POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Systems and Methods for Big and Unstructured Data Project

Author(s): **Tomaso Stefanizzi**

Personal Code: **10713635**

Academic Year: 2023-2024

Contents

1 | Introduction

My idea for this project was to develop a dashboard in the context of an Amazon-style bookshop. The dataset at the core of this project comprises information about books, including details such as category, authors, user ratings, and other relevant attributes. To effectively model the relationships within this dataset, Neo4j, a graph database technology, was selected as the preferred database management system. In fact, the interconnected nature of books, authors, categories, users and reviews, aligns perfectly with the graph structure.

1.1. Data Wrangling/Data Generation

To work with a relatively small dataset, the original data underwent significant filtering through Python code:

- The data was cleaned and filtered to remove null values
- Certain columns were excluded to focus on essential attributes
- Books with multiple authors have been excluded for simplicity
- One column have been added to represent the id of a Review

1.2. Dataset

1.2.1. Original Dataset

The original data is divided into two separated files, one for the Ratings and one for the Books. In the following tables I provide their features.

Ratings

Feature	Description
id	The Id of Book
Title	Book Title
Price	The price of Book
User_id	Id of the user who rates the book
profileName	Name of the user who rates the book
review/helpfulness	helpfulness rating of the review, e.g. 2/3
review/score	rating from 0 to 5 for the book
review/time	time of given the review
review/summary	the summary of a text review
review/text	the full text of a review

Books

Feature	Description
Title	Book Title
description	description of book
authors	Name of book authors
image	url for book cover
previewLink	link to access this book on google Books
publisher	name of the publisher
publishedDate	the date of publish
categories	genres of books
ratingsCount	averaging rating for book

1.2.2. Neo4j

After the Data Wrangling process, I created the different entities and relationships in Neo4j:

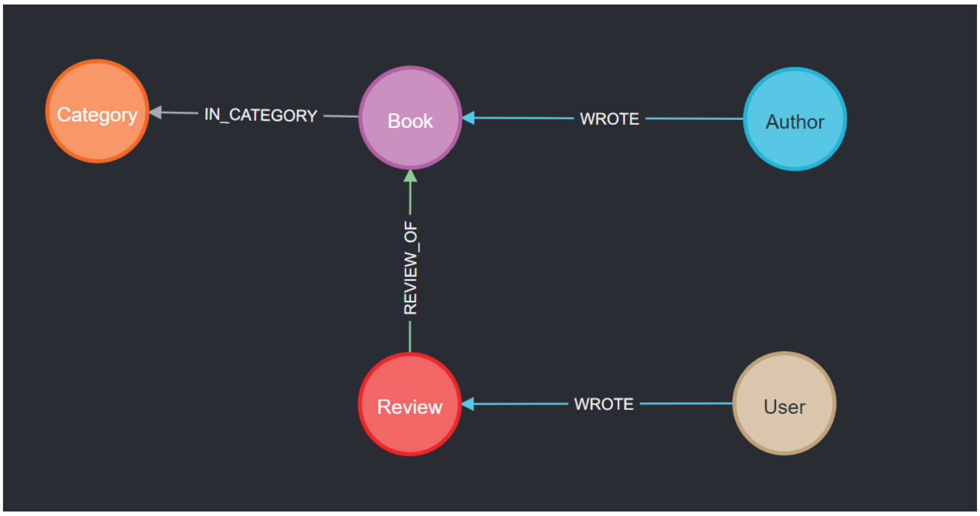


Figure 1.1: Database Schema

Node Type	Properties
Book	title, ratingsCount, publisher, publishedDate, description
Author	name
Category	name
Review	id, time, text, summary, score, price, helpfulness

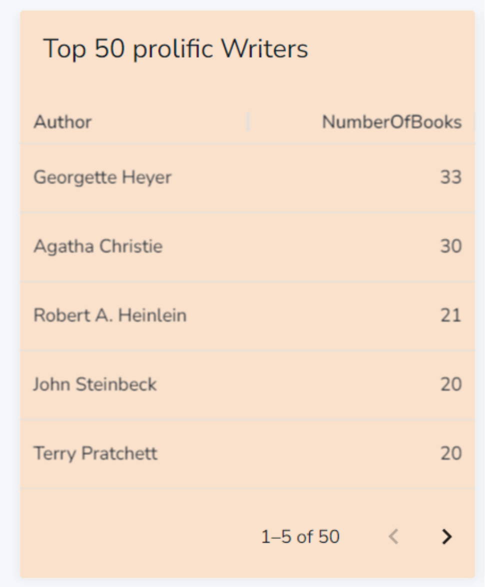
2 | Queries

In this section I list 10 of my queries for this project.

