JUDGING A BOOK BY ITS COVER

WHAT MAKES A COVER SELL?
A VISUAL ANALYSIS OF BESTSELLING BOOKS

PATRICIA GIMENEZ

AGENDA

1.Project Overview

2.Data Sources

3.Dataset Summary

4.Python Web Scraping Process

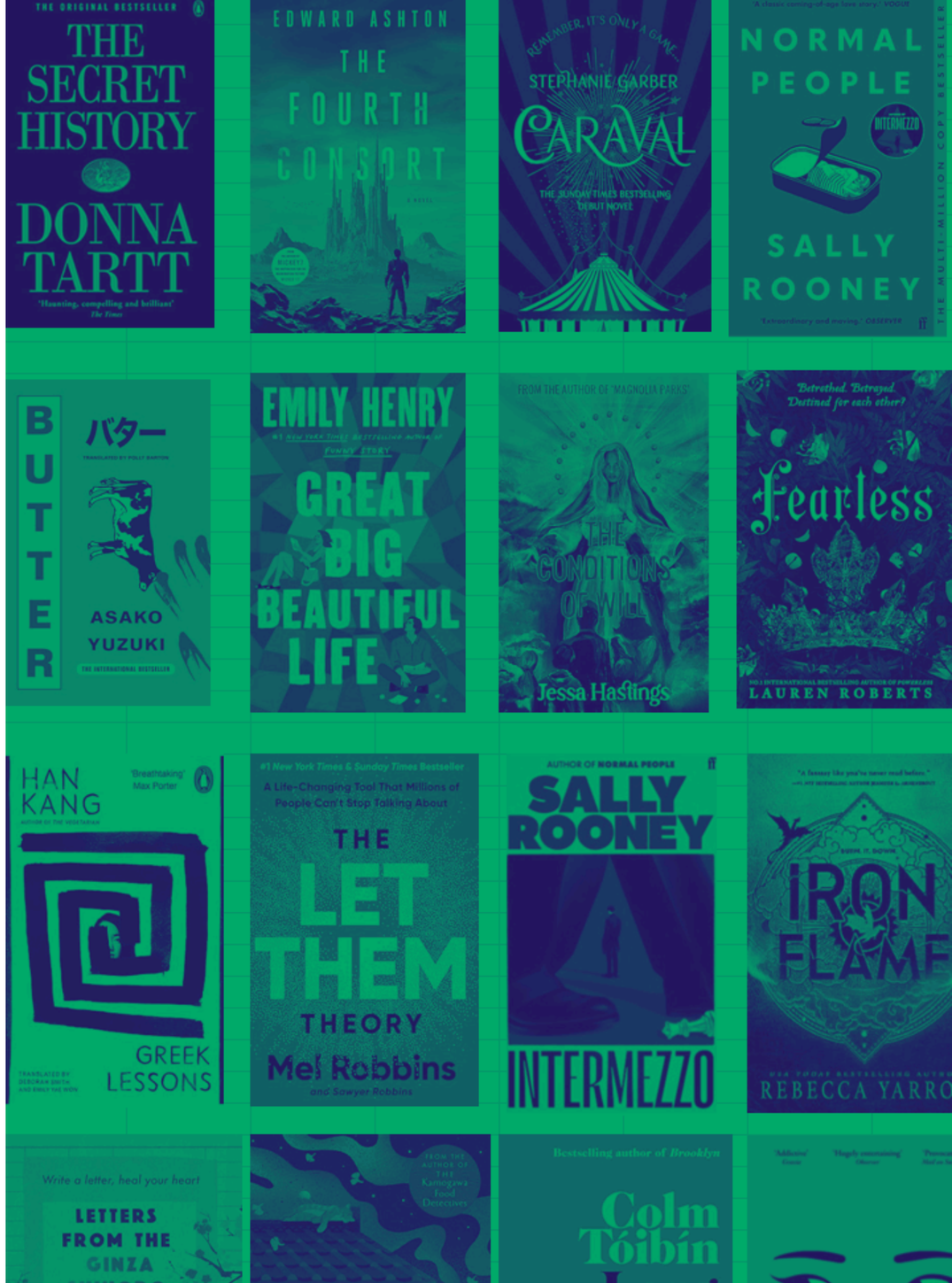
5.Exploratory Data Analysis (EDA)

