Movies project

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
import plotly.express as px
warnings.filterwarnings('ignore')
```

In [2]:
```
data = pd.read_csv("C:/Users/Allewaa/Documents/movies.csv")
```

Out[2]:

	name	rating	genre	year	released	score	votes	director	writer
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.7	1200000.0	Irvin Kershner	Leigh Brackett
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray
...	...	...	...	...	...	...	...	...	...
7663	More to Life	NaN	Drama	2020	October 23, 2020 (United States)	3.1	18.0	Joseph Ebanks	Joseph Ebanks
7664	Dream Round	NaN	Comedy	2020	February 7, 2020 (United States)	4.7	36.0	Dusty Dukatz	Lisa Huston
7665	Saving Mbango	NaN	Drama	2020	April 27, 2020 (Cameroon)	5.7	29.0	Nkanya Nkwai	Lynno Lovert
7666	It's Just Us	NaN	Drama	2020	October 1, 2020 (United States)	NaN	NaN	James Randall	James Randall
7667	Tee em el	NaN	Horror	2020	August 19, 2020 (United States)	5.7	7.0	Pereko Mosia	Pereko Mosia

7668 rows × 15 columns

In [3]:
```
for col in data.columns:
    pct_missing = np.mean(data[col].isnull())
```

```
name-0.0%
rating-0.010041731872717789%
genre-0.0%
year-0.0%
released-0.00026082420448617 63%
score-0.00039123630672926 45%
votes-0.00039123630672926 45%
director-0.0%
writer-0.00039123630672926 45%
star-0.00013041210224308815%
country-0.00039123630672926 45%
budget-0.2831246739697444%
gross-0.02464788732394366%
company-0.002217005738132499%
runtime-0.0005216484089723526%
```

In [4]:
```
```

Out[4]:
```
name         object
rating       object
genre        object
year          int64
released     object
score       float64
votes       float64
director     object
writer       object
star         object
country      object
budget      float64
gross       float64
company      object
runtime     float64
dtype: object
```

In [5]:
```
data1=data.drop(columns=['name','genre','director','writer','star','released'])
```

Out[5]:

	rating	year	score	votes	country	budget	gross	company	runtime
0	R	1980	8.4	927000.0	United Kingdom	19000000.0	46998772.0	Warner Bros.	146.0
1	R	1980	5.8	65000.0	United States	4500000.0	58853106.0	Columbia Pictures	104.0
2	PG	1980	8.7	1200000.0	United States	18000000.0	538375067.0	Lucasfilm	124.0
3	PG	1980	7.7	221000.0	United States	3500000.0	83453539.0	Paramount Pictures	88.0
4	R	1980	7.3	108000.0	United States	6000000.0	39846344.0	Orion Pictures	98.0
...	...	...	...	...	...	...	...	...	...
7663	NaN	2020	3.1	18.0	United States	7000.0	NaN	NaN	90.0
7664	NaN	2020	4.7	36.0	United States	NaN	NaN	Cactus Blue Entertainment	90.0
7665	NaN	2020	5.7	29.0	United States	58750.0	NaN	Embi Productions	NaN
7666	NaN	2020	NaN	NaN	United States	15000.0	NaN	NaN	120.0
7667	NaN	2020	5.7	7.0	South Africa	NaN	NaN	PK 65 Films	102.0

7668 rows × 9 columns

In [6]:
```
data1.dropna(inplace=True)
```

Out[6]:

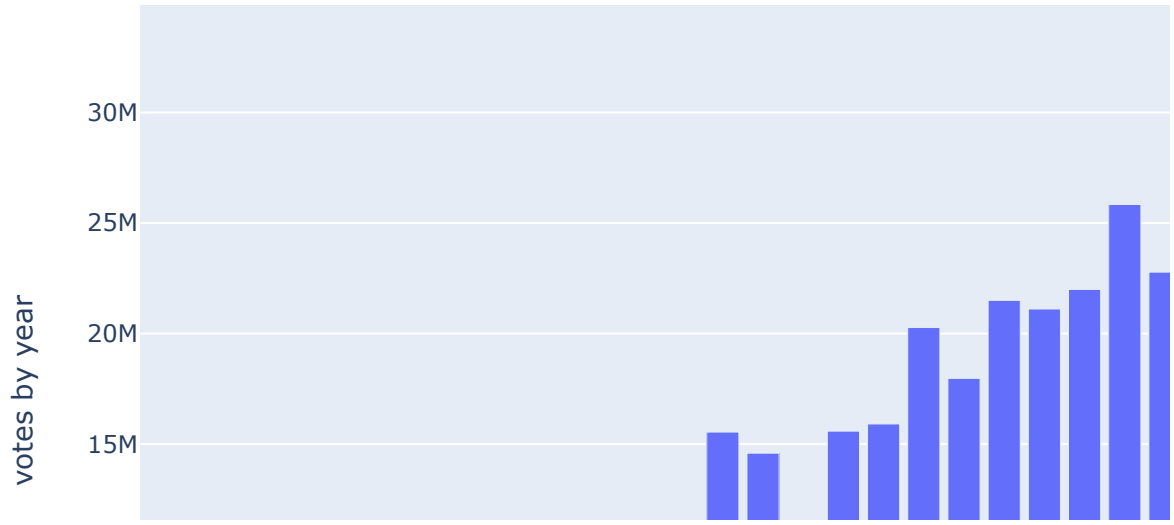
	rating	year	score	votes	country	budget	gross	company	runtime
0	R	1980	8.4	927000.0	United Kingdom	19000000.0	46998772.0	Warner Bros.	146.0
1	R	1980	5.8	65000.0	United States	4500000.0	58853106.0	Columbia Pictures	104.0
2	PG	1980	8.7	1200000.0	United States	18000000.0	538375067.0	Lucasfilm	124.0
3	PG	1980	7.7	221000.0	United States	3500000.0	83453539.0	Paramount Pictures	88.0
4	R	1980	7.3	108000.0	United States	6000000.0	39846344.0	Orion Pictures	98.0
...	...	...	...	...	...	...	...	...	...
7648	R	2020	6.6	140000.0	United States	90000000.0	426505244.0	Columbia Pictures	124.0
7649	PG	2020	6.5	102000.0	United States	85000000.0	319715683.0	Paramount Pictures	99.0
7650	PG	2020	5.6	53000.0	United States	175000000.0	245487753.0	Universal Pictures	101.0
7651	PG	2020	6.8	42000.0	Canada	135000000.0	111105497.0	20th Century Studios	100.0
7652	Not Rated	2020	6.8	3700.0	China	80000000.0	461421559.0	Beijing Diqi Yinxiang Entertainment	149.0

5421 rows × 9 columns

In [7]:
```
```

Out[7]:
```
Universal Pictures                        330
Columbia Pictures                         302
Warner Bros.                              298
Paramount Pictures                        279
Twentieth Century Fox                     209
                                         ...
Cinépix Film Properties (CFP)               1
Intermedia Films                            1
Dollface                                    1
Calimari Productions                        1
Beijing Diqi Yinxiang Entertainment         1
Name: company, Length: 1475, dtype: int64
```

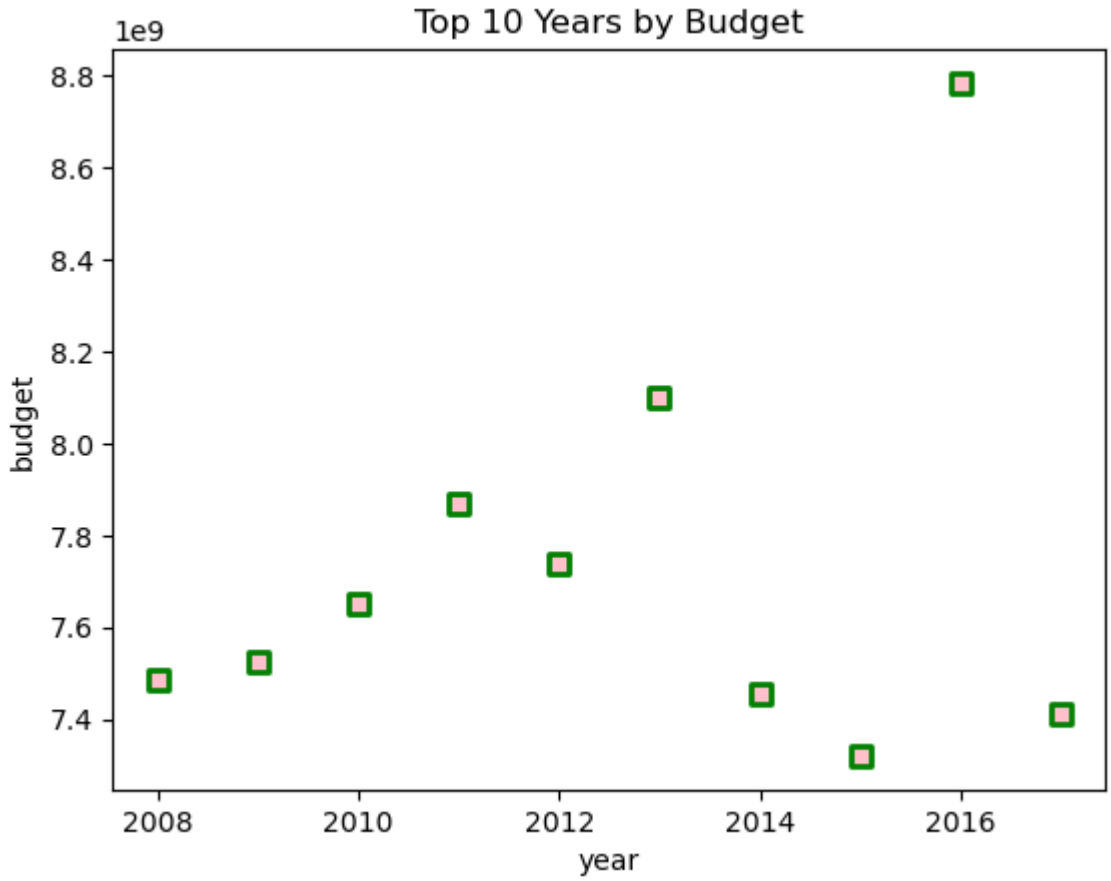
In [8]:
```
fig = px.bar(data_frame=data1,
     x=data.groupby('year').sum()['votes'].index,
    y=data.groupby('year').sum()['votes'].values
      )
fig.update_xaxes(title='year')
fig.update_yaxes(title='votes by year')
```



In [9]:

Out[9]:
```
year
2016    8.785000e+09
2013    8.101670e+09
2011    7.868730e+09
2012    7.737295e+09
2010    7.652750e+09
2009    7.522500e+09
2008    7.483830e+09
2014    7.454300e+09
2017    7.409700e+09
2015    7.319726e+09
Name: budget, dtype: float64
```

In [10]:
```
plt.scatter(data1.groupby('year')['budget'].sum().sort_values(ascending=False)
            linewidths = 2,
            marker ="s",
            edgecolor ="green",
            s = 50)
plt.xlabel('year')
plt.ylabel('budget')
plt.title('Top 10 Years by Budget')
plt.show()
```

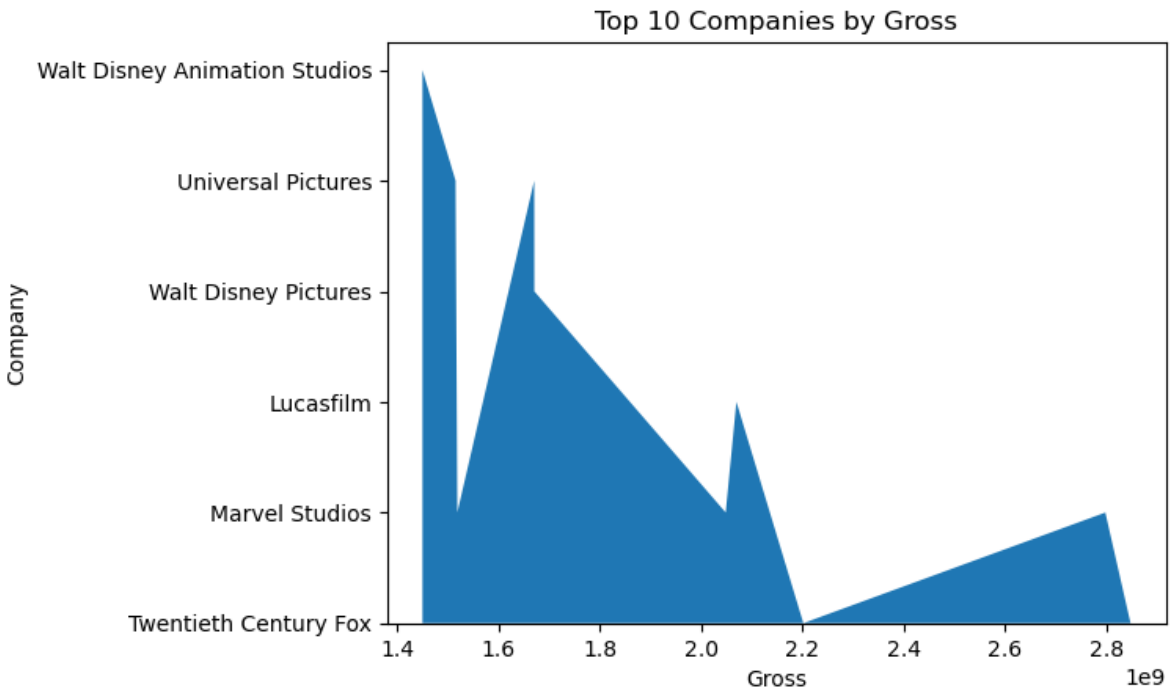


In [11]:
```
top_5_countries = data1.groupby('country')['gross'].sum().nlargest(5)
```

```
country
United States     4.707947e+11
United Kingdom    4.535171e+10
France            6.568533e+09
China             6.305919e+09
New Zealand       6.278355e+09
Name: gross, dtype: float64
```

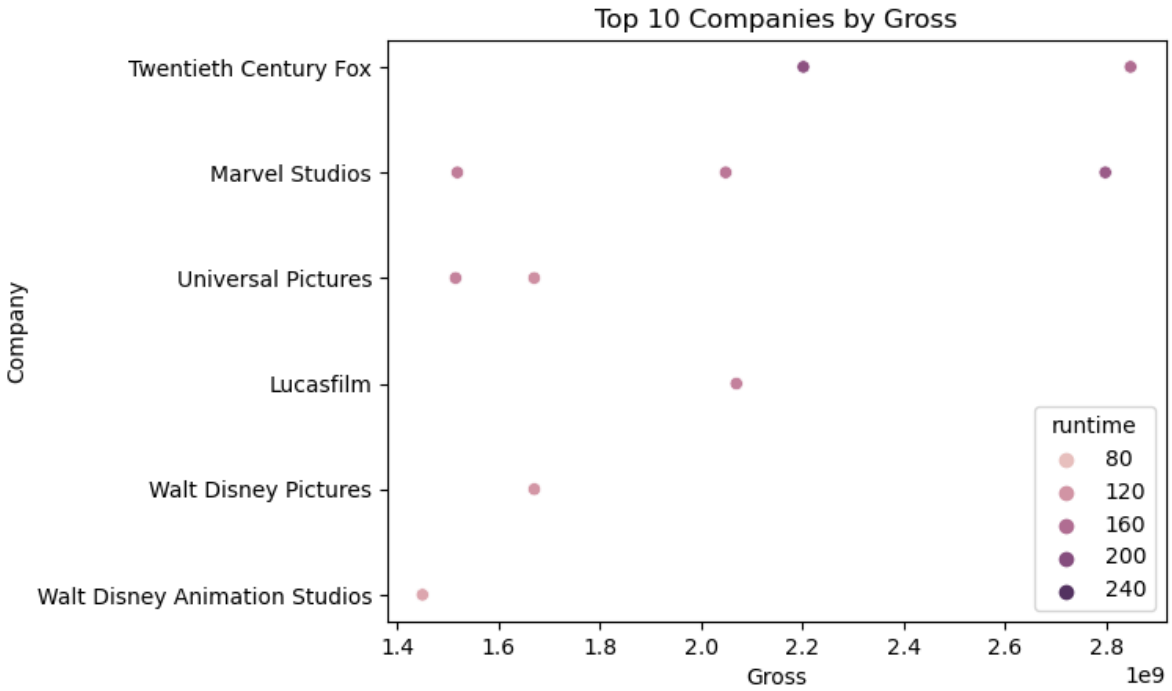
In [12]:

In [13]:
```
fig, ax = plt.subplots()
ax.stackplot(top_10['gross'], top_10['company'])
plt.xlabel('Gross')
plt.ylabel('Company')
plt.title('Top 10 Companies by Gross')
```

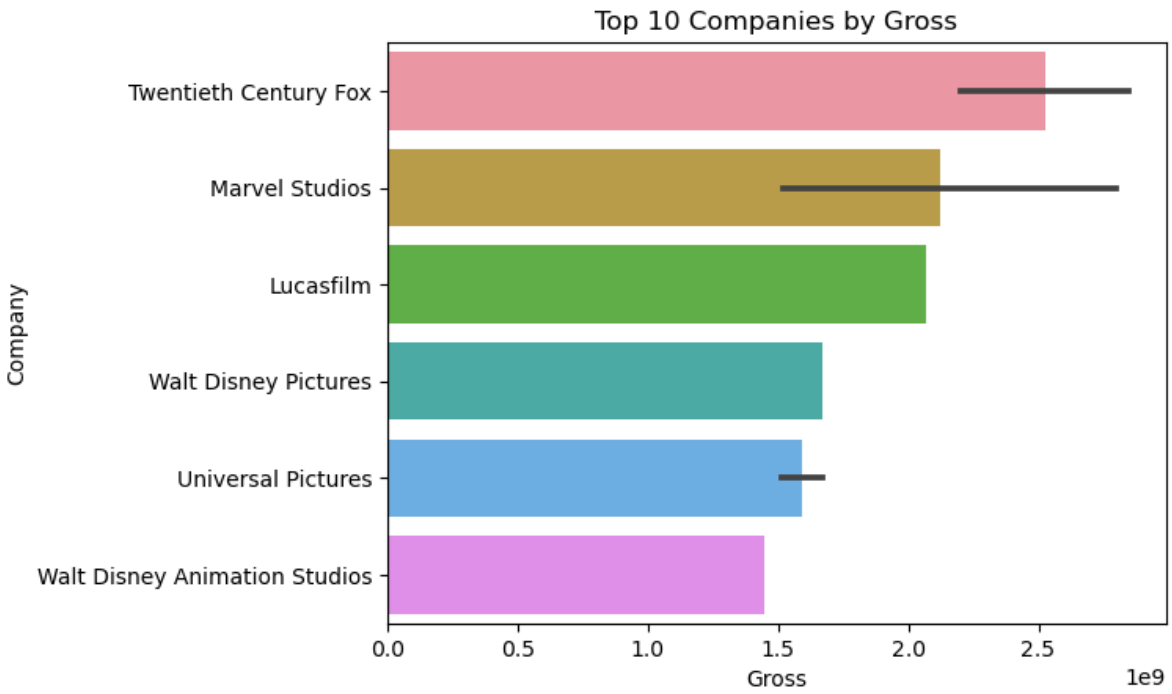


In [14]:

```
custom_palette = ['purple', 'lavender', 'violet']
sns.set_palette(custom_palette)
fig, ax = plt.subplots()
sns.scatterplot(data=data1,x=top_10['gross'],y=top_10['company'],hue='runtime'
plt.xlabel('Gross')
plt.ylabel('Company')
plt.title('Top 10 Companies by Gross')
```

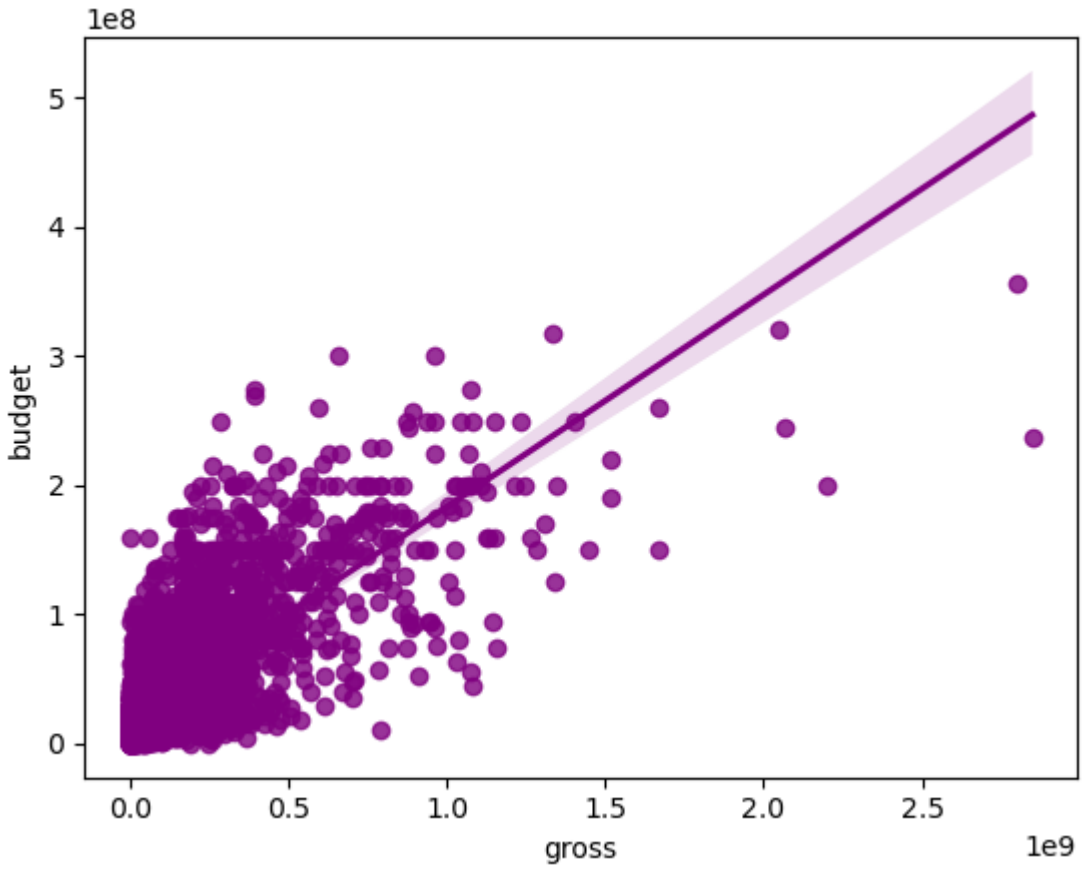


In [15]:
```
fig, ax = plt.subplots()
sns.barplot(data=data1,x=top_10['gross'],y=top_10['company'])
plt.xlabel('Gross')
plt.ylabel('Company')
plt.title('Top 10 Companies by Gross')
```



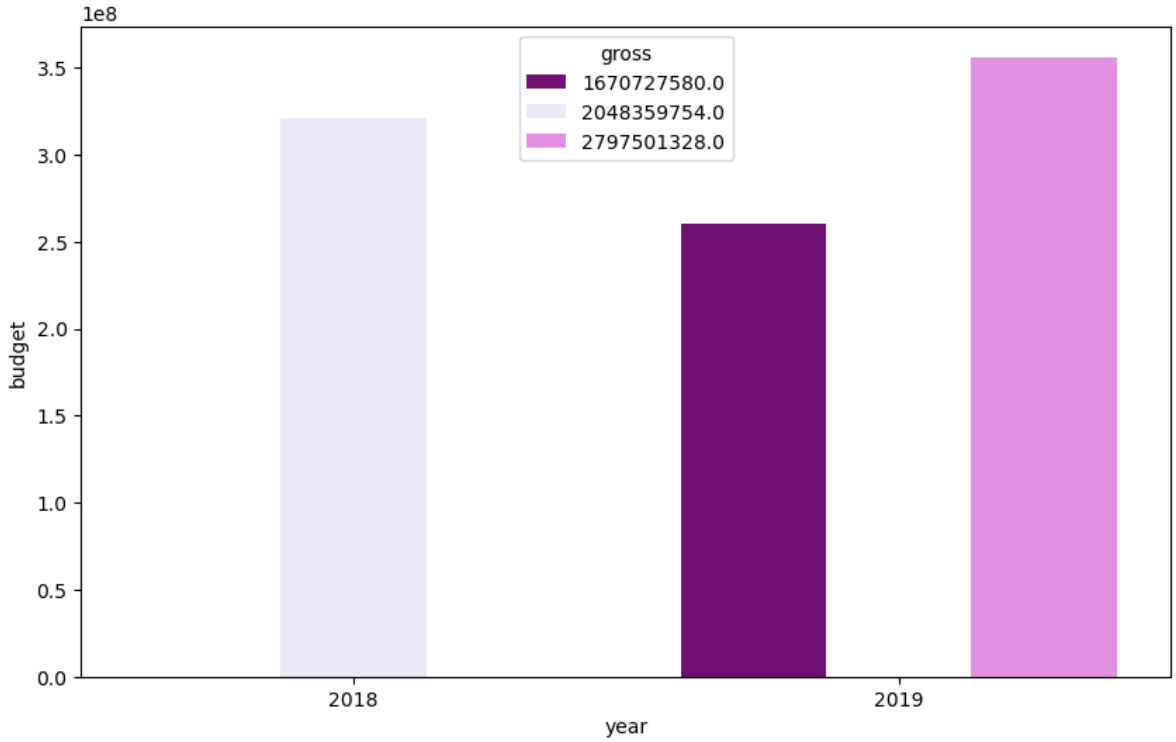
In [16]:

Out[16]: <Axes: xlabel='gross', ylabel='budget'>



In [17]:
```
budget_top_10 = data1['budget'].nlargest(10)
gross_top_10 = data1['gross'].nlargest(10)
filtered_data = data1[data1['budget'].isin(budget_top_10) & data1['gross'].isi
plt.figure(figsize=(10,6))
sns.barplot(data=filtered_data, x='year', y='budget', hue='gross')
```

Out[17]: <Axes: xlabel='year', ylabel='budget'>



In [ ]: