Plan

1. The plan is to look at the dataset we collected earlier and try to see if any other column that we dropped can be used to be in 50x5 dataset limit.
2. If we get one column to be in the limit and another one is not.
3. Like Student Teacher Ratio had 71 missing values, One AP Test had 69 missing values, College Readiness had 70 missing values and by dropping their NaN rows would make us go less than 50x5.
4. We can put mean values from the column in the of the rows to get the desired dataset size.
5. Making the dataset rows to be more than 50 and the attributes be to 9.
6. By keeping the original 6 columns and dropping the NaN rows from it but keeping the attributes to 9 columns as student teacher ratio cannot be converted into numerical and be scaled.
7. This means that the rows will be 63x6 from the transformed dataset but by adding the 3 other attributes will result in a dataset of 62x9 as there is 1 duplicate value present which has been merged.
8. After putting this new transformed dataset into a linear regression model, we will check if the results are same/better or worse.
9. Then we can perform bagging on the linear regression model to get better results as bagging can be used to understand the dataset trends and correlations better.

**Before Bagging**

Mean Squared Error (MSE) - Measures the average of the squares of the errors

0.027787678827900356

Root Mean Squared Error (RMSE)- Represents the square root of the MSE

0.1666963671706746

R-squared (R²) - Indicates the proportion of variance in the dependent variable that is predictable from the independent variables

0.09742671988922114

Coefficient Interpretation:

A one-unit increase in Total Enrolments is associated with a decrease of in Math Scoring by - 0.0586

A one-unit increase in One AP Test is associated with a decrease of in Math Scoring by - 0.0260

A one-unit increase in English Scoring is associated with an increase of in Math Scoring by - 0.6945

A one-unit increase in Graduation Rate is associated with a decrease of in Math Scoring by - 0.0412

A one-unit increase in Minority Enrolment is associated with a decrease of in Math Scoring by - 0.0545

A one-unit increase in Free Lunch Program is associated with an increase of in Math Scoring by - 0.0170

A one-unit increase in College Readiness is associated with an increase of in Math Scoring by - 0.1118

Intercept (baseline Math Scoring when all other factors are zero): 0.043759542802430995

**After Bagging (FINAL RESULTS)**

Mean Squared Error (MSE) - Measures the average of the squares of the errors

0.027787678827900356

Root Mean Squared Error (RMSE)- Represents the square root of the MSE

0.1666963671706746

R-squared (R²) - Indicates the proportion of variance in the dependent variable that is predictable from the independent variables

0.09742671988922114

Coefficient Interpretation:

A one-unit increase in Total Enrolments is associated with a decrease of in Math Scoring by - 0.0586 decrease

A one-unit increase in One AP Test is associated with a decrease of in Math Scoring by - 0.0260 decrease

A one-unit increase in English Scoring is associated with an increase of in Math Scoring by - 0.6945 increase

A one-unit increase in Graduation Rate is associated with a decrease of in Math Scoring by - 0.0412 decrease

A one-unit increase in Minority Enrolment is associated with a decrease of in Math Scoring by - 0.0545 decrease

A one-unit increase in Free Lunch Program is associated with an increase of in Math Scoring by - 0.0170 increase

A one-unit increase in College Readiness is associated with an increase of in Math Scoring by - 0.1118 increase

Intercept (baseline Math Scoring when all other factors are zero): 0.043759542802430995

