ASSIGNMENT 7.2 Student Survey

Kausik Chattapadhyay

October 12 2022

2. Student Survey

As a data science intern with newly learned knowledge in skills in statistical correlation and R programming, you will analyze the results of a survey recently given to college students. You learn that the research question being investigated is: "Is there a significant relationship between the amount of time spent reading and the time spent watching television?" You are also interested if there are other significant relationships that can be discovered? The survey data is located in
this StudentSurvey.csv file.

Load the Student Survey dataset

```
##   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
```

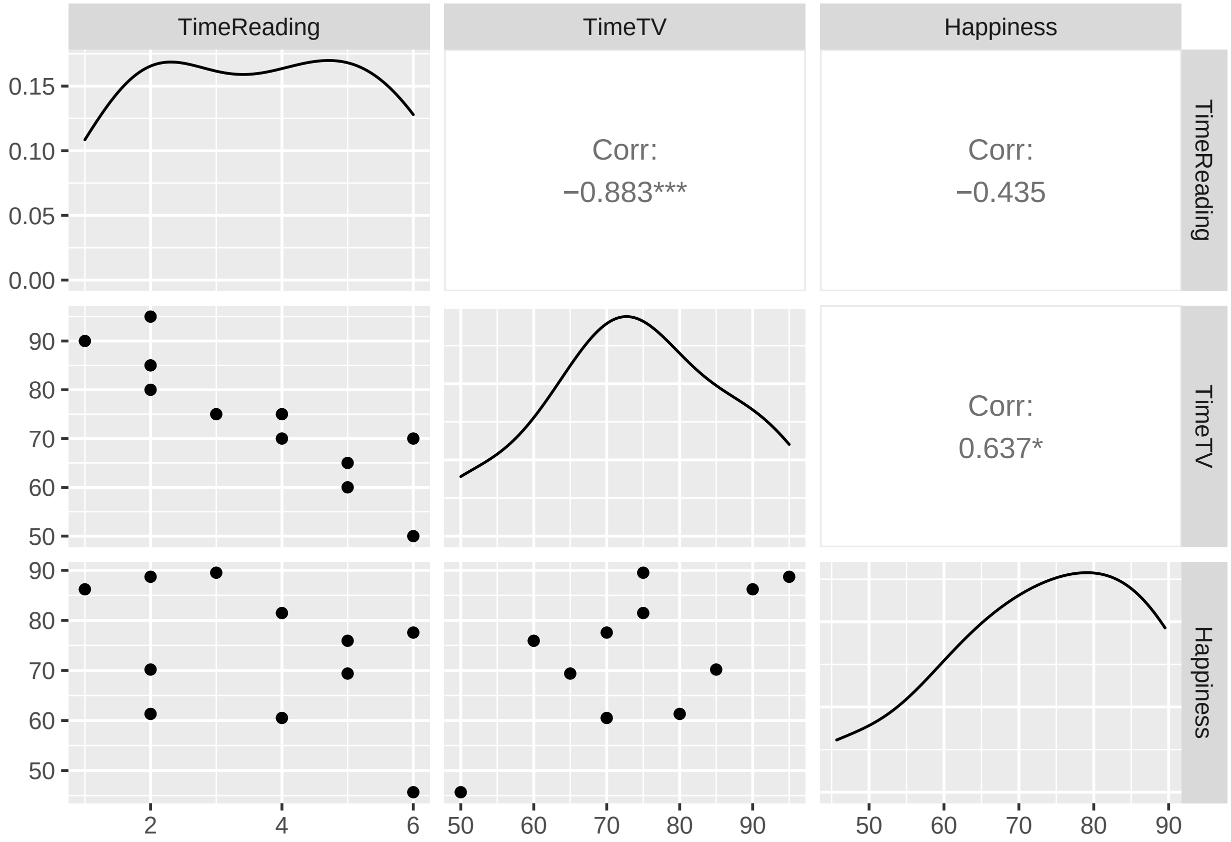
i.

Use R to calculate the covariance of the Survey variables and provide an explanation of why you would use this calculation and what the results indicate.

```
#create the matrix of the student data for variables TimeReading, TimeTV and Happiness

cor(students_df[, c("TimeReading", "TimeTV", "Happiness")])
```

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

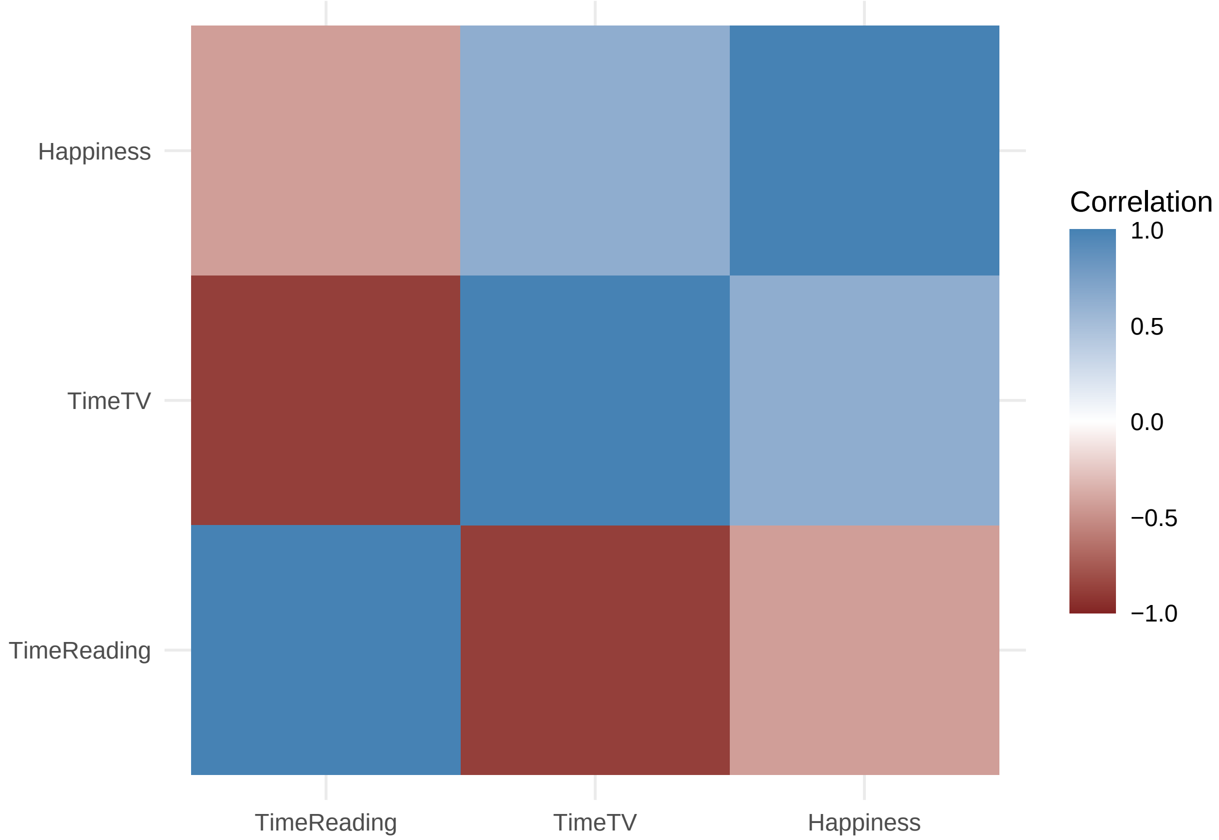
```
#lets draw a graph for the correlation
GGally::ggpairs(students_df[, c("TimeReading", "TimeTV", "Happiness")])
```



```
#Heatmap
# load the reshape package for melting the data
library(reshape2)

# load the scales package for some extra plotting features
library(scales)

# build the correlation matrix
studentCor <- cor(students_df[, c("TimeReading", "TimeTV", "Happiness")])

# melt it into the long format
studentMelt <- melt(studentCor, varnames=c("x", "y"), value.name="Correlation")

# order it according to the correlation
studentMelt <- studentMelt[order(studentMelt$Correlation), ]

# display the melted data
studentMelt
```

```
##             x           y Correlation
## 2      TimeTV TimeReading  -0.8830677
## 4 TimeReading      TimeTV  -0.8830677
## 3   Happiness TimeReading  -0.4348663
## 7 TimeReading   Happiness  -0.4348663
## 6   Happiness      TimeTV   0.6365560
## 8      TimeTV   Happiness   0.6365560
## 1 TimeReading TimeReading   1.0000000
## 5      TimeTV      TimeTV   1.0000000
## 9   Happiness   Happiness   1.0000000
```

```
## plot it with ggplot
# initialize the plot with x and y on the x and y axes
ggplot(studentMelt, aes(x=x, y=y)) + geom_tile(aes(fill=Correlation)) +
        scale_fill_gradient2(low=muted("red"), mid="white",
                                          high="steelblue",
                                         guide=guide_colorbar(ticks=FALSE, barheight=10),
                                         limits=c(-1, 1)) +  theme_minimal() + labs(x=NULL, y=NULL)
```



Ans :

1. I am using Pearson correlation for this calculation as it ony requires that data are interval for it to be an accurate measure of the linear relationship between two variables.

2. we see that the TimeReading is negatively related to TimeTv with pearson correlation of r = -0.883, this is a reasonably big effect, so we can conclude that as Tv Time increases the Reading time decreases.

3. Also, we see that the TimeReading is negatively related to Happiness with pearson correlation of r =-0.434, again this is a big effect, so we can conclude that as Reading time increases the happiness decreases

ii.

Examine the Survey data variables. What measurement is being used for the variables? Explain what effect changing the measurement being used for the variables would have on the covariance calculation. Would this be a problem? Explain and provide a better alternative if needed.

```
# Ans: In the Survey data variables, there are following variables with the mentioned # measurement as f
# TimeReading - numeric value (hours)
# TimeTV      - numeric value (minutes)
# Happiness   - float value (int percentage)
# Gender      - numeric value ( 1 represents male and 0 female)

# Effect of changing the measurement in covariance
# lets convert the TimeReading to Minutes and get the covariance

students_df_new <- students_df
students_df_new$TimeReading <- students_df_new$TimeReading * 60
students_df_new
```

```
##    TimeReading TimeTV Happiness Gender
## 1           60     90     86.20      1
## 2          120     95     88.70      0
## 3          120     85     70.17      0
## 4          120     80     61.31      1
## 5          180     75     89.52      1
## 6          240     70     60.50      1
## 7          240     75     81.46      0
## 8          300     60     75.92      1
## 9          300     65     69.37      0
## 10         360     50     45.67      0
## 11         360     70     77.56      1
```

```
cor(students_df_new[, c("TimeReading", "TimeTV", "Happiness")])
```

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

```
#clearly we can see that there is no effect after changing the measurement
```

iii.

Choose the type of correlation test to perform, explain why you chose this test, and make a prediction if the test yields a positive or negative correlation?

```
# I chose the pearson method with .95 level of confidence with the prediction that the # correlation of

cor.test(students_df_new$TimeReading, students_df_new$TimeTV, alternative = "less",     method = "pearso
##
##  Pearson's product-moment correlation
##
## data:  students_df_new$TimeReading and students_df_new$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
```

```
# pearson method with .95 level of confidence with the prediction that the correlation # of Reading tim
cor.test(students_df_new$TimeReading, students_df_new$Happiness, alternative = "less",   method = "pear
```

```
##
##  Pearson's product-moment correlation
##
## data:  students_df_new$TimeReading and students_df_new$Happiness
## t = -1.4488, df = 9, p-value = 0.09067
## alternative hypothesis: true correlation is less than 0
## 95 percent confidence interval:
##  -1.0000000  0.1151482
## sample estimates:
##        cor
## -0.4348663
```

```
# pearson method with .95 level of confidence with the prediction that the correlation # of Tv Time tim
cor.test(students_df_new$TimeTV, students_df_new$Happiness, alternative = "greater",    method = "pears
```

```
##
##  Pearson's product-moment correlation
##
## data:  students_df_new$TimeTV and students_df_new$Happiness
## t = 2.4761, df = 9, p-value = 0.01761
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  0.1691762 1.0000000
## sample estimates:
##      cor
## 0.636556
```

iv. Perform a correlation analysis of:

1. All variables

```
cor(students_df_new, use = "complete.obs", 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
```

2. A single correlation between two a pair of the variables

```
cor(students_df_new$TimeReading, students_df_new$TimeTV, use = "complete.obs", method = "pearson")
```

```
## [1] -0.8830677
```

3. Repeat your correlation test in step but set the confidence interval at 99%

```
cor.test(students_df_new$TimeReading, students_df_new$TimeTV, alternative = "less", method = "pearson",
```

```
##
##  Pearson's product-moment correlation
##
## data:  students_df_new$TimeReading and students_df_new$TimeTV
## t = -5.6457, df = 9, p-value = 0.0001577
## alternative hypothesis: true correlation is less than 0
## 99 percent confidence interval:
##  -1.0000000 -0.5131843
## sample estimates:
##        cor
## -0.8830677
```

4. Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.

```
# The calculation suggest that the Reading time is inversely related to TV time at .99 # confidence lev
```

v.

Calculate the correlation coefficient and the coefficient of determination, describe what you conclude about the results.

```
library(Hmisc)
rcorr(as.matrix(students_df_new[, c("TimeReading", "TimeTV", "Happiness")]))
```

```
##             TimeReading TimeTV Happiness
## TimeReading        1.00  -0.88     -0.43
## TimeTV            -0.88   1.00      0.64
## Happiness         -0.43   0.64      1.00
##
## n= 11
##
##
## P
##             TimeReading TimeTV Happiness
## TimeReading             0.0003 0.1813
## TimeTV      0.0003             0.0352
## Happiness   0.1813      0.0352
```

```
# The output of the above correlation shows that
# Time TV  is negatively related to Reading Time with a Pearson correlation
# coefficient of r = -0.88 and the significance value is 0.0003 (close to 0).
# This significance value tells us that the probability of getting a correlation
# coefficient this big in a sample of 11 people if the null hypothesis
# were true (there was no relationship between these variables) is very low
# Hence, we can gain confidence that there is a genuine relationship between TVTime # # and ReadingTime
# so we can say that all of the correlation coefficients are significant.

# Coefficient of Determination for TimeTv vs TimeReading
coeffDet <- (-0.88) * (-0.88) * 100
coeffDet
```

```
## [1] 77.44
```

```
# the coefficient of determination came out to be 77.44%, this means that TVTime  is #highly correlated

# For all variables
cor(students_df_new)^2 * 100
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading 100.0000000  77.98085292  18.910873   0.80357143
## TimeTV       77.9808529 100.00000000  40.520352   0.00435161
## Happiness    18.9108726  40.52035234 100.000000   2.46527174
## Gender        0.8035714   0.00435161   2.465272 100.00000000
```

```
# TimeTv account for 40.52% of variation in Happiness
# Happiness accounts for 18.91% variation in TimeReading
```

vi.

Based on your analysis can you say that watching more TV caused students to read less? Explain.

Yes, based on the analysis on the calculation done in previous step we can say that watching more TV caused students to read less.

vii.

Pick three variables and perform a partial correlation, documenting which variable you are "controlling". Explain how this changes your interpretation and explanation of the results.

```
library(ggm)
students_df2 <- students_df_new[, c("TimeReading", "TimeTV", "Happiness")]

#Partial coorelation between TimeReading and TimeTV controlling Happiness
pc <- pcor(c("TimeTV", "TimeReading", "Happiness"), var(students_df2))
pc
```

```
## [1] -0.872945
```

```r
pc <- pc^2

pc
```

```
## [1] 0.762033
```

```r
pcor.test(pc, 1, 11)
```

```
## $tval
## [1] 3.328537
##
## $df
## [1] 8
##
## $pvalue
## [1] 0.01040702
```

```r
# so we see that the partial corelation between TimeReading and TimeTV came nearly
# same 76.20% keeping Happiness in control

#Partial coorelation between TimeTV and Happiness controlling TimeReading
pc2 <- pcor(c("TimeTV", "Happiness", "TimeReading"), var(students_df2))
pc2
```

```
## [1] 0.5976513
```

```r
pc2 <- pc2^2

pc2
```

```
## [1] 0.3571871
```

```r
pcor.test(pc2, 1, 11)
```

```
## $tval
## [1] 1.08163
##
## $df
## [1] 8
##
## $pvalue
## [1] 0.3109403
```

```r
# so we see that the partial corelation between TimeTV and Happiness came around
# 35.71% keeping TimeReading in control
```