cwy_ 1_part2

March 27, 2023

[ ]:
```
# Movies Dataset from Pirated Sites
!wget -O data/data.zip "https://storage.googleapis.com/kaggle-data-sets/2922294/
 ↪5035020/compressed/movies_dataset.csv.zip?
 ↪X-Goog-Algorithm=GOOG4-RSA-SHA256&X-Goog-Credential=gcp-kaggle-com%40kaggle-161607.
 ↪iam.gserviceaccount.
 ↪com%2F20230327%2Fauto%2Fstorage%2Fgoog4_request&X-Goog-Date=20230327T091732Z&X-Goog-Expires
!unzip -o data/data.zip -d data/
```

```
--2023-03-27 21:59:04--  https://storage.googleapis.com/kaggle-data-
sets/2922294/5035020/compressed/movies_dataset.csv.zip?X-Goog-
Algorithm=GOOG4-RSA-SHA256&X-Goog-Credential=gcp-kaggle-com%40kaggle-161607.iam.
gserviceaccount.com%2F20230327%2Fauto%2Fstorage%2Fgoog4_request&X-Goog-
Date=20230327T091732Z&X-Goog-Expires=259200&X-Goog-SignedHeaders=host&X-Goog-Sig
nature=48b6af847a1e1fc622f2e7683240150828e9947bfd5f4f3096d058dda933afe810d3ea2ac
e821dbff97549dc9c8eb7152fed0fdbe2059363338700a8b0d8015dc525e63b3579372ab43568322
838f57c3f257f5ba34998b897a4cb61476163311b17cbc76c1ffa6cb35532ab769caca3324205bdc
450518868eb6978962a91299a7a429267f4e560fe11c31caa110d348ccd4934a6d97ecc8e0d0aee1
fbd7146ab0f7939050a48a13c688ca2753b0766e27d22efdd94b9c06e4588efec00754b1ad00df4d
6a9380ee27852ca69f64ef1444dab2a5ede76b8f6495869b9cd3d0019821e58820aa4777515818c1
667ceb382f2ab65b5de07152f3e7e678fd10c70
Resolving storage.googleapis.com (storage.googleapis.com)… 216.58.200.240,
142.251.222.208, 142.250.66.80, …
Connecting to storage.googleapis.com
(storage.googleapis.com)|216.58.200.240|:443… connected.
HTTP request sent, awaiting response… 200 OK
Length: 3713705 (3.5M) [application/zip]
Saving to: ‘data/data.zip’

data/data.zip       100%[===================>]   3.54M  11.0MB/s    in 0.3s

2023-03-27 21:59:05 (11.0 MB/s) - ‘data/data.zip’ saved [3713705/3713705]

Archive:  data/data.zip
  inflating: data/movies_dataset.csv
```

[ ]:
```
import pandas as pd
import matplotlib.pyplot as plt
```

```python
#
df = pd.read_csv("data/movies_dataset.csv")

print(df.head(5))
#
print("   ")
print(df.describe())
```

```
   Unnamed: 0  IMDb-rating appropriate_for      director downloads      id  \
0           0          4.8               R     John Swab       304  372092
1           1          6.4           TV-PG   Paul Ziller        73  372091
2           2          5.2               R  Ben Wheatley     1,427  343381
3           3          8.1             NaN  Venky Atluri     1,549  372090
4           4          4.6             NaN  Shaji Kailas       657  372089

              industry       language   posted_date release_date  run_time  \
0  Hollywood / English        English  20 Feb, 2023  Jan 28 2023       105
1  Hollywood / English        English  20 Feb, 2023  Feb 05 2023        84
2  Hollywood / English  English,Hindi  20 Apr, 2021  Jun 18 2021  1h 47min
3           Tollywood          Hindi  20 Feb, 2023  Feb 17 2023       139
4           Tollywood          Hindi  20 Feb, 2023  Jan 26 2023       122

                                          storyline  \
0  Doc\r\n facilitates a fragile truce between th…
1  Caterer\r\n Goldy Berry reunites with detectiv…
2  As the world searches for a cure to a disastro…
3  The life of a young man and his struggles agai…
4  A man named Kalidas gets stranded due to the p…

                                    title   views                 writer
0                            Little Dixie   2,794              John Swab
1  Grilling Season: A Curious Caterer Mystery   1,002  John Christian Plummer
2                            In the Earth  14,419           Ben Wheatley
3                                  Vaathi   4,878           Venky Atluri
4                                   Alone   2,438       Rajesh Jayaraman

         Unnamed: 0   IMDb-rating             id
count  20548.000000  19707.000000   20548.000000
mean   10273.500000      5.762151  222351.199776
std     5931.841001      1.374041  138422.327931
min        0.000000      1.100000       1.000000
25%     5136.750000      4.800000   96122.250000
50%    10273.500000      5.700000  264457.500000
75%    15410.250000      6.600000  354561.250000
max    20547.000000      9.900000  372092.000000
```

