Practical 2

Jumping Rivers

DataFrames

For this set of questions we will use the movies data from the IMDB database. This data is contained in the course package `jrpyml`. To load the movies data as a `DataFrame` called movies you can use the following code:

```
import jrpyml.datasets as dat
movies = dat.load_movies()
```

1. Use the `.head()` method to inspect the top of the data. This can help give you a feel for what the data looks like and what variables are contained within the data.
2. How many films and variables are there in this dataset?
3. What is the mean and median film length?
4. What year is the oldest film in the data set from?
5. How long are the longest and shortest films?
6. Calculate the standard deviation of the ratings by using the **numpy** `std()` function.
7. Now calculate the standard deviation using the `DataFrame` member method `.std()`. Is there a difference? If so, why do you think that is?
8. How many action films are in the data? (There is a 1 in the Action column whenever a film belongs to that genre.)