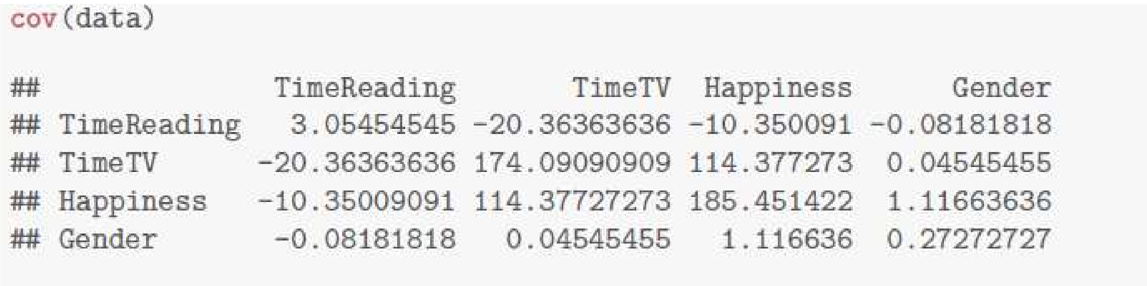
The Github repo link for the linked R file:

https://github.com/tang-xin72/DSC520/tree/main/completed/assignment05

Week7 exercise:

Student Survey Assignment :

1. 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.



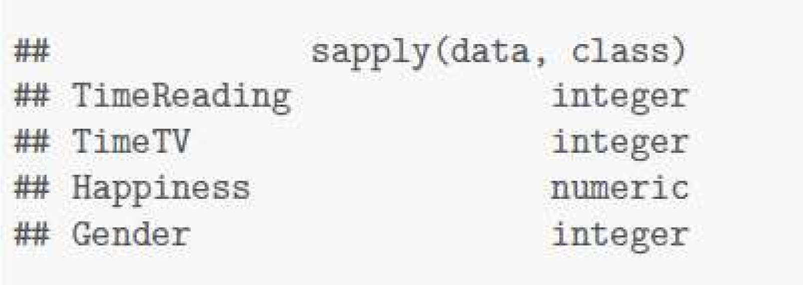
```
cov(data)

##              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
```

A: Covariance is a measure of the relationship between two random variables and to what extent they change together. So this give us a rough idea of the relationship between the listed variables.

As shown above,

- The time reading and time TV have a strong inverse relationship. One increase, the other will decrease.
- The time TV have strong positive relationship with Happiness. One increase, the other increase as well.
- Gender had little relation with any other variables.

2. 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.



```
##              sapply(data, class)
## TimeReading            integer
## TimeTV                 integer
## Happiness              numeric
## Gender                 integer
```

A: From data we can see the time of Reading is likely measured by hour while time on TV is measured by minutes). Happiness was measured as numeric and bear big values.

Since the covariance is impacted by variable scales. So the timeReading variable has a weaker effect when compared to TimeTV variable. Also the happiness has big value so if it will also have a dominating effect on other variables. Last Gender has 1 or 0, so it will play a very weak role in relationship.

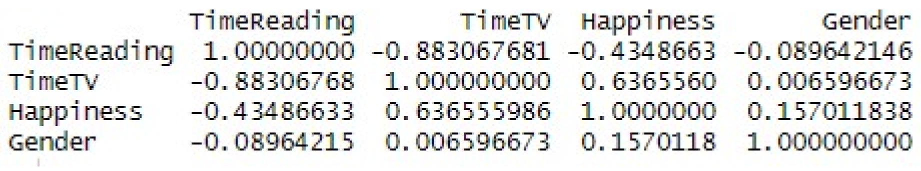
I would like to use same unit to measure reading and TV time, also convert happiness into a small number, so it will have similar weight when calculating covariance.

3. 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?

A: I do not know which model is the best, so I chose all 3 methods (pearson", "kendall", "spearman. Based on common sense, I predict time on reading and TV has negative correlation, also TV time and happiness are positive correlated.

4. Perform a correlation analysis of:
   - All variables
   - A single correlation between two a pair of the variables
   - Repeat your correlation test in step 2 but set the confidence interval at 99%
   - Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.

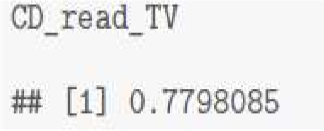
A: see R knit pdf for the calculation result, which shown below. As predicted, the reading and TV time has negative correlation; TV time and happiness are positive correlated, it also have a stronger correlation with happiness than reading. Gender has little effect.



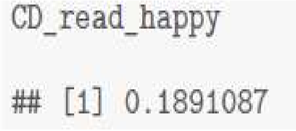
```
            TimeReading       TimeTV  Happiness       Gender
TimeReading  1.00000000 -0.883067681 -0.4348663 -0.089642146
TimeTV      -0.88306768  1.000000000  0.6365560  0.006596673
Happiness   -0.43486633  0.636555986  1.0000000  0.157011838
Gender      -0.08964215  0.006596673  0.1570118  1.000000000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

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

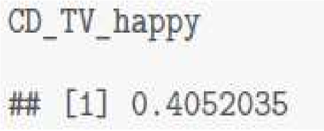
   A: see R knit pdf for the calculation result. The coefficient of determination is the square of correlation coefficient, which turn everything into positive and magnify the impact (since the coefficient is between -1 and 1)



```
CD_read_TV

## [1] 0.7798085
```



```
CD_read_happy

## [1] 0.1891087
```



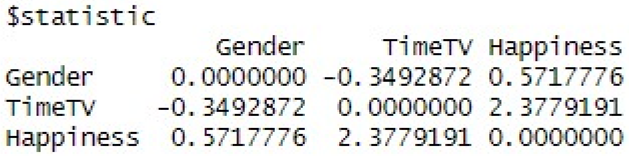
```
CD_TV_happy

## [1] 0.4052035
```

As result shown here. The reading time and TV time are still strongly correlated. The impact of TV time to happiness is stronger than reading time.

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

   A: Yes. With correlation coefficient as -0.88 and coefficient of determination as 0.78. watching more TV caused students to read less.

7. 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.



```
$statistic
              Gender     TimeTV Happiness
Gender     0.0000000 -0.3492872 0.5717776
TimeTV    -0.3492872  0.0000000 2.3779191
Happiness  0.5717776  2.3779191 0.0000000
```

   A: I am using gender as controlling variables. As shown here. The TV time on happiness become much stronger.