exercise_0702_PhillipsEmily

Emily Phillips

7/25/2021

Calculate the covariance of the Survey variables

```
##              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
```

The covariance between variables allows us to measure the relationship between them. If there were a relationship between any two variables in a data set, then as one variable deviates from its mean, the other variable should deviate from its mean in the same or the opposite way.

There are some greater covariance values between some of the variables, both in the positive and negative direction. For example with TimeTV & Happiness, they have a covariance of 114.377, which indicates that as time watching TV deviates from the mean, the happiness of the consumer deviates in the same direction, which could signify a relationship.

Also, for a negative covariance example, TimeReading and TimeTV have a relatively high (negative) value which indicates that as time watching TV deviates from the mean, the time spent reading deviates in the opposite direction from the mean, possibly indicating an inverse relationship.

Examine the Survey variables

```
##   TimeReading        TimeTV         Happiness         Gender
##  Min.   :1.000   Min.   :0.8333   Min.   :45.67   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:1.1250   1st Qu.:65.34   1st Qu.:0.0000
##  Median :4.000   Median :1.2500   Median :75.92   Median :1.0000
##  Mean   :3.636   Mean   :1.2348   Mean   :73.31   Mean   :0.5455
##  3rd Qu.:5.000   3rd Qu.:1.3750   3rd Qu.:83.83   3rd Qu.:1.0000
##  Max.   :6.000   Max.   :1.5833   Max.   :89.52   Max.   :1.0000
```

```
##   TimeReading   TimeTV Happiness Gender
## 1           1 1.500000     86.20      1
## 2           2 1.583333     88.70      0
## 3           2 1.416667     70.17      0
## 4           2 1.333333     61.31      1
## 5           3 1.250000     89.52      1
## 6           4 1.166667     60.50      1
```

I am assuming that each record in the dataset represents a weekly amount of hours spent reading & watching TV.

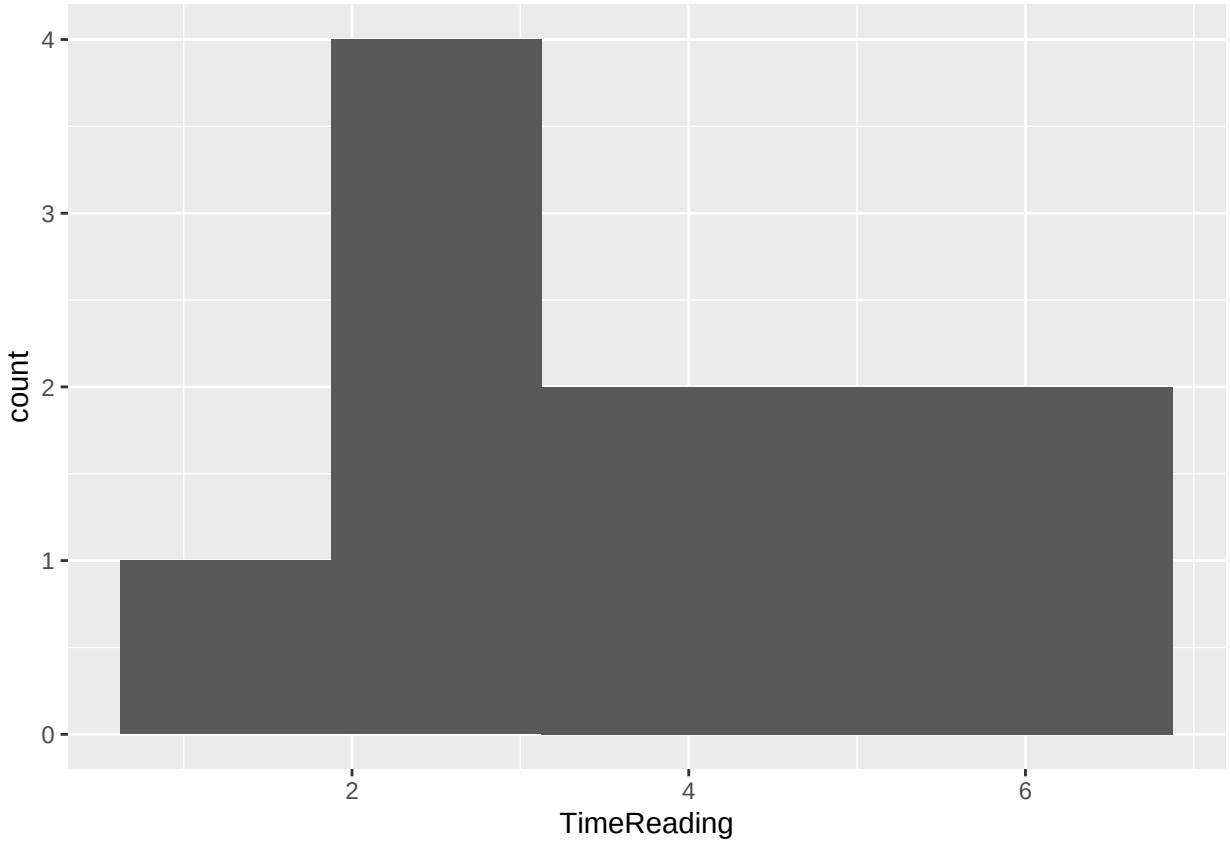
Measurements: - TimeReading: hours spent reading - TimeTV: hours spent watching TV - Happiness: percentage score of one's happiness - Gender: binary value for gender (0 or 1)

