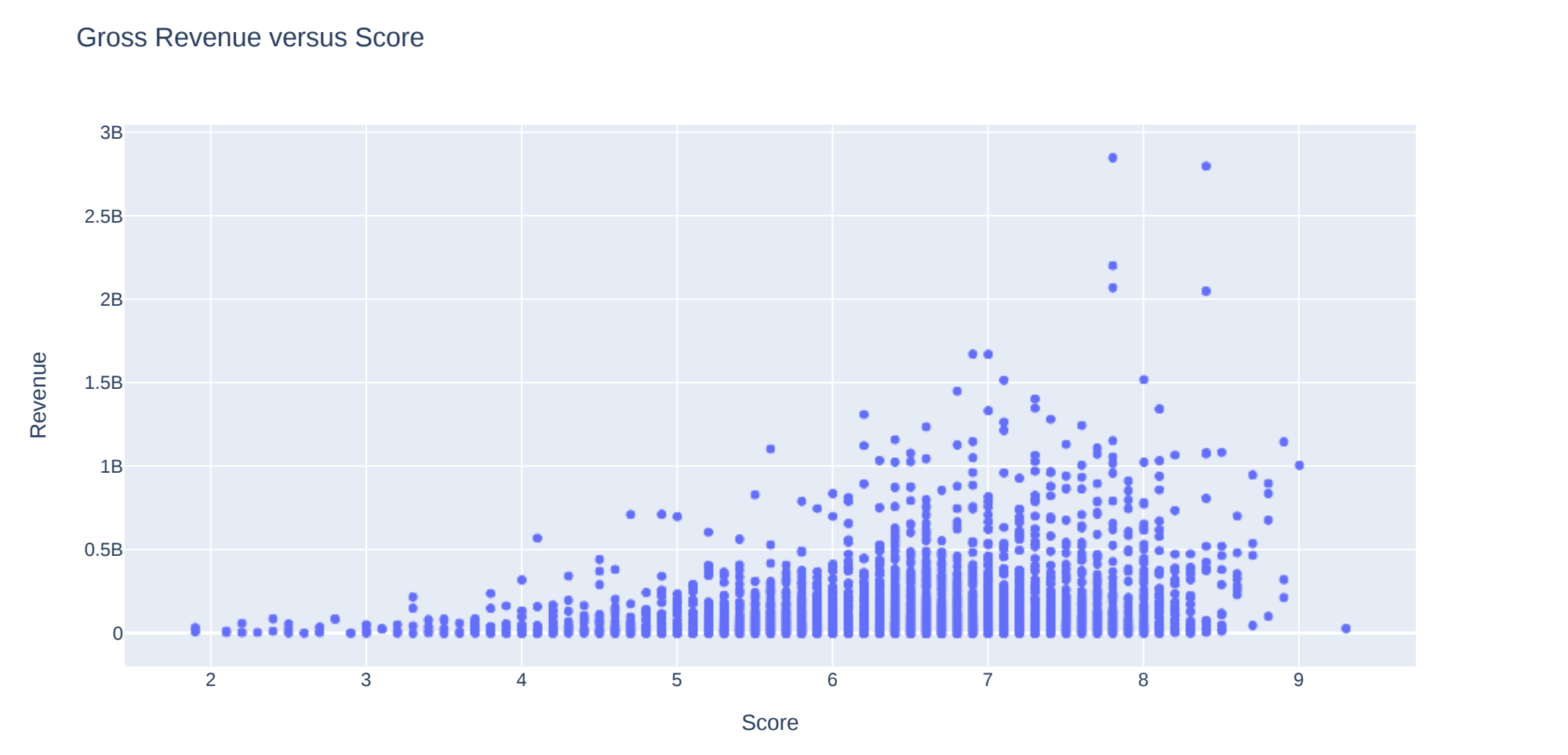
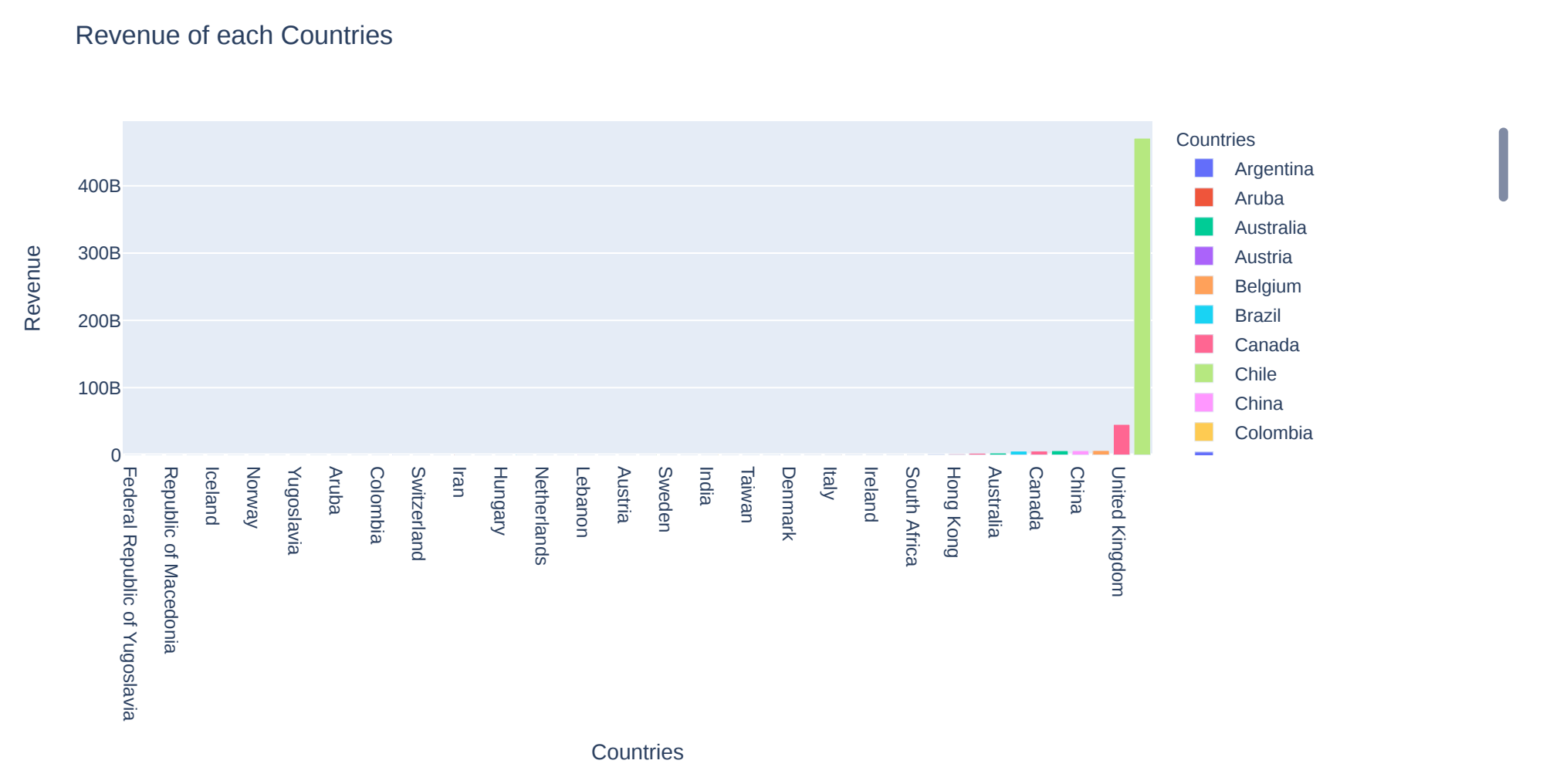
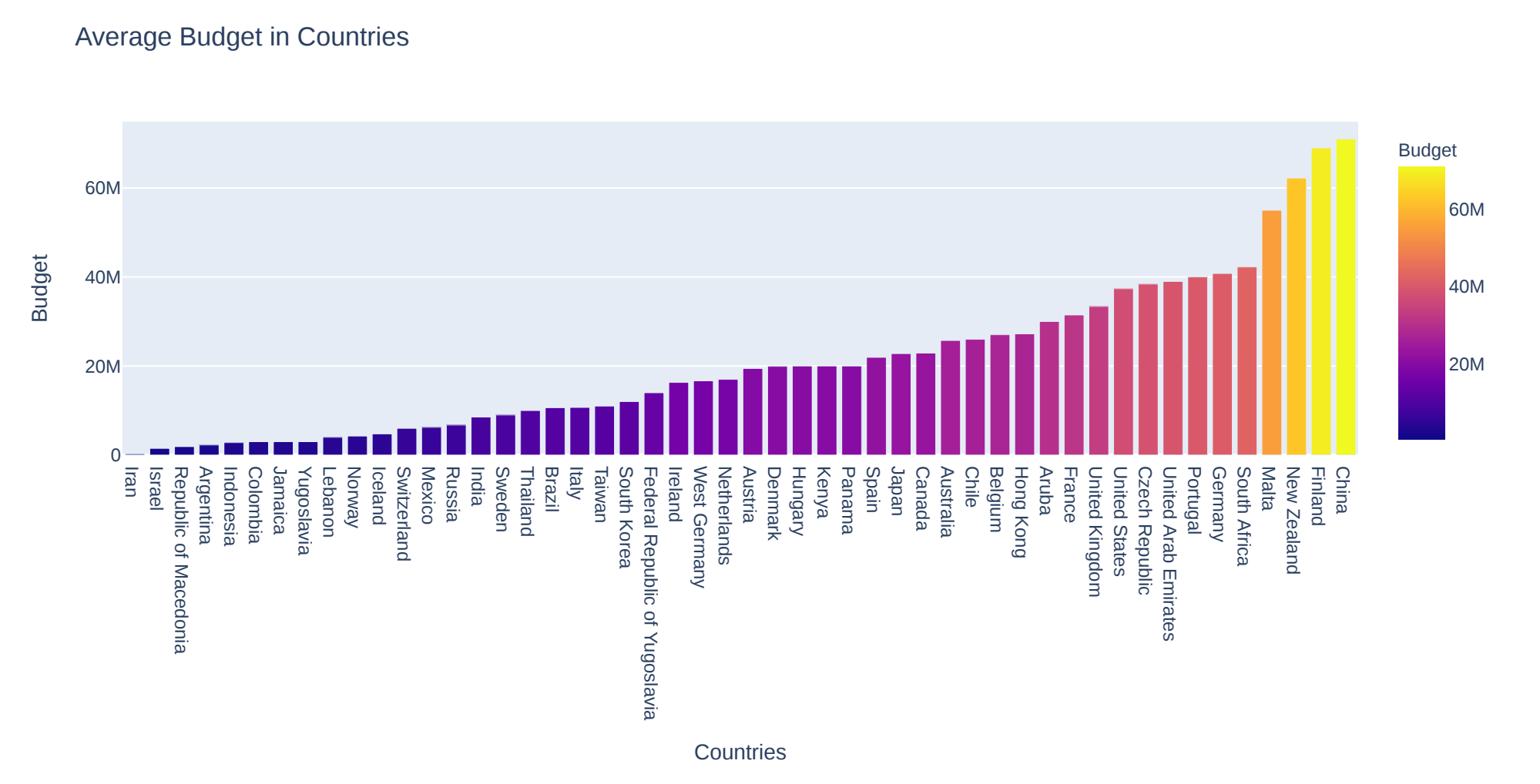
In [1]:
```python
#https://www.kaggle.com/datasets/danielgrijalvas/movies
#Project #3, LinhPhan, Nov 28
import pandas as pd
import plotly.express as px

#import csv file
rawdf = pd.read_csv('movies.csv')

#clean data
cleanedDF = rawdf.dropna()

#MAIN QUESTION: How do Score, Country, Budget, or Genre affect Revenue?

#FACTOR 1: SCORE
#Ques1: Does movies with higher score have higher revenue?
fig1 = px.scatter(cleanedDF, x='score', y='gross',
                  labels={
                      'score': 'Score',
                      'gross': 'Revenue'
                  },
                  title='Gross Revenue versus Score')
fig1.show()
```



In [2]:
```python
#RESULT: There is no relationship between Score and Revenue.

#FACTOR 2: COUNTRY
#Sub2a: Which country generates highest revenue?
