Assignment: ASSIGNMENT 7.2 Student Survey

Name: Shekhar, Manish

Date: 2021-04-26

1. Is there a significant relationship between the amount of time spent reading and the time spent watching television?

2. What are other significant relationships that can be discovered?

```
## Load StudentSurvey.csv
setwd('/Users/mshekhar/Desktop/R Programming/DSC520/stats_for_data_science/stats_for_data_science')
studentData <- read.csv('./student-survey.csv')
head(studentData)
```

```
##   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
```

```
str(studentData)
```

```
## 'data.frame':    11 obs. of  4 variables:
##  $ TimeReading: int  1 2 2 2 3 4 4 5 5 6 ...
##  $ TimeTV     : int  90 95 85 80 75 70 75 60 65 50 ...
##  $ Happiness  : num  86.2 88.7 70.2 61.3 89.5 ...
##  $ Gender     : int  1 0 0 1 1 1 0 1 0 0 ...
```

```
summary(studentData)
```

```
##   TimeReading        TimeTV        Happiness         Gender      
##  Min.   :1.000   Min.   :50.00   Min.   :45.67   Min.   :0.0000  
##  1st Qu.:2.000   1st Qu.:67.50   1st Qu.:65.34   1st Qu.:0.0000  
##  Median :4.000   Median :75.00   Median :75.92   Median :1.0000  
##  Mean   :3.636   Mean   :74.09   Mean   :73.31   Mean   :0.5455  
##  3rd Qu.:5.000   3rd Qu.:82.50   3rd Qu.:83.83   3rd Qu.:1.0000  
##  Max.   :6.000   Max.   :95.00   Max.   :89.52   Max.   :1.0000
```

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.

```
## I would exclude gender as though gender is interpreted as integer it is more of a factor variable wi
## check if there are any NAs
sum(is.na(studentData$TimeReading))
```

```
## [1] 0
```

```
sum(is.na(studentData$TimeTV))
```

```
## [1] 0
```

```
sum(is.na(studentData$Happiness))
```

```
## [1] 0
```

```
## create a matrix of three variables as cov() needs matrix as input
studentMatrix <- as.matrix(studentData[,1:3])
studentMatrix
```

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

```
# As there are no NAs we do not need to specify use argument in the cov() function
cov(studentMatrix, method = "pearson")
```

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

```
cov(studentMatrix, method = "spearman")
```

```
##             TimeReading TimeTV Happiness
## TimeReading      10.650 -9.775      -4.4
## TimeTV           -9.775 10.900       6.2
## Happiness        -4.400  6.200      11.0
```

```
cov(studentMatrix, method = "kendall")
```

```
##             TimeReading TimeTV Happiness
## TimeReading          98    -82       -30
## TimeTV              -82    106        50
## Happiness           -30     50       110
```

Now though covariance results indicate below -

1. Time of reading is negatively related to Time of watching TV, which means that if one deviates from mean in positive direction other would deviate in negative direction.

2. Time of reading is negatively related to Happiness, which means that if one deviates from mean in positive direction other would deviate in negative direction.

3. Time of watching TV is positively related to Happiness, which means that if one deviates from mean in positive direction other would deviate in the same direction.

Though these results indicate direction of relationship we cannot use covariance values to determine the strength of relationship as covariance assumes variables to be measured using

same units of measurements. In this case Time of reading and Time of watching TV could be measured in minutes while Happiness cannot be measured in minutes and thus we cannot exactly determine the strength of relationship, for which we will need to calculate correlation coefficients.

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.

```
str(studentData)
```

Let's look at the data again

```
## 'data.frame':    11 obs. of  4 variables:
##  $ TimeReading: int  1 2 2 2 3 4 4 5 5 6 ...
##  $ TimeTV     : int  90 95 85 80 75 70 75 60 65 50 ...
##  $ Happiness  : num  86.2 88.7 70.2 61.3 89.5 ...
##  $ Gender     : int  1 0 0 1 1 1 0 1 0 0 ...
```

```
head(studentData)
```

```
##   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
```

```
summary(studentData)
```

```
##   TimeReading        TimeTV        Happiness         Gender
##  Min.   :1.000   Min.   :50.00   Min.   :45.67   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:67.50   1st Qu.:65.34   1st Qu.:0.0000
##  Median :4.000   Median :75.00   Median :75.92   Median :1.0000
##  Mean   :3.636   Mean   :74.09   Mean   :73.31   Mean   :0.5455
##  3rd Qu.:5.000   3rd Qu.:82.50   3rd Qu.:83.83   3rd Qu.:1.0000
##  Max.   :6.000   Max.   :95.00   Max.   :89.52   Max.   :1.0000
```

I am assmuning that unit of measurement used for TimeReading and TimeTV is same and lets suppose it's number of minutes. Looking at TimeReading 6 minutes is the maximum reading time and 1 minute being minimum. Time watching time is way more with minimum of 50 minutes and maximum of 95 minutes.

Looking at Happiness, clearly it cannot be measured in minutes and appears to be a % value which ranges from minimum of 45.67 to maximum of 89.52.

Gender is again not in minutes but is a nominal variable with two possible values 0 and 1, for male and female.

As variables are in mostly measured using different units, we cannot just rely on "covariance" to guage the strength of the relationship between them. "Covariance" can help with understanding the direction of relationship though.

```
studentDataSec <- data.frame(studentData$TimeReading*60, studentData$TimeTV*60)
colnames(studentDataSec) <- c("TimeReadingSec","TimeTVSec")
studentDataSec
```

Also, if we change the units of measurment the covariance value changes thus to understand the strength of relationship we use "Correlation". Let's see this with an example. Let's derive another data frame with TimeTv and TimeReading in seconds.

```
##    TimeReadingSec TimeTVSec
## 1              60      5400
## 2             120      5700
## 3             120      5100
## 4             120      4800
## 5             180      4500
## 6             240      4200
## 7             240      4500
## 8             300      3600
## 9             300      3900
## 10            360      3000
## 11            360      4200
```

```
## Covariance between TimeTV and TimeReading from original data frame
cov(studentData$TimeReading, studentData$TimeTV)
```

```
## [1] -20.36364
```

```
## Covariance between TimeTVSec and TimeReadingSec from new derived data frame
## with seconds as unit of measurement
cov(studentDataSec$TimeReadingSec, studentDataSec$TimeTVSec)
```

```
## [1] -73309.09
```

We can observe that magnitude of covariance has significantly increased with change in unit of measurement thus we really do not know the actual strength of relationship between two variables. We need "Correlation" to get correct guage of relationship strength along with direction. "Correlation" coefficient "r" is defined as Covariance(x,y)/[Standard deviation(x)*Standard deviation(y)]. Denominator here takes care of keeping the magnitude in check mostly creates an output between -1 and +1.

```
cor(studentData$TimeReading, studentData$TimeTV)
```

Let's try calculating correlation to see no impact because of unit of measurement change.

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

```
cor(studentDataSec$TimeReadingSec, studentDataSec$TimeTVSec)
```

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

We can see that irrespective of magnitude the correlation between two variables is same. Negative 0.88 shows that they are highly inversely correlated i.e. positive change in one will cause negative change in other.

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?

```
# run Shapiro-Wilk test to check normality
shapiro.test(studentData$TimeReading)
```

checking if the data distribution is normal

```
##
##  Shapiro-Wilk normality test
##
## data:  studentData$TimeReading
## W = 0.92093, p-value = 0.3265
```

```
shapiro.test(studentData$TimeTV)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  studentData$TimeTV
## W = 0.98681, p-value = 0.9923
```

```
shapiro.test(studentData$Happiness)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  studentData$Happiness
## W = 0.9412, p-value = 0.5347
```

As we can see p-value from Shapiro-Wilk test for each vaiable in the dataset is $>$ 0.05 and thus we can say that data is normally distributed. As data is normally distributed we can use parametric test i.e. Perason's correlation coefficient to check the correlation between variables.

Also, I would choose cor.test() function to perform correlation test because it allows additional flexibility i.e. -

1. It let's us specify if we want to perform 2 tailed test or one tailed test i.e. if we know about of alternate hypothesis being unknown or less than or greater than some value respectively. This can controlled using "alternate" argument's values. Default value is "two.sided". Other values are "less" or "greater".

2. It also let's us control "confidence interval" using another argument called "conf.level". Default width of confidence of correlation is 0.95. This is produced only for Pearson's correlation coefficient.

3. In addition like in cor() we can also control how NAs in the data should be considered while calculating correlation coefficient.

4. It can output all three correlations i.e. pearson's product momemt correlation, spearman's rho, and kendall's tau.

5. Output also is much detailed showing confidence level boundary, alternate hyposthesis being true or false (i.e. true if correlation exists), and p-value.

```
cor.test(studentData$TimeReading, studentData$TimeTV)
```

I can predict that test will show negative correlation between TimeTV and TimeReading variables. Also, correlation between TimeReading and Happiness could be negative, and that between TimeTV and Happiness should be positive :).

```
##
##  Pearson's product-moment correlation
##
## data:  studentData$TimeReading and studentData$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
```

```
cor.test(studentData$TimeReading, studentData$Happiness)
```

```
##
##  Pearson's product-moment correlation
##
## data:  studentData$TimeReading and studentData$Happiness
## t = -1.4488, df = 9, p-value = 0.1813
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8206596  0.2232458
## sample estimates:
##        cor
## -0.4348663
```

```
cor.test(studentData$TimeTV, studentData$Happiness)
```

```
##
##  Pearson's product-moment correlation
##
## data:  studentData$TimeTV and studentData$Happiness
## t = 2.4761, df = 9, p-value = 0.03521
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05934031 0.89476238
## sample estimates:
##      cor
## 0.636556
```

There exists significant negative correlation between TimeReading and TimeTV, with Pearson's correlation coefficient = -0.88 with p-value significantly < 0.05.

I predicted that correlation between TimeReading and Happiness will be negative. Per Peason's correlation coefficient = -0.43 it appears correct. But with p-value > 0.05 (p-value = 0.1813) it's insignificant and we cannot say mathematiocally that correlation exists.

I predicted that there exists positive correlation between TimeTV and Happiness. Looking at Pearson's correlation coefficient of 0.63 and p-value of 0.03 (<0.05) this seems statistically significant and correct.

Perform a correlation analysis of:

1. All variables

```
# Correlation analysis of all variables
cor(studentData)
```

```
##             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

```
# correlation between two variables
# cor() by default shows Pearson's product moment correlation coefficient
cor(studentData$TimeReading, studentData$Happiness)
```

```
## [1] -0.4348663
```

3. Repeat your correlation test in step 2 but set the confidence interval at 99%

```
# to set the confidence interval we need to use cor.test()
# by default confidence level is set to .95
cor.test(studentData$TimeReading, studentData$Happiness, conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  studentData$TimeReading and studentData$Happiness
## t = -1.4488, df = 9, p-value = 0.1813
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.8801821  0.4176242
## sample estimates:
##        cor
## -0.4348663
```

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

```
cor(studentData)
```

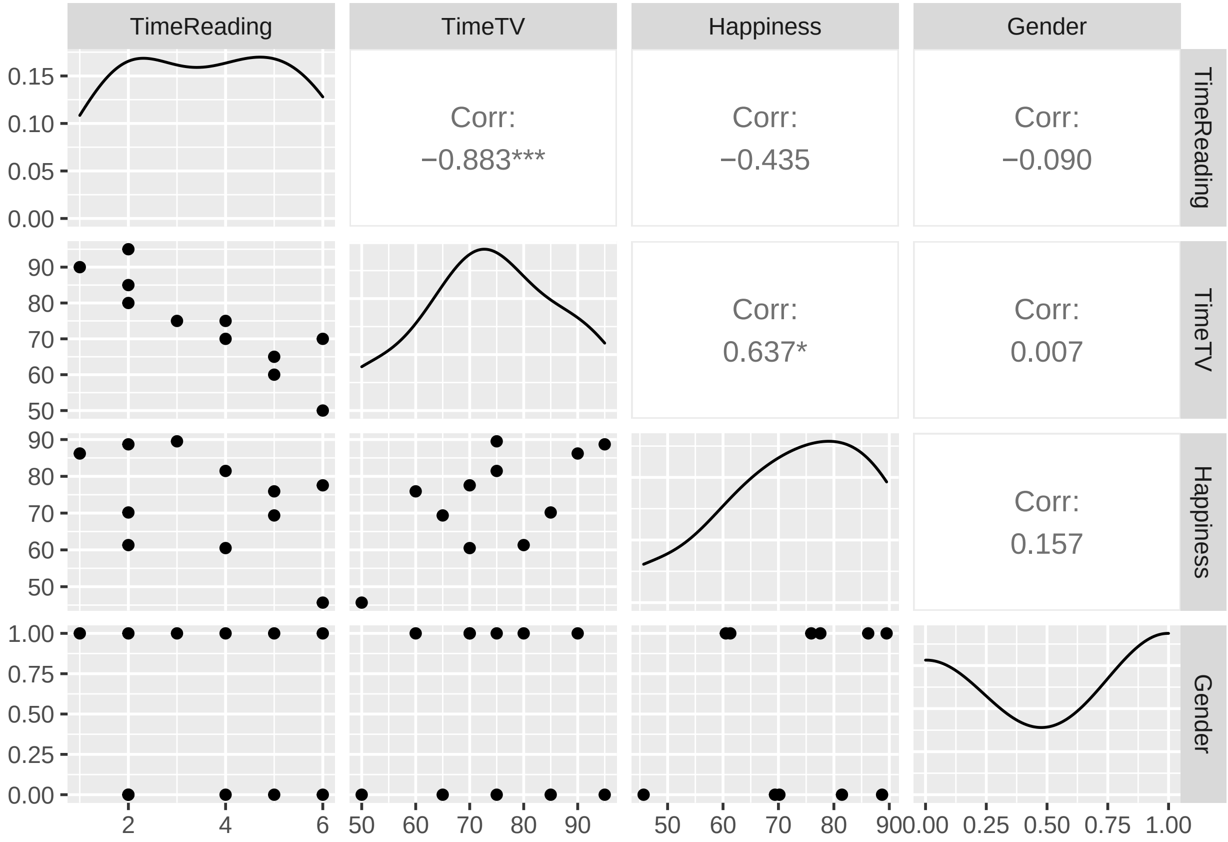
```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading  1.00000000 -0.883067681 -0.4348663 -0.089642146
## TimeTV      -0.88306768  1.000000000  0.6365560  0.006596673
## Happiness   -0.43486633  0.636555986  1.0000000  0.157011838
## Gender      -0.08964215  0.006596673 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

```
# Plotting correlations using GGally
library(GGally)
```

```
## Loading required package: ggplot2
```

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

```
library(ggplot2)
GGally::ggpairs(studentData)
```



Correlation matrix shows the mapping of correlation coefficient values between variables in the dataset. Correlation coefficient between a variable and itself is 1 i.e. completely positively correlated. If correlation coefficient is < 0 that would signify negative correlation. For e.g. correlation coeffient between TimeReading and Happiness is -0.43 which is significant negative correlation implying that reading time increase will bring happiness value down (this is not causal relationship proof). If correlation coeffient is exactly 0 or very close it would mean no significant relationship between two variables.

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

```
# calculating correlation coefficient (r) between variables
cor(studentData)
```

```
##             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
```

```
# calculating coefficient of determination (r^2) between two variables
cor(studentData)^2
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading 1.000000000 0.7798085292 0.18910873 0.0080357143
## TimeTV      0.779808529 1.0000000000 0.40520352 0.0000435161
## Happiness   0.189108726 0.4052035234 1.00000000 0.0246527174
## Gender      0.008035714 0.0000435161 0.02465272 1.0000000000
```

While correlation coefficient does not determine direction of causality, coefficient of determination (r^2) helps to understand "shared amount of variability" between two variables. For e.g. coefficient of determination between TimeReading and Happiness is 0.18, if we represent it as % it will be 18% and thus we can say that variability in TimeReading only accounts for 18% variability is Happiness, while remaining 82% should be caused by some other variables.

Looking at coeffiecient of determination between TimeTV and Happiness shared variability is about 40% which would imply that TV time variability effects Happiness upto 40% only, while remaining 60% variability in Happiness must be caused by some other variable.

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

```

cor(studentData$TimeReading, studentData$TimeTV)
```

Let's look at both correlation coefficient (r) and coefficient of determination (r^2) between TimeReading and TimeTV.

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

```
cor(studentData$TimeReading, studentData$TimeTV)^2
```

```
## [1] 0.7798085
```

As we know that using correlation coefficient we cannot determine the direction of causality between two variables though TimeReading and TimeTV are inversely proportional we cannot say that if increase in TimeReading caused TimeTV to decrease or if increase in TimeTV caused TimeReading to decrease.

Looking at coefficient of determination (r^2) we can say that variability in TimeReading can cause upto 77% variability in TimeTV but still there could be other variables that may cause rest 23% variability in TimeTV.

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.

```
# Picking up TimeReading, TimeTV, and Happiness to perform partial correlation
# creating new data frame with just above three variables
studentData2 <- studentData[,1:3]
```

```r
# Run partial correlation between TimeTV and Happiness while controlling TimeReading
library(ggm)
pc <- pcor(c("TimeTV","Happiness","TimeReading"), var(studentData2))
pc
```

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

```r
# R^2 or coefficient of determination using partial correlation is
pc^2
```

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

```r
# Original correlation between "TimeTV" and "Happiness" was
corr <- cor(studentData$TimeTV, studentData$Happiness)
corr
```

```
## [1] 0.636556
```

```r
# R^2 or coefficient of determination using cor() is
corr^2
```

```
## [1] 0.4052035
```

We can see that originally correlation coefficient between TimeTV and Happiness was 0.63 and coeffecient of determination was 40% i.e. variation in TimeTV should account for 40% variation in Happiness (shared variation).

With partial correlation keeping "TimeReading" in control we can see that correlation coefficient has decreased to 0.59 and coefficient of determination has decreased to 35%. This decrease suggests that variation in Happiness was also effected positively by TimeReading by about 5%.

```r

# run pcor.test() to get the pvalue associated with partial correlation
# pass partial correlation object, number of control variables, and number of observations
pcor.test(pc, 1, 11)
```

Let's see if new partial correlation between TimeTV and Happiness is statistically significant (check p-value < 0.05)

```
## $tval
## [1] 2.108388
##
## $df
## [1] 8
##
## $pvalue
## [1] 0.06804372
```

p-value of 0.06 which is > 0.05 suggests that we cannot reject the null hypothesis i.e. we cannot say with certainty mathematically that correlation exists between TimeTV and Happiness.