For this project we are going to perform a number analytical tasks on the movies.csv and main_genre.csv files.

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.
1

3. Apply an appropriate visualization technique to display the outliers in movie duration (Run-time). 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.