SHOULD WE JUDGE A BOOK BY ITS COVER?

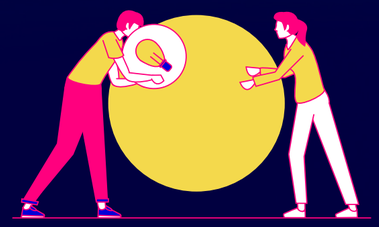
By Emotional Bots

Ashish Kumar,
Divyanshu Torlikonda,
Ravindra Manne,
Vaishnavi M R

Project Outline

- Problem statement
- Data Scraping
- Data Cleaning
- Dataset Description
- Data Visualization
- EDA
- Cover Image color classification
- Conclusions and Future scope

Problem Statement



- Analyzing the dataset of the best-selling books on the Amazon website in order to find more about the characteristics and trends of the most popular books across different genres.

- To observe if there is any relationship between the best selling books and their respective cover pages

How did we get the data ?

- Source: amazon books
  https://www.amazon.in/gp/bestsellers/books/ref=zg_bs_unv_books
- Using Scraper API, we parsed the data from 'Bestseller books' section of Amazon website
- The data spanned over top 100 books across 54 genres as listed on the website
- Used Requests Library to scrape the HTML data
- Used BeautifulSoup to parse the data

Dataset Description

- Our dataset contains a total of 12 features.
- 'cover_img_url' : Contains URLs of book cover images
- 'Genre' represents the genre of respective books
- 'Title' represents the book's title

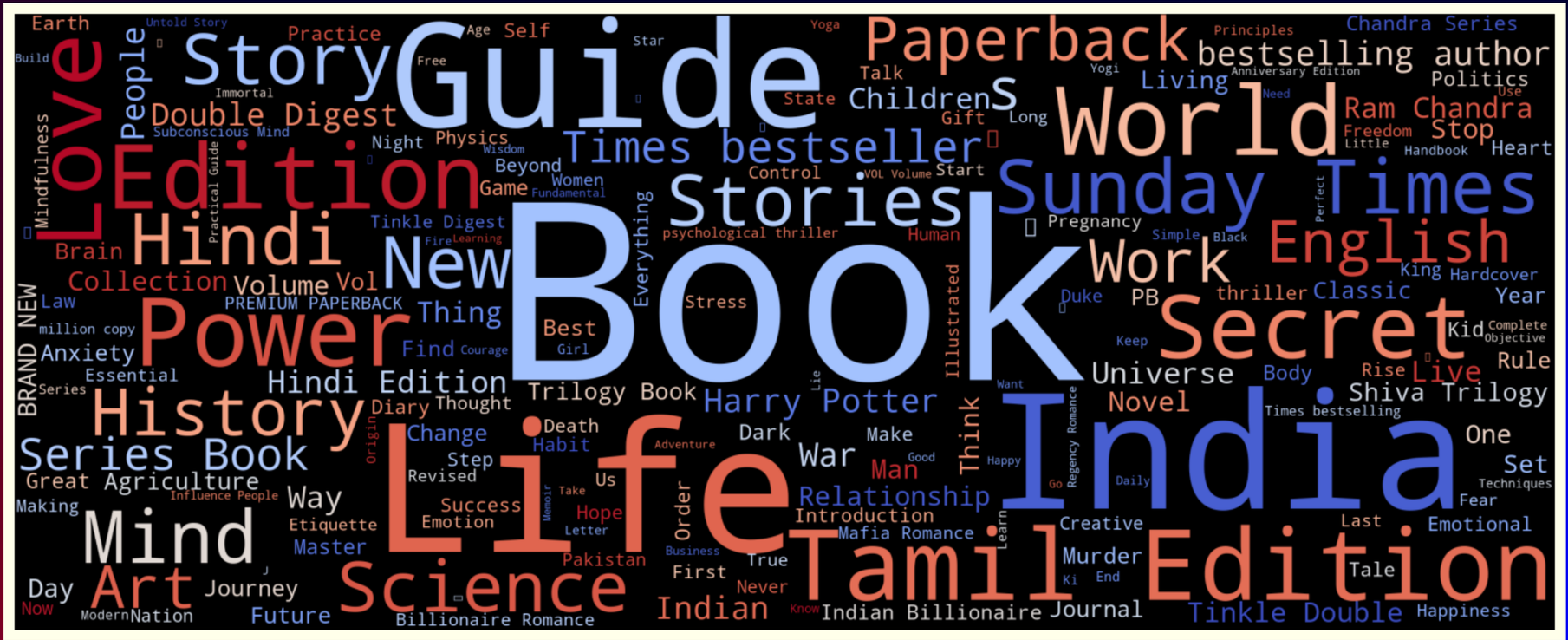
Data columns (total 12 columns):			
#	Column names	No of data points	Data type
1	Genre	4680	object
2	URL	4680	object
3	Title	4680	object
4	Price	4680	float64
5	five_star_rating	4680	float64
6	four_star_rating	4680	float64
7	three_star_rating	4680	float64
8	two_star_rating	4680	float64
9	one_star_rating	4680	float64
10	overall_rating	4680	float64
11	No_of_ratings	4680	float64
12	cover_img_url	4680	object

Rest all columns are concerned with the ratings provided by the buyers

Data Cleaning

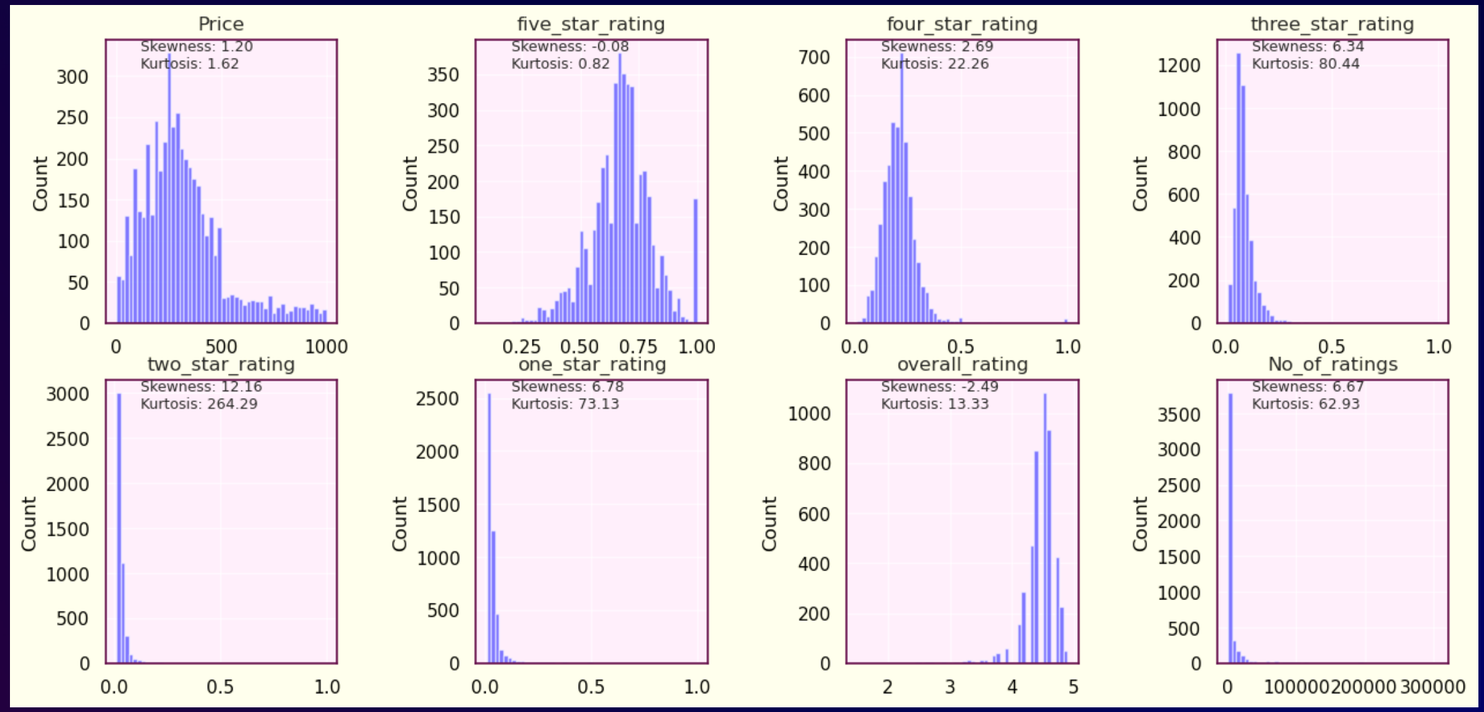
- Dropped rows which contain null values in 'cover_img_url' and 'Title' columns
- Order of of best-selling books during scraping determined the ranking
- Specified the rankings of the books for each genre on the basis of available data hierarchy

Data Visualization



We can deduce that the words like : Book, Life , India, Guide, Edition , etc are most occuring words in the Title of all the books across all genres

Data Visualization



- Many of the books are not rated by the buyers
- This can be because of the reason that not many books were highly negatively rated

Exploratory Data Analysis

We can deduce that :

Overall rating is -

->Number of five star rating is high implies overall rating is high

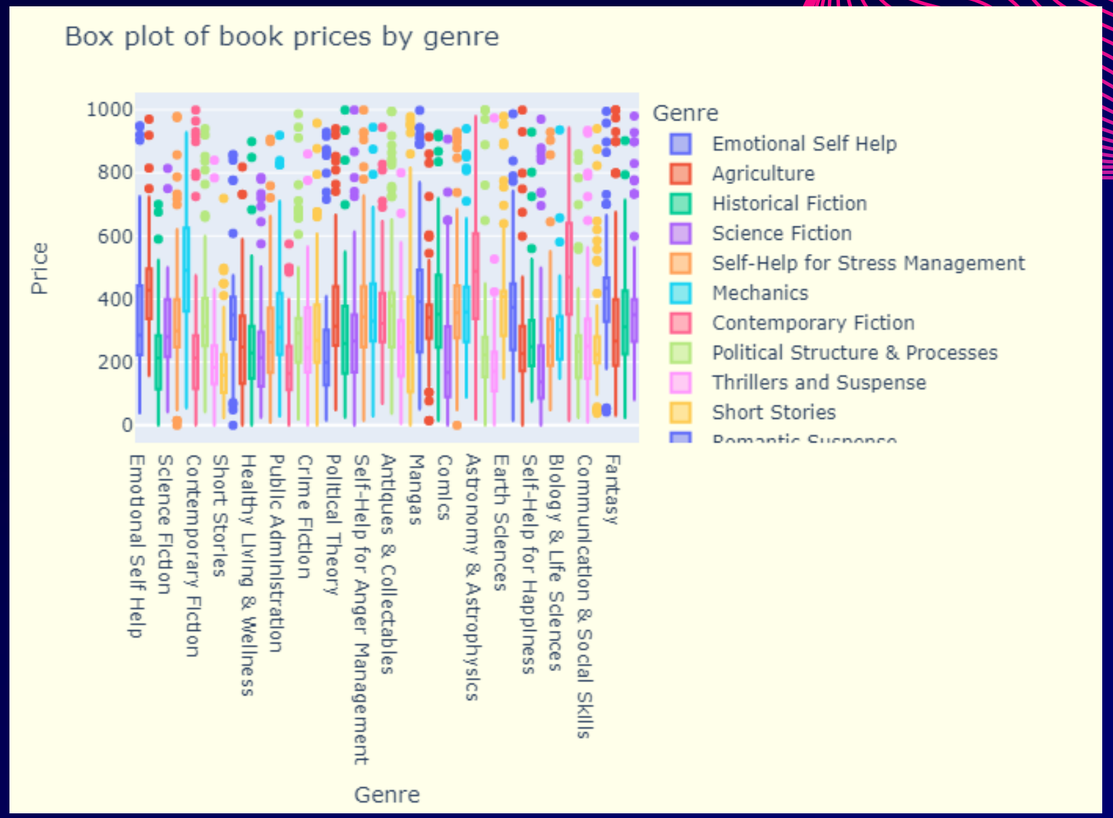
->Number of one star rating is high implies overall rating is low

Exploratory Data Analysis

1. Books in the genres like : Maps and Atlases, Exam Preparation, Business Economics, Crime Thriller and Mystery , Society and Social Sciences, Textbooks and Study Guides , Reference and Romance are having median < Rs. 200.

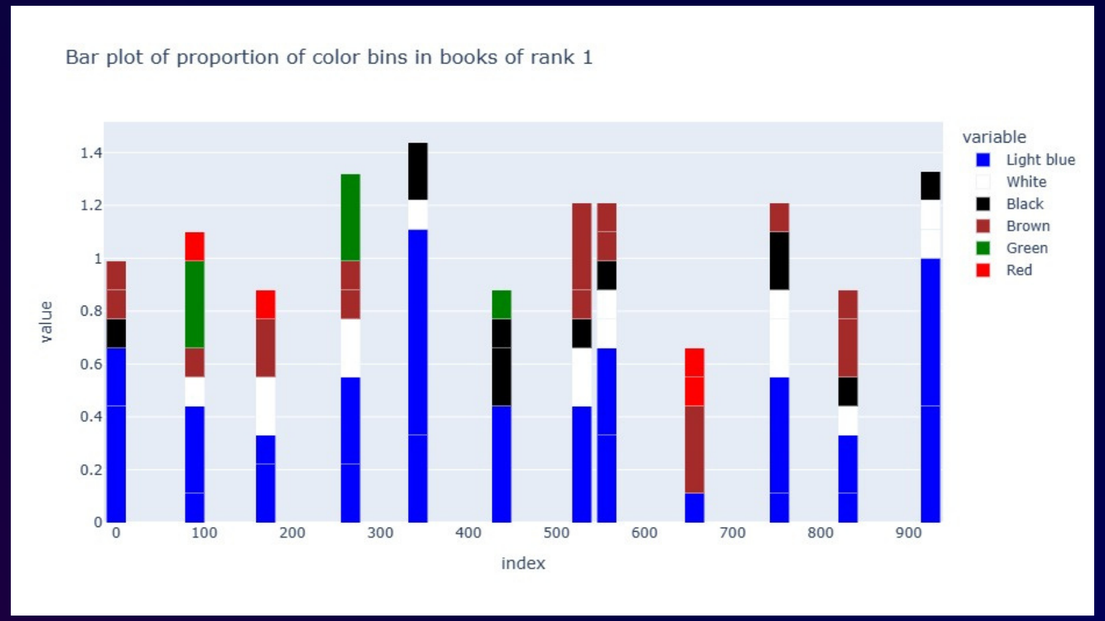
It means that many of the bestsellers are in this genres are highly affordable

2. There is no fixed range for the pricing of the books in the genres Children and Young Adult.



Box plot

Exploratory Data Analysis



1. From the stacked plots above, we can observe that light blue color is present in the greatest amount for the bestsellers.

2. This information is insufficient to draw conclusions if the cover pages really affect the popularity of a book.

Conclusion & Future Scope

1. We hypothesize that the cover images do not really affect the popularity of a given book. To prove our claim, we need to further devise CNNs on the collected data

2. We can also use NLP techniques to analyze the sentiments of customers and see which factors really do affect the popularity of a book


THANKYOU