2.0.1. Top 50 prolific Writers

Returns the list of the top 50 Authors with the highest number of written books

```
MATCH (a:Author)-[:WROTE]->(b:Book)
RETURN a.name AS Author, count(b) AS NumberOfBooks
ORDER BY NumberOfBooks DESC LIMIT 50
```



Top 50 prolific Writers

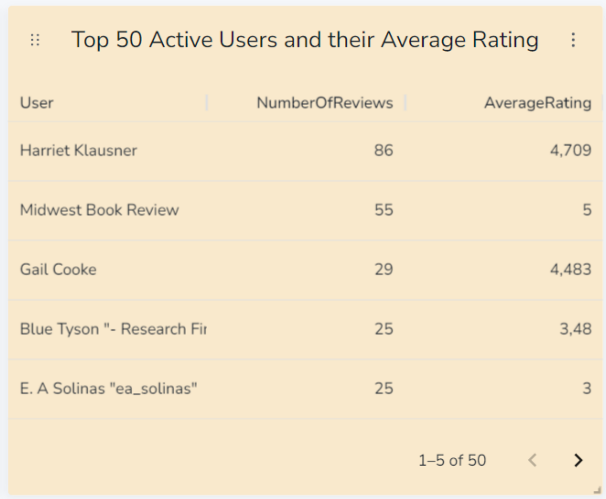
Author	NumberOfBooks
Georgette Heyer	33
Agatha Christie	30
Robert A. Heinlein	21
John Steinbeck	20
Terry Pratchett	20

2.0.2. Top 50 Active Users and their Average Rating

Returns the top 50 Users with the highest number of written reviews

```
MATCH (u:User)-[:WROTE]->(r:Review)
WITH u.id as ID, u.profileName as User, avg(r.score) AS
   AverageRating, count(r) AS NumberOfReviews
RETURN User, NumberOfReviews,AverageRating
```

```
ORDER BY NumberOfReviews DESC LIMIT 50
```



Top 50 Active Users and their Average Rating

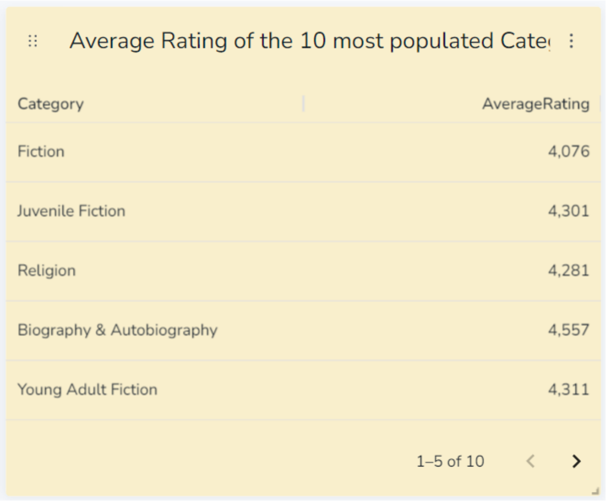
User	NumberOfReviews	AverageRating
Harriet Klausner	86	4,709
Midwest Book Review	55	5
Gail Cooke	29	4,483
Blue Tyson "- Research Fir	25	3,48
E. A Solinas "ea_solinas"	25	3

1–5 of 50

2.0.3. Average Rating of the 10 most populated Categories

Returns the 10 most populated categories in the dataset, with the average rating of all the books of that specific category

```
MATCH (c:Category)<-[:IN_CATEGORY]-(b:Book)
WITH c, count(b) AS NumberOfBooks
ORDER BY NumberOfBooks DESC
LIMIT 10
WITH collect(c.name) AS TopCategories
UNWIND TopCategories AS Category
MATCH (c:Category {name: Category})<-[:IN_CATEGORY]-(b:Book)<-[:
   REVIEW_OF]-(r:Review)
RETURN c.name AS Category, avg(r.score) AS AverageRating
```



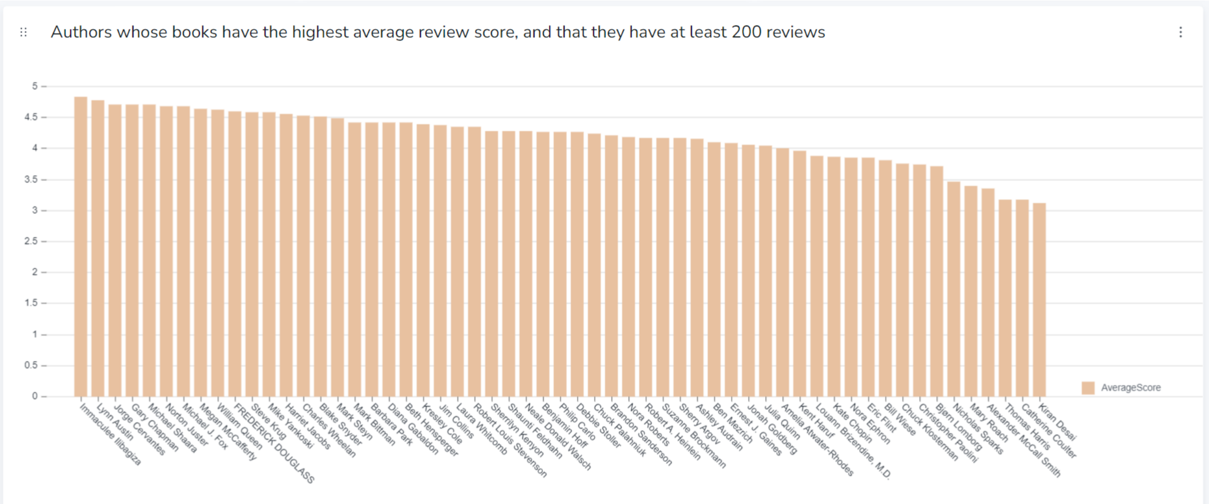
Average Rating of the 10 most populated Cate

Category	AverageRating
Fiction	4,076
Juvenile Fiction	4,301
Religion	4,281
Biography & Autobiography	4,557
Young Adult Fiction	4,311

2.0.4. Authors whose books have the highest average review score

Returns the Authors with the their average ratings. Select only the authors with at least 200 reviews.

```
MATCH (a:Author)-[:WROTE]->(b:Book)<-[:REVIEW_OF]-(r:Review)
WITH a, avg(r.score) AS AverageScore, count(r) AS NumberOfReviews
WHERE NumberOfReviews >= 200
RETURN a.name AS Author, AverageScore
ORDER BY AverageScore DESC
```

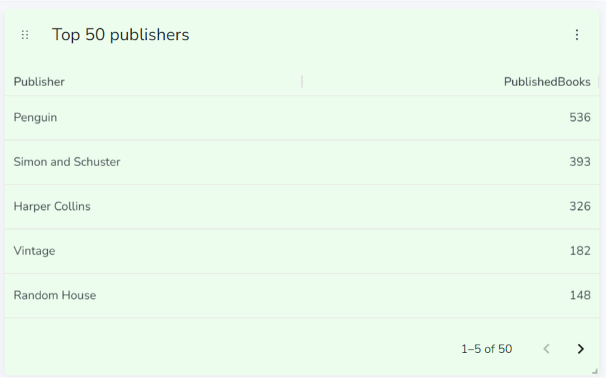


Authors whose books have the highest average review score, and that they have at least 200 reviews

2.0.5. Top 50 publishers

Returns the 50 Publishers that published the highest number of books

```
MATCH (b: Book)
WITH b.publisher as Publisher, count(*) as PublishedBooks
RETURN Publisher, PublishedBooks
ORDER BY PublishedBooks DESC limit 50
```



Top 50 publishers

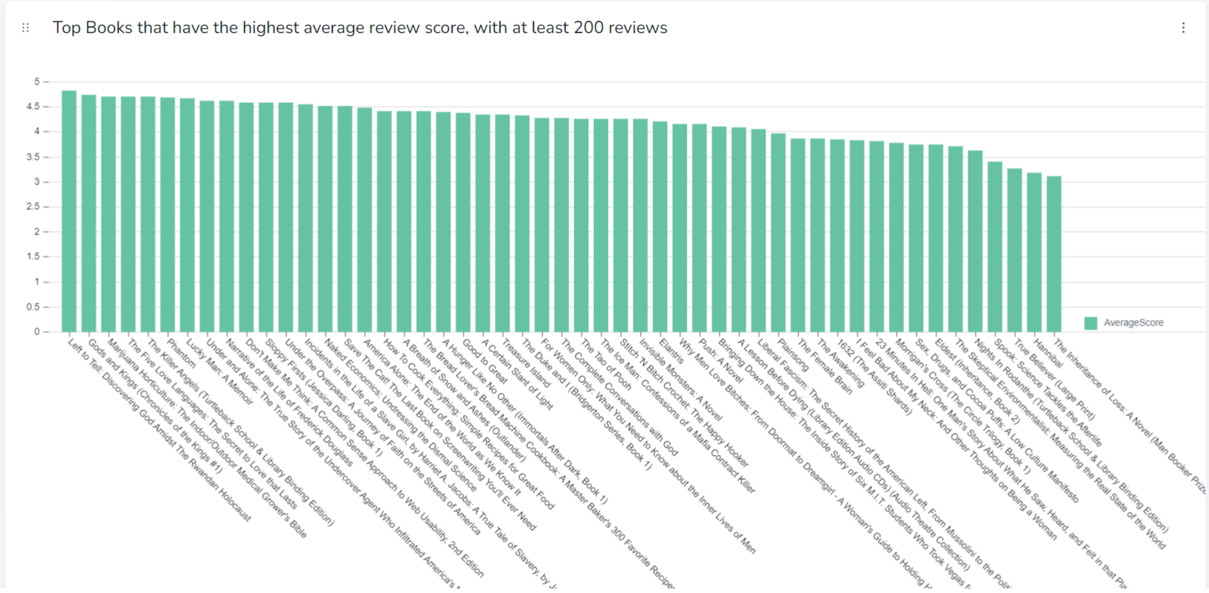
Publisher	PublishedBooks
Penguin	536
Simon and Schuster	393
Harper Collins	326
Vintage	182
Random House	148

1–5 of 50

2.0.6. Books with the highest average review score

Returns the books ordered by their average score. Select only the books with at least 200 reviews.

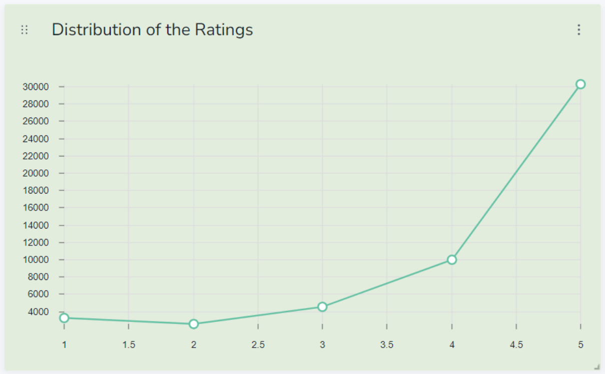
```
MATCH (b:Book)<-[:REVIEW_OF]-(r:Review)
WITH b, avg(r.score) AS AverageScore, count(r) AS NumberOfReviews
WHERE NumberOfReviews >= 200
RETURN b.title AS Book, AverageScore
ORDER BY AverageScore DESC
```



2.0.7. Ratings distribution

Returns each score and its count to plot the rating distribution

```
MATCH (r:Review)
RETURN r.score AS Rating, count(*) AS Count
ORDER BY Rating
```

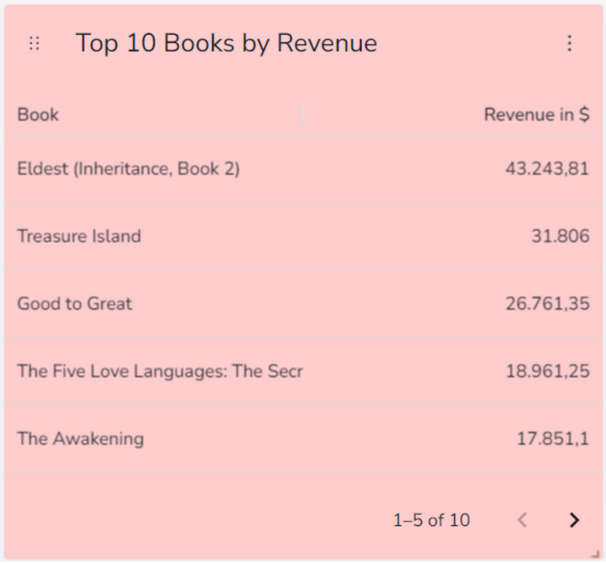


2.0.8. Top 10 Books by Revenue

Returns the top 10 Books ordered by their revenue.

```
MATCH (b:Book)<-[:REVIEW_OF]-(r:Review)
WITH b, sum(r.price) AS Revenue
RETURN b.title as Book, round(Revenue,2) as ‘Revenue in $‘
```

```
ORDER BY Revenue DESC
LIMIT 10
```



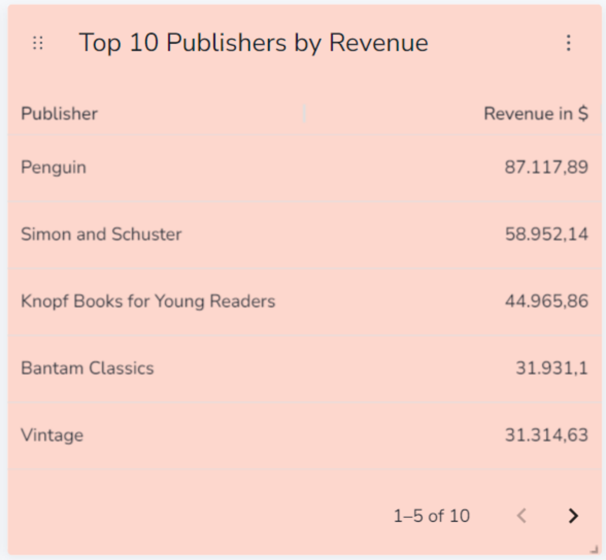
Top 10 Books by Revenue

Book	Revenue in $
Eldest (Inheritance, Book 2)	43.243,81
Treasure Island	31.806
Good to Great	26.761,35
The Five Love Languages: The Secr	18.961,25
The Awakening	17.851,1

2.0.9. Top 10 Publishers by Revenue

Returns the top 10 Publishers ordered by their revenue.

```
MATCH (b:Book)<-[:REVIEW_OF]-(r:Review)
WITH b.publisher AS Publisher, sum(r.price) AS ‘Revenue in $‘
RETURN Publisher, ‘Revenue in $‘
ORDER BY ‘Revenue in $‘ DESC
LIMIT 10
```



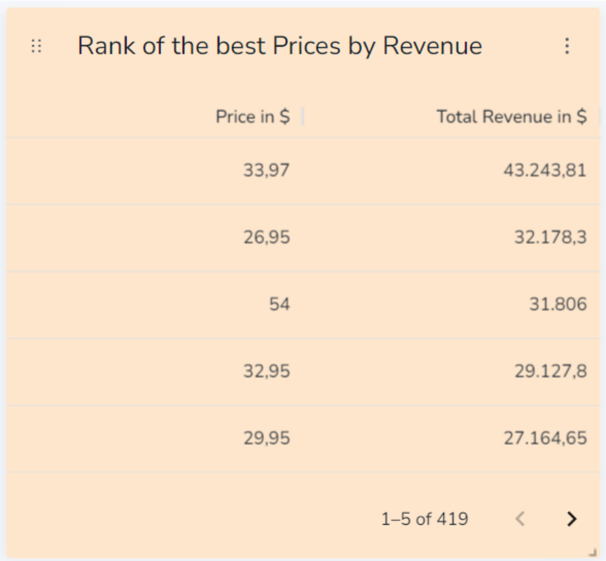
Top 10 Publishers by Revenue

Publisher	Revenue in $
Penguin	87.117,89
Simon and Schuster	58.952,14
Knopf Books for Young Readers	44.965,86
Bantam Classics	31.931,1
Vintage	31.314,63

2.0.10. Price Popularity

Returns the prices that produced more revenue.

```
MATCH (b:Book)<-[:REVIEW_OF]-(r:Review)
WITH b.publisher AS Publisher, sum(r.price) AS ‘Revenue in $‘
RETURN Publisher, ‘Revenue in $‘
ORDER BY ‘Revenue in $‘ DESC
LIMIT 10
```



Rank of the best Prices by Revenue

Price in $	Total Revenue in $
33,97	43.243,81
26,95	32.178,3
54	31.806
32,95	29.127,8
29,95	27.164,65

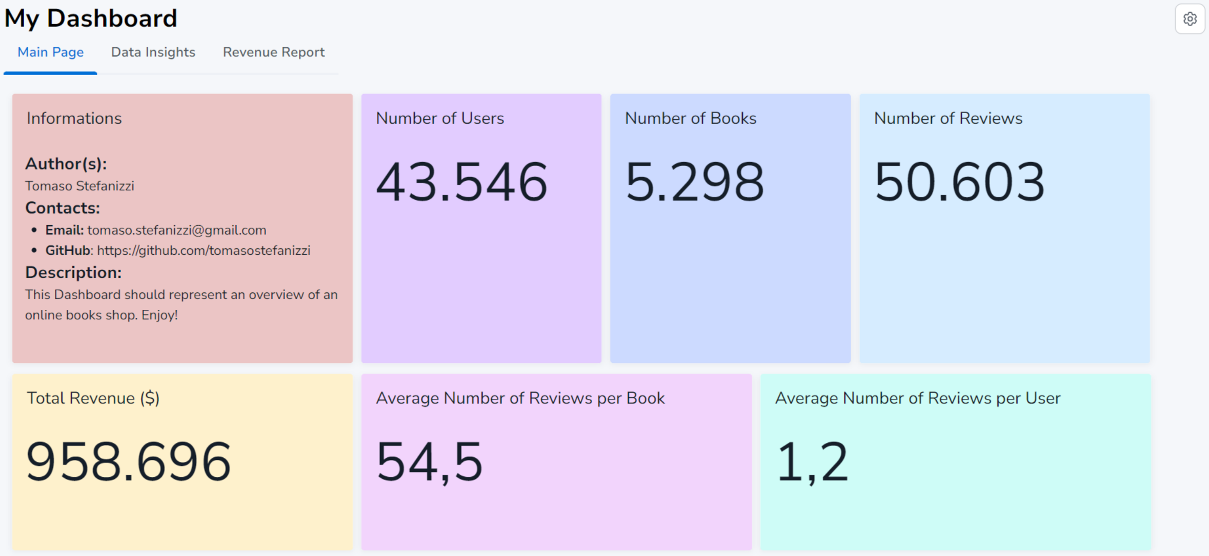
3 | Dashboard

In order to provide an overview of the system, I created a Dashboard with useful data visualizations.

