Programming for Data Analytic
SOFT8032
First Assessment

November 2021

1 Second Assessment. First Project

This project contributes 30% in your final mark. This is an individual project and has to be all done by yourself. You may be called for a zoom meeting to explain different parts of your submission, if needed.

Any question regarding the project should be communicated with farshad.toosi@mtu.ie or Canvas message.

1.1 Dataset Overview

For this project we are going to perform a number analytic tasks on the **movie_methadata.csv** file. Each row of the dataset contains 28 pieces of information about a movie. Some of those include:

- color
- director_name
- duration
- actor_2_name
- gross
- genres
- actor_1name
- movie_title
- num_voted_users
- actor_3_name
- plot_keywords
- num_user_for_reviews

- language
- country
- content_rating
- budget
- title_year
- imdb_score
- aspect_ratio

1.2 Project Specification

The objective of this project is to provide an insight into some of the relationships and trends that exist within this dataset. Please note that you can only use Pandas, Numpy and Matplotlib as means of analysing data within this dataset. Please download the template file and complete your assignment in the template file. Make sure to type your name, ID and your cohort in the file. Once finished, rename the file as follows: SR12345678 where 12345678 is your Student ID. Please only submit the .py file and nothing else.

1.2.1 Details of the project specified as a number of tasks

1. Complete the **Task1** function in the template file and use groupby in pandas to return the name of those actors/actresses that have played at least in two Black and White movies. Actors/actresses' name can be found in a column with header: actor_1_name.

   Data cleansing: Some values in the color column have extra space at the beginning, remove the spaces from the beginning and the end of each value in the cell.

2. Complete the **Task2** function in the template file and extract those country names that have at least one movie that is longer than 150 and the language is not English.

3. Complete the **Task3** function in the template file.

   Calculate the the amount of income that is gained in each year. Use an appropriate visualization technique and show the yearly income trend from the earliest year until the latest year in the file. Each movie has a column called gross that can be assumed as the income.

   Data Cleansing: Some movies do not have a known income (gross); fill the empty cells in gross column with the average value of the gross column.

4. Complete the **Task4** function in the template file.

   Each movie has a column called year (titleyear) that indicates the year that the movie was made.

   For each year, extract the percentage of those movies that their income (gross) is higher than the doubled of the budget.

   Only consider the years after 1989.

   Use an appropriate visualization technique to visualize this information. See Figure 1.

Data Cleansing: Remove all the rows from the data frame where they have an empty cell in gross or budget.

5. Complete the **Task5** function in the template file.

   Use an appropriate visualization technique that visually depicts the number of movies at each country. Display the percentage of the number of movies for each country on the visualization.

   The USA and the UK and those countries with less then 30 movies should not be considered for this task.

6. Complete the **Task6** function in the template file.

   Each movie has a length (duration). Apply an appropriate visualization technique and visually depict what durations (movie lengths) are more common and popular among all movies in the file.

   Use comment and indicate the common and popular movie lengths.

   Data Cleansing: Some movies do not have a known duration, those movies (rows) should be ignored for this task.

Note1: You will need to provide proper and complete set of visualization features (e.g., legend, axis label etc, if applicable) for those tasks that require a visualization.

Note2: Apply the most efficient approaches for the above tasks. E.g., avoid loops if it is possible.

Note3: Please only apply the Pandas, Numpy and MatPlotlib functions that have been discussed during the lectures and labs.

2 Rubric

This rubric is subject to change.

1. Correct task implementation (visualization) with meaningful labels, annotation (if needed) legend, comment and etc. (100%)

2. Correct task implementation (visualization) with less/minimum meaningful labels, annotation, comment and etc. (80%)

3. Partly correct task implementation (visualization) with partly meaningful labels, annotation (if needed) legend, comment and etc. (50%)

4. Tasks are attempted towards the correct solution but no result is obtained. (20%)

5. Wrong task implementation (visualization). (0%)

3 Submission

There is template file (Python file) provided for you on Canvas. You are required to complete your project in that file. Each task needs to be implemented as a separated function with one line interpretation as a comment below the function.

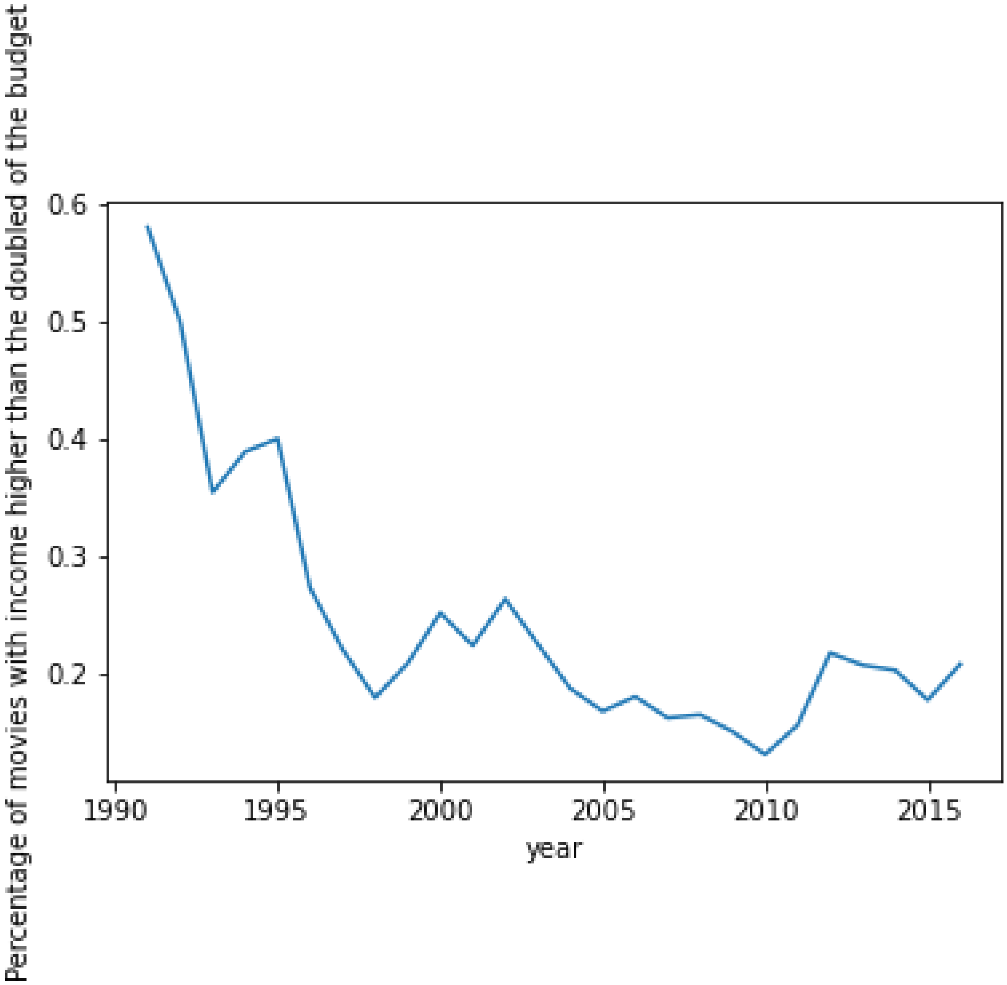

Figure 1: Income

Please write your name, student ID and your course name (Cohort) as a comment in the designated area in the provided template python file.

The template file should be re-named at the end using your student ID followed by letter SR, for example if your student ID is: 1234567 the the python file should be named: SR1234567.py

The deadline for this project is 28th of November 2021 at 23:59. One week late submission is accepted with 10 marks penalty and the deadline for the late submission is 5th of December at 23:59. Two weeks submission with 20 marks penalty is also accepted and the deadline is 12th December at 23:59.

Please submit your project via Canvas as one single python file ONLY.