To what extent can we predict book ratings using the attributes of a book and a user

Presented by Emily, Erich, Raphael

Why are ratings important?

Bookstores often collect data on book ratings that have been given by their customers for several reasons:

- Provide better recommendations to customers
- Identify which books are better liked than others
- Avoid recommending books with low ratings

What about the books with no ratings?

How would bookstore managers know which books to recommend to a customer?

Use our recommendation system that can predict how each customer would rate these new books!

Data pre-processing first!

Scaling

All numerical data must not have any outliers or numbers not in range:

- **Ratings**: Need to be in range of 1 to 10
- **Ages**: None below 6 years old (children start reading at 6)
  None above 100 (considered unrealistic)
- **Years of publishing**: No years beyond 2024 (future)

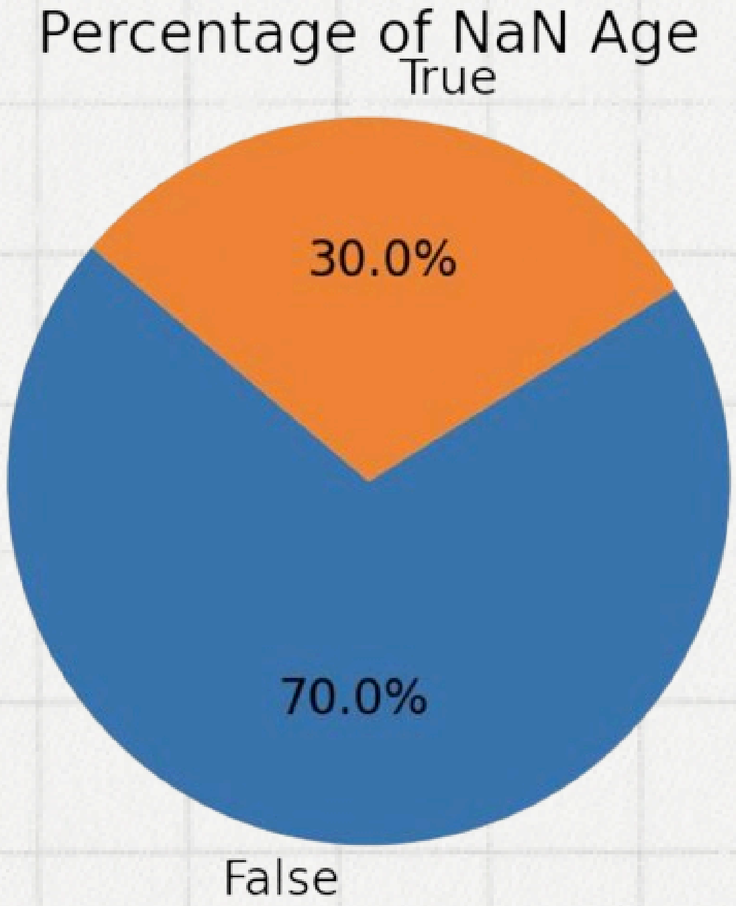


Figure 1: Pie chart showing the percentage of NaN User Age in the dataset

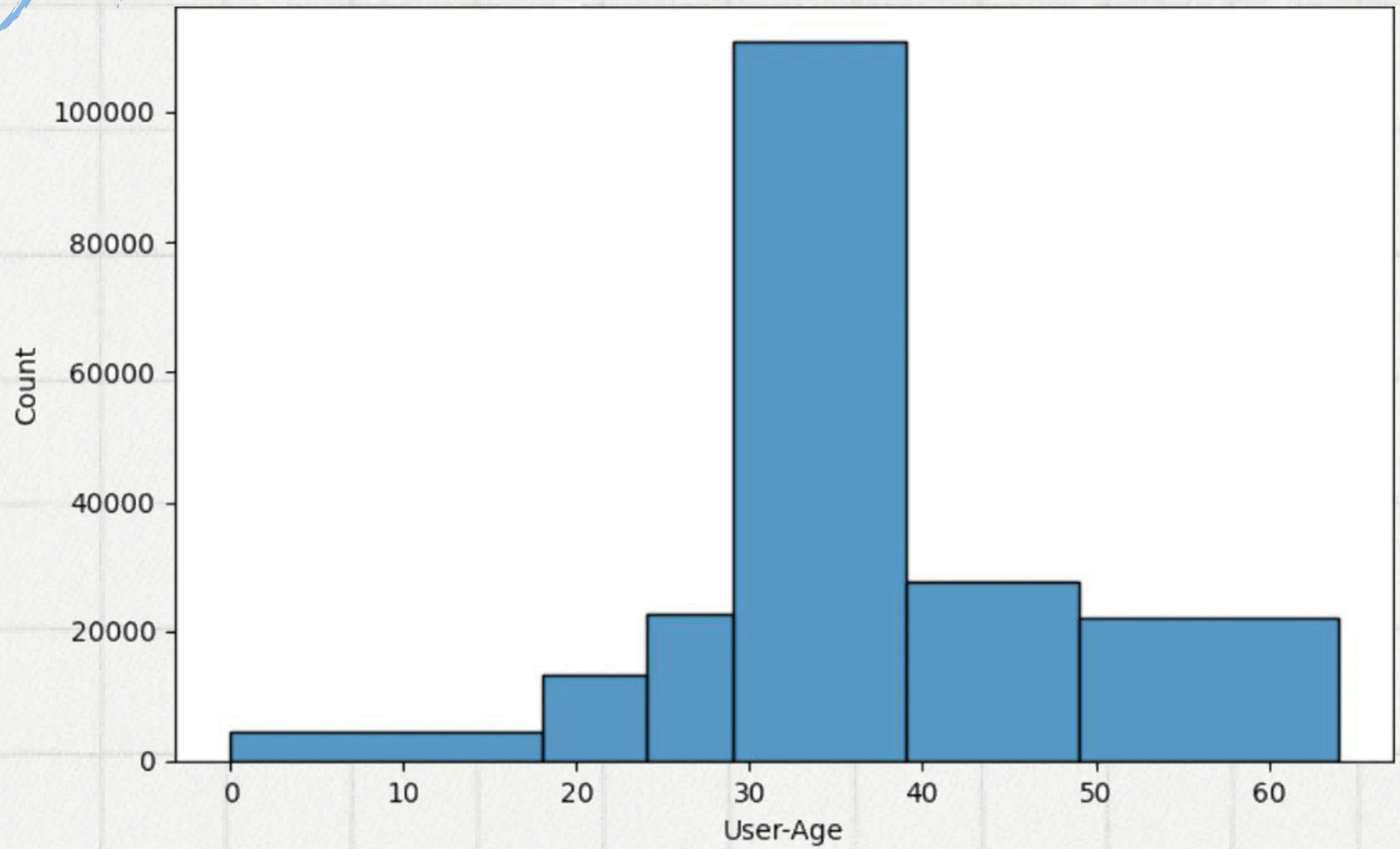


Figure 2: Histogram of the User Age frequency in bins after filling the NaN values with median

K-nearest neighbor

Fill each missing age of a user with the average age of 5 nearest neighbors based on their location.
Location = Coordinate points of user's country latitude and longitude
(added 2 new columns in dataset)

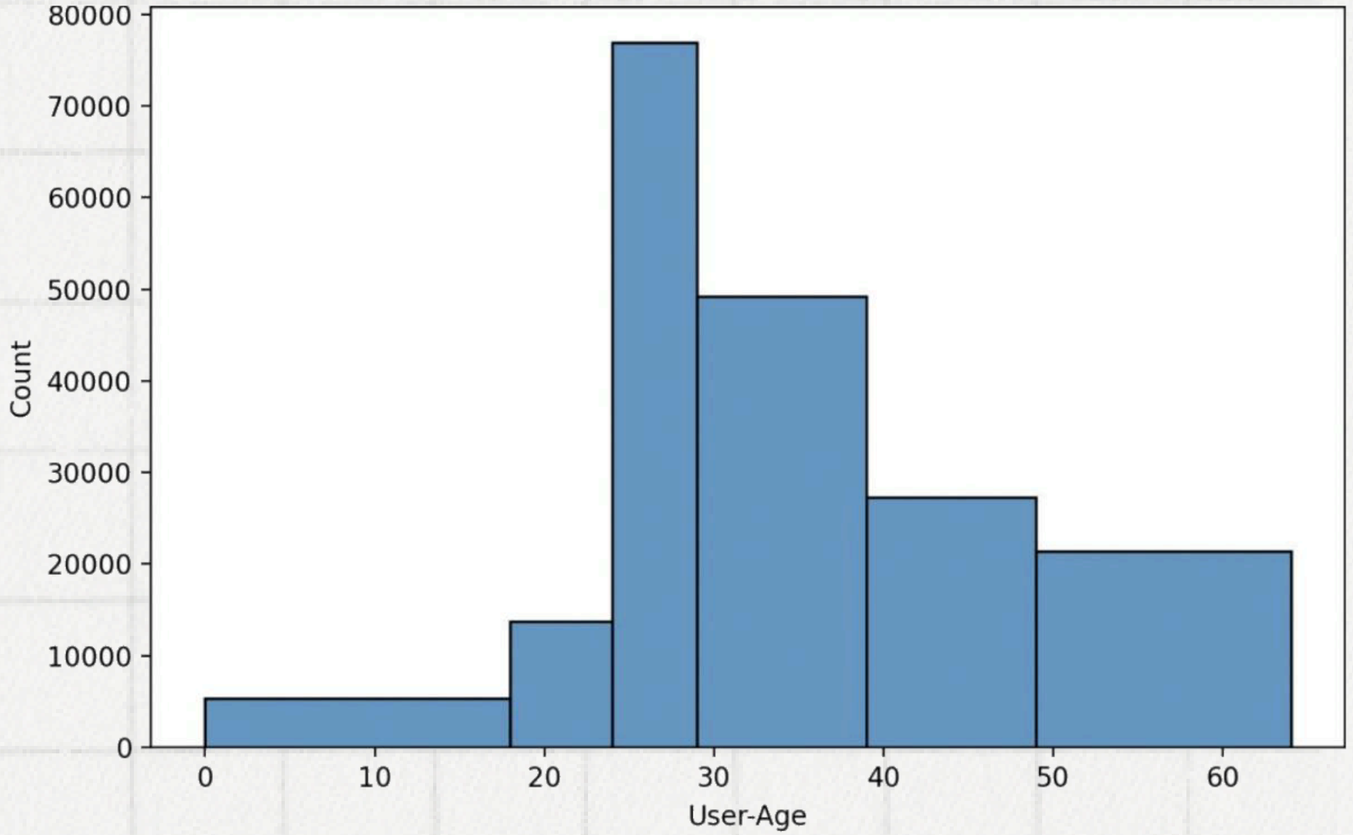


Figure 3: Histogram of User Age frequency in bins after filling the NaN values using KNN

Clustering

Assuming book titles can show similarities between books :

- **Calculate TF-IDF** vectors for each book title in the dataset
- Use **K-means clustering** for TF-IDF book titles
- Find best K value from **elbow method**

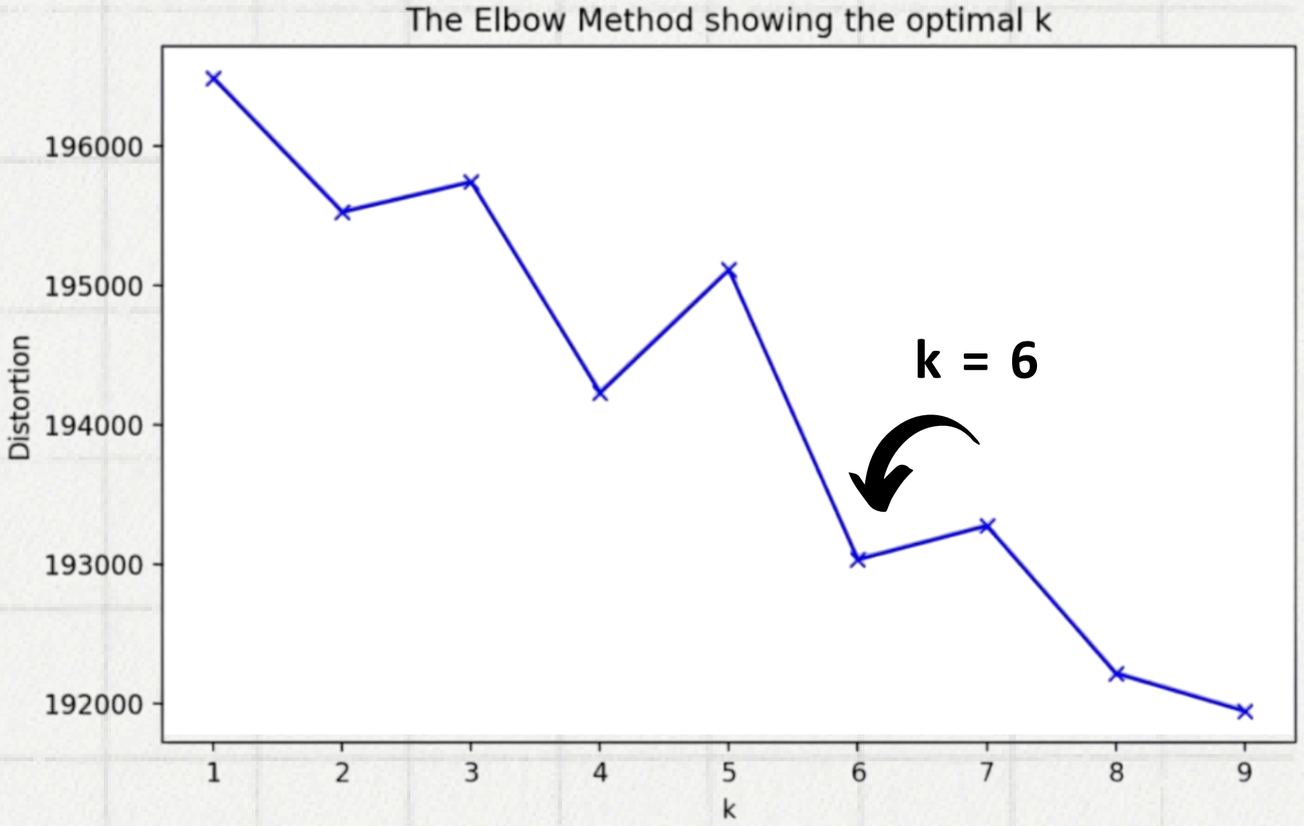


Figure 6: The Distortion Plot of K-Means for given k

Correlation Analysis

$$H(\boldsymbol{X}) = -\sum_{i=1}^{k} p_i \log_2 p_i$$

p_i: proportion of points in the i-th category (bin)

- NMI(X,Y) = MI(X,Y) / min(H(X),H(Y))
- NMI(X,Y) = MI(X,Y) / max(H(X),H(Y))
- NMI(X,Y) = MI(X,Y) / mean(H(X),H(Y))

01

Entropy

02

Mutual Information

$$MI(X,Y) = H(Y) - H(Y|X)$$
$$= H(X) - H(X|Y)$$
$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x)$$

where X and Y are features (columns) in a dataset

03

Normalized Mutual Information

04

Information Gain

$$\mathrm{IG}(X, y) = H(X) - H(X|y) = H(y) - H(y|X)$$

Methods & Implementations

- **Methods undergone:**
  a. Match the new book's user and its title cluster with the old dataset's
  b. Predict new books' rating looking at users' mean rating of the matching cluster
  c. Fill any gap using overall user mean rating
- **Implementation**:
  - Personalized Recommendation
  - Low accuracy = Lose trust
  - Publicity for authors & publishers

Title-Cluster	0	1	2	3	4	5	6
User-ID							
57402	NaN	NaN	8.0	NaN	NaN	NaN	NaN
41932	NaN	NaN	NaN	NaN	NaN	6.000000	NaN
230689	NaN	NaN	NaN	NaN	NaN	8.000000	NaN
218930	NaN	NaN	NaN	NaN	NaN	7.000000	NaN
139822	NaN	NaN	NaN	NaN	NaN	8.500000	NaN
127417	NaN	NaN	NaN	NaN	NaN	7.600000	8.0
193961	NaN	NaN	NaN	NaN	NaN	7.000000	NaN
36169	NaN	NaN	NaN	NaN	NaN	7.333333	NaN
13733	NaN	NaN	NaN	NaN	NaN	7.000000	NaN
91931	NaN	NaN	NaN	NaN	NaN	9.100000	NaN

Figure 7: Pivot Table of Mean Rating

Result

- **Accuracy around 62.4%**

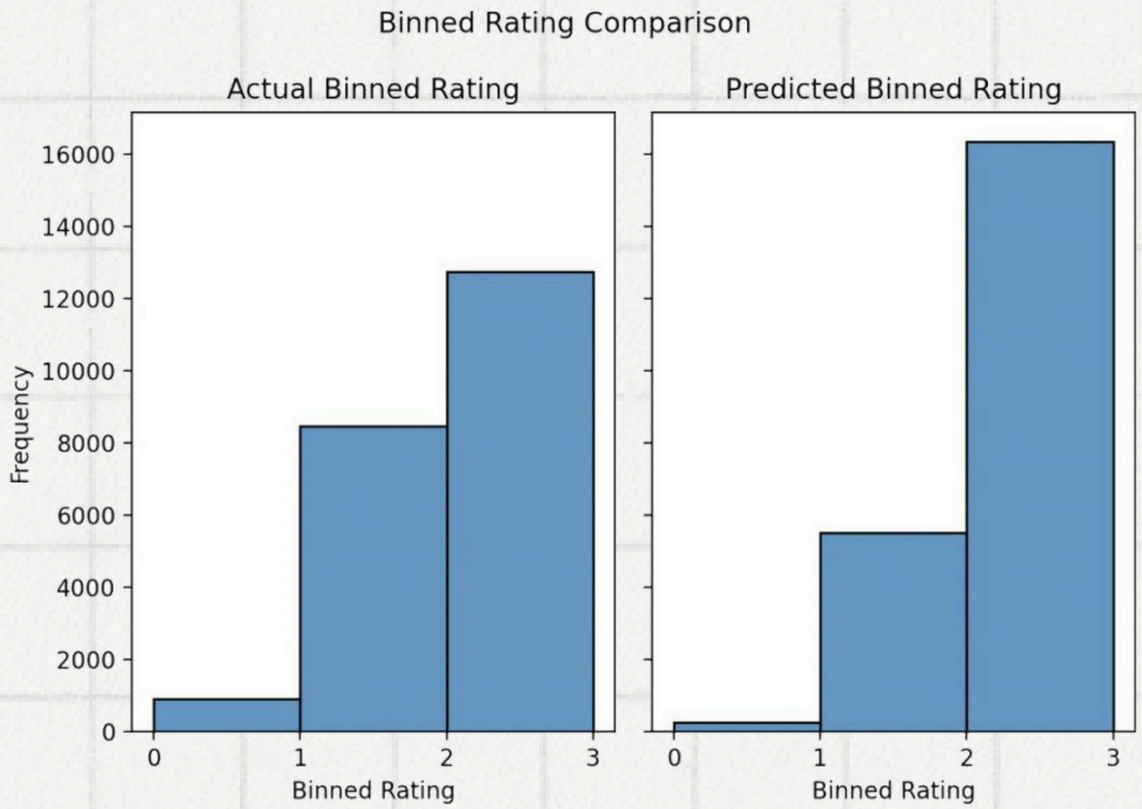


Figure 8: Histogram Comparison of Binned Rating

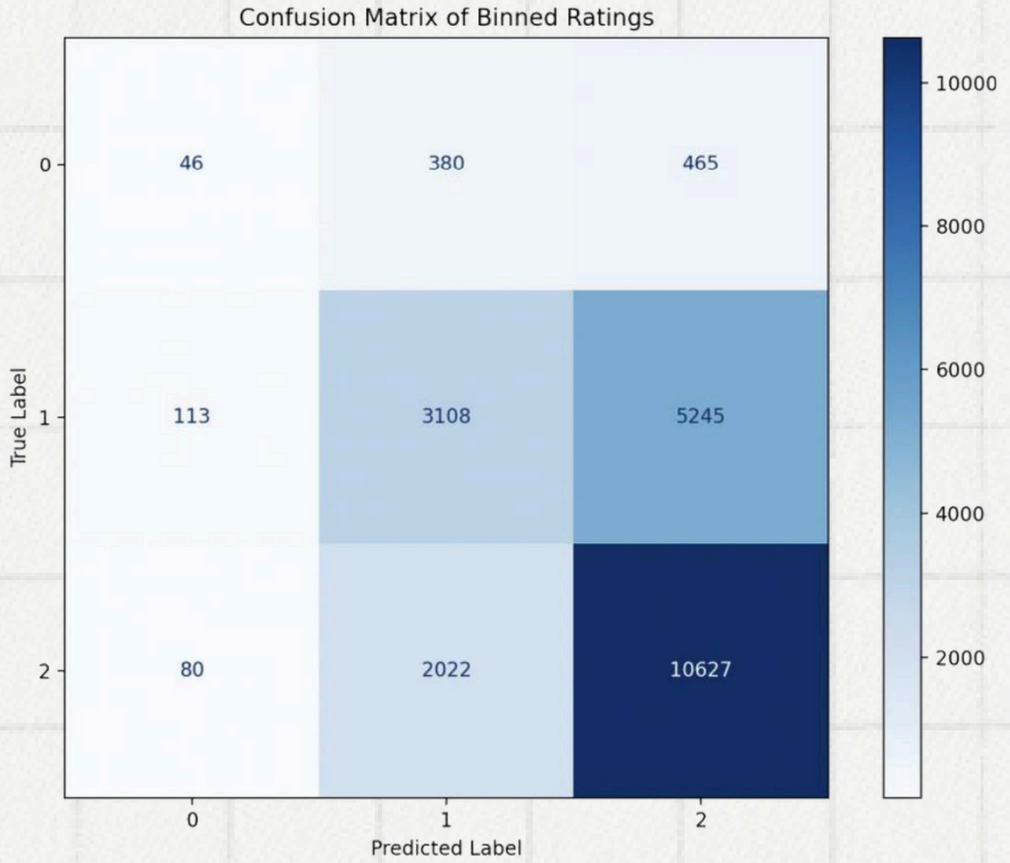


Figure 9: Confusion Matrix Of Predicted Binned Rating and the Actual Binned Rating

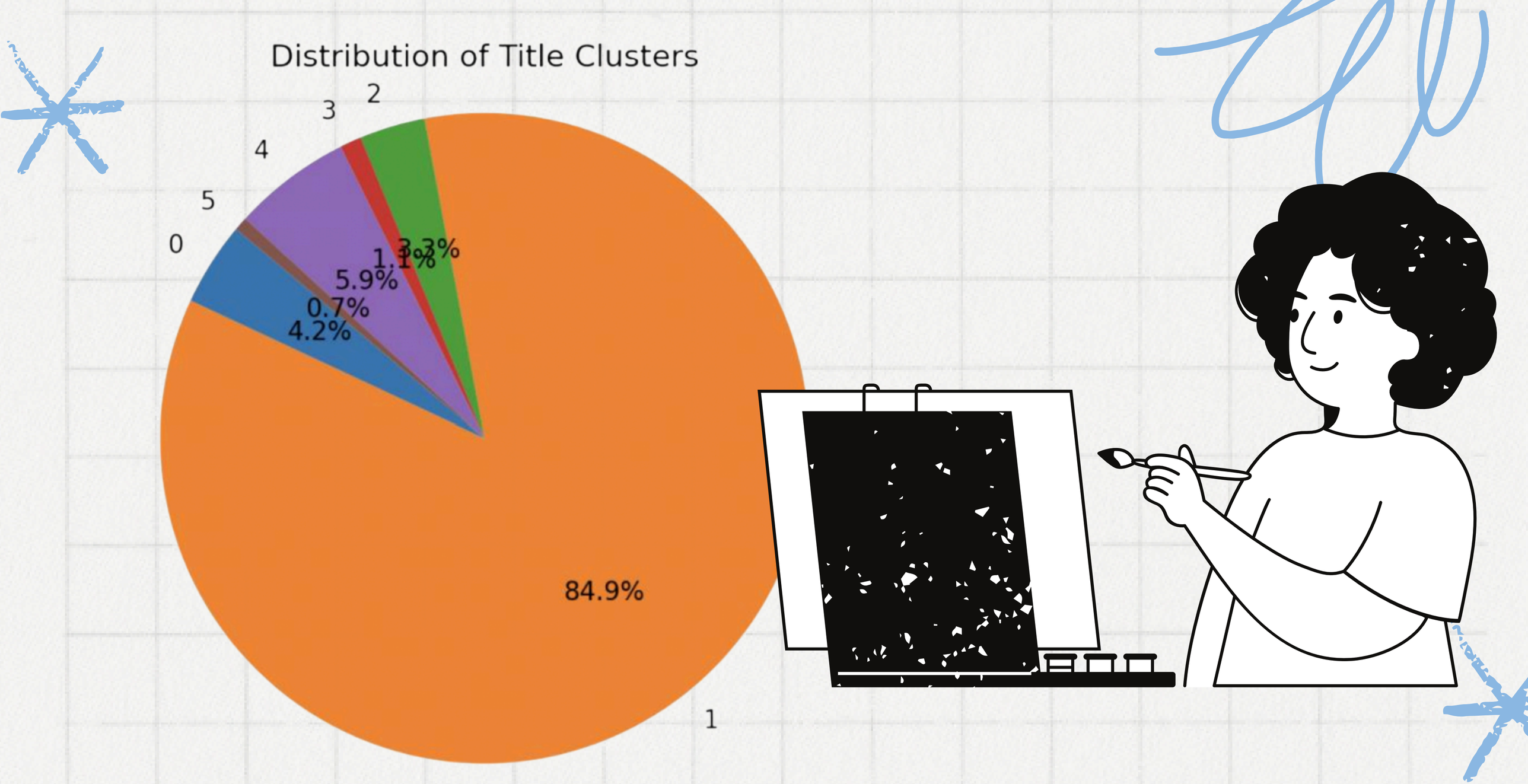


Figure 10: Pie chart showing the Distribution of Title Cluster

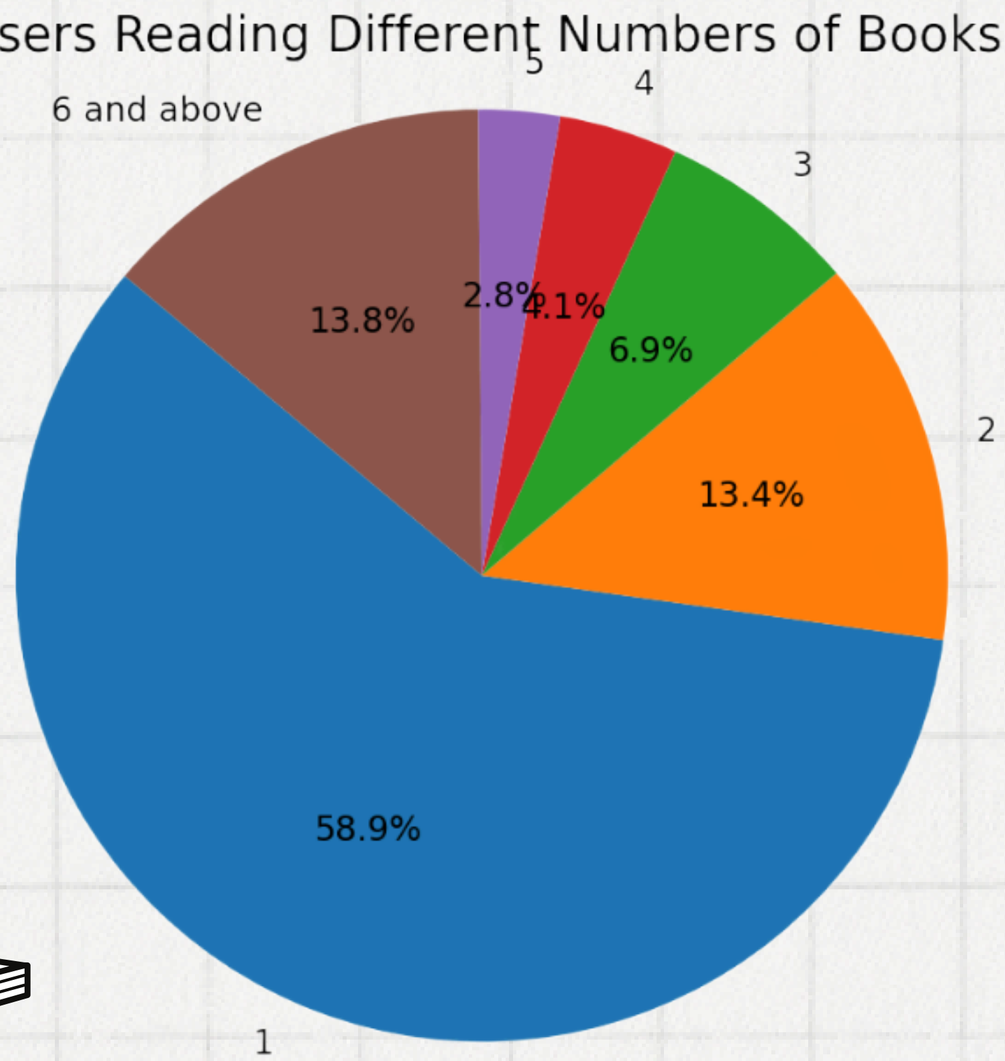
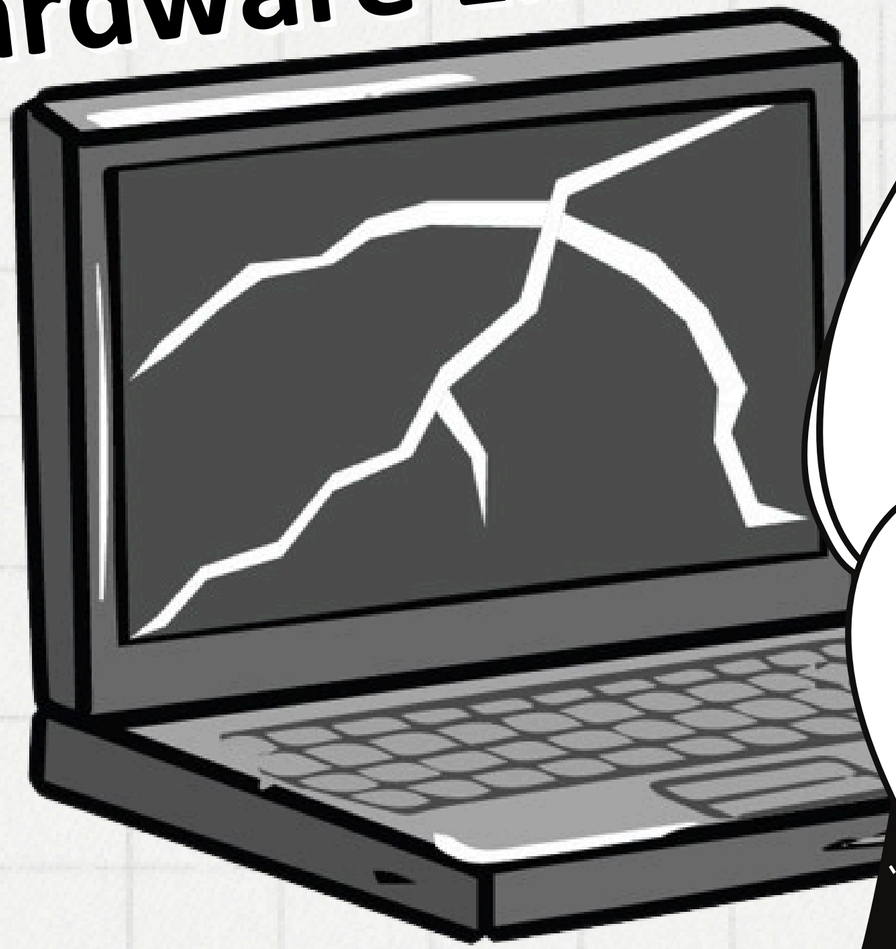
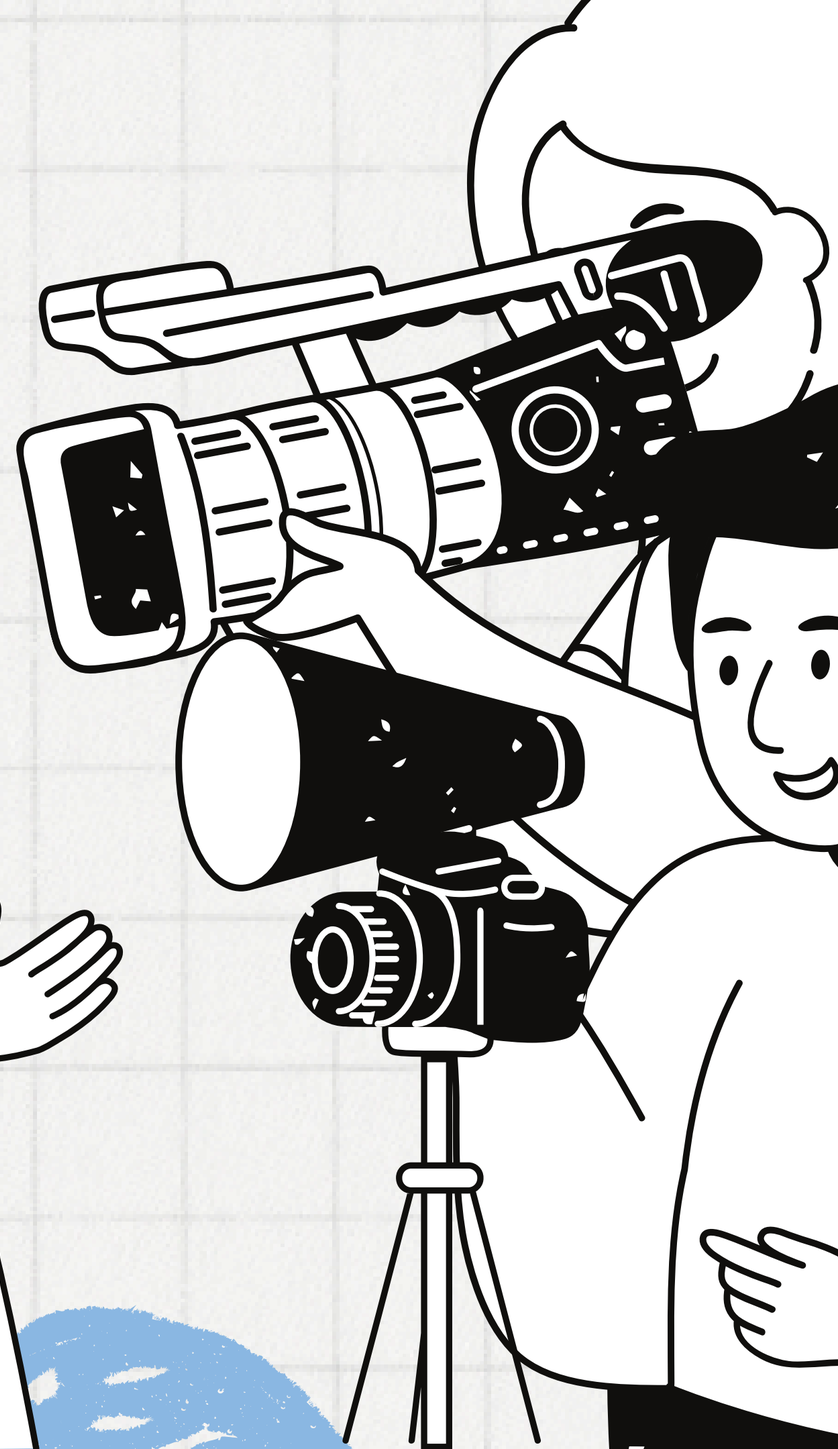


Figure 11: Pie chart showing the Distribution of the amount books read by users

Hardware Limitation





Thank you very much!

For listening