ASSIGNMENT 5 STUDENT SURVEY

Rajasekhar Reddy Karna

2020-10-03

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.

Table 1: Student Survey

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

a. 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.

Table 2: Covariance Matrix for Student Survey data set

	TimeReading	TimeTV	Happiness	Gender
TimeReading	3.0545455	-20.3636364	-10.350091	-0.0818182
TimeTV	-20.3636364	174.0909091	114.377273	0.0454545
Happiness	-10.3500909	114.3772727	185.451422	1.1166364
Gender	-0.0818182	0.0454545	1.116636	0.2727273

Calculating the covariance is an option to assess how two variables are related. Positive Covariance: Indicates variable deviates from the mean, the other variable deviates in the same direction. Negative Covariance: Indicates variable deviates from the mean (e.g., increases), the other deviates from the mean in the opposite direction (e.g., decreases). In student survey data set, covariance between 'Happiness' and 'TimeReading' shows that time of reading "TimeReading" is negatively impacting time being "Happiness". It has the negative covariance of -10.35. So if user being spend more timeon Time Reading, the reducing of Happiness like

from watching TV etc... to their daily routine.

b. 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.

** TimeReading: Using to represent how much time each student spends reading a day. TimeTV: Using to represent how much time each student spends watching TV a day. Happiness: Using to represent how much time each student spends being happy. DO not have specific scale to define how students being stay happy, like playing games, having fun/cultural activity, wathcing TV etc.. So has opporunity to factor this. Gender: Using to represents the gender of each student. We do not have exact represenstation of gender, like 0 mean men and 1 mean women or vice versa. So have opportunity to be converted as a factor.

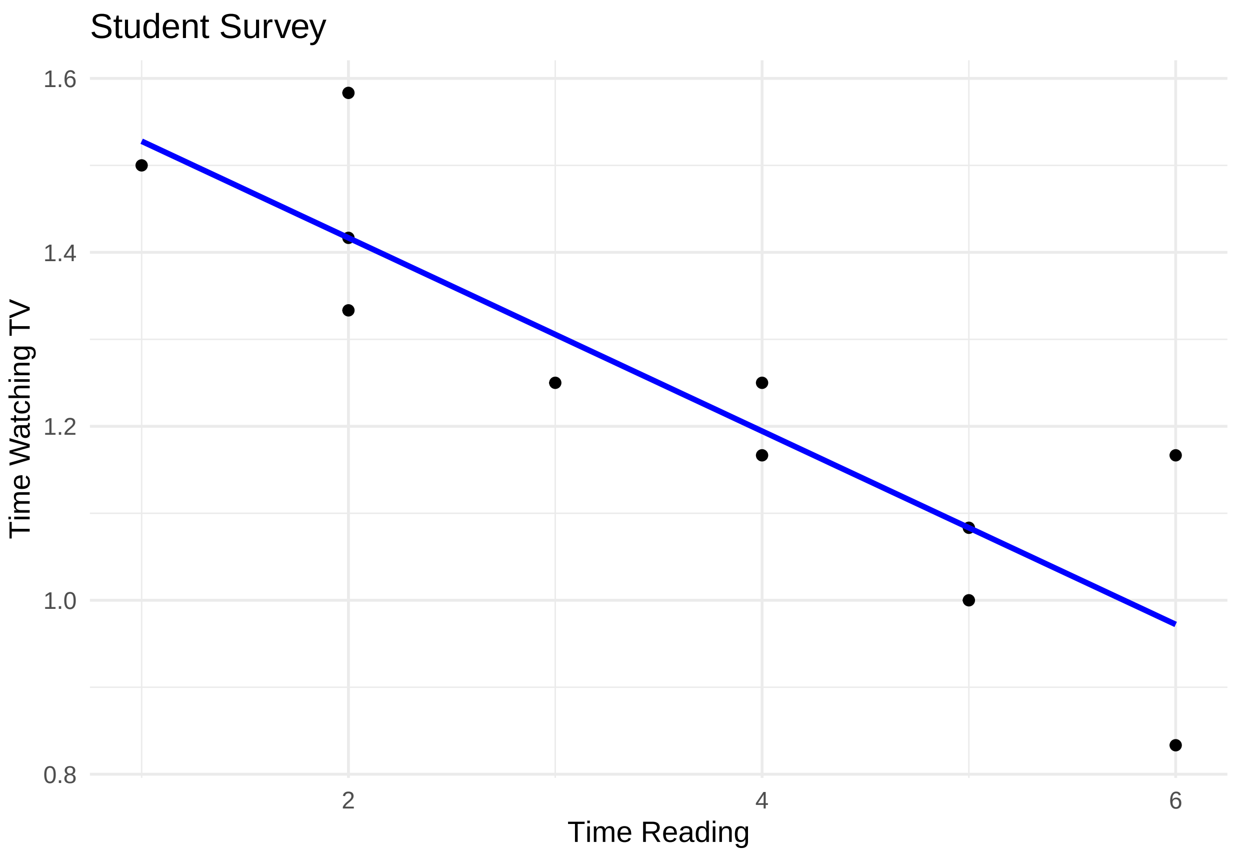
With the assumption of TimeTV given as in minutes and TimeReading given as in hours, using logic, has opporunity to conver TimeTV watching into hours.

Table 3: Covariance Matrix for Student Survey data set

	TimeReading	TimeTV	Happiness	Gender
TimeReading	3.0545455	-0.3393939	-10.350091	-0.0818182
TimeTV	-0.3393939	0.0483586	1.906288	0.0007576
Happiness	-10.3500909	1.9062879	185.451422	1.1166364
Gender	-0.0818182	0.0007576	1.116636	0.2727273

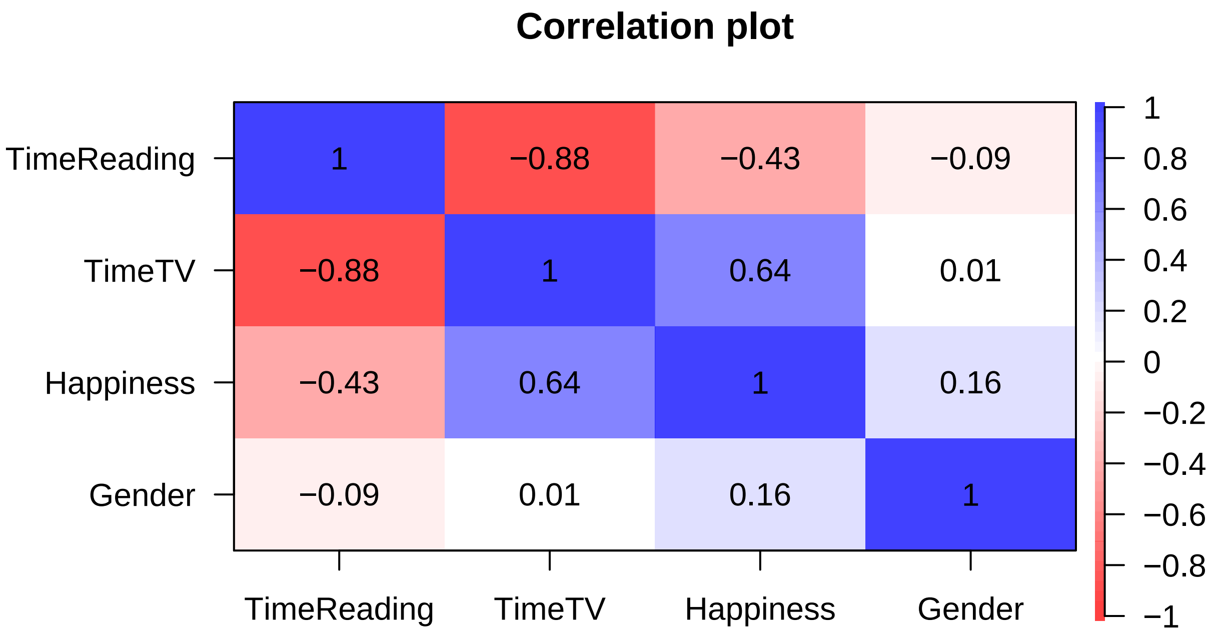
c. Choose the type of correlation test to perform, explain why you chose this test, and make a prediction if the test yields a positive or negative correlation?

