In [74]:

```
import pandas as pd
import numpy as np
import seaborn as sns

df = pd.read_csv("F:/COMPANY/VYAKYA/Technical Assessment/task1/books.csv", error_bad_lines=

df.head()
```

```
b'Skipping line 3350: expected 12 fields, saw 13\nSkipping line 4704: expect
ed 12 fields, saw 13\nSkipping line 5879: expected 12 fields, saw 13\nSkippi
ng line 8981: expected 12 fields, saw 13\n'
```

Out[74]:

	bookID	title	authors	average_rating	isbn	isbn13	language_code
0	1	Harry Potter and the Half-Blood Prince (Harry ...	J.K. Rowling/Mary GrandPré	4.57	0439785960	9780439785969	eng
1	2	Harry Potter and the Order of the Phoenix (Har...	J.K. Rowling/Mary GrandPré	4.49	0439358078	9780439358071	eng
2	4	Harry Potter and the Chamber of Secrets (Harry...	J.K. Rowling	4.42	0439554896	9780439554893	eng
3	5	Harry Potter and the Prisoner of Azkaban (Harr...	J.K. Rowling/Mary GrandPré	4.56	043965548X	9780439655484	eng
4	8	Harry Potter Boxed Set Books 1-5 (Harry Potte...	J.K. Rowling/Mary GrandPré	4.78	0439682584	9780439682589	eng

In [76]:

```
df.tail()
```

Out[76]:

	bookID	title	authors	average_rating	isbn	isbn13	langua
11118	45631	Expelled from Eden: A William T. Vollmann Reader	William T. Vollmann/Larry McCaffery/Michael He...	4.06	1560254416	9781560254416	
11119	45633	You Bright and Risen Angels	William T. Vollmann	4.08	0140110879	9780140110876	
11120	45634	The Ice-Shirt (Seven Dreams #1)	William T. Vollmann	3.96	0140131965	9780140131963	
11121	45639	Poor People	William T. Vollmann	3.72	0060878827	9780060878825	
11122	45641	Las aventuras de Tom Sawyer	Mark Twain	3.91	8497646983	9788497646987	

In [78]:

```
df.shape
```

Out[78]:

```
(11123, 12)
```

In [80]:

```
df.describe()
```

Out[80]:

	bookID	average_rating	isbn13	num_pages	ratings_count	text_reviews_cc
count	11123.000000	11123.000000	1.112300e+04	11123.000000	1.112300e+04	11123.000
mean	21310.856963	3.934075	9.759880e+12	336.405556	1.794285e+04	542.048
std	13094.727252	0.350485	4.429758e+11	241.152626	1.124992e+05	2576.619
min	1.000000	0.000000	8.987060e+09	0.000000	0.000000e+00	0.000
25%	10277.500000	3.770000	9.780345e+12	192.000000	1.040000e+02	9.000
50%	20287.000000	3.960000	9.780582e+12	299.000000	7.450000e+02	47.000
75%	32104.500000	4.140000	9.780872e+12	416.000000	5.000500e+03	238.000
max	45641.000000	5.000000	9.790008e+12	6576.000000	4.597666e+06	94265.000

In [81]:

```
df.columns
```

Out[81]:

```
Index(['bookID', 'title', 'authors', 'average_rating', 'isbn', 'isbn13',
       'language_code', '  num_pages', 'ratings_count', 'text_reviews_coun
t',
       'publication_date', 'publisher'],
      dtype='object')
```

In [82]:

```
df.nunique()
```

Out[82]:

```
bookID               11123
title                10348
authors               6639
average_rating         209
isbn                 11123
isbn13               11123
language_code           27
  num_pages            997
ratings_count         5294
text_reviews_count    1822
publication_date      3679
publisher             2290
dtype: int64
```

In [85]:

```
df['title'].unique()
```

Out[85]:

```
array(['Harry Potter and the Half-Blood Prince (Harry Potter  #6)',
       'Harry Potter and the Order of the Phoenix (Harry Potter  #5)',
       'Harry Potter and the Chamber of Secrets (Harry Potter  #2)', ...,
       'The Ice-Shirt (Seven Dreams #1)', 'Poor People',
       'Las aventuras de Tom Sawyer'], dtype=object)
```

In [ ]:

```
#cleaning the dataset
```

In [86]:

```
df.isnull().sum()
```

Out[86]:

```
bookID                0
title                 0
authors               0
average_rating        0
isbn                  0
isbn13                0
language_code         0
  num_pages           0
ratings_count         0
text_reviews_count    0
publication_date      0
publisher             0
dtype: int64
```

In [ ]:

```
#relationship analysis
```

In [87]:

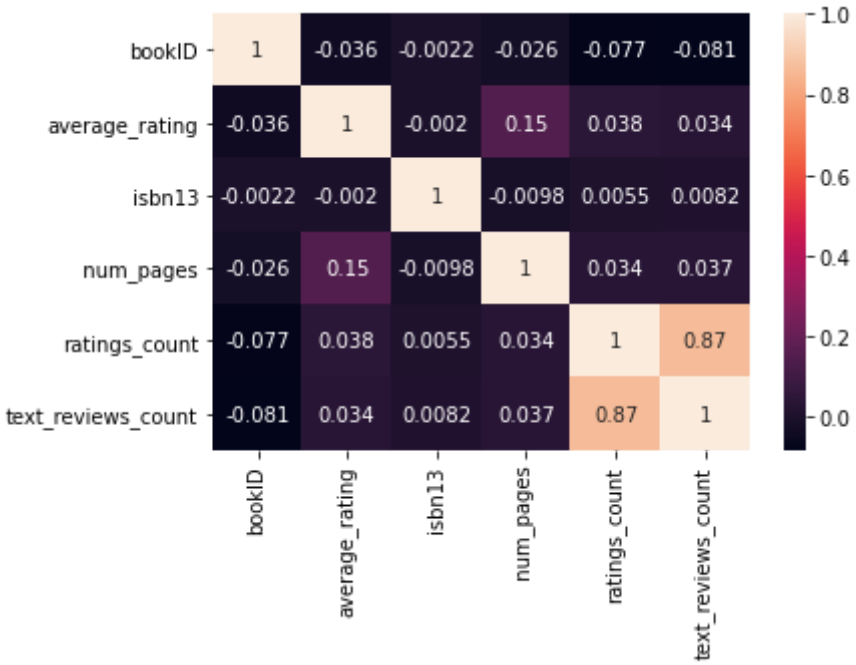
```
corelation = df.corr()
```

In [88]:

```
sns.heatmap(corelation, xticklabels=corelation.columns, yticklabels=corelation.columns, ann
```

Out[88]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x2558350c9a0>
```

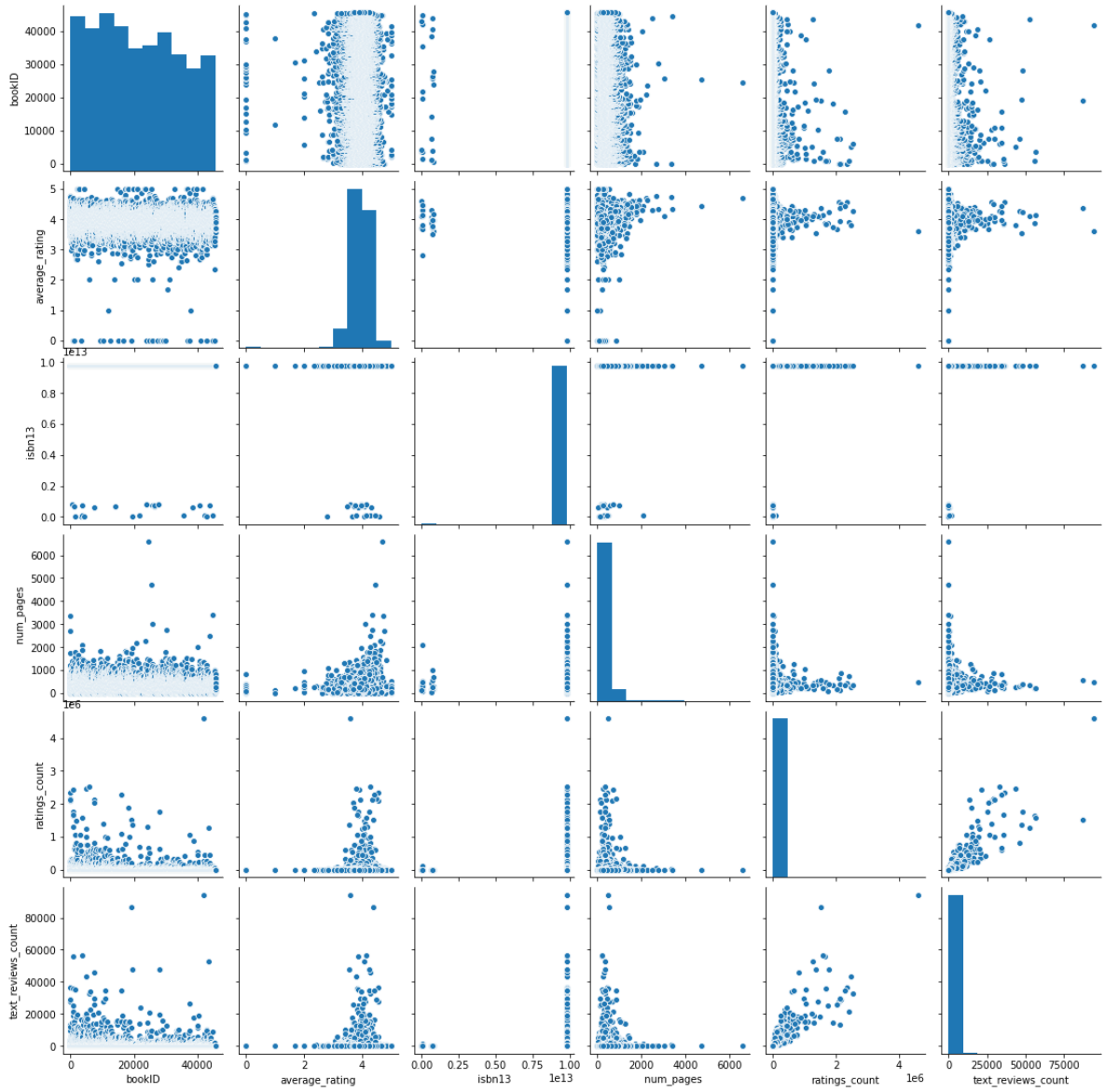


In [89]:

```
sns.pairplot(df)
```

Out[89]:

```
<seaborn.axisgrid.PairGrid at 0x255839dba00>
```

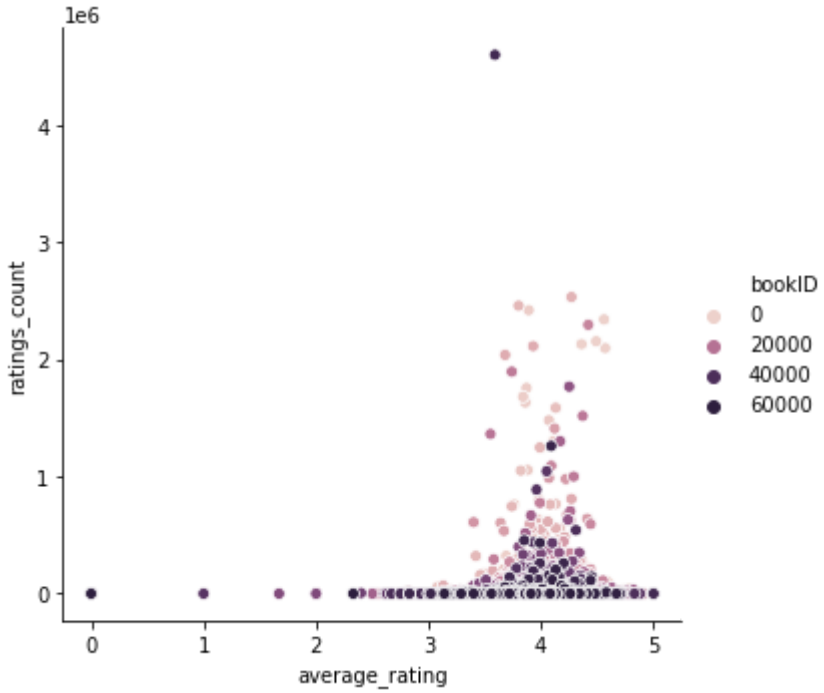


In [90]:

```
sns.relplot(x= 'average_rating', y='ratings_count', hue='bookID', data=df)
```

Out[90]:

```
<seaborn.axisgrid.FacetGrid at 0x25586117f40>
```

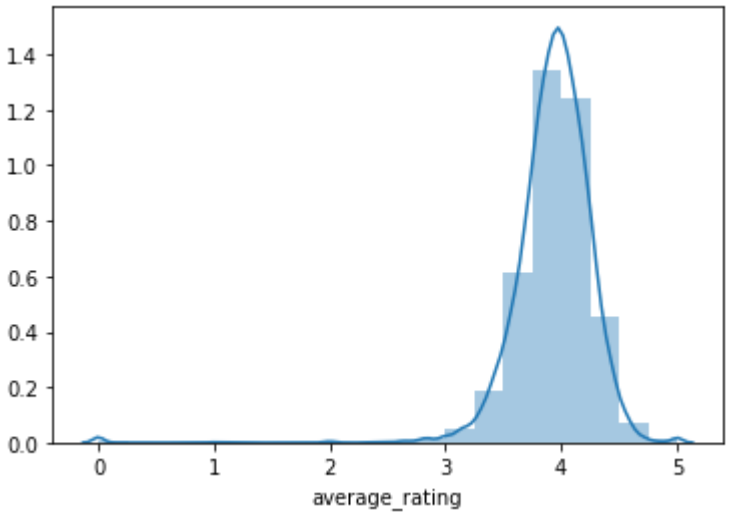


In [91]:

```
sns.distplot(df['average_rating'], bins=20)
```

Out[91]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x255861ae250>
```

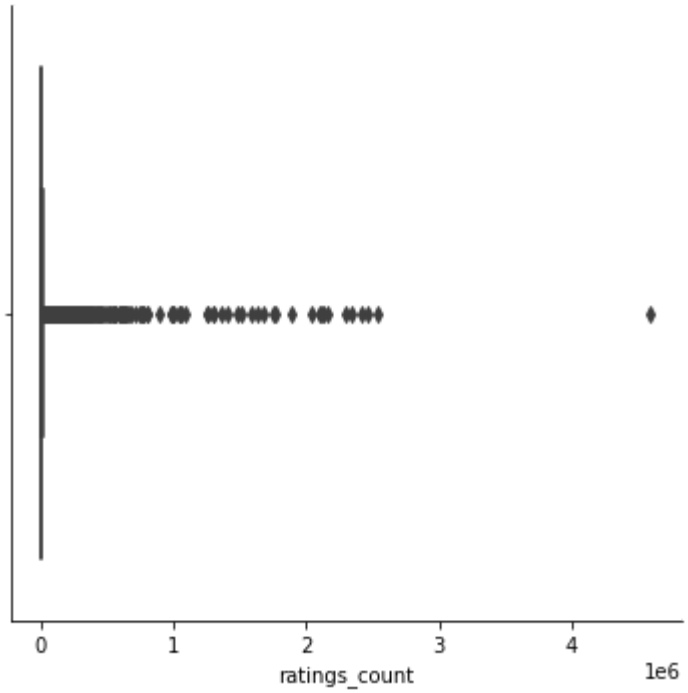


In [95]:

```
sns.catplot(x='ratings_count', kind='box', data=df)
```

Out[95]:

```
<seaborn.axisgrid.FacetGrid at 0x2558691ec10>
```

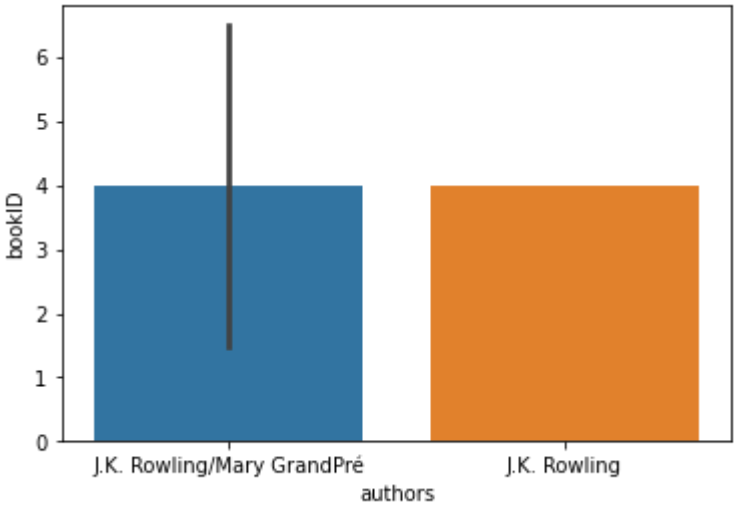


In [ ]:

```
```

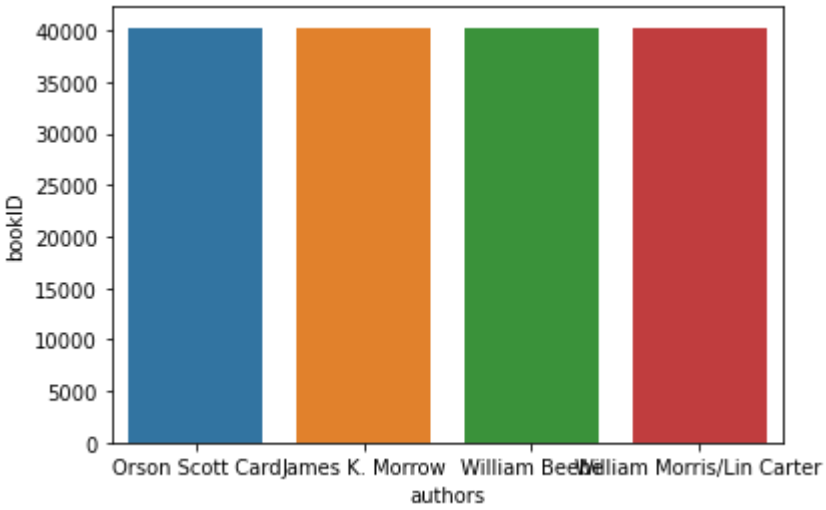
In [116]:

```
from matplotlib import pyplot as plt
sns.barplot(x="authors", y="bookID", data=df[0:5])
plt.show()
```



In [127]:

```
sns.barplot(x="authors", y="bookID", data=df[9999:10004])
plt.show()
```

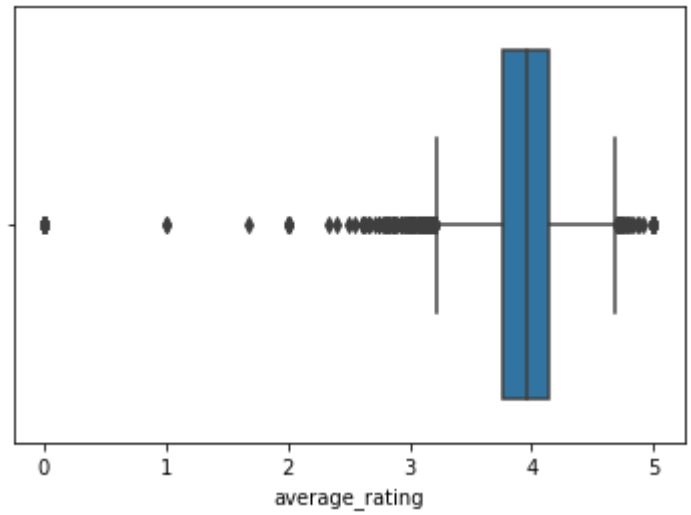


In [132]:

```
sns.boxplot(x=df["average_rating"])
```

Out[132]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x255f9684c10>
```



In [143]:

```
df.groupby('bookID')['authors'].value_counts()
```

Out[143]:

```
bookID  authors
1       J.K. Rowling/Mary GrandPré                                  1
2       J.K. Rowling/Mary GrandPré                                  1
4       J.K. Rowling                                                1
5       J.K. Rowling/Mary GrandPré                                  1
8       J.K. Rowling/Mary GrandPré                                  1
                                                                   ..
45631   William T. Vollmann/Larry McCaffery/Michael Hemmingson      1
45633   William T. Vollmann                                         1
45634   William T. Vollmann                                         1
45639   William T. Vollmann                                         1
45641   Mark Twain                                                  1
Name: authors, Length: 11123, dtype: int64
```

In [ ]:

```

```