



Group 2

**i** Sanjana Sankar, Kelly Lyons, Alisha Gunadharma Hartarto, Abigail Neoma, Leanne Chung, Yoojeong Seo

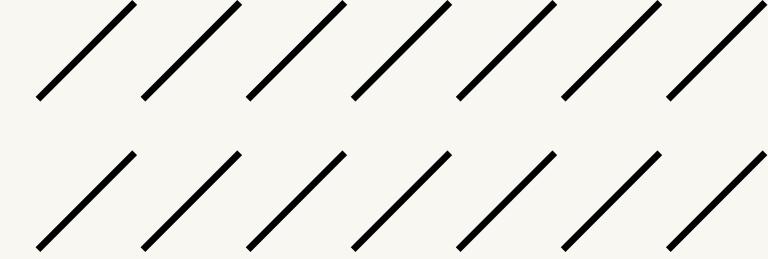
# ANALYZING THE NYT BESTSELLER LIST

APAN 5400 Data Engineering



# *Contents*

01	-----	Introduction
02	-----	Research Questions
03	-----	<i>Neo4j Visuals</i>
04	-----	<i>Feature Analysis</i>
05	-----	<i>Publisher Analysis</i>
06	-----	<i>Sentiment Analysis</i>
07	-----	<i>Seasonal Analysis</i>
08	-----	<i>Lifespan Analysis</i>
09	-----	<i>Co-Author Analysis</i>
010	-----	Appendix



# INTRODUCTION

Every Wednesday, The New York Times publishes its Best Sellers list, an authoritatively ranking of the top-selling books in the United States.

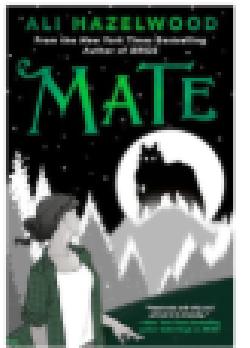
The list is organized by both **format** (i.e., hardcover, paperback, e-book, etc.) and **genre**. These lists serve as one of the most influential indicators of literary popularity and commercial success in the publishing industry. The ranked titles subsequently appear in The New York Times Book Review eleven days later, where a summary and review is printed.

From the enduring appeal of thrillers and romance to the recent rise of celebrity memoirs, the list reflects **weekly consumer preferences** across formats and genres, offering valuable insight into national reading trends.

FICTION | NONFICTION | CHILDREN'S | MONTHLY LISTS

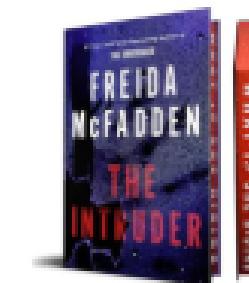
## Combined Print & E-Book Fiction >

1



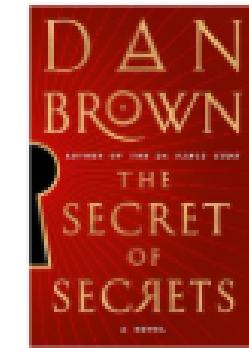
NEW THIS WEEK  
**MATE**  
by Ali Hazelwood

2



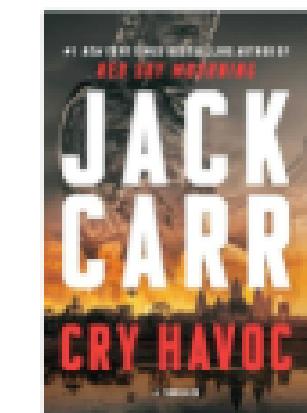
NEW THIS WEEK  
**THE INTRUDER**  
by Freida McFadden

3



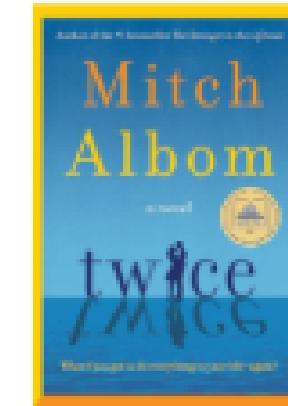
5 WEEKS ON THE LIST  
**THE SECRET OF SECRETS**  
by Dan Brown  
[Read Review](#)

