CS 6375
ASSIGNMENT 1

Names of students in your group:
James Hooper
Hritik Panchasara

Number of free late days used: 0

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

For Lin Reg model, metrics, & preprocessing data split
https://scikit-learn.org/stable/
For preprocessing outliers
https://machinelearningmastery.com/how-to-use-statistics-to-identify-outliers-in-data/
For graphing data
https://seaborn.pydata.org/
https://matplotlib.org/
For data manipulation
https://pandas.pydata.org/
https://numpy.org/

Part 1 of Logs: Tuning the Gradient Descent Model for best Parameters

Part 1: Gradient Descent
Parameters Used:
State: 4
Standard Deviations for Outlier Removal: 4
Learning Rate: 1e-06
Iterations: 10000
Coefficients:
[0.11928264595830723, 0.10363271225307923, 0.08608352708856432, -0.215357831473916, 0.07307720765591492, 0.008023474755255763, 0.017534096270358745, 0.1772805781677691]
Train Accuracy:
Mean Squared Error: 95.32110584763825
R^2 Value: 0.48183252833057943
Test Accuracy:
Mean Squared Error: 73.99337096249627
R^2 Value: 0.7023277579923453

Part 1: Gradient Descent
Parameters Used:
State: 4
Standard Deviations for Outlier Removal: 4
Learning Rate: 1e-06
Iterations: 50000
Coefficients:
[0.1164429638989648, 0.09989996580909491, 0.07953100319261813, -0.20423221706825073, 0.1878986115809362, 0.008408278807793712, 0.015409554989700978, 0.1767887289439327]
Train Accuracy:
Mean Squared Error: 94.64063519917634
R^2 Value: 0.4866136421663171
Test Accuracy:
Mean Squared Error: 75.63947079780894
R^2 Value: 0.7001860528658519

Part 1: Gradient Descent
Parameters Used:
State: 4
Standard Deviations for Outlier Removal: 5
Learning Rate: 1e-07
Iterations: 10000
Coefficients:
[0.11100745157694752, 0.08506917134090101, 0.0761917402946925, -0.07833755133887067, 0.02274130432140025, -0.0037499970287976536, 0.009144790638065158, 0.1136446189035902]
Train Accuracy:
Mean Squared Error: 112.82941279769462
R^2 Value: 0.11931835530706003
Test Accuracy:
Mean Squared Error: 108.48393625502074
R^2 Value: 0.3135650357530867

Part 1: Gradient Descent
Parameters Used:
State: 4
Standard Deviations for Outlier Removal: 6
Learning Rate: 1e-06
Iterations: 10000
Coefficients:
[0.11904324516221096, 0.10317937744265458, 0.08977279735113078, -0.20592047153596504, 0.062367879265086384, 0.009651498872081577, 0.015990455657314155, 0.11455208839988694]
Train Accuracy:
Mean Squared Error: 103.7274013858669
R^2 Value: 0.39324962756991466
Test Accuracy:
Mean Squared Error: 145.0591307896913
R^2 Value: 0.1395237842831809

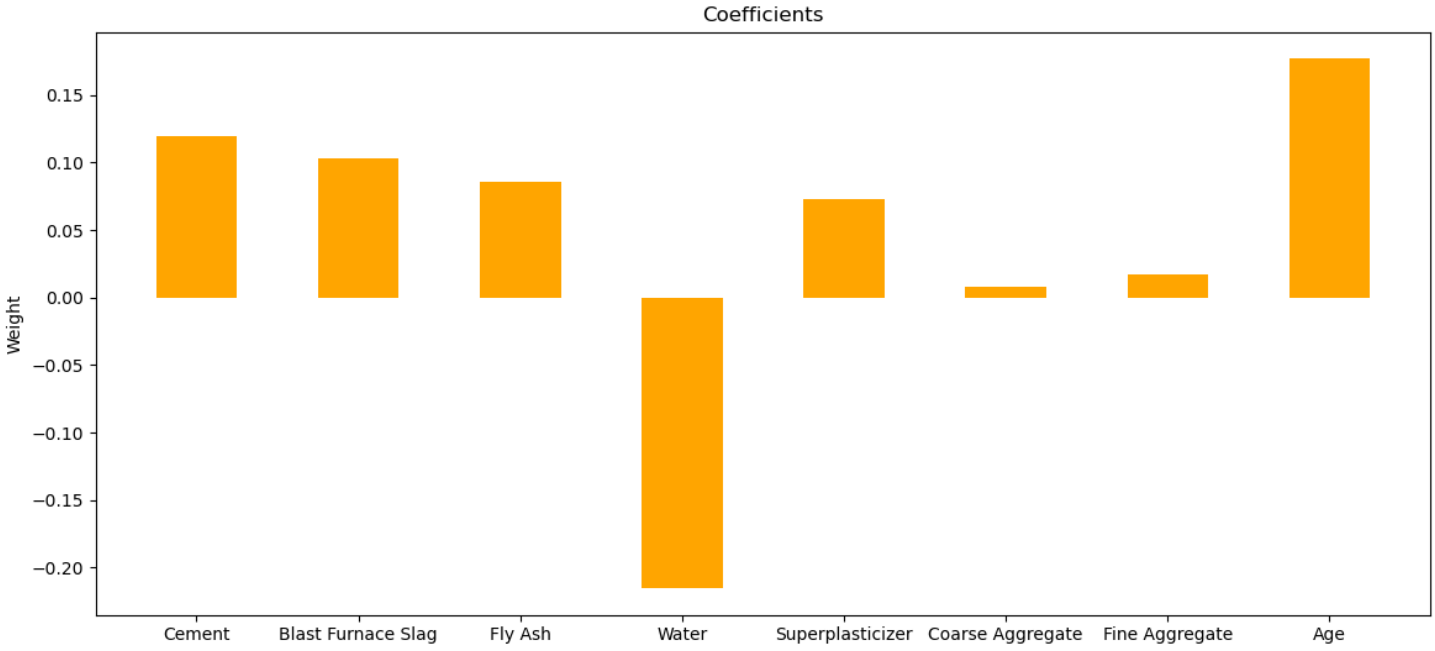
Part 1: Gradient Descent
Parameters Used:
State: 4
Standard Deviations for Outlier Removal: 4
Learning Rate: 1e-07
Iterations: 50000
Coefficients:
[0.119161753064774, 0.10295411710771953, 0.08713680516620799, -0.2047882658796393, 0.053381075015879714, 0.00654162279427648, 0.01723002903981783, 0.17517799721087254]
Train Accuracy:
Mean Squared Error: 95.56714696258463
R^2 Value: 0.46738519134778667
Test Accuracy:
Mean Squared Error: 73.10088446898298
R^2 Value: 0.6977969553984149

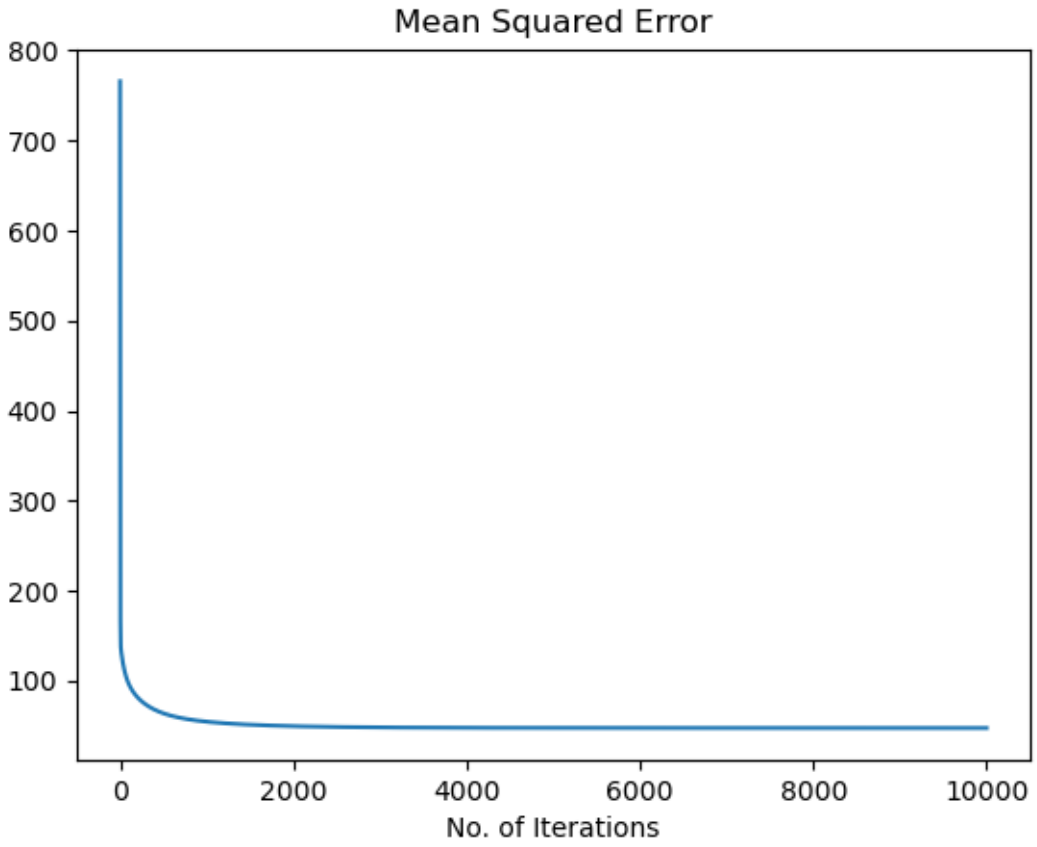
Part 1: Gradient Descent
Parameters Used:
State: 4
Standard Deviations for Outlier Removal: 4
Learning Rate: 1e-07
Iterations: 10000
Coefficients:
[0.11016760209380368, 0.08532314820045674, 0.07579903023284083, -0.08217984586694421, 0.0233342403592557, -0.004984507833790944, 0.01114501018071378, 0.1344575555092756]
Train Accuracy:
Mean Squared Error: 109.34223488277999
R^2 Value: 0.14830063658573323
Test Accuracy:
Mean Squared Error: 79.05739969986148
R^2 Value: 0.5400015785294462

