evaluating RMSE of algorithm NormalPredictor on 5 split(s).

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
RMSE (testset)	1.9353	1.8834	1.9800	1.8707	1.9953	1.9330	0.0499
Fit time	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Test time	0.00	0.00	0.00	0.01	0.01	0.01	0.01
CPU times: total: 0 ns							
Wall time: 42.6 ms							

NMF Method

Decomposition of matrix

RMSE: 1.7031

```
47 user_id = 'urn:uuid:070e5593-a9d9-41d1-b59a-1d1415e5de20'  
48 get_recommendation_npred(user_id, 5)  
49
```

	User ID	Title	Ratings	Popularity
214	urn:uuid:5f3b4ab0-6db2-4ac9-8b83-30b85f6e2a5d	The Evolution of the Sensitive Soul	4.840444	1
222	urn:uuid:5f3b4ab0-6db2-4ac9-8b83-30b85f6e2a5d	The Evolution of the Sensitive Soul	4.959979	1
223	urn:uuid:5f3b4ab0-6db2-4ac9-8b83-30b85f6e2a5d	The Evolution of the Sensitive Soul	2.909735	1
1957	urn:uuid:b6c8b29e-ca1d-47ed-b149-0104e8220207	Einstein The Life and Times	4.169915	1
1961	urn:uuid:b6c8b29e-ca1d-47ed-b149-0104e8220207	Einstein The Life and Times	4.876439	1
1967	urn:uuid:b6c8b29e-ca1d-47ed-b149-0104e8220207	Einstein The Life and Times	4.773053	1
1968	urn:uuid:b6c8b29e-ca1d-47ed-b149-0104e8220207	Einstein The Life and Times	1.365874	1
2173	urn:uuid:8e5a78d7-481f-43bd-956e-2fc6b299872f	Computer graphics	3.675160	1
2174	urn:uuid:8e5a78d7-481f-43bd-956e-2fc6b299872f	Computer graphics	4.985997	1
2179	urn:uuid:8e5a78d7-481f-43bd-956e-2fc6b299872f	Computer graphics	3.952416	1
2516	urn:uuid:4d1728ed-3ca3-4906-a916-a93e36cc30e8	Lean analytics	4.755025	2
2520	urn:uuid:4d1728ed-3ca3-4906-a916-a93e36cc30e8	Lean analytics	3.559595	2
4822	urn:uuid:39f58ec1-fa71-4a3d-b17b-0a2306b17079	The business case for AI	1.199519	1
4825	urn:uuid:39f58ec1-fa71-4a3d-b17b-0a2306b17079	The business case for AI	4.873103	1
5139	urn:uuid:8fa63524-4fd9-4895-b34b-32b5bfce4b7c	Lean analytics	4.387234	2
5149	urn:uuid:8fa63524-4fd9-4895-b34b-32b5bfce4b7c	Lean analytics	4.563909	2

KNN

User based collaborative filtering - cosine similarity

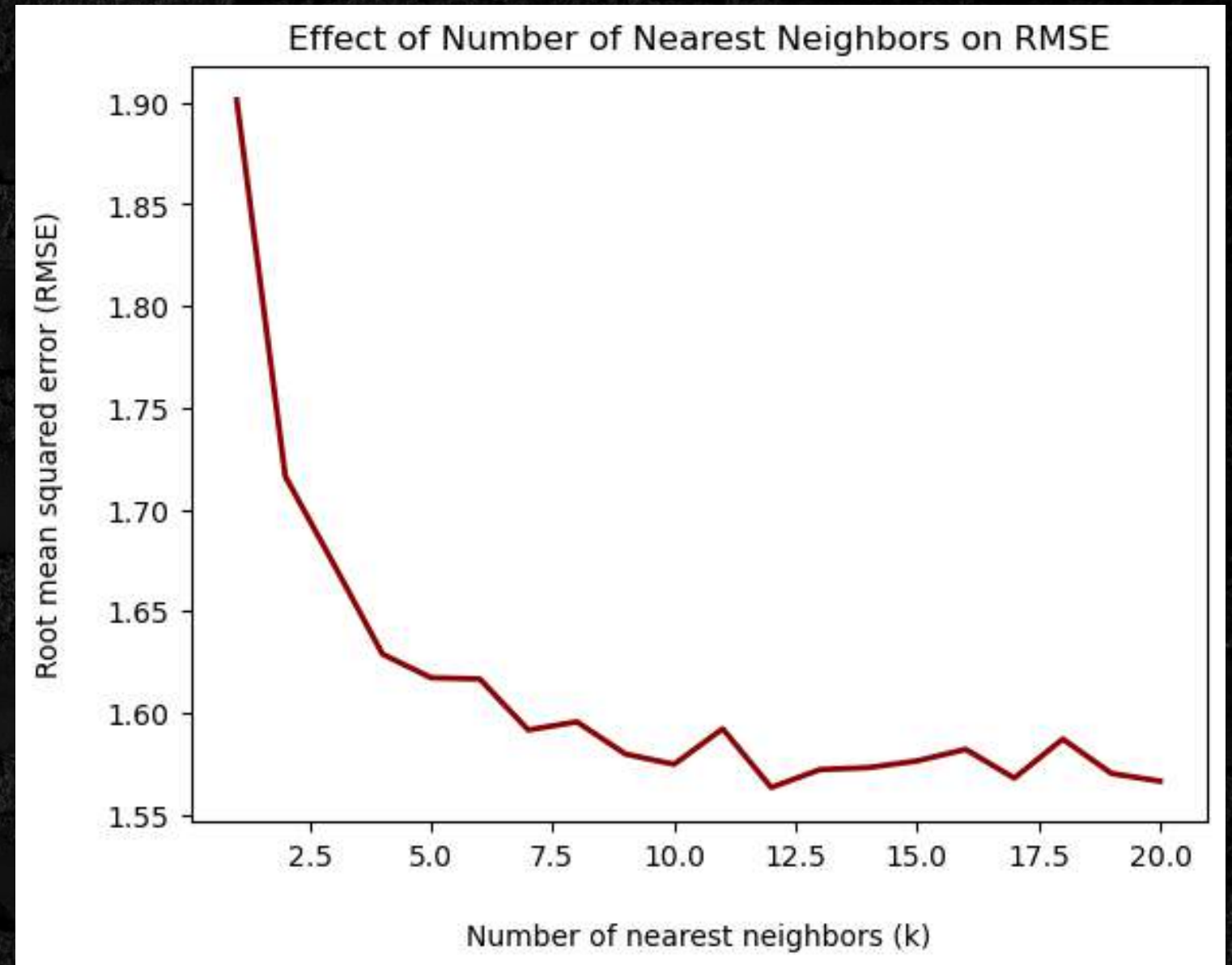
Computing the cosine similarity matrix...

