Analyzing Book Checkout Trends at the Seattle Public Library

Sarvesh Fotedar, Ekrem Kizilkaya, Abbas Shaikh, Cody VanZandt, Andy Wang

Contents

1 Introduction

From classic literature to contemporary bestsellers, the Seattle Public Library's book collection offers a rich window into its surrounding community's literary tastes and interests. By analyzing checkout trends and book movement, we can uncover which novels and authors are most beloved among library patrons, giving us remarkable insight into the community's unique cultural and social context.

This report conducts a preliminary examination of 11,749,255 checkouts from the Seattle Public Library system over the five years between 2018 and 2022. We intend to combine this information with Goodreads review data from the UCSD Book Graph to answer general questions about the contours of contemporary American readership and even perform specific analyses on the competitive dynamics of the book authorship and publishing industry.

From this point forward, the following questions to guide our exploration and report:

1. Why do some books enjoy widespread acclaim while others fizzle?
2. Do popular books explode onto the scene or accumulate readers more gradually?
3. How do the popular success and critical acclaim of novels differ?
4. How has recent conglomeration of publishers changed the literary marketplace?
5. And, ultimately, to what degree can the success or failure of a book be predicted?

From the most popular genres and titles to the unique factors influencing readers' choices, book checkout data offers a rich window into the complex relationship between readers, libraries, and the larger cultural context. In this paper, we will examine the book checkout trends at the Seattle Public Library in detail, drawing on both statistical and literary analysis to shed light on the unique reading culture of this vibrant community.

2 Datasets

2.1 Seattle Public Library

Our primary dataset comes directly from the Seattle Public Library and has approximately twelve million rows for book checkouts spanning the five years between 2018 and 2022. Specifically, each row corresponds to a monthly count of checkouts for the physical or electronic version of an item. Items are not just limited to books; the Seattle Public Library also makes video discs, e-books, and sound discs available to patrons, to name a few. Each row is described by eleven variables, which are described by Table 1.

Variable	**Description**
UsageClass	Whether the item was physical or digital
CheckoutType	The tool or vendor that was used for checkout
MaterialType	The item type (ex: book, video disc, etc)
CheckoutMonth	The four digit checkout year
CheckoutYear	The month of checkout
Checkouts	The number of times that the item was checked out within the checkout month
Title	The full title and subtitle
Creator	The author or entity responsible for creating
Subjects	The subjects as they appear in the library catalog
Publisher	The publisher of the title
PublicationYear	The year that the item was published, printed, or copyrighted

Table 1: Seattle Public Library Variables

2.2 UCSD Book Graph

Our supplementary dataset comes from the UCSD Book Graph initiative and contains over fifteen million reviews for approximately two million books from 465,000 users. Each row of the dataset is a JSON object which represents a single review and its associated metadata. The nine attributes of each JSON-ified review are described in Table 2

Attribute	**Description**
user_id	The UUID of the reviewing user
book_id	The unique numerical id for the reviewed book
review_id	The UUID of the review itself
rating	The numerical rating of the book out of 5
review_text	The text content of the review
date_added	The date the review was published
date_updated	The date the review was updated, if applicable
n_votes	The number of votes endorsing the review
n_comments	The number of comments for the review

Table 2: UCSD Book Review Object Attributes

3 Exploratory Analysis

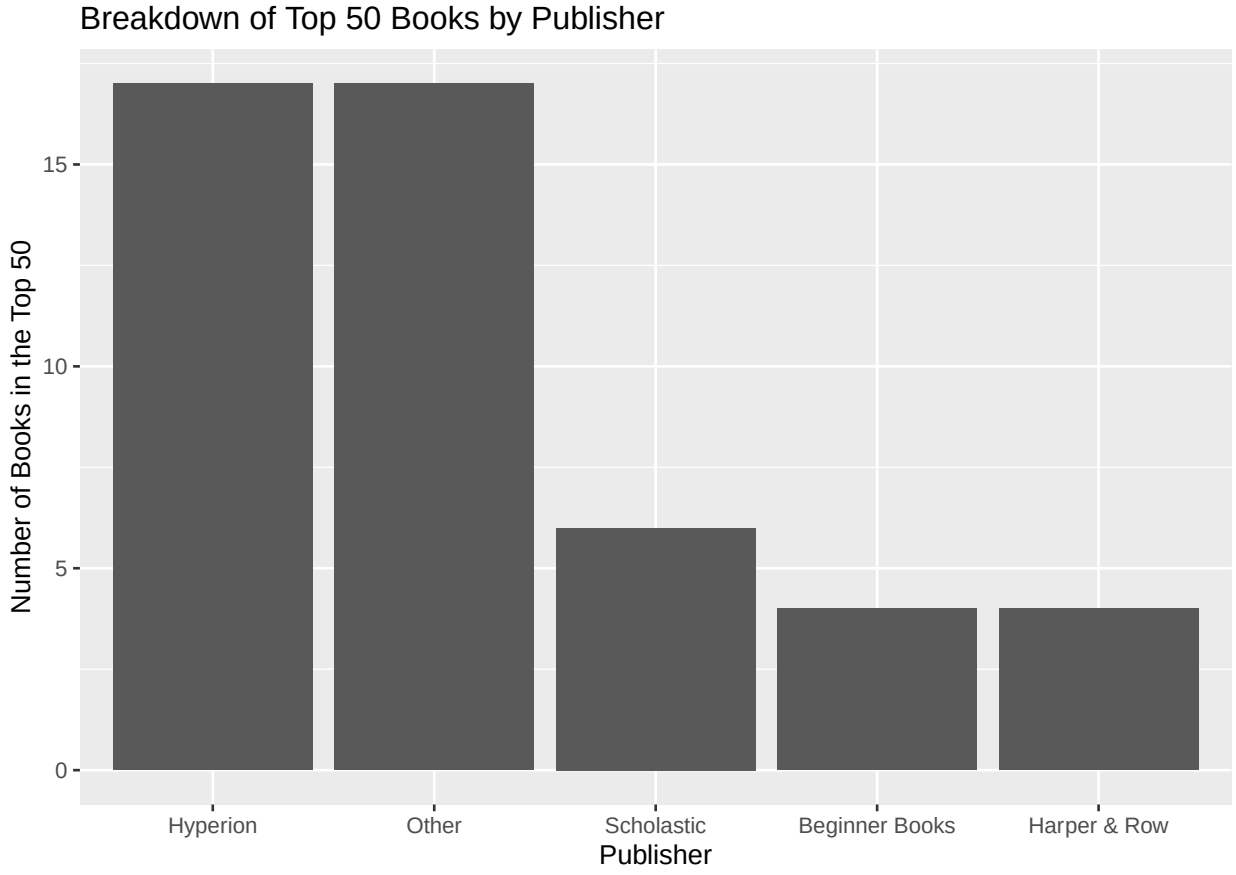
3.1 Seattle Public Library Dataset

3.2 UCSD Book Graph

4 Add Sections Here (One Section per Question/Topic being addressed)

5 Analyzing properties of book popularity

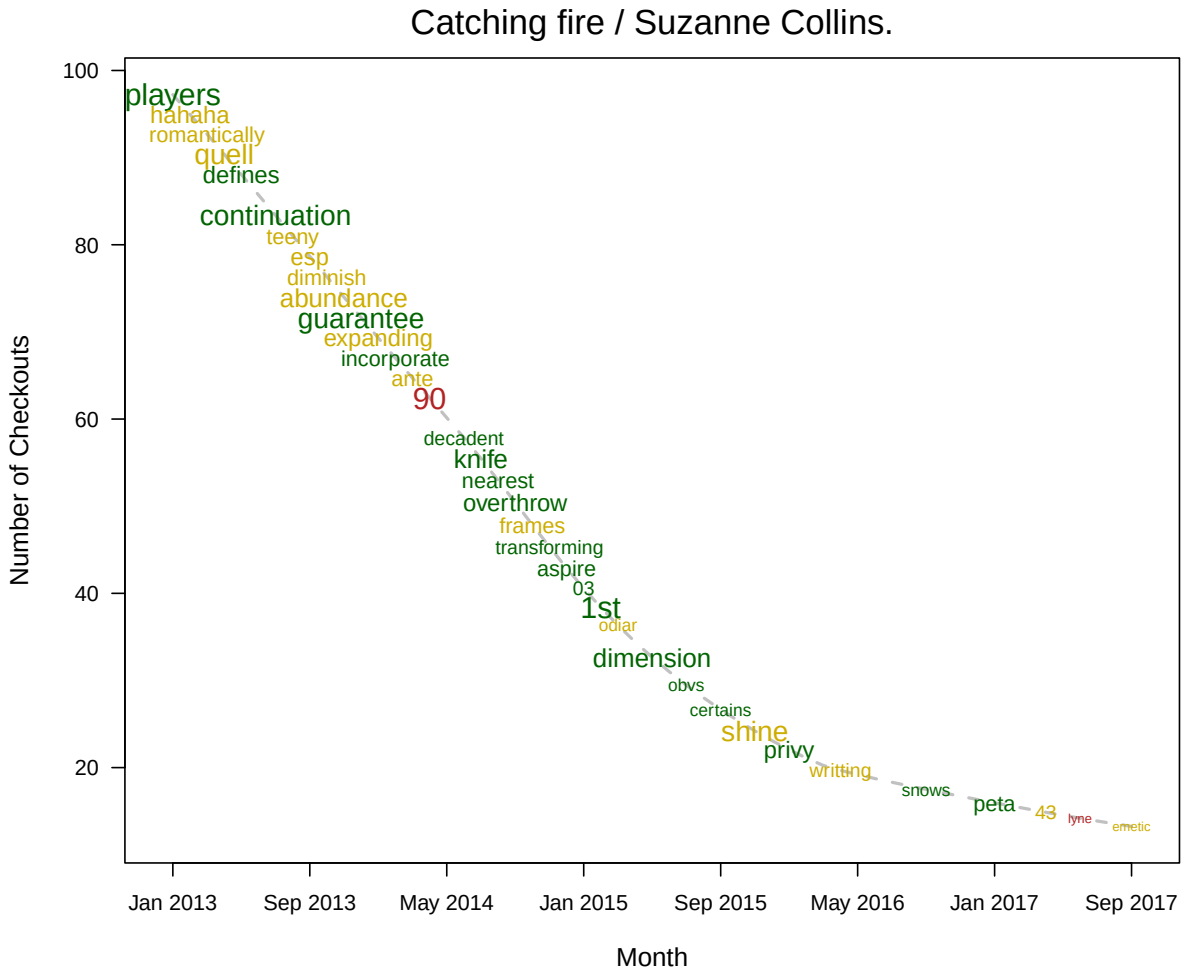
We first pose the question, "why do some books enjoy widespread acclaim while others fizzle?" More broadly, we are interested in what specific characteristics of books make them more popular than others. We begin this exploration by analyzing the publishers of the top fifty books checked out at the Seattle Public Library. Specifically, we first filter for the top 50 books by aggregate number of checkouts from 2005 to 2022, then keep a count of their associated publishers. We found that most publishers only have a single "popular" book, so we group any publisher with less than 3 books together in an 'Other' category.



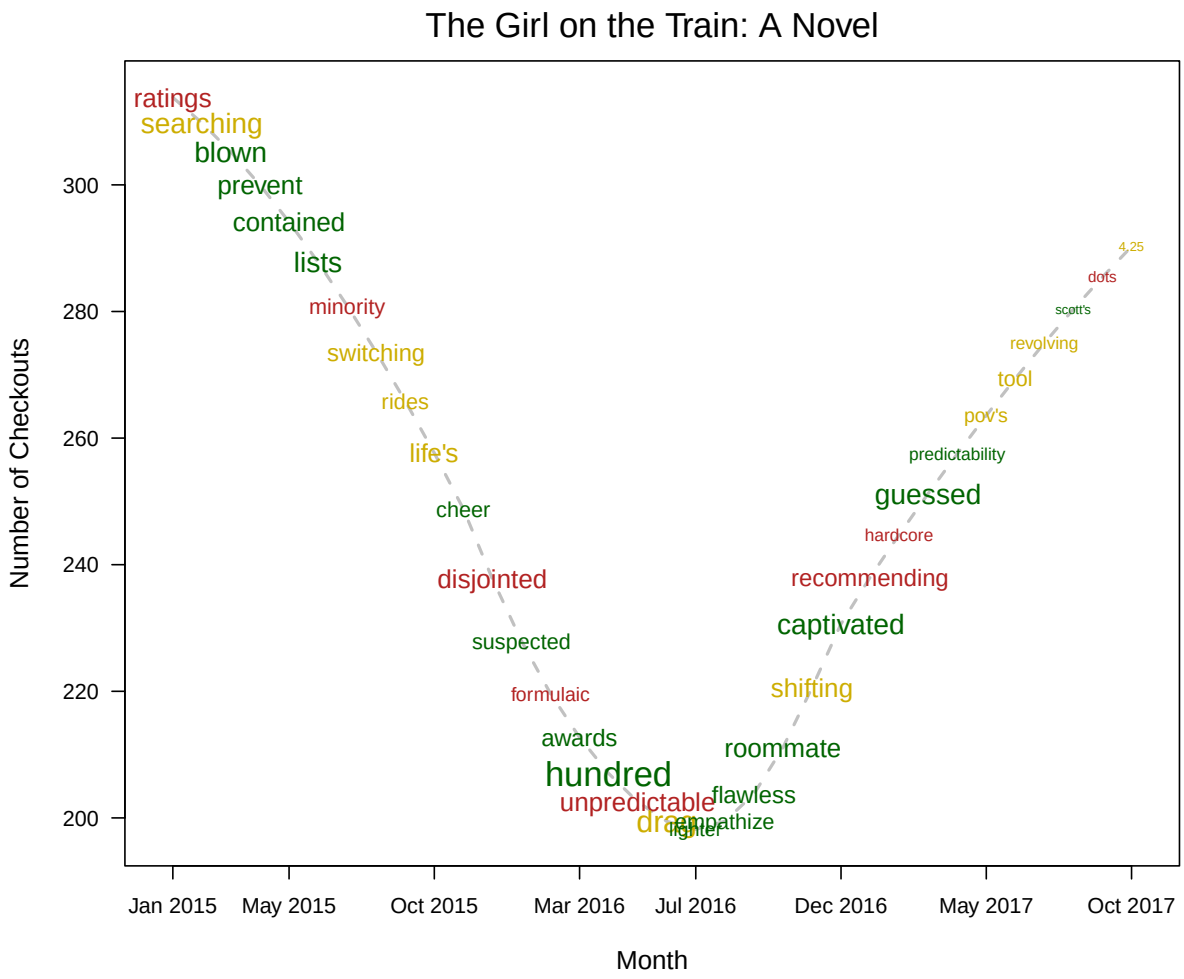
The bar plot shows that Hyperion has published the highest number of "popular" books at the Seattle Public Library. However, they are still tied with the 'Other' category, which implies that the publisher doesn't really have an effect on the popularity of the book. This seems to makes sense; most readers don't consider the publishing company when deciding what book to read.

6 Killer Plot

6.1 Catching Fire - Suzanne Collins



6.2 The Girl on the Train: A Novel - Paula Hawkins


The Girl on the Train: A Novel
Number of Checkouts
300
280
260
240
220
200
Jan 2015
May 2015
Oct 2015
Mar 2016
Jul 2016
Dec 2016
May 2017
Oct 2017
Month
ratings
searching
blown
prevent
contained
lists
minority
switching
rides
life's
cheer
disjointed
suspected
formulaic
awards
hundred
unpredictable
dra
empathize
fighter
flawless
roommate
shifting
captivated
recommending
hardcore
guessed
predictability
pov's
tool
revolving
scott's
dots
4.25

7 Appendix

7.1 Fuzzy Title Matching

To make best use of our UCSD Book Graph Goodreads data, we need to match unique book IDs from Goodreads reviews to book checkouts from the Seattle Public Library. The Seattle data, however, is messy. Author information is often encoded in the title field, which is itself often full of misspellings and other typographical oddities. Dataset alignment, then, is a challenging problem in its own. As a first attempt, we employ a fuzzy string matching algorithm that computes pairwise similarity between the Goodreads and Seattle book titles and reports a match when that similarity crosses a given threshold. Our current similarity measure is Jaro-Winkler, which we selected primarily because it weighs matches at the beginning of a string more heavily than matches towards the end. Given the erroneous additions frequently appended to the end of Seattle library book titles, Jaro-Winkler ought to perform well. After parameter optimization, we report that approximately thirty percent of Seattle titles have been matched to Goodreads book IDs. With more nuanced title preprocessing and multivariable matching across author and publisher, we anticipate that the match rate could approach 50 percent.

7.2 Popularity Curve Clustering

Beyond string matching, we also report some initial research into time series dimensionality reduction and clustering. In considering book popularity over time, it is natural to ask if there are certain popularity patterns that recur across books, publishers, and genres. If these common patterns can be identified, then perhaps they can be predicted. After consulting the literature on time series clustering, we present a pipeline that we believe could identify clusters of books that exhibit similar popularity patterns. This pipeline computes monthly popularity time series curves and transforms them through z-normalization, discrete cosine transform, dynamic time warping, and k-medoids clustering.

Z-normalization makes the shape of popularity curves comparable across books with dramatically different raw checkout numbers. Discrete cosine transform reduces the dimensionality by decomposing and recomposing the time series using a smaller number of cosine waves. Dynamic time warping defines a similarity score for popularity curves with potentially non-overlapping time domains. And finally, k-medoids clustering modififes the more familiar k-medoids clustering by replacing average-computed centroids with median-computed centroids. K-medoids is a less common choice than k-means, so the selection is worth commenting upon. Compared with k-means, k-medoids is less sensitive to outliers and replaces the creation of artificial average-based centroids with median-based ones This sidesteps the need to average different popularity curves together during centroid creation – a questionable practice to be sure.

We hope that this unsupervised machine learning pipeline will, once implemented, make possible a variety of supervised and predictive modeling tasks.