To prepare the data for our Machine Learning model we must first start by importing our data into the model and creating a data frame. Our team cleaned the data before importing it to better fit the model and produce more accurate results, since the data is pulled from a public website ([www.Kaggle.com](http://www.Kaggle.com)) we must ensure that we standardize the data to ensure the integrity of our analysis. We found that the data was mainly in the Wrong Form and with some work could be standardized or repaired. As an example, there were inconsistencies with the player column, and we decided to drop it since it would not have an impact on the outcome of our analysis.

Feature engineering is an important part of the Machine Learning process as taking the time to apply the process will improve the quality of our results. Some of the features that we will extract from the raw data include establishing an overall score by combining the Critic Score and the User Score and assessing how the various scores compare to global sales.

* + Multiply the Critic Score and the User Score to establish an Overall Score (out of 1000 = 100 critics x 10 users)
  + Explore evaluating the correlation between the scores and Global Sales

We will use the holdout validation technique that will randomly split the data into training and testing data sets. There will be a sample set of data that is excluded to ensure efficiency during testing. The complete set of data will be used once the machine learning model is completed to confirm accuracy of the final model.

We decided as a team to use a Linear Regression Model for out Machine Learning analysis. Linear Regression is one of the most common algorithms for the regression task; we believe the model to be very straight forward and very easy to update. This model can make predictions for years of experience beyond the range of the current data which is why we selected Linear Regression to make accurate predictions from data that is a few years old. Limitations of this model are that there must be a linear relationship in the data, if there are non-linear relationships in the data it would cause the model to perform poorly. We do not anticipate having this issue as it appears we do have a linear relationship, as an example between our scores and sales in our data set.