

*What Predictors Help in Predicting Scores for the 2024
JAMB Examination Session?*

Group 12
Project B
STA4164
Fall 2024

Table of Contents

<i>Abstract</i>	3
<i>Introduction and Motivation</i>	4
<i>Data Description</i>	6
<i>Exploratory Data Analysis</i>	8
<i>Model Diagnostics and Model Selection</i>	11
<i>Model Reliability</i>	19
<i>Results, Summary, and Interpretations</i>	20
<i>Conclusion and Limitations</i>	22
<i>Works Cited</i>	24
<i>Project Member Effort</i>	25

I. Abstract

The data set utilized in this research was a data simulation created by Idowu Adamo using the pandas and NumPy libraries in Python and was drawn on the JAMB 2024 exam session and was published onto Kaggle. The dataset includes 5,000 observations and recorded various factors such as whether a student attended public or private school as well as how heavily parents were involved in their child's education and saw how they would impact a student's score. With this data set, we had aimed to create a reliable multiple linear regression to predict a student's score on the JAMB given their educational and home background. This was driven by our motivation as students as we are aware of how various structural forces can impact a student's performance on exams such as a student's socioeconomic status and access to extra learning resources which can aid in getting a higher score on exams. After creating training and testing sets within the data and using backwards selection, a final model of 7 predictors was created, with 4 numerical and 3 categorical variables. The final model had a mean absolute error (MAE) of 31.00 points for its training set and 31.62 for its testing set, making it a reliable model to predict JAMB score as its MAE values are similar. The final model also contained an MSE value of 1481.780, which within the context of the large dataset, helps in validating the linear model as reliable since it is considered low.

II. Introduction and Motivation

The Unified Tertiary Matriculation Examination, more popularly known as the JAMB Exam is a 180-question exam that Nigerian students take within 2 hours in order to obtain a score from 0 to 400 in order to be placed into tertiary institutions such as college, with a minimum score of 140 needed in order to be admitted to university (Joint Admissions and Matriculation Board). This exam is similar to what high school students experience in the United States with the SAT (Scholastic Aptitude Test) as they submit their scores from the exam along with their college application for admission into an undergraduate university. From the dataset obtained for the 2024 JAMB examination session, the scores of 5000 students were released along with several potential factors that could have influenced students' scores such as the type of school a student attended (public or private), how many hours spent studying, and parent involvement (low, medium, or high) (Adamo). The purpose for this research is to predict how these given factors contribute to the overall score on the JAMB exam so that a plausible prediction can be made for a student's score on the exam given their background for upcoming examination sessions.

There are a variety of implications that can be made based upon this research as it can mostly be useful for providing better quality education to students. One of the main benefits that will come from this research, is that it would allow Nigerians students to be able to predict their JAMB score given their educational and home background which they can then use to determine what else they can do in order to help boost their score. This research can also help with providing performance data to the Nigerian government and also international groups such as the International Finance Facility for Education, or IFFed, which provides grants toward education to lower-income countries to help in delivering educational funding and provide possible demand

for a reform in Nigeria's education system as the data analysis will allow them to observe and analyze patterns of student performance across different socioeconomic backgrounds and help in providing funding and potential education reform to those in need so that they can improve their scores. Furthermore, by highlighting which key factors influence the prediction of JAMB scores, educators and policy makers are able to determine what changes can be made in order to improve student performance in education through curriculum updates, teacher training, and funding to provide educational reform within Nigeria.

III. Data Description

The dataset we have selected to use on this project was sourced from Kaggle, an online forum where people can share and download a large variety of datasets and is a simulated dataset. The set uses a number of variables and statistics collected from a sample of students in Nigeria, along with their overall performance on a standardized test referred to as the JAMB. The dataset includes variables such as: study hours per week, attendance rate, teacher quality (rated from 1 to 4), the distance between the students home and the school, the type of school (public or private), school location (urban or rural), whether the student received additional tutoring or not, the students access to learning and materials, the level of parent involvement in the students education (low, medium, or high), the student's computer proficiency, student ID, age, gender, socioeconomic status, parental education level, and the assignment completion rate of each student.

While there are many valuable variables provided for predicting a student's test score, there are likely some irrelevant ones included in the dataset as well. Certain variables, like student ID and potentially gender is more than likely irrelevant as a predictor. Additionally, there may be some concern about age as it potentially creates collinearity, as it often does in datasets. Further testing for association between age, as well as other variables, will be conducted to assure that the variables are not collinear or confounding the prediction model. It is possible that some association exists between things like parental involvement and parental education level, as well as socioeconomic status and computer proficiency. These associations are not certain, but should be considered with caution.