4



NEW THIS WEEK  
**CRY HAVOC**  
by Jack Carr

5



NEW THIS WEEK  
**TWICE**  
by Mitch Albom

2

# RESEARCH QUESTIONS

***By addressing these questions, we aim to generate insights that help publishers and authors understand the key drivers of bestseller success and inform strategic decisions in book selection, marketing, and publication timing.***

01 Which attributes of books have the strongest association with placement on the NYT bestseller list?

02 Which publishers have books that remain at the top of the list versus those that fluctuate often?

03 Are there identifiable seasonal trends or social events that influence list rankings?

04 What is the relationship between a book's description, the overall sentiment, and genre?

05 When many major publishers launch at the same time, does heavy competition on a bestseller's list debut lead to shorter lifespans for books on that list?

06 How do co-authored books perform compared to single-author books on the NYT Best Sellers list?





## Neo4j Visualization

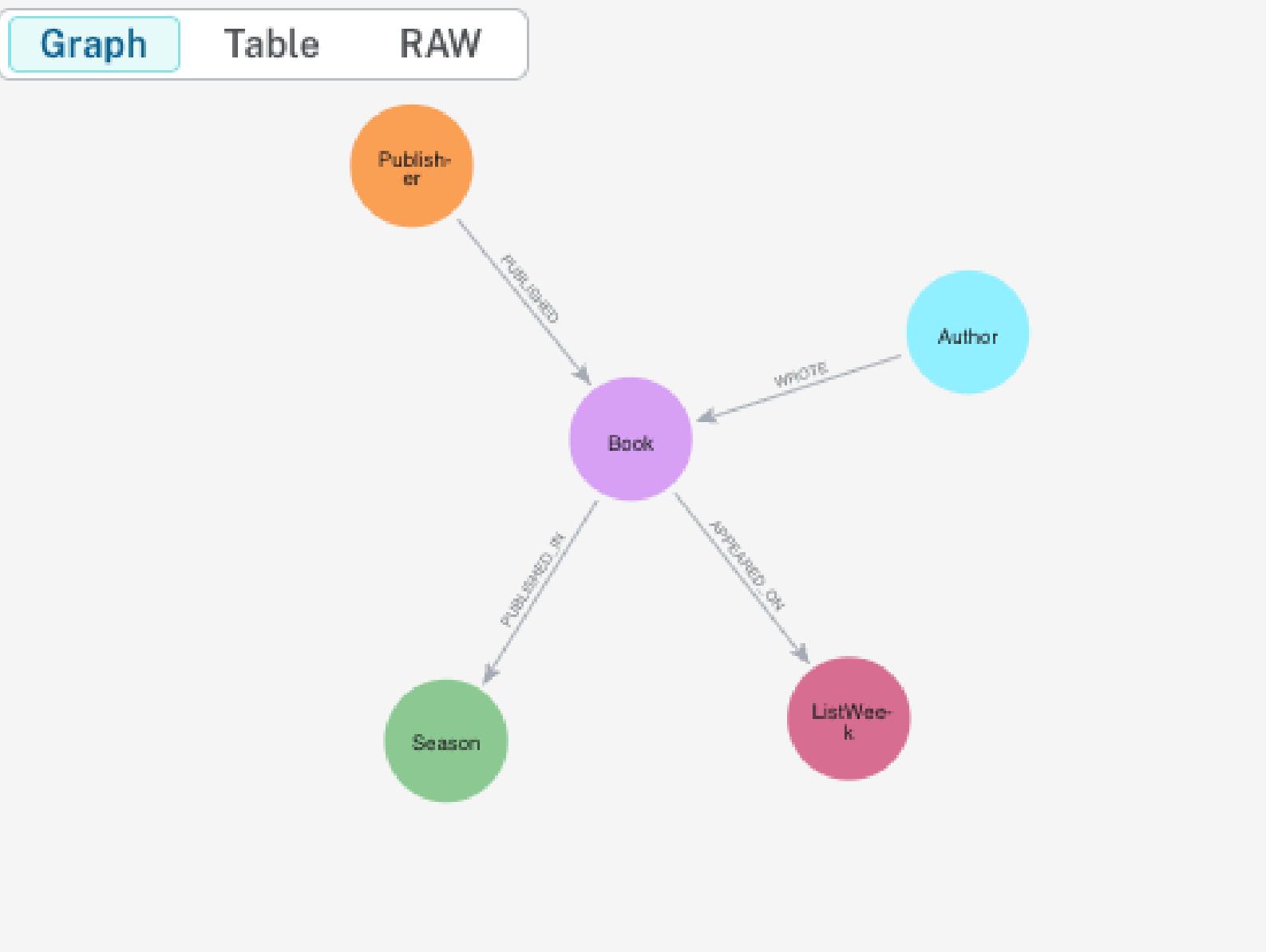
Neo4j

Each Book's Relationship

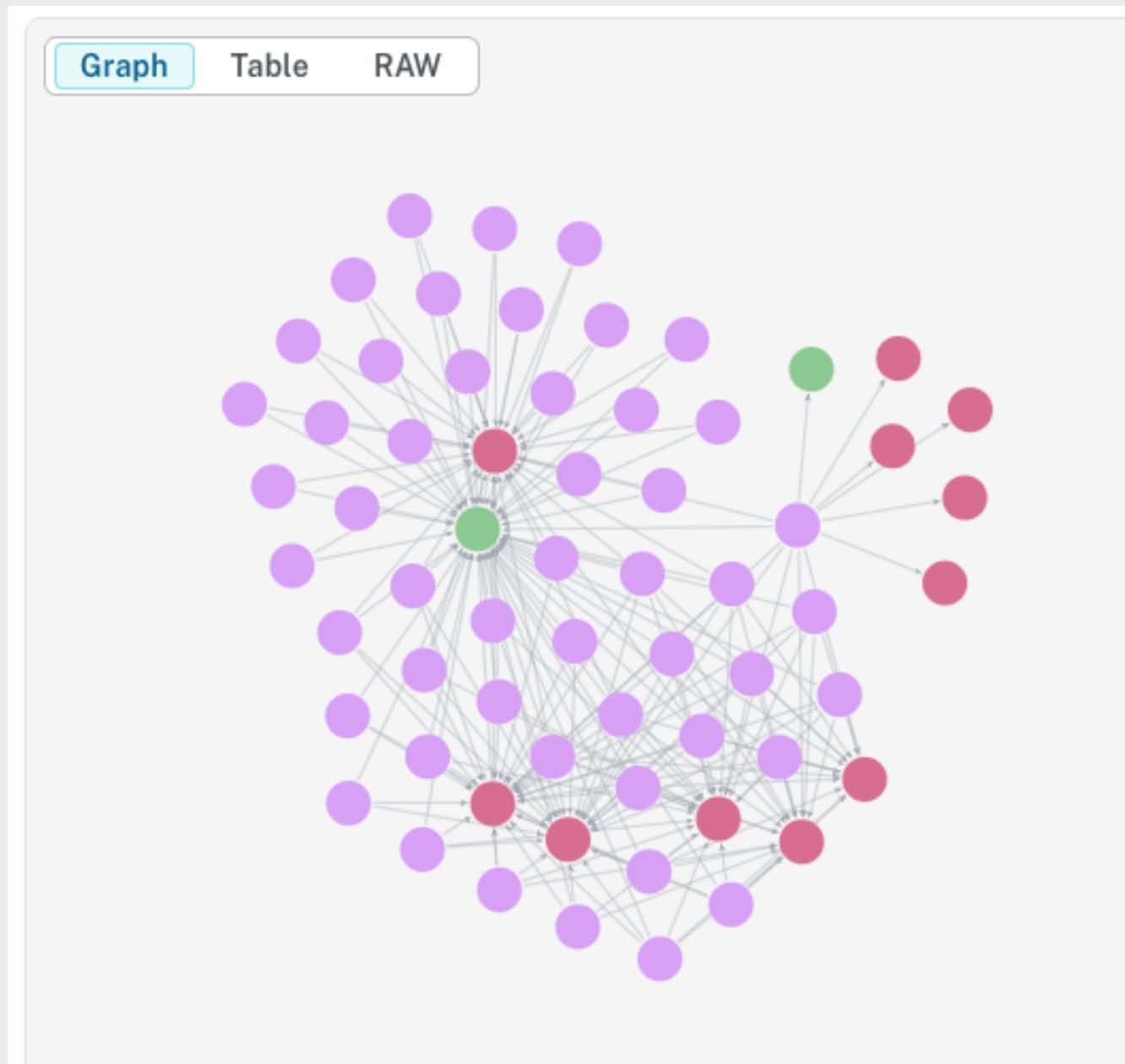
SQL

Mongo DB

Streamlit



Match Books → Season → ListWeek  
Limit 200





## RQ 1 Which attributes of books have the strongest association with placement on the NYT bestseller list?

Q1

Q2

Q3

Q4

Q5

Q6



### Statistically Significant Attributes

Attribute	Coefficient	P> t
rank_last_week	-0.276380	0.000
weeks_on_list	-0.028330	0.000
days_since_publication	-1.222655	0.000
author_popularity	-0.037579	0.000
publisher_strength	-0.004204	0.016
holiday_release	-1.985049	0.000
title_length	0.051204	0.001
description_length	-0.039958	0.000

\*strongest predictors to improve a book's rank

### Interpretation

Only a small subset of attributes significantly influences bestseller rank, suggesting that success depends on additional external factors not captured in the dataset.

External factors may include:

- Marketing efforts
- Media coverage
- Author reputation

### Rolling error

Our rolling error analysis indicates that errors remain low and stable for mid-ranked books (around ranks 10-20), suggesting that the model effectively captures central ranking behavior.

However, prediction error rises for top-ranked books (ranks 1-5), suggesting that true bestseller success depends on additional non-textual drivers not captured in the dataset.

\*In-depth graphs available on Streamlit



## RQ 2: Which publishers have books that remain at the top of the list versus those that fluctuate often?

Q1

Q2

Q3

Q4

Q5

Q6



### Publisher Stability vs. Performance - Bubble Chart

Tracks the trade-off between a publisher's average rank (performance) and consistency (stability)

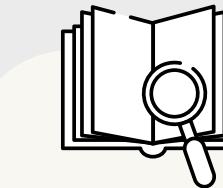
**Random House** and **Penguin** stand out in having consistent high ranking with low standard deviation, indicating stable performance. In contrast, publishers like **Crown** or **Simon & Schuster** show higher variability in rank, indicating their success is less predictable.

**Vintage** and **Grand Central** are moderately stable but don't achieve top ranks as often, suggesting a more *niche presence* rather than blockbuster hits.

### Top Publishers by Ranking Performance Over Time - Scatterplot

Tracks the weekly rank of the 5 publishers with the highest number of overall appearances.

This analysis highlights how certain publishers maintain a steady presence at the top of the bestseller lists, while others experience more fluctuations, reflecting different approaches to publishing and market influence.



### Marketable Insights

#### Publisher Selection Strategy for Authors

Authors seeking predictable, sustained success should prioritize publishers like Random House and Penguin, which demonstrate consistent top-ranking placements.

Authors with potentially breakout books might benefit from publishers like Crown or Simon & Schuster, where high variability suggests they're willing to take bigger risks on potential blockbusters.

#### Investment & Marketing Allocation

- Publishers with low standard deviation (stable performers) likely have:
  - Systematic editorial selection processes
  - Consistent marketing budgets and strategies
  - Established distribution networks
  - Strong brand recognition with readers
- Publishers with high variability may be:
  - More dependent on individual title marketing spends
  - Taking bigger creative risks
  - Relying on author platform/celebrity over publisher brand





## RQ 3: Are there identifiable seasonal trends or social events that influence list rankings?

Q1

Q2 Both seasonality and social events influence the list of ranking in the NYT Bestseller List, since:

Q3

F-statistic: 3.2322

Q4

P-value: 0.0216

Q5

### Takeaways for Publishers:

Q6

1. Time releases to match seasonal strengths.
2. Use summer as a power-staying opportunity.
3. Leverage holiday moments intentionally.
4. Don't underestimate 'regular' weeks.



### Key Insights

#### Most Popular Season: Winter

Winter sees the biggest wave of new titles hitting the NYT list.

#### Books Stay Longest In: Summer

Less books are released in the summer, but the ones that do tend to stay on the list longer.

#### Most Popular Holiday: Father's Day

Titles released around Father's Day stay on the list longer than other holidays.

#### Most Popular Period: Regular Period

Regular periods consistently introduce the highest number of new titles in the NYT list.



Q1

## RQ 4: What is the relationship between a book's description, its overall sentiment, and genre?

Q2

Q3

Q4

Q5

Q6



### Genre Mapping

- Fiction
- Nonfiction



### Sentiment Category

- Polarity
- Subjectivity
- Category



### Genre Sentiment Analysis

- Donut charts
- Ridge Chart
- Bar Chart
- Scatter Plot

#### Polarity

- Positive:  $> 0.1$
- Neutral:  $-0.1 \leq P \leq 0.1$
- Negative:  $< -0.1$

#### Subjectivity

- 0.0 = Objective/factual  
1.0 = Subjective/opinion-based

#### Fiction

- No relationship between book's description, its overall sentiment and genre
- Relatively has a short-lasting on the list
- Sentiment does not predict success
- The publisher needs to check other variables
  - Genre convention
  - Author reputation
  - Story quality

#### Nonfiction

- Has a strong positive relationship between a book's description, overall sentiment and genre
- Books that have solution-oriented and optimistic language in the description stay longer in the NYT list.
- Non-fiction stays longer than fiction



RQ 5 When many major publishers launch at the same time, does heavy competition on a bestseller's list debut lead to shorter lifespans for books on that list?

**Answer:** Yes, heavy competition leads to shorter lifespans and ranking. Having a less competitive entry-point confers a durable competitive edge, improving a book's visibility, driving higher sales, and extending its lifespan on the market.

Outcomes Measured

- **Debut Rank:** rank, is\_debut, competition\_level, list\_name, bestsellers\_date
- **Crowded/Quiet:** rank\_last\_week → is\_debut, grouped by list\_name + bestsellers\_date
- **Longevity:** weeks\_on\_list, competition\_level

Analysis

Competition Level	Average Rank	Average Persistence
Crowded	17.5	18.9
Quiet	9.31 ↑	45.4 ↑





## RQ 6 How do co-authored books perform compared to single-author books on the NYT Best Sellers list?

Q1

**For Publishers:** Learn how to strategically prioritize resource allocation based on a book's predicted market potential & sustained visibility

Q2

**For Authors:** Set realistic career goals & understand the market dynamics of solo vs collaborative work

Q3

**For Marketers:** Determine the optimal campaign duration & intensity based on the predicted market lifecycle & longevity of the book

Q4

### KEY FINDINGS

METRIC	SINGLE AUTHOR	CO-AUTHORED
Longevity	33.09 weeks	11.81 weeks
Initial Success	7.68 / 10	7.66 / 10
Market Volume	2,865 books	245 books

\*Based on 2015-2025 data



*Co-authored books initially secured a \*slightly\* higher rank; however, single-authored books dominated the NYT Best Sellers list overall, appearing 10X more frequently and remaining on the list 3X Longer.*

# Questions?

# APPENDIX A: PROJECT RESOURCES

---

“AAPI Reference.” Streamlit Documentation, Streamlit, “[API Reference.](#)” Streamlit Documentation, Streamlit, “Books API Overview.” NYT Developer Network, The New York Times, API - <https://developer.nytimes.com/docs/books-product/1/overview>

Review of Data: [https://github.com/NYTimes/public\\_api\\_specs/blob/master/books%20api/books%20api.md](https://github.com/NYTimes/public_api_specs/blob/master/books%20api/books%20api.md)

Preliminary Slideshow with brainstorming information - <https://tinyurl.com/3fsb6az2>

Google Drive containing files - [https://drive.google.com/drive/folders/11HuM1EOzLT-Nxkvj08\\_DcR6qFEpG3mxI](https://drive.google.com/drive/folders/11HuM1EOzLT-Nxkvj08_DcR6qFEpG3mxI)

Github repository - [https://github.com/SSankar17/nyt\\_dashboard](https://github.com/SSankar17/nyt_dashboard)

## Project Documentation

The following resources are internal files and repositories created during the project development lifecycle.

- Preliminary Slideshow with Brainstorming Information:
  - Source: Google Slides
  - Link: <https://tinyurl.com/3fsb6az2>
- Google Drive Containing Files: [Project data and working files](#).
- GitHub Repository: [Codebase, scripts, and version control](#).
- Streamlit App: The main application dashboard.

---

# APPENDIX B: PROPOSAL PAGES

---

# DATA SOURCE SPECIFICS

---

The New York Times Best Sellers dataset is sourced from the New York Times Developer API, specifically the Books API. This API provides **structured information on the top-selling books** featured on the New York Times Best Sellers lists. The dataset includes both weekly and monthly rankings across multiple categories such as Fiction, Nonfiction, Advice & Miscellaneous, Children's Books, and more. Data is accessed through **HTTP requests to the NYT Books API endpoints**, and the API responses are returned in JSON format, making it easy to parse and integrate into data workflows.

In order to maintain data integrity, we will be implementing *standardized* data collection and validation procedures by procuring data from **January 2019 to October 2025**.

# PROCUREMENT DETAILS

## 01 Main Performance

Accessing data using the NYT Developer Portal by conducting REST API calls.

## 02 Limitations

API keys creation is limited to one week per call & API rate limits.

## 03 Solutions

1. Store our data locally in a SQL or NoSQL database.
2. Procure data by converting API information to CSV.

Approximate DataFrame size: 1.36 GB

### NYT Best Sellers Mock Data

fact_url	age_group	book_review_link	sunday_review_link	corrections
36734: nyt.com/book	Adult	<a href="https://nyt.com/review">https://nyt.com/review</a>	<a href="https://nyt.com/sunday_review">https://nyt.com/sunday_review</a>	[object Object]
235502: nyt.com/book	Young Adult	<a href="https://nyt.com/review">https://nyt.com/review</a>	<a href="https://nyt.com/sunday_review">https://nyt.com/sunday_review</a>	[object Object]
480840: nyt.com/book	Young Adult	<a href="https://nyt.com/review">https://nyt.com/review</a>	<a href="https://nyt.com/sunday_review">https://nyt.com/sunday_review</a>	[object Object]
716194: nyt.com/book	Children	<a href="https://nyt.com/review">https://nyt.com/review</a>	<a href="https://nyt.com/sunday_review">https://nyt.com/sunday_review</a>	[object Object]
133683: nyt.com/book	Young Adult	<a href="https://nyt.com/review">https://nyt.com/review</a>	<a href="https://nyt.com/sunday_review">https://nyt.com/sunday_review</a>	[object Object]
561258: nyt.com/book	Adult	<a href="https://nyt.com/review">https://nyt.com/review</a>	<a href="https://nyt.com/sunday_review">https://nyt.com/sunday_review</a>	[object Object]
325759: nyt.com/book	Children	<a href="https://nyt.com/review">https://nyt.com/review</a>	<a href="https://nyt.com/sunday_review">https://nyt.com/sunday_review</a>	[object Object]
573999: nyt.com/book	Children	<a href="https://nyt.com/review">https://nyt.com/review</a>	<a href="https://nyt.com/sunday_review">https://nyt.com/sunday_review</a>	[object Object]
449297: nyt.com/book	Adult	<a href="https://nyt.com/review">https://nyt.com/review</a>	<a href="https://nyt.com/sunday_review">https://nyt.com/sunday_review</a>	[object Object]
181607: nyt.com/book	Children	<a href="https://nyt.com/review">https://nyt.com/review</a>	<a href="https://nyt.com/sunday_review">https://nyt.com/sunday_review</a>	[object Object]
934000: nyt.com/book	Children	<a href="https://nyt.com/review">https://nyt.com/review</a>	<a href="https://nyt.com/sunday_review">https://nyt.com/sunday_review</a>	[object Object]

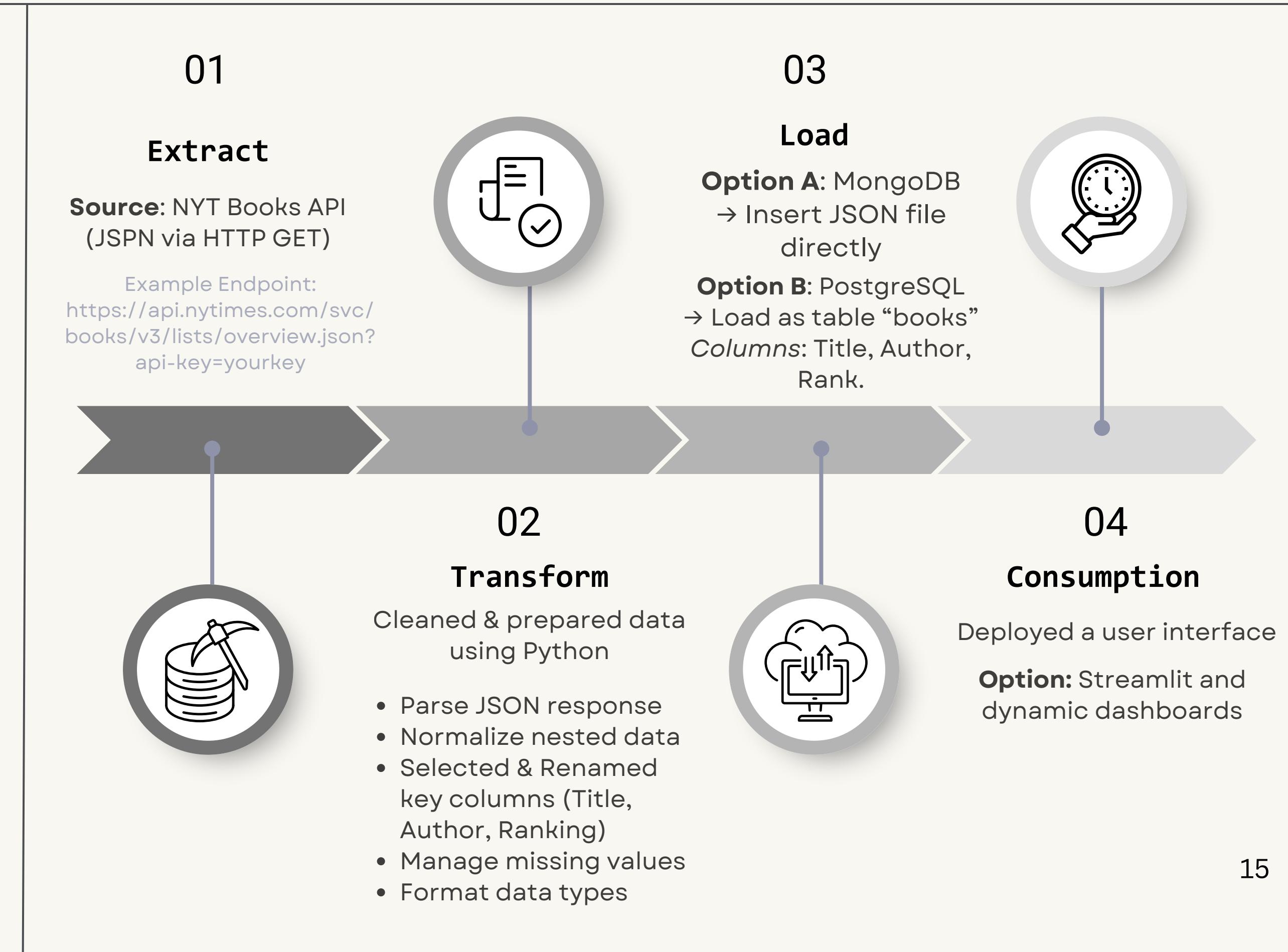
## 04 Data Validation

Checking for duplicates, missing timestamps, inaccurate data types for rows.

