What Factors Influence A Student's Academic Success?

Nour Kahky

900221042

Omar Moustafa

900222400

MACT 4231

Fall 2024

Statement of the Problem:

What Question(s) Can Be Answered by the Analysis of the Data?

1. Which factors have the greatest influence on a student’s academic performance?
2. How do social and economic factors influence a student’s academic performance?
3. How do school-related factors outside a student's control influence their academic performance?
4. To what extent do lifestyle factors influence a student’s academic performance?
5. Does participating in extracurricular activities truly help a student’s academic performance?
6. Does parental involvement help a student achieve higher academic success, specifically higher final exam scores?
7. Are there existing relationships, tendencies, or patterns between the categorical variables and final exam results?

Background or Literature Review:

Academic performance has been a globally studied topic in countless fields where studies have revealed that certain factors influence the extent of success that student achieves in their examinations such as the effort they exert in their studies and attendance, their economic status and lifestyle factors, parental involvement, the influence of their surrounding peers, the quality of their school and teachers, and many others. The dataset utilized throughout this investigation combines the previously listed factors to offer an overarching perspective of what influences student success. The study works to confirm relationships that are publicly considered to be existing and accurate and determine new patterns revealed by this dataset.

Complete Description of the Data:

Complete Source of the Data:

The dataset that will be used and analyzed is called “StudentPerformanceFactors.csv” which consists of a long list of variables that are well-known to influence the academic performances of students. These influential variables range from the genuine effort and preparation of the student to their previous academic performances to their lifestyle and economic status to the qualities of their school and teachers to many more. Additionally, the dataset includes a total of 20 features along with a total of 6607 observations.

What Are The Observations?

Each observation captures a student's academic, lifestyle, and demographic characteristics. These observations help one understand the relationships between the predictors and the response variable, which is Exam_Score. They also help conclude what factors impact a student’s academic performance and formulate the ideal scenario for their academic success.

Full Definition of the List of Variables:

Variable Name:	Variable Type:	Unit of Measurement:	Predictor or Response Variable?	Hypothesized Relationship w/ the Response:
Exam_Score	Quantitative	Percentage	Response Variable	*This is the response variable*
Hours_Studied	Quantitative	Hours	Predictor	Positive relationship as it indicates effort put into the final exam preparation

Attendance	Quantitative	Percentage	Predictor	Positive relationship as it indicates exposure to the final exam content
Parental_Involvement	Ordinal Categorical	N/A	Predictor	Positive relationships would help their offspring study well
Access_to_Resources	Ordinal Categorical	N/A	Predictor	Positive relationship as it indicates exposure to the final exam content
Extracurricular_ Activities	Nominal Categorical	N/A	Predictor	Positive relationship as it would lead to more educational exposure
Sleep_Hours	Quantitative	Hours	Predictor	Positive relationship as the more sleep before an exam, the more focused the student will be
Previous_Scores	Quantitative	Percentage	Predictor	Positive relationship as one that scored well before and has the habit of scoring well again on their final
Motivation_Level	Ordinal Categorical	N/A	Predictor	Positive relationship as it indicates the

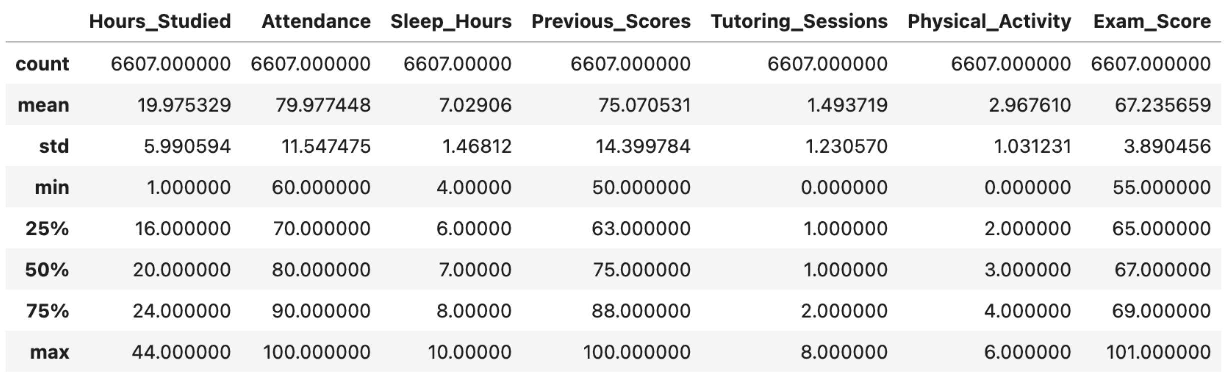
				willingness of the student to exert effort in studying
Internet_Access	Nominal Categorical	N/A	Predictor	Positive relationship as it would lead to more educational exposure
Tutoring_Sessions	Quantitative	Sessions	Predictor	Positive relationship as it indicates effort put into the final exam preparation
Family_Income	Ordinal Categorical	N/A	Predictor	Positive relationship as higher family income is associated with better academic performance
Teacher_Quality	Ordinal Categorical	N/A	Predictor	Positive relationship as the better the teacher, the more the student should be learning
School_Type	Nominal Categorical	N/A	Predictor	Positive relationships as the better the school, the more the student should be learning
Peer_Influence	Nominal Categorical	N/A	Predictor	Positive as positive

				peer influence can help one gain motivation in their studies and performances
Physical_Activity	Quantitative	Hours	Predictor	Positive relationships as physical activity lead to better concentration and motivation
Learning_Disabilities	Nominal Categorical	N/A	Predictor	Negative relationship as it would prevent the student from fathoming the material as well as possible
Parental_Educational_Level	Ordinal Categorical	N/A	Predictor	Positive relationships higher educated parents will push their children to study well
Distance_from_Home	Ordinal Categorical	N/A	Predictor	Negative relationship as the further someone lives from school, the less time they have to study or revise
Gender	Nominal Categorical	The gender of the student	Predictor	N/A

Table 1.1

Data Analysis

Descriptive Statistics



	Hours_Studied	Attendance	Sleep_Hours	Previous_Scores	Tutoring_Sessions	Physical_Activity	Exam_Score
count	6607.000000	6607.000000	6607.00000	6607.000000	6607.000000	6607.000000	6607.000000
mean	19.975329	79.977448	7.02906	75.070531	1.493719	2.967610	67.235659
std	5.990594	11.547475	1.46812	14.399784	1.230570	1.031231	3.890456
min	1.000000	60.000000	4.00000	50.000000	0.000000	0.000000	55.000000
25%	16.000000	70.000000	6.00000	63.000000	1.000000	2.000000	65.000000
50%	20.000000	80.000000	7.00000	75.000000	1.000000	3.000000	67.000000
75%	24.000000	90.000000	8.00000	88.000000	2.000000	4.000000	69.000000
max	44.000000	100.000000	10.00000	100.000000	8.000000	6.000000	101.000000

Table 1.2

The table above provides a brief yet valuable summary of the distribution of the quantitative independent variables with the last column being the response or dependent variable. From the table, it can firstly be interpreted that the row "count" reveals that there are 6607 observations for each variable, confirming a statement just previously made. Secondly, the row "mean" gives the average value of each of these variables, providing a sense of the central tendency. After that, the row "std" gives the standard deviation of each of these variables, providing a measurement of the spread of the observations around the average. As for the rest of the rows, these work together to state the values of the range, median, and upper and lower quartiles of each of these variables. Therefore, by analyzing this initial table that is filled with descriptive statistics, valuable insights about the distribution of the quantitative predictor variables can be interpreted and certain patterns and trends can be revealed, where these can come in handy when making predictions and conclusions about the factors that truly influence a student's academic performance.

Graphs Before Fitting a Model to the Data

1. *Assumption #1: Checking Linearity*

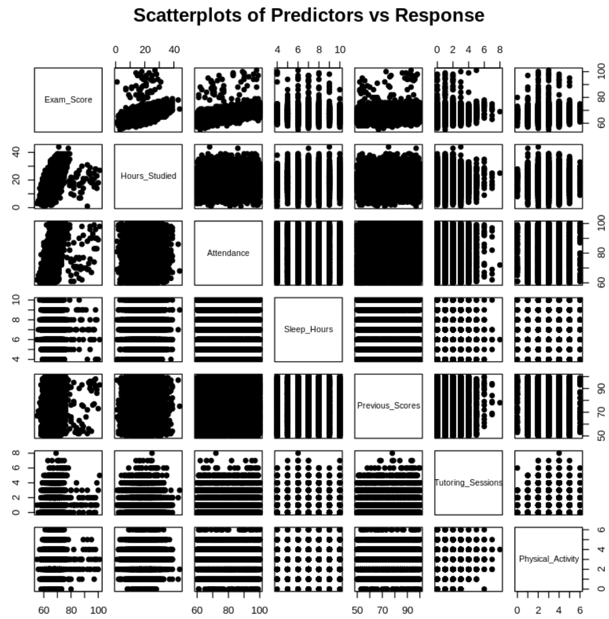


Figure 1.1

Comments on Figure 1.1:

The first assumption to check is Linearity. Therefore, this figure represents a pairwise scatterplot of the quantitative predictors against the response variable, Exam_Score. Giving a brief analysis of this figure, it shows that all of the predictors have linear relationships with the response variable, except for two, which are Sleep_Hours and Physical_Activity. For example, as Attendance increases, so does Exam_Score, but it is not as straightforward with Study_Hours or Physical_Activity. Therefore, square root transformations were applied to Study_Hours and Physical_Activity so that now all quantitative variables would satisfy the linearity assumption.

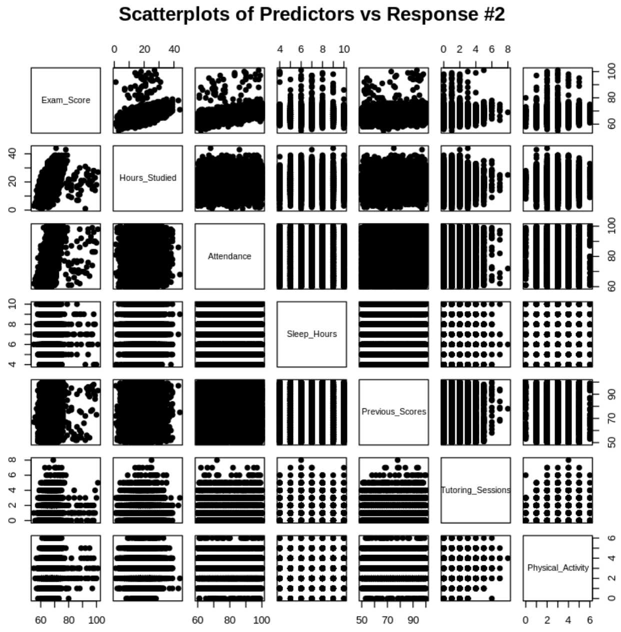
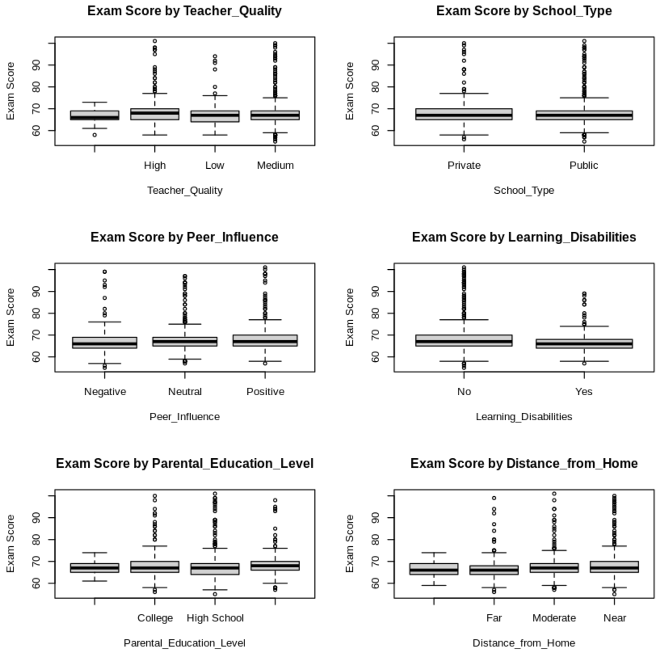
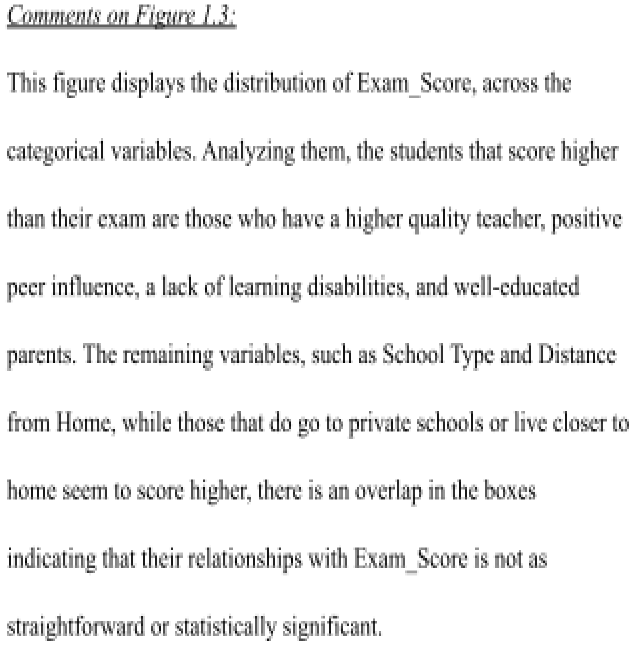


Figure 1.2

Comments on Figure 1.2:

The figure on the left is an updated version of the previous figure which this updated version, unlike the previous one, full satisfies the linearity assumption of standard regression as it illustrates the scatter plots after the square root transformation was applied to Sleep_Hours and Physical_Activity as they were the two variables that had previously not satisfied the assumption of linearity whereas not they do.





__Comments on Figure 1.3:__

This figure displays the distribution of Exam_Score, across the categorical variables. Analyzing them, the students that score higher than their exam are those who have a higher quality teacher, positive peer influence, a lack of learning disabilities, and well-educated parents. The remaining variables, such as School Type and Distance from Home, while those that do go to private schools or live closer to home seem to score higher, there is an overlap in the boxes indicating that their relationships with Exam_Score is not as straightforward or statistically significant.

Figure 1.3

2. ***__Assumption #2: Checking Independence of Observations__***

Moving on, the second assumption is to check the independence of the observations. To achieve this, the Durbin-Watson test was conducted and the output was as follows:

```
        Durbin-Watson test

data:  lm(Exam_Score ~ 1, data = df)

DW = 1.9718, p-value = 0.1255

alternative hypothesis: true autocorrelation is greater than 0
```

Analyzing the above output, it can be interpreted that the Durbin-Watson, DW, statistic is equal to 1.9718, which is very close to being equal to 2 which satisfies the assumption as “a value of DW = 2 indicates that there is no autocorrelation” (CFI Team). Upon that, the p-value is equal to 0.1255 which is greater than the typical significance level of α = 0.05 further satisfying the assumption. Therefore, these together imply that autocorrelation in the residuals is not evident and unlikely to be the case, satisfying the second assumption.

3. ***Assumption #3: Checking Homoscedasticity***

The next check was about the homoscedasticity of the model. To do this, the following plot was reached and analyzed to see if the data met this assumption or required a transformation:

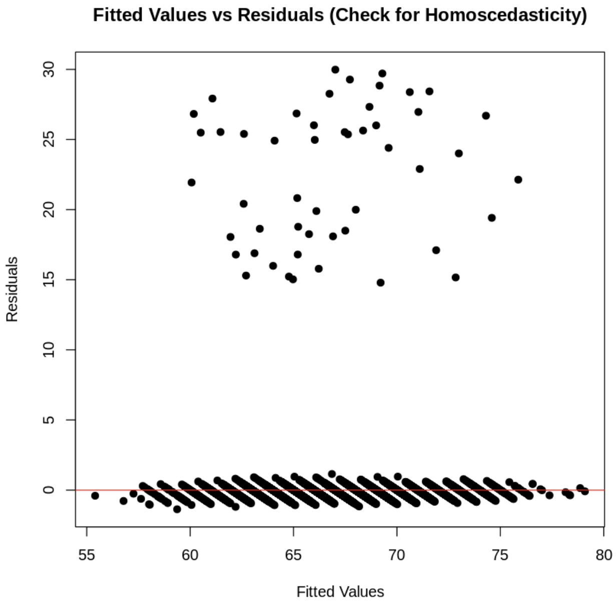
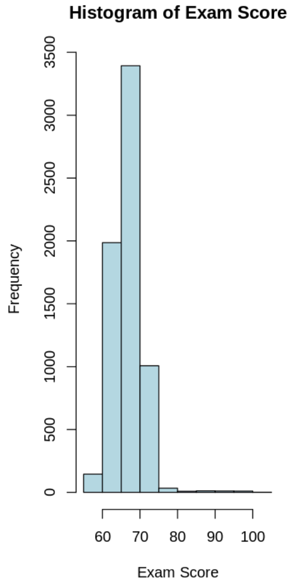


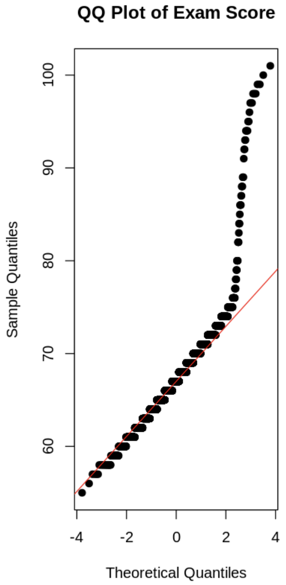
Figure 1.4

Comments on Figure 1.4:

Based on the figure to the left which is a plot that has the Fitted Values along the x-axis and the Residuals along y-axis to allow one to investigate if the assumption of homoscedasticity is satisfied or not. From the plot, it can be observed that the residuals are randomly scattered around the red horizontal line at zero without any real pattern or trend that can be deciphered or interpreted. This leads into the second aspect of homoscedasticity which is that the variance is indeed constant across the various levels of the Fitted Values. Combining these two together, it is reasonable to say that the assumption is indeed satisfied.

4. ***Assumption #4: Checking the Normality of the Response Variable***





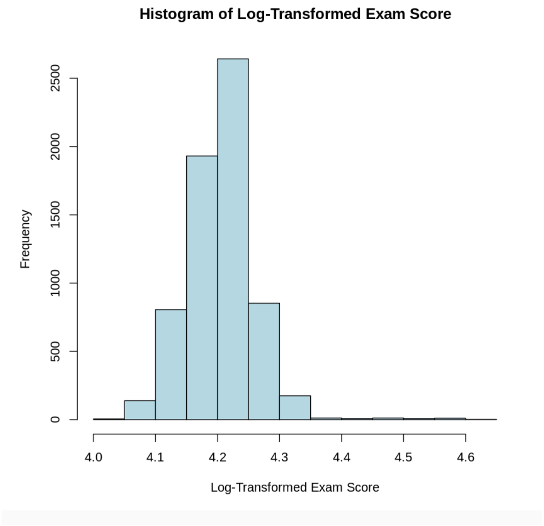
Comments on Figures 1.5 & 1.6:

The fourth of the ten standard assumptions of regression constitutes checking the normality of the response variable. The following visualizations were the output plots to check interpret whether or not the response variable is normally distributed, but as it can be seen, it was not the case as it can be seen in the two plots directly on the left. Specifically regarding the histogram, Firstly, a histogram was chosen to visualize the response variable because the data has a total of 6607 observations which is well over the 150 minimum for creating histograms. Analyzing the histogram itself, it can be seen that most of the scores fall between 65-70 and the most of the remaining scores are either between 60-65 and 70-75. A very small number of students scored 75 and above and a very small number of students scored below 65. Overall, the histogram of final exam scores shows a slightly right-skewed distribution, indicating that a majority of students performed moderately well and passed, while a small number of students achieved either much higher or lower scores. Due to this particular skewness, a transformation is necessary.

Figures 1.5 & 1.6 (Before Transformation)

Figure 1.5 **Figure 1.6**

Therefore, the response variable, Exam_Score, was put through a log transformation for it now satisfies the normality assumption, which was indeed achieved as seen in the two plots below. From these plots, which are new and updated versions of the Histogram and Q-Q plot above, it can be observed that the log transformation successfully improved the normality of the response variable which was necessary as normality is crucial for linear regression and several other statistical tests and models.



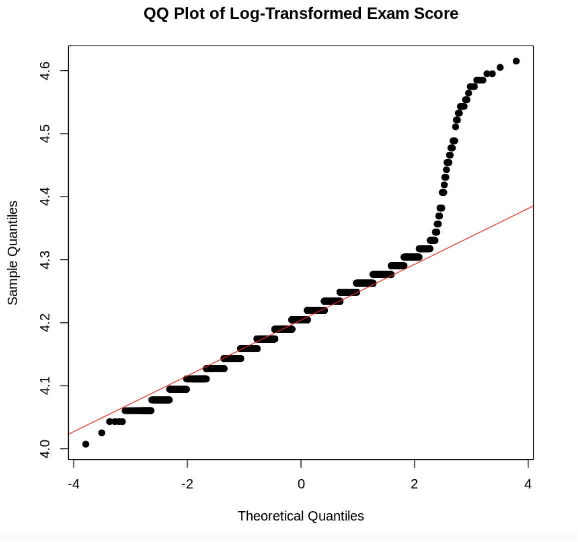


Figure 1.7 **Figure 1.8**

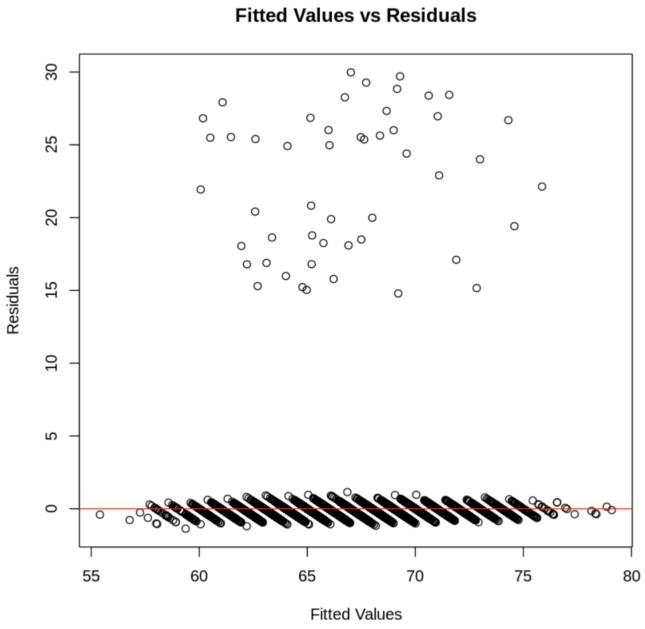
5. ***Assumption #5: No Multicollinearity***

Next, it needed to be checked that there was a lack of multicollinearity. For this, computing the correlation matrix is a reasonable approach as it would reveal whether or not two or more quantitative variables are highly correlated with one another. The output shows that the values are quite small suggesting that the correlation between the predictors falls somewhere between low and moderate, not high. Building on this, it points to the circumstance there is no major issue regarding multicollinearity within the model and therefore, the fifth standard assumption is indeed satisfied.

```
                  Hours_Studied   Attendance   Sleep_Hours Previous_Scores
Hours_Studied       1.000000000 -0.009907859  0.0109766893      0.02484578
Attendance         -0.009907859  1.000000000 -0.0159178258     -0.02018610
Sleep_Hours         0.010976689 -0.015917826  1.0000000000     -0.02175034
Previous_Scores     0.024845782 -0.020186103 -0.0217503431      1.00000000
Tutoring_Sessions  -0.014282264  0.014323509 -0.0122161120     -0.01312233
Physical_Activity   0.004624390 -0.022434703 -0.0003780652     -0.01127373
                   Tutoring_Sessions Physical_Activity
Hours_Studied            -0.01428226      0.0046243903
Attendance                0.01432351     -0.0224347027
Sleep_Hours              -0.01221611     -0.0003780652
Previous_Scores          -0.01312233     -0.0112737339
Tutoring_Sessions         1.00000000      0.0177329453
Physical_Activity         0.01773295      1.0000000000
```

6. ***Assumption #6: Checking No Autocorrelation of the Residuals***

Next, it was time to check that there was no autocorrelation of residuals. To check this particular assumption, a plot with Fitted Values as the x-axis and the Residuals as the y-axis was constructed.



Comments on Figure 1.9:

The sixth assumption of the ten, which states that there should be no autocorrelation of residuals is indeed satisfied as it can be seen from the plot to the left, there is no real or perceivable pattern or trend in the residuals which points towards the perspective that they are independent of each other. This plot supports the perspective that there exists little to no autocorrelation in the data

Figure 1.9

In addition to the Fitted Values vs Residuals plot above, the code output mentioned under *Assumption 2* also helps satisfy that there is autocorrelation of the residuals as the output value of the Durbin-Waston test was only slightly under 2, indicating a lack of autocorrelation, as previously mentioned.

7. ***Assumption #7: Checking Outliers & Leverages***

The following step was to check whether or not outliers were present throughout the dataset. This check was pursued by drawing box plots of the variables. In the case of numeric variables, there was nothing plotted against the y-axis, and in the case of categorical variables, they were plotted against the response variable. This would lead to investigating the spread of the data becoming a more straightforward process as the quantiles and minimum and maximum values allow one to highlight that if a point does not fall within a certain range, it is most likely an outlier.

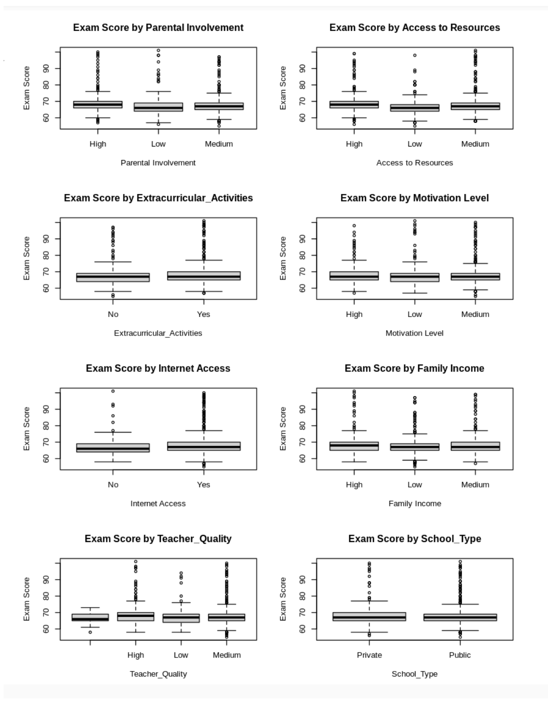
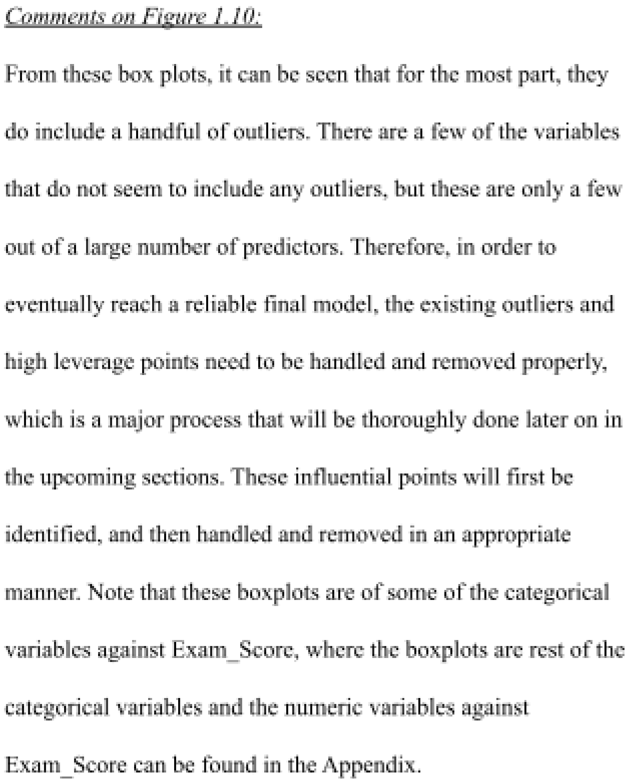


Figure 1.10



Comments on Figure 1.10:

From these box plots, it can be seen that for the most part, they do include a handful of outliers. There are a few of the variables that do not seem to include any outliers, but these are only a few out of a large number of predictors. Therefore, in order to eventually reach a reliable final model, the existing outliers and high leverage points need to be handled and removed properly, which is a major process that will be thoroughly done later on in the upcoming sections. These influential points will first be identified, and then handled and removed in an appropriate manner. Note that these boxplots are of some of the categorical variables against Exam_Score, where the boxplots are rest of the categorical variables and the numeric variables against Exam_Score can be found in the Appendix.

8. ***Assumption #8: Alternative Check For Homoscedasticity***

Upon checking if outliers exist within the dataset, it was time to conduct an alternative check for homoscedasticity. This alternative check is based on drawing scatter plots of the numeric variables against the response variable and seeing whether or not the residuals were randomly scattered and whether or not the spread of the residuals was constant throughout the range of fitted values.

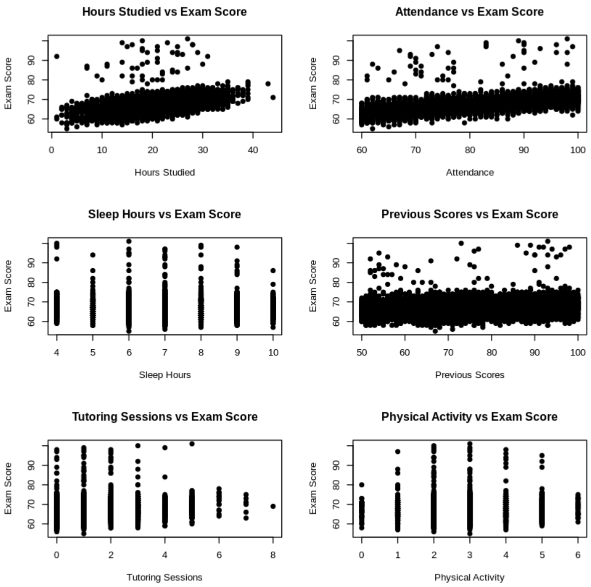
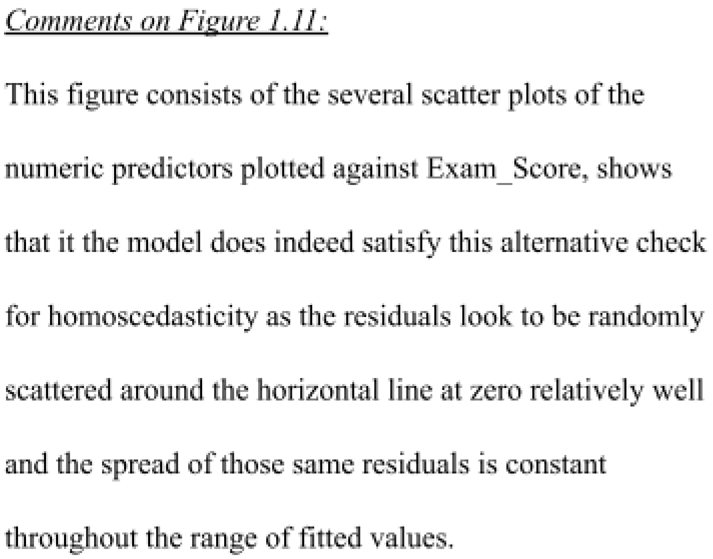


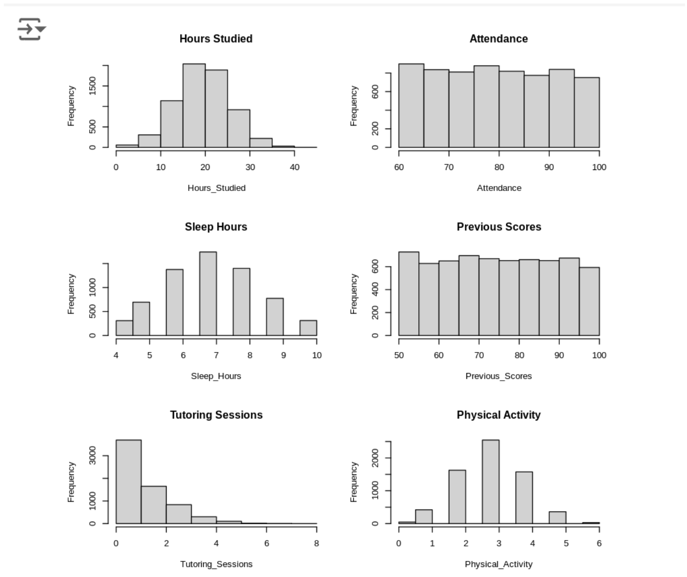
Figure 1.11

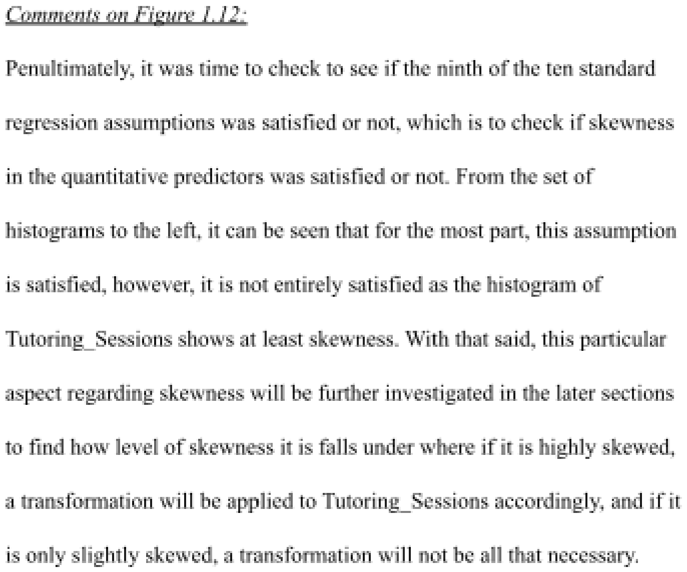


Comments on Figure 1.11:

This figure consists of the several scatter plots of the numeric predictors plotted against Exam_Score, shows that it the model does indeed satisfy this alternative check for homoscedasticity as the residuals look to be randomly scattered around the horizontal line at zero relatively well and the spread of those same residuals is constant throughout the range of fitted values.

9. ***Assumption #9: Checking For Skewness in the Predictors***





Comments on Figure 1.12:

Penultimately, it was time to check to see if the ninth of the ten standard regression assumptions was satisfied or not, which is to check if skewness in the quantitative predictors was satisfied or not. From the set of histograms to the left, it can be seen that for the most part, this assumption is satisfied, however, it is not entirely satisfied as the histogram of Tutoring_Sessions shows at least skewness. With that said, this particular aspect regarding skewness will be further investigated in the later sections to find how level of skewness it is falls under where if it is highly skewed, a transformation will be applied to Tutoring_Sessions accordingly, and if it is only slightly skewed, a transformation will not be all that necessary.

Figure 1.12

10. *Assumption #10: Checking the Model Fit*

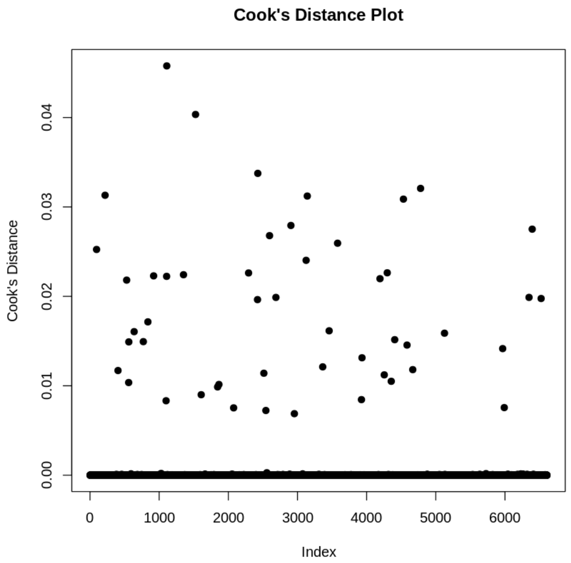


Figure 1.13

Comments on Figure 1.13:

Finally, the tenth assumption was to check the model fit which was done using Cook's Distance plot as seen to the left. Interpreting this figure, it can be seen that the majority of the points have quite low Cook's Distances values which means that they have little-to-no influence on the fitted values of the model. In addition to displaying low Cook's Distances values, there does not seem to be any existing outliers that are the few exceptions that turned out to have high Cook's Distances values compared to the rest. Putting these two interpretable traits together, it can be stated that the tenth standard assumption regression is also satisfied indicating that the model is a reasonable fit for the data.

Initial Model Fit to the Data

The initial model fit to the data was a standard linear regression model with all of the predictors. The quantitative predictors were included as is whereas the categorical variables were set as factors using the built-in function, "`as.factor`," which "is used to convert a vector object to a factor" (Educative). This is a crucial step before conducting the initial regression model; only then can the model assign numerical values to the categorical predictors, estimate the effect of each of their categories on the dependent variable, and point out any major differences between the categorical variables and their categories.

Below are the important findings from this initial model fit to the data:

1. `Residual standard error: 2.036 on 6576 degrees of freedom`

    a. This reveals that, on average, the model's predictors of a student's score on their final exam, or Exam_Score, is off by approximately 2.036 points or percent.

2. `Multiple R-squared:  0.7275, Adjusted R-squared:  0.7262`

    a. This reveals that the proportion of variability in the response variable, Exam_Score, that is explained by the predictors is approximately 0.72, or 72%.

3. `F-statistic: 585.1 on 30 and 6576 DF,  p-value: < 2.2e-16`

    a. This indicates that the overall model is statistically significant where at least one of the predictors is significantly related to the dependent variable, Exam_Score.

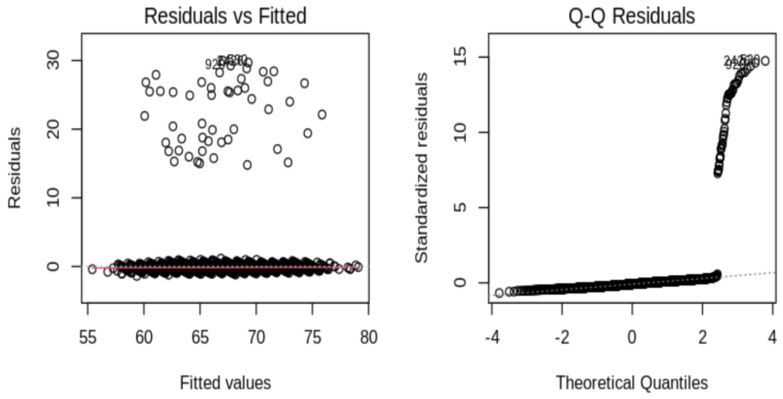


Figure 1.14

Comments on Figure 1.14:

The figure to left shows that for this first set of plots, the data well and truly includes outliers and highly influential observations. This is due to how their impact can be perceived as the residuals go all the way up to 30, highlighting that these outliers took over the scaling and axes of the plots. This leaves that there is room for more transformations to take place to adjust such a circumstance.

More Transformation of Variables If Skewness Values Deem It Necessary:

The next stage of the data analysis was to check if there was any need to transform the numeric variables, and if so, to do so. To conduct this specific check, code was written and run to see if any quantitative predictors had a skewness value greater than 1 or less than -1. This is because skewness values that fall in the interval [-1, -0.5] are considered "negatively skewed" and those falling in the interval [0.5, 1] are considered "positively skewed." Building on this, variables that have skewness values that are either less than -1 or greater than 1 are referred to as "highly skewed," and those are the ones that require transformation (Suvarna). Hence, why the code was run to check if any variables had skewness values satisfying these criteria and were candidates for transformation. The following is the output of the code which investigated the skewness of the quantitative variables:

```
Hours_Studied        Attendance       Sleep_Hours   Previous_Scores
      0.013492780       0.013659655      -0.023794629      -0.003734837
Tutoring_Sessions Physical_Activity
      0.815159338      -0.031350472
[1] "Quantitative Variables With Skewness > 1:  "
[1] "Quantitative Variables With Skewness < -1:  "
```

The above output shows that none of the predictors are “highly skewed,” since there are no skewness values less than -1 or greater than 1. This confirms that no transformation of the variables is needed. An additional piece of analysis is that this output shows that only one of the quantitative variables is “positively skewed,” which is Tutoring_Sessions, as it has a skewness value of approximately 0.815, which again, is a skewness value that does not show the necessity of doing any transformations.

Identification of Outliers, Leverage Points, & Influential Observations

Starting with the outliers, the procedure that was taken was to identify the observations that had “a standardized residual that is larger than 3 (in absolute value)” (PennState). To be concise, the following is a shortened list of the observations that were identified as outliers:

```
[1] "Outliers (Those With |Standardized Residual| > 3):"
  95  218  405  530  559  561  638  771 . . . 6523
```

Moving on to the high-leverage points, these observations have “‘extreme’ predictor x values,” which are simply put, outliers with respect to the predictors. (PennState) To be concise, The following is a shortened list of the observations that were identified as high-leverage points:

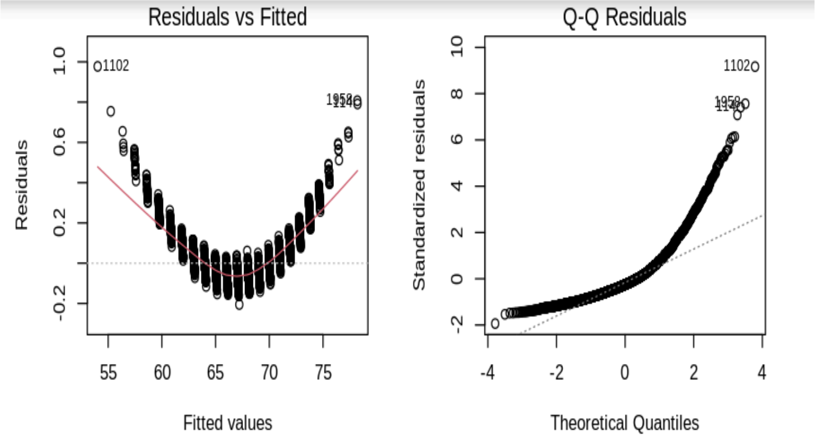
```
[1] "High Leverage Points: "
  34  128  241  276  317  360  381  397 . . . 6597
```

As for the influential observations, these were identified through the use of Cook’s Distance as the general rule of thumb states “that D(i) > 4/n is a good threshold for determining highly influential points" (Yellowbrick). To be concise, the following is a shortened list of the observations that were identified as influential observations:

```
[1] "Influential Points (Those With A Cook's Distance > 4/n)"
  95  218  405  530  559  561  638  77 . . . 6523
```

Graphs After Fitting a Model to the Data

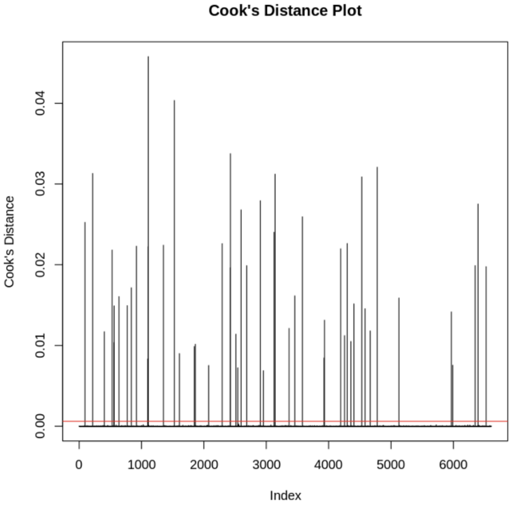
After identifying the outliers, high leverage points, and influential observations in the previous section, it was time to handle and remove them and then refit the model to have updated versions of those in the *Initial Model Fit to the Data* section. One major reason for this process is that these outliers and influential observations do not represent the population. This is because their inclusion in the initial models causes disproportionate changes to the model parameters to the point where these initial models are quite unreliable and do not accurately represent the population. Therefore, removing them and then refitting the model would lead to more representative and reliable visualizations. Another critical reason for doing so is that such influential observations can also invalidate the model by violating some of the standard regression assumptions, such as linearity, homoscedasticity, and normality of residuals, just to name a few out of the ten. So, handling and removing them appropriately would allow the data to be more satisfactory of these assumptions and eventually lead to reliable findings and conclusions. Below are the visualizations after the model was fit to the data and the influential observations were dealt with appropriately.



Comments on Figure 1.15:

From this figure, it can be seen that the residuals are all close to zero and are also randomly distributed, and their variance looks to be constant across the range of Fitted Values. Moving to the Q-Q plot, it also hints at the residuals being normally distributed. These aspects put together indicate the new and updated model is a good fit for the data where reasonable predictions and conclusions can be made. These two plots represent much more stable and better-fits compared to their initial counterparts where there was a clear presence of impactful outliers and highly influential points.

Figure 1.15



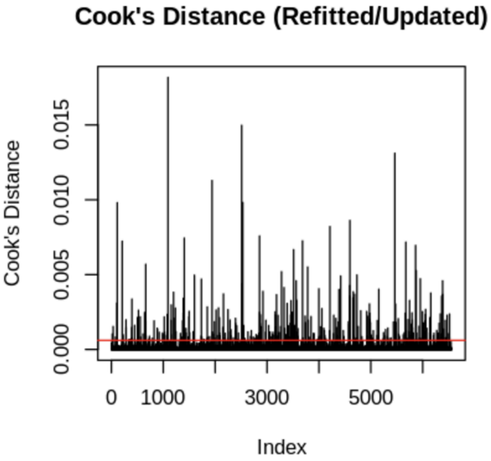


Figure 1.16

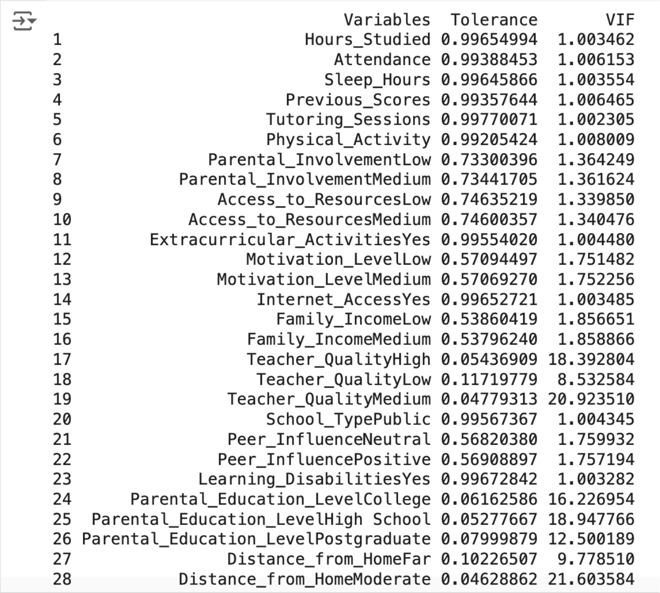
Figure 1.17

Finally, the two figures above are two Cook's Distances plots respectively displaying the before and after conducting the necessary transformations and properly handling the outliers. To be more specific, Figure 1.16 is of the initially fitted model, and Figure 1.17 is the updated model and goes along with grouped plots in Figure 1.15. Analyzing these two plots, Figure 1.16 shows multiple plots going beyond the threshold line signaling the presence of outliers and influential observations. Whereas in Figure 1.17, the points are all much closer to zero highlighting the lesser impact of the influential observations on the model itself now that they have been removed. Accordingly, the figure on the right shows an improved version where the model has become more stable demonstrating the importance of the processes implemented to go from the figure on the left to the one on the right, namely the removal of the outliers and the transformation of certain variables.

To reiterate, with these newly refitted models now that certain transformations and handling the outliers have taken place, these updated plots of the Residuals vs Fitted, Q-Q Residuals, and Cook's Distance will be better servants when it comes to providing a better representation and reflection of the remaining observations, and ultimately of the population, as there are no unusual observations negatively impacting the data and altering the plots in ways which prevent on from reaching reliable conclusions and insights.

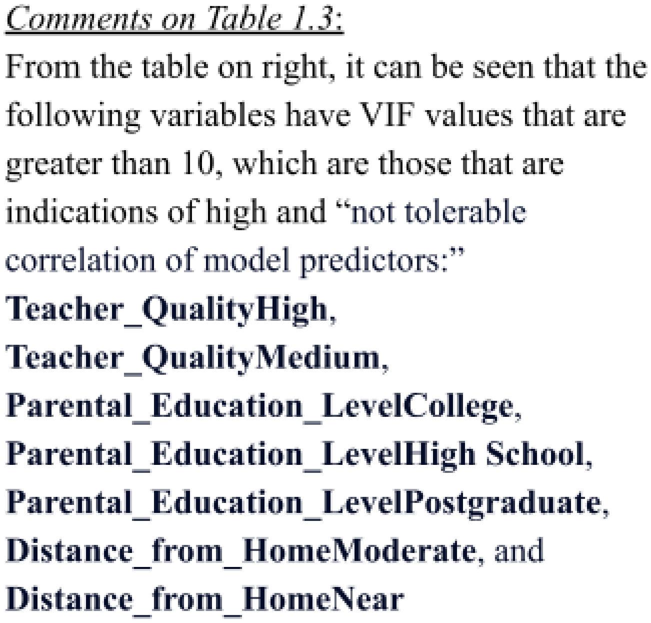
Model Criticism & Reformulation

The final part of the data analysis was to criticize and reformulate the model. This began with calculating the Tolerance and VIF values of all predictors. If a variable has a VIF value greater than 10, this suggests high collinearity making it obligatory to remove it. The following table shows the output to the line of code: `ols_vif_tol(initial_reg_model)`



```
                                 Variables  Tolerance       VIF
1                            Hours_Studied 0.99654994  1.003462
2                               Attendance 0.99388453  1.006153
3                              Sleep_Hours 0.99645866  1.003554
4                          Previous_Scores 0.99357644  1.006465
5                        Tutoring_Sessions 0.99770071  1.002305
6                        Physical_Activity 0.99205424  1.008009
7                  Parental_InvolvementLow 0.73300396  1.364249
8               Parental_InvolvementMedium 0.73441705  1.361624
9                   Access_to_ResourcesLow 0.74635219  1.339850
10               Access_to_ResourcesMedium 0.74600357  1.340476
11           Extracurricular_ActivitiesYes 0.99554020  1.004480
12                     Motivation_LevelLow 0.57094497  1.751482
13                  Motivation_LevelMedium 0.57069270  1.752256
14                      Internet_AccessYes 0.99652721  1.003485
15                        Family_IncomeLow 0.53860419  1.856651
16                     Family_IncomeMedium 0.53796240  1.858866
17                     Teacher_QualityHigh 0.05436909 18.392804
18                      Teacher_QualityLow 0.11719779  8.532584
19                   Teacher_QualityMedium 0.04779313 20.923510
20                       School_TypePublic 0.99567367  1.004345
21                   Peer_InfluenceNeutral 0.56820380  1.759932
22                  Peer_InfluencePositive 0.56908897  1.757194
23                Learning_DisabilitiesYes 0.99672842  1.003282
24         Parental_Education_LevelCollege 0.06162586 16.226954
25     Parental_Education_LevelHigh School 0.05277667 18.947766
26    Parental_Education_LevelPostgraduate 0.07999879 12.500189
27                   Distance_from_HomeFar 0.10226507  9.778510
28              Distance_from_HomeModerate 0.04628862 21.603584
```

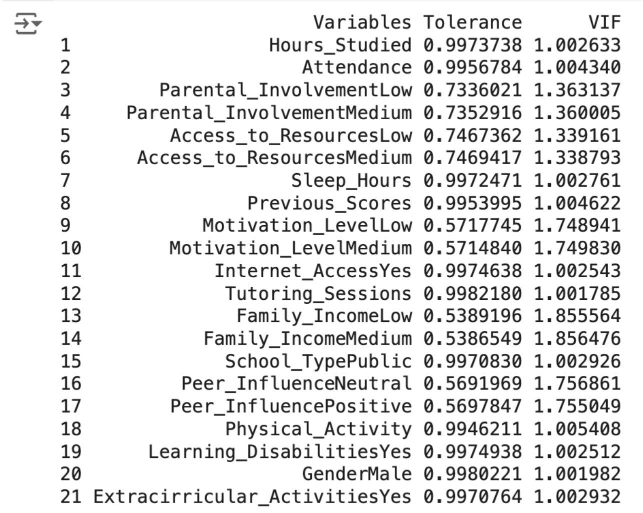
Table 1.3



Comments on Table 1.3:

From the table on right, it can be seen that the following variables have VIF values that are greater than 10, which are those that are indications of high and "not tolerable correlation of model predictors:" **Teacher_QualityHigh**, **Teacher_QualityMedium**, **Parental_Education_LevelCollege**, **Parental_Education_LevelHigh School**, **Parental_Education_LevelPostgraduate**, **Distance_from_HomeModerate**, and **Distance_from_HomeNear**

Therefore, removing this list of variables will allow for a lower residual standard error value and a higher Adjusted R-squared value, representing a better fit and much more interpretable model.

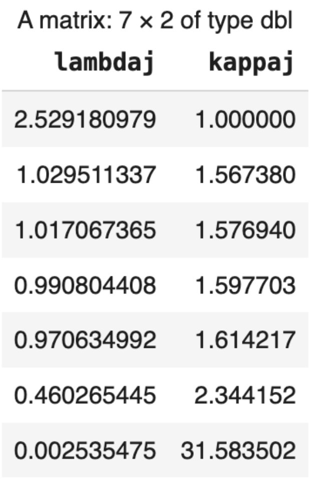


```
                       Variables Tolerance      VIF
1                  Hours_Studied 0.9973738 1.002633
2                     Attendance 0.9956784 1.004340
3         Parental_InvolvementLow 0.7336021 1.363137
4      Parental_InvolvementMedium 0.7352916 1.360005
5           Access_to_ResourcesLow 0.7467362 1.339161
6        Access_to_ResourcesMedium 0.7469417 1.338793
7                    Sleep_Hours 0.9972471 1.002761
8                Previous_Scores 0.9953995 1.004622
9             Motivation_LevelLow 0.5717745 1.748941
10         Motivation_LevelMedium 0.5714840 1.749830
11             Internet_AccessYes 0.9974638 1.002543
12               Tutoring_Sessions 0.9982180 1.001785
13                Family_IncomeLow 0.5389196 1.855564
14             Family_IncomeMedium 0.5386549 1.856476
15               School_TypePublic 0.9970830 1.002926
16           Peer_InfluenceNeutral 0.5691969 1.756861
17          Peer_InfluencePositive 0.5697847 1.755049
18               Physical_Activity 0.9946211 1.005408
19        Learning_DisabilitiesYes 0.9974938 1.002512
20                      GenderMale 0.9980221 1.001982
21 Extracirricular_ActivitiesYes 0.9970764 1.002932
```

Table 1.4

Comments on Table 1.4:

This table represents the now just-updated table of Tolerance and VIF values after the removal of those that had VIF values greater than 10 in the initial table of Tolerance and VIF (*Table 1.3)*. Now the Model Criticism & Formulation can be continued as the variables that were indicative of high collinearity have been removed.



A matrix: 7 × 2 of type dbl

lambdaj	kappaj
2.529180979	1.000000
1.029511337	1.567380
1.017067365	1.576940
0.990804408	1.597703
0.970634992	1.614217
0.460265445	2.344152
0.002535475	31.583502

Table 1.5

Comments on Table 1.5:

This table shows that has "lambdaj" as the left column being the eigenvalues of the principal components and "kappaj" as the right column being the condition indices, which are the telling aspect of whether or not there are potential multicollinearity issues that need to be dealt with. If there is a kappaj value greater than 30, it is a sign that a multicollinearity issue needs to be handled which is the case here as the final row has a kappaj value of about 31.58. This was dealt with using PCA methods which are effective for handling such multicollinearity issue as they are based on conducting a transformation on the original correlated predictors where they now become uncorrelated PC's so that the model is now much more stable and explicable.

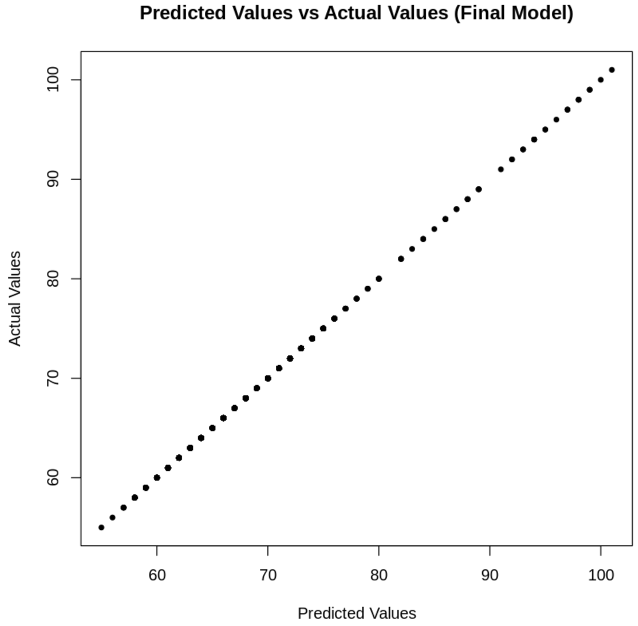
Following this, code that was written to further reformulate the model where the main takeaway of the output summary of the latest regression model was the following:

```
Residual standard error: 1.149e-13 on 6599 degrees of freedom
Multiple R-squared:      1,      Adjusted R-squared:      1
F-statistic: 1.082e+30 on 7 and 6599 DF,  p-value: < 2.2e-16
```

This output reveals that the model now has an R-squared value of 1 indicating a perfect fit. This is very unusual and unlikely to be the case with the vast majority of real-world data leading to the intuition that when pursuing the fitting procedures, they could have been overdone, leading to such an outcome. Another potential reason for such a perfect fit is that the data had included existing deterministic relationships leaving it with little to no room for error or randomness (Intersect Technologies). In addition to having a perfect fit highlighted by the R-squared value, the value of the residual standard error being very close to zero highlights that there could be hardly any unexplained variability, which further implies the chances that the fitting was done to an excessive extent.

Summary and Conclusions

Final Model



Comment on Figure 1.18:
The following plot shows the predicted values plotted against the actual values, which are the actual values of the group of students' final exam scores, where this particular plot represents the final model that has been reached. It can be seen that it reflects the perfect fit and correlations mentioned in the previous section where the R-squared value was equal to 1. To put it quite simply, this final model that has been reached shows that the predicted values turned out to perfectly match the actual results.

Figure 1.18

Further investigating this plot will lead to observing that there are two masked points. Although they are easily distinguishable, these two particular points seem to be the case where while the predicted values did not perfectly match with the actual results, they were very close, hence why they line up very closely with the diagonal. Nonetheless, why these two masked points are observable, it is clear that their impact is not significant enough to prevent the final model from representing and displaying a perfect fit, as seen right above.

To re-emphasize, this perfect model could have been reached due to overfitting. While it is desirable to have a perfectly fit model, it is important to note that it is quite unrealistic and unlikely to occur in most cases. Having a perfect fit is an extremely rare occurrence where it could be seen as an alert that while fitting the model, it may capture some deterministic relationships causing such a final model to be reached.

Conclusions

In conclusion, the final model that was reached and described in the previous section achieved a perfect R-squared value as well as a very minimal residual standard error value, suggesting an ideal fit that the predictors explain all of the variability in the response variable, Exam_Score. With that said, it is important to note that when examining and interpreting a perfect linear model like the one displayed above, there could be a case of overfitting or other potential data-related issues that were at least partly responsible for reaching such a perfect final model.

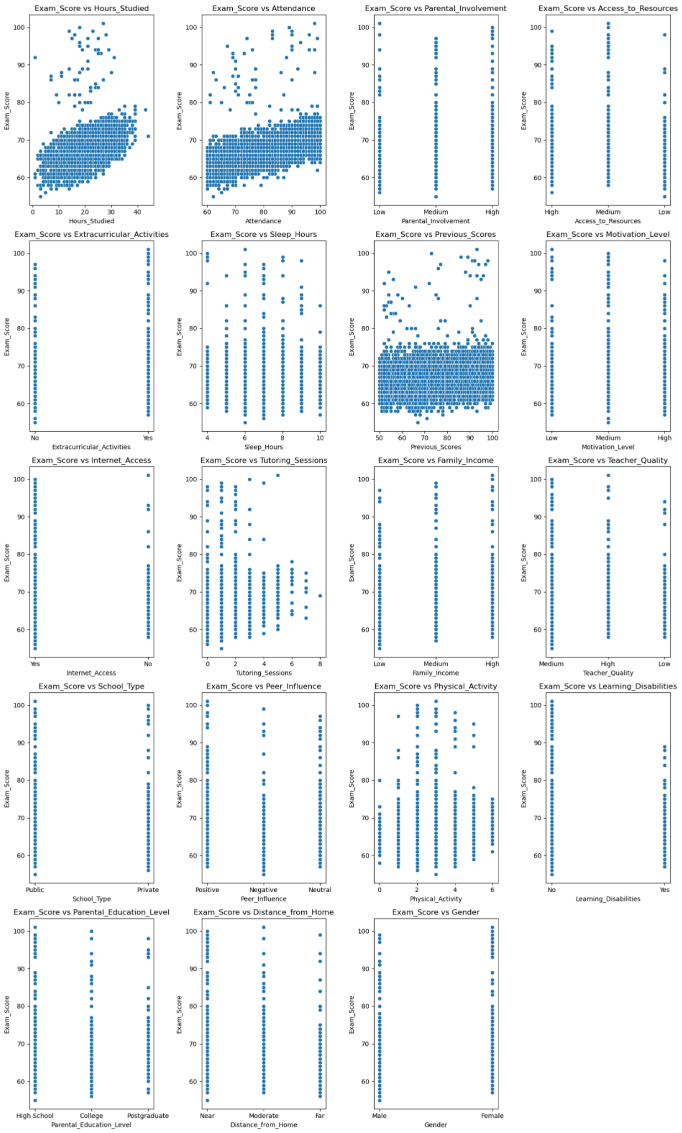
Overall, the results of this thoroughly conducted data analysis provide one with valuable insights into the long list of various factors that potentially play into a student's academic performance. The predictors that were untransformed and remaining by the final model can be heavily utilized to formulate the ideal preparation and scenario as well as certain academic restrictions to allow a student to thrive as much as possible in their academics. With all of that said, it is essential to keep in mind that all the models, from the initial to the final one, are based on a particular collection of data and observations where it may not always be possible and that correlation does not equal causation.

Appendix

List of the data set

Dataset URL: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors

Annotated Computer Outputs (including numbered Tables and Figures)



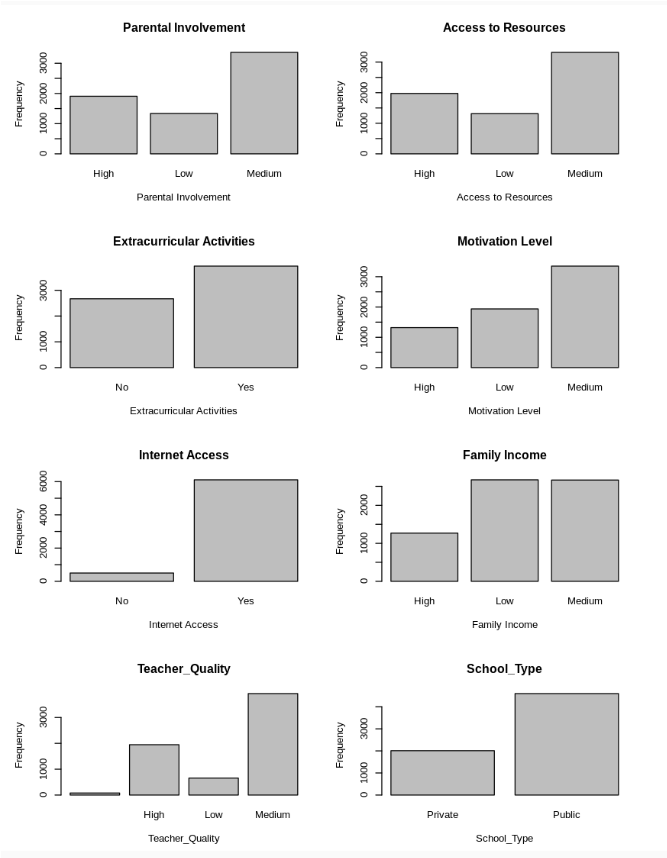


Figure 1.19 **Figure 1.20**

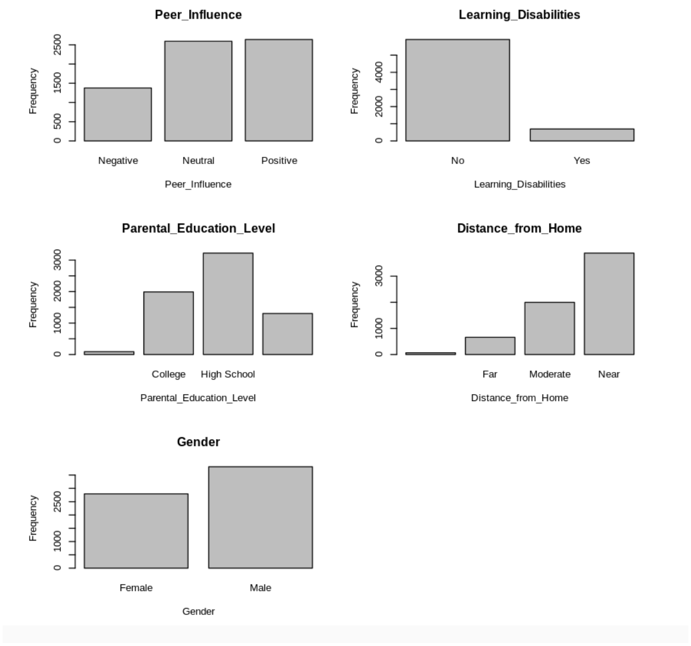


Figure 1.21

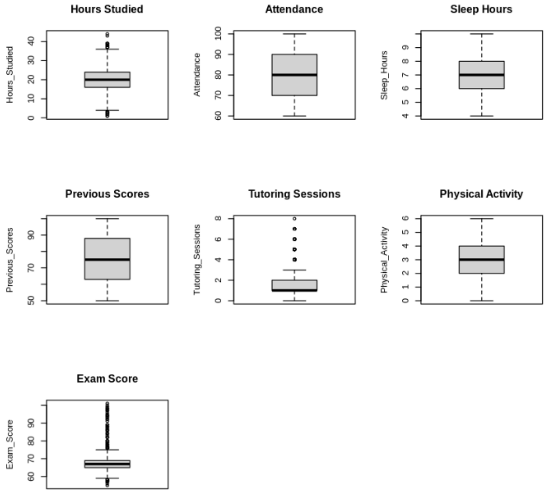


Figure 1.22

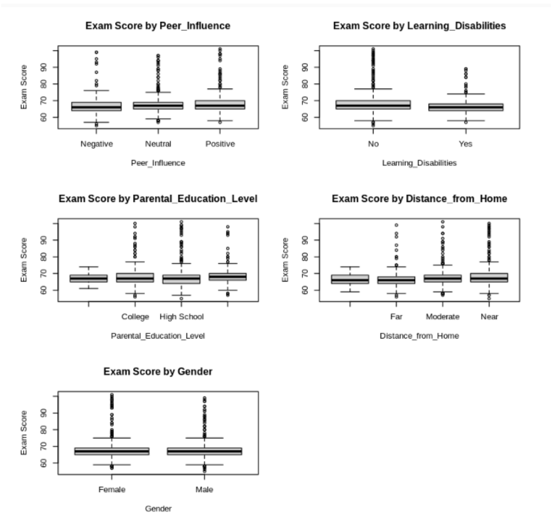


Figure 1.23

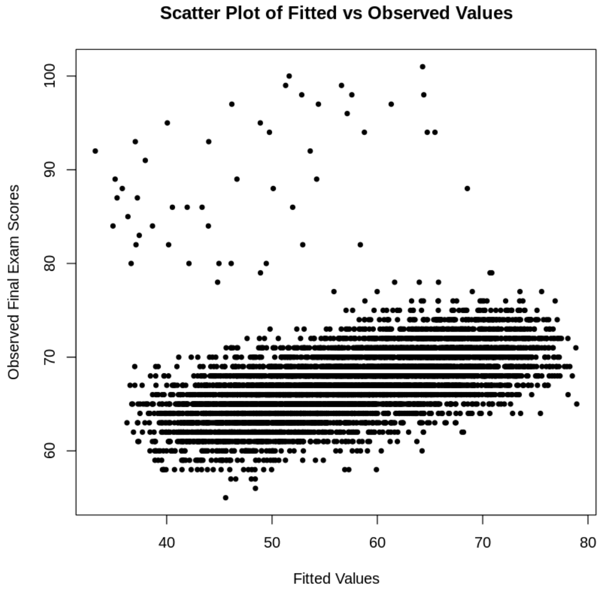


Figure 1.24

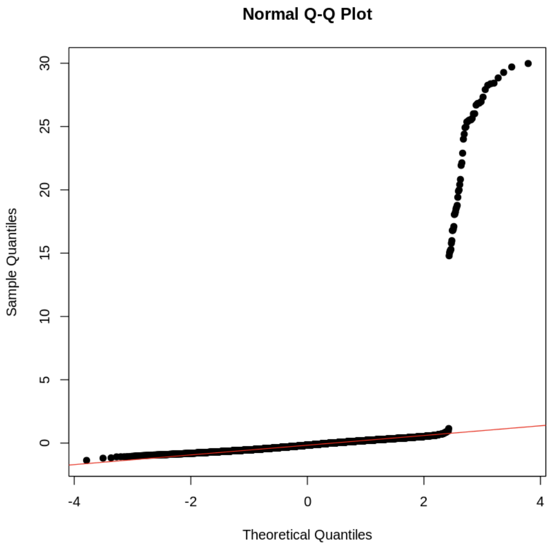

Normal Q-Q Plot
Sample Quantiles
30
25
20
15
10
5
0
-4
-2
0
2
4
Theoretical Quantiles

Figure 1.25

References

"9.1 - Distinction Between Outliers and High Leverage Observations." *PennState Eberly College of Science.* https://online.stat.psu.edu/stat462/node/170/#:~:text=A%20data%20point%20has%20high,is%20particularly%20high%20or%20low. Accessed December 8, 2024.

"9.3 - Identifying Outliers (Unusual Y Values)." *PennState Eberly College of Science.* https://online.stat.psu.edu/stat462/node/172/#:~:text=The%20good%20thing%20about%20standardized,some%20to%20be%20an%20outlier. Accessed December 7, 2024.

CFI Team. "Durbin Watson Statistic." *CFI.* https://corporatefinanceinstitute.com/resources/data-science/durbin-watson-statistic/#:~:text=Interpreting%20the%20Durban%20Watson%20Statistic,-The%20Durban%20Watson&text=0%20and%204.-,A%20value%20of%20DW%20%3D%202%20indicates%20that%20there%20is%20no,indicates%20a%20negative%20serial%20correlation. Accessed December 6, 2024.

"Cook's Distance." *Yellowbrick.* https://www.scikit-yb.org/en/latest/api/regressor/influence.html#:~:text=Because%20of%20this%2C%20Cook's%20Distance,that%20is%20above%20that%20threshold. Accessed December 7, 2024.

Ludecke, Daniel. (et al.) "Check for multicollinearity of model terms." *R easystats.* https://easystats.github.io/performance/reference/check_collinearity.html#:~:text=Interpretation%20of%20the%20Variance%20Inflation%20Factor&text=A%20VIF%20less%20than%205,2013. Accessed December 6, 2024.

“Probabilistic vs. Deterministic Data: An Intersect Technologies White Paper.” *Intersect Technologies*. https://www.intersecttechnologies.com/probabilistic-vs-deterministic-data/#:~:text=Section%20Two%3A%20Deterministic%20Data&text=It%20assumes%20that%20events%20can,little%20or%20no%20inherent%20randomness. Accessed December 7, 2024.

Suvarna. “Difference Between Skewness and Kurtosis.” *Analytics Vidhya.* November 29, 2024. https://www.analyticsvidhya.com/blog/2021/05/shape-of-data-skewness-and-kurtosis/#:~:text=Skewness%20values%20within%20the%20range,skewed. Accessed December 6, 2024.

“What is the as.factor() function in R?” *Educative.* https://www.educative.io/answers/what-is-the-asfactor-function-in-r. Accessed December 6, 2024.