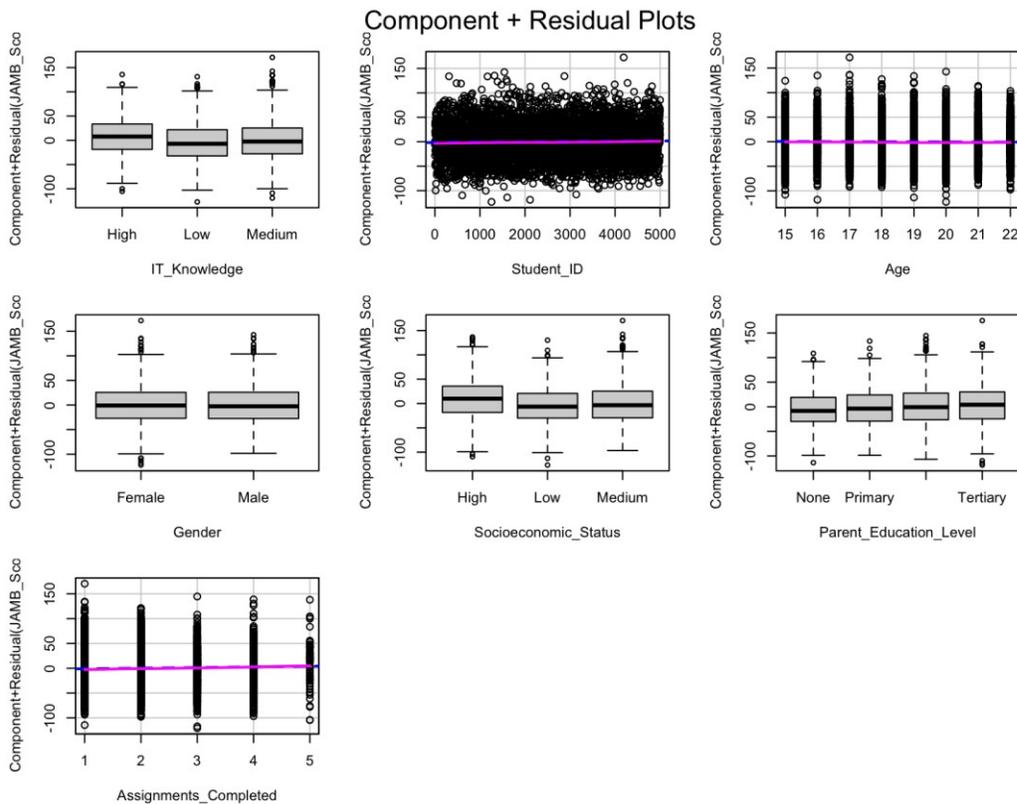
The dataset being used was generated using statistics from the 2024 Joint Admissions and Matriculation Board (JAMB) examination to predict students' performance. This notion is important to the reliability and accuracy of the data gathered, as it is compiled directly from the

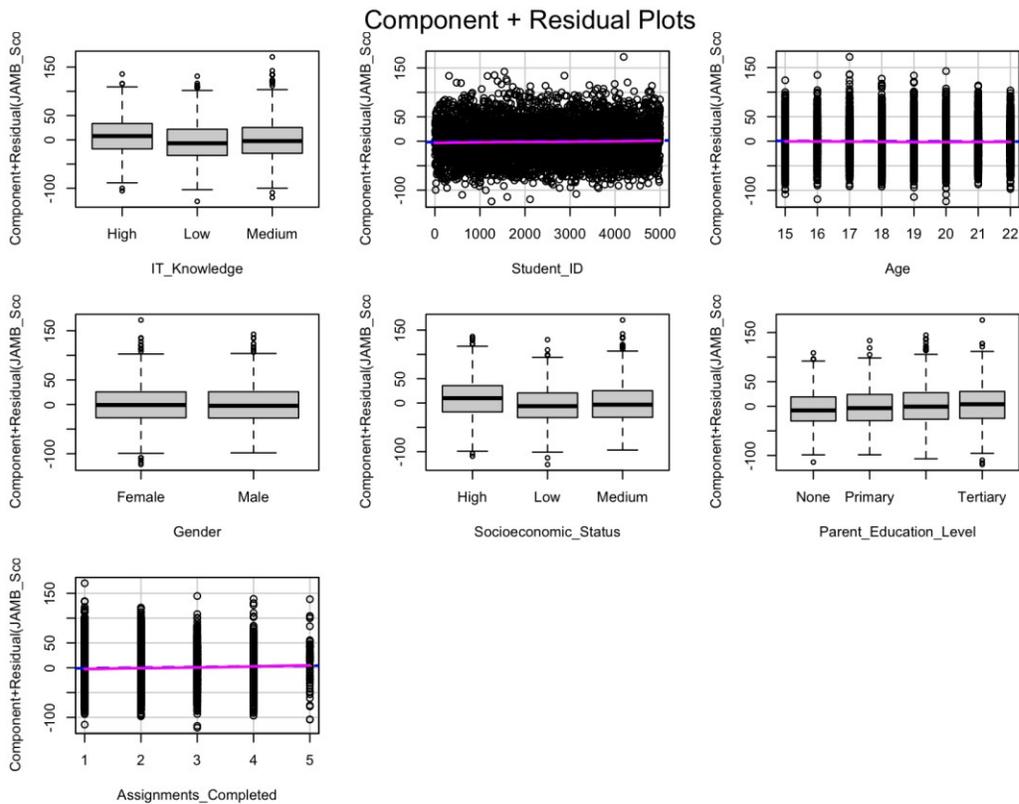
source of interest. The dataset is useful to our study's motivation because it provides information directly related to a student's performance on the JAMB, which is what we are aiming to predict with a regression model. By using this dataset as reference, we should be able to effectively conclude what is inherently important to the results a student produces on the test, thus allowing the Nigerian school system to better understand what is most important to a student's academic success.

IV. Exploratory Data Analysis

It appears that within the dataset, the lowest score for the 2024 exam session was 100 and the highest being 367 with the average student scoring 174.1 points on their exam. This shows that there are no unlikely or impossible values within the dataset, showing no potential outliers. To determine if any transformations are necessary to improve the model, partial residual plots were created to determine that the model would need any form of transformations to improve the initial full model.

Figure 1: Component and Residual Plots for Potential Predictors in the Initial Model





As seen by the plots, no transformations to the model are not necessary as the pink and blue lines in the plots overlap. Additionally, we are including three interaction terms to the model due to reasoning to believe that they could help in improving the linear model. The three terms we are adding are: attendance rate and teacher quality, parent involvement and attendance rate, and socioeconomic status and IT knowledge. Our reasoning for adding these interaction terms is due to the close relationship that they have with each other. For instance, attendance rate and teacher quality are being considered because some students may be more likely to go to class if they have a higher quality teacher that keeps students engaged in learning, which encourages student involvement. Attendance Rate and Parent Involvement is being considered because some students rely on their parents for transportation and if their parent is unable to take their child such as having to go to work or does not have a form of transportation, then their child will miss out on important instructional hours. Additionally, socioeconomic status and IT knowledge are

being considered because if a student's family is wealthier, then they are able to afford access to computers and online resources that could benefit their learning and understanding of educational concepts.

These inclusions to the model and the lack of transformations are an approach to keep the model as simple as possible while also being reliable and accurate. If we were to do any transformation or include other interaction terms, then the model might possibly become overfit and have a much higher chance of causing collinearity to appear, affecting the overall reliability and practicality of the model.

V. Model Diagnostics and Model Selection

In order to build the final full model, we first created a linear model that contained all the predictors that were in the dataset and observed the partial regression plots. From observing the plots, there seemed to be no obvious outliers within the dataset that would be necessary to remove and noted that no transformations would be necessary for the model. It was also noticed that constant variance was violated from the model as the cone-shaped figure seen in *Figure 2*, shows a violation. However, it appears that no other assumptions were violated within the plots.

Figure 2: *Residuals vs. Fitted Values Chart of Initial Full Model*

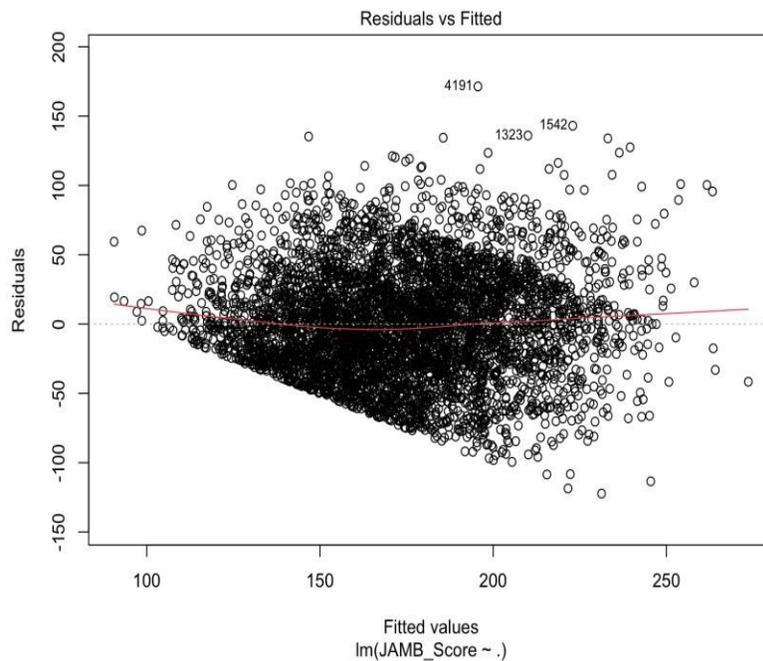
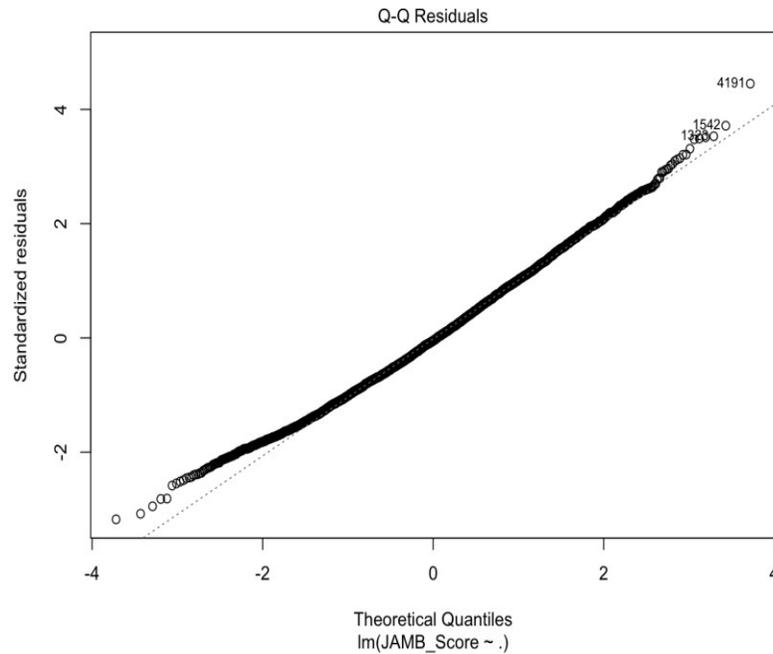


Figure 3: Normal Probability Plot of the Initial Full Model



After observing the plots, we considered using interaction terms between: attendance rate and teacher quality, parent involvement and attendance rate, and socioeconomic status and IT knowledge. These interaction terms were deemed potentially useful as there could be a potential relationship between the predictors and ergo, creating more accurate score predictions. After deciding what interaction terms to include for the potential final model, a second linear model was created using the first full model with the addition of the potential interaction terms. To check if any assumptions were violated within the new model, we observed the residual and normal probability plots of the new model and found that there were no violations to the assumptions. After checking the residual and normal probability plots, it appears that only constant variances seemed to have still been violated. To check for outliers within the residual plots, jackknife residuals, leverage, and cook's distance were used to find any points that would have needed to be removed from the data. In total, there were a total of 43 violations within the data set which were kept in the model due to the data set having 5,000 observations and how all

Figure 5: Leverage of Initial Full Model

```

> # 2(16)/5000 = 0.0064
> tail(sort(hatvalues(full_model)),n=1200) # 1,184 Violations
  390      3108      589      4365      2976      209      3794      3609      2189
0.006377742 0.006380134 0.006380780 0.006381798 0.006385119 0.006387606 0.006387882 0.006387967 0.006392514
  2527      731      3197      1783      4948      1344      1991      1075      4962
0.006393592 0.006394093 0.006395548 0.006396584 0.006398590 0.006398916 0.006399884 0.006402281 0.006403154
  3175      1441      424      4289      2607      4586      2029      2748      4295
0.006403447 0.006403766 0.006404261 0.006404400 0.006406757 0.006407563 0.006408191 0.006410636 0.006412067
  946      639      4246      2600      1474      1879      3222      4341      686
0.006412423 0.006412459 0.006413407 0.006413712 0.006415570 0.006415877 0.006417854 0.006418039 0.006418479
  1597      1832      3957      3778      2270      3838      4139      4049      3339
0.006418545 0.006419119 0.006421034 0.006421949 0.006424627 0.006425644 0.006425793 0.006426078 0.006426280
  96      2429      4432      1635      3531      1664      2650      4536      4932
0.006429971 0.006430341 0.006430475 0.006432534 0.006433554 0.006433850 0.006435464 0.006435743 0.006436958
  871      664      4322      2386      1554      2633      3009      2079      1258
0.006437532 0.006437654 0.006439116 0.006440419 0.006442379 0.006442637 0.006444416 0.006445717 0.006448114
  2778      1216      4960      2982      955      4490      822      4449      1524
0.006448178 0.006449529 0.006449676 0.006452000 0.006452369 0.006452967 0.006453547 0.006454489 0.006454627
  2328      4362      2333      3594      2936      2906      1516      3670      2754
0.006454667 0.006456359 0.006457717 0.006458935 0.006460163 0.006462413 0.006462479 0.006463471 0.006465300
  4026      3006      3533      4884      2173      2771      3646      4467      3390
0.006467800 0.006468161 0.006469020 0.006469302 0.006469549 0.006470864 0.006470940 0.006474060 0.006474136
  223      1577      2918      786      1631      2208      1213      832      4489
0.006474917 0.006476198 0.006477780 0.006480197 0.006481541 0.006482580 0.006483880 0.006484090 0.006485010
  4781      87      457      3747      3034      4509      94      3588      1169
0.006485627 0.006485988 0.006486305 0.006486595 0.006486783 0.006488711 0.006489186 0.006491158 0.006491735
  4358      4300      736      2900      4462      1360      2782      3176      4321
0.006492121 0.006493894 0.006494242 0.006495413 0.006495559 0.006496063 0.006500406 0.006501798 0.006502088
  2740      3148      1340      3978      1341      1600      1752      4981      4904
0.006502985 0.006503351 0.006503467 0.006504882 0.006505173 0.006506022 0.006508565 0.006509845 0.006513051
  2557      2730      2894      3461      3224      4913      4311      1463      4927
0.006514517 0.006514823 0.006514873 0.006516149 0.006517162 0.006518112 0.006518511 0.006519352 0.006519435

```

Figure 6: Cook's Distance of Initial Full Model

```

> tail(sort(cooks.distance(full_model)),n=10) # No violation
  1323      1606      1770      2875      1542      1392      1168      1257      1587
0.002179852 0.002325663 0.002402934 0.002431671 0.002614819 0.002717890 0.003018001 0.003037426 0.003057904
  4191
0.003356078

```

While there were violations within leverage and the jackknife residuals, cook's distance had none so all points were kept despite some violations as with having a large dataset, it was expected that there would be violations. To improve the line's reliability and accuracy, backwards selection was conducted to create a final model that would be significant in predicting JAMB scores. Since the dataset contains 5,000 observations, 20% of the observations were randomly selected and placed into a training set while the other observations were placed into a testing set. Subsequently, using the `ols_step_backward_p()` function in R with a p-value of 0.1, the predictors: parent education level, parent involvement, gender, age, assignments completed,

school type, school location, distance to school, and the interaction terms attendance rate and teacher quality and attendance rate and parent involvement were removed from the model.

Figure 7: Parameter estimates of the Final Model

```

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      47.0084    12.3355   3.811 0.000147 ***
Study_Hours_Per_Week  1.7728     0.1317  13.462 < 2e-16 ***
Attendance_Rate    1.0629     0.1316   8.078 1.91e-15 ***
Teacher_Quality    9.3306     1.2636   7.384 3.25e-13 ***
Distance_To_School -0.5380     0.2546  -2.113 0.034852 *
Extra_TutorialsYes  5.2254     2.4594   2.125 0.033861 *
IT_KnowledgeLow   -13.2484    3.2441  -4.084 4.79e-05 ***
IT_KnowledgeMedium -2.4302     3.2138  -0.756 0.449720
Socioeconomic_StatusLow -16.5490    3.3428  -4.951 8.69e-07 ***
Socioeconomic_StatusMedium -17.6086    3.3363  -5.278 1.61e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.69 on 990 degrees of freedom
Multiple R-squared:  0.3267,    Adjusted R-squared:  0.3205
F-statistic: 53.37 on 9 and 990 DF,  p-value: < 2.2e-16

```

Finally, the final model was created, composed of seven predictors, four numerical and three categorical variables. The final model removed the interaction terms we had added, showing that they were not significant in predicting student's JAMB scores. The final model is also more simplified than the initial model, removing most of the predictors, a majority of them being categorical.

Figure 8: VIF Values of Predictors on Final Full Model

```

Tolerance and Variance Inflation Factor
-----
                Variables Tolerance    VIF
1      Study_Hours_Per_Week 0.9525940  1.049765
2      Attendance_Rate    0.9706800  1.030206
3      Teacher_Quality    0.9683918  1.032640
4      Distance_To_School 0.9918944  1.008172
5      Extra_TutorialsYes 0.9941604  1.005874
6      IT_KnowledgeLow    0.6081023  1.644460
7      IT_KnowledgeMedium 0.6069273  1.647644
8      Socioeconomic_StatusLow 0.5595141  1.787265
9      Socioeconomic_StatusMedium 0.5612289  1.781804

```

While observing the VIF values of the model's predictors, it appears that collinearity is not present within the model as none of the VIF values are greater than 10, showing no collinearity between predictors. Hence, the reliability of the model is further enhanced as the model is free from any collinearity issues.

Figure 9: Residual vs. Fitted Plot of Final Full Model

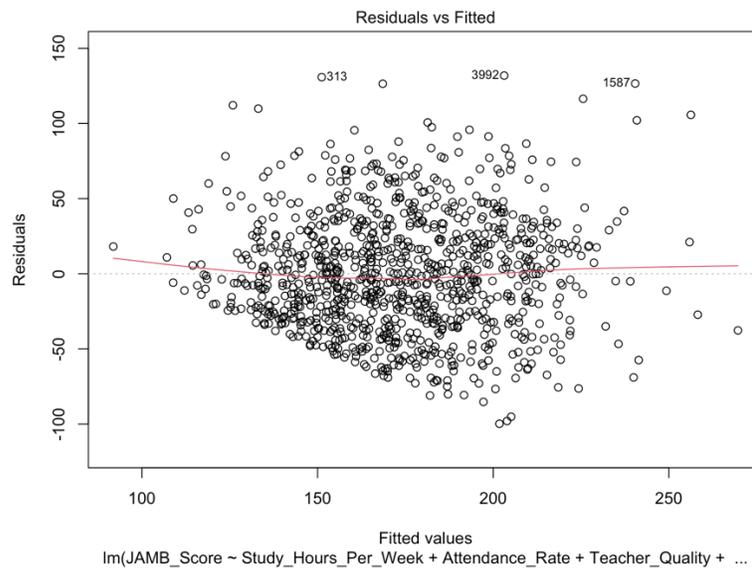
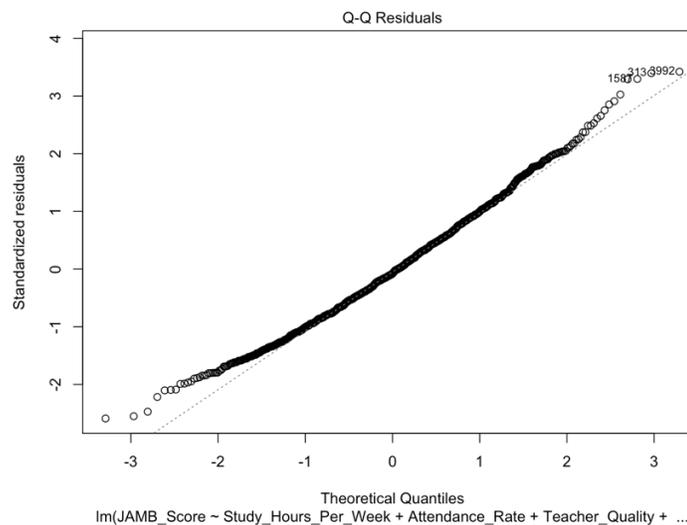


Figure 10: Normal Probability Plot of the Final Full Model



Constant variance is still violated within the full model as there is still a cone-shaped pattern within the residual plot. To remedy this, we conducted a box transformation that recommended a logarithmic transformation on Y as shown below:

Figure 11: Box Cox Graph

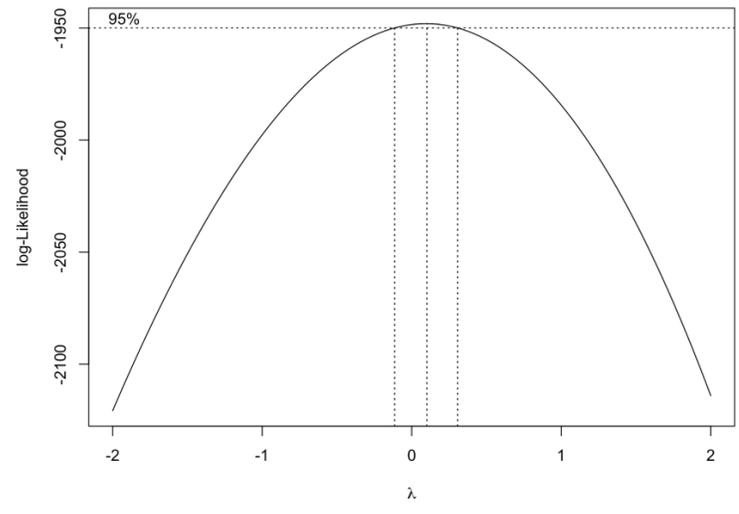


Figure 12: Residual vs. Fitted Values Plot of Transformed Model

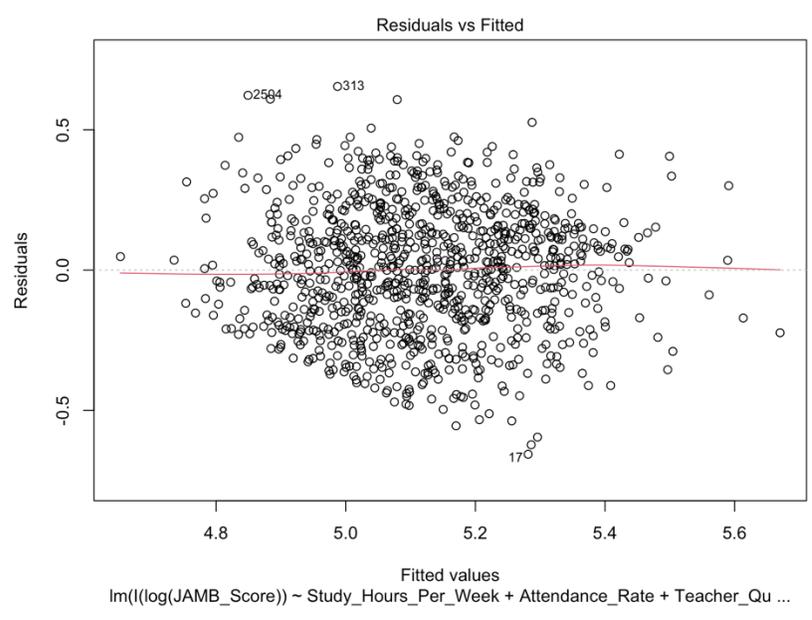
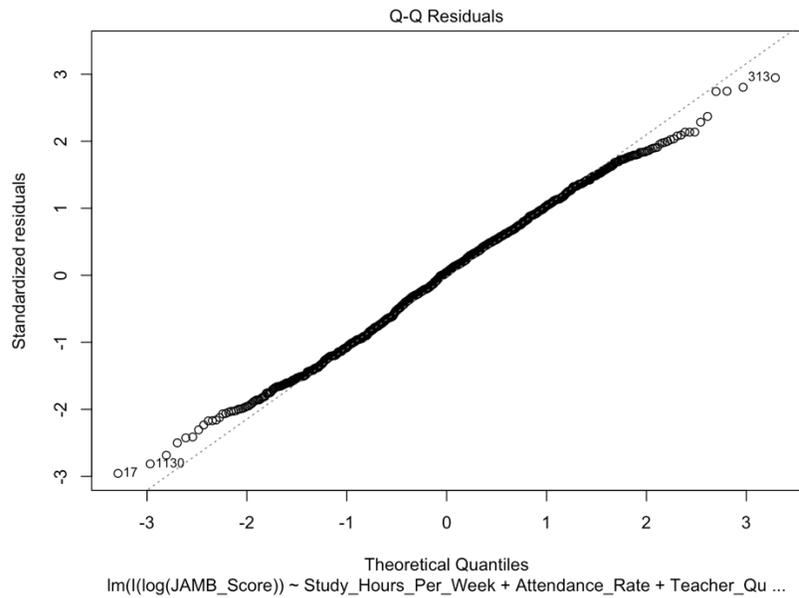


Figure 13: Normal Probability Plot of the Transformed Model



It appears that even with the transformation, constant variance is still violated. To continue in our effort in maintaining a simple model, the normal final model will be used for our JAMB score predictions instead of the one recommended by the Box Cox transformation.

VI. Model Reliability

In order to determine if the final model built is reliable to predict a student's performance in the JAMB 2024 examination session, it was important to make sure that the model produced was deemed overfit which could cause inaccuracies with its predictions. To confirm that the model was reliable enough, we utilized the `mae()` function within R to calculate the mean absolute error of the training and testing sets to see if there was a vast difference between the two. The MAE of the training set was 31.00 points while the testing set had a mean absolute error of 31.62 points. Within the context of scoring between 0 to 400 points on the exam, those values are considered well and also show that the model is reliable as they are similar in value.

Since both MAE values are similar, the model is deemed reliable and therefore, can be used to help with predicting a student's exam score for the 2024 examination session. Additionally, by observing the model's R^2 , which has a value of .3205, it shows that 32.05% of the variability in JAMB scores can be explained by the predictors within the model. Despite it being considered a "low" value the model itself is predicting the values of the JAMB scores well.

VII. Results, Summary, and Interpretation

The final model for the dataset is given below:

$$\hat{y} = 47.01 + 1.77X_1 + 1.06X_2 + 9.33X_3 - 0.54X_4 + 5.23Z_1 - 13.25Z_2 - 2.43Z_3 \\ -16.55Z_4 - 17.61Z_5$$

$$Z_1 = \begin{cases} 1 & \text{if Yes} \\ 0 & \text{if Otherwise} \end{cases} \quad Z_2 = \begin{cases} 1 & \text{if Low IT Knowledge} \\ 0 & \text{if Otherwise} \end{cases} \quad Z_3 = \begin{cases} 1 & \text{if Medium IT Knowledge} \\ 0 & \text{if Otherwise} \end{cases}$$

$$Z_4 = \begin{cases} 1 & \text{if Low Socioeconomic Status} \\ 0 & \text{if Otherwise} \end{cases} \quad Z_5 = \begin{cases} 1 & \text{if Medium Socioeconomic Status} \\ 0 & \text{if Otherwise} \end{cases}$$

This model includes predictors with seven predictors that includes three categorical variables. The final model also has predictors which have p-values that are significant at the $\alpha = 0.05$ significance level, except for a student who has a medium socioeconomic status. This is a vital subject for interpretation as it shows that the predictors within the model are significant in predicting student's scores for the 2024 exam.

The model also reveals how certain aspects of a student's background can positively or negatively impact their score. For instance, given that a student had access to extra tutorials, their JAMB score would be 5.23 points higher than a student who did not receive extra tutorials, while holding all other variables constant. Similarly, if a student had low IT knowledge, they would be 13.25 points lower than a student with a high level of IT knowledge while a student with medium knowledge would only be 2.43 points lower than a student with high knowledge. This information could be deemed extremely important to educators and government officials as this gives possible suggestions in ways to improve scores for students and create new ideas and resources to implement within schools and communities so that student scores are improving such as giving more resources to lower income areas and providing education on computers to students so that they can access online resources. It is important to note that this study is

observational so we cannot make any causal inference as the recommendations are purely suggestions and have no guarantee that they will affect any outcome.

VIII. Conclusions and Limitations

Throughout this study, we have learned many lessons on how to conduct tests and interpret and analyze data. This study taught us a lot about the research process and the steps that must be taken in order to ensure that a reliable, accurate, and helpful model is created that can help in better understanding the world around us. The final model that was created was a result of our group's effort in building a reliable model that was appropriate to use and analyze. This research taught us what it means to work in a group and the importance of communication in order to get ideas across so that we can share potential ideas and solutions to help make our model stronger and more accurate.

While this research did build a model that is good at predicting student's scores, there were some limitations in the study that would need improvement to build a better model. The main limitation that we encountered was that the given data set is a simulation that was created using the pandas and NumPy library within Python along with actual results from the 2024 examination session. This caused the model to have a "clean" dataset that did not have any other assumption violations besides a violation of constant variance and any issues such as residual values that would have violated cook's distance. This prevented the production of a more realistic model that would have been much more accurate in predicting scores that potentially would have needed transformation or the addition or removal of variables and interaction terms. Further potential improvements would have been in creating new interaction terms such as socioeconomic status and access to learning materials which would suggest other potential reasons as to why some students score higher or lower than others. However, due to time constraints and as an effort to keep the model simple and not complex and potentially overfit, only three interaction terms were picked.

Despite the limitations within the study, a reliable model was built through our collaboration as a group. While this study is observational, there are some recommendations that can be suggested within our model that potentially highlight the importance that some of these predictors have on a student's outcome on the JAMB exam. Through these suggestions, we can possibly create and harness solutions that better help out students in their educational career so that they will have the tools and resources they need in order to thrive.

Works Cited

Adamo, Idowu. "Students Performance in 2024 JAMB." *Kaggle*,

[https://www.kaggle.com/datasets/idowuadamo/students-performance-in-2024-](https://www.kaggle.com/datasets/idowuadamo/students-performance-in-2024-jamb?resource=download&select=jamb_exam_results.csv)

[jamb?resource=download&select=jamb_exam_results.csv](https://www.kaggle.com/datasets/idowuadamo/students-performance-in-2024-jamb?resource=download&select=jamb_exam_results.csv).

Joint Admissions and Matriculation Board. *JOINT ADMISSIONS AND MATRICULATION BOARD*,

<https://www.jamb.gov.ng/>. Accessed 19 November 2024.

PROJECT MEMBER EFFORT

We, the project teams members, certify that below is an accurate account of the percentage of effort contributed by each team member in the project and report

Project Team Member	Percentage of Total Effort
Reymond Ramirez	33.33%
Max Schneidt	33.33%
Ty Spradley	33.33%
Nathaniel Darch	0%