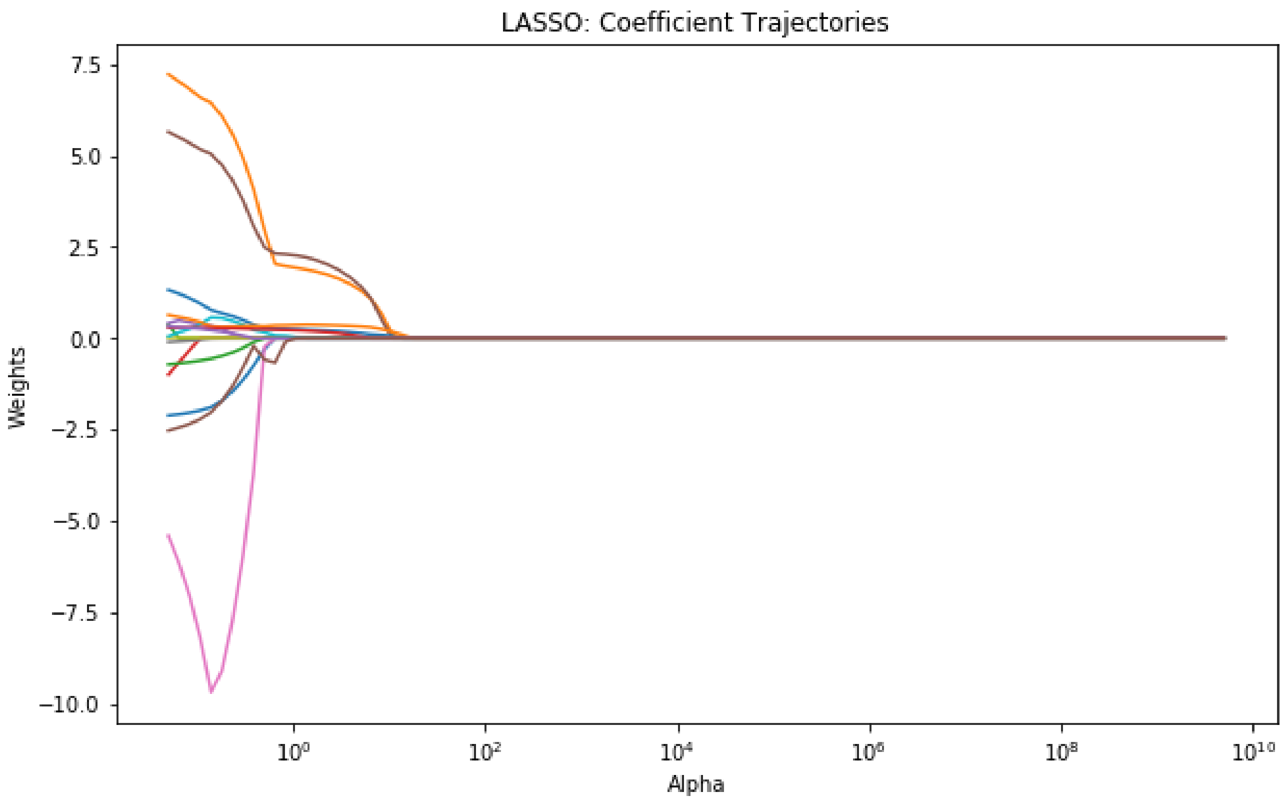
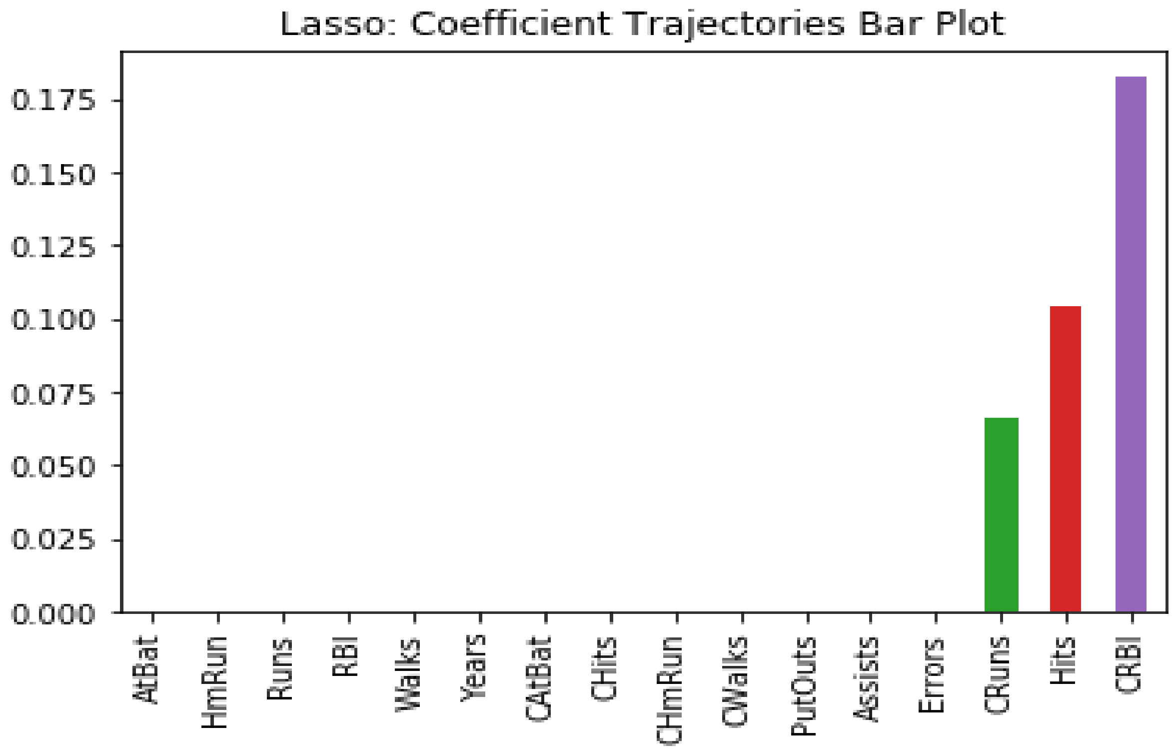
1.1) **<u>LASSO Regression:</u> Create a visualization of the coefficient trajectories. Comment on which are the final three predictors that remain in the model.**

When we use LASSO regression with a penalty of 10.77, the final three predictors that remain in the model are CRBI, Hits, and CRuns.



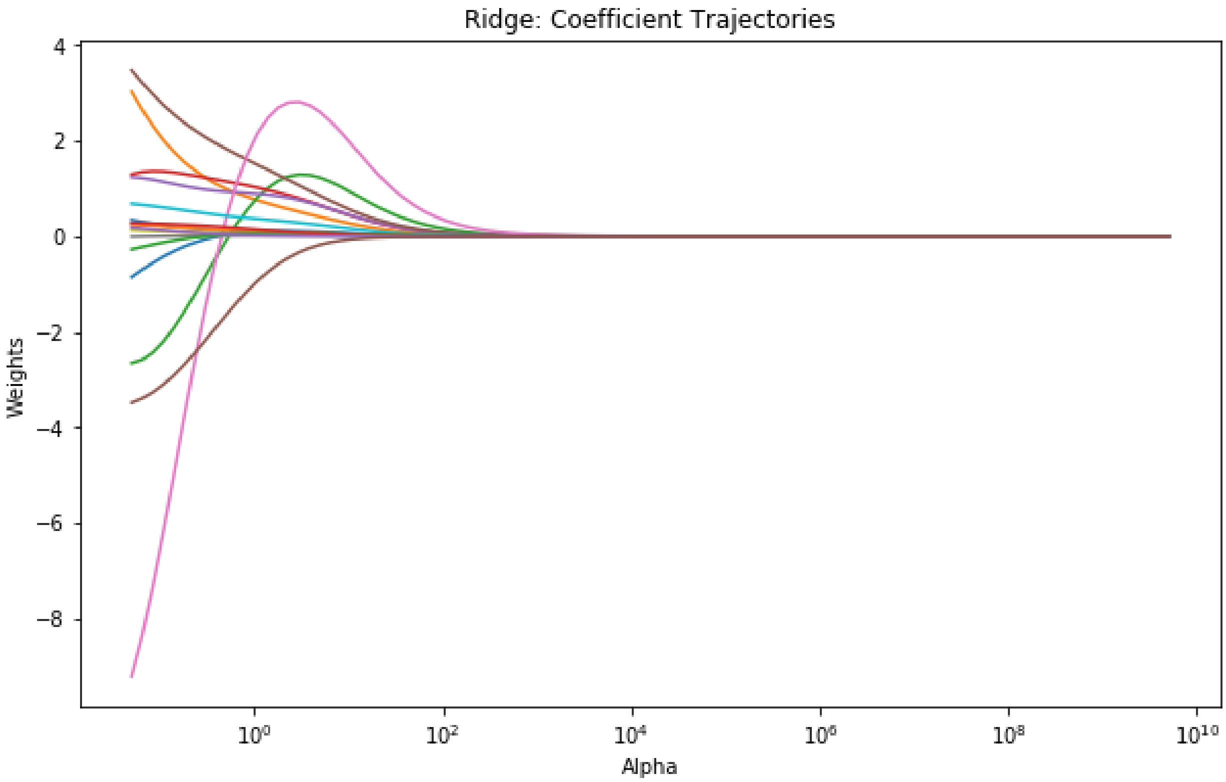


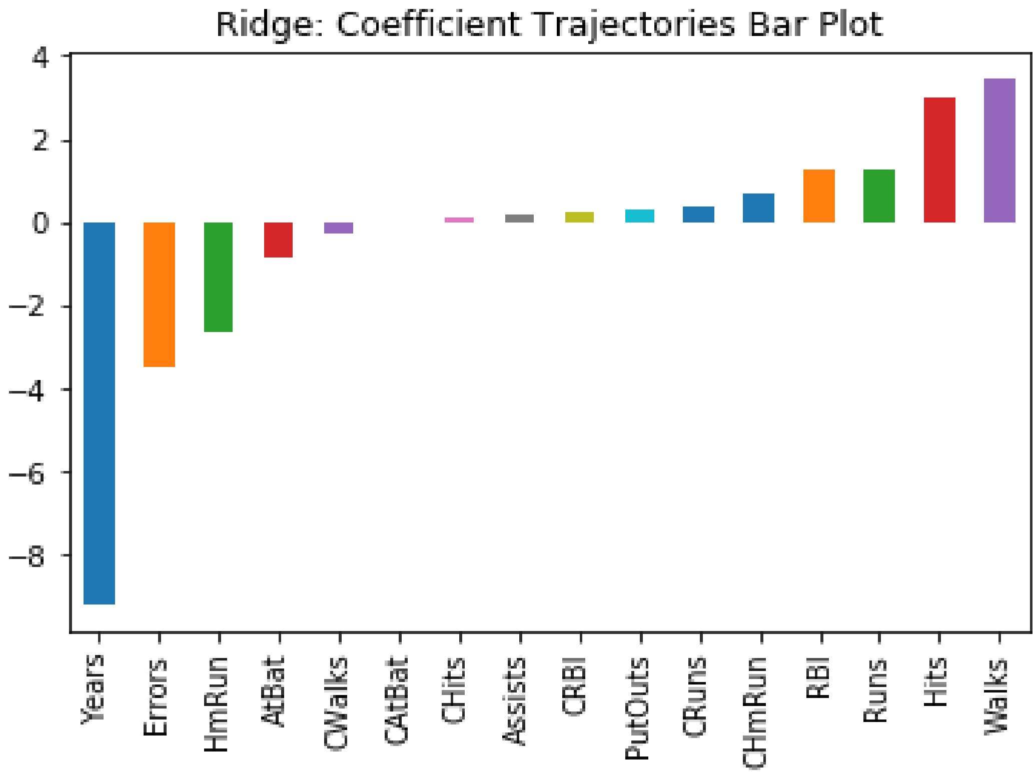
<u>LASSO Regression:</u> Use cross-validation to find the optimal value of the regularization penalty. How many predictors are left in that model?

When we use 10-fold cross-validation for the LASSO regression, the optimal value of the regularization penalty is approximately 0.0635. Using the optimal value of alpha, we see that there are 14 predictors left in the LASSO Regression model. We can see that both HMRun and CHits have their coefficients shrunk to 0, which leaves us with 14 predictors.

1.2) **<u>Ridge Regression:</u> Create a visualization of the coefficient trajectories. Comment on which are the final three predictors that remain in the model.**

From the visualization above, when we use Ridge regression the top three predictors are Walks, Hits, and Runs. Unlike LASSO, Ridge Regression cannot shrink any coefficients to zero.


Ridge: Coefficient Trajectories
Weights
Alpha
4
2
0
−2
−4
−6
−8
10^0
10^2
10^4
10^6
10^8
10^{10}



<u>Ridge Regression:</u> Use cross-validation to find the optimal value of the regularization penalty.

With 10-fold cross-validation for the Ridge Regression, the optimal value of the regularization penalty is: 0.05.

2a) Explain in your own words the bias-variance tradeoff.

The Bias-Variance Tradeoff is a phenomenon that appears in machine learning. Bias is the difference between the model's expected predictions and the true values. Variance refers to a machine learning algorithm's sensitivity to different training sets. The goal of any machine learning algorithm is to achieve low bias and low variance. The Bias-Variance Tradeoff states that we cannot achieve an algorithm that minimizes both.

To illustrate, as a machine learning algorithm becomes more complex, then the chances of it creating a model that overfits also increases. This means that a more complex algorithm is likely to design a model that does well on the training set, and thus has a low bias, but it could also be learning the noise as well. If we feed in a different training set to the same algorithm, then it would likely produce another model that has memorized that training set, and also has low bias, but may make predictions that are very different/inconsistent from those of the previously learned model. The algorithm would thus be said to have high variance while having low bias.

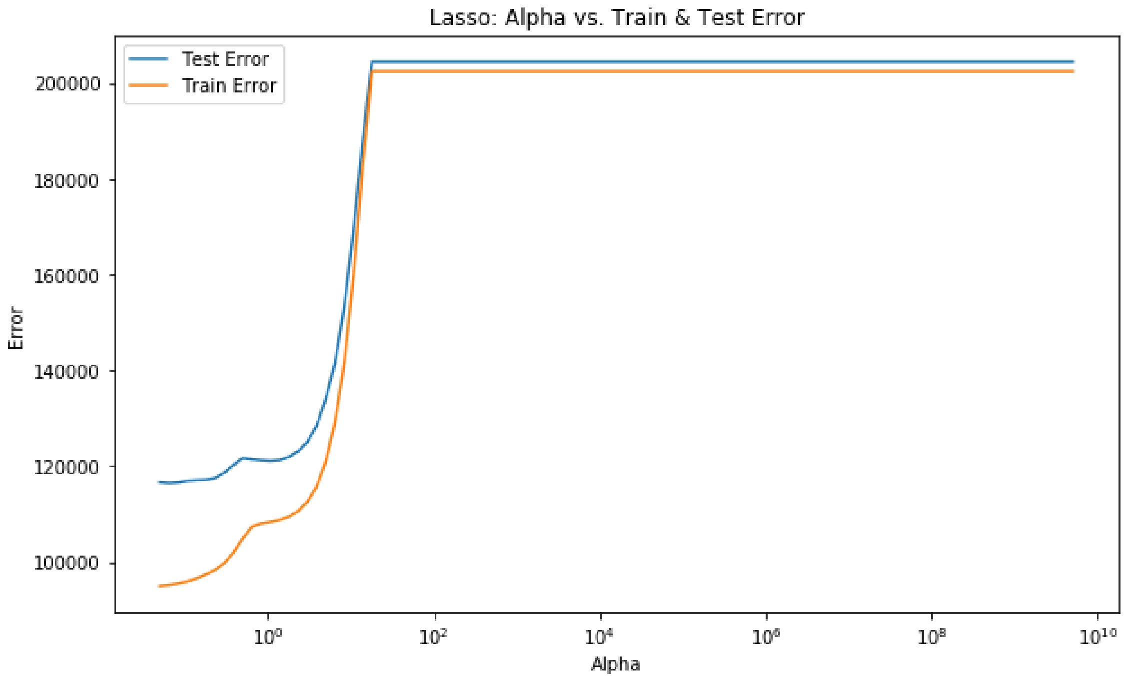
On the other hand, as a machine learning algorithm becomes simpler, then it is likely to produce a model that that underfits. As the algorithm is now simpler, it would have a harder time capturing the underlying trend of the data and may ignore important features of the dataset. If we were to pass such an algorithm a different training set, it would develop a model that would make similar predictions to the previously learned model, but the predictions may not capture the underlying trend of the data that well. Such a machine learning algorithm would be said to have low variance, but high bias.

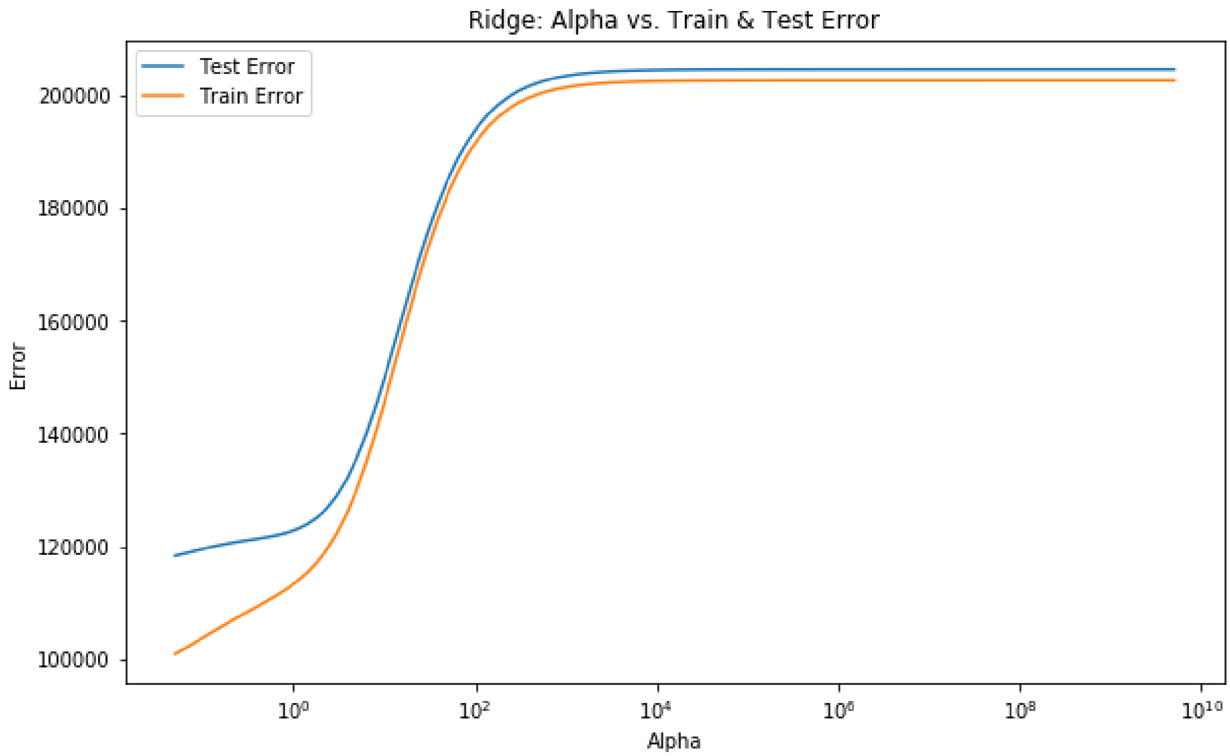
We can see that there is a tradeoff between bias and variance. We cannot create a model that accurately learns the irregularities in its training data and also generalizes unseen data well. A high variance, low bias algorithm may do well on the training set, but may over-generalize unseen data. On the other hand, a low variance, high bias algorithm may produce models that make similar predictions, but may not fit the data well.

2b) What role does regularization play in this tradeoff? Make reference to your findings in number (1) to describe models of high/low bias and variance.

The goal of regularization is to reduce the overfitting of a machine learning model by adding bias. Regularization regularizes or shrinks the coefficient estimates towards zero in an attempt to discourage a more complex or flexible model, to avoid overfitting. Models that tend to have larger coefficients tend to have a higher variance, and vice versa. For both Ridge and LASSO regression, the value of alpha controls the amount of shrinkage of the coefficients. A higher value of alpha will cause more coefficients to shrink and will thus yield a simpler model, which will have a higher bias but lower variance. Conversely, a lower value of alpha means less shrinkage and will thus yield a more complex model, which will have high variance, but low bias.

In both of the graphs, we can see the training and testing error for each value of alpha for both LASSO and Ridge Regression. We see that as the alpha values decrease (and the model complexity increases), the training error decreases but the test error starts to increase. The widening gap between the train and test errors indicates overfitting, as the model seems to do very well for the training set, but does a lot worse with the testing set. Here, the model is said to have high variance and low bias. In contrast, as we increase the value of alpha and thus increase shrinkage (reducing the model complexity), we see that not only do the training and testing errors of the model increase, but also that the gap between the training and testing error decreases. This indicates underfitting in which there is low variance and high bias. Thus, we can see how regularization comes into play in terms of the Bias-Variance Tradeoff.




Ridge: Alpha vs. Train & Test Error
Test Error
Train Error
Error
Alpha
100000
120000
140000
160000
180000
200000
10^0
10^2
10^4
10^6
10^8
10^{10}