Title: Correlation

Introduction

In this assignment, I will try to find any kind of correlation between two values. As the main dataset I choose the education_7.csv file. From this dataset I selected only major subject Engineering and sorted data by TestScore.

Methods

The main method of finding correlation is *Pearson correlation coefficient,* ranging from -1 to 1. Analyzing correlation can reveal potential linear relationships that influence model performance. So if our calculation shows near +1 number so that means perfect positive correlation and if near to -1 this will be perfect negative correlation. the number about 0 seems to be no correlation.

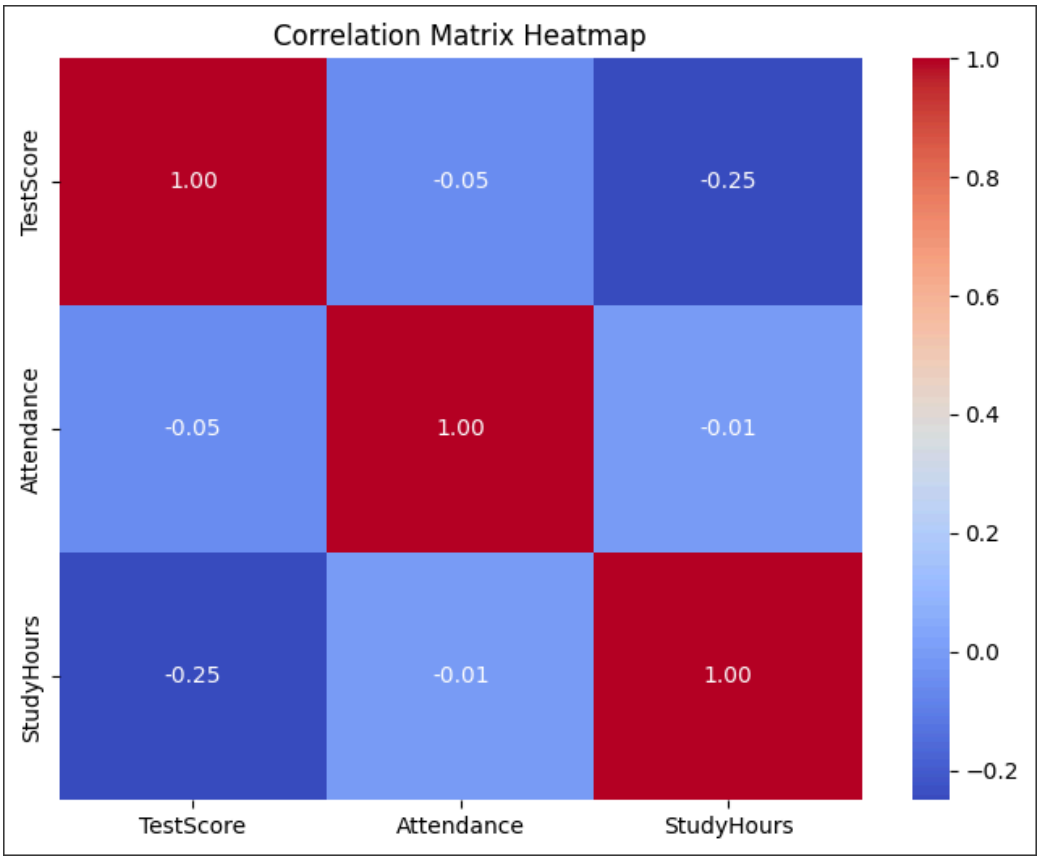
Findings

Firstly Lets see to head of our sorted data

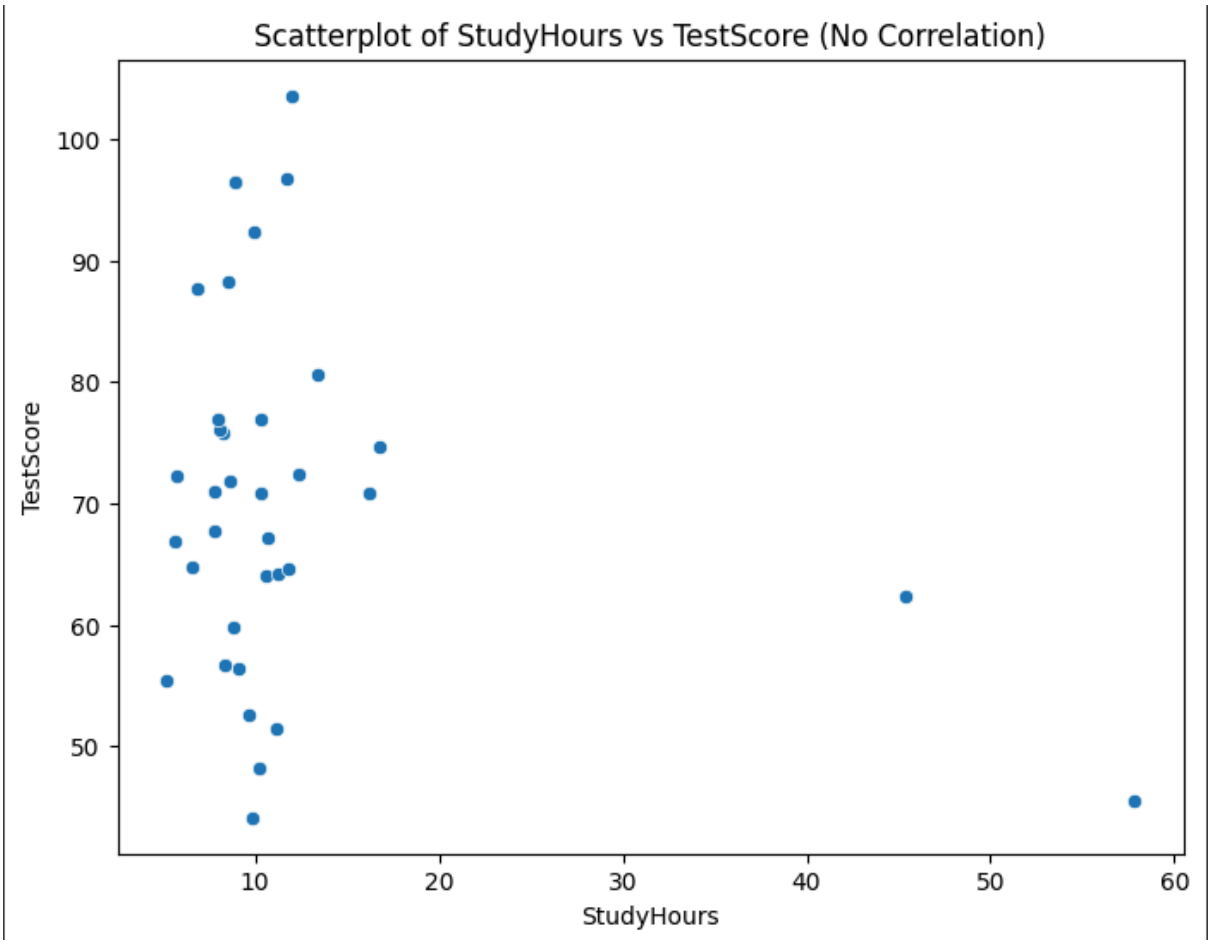
index	TestScore	Attendance	StudyHours
63	44.10576096502484	90.43775620692595	9.84229811119136
66	45.54702479550933	84.99892326305218	57.85836535751308
25	48.18451488101853	89.25182729836165	10.189785982600997
73	51.47761269469521	86.74487203349926	11.17027996806378
80	52.52275238824966	88.23003044373257	9.668378028302582

From this data we can find that the data is scattered and there seems to be no correlation but let's try to find it.

We can also try to create heatmap of this three features



From this it is clear that correlation between StudyHours and TestScore is better than other ones. Lets took these two values, as x-axis StudyHours and as y-axis TestScore and look to Scatterplot



So, as I said before there is no any kind of correlation.

Conclusion

Summarizing all my actions in this assignment can be said that to be any kind of correlation x and y axes should go down or up parallel or x and y axes go inversely to each other. But looking at our scatterplot there is clear that as y axis go up the x axis varies only between 3 and 20. So it is kind of a straight line not diagonal.