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.
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. Median was chosen as the substitute due to its independency of outliers.
In many numeric columns, there may be significant number of outliers that may need to be handled in order not to disrupt plotting and learning the data later. Those could be handled early on by creating histograms of numerical data columns and removing outliers below or beyond specified percentile of their respective columns