Project Description

My project will involve me working with a dataset from Kaggle. The dataset includes the most popular 10,000 movies based on TMDB ratings. A link to this dataset will be included in this document. Further information may be found from accessing the link below.

The reason why I chose this dataset was because I believed it would be straightforward to analyze and make sense of the various columns, which turned out to be true. More importantly, this kind of dataset is important for consumer analytics. That is, analyzing datasets like this may help businesses and other individuals make sense of consumer demand, especially in this case, as it relates to the film industry. Thinking more broadly, consumer-related data involves data that represents their opinions about products, services and systems. It is necessary to be analyze and evaluate these sorts of datasets so that future products and services can be developed and advertised in potentially more optimal ways. It is critically important to be aware of what consumers enjoy as consumers are more likely to buy products, use services and make subscriptions when they are interacting with enjoyable content. Moreover, working with these sorts of straightforward and accessible datasets are an interesting and non-stressful sort of way for aspiring data analysts to familiarize themselves with the suite of tools and methods at their disposal, such as in the case of this project, the Pandas and Seaborn libraries. Overall, this project was very informative and fruitful.

Link to the Dataset

Dataset link: <https://www.kaggle.com/omkarborikar/top-10000-popular-movies>

The primary problems I want to solve from this dataset are the following:

1.    What is the most popular movie genre?

2.    Are the most popular movies ones with a single genre or many genres?

3.    Is there a relationship between movie release date and the genre of the film?

4.    Is there a relationship between movie release date and the popularity of the film?

5.    Does having more than one genre make a film more popular?

6.    EDA of the dataset (max popularity, min popularity, average popularity score, etc.)

The reason why I want to solve the aforementioned problems is because I want to get a better idea of potential trends from the dataset. For example, what variables are most associated with the popularity score: film genre or release date?

I have used the Seaborn Python library in order to create my data visualization graphics for this project. I will be using Seaborn due to its ease-of-use and because of its versatile customizability options due to its connection with the Matplotlib Python library. I have created two graphs, which were a scatterplot and a distribution plot (histogram).

Solutions to the primary problems

Problem 1: What is the most popular movie genre?

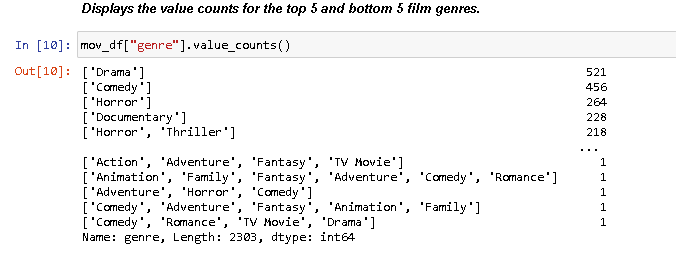
Answer 1:

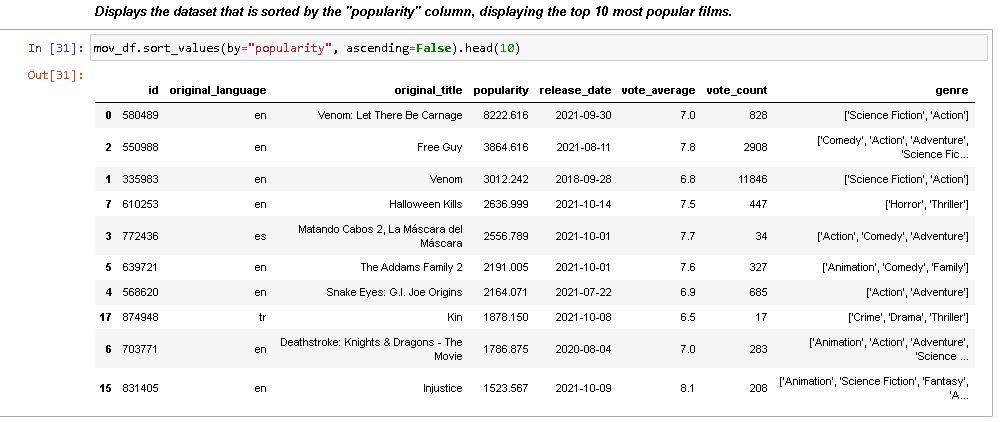
There is no single most popular movie genre. Instead, films with a combination of genres tend to be the most popular. The most frequent genres are “Drama”, “Comedy”, “Horror”, “Documentary” and “Horror and Thriller”, in that order.

However, the frequency of film genre does not appear to be correlated with the popularity of the film. That means, just because “Drama”, “Comedy” and “Horror” films make up 12%1 of the films in the dataset, none of the most popular films are ones that just have either of those genres. Instead, the five most popular films all have a combination of genres. In fact, four of the top five most popular films have the “Action” genre.

With that said, it does appear that “Action” films appear to be generally popular, but that feature alone does not result in higher popularity scores.

Relevant images:



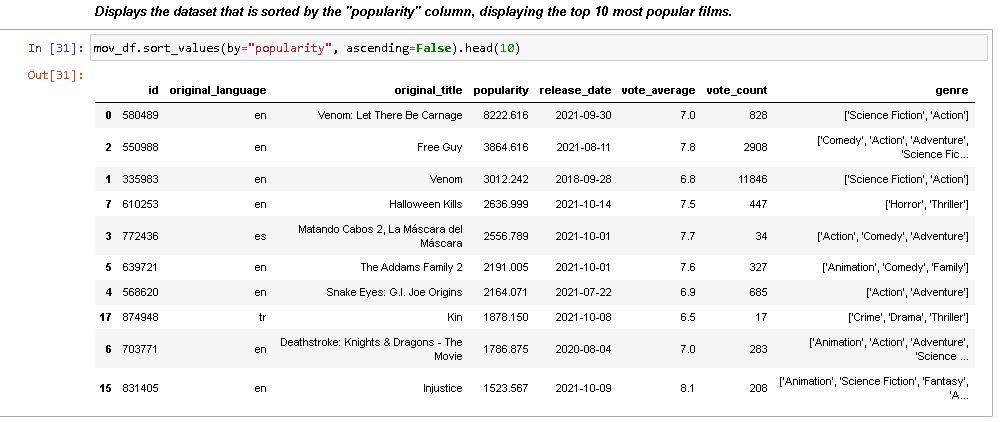


Problem 2: Are the most popular movies ones with a single genre or many genres?

Answer 2:

Given the previous answer, it is clear that the most popular movies contain more than one genre. That said, it is not clear how many genres are associated with the most popular movies, but it does appear to be at least two. In fact, four out of the ten most popular films are ones with only two genres, while the remaining six most popular films are ones with more than two genres.

Relevant images:

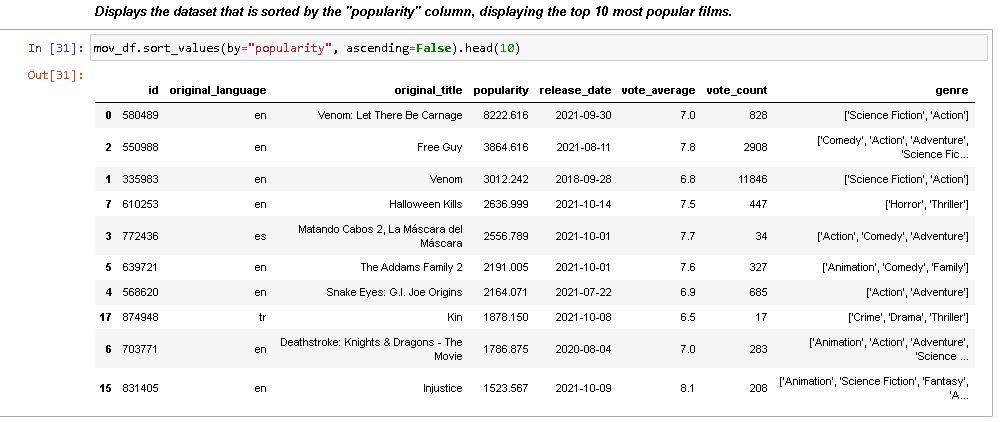


Problem 3: Is there a relationship between movie release date and the genre of the film?

Answer 3: There does not appear to be a clear correlation between release date and genre of the film. That is, given some year or some month, neither of these factors are associated with whether or not a movie will be, for example, “Action” or “Horror”.

That being said, there may potentially be some correlation between the two variables if, instead of looking at year, we look at month. That is, it is possible that certain kinds of films, which may be associated with certain genres, tend to be associated with certain months. For example, “Romance” films may be associated with the month of February due to the holiday of Valentine’s Day. In addition, “Family” films may be associated with November and December due to the Thanksgiving holidays and Christmas. Finally, “Horror” and “Thriller” films may be associated with October due to the holiday of Halloween. However, these are just speculative claims that are not directly supported by the data itself. For that reason, we should be hesitant to jump to conclusions.

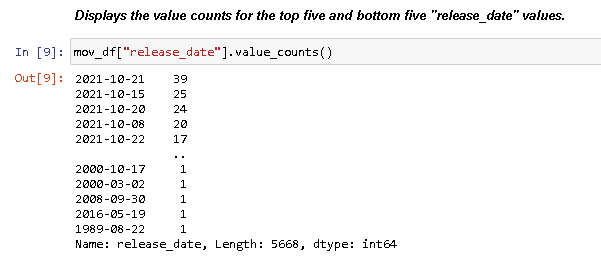
Relevant images:

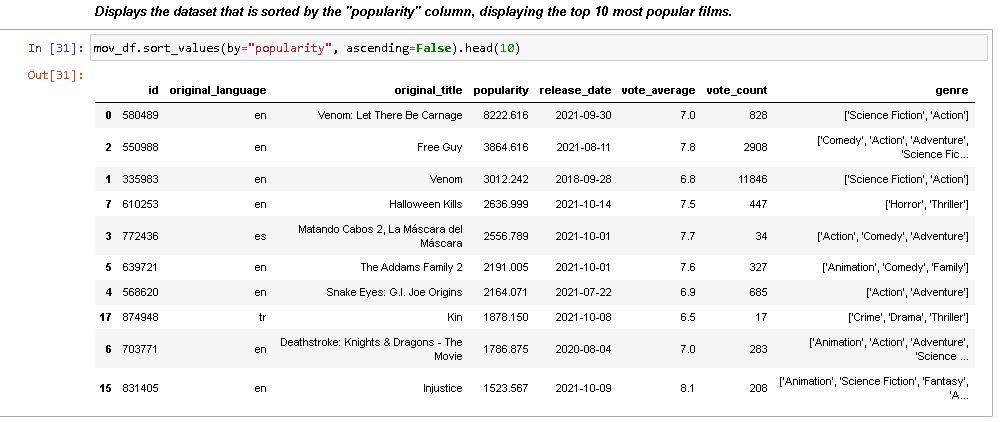


Problem 4: Is there a relationship between movie release date and the popularity of the film?

Answer 4: Given the previous answer, there does appear to be some correlation between release date and popularity. That is, when we look at the top 10 most popular films, eight out of ten of them were released in the period of 2021, with at least five also being released in October. The remaining two films were released in 2018 and 2020. While it is not possible, given the data itself, to understand why this is, it is possible that the recency effect could be influential here. That is, things that are recent are things that we tend to remember more and engage with more, which could drive up popularity scores.

Relevant images:



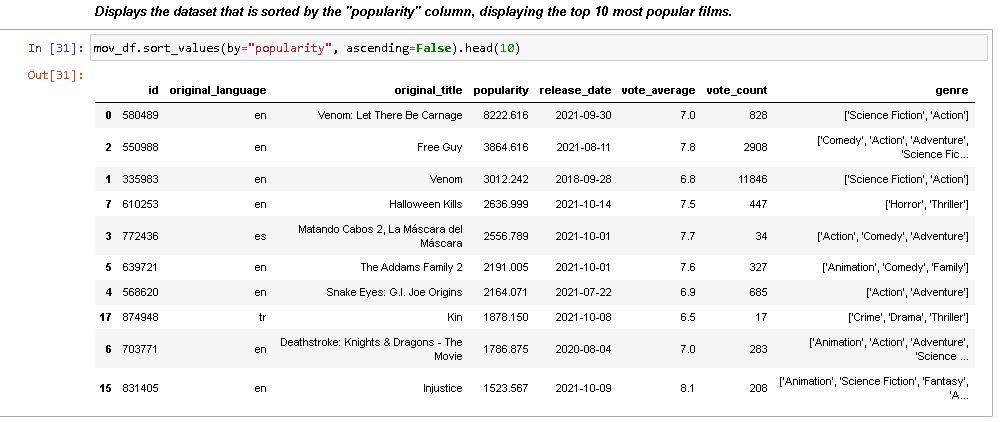


Problem 5: Does having more than one genre make a film more popular?

Answer 5:

Given the answers to Problem 1 and Problem 2, we can be confident that there is some correlation between number of genre and the popularity of the film. However, this correlation does not appear to be linear. That is, simply adding having more genres in a film does not result in that film being more popular. So, instead, there may just be some kind of correlation between already popular genres and the popularity of the film as a whole. With that said, it must be stressed, there are many factors besides genre that could make a film popular. Because of this, it would be inappropriate to jump to the conclusion and declare definitively that number of genres, let alone any genre for that matter, is directly responsible for a change in the popularity of a film.

Relevant images:

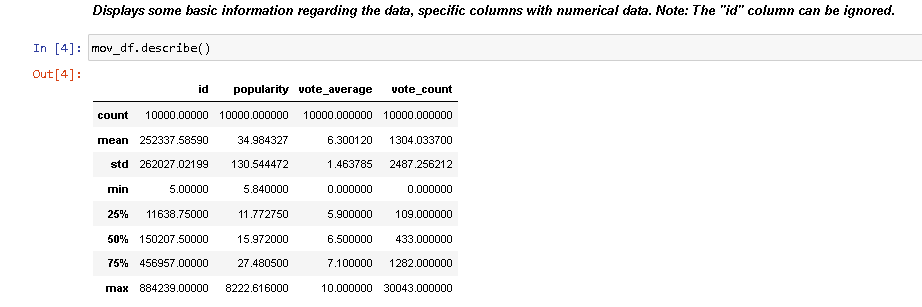


Problem 6: EDA of the dataset (max popularity, min popularity, average popularity score, etc.)

