**Describe the data:**

We were both involved in the search for the data.

In the process of finding data, we found that the film is easier to find the main data, such as name, category, release time, rating and so on, which could be found at <https://www.imdb.com/interfaces/>. But the movie box office data is more difficult to find, we tried many websites but cannot succeed, either because the box office data we found was difficult to match with the movie data we already have, or because the data on the website was not available to download. Therefore, we finally tried the method taught in the class, using ‘Rvest’ to do the web scraping. Finally, we successfully obtained the box office data we needed, and combined it with the previous one to form a more complete and comprehensive data set, which was convenient for us to conduct further analysis.

**Basic information:** Our data set contains tens of thousands of pieces of data covering global film data from 2000 to 2020, including film title, release time, film category, film duration, lead actor, rating, total box office, etc. Later, we will focus on the period from 2018 to 2020. The reason why we choose this period is to discuss the impact of the epidemic on the film industry, which will also be part of our report.

**Problems with the data:** Since we use ‘Rvest’ to get the movie box office data, there are many missing values of box office in our data set. Missing values will be processed in the subsequent data processing process, so as to ensure the integrity of our data set and avoid the influence of missing values on our plotting.