"Does screen time affect the amount and quality of sleep people get?"

By Betel Mekonnen

STAT 430

Introduction

In this project, I am looking to answer the question "Does screen time affect the amount and quality of sleep people get?". The quantitative variables I have used are Age, Screen Time, and Sleep hours. So to clarify, I collected each participant's age, the average number of hours they spent on their phones per week, and the average number of hours they slept per night. For my qualitative variable, I collected each participant's Gender, primary phone activity, their most used app, and their sleep quality. For Gender, primary phone activity, and sleep quality, I categorized them in the survey and gave them options. For "Gender: I provided the categories: "Male", "Female", and "Non-binary/other". For primary phone activity, I had "Socials", "Messaging", "Work/study", "Streaming", and "other". And finally for sleep quality, I had them rate their quality of sleep using "Very good", "Good", and "Fair"

Statistical summary

The mean age of the participants was 22.6 years, a median of 22 years and a standard deviation of 3.53 years. And looking at the distribution of Age, you can see that it is slightly skewed. And you can visually and numerically see that the mean is slightly larger than the median so it means the distribution is skewed right. Since it is slightly skewed right, the median would be the best measure of spread since we can visually see that there some outliers that are skewing the data, so that can impact the mean and range. So the median and the interquartile range are the best measure of center and spread.

The mean screen time the participants had was 6.68 hours, a median of 6.45 hours and a standard deviation of 2.86 hours. And looking at the distribution of Screentime, you can see that it is slightly skewed. And you can visually and numerically see that the mean is slightly larger than the median so it means the distribution is skewed right. Since it is slightly skewed right, the median would be the best measure of spread since we can visually see that there some outliers that are skewing the data, so that can impact the mean and range. So the median and the interquartile range are the best measure of center and spread.

The mean of the amount of sleep our participants get was 7.06 hours, a median of 7.06 hours and a standard deviation of 0.94 hours. And looking at the distribution of Sleephour, you can see that it looks pretty normal, one of the bins is just missing values. And you can confirm numerically by seeing that the mean and the median are not too far off from each other. In this case the mean and standard deviation would be a pretty good measure of center.

[Proc means output]:

The MEANS Procedure

Variable	Mean	Median	Std Dev	Quartile Range	Maximum	Minimum
Age	22.55	22.00	3.53	4.00	38.00	17.00
Screentime	6.68	6.45	2.86	3.66	18.00	2.00
Sleephours	7.06	7.00	0.94	1.25	10.00	5.00

[Proc univariate output]:

Distribution of Age

Distribution of Screentime

Distribution of Sleephours
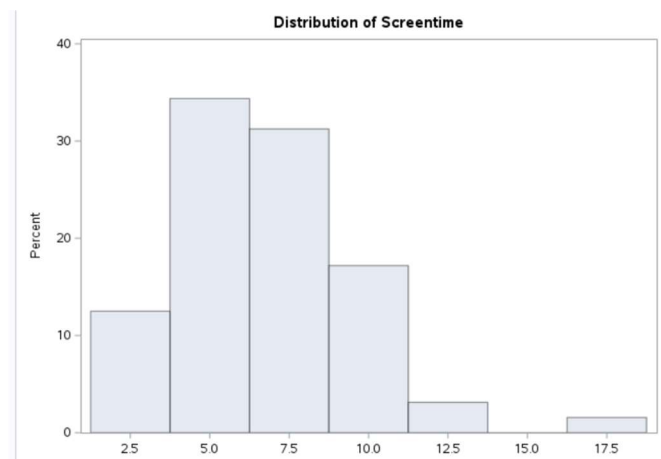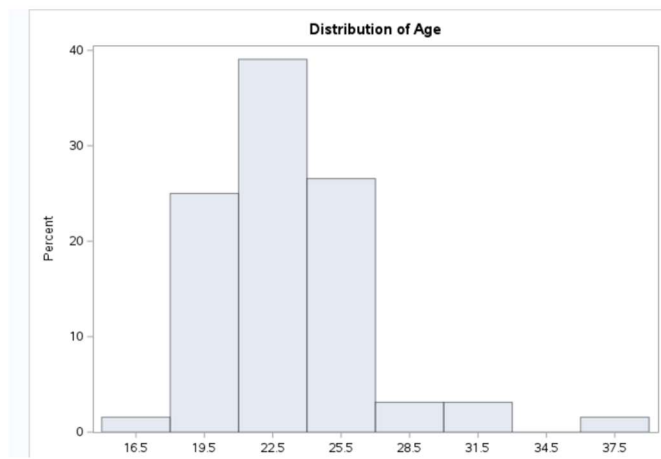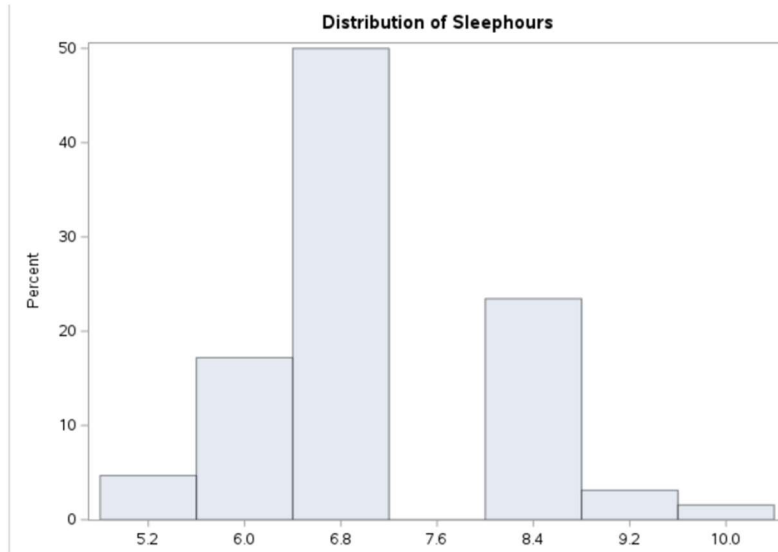
[Proc freq output]:

The FREQ Procedure

Gender	Frequency	Percent
Female	37	57.81
Male	24	37.50
Other	3	4.69

Activity	Frequency	Percent
Gaming	2	3.13
Messaging	7	10.94
Socials	32	50.00
Streaming	15	23.44
Work	8	12.50

App	Frequency	Percent
Chrome	5	7.81
Hulu	2	3.13
Instagram	8	12.50
Kindle	1	1.56
MAX	1	1.56
Messages	7	10.94
Netflix	1	1.56
Overmorta	1	1.56
Pokemon G	1	1.56
Reddit	3	4.69
Safari	1	1.56
Tik tok	20	31.25
Twitter	4	6.25
Youtube	9	14.06

Sleepquality	Frequency	Percent
Fair	22	34.38
Good	23	35.94
Very good	19	29.69

Looking at the qualitative data, I noticed that most of our participants, specifically 58% were female. I also noticed that the quality of sleep ratings was evenly spread out between our three categories but a majority of our participants rated their sleep quality as "good", followed closely by 22% of our participants who rated their sleep as "Fair".

Correlation

The correlation coefficient between Age and Sleephours was -0.313, telling us that there is a weak negative relationship between the two variables, which could indicate that as the age of our participants increases, the amount of sleep they get decreases.

Similarly, the correlation coefficient between Screentime and Sleephours is -0.2514, telling us that there is a weak negative relationship between the two variables, which could indicate that as the as screen time increases, the amount of sleep they get decreases. Which is what we initially sought to test but since the correlation is so low, we have to pivot and look at something else.

And the correlation coefficient between Age and Screentime was 0.30808, which tells us there is a weak but positive relationship between the age of our participants and the amount of time they spend on their phones. These two variables have the best correlation compared to the rest of the variables so we will look at the relationship between them further.

[Proc Corr results and interpretation]

The CORR Procedure

3 Variables:	Age Screentime Sleephours

Pearson Correlation Coefficients, N = 64
Prob > |r| under H0: Rho=0

	Age	Screentime	Sleephours
Age	1.00000	0.30808 0.0133	-0.31324 0.0117
Screentime	0.30808 0.0133	1.00000	-0.25139 0.0451
Sleephours	-0.31324 0.0117	-0.25139 0.0451	1.00000

Spearman Correlation Coefficients, N = 64
Prob > |r| under H0: Rho=0

	Age	Screentime	Sleephours
Age	1.00000	0.01861 0.8840	-0.25127 0.0452
Screentime	0.01861 0.8840	1.00000	-0.22289 0.0767
Sleephours	-0.25127 0.0452	-0.22289 0.0767	1.00000

Linear Regression Model

Regression 1: All numeric data

Considering the assumptions for the regression, the data is normal when looking at the residual distribution, as it does have a clearly normal distribution, and the data is independent as one point has no way of influencing the other.

