Ridge and Lasso Regression for Beam Calculations

The Strain dataset for Alex's Dataset was only studied here and not the other datasets, primarily due to lack of access to the geometrical data of each run, for example the void radius and distance from the edges. As discussed later, these geometrical data have been indicative of improving the Ridge and Lasso Regression runs, and incorporating the external datasets (Dataset1 and Dataset2) without the geometry information may not show an improved Regression model, and potentially worse performance.

Dataset	Volume Fraction	Holes_100N	MatA_100N	MatB_100N	Holes_500N	MatA_500N	MatB_500N
Alex	0.1	0.88	0.81	0.63	4.41	4.06	3.14
Alex	0.15	0.76	0.73	0.65	3.82	3.64	3.23
Alex	0.2	1.16	0.95	0.6	5.79	4.74	3.02
Alex	0.25	0.97	0.84	0.62	4.85	4.2	3.1
Alex	0.3	1.11	0.9	0.6	5.53	4.48	3.02
Alex	0.35	1.3	0.96	0.59	6.48	4.82	2.94
Alex	0.4	1.63	1.13	0.57	8.14	5.65	2.86
Alex	0.45	2.68	1.24	0.55	13.38	6.2	2.74
Alex	0.5	1.54	1.03	0.56	7.68	5.17	2.82
Alex	0.55	3.87	1.69	0.5	19.2	8.46	2.51
Alex	0.6	3.08	1.36	0.52	15.41	6.78	2.59
Alex	0.65	7.38	1.54	0.49	36.89	7.72	2.46
Alex	0.7	9.17	1.91	0.46	45.84	9.52	2.32

Figure 1 Raw Data of Displacement Runs

Figure 1 displays the raw data obtained from Abaqus, which organizes the associated maximum displacement result for each combination of void material, volume fraction and initial load. Additionally, physical information about the model such as the number of voids and shortest distance from the void to the beam's edge was obtained as shown below.

Volume Fraction	Number of Largest Holes	Largest Hole Radius	Number of Smallest Holes	Smallest Hole Radius	Shortest Distance to Edge
0.1	8	2.82	8	2.82	2.18
0.15	4	4.88	4	4.88	2.62
0.2	8	3.89	8	3.89	1.11
0.25	7	4.76	7	4.76	1.49
0.3	4	6.9	4	6.9	1.85
0.35	4	7.46	4	7.46	1.29
0.4	9	5.31	9	5.31	0.94
0.45	4	8.46	4	8.46	0.28
0.5	5	7.97	5	7.97	0.78
0.55	7	5.64	32	2	0.36
0.6	4	9.01	10	2	0.5
0.65	4	9.01	18	2	0.5
0.7	5	9.76	15	1	0.25

Figure 2 Geometry Data of Voids

The independent variables were chosen to be the variables that the user changed for every combination of a beam run, which in this case, was the Volume Fraction, whether the voids were not filled with a material, Material A and B composition, loads I.E 100N or 500N, and the corresponding hole geometry for each volume fraction, as shown in Figure 2. The response variable was chosen to be the displacement values, which were results from the Abaqus runs. The two datasets would not be easy to input as is for the Regression models to interpret, so a Python script was developed to reorganize the data such that all the independent variables were on the leftmost columns, and only the response variable was on the rightmost column.

Volume Fraction	Load	Youngs_Modulus	Poisson_Ratio	Number of Largest Holes	Largest Hole Radius	Number of Smallest Holes	Smallest Hole Radius	Shortest Distance to Edge	Displacement_Values
0.1	100	0	0	8	2.82	8	2.82	2.18	0.88
0.15	100	0	0	4	4.88	4	4.88	2.62	0.76
0.2	100	0	0	8	3.89	8	3.89	1.11	1.16
0.25	100	0	0	7	4.76	7	4.76	1.49	0.97
0.3	100	0	0	4	6.9	4	6.9	1.85	1.11
0.35	100	0	0	4	7.46	4	7.46	1.29	1.3
0.4	100	0	0	9	5.31	9	5.31	0.94	1.63
0.45	100	0	0	4	8.46	4	8.46	0.28	2.68
0.5	100	0	0	5	7.97	5	7.97	0.78	1.54
0.55	100	0	0	7	5.64	32	2	0.36	3.87
0.6	100	0	0	4	9.01	10	2	0.5	3.08
0.65	100	0	0	4	9.01	18	2	0.5	7.38
0.7	100	0	0	5	9.76	15	1	0.25	9.17
0.1	100	10000	0.34	8	2.82	8	2.82	2.18	0.81
0.15	100	10000	0.34	4	4.88	4	4.88	2.62	0.73
0.2	100	10000	0.34	8	3.89	8	3.89	1.11	0.95
0.25	100	10000	0.34	7	4.76	7	4.76	1.49	0.84
0.3	100	10000	0.34	4	6.9	4	6.9	1.85	0.9
0.35	100	10000	0.34	4	7.46	4	7.46	1.29	0.96
0.4	100	10000	0.34	9	5.31	9	5.31	0.94	1.13
0.45	100	10000	0.34	4	8.46	4	8.46	0.28	1.24
0.5	100	10000	0.34	5	7.97	5	7.97	0.78	1.03
0.55	100	10000	0.34	7	5.64	32	2	0.36	1.69
0.6	100	10000	0.34	4	9.01	10	2	0.5	1.36
0.65	100	10000	0.34	4	9.01	18	2	0.5	1.54
0.7	100	10000	0.34	5	9.76	15	1	0.25	1.91

Figure 3 Reorganized Datasheet

A heatmap of the independent and response variables above was produced, and it can be shown that there is a somewhat strong correlation between the Displacement values and the hole geometry data; thus, supporting the hypothesis to only run Alex's dataset as it has the complete information for the Regression models.

