

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

R² Value: 0.48183252833057943

Test Accuracy:

Mean Squared Error: 73.99337096249627

R² 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

R² Value: 0.4866136421663171

Test Accuracy:

Mean Squared Error: 75.63947079780894

R² 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

R² Value: 0.11931835530706003

Test Accuracy:

Mean Squared Error: 108.48393625502074

R² 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

R² Value: 0.39324962756991466

Test Accuracy:

Mean Squared Error: 145.0591307896913

R² Value: 0.1395237842831809

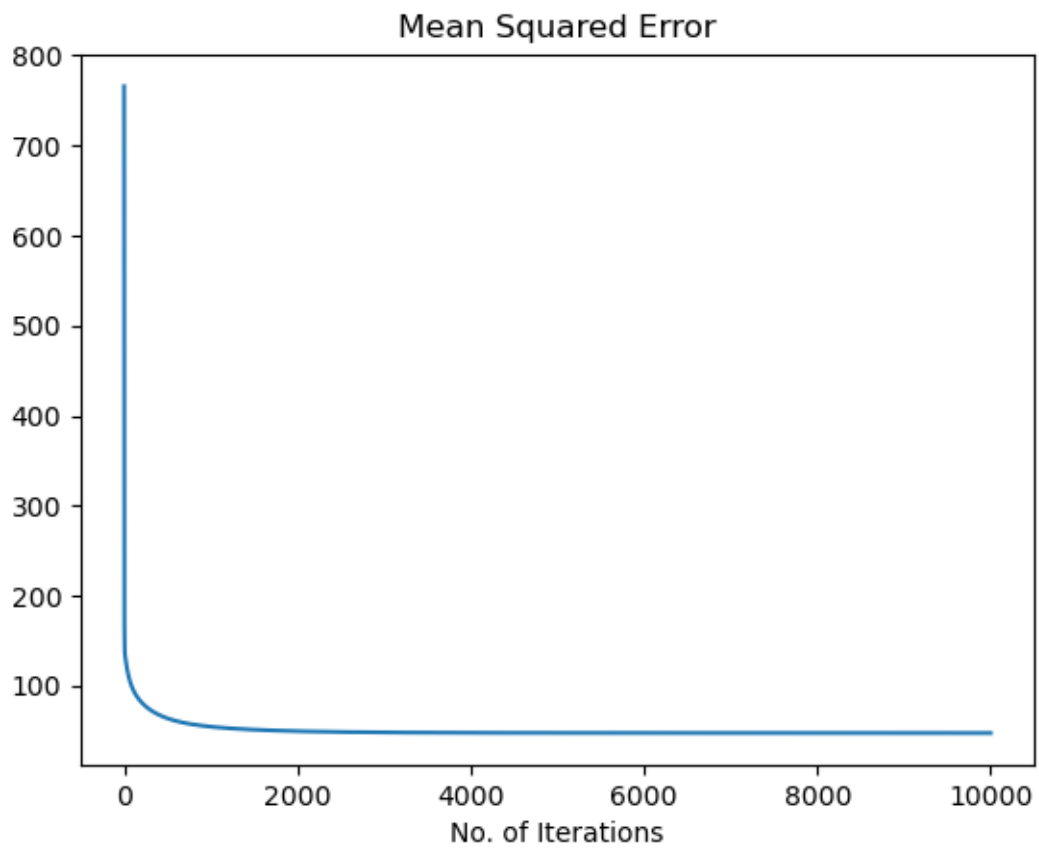
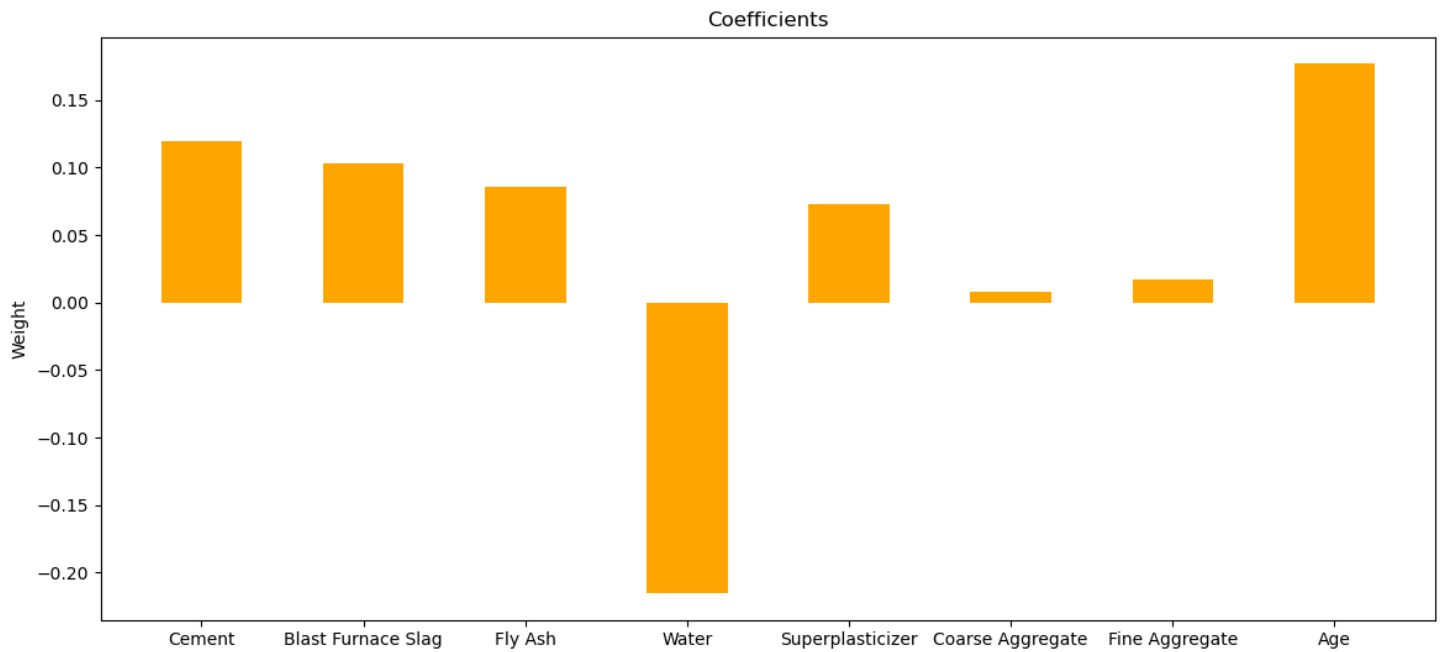
Part 1: Gradient Descent
Parameters Used:
State: 4
Standard Deviations for Outlier Removal: 4
Learning Rate: 1e-07
Iterations: 50000
Coefficients:
[0.119161753064774, 0.10295411710771953, 0.08713680516620799, -0.2047882658796393,
0.053381075015879714, 0.00654162279427648, 0.01723002903981783, 0.17517799721087254]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

