CS 24200 Project 3 Report

Matthew Sindac

November 20, 2023

1 Summary

This technical report investigates the Book-Crossing dataset which is comprised of user ratings for books sourced from many users. With approximately one million ratings for around 270,000 books from 280,000 users, there are many insights this dataset contains. The primary goal of this project is to uncover patterns in user behavior, identify popular books, and discern clusters of users based on their rating patterns.

The analyses in this project have provided key insights into the dataset. The data has revealed the top-rated books based on user frequency, showcasing the literary preferences of the users. Additionally, clustering analysis has uncovered distinct groups of users, each with unique patterns in their book ratings.

Overall, this report serves as a comprehensive exploration of the Book-Crossing dataset, offering valuable insights into user preferences and behavior. The findings have the potential to inform recommendation systems and enhance our understanding of user interactions with literary content. With this analysis, contributing to the improvement of book recommendation strategies is possible.

2 Introduction

In our exploration of the Book-Crossing dataset, our mission is discover valuable insights of the data using a diverse set of analytical tools, visualization techniques, and comparative analyses. By using a comprehensive approach, we want to find information within this dataset, comprising user ratings for a wide variety of books books.

This analysis is dedicated to identifying patterns and relationships within the data, with the ultimate goal to provide insights into user preferences and behavior in regards to literary content. By examining various data points and using many different analytical methods, our objective is to reveal correlations that contribute to a deeper understanding of the Book-Crossing dataset. Through this exploration, we will use these data-driven insights to improve the effectiveness of book recommendation systems and possibly contribute knowledge to the broader field of literature and user engagement.

3 Methodology

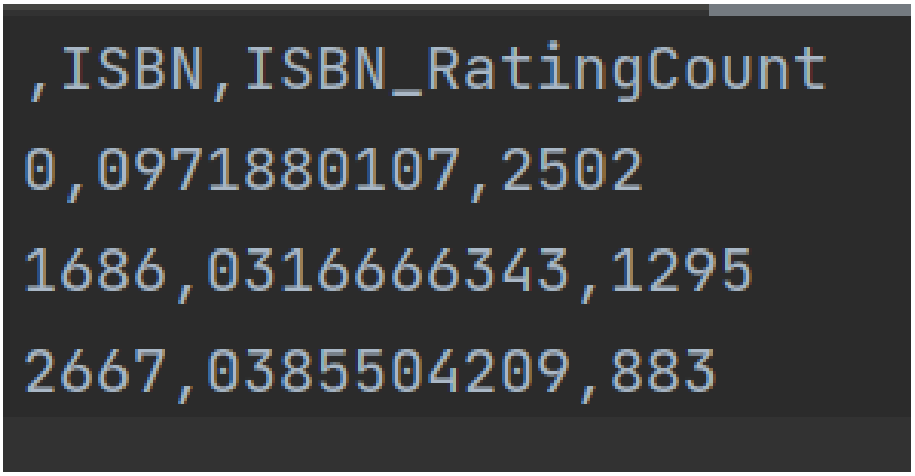
For this project, I utilized all three of the Book-Crossing datasets which contained three csv files which were the following:

- BX-Book-Ratings.csv: you are provided with ratings (on a scale from 1 to 10) for books by different users. Both users and books are given in the form of ID values. It contains around 1M ratings for across around 270K books from around 280K users.
- BX-Books.csv: the books are identified with their ISBN and content related information is associated obtained from AWS
- BX-Users.csv: demographics data is provided with the users anonymized.

3.1 Transforming Data

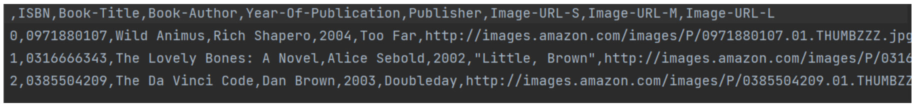
For transforming the data, I focused on the BX-Book-Ratings.csv that contained user ratings with one book/user combination per line. I loaded in the data set using pandas and replaced zero ratings with the mean rating of the book. After this, I selected a subset of the data with books that have been rated 200 times and users that have rated 5 times. I transformed this subset of data into a user-book ratings matrix which will be used later on in this project.

From the subset, we are able to gather the top three books in terms of the number of users that have rated them which is the following:



```
,ISBN,ISBN_RatingCount
0,0971880107,2502
1686,0316666343,1295
2667,0385504209,883
```

(Top 3 books in terms of number of users that have rated them)



```
,ISBN,Book-Title,Book-Author,Year-Of-Publication,Publisher,Image-URL-S,Image-URL-M,Image-URL-L
0,0971880107,Wild Animus,Rich Shapero,2004,Too Far,http://images.amazon.com/images/P/0971880107.01.THUMBZZZ.jpg
1,0316666343,The Lovely Bones: A Novel,Alice Sebold,2002,"Little, Brown",http://images.amazon.com/images/P/0316
2,0385504209,The Da Vinci Code,Dan Brown,2003,Doubleday,http://images.amazon.com/images/P/0385504209.01.THUMBZZ
```

(The top 3 books titles)

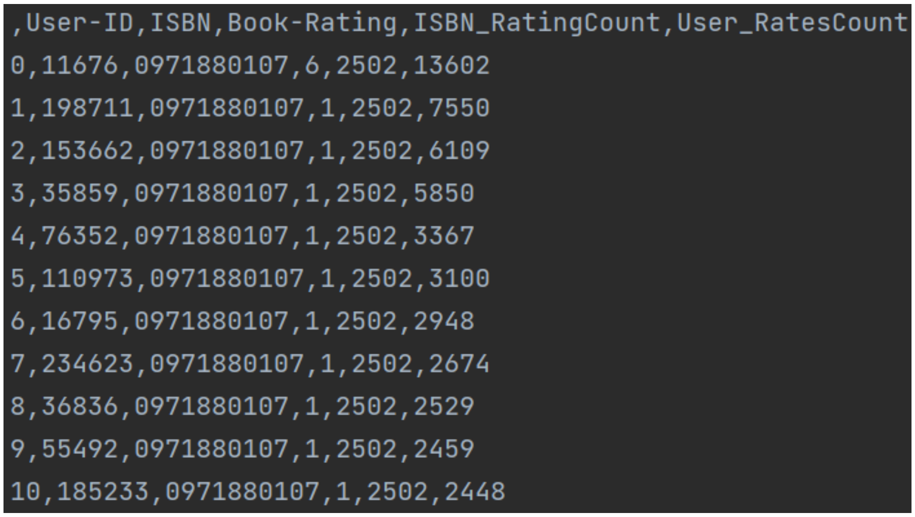
So, the titles of the top 3 books were the following:

- Wild Animus
  Number of votes: 2502

- The Lovely Bones: A Novel
  Number of votes: 1295

- The Da Vinci Code
  Number of votes: 883

From the subset, we are also able to gather the top three users in terms of how many books they have rated which is the following:



```
,User-ID,ISBN,Book-Rating,ISBN_RatingCount,User_RatesCount
0,11676,0971880107,6,2502,13602
1,198711,0971880107,1,2502,7550
2,153662,0971880107,1,2502,6109
3,35859,0971880107,1,2502,5850
4,76352,0971880107,1,2502,3367
5,110973,0971880107,1,2502,3100
6,16795,0971880107,1,2502,2948
7,234623,0971880107,1,2502,2674
8,36836,0971880107,1,2502,2529
9,55492,0971880107,1,2502,2459
10,185233,0971880107,1,2502,2448
```

(Top 3 users in terms of number of ratings they have given)

So, the top 3 users along with how many ratings they have given are the following:

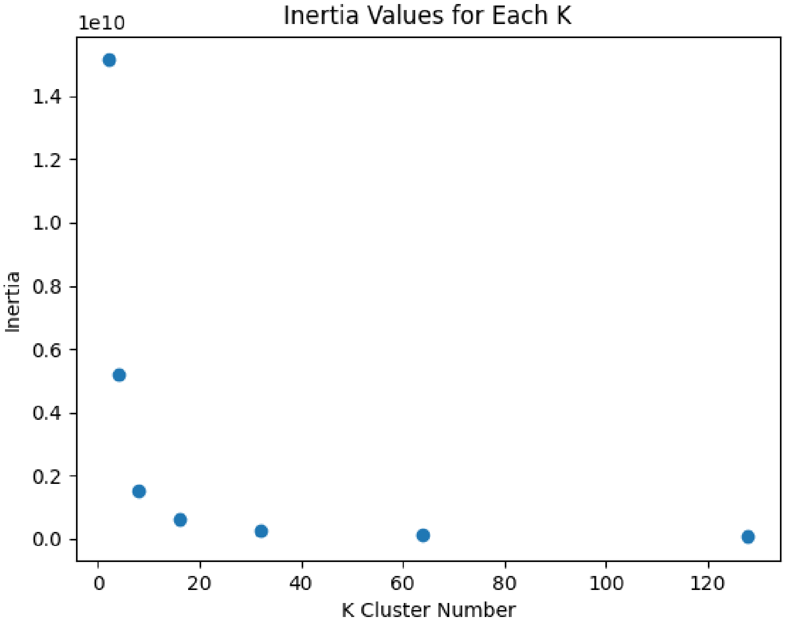
- User: 11676
  Number of ratings: 13602