Done computing similarity matrix.

RMSE: 1.5691

RMSE: 1.5691391050922812

Top 5 neighbors for user urn:uuid:070e5593-a9d9-41d1-b59a-1d1415e5de20: ['urn:uuid:417ff77a-fa5c-4759-8cac-f22cbcccf77f', 'urn:uuid:4665b802-4668-493a-962a-d97263277a75', 'urn:uuid:bdd062f3-d1bb-4979-aaaf-c30b35d77752', 'urn:uuid:01dd2320-c959-4496-bddc-66e75ed9c5a1', 'urn:uuid:1155380b-5f6a-4f8e-b95f-129f3bb1c59d', 'urn:uuid:441e697f-0f59-48db-9fda-9d5fc6784284', 'urn:uuid:766d7aea-fa2f-48ba-a849-8402fc69a2e2']



Hybrid Model

Content based + SVD

KNN Baseline

RMSE: 0.6321

MAE: 0.4921

SVD

RMSE: 0.4532

MAE: 0.3141

```
38          Title      est      Model
3  Introduction to computation and programming us...  3.781302  SVD + CF
11 Introduction to computation and programming us...  3.781302  SVD + CF
33 Introduction to computation and programming us...  3.781302  SVD + CF
['Algorithm design', 'Introduction to computation and programming using python ', 'Introduction to computation and programming using python ', 'Introduction to computation and program
ALERT: Multiple values
ALERT: Multiple values
ALERT: Multiple values
ALERT: Multiple values
```

level_0	index	User ID	Title	AuthorFullName	Transaction	Barcode	Issue & Return	Genre	Popularity	Ratings
0	0	urn:uuid:070e5593-a9d9-41d1-b59a-1d1415e5de20	Introduction to computation and programming us...	John V. Guttag	Check out	2358	2022-03-01 11:45:31	Computer Science	31	3.80
1	1	urn:uuid:070e5593-a9d9-41d1-b59a-1d1415e5de20	The Psychology of money	Morgan Housel	Check in	264	2022-03-05 10:47:54	Business & Economics	7	2.73
2	2	urn:uuid:070e5593-a9d9-41d1-b59a-1d1415e5de20	Introduction to computation and programming us...	John V. Guttag	Check in	2358	2022-03-15 20:15:23	Computer Science	31	3.80
3	3	urn:uuid:070e5593-a9d9-41d1-b59a-1d1415e5de20	Introduction to computation and programming us...	John V. Guttag	Check out	2358	2022-03-15 20:16:03	Computer Science	31	3.80
4	4	urn:uuid:070e5593-a9d9-41d1-b59a-1d1415e5de20	Introduction to probability and statistics for...	Sheldon M Ross	Check in	2394	2022-03-16 18:00:46	Mathematics	41	3.79
...
5223	5223	urn:uuid:b0da4d10-ec68-4b37-9c51-c053de2e314b	Deep Work	Cal Newport	Renew	G000008	2022-03-10 19:00:40	Business & Economics	3	3.82
5224	5224	urn:uuid:b0da4d10-ec68-4b37-9c51-c053de2e314b	What would keynes do?	Tejvan Pettinger	Renew	592	2022-03-10 19:00:40	Fiction	5	1.03
5225	5225	urn:uuid:b0da4d10-ec68-4b37-9c51-c053de2e314b	Deep Work	Cal Newport	Renew	G000008	2022-03-16 15:40:23	Business & Economics	3	3.82
5226	5226	urn:uuid:b0da4d10-ec68-4b37-9c51-c053de2e314b	Factfulness	Hans. Rosling	Renew	541	2022-03-29 16:18:43	Self-Help	1	1.52
5227	5227	urn:uuid:b0da4d10-ec68-4b37-9c51-c053de2e314b	What would keynes do?	Tejvan Pettinger	Renew	592	2022-03-29 16:18:43	Fiction	5	1.03

5228 rows x 11 columns



Challenges

Future Scope

THANK

you

