Amazon Book Sales Data

Team - 50 : Snigdha Raghavan Pradhipa, Vaishnavi Narasimhaiah Sathish,
Samrudhi Ramesh Rao, Kongarasan Sathiya Moorthy

June 1, 2025

1 Dataset

Dataset 1: Amazon Bestsellers (Dataset1_Amazon.xlsx)

- **Source**: Kaggle - Amazon Top 100 Bestsellers (2009–2021)
- **Link:** Amazon Bestsellers
- **Attributes**: ['price', 'ranks', 'title', 'no_of_reviews', 'ratings', 'author', 'cover_type', 'year', 'genre']
- **Description**: Ranked list of top-selling books (2009-2021) with prices, ratings, and sales ranks.

Dataset 2: Goodreads Choice Awards 2023 (Dataset2_GoodreadsAwards.csv)

- **Source**: Kaggle - Goodreads "Best Books of 2023" Winners/Nominees
- **Link:** Good Reads
- **Attributes**: ['source_URL', 'Readers Choice Votes', 'Readers Choice Category', 'Title', 'Author', 'Total Avg Rating', 'Number of Ratings', 'Number of Reviews', 'Number of Pages', 'Edition', 'Book Description', 'First Published date', 'Kindle Version and Price', 'Kindle Price', 'About the Author']
- **Description**: 2023 award-winning books with reader ratings, categories, and Kindle prices.

Dataset 3: Kindle Store (Dataset3_Kindle.csv)

- **Source**: Kaggle - Amazon Kindle Catalog (2023)
- **Link:** Kindle
- **Attributes**: ['asin', 'title', 'author', 'soldBy', 'imgUrl', 'productURL', 'stars', 'reviews', 'price', 'isKindleUnlimited', 'category_id', 'isBestSeller', 'isEditorsPick', 'isGoodReadsChoice', 'publishedDate', 'category_name']
- **Description**: Current Kindle catalog with pricing, star ratings, and award status flags.

2 Competency Questions

Competency questions serve as a benchmark to evaluate the effectiveness and completeness of the mediated schema. These questions are designed to reflect realistic and meaningful information needs that a user might have when interacting with a unified view of multiple data sources.

The following set of 10 competency questions has been formulated to cover a diverse range of aspects in the book domain, including book availability, ratings, pricing, author trends, and platform-specific features. They are intended to validate the correctness of data integration across the Amazon, Goodreads, and Kindle datasets and guide the formulation of queries in a conjunctive form.

1. Which books have more than 50,000 reviews on Amazon and are part of a book series?

2. Which books published before 2010 are bestsellers on Kindle?

3. Which books with a Goodreads average rating of 4.5 or higher are also available on the Kindle platform?

4. What is the distribution of prices (grouped into ranges) for books that appear in both the Amazon Bestsellers and Kindle Store datasets, and what are their average ratings per price range?

5. Which Amazon bestsellers are also available in the Kindle store, and how do their prices compare?

6. Which Goodreads award-winning books are available in the Kindle store, and what are their ratings on both platforms?

7. What are the common genres or categories between the top-selling Amazon books and the Goodreads Choice Awards 2023 winners?

8. Do Goodreads Choice Award-winning books (2023) have higher average Kindle prices than non-award-winning books in the 2023 Amazon Kindle dataset?

9. Which books listed on the Amazon Kindle dataset have the highest rating values but did not win a Goodreads Choice Award in 2023?

10. Which "Fiction" books are available in "Paperback" format in both the Goodreads Awards and Amazon Bestsellers?

11. Which books are priced under 10 dollars and sold by "Amazon.com Services LLC"?

12. Which books won Goodreads Choice Awards in 2023 and have an Amazon sales rank within the top 50?

3 Conjunctive Queries

Conjunctive Queries are formal, logic-based query format used to express information needs in data integration systems. Based on the three datasets and considering their Competency Questions, given below are the conjuctive queries:

1. **Which books have more than 50,000 reviews on Amazon and are part of a book series?**

```
Q(Title, Author, Reviews) :-
AmazonBestsellers(_,_,title,no_of_reviews,_,author,_,_,_true),
no_of_reviews > 50,000
```

2. **Which books published before 2010 are bestsellers on Kindle?**

```
Q(Title, Author, PublishDate) :-
KindleStore(_,title,author,_,_,_,_,_,_,_,_,true,_,_,publishDate,_),
publishDate < 2010-01-01
```

3. **Which books with a Goodreads average rating of 4.5 or higher are also available on the Kindle platform?**

```
Q(Title, Author, PublishDate) :-
GoodreadsAwards(_,_,_,Title,Author,TotalAvgRating,_,_,_,_,_,_,_),
KindleStore(_,title,author,_,_,_,_,_,_,_,_,_,_,true,_,_),
Total_Avg_Rating >= 4.5
```

