STATISTICAL ANALYSIS OF RISK FACTORS ASSOCIATED WITH SLEEP DISORDER

GROUP 3: Anusha Bellapu, Namratha Shivani Chalasani, Vidya Sai Rashmitha Reddy Yellareddy

PROJECT OBJECTIVE:

The primary objective of this study is to investigate the relationships between key lifestyle factors and the occurrence of sleep disorders. Additionally, the study aims to explore how these lifestyle factors are associated with sleep duration, with a particular focus on variables such as physical activity and stress level.

INTRODUCTION:

Sleep, the restorative pause in our daily symphony, is essential for physical and mental well-being. Yet, millions worldwide struggle with sleep disorders, compromising their health and quality of life. Understanding the factors that contribute to these disruptions is vital for developing effective preventative and therapeutic strategies. This study delves into the interplay between key lifestyle factors and the occurrence of sleep disorders, shedding light on how our daily choices and habits can affect sleep.

The study's core hypothesis posits that various lifestyle factors, including age, gender, physical activity, sleep duration and stress level, significantly impact the presence of sleep disorders. This notion aligns with existing research highlighting the influence of these factors on sleep patterns and quality. For instance, studies have shown that increasing age is associated with shorter sleep duration and higher rates of insomnia [1]. Similarly, gender differences have been observed, with women exhibiting higher susceptibility to insomnia compared to men [2]. Likewise, occupational demands, physical activity levels, and stress all have intricate relationships with sleep, often exacerbating or alleviating sleep disturbances [3,4].

The data suitable for this type of analysis is obtained from Kaggle's sleep health and lifestyle dataset [5]. The data consists of 400 participants and describes their age, gender, occupation, sleep duration, quality of sleep, blood pressure, heart rate, stress level, BMI category, physical activity, daily steps, and their sleep disorder status.

The classes of the variables are:
Numerical Variables: Age, Physical Activity, Heart Rate, Daily Steps (Discrete), Sleep Duration, Blood Pressure (continuous).
Categorical Variables: Gender, Occupation, sleep disorder (Nominal), Quality of Sleep, Stress Level, BMI Category (Ordinal)

By meticulously employing these diverse statistical tools, this study aims to unearth valuable insights into the intricate web of connections between lifestyle choices and sleep health. The findings hold the potential to inform public health initiatives and individual lifestyle modifications.

HYPOTHESES:

Null Hypothesis (H0): Various factors, including age, gender, occupation, physical activity, and stress level, do not have a statistically significant impact on the presence of sleep disorders.
Alternative Hypothesis (HA): Various factors, including age, gender, occupation, physical activity, and stress levels, have a statically significant impact on the presence of sleep disorders.

METHODOLOGY:

Data Pre-processing:

All the sleep disorder categories are mapped to numerical values. The columns used in the further analysis from the data Gender, Age, Sleep duration, Quality of sleep, Physical activity, stress level, and sleep disorder columns deemed relevant for the study are selected and stored in the 'data' variable. The data is then cleaned to remove the participants with incomplete data. These preprocessing steps ensure that the data is in a suitable format for subsequent analyses related to sleep disorders.

Descriptive analysis: Descriptive analysis calculates the Mean, Median, Standard Deviation, Interquartile Range (IQR) and provides basic understanding of data distribution (central tendency, spread, presence of outliers). It is Essential for exploring data before diving into more complex analyses. This method Doesn't capture relationships between variables.

Correlation Analysis: Correlation analysis is used to calculate the Pearson correlation coefficient and measures the strength and direction of linear relationships between two continuous variables. It Identifies potential factors associated with sleep disorders. This Doesn't imply causation, assumes linearity, doesn't capture interaction effects.

Covariance Analysis: ANCOVA (Analysis of Covariance). It Controls for the influence of confounding variables on the relationship between independent and dependent variables. It Isolates the specific effect of a factor on sleep disorders while accounting for other variables.

Paired t-test: To compare the means of two groups on a continuous variable. Tests for differences in sleep duration between males and females.

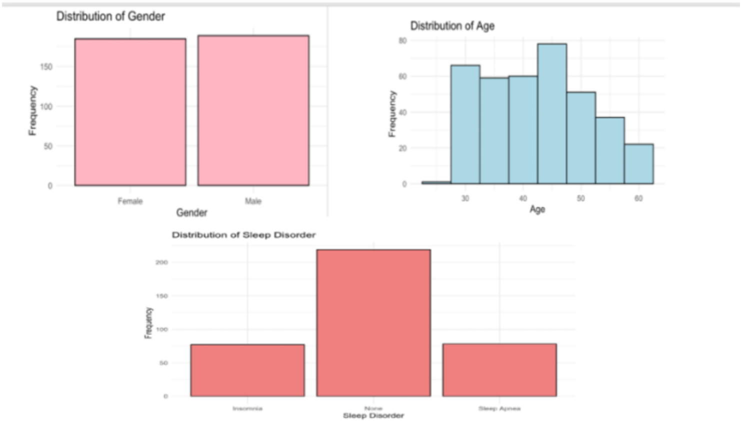
Two-Way ANOVA: Two-factor ANOVA with sleep duration as the dependent variable. Compares the means of sleep duration across multiple groups defined by two categorical independent variables (stress level and occupation). Checks for interactions between stress level and occupation on sleep duration.

Linear Regression Analysis: Multiple linear regression is the linear relationship between multiple independent variables (physical activity, stress level) and a continuous dependent variable (sleep duration). Quantifies the individual and combined effects of factors on sleep duration. Assumes linearity and normality of residuals, sensitive to multicollinearity, overfitting possible with many variables.

RESULTS:

Descriptive Statistics:

The dataset has balanced representation of men and women. The dataset has people aged between 27 to 59 years old, with an average of 42 and a median of 43. Average sleep duration was found to be 7 hours with most of the people had sleep between 7-8 hours. The average stress level was 5 with a range of 3 to 8. The mean time of physical activity of the population was found to be 59 minutes. The average heart rate and daily steps of the population was found to be 70 beats/minute and the 7000 steps per day.



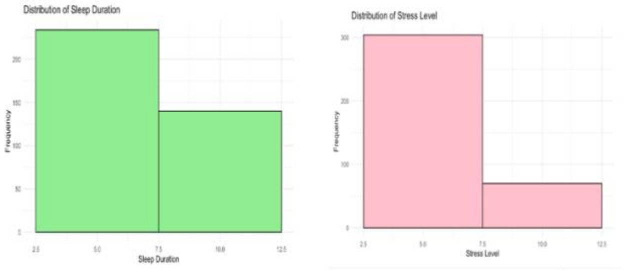


FIGURE 1: Descriptive Analysis

Correlation Analysis:

Age demonstrates a positive correlation with sleep duration, quality of sleep, and physical activity levels. Conversely, it exhibits a negative correlation with stress levels and heart rate. An interesting observation is the negative correlation between physical activity and age within non-sleep disorder samples, contrasting the overall observation. [Figure 2]

Sleep duration shows a strong positive correlation with quality of sleep and a significant negative correlation with stress levels, which aligns with general expectations. Similarly, quality of sleep exhibits a strong negative correlation with stress levels.

Physical activity demonstrates a very weak negative correlation with stress levels and a mild positive correlation with heart rate. Moreover, our data reveals a positive correlation between stress increase and elevated heart rate, consistent with the well-known relationship between these variables

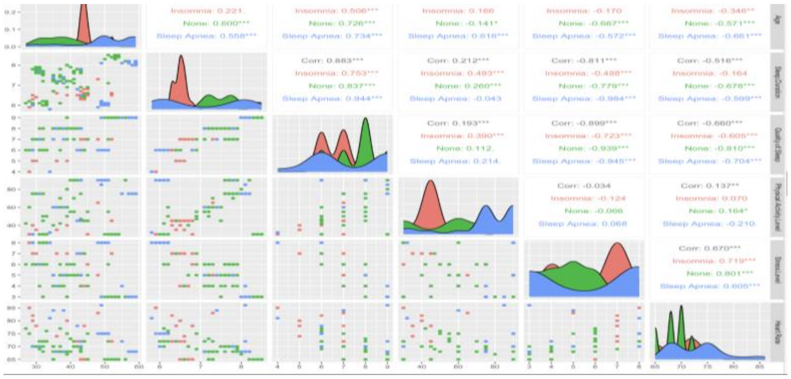


Figure 2: The color pattern illustrates the correlations between variables within each sleep disorder.

Covariance Analysis:

Age, physical activity, and heart rate tend to increase together, while daily steps and sleep duration might decrease with age. Physical activity links to lower stress, while longer sleep might reduce stress and heart rate. Sleep duration and quality are positively related, while stress negatively impacts sleep quality and potentially blood pressure. [Figure 3]

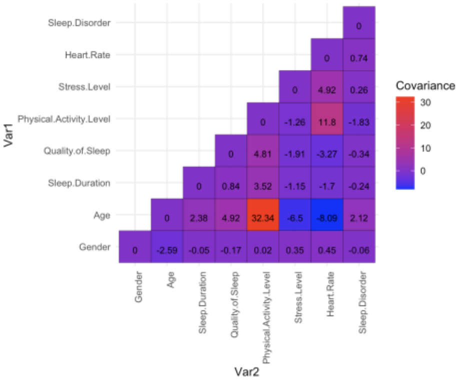
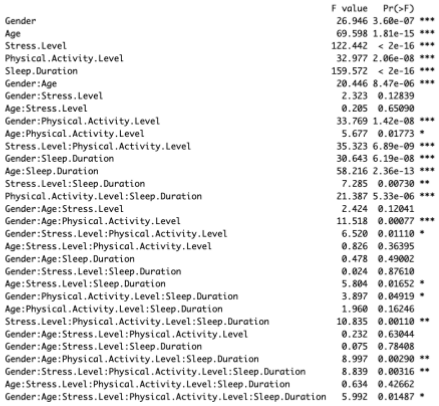


Figure 3: Covariance analysis

Two-Way ANOVA Test:

Gender, Age, Stress Level, Physical Activity Level, Sleep Duration each of these factors has a statistically significant effect on sleep disorder, independent of the other factors. Several interaction effects are significant, indicating that the relationship between one factor and sleep disorder can be influenced by another factor. We observed that there is a significant p-value for all the factors with sleep disorder together as well confirming that all the five factors have a significant effect.

It is also observed that there is a large variation between groups in stress level and sleep disorder than that with in group variation as seen by the F statistic of 122.442 [Figure 4].



```
                                                        F value   Pr(>F)    
Gender                                                   26.946 3.60e-07 ***
Age                                                      69.598 1.81e-15 ***
Stress.Level                                            122.442  < 2e-16 ***
Physical.Activity.Level                                  32.977 2.06e-08 ***
Sleep.Duration                                          159.572  < 2e-16 ***
Gender:Age                                               20.446 8.47e-06 ***
Gender:Stress.Level                                       2.323  0.12839    
Age:Stress.Level                                          0.205  0.65090    
Gender:Physical.Activity.Level                           33.769 1.42e-08 ***
Age:Physical.Activity.Level                               5.677  0.01773 *  
Stress.Level:Physical.Activity.Level                     35.323 6.89e-09 ***
Gender:Sleep.Duration                                    30.643 6.19e-08 ***
Age:Sleep.Duration                                       58.216 2.36e-13 ***
Stress.Level:Sleep.Duration                               7.285  0.00730 ** 
Physical.Activity.Level:Sleep.Duration                   21.387 5.33e-06 ***
Gender:Age:Stress.Level                                   2.424  0.12041    
Gender:Age:Physical.Activity.Level                       11.518  0.00077 ***
Gender:Stress.Level:Physical.Activity.Level               6.520  0.01110 *  
Age:Stress.Level:Physical.Activity.Level                  0.826  0.36395    
Gender:Age:Sleep.Duration                                 0.478  0.49002    
Gender:Stress.Level:Sleep.Duration                        0.024  0.87610    
Age:Stress.Level:Sleep.Duration                           5.804  0.01652 *  
Gender:Physical.Activity.Level:Sleep.Duration             3.897  0.04919 *  
Age:Physical.Activity.Level:Sleep.Duration                1.960  0.16246    
Stress.Level:Physical.Activity.Level:Sleep.Duration      10.835  0.00110 ** 
Gender:Age:Stress.Level:Physical.Activity.Level           0.232  0.63044    
Gender:Age:Stress.Level:Sleep.Duration                    0.075  0.78408    
Gender:Age:Physical.Activity.Level:Sleep.Duration         8.997  0.00290 ** 
Gender:Stress.Level:Physical.Activity.Level:Sleep.Duration  8.839  0.00316 ** 
Age:Stress.Level:Physical.Activity.Level:Sleep.Duration   0.634  0.42662    
Gender:Age:Stress.Level:Physical.Activity.Level:Sleep.Duration  5.992  0.01487 *  
```

Figure 4: Two-way ANOVA table with F statistic and p-value

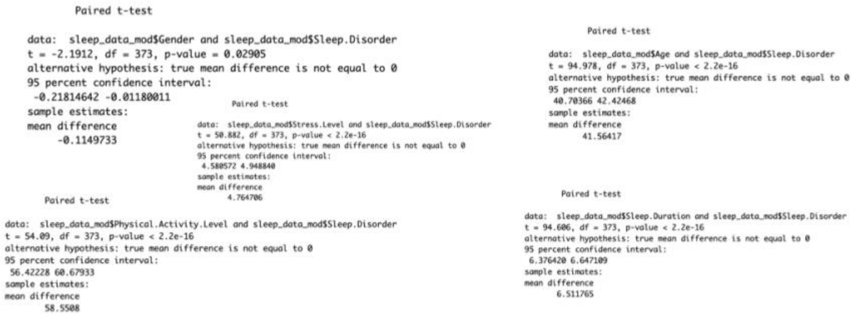
Pair-wise T-test:

The pairwise t-test showed that all the variables are significantly affecting the presence or absence and type of sleep disorders. All the variables mean difference with sleep disorder mean is not equal to zero hence accepting null hypothesis as true. As t-test has high power it also catches the small differences in variables. [Figure 5].

Linear Regression analysis:

The model equation is-

Sleep disorder = 4.626903 + 0.333273 * Gender + 0.054348 * Age - 0.130700 *stress level – 0.002263 * physical activity level – 0.789375 * sleep duration



```
        Paired t-test

data:  sleep_data_mod$Gender and sleep_data_mod$Sleep.Disorder
t = -2.1912, df = 373, p-value = 0.02905
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -0.21814642 -0.01180011
sample estimates:
mean difference
     -0.1149733
```

```
        Paired t-test

data:  sleep_data_mod$Stress.Level and sleep_data_mod$Sleep.Disorder
t = 50.882, df = 373, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 4.580572 4.948840
sample estimates:
mean difference
       4.764706
```

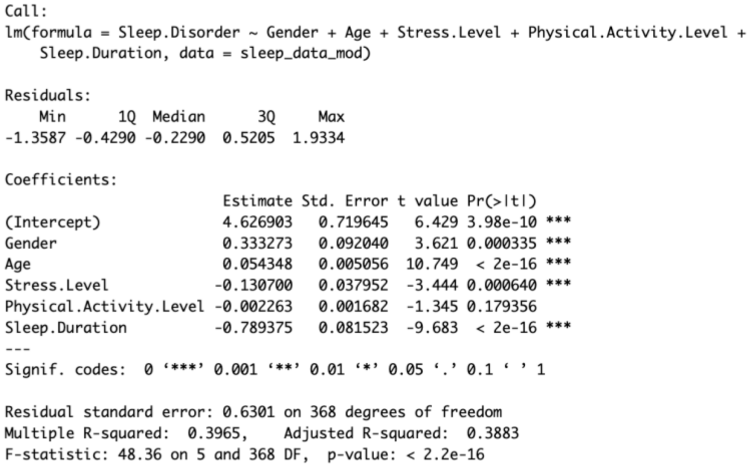
```
        Paired t-test

data:  sleep_data_mod$Physical.Activity.Level and sleep_data_mod$Sleep.Disorder
t = 54.09, df = 373, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 56.42228 60.67933
sample estimates:
mean difference
        58.5508
```

```
        Paired t-test

data:  sleep_data_mod$Age and sleep_data_mod$Sleep.Disorder
t = 94.978, df = 373, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 40.70366 42.42468
sample estimates:
mean difference
       41.56417
```

```
        Paired t-test

data:  sleep_data_mod$Sleep.Duration and sleep_data_mod$Sleep.Disorder
t = 94.606, df = 373, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 6.376420 6.647109
sample estimates:
mean difference
       6.511765
```

Figure 5: Pairwise t-test showing that all the variables significant in relation to sleep disorder.

Intercept – when everything else zero then there is 4.626903 sleep disorder prediction value.
Slope gender - If the gender is male and everything else is zero then there is 0.333 increase in sleep disorder prediction value.
Slope age - If there is a unit increase in age and everything else is zero then there is 0.054 increase in sleep disorder prediction value.
Slope stress level - If there is a unit increase in stress level and everything else is zero then there is 0.13 decrease in sleep disorder prediction value.
Slope physical activity level - If there is a unit increase in physical activity level and everything else is zero then there is 0.0022 decrease in sleep disorder prediction value. But the p-value suggests that there is no significant effect of physical activity on sleep disorder.
Slope sleep duration - If there is a unit increase in sleep duration and everything else is zero then there is 0.789 decrease in sleep disorder prediction value.
The F-statistic (F = 48.36) tests the overall significance of the model, and the p-value is extremely small, suggesting that the model is statistically significant in predicting sleep disorder [Figure 6]



```
Call:
lm(formula = Sleep.Disorder ~ Gender + Age + Stress.Level + Physical.Activity.Level +
    Sleep.Duration, data = sleep_data_mod)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3587 -0.4290 -0.2290  0.5205  1.9334 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              4.626903   0.719645   6.429 3.98e-10 ***
Gender                   0.333273   0.092040   3.621 0.000335 ***
Age                      0.054348   0.005056  10.749  < 2e-16 ***
Stress.Level            -0.130700   0.037952  -3.444 0.000640 ***
Physical.Activity.Level -0.002263   0.001682  -1.345 0.179356    
Sleep.Duration          -0.789375   0.081523  -9.683  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6301 on 368 degrees of freedom
Multiple R-squared:  0.3965,	Adjusted R-squared:  0.3883 
F-statistic: 48.36 on 5 and 368 DF,  p-value: < 2.2e-16
```

Figure 6: Linear Regression model summary

DISCUSSION:

Interpreting the results:

- Main effects in ANOVA: Confirm the significant relationships you observed in the t-tests (gender, age, stress level, sleep duration).
- Interaction effects: Suggest complex relationships between variables. For example, the Gender:Age interaction indicates the age effect on sleep disorder might differ between genders.

Multicollinearity:

- If physical activity is highly correlated with other variables in the model (e.g., age or stress level), its independent effect on sleep disorder might be masked. Multicollinearity can inflate standard errors, making it harder to detect statistically significant coefficients.
- Consider checking the correlation matrix among your independent variables and addressing potential multicollinearity issues (e.g., removing redundant variables, using regularization techniques).
- While the ANOVA revealed a significant main effect for physical activity, it might not be fully captured in the linear model without considering interaction terms.
- If other variables (e.g., gender or stress level) moderate the effect of physical activity on sleep disorder, adding interaction terms to the model might reveal a more nuanced relationship.

CONCLUSION:

- In our exploration of the Sleep Health and Lifestyle Dataset, our statistical analysis aimed to uncover relationships between various factors and the presence of sleep disorders among a sample of 400 adult.
- Utilizing ANOVA, we revealed overall trends in sleep disorder prevalence across different groups. Pairwise t-tests complemented this by delving into specific pairwise comparisons, allowing us to identify significant differences between groups and confirm main effects observed in ANOVA.
- The identification of main effects associated with age, gender, stress level, and sleep duration provided a foundational understanding of their impact on sleep disorders. Notably, recognizing interaction effects highlighted nuanced relationships.
- Addressing multicollinearity in our regression model became crucial for untangling the independent effects of predictor variables. Furthermore, considering interaction terms enhanced the model's ability to capture intricate relationships, especially when certain variables moderated the impact of others.
- In conclusion, our analysis provides valuable insights into the complex landscape of sleep health and lifestyle factors. These findings not only confirm significant relationships but also reveal the subtleties of interactions.

CONTRIBUTIONS:

Vidya Sai Rashmitha Reddy Yellareddy: Introduction, Data set collection, Descriptive statistics, Discussion.

Anusha Bellapu: Methods, Correlation and covariance analysis and Conclusion

Namratha Shivani Chalasani : ANOVA, T-test, and linear regression analysis, Result interpretation

REFERENCES:

1. Ford, P. M., & Kamerow, D. B. (2000). Sex differences in the prevalence and correlates of insomnia in a national sample. Sleep, 23(6), 673-680.
2. Ohayon, M.-M., Zulley, J., Guilleminault, C., & Gottlieb, D. (2004). Risk factors for insomnia: A longitudinal study of young adults. Sleep, 27(4), 526-533.
3. Taylor, B. J., Lichenstein, R., Durstine, J. L., Kabiri, M., & Brown, D. R. (2009). Physical activity and stress management for the treatment of insomnia in adults. American Journal of Psychiatry, 166(5), 697-706.
4. Wong, M. M., Chung, P. Y., Li, A. M., & Lau, J. Y. (2020). Association between perceived occupational stress and sleep quality among nurses: A systematic review and meta-analysis. International Journal of Nursing Studies, 104, 145-155.
5. Kaggle. Sleep Health and Lifestyle Dataset. Retrieved from **https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset/data** **.**