

Social Media's Impact on the Lives of Students during Covid-19

Abstract

During the Corona virus pandemic, several schools switched from in person learning to online learning. This switch greatly impacted the lives of students since they no longer had to account for the time it took for transportation to and from school along with after-school activities which lead to students generally having more time on their hands. Since students couldn't interact with each other in person, one of the main methods of communication was through social media. Students also had to decide on how they should manage their free time and it usually amounted to spending time on social media, sleeping, fitness, and self-study. This is a study conducted on a dataset containing student reviews that include information on how they spent their time during the lockdown to see the impact social media had on the way student's spent their day.

Collect and organize the data

After loading the covid_student_responses.csv file into a dataframe and printing it, I saw that there was 1182 rows and 19 columns, which meant that a total of 1182 students were in this dataset. However, I did not need all 19 columns so I dropped the columns that were irrelevant to the study and made two data frames. reviewData will be used to create a model and make predictions and the other data frame, studentLifeData, will be used to analyze any patterns, connections, or relations within the dataset.

Original Data frame of Dataset:

```
PS C:\Users\19172\Desktop> python midterm_csc221.py
   ID Region of residence Age of Subject ... Time utilized Do you find yourself more connected with your family, close friends , relatives ? What you miss the most
0    R1      Delhi-NCR      21 ...      YES      YES      School/college
1    R2      Delhi-NCR      21 ...      YES      NO      Roaming around freely
2    R3      Delhi-NCR      20 ...      NO      YES      Travelling
3    R4      Delhi-NCR      20 ...      NO      NO      Friends , relatives
4    R5      Delhi-NCR      21 ...      NO      NO      Travelling
... ..
1177 R1191      Delhi-NCR      12 ...      YES      YES      Travelling
1178 R1192      Delhi-NCR      14 ...      YES      YES      Friends , relatives
1179 R1193      Delhi-NCR      13 ...      NO      YES      School/college
1180 R1194      Delhi-NCR      14 ...      YES      YES      School/college
1181 R1195      Delhi-NCR      13 ...      YES      YES      School/college
[1182 rows x 19 columns]
```

reviewData:

```
PS C:\Users\19172\Desktop> python midterm_csc221.py
Age of Subject  Time spent on self study  Time spent on fitness  Time spent on sleep  Time spent on social media  Preferred social media platform  Stress busters  What you miss the most
0              21              4.0              0.0              7.0              3.0              LinkedIn              Cooking              School/college
1              21              0.0              2.0              10.0             3.0              Youtube              Scrolling through social media  Roaming around freely
2              20              3.0              0.0              6.0              2.0              LinkedIn              Listening to music              Travelling
3              20              2.0              1.0              6.0              5.0              Instagram              Watching web series              Friends , relatives
4              21              3.0              1.0              8.0              3.0              Instagram              Social Media              Travelling
...           ...           ...           ...           ...           ...           ...           ...
1177          12              4.0              1.0              8.0              1.0              Instagram              Dancing              Travelling
1178          14              4.0              1.0              9.0              1.0              Whatsapp              Listening to music              Friends , relatives
1179          13              0.0              0.5              8.0              3.0              Youtube              Online gaming              School/college
1180          14              3.5              1.0              8.0              0.5              Youtube              Reading books              School/college
1181          13              2.0              0.5              7.0              1.0              Whatsapp              Talking              School/college

[1182 rows x 8 columns]
```

studentLifeData:

```
PS C:\Users\19172\Desktop> python midterm_csc221.py
Time spent on self study  Time spent on fitness  Time spent on sleep  Time spent on social media  Preferred social media platform
0              4.0              0.0              7.0              3.0              LinkedIn
1              0.0              2.0              10.0             3.0              Youtube
2              3.0              0.0              6.0              2.0              LinkedIn
3              2.0              1.0              6.0              5.0              Instagram
4              3.0              1.0              8.0              3.0              Instagram
...           ...           ...           ...           ...
1177          4.0              1.0              8.0              1.0              Instagram
1178          4.0              1.0              9.0              1.0              Whatsapp
1179          0.0              0.5              8.0              3.0              Youtube
1180          3.5              1.0              8.0              0.5              Youtube
1181          2.0              0.5              7.0              1.0              Whatsapp

[1182 rows x 5 columns]
```

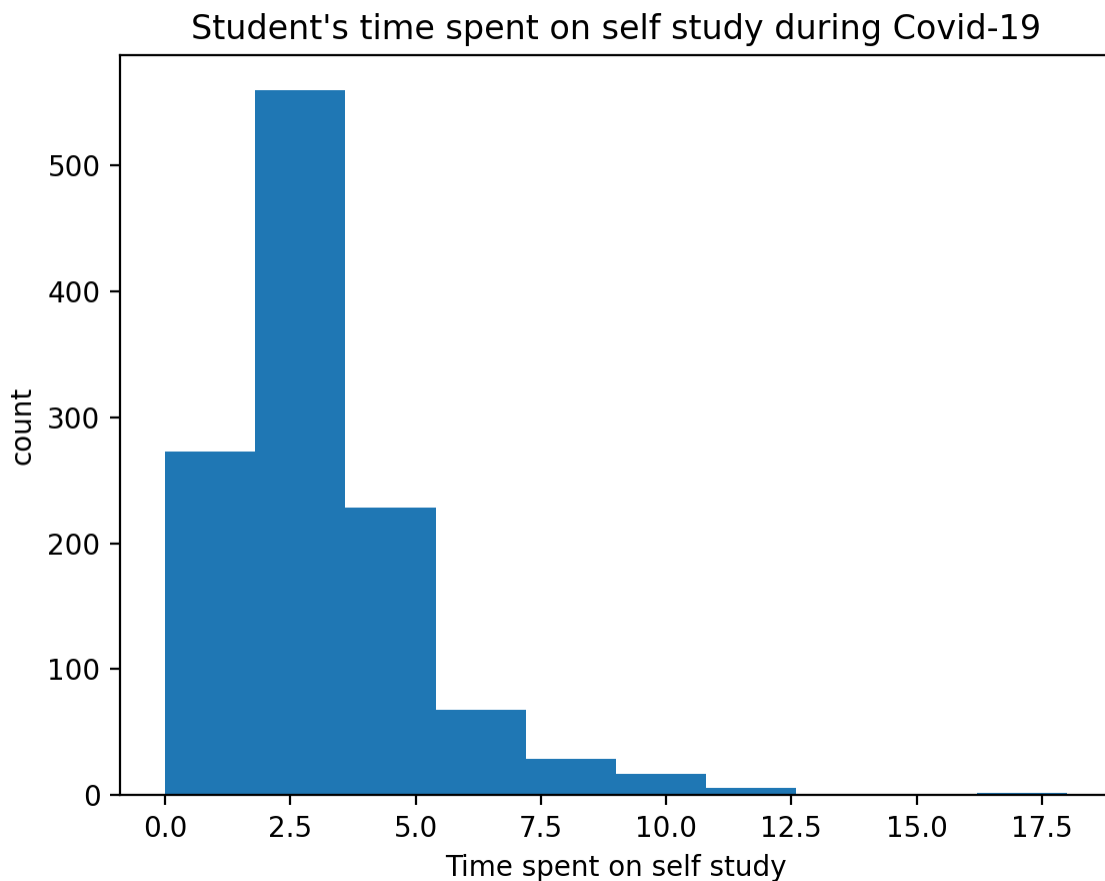
To check if there were any null values: `reviewData.isnull().any()` and `studentLifeData.isnull().any()`. Since all columns showed False, there were no null values in the data frame. The results are shown below where the top is `reviewData` and bottom is `studentLifeData`.

```
-----
Age of Subject              False
Time spent on self study    False
Time spent on fitness       False
Time spent on sleep         False
Time spent on social media  False
Preferred social media platform False
Stress busters              False
What you miss the most      False
dtype: bool
-----
Time spent on self study    False
Time spent on fitness       False
Time spent on sleep         False
Time spent on social media  False
Preferred social media platform False
dtype: bool
```

Explore the Data Statistically

Time Spent on Self Study

Dataset Dimension:



The histogram produced is skewed to the right where it seems that most students spend around 2 hours studying and there are a few that don't study at all. Although it is hard to see in the image above, there is an outlier at 18, where a student stated in their review that they had studied 18 hours a day.

Central Tendency

Mean: 2.911591
Median: 2.0
Mode: 2.0

The values for the mean and median suggest a graph that is skewed to the right because the value for the mean is greater than the value for the median. This can also be seen in the histogram above where the graph is also skewed to the right.

Standard Deviation

The standard deviation for this time spent on self study is 2.140590.

5 Number Summary

Minimum: 0.000000

1st Quartile: 2.000000

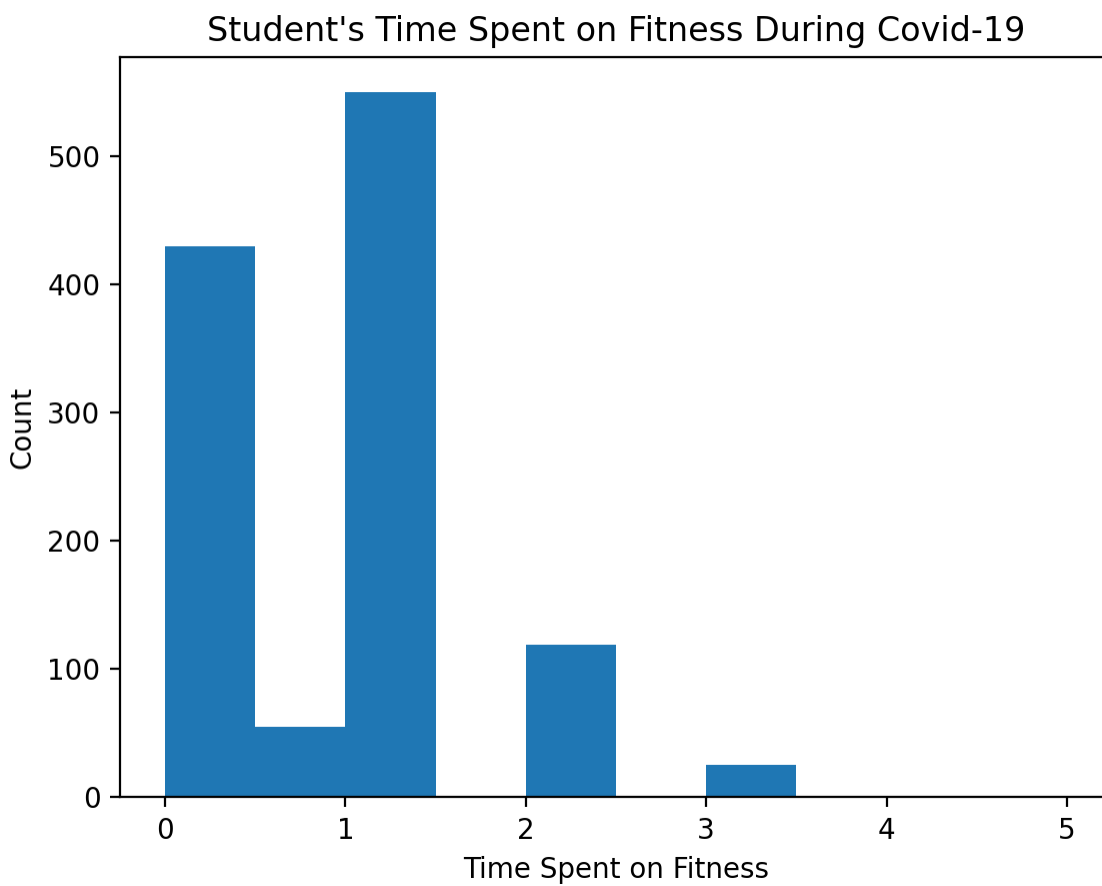
Median: 2.000000

3rd Quartile: 4.000000

Maximum: 18.000000

Time Spent on Fitness

Dataset Dimension:



The histogram produced doesn't show a specific shape and this can be because there were several people that either spent 0 or 1 hour on fitness since both the minimum values and median value was high in count. It seems uncommon for students to spend more than 3 hours on fitness.

Central Tendency

Mean: 0.765821
Median: 1.000000
Mode: 1.0

The values for the mean and median suggest a graph that is skewed to the left because the value for the median is greater than the value for the mean. However, the histogram produced doesn't exactly show this and the reason for this can be due to the values being close to one another.

Standard Deviation

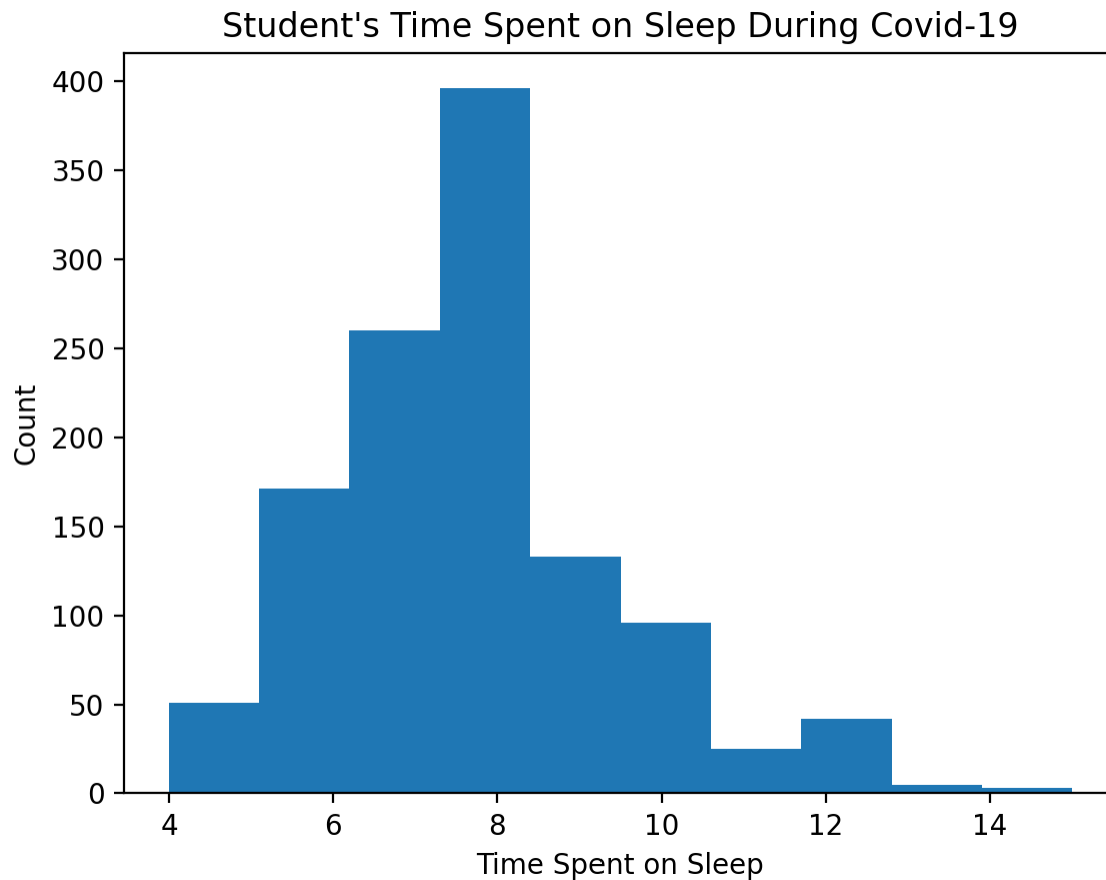
The standard deviation for this time spent on fitness is 0.724451

5 Number Summary

Minimum: 0.000000
1st Quartile: 0.000000
Median: 1.000000
3rd Quartile: 1.000000
Maximum: 5.000000

Time Spent on Sleep

Dataset Dimension:



The histogram produced seems to look symmetrical. Based off of the histogram most students get 8 hours of sleep and it is uncommon for students to be sleeping more than 11 hours.

Central Tendency

Mean: 7.871235

Median: 8.000000

Mode: 8.000000

The values for the mean and median suggest a graph that is skewed to the left because the value for the median is greater than the value for the mean. However, the histogram produced doesn't exactly show this and the reason for this can be due to the values being close to one another. In addition to that, it may also be due to the mode value and median value being equal, which may contribute towards why the graph looks a little symmetric.

Standard Deviation

The standard deviation for this time spent on fitness is 1.615762.

5 Number Summary

Minimum: 4.000000

1st Quartile: 7.000000

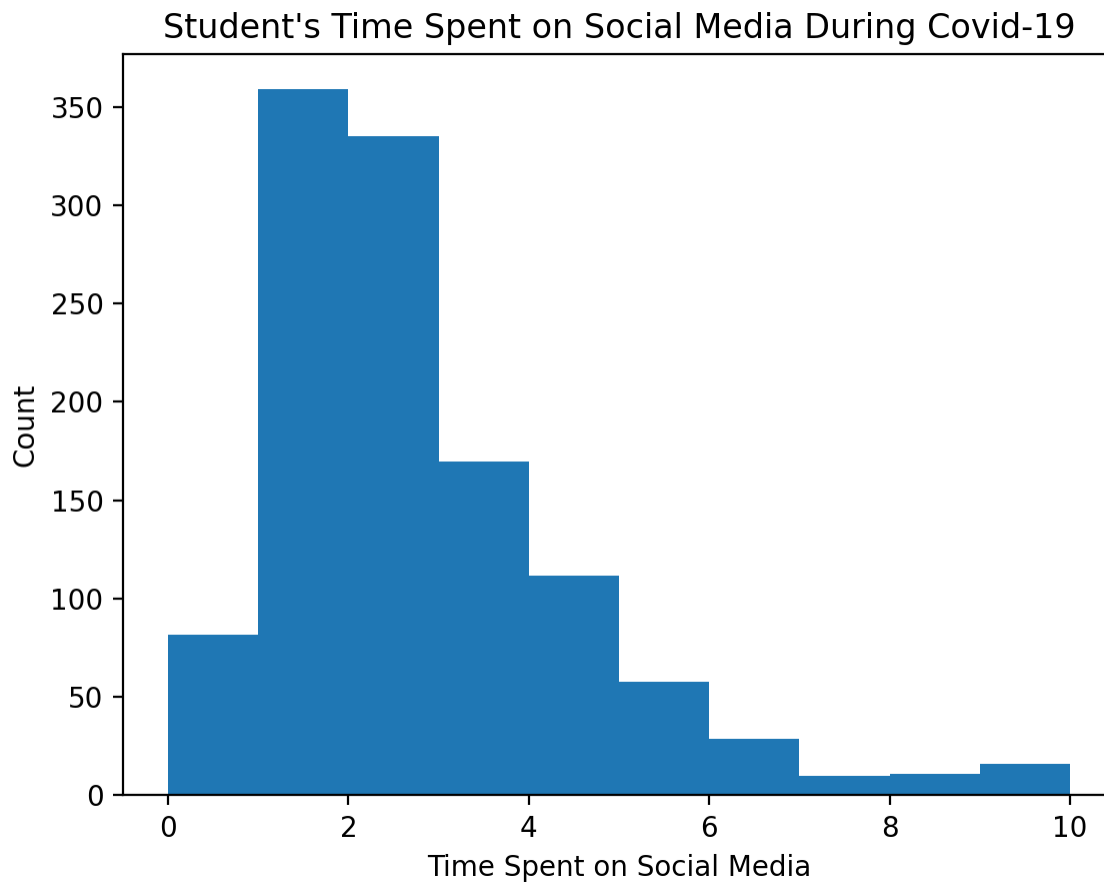
Median: 8.000000

3rd Quartile: 9.000000

Maximum: 15.000000

Time Spent on Social Media

Dataset Dimension:



The histogram produced is skewed to the right where it seems that most students spend around 2 hours on social media. It surprises me to see that there are a handful of students that don't spend

any time on social media at all, along with there being an outlier that spends 10 hours a day on social media.

Central Tendency

Mean: 2.365694

Median: 2.000000

Mode: 1.000000

The values for the mean and median suggest a graph that is skewed to the right because the value for the mean is greater than the value for the median which is greater than the value for the mode. This can also be seen in the histogram above where the graph is also skewed to the right.

Standard Deviation

The standard deviation for this time spent on fitness is 1.767336

5 Number Summary

Minimum: 0.000000

1st Quartile: 1.000000

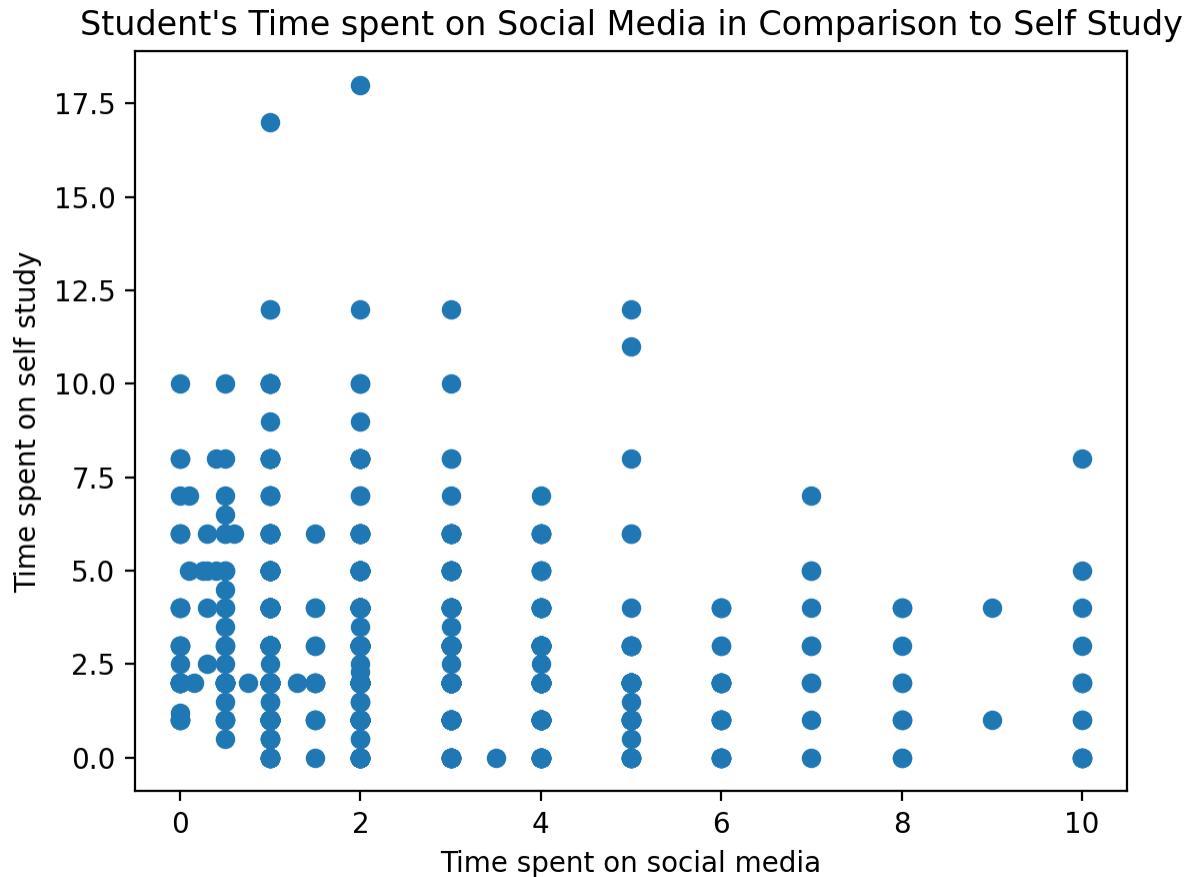
Median: 2.000000

3rd Quartile: 3.000000

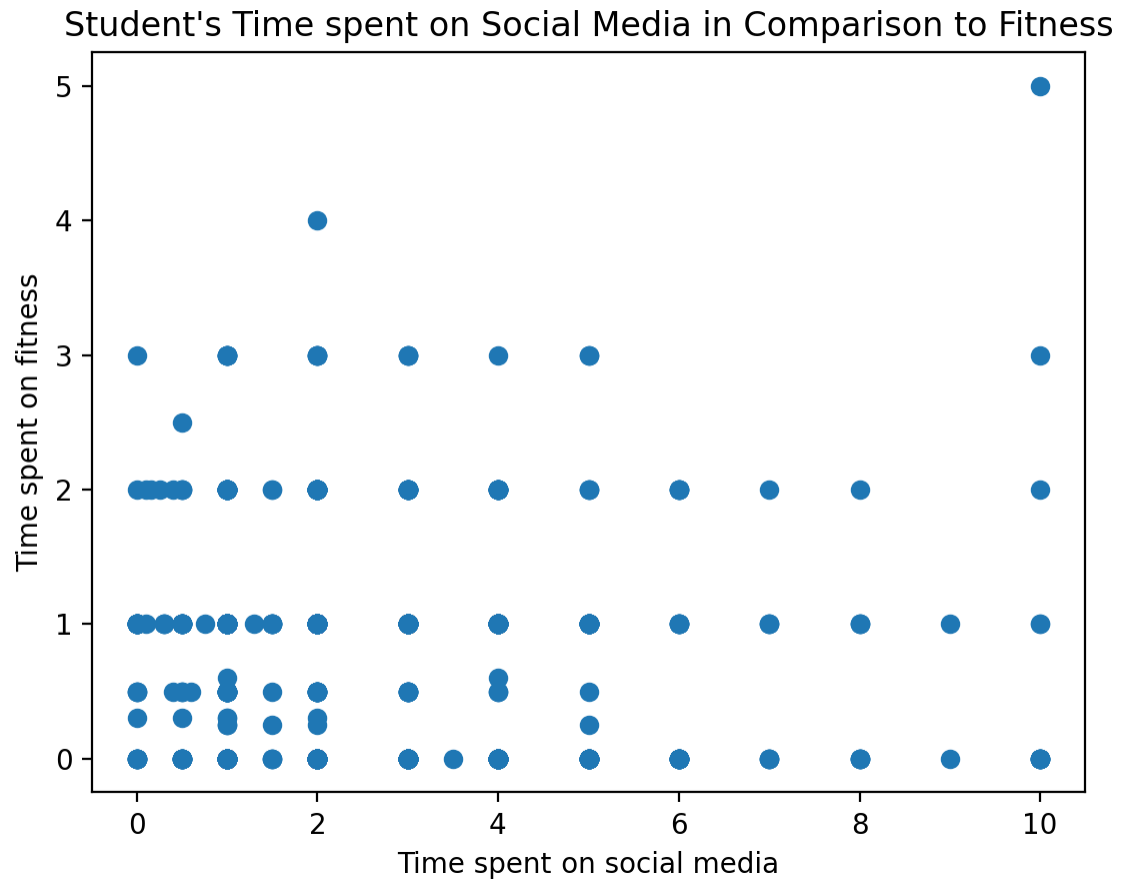
Maximum: 10.000000

Correlation (Connection)

Correlation between time spent on social media and time spent on self-study



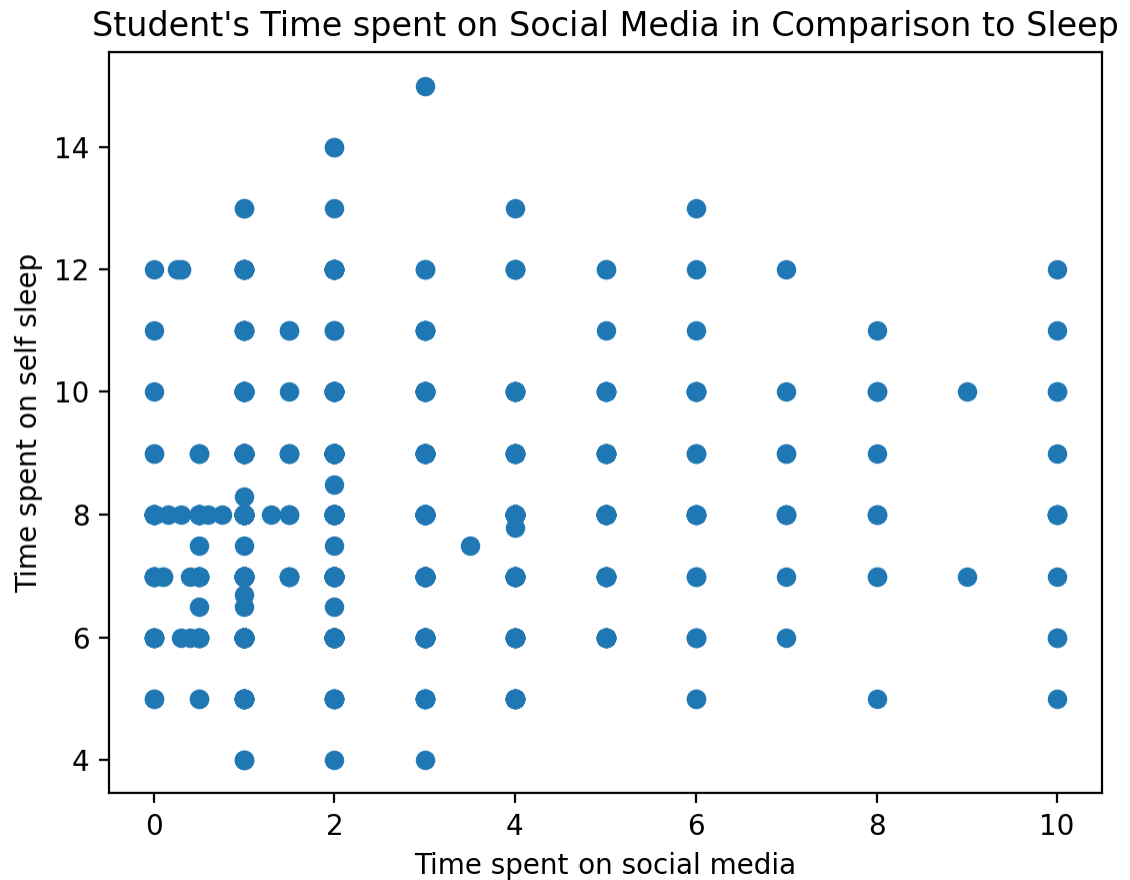
Correlation between time spent on social media and time spent on fitness



The Pearson's correlation coefficient r was -0.04285516491983569 . The value suggested a very weak negative correlation between the time students spend on social media and the time students spent on fitness since the value for r is almost zero. As a result, by looking at the scatterplot, it is hard to tell since the correlation is very weak and really close to showing no correlation. However, the correlation coefficient suggests that there is very little correlation showing that as the time spent on social media increases, the time spent on fitness will decrease.

I found these results to be reasonable because after students finish their classwork and homework, they are bound to figure out how they will spend their free time. As a result, the more time a student spends on social media, the less time they will have for fitness. Since the correlation is very weak, we can't really say that more time spent on social media leads to less time spent on fitness. The very weak correlation can also be due to the fact that some people have fixed times for the time spent on fitness everyday, resulting in the time spent on fitness to be the same regardless of they spent more or less time on social media.

Correlation between time spent on social media and time spent on sleep



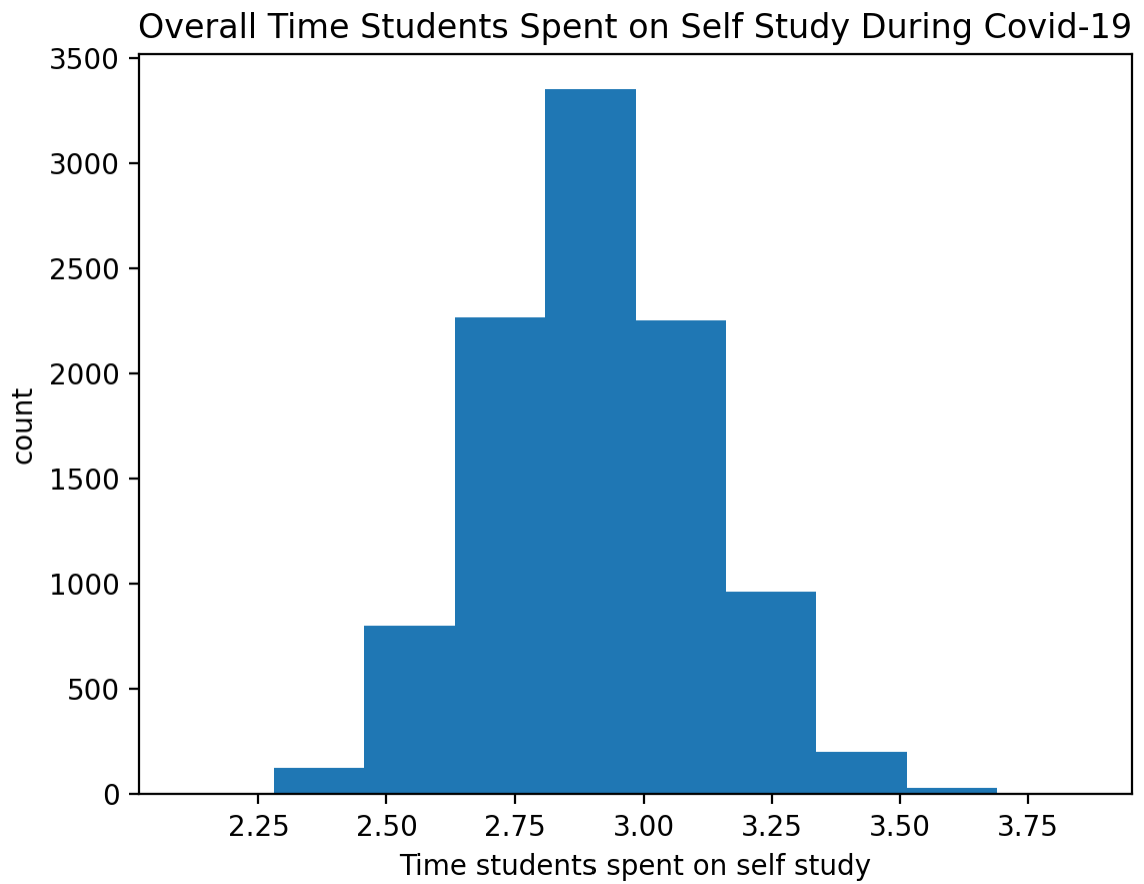
The Pearson's correlation coefficient r was 0.08566707624795465. The value suggested a weak positive correlation between the time students spend on social media and the time students spent on sleep. Since the r value is very close to zero, it makes the scatterplot to look like there is no correlation between the two variables when in actuality, there's a weak positive correlation. The positive coefficient value signifies that as the time spent on social media is increased, the time spent on sleep is increased, however we can't say this for certain due to the r value being so low.

I found these results to be shocking because, I'd assume that students tend to scroll through Instagram, YouTube, Facebook, and Twitter before they go to sleep and what may be intended to be 10 minutes of scrolling through social media turns out to be an hour or more. However, it surprises me to see that there are some students that spend both a lot of time sleeping and a lot of time on social media. This can also be a reason to why the correlation is weak.

Findings: All three studies of the correlations between the time students spent on social media compared to the time students spent on self study, sleep, and fitness were weak. Amongst the three, the strongest correlation was between the time spent on social media and the time spent on self study.

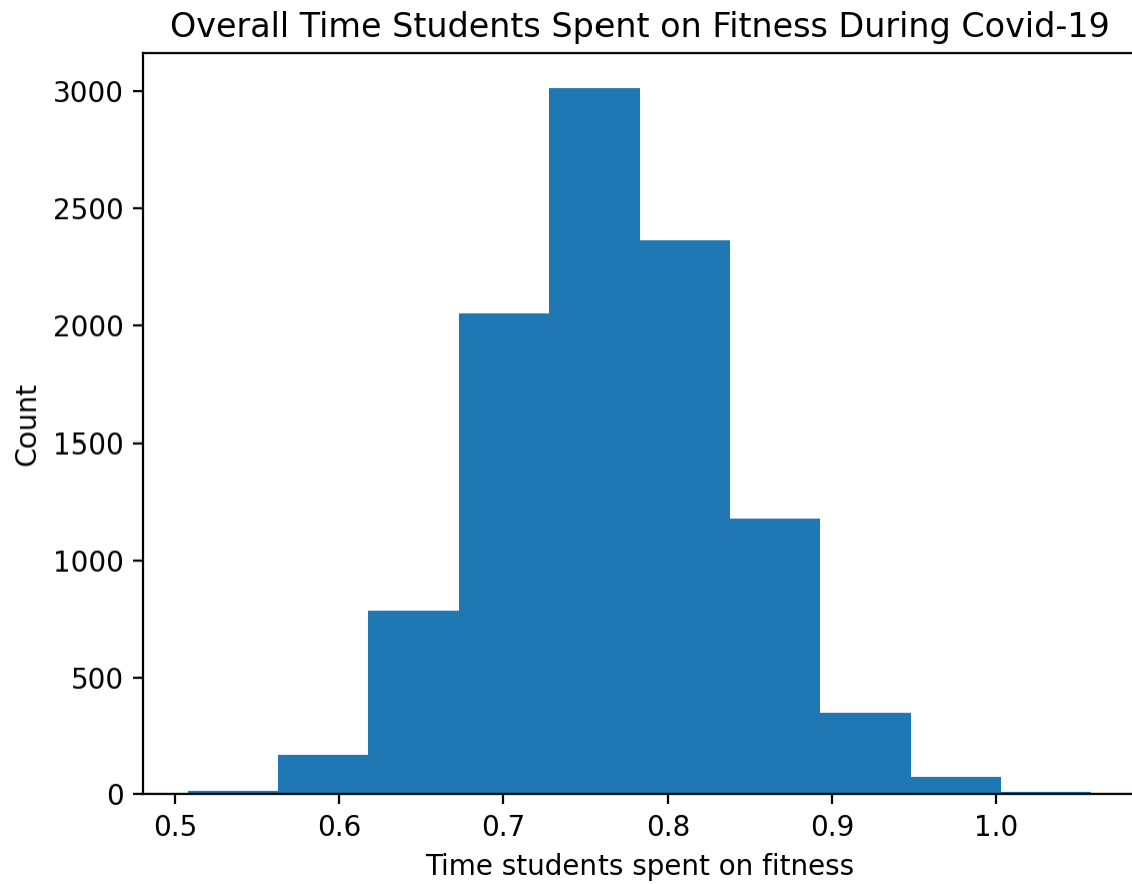
Normal Distributions

Normal Distribution for time spent on self study



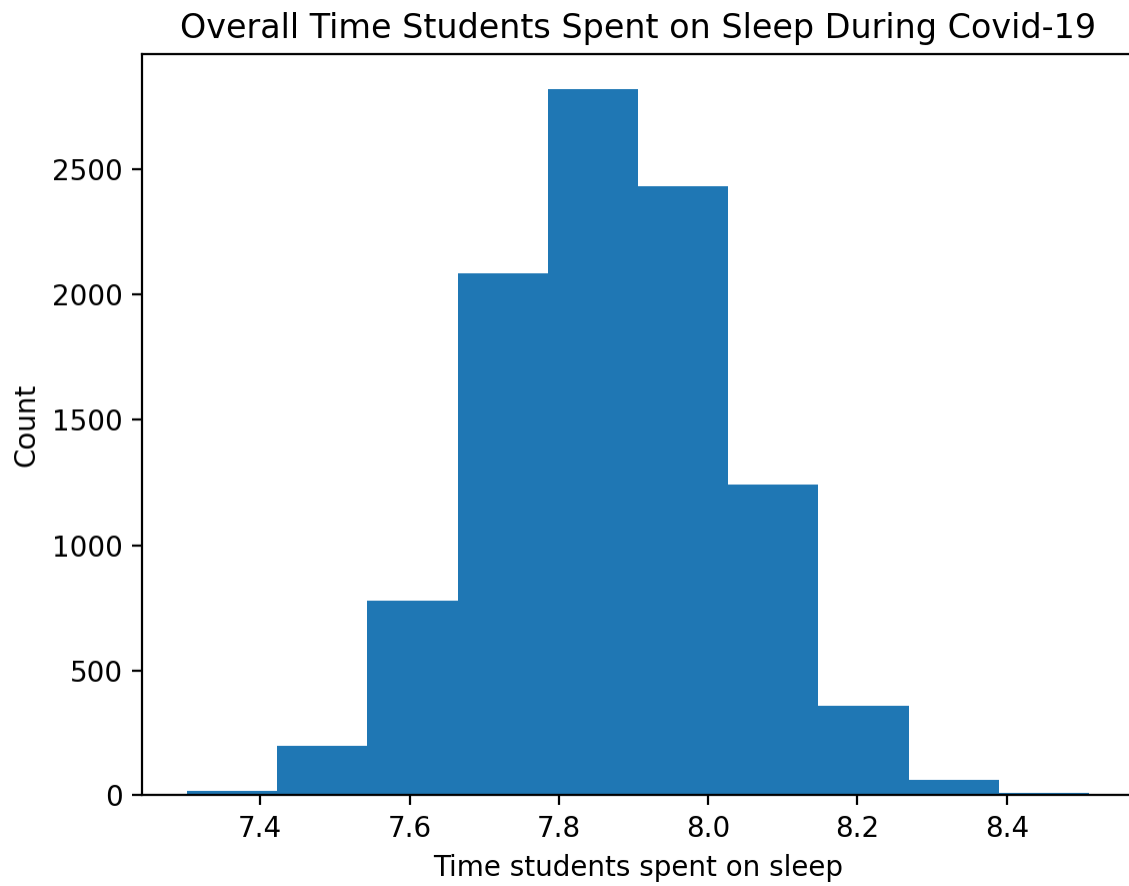
After applying the central limit theorem, I get a sampling mean value of 2.9097146

Normal Distribution for time spent on fitness



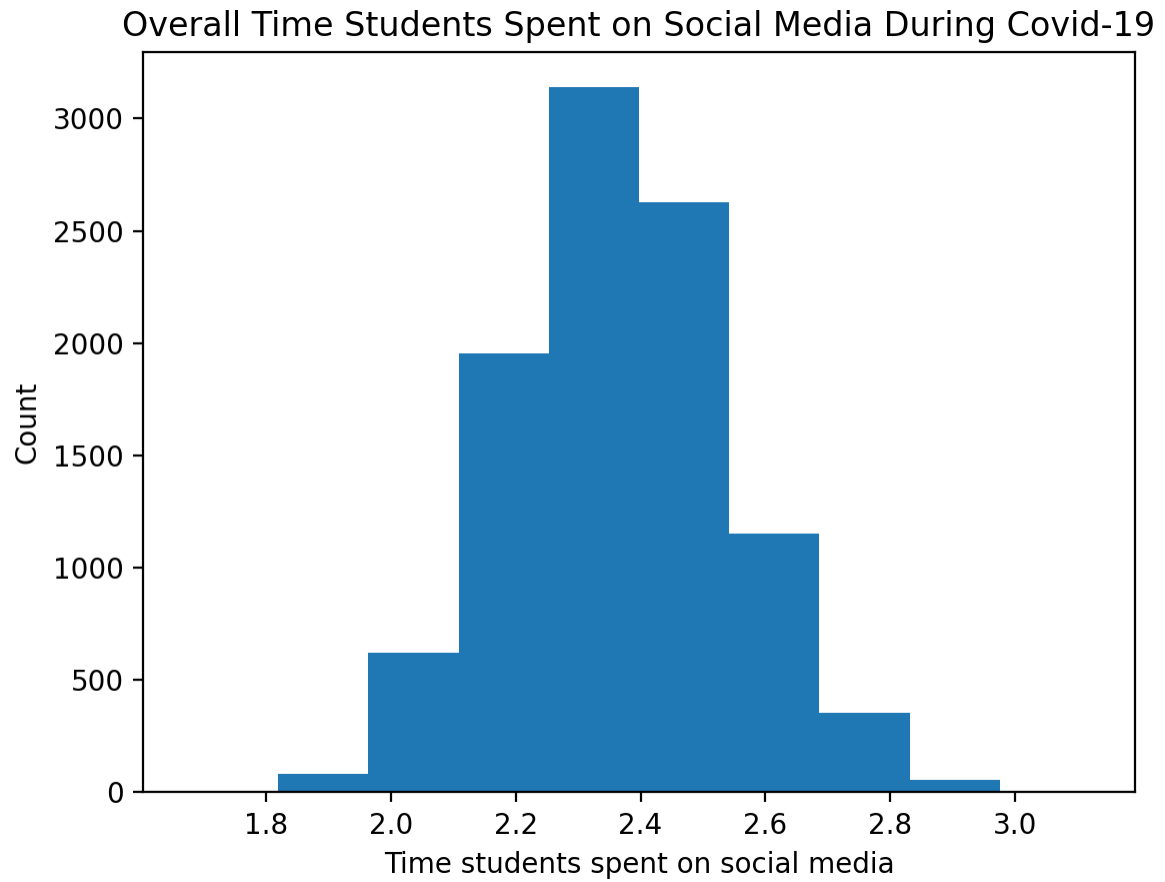
After applying the central limit theorem, I get a sampling mean value of 0.7665098

Normal Distribution for time spent on sleep



After applying the central limit theorem, I get a sampling mean value of 7.8699322

Normal Distribution for time spent on social media



After applying the central limit theorem, I get a sampling mean value of 2.3644672000000004

Look for Relationships Amongst the Dataset using Hypothesis Testing

Study comparing the study time amongst students that have high social media time versus low social media time

Question 1: Are the times spent on self study where time spent on social media is high(greater than or equal to 2.0) different than the times spent on self study where the time spent of social media is low(less than 2.0)?

Null Hypothesis: The time students spent on self study is the same whether or not the time spent on social media is high or low.

Alternative Hypothesis: The time students spent on self study is different for when the time on social media is high compared to when the time on social media is low.

```
Question 1: Are the times spent on self study where time spent on social media is high(greater than or equal to 2.0) different than the times spent on self study where the time spent of social media is low(less than 2.0)?
```

```
H0: The time students spent on self study is the same whether or not the time spent on social media is high or low.
```

```
HA: The time students spent on self study is different for when the time on social media is high compared to when the time on social media is low.
```

```
The p value for this case is: 0.0010966290268165607
```

```
Conclusion: Reject H0 and accept HA to say that the time students spent on self study is different for when the time on social media is high compared to when the time on social media is low.
```

To solve this question, I used a T-Test and used the `ttest_ind()` function since we are looking at 2 groups in the same column. Here the two groups are the time spent on social media is high and the time spent of social media is low which both belong in the social media columns. Amongst the two groups, a study is conducted analyzing the amount spent studying in those groups.

This function results in the p- value being 0.0010966290268165607. Since the p-value is very close to zero, we must reject the null hypothesis and accept the alternative hypothesis to say that the time students spent on self study is different for when the time on social media is high compared to when the time on social media is low

Question 2: Amongst students who chose Instagram as the medium for social media platforms, the time spent on sleep is lower where the time spent on social media is high compared to the times where social media time is low.

Null Hypothesis: Amongst students that chose Instagram as their preferred medium, the time spent on sleep is the same for students that spent high amounts of time on social media as the students that spent low amount of time on social media.

Alternative Hypothesis: Amongst students that chose Instagram as their preferred medium, the time spent on sleep is the lower for students that spent high amounts of time on social media compared to the students that spent low amount of time on social media.

```
The p value for this case is: 0.8197238302305175  
Conclusion: Accept H0 and say that amongst students that chose Instagram as their preferred medium, the time spent on sleep is the same for students that spent high amounts of time on social media as the students that spent low amount of time on social media.
```

This is a study to see whether students that spend a lot of time on Instagram sacrifice their sleep to be on social media. During Covid times, Instagram was one of the most popular forms of communication and self-expression. Some students wouldn't sleep on time since they may not need to wake up earlier for transportation.

To solve this question, I used a T-Test and used the `ttest_rel()` function to evaluate the difference between two independent means. The `ttest_rel()` was used because unlike Question 1, this question requires us to look at two different columns. For this question, a low amount of time for social media are those less than the median which is 2.0 and a high amount of time for social media are those greater than or equal to 2.0.

This function results in the p-value being 0.8197238302305175. Since the p-value is greater than 5%, we must accept the null hypothesis and say that amongst students that chose Instagram as their preferred medium, the time spent on sleep is the same for students that spent high amounts of time on social media as the students that spent low amount of time on social media.

Compare the effects of Instagram and LinkedIn on student study time given students that spend a lot of time on social media

Question 3: Amongst students who chose Instagram as the medium for social media platforms, is the time spent on self study is lower where the time spent on social media is high compared to the times where social media time is low.

Null Hypothesis: Amongst students who chose Instagram as the medium for social media platforms, the time spent on self study is the same where the time spent on social media is high compared to the times where social media time is low.

Alternative Hypothesis: Amongst students who chose Instagram as the medium for social media platforms, the time spent on self study is lower where the time spent on social media is high compared to the times where social media time is low.

```
The p value for this case is: 0.0022602550063179263
Conclusion: Reject H0 and accept HA to say that amongst students who chose Instagram as the medium for social media platforms, the time spent on self study is lower where the time spent on social media is high compared to the times where social media time is low.
```

To solve this question, I used a T-Test and used the `ttest_rel()` function to evaluate the difference between two independent means. Similar to Question 2, a low amount of time for social media are those less than the median which is 2.0 and a high amount of time for social media are those greater than or equal to 2.0.

This function results in the p-value being 0.0022602550063179263. Since the p-value is less than 5%, we must reject null hypothesis and accept alternate hypothesis to say that amongst students who chose Instagram as the medium for social media platforms, the time spent on self study is lower where the time spent on social media is high compared to the times where social media time is low.

Question 4: Amongst students who chose LinkedIn as the medium for social media platforms, the time spent on self study is lower where the time spent on social media is high compared to the times where social media time is low.

Null Hypothesis: Amongst students who chose LinkedIn as the medium for social media platforms, the time spent on self study is the same where the time spent on social media is high compared to the times where social media time is low.

Alternative Hypothesis: Amongst students who chose LinkedIn as the medium for social media platforms, the time spent on self study is lower where the time spent on social media is high compared to the times where social media time is low.

```
The p value for this case is: 0.08293087317322734
Conclusion: Accept H0 and say that amongst students who chose Instagram as the medium for social
media platforms, the time spent on self study is the same where the time spent on social media
is high compared to the times where social media time is low.
```

To solve this question, I used a T-Test and used the `ttest_rel()` function to evaluate the difference between two independent means. Similar to Question 2, a low amount of time for social media are those less than the median which is 2.0 and a high amount of time for social media are those greater than or equal to 2.0.

This function results in the p- value being 0.08293087317322734. Since the p-value is greater than 5%, we must accept the null hypothesis and say that amongst students who chose LinkedIn as the medium for social media platforms, the time spent on self study is the same where the time spent on social media is high compared to the times where social media time is low.

Findings

Questions 3 and 4 were completed to help analyze whether the social media platform a student uses would have an impact on whether or not they would spend more time studying. From the study conducted, we can see that students that are on social media a lot that choose to use Instagram spend less time on self study. However, we can't say for certain that this is true since all we know for sure is that the null hypothesis stating that students spending a lot of time on Instagram study the same amount as students that spend less time on Instagram is false. On the other hand, we see that the null hypothesis for this same study for students that spend little time on LinkedIn versus students that spend a lot of time on LinkedIn spend around the same amount of time for self study.

Independence between columns

Social Media time and Fitness time

Question 5a: Is the student's time spent on social media(greater or less than 2.0) independent from the time students spend on fitness(less than or greater than 1)

Null Hypothesis: The two variables, student's time on social media and the student's time doing fitness are independent

Alternative Hypothesis: The two variables, student's time on social media and the student's time doing fitness are dependent

To analyze the independence of two variables, chi-square testing was utilized.

I was able to count the amount of occurrences there was of high social media time and high fitness time, high social media time and low fitness time, low social media time and high fitness time, and low social media time and low fitness time and created a table.

```
Testing for independence between students social media time and fitness time
              Fitness Time greater than or equal to 2  Media Time less than 2
Media Time greater than or equal to 2                424                317
Media Time less than 2                                273                168
p value is 0.12789738176130594
Conclusion: The two variables are dependent(reject H0)
```

After creating this table, I was able to do chi-testing using the `chi2_contingency()` function and got a p-value of 0.12789738176130594. Since the p value was very small and close to zero, we must reject the null hypothesis and say that the two variables are dependent on each other.

I found this surprising because looking back at the scatter plot data for these two columns, we see that the correlation between these two variables is very weak. Although the correlation may be weak, there still seems to be dependence between the two variables.

Social Media time and Self Study time

Question 5a: Is the student's time spent on social media(greater or less than 2.0) independent from the time students spend on self study(less than or greater than 2.0)

Null Hypothesis: The two variables, student's time on social media and the student's time doing self study are independent

Alternative Hypothesis: The two variables, student's time on social media and the student's time doing self study are dependent

To analyze the independence of two variables, chi-square testing was utilized.

I was able to count the amount of occurrences there was of high social media time and high self study time, high social media time and low self study time, low social media time and high self study time, and low social media time and low self study time and created a table.

```
Testing for independence between students social media time and self study time
                                     Fitness Time greater than or equal to 2  Media Time less than 2
Media Time greater than or equal to 2                                     555                                186
Media Time less than 2                                                  354                                87
p value is 0.040505601914242344
Conclusion: The two variables are dependent(reject H0)
```

After creating this table, I was able to do chi-testing using the `chi2_contingency()` function and got a p-value of 0.040505601914242344. Since the p value was small, we must reject the null hypothesis and say that the two variables are dependent on each other.

I found this reasonable due to the fact that we were able to see more correlation between these two columns of data when analyzing the scatterplot produced by these two variables.

Social Media time and Sleep time

Question 5a: Is the student's time spent on social media(greater or less than 2.0) independent from the time students spend on sleep(less than or greater than 8)

Null Hypothesis: The two variables, student's time on social media and the student's time doing sleep are independent

Alternative Hypothesis: The two variables, student's time on social media and the student's time doing sleep are dependent

To analyze the independence of two variables, chi-square testing was utilized.

I was able to count the amount of occurrences there was of high social media time and high sleep time, high social media time and low sleep time, low social media time and high sleep time, and low social media time and low sleep time and created a table.

```
Testing for independence between students social media time and sleep time
                                Fitness Time greater than or equal to 2  Media Time less than 2
Media Time greater than or equal to 2                                458                                283
Media Time less than 2                                              237                                204
p value is 0.007721241682365073
Conclusion: The two variables are dependent(reject H0)
```

After creating this table, I was able to do chi-testing using the `chi2_contingency()` function and got a p-value of 0.007721241682365073. Since the p value is very small, we must reject the null hypothesis and say that the two variables are dependent on each other.

Goal: Predict if the next student to write a review will choose Instagram to be their preferred social media platform given information about the age and how they spend their day.

Extract features

The first step was to extract all the necessary features and split the data into train and test data.

First, I set the independent variable to be a data frame containing the age, time spent studying, fitness, sleep, social media, stress buster, what they miss most.

Then I set the target variable, y, to be preferred social media platform column.

Train the data so 80% will be training data and 20% will be test data

We get the first five rows of the trained data to be:

```
[ [20.  3.  0.  8.  4. ]  
  [19.  4.  1.  6.  5. ]  
  [12.  4.  1.  8.  1. ]  
  [20.  4.  1.  8.  1.5]  
  [20.  1.  0. 10.  1. ]]
```

Standardization

Did `sc.fit_transform` to our `x_train` dataframe to fit it over the data and transform it.

After that, did `sc.transform(x_test)` to perform centering and scaling for the `x_test`.

Import our Gaussian Model

Now it was time to train the model using the training set using the `gnb.fit()` function.

Now to predict the target response for our test dataset and store it in `y_predict` which looks like:

```
-----  
[ 2  1  2  1  5  1  2  6  0  1  3  4  1  5  1  2  2  1  1  3  3  3  2  4  
 4  2  2  1  1  1  1  1  0  1  3  1  2  1  2  2  2  3  1  4  7  4  3  2  
 2  2  4  1  4  1  0  4  6  0  2  2  3  1  2  4  5  1  1  1  1 10  1  2  
 1  3  1  5  2  2  1  1  4  3  7  1  2  4  2  2  2  2  0  2  2  1  2  1  
 0  1  4  3  7  2  1  8  1  1  3  3  1  1  1  2  0  5  3  2  4  1  4  2  
 5  2  2  0  4  4  6  1  2  1  2  2  2  2  1  1  2  1  1  3  1  1  1  1  
 3  1  4  3  1  4  1  2  3  1  4  2  2  6  2  1  1  3  3  4  1  2  6  3  
 1  2  2  1  4  2  3  2  5  1  6  3  5  3  3  2  1  2  4  1  0  1  
 1  1  0  1  3  1  3  7  4  3  1  2  6  2 10  2  3  0  2  5  1  1  2  2  
 1  1  4  2  2  2  4  1  1  1  3  2  2  2  2  1  1  3  2  6  4]
```

```
PS C:\Users\19172\Desktop>
```

EVALUATE THE MODEL'S OUTPUT

Accuracy measures the correctly classified observations in the model. The program outputted a value of 0.99578 for our Gaussian model which means that the model is looking good.

```
-----  
The accuracy of the Gaussian Model is: 0.9957805907172996  
-----
```


Additionally, we can evaluate the model's output by looking at the confusion matrix to see the amount of false positives and negatives there are. We can also use this matrix to calculate the sensitivity and specificity values. The sensitivity measures the proportion of positives that are correctly identified in the model and for the model produced, the sensitivity value is 0.916. The specificity value measures the proportion of negatives that are correctly identified in the model. The model produced by this program, we get a specificity value of 1.0. These values for sensitivity and specificity are high and reveal that the model is good and that it was correctly able to identify most information.

```
-----  
Sensitivity:  0.9166666666666666  
-----  
Specificity:  1.0  
-----
```

CONCLUSION

Result: [0]

After evaluating the model's outputs, I concluded that the model seemed to be good enough to predict what platform of social media then next student to write a review will pick. In order to do so, I used the `gnb.predict` function and got a result of 0 which is WhatsApp. To ensure that it was WhatsApp, I looked back to see which row had a 1 for that column after I used the `labelencoder` and found that row 1177 had a 1 for that column. After that, I went to the dataset and saw that 1177 preferred platform for social media was WhatsApp. As a results, 1 represented WhatsApp.

More prediction were conducted on the dataset to determine how the student spends their day and the results we got is as follows:

```
[0]  
Age:  14.79037778668566  
Time spent on self study:  0.7612324670678978  
Time spent on fitness:  0.03609056003487021  
Time spent on sleep:  6.2175421091024194  
Time spent on social media:  0.5774841172558074  
PS C:\Users\19172\Desktop> █
```