## 05 Data Retrieval & Visualization

Use Streamlit: display weekly updates & stored data interactively alongside our graphics and analysis.

# ETL DIAGRAM





Python

SQL

Mongo DB

Streamlit



O1

### WORKFLOW



Data Extraction,  
Transformation,  
Analysis.



O2

### VERSATILITY



Python provides a unified environment for data extraction (via APIs), transformation (ETL pipelines), and analysis.



O3

### EASE OF INTEGRATION

Python seamlessly connects with SQL databases, NoSQL stores (MongoDB), and visualization tools.

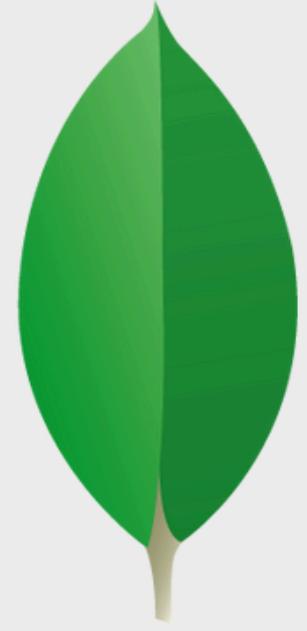


Python

SQL

Mongo DB

Streamlit



# mongoDB



O1

## FLEXIBILITY

Handles nested JSON objects directly from the NYT Books API without needing strict schemas.



O2

## SCALABILITY

Ideal for large, evolving datasets like ours where entries or metadata structures change constantly.



O3

## RAPID DEVELOPMENT

Faster prototyping and ingestion of semi-structured API data.



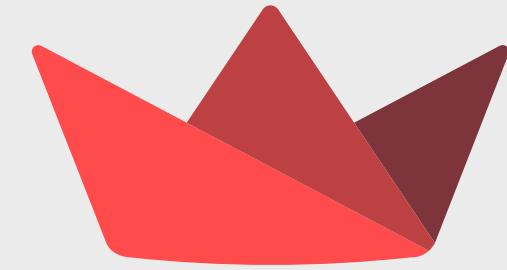


Python

SQL

Mongo DB

Streamlit



# Streamlit



01

## WORKFLOW

Interactive User Interface / Demo.



02

## ACCESSIBILITY

Provides an interactive dashboard for non-technical audiences to explore insights without coding.



03

## INTEGRATION WITH PYTHON

Both Streamlit and Flask run seamlessly with Python analytics script.



# SCALABILITY

---

We are hoping to scale this product in the future with **AWS/Microsoft Azure** and **Tableau/PowerBI** so that larger publishing companies can utilize our dashboards to predict or forecast the performance of upcoming books as well as their ability to achieve success.

Cloud data lakes such as Amazon S3 or Google Cloud Storage provide elastic and low-cost storage for raw JSON files and processed datasets. Structured data can be managed using PostgreSQL which automatically scale with query load. ETL automation can be implemented through AWS Lambda or Google Cloud Functions, allowing serverless batch jobs to periodically extract new best-seller data from the NYT API and update the warehouse without requiring dedicated computing resources. Analytical modeling and visualization can then be deployed on scalable platforms such as Google Colab, Vertex AI, or Tableau Cloud, ensuring accessibility and performance across larger workloads.

The associated costs for small-scale implementation are typically **\$25-\$70** per month and this would cover storage, computation, and visualization tools. As the project grows, these costs vary based on data size and usage. This architecture not only supports future data growth but also enables the seamless integration of additional APIs (e.g., Goodreads) to enhance predictive modeling of author branding and repeat best-seller success

Limitations of scaling would include the API call limitation, cloud storage, data warehousing, and accuracy of the NYT API data. Our most pressing limitation would be storage. Cloud data lakes such as Amazon S3 or Google Cloud Storage provide elastic and low-cost storage for raw JSON files and processed datasets.