main

November 7, 2023

[282]:

```
import pandas as pd
import numpy as np
```

[283]:

```
data = pd.read_csv('movie_metadata.csv')
```

[284]:

```
print(data.columns)
set([data[col].dtype for col in data.columns])
```

```
Index(['color', 'director_name', 'num_critic_for_reviews', 'duration',
       'director_facebook_likes', 'actor_3_facebook_likes', 'actor_2_name',
       'actor_1_facebook_likes', 'gross', 'genres', 'actor_1_name',
       'movie_title', 'num_voted_users', 'cast_total_facebook_likes',
       'actor_3_name', 'facenumber_in_poster', 'plot_keywords',
       'movie_imdb_link', 'num_user_for_reviews', 'language', 'country',
       'content_rating', 'budget', 'title_year', 'actor_2_facebook_likes',
       'imdb_score', 'aspect_ratio', 'movie_facebook_likes'],
      dtype='object')
```

[284]:

```
{dtype('int64'), dtype('float64'), dtype('O')}
```

Fill every NaN value with the mean of the column.

[285]:

```
for col in data.select_dtypes(['int', 'float']).columns:
    mean = data[col].mean()
    data[col].fillna(mean, inplace=True)
```

Remove duplicate rows

[286]:

```
data = data.drop_duplicates()
```

Add new column `profit`

[287]:

```
data['profit'] = data['gross'] - data['budget']
```

Remove spaces from and lower the titles

[288]:

```
data['movie_title'] = data['movie_title'].apply(lambda x: x.replace(' ', '').
 ↪lower())
data['movie_title']
```

```
[288]: 0                                          avatar
       1         piratesofthecaribbean:atworld'send
       2                                         spectre
       3                              thedarkknightrises
       4        starwars:episodevii-theforceawakens
                                   …
       5038                   signedsealeddelivered
       5039                            thefollowing
       5040                       aplaguesopleasant
       5041                         shanghaicalling
       5042                          mydatewithdrew
       Name: movie_title, Length: 4998, dtype: object
```

Sort by `imdb_score` in descending order

```
[289]: data.sort_values(by='imdb_score', ascending=False)
```

```
[289]:       color         director_name  num_critic_for_reviews  duration  \
       2765  Color        John Blanchard              140.194272      65.0
       1937  Color        Frank Darabont              199.000000     142.0
       3466  Color  Francis Ford Coppola              208.000000     175.0
       3207  Color                   NaN               53.000000      55.0
       2824  Color                   NaN               53.000000      55.0
       …     …                   …                     …       …
       4605  Color         A. Raven Cruz                3.000000      97.0
       2295  Color             Bob Clark               32.000000      88.0
       2268  Color       Jason Friedberg              111.000000      88.0
       1136  Color     Lawrence Kasanoff               12.000000      91.0
       2834  Color            Jon M. Chu               84.000000     115.0

             director_facebook_likes  actor_3_facebook_likes     actor_2_name  \
       2765                 0.000000                   176.0    Andrea Martin
       1937                 0.000000                   461.0   Jeffrey DeMunn
       3466                 0.000000                  3000.0    Marlon Brando
       3207               686.509212                     2.0  Olaf Lubaszenko
       2824               686.509212                     2.0  Olaf Lubaszenko
       …                       …                     …              …
       4605                 0.000000                    94.0      Vanilla Ice
       2295                84.000000                   177.0    Vanessa Angel
       2268                82.000000                   329.0         Tony Cox
       1136                11.000000                   500.0     Larry Miller
       2834               209.000000                    41.0    Sean Kingston

             actor_1_facebook_likes         gross  \
       2765                   770.0  4.846841e+07
       1937                 11000.0  2.834147e+07
       3466                 14000.0  1.348220e+08
       3207                    20.0  4.470930e+05
```

```
2824                    20.0  4.470930e+05
…                      …           …
4605                   639.0  4.846841e+07
2295                   650.0  9.109322e+06
2268                   869.0  1.417465e+07
1136                   719.0  4.846841e+07
2834                   569.0  7.300094e+07

                                      genres  … language  country  \
2765                                  Comedy  …  English   Canada
1937                             Crime|Drama  …  English      USA
3466                             Crime|Drama  …  English      USA
3207                                   Drama  …   Polish   Poland
2824                                   Drama  …   Polish   Poland
…                                      …   …      …      …
4605  Action|Adventure|Comedy|Fantasy|Sci-Fi  …  English      USA
2295                    Comedy|Family|Sci-Fi  …  English  Germany
2268                                  Comedy  …  English      USA
1136  Action|Animation|Comedy|Family|Fantasy  …  English      USA
2834                       Documentary|Music  …  English      USA

      content_rating        budget   title_year  actor_2_facebook_likes  \
2765             NaN  3.975262e+07  2002.470517                   179.0
1937               R  2.500000e+07  1994.000000                   745.0
3466               R  6.000000e+06  1972.000000                 10000.0
3207           TV-MA  3.975262e+07  2002.470517                     3.0
2824           TV-MA  3.975262e+07  2002.470517                     3.0
…              …           …          …                     …
4605               R  1.000000e+06  2005.000000                   361.0
2295              PG  2.000000e+07  2004.000000                   384.0
2268           PG-13  2.500000e+07  2008.000000                   624.0
1136              PG  6.500000e+07  2012.000000                   611.0
2834               G  1.300000e+07  2011.000000                    69.0

     imdb_score aspect_ratio  movie_facebook_likes        profit
2765        9.5     1.330000                     0  8.715787e+06
1937        9.3     1.850000                108000  3.341469e+06
3466        9.2     1.850000                 43000  1.288220e+08
3207        9.1     1.330000                     0 -3.930553e+07
2824        9.1     1.330000                     0 -3.930553e+07
…          …          …                    …           …
4605        1.9     1.780000                   128  4.746841e+07
2295        1.9     2.350000                     0 -1.089068e+07
2268        1.9     1.850000                     0 -1.082535e+07
1136        1.7     2.220403                     0 -1.653159e+07
2834        1.6     1.850000                 62000  6.000094e+07
```

```
[4998 rows x 29 columns]
```

Substitute null language with the most spoken one

```
[290]: most_lang = data.value_counts('language', ascending=False).idxmax(0)
       data['language'] = data['language'].fillna(most_lang)
```

7 successful_movie imdb_score 7.5 profit 0 True False

```
[291]: data['successful_movie'] = (data.imdb_score > 7.5) & (data.profit > 0)
```

8 genres One-Hot

```
[292]: oh = pd.get_dummies(data['genres'].str.split('|', expand=True).
         ↪stack(dropna=True))
       oh = oh.groupby(level=0,axis=0).max()
       oh
```

```
[292]:       Action  Adventure  Animation  Biography  Comedy  Crime  Documentary  \
       0          1          1          0          0       0      0            0
       1          1          1          0          0       0      0            0
       2          1          1          0          0       0      0            0
       3          1          0          0          0       0      0            0
       4          0          0          0          0       0      0            1
       …       …       …       …       …     …    …        …
       5038       0          0          0          0       1      0            0
       5039       0          0          0          0       0      1            0
       5040       0          0          0          0       0      0            0
       5041       0          0          0          0       1      0            0
       5042       0          0          0          0       0      0            1

             Drama  Family  Fantasy  …  Mystery  News  Reality-TV  Romance  Sci-Fi  \
       0         0       0        1  …        0     0           0        0       1
       1         0       0        1  …        0     0           0        0       0
       2         0       0        0  …        0     0           0        0       0
       3         0       0        0  …        0     0           0        0       0
       4         0       0        0  …        0     0           0        0       0
       …     …     …      …  …      …    …     …     …      …
       5038      1       0        0  …        0     0           0        0       0
       5039      1       0        0  …        1     0           0        0       0
       5040      1       0        0  …        0     0           0        0       0
       5041      1       0        0  …        0     0           0        1       0
       5042      0       0        0  …        0     0           0        0       0

             Short  Sport  Thriller  War  Western
       0         0      0         0    0        0
       1         0      0         0    0        0
       2         0      0         1    0        0
       3         0      0         1    0        0
```

```
4         0       0         0    0        0
…     …    …        …  …      …
5038      0       0         0    0        0
5039      0       0         1    0        0
5040      0       0         1    0        0
5041      0       0         0    0        0
5042      0       0         0    0        0

[4998 rows x 26 columns]
```