Answer 6:

A basic set of the EDA of the dataset can be seen from the following image. That said, other sections of the Jupyter Notebook file help us to make sense of the distribution of data related to the various columns of the dataset. That is, we are mostly looking at the frequencies of data, the top-most data values and the bottom-most data values.

Relevant images:



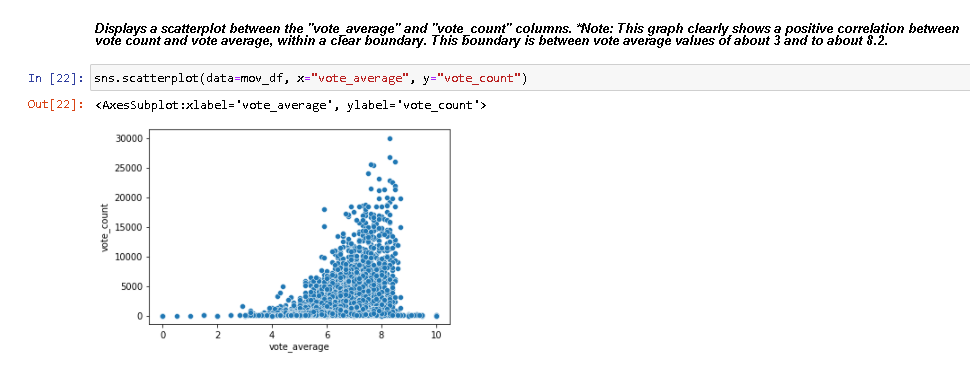
General comments:

Note: images are used more than once as part of illustrating the answers to the problems of this project. This redundancy is meant to reduce confusion and ambiguity as to how answers were derived.

As I have mentioned in some of my answers, it is extremely important that we don’t jump to conclusions whenever we see trends in the data. We cannot always be certain why some correlations appear to be the way they are. This is true even for small datasets like mine, whose shape is only 10000 rows by 8 columns.

What’s more, a crucial element of this dataset is the “popularity” column, which is not explained from the source webpage (see the dataset link for more information). It is not clear to me how popularity was derived and to what real influence it has. For example, the most popular film in the dataset is “Venom: Let There Be Carnage” but this film only has a vote average of 7.0 and a vote count of 828. Therefore, presumably, that means that only 828 unique users voted on that film, to which the average vote for said film was 7.0. While this may be impressive in its own right, it hardly demonstrates or validates that the film is the most popular. Ultimately, what I am getting at is, something as subjective and personal as “popularity” may not be best captured as a seemingly static number. Moreover, we cannot be certain that the data from the “vote count” column is reliable. It may be possible that non-human users (bots) voted on particular films. This is entirely possible due to automation. However, that is a speculative point and one that does not seem particularly the case given how much variation there is in the “vote count” column.

Having said all of this, what appears most clear to me is that there does appear to be a strong correlation between the “vote count” column and the “vote average” column. This is captured by the following image:



While there appear to be some outliers and “noise”, the overall shape of the scatterplot distribution is clear and quite compact, which suggests general consistency and uniformity in the data.

Finally, the last part of the Jupyter Notebook file includes an attempt at creating a simple “Recommender” system. This was done without the use of any explicit ML libraries, such as sci-kit learn. That said, the system was developed by taking a sub-section of the original dataframe, “mov\_df”, focusing exclusively on data which were within the 90th quartile of the “vote\_count” column. This is presumably to ascertain while films received the most votes and thereby likely to be truly popular films. However, as I mentioned in the Jupyter Notebook file itself, it is important to acknowledge that the system does not interact with or utilize the “popularity” column data from the original dataframe. For that reason, the recommender system may not be necessarily be capturing the most popular films any more than the original dataset captures with its “popularity” column.

\*Added Commentary: I believe this project could be improved upon by using ML-related tools and techniques to glean insight from the data. Improvements in gleaning insights from the data probably would also come about as using visualization software/services such as Tableau or Plotly. As I learn such tools and software, I will work to implement them to this project.

Notes:

1. The “12% of films” figure, as mentioned in Answer 1 of Problem 1, is the result of adding the frequencies of films with the “Drama” (521), “Comedy” (456), and “Horror” (264) genres to get 1,241 and then dividing that number by the total number of films in the dataset (10,000) and then multiplying that value by 100.