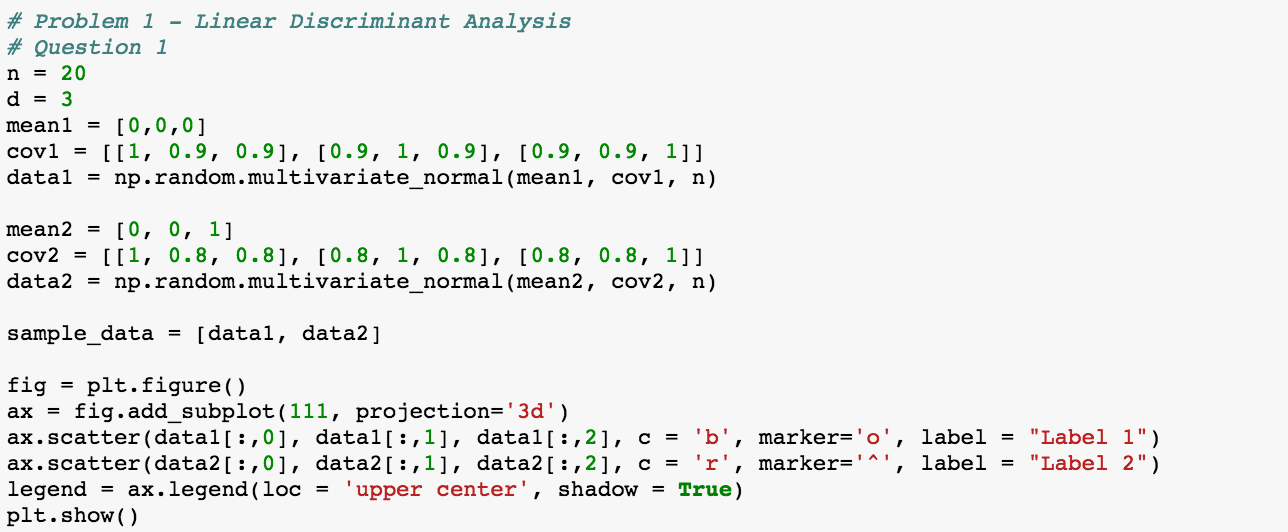
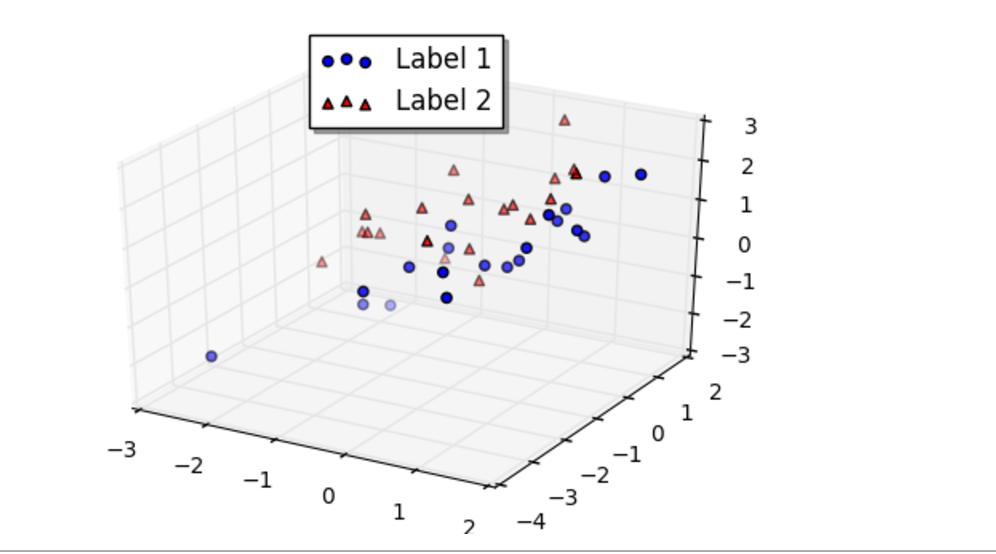
Problem 1: Linear Discriminant Analysis





This plot makes sense because it has both distributions that appear to follow the same axis but the points in red, Label 2, are more tightly packed than those in blue, Label 1.

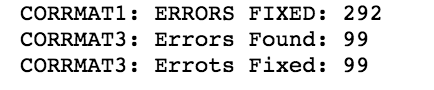
Problem 2: Using Low Rank Structure for Corrupted Entries

First, I plotted each of the features on a scatter plot against a linear index and inspected them. By doing this, it was clear that the value of the corrupted entries was about 10,000 because most of the plots were randomly distributed and then there would be a line of data points at 10,000. By looking at the data, it was easy to see that the actual corrupted value was 9999.

For corrMat1, I first found the rank of the matrix to be 36. This was not helpful. Then, I removed all of the features that contain the corrupted value (9999) and re-ran the rank. This gave a rank of 1 and the RREF was a series of fractions. This indicates that all 75 remaining rows are multiples of each other. I then extrapolated this to the features I had removed to correct the corrupted values. I also ran a similar analysis where I removed all of the rows that contained corrupted entries and the same rank / REEF combination was discovered.

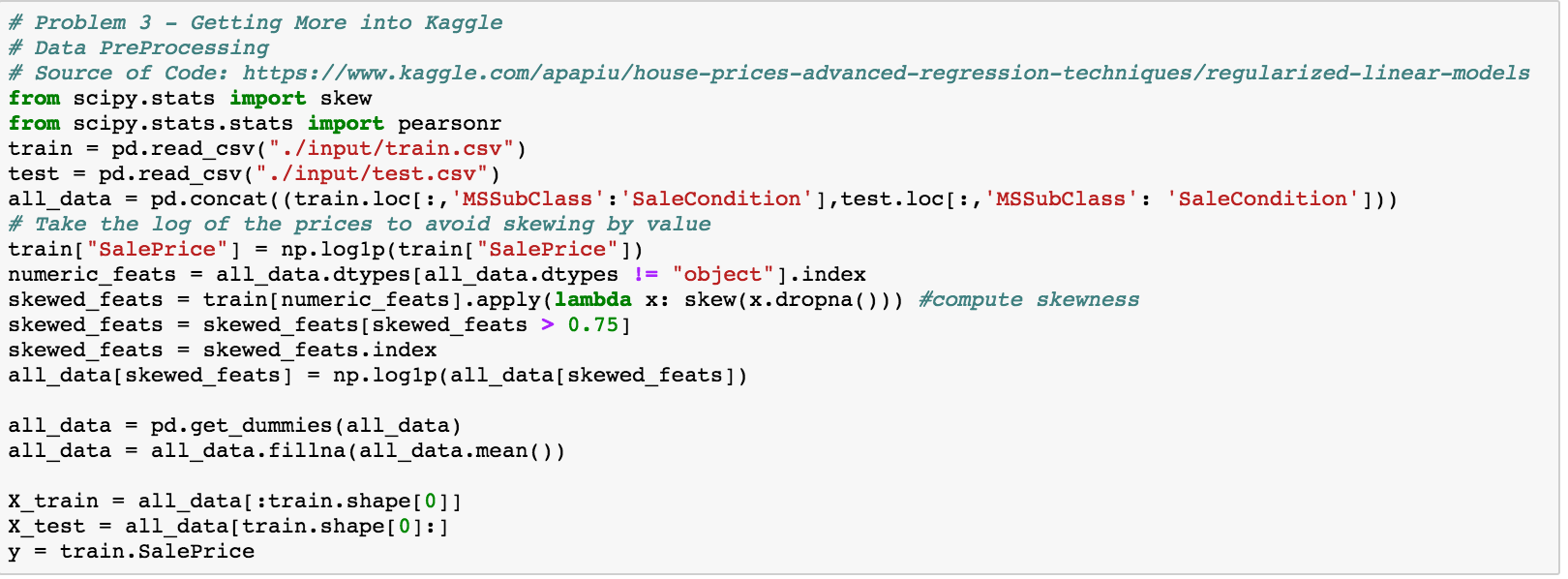
For corrMat3, I first found that the rank of the matrix which is 100. This means that each feature is linearly independent. Then, I calculated the row reduced echelon form of the matrix. This returned the identity matrix which indicates that corrMat3 is symmetrical along its diagonal. With a simple loop, I replaced each value that was not equal to its counterpart with the value that was not the error value (9999). This will recover all of the data points where the entry is corrupted on only one side of the diagonal. If it is corrupted on both sides, or on the diagonal, then it will be remedied. However, because the matrix is linearly independent and symmetric across the diagonal, having one corrupted entry on the diagonal is not a problem because it is still independent. Having more than one would affect the rank of the matrix.

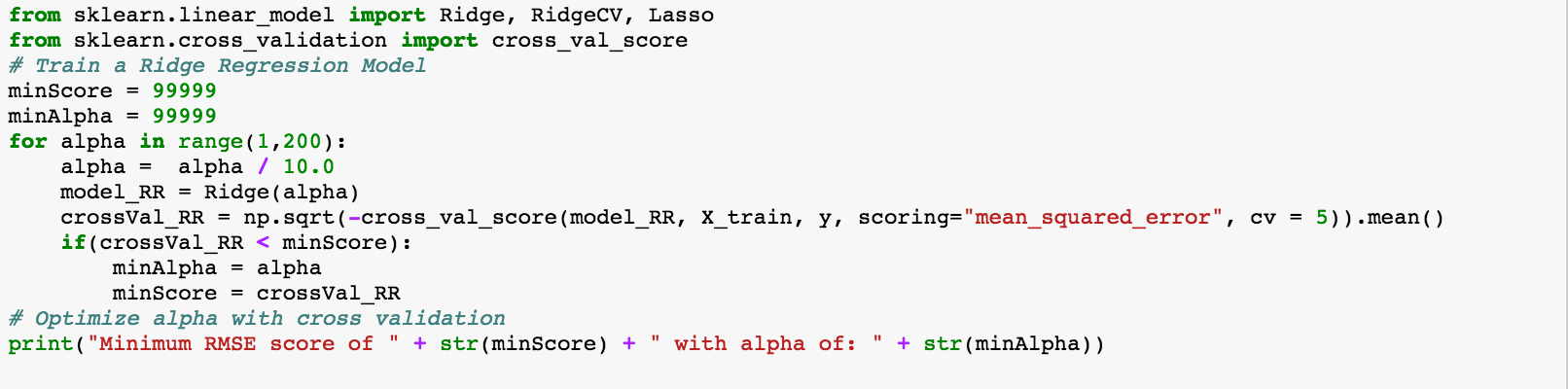
Results from running the algorithm:



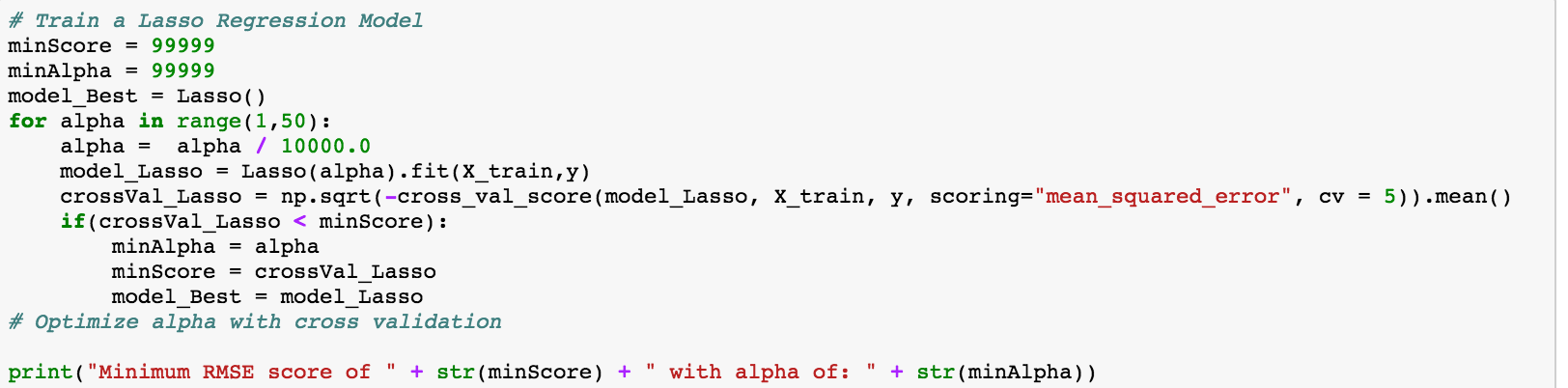


Problem 3: Getting more into Kaggle



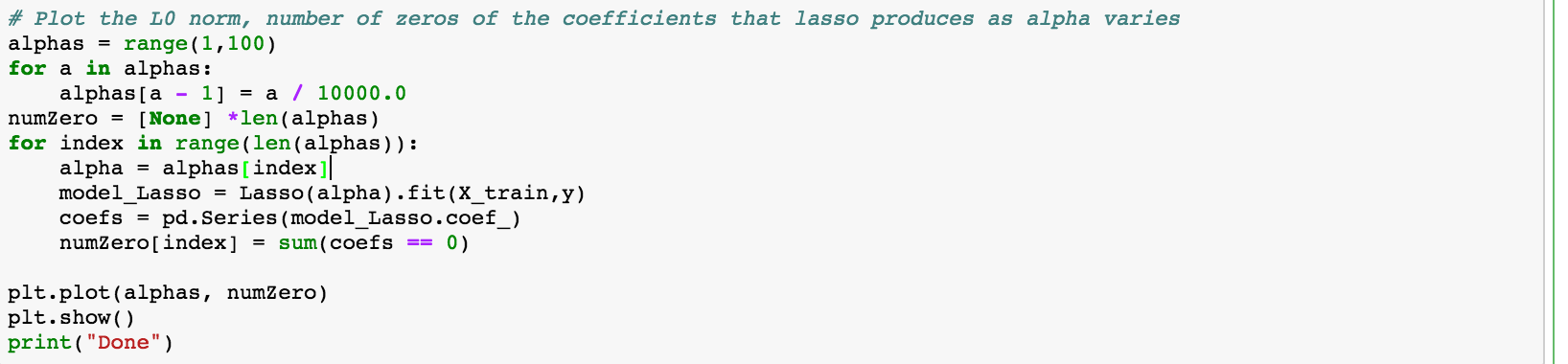


The best score from Ridge Regression and results of RMSE cross validation:../../../../../../Desktop/Screen%20Shot%202017-02-14%20at%204.14.17%20P



The best score from Lasso Regression and results of RMSE cross validation:

../../../../../../Desktop/Screen%20Shot%202017-02-14%20at%204.14.46%20P



Plot of coefficients equal to zero as alpha varies:

