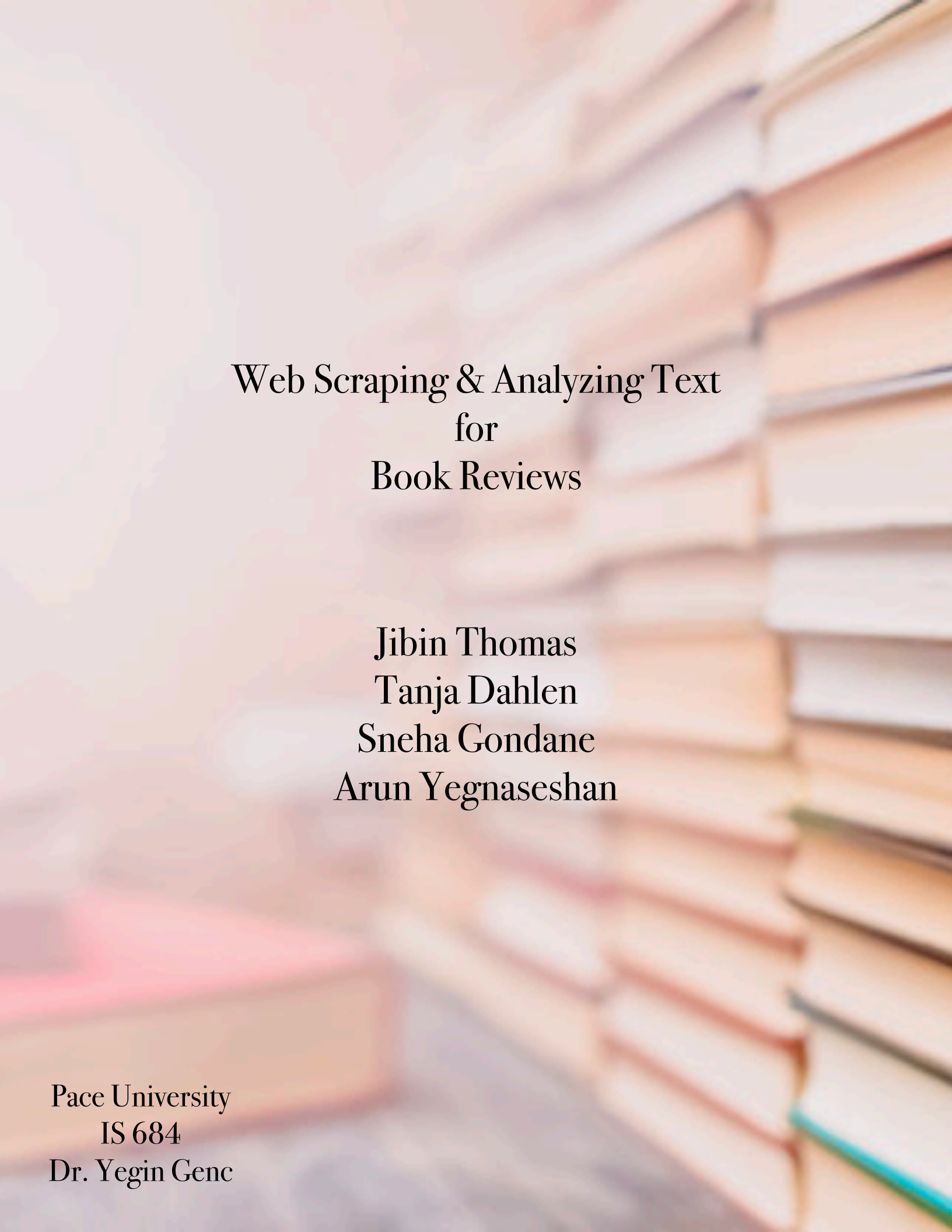


# Web Scraping & Analyzing Text for Book Reviews

Jibin Thomas
Tanja Dahlen
Sneha Gondane
Arun Yegnaseshan

Pace University
IS 684
Dr. Yegin Genc

## Table of Contents

# Introduction

Book reviews make books a known quantity. They decrease the risk to readers that a particular book will be not what they had in mind at all. In fact, book reviews help potential readers become familiar with what a book is about, give them an idea of how they themselves might react to it and determine whether this particular book will be the right book for them right now.

Book reviews save readers time, prepare them for what they will find and offer them a greater chance of connecting with a particular book, even before they read the first page. Book reviews give books greater visibility and a greater chance for the authors getting found by more readers. Reviews help books to be more discoverable and credible and thereby boost sales and generate revenue for publishers and authors.

Hence it really crucial to examine these reviews and analyze their content in order to understand book sales and predict their market performance.

## Project Overview

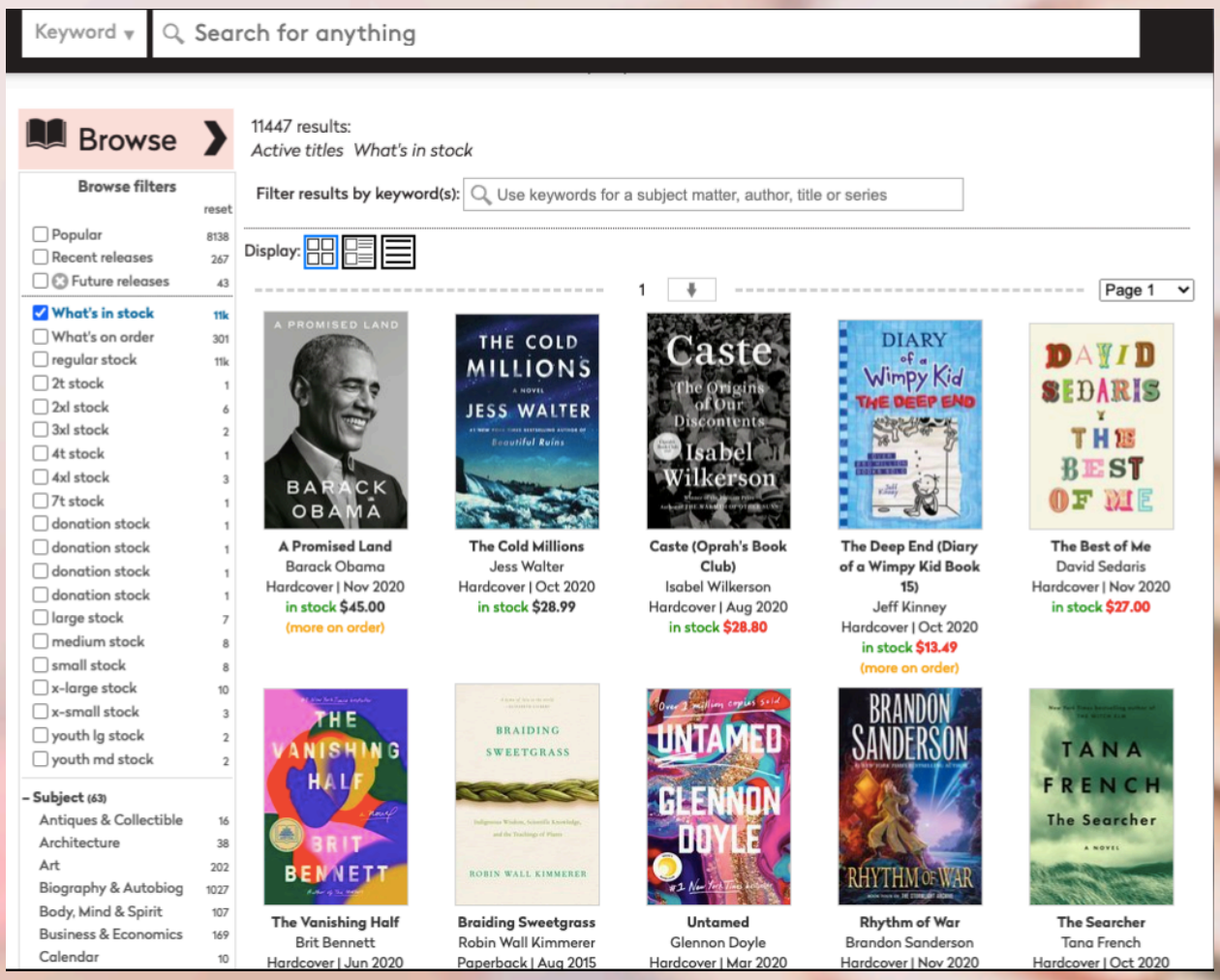
The website www.booksaremagic.com is for a local bookstore in Cobble Hill, Brooklyn. Owned by best-selling author Emma Straub and her husband Michael Fusco-Straub, Books Are Magic is home to exciting new releases and beloved classics, hidey-holes for children and books to read in them, gumballs filled with poetry, readings and panels almost every night of the week and storytimes on the weekends, and yes, plenty of magic.

