# Movies Dataset from Pirated Sites

March 18, 2024

[1]:
```
import pandas as pd
import warnings
import numpy as np
import matplotlib.pyplot as plt
warnings.filterwarnings('ignore')
```

[2]:
```
file = 'input/movies_dataset.csv'
data = pd.read_csv(file)
```

3.1

5

[3]:
```
Nominal_Attribute = ['director', 'industry', 'language', 'title', 'writer']
print('   :')
for attribute in Nominal_Attribute:
    print('--------------------------------------------------')
    print(attribute + ":")
    print(data[attribute].value_counts())
```

```
  :
--------------------------------------------------
director:
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: count, Length: 9672, dtype: int64
--------------------------------------------------
industry:
```

```
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: count, dtype: int64
-------------------------------------------------
language:
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: count, Length: 1167, dtype: int64
-------------------------------------------------
title:
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: count, Length: 16572, dtype: int64
-------------------------------------------------
writer:
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: count, Length: 13603, dtype: int64
```

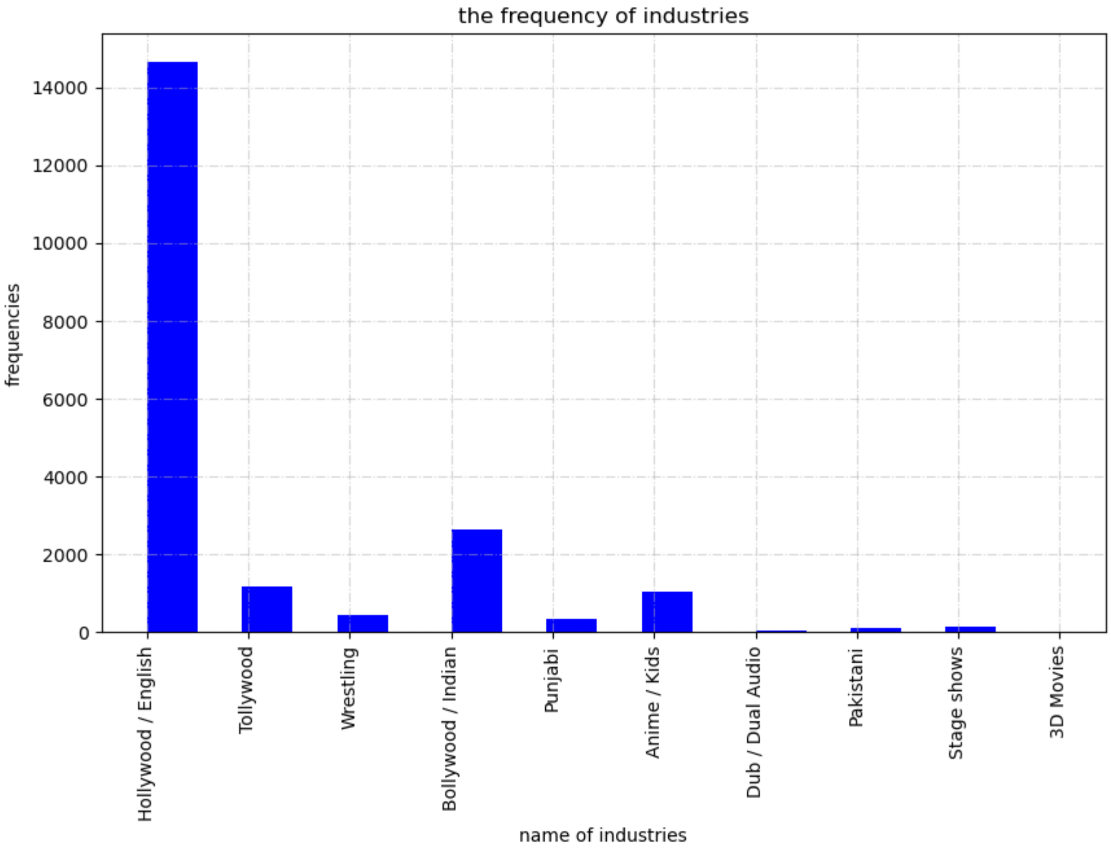
[4]:

```
Number_Attribute = ['IMDb-rating', 'downloads', 'id', 'views']
def five_number(data):
    Min_num = min(data)
    Max_num = max(data)
    Q1_num = np.percentile(data, 25)
    Median_num = np.median(data)
    Q3_num = np.percentile(data, 75)
    return Min_num, Max_num, Q1_num, Median_num, Q3_num

print('   :')

data['downloads'] = data['downloads'].str.replace(',', '').astype(float)
data['views'] = data['views'].str.replace(',', '').astype(float)

for attribute in Number_Attribute:
    print('-------------------------------------------------')
    print(attribute + ":")
    print('    ')
    print(data[attribute].isnull().sum())
    print('   :')
    print(five_number(data.dropna(subset=[attribute])[attribute]))
```

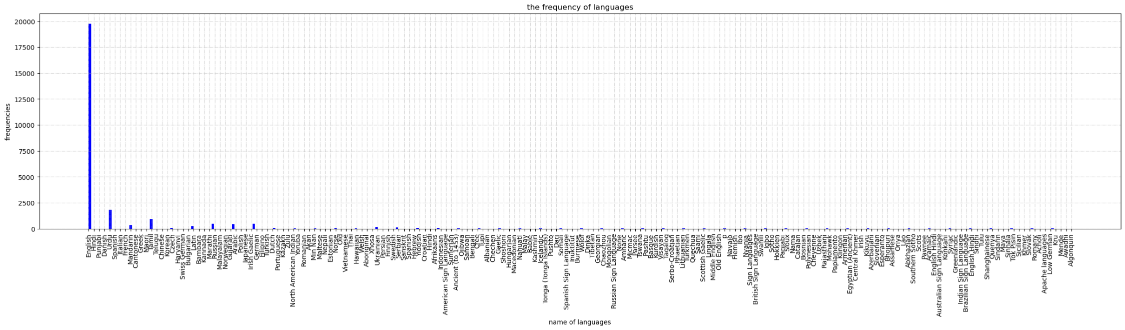
```
  :
-------------------------------------------------
IMDb-rating:

841
  :
(1.1, 9.9, 4.8, 5.7, 6.6)
-------------------------------------------------
downloads:

1
  :
(0.0, 391272.0, 855.5, 2716.0, 10070.0)
-------------------------------------------------
id:
```

```
0
  :
(1, 372092, 96122.25, 264457.5, 354561.25)
--------------------------------------------------
views:

1
  :
(667.0, 1638533.0, 7571.5, 15222.0, 36571.0)
```

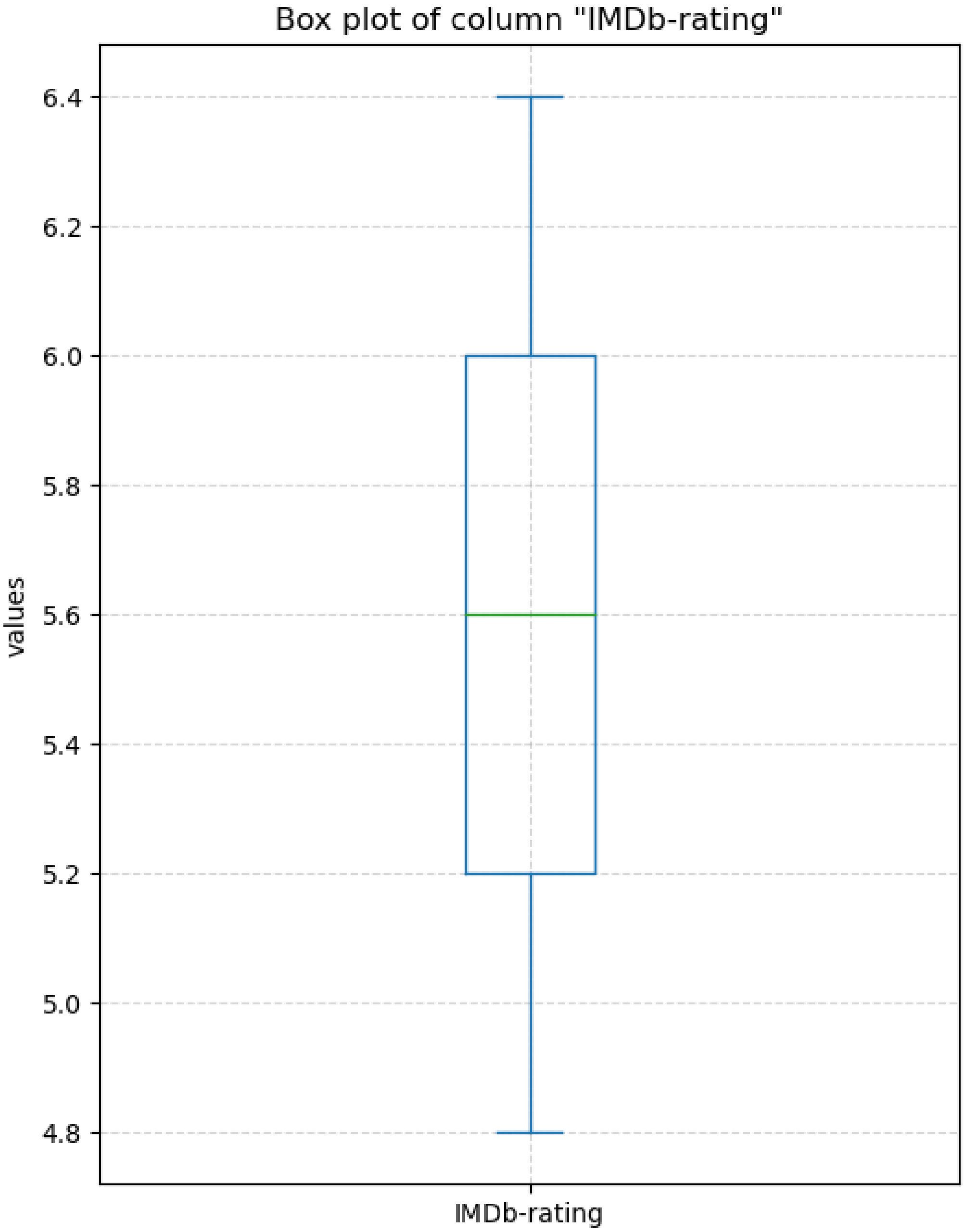
[5]:
```
# histogram
# industry
plt.figure(num=1, figsize=(10,6))
sub_data = data.dropna(subset=['industry'])['industry']
plt.xticks(rotation=90)
plt.hist(sub_data, bins=48, width=0.5, color='blue')
plt.grid(alpha=0.5, linestyle='-.')
plt.xlabel('name of industries')
plt.ylabel('frequencies')
plt.title(r'the frequency of industries')
plt.show()
```