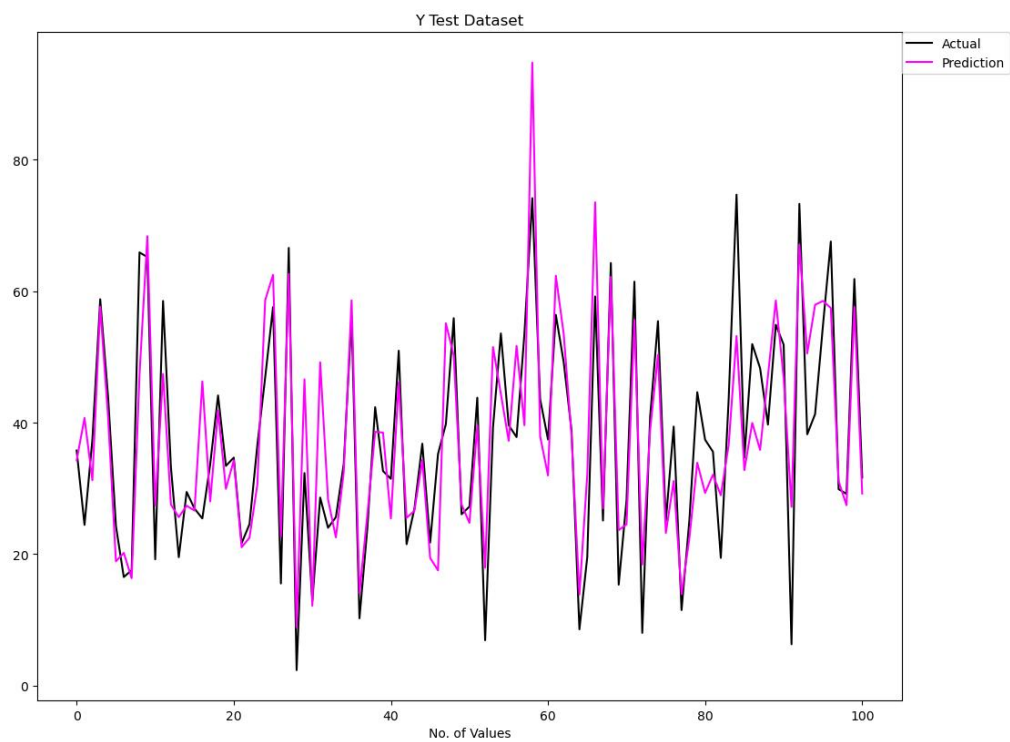
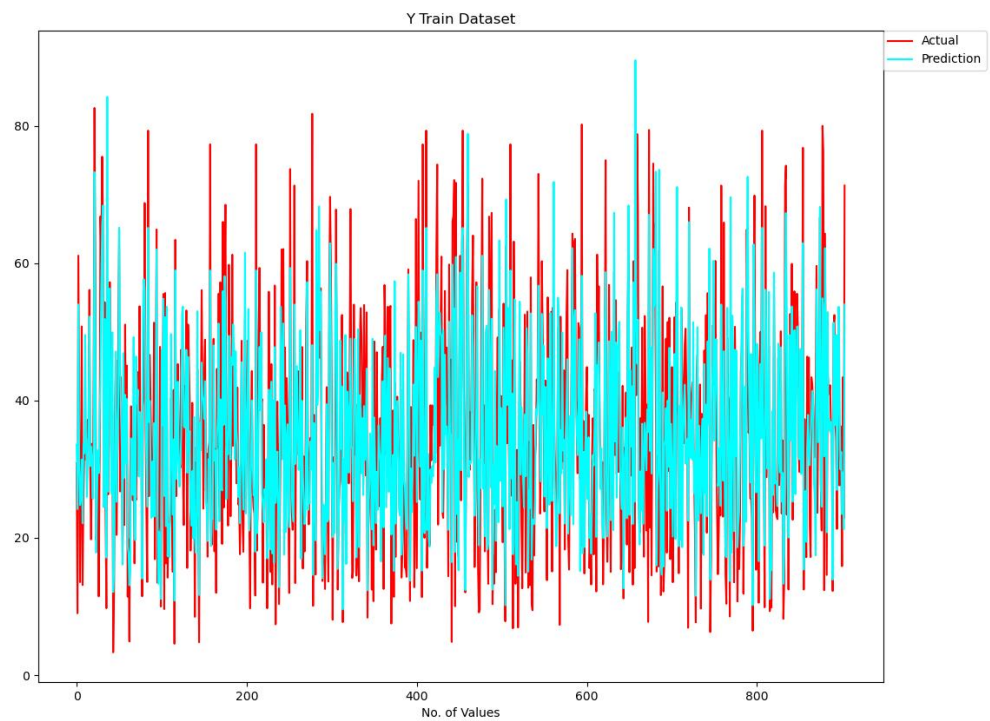
Part 1: Gradient Descent
Parameters Used:
State: 4
Standard Deviations for Outlier Removal: 4
Learning Rate: 1e-07
Iterations: 10000
Coefficients:
[0.11016760209380368, 0.08532314820045674, 0.07579903023284083, -0.08217984586694421,
0.0233342403592557, -0.004984507833790944, 0.01114501018071378, 0.1344575555092756]
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

