Data Cleaning

The very first step is to import necessary libraries to wrangle the data. For example, we need to import pandas and numpy with their standard aliases. Others will be imported later as needed We then load the dataset as pandas dataframe to handle it. The next step is to inspect the dataset especially the first few rows, the column labels and their data types. This is important for the upcoming steps to have a good idea about the dataset. It helps identify, and ultimately drop, useless columns such as ID, codes, etc... This can be done using .head(), .describe(), and .info() methods of the dataframe to inspect it, .columns for column labels, and .dtypes for data types of columns. In this step, release date column (Year-Month-Day format) will be split to separate year, month, and day numerical columns so that the date could be inputted inside ML model(s)

Next, we need to inspect for nulls and missing values in the dataset. In our case, there were no nulls although there were some zeros in the budget column. Going back to the dataset description, it was mentioned that the dataset had 0 budget for films for which the budget was unknown. In that case, we had to replace the zero values with some better representative values. There were about 2200 movies with missing budget. Because budget may be a critical variable in estimating gross, we will filter out those movies instead of substituting their budget with a statistic. Later, we will get back to those movies, estimate their budgets by other means, and append them back to the original dataframe

In many numeric columns, there may be significant number of outliers that could to be handled in order not to disrupt plotting and visualization later. Those could be handled early on by creating histograms of numerical data columns and removing outliers below or beyond 95% percentile of their respective columns, if needed.