1. All entries from the year 2005 onwards.

A screenshot of a computer

Description automatically generated

1. This is the testing dataset (5,356 entries out of 21,428-25%):

A close up of a computer

Description automatically generated

This is the training data set (16,072 entries out of 21,428-75%):A close up of a computer

Description automatically generated

1. The p-value on the Home Run variable is 0.9529. The p-value on the At Bat variable is 0.8278. The partitions are validated.
2. A picture containing computer

   Description automatically generatedBelow is the scatterplot matrix.
3. Below is the correlation matrix. The majority of p-values show zero, and this can be seen later in the stepwise progression step, that the majority of p-values fall below 2.2e-16.

A screenshot of a computer

Description automatically generated

1. The results of the multiple regression are below.A close up of text on a black background

   Description automatically generated
2. A stepwise regression was performed comparing Home Runs to the variables: Games, At Bats, Runs, Hits, Doubles, Triples, Runs Batted Ins, Stolen Bases, Caught Stealing, Base on Balls, Strike Outs, and Year. The p-values for each variable except year fall below 2.2e-16, and the p-value for year was 0.02611. Outputs for each regression can be seen when the code is run.
3. A new multiple regression was made using Games, RBIs, Strike Outs, and Year. The equation for the new model is:

HR = 0.28037(RBI) - 0.03757(G) + 0.03843(SO) + 0.04703(yearID) - 94.24990

1. The VIF for Games is 4.605026, for RBIs it is 6.432914, for Strike Outs it is 6.028173, and for Year it is 1.028050. For Games, RBIs, and Strike outs, there is moderate multicollinearity. For year the VIF is acceptable low, being slightly above 1.
2. 89.58% of variability in home runs is accounted for by this model.
3. Personally, I would use this model to estimate for making predictions on number of Home Runs a player will hit. A prediction this model makes will only be about 2.397 Home Runs off from the actual value.
4. I think this model is appropriate for interpretation because of the low residual standard error, relatively low multicollinearity, very low p-value, and 89.58% or variability being due to this model.
5. When the same model is created using the testing set instead of the training set, the same variables are significant because each value (adjusted R2, multicollinearity values, Residual Standard Error, etc.) is very similar to the values produced by the model built using the training data.