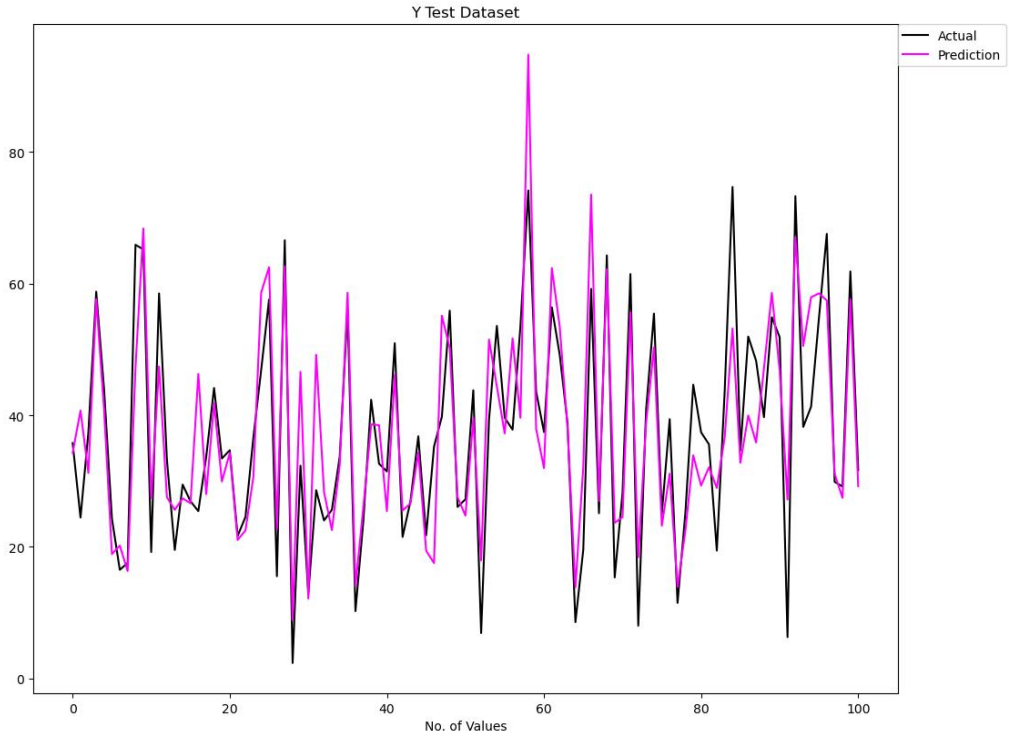
Part 1: Gradient Descent
Parameters Used:
State: 4
Standard Deviations for Outlier Removal: 4
Learning Rate: 1e-07
Iterations: 100000
Coefficients:
[0.11928260069631458, 0.10363260776891696, 0.08608353812264709, -0.21535677958316563, 0.0730769098334549, 0.008023351072429961, 0.017534043892905422, 0.17728039057195577]
Train Accuracy:
Mean Squared Error: 95.32110807948871
R^2 Value: 0.4818313958245929
Test Accuracy:
Mean Squared Error: 73.9933003302369
R^2 Value: 0.7023273881994512

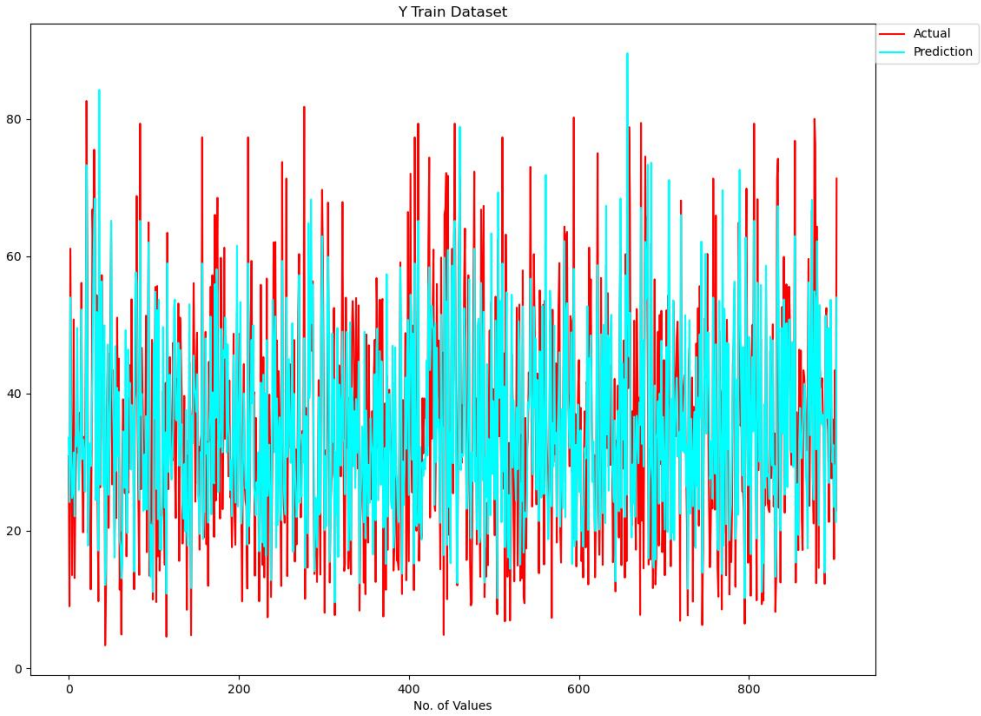
- The states are kept consistent for proper comparison of the parameters.
- Overall, the best tuned trial was the first one shown with the following parameters:
    - Standard Deviations for Outlier Removal: 4
    - Learning Rate: 1e-06
    - Iterations: 10000

Here are the plots for the trial with the best parameters found.

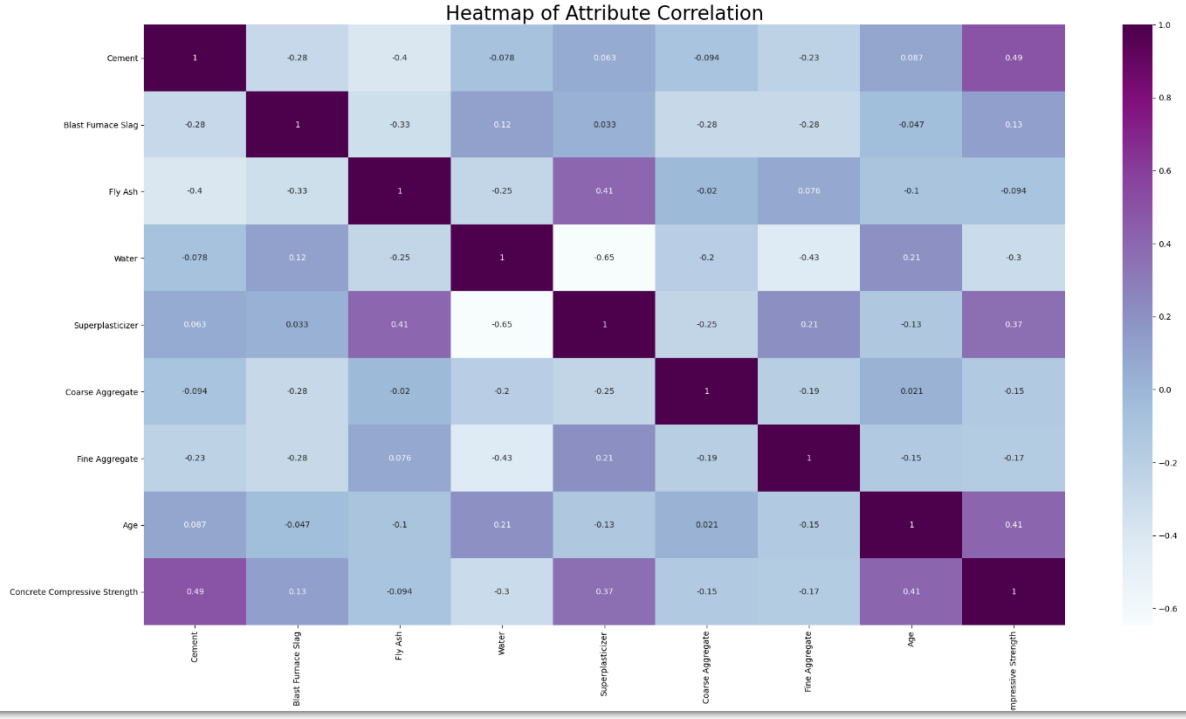


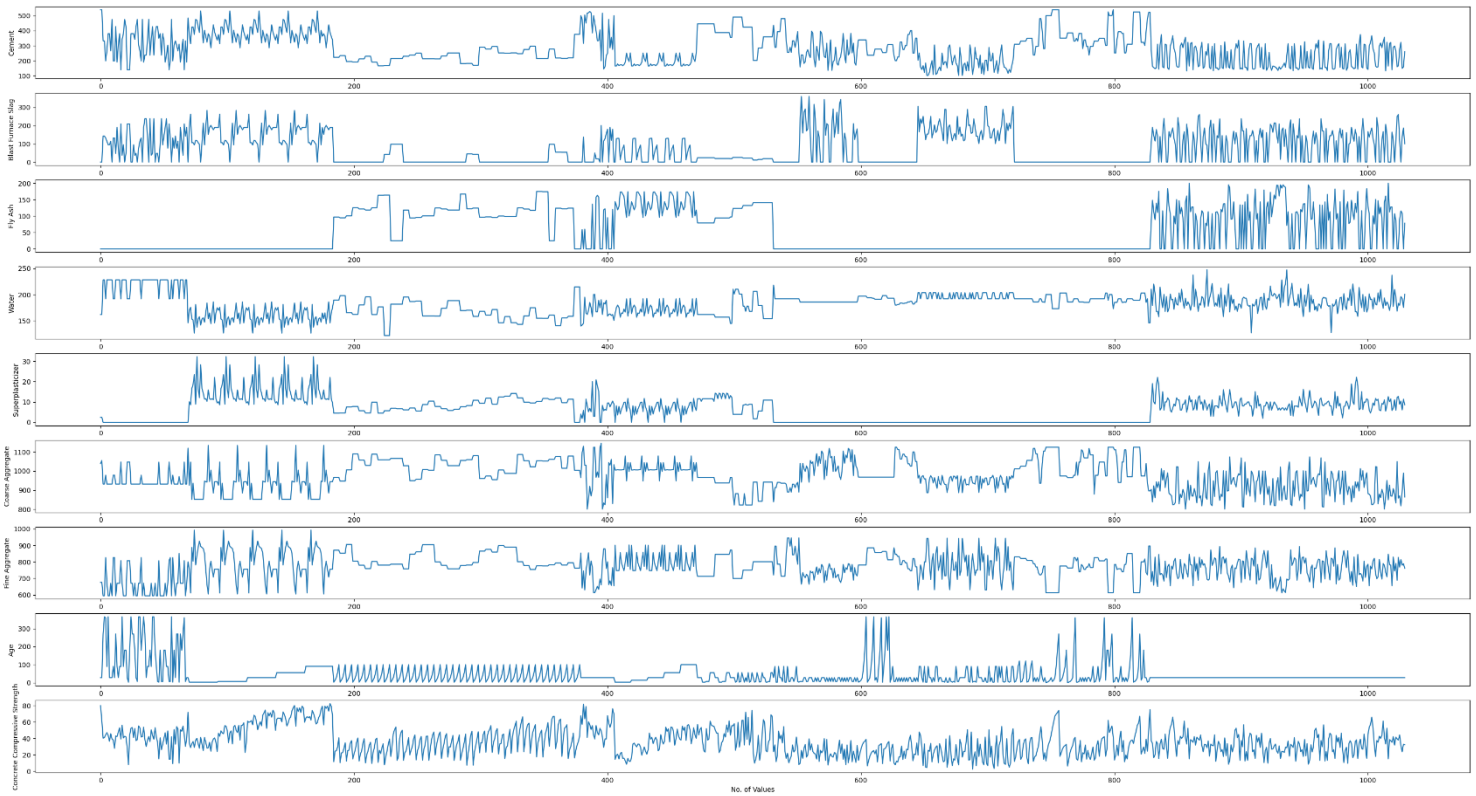



Y Test Dataset
Actual
Prediction
0
20
40
60
80
0
20
40
60
80
100
No. of Values


Y Train Dataset
Actual
Prediction
0
20
40
60
80
0
200
400
600
800
No. of Values

Here are the plots for the overall dataset after removing outliers.





Input/Output Attribute Values