[ ]:
```
#
for column in df.columns:
    if df[column].dtype == 'object':
        print(f"\n   {column}   ")
        print(df[column].value_counts())
```

```
   appropriate_for
R                 4384
Not Rated         2142
PG-13             1968
PG                 886
TV-14              694
TV-MA              406
G                  152
Unrated            132
TV-PG              115
TV-G                99
TV-Y7               45
TV-Y                25
Approved             9
NC-17                4
TV-Y7-FV             3
Passed               3
MA-17                1
TV-13                1
Drama                1
Drama, Romance       1
18+                  1
Name: appropriate_for, dtype: int64

   director
Venky Atluri                                   405
Simone Stock                                   403
Xavier Manrique                                403
John Swab                                      205
Neil Jordan                                    205
                                              …
Agnieszka Smoczynska                             1
Dylan Thomas Ellis                               1
Sunil Thakur, Sunil Dhawan, Shivani Thakur       1
Suman Mukhopadhyay                               1
Shea Sizemore                                    1
Name: director, Length: 9672, dtype: int64

   downloads
75        403
```

```
622       212
378       209
1,782     187
466       170
         …
34,781      1
3,721       1
13,947      1
51,963      1
3,276       1
Name: downloads, Length: 10625, dtype: int64
```

```
   industry
Hollywood / English    14649
Bollywood / Indian      2645
Tollywood               1172
Anime / Kids            1049
Wrestling                433
Punjabi                  332
Stage shows              129
Pakistani                 92
Dub / Dual Audio          45
3D Movies                  1
Name: industry, dtype: int64
```

```
   language
English                                  12657
Hindi                                     2558
English,Spanish                            391
Punjabi                                    310
English,Hindi                              304
                                          …
English,Korean,Spanish                       1
Norwegian,Swedish                            1
Spanish,Chinese,English,Maori,French         1
Urdu,Punjabi,English                         1
Spanish,German,English                       1
Name: language, Length: 1168, dtype: int64
```

```
   posted_date
13 Feb, 2023    812
20 Feb, 2023    607
15 Feb, 2023    607
10 Feb, 2023    485
16 Feb, 2023    406
               …
12 Sep, 2009      1
08 Sep, 2009      1
```

```
01 Sep, 2009      1
18 Aug, 2009      1
30 Nov, 2011      1
Name: posted_date, Length: 4123, dtype: int64

   release_date
Jan 01 1970    962
Feb 03 2023    616
Feb 17 2023    607
Feb 10 2023    410
Feb 11 2023    402
               …
Sep 05 2003      1
Dec 29 2022      1
Aug 24 2013      1
Jan 12 2014      1
Mar 28 1958      1
Name: release_date, Length: 4886, dtype: int64

   run_time
93          652
88          622
101         568
139         454
95          454
           …
74 min        1
288           1
220           1
49min         1
3h 13min      1
Name: run_time, Length: 415, dtype: int64

   storyline
The life of a young man and his struggles against the privatization of
education.
402
Follows\r\n a New York City family hiding out in the Hamptons whose bubble is
\r\npopped when a Bloody Mary-swilling, pot-smoking 'Charlie' comes to bring\r\n
a lifetime of hurt that might heal them all.
402
It follows Kara Robinson as she survives an abduction and ultimately brings down
a serial killer.
402
Doc\r\n facilitates a fragile truce between the Governor and Cartel, trading
\r\nprosecutorial leniency for finance. With no more truce, Doc is left to
\r\nfend for himself and protect the one untainted thing in his life: his
\r\ndaughter, Little Dixie.
```

```
202
A\r\n young, gay Black man, rejected by his mother and with few options for
\r\nhis future, decides to join the Marines, doing whatever it takes to
\r\nsucceed in a system that would cast him aside.
202
                                                                        …
Four waves of increasingly deadly attacks have left most of Earth in ruin.
Against a backdrop of fear and distrust, Cassie is on the run, desperately
trying to save her younger brother. As she prepares for the inevitable and
lethal fifth wave, Cassie teams up with a young man who may become her final
hope - if she can only trust him.
1
Yamuna along with her son Laxman locates to Mumbai leaving behind her abusive
husband.She takes shelter in the house of her aunt Chandra whom she calls
Akka.Yamuna's only aim is to give a better education to her son.Chandra finds
her a job as sweeper in a art school.Yamuna finds that Chandra poses as a nude
model to the students of the school.Chandra confines Yamuna to take up the job
being nude out there the students don't look at you in lust but as a project.
1
A young violinist struggles to assert her individuality amidst the intense
pressure of her pianist father, and the weight of her own musical ability.
1
A right wing talk show host's life takes a sudden turn when his 16 year old
niece comes crashing into his life.
1
While driving his car on a rainy night, Anand's car breaks down, and he goes to
seek shelter in a nearby house. He is let into the house by the servant, and he
is permitted to stay until the rains stop be able to get his car fixed. It is
here that he will find out about his previous birth, his true love, Madhumati,
their ill-fated, star-crossed and tragic romance, and how events in his previous
birth are going to effect him in this life-time.                         1
Name: storyline, Length: 15748, dtype: int64

   title
The Girl Who Escaped: The Kara Robinson Story    402
Vaathi                                           402
Who Invited Charlie?                             402
Little Dixie                                     202
The Inspection                                   202
                                                …
Kesari                                             1
Old Boys                                           1
American Exit                                      1
Adventures of Aladdin                              1
Madhumati                                          1
Name: title, Length: 16572, dtype: int64

   views
```

