The Effects of Sleep on Cognitive Performance

Group 15

11-5-25

Abstract

In this project we seek to analyze how sleep related behaviors and lifestyle factors have a profound effect on a subject's cognitive performance. More broadly, the analysis investigates whether cognitive performance can be reliably predicted using everyday habits such as sleep, stress, caffeine consumption, physical activity, and reaction-time measures. By comparing a simple model (sleep duration only) to a full multivariable model, we determined both the unique contribution of sleep and the relative importance of other factors. Our primary question was whether the surrounding attributes of a person's sleep (hours slept, caffeine intake, etc.) have a direct effect on a persons ability to perform throughout the day? In a simple linear regression model, our predictive power was rather weak. Developing a full model through multiple regression using backwards selection, we obtained an adjusted R-squared value of 0.8664. Our data analysis yielded that Age and Gender were the least statistically significant influences (which is good to hear as they are not very easy to control), and displayed that we can reasonably control our cognitive performance with our sleep and lifestyle choices.

1 Introduction

Every student follows a daily routine, employing a unique strategy to prepare for mental engagement in school. However, not all of these strategies are equal and performance varies among individuals in both grades and in-class comprehension. Thus, in pursuit of sturdy academic conduct, it is vital to understand how sleep-related habits and lifestyle factors influence their activity. Specifically, in what way do the attributes of one's sleep, along with behavioral and physiological variables, predict cognitive performance?

This analysis may benefit researchers, students, and health professionals interested in understanding how daily behaviors interact to affect cognitive functioning. It also provides insight for individuals seeking to optimize academic or work performance, as the results clarify which aspects of lifestyle matter most. While the project includes hypothesis testing (for example, whether sleep significantly predicts cognitive score), the broader goal is predictive modeling and identifying which variables best explain cognitive performance. This helps determine whether a linear model can reliably estimate cognitive ability from behavioral data and highlights which factors have the strongest practical impact.

2 Data

The data set in use is "human_cognitive_performance.csv" pulled from Kaggle, which contains the observed attributes of individuals', aged 18 to 59, sleep and lifestyle, and their results in cognitive performance tests. The original data set contained 80,000 entries with 11 observational variables. 3 of the variables were measures of cognitive performance, while 8 were the aforementioned attributes observed of the individuals. As our interests pertains to students, we extracted only the results of individuals aged 18 to 25 (a reasonable age for undergraduate and masters students). Furthermore, we removed two of the variables which

scored cognitive performance such that our data set fed directly into building the linear regression model. The resulting data frame presents 1 dependent variable in “Cognitive_Score” and 8 independent in “Age”, “Gender”, “Sleep_Duration”, “Stress_Level”, “Diet_Type”, “Daily_Screen_Time”, “Exercise_Frequency”, “Caffeine_Intake”, and “Reaction_Time”.

3 Visualization

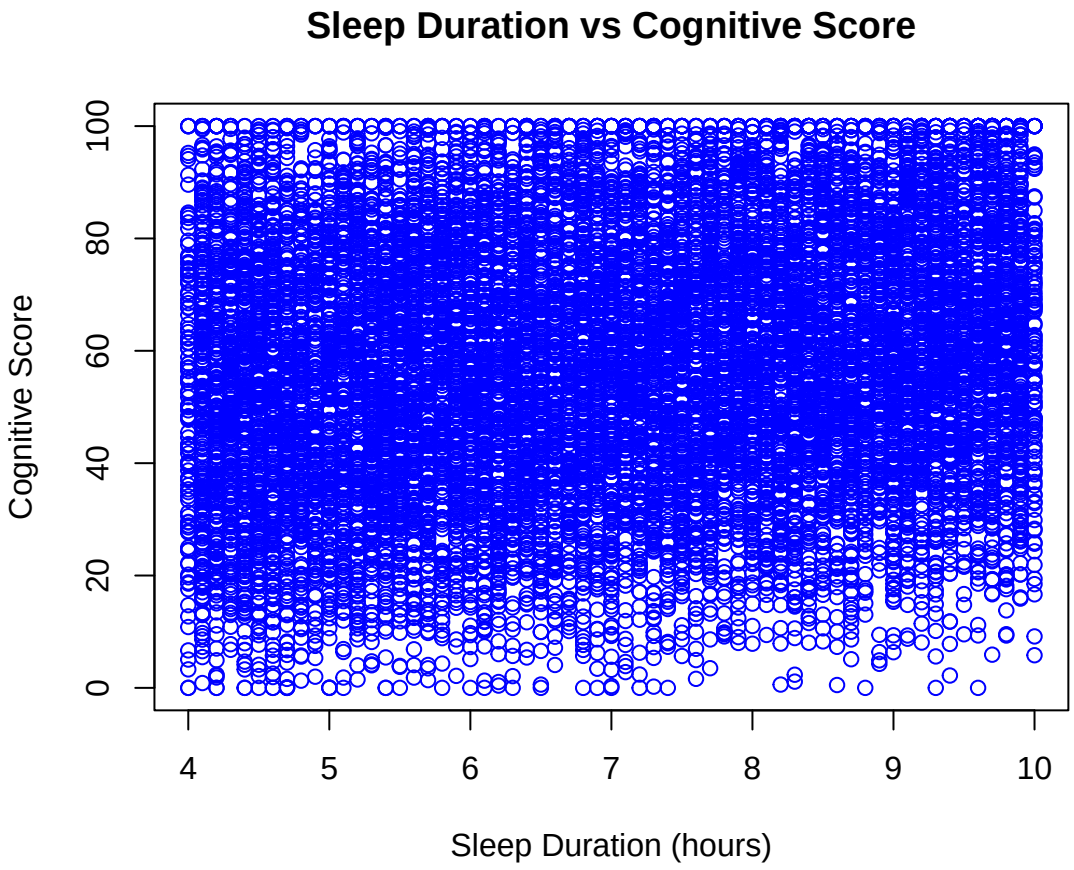

Figure 1: Figure 0: Sleep Duration plotted against Cognitive Score

Figure 0 is a plot of solely the hours of sleep an individual got versus their received cognitive score. From the density of the points, a weak positive linear relationship can be made out.

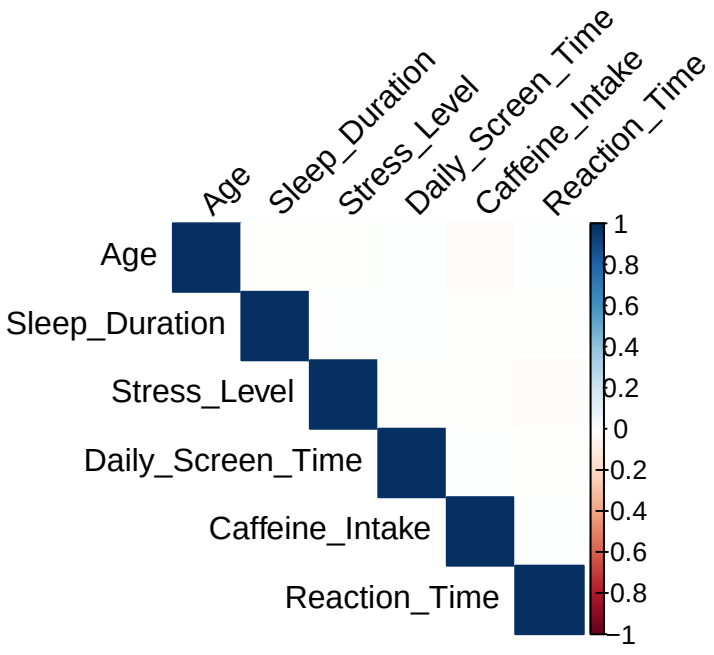

Figure 2: Figure 1: Heatmap of pairwise correlations between key variables.

Figure 1 shows the correlation plot of the numeric variables we intend to use to predict Cognitive Score. As indicated by the lack of color, there is no significant coorelation between any of the numeric independent variables. This is a sign that there is not any collinearity within our predictive variables which will allow for better model performance and independence in said variables.

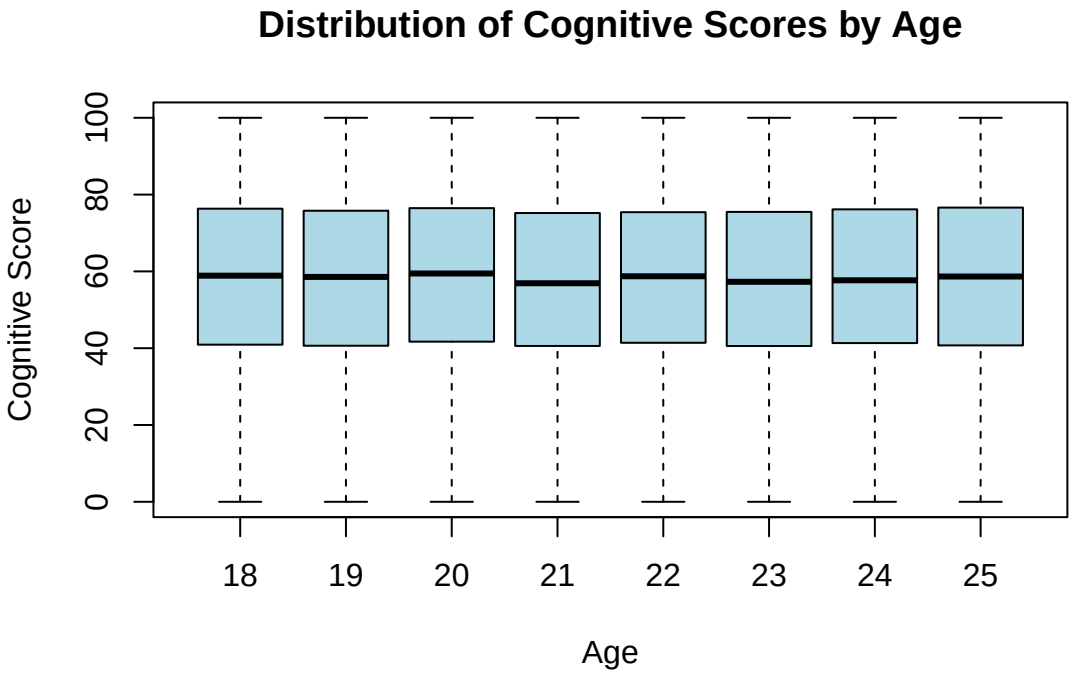

Figure 3: Figure 2: Boxplots showing cognitive scores by age.

Figure 2 displays the box plot distribution of measured cognitive scores of each age. These boxplot graphs are an effective way to visually examine the spread of Cognitive Score among each age. we can see that each distribution appears normal, with roughly the same mean, so it is unlikely that Age will be much of an influential factor in determining Cognitive Performance.

4 Analysis

As a baseline, we attempt simple linear regression to predict cognitive score based on only sleep duration.

```
##
## Call:
## lm(formula = Cognitive_Score ~ Sleep_Duration, data = sleep_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -63.43 -17.24  -0.18  17.77  47.81
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     44.1648     0.7754   56.95   <2e-16 ***
## Sleep_Duration   2.0068     0.1074   18.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.9 on 15084 degrees of freedom
```

```
## Multiple R-squared:  0.02264,    Adjusted R-squared:  0.02257
## F-statistic: 349.4 on 1 and 15084 DF,  p-value: < 2.2e-16
```

The resulting equation:

$$\widehat{\text{Cognitive Score}} = 44.1648 + 2.0068 \cdot \text{Sleep Duration}$$

From this we gain insight with the y intercept being equal to 44.1648; the predicted Cognitive Score when the individual got no sleep. The slope is 2.0068 which we can interpret as fan increase of about 2 points in predicted Cognitive Score for every additional hour of sleep. Now for the significance of this predictor we observe a value of <2e-16, less than 0.05. From this we are able to conclude that this sleep duration variable in our data set is a statistically significant predictor of the cognitive score. Although this is a statistically significant predictor the variance explained by this variable is shown through the adjusted R-squared value: 0.02257. From this, sleep duration only explains about 2.3% of the variability in cognitive score. Thus, although a significant predictor of cognitive score, it is not sufficient to fully predict cognitive score alone.