[293]:
```
data = data.merge(oh, left_index=True, right_index=True)
data
```

[293]:
```
      color      director_name  num_critic_for_reviews    duration  \
0     Color      James Cameron              723.000000  178.000000
1     Color     Gore Verbinski              302.000000  169.000000
2     Color         Sam Mendes              602.000000  148.000000
3     Color  Christopher Nolan              813.000000  164.000000
4       NaN        Doug Walker              140.194272  107.201074
…     …                 …                     …         …
5038  Color        Scott Smith                1.000000   87.000000
5039  Color                NaN               43.000000   43.000000
5040  Color   Benjamin Roberds               13.000000   76.000000
5041  Color        Daniel Hsia               14.000000  100.000000
5042  Color           Jon Gunn               43.000000   90.000000

      director_facebook_likes  actor_3_facebook_likes      actor_2_name  \
0                    0.000000              855.000000  Joel David Moore
1                  563.000000             1000.000000     Orlando Bloom
2                    0.000000              161.000000      Rory Kinnear
3                22000.000000            23000.000000    Christian Bale
4                  131.000000              645.009761        Rob Walker
…                        …                     …                …
5038                 2.000000              318.000000     Daphne Zuniga
5039               686.509212              319.000000     Valorie Curry
5040                 0.000000                0.000000     Maxwell Moody
5041                 0.000000              489.000000     Daniel Henney
5042                16.000000               16.000000  Brian Herzlinger

      actor_1_facebook_likes         gross                           genres  \
0                     1000.0  7.605058e+08  Action|Adventure|Fantasy|Sci-Fi
1                    40000.0  3.094042e+08         Action|Adventure|Fantasy
2                    11000.0  2.000742e+08        Action|Adventure|Thriller
3                    27000.0  4.481306e+08                  Action|Thriller
4                      131.0  4.846841e+07                      Documentary
…                       …           …                               …
5038                   637.0  4.846841e+07                     Comedy|Drama
```

```
5039                   841.0  4.846841e+07      Crime|Drama|Mystery|Thriller
5040                     0.0  4.846841e+07             Drama|Horror|Thriller
5041                   946.0  1.044300e+04              Comedy|Drama|Romance
5042                    86.0  8.522200e+04                       Documentary

      … Mystery News  Reality-TV  Romance Sci-Fi  Short Sport Thriller  War   \
0     …       0    0           0        0      1      0     0        0    0
1     …       0    0           0        0      0      0     0        0    0
2     …       0    0           0        0      0      0     0        1    0
3     …       0    0           0        0      0      0     0        1    0
4     …       0    0           0        0      0      0     0        0    0
…   …     …  …         …      …    …    …   …      …  …
5038  …       0    0           0        0      0      0     0        0    0
5039  …       1    0           0        0      0      0     0        1    0
5040  …       0    0           0        0      0      0     0        1    0
5041  …       0    0           0        1      0      0     0        0    0
5042  …       0    0           0        0      0      0     0        0    0

     Western
0          0
1          0
2          0
3          0
4          0
…      …
5038       0
5039       0
5040       0
5041       0
5042       0

[4998 rows x 56 columns]
```

9 avg_director_score

```
[294]: direc_score = data.groupby('director_name')['imdb_score'].mean().reset_index().
         ↪rename(columns={'imdb_score': 'avg_director_score'})
       direc_score
```

```
[294]:            director_name  avg_director_score
       0         A. Raven Cruz                 1.9
       1            Aaron Hann                 6.0
       2       Aaron Schneider                 7.1
       3         Aaron Seltzer                 2.7
       4          Abel Ferrara                 6.6
       …                  …                 …
       2393      Zoran Lisinac                 7.1
       2394 Álex de la Iglesia                 6.1
```

```
2395    Émile Gaudreault                  6.7
2396        Éric Tessier                  6.6
2397       Étienne Faure                  4.3

[2398 rows x 2 columns]
```

[295]:

```
data = data.join(direc_score.set_index('director_name'), on='director_name')
data
```