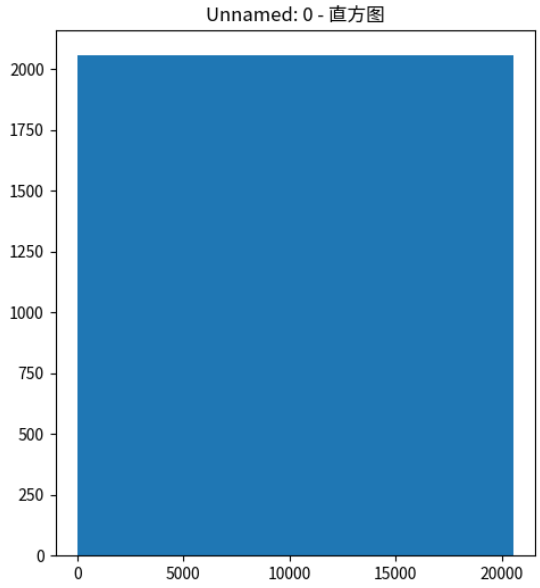
```
6,259      6
5,926      5
5,859      5
5,603      5
4,288      5
          ..
142,470    1
104,507    1
152,768    1
6,460      1
1,419      1
Name: views, Length: 16821, dtype: int64

   writer
Nicholas Schutt                             403
Venky Atluri                                402
Haley Harris                                402
John Swab                                   205
Elegance Bratton                            202
                                           ...
Barbara Samuels, Joseph Boyden                1
Maria Allred                                  1
Pia Mechler                                   1
Paul Flannery, David Ryan Keith               1
Khwaja Ahmad Abbas, Khwaja Ahmad Abbas        1
Name: writer, Length: 13603, dtype: int64
```

