جامعة الملك سعود
King Saud University

BOOK SALES TRENDS

IT362- Data Science

Supervised by Dr.Mashael Aldayel

INTRODUCTION

This project explores book data to identify factors that contribute to bestseller status, such as reviews, genres, and reader preferences. By uncovering key trends, the analysis offers insights to help publishers and authors improve marketing strategies and boost book popularity.

QUESTIONS

How do ratings and the number of reviews vary among bestsellers?

Are certain authors more likely to have their books become bestsellers?

What visual design patterns are consistently associated with bestseller book covers?

What genres are most represented among bestsellers?

What is the relationship between price and bestseller books? , What is the price range of bestseller books?

OUTLINE:

Data collection

Data Preprocessing

EDA

Model Evaluation

Technical hurdles

DATA COLLECTION METHODS

WEB SCRAPER

- Used to bypass server restrictions on Amazon and extract structured data from the website efficiently.

INSTANT DATA SCRAPER

- Used to quickly extract bestseller book URLs from the Jarir bookstore website without needing manual coding.

REQUESTS (PYTHON LIBRARY)

BEAUTIFULSOUP (PYTHON LIBRARY)

SELENIUM ,PANDAS AND TIME

- python libraries , tools and modules .

DATA PREPROCESSING

- Handling Missing Values
- Cleaning and Formatting Numerical Columns
- Handling Duplicate Rows and Data Aggregation
- Cleaning and Standardization
- language unification

IMAGE PROCESSING

We analyzed book covers to understand how visual features affect their appeal. Since many websites block direct image downloads, we used Selenium to simulate user browsing and capture cover images. Key features like contrast, edge complexity, color variance, and saturation were extracted using image processing techniques. We then applied K-Means Clustering to classify covers into “interesting” or “not interesting” categories. This approach provides a data-driven way to evaluate and improve book cover design.

EDA

- We analyzed Amazon and Jarir datasets to explore factors affecting book popularity, such as ratings, reviews, and other features.
- We visualized distributions and correlations.
- Amazon showed higher engagement, while Jarir had more attractive covers.
- Correlations were weak, highlighting platform-specific trends.

Tools Used:

- **Matplotlib** – Basic plots and figure layout
- **Seaborn** – Statistical visualizations (e.g., bar charts, violin plots)
- **Plotly** – Interactive rating distribution plots

MODEL EVALUATION

We applies clustering algorithms to identify visual design patterns in book covers using features like contrast, sharpness, and color properties. Key steps include preprocessing, applying K-Means, and Spectral Clustering, optimizing cluster count via Silhouette methods, and visualizing results with PairPlots and t-SNE.

MODEL EVALUATION

Model building Algorithms

Why chosen?

Evaluation Matrics

why chosen?
what do they reveal?

Model Improvement

How models could improved?

ALGORITHMS

SELECTED FOR MODEL BUILDING

K-Means Clustering

- Rationale: A baseline clustering algorithm that is simple yet effective when data exhibits spherical clusters.
- Implementation Details: Applied to standardized visual features, with the optimal value of k determined using the Silhouette score Method .

Spectral Clustering

- Rationale: Well-suited for non-convex clusters and works well on data manifolds like visual features reduced via t-SNE.
- Implementation Details: Used affinity='nearest_neighbors' to capture local structure in the visual design space.

EVALUATION METRICS

Silhouette Score

Why Chosen:

Measures how similar an object is to its own cluster compared to other clusters; ranges from -1 to 1.

Interpretation:

- Score near 1: Well-clustered.
- Score near 0: Overlap between clusters.
- Score near -1: Misclassified points.

Findings:

K-Means and Spectral Clustering both showed reasonable scores, supporting meaningful cluster formation in the visual space

Visual Inspection via PairPlots and t-SNE

Why Chosen:

To interpret cluster separability and coherence.

Findings:

Clusters revealed distinct groupings based on visual style parameters, validating algorithmic output.

HOW TO IMPROVE THE MODEL

Use better feature

- Add more useful information, like the popularity of the author, the type of words in the book title, or smarter image features using deep learning.

TECHNICAL HURDLES

Time-Consuming Web Scraping

Collecting data from multiple sources was slow and complex due to dynamic content and rate limits.

Unclear and Inconsistent HTML Structures

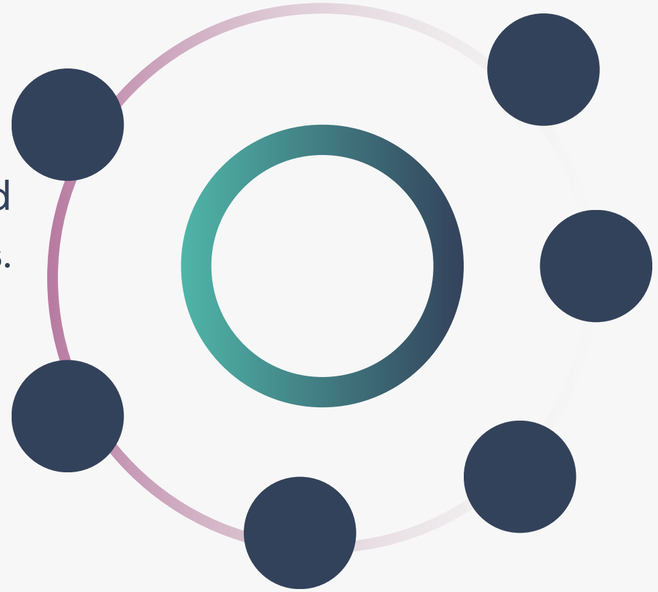

Image Quality & Cover Classification Challenges

Incomplete and Inconsistent Data

Choosing correct Model

Dynamic JavaScript-Loaded Content

Some data wasn’t present in the initial HTML, requiring extra handling.

HOW WE OVERCOME TECHNICAL HURDLES

Optimized Scraping Tools & Techniques

Used BeautifulSoup with asynchronous requests, Selenium for dynamic content, and added time delays to avoid blocks.

Focused on HTML Structure & Attributes

Targeted IDs and structural patterns instead of class names; used inspection tools to streamline extraction.

Enhanced Image Quality

Try several models and test them

Explored ML Techniques & Visualization Tools

Conducted literature review and used unsupervised learning for cover interest; selected graphs based on variable types.

Data Enrichment

Combined multiple data sources and performed manual reviews to fill gaps and ensure consistency.

THANK YOU!

Group #4
Razan Aldosari 444201215
Mona Alnajjar 444200091
Remas Al-subaie 444200712
Rima Alsonbul 444200524
Jana Alruzuq 444201157