Emily Huang
efh34
No Partners

Final Project Report

Project Definition:

- What problem are you solving?

With the increase of technological innovations, individuals seemingly lack undisturbed and peaceful nights of sleep. One of the most fundamental reasons why this issue arises is because of the use of technology and its interruption of our natural sleep cycles. Widespread technological usage before the hours in bed can greatly impact the quality of sleep individuals get overnight. This project aims to solve the issue that extends from the relationship between technology use and sleep quality by using a data management system. By using this system, it can better analyze how much quality sleep individuals are going to get based on their screen times.

- What strategic aspects are involved? How does your project relate to the lectures/papers we discussed?

The project proposal uses various strategic aspects that are aligned with previously discussed lectures and papers. This project will take multiple incomes of data, such as screen time of individuals and sleep data. From the coursework and lessons, there are several aspects that can be incorporated into this project. Making a simulated dataset aligns with organizing and collecting data- something that has been incorporated through each lesson. The labs taught students not only how to organize and collect data, but also how to clean, test, and train. By using DataFrames, it allows the data to be cleaned and split. The topics taught in class also showed how to build a linear regression model, which will be beneficial to test the data points simulated in this project. It also allows the data to be analyzed for relationships and further predictions. Using a linear regression model also has to be evaluated, and that can be done using the techniques learned in class such as using mean squared error and R squared to make sure the project model is accurate. Overall, this project will implement many skills taught throughout the semester through Python coding and the use of its libraries, which are crucial to processing files and handling data, as well as being able to create a logical and dependable prediction model.

Novelty and Importance:

- Why is your project important?

This project is important because this issue has been ongoing and continues to impact more individuals as time goes on. Using technology, whether it is for personal time, academia, or work, has become the new standard. Especially after the pandemic, many people now rely on digital technology to complete their work, as seen in hybrid or virtual classes and remote or hybrid environments for work. In this generation, it is much easier to communicate with others through video chats and calls, texting, or social media. Dependencies on technology have made it harder for others to put devices down and focus on getting a good night's rest. Previously, sleep trackers would tell individuals to eat less before bed, to drink tea, or to block out light and noise to better their quality of sleep. This project not only addresses the issue of the lack of integration

of technology in sleep cycles but it also pinpoints when to sleep for each individual based on their usage.

- Why are you excited about it?

I'm extremely excited about the concept of this project because, like many others who grew up in the digital age, it's difficult to separate screen time from the time you need for adequate sleep. As a person who is constantly busy, I often find common ground with others who agree that when there is finally time for yourself at the end of the day, laying in bed and scrolling through your phone feels like needed relaxation. The issue with this is that it can be tricky to not have that alone time on your device seep into the necessary hours of sleep for health and function. I believe that this project is important and interesting because I deal with this situation myself, and I find that the majority of people do as well. I understand how frustrating it is to wake up sluggish and tired, and how no matter how long I sleep it feels inadequate. Relating on a personal level makes this project even more important.

- What are some existing issues in current data management practices? Are there any prior related works?

Some current issues in current data management practices such as using Apple Health include movements and physical barriers but lack the issue of combatting screen time for sleep. While it detects several factors such as heartbeat and how long the individual has slept, it does not fully incorporate all the digital screen time through all platforms. There have been studies that show screen time has an impact on sleep, such as it being due to the blue light being emitted, but they focus on raw data rather than combating the issue. There are also apps that track and manage screen time, as many phones have settings that are built in to do so. Despite all of these newfound developments, it would be beneficial for applications to merge the two (sleep tracking and screen time data) to help individuals manage their devices for their sleep cycle.

- What kind of data did you use? How did you get it? Will you create or simulate it?

This data will use simulated data that replicates real-world sleep patterns and sleep durations. By using simulated data it will allow the project to move forward to test and evaluate the prediction model. The simulated data would include screen time numbers primarily during the evening which would be based on typical screen time patterns and sleep duration with either a negative or positive correlation depending on screen time. This data will be able to correlate the quality of one's sleep to their usage of technology.

- What techniques/models/algorithms do you plan to use or develop?

For my simulated data, there would be a dataset that includes columns such as screen time in hours and sleep duration in hours. These will be randomly generated by either using random or numpy from Python, and that would contribute a couple of hundred data points. This project would benefit from using a linear regression model to predict sleep duration from screen time. This model is best to compare the relationship between two different variables and to develop a prediction model from the independent and dependent variables, which in this case would be screen time and length of sleep. To create this, it would be efficient to use pandas, matplotlib, and scikit-learn from Python. The project will implement an SQL database system to store and manage data, using a table such as screen_time to log usage for all devices and a sleep table such

as sleep_data to record the length of sleep. These queries would include ones such as calculating the average sleep for different screen times. This database structure will allow us to see similarities using SQL queries and data aggregation.

- What experiments did you design?

Throughout this project, there were a number of experiments that were designed. These experiments were to explore the relationship between screen time and sleep quality, as said in the proposal report. Firstly, I created a dataset of 1,000 records that included data on screen time, sleep quality, and an indicator if the individual did or did not experience bad sleep, which was if their sleep quality was below 5, giving me a controlled dataset. I then constructed a database setup that allowed me to manage and analyze the data using both SQLite and attempting to use MongoDB. I tried to implement MongoDB as taught in the last recitation, aiming to build a more complex and efficient project. This allowed me to do different queries and data manipulation, as well as testing within the database. I ran into some complications, so I made sure to write in the fallback into SQLite if MongoDB didn't work. I wrote a mix of basic and complicated queries to analyze the data as well, such as monthly statistics and averages. While doing data analysis, it involved calculating average, min, and max for screen time and sleep quality as well as evaluating correlations. For the machine learning section of the project, I built three models. One of which was a linear regression model to predict sleep quality from screen time. I also built a decision tree and a logistic regression model to analyze accuracies. My visualizations were built to communicate the findings of the data, including predicted versus actual sleep quality, box plots, and histograms for sleep quality distribution, scatter plots with a regression line for screen time versus sleep quality, and etcetera. These experiments allowed me to analyze and effectively use the data to build and communicate my findings.

- What are the key findings or results from your project? Did they verify or refute your original hypothesis? How did you evaluate your method?

The key findings or results of the experiments verified and broadened the hypothesis that was proposed in the first report- increased technology use does negatively affect the quality of sleep that individuals can get. The regression models' predictions and outputs of the visualizations further prove this hypothesis is correct. These were the outputs of the model:

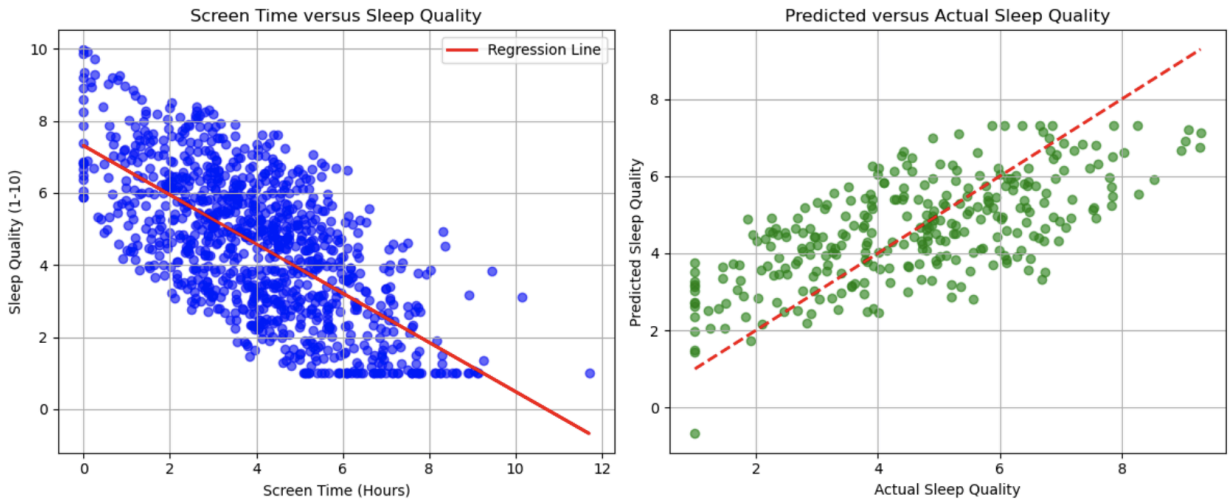
```
Sleep Quality Rank by Device:
DeviceType  SleepQualityRank
     Phone                 1
  Computer                 2
        TV                 3
    Tablet                 4
Monthly Sleep Quality and Screen Time:
      Month  PoorSleepPercentage  MaxScreenTime  MinScreenTime  TotalRecords
0   2023-01            70.967742       7.818833       0.024862            31
1   2023-02            45.000000       7.102304       1.028879            20
2   2023-03            66.666667       8.306365       1.139717            27
3   2023-04            72.000000       9.264764       1.244661            25
4   2023-05            60.000000      11.705463       0.787107            30
5   2023-06            62.962963       7.264823       0.000000            27
6   2023-07            52.500000       7.793586       0.000000            40
7   2023-08            56.521739       6.533822       1.239797            23
8   2023-09            59.259259       7.129287       1.103831            27
9   2023-10            65.714286      10.157762       0.000000            35
10  2023-11            59.375000       8.244312       0.000000            32
11  2023-12            56.000000       8.326509       0.593235            25
12  2024-01            63.636364       8.629317       0.000000            33
13  2024-02            73.333333       8.926484       0.000000            30
14  2024-03            64.864865       9.146720       0.000000            37
15  2024-04            52.173913       9.120169       0.258416            23
16  2024-05            48.571429       9.053865       0.676960            35
17  2024-06            59.375000       7.693415       0.000000            32
18  2024-07            66.666667       7.772372       0.000000            30
19  2024-08            48.275862       6.061999       0.579663            29
20  2024-09            48.484848       7.330949       0.442560            33
```

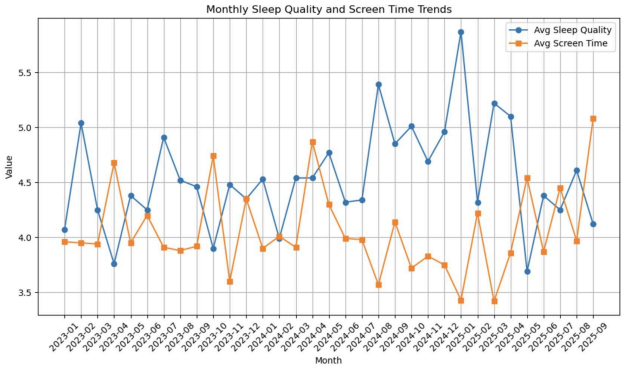
```
Device Screen Time and Sleep Quality:
  DeviceType  AvgScreenTime  AvgSleepQuality  MaxScreenTime  MinScreenTime
0   Computer       3.996503         4.530341       8.380911            0.0
1      Phone       3.905554         4.627389       9.264764            0.0
2         TV       4.124307         4.514353      10.157762            0.0
3     Tablet       4.157512         4.480015      11.705463            0.0

Screen and Sleep Quality Correlation by Device:
  DeviceType  CorrelationCoefficient
0   Computer                -0.6668
1      Phone                -0.6352
2         TV                -0.6451
3     Tablet                -0.6726

Monthly Sleep Quality and Screen Time Trends:
      Month  AVGSleepQuality  AVGScreenTime
0   2023-01             4.07           3.96
1   2023-02             5.04           3.95
2   2023-03             4.25           3.94
3   2023-04             3.76           4.68
4   2023-05             4.38           3.95
5   2023-06             4.25           4.20
6   2023-07             4.91           3.91
7   2023-08             4.52           3.88
8   2023-09             4.46           3.92
9   2023-10             3.90           4.74
10  2023-11             4.48           3.60
11  2023-12             4.35           4.35
12  2024-01             4.53           3.90
13  2024-02             3.99           4.01
14  2024-03             4.54           3.91
15  2024-04             4.54           4.87
16  2024-05             4.77           4.30
```

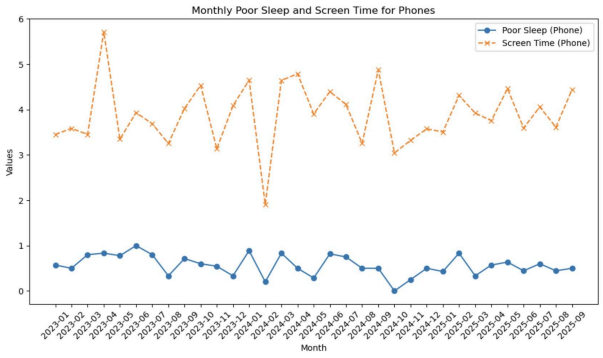
```
Mean Squared Error: 2.0623405522770795
Regression Equation for Sleep Quality = 7.31 + -0.68 * Screen Time
R^2 value: 0.4380
Decision Tree Accuracy: 0.63
Logistic Regression:
              precision    recall  f1-score   support

           0       0.67      0.62      0.64       126
           1       0.74      0.78      0.76       174

    accuracy                           0.71       300
   macro avg       0.71      0.70      0.70       300
weighted avg       0.71      0.71      0.71       300
```

As seen in the screenshot, the sleep quality is slightly higher with devices having a lower average screen time, as well as the correlation coefficients being negative (the strongest being computers at -0.6668). This implies that there is an inverse relationship between the two variables which both support the hypothesis that sleep quality is worse with higher screen times. The monthly screen and sleep quality trends were also printed, and in the snippet above these graphs display the maximum and minimum screen time per month along with sleep quality. While it is not perfectly linear, there is an inverse relationship between the two. The table shows the relative sleep quality of different devices, with phones being the highest ranking and tablets being the most negative on sleep quality, suggesting there might be an impact on sleep by device type. The MSE of 2.062 would indicate that the predictions from the model were accurate to the data, and the regression equation implies that for every hour of screen time, the sleep quality decreases by 0.68. The R-squared value of 0.4380 means that 43.80% of the variation can be explained by the screen time variable. This implies a strong correlation, meaning that screen time impacts sleep quality greatly. The decision tree accuracy is 0.63, which means it correctly predicted sleep quality 63% of the time, which is likely due to other factors not within the dataset. The logistic regression analyzed classes 0 and 1, 0 being good sleep and 1 being bad sleep. This means 67% of predictions for good sleep were good sleep and 74% of predictions for bad sleep were bad sleep. Recalling, which means the measure of real positives, had a high of 78% for bad sleep and a lower score of 62% for good sleep. The F1- score shows the balance between the two, with bad sleep being a higher percentage again. The overall accuracy of the model would be 71%, and the graph also provides the weighted and unweighted scores. As the score is 71% accuracy, the model would be reasonably correct at predicting. These many parts of the output would further help prove the hypothesis is correct.

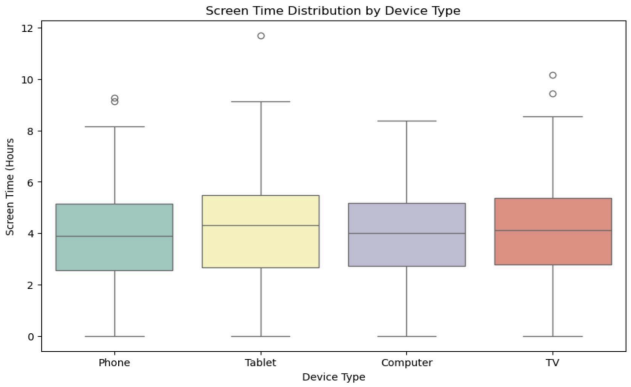
As seen in the visualizations below, these graphs show different ways of visualizing the data where there is an inverse relationship between screen time and sleep quality, as higher screen time is associated with lower sleep quality.

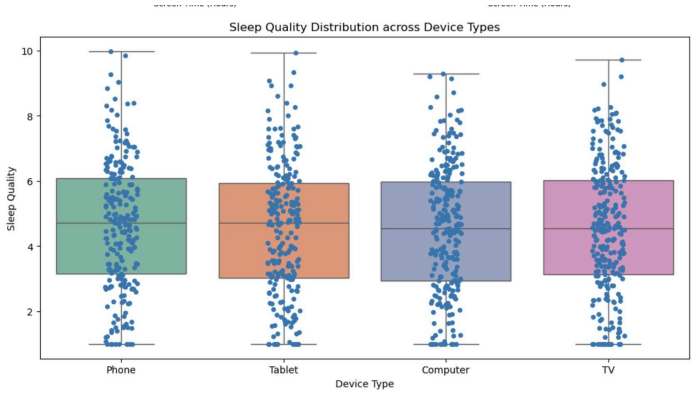






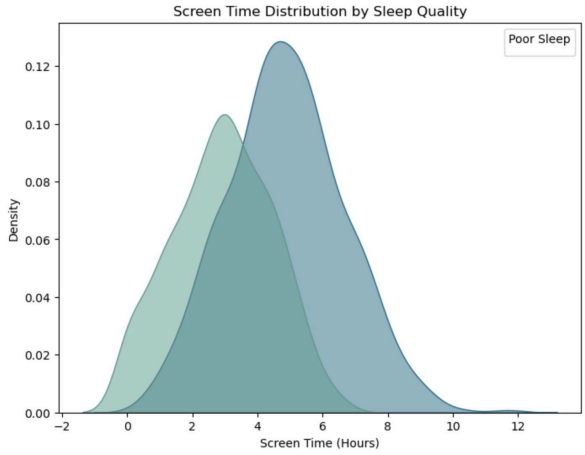
Visualizations #1 and #2 are both scatterplots with a regression line. Visualization #1 shows the inverse relationship between screen time in hours and sleep quality on a scale from 1-10. The red regression line emphasized the inverse relationship showing that on average an additional hour of screen time is correlated to declined sleep quality. The data has a clear downward slope, indicating that when screen time increases, sleep quality tends to decrease. Visualization #2 goes a bit further and represents the individual observations as the x-axis shows the actual sleep quality and the y-axis shows predicted sleep quality based on the regression model. Since the data points and regression line are well matched, it creates further possibility that the correlation is true for the observed and predicted data. Visualization #3 is a line chart that tracks the monthly averages for sleep quality and screen time, showing fluctuations in both variables over an extended period of time. Looking at this graph, the inverse relationship between the two shows through periods of high screen time corresponding to drops in sleep quality as well as the other way around. Visualization #4 goes into more specific detail, focusing on just phones rather than any other device. Though this is one out of four possible devices, the phone screen data is very similar to the overall screen time trends.

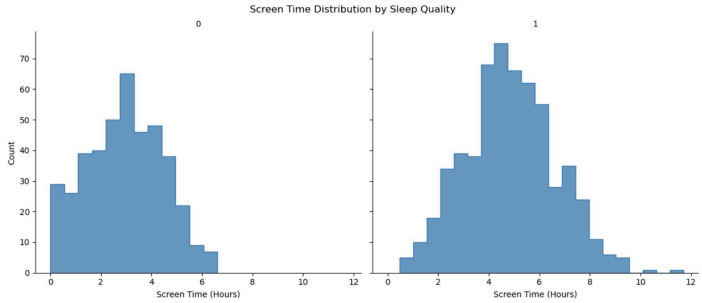




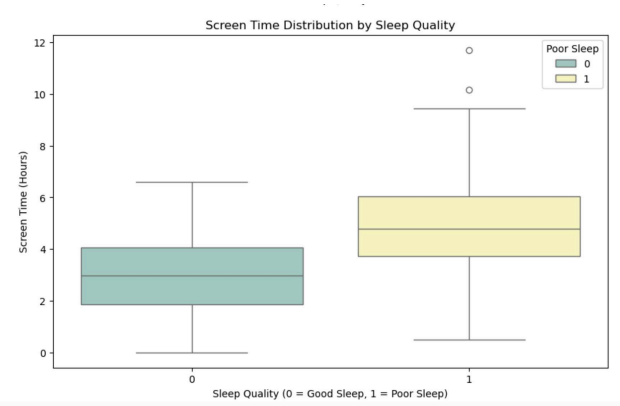
Visualizations #5 and #6 both show the screen time distribution across all four device types. I thought it would be beneficial to have two different graphs with one having scatter points to show the distribution of screen time usage across these devices. By implementing one visualization with a scatter plot, it allows us to see the individual data points and how accurately they match the actual sleep quality. This also shows a more visualized concept of the outliers and

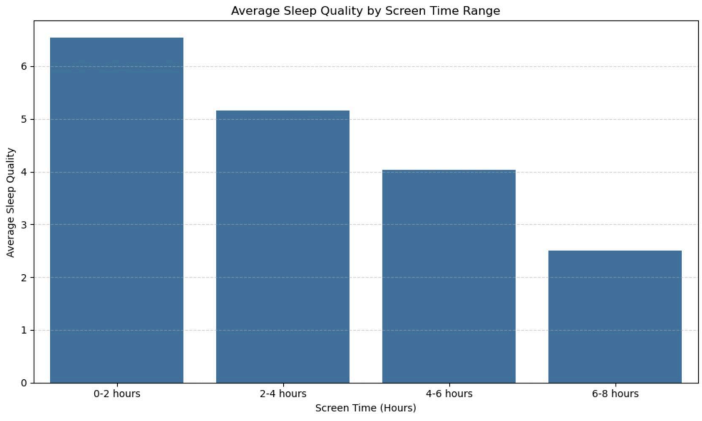
shows if it was more complex than linear as well as the other graph showing a more overall trend of the devices. The graph shows The average screen times and their distributions are all very similar, with tablet averages being a bit higher, and computer and TV's having narrower ranges. This suggests that device type is not the factor that primarily impacts poor quality sleep within individuals and that the type of screen is not a large indicator.





Visualizations #7 and #8 both depict the screen time distribution for sleep quality, As shown in the visualizations, the subset of those with higher screen times and poor sleep quality are at the right end of the distribution with a higher peak which shows that individuals who have poor sleep quality tend to have higher screen times on average.





Visualization #9 is another general overview of sleep quality and how it is distributed. The blue curve represents good sleep, peaking at around 6 on the quality scale, and the orange represents bad sleep, peaking at about 5- reinforcing the hypothesis. Visualization #10 is also another graph to show the relationship between sleep quality and screen time. The green graph represents good sleep with their sleep quality equalling 0, and the yellow graph represents bad sleep which equals 1. The average screen time of those with good sleep is around 3, while the ones with poor sleep average about 5 hours of screen time. The ones with lower screen times have better sleep, and the ones with higher screen times have poorer sleep, similar to the other visualizations previously. Visualization #11 displays the average sleep quality score across different screen time ranges: 0-2, 2-4, 4-6, and 6-8. As screen time increases, the average sleep quality decreases linearly and those with only 0-2 hours of screen time have better sleep (over 6 on the scale) than those with 6-8 hours of screen time (under 3 on the scale).

Overall, these visualizations suggest the same hypothesis as the first report- that individuals who have a higher amount of screen time are more likely to have much worse sleep quality than those who balance their screen time within a lower number of hours. The relationship between screen time and sleep quality are closely related as well as the predicted numbers being accurate to the actual values. This means that not only do the findings emphasize the hypothesis but also that the prediction model is mainly accurate to the actual numbers.

- Discuss the advantages and limitations of your approach.

Some advantages of this approach were that using simulated data made it a controlled and easier environment to manage and model the data without having to deal with real-world data issues. Using both SQLite and attempting MongoDB combines their strengths for enhancing query capabilities. By having more generic and complex queries, it easily showed many patterns with the data. Also being able to use not only the linear regression model but also the decision tree and logistic regression models helped to construct multiple dependable models to predict sleep quality from screen times. Using over ten visualizations helped highlight the findings and make them more clear to see, further proving the original hypothesis. The limitations of this approach were not being able to fully include the other factors that go into real-world screen time and sleep quality such as lifestyle habits. Since it is simulated data, it would not have the complexities that real-world data would have. As the project only covered monthly reports, there was no accuracy for specific days such as weekends or holidays, which can alter screen use.

- - Changes After Proposal: If your final report differs from your proposed project, discuss the differences, why you made certain changes, and the bottlenecks that prevented you from proceeding with the proposed project.

The proposal and the final code were relatively the same, except for a couple of changes. I discussed in the first report that I would focus mainly on nighttime screen habits, but I realized that it could overlook crucial insights for overall screen usage. I also had issues with implementing MongoDB. I wanted to be able to test myself to make the code more advanced, but I kept receiving errors, so I had to make sure it fell back to SQLite. Expanding the scope of my phone usage hours helped me understand how accruing screen time helps overall sleep quality. I also added other factors such as monthly sleep quality and type of device, which I thought could be beneficial to add for a more specific analysis. This would allow for more analysis of

device use habits and long-term patterns. These changes were made so that I could implement more things I've learned throughout the course.