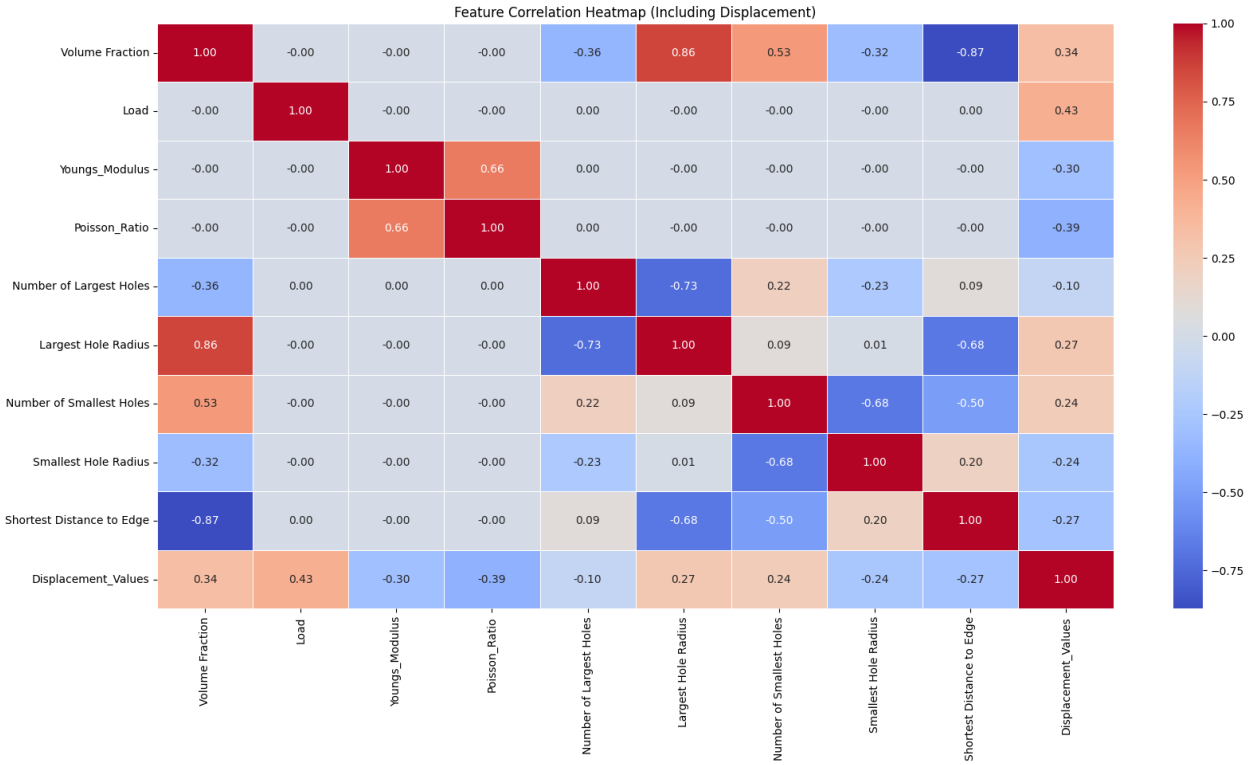

Figure 4 Heatmap of Variables for Displacement Run

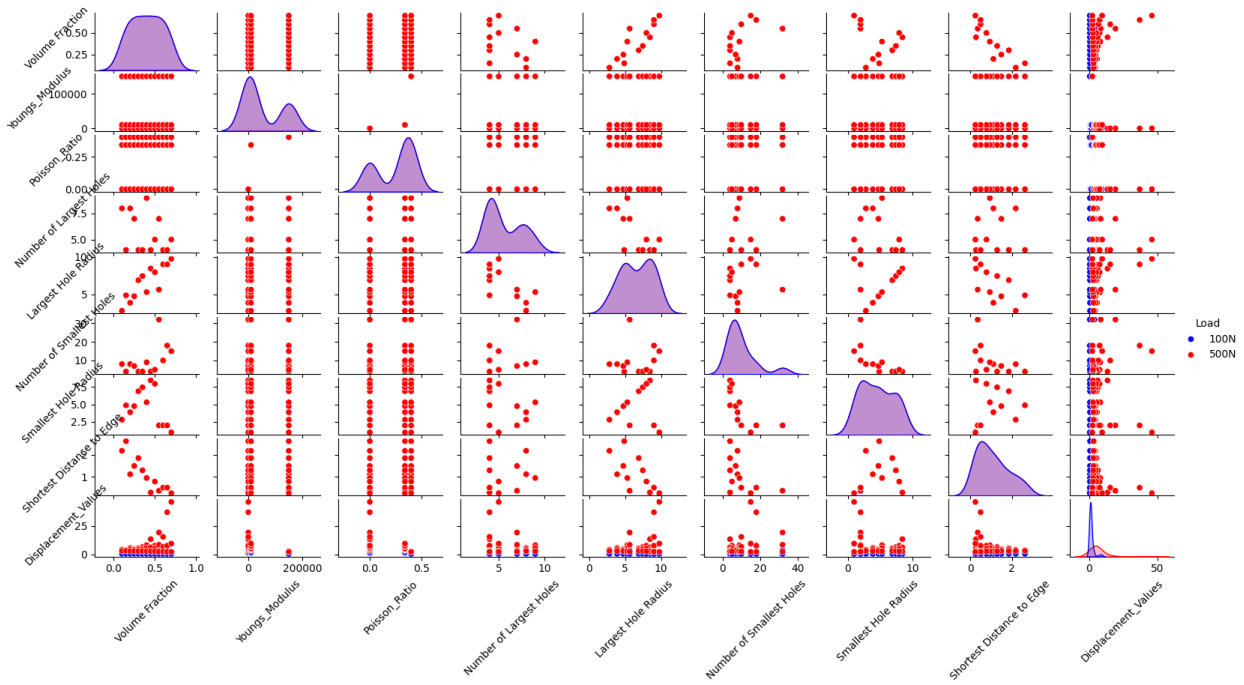

Figure 5 Pairplot Trend of Variables

At this point, data visualization of the dataset was needed, to check for outliers, any skewness and the distribution. The distribution of the displacement values was created below.

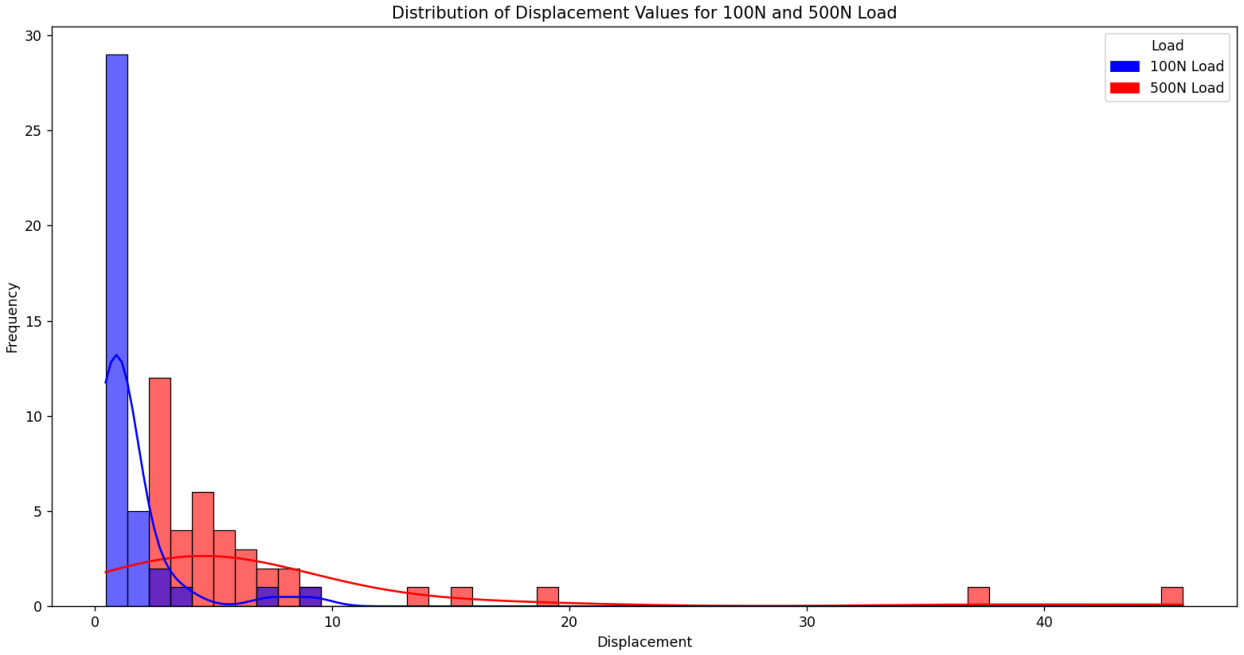

Figure 6 Distribution of Displacement results

Both distributions for the 100N and 500N load appear to be right skewed, with the 100N load having a close spread of values, while the 500N load has a broader spread, which is expected as greater loads induce greater displacement values. The long tail of the 500N load indicates there may be some outliers but this is not unexpected as the 0.70 volume fraction runs can induce unstable results, with the Geometry Data of Voids datasheet that should capture this trend for the model. However, these outliers should be kept in the experiment since they may contain useful information about performance of the beams with large voids. As expected, when calculating the **Skewness** and **Kurtosis** of Displacement, the results were 4.153 and 19.521 respectively. This information is critical, seeing as the Regression models here work best under a near Gaussian distribution. This indicates that a transformation of either Log or a Box Cox of the data is needed at some point, but an initial pass of the data was performed first for the Regression Models.

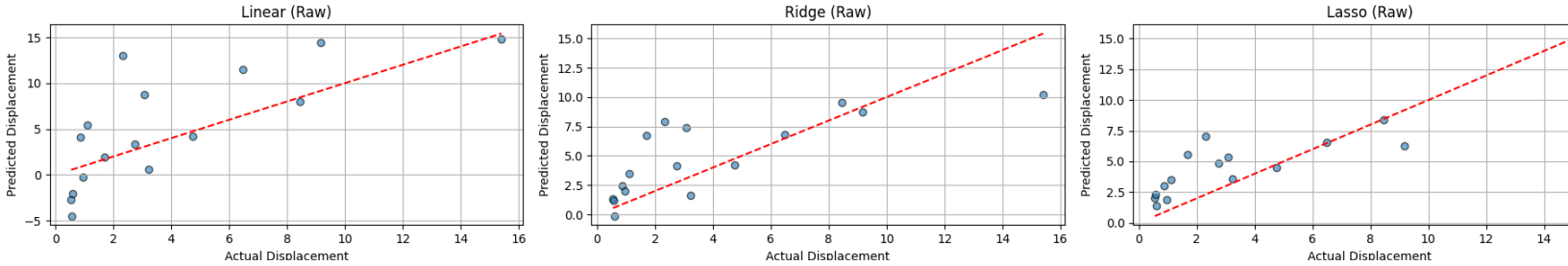

Figure 7 Visuals of Regression Runs

```
================================================
 🔹 Best Alpha for Ridge Regression: 79.06043210907701
 🔹 Best Alpha for Lasso Regression: 1.5264179671752334
================================================

**Model Performance Summary (Train & Test)**
=============================================================================
  Metric  Linear Regression (Train)  Linear Regression (Test)  Ridge Regression (Train)  Ridge Regression (Test)  Lasso Regression (Train)  Lasso Regression (Test)
     MSE                 27.832061                 17.652315                 35.844755                 6.666659                 38.435521                 6.932704
     MAE                  3.447583                  3.218153                  2.718528                 1.934017                  2.728439                 2.188207
R2 Score                  0.508592                 -0.098715                  0.367119                 0.585054                  0.321376                 0.568495
=============================================================================
```

Figure 8 Metric results of the Regression Runs, MSE, MAE, R2 Score

The models don't perform too well with the raw data, and most likely due to the high Skewness and Kurtosis values as the R squared values for Ridge and Lasso were 0.585 and 0.568 respectively.

The data was then transformed, with a Log and Box Cox separately, to assess model performance. The new data histogram plots below indicate a better distribution which would most likely be suitable for the Regression models.

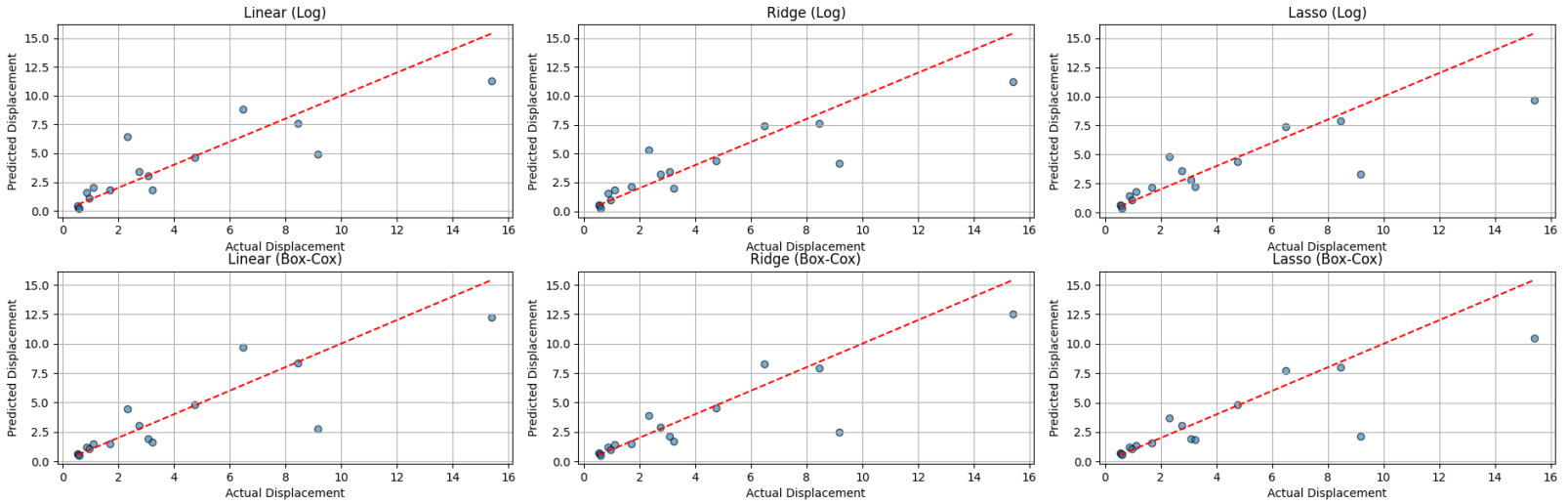

Figure 10 Plots of improved Regression models

```
🔹 Best Alpha Values for Each Transformation:
Raw: Ridge Alpha = 19.306977, Lasso Alpha = 0.868511
Log: Ridge Alpha = 1.526418, Lasso Alpha = 0.022230
Box-Cox: Ridge Alpha = 0.868511, Lasso Alpha = 0.022230
```

Figure 9 Optimized Alpha values for each Regression run

As it can be seen, the model's performance significantly improves, with the highest R squared value being 0.7812 for the Ridge Regression model after performing a Log transformation. The independent variables were scaled using RobustScaler to mitigate the impact of outliers while preserving important data trends. Unlike StandardScaler, which normalizes based on mean and standard deviation, RobustScaler scales data using interquartile range making it more resilient to extreme values. Testing confirmed that StandardScaler led to a lower R^2 score, suggesting that extreme displacement values may contribute more noise than useful predictive information.

Model Performance Comparison

	MSE			MAE			R2 Score		
Model Transformation	Lasso	Linear	Ridge	Lasso	Linear	Ridge	Lasso	Linear	Ridge
Box-Cox	5.0547	4.4112	3.8823	1.1915	1.2130	1.0865	0.6854	0.7254	0.7584
Log	4.8597	3.8923	3.5149	1.2648	1.2887	1.1670	0.6975	0.7577	0.7812
Raw	7.7416	17.6523	7.4925	2.0766	3.2182	2.0437	0.5181	-0.0987	0.5337

Figure 12 Model performance table of each Regression model

```
🔹 Log Transformed Skewness: 0.933
🔹 Log Transformed Kurtosis: 0.843
🔹 Box-Cox Transformed Skewness: 0.045
🔹 Box-Cox Transformed Kurtosis: -0.861
```

Figure 11 Updated Skew and Kurtosis values after transformation

The alpha parameter in Ridge and Lasso regression is optimized using GridSearchCV, which utilizes cross-validation (CV) to find the best regularization strength. A logarithmically spaced range of alpha values (10^{-3} to 10^{3}) is tested, where the dataset is split into training and testing (20/80) folds. The model is trained for each α value, and its performance is evaluated using the R^2 score. The best alpha is the one that maximizes R squared across the CV folds, balancing

model complexity and generalization. Higher alpha values increase regularization by shrinking coefficients, reducing overfitting, while lower alpha values allow more flexibility but risk capturing noise.

Regarding the limited dataset, visually that the model predicted data converges the most near the lower displacement values but in the regions with higher displacement values, the model consistently diverged from the ideal displacement line. What this indicates initially is that more runs with the 500N loads is needed to better capture the performance of the displacement values at this region.

To ensure that the geometry data of the voids is essential to keep as independent variables, another run was devised that excluded the void geometry variables but kept all other aspects identical (such as data transformation).

