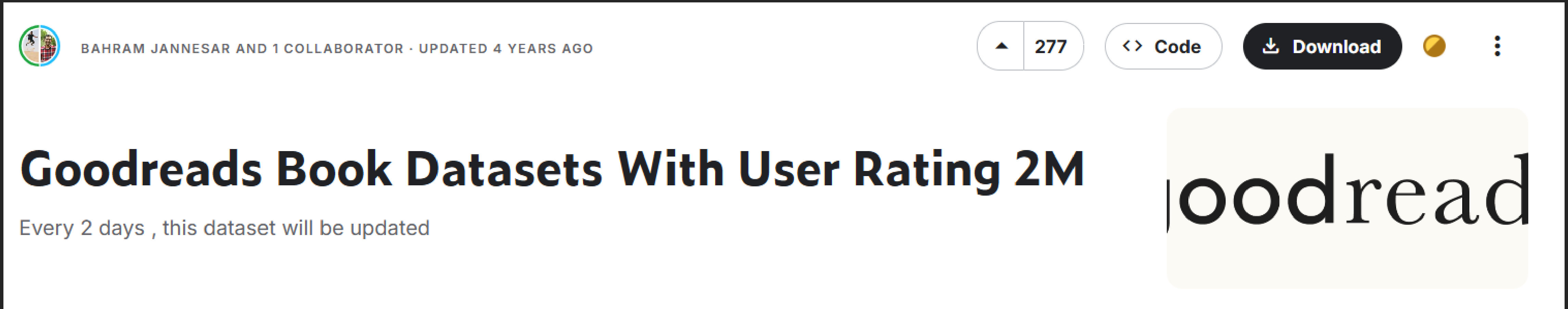
Checkpoint 1

Nikolina Krce

Izabrani dataset

- Goodreads Book Datasets With User Rating 2M
- https://www.kaggle.com/datasets/bahramjannesarr/goodreads-book-datasets-10m?select=book1000k-1100k.csv

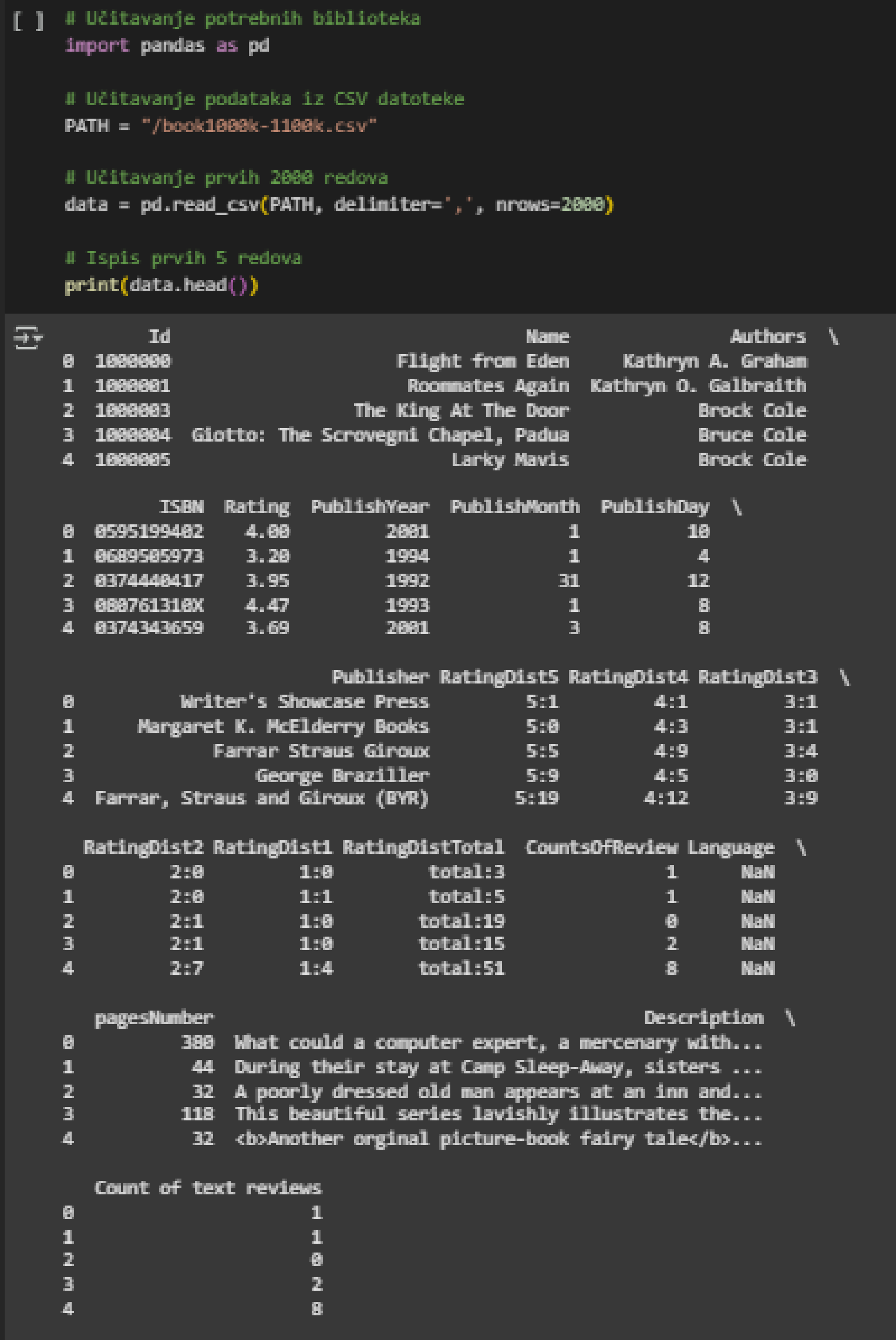


BAHRAM JANNESAR AND 1 COLLABORATOR · UPDATED 4 YEARS AGO

Goodreads Book Datasets With User Rating 2M

Every 2 days , this dataset will be updated

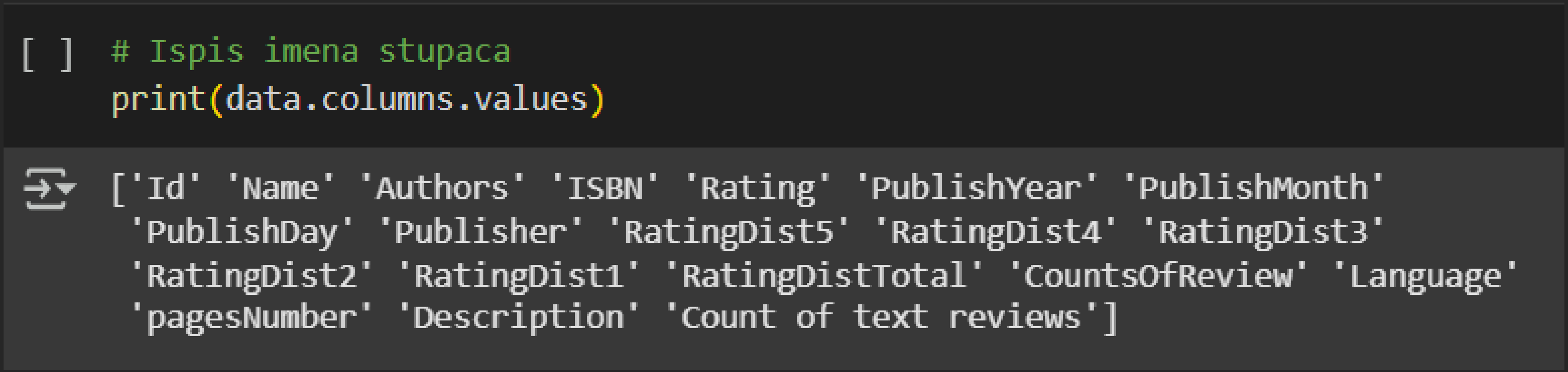
- učitavanje iz csv u dataframe (pandas)
- pregled prvih 5 redaka



```
# Učitavanje potrebnih biblioteka
import pandas as pd

# Učitavanje podataka iz CSV datoteke
PATH = "/book1000k-1100k.csv"

# Učitavanje prvih 2000 redova
data = pd.read_csv(PATH, delimiter=',', nrows=2000)

# Ispis prvih 5 redova
print(data.head())
```

```
        Id                                 Name               Authors  \
0  1000000                     Flight from Eden     Kathryn A. Graham   
1  1000001                      Roommates Again  Kathryn O. Galbraith   
2  1000003                 The King At The Door            Brock Cole   
3  1000004  Giotto: The Scrovegni Chapel, Padua            Bruce Cole   
4  1000005                          Larky Mavis            Brock Cole   

         ISBN  Rating  PublishYear  PublishMonth  PublishDay  \
0  0595199402    4.00         2001             1          10   
1  0689505973    3.20         1994             1           4   
2  0374440417    3.95         1992            31          12   
3  080761310X    4.47         1993             1           8   
4  0374343659    3.69         2001             3           8   

                         Publisher RatingDist5 RatingDist4 RatingDist3  \
0          Writer's Showcase Press         5:1         4:1         3:1   
1     Margaret K. McElderry Books         5:0         4:3         3:1   
2              Farrar Straus Giroux         5:5         4:9         3:4   
3                  George Braziller         5:9         4:5         3:0   
4  Farrar, Straus and Giroux (BYR)        5:19        4:12         3:9   

  RatingDist2 RatingDist1 RatingDistTotal  CountsOfReview Language  \
0         2:0         1:0         total:3               1      NaN   
1         2:0         1:1         total:5               1      NaN   
2         2:1         1:0        total:19               0      NaN   
3         2:1         1:0        total:15               2      NaN   
4         2:7         1:4        total:51               8      NaN   

   pagesNumber                                        Description  \
0          380  What could a computer expert, a mercenary with...   
1           44  During their stay at Camp Sleep-Away, sisters ...   
2           32  A poorly dressed old man appears at an inn and...   
3          118  This beautiful series lavishly illustrates the...   
4           32  <b>Another orginal picture-book fairy tale</b>...   

   Count of text reviews  
0                      1  
1                      1  
2                      0  
3                      2  
4                      8  
```

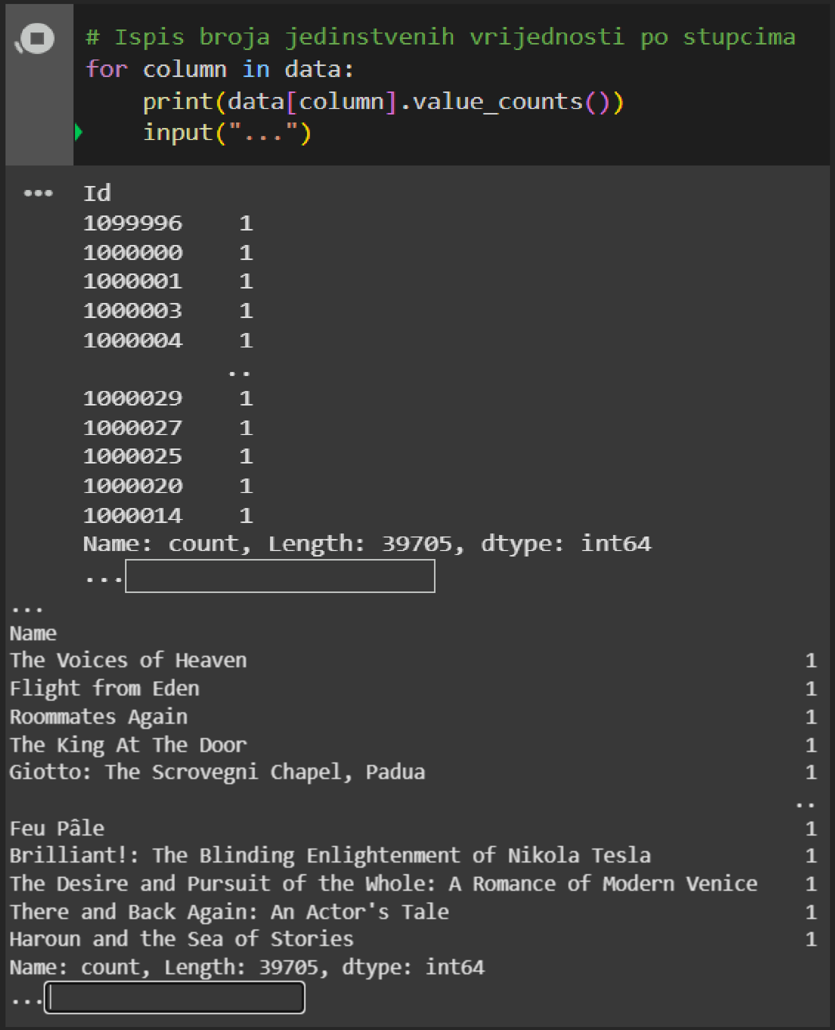
Nazivi stupaca



```
[ ]  # Ispis imena stupaca
     print(data.columns.values)
```

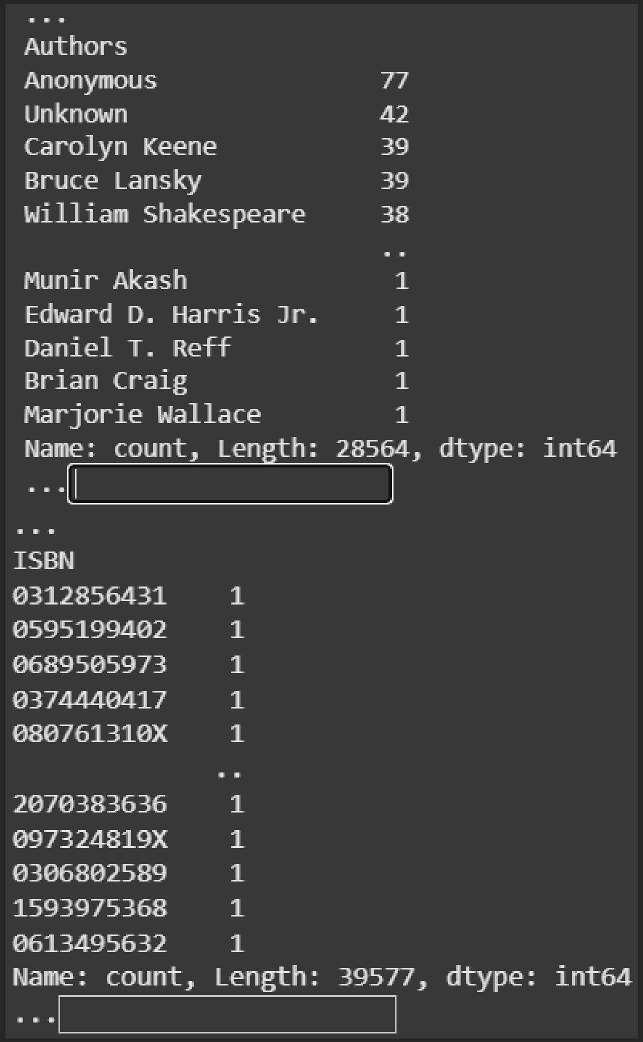
```
['Id' 'Name' 'Authors' 'ISBN' 'Rating' 'PublishYear' 'PublishMonth'
 'PublishDay' 'Publisher' 'RatingDist5' 'RatingDist4' 'RatingDist3'
 'RatingDist2' 'RatingDist1' 'RatingDistTotal' 'CountsOfReview' 'Language'
 'pagesNumber' 'Description' 'Count of text reviews']
```

Frekvencije vrijednosti po stupcu

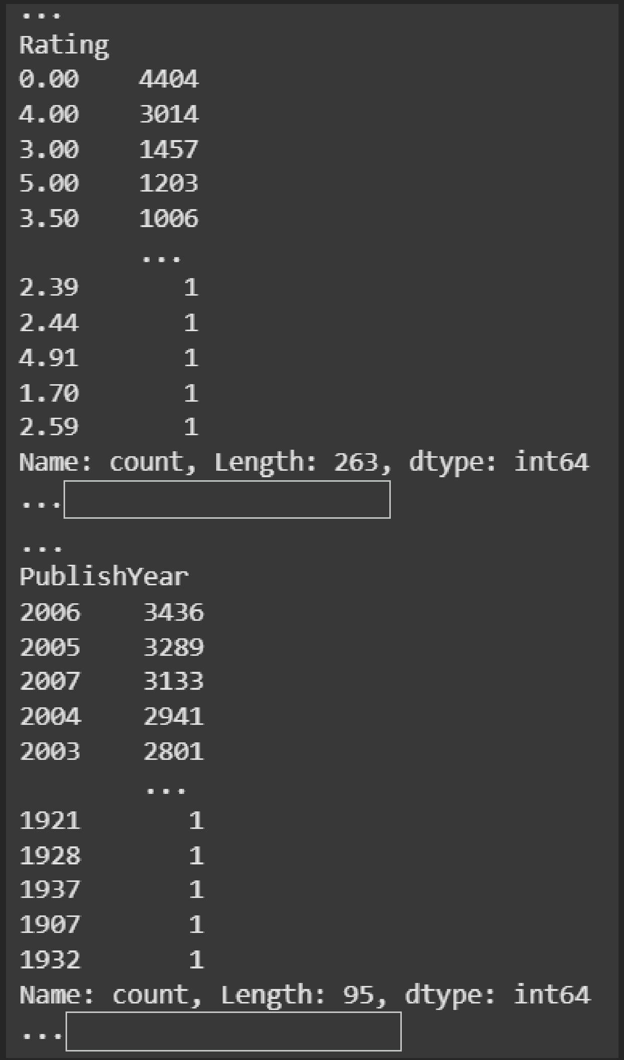


```
# Ispis broja jedinstvenih vrijednosti po stupcima
for column in data:
    print(data[column].value_counts())
    input("...")
```

```
... Id
    1099996    1
    1000000    1
    1000001    1
    1000003    1
    1000004    1
              ..
    1000029    1
    1000027    1
    1000025    1
    1000020    1
    1000014    1
    Name: count, Length: 39705, dtype: int64
    ...
...
Name
The Voices of Heaven                                       1
Flight from Eden                                           1
Roommates Again                                            1
The King At The Door                                       1
Giotto: The Scrovegni Chapel, Padua                        1
                                                          ..
Feu Pâle                                                   1
Brilliant!: The Blinding Enlightenment of Nikola Tesla     1
The Desire and Pursuit of the Whole: A Romance of Modern Venice   1
There and Back Again: An Actor's Tale                      1
Haroun and the Sea of Stories                              1
Name: count, Length: 39705, dtype: int64
...
```

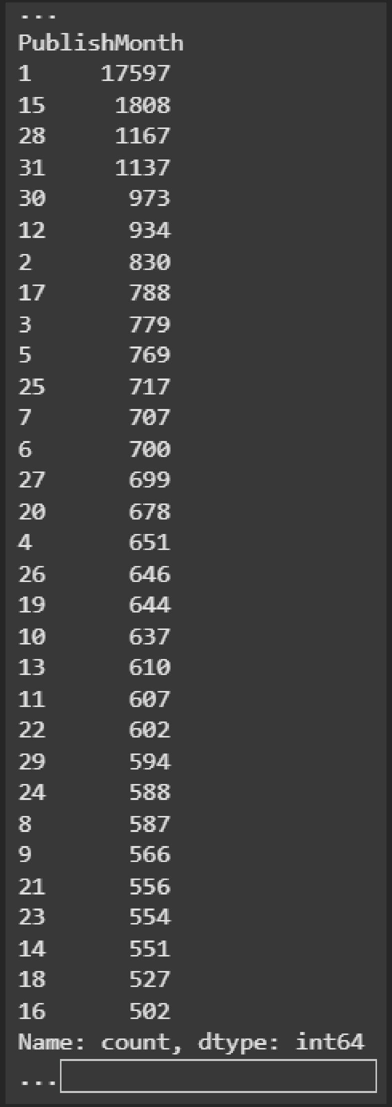


```
...
Authors
Anonymous                77
Unknown                  42
Carolyn Keene            39
Bruce Lansky             39
William Shakespeare      38
                         ..
Munir Akash               1
Edward D. Harris Jr.      1
Daniel T. Reff            1
Brian Craig               1
Marjorie Wallace          1
Name: count, Length: 28564, dtype: int64
...
...
ISBN
0312856431    1
0595199402    1
0689505973    1
0374440417    1
080761310X    1
             ..
2070383636    1
097324819X    1
0306802589    1
1593975368    1
0613495632    1
Name: count, Length: 39577, dtype: int64
...
```

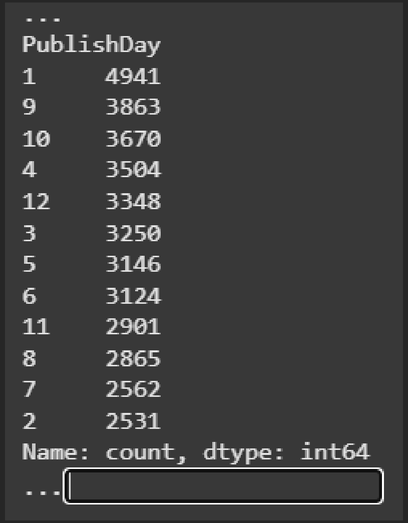


```
...
Rating
0.00    4404
4.00    3014
3.00    1457
5.00    1203
3.50    1006
        ...
2.39       1
2.44       1
4.91       1
1.70       1
2.59       1
Name: count, Length: 263, dtype: int64
...
...
PublishYear
2006    3436
2005    3289
2007    3133
2004    2941
2003    2801
        ...
1921       1
1928       1
1937       1
1907       1
1932       1
Name: count, Length: 95, dtype: int64
...
```

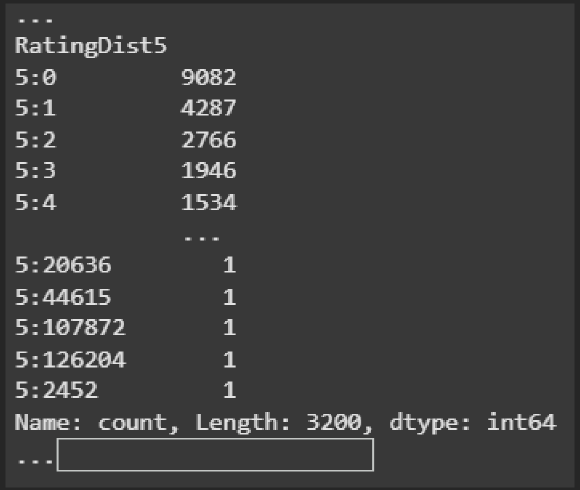
Frekvencije vrijednosti po stupcu



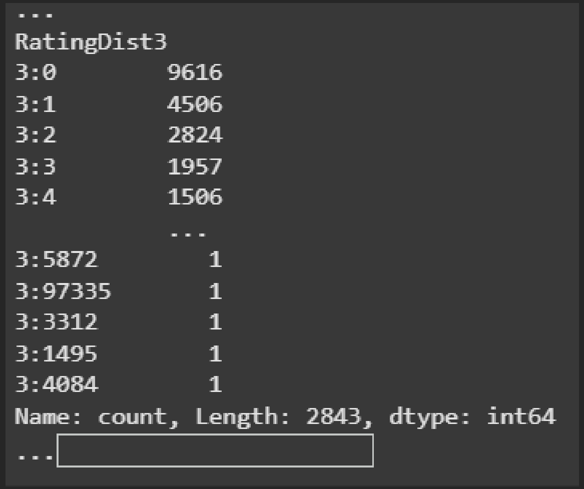
```
...
PublishMonth
1      17597
15      1808
28      1167
31      1137
30       973
12       934
2        830
17       788
3        779
5        769
25       717
7        707
6        700
27       699
20       678
4        651
26       646
19       644
10       637
13       610
11       607
22       602
29       594
24       588
8        587
9        566
21       556
23       554
14       551
18       527
16       502
Name: count, dtype: int64
...
```



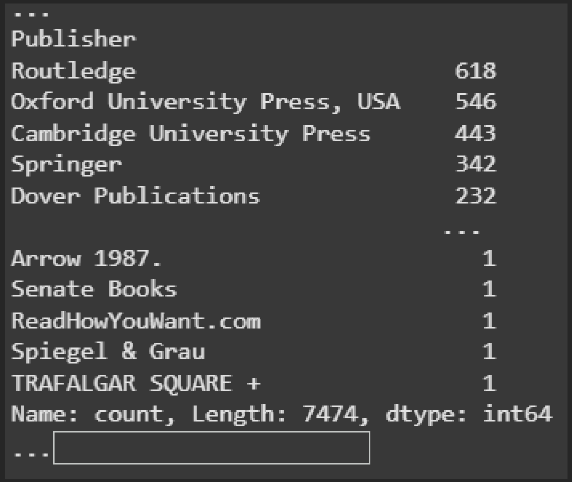
```
...
PublishDay
1     4941
9     3863
10    3670
4     3504
12    3348
3     3250
5     3146
6     3124
11    2901
8     2865
7     2562
2     2531
Name: count, dtype: int64
...
```



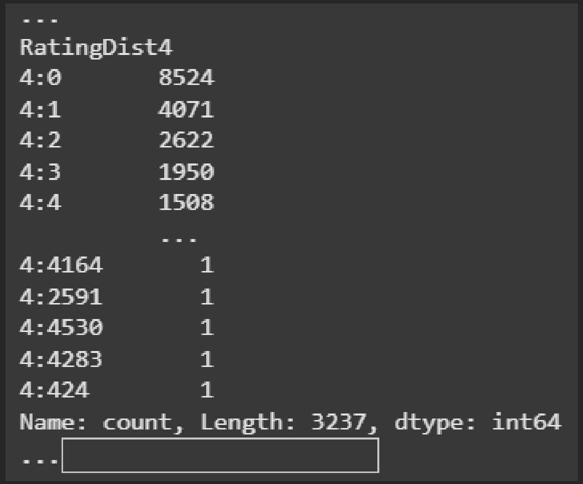
```
...
RatingDist5
5:0         9082
5:1         4287
5:2         2766
5:3         1946
5:4         1534
            ...
5:20636        1
5:44615        1
5:107872       1
5:126204       1
5:2452         1
Name: count, Length: 3200, dtype: int64
...
```



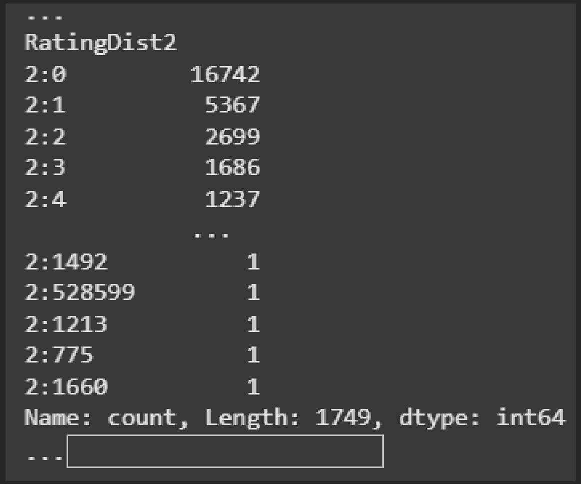
```
...
RatingDist3
3:0        9616
3:1        4506
3:2        2824
3:3        1957
3:4        1506
           ...
3:5872        1
3:97335       1
3:3312        1
3:1495        1
3:4084        1
Name: count, Length: 2843, dtype: int64
...
```



```
...
Publisher
Routledge                          618
Oxford University Press, USA       546
Cambridge University Press         443
Springer                           342
Dover Publications                 232
                                  ...
Arrow 1987.                          1
Senate Books                         1
ReadHowYouWant.com                   1
Spiegel & Grau                       1
TRAFALGAR SQUARE +                   1
Name: count, Length: 7474, dtype: int64
...
```

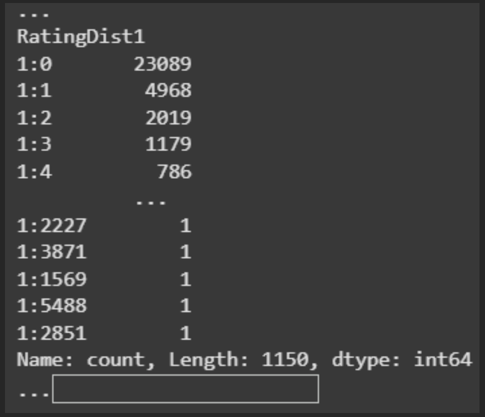


```
...
RatingDist4
4:0       8524
4:1       4071
4:2       2622
4:3       1950
4:4       1508
          ...
4:4164       1
4:2591       1
4:4530       1
4:4283       1
4:424        1
Name: count, Length: 3237, dtype: int64
...
```

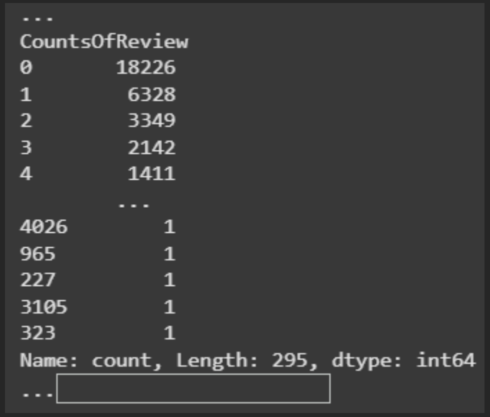


```
...
RatingDist2
2:0         16742
2:1          5367
2:2          2699
2:3          1686
2:4          1237
            ...
2:1492          1
2:528599        1
2:1213          1
2:775           1
2:1660          1
Name: count, Length: 1749, dtype: int64
...
```

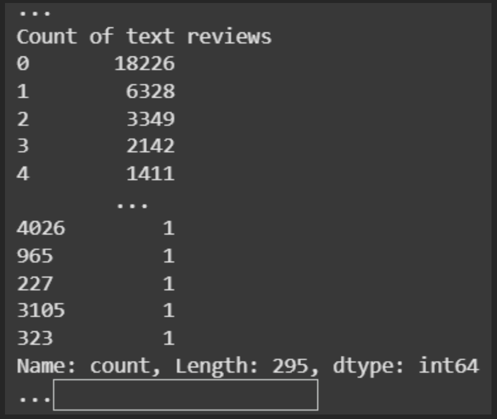
Frekvencije vrijednosti po stupcu



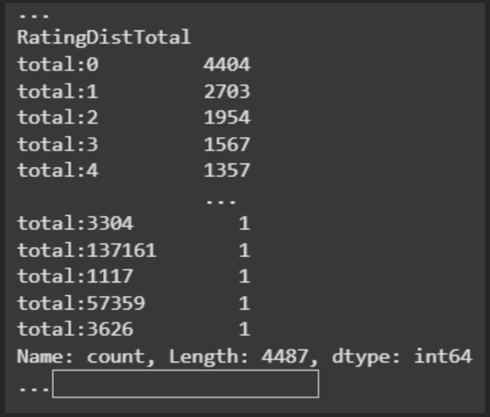
```
...
RatingDist1
1:0       23089
1:1        4968
1:2        2019
1:3        1179
1:4         786
          ...
1:2227        1
1:3871        1
1:1569        1
1:5488        1
1:2851        1
Name: count, Length: 1150, dtype: int64
...
```



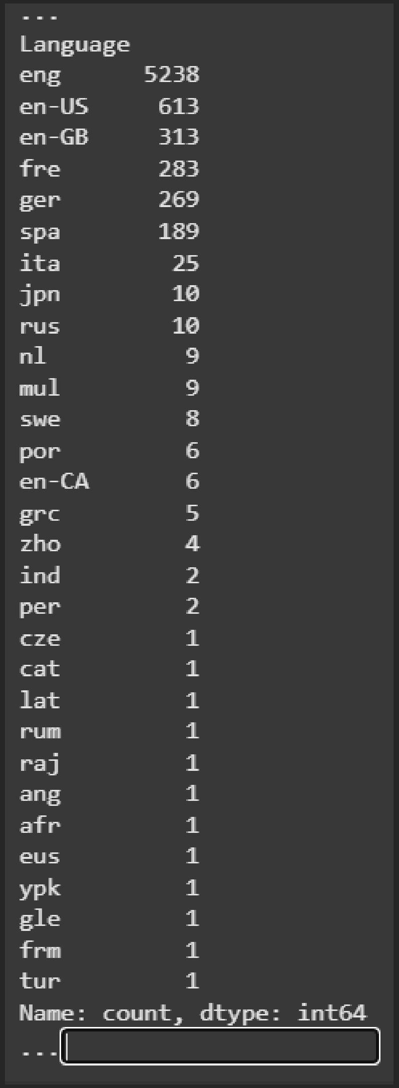
```
...
CountsOfReview
0       18226
1        6328
2        3349
3        2142
4        1411
        ...
4026        1
965         1
227         1
3105        1
323         1
Name: count, Length: 295, dtype: int64
...
```



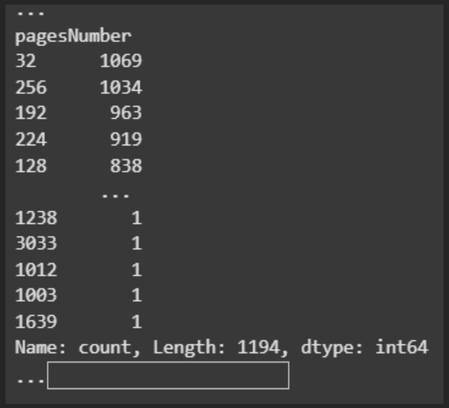
```
...
Count of text reviews
0       18226
1        6328
2        3349
3        2142
4        1411
        ...
4026        1
965         1
227         1
3105        1
323         1
Name: count, Length: 295, dtype: int64
...
```



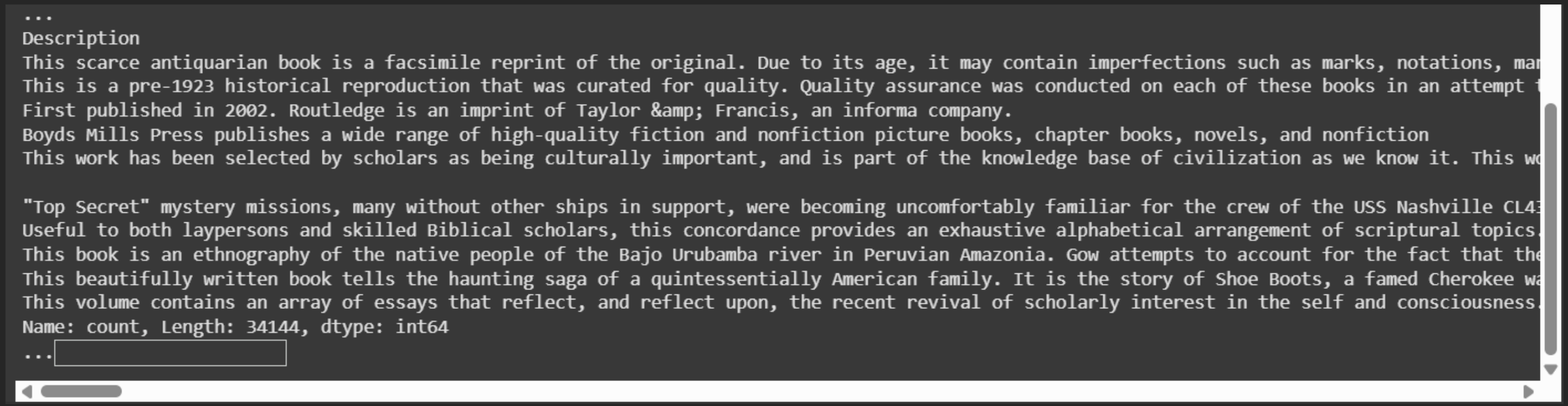
```
...
RatingDistTotal
total:0         4404
total:1         2703
total:2         1954
total:3         1567
total:4         1357
                ...
total:3304         1
total:137161       1
total:1117         1
total:57359        1
total:3626         1
Name: count, Length: 4487, dtype: int64
...
```



```
...
Language
eng      5238
en-US     613
en-GB     313
fre       283
ger       269
spa       189
ita        25
jpn        10
rus        10
nl          9
mul         9
swe         8
por         6
en-CA       6
grc         5
zho         4
ind         2
per         2
cze         1
cat         1
lat         1
rum         1
raj         1
ang         1
afr         1
eus         1
ypk         1
gle         1
frm         1
tur         1
Name: count, dtype: int64
...
```



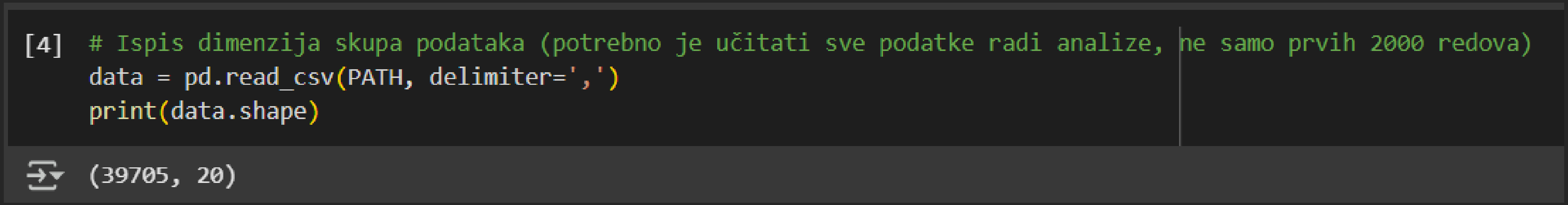
```
...
pagesNumber
32      1069
256     1034
192      963
224      919
128      838
        ...
1238       1
3033       1
1012       1
1003       1
1639       1
Name: count, Length: 1194, dtype: int64
...
```



```
...
Description
This scarce antiquarian book is a facsimile reprint of the original. Due to its age, it may contain imperfections such as marks, notations, mar
This is a pre-1923 historical reproduction that was curated for quality. Quality assurance was conducted on each of these books in an attempt t
First published in 2002. Routledge is an imprint of Taylor &amp; Francis, an informa company.
Boyds Mills Press publishes a wide range of high-quality fiction and nonfiction picture books, chapter books, novels, and nonfiction
This work has been selected by scholars as being culturally important, and is part of the knowledge base of civilization as we know it. This wo

"Top Secret" mystery missions, many without other ships in support, were becoming uncomfortably familiar for the crew of the USS Nashville CL4
Useful to both laypersons and skilled Biblical scholars, this concordance provides an exhaustive alphabetical arrangement of scriptural topics.
This book is an ethnography of the native people of the Bajo Urubamba river in Peruvian Amazonia. Gow attempts to account for the fact that the
This beautifully written book tells the haunting saga of a quintessentially American family. It is the story of Shoe Boots, a famed Cherokee wa
This volume contains an array of essays that reflect, and reflect upon, the recent revival of scholarly interest in the self and consciousness.
Name: count, Length: 34144, dtype: int64
...
```

Pitanja:

- Je li je skup podataka dovoljno velik? (Optimalno >15 000 redaka i >20 stupaca)

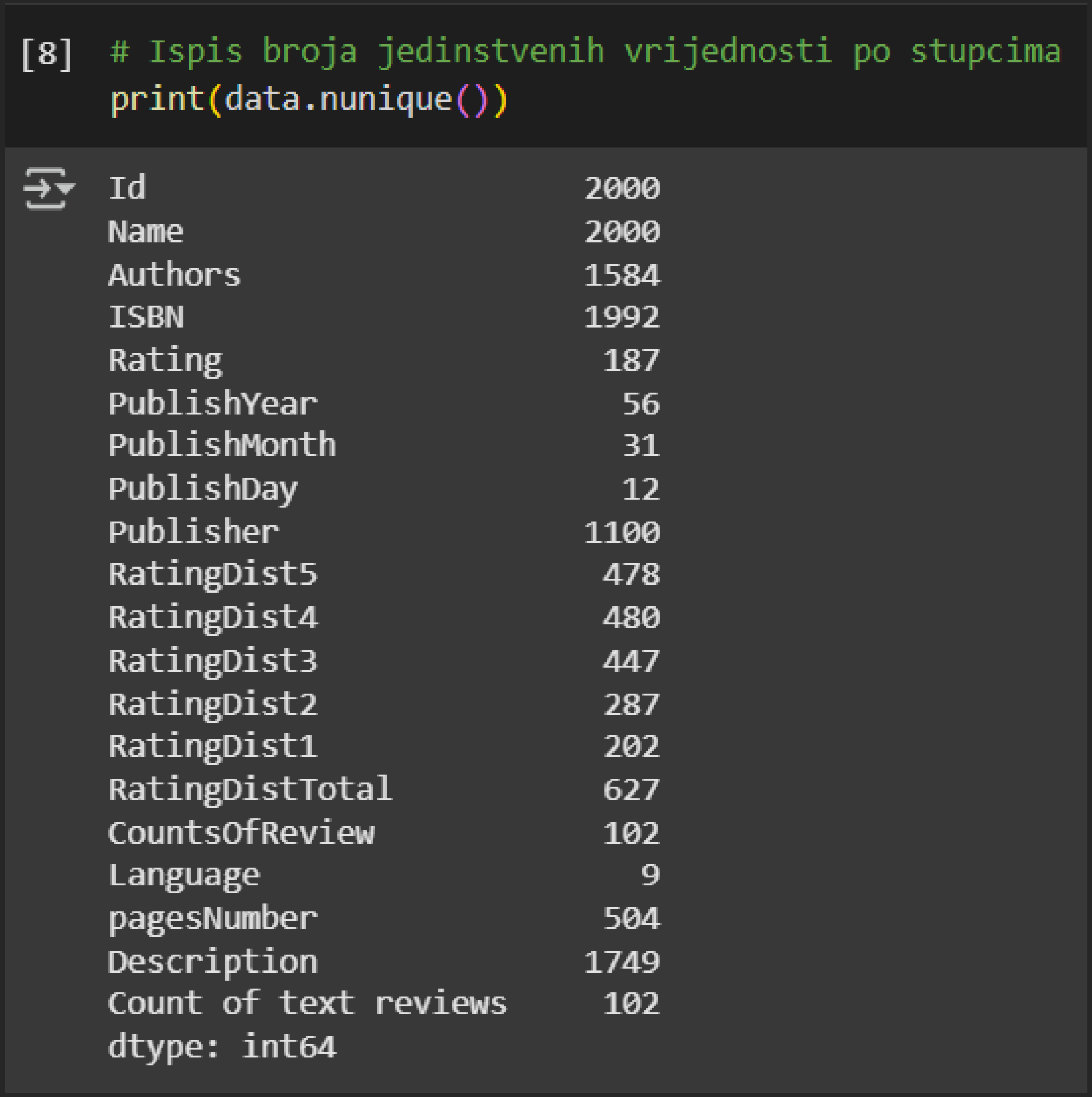


```python
# Ispis dimenzija skupa podataka (potrebno je učitati sve podatke radi analize, ne samo prvih 2000 redova)
data = pd.read_csv(PATH, delimiter=',')
print(data.shape)
```

```
(39705, 20)
```

Pitanja:

- Ima li skup dovoljno različite podatke?



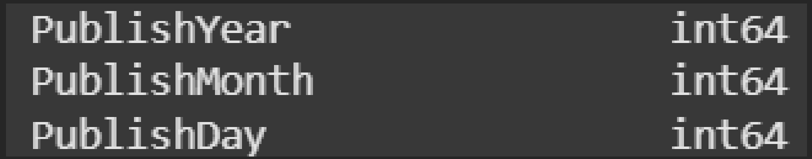
[8]

```
# Ispis broja jedinstvenih vrijednosti po stupcima
print(data.nunique())
```

```
Id                       2000
Name                     2000
Authors                  1584
ISBN                     1992
Rating                    187
PublishYear                56
PublishMonth               31
PublishDay                 12
Publisher                1100
RatingDist5               478
RatingDist4               480
RatingDist3               447
RatingDist2               287
RatingDist1               202
RatingDistTotal           627
CountsOfReview            102
Language                    9
pagesNumber               504
Description              1749
Count of text reviews     102
dtype: int64
```

Pitanja:

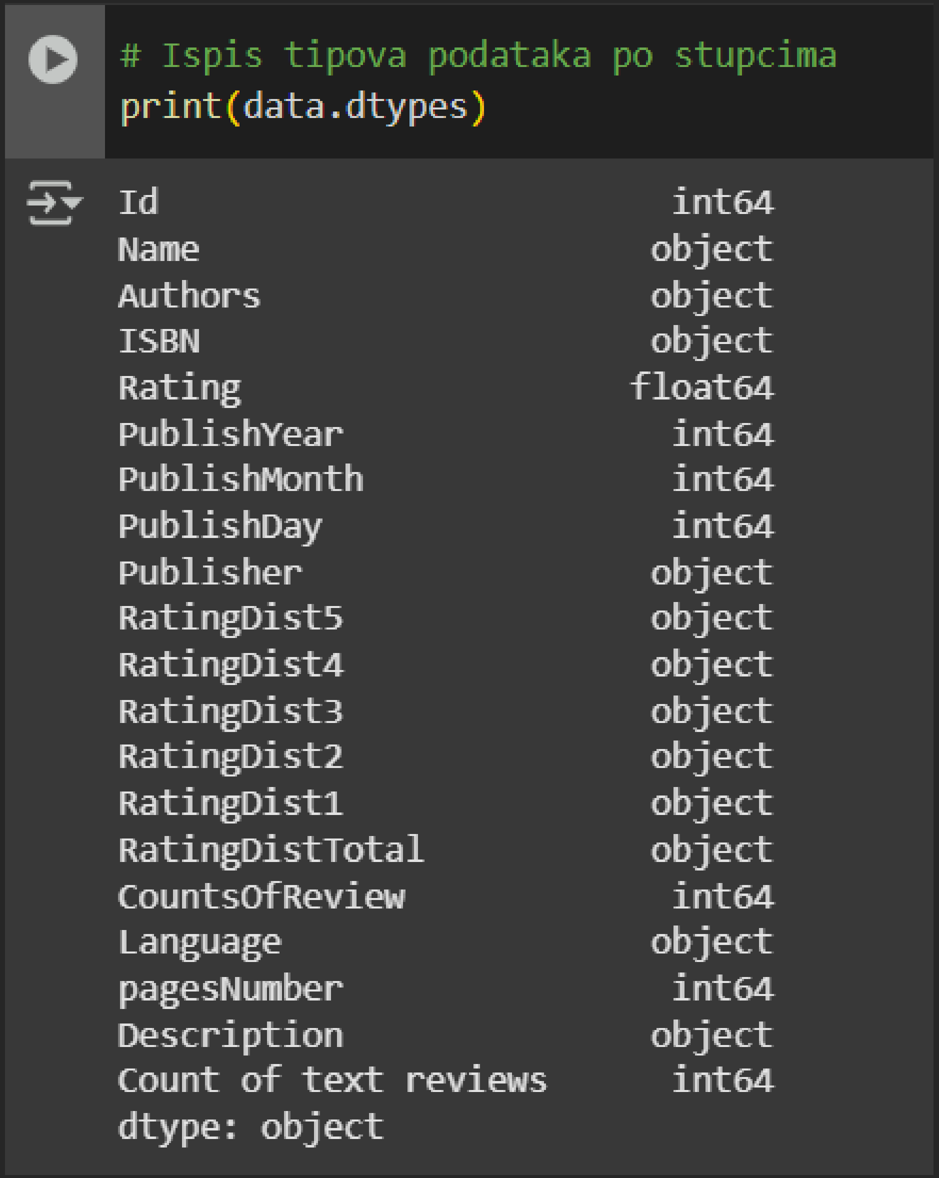
- Ima li skup vremensku dimenziju?



```
PublishYear                 int64
PublishMonth                int64
PublishDay                  int64
```

Pitanja:

- Ima li skup kvantitativne i kvalitativne podatke?
- Kvalitativni podatci (opisni, nemjerljivi) → Name, Authors, ISBN,Publisher…
- Kvantitativni podatci (numerički, mjerljivi) → Rating, Count of text reviews, pagesNumber…

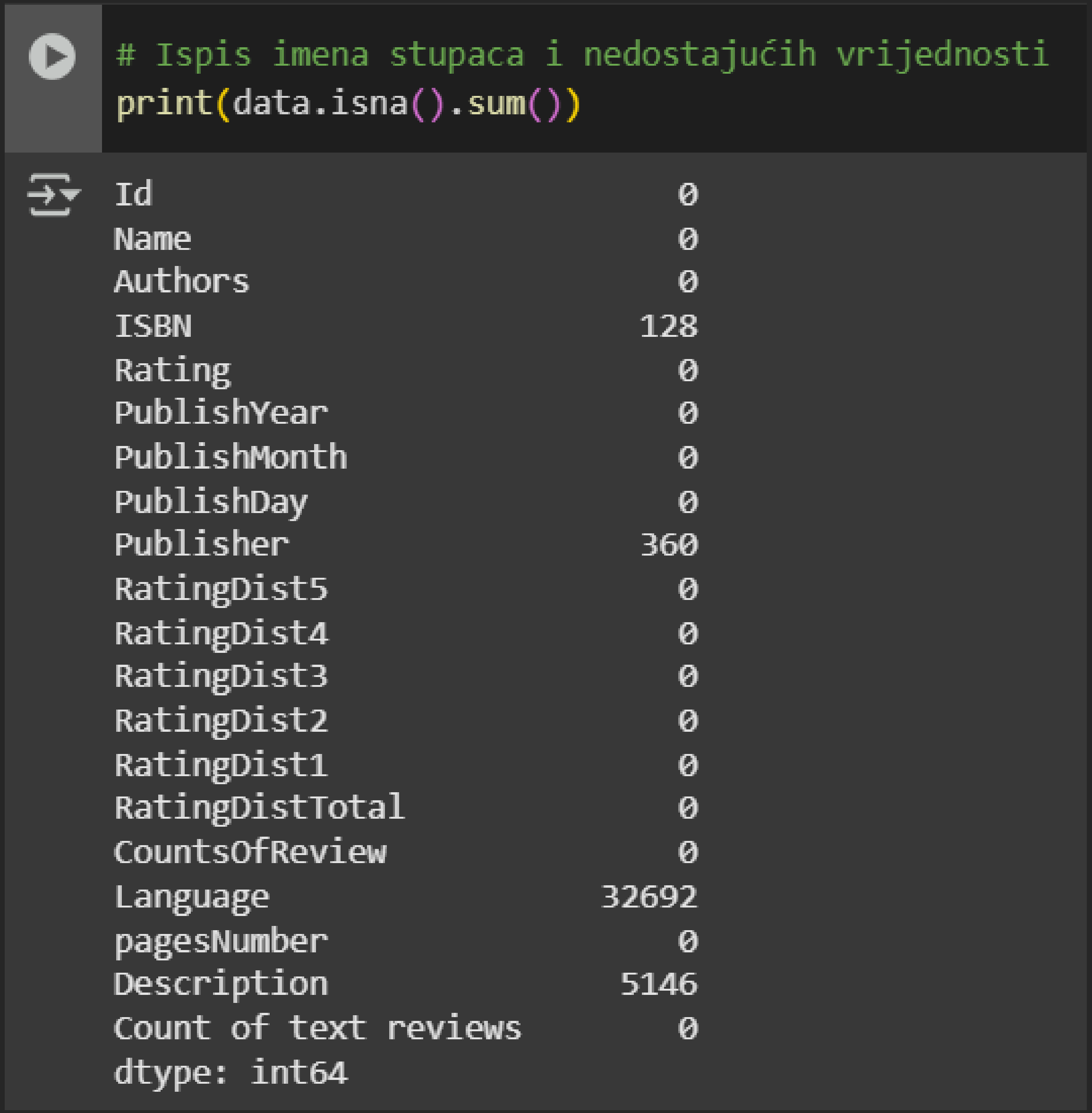


```python
# Ispis tipova podataka po stupcima
print(data.dtypes)
```

```
Id                         int64
Name                      object
Authors                   object
ISBN                      object
Rating                   float64
PublishYear                int64
PublishMonth               int64
PublishDay                 int64
Publisher                 object
RatingDist5               object
RatingDist4               object
RatingDist3               object
RatingDist2               object
RatingDist1               object
RatingDistTotal           object
CountsOfReview             int64
Language                  object
pagesNumber                int64
Description               object
Count of text reviews      int64
dtype: object
```

Pitanja:

- Ima li skup puno nedostajućih vrijednosti?



```python
# Ispis imena stupaca i nedostajućih vrijednosti
print(data.isna().sum())
```

```
Id                          0
Name                        0
Authors                     0
ISBN                      128
Rating                      0
PublishYear                 0
PublishMonth                0
PublishDay                  0
Publisher                 360
RatingDist5                 0
RatingDist4                 0
RatingDist3                 0
RatingDist2                 0
RatingDist1                 0
RatingDistTotal             0
CountsOfReview              0
Language                32692
pagesNumber                 0
Description              5146
Count of text reviews       0
dtype: int64
```