Our project goal is to scrape book attributes and reviews for the top 10,000 most popular and available books from www.booksaremagic.com and to analyze the generated datasets. The main idea is to answer whether critic reviews have an influence on readers and a subsequent effect on book sales?

The roadmap for the project consists of the following steps: Scrape data from the website, Cleanse it through appropriate operations, Perform Exploratory Data Analysis using Tableau, Perform Text Analysis using python libraries such Natural Language ToolKit (NLTK), TextBlob etc. to gain meaningful insights and discover valuable results

# Web Scraping

To scrape "BooksAreMagic.com", we used Selenium - an automated scraping tool, PhantomJS - a headless browser and BeautifulSoup - a python library. For web scraping, we analyzed the website and selected only books from the Popular book list and all books that are available and in stock. We decided to exclude all future releases since those books won't have any reviews.



## Part 1: Scraping book attributes

We selected the top 10,000 most popular available books using selenium for book selection and page navigation. Using Beautiful Soup, we extracted attributes for each book. The following attributes were scraped: Book Title, Author, Price, Sales Rank (Ranking based on the number of copies of sold. Lower the rank, higher the number of books sold by the respective publisher/author), Genre, Published Month, Published Year.

After successfully scraping 200 pages from the website, the dataset generate consists of 7 columns & 10,000 rows.

```
In [227]:
scraped_books
#Approx 10K books scraped with 7 required attribute columns
```

| | Title | Author | Price | Sales Rank | Genre/Theme | Published Month | Published Year |
|---|---|---|---|---|---|---|---|
| 0 | A Promised Land | Barack Obama | 45.00 | 1 | History | Nov | 2020 |
| 1 | Caste (Oprah's Book Club): The Origins of Our... | Isabel Wilkerson | 28.80 | 2 | History | Aug | 2020 |
| 2 | The Cold Millions: A Novel | Jess Walter | 28.99 | 3 | Fiction | Oct | 2020 |
| 3 | The Deep End (Diary of a Wimpy Kid Book 15) | Jeff Kinney | 13.49 | 4 | Juvenile Fiction | Oct | 2020 |
| 4 | The Vanishing Half: A Novel | Brit Bennett | 24.30 | 5 | Fiction | Jun | 2020 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 9965 | The Young Bride: A Novel | Alessandro Baricco | 16.00 | NA | Fiction | Jul | 2016 |
| 9966 | Gris Grimly's Tales from the Brothers Grimm | Jacob and Wilhelm Grimm | 17.99 | NA | Juvenile Fiction | Jul | 2016 |
| 9967 | Bye Bye Blondie | Virginie Despentes | 17.95 | NA | Fiction | Jul | 2016 |
| 9968 | Let's Be Less Stupid: An Attempt to Maintain ... | Patricia Marx | 14.99 | NA | Self-Help | Jul | 2016 |
| 9969 | Lion Lessons | Jon Agee | 17.99 | NA | Juvenile Fiction | Jul | 2016 |

9970 rows × 7 columns

## Part 2: Scraping book reviews

On Inspecting the website, it can be observed that all elements under the review pane originate from hyperlinks where the review content are saved and loaded in json format. Using Beautiful Soup, all reviews were scraped and the review column consisted of all reviews, merged together for a single book. For text analysis it was essential to split all the reviews into separate rows for each book using pandas and other python libraries.

The scraped dataset consists of top 10,000 book titles, Links and all their reviews merged into a single row.

```
In [532]:
scraped_reviews
#All reviews are scraped into one row for each book
```

| | Book Title | Book Link | Book Reviews |
|---|---|---|---|
| 0 | A Promised Land | https://www.booksaremagic.net/?q=h.reports.tin... | "Barack Obama is as fine a writer as they come... |
| 1 | Caste (Oprah's Book Club) | https://www.booksaremagic.net/?q=h.reports.tin... | "Magnificent . . . a trailblazing work on the ... |
| 2 | The Cold Millions | https://www.booksaremagic.net/?q=h.reports.tin... | "Vibrant.... Filled with a gusto that honors the... |
| 3 | The Deep End (Diary of a Wimpy Kid Book 15) | https://www.booksaremagic.net/?q=h.reports.tin... | "The Wimpy Kid books run like a well-oiled mac... |
| 4 | The Vanishing Half | https://www.booksaremagic.net/?q=h.reports.tin... | Named a BEST BOOK OF THE YEAR by Entertainment... |
| ... | ... | ... | ... |
| 9995 | The Invaders | https://www.booksaremagic.net/?q=h.reports.tin... | Book has no reviews |
| 9996 | Coconut | https://www.booksaremagic.net/?q=h.reports.tin... | Book has no reviews |
| 9997 | Where Are the Galapagos Islands? | https://www.booksaremagic.net/?q=h.reports.tin... | Book has no reviews |
| 9998 | Typewriters, Bombs, Jellyfish | https://www.booksaremagic.net/?q=h.reports.tin... | "[Tom McCarthy] is one of the few writers who ... |
| 9999 | Big Bad Bubble | https://www.booksaremagic.net/?q=h.reports.tin... | * "Rubin's voice-over narrator counsels Yerbur... |

10000 rows × 3 columns

After splitting and separating the review for each book, the dataset finally consisted of approximately 112,000 rows for text analysis.

```
In [531]:
1  scraped_reviews_copy
2  #More than 112K reviews scraped
```

| | Book Title | Critic Reviews |
|---|---|---|
| 0 | A Promised Land | "Barack Obama is as fine a writer as they come... |
| 1 | A Promised Land | Chimamanda Ngozi Adichie, The New York Times B... |
| 3 | Caste (Oprah's Book Club) | "Magnificent . . . a trailblazing work on the ... |
| 4 | Caste (Oprah's Book Club) | O: The Oprah Magazine"This book has the reverb... |
| 5 | Caste (Oprah's Book Club) | Dwight Garner, The New York Times"A surprising... |
| ... | ... | ... |
| 114770 | Typewriters, Bombs, Jellyfish | [...less] |
| 114771 | Big Bad Bubble | * "Rubin's voice-over narrator counsels Yerbur... |
| 114772 | Big Bad Bubble | that moment when the monsters |
| 114773 | Big Bad Bubble | or whatever lurks under their particular beds |
| 114774 | Big Bad Bubble | start to worry them."—The New York Times |

112894 rows × 2 columns

# Exploratory Analysis

We performed EDA using Tableau to answer 5 most relevant questions.

**Question 1:**

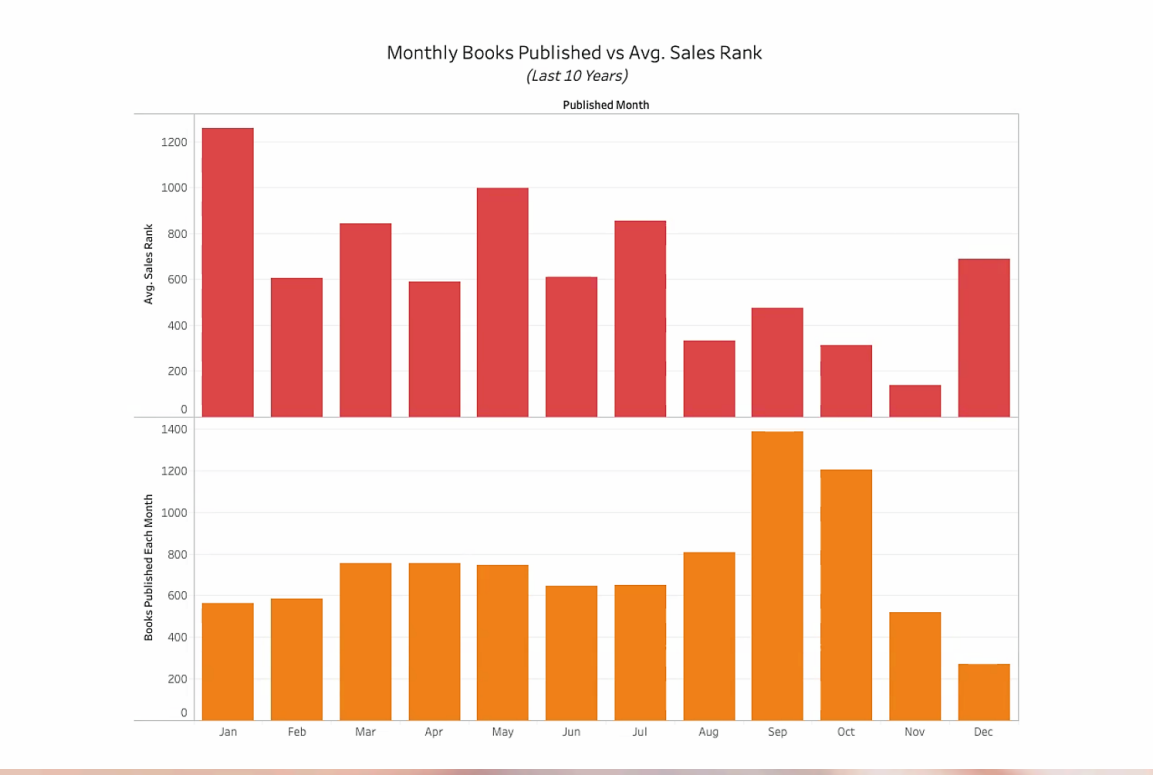
Which month has the most number of books published?

Ans: As evident from the graph, the month September followed by the month of October has the most number of books published. A possible explanation for this rise might be that these months are pre cursor to the holiday season where more books are sold.

**Question 2:**

Which month has the highest sales rank for books?

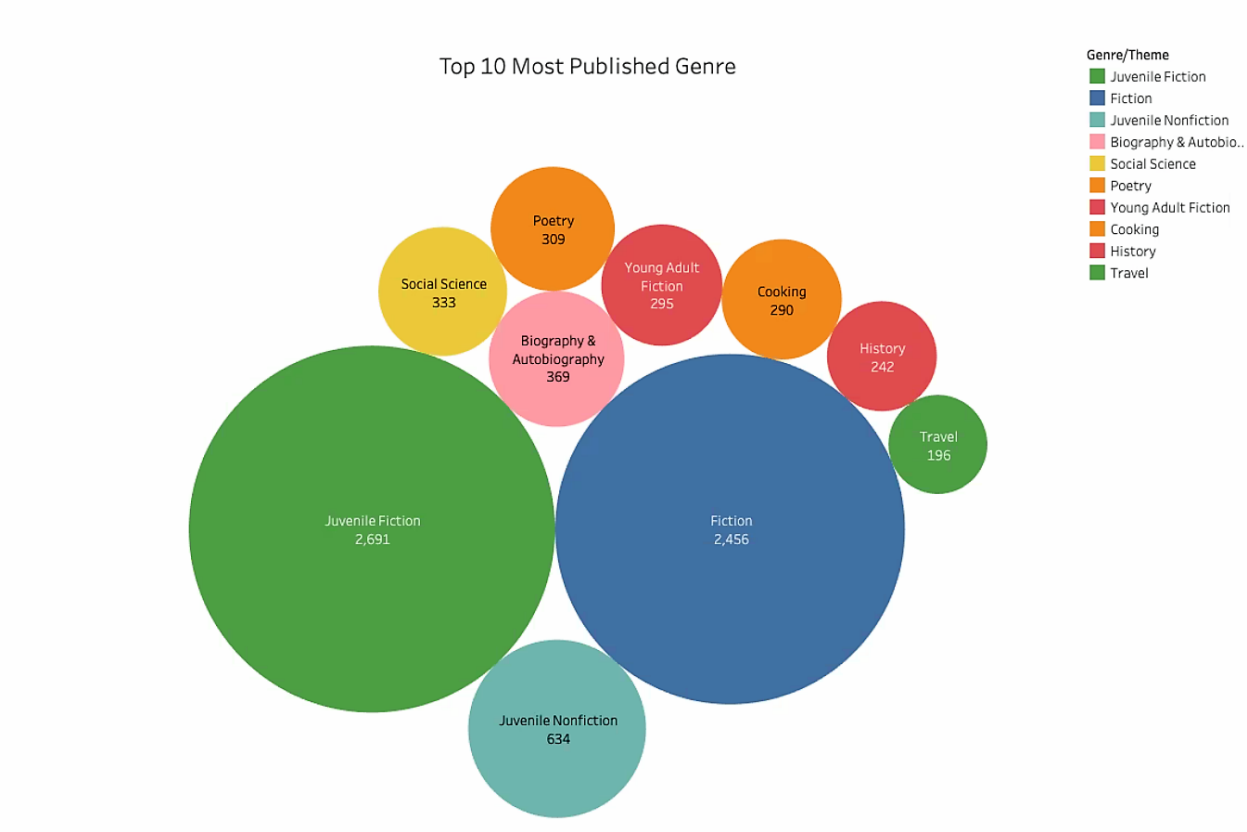
Ans:The month of November produces the book with highest sales rank



**Question 3:**

What are the most published genres?

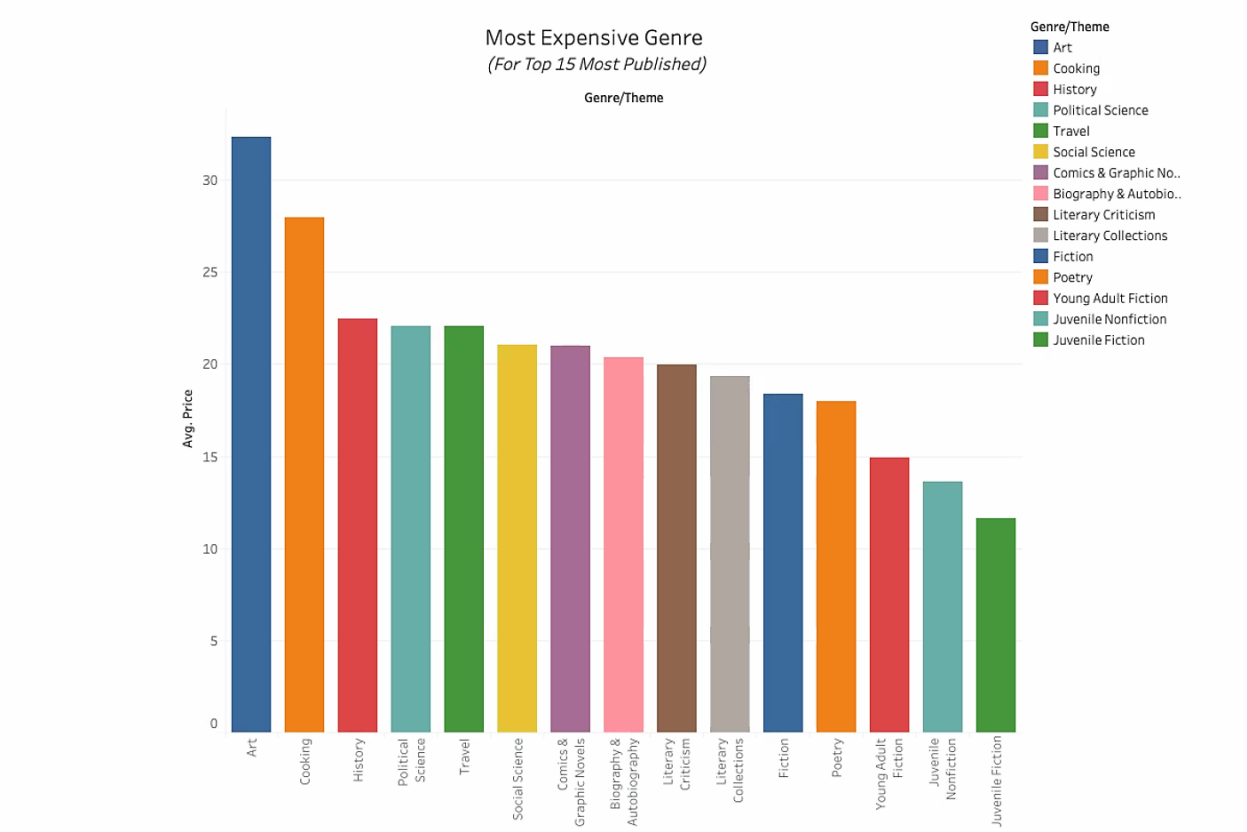
Ans: Juvenile Fiction, Fiction, Juvenile Non-Fiction, Biography & Autography are few of the most published genres



**Question 4:**

What are the most expensive priced genres?

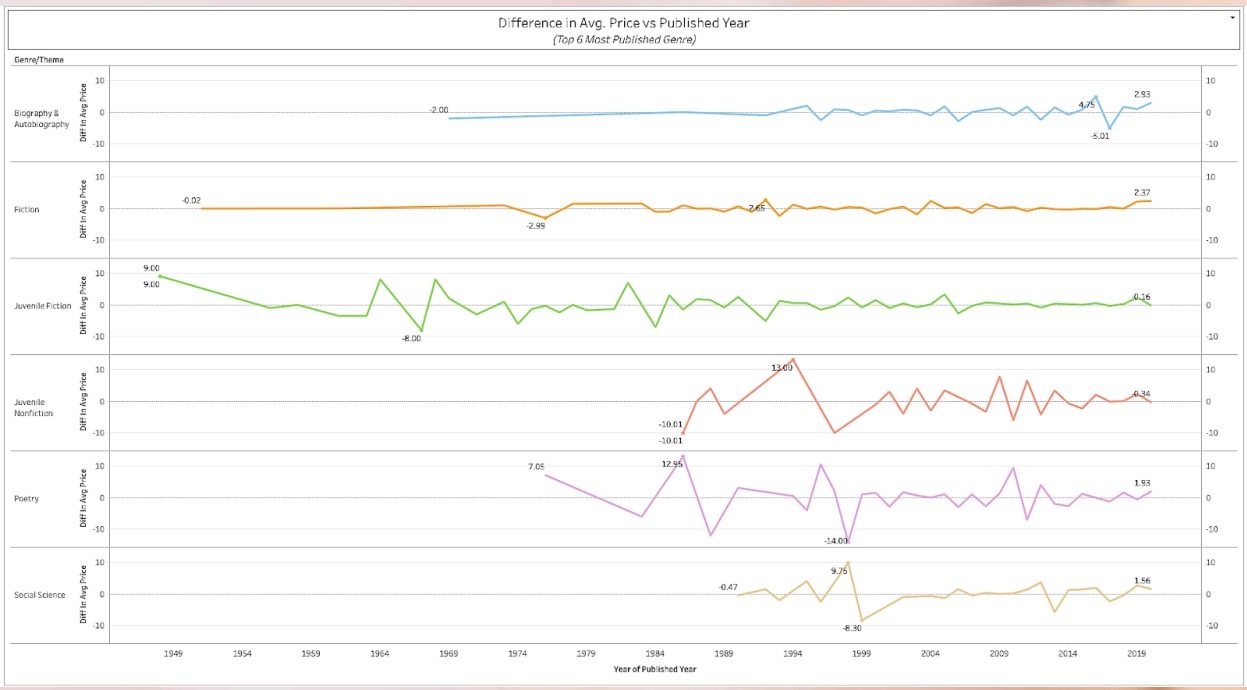
Ans: The most expensive book genre is art followed by cooking.



