Data Wrangling for Capstone 1 Project

For my capstone 1 project, I’ve decided to develop a predictive model for any given video game Metacritic review score. I began by locating all the historical review score from Metacritic on video games. I was able to find a Dataset from Kaggle that contained sales data and Metacritic review score for all video games before December 2016. To build the best predictive model, my data set will only include video games that have been released after 2012. Video games review scores are rated differently now because video game console have more capabilities. With that in mind, it would be best to try to keep the data used to construct the model as current as possible. After importing the data set, I noticed several columns contained sales data which is irrelevant in terms of the project goal. I proceeded to drop all columns containing sales data from the data frame. Next, I discovered that there were several missing values in the user count and the VGR.reset\_index(drop=True)user rating columns. I felt the best way to quickly deal with this problem would be to calculate the mean of each column and replace those missing values with mean of all know values. Next, The rating column contains some missing values as well. The ratings column contains data about each games ESRB rating. After some brainstorming, I couldn’t come up with a method to fill the missing values. The best thing to do in this situation is drop the video games that have missing values in the rating column. Finally, the dataset is ready for analysis.