Analyzing Reading Habit using Goodreads Data

Background

This data analysis project is an exploration of my **Goodreads Library data**. Goodreads, a popular book recommendation and cataloging website, provides a platform for readers to track their reading progress, rate books, write reviews, and interact with other readers. As a user of Goodreads, I have amassed a good amount of data about my reading habits over the years, from the books I read, to the books that I want to read.

Business Task Definition

The goal of this project is to extract and visualize trends, patterns, and insights on my reading habits. The analysis will cover various aspects such as the number of books I have read over a certain period of time, my preferences in genre based on the books I read, and the speed at which I read.

Data Source

The primary data source for this project is `goodreads_library_export.csv` . This data was exported directly from my Goodreads account, which I have been using to track my reading progress, rate books, write reviews, and interact with other readers. The exported Goodreads CSV file was relatively small, with a data volume of approximately 39 kilobytes (kB) and contains information about my Goodreads library data in across 154 observations with 24 features for each book. Comics, web novels, and similar categories are not included in this analysis.

Data Collection

The data was collected using the **export feature** available in Goodreads. This feature allows users to download a CSV file containing information about all the books in their library.

Data Features

The exported data includes several features for each book:

- **Book Id**: A unique identifier for each book.
- **Title**: The title of the book.
- **Author**: The author of the book.
- **Author l-f**: The author’s name in “last name, first name” format.

- **Additional Authors**: Any additional authors of the book.
- **ISBN**: The International Standard Book Number of the book.
- **ISBN13**: The 13-digit International Standard Book Number of the book.
- **My Rating**: My rating for the book on a scale of 1 to 5.
- **Average Rating**: The average rating of the book by all Goodreads users.
- **Publisher**: The publisher of the book.
- **Binding**: The type of book binding (e.g., paperback, hardcover).
- **Number of Pages**: The number of pages in the book.
- **Year Published**: The year the book edition was published.
- **Original Publication Year**: The year the book was originally published.
- **Date Read**: The date I finished reading the book.
- **Date Added**: The date I added the book to my Goodreads library.
- **Bookshelves**: The Goodreads shelves (categories) I have placed the book on.
- **Bookshelves with positions**: The position of the book on my Goodreads shelves.
- **Exclusive Shelf**: The exclusive shelf (Read, Currently Reading, Want to Read) the book is on.
- **My Review**: My review of the book.
- **Spoiler**: Whether my review contains spoilers.
- **Private Notes**: Any private notes I have made about the book.
- **Read Count**: The number of times I have read the book.
- **Owned Copies**: The number of copies of the book I own.

Data Quality

The data exported from Goodreads is generally clean and well-structured. However, there may be some unnecessary, missing or inconsistent data, These issues will be addressed during the data cleaning process.

Data Privacy

To ensure privacy, any sensitive information in the dataset, such as personal notes or private reviews, has been removed before the analysis.

Deliverables

1. **A Cleaned and Prepared Dataset**: The Goodreads library data is cleaned and prepared for analysis. This includes handling missing values, dealing with inconsistencies, ad removing irrelevant information.

2. **Data Analysis Report**: A report documenting the analysis process and findings. This includes the methodology used for the analysis, the results, and any insights gained from the analysis.
3. **Visualizations or Graphs**: Visual representations of the analyzed data showing key insights. These could include bar charts, line graphs, pie charts, etc., depending on the type of data and the insights being presented.
4. **Presentation**: A presentation summarizing the project and its results. This includes an overview of the project, the methodology used, the key findings, and the insights gained from the analysis.

Tools Used

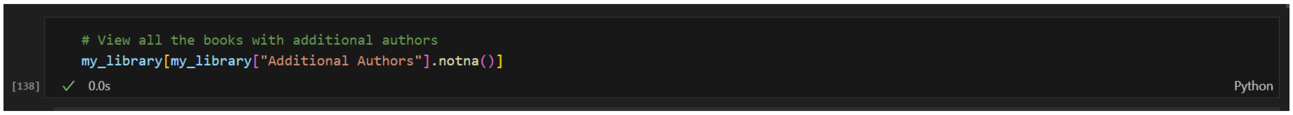
This project utilizes a variety of tools for data cleaning, processing, analysis, and visualization:

- **Python**: The main programming language used for this project.
- **Pandas**: A Python library used for data cleaning and processing.
- **Tableau**: A powerful data visualization tool used for creating interactive graphs, charts, maps, and more for the analysis.