Looking a closer look at our outputs, this is not a very good model. The R squared values is 0.1256 which is very low which tells us that this isn't a very a great model for predicting values. Our R squared value needs to be closer to 1 and this is a very low R squared value. Looking at the P values of our variables, Age is significant since it is lower than our alpha or 0.05 which lets us know that we should keep this variable in our model. But screentime is not significant with a p value of 0.1791, so it might make our model better to remove this variable.

And finally, looking at our residuals for screentime and Age, there is no clear pattern which let's us know that it is a good fit. But when looking at the residuals for the regression there is a very clear pattern which indicates it is not a good fit, There are a few outliers but only one outlier that seems like it might be an influential observation. And I remember there was a data point where the participant was 38 years old and I knew it was an outlier but removing it made the correlation go drastically go down so I had to leave it in. I know this isn't the best data but I had to leave it in so I can at least have an okay data to analyze. So I am sure that 38 year old is that outlier.

Summary:

Equation: Sleephours = 8.99648 - 0.06917 (Age) - 0.05608 (Screentime)

R squared= 0.1246

Adj R squared= 0.0959

P Values:

Age: 0.0428

Screentime: 0.1791

[Proc reg output with all the numeric data]

The REG Procedure
Model: MODEL1
Dependent Variable: Sleephours

Number of Observations Read	64
Number of Observations Used	64

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6.88548	3.44274	4.34	0.0173
Error	61	48.36452	0.79286		
Corrected Total	63	55.25000			

Root MSE	0.89043	R-Square	0.1246
Dependent Mean	7.06250	Adj R-Sq	0.0959
Coeff Var	12.60782		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.99648	0.72696	12.38	<.0001
Age	1	-0.06917	0.03343	-2.07	0.0428
Screentime	1	-0.05608	0.04126	-1.36	0.1791

Fit Diagnostics for Sleephours

Residual by Regressors for Sleephours
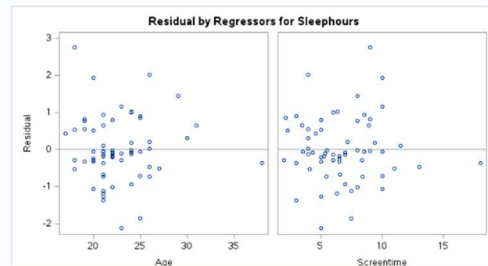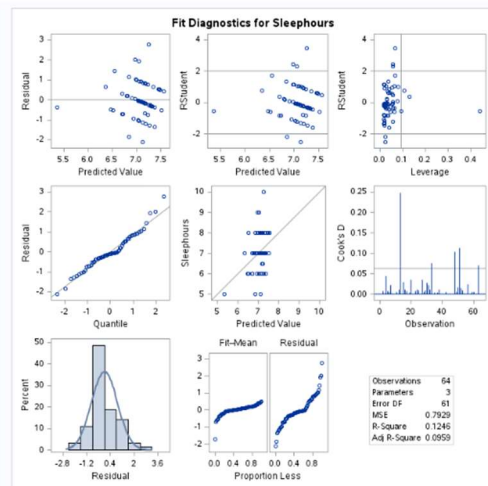
Regression 2: Numeric variables and one qualitative variable

Considering the assumptions for the regression, the data is normal when looking at the residual distribution, as it does have a clearly normal distribution, and the data is independent as one point has no way of influencing the other.

Looking a closer look at our outputs, this is not a very good model. The R squared values is 0.1499 which is very low which tells us that this isn't a very a great model for predicting values. Our R squared value needs to be closer to 1 and this is a very low R squared value. Looking at the P values of our variables, Age is significant with a p value of 0.0329 since it is lower than our alpha or 0.05 which lets us know that we should keep this variable in our model. But screentime, Gender_Male, Gender_other is not significant as their p values were much higher than our alpha of 0.05, so it might make our model better if we removed these variables.

And finally, looking at our residuals for screentime and Age, there is no clear pattern which let's us know that it is a good fit. But when looking at the residuals for Gender_male and Gender_other there is a clear pattern which is not good. And similarly to our previous regression, when looking at the residuals for the regression there is a very clear pattern which indicates it is not a good fit, There are a few outliers but only one outlier that seems like it might be an influential observation. And I left that variable in as it was helping with my correlation but if I had collected the data better I would definitely get rid of this outlier as it is drastically influencing the model.

