KnowlEDGE Streams

Intermediate Data Science & Machine Learning

Mini Project

Topics Covered

- Numpy
- Pandas

Project Overview:

You'll be provided with data on the 100 highest-rated movies from the last decade and details about the films, their actors, and the online voters who rated them. Your task is to uncover compelling insights about these movies and their audience using Python.

What we expect:

This is a compulsory individual mini project wherein you will perform EDA on a movie dataset, write Python code to explore the data, gain insights into the movies, actors, votes, ratings and collections, and submit the code on a jupyter notebook file.

- The dataset will be shared on the WhatsApp group.

On the next page, you will find instructions along with the questions that you need to do in order to complete the assigned work.

Best of Luck!

Questions

Task 1: Reading the DATA

1.1. Read the “Movies” Data.

1.2. Inspect the Dataframe for dimensions, null-values, and summary of different numeric columns.

Task 2: DATA Analysis

2.1. Reduce the digits in “budget” and “gross” for readability **(See notebook for details)**

2.2.

- Create a new column called profit which contains the difference of “**gross**” and “**budget**”
- **Sort** the data frame using the profit column as a reference.
- **Extract** the top ten profiting movies in descending order and store them in a new data frame named “top10”
- **Record** your observations
- Extract the movies with a **negative** profit and store them in a new data frame named “**neg_profit**”

2.3. You might have noticed the column “**MetaCritic**” in this dataset. Second, you also have another column “**IMDb_rating**” which tells you the IMDb rating of a movie. Your task is to find out the **highest rated** movies which have been **liked** by critics and audiences alike.

1. Firstly you will notice that the MetaCritic score is on a scale of 100 whereas the IMDb_rating is on a scale of 10. First convert the MetaCritic column to a scale of 10.
2. Now, you have to find out the movies which have been liked by both critics and audiences alike and also have a high rating overall. **(See notebook for details)**

2.4. Find out the **top 5** popular trios, and output their names in a list. **(See notebook for details)**

2.5. In the previous subtask you found the popular trio based on the total number of Facebook likes. Let's add a small condition to it and make sure that all three actors are **popular**. The

condition is none of the three actors' Facebook likes should be less than half of the other two. For example, the following is a valid combo:

actor_1_facebook_likes: 70000
actor_2_facebook_likes: 40000
actor_3_facebook_likes: 50000
But the below one is not:
actor_1_facebook_likes: 70000
actor_2_facebook_likes: 40000
actor_3_facebook_likes: 30000

since in this case, actor_3_facebook_likes is 30000, which is less than half of actor_1_facebook_likes. You can do a manual inspection of the top 5 popular trios you have found in the previous subtask and check how many of those trios satisfy this condition. Also, which is the most popular trio after applying the condition above? Write your answers in the markdown cell provided in the jupyter file. **(See notebook for details)**

2.6. Although R rated movies are restricted movies for the under 18 age group, still there are vote counts from that age group. Among all the R rated movies that have been voted by the under-18 age group, find the top 10 movies that have the highest number of votes i.e., CVotesU18 from the movies dataframe. Store these in a dataframe named **PopularR**.

Task 3: Demographic Analysis

3.1. Combine the data frames by genre **(See notebook for details)**

With the above subtask, your assignment is over. In your free time, do explore the dataset further on your own and see what kind of other insights you can get across various other columns.