Elijah Chan
STA2260-6

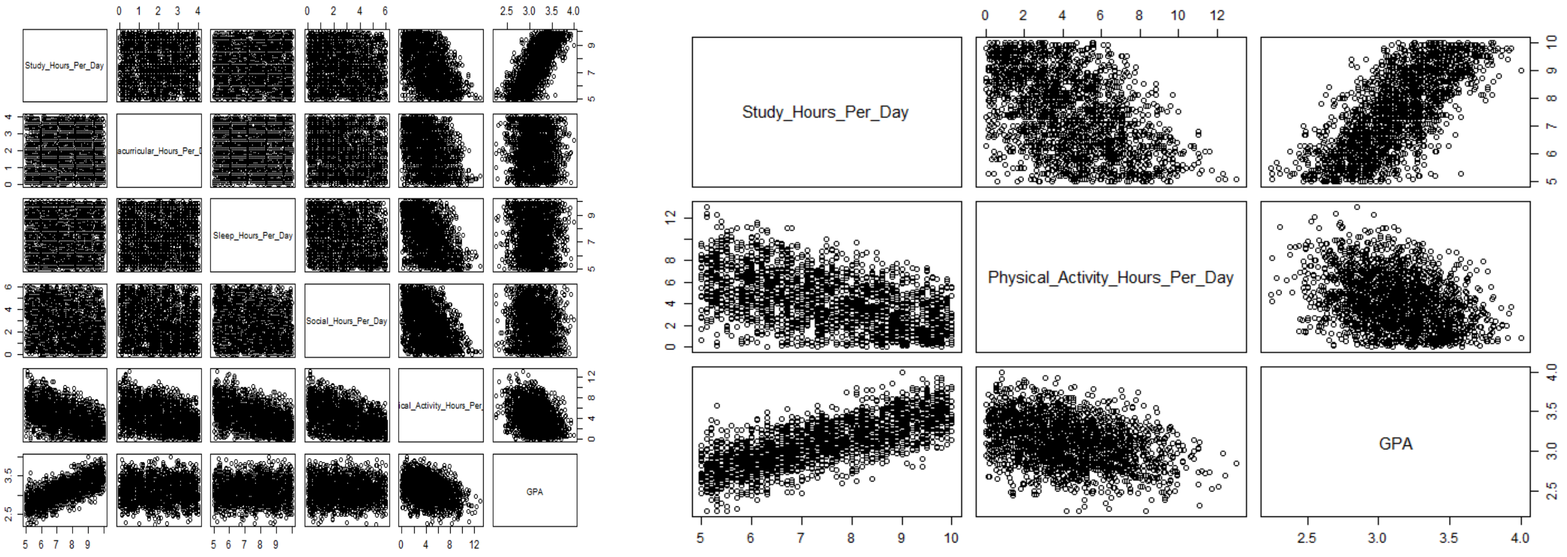
STA2260 Final Project

For this project, I will be looking into many lifestyle factors that affect a student's academic performance. I used a dataset from kaggle (linked below) that gave 2000 accounts of students and their many lifestyle habits: studying hours, extracurricular hours, sleeping hours, social hours, physical activity hours. It also gives their stress levels and GPA. I will primarily be focusing on the GPA as the response variable for the models as it is a better interpretation of academic performance. Stress level, according to the dataset report, is just a derivative of study and sleep hours so I will be ignoring it for this report.

Dataset I used: student lifestyle dataset

1. Finding Relevant Factors

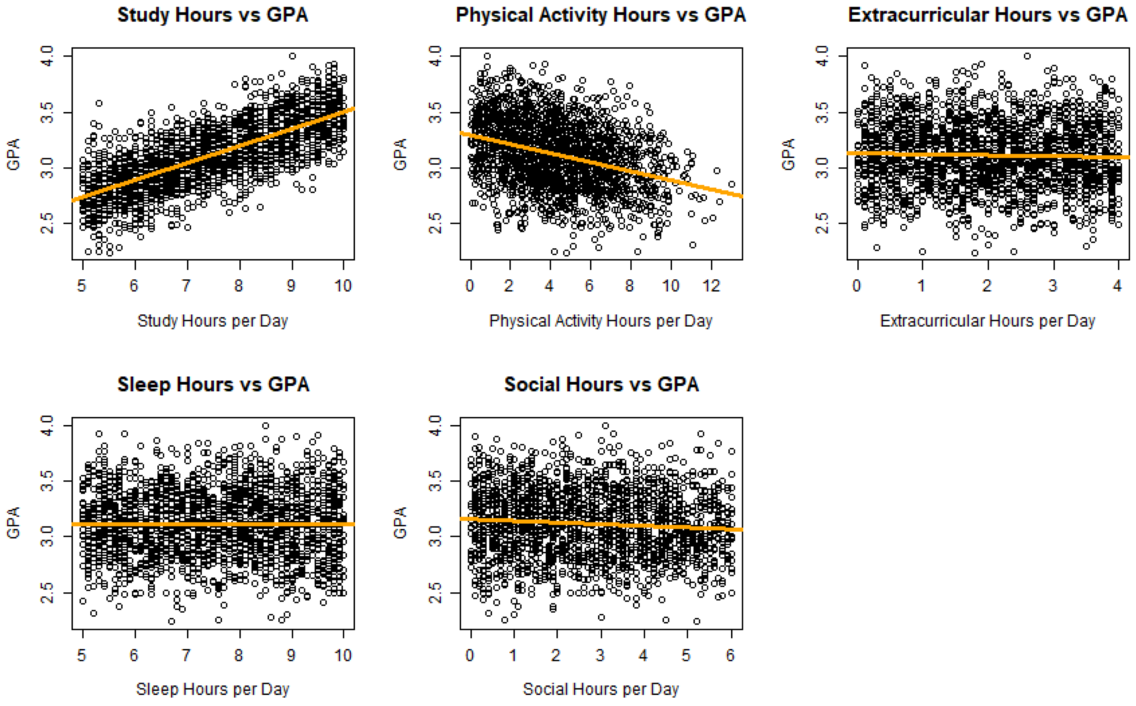
I first wanted to see an overview of all the factors. Many of the plots were just a giant cluster of dots filling in the entire plot so I knew those were not the factors I should be focusing on. I cut it down to just 2 main factors that showed an obvious relationship with GPA: study hours per day & physical activity hours per day.



Left: overview look of all the plots including irrelevant ones.
Right: the two most relevant factors to GPA with trends.

2. Building Model 1

I first compared all 5 relevant factors and plotted them against GPA with a linear regression.



The two relevant factors I found previously held true as the others three didn't really seem to be good at determining GPA. I then further looked into this and compared the adjusted R^2 value and AIC of each of these variables.

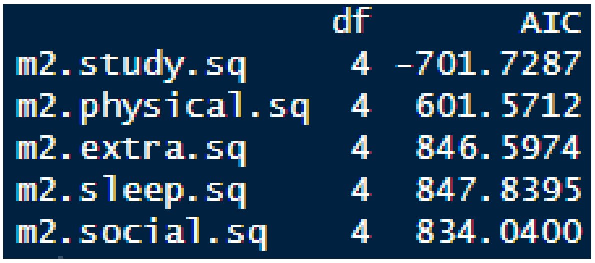
Variable	adjusted R^2 value	AIC
Study Hours per Day	0.5392	-703.5010
Phy. Activity Hours per Day	0.1159	599.6696
Extracurr. Hours per Day	0.0005352	845.0659
Sleep Hours per Day	-0.0004822	847.1007
Social Hours per Day	0.006844	832.4020

From both the graphs and the values for R^2 & AIC, I concluded that using Study Hours per Day as the predictor was the best as it has the lowest AIC and the highest R^2 value. From the previous segment, I found that the top two most correlated were study hours and physical activity. Looking at the numbers, though, show just how significantly different the two are in terms of accuracy. In the end, study hours also seems the most compact, in terms of visualization, and best represents the relationship between itself and GPA, according to the R^2 & AIC values.

3. Building Model 2

For the first new model, 2.1, I decided to use a quadratic relationship to represent study hours to GPA. It uses this equation:

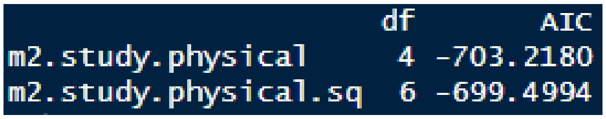
$$y = \beta_0 + \beta_1 \times x + \beta_2 \times x^2 + \varepsilon$$



```
                   df       AIC
m2.study.sq        4  -701.7287
m2.physical.sq     4   601.5712
m2.extra.sq        4   846.5974
m2.sleep.sq        4   847.8395
m2.social.sq       4   834.0400
```

Variable	adjusted R^2 value
Study Hours	0.539
Phy. Activity Hours	0.1155
Extracurr. Hours	0.0002689
Sleep Hours	-0.0003522
Social Hours	0.006526

For the second model, 2.2, I decided on using multi regression with the top two correlating factors, study hours and physical activity hours. I modeled it with both a standard linear and a quadratic relationship.



```
                      df       AIC
m2.study.physical      4 -703.2180
m2.study.physical.sq   6 -699.4994
```

Variables	adjusted R^2 value
Multi Linear	0.5394
Multi Quadratic	0.539

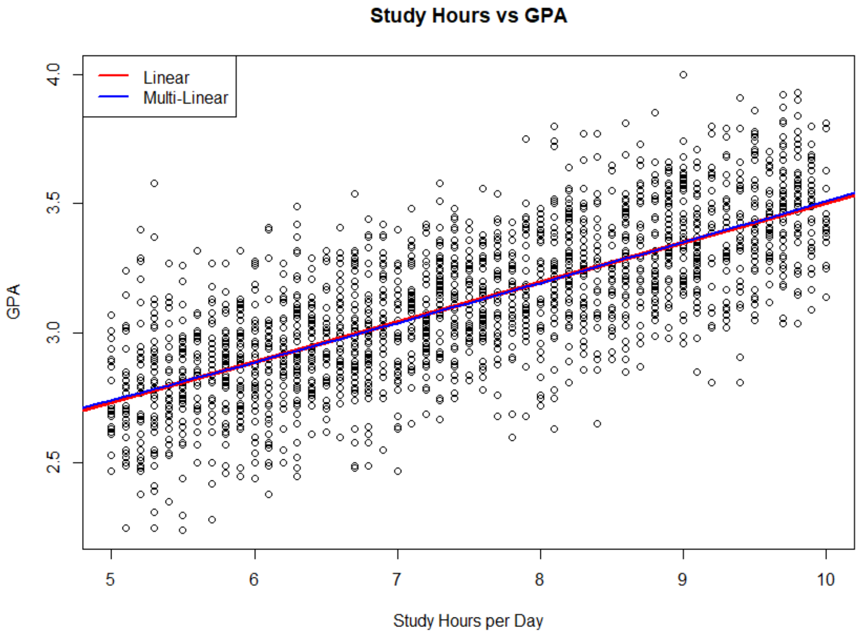
I then collected the relationship with the best adjusted R^2 value & AIC from both models.

Model	adjusted R^2 value	AIC
2.1 Quadratic (study)	0.539	-701.7287
2.2 Multi (linear)	0.5394	-703.2180

With the information from the table above, Model 2.2 has the higher R^2 value and the lower AIC value. Decisively, I concluded that Model 2.2, the multi-linear regression, was better.

4. Comparing Model 1 & Model 2

Now for the final step, I first wanted to visually compare them to see what I was working with. I plotted both the linear and multi-linear model onto one graph.



As I could interpret from this graph, both models were extremely close as they were basically stacked right on top of each other with only tiny differences between them. As a result, this meant that I needed to look at the numbers in order to better understand which model was the most accurate representation of the data.

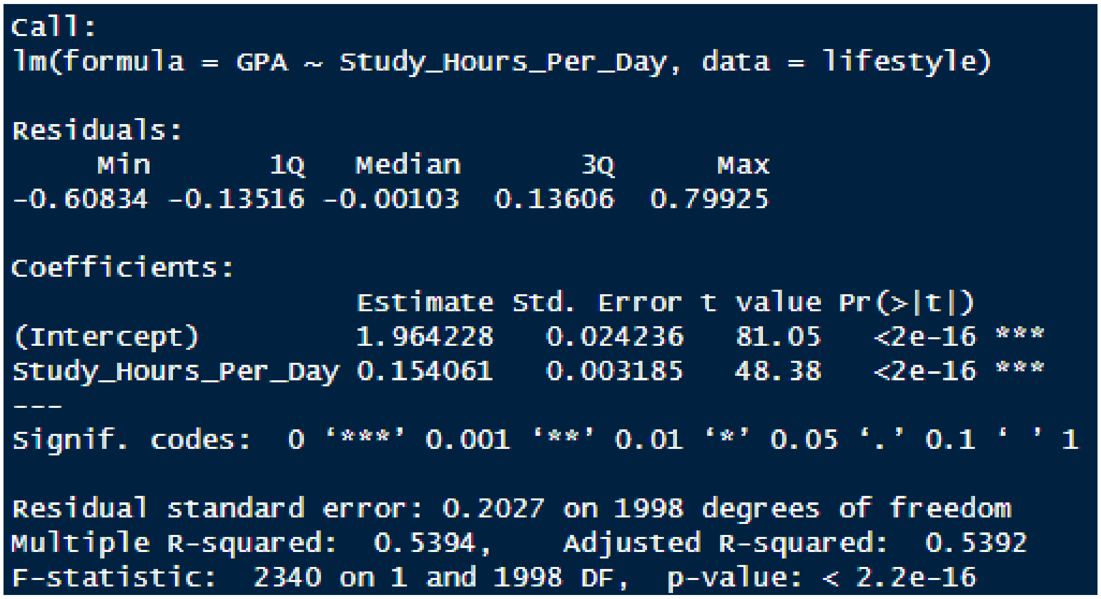
Model	adjusted R^2 value	AIC
1 : Linear	0.5392	-703.5010
2 : Multi-Linear	0.5394	-703.2180

Here, we can see that both models have nearly the same R^2 values so it would be very hard to compare them based on just that. Looking at AIC, however, we can see that Model 1, the linear model, has the lowest in comparison to Model 2. Despite the R^2 value being 0.0002 smaller, I would say that the 0.283 difference in AIC is enough to make up for it making Model 1 come out on top.

Ultimately, Model 1, the linear model, was the best model as it achieved the best overall score in consideration to only R^2 and AIC values. Despite marginally less variability being accounted for by the linear model, the lower AIC meant that it was a better representation of the relationship between study hours per day and GPA.

5. Model Summary

For Model 1:

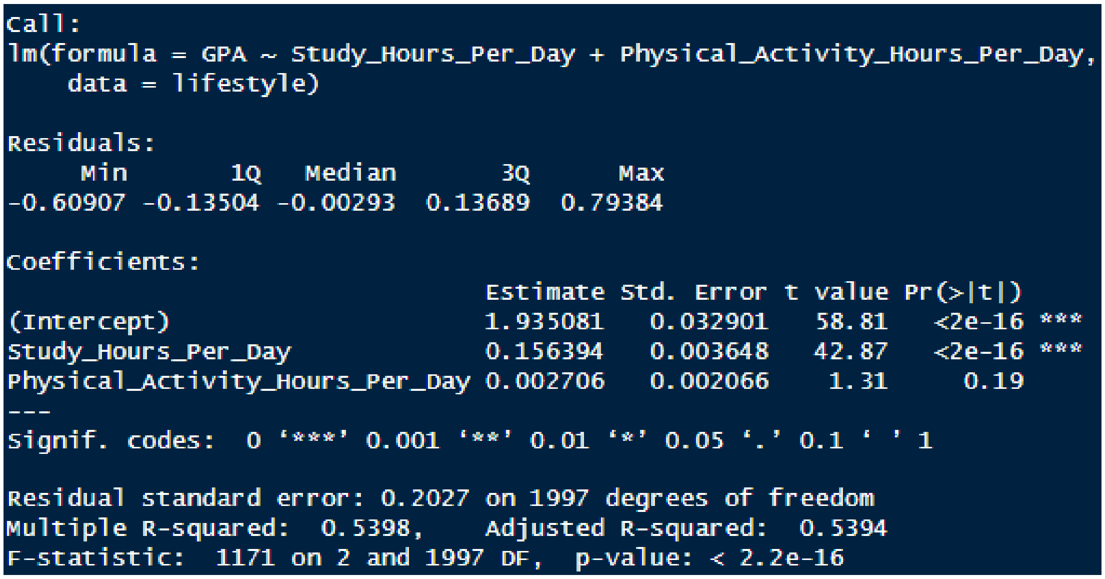


```
Call:
lm(formula = GPA ~ Study_Hours_Per_Day, data = lifestyle)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.60834 -0.13516 -0.00103  0.13606  0.79925 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         1.964228   0.024236   81.05   <2e-16 ***
Study_Hours_Per_Day 0.154061   0.003185   48.38   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2027 on 1998 degrees of freedom
Multiple R-squared:  0.5394,	Adjusted R-squared:  0.5392 
F-statistic:  2340 on 1 and 1998 DF,  p-value: < 2.2e-16
```

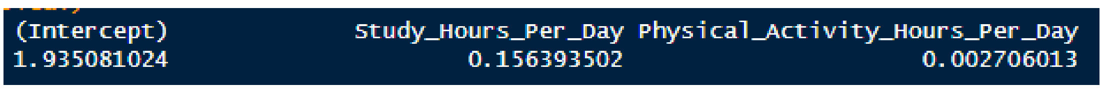
Equation for the model:

$$y = 0.1540613x + 1.9642282$$

For Model 2.2:



```
Call:
lm(formula = GPA ~ Study_Hours_Per_Day + Physical_Activity_Hours_Per_Day, 
    data = lifestyle)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.60907 -0.13504 -0.00293  0.13689  0.79384 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     1.935081   0.032901   58.81   <2e-16 ***
Study_Hours_Per_Day             0.156394   0.003648   42.87   <2e-16 ***
Physical_Activity_Hours_Per_Day 0.002706   0.002066    1.31     0.19    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2027 on 1997 degrees of freedom
Multiple R-squared:  0.5398,	Adjusted R-squared:  0.5394 
F-statistic:  1171 on 2 and 1997 DF,  p-value: < 2.2e-16
```



```
 (Intercept)             Study_Hours_Per_Day Physical_Activity_Hours_Per_Day 
1.935081024                     0.156393502                     0.002706013 
```