**Data Wrangling**

There were a couple of things that needed to be done. First, I had to parse the dates from the **‘release\_date’** column to datetime objects. The second change, involved dropping unnecessary columns, rows with null values, and rows with zeros.

Parsing the dates of this dataset was trickier than usual. I would normally use the **parse\_dates** argument of the **pd.read\_csv()** function, or the **pd.to\_datetime** function, but they proved unhelpful here. Instead of properly parsing the dates, I received some future years. This is because this dataset roughly spans the years of 1960-2015. To make matters worse, the **‘release\_date’** column does not specify the century of the year; for example, the entry 10-31-65 represents October 31st of 1965, but Python incorrectly parses the year to be 2065. To correct this, I created a Boolean mask to identify all years in the 1900s, then inserted a ‘19’ marker in front of the year to clearly indicate that they belong in the 1900s.

Dropping columns didn’t prove to be nearly as difficult. The first columns I decided to drop were the **‘budget’** and **‘revenue’** columns, and replace them with the **‘budget\_adj’** and **‘revenue\_adj’** columns, as they represent the budget and revenue values adjusted for inflation. Next, I dropped rows with null values, and zero values. Lastly, I only needed the following columns, so I updated the DataFrame accordingly: [**'original\_title'**, **'cast'**, **'director'**, **'genres'**, **'production\_companies'**, **'release\_date'**, **'vote\_average'**, **'release\_year'**, **'budget'**].