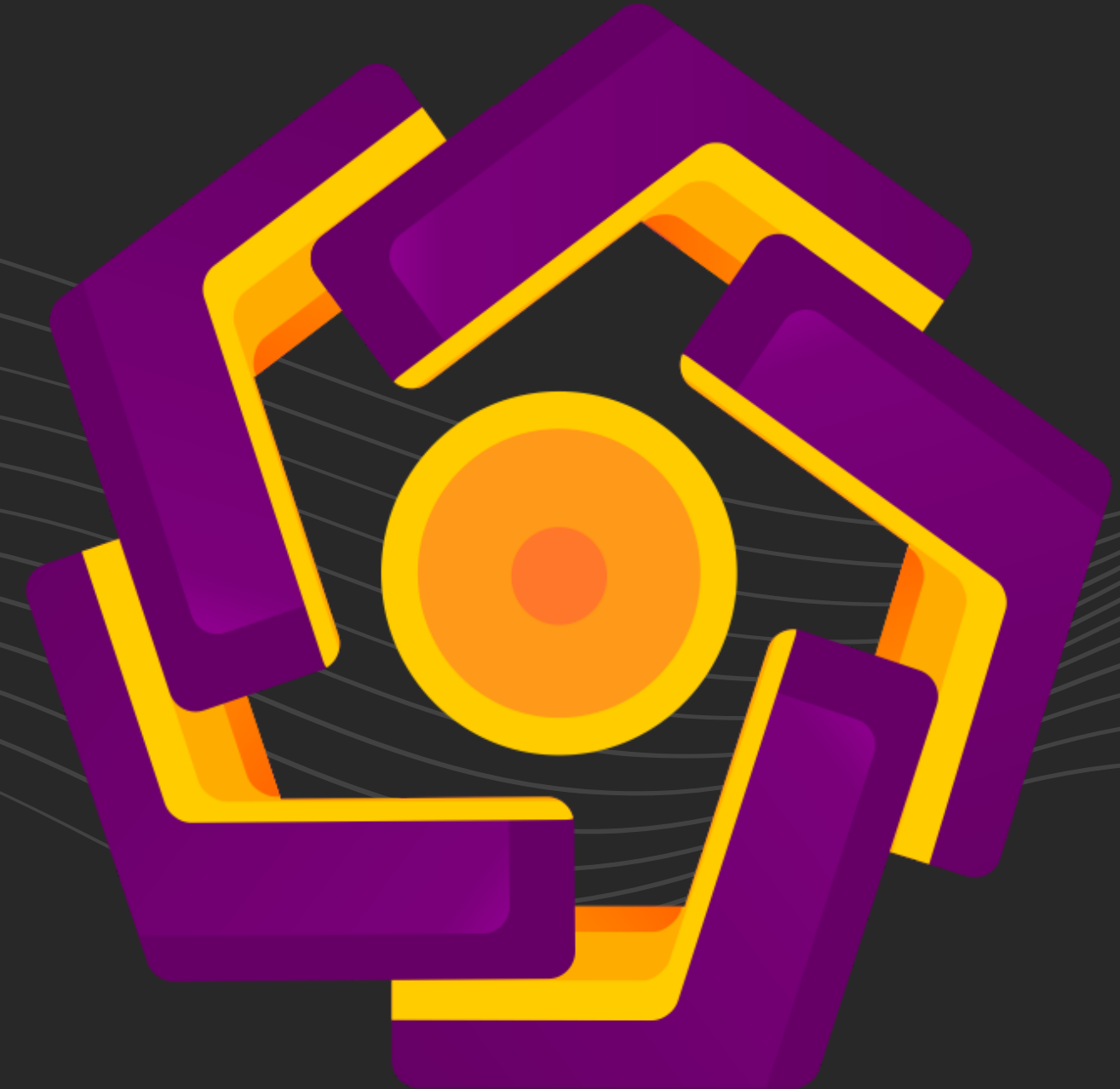


# Final Project BDL : Book Dataset

## Kelompok Banteng aja :


Nama Anggota :

1. Rizqi Bio Janhefi (19.11.2899)
2. M.Muh.Akhyar.Z (19.11.2892)
3. Abdul Rafiq (19.11.2894)
4. Vicky Reza Fadillah (19.11.2904)
5. Meisya Varadilla S (19.11.2922)



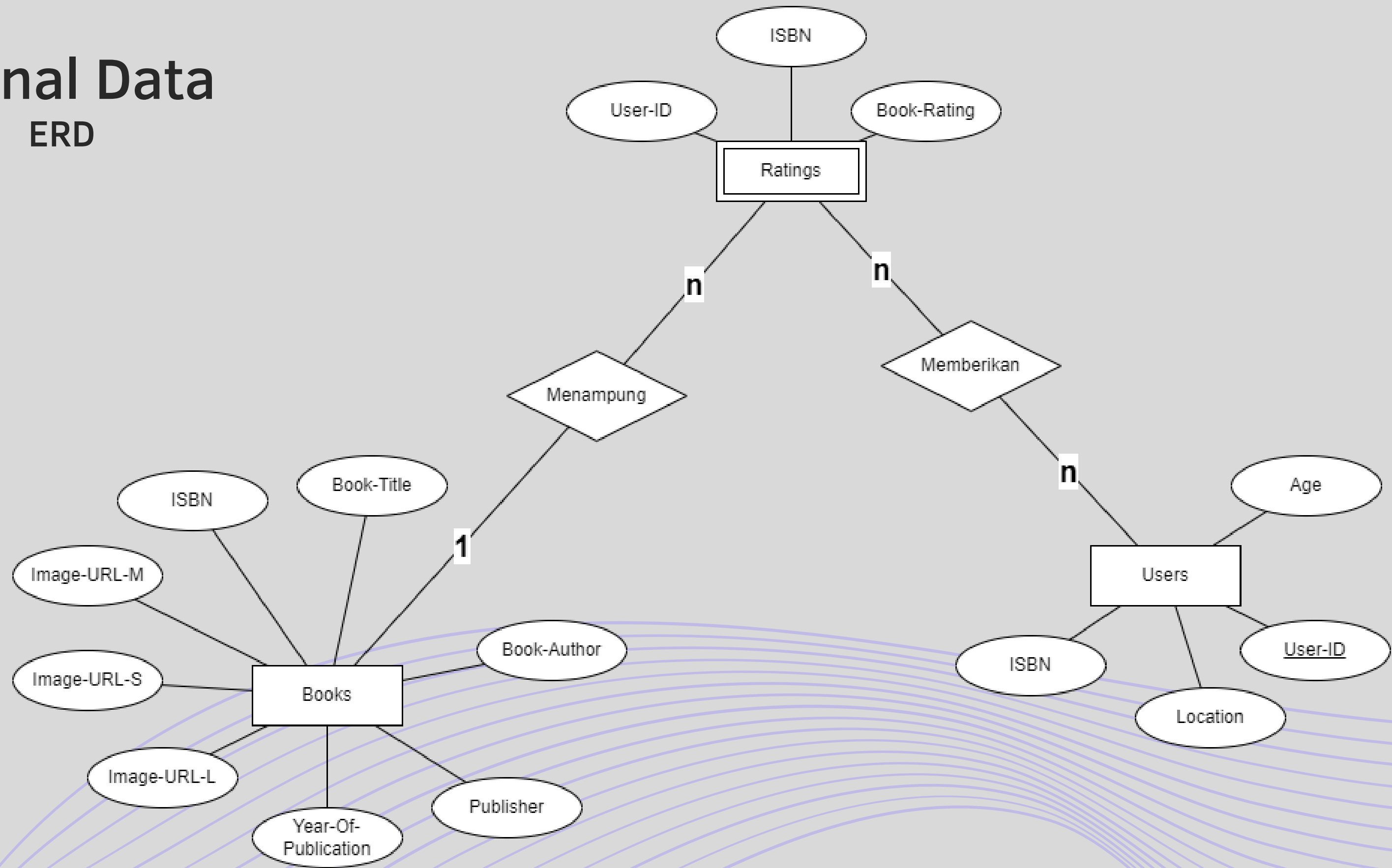
# Informasi Dataset

Dataset yang kami gunakan adalah dataset "Book Dataset" , yang mana dataset tersebut kami ambil dari kaggle. Dataset ini dikumpulkan oleh Cai-Nicolas Ziegler dari komunitas Book-Crossing. Dataset ini memiliki format csv berisi 278.858 users/pembaca yang memberikan 1.149.780 rating terhadap 271.379 buku yang mereka baca.



# Relasional Data

## ERD



# Relational Data

Relationship : Document Reference

```
1 {
2   $match: {
3     'Book-Title': 'Flesh Tones: A Novel'
4   }
5 }, {
6   $lookup: {
7     from: 'ratings',
8     localField: 'ISBN',
9     foreignField: 'ISBN',
10    as: 'Ratings'
11  }
12 }
```



```
_id: ObjectId('62cd706aa5e18aeffdc34d5a')
ISBN: "034545104X"
Book-Title: "Flesh Tones: A Novel"
Book-Author: "M. J. Rose"
Year-Of-Publication: "2002"
Publisher: "Ballantine Books"
Image-URL-S: "http://images.amazon.com/images/P/034545104X.01.THUMBZZZ.jpg"
Image-URL-M: "http://images.amazon.com/images/P/034545104X.01.MZZZZZZZ.jpg"
Image-URL-L: "http://images.amazon.com/images/P/034545104X.01.LZZZZZZZ.jpg"
✓ Ratings: Array
  > 0: Object
  > 1: Object
  > 2: Object
  > 3: Object
```

# Relasional Data

Relationship : Merge

Pada python kami menggunakan merge untuk menghubungkan tiap-tiap collection pada database

```
"MERELASIKAN COLLECTION"
```

```
ratings_books = ratings.merge(books,on='ISBN').merge(users,on='User-ID')  
ratings_books.head()
```

# Konversi Dataset ke MongoDB Cluster

```
mongodb+srv://bantengaja:<password>@cluster0.dornyly.mongodb.net/?
retryWrites=true&w=majority
```

(1) Melakukan koneksi ke MongoDB Cluster yang telah dibuat

(3) Melakukan import file dataset pada tiap-tiap collection yang dibuat serta melakukan penyesuaian terhadap tipe data pada tiap-tiap atribut

Select File

Books.csv

Select Input File Type

JSON CSV

Options

Select delimiter COMMA

☒ Ignore empty strings

☐ Stop on errors

	Book-Author	Year-Of-Publication	Publisher	Image-URL-S
	<input checked="" type="checkbox"/> String	<input checked="" type="checkbox"/> Int32	<input checked="" type="checkbox"/> String	<input checked="" type="checkbox"/> String
	Mark P. O. Morford	200	xford University Press	http://images.ar
	Richard Bruce Wright	200	arnerElamingo Canada	http://images.ar

(2) Membuat database beserta collection

Book

- books
- ratings
- users

# Data Mining (Data Cleaning)

```
ratings_books.columns.values
```

```
array(['_id_x', 'User-ID', 'ISBN', 'Book-Rating', '_id_y', 'Book-Title',  
      'Book-Author', 'Year-Of-Publication', 'Publisher', 'Image-URL-S',  
      'Image-URL-M', 'Image-URL-L', '_id', 'Location', 'Age'],  
      dtype=object)
```

```
"DROP COLOM YANG TIDAK DIGUNAKAN"
```

```
column_drop = [  
    '_id_x', 'User-ID', '_id_y', 'ISBN', 'Book-Author',  
    'Year-Of-Publication', 'Publisher', 'Image-URL-S',  
    'Image-URL-M', 'Image-URL-L', '_id', 'Location']  
ratings_books.drop(column_drop, inplace=True, axis=1)  
ratings_books.columns.values
```

```
array(['Book-Rating', 'Book-Title', 'Age'], dtype=object)
```

```
ratings_books.head()
```

	Book-Rating	Book-Title	Age
0	0	Flesh Tones: A Novel	NaN
1	5	Flesh Tones: A Novel	23.0
2	9	Ender's Game (Ender Wiggins Saga (Paperback))	23.0
3	8	In Cold Blood (Vintage International)	23.0
4	9	Divine Secrets of the Ya-Ya Sisterhood : A Novel	23.0

Untuk tahapan awal, tentukan terlebih dahulu kolom-kolom yang digunakan dalam perhitungan. Lalu untuk kolom yang tidak digunakan atau tidak diperlukan maka kolom tersebut akan di drop.

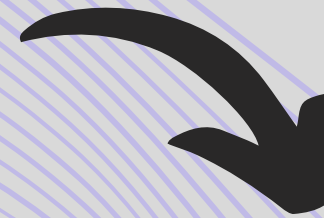


# Data Mining (Data Cleaning)

```
print(df_age.isna().sum())
```

```
Book-Rating    0  
Book-Title     0  
Age           361  
dtype: int64
```

Proses data cleaning terdapat pada sebuah dataframe, yang mana terdapat Missing Value pada value field "Age"

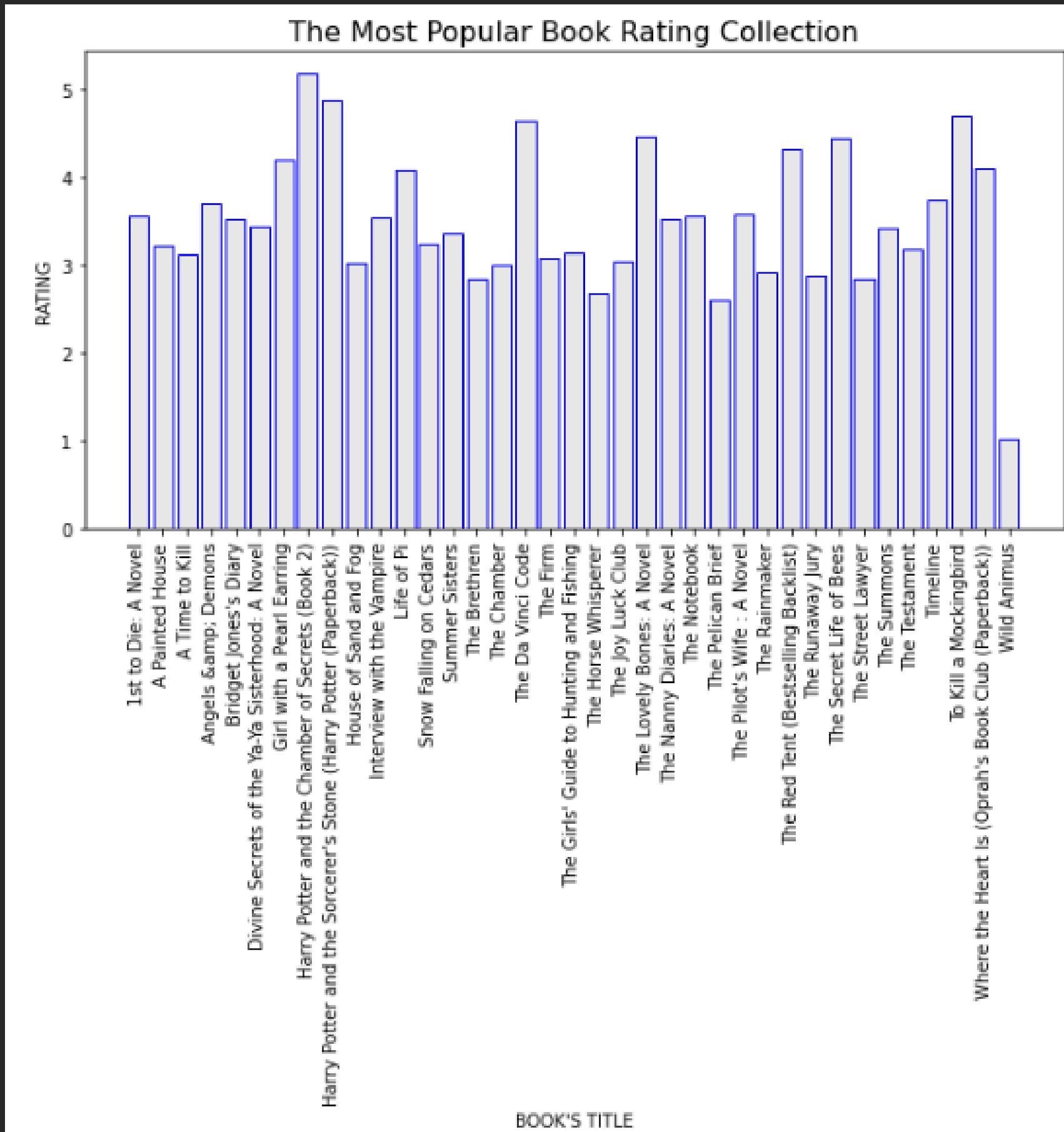


Melakukan pengisian data Missing Value dengan Value rata-rata umur

```
mean_age = df_age['Age'].astype(float).mean()  
df_distribution = df_age.fillna({'Age':mean_age})
```



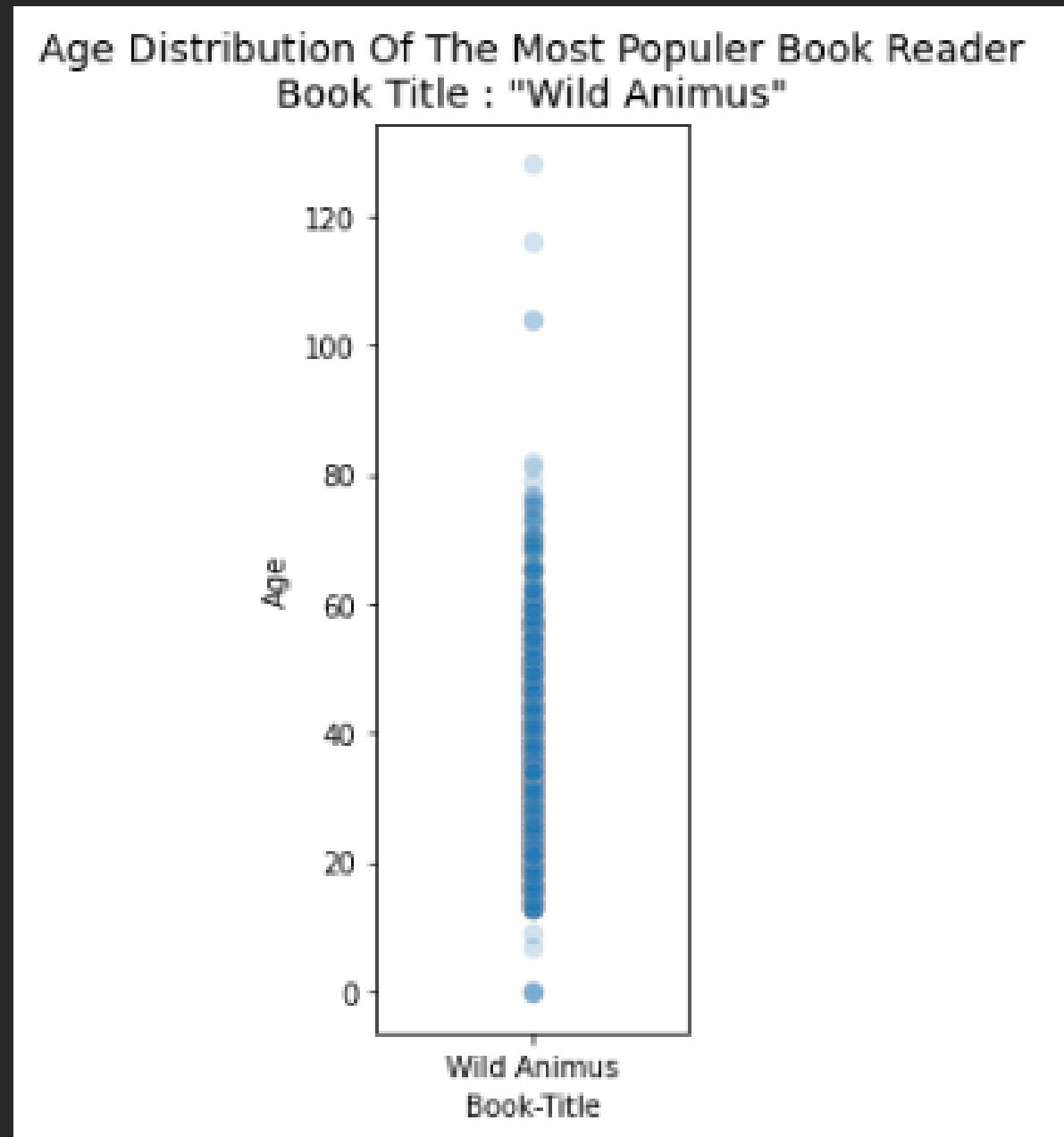
# Data Mining (Representasi Data )



Dengan menggunakan Histogram Diagram kami merepresentasikan data Rating Buku Populer dari dataset Book.

Pada diagram dapat dilihat tinggi rendahnya bar menunjukan besar atau kecilnya rating dari tiap-tiap buku.

# Data Mining (Representasi Data )



Representasi data sebaran umur pembaca buku paling populer "Wild Animus".

Pada diagram dapat kita lihat semakin terang titik biru menunjukkan sebaran umur pembaca.

# Kesimpulan

Jadi kesimpulannya adalah dari dataset book yang berisi 278.858 users/pembaca yang memberikan 1.149.780 rating terhadap 271.379 buku yang mereka baca, setelah melakukan proses data mining didapatkan beberapa informasi data rating buku terpopuler dengan melihat ukuran dari tiap-tiap bar kita dapat mengetahui tinggi rendahnya rating buku.