6.Univariate Analysis

7.Multivariate Analysis

8.Inferential Statistics

9.10 Guidelines for Bestsellers



PROJECT
OVERVIEW

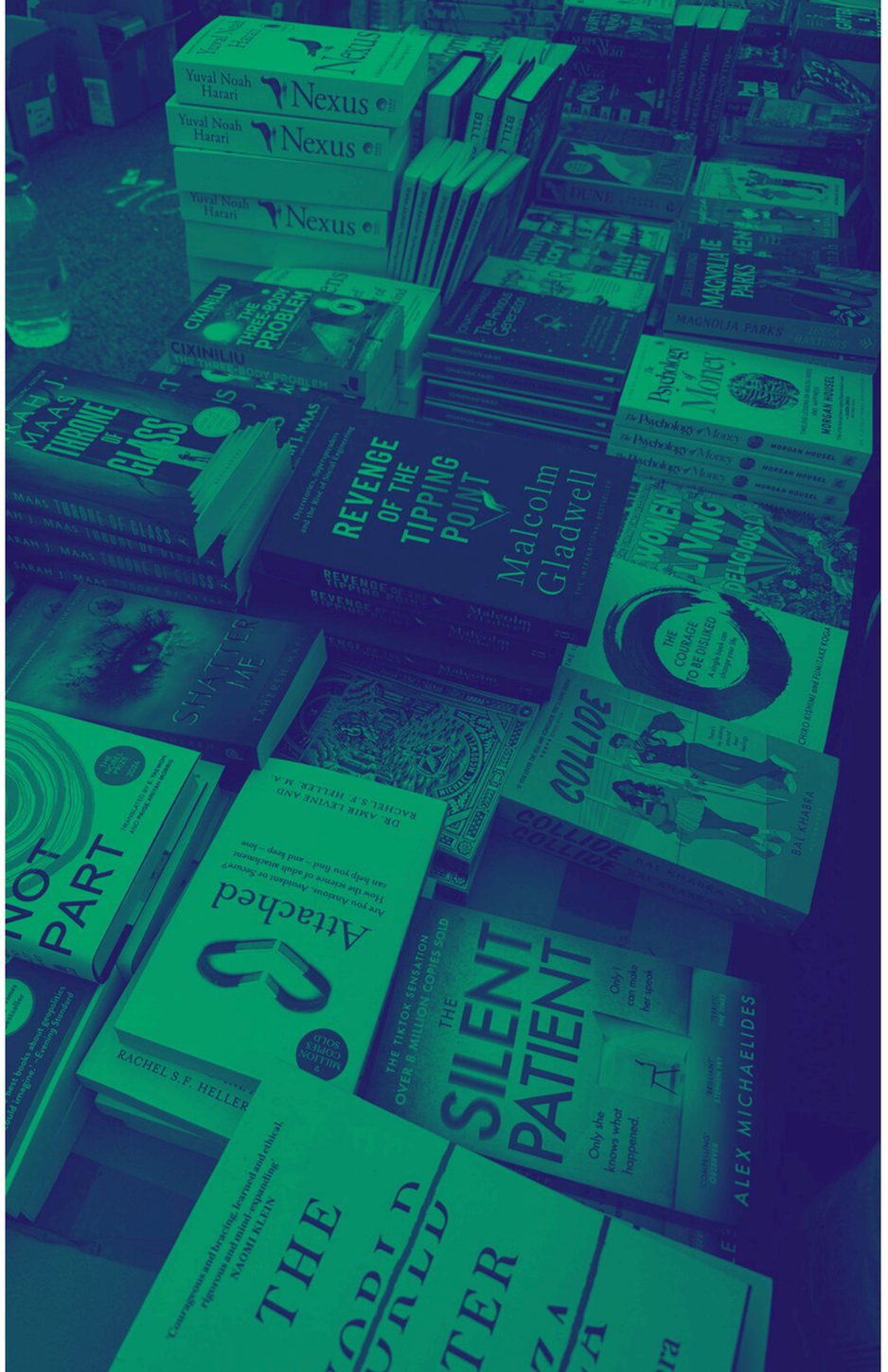
This project investigates whether certain **visual elements** are more **common** among **popular book covers.**

I manually collected **80 book covers** from bestseller and recommended sections in physical bookstores. For each book, I recorded key **visual features** such as title size, dominant color, font type, and illustration style.

To enrich this primary dataset, I performed **web scraping on Goodreads** to retrieve additional metadata:

- Page count
- Rating
- Top 2 genres
- Publisher
- First published date

The goal is to explore whether some visual patterns consistently appear in successful or widely recommended books.



DATA *SOURCES*

Primary

- 80 book covers selected from bestseller and recommended sections in physical bookstores
- Visual elements were recorded manually: dominant color, visual type, etc

Scraped

Extracted using Python (BeautifulSoup) due to limited access to the Goodreads public API

- Page count
- Average rating
- Top 2 genres
- Publisher
- First published date

Derived/Calculated

Calculated fields in Excel

- Title length (word count)
- Estimated reading time (in hours, calculated from page count using average reading speed)

Title	Author	Publisher	Genre	Price	Promoted	Dominant_color	Cover_type	Visual_styl	Title_word_count	hor_name_prominent	Award_recognition	age_count	Estimated_reading_time_hours	ting	First_published_date
1984	George Orwell	Plume	Classics, Fiction	10.7	1	black	paperback	illustration	1	1	0	368	8	4.2	08/06/1949
A Court of Thorns and F	Sarah J. Maas	Bloomsbury Publishing	Fantasy, Young Adult	9.5	0	red	paperback	illustration	6	0	1	419	9	4.17	05/05/2015
A Palace Near the Wing	Ai Jiang	Titan Books	Fantasy, Novella	17.1	0	blue	hardcover	illustration	5	1	1	192	4	3.35	15/04/2025
All About Love	Bell Hooks	William Morrow	Nonfiction, Feminism	15.1	0	red	paperback	typographic	3	1	1	240	5	4.03	22/12/1999
Atomic Habits	James Clear	Avery	Nonfiction, Self Help	19.2	1	white	paperback	typographic	2	0	1	319	7	4.34	18/10/2018
Babel	Rebecca F. Kuang	Harper Voyager	Fantasy, Historical Fiction	14	0	black	paperback	illustration	1	1	1	544	11	4.16	23/08/2022
Baumgartner	Paul Auster	Grove Atlantic	Fiction, Audiobook	20.9	0	grey	hardcover	photo	1	1	0	208	4	3.74	07/11/2023

