Book recommender system

In [9]:
```
import numpy as np
import pandas as pd
```

In [10]:
```
#importing data
#to ksip lines with errors, "error bad lines" is used.
df = pd.read_csv(r"C:\RESEARCH\books.csv", error_bad_lines=False)
```

```
C:\Users\PRASANTA\AppData\Local\Temp\ipykernel_15968\1557645421.py:3: FutureWarning: The error_bad_lines argument has been
deprecated and will be removed in a future version. Use on_bad_lines in the future.


  df = pd.read_csv(r"C:\RESEARCH\books.csv", error_bad_lines=False)
b'Skipping line 3350: expected 12 fields, saw 13\nSkipping line 4704: expected 12 fields, saw 13\nSkipping line 5879: expec
ted 12 fields, saw 13\nSkipping line 8981: expected 12 fields, saw 13\n'
```

In [11]:
```
df.head()
```

Out[11]:

	title	authors	average_rating	isbn	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	publi
0	Harry Potter and the Half-Blood Prince (Harry ...	J.K. Rowling/Mary GrandPré	4.57	0439785960	9780439785969	eng	652	2095690	27591	9/16/2006	Schol
1	Harry Potter and the Order of the Phoenix (Har...	J.K. Rowling/Mary GrandPré	4.49	0439358078	9780439358071	eng	870	2153167	29221	9/1/2004	Schol
2	Harry Potter and the Chamber of Secrets (Harry...	J.K. Rowling	4.42	0439554896	9780439554893	eng	352	6333	244	11/1/2003	Schol
3	Harry Potter and the Prisoner of Azkaban (Harr...	J.K. Rowling/Mary GrandPré	4.56	043965548X	9780439655484	eng	435	2339585	36325	5/1/2004	Schol
4	Harry Potter Boxed Set Books 1-5 (Harry Potte...	J.K. Rowling/Mary GrandPré	4.78	0439682584	9780439682589	eng	2690	41428	164	9/13/2004	Schol

In [12]:
```
df.tail()
```

Out[12]:

	title	authors	average_rating	isbn	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	publi
11122	xpelled from :den: A lliam T. ıllmann Reader	William T. Vollmann/Larry McCaffery/Michael He...	4.06	1560254416	9781560254416	eng	512	156	20	12/21/2004	Da (F
11123	You Bright and Risen Angels	William T. Vollmann	4.08	0140110879	9780140110876	eng	635	783	56	12/1/1988	Per B
11124	he Ice-Shirt (Seven)reams #1)	William T. Vollmann	3.96	0140131965	9780140131963	eng	415	820	95	8/1/1993	Per B
11125	Poor People	William T. Vollmann	3.72	0060878827	9780060878825	eng	434	769	139	2/27/2007	
11126	Las enturas de Tom Sawyer	Mark Twain	3.91	8497646983	9788497646987	spa	272	113	12	5/28/2006	Ec L

In [13]:
```
df.isnull().sum()
```

Out[13]:
```
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 [14]:
```
df.shape
```

Out[14]:
```
(11123, 12)
```

In [15]:
```
df.duplicated().sum()
```

Out[15]:
```
0
```

In [16]:
```
df['authors'].unique()
```

Out[16]:
```
array(['J.K. Rowling/Mary GrandPré', 'J.K. Rowling',
       'W. Frederick Zimmerman', ..., 'C.S. Lewis/Ana Falcão Bastos',
       'C.S. Lewis/Pauline Baynes/Ana Falcão Bastos',
       'William T. Vollmann/Larry McCaffery/Michael Hemmingson'],
      dtype=object)
```

In [18]:
```
# as j.k rowling appears with another author we have to replace it
df.replace(to_replace='J.K. Rowling/Mary GrandPré',value='J.K. Rowling',inplace=True)
```

In [20]:
```
df['authors'].value_counts()
```

Out[20]:
```
Stephen King                                          40
P.G. Wodehouse                                        40
Rumiko Takahashi                                      39
Orson Scott Card                                      35
Agatha Christie                                       33
                                                      ..
Ian Glasper                                            1
Legs McNeil/Gillian McCain                             1
Adam Woog                                              1
Mikal Gilmore                                          1
William T. Vollmann/Larry McCaffery/Michael Hemmingson 1
Name: authors, Length: 6638, dtype: int64
```

In [22]:
```
df['title'].duplicated().sum()
```

Out[22]:
```
775
```

In [23]:
```
#droping irrelevant columns
df.drop(['bookID', 'isbn', 'isbn13'],axis=1,inplace=True)
```

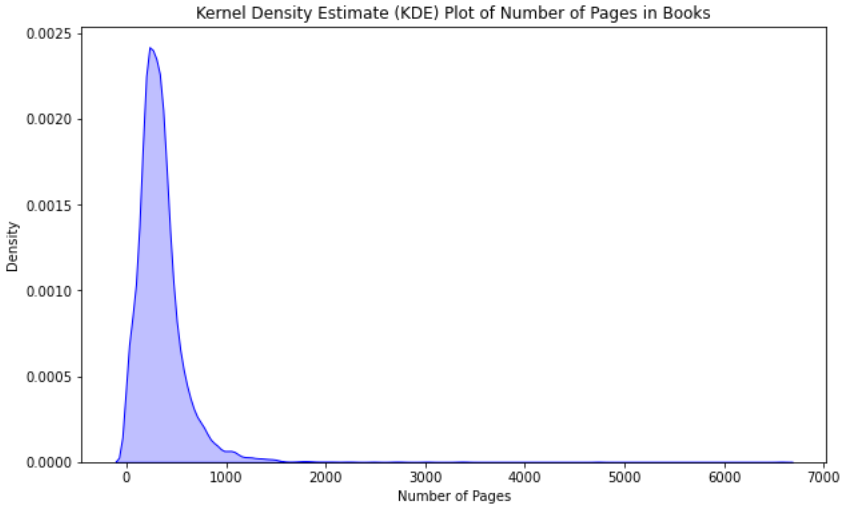
In [24]:
```
df.columns
```

Out[24]:
```
Index(['title', 'authors', 'average_rating', 'language_code', '  num_pages',
       'ratings_count', 'text_reviews_count', 'publication_date', 'publisher'],
      dtype='object')
```

EDA

In [28]:
```
df.rename(columns={'  num_pages':'Total_page'},inplace=True)
```

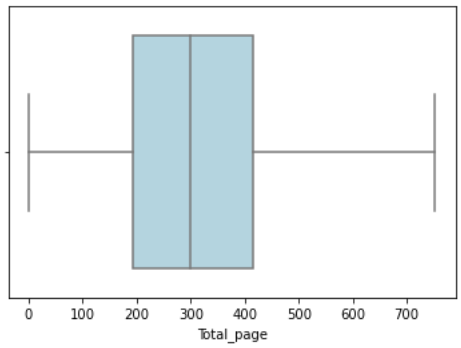
In [32]:
```
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

# Create KDE plot
plt.figure(figsize=(10, 6))
sns.kdeplot(df['Total_page'].dropna(), color='blue', fill=True)

# Adding labels and title
plt.xlabel('Number of Pages')
plt.ylabel('Density')
plt.title('Kernel Density Estimate (KDE) Plot of Number of Pages in Books')

# Display the plot
plt.show()
```



In [33]:
```
sns.boxplot("Total_page", data=df,
                palette=["lightblue"],sym='')
```

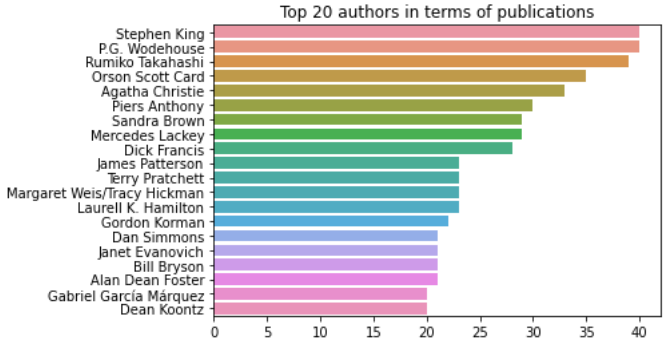
```
C:\Users\PRASANTA\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a ke
yword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an
explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[33]:
```
<AxesSubplot:xlabel='Total_page'>
```



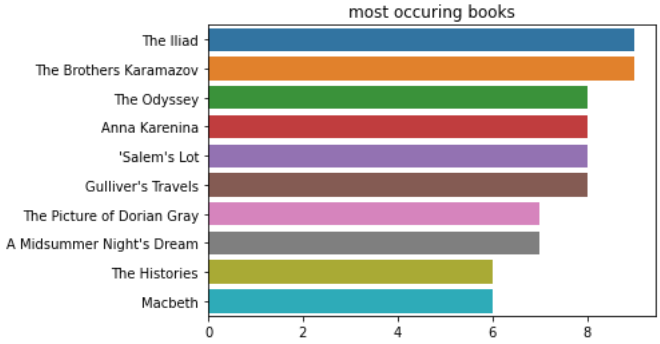
In [37]:
```
#Top Authors based on number of books
top_20_author=df['authors'].value_counts()[:20]

sns.barplot(top_20_author.values, top_20_author.index, alpha=1.).set_title('Top 20 authors in terms of publications')

plt.show()
```

```
C:\Users\PRASANTA\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as key
word args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without
an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



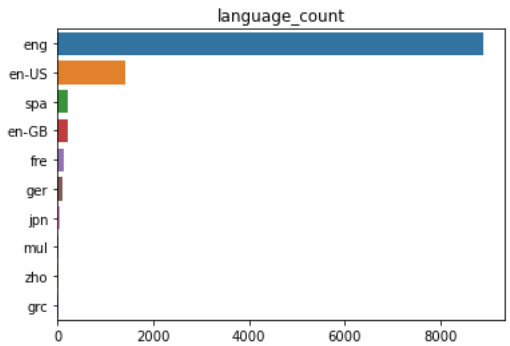
In [40]:
```
#Most occuring books
most_occuring_books=df['title'].value_counts()[:10]
sns.barplot(most_occuring_books.values,most_occuring_books.index,alpha=1.).set_title('most occuring books')
plt.show()
```

```
C:\Users\PRASANTA\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as key
word args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without
an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



In [43]:
```
# No of books in different language
language_count=df['language_code'].value_counts()[:10]
sns.barplot(language_count.values,language_count.index,alpha=1).set_title("language_count")
plt.show()
```

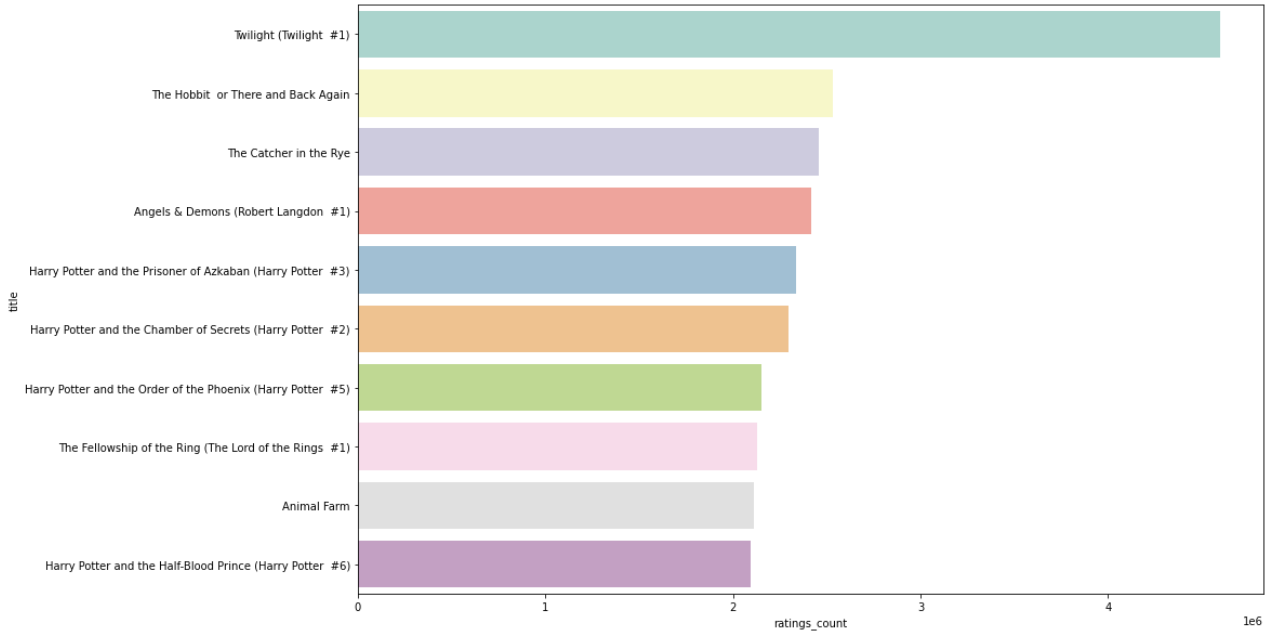
```
C:\Users\PRASANTA\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as key
word args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without
an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



In [46]:
```
#Top books on rating counts
most_rated = df.sort_values('ratings_count', ascending = False).head(10).set_index('title')
plt.figure(figsize=(15,10))
sns.barplot(most_rated['ratings_count'], most_rated.index, alpha=.8,palette='Set3')
```

```
C:\Users\PRASANTA\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as key
word args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without
an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

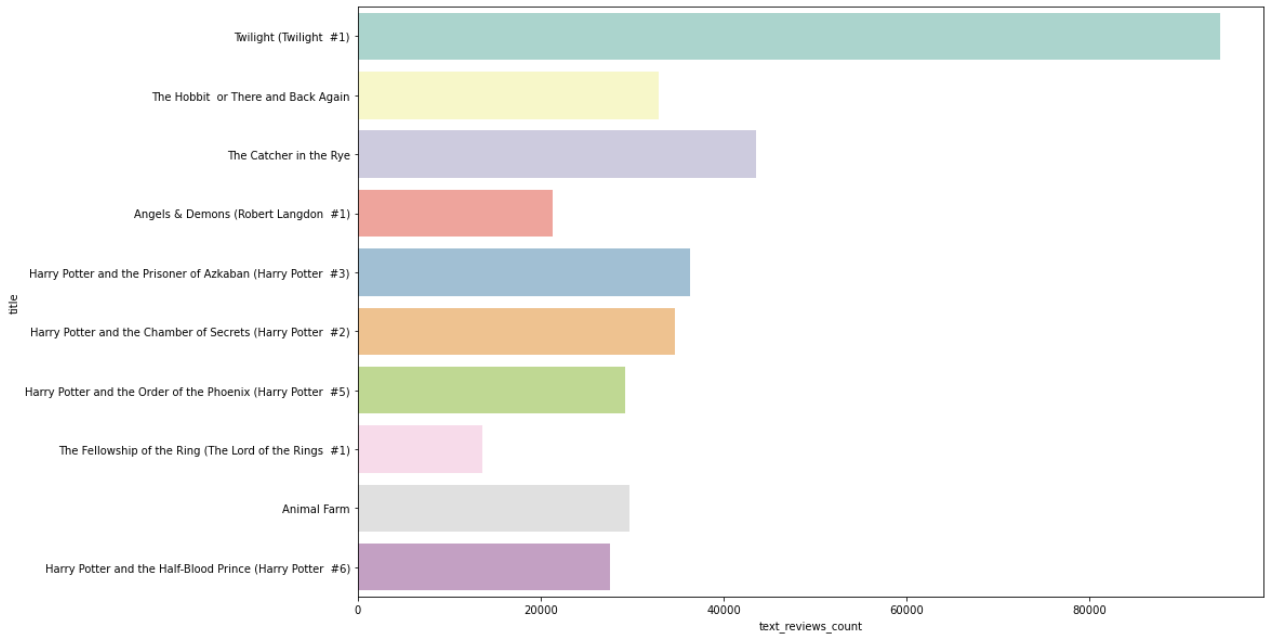
Out[46]:
```
<AxesSubplot:xlabel='ratings_count', ylabel='title'>
```



In [48]:
```
#Top books on text reviews
most_reviews = df.sort_values('text_reviews_count', ascending = False).head(10).set_index('title')
plt.figure(figsize=(15,10))
sns.barplot(most_rated['text_reviews_count'], most_rated.index, alpha=.8,palette='Set3')
```

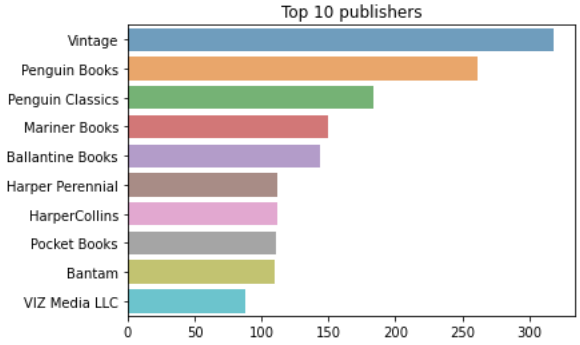
```
C:\Users\PRASANTA\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as key
word args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without
an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[48]:
```
<AxesSubplot:xlabel='text_reviews_count', ylabel='title'>
```



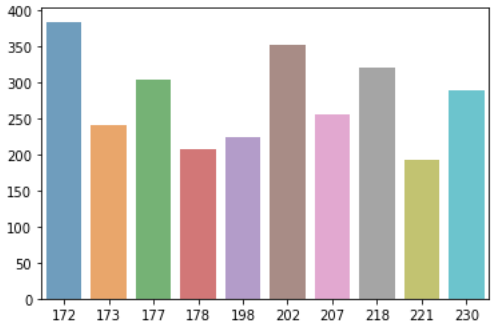
In [49]:
```
#Top 10 publishers
publisher=df['publisher'].value_counts()[:10]

sns.barplot(publisher.values, publisher.index, alpha=.7).set_title('Top 10 publishers')

plt.show()
```

```
C:\Users\PRASANTA\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as key
word args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without
an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```



In [50]:
```
#10 books on Total pages
pages=df['Total_page'].value_counts()[:10]

sns.barplot(pages.values, pages.index, alpha=.7).set_title(' ')

plt.show()
```

```
C:\Users\PRASANTA\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as key
word args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without
an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

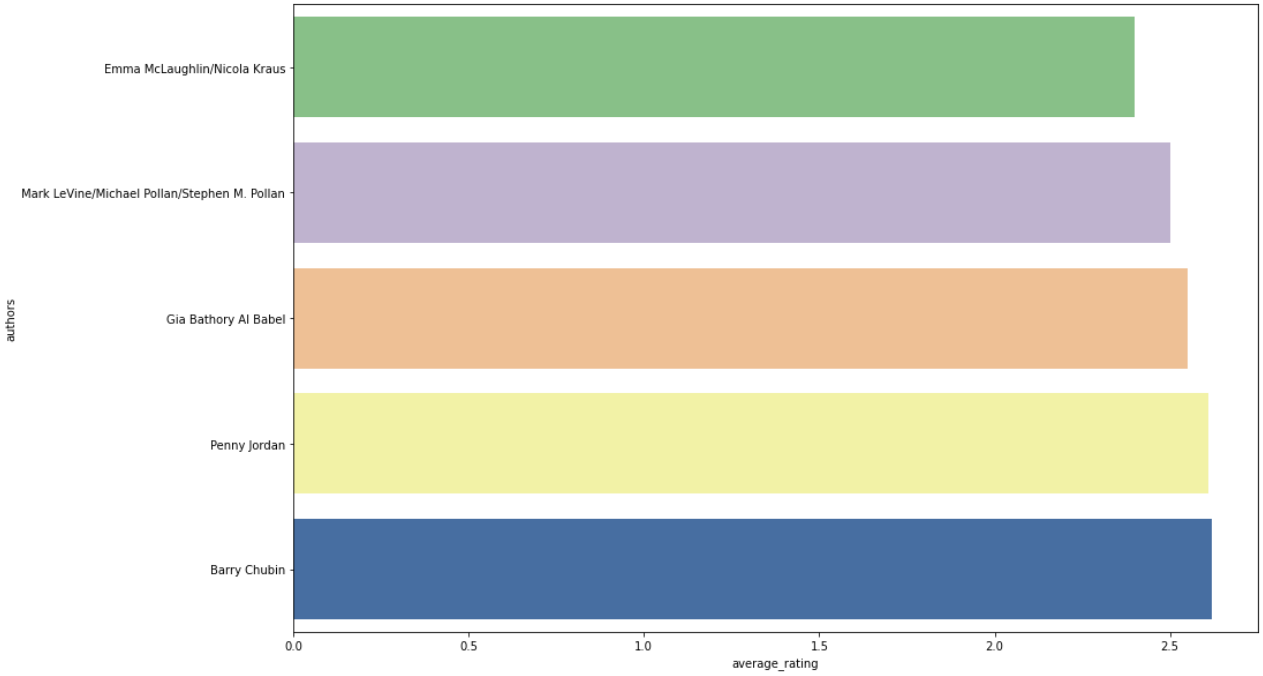


In [58]:
```
# Relationship between authors and averating rating on text review count and rating count where threshold for rating count is
```

In [59]:
```
rev=df['text_reviews_count']
ratings=df['ratings_count']>3
dff=pd.DataFrame(df[rev & ratings].sort_values('average_rating', ascending = True).head(5))

plt.figure(figsize=(15,10))
sns.barplot(y=dff['authors'], x=dff['average_rating'], palette='Accent')
dff.head()
```

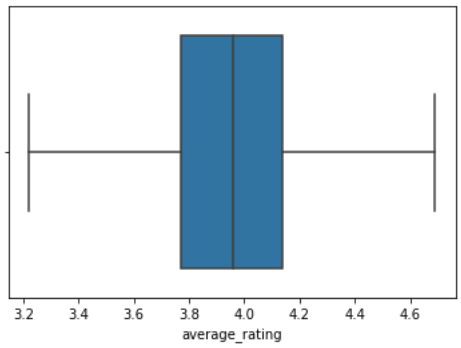
Out[59]:

	title	authors	average_rating	language_code	Total_page	ratings_count	text_reviews_count	publication_date	publisher
8812	Citizen Girl	Emma McLaughlin/Nicola Kraus	2.40	en-US	305	5415	577	10/4/2005	Washington Square Press
6502	Field Guide to Home Buying in America	Mark LeVine/Michael Pollan/Stephen M. Pollan	2.50	eng	256	4	1	5/15/1988	Touchstone
4452	The Trouble with the Pears: An Intimate Portra...	Gia Bathory Al Babel	2.55	eng	272	33	7	2/14/2006	Authorhouse
8965	A Matter of Trust	Penny Jordan	2.61	eng	224	36	5	12/23/1994	Harlequin Presents
8953	13th Directorate	Barry Chubin	2.62	eng	0	8	1	2/28/1989	Ivy Books



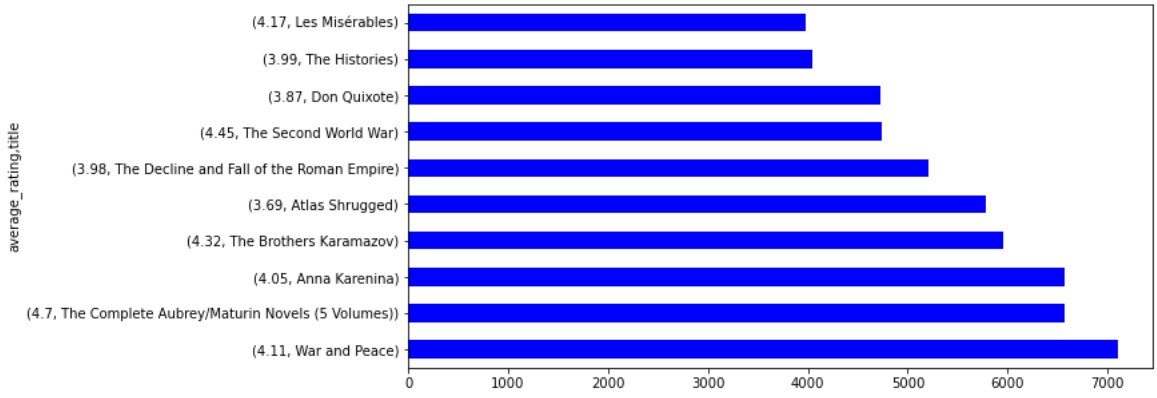
In [52]:
```
sns.boxplot(x=df['average_rating'],sym='')
```

Out[52]: <AxesSubplot:xlabel='average_rating'>



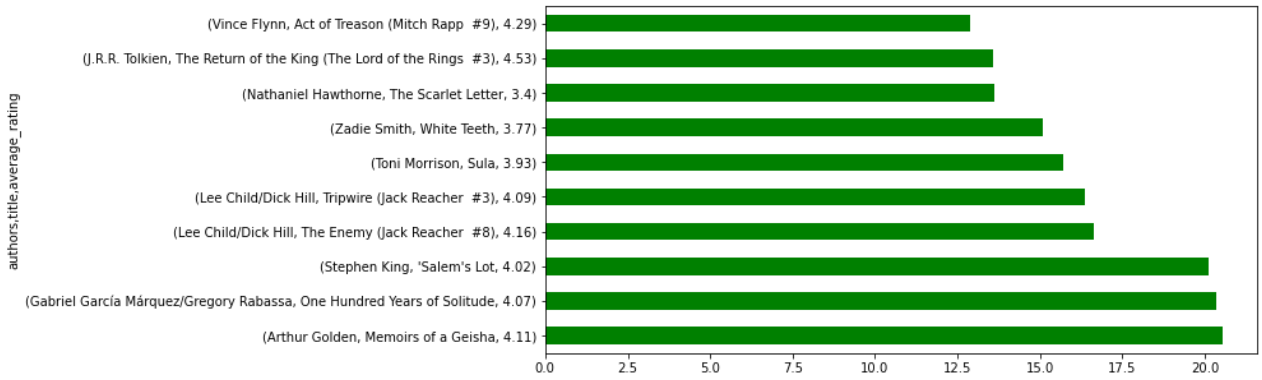
In [53]:
```
plt.figure(figsize=(10,5))

df.groupby(['average_rating','title']).Total_page.sum().nlargest(10).plot(kind='barh',color='b')
```

Out[53]: <AxesSubplot:ylabel='average_rating,title'>



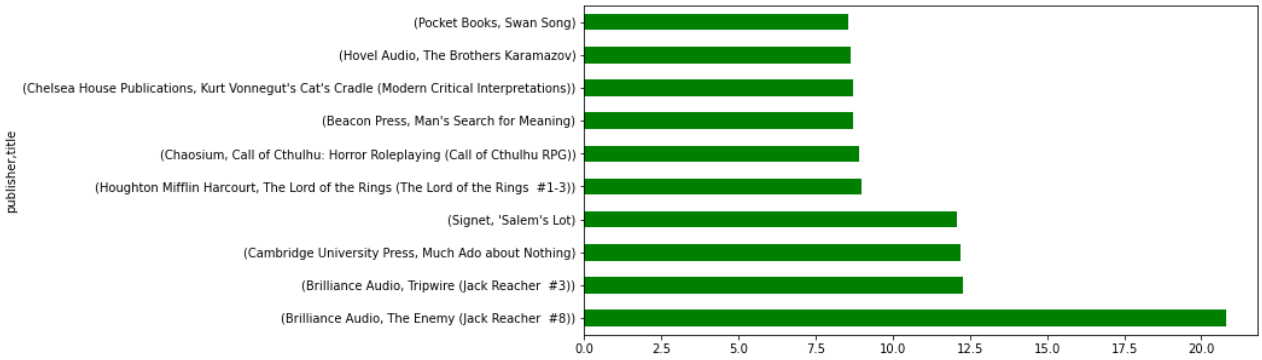
In [54]:
```
plt.figure(figsize=(10,5))

df.groupby(['authors','title','average_rating']).average_rating.sum().nlargest(10).plot(kind='barh',color='g')
```

Out[54]: <AxesSubplot:ylabel='authors,title,average_rating'>



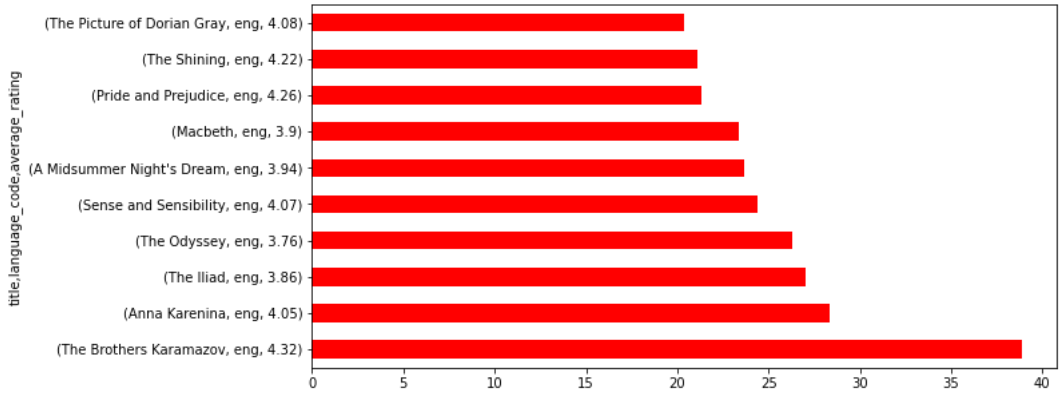
In [55]:
```
plt.figure(figsize=(10,5))

df.groupby(['publisher','title']).average_rating.sum().nlargest(10).plot(kind='barh',color='g')
```

Out[55]: <AxesSubplot:ylabel='publisher,title'>

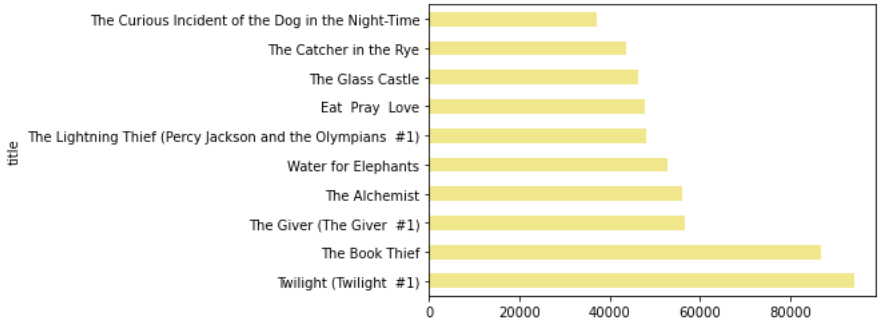


In [56]:
```
plt.figure(figsize=(10,5))

df.groupby(['title','language_code','average_rating']).average_rating.sum().nlargest(10).plot(kind='barh',color='r')
```

Out[56]: <AxesSubplot:ylabel='title,language_code,average_rating'>



In [57]:
```
ost_reviews=df.groupby('title')['text_reviews_count'].sum().sort_values(ascending=False).head(10).plot(kind='barh',color='kha
```



In [60]:
```
df.head()
```

Out[60]:

	title	authors	average_rating	language_code	Total_page	ratings_count	text_reviews_count	publication_date	publisher
0	Harry Potter and the Half-Blood Prince (Harry ...	J.K. Rowling	4.57	eng	652	2095690	27591	9/16/2006	Scholastic Inc.
1	Harry Potter and the Order of the Phoenix (Har...	J.K. Rowling	4.49	eng	870	2153167	29221	9/1/2004	Scholastic Inc.
2	Harry Potter and the Chamber of Secrets (Harry...	J.K. Rowling	4.42	eng	352	6333	244	11/1/2003	Scholastic
3	Harry Potter and the Prisoner of Azkaban (Harr...	J.K. Rowling	4.56	eng	435	2339585	36325	5/1/2004	Scholastic Inc.
4	Harry Potter Boxed Set Books 1-5 (Harry Potte...	J.K. Rowling	4.78	eng	2690	41428	164	9/13/2004	Scholastic

In [64]:
```
popular_df=df[df['ratings_count']>=250].sort_values('average_rating',ascending=False).head(50)
```

In [65]:
```
popular_df
```

Out[65]:

	title	authors	average_rating	language_code	Total_page	ratings_count	text_reviews_count	publication_date	publisher
6587	The Complete Calvin and Hobbes	Bill Watterson	4.82	eng	1456	32213	930	9/6/2005	Andrews McMeel Publishing
4	Harry Potter Boxed Set Books 1-5 (Harry Potte...	J.K. Rowling	4.78	eng	2690	41428	164	9/13/2004	Scholastic
6589	It's a Magical World (Calvin and Hobbes #11)	Bill Watterson	4.76	eng	176	23875	303	9/1/1996	Andrews McMeel Publishing
6	Harry Potter Collection (Harry Potter #1-6)	J.K. Rowling	4.73	eng	3342	28242	808	9/12/2005	Scholastic
	Homicidal								

In [66]:
```
popular_df.drop_duplicates('title')
```

	title	authors	average_rating	language_code	Total_page	ratings_count	text_reviews_count	publication_date	publisher
10937	The Price of the Ticket: Collected Nonfiction ...	James Baldwin	4.70	eng	712	404	30	9/15/1985	St. Martin's Press
6497	The Complete Aubrey/Maturin Novels (5 Volumes)	Patrick O'Brian	4.70	eng	6576	1338	81	10/17/2004	W. W. Norton Company
7042	The Sibley Field Guide to Birds of Western Nor...	David Allen Sibley	4.69	en-US	473	730	36	4/29/2003	Alfred A. Knopf
6591	The Days Are Just Packed	Bill Watterson	4.69	eng	176	20308	244	9/1/1993	Andrews McMeel Publishing
1530	The Life and Times of Scrooge McDuck	Don Rosa	4.67	eng	266	2467	149	6/1/2005	Gemstone Publishing

In [67]:
```
popular_df.shape
```

Out[67]: (50, 9)

In [71]:
```
def segregate(df):
    values = []
    for val in df.average_rating:
        if val >= 0 and val <= 1:
            values.append("Between 0 and 1")
        elif val > 1 and val <= 2:
            values.append("Between 1 and 2")
        elif val > 2 and val <= 3:
            values.append("Between 2 and 3")
        elif val > 3 and val <= 4:
            values.append("Between 3 and 4")
        elif val > 4 and val <= 5:
            values.append("Between 4 and 5")
        else:
            values.append("NaN")
    return values
```

In [72]:
```
df['ratings_dist'] = segregate(df)
df.head()
```

Out[72]:

	title	authors	average_rating	language_code	Total_page	ratings_count	text_reviews_count	publication_date	publisher	ratings_dist
0	Harry Potter and the Half-Blood Prince (Harry ...	J.K. Rowling	4.57	eng	652	2095690	27591	9/16/2006	Scholastic Inc.	Between 4 and 5
1	Harry Potter and the Order of the Phoenix (Har...	J.K. Rowling	4.49	eng	870	2153167	29221	9/1/2004	Scholastic Inc.	Between 4 and 5
2	Harry Potter and the Chamber of Secrets (Harry...	J.K. Rowling	4.42	eng	352	6333	244	11/1/2003	Scholastic	Between 4 and 5
3	Harry Potter and the Prisoner of Azkaban (Harr...	J.K. Rowling	4.56	eng	435	2339585	36325	5/1/2004	Scholastic Inc.	Between 4 and 5
4	Harry Potter Boxed Set Books 1-5 (Harry Potte...	J.K. Rowling	4.78	eng	2690	41428	164	9/13/2004	Scholastic	Between 4 and 5

In [74]:
```
print(df['ratings_dist'].value_counts().index)
print(df['ratings_dist'].value_counts().values)
```

```
Index(['Between 3 and 4', 'Between 4 and 5', 'Between 2 and 3',
       'Between 0 and 1', 'Between 1 and 2'],
      dtype='object')
[6285 4735   69   27    7]
```

In [77]:
```
df['ratings_dist'] = df['ratings_dist'].replace({
    'Between 0 and 1':0,
    'Between 1 and 2':1,
    'Between 2 and 3':2,
    'Between 3 and 4':3,
    'Between 4 and 5':4
});
df.head()
```

Out[77]:

	title	authors	average_rating	language_code	Total_page	ratings_count	text_reviews_count	publication_date	publisher	ratings_dist
0	Harry Potter and the Half-Blood Prince (Harry ...	J.K. Rowling	4.57	eng	652	2095690	27591	9/16/2006	Scholastic Inc.	4
1	Harry Potter and the Order of the Phoenix (Har...	J.K. Rowling	4.49	eng	870	2153167	29221	9/1/2004	Scholastic Inc.	4
2	Harry Potter and the Chamber of Secrets (Harry...	J.K. Rowling	4.42	eng	352	6333	244	11/1/2003	Scholastic	4
3	Harry Potter and the Prisoner of Azkaban (Harr...	J.K. Rowling	4.56	eng	435	2339585	36325	5/1/2004	Scholastic Inc.	4
4	Harry Potter Boxed Set Books 1-5 (Harry Potte...	J.K. Rowling	4.78	eng	2690	41428	164	9/13/2004	Scholastic	4

creating the model

In [78]:
```
# Creating an instance of the NearestNeighbors model
from sklearn.neighbors import NearestNeighbors
model = NearestNeighbors(n_neighbors=6, algorithm='ball_tree')

# Fitting the model to the feature matrix of books
model.fit(df[['average_rating','Total_page','ratings_count','text_reviews_count','ratings_dist']])

# Querying the model to find the nearest neighbors
distance, indices = model.kneighbors(df[['average_rating','Total_page','ratings_count','text_reviews_count','ratings_dist']])
```

In [79]:
```
indices.shape
```

Out[79]:
```
(11123, 6)
```

In [80]:
```
df['indices'] = indices.tolist()
df['distance']= distance.tolist()
df.head()
```

Out[80]:

	title	authors	average_rating	language_code	Total_page	ratings_count	text_reviews_count	publication_date	publisher	ratings_dist	indices
0	Harry Potter and the Half-Blood Prince (Harry ...	J.K. Rowling	4.57	eng	652	2095690	27591	9/16/2006	Scholastic Inc.	4	[0, 2114, 23, 1, 2116, 4415]
1	Harry Potter and the Order of the Phoenix (Har...	J.K. Rowling	4.49	eng	870	2153167	29221	9/1/2004	Scholastic Inc.	4	[1, 23, 2114, 0, 2116, 4415]
2	Harry Potter and the Chamber of Secrets (Harry...	J.K. Rowling	4.42	eng	352	6333	244	11/1/2003	Scholastic	4	[2, 254, 8166, 839, 10615, 7824]
3	Harry Potter and the Prisoner of Azkaban (Harr...	J.K. Rowling	4.56	eng	435	2339585	36325	5/1/2004	Scholastic Inc.	4	[3, 4415, 307, 1462, 1, 1697]
4	Harry Potter Boxed Set Books 1-5 (Harry Potte...	J.K. Rowling	4.78	eng	2690	41428	164	9/13/2004	Scholastic	4	[4, 1643, 7870, 4143, 5456, 3624]

In [85]:
```
import re
class BookQuest:
    def __init__(self, dataframe, indices):
        self.df = dataframe
        self.indices = indices
        self.all_books_names = list(self.df["title"].values)

    def find_id(self,name):
        for index,string in enumerate(self.all_books_names):
            if re.search(name,string):
                index=index;
                break;
        return(index)

    def print_similar_books(self, query=None):
        if query:
            found_id = self.find_id(query)
            for id in self.indices[found_id][1:]:
                print(self.df.iloc[id]["title"])
```

In [86]:
```
recsys = BookQuest(df,df.indices)
recsys.print_similar_books("The Hobbit or There and Back Again")
print("-----------------------------------------------")
recsys.print_similar_books("Harry Potter")
```

```
Living La Dolce Vita: Bring the Passion  Laughter  and Serenity of Italy Into Your Daily Life
Penguin Book Of Norse Myths: Gods Of The Vikings
Body Language (Mark Manning Mystery  #3)
The Essential Augustine
Tales From Shakespeare
-----------------------------------------------
Animal Farm
The Fellowship of the Ring (The Lord of the Rings  #1)
Harry Potter and the Order of the Phoenix (Harry Potter  #5)
Lord of the Flies
Harry Potter and the Chamber of Secrets (Harry Potter  #2)
```

In [ ]: