The first cleaning step I performed was removing all null values. I checked beforehand that none of the null values came from irrelevant columns (regional sales). My question requires complete cases and for the sake of simplicity, imputation is not used. The data set is large enough for imputation to not be impactful.

I noticed some of the columns were of the wrong data type. User\_Score was an object instead of a float64 value. I changed it to a float64 data type. I also changed Platform, Genre, Rating, and Publisher as a category instead of an object.

Under Platform, DC (Dreamcast) has a fairly low sample size. I combine it with the WiiU category and renamed it ‘Other’.

There are many categories under the Publisher column. I found the first Publisher with less than 30 observations and then took everything below that and combined it into the ‘Other’ category.

The Developer column was done the same since there many distinct categories. I took the top 50 categories with the remaining ones collapsed in the ‘Other’ category.

Next, I created a new column which shows if the Publisher is the same value as the Developer.

I notice the Rating column has only 1 count for RP, K-A, and AO. The game tagged as AO was GTA: San Andreas. There was a controversy about a scene in that game which made it AO several years after release. Originally, it was tagged as M, so I collapse it with the M category. K-A is kids to adults, so that belongs in the E category. RP means Rating Pending. I collapse it with the T category because it is the most populated one.

I examine how many rows have Critic\_Count or User\_Count less than 10. These are the number of votes that contribute to Critic\_Score and User\_Score respectively. There are 1954 rows. I don’t remove any because there are so many.

I change the year\_of\_release column to years\_since\_release. The data was collected at the end of 2016, so I assume that any game released in 2016 has been out for a year, 2015 for 2 years, etc.

Finally, I created a new column to determine if a game is good or not. The definition I used is if User\_Score >= 8, the game is good.

There are outliers, however, they are valid, so I don’t remove them. For an example, one of the outliers is the highest Global\_Sales contributor, Wii Sports. It has over 2 times the value of the 2nd highest Global\_Sales. The outliers can be examined further than the analysis to see what effect they have if neccesary.