(1) Missing value:

From the plot of missing values, we can see if we plot the entire data set, we cannot see the names of the variables below the graph, so we take a subset of the data set, which keeps all missing values. It turns out we have about half the complete data. More than 80% of data missing occurs in the three variables of Average Rating, Number of Votes and Runtime. At the same time, we also found some patterns between different types of variables: Average Rating and Number of Votes will be missing at the same time and the number of missing is almost the same, which may be caused by some force majeure factors in the data statistics process, for example, a certain film has not publicly released voting and scoring channels.

(2) Time series:

We want to find the time trends in the number of films in different types. From the plot of missing values, we can see the total number of documentary films has been almost consistently the highest in 20 years, while the number of adventure and crime films has been the lowest. This may be because documentaries are more popular in the market and cheaper to make. In addition, we can see that from 2020 to 2018, the number of films released in different categories increased steadily, but there was a relatively large decline in 2020, which is probably due to the impact of COVID-19 in 2019-2020, which caused the entire film industry to enter a winter period. Rising filming costs, falling cinema attendance, the absence of actors due to the pandemic and stringent controls in different countries are all important factors. However, we can also see new opportunities for the film industry in 2021, with an increase in the number of films of all types, which cannot be achieved without the effective control of the epidemic worldwide and the research and development of vaccines.

(3) line plot:

We want to find whether the movie rating is related to the number of movies. From the line plot of films with different scores from 2020 to 2021, we can see all categories saw a decline in the number of films in 2020 because of the epidemic. Between 2000 and 2020, the total number of movies that scored more than 2 increased. Interestingly, the number of films with a score of less than 2 fluctuated more significantly during the 20 years, which may be due to the low public evaluation of low-scoring films, and the return on investment would change more sharply with different years, thus affecting the number of films released.

(4) Cleveland plot:

We want to explore the relationship between the average rating of different types of movies over the past 20 years. As can be seen from the graph, talk show, documentary and music films occupy the top three spots, while short, horror and science films occupy the bottom three spots. Other categories were evenly distributed between 5.5 and 7 score. This may be because people prefer light-hearted movies in stressful lives. Talk shows and musical movies satisfy this need, while horror movies are the opposite. Science and short movies have lower average scores maybe because they are not so fancy.

(5) mosaic plot:(startyear)

Here, we want to explore whether there is a relationship between the release time of a film and the type of film. As shown in the figure, in six different years from 1990 to 2015, the proportion of films of each type is basically the same, so we can think that these two variables are independent of each other. We can also see that documentaries have increased significantly in the past 25 years, which may be due to the fact that it may take several years to shoot a documentary, and the improvement of filming technology and easier access to information are both reasons for this phenomenon.

(6) mosaic plot: (timelong)

Here, we want to explore whether the length of a movie is related to the genre. As shown in the figure, we can see that these two variables are not independent. Documentaries take up the largest proportion of 60 minutes, while action films take up the largest proportion of more than 120 minutes. The proportion of films between 60 and 120 minutes is the largest for all genres. This may be because action movies need more foreshadowing and development of the story, while documentaries only narrate the story.

(7) time series of box office:

Here, we want to explore the box office trends for each year from 1980 to 2021. As can be seen from the plot, the box office grew steadily from 1980 to 2020, but fell off a cliff in 2020, which may also be due to the impact of the epidemic on the film industry. Fewer films, fewer moviegoers and a large number of theater closures were among the main reasons for the decline. We can also see that the box office in 2020 dropped to the level of 1980, which shows that the epidemic has a huge impact on the film industry.

(8) heatmap:

Here, we want to explore the relationship between average gross box office and genre between 2017 and 2021. As can be seen from the chart, the average box office in 2017-2019 is higher than that in 2020-2021, which may be due to the impact of the epidemic on the film industry. Science films and music films are the top grossing films. As can be seen from the previous figure, the total number of sci-fi movies is relatively low and the total box office is relatively high, which may be due to the high unit price of such movies. Because sci-fi movies are usually expensive and use IMAX and 3D technology, the high unit price is reasonable. The missing data in the chart is because we only analyze the top 200 films at the box office, while the missing data in the western genre is because the number of films in this genre has declined significantly since the beginning of the century. Some years in the top 200 list have no such films, and the number of moviegoers has decreased. The documentary's missing figures for 2017 and 2019 are also due to its absence from the top 200 grossing lists for that year, and documentaries are not released in theaters every year.

(9) boxplot:

Here, we want to explore the distribution of total box office data by genre between 2017 and 2021. For the sake of illustration and analysis, we excluded movies that made more than $50 million. As we can see, horror films have the highest median box office, while mystery films have the lowest. Action movies have the most outliers, which may be because the quality of action movies is uneven. High-quality action movies such as 《The Avengers》 can reach the top of the total box office, but there are few such excellent action movies, and most action movies are concentrated in the low box office range.

(10) parallel:

Here, we want to explore the relationship between different variables in the film box office data set from 2017 to 2021. The variables include movie duration, average score, total vote count, total box office and domestic box office. We can see that there are distinct clusters in the figure. At the same time, there is some positive correlation between the number of people rating movies and the total box office, which may be because movies that are watched by more people have higher box office and therefore have more people rating them. Interestingly, the longer the film is, the higher the average rating is, which may be because the longer the film is, the richer the film content and plot can be displayed. Therefore, the story will be more complete, giving the audience a better viewing experience and therefore higher scores.