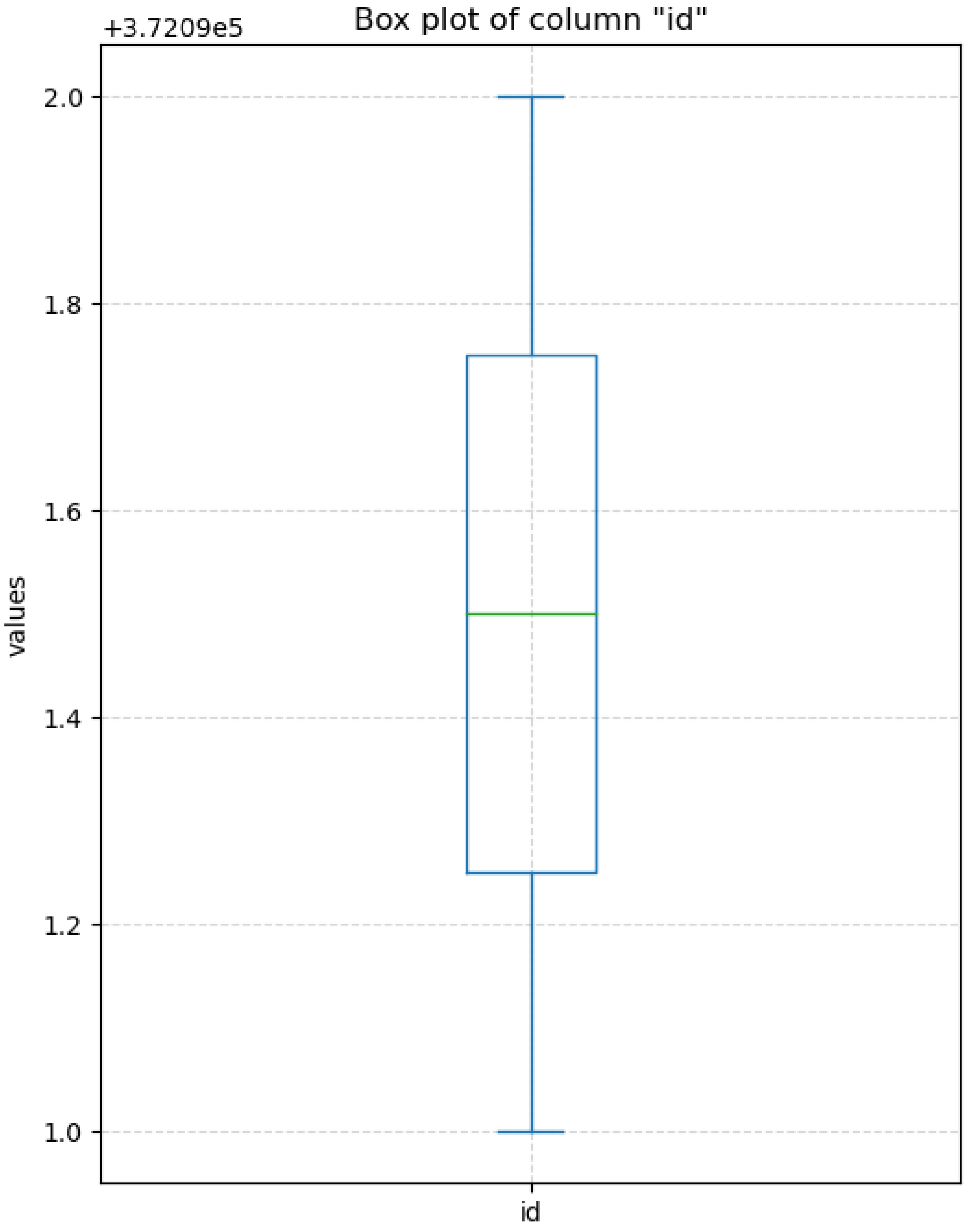
[6]:
```python
def generate_list(data, name):
    attr_list = []
    Attribute_list = data.dropna(subset=[name])[name]
    for attribute_list in Attribute_list:
        Attributes = attribute_list.strip().split(',')
        for attr in Attributes:
            attr_list.append(attr)
    return attr_list

# language
plt.figure(num=2, figsize=(30,6))
attr_list = generate_list(data, 'language')
plt.xticks(rotation=90)
plt.hist(attr_list, bins=48, width=0.5, color='blue')
plt.grid(alpha=0.5, linestyle='-.')
plt.xlabel('name of languages')
plt.ylabel('frequencies')
plt.title(r'the frequency of languages')
plt.show()
```



[37]:
```python
# box-plot picture
# IMDb-rating
plt.figure(num=3, figsize=(6, 8))
sub_data = data.dropna(subset=['IMDb-rating'])['IMDb-rating'][:2]
sub_data.plot.box(title='box plot')
plt.grid(linestyle="--", alpha=0.5)
plt.ylabel('values')
plt.title(r'Box plot of column "IMDb-rating"')
plt.show()
```

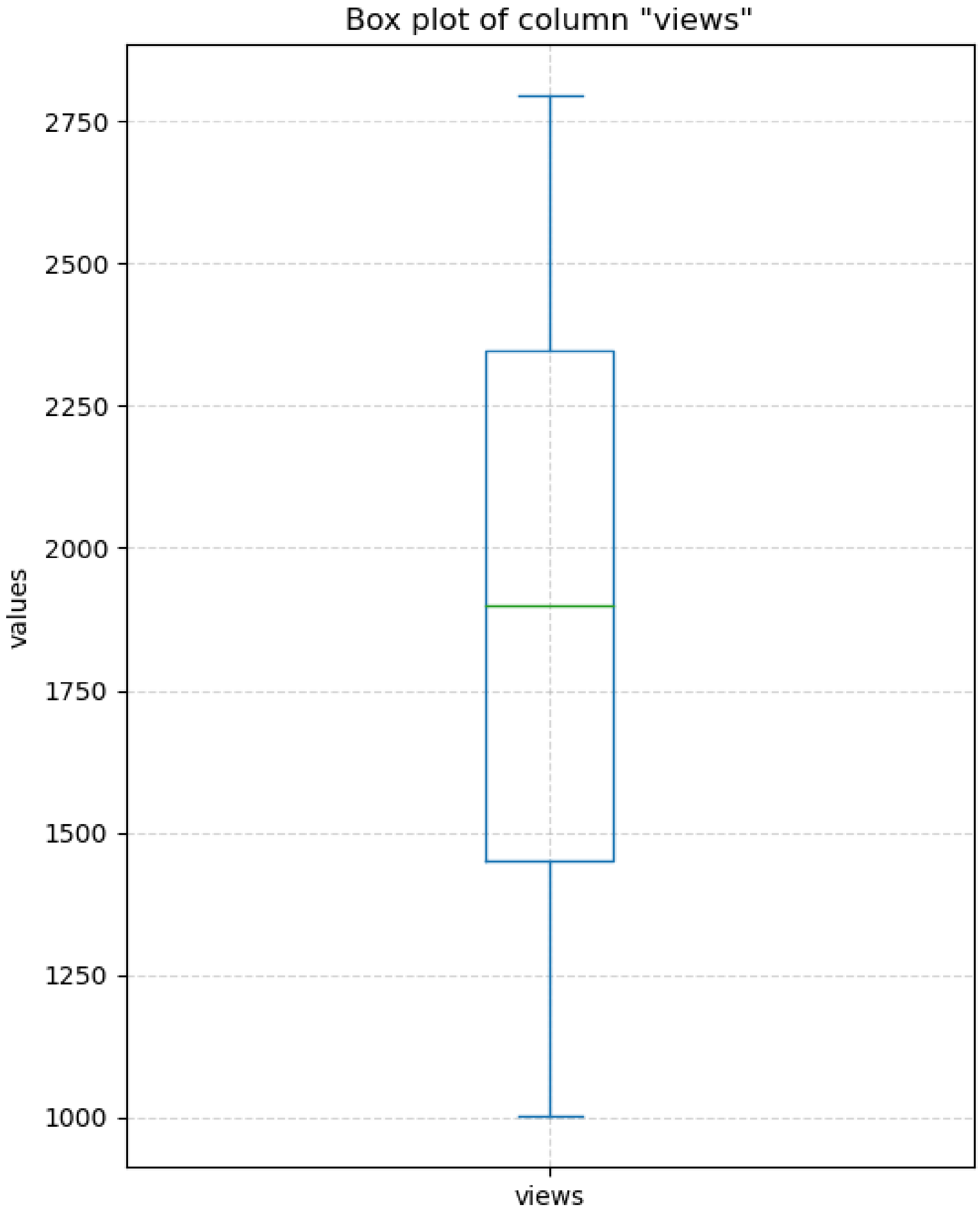


[36]:
```
# id
plt.figure(num=5, figsize=(6, 8))
#data['id'] = data['id'].str.replace(',', '').astype(float)
sub_data = data.dropna(subset=['id'])['id'][:2]
sub_data.plot.box(title='box plot')
plt.grid(linestyle="--", alpha=0.5)
plt.ylabel('values')
plt.title(r'Box plot of column "id"')
plt.show()
```



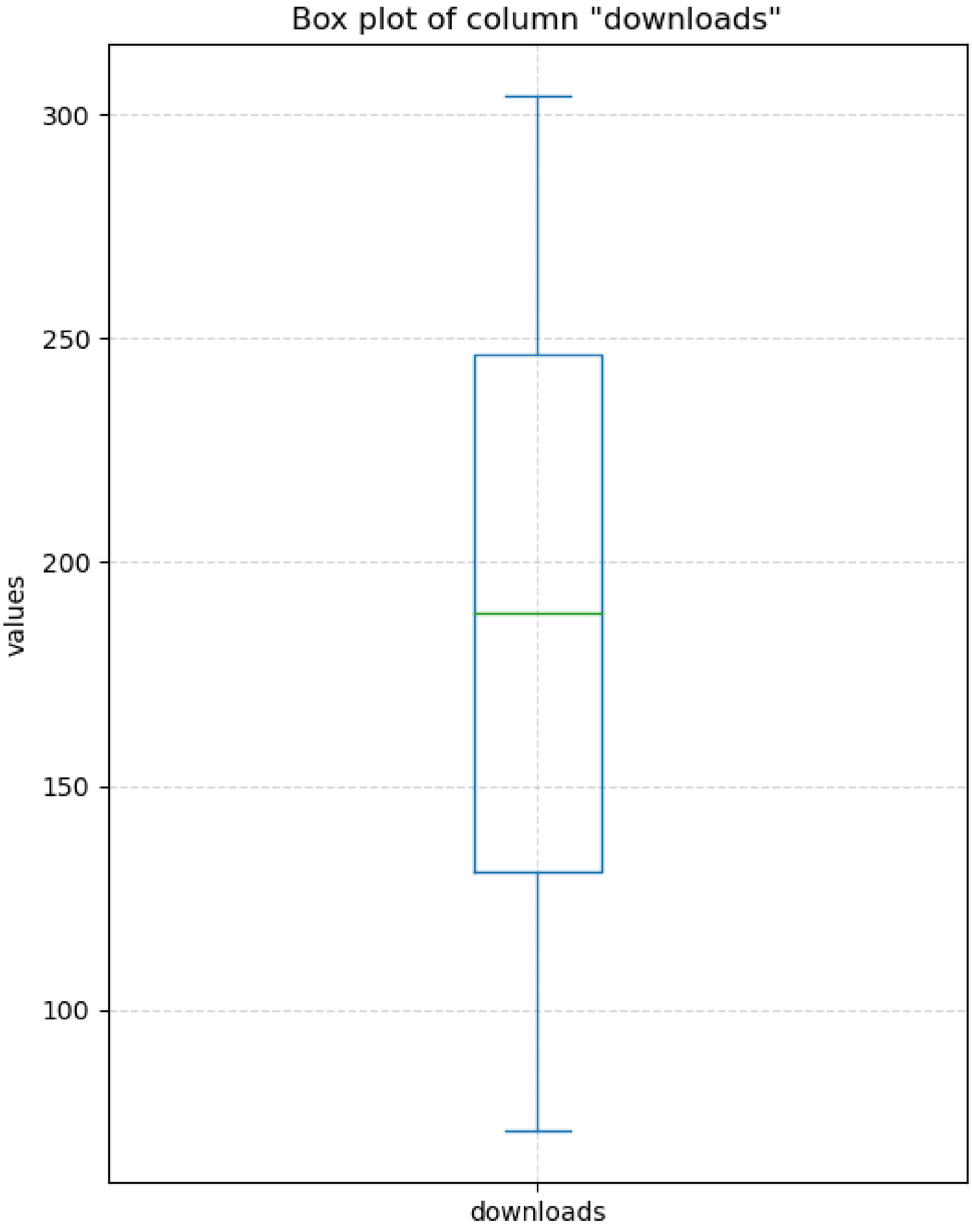
[34]:

```
# views
plt.figure(num=6, figsize=(6, 8))
sub_data = data['views'][:2]
sub_data.plot.box(title='box plot')
plt.grid(linestyle="--", alpha=0.5)
plt.ylabel('values')
plt.title(r'Box plot of column "views"')
plt.show()
```



[35]:
```
# downloads
plt.figure(num=4, figsize=(6, 8))
sub_data = data.dropna(subset=['downloads'])['downloads'][:2]
```

```
sub_data.plot.box(title='box plot')
plt.grid(linestyle="--", alpha=0.5)
plt.ylabel('values')
plt.title(r'Box plot of column "downloads"')
plt.show()
```



3.2 :

```
[17]: #
      print('    ')
      data[data[['IMDb-rating']].isnull().T.any()].iloc[:10,]
```

```
[17]:     Unnamed: 0  IMDb-rating appropriate_for           director  downloads  \
      6            6          NaN           TV-PG                NaN     5332.0
      12          12          NaN             NaN                NaN     2253.0
      16          16          NaN             NaN                NaN     2785.0
      18          18          NaN             NaN                NaN      171.0
      24          24          NaN             NaN      Sumeet Nagdev     1299.0
      44          44          NaN             NaN                NaN     2609.0
      55          55          NaN             NaN                NaN     1781.0
      62          62          NaN             NaN  Anil Bhoop Sharma     1373.0
      63          63          NaN             NaN                NaN      309.0
      77          77          NaN             NaN                NaN     3108.0

              id            industry language   posted_date release_date run_time  \
      6   372059           Wrestling  English  19 Feb, 2023  Feb 18 2023      200
      12  372038           Wrestling  English  18 Feb, 2023  Feb 17 2023      NaN
      16  371990             Punjabi  Punjabi  17 Feb, 2023  Feb 16 2023      NaN
      18  371988           Wrestling  English  17 Feb, 2023  Feb 16 2023      NaN
      24  371932  Bollywood / Indian    Hindi  16 Feb, 2023  Jan 21 2023      142
      44  371816           Wrestling  English  14 Feb, 2023  Feb 13 2023      NaN
      55  371740           Wrestling  English  13 Feb, 2023  Feb 10 2023      NaN
      62  371670  Bollywood / Indian    Hindi  10 Feb, 2023  Sep 08 2022       75
      63  371669           Wrestling  English  10 Feb, 2023  Feb 09 2023      NaN
      77  371517           Wrestling  English  07 Feb, 2023  Feb 06 2023      NaN

                                                  storyline  \
      6   Undisputed\r\n WWE Universal title: Reigns vs …
      12                                                NaN
      16                                                NaN
      18                                                NaN
      24  21-year\r\n old Vicky is diagnosed with mental…
      44                                                NaN
      55                                                NaN
      62                                                NaN
      63                                                NaN
      77                                                NaN

                             title    views                               writer
      6     WWE Elimination Chamber  11978.0                                  NaN
      12   WWE Smackdown 2023-02-17   5468.0                                  NaN
      16      Sab Fadey Jaange.2023  12968.0                                  NaN
      18      TNA.Impact 2023-02-16    667.0                                  NaN
      24                 Ho Ja Mukt  10891.0                        Sumeet Nagdev
```

```
44        WWE Raw 2023-02-13   6605.0                                  NaN
55  WWE Smackdown 2023-02-10   4736.0                                  NaN
62                   Subhagi  10318.0  Munshi Premchand, Anil Bhoop Sharma
63     TNA.Impact 2023-02-09   1337.0                                  NaN
77        WWE Raw 2023-02-06   8205.0                                  NaN
```

[18]:

```
#
data1 = data.dropna()
print('    ')
data1[data[['IMDb-rating']].isnull().T.any()].iloc[:10,]
```

[18]:

```
Empty DataFrame
Columns: [Unnamed: 0, IMDb-rating, appropriate_for, director, downloads, id,
industry, language, posted_date, release_date, run_time, storyline, title,
views, writer]
Index: []
```

[19]:

```
#
data2 = data.copy(deep=True)
data2['IMDb-rating'] = data2['IMDb-rating'].fillna(np.median(data2.
 ↪dropna(subset=['IMDb-rating'])['IMDb-rating']),inplace=False)
print('      ')
data2[data[['IMDb-rating']].isnull().T.any()].iloc[:10,]
```

[19]:

```
    Unnamed: 0  IMDb-rating appropriate_for           director  downloads  \
6            6          5.7           TV-PG                NaN     5332.0
12          12          5.7             NaN                NaN     2253.0
16          16          5.7             NaN                NaN     2785.0
18          18          5.7             NaN                NaN      171.0
24          24          5.7             NaN      Sumeet Nagdev     1299.0
44          44          5.7             NaN                NaN     2609.0
55          55          5.7             NaN                NaN     1781.0
62          62          5.7             NaN  Anil Bhoop Sharma     1373.0
63          63          5.7             NaN                NaN      309.0
77          77          5.7             NaN                NaN     3108.0

        id             industry language   posted_date release_date run_time  \
6   372059            Wrestling  English  19 Feb, 2023  Feb 18 2023      200
12  372038            Wrestling  English  18 Feb, 2023  Feb 17 2023      NaN
16  371990              Punjabi  Punjabi  17 Feb, 2023  Feb 16 2023      NaN
18  371988            Wrestling  English  17 Feb, 2023  Feb 16 2023      NaN
24  371932   Bollywood / Indian    Hindi  16 Feb, 2023  Jan 21 2023      142
44  371816            Wrestling  English  14 Feb, 2023  Feb 13 2023      NaN
55  371740            Wrestling  English  13 Feb, 2023  Feb 10 2023      NaN
```

```
62  371670  Bollywood / Indian    Hindi  10 Feb, 2023  Sep 08 2022       75
63  371669           Wrestling  English  10 Feb, 2023  Feb 09 2023      NaN
77  371517           Wrestling  English  10 Feb, 2023  Feb 06 2023      NaN

                                            storyline  \
6   Undisputed\r\n WWE Universal title: Reigns vs …
12                                                NaN
16                                                NaN
18                                                NaN
24  21-year\r\n old Vicky is diagnosed with mental…
44                                                NaN
55                                                NaN
62                                                NaN
63                                                NaN
77                                                NaN

                       title    views                               writer
6    WWE Elimination Chamber  11978.0                                  NaN
12  WWE Smackdown 2023-02-17   5468.0                                  NaN
16     Sab Fadey Jaange.2023  12968.0                                  NaN
18     TNA.Impact 2023-02-16    667.0                                  NaN
24                Ho Ja Mukt  10891.0                        Sumeet Nagdev
44        WWE Raw 2023-02-13   6605.0                                  NaN
55  WWE Smackdown 2023-02-10   4736.0                                  NaN
62                   Subhagi  10318.0  Munshi Premchand, Anil Bhoop Sharma
63     TNA.Impact 2023-02-09   1337.0                                  NaN
77        WWE Raw 2023-02-06   8205.0                                  NaN
```

```
[20]: #
data3 = data.copy(deep=True)
data3['IMDb-rating'] = data3['IMDb-rating'].interpolate()
print('      ')
data3[data[['IMDb-rating']].isnull().T.any()].iloc[:10,]
```

```
[20]:     Unnamed: 0  IMDb-rating appropriate_for           director  downloads  \
      6            6     5.950000           TV-PG                NaN     5332.0
      12          12     7.800000             NaN                NaN     2253.0
      16          16     6.450000             NaN                NaN     2785.0
      18          18     5.550000             NaN                NaN      171.0
      24          24     6.350000             NaN      Sumeet Nagdev     1299.0
      44          44     5.350000             NaN                NaN     2609.0
      55          55     6.850000             NaN                NaN     1781.0
      62          62     6.666667             NaN  Anil Bhoop Sharma     1373.0
      63          63     6.033333             NaN                NaN      309.0
      77          77     5.900000             NaN                NaN     3108.0
```

```
        id            industry language   posted_date release_date run_time  \
6   372059           Wrestling  English  19 Feb, 2023  Feb 18 2023      200
12  372038           Wrestling  English  18 Feb, 2023  Feb 17 2023      NaN
16  371990             Punjabi  Punjabi  17 Feb, 2023  Feb 16 2023      NaN
18  371988           Wrestling  English  17 Feb, 2023  Feb 16 2023      NaN
24  371932  Bollywood / Indian    Hindi  16 Feb, 2023  Jan 21 2023      142
44  371816           Wrestling  English  14 Feb, 2023  Feb 13 2023      NaN
55  371740           Wrestling  English  13 Feb, 2023  Feb 10 2023      NaN
62  371670  Bollywood / Indian    Hindi  10 Feb, 2023  Sep 08 2022       75
63  371669           Wrestling  English  10 Feb, 2023  Feb 09 2023      NaN
77  371517           Wrestling  English  07 Feb, 2023  Feb 06 2023      NaN

                                           storyline  \
6   Undisputed\r\n WWE Universal title: Reigns vs …
12                                               NaN
16                                               NaN
18                                               NaN
24  21-year\r\n old Vicky is diagnosed with mental…
44                                               NaN
55                                               NaN
62                                               NaN
63                                               NaN
77                                               NaN

                       title    views                                  writer
6    WWE Elimination Chamber  11978.0                                     NaN
12  WWE Smackdown 2023-02-17   5468.0                                     NaN
16     Sab Fadey Jaange.2023  12968.0                                     NaN
18     TNA.Impact 2023-02-16    667.0                                     NaN
24                Ho Ja Mukt  10891.0                          Sumeet Nagdev
44        WWE Raw 2023-02-13   6605.0                                     NaN
55  WWE Smackdown 2023-02-10   4736.0                                     NaN
62                   Subhagi  10318.0  Munshi Premchand, Anil Bhoop Sharma
63     TNA.Impact 2023-02-09   1337.0                                     NaN
77        WWE Raw 2023-02-06   8205.0                                     NaN
```

[21]:

```
#
data4 = data.copy(deep=True)
data4['IMDb-rating'] = data4['IMDb-rating'].fillna(np.
 ↪mean(data4['IMDb-rating']),inplace=False)
print('     : ')
data4[data[['IMDb-rating']].isnull().T.any()].iloc[:10,]
```

```
     :
```

[21]:

```
    Unnamed: 0  IMDb-rating appropriate_for           director  downloads  \
6            6     5.762151           TV-PG                NaN     5332.0
12          12     5.762151             NaN                NaN     2253.0
```

```
16          16     5.762151             NaN                 NaN     2785.0
18          18     5.762151             NaN                 NaN      171.0
24          24     5.762151             NaN       Sumeet Nagdev     1299.0
44          44     5.762151             NaN                 NaN     2609.0
55          55     5.762151             NaN                 NaN     1781.0
62          62     5.762151             NaN   Anil Bhoop Sharma     1373.0
63          63     5.762151             NaN                 NaN      309.0
77          77     5.762151             NaN                 NaN     3108.0

        id            industry language   posted_date release_date run_time  \
6   372059           Wrestling  English  19 Feb, 2023  Feb 18 2023      200
12  372038           Wrestling  English  18 Feb, 2023  Feb 17 2023      NaN
16  371990             Punjabi  Punjabi  17 Feb, 2023  Feb 16 2023      NaN
18  371988           Wrestling  English  17 Feb, 2023  Feb 16 2023      NaN
24  371932  Bollywood / Indian    Hindi  16 Feb, 2023  Jan 21 2023      142
44  371816           Wrestling  English  14 Feb, 2023  Feb 13 2023      NaN
55  371740           Wrestling  English  13 Feb, 2023  Feb 10 2023      NaN
62  371670  Bollywood / Indian    Hindi  10 Feb, 2023  Sep 08 2022       75
63  371669           Wrestling  English  10 Feb, 2023  Feb 09 2023      NaN
77  371517           Wrestling  English  07 Feb, 2023  Feb 06 2023      NaN

                                            storyline  \
6   Undisputed\r\n WWE Universal title: Reigns vs …
12                                                NaN
16                                                NaN
18                                                NaN
24  21-year\r\n old Vicky is diagnosed with mental…
44                                                NaN
55                                                NaN
62                                                NaN
63                                                NaN
77                                                NaN

                       title    views                               writer
6    WWE Elimination Chamber  11978.0                                  NaN
12  WWE Smackdown 2023-02-17   5468.0                                  NaN
16     Sab Fadey Jaange.2023  12968.0                                  NaN
18     TNA.Impact 2023-02-16    667.0                                  NaN
24                Ho Ja Mukt  10891.0                        Sumeet Nagdev
44        WWE Raw 2023-02-13   6605.0                                  NaN
55  WWE Smackdown 2023-02-10   4736.0                                  NaN
62                   Subhagi  10318.0  Munshi Premchand, Anil Bhoop Sharma
63     TNA.Impact 2023-02-09   1337.0                                  NaN
77        WWE Raw 2023-02-06   8205.0                                  NaN
```

```
[ ]:
```