Summary:

R Squared = 0.1499

Adj R Squared = 0.0923

Equation: Sleephours = 9.11467 - 0.0736 (Age) - 0.03981 (Screentime) – 0.32309 (Gender_Male) – 0.10041 (Gender_other)
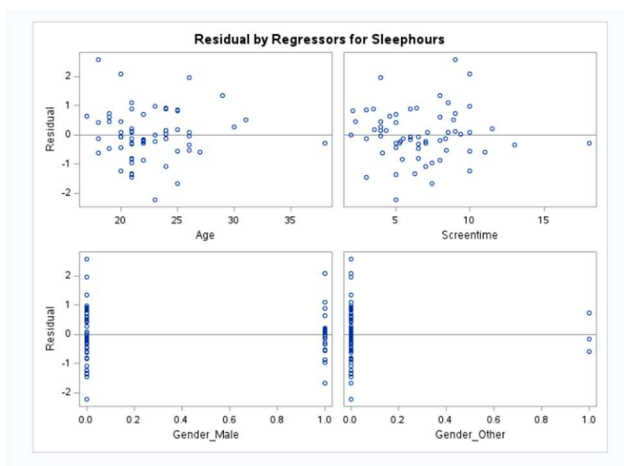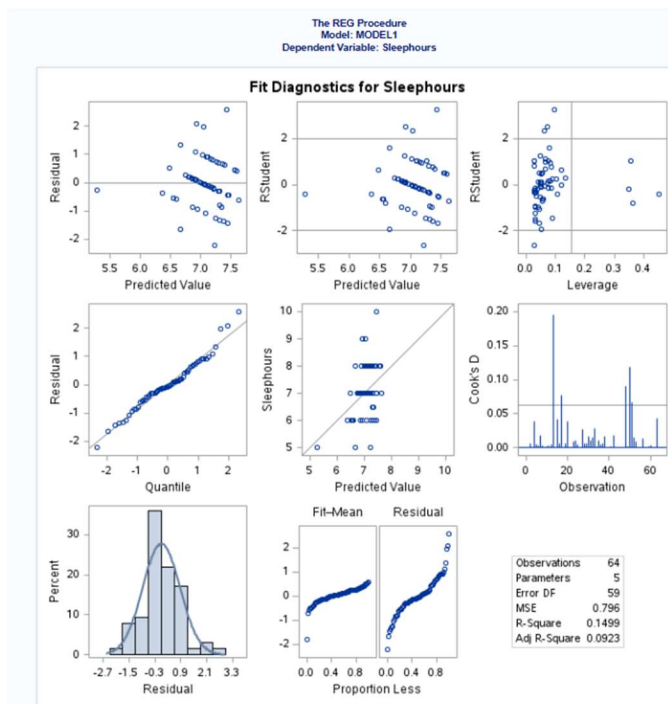
P values:

Screentime: 0.3675

Age: 0.0329

Gender(Male): 0.1918

Gender(Other): 0.8554

[Proc reg output with all the numeric data and one qualitative variable]

The REG Procedure
Model: MODEL1
Dependent Variable: Sleephours

Number of Observations Read	64
Number of Observations Used	64

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	8.28373	2.07093	2.60	0.0450
Error	59	46.96627	0.79604		
Corrected Total	63	55.25000			

Root MSE	0.89221	R-Square	0.1499
Dependent Mean	7.06250	Adj R-Sq	0.0923
Coeff Var	12.63306		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.11467	0.73413	12.42	<.0001
Age	1	-0.07364	0.03371	-2.18	0.0329
Screentime	1	-0.03981	0.04383	-0.91	0.3675
Gender_Male	1	-0.32309	0.24471	-1.32	0.1918
Gender_Other	1	-0.10041	0.54853	-0.18	0.8554

The REG Procedure
Model: MODEL1
Dependent Variable: Sleephours

Fit Diagnostics for Sleephours

Observations	64
Parameters	5
Error DF	59
MSE	0.796
R-Square	0.1499
Adj R-Square	0.0923

Residual by Regressors for Sleephours

Regression 3: All variables

When running a regression I chose to leave out my qualitative variable App because I were to put it into smaller categories, it would look identical to my variable "Activity". Just to clarify, my variable "App" contains the name of the specific apps people spent majority of the time. And the variable "Activity" is the category the main app the participants spent the most time on goes under. So it felt redundant to run a regression on the variable "App" for that reason.

