MACHINE LEARNING ASSIGNMENT 1

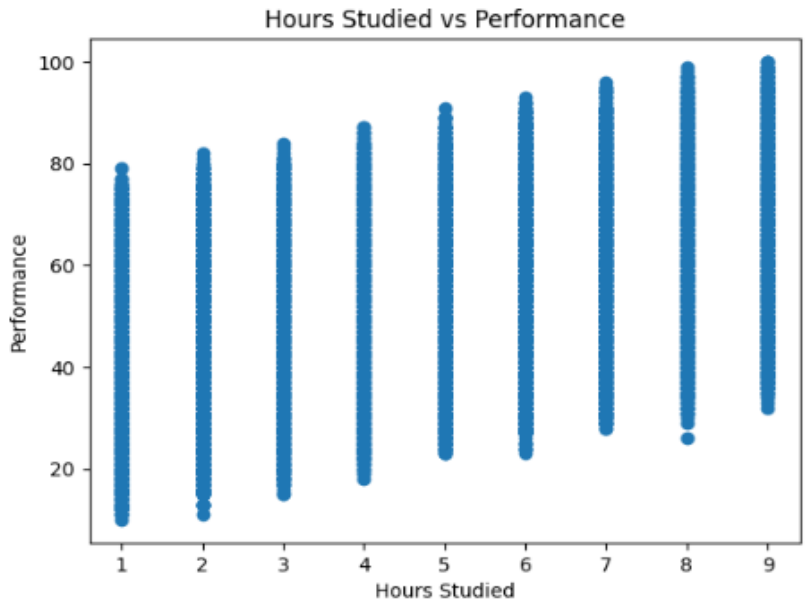
1.1 Here we need to find the factors influencing academic students performance. So here the dependent variable is **Performance** .And the five features are independent variables. Among the five **'Hours Studied'** and **'Previous scores',** correlation coefficient values of 0.37 and 0.92, respectively, were found with performance. So both are more influential for performance.

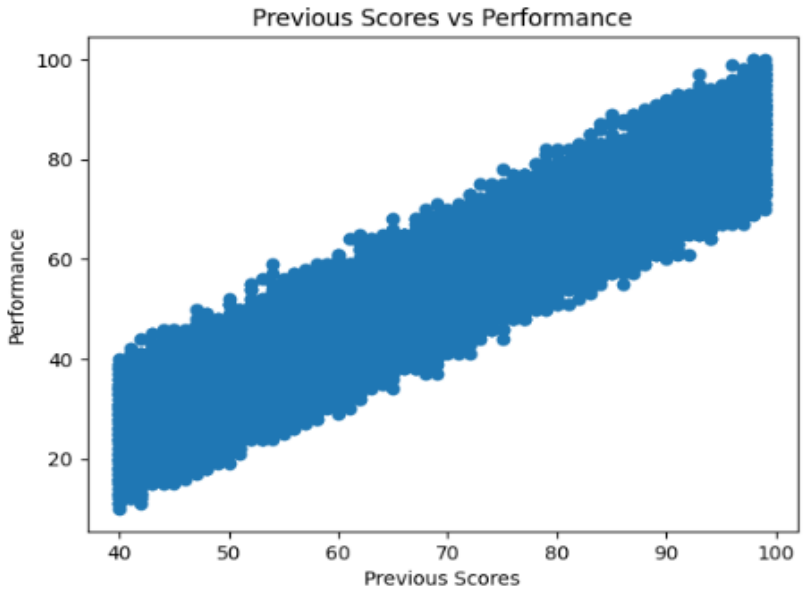
Link: Task1.pynb

1.2 There is no missing data . For convenience, i have encoded the Extracurricular Activities attributes to binary as 1 for 'Yes' and 0 for 'No'. On analyzing performance with other attributes, below are some scatter plots and central tendency table.

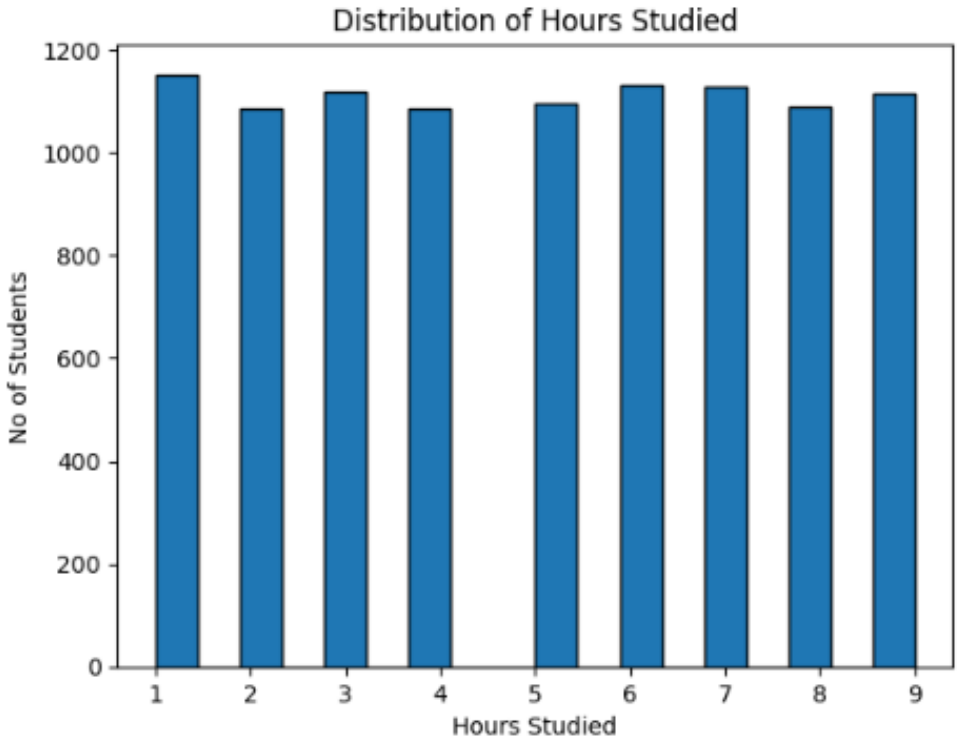
Link: Task2.ipynb

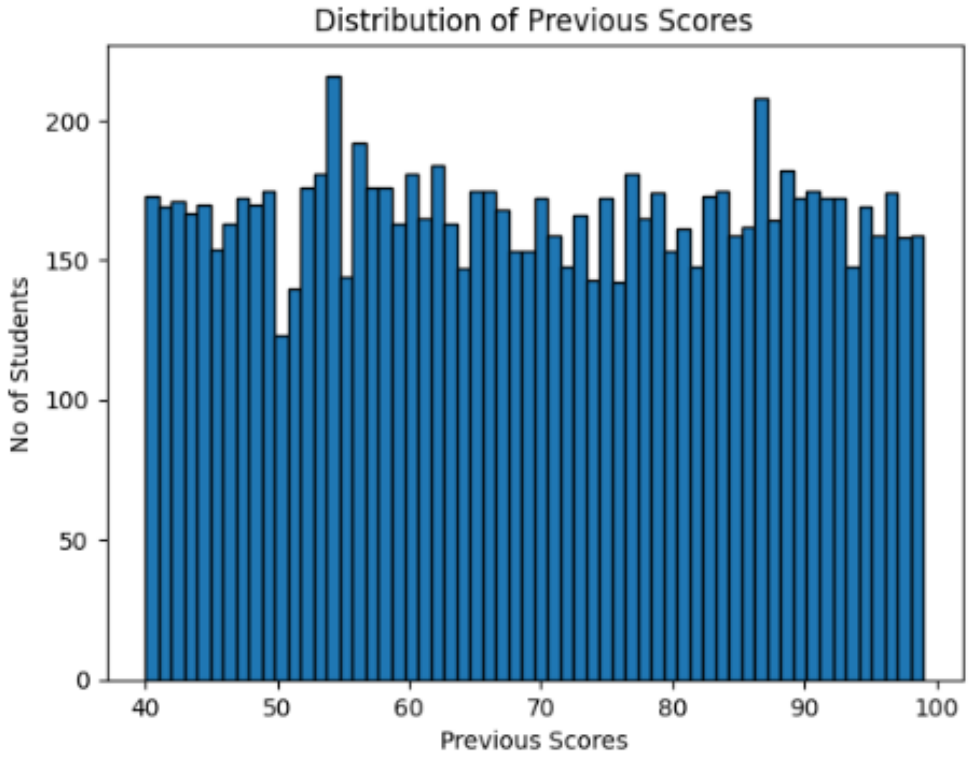
	Hours Studied	Previous Scores	Extracurricular Activities	Duration of Sleep	Sample Question Papers Practiced	Performance
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	4.992900	69.445700	0.494800	6.530600	4.583300	55.224800
std	2.589309	17.343152	0.499998	1.695863	2.867348	19.212558
min	1.000000	40.000000	0.000000	4.000000	0.000000	10.000000
25%	3.000000	54.000000	0.000000	5.000000	2.000000	40.000000
50%	5.000000	69.000000	0.000000	7.000000	5.000000	55.000000
75%	7.000000	85.000000	1.000000	8.000000	7.000000	71.000000
max	9.000000	99.000000	1.000000	9.000000	9.000000	100.000000

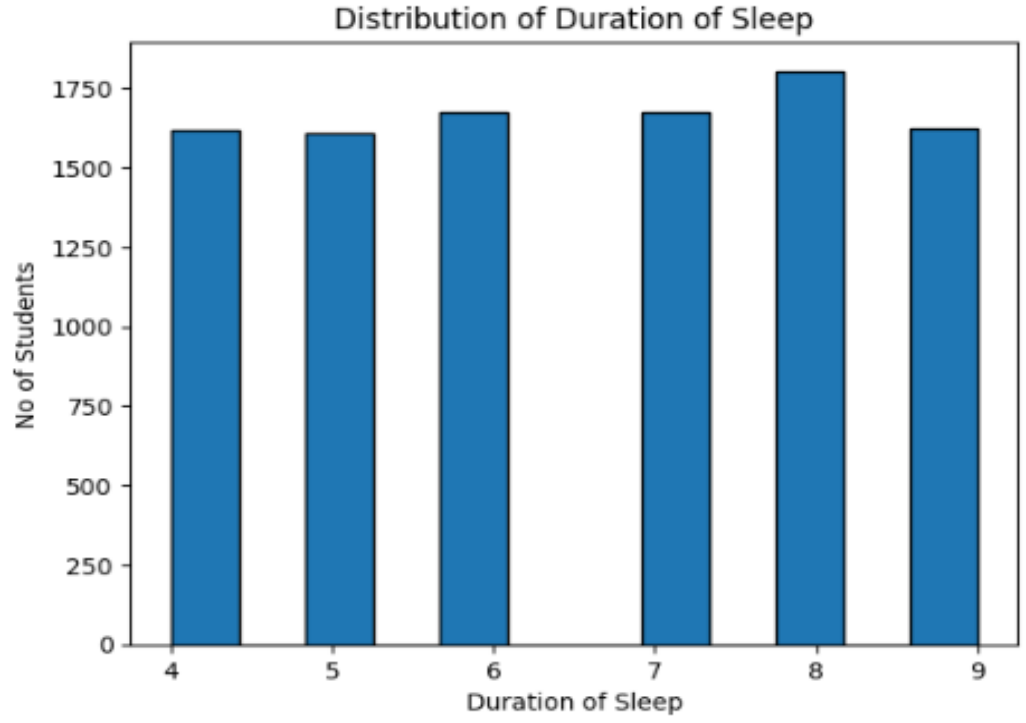


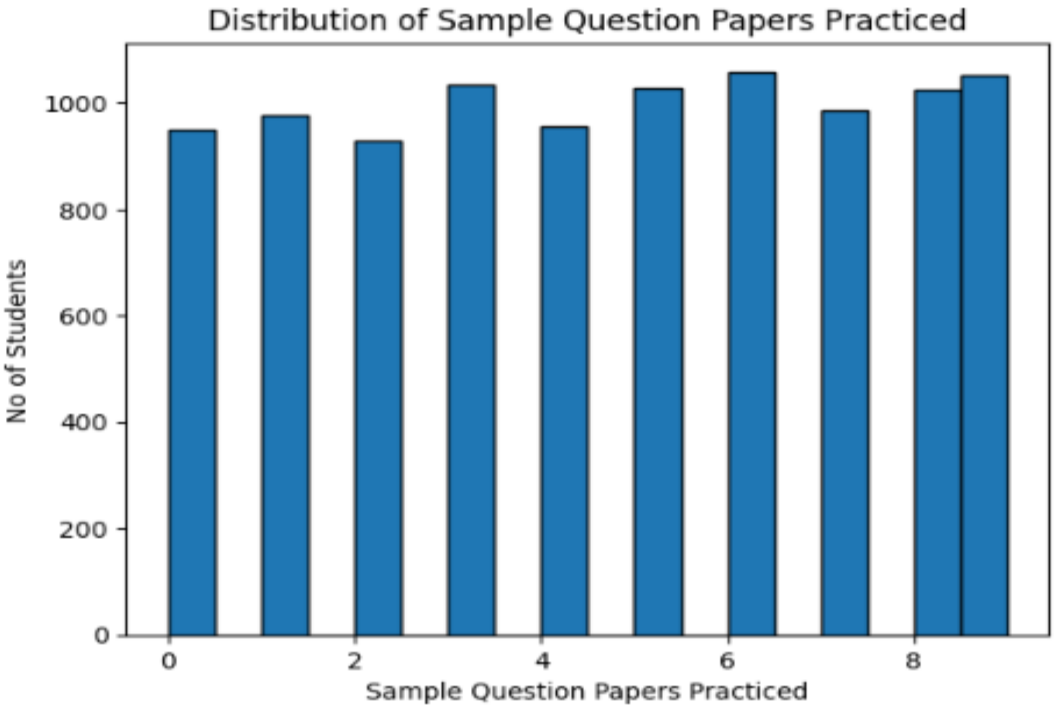


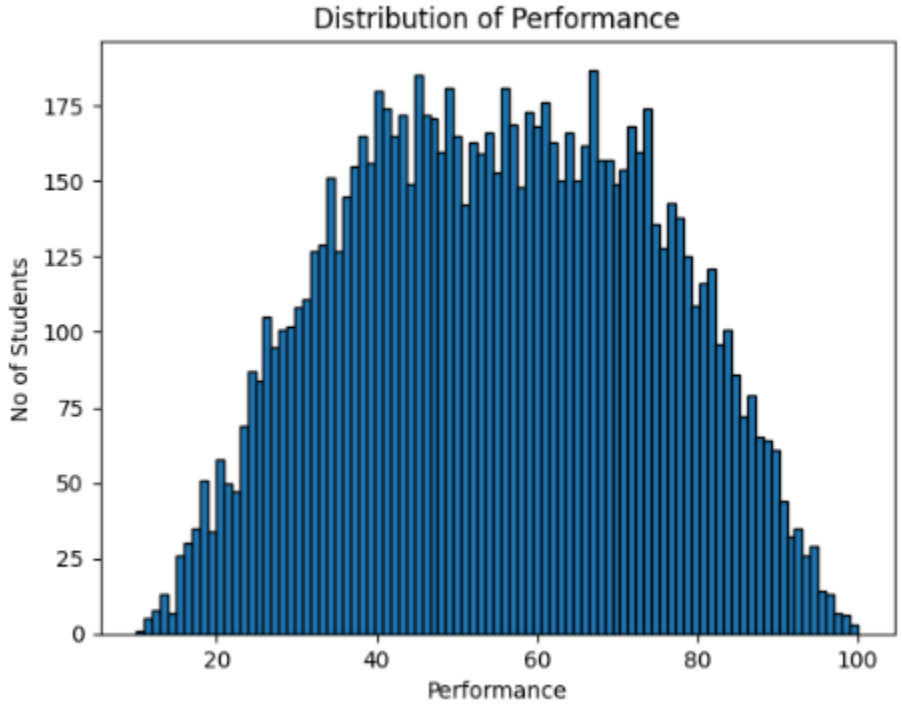
Below are some frequency distributions for independent attributes.




Distribution of Previous Scores
No of Students
Previous Scores
0
50
100
150
200
40
50
60
70
80
90
100


Distribution of Duration of Sleep
No of Students
Duration of Sleep
0
250
500
750
1000
1250
1500
1750
4
5
6
7
8
9





1.3 Dataset is split for training and testing in 80:20 using random shuffling

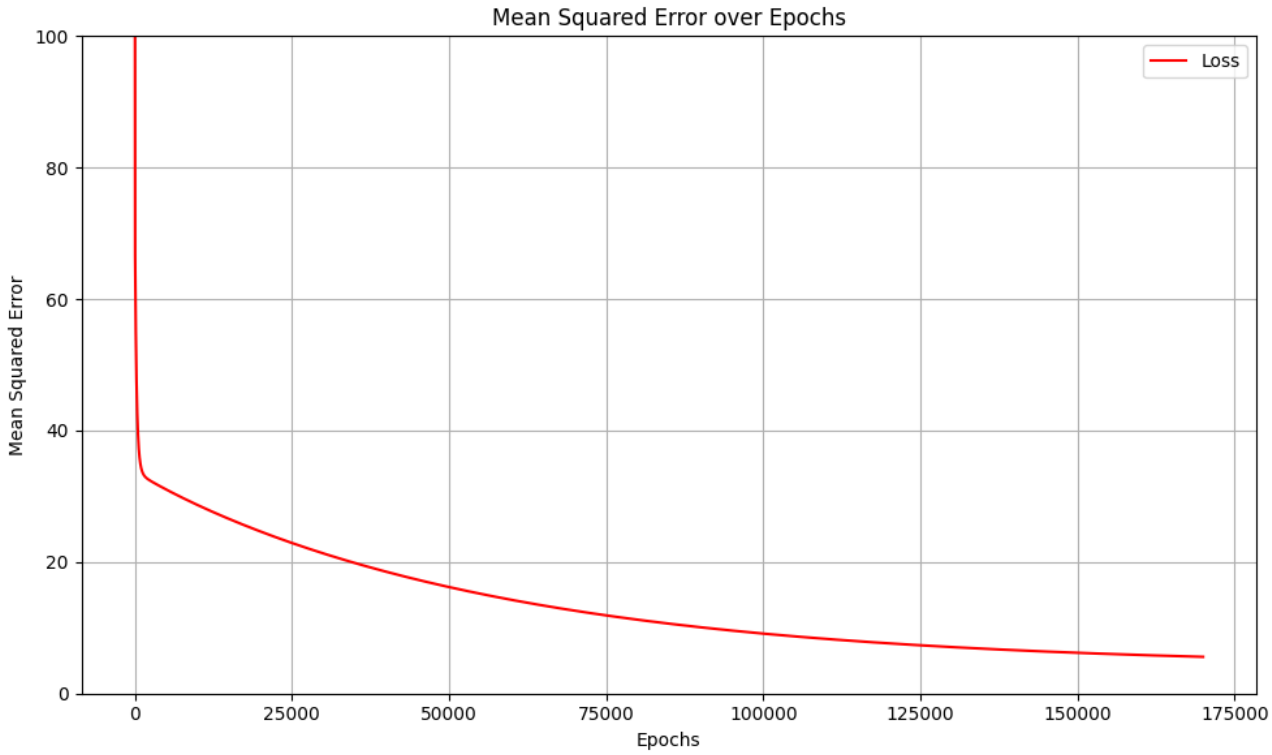
Link: **Task3.ipynb**

1.4

Link: Task4.ipynb

1.5

Link: Task5.ipynb



1.6 On the basis of other given attributes, student performance is 85.99

Link : Task6.ipynb

1.7 Mean square error is 5.88, while the R2 Score is 0.98. The low MSE and high R2 score indicate that the model's predictions are accurate.

Link: Task7.ipynb

2.1 Comparison between regression library sklearn and numpy

Library	Mean Squared Error	Executation time(second)
Pandas and Numpy	5.88	20.5
sklearn	4.09	0.39

So improvement can be achieved as

- By implementing feature scaling to bring all features to a similar scale.
- Implementation of regularization techniques like L1 or L2 can improve the generalization of the model.
- Can use other regression models like Random forest regression, and Vector regression to see if better performance can be achieved.

Link: **Task8.ipynb**

Task4.ipynb

Note- Each task can be executed separately. Upload the dataset from the local device before execution.