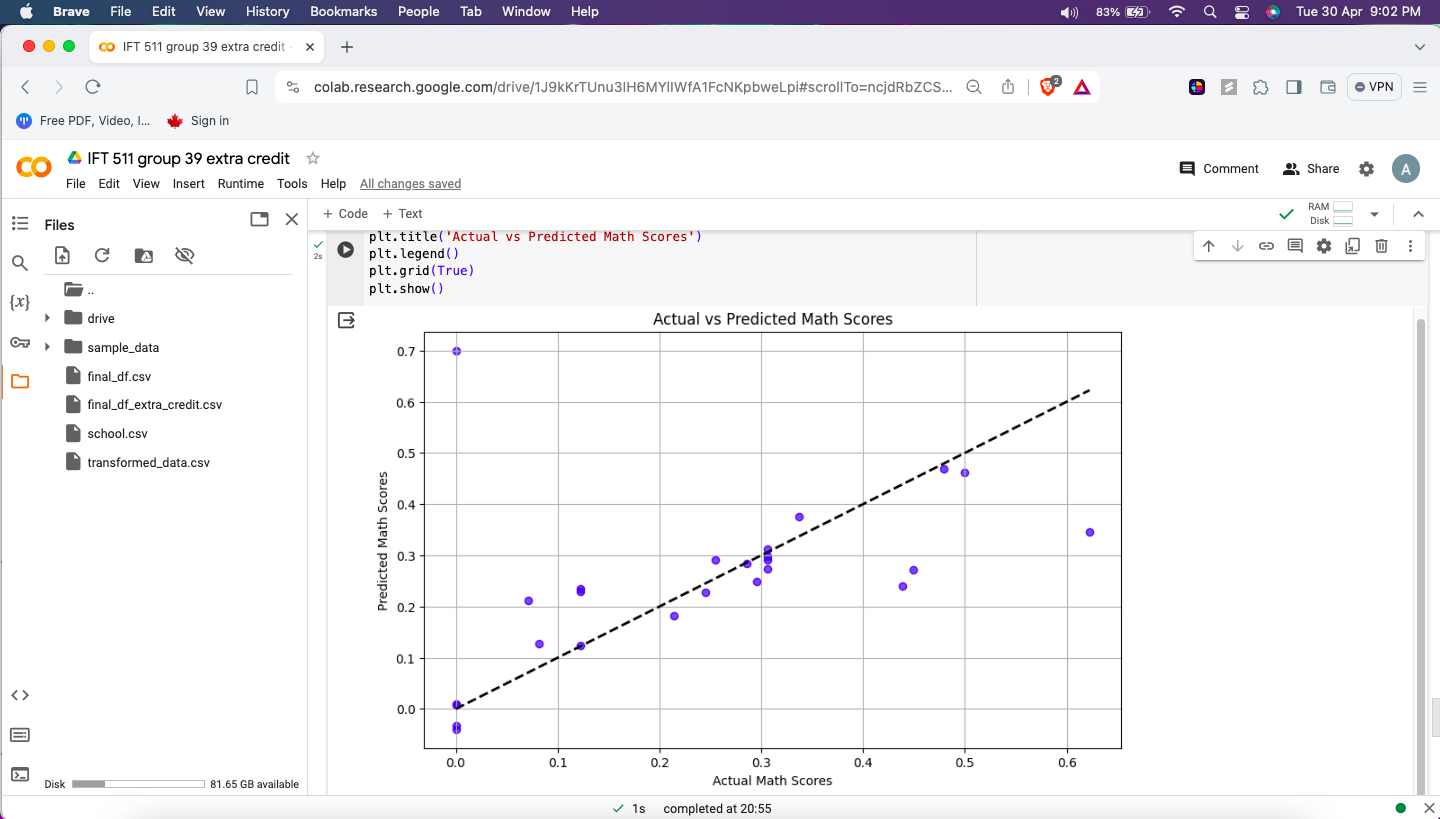
Mean Squared Error (MSE): The MSE in the new results (0.0278) is lower than the MSE in the initial results (0.0563). A lower MSE indicates that the model's predictions are closer to the actual values, reflecting better accuracy.

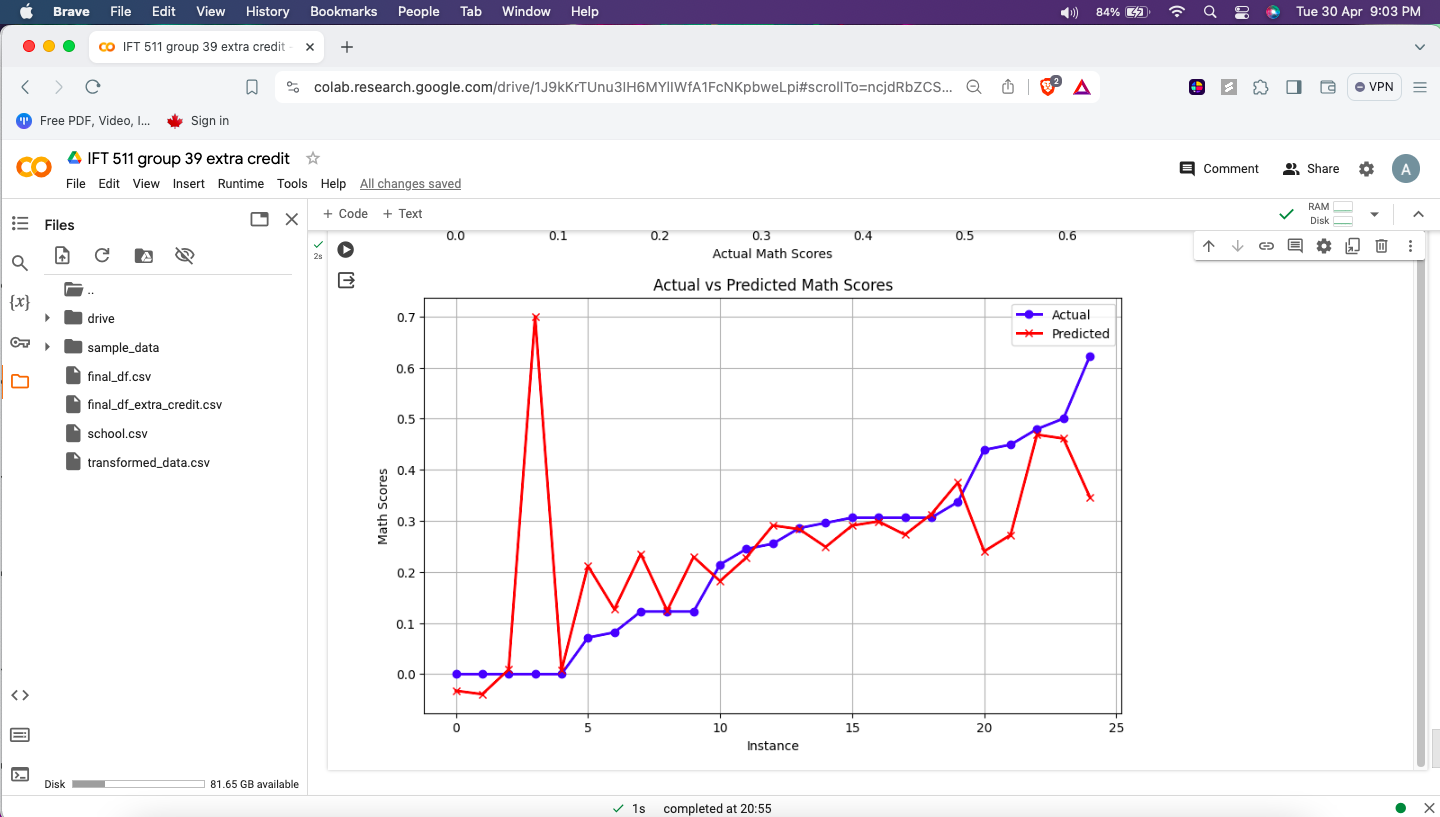
Root Mean Squared Error (RMSE): Similarly, the RMSE in the new results (0.1667) is lower than the RMSE in the initial results (0.2372). A lower RMSE indicates that the model's predictions have less variance from the actual values, indicating better performance.

R-squared (R²): The R² value in the new results (0.0974) is lower than the R² value in the initial results (0.1518). However, this does not necessarily mean worse performance. R² measures the proportion of variance in the dependent variable explained by the independent variables. In this case, a lower R² might indicate that the model is not explaining as much variance, but it does not necessarily mean worse predictive performance overall.

Coefficient Interpretation: The interpretations of the coefficients remain consistent across both sets of results. However, the magnitude of the coefficients differs between the two models, indicating different levels of influence of the independent variables on the target variable.

Overall, the new results demonstrate improved performance in terms of MSE and RMSE, suggesting better predictive accuracy compared to the initial results. However, the lower R² value indicates that the model may not explain as much variance in the target variable compared to the initial model.

****

****

**Q - new results and explanation of the new results.**

**A -** Mean Squared Error (MSE) - Measures the average of the squares of the errors

0.027787678827900356

Root Mean Squared Error (RMSE)- Represents the square root of the MSE

0.1666963671706746

R-squared (R²) - Indicates the proportion of variance in the dependent variable that is predictable from the independent variables

0.09742671988922114

Coefficient Interpretation:

A one-unit increase in Total Enrolments is associated with a decrease of in Math Scoring by - 0.0586 decrease

A one-unit increase in One AP Test is associated with a decrease of in Math Scoring by - 0.0260 decrease

A one-unit increase in English Scoring is associated with an increase of in Math Scoring by - 0.6945 increase

A one-unit increase in Graduation Rate is associated with a decrease of in Math Scoring by - 0.0412 decrease

A one-unit increase in Minority Enrolment is associated with a decrease of in Math Scoring by - 0.0545 decrease

A one-unit increase in Free Lunch Program is associated with an increase of in Math Scoring by - 0.0170 increase

A one-unit increase in College Readiness is associated with an increase of in Math Scoring by - 0.1118 increase

Intercept (baseline Math Scoring when all other factors are zero): 0.043759542802430995

Mean Squared Error (MSE): This metric measures the average of the squares of the errors between the predicted values and the actual values. In this case, the MSE is 0.0278, indicating that, on average, the squared difference between the predicted and actual Math Scoring values is 0.0278.

Root Mean Squared Error (RMSE): The RMSE is the square root of the MSE. It represents the standard deviation of the residuals, which are the differences between the predicted and actual values. Here, the RMSE is 0.1667, indicating that, on average, the model's predictions deviate from the actual values by approximately 0.1667 units of Math Scoring.

R-squared (R²): R² measures the proportion of variance in the dependent variable (Math Scoring) that is predictable from the independent variables (Total Enrolments, One AP Test, English Scoring, Graduation Rate, Minority Enrolment, Free Lunch Program, College Readiness). An R² of 0.0974 suggests that approximately 9.74% of the variance in Math Scoring can be explained by the independent variables included in the model.

Coefficient Interpretation: The coefficients represent the change in Math Scoring associated with a one-unit increase in each independent variable, holding all other variables constant. For example:

A one-unit increase in Total Enrolments is associated with a decrease of 0.0586 in Math Scoring.

A one-unit increase in One AP Test is associated with a decrease of 0.0260 in Math Scoring.

A one-unit increase in English Scoring is associated with an increase of 0.6945 in Math Scoring.

A one-unit increase in Graduation Rate is associated with a decrease of 0.0412 in Math Scoring.

A one-unit increase in Minority Enrolment is associated with a decrease of 0.0545 in Math Scoring.

A one-unit increase in Free Lunch Program is associated with an increase of 0.0170 in Math Scoring.

A one-unit increase in College Readiness is associated with an increase of 0.1118 in Math Scoring.

Intercept: The intercept represents the baseline Math Scoring when all other independent variables are zero. In this case, it is 0.0438, indicating the expected Math Scoring when none of the other factors are considered.

**Description of Final Dataset -**

1. **School Name:**

Represents the name of the school in Phoenix, Arizona.

1. **Total Enrollments:**

Shows the total number of students enrolled in the school.

1. **One AP Test:**

Determines if the institution provides at least one Advanced Placement (AP) test. AP examinations are difficult exams that can grant students college credit.

1. **Math Scoring:**

This represents students' arithmetic performance or scores. It may be based on standardized exams or other measures.

1. **English Scoring:**

Represents student achievement or scores in English. Similar to Math Scoring, it may be based on standardized examinations or other evaluations.

1. **Graduation Rate:**

The proportion of students that successfully graduate from the school. A greater graduation rate is often seen positively.

1. **Minority Enrollment:**

The percentage of enrolled students from minority groups. It gives information about the variety of the student body.

1. **Free Lunch Program:**

Indicates if the school participates in the free lunch programme, which is frequently tied to socioeconomic status.

1. **College Readiness:**

This represents the school's efforts to prepare students for college. It may involve advanced coursework, counseling, or other college preparation programmes.