Student Survey Dataset Assignment

David Berberena

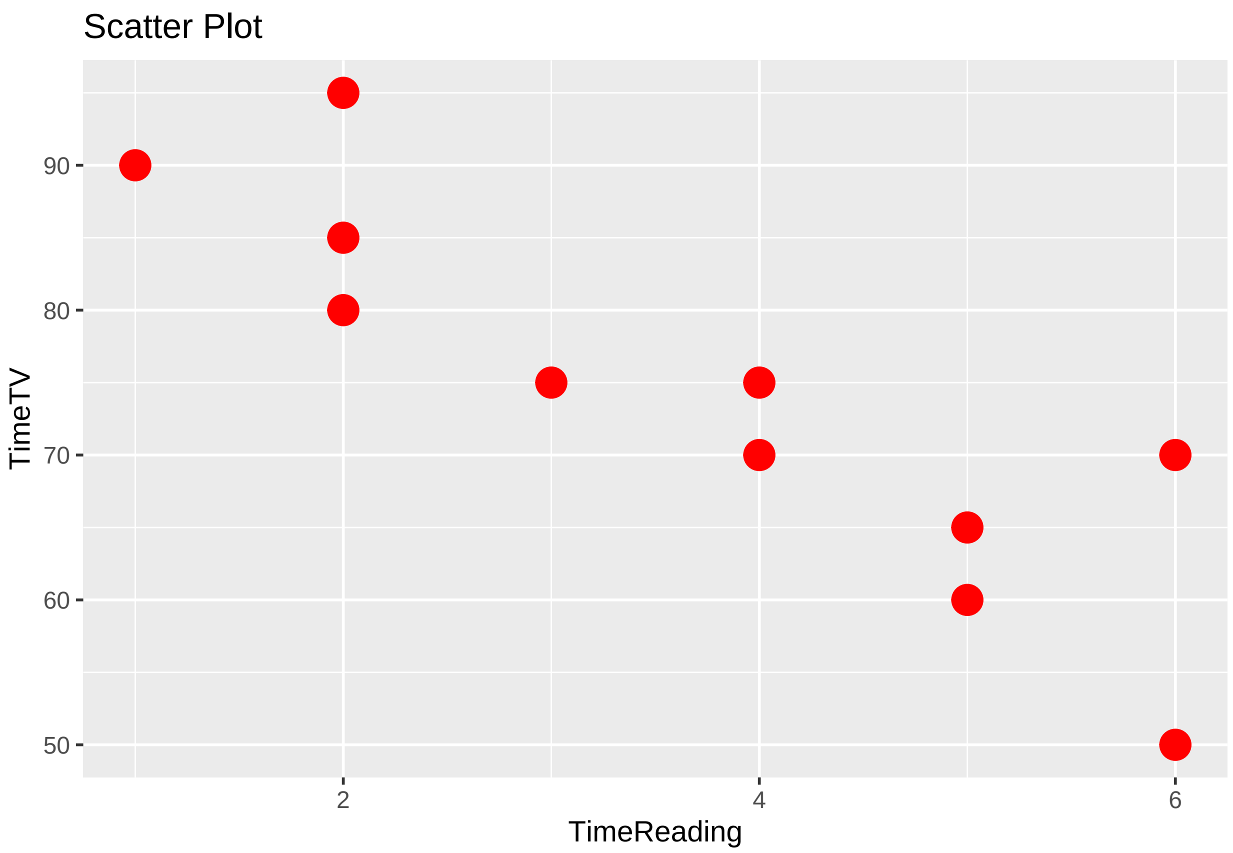
01-28-2024

Student Survey Dataset Assignment

```r
# Assignment Start

# Upload readr library for file importing

library(readr)

# set working directory for smooth file importing

setwd("C:/Users/dbzda/Documents/School/DSC 520 Statistics for Data Science")

# Import the converted Student Survey CSV file to view its properties

student <- read_csv("Student.csv", show_col_types = FALSE)
```

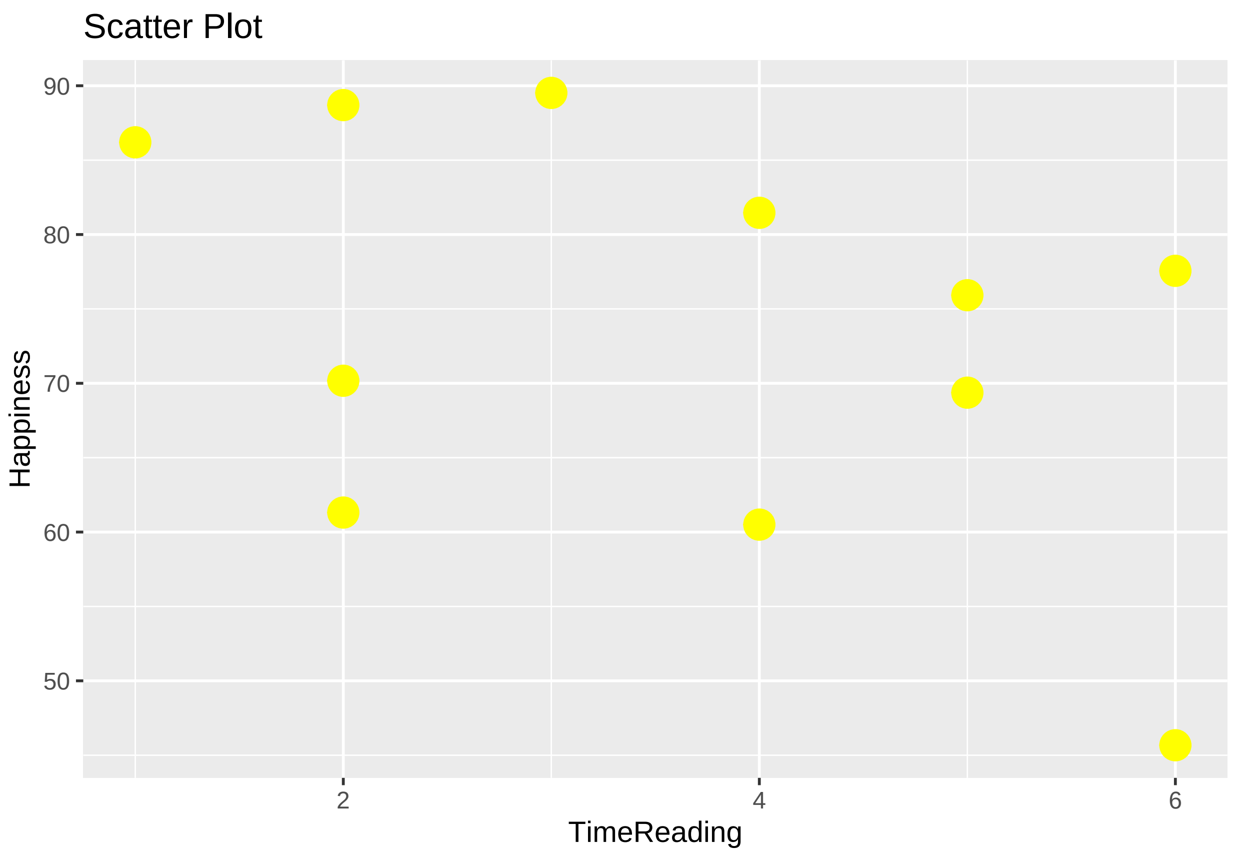
Survey Variable Plots

```r
# ggplot2 package is loaded to create various plots for the dataset

library(ggplot2)

# First plot for TimeReading and TimeTV

ggplot(student, aes(x = TimeReading, y = TimeTV)) +
  geom_point(color = "red", size = 5) +
  labs(title = "Scatter Plot", x = "TimeReading", y = "TimeTV")
```