Considering the assumptions for the regression, the data is normal when looking at the residual distribution, as it does have a clearly normal distribution, and the data is independent as one point has no way of influencing the other.

Looking a closer look at our outputs, it is an okay model but it is definitely the best model we have so far. The R squared values is 0.4121 which is okay. 41% of our sleeping time data can be explained by the variables we are testing against. Our R squared value needs to be closer to 1 to be a good model that predicts data so 0.4 is not so bad but it definitely could be better. When looking at our regression, you can see in the predicted values output that data is not on the line, so our model is not doing a good job of predicting values. Looking at the P values of our variables, Age is significant with a p value of 0.0385 since it is lower than our alpha or 0.05 which lets us know that we should keep this variable in our model. Similarly, S2 (the rating of "Very good" in sleep quality) is also significant since it has a low p value of 0.0057. Similarly S1 (the rating of " Good" sleep quality) is also significant with a low p value of 0.0556. But screentime, Gender_Male, Gender_other and the dummy variables for activity (A1, A2, A3, A4) are not significant as their p values were much higher than our alpha of 0.05, so it might make our model better if we removed these variables. And finally, looking at our residuals, there is no clear pattern which let's us know that it is a good fit. And there are a few outlier.

And finally, looking at our residuals for screentime and Age, there is no clear pattern which let's us know that it is a good fit. But when looking at the residuals for every other predictor there is a clear pattern which is not good. And similarly to our previous regression, when looking at the residuals for the regression there is a very clear pattern which indicates it is not a good fit, There are a few outliers but only one outlier that seems like it might be an influential observation. And I left that variable in as it was helping with my correlation but if I had collected the data better I would definitely get rid of this outlier as it is drastically influencing the model.There seems to be more outliers in this model than the rest, which makes sense since we have included more predicator.

Summary:

R Squared = 0.4121

Adj R Squared = 0.3011

Equation: Sleephours = 8.74379 - 0.06673 (Age) - 0.03153 (Screentime) – 0.49075 (S1) + 0.73305 (S2) + 0.52572 (A1) +0.24621 (A2) + 0.12918 (A3) - 0.27540(A4) - 0.22596 (Gender_Male) + 0.20373 (Gender_Other)

P values:

Age : 0.0385

Screentime: 0.4213

S1: 0.0556

S2: 0.0057

A1: 0.4285

A2: 0.3387

A3: 0.6875

A4: 0.4181

Gender_male: 0.3078

Gender_other: 0.7065

Best regression model

Comparing out three R squared values, we can see that when we performed a regression on all of the variables, we go an R squared value of 0.4121, which was the highest R squared compared to the other two regressions. This makes sense since this includes variables like Age and sleepquality (specifically our dummy variable S2) which were significant when looking at their p values which were lower that our presumed alpha of 0.05. And of course, we can further improve this model by removing the variables that were not significant. To further confirm our choice, I also noticed that our MSE of regression 3 (with all the variables), you can see that it is 0.1 less than the other models MSE, indicating that it is a better model.

**[Proc reg output with all the data]**

The REG Procedure
### Model: MODEL1
Dependent Variable: Sleephours

Number of Observations Read	64
Number of Observations Used	64

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	22.76643	2.27664	3.71	0.0008
Error	53	32.48357	0.61290		
Corrected Total	63	55.25000			

Root MSE	0.78288	R-Square	0.4121
Dependent Mean	7.06250	Adj R-Sq	0.3011
Coeff Var	11.08499		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.74379	0.67514	12.95	<.0001
Age	1	-0.06673	0.03144	-2.12	0.0385
Screentime	1	-0.03153	0.03891	-0.81	0.4213
S1	1	-0.49075	0.25071	-1.96	0.0556
S2	1	0.73305	0.25457	2.88	0.0057
A1	1	0.52572	0.65893	0.80	0.4285
A2	1	0.24621	0.25502	0.97	0.3387
A3	1	0.12918	0.31937	0.40	0.6875
A4	1	-0.27540	0.33747	-0.82	0.4181
Gender_Male	1	-0.22596	0.21944	-1.03	0.3078
Gender_Other	1	0.20373	0.53817	0.38	0.7065

Fit Diagnostics for Sleephours

Residual by Regressors for Sleephours
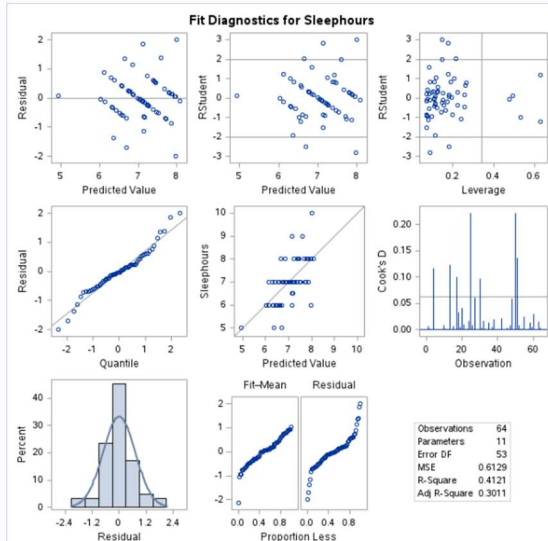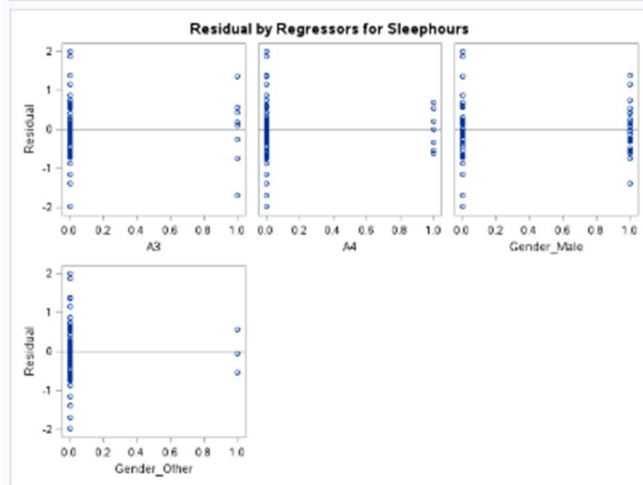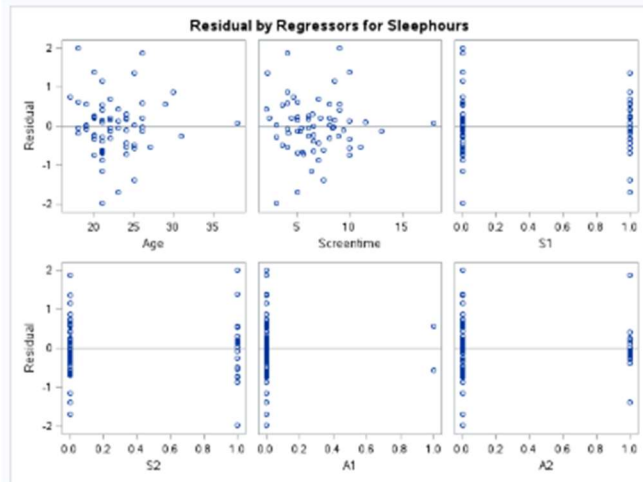
My survey:

Screen Time and Sleep Hours Survey

Hello, I am conducting a survey for my STAT430 class to collect data on young adults, specifically focusing on their screen time and sleep hours. This survey should take no more than 2 minutes, thank you for your time.
Here are instructions on how to get your screen time data:

For iPhone: Go to **Settings** > **Screen Time** > Click **"See All App & Website Activity"** > click on **"Week"**
For Android: Go to **Settings** > **Digital Wellbeing & Parental Controls** > Tap on **Dashboard** (click **STAT**)to view your screen time statistics.

If you have any questions, please reach out to me at bmekonne@umd.edu. Thanks again!

betelmekonnen@gmail.com Switch account

☒ Not shared

* Indicates required question

What is your age? *

Your answer

What is your Gender? *

○ Male

○ Female

○ Non-binary/other

○ Prefer not to say

How many hours of screen time do you average per week on your phone? (Check * your phone settings and type the number of hours averaged in the week **NOT** the total, **instructions are at the top**)

Your answer

What is the primary activity you use your phone for (according to your settings) *

○ Social Media

○ Messaging apps

○ Work/Study

○ Watching videos/streaming

○ Gaming

○ Other: _____

What app did you spend the most time on? *

Your answer

How many hours of sleep do you get per night? (Please type the number of hours) *

Your answer

How would you rate your overall sleep quality? *

○ Poor

○ Fair

○ Good

○ Very good

○ Excellent

Submit   Clear form

Code: