COGS 109 Final Project

SAT Scores: A Culmination of Contributing Factors or an Isolated Score?

Presented by: Bingzhe Wang, Alexander Schonken, Jolene Leung

Backgrounds

To get a head start of the project, we first raised our research question: Can you predict a US student's Math SAT Score based on his/her Mathematics GPA and Music GPA? To further dig into our topic and try to find the answer to our research question, we decided to investigate the School Scores Dataset from the CORGIS Dataset Project, which is a dataset of 577 samples and 99 features. Each sample is attached to a specific state in the United States of America and a range of multiple years (2005-2015). The features has a very wide variety, which ranges in number of test takers of each gender (male and female), the number of test takers in different ranges of math and verbal scores (from 200 to 800) and different genders (male and female), the average GPA of test takers in different subject, the average time of test takers spending on these subjects (in years), and a range of family income of test takers (from less than 20K to more than 100K a year). From all the features we can see that this dataset does deal with data over many different properties, including different states, time spent on each subject, income, gender, average gpa for each subject and multiple ranges of SAT scores for each subject. Therefore, we will carefully work around and select appropriate feature into our analysis as we handle the dataset and try to build our desired linear regression model. The dataset can be downloaded from the link: https://corgis-edu.github.io/corgis/csv/school_scores/

Methods

As we discussed above in the backgrounds, there were many different features that cover many different aspects, which were not necessary to our study. Since we were studying the possible relationship between a US student's Math SAT score and his/her Math GPA and Music GPA, we had to clean and sort our dataset first.

For the first analysis, we took out the data of average Math SAT scores of California, Mississippi and Illinois and plotted the average of Math SAT scores over time of these three states so that we could see if time had any effect on the average of Math SAT scores.

For the second analysis, we pulled out the data of income and average SAT scores of CAlifornia, Arkansas, New York and North Dakota and plot bar plots of average of Math and Verbal scores over different ranges of income of these four states. The reason we conducted such analysis was that we believed there would be solid relationship between family income and test takers' performance.

For the third analysis, we extracted the average Math SAT score and the GPAs of the same four states as in the second analysis. We tried to compare the average Math SAT score and the GPAs because we believed that higher GPA would naturally lead to higher SAT scores.

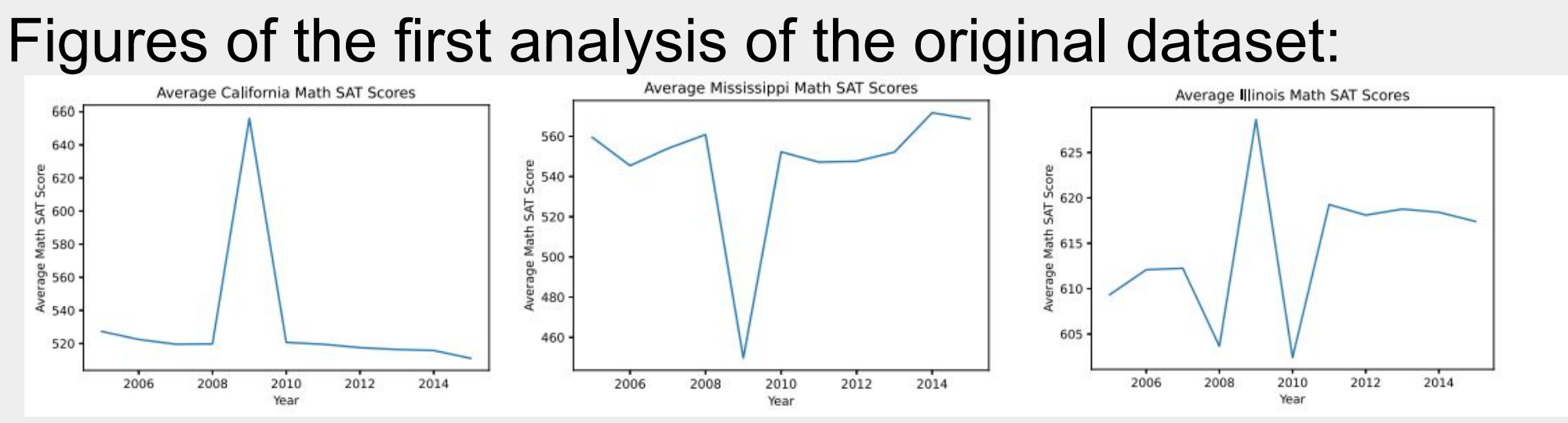
For the fourth analysis, we conducted analysis on family average Math SAT scores in higher income ranges in California. Therefore, we got corresponding data of California and plotted the family average Math SAT scores of each range of relatively higher family income (60-80k and 100k+) over time.

After these procedures, we raised three possible linear regression models: Linear Regression Model 1: math_sat_score = w0 + w1 * math_gpa + w2 * music_gpa; Linear regression Model 2: math_sat_score = w0 + w1 * math_gpa; Linear regression Model 3: math_sat_score = w0 + w1 * music_gpa;

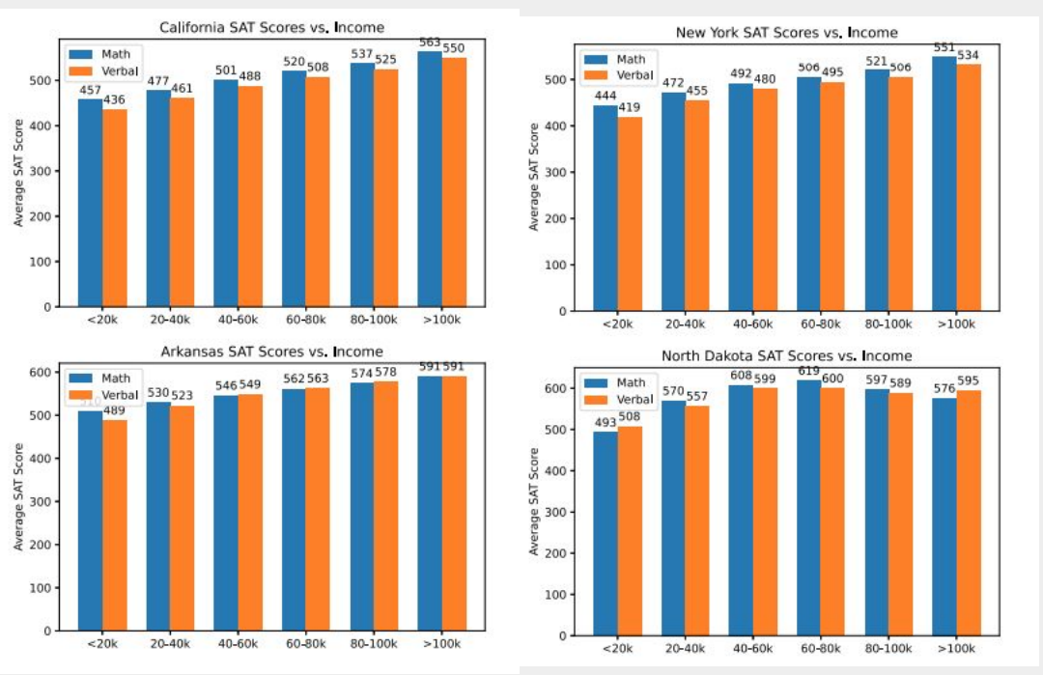
For each model, we got the corresponding data ready and then trained the data. We used techniques such as including bias and creating training and testing sets in order to find weight for each linear regression model. By doing so, we finally got functions of all the three models. After that, we will calculate the training SSE of each model so that we can select the best model for our project.

Results

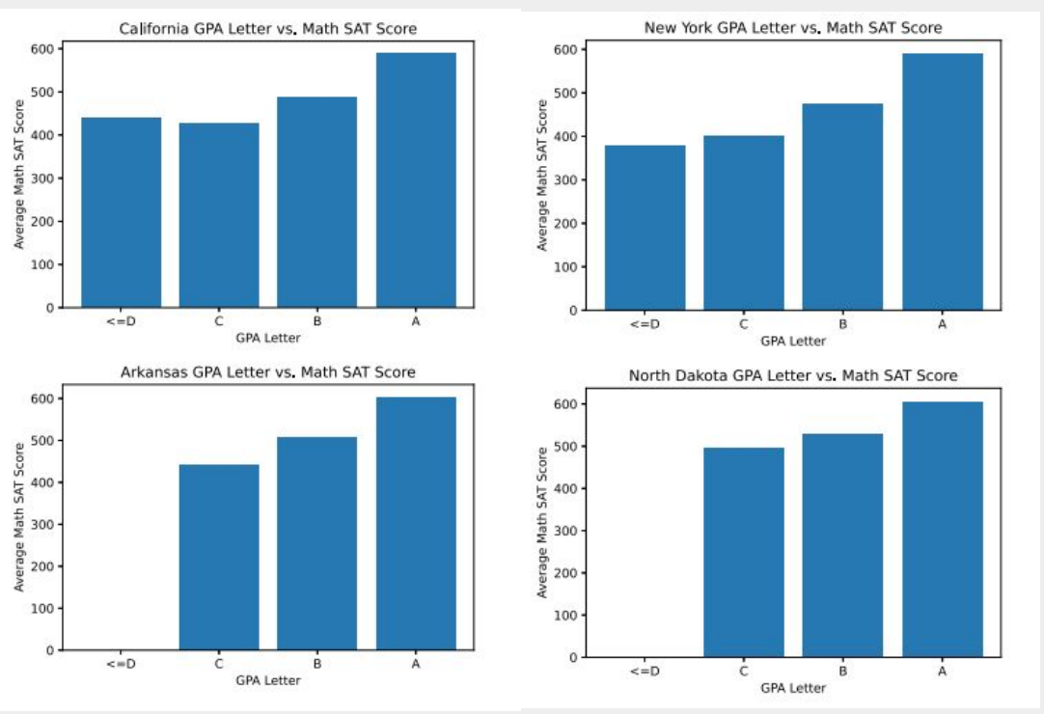
Figures of the first analysis of the original dataset:



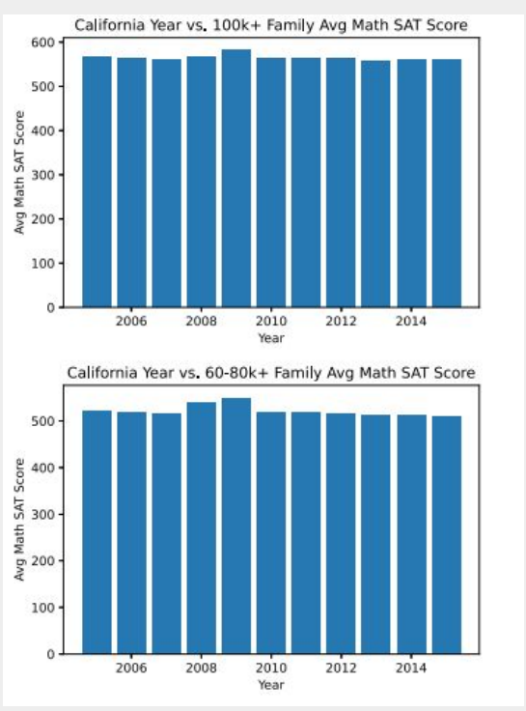
Figures of the second analysis of the original dataset:



Figures of the third analysis of the original dataset:



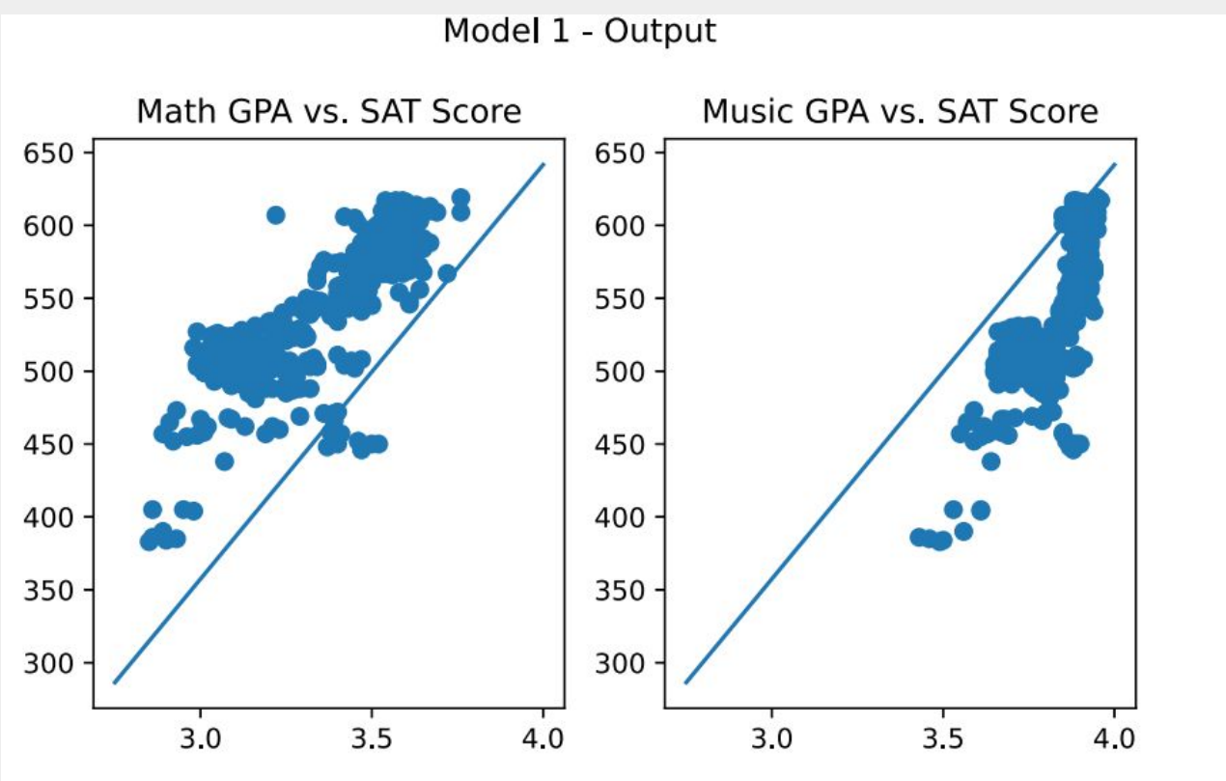
Figures of the fourth analysis of the original dataset:



From all these figures generated from our analysis, we can conclude that:

1. One or multiple events between year 2008 and 2010 interfered test takers' performance in their SAT test, resulting in a huge fluctuate in their Math SAT scores.
2. Higher family income would lead to better performance in SAT Test (higher SAT score).
3. Higher GPA (better letter grade) would also lead to better performance in SAT Test (higher SAT score). This means GPA is positively correlated to SAT score.

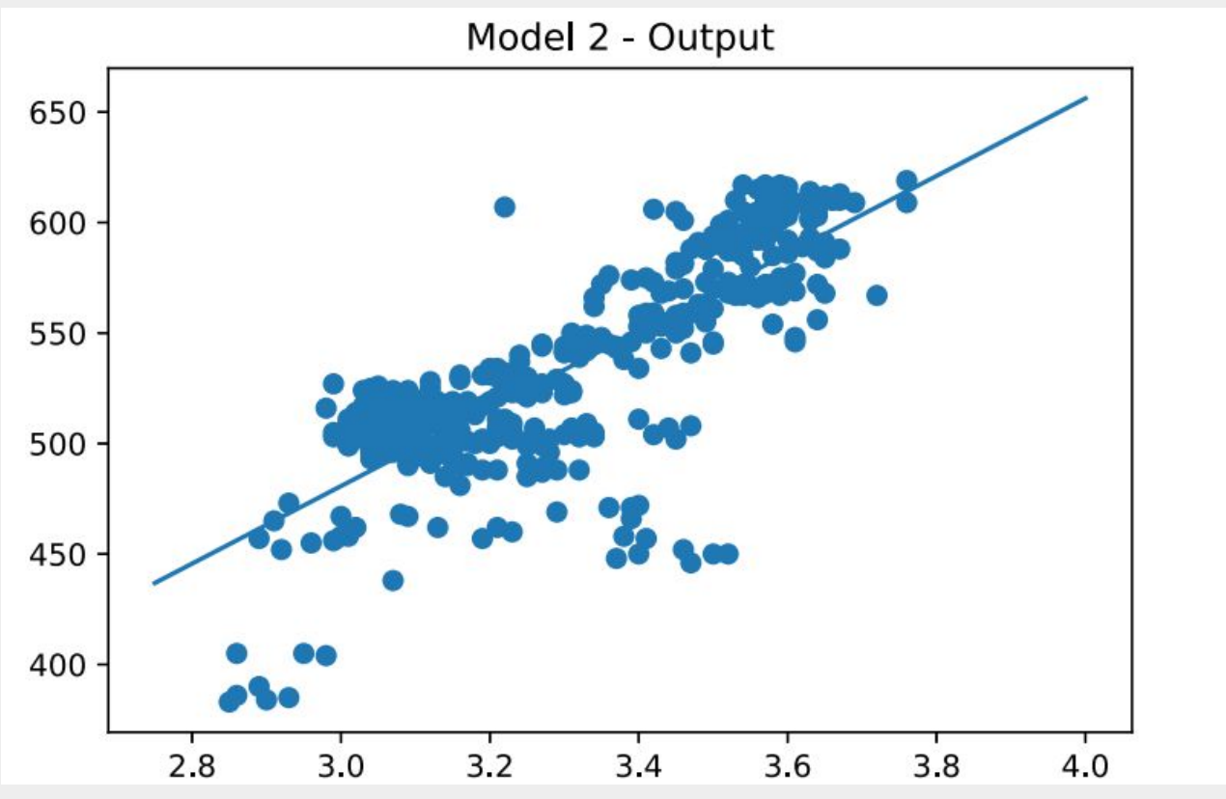
Figure of the first linear regression model:



Function of the first linear regression model:

math_sat_score = [-395.21776422] + [128.28079375] * math_gpa + [132.69110418] * music_gpa

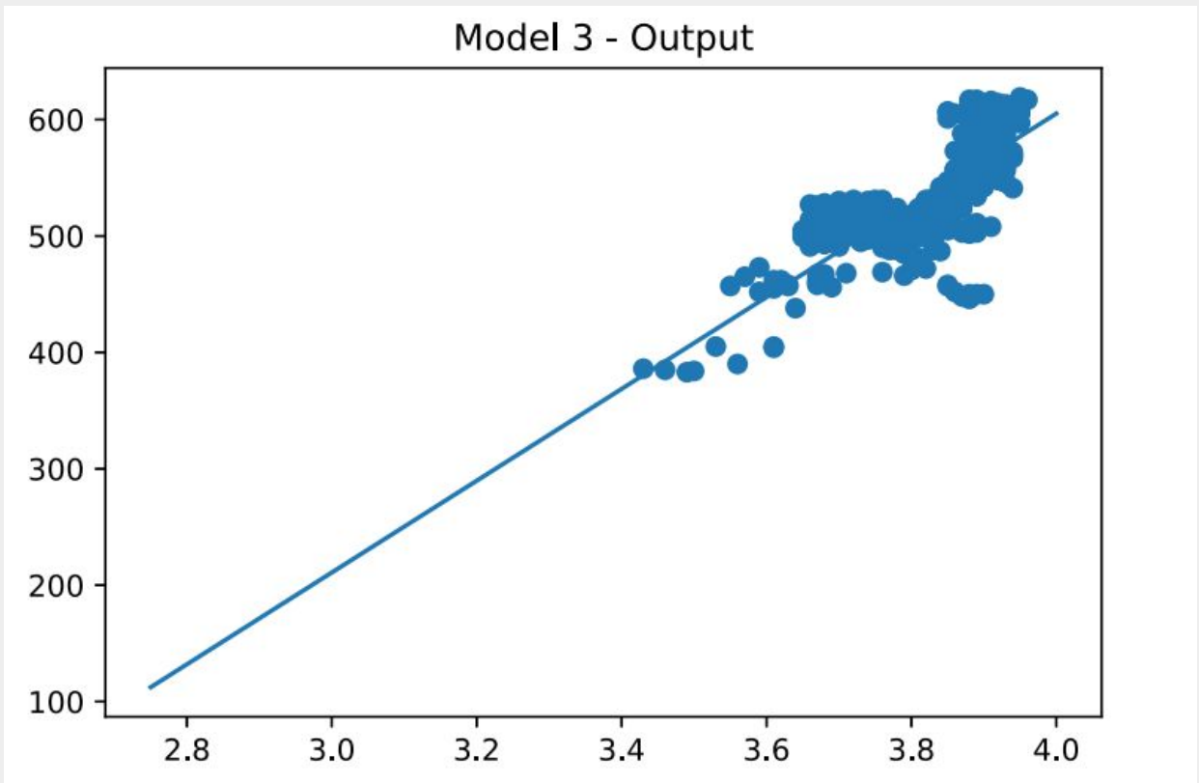
Figure of the second linear regression model:



Function of the second linear regression model:

math_sat_score = [-51.80710101] + [177.77622533] * math_gpa

Figure of the third linear regression model:

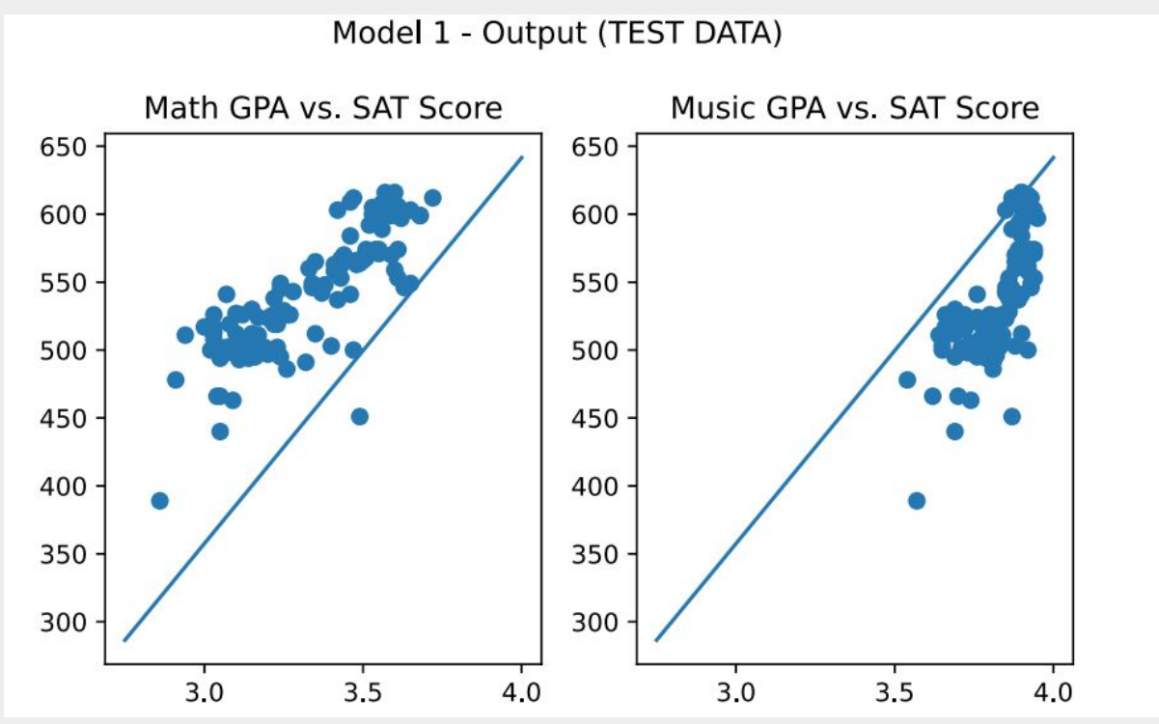


Function of the third linear regression model:

math_sat_score = [-983.79740905] + [397.75669016] * music_gpa

Training SSE for each model:
Model 1: Training SSE: 315203.6307403873
Model 2: Training SSE: 344557.2348631661
Model 3: Training SSE: 375899.79992522375

By looking at the training SSE and the plots of each model, we think that **model 1** would be the most appropriate model because it has the smallest training SSE among all the three models and also seems to generalize better when looking at 2D plots of the regression line.



Model 1 test RMSE: 26.538549090806253

Discussion

Based on the results we generated, we can affirm that a US student's Math and Music GPA can be used to predict his/her Math SAT score. Our desired model for this prediction would be "**math_sat_score = [-395.21776422] + [128.28079375] * math_gpa + [132.69110418] * music_gpa**". The reason we choose this model is that it has lowest training SSE among all the models we selected and its test RMSE is only 26.53, which is an acceptable value when the SAT Math scores range from 0-800 because it is only ~3% of the total values.

Nevertheless, there are still limitations in this model. Our model is too general that it mainly focuses on average SAT scores of a range of students from different features. It does not give samples on a per student basis so that the result and data generated from this dataset are not specific enough. Therefore, this model is not completely able to give a definitive and specific answer for each student but instead only generate a general prediction of how students will perform on average given their math and music GPAs.

Future work with this dataset would greatly benefit from the availability of individual data so that each student's scores can be analyzed. However, our outcome from this investigation can be used as a reminder that math alone is not the only factor in achieving a high math SAT score. Music GPA can also contribute to a student's performance with a high correlation. Schools or parents need to reconsider their decision if they want to cut off music program for teenagers.