4. **What is the distribution of prices (grouped into ranges) for books that appear in both the Amazon Bestsellers and Kindle Store datasets, and what are their average ratings per price range?**

```
Q(PriceRange, BookCount, AvgRating) :-
AmazonBestsellers(Title, Author, _, _, _, _, _, _, _, _),
KindleStore(Title, Author, Price, Rating, _, _, _, _, _, _,_, _, _),
PriceRange = CASE
WHEN Price < 5 THEN "0-5"
WHEN Price BETWEEN 5 AND 10 THEN "5-10"
WHEN Price > 10 THEN "10+"
END,
GROUP BY(PriceRange),
BookCount = COUNT(Title),
AvgRating = AVG(Rating)
```

5. **Which Amazon bestsellers are also available in the Kindle store, and how do their prices compare?**

```
Q(Title, Author, AmazonRank, AmazonPrice, KindlePrice, PriceDifference)
    :-
AmazonBestsellers(Title, Author, AmazonRank, AmazonPrice, _, _, _, _, _,
    _),
KindleStore(Title, Author, KindlePrice, _, _, _, _, _, _, _, _, _, _),
AmazonPrice > 0,
KindlePrice > 0,
PriceDifference = AmazonPrice - KindlePrice,
AmazonRank <= 100,
ORDER BY(AmazonRank)
```

6. **Which Goodreads award-winning books are available in the Kindle store, and what are their ratings on both platforms?**

```
Q(Title, Author, AwardCategory, GoodreadsRating, KindleRating,
    RatingDifference) :-
GoodreadsAwards(Title, Author, AwardCategory, GoodreadsRating, _, _, _,
    _, _),
KindleStore(Title, Author, _, KindleRating, _, _, _, _, _, _, _, _, _),
GoodreadsRating > 0,
KindleRating > 0,
RatingDifference = GoodreadsRating - KindleRating,
ORDER BY(RatingDifference DESC)
```

7. **What are the common genres or categories between the top-selling Amazon books and the Goodreads Choice Awards 2023 winners?**

```
Q(CategoryName) :-
AmazonBestSellers(_, _, Title1, _, _, Author1, _, Year, Genre),
GoodreadsAwards(_, _, ReadersChoiceCategory, Title2, Author2, _, _, _, _
    , _, _, _, _, _),
Year >= 2009,
Year =< 2021,
Genre = ReadersChoiceCategory.
```

8. **Do Goodreads Choice Award-winning books (2023) have higher average Kindle prices than non-award-winning books in the 2023 Amazon Kindle dataset?**

```
Q_AvgPrice_Winner(AVG(KindlePrice)) :-
KindleStore(Title, Author, _, _, _, _, _, KindlePrice, _, _, _, _,
    isGoodReadsChoice, PublishedDate, _),
GoodreadsAwards(_, _, _, Title, Author, _, _, _, _, _, _, _, _, _),
isGoodReadsChoice = 'Yes',
PublishedDate >= 2023.

Q_AvgPrice_NonWinner(AVG(KindlePrice)) :-
KindleStore(Title, Author, _, _, _, _, _, KindlePrice, _, _, _, _,
    isGoodReadsChoice, PublishedDate, _),
NOT GoodreadsAwards(_, _, _, Title, Author, _, _, _, _, _, _, _, _, _),
isGoodReadsChoice = 'No',
PublishedDate >= 2023.

Q_Compare(WinnerAvg, NonWinnerAvg, PriceDifference) :-
Q_AvgPrice_Winner(WinnerAvg),
Q_AvgPrice_NonWinner(NonWinnerAvg),
```

```
PriceDifference = WinnerAvg - NonWinnerAvg,
ORDER BY(PriceDifference DESC).
```

9. **Which books listed on the Amazon Kindle dataset have the highest rating values but did not win a Goodreads Choice Award in 2023?**

```
Q(Title, Author, KindleRating) :-
KindleStore(Title, Author, _, _, _, KindleRating, _, _, _, _, _, _,
    isGoodReadsChoice, PublishedDate, _),
KindleRating > 4.5,
isGoodReadsChoice = 'No',
PublishedDate >= 2023,
NOT GoodreadsAwards(_, _, _, Title, Author, _, _, _, _, _, _, _, _, _),
ORDER BY(KindleRating DESC).
```

10. **Which "Fiction" books are available in "Paperback" format in both the Goodreads Awards and Amazon Bestsellers?**

```
Q(Title, Author, Year) :-
GoodreadsAwards(Title, Author, Genre, CoverType, Year, _, _, _, _),
AmazonBestsellers(Title, Author, _, _, _, _, _, Edition),
Genre = "Fiction",
CoverType = "Paperback",
Edition = "Paperback",
ORDER BY(Year ASC)
```

11. **Which books are priced under 10 dollars and sold by "Amazon.com Services LLC?**

```
Q(Title, Author, Price) :-
KindleStore(Title, Author, Genre, Rating, Price, Seller),
Price < 10,
Seller = "Amazon.com Services LLC",
ORDER BY(Price DESC).
```

12. **Which books won Goodreads Choice Awards in 2023 and have an Amazon sales rank within the top 50?**

```
Q(Title, Author, Genre, AmazonRank) :-
GoodreadsAwards(Title, Author, Genre, _,  _, _, Year),
Year = 2023,
AmazonBestsellers(Title, Author, AmazonRank,  _, _, _, _),
AmazonRank =< 50,
ORDER BY(AmazonRank ASC).
```

4 Mediated Schema

The mediated schema represents a unified, abstract layer that integrates the heterogeneous data from Amazon, Goodreads, and Kindle datasets into a common structure. It serves as a global schema through which users can query across all three sources using a consistent vocabulary, regardless of differences in column names, formats, or structure in the original datasets.

It maps different attribute names- like ratings, reviews, and genre - into unified schema terms. Fields that are unique to each source are matched to common ones like award_status, review_count, and sales_rank. The result is a cohesive schema that enables smooth integration, comparison, and querying across all three datasets.

4.1 Dataset Schema and Tables:

AmazonBestsellers('price', 'ranks', 'title', 'no_of_reviews', 'ratings', 'author', 'cover_type', 'year', 'genre')

GoodreadsAwards('source_URL', 'Readers Choice Votes', 'Readers Choice Category', 'Title', 'Author', 'Total Avg Rating', 'Number of Ratings', 'Number of Reviews', 'Number of Pages', 'Edition', 'Book Description', 'First Published date', 'Kindle Version and Price', 'Kindle Price', 'About the Author')

KindleStore('asin', 'title', 'author', 'soldBy', 'imgUrl', 'productURL', 'stars', 'reviews', 'price', 'isKindleUnlimited', 'category_id', 'isBestSeller', 'isEditorsPick', 'isGoodReadsChoice', 'publishedDate', 'category_name')

4.2 Mediated Schema:

Table 1: Attribute Mapping to Mediated Schema

Amazon Bestsellers	Goodreads Awards	Kindle Store	Mediated Schema Attribute
title	Title	title	title
author	Author	author	author
ratings	Total Avg Rating	stars	rating
no_of_reviews	Number of Reviews	reviews	review_count
price	Kindle Price	price	price
ranks	-	-	sales_rank
genre	-	category_name	genre
-	Award Category	isGoodReadsChoice	award_status
year	First Published date	publishedDate	publication_year

Mediated schema for the chosen dataset is as shown below;

Books('title','author','publishDate','price')

Rating('ratings','reviews')

4.3 Heterogeneity

Heterogeneity refers to the differences that arise when multiple people or organizations create schemas or data systems, even when they aim to represent the same subject or domain. These differences are inevitable and can pose significant challenges during data integration and interoperability. Understanding and addressing these types of heterogeneity is critical for effective data integration and ensuring that data from diverse sources can be accurately aligned and used together.

	price	ranks	title	no_of_reviews	ratings	author	cover_type	year	genre
0	12.49	1	The Lost Symbol	16,118	4.4	Dan Brown	Hardcover	2009	Fiction

Figure 1: Dataset 1: Amazon Bestsellers (Dataset1_Amazon.xlsx)

source_URL	Readers Choice Votes	Readers Choice Category	Title	Author	Total Avg Rating	Number of Ratings	Number of Reviews	Number of Pages	Edition	Book Description	First Published date	Kindle Version and Price	Kindle Price	About the Author
https://www.goodreads.com/book/show/82047984-yellowface?from_choice=true	200722	Fiction	Yellowface	R.F. Kuang	3.87	2,52,189	37,032	336	Hardcover	Athena Liu is a literary darling and June Hayward is literally nobody....	29-05-2023	Kindle $14.99	14.99	Rebecca F. Kuang is a Marshall Scholar, translator, and award-winning, ...

Figure 2: Dataset 2: Goodreads Choice Awards 2023 (Dataset2_GoodreadsAwards.csv)

asin	title	author	soldBy	imgUrl	productURL	stars	reviews	price	isKindleUnlimited	category_id	isBestSeller	isEditorsPick	isGoodReadsChoice	publishedDate	category_name
B00TZE8754	Adult Children of Emotionally Immature Parents: How to Heal from Distant, Rejecting, or Self-Involved Parents	Lindsay C. Gibson	Amazon.com Services LLC	https://m.media-amazon.com/images/I/713KZ1aWYjL._AC_UY218_.jpg	https://www.amazon.com/dp/B00TZE8754	4.8	0	9.99	FALSE	6	TRUE	FALSE	FALSE	01-06-2015	Parenting & Relationships

Figure 3: Dataset 3: Kindle Store (Dataset3_Kindle.csv)

5 Contributions

Snigdha Raghavan Pradhipa

- Dataset collection from Kaggle
- Competency question design
- Datalog query formulation
- Attribute mapping
- Mediated schema development
- Quality assurance & Report documentation

Vaishnavi Narasimhaiah Sathish

- Dataset collection from Kaggle
- Competency question design

- Datalog query formulation
- Attribute mapping
- Mediated schema development
- Quality assurance & Report documentation

Samrudhi Ramesh Rao

- Dataset collection from Kaggle
- Competency question design
- Datalog query formulation
- Attribute mapping
- Mediated schema development
- Quality assurance & Report documentation

Kongarasan Sathiya Moorthy

- Dataset collection from Kaggle
- Competency question design
- Datalog query formulation
- Attribute mapping
- Mediated schema development
- Quality assurance & Report documentation