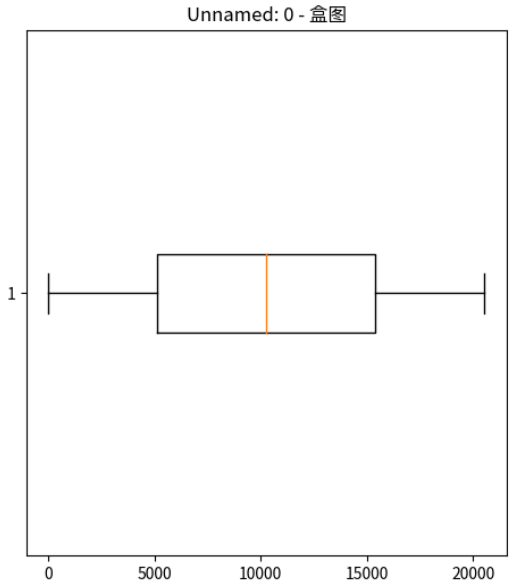
[ ]:
```python
#    5
num_attributes = df.select_dtypes(include=['number'])
for column in num_attributes.columns:
    print(f"\n   {column} 5  ")
    print(df[column].describe())
    print(f"  {column}     {df[column].isna().sum()}")
```

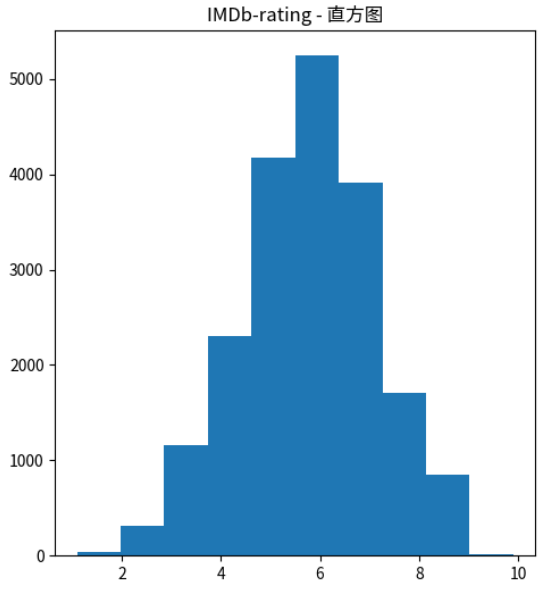
```
   Unnamed: 0  5
count    20548.000000
mean     10273.500000
std       5931.841001
min          0.000000
25%       5136.750000
50%      10273.500000
75%      15410.250000
max      20547.000000
Name: Unnamed: 0, dtype: float64
  Unnamed: 0     0

   IMDb-rating  5
```

