A Big Data project

Spark Your Imagination

Madhuri Mahalingam - mm13575
Swarali Dabhadkar - sd5664
Kartikey Sharma - ks7154

Outline

01.
Temporal Analysis
Analysis of goodreads datasets, including genre and author trends over time

02.
Genre Cluster Analysis
Analysis of books based on genre-tagging

03.
Sentiment Analysis
On book reviews based on different factors

04.
Recommendation system
Content based and collaborative

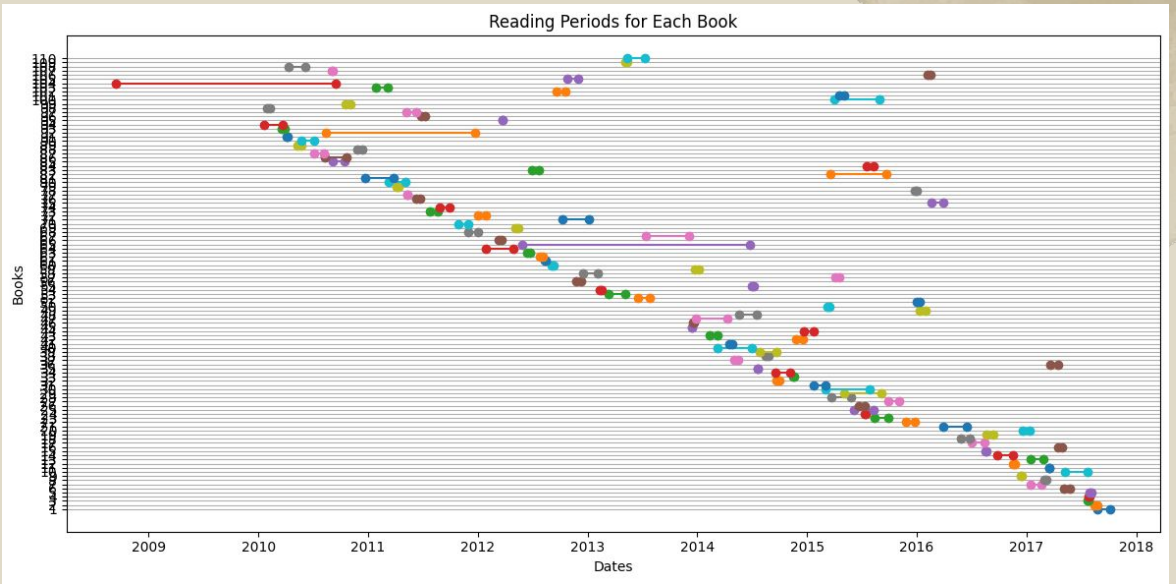
05.
Conclusion and Future Scope
Results and the big picture

01

Temporal Analysis

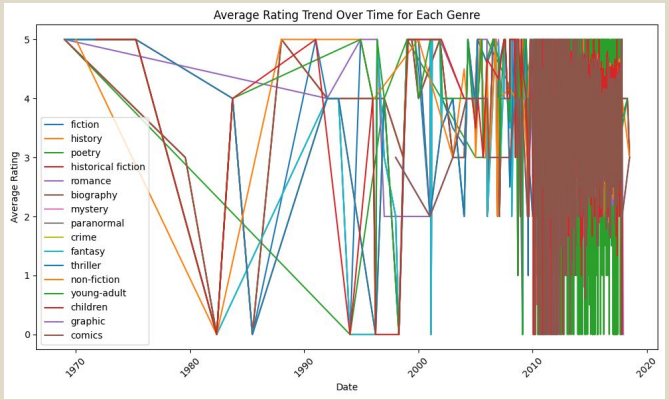
User history

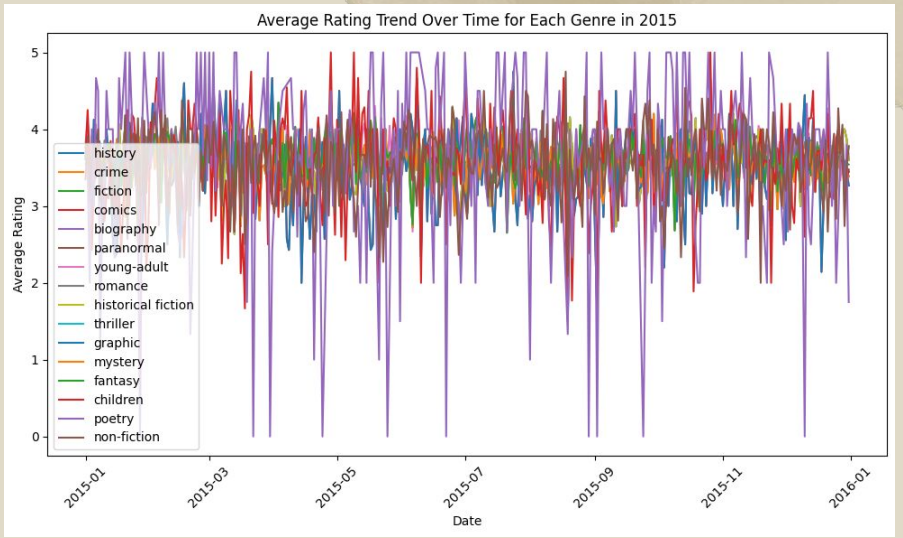
Visualize the user's reading trends over time, sort of a 'GOODREADS WRAPPED'
Can be extrapolated to showcase genres too



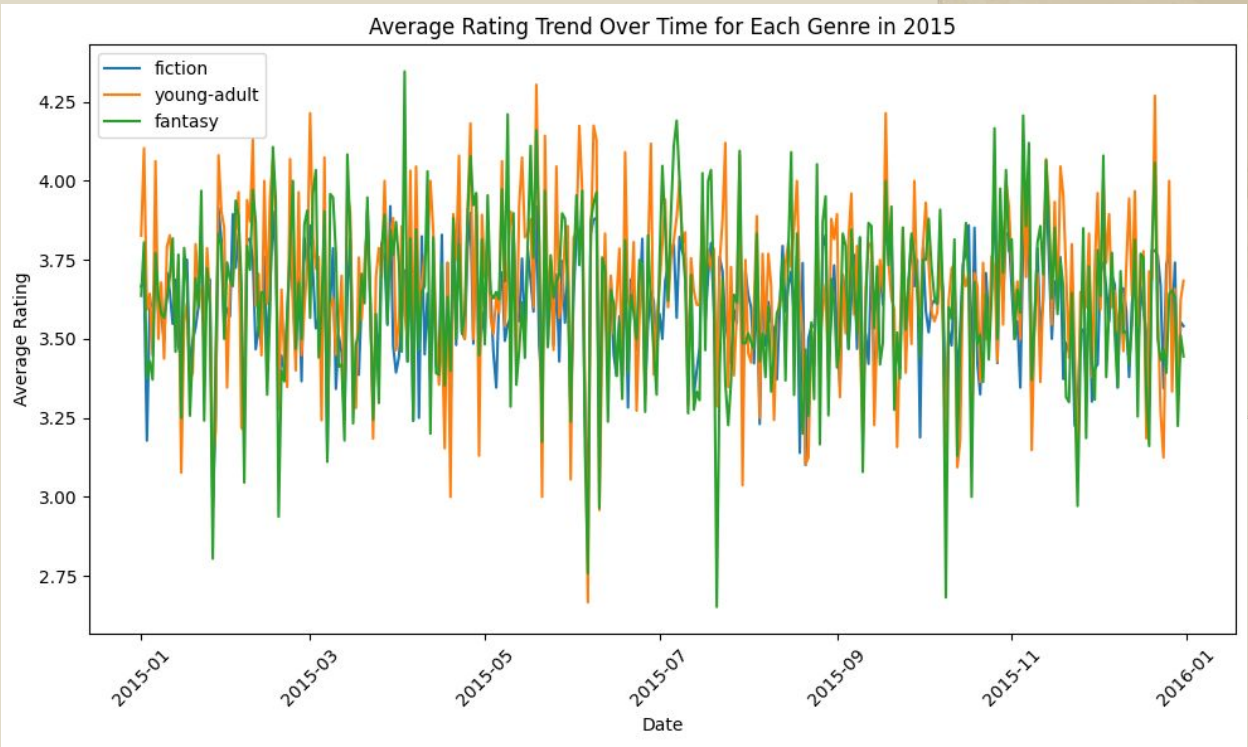
Genre based Analysis

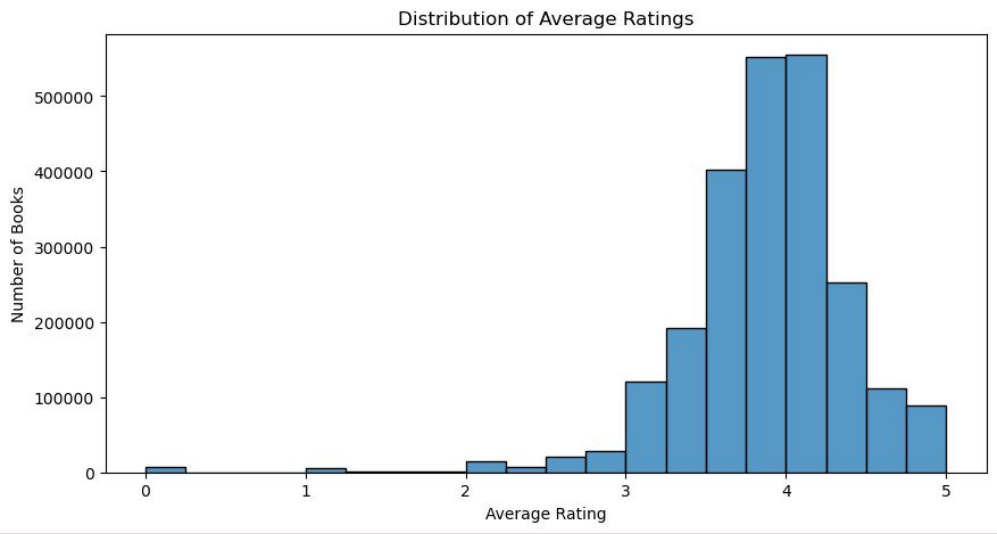
- Significant variability in average ratings over time for all genres.
- Sharp spikes could indicate data anomalies, or actually represent the release dates of popular books.

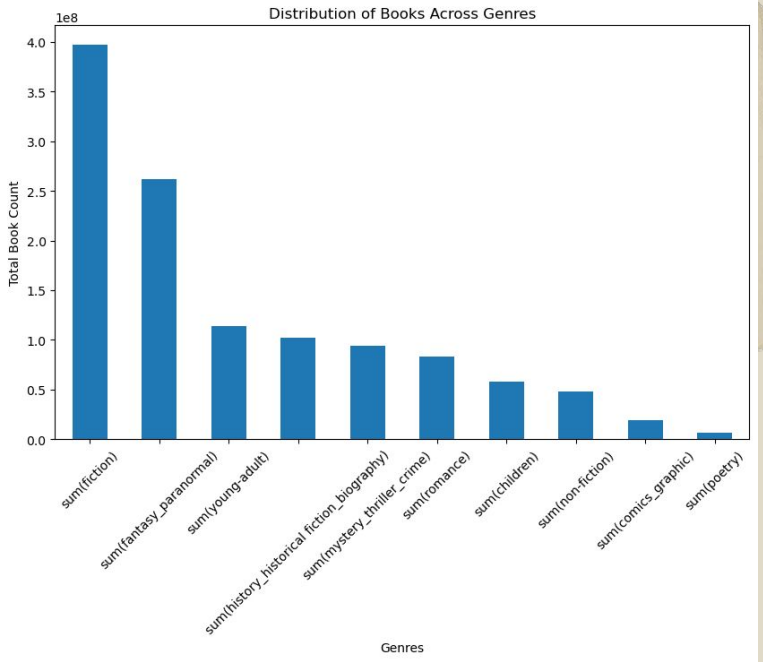


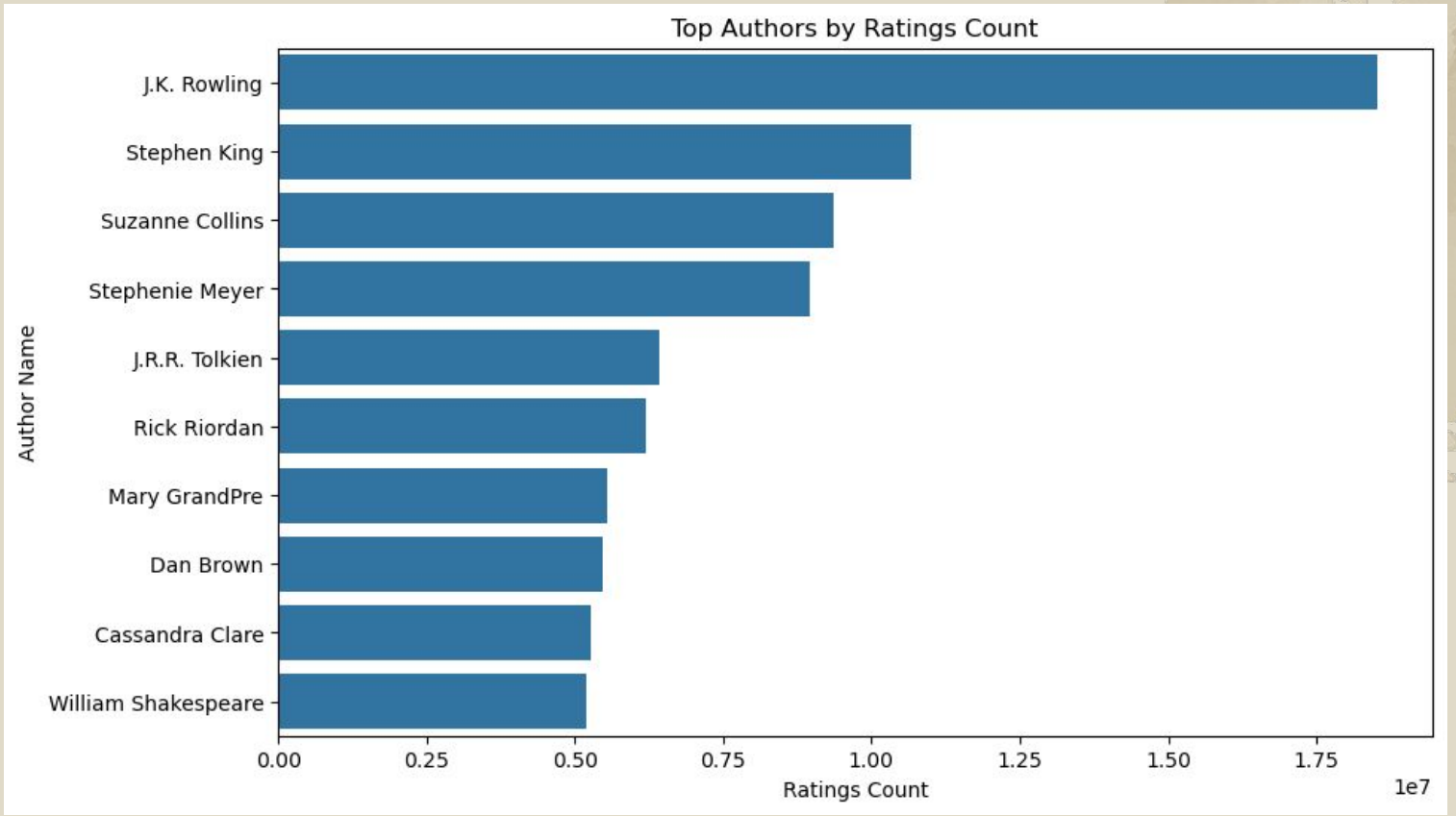


Sharp spikes or drops in ratings could indicate data anomalies or they could represent moments when a particularly influential book was published.



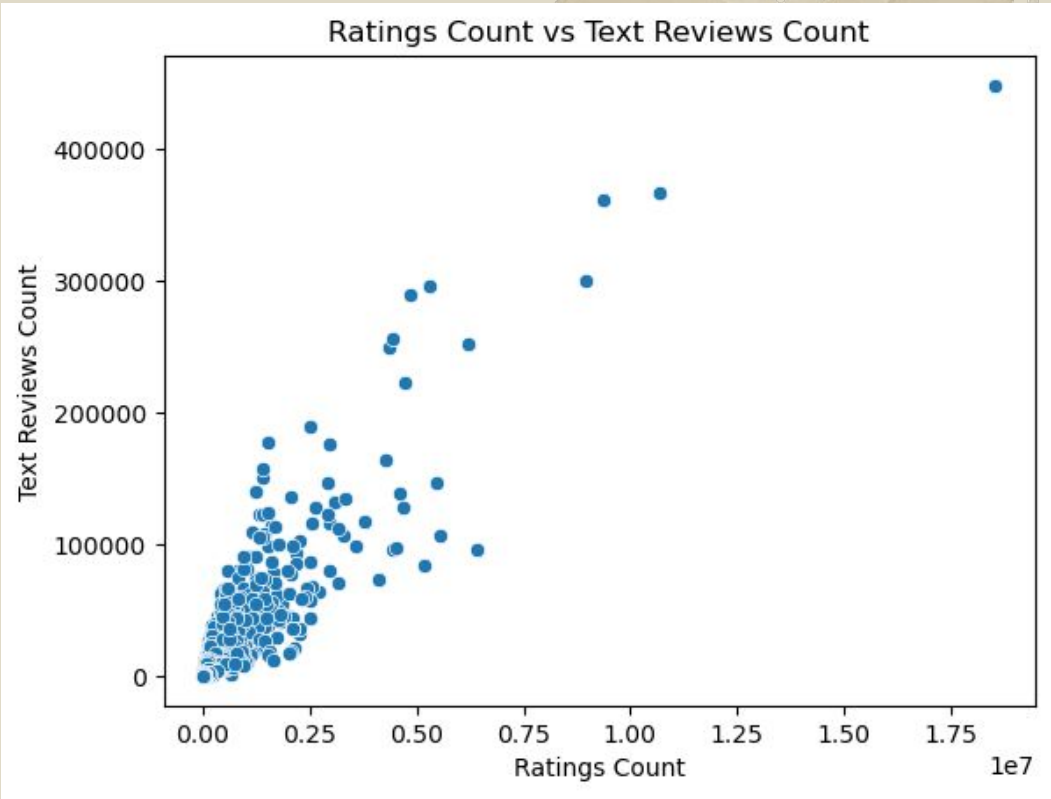

Distribution of Average Ratings
Number of Books
Average Rating
0
100000
200000
300000
400000
500000
0
1
2
3
4
5


Distribution of Books Across Genres
1e8
Total Book Count
Genres
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
sum(fiction)
sum(fantasy_paranormal)
sum(young adult)
sum(history_historical fiction_biography)
sum(mystery_thriller_crime)
sum(romance)
sum(children)
sum(non-fiction)
sum(comics_graphic)
sum(poetry)

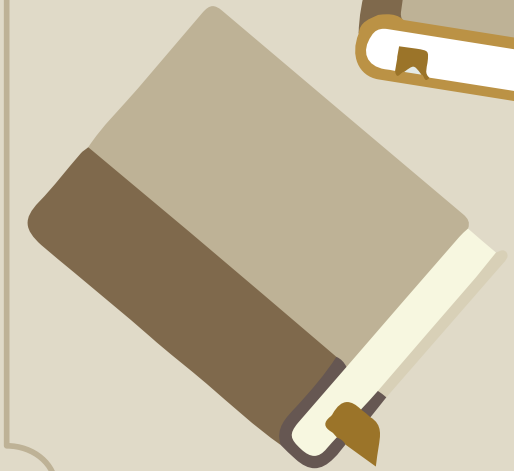

Top Authors by Ratings Count
J.K. Rowling
Stephen King
Suzanne Collins
Stephenie Meyer
J.R.R. Tolkien
Rick Riordan
Mary GrandPre
Dan Brown
Cassandra Clare
William Shakespeare
Author Name
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Ratings Count
1e7

X increase => Y increase (generally)

Dense data points on the lower side of the graph



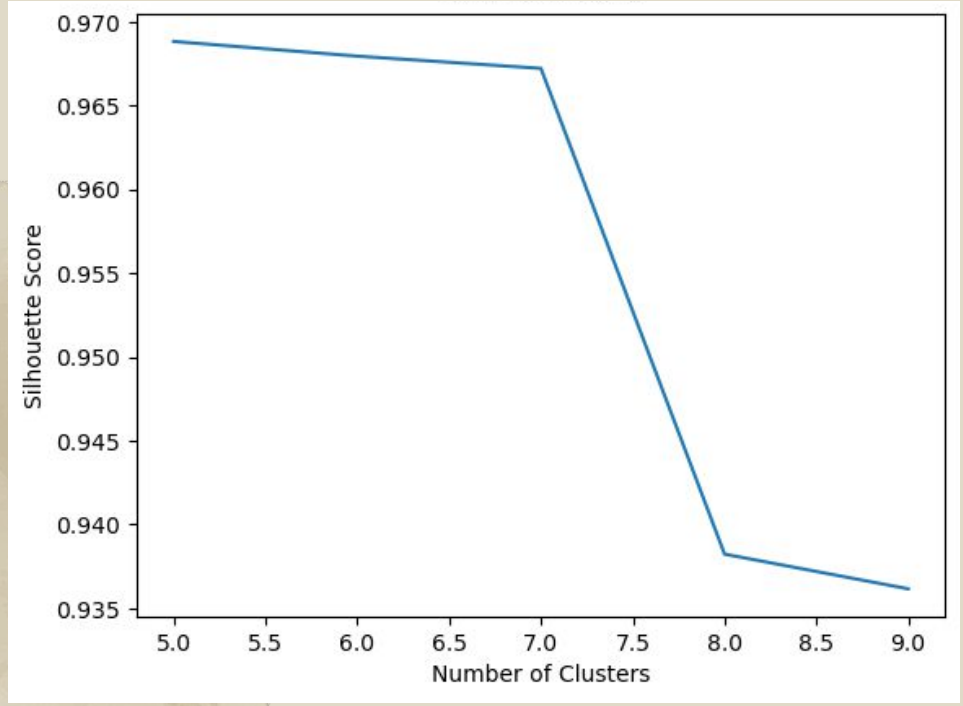
02 Genre Cluster Analysis

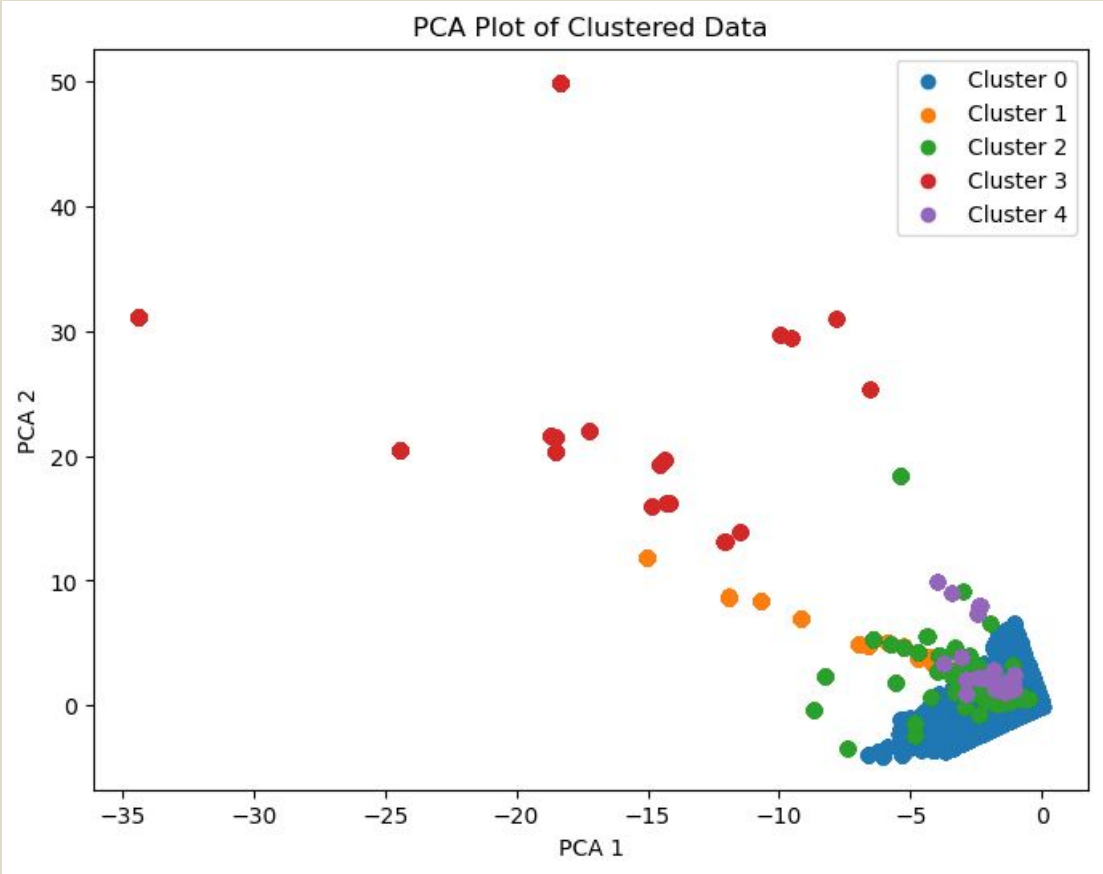


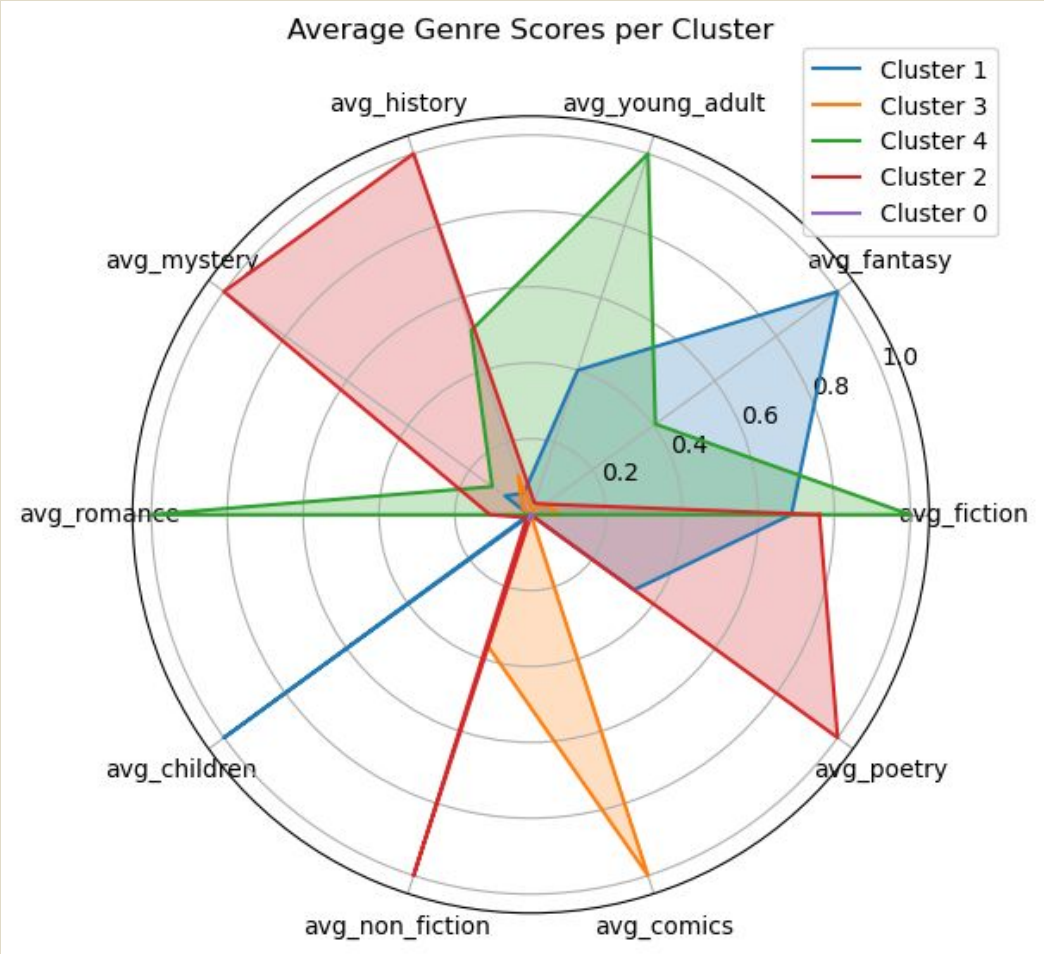
- After EDA, we found that the dataset contains 10 distinct genres, however each book is tagged under multiple genres by users
- We attempt to find some clustering of books according to genres to gain insights
- Spark libraries used: ClusteringEvaluator from pyspark.ml.evaluation, Kmeans from pyspark.ml.clustering, PCA from pyspark.ml.feature

Optimization for k

- Start with k = 10 as there are 10 distinct genres
- Apply silhouette method to find an optimal ‘k’
- Visualize the clusters to gain insights regarding related genres




PCA Plot of Clustered Data
Cluster 0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
PCA 2
PCA 1
50
40
30
20
10
0
−35
−30
−25
−20
−15
−10
−5
0


Average Genre Scores per Cluster
Cluster 1
Cluster 3
Cluster 4
Cluster 2
Cluster 0
avg_history
avg_young_adult
avg_mystery
avg_fantasy
0.2
0.4
0.6
0.8
1.0
avg_romance
avg_fiction
avg_children
avg_poetry
avg_non_fiction
avg_comics

03 Sentiment Analysis

Sentiment analysis performed by:

- Rating
- Review length
- Number of votes and comments

- Library used: A UDF on TextBlob.sentiment.polarity
- The provided analyses give insights into the sentiment, review length, and reader engagement patterns for reviews of different ratings on Goodreads.

Observations

rating	avg(sentiment)
0	0.16896166058394 17
1	-0.04232273559161615
2	0.045763957237147 33
3	0.12798248384417 61
4	0.19277641212956653
5	0.25372733307540 35

rating	avg(review_length)
0	463.11227154047
1	577.0843373493976
2	585.0147058823529
3	518.5528775209051
4	742.3339800443459
5	760.4680013127667

Observations

rating	avg(n_votes)
0	0.720626631853786
1	0.5421686746987951
2	0.5632352941176471
3	0.41859321200196753
4	1.3586474501108647
5	2.21168362323597

rating	avg(n_comments)
0	0.4177545691906005
1	0.3453815261044177
2	0.2352941176470588
3	0.11018199704869651
4	0.20953436807095344
5	0.373153921890384

04 Recommendation System

Content based

Tokenized descriptions and applied HashingTF and IDF for feature extraction

Collaborative

Used ALS algorithm and transformed data into a user=item interaction matrix

Content-based recommendation

- Approximate similarity joins were performed using LSH
- Why LSH?
  - Efficient for large-scale datasets
  - Approximate nearest neighbors searches without exhaustive pairwise comparisons.

	book_id_A	title_A	description_A	book_id_B	title_B	description_B	distCol
0	13598461	Labor and Desire: Women's Revolutionary Fictio...	"This critical, historical, and theoretical st...	2250580	A.I. Revolution, Vol. 1	Like everyone else in the future, Sui's used t...	1.389598
1	13598461	Labor and Desire: Women's Revolutionary Fictio...	"This critical, historical, and theoretical st...	10862067	Crisis en Tierras Infinitas Absolute	!Edicion especial 20 aniversario de Crisis en ...	1.414214
2	13598461	Labor and Desire: Women's Revolutionary Fictio...	"This critical, historical, and theoretical st...	28422525	B.P.R.D. Inferno sulla Terra, Vol. 3: Russia (...	Un cimitero di non-morti viene scoperto nelle ...	1.413918
3	13598461	Labor and Desire: Women's Revolutionary Fictio...	"This critical, historical, and theoretical st...	6728819	O Principezinho	Considerado um dos grandes classicos da litera...	1.413371
4	13598461	Labor and Desire: Women's Revolutionary Fictio...	"This critical, historical, and theoretical st...	13611715	El exilio (Reinos Olvidados, El Elfo Oscuro, # 2)	Hostiles hasta tal punto que un habitante de l...	1.413912

Collaborative filtering Recommendation

- Collaborative filtering provides personalized recos based on a user's past behavior and the behavior of similar users.
- (RMSE) of 1.61645

```
+-------------+--------------------------------------------------------------------------------------------------------------+
|user_id_index|recommendations                                                                                               |
+-------------+--------------------------------------------------------------------------------------------------------------+
|0            |[{460548, 5.0003824}, {304889, 5.0003824}, {128029, 5.0003824}, {119322, 5.0003824}, {99107, 5.0003824}] |
|1            |[{18767539, 4.995974}, {17939501, 4.995974}, {2718668, 4.995974}, {2088385, 4.995974}, {15881, 4.995974}]|
+-------------+--------------------------------------------------------------------------------------------------------------+
```

Recommendations Dashboard

https://sparkyourimagination.streamlit.app/

Navigation

- Home
- Book Recommender (Collaborative)
- Content-based Recommender

Choose an option from the sidebar to get started.

Limitations and Future Scope

- Data has been sampled in some cases due to resource limitations
- The data was available only till 2017
- GPT integration for an interactive recommendation system
- Real-Time Analysis: Potential integration of real-time streaming data using Kafka.
- Preparing for increasing data volumes and real-time analytics.
- Scale up to a full-fledged recommendation system

Conclusion

Performed various analysis on ~ 100 GB of Goodreads data to gain insights and develop a recommendation system

Why is it Big Data?

- Diverse range of data types
- Data size
- Complex and computationally intensive methods like sentiment analysis and k-means applied to a large dataset

```
Py4JJavaError: An error occurred while calling o158.fit.
: org.apache.spark.SparkException: Job aborted due to stage failure: Task 0 in stage 9.0 failed 1 times, most recent failure: Lost task 0.0 in
ter-mm13575 executor driver): org.apache.spark.SparkException: Kryo serialization failed: Buffer overflow. Available: 0, required: 13619056
Serialization trace:
_values$mcJ$sp (org.apache.spark.util.collection.OpenHashMap$mcJ$sp). To avoid this, increase spark.kryoserializer.buffer.max value.
```

Big Data technology used: Spark

Thanks

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**