Laporan Praktikum Processing SPSS & Python

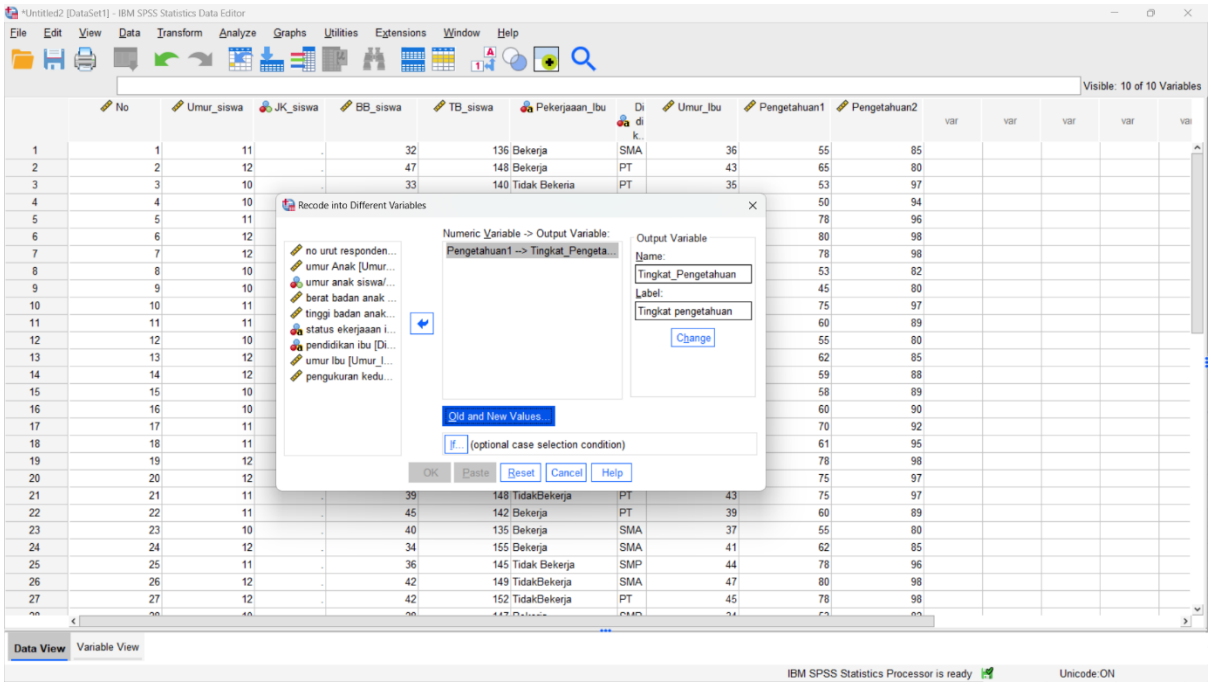
Dosen : Edi Satriyanto S.Si., M.Si

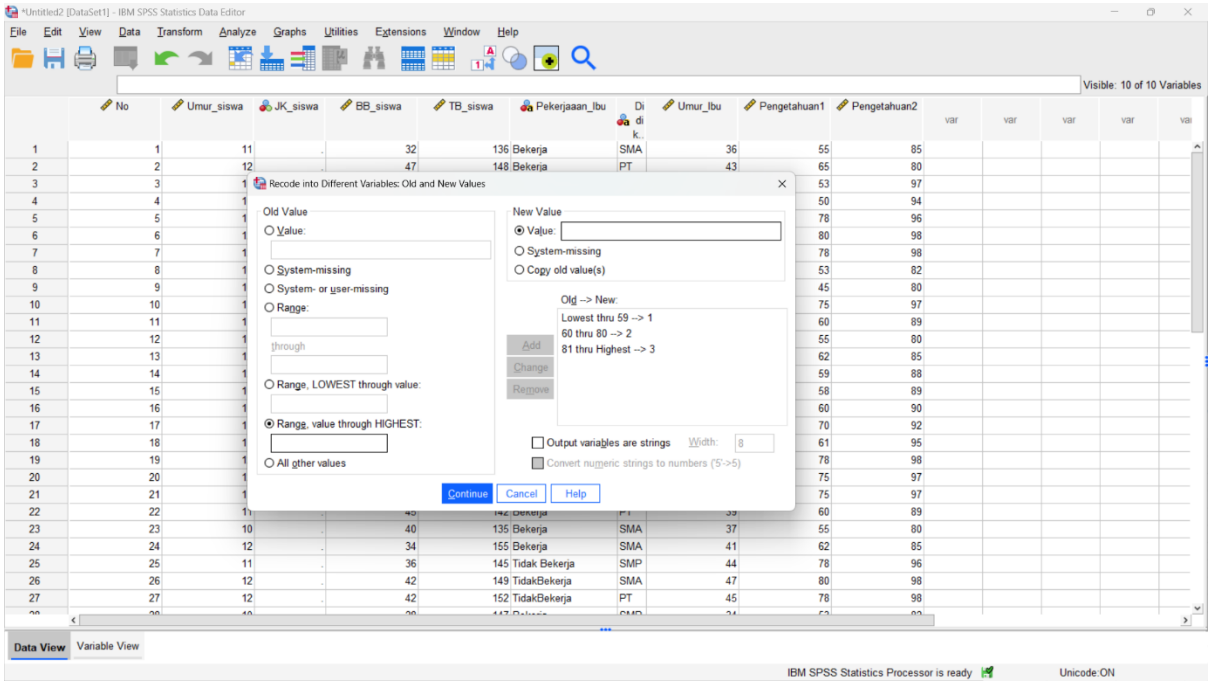
Disusun Oleh:

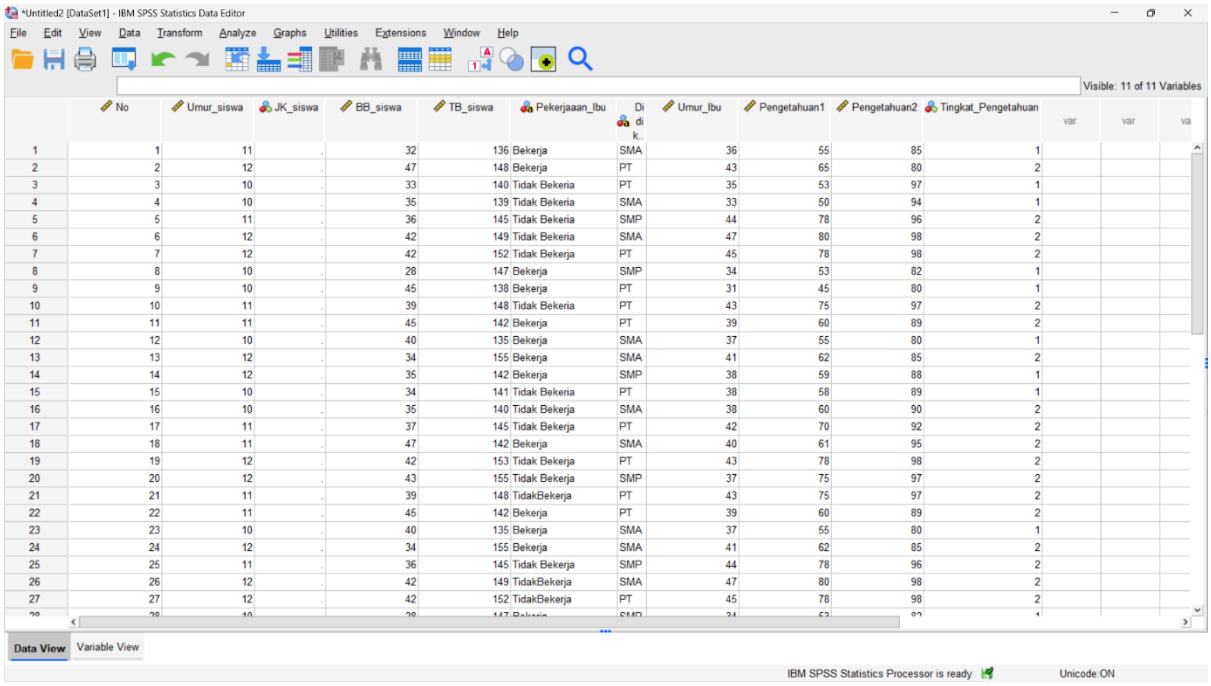
Wahyu Ikbal Maulana (3323600056)

POLITEKNIK ELEKTRONIKA NEGERI SURABAYA

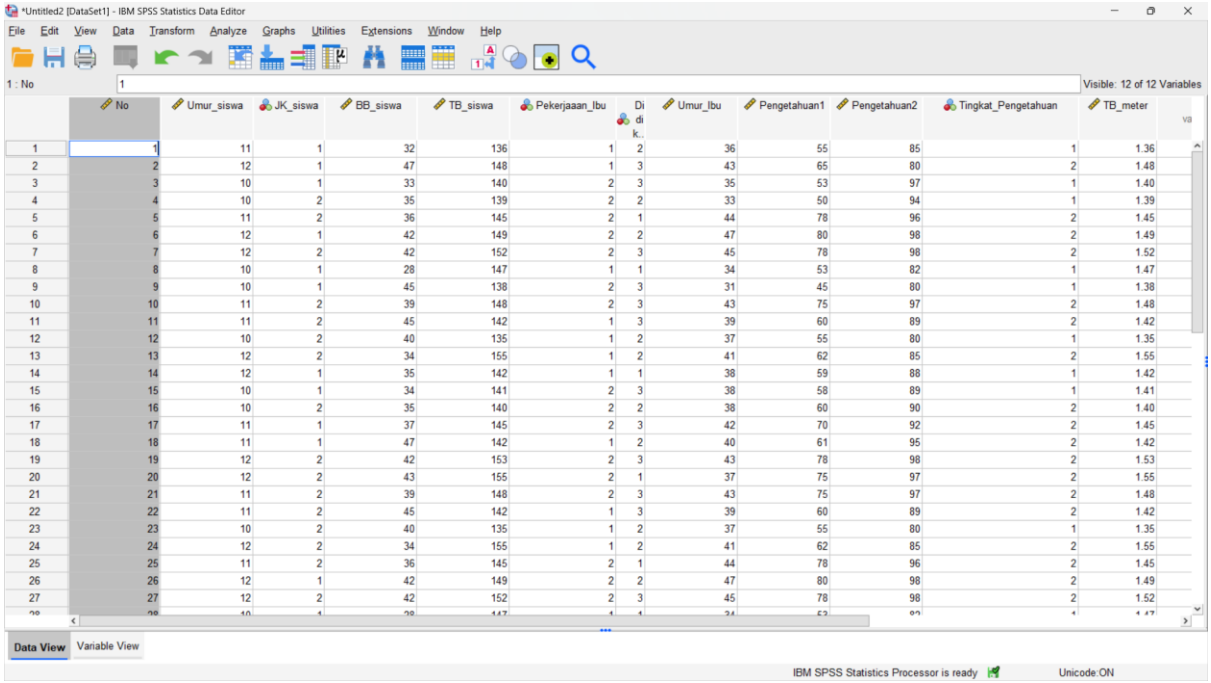
PENGOLAHAN DATA MENGGUNAKAN SPSS







Setelah dilakukan preprocessing sesuai dengan tutorial di pdf, berikut hasil akhir yang saya peroleh, berupa data full numeric yang dapat terbaca oleh komputer



Dikarenakan saya tidak terlalu banyak menscreenshot saat sedang melakukan preprocessing, jadi cukup sedikit yang saya dokumentasikan, selebihnya sudah ada di Pengolahan Data SPSS.pdf

Data preprocessing Movie_sample_dataset using Python

mengimport library

disini kita mengimport beberapa library yang dibutuhkan untuk melakukan preprocessing data

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

mengimport the Dataset untuk kemudian dianalisa

disini kita mengimport untuk membaca dataset movie sample dataset yang kemudian selanjutnya kita prepreocessing

In [ ]:
```
df = pd.read_csv('movie_sample_dataset.csv', encoding='utf-8')
df.head()
```

Out[ ]:

	color	director_name	duration	gross	genres	movie_title	title_y
0	Color	Martin Scorsese	240	116866727.0	Biography|Comedy|Crime|Drama	The Wolf of Wall Street	2(
1	Color	Shane Black	195	408992272.0	Action|Adventure|Sci-Fi	Iron Man 3	2(
2	color	Quentin Tarantino	187	54116191.0	Crime|Drama|Mystery|Thriller|Western	The Hateful Eight	2(
3	Color	Kenneth Lonergan	186	46495.0	Drama	Margaret	2(
4	Color	Peter Jackson	186	258355354.0	Adventure|Fantasy	The Hobbit: The Desolation of Smaug	2(

Menjabarkan dataset

In [ ]:
```
df.drop(['color','language'], axis=1, inplace=True)
df.head()
```

Out[ ]:

	director_name	duration	gross	genres	movie_title	title_year	co
0	Martin Scorsese	240	116866727.0	Biography|Comedy|Crime|Drama	The Wolf of Wall Street	2013	
1	Shane Black	195	408992272.0	Action|Adventure|Sci-Fi	Iron Man 3	2013	
2	Quentin Tarantino	187	54116191.0	Crime|Drama|Mystery|Thriller|Western	The Hateful Eight	2015	
3	Kenneth Lonergan	186	46495.0	Drama	Margaret	2011	
4	Peter Jackson	186	258355354.0	Adventure|Fantasy	The Hobbit: The Desolation of Smaug	2013	

Mencari missing Values

In [ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   director_name         88 non-null     object
 1   duration              99 non-null     int64
 2   gross                 91 non-null     float64
 3   genres                98 non-null     object
 4   movie_title           99 non-null     object
 5   title_year            99 non-null     int64
 6   country               99 non-null     object
 7   budget                95 non-null     float64
 8   imdb_score            99 non-null     float64
 9   actors                99 non-null     object
 10  movie_facebook_likes  99 non-null     int64
dtypes: float64(3), int64(3), object(5)
memory usage: 8.6+ KB
```

Menangani error dan missing Values

Mengekstrack nilai kosong pada tiap tiap variabel yang ada pada dataset

1. director_name (11 Missing Values)

In [ ]:
```
#Total Number of Missing NA
print(df.isnull().sum())
```

```
director_name           11
duration                 0
gross                    8
genres                   1
movie_title              0
title_year               0
country                  0
budget                   4
imdb_score               0
actors                   0
movie_facebook_likes     0
dtype: int64
```

2. Duration (Outliers)

In [ ]:
```
df['director_name'].unique()
```

Out[ ]:
```
array(['Martin Scorsese', 'Shane Black', 'Quentin Tarantino',
       'Kenneth Lonergan', 'Peter Jackson', nan, 'Edward Hall',
       'Joss Whedon', 'Tom Tykwer', 'Null', 'Christopher Spencer',
       'Christopher Nolan', 'F. Gary Gray', 'Richard Linklater',
       'Michael Bay', 'Tom Hooper', 'Kathryn Bigelow', 'Ridley Scott',
       'Denis Villeneuve', 'Gnana Rajasekaran', 'Marc Webb', 'Nan',
       'Mike Leigh', 'Gore Verbinski', 'David Fincher', 'Bryan Singer',
       'Jay Oliva', 'Paul Thomas Anderson', 'Sam Mendes',
       'Michael Patrick King', 'Tate Taylor', 'Francis Lawrence',
       'Steven Spielberg', 'Guillaume Canet', 'Adam McKay', 'Zack Snyder',
       'Baz Luhrmann', 'Timur Bekmambetov', 'Justin Chadwick',
       'Oliver Stone', 'David Dobkin', 'Ryan Murphy', 'James Wan',
       'Derek Cianfrance', "Gavin O'Connor", 'Gary Ross',
       'Terrence Malick', 'Robert Zemeckis', 'Darren Aronofsky',
       'James Mangold', 'Daniel Espinosa', 'Walter Salles',
       'Angelina Jolie Pitt', 'Seth MacFarlane', 'Anthony Russo',
       'Rob Marshall', 'Adam Shankman', 'Sadyk Sher-Niyaz', 'Tony Gilroy',
       'Steve McQueen', 'Richard J. Lewis', 'Paul Greengrass',
       'David Ayer', 'Clint Eastwood'], dtype=object)
```

3. Gross (8 Missing Values)

In [ ]:
```
df['duration'].unique()
```

Out[ ]:
```
array([240, 195, 187, 186, 183, -50, 180, 173, 172, 158, 170, 169, 167,
       165, 580, 164, 157, 156, 154, 153, 151, 150, 650, 149, 148, 147,
       146, 144, 143, 142, 141, 140, 139, 138, 137, 136, 135, 134,   5],
      dtype=int64)
```

4. genres (1 Missing Values)

In [ ]:
```
df['gross'].unique()
```

Out[ ]:
```
array([1.16866727e+08, 4.08992272e+08, 5.41161910e+07, 4.64950000e+04,
       2.58355354e+08, 3.30249062e+08, 3.03001229e+08,            nan,
       6.23279547e+08, 2.70985800e+07, 1.02515793e+08, 5.96961760e+07,
       1.87991439e+08, 1.61029270e+08, 2.53592000e+07, 1.62804648e+08,
       2.45428137e+08, 4.48130642e+08, 2.55108370e+08, 1.48775460e+08,
       9.57207160e+07, 1.05219735e+08, 1.83635922e+08, 3.52358779e+08,
       6.09628780e+07, 2.62030663e+08, 2.28430993e+08, 6.50070450e+07,
       1.82204440e+08, 3.95850000e+06, 8.92899100e+07, 1.67735396e+08,
       2.33914986e+08, 2.92568851e+08, 8.09331800e+06, 2.00074175e+08,
       4.07197282e+08, 9.53289370e+07, 1.69705587e+08, 4.24645577e+08,
       7.98833590e+07, 5.28224180e+07, 4.12290000e+04, 1.63772740e+07,
       1.54985087e+08, 2.17531200e+06, 2.91021565e+08, 3.04360277e+08,
       1.44812796e+08, 7.23060650e+07, 2.02853933e+08, 4.07999255e+08,
       4.58991599e+08, 8.32474800e+06, 4.73075500e+07, 4.71050850e+07,
       8.05740100e+07, 3.50034110e+08, 2.13832980e+07, 1.36516620e+07,
       1.50832203e+08, 2.03899670e+07, 3.05139400e+07, 1.33033190e+07,
       1.50117807e+08, 9.37492030e+07, 1.01160529e+08, 1.27968405e+08,
       1.69693900e+07, 1.32550960e+08, 1.20613500e+06, 3.73049500e+07,
       7.17753000e+05, 2.81666058e+08, 1.15603980e+08, 4.26156850e+07,
       2.59746958e+08, 2.41063875e+08, 3.85093420e+07, 5.24746160e+07,
       1.13165635e+08, 5.66678700e+07, 7.50140400e+06, 1.07100855e+08,
       8.57071160e+07, 4.70342720e+07])
```

5. title_year (Outliers)

In [ ]:
```
df['gross'].fillna(0, inplace=True)
```

6. Country (Uppercase, Different Format)

In [ ]:
```
df['genres'].unique()
```

Out[ ]:
```
array(['Biography|Comedy|Crime|Drama', 'Action|Adventure|Sci-Fi',
       'Crime|Drama|Mystery|Thriller|Western', 'Drama',
       'Adventure|Fantasy', 'Drama|Romance', 'Drama|Sci-Fi',
       'Crime|Drama|Mystery|Thriller', nan, 'Adventure|Drama|Sci-Fi',
       'Biography|Crime|Drama|History|Music', 'Drama|Western',
       'Action|Thriller', 'Drama|Musical|Romance',
       'Drama|History|Thriller', 'Action|Adventure|Drama|History',
       'Adventure|Drama|Thriller|Western', 'Biography|Drama|History',
       'Action|Adventure|Fantasy', 'Action|Adventure|Drama',
       'Biography|Drama|History|War', 'Action|Adventure|Western',
       'Action|Adventure|Fantasy|Sci-Fi|Thriller',
       'Action|Animation|Crime|Sci-Fi|Thriller',
       'Action|Adventure|Sci-Fi|Thriller',
       'Comedy|Crime|Drama|Mystery|Romance', 'Action|Adventure|Thriller',
       'Comedy|Drama|Romance', 'Adventure|Sci-Fi|Thriller', 'Drama|War',
       'Action|Drama|Thriller|War', 'Crime|Drama|Thriller', 'Comedy',
       'Action|Adventure|Fantasy|Sci-Fi', 'Drama|Horror|Thriller',
       'Adventure|Drama|Sci-Fi|Thriller', 'Adventure|Drama|History',
       'Crime|Drama', 'Action|Crime|Thriller', 'Drama|Sport',
       'Adventure|Mystery|Sci-Fi', 'Action|Biography|Drama|History|War',
       'Biography|Drama|Music', 'Drama|Fantasy', 'Drama|Thriller',
       'Mystery|Thriller', 'Biography|Crime|Drama', 'Adventure|Drama',
       'Adventure|Sci-Fi', 'Biography|Drama|Sport|War', 'Comedy|Western',
       'Comedy|Drama|Musical|Romance', 'Action|Biography|Drama|History',
       'Comedy|Drama', 'Biography|Drama|Thriller', 'Action|Drama|War',
       'Biography|Drama|Music|Musical'], dtype=object)
```

In [ ]:
```
pd.set_option('display.max_rows', None)
df['genres']
```

Out[ ]:

0	Biography|Comedy|Crime|Drama
1	Action|Adventure|Sci-Fi
2	Crime|Drama|Mystery|Thriller|Western
3	Drama
4	Adventure|Fantasy
5	Action|Adventure|Sci-Fi
6	Adventure|Fantasy
7	Drama|Romance
8	Action|Adventure|Sci-Fi
9	Action|Adventure|Sci-Fi
10	Drama|Sci-Fi
11	Crime|Drama|Mystery|Thriller
12	NaN
13	Adventure|Drama|Sci-Fi
14	Biography|Crime|Drama|History|Music
15	Drama
16	Drama|Western
17	Action|Adventure|Sci-Fi
18	Action|Thriller
19	Adventure|Fantasy
20	Drama|Musical|Romance
21	Drama|Musical|Romance
22	Drama|History|Thriller
23	Action|Adventure|Drama|History
24	Adventure|Drama|Thriller|Western
25	Action|Adventure|Sci-Fi
26	Crime|Drama|Mystery|Thriller
27	Biography|Drama|History
28	Action|Adventure|Fantasy
29	Adventure|Drama|Sci-Fi
30	Action|Adventure|Drama
31	Action|Adventure|Drama
32	Biography|Drama|History|War
33	Biography|Drama|History
34	Action|Adventure|Western
35	Crime|Drama|Mystery|Thriller
36	Action|Adventure|Fantasy|Sci-Fi|Thriller
37	Action|Animation|Crime|Sci-Fi|Thriller
38	Action|Adventure|Sci-Fi|Thriller
39	Comedy|Crime|Drama|Mystery|Romance
40	Action|Adventure|Thriller
41	Action|Adventure|Sci-Fi
42	Comedy|Drama|Romance
43	Drama
44	Adventure|Sci-Fi|Thriller
45	Drama|War
46	Action|Drama|Thriller|War
47	Crime|Drama|Thriller
48	Drama
49	Action|Adventure|Sci-Fi
50	Comedy
51	Action|Adventure|Fantasy|Sci-Fi
52	Action|Adventure|Thriller
53	Action|Adventure|Thriller
54	Drama|Romance
55	Drama|Romance
56	Drama|Horror|Thriller

```
57                    Drama|History|Thriller
58             Action|Adventure|Fantasy|Sci-Fi
59             Adventure|Drama|Sci-Fi|Thriller
60                     Action|Adventure|Sci-Fi
61                     Adventure|Drama|History
62                     Adventure|Drama|History
63                     Adventure|Drama|History
64                     Biography|Drama|History
65                        Crime|Drama|Thriller
66                                 Crime|Drama
67                               Drama|Romance
68                       Action|Crime|Thriller
69                        Crime|Drama|Thriller
70                                 Drama|Sport
71                   Adventure|Mystery|Sci-Fi
72           Action|Biography|Drama|History|War
73                       Biography|Drama|Music
74                               Drama|Fantasy
75                                 Crime|Drama
76                              Drama|Thriller
77                     Action|Adventure|Drama
78                           Mystery|Thriller
79                        Crime|Drama|Thriller
80              Action|Adventure|Sci-Fi|Thriller
81                        Crime|Drama|Thriller
82                       Biography|Crime|Drama
83                             Adventure|Drama
84                             Adventure|Sci-Fi
85                   Biography|Drama|Sport|War
86                   Biography|Drama|Sport|War
87                               Comedy|Western
88                     Action|Adventure|Sci-Fi
89                     Action|Adventure|Fantasy
90                 Comedy|Drama|Musical|Romance
91                                       Drama
92               Action|Biography|Drama|History
93                    Action|Adventure|Thriller
94                     Biography|Drama|History
95                                 Comedy|Drama
96                    Biography|Drama|Thriller
97                             Action|Drama|War
98                 Biography|Drama|Music|Musical
Name: genres, dtype: object
```

In [ ]:
```
df['title_year'].unique()
```

Out[ ]:
```
array([2013, 2015, 2011,  202, 2012, 2014, 2010, 2016,  205], dtype=int64)
```

In [ ]:
```
df['country'].unique()
```

Out[ ]:
```
array(['USA', 'usa', 'UK', 'Germany', 'New Zealand', 'India',
       'United States', 'France', 'Australia', 'Czech Republic',
       'Kyrgyzstan', 'Canada'], dtype=object)
```

In [ ]:
```
df['country'].unique()
```

Out[ ]:
```
array(['USA', 'usa', 'UK', 'Germany', 'New Zealand', 'India',
       'United States', 'France', 'Australia', 'Czech Republic',
       'Kyrgyzstan', 'Canada'], dtype=object)
```

7. Budget (Outlier)

Disini kita melihat garis besar outlier dari rentang dari budget

In [ ]:
```
df['budget'].unique()
```

Out[ ]:
```
array([1.000e+08, 2.000e+08, 4.400e+07, 1.400e+07, 2.250e+08, 2.500e+08,
       1.800e+08,       nan, 2.200e+08, 1.020e+08, 9.000e+07, 2.200e+07,
       1.650e+08, 2.800e+07, 4.000e+06, 2.100e+08, 6.100e+07, 4.000e+07,
       1.350e+08, 1.950e+08, 4.600e+07, 2.300e+08, 1.080e+08, 1.400e+08,
       6.500e+07, 2.150e+08, 3.500e+06, 1.600e+08, 2.000e+07, 2.450e+08,
       2.500e+07, 1.300e+08, 6.600e+07, 5.000e+07, 2.550e+07, 3.200e+07,
       1.780e+08, 1.050e+08, 1.735e+04, 7.800e+07, 3.500e+07, 4.500e+07,
       6.000e+07, 1.900e+08, 1.500e+07, 8.500e+07, 3.000e+07, 3.100e+07,
       1.250e+08, 8.000e+07, 1.200e+08, 1.700e+08, 7.500e+07, 7.000e+07,
       1.400e+06, 5.500e+07, 6.800e+07])
```

8. Imdb_score (Outlier)

Disini kita melihat garis besar outlier dari rentang dari imdb_score

In [ ]:
```
df['imdb_score'].unique()
```

Out[ ]:
```
array([ 8.2,  7.2,  7.9,  6.5,  6.9,  8.1, -7.5,  7.8,  5.6,  8.6,  8. ,
        8.5,  5.7,  7.5,  7.6,  7.4,  6.7,  6.3,  7. ,  6.1,  6.8,  8.4,
        8.8,  4.3,  7.1,  7.3,  3. ,  6. ,  5.8,  5.3,  6.4,  6.6, -1.2,
        5.9,  8.7])
```

In [ ]:
```
df['imdb_score'].unique()
```

Out[ ]:
```
array([ 8.2,  7.2,  7.9,  6.5,  6.9,  8.1, -7.5,  7.8,  5.6,  8.6,  8. ,
        8.5,  5.7,  7.5,  7.6,  7.4,  6.7,  6.3,  7. ,  6.1,  6.8,  8.4,
        8.8,  4.3,  7.1,  7.3,  3. ,  6. ,  5.8,  5.3,  6.4,  6.6, -1.2,
        5.9,  8.7])
```