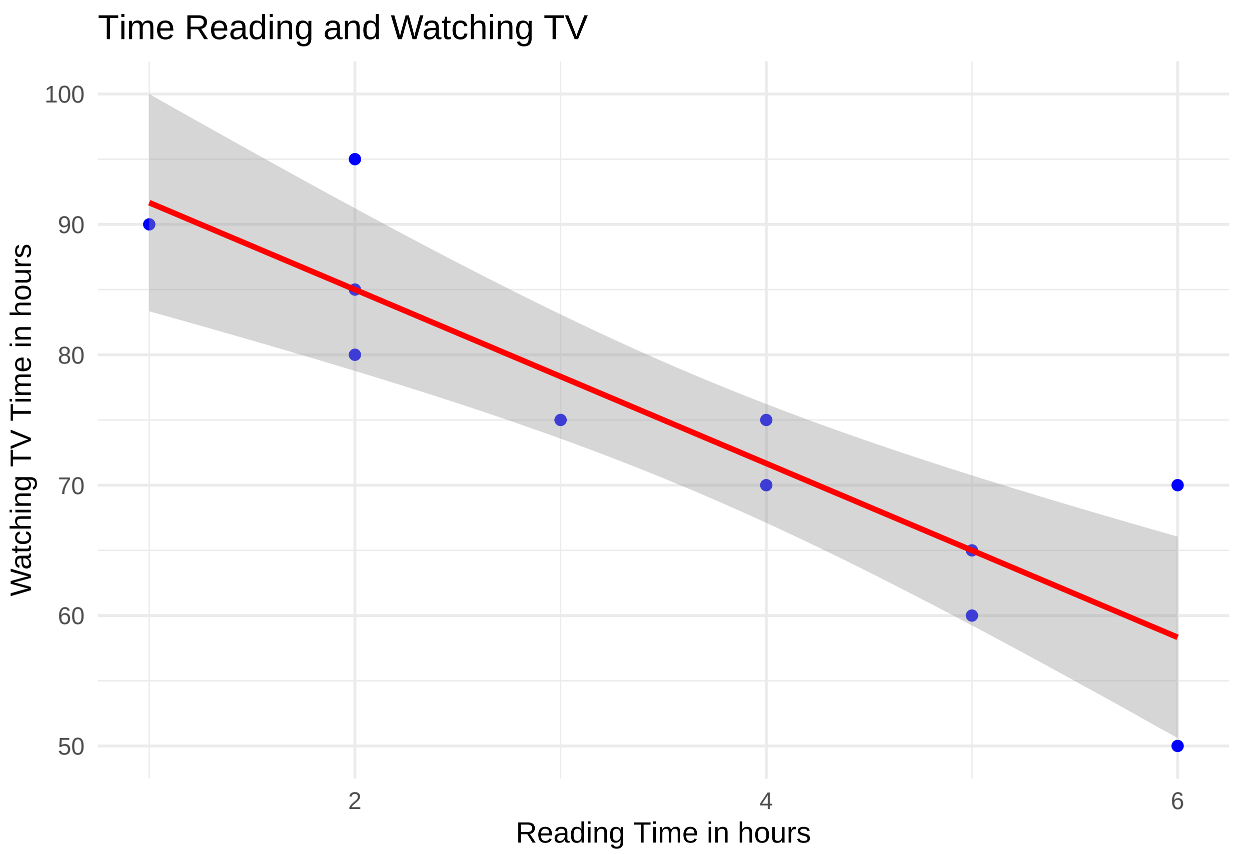
Week7_7.2_Exercise

Maxim Bilenkin

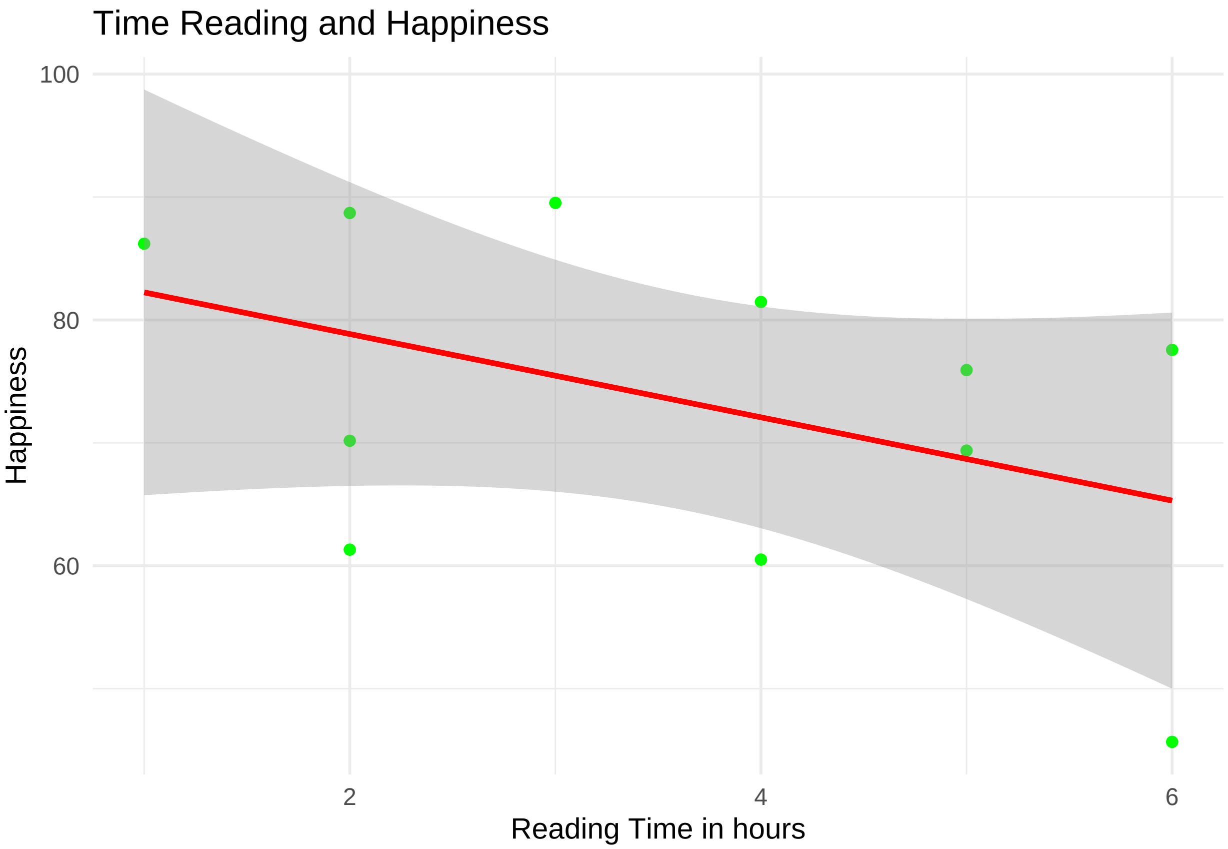
2025-01-22

1. Using the following file: student-survey complete the below steps.
2. Create the following plots of survey variables (first variable is x-axis, second is y-axis):
   a. TimeReading, TimeTV
   b. TimeReading, Happiness
   c. TimeTV, Happiness

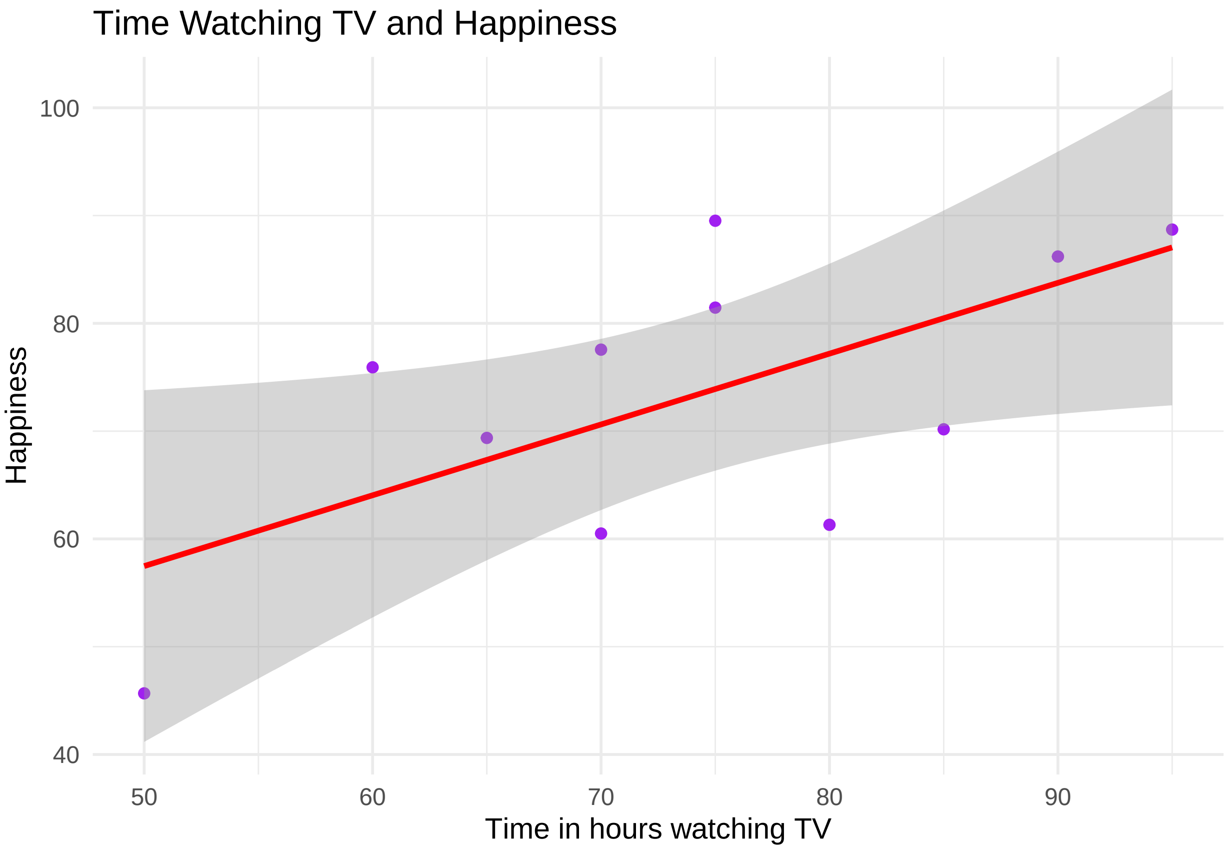
```
# Loading data file "student-survey and saving to a variable
student_survey_data <-
read.csv("C:/Users/maxim/OneDrive/Desktop/Bellevue University/DSC 520/student-survey.csv")

# Loading library
library(ggplot2)

# Creating First plot and saving to a variable
first_plot <- ggplot(student_survey_data, aes(x = TimeReading, y = TimeTV)) +
    geom_point(color = "blue") +
    geom_smooth(method = "lm", formula = y ~ x, color = "red") +
    labs(title = "Time Reading and Watching TV",
         x = "Reading Time in hours",
         y = "Watching TV Time in hours") +
    theme_minimal()

# Creating Second plot and saving to a variable
second_plot <- ggplot(student_survey_data, aes(x = TimeReading, y = Happiness)) +
    geom_point(color = "green") +
    geom_smooth(method = "lm", formula = y ~ x, color = "red") +
    labs(title = "Time Reading and Happiness",
         x = "Reading Time in hours",
         y = "Happiness") +
    theme_minimal()

# Creating Third plot and saving to a variable
third_plot <- ggplot(student_survey_data, aes(x = TimeTV, y = Happiness)) +
    geom_point(color = "purple") +
    geom_smooth(method = "lm", formula = y ~ x, color = "red") +
    labs(title = "Time Watching TV and Happiness",
         x = "Time in hours watching TV",
         y = "Happiness") +
    theme_minimal()

# Displaying all three plots in order
print(first_plot)
```



```r
print(second_plot)
```



```
print(third_plot)
```



3. Based on the plots you created, provide a rough estimate of the relationship between the variables; do the slopes indicate a positive or negative relationship?

Answer:

For the first scatter plot named "Time Reading and Watching TV" the relationship is negative. We can see that the relationship between "Reading Time" and "Watching TV Time" is downward sloping. Also, the red line clearly indicates downward sloping.

For the second scatter plot named "Time Reading and Happiness", it looks like the relationship is negative between the two variables of "Reading Time in hours" and "Happiness" because the red line is downward sloping.

For the third scatter plot named "Time Watching TV and Happiness", the relationship is positive. We can see how the red line upward sloping. So, there is a positive correlation between the two variables.

4. Create a covariance matrix with TimeReading, TimeTV, and Happiness variables. Explain the relationship between the variables using the matrix.

```
# Selecting needed column names to create covariance matrix and assigning to a variable
selected_column_names <- student_survey_data[, c("TimeReading", "TimeTV", "Happiness")]

# Calculation covariance using cov() method
covariance_matrix <- cov(selected_column_names)

# Rounding the covariance matrix output to two decimals for better readability
formatted_to_two_dec <- round(covariance_matrix, 2)

# Displaying result of calculated covariance matrix
print(formatted_to_two_dec)
```

```
##             TimeReading TimeTV Happiness
## TimeReading        3.05 -20.36    -10.35
## TimeTV           -20.36 174.09    114.38
## Happiness        -10.35 114.38    185.45
```

Explanation for relationship between the variables. As we can see the some output is negative and some positive. The correlation of -20.36 between TimeTV and TimeReading is negative which means the two variable have inverse relationship. As TimeReading increases, the TimeTV decreases. On the other hand, the covariance of 114.38 between TimeTV and Happiness has a strong positive relationship relative to other variable. As TimeTV increases, Happiness also increases. Lastly, the covariance of -10.35 indicating negative weak relationship between the two variable. As the TimeReading increases the Happiness tends to decrease but not a lot relative to other variables that have negative relationships.

5. Now, create a correlation matrix with the same variables. Again, explain the relationship between the variables using the matrix. Additionally, in your personal opinion, are the relationships between variables easier interpret using the covariance or correlation matrix? Explain your answer.

```r
# Selecting relevant columns for correlation analysis
selected_data <- student_survey_data[, c("TimeReading", "TimeTV", "Happiness")]

# Calculating the correlation matrix using cor()
correlation_matrix <- cor(selected_data)

# Rounding the correlation  n matrix to 2 decimal places for better readability
formatted_correlation_matrix <- round(correlation_matrix, 2)

# Displaying the formatted correlation matrix
print(formatted_correlation_matrix)
```

```
##             TimeReading TimeTV Happiness
## TimeReading        1.00  -0.88     -0.43
## TimeTV            -0.88   1.00      0.64
## Happiness         -0.43   0.64      1.00
```

Explanation for relationship between the variables using correlation matrix. The correlation of -0.88 between TimeReading and TimeTV variables indicate a strong negative correlation. So, as the time spent reading increases, the time spent watching television decreases substantially.

The correlation of -0.43 for TimeReading and Happiness variables has moderate negative correlation. We can think of it as more reading time has less association with lower happiness.

The correlation of 0.64 between TimeTV and Happiness has moderate positive correlation. In other words, people that are spend more time watching more television having higher happiness.

For me personally, correlation matrix is easier to interpret because the number bound by -1 and 1 on negative and positive side respectively. Also, correlation presents the highest available number either -1 or 1 in a single number. On the other hand, covariance present numbers in multiple digits. Also, when I look at correlation, I think of it in percentage terms by multiplying by 100. For example, 1 means 100% or 0.64 means 64% relationship. So, if one variable moves then the other variable will be impacted by 64% of a time. It shows the strength of two variables. Obviously, 1 or 100% has a very strong relationship strength of two variables.

6. Perform a correlation test on TimeReading and TimeTV. What does the correlation value tell you about the relationship between them? Can you say that TimeReading has an effect on TimeTV?

We can see that the correlation of -0.88 has a strong or perhaps very strong negative relationship between TimeReading and TimeTV variables. We can think of it as people who are spending more time of reading tend to spend less time watching television. So, yes we can clearly say that TimeReading has an effect on TimeTV if we look at -.88 value. However, it is very important to mention that because two variables have strong relationship according to the calculated value doesn't imply causation. So, because two variables have strong or negative correlation we cannot for sure to say that one variable impacts the other.