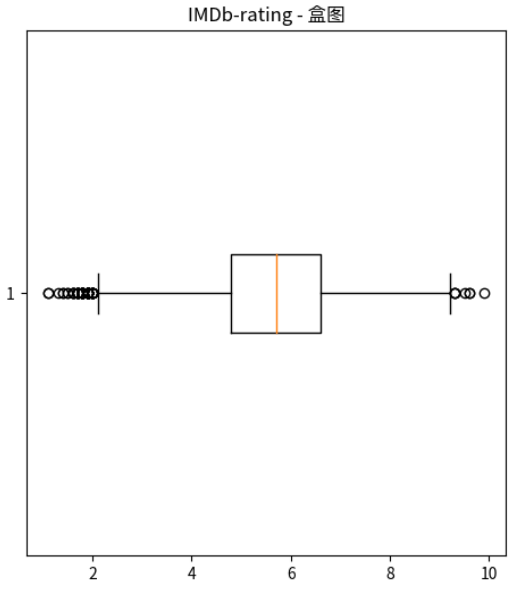
```
count    19707.000000
mean         5.762151
std          1.374041
min          1.100000
25%          4.800000
50%          5.700000
75%          6.600000
max          9.900000
Name: IMDb-rating, dtype: float64
  IMDb-rating     841

   id  5
count     20548.000000
mean     222351.199776
std      138422.327931
min           1.000000
25%       96122.250000
50%      264457.500000
75%      354561.250000
max      372092.000000
Name: id, dtype: float64
  id     0
```

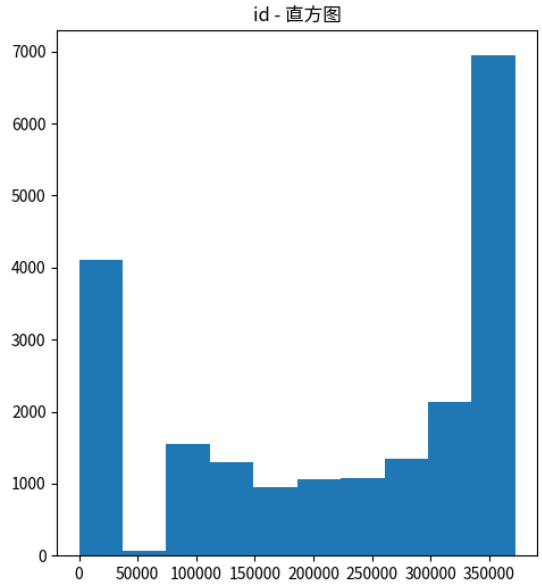
[ ]:
```
from pathlib import Path
import matplotlib as mpl
from matplotlib import font_manager
font_path = Path('/usr/share/fonts/opentype/noto')

#   matplotlib
mpl.rcParams['font.family'] = font_manager.FontProperties(fname="/usr/share/
 ↪fonts/opentype/noto/NotoSansCJK-Regular.ttc").get_name()
mpl.rcParams['axes.unicode_minus'] = False

#     -
for column in num_attributes.columns:
    plt.figure(figsize=(12, 6))

    plt.subplot(121)
    plt.hist(df[column].dropna())
    plt.title(f"{column} -   ")

    plt.subplot(122)
    plt.boxplot(df[column].dropna(), vert=False)
    plt.title(f"{column} -   ")

    plt.show()
```

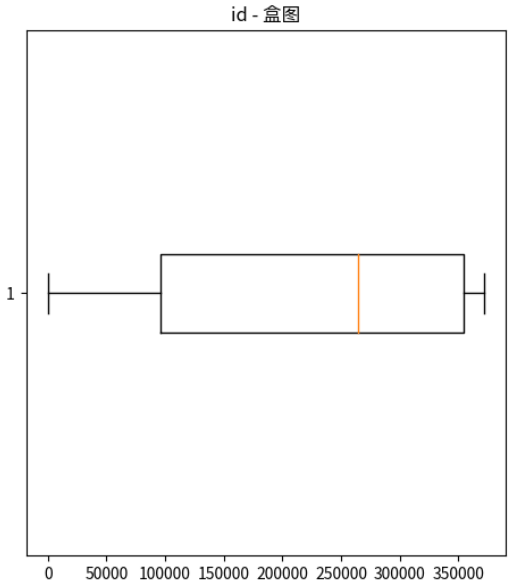

Unnamed: 0 - 直方图
2000
1750
1500
1250
1000
750
500
250
0
0
5000
10000
15000
20000


Unnamed: 0 - 盒图
1
0
5000
10000
15000
20000

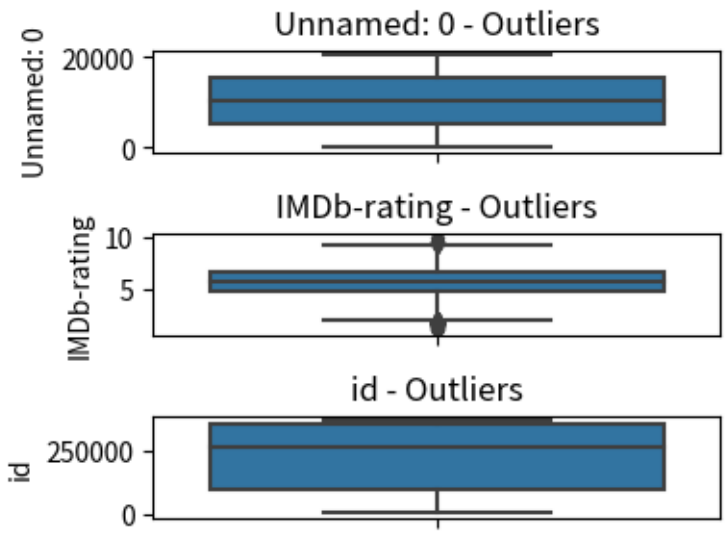

IMDb-rating - 直方图
5000
4000
3000
2000
1000
0
2
4
6
8
10


IMDb-rating - 盒图
1
2
4
6
8
10





[ ]:
```python
import seaborn as sns

#
numeric_columns = df.select_dtypes(include=['number']).columns

#        boxplot
n_plots = len(numeric_columns)
fig, axes = plt.subplots(nrows=n_plots, ncols=1, figsize=(4, 1 * n_plots))

for idx, column in enumerate(numeric_columns):
    sns.boxplot(y=df[column], ax=axes[idx])
    axes[idx].set_title(f'{column} - Outliers')
    axes[idx].set_ylabel(column)

#
plt.tight_layout()
plt.show()
```



[ ]:
```
#
missing_values = df.isnull().sum()
print(missing_values)
```

```
Unnamed: 0            0
IMDb-rating         841
appropriate_for    9476
director           1938
downloads             1
id                    0
industry              1
language            542
posted_date           1
release_date          1
run_time           1768
storyline          1701
title                 1
views                 1
writer             2192
dtype: int64
```

