COMMENT - This is a detailed multiple regression analysis project from start to end using Microsoft Excel. The dataset that is in use here is obtained from kaggle. The link is https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression .

Multiple Linear Analysis is still very much relevant because of its high interpretability. Using MLR model, predictions can be made and it also helps to understand the data better.

About The data set, the data set has categorical and numerical variables. We are trying to build a model to predict what is the performance index.

Let us see the data first.

- **Step 1** - Let us see first the distribution of the performance index data by putting a whisker-box plot. It is seen that in the distribution here, between Q1 and Q3 , that is between the 50 percent of the data lies between 40 and 71. The dispersion is narrow, it is fairly a normally distributed data. The highest and lowest values are 100 and 10 respectively.
  Now let us try to see, how the distribution of the data varies across different categories. Let us take the extracurricular activity column and create the same plot. The distribution is well distributed normally here also.
  Let us see the distribution of the categorical variable, the extracurricular activity column by putting a bar chart.

- **Step 2 –** Now let us try to find out the relationship between the variables. The concept of *Multicollinearity* is used here. Multicollinearity occurs when two or more independent variables are highly correlated with each other, so the model can not tell which one is affecting the independent variable. If it exists, then the model might not be as reliable. Then we would have to take certain steps to correct the situation.
  To do this, first take the categorical variable to one side and all the numerical variables to another. The multicollinearity test suggested that there is minimal correlation between the independent variables. This is good sign.

- **Step -3 –** Creating the first Model is the next target. From the data analysis option, select regression and **put the variable that you want to predict on the y axis .** In this case it is the Performance index. X variables are the variables that are going to be used to predict the Y variable, that is the performance index in this case.

  Explanation of the Initial model – 1) *Multiple R* shows that there is a linear relationship between the dependent and independent variables. This is desirable.

  2*) R square* shows the proportion of variance in the dependent variable that is explained by your independent variables.Values range from 0 to 1. R^2 = 0.8 means 80% of the variation in the dependent variable is explained by the model. It always increases (or stays the same) when you add more predictors — even if they are useless.

  The Problem with R square : It can be misleading. More variables → higher R square → falsely better model.

  3) *Adjusted R square* shows how good the model is. Value of adjusted R square changes based on the number of predictors and sample size.

Also, adjusted R square penalizes the model for adding irrelevant variables. It can increase or decrease depending on whether new predictors actually improve the model. In this case R square is around 84, so it means the independent variables can explain 84 percent of variability of the dependent variable. This is desirable.

4) *P value* tells whether the individual predictor has a statistically significant relationship with the dependent variable. If P value is less than .05, we can reject the null hypothesis. In general, for every variable, the null hypothesis is that the coefficient does not have any statistical significance. In this case P values are less than .05 for each of the independent variables. So, the coefficients are statistically significant. This means, for every increase of 1 unit of the X variable, the performance index increases by the coefficient amount.

- **STEP – 4** – Let us see if we can further improve our model and further improve our accuracy. In this new model we are aiming to include two additional variables, extracurricular activity and Hours studies. To incorporate the categorical variable that is the Extracurricular activity, first we have to turn the YES/No data into binary 1/0 f form. Run the regression model again. Adjusted R squared increased to 98 percent. This is highly accurate and shows strong linear relationship. The important observation , maybe the most important one is it is seen in the updated model, MODDEL – 2, every extra hour studies by the student increases the performance index by 2.85, keeping all the variables at the same level. Let us take a look at one more statistic, that is , with around 95 percent accuracy, we are able to predict that the range will be between 2.83 and 2.86. So it can definitely be said that hours studies surely has a great impact on improving performance index.

- **STEP – 5** – Predicting the performance index, use the

  model equation =
  -34.0699+0.616*Extracurricular_Yes+2.85* Hours Studied + 1.01* Previous Scores +0.48 Sleep Hours +0.19*Sample question Paper Practiced

  This eq can be used to predict future performance index if new data arrives.