**Pearson's Correlation: Though all types of correlation tests all return p-values. Indicating a high level of correlation in Pearson's Correlation may show confidence intervals as well as interval data



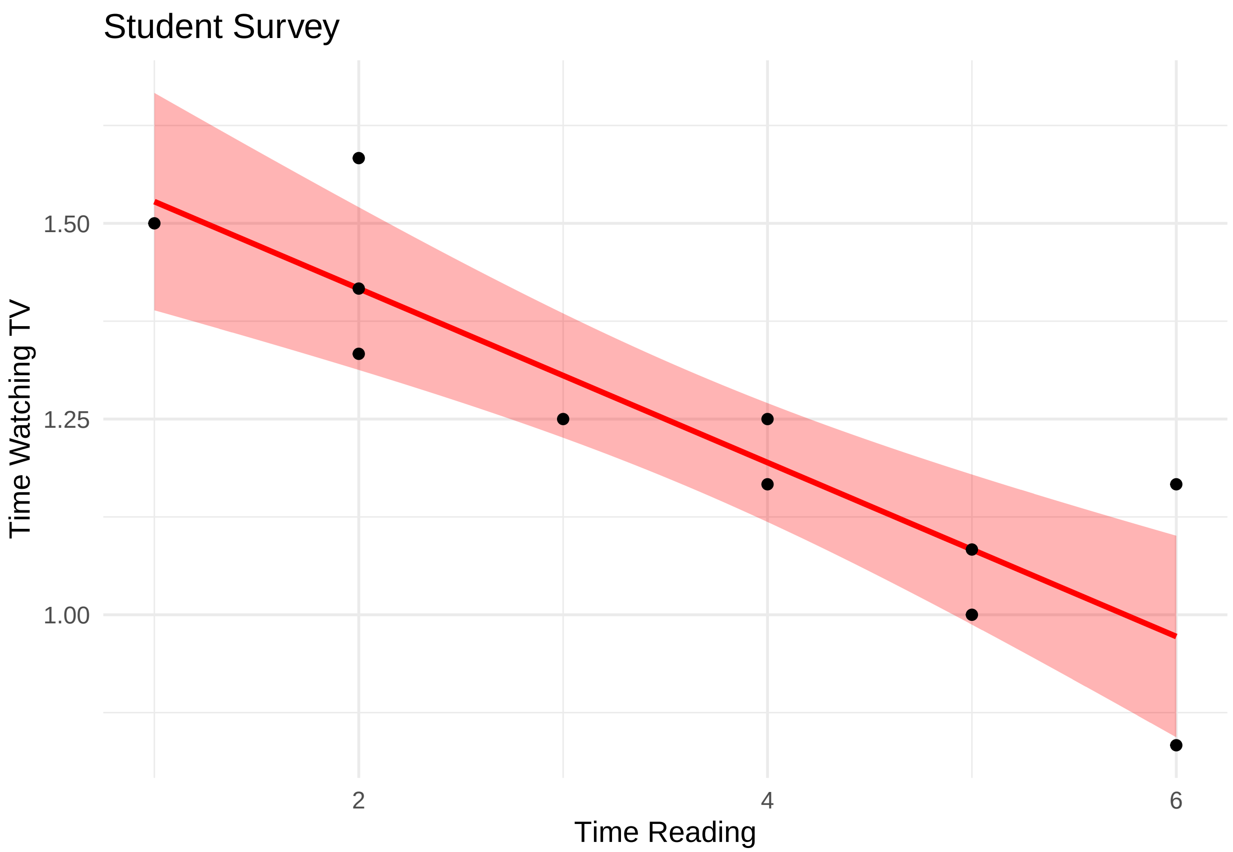
d. Perform a correlation analysis of:

```
1. All variables
```



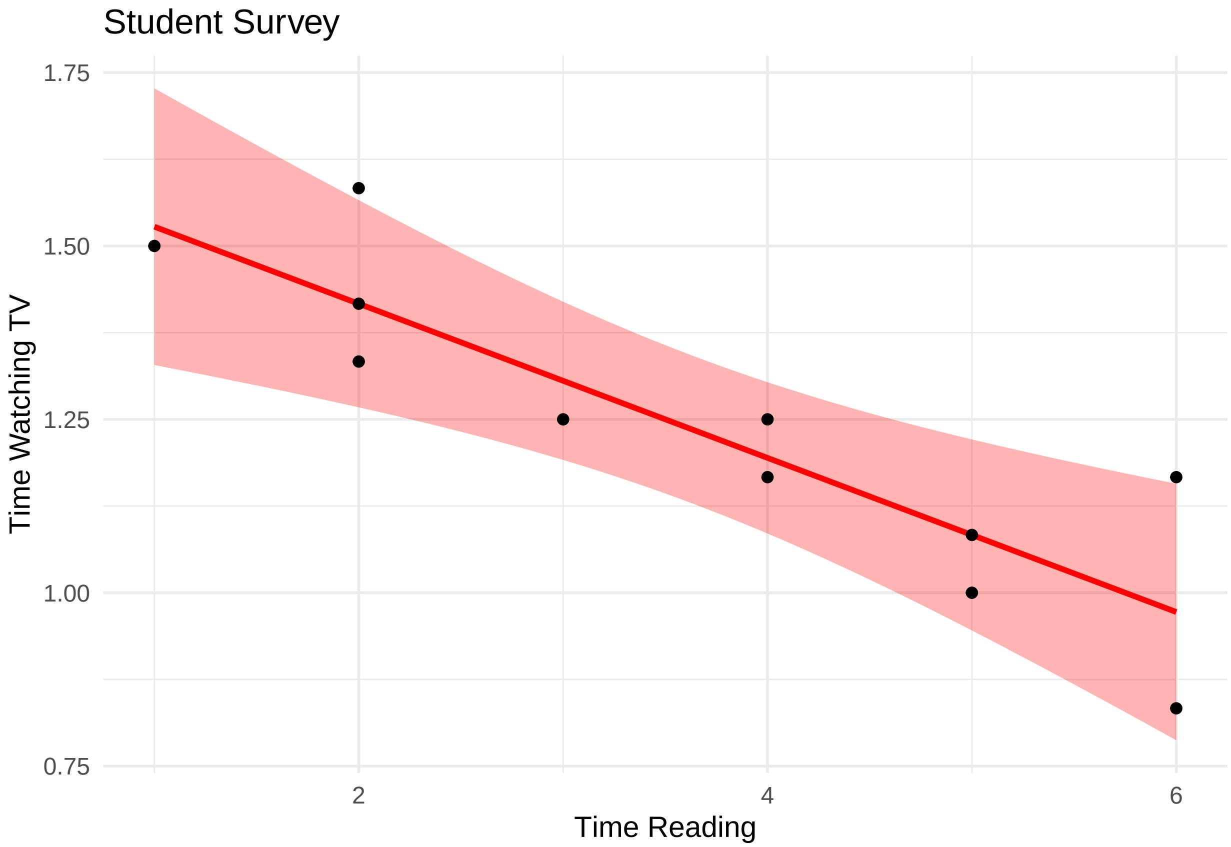
2. A single correlation between two a pair of the variables

```
##
##  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
```



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

```
## 
##  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
## 99 percent confidence interval:
##  -0.9801052 -0.4453124
## sample estimates:
##        cor 
## -0.8830677
```



```
4. Describe what the calculations in the correlation matrix suggest about the relationship between the
```

**The correlation between `TimeTV` and `TimeReading` are highly negatively correlated. Means if students read more they are less happy and watching TV student are more happier since they had positive correlation. Also `Gender` is also some what correlated with `Happiness`.

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

```
##              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
```

```
##              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
```

** Correlation Coefficient: Indicating a high level of negative correlation. Coefficident of Determination: Indicating correlation coefficient squared, is a measure of the amount of variability in one variable that is shared by the other. In student survey correlation coefficient describes watching TV is negatively related to reading. With Coefficient of Determination shows the percent of reading is affected by watching TV. Matrix shows that the ~77% of the time the reading is affected by watching TV.

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

**Based on correlation test on student survey report, at high confidence that students wathcing more TV invest less time on reading. Also based on coefficient of determination determines as much as ~77% of the time reading time is affected by watching TV.

g. 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.

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

**Partial correlation analysis on attributes TimeTv, TimeReading and Happiness represents that the time watching TV is negatively affecting reading time as it showed negative correlation.

References

(1) Slake for students feedback and reference comments.
(2) DISCOVERING STATISTICS USING R – ANDY FIELD | JEREMY MILES | ZOË FIELD
(3) R for Everyone – Second Edition