sum = cleanedDF.groupby('country')['gross'].sum().reset_index()

fig2a = px.bar(sum, x='country', y='gross', color='country',
               labels={
                   'country': 'Countries',
                   'gross': 'Revenue'
               },
               title='Revenue of each Countries')
fig2a.update_layout(barmode='stack', xaxis={'categoryorder':'total ascending'})
fig2a.show()
```



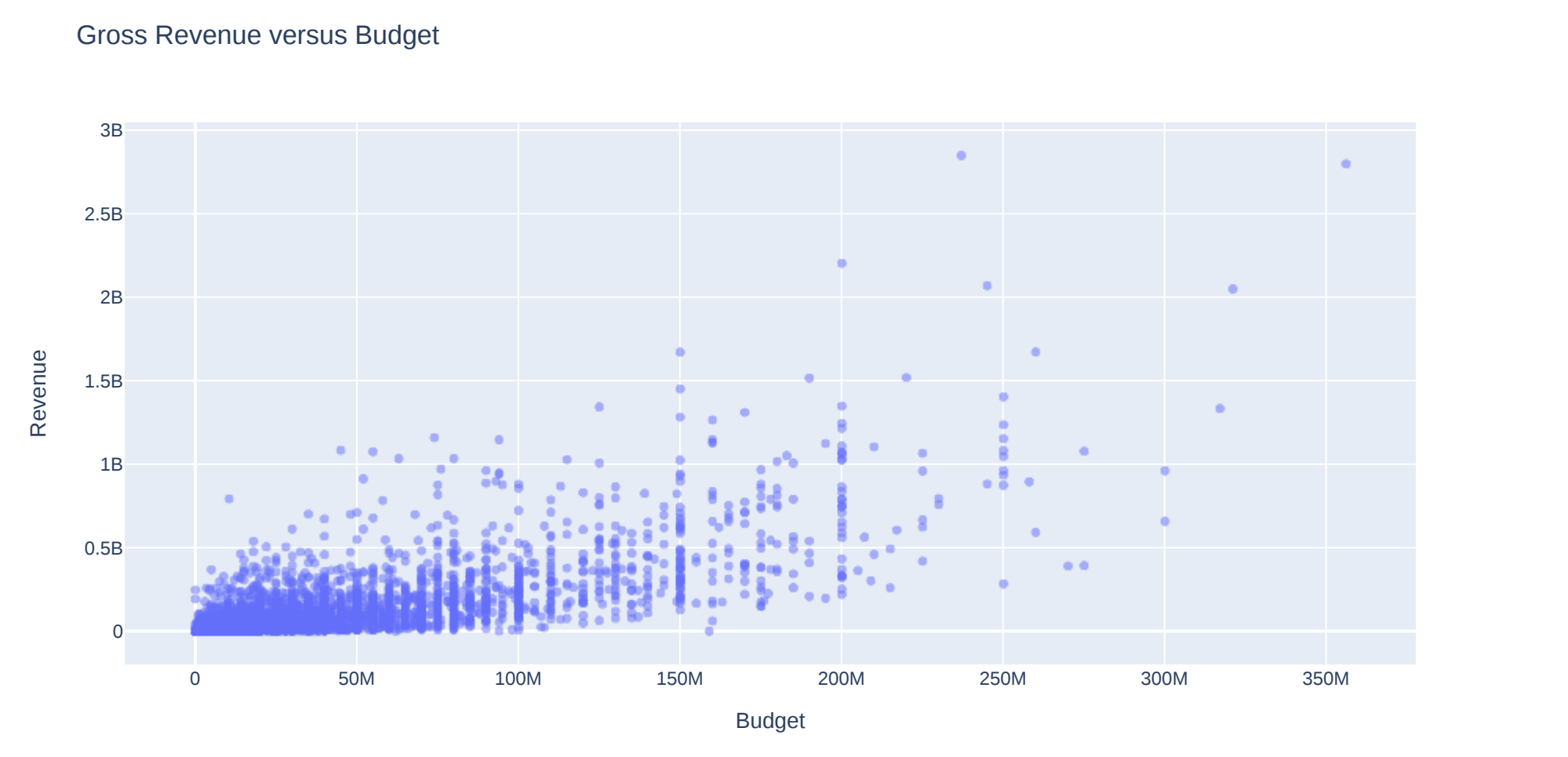
In [3]:
```python
#Answer: United States generates highest revenue.

#Sub2b: Is it because the US having the highest budget?