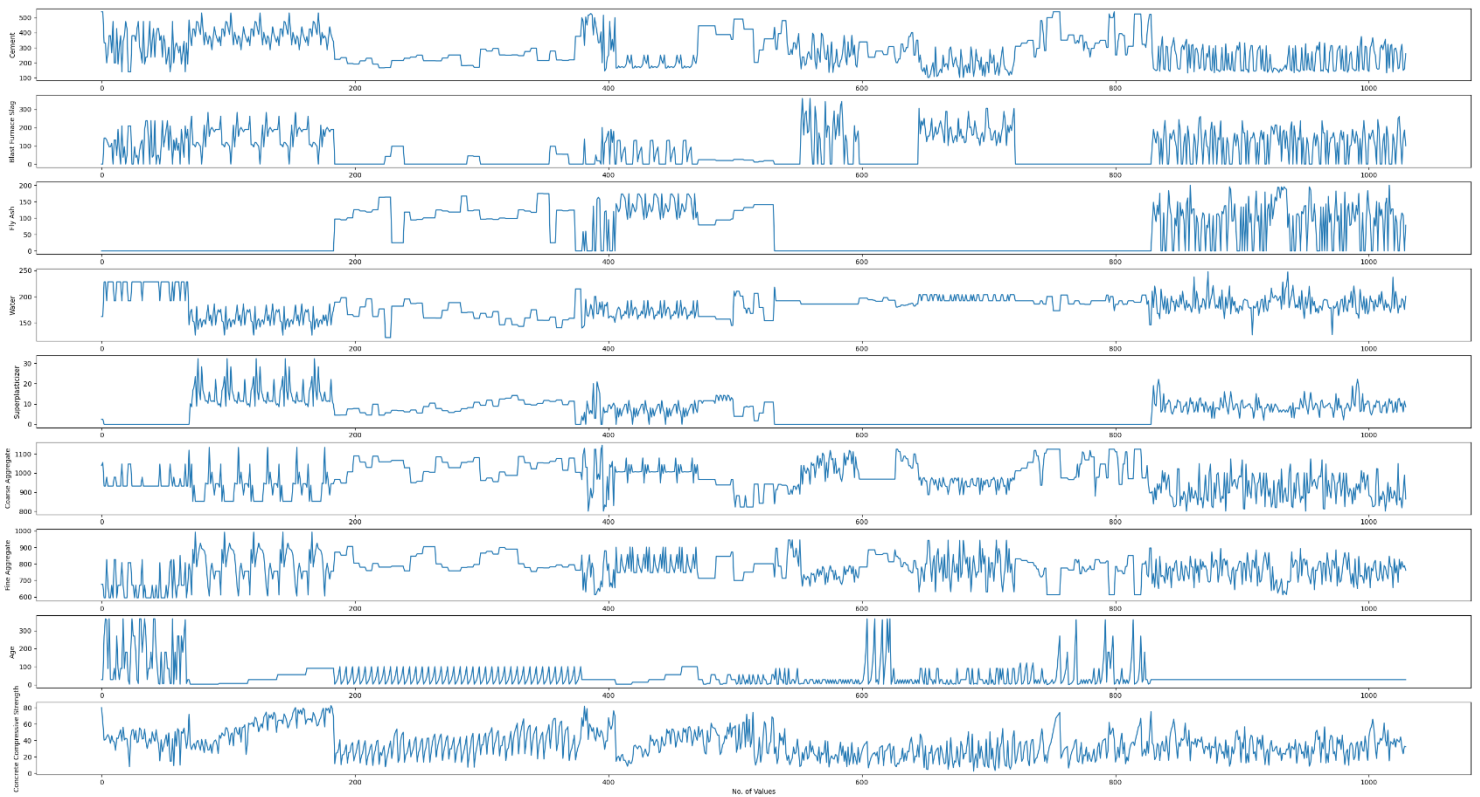
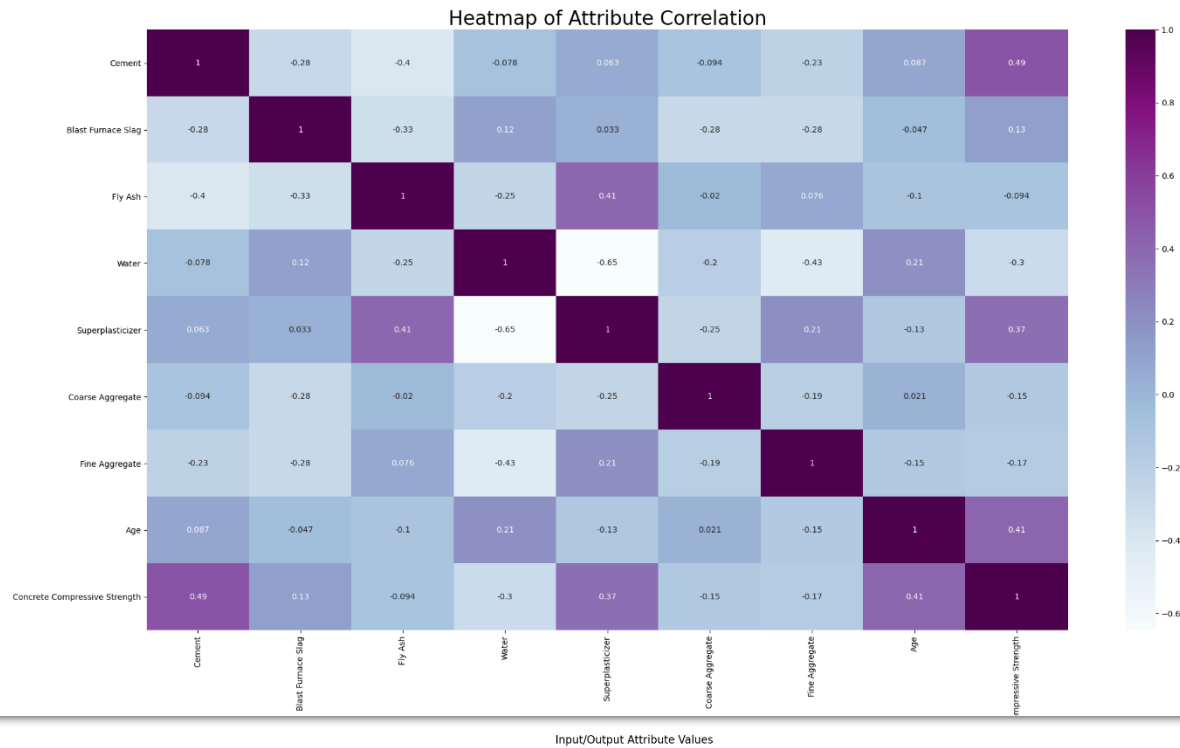
- The states are kept consistent for proper comparison of the parameters.
- Standard Deviations value of 6 is basically removing no rows. Where a value of 2 is removing almost all rows. Overall, a value of 4 is the sweet point for this data set.
- Overall, the best tuned trial was the first one shown with the following parameters:
 - o Standard Deviations for Outlier Removal: 4
 - o Learning Rate: 1e-06
 - o Iterations: 10000

