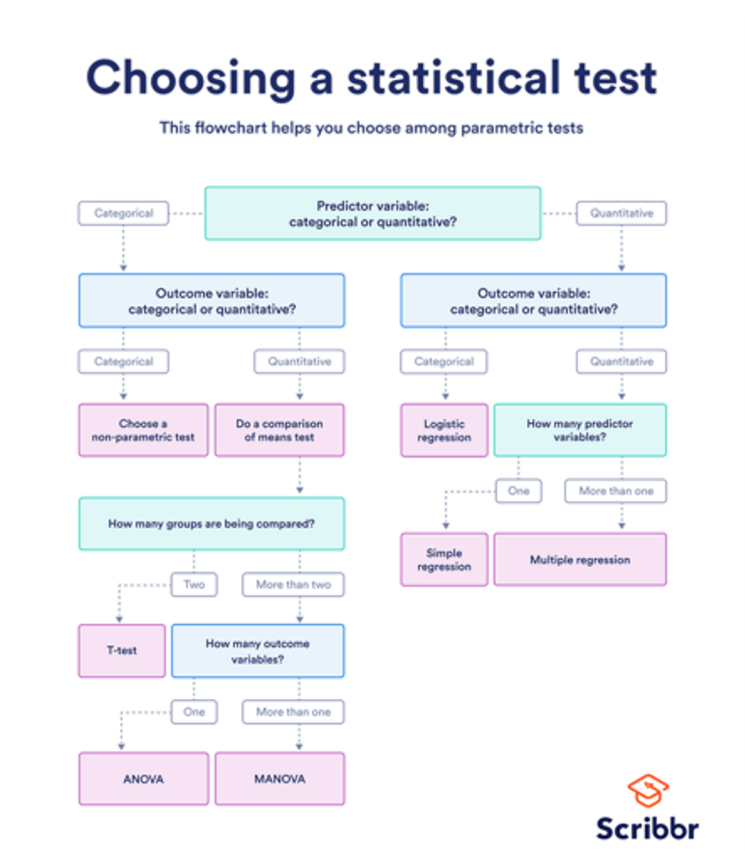
Sleep Health and Lifestyle Dataset & Student Performance Factors Dataset

Mobina Kargar

1 Introduction

The purpose of this assignment is to practice basic data pre-processing skills and apply different statistical tests. Two datasets given: the Sleep Health and Lifestyle Dataset and the Student Performance Factors dataset.Before I start to code and know what should I do I search this and get info to how to write it's statistical tests.



2 Features(Sleep Health and Lifestyle Dataset)

The data is composed of 31 columns and 374 entries (Full train dataset shape is (374, 31)). We can see all 31 dimensions of our dataset by printing out the first 3 entries:

Table 1: train dataset (3 rows x 31 columns)

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	...	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
0	1	Male	27	Software Engineer	6.1	6	...	6	Overweight	126/83	77	4200	-
1	2	Male	28	Doctor	6.2	6	...	8	Normal	125/80	75	10000	-
2	3	Male	28	Doctor	6.2	6	...	8	Normal	125/80	75	10000	Sleep Apnea

Table 2: Data columns & Their Types (total 31 columns):

```
Person ID                 int64
Gender                   object
Age                       int64
Occupation               object
Sleep Duration          float64
Quality of Sleep          int64
Physical Activity Level   int64
Stress Level              int64
BMI Category             object
Blood Pressure           object
Heart Rate                int64
Daily Steps               int64
Sleep Disorder           object
```

There are many choices to remove missing or null values we can use dropna() to remove all missing and invalid values, or we can use fillna() to replace them. However, in this step, we need to understand the data types of each column.

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical A
0	False	False	False	False	False	False	
1	False	False	False	False	False	False	
2	False	False	False	False	False	False	
3	False	False	False	False	False	False	
4	False	False	False	False	False	False	
5	False	False	False	False	False	False	
6	False	False	False	False	False	False	
7	False	False	False	False	False	False	
8	False	False	False	False	False	False	
9	False	False	False	False	False	False	

374 rows x 13 cols 10 per page Page 1 of 38

3 Distribiution

In this part we are checking the distribution to make sure the data is normaly distributed if it's p value less than 0.05 it Reject null hypothesis and said The data is not normally distributed an if it is upper or equal to 0.05 then it said that it Fail to reject null hypothesis and it means The data is normally distributed.

3. 1 Sleep Duration Distribution

look at how the Sleep Duration are distributed.

count	374.0
mean	7.132085561497325
std	0.7956567308898189
min	5.8
25%	6.4
50%	7.2
75%	7.8
max	8.5

Name: Sleep Duration, dtype: float64

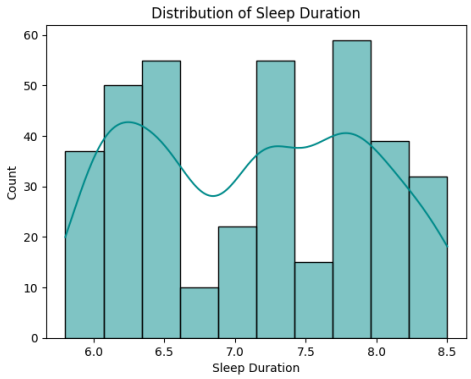


Figure 1: distributed of Sleep Duration shows this plot is not normal

3.2 Stress Level by Occupation

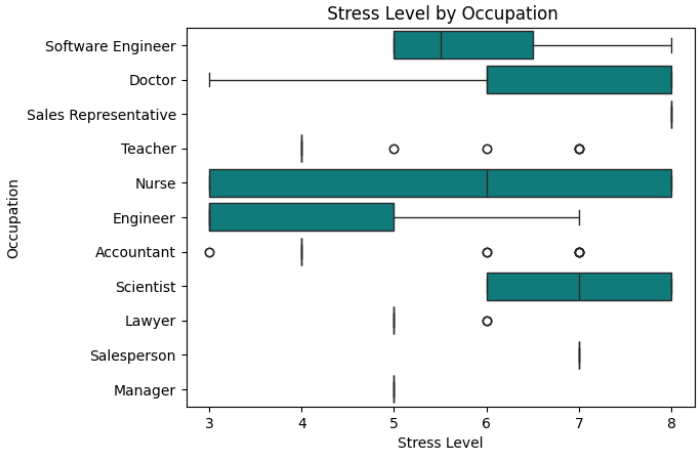


Figure 2: distributed of Sleep Duration shows this plot is not normal

3.3 Correlation for Each Column

In this part we use heatmap to see multiple numerial variables.

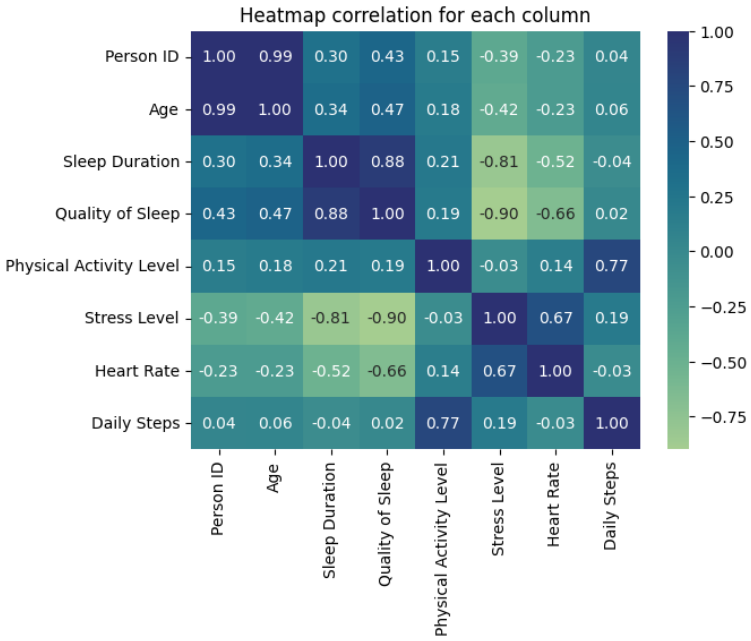


Figure 3: the upper of this plot it shows that how it is distributed

4 Exploratory Data Analysis(Sleep Health and Lifestyle Dataset)

[A] Does women's sleep duration follow a normal distribution?

No, as we can see from the plot below and by checking normality using the Shapiro-Wilk test, we find that the data is not normally distributed. We also checked which statistical test to use and the Shapiro-Wilk test is proper for testing normality. The p-value of 0.0000 shows that based on the Shapiro-Wilk test the sleep duration data for women is not normally distributed. The plot below confirm this with a skewed distribution and other signs of non-normality

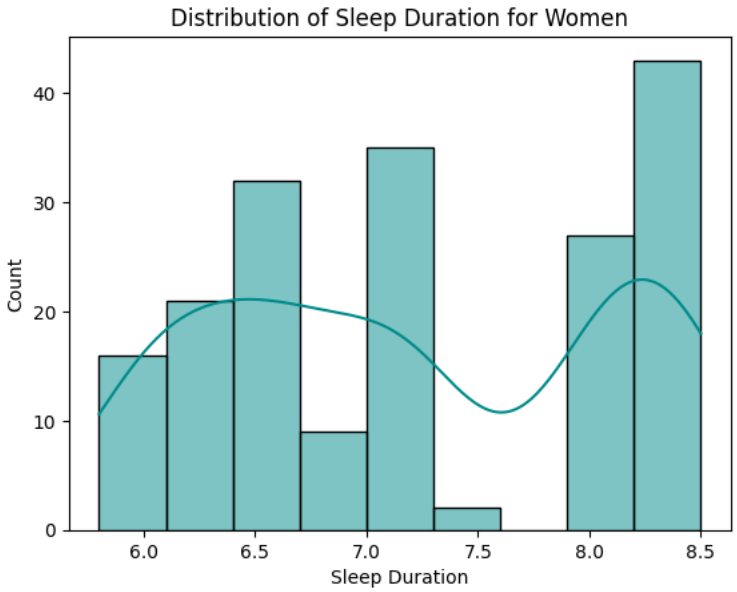


Figure 4: p-value: 0.0000 so Reject null hypothesis >> The data is not normally distributed

[B] Is having higher daily steps a contributing factor into better sleep?

	Daily Steps	Quality of Sleep
Daily Steps	1.000000	0.016791
Quality of Sleep	0.016791	1.000000

Result = 0.01679141492471579
p-value: 0.7461906652960792
Fail to reject null hypothesis

As you see the result I hardly say the answer is write because the result is suggests that daily steps have little to no effect on sleep quality and any potential connection is likely not strong enough to be considered a significant contributing factor.

[C] Is stress level different among different occupations?

H0: $\mu_1=\mu_2=\mu_3$ or The mean of the samples is the same.
H1: At least one of them is different.

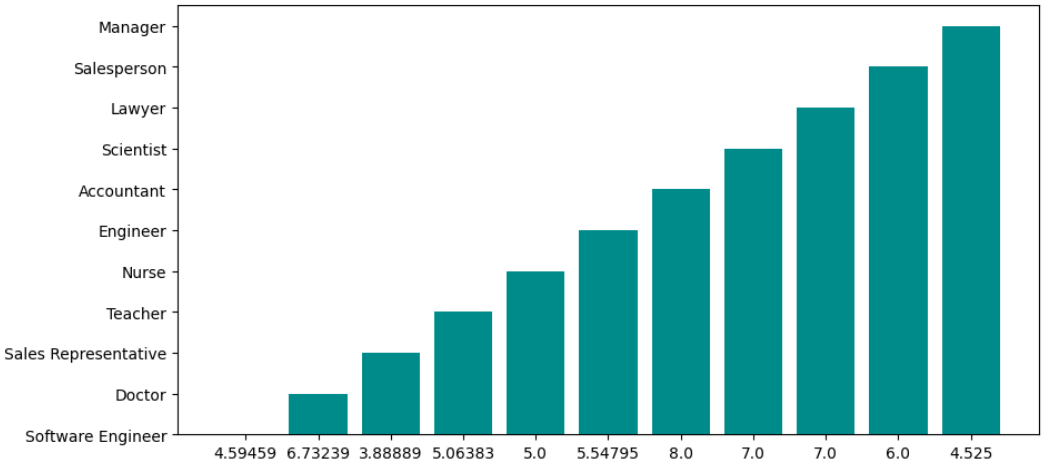
H0: The data is normally distributed.
H1: The data is not normally distributed.

H0: The variances of the samples are the same.
H1: The variances of the samples are different.
The stress levels vary across different occupation but since the data is not normally distribute further non-parametric testing would be more appropriate And Based on the tests for variance the variances in stress levels are consistent across occupations.

just for stress level::
p-value: 0.0000
Reject null hypothesis >> The data is not normally distributed

Occupation
Accountant 4.59459
Doctor 6.73239
Engineer 3.88889
Lawyer 5.06383
Manager 5.00000
Nurse 5.54795
Sales Representative 8.00000
Salesperson 7.00000
Scientist 7.00000
Software Engineer 6.00000
Teacher 4.52500
Name: Stress Level, dtype: float64

p-value: 0.7075
Fail to reject null hypothesis >> The data is normally distributed



[D] Are different BMI categories significantly different given their blood pressure?

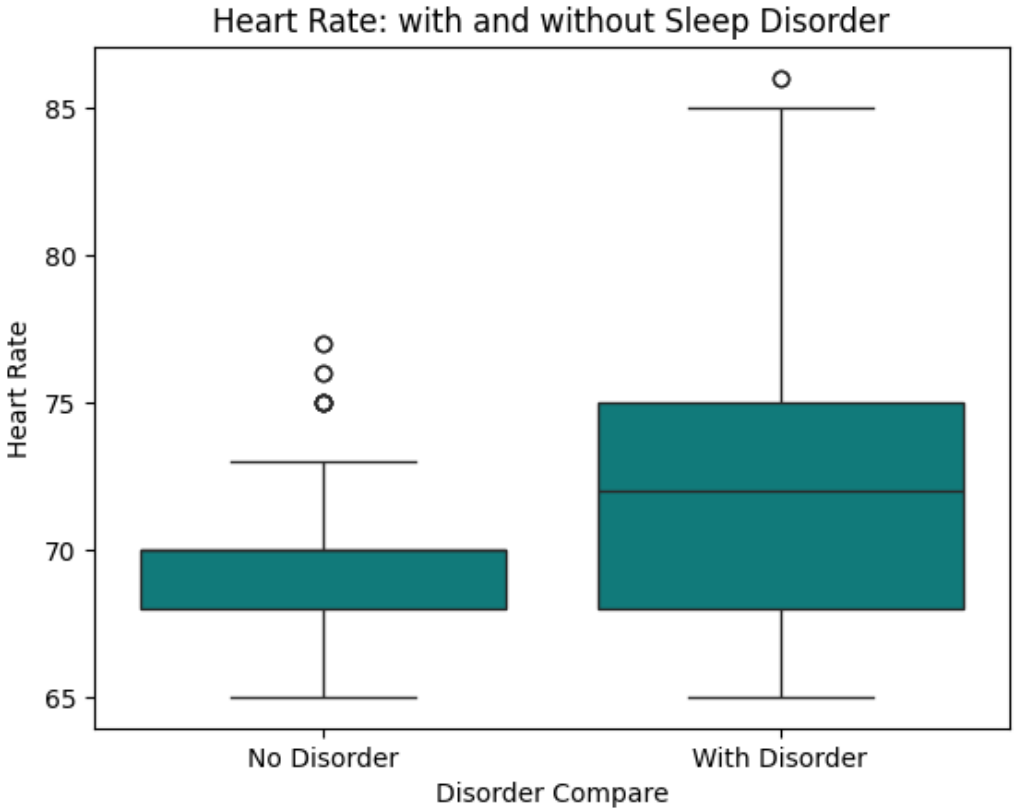
As we see the result Both tests for [Upper & Below] Blood Pressure have extremely small p-values (far less than 0.05). that we reject the null hypothesis for both tests.
And we use ANOVA test because we have more than two groups.

ANOVA for Upper Blood Pressure: F_onewayResult(statistic=163.545195439236, pvalue=1.7438768080834315e-67)
ANOVA for Below Blood Pressure: F_onewayResult(statistic=179.09700388935002, pvalue=1.0189415858614994e-71)

[E] Do people with sleep disorders have higher heart rates than those without any sleep disorder?

As we know T-statistic shows the size of the difference between the means in units of the standard error. If it's large it suggests a large difference between groups.

T-statistic = 6.10252277092434
P-value = 4.919396753345923e-09
People with sleep disorders have higher heart rates



[Bonus]

[F] Does physical activity level influence heart rate?

Mann-Whitney U Test , Tests whether the distributions of two independent samples are equal or not.

Assumptions

Observations in each sample are independent and identically distributed (iid).

Observations in each sample can be ranked.

Interpretation

H0: the distributions of both samples are equal.

H1: the distributions of both samples are not equal

Mann-Whitney U Statistic: 17895.0000

P-value: 0.4615

Fail to reject null hypothesis >> Physical activity level doesn't significantly influences heart rate

5 Features(Student Performance Factors Dataset)

The data is composed of 20 columns and 6607 entries (Full train dataset shape is (6607, 20)). We can see all 20 dimensions of our dataset by printing out the first 3 entries:

Table 3: train dataset (3 rows x 20 columns)

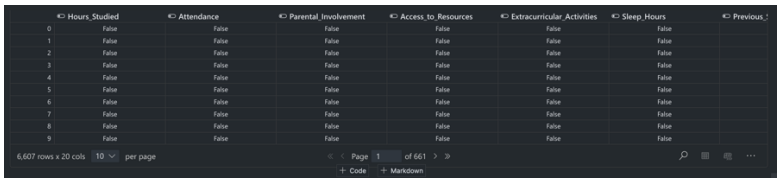
	Hours_Studied	Attendance	Parental_Involvement	Access_to_Resources	Extracurricular_Activities	...	Learning_Disabilities	BMI Category	Distance_from_Home	Gender	Exam_Score
0	23	84	Low	High	No	...	No	High School	Near	Male	67
1	19	64	Low	Medium	No	...	No	College	Moderate	Female	61
2	24	98	Medium	Medium	Yes	...	No	Postgraduate	Near	Male	74

Table 4: Data columns & Their Types (total 31 columns):

```
Hours_Studied                 int64
Attendance                    int64
Parental_Involvement         object
Access_to_Resources          object
Extracurricular_Activities   object
Sleep_Hours                   int64
Previous_Scores               int64
Motivation_Level             object
Internet_Access              object
Tutoring_Sessions             int64
Family_Income                object
Teacher_Quality              object
School_Type                  object
Peer_Influence               object
Physical_Activity             int64
Learning_Disabilities        object
Parental_Education_Level     object
Distance_from_Home           object
Gender                       object
Exam_Score                    int64
          dtype: object
```

There are many choices to remove missing or null values we can use dropna() to remove all missing and invalid values, or we can use fillna() to replace them. However, in this step, we need to understand the data types of each column.

latest >> : (6529, 20)
new >> : (6378, 20)



6 Distribiution

In this part we are checking the distribution to make sure the data is normaly distributed if it's p value less than 0.05 it Reject null hypothesis and said The data is not normally distributed an if it is upper or equal to 0.05 then it said that it Fail to reject null hypothesis and it means The data is normally distributed.

6. 1 Sleep Hours Distribution

look at how the Sleep Hours are distributed.

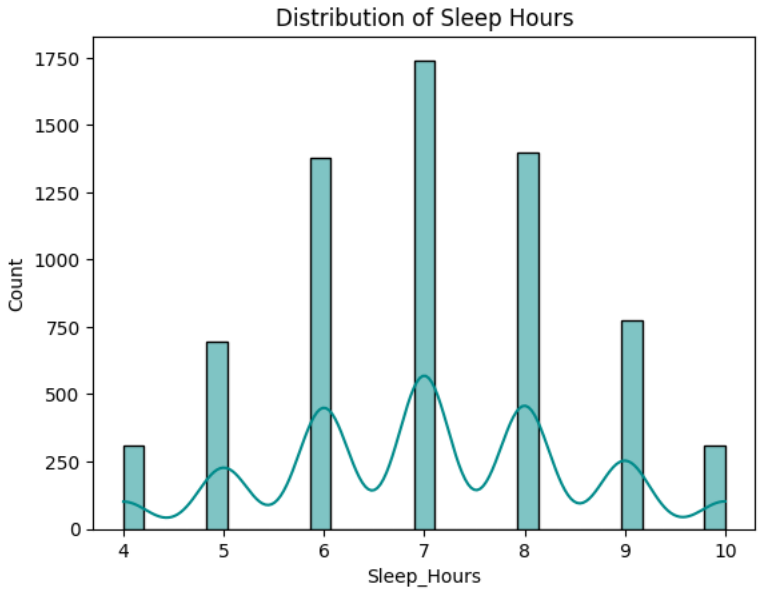
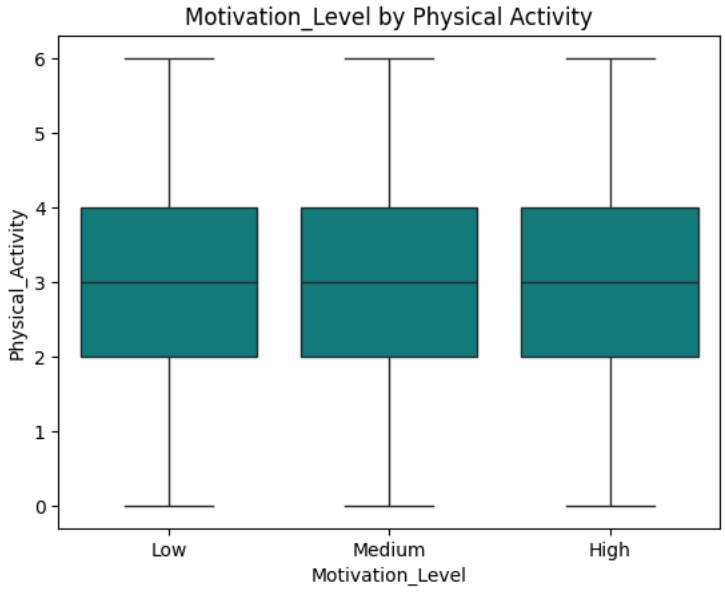


Figure 4: distributed of Sleep hours shows this plot is not normal

6.2 Motivation Level by Physical Activity



6.3 Correlation for Each Column

In this part we use heatmap to see multiple numerial variables.

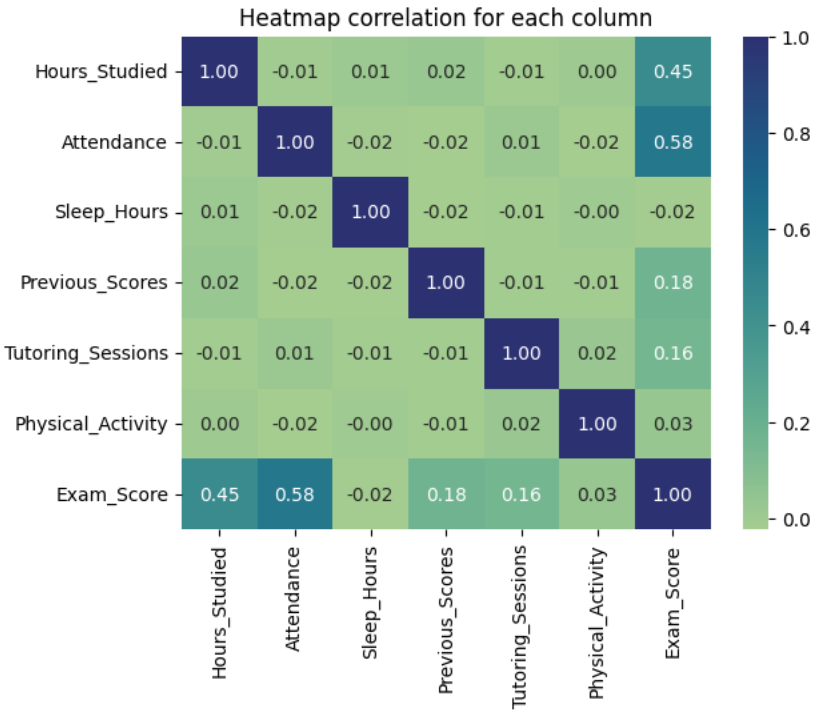
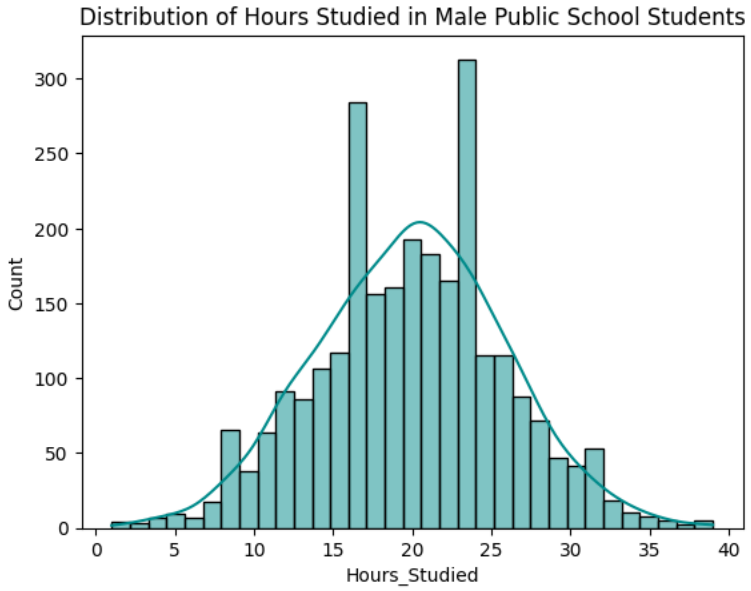


Figure 5: the upper of this plot it shows that how it is distributed

7 Exploratory Data Analysis(Student Performance Factors Dataset)

[A] Do the hours spent in public school by males follow a normal distribution?

p-value: 0.0001
Reject null hypothesis >> The data is not normally distributed

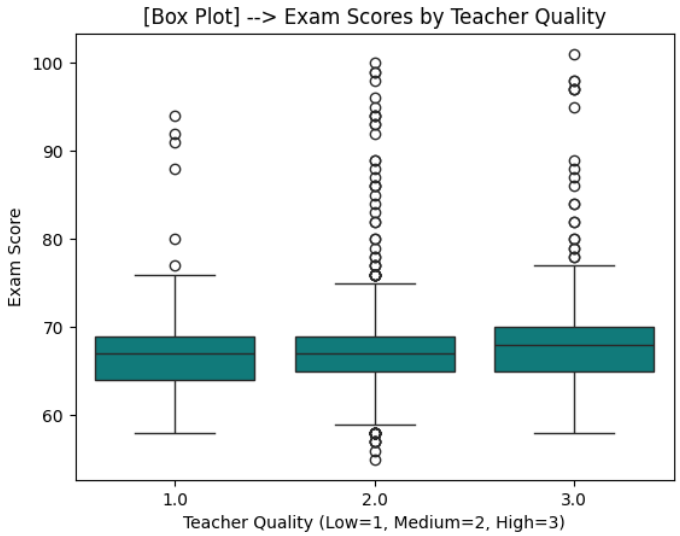


[B] Is having higher teacher quality a contributing factor to higher exam scores?

First I convert the categorical values in the Teacher_Quality column ("Low", "Medium", "High") into numer ic values (1, 2, 3) to make it easier to perform statistical analysis like correlation.
The result shown that teacher quality has little to no impact on exam scores

	Teacher_Quality	Exam_Score
Teacher_Quality	1.000000	0.076387
Exam_Score	0.076387	1.000000

Result = 0.07638720155795795
p-value: 6.405898549878319e-10
Reject null hypothesis



[C] Does the number of tutoring sessions significantly impact exam scores?

Hypothesis for Pearson's Correlation Test

H0 : (Null Hypothesis): There is no correlation between tutoring sessions and exam scores.
H1: (Alternative Hypothesis): More tutoring sessions correlate with higher exam scores.

Pearson Correlation=0.15652518539225324, p=1.6508174156341984e-37

It affects exam scores

[D] Is there a significant difference in the distribution of Peer Influence and Motivation Level?

Wilcoxon Signed-Rank Test

Tests whether the distributions of two paired samples are equal or not.

<u>Assumptions</u>

Observations in each sample are independent and identically distributed (iid).

Observations in each sample can be ranked.

Observations across each sample are paired.

<u>Interpretation</u>

H0: the distributions of both samples are equal.

H1: the distributions of both samples are not equal.

Wilcoxon Test Statistic = 2976006.000, p-value = 0.000

Reject null hypothesis >> Distributions are not equal

[E] Is there a significant difference in the distribution of Peer Influence, Motivation Level, and Stress Level across different age groups?

Tests whether the distributions of two or more paired samples are equal or not.

Assumptions

Observations in each sample are independent and identically distributed (iid).

Observations in each sample can be ranked.

Observations across each sample are paired.

Interpretation

H0: the distributions of all samples are equal.

H1: the distributions of one or more samples are not equal.

School Type: Public

Friedman Test Statistic = 2448.289542345394, p-value = 0.0

Reject null hypothesis >> Distributions are not equal

School Type: Private

Friedman Test Statistic = 1119.7454780361743, p-value = 7.085310324959533e-244

Reject null hypothesis >> Distributions are not equal