# Sean McLean

# ALY 6010

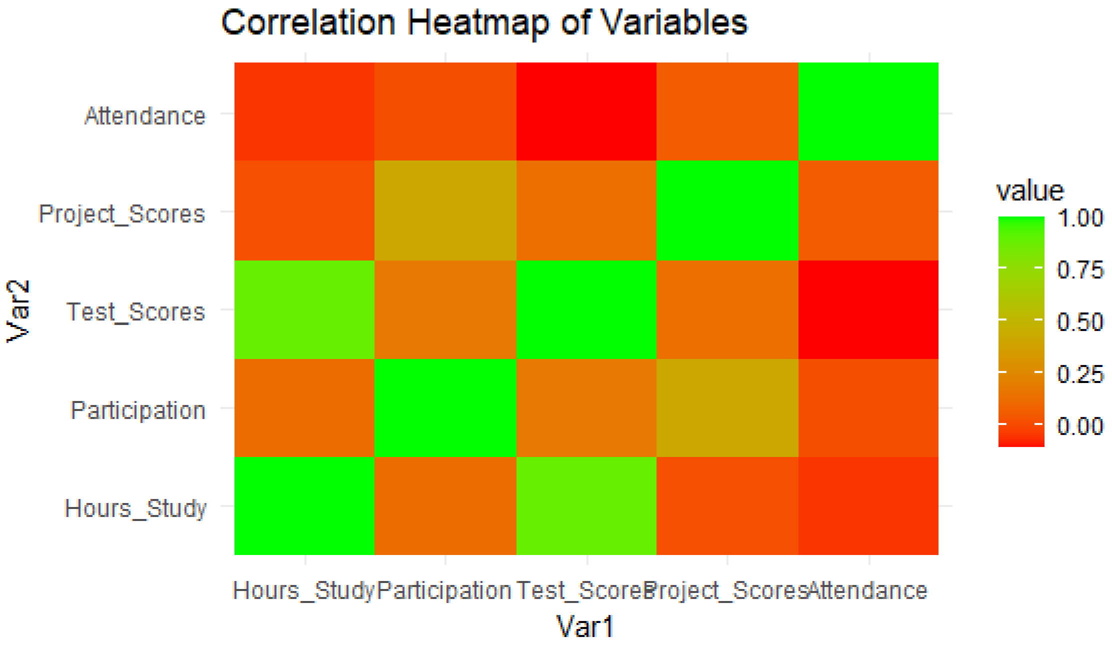
# Module 5 R Practice

## Introduction- Question 1

The correlation table used to show the correlation coefficients between the five variables in the dataset is a heatmap. The five variables in the dataset that are being analyzed are attendance, project scores, test scores, participation, and hours of study per student. The dataset contains 100 students and the focus is to look at which variables in the dataset have positive or negative correlations between each other. Using only five variables makes the correlation table more self-explanatory and prevents any issues with finding patterns and trends if more variables are incorporated into the correlation table.

## Findings/Conclusion

Overall, there are 10 possible variable correlations and every one of them has either a positive correlation or close to no correlation. The strongest correlations in the heatmap are between the test scores and hours studied and also the correlation between project scores and participation. The variable comparisons that have almost no correlation are between attendance and test scores and also attendance and hours studied. When analyzing each variable individually, attendance had the least impact as there was less correlation when compared with the other variables. The participation variable and hours studied variable fared better with a more positive correlation when compared with the other variables like test scores and project scores. It appears from the heat map that attendance does not affect academic performance as much as hours studied and participation. While I assumed that attendance and participation would have a high correlation since they seem to be related to each other in some aspects, the heatmap indicates that there is not much of a correlation between the variables.



## Introduction – Question 2

The two regression tables used for the dataset use test scores as the outcome variable for one of the tables and project scores as the outcome variable in the other table. The predictor variables used for both regression table are hours studied, attendance, and participation. Using regression tables for analysis differs from correlation analysis in that instead of looking at the relationships between the correlation variables, regression analysis seeks to predict the value of the outcome variable based off the independent variables in the table.

## Findings/Conclusion

Looking at how the predictor variables interact with the dependent variable test scores, the one variable that is by far the most impactful is hours studied. It has a much higher estimate value and very small p-value which would indicate a strong positive relationship between the variables. Overall, the r-squared value is high, the residual standard error value is small, and the p-value is very small, concluding that the predictor variables overall have a strong relationship with the outcome variable. The combination of hours studied, attendance, and participation does have a positive effect on student’s test scores.

```
Call:
lm(formula = Test_Scores ~ Hours_Study + Attendance + Participation, 
    data = reg)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.3840  -8.4884   0.7619   6.4620  21.7605 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   68.98828   49.49472   1.394    0.167    
Hours_Study    9.37846    0.50508  18.568   <2e-16 ***
Attendance    -0.75095    0.51901  -1.447    0.151    
Participation  0.07489    0.04533   1.652    0.102    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.08 on 96 degrees of freedom
Multiple R-squared:  0.7923,	Adjusted R-squared:  0.7858 
F-statistic: 122.1 on 3 and 96 DF,  p-value: < 2.2e-16
```

The second regression table looked at how hours studied, attendance, and participation can impact project scores. What sticks out is the participation variable which has a very small p-value where the other two predictor variables have very high p-values. This strong relationship makes sense in that doing projects requires participating to get the job done. The r-squared value and f-statistic values are pretty low which means it lacks significance in the regression model. Despite this, the residual standard error value is very low, indicating a better fit of the data in the model and that the observed values are close to the predicted values. Collectively, the combination of hours studied, attendance, and participation does have a somewhat positive effect on student’s project scores.

```
Call:
lm(formula = Project_Scores ~ Hours_Study + Attendance + Participation,
    data = reg)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5416 -2.9216 -0.4885  2.8174 14.7849

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -10.93951   20.08108  -0.545    0.587
Hours_Study    -0.07454    0.20492  -0.364    0.717
Attendance      0.12871    0.21057   0.611    0.542
Participation   0.08493    0.01839   4.618  1.2e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.088 on 96 degrees of freedom
Multiple R-squared:  0.1855,	Adjusted R-squared:   0.16
F-statistic: 7.287 on 3 and 96 DF,  p-value: 0.0001872
```

## References:

Bluman, A. G. (2017). Elementary Statistics: A Step-by-Step Approach. 10th edition. McGraw-Hill Education.

Chat GPT. (2023, December 10th). Default (GPT 3.5)