```
import numpy as np
import pandas as pd
import matplotlib as mpl
import matplotlib.pyplot as plt
%matplotlib inline
%load_ext autoreload
%autoreload 2
%graphs_dir = "C:\Users\Tomasz Darmetko (darmetkow)\goodreadbooks\"
```

```
goodreads.head()
```

	bookID	title	authors	average_rating	isbn	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	publisher
0	1	Harry Potter and the Half-Blood Prince (Harry Potter, #6)	J.K. Rowling/Mary GrandPré	4.57	0439785960	9780439785960	eng	652	2095690	27591	9/16/2004	Scholastic Inc.
1	2	Harry Potter and the Order of the Phoenix (Harry Potter, #5)	J.K. Rowling/Mary GrandPré	4.49	0439358078	9780439358078	eng	870	2339585	36325	09/01/2004	Scholastic Inc.
2	4	Harry Potter and the Prisoner of Azkaban (Harry Potter, #3)	J.K. Rowling/Mary GrandPré	4.57	0439655438	9780439655438	eng	352	2041594	24768	11/01/2003	Scholastic Inc.
3	5	Harry Potter and the Chamber of Secrets (Harry Potter, #2)	J.K. Rowling	4.42	0439554896	9780439554896	eng	352	2548483	26326	09/01/2004	Scholastic Inc.
4	8	Harry Potter and the Goblet of Fire (Harry Potter, #4)	J.K. Rowling/Mary GrandPré	4.56	0439139597	9780439139597	eng	734	2096690	24768	9/15/2004	Scholastic Inc.

```
goodreads.shape
```

```
(11127, 12)
```

```
goodreads.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11127 entries, 0 to 11126
Data columns (total 12):
 bookID  11127 non-null  int64
 title  11127 non-null  object
 authors  11127 non-null  object
 average_rating  11127 non-null  float64
 isbn  11127 non-null  object
 isbn13  11127 non-null  int64
 language_code  11127 non-null  object
 num_pages  11127 non-null  int64
 ratings_count  11127 non-null  int64
 text_reviews_count  11127 non-null  int64
 publication_date  11127 non-null  object
 publisher  11127 non-null  object
dtype: float64, int64, object
memory usage: 1.0+ MB
```

```
goodreads.describe()
```

	bookID	average_rating	isbn13	num_pages	ratings_count	text_reviews_count
count	11127.000000	11127.000000	1.112700e+04	11127.000000	1.112700e+04	11127.000000
mean	21310.830987	3.934075	9.731904e+12	336.405016	17942.859960	542.048967
std	14041.094491	0.350485	4.687320e+11	241.896277	112576.950019	2804.578893
min	1.000000	2.605000	9.000000e+12	0.000000	0.000000	0.000000
25%	10700.500000	3.770000	9.780345e+12	214.000000	104.000000	9.000000
50%	20987.000000	3.960000	9.780451e+12	337.000000	745.000000	39.000000
75%	30148.000000	4.094000	9.781707e+12	417.000000	4993.000000	237.000000
max	45641.000000	4.820000	9.789766e+12	3456.000000	4597666.000000	94265.000000

```
goodreads.columns
```

```
Index(['bookID', 'title', 'authors', 'average_rating', 'isbn', 'isbn13',
       'language_code', 'num_pages', 'ratings_count', 'text_reviews_count',
       'publication_date', 'publisher'],
      dtype='object')
```

Sorting by average rating

```
goodreads.sort_values('average_rating', ascending=False)
```

	bookID	title	authors	average_rating	isbn	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	publisher
10798	45641	Steve Jobs	Walter Isaacson	4.82	9780307474278	9780307474278	eng	612	5	1	12/07/2009	Simon and Schuster
1267	1207	Emotional Intelligence: 10th Anniversary: With new chapter, notes, and interviews	Daniel Goleman	4.78	0553383690	9780553383690	eng	128	2	0	01/7/2022	Random Publishing Group
9240	25204	Steve Jobs: The Exclusive Biography	Jeffrey S. Young/William L. Simon	4.76	0471653152	9780471653152	eng	272	2	0	5/24/2006	Wiley
8778	25076	Steve Jobs: The Man Who Thought Different: A Book About Vision, Obsession, Passion, and Perfection	Jeffrey S. Young/William L. Simon	4.76	0615342042	9780615342042	eng	197	1	0	9/22/2007	FastLane
4925	1765	The Power of Now: A Guide to Spiritual Enlightenment	Eckhart Tolle	4.76	0931432476	9780931432476	eng	152	3	0	01/01/1970	Namaste Publishing
7480	20986	The Lucifer Effect: Understanding How Good People Turn Evil	Philip Zimbardo	4.75	0385506120	9780385506120	eng	530	5	1	06/01/2008	Random House
4752	1454	Steve Jobs	Walter Isaacson	4.75	1451648537	9781451648537	eng	490	35	3	04/06/2014	Simon and Schuster
8979	25395	Steve Jobs: The Journey Is the Reward	Brown	Dan Lyons	1	0517407124	9780517407124	eng	45	1	0	05/01/1977
9077	25128	Steve Jobs: The Journey Is the Reward	Lynn Dralle	4.75	0595365204	9780595365204	eng	240	36	4	1/15/2006
10	45	Steve Jobs	Walter Isaacson	4.75	9780743264808	9780743264808	eng	656	30	5	4/25/2004

11127 rows × 12 columns

Sorting by id for authors and ratings count

```
sorting_bookID = goodreads.sort_values('authors', 'ratings_count', ascending=[False, False])
sorting_bookID.head(20)
```

	bookID	title	authors	average_rating	isbn	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	publi
1204	1014	Wuthering Heights (Penguin Classics)	Zadie Smith/Emily Brontë	3.44	0141441054	9780141441054	fre	40	5	0	01/02/2011	Penguin
7276	20407	Sacrificial Violence: The Sacred in Modern Theory	Zoran Grozdanovski	4.15	1934752071	9781934752071	eng	302	23124	750	12/09/2010	Wipf and
2103	1756	The Resistance	Zoran Grozdanovski	3.99	1933392022	9781933392022	eng	454	1977	151	7/22/2008	Lulu.com
7280	20412	The Lucifer Effect: Understanding How Good People Turn Evil	Zoran Grozdanovski	3.99	1933392022	9781933392022	eng	454	1625	262	09/01/2008	Lulu.com
7280	20410	La Débâcle	Zoran Grozdanovski	4.01	0202305304	9780202305304	eng	592	921	52	9/28/2003	Aldine
7280	20417	Lucifer Effect: Understanding How Good People Turn Evil	Zoran Grozdanovski	4.00	0970420204	9780970420204	fre	452	630	152	5/18/2011	Zuhr
2071	1684	The Gulag Archipelago, 1918-1956: An Experiment in Literary Investigation, I-II	Zoran Grozdanovski	3.91	0671189014	9780671189014	eng	367	8	1	12/01/1985	Harper
7280	20420	The Crucible of Terror	Zoran Grozdanovski	3.9	1933392022	9781933392022	eng	454	1	0	7/22/2008	Lulu.com
7280	20411	Les Misérables: Complete and Unabridged	Zoran Grozdanovski	3.7	1598000007	9781598000007	eng	354	44	9	9/28/2009	BOTA
2074	1616	The Revolution Betrayed	Boris Grozdanovski	3.77	0202304017	9780202304017	en-US	288	34841	2473	10/06/2004	Lulu.com
8740	24980	El Manual de la Guerra Gitana	Zoran Grozdanovski	3.74	0970420204	9780970420204	spa	280	132	19	12/22/2007	Zuhr
8740	24981	Die Zerstörung der Vernunft: Der Irrationalismus in der Philosophie Heute	Zoran Grozdanovski	3.74	0800034346	9780800034346	ger	301	1	0	10/01/2004	Fortress
4072	1785	Homecoming	Zoran Grozdanovski	4.18	7880406001	9787880406001	eng	32	153	15	12/22/2007	Eos
4072	1787	All Souls Come Home: Stories	Zoran Grozdanovski	4.05	0891800014	9780891800014	eng	240	2909	114	12/22/2007	Eos
7280	27981	Les Misérables: Complete and Unabridged	Zoran Grozdanovski	4.07	4194000006	9784194000006	jpn	216	5162	325	9/14/2004	Rosé
4072	1783	Fugitive Days: A Memoir	Zoran Grozdanovski	3.96	7534000001	9787534000001	eng	189	2574	322	12/22/2007	Eos
4072	1782	Memoirs of a Revolutionary: Essays in Marxist Social Philosophy	Zoran Grozdanovski	4.06	0900372333	9780900372333	eng	250	19574	1108	09/01/2004	Pluto

Top 10 good authors

```
top10_authors = goodreads.groupby('authors').agg({'bookID':'count'})[:].reset_index().sort_values(by='count', ascending=False).head(10)
```

	authors	count
4750	P.G. Wodehouse	40
4750	Stephen King	40
5273	Agatha Christie	39
4750	Terry Pratchett	37
4750	Agatha Christie	36
2906	Piers Anthony	34
4750	Nicholas Lantos	34
5501	Sandra Brown	33
4750	Mercedes Lackey	32
4891	Margaret Weis/Tracy Hickman	31
2700	James Patterson	31
5000	Laurell K. Hamilton	30
6350	Terry Pratchett	29
4750	Anne McCaffrey	29
4750	Dean Koontz	29
4750	Ann McCaffrey	29
2700	Mary Higgins Clark	28
2300	Carolyn Keene	28
2300	Dan Simmons	25
2300	Gene Wolfe/Gene Wolfe	25
2300	Jane Austen	25

```
goodreads['authors'].value_counts()
```

```
J.K. Rowling/Mary GrandPré    7
Rick Riordan                  6
Stephenie Meyer               5
Gail Carriger                 5
J.R. Ward                     5
Brandon Sanderson             5
Jim Butcher                   5
Patricia Briggs               5
Ilona Andrews                 5
Kresley Cole                  5
Lynsay Sands                  5
Sherrilyn Kenyon              5
Dean Koontz                   5
Name: authors, Length: 3380, dtype: int64
```

```
print('{:1} / {:1.1f} Books/Author', '{:1} Rating', '{:1} Reviews/Book', '{:1} Pages/Book', '{:1} Ratings/Book', '{:1} Reviews/Rating', '{:1} Pages/Rating', '{:1} Ratio Pages/Reviews', '{:1} Ratio Ratings/Reviews', '{:1} Ratio Pages/Ratings').
format(len(goodreads.authors.value_counts()),
       goodreads.shape[0]/len(goodreads.authors.value_counts()),
       goodreads.average_rating.sum(),
       goodreads.text_reviews_count.sum()/goodreads.ratings_count.sum(),
       goodreads.num_pages.sum()/goodreads.text_reviews_count.sum(),
       goodreads.num_pages.sum()/goodreads.ratings_count.sum(),
       goodreads.text_reviews_count.sum()/goodreads.average_rating.sum(),
       goodreads.num_pages.sum()/goodreads.average_rating.sum(),
       goodreads.text_reviews_count.sum()/goodreads.num_pages.sum(),
       goodreads.ratings_count.sum()/goodreads.text_reviews_count.sum(),
       goodreads.ratings_count.sum()/goodreads.num_pages.sum)
```

```
3380 / 3.3 Books/Author
43091.4 Rating
1.6 Reviews/Book
336.4 Pages/Book
17942.9 Ratings/Book
0.0 Reviews/Rating
0.0 Pages/Rating
0.0 Ratio Pages/Reviews
0.0 Ratio Ratings/Reviews
0.0 Ratio Pages/Ratings.
```

Grouping authors

```
authors = goodreads.groupby('authors')
```

Grouping authors and languages

```
authors_lang = goodreads.groupby(['authors', 'language_code'])
```

Grouping languages

```
lang = goodreads.groupby('language_code')
```

Grouping publishers

```
publisher = goodreads.groupby('publisher')
```

Checking for any missing values

```
goodreads.isnull().sum()
```

```
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
```

Count missing values

```
goodreads.isnull().sum().sum()
```

```
0
```

```
goodreads.shape
```

```
(11127, 12)
```

```
goodreads.dtypes
```

```
bookID                int64
title                 object
authors               object
average_rating        float64
isbn                  object
isbn13                int64
language_code         object
num_pages             int64
ratings_count         int64
text_reviews_count    int64
publication_date      object
publisher             object
dtype: object
```

Checking unique values

```
goodreads.nunique()
```

```
bookID                11127
title                 10346
authors               3380
average_rating        229
isbn                  11127
isbn13                11127
language_code         27
num_pages             1191
ratings_count         4690
text_reviews_count    1093
publication_date      2529
publisher             1329
dtype: int64
```

Checking missing values

```
goodreads.isnull().sum()
```

```
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
```

Checking missing values

```
goodreads.isnull().sum()[goodreads.isnull().sum()!=0]
```

```
Series([], dtype: int64)
```

```
goodreads.publication_date.value_counts[:20].plot.barh(rot=0)
```

```
<Function matplotlib.pyplot.show(*args,*kwargs, block=None)>
```

Checking missing values

```
goodreads.isnull()
```

	bookID	title	authors	average_rating	isbn	isbn13	language_code	num_pages	ratings_count	text_reviews_count	publication_date	publisher
0	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...
11123	False	False	False	False	False	False	False	False	False	False	False	False
11124	False	False	False	False	False	False	False	False	False	False	False	False
11125	False	False	False	False	False	False	False	False	False	False	False	False
11126	False	False	False	False	False	False	False	False	False	False	False	False

11127 rows × 12 columns

Top 10 Publisher

```
publisher = goodreads.value_counts('publisher').sort_values(ascending=False).head(10)
publisher.plot.barh(rot=0, figsize=(8,5))
```

```
<AxesSubplot:ylabel='publisher'>
```

Top 10 Language

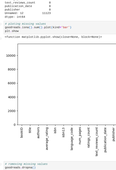
```
plt.figure(figsize = (8,5))
sns.barplot(lang.size().sort_values(ascending=False)[:10].index, lang.size().sort_values(ascending=False)[:10])
```

```
<AxesSubplot:>
```

Most Occuring Books

```
sns.set_context('poster')
plt.figure(figsize = (8,5))
books = goodreads[goodreads.ratings_count>1].sort_values(ascending=False).head(20)
sns.barplot(books.ratings_count, books.title)
sns.despine(left=True, bottom=True)
plt.xlabel('Number of occurances')
plt.ylabel('Books')
plt.show()
```

Top 10 Publication

```
plt.figure(figsize = (8,5))
sns.barplot(publication.size().sort_values(ascending=False)[:10].index, publication.size().sort_values(ascending=False)[:10])
```

```
<AxesSubplot:>
```