Name : Loreen Mohamed Saeed

Department : AI – Third level

ID : 20221377356

Books analysis results

1. First, I did some data preprocessing to remove unnecessary columns like images as they won't affect our analysis, and also combined similar columns (ratings_1, ratings_2...) in one column (ratings) for easier analysis.

```
data = data.drop(columns=['image_url', 'small_image_url'])

data['ratings'] = data['ratings_1'] + data['ratings_2'] + data['ratings_3'] + data['ratings_4'] + data['ratings_5']
data = data.drop(columns=['ratings_1', 'ratings_2', 'ratings_3', 'ratings_4', 'ratings_5'])

data.head()
```

Python

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	title	language_co
0	1	2767052	2767052	2792775	272	439023483	9.780439e+12	Suzanne Collins	2008.0	The Hunger Games	The Hunger Games (The Hunger Games, #1)	e
1	2	3	3	4640799	491	439554934	9.780440e+12	J.K. Rowling, Mary GrandPré	1997.0	Harry Potter and the Philosopher's Stone	Harry Potter and the Sorcerer's Stone (Harry P...	e
2	3	41865	41865	3212258	226	316015849	9.780316e+12	Stephenie Meyer	2005.0	Twilight	Twilight (Twilight, #1)	en-l
3	6	11870085	11870085	16827462	226	525478817	9.780525e+12	John Green	2012.0	The Fault in Our Stars	The Fault in Our Stars	e
4	12	13335037	13335037	13155899	210	62024035	9.780062e+12	Veronica Roth	2011.0	Divergent	Divergent (Divergent, #1)	e

2. Then choose Harry Potter books as they are what we care about in our analysis.

```
subTitle = 'Harry Potter'
harryPotter_data = []
length = len(data['book_id'])

for i in range(length):
    name = data.iloc[i]['title']
    if subTitle in name:
        harryPotter_data.append(data.iloc[i])


harryPotter_data = pd.DataFrame(harryPotter_data)
harryPotter_data.head()
```

Python

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	title	language_co
1	2	3	3	4640799	491	439554934	9.780440e+12	J.K. Rowling, Mary GrandPré	1997.0	Harry Potter and the Philosopher's Stone	Harry Potter and the Sorcerer's Stone (Harry P...	e
6	18	5	5	2402163	376	043965548X	9.780440e+12	J.K. Rowling, Mary GrandPré, Rufus Beck	1999.0	Harry Potter and the Prisoner of Azkaban	Harry Potter and the Prisoner of Azkaban (Harr...	e
8	21	2	2	2809203	307	439358078	9.780439e+12	J.K. Rowling, Mary GrandPré	2003.0	Harry Potter and the Order of the Phoenix	Harry Potter and the Order of the Phoenix	e

3. I did further analysis on Harry Potter dataset as that is easier than if I did that on whole dataset. I checked for null values then replaced them with a value I found suitable for this null value. We had only one null value in “original_title” column

```
harryPotter_data.isna().sum()  # check for null values
```

```
book_id                    0
goodreads_book_id          0
best_book_id               0
work_id                    0
books_count                0
isbn                       0
isbn13                     0
authors                    0
original_publication_year  0
original_title             1
title                      0
language_code              0
average_rating             0
ratings_count              0
work_ratings_count         0
work_text_reviews_count    0
ratings                    0
dtype: int64
```

```
# I saw that null value only existed in original title column
# so thought best practise is to just replace nan value with title = "Harry Potter"
harryPotter_data = harryPotter_data.fillna(subTitle)
harryPotter_data
```

4. The result after removing the null values

```
harryPotter_data.isna().sum()   # removed null value
```

```
book_id                     0
goodreads_book_id           0
best_book_id                0
work_id                     0
books_count                 0
isbn                        0
isbn13                      0
authors                     0
original_publication_year   0
original_title              0
title                       0
language_code               0
average_rating              0
ratings_count               0
work_ratings_count          0
work_text_reviews_count     0
ratings                     0
dtype: int64
```

5. Here I did last analysis requirement by getting most sold Harry Potter books and the average rating of these books.

Most sold books:

```
harryPotter_data = harryPotter_data.sort_values(by=['books_count'], ascending=False)
most_sold = [name for name in harryPotter_data['title']]
print("Most sold books in Harry potter series are:", most_sold[0:3])
```

Python

```
Most sold books in Harry potter series are: ["Harry Potter and the Sorcerer's Stone (Harry Potter, #1)", 'Harry Potter and the Chamber of Secrets (Harry Potter,
```

Average rating:

```
summ = 0
length = len(harryPotter_data['isbn'])
for i in range(length):
    summ += harryPotter_data.iloc[i]['average_rating']

average_rating = round(summ / length, 3)
print("Average rating of the Harry Potter books =", average_rating)
```

Python

```
Average rating of the Harry Potter books = 4.483
```