

```

#Import necessary library

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats

''' Loading movies dataset where information is related to budget
genres homepage
which is essential to movie production. To decide whether the movie is
performing well or whats the reason for its
performance. '''

df = pd.read_csv('/content/movies.csv')
df
{"type": "dataframe", "variable_name": "df"}
df.head()
{"type": "dataframe", "variable_name": "df"}
df.columns
Index(['index', 'budget', 'genres', 'homepage', 'id', 'keywords',
      'original_language', 'original_title', 'overview',
      'popularity',
      'production_companies', 'production_countries', 'release_date',
      'revenue', 'runtime', 'spoken_languages', 'status', 'tagline',
      'title',
      'vote_average', 'vote_count', 'cast', 'crew', 'director'],
      dtype='object')

df.shape
(4803, 24)

''' If we observe the dataset carefully there are null value within
the dataset
so before we move into analysis part its important to deal with null
values. '''

df.isnull().sum()

```

index	0
budget	0
genres	28
homepage	3091
id	0
keywords	412

```
original_language    0
original_title       0
overview            3
popularity           0
production_companies 0
production_countries 0
release_date        1
revenue             0
runtime             2
spoken_languages     0
status              0
tagline             844
title               0
vote_average         0
vote_count           0
cast                43
crew                 0
director            30
dtype: int64
```

```
df.dtypes
```

```
index                int64
budget              int64
genres              object
homepage            object
id                  int64
keywords            object
original_language   object
original_title      object
overview            object
popularity          float64
production_companies object
production_countries object
release_date        object
revenue             int64
runtime             float64
spoken_languages    object
status              object
tagline             object
title               object
vote_average        float64
vote_count          int64
cast                object
crew                object
director            object
dtype: object
```

*''' Here firstly there is one unnecessary column which is index but we have row numbers so that column is not needed.*

Similarly we dropped the homepage since its just the web link. Then original title is dropped because we have column related to title so it makes no sence to have two column with similar meaning. We also drop Overview column and prodution countries.'''

```
columns =
['homepage', 'index', 'original_title', 'overview', 'production_countries']
```

```
df.drop(columns,axis=1,inplace=True)
```

```
df.head()
```

```

{"summary":{"\n  \"name\": \"df\",\n  \"rows\": 4803,\n  \"fields\": [\n    {\n      \"column\": \"budget\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 40722391,\n        \"min\": 0,\n        \"max\": 380000000,\n        \"num_unique_values\": 436,\n        \"samples\": [\n          439000,\n          68000000,\n          700000\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\",\n        \"genres\": {\n          \"properties\": {\n            \"dtype\": \"category\",\n            \"num_unique_values\": 1168,\n            \"samples\": [\n              \"Adventure Family\",\n              \"Comedy Mystery Thriller\",\n              \"Drama Action Adventure Fantasy\"\n            ],\n            \"semantic_type\": \"\",\n            \"description\": \"\"\n          },\n          \"column\": \"id\",\n          \"properties\": {\n            \"dtype\": \"number\",\n            \"std\": 88694,\n            \"min\": 5,\n            \"max\": 459488,\n            \"num_unique_values\": 4803,\n            \"samples\": [\n              8427,\n              13006,\n              18041\n            ],\n            \"semantic_type\": \"\",\n            \"description\": \"\"\n          },\n          \"column\": \"keywords\",\n          \"properties\": {\n            \"dtype\": \"string\",\n            \"num_unique_values\": 4219,\n            \"samples\": [\n              \"revolution political activism loser\",\n              \"riddle captain treasure liquor treasure hunt\",\n              \"male nudity parents kids relationship camping cutting the cord self-discovery\"\n            ],\n            \"semantic_type\": \"\",\n            \"description\": \"\"\n          },\n          \"column\": \"original_language\",\n          \"properties\": {\n            \"dtype\": \"category\",\n            \"num_unique_values\": 37,\n            \"samples\": [\n              \"ta\",\n              \"es\"\n            ],\n            \"semantic_type\": \"\",\n            \"description\": \"\"\n          },\n          \"column\": \"popularity\",\n          \"properties\": {\n            \"dtype\": \"number\",\n            \"std\": 31.816649749537802,\n            \"min\": 0.0,\n            \"max\": 875.581305,\n            \"num_unique_values\": 4802,\n            \"samples\": [\n              13.267631,\n              0.010909,\n              5.842299\n            ],\n            \"semantic_type\": \"\",\n            \"description\": \"\"\n          },\n          \"column\": \"production companies\",\n          \"properties\": {\n            \"dtype\": \"string\"

```

```

\"string\", \n          \"num_unique_values\": 3697, \n
\"samples\": [ \n          \"[{\\\"name\\\": \\\"Paramount
Pictures\\\", \\\"id\\\": 4}, {\\\"name\\\": \\\"Cherry Alley
Productions\\\", \\\"id\\\": 2232}]\", \n          \"[{\\\"name\\\":
\\\"Twentieth Century Fox Film Corporation\\\", \\\"id\\\": 306},
{\\\"name\\\": \\\"Dune Entertainment\\\", \\\"id\\\": 444},
{\\\"name\\\": \\\"Regency Enterprises\\\", \\\"id\\\": 508},
{\\\"name\\\": \\\"Guy Walks into a Bar Productions\\\", \\\"id\\\":
2645}, {\\\"name\\\": \\\"Deep River Productions\\\", \\\"id\\\":
2646}, {\\\"name\\\": \\\"Friendly Films (II)\\\", \\\"id\\\":
81136}]\", \n          \"[{\\\"name\\\": \\\"Twentieth Century Fox Film
Corporation\\\", \\\"id\\\": 306}]\" \n          ], \n
\"semantic_type\": \"\", \n          \"description\": \"\" \n          } \n
}, \n          { \n          \"column\": \"release_date\", \n
\"properties\": { \n          \"dtype\": \"object\", \n
\"num_unique_values\": 3280, \n          \"samples\": [ \n
\"1966-10-16\", \n          \"1987-07-31\", \n          \"1993-09-23\" \n
], \n          \"semantic_type\": \"\", \n          \"description\": \"\" \n
} \n          }, \n          { \n          \"column\": \"revenue\", \n
\"properties\": { \n          \"dtype\": \"number\", \n          \"std\":
162857100, \n          \"min\": 0, \n          \"max\": 2787965087, \n
\"num_unique_values\": 3297, \n          \"samples\": [ \n
11833696, \n          10462500, \n          17807569 \n
], \n          \"semantic_type\": \"\", \n          \"description\": \"\" \n
} \n          }, \n          { \n          \"column\": \"runtime\", \n
\"properties\": { \n          \"dtype\": \"number\", \n          \"std\":
22.611934588844207, \n          \"min\": 0.0, \n          \"max\": 338.0, \n
\"num_unique_values\": 156, \n          \"samples\": [ \n
74.0, \n
85.0, \n
170.0 \n
], \n          \"semantic_type\": \"\", \n
\"description\": \"\" \n          }, \n          { \n          \"column\":
\"spoken_languages\", \n          \"properties\": { \n          \"dtype\":
\"category\", \n          \"num_unique_values\": 544, \n
\"samples\": [ \n          \"[{\\\"iso_639_1\\\": \\\"es\\\",
\\\"name\\\": \\\"Espa\\\"\\u00f1ol\\\"}, {\\\"iso_639_1\\\":
\\\"en\\\", \\\"name\\\": \\\"English\\\"}, {\\\"iso_639_1\\\":
\\\"fr\\\", \\\"name\\\": \\\"Fran\\\"\\u00e7ais\\\"},
{\\\"iso_639_1\\\": \\\"hu\\\", \\\"name\\\": \\\"Magyar\\\"}]\", \n
\"[{\\\"iso_639_1\\\": \\\"en\\\", \\\"name\\\": \\\"English\\\"},
{\\\"iso_639_1\\\": \\\"it\\\", \\\"name\\\": \\\"Italiano\\\"},
{\\\"iso_639_1\\\": \\\"pt\\\", \\\"name\\\": \\\"Portugu\\\"\\u00eas\\\"}]\", \n
          \"[{\\\"iso_639_1\\\": \\\"de\\\",
\\\"name\\\": \\\"Deutsch\\\"}, {\\\"iso_639_1\\\": \\\"it\\\",
\\\"name\\\": \\\"Italiano\\\"}, {\\\"iso_639_1\\\": \\\"la\\\",
\\\"name\\\": \\\"Latin\\\"}, {\\\"iso_639_1\\\": \\\"pl\\\",
\\\"name\\\": \\\"Polski\\\"}]\" \n          ], \n
\"semantic_type\": \"\", \n          \"description\": \"\" \n          } \n
}, \n          { \n          \"column\": \"status\", \n
\"properties\": { \n          \"dtype\": \"category\", \n          \"num_unique_values\":
3, \n          \"samples\": [ \n          \"Released\", \n          \"Post

```

```

Production\", \n          \"Rumored\" \n          ], \n
\"semantic_type\": \"\", \n          \"description\": \"\" \n          } \n
    }, \n    { \n          \"column\": \"tagline\", \n          \"properties\":
{ \n          \"dtype\": \"string\", \n          \"num_unique_values\":
3944, \n          \"samples\": [ \n          \"When you're 17, every day
is war.\", \n          \"An Unspeakable Horror. A Creative Genius.
Captured For Eternity.\", \n          \"May the schwartz be with you\" \n
          ], \n          \"semantic_type\": \"\", \n
          \"description\": \"\" \n          } \n    }, \n    { \n          \"column\":
\"title\", \n          \"properties\": { \n          \"dtype\": \"string\", \n
          \"num_unique_values\": 4800, \n          \"samples\": [ \n          \"I
Spy\", \n          \"Who's Your Caddy?\", \n          \"Sleepover\" \n
          ], \n          \"semantic_type\": \"\", \n          \"description\": \"\" \n
          } \n    }, \n    { \n          \"column\": \"vote_average\", \n
          \"properties\": { \n          \"dtype\": \"number\", \n          \"std\":
1.1946121628478925, \n          \"min\": 0.0, \n          \"max\": 10.0, \n
          \"num_unique_values\": 71, \n          \"samples\": [ \n          5.1, \n
          7.2, \n          4.0 \n          ], \n          \"semantic_type\": \"\", \n
          \"description\": \"\" \n          } \n    }, \n    { \n          \"column\":
\"vote_count\", \n          \"properties\": { \n          \"dtype\":
\"number\", \n          \"std\": 1234, \n          \"min\": 0, \n
          \"max\": 13752, \n          \"num_unique_values\": 1609, \n
          \"samples\": [ \n          7604, \n          3428, \n          225 \n
          ], \n          \"semantic_type\": \"\", \n          \"description\": \"\" \n
          } \n    }, \n    { \n          \"column\": \"cast\", \n          \"properties\":
{ \n          \"dtype\": \"string\", \n          \"num_unique_values\":
4741, \n          \"samples\": [ \n          \"Al Gore Billy West Ronald
Reagan George W. Bush George H. W. Bush\", \n          \"David
Schwimmer Gwyneth Paltrow Michael Rapaport Toni Collette Carol
Kane\", \n          \"Paul Scofield Wendy Hiller Leo McKern Robert Shaw
Orson Welles\" \n          ], \n          \"semantic_type\": \"\", \n
          \"description\": \"\" \n          } \n    }, \n    { \n          \"column\":
\"crew\", \n          \"properties\": { \n          \"dtype\": \"string\", \n
          \"num_unique_values\": 4776, \n          \"samples\": [ \n
          \"[{ 'name': 'Mark Goldblatt', 'gender': 2, 'department': 'Editing',
'job': 'Editor', 'credit_id': '538d82720e0a26670e005e83', 'id': 898},
{ 'name': 'Mark Helfrich', 'gender': 2, 'department': 'Editing', 'job':
'Editor', 'credit_id': '538d82860e0a266702005de5', 'id': 1098},
{ 'name': 'Ralph Winter', 'gender': 0, 'department': 'Production',
'job': 'Producer', 'credit_id': '538d820c0e0a266708006129', 'id':
2095}, { 'name': 'Richard Pagano', 'gender': 2, 'department':
'Production', 'job': 'Casting', 'credit_id':
'538d82d30e0a266712005def', 'id': 2624}, { 'name': 'Ed Verreaux',
'gender': 2, 'department': 'Art', 'job': 'Production Design',
'credit_id': '538d82ea0e0a266715005e74', 'id': 4953}, { 'name': 'Heike
Brandstätter', 'gender': 1, 'department': 'Production', 'job':
'Casting', 'credit_id': '538d82ad0e0a26670800614d', 'id': 5362},
{ 'name': 'Coreen Mayrs', 'gender': 1, 'department': 'Production',
'job': 'Casting', 'credit_id': '538d82c00e0a266705005fa2', 'id':

```

5363}, {'name': 'Dan Moore', 'gender': 0, 'department': 'Costume & Make-Up', 'job': 'Costume Supervisor', 'credit\_id': '538d845ec3a368713c0047d2', 'id': 6924}, {'name': 'Lauren Shuler Donner', 'gender': 1, 'department': 'Production', 'job': 'Producer', 'credit\_id': '538d81ef0e0a266718005cb8', 'id': 7200}, {'name': 'Stan Lee', 'gender': 2, 'department': 'Production', 'job': 'Executive Producer', 'credit\_id': '576710bec3a36806720004b0', 'id': 7624}, {'name': 'Avi Arad', 'gender': 2, 'department': 'Production', 'job': 'Producer', 'credit\_id': '538d81d80e0a266715005e4d', 'id': 7626}, {'name': 'Kevin Feige', 'gender': 2, 'department': 'Production', 'job': 'Executive Producer', 'credit\_id': '5767101092514165c20000db', 'id': 10850}, {'name': 'Judianna Makovsky', 'gender': 1, 'department': 'Costume & Make-Up', 'job': 'Costume Design', 'credit\_id': '538d83ba0e0a26670b005ded', 'id': 10970}, {'name': 'Justin Scoppa Jr.', 'gender': 2, 'department': 'Art', 'job': 'Art Direction', 'credit\_id': '538d83620e0a266715005e82', 'id': 10549}, {'name': 'John Powell', 'gender': 2, 'department': 'Sound', 'job': 'Original Music Composer', 'credit\_id': '5626465bc3a3680e0e01704b', 'id': 11098}, {'name': 'Simon Kinberg', 'gender': 2, 'department': 'Writing', 'job': 'Screenplay', 'credit\_id': '52fe45ff9251416c910459e1', 'id': 11092}, {'name': 'Bill Corso', 'gender': 2, 'department': 'Crew', 'job': 'Makeup Effects', 'credit\_id': '56abe4ecc3a3681c47003aa3', 'id': 15017}, {'name': 'Dante Spinotti', 'gender': 2, 'department': 'Camera', 'job': 'Director of Photography', 'credit\_id': '538d82560e0a266705005f94', 'id': 11099}, {'name': 'Zak Penn', 'gender': 2, 'department': 'Writing', 'job': 'Screenplay', 'credit\_id': '52fe45ff9251416c910459e7', 'id': 11011}, {'name': 'Brett Ratner', 'gender': 2, 'department': 'Directing', 'job': 'Director', 'credit\_id': '52fe45ff9251416c910459db', 'id': 11091}, {'name': 'Helen Jarvis', 'gender': 1, 'department': 'Art', 'job': 'Art Direction', 'credit\_id': '538d834f0e0a266712005dfe', 'id': 22061}, {'name': 'Elizabeth Wilcox', 'gender': 1, 'department': 'Art', 'job': 'Set Decoration', 'credit\_id': '538d83a50e0a266715005e8b', 'id': 11021}, {'name': 'Chad S. Frey', 'gender': 0, 'department': 'Art', 'job': 'Art Direction', 'credit\_id': '538d83220e0a266702005df6', 'id': 11102}, {'name': 'Geoff Hubbard', 'gender': 0, 'department': 'Art', 'job': 'Supervising Art Director', 'credit\_id': '570b8dde9251412c740020d0', 'id': 46082}, {'name': 'Lisa Tomczeszyn', 'gender': 2, 'department': 'Costume & Make-Up', 'job': 'Costume Design', 'credit\_id': '538d83ce0e0a266702005e02', 'id': 11106}, {'name': 'James Bamford', 'gender': 0, 'department': 'Crew', 'job': 'Stunts', 'credit\_id': '560046499251414e4d000197', 'id': 113194}, {'name': 'Julia Wong', 'gender': 1, 'department': 'Editing', 'job': 'Editor', 'credit\_id': '538d82990e0a266712005de9', 'id': 11100}, {'name': 'James M. Freitag', 'gender': 2, 'department': 'Production', 'job': 'Co-Producer', 'credit\_id': '57671042c3a36806b3000104', 'id': 11095}, {'name': 'Ross Fanger', 'gender': 2, 'department': 'Production', 'job': 'Co-Producer', 'credit\_id': '57670fdbbc3a3680722000106', 'id': 11014}, {'name': 'Sandi Tanaka', 'gender': 0, 'department': 'Art', 'job': 'Art

```

Direction', 'credit_id': '538d83750e0a26670b005de5', 'id': 223991},
{'name': 'John Palermo', 'gender': 0, 'department': 'Production',
'job': 'Executive Producer', 'credit_id': '576710f192514133c5001213',
'id': 1182092}, {'name': 'John Bruno', 'gender': 0, 'department':
'Visual Effects', 'job': 'Visual Effects Supervisor', 'credit_id':
'55402d2fc3a3681da0000633', 'id': 1395269}, {'name': 'Guy Williams',
'gender': 0, 'department': 'Crew', 'job': 'Digital Effects
Supervisor', 'credit_id': '563a9ed79251416d520145a0', 'id': 1401803},
{'name': 'Kurt Williams', 'gender': 0, 'department': 'Production',
'job': 'Co-Producer', 'credit_id': '5767112fc3a3680722000132', 'id':
1432021}, {'name': 'Rachel Wyn Dunn', 'gender': 0, 'department':
'Crew', 'job': 'Compositors', 'credit_id': '55423048c3a36829e5003bc8',
'id': 1447557}, {'name': 'Joel Fletcher', 'gender': 0, 'department':
'Visual Effects', 'job': 'Animation', 'credit_id':
'552986b4c3a368596e001a79', 'id': 1452932}, {'name': 'Lee Cleary',
'gender': 0, 'department': 'Production', 'job': 'Co-Producer',
'credit_id': '57670f92925141655a000109', 'id': 1477146}, {'name':
'David Gorder', 'gender': 0, 'department': 'Production', 'job':
'Associate Producer', 'credit_id': '5767107c925141655a00012c', 'id':
1477147}, {'name': 'Luca Marco Paracels', 'gender': 2, 'department':
'Production', 'job': 'Associate Producer', 'credit_id':
'5767105bc3a36850c4000573', 'id': 1477203}]]", "\n" [{"name':
'Jeanne McCarthy', 'gender': 1, 'department': 'Production', 'job':
'Casting', 'credit_id': '58dd0e90c3a368223e00b59c', 'id': 546},
{'name': 'John Davis', 'gender': 2, 'department': 'Production', 'job':
'Producer', 'credit_id': '52fe43bb9251416c7501bf83', 'id': 2043},
{'name': 'David Newman', 'gender': 2, 'department': 'Sound', 'job':
'Original Music Composer', 'credit_id': '5539fe35c3a3685cf8002f1e',
'id': 3393}, {'name': "\\\"Aldric La'Auli Porter\\\"", 'gender': 0,
'department': 'Production', 'job': 'Associate Producer', 'credit_id':
'58dd0d3fc3a368228a00aa50', 'id': 6187}, {'name': "\\\"Aldric La'Auli
Porter\\\"", 'gender': 0, 'department': 'Directing', 'job': 'First
Assistant Director', 'credit_id': '5990427f925141672302ba00', 'id':
6187}, {'name': 'William Sandell', 'gender': 2, 'department': 'Art',
'job': 'Production Design', 'credit_id': '58dd0efdc3a368228300b438',
'id': 6877}, {'name': 'Robert Gould', 'gender': 2, 'department':
'Art', 'job': 'Set Decoration', 'credit_id':
'58dd0f56925141411000c340', 'id': 6880}, {'name': 'Brad Ricker',
'gender': 2, 'department': 'Art', 'job': 'Art Direction', 'credit_id':
'58dd0f1fc3a36822a600c1dd', 'id': 8705}, {'name': 'Dan Yale',
'gender': 0, 'department': 'Sound', 'job': 'Foley Editor',
'credit_id': '5990490b925141673e02d2f5', 'id': 8763}, {'name': 'Neil
A. Machlis', 'gender': 2, 'department': 'Production', 'job':
'Executive Producer', 'credit_id': '58dd0d3192514140f600b915', 'id':
8880}, {'name': 'Juel Bestrop', 'gender': 1, 'department':
'Production', 'job': 'Casting', 'credit_id':
'58dd0e0bc3a368229300bdf6', 'id': 14377}, {'name': 'Hugo Santiago',
'gender': 0, 'department': 'Art', 'job': 'Set Designer', 'credit_id':
'599045b6925141672302bdac', 'id': 12639}, {'name': 'Ruth E. Carter',

```



'gender': 1, 'department': 'Costume & Make-Up', 'job': 'Costume Design', 'credit\_id': '58dd0f69c3a368228300b48f', 'id': 15524}, {'name': 'Craig Herring', 'gender': 2, 'department': 'Editing', 'job': 'Editor', 'credit\_id': '52fe43bb9251416c7501bf9b', 'id': 18276}, {'name': 'Bruce Crone', 'gender': 2, 'department': 'Art', 'job': 'Supervising Art Director', 'credit\_id': '58dd0f2f92514140fa00c653', 'id': 13434}, {'name': 'Douglas Smith', 'gender': 0, 'department': 'Visual Effects', 'job': 'Visual Effects Supervisor', 'credit\_id': '59904bc3a368744202b9fb', 'id': 27100}, {'name': 'Larry Levin', 'gender': 2, 'department': 'Writing', 'job': 'Screenplay', 'credit\_id': '52fe43bb9251416c7501bf7d', 'id': 31026}, {'name': 'Michele Imperato', 'gender': 1, 'department': 'Production', 'job': 'Co-Producer', 'credit\_id': '58dd0c8592514140fa00c441', 'id': 49907}, {'name': 'Michele Imperato', 'gender': 1, 'department': 'Production', 'job': 'Unit Production Manager', 'credit\_id': '58dd0fb0c3a368227300bfad', 'id': 49907}, {'name': 'Steve Carr', 'gender': 2, 'department': 'Directing', 'job': 'Director', 'credit\_id': '52fe43bb9251416c7501bfb1', 'id': 52112}, {'name': 'R.J. Kizer', 'gender': 2, 'department': 'Sound', 'job': 'Dialogue Editor', 'credit\_id': '5990471cc3a36874a302bbdc', 'id': 52193}, {'name': 'Daryn Okada', 'gender': 2, 'department': 'Camera', 'job': 'Director of Photography', 'credit\_id': '52fe43bb9251416c7501bf95', 'id': 56324}, {'name': 'Joseph Singer', 'gender': 2, 'department': 'Production', 'job': 'Producer', 'credit\_id': '52fe43bb9251416c7501bf89', 'id': 56053}, {'name': 'Peter S. Elliot', 'gender': 2, 'department': 'Crew', 'job': 'Visual Effects Editor', 'credit\_id': '59904a06925141676602b701', 'id': 57769}, {'name': 'Heidi Santelli', 'gender': 1, 'department': 'Production', 'job': 'Co-Producer', 'credit\_id': '58dd0d4c92514140fa00c4da', 'id': 66921}, {'name': 'Tom Villano', 'gender': 0, 'department': 'Sound', 'job': 'Music Editor', 'credit\_id': '59905155c3a36874b402a906', 'id': 91324}, {'name': 'Judy Murdock', 'gender': 0, 'department': 'Costume & Make-Up', 'job': 'Makeup Artist', 'credit\_id': '5990424a925141676602adf3', 'id': 168214}, {'name': 'Bruce McBroom', 'gender': 2, 'department': 'Camera', 'job': 'Still Photographer', 'credit\_id': '59904eae925141673e02da30', 'id': 555085}, {'name': 'Kris Evans', 'gender': 0, 'department': 'Costume & Make-Up', 'job': 'Makeup Department Head', 'credit\_id': '599041d0c3a36874a302b5e9', 'id': 1015965}, {'name': 'Ido Lampton Enochs Jr.', 'gender': 0, 'department': 'Production', 'job': 'Location Manager', 'credit\_id': '58dd11a2c3a368227300c14b', 'id': 1113659}, {'name': 'Liz Radley', 'gender': 0, 'department': 'Crew', 'job': 'CG Supervisor', 'credit\_id': '59904b71925141675f02b7e7', 'id': 1178615}, {'name': 'Donna Marcione Pollack', 'gender': 0, 'department': 'Costume & Make-Up', 'job': 'Costume Supervisor', 'credit\_id': '59904fe5c3a368744202bf0c', 'id': 1319193}, {'name': 'Frank McElDowney', 'gender': 0, 'department': 'Art', 'job': 'Greensman', 'credit\_id': '599044c4925141672302bc97', 'id': 1364403}, {'name': 'Ted Caplan', 'gender': 0, 'department': 'Sound', 'job': 'Dialogue Editor',



'credit\_id': '59904694925141673e02cfd2', 'id': 1367505}, {'name': 'Gerald Lehtola', 'gender': 0, 'department': 'Art', 'job': 'Art Department Coordinator', 'credit\_id': '5990439d925141675902ac1b', 'id': 1377216}, {'name': 'Mark Wade', 'gender': 0, 'department': 'Art', 'job': 'Property Master', 'credit\_id': '599045fb925141675f02b16c', 'id': 1378165}, {'name': 'Patricia Klawonn', 'gender': 1, 'department': 'Art', 'job': 'Set Designer', 'credit\_id': '58dd0fd0c3a368228a00abfb', 'id': 1378749}, {'name': 'Bill Taliaferro', 'gender': 0, 'department': 'Art', 'job': 'Set Designer', 'credit\_id': '58dd1023c3a368227300c025', 'id': 1389585}, {'name': 'Dave Kulczycki', 'gender': 0, 'department': 'Sound', 'job': 'Sound Effects Editor', 'credit\_id': '59904785c3a3687493029126', 'id': 1390527}, {'name': 'Elizabeth Frank', 'gender': 0, 'department': 'Costume & Make-Up', 'job': 'Set Costumer', 'credit\_id': '59905010925141675f02bd38', 'id': 1399322}, {'name': 'Casey Hotchkiss', 'gender': 2, 'department': 'Camera', 'job': 'Camera Operator', 'credit\_id': '59904d81925141670b02cf99', 'id': 1400374}, {'name': 'Zack Davis', 'gender': 0, 'department': 'Sound', 'job': 'ADR Editor', 'credit\_id': '599046d2c3a3687493029074', 'id': 1404841}, {'name': 'Erma Kent', 'gender': 0, 'department': 'Costume & Make-Up', 'job': 'Hair Department Head', 'credit\_id': '59904214c3a368746102979f', 'id': 1408278}, {'name': 'Donald Sylvester', 'gender': 0, 'department': 'Sound', 'job': 'Supervising Sound Editor', 'credit\_id': '599048b2925141670b02c97f', 'id': 1416153}, {'name': 'John Kilkenny', 'gender': 0, 'department': 'Visual Effects', 'job': 'Visual Effects Producer', 'credit\_id': '59904ad6c3a36874b402a1cc', 'id': 1416167}, {'name': 'Camille Friend', 'gender': 1, 'department': 'Costume & Make-Up', 'job': 'Hairstylist', 'credit\_id': '56ae7d849251416357002915', 'id': 1424894}, {'name': 'Vanessa Lapato', 'gender': 0, 'department': 'Sound', 'job': 'ADR Editor', 'credit\_id': '599047ab925141674c02b48e', 'id': 1425343}, {'name': 'Alan E. Lorimer', 'gender': 0, 'department': 'Crew', 'job': 'Special Effects Coordinator', 'credit\_id': '59904951925141673902ae4e', 'id': 1442509}, {'name': 'Bryan McBrien', 'gender': 0, 'department': 'Art', 'job': 'Greensman', 'credit\_id': '59904427925141675902acc6', 'id': 1451398}, {'name': 'Butch West', 'gender': 0, 'department': 'Art', 'job': 'Construction Coordinator', 'credit\_id': '59904629c3a368744202b2d6', 'id': 1453276}, {'name': 'Hugh Lofting', 'gender': 2, 'department': 'Crew', 'job': 'Creator', 'credit\_id': '599034bcc3a3687493027d17', 'id': 1454647}, {'name': 'Nancy Karlin', 'gender': 1, 'department': 'Directing', 'job': 'Script Supervisor', 'credit\_id': '58dd11flc3a36822ae00bc75', 'id': 1457729}, {'name': 'Mitchell Ferm', 'gender': 0, 'department': 'Visual Effects', 'job': 'Visual Effects Coordinator', 'credit\_id': '59904a8d925141675f02b6d8', 'id': 1496418}, {'name': 'Jim Harrison', 'gender': 2, 'department': 'Sound', 'job': 'Music Editor', 'credit\_id': '599050e6925141674c02beef', 'id': 1550058}, {'name': 'Jasan Sherman', 'gender': 0, 'department': 'Art', 'job': 'Location Scout', 'credit\_id': '58dd11b0c3a368223e00b7db', 'id': 1585159},

```
{'name': 'Thomas Meikle', 'gender': 0, 'department': 'Art', 'job':  
'Sculptor', 'credit_id': '5990453a925141673902a8ec', 'id': 1586888},  
{'name': 'Spring Aspers', 'gender': 0, 'department': 'Sound', 'job':  
'Music Supervisor', 'credit_id': '5990506dc3a368747302c6e3', 'id':  
1587383}, {'name': 'Craig Anthony', 'gender': 0, 'department':  
'Costume & Make-Up', 'job': 'Key Costumer', 'credit_id':  
'59904fbb925141672302c9ed', 'id': 1610228}, {'name': 'Harrison D.  
Marsh', 'gender': 2, 'department': 'Sound', 'job': 'Boom Operator',  
'credit_id': '599047e3925141674c02b4cf', 'id': 1737829}, {'name':  
'Nadia Lubbe Simon', 'gender': 1, 'department': 'Production', 'job':  
'Casting Assistant', 'credit_id': '58dd10b2c3a368228300b588', 'id':  
1763970}, {'name': 'Diana Jaher', 'gender': 1, 'department':  
'Production', 'job': 'Casting Associate', 'credit_id':  
'58dd1052925141413c00beac', 'id': 1786614}, {'name': 'Craig W. van  
Gundy', 'gender': 0, 'department': 'Production', 'job': 'Location  
Manager', 'credit_id': '58dd11d5c3a368229d00b6a0', 'id': 1786616},  
{'name': 'Kris Fullan Martinez', 'gender': 1, 'department':  
'Production', 'job': 'Production Coordinator', 'credit_id':  
'58dd1254c3a368227300cle6', 'id': 1786618}, {'name': 'Joe A.  
Hawthorne', 'gender': 0, 'department': 'Art', 'job': 'Painter',  
'credit_id': '599042f4c3a368744202aefa', 'id': 1868861}, {'name':  
'Robert McDonald', 'gender': 0, 'department': 'Crew', 'job':  
'Propmaker', 'credit_id': '5990448e925141675f02afa7', 'id': 1868865},  
{'name': 'Robert Prchal', 'gender': 0, 'department': 'Crew', 'job':  
'Propmaker', 'credit_id': '59904571925141670b02c572', 'id': 1868867},  
{'name': 'Hal Schlax', 'gender': 0, 'department': 'Crew', 'job':  
'Propmaker', 'credit_id': '599045da925141673902a9d7', 'id': 1868868},  
{'name': 'John Molo', 'gender': 2, 'department': 'Crew', 'job': 'Stunt  
Coordinator', 'credit_id': '59904c34925141675f02b8bc', 'id': 1868877},  
{'name': 'Todd Dos Reis', 'gender': 0, 'department': 'Camera', 'job':  
'Camera Operator', 'credit_id': '59904cac925141673e02d7a8', 'id':  
1868878}, {'name': 'Jason D. Hodges', 'gender': 0, 'department':  
'Camera', 'job': 'Dolly Grip', 'credit_id':  
'59904d04925141676602ba9a', 'id': 1868879}, {'name': 'Mike Hodges',  
'gender': 0, 'department': 'Camera', 'job': 'Key Grip', 'credit_id':  
'59904d3bc3a36874b402a4b0', 'id': 1868880}, {'name': 'Kal Manning',  
'gender': 0, 'department': 'Lighting', 'job': 'Gaffer', 'credit_id':  
'59904df0925141672302c7da', 'id': 1868881}, {'name': 'Larry McCarron',  
'gender': 0, 'department': 'Camera', 'job': 'Dolly Grip', 'credit_id':  
'59904ede925141675f02bbe6', 'id': 1868889}, {'name': 'Kim Strand',  
'gender': 0, 'department': 'Sound', 'job': 'Music Editor',  
'credit_id': '59905109925141673e02dcde', 'id': 1868892}]\n",\n\n"[{'name': 'Harry Manfredini', 'gender': 2, 'department': 'Sound',  
'job': 'Original Music Composer', 'credit_id':  
'52fe44479251416c7502f0c3', 'id': 20953}, {'name': 'Robin D. Cook',  
'gender': 0, 'department': 'Production', 'job': 'Casting',  
'credit_id': '534a8875c3a3682df60016a4', 'id': 23545}, {'name': 'Sean  
S. Cunningham', 'gender': 2, 'department': 'Production', 'job':  
'Executive Producer', 'credit_id': '52fe44479251416c7502f0ed', 'id':
```

```

35475}, {'name': 'Derick V. Underschultz', 'gender': 2, 'department':
'Camera', 'job': 'Director of Photography', 'credit_id':
'52fe44479251416c7502f0cf', 'id': 55642}, {'name': 'Todd Farmer',
'gender': 2, 'department': 'Writing', 'job': 'Screenplay',
'credit_id': '52fe44479251416c7502f0b7', 'id': 61117}, {'name': 'David
Handman', 'gender': 2, 'department': 'Editing', 'job': 'Editor',
'credit_id': '52fe44479251416c7502f0c9', 'id': 64748}, {'name': 'James
Isaac', 'gender': 2, 'department': 'Directing', 'job': 'Director',
'credit_id': '52fe44479251416c7502f0b1', 'id': 69579}, {'name': 'Noel
Cunningham', 'gender': 2, 'department': 'Production', 'job':
'Producer', 'credit_id': '52fe44479251416c7502f0bd', 'id': 69585},
{'name': 'Maxyne Baker', 'gender': 0, 'department': 'Costume & Make-
Up', 'job': 'Costume Design', 'credit_id': '534a888cc3a3682df60016a8',
'id': 958703}]\\n          ],\\n          \\\"semantic_type\\\": \\\"\\\",\\n
\\\"description\\\": \\\"\\\"\\n          }\\n          },\\n          {\\n          \\\"column\\\":
\\\"director\\\",\\n          \\\"properties\\\": {\\n          \\\"dtype\\\":
\\\"category\\\",\\n          \\\"num_unique_values\\\": 2349,\\n
\\\"samples\\\": [\\n          \\\"Hart Bochner\\\",\\n          \\\"David
Wain\\\",\\n          \\\"Ham Tran\\\"\\n          ],\\n
\\\"semantic_type\\\": \\\"\\\",\\n          \\\"description\\\": \\\"\\\"\\n          }\\n
          }\\n          ]\\n          }\", \"type\": \"dataframe\", \"variable_name\": \"df\"}

```

*#Changing column names for better understanding.*

```

header = ['Budget', 'Genres', 'Id', 'Keywords', 'Original language',
'Popularity', 'Production companies', 'Release
date', 'Revenue', 'Runtime', 'Spoken languages',
'Status', 'Tagline', 'Title', 'Vote average', 'Vote
count', 'Cast', 'Crew', 'Director']

```

```
df.columns = header
```

```
df.head()
```

```

{"summary": "{\\n  \\\"name\\\": \\\"df\\\",\\n  \\\"rows\\\": 4803,\\n  \\\"fields\\\":
[\\n    {\\n      \\\"column\\\": \\\"Budget\\\",\\n      \\\"properties\\\": {\\n
\\\"dtype\\\": \\\"number\\\",\\n      \\\"std\\\": 40722391,\\n      \\\"min\\\":
0,\\n      \\\"max\\\": 380000000,\\n      \\\"num_unique_values\\\": 436,\\n
\\\"samples\\\": [\\n        439000,\\n        68000000,\\n
700000\\n      ],\\n      \\\"semantic_type\\\": \\\"\\\",\\n
\\\"description\\\": \\\"\\\"\\n    }\\n    },\\n    {\\n      \\\"column\\\":
\\\"Genres\\\",\\n      \\\"properties\\\": {\\n      \\\"dtype\\\":
\\\"category\\\",\\n      \\\"num_unique_values\\\": 1168,\\n
\\\"samples\\\": [\\n        \\\"Adventure Family\\\",\\n        \\\"Comedy
Mystery Thriller\\\",\\n        \\\"Drama Action Adventure Fantasy\\\"\\n
],\\n      \\\"semantic_type\\\": \\\"\\\",\\n      \\\"description\\\": \\\"\\\"\\n
}\\n    },\\n    {\\n      \\\"column\\\": \\\"Id\\\",\\n      \\\"properties\\\": {\\n
\\\"dtype\\\": \\\"number\\\",\\n      \\\"std\\\": 88694,\\n      \\\"min\\\": 5,\\n
\\\"max\\\": 459488,\\n      \\\"num_unique_values\\\": 4803,\\n
\\\"samples\\\": [\\n        8427,\\n        13006,\\n        18041\\n

```

```

],\n      \"semantic_type\": \"\", \n      \"description\": \"\" \n}\n    },\n    {\n      \"column\": \"Keywords\", \n      \"properties\": {\n        \"dtype\": \"string\", \n        \"num_unique_values\": 4219, \n        \"samples\": [\n          \"revolution political activism loser\", \n          \"riddle captain treasure liquor treasure hunt\", \n          \"male nudity parents kids relationship camping cutting the cord self-discovery\" \n        ], \n        \"semantic_type\": \"\", \n        \"description\": \"\" \n      } \n    },\n    {\n      \"column\": \"Original language\", \n      \"properties\": {\n        \"dtype\": \"category\", \n        \"num_unique_values\": 37, \n        \"samples\": [\n          \"xx\", \n          \"ta\", \n          \"es\" \n        ], \n        \"semantic_type\": \"\", \n        \"description\": \"\" \n      } \n    },\n    {\n      \"column\": \"Popularity\", \n      \"properties\": {\n        \"dtype\": \"number\", \n        \"std\": 31.816649749537802, \n        \"min\": 0.0, \n        \"max\": 875.581305, \n        \"num_unique_values\": 4802, \n        \"samples\": [\n          13.267631, \n          0.010909, \n          5.842299 \n        ], \n        \"semantic_type\": \"\", \n        \"description\": \"\" \n      } \n    },\n    {\n      \"column\": \"Production companies\", \n      \"properties\": {\n        \"dtype\": \"string\", \n        \"num_unique_values\": 3697, \n        \"samples\": [\n          \"[{'name': 'Paramount Pictures', 'id': 4}, {'name': 'Cherry Alley Productions', 'id': 2232}]\", \n          \"[{'name': 'Twentieth Century Fox Film Corporation', 'id': 306}, {'name': 'Dune Entertainment', 'id': 444}, {'name': 'Regency Enterprises', 'id': 508}, {'name': 'Guy Walks into a Bar Productions', 'id': 2645}, {'name': 'Deep River Productions', 'id': 2646}, {'name': 'Friendly Films (II)', 'id': 81136}]\", \n          \"[{'name': 'Twentieth Century Fox Film Corporation', 'id': 306}]\" \n        ], \n        \"semantic_type\": \"\", \n        \"description\": \"\" \n      } \n    },\n    {\n      \"column\": \"Release date\", \n      \"properties\": {\n        \"dtype\": \"object\", \n        \"num_unique_values\": 3280, \n        \"samples\": [\n          \"1966-10-16\", \n          \"1987-07-31\", \n          \"1993-09-23\" \n        ], \n        \"semantic_type\": \"\", \n        \"description\": \"\" \n      } \n    },\n    {\n      \"column\": \"Revenue\", \n      \"properties\": {\n        \"dtype\": \"number\", \n        \"std\": 162857100, \n        \"min\": 0, \n        \"max\": 2787965087, \n        \"num_unique_values\": 3297, \n        \"samples\": [\n          11833696, \n          10462500, \n          17807569 \n        ], \n        \"semantic_type\": \"\", \n        \"description\": \"\" \n      } \n    },\n    {\n      \"column\": \"Runtime\", \n      \"properties\": {\n        \"dtype\": \"number\", \n        \"std\": 22.611934588844207, \n        \"min\": 0.0, \n        \"max\": 338.0, \n        \"num_unique_values\": 156, \n        \"samples\": [\n          74.0, \n

```

```

85.0,\n          170.0\n          ],\n          \"semantic_type\": \"\", \n\n\"description\": \"\" \n          },\n          {\n          \"column\": \n\n\"Spoken languages\", \n          \"properties\": {\n          \"dtype\": \n\n\"category\", \n          \"num_unique_values\": 544, \n\n\"samples\": [\n          \"[{\n\n\"iso_639_1\": \"es\", \n\n\"name\": \"Espa\n\nu00f1ol\", \n\n\"iso_639_1\": \n\n\"en\", \n\n\"name\": \"English\", \n\n\"iso_639_1\": \n\n\"fr\", \n\n\"name\": \"Fran\n\nu00e7ais\", \n\n\"iso_639_1\": \"hu\", \n\n\"name\": \"Magyar\"}]\", \n\n\"[{\n\n\"iso_639_1\": \"en\", \n\n\"name\": \"English\", \n\n\"iso_639_1\": \"it\", \n\n\"name\": \"Italiano\", \n\n\"iso_639_1\": \"pt\", \n\n\"name\": \"Portugu\n\nu00eas\"}]\", \n\n\"[{\n\n\"iso_639_1\": \"de\", \n\n\"name\": \"Deutsch\", \n\n\"iso_639_1\": \"it\", \n\n\"name\": \"Italiano\", \n\n\"iso_639_1\": \"la\", \n\n\"name\": \"Latin\", \n\n\"iso_639_1\": \"pl\", \n\n\"name\": \"Polski\"}]\", \n\n\"semantic_type\": \"\", \n\n\"description\": \"\" \n\n          },\n          {\n          \"column\": \"Status\", \n          \"properties\": \n\n{\n          \"dtype\": \"category\", \n          \"num_unique_values\": \n\n3, \n          \"samples\": [\n          \"Released\", \n          \"Post \n\nProduction\", \n          \"Rumored\" \n          ], \n          \"semantic_type\": \"\", \n          \"description\": \"\" \n\n          },\n          {\n          \"column\": \"Tagline\", \n          \"properties\": \n\n{\n          \"dtype\": \"string\", \n          \"num_unique_values\": \n\n3944, \n          \"samples\": [\n          \"When you're 17, every day \n\nis war.\", \n          \"An Unspeakable Horror. A Creative Genius. \n\nCaptured For Eternity.\", \n          \"May the schwartz be with you\" \n\n          ], \n          \"semantic_type\": \"\", \n          \"description\": \"\" \n\n          },\n          {\n          \"column\": \n\n\"Title\", \n          \"properties\": {\n          \"dtype\": \"string\", \n          \"num_unique_values\": 4800, \n          \"samples\": [\n          \"I \n\nSpy\", \n          \"Who's Your Caddy?\", \n          \"Sleepover\" \n          ], \n          \"semantic_type\": \"\", \n          \"description\": \"\" \n\n          },\n          {\n          \"column\": \"Vote average\", \n          \"properties\": {\n          \"dtype\": \"number\", \n          \"std\": \n\n1.1946121628478925, \n          \"min\": 0.0, \n          \"max\": 10.0, \n          \"num_unique_values\": 71, \n          \"samples\": [\n          5.1, \n          7.2, \n          4.0 \n          ], \n          \"semantic_type\": \"\", \n          \"description\": \"\" \n\n          },\n          {\n          \"column\": \n\n\"Vote count\", \n          \"properties\": {\n          \"dtype\": \n\n\"number\", \n          \"std\": 1234, \n          \"min\": 0, \n          \"max\": 13752, \n          \"num_unique_values\": 1609, \n          \"samples\": [\n          7604, \n          3428, \n          225 \n          ], \n          \"semantic_type\": \"\", \n          \"description\": \"\" \n\n          },\n          {\n          \"column\": \"Cast\", \n          \"properties\": \n\n{\n          \"dtype\": \"string\", \n          \"num_unique_values\": \n\n4741, \n          \"samples\": [\n          \"Al Gore Billy West Ronald \n\nReagan George W. Bush George H. W. Bush\", \n          \"David

```



```

Schwimmer Gwyneth Paltrow Michael Rapaport Toni Collette Carol
Kane\", \n          \"Paul Scofield Wendy Hiller Leo McKern Robert Shaw
Orson Welles\\\" \n          ], \n          \"semantic_type\\\": \"\", \n
\\\"description\\\": \"\" \n          } \n          { \n          \"column\\\":
\\\"Crew\\\", \n          \"properties\\\": { \n          \"dtype\\\": \"string\", \n
\\\"num_unique_values\\\": 4776, \n          \"samples\\\": [ \n
\\\"[{ 'name': 'Mark Goldblatt', 'gender': 2, 'department': 'Editing',
'job': 'Editor', 'credit_id': '538d82720e0a26670e005e83', 'id': 898},
{ 'name': 'Mark Helfrich', 'gender': 2, 'department': 'Editing', 'job':
'Editor', 'credit_id': '538d82860e0a266702005de5', 'id': 1098},
{ 'name': 'Ralph Winter', 'gender': 0, 'department': 'Production',
'job': 'Producer', 'credit_id': '538d820c0e0a266708006129', 'id':
2095}, { 'name': 'Richard Pagano', 'gender': 2, 'department':
'Production', 'job': 'Casting', 'credit_id':
'538d82d30e0a266712005def', 'id': 2624}, { 'name': 'Ed Verreaux',
'gender': 2, 'department': 'Art', 'job': 'Production Design',
'credit_id': '538d82ea0e0a266715005e74', 'id': 4953}, { 'name': 'Heike
Brandstätter', 'gender': 1, 'department': 'Production', 'job':
'Casting', 'credit_id': '538d82ad0e0a26670800614d', 'id': 5362},
{ 'name': 'Coreen Mayrs', 'gender': 1, 'department': 'Production',
'job': 'Casting', 'credit_id': '538d82c00e0a266705005fa2', 'id':
5363}, { 'name': 'Dan Moore', 'gender': 0, 'department': 'Costume &
Make-Up', 'job': 'Costume Supervisor', 'credit_id':
'538d845ec3a368713c0047d2', 'id': 6924}, { 'name': 'Lauren Shuler
Donner', 'gender': 1, 'department': 'Production', 'job': 'Producer',
'credit_id': '538d81ef0e0a266718005cb8', 'id': 7200}, { 'name': 'Stan
Lee', 'gender': 2, 'department': 'Production', 'job': 'Executive
Producer', 'credit_id': '576710bec3a36806720004b0', 'id': 7624},
{ 'name': 'Avi Arad', 'gender': 2, 'department': 'Production', 'job':
'Producer', 'credit_id': '538d81d80e0a266715005e4d', 'id': 7626},
{ 'name': 'Kevin Feige', 'gender': 2, 'department': 'Production',
'job': 'Executive Producer', 'credit_id': '5767101092514165c20000db',
'id': 10850}, { 'name': 'Judianna Makovsky', 'gender': 1, 'department':
'Costume & Make-Up', 'job': 'Costume Design', 'credit_id':
'538d83ba0e0a26670b005ded', 'id': 10970}, { 'name': 'Justin Scoppa
Jr.', 'gender': 2, 'department': 'Art', 'job': 'Art Direction',
'credit_id': '538d83620e0a266715005e82', 'id': 10549}, { 'name': 'John
Powell', 'gender': 2, 'department': 'Sound', 'job': 'Original Music
Composer', 'credit_id': '5626465bc3a3680e0e01704b', 'id': 11098},
{ 'name': 'Simon Kinberg', 'gender': 2, 'department': 'Writing', 'job':
'Screenplay', 'credit_id': '52fe45ff9251416c910459e1', 'id': 11092},
{ 'name': 'Bill Corso', 'gender': 2, 'department': 'Crew', 'job':
'Makeup Effects', 'credit_id': '56abe4ecc3a3681c47003aa3', 'id':
15017}, { 'name': 'Dante Spinotti', 'gender': 2, 'department':
'Camera', 'job': 'Director of Photography', 'credit_id':
'538d82560e0a266705005f94', 'id': 11099}, { 'name': 'Zak Penn',
'gender': 2, 'department': 'Writing', 'job': 'Screenplay',
'credit_id': '52fe45ff9251416c910459e7', 'id': 11011}, { 'name': 'Brett
Ratner', 'gender': 2, 'department': 'Directing', 'job': 'Director',

```



'credit\_id': '52fe45ff9251416c910459db', 'id': 11091}, {'name': 'Helen Jarvis', 'gender': 1, 'department': 'Art', 'job': 'Art Direction', 'credit\_id': '538d834f0e0a266712005dfe', 'id': 22061}, {'name': 'Elizabeth Wilcox', 'gender': 1, 'department': 'Art', 'job': 'Set Decoration', 'credit\_id': '538d83a50e0a266715005e8b', 'id': 11021}, {'name': 'Chad S. Frey', 'gender': 0, 'department': 'Art', 'job': 'Art Direction', 'credit\_id': '538d83220e0a266702005df6', 'id': 11102}, {'name': 'Geoff Hubbard', 'gender': 0, 'department': 'Art', 'job': 'Supervising Art Director', 'credit\_id': '570b8dde9251412c740020d0', 'id': 46082}, {'name': 'Lisa Tomczeszyn', 'gender': 2, 'department': 'Costume & Make-Up', 'job': 'Costume Design', 'credit\_id': '538d83ce0e0a266702005e02', 'id': 11106}, {'name': 'James Bamford', 'gender': 0, 'department': 'Crew', 'job': 'Stunts', 'credit\_id': '560046499251414e4d000197', 'id': 113194}, {'name': 'Julia Wong', 'gender': 1, 'department': 'Editing', 'job': 'Editor', 'credit\_id': '538d82990e0a266712005de9', 'id': 11100}, {'name': 'James M. Freitag', 'gender': 2, 'department': 'Production', 'job': 'Co-Producer', 'credit\_id': '57671042c3a36806b3000104', 'id': 11095}, {'name': 'Ross Fanger', 'gender': 2, 'department': 'Production', 'job': 'Co-Producer', 'credit\_id': '57670fdbcb3a3680722000106', 'id': 11014}, {'name': 'Sandi Tanaka', 'gender': 0, 'department': 'Art', 'job': 'Art Direction', 'credit\_id': '538d83750e0a26670b005de5', 'id': 223991}, {'name': 'John Palermo', 'gender': 0, 'department': 'Production', 'job': 'Executive Producer', 'credit\_id': '576710f192514133c5001213', 'id': 1182092}, {'name': 'John Bruno', 'gender': 0, 'department': 'Visual Effects', 'job': 'Visual Effects Supervisor', 'credit\_id': '55402d2fc3a3681da0000633', 'id': 1395269}, {'name': 'Guy Williams', 'gender': 0, 'department': 'Crew', 'job': 'Digital Effects Supervisor', 'credit\_id': '563a9ed79251416d520145a0', 'id': 1401803}, {'name': 'Kurt Williams', 'gender': 0, 'department': 'Production', 'job': 'Co-Producer', 'credit\_id': '5767112fc3a3680722000132', 'id': 1432021}, {'name': 'Rachel Wyn Dunn', 'gender': 0, 'department': 'Crew', 'job': 'Compositors', 'credit\_id': '55423048c3a36829e5003bc8', 'id': 1447557}, {'name': 'Joel Fletcher', 'gender': 0, 'department': 'Visual Effects', 'job': 'Animation', 'credit\_id': '552986b4c3a368596e001a79', 'id': 1452932}, {'name': 'Lee Cleary', 'gender': 0, 'department': 'Production', 'job': 'Co-Producer', 'credit\_id': '57670f92925141655a000109', 'id': 1477146}, {'name': 'David Gorder', 'gender': 0, 'department': 'Production', 'job': 'Associate Producer', 'credit\_id': '5767107c925141655a00012c', 'id': 1477147}, {'name': 'Luca Marco Paracels', 'gender': 2, 'department': 'Production', 'job': 'Associate Producer', 'credit\_id': '5767105bc3a36850c4000573', 'id': 1477203}], [{"name": "Jeanne McCarthy", "gender": 1, "department": "Production", "job": "Casting", "credit\_id": "58dd0e90c3a368223e00b59c", "id": 546}, {"name": "John Davis", "gender": 2, "department": "Production", "job": "Producer", "credit\_id": "52fe43bb9251416c7501bf83", "id": 2043}, {"name": "David Newman", "gender": 2, "department": "Sound", "job": "Original Music Composer", "credit\_id": "5539fe35c3a3685cf8002fle", "id": 11091}

```
'id': 3393}, {'name': '\\\\\"Aldric La'Auli Porter\\\\\", 'gender': 0,
'department': 'Production', 'job': 'Associate Producer', 'credit_id':
'58dd0d3fc3a368228a00aa50', 'id': 6187}, {'name': '\\\\\"Aldric La'Auli
Porter\\\\\", 'gender': 0, 'department': 'Directing', 'job': 'First
Assistant Director', 'credit_id': '5990427f925141672302ba00', 'id':
6187}, {'name': 'William Sandell', 'gender': 2, 'department': 'Art',
'job': 'Production Design', 'credit_id': '58dd0efdc3a368228300b438',
'id': 6877}, {'name': 'Robert Gould', 'gender': 2, 'department':
'Art', 'job': 'Set Decoration', 'credit_id':
'58dd0f56925141411000c340', 'id': 6880}, {'name': 'Brad Ricker',
'gender': 2, 'department': 'Art', 'job': 'Art Direction', 'credit_id':
'58dd0f1fc3a36822a600cldd', 'id': 8705}, {'name': 'Dan Yale',
'gender': 0, 'department': 'Sound', 'job': 'Foley Editor',
'credit_id': '5990490b925141673e02d2f5', 'id': 8763}, {'name': 'Neil
A. Machlis', 'gender': 2, 'department': 'Production', 'job':
'Executive Producer', 'credit_id': '58dd0d3192514140f600b915', 'id':
8880}, {'name': 'Juel Bestrop', 'gender': 1, 'department':
'Production', 'job': 'Casting', 'credit_id':
'58dd0e0bc3a368229300bdf6', 'id': 14377}, {'name': 'Hugo Santiago',
'gender': 0, 'department': 'Art', 'job': 'Set Designer', 'credit_id':
'599045b6925141672302bdac', 'id': 12639}, {'name': 'Ruth E. Carter',
'gender': 1, 'department': 'Costume & Make-Up', 'job': 'Costume
Design', 'credit_id': '58dd0f69c3a368228300b48f', 'id': 15524},
{'name': 'Craig Herring', 'gender': 2, 'department': 'Editing', 'job':
'Editor', 'credit_id': '52fe43bb9251416c7501bf9b', 'id': 18276},
{'name': 'Bruce Crone', 'gender': 2, 'department': 'Art', 'job':
'Supervising Art Director', 'credit_id': '58dd0f2f92514140fa00c653',
'id': 13434}, {'name': 'Douglas Smith', 'gender': 0, 'department':
'Visual Effects', 'job': 'Visual Effects Supervisor', 'credit_id':
'59904bcd3a368744202b9fb', 'id': 27100}, {'name': 'Larry Levin',
'gender': 2, 'department': 'Writing', 'job': 'Screenplay',
'credit_id': '52fe43bb9251416c7501bf7d', 'id': 31026}, {'name':
'Michele Imperato', 'gender': 1, 'department': 'Production', 'job':
'Co-Producer', 'credit_id': '58dd0c8592514140fa00c441', 'id': 49907},
{'name': 'Michele Imperato', 'gender': 1, 'department': 'Production',
'job': 'Unit Production Manager', 'credit_id':
'58dd0fb0c3a368227300bfad', 'id': 49907}, {'name': 'Steve Carr',
'gender': 2, 'department': 'Directing', 'job': 'Director',
'credit_id': '52fe43bb9251416c7501bfb1', 'id': 52112}, {'name': 'R.J.
Kizer', 'gender': 2, 'department': 'Sound', 'job': 'Dialogue Editor',
'credit_id': '5990471cc3a36874a302bbdc', 'id': 52193}, {'name': 'Daryn
Okada', 'gender': 2, 'department': 'Camera', 'job': 'Director of
Photography', 'credit_id': '52fe43bb9251416c7501bf95', 'id': 56324},
{'name': 'Joseph Singer', 'gender': 2, 'department': 'Production',
'job': 'Producer', 'credit_id': '52fe43bb9251416c7501bf89', 'id':
56053}, {'name': 'Peter S. Elliot', 'gender': 2, 'department': 'Crew',
'job': 'Visual Effects Editor', 'credit_id':
'59904a06925141676602b701', 'id': 57769}, {'name': 'Heidi Santelli',
'gender': 1, 'department': 'Production', 'job': 'Co-Producer',
```

'credit\_id': '58dd0d4c92514140fa00c4da', 'id': 66921}, {'name': 'Tom Villano', 'gender': 0, 'department': 'Sound', 'job': 'Music Editor', 'credit\_id': '59905155c3a36874b402a906', 'id': 91324}, {'name': 'Judy Murdock', 'gender': 0, 'department': 'Costume & Make-Up', 'job': 'Makeup Artist', 'credit\_id': '5990424a925141676602adf3', 'id': 168214}, {'name': 'Bruce McBroom', 'gender': 2, 'department': 'Camera', 'job': 'Still Photographer', 'credit\_id': '59904eae925141673e02da30', 'id': 555085}, {'name': 'Kris Evans', 'gender': 0, 'department': 'Costume & Make-Up', 'job': 'Makeup Department Head', 'credit\_id': '599041d0c3a36874a302b5e9', 'id': 1015965}, {'name': 'Iddo Lampton Enochs Jr.', 'gender': 0, 'department': 'Production', 'job': 'Location Manager', 'credit\_id': '58dd11a2c3a368227300c14b', 'id': 1113659}, {'name': 'Liz Radley', 'gender': 0, 'department': 'Crew', 'job': 'CG Supervisor', 'credit\_id': '59904b71925141675f02b7e7', 'id': 1178615}, {'name': 'Donna Marcione Pollack', 'gender': 0, 'department': 'Costume & Make-Up', 'job': 'Costume Supervisor', 'credit\_id': '59904fe5c3a368744202bf0c', 'id': 1319193}, {'name': 'Frank McEldowney', 'gender': 0, 'department': 'Art', 'job': 'Greensman', 'credit\_id': '599044c4925141672302bc97', 'id': 1364403}, {'name': 'Ted Caplan', 'gender': 0, 'department': 'Sound', 'job': 'Dialogue Editor', 'credit\_id': '59904694925141673e02cfd2', 'id': 1367505}, {'name': 'Gerald Lehtola', 'gender': 0, 'department': 'Art', 'job': 'Art Department Coordinator', 'credit\_id': '5990439d925141675902ac1b', 'id': 1377216}, {'name': 'Mark Wade', 'gender': 0, 'department': 'Art', 'job': 'Property Master', 'credit\_id': '599045fb925141675f02b16c', 'id': 1378165}, {'name': 'Patricia Klawonn', 'gender': 1, 'department': 'Art', 'job': 'Set Designer', 'credit\_id': '58dd0fd0c3a368228a00abfb', 'id': 1378749}, {'name': 'Bill Taliaferro', 'gender': 0, 'department': 'Art', 'job': 'Set Designer', 'credit\_id': '58dd1023c3a368227300c025', 'id': 1389585}, {'name': 'Dave Kulczycki', 'gender': 0, 'department': 'Sound', 'job': 'Sound Effects Editor', 'credit\_id': '59904785c3a3687493029126', 'id': 1390527}, {'name': 'Elizabeth Frank', 'gender': 0, 'department': 'Costume & Make-Up', 'job': 'Set Costumer', 'credit\_id': '59905010925141675f02bd38', 'id': 1399322}, {'name': 'Casey Hotchkiss', 'gender': 2, 'department': 'Camera', 'job': 'Camera Operator', 'credit\_id': '59904d81925141670b02cf99', 'id': 1400374}, {'name': 'Zack Davis', 'gender': 0, 'department': 'Sound', 'job': 'ADR Editor', 'credit\_id': '599046d2c3a3687493029074', 'id': 1404841}, {'name': 'Erma Kent', 'gender': 0, 'department': 'Costume & Make-Up', 'job': 'Hair Department Head', 'credit\_id': '59904214c3a368746102979f', 'id': 1408278}, {'name': 'Donald Sylvester', 'gender': 0, 'department': 'Sound', 'job': 'Supervising Sound Editor', 'credit\_id': '599048b2925141670b02c97f', 'id': 1416153}, {'name': 'John Kilkenny', 'gender': 0, 'department': 'Visual Effects', 'job': 'Visual Effects Producer', 'credit\_id': '59904ad6c3a36874b402a1cc', 'id': 1416167}, {'name': 'Camille Friend', 'gender': 1, 'department': 'Costume & Make-Up', 'job': 'Hairstylist',

'credit\_id': '56ae7d849251416357002915', 'id': 1424894}, {'name': 'Vanessa Lapato', 'gender': 0, 'department': 'Sound', 'job': 'ADR Editor', 'credit\_id': '599047ab925141674c02b48e', 'id': 1425343}, {'name': 'Alan E. Lorimer', 'gender': 0, 'department': 'Crew', 'job': 'Special Effects Coordinator', 'credit\_id': '59904951925141673902ae4e', 'id': 1442509}, {'name': 'Bryan McBrien', 'gender': 0, 'department': 'Art', 'job': 'Greensman', 'credit\_id': '59904427925141675902acc6', 'id': 1451398}, {'name': 'Butch West', 'gender': 0, 'department': 'Art', 'job': 'Construction Coordinator', 'credit\_id': '59904629c3a368744202b2d6', 'id': 1453276}, {'name': 'Hugh Lofting', 'gender': 2, 'department': 'Crew', 'job': 'Creator', 'credit\_id': '599034bcc3a3687493027d17', 'id': 1454647}, {'name': 'Nancy Karlin', 'gender': 1, 'department': 'Directing', 'job': 'Script Supervisor', 'credit\_id': '58dd11f1c3a36822ae00bc75', 'id': 1457729}, {'name': 'Mitchell Ferm', 'gender': 0, 'department': 'Visual Effects', 'job': 'Visual Effects Coordinator', 'credit\_id': '59904a8d925141675f02b6d8', 'id': 1496418}, {'name': 'Jim Harrison', 'gender': 2, 'department': 'Sound', 'job': 'Music Editor', 'credit\_id': '599050e6925141674c02beef', 'id': 1550058}, {'name': 'Jasan Sherman', 'gender': 0, 'department': 'Art', 'job': 'Location Scout', 'credit\_id': '58dd11b0c3a368223e00b7db', 'id': 1585159}, {'name': 'Thomas Meikle', 'gender': 0, 'department': 'Art', 'job': 'Sculptor', 'credit\_id': '5990453a925141673902a8ec', 'id': 1586888}, {'name': 'Spring Aspers', 'gender': 0, 'department': 'Sound', 'job': 'Music Supervisor', 'credit\_id': '5990506dc3a368747302c6e3', 'id': 1587383}, {'name': 'Craig Anthony', 'gender': 0, 'department': 'Costume & Make-Up', 'job': 'Key Costumer', 'credit\_id': '59904fbb925141672302c9ed', 'id': 1610228}, {'name': 'Harrison D. Marsh', 'gender': 2, 'department': 'Sound', 'job': 'Boom Operator', 'credit\_id': '599047e3925141674c02b4cf', 'id': 1737829}, {'name': 'Nadia Lubbe Simon', 'gender': 1, 'department': 'Production', 'job': 'Casting Assistant', 'credit\_id': '58dd10b2c3a368228300b588', 'id': 1763970}, {'name': 'Diana Jaher', 'gender': 1, 'department': 'Production', 'job': 'Casting Associate', 'credit\_id': '58dd1052925141413c00beac', 'id': 1786614}, {'name': 'Craig W. van Gundy', 'gender': 0, 'department': 'Production', 'job': 'Location Manager', 'credit\_id': '58dd11d5c3a368229d00b6a0', 'id': 1786616}, {'name': 'Kris Fullan Martinez', 'gender': 1, 'department': 'Production', 'job': 'Production Coordinator', 'credit\_id': '58dd1254c3a368227300cle6', 'id': 1786618}, {'name': 'Joe A. Hawthorne', 'gender': 0, 'department': 'Art', 'job': 'Painter', 'credit\_id': '599042f4c3a368744202aefa', 'id': 1868861}, {'name': 'Robert McDonald', 'gender': 0, 'department': 'Crew', 'job': 'Propmaker', 'credit\_id': '5990448e925141675f02afa7', 'id': 1868865}, {'name': 'Robert Prchal', 'gender': 0, 'department': 'Crew', 'job': 'Propmaker', 'credit\_id': '59904571925141670b02c572', 'id': 1868867}, {'name': 'Hal Schlax', 'gender': 0, 'department': 'Crew', 'job': 'Propmaker', 'credit\_id': '599045da925141673902a9d7', 'id': 1868868}, {'name': 'John Molo', 'gender': 2, 'department': 'Crew', 'job': 'Stunt

```
Coordinator', 'credit_id': '59904c34925141675f02b8bc', 'id': 1868877},
{'name': 'Todd Dos Reis', 'gender': 0, 'department': 'Camera', 'job':
'Camera Operator', 'credit_id': '59904cac925141673e02d7a8', 'id':
1868878}, {'name': 'Jason D. Hodges', 'gender': 0, 'department':
'Camera', 'job': 'Dolly Grip', 'credit_id':
'59904d04925141676602ba9a', 'id': 1868879}, {'name': 'Mike Hodges',
'gender': 0, 'department': 'Camera', 'job': 'Key Grip', 'credit_id':
'59904d3bc3a36874b402a4b0', 'id': 1868880}, {'name': 'Kal Manning',
'gender': 0, 'department': 'Lighting', 'job': 'Gaffer', 'credit_id':
'59904df0925141672302c7da', 'id': 1868881}, {'name': 'Larry McCarron',
'gender': 0, 'department': 'Camera', 'job': 'Dolly Grip', 'credit_id':
'59904ede925141675f02bbe6', 'id': 1868889}, {'name': 'Kim Strand',
'gender': 0, 'department': 'Sound', 'job': 'Music Editor',
'credit_id': '59905109925141673e02dcde', 'id': 1868892}}\n\n
\"[{ 'name': 'Harry Manfredini', 'gender': 2, 'department': 'Sound',
'job': 'Original Music Composer', 'credit_id':
'52fe44479251416c7502f0c3', 'id': 20953}, { 'name': 'Robin D. Cook',
'gender': 0, 'department': 'Production', 'job': 'Casting',
'credit_id': '534a8875c3a3682df60016a4', 'id': 23545}, { 'name': 'Sean
S. Cunningham', 'gender': 2, 'department': 'Production', 'job':
'Executive Producer', 'credit_id': '52fe44479251416c7502f0ed', 'id':
35475}, { 'name': 'Derick V. Underschultz', 'gender': 2, 'department':
'Camera', 'job': 'Director of Photography', 'credit_id':
'52fe44479251416c7502f0cf', 'id': 55642}, { 'name': 'Todd Farmer',
'gender': 2, 'department': 'Writing', 'job': 'Screenplay',
'credit_id': '52fe44479251416c7502f0b7', 'id': 61117}, { 'name': 'David
Handman', 'gender': 2, 'department': 'Editing', 'job': 'Editor',
'credit_id': '52fe44479251416c7502f0c9', 'id': 64748}, { 'name': 'James
Isaac', 'gender': 2, 'department': 'Directing', 'job': 'Director',
'credit_id': '52fe44479251416c7502f0b1', 'id': 69579}, { 'name': 'Noel
Cunningham', 'gender': 2, 'department': 'Production', 'job':
'Producer', 'credit_id': '52fe44479251416c7502f0bd', 'id': 69585},
{ 'name': 'Maxyne Baker', 'gender': 0, 'department': 'Costume & Make-
Up', 'job': 'Costume Design', 'credit_id': '534a888cc3a3682df60016a8',
'id': 958703}]\n\n      ],\n      \"semantic_type\": \"\", \n
\"description\": \"\", \n      },\n      {\n      \"column\":
\"Director\", \n      \"properties\": {\n      \"dtype\":
\"category\", \n      \"num_unique_values\": 2349, \n
\"samples\": [\n      \"Hart Bochner\", \n      \"David
Wain\", \n      \"Ham Tran\", \n      ], \n
\"semantic_type\": \"\", \n      \"description\": \"\" \n      } \n
      ] \n      }\", \"type\": \"dataframe\", \"variable_name\": \"df\"}
```

*#Now we need to deal with null value.*

```
df.isnull().sum()
```

Budget	0
Genres	28
Id	0



Keywords	412
Original language	0
Popularity	0
Production companies	0
Release date	1
Revenue	0
Runtime	2
Spoken languages	0
Status	0
Tagline	844
Title	0
Vote average	0
Vote count	0
Cast	43
Crew	0
Director	30

dtype: int64

```
df['Genres'].isnull().sum()
```

28

```
df.dropna(subset = ['Genres'],inplace=True)
```

```
df['Genres'].isnull().sum()
```

0

*#Here we will replace null value with maximum value from the column  
i.e independent film*

```
df['Keywords'].value_counts()
```

Keywords	
independent film	55
woman director	41
duringcreditsstinger	15
sport	13
independent film woman director	10
..	
broken neck fbi agent wall safe trashed house aerial shot	1
superhero kids and family	1
1970s skyscraper based on true story tightrope planning	1
american football mythology chinese food kangaroo	1
obsession camcorder crush dream girl	1

Name: count, Length: 4218, dtype: int64

```
df['Keywords'] = df['Keywords'].fillna(df['Keywords'].mode()[0])
```

```
df['Keywords'].isnull().sum()
```

0



```
df.dropna(subset = ['Release date'],inplace=True)
```

```
df.isnull().sum()
```

```
Budget          0
Genres          0
Id              0
Keywords        0
Original language 0
Popularity      0
Production companies 0
Release date    0
Revenue         0
Runtime         2
Spoken languages 0
Status          0
Tagline         819
Title           0
Vote average    0
Vote count      0
Cast           27
Crew            0
Director        16
dtype: int64
```

*#For tagline we cant actually use maximum value from the column to substitute with null so we will eliminate null value.*

```
df['Tagline'].isnull().sum()
```

```
819
```

```
df['Tagline'].value_counts()
```

```
Tagline
Based on a true story.
3
Who is John Galt?
2
There are two sides to every love story.
2
Be careful what you wish for.
2
What could go wrong?
2
..
The sort-of sequel to 'Knocked Up'
1
Life is Not Child-Proof.
1
```

```

Every war has a beginning.
1
First came love... then came Reverend Frank.
1
A newlywed couple's honeymoon is upended by the arrivals of their
respective sisters.      1
Name: count, Length: 3941, dtype: int64

```

```
df.dropna(subset = ['Tagline'], inplace=True)
```

```
df.isnull().sum()
```

```

Budget          0
Genres          0
Id              0
Keywords        0
Original language  0
Popularity      0
Production companies  0
Release date    0
Revenue         0
Runtime         0
Spoken languages  0
Status          0
Tagline         0
Title          0
Vote average    0
Vote count     0
Cast           8
Crew           0
Director       7
dtype: int64

```

```
df.shape
```

```
(3956, 19)
```

```
#Here we will directly eliminate column
```

```
df.dropna(subset = ['Cast', 'Director'], inplace = True)
```

```
df.isnull().sum()
```

```

Budget          0
Genres          0
Id              0
Keywords        0
Original language  0
Popularity      0
Production companies  0
Release date    0

```

Revenue	0
Runtime	0
Spoken languages	0
Status	0
Tagline	0
Title	0
Vote average	0
Vote count	0
Cast	0
Crew	0
Director	0

dtype: int64

df.dtypes

Budget	int64
Genres	object
Id	int64
Keywords	object
Original language	object
Popularity	float64
Production companies	object
Release date	object
Revenue	int64
Runtime	float64
Spoken languages	object
Status	object
Tagline	object
Title	object
Vote average	float64
Vote count	int64
Cast	object
Crew	object
Director	object

dtype: object

df.describe()

```
{"summary": "{\n  \"name\": \"df\",\n  \"rows\": 8,\n  \"fields\": [\n    {\n      \"column\": \"Budget\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 128358713.1618407,\n        \"min\": 0.0,\n        \"max\": 380000000.0,\n        \"num_unique_values\": 8,\n        \"samples\": [\n          33550053.66903193,\n          20000000.0,\n          3946.0\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Id\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 145830.39952210756,\n        \"min\": 5.0,\n        \"max\": 433715.0,\n        \"num_unique_values\": 8,\n        \"samples\": [\n          49695.15686771414,\n          12460.5,\n          3946.0\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    }\n  ]\n}
```

```

\"semantic_type\": \"\", \n      \"description\": \"\" \n    } \n  }, \n  { \n    \"column\": \"Popularity\", \n    \"properties\": { \n      \"dtype\": \"number\", \n      \"std\": 1378.1611669336598, \n      \"min\": 0.000372, \n      \"max\": 3946.0, \n      \"num_unique_values\": 8, \n      \"samples\": [ \n        24.771259384693362, \n        16.35253, \n        3946.0 \n      ], \n      \"semantic_type\": \"\", \n      \"description\": \"\" \n    } \n  }, \n  { \n    \"column\": \"Revenue\", \n    \"properties\": { \n      \"dtype\": \"number\", \n      \"std\": 966644446.4865726, \n      \"min\": 0.0, \n      \"max\": 2787965087.0, \n      \"num_unique_values\": 8, \n      \"samples\": [ \n        97217241.08616321, \n        32887555.0, \n        3946.0 \n      ], \n      \"semantic_type\": \"\", \n      \"description\": \"\" \n    } \n  }, \n  { \n    \"column\": \"Runtime\", \n    \"properties\": { \n      \"dtype\": \"number\", \n      \"std\": 1359.3967126771254, \n      \"min\": 0.0, \n      \"max\": 3946.0, \n      \"num_unique_values\": 8, \n      \"samples\": [ \n        108.12544348707551, \n        104.0, \n        3946.0 \n      ], \n      \"semantic_type\": \"\", \n      \"description\": \"\" \n    } \n  }, \n  { \n    \"column\": \"Vote average\", \n    \"properties\": { \n      \"dtype\": \"number\", \n      \"std\": 1393.316548274669, \n      \"min\": 0.0, \n      \"max\": 3946.0, \n      \"num_unique_values\": 8, \n      \"samples\": [ \n        6.187911809427268, \n        6.25, \n        3946.0 \n      ], \n      \"semantic_type\": \"\", \n      \"description\": \"\" \n    } \n  }, \n  { \n    \"column\": \"Vote count\", \n    \"properties\": { \n      \"dtype\": \"number\", \n      \"std\": 4660.128856311597, \n      \"min\": 0.0, \n      \"max\": 13752.0, \n      \"num_unique_values\": 8, \n      \"samples\": [ \n        810.7544348707552, \n        331.0, \n        3946.0 \n      ], \n      \"semantic_type\": \"\", \n      \"description\": \"\" \n    } \n  } \n] \n} \", \"type\": \"dataframe\"}

```

*#since data is cleaned we will move to outlier detection*

df

```

{ \"summary\": \" { \n  \"name\": \"df\", \n  \"rows\": 3946, \n  \"fields\": [ \n    { \n      \"column\": \"Budget\", \n      \"properties\": { \n        \"dtype\": \"number\", \n        \"std\": 42554177, \n        \"min\": 0, \n        \"max\": 380000000, \n        \"num_unique_values\": 393, \n        \"samples\": [ \n          87000000, \n          2200000, \n          85000 \n        ], \n        \"semantic_type\": \"\", \n        \"description\": \"\" \n      } \n    }, \n    { \n      \"column\": \"Genres\", \n      \"properties\": { \n        \"dtype\": \"category\", \n        \"num_unique_values\": 1061, \n        \"samples\": [ \n          \"Action Drama Horror Science Fiction Thriller\", \n          \"Action Adventure Thriller War\", \n          \"Adventure Drama Romance\" \n        ], \n        \"semantic_type\": \"\", \n        \"description\": \"\" \n      } \n    } \n  ] \n} \", \n  \"description\": \"\" \n} \n} \", \n  { \n    \"column\": \"Budget\", \n    \"properties\": { \n      \"dtype\": \"number\", \n      \"std\": 42554177, \n      \"min\": 0, \n      \"max\": 380000000, \n      \"num_unique_values\": 393, \n      \"samples\": [ \n        87000000, \n        2200000, \n        85000 \n      ], \n      \"semantic_type\": \"\", \n      \"description\": \"\" \n    } \n  }, \n  { \n    \"column\": \"Genres\", \n    \"properties\": { \n      \"dtype\": \"category\", \n      \"num_unique_values\": 1061, \n      \"samples\": [ \n        \"Action Drama Horror Science Fiction Thriller\", \n        \"Action Adventure Thriller War\", \n        \"Adventure Drama Romance\" \n      ], \n      \"semantic_type\": \"\", \n      \"description\": \"\" \n    } \n  } \n] \n} \", \"type\": \"dataframe\"}

```

```

\"column\": \"Id\", \n      \"properties\": { \n          \"dtype\":
\"number\", \n          \"std\": 80647, \n          \"min\": 5, \n
\"max\": 433715, \n          \"num_unique_values\": 3946, \n
\"samples\": [ \n              14369, \n              12279, \n              288980 \n
], \n          \"semantic_type\": \"\", \n          \"description\": \"\" \n
} \n      }, \n      { \n          \"column\": \"Keywords\", \n
\"properties\": { \n          \"dtype\": \"string\", \n
\"num_unique_values\": 3672, \n          \"samples\": [ \n              \"wolf
suspicion little red riding hood investigation burglary\", \n
\"cat photographer nightmare hallucination hypnosis\", \n
\"child abuse detective social worker supernatural murder\" \n
          ], \n          \"semantic_type\": \"\", \n
\"description\": \"\" \n      } \n      }, \n      { \n          \"column\":
\"Original language\", \n          \"properties\": { \n          \"dtype\":
\"category\", \n          \"num_unique_values\": 29, \n
\"samples\": [ \n              \"ps\", \n              \"pt\", \n
\"th\" \n          ], \n          \"semantic_type\": \"\", \n
\"description\": \"\" \n      } \n      }, \n      { \n          \"column\":
\"Popularity\", \n          \"properties\": { \n          \"dtype\":
\"number\", \n          \"std\": 33.895461448802436, \n          \"min\":
0.000372, \n          \"max\": 875.581305, \n
\"num_unique_values\": 3945, \n          \"samples\": [ \n
0.29245, \n              19.311572, \n              8.426589 \n          ], \n
\"semantic_type\": \"\", \n          \"description\": \"\" \n      } \n
      }, \n      { \n          \"column\": \"Production companies\", \n
\"properties\": { \n          \"dtype\": \"string\", \n
\"num_unique_values\": 3162, \n          \"samples\": [ \n
\"[{\\\"name\\\": \\\"Universal Pictures\\\", \\\"id\\\": 33},
\\\"name\\\": \\\"Twentieth Century Fox Film Corporation\\\",
\\\"id\\\": 306}, {\\\"name\\\": \\\"Samuel Goldwyn Films\\\",
\\\"id\\\": 9118}, {\\\"name\\\": \\\"Miramax\\\", \\\"id\\\":
53009}]\", \n          \"[{\\\"name\\\": \\\"DreamWorks SKG\\\",
\\\"id\\\": 27}, {\\\"name\\\": \\\"Gran Via Productions\\\",
\\\"id\\\": 2605}]\", \n          \"[{\\\"name\\\": \\\"Paramount
Pictures\\\", \\\"id\\\": 4}, {\\\"name\\\": \\\"Metro-Goldwyn-Mayer
(MGM)\\\", \\\"id\\\": 8411}, {\\\"name\\\": \\\"Sean Daniel
Company\\\", \\\"id\\\": 19857}, {\\\"name\\\": \\\"LightWorkers
Media\\\", \\\"id\\\": 23920}]\", \n          ], \n
\"semantic_type\": \"\", \n          \"description\": \"\" \n
      } \n      }, \n      { \n          \"column\": \"Release date\", \n
\"properties\": { \n          \"dtype\": \"object\", \n
\"num_unique_values\": 2877, \n          \"samples\": [ \n
\"2002-07-26\", \n          \"2008-08-29\", \n          \"2013-03-08\" \n
], \n          \"semantic_type\": \"\", \n          \"description\": \"\" \n
      } \n      }, \n      { \n          \"column\": \"Revenue\", \n
\"properties\": { \n          \"dtype\": \"number\", \n          \"std\":
174101155, \n          \"min\": 0, \n          \"max\": 2787965087, \n
\"num_unique_values\": 3029, \n          \"samples\": [ \n
96047633, \n              87713825, \n              25754284 \n          ], \n

```

```

{"semantic_type": "", "description": "", "column": "Runtime", "properties": {"dtype": "number", "std": 20.6262713198773, "min": 0.0, "max": 338.0, "num_unique_values": 141, "samples": [88.0, 188.0, 103.0], "semantic_type": ""}, {"column": "Spoken languages", "properties": {"dtype": "category", "num_unique_values": 469, "samples": [{"iso_639_1": "en", "name": "English"}, {"iso_639_1": "zh", "name": "\u666e\u901a\u8bdd"}, {"iso_639_1": "es", "name": "Espa\u00f1ol"}, {"iso_639_1": "en", "name": "English"}, {"iso_639_1": "ga", "name": "Gaeilge"}, {"iso_639_1": "la", "name": "Latin"}, {"iso_639_1": "zh", "name": "\u666e\u901a\u8bdd"}, {"iso_639_1": "de", "name": "Deutsch"}, {"iso_639_1": "en", "name": "English"}, {"iso_639_1": "it", "name": "Italiano"}, {"iso_639_1": "zh", "name": "\u666e\u901a\u8bdd"}, {"iso_639_1": "cs", "name": "\u010cesk\u00fd"}, {"iso_639_1": "cn", "name": "\u5e7f\u5dde\u8bdd / \u5ee3\u5dde\u8a71"}, {"iso_639_1": "en", "name": "English"}, {"iso_639_1": "it", "name": "Italiano"}, {"iso_639_1": "zh", "name": "\u666e\u901a\u8bdd"}, {"iso_639_1": "cs", "name": "\u010cesk\u00fd"}, {"iso_639_1": "cn", "name": "\u5e7f\u5dde\u8bdd / \u5ee3\u5dde\u8a71"}], "semantic_type": ""}, {"column": "Status", "properties": {"dtype": "category", "num_unique_values": 3, "samples": ["Released", "Post Production", "Rumored"], "semantic_type": ""}, {"column": "Tagline", "properties": {"dtype": "string", "num_unique_values": 3931, "samples": ["She's chasing a dream... they're chasing her.", "The call of the wild, the thrill of adventure. The mistake of a lifetime.", "When you follow your dream, there's no telling what you'll become."], "semantic_type": ""}, {"column": "Title", "properties": {"dtype": "string", "num_unique_values": 3943, "samples": ["Slam", "Half Baked", "In the Valley of Elah"], "semantic_type": ""}, {"column": "Vote average", "properties": {"dtype": "number", "std": 0.9756234746750709, "min": 0.0, "max": 10.0, "num_unique_values": 65, "samples": [2.9, 2.7, 7.2], "semantic_type": ""}}

```



```

"description\": \"\\"\n    }\n  },\n  {\n    \"column\":
\"Vote count\", \n    \"properties\": {\n      \"dtype\":
\"number\", \n      \"std\": 1317, \n      \"min\": 0, \n
\"max\": 13752, \n      \"num_unique_values\": 1587, \n
\"samples\": [\n        1429, \n        3220, \n        1477 \n
], \n      \"semantic_type\": \"\", \n      \"description\": \"\\"\n
}\n  }, \n  {\n    \"column\": \"Cast\", \n    \"properties\":
{\n      \"dtype\": \"string\", \n      \"num_unique_values\":
3930, \n      \"samples\": [\n        \"Sarah Jessica Parker Pierce
Brosnan Sarah Shahi Busy Philipps Olivia Munn\", \n        \"Peter
Sallis Helena Bonham Carter Ralph Fiennes Nicholas Smith Liz Smith\", \n
        \"Allan Corduner Jim Broadbent Timothy Spall Dexter
Fletcher Shirley Henderson\" \n      ], \n      \"semantic_type\":
\"\", \n      \"description\": \"\\"\n    }\n  }, \n  {\n    \"column\": \"Crew\", \n    \"properties\": {\n      \"dtype\":
\"string\", \n      \"num_unique_values\": 3946, \n
\"samples\": [\n      \"[{'name': 'Brendan Malloy', 'gender': 2,
'department': 'Directing', 'job': 'Director', 'credit_id':
'52fe45eb9251416c75066e71', 'id': 82221}, {'name': 'Emmett Malloy',
'gender': 2, 'department': 'Directing', 'job': 'Director',
'credit_id': '52fe45eb9251416c75066e77', 'id': 82222}, {'name': 'Jon
Zack', 'gender': 2, 'department': 'Writing', 'job': 'Writer',
'credit_id': '52fe45eb9251416c75066e7d', 'id': 85781}, {'name': 'J.J.
Makaro', 'gender': 2, 'department': 'Crew', 'job': 'Stunt
Coordinator', 'credit_id': '5643599b9251410a49001f90', 'id': 1007395},
{'name': 'Cara Fay', 'gender': 1, 'department': 'Crew', 'job': 'Post
Production Supervisor', 'credit_id': '5899d00d92514113c1004fe0', 'id':
1752658}]]\", \n      \"[{'name': 'Robert Rodriguez', 'gender': 0,
'department': 'Writing', 'job': 'Screenplay', 'credit_id':
'52fe44d89251416c75042f83', 'id': 2294}, {'name': 'Robert Rodriguez',
'gender': 0, 'department': 'Directing', 'job': 'Director',
'credit_id': '52fe44d89251416c75042f6b', 'id': 2294}, {'name': 'Robert
Rodriguez', 'gender': 0, 'department': 'Camera', 'job': 'Director of
Photography', 'credit_id': '52fe44d89251416c75042f77', 'id': 2294},
{'name': 'Robert Rodriguez', 'gender': 0, 'department': 'Editing',
'job': 'Editor', 'credit_id': '52fe44d89251416c75042f89', 'id': 2294},
{'name': 'Robert Rodriguez', 'gender': 0, 'department': 'Production',
'job': 'Producer', 'credit_id': '52fe44d89251416c75042f71', 'id':
2294}, {'name': 'Robert Rodriguez', 'gender': 0, 'department':
'Sound', 'job': 'Original Music Composer', 'credit_id':
'52fe44d89251416c75042f7d', 'id': 2294}, {'name': 'Bob Weinstein',
'gender': 2, 'department': 'Production', 'job': 'Executive Producer',
'credit_id': '52fe44d89251416c75042f59', 'id': 1307}, {'name': 'Harvey
Weinstein', 'gender': 2, 'department': 'Production', 'job': 'Executive
Producer', 'credit_id': '52fe44d89251416c75042f5f', 'id': 59839},
{'name': 'Elizabeth Avellan', 'gender': 1, 'department': 'Production',
'job': 'Producer', 'credit_id': '52fe44d89251416c75042f53', 'id':
5911}, {'name': 'Rebecca Rodriguez', 'gender': 0, 'department':
'Sound', 'job': 'Original Music Composer', 'credit_id':

```

```
'52fe44d89251416c75042f65', 'id': 72030}]]\",\\n          \\\"[{ 'name':
'Timothy A. Chey', 'gender': 2, 'department': 'Writing', 'job':
'Writer', 'credit_id': '577a83c1c3a368770a000b94', 'id': 238822},
{'name': 'Peter Cousins', 'gender': 2, 'department': 'Directing',
'job': 'Director', 'credit_id': '53fd6e8a0e0a262de8000f2a', 'id':
1219769}]]\\n          ],\\n          \\\"semantic_type\\\": \\\"\\\",\\n
\\\"description\\\": \\\"\\\"\\n          }\\n          },\\n          {\\n          \\\"column\\\":
\\\"Director\\\",\\n          \\\"properties\\\": {\\n          \\\"dtype\\\":
\\\"category\\\",\\n          \\\"num_unique_values\\\": 1868,\\n
\\\"samples\\\": [\\n          \\\"Shimit Amin\\\",\\n          \\\"Henry
Hathaway\\\",\\n          \\\"Bart Freundlich\\\"\\n          ],\\n
\\\"semantic_type\\\": \\\"\\\",\\n          \\\"description\\\": \\\"\\\"\\n          }\\
n          }\\n          ]\\n}\\\", \"type\": \"dataframe\", \"variable_name\": \"df\"}
```

```
df.head()
```

```
{ \"summary\": \"{\\n  \\\"name\\\": \\\"df\\\",\\n  \\\"rows\\\": 3946,\\n  \\\"fields\\\":
[\\n    {\\n      \\\"column\\\": \\\"Budget\\\",\\n      \\\"properties\\\": {\\n
\\\"dtype\\\": \\\"number\\\",\\n      \\\"std\\\": 42554177,\\n      \\\"min\\\":
0,\\n      \\\"max\\\": 380000000,\\n      \\\"num_unique_values\\\": 393,\\n
\\\"samples\\\": [\\n      87000000,\\n      2200000,\\n
85000\\n      ],\\n      \\\"semantic_type\\\": \\\"\\\",\\n
\\\"description\\\": \\\"\\\"\\n      }\\n      },\\n      {\\n      \\\"column\\\":
\\\"Genres\\\",\\n      \\\"properties\\\": {\\n      \\\"dtype\\\":
\\\"category\\\",\\n      \\\"num_unique_values\\\": 1061,\\n
\\\"samples\\\": [\\n      \\\"Action Drama Horror Science Fiction
Thriller\\\",\\n      \\\"Action Adventure Thriller War\\\",\\n
\\\"Adventure Drama Romance\\\"\\n      ],\\n      \\\"semantic_type\\\":
\\\"\\\",\\n      \\\"description\\\": \\\"\\\"\\n      }\\n      },\\n      {\\n
\\\"column\\\": \\\"Id\\\",\\n      \\\"properties\\\": {\\n      \\\"dtype\\\":
\\\"number\\\",\\n      \\\"std\\\": 80647,\\n      \\\"min\\\": 5,\\n
\\\"max\\\": 433715,\\n      \\\"num_unique_values\\\": 3946,\\n
\\\"samples\\\": [\\n      14369,\\n      12279,\\n      288980\\n
],\\n      \\\"semantic_type\\\": \\\"\\\",\\n      \\\"description\\\": \\\"\\\"\\n
}\\n      },\\n      {\\n      \\\"column\\\": \\\"Keywords\\\",\\n
\\\"properties\\\": {\\n      \\\"dtype\\\": \\\"string\\\",\\n
\\\"num_unique_values\\\": 3672,\\n      \\\"samples\\\": [\\n      \\\"wolf
suspicion little red riding hood investigation burglary\\\",\\n
\\\"cat photographer nightmare hallucination hypnosis\\\",\\n
\\\"child abuse detective social worker supernatural murder\\\"\\
n      ],\\n      \\\"semantic_type\\\": \\\"\\\",\\n
\\\"description\\\": \\\"\\\"\\n      }\\n      },\\n      {\\n      \\\"column\\\":
\\\"Original language\\\",\\n      \\\"properties\\\": {\\n      \\\"dtype\\\":
\\\"category\\\",\\n      \\\"num_unique_values\\\": 29,\\n
\\\"samples\\\": [\\n      \\\"ps\\\",\\n      \\\"pt\\\",\\n
\\\"th\\\"\\n      ],\\n      \\\"semantic_type\\\": \\\"\\\",\\n
\\\"description\\\": \\\"\\\"\\n      }\\n      },\\n      {\\n      \\\"column\\\":
\\\"Popularity\\\",\\n      \\\"properties\\\": {\\n      \\\"dtype\\\":
\\\"number\\\",\\n      \\\"std\\\": 33.895461448802436,\\n      \\\"min\\\":
0.000372,\\n      \\\"max\\\": 875.581305,\\n
```

```

\"num_unique_values\": 3945,\n          \"samples\": [\n
0.29245,\n          19.311572,\n          8.426589\n          ],\n
\"semantic_type\": \"\", \n          \"description\": \"\"\n
n      },\n      {\n          \"column\": \"Production companies\", \n
\"properties\": {\n          \"dtype\": \"string\", \n
\"num_unique_values\": 3162,\n          \"samples\": [\n
\"[{\\\"name\\\": \\\"Universal Pictures\\\", \\\"id\\\": 33},\n
{\\\"name\\\": \\\"Twentieth Century Fox Film Corporation\\\", \n
\\\"id\\\": 306}, {\\\"name\\\": \\\"Samuel Goldwyn Films\\\", \n
\\\"id\\\": 9118}, {\\\"name\\\": \\\"Miramax\\\", \\\"id\\\": \n
53009}]\", \n          \"[{\\\"name\\\": \\\"DreamWorks SKG\\\", \n
\\\"id\\\": 27}, {\\\"name\\\": \\\"Gran Via Productions\\\", \n
\\\"id\\\": 2605}]\", \n          \"[{\\\"name\\\": \\\"Paramount \n
Pictures\\\", \\\"id\\\": 4}, {\\\"name\\\": \\\"Metro-Goldwyn-Mayer \n
(MGM)\\\", \\\"id\\\": 8411}, {\\\"name\\\": \\\"Sean Daniel \n
Company\\\", \\\"id\\\": 19857}, {\\\"name\\\": \\\"LightWorkers \n
Media\\\", \\\"id\\\": 23920}]\" \n          ],\n
\"semantic_type\": \"\", \n          \"description\": \"\" \n
n      },\n      {\n          \"column\": \"Release date\", \n
\"properties\": {\n          \"dtype\": \"object\", \n
\"num_unique_values\": 2877,\n          \"samples\": [\n
\"2002-07-26\", \n          \"2008-08-29\", \n          \"2013-03-08\" \n
],\n          \"semantic_type\": \"\", \n          \"description\": \"\" \n
n      },\n      {\n          \"column\": \"Revenue\", \n
\"properties\": {\n          \"dtype\": \"number\", \n          \"std\": \n
174101155,\n          \"min\": 0,\n          \"max\": 2787965087,\n
\"num_unique_values\": 3029,\n          \"samples\": [\n
96047633,\n          87713825,\n          25754284 \n          ],\n
\"semantic_type\": \"\", \n          \"description\": \"\" \n
n      },\n      {\n          \"column\": \"Runtime\", \n          \"properties\": \n
{\n          \"dtype\": \"number\", \n          \"std\": 20.6262713198773,\n
          \"min\": 0.0,\n          \"max\": 338.0,\n
\"num_unique_values\": 141,\n          \"samples\": [\n
188.0,\n          103.0 \n          ],\n          \"semantic_type\": \"\", \n
          \"description\": \"\" \n          },\n      {\n
\"column\": \"Spoken languages\", \n          \"properties\": {\n
\"dtype\": \"category\", \n          \"num_unique_values\": 469,\n
\"samples\": [\n          \"[{\\\"iso_639_1\\\": \\\"en\\\", \n
\\\"name\\\": \\\"English\\\", {\\\"iso_639_1\\\": \\\"zh\\\", \n
\\\"name\\\": \\\"\\\"u666e\\\"u901a\\\"u8bdd\\\", \n
{\\\"iso_639_1\\\": \\\"es\\\", \\\"name\\\": \\\"Espa\\\" \n
u00f1ol\\\"}]\", \n          \"[{\\\"iso_639_1\\\": \\\"en\\\", \n
\\\"name\\\": \\\"English\\\", {\\\"iso_639_1\\\": \\\"ga\\\", \n
\\\"name\\\": \\\"Gaeilge\\\", {\\\"iso_639_1\\\": \\\"la\\\", \n
\\\"name\\\": \\\"Latin\\\", {\\\"iso_639_1\\\": \\\"zh\\\", \n
\\\"name\\\": \\\"\\\"u666e\\\"u901a\\\"u8bdd\\\"}]\", \n
\"[{\\\"iso_639_1\\\": \\\"de\\\", \\\"name\\\": \\\"Deutsch\\\", \n
{\\\"iso_639_1\\\": \\\"en\\\", \\\"name\\\": \\\"English\\\", \n
{\\\"iso_639_1\\\": \\\"it\\\", \\\"name\\\": \\\"Italiano\\\",

```

```

{\\\\"iso_639_1\\": \\\"zh\\\", \\\"name\\\": \\\"\\\"\\\"u666e\\\"\\\"
u901a\\\"\\\"u8bdd\\\"}, {\\\\"iso_639_1\\": \\\"cs\\\", \\\"name\\\":
\\\"\\\"\\\"u010cesk\\\"\\\"u00fd\\\"}, {\\\\"iso_639_1\\": \\\"cn\\\",
\\\"name\\\": \\\"\\\"\\\"u5e7f\\\"\\\"u5dde\\\"\\\"u8bdd / \\\"\\\"u5ee3\\\"\\\"u5dde\\\"\\\"
u8a71\\\"}\\\"\\n
\\\"semantic_type\\\": \\\"\\\",\\n
\\\"description\\\": \\\"\\\"\\n
\\\"}\\n
\\\"}\\n
\\\"column\\\":
\\\"Status\\\",\\n
\\\"properties\\\": {\\n
\\\"dtype\\\":
\\\"category\\\",\\n
\\\"num_unique_values\\\": 3,\\n
\\\"samples\\\":
[\\n
\\\"Released\\\",\\n
\\\"Post Production\\\",\\n
\\\"Rumored\\\"\\n
\\\"}\\n
\\\"semantic_type\\\": \\\"\\\",\\n
\\\"description\\\": \\\"\\\"\\n
\\\"}\\n
\\\"}\\n
\\\"column\\\":
\\\"Tagline\\\",\\n
\\\"properties\\\": {\\n
\\\"dtype\\\": \\\"string\\\",\\n
\\\"num_unique_values\\\": 3931,\\n
\\\"samples\\\": [\\n
\\\"She's chasing a dream... they're chasing her.\\\",\\n
\\\"The
call of the wild, the thrill of adventure. The mistake of a
lifetime.\\\",\\n
\\\"When you follow your dream, there's no
telling what you'll become.\\\"\\n
\\\"}\\n
\\\"semantic_type\\\":
\\\"\\\",\\n
\\\"description\\\": \\\"\\\"\\n
\\\"}\\n
\\\"}\\n
\\\"column\\\": \\\"Title\\\",\\n
\\\"properties\\\": {\\n
\\\"dtype\\\":
\\\"string\\\",\\n
\\\"num_unique_values\\\": 3943,\\n
\\\"samples\\\": [\\n
\\\"Slam\\\",\\n
\\\"Half Baked\\\",\\n
\\\"In the Valley of Elah\\\"\\n
\\\"}\\n
\\\"semantic_type\\\":
\\\"\\\",\\n
\\\"description\\\": \\\"\\\"\\n
\\\"}\\n
\\\"}\\n
\\\"column\\\": \\\"Vote average\\\",\\n
\\\"properties\\\": {\\n
\\\"dtype\\\": \\\"number\\\",\\n
\\\"std\\\": 0.9756234746750709,\\n
\\\"min\\\": 0.0,\\n
\\\"max\\\": 10.0,\\n
\\\"num_unique_values\\\":
65,\\n
\\\"samples\\\": [\\n
2.9,\\n
2.7,\\n
7.2\\n
\\\"}\\n
\\\"semantic_type\\\": \\\"\\\",\\n
\\\"description\\\": \\\"\\\"\\n
\\\"}\\n
\\\"}\\n
\\\"column\\\":
\\\"Vote count\\\",\\n
\\\"properties\\\": {\\n
\\\"dtype\\\":
\\\"number\\\",\\n
\\\"std\\\": 1317,\\n
\\\"min\\\": 0,\\n
\\\"max\\\": 13752,\\n
\\\"num_unique_values\\\": 1587,\\n
\\\"samples\\\": [\\n
1429,\\n
3220,\\n
1477\\n
\\\"}\\n
\\\"semantic_type\\\": \\\"\\\",\\n
\\\"description\\\": \\\"\\\"\\n
\\\"}\\n
\\\"}\\n
\\\"column\\\": \\\"Cast\\\",\\n
\\\"properties\\\":
{\\n
\\\"dtype\\\": \\\"string\\\",\\n
\\\"num_unique_values\\\":
3930,\\n
\\\"samples\\\": [\\n
\\\"Sarah Jessica Parker Pierce
Brosnan Sarah Shahi Busy Philipps Olivia Munn\\\",\\n
\\\"Peter
Sallis Helena Bonham Carter Ralph Fiennes Nicholas Smith Liz Smith\\\",\\n
\\\"
\\\"Allan Corduner Jim Broadbent Timothy Spall Dexter
Fletcher Shirley Henderson\\\"\\n
\\\"}\\n
\\\"semantic_type\\\":
\\\"\\\",\\n
\\\"description\\\": \\\"\\\"\\n
\\\"}\\n
\\\"}\\n
\\\"column\\\": \\\"Crew\\\",\\n
\\\"properties\\\": {\\n
\\\"dtype\\\":
\\\"string\\\",\\n
\\\"num_unique_values\\\": 3946,\\n
\\\"samples\\\": [\\n
\\\"[{ 'name': 'Brendan Malloy', 'gender': 2,
'department': 'Directing', 'job': 'Director', 'credit_id':
'52fe45eb9251416c75066e71', 'id': 82221}, { 'name': 'Emmett Malloy',
'gender': 2, 'department': 'Directing', 'job': 'Director',
'credit_id': '52fe45eb9251416c75066e77', 'id': 82222}, { 'name': 'Jon

```

```

Zack', 'gender': 2, 'department': 'Writing', 'job': 'Writer',
'credit_id': '52fe45eb9251416c75066e7d', 'id': 85781}, {'name': 'J.J.
Makaro', 'gender': 2, 'department': 'Crew', 'job': 'Stunt
Coordinator', 'credit_id': '5643599b9251410a49001f90', 'id': 1007395},
{'name': 'Cara Fay', 'gender': 1, 'department': 'Crew', 'job': 'Post
Production Supervisor', 'credit_id': '5899d00d92514113c1004fe0', 'id':
1752658}]]", "\n" [{"name": 'Robert Rodriguez', 'gender': 0,
'department': 'Writing', 'job': 'Screenplay', 'credit_id':
'52fe44d89251416c75042f83', 'id': 2294}, {'name': 'Robert Rodriguez',
'gender': 0, 'department': 'Directing', 'job': 'Director',
'credit_id': '52fe44d89251416c75042f6b', 'id': 2294}, {'name': 'Robert
Rodriguez', 'gender': 0, 'department': 'Camera', 'job': 'Director of
Photography', 'credit_id': '52fe44d89251416c75042f77', 'id': 2294},
{'name': 'Robert Rodriguez', 'gender': 0, 'department': 'Editing',
'job': 'Editor', 'credit_id': '52fe44d89251416c75042f89', 'id': 2294},
{'name': 'Robert Rodriguez', 'gender': 0, 'department': 'Production',
'job': 'Producer', 'credit_id': '52fe44d89251416c75042f71', 'id':
2294}, {'name': 'Robert Rodriguez', 'gender': 0, 'department':
'Sound', 'job': 'Original Music Composer', 'credit_id':
'52fe44d89251416c75042f7d', 'id': 2294}, {'name': 'Bob Weinstein',
'gender': 2, 'department': 'Production', 'job': 'Executive Producer',
'credit_id': '52fe44d89251416c75042f59', 'id': 1307}, {'name': 'Harvey
Weinstein', 'gender': 2, 'department': 'Production', 'job': 'Executive
Producer', 'credit_id': '52fe44d89251416c75042f5f', 'id': 59839},
{'name': 'Elizabeth Avellan', 'gender': 1, 'department': 'Production',
'job': 'Producer', 'credit_id': '52fe44d89251416c75042f53', 'id':
5911}, {'name': 'Rebecca Rodriguez', 'gender': 0, 'department':
'Sound', 'job': 'Original Music Composer', 'credit_id':
'52fe44d89251416c75042f65', 'id': 72030}]]", "\n" [{"name":
'Timothy A. Chey', 'gender': 2, 'department': 'Writing', 'job':
'Writer', 'credit_id': '577a83c1c3a368770a000b94', 'id': 238822},
{'name': 'Peter Cousins', 'gender': 2, 'department': 'Directing',
'job': 'Director', 'credit_id': '53fd6e8a0e0a262de8000f2a', 'id':
1219769}]]", "\n" ], "\n" "semantic_type": "\n", "\n
"description": "\n", "\n" }, "\n" {"column":
"Director", "\n" "properties": {"dtype":
"category", "\n" "num_unique_values": 1868, "\n
"samples": [\n "Shimit Amin", "\n "Henry
Hathaway", "\n "Bart Freundlich", "\n ], "\n
"semantic_type": "\n", "\n" "description": "\n", "\n" } \n
} \n ] \n }", "type": "dataframe", "variable_name": "df"}

```

```
df.describe()
```

```

{"summary": "{\n "name": "\n df", "\n "rows": 8, "\n "fields": [\n
{\n "column": "\n Budget", "\n "properties": {\n
"dtype": "\n number", "\n "std": 128358713.1618407, "\n
"min": 0.0, "\n "max": 380000000.0, "\n
"num_unique_values": 8, "\n "samples": [\n
33550053.66903193, "\n 20000000.0, "\n 3946.0 \n

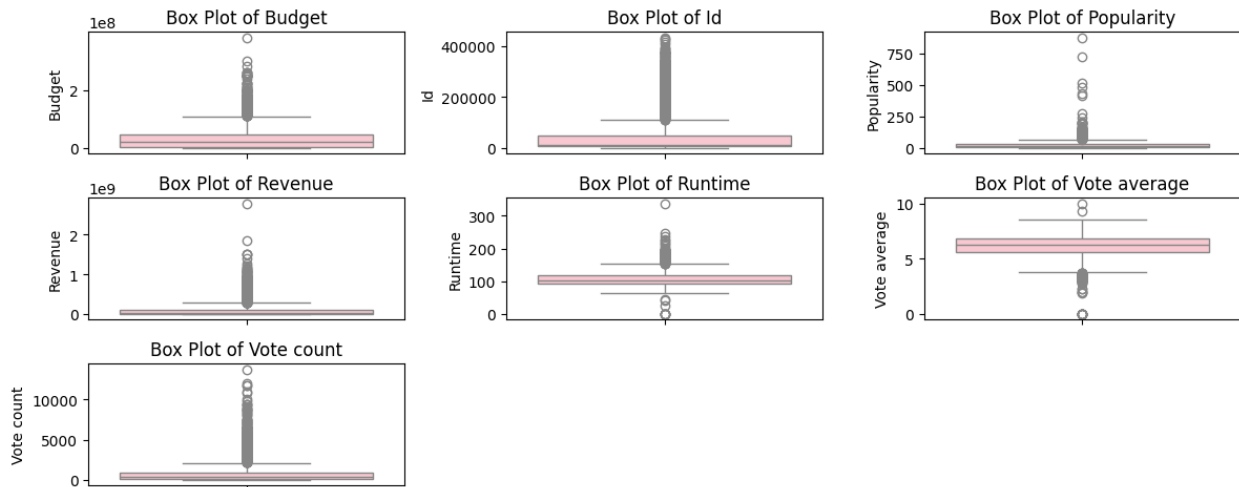
```





```
plt.subplot(num_rows, num_cols, i) # Update subplot grid
sns.boxplot(y=df[column], color='pink')
plt.title(f'Box Plot of {column}')
```

```
plt.tight_layout() # Adjust subplot parameters for a tight layout
plt.show()
```



'''After taking look at graph  
the distribution is highly right-skewed, meaning most movies have  
relatively low budgets, but a few movies have exceptionally high  
budgets  
since it is heavily right-skewed so we use MAD median absolute  
deviation to deal with outliers.'''

```
plt.figure(figsize=(10,5))
sns.distplot(df[['Budget']])
plt.show()
```

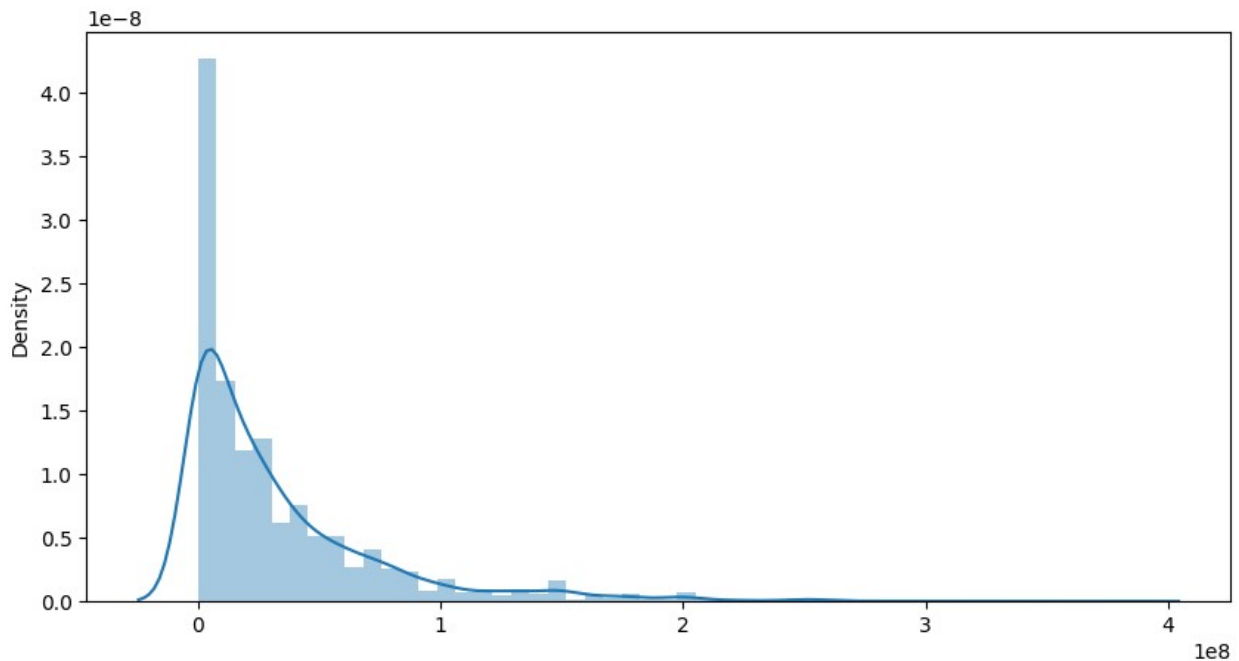
<ipython-input-253-8f9b51d1048d>:6: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see <https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751>

```
sns.distplot(df[['Budget']])
```



```

median = df[['Budget']].median()
median
Budget    20000000.0
dtype: float64

mad = np.abs(df[['Budget']] - median).median()
mad
Budget    18636100.0
dtype: float64

threshold = 3

outlier = (np.abs(df[['Budget']] - median) / mad >
threshold).any(axis=1)
outlier
0      True
1      True
2      True
3      True
4      True
...
4793   False
4795   False
4796   False
4798   False

```

```

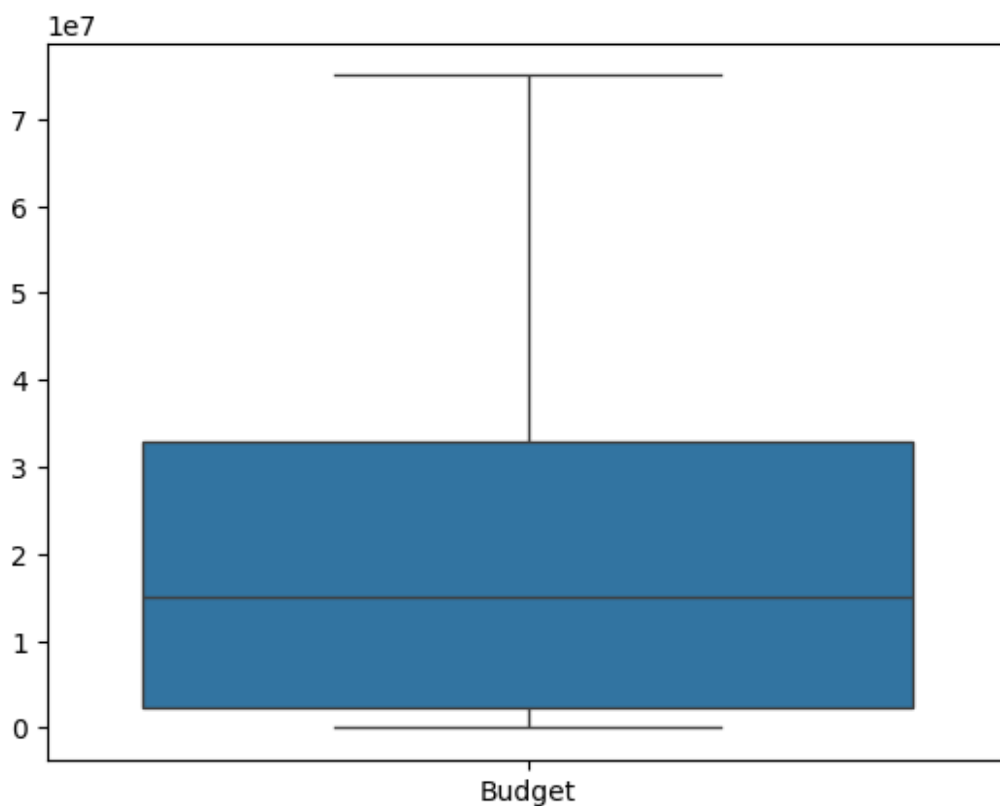
4799      False
Length: 3946, dtype: bool

df = df[~outlier]

df
{"repr_error": "0", "type": "dataframe", "variable_name": "df"}
#From plot we can see that budget column is free from outliers

sns.boxplot(df[['Budget']])
plt.show()

```



*#Since graph suggest that Id is right skewed so we will use same MAD to deal with outlier*

```

sns.distplot(df['Id'])
plt.show()

```

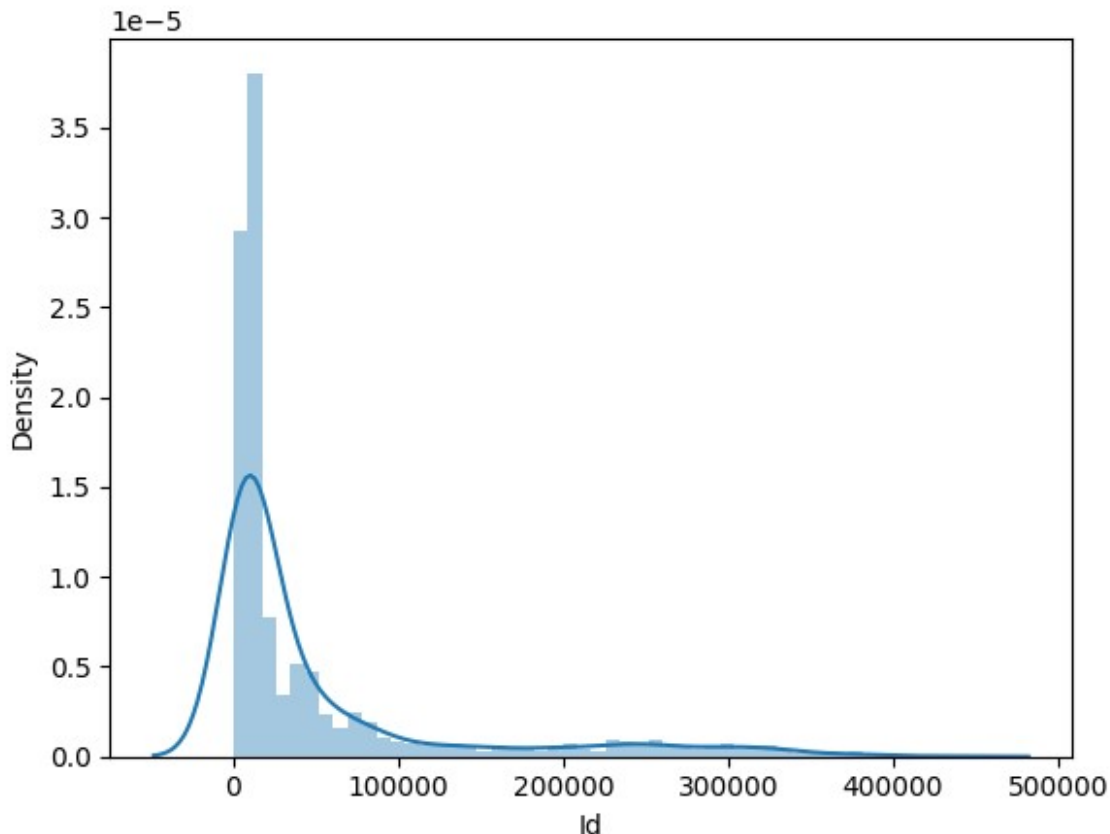
<ipython-input-262-ee06f63d9f98>:3: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see <https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751>

```
sns.distplot(df['Id'])
```



```
median1 = df[['Id']].median()

mad1 = np.abs(df[['Id']] - median1).median()
mad1

Id      10882.0
dtype: float64

threshold = 3
outliers_id = (np.abs(df[['Id']] - median1) / mad1 >
threshold).any(axis=1)

outliers_id
```

```
83      True
97      True
99     False
151     False
180     False
```

```
...
4793    True
4795    True
4796    False
4798    False
4799    True
```

```
Length: 3463, dtype: bool
```

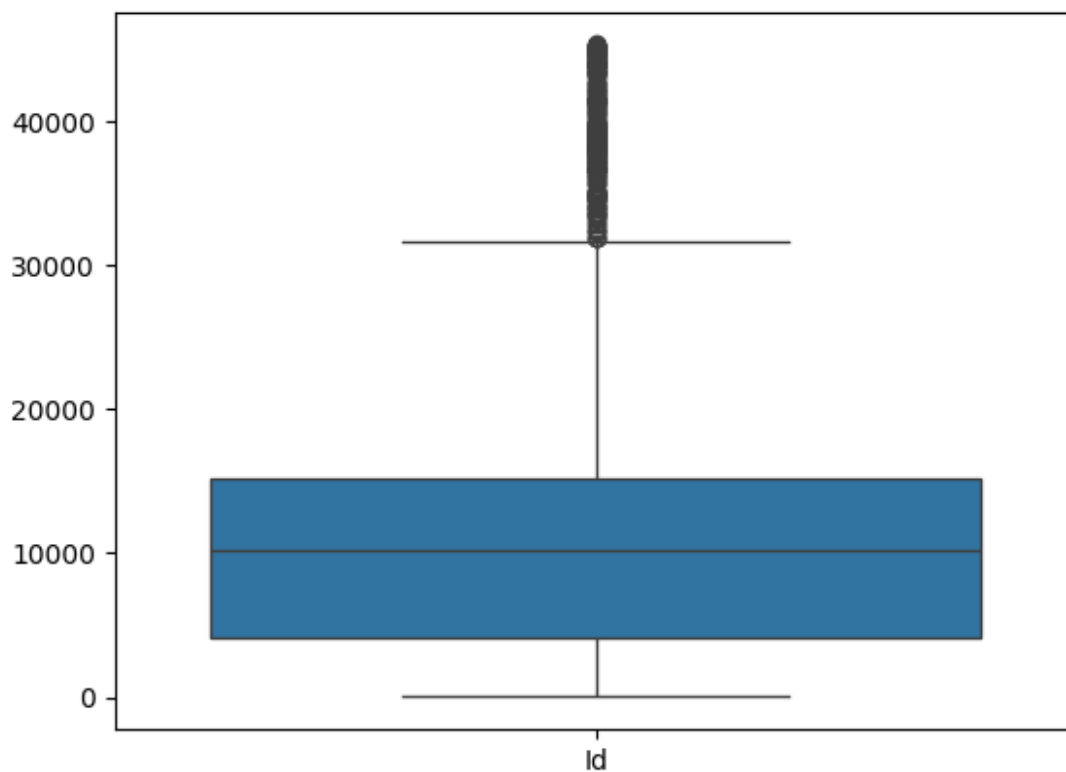
```
df_id = df[~outliers_id]
```

```
df_id
```

```
{"repr_error": "0", "type": "dataframe", "variable_name": "df_id"}
```

*#Here we have reduced the effect of outlier but we can try more methods which are less effective because graph is highly skewed to right.*

```
sns.boxplot(df_id[['Id']])
plt.show()
```



```
#Here similar to previous column we will use distplot to check skewness and then try to figure out how to deal with outliers  
df_id['Popularity']
```

```
99      6.909942  
151     35.601665  
180     45.381501  
207     43.129703  
265     18.251129
```

```
...  
4788    4.553644  
4789    1.464566  
4792    0.212443  
4796    23.307949  
4798    14.269792
```

```
Name: Popularity, Length: 2559, dtype: float64
```

```
# After observing the plot again the graph is highly skewed towards right we will use MAD method to deal with outliers  
#The density gradually decreases as the popularity score increases, but a few movies still have very high popularity.
```

```
sns.distplot(df_id[['Popularity']])  
plt.show()
```

```
<ipython-input-271-222f93e81df3>:4: UserWarning:
```

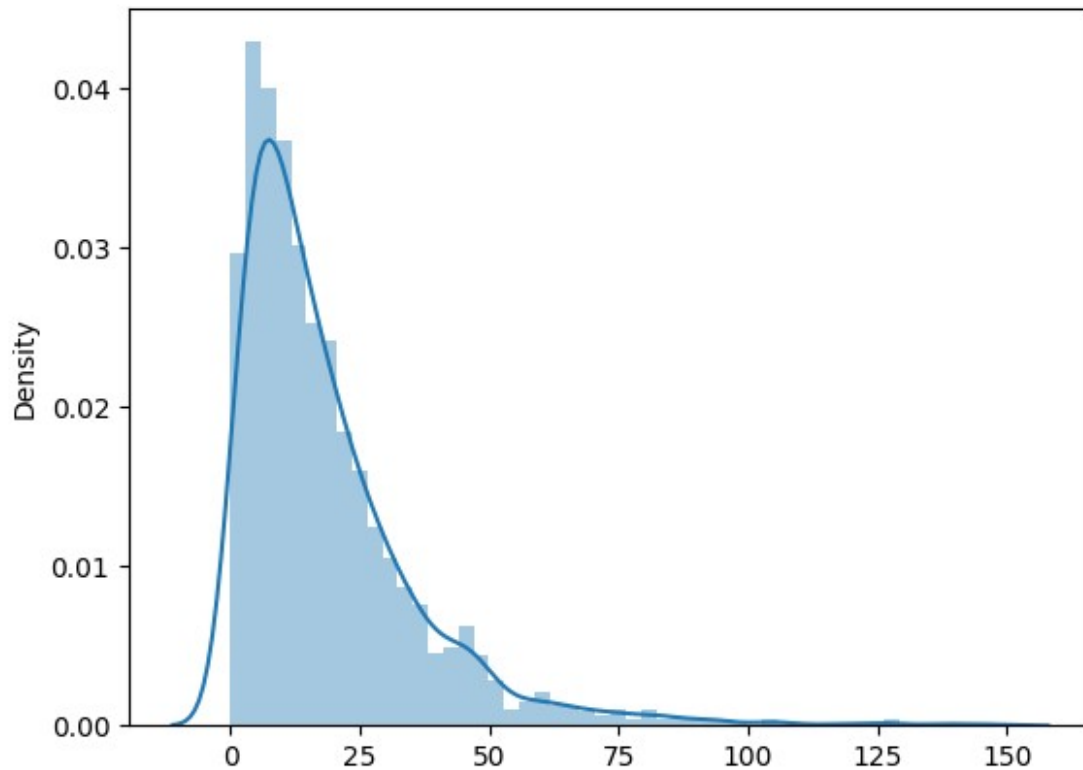
```
`distplot` is a deprecated function and will be removed in seaborn v0.14.0.
```

```
Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
```

```
For a guide to updating your code to use the new functions, please see https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751
```

```
sns.distplot(df_id[['Popularity']])
```





```
median2 = df_id[['Popularity']].median()
median2
```

```
Popularity    13.786075
dtype: float64
```

```
mad2 = np.abs(df_id[['Popularity']] - median2).median()
mad2
```

```
Popularity    8.288231
dtype: float64
```

```
outliers_p = (np.abs(df_id[['Popularity']] - median2) / mad2 >
threshold).any(axis=1)
outliers_p
```

```
99      False
151     False
180      True
207      True
265     False
...
4788    False
4789    False
4792    False
4796    False
```

```

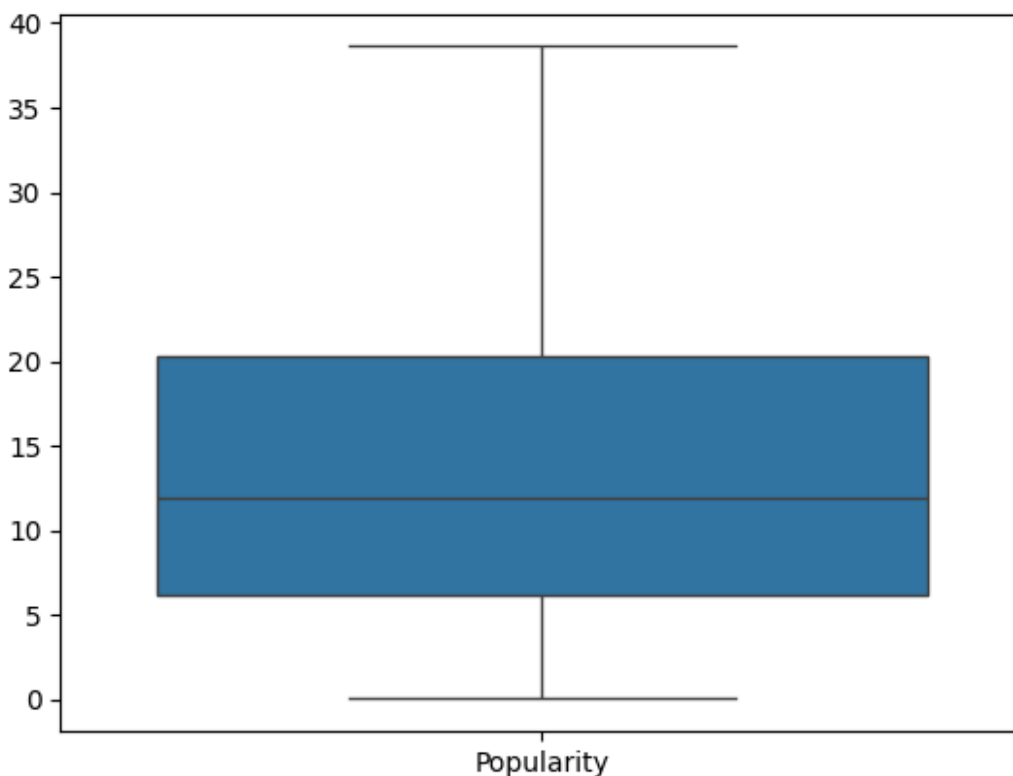
4798     False
Length: 2559, dtype: bool

df_p = df_id[~outliers_p]
df_p
{"repr_error": "0", "type": "dataframe", "variable_name": "df_p"}

#So from boxplot we conclude that popularity graph is also free from outliers

sns.boxplot(df_p[['Popularity']])
plt.show()

```



*'''From this plot we can say that maximum number of movies have high revenue but here density decreases gradually but similar to other two column graph is right skewed'''*

```

sns.distplot(df_p['Revenue'])
plt.show()

```

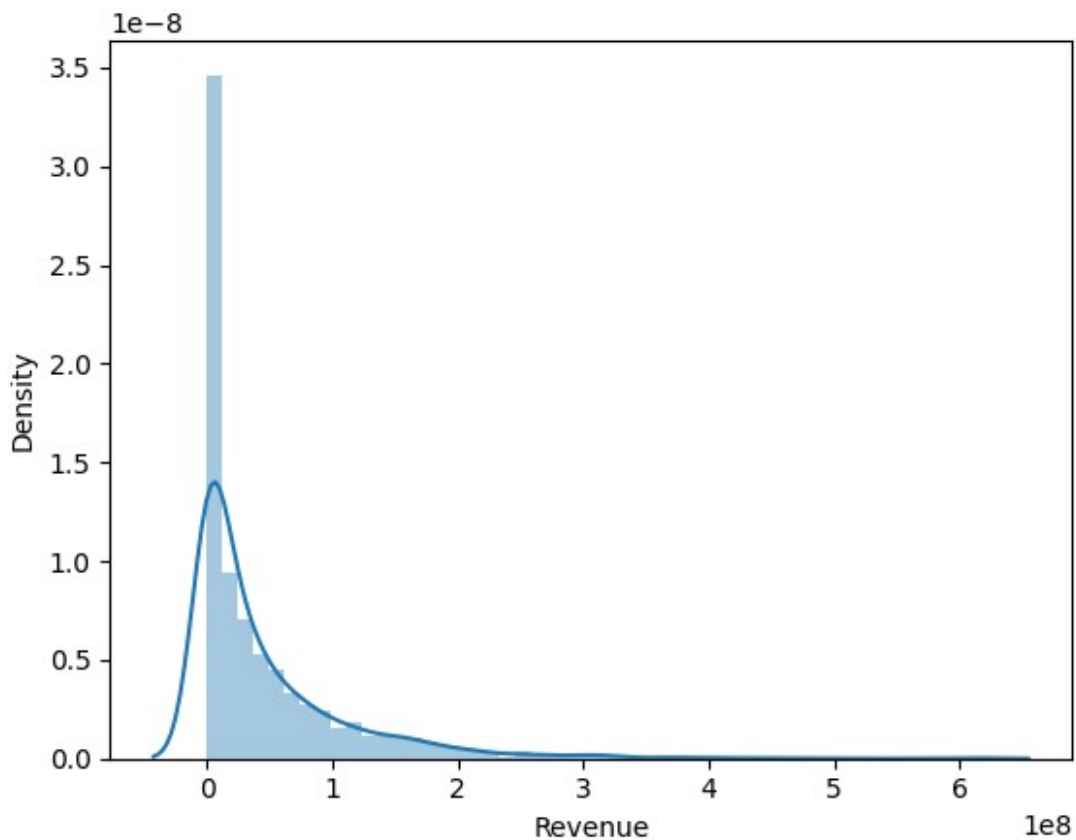
<ipython-input-278-f3ff9544b3df>:4: UserWarning:

``distplot` is a deprecated function and will be removed in seaborn v0.14.0.`

Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see <https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751>

```
sns.distplot(df_p['Revenue'])
```



```
median3 = df_p[['Revenue']].median()  
median3
```

```
Revenue    19870567.0  
dtype: float64
```

```
mad3 = np.abs(df_p[['Revenue']] - median3).median()  
mad3
```

```
Revenue    19870567.0  
dtype: float64
```

```

outliers_r = (np.abs(df_p[['Revenue']] - median3) / mad3 >
threshold).any(axis=1)
outliers_r

99      True
151     True
265    False
314    False
327     True
...
4788   False
4789   False
4792   False
4796   False
4798   False
Length: 2275, dtype: bool

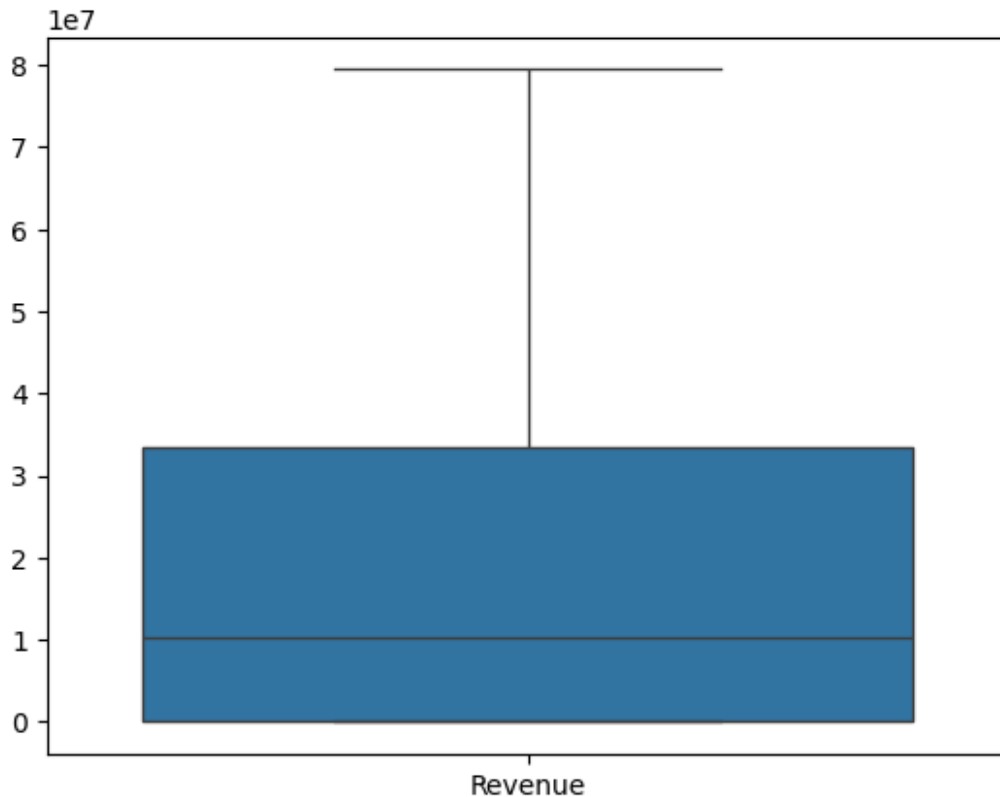
df_r = df_p[~outliers_r]
df_r

{"repr_error": "0", "type": "dataframe", "variable_name": "df_r"}

#Here revenue is free from outliers

sns.boxplot(df_r[['Revenue']])
plt.show()

```



*''' From this graph we can observe that the graph is not perfectly symmetric but also it is slightly skewed so if we take closer look at graph it is more skewed toward right but is less skewed towards left but here we cant say it is right skewed so here we can use any method but i have prefered to use MAD method to deal with outliers'''*

```
sns.distplot(df_r['Runtime'])  
plt.show()
```

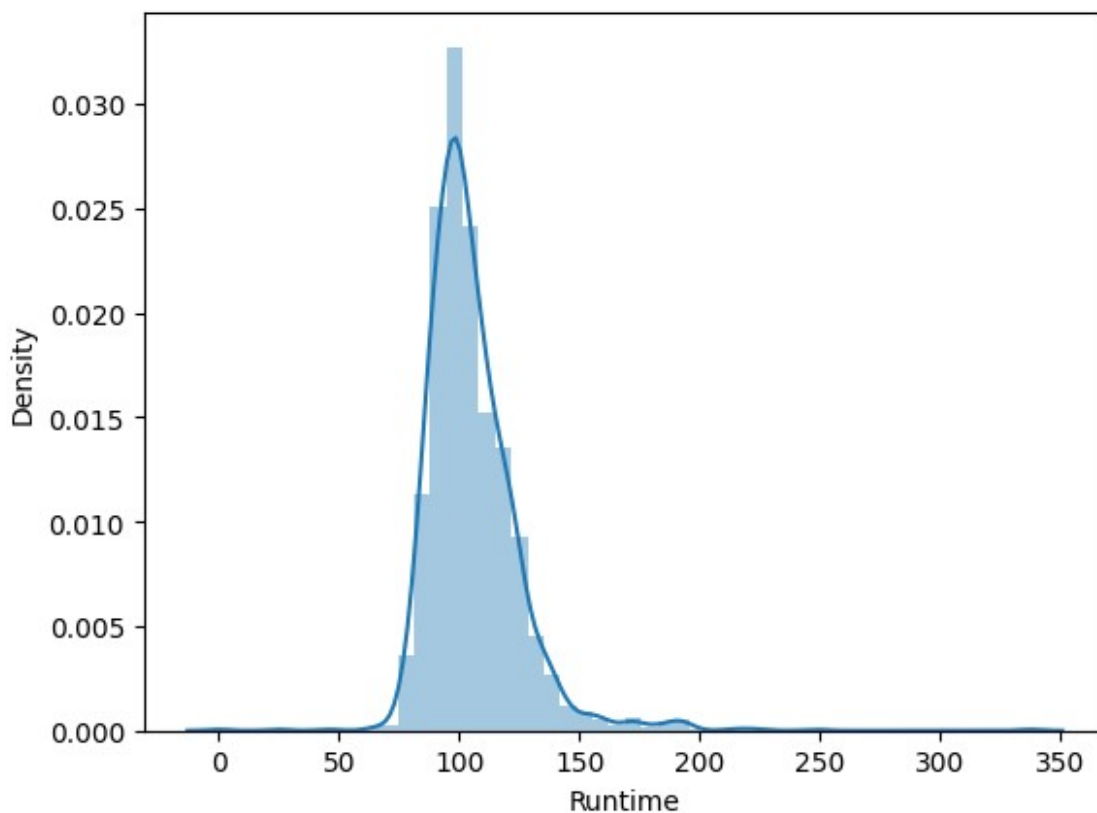
<ipython-input-284-3be68de0e709>:5: UserWarning:

``distplot`` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either ``displot`` (a figure-level function with similar flexibility) or ``histplot`` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see <https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751>

```
sns.distplot(df_r['Runtime'])
```



```
median4 = df_r[['Runtime']].median()
median4

Runtime      102.0
dtype: float64

mad4 = np.abs(df_r[['Runtime']] - median4).median()
mad4

Runtime      10.0
dtype: float64

outliers_run = (np.abs(df_r[['Runtime']] - median4) / mad4 >
threshold).any(axis=1)
outliers_run

265      False
314      False
379      False
409      False
412       True
...
4788     False
4789     False
4792     False
```



```

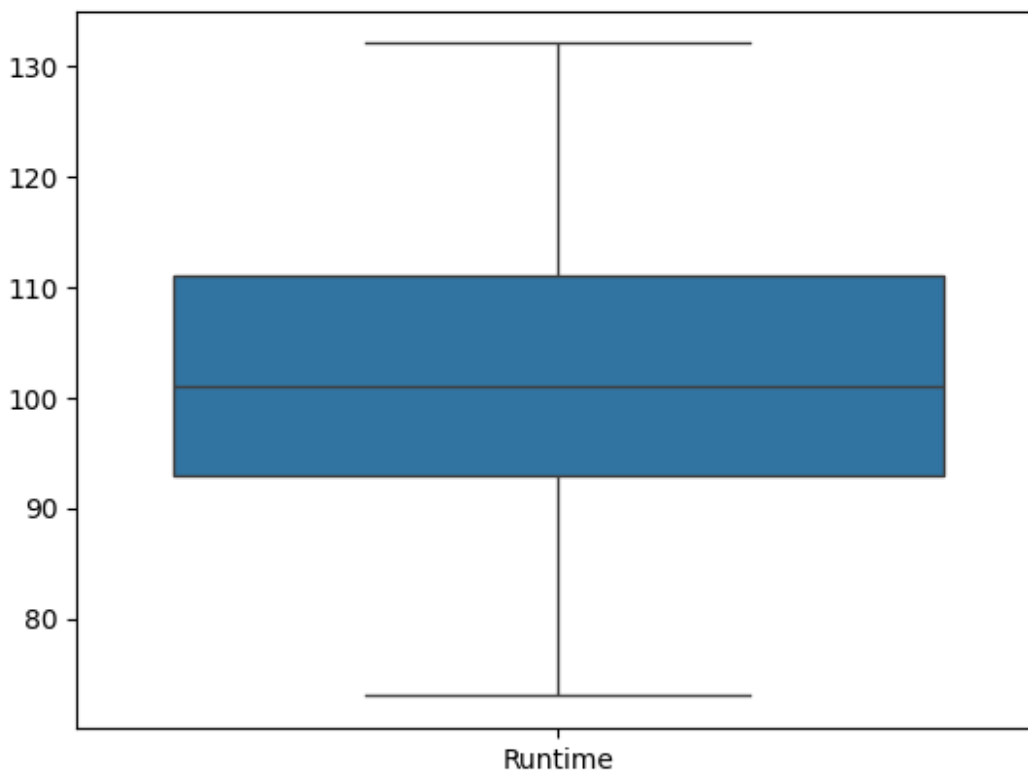
4796     False
4798     False
Length: 1826, dtype: bool

df_run = df_r[~outliers_run]
df_run

{"repr_error": "0", "type": "dataframe", "variable_name": "df_run"}

#The column is free from outliers
sns.boxplot(df_run[['Runtime']])
plt.show()

```



*#From this graph the conclusion is that it is right skewed but the extending towards right also data is more concentrated towards left*  
*#Here er will use zscore method to detect and eliminate outliers*

```

sns.distplot(df_run['Vote average'])
plt.show()

```

<ipython-input-290-93054cfb6060>:4: UserWarning:

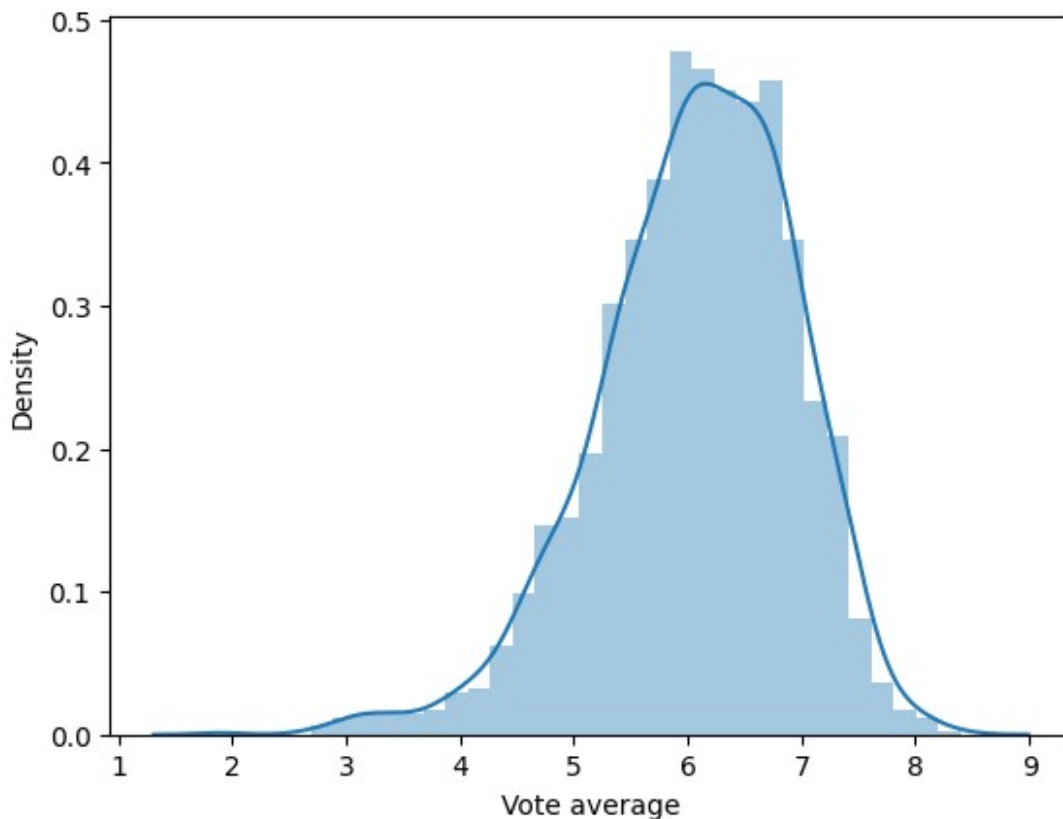
`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level

function with similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see <https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751>

```
sns.distplot(df_run['Vote average'])
```



```
z_scores = np.abs(stats.zscore(df_run[['Vote average']]))
z_scores

{"summary": "{\n  \"name\": \"z_scores\",\n  \"rows\": 1698,\n  \"fields\": [\n    {\n      \"column\": \"Vote average\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 0.6187066826313937,\n        \"min\": 0.031049039221791475,\n        \"max\": 4.772190182342995,\n        \"num_unique_values\": 57,\n        \"samples\": [\n          1.3413050240824322,\n          0.5407651538216355,\n          2.2562077329519163\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    }\n  ],\n  \"type\": \"dataframe\",\n  \"variable_name\": \"z_scores\"}
```

```
outliers_avg = (z_scores > threshold).any(axis=1)
outliers_avg
```

```

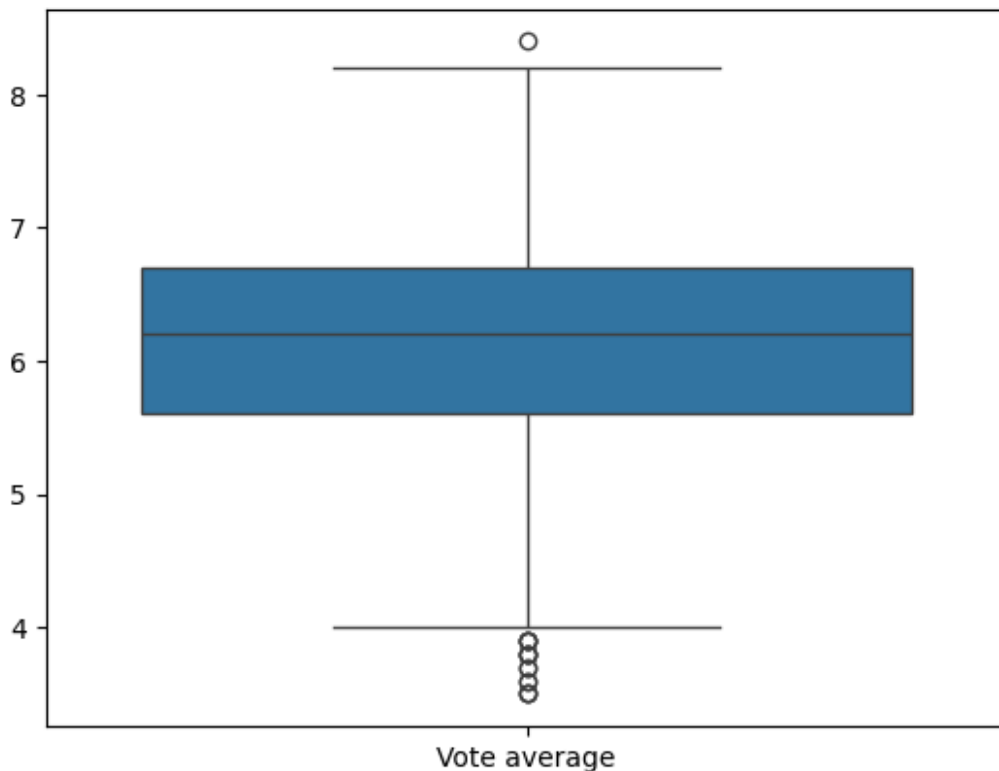
265     False
314     False
379     False
409     False
417     False
...
4788    False
4789    False
4792    False
4796    False
4798    False
Length: 1698, dtype: bool

df_avg = df_run[~outliers_avg]
df_avg

{"repr_error": "0", "type": "dataframe", "variable_name": "df_avg"}

#Here we have reduced the effect of the outlier to a certain amount
sns.boxplot(df_avg[['Vote average']])
plt.show()

```



*#Here the graph suggest that data is more concentrated towards right and also tail is extending towards right*  
*#so similar to other it is highly right skewed so we will use MAD*

*method to detect and deal with outlier*

```
sns.distplot(df_avg['Vote count'])  
plt.show()
```

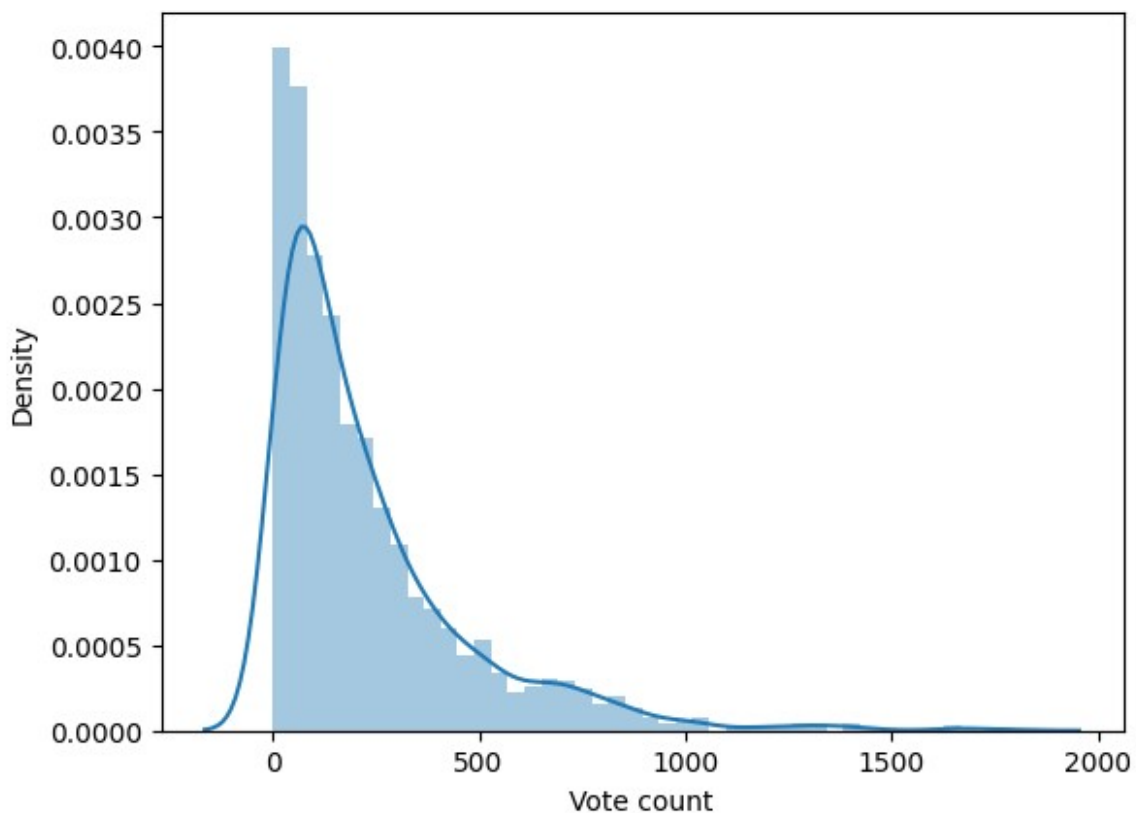
<ipython-input-295-c681c6171998>:4: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see <https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751>

```
sns.distplot(df_avg['Vote count'])
```



```
median5 = df_avg[['Vote count']].median()  
median5
```

```

Vote count    151.0
dtype: float64

mad5 = np.abs(df_avg[['Vote count']] - median5).median()
mad5

Vote count    104.0
dtype: float64

outliers = (np.abs(df_avg[['Vote count']] - median5) / mad5 >
threshold).any(axis=1)
outliers

265      False
314      False
379       True
409      False
417      False
...
4788     False
4789     False
4792     False
4796      True
4798     False
Length: 1681, dtype: bool

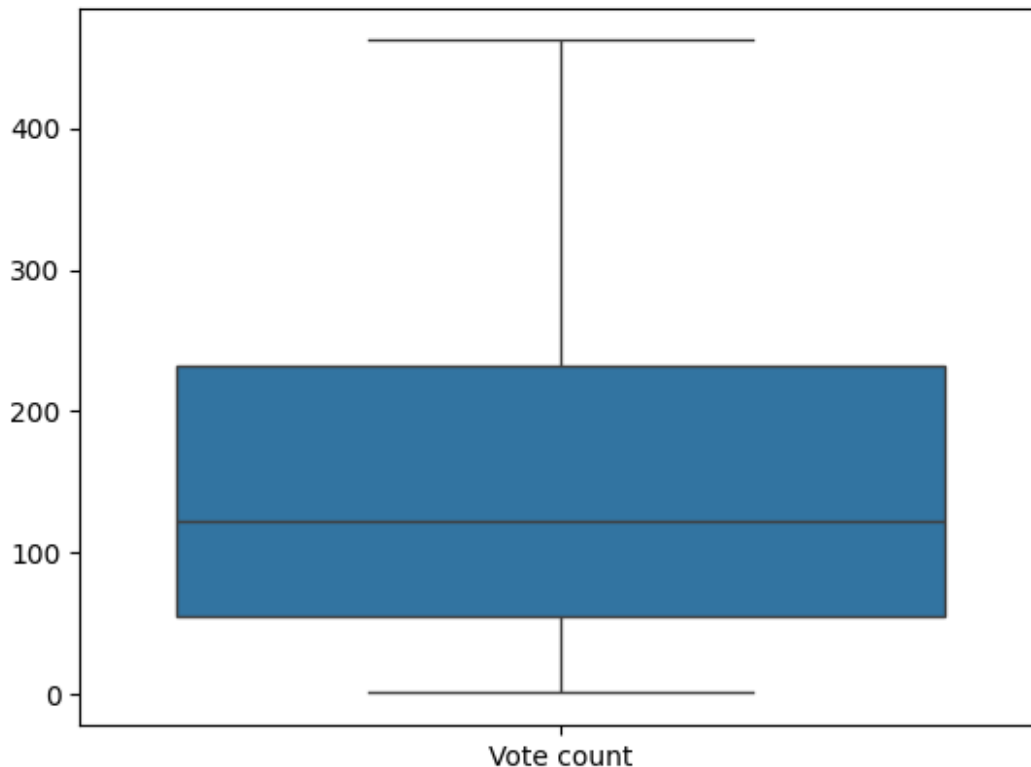
df = df_avg[~outliers]

df

{"repr_error": "0", "type": "dataframe", "variable_name": "df"}

#Vote count are free from outliers
sns.boxplot(df[['Vote count']])
plt.show()

```



*#Now the data is free from outlier we move to visualization part.*  
df

```
{"repr_error": "0", "type": "dataframe", "variable_name": "df"}
```

```
df.head()
```

```
{"repr_error": "0", "type": "dataframe", "variable_name": "df"}
```

```
df.describe()
```

```
{"summary": "{\n  \"name\": \"df\",\n  \"rows\": 8,\n  \"fields\": [\n    {\n      \"column\": \"Budget\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 24898588.956824202,\n        \"min\": 0.0,\n        \"max\": 75000000.0,\n        \"num_unique_values\": 8,\n        \"samples\": [\n          15149295.003457814,\n          9000000.0,\n          1446.0\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      },\n      \"column\": \"Id\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 14075.997423080373,\n        \"min\": 66.0,\n        \"max\": 45324.0,\n        \"num_unique_values\": 8,\n        \"samples\": [\n          14704.965421853389,\n          11533.5,\n          1446.0\n        ],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      },\n      \"column\": \"Popularity\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
```

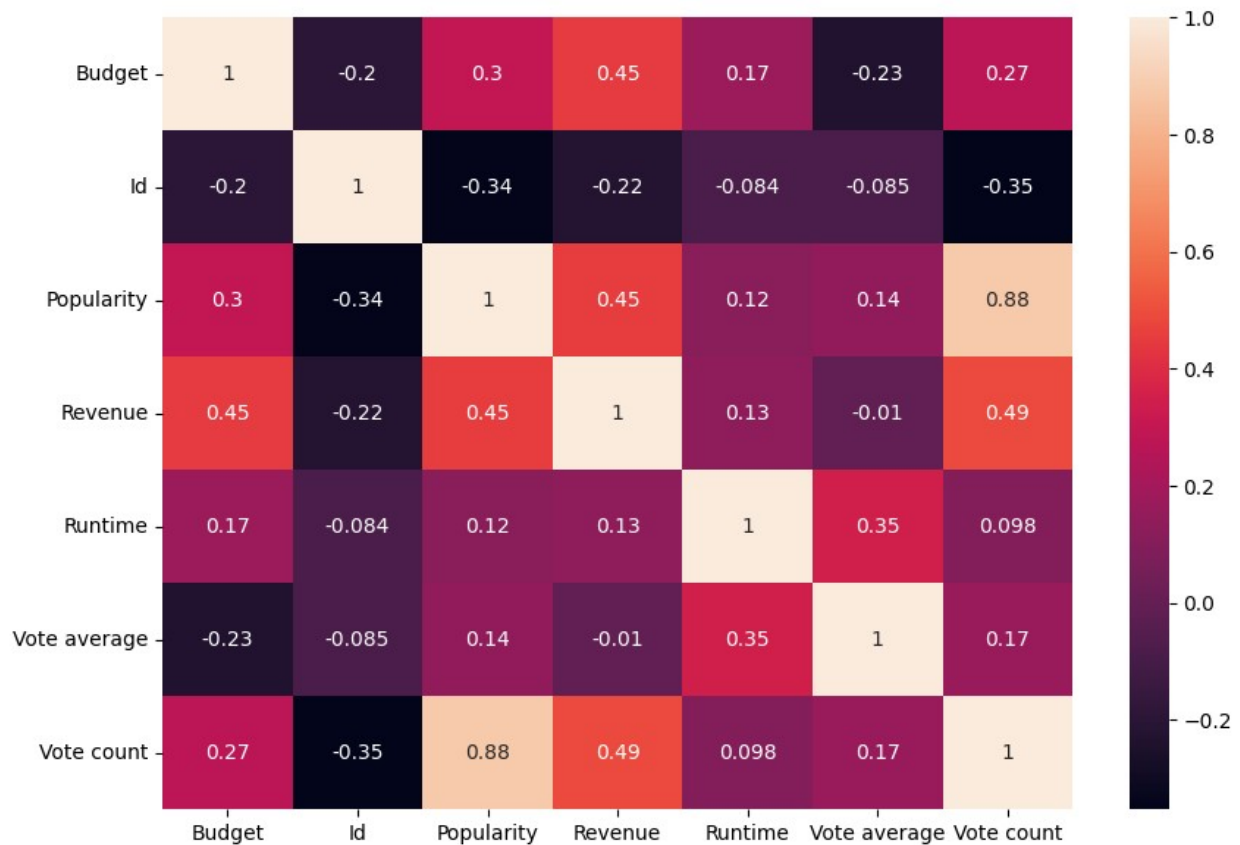


## #Heatmap

There is positive correlation between popularity and vote count which suggest that movies with more vote count are more popular. There is another strong correlation between Revenue and Popularity which suggest that movies tends to generate high revenue when they popular.

Lastly we can see that there is less correlation between Revenue and vote average means slight longer movie can be precieved better.

```
plt.figure(figsize=(10,7))
sns.heatmap(df.corr(numeric_only = True),annot=True)
plt.show()
```



*#Here this graph helps understand which language are preferred by views, audience or even maker to increase revenue*

```
language = df['Original language'].value_counts()
```

```
top_5=language.head()
```

```
top_5
```

```
Original language
```

```
en      1374
```

```
fr       21
```

```
es       11
```

```
zh        6
```

```
it        5
```

```
Name: count, dtype: int64
```

*'''From graph english language is used most in movies or is preferred by viewer*

*since English is one of the language that is being used across the continents.*

*Then next language is french followed by spain, chinese and italian '''*

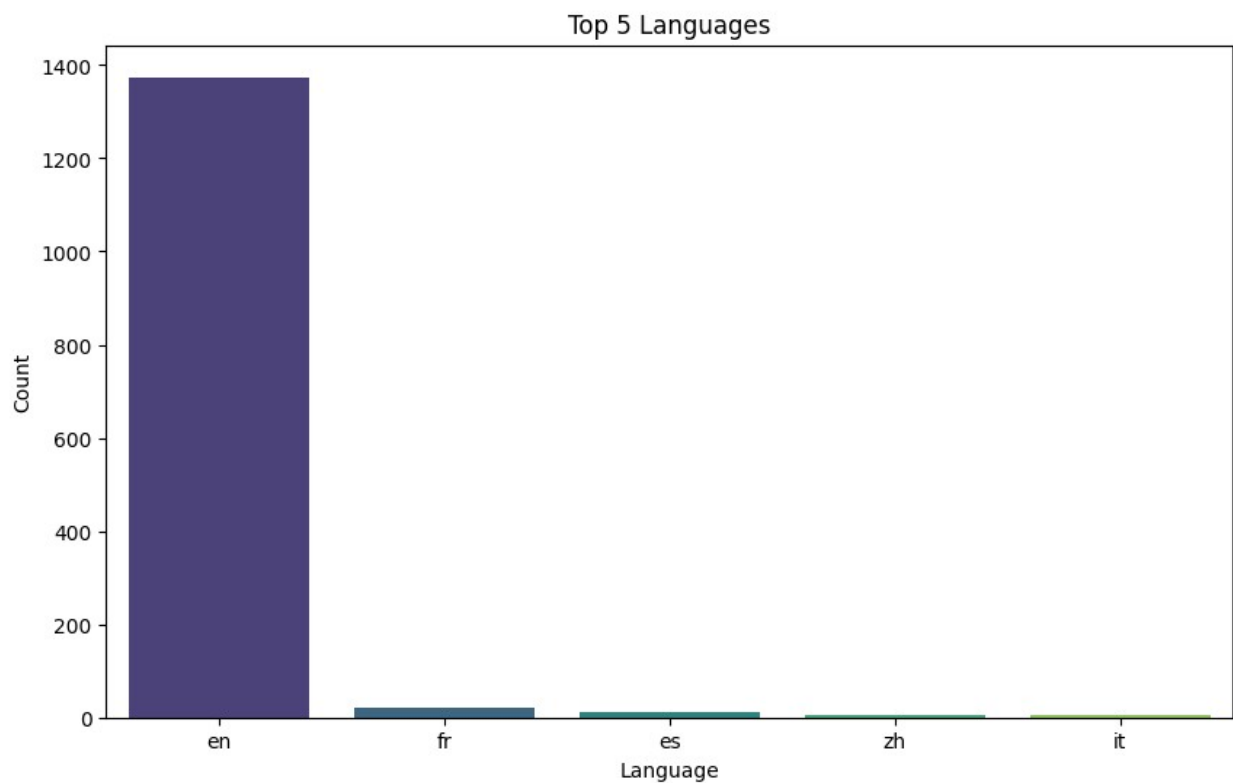
```
plt.figure(figsize=(10,6))
```

```
sns.barplot(top_5,palette='viridis')
plt.xlabel('Language')
plt.ylabel('Count')
plt.title('Top 5 Languages')
plt.show()
```

<ipython-input-309-01c9f80ab44c>:6: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

```
sns.barplot(top_5,palette='viridis')
```



*#Here graph explain which are top 5 movies with high popularity*

```
df_sort = df.sort_values(by=['Popularity'],ascending=False)
```

```
top_5_pop = df_sort.head()
```

```
top_5_pop
```

```
{"repr_error": "0", "type": "dataframe", "variable_name": "top_5_pop"}
```

*''' From the graph Dogtooth is most popular movie followed by  
The Count of Monte Cristo, Bill & Ted's Excellent Adventure, Wild*

*Things, and Astro Boy where Astro boy has least popularity. This popularity depends on some factors such as maketing strategy or vote views even preference of the audience.*

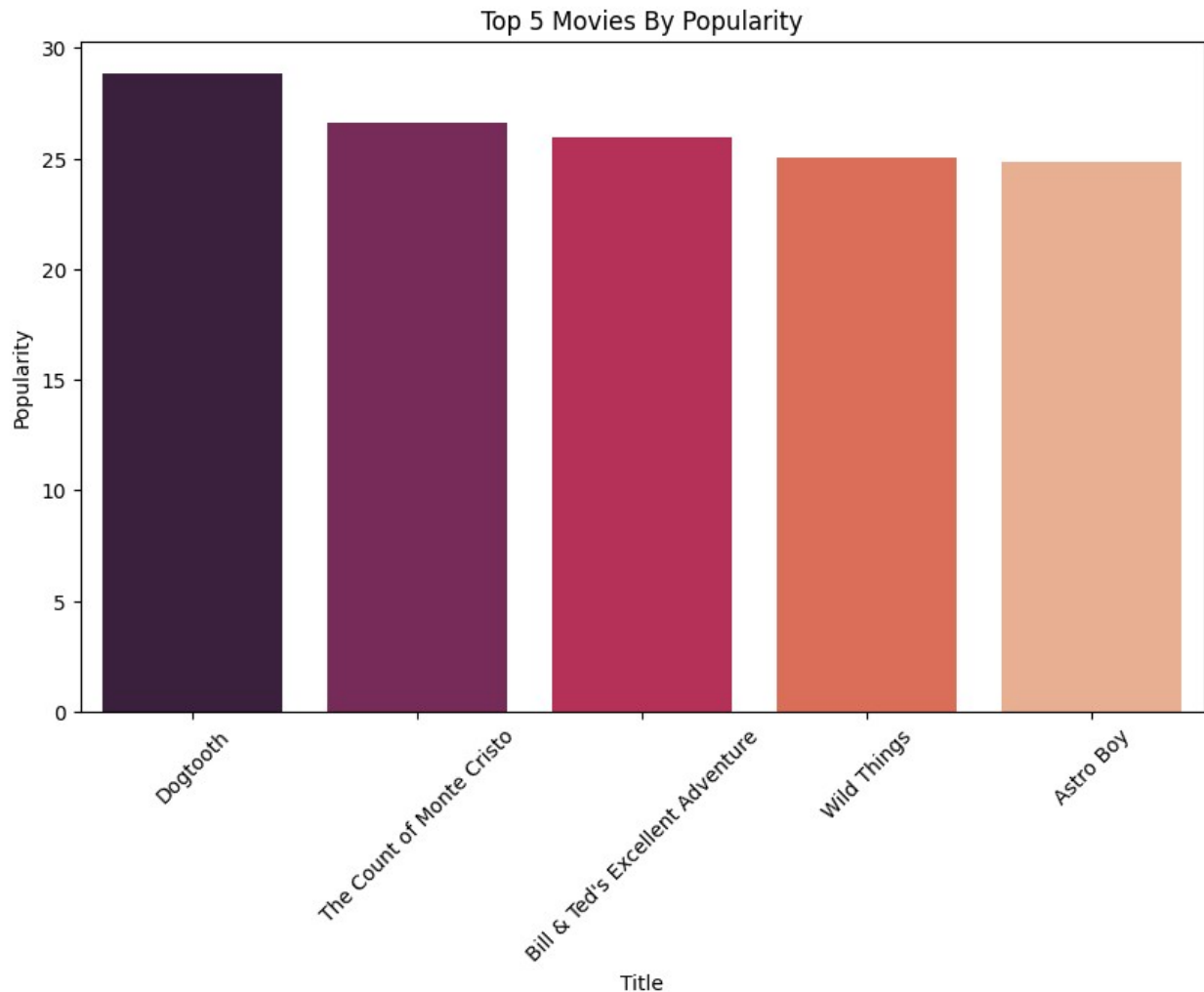
```
'''
```

```
plt.figure(figsize=(10,6))
sns.barplot(x='Title',y='Popularity',data =
top_5_pop,palette='rocket')
plt.xticks(rotation=45)
plt.xlabel('Title')
plt.ylabel('Popularity')
plt.title('Top 5 Movies By Popularity')
plt.show()
```

<ipython-input-311-54c0c9829093>:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

```
sns.barplot(x='Title',y='Popularity',data =
top_5_pop,palette='rocket')
```



```
#Here graphs explain which of top 5 movies has highest Revenue.
df_r = df.sort_values(by = ['Revenue'],ascending=False)
top_5_revenue = df_r.head()
top_5_revenue
{"repr_error":"0","type":"dataframe","variable_name":"top_5_revenue"}
'''Teenage Mutant Ninja Turtles II: The Secret of the Ooze is the
highest-grossing movie with the tallest bar.
This graph also suggest top 5 movies with highest revenue that is the
movies that was successful on box office.
This increase in revenue depends on marketing, mouth to mouth
ppublicity of movie even popularity of the movie.
'''

plt.figure(figsize=(10,6))
```

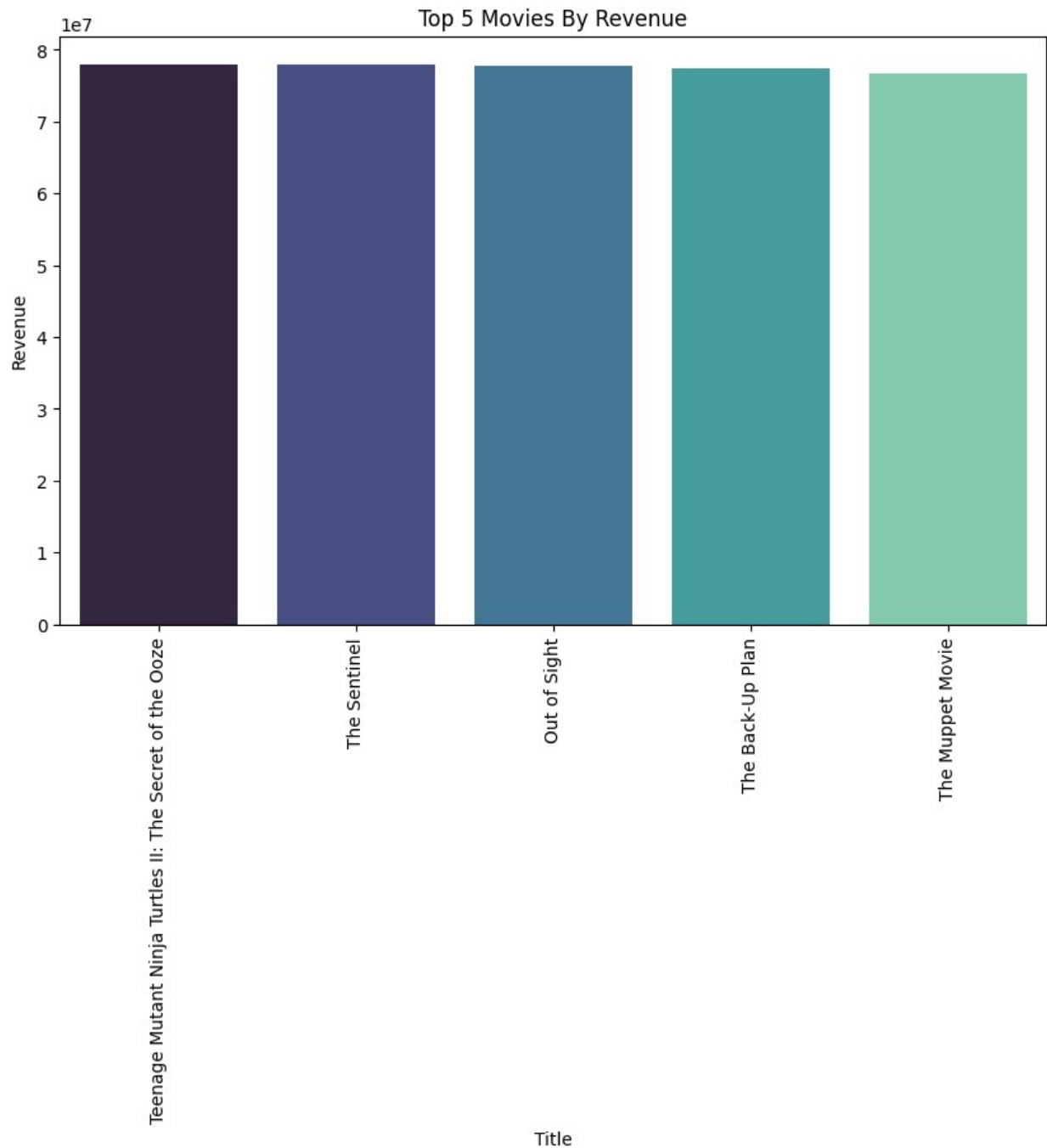
```
sns.barplot(x="Title", y='Revenue', data=top_5_revenue,  
palette='mako')  
plt.xticks(rotation=90)  
plt.xlabel('Title')  
plt.ylabel('Revenue')  
plt.title('Top 5 Movies By Revenue')  
plt.show()
```

<ipython-input-313-d4a9b528a3a4>:7: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

```
sns.barplot(x="Title", y='Revenue', data=top_5_revenue,  
palette='mako')
```





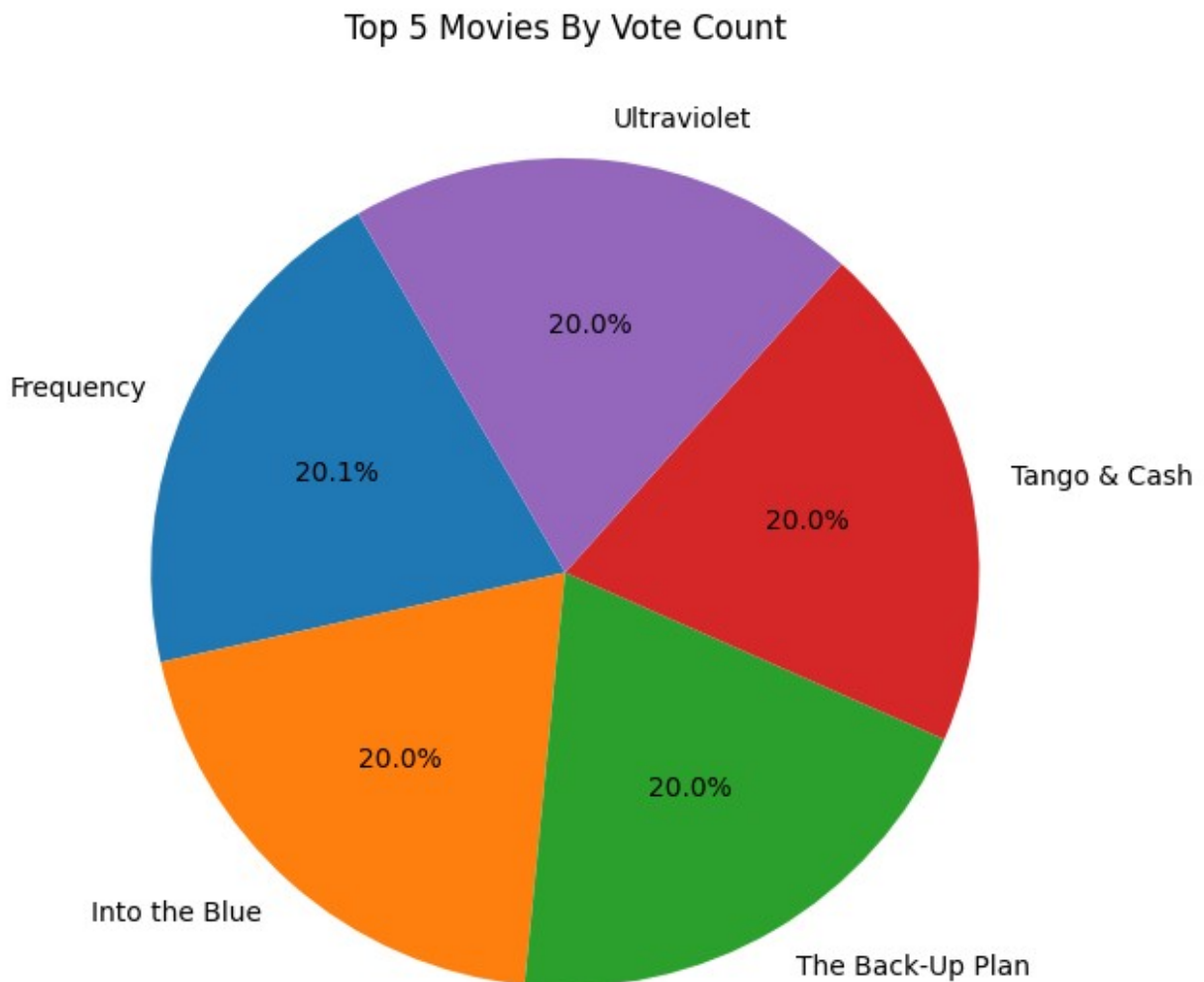
*#Here top 5 movies have similar vote count but Frequency has little more than compard with other.  
#The vote count can reflect audience interest and engagement with these movies.*

```
df_pie = df.sort_values(by = ['Vote count'], ascending = False)
top_5_voted = df_pie.head()
```

```

top_5_voted
{"repr_error": "0", "type": "dataframe", "variable_name": "top_5_voted"}
plt.figure(figsize=(10,7))
plt.pie(top_5_voted['Vote count'], labels=top_5_voted['Title'],
autopct='%1.1f%%', startangle=120)
plt.title('Top 5 Movies By Vote Count')
Text(0.5, 1.0, 'Top 5 Movies By Vote Count')

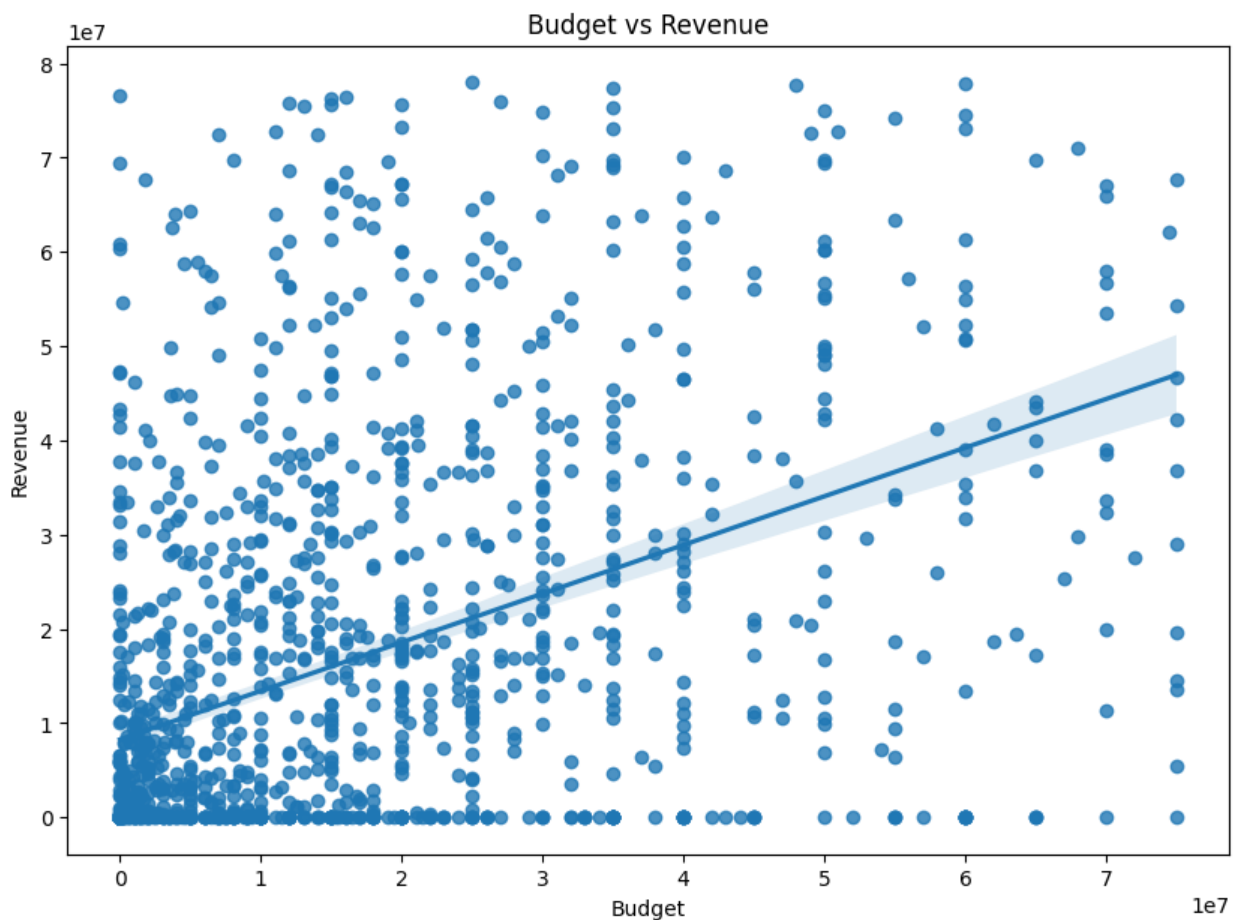
```



'''The line drawn through the data represents a linear regression model.  
It's an attempt to fit a straight line that best represents the overall trend in the data.

The data points are scattered, indicating that while there is a general trend, there is also a lot of variability. Some high-budget movies might not have high revenue, and some low-budget movies might be very successful. This scatter plot suggest that budget is a factor that can influence the revenue of the movie. But it cant be a solely factor there must be different factors like marketing, quality of movie and audience also play significant important role.

```
plt.figure(figsize=(10,7))
sns.regplot(x='Budget',y='Revenue',data=df)
plt.xlabel('Budget')
plt.ylabel('Revenue')
plt.title('Budget vs Revenue')
plt.show()
```



```
df['Runtime'].unique()
array([ 82., 110., 123., 120., 117.,  84.,  78., 100.,  94.,  99.,
        93.,
```

```

113., 109., 125., 111., 116., 97., 95., 107., 114., 124.,
103.,
127., 121., 102., 104., 105., 87., 115., 89., 98., 119.,
108.,
96., 101., 131., 90., 86., 118., 91., 92., 132., 88.,
75.,
80., 112., 106., 85., 126., 122., 128., 130., 81., 76.,
83.,
77., 73., 129., 74., 79.])

```

```
df['Runtime'].max()
```

```
132.0
```

```
df['Runtime'].min()
```

```
73.0
```

```
'''converting runtime into bins of minutes to get more accurate
result.
'''
```

```
bins = [0, 70, 80, 90, 100, 110, 120, 130, 140]
```

```
labels = ['0-70', '70-80', '80-90', '90-100', '100-110', '110-120', '120-130', '130-140']
```

```
df['Runtime Group'] = pd.cut(df['Runtime'], bins, labels=labels,
right=False)
```

```
df.head()
```

```
<ipython-input-338-d3bf28c76d42>:8: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
```

```
df['Runtime Group'] = pd.cut(df['Runtime'], bins, labels=labels,
right=False)
```

```
{"repr_error": "0", "type": "dataframe", "variable_name": "df"}
```

```
df['Runtime Group'].value_counts()
```

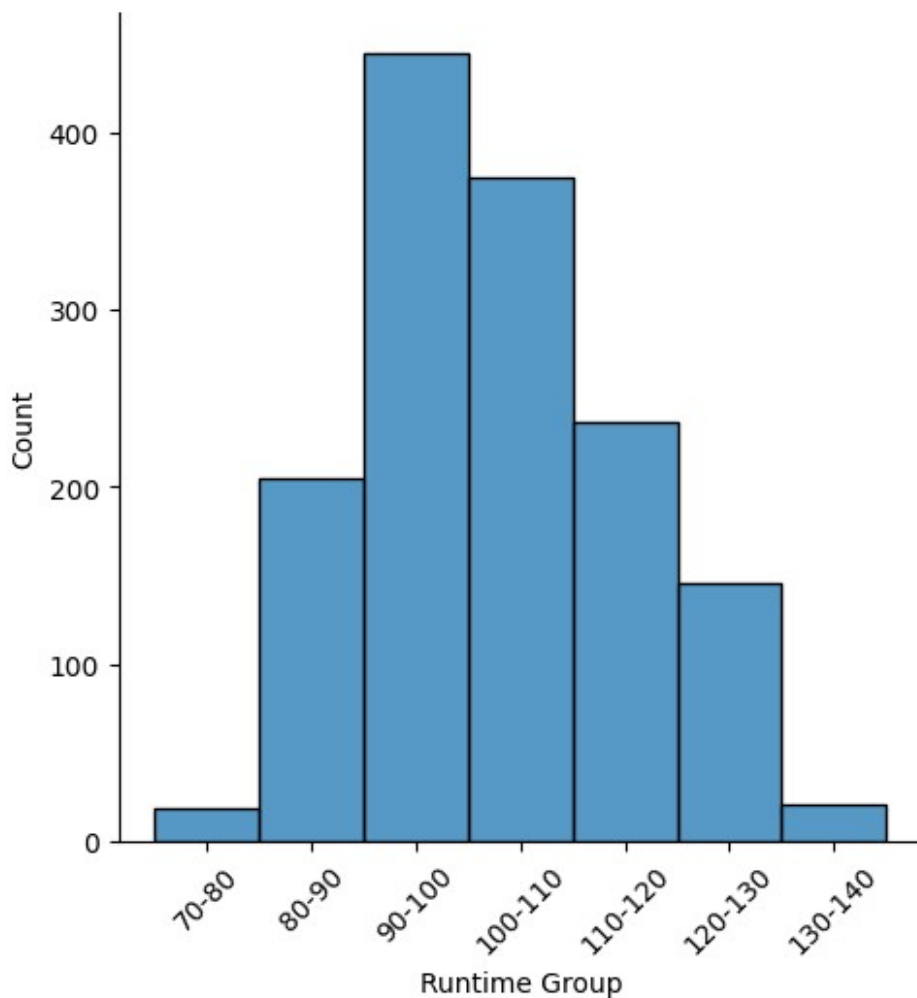
```
Runtime Group
90-100      445
100-110     374
110-120     236
80-90       205
120-130     146
```

```
130-140    21
70-80      19
0-70        0
Name: count, dtype: int64
```

```
'''From the graph we can observe that most of the movie has runtime
within 90-100 minutes.
This suggest that its the time span which is mostly preffered by the
audience.
'''
```

```
plt.figure(figsize=(10,6))
sns.displot(df['Runtime Group'])
plt.xlabel('Runtime Group')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```

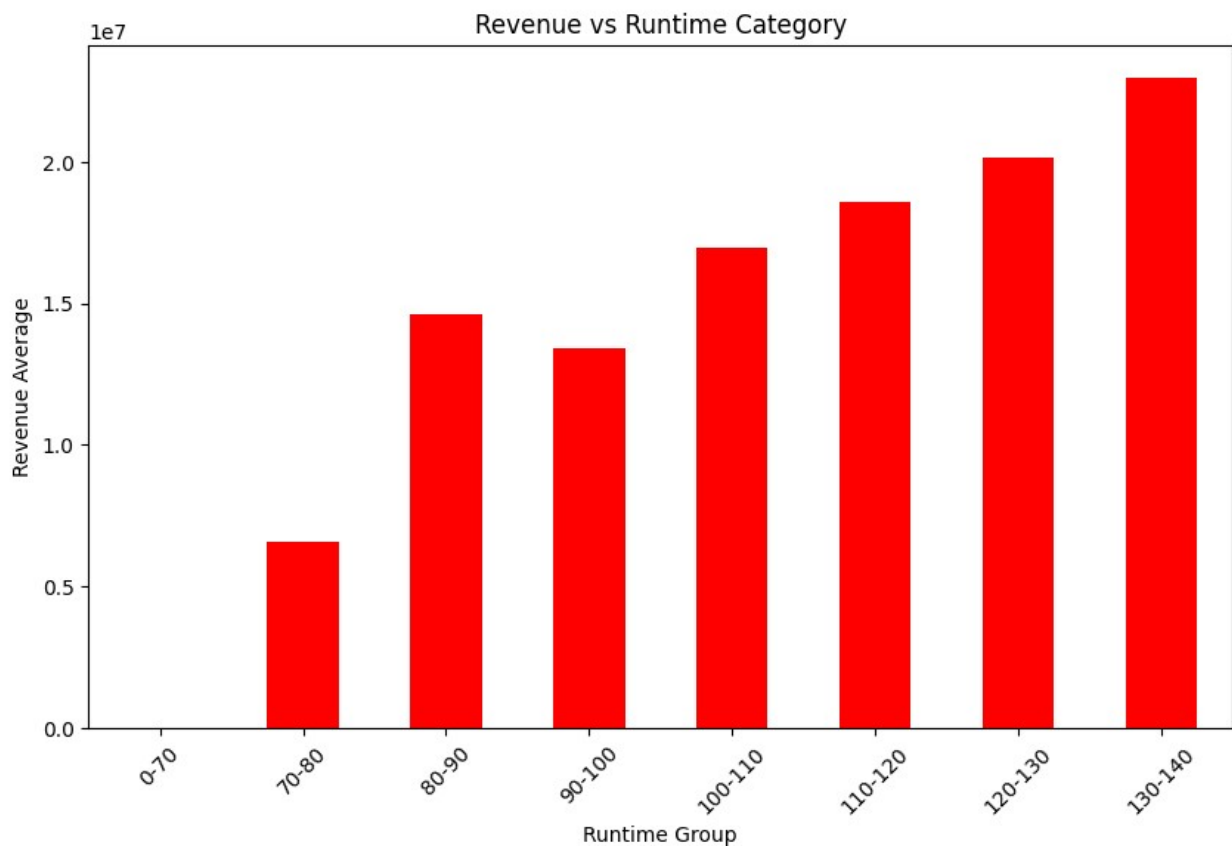
```
<Figure size 1000x600 with 0 Axes>
```



```
'''Average Revenue generated by movie is high for movie with timespan  
or runtime more than 130 minutes.  
The reason can be due to detail explanation of movie or topic related  
to movie.  
'''
```

```
plt.figure(figsize=(10, 6))  
df.groupby('Runtime Group')['Revenue'].mean().plot(kind='bar',  
color='red')  
plt.xlabel('Runtime Group')  
plt.ylabel('Revenue Average')  
plt.title('Revenue vs Runtime Category')  
plt.xticks(rotation=45)  
plt.show()
```

```
<ipython-input-347-710384b8a155>:6: FutureWarning: The default of  
observed=False is deprecated and will be changed to True in a future  
version of pandas. Pass observed=False to retain current behavior or  
observed=True to adopt the future default and silence this warning.  
df.groupby('Runtime Group')['Revenue'].mean().plot(kind='bar',  
color='red')
```



```
'''Similar to Revenue popularity of movie is high for movie with  
timespan or runtime between 120-140 minutes.  
This concludes that movie with higher runtime are preferable by  
audience.  
'''
```

```
plt.figure(figsize=(10,6))  
df.groupby('Runtime Group')  
['Popularity'].mean().plot(kind='bar',color='blue')  
plt.xlabel('Runtime Group')  
plt.ylabel('Popularity Average')  
plt.title('Popularity vs Runtime Category')  
plt.xticks(rotation=45)  
plt.show()
```

```
<ipython-input-348-2ab5e35c776d>:6: FutureWarning: The default of  
observed=False is deprecated and will be changed to True in a future  
version of pandas. Pass observed=False to retain current behavior or  
observed=True to adopt the future default and silence this warning.  
df.groupby('Runtime Group')  
['Popularity'].mean().plot(kind='bar',color='blue')
```

