Dataset Overview:

The IMDB movie dataset hosted by Kaggle 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 and report 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** as a means of analysing the data. Where requested please incorporate visualisation as a method of illustrating your results. Your final submission should include code and a short report. Your report will provide a summary of your findings, mainly showing the graphical output generated by your code. Please be aware that the dataset does contain missing values. Depending on how you implement the application you may have to remove some rows that contain 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. Genre Analysis
    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.

[20 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.

[20 Marks]

3. Menu Option 3 – Analyse the distribution of gross earnings

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.

[20 Marks]

4. Menu Option 4 – Genre Analysis

If the user selects “Genre Analysis” the program should present a listing of all unique genres included in the dataset. The user should then be asked to input a specific genre and your program will output the mean IMDB score of all films within that genre.

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")
```

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.

[20 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.

Guidelines and Submission Instructions:

1. The project is worth 40% of your overall module grade.

2. You will produce a **jupyter notebook file** containing all your code and output for each of the operations outlined the project.

3. Upload your solution jupyter file to Canvas before on or before 20 **December, 2021.**

4. Go to the Assignment Project -I in Canvas to upload your file.

5. 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.

6. Please make sure you **fully comment your code**. You should clearly be explaining the operation of important lines of code.

7. Please put your student name and number as comments at the top of your file.