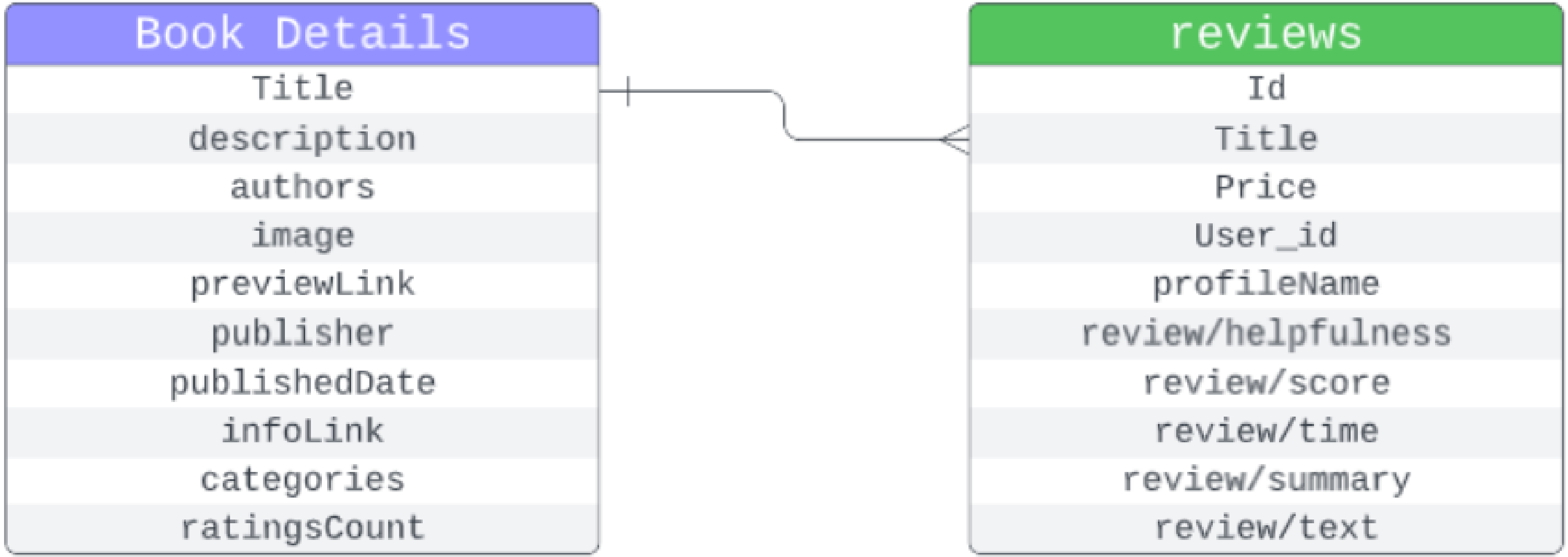
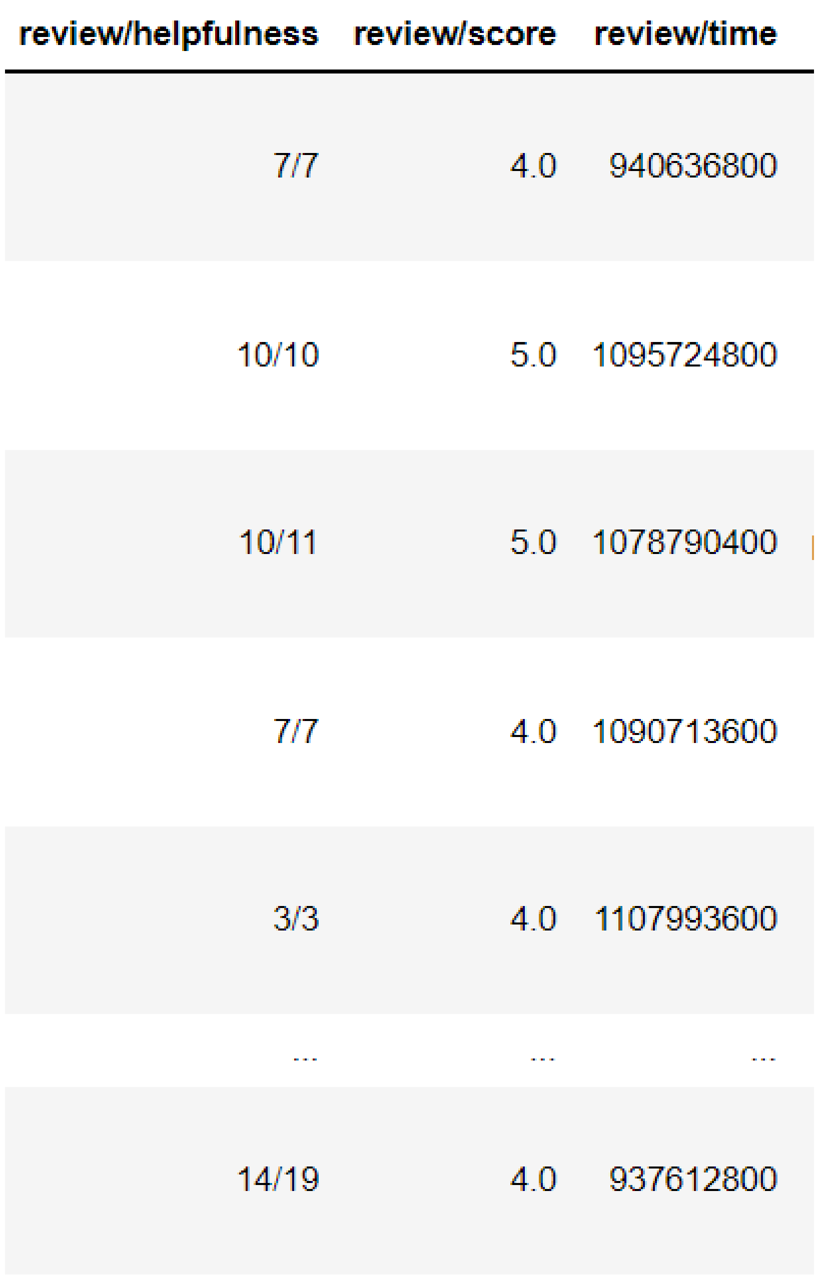
Sentiment Analysis on the Amazon Book Review Dataset

Hazel Kim

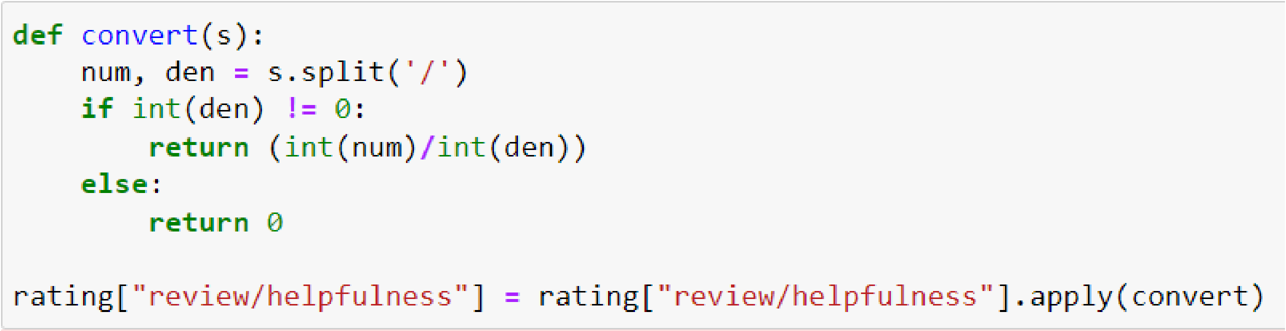


Book Details
Title
description
authors
image
previewLink
publisher
publishedDate
infoLink
categories
ratingsCount

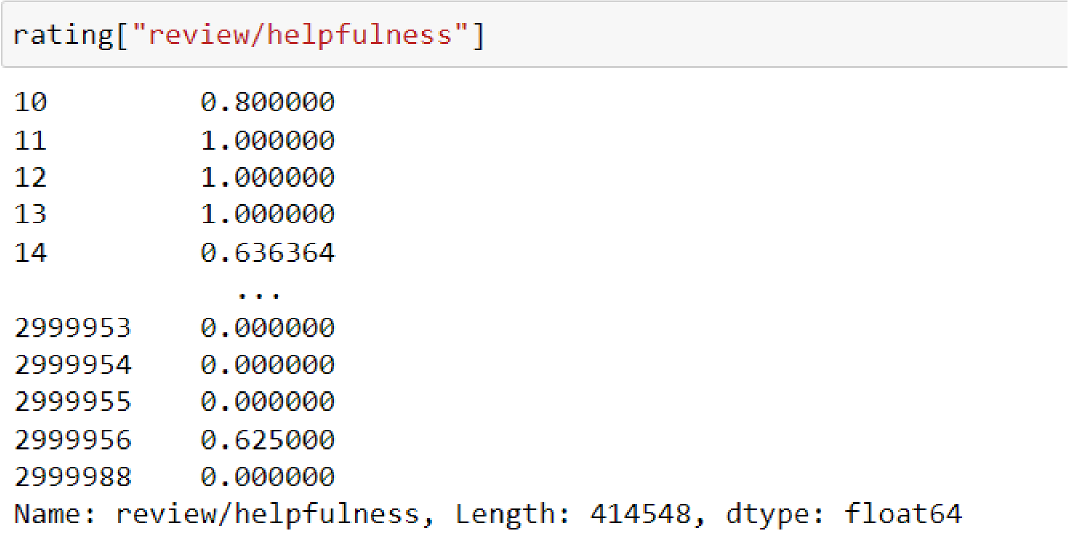
reviews
Id
Title
Price
User_id
profileName
review/helpfulness
review/score
review/time
review/summary
review/text



review/helpfulness	review/score	review/time
7/7	4.0	940636800
10/10	5.0	1095724800
10/11	5.0	1078790400
7/7	4.0	1090713600
3/3	4.0	1107993600
...	...	...
14/19	4.0	937612800

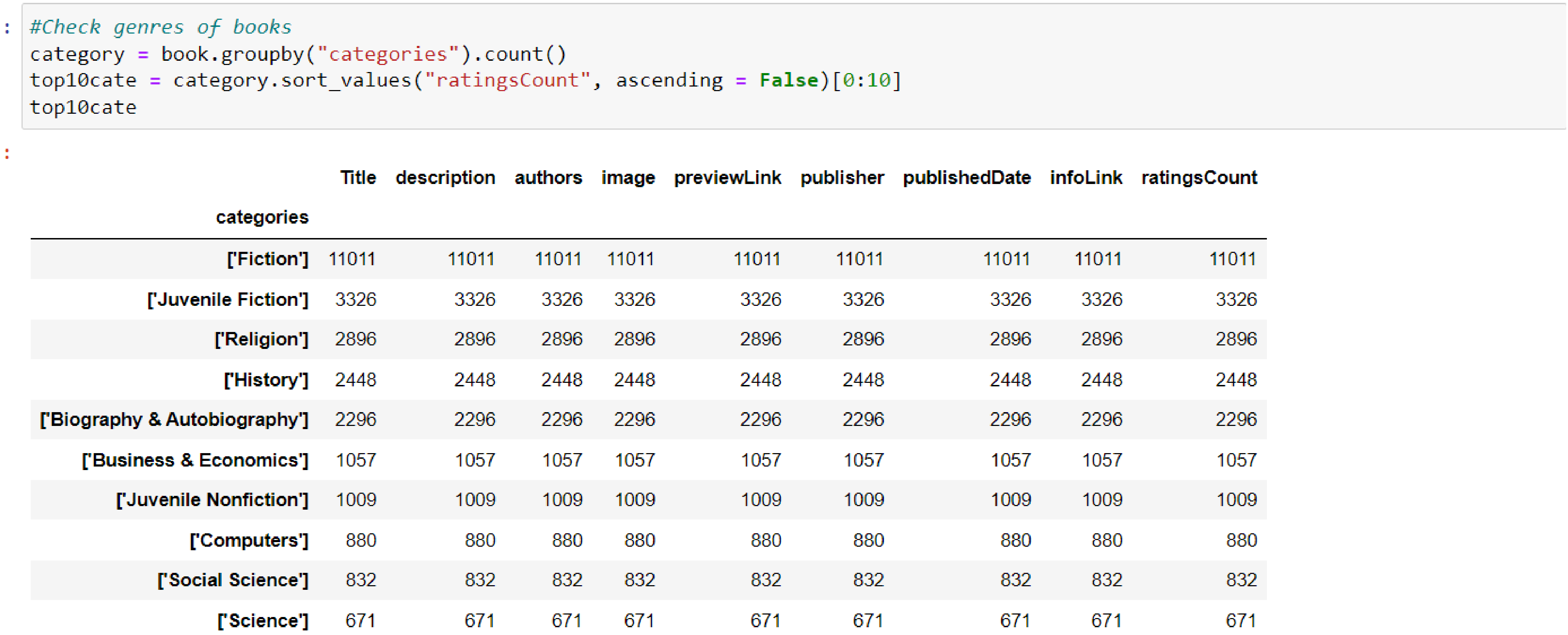


```python
def convert(s):
    num, den = s.split('/')
    if int(den) != 0:
        return (int(num)/int(den))
    else:
        return 0

rating["review/helpfulness"] = rating["review/helpfulness"].apply(convert)
```



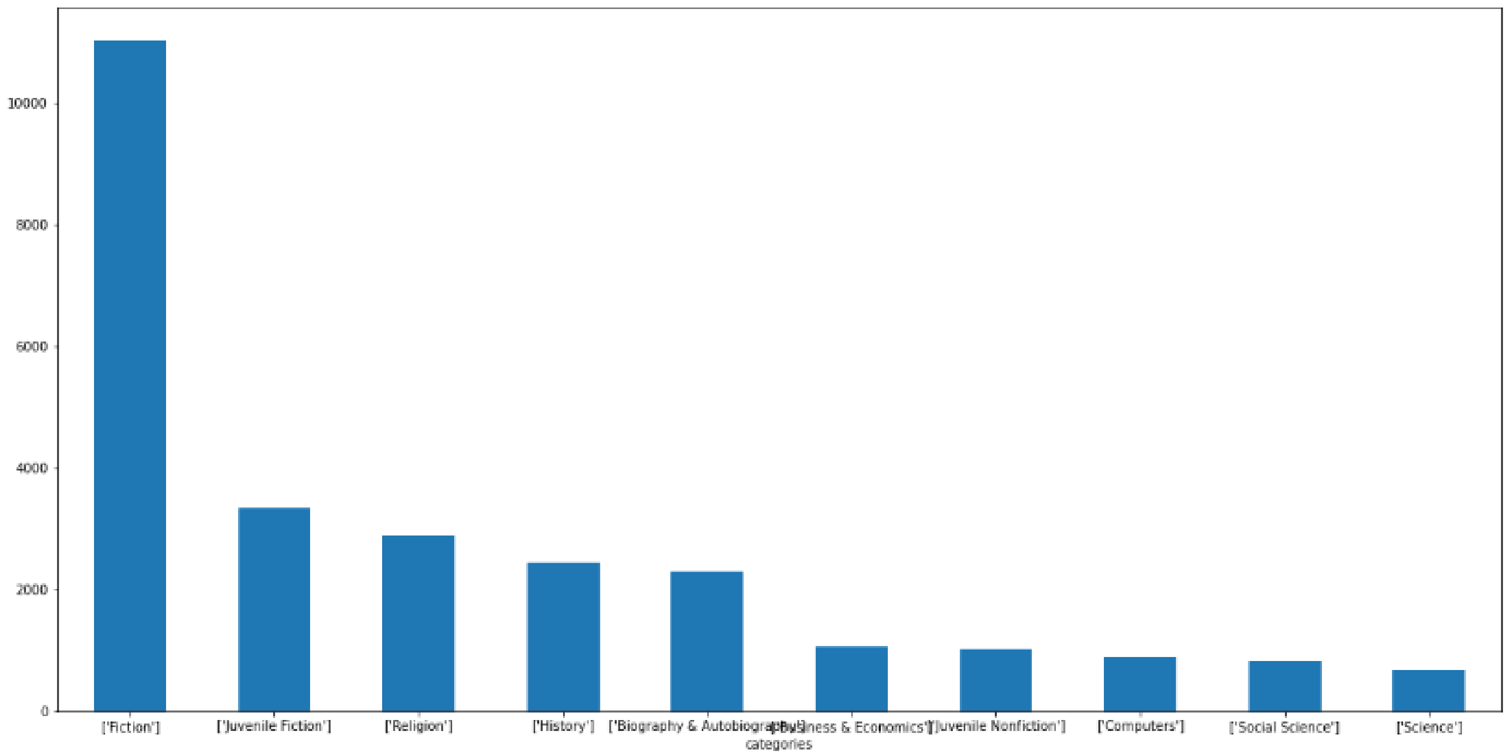
```python
rating["review/helpfulness"]
```

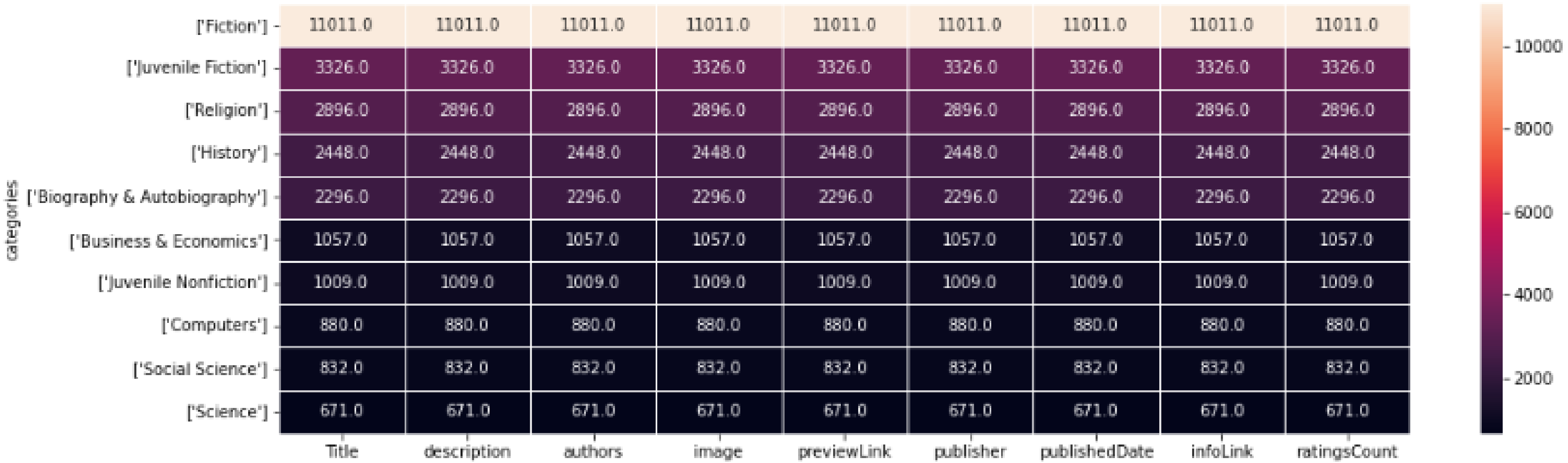
```
10         0.800000
11         1.000000
12         1.000000
13         1.000000
14         0.636364
             ...
2999953    0.000000
2999954    0.000000
2999955    0.000000
2999956    0.625000
2999988    0.000000
Name: review/helpfulness, Length: 414548, dtype: float64
```



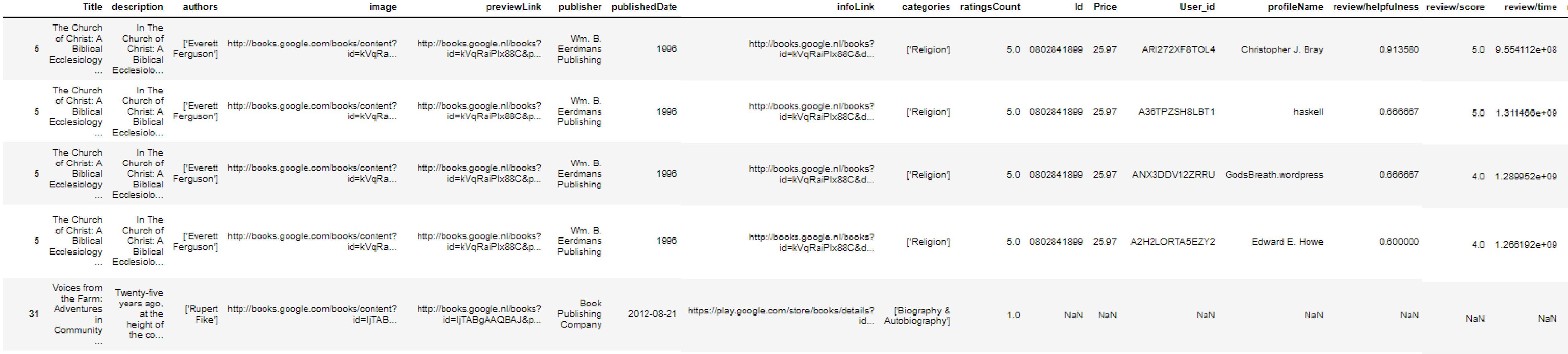
```
#Check genres of books
category = book.groupby("categories").count()
top10cate = category.sort_values("ratingsCount", ascending = False)[0:10]
top10cate
```

	Title	description	authors	image	previewLink	publisher	publishedDate	infoLink	ratingsCount
categories									
['Fiction']	11011	11011	11011	11011	11011	11011	11011	11011	11011
['Juvenile Fiction']	3326	3326	3326	3326	3326	3326	3326	3326	3326
['Religion']	2896	2896	2896	2896	2896	2896	2896	2896	2896
['History']	2448	2448	2448	2448	2448	2448	2448	2448	2448
['Biography & Autobiography']	2296	2296	2296	2296	2296	2296	2296	2296	2296
['Business & Economics']	1057	1057	1057	1057	1057	1057	1057	1057	1057
['Juvenile Nonfiction']	1009	1009	1009	1009	1009	1009	1009	1009	1009
['Computers']	880	880	880	880	880	880	880	880	880
['Social Science']	832	832	832	832	832	832	832	832	832
['Science']	671	671	671	671	671	671	671	671	671



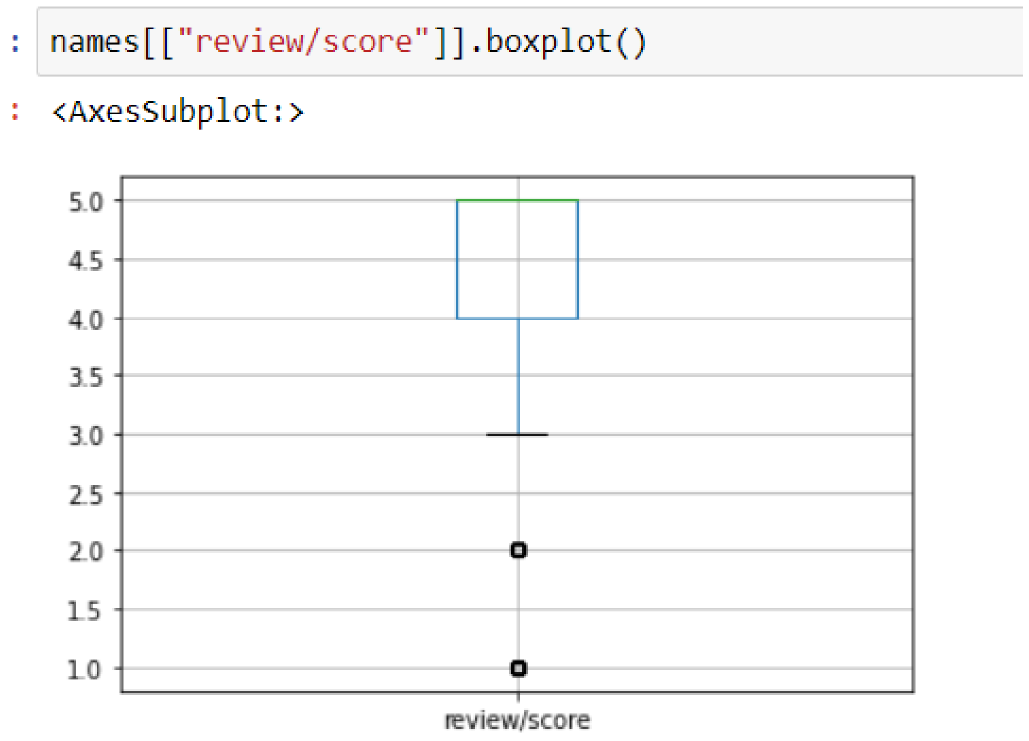


categories	Title	description	authors	image	previewLink	publisher	publishedDate	infoLink	ratingsCount
['Fiction']	11011.0	11011.0	11011.0	11011.0	11011.0	11011.0	11011.0	11011.0	11011.0
['Juvenile Fiction']	3326.0	3326.0	3326.0	3326.0	3326.0	3326.0	3326.0	3326.0	3326.0
['Religion']	2896.0	2896.0	2896.0	2896.0	2896.0	2896.0	2896.0	2896.0	2896.0
['History']	2448.0	2448.0	2448.0	2448.0	2448.0	2448.0	2448.0	2448.0	2448.0
['Biography & Autobiography']	2296.0	2296.0	2296.0	2296.0	2296.0	2296.0	2296.0	2296.0	2296.0
['Business & Economics']	1057.0	1057.0	1057.0	1057.0	1057.0	1057.0	1057.0	1057.0	1057.0
['Juvenile Nonfiction']	1009.0	1009.0	1009.0	1009.0	1009.0	1009.0	1009.0	1009.0	1009.0
['Computers']	880.0	880.0	880.0	880.0	880.0	880.0	880.0	880.0	880.0
['Social Science']	832.0	832.0	832.0	832.0	832.0	832.0	832.0	832.0	832.0
['Science']	671.0	671.0	671.0	671.0	671.0	671.0	671.0	671.0	671.0



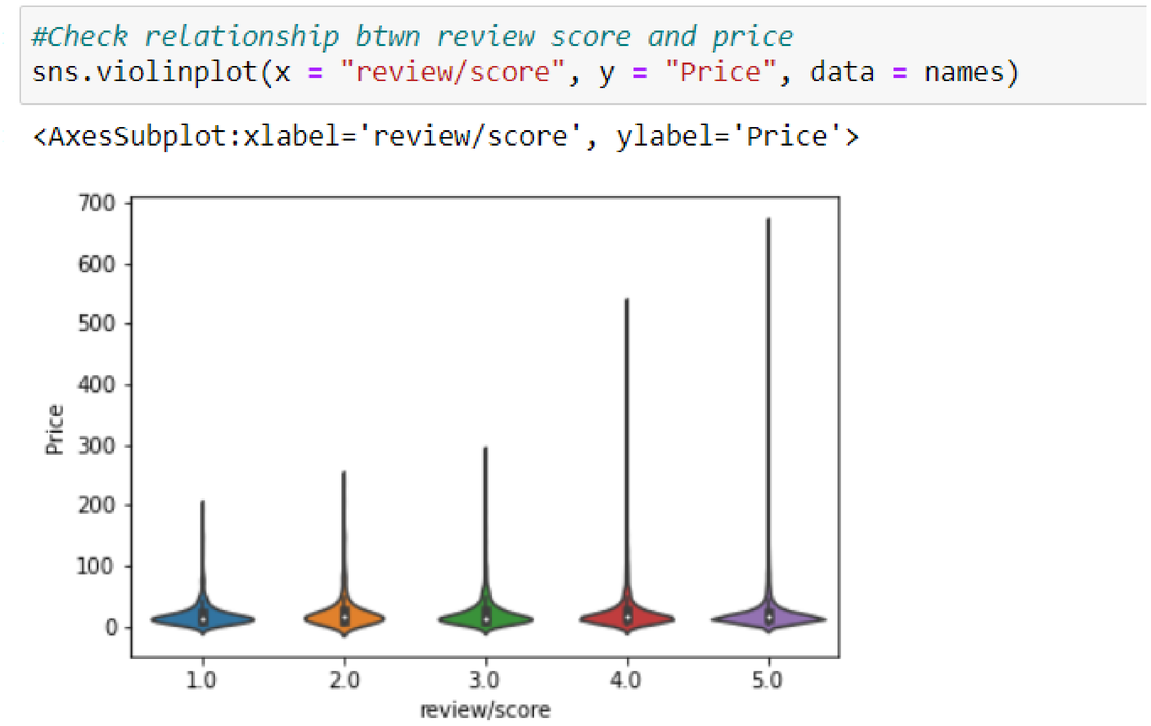
	Title	description	authors	image	previewLink	publisher	publishedDate	infoLink	categories	ratingsCount	Id	Price	User_id	profileName	review/helpfulness	review/score	review/time
5	The Church of Christ: A Biblical Ecclesiology ...	In The Church of Christ: A Biblical Ecclesiolo...	['Everett Ferguson']	http://books.google.com/books/content?id=kVqRa...	http://books.google.nl/books?id=kVqRaiPlx88C&p...	Wm. B. Eerdmans Publishing	1996	http://books.google.nl/books?id=kVqRaiPlx88C&d...	['Religion']	5.0	0802841899	25.97	ARI272XF8TOL4	Christopher J. Bray	0.913580	5.0	9.554112e+08
5	The Church of Christ: A Biblical Ecclesiology ...	In The Church of Christ: A Biblical Ecclesiolo...	['Everett Ferguson']	http://books.google.com/books/content?id=kVqRa...	http://books.google.nl/books?id=kVqRaiPlx88C&p...	Wm. B. Eerdmans Publishing	1996	http://books.google.nl/books?id=kVqRaiPlx88C&d...	['Religion']	5.0	0802841899	25.97	A36TPZSH8LBT1	haskell	0.666667	5.0	1.311466e+09
5	The Church of Christ: A Biblical Ecclesiology ...	In The Church of Christ: A Biblical Ecclesiolo...	['Everett Ferguson']	http://books.google.com/books/content?id=kVqRa...	http://books.google.nl/books?id=kVqRaiPlx88C&p...	Wm. B. Eerdmans Publishing	1996	http://books.google.nl/books?id=kVqRaiPlx88C&d...	['Religion']	5.0	0802841899	25.97	ANX3DDV12ZRRU	GodsBreath.wordpress	0.666667	4.0	1.289952e+09
5	The Church of Christ: A Biblical Ecclesiology ...	In The Church of Christ: A Biblical Ecclesiolo...	['Everett Ferguson']	http://books.google.com/books/content?id=kVqRa...	http://books.google.nl/books?id=kVqRaiPlx88C&p...	Wm. B. Eerdmans Publishing	1996	http://books.google.nl/books?id=kVqRaiPlx88C&d...	['Religion']	5.0	0802841899	25.97	A2H2LORTA5EZY2	Edward E. Howe	0.600000	4.0	1.266192e+09
31	Voices from the Farm: Adventures in Community ...	Twenty-five years ago, at the height of the co...	['Rupert Fike']	http://books.google.com/books/content?id=IjTAB...	http://books.google.nl/books?id=IjTABgAAQBAJ&p...	Book Publishing Company	2012-08-21	https://play.google.com/store/books/details?id...	['Biography & Autobiography']	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

205031 rows × 19 columns



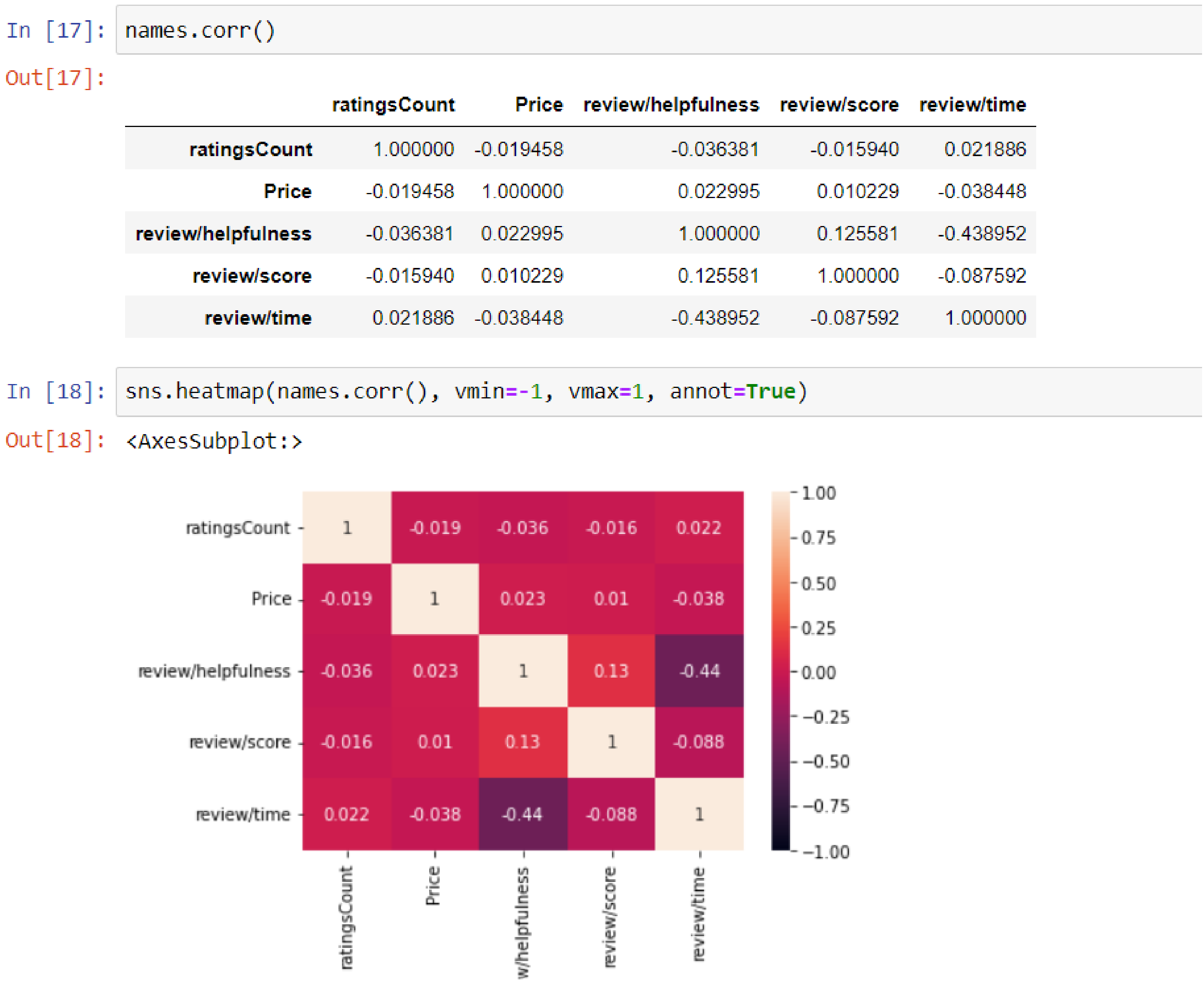
```
: names[["review/score"]].boxplot()
```

```
: <AxesSubplot:>
```



```
#Check relationship btwn review score and price
sns.violinplot(x = "review/score", y = "Price", data = names)
```

```
<AxesSubplot:xlabel='review/score', ylabel='Price'>
```

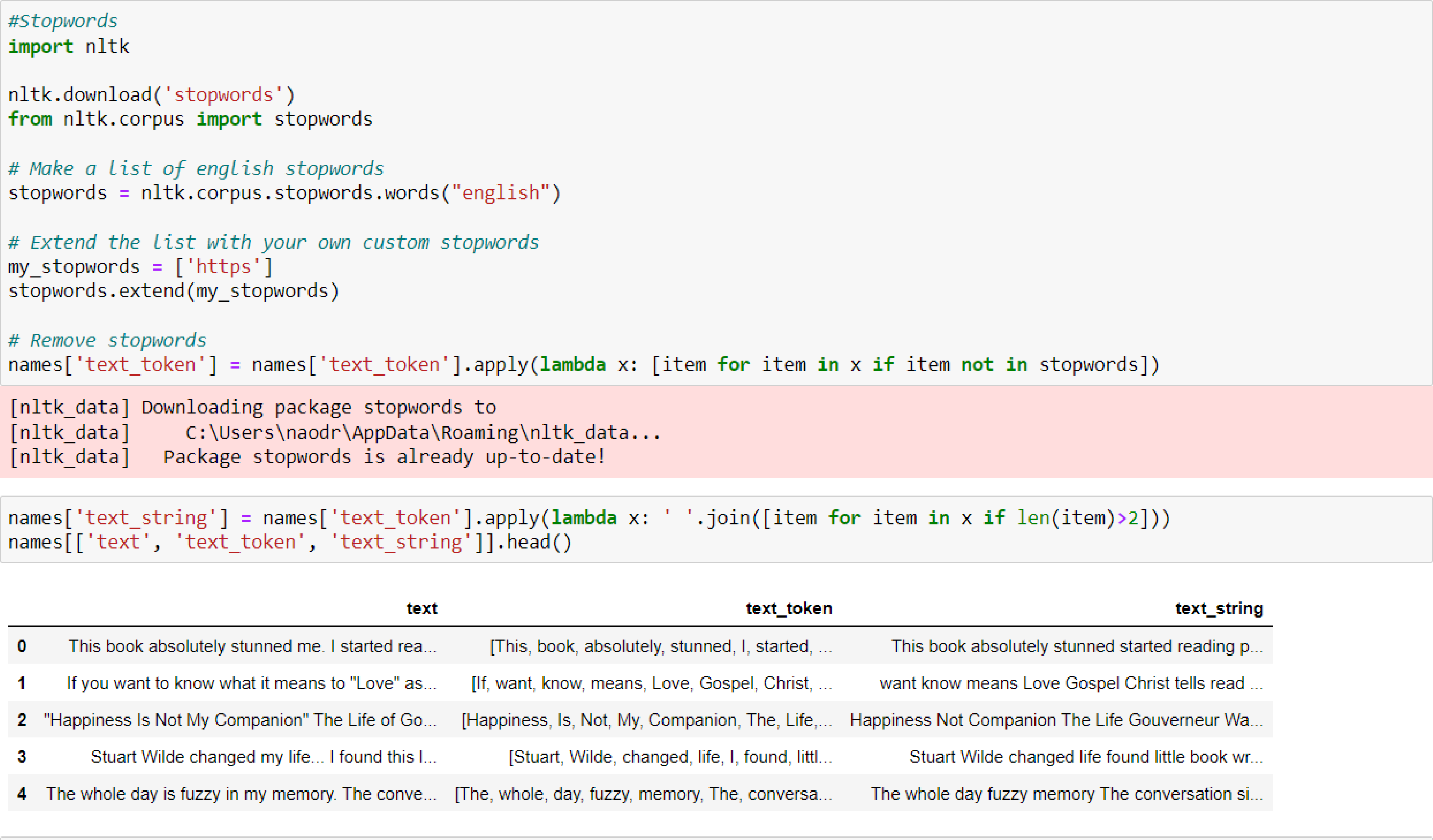


```
In [17]: names.corr()
```

Out[17]:

	ratingsCount	Price	review/helpfulness	review/score	review/time
ratingsCount	1.000000	-0.019458	-0.036381	-0.015940	0.021886
Price	-0.019458	1.000000	0.022995	0.010229	-0.038448
review/helpfulness	-0.036381	0.022995	1.000000	0.125581	-0.438952
review/score	-0.015940	0.010229	0.125581	1.000000	-0.087592
review/time	0.021886	-0.038448	-0.438952	-0.087592	1.000000

```
In [18]: sns.heatmap(names.corr(), vmin=-1, vmax=1, annot=True)

Out[18]: <AxesSubplot:>
```

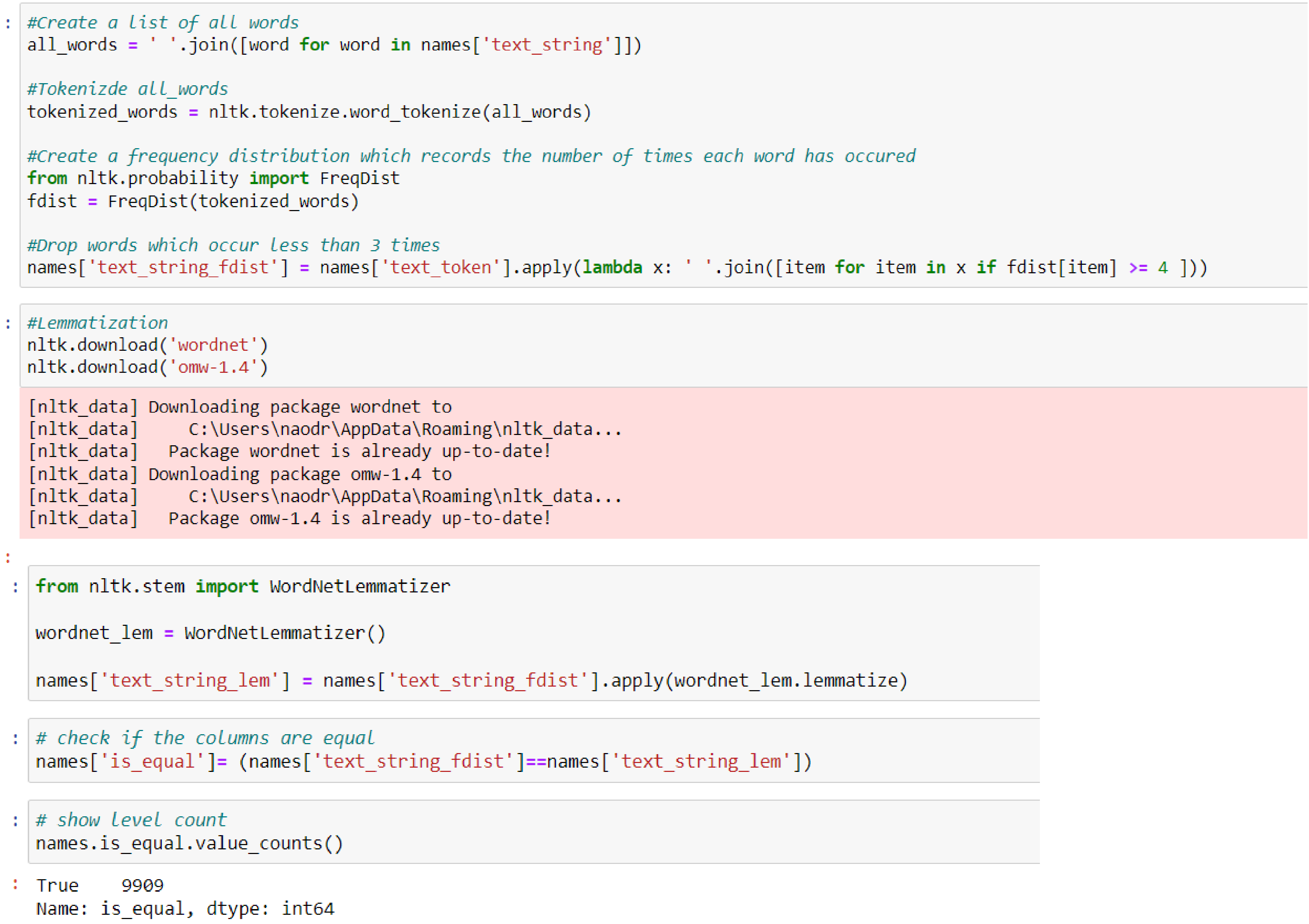


```
#Stopwords
import nltk

nltk.download('stopwords')
from nltk.corpus import stopwords

# Make a list of english stopwords
stopwords = nltk.corpus.stopwords.words("english")

# Extend the list with your own custom stopwords
my_stopwords = ['https']
stopwords.extend(my_stopwords)

# Remove stopwords
names['text_token'] = names['text_token'].apply(lambda x: [item for item in x if item not in stopwords])
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\naodr\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
names['text_string'] = names['text_token'].apply(lambda x: ' '.join([item for item in x if len(item)>2]))
names[['text', 'text_token', 'text_string']].head()
```

	text	text_token	text_string
0	This book absolutely stunned me. I started rea...	[This, book, absolutely, stunned, I, started, ...	This book absolutely stunned started reading p...
1	If you want to know what it means to "Love" as...	[If, want, know, means, Love, Gospel, Christ, ...	want know means Love Gospel Christ tells read ...
2	"Happiness Is Not My Companion" The Life of Go...	[Happiness, Is, Not, My, Companion, The, Life,...	Happiness Not Companion The Life Gouverneur Wa...
3	Stuart Wilde changed my life... I found this l...	[Stuart, Wilde, changed, life, I, found, littl...	Stuart Wilde changed life found little book wr...
4	The whole day is fuzzy in my memory. The conve...	[The, whole, day, fuzzy, memory, The, conversa...	The whole day fuzzy memory The conversation si...



```
#Create a list of all words
all_words = ' '.join([word for word in names['text_string']])

#Tokenizde all_words
tokenized_words = nltk.tokenize.word_tokenize(all_words)

#Create a frequency distribution which records the number of times each word has occured
from nltk.probability import FreqDist
fdist = FreqDist(tokenized_words)

#Drop words which occur less than 3 times
names['text_string_fdist'] = names['text_token'].apply(lambda x: ' '.join([item for item in x if fdist[item] >= 4 ]))
```

```
#Lemmatization
nltk.download('wordnet')
nltk.download('omw-1.4')
```

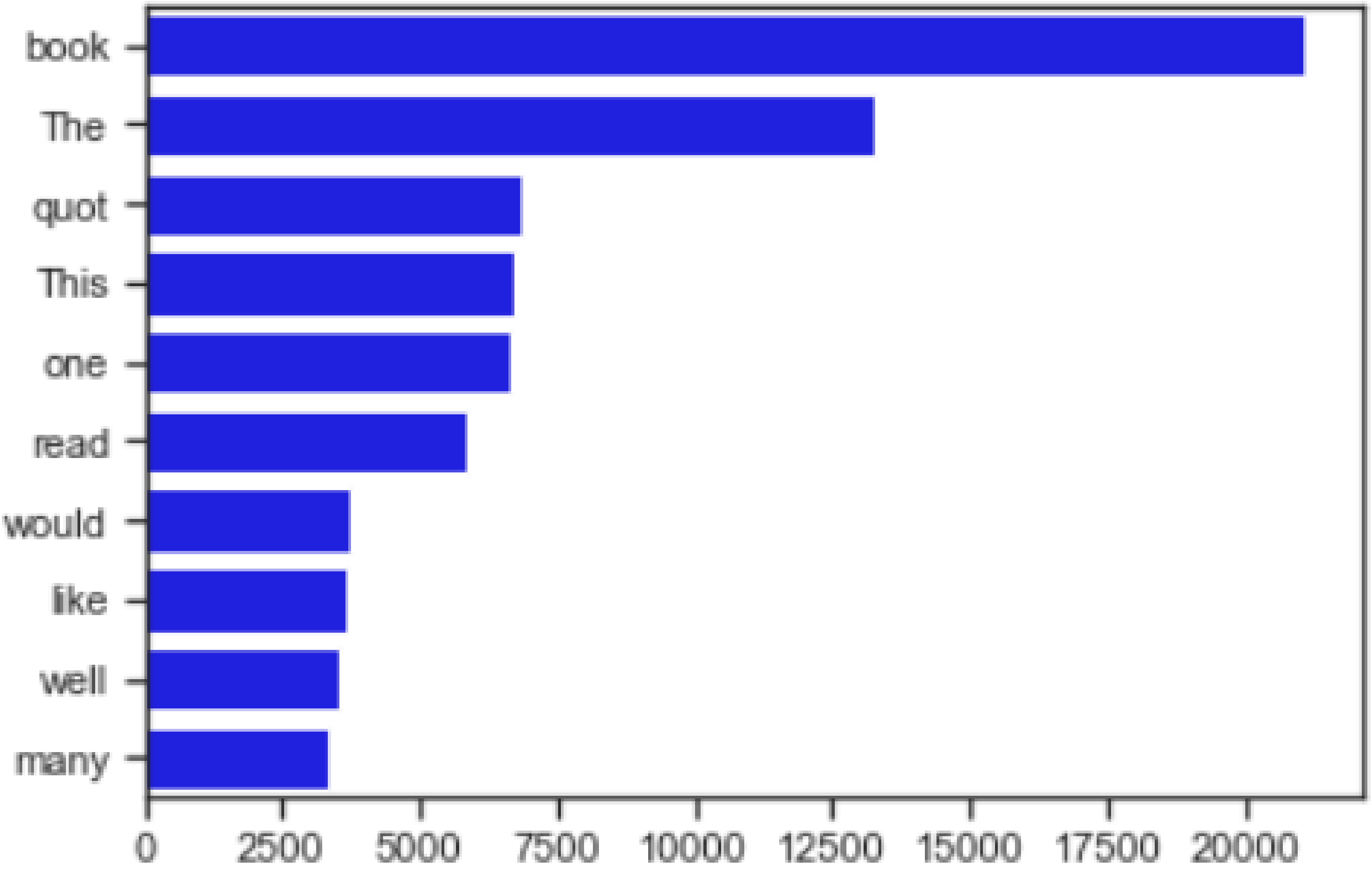
```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\naodr\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\naodr\AppData\Roaming\nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
```

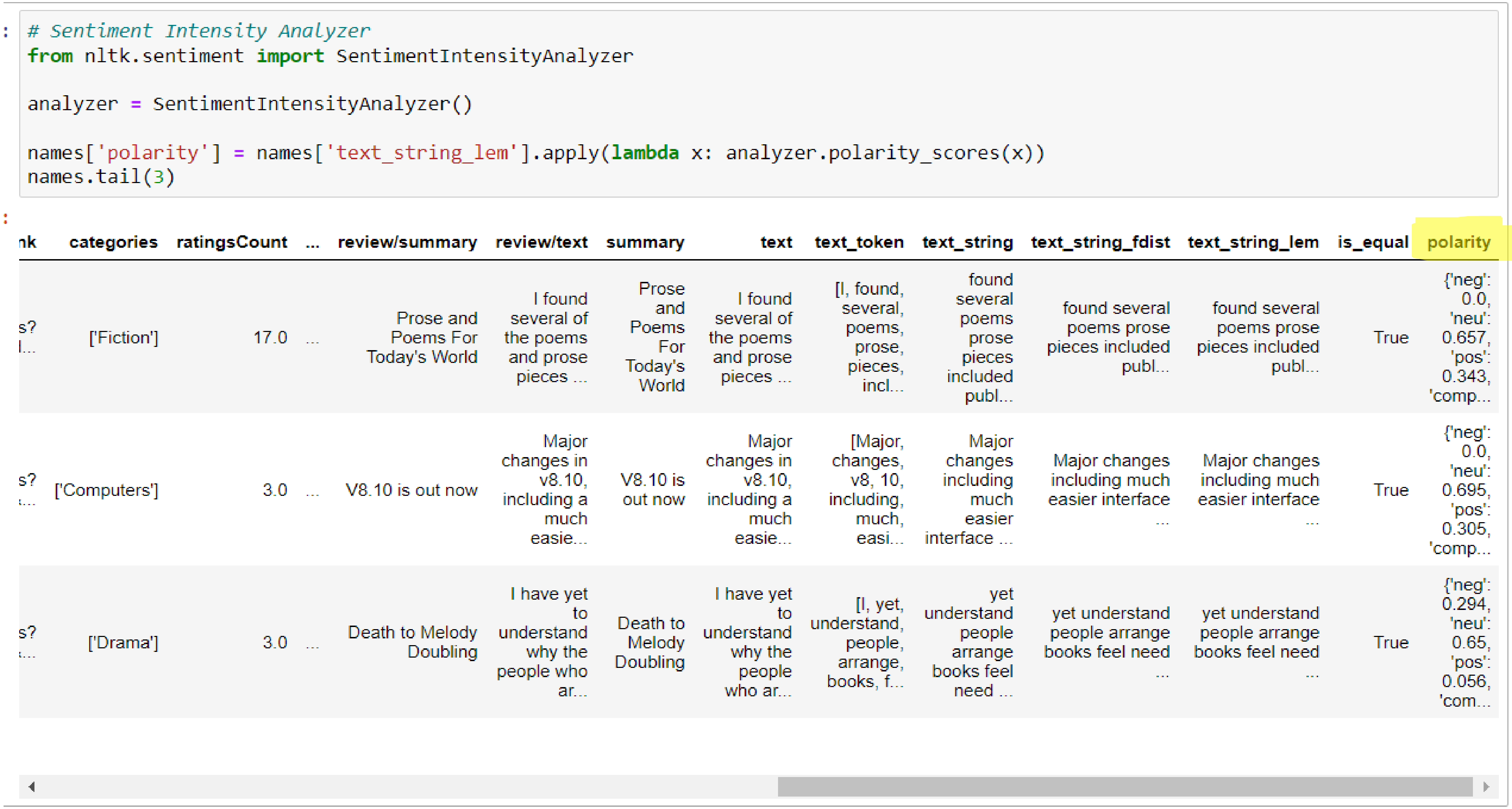
```
from nltk.stem import WordNetLemmatizer

wordnet_lem = WordNetLemmatizer()

names['text_string_lem'] = names['text_string_fdist'].apply(wordnet_lem.lemmatize)
```

```
# check if the columns are equal
names['is_equal']= (names['text_string_fdist']==names['text_string_lem'])
```

```
# show level count
names.is_equal.value_counts()
```

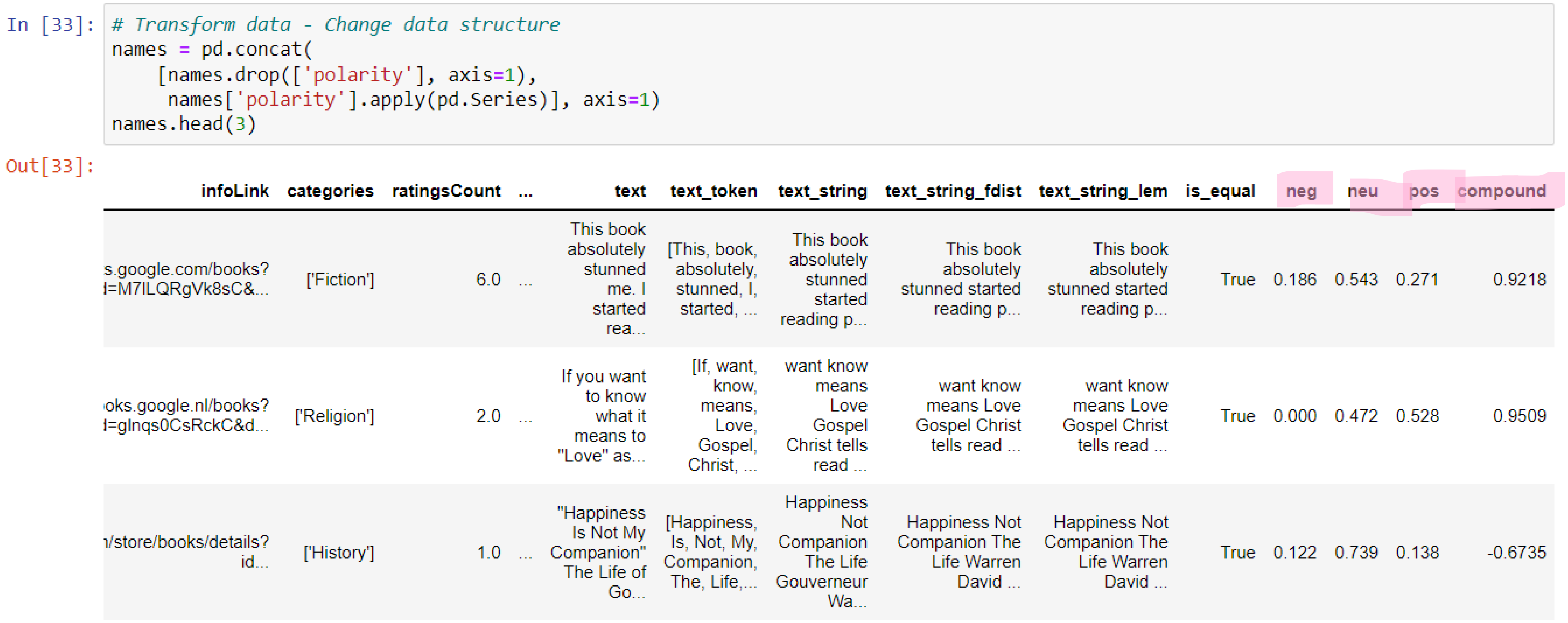
```
True     9909
Name: is_equal, dtype: int64
```


book
The
quot
This
one
read
would
like
well
many
0
2500
5000
7500
10000
12500
15000
17500
20000



```python
# Sentiment Intensity Analyzer
from nltk.sentiment import SentimentIntensityAnalyzer

analyzer = SentimentIntensityAnalyzer()

names['polarity'] = names['text_string_lem'].apply(lambda x: analyzer.polarity_scores(x))
names.tail(3)
```

nk	categories	ratingsCount	...	review/summary	review/text	summary	text	text_token	text_string	text_string_fdist	text_string_lem	is_equal	polarity
s? l...	['Fiction']	17.0	...	Prose and Poems For Today's World	I found several of the poems and prose pieces ...	Prose and Poems For Today's World	I found several of the poems and prose pieces ...	[I, found, several, poems, prose, pieces, incl...	found several poems prose pieces included publ...	found several poems prose pieces included publ...	found several poems prose pieces included publ...	True	{'neg': 0.0, 'neu': 0.657, 'pos': 0.343, 'comp...
s? ...	['Computers']	3.0	...	V8.10 is out now	Major changes in v8.10, including a much easie...	V8.10 is out now	Major changes in v8.10, including a much easie...	[Major, changes, v8, 10, including, much, easi...	Major changes including much easier interface ...	Major changes including much easier interface ...	Major changes including much easier interface ...	True	{'neg': 0.0, 'neu': 0.695, 'pos': 0.305, 'comp...
s? ...	['Drama']	3.0	...	Death to Melody Doubling	I have yet to understand why the people who ar...	Death to Melody Doubling	I have yet to understand why the people who ar...	[I, yet, understand, people, arrange, books, f...	yet understand people arrange books feel need ...	yet understand people arrange books feel need ...	yet understand people arrange books feel need ...	True	{'neg': 0.294, 'neu': 0.65, 'pos': 0.056, 'com...

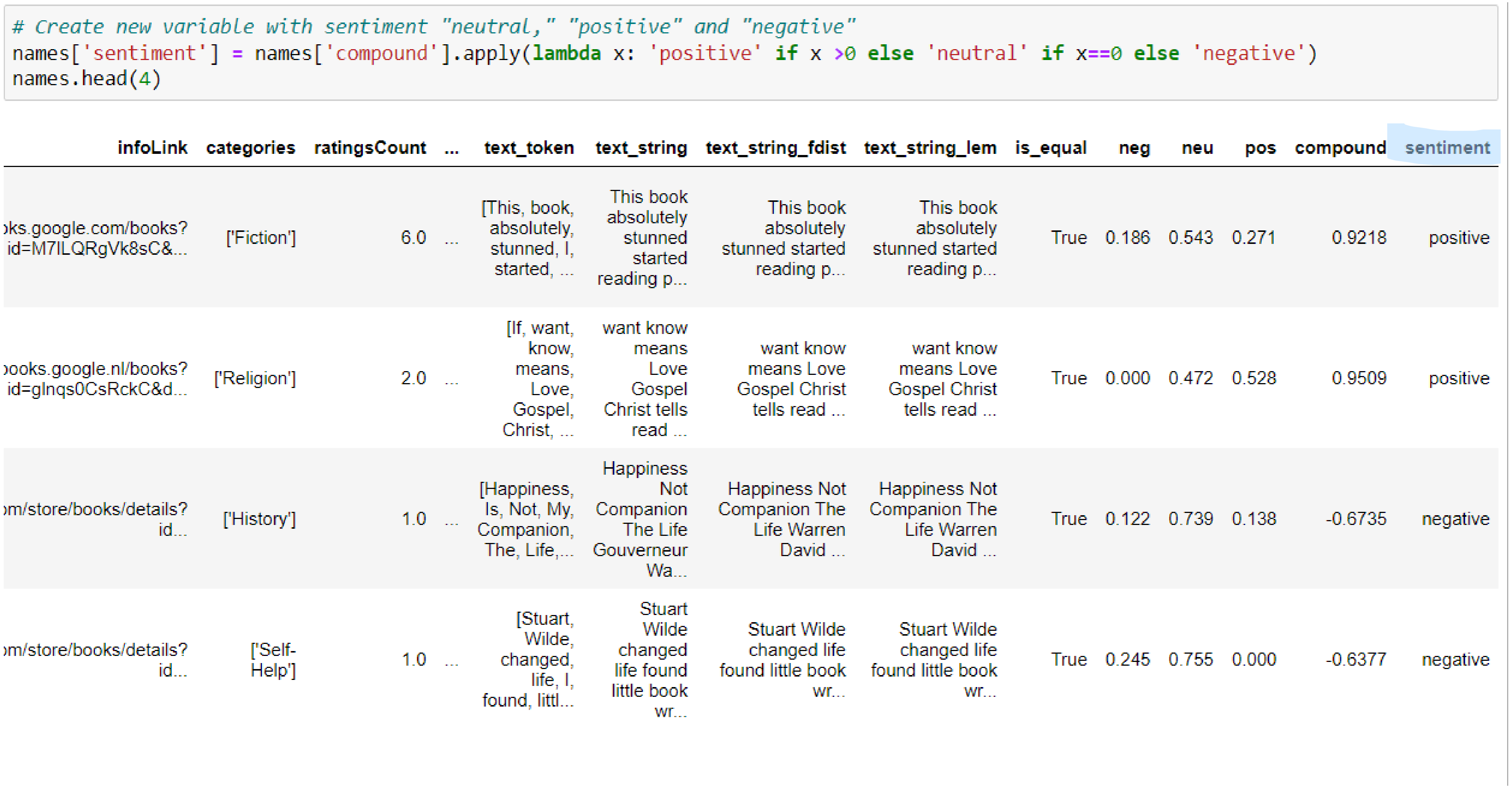


In [33]:

```
# Transform data - Change data structure
names = pd.concat(
    [names.drop(['polarity'], axis=1),
     names['polarity'].apply(pd.Series)], axis=1)
names.head(3)
```

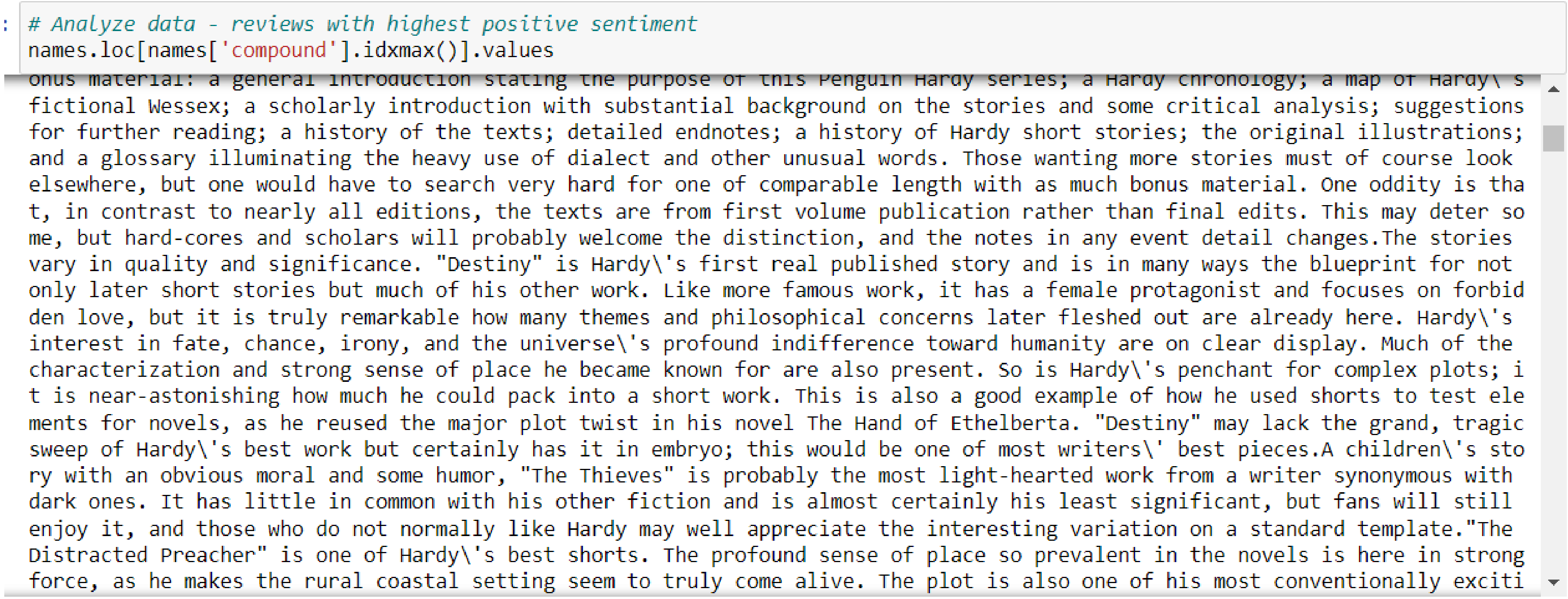
Out[33]:

infoLink	categories	ratingsCount	...	text	text_token	text_string	text_string_fdist	text_string_lem	is_equal	neg	neu	pos	compound
s.google.com/books? l=M7lLQRgVk8sC&...	['Fiction']	6.0	...	This book absolutely stunned me. I started rea...	[This, book, absolutely, stunned, I, started, ...	This book absolutely stunned started reading p...	This book absolutely stunned started reading p...	This book absolutely stunned started reading p...	True	0.186	0.543	0.271	0.9218
oks.google.nl/books? l=glnqs0CsRckC&d...	['Religion']	2.0	...	If you want to know what it means to "Love" as...	[If, want, know, means, Love, Gospel, Christ, ...	want know means Love Gospel Christ tells read ...	want know means Love Gospel Christ tells read ...	want know means Love Gospel Christ tells read ...	True	0.000	0.472	0.528	0.9509
ı/store/books/details? id...	['History']	1.0	...	"Happiness Is Not My Companion" The Life of Go...	[Happiness, Is, Not, My, Companion, The, Life,...	Happiness Not Companion The Life Gouverneur Wa...	Happiness Not Companion The Life Warren David ...	Happiness Not Companion The Life Warren David ...	True	0.122	0.739	0.138	-0.6735



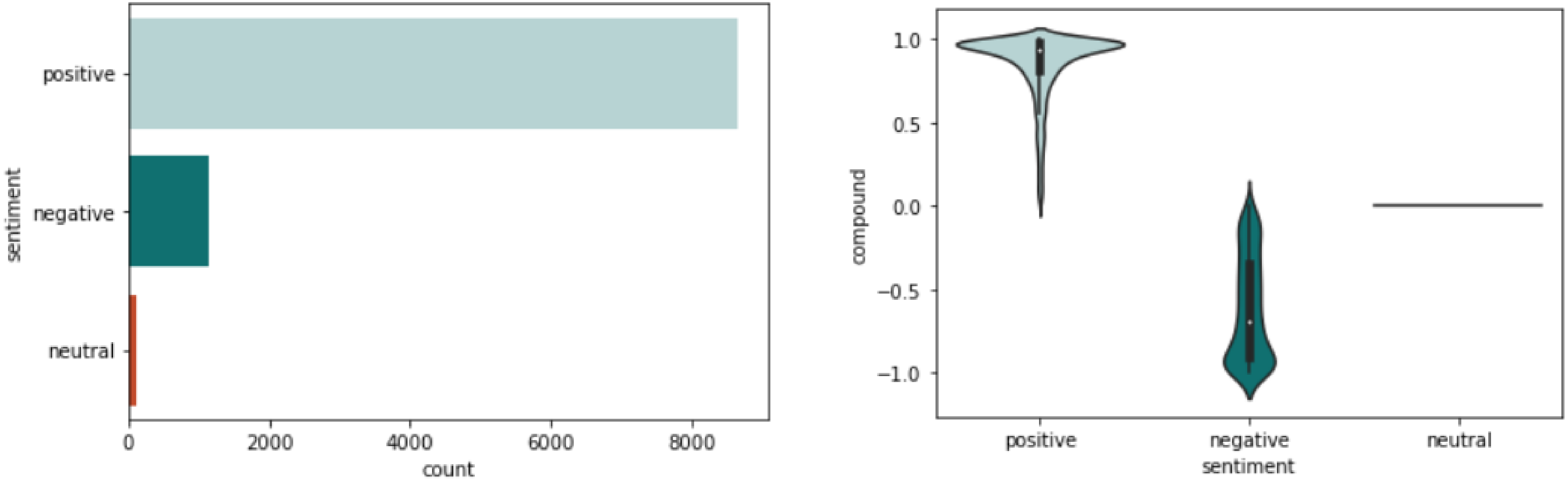
```
# Create new variable with sentiment "neutral," "positive" and "negative"
names['sentiment'] = names['compound'].apply(lambda x: 'positive' if x >0 else 'neutral' if x==0 else 'negative')
names.head(4)
```

infoLink	categories	ratingsCount	...	text_token	text_string	text_string_fdist	text_string_lem	is_equal	neg	neu	pos	compound	sentiment
ks.google.com/books?id=M7ILQRgVk8sC&...	['Fiction']	6.0	...	[This, book, absolutely, stunned, I, started, ...	This book absolutely stunned started reading p...	This book absolutely stunned started reading p...	This book absolutely stunned started reading p...	True	0.186	0.543	0.271	0.9218	positive
ooks.google.nl/books?id=glnqs0CsRckC&d...	['Religion']	2.0	...	[If, want, know, means, Love, Gospel, Christ, ...	want know means Love Gospel Christ tells read ...	want know means Love Gospel Christ tells read ...	want know means Love Gospel Christ tells read ...	True	0.000	0.472	0.528	0.9509	positive
m/store/books/details?id...	['History']	1.0	...	[Happiness, Is, Not, My, Companion, The, Life,...	Happiness Not Companion The Life Warren Gouverneur Wa...	Happiness Not Companion The Life Warren David ...	Happiness Not Companion The Life Warren David ...	True	0.122	0.739	0.138	-0.6735	negative
m/store/books/details?id...	['Self-Help']	1.0	...	[Stuart, Wilde, changed, life, I, found, littl...	Stuart Wilde changed life found little book wr...	Stuart Wilde changed life found little book wr...	Stuart Wilde changed life found little book wr...	True	0.245	0.755	0.000	-0.6377	negative



```
# Analyze data - reviews with highest positive sentiment
names.loc[names['compound'].idxmax()].values
```

```
onus material: a general introduction stating the purpose of this Penguin Hardy series; a Hardy chronology; a map of Hardy\'s
fictional Wessex; a scholarly introduction with substantial background on the stories and some critical analysis; suggestions
for further reading; a history of the texts; detailed endnotes; a history of Hardy short stories; the original illustrations;
and a glossary illuminating the heavy use of dialect and other unusual words. Those wanting more stories must of course look
elsewhere, but one would have to search very hard for one of comparable length with as much bonus material. One oddity is tha
t, in contrast to nearly all editions, the texts are from first volume publication rather than final edits. This may deter so
me, but hard-cores and scholars will probably welcome the distinction, and the notes in any event detail changes.The stories
vary in quality and significance. "Destiny" is Hardy\'s first real published story and is in many ways the blueprint for not
only later short stories but much of his other work. Like more famous work, it has a female protagonist and focuses on forbid
den love, but it is truly remarkable how many themes and philosophical concerns later fleshed out are already here. Hardy\'s
interest in fate, chance, irony, and the universe\'s profound indifference toward humanity are on clear display. Much of the
characterization and strong sense of place he became known for are also present. So is Hardy\'s penchant for complex plots; i
t is near-astonishing how much he could pack into a short work. This is also a good example of how he used shorts to test ele
ments for novels, as he reused the major plot twist in his novel The Hand of Ethelberta. "Destiny" may lack the grand, tragic
sweep of Hardy\'s best work but certainly has it in embryo; this would be one of most writers\' best pieces.A children\'s sto
ry with an obvious moral and some humor, "The Thieves" is probably the most light-hearted work from a writer synonymous with
dark ones. It has little in common with his other fiction and is almost certainly his least significant, but fans will still
enjoy it, and those who do not normally like Hardy may well appreciate the interesting variation on a standard template."The
Distracted Preacher" is one of Hardy\'s best shorts. The profound sense of place so prevalent in the novels is here in strong
force, as he makes the rural coastal setting seem to truly come alive. The plot is also one of his most conventionally exciti
```


positive
negative
neutral
sentiment
0
2000
4000
6000
8000
count
1.0
0.5
0.0
−0.5
−1.0
compound
positive
negative
neutral
sentiment

Further Development?

- TF-IDF
- CNN
- GPT 3

$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$

$$IDF(t) = log\frac{N}{1+df}$$

$$TF-IDF(t,d) = TF(t,d) * IDF(t)$$

$$\mathbf{tf}(t,d) = \frac{f_d(t)}{\max\limits_{w \in d} f_d(w)}$$

$$\mathbf{idf}(t,D) = \ln\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

$$\mathbf{tfidf}(t,d,D) = \mathbf{tf}(t,d) \cdot \mathbf{idf}(t,D)$$

$$\mathbf{tfidf}'(t,d,D) = \frac{\mathbf{idf}(t,D)}{|D|} + \mathbf{tfidf}(t,d,D)$$

$f_d(t) :=$ frequency of term t in document d

$D :=$ corpus of documents

Thank you☺