A hand holding a pencil over a piece of paper

Description automatically generated

**Marwan Mohamed Hussien – 22010248**

**Emad Mohamed Fawzy – 20221461122**

**Mohamed Ahmed Mohamed – 22011600**

**Marwan Gaber Ahmed – 22010246**

**Mahmoud Safwat Abd El-Hamid – 22011592**

**Ahmed Atia Ahmed – 22010025**

**Islam Mohamed El-Said – 22010050**

**Ahmed Tarek Hassan – 222012050**

**Ahmed Sherif Abd El-Hamid – 20221321942**

**Radwa Mohamed Ahmed – 22010089**

**Abd El-Rahman Amr Abd El-Salam - 20201447533**

Regression project

Exam dataset

# **Introduction**

In this project we aim the determine the factors which affects the students’ performance during exams.

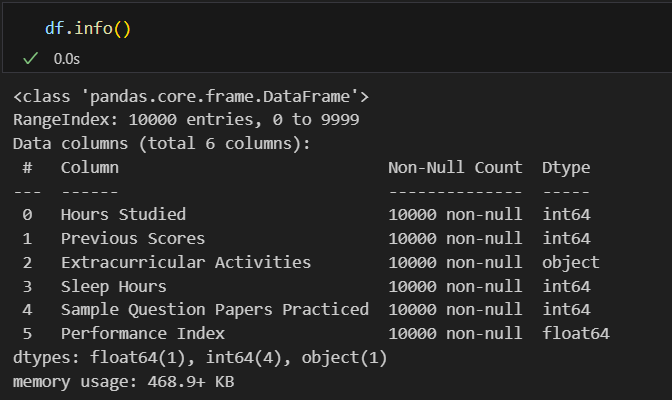
To achieve our goal, we are going to use the dataset at:

<https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression>

We are using python as our programming language of choice along with pandas for data storing , matplotlib and seaborn for data visualization , and SK-Learn for the regression process itself.

# **Data Cleaning**

We start by importing the data into a pandas data frame and then we see the make of the dataset.



We see that we have 6 columns : Hours studied, Previous Scores, Extracurricular Activities, Sleep Hours, Sample Question Papers Practiced, and Performance Index.

A screenshot of a computer

Description automatically generatedWe print the shape of the dataset and see that we have 10,000 rows and six columns.

We then check for null values of which there were none.

A screenshot of a computer program

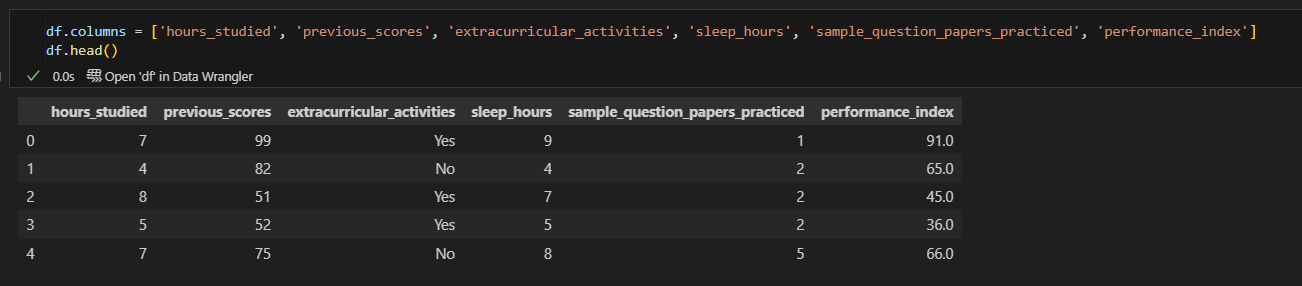
Description automatically generated

Next, we checked for duplicate values which there were 127.

A screen shot of a computer

Description automatically generated

We removed the null values and then the data were reduced to 9873 entries.

After that, we rename the columns to remove the spaces.

A screenshot of a computer program

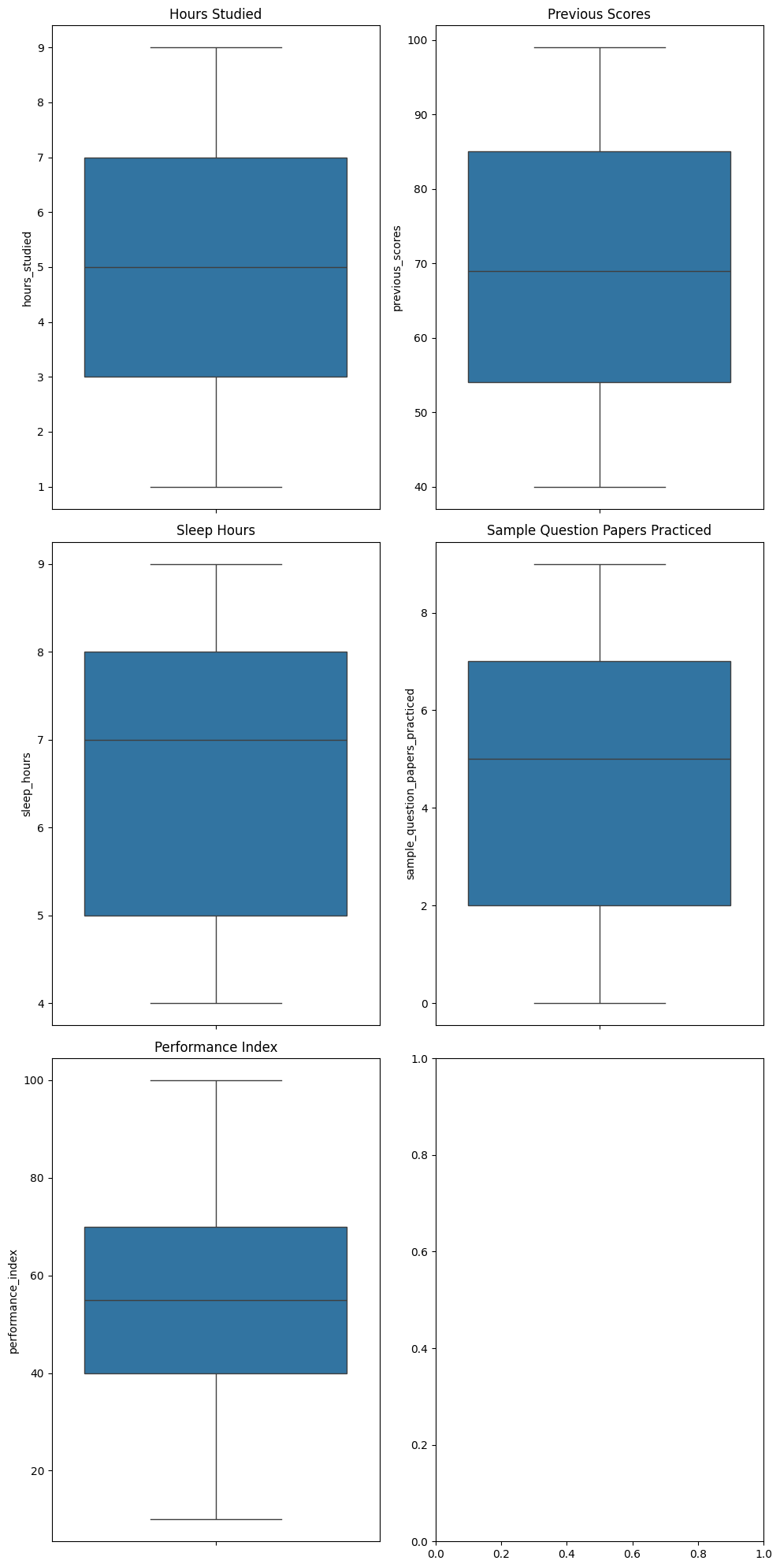
Description automatically generatedWe also edit the extracurricular\_activities column to numeric values instead of yes or no.

Then we check for zero or below zero values.

A screenshot of a computer program

Description automatically generatedThe columns with zero values are normal to have them.

Lastly, we check for outliers with box plots.

The box plots show no outlier values, so we are good to move ahead with the regression.

# **Feature selection**

In this section we are selecting the features that affect the outcome of the performance of the students.

The first thing is to draw a scatter of the variables and visualize their relation to the target.A graph of different activities

Description automatically generated with medium confidence

The plot shows a big correlation with the previous scores specially making it the dominant variable with the others contributing to smaller amounts.

We have two methods to choose our features.

# **Correlation**

To visualize the correlation between our variables we will draw a correlation matrix. A screenshot of a computer screen

Description automatically generated

We can summarize from the matrix that the hours studied and the previous scores have the most impact on the outcome and that most of the independent variables are not correlated with one another.

We will choose to remove any variables with correlation with the target of less than 0.2.A screenshot of a computer program

Description automatically generated

This leaves us with previous scores and hours studied.

# **Anova Table**

We start by constructing the table to find the p-values of each variable.A screenshot of a computer screen

Description automatically generated

As we can see all the variables have very low p-values which indicates that all of them are needed for the regression.

Taking a 95% confidence level we get that all the variables are needed.A screen shot of a computer program

Description automatically generated

Now that we have two different outcomes we will do the model with both and then decide which one is better.

# **Models**

The first model we tried was without removing any features and the results we got was as follows:

The Mean Squared Error is 4.305900938538476.

The R2 score is 0.9884301209927054.

A diagram of blue dots

Description automatically generatedA graph with blue lines and white text

Description automatically generatedThe adjusted R2 score is 0.9884007409038094.

Which indicates a very well fitted model.

The second model we tried we only kept the hours studied and the previous scores.

The results were :

The Mean Squared Error is 5.573764139536915.

The R2 score is 0.9850233951895029.

A diagram of blue dots

Description automatically generatedA graph with blue dots and a red line

Description automatically generatedThe adjusted R2 score is 0.9850082059351312.

Which does not vary a lot from the previous model.

# **Conclusion**

After trying both models we reached the following points:

1. The two most significant variables were hours studied and previous scores as the model without them got an R2 adjusted with a difference of just 0.003.
2. The relation with the variables was very linear as the model was very well fitted.
3. The correlation between the previous scores and higher performance is because if a student is already excellent, he is more likely to stay so and vice versa.
4. While some variables like hours slept did not affect the performance in the regression, it does not mean that they do not have an outright impact on the longer term performance or other health reasons.