Capstone Project 1: Data Wrangling Report

The dataset used for this capstone project came from a Kaggle competition on video game sales with Metacritic ratings:

https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings

The dataset was in a clean and well structured state because it came from this source. The only real problem here that needed to be dealt with were the NaN values and duplicate values as well as filtering and grouping of columns and rows.

After working through the dataset a bit, it was noticed that there were place-holder values of 'tbd' under the scores fields. Those values were converted to NaNs during the importing of the data with the read_csv function of pandas(pd.read_csv(na_values='tbd')). A problem that was noticed in this dataset was that it included upcoming games that had yet to be released, given the year this data was compiled. Any games that had years of release past 2016 were filtered out. The video games that did not have Metacritic scores and 0 units sold were also filtered out. The data type of the year column was of float64. Using the astype method for a pandas series, the Year_of_Release column was converted to a data type of int64.

Using the pandas DataFrame describe method provides a statistical summary of all the numerical columns. The global sales column was seen to contain a very large outlier. The 3rd quartile had a value of 0.75 while the max value was 82.53. Upon further investigation of the data, it was seen that the global sales value of 82.53 came from a Nintendo game called "Wii Sports". This was seen by sorting the dataset by it's global sales values. Throughout the gaming community, it is well known that this video game performed quite well upon its release because it was bundled with the release of the Wii console. This video game will be kept in the dataset and will be removed if necessary during the analysis phase.

The next thing that was looked into was whether or not there were any duplicate occurrences of video games. Using the pandas series value_counts method, the video games that had multiple occurrences were counted up. Some games had upwards of 8 appearances in this dataset. Looking deeper into the data, it is seen that a video game could have multiple release dates and different platform releases(ie. PS2, Xbox, PC...). Some publishers re-release their games on different years or remake their old games using the same name. In order to work with this dataset, the duplicate games had to be dealt with. It was decided that the most relevant columns to keep for this capstone project were name, year of release, publisher, global sales, and the critic and user score. The dataset would have to be sorted by descending global sales and then grouped by the video game names. Aggregations were performed on year of release by first value, publisher by mode, global sales by sum, critic and user scores by max, and genre by mode.

The final data cleaning and wrangling steps to be performed on this dataset were easy ones. The user score was on a scale of 0-10 while the critic score was on a scale of 0-100. The user scores were multiplied by 10 in order to bring the two scores on the same scale. Then the entire dataset was sorted by the descending global sales value. This cleaned dataset was then exported to a csv file in order to perform exploratory data analysis in the next step of this capstone project.