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





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



Part 2 of Logs: Comparing Gradient Descent & Sklearn Linear Regression Models

- In this part of the logs we are going to compare results from the two models over the same respective states. The state can be seen as the seed for repeatable datasets that come from the train-test-split sklearn function. We will be covering states 0 through 4.
- The graphs displayed for part1.py and part2.py are the overall about the same. Figure 1 is the Heatmap. Figure 2 is the Input/Output Attribute Values plots, Figure 3 is the Y train Dataset comparing the actual versus predicted, and Figure 4 is the Y test Dataset comparing the actual versus predicted. For part1.py, Figure 5 is the graph of the Mean Squared Error and Figure 6 is the Weight Coefficients bar graph. For part2.py, Figure 5 is the Weight Coefficients bar graph. Since the sklearn Linear Regression model doesn't have a MSE iteration list/array we cannot graph this value such as the case with the Gradient Descent model.
- The log text should be sufficient for the majority of the comparisons. We will showcase 1 set of images for one comparison of states (for state 4) between the models.
- Since we are showing images for state four with the best parameters. The images for the Gradient Descent model have already been shown above.

Part 1: Gradient Descent

Parameters Used:

State: 0

Standard Deviations for Outlier Removal: 4

Learning Rate: 1e-06

Iterations: 10000

Coefficients:

[0.11742295248024885, 0.10279389644451296, 0.08934583664981885, -0.20347575707900098, 0.06422880733953612, 0.007748876444912093, 0.0157987088918906, 0.17500426642117306]

Train Accuracy:

Mean Squared Error: 94.4887609420638

R² Value: 0.48584041227217545

Test Accuracy:

Mean Squared Error: 82.11126979610026

R² Value: 0.5361910895307058

Part 2: Sklearn Linear Regression

Parameters Used:

State: 0

Standard Deviations for Outlier Removal: 4

Coefficients:

[0.114753 0.09979918 0.08096528 -0.16302816 0.30367326 0.01300971
 0.01628776 0.17435689]

Train Accuracy:

Mean Squared Error: 93.82637192635477

R² Value: 0.49201211043493076

Test Accuracy:

Mean Squared Error: 81.42082202348581

R² Test: 0.5381103065959922

Part 1: Gradient Descent

Parameters Used:

State: 1

Standard Deviations for Outlier Removal: 4

Learning Rate: 1e-06

Iterations: 10000

Coefficients:

[0.11814771005570632, 0.10186101382070192, 0.08902474486233834, -0.2107879304545916, 0.06517693067640419, 0.006533276732241287, 0.018708741304711916, 0.17379603949238087]

Train Accuracy:

Mean Squared Error: 92.02946880954119

R² Value: 0.5006581557142622

Test Accuracy:

Mean Squared Error: 103.96175332933761

R² Value: 0.42875598685152794

Part 2: Sklearn Linear Regression

Parameters Used:

State: 1

Standard Deviations for Outlier Removal: 4

Coefficients:

[0.11564438 0.09865819 0.08033023 -0.1703992 0.30779379 0.01149019
0.01905914 0.17307286]

Train Accuracy:

Mean Squared Error: 91.35828102156856

R² Value: 0.506887806836013

Test Accuracy:

Mean Squared Error: 103.35342572082791

R² Test: 0.44815337790065

Part 1: Gradient Descent

Parameters Used:

State: 2

Standard Deviations for Outlier Removal: 4

Learning Rate: 1e-06

Iterations: 10000

Coefficients:

[0.11908850292888186, 0.1030791609966054, 0.0884076391394888, -0.2054135447237058, 0.06684723337446108, 0.005413060104025423, 0.01886905867433992, 0.1731862941313682]

Train Accuracy:

Mean Squared Error: 93.3652653798686

R² Value: 0.4956621387854945

Test Accuracy:

Mean Squared Error: 92.32994505018246

R² Value: 0.5004191819318058

Part 2: Sklearn Linear Regression

Parameters Used:

State: 2

Standard Deviations for Outlier Removal: 4

Coefficients:

[0.11898089 0.10304339 0.08339464 -0.14765264 0.33680822 0.01421988
0.02316308 0.17213721]

Train Accuracy:

Mean Squared Error: 92.54581376527292

R² Value: 0.5033047004230051

Test Accuracy:

Mean Squared Error: 93.3306992706091

R² Test: 0.48624412949162243

Part 1: Gradient Descent

Parameters Used:

State: 3

Standard Deviations for Outlier Removal: 4

Learning Rate: 1e-06

Iterations: 10000

Coefficients:

[0.1175067847672236, 0.10173110724975126, 0.0874998463171048, -0.20282701142941403, 0.06623169113809754, 0.00601403825317352, 0.017944464337931683, 0.17529552671566662]

Train Accuracy:

Mean Squared Error: 95.06928711247815

R² Value: 0.4789135118725113

Test Accuracy:

Mean Squared Error: 76.54639737869816

R² Value: 0.5860052326310177

Part 2: Sklearn Linear Regression

Parameters Used:

State: 3

Standard Deviations for Outlier Removal: 4

Coefficients:

[0.11589608 0.09979966 0.0801185 -0.15407964 0.32915113 0.01292325
0.01991372 0.17441678]

Train Accuracy:

Mean Squared Error: 94.27697897215714

R² Value: 0.48641720889767615

Test Accuracy:

Mean Squared Error: 77.0967452186757

R² Test: 0.587651290387048

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

R² Value: 0.48183252833057943

Test Accuracy:

Mean Squared Error: 73.99337096249627

R² Value: 0.7023277579923453

Part 2: Sklearn Linear Regression

Parameters Used:

State: 4

Standard Deviations for Outlier Removal: 4

Coefficients:

[0.11356717 0.09581532 0.07095776 -0.17284843 0.39780161 0.01153119
0.01460035 0.17576968]

Train Accuracy:

Mean Squared Error: 94.167898054493

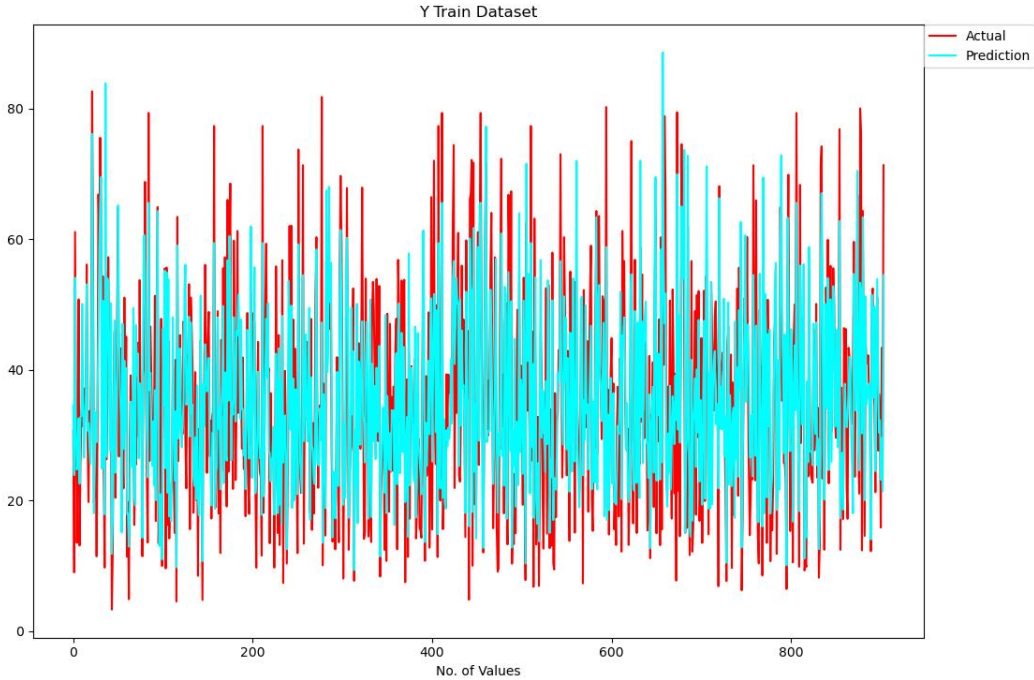
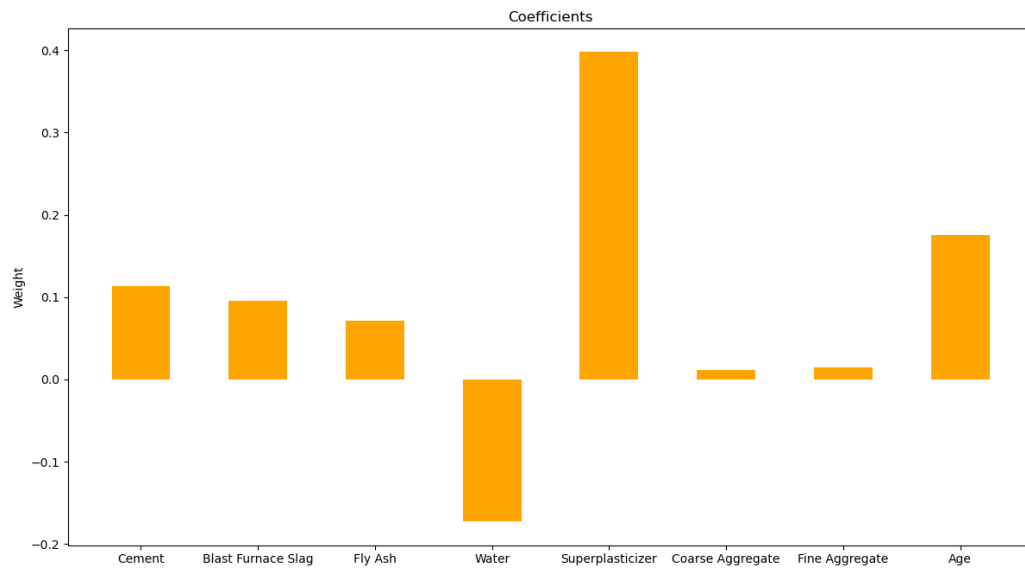
R² Value: 0.49210164952908153

