Scientific Programming in Python Project

Dataset Overview:

The attached IMDB movie dataset (*movie_meatadata.csv*) contains the details of over 5043 movies scraped from the IMDB movie website. The dataset attempted to collect 28 features describing each movie.

1. "movie_title"
2. "color"
3. "num_critic_for_reviews"
4. "movie_facebook_likes"
5. "duration"
6. "director_name"
7. "director_facebook_likes"
8. "actor_3_name"
9. "actor_3_facebook_likes"
10. "actor_2_name"
11. "actor_2_facebook_likes"
12. "actor_1_name"
13. "actor_1_facebook_likes"
14. "gross"
15. "genres"
16. "num_voted_users"
17. "cast_total_facebook_likes"
18. "facenumber_in_poster"
19. "plot_keywords"
20. "movie_imdb_link"
21. "num_user_for_reviews"
22. "language"
23. "country"
24. "content_rating"
25. "budget"
26. "title_year"
27. "imdb_score"
28. "aspect_ratio"

Project Specification.

The objective of this project is to produce an application that allows the user to explore some of the most interesting aspects of the imdb dataset. Please note that where possible you should use **Pandas** a as a means of analysing the data. Where requested please incorporate visualisation as a method of illustrating your results. Your report should comprehensively address the following questions about the dataset. Please be aware that the dataset does contain some missing values.

When you run your program it should display the following menu:

> Please select one of the following options:
>
> 1. Most successful directors or actors
> 2. Film comparison
> 3. Analyse the distribution of gross earnings
> 4. Self-Directing
> 5. Earnings and IMDB scores
> 6. Exit

1. **Menu Option 1 – Most successful directors or actors**

   When the user selects the first option ("Most successful directors or actors") they should be presented with the following menu.

   i. Top Directors
   ii. Top Actors

   If the user selects "Top Directors" they will be asked to enter an integer value specifying the number of directors they want to return. If, for example, the user enters the value 10 then the ten most successful directors (based on gross film earnings) will be outputted.

   If the user selects "Top Actors" they will similarly be asked to enter the number of actors they wish to display. If, for example, the user enter the value 8, they will be shown the top eight most successful actors (based on gross film earnings).

   In each case the information should be conveyed using a horizontal bar graph.

   You should provide some basic error checking in your code. You should make sure that the user cannot enter a negative value or cannot select an invalid integer value for the number of directors/actors that is greater than the number of directors/actors in the dataset. If the user does select an invalid value, your code should output an error message and ask the user to re-enter a valid value. Your code should continue to do this until the user enters a valid value.

Please note that if the user selects a very large integer value (specifying the number of directors/actors) the resulting graph may look quite cluttered, which would typically necessitate resizing your image. Your solution for this question does not need to deal with this issue.

Your report should contain one graph generated for the actors and one for the directors.

[16 Marks]

2. Menu Option 2 – Film Comparison

If the user selects "Film Comparison" your code should ask the user to enter the name of two films from the dataset. Your code should then provide basic error checking. If the user enters the name of a film not contained in the dataset it should repeatedly ask the user to enter a valid film name.

Once the user has entered two valid film names they should be presented with the following options.

i. IMDB Scores
ii. Gross Earning
iii. Movie Facebook Like

The user will select one of these options and your application will display a simple bar graph comparing the two films using the option selected. For example, if the user selects the first option then then a bar graph containing the IMDB scores for each film should be generated.

Your report should contain one bar graph generated for each of the above options.

[16 Marks]

3. Menu Option 3 – Gross earnings analysis

If the user selects "Analyse the distribution of gross earnings" then the program should ask the user for a start year and then for an end year for the analysis. Using a line graph it should then display the min, average and max gross earnings achieved by the films for each year between the start year and end year inclusive. The line graph should have three lines, one for max, one for average and one for min.

Your report should contain the line graph that is outputted when the user enters a specific start and end year.

[16 Marks]

4. Menu Option 4 – Self-Directing

If the user selects “Self-Directing” the program should present a listing of all actors (i.e. those which are also listen under either under actor_1 or actor_2 or actor_3 categories included in the dataset) who directed at least one of their own movies. The list should not contain duplicates. Your program should also output the top five directors who have self-directed most often, together with the number of times they directed themselves.

[16 Marks]

5. Menu Option 5 – Earnings and IMDB scores

We are interested in building a model that will predict the IMDB score of each film with a reasonable level of accuracy. To assist in the process you have been asked to examine the relationship between the numerical IMDB scores and other numerical columns in your dataset. Generate graphs that will best help illustrate the relationship or lack of relationship between the IMDB column and the other features.

Your code should output the graphs that best supports your analysis. These graphs should also be included in your report and should be accompanied by a written account of your findings.

The following should be of use in helping you to solve this problem.

The code below will take each string element in a specific dataframe column and split the string into a list (using ‘|’ as a delimiter). It returns a Series object where each element is now a list of strings.

```
test = df['column_name’].str.split('|')
```

The code below will check each String element in a column to determine if it contains, even in part, the word “Test”. The code returns a Series object containing Boolean values (which can then be used for array based indexing).

```
result = df['column_name’].str.contains("Test")
```

[16 Marks]

6. Menu Option 6 – Exit

If the user selects an option 1-5 then your program should display the associated output and will subsequently display the main menu again.

If the user selects option 6 the application should exit.

The final 20 marks will be going on the report as well as general coding, imports, functions usage, comments, data loading, handling of missing data, variable naming, menu.

Note: You will be submitting both a report (.pdf format) as well as a python file. It is very important to make sure you include a clear reference between the results presented in the report and the python file. I should be able to easily find the code that was used to generate the results presented in the report.

Academic Integrity

This is an individual assignment. The work you submit must be your own. In no way, shape or form should you submit work as if it were your own when some or all of it is not. Any online source that is used must be cited, and a full citation given, e.g. do NOT give ``stackoverflow", but the full citation, e.g. https://meta.stackoverflow.com/questions/339152/plagiarism-and-using-copying-code-from-stack-overflow-and-submitting-it-in-an-as. If you are unsure on whether something should be cited, general rule of thumb is to err on the side of caution and include the citation. You can also ask me via email

Given how much freedom there is in the assignment, everybody's work will be different. It will be obvious if there is collusion. All parties to collusion will be penalized.

Deliberate plagiarism: You must not plagiarise the programs, results, writings or other efforts of another student or any other third-party. Plagiarism will meet with severe penalties, which can include exclusion from the University.

Your report will be checked for signs of collusion, plagiarism, falsification and fabrication. You may be called to discuss your submission and implementation with me and this will inform the grading, any penalties and any disciplinary actions.

You may, of course, ask me questions. I may share questions and answers with the class, if I feel they are general matters, for example, of clarification. But I will also discuss with you questions that relate to your own Python, your experiments and your reading and, in the interest of giving you proper credit for your endeavours, these will not be shared with the class.

A declaration must be included in your submission to state that the work included is all your own work.

Submission Instructions:

1. Submit your report (.pdf format) on Canvas (via the *submit pdf for turnitin* link in the Assignments unit) before **20:00 on Sunday Dec 19th**. Submit your solution python file (.py file) via the *submit code for project* link in the Assignments unit.
   Note the late submission penalty for the first two weeks will be waived. However, as per CIT regulations, submitting later than 14 days after the due date will result in a 100% penalty applied.
   So you may submit up until January 2nd penalty free. (This is a hard dealine, no extensions will be possible.) Note also this is a one time submission
   Your file names MUST be *Surname_StudentNo*_Report.pdf and *Surname_StudentNo*_Project.py.

2. Once you have submitted your files you should verify that you have correctly uploaded them. **It is your responsibility to make sure you upload the correct files**.

3. Please make sure you **fully comment** your code.

4. Please also put your student name and number as comments at the top of your python file.