Wrangle Report

The gather phase of the project was done in the following steps: I read the file ' Video_Games_Sales_as_at_22_Dec_2016.csv' into a dataframe called 'df_original'. That csv file can be found at the following link: https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings. I then made a copy of that dataframe to work with called 'df'.

During the assessment and cleaning phases, I found and fixed the following issues:

1. There were 2 rows that were missing names, ones of which was also missing many other attributes. They were at indexes 659 and 14246. Since there was no way to tell what this data was, I deleted the 2 rows.

2. There were 4 rows that contained games that were released after 2016. They were at indexes 5936, 14086, 16222, and 16,385. Since this dataset is only supposed to contain games released up until the end of 2016, this mistake must have been made during the web scraping process. I then deleted the 4 rows.

3. The column name 'User_Score' was of the datatype 'object'. In order to convert this to a float, I also had to deal with the 2506 rows that contained the string 'tbd'. In order to do this, I used the function 'to_numeric', and forced all non-number columns to be converted to NaNs via the 'coerced' argument.

4. I found that within the subset of the columns 'Name', 'Platform', and 'Year_of_Release', there were 2 duplicated rows. The first was the index 14246 that was removed during the cleaning of assessment 1. The second was the index 16233. I removed the second row.

I then stored the cleaned dataframe into a csv file called 'Video_Games_Sales_as_at_22_Dec_2016_master.csv'.