average = cleanedDF.groupby('country').mean().reset_index()
average = cleanedDF.groupby('country')['budget'].mean().reset_index()

fig2b = px.bar(average, x='country', y='budget', color='budget',
               labels={
                   'country': 'Countries',
                   'budget': 'Budget'
               },
               title='Average Budget in Countries')
fig2b.update_layout(barmode='stack', xaxis={'categoryorder':'total ascending'})
fig2b.show()
```



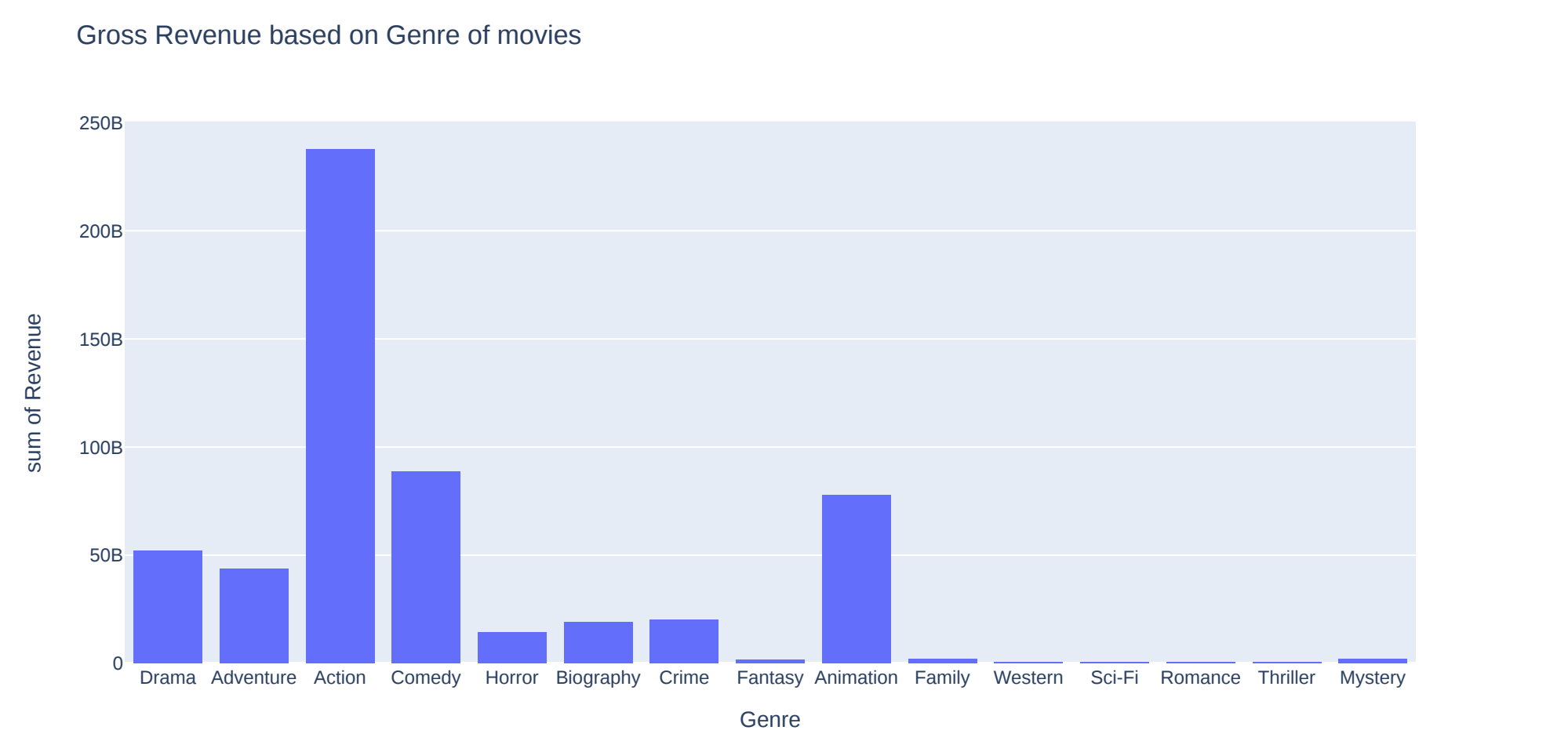
In [4]:
```python
#Answer: US (highest revenue) is not the one having highest budget.
        # Meaning budget is not the reason.

#Sub2c: Is it because the US produce the highest number of movies?
countUS = cleanedDF.groupby('country')['year'].count().reset_index()
countUS = countUS.sort_values('year')
renamed_US = countUS.rename(columns={'year':'No of movies'})
print(renamed_US)
```

```
                          country  No of movies
49                     Yugoslavia             1
25                        Jamaica             1
27                          Kenya             1
17                        Hungary             1
28                        Lebanon             1
29                          Malta             1
12  Federal Republic of Yugoslavia            1
34                         Panama             1
9                        Colombia             1
23                         Israel             1
7                           Chile             1
36          Republic of Macedonia             1
44                       Thailand             1
1                           Aruba             1
35                       Portugal             1
13                        Finland             2
3                         Austria             2
18                        Iceland             2
45           United Arab Emirates             2
20                      Indonesia             2
48                   West Germany             3
31                    Netherlands             3
43                         Taiwan             3
33                         Norway             3
0                       Argentina             3
21                           Iran             3
5                          Brazil             3
4                         Belgium             4
42                    Switzerland             4
38                   South Africa             4
37                         Russia             4
41                         Sweden             5
10                 Czech Republic             7
19                          India             8
11                        Denmark            10
30                         Mexico            11
39                    South Korea            13
40                          Spain            18
32                    New Zealand            18
24                          Italy            19
22                        Ireland            21
16                      Hong Kong            24
8                           China            26
26                          Japan            28
2                       Australia            42
15                        Germany            82
14                         France           105
6                          Canada           111
46                 United Kingdom           492
47                  United States          4319
```

In [5]:
```python
#Answer: US produces the most movies.
        # Meaning United States with highest movie productions generates highest revenue.
#RESULT: Country might affect Revenue, need more researchs for this.
#ExtraCredit #1: We need to find the common characteristics of countries with highest revenue.
# If those characteristics exist, we can conclude that Country affect Revenue. In order to find the answer,
# we need to combine with other datasets such as GDP, Population, Education, etc. of different countries in the worl
d.

#FACTOR 3: BUDGET
#Question 3 (ExtraCredit #2): Does movies with higher budget have higher revenue?
fig3 = px.scatter(cleanedDF, x='budget', y='gross', opacity=0.5,
                  labels={
                      'budget': 'Budget',
                      'gross': 'Revenue'
                  },
                  title='Gross Revenue versus Budget')
fig3.show()
```

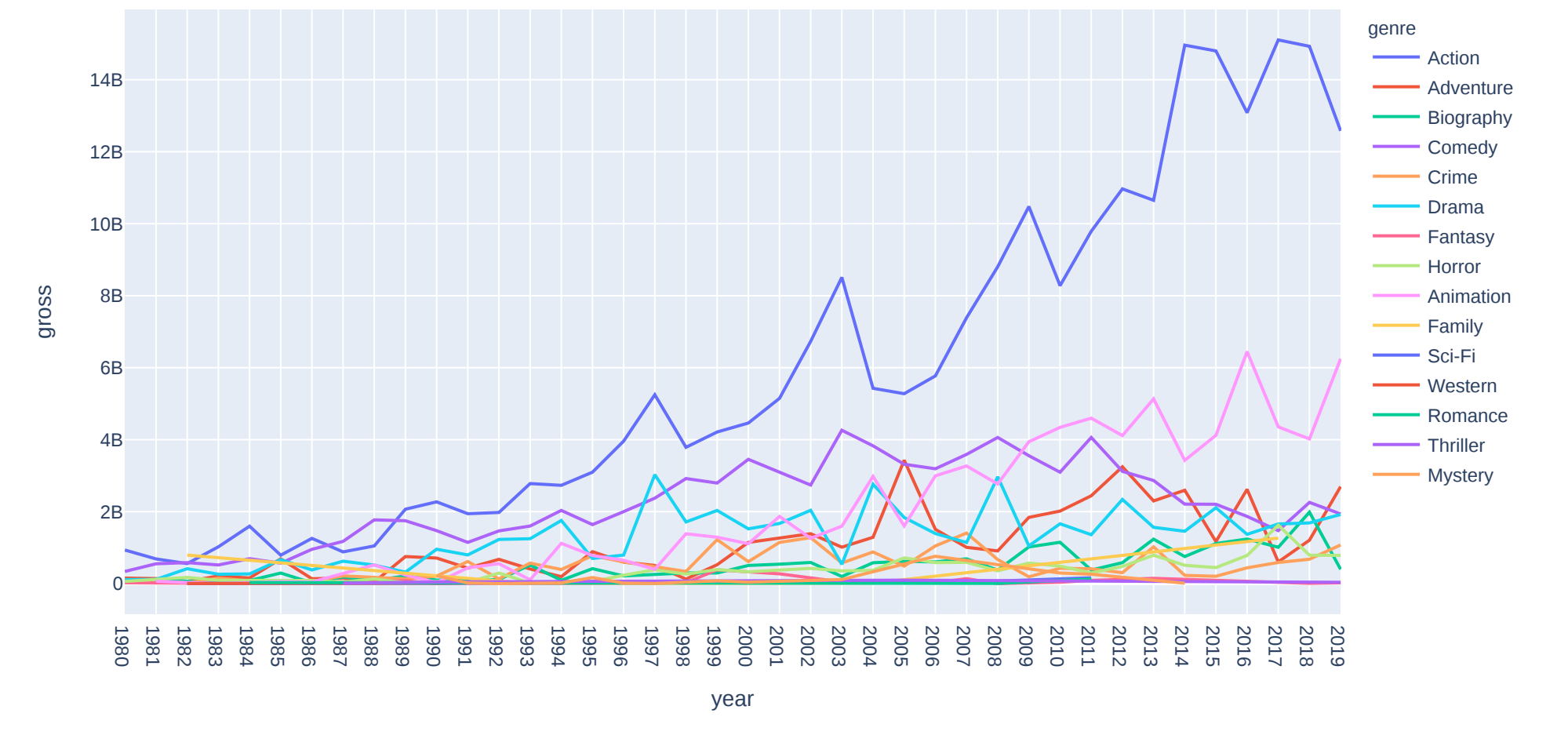


In [6]:
```python
#Answer: Movies with higher budget generate higher revenue while movies with lower budget generate less revenue,
# with a few exceptions (outliers).
#RESULY: Budget affects Revenue.

#FACTOR 4: GENRE
#Sub4a: Gross revenue for each genre of movies?
fig4a = px.histogram(cleanedDF, x='genre', y='gross',
                     labels={
                     'genre': 'Genre',
                     'gross': 'Revenue'
                     },
                     title='Gross Revenue based on Genre of movies')
fig4a.show()
```



In [8]:
```python
#Answer: Overall, Action generates highest revenue, then Comedy and Animation rank 2nd and 3rd.

#Sub4b (ExtraCredit #3): How does the revenue vary from time to time, based on Genre?
cleanedDF = cleanedDF.query('year <= 2019')
cleanedDF['year'] = cleanedDF['year'].apply(str)
cleanedDF = cleanedDF.groupby(['year', 'genre']).sum().reset_index()
print(cleanedDF)
fig4b = px.line(cleanedDF, x="year", y="gross", color='genre')
fig4b.show()
```

```
     year      genre  score       votes        budget         gross  runtime
0    1980     Action   69.5   1563800.0  1.970000e+08  9.335692e+08   1277.0
1    1980  Adventure   31.0    164300.0  7.950000e+07  1.531924e+08    643.0
2    1980  Biography   15.2    340000.0  2.800000e+07  3.919762e+07    229.0
3    1980     Comedy   65.8    426615.0  6.500000e+07  3.381051e+08    998.0
4    1980      Crime   27.0     96000.0  3.130000e+07  1.115789e+08    454.0
5    1980      Drama   33.4   1004150.0  4.030000e+07  1.120693e+08    564.0
6    1980    Fantasy    5.3     12000.0  2.000000e+07  2.276257e+07     96.0
7    1980     Horror   13.2    189000.0  1.550000e+06  6.120338e+07    184.0
8    1981     Action   85.9   1468312.0  1.765000e+08  6.862207e+08   1385.0
9    1981  Adventure   35.9    147700.0  6.100000e+07  1.405919e+08    683.0
10   1981  Animation    7.3     87000.0  1.200000e+07  6.345699e+07     83.0
11   1981  Biography   21.2     92000.0  4.250000e+07  1.187183e+08    449.0
12   1981     Comedy  102.4    319661.0  1.613000e+08  5.509042e+08   1646.0
13   1981      Crime   40.7     93900.0  6.010000e+07  5.350379e+07    717.0
14   1981      Drama   48.3     71859.0  1.487000e+08  1.165428e+08    932.0
15   1981     Horror   57.0    419000.0  3.080000e+07  1.088187e+08    849.0
16   1982     Action   94.8   1469800.0  1.777000e+08  5.563448e+08   1590.0
17   1982  Adventure   14.5     71200.0  1.850000e+07  6.228985e+07    195.0
18   1982  Animation   15.0     62000.0  1.050000e+07  2.112106e+07    174.0
19   1982  Biography   15.7    239000.0  3.150000e+07  6.676789e+07    313.0
20   1982     Comedy  141.7    582921.0  3.282500e+08  5.833238e+08   2543.0
21   1982      Crime   19.6     25400.0  4.200000e+07  1.820930e+07    318.0
22   1982      Drama   84.8    491400.0  1.455000e+08  4.145846e+08   1492.0
23   1982     Family    7.8    381000.0  1.050000e+07  7.929106e+08    115.0
24   1982    Fantasy    6.2     20000.0  1.800000e+07  7.000000e+06    118.0
25   1982     Horror   47.6    654400.0  3.702500e+07  1.730389e+08    781.0
26   1982     Sci-Fi    6.2      1100.0  5.000000e+06  3.792188e+06     87.0
27   1982    Western    6.4      1500.0  1.100000e+07  1.736123e+06     90.0
28   1983     Action  127.0   1456100.0  3.228070e+08  1.022821e+09   2236.0
29   1983  Adventure   38.7    209300.0  1.170000e+08  2.118289e+08    802.0
..    ...        ...    ...         ...           ...           ...      ...
349  2016      Drama  117.8   2806000.0  4.594000e+08  1.368792e+09   2016.0
350  2016     Horror   48.0   1001000.0  9.100000e+07  7.825563e+08    823.0
351  2017     Action  250.7  10081000.0  4.358200e+09  1.510608e+10   4619.0
352  2017  Adventure   26.4    540000.0  6.570000e+07  6.031200e+08    434.0
353  2017  Animation   81.4   1124900.0  1.032500e+09  4.354454e+09   1264.0
354  2017  Biography  104.5   1259900.0  3.934000e+08  1.009730e+09   1811.0
355  2017     Comedy  133.0   1791000.0  4.483000e+08  1.472826e+09   2149.0
356  2017      Crime   52.9    781000.0  1.815000e+08  5.871441e+08    904.0
357  2017      Drama  169.3   1810600.0  5.556000e+08  1.656949e+09   2963.0
358  2017     Family    7.1    283000.0  1.600000e+08  1.264435e+09    129.0
359  2017     Horror   40.7   1676000.0  2.145000e+08  1.625070e+09    665.0
360  2018     Action  308.2   8051500.0  3.897000e+09  1.492840e+10   5531.0
361  2018  Adventure   37.2    716000.0  5.370000e+08  1.206474e+09    659.0
362  2018  Animation   68.3   1072000.0  8.435000e+08  4.019818e+09    999.0
363  2018  Biography  136.5   2027900.0  4.625000e+08  1.985195e+09   2320.0
364  2018     Comedy  149.9   1337000.0  6.562000e+08  2.256646e+09   2541.0
365  2018      Crime   67.2    719800.0  2.450000e+08  6.773376e+08   1254.0
366  2018      Drama  101.8   1654400.0  2.810800e+08  1.686406e+09   1674.0
367  2018    Fantasy    6.7     66000.0  2.000000e+07  7.942093e+06    152.0
368  2018     Horror   47.8    422000.0  7.070000e+07  7.903033e+08    865.0
369  2019     Action  215.8   5524095.0  3.241300e+09  1.258054e+10   3959.0
370  2019  Adventure   72.1    899000.0  7.742000e+08  2.695500e+09   1174.0
371  2019  Animation   86.3    967100.0  1.328100e+09  6.246113e+09   1275.0
372  2019  Biography   61.5    713200.0  2.930000e+08  3.973810e+08   1117.0
373  2019     Comedy  138.8   3102700.0  4.590000e+08  1.935893e+09   2361.0
374  2019      Crime    8.4   1000000.0  5.500000e+07  1.074427e+09    122.0
375  2019      Drama  105.1   1657100.0  3.823500e+08  1.920907e+09   1883.0
376  2019    Fantasy   12.4     26500.0  1.100000e+07  2.150964e+07    188.0
377  2019     Horror   29.3    487000.0  8.500000e+07  7.837949e+08    515.0
378  2019   Thriller    5.6     15000.0  8.000000e+06  3.659936e+07    102.0

[379 rows x 7 columns]
```



In [7]:
```python
#Answer: Revenue of each Genre fluctuates from 1980 to 2019, with Action generates highest revenue most of the year
s,
# followed by Animation and Comedy.

#Sub2c: How many movies produced in each kind of genres?
countGE = cleanedDF.groupby('genre')['year'].count().reset_index()
countGE = countGE.sort_values('year')
renamed_GE = countGE.rename(columns={'year':'count'})
print(renamed_GE)
```

```
        genre  count
14    Western      2
7      Family      4
11    Romance      5
12     Sci-Fi      6
13   Thriller      7
10    Mystery     17
8     Fantasy     41
9      Horror    251
2   Animation    277
3   Biography    311
1   Adventure    327
5       Crime    399
6       Drama    863
0      Action   1415
4      Comedy   1496
```

In [ ]:
```python
#Answer: The most produced movie genres are also the ones generating higher revenues.
#RESULT: Genre affects Revenue.
```