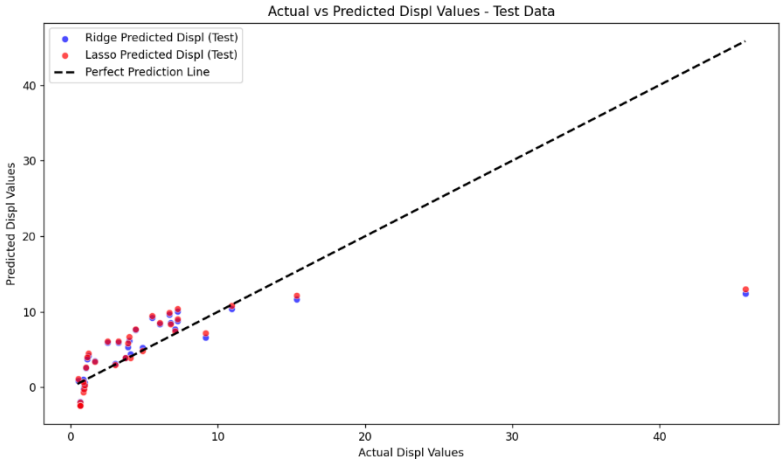

Figure 13 Visual of test run for geometry variables

Both Ridge and Lasso Regression performed not as well as the run with the geometry inputs. The R squared for Ridge and Lasso were 0.389 and 0.401 respectively.

This strongly indicates that geometry data of the voids provides useful and critical input for the Regression models to build upon and interpret.

```
Ridge Regression - Best Alpha: 0.46415888336127775
Ridge Regression - Test Data MSE: 39.37833123914268
Ridge Regression - Test Data R-squared: 0.389905802482135
Ridge Regression - Training Data MSE: 15.002251492187927
Ridge Regression - Training Data R-squared: 0.5206909904879748

Lasso Regression - Best Alpha: 0.001
Lasso Regression - Test Data MSE: 38.61399927715754
Lasso Regression - Test Data R-squared: 0.4017477084316958
Lasso Regression - Training Data MSE: 14.909002163797016
Lasso Regression - Training Data R-squared: 0.5236702261881618
```

Figure 14 Poor performance results excluding the geometry results

For the Stress runs, a similar approach to the Displacement was performed. After looking at the histogram of the resulting displacement values, the skewness isn't as dramatic as the Displacement runs but it is still a positive skew.

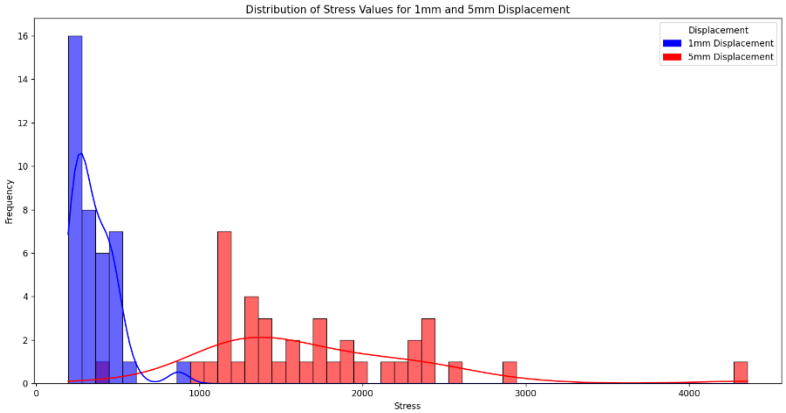

Figure 15 Distribution of Stress values

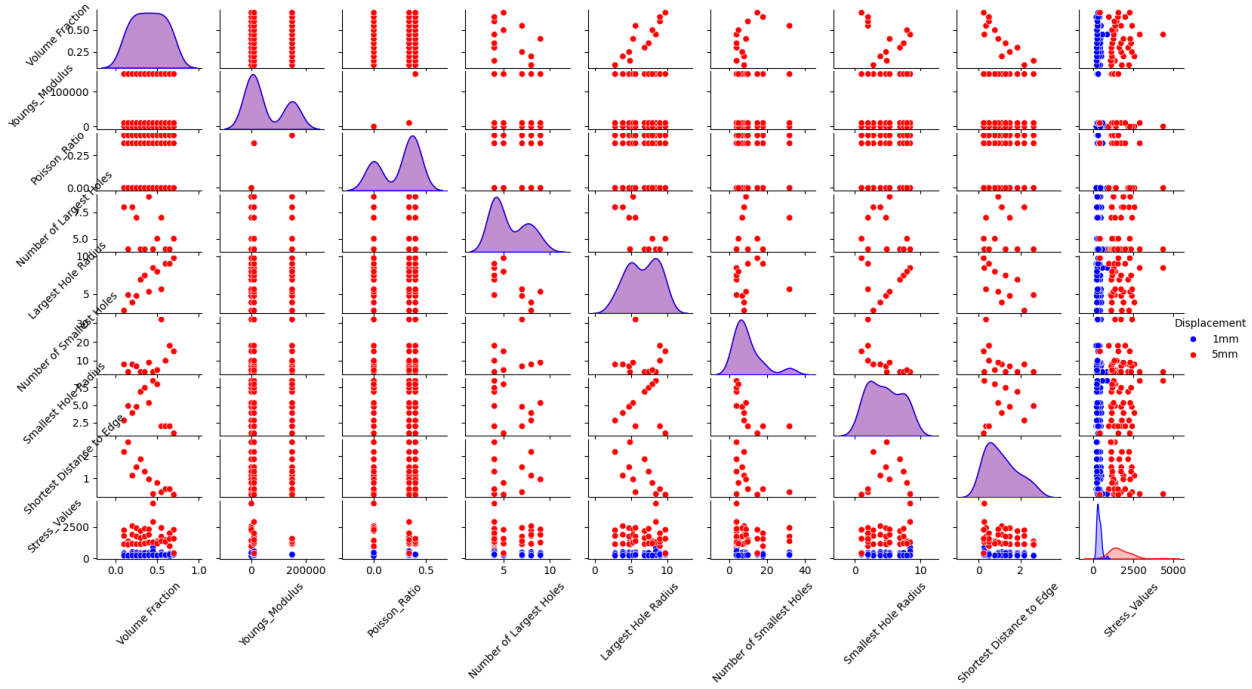

Figure 17 Pairplot of Stress data run

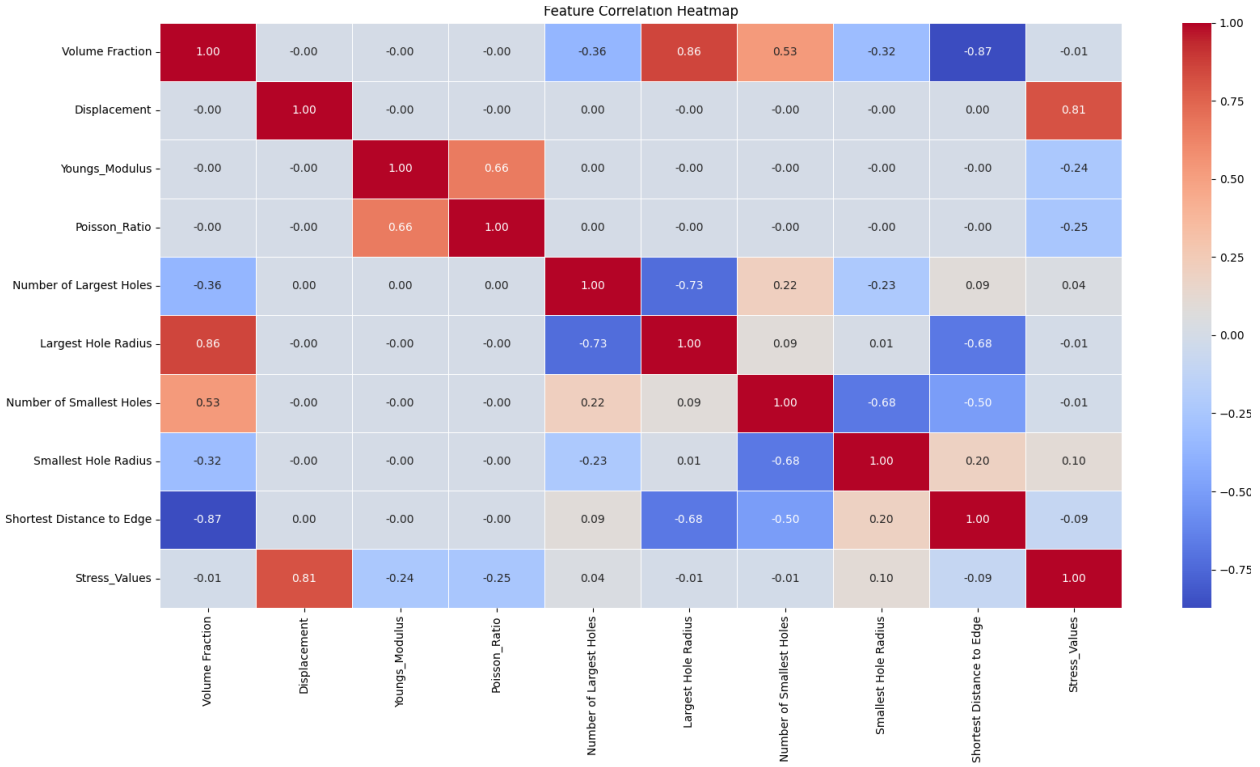

Figure 16 Heatmap of Stress run data

Although the correlation heatmap suggests a weaker relationship between geometry variables and stress, these variables were retained in the Regression Models because the correlation measure only captures **linear relationships** and may not reflect **nonlinear interactions**.

Additionally, LASSO regression naturally reduces the impact of irrelevant features by shrinking their coefficients toward zero. If geometry truly had no predictive value, its coefficients would be minimized, but keeping these features ensures that any **potential higher-order effects** are not discarded.

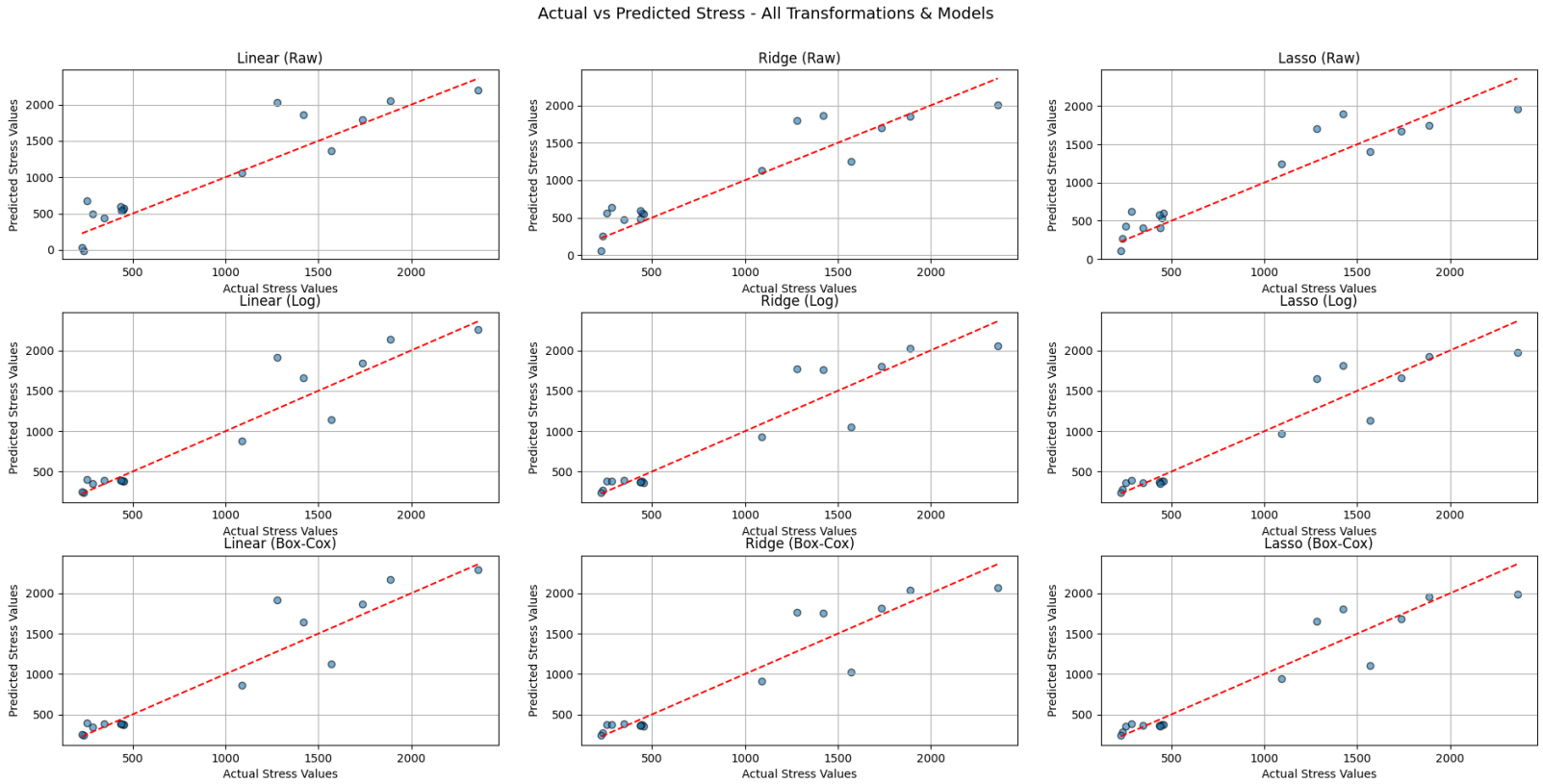

Figure 18 Visual plots of results of running Regression model on different transformed data

For the data preprocessing, the stress values were log transformed/Box Cox transformed due to the data skewness and RobustScaler was used for the independent variables.

	MSE			MAE			R2 Score		
Model	Lasso	Linear	Ridge	Lasso	Linear	Ridge	Lasso	Linear	Ridge
Transformation									
Box-Cox	44924.0238	52559.7786	52125.7592	153.5804	157.7998	167.5528	0.9046	0.8883	0.8893
Log	43690.7274	50388.0860	51013.0675	149.6888	153.8893	163.7944	0.9072	0.8930	0.8916
Raw	52376.4353	78433.8322	62706.3955	182.2544	216.5209	195.0086	0.8887	0.8334	0.8668

Figure 19 MSE, MAE, R2 score for each transformation/model run

```
Best Alpha Values for Each Transformation:
Raw: Ridge Alpha = 6.250552, Lasso Alpha = 33.932218
Log: Ridge Alpha = 0.868511, Lasso Alpha = 0.009541
Box-Cox: Ridge Alpha = 0.868511, Lasso Alpha = 0.005429
```

Figure 20 Optimized alpha scores for each transformation

However, even by running the Regression models with the raw data, there appears to be a relatively strong linear relationship between the independent and response variables, as the R squared values for the Linear Regression runs are high for the raw data, but the best performing Model with the highest R squared value and lowest MSE/MAE value ended up being the Lasso Regression model, Log transformed, with an R squared value of 0.9072, MAE value of 149.68 and MSE value of 43690.

```
Original Skewness: 1.193
Original Kurtosis: 1.659
Box-Cox Lambda: -0.069
Log Transformed Skewness: 0.078
Log Transformed Kurtosis: -1.445
Box-Cox Transformed Skewness: 0.029
Box-Cox Transformed Kurtosis: -1.475
```

Figure 21 Before and after skewness comparison for transformation

Regarding the limited dataset, visually that the model predicted data converges the most near the lower Stress values but in the regions with higher stress values, the model consistently diverged from the ideal displacement line. What this indicates initially is that more runs with the upper load results is needed to better capture the performance of the stress values at this region, most likely due to the stress results becoming unstable as the volume fraction of the voids increases. More data in this region would be needed for future studies.

Ridge and Lasso Regression for External Dataset Calculations

The dataset consists of fatigue stress-controlled and strain-controlled data, used to analyze fatigue life (Nf). In the stress-controlled dataset, cyclic stress is applied to a material sample, and the number of cycles to failure is recorded.

This dataset includes columns for material properties (elastic modulus, tensile strength, yield strength, Poisson ratio), applied stress, and fatigue life, log transformed, (Nf). The data for the strain-controlled data set, appears to be mostly low cycle fatigue, with cycles being less than 10,000 and the data for the stress-controlled data set, is a mix of low and high cycle fatigue, but with more high cycle fatigue counts than low.

Both datasets contain CSV files with raw stress-strain data over time, which are preprocessed into standardized feature vectors. The first column represents the uniaxial stress or strain, and the second column represents the shear stress or strain.

The goal is to analyze correlations between stress/strain and fatigue life and apply regression models to improve predictive accuracy.

There were about 40 different metal alloys in the dataset, each with having different amplitude runs, so Python code was developed to organize the dataset into Uniaxial, Pure Shear, Proportional and Nonproportional by reading off the slope of each column.

Additionally, the individual metal alloys were broken down and organized to be associated with each .CSV file.

Overall, the independent variables were the material properties, the .CSV file that contains the uniaxial and shear results transposed into a single array, amplitude category and metal alloy name/category and the response variable was the fatigue life. The amplitude category and metal alloy used OneHotEncoding to properly categorize as independent variables for the Regression Models to interpret.

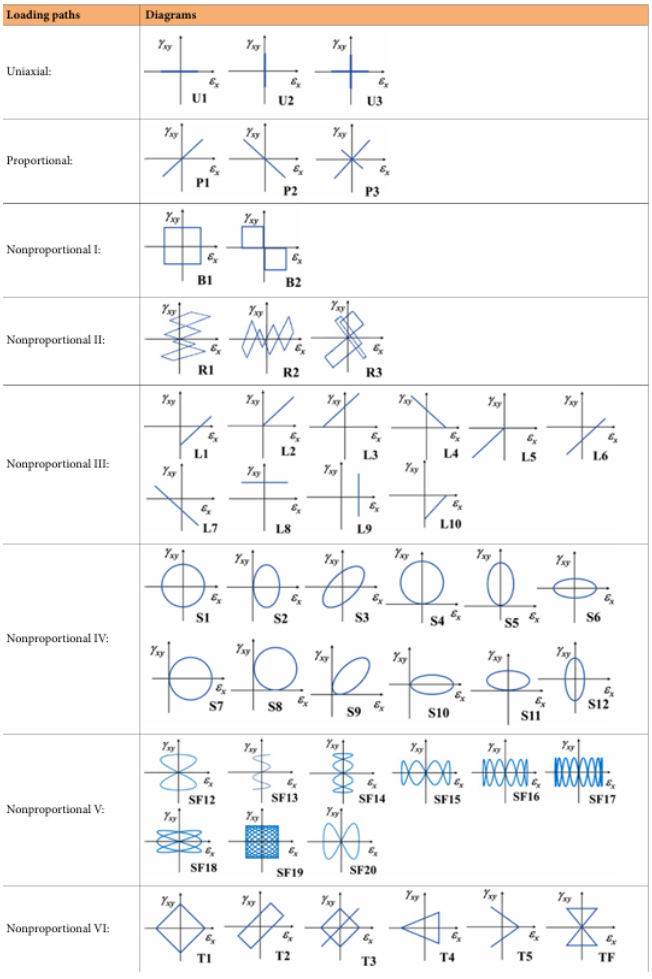

Figure 22 Load visualization for multiaxial fatigue life set performed

Considering how large the datasets are, it was difficult to determine the distribution of the data, or to even evaluate how strong the linear relationship was between the independent and response variables. Eventually a method using Pearson and Spearmans Coefficients was utilized here.

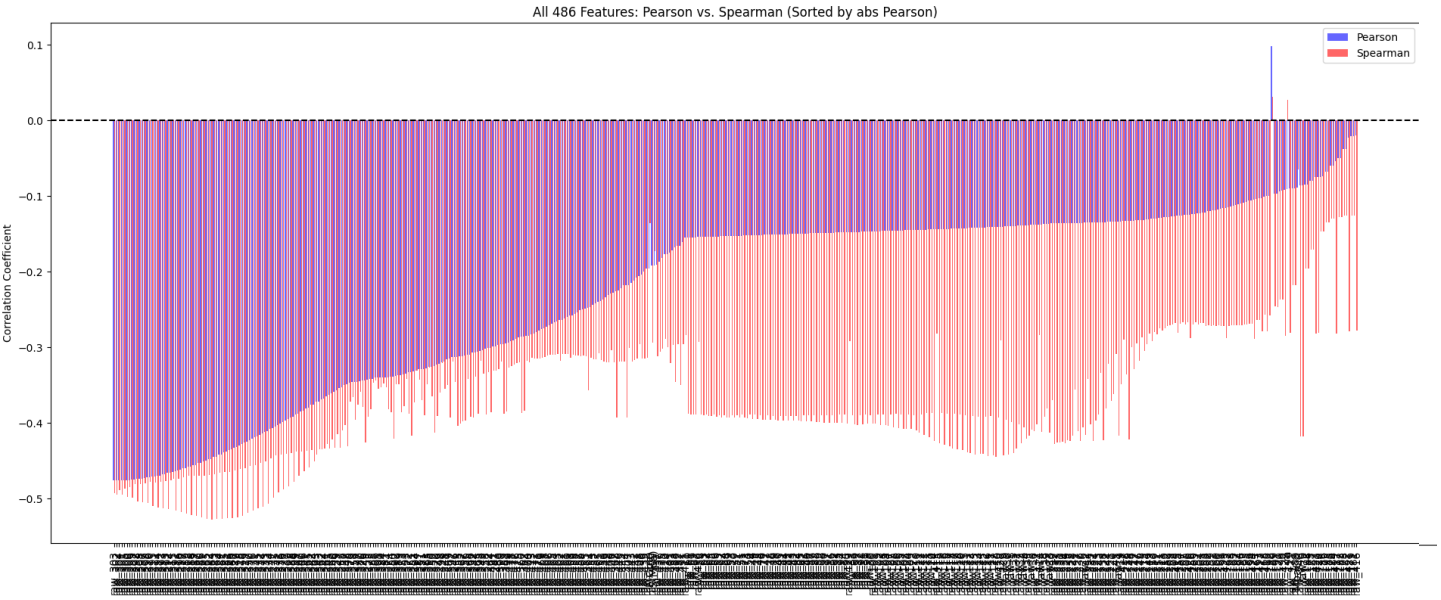

Figure 23 Pearson v Spearman plot for Stress-Fatigue

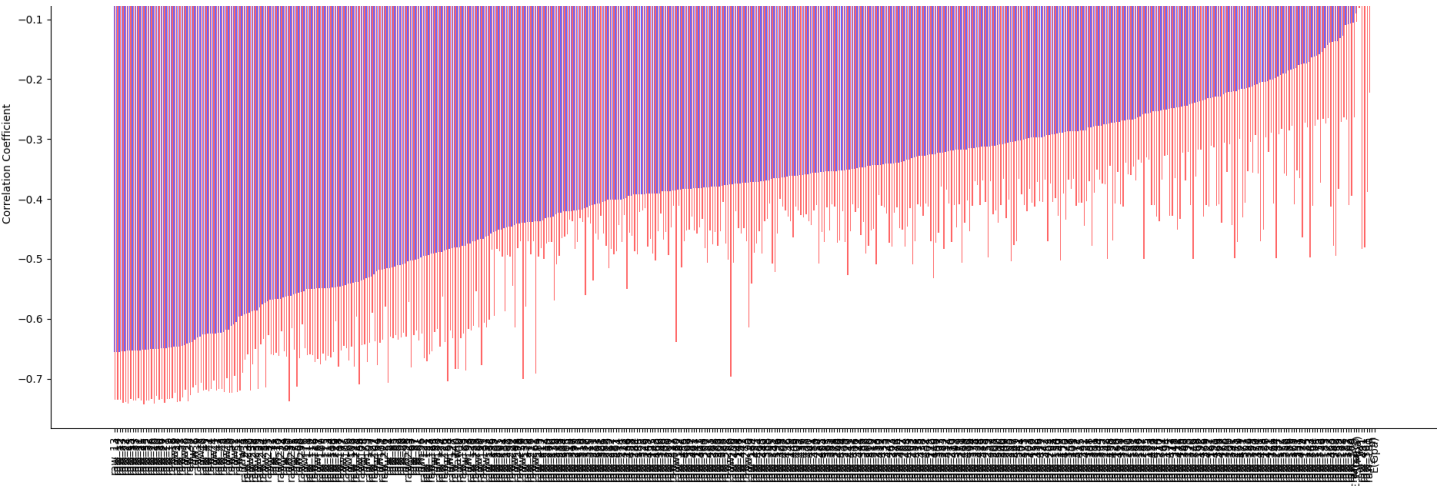

Figure 24 Pearson v Spearman plot for Strain-Fatigue

Above are the correlation plots for the Strain-Fatigue and Stress-Fatigue runs.

These charts were produced by calculating a Pearson/Spearman Coefficient for each variable against the associated fatigue life for each .CSV file.

There are several interesting takeaways from this. For the Stress-Fatigue runs, the overall correlation between the independent variables and response variable is lower (absolute max value of around 0.53) but for the Strain-Fatigue runs, the overall correlation is stronger (absolute max value of around 0.74). Across both charts, the Spearman Coefficient is **consistently** higher than the Pearson Coefficient, indicating that the relationship between the independent and response variables may be more monotonic than linear. From these charts prior to running any of

the Regression models, it is expected to obtain a better performance of the Strain-Fatigue Regression runs than then Stress-Fatigue Regression runs.

As an attempt to investigate the distribution of both independent and response variables, several histograms were developed. First, the Strain .CSV files were all analyzed for overall skewness. It was found that the majority of the .CSV files had very low skew values, with only two .CSV files having an absolute skew value greater than 1. The Stress .CSV files performed similarly; the charts are below.

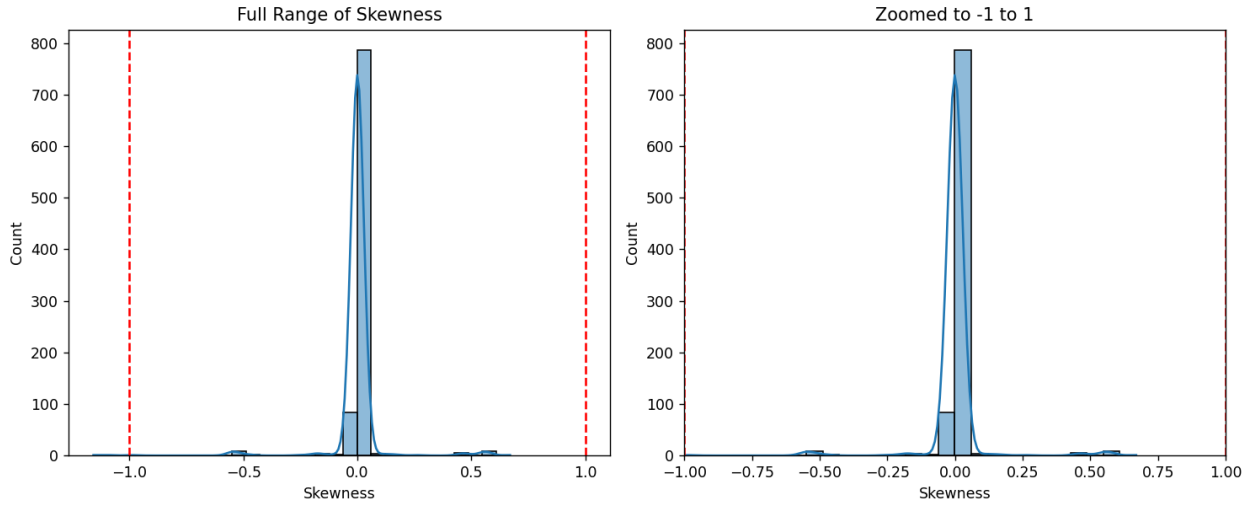

Figure 25 Strain .CSV file skew histogram

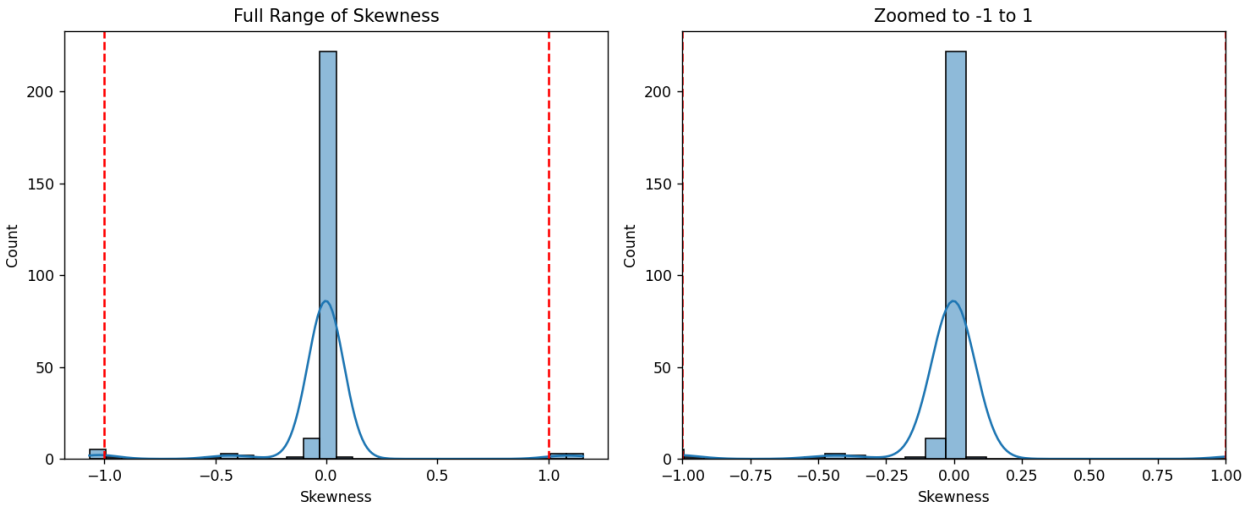

Figure 26 Stress.CSV file skew histogram

Initially at a first glance, it appears that for the majority of the .CSV files, there appears to be no significant skewness for the Stress/Strain values. The best results from the Regression models

were from only standardizing the raw data as inputs for the model, the results (R squared, MSE, MAE). Any further transformation such as log, of the raw data made the model results worse.

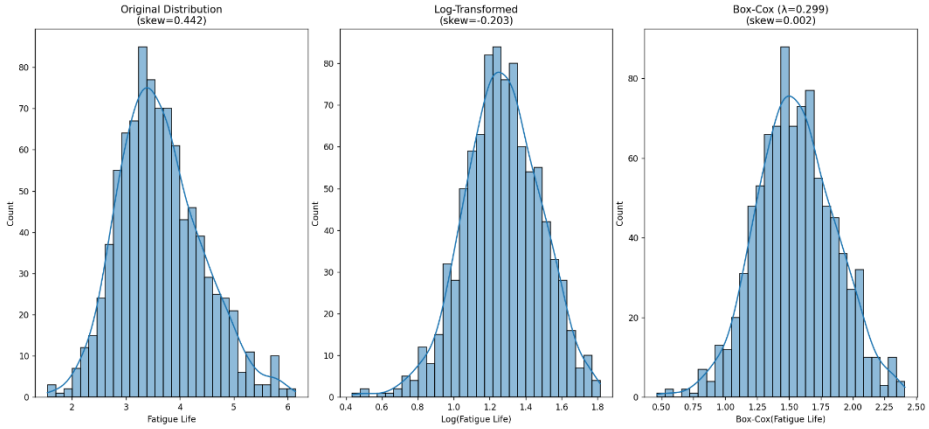

Figure 27 Strain Fatigue skew diagram

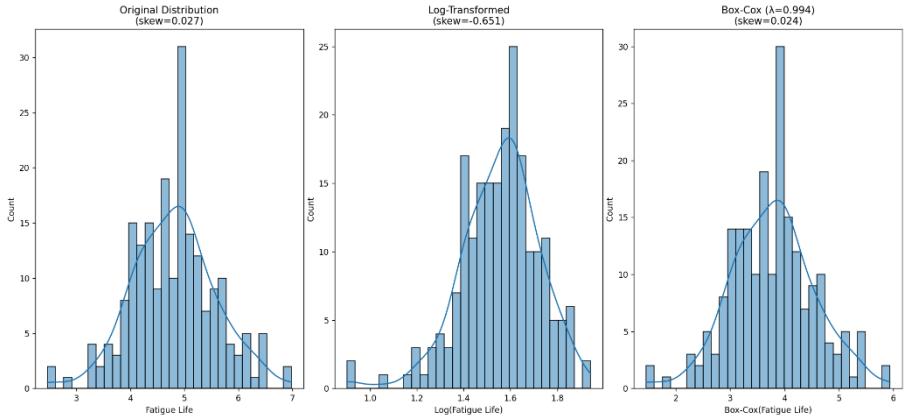

Figure 28 Stress Fatigue skew diagram

The fatigue skew displays a similar distribution, with the skewness being near 0, with no obvious bias. Any transformation here yielded similar results for the most part, however, after experimenting with the Regression models, the best R squared/MSE/MAE results were with the raw fatigue data. This is reasonable, given the fatigue is already log transformed, and doing another log transformation would interrupt any linear relationship that already existed with the independent variables.

For the Stress-Fatigue run, several attempts were made to better prepare the data, such as log transforming, adjusting any negative and zero values, performing a Box-Cox transformation, however, the best result yielded was with only Standardizing the independent variables. Again, these runs optimize the best alpha score and perform a similar evaluation as to the Beam Regression Models earlier discussed where the dataset is split into training and testing (20/80) folds as well.

```
Linear Regression Results:
  R² Score: -17435025550648.3965
  MAE: 527946.4480
  MSE: 11706531438209.1680

Ridge Regression Results:
  Best Alpha: 0.1
  R² Score: 0.3583
  MAE: 0.4795
Lasso Regression Results:
  Best Alpha: 0.001
  R² Score: 0.3273
  MAE: 0.4973
  MSE: 0.4517
```

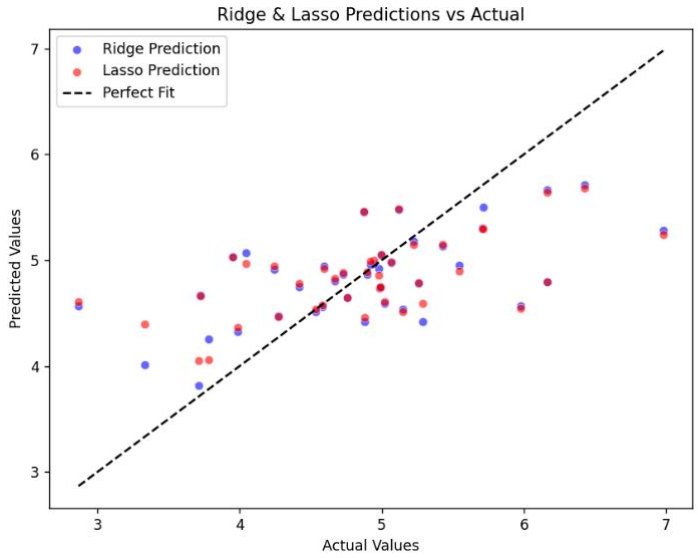

Figure 29 Stress-Fatigue Regression Results

The best R squared value was 0.3583 with the Ridge Regression model. Visually, there is a poor fit of the data with the ideal fit line, strongly indicating that the relationship between the Strain-Life Fatigue dataset is not strong enough for the Regression models to pick up on.

Now, when running the Strain-Fatigue dataset, the results are much better, with an R squared value of 0.7765 for the Lasso Regression model. The independent variables were only Standardized, since this also produced the best results

```
Ridge Regression Results:
  R² Score: 0.7635
  MAE: 0.2683
  MSE: 0.1495

Lasso Regression Results:
  R² Score: 0.7765
  MAE: 0.2706
  MSE: 0.1413
```

Best Ridge Alpha: 2.1554

Best Lasso Alpha: 0.001

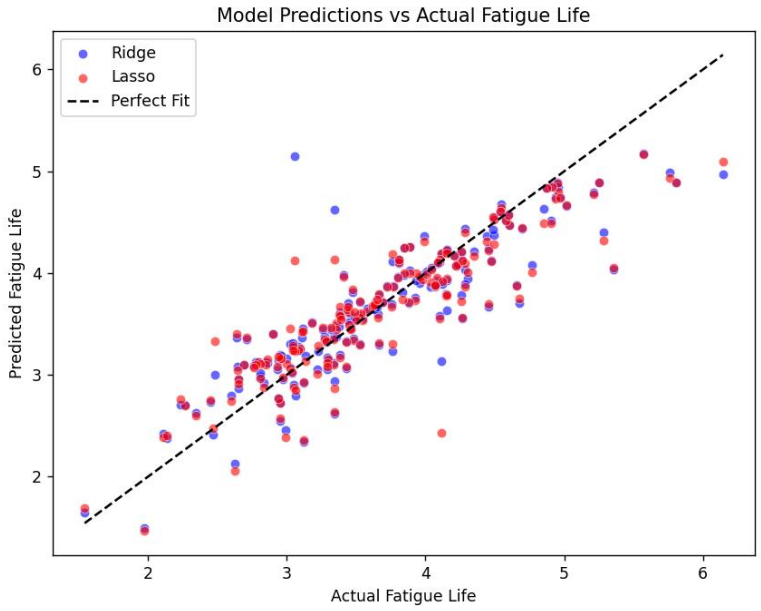

Figure 30 Strain-Fatigue Regression Results

The Regression model performances align well with the earlier Spearman/Pearson correlation analysis. For the strain-fatigue dataset, the regression models capture a strong enough linear relationship that about 78% of the variance in the fatigue life is explained, leaving roughly 22% unexplained. In contrast, the stress-fatigue dataset exhibits a much weaker linear relationship, with only about 36% of the variability being accounted for by the model, meaning approximately 64% of the variance remains unexplained. This indicates that, at least within these datasets, strain is a more reliable predictor of fatigue life than stress.

Another interesting note about the dataset size here, the Stress-Fatigue runs only consisted of a variety of Steel and Aluminum alloys, so the total CSV file dataset was limited to 254 .CSV files but the Strain-Fatigue runs consisted of 915 .CSV files as it consisted of approximately 40 different metal alloys such as Titanium or Inconel. Whether it is by coincidence or not, the Strain Fatigue dataset had the most datapoints and performed the best, and the Stress Fatigue dataset lacked in datapoints comparatively and performed the worst.

The lower predictive performance of the stress-fatigue models suggests that the measurement devices may not have captured key stress-related variables needed for strong correlation detection. This may have resulted in missing or underrepresented features critical for accurately modeling fatigue life. In contrast, the strain-fatigue dataset provided more comprehensive information, leading to significantly better predictive accuracy. The higher R^2 values in strain-based models indicate that strain may inherently be a more reliable predictor of fatigue life compared to stress for multiaxial runs.