Somia Tarek
22011639
health care

Import and loading dataset:

Dataset analysis.ipynb

Code

Open in...

Python 3 (ipykernel)

```
df.head()
```

[2]:

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication
0	1	2767052	2767052	2792775	272	439023483	9.780439e+12	Suzanne Collins	2
1	2	3	3	4640799	491	439554934	9.780440e+12	J.K. Rowling, Mary GrandPré	1
2	3	41865	41865	3212258	226	316015849	9.780316e+12	Stephenie Meyer	2
3	6	11870085	11870085	16827462	226	525478817	9.780525e+12	John Green	2
4	12	13335037	13335037	13155899	210	62024035	9.780062e+12	Veronica Roth	2

5 rows × 23 columns

Data cleaning :

```python
# Filter the dataset to include only the Harry Potter books
harry_potter_df = books_df[books_df['title'].str.contains('Harry Potter')]

# Find the most selling books within the Harry Potter series
most_selling_books = harry_potter_df.groupby('title')['ratings_count'].sum().sort_values(ascending=False)
print("Most selling Harry Potter books:")
print(most_selling_books.head())

# Calculate the average rating of the Harry Potter books
average_rating = harry_potter_df['average_rating'].mean()
print("Average rating of Harry Potter books:", average_rating)
```

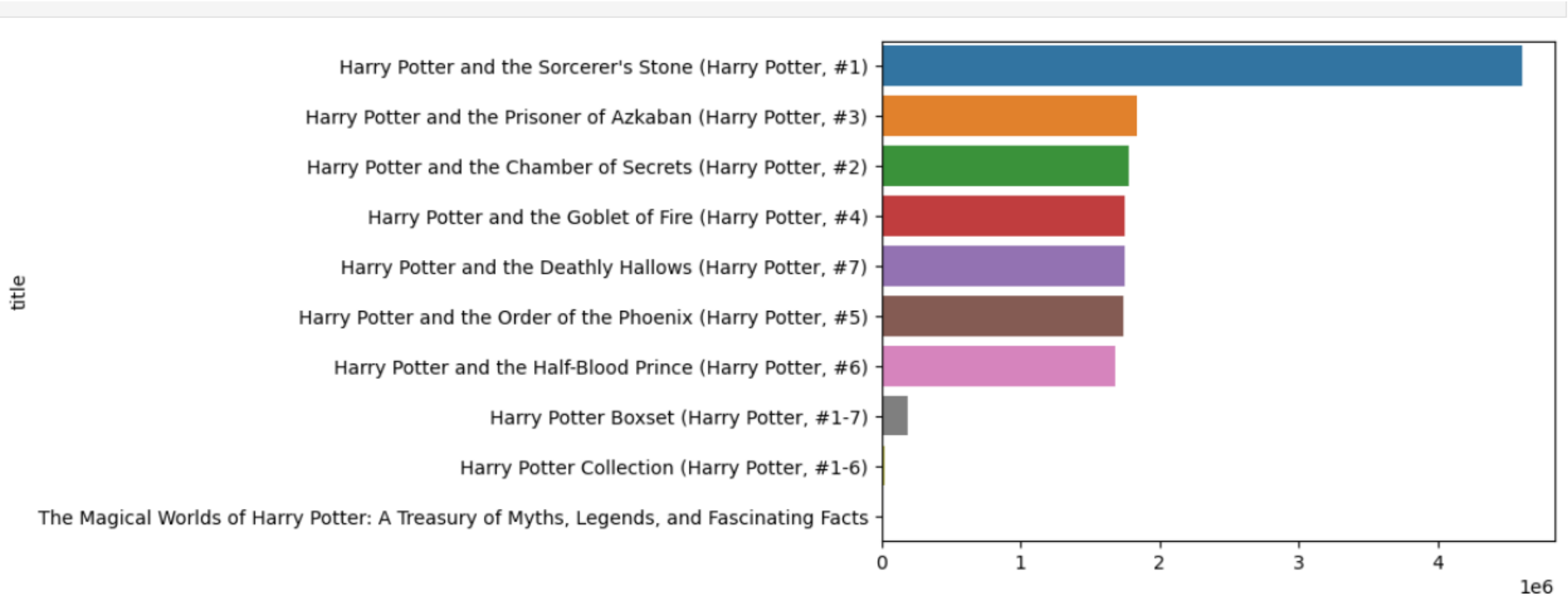
```
Most selling Harry Potter books:
title
Harry Potter and the Sorcerer's Stone (Harry Potter, #1)       4602479
Harry Potter and the Prisoner of Azkaban (Harry Potter, #3)    1832823
Harry Potter and the Chamber of Secrets (Harry Potter, #2)     1779331
Harry Potter and the Goblet of Fire (Harry Potter, #4)         1753043
Harry Potter and the Deathly Hallows (Harry Potter, #7)        1746574
Name: ratings_count, dtype: int64
Average rating of Harry Potter books: 4.482727272727273
```

```python
# Load the dataset into books_df
books_df = pd.read_csv('books.csv')
# Remove unnecessary columns
columns_to_drop = ['book_id', 'goodreads_book_id', 'best_book_id', 'work_id',
                   'isbn', 'isbn13', 'image_url', 'small_image_url']
books_df = books_df.drop(columns=columns_to_drop)

# Check for missing values
missing_values = books_df.isnull().sum()
print(missing_values)
# Handle missing values if any
```

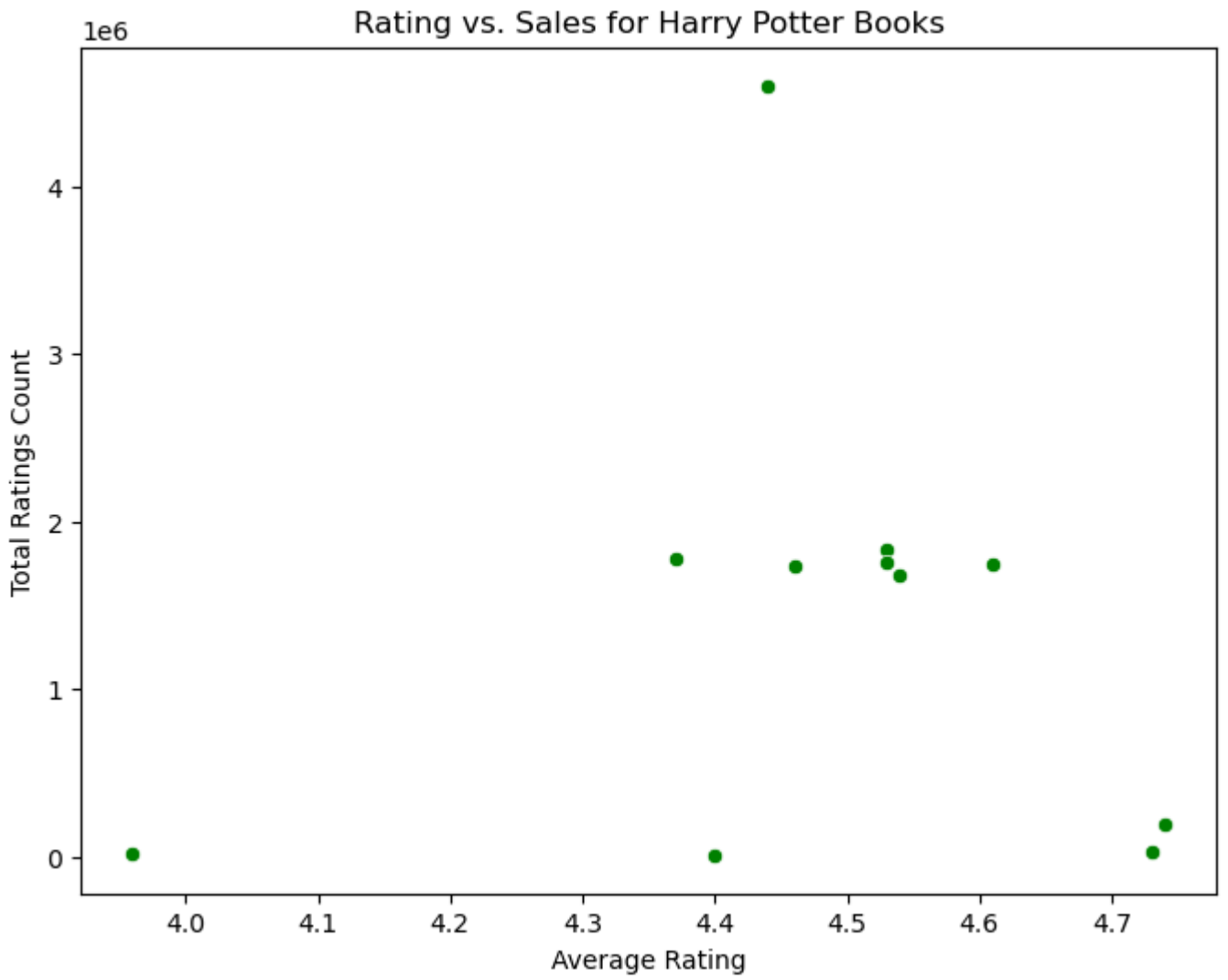
```
books_count                  0
authors                      0
original_publication_year    3
original_title              52
title                        0
language_code              109
average_rating               0
ratings_count                0
work_ratings_count           0
work_text_reviews_count      0
```

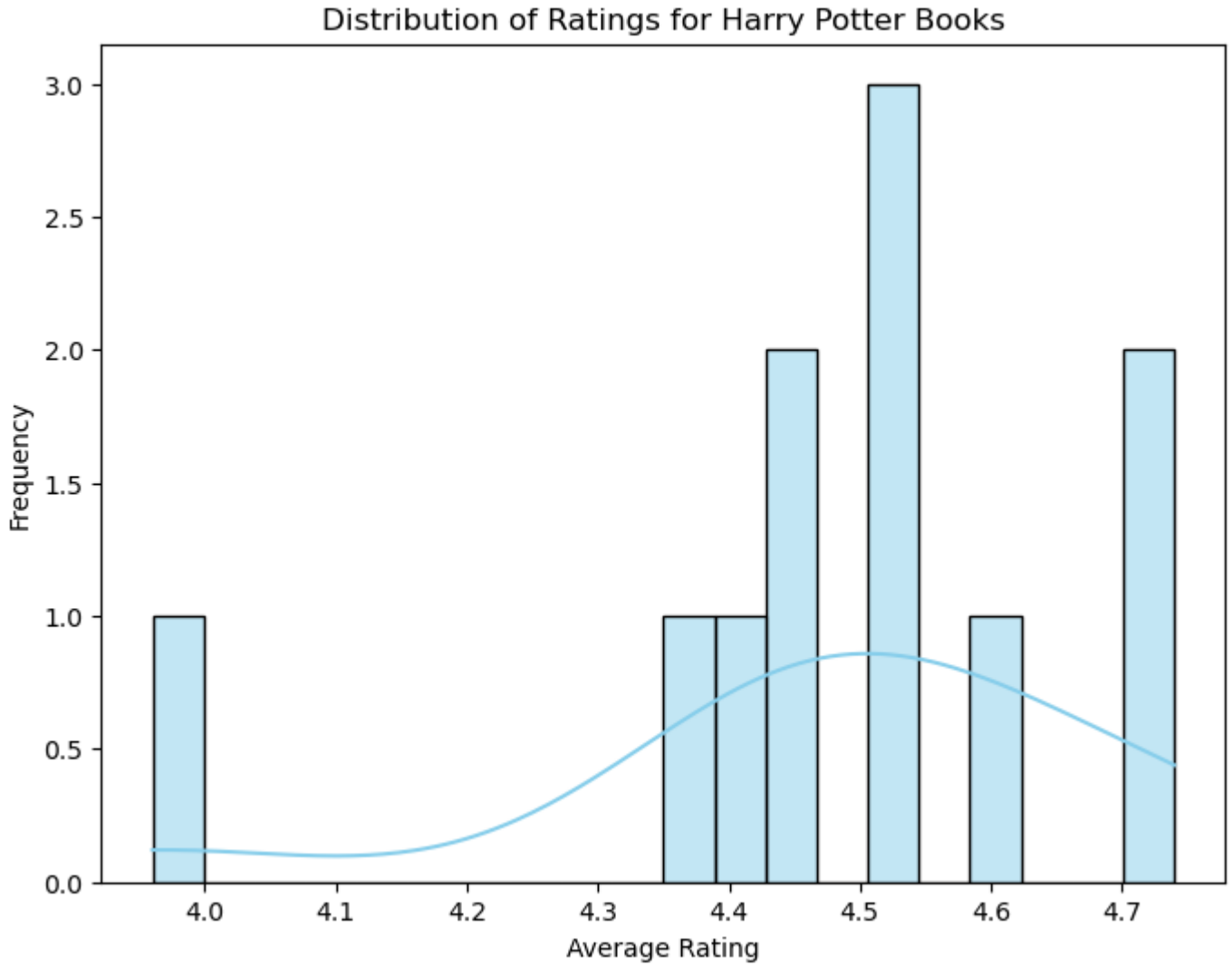
Data analysis of the rating count of Harry Potter Books:




```
<Figure size 1200x600 with 0 Axes>
```

Comparison between ratings and sales of Harry potter Books




Distribution of Ratings for Harry Potter Books
Frequency
Average Rating
0.0
0.5
1.0
1.5
2.0
2.5
3.0
4.0
4.1
4.2
4.3
4.4
4.5
4.6
4.7