[295]:

```
      color      director_name  num_critic_for_reviews    duration  \
0     Color      James Cameron              723.000000  178.000000
1     Color     Gore Verbinski              302.000000  169.000000
2     Color         Sam Mendes              602.000000  148.000000
3     Color  Christopher Nolan              813.000000  164.000000
4       NaN        Doug Walker              140.194272  107.201074
…       …                  …                      …           …
5038  Color        Scott Smith                1.000000   87.000000
5039  Color                NaN               43.000000   43.000000
5040  Color   Benjamin Roberds               13.000000   76.000000
5041  Color        Daniel Hsia               14.000000  100.000000
5042  Color           Jon Gunn               43.000000   90.000000

      director_facebook_likes  actor_3_facebook_likes      actor_2_name  \
0                    0.000000              855.000000  Joel David Moore
1                  563.000000             1000.000000     Orlando Bloom
2                    0.000000              161.000000      Rory Kinnear
3                22000.000000            23000.000000    Christian Bale
4                  131.000000              645.009761        Rob Walker
…                         …                       …               …
5038                 2.000000              318.000000     Daphne Zuniga
5039               686.509212              319.000000     Valorie Curry
5040                 0.000000                0.000000     Maxwell Moody
5041                 0.000000              489.000000    Daniel Henney
5042                16.000000               16.000000  Brian Herzlinger

      actor_1_facebook_likes         gross                           genres  \
0                     1000.0  7.605058e+08  Action|Adventure|Fantasy|Sci-Fi
1                    40000.0  3.094042e+08         Action|Adventure|Fantasy
2                    11000.0  2.000742e+08        Action|Adventure|Thriller
3                    27000.0  4.481306e+08                  Action|Thriller
4                      131.0  4.846841e+07                      Documentary
…                        …           …                              …
5038                   637.0  4.846841e+07                     Comedy|Drama
5039                   841.0  4.846841e+07     Crime|Drama|Mystery|Thriller
5040                     0.0  4.846841e+07            Drama|Horror|Thriller
5041                   946.0  1.044300e+04             Comedy|Drama|Romance
5042                    86.0  8.522200e+04                      Documentary
```

```
      … News Reality-TV  Romance  Sci-Fi Short  Sport Thriller War  Western  \
0     …    0          0        0       1     0      0        0   0        0
1     …    0          0        0       0     0      0        0   0        0
2     …    0          0        0       0     0      0        1   0        0
3     …    0          0        0       0     0      0        1   0        0
4     …    0          0        0       0     0      0        0   0        0
…   …  …         …      …      …   …    …      …  ..       …
5038  …    0          0        0       0     0      0        0   0        0
5039  …    0          0        0       0     0      0        1   0        0
5040  …    0          0        0       0     0      0        1   0        0
5041  …    0          0        1       0     0      0        0   0        0
5042  …    0          0        0       0     0      0        0   0        0

     avg_director_score
0              7.914286
1              6.985714
2              7.500000
3              8.425000
4              7.100000
…                   …
5038           7.700000
5039                NaN
5040           6.300000
5041           6.300000
5042           6.033333

[4998 rows x 57 columns]
```

title_year year_group

[296]:
```
years = pd.cut(data['title_year'], bins=[1910 + k * 10 for k in range(13)],␣
 ↪labels=[1910 + k * 10 for k in range(12)]).reset_index()
years = years.rename(columns={'title_year': 'year_group'})['year_group']
years
```

[296]:
```
0       2000
1       2000
2       2010
3       2010
4       2000
        …
4993    2010
4994    2000
4995    2010
4996    2010
4997    2000
Name: year_group, Length: 4998, dtype: category
```

```
Categories (12, int64): [1910 < 1920 < 1930 < 1940 … 1990 < 2000 < 2010 <
2020]
```

[297]:

```python
data = data.merge(years, left_index=True, right_index=True)
data
```

[297]:

```
      color       director_name  num_critic_for_reviews    duration  \
0     Color       James Cameron              723.000000  178.000000
1     Color      Gore Verbinski              302.000000  169.000000
2     Color          Sam Mendes              602.000000  148.000000
3     Color   Christopher Nolan              813.000000  164.000000
4       NaN         Doug Walker              140.194272  107.201074
…       …               …                         …          …
4993  Color      William Eubank              161.000000   97.000000
4994  Color      Patrick Meaney                7.000000   81.000000
4995  Color        Chad Hartigan               34.000000   83.000000
4996  Color      Malcolm Goodwin             140.194272   96.000000
4997  Color  David Gordon Green               75.000000   90.000000

      director_facebook_likes  actor_3_facebook_likes      actor_2_name  \
0                         0.0              855.000000  Joel David Moore
1                       563.0             1000.000000     Orlando Bloom
2                         0.0              161.000000      Rory Kinnear
3                     22000.0            23000.000000    Christian Bale
4                       131.0              645.009761        Rob Walker
…                         …                     …               …
4993                     18.0              236.000000      Olivia Cooke
4994                      3.0               18.000000    Greg Aronowitz
4995                      3.0               69.000000     Paul Eenhoorn
4996                    117.0              281.000000         Jon Gries
4997                    234.0               15.000000       Eddie Rouse

      actor_1_facebook_likes         gross                           genres  \
0                     1000.0  7.605058e+08  Action|Adventure|Fantasy|Sci-Fi
1                    40000.0  3.094042e+08         Action|Adventure|Fantasy
2                    11000.0  2.000742e+08        Action|Adventure|Thriller
3                    27000.0  4.481306e+08                  Action|Thriller
4                      131.0  4.846841e+07                      Documentary
…                         …           …                               …
4993                   852.0  4.846841e+07                  Sci-Fi|Thriller
4994                    26.0  4.846841e+07             Biography|Documentary
4995                   695.0  4.846841e+07                            Drama
4996                   948.0  4.846841e+07                           Comedy
4997                   552.0  2.418160e+05                            Drama

      … Reality-TV Romance  Sci-Fi  Short Sport  Thriller War Western  \
0     …          0       0       1      0     0         0   0       0
```

```
1     …          0       0       0      0     0         0   0       0
2     …          0       0       0      0     0         1   0       0
3     …          0       0       0      0     0         1   0       0
4     …          0       0       0      0     0         0   0       0
…   …        …     …     …    …   …       …  . .    …
4993  …          0       0       1      0     0         1   0       0
4994  …          0       0       0      0     0         0   0       0
4995  …          0       0       0      0     0         0   0       0
4996  …          0       0       0      0     0         0   0       0
4997  …          0       0       0      0     0         0   0       0

      avg_director_score year_group
0               7.914286       2000
1               6.985714       2000
2               7.500000       2010
3               8.425000       2010
4               7.100000       2000
…                  …        …
4993            6.100000       2010
4994            7.400000       2000
4995            6.600000       2010
4996            5.500000       2010
4997            6.575000       2000

[4953 rows x 58 columns]
```

[298]:
```
data['title_year'].unique()
```

[298]:
```
array([2009.        , 2007.        , 2015.        , 2012.        ,
       2002.47051672, 2010.        , 2016.        , 2006.        ,
       2008.        , 2013.        , 2011.        , 2014.        ,
       2005.        , 1997.        , 2004.        , 1999.        ,
       1995.        , 2003.        , 2001.        , 2002.        ,
       1998.        , 2000.        , 1990.        , 1991.        ,
       1994.        , 1996.        , 1982.        , 1993.        ,
       1979.        , 1992.        , 1989.        , 1984.        ,
       1988.        , 1978.        , 1962.        , 1980.        ,
       1972.        , 1981.        , 1968.        , 1985.        ,
       1940.        , 1963.        , 1987.        , 1986.        ,
       1973.        , 1983.        , 1976.        , 1977.        ,
       1970.        , 1971.        , 1969.        , 1960.        ,
       1965.        , 1964.        , 1927.        , 1974.        ,
       1937.        , 1975.        , 1967.        , 1951.        ,
       1961.        , 1946.        , 1953.        , 1954.        ,
       1959.        , 1932.        , 1947.        , 1956.        ,
       1945.        , 1952.        , 1930.        , 1966.        ,
       1939.        , 1950.        , 1948.        , 1958.        ,
       1957.         , 1943.         , 1944.         , 1938.         ,
       1949.         , 1936.         , 1941.         , 1955.         ,
       1942.         , 1929.         , 1935.         , 1933.         ,
       1916.         , 1934.         , 1925.         , 1920.         ])
```

1 Part 2

1. NumPy duration
2. budget 3
3. NumPy profit
4. NumPy movie_title
5. imdb_score 25% 75% 25% 75%
6. NumPy num_voted_users
7. language
8. NumPy actor_1_facebook_likes
9. NumPy num_user_for_reviews
10. NumPy movie_facebook_likes

[301]:
```
data = pd.read_csv('movie_metadata.csv')
```

[312]:
```
# Q1
np.mean(data['duration'])
np.mean(data['duration'][data['duration'].notna()])
```

[312]:
```
107.2010739856802
```

[321]:
```
# Q2
budgets = data['budget'][data['budget'].notna()]
mean, std = np.mean(budgets), np.std(budgets)
budgets[(budgets > mean + 3 * std) | (budgets < mean - 3 * std)]
```

[321]:
```
2323    2.400000e+09
2334    2.127520e+09
2988    1.221550e+10
3005    2.500000e+09
3075    7.000000e+08
3423    1.100000e+09
3851    7.000000e+08
3859    4.200000e+09
4542    1.000000e+09
Name: budget, dtype: float64
```

[334]:
```
# Q3
idx = data['budget'].notna() & data['gross'].notna()
profit = np.array(data[idx]['budget'].array)
data.loc[idx, 'profit'] = profit
data
```

```
[334]:
      color      director_name  num_critic_for_reviews  duration  \
0     Color      James Cameron                   723.0     178.0
1     Color     Gore Verbinski                   302.0     169.0
2     Color         Sam Mendes                   602.0     148.0
3     Color  Christopher Nolan                   813.0     164.0
4       NaN        Doug Walker                     NaN       NaN
…     …                 …                     …       …
5038  Color        Scott Smith                     1.0      87.0
5039  Color                NaN                    43.0      43.0
5040  Color   Benjamin Roberds                    13.0      76.0
5041  Color        Daniel Hsia                    14.0     100.0
5042  Color           Jon Gunn                    43.0      90.0

      director_facebook_likes  actor_3_facebook_likes      actor_2_name  \
0                         0.0                   855.0  Joel David Moore
1                       563.0                  1000.0     Orlando Bloom
2                         0.0                   161.0      Rory Kinnear
3                     22000.0                 23000.0    Christian Bale
4                       131.0                     NaN        Rob Walker
…                        …                     …               …
5038                      2.0                   318.0     Daphne Zuniga
5039                      NaN                   319.0     Valorie Curry
5040                      0.0                     0.0     Maxwell Moody
5041                      0.0                   489.0     Daniel Henney
5042                     16.0                    16.0  Brian Herzlinger

      actor_1_facebook_likes        gross                           genres  \
0                     1000.0  760505847.0  Action|Adventure|Fantasy|Sci-Fi
1                    40000.0  309404152.0         Action|Adventure|Fantasy
2                    11000.0  200074175.0        Action|Adventure|Thriller
3                    27000.0  448130642.0                  Action|Thriller
4                      131.0          NaN                      Documentary
…                       …          …                              …
5038                   637.0          NaN                     Comedy|Drama
5039                   841.0          NaN     Crime|Drama|Mystery|Thriller
5040                     0.0          NaN            Drama|Horror|Thriller
5041                   946.0      10443.0             Comedy|Drama|Romance
5042                    86.0      85222.0                      Documentary

      … language country  content_rating       budget title_year  \
0     …  English     USA           PG-13  237000000.0     2009.0
1     …  English     USA           PG-13  300000000.0     2007.0
2     …  English      UK           PG-13  245000000.0     2015.0
3     …  English     USA           PG-13  250000000.0     2012.0
4     …      NaN     NaN             NaN          NaN        NaN
…   …       …     …             …          …        …
5038  …  English  Canada             NaN          NaN     2013.0
```

```
5039  …  English     USA           TV-14          NaN        NaN
5040  …  English     USA             NaN       1400.0      2013.0
5041  …  English     USA           PG-13          NaN      2012.0
5042  …  English     USA              PG       1100.0      2004.0

      actor_2_facebook_likes imdb_score aspect_ratio  movie_facebook_likes  \
0                      936.0        7.9         1.78                 33000
1                     5000.0        7.1         2.35                     0
2                      393.0        6.8         2.35                 85000
3                    23000.0        8.5         2.35                164000
4                       12.0        7.1          NaN                     0
…                      …        …          …                   …
5038                   470.0        7.7          NaN                    84
5039                   593.0        7.5        16.00                 32000
5040                     0.0        6.3          NaN                    16
5041                   719.0        6.3         2.35                   660
5042                    23.0        6.6         1.85                   456

           profit
0     237000000.0
1     300000000.0
2     245000000.0
3     250000000.0
4             NaN
…           …
5038          NaN
5039          NaN
5040          NaN
5041          NaN
5042       1100.0

[5043 rows x 29 columns]
```

[343]:

```
# Q4
data['movie_title'] = np.vectorize(lambda x: x.upper())(np.
 ↪array(data['movie_title']))
data
```

[343]:

```
      color      director_name  num_critic_for_reviews  duration  \
0     Color      James Cameron                   723.0     178.0
1     Color     Gore Verbinski                   302.0     169.0
2     Color         Sam Mendes                   602.0     148.0
3     Color  Christopher Nolan                   813.0     164.0
4       NaN        Doug Walker                     NaN       NaN
…     …                …                      …       …
5038  Color        Scott Smith                     1.0      87.0
5039  Color                NaN                    43.0      43.0
```

```
5040  Color   Benjamin Roberds                    13.0      76.0
5041  Color        Daniel Hsia                    14.0     100.0
5042  Color           Jon Gunn                    43.0      90.0

      director_facebook_likes  actor_3_facebook_likes      actor_2_name  \
0                         0.0                   855.0  Joel David Moore
1                       563.0                  1000.0     Orlando Bloom
2                         0.0                   161.0      Rory Kinnear
3                     22000.0                 23000.0    Christian Bale
4                       131.0                     NaN        Rob Walker
…                       …                     …               …
5038                      2.0                   318.0     Daphne Zuniga
5039                      NaN                   319.0     Valorie Curry
5040                      0.0                     0.0     Maxwell Moody
5041                      0.0                   489.0     Daniel Henney
5042                     16.0                    16.0  Brian Herzlinger

      actor_1_facebook_likes        gross                             genres  \
0                     1000.0  760505847.0  Action|Adventure|Fantasy|Sci-Fi
1                    40000.0  309404152.0         Action|Adventure|Fantasy
2                    11000.0  200074175.0        Action|Adventure|Thriller
3                    27000.0  448130642.0                  Action|Thriller
4                      131.0          NaN                      Documentary
…                      …           …                              …
5038                   637.0          NaN                     Comedy|Drama
5039                   841.0          NaN     Crime|Drama|Mystery|Thriller
5040                     0.0          NaN            Drama|Horror|Thriller
5041                   946.0      10443.0             Comedy|Drama|Romance
5042                    86.0      85222.0                      Documentary

      … language country  content_rating       budget title_year  \
0     …  English     USA           PG-13  237000000.0     2009.0
1     …  English     USA           PG-13  300000000.0     2007.0
2     …  English      UK           PG-13  245000000.0     2015.0
3     …  English     USA           PG-13  250000000.0     2012.0
4     …      NaN     NaN             NaN          NaN        NaN
…   …      …     …             …          …        …
5038  …  English  Canada             NaN          NaN     2013.0
5039  …  English     USA           TV-14          NaN        NaN
5040  …  English     USA             NaN       1400.0     2013.0
5041  …  English     USA           PG-13          NaN     2012.0
5042  …  English     USA              PG       1100.0     2004.0

      actor_2_facebook_likes imdb_score aspect_ratio  movie_facebook_likes  \
0                      936.0        7.9         1.78                 33000
1                     5000.0        7.1         2.35                     0
2                      393.0        6.8         2.35                 85000
```

```
3                    23000.0        8.5         2.35                164000
4                       12.0        7.1          NaN                     0
…                      …        …          …                   …
5038                   470.0        7.7          NaN                    84
5039                   593.0        7.5        16.00                 32000
5040                     0.0        6.3          NaN                    16
5041                   719.0        6.3         2.35                   660
5042                    23.0        6.6         1.85                   456

           profit
0     237000000.0
1     300000000.0
2     245000000.0
3     250000000.0
4             NaN
…           …
5038          NaN
5039          NaN
5040          NaN
5041          NaN
5042       1100.0

[5043 rows x 29 columns]
```

[360]:
```
# Q5
idx = data['imdb_score'].notna()
scores = np.array(data['imdb_score'][idx])
scores2 = np.vectorize(lambda x: np.median(scores) if x > np.quantile(scores,␣
 ↪q=0.75) or x < np.quantile(scores, q=0.25) else x)(scores)
data.loc[idx, 'imdb_score'] = scores2
scores2
```

[360]:
```
array([6.6, 7.1, 6.8, …, 6.3, 6.3, 6.6])
```

[364]:
```
# Q6
np.sum(data['num_voted_users'])
```

[364]:
```
421938535
```

[378]:
```
# Q7
counts = {}
for l in np.array(data['language']):
    if l not in counts:
        counts[l] = 0
    counts[l] += 1
sorted(list(counts.keys()), key=lambda x: counts[x], reverse=True)[0]
```

[378]:
```
'English'
```

[380]:
```
# Q8
data['actor_1_facebook_likes'].median()
```

[380]:
```
988.0
```

[381]:
```
# Q9
data['num_user_for_reviews'].mean()
```

[381]:
```
272.77080844285143
```

[382]:
```
# Q10
data['movie_facebook_likes'].std()
```

[382]:
```
19320.445109946588
```

1. imdb_score
2.
3.
4. year_group imdb_score
5.
6. language gross
7. 1 actor_1_name Facebook
8. country imdb_score
9. gross
10. content_rating

[383]:
```
data = pd.read_csv('movie_metadata.csv')
```

[384]:
```
# Q1
data['imdb_score'].describe()
```

[384]:
```
count    5043.000000
mean        6.442138
std         1.125116
min         1.600000
25%         5.800000
50%         6.600000
75%         7.200000
max         9.500000
Name: imdb_score, dtype: float64
```

[398]:
```
# Q2
data.sort_values('imdb_score', ascending=False).head(1)[['movie_title',␣
 ↪'director_name']]
```

[398]:
```
                        movie_title    director_name
2765  Towering Inferno                 John Blanchard
```

[414]:
```
# Q3
pd.get_dummies(data['genres'].str.split('|', expand=True).stack()).
 ↪groupby(level=0).sum().sum(axis=0).sort_values(ascending=False)
```

[414]:
```
Drama          2594
Comedy         1872
Thriller       1411
Action         1153
Romance        1107
Adventure       923
Crime           889
Sci-Fi          616
Fantasy         610
Horror          565
Family          546
Mystery         500
Biography       293
Animation       242
Music           214
War             213
History         207
Sport           182
Musical         132
Documentary     121
Western          97
Film-Noir         6
Short             5
News              3
Reality-TV        2
Game-Show         1
dtype: int64
```

[422]:
```
# Q4
years = pd.cut(data['title_year'], bins=[1910 + k * 10 for k in range(13)],␣
 ↪labels=[1910 + k * 10 for k in range(12)]).reset_index()
years = years.rename(columns={'title_year': 'year_group'})['year_group']
data.merge(years, left_index=True, right_index=True)[['year_group',␣
 ↪'imdb_score']].groupby('year_group').mean('imdb_score')[:-1] # 2020 NaN
```

[422]:
```
            imdb_score
year_group
1910          6.400000
1920          7.740000
1930          7.610526
```

```
1940          7.428571
1950          7.550000
1960          7.336585
1970          7.102381
1980          6.635690
1990          6.460949
2000          6.355118
2010          6.246961
```

[426]:

```
# Q5
data.sort_values(by='num_voted_users', ascending=False).head(1)[['movie_title',␣
 ↪'director_name']]
```

[426]:

```
                    movie_title   director_name
1937  The Shawshank Redemption   Frank Darabont
```

[428]:

```
# Q6
data[['language', 'gross']].groupby('language').mean('gross')
```

[428]:

```
                   gross
language
Aboriginal  3.934039e+07
Arabic      8.409155e+05
Aramaic     4.992630e+05
Bosnian     3.013050e+05
Cantonese   6.429425e+06
Chinese     5.000000e+04
Czech       6.172280e+05
Danish      8.012857e+05
Dari        8.462619e+06
Dutch       1.884888e+06
Dzongkha    5.052950e+05
English     5.102552e+07
Filipino    1.016650e+07
French      4.852977e+06
German      2.916576e+06
Greek       1.101970e+05
Hebrew      1.088493e+06
Hindi       2.217130e+06
Hungarian   1.958880e+05
Icelandic   1.183500e+04
Indonesian  2.294672e+06
Italian     4.697477e+06
Japanese    4.768039e+06
Kannada              NaN
Kazakh      7.723100e+04
Korean      1.100612e+06
```

```
Mandarin    9.089529e+06
Maya        5.085989e+07
Mongolian   5.701643e+06
None        2.601847e+06
Norwegian   4.511372e+05
Panjabi              NaN
Persian     2.284408e+06
Polish      1.573547e+06
Portuguese  2.262183e+06
Romanian    1.185783e+06
Russian     7.237200e+05
Slovenian            NaN
Spanish     8.577084e+06
Swahili              NaN
Swedish     9.939000e+04
Tamil                NaN
Telugu      6.498000e+06
Thai        4.153943e+06
Urdu                 NaN
Vietnamese  6.389510e+05
Zulu        2.912363e+06
```

[430]:
```
# Q7
data.sort_values('actor_1_facebook_likes', ascending=False).
 ↪head(1)[['movie_title', 'director_name']]
```

[430]:
```
                                 movie_title director_name
1902  Anchorman: The Legend of Ron Burgundy     Adam McKay
```

[429]:
```
data.columns
```

[429]:
```
Index(['color', 'director_name', 'num_critic_for_reviews', 'duration',
       'director_facebook_likes', 'actor_3_facebook_likes', 'actor_2_name',
       'actor_1_facebook_likes', 'gross', 'genres', 'actor_1_name',
       'movie_title', 'num_voted_users', 'cast_total_facebook_likes',
       'actor_3_name', 'facenumber_in_poster', 'plot_keywords',
       'movie_imdb_link', 'num_user_for_reviews', 'language', 'country',
       'content_rating', 'budget', 'title_year', 'actor_2_facebook_likes',
       'imdb_score', 'aspect_ratio', 'movie_facebook_likes'],
      dtype='object')
```

[433]:
```
# Q8
data[['country', 'imdb_score']].sort_values(by='country', ascending=False).
 ↪groupby('country').mean('imdb_score')
```

[433]:
```
                      imdb_score
country
```

```
Afghanistan             7.400000
Argentina               7.500000
Aruba                   4.800000
Australia               6.514545
Bahamas                 4.400000
…                          …
Turkey                  6.000000
UK                      6.818304
USA                     6.367428
United Arab Emirates    8.200000
West Germany            7.266667

[65 rows x 1 columns]
```

```
[434]: # Q9
       subdata = data[data['num_voted_users'].notna() & data['gross'].notna()]
       np.corrcoef(np.array(data['num_voted_users'], data['gross']))
```

```
[434]: 1.0
```

```
[453]: # 10
       data['content_rating'].value_counts().sort_values(ascending=True)
```

```
[453]: TV-Y7           1
       TV-Y            1
       M               5
       GP              6
       NC-17           7
       Passed          9
       TV-G           10
       X              13
       TV-PG          13
       TV-MA          20
       TV-14          30
       Approved       55
       Unrated        62
       G             112
       Not Rated     116
       PG            701
       PG-13        1461
       R            2118
       Name: content_rating, dtype: int64
```

```
[ ]:
```