Final Project

Professor Barriga – Stat 2223-004

12.6.2024

Angela Iraya, Cassidy Lowrance, Faith Akujobi-Robers, LK Stewart, Salma Kulatein

Github Repository : *cleanData.csv:*

INTRODUCTION

Background + Interest

Studying is influenced by various factors, from personal habits to the time when studying. One factor that we believe that is common is what types of music people choose to listen to while studying. Students often include music in their study routine, with preferences varying significantly - from instrumental tunes to upbeat lyrical tracks.

The relationship between music and study habits isn't completely clear. Some believe that music can improve focus, while others can find it disruptive. Some genres might be more efficient at helping students stay focused during study sessions. Our project seeks to address a specific aspect which is the relationship between music genres and their impact on study habits of Computer Science and Data Science students. These students usually deal with tasks that require concentration, as they deal with problem solving tasks that require logical thinking. So making the choice of the right study environment including background music, could potentially be a significant factor impacting how effective their studying is.

- What we want to predict - Can the genre of music listened to while studying predict total study time ?

Objectives (statement of problems, what we want to predict, why it's important to us)

Studying is a crucial part of the academic experience for many students. It is influenced by various factors, ranging from personal habits to the timing of study sessions. One common factor that often plays a role is music. Many students often incorporate music into their study routines, with preferences varying from calming instrumental tunes to upbeat lyrical tracks. However, the relationship between music and study habits remains uncertain. While some believe that music enhances concentration and focus, others argue that it's more distracting. Certain genres, in particular, might be more efficient at helping students stay focused during study sessions.

Our goal is to predict *whether different genres of music listened to while studying can influence total study time.* As students, we are interested in exploring how other students' music tastes affect their studying habits. Specifically, addressing questions like: *Does listening to music during study sessions correlate with increased study time? Are there specific genres of music that contribute to longer or more effective study periods?*

Understanding these influences and underlying factors can help students make better decisions about their study habits, which could improve academic performance and productivity.

DATA COLLECTION

Data Collection Process (how and who we collected it from)

To gather our data we created a Google Form with questions related to people's music preferences and study habits. We each collected responses from other students in our classes or that we knew personally. After gathering our responses we combined them into a single spreadsheet.

We found that the majority of our responses were from Computer Science and Data Science majors. In order to reduce sampling bias we chose to narrow our research question to only focus on Computer Science and Data Science majors and removed all responses from students with other majors. Over the next several days we gathered more responses specifically from Computer Science and Data Science majors to replace the responses we had removed. We then once again combined them into a single dataset and then began the cleaning and standardization processes.

Data Cleaning

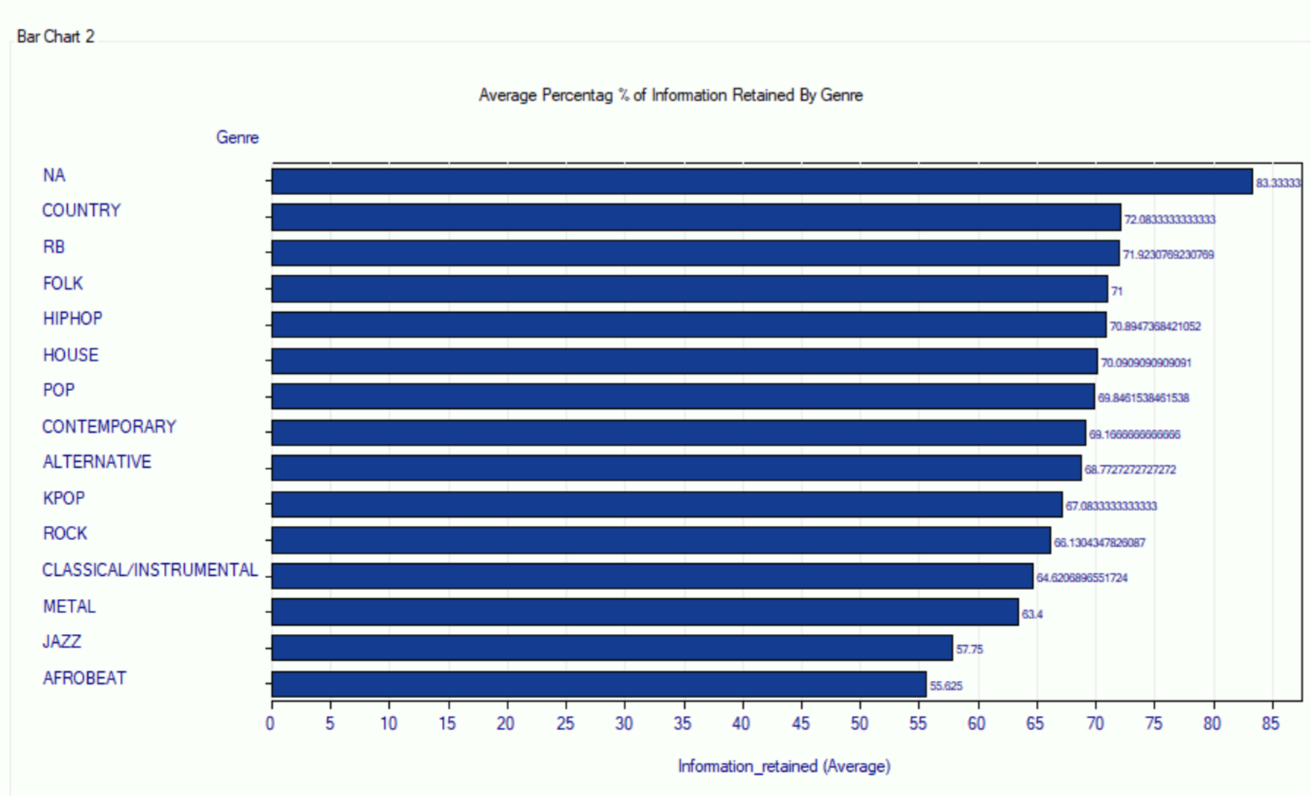
The first step in our data cleaning process was to standardize the language used in our dataset. For the variable *Major* all variations of Computer Science such as, cs, compsci, or comp-sci were converted to Computer Science. The same was done with Data Science. Next we standardized the *Genres* variable. In our survey we had allowed respondents to select one or more genres from a list, but they were also able to enter their own. This resulted in many genres being entered that could be generalized into one of the ones on our list. For instance, christian rock, alt rock, and classic rock could all be categorized under the more general rock genre.

The previous steps in our data cleaning processes had to be done by hand but we realized the majority of our standardization and cleaning could be done much faster programmatically using SAS. Even after standardizing the *Genres* variable by hand there were still several very specific or unique genres that we did not feel we could fit into our existing categories. We did not want to remove these responses and reduce the size of our dataset so we instead grouped them as 'Other' and used that as our reference group for the *Genres* variable. The *Information Retained* variable included unwanted percent (%) and dash (-) symbols which were stripped out using SAS. As our data had many categorical variables we also used SAS to convert some of them into numeric levels due to the fact that some of the visualization procedures in SAS such as sgscatter can only handle numeric data.

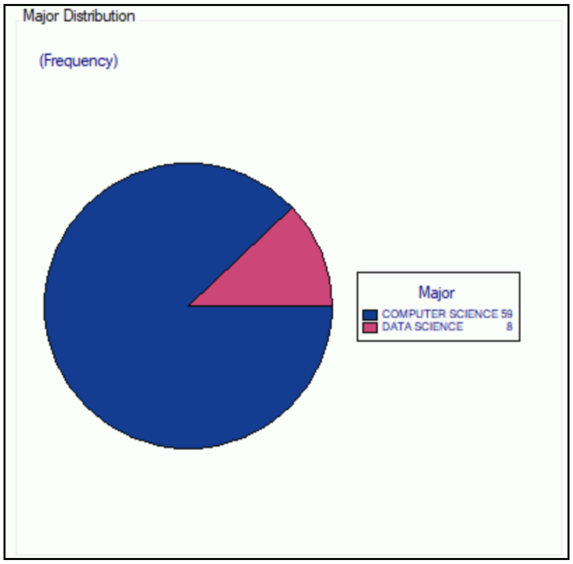
Our data had a high number of categorical variables, one of which had 15 levels and two of our categorical variables allowed multiple selections, meaning that for example, a respondent could select levels 1, 2, and 3 of the variable. Due to needing so many dummy variables we chose to generate them programmatically using SAS as well. Once we had reviewed the resulting dataset and checked that all the dummy variables were being generated correctly we saved our dataset in a new file called *cleanData.csv* which can be found in our project repository.

EXPLORATORY ANALYSIS

Initial analysis shows that not listing to music leads to the most information retained when studying, however only three respondents did not list any genres that they listened to when studying.

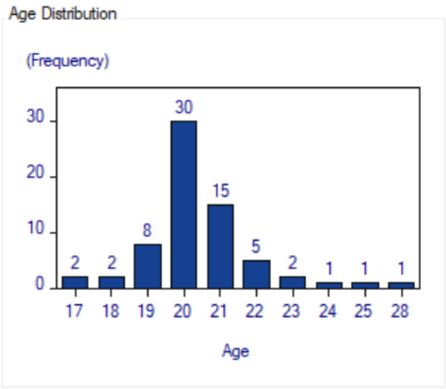


Assumptions and Limitations of the Data



The dataset is confined to students who are majoring in computer science and data science, which narrows the extent of our analysis. By focusing on these two fields, the findings may not reflect the study habits or the preferences of music from students in other majors. Additionally relatively few of the respondents were data science majors, however given how similar the two majors are we do not think this adds a significant amount of biases to our sample.

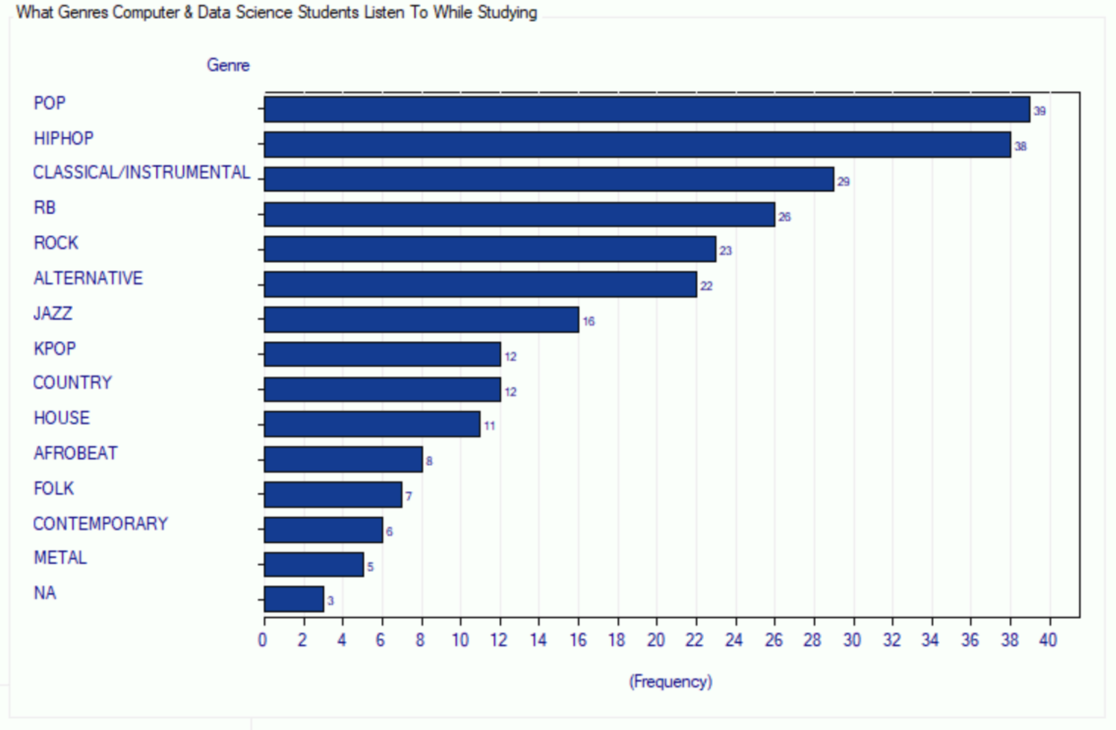
Due to the fact that we collected data from college students, our sample mostly includes college age students between 19 and 22. Given this, any conclusions we make from this data will primarily apply to students within that age group.



The vast majority of participants reported listening to music while studying, meaning that our data may not accurately represent students who prefer silence or other methods to increase their focus while studying. Only three respondents didn't list any genres of music they listened to while studying.

Our dataset contains 67 responses, which is likely too small to be representative of something as broad as music preferences and study habits. This can limit the ability to see if there are larger patterns or trends that might appear in a wider population. Even though 67 responses gives us a good starting point, increasing the sample size can ensure that there is more variety in the sample and would likely improve the reliability of the findings.

Another limitation of our dataset is that some genres were overrepresented while others were far more niche which may have skewed our results. For example 39 respondents included Pop as one of their study genres and 38 included Hip-Hop while only 16 included Jazz, 11 included house music, and 8 included Afrobeat.



We are relying on qualitative self-reported data, which can be inaccurate. Participants may have provided incorrect information unintentionally about their study habits or not remember the details about their preferred music genres to study to or misinterpreted the genre categories. These inaccuracies could lead to the results being skewed and decrease the reliability of the conclusions from this study.

Our research also didn't include factors such as deadlines or stress. These factors could influence whether people choose to listen to music, as some people tend to avoid it when stressed or facing a deadline. For some people, stress can alter their ability to focus and, as a result, their music preferences may be altered as well. Some people may find that music helps them relax and concentrate, while others may feel that it adds to their anxiety, especially during high-pressure times like deadlines.

Lastly, one of the largest assumptions in our data is that for students who listen to multiple genres of music while studying (which were the vast majority) our data assumes that they retain the same amount of information regardless of which of their preferred genres they are listening to. Our dataset assumes for example, that a student who feels they retain 75% of what they study and listens to Metal, Pop, and Hip-Hop, retains 75% regardless of which of the 3 they are listening to.

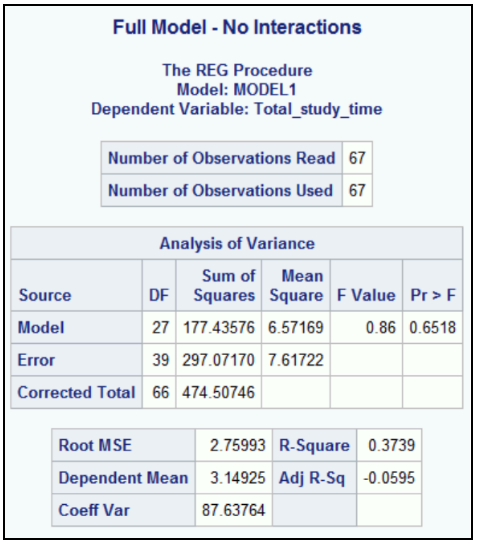
We are assuming for all models that all the random errors ε will follow independent and identically normal distribution with mean 0 and variance σ^2.

ANALYSIS

Initial Model

Our initial model included a total of 27 variables, broken down below:

- 2 continuous variables (age, information retained)
- 3 categorical variables with two levels (major, listens to music, studies with music)
- 1 categorical variable with three levels (study music has lyrics)
- 2 categorical variable with four levels (preferred study times, study efficacy)
- 1 categorical variable with fourteen levels (genres)



Full Model - No Interactions

The REG Procedure
Model: MODEL1
Dependent Variable: Total_study_time

Number of Observations Read	67
Number of Observations Used	67

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	27	177.43576	6.57169	0.86	0.6518
Error	39	297.07170	7.61722		
Corrected Total	66	474.50746			

Root MSE	2.75993	R-Square	0.3739
Dependent Mean	3.14925	Adj R-Sq	-0.0595
Coeff Var	87.63764		

The Initial model did not appear to be predictive for our response variable total time spent studying due to the model's very high p-value of 0.6518. Additionally, this model had an R-Squared value of 0.3739 meaning that our model explained only 37.39% of the variance in our data. We thought the large number of variables in our model might have been part of the problem given that we were not accounting for interactions. We conducted hypothesis testing on each of our variables, against the null hypothesis that they were not statistically significant to the model.

Null Hypothesis: H_0: $\beta_1 ... \beta_{27} = 0$

Alternative Hypothesis: H_0: $\beta_1 ... \beta_{27} \neq 0$

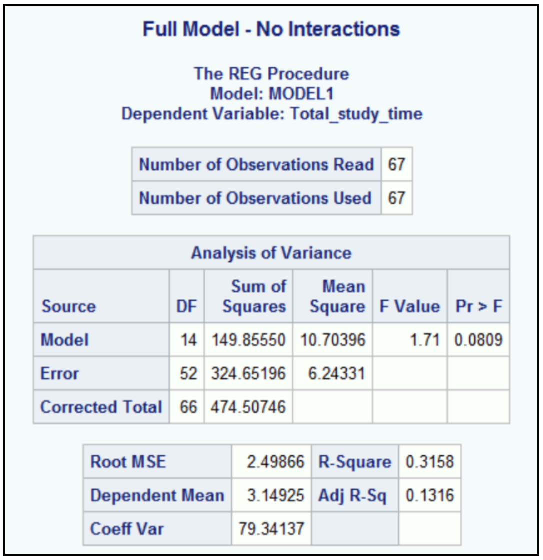
Determining Significant Predictors

Our initial hypothesis testing showed that the only variable with a significant P-value was whether students listened to music while they studied. That variable had a P-value of 0.0108 while all other variables had values over 0.05, with most of them being over 0.1. We chose to drop predictors incrementally so that we could see how the model behaved and whether removing certain predictors made the overall model more useful.

After removing 2-3 predictors with the highest P values we reran the model to observe any changes. We observed a gradual but steady decrease in both our model's P-value and R-squared value. So while our model was becoming less accurate, it was also getting closer to being useful in predicting total time per day that computer science students spend studying.

After dropping the following variables our model still did not have a p-value under 0.05 and the r-squared value had decreased by approximately 0.5 to 0.3158 as can be seen in the figure to the right. The following variables were dropped from our model:

- information retained
- All levels of preferred study times
- 3rd level of study efficacy
- Genre levels: Pop, Classical/instrumental, country, R&B, Kpop
- Listens to music (regardless of whether or not they are studying)
- 3rd level of study music: study music sometimes has lyrics (studyL_some_D)
- Major



Full Model - No Interactions

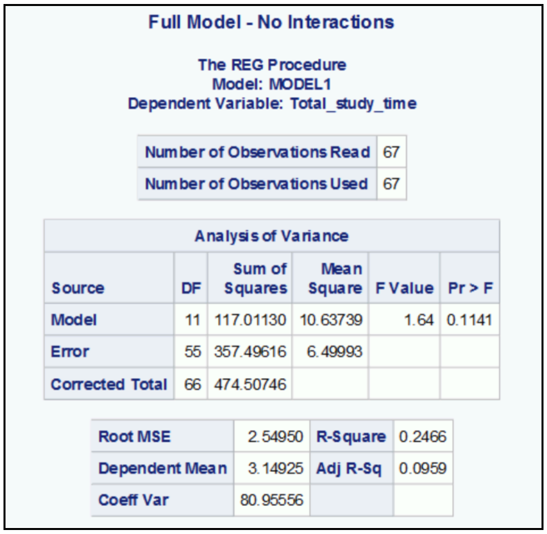
The REG Procedure
Model: MODEL1
Dependent Variable: Total_study_time

Number of Observations Read	67
Number of Observations Used	67

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	149.85550	10.70396	1.71	0.0809
Error	52	324.65196	6.24331		
Corrected Total	66	474.50746			

Root MSE	2.49866	R-Square	0.3158
Dependent Mean	3.14925	Adj R-Sq	0.1316
Coeff Var	79.34137		



Full Model - No Interactions

The REG Procedure
Model: MODEL1
Dependent Variable: Total_study_time

Number of Observations Read	67
Number of Observations Used	67

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	117.01130	10.63739	1.64	0.1141
Error	55	357.49616	6.49993		
Corrected Total	66	474.50746			

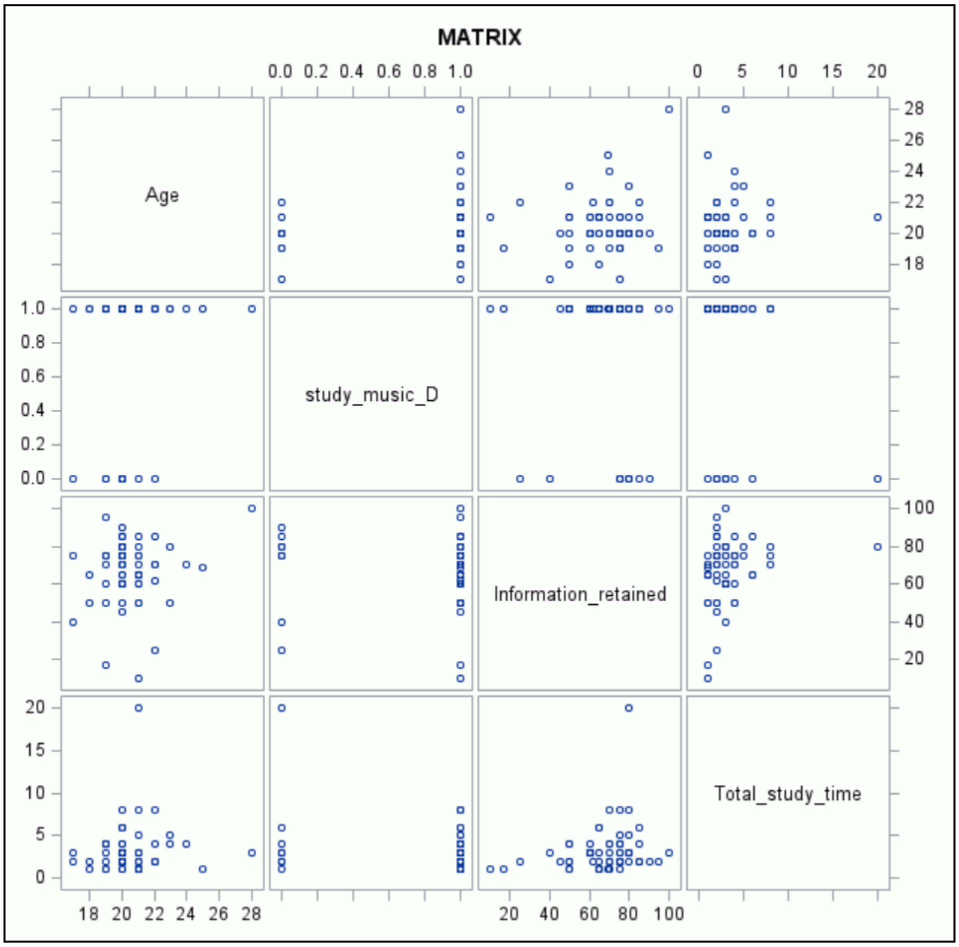
Root MSE	2.54950	R-Square	0.2466
Dependent Mean	3.14925	Adj R-Sq	0.0959
Coeff Var	80.95556		

The next three variables we planned to drop were the genre levels Hip-Hop, Jazz, and afrobeat; however, after dropping them our model became significantly less accurate and moved further away from being useful in predicting our response variable. Removing these three predictors increased the model's p-value from 0.0809 to 0.1141 and decreased the R-squared value from .3158 down to 0.2466.

We experimented with leaving these variables and removing others with p-values above 0.05 but found that the model's p-value continued to increase while R-squared decreased. We concluded that there was no way for this model to be useful in predicting our response variable.

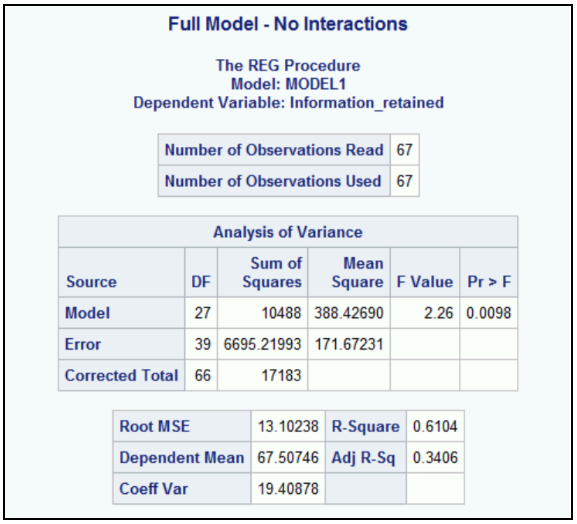
We chose to try to create another model using "information_retained" as the response variable in the hope that it would be more predictive than our first model. The decision to choose “information_retained” as the new response variable is due to the knowledge acquired from the scatter plot matrix, where there is a slight correlation observed between this variable and several predictors. The “information_retained” shows a relationship like study_music_D and Total_study_time, showing that these variables could help in explaining how much information that the students retained. However, variables like Age didn't seem to have a significant relationship to “information_retained”, as displayed by the randomness of the scattering of points in the matrix.

Our decision to switch our focus was due to the shortcomings of the initial response variable (“Total_study_time”). That model didn’t have significant predictors and a low R-squared value. By changing the response variable to “information_retained”, we are focusing more on a variable with a clearer relationship to the predictors, which might result in a more beneficial model. This change will improve the overall standard of the analysis.



New Model With Information_retained As Response Variable

The new model with “information_retained” as the response variable, displayed an improvement versus the initial model. The new model has an R-squared value of 0.6104, shows that it describes 61% of the variance that is in the data. The adjusted R-squared (0.3406) was a little bit lower and it still showed a better fit versus the initial model. The model is significant with an F-value of 2.26 and the p-value of 0.0098, indicating that the predictors contributed to describing the response variable.



Full Model - No Interactions

The REG Procedure
Model: MODEL1
Dependent Variable: Information_retained

Number of Observations Read	67
Number of Observations Used	67

Analysis of Variance

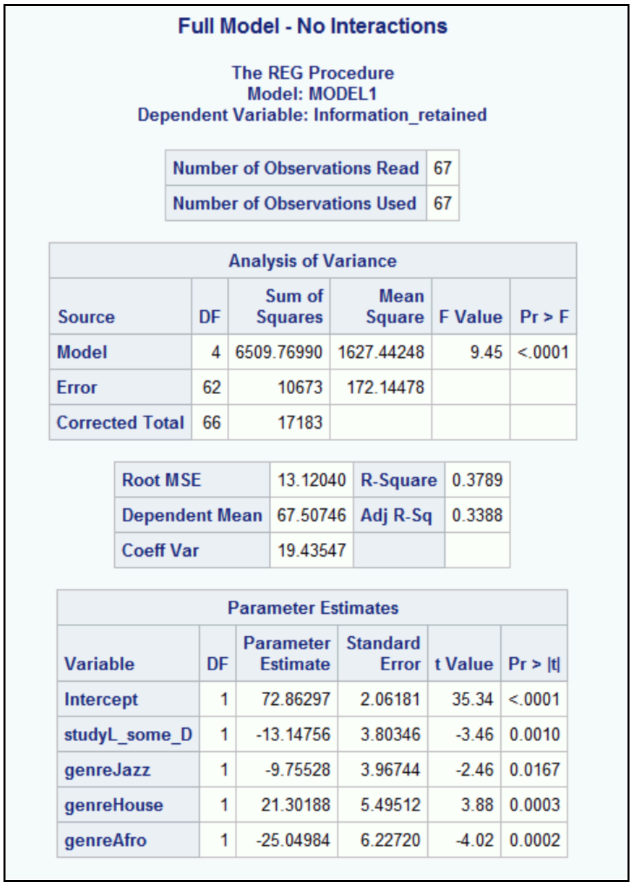
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	27	10488	388.42690	2.26	0.0098
Error	39	6695.21993	171.67231		
Corrected Total	66	17183			

Root MSE	13.10238	R-Square	0.6104
Dependent Mean	67.50746	Adj R-Sq	0.3406
Coeff Var	19.40878		

As with our first model we dropped predictors incrementally, and reran the model after dropping a few with the highest p-values. We eventually reached a model with only significant predictors, however it did also result in a significant decrease in the model’s accuracy from 61.04% to 37.89%.

The following predictors were dropped from our model:

-
- Total study time
- Listens to music (regardless of whether or not they are studying)
- All 3 levels of preferred studying times
- All levels of study efficacy
- Genre levels: Metal, Kpop, Contemporary, folk, classical/instrumental, Pop, R&B, rock, alternative, country, hip-ho
- Major
- Study music type level 2 (study music does have lyrics)
- Listens to music when studying (study_music_D)
- Age



Full Model - No Interactions

The REG Procedure
Model: MODEL1
Dependent Variable: Information_retained

Number of Observations Read	67
Number of Observations Used	67

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	6509.76990	1627.44248	9.45	<.0001
Error	62	10673	172.14478		
Corrected Total	66	17183			

Root MSE	13.12040	R-Square	0.3789
Dependent Mean	67.50746	Adj R-Sq	0.3388
Coeff Var	19.43547		

Parameter Estimates

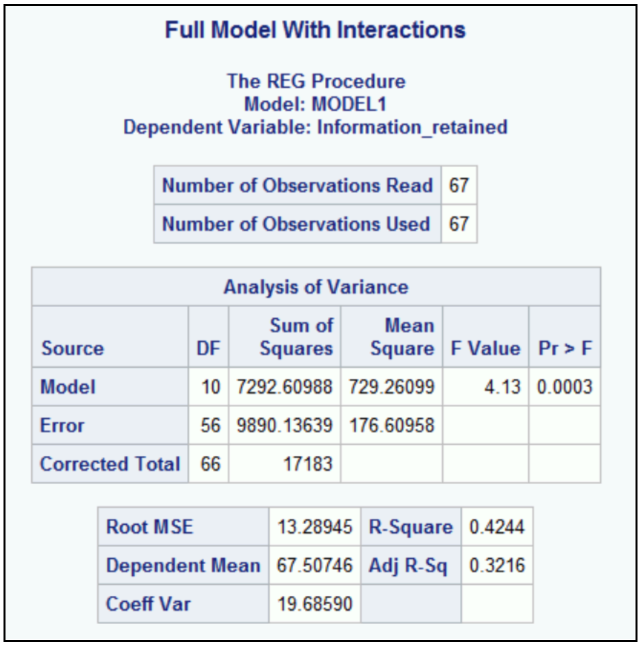
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	72.86297	2.06181	35.34	<.0001
studyL_some_D	1	-13.14756	3.80346	-3.46	0.0010
genreJazz	1	-9.75528	3.96744	-2.46	0.0167
genreHouse	1	21.30188	5.49512	3.88	0.0003
genreAfro	1	-25.04984	6.22720	-4.02	0.0002

Model Equation

Information_retained = 72.86297 - 13.14756(studyL_some_D) - 9.75528(genreJazz) + 21.30188(genreHouse) - 25.04984(genreAfro)

Model of Significant Predictors With Interactions

Adding interactions into the model raised the p-value only slightly but increased the accuracy of the model by nearly five percent. The R-square value increased from 0.3789 to 0.4244. Though adding interactions increased the accuracy of the model, when we conducted hypothesis testing none of the interactions were statistically significant.



Full Model With Interactions

The REG Procedure
Model: MODEL1
Dependent Variable: Information_retained

Number of Observations Read	67
Number of Observations Used	67

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	7292.60988	729.26099	4.13	0.0003
Error	56	9890.13639	176.60958		
Corrected Total	66	17183			

Root MSE	13.28945	R-Square	0.4244
Dependent Mean	67.50746	Adj R-Sq	0.3216
Coeff Var	19.68590		

Final Fitted Model

Information_retained = 72.86297 - 13.14756(studyL_some_D) - 9.75528(genreJazz) + 21.30188(genreHouse) - 25.04984(genreAfro)

CONCLUSION

Our first model did not indicate that any of our predictors had a significant effect on the total time computer and data science students spend studying. After switching our response variable to be the percentage of information students felt they retained from their studying we were able to find a predictive model. In that model the significant predictors were whether students listened to one of the music genres Jazz, House music, or Afrobeat, and whether students listened to both music with lyrics and music without while studying.

Switching between music with lyrics and without, listening to Jazz, and listening to

Afrobeats while studying were negatively correlated with information retention while house music was positively correlated. While all three genres include both music with and without lyrics both Jazz and Afrobeat songs tend to be busier and have more elements and variety than House music. Repetition and a strong beat that is equally loud if not louder than any lyrics are primary elements of most house music. It could be that the repetitive elements and strong beat are less distracting for many students, or that it actively helps some students study more effectively.

Even though all four predictors had very low P-values, indicating that they were highly influential on the model, the final model itself only accounted for 37.89% of the variance in the dataset. The low accuracy can be partially explained by the fact that our data relies heavily on self-reported, subjective data but there are also many factors that impact our response variable (both the first and the second) which we did not account for. For example, information retained may be highly dependent on the subject being studied, other external factors such as sleep, caffeine consumption or stress levels. Additionally we did not account for other outside factors such as students who listen to multiple genres while studying finding that they retain different amounts of information when listening to their different preferred genres.

If we had the chance to do this project again, we would choose not to use as many narrow categorical variables such as specific genres, preferred study windows, or categorical scales. Instead we would try to include only a few categorical levels with minimal levels and focus more on continuous quantitative variables.

As a result, we would also choose to broaden our research question to *“What factors are most useful in predicting how much information students feel they retain from studying”*. Additionally we would spend more time exploring possible predictors before collecting data in order to hopefully create a dataset that encompassed a greater number of possible predictors.