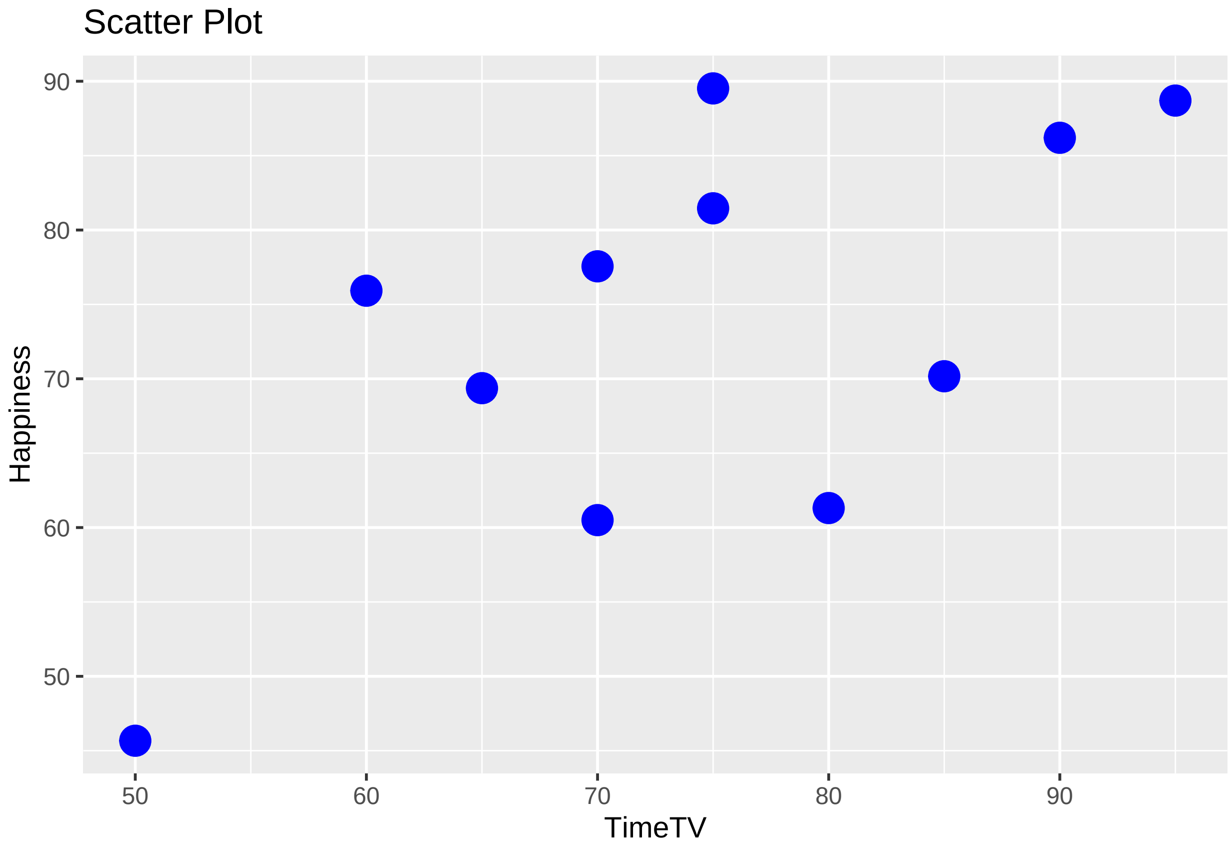
Based on the scatter plot generated for the relationship between TimeReading and TimeTV, the downward slope of the plot from left to right indicates a negative relationship between the two variables. This negative relationship is strong as the data points are clustered tightly around each other yet still maintain the telltale decreasing slope.

```r
# Second plot for TimeReading and Happiness

ggplot(student, aes(x = TimeReading, y = Happiness)) +
  geom_point(color = "yellow", size = 5) +
  labs(title = "Scatter Plot", x = "TimeReading", y = "Happiness")
```



Based on this scatter plot of TimeReading and Happiness, it is hard to discern a positive or negative relationship between the two. The data points are very spread out and do not create a definitive slope to analyze, so I would say the variables are not linearly related, and the relationship is weak due to the spread of the data points.

```
# Third plot for TimeTV and Happiness

ggplot(student, aes(x = TimeTV, y = Happiness)) +
  geom_point(color = "blue", size = 5) +
  labs(title = "Scatter Plot", x = "TimeTV", y = "Happiness")
```



This scatter plot shows a positive relationship between TimeTV and Happiness. The plot shows an increase in slope from left to right, which is the key indicator of a positive relationship. I would say that the strength of this positive relationship is rather weak though since the data points are a little scattered and not closely packed together to create a strong positively linear line.

Covariance and Correlation

```r
# Covariance matrix of all three variables

# Each variable vector needs to be extracted from the dataset using the
# $ operator and bound together into a matrix which will be made into a
# covariance matrix using the built-in cov() function

TimeReading <- student$TimeReading

TimeTV <- student$TimeTV

Happiness <- student$Happiness

# creation of matrix using cbind()

matrix <- cbind(TimeReading, TimeTV, Happiness)
covar_matrix <- cov(matrix)
covar_matrix
```

```
##             TimeReading    TimeTV Happiness
## TimeReading    3.054545 -20.36364 -10.35009
## TimeTV       -20.363636 174.09091 114.37727
## Happiness    -10.350091 114.37727 185.45142
```

Looking at the covariance values for each variable pairing, I can see the following:

- TimeReading and TimeTV have the lowest negative covariance value, meaning that as TimeReading increases, TimeTV will decrease and vice-versa.
- TimeReading and Happiness also has a negative covariance coefficient, meaning it will exhibit the same behavior as TimeReading and TimeTV. However, as the coefficient here is not as low as the previous comparison, when the first variable increases, the decrease of the second variable will not happen as drastically.
- Happiness and TimeTV have a positive covariance coefficient, meaning that the variables will change in the same direction (when one increases, so will the other, and the same holds true for a decrease).

The other numbers within the matrix are the levels of variance for each variable. These numbers show that TimeReading has the lowest level of variance while TimeTV and Happiness have high levels of variance, with Happiness having the highest variance levels among the variables being compared.

```
# Correlation matrix of all three variables

corel_matrix <- cor(matrix)
corel_matrix
```

```
##             TimeReading     TimeTV  Happiness
## TimeReading   1.0000000 -0.8830677 -0.4348663
## TimeTV       -0.8830677  1.0000000  0.6365560
## Happiness    -0.4348663  0.6365560  1.0000000
```

Looking at the correlation coefficients for each variable pairing, I can see the following:

- TimeReading and TimeTV have a strong negative correlation. This means that their linear relationship has a steep negative slope, so as one variable increases, the other sharply decreases. Its negative strength is seen in its value of -0.88.
- TimeReading and Happiness have a medium negative correlation, as outlined by its value of -0.43. These two variables also have a negative linear relationship, yet the steepness of the negative slope is not as severe as the previous comparison.
- TimeTV and Happiness have a medium positive correlation, as the correlation coefficient is a positive number. Its strength of medium is defined by its value of 0.63. These variables will have a positive linear relationship, which means that as one variable increases or decreases, the second variable will perform in the same way.
- The correlation coefficients of 1 for each of the variables against themselves shows that each variable is perfectly linearly related to itself, as it is just being compared to itself.

Regarding the ease of which I can interpret the relationship between each variable pair using either covariance or correlation, I believe that correlation is much easier to deal with and decipher. Correlation coefficients are designed to fit within the scale of -1 to 1. This guideline easily allows someone to see to what degree a variable pair is linearly related and in what direction. While covariance allows the viewer to discern almost the same thing, the scale of which covariance values appear feels arbitrary in its measurement of linear strength. I can't really tell how much stronger a variable pair with 1000 as a covariance coefficient is linearly than a variable whose covariance value is 400 when there is no scale to put those numbers up against.

```
# Correlation test

# Correlation tests can be conducted using the cor.test() function

pearson_test <- cor.test(TimeReading, TimeTV, method = 'pearson')
pearson_test
```

```
##
##  Pearson's product-moment correlation
##
## data:  TimeReading and TimeTV
## t = -5.6457, df = 9, p-value = 0.0003153
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9694145 -0.6021920
## sample estimates:
##        cor
## -0.8830677
```

While the Pearson correlation test tells the viewer much information, such as the t-test value (measure significance of difference in variation of the data), degrees of freedom (number of values that are allowed to vary), the p-value (the likeliness of a group difference being due to chance), the 95% confidence interval range, and an alternative hypothesis (in relation to the null hypothesis), the important value here is the correlation coefficient generated by the Pearson test, which is -0.8830677. This value means that the two variables are strongly negatively correlated, meaning that when one variable increases linearly, the other drastically decreases. As much as the data shows that there is a strong negative correlation between TimeReading and TimeTV, I cannot say for certain that TimeReading has an effect on TimeTV. This is because of one of the main defining aspects of correlation: it does not imply causation. More tests would have to be run to say for sure that TimeReading has a direct effect on TimeTV.