- User: 198711
  Number of ratings: 7550

- User: 153662
  Number of ratings: 6109

3.2 Clustering

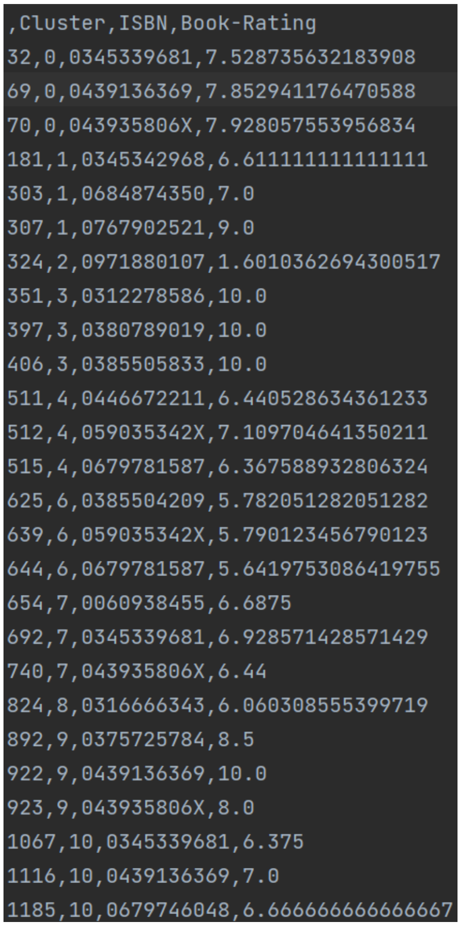
For clustering, I applied k-means clustering to the data subset from the previous question for values of k = [2, 4, 8, 16, 32, 64, 128] and measured the inertia for each value of k which came out to [15147595257.786577, 5202373085.26326, 1527631649.5550025, 594481572.7310531, 274139265.3758855, 119623863.28015089, 53784983.80097392] respectively. This resulted in the following visualization:



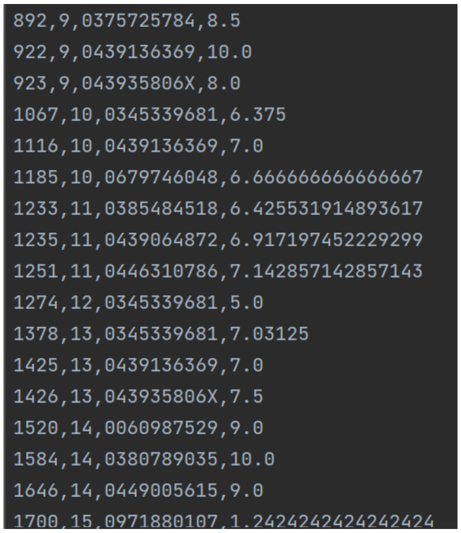
(Plot of the inertia scores)

From the results above, I believe the most appropriate value of k to use for this data is 16. I reached this conclusion since looking at the visualization above, the elbow point, which is used to find the optimal k value, is at k = 16. Using this rule, I concluded that the k value of 16 is the most appropriate k value for this data.

Using k = 16, I found the following results in regards to the top three books that are highest rated on average by users within the cluster:

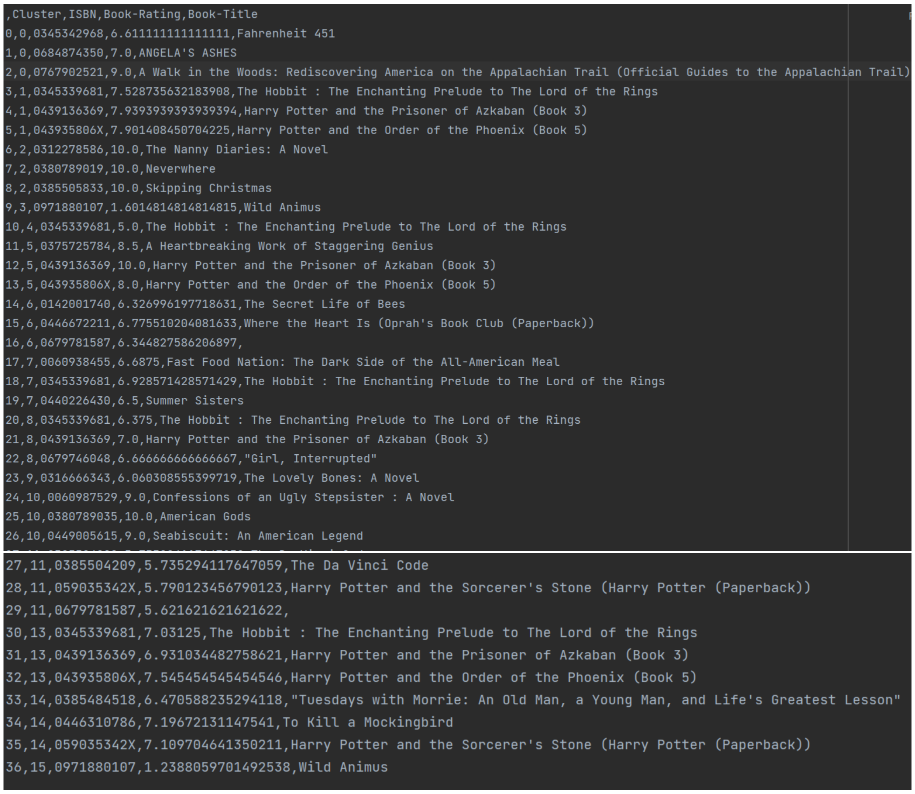


```
,Cluster,ISBN,Book-Rating
32,0,0345339681,7.528735632183908
69,0,0439136369,7.852941176470588
70,0,043935806X,7.928057553956834
181,1,0345342968,6.611111111111111
303,1,0684874350,7.0
307,1,0767902521,9.0
324,2,0971880107,1.6010362694300517
351,3,0312278586,10.0
397,3,0380789019,10.0
406,3,0385505833,10.0
511,4,0446672211,6.440528634361233
512,4,059035342X,7.109704641350211
515,4,0679781587,6.367588932806324
625,6,0385504209,5.782051282051282
639,6,059035342X,5.790123456790123
644,6,0679781587,5.6419753086419755
654,7,0060938455,6.6875
692,7,0345339681,6.928571428571429
740,7,043935806X,6.44
824,8,0316666343,6.060308555399719
892,9,0375725784,8.5
922,9,0439136369,10.0
923,9,043935806X,8.0
1067,10,0345339681,6.375
1116,10,0439136369,7.0
1185,10,0679746048,6.666666666666667
```



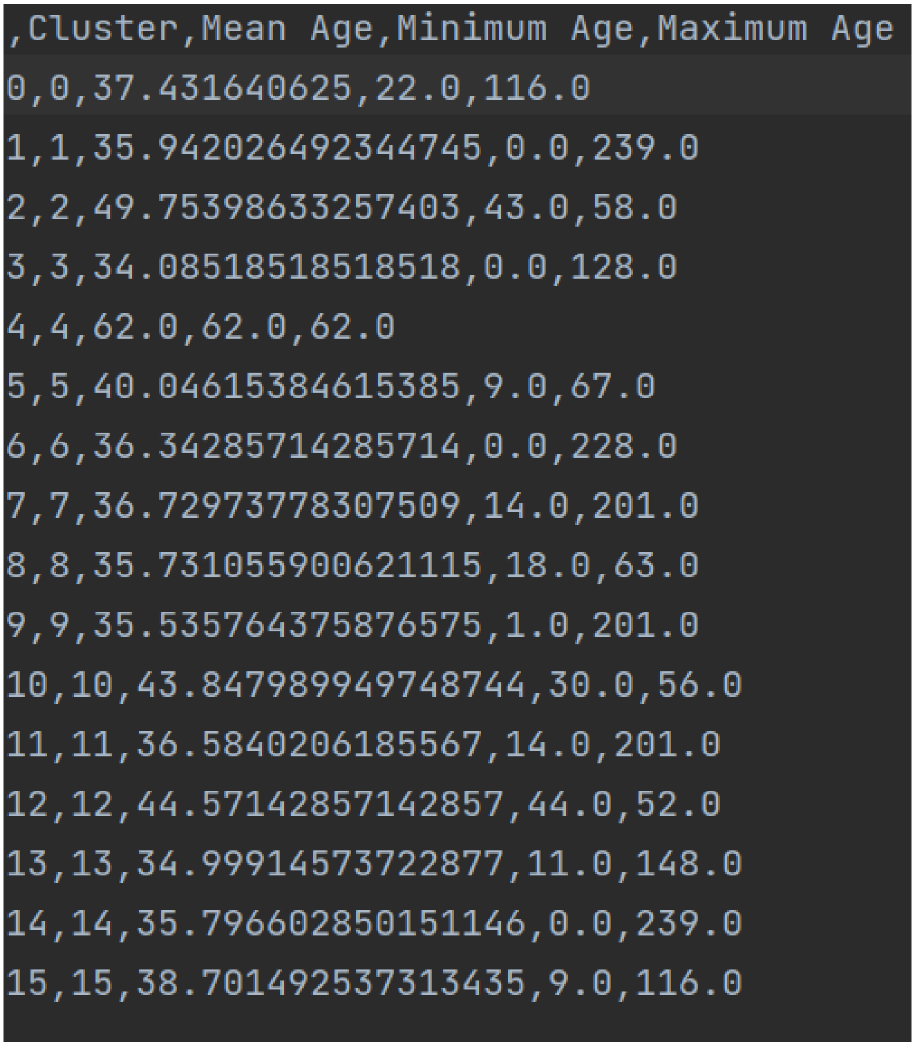
```
892,9,0375725784,8.5
922,9,0439136369,10.0
923,9,043935806X,8.0
1067,10,0345339681,6.375
1116,10,0439136369,7.0
1185,10,0679746048,6.666666666666667
1233,11,0385484518,6.425531914893617
1235,11,0439064872,6.917197452229299
1251,11,0446310786,7.142857142857143
1274,12,0345339681,5.0
1378,13,0345339681,7.03125
1425,13,0439136369,7.0
1426,13,043935806X,7.5
1520,14,0060987529,9.0
1584,14,0380789035,10.0
1646,14,0449005615,9.0
1700,15,0971880107,1.2424242424242424
```

(Top 3 scoring ISBN's in each cluster)



```
,Cluster,ISBN,Book-Rating,Book-Title
0,0,0345342968,6.611111111111111,Fahrenheit 451
1,0,0684874350,7.0,ANGELA'S ASHES
2,0,0767902521,9.0,A Walk in the Woods: Rediscovering America on the Appalachian Trail (Official Guides to the Appalachian Trail)
3,1,0345339681,7.528735632183908,The Hobbit : The Enchanting Prelude to The Lord of the Rings
4,1,0439136369,7.9393939393939394,Harry Potter and the Prisoner of Azkaban (Book 3)
5,1,043935806X,7.901408450704225,Harry Potter and the Order of the Phoenix (Book 5)
6,2,0312278586,10.0,The Nanny Diaries: A Novel
7,2,0380789019,10.0,Neverwhere
8,2,0385505833,10.0,Skipping Christmas
9,3,0971880107,1.6014814814814815,Wild Animus
10,4,0345339681,5.0,The Hobbit : The Enchanting Prelude to The Lord of the Rings
11,5,0375725784,8.5,A Heartbreaking Work of Staggering Genius
12,5,0439136369,10.0,Harry Potter and the Prisoner of Azkaban (Book 3)
13,5,043935806X,8.0,Harry Potter and the Order of the Phoenix (Book 5)
14,6,0142001740,6.326996197718631,The Secret Life of Bees
15,6,0446672211,6.775510204081633,Where the Heart Is (Oprah's Book Club (Paperback))
16,6,0679781587,6.344827586206897,
17,7,0060938455,6.6875,Fast Food Nation: The Dark Side of the All-American Meal
18,7,0345339681,6.928571428571429,The Hobbit : The Enchanting Prelude to The Lord of the Rings
19,7,0440226430,6.5,Summer Sisters
20,8,0345339681,6.375,The Hobbit : The Enchanting Prelude to The Lord of the Rings
21,8,0439136369,7.0,Harry Potter and the Prisoner of Azkaban (Book 3)
22,8,0679746048,6.666666666666667,"Girl, Interrupted"
23,9,0316666343,6.060308555399719,The Lovely Bones: A Novel
24,10,0060987529,9.0,Confessions of an Ugly Stepsister : A Novel
25,10,0380789035,10.0,American Gods
26,10,0449005615,9.0,Seabiscuit: An American Legend
27,11,0385504209,5.735294117647059,The Da Vinci Code
28,11,059035342X,5.790123456790123,Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))
29,11,0679781587,5.621621621621622,
30,13,0345339681,7.03125,The Hobbit : The Enchanting Prelude to The Lord of the Rings
31,13,0439136369,6.931034482758621,Harry Potter and the Prisoner of Azkaban (Book 3)
32,13,043935806X,7.545454545454546,Harry Potter and the Order of the Phoenix (Book 5)
33,14,0385484518,6.470588235294118,"Tuesdays with Morrie: An Old Man, a Young Man, and Life's Greatest Lesson"
34,14,0446310786,7.19672131147541,To Kill a Mockingbird
35,14,059035342X,7.109704641350211,Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))
36,15,0971880107,1.2388059701492538,Wild Animus
```