**Question 5:**

How does the price vary across the years for genres?

Ans: The graph depicts how the average price for genres has fluctuated across the years. Poetry and Juvenile fiction based books have had the most fluctuation in prices through the years.



# Text Analysis

We performed text analysis on the scraped dataset from web scarping part 2. We used the 112,000 rows of reviews which was scraped for the top 10,000 books. The main idea was to get the sentiment measures for each review, group it by the book title, compute average sentiment value and merge it with the book attributes dataset for further analysis.

The following python libraries were used for text analysis:

**Re** - To deal with all string and text related regular expression operations.

**NLTK** - For symbolic & statistical natural language processing of the English language.

**ScikitLearn** - To represent the text and string data into arrays and vectors form.

**TextBlob** - For processing textual data, speech tagging, noun phrase extraction, and sentiment analysis.

**WordCloud** - Open source library to visually represent textual data.

The following preprocessing operation were carried out before text analysis :

Cleansing text, Lemmatization, Stemmatization and Removing stopwords.

For text analysis, we used the 112,000 rows of reviews which was scraped.

```
In [531]:
```

```
scraped_reviews_copy
#More than 112K reviews scraped
```

| | Book Title | Critic Reviews |
|---|---|---|
| 0 | A Promised Land | "Barack Obama is as fine a writer as they come... |
| 1 | A Promised Land | Chimamanda Ngozi Adichie, The New York Times B... |
| 3 | Caste (Oprah's Book Club) | "Magnificent . . . a trailblazing work on the ... |
| 4 | Caste (Oprah's Book Club) | O: The Oprah Magazine"This book has the reverb... |
| 5 | Caste (Oprah's Book Club) | Dwight Garner, The New York Times"A surprising... |
| ... | ... | ... |
| 114770 | Typewriters, Bombs, Jellyfish | [...less] |
| 114771 | Big Bad Bubble | * "Rubin's voice-over narrator counsels Yerbur... |
| 114772 | Big Bad Bubble | that moment when the monsters |
| 114773 | Big Bad Bubble | or whatever lurks under their particular beds |
| 114774 | Big Bad Bubble | start to worry them."--The New York Times |

112894 rows × 2 columns

Reviews with sentiment value.

| | Book Title | Critic Reviews | Clean_Critic Reviews | sentiment | Polarity | Subjectivity |
|---|---|---|---|---|---|---|
| 0 | A Promised Land | “Barack Obama is as fine a writer as they come... | barack obama fine writer come promised land n... | (0.2630952380952381, 0.47857142857142854) | 0.263095 | 0.478571 |
| 1 | A Promised Land | Chimamanda Ngozi Adichie, The New York Times B... | chimamanda ngozi adichie new york time book re... | (0.13636363636363635, 0.45454545454545453) | 0.136364 | 0.454545 |
| 2 | Caste (Oprah's Book Club) | “Magnificent . . . a trailblazing work on the ... | magnificent trailblazing work birth inequalit... | (0.75, 0.8) | 0.750000 | 0.800000 |
| 3 | Caste (Oprah's Book Club) | O: The Oprah Magazine“This book has the reverb... | oprah magazine book reverberating patriotic sl... | (0.32857142857142857, 0.28214285714285714) | 0.328571 | 0.282143 |
| 4 | Caste (Oprah's Book Club) | Dwight Garner, The New York Times“A surprising... | dwight garner new york time surprising arresti... | (0.21818181818181817, 0.45227272727272727) | 0.218182 | 0.452273 |
| ... | ... | ... | ... | ... | ... | ... |
| 112889 | Typewriters, Bombs, Jellyfish | [...less] | le | (0.0, 0.0) | 0.000000 | 0.000000 |
| 112890 | Big Bad Bubble | * "Rubin's voice-over narrator counsels Yerbur... | rubin voice narrator counsel yerburt froofle ... | (0.12095238095238096, 0.5780952380952381) | 0.120952 | 0.578095 |
| 112891 | Big Bad Bubble | that moment when the monsters | moment monster | (0.0, 0.0) | 0.000000 | 0.000000 |
| 112892 | Big Bad Bubble | or whatever lurks under their particular beds | whatever lurks particular bed | (0.16666666666666666, 0.3333333333333333) | 0.166667 | 0.333333 |
| 112893 | Big Bad Bubble | start to worry them."--The New York Times | start worry new york time | (0.13636363636363635, 0.45454545454545453) | 0.136364 | 0.454545 |