DATASET SUMMARY

df.shape

Total columns: 16
Total books/rows: 80

Column Name	Type	Description	Source
Title	Text	Full title of the book Used as primary key for matching external data	Primary
Author	Text	Author’s name as it appears on the book cover	Primary
Publisher	Text	Name of the publishing house, scraped from Goodreads.	External
Genre	Text	Top 2 genres listed on Goodreads for the book (e.g. Fiction, Fantasy, Romance).	External
Price	Numeric	Book price observed in-store or online (€)	Primary
Promoted	Binary	1 = Displayed face-out (front cover visible), 0 = Placed spine-out on a regular shelf	Primary
Dominant_color	Text	Main color seen on the cover (e.g. red, blue, white…)	Primary
Cover_type	Categorical	Format of the physical edition: hardcover or paperback	Primary
Visual_style	Categorical	Type of visual element used on the cover: - illustration: drawn or graphic art - photo: photographic image - typographic: text-only design - symbolic: icons, symbols, or abstract minimal elements	Primary
Title_word_count	Numeric	Number of words in the book’s main title, calculated field	Derived
Author_name_prominent	Binary	1 = author’s name is visually prominent on the cover, 0 = small or secondary prominent (1) if at least one of the following is true: - The author’s name covers 30–40% or more of the cover width - The font size of the author’s name is equal to or larger than the title -The author’s name is placed at the top of the cover - The author’s name is displayed in a bold, bright or contrasting color that draws attent	Primary
Award_recognition	Binary	1 = a literary prize or printed recognition is visibly shown on the cover, 0 = no badge or mention	Primary
Page_count	Numeric	Total number of pages in the book, as recorded on Goodreads	External
Estimated_reading_time	Numeric	Estimated reading time in hours, based on 275 words per page and a reading speed of 200 wpm	Derived
Rating	Numeric	Average rating of the book on Goodreads, on a scale from 0 to 5	External
First_published_date	Date	Full original publication date of the book, as recorded on Goodreads. The date has been formatted for compatibility with spreadsheet	External

PYTHON *WEB SCRAPING PROCESS*

Search & Match

Locate book pages on Goodreads by using the full book title.
We use a helper function to search and return the first relevant book link automatically

Scrape Details

Once on the book page, we use requests and BeautifulSoup to extract key metadata:

- Page count
- Average rating
- Top 2 genres
- First published date
- Publisher

Clean & Store

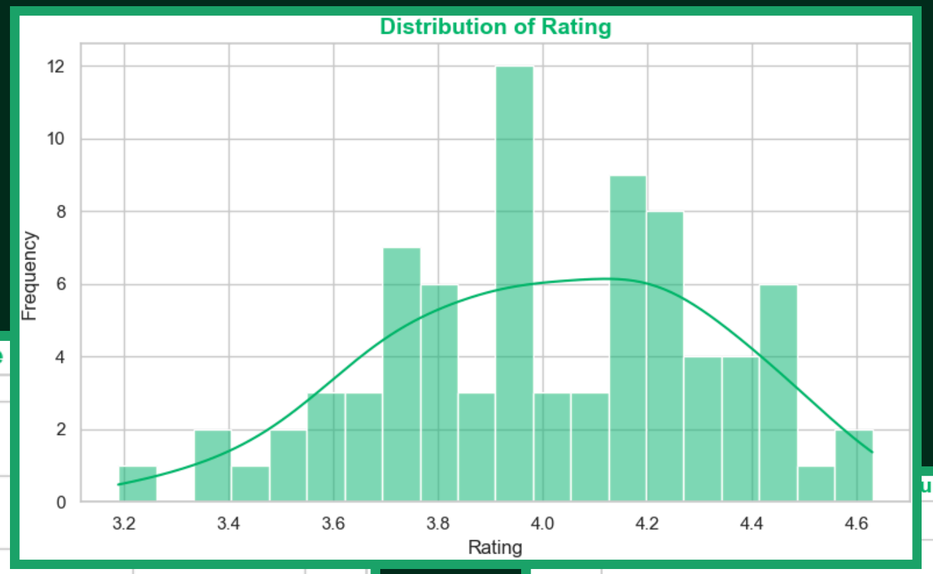
The extracted data is cleaned, formatted, and stored in new lists.

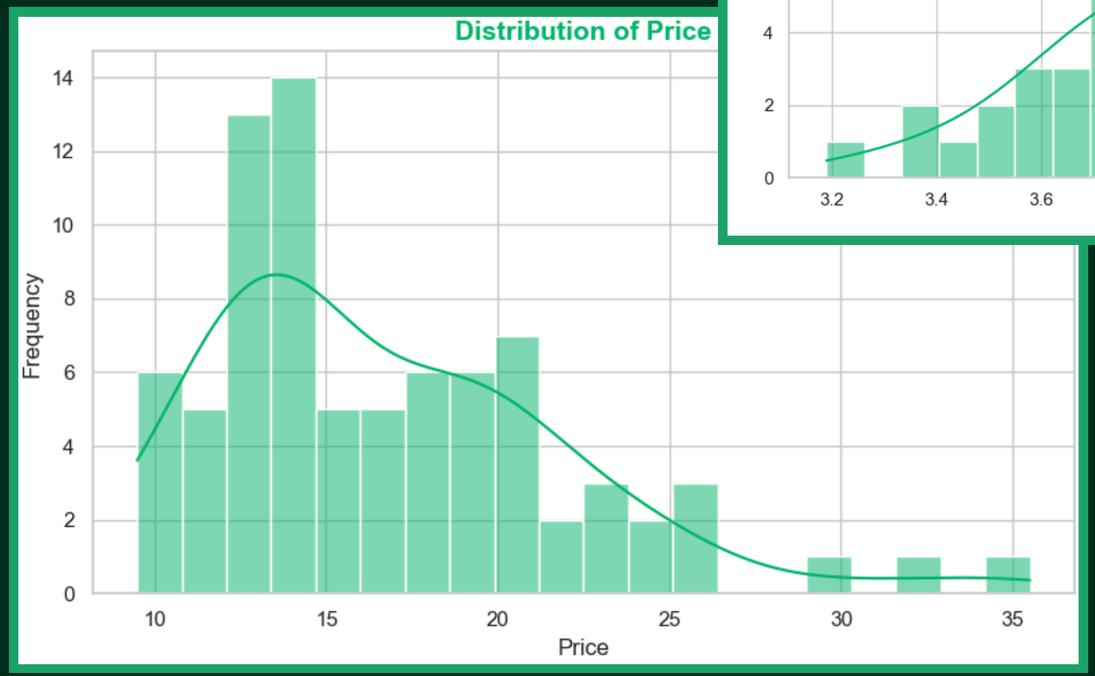
- Dates are formatted for Excel compatibility
- Null values are handled
- Results are saved into new .csv file

EXPLORATORY *DATA* ANALYSIS

01	Understand the Data	Reviewed a curated dataset of 80 bestselling and recommended books. Explored variable types (text, numeric, binary) and their relevance
02	Normalize & Clean	Created a new main genre column using keyword rules for simplified grouping. Converted column types (e.g., price to numeric, dates to datetime) for accurate analysis.
03	Univariate Analysis	Analyzed key features individually: price ranges, typical page counts, rating distributions, and dominant cover colors. Identified patterns aligned with commercial trends
04	Multivariate Analysis	Explored variable combinations to reveal insights. Used scatterplots, heatmaps, and stacked bars to highlight trends.

UNIVARIATE *ANALYSIS*

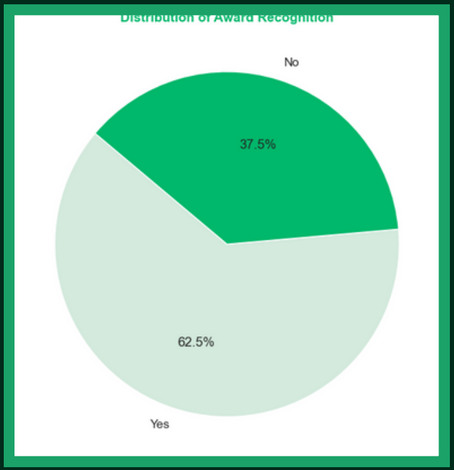


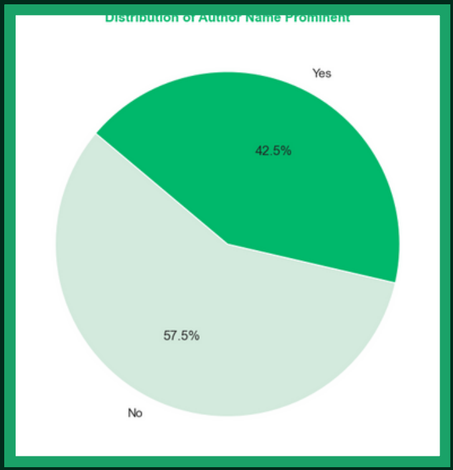


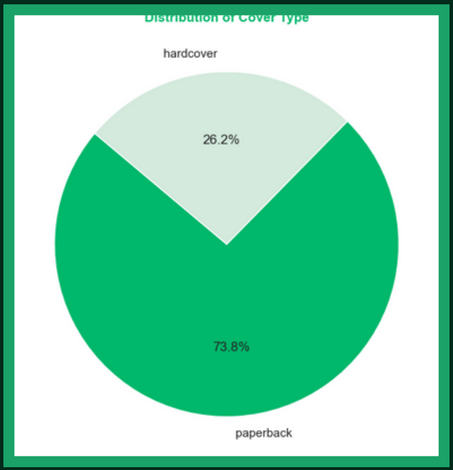
Numerical Variables

- **Price:** Most books cost between €10 and €20; few exceed €30
- **Page Count:** Typically between 300–500 pages; one outlier >1300
- **Rating**: Mostly between 3.5 and 4.5, indicating strong reception
- **Estimated Reading Time:** Often ranges from 4 to 12 hours
- **Title Word Count:** Short titles are common (2–4 words)itle Word Count: Short titles are common (2–4 words)
- **First Published Date**: Clear peak in books from 2022–2024, highlighting recent or upcoming titles

UNIVARIATE *ANALYSIS*

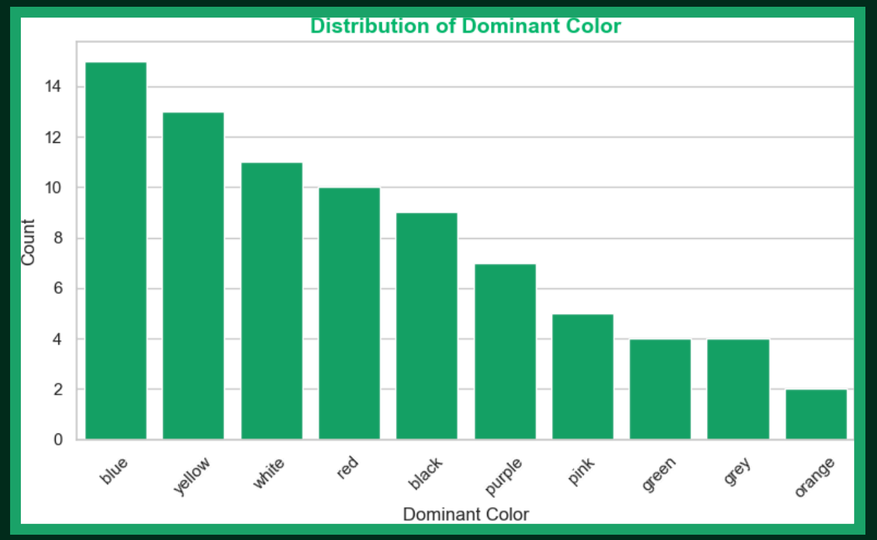


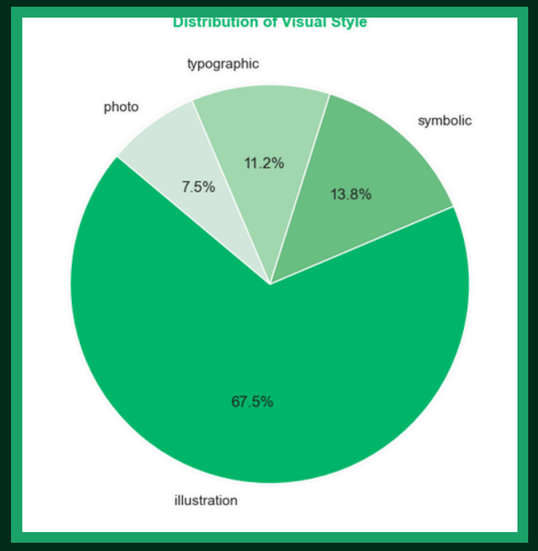


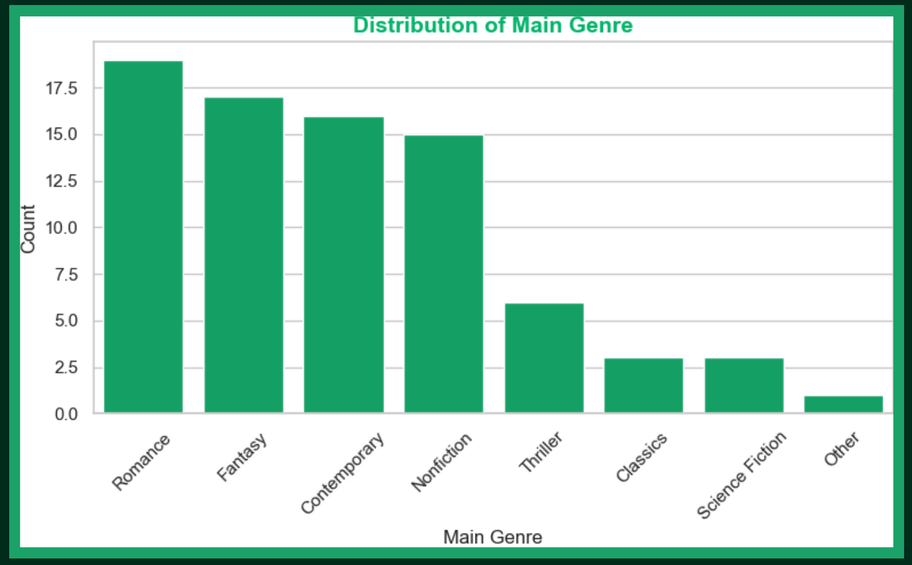


Categorical Variables

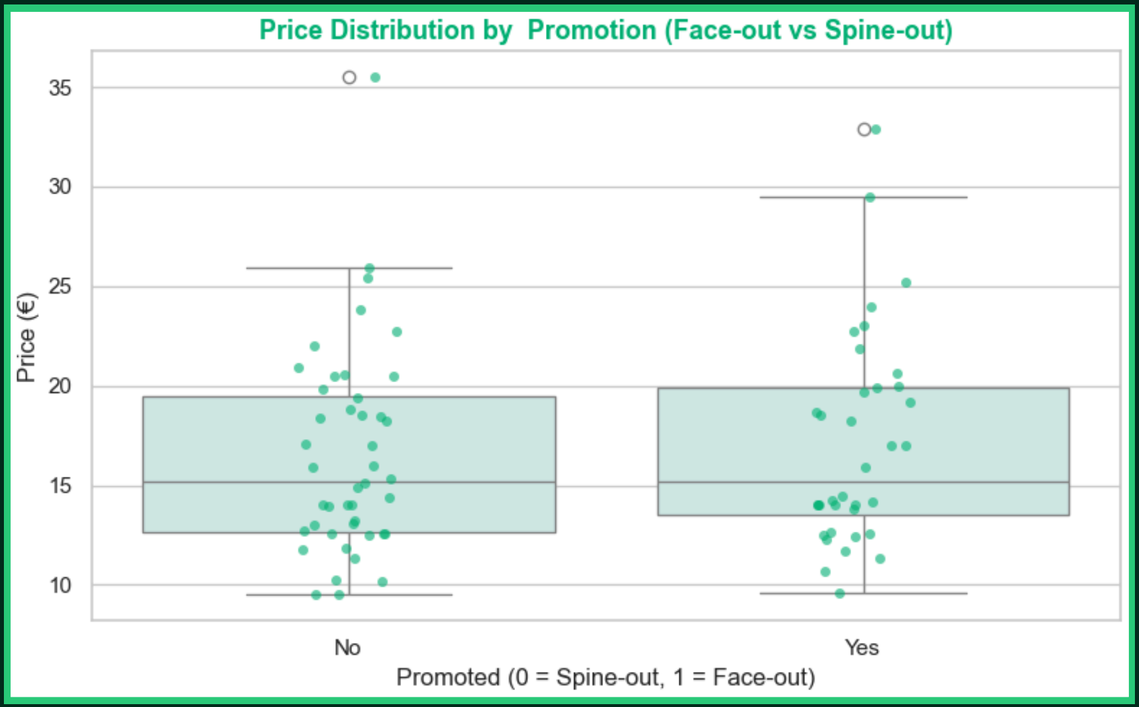
- **Promoted:** 45% of books are face-out (promoted), 55% spine-out
- **Visual Style:** Illustration is the dominant cover style
- **Dominant Color:** Most common colors are white, yellow, blue
- **Genre:** Top genres include Fantasy, Romance, Nonfiction
- **Cover Type:** Paperback is more frequent than hardcover
- **Author Prominence:** 42.5% of covers highlight the author visually
- **Award Recognition:** 62.5% of books show awards or recognitions

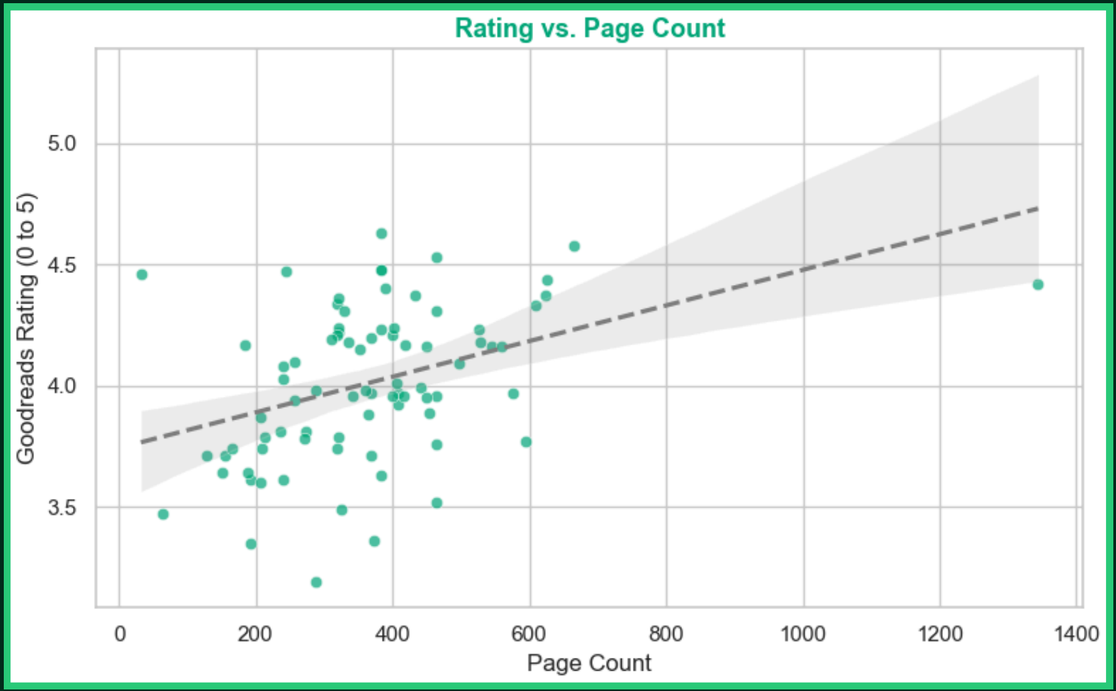






MULTIVARIATE ANALYSIS





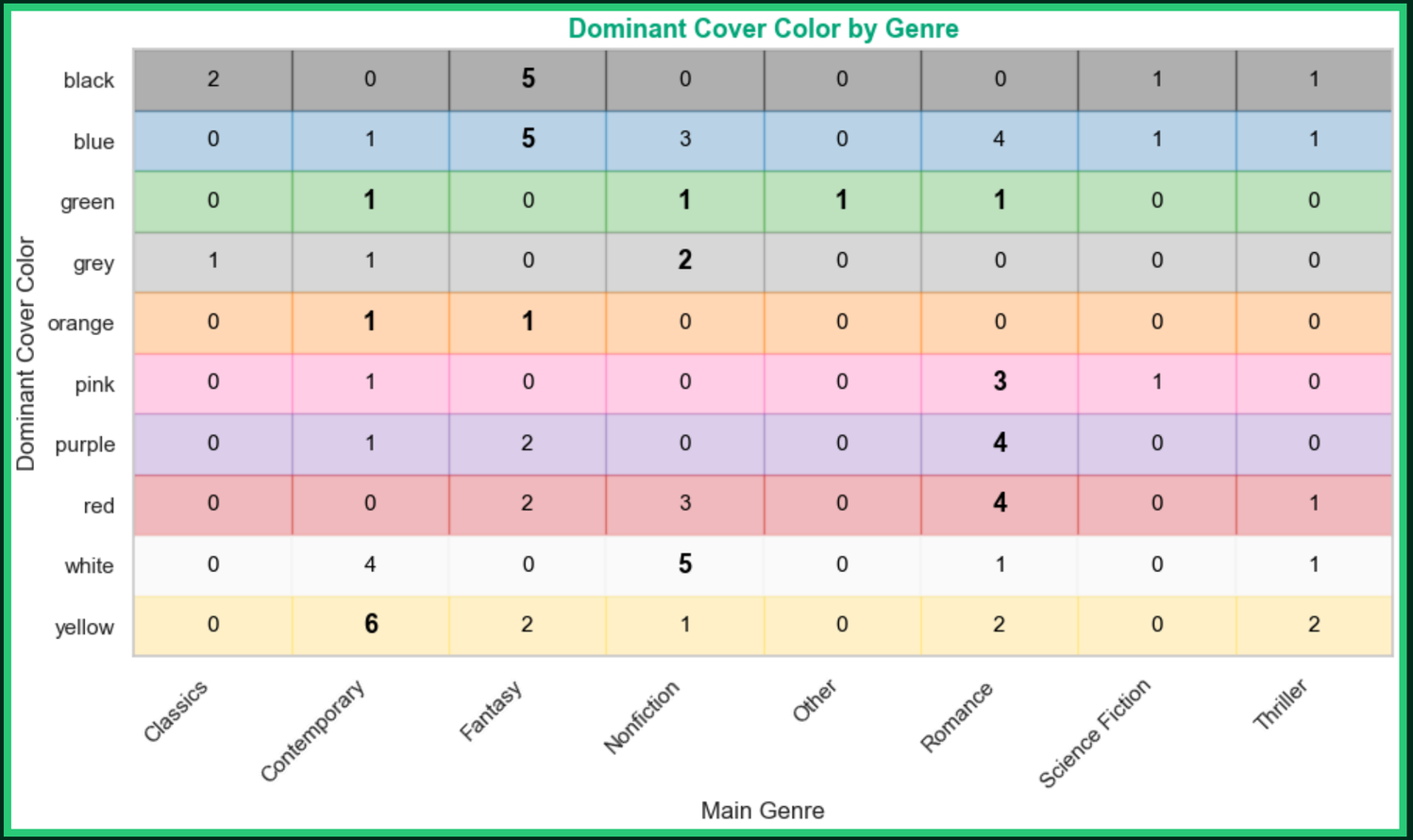
Promoted books are not significantly more expensive, though the highest-priced titles are slightly more likely to be promoted.

There's a very weak positive correlation—longer books tend to get slightly higher ratings, but page count alone is not a strong predictor.

Are promoted books more expensive?

Do longer books receive lower ratings due to reader fatigue?

MULTIVARIATE ANALYSIS

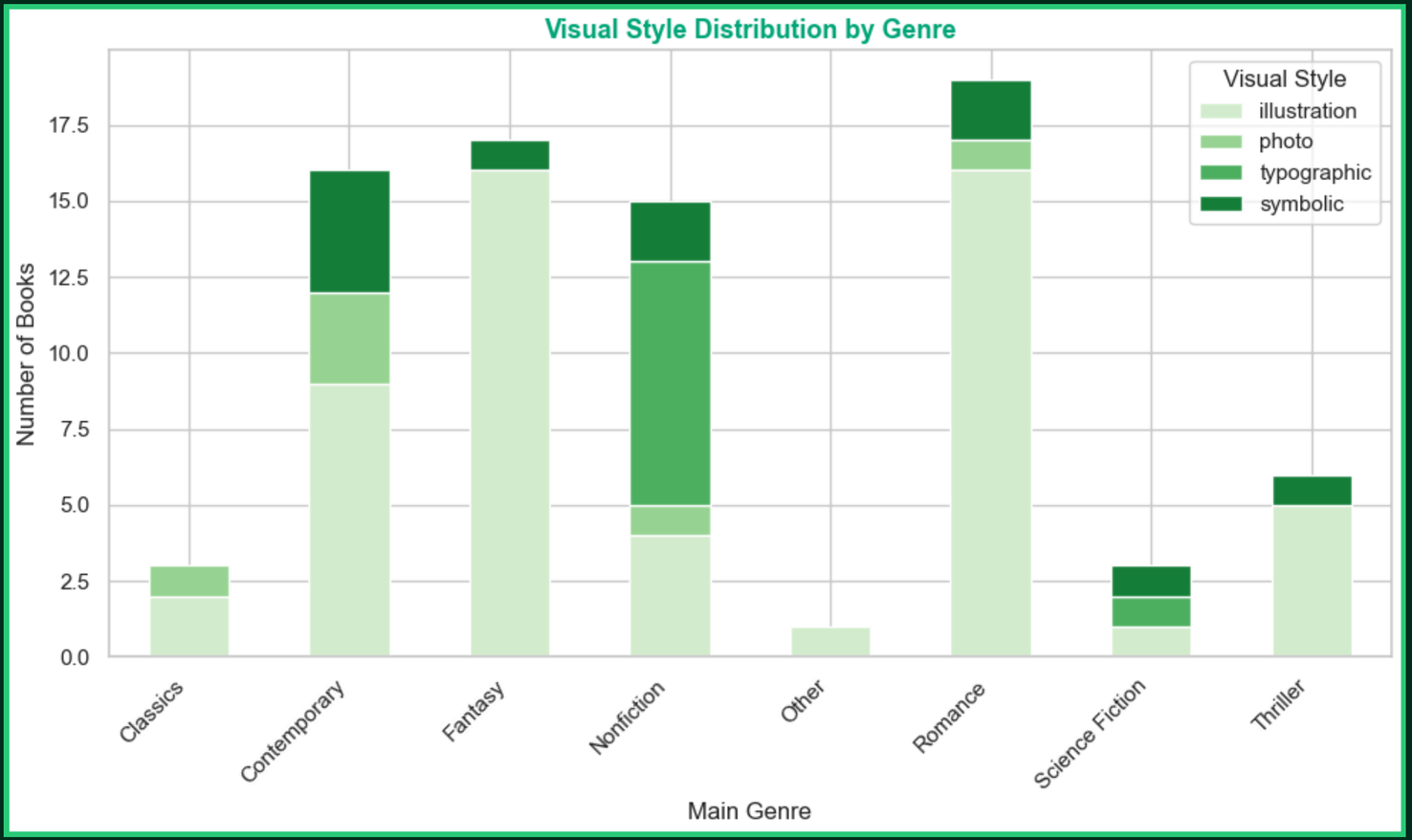


The matrix reveals several notable patterns in the relationship between cover color and book genre:

- **Fantasy books are frequently associated with black and blue covers, suggesting a strong stylistic convention in this genre.**
- **Romance titles often use red, pink, or purple, reinforcing their connection to emotional and intimate themes.**
- **Nonfiction books show a clear preference for white, which may reflect a clean and neutral design style.**
- **Contemporary fiction tends to use yellow more than other genres, possibly to stand out visually.**

Is blue often used in Fantasy covers?

MULTIVARIATE ANALYSIS



The stacked bar chart reveals strong genre-based preferences for certain visual styles on book covers:

- **Illustration dominates across nearly all genres, especially in Fantasy and Romance, where it's the overwhelming choice. This aligns with their tendency to evoke imagination and emotion.**
- **Contemporary books show more diversity, incorporating photo, typographic, and symbolic elements more frequently than other genres.**
- **Nonfiction stands out for its typographic and symbolic styles, likely due to their emphasis on clarity and authority.**
- **Science Fiction and Thriller also feature a mix of styles, but tend to favor illustration and symbolic approaches.**

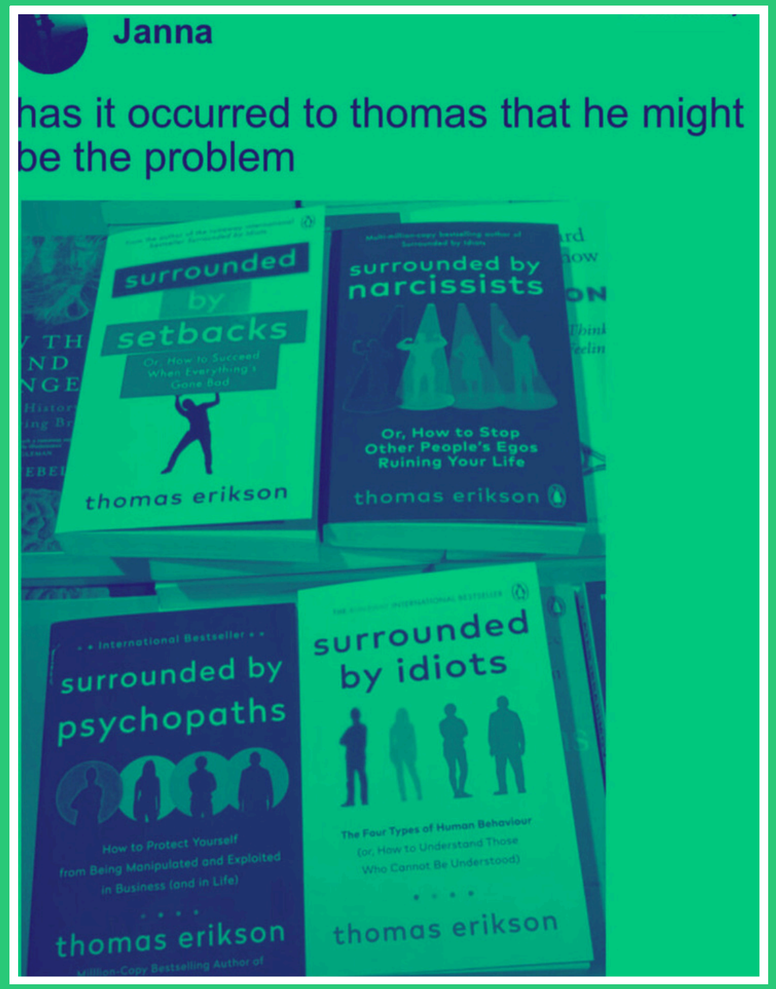
Which visual styles dominate each genre?

INFERENTIAL *STATISTICS*

CONFIDENCE INTERVALS: HOW CONSISTENT ARE GOODREADS RATINGS ACROSS BESTSELLING BOOKS?

Methodology	Results
• Computed 95% confidence interval for average Goodreads rating using t-distribution. • Based on a curated sample of 80 bestselling titles. • Repeated the process for the top 4 genres separately: Romance, Fantasy, Contemporary, Nonfiction.	Overall Rating: 4.01 • → CI: [3.94, 4.08] → ✓ Includes benchmark value 4.0 Genre Ratings (95% CI): • Romance: [3.88, 4.19] • Fantasy: [3.82, 4.28] • Nonfiction: [4.02, 4.23] • Contemporary: [3.74, 3.97]

Bestseller books tend to receive consistently high Goodreads ratings around 4.0. Nonfiction and Fantasy may score slightly higher than other genres.



INFERENTIAL *STATISTICS*

HYPOTHESIS TESTING: WHAT PATTERNS ARE STATISTICALLY SIGNIFICANT IN BESTSELLER COVER DESIGN?

Hypothesis	Test Type	Result
Promoted books are more recent	Mann-Whitney U (median year)	❌ Not significant
Awarded/ recognition books have shorter titles	Mann-Whitney U (median word count)	❌ Not significant
Cover style depends on genre	Chi-squared test	✅ Significant association

- Some industry assumptions (like using shorter titles with awards) don't hold statistically
- However, visual style is strongly genre-dependent, confirming editorial design patterns

10 GUIDELINES FOR *BESTSELLER COVERS*

Based on the analysis of 80 recent bestsellers, we propose a data-backed set of cover design guidelines

These guidelines can help publishers or designers optimize cover decisions aligned with genre and market expectation

1. 📏 **Keep Titles Short and Punchy**
   Most bestsellers use titles of **2 to 4 words**, making them easy to remember and visually strong on shelves.
2. 🎨 **Choose Illustration Over Photography**
   67.5% of covers use **illustration** as their main visual style—especially in genres like Romance and Fantasy.
3. 🎯 **Use Color to Match Genre**

- **Fantasy:** Black and Blue
- **Romance:** Red, Pink, Purple
- **Nonfiction:** White (neutral, clean)
- **Contemporary:** Yellow (stands out)

4. 📚 **Paperback Wins**
   74% of the books are paperback—an accessible format for mass-market success.
5. 💬 **Feature Awards or Badges**
   62.5% show a badge or recognition label on the cover—these act as **trust signals** for readers.

10 GUIDELINES FOR *BESTSELLER COVERS*

Based on the analysis of 80 recent bestsellers, we propose a data-backed set of cover design guidelines

These guidelines can help publishers or designers optimize cover decisions aligned with genre and market expectation

6. 👤 **Highlight the Author's Name (if famous)**
   43% of covers emphasize the author—especially effective for returning readers or established names.
7. 💸 **Stick to the €10–€20 Range**
   The sweet spot for pricing: **affordable but not cheap**, and appealing for impulse buyers.
8. 🗓️ **Go Recent or Go Home**
   Books published from **2022 onwards** dominate the sample, reinforcing the importance of recency in bookstore success.
9. 🕓 **Optimize for Medium Reading Time**
   The majority of bestsellers fall between **4 and 12 hours** of reading—engaging, but not overwhelming.
10. 📖 **Genre Dictates Design**
    Genres follow **clear style rules**:

- **Fantasy & Romance:** Illustrated
- **Contemporary:** Mixed styles
- **Nonfiction:** Typography or symbols for authority


THANK YOU!
THE SUNDAY TIMES BESTSELLER
From behind closed doors, she sees everything.
THE SUNDAY TIMES BESTSELLER
This perfect home has one rule. Don't look behind closed doors ...
BOOK OF THE YEAR
THE SEVEN HUSBANDS OF EVELYN HUGO
BLACKWATER IV
BLACKWATER V
BLACKWATER VI
Kathy Reichs
SALLY ROONEY
NORMAL
Harry Potter