Programming for Data Analytic
SOFT8032
Second Examination

October 2023

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 are not allowed to disclose your code to anyone else.

You will be called to present your submission.

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 analytical tasks on the **movies.csv** and **main_genre.csv** files.

1.2 Project Specification

The objective of this project is mainly to provide an insight into the underlying pattern of the dataset in **movies.csv** such as statistical details of different features and etc. Please perform the following tasks:

1. How many main-genres exist in **movies.csv**, and which one is the most popular, and which one is the least popular?

   How the results should be displayed? In three different lines print 1) The total number of unique main-Genres, 2) The most popular main-genre, 3) The least popular main-genre. Additionally, display the top 8 popular genres using an appropriate visualization technique. Do not print or report anything else.

2. What is the most and least common genre? Note that there are two columns related to genres: 'genre' and 'main_genre.' For this task, the 'genre' attribute is the focus, not 'main_genre.

   How the results should be displayed? Only print the most and the least common genres and nothing else.

3. Apply an appropriate visualization technique to display the outliers in movie duration (Runtime). Print the names of the movies for which the duration is considered an outlier.

   How the results should be displayed? Only print the Title of the movies that belong to the outliers. Also display the visualization, no need to save the visualization in any file.

4. Apply an appropriate visualization technique to analyze the relationship between the 'number of votes' and the 'rating'. Report if there are any null values in either of the mentioned attributes. If any null values are found, they should be filled with the average of the existing values for each attribute prior to the visualization. Note the difference scale of the two attributes, 'number of votes' and the 'rating'.

   How the results should be displayed? Write a short comment below this task's function and explain the the existence of null values in those attributes/columns. Also display the figure. No need to save the visualization in any file.

5. The **main_genre.csv** file contains various main genres (see Column headers). Each main genre (column header) is associated with multiple terms. For instance, **fantasy** is associated with *Imagination*, *Reverie*, *Dream*, *Delusion*, and more. Please open the file to view its contents.

   Your task is to read the **main_genre.csv** file and, for each main-genre, select a group of movies in the (Movies.csv) file whose synopses contain one or more terms associated with the given main genre in **main_genre.csv**. After forming this group, further analysis is required to determine which *main_genre* in Movies.csv in that group has the highest frequency.

   Please note that the words in the Synopsis need to be lower-cased and cleansed by removing the following noises: [',', "', '.', '-']. The terms from the main-Genres file should also be lower-cased.

   How the results should be displayed? There are 8 main-Genres in the **main_genre.csv**. For each main-Genre, print the main-Genre (the one in **main_genre.csv**, column header) itself, and next to it, print the most frequent *main_genre* related to the group of movies from Movies.csv. For example, the 'fantasy' main-Genre in **main_genre.csv** appears in many movie synopses where the *main_genre* of those movies is also 'fantasy.' Do not print or output anything else. Only 8 main-Genres, and for each main-Genre, the *main_genre* with the highest frequency.

6. Apply one analytical task of your choice. Make sure the chosen task is useful for people in this industry and also complex enough. Use comment section and explain the idea of your task.

   How the results should be displayed? Please comment below this task's function and explain the expected output. Do not generate additional output.

Note that all visualization plots need to have proper labels and annotations if needed. Lack of visualization features attracts penalty.

Efficiency is crucial for this project, as a more efficient code can expedite the process and reduce the likelihood of errors in the analysis. For example, avoid unnecessary loops and hard coded values, and so forth.

1.3 Submission and Deadline

Please use the python file template that is provided for you and complete your project in that file. Each task needs to be implemented as a separated function with interpretation as a comment below the function. You may define extra functions if needed.

You will be required to present your code to me during the labs in week 10. The failure to explain your code during the lab would attract penalty.

Please write your name and student ID as a comment in the designated area in the provided template python file.

The deadline for this project is 17th Nov 2023 at 23:59. One-week late submission with 10 marks penalty would be accepted and the deadline would be 24th Nov 2023 at 23:59.

Please note that you are required to present your project to me individually, not to the entire class. You should also be prepared to answer any questions I might have about your code, pandas, numpy, and matplotlib during the lab. The actual presentation will take place during the labs in week 10.

Any question about this project should be communicated with Farshad Ghassemi Toosi farshad.toosi@mtu.ie or via Canvas.

1.4 Rubric

This rubric is subject to change.

1. Correct task implementation (Data extraction, data cleaning, correct visualization if needed etc.). (100%)
2. Relatively correct task implementation (Data extraction, data cleaning, correct visualization if needed etc.). (70%)
3. Partly correct task implementation (Data extraction, data cleaning, correct visualization if needed etc.). (40%)
4. Wrong task implementation. (0%)