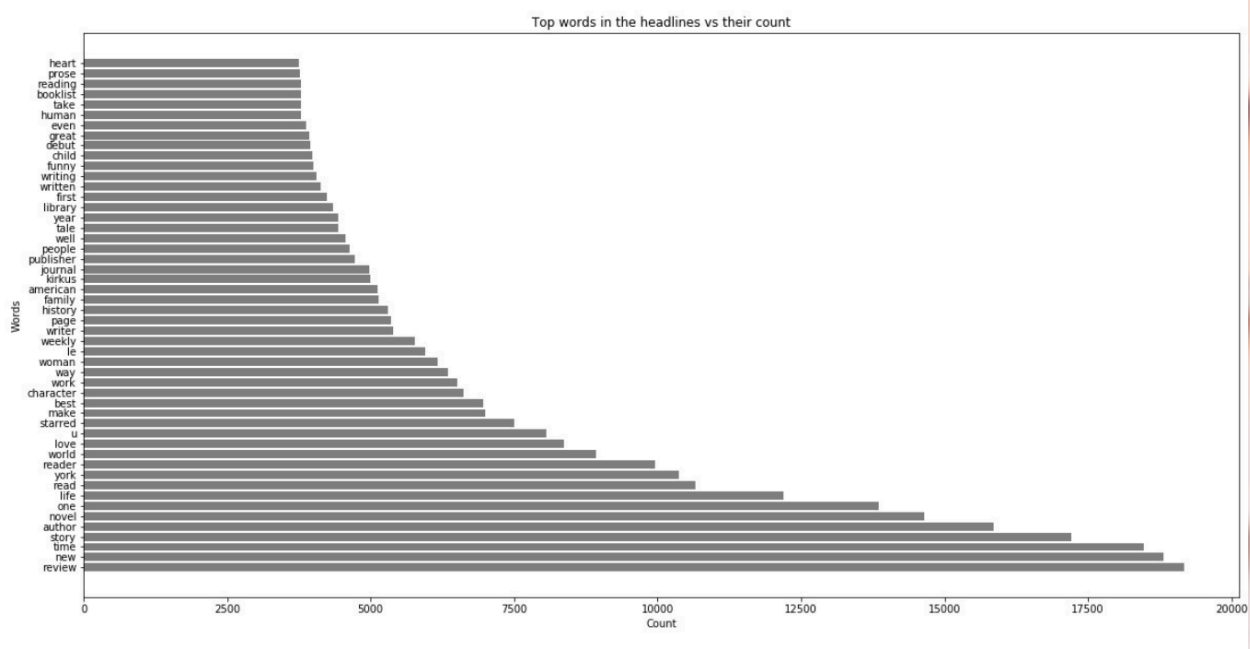
112894 rows × 6 columns

Book attribute dataset and Book reviews dataset merged together

| | Title | Author | Price | Sales Rank | Genre/Theme | Published Month | Published Year | Mean_Polarity | Mean_Subjectivity |
|---|---|---|---|---|---|---|---|---|---|
| 0 | A Promised Land | Barack Obama | 45.00 | 1 | History | Nov | 2020 | 0.199729 | 0.466558 |
| 1 | The Cold Millions | Jess Walter | 28.99 | 2 | Fiction | Oct | 2020 | 0.202203 | 0.639696 |
| 2 | Caste (Oprah's Book Club) | Isabel Wilkerson | 28.80 | 3 | History | Aug | 2020 | 0.181603 | 0.431806 |
| 3 | The Deep End (Diary of a Wimpy Kid Book 15) | Jeff Kinney | 13.49 | 4 | Juvenile Fiction | Oct | 2020 | 0.188393 | 0.349554 |
| 4 | Rhythm of War | Brandon Sanderson | 34.99 | 5 | Fiction | Nov | 2020 | 0.254038 | 0.486132 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 7495 | The Bed Moved | Rebecca Schiff | 15.95 | 79327 | Fiction | Feb | 2017 | 0.265144 | 0.513147 |
| 7496 | The Yellow Bird Sings | Jennifer Rosner | 25.99 | 79344 | Fiction | Mar | 2020 | 0.243665 | 0.519479 |
| 7497 | The Flapper Queens | Trina Robbins | 34.99 | 79349 | History | Aug | 2020 | 0.000000 | 0.000000 |
| 7498 | Nico Bravo and the Hound of Hades | Mike Cavallaro | 12.99 | 79350 | Juvenile Fiction | Apr | 2019 | 0.151562 | 0.253125 |
| 7499 | Life | Keith Richards | 19.99 | 79351 | Biography & Autobiography | May | 2011 | 0.118628 | 0.339715 |

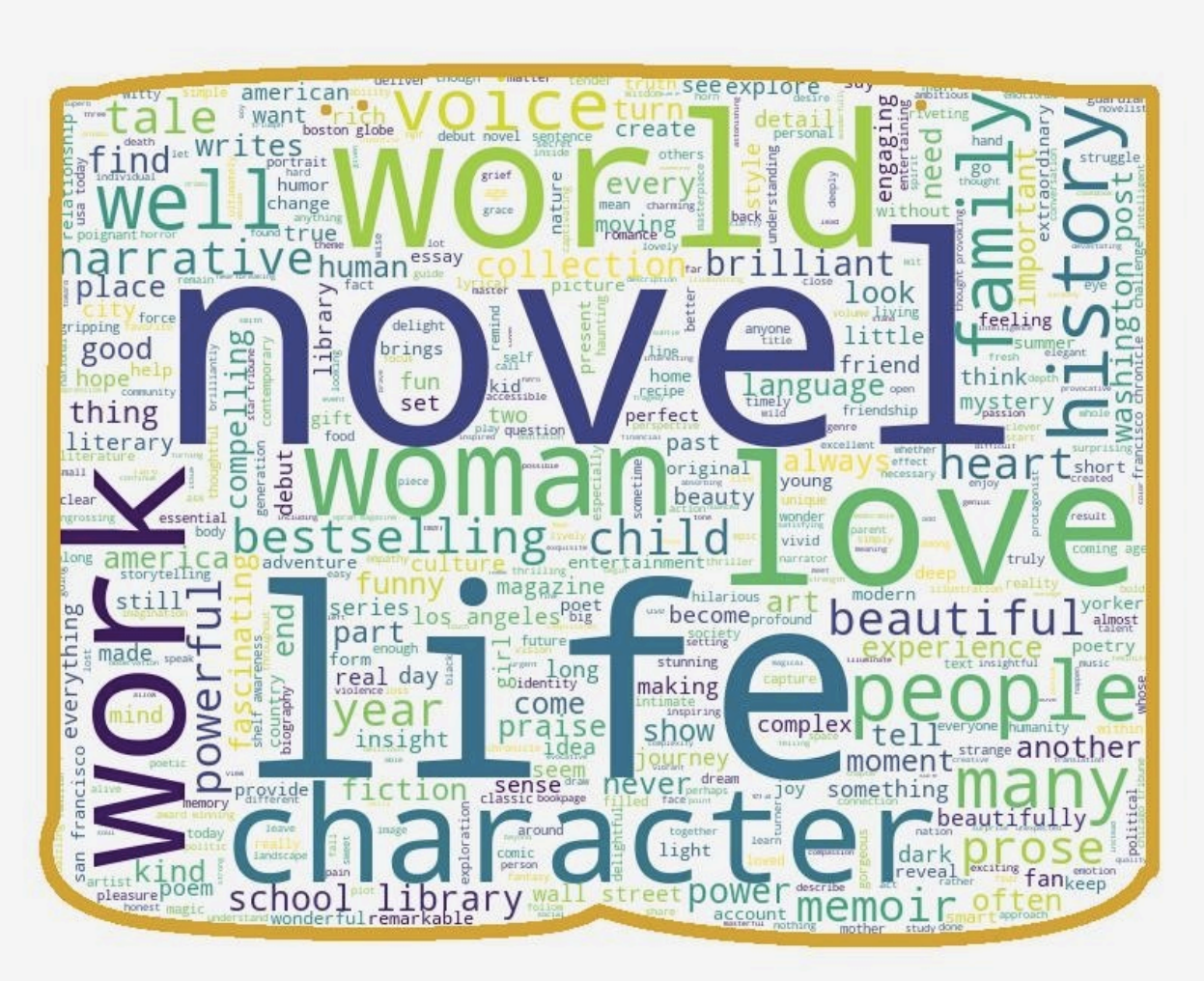
7500 rows × 9 columns

The most frequent words in the reviews column are displayed and meaningless words were added to create a custom stopwords list.



## Word Cloud

The image depicts the most frequent words in the reviews section. Bigger and bolder words mean more repetitive words.



## Sentiment Analysis

Using TextBlob Library, we computed to the sentiment measures values for each review. The key aspect of sentiment analysis is to analyze a body of text for understanding the opinion expressed by it. The sentiment measures consist of polarity and subjectivity values.

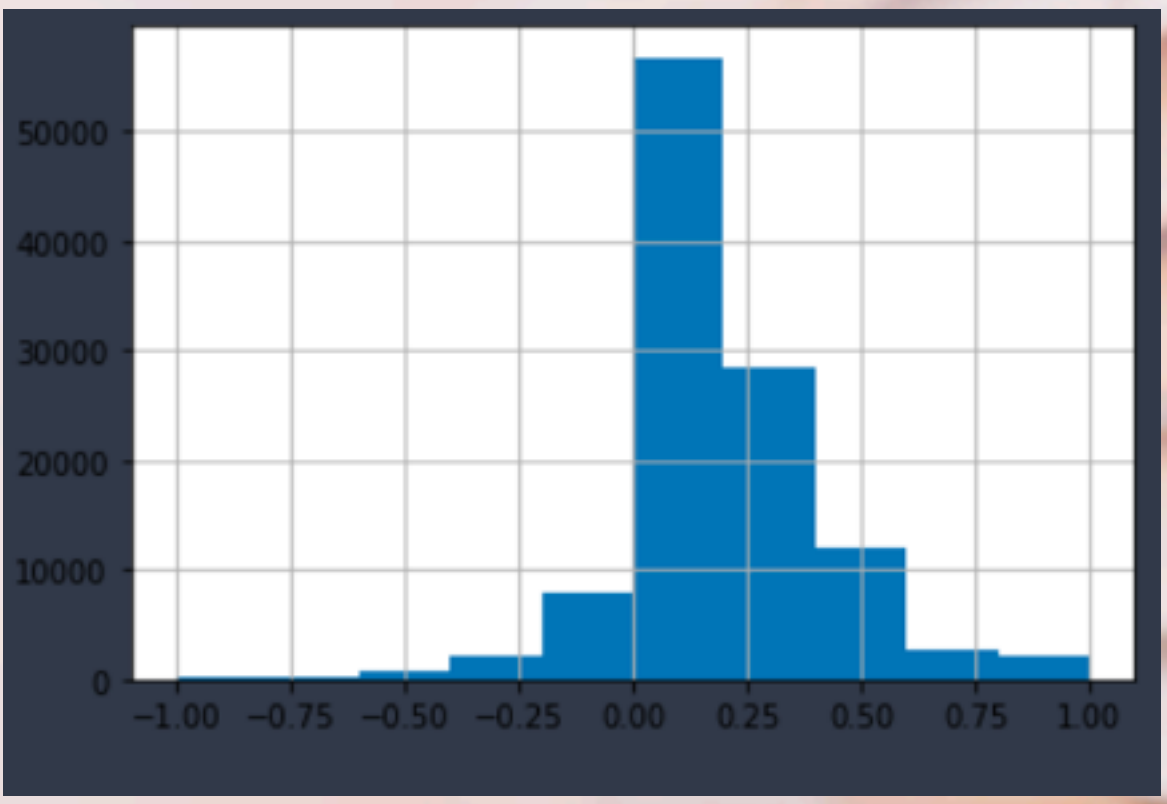
Polarity is a measure which represents the levels of positivity in a text. Its a float value ranging from -1 to 1.

Subjectivity is a measure which represents the levels of subject relevance in a text. Its a float value ranging from 0 to 1. For our project, we decided to filter out all reviews below a subjectivity of 0.1.

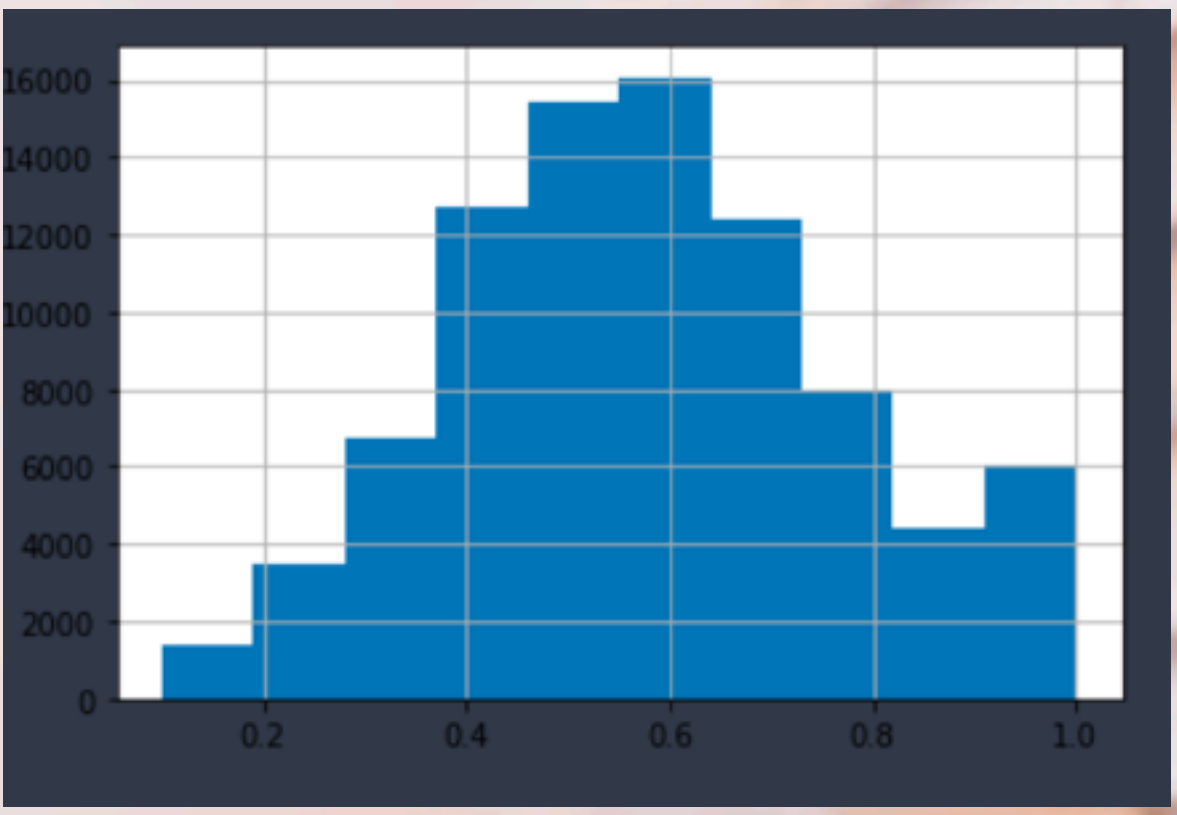
We grouping the sentiment value for each reviews by Book Title to get an aggregated value which was later merged to the book attribute dataset.

After sentiment analysis, it was observed that most of the reviews were positive and with high subjectivity. The following graph depicts the polarity and subjectivity measures.

## Polarity Graph:



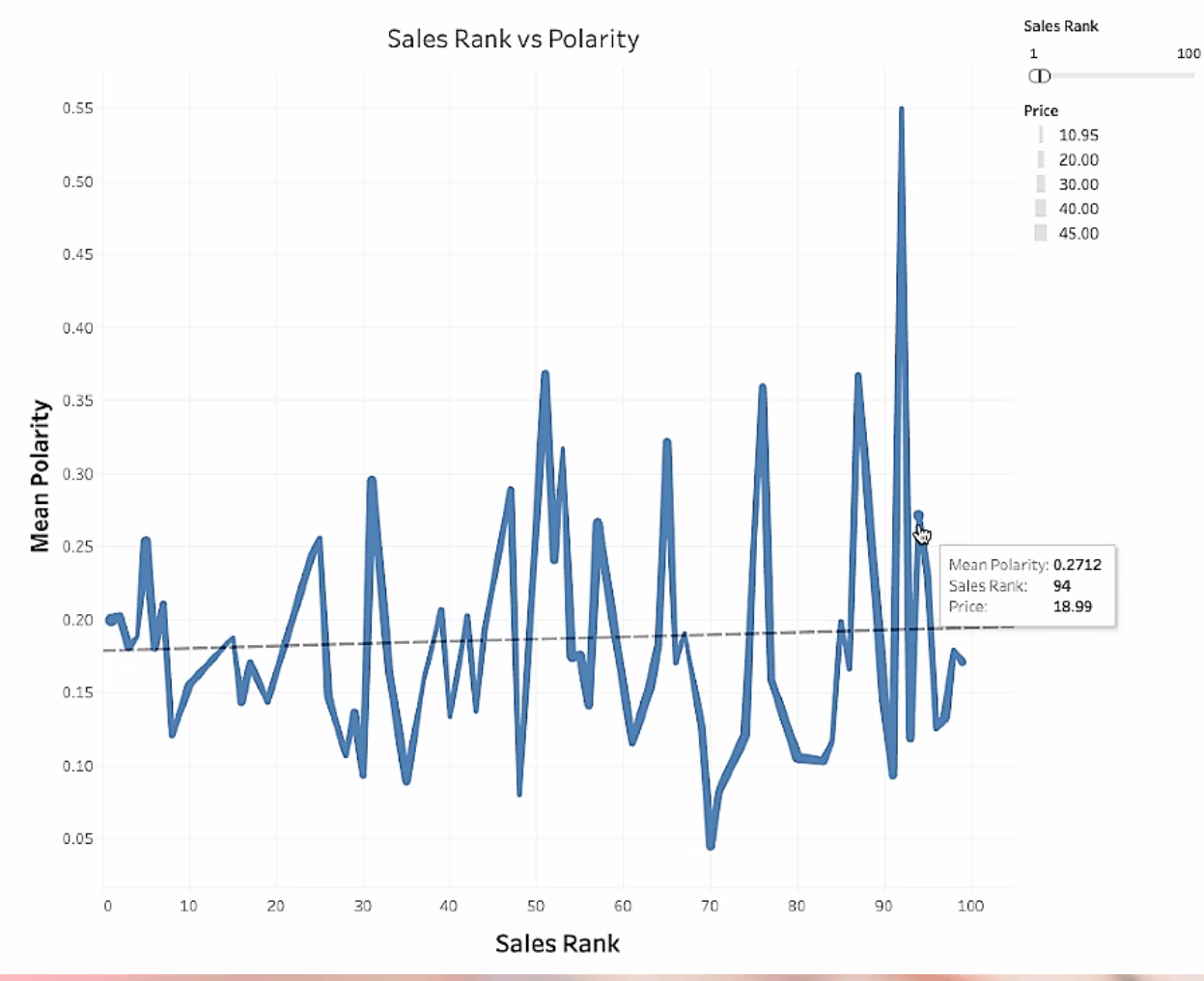
## Subjectivity Graph:



# Results

In order to analyze the sentiment measures and compare with sales performance for each book, we grouped the sentiment value for each reviews by its Book Title to get an aggregated value and merged it to the book attribute dataset. To derive final results, we compared the polarity values for each book to its sales rank using Tableau.

The following graph compares the mean polarity versus the sales rank for the top 100 ranked books. Ideally, the polarity values should have decreased as the sales rank increased. But as observed from the graph, the general trend line depicts a gradual increase in polarity values as sales rank increase. Thus it can be summarized that a **positive critic review doesn't guarantee a success for a book in terms of sales and alternatively, critic reviews don't have a major influence in book sales.**



Using tableau we analyzed further to inspect whether critic reviews are genre/theme subjective.


It's interesting from the graph to observe **that critic reviews are subjective when it comes to reviews across different genres. For genres, politics and poetry, the general trend line decreases ie Polarity value decreases as rank increases which means critic reviews do have an influence over sales for these genres.**

# Conclusion

This project was successfully implemented and the following goals were achieved:

1) Web scraping www.booksaremagic.com for Books Attributes & Book Reviews.

2) Performed EDA using the scraped dataset.

3) Answered key questions using text analysis:

   a) Do critic reviews have a major influence in sales?

   b) Are critics reviews genre - biased?

Few challenges faced were:

1) The volume of data scraped was large and scraping dynamic JavaScript elements along with book attributes was very time consuming. This problem was solved by automating selenium using headless Phantom JS browser.
2) Dealing with fields that have no book attributes/no reviews.
3) As the scraping consisted of two parts and two different datasets were generated, combining the scraped book attribute dataset and the book review dataset for analysis was tedious

Future Scope:

Due to the limited time, this project was carried out only for book reviews. The project can be extended by scraping and performing same analysis operation on author description and author reviews section to yield interesting results.

# References:


1) www.booksaremagic.net/ : Website from datasets were scraped for top 10,000 books
2) Sentimental Analysis of Book Reviews using Unsupervised Semantic Orientation and Supervised Machine Learning Approaches https://ieeexplore.ieee.org/document/8753089
3) https://thisiswriting.com/why-are-book-reviews-important-for-authors/
4) crummy.com/software/BeautifulSoup/bs4/doc/
5) https://www.nltk.org/