Data Cleaning and Processing

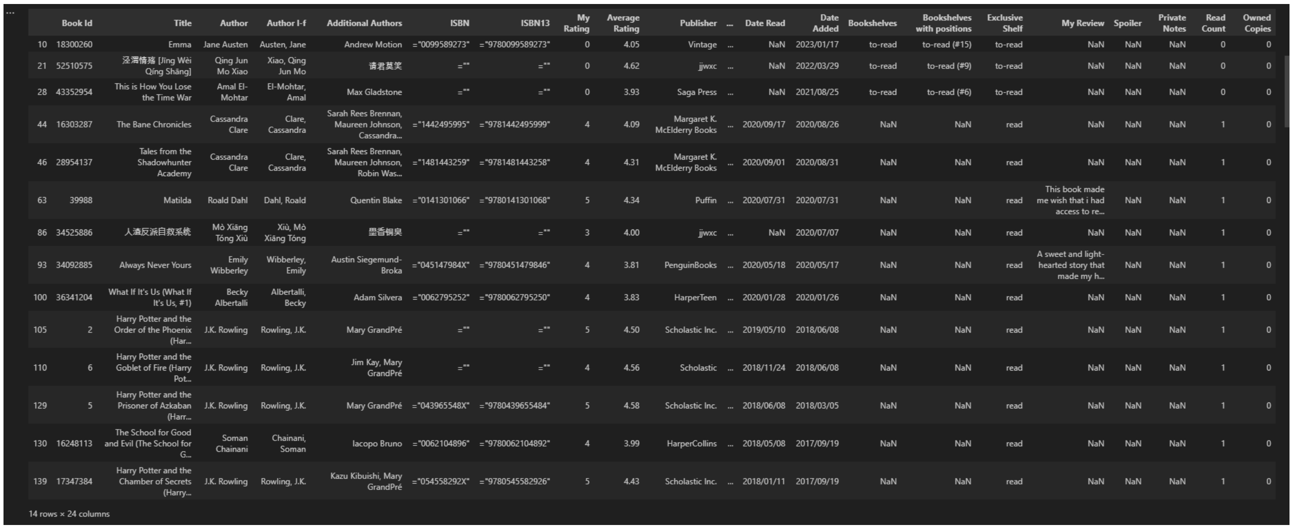
The data set is exported from Goodreads raw, so there are more chances of blank or dirty data. In order to clean obtained data, following cleaning tasks have been executed:

- Checking for books with additional authors.



```
# View all the books with additional authors
my_library[my_library["Additional Authors"].notna()]
```

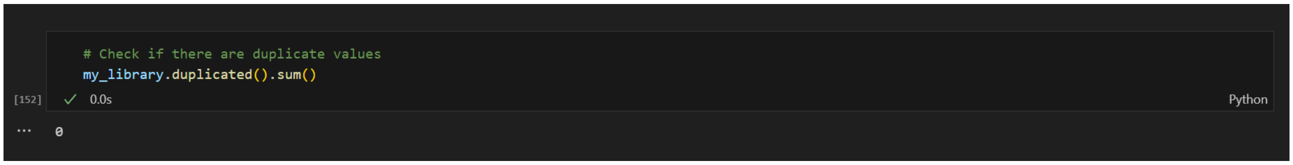
Result:



	Book Id	Title	Author	Author l-f	Additional Authors	ISBN	ISBN13	My Rating	Average Rating	Publisher	...	Date Read	Date Added	Bookshelves	Bookshelves with positions	Exclusive Shelf	My Review	Spoiler	Private Notes	Read Count	Owned Copies
10	18300260	Emma	Jane Austen	Austen, Jane	Andrew Motion	="0099589273"	="9780099589273"	0	4.05	Vintage	...	NaN	2023/01/17	to-read	to-read (#15)	to-read	NaN	NaN	NaN	0	0
21	52510575	泾渭情殇 [Jīng Wèi Qíng Shāng]	Qing Jun Mo Xiao	Xiao, Qing Jun Mo	请君莫笑	=""	=""	0	4.62	jjwxc	...	NaN	2022/03/29	to-read	to-read (#9)	to-read	NaN	NaN	NaN	0	0
28	43352954	This is How You Lose the Time War	Amal El-Mohtar	El-Mohtar, Amal	Max Gladstone	=""	=""	0	3.93	Saga Press	...	NaN	2021/08/25	to-read	to-read (#6)	to-read	NaN	NaN	NaN	0	0
44	16303287	The Bane Chronicles	Cassandra Clare	Clare, Cassandra	Sarah Rees Brennan, Maureen Johnson, Cassandra...	="1442495995"	="9781442495999"	4	4.09	Margaret K. McElderry Books	...	2020/09/17	2020/08/26	NaN	NaN	read	NaN	NaN	NaN	1	0
46	28954137	Tales from the Shadowhunter Academy	Cassandra Clare	Clare, Cassandra	Sarah Rees Brennan, Maureen Johnson, Robin Was...	="1481443259"	="9781481443258"	4	4.31	Margaret K. McElderry Books	...	2020/09/01	2020/08/31	NaN	NaN	read	NaN	NaN	NaN	1	0
63	39988	Matilda	Roald Dahl	Dahl, Roald	Quentin Blake	="0141301066"	="9780141301068"	5	4.34	Puffin	...	2020/07/31	2020/07/31	NaN	NaN	read	This book made me wish that i had access to re...	NaN	NaN	1	0
86	34525886	人渣反派自救系统	Mò Xiāng Tóng Xiù	Xiù, Mò Xiāng Tóng	墨香铜臭	=""	=""	3	4.00	jjwxc	...	NaN	2020/07/07	NaN	NaN	read	NaN	NaN	NaN	1	0
93	34092885	Always Never Yours	Emily Wibberley	Wibberley, Emily	Austin Siegemund-Broka	="045147984X"	="9780451479846"	4	3.81	PenguinBooks	...	2020/05/18	2020/05/17	NaN	NaN	read	A sweet and light-hearted story that made my h...	NaN	NaN	1	0
100	36341204	What If It's Us (What If It's Us, #1)	Becky Albertalli	Albertalli, Becky	Adam Silvera	="0062795252"	="9780062795250"	4	3.83	HarperTeen	...	2020/01/28	2020/01/26	NaN	NaN	read	NaN	NaN	NaN	1	0
105	2	Harry Potter and the Order of the Phoenix (Har...	J.K. Rowling	Rowling, J.K.	Mary GrandPré	=""	=""	5	4.50	Scholastic Inc.	...	2019/05/10	2018/06/08	NaN	NaN	read	NaN	NaN	NaN	1	0
110	6	Harry Potter and the Goblet of Fire (Harry Pot...	J.K. Rowling	Rowling, J.K.	Jim Kay, Mary GrandPré	=""	=""	4	4.56	Scholastic	...	2018/11/24	2018/06/08	NaN	NaN	read	NaN	NaN	NaN	1	0
129	5	Harry Potter and the Prisoner of Azkaban (Harr...	J.K. Rowling	Rowling, J.K.	Mary GrandPré	="043965548X"	="9780439655484"	5	4.58	Scholastic Inc.	...	2018/06/08	2018/03/05	NaN	NaN	read	NaN	NaN	NaN	1	0
130	16248113	The School for Good and Evil (The School for G...	Soman Chainani	Chainani, Soman	Iacopo Bruno	="0062104896"	="9780062104892"	4	3.99	HarperCollins	...	2018/05/08	2017/09/19	NaN	NaN	read	NaN	NaN	NaN	1	0
139	17347384	Harry Potter and the Chamber of Secrets (Harry...	J.K. Rowling	Rowling, J.K.	Kazu Kibuishi, Mary GrandPré	="054558292X"	="9780545582926"	5	4.43	Scholastic Inc.	...	2018/01/11	2017/09/19	NaN	NaN	read	NaN	NaN	NaN	1	0

14 rows × 24 columns

- Checking for duplicate values.

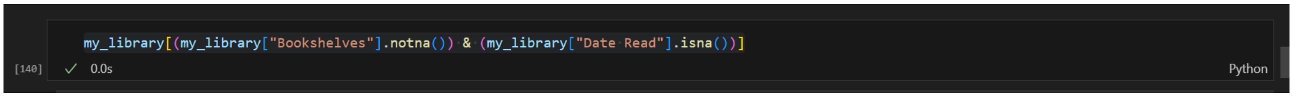


```
# Check if there are duplicate values
my_library.duplicated().sum()
```

[152] ✓ 0.0s Python

```
0
```

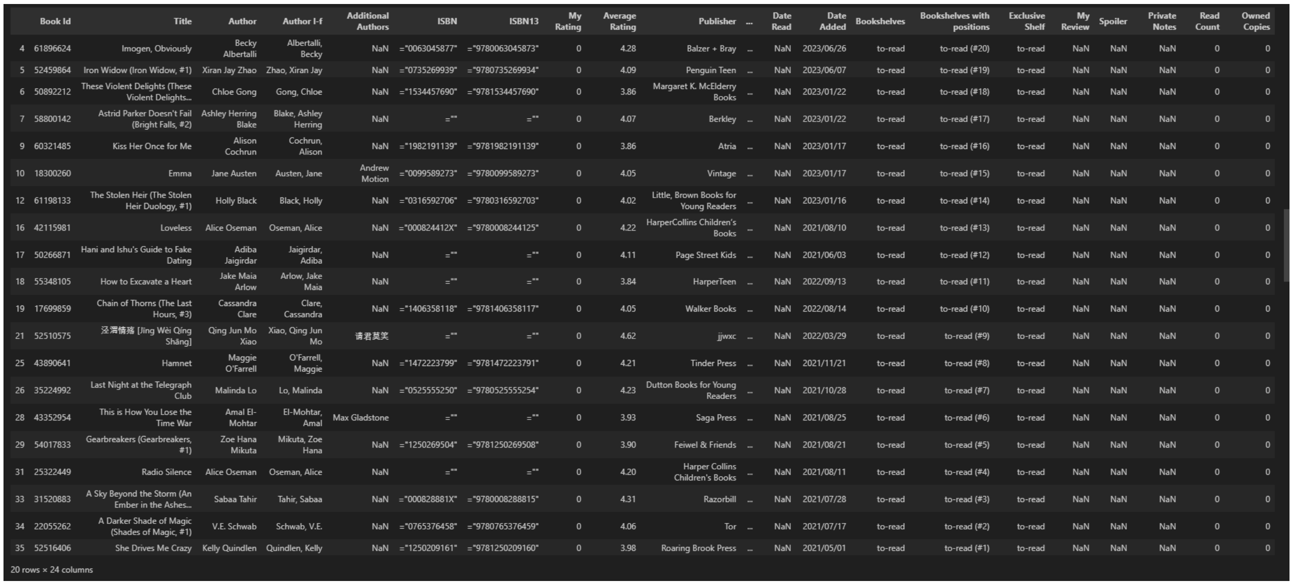
- Checking for Books that are not "Read".



```
my_library[(my_library["Bookshelves"].notna()) & (my_library["Date Read"].isna())]
```

[140] ✓ 0.0s Python

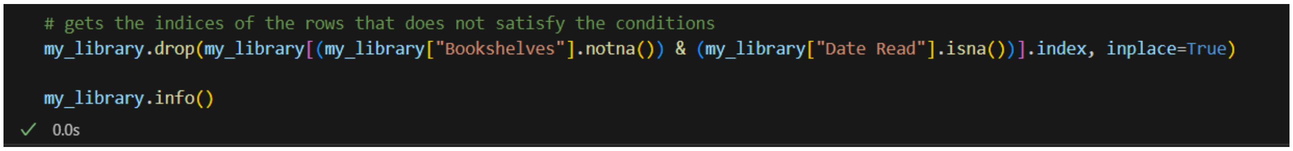
Result:



	Book Id	Title	Author	Author l-f	Additional Authors	ISBN	ISBN13	My Rating	Average Rating	Publisher	...	Date Read	Date Added	Bookshelves	Bookshelves with positions	Exclusive Shelf	My Review	Spoiler	Private Notes	Read Count	Owned Copies
4	61896624	Imogen, Obviously	Becky Albertalli	Albertalli, Becky	NaN	="0063045877"	="9780063045873"	0	4.28	Balzer + Bray	...	NaN	2023/06/26	to-read	to-read (#20)	to-read	NaN	NaN	NaN	0	0
5	52459864	Iron Widow (Iron Widow, #1)	Xiran Jay Zhao	Zhao, Xiran Jay	NaN	="0735269939"	="9780735269934"	0	4.09	Penguin Teen	...	NaN	2023/06/07	to-read	to-read (#19)	to-read	NaN	NaN	NaN	0	0
6	50892212	These Violent Delights (These Violent Delights...	Chloe Gong	Gong, Chloe	NaN	="1534457690"	="9781534457690"	0	3.86	Margaret K. McElderry Books	...	NaN	2023/01/22	to-read	to-read (#18)	to-read	NaN	NaN	NaN	0	0
7	58800142	Astrid Parker Doesn't Fail (Bright Falls, #2)	Ashley Herring Blake	Blake, Ashley Herring	NaN	=""	=""	0	4.07	Berkley	...	NaN	2023/01/22	to-read	to-read (#17)	to-read	NaN	NaN	NaN	0	0
9	60321485	Kiss Her Once for Me	Alison Cochrun	Cochrun, Alison	NaN	="1982191139"	="9781982191139"	0	3.86	Atria	...	NaN	2023/01/17	to-read	to-read (#16)	to-read	NaN	NaN	NaN	0	0
10	18300260	Emma	Jane Austen	Austen, Jane	Andrew Motion	="0099589273"	="9780099589273"	0	4.05	Vintage	...	NaN	2023/01/17	to-read	to-read (#15)	to-read	NaN	NaN	NaN	0	0
12	61198133	The Stolen Heir (The Stolen Heir Duology, #1)	Holly Black	Black, Holly	NaN	="0316592706"	="9780316592703"	0	4.02	Little, Brown Books for Young Readers	...	NaN	2023/01/16	to-read	to-read (#14)	to-read	NaN	NaN	NaN	0	0
16	42115981	Loveless	Alice Oseman	Oseman, Alice	NaN	="000824412X"	="9780008244125"	0	4.22	HarperCollins Children's Books	...	NaN	2021/08/10	to-read	to-read (#13)	to-read	NaN	NaN	NaN	0	0
17	50266871	Hani and Ishu's Guide to Fake Dating	Adiba Jaigirdar	Jaigirdar, Adiba	NaN	=""	=""	0	4.11	Page Street Kids	...	NaN	2021/06/03	to-read	to-read (#12)	to-read	NaN	NaN	NaN	0	0
18	55348105	How to Excavate a Heart	Jake Maia Arlow	Arlow, Jake Maia	NaN	=""	=""	0	3.84	HarperTeen	...	NaN	2022/09/13	to-read	to-read (#11)	to-read	NaN	NaN	NaN	0	0
19	17699859	Chain of Thorns (The Last Hours, #3)	Cassandra Clare	Clare, Cassandra	NaN	="1406358118"	="9781406358117"	0	4.05	Walker Books	...	NaN	2022/08/14	to-read	to-read (#10)	to-read	NaN	NaN	NaN	0	0
21	52510575	泾渭情殇 [Jīng Wèi Qíng Shāng]	Qing Jun Mo Xiao	Xiao, Qing Jun Mo	请君莫笑	=""	=""	0	4.62	jjwxc	...	NaN	2022/03/29	to-read	to-read (#9)	to-read	NaN	NaN	NaN	0	0
25	43890641	Hamnet	Maggie O'Farrell	O'Farrell, Maggie	NaN	="1472223799"	="9781472223791"	0	4.21	Tinder Press	...	NaN	2021/11/21	to-read	to-read (#8)	to-read	NaN	NaN	NaN	0	0
26	35224992	Last Night at the Telegraph Club	Malinda Lo	Lo, Malinda	NaN	="0525555250"	="9780525555254"	0	4.23	Dutton Books for Young Readers	...	NaN	2021/10/28	to-read	to-read (#7)	to-read	NaN	NaN	NaN	0	0
28	43352954	This is How You Lose the Time War	Amal El-Mohtar	El-Mohtar, Amal	Max Gladstone	=""	=""	0	3.93	Saga Press	...	NaN	2021/08/25	to-read	to-read (#6)	to-read	NaN	NaN	NaN	0	0
29	54017833	Gearbreakers (Gearbreakers, #1)	Zoe Hana Mikuta	Mikuta, Zoe Hana	NaN	="1250269504"	="9781250269508"	0	3.90	Feiwel & Friends	...	NaN	2021/08/21	to-read	to-read (#5)	to-read	NaN	NaN	NaN	0	0
31	25322449	Radio Silence	Alice Oseman	Oseman, Alice	NaN	=""	=""	0	4.20	Harper Collins Children's Books	...	NaN	2021/08/11	to-read	to-read (#4)	to-read	NaN	NaN	NaN	0	0
33	31520883	A Sky Beyond the Storm (An Ember in the Ashes...	Sabaa Tahir	Tahir, Sabaa	NaN	="000828881X"	="9780008288815"	0	4.31	Razorbill	...	NaN	2021/07/28	to-read	to-read (#3)	to-read	NaN	NaN	NaN	0	0
34	22055262	A Darker Shade of Magic (Shades of Magic, #1)	V.E. Schwab	Schwab, V.E.	NaN	="0765376458"	="9780765376459"	0	4.06	Tor	...	NaN	2021/07/17	to-read	to-read (#2)	to-read	NaN	NaN	NaN	0	0
35	52516406	She Drives Me Crazy	Kelly Quindlen	Quindlen, Kelly	NaN	="1250209161"	="9781250209160"	0	3.98	Roaring Brook Press	...	NaN	2021/05/01	to-read	to-read (#1)	to-read	NaN	NaN	NaN	0	0

20 rows × 24 columns

- Dropping books that has not been read.

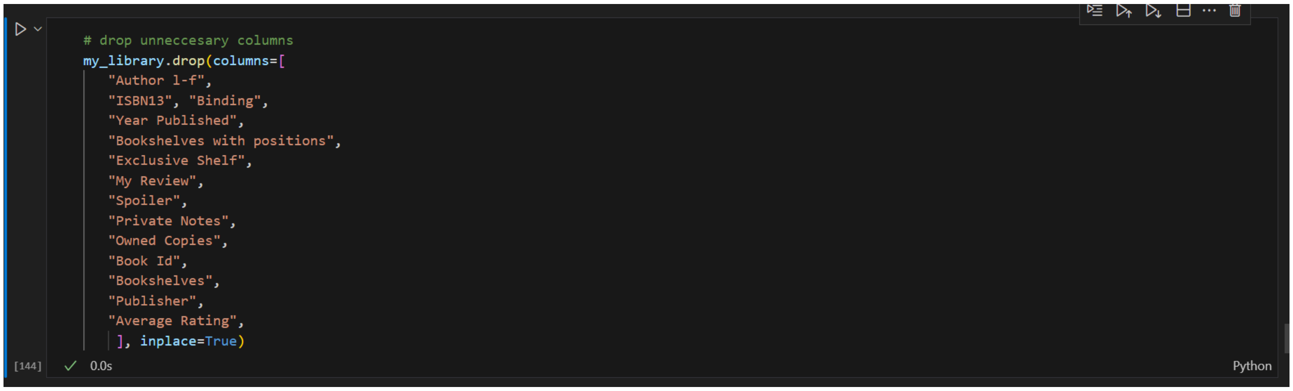


```
# gets the indices of the rows that does not satisfy the conditions
my_library.drop(my_library[(my_library["Bookshelves"].notna()) & (my_library["Date Read"].isna())].index, inplace=True)

my_library.info()
```

✓ 0.0s

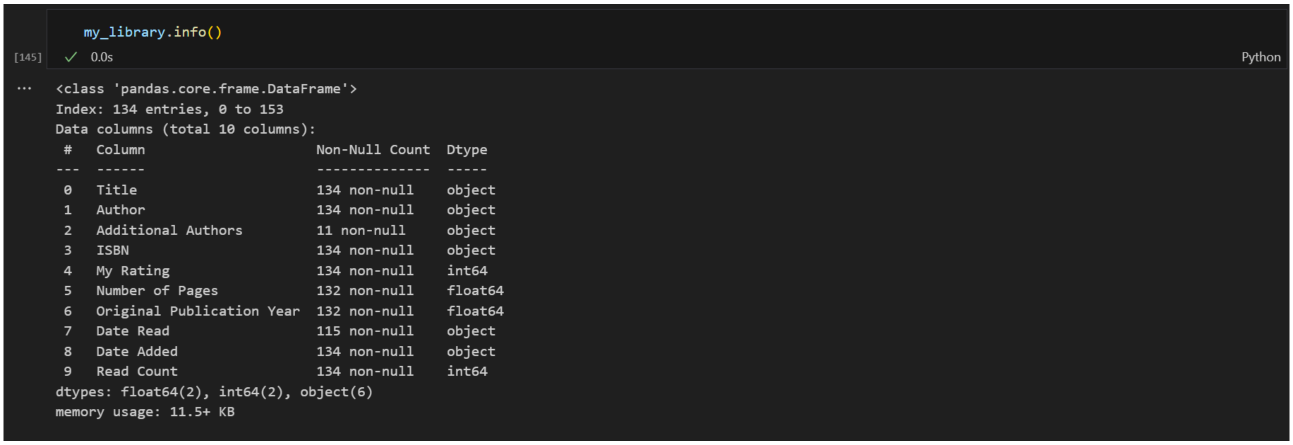
- Dropping unnecessary columns:



```
# drop unneccesary columns
my_library.drop(columns=[
    "Author l-f",
    "ISBN13", "Binding",
    "Year Published",
    "Bookshelves with positions",
    "Exclusive Shelf",
    "My Review",
    "Spoiler",
    "Private Notes",
    "Owned Copies",
    "Book Id",
    "Bookshelves",
    "Publisher",
    "Average Rating",
     ], inplace=True)
```

[144] ✓ 0.0s Python

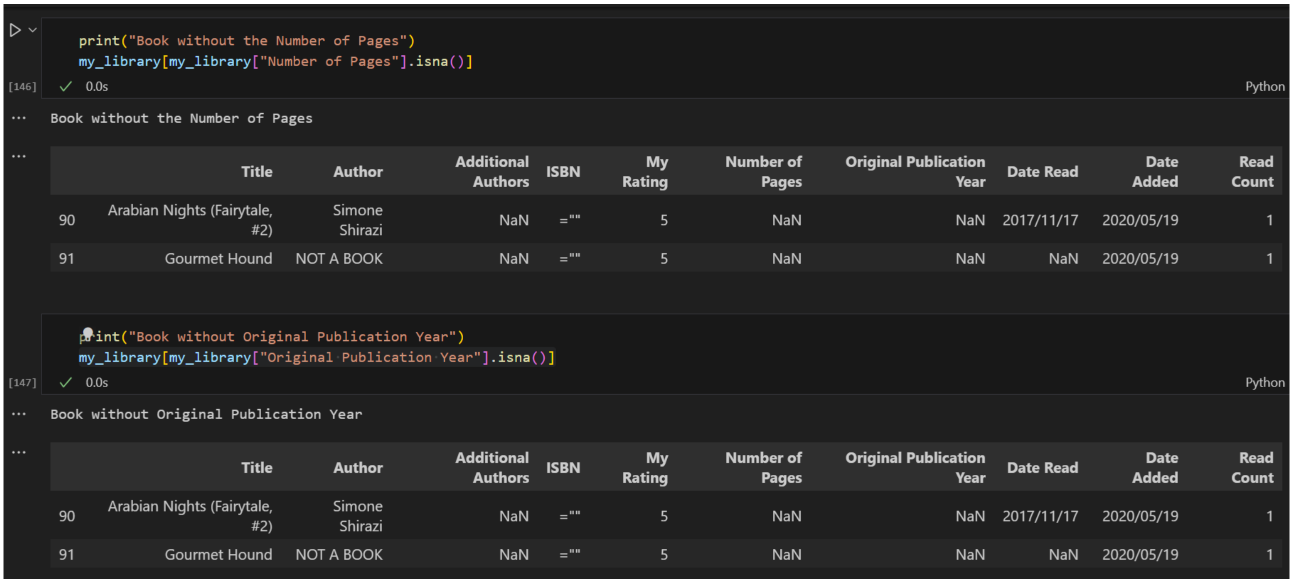
- Checking the new information about the dataset after initial dropping.



```
my_library.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 134 entries, 0 to 153
Data columns (total 10 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Title                      134 non-null    object
 1   Author                     134 non-null    object
 2   Additional Authors         11 non-null     object
 3   ISBN                       134 non-null    object
 4   My Rating                  134 non-null    int64
 5   Number of Pages            132 non-null    float64
 6   Original Publication Year  132 non-null    float64
 7   Date Read                  115 non-null    object
 8   Date Added                 134 non-null    object
 9   Read Count                 134 non-null    int64
dtypes: float64(2), int64(2), object(6)
memory usage: 11.5+ KB
```

- During data cleaning, many rows had null values. However, 'Additional Authors' and 'Date Read' columns will be kept even with null values. However, the number of pages and original publication year should be checked.



```
print("Book without the Number of Pages")
my_library[my_library["Number of Pages"].isna()]
```

```
Book without the Number of Pages
```

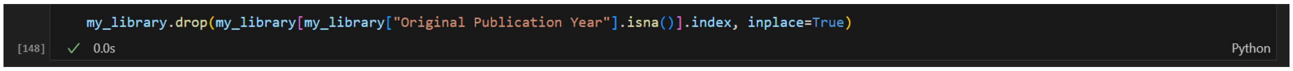
	Title	Author	Additional Authors	ISBN	My Rating	Number of Pages	Original Publication Year	Date Read	Date Added	Read Count
90	Arabian Nights (Fairytale, #2)	Simone Shirazi	NaN	=""	5	NaN	NaN	2017/11/17	2020/05/19	1
91	Gourmet Hound	NOT A BOOK	NaN	=""	5	NaN	NaN	NaN	2020/05/19	1

```
print("Book without Original Publication Year")
my_library[my_library["Original Publication Year"].isna()]
```

```
Book without Original Publication Year
```

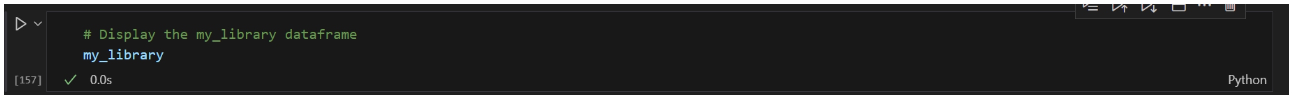
	Title	Author	Additional Authors	ISBN	My Rating	Number of Pages	Original Publication Year	Date Read	Date Added	Read Count
90	Arabian Nights (Fairytale, #2)	Simone Shirazi	NaN	=""	5	NaN	NaN	2017/11/17	2020/05/19	1
91	Gourmet Hound	NOT A BOOK	NaN	=""	5	NaN	NaN	NaN	2020/05/19	1

- Based on the filter, two observations that are displayed are not needed for analysis due to the nature of the observations (a web comic and a web novel) and therefore, will be dropped.



```
my_library.drop(my_library[my_library["Original Publication Year"].isna()].index, inplace=True)
```

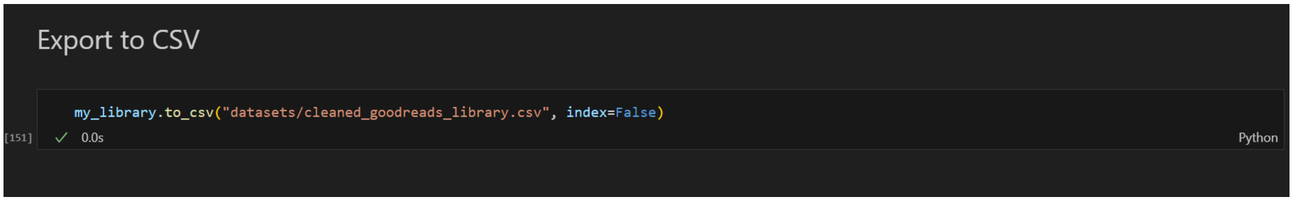
- Displaying the new final cleaned data,



```
# Display the my_library dataframe
my_library
```

This query resulted to 132 rows and 10 columns of read books data.

- Exporting cleaned data into a new CSV file.



Export to CSV

```
my_library.to_csv("datasets/cleaned_goodreads_library.csv", index=False)
```

Analyzing and Visualizing Data

As per available data, the following analysis is done:

Number of Books Read over the Years