[ ]:
```
# uncommon  strategy
#

# strategy="del"
# strategy="max_fre"
# strategy="atri_rela"  #
strategy="sim"    #
if strategy=="del":
```

```
    #
    df_dropna = df.dropna()
    print(df_dropna.describe())
    print(df.compare(df_dropna))
elif strategy=="max_fre":
    #
    df_filled_max_frequency = df.copy()
    for column in df_filled_max_frequency.columns:
        column_mode = df_filled_max_frequency[column].mode()
        if not column_mode.empty:
            most_frequent_value = column_mode.iloc[0]
            df_filled_max_frequency[column] = df_filled_max_frequency[column].
 ↪fillna(most_frequent_value)
    print(df_filled_max_frequency.describe())
    print(df.compare(df_filled_max_frequency))
elif strategy=="atri_rela":
    from sklearn.linear_model import LinearRegression
    # replace
    df['downloads'] = df['downloads'].str.replace(',', '')
    df = df[df['downloads'].notna()]
    #
    df['downloads'] = df['downloads'].astype(int)
    #
    df_regression = df.dropna(subset=['downloads', 'IMDb-rating'])
    #
    X = df_regression['downloads'].values.reshape(-1, 1)
    y = df_regression['IMDb-rating']
    model = LinearRegression().fit(X, y)

    #
    missing_rows = df['IMDb-rating'].isnull()
    missing_data_values = df.loc[missing_rows, 'downloads'].values.reshape(-1,␣
 ↪1)
    predicted_values = model.predict(missing_data_values)
    df_filled_regression = df.copy()
    df_filled_regression.loc[missing_rows, 'IMDb-rating'] = predicted_values
    print(df_filled_regression.describe())
    print(df.compare(df_filled_regression))
    pass
elif strategy=="sim":
    from sklearn.experimental import enable_iterative_imputer
    from sklearn.impute import SimpleImputer, KNNImputer
    #
    mapping_dict = {}

    #           SimpleImputer
    for col in df:
```

```
        df[col] = df[col].astype('category')
        mapping_dict[col] = dict(enumerate(df[col].cat.categories))
        df[col] = df[col].cat.codes

    #  SimpleImputer
    # numerical_imputer = SimpleImputer(strategy='mean')
    # imputed_data = numerical_imputer.fit_transform(df)
    # df_filled = pd.DataFrame(imputed_data, columns=df.columns)
    #  KNNImputer
    knn_imputer = KNNImputer(n_neighbors=5)
    imputed_data = knn_imputer.fit_transform(df)
    df_filled = pd.DataFrame(imputed_data, columns=df.columns)
    #
    for col in mapping_dict:
        reversed_mapping = {v: k for k, v in mapping_dict[col].items()}
        df_filled[col] = df_filled[col].round(0).astype('int').
↪map(reversed_mapping).astype('category')
    print(df_filled.describe())
    print(df.compare(df_filled))
    pass
```

```
      IMDb-rating       appropriate_for       director       downloads        \
             self other             self other     self other      self other
0              36   NaN              11   NaN     4409   NaN      5974   NaN
1              52   NaN              16   NaN     6903   NaN      9487   NaN
2              40   NaN              11   NaN     1071   NaN       405   NaN
3              69   NaN              -1   NaN     9322   NaN       517   NaN
4              34   NaN              -1   NaN     8262   NaN      8918   NaN
…             …    …                …    …       …    …        …    …
20543          -1   NaN              -1   NaN       -1   NaN       930   NaN
20544          65   NaN              -1   NaN     1175   NaN      8391   NaN
20545          68   NaN              -1   NaN       -1   NaN      5427   NaN
20546          -1   NaN              -1   NaN       -1   NaN      5979   NaN
20547          -1   NaN              -1   NaN       -1   NaN      3832   NaN

          id          … run_time       storyline        title        views  \
        self    other  …     self other      self other   self other   self
0      17085      NaN  …       12   NaN      6512   NaN   8014   NaN   5503
1      17084      NaN  …      383   NaN      5956   NaN   5672   NaN      0
2      14223      NaN  …      195   NaN      5221   NaN   6599   NaN   2937
3      17083      NaN  …       75   NaN     12955   NaN  15653   NaN  10619
4      17082      NaN  …       46   NaN      1929   NaN    989   NaN   5392
…      …      …   …       …    …        …    …     …    …     …
20543   3883      NaN  …       -1   NaN        -1   NaN   2109   NaN  13131
20544   3884      NaN  …      106   NaN      6453   NaN   3965   NaN   3949
20545   3953   1363.0  …      199   NaN     15438   NaN   8348   NaN  14198
20546  17038      NaN  …       -1   NaN        -1   NaN  12547   NaN     88
```

```
20547  17055     NaN  …       -1   NaN        -1   NaN  15863   NaN  13420
```

```
            writer
      other   self other
0       NaN   6231   NaN
1       NaN   6093   NaN
2       NaN   1461   NaN
3       NaN  13092   NaN
4       NaN  10204   NaN
…     …    …   …
20543   NaN     -1   NaN
20544   NaN  10212   NaN
20545   NaN     -1   NaN
20546   NaN     -1   NaN
20547   NaN     -1   NaN

[20548 rows x 28 columns]
```

[ ]: