# Week7Assignment_RiazAhmed_TamimAnsari.r

Riaz

2023-10-15

```
# title: "Week_7_Excercise_Riaz"
# author: "Riaz Ahmed Tamim Ansari"
# date: "2023-10-15"


#Load the data and ggplot2
df <- read.delim("C:\\Users\\Riaz\\Desktop\\MSDS\\Introduction to Statistics\\Week7\\student-survey.csv
df
```

```
##    TimeReading TimeTV Happiness Gender
## 1            1     90     86.20      1
## 2            2     95     88.70      0
## 3            2     85     70.17      0
## 4            2     80     61.31      1
## 5            3     75     89.52      1
## 6            4     70     60.50      1
## 7            4     75     81.46      0
## 8            5     60     75.92      1
## 9            5     65     69.37      0
## 10           6     50     45.67      0
## 11           6     70     77.56      1
```

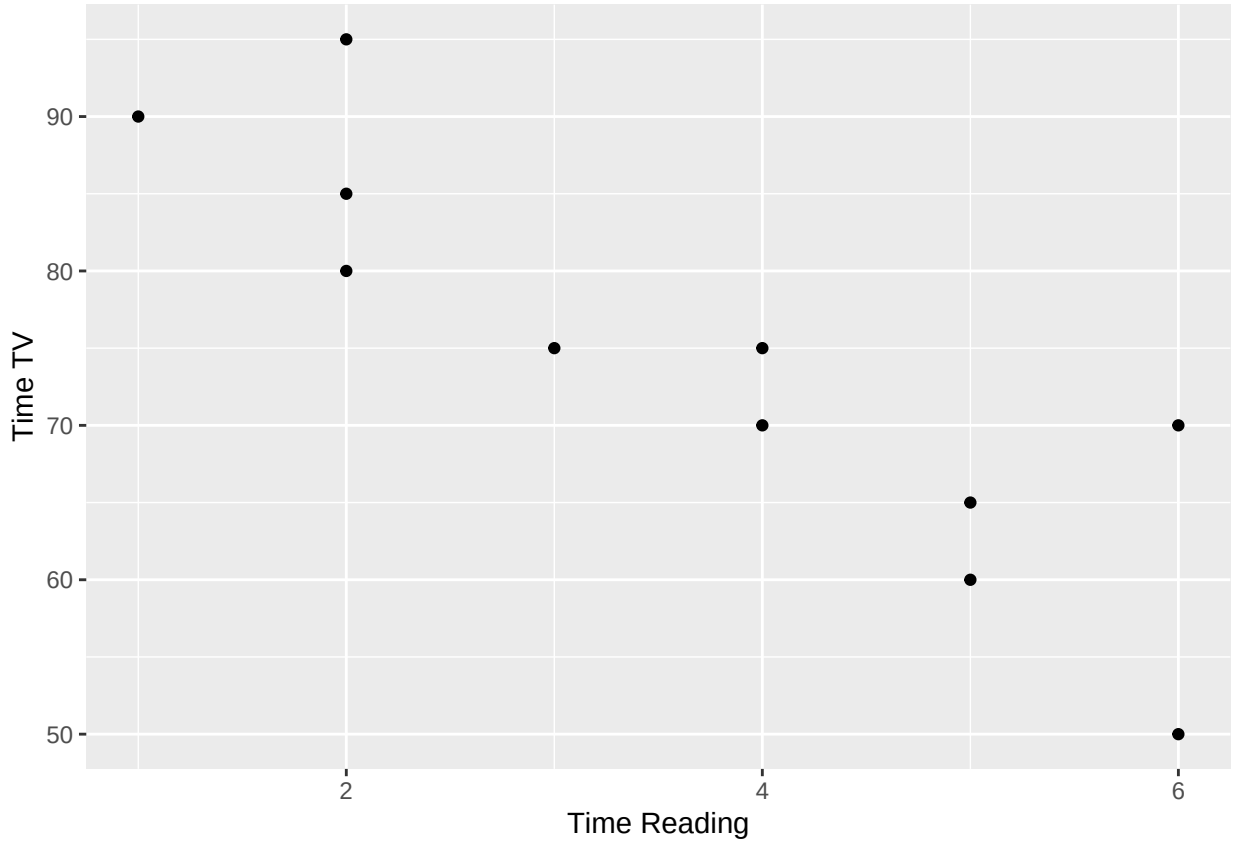
```
library(ggplot2)

#1.  Create the following plots of survey variables (first variable is x-axis, second is y-axis):
#2.  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?

tvplot1 <- ggplot(df,aes(df$TimeReading,df$TimeTV))
tvplot1 + geom_point() + labs(x="Time Reading", y="Time TV")
```

```
## Warning: Use of `df$TimeReading` is discouraged.
## i Use `TimeReading` instead.
```

```
## Warning: Use of `df$TimeTV` is discouraged.
## i Use `TimeTV` instead.
```

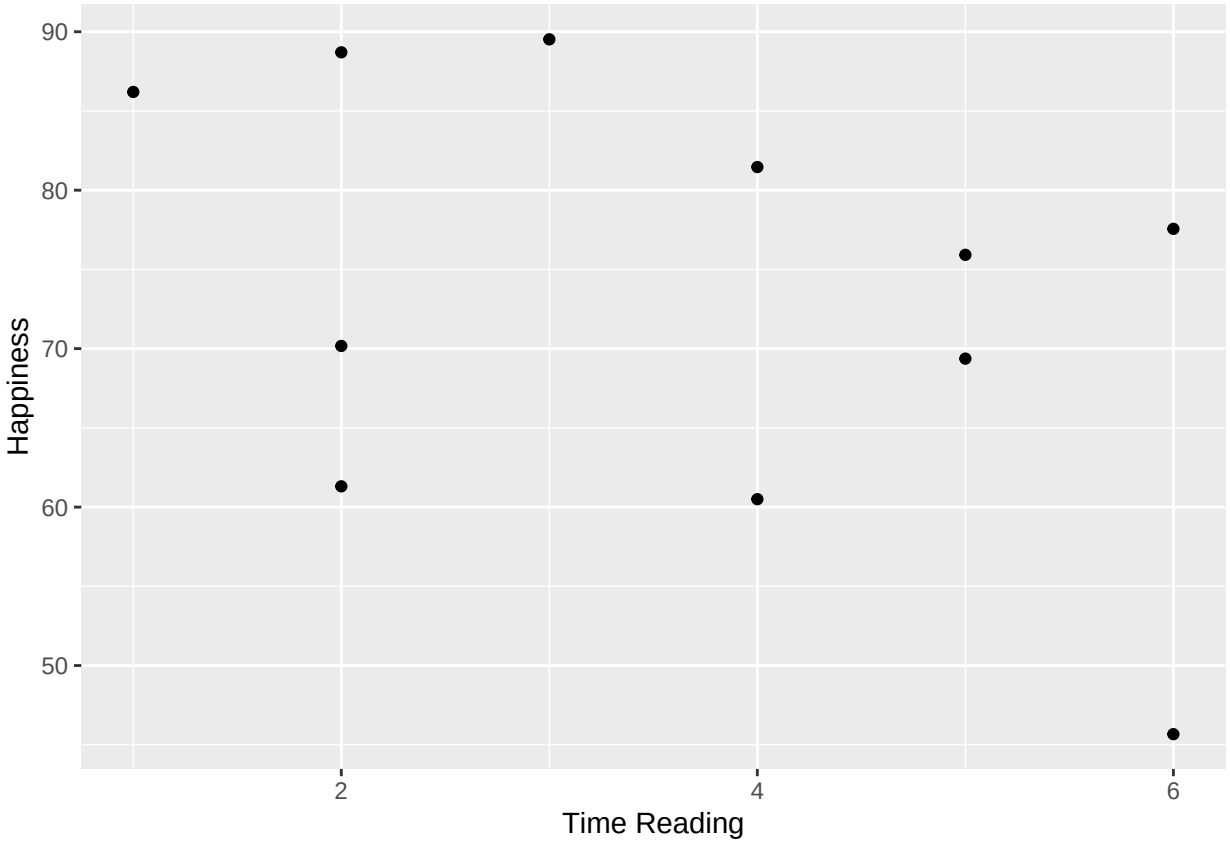



```
#Based on the above plot there is a negative relationship between TimeReading and TimeTV

tvplot2 <- ggplot(df,aes(df$TimeReading,df$Happiness))
tvplot2 + geom_point() + labs(x="Time Reading", y="Happiness")
```

```
## Warning: Use of ‘df$TimeReading‘ is discouraged.
## i Use ‘TimeReading‘ instead.
```

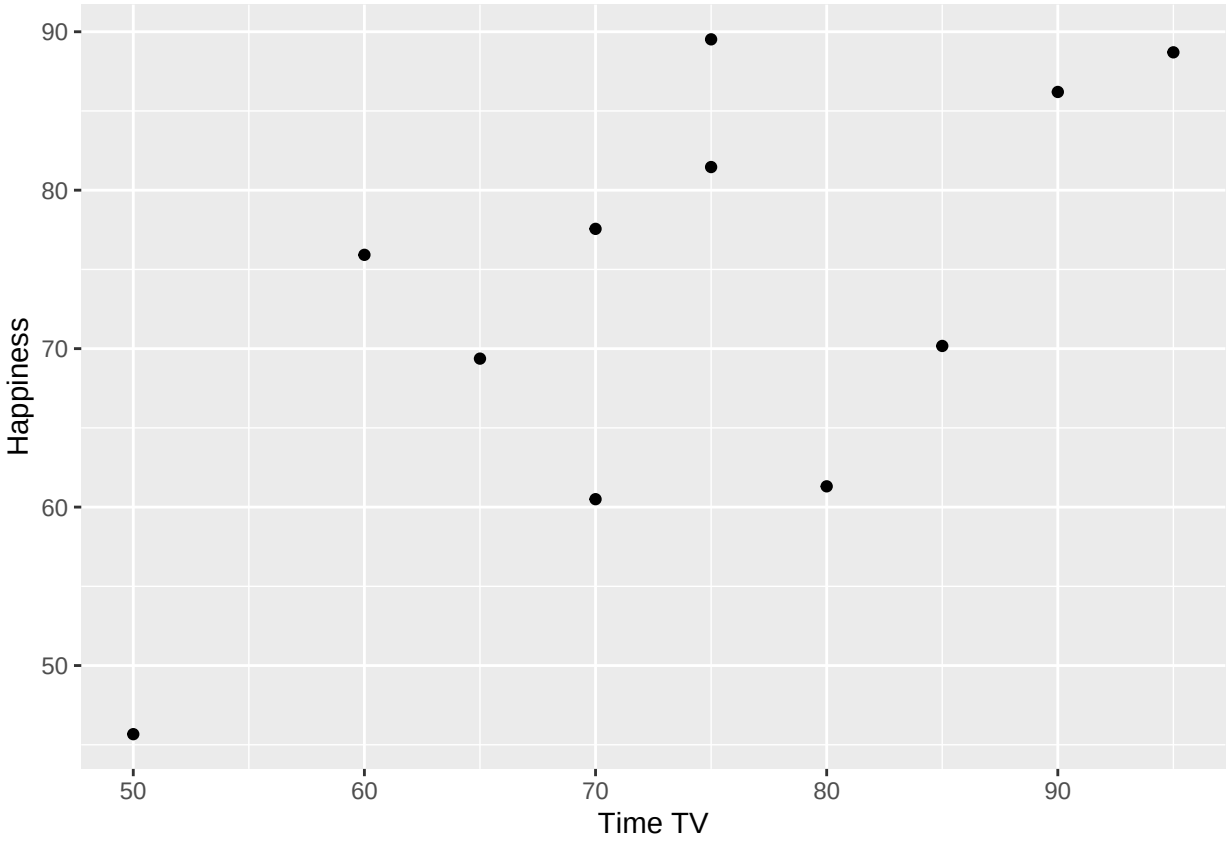
```
## Warning: Use of ‘df$Happiness‘ is discouraged.
## i Use ‘Happiness‘ instead.
```




```
#Based on the above plot, I am not able to find any relationship between TimeReading and Happiness
#The plot just looks scattered and cannot infer much.

tvplot3 <- ggplot(df,aes(df$TimeTV,df$Happiness))
tvplot3 + geom_point() + labs(x="Time TV", y="Happiness")
```

```
## Warning: Use of ‘df$TimeTV‘ is discouraged.
## i Use ‘TimeTV‘ instead.
## Use of ‘df$Happiness‘ is discouraged.
## i Use ‘Happiness‘ instead.
```

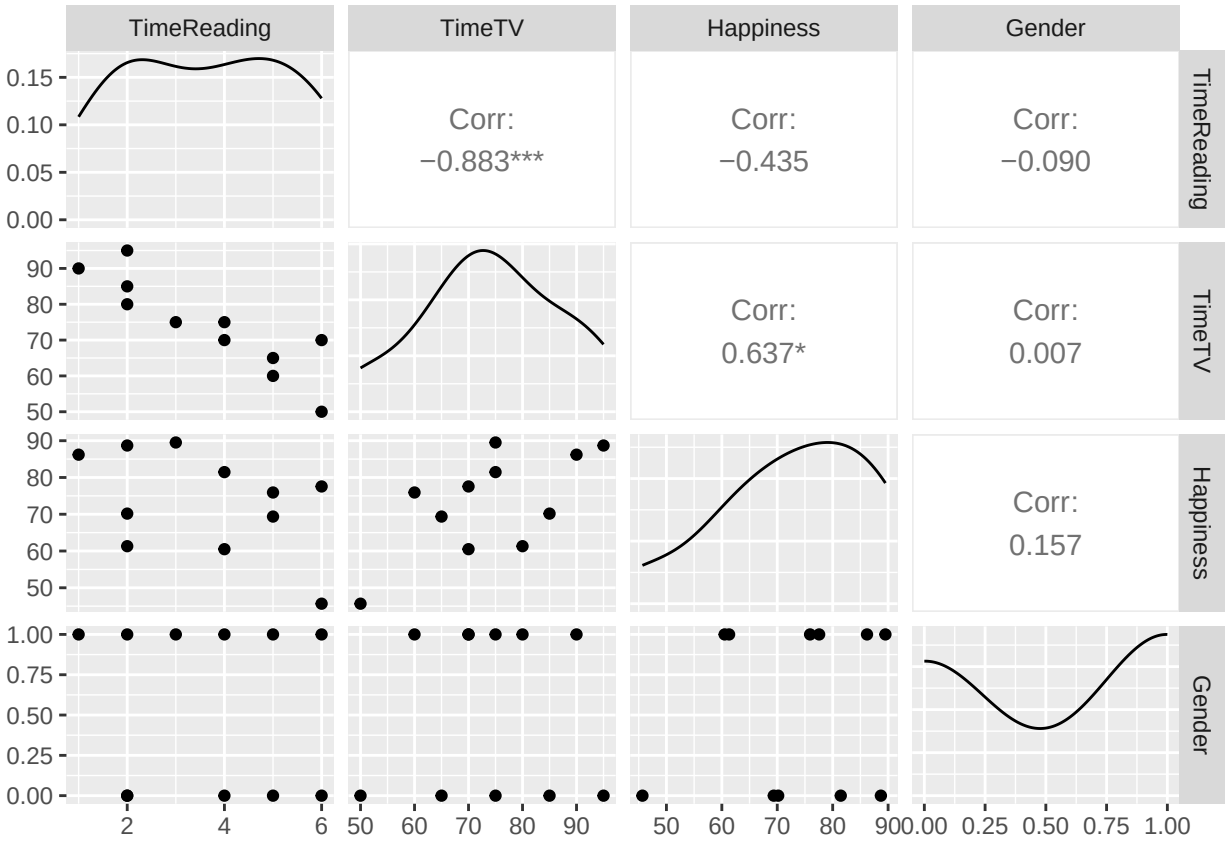



```
#Based on the above plot, I am not able to find any relationship between TimeTV and Happiness
#The plot just looks scattered and cannot infer much.

# We can also visualize using ggpairs function from GGally
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
GGally::ggpairs(df)
```




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

cov(df)
```

```
##              TimeReading       TimeTV  Happiness      Gender
## TimeReading   3.05454545 -20.36363636 -10.350091 -0.08181818
## TimeTV      -20.36363636 174.09090909 114.377273  0.04545455
## Happiness   -10.35009091 114.37727273 185.451422  1.11663636
## Gender       -0.08181818   0.04545455   1.116636  0.27272727
```

```
#Covariance can vary from -infinity to +infinity.
#The diagonal represents variance, of the respective variables.
#There is a negative covariance between TimeReading to TimeTV/Happiness/Gender
#There is a positive covariance between TimeTV to Happiness
#There is a positive covariance between  Happiness to TimeTV
#However covariance is not a standardized unit of measure.  For that we need to find correlation coeffi

#4. 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.

cor(df,method = "pearson")
```

```
##              TimeReading       TimeTV  Happiness       Gender
```

```
## TimeReading  1.00000000 -0.883067681 -0.4348663 -0.089642146
## TimeTV      -0.88306768  1.000000000  0.6365560  0.006596673
## Happiness   -0.43486633  0.636555986  1.0000000  0.157011838
## Gender      -0.08964215  0.006596673  0.1570118  1.000000000
```

```
#The diagonal would be always 1 as it is the correlation for the same variable.
#Correlation can vary from -1 to 1
#There is a negative correlation between TimeReading to TimeTV/Happiness/Gender
#There is a positive correlation between TimeTV to Happiness
#There is a positive correlation between  Happiness to TimeTV
#I find it is easier to interpret using correlation matrix as this is a standardized unit of measure.
#Here we define the standardized measure by dividing by the Std deviation of both the variables.
#Also looking at the TimeReading and TimeTV there is a strong negative correlation
#Also looking at Happiness and TimeTV there is a strong postive correlation.


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

cor.test(df$TimeReading,df$TimeTV,alternative = "less")
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$TimeReading and df$TimeTV
## t = -5.6457, df = 9, p-value = 0.0001577
## alternative hypothesis: true correlation is less than 0
## 95 percent confidence interval:
##  -1.0000000 -0.6684786
## sample estimates:
##        cor
## -0.8830677
```

```
#Doing a correlational test between these two variables, I infer that the 95% CI is inbetween -1.0 and
#And this does not cross 0, so we are sure that they are negatively correlated.
#AS the p value is less than .05, we can reject the null hypothesis. Alternative hypothesis would be
#TimeReading and TimeTV correlation is less than 0, and they are strongly negatively correlated
#and it has the effect.
```