Solutions 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.

```
print(movies.head())
```

```
##                       title  year  length  ...  Documentary  Romance  Short
## 0                         $  1971     121  ...            0        0      0
## 1          $1000 a Touchdown  1939      71  ...            0        0      0
## 2    $21 a Day Once a Month  1941       7  ...            0        0      1
## 3                   $40,000  1996      70  ...            0        0      0
## 4  $50,000 Climax Show, The  1975      71  ...            0        0      0
##
## [5 rows x 24 columns]
```

2. How many films and variables are there in this dataset?

```
print(movies.shape)
# 58788 films, 24 variables
```

```
## (58788, 24)
```

3. What is the mean and median film length?

```
# either
print(
  movies.length.mean()
)


# or
```

```
## 82.33787507654624
```

```
import numpy as np
print(
  np.mean(movies.length)
)
```

```
## 82.33787507654624
```

4. What year is the oldest film in the data set from?

```
print(
  movies.year.min()
)
```

```
## 1893
```

5. How long are the longest and shortest films?

```
# I have used different syntax here to highlight there is more than
# one way to extract columns from a data frame.
print(
  movies['length'].max()
)
```

```
## 5220
```

```
print(
  movies.loc[ : , 'length'].min()
)
```

```
## 1
```

6. Calculate the standard deviation of the ratings by using the **numpy** `std()` function.

```
print(
  np.std(movies.rating)
)
```

```
## 1.553017591358266
```

7. Now calculate the standard deviation using the `DataFrame` member method `.std()`. Is there a difference? If so, why do you think that is?

```
print(
  movies.rating.std()
)
# There is a small difference. This is because np.std calculates
# population standard deviation by default wheras DataFrame.std
# calculates sample standard deviation.
```

```
## 1.5530308001543176
```

8. How many action films are in the data? (There is a 1 in the Action column whenever a film belongs to that genre.)

```
print(
  movies.Action.sum()
)
```

```
## 4688
```