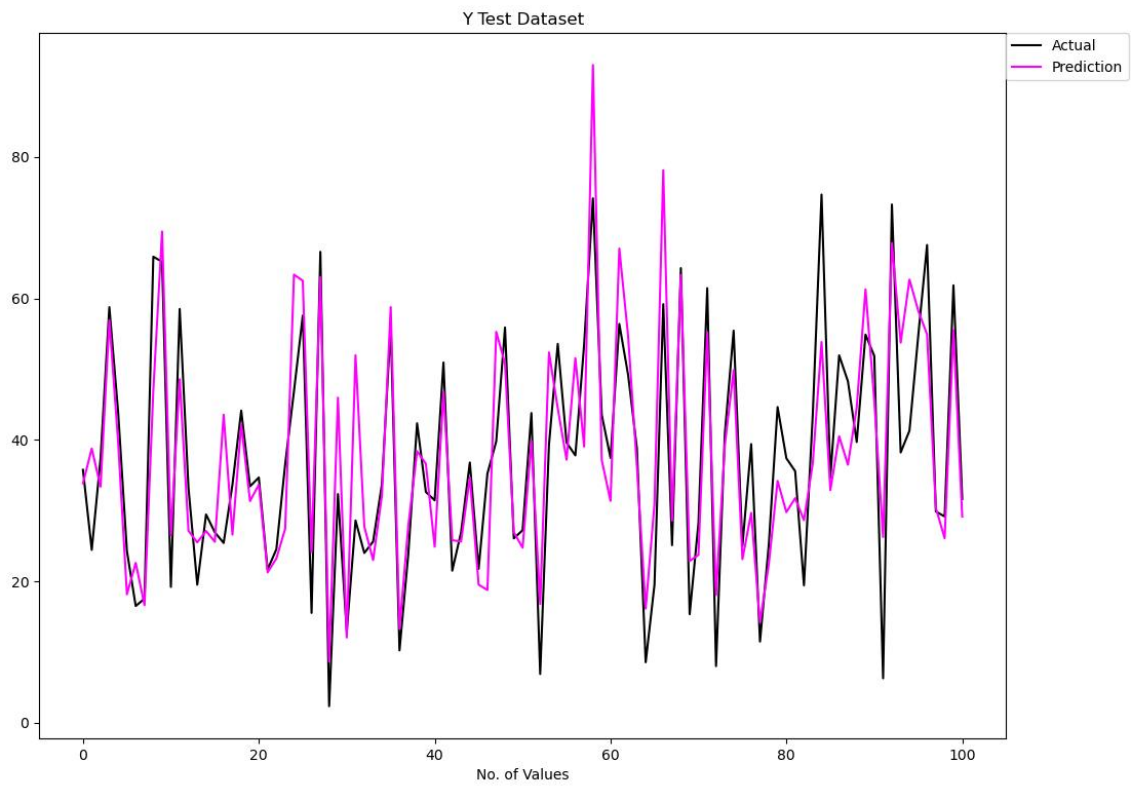
Test Accuracy:

Mean Squared Error: 79.67284654789876

R² Test: 0.6932652802404137

Here are the plots for the Sklearn Linear Regression Model in State 4.
Reminder that the images for the Gradient Descent Model in State 4 are above in the log.





Answers to Questions

Are you satisfied that you have found the best solution? Explain.

- Overall, we are satisfied that we have found the best solution in comparison to the sklearn Linear Regression model. Our results show a clear similarity and on occasion our accuracy is even better. There could be a better way to optimize the Gradient Descent model such as taking away certain attributes that do not correlate with the rest of the data, finding even better parameters, altering the starting weights, or even using other preprocessing methods such as Normalizing or Standardizing. But from what we were able to do we seem to have found a sweet spot of comparing the two models. The satisfaction comes from the goal of having accuracy comparable, if not better than the Linear Regression models results. Sure, the overall accuracy could probably be better given better preprocessing, but the tuning of the Gradient Descent model seems to be satisfactory.

Are you satisfied that the package has found the best solution? Explain.

- Similar to the answer to the last question there are of course preprocessing methods that can be taken to acquire a higher test accuracy with either the Gradient Descent or Linear Regression model. There could also be better models other than these types that may lead to better results. We aren't so quick to judge the models we have evaluated as the best overall solution, but for the purpose of this assignment we feel comfort knowing our Gradient Descent model is comparable to the sklearn packages Linear Regression model.