BUSINESS INTELLIGENCE &
BIG DATA ANALYTICS PROJECT

Book Dataset Analysis

By team 13
Department of Management Science & Technology
Athens University of Business & Economics

Table of Contents

Book Dataset :

3 files

book.csv - >271,379

users.csv - >278,859

ratings.csv - >1,149,780

Source:
Kaggle.com

License:
Public Domain Dedication

Book Dataset :

3 files

book.csv -
>271,379

users.csv -
>278,859

ratings.csv -
>1,149,780

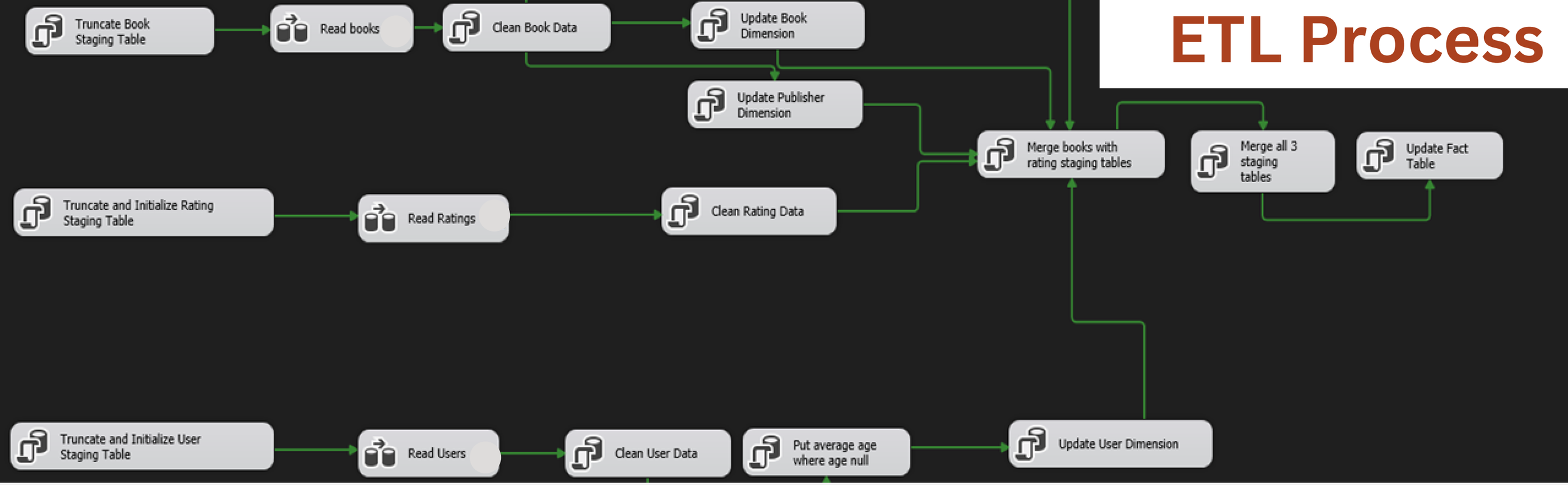
Columns

ISBN, Book Title, Book Author,
Publisher, Year of Publication

Reader id, Location, Age

ISBN, Reader id, Rating
Rating = 0 --> sale (implicit)
Rating in [1,10] --> score (explicit)

ETL Process



Truncate staging tables	Load CSV files into sql staging tables	Data cleaning	Create and Update Dimension Tables	Create and Update Fact Table
✓	✓	✓	✓	✓

Data Cleaning Books

ISBN	Accurate ISBNs consiist of eithe 10 or 13 characters. Must be unique.
Year of Publication	Should contain only numeric characters with length equal to 4
Publisher , Author	Should not contain null values

Data Cleaning
Users

Age	Keep ages between 5 and 90 and set other null Where Age is null put the average Age Round Age to the nearest 5
Locstion	Keep last part of location that is country Delete values with low frequency Choose the most popular name for counties who appear in different ways If state put it as USA

Star Schema

Fact Table:

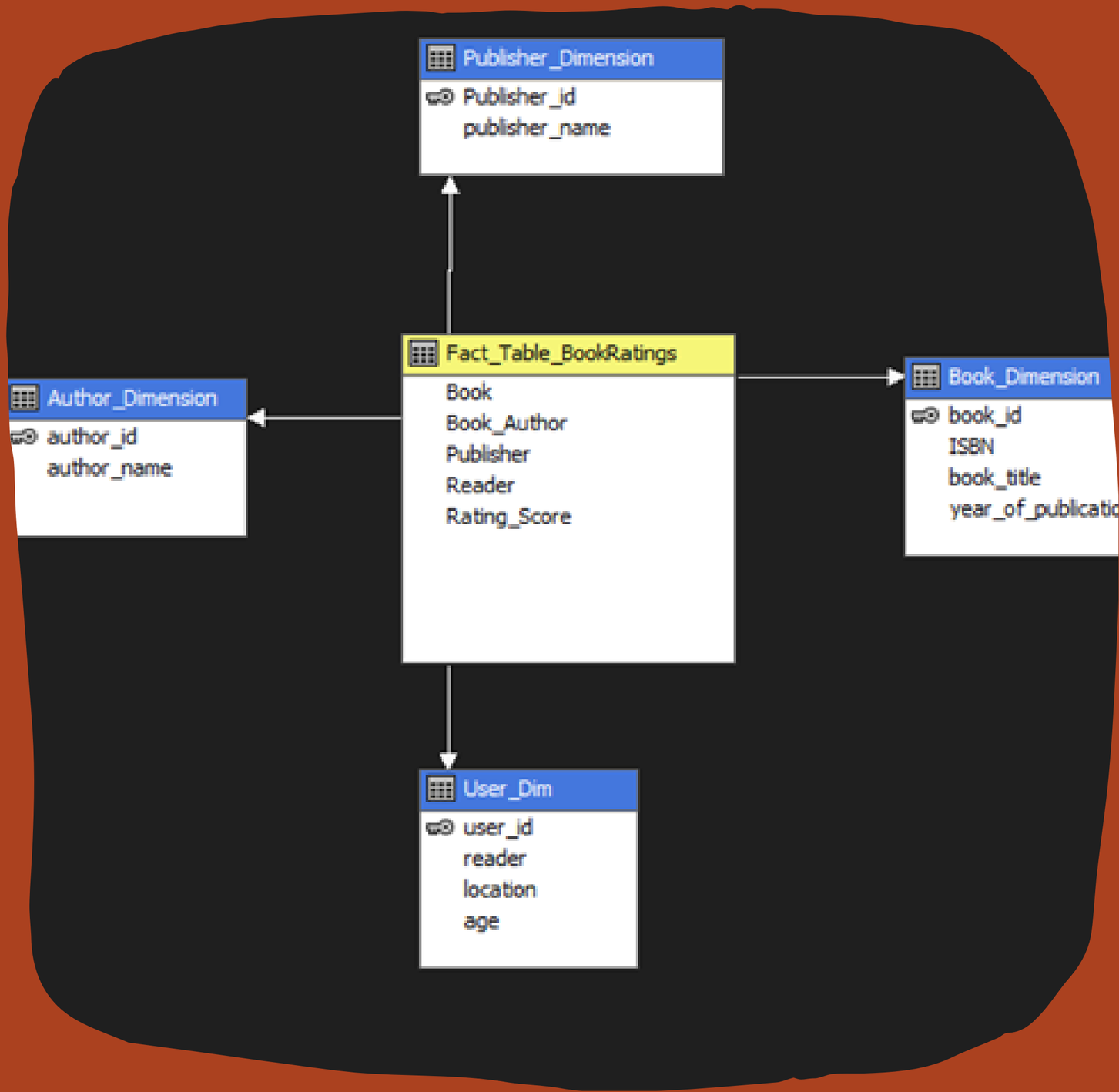
Contains all the Book Ratings (732,242)

Dimensions:

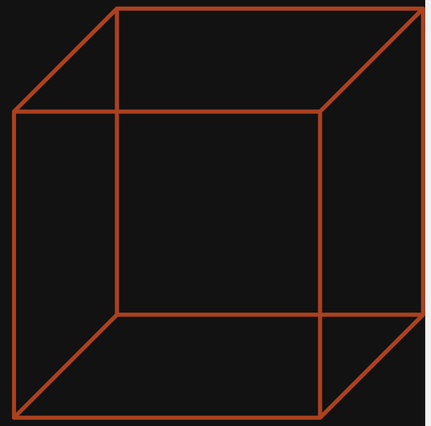
- Book (book id, isbn, book_title,year of publication
- Book_Author (author_id,author_name)
- Publisher (publisher_id, publisher_name)
- Reader (user_id, reader, location, age)

Measure:

Rating Score in range [0-10]



Cube building and deployment



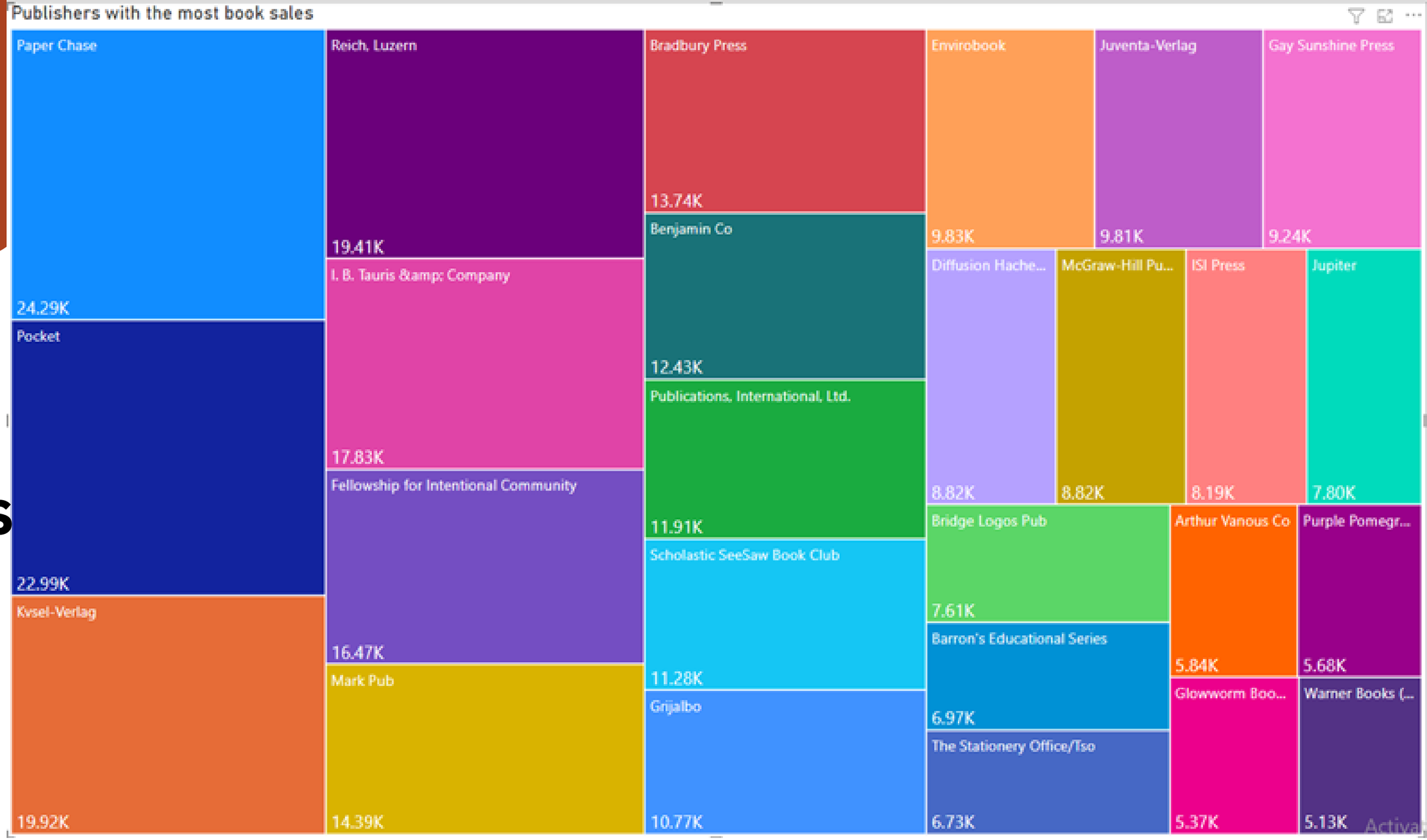
Initial Measures:
- Rating Score

Extra Measures:
-Count of Ratings
-Average of Ratings

Steps to create Cube	
✓	Create a MMSQL Analysisis Services Project
✓	Select as Data Source the Data Warehouse we created
✓	Create Data Source View by loading the Fact Table and the related to it dimensions tables.
✓	Create cube with Fact Table as the measure group and the dimension tables as dimensions
✓	Add all the attributes of every dimension and process

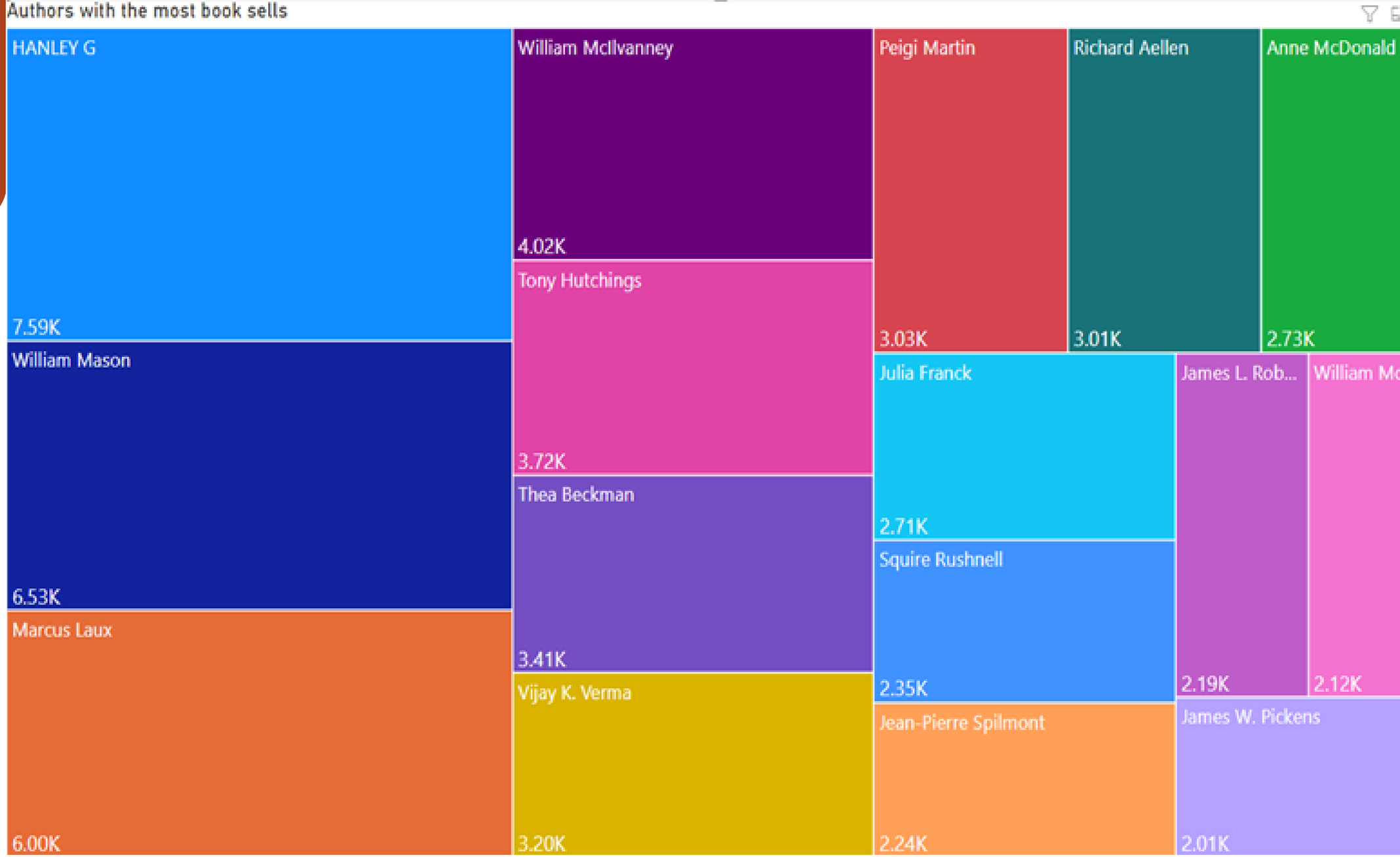
Power BI Dashboards

Best selling publishers



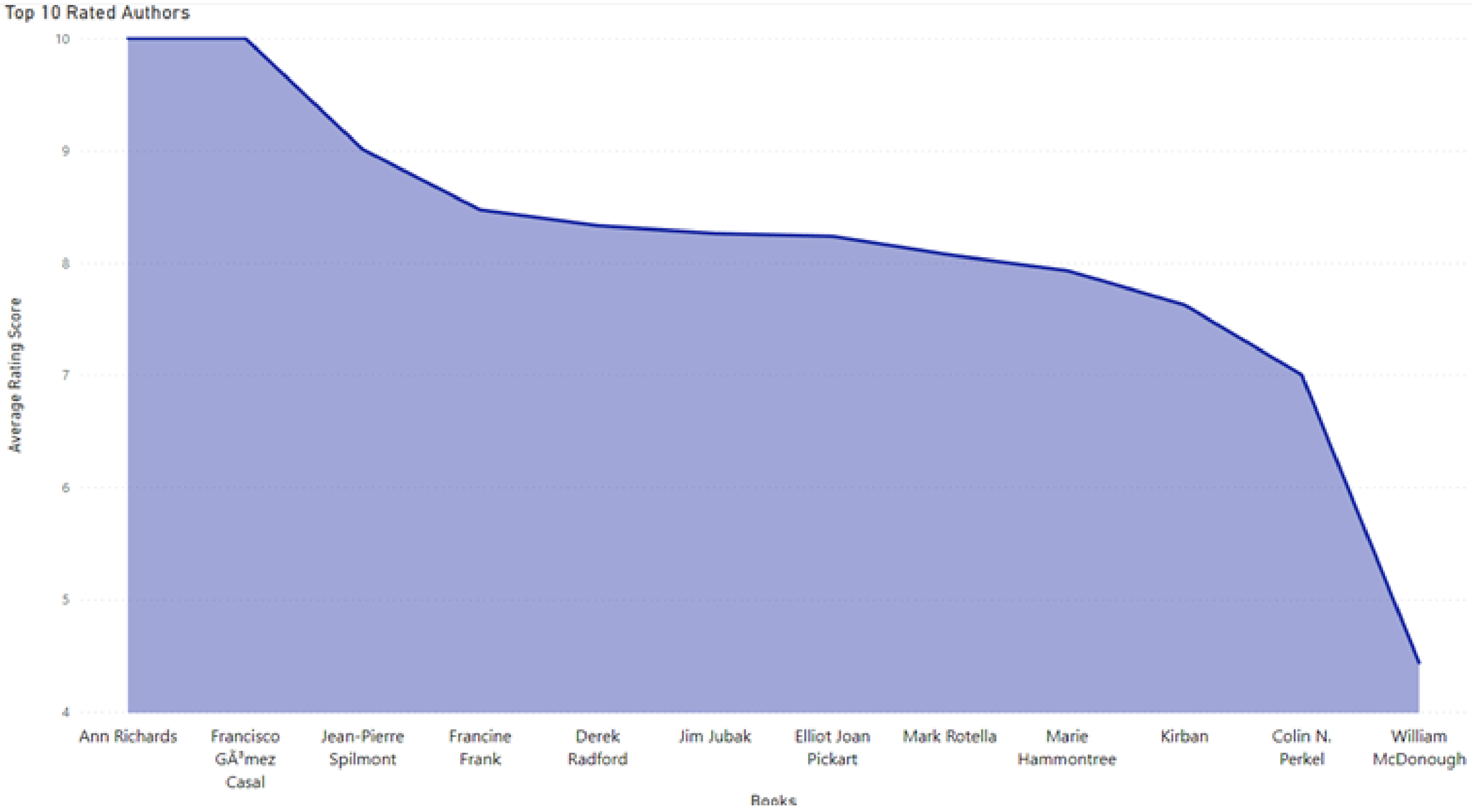
Power BI Dashboards

Best selling authors



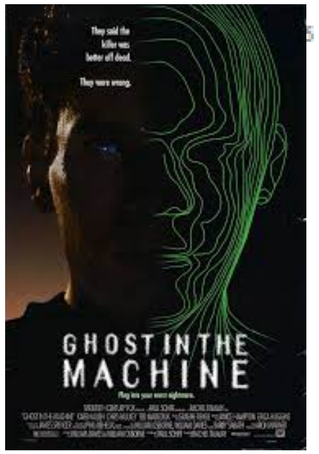
Power BI Dashboards

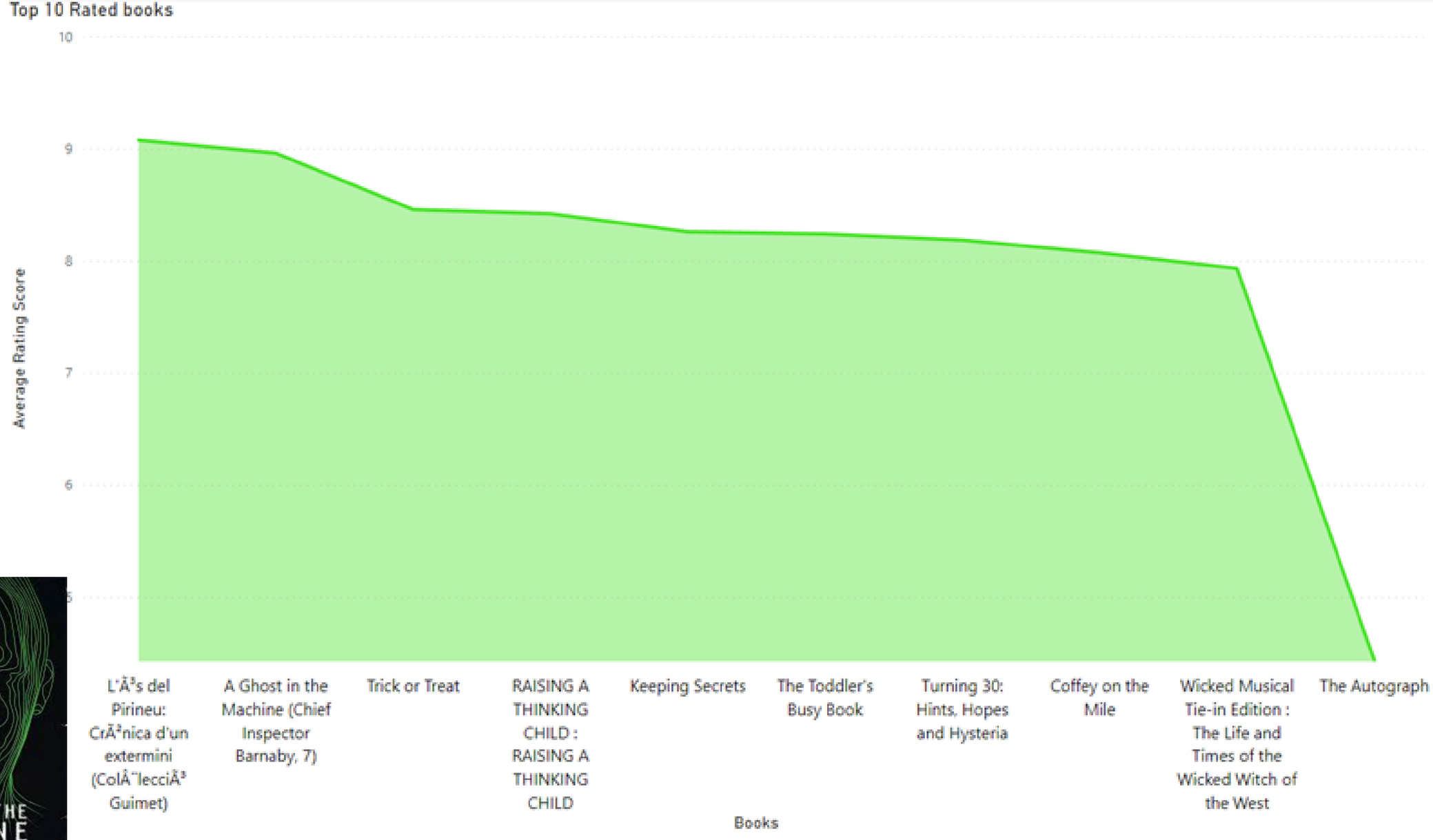
Top 10 rated authors



Power BI Dashboards

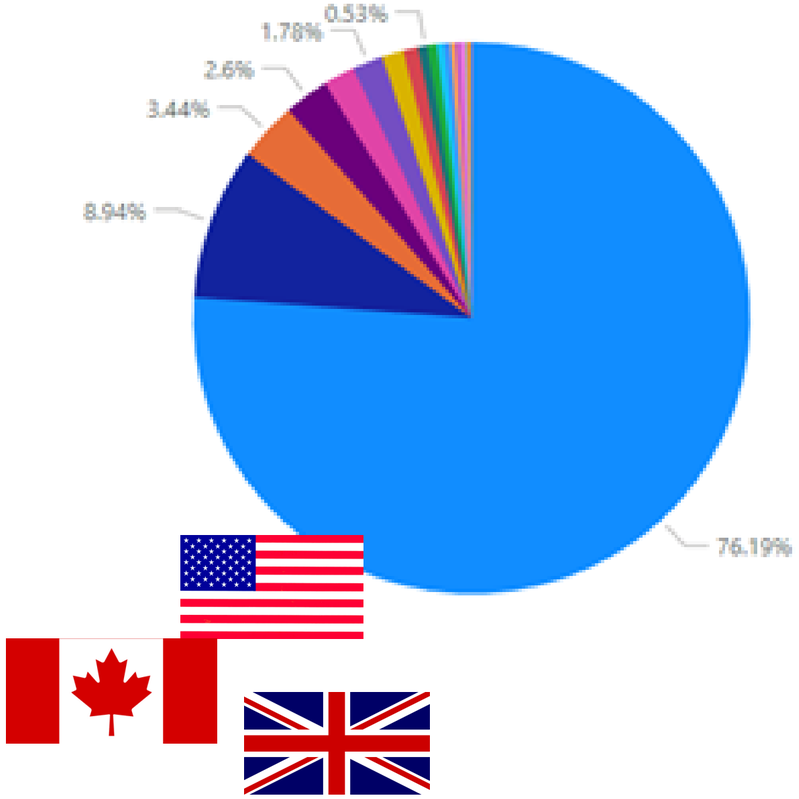
Top 10 rated authors





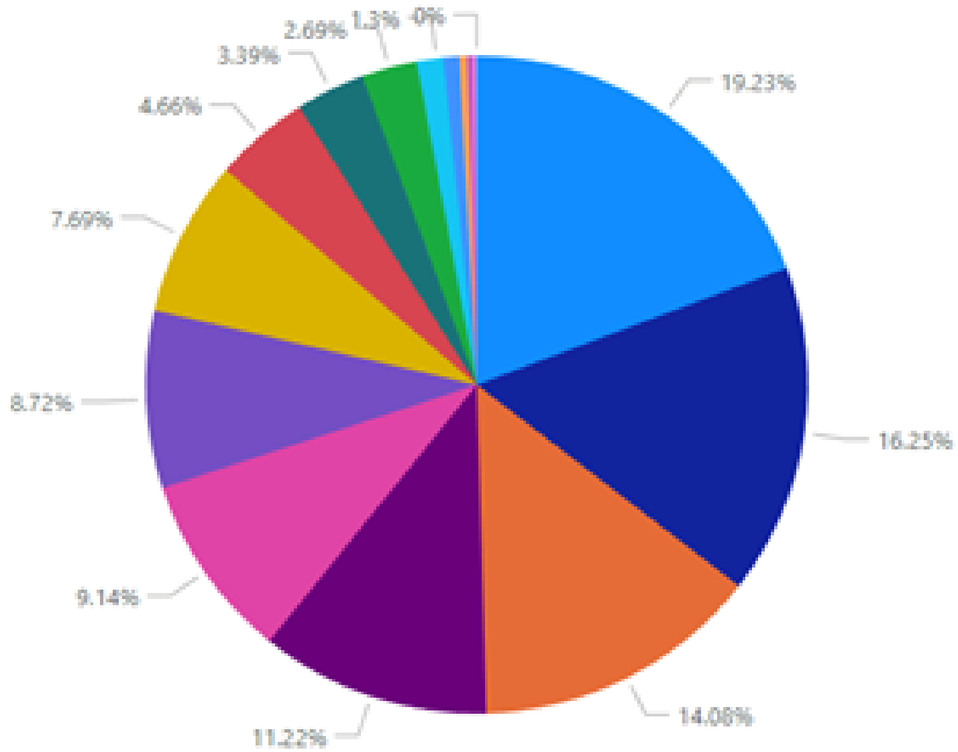
Power BI Dashboards

Books bought by county



Location
- usa
- canada
- united kingdom
- germany
- australia
- spain
- france
- portugal
- malaysia
- netherlands
- switzerland
- new zealand
- austria
- italy
- iran
- finland
- romania
- szech republic

Books bought by Age



Age
- 30
- 25
- 35
- 40
- 45
- 20
- 50
- 55
- 15
- 60
- 65
- 10
- 5
- 70
- 75
- 80
- 85
- 90

Association Rules Analysis

Information on the combinations of books purchased

- Apriori Algorithm
  1. Support
  2. Confidence
  3. Lift

Harry Potter and the Goblet of Fire(4)
Harry Potter and the Sorcerer's Stone(1)
Harry Potter and the Prisoner of Azkaban(3)

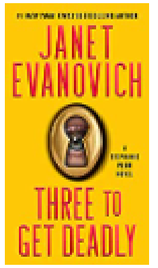
Harry Potter and the Chamber of Secrets (2)

Association Rules Analysis

Information on successful/appealing combinations of books purchased

- Apriori Algorithm
- Purpose the suggestion of likable books

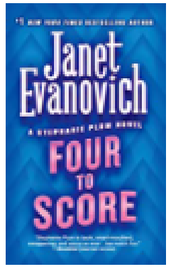
3



Three to Get Deadly (Stephanie Plum, No. 3): A Stephanie Plum Novel

› Janet Evanovich

4



Four to Score (Stephanie Plum, No. 4): A Stephanie Plum Novel

› Janet Evanovich

1.Three To Get Deadly : A Stephanie Plum Novel
2.Hot Six : A Stephanie Plum Novel
3.One for the Money : A Stephanie Plum Novel

1. Four To Score (A Stephanie Plum Novel)
2. High Five (A Stephanie Plum Novel)

Association Rules Analysis

Achieve cross sellings!

How can a big bookstore use this?	
✓	Combined offers
✓	Offers for series books
✓	Personalized recommendation pop - ups
✓	Re-orderng the placement of the books in case of physical store

Book Classification

1
Widely liked

Books with Average Rating > 6
&
Rating Variance < 1.67

2
Debateable

Books with
Rating Variance > 1.67

3
Widely disliked

Books with Average Rating < 6
&
Rating Variance < 1.67

Book Classification

1
Widely liked

Mass advertisement for classical books, no need for personalization.

Safe to renew these books supplies

2
Debateable

We need to find the correct audience for these books.

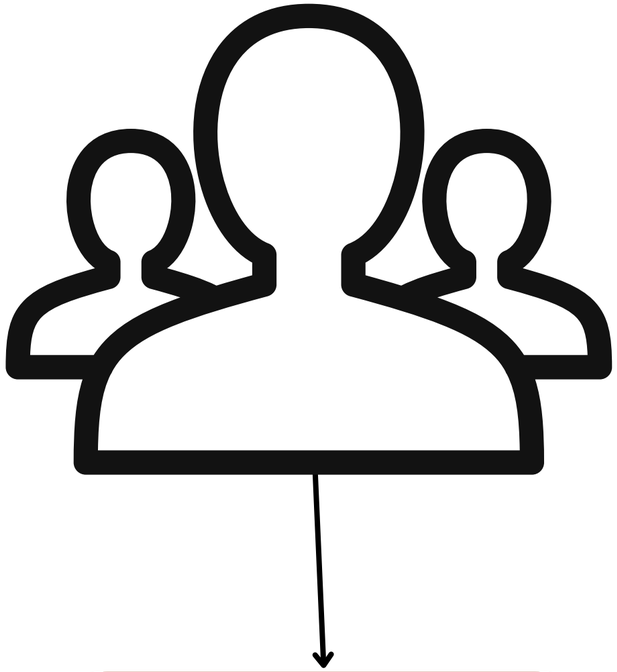
Personalized > Mass advertisement

3
Widely disliked

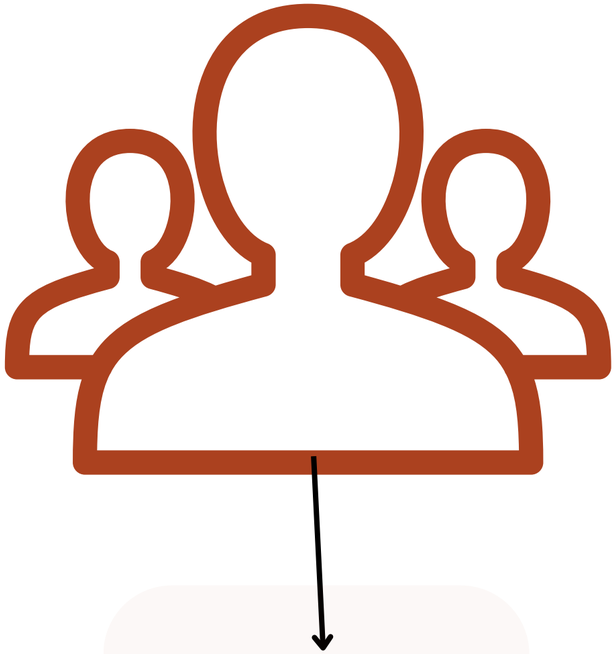
Not profitable to keep selling these books.
Low demand
Bad Ratings fot book --> Bad ratings for the book store.

Customer Segmentation Analysis

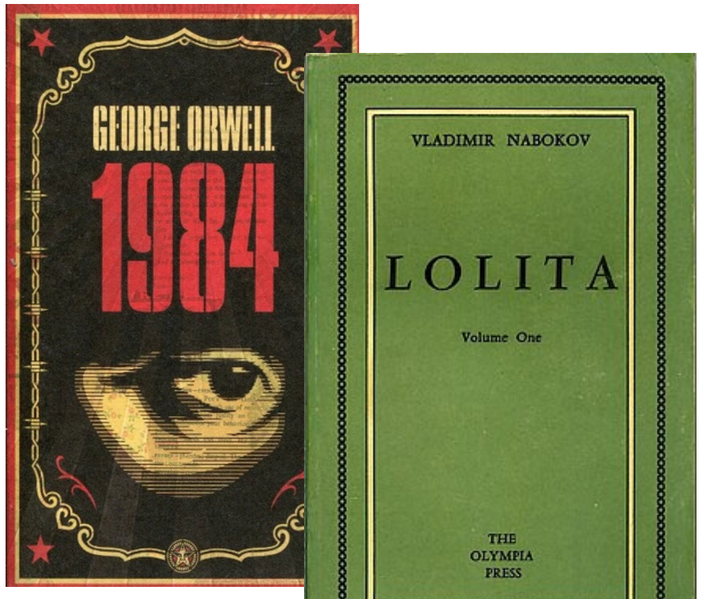
Based on Book Classification



Classic books



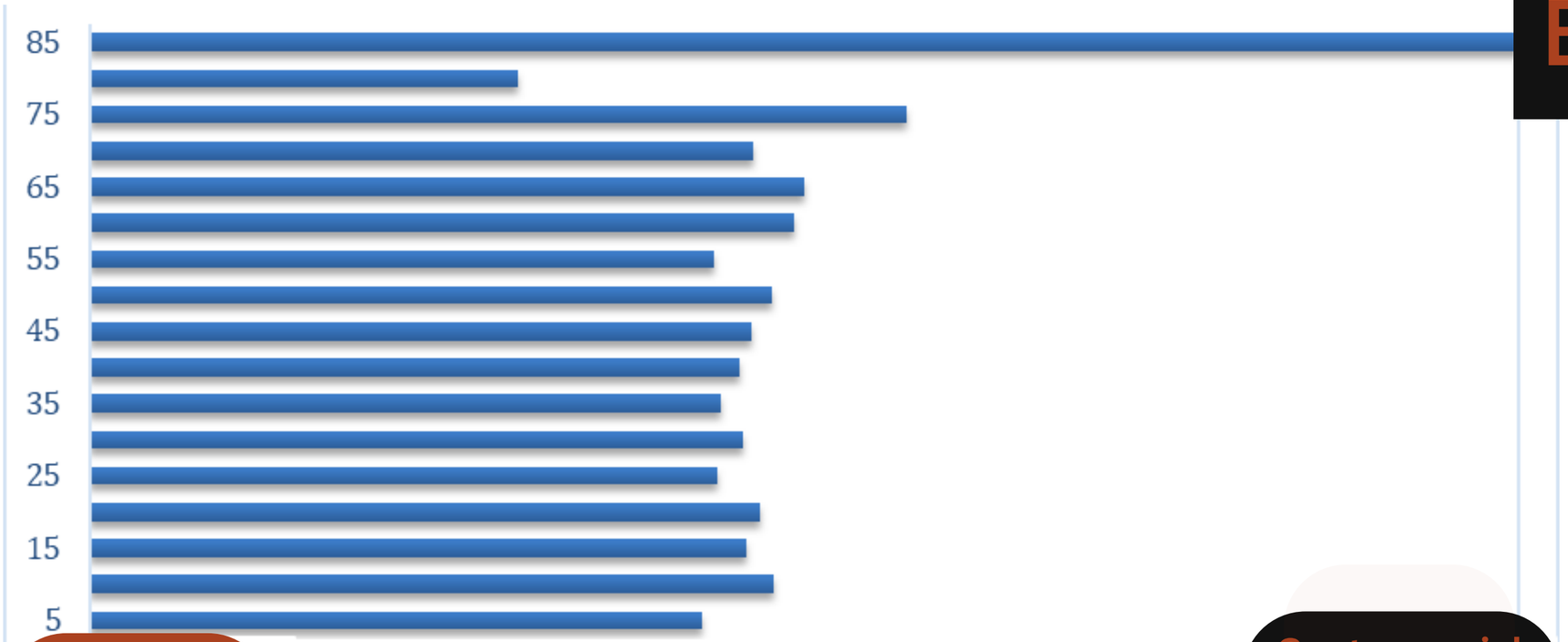
Controversial books



Customer Segmentation Analysis

Based on Book Classification

Age Participation to each segment



Classic books

Controversial books

Do you have any questions?

Team members :

Boura Angeliki
t8190118@aueb.gr

Sotiropoulou Sofia
t8190159@aueb.gr