Programming Assignment 6

Getting started

Review class handouts and examples, work on the reading and practice assignments posted on the course schedule. This assignment is designed to practice data manipulation with Pandas and plotting with Matplotlib.

Programming Project: Plotting **worth: 20 points**

Create plots visualizing movie data.

Data and program overview

In this assignment you will be working with data on movies and people's ratings of these movies. The task will be to create visualizations of the movie data.

The following data will be provided using csv files:

- A table with movie information (*IMDB.csv*); we will call this the *movie data*.
- A table documenting the movie data columns (*IMDB-dataDict.csv*).

The files are supplied in a zip file, which will create a *data* subfolder, when unpacked.

Your program should produce two plots when it is run. The plots should follow the specification provided below. Note that creating the second plot will require your independent exploration of Matplotlib for generating a bar graph.

Plot 1: Ratings by age group

This plot displays the average ratings by voters in the four age groups identified in the data for a set of movies selected by the user. A sample is presented in the first figure in the sample interaction. The four ***age groups*** are: voters under 18 (*VotesU18* column), between 18 and 29 years old, inclusively (*Votes1829*), between 30 and 44, inclusively (*Votes3044*), and voters 45 and older (*Votes45A*).

To generate the plot, the user will first need to specify which movies data to include, as described below:

Step 1. The program should ask how many movies should be included. Assume the user will enter a valid number between 1 and the total number of movies in the data file.

Step 2. The program should let the user pick movies by executing the following procedure for each selection:

a. Ask the user to specify a title keyword and identify the movies that have the keyword in the title (the search should be case insensitive).

b. If there is a single movie with the provided keyword (as shown in the interaction, while selecting the movie for Plot 2), that movie becomes the selected movie.

   Otherwise, if there is more than one matching movie, all matching titles should be displayed, along with a 1-based number identifying each title (see interaction for Plot 1). The user should be invited to select one of these titles by specifying a number (you may assume user input will be valid here). The corresponding movie becomes the selected movie.

   If there are no titles containing the specified keyword, the user should be asked to provide another keyword and steps 2a and 2b must be repeated.

Step 3. Save the plot as *plot1.jpg* file using `plt.savefig()` function.

Other Plot 1 Requirements: different movie data must be shown in different colors, displaying a marked point for each of the four values, and a line, connecting the points. Axes must be clearly labeled and the movie title and genre must be displayed near the first point of the respective graph (see figure). The plot must include a dashed grid and a title. See the sample interaction for an illustration.

Plot 2: Percentage of raters within different gender-age groups

The movie data contains eight columns specifying the number of female and male voters within each of the age groups identified in the description of Plot 1. We will refer to these columns as ***age-gender*** columns (they are titled *CVotesU18M, CVotesU18F* and so on). **Note:** the total sum of values in the age-gender columns does not equal the value supplied by the *TotalVotes* column; perhaps this is because the *TotalVotes* count also includes people who did not identify their age or gender, and thus are not accounted for by the age-gender columns. We will refer to the sum of values within the age-gender columns for each row as ***all votes*** for the movie.

Plot 2 displays a bar graph, showing what percentage of all votes for a movie are contributed by the specific age-gender group. As with Plot 1, the user must be allowed to select a movie, using the sample procedure as specified in Step 2 of Plot 1 description, except run for a single movie. The plot must be saved by your program as *plot2.jpg*.

Other Plot 2 Requirements: Bars for different genders must be shown in different colors, with the corresponding percentage value displayed clearly on top of the bar. Axes must be clearly labeled as shown, and the values along the x axis must be slanted by 30 degrees. Plot title must include the title of the movie. Refer to the sample interaction for an illustration.

Required Functions

Include `main()`, function `pickMovieWithKeyword()` that will run the movie selection procedure described in Step 2 of Plot 1 description. Functions `plot1()` and `plot2()` should each generate and save the respective plot. Pick function parameters and return values as you see fit, and define other functions as needed.

General Requirements

- You can assume that the provided file will have all of the columns involved in the required computations, but the number of records and order of columns may be different.
- Your program should have no code outside of function definitions, except for a single call to `main()` and definitions described in the next bullet.
- In order to make the code easier to modify for a different set of column names, define global variables that store the names of columns and use them throughout, e.g. `TITLE = 'Title'`.
- All file related operations must use device-independent handling of paths (use `os.getcwd()` and `os.path.join()` functions to create paths, instead of hardcoding them).

Submission and Grading

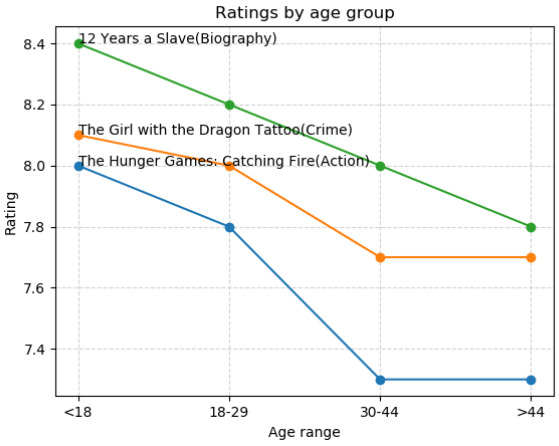
Submit your code along with two image files that your program will generate for the input data contained in the sample interaction. Grading will be based on the accuracy (conforming to all the requirements and format of the interaction), generality of code and the appropriate use of pandas/numpy/matplotlib resources (data structures and functions). Each plot will be worth 9 points and two points will be awarded for programming style.

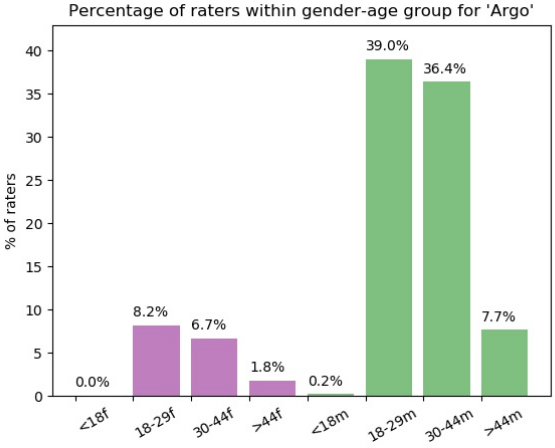
The sample interaction that follows demonstrates one full execution of the program with the generated plots. Note that the location of the data file must be requested from the user first.

Sample interaction

```
Please enter the name of the subfolder with the data file: data

Plot1:  ratings by age group
How many of the 118 movies would you like to consider? 3
Select 3 movies

Enter movie keyword: game
Which of the following movies would you like to pick (enter number)
        1 The Hunger Games: Catching Fire
        2 The Imitation Game
enter a number: 1
Movie #1: 'The Hunger Games: Catching Fire'

Enter movie keyword: girl
Which of the following movies would you like to pick (enter number)
        1 Gone Girl
        2 Me and Earl and the Dying Girl
        3 The Girl with the Dragon Tattoo
enter a number: 3
Movie #2: 'The Girl with the Dragon Tattoo'

Enter movie keyword: 12
Which of the following movies would you like to pick (enter number)
        1 12 Years a Slave
        2 127 Hours
        3 Short Term 12
Enter a number: 1
Movie #3: '12 Years a Slave'
```

```
Plot2: Percentage of raters within gender-age. Select a movie:

Enter movie keyword: argo
Movie #1: 'Argo'
```





Created by Tamara Babaian on December 1, 2018