As we seek to determine which variables are of statistical importance in predicting an individual's performance, we create a full model containing all 8 independent variables and use backwards selection, eliminating variables based on which model produces the lowest AIC score, to determine an optimal multiple linear regression model.

```
## Start:  AIC=64467.28
## Cognitive_Score ~ Age + Gender + Sleep_Duration + Stress_Level +
##     Daily_Screen_Time + Exercise_Frequency + Caffeine_Intake +
##     Reaction_Time
##
##                       Df Sum of Sq     RSS   AIC
## - Gender               2       128 1081113 64465
## - Age                  1       104 1081089 64467
## <none>                              1080984 64467
## - Caffeine_Intake      1    108792 1189776 65912
## - Sleep_Duration       1    173127 1254112 66706
## - Daily_Screen_Time    1    298911 1379895 68148
## - Stress_Level         1    464982 1545966 69863
## - Exercise_Frequency   2    515169 1596153 70343
## - Reaction_Time        1   5425362 6506346 91543
##
## Step:  AIC=64465.07
## Cognitive_Score ~ Age + Sleep_Duration + Stress_Level + Daily_Screen_Time +
##     Exercise_Frequency + Caffeine_Intake + Reaction_Time
##
##                       Df Sum of Sq     RSS   AIC
## - Age                  1       105 1081218 64465
## <none>                              1081113 64465
## - Caffeine_Intake      1    108722 1189834 65909
## - Sleep_Duration       1    173105 1254218 66704
## - Daily_Screen_Time    1    299096 1380209 68148
## - Stress_Level         1    465337 1546449 69863
## - Exercise_Frequency   2    515125 1596238 70339
## - Reaction_Time        1   5425333 6506445 91539
##
## Step:  AIC=64464.54
## Cognitive_Score ~ Sleep_Duration + Stress_Level + Daily_Screen_Time +
##     Exercise_Frequency + Caffeine_Intake + Reaction_Time
```

```
##
##                       Df Sum of Sq     RSS   AIC
## <none>                              1081218 64465
## - Caffeine_Intake      1    108644 1189862 65907
## - Sleep_Duration       1    173183 1254401 66704
## - Daily_Screen_Time    1    299226 1380443 68148
## - Stress_Level         1    465243 1546461 69862
## - Exercise_Frequency   2    515150 1596368 70339
## - Reaction_Time        1   5425586 6506804 91538
```

```
##
## Call:
## lm(formula = Cognitive_Score ~ Sleep_Duration + Stress_Level +
##     Daily_Screen_Time + Exercise_Frequency + Caffeine_Intake +
##     Reaction_Time, data = sleep_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.9292  -7.2430  -0.0589   7.1638  15.8997
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.426e+02  4.592e-01  310.52   <2e-16 ***
## Sleep_Duration             1.951e+00  3.970e-02   49.14   <2e-16 ***
## Stress_Level              -1.947e+00  2.418e-02  -80.55   <2e-16 ***
## Daily_Screen_Time         -1.409e+00  2.182e-02  -64.60   <2e-16 ***
## Exercise_FrequencyLow     -1.466e+01  1.892e-01  -77.47   <2e-16 ***
## Exercise_FrequencyMedium -4.860e+00  1.880e-01  -25.86   <2e-16 ***
## Caffeine_Intake           -1.856e-02  4.767e-04  -38.92   <2e-16 ***
## Reaction_Time             -1.645e-01  5.982e-04 -275.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.468 on 15078 degrees of freedom
## Multiple R-squared:  0.8665, Adjusted R-squared:  0.8664
## F-statistic: 1.398e+04 on 7 and 15078 DF,  p-value: < 2.2e-16
```

After fitting the full model with six predictors, eliminating Age and Gender, the summary gives us a picture of how these variables together help explain differences in Cognitive Score. The adjusted R-squared value of the model is 0.8664, meaning that 86.64% of variance in cognitive score is able to be explained by the model. The model also has an F-statistic of 1.398e+04, corresponding to a p-value of 2.2e-16, conveying statistical significance in its predictive power.

The model estimates a starting point (intercept) of about 142.7 for the predicted Cognitive Score when all predictors are zero. From there, each predictor shifts the cognitive score according to it's respective coefficient. Sleep Duration has a positive coefficient of about 1.92, meaning that each additional hour of sleep is associated with nearly 2 more points on the cognitive score, holding everything else constant. Stress Level goes the opposite way, decreasing cognitive score by about 1.93 points for each unit increase in stress.

Other daily habits also contribute. More time spent on screens predicts lower scores, dropping by about 1.45 points for each hour. Caffeine Intake also has a very small but negative association, where increasing caffeine slightly lowers predicted performance. Reaction Time has a larger negative impact: slower reaction times correspond to noticeably lower cognitive scores, which makes sense because reaction time is itself tied to attention and cognitive processing. The exercise frequency categories also show meaningful differences. Compared to the reference group (likely "High" exercise), people who exercise "Low" score about 14 points

lower, and those in the "Medium" group score about 5 points lower. This suggests that physical activity plays a noticeable role in cognitive outcomes for this data set.

After fitting the full model we want to be able to run a few test on this model and be able to determine if our data set contains proper linearity, spread of residuals, normality of residuals, influential points, and homoscedasticity.

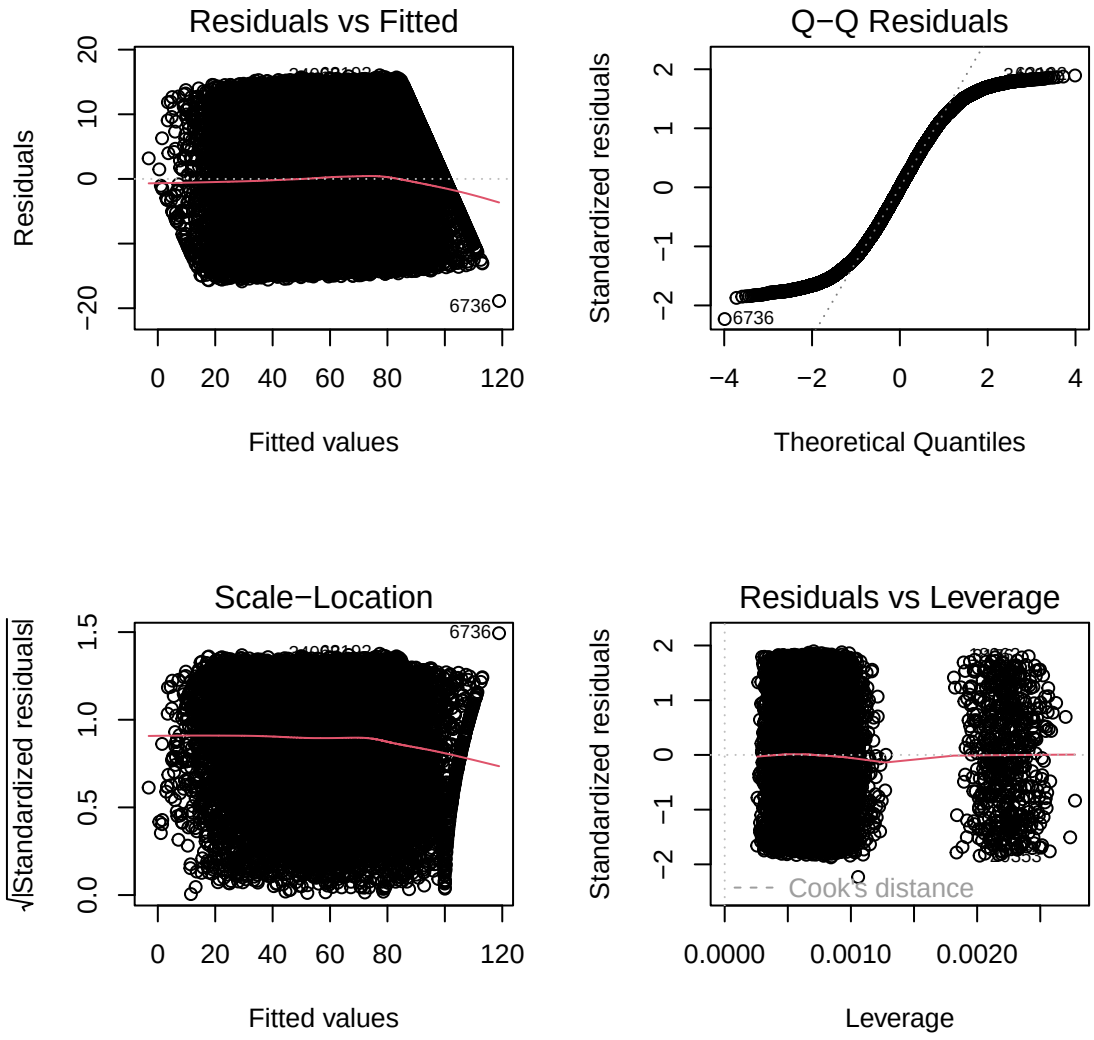

Figure 4: Figure 4: Residual diagnostics for the multiple linear regression model predicting Cognitive Score

The first plot residuals vs fitted values of the full model checks for linearity meaning that it determines whether the relationship between the predictor variables and the response variable follows a true linear relationship. What our data shows is that most observations fall between +15 and -15. This plot does allow us to observe a certain flaw: Instead of a random cloud of points, there is a clear pattern of diagonal lines being shown in the residuals. This is likely due to our response variable not being entirely continuous.

The second plot, Q-Q plot of Residuals, allows us to check for normality of the residuals. If the points do follow a normal residual distribution then these points are to form a relatively straight line. In our data we observed an s shaped plot with heavy tails on each end. This is due to the nature of the test scores, in that

they are bound to the interval of [0, 100].

The third plot is Scale-Location Plot which plots Standardized Residuals versus Fitted. This plot is utilized to check for homoscedasticity or the constant variance of residual. Our data set shows that most points are centered at around 0.7 and almost all points falling plus or minus 0.7 of that center. Again we see one of the three reoccurring outlines showing up in this plot. This shows us that the variance is constant with the exclusion of a few outliers in the top right. Here we are able to conclude that the constant variance of residuals is satisfied.

The Residuals vs leverage plot allows us to determine whether our data set contains influential points that may cause a disproportional affect on the model. We see here that most observation have a minimal leverage except for the three major outliers that hold extreme leverage over the rest of our data point in the model. This allows us to see why are model might have some distortion within their coefficients that are affecting the full models normality and residuals.

Overall, the conditions for multiple linear regression are not satisfied, and a linear model is not appropriate for this data.

5 Conclusion