3.1. Structure

I created 3 pages:

- **Main Page**: It contains general informations of the system.
- **Data Insights**: It contains some tables/graphs of the data distribution. On the bottom, I added an interactive plot: for a selected Author, there is a plot of all their books in a graph and in a list on the side.
- **Revenue Report**: It contains some informations related to the revenue. Also here, I provided an interactive plot: for a selected User, we can see his overall purchases (in $), a graph with all his Reviews and also a plot with the price profiling (it's the distribution of all the prices on his purchases).



My Dashboard

Informations

Author(s):
Tomaso Stefanizzi
Contacts:

- **Email:** tomaso.stefanizzi@gmail.com
- **GitHub:** https://github.com/tomasostefanizzi

Description:
This Dashboard should represent an overview of an online books shop. Enjoy!

Number of Users
43.546

Number of Books
5.298

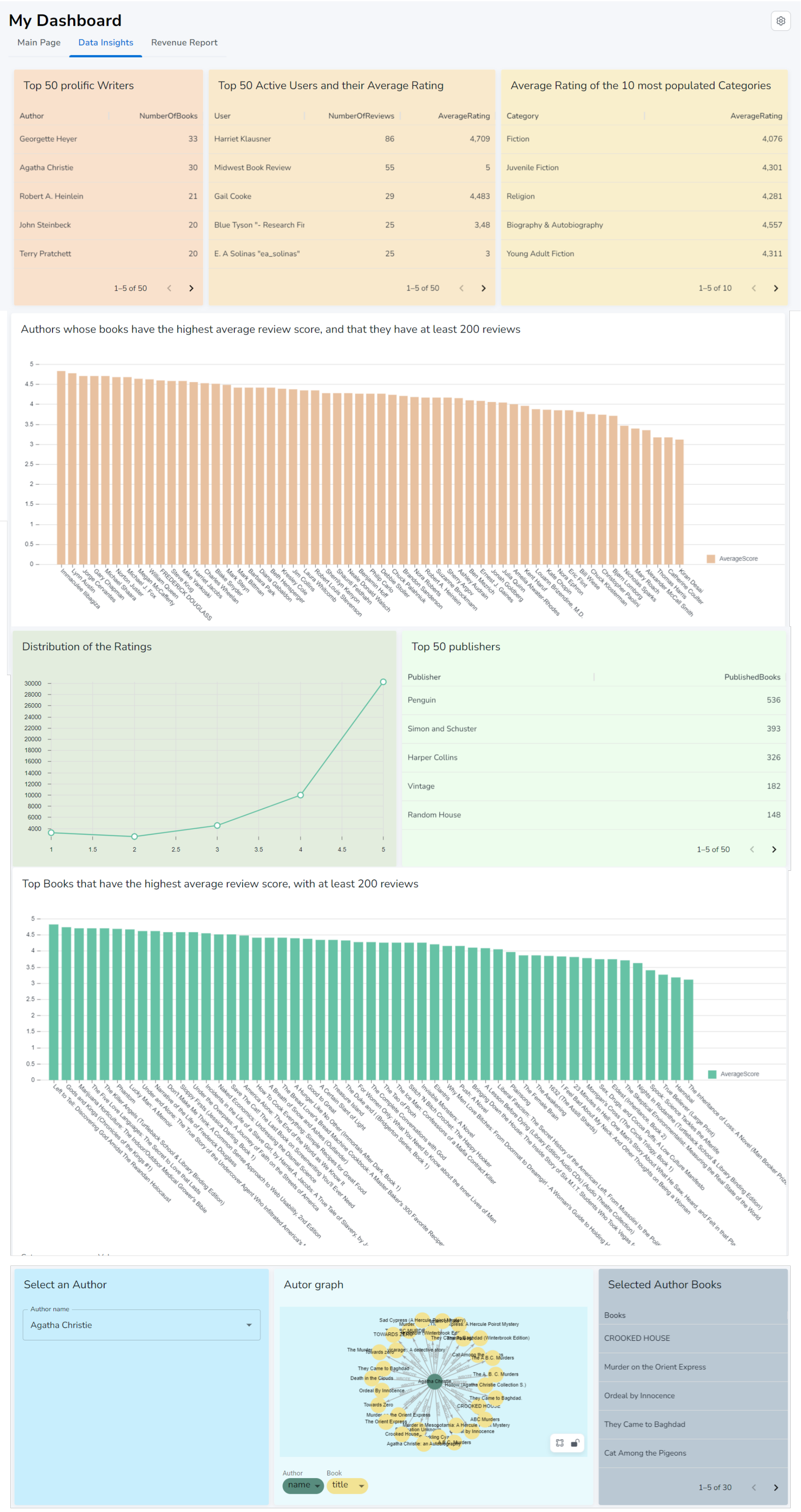
Number of Reviews
50.603

Total Revenue ($)
958.696

Average Number of Reviews per Book
54,5

Average Number of Reviews per User
1,2

Figure 3.1: Main Page



My Dashboard

Main Page | Data Insights | Revenue Report

Top 50 prolific Writers

Author	NumberOfBooks
Georgette Heyer	33
Agatha Christie	30
Robert A. Heinlein	21
John Steinbeck	20
Terry Pratchett	20

1–5 of 50

Top 50 Active Users and their Average Rating

User	NumberOfReviews	AverageRating
Harriet Klausner	86	4,709
Midwest Book Review	55	5
Gail Cooke	29	4,483
Blue Tyson "- Research Fi	25	3,48
E. A Solinas "ea_solinas"	25	3

1–5 of 50

Average Rating of the 10 most populated Categories

Category	AverageRating
Fiction	4,076
Juvenile Fiction	4,301
Religion	4,281
Biography & Autobiography	4,557
Young Adult Fiction	4,311

1–5 of 10

Authors whose books have the highest average review score, and that they have at least 200 reviews

Distribution of the Ratings

Top 50 publishers

Publisher	PublishedBooks
Penguin	536
Simon and Schuster	393
Harper Collins	326
Vintage	182
Random House	148

1–5 of 50

Top Books that have the highest average review score, with at least 200 reviews

5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
AverageScore
Left to Tell: Discovering God Amidst The Rwandan Holocaust
Gods and Kings (Chronicles of the Kings #1)
Marijuana Horticulture: The Indoor/Outdoor Medical Grower's Bible
The Five Love Languages: The Secret to Love that Lasts
The Killer Angels (Turtleback School & Library Binding Edition)
Phantom
Lucky Man: A Memoir
Under and Alone: The True Story of the Undercover Agent Who Infiltrated America's
Narrative of the Life of Frederick Douglass
Don't Make Me Think: A Common Sense Approach to Web Usability, 2nd Edition
Sloppy Firsts (Jessica Darling, Book 1)
Under the Overpass: A Journey of Faith on the Streets of America
Incidents in the Life of a Slave Girl, by Harriet A. Jacobs: A True Tale of Slavery, by J
Naked Economics: Undressing the Dismal Science
Save The Cat! The Last Book on Screenwriting You'll Ever Need
America Alone: The End of the World as We Know It
How To Cook Everything: Simple Recipes for Great Food
A Breath of Snow and Ashes (Outlander)
The Bread Lover's Bread Machine Cookbook: A Master Baker's 300 Favorite Recipe
A Hunger Like No Other (Immortals After Dark, Book 1)
Good to Great
A Certain Slant of Light
Treasure Island
The Duke and I (Bridgerton Series, Book 1)
For Women Only: What You Need to Know about the Inner Lives of Men
The Complete Conversations with God
The Tao of Pooh
The Ice Man: Confessions of a Mafia Contract Killer
Stitch 'N Bitch Crochet: The Happy Hooker
Invisible Monsters: A Novel
Elantris
Why Men Love Bitches: From Doormat to Dreamgirl - A Woman's Guide to Holding H
Push: A Novel
Bringing Down the House: The Inside Story of Six M.I.T. Students Who Took Vegas f
A Lesson Before Dying (Library Edition Audio CDs) (Audio Theatre Collection)
Liberal Fascism: The Secret History of the American Left, From Mussolini to the Polit
Plainsong
The Female Brain
The Awakening
1632 (The Assiti Shards)
I Feel Bad About My Neck: And Other Thoughts on Being a Woman
23 Minutes In Hell: One Man's Story About What He Saw, Heard, and Felt in that Pla
Morrigan's Cross (The Circle Trilogy, Book 1)
Sex, Drugs, and Cocoa Puffs: A Low Culture Manifesto
Eldest (Inheritance, Book 2)
The Skeptical Environmentalist: Measuring the Real State of the World
Nights In Rodanthe (Turtleback School & Library Binding Edition)
Spook: Science Tackles the Afterlife
True Believer (Large Print)
Hannibal
The Inheritance of Loss: A Novel (Man Booker Prize

Select an Author

Author name: Agatha Christie

Autor graph

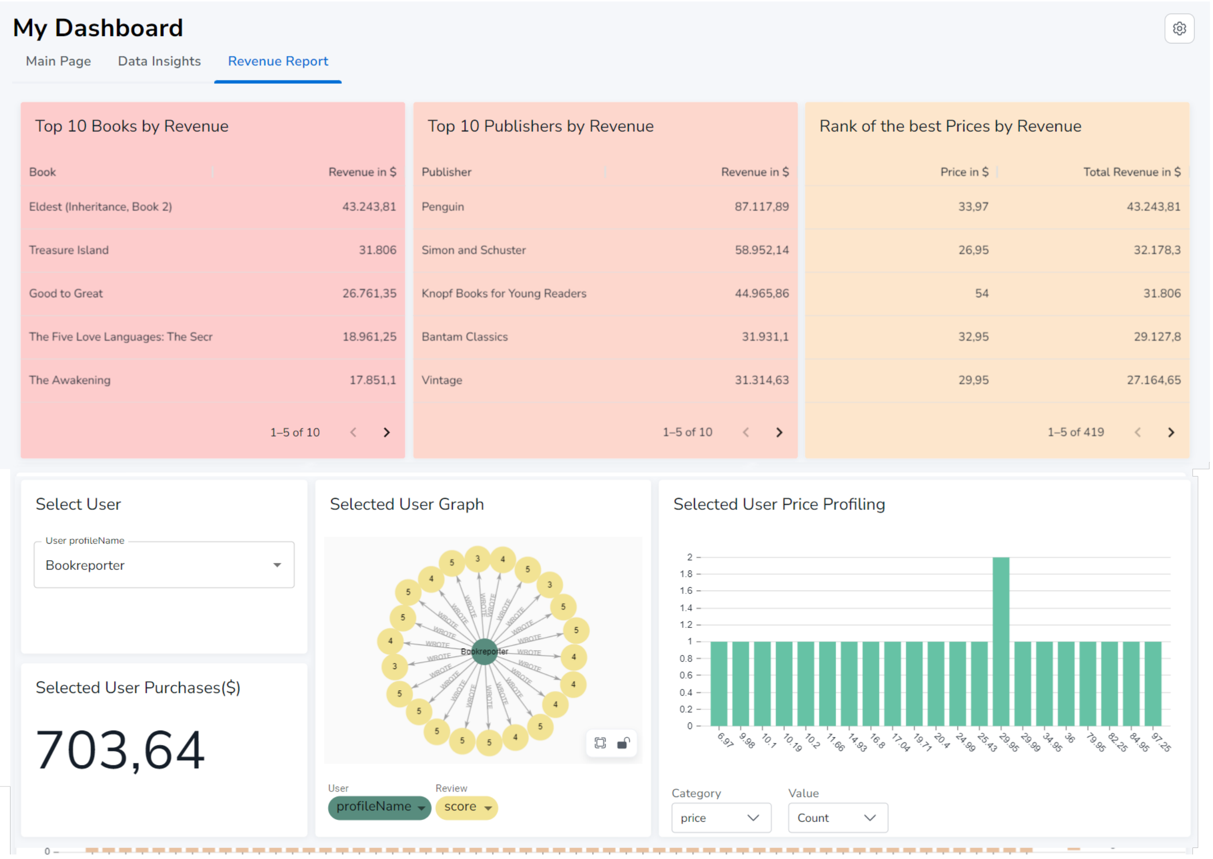
Agatha Christie
Sad Cypress (A Hercule Poirot Mystery)
Murder on the Orient Express: A Hercule Poirot Mystery
TOWARDS ZERO
The Hollow (Winterbrook Edition)
They Came to Baghdad (Winterbrook Edition)
The Murder at the Vicarage: A detective story
Towards zero
Cat Among the Pigeons
The A.B.C. Murders
They Came to Baghdad
Death in the Clouds
The A. B. C. Murders
Hollow (Agatha Christie Collection S.)
Ordeal By Innocence
They Came to Baghdad.
Towards Zero
CROOKED HOUSE
Murder on the Orient Express
The Orient Express
ABC Murders
Murder in Mesopotamia: A Hercule Poirot Mystery
Ordeal by Innocence
Crooked House
Agatha Christie: an Autobiography
A.B.C. Murders
WROTE
Author
name
Book
title

Selected Author Books

Books
CROOKED HOUSE
Murder on the Orient Express
Ordeal by Innocence
They Came to Baghdad
Cat Among the Pigeons

1–5 of 30

Figure 3.2: Data Insights



My Dashboard

Main Page | Data Insights | Revenue Report

Top 10 Books by Revenue

Book	Revenue in $
Eldest (Inheritance, Book 2)	43.243,81
Treasure Island	31.806
Good to Great	26.761,35
The Five Love Languages: The Secr	18.961,25
The Awakening	17.851,1

1–5 of 10

Top 10 Publishers by Revenue

Publisher	Revenue in $
Penguin	87.117,89
Simon and Schuster	58.952,14
Knopf Books for Young Readers	44.965,86
Bantam Classics	31.931,1
Vintage	31.314,63

1–5 of 10

Rank of the best Prices by Revenue

Price in $	Total Revenue in $
33,97	43.243,81
26,95	32.178,3
54	31.806
32,95	29.127,8
29,95	27.164,65

1–5 of 419

Select User

User profileName: Bookreporter

Selected User Purchases($)

703,64

Selected User Graph

Selected User Price Profiling

Category: price

Value: Count

Figure 3.3: Revenue Report