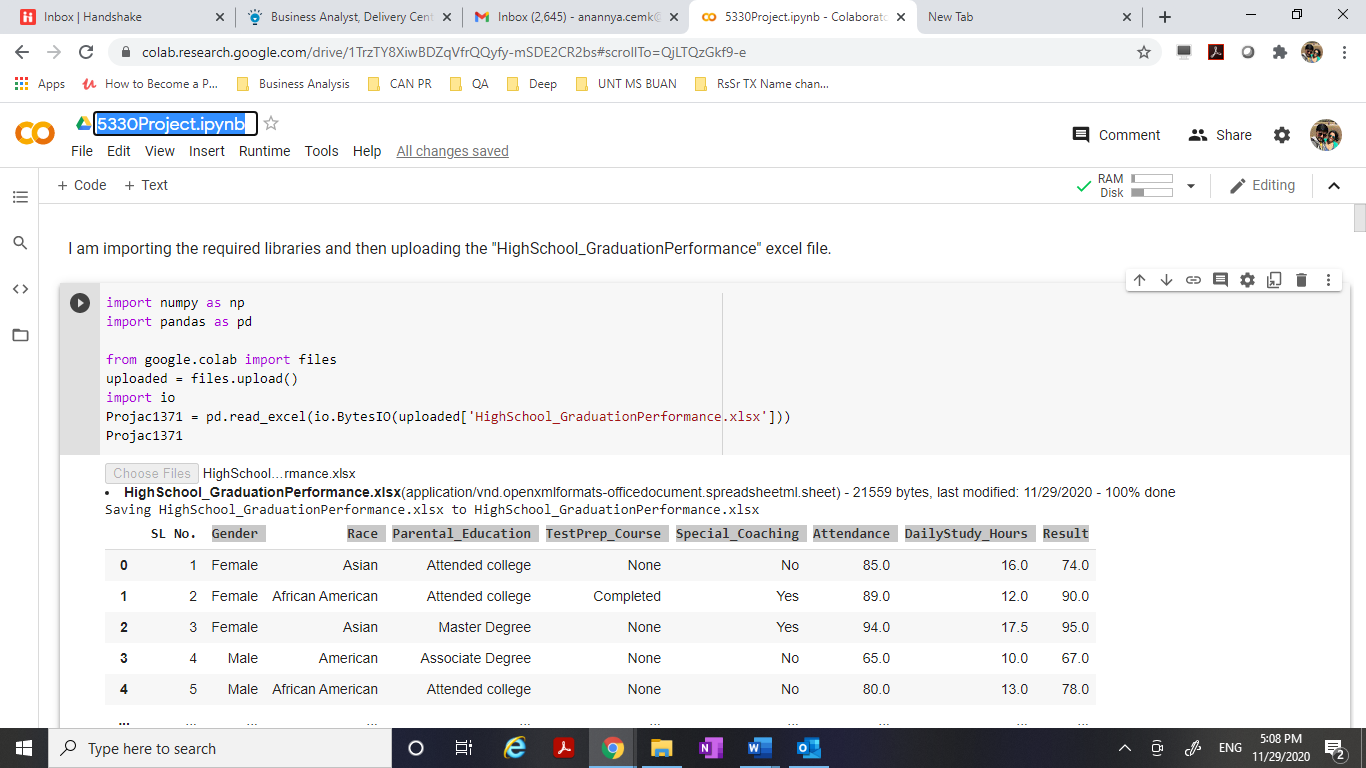
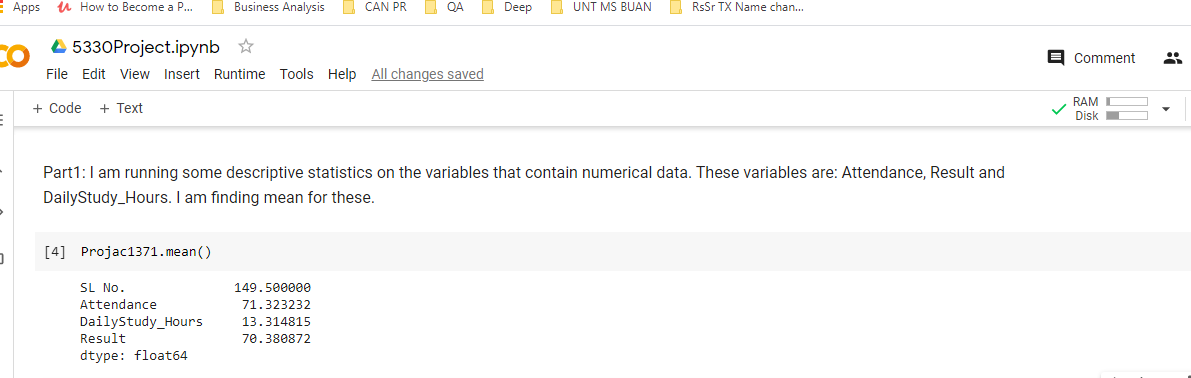
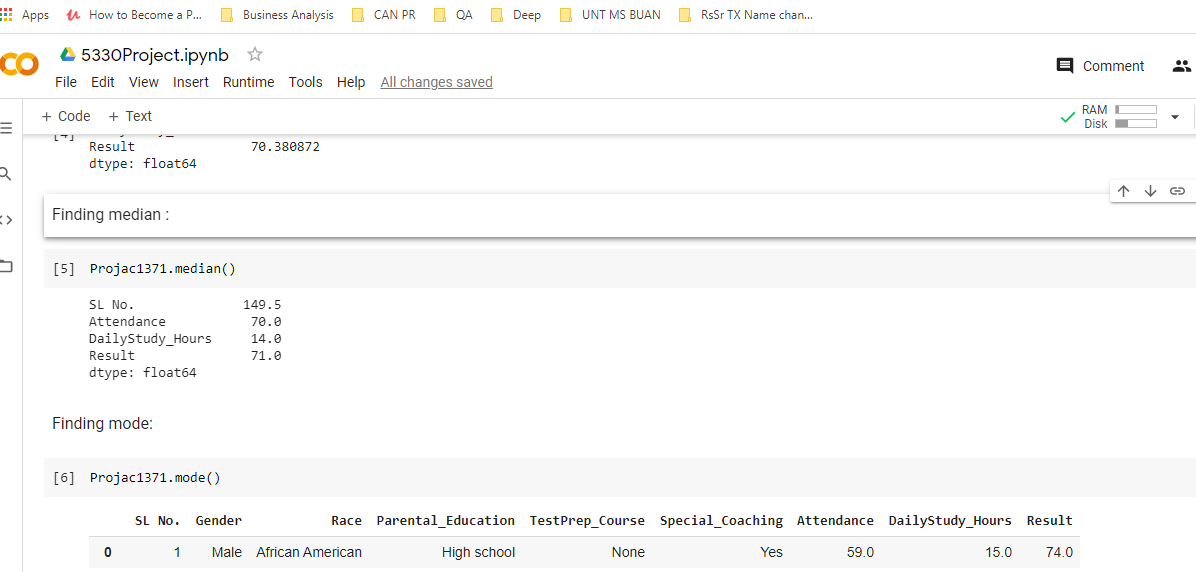
**5330 Project: Snapshots of the Python Code along with the output and explanation from the Google colab environment.**

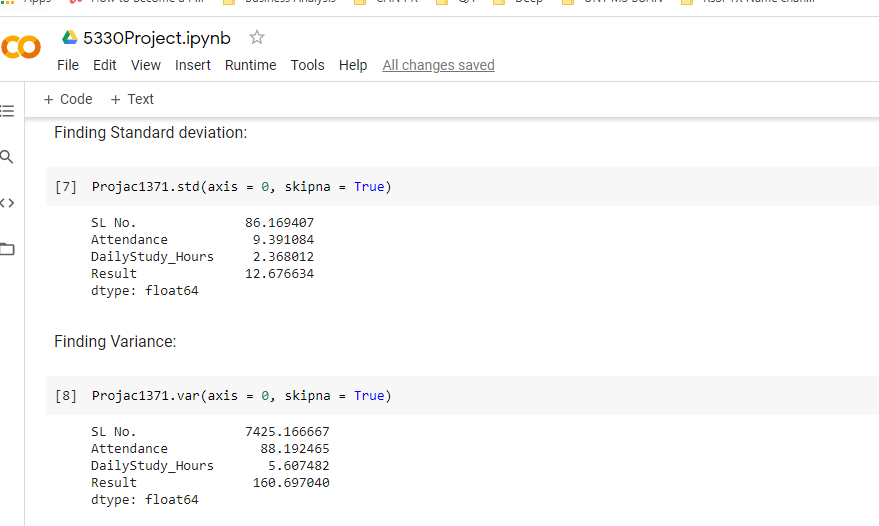
**Data Preparation:**

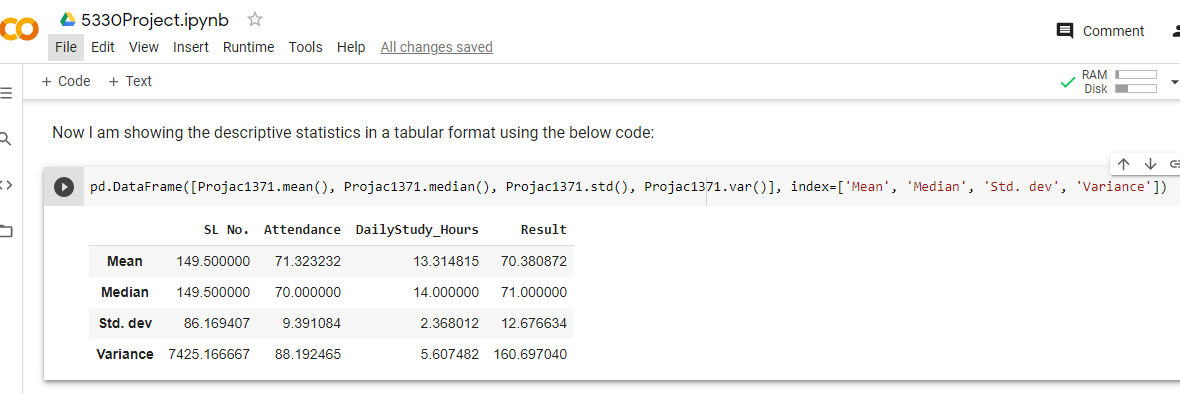
**Part 1: Running descriptive statistics on the various variables.**





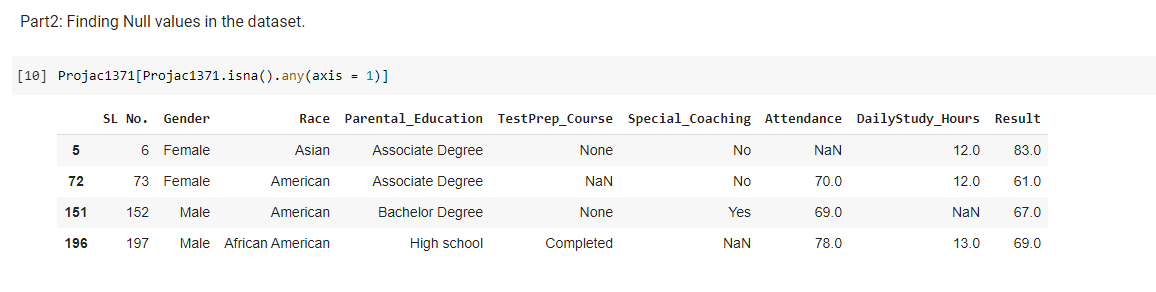






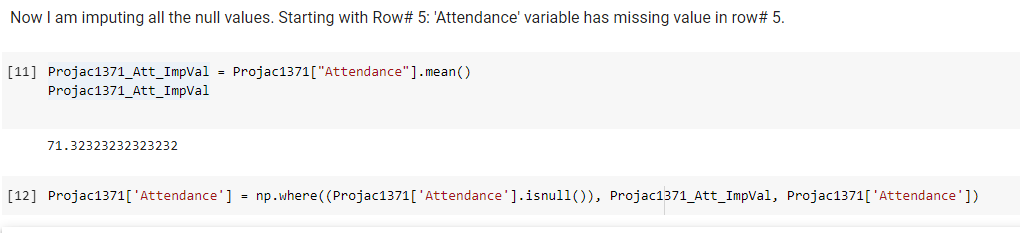
**Part 2: Here, we are taking care of the missing values in the data and following the process of imputation**.

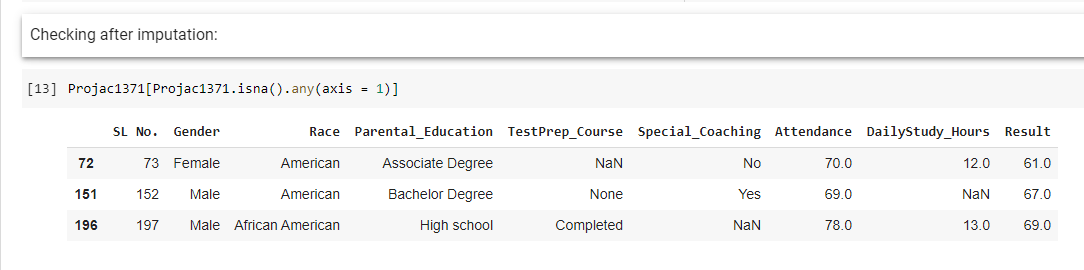
I have used the Python code to find the rows with missing values. Below is the snapshot for the same.



**Row# 5:**

Attendance variable has missing value in row# 5. I am imputing that with the mean of the values in Attendance column.

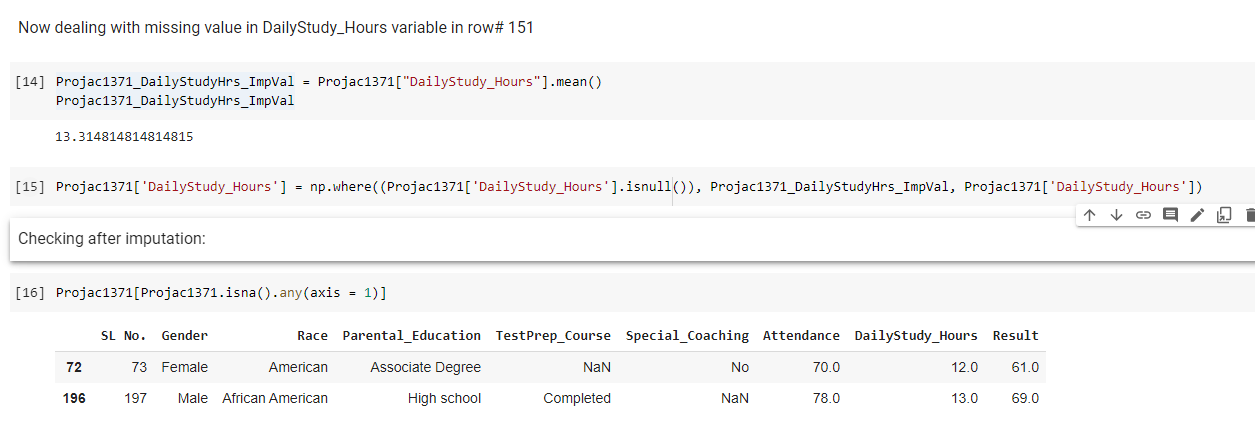




Row # 5 did not appear in the result set because the null value in that record got imputed successfully.

**Row# 151:**

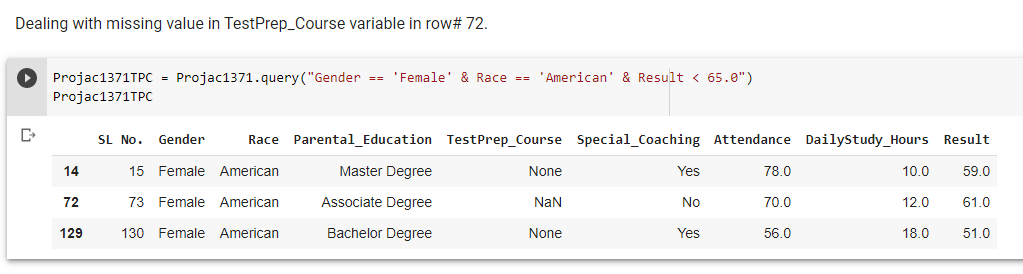
DailyStudy\_Hours variable has missing value in row# 151. I am imputing that with the mean of the values in DailyStudy\_Hours column.



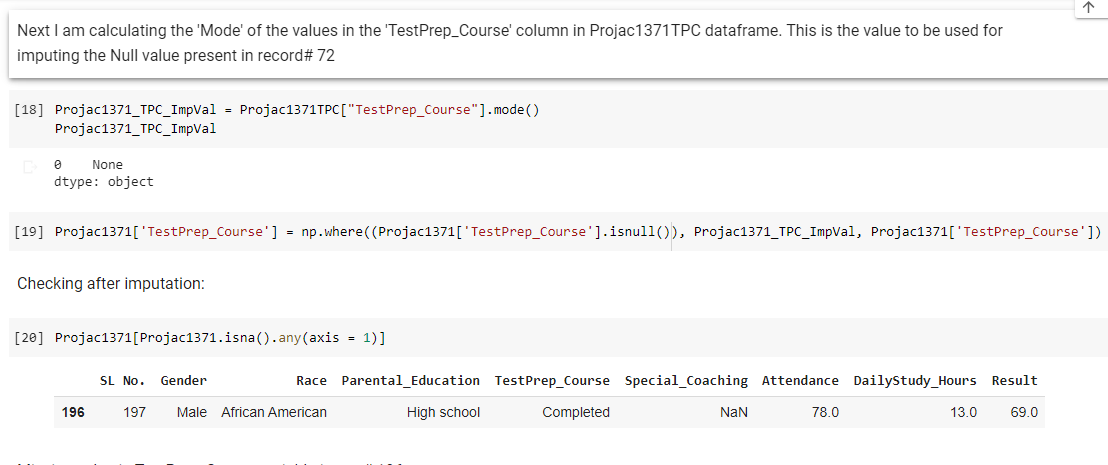
Row # 151 did not appear in the result set because the null value in that record got imputed successfully.

**Row# 72:**

Let me take Record# 72 into consideration. I want to query for similar students who are Female, American and have secured less than 65% in Result.



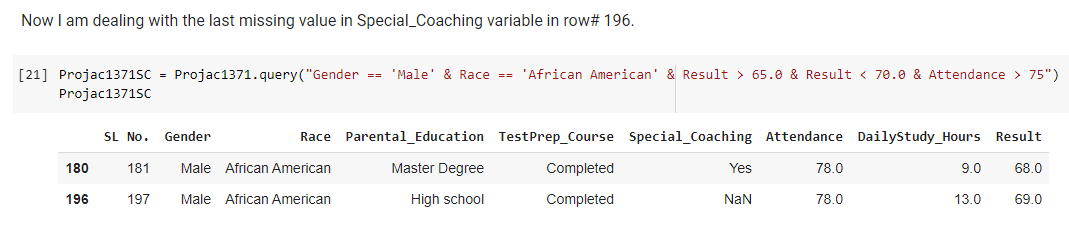
Next, I am calculating the 'Mode' of the values in the 'TestPrep\_Course' column in Projac1371TPC dataframe. This is the value to be used for imputing the Null value present in record# 72.



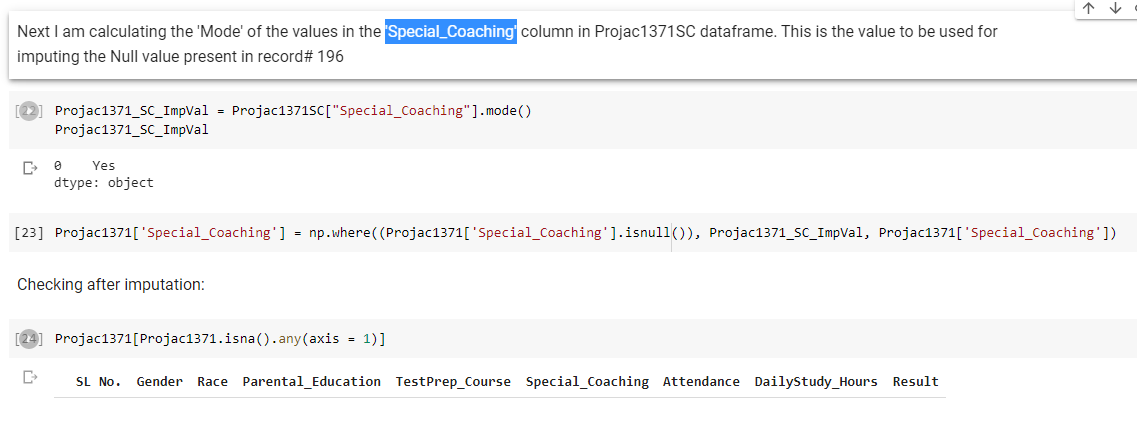
Row # 72 did not appear in the result set because the null value in that record got imputed successfully.

**Row# 196:**

Let me take Record# 196 into consideration. I want to query for similar students who are Male, African American and have secured Result percentage between 65 and 70 with an attendance percentage gtreater than 75.



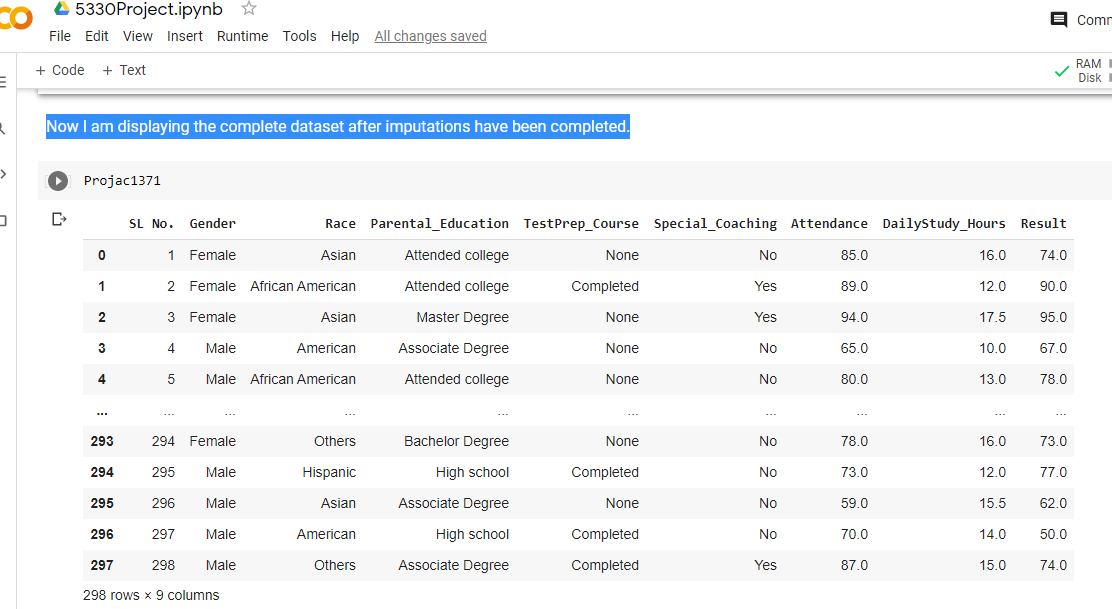
Next, I am calculating the 'Mode' of the values in the “Special\_Coaching'' column in Projac1371SC dataframe. This is the value to be used for imputing the Null value present in record# 196.



**So, we finally see now that there is no variable that has any missing value in it.**

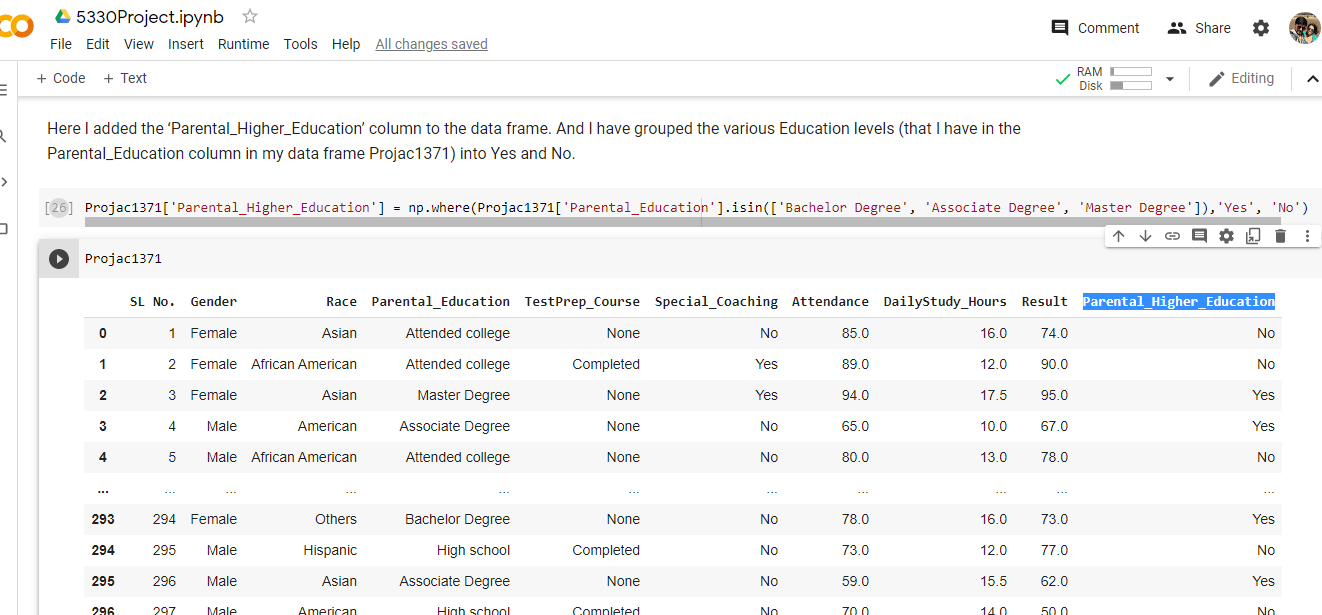
**Part 3: Here, we are transforming our various categorical variables.**

I am displaying here the complete dataset after imputations have been completed.

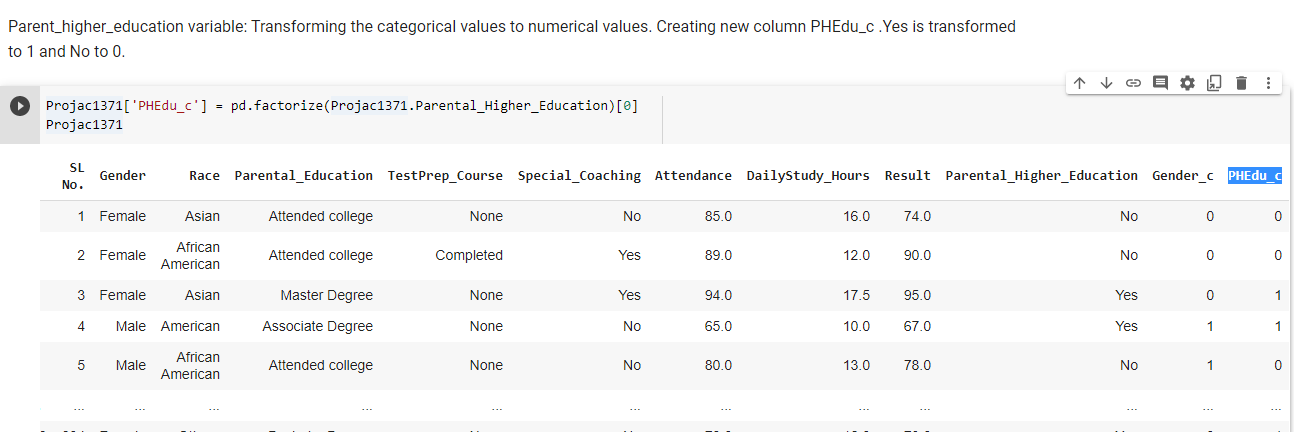


Now I have added the ‘**Parental\_Higher\_Education**’ column to the data frame. I have grouped the various Education levels in the Parental\_Education column in my data frame (Projac1371) into Yes and No. **The Parental Education levels that are either ‘Associate Degree’, ‘Bachelor Degree’ or ‘Master Degree’ are categorized as ‘Yes’. The Parental Education levels that are either ‘Attended college’ or ‘High school’ are categorized as ‘No’.**

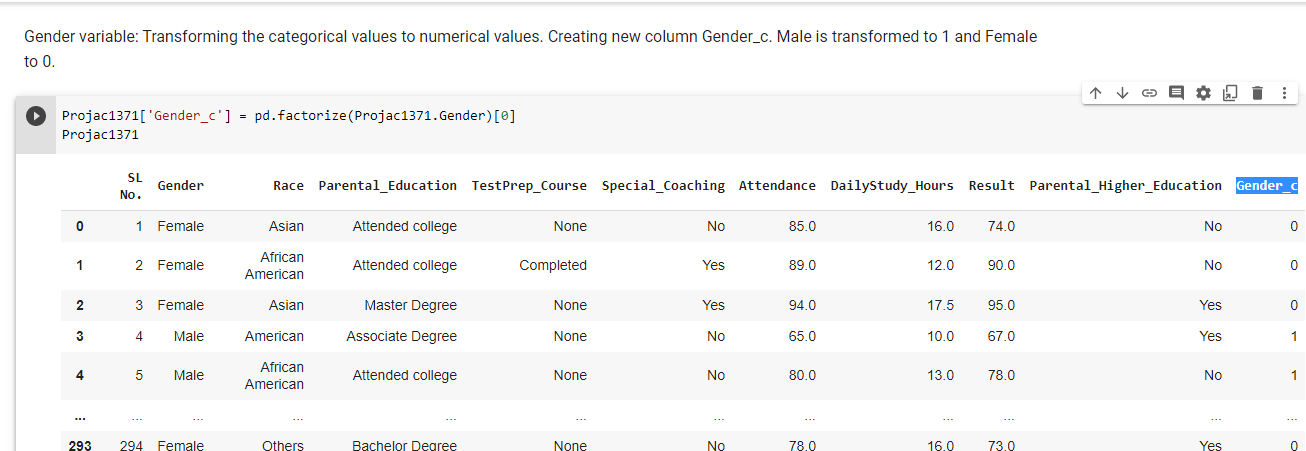
**Below is the Python code.**



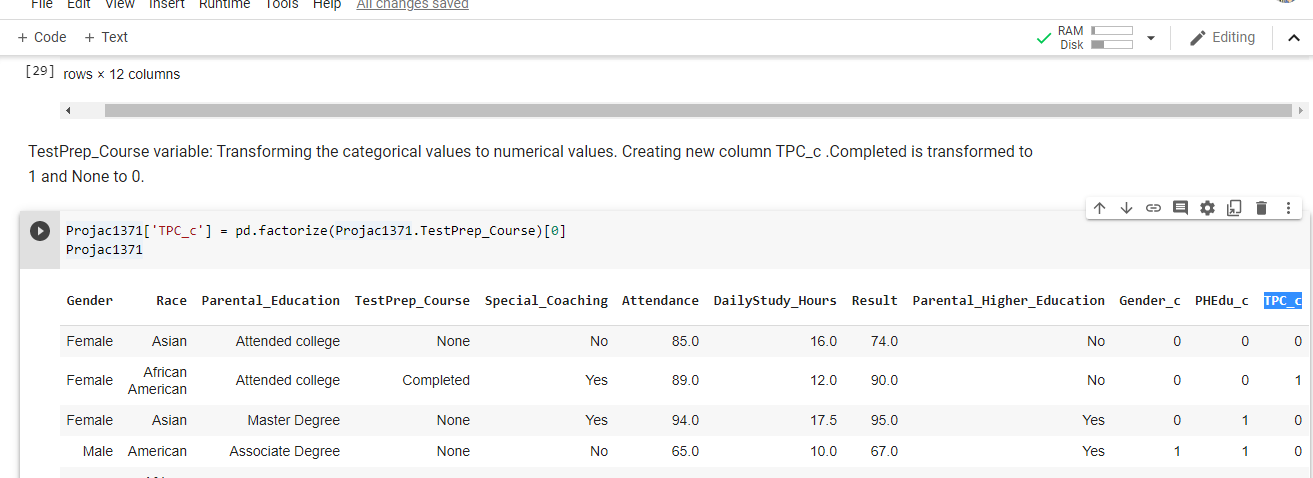
**Next, I am transforming the categorical values for the ‘Parent\_Higher\_Education’ variable to numerical values. I created new column PHEdu\_c. Yes is transformed to 1 and No to 0.**



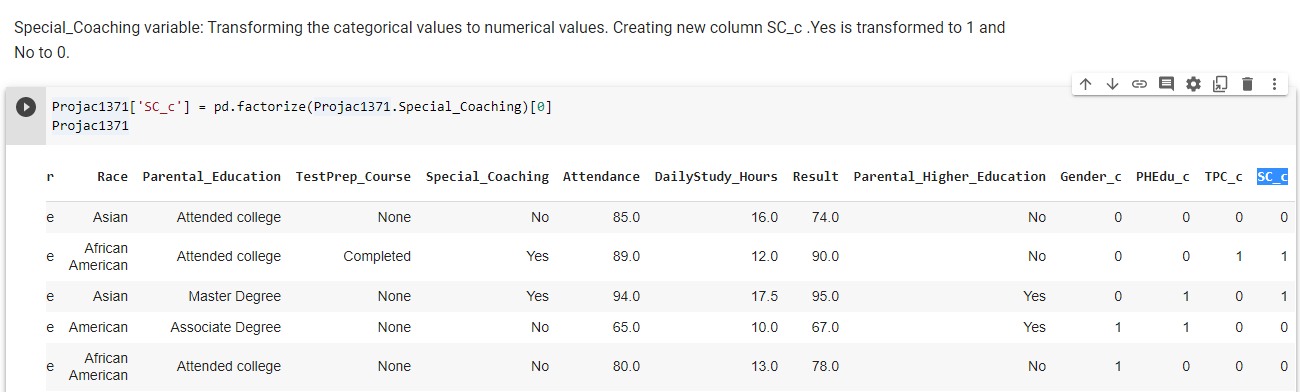
**Now I am working on the Gender variable. I am transforming the categorical values to numerical values. I created new column Gender\_c. Male is transformed to 1 and Female to 0.**



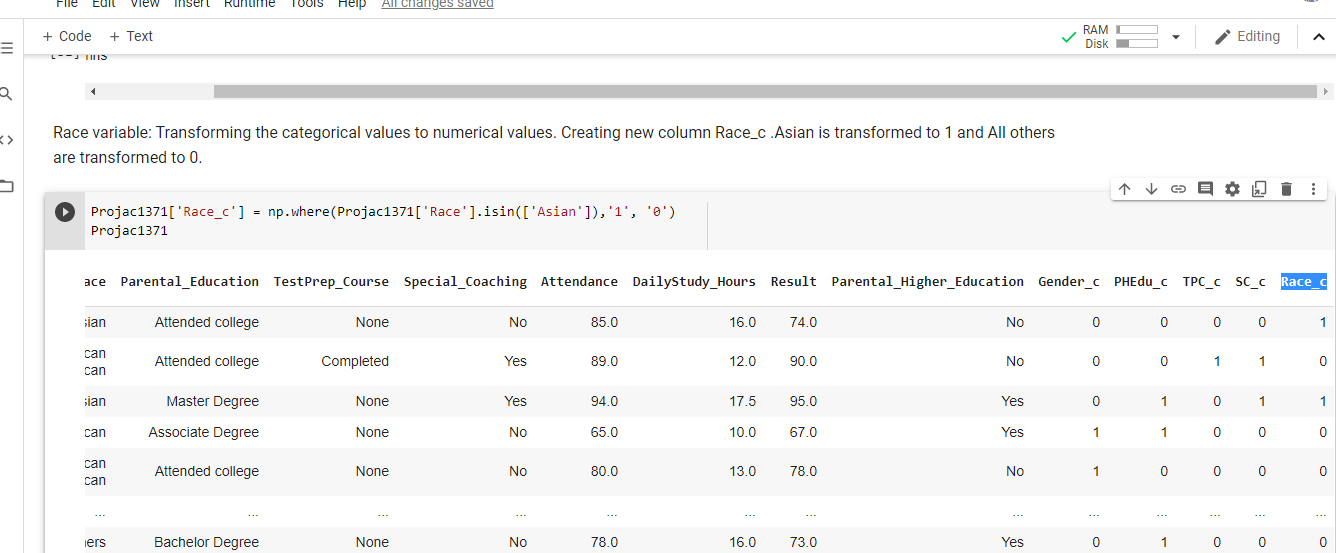
**Now I am working on the TestPrep\_Course variable. I am transforming the categorical values to numerical values. I created new column TPC\_c. Completed is transformed to 1 and None to 0.**



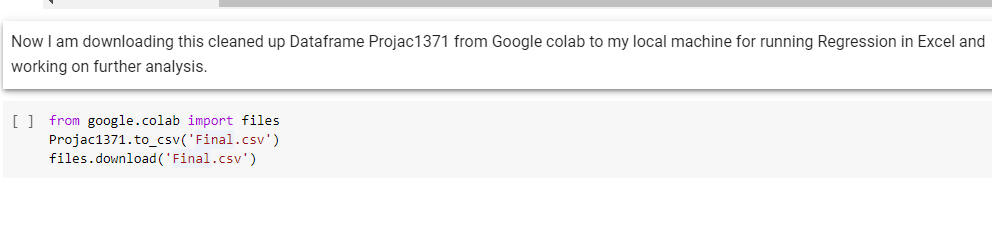
**Now I am working on the Special\_Coaching variable. I am transforming the categorical values to numerical values. I created new column SC\_c. Yes is transformed to 1 and No to 0.**



**Now** **I am working on the Race variable. I am transforming the categorical values to numerical values. Since I am analyzing the ‘Asian’ category of the Race variable baselining all the other 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.**



**Next, I have downloaded this cleaned up Data frame Projac1371 from Google colab to my local machine (named the file ‘Final.csv’) for running Regression in Excel and working on further analysis.**



**Next, I have run a preliminary regression model with all the variables in Excel.** **The various Independent variables are:**

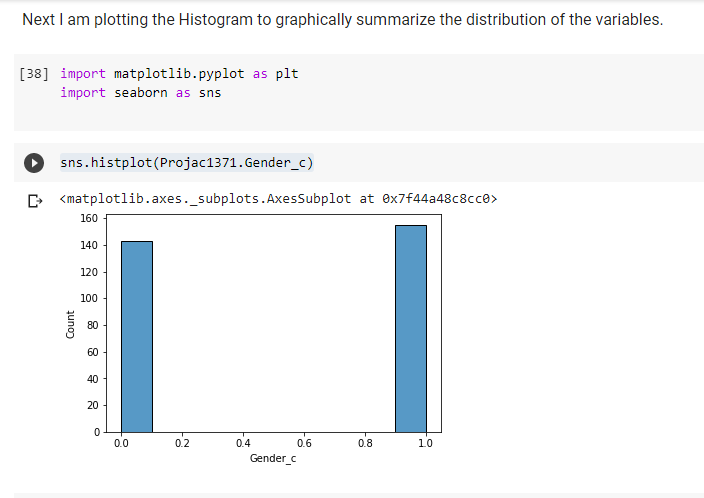
**Gender\_c, Race\_c, PHEdu\_c, TPC\_c, SC\_c, Attendance and DailyStudy\_Hours**

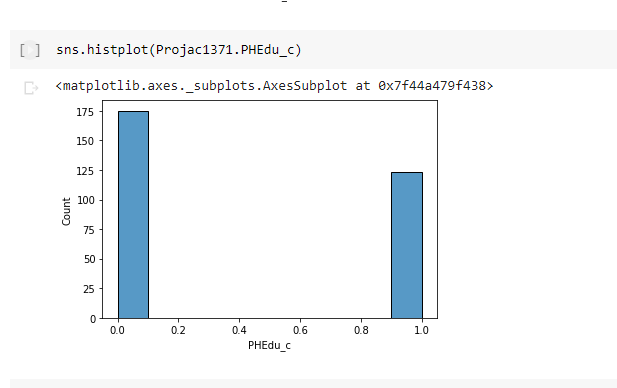
**And the Dependent variable is Result.**

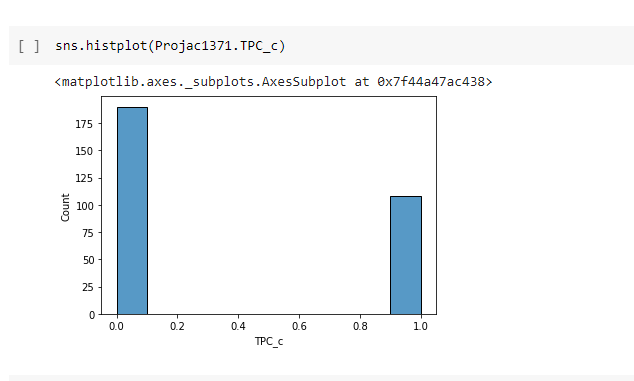
**After running the preliminary regression model and paying close attention to the statistically insignificant variables, practically and statistically significant variables and practically insignificant but statistically significant variables, I have reduced the set of independent variables to Gender\_c, PHEdu\_c, TPC\_c and Attendance. The output or dependent variable is Result.**

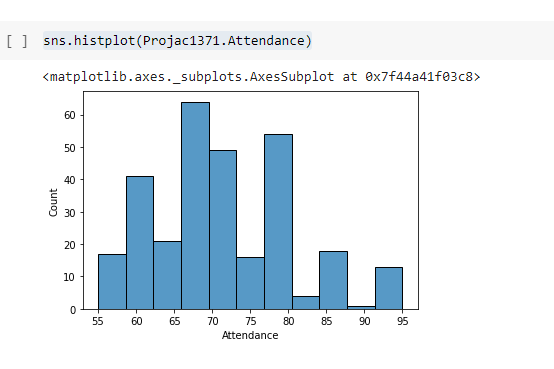
**After having the reduced set of variables with me now, I have looked at the histograms of all variables (including the dependent variable).**

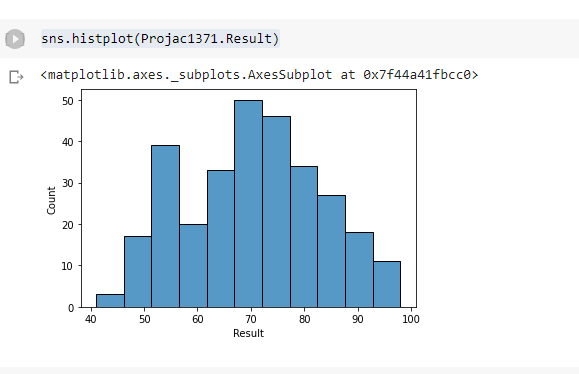
**The snapshots are given below.**



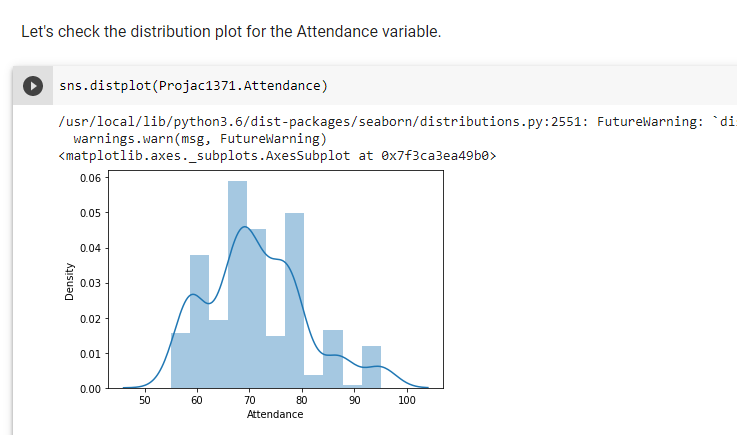




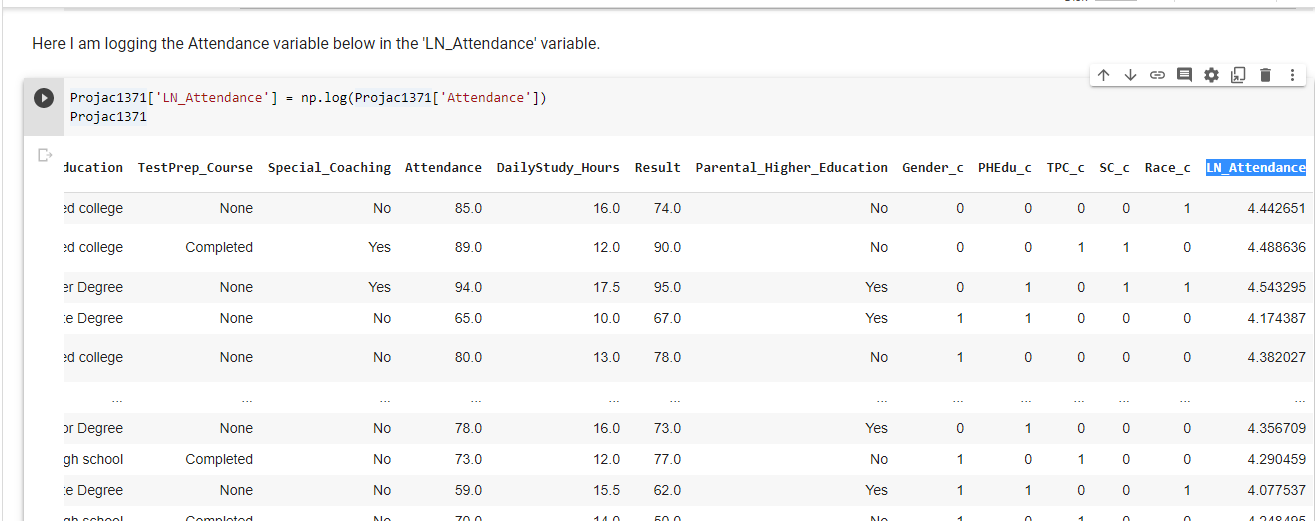




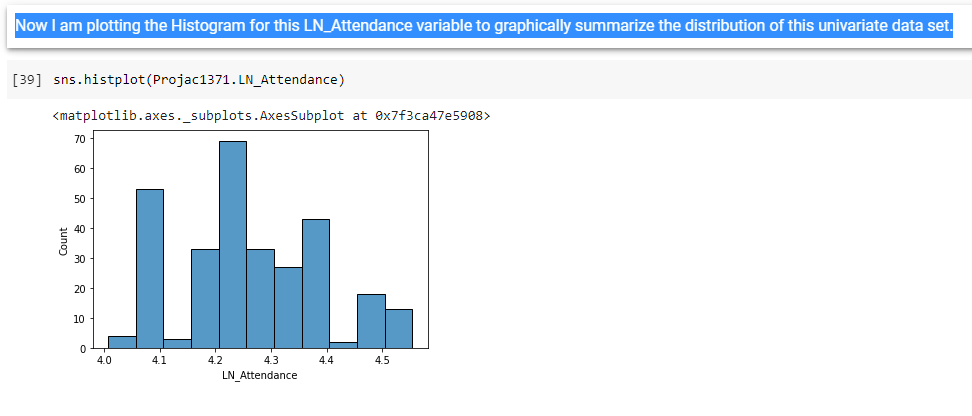
**Looking at the Attendance histogram, I thought of creating the distribution plot for this variable.**



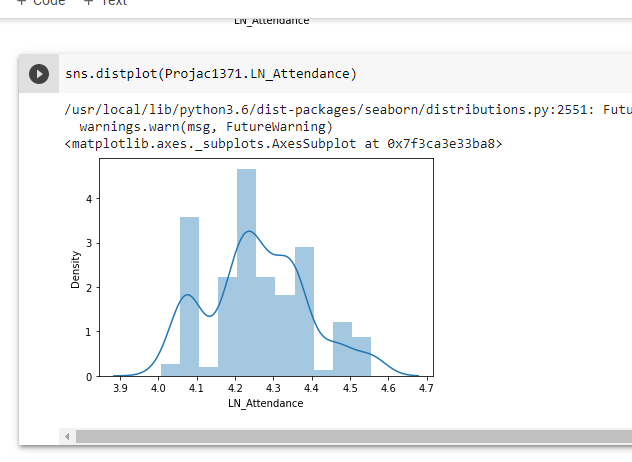
**The plot seemed to be skewed towards the right side and hence I had to Log the Attendance variable.**



**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.**



**Also, I have looked at the scatter plots of these variables taking each independent variable (including LN\_Attendance) with “Result” at a time).**

**Next, I have created three different Regression models in Excel and carried on further analysis over there for comparing those three models.**

-------------------------------- xxxxxxxxx------------------------------------