BOOKLAB

HELPING YOU DISCOVER BOOKS YOU WILL LOVE TO READ

OUTLINE

- Project Motivation
- Dataset & Pipeline
- Reader Insights [EDA]
- Supervised ML: Predicting Book Approval Rating
- Unsupervised ML: Book Recommendation Algorithm
- Live Prototype Demo
- Business Application

http://blog.nycdatascience.com/student-works/book-rating-prediction-recommendation-engine/

PROJECT MOTIVATION

- Everybody Loves a Good Story
- Target Audience
- Goals
- Languages, Tools, Platforms

MARKET-LEADER VS TRADITIONAL BRICK & MORTAR

www.barnesandnoble.com/s/harry+potter/_/N-8q8?_requestid=1643192

BARNES & NOBLE

harry potter Books

Spend $25, Get Free Shipping

Sign In My Account

Membership Gift Cards Stores & Events Help

Books NOOK Books NOOK Textbooks Newsstand Teens Kids Toys & Games Hobbies & Collectibles Home & Gifts Movies & TV Music Sale

Order Your Gifts in Time for the Holidays

STANDARD DELIVERY 00:16:15:14 DAYS HRS MINS SEC

EXPRESS DELIVERY 01:16:15:14 DAYS HRS MINS SEC

EXPEDITED DELIVERY 03:16:15:14 DAYS HRS MINS SEC

× Books

1 - 20 of 1111 results for harry potter Sort by: Top Matches Results: 20

SUBJECTS

- Activity & Game Books
- Antiques & Collectibles
- Awards
- Bibles & Christianity
- Biography
- Cookbooks, Food & Wine
- Fiction
- Kids
- Literature
- Music, Film & Performing Arts
- Reference
- Religion

THE HARRY POTTER COLLECTION

Harry Potter and the Chamber…
by J. K. Rowling
★★★★★

Harry Potter and the Cursed…
by J. K. Rowling
★★★½☆

Harry Potter and the…
by J. K. Rowling
★★★★★

Harry Potter Paperback Boxed…
by J. K. Rowling
★★★★½

The Unofficial Harry Potter…
by Dinah Buchotz
★★★★☆

www.strandbooks.com/books

18 MILES OF BOOKS STRAND NEW YORK CITY • EST. 1927

ABOUT STRAND HOURS & LOCATIONS SELL YOUR BOOKS WISH LISTS MY ACCOUNT

SEARCH ADVANCED SEARCH

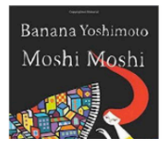
SHOPPING BAG (0)

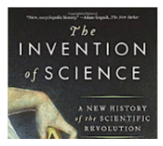
TOP PICKS | BROWSE BOOKS & MEDIA | RARE & COLLECTIBLES | KIDS & Y.A. | BOOKS BY THE FOOT | STAFF PICKS | GIFTS | GIFT CARDS | EVENTS

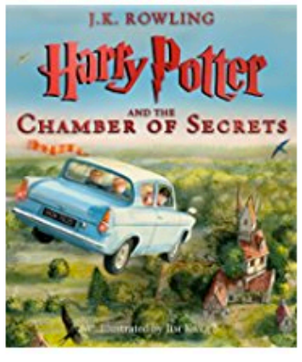
- NEW ARRIVALS IN BOOKS
- NEW ARRIVALS IN GIFTS
- BESTSELLERS
- SIGNED NEW EDITIONS
- STRAND SUBSCRIPTIONS
- STAFF PICKS
- LOWER PRICED THAN E-BOOKS
- THE AUTHOR 'S BOOKSHELF
- STRAND 80
- UPCOMING RELEASES

New Arrivals

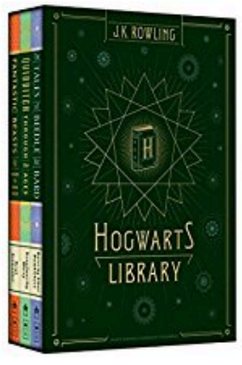
See All New Arrivals ▶



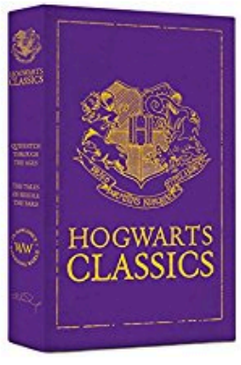




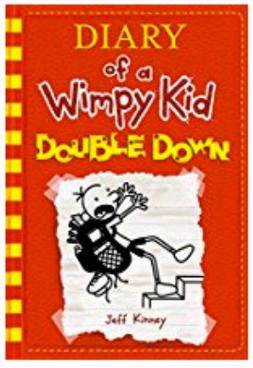
Harry Potter and the Chamber of Secrets: The Illustrated Edition…
› J.K. Rowling
★★★★★ 305
$23.99 Prime



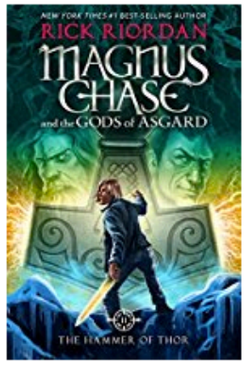
Hogwarts Library
› J.K. Rowling
★★★★★ 911
$24.85 Prime



Hogwarts Classics (Harry Potter)
› J.K. Rowling
★★★★½ 165
$11.99 Prime



Double Down (Diary of a Wimpy Kid Book 11)
New Release
› Jeff Kinney
★★★★½ 433
$7.52



Magnus Chase and the Gods of Asgard, Book 2: The Hammer of Thor
› Rick Riordan
★★★★½ 394
$9.99

GOAL

help Barnes and Noble's customers find great books that will inspire them, make them laugh, make them cry, and invoke their curiosity.

METHODOLOGY & TECHNIQUES

- Use analyze reader behavior and preferences using EDA and clustering
- develop a machine learning algorithm that predicts the reader satisfaction rate for books
- create a recommendation engine algorithm to select the top matches for their needs design a customer friendly interface that can be used by Bookstore specialists and customers

LANGUAGES , TOOLS, PLATFORMS

- Languages: R, Python
- Platform: Spark , Data Science Studio, Graph Lab
- API: Google , GoodReads

SOURCES

- Data: Universität Freiburg, bookcrossing.com
- Collaborative Intelligence

DATASET

FEATURES	TOTAL	MISSING	UNIQUE
ISBN	1,149,780		340,556 271,379
TITLES	271,379		242,154
AUTHORS	271,379		102,028
PUBLISHERS	271,379		16,806
PUBLICATION YEAR	271,377	2	116
USERID	1,149,780		105,283
AGE	168,096	110,762	
LOCATION	278,858		
RATINGS	1,149,780		
BOOK IMAGES	271,379		271,063

BOOK-CROSSING BACKGROUND

members connect with other readers, journal and review literature and trade and follow their books as lives are changed through “reading and releasing“

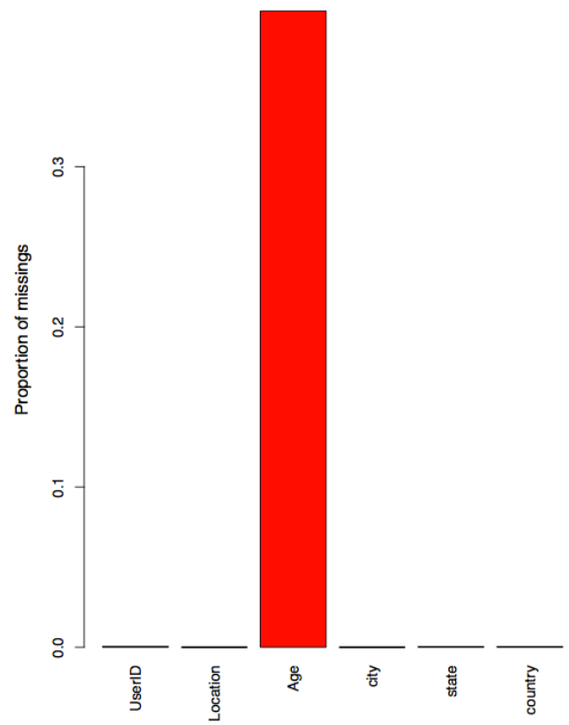
PIPELINE

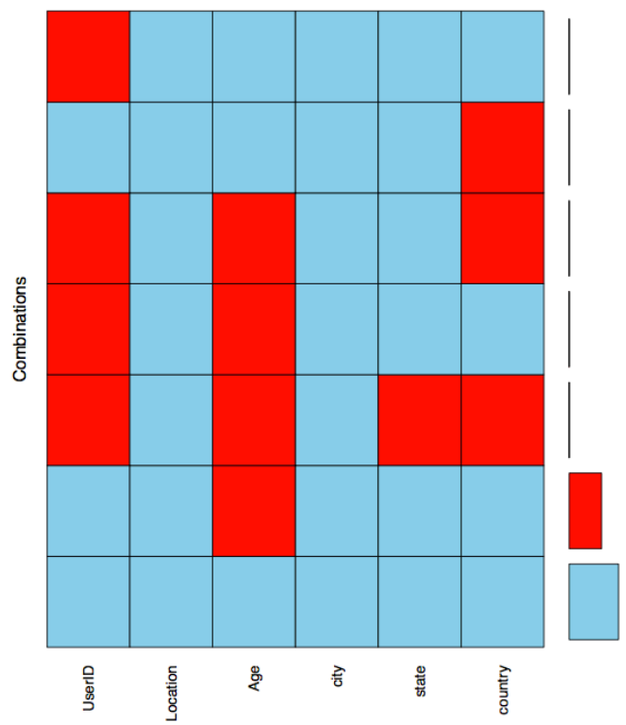
- **Preprocessing**: cleaning, split rated and un-rated
- **EDA** : original, cleaned, supplemented set
- **Feature Creation**: API
- **Supervised Machine Learning:** Predicting Book Approval Rating
- **Unsupervised Machine Learning**: Book Recommendation engine

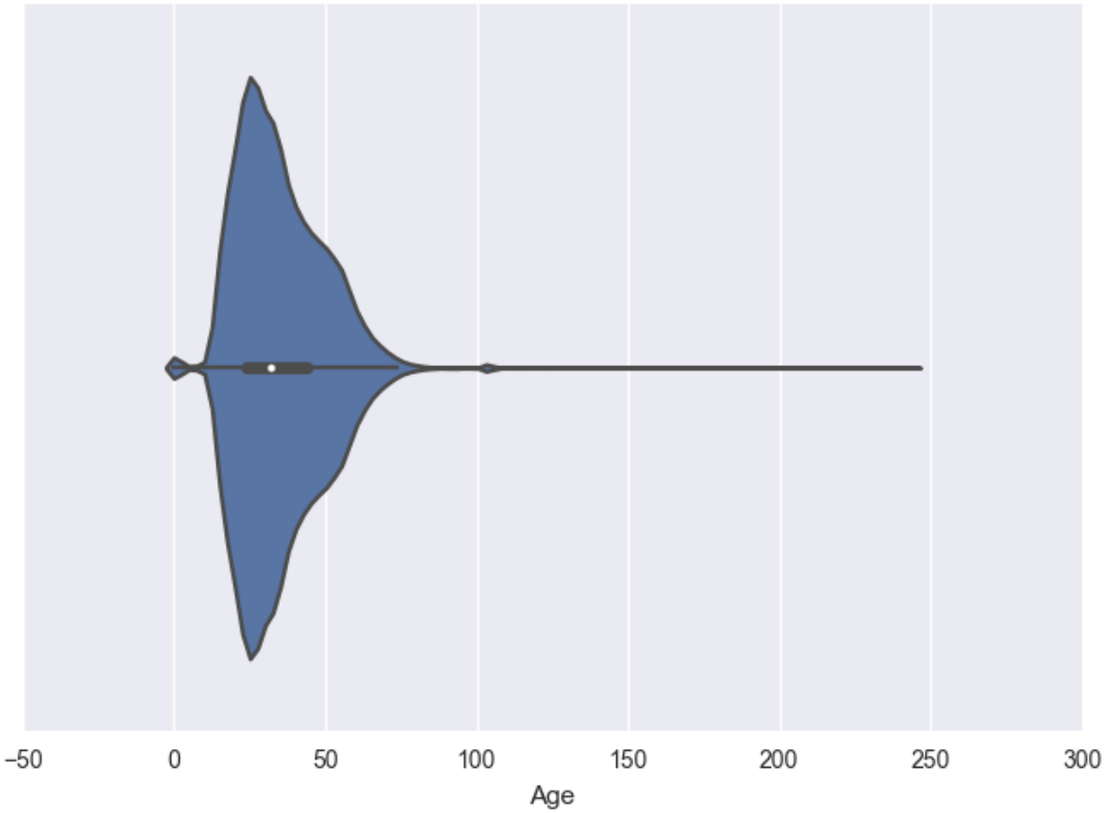
DATA CHALLENGES SUMMARY

- formatting issues such as misspelled City, State, and Country information
- Book title issues especially for non-English titles
- Missing user data such as age and location (City, State, and Country)
- Read but unrated books exceeding read and rated books
- Unreal user age exceeding 100 years old all the way to 250 years old

DATA CHALLENGES: MISSINGNESS







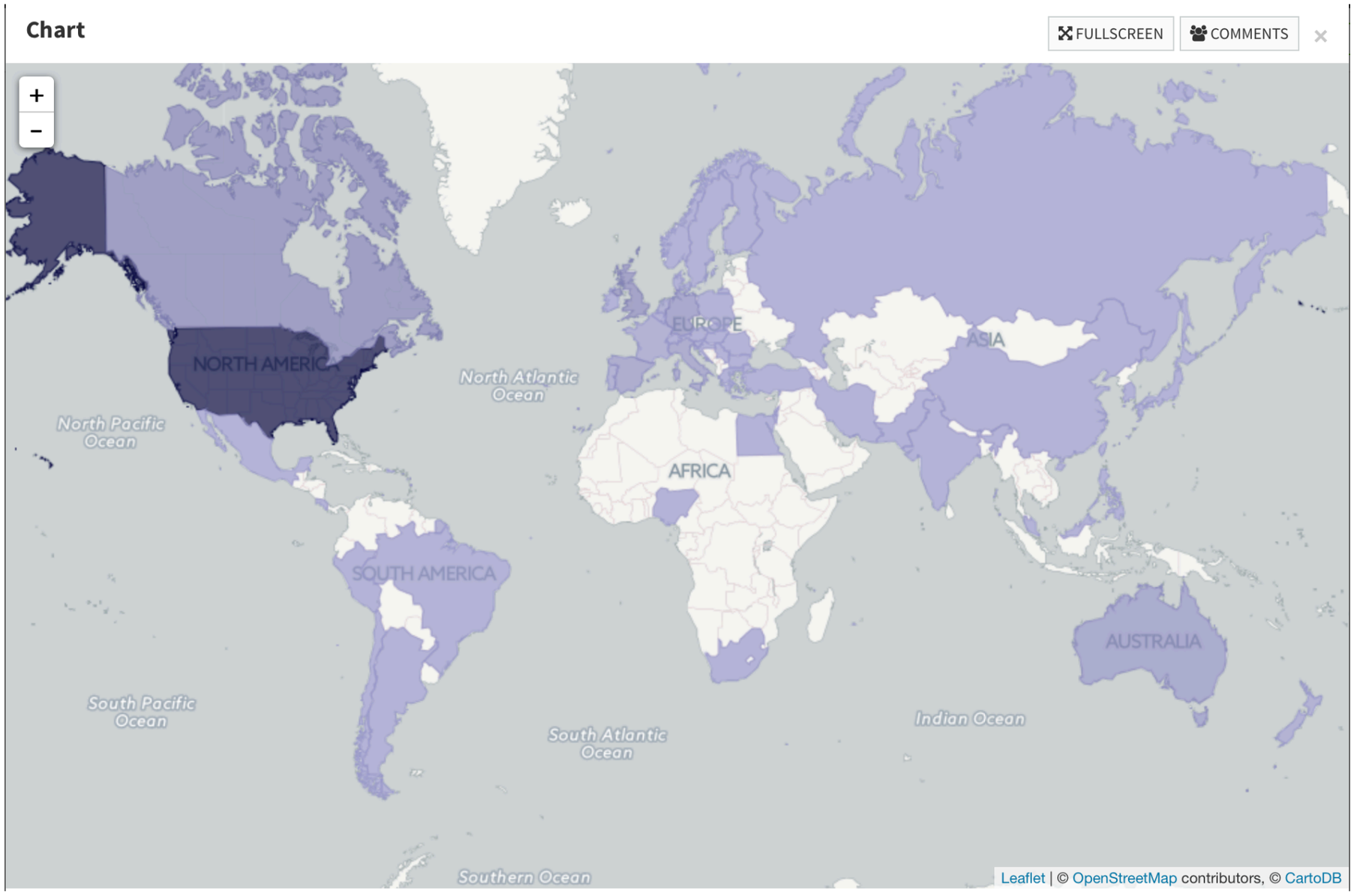
- Age: Missing at random
- Similar to any social media websites, users are not incentivized for providing truthful information nor are they penalized for reporting false ones.

MORE DATA WITH GOOGLE API

- Book Genre
- Page Count
- Maturity

READER INSIGHTS: GLOBAL READERS

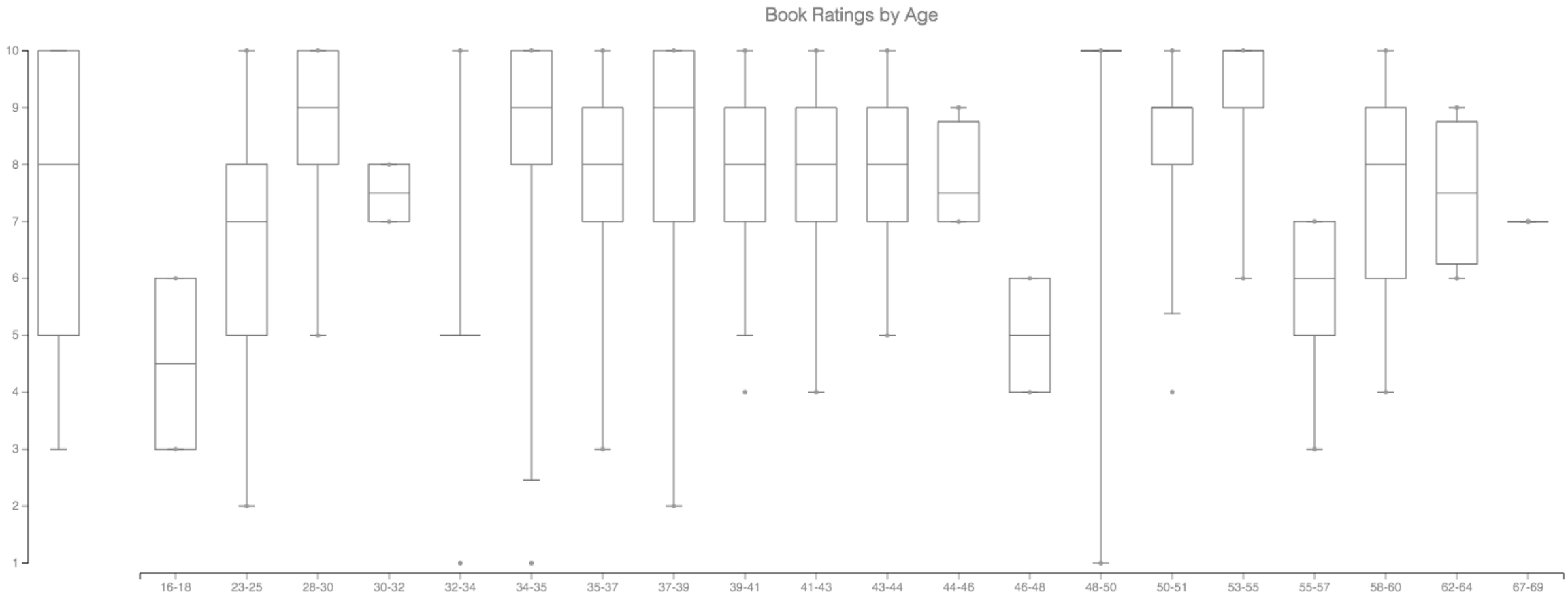
Book readers from all over the world are attracted to join online social book-sharing websites



- Majority from US
- Some readers from African countries

READER INSIGHTS: EDA

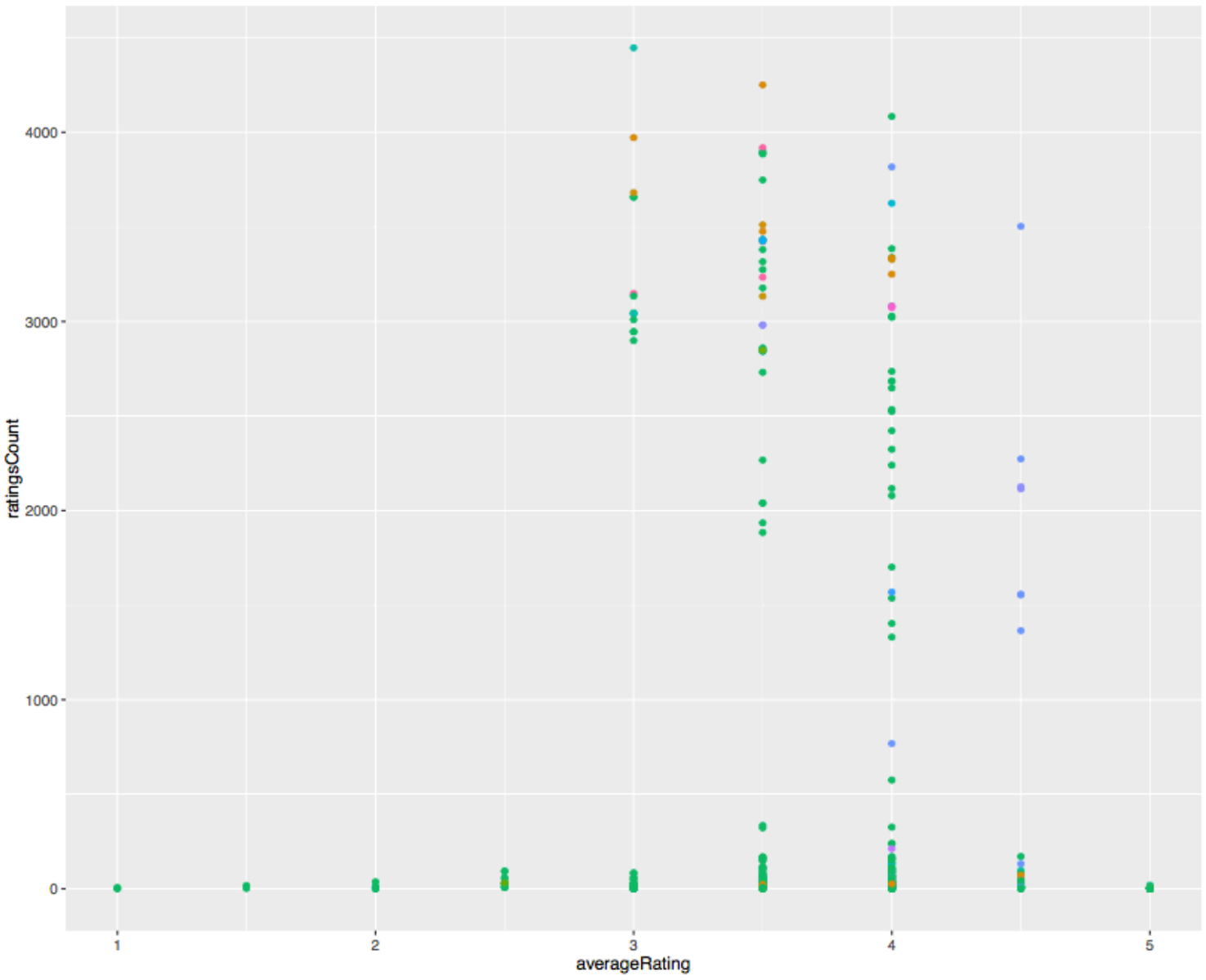
Young and Dissatisfied, 30s and happy?



- on average most of the low ratings came from Book-crossing users between the ages of 16-18.
- On the other hand, most readers in their 30s rated their books higher on average.

READER INSIGHTS: EDA

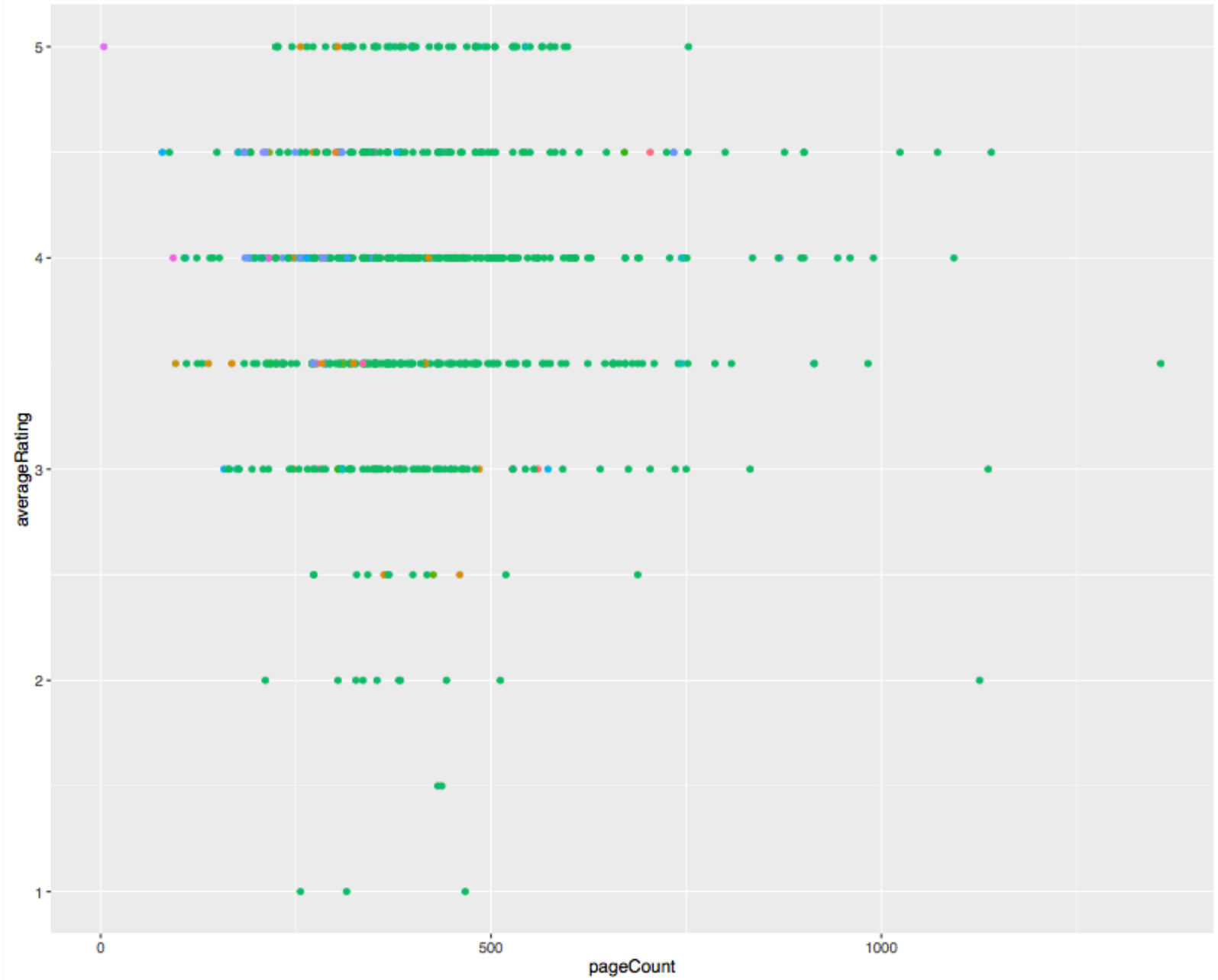
readers happier to escape reality



- Fiction represented by green points seemed to be the most read & rated genre by book-crossing members.

READER INSIGHTS: EDA

SHORT VS LONG STORIES

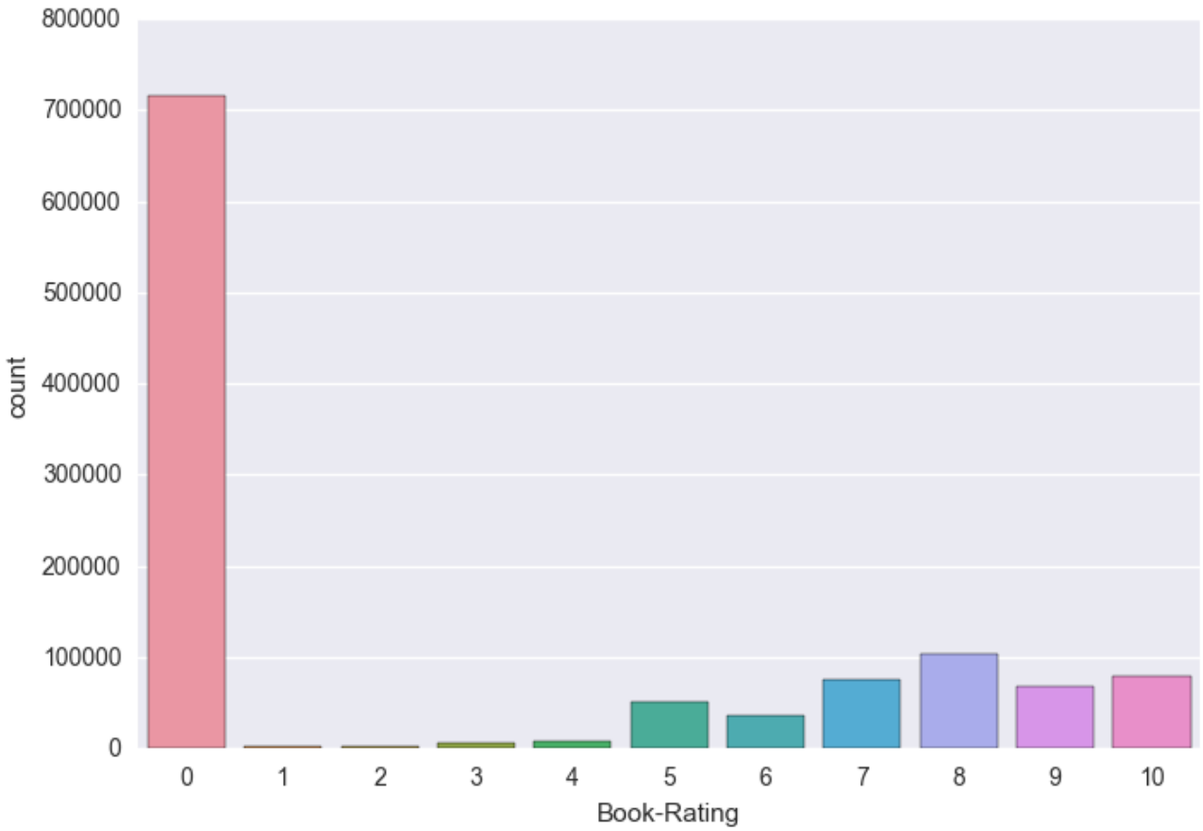


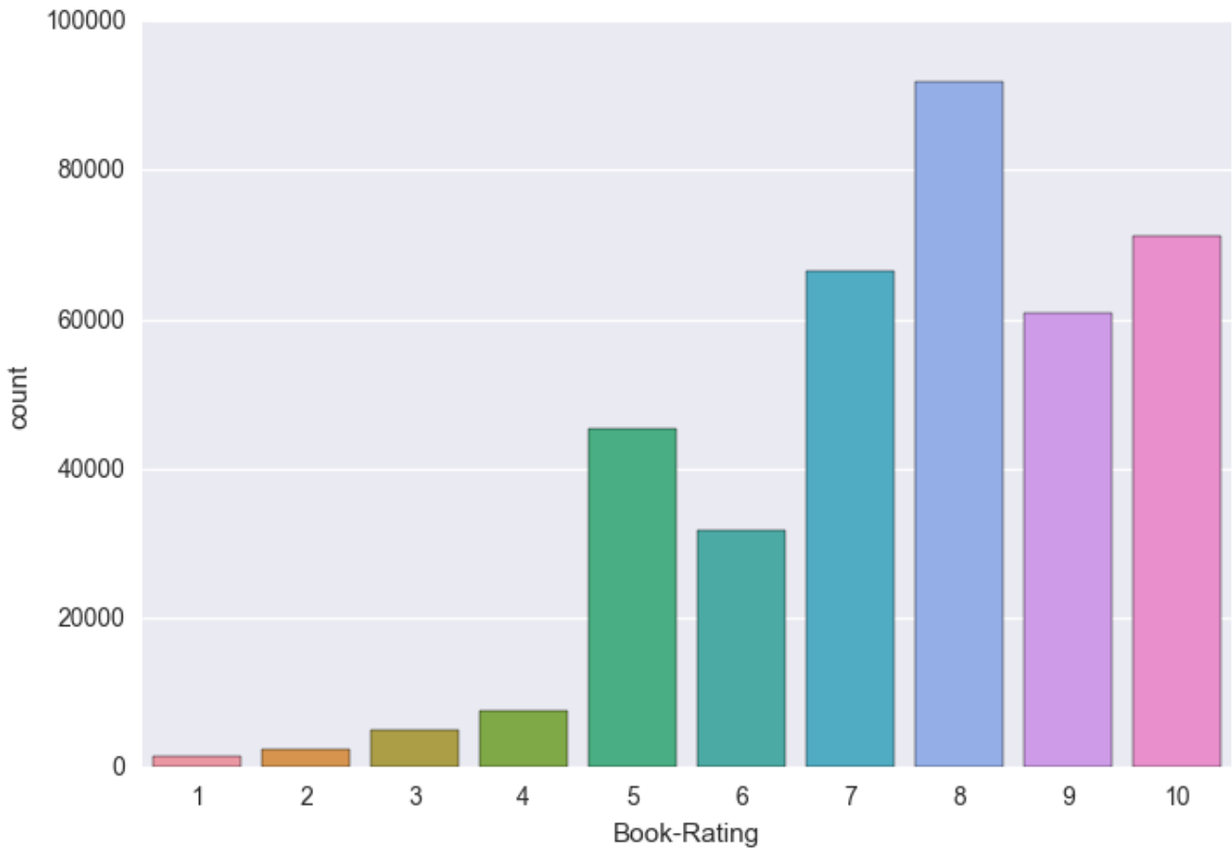
- more common for shorter books with less than 250 pages to be rated high. Same as 750 up

STORY LENGTH GUIDELINE

FEATURES	# WORDS	# PAGES
MICRO FICTION	up to 100	0.4
FLASH FICTION	100 - 1,000	4
SHORT STORY	1,000 - 7,500	30
NOVELLETTE	7,500 - 20,000	80
NOVELA	20,000 - 50,000	200
NOVEL	50,000 - 110,000	440
EPICS & SEQUEL	OVER 110,000	500

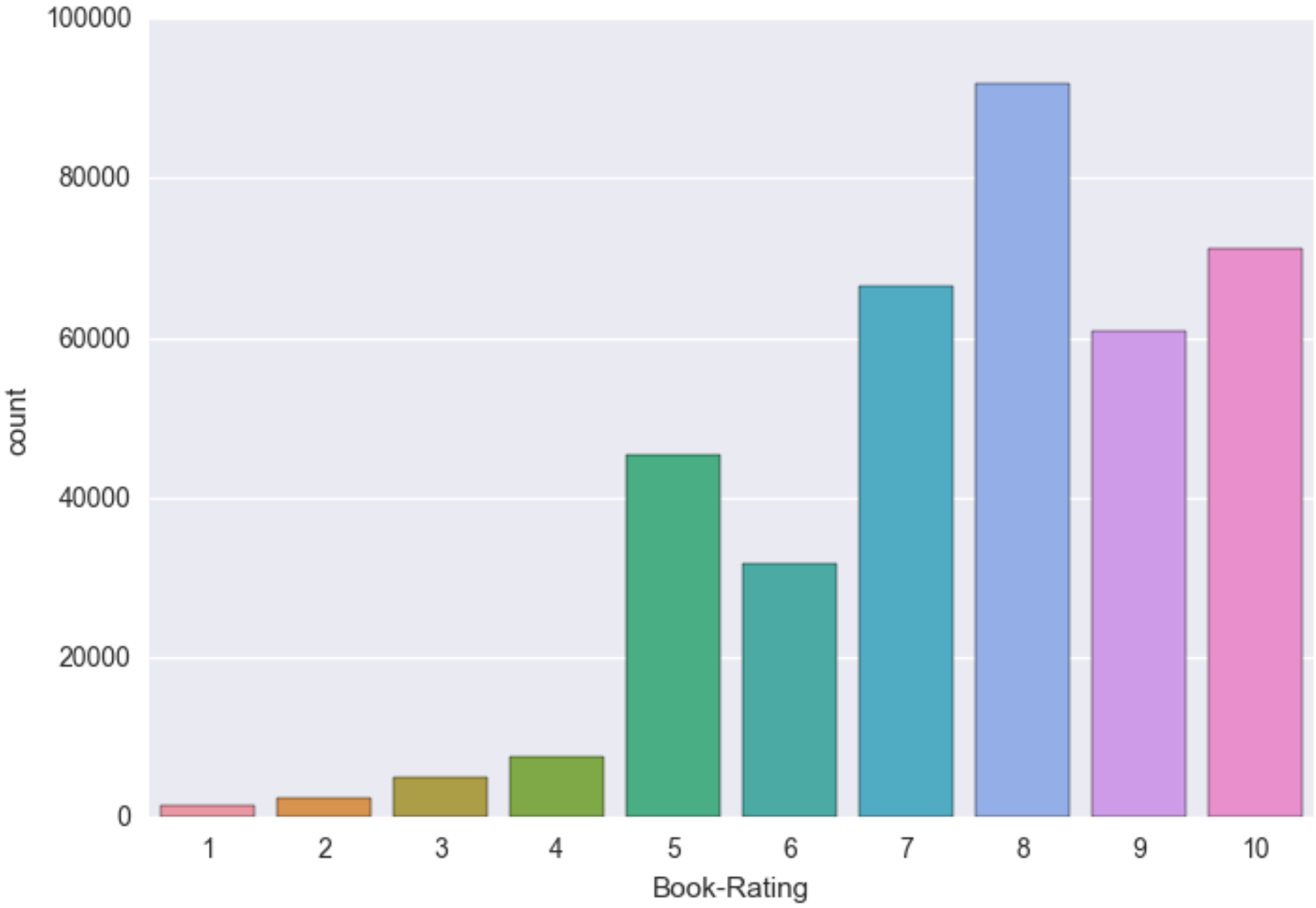
SUPERVISED LEARNING: PREDICTING BOOK APPROVAL RATING CLASSIFICATION





- Majority of books are un-rated
- Split into 2 datasets: rated and unrated
- Use rated for ML, and un-rated as new user dataset

CROSS-VALIDATION: IMBALANCED CLASSES



- most of the reviews were 8-10 and realized that we were faced with an imbalanced target class challenge. First, we performed a K-Fold cross-validation split on the entire rated dataset
- Performed cross-validation and then model fitting on the rated dataset with 10 classes for prediction resulted to low accuracy rate, low sensitivity, and low specificity.

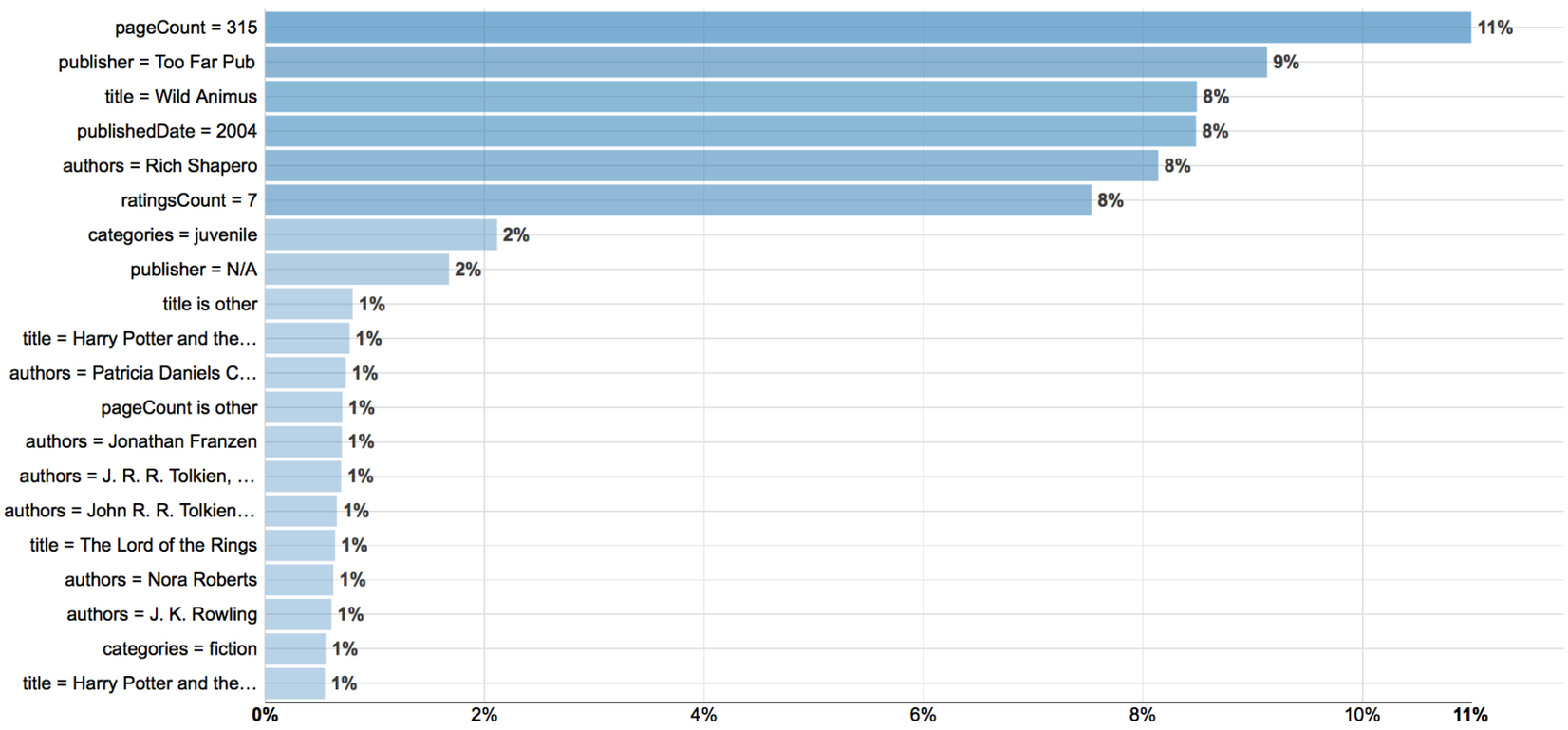
MODELS: LOGISTIC REGRESSION VS RANDOM TREES

Capstone Project | ANALYSES

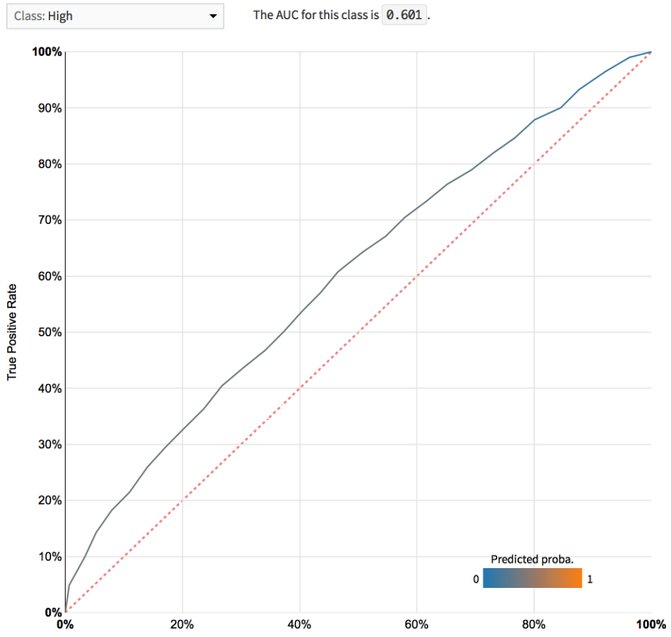
Analyze BX_Book_Ratings_Categories_LMH… | Summary | Script | Charts | Models | ACTIONS

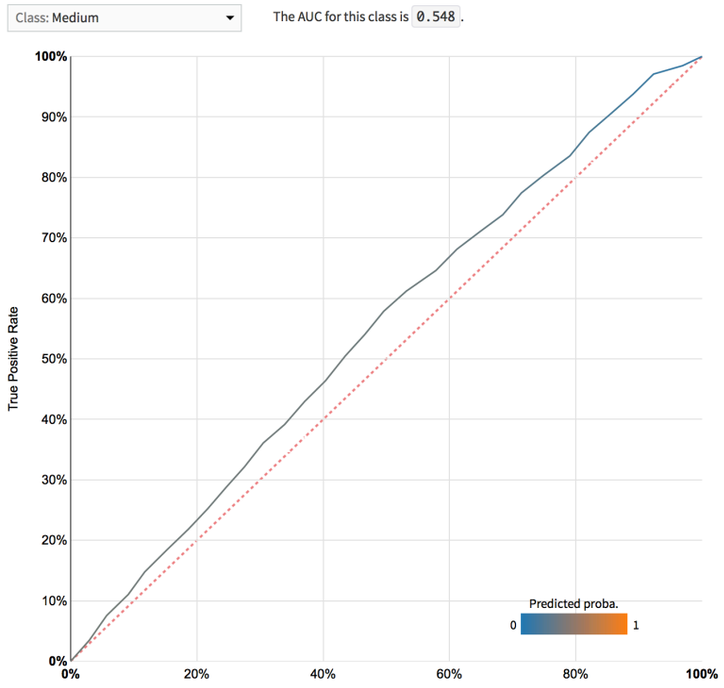
Predict Book-Rating-Category (Multiclass classification) | TRAIN | LIST | TABLE | Settings

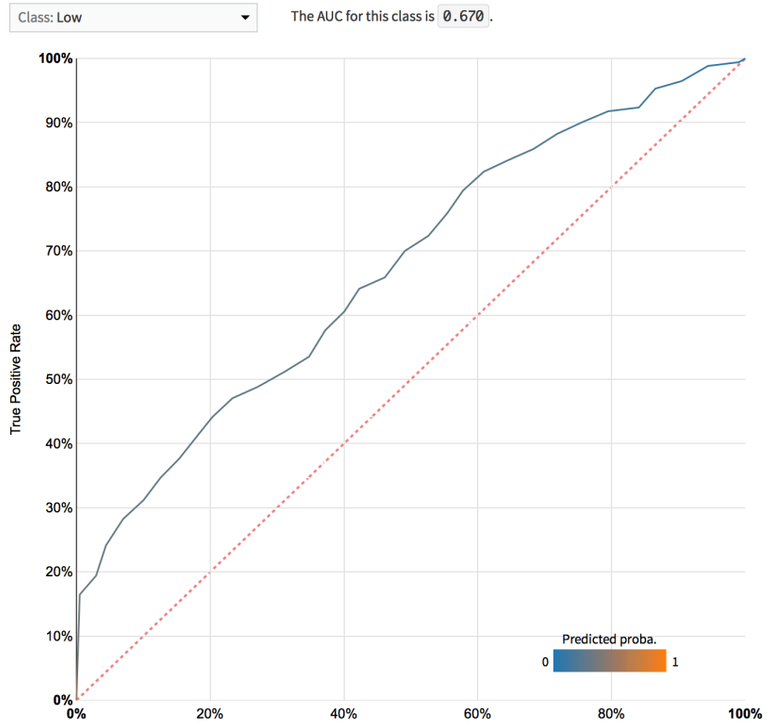
Name	Trained	Accuracy	Precision	Recall	F1 Score	Log Loss	ROC AUC
Decision Tree	2016-12-15 17:05:09	0.64	0.44	0.39	0.35	0.76	0.56
Decision Tree	2016-12-15 17:03:54	0.65	0.72	0.39	0.38	1.66	0.52
Logistic Regression	2016-12-15 17:05:13	0.64	0.41	0.39	0.35	0.75	0.58
Logistic Regression	2016-12-15 17:03:55	0.64	0.21	0.33	0.26	0.87	0.50
Random Forest with 84 trees	2016-12-15 17:05:09	0.46	0.41	0.43	0.38	1.02	0.63
Random Forest with 84 trees	2016-12-15 17:03:54	0.57	0.42	0.37	0.38	0.91	0.55



MODELS: LOGISTIC REGRESSION VS RANDOM TREES







Display: % of predicted classes

	Predicted		
Actual	High	Medium	Low
High	73 %	61 %	46 %
Medium	25 %	35 %	38 %
Low	2 %	4 %	16 %
	100 %	100 %	100 %

UNSUPERVISED LEARNING: RECOMMENDATION IPYTHON DEMO GRAPH LAB

```
Recsys training: model = ranking_factorization_recommender
Preparing data set.
    Data has 121367 observations with 30188 users and 847 items.
    Data prepared in: 0.174032s
Training ranking_factorization_recommender for recommendations.
+--------------------------------+--------------------------------------------------+----------+
| Parameter                      | Description                                      | Value    |
+--------------------------------+--------------------------------------------------+----------+
| num_factors                    | Factor Dimension                                 | 32       |
| regularization                 | L2 Regularization on Factors                     | 1e-09    |
| solver                         | Solver used for training                         | sgd      |
| linear_regularization          | L2 Regularization on Linear Coefficients         | 1e-09    |
| ranking_regularization         | Rank-based Regularization Weight                 | 0.25     |
| max_iterations                 | Maximum Number of Iterations                     | 25       |
+--------------------------------+--------------------------------------------------+----------+
  Optimizing model using SGD; tuning step size.
  Using 15170 / 121367 points for tuning the step size.
```

UNSUPERVISED LEARNING: RECOMMENDATION IPYTHON DEMO GRAPH LAB

In [37]:
```
print similar_items
similar_items_predict_rating
```

```
+------------+------------+------------------+------+
|    ISBN    |  similar   |      score       | rank |
+------------+------------+------------------+------+
| 0971880107 | 0316666343 | 0.0523138642311  |  1   |
| 0971880107 | 044023722X | 0.0348584055901  |  2   |
| 0971880107 | 0446364193 | 0.0329024791718  |  3   |
| 0971880107 | 0142001740 | 0.0323660969734  |  4   |
| 0971880107 | 0380731851 | 0.0318396091461  |  5   |
| 0971880107 | 0316601950 | 0.0306122303009  |  6   |
| 0971880107 | 059035342X | 0.0295790433884  |  7   |
| 0971880107 | 0553572997 | 0.0288350582123  |  8   |
| 0971880107 | 0671003755 | 0.0275545120239  |  9   |
| 0971880107 | 0446605484 | 0.0260047316551  |  10  |
+------------+------------+------------------+------+
[180 rows x 4 columns]
Note: Only the head of the SFrame is printed.
```

UNSUPERVISED LEARNING: RECOMMENDATION IPYTHON DEMO GRAPH LAB

In [38]:
```
print ISBNlookup('0971880107')
ISBNlookup('037570504X')
```

```
+------------+--------------+-----------+-------------------------------+-----------+
|    ISBN    |   authors    | categories |         imageThumbnail        | pageCount |
+------------+--------------+-----------+-------------------------------+-----------+
| 0971880107 | Rich Shapero |  Fiction  | http://books.google.com/bo... |    315    |
+------------+--------------+-----------+-------------------------------+-----------+
+---------------+-------------+-------------------------------+-------------+
| publishedDate |  publisher  |          textSnippet          |    title    |
+---------------+-------------+-------------------------------+-------------+
|      2004     | Too Far Pub | Newly graduated from colle... | Wild Animus |
+---------------+-------------+-------------------------------+-------------+
[? rows x 9 columns]
Note: Only the head of the SFrame is printed. This SFrame is lazily evaluated.
You can use sf.materialize() to force materialization.
```

Out[38]:

ISBN	authors	categories	imageThumbnail	pageCount	publishedDate	publisher
037570504X	Edwidge Danticat	Fiction	http://books.google.com/books/content?id=EcA-R ...	234	1998	Vintage

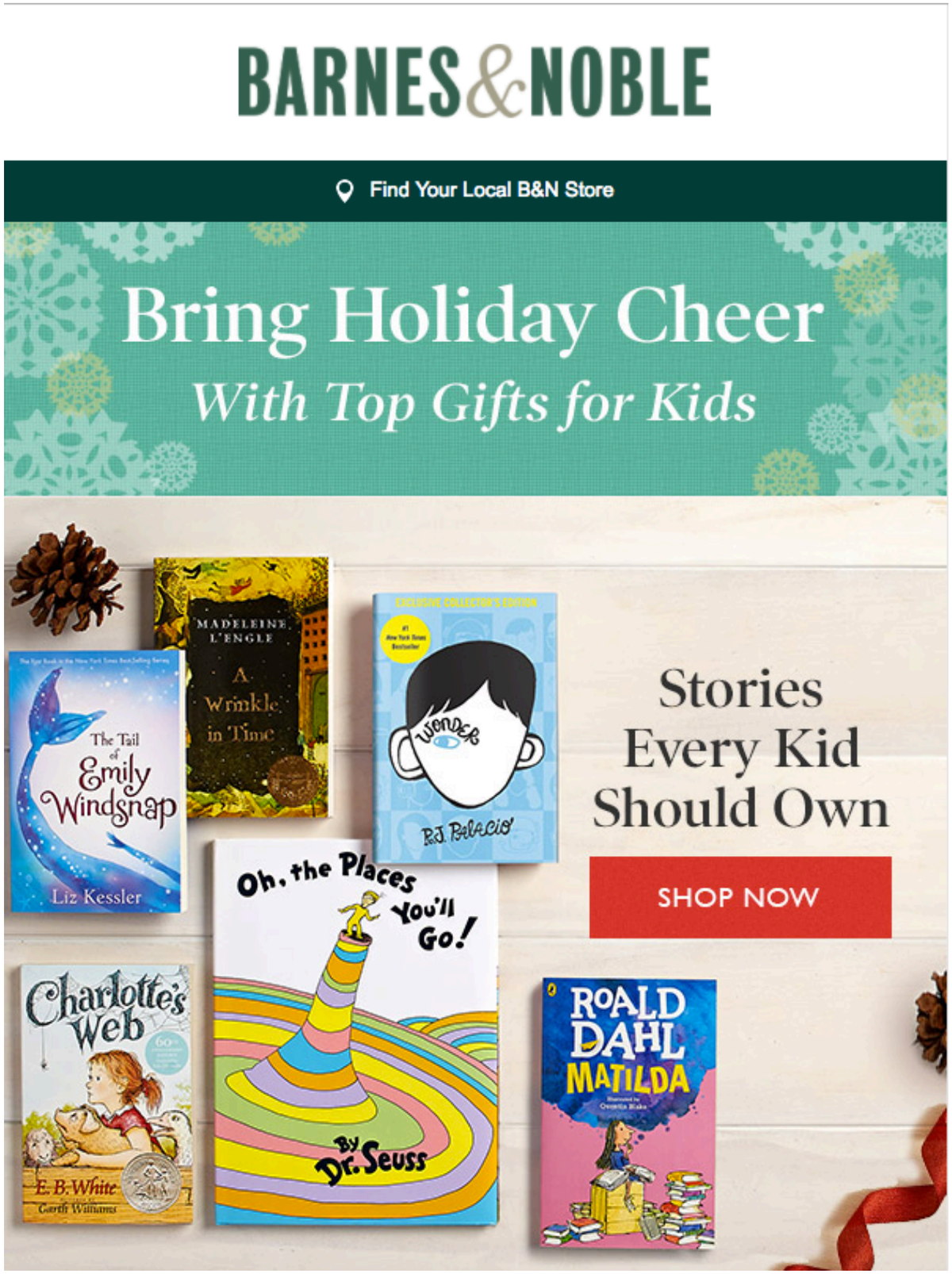
textSnippet	title
Oprah's Book Club.	Breath, Eyes, Memory

[? rows x 9 columns]

Note: Only the head of the SFrame is printed. This SFrame is lazily evaluated.

You can use sf.materialize() to force materialization.

BUSINESS APPLICATION



- BookLab can be implemented by Barnes and Noble using their own proprietary dataset. We expect that with the rich dataset they have from their BN Members and over 6 million books, BookLab will render more accurate book classification and recommendation results compared to the more limited dataset the team used for this capstone.

- The recommendation algorithm can also be used for more personalized email marketing campaigns to BN members wherein every month TopMatch books alerts will be sent instead of generic email ads.

FUTURE WORK

- Perform SMOTE data balancing and other penalizing models to check if better ROC, sensitivity, and specificity can be achieved
- Add more book features such as pricing via scraping to understand price sensitivity of customers
- Identify interesting customer clusters after the addition of more features
- Develop title and author search to improve user experience of the app
- Enhance the recommendation model through the reformulated Linear Regression model by Yehuda Koren et al.

SOURCES

- Barnes & Noble
- Amazon
- Quora
- Collective Intelligence

APPENDIX

- BRCF.1.Baseline.ipynb
- BRCF.2.CleanerData.ipynb
- BRCF.3.VerifyData.ipynb
- BRCF.4.ImputeImplicit.ipynb
- BRCF.5.DataImpact.ipynb

Thank you!

geekCentroids

Chris Valle

Jhonnastan Regalado

Conred Wang