(Book titles of the top 3 scoring ISBN's in each cluster)



```
,Cluster,Mean Age,Minimum Age,Maximum Age
0,0,37.431640625,22.0,116.0
1,1,35.942026492344745,0.0,239.0
2,2,49.75398633257403,43.0,58.0
3,3,34.08518518518518,0.0,128.0
4,4,62.0,62.0,62.0
5,5,40.04615384615385,9.0,67.0
6,6,36.342857142857714,0.0,228.0
7,7,36.72973778307509,14.0,201.0
8,8,35.731055900621115,18.0,63.0
9,9,35.535764375876575,1.0,201.0
10,10,43.847989949748744,30.0,56.0
11,11,36.5840206185567,14.0,201.0
12,12,44.57142857142857,44.0,52.0
13,13,34.99914573722877,11.0,148.0
14,14,35.796602850151146,0.0,239.0
15,15,38.701492537313435,9.0,116.0
```

(Minimum, Mean, and Maximum ages of each clusters)

The results seem to be pretty reasonable. Many of the books seen are extremely popular, such as the Harry Potter series, The Da Vinci Code, To Kill A Mocking Bird, The Hobbit and more. After seeing the titles of the books with the highest average ratings in its respective cluster, the results are pretty reasonable.

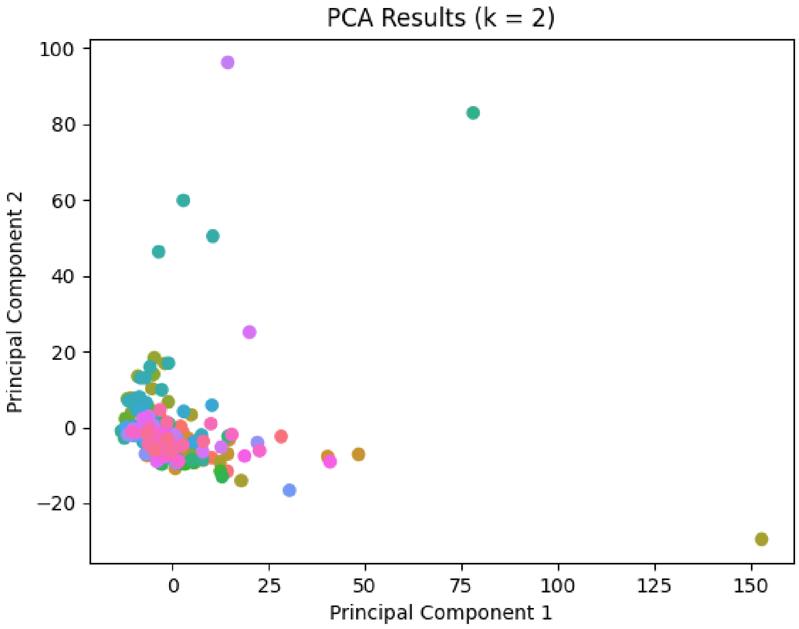
It seems that the ages of each cluster varies. Some of the minimum and maximums of each cluster seems to be way outside normal age range so I took the mean age of each cluster in order to see if we could see more patterns. It seems that cluster 4 has the oldest age of 62 years old with cluster 2 taking the next spot with a mean age of 49. Most of the ages seem to vary between 35 - 45.

3.3 Principle Component Analysis

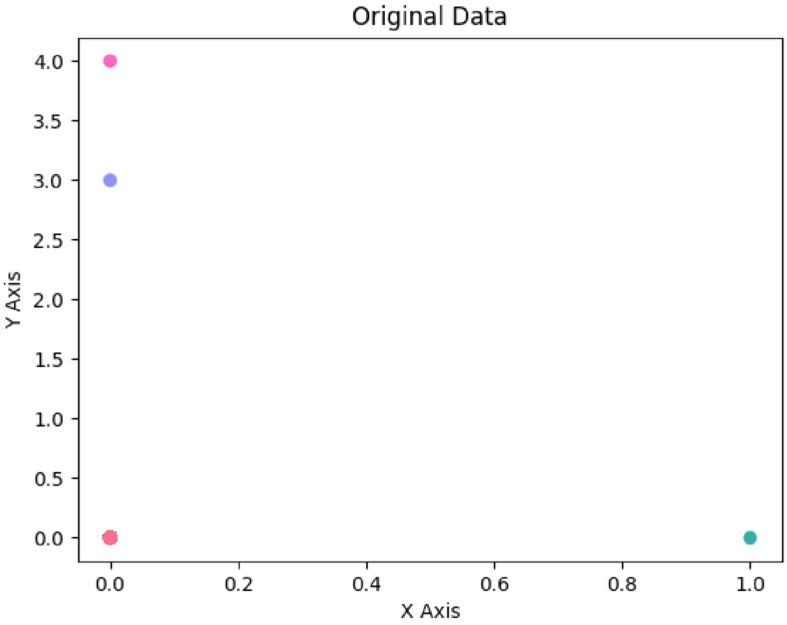
For the Principle Component Analysis, I first transposed the matrix from question one so that the rows would refer to books and columns would refer to users. I mean centered the data in order to ensure that the influence of the mean is removed and to find the most significant patterns in the dataset.

After this, I applied PCA with the number of components k = 2 to reduce the dimensionality of the books in order to gather more insight in regards to the data.

I then plotted the results and colored each one based on the book genre of the book which is seen in the following:



(PCA results)



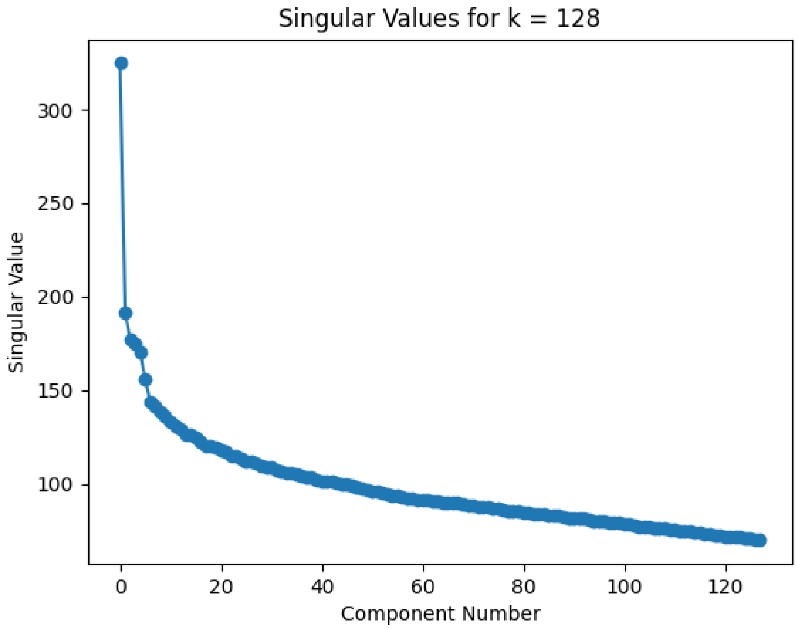
(Original data)

The patterns seen within the plot of the PCA results from the books is very condensed. It seems to have a slight negative linear correlation, however, overall there doesn't seem to have much variance in the data. But, if we compare this to the original data, we can see a significant difference in the variance. This could mean that the data that was collected concentrated on a specific genre of books or maybe we should use a higher number of components within the data.

The number of principle components needed to explain 80 percent of the variance of the data is 1 and also the number of principle components needed to explain 40 percent of the variance of the data is also 1. I found this using the explained variance ratio. This suggests that the data has a low intrinsic dimensionality and a single principle component captures a substantial amount of the variability. In comparison to k = 2, the k value of 1 might be more beneficial in order to produce a visualization regarding the PCA results.

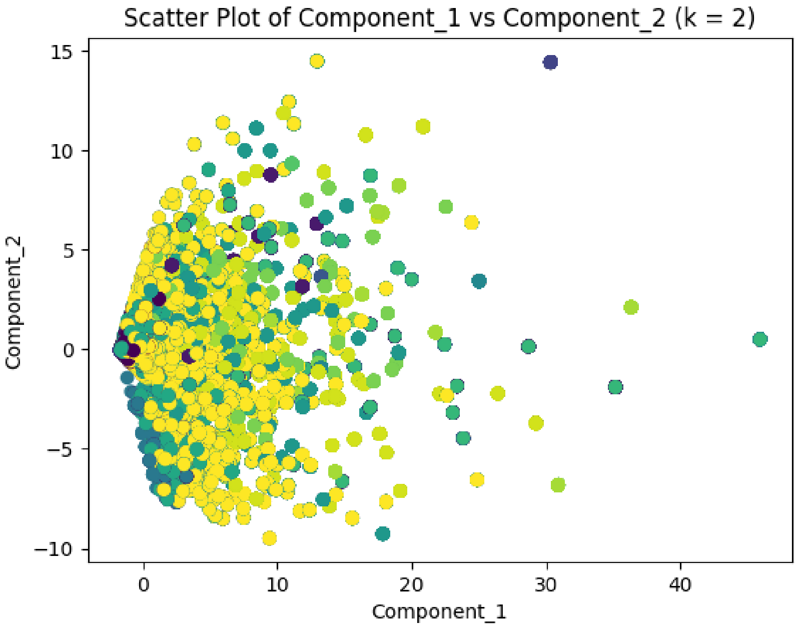
3.4 Singular Value Decomposition

For the Singular Value Decomposition analysis, I utilized the matrix from Q1 and applied SVD with number of components in k = 128 which resulted in the following plot of singular_values:



(k = 128 results)

For each of the values of k = [2, 4, 8, 16, 32, 64, 128], the sum of the explained_variance_ratio resulted in the [0.08676927318015909, 0.12460236107774864, 0.18188566173159, 0.265204798979922944, 0.38980408051954246, 0.5765546556100236, 0.8345519955655069], in which each entry corresponds to its respective k value. In comparison to the inertia values computed in Q2, the values have a significant change on k = 16 which supports my answer of k = 16. It is also seen in the graph where the points start to plateau after k = 16.



(Coloring according to clusters results)

It seems that the SVD plot analysis is pretty similar to the PCA plot. The data points seem to spread out from 0 which suggests a low intrinsic dimensionality. This is supported by the results of the explained variance ratio, where a small number of components can capture a significant proportion of the variability in the data.

The SVD analysis reinforces the findings from previous analyses, providing insights into the intrinsic dimensionality and patterns within the data. The close agreement between the inertia values, explained variance, and clustering results adds confidence to the choice of k = 16 in capturing meaningful information within the dataset.

4 Results

4.1 Transforming Data

The data transformation process involved handling the BX-Book-Ratings.csv dataset. Zero ratings were replaced with the mean rating of the respective book. A subset of the data was selected, focusing on books rated at least 200 times and users who had rated at least 5 books. The resulting user-book ratings matrix formed the basis for further analysis.

Key Findings:

- The top three books based on the number of users who rated them were "Wild Animus," "The Lovely Bones: A Novel," and "The Da Vinci Code."
- The top three users with the highest number of ratings were User 11676, User 198711, and User 153662.

4.2 Clustering

Applying k-means clustering to the dataset with values of k ranging from 2 to 128 revealed a clear elbow point at k = 16. Using this value, the top three books with the highest average ratings within each cluster were identified. Additionally, an analysis of the age distribution within each cluster showcased variations in age ranges.

Key Findings:

- Cluster 4 had the oldest mean age at 62, and cluster 2 followed with a mean age of 49.
- Most clusters exhibited a mean age ranging between 35 and 45.

4.3 Principle Component Analysis

The PCA step involved transposing the matrix, mean-centering the data, and applying PCA with k = 2. The resulting plot, color-coded by book genre, revealed a condensed pattern with a slight negative linear correlation. However, comparison with the original data plot indicated significant differences in variance. The intrinsic dimensionality of the data was found to be low, suggesting the potential benefit of using a single principal component for visualization.

Key Findings:

- A single principal component could capture 80 percent and 40 percent of the variance.

4.4 Singular Value Decomposition

SVD analysis, performed on the matrix from the initial transformation with k = 128, provided insights into the intrinsic dimensionality of the data. The sum of explained variance ratios for various values of k supported the choice of k = 16, consistent with clustering results. The SVD analysis showcased the low intrinsic dimensionality observed in the PCA results.

- The choice of k = 16 was supported by both inertia values and explained variance ratios.

4.5 Summary

In summary, the results from data transformation, clustering, PCA, and SVD provided an understanding of the Book-Crossing datasets. The chosen value of k = 16 was consistently chosen as the best k value across the analyses. The age analysis within clusters, along with insights into intrinsic dimensionality, contributes valuable information for future recommendations and system improvements. The findings lay the groundwork for more understanding and potential enhancements in the book recommendation system.

5 Conclusion:

In this investigation of the Book-Crossing datasets, we obtained many different insights into user behavior and book preferences. The identified clusters, particularly with a focus on age and other demographics, provided a foundation for future book recommendation systems. The low intrinsic dimensionality observed in both PCA and SVD analyses suggests the potential effectiveness of a streamlined approach to capture user preferences. These findings not only contribute to the creation of a more efficient book recommendation system but also gives insights in regards to how readers engage with literary content. By using these insights, the goal is to improve user satisfaction and encourage exploration of different literary genres.

6 References

Improving Recommendation Lists Through Topic Diversification, Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, Georg Lausen; Proceedings of the 14th International World Wide Web Conference (WWW '05), May 10-14, 2005, Chiba, Japan. To appear

Ziegler, C. (n.d.). Book-Crossing Dataset. University of Freiburg. Retrieved November 20, 2023, from http://www2.informatik.uni-freiburg.de/ cziegler/BX/