REBECA PRACTICE: DATA SCIENTIST SOLUTIONS

SOLUTION TO TASK 3

In [ ]:
```
# i) First I am interested in the statistics
# I want to know the mean and standard deviation
# of each of the features
movies_data_filtered.describe()
```

In [ ]:
```
# Correlation coefficients
movies_data_filtered.corr()
```

In [ ]:
```
# I decide to also drop the US_Gross for modelling,
# as it is almost identical to the worldwide gross
movies_data_filtered.drop(columns = ['US_Gross'], inplace = True)
```

In [ ]:
```
# Visualize
sns.pairplot(movies_data_filtered)
```

The worldwide gross and the US gross are highly correlated, which is trivial and expected. I also dropped the US Gross for this reason.

The production budget and the US gross are also quite correlated, which is also expected.

Suprisingly, the IMDB rating (=user rating) is very weakly correlated with the US gross. -> This is something to look into!