**5330 Project Report**

Anannya Chatterjee – UNT ID# ac1371

**Multiple Linear Regression Analysis to determine validity of relationship between High school graduation performance and its various deciding factors**

**Theory about the data:**

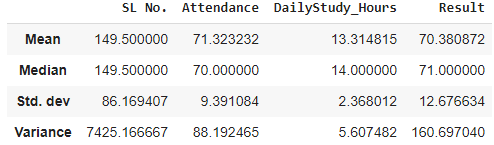
The technology revolution and digitization of 21st century has brought in new challenges for the teachers, parents, and mentors to understand and quantify multiple factors that influence a student’s conduct in various subjects which becomes a deciding factor for their grades/marks in the final year of graduation. Creating data models to follow students throughout their high school to unlock key issues that lead to success or failure of post-middle school education have the potential to not only help the national discussion regarding post-middle school success, but in this age of spiraling college tuition costs, may assist families in making more thoughtful decisions about how best to prepare their children for productive and successful educational experiences after high school graduation.

As part of this project, I have introspected the “[HighSchool\_GraduationPerformance.xlsx](https://unt.instructure.com/courses/38902/files/8963402/download?wrap=1) “dataset for determining the factors that possibly influence the performance of a student . There were all total 9 variables (Gender, Race, Parental\_Education, TestPrep\_Course, Special\_Coaching, Attendance, DailyStudy\_Hours and Result) and 298 observations. Result is considered as the dependent variable and all others are considered independent variables.

Looking at the data, I hypothesized that Parent higher education, and completion of Test preparation courses would have a positive impact on the student’s graduation performance.

I started with the process of **Data preparation and Cleansing.**

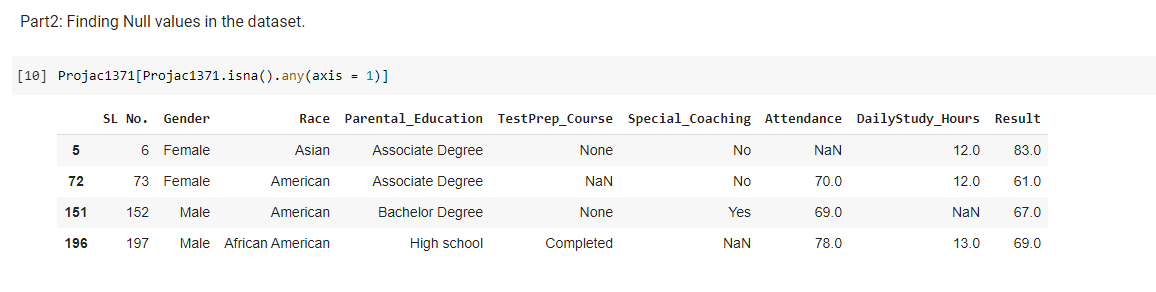
First, I have uploaded the "HighSchool\_GraduationPerformance" excel file to google colab environment. Also, I have introspected the data by running descriptive statistics on it. I found out the mean, median, standard deviation, variance of the data present in the numerical columns of Attendance, DailyStudy\_Hours and Result.



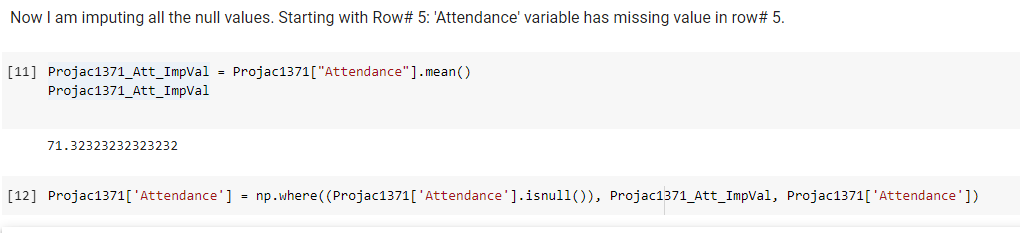
The mode is described as:

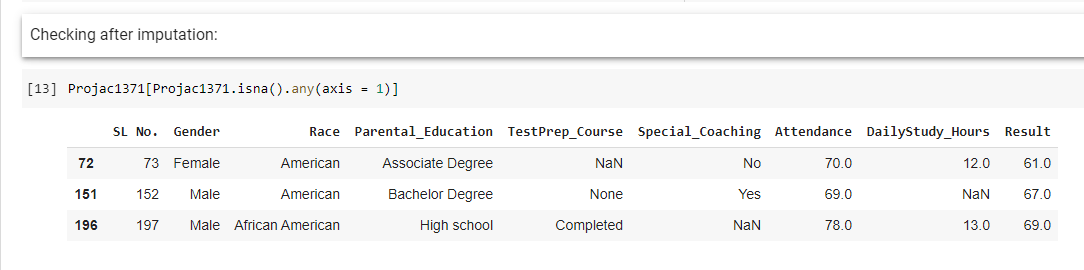


Upon further inspection I found that there were 4 rows in the dataset with missing/ NULL values in them (Row# 5, 72, 151 and 196).

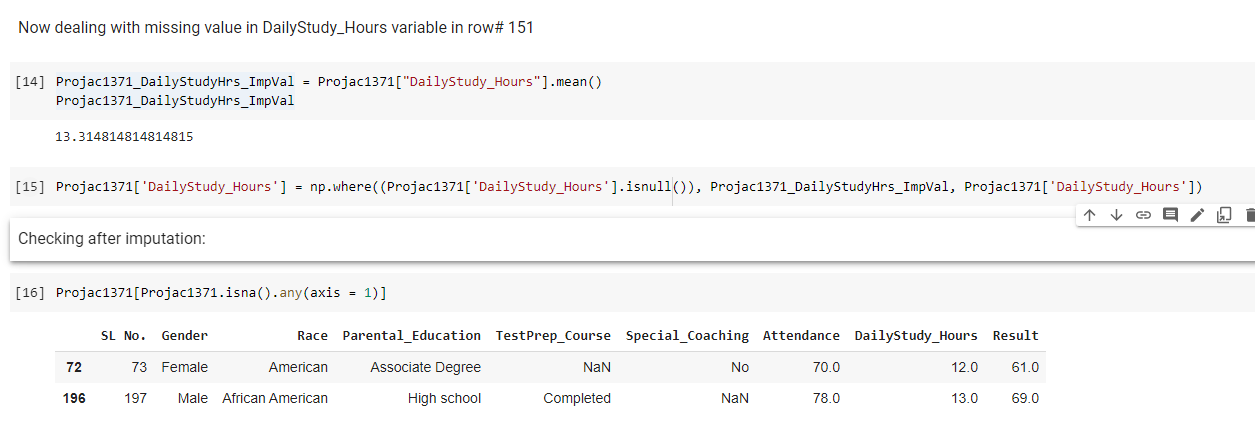


I imputed the missing values in Attendance variable (**row#5**) with the mean values of the Attendance column.

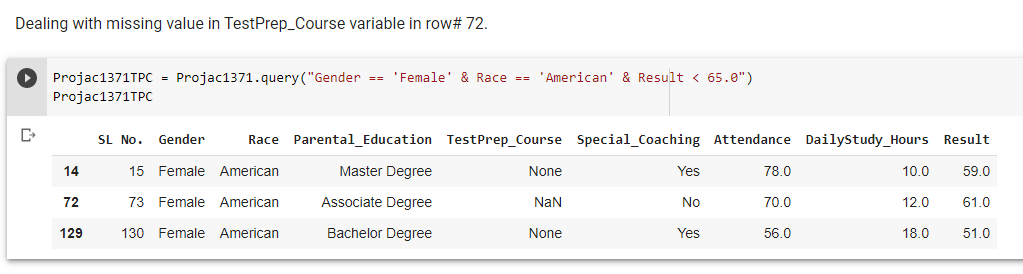


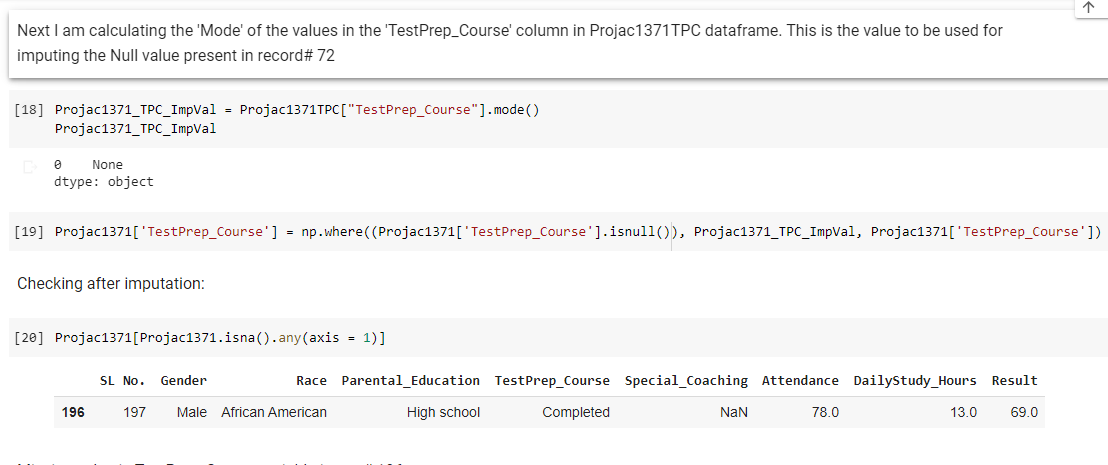


I imputed the missing values in DailyStudy\_Hours variable (**row#151**) with the mean value of the DailyStudy\_Hours column.

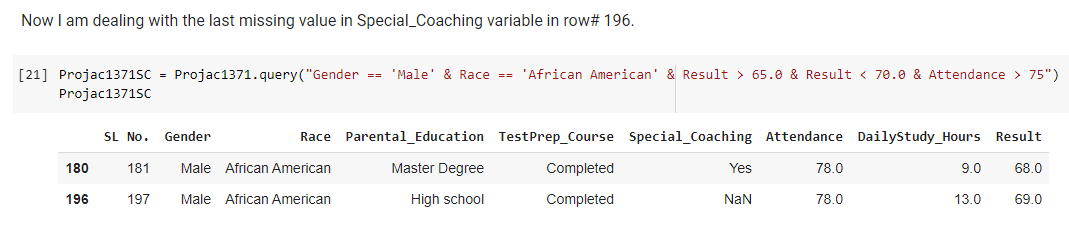


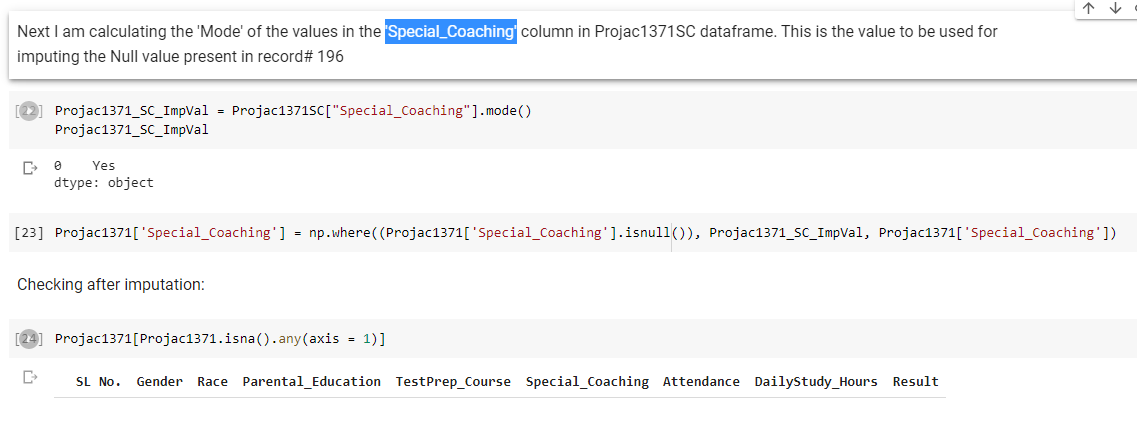
For **row#72**, TestPrep\_Course variable was having missing value. Hence, I queried for similar students who are Female, American and have secured less than 65% in Result. Based on the search result, I found the mode value of the TestPrep\_Course column. This mode value was used for imputation.





For **row#196**, Special\_Coaching variable was having missing value. Hence, I queried for similar students who are Male, African American and have secured Result percentage between 65 and 70 with an attendance percentage greater than 75. Based on the search result, I found the mode value of the Special\_Coaching column. This mode value was used for imputation.

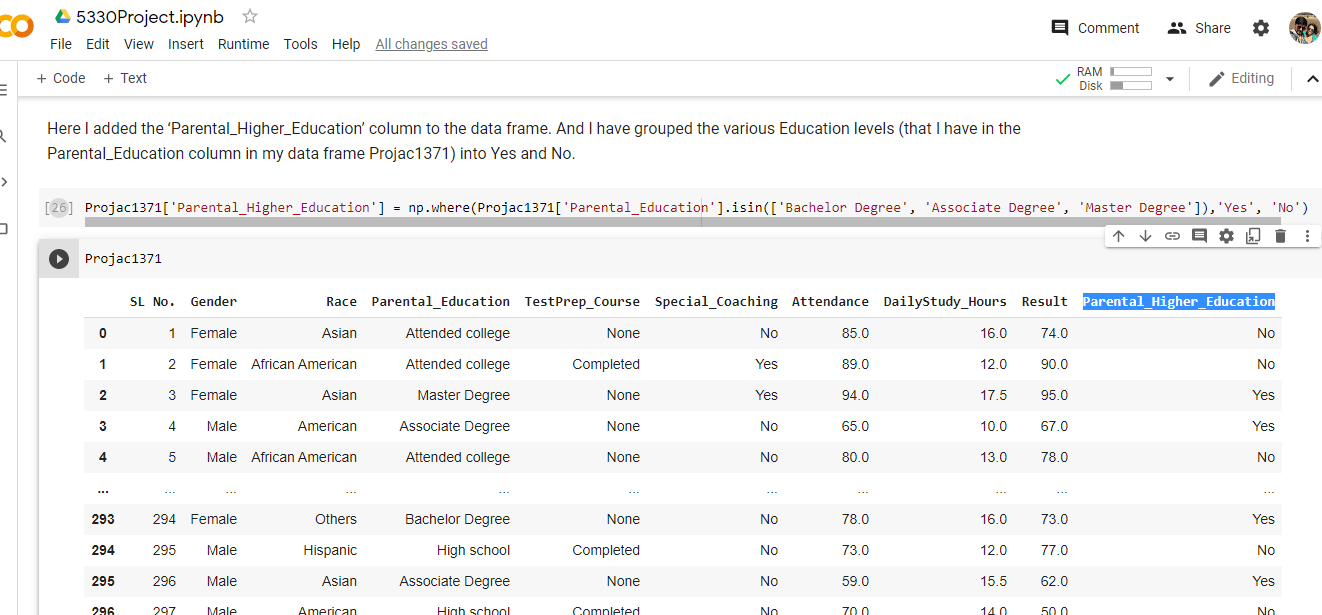


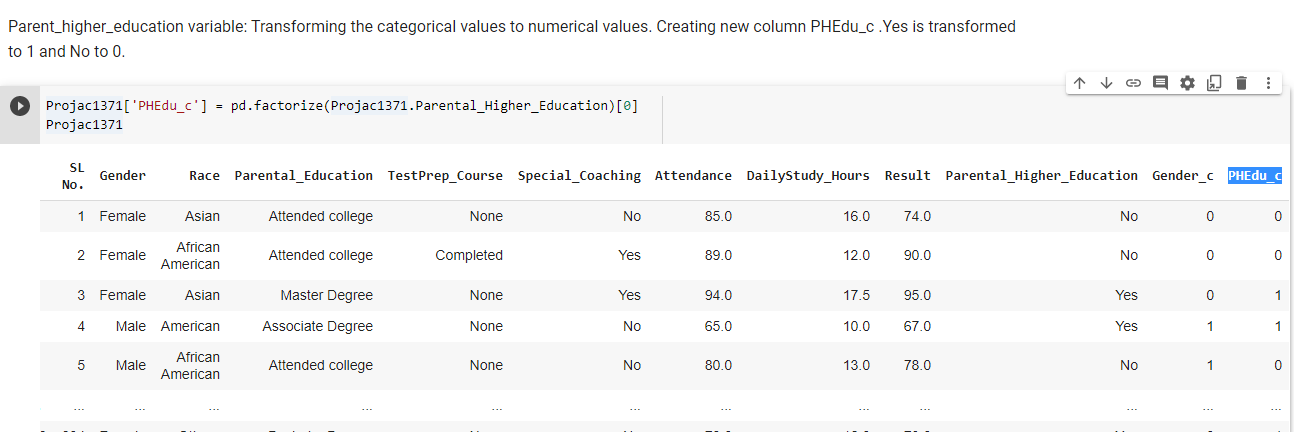


So, we see that all the missing values in the dataset were taken care off in this manner.

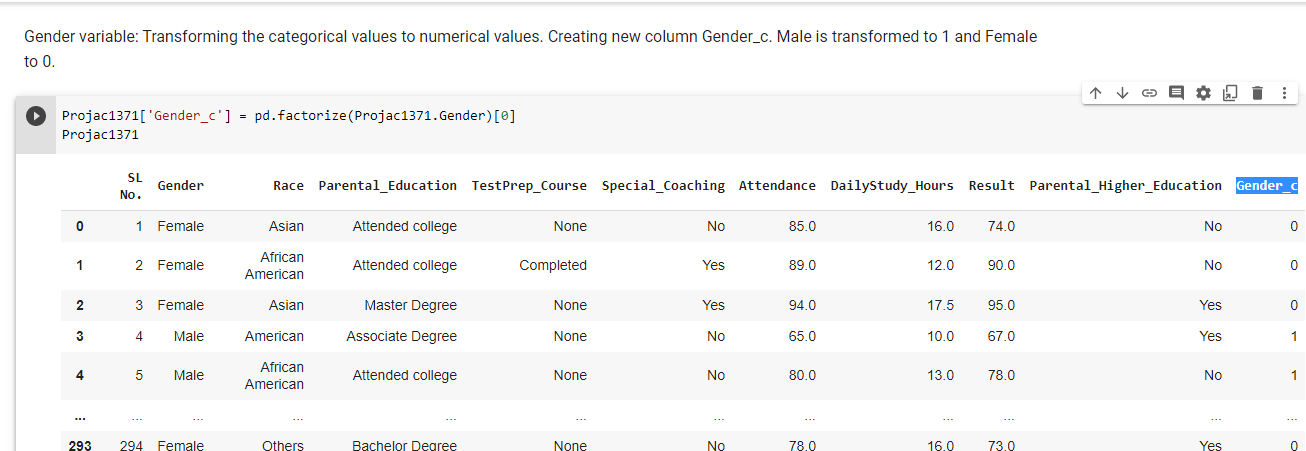
Lastly, I worked on transforming the various categorical variables. They are Gender, Race, Parental\_Education, TestPrep\_Course and Special\_Coaching.

Now I have added the ‘**Parental\_Higher\_Education**’ column to the data frame. I have grouped the various Education levels in the Parental\_Education column in my data frame (Projac1371) into Yes and No. The Parental Education levels that are either ‘Associate Degree’, ‘Bachelor Degree’ or ‘Master Degree’ are categorized as ‘Yes’. The Parental Education levels that are either ‘Attended college’ or ‘High school’ are categorized as ‘No’. **Next, I have transformed the categorical values for the ‘Parent\_Higher\_Education’ variable to numerical values. I created new column PHEdu\_c. Yes is transformed to 1 and No to 0.**

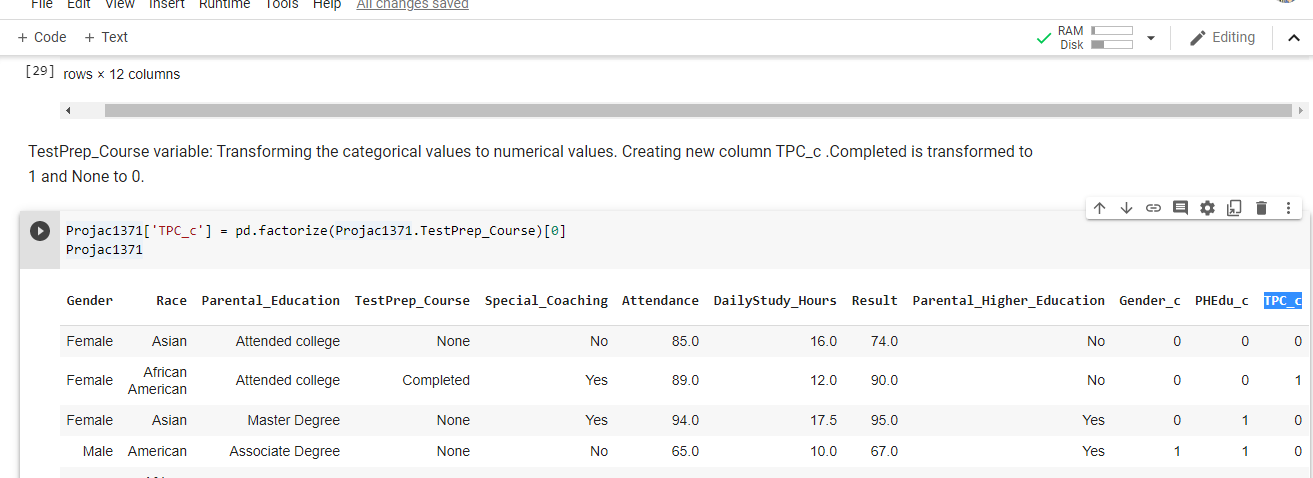




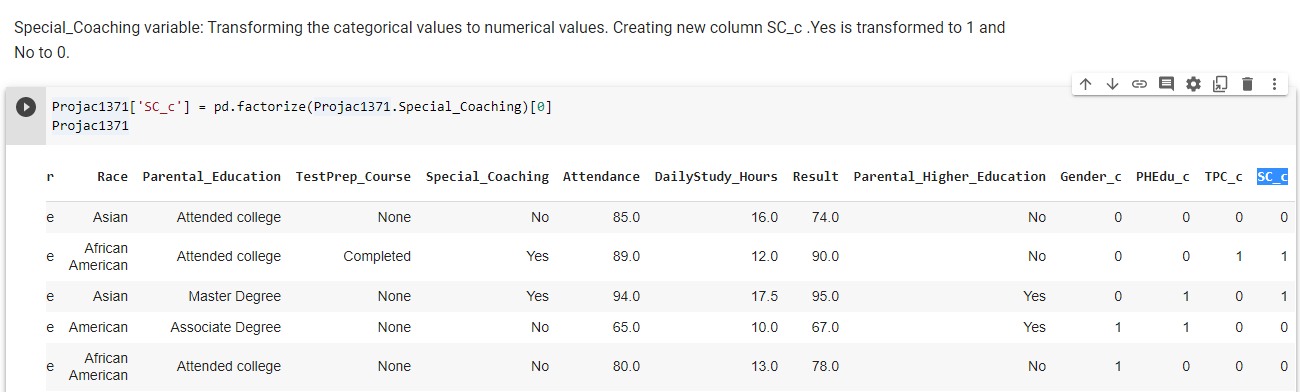
Next, I have worked on the Gender variable. I transformed the categorical values to numerical values. I created new column **Gender\_c**. Male is transformed to 1 and Female to 0.



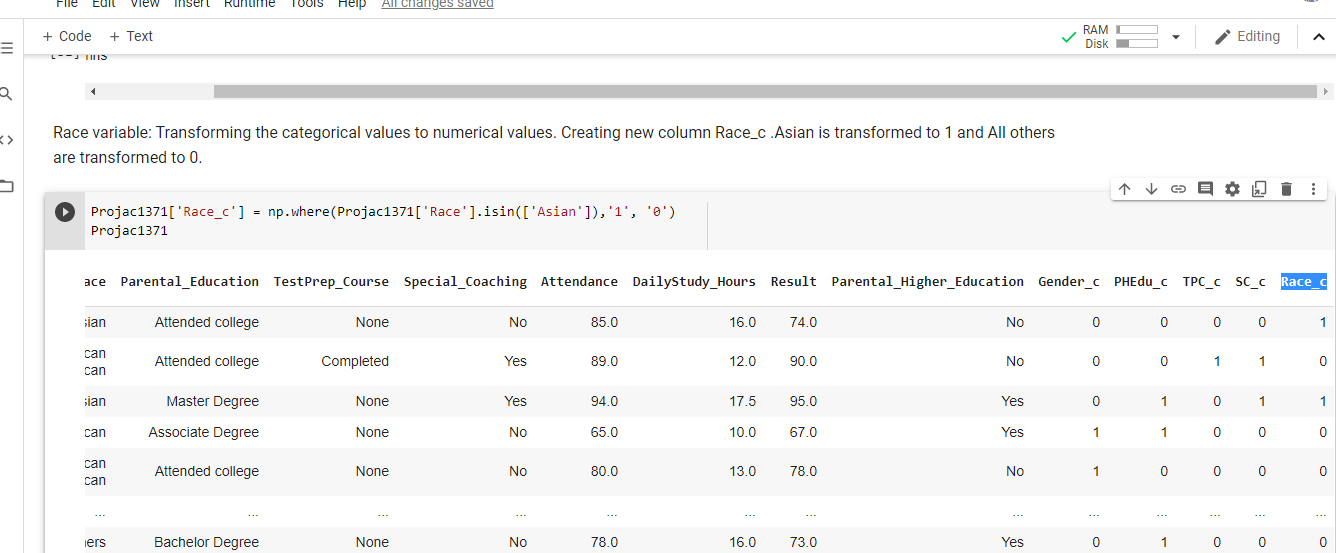
Then, I worked on the TestPrep\_Course variable and had transformed the categorical values to numerical values. I created new column **TPC\_c.** Completed is transformed to 1 and None to 0.



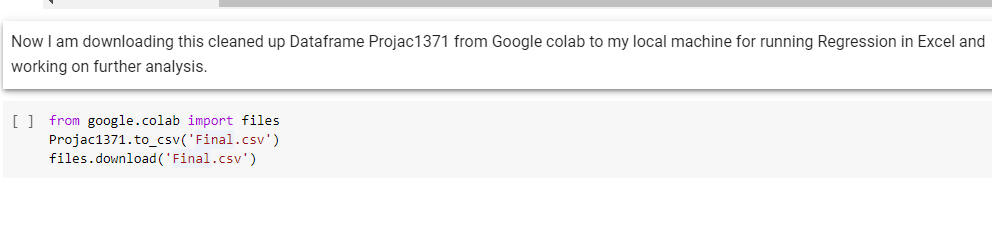
**After that, I worked on the Special\_Coaching variable and transformed the categorical values to numerical values. I created new column SC\_c. ‘Yes’ is transformed to 1 and No to 0.**



Considering **the Race variable. I transformed the categorical values to numerical values. Since I am analyzing the performance of the ‘Asian’ category of the Race variable baselining all the other Race categories, hence I have created new column Race\_c, and “Asian” is transformed to 1 and all other categories in Race variable are transformed to 0.**

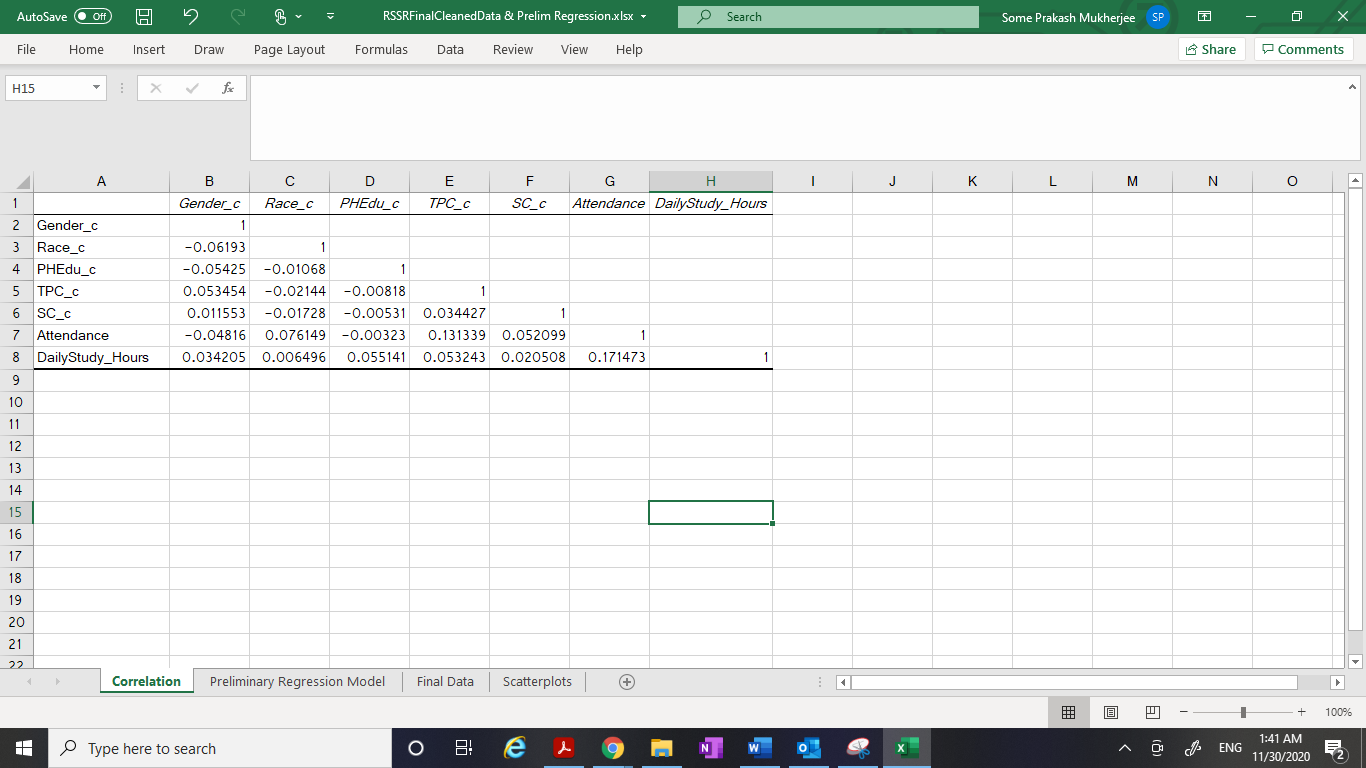


The data preparation and cleansing got concluded and then I downloaded this cleaned up Data frame ‘Projac1371’ from Google colab environment to the local machine for running Regression in Excel and working on further analysis.



**Next, I found the correlation between all the ‘X’ or independent variables.**

This is a pre regression testing that I did for checking on the possible multicollinearity. I completed the correlation test among all my independent variables which are: **Gender\_c, Race\_c, PHEdu\_c, TPC\_c, SC\_c, Attendance and DailyStudy\_Hours. It is done in** Excel as displayed in the below snapshot.

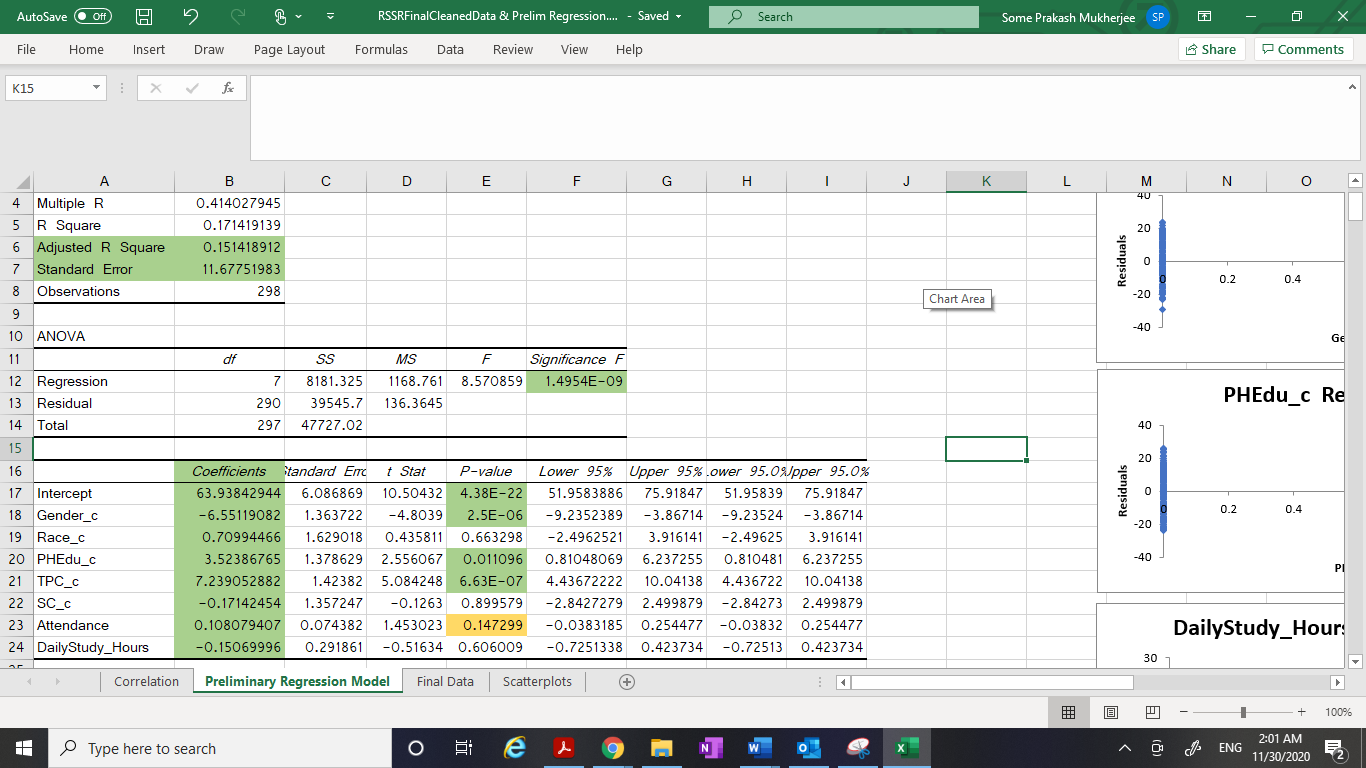


The above correlation matrix states that there **is no correlation between the independent variables.** As we know that a value of 0.8 or above indicates possible multicollinearity, so from the findings, it can be stated that Multicollinearity does not exist between the independent variables here.

**The process of Regression starts now. I have divided that into multiple parts for better explanation.**

**Part 1: Running a preliminary regression model considering all the variables.**

I have run a preliminary regression model considering all the variables as shown in the below screenshot.



I found that this preliminary regression is significant as the ‘p’ value for the overall F test is 1.4954E-09 which is much lower than 0.05.

**Part 2 and 3: Paying close attention to the statistical and practical significance levels of the variables and coming up with the reduced set of variables .**

After I paid close attention to the coefficients and the p-values (probability) for the respective independent variables and the intercepts in the summary output, I came up with the below points that needed to be considered.

1. I found that **Race\_c, SC\_c and DailyStudy\_Hours are statistically insignificant variables** because those variables are with p values much higher than 0.05 (significance level, Alpha = 0.05)
2. **Gender\_c, PHEdu\_c and TPC\_c are both practically and statistically significant variables** because those variables are with p values much lower than 0.05 and they are having higher beta weights.
3. **Intercept is statistically significant** because it is with p values much lower than 0.05, but do not have practical significance.
4. If I now look at **Attendance variable,** thep value here is not exceptionally higher than 0.05. So, **although it may seem** **statistically insignificant,** **but it can have practical significance.**

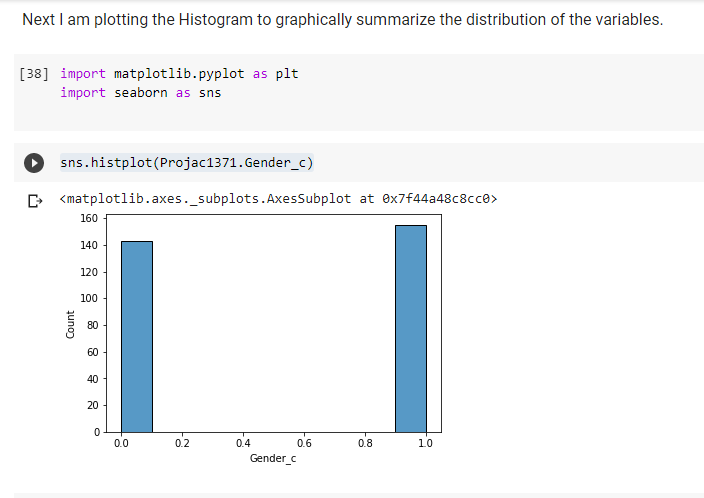
Hence based on the above information, I decided to **remove Race\_c, SC\_c and DailyStudy\_Hours** variables from taking into consideration for creating regression models. But I would still **not remove Attendance variable** as thep value here is not too much higher than 0.05.

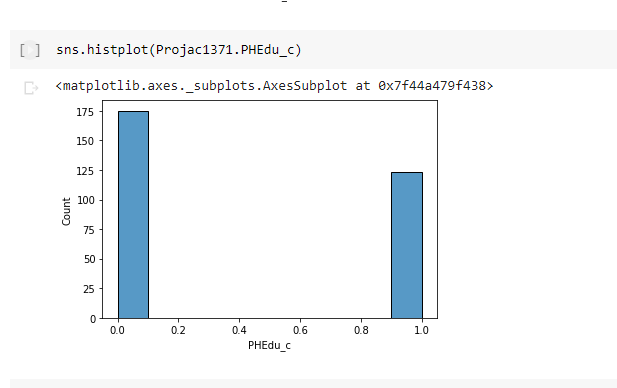
All these findings brought me to the reduced set of independent variables for working on. This reduced set contains: **Gender\_c, PHEdu\_c , TPC\_c and Attendance as the independent variables to work on along with Result as the dependent variable.**

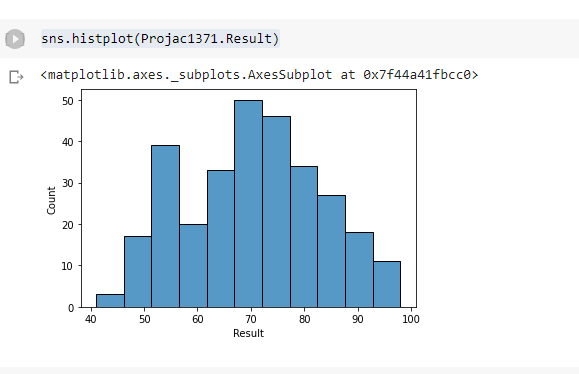
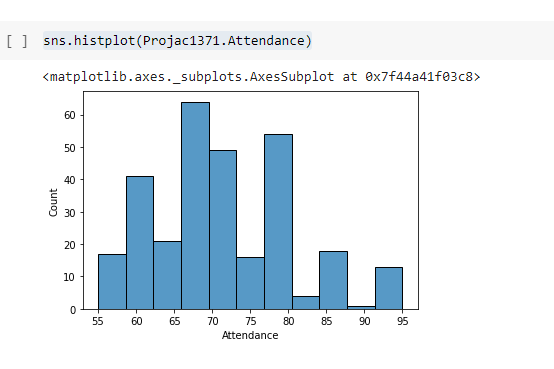
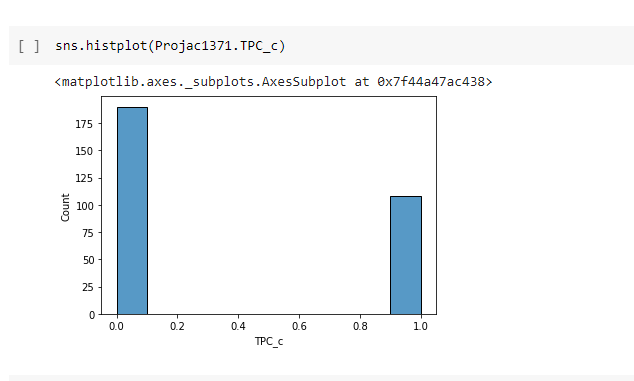
**Part 4: Plotting Histograms of the independent variables and the dependent variable.**

The purpose of a **histogram** is to graphically summarize the distribution of a variable.

I have plotted the histograms of all the independent variables as well as the dependent variable. Below are the snapshots.



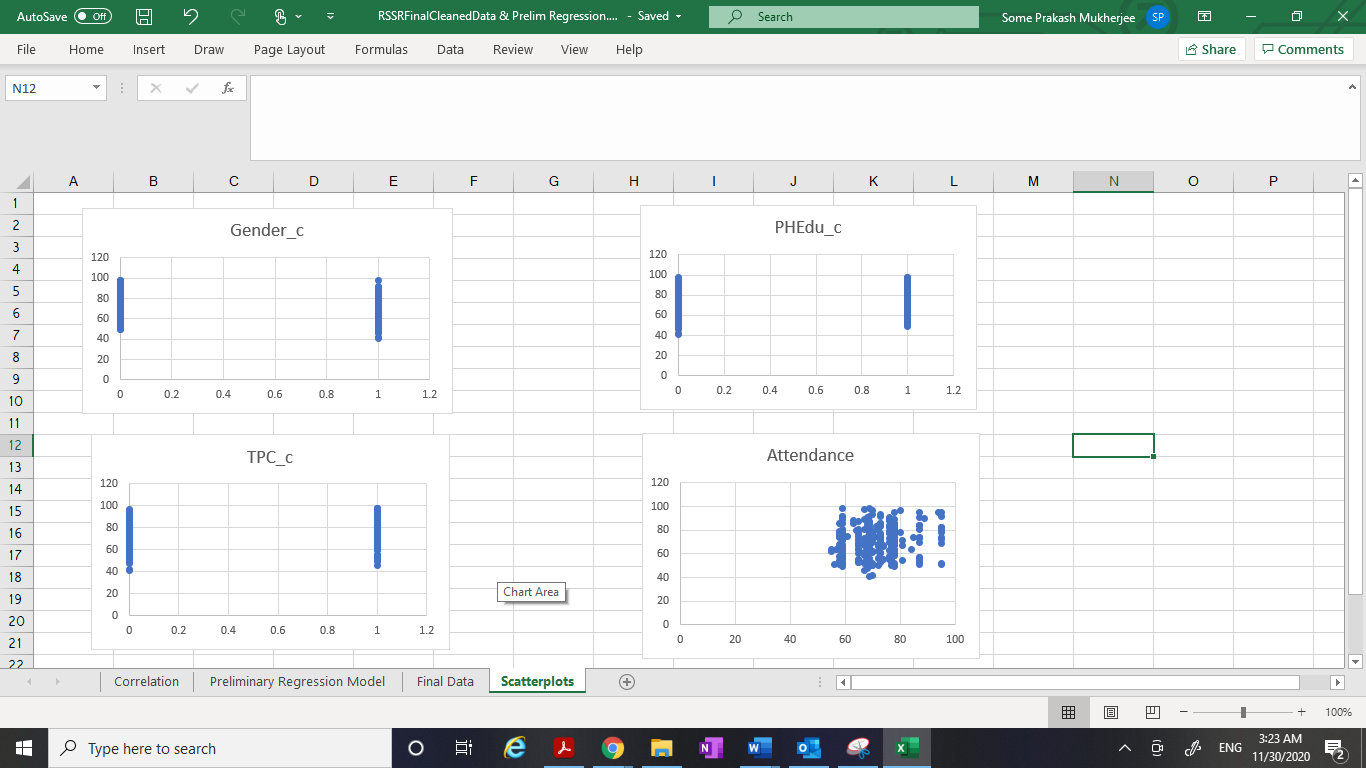




**Gender\_c, PHEdu\_c and TPC\_c** are having Binary data (0 or 1) which gets reflected from the histograms. Also, we saw that Result is normally distributed.

**Part 5: Looking at the scatter plots.**

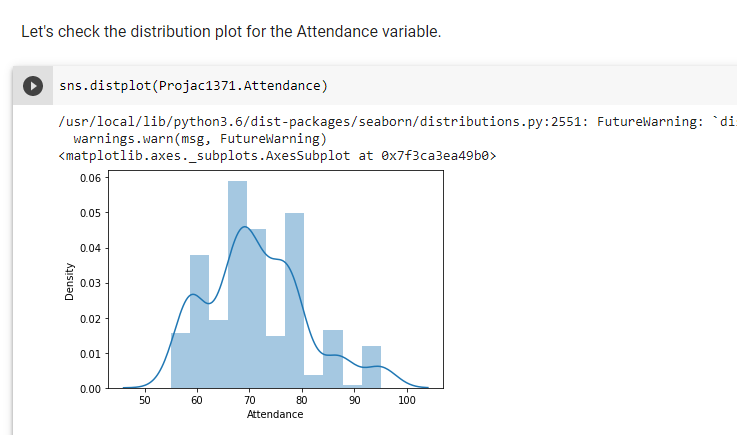
Then, I looked at the scatterplots plotted between the Result (dependent variable) with individual independent variables respectively. I completed this in Excel which are shown in the below snapshot.



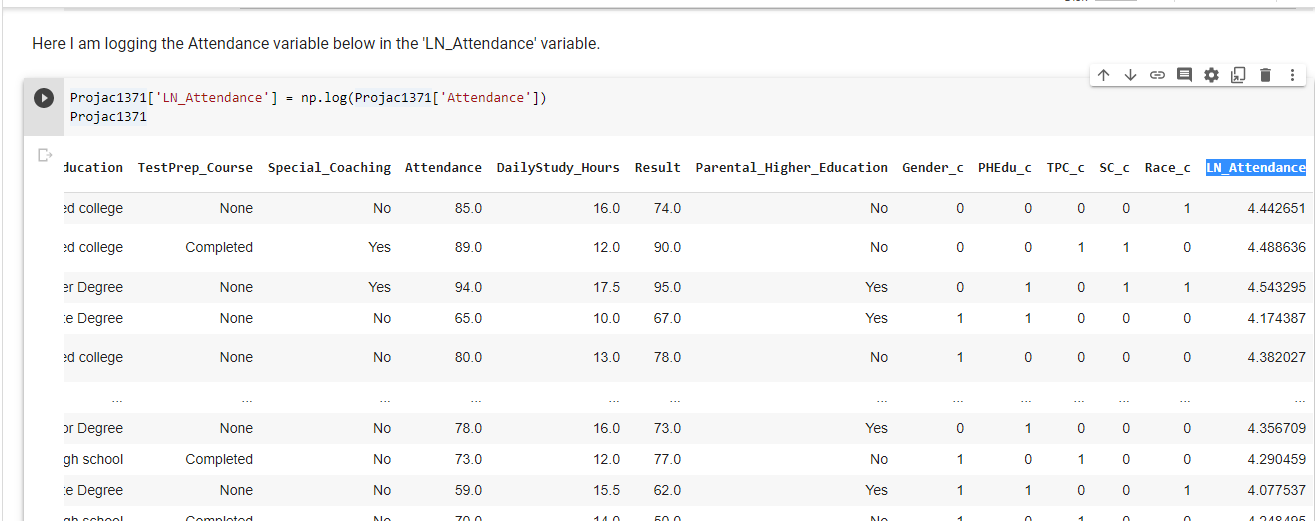
**Part 6: Finding out the correct functional form of the variables.**

Based on the histograms and the scatterplots that I already plotted in Part 4 and Part 5, I tried to find the correct functional form of the variables.

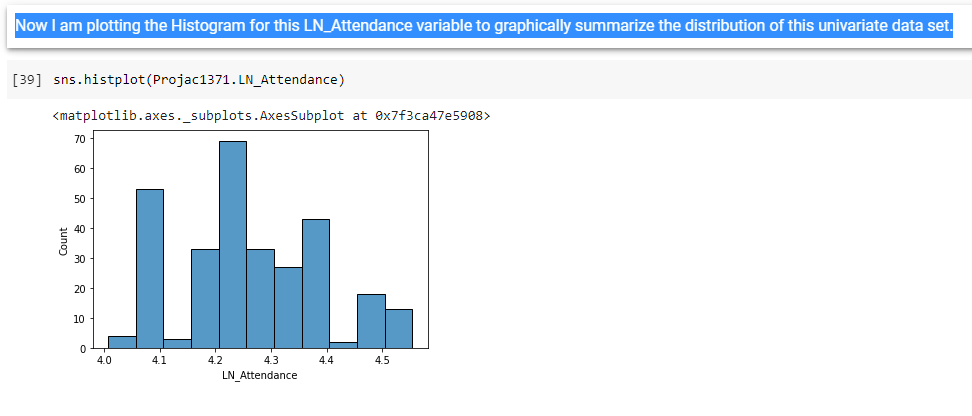
When I looked at the **Attendance histogram, I thought of creating the distribution plot for this variable to analyze more on its distribution.**



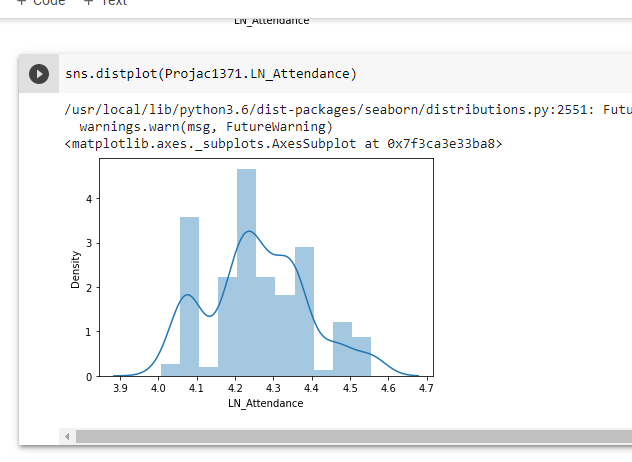
**The distribution plot seemed to be skewed towards the right side and hence I decided to Log the Attendance variable. I created new variable LN\_Attendance.**



**Then, I again plotted the Histogram for this LN\_Attendance variable to graphically summarize the distribution of this univariate data set.**



**Also, I did the Distribution plot for this LN\_Attendance variable and now I can find a better plot in terms of distribution as shown below.**

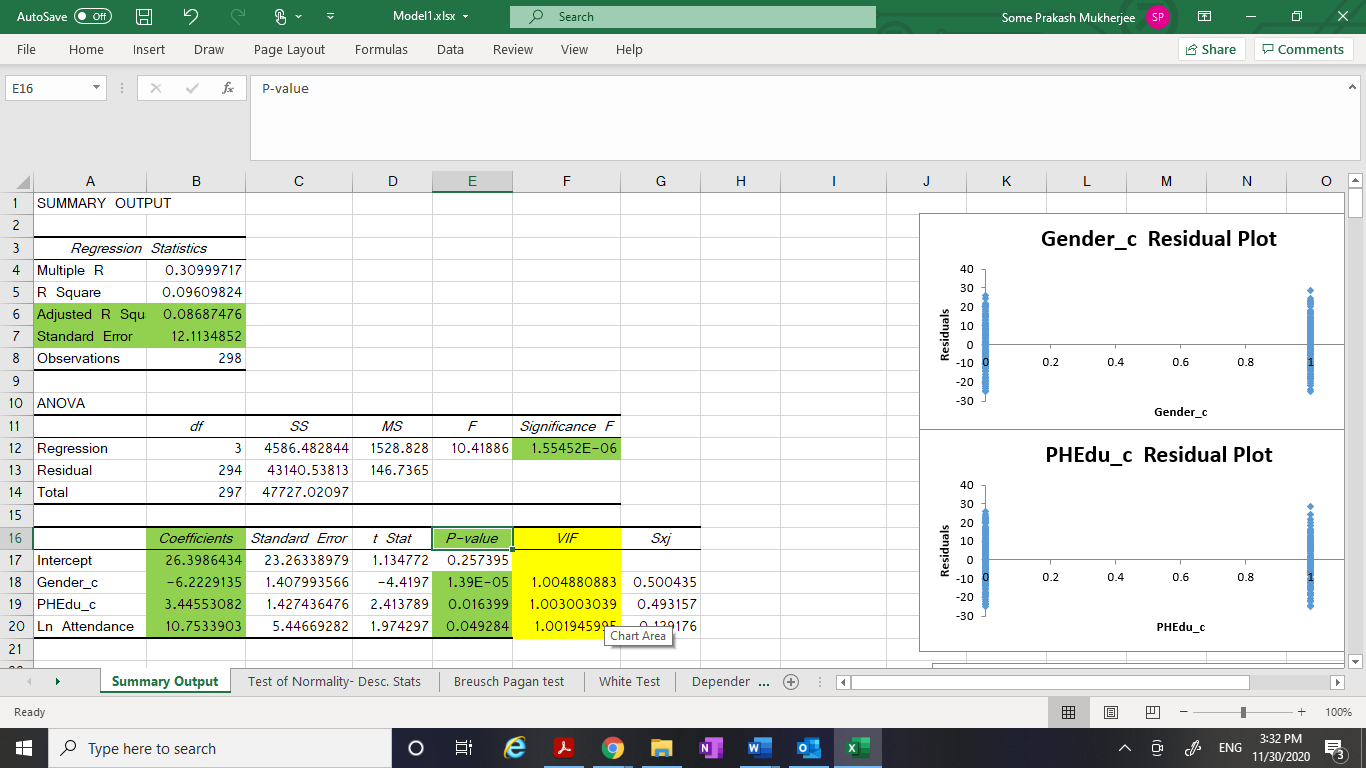


**Part 7: Creation of 3 different Regression models.**

**Model#1:**

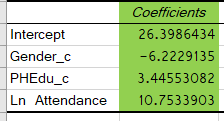
Here I have considered **Gender\_c, PHEdu\_c, Ln\_Attendance** as the independent variables and **Result** as the dependent variable and ran the Regression in the Excel.

Attaching the snapshots here with.

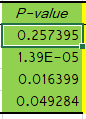


From this Model# l, I came up with below findings:

1. R^2 = 0.096. The model#1 explains **9.6%** of the total variance of the dependent variable.
2. The **Adjusted R Square** is **‘0.0868’**
3. The **Standard Error** is **12.11**
4. The ‘p’ value for the overall F test is **1.5545183E-06** which is significant.
5. The various coefficients below explain about practical significance:



1. The ‘p’ value for the partial slopes that are used in partial slope’s ‘t’ test are:



1. Here I see that the p-value for Intercept is more than 0.05 making it statistically insignificant. But p-value for Gender\_c, PHEdu\_c and Ln Attendance are lesser than 0.05 (Alpha = 0.05) which makes these significant.
2. From all these findings, I came up with **predicted equation** for **Model#1.**

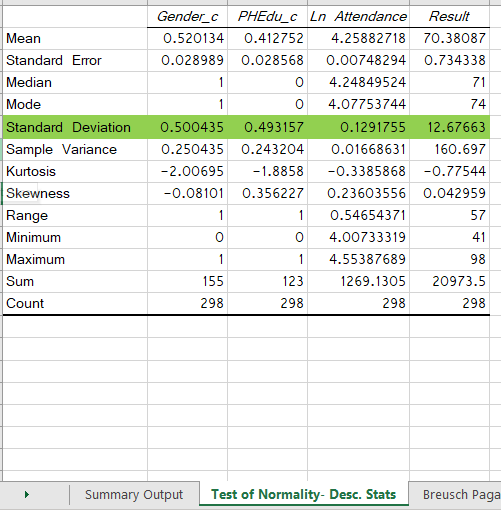
**Y(hat) = B0 + B1\*X1 + B2\*X2 +B3\*Ln(X3)**

Where B0 is the intercept and B1, B2 and B3 are the various respective coefficients. Then I had filled the equation with the actual coefficients values that I got after performing the Regression. Then I got the equation as:

**Result(hat) = 26.398 - 6.222\*Gender\_c + 3.445\*PHEdu\_c +**

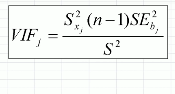
**10.7533\*Ln\_Attendance**

1. I have performed the **‘Test of Normality’** for the Variables in Model# 1. Please find the attached snapshots.

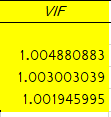


1. Also, I have performed the ‘**test for multicollinearity’** by finding out the **VIF (Variance Inflation Factor)** for the independent variables that were considered in Model#1 in Excel. Please find the attached snapshots.

I have made use of the below formula:

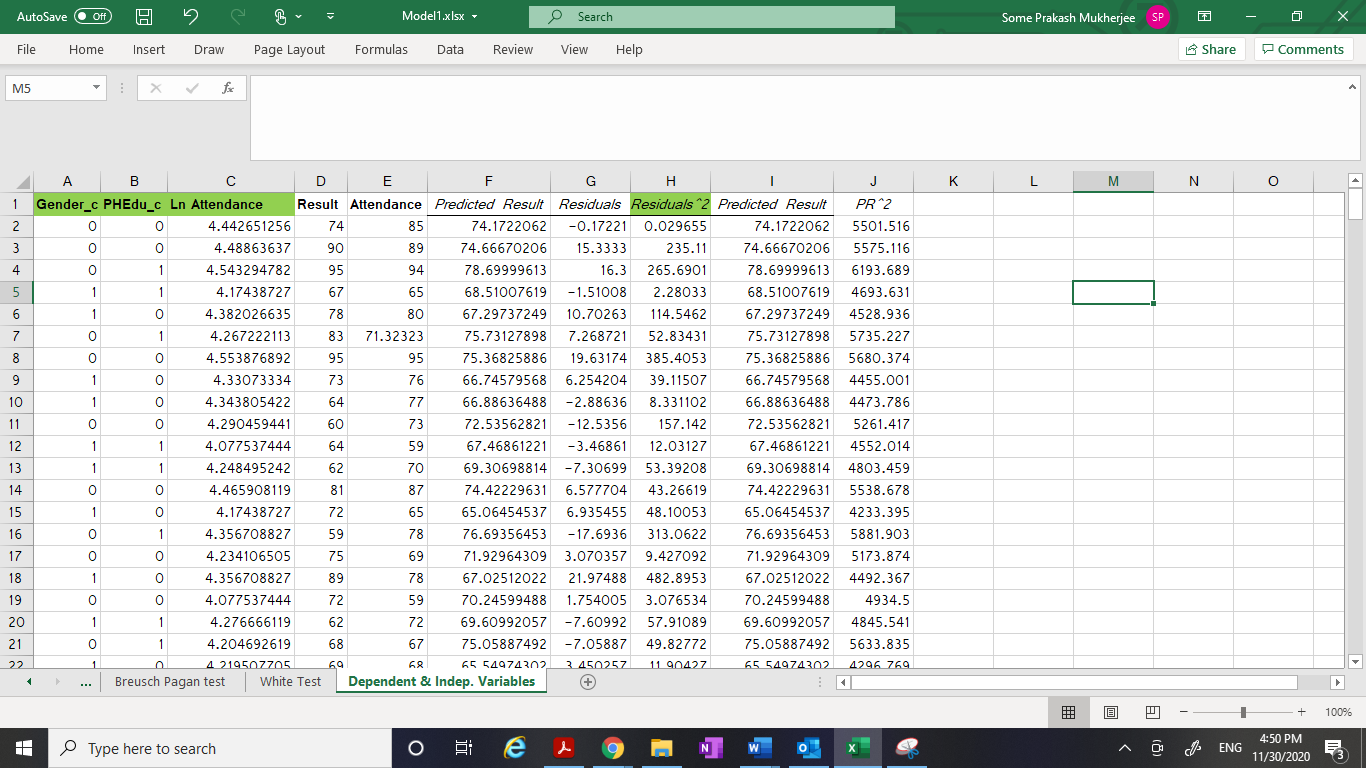


I have got the values as 1.004, 1.003, 1.001 for Gender\_c, PHEdu\_c, Ln( Attendance) respectively. As we know that values of VIF that exceed 10 are often regarded as indicating multicollinearity. **So, in our case, it clearly indicates that there does not exist any multicollinearity in my Model 1.**

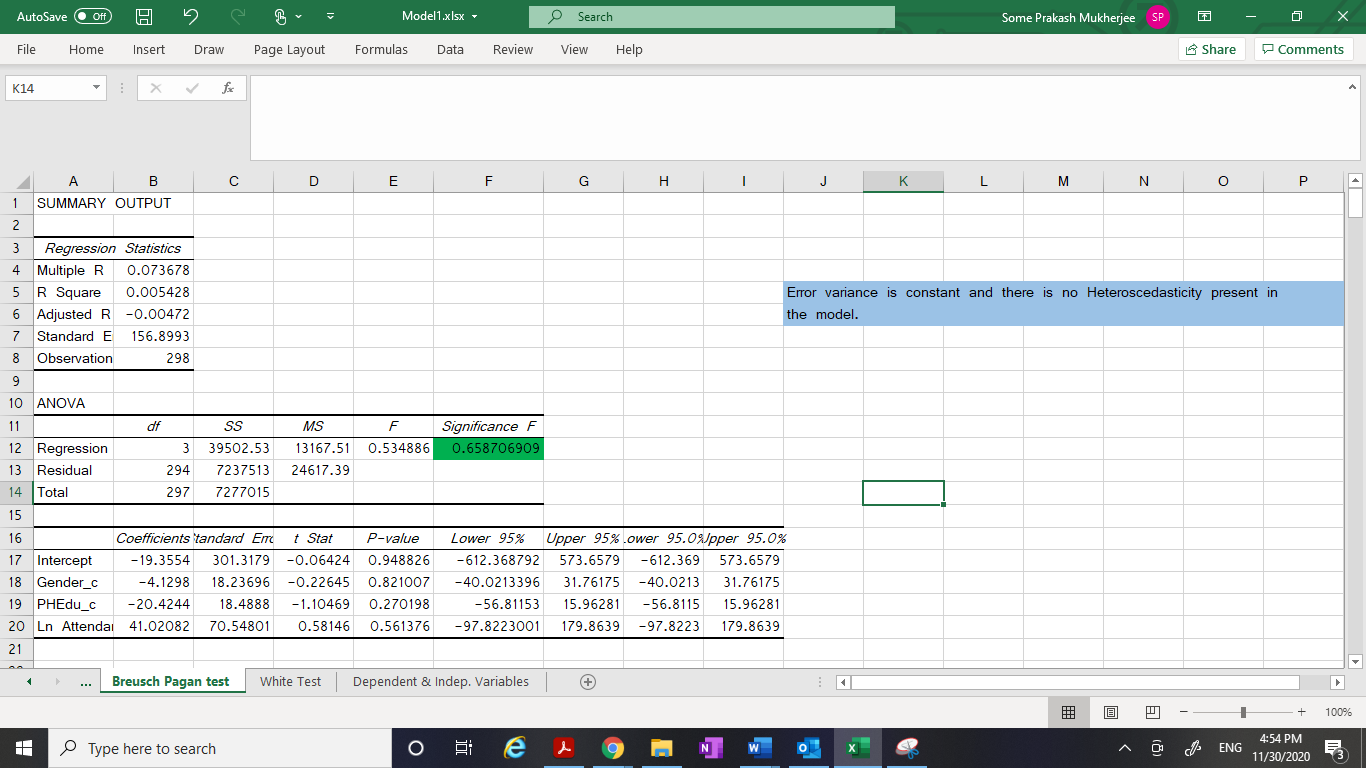


1. I have performed the **‘Test of heteroscedasticity’** for the Model# 1. Please find the attached snapshot. I had looked at the Scatter plot, for the ‘Predicted Result’ in x-axis and ‘Residual^2’ in the y-axis and found that there was not much increase in Residual^2 as the Result value was going higher. This suggests that **we might not have Heteroscedasticity** in our model 1. Here goes the excel snapshot for this.

For more confirmation I applied **Breusch Pagan test** to the Model# 1. This is essentially where I regressed the ‘Residual^2’ on all the independent variables that I had in the model 1. Please find the attached snapshot.



We have now run our regression taking the Residual^2 as the dependent variable and all ‘x’ variables of model 1 as independent. We got the below screen shot.



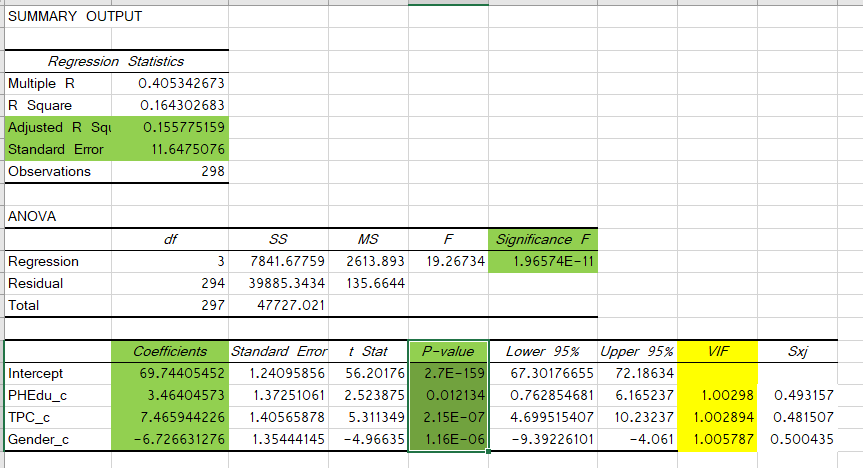
I see here that the **p-value for the F test is “0.658” which is much higher than 0.05. This finally concludes that we fail to reject H0 that our model is homoscedastic. There is no heteroscedasticity in my Model# 1.**

====================================================

**Model#2:**

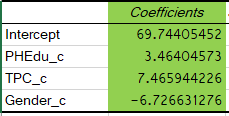
Here I have considered **PHEdu\_c, TPC\_c and Gender\_c** as the independent variables and **Result** as the dependent variable and ran the Regression in the Excel.

Attaching the snapshots here with.

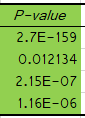


From this Model# 2, I came up with below findings:

1. R^2 = 0.164. The model#1 explains **16.4%** of the total variance of the dependent variable.
2. The **Adjusted R Square** is **‘0.1557’**
3. The **Standard Error** is **11.64**
4. The ‘p’ value for the overall F test is **1.96574E-11** which is much significant.
5. The various coefficients below explain about practical significance:



1. The ‘p’ value for the partial slopes that are used in partial slope’s ‘t’ test are:



1. Here I see that the p-value for Intercept, PHEdu\_c, TPC\_c and Gender\_c and are much lesser than 0.05 (Alpha = 0.05) which make these highly significant.
2. From all these findings, I came up with **predicted equation** for **Model#2.**

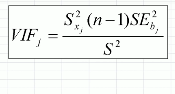
**Y(hat) = B0 + B1\*X1 + B2\*X2 +B3\*X3**

Where B0 is the intercept and B1, B2 and B3 are the various respective coefficients. Then I had filled the equation with the actual coefficients values that I got after performing the Regression. Then I got the equation as:

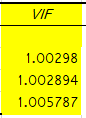
**Result(hat) = 69.744 + 3.464\*PHEdu\_c + 7.465\*TPC\_c - 6.726\*Gender\_c**

1. Also, I have performed the ‘**test for multicollinearity’** by finding out the **VIF (Variance Inflation Factor)** for the independent variables that were considered in Model#2 in Excel. Please find the attached snapshots.

I have made use of the below formula:

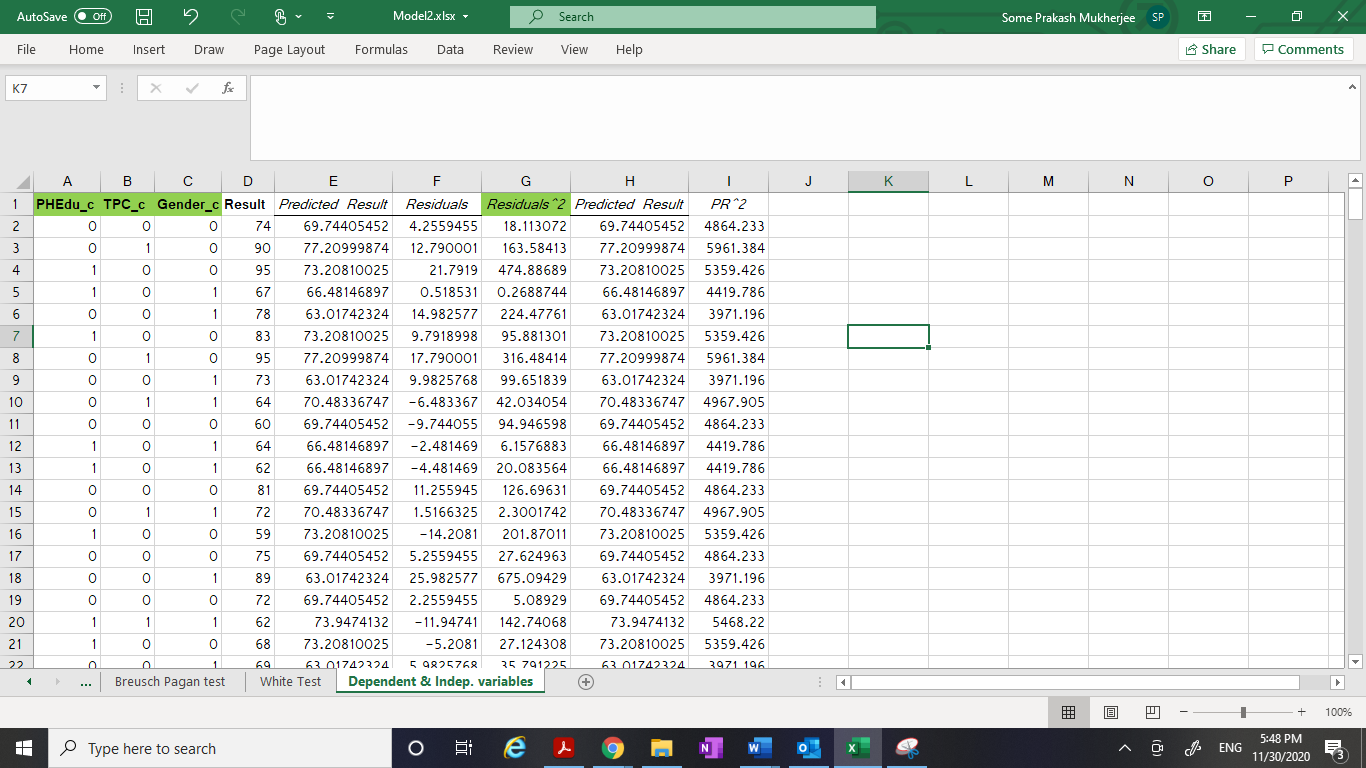


I have got the values as 1.002, 1.002, 1.005 for PHEdu\_c, TPC\_c and Gender\_c respectively. As we know that values of VIF that exceed 10 are often regarded as indicating multicollinearity. **So, in our case, it clearly indicates that there does not exist any multicollinearity in the Model# 2.**

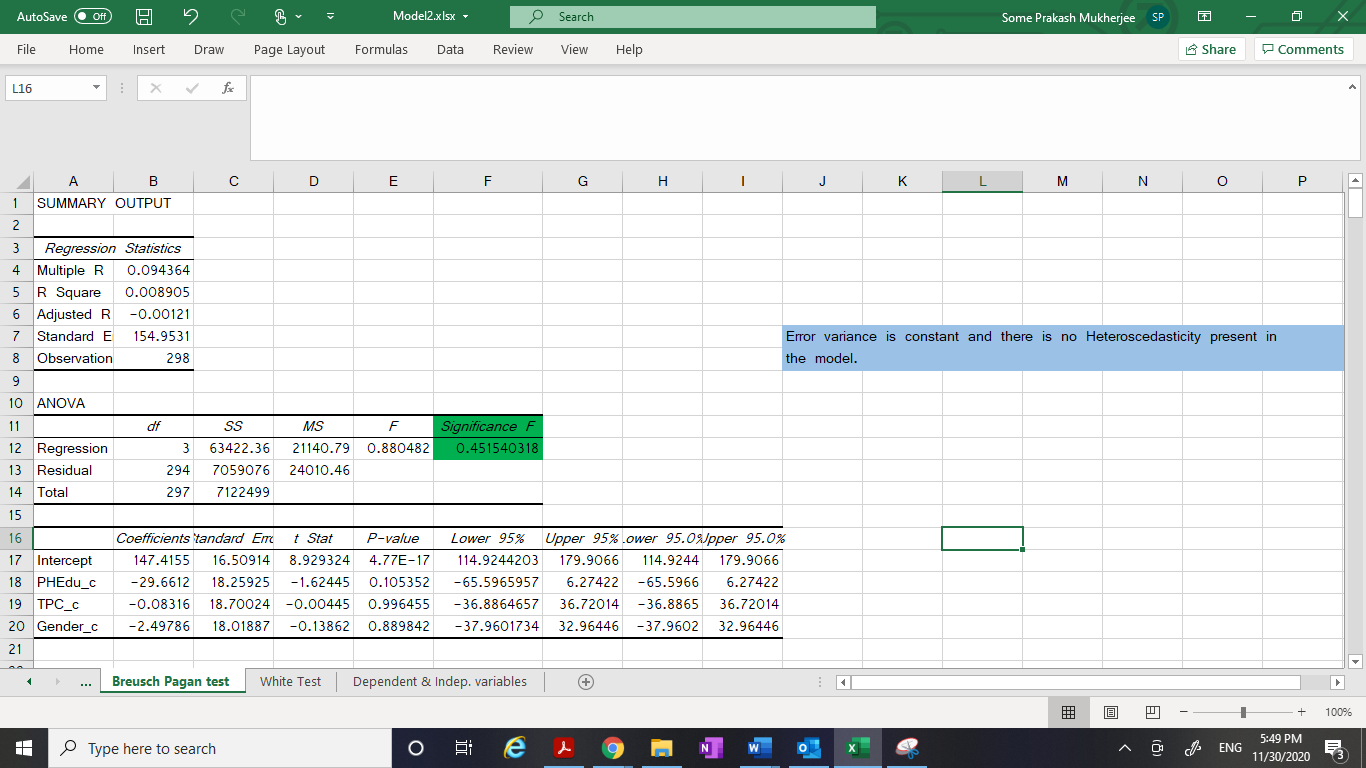


1. I have performed the **‘Test of heteroscedasticity’** for the Model# 2. Please find the attached snapshot. I had looked at the Scatter plot, for the ‘Predicted Result’ in x-axis and ‘Residual^2’ in the y-axis and found that there was not much increase in Residual^2 as the Result value was going higher. This suggests that **we might not have Heteroscedasticity** in our model 2. Here goes the excel snapshot for this.

For more confirmation I applied **Breusch Pagan test** to the Model# 2. This is essentially where I regressed the ‘Residual^2’ on all the independent variables that I had in the model 2. Please find the attached snapshot.



We have now run our regression taking the Residual^2 as the dependent variable and all ‘x’ variables of model 2 as independent variables. We got the below screen shot.

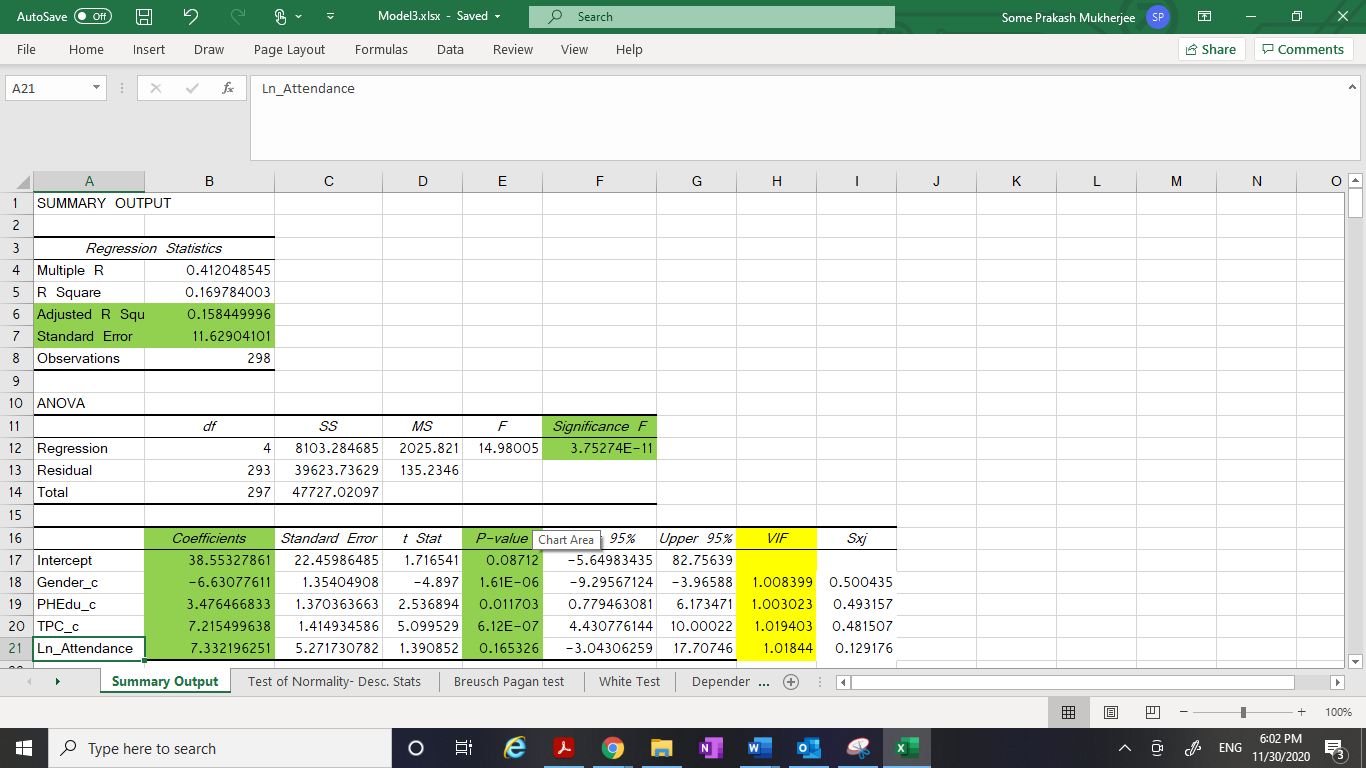


I see here that the **p-value for the F test is “0.451” which is much higher than 0.05. This finally concludes that we fail to reject H0 that our model is homoscedastic. There is no heteroscedasticity in my Model# 2.**

**Model#3:**

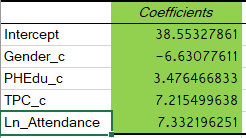
Here I have considered **Gender\_c, PHEdu\_c, TPC\_c and Ln\_Attendance** as the independent variables and **Result** as the dependent variable and ran the Regression in the Excel.

Attaching the snapshots here with.

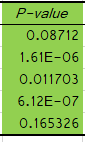


From this Model# 3, I came up with below findings:

1. R^2 = 0.169. The model#1 explains **16.9%** of the total variance of the dependent variable.
2. The **Adjusted R Square** is **‘0.158’**
3. The **Standard Error** is **11.629**
4. The ‘p’ value for the overall F test is **3.75274E-11** which is much significant.
5. The various coefficients below explain about practical significance:



1. The ‘p’ value for the partial slopes that are used in partial slope’s ‘t’ test are:



1. Here I saw that the p-value for Intercept and Ln\_Attendance are more than 0.05 making it statistically insignificant. But p-value for Gender\_c, PHEdu\_c and TPC\_c is much lesser than 0.05 (Alpha = 0.05) which makes these highly significant.
2. From all these findings, I came up with **predicted equation** for **Model#3.**

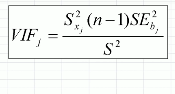
**Y(hat) = B0 + B1\*X1 + B2\*X2 +B3\*X3 + B4\*Ln(X4)**

Where B0 is the intercept and B1, B2, B3 and B4 are the various respective coefficients. Then I had filled the equation with the actual coefficients values that I got after performing the Regression. Then I got the equation as:

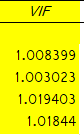
**Result(hat) = 38.553 - 6.630\*Gender\_c + 3.476\*PHEdu\_c + 7.215\*TPC\_c + 7.332\*Ln\_Attendance**

1. Also, I have performed the ‘**test for multicollinearity’** by finding out the **VIF (Variance Inflation Factor)** for the independent variables that were considered in Model#3 in Excel. Please find the attached snapshots.

I have made use of the below formula:

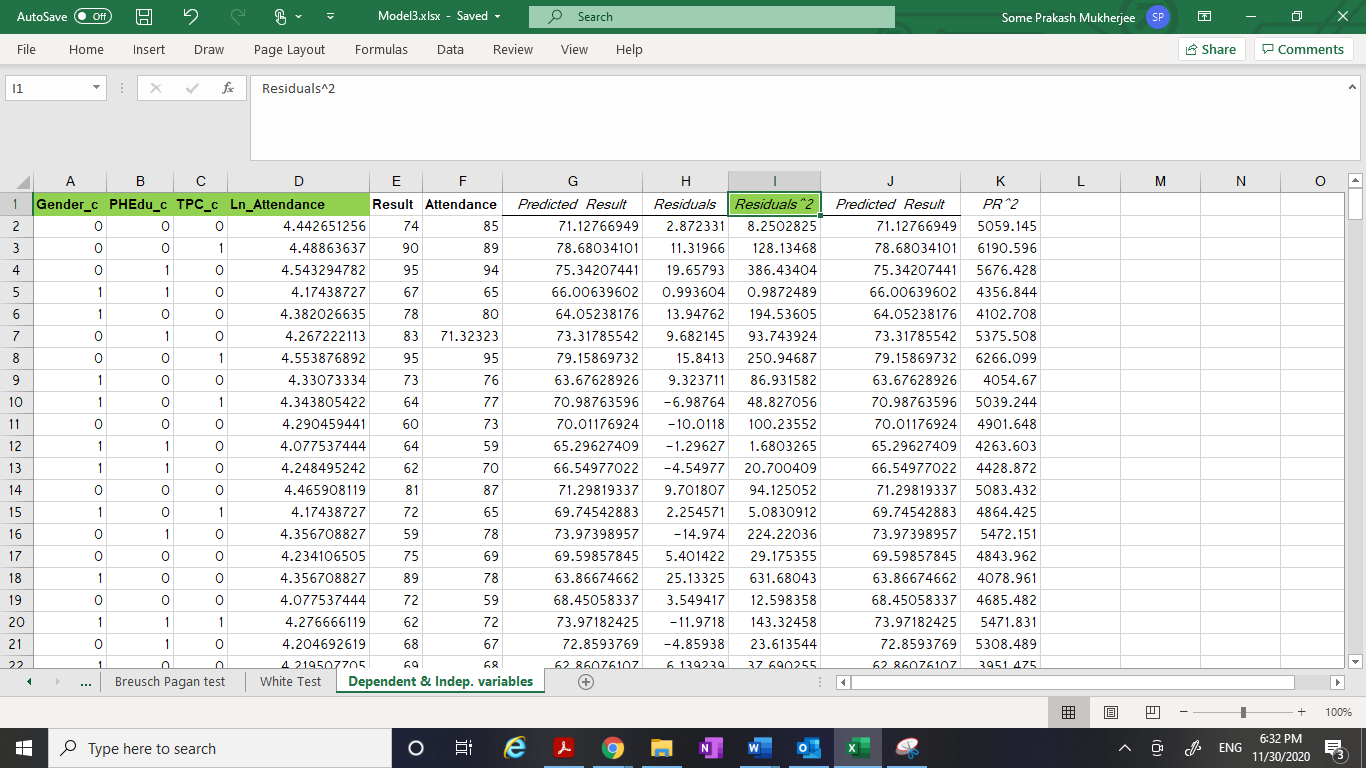


I have got the values as 1.008, 1.003, 1.019 and 1.018 for Gender\_c, PHEdu\_c, TPC\_c and Ln\_Attendance respectively. As we know that values of VIF that exceed 10 are often regarded as indicating multicollinearity. **So, in our case, it clearly indicates that there does not exist any multicollinearity in the Model 3.**

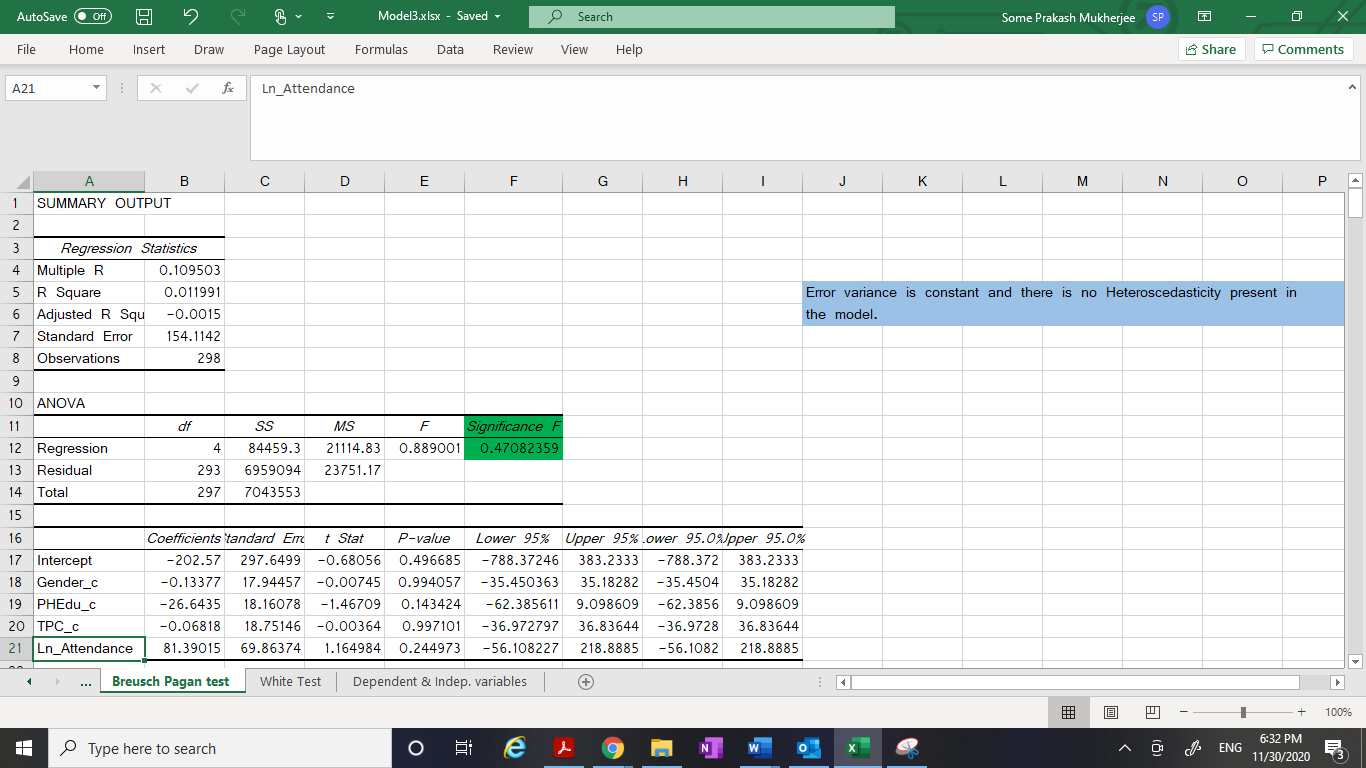


1. I have performed the **‘Test of heteroscedasticity’** for the Model# 3. Please find the attached snapshot. I had looked at the Scatter plot, for the ‘Predicted Result’ in x-axis and ‘Residual^2’ in the y-axis and found that there was not much increase in Residual^2 as the Result value was going higher. This suggests that **we might not have Heteroscedasticity** in our model 3. Here goes the excel snapshot for this.

For more confirmation I applied **Breusch Pagan test** to the Model# 3. This is essentially where I regressed the ‘Residual^2’ on all the independent variables that I had in the model 3. Please find the attached snapshot.



We have now run our regression taking the Residual^2 as the dependent variable and all ‘x’ variables of model 3 as independent variables. We got the below screen shot.



I saw here that the **p-value for the F test is “0.470” which is much higher than 0.05. This finally concludes that we fail to reject H0 that our model is homoscedastic. There is no heteroscedasticity in my Model# 3.**

**Part 8: Picking up the best model**

I have compared the values of “Adjusted R Square “and “Standard Error” among all the three different models. **Adjusted R**-**squared** increases only when independent variable is significant and affects dependent variable. **Since the “Adjusted R Square” is highest in Model 3 (compared to model 1 and 2) and “Standard Error” is least in “Model 3” compared to Model 1 and 2, so I predict that Model 3 is the best among all the three models.**

|  |  |  |
| --- | --- | --- |
|  | Adjusted R Square | Standard Error |
| Model 1 | 0.086874755 | 12.11348522 |
| Model 2 | 0.155775159 | 11.6475076 |
| Model 3 | 0.158449996 | 11.62904101 |

**Part 9: Output explanation and findings from the best Model (Model# 3):**

1. There was total 298 samples.
2. The NULL hypothesis or H0 for regression states that:

B0 = B1 = B2 =B3=B4 =0. It states that there is no relationship between the output and input variables which can be described with any model.

The alternate hypothesis H(a) states that at least one of the ‘B’s is not equal to zero and there exists some form of relationship between the output and input variables and hence, the model exists that explains this relationship.

Now for proving that there exists some relationship between the variables, we had to reject Null hypothesis (H0).

1. ANOVA is the Analysis of Variance of the dependent variable. It is especially important because it tells whether the model is useful or not. It is breaking the variance into what is explained by the model and what is not explained by the model.

The total variance of the ‘Result’ variable is 47727.02 and out of that, Model 3 explains 8103.28 and the unexplained part is 39623.7 given by Residual.

1. R^2 = SSR/SST = 0.169
2. Degree of Freedom is n-1 = 298 – 1 = 297
3. Mean Square (MS) = SS/ df
4. The probability of significance of F is **3.75274e-11**. Alpha is the level of significance. At 95% confidence interval, value of alpha is 0.05.

So here the p-value of F test is much lesser than 0.05 which means that we are in the rejection region and we are confident to reject H0.

This concludes that the model 3 is better than the H0 and there is at least one coefficient which is not 0.

1. The Standard error of the regression Model1 is 11.629
2. **Adjusted R-squared** increases only when independent variable is significant and affects the dependent variable. Here the **Adjusted R-squared** is 0.158
3. The coefficients section helped in describing about the individual variables also. We saw that:
4. For Intercept (B0): the value ranges from -5.649 to 82.75 which means it contains 0 in the range. Also, the p-value for t test here is 0.08 which is more than 0.05 which is the threshold value. **So, Intercept is statistically not significant.**
5. For Gender\_c(B1): the value ranges from -9.29 to -3.9 which means it does not include 0 in the range. Also, the p-value for t test here is 1.61e-06 which is much less than 0.05 which is the threshold value. So, **Gender\_c is statistically significant.**
6. For PHEdu\_c (B3): the value ranges from 0.77 to 6.17 which means it does not contain 0 in the range. Also, the p-value for t test here is 0.011 which is very lesser than 0.05 which is the threshold value. So, **PHEdu\_c is statistically significant.**
7. For TPC\_c (B3): the value ranges from 4.43 to 10.00 which means it does not contain 0 in the range. Also, the p-value for t test here is 6.12e-07 which is very lesser that 0.05 which is the threshold value. So, **TPC\_c is statistically significant.**
8. For Ln\_Attendance (B4): the value ranges from -3.0 to 17.70 which means it includes 0 in the range. Also, the p-value for t test here is 0.16 which is higher than 0.05 which is the threshold value. So, **Ln\_Attendance is statistically not significant.**

From the Model# 3, I can state that H0 is rejected and Model# 3 is better than H0. Hence, we have proved that there is a relationship between the variables that is explained by Model 3.

The model#3 states the relationship among the variables as follows:

**Result(hat) = 38.553 - 6.630\*Gender\_c + 3.476\*PHEdu\_c + 7.215\*TPC\_c + 7.332\*Ln\_Attendance**

Where ‘Result’denotes the **Graduation Result Percentage**, ‘Gender\_c’ tells about the gender, ‘PHEdu\_c’ denotes whether parent is highly educated or not, ‘TPC\_c’ denotes whether Test Preparation Course was completed or not and ‘Attendance’ denotes the percentage of attendance.

So, parent higher education, and completion of test preparation courses indeed had a positive impact on the student’s graduation performance.

**Part 10: Interpretation of the results and explaining the social implications of it.**

As compared to Female student, the Graduation Result percentage decreases by 6.63 for Male student, while everything else is constant.

As compared to the student whose parents are not highly qualified, the Graduation Result percentage increases by 3.476 for the students whose parents are highly qualified, while everything else is constant.

As compared to the students who did not complete Test preparation courses, the Graduation Result percentage increases by 7.215 for the students who completed Test preparation courses, while everything else is constant.

This Model #3 is a **semi log model** as we have one logarithmic form of independent variable (Attendance) and our dependent variable is not in logarithmic form. As we know that whenever we are using Logarithmic form for any of the variable in the regression model, the coefficient (B value) tells the percentage change for that variable. That is the increase or decrease is measured as B%. Here Attendance variable had been logged and we had Ln(Attendance) in our Model.

So, the model speaks the below statement about the Attendance variable:

For 1 percent increase in Attendance percentage, Graduation Result percentage increases by 0.07332 (i.e 7.332/100), while everything else is constant.

Here I am describing the **social implications** of the findings.

Based on the data analysis, it is quite imperative that teenager female students remain more focused on studies compared to male counterparts who get easily distracted on a wide array of things ranging from sports, outdoor activities, gaming, music and other extra-curricular activities. The thread which ties together the attention in students easily loosens up in male student compared to female and hence we see a drop in their performance during graduation.

Similarly, parents who are highly educated and earned higher degrees have the natural tendency to pass on the value of education to their off springs. Additionally, highly qualified parents tend to provide their kids with an atmosphere to learn and grow and they set it all by example which makes it easy for kids to interpret.

Also, the test preparatory courses provide an edge in the success of the students through giving them a look and feel of the actual exam pattern and environment. Students who complete those get equipped with the techniques required to set themselves apart from rest of the students in high school graduation exam

Last but not the least, attending class sessions helps students to stay on track, understand expectations, foster important peer social interactions, and generally promote a sense of connectedness. Increasingly, attendance is being understood as a precursor and leading indicator for student success. Attendance improves performance.

**Part 11: Forecast future values:**

The estimated equation could be better understood by using that in forecasting a future value of the Result variable. In this case, the estimated equation is:

**Result(hat) = 38.553 - 6.630\*Gender\_c + 3.476\*PHEdu\_c + 7.215\*TPC\_c + 7.332\*Ln\_Attendance**

Suppose high schooler Tom who maintains a 90% attendance record has reasons to believe that Test preparation Courses hardly impacts his performance. His highly educated parents are really worried about his graduation results.

The regression equation can be used to predict the result percentage of Tom in both the scenarios (with and without completing Test preparation Courses).

Case 1: Tom considers NOT to complete Test Preparation Courses.

Plugging in the values of Gender\_c = 1, PHEdu\_c = 1, TPC\_c = 0 and Attendance = 90 in the above equation, I get:

Result = 38.553 – 6.630\*1 + 3.476\*1 + 7.215\*0 + 7.332\*ln(90)

= 68.33

Case 2: Tom considers completing Test Preparation Courses.

Plugging in the values of Gender\_c = 1, PHEdu\_c = 1, TPC\_c = 1 and Attendance = 90 in the above equation, I get:

Result = 38.553 – 6.630\*1 + 3.476\*1 + 7.215\*1 + 7.332\*ln(90)

= 75.55

So, the regression equation could really help in estimating the result of students that is impacted by multiple variables.

**Part 12: Impact of Social desirability biasness on the regression analysis**

Social-desirability bias is a type of response bias that is the tendency of survey respondents to answer questions in a manner that will be viewed favorably by others. It can take the form of over-reporting "good behavior" or undesirable behavior. The dataset that I worked on to determine validity of relationships between the Student Performance and its various deciding factors (using Regression Analysis) is social desirability biased as respondents often hide their true attitudes - to impress the researcher or interviewer or to preserve one’s self-esteem. That is why the coefficient of variation (also known as R2) that is used to determine how closely a regression model “fits” or explains the relationship between all the independent variables (Parent higher education, Gender, Test preparation course and Attendance) and the dependent variable (student Performance) has emerged low. Moreover, some other factors like inherent talent, genetic aptitude also determine the performance of a student that are very unlikely determined by any mathematical or statistical models.

------------------------------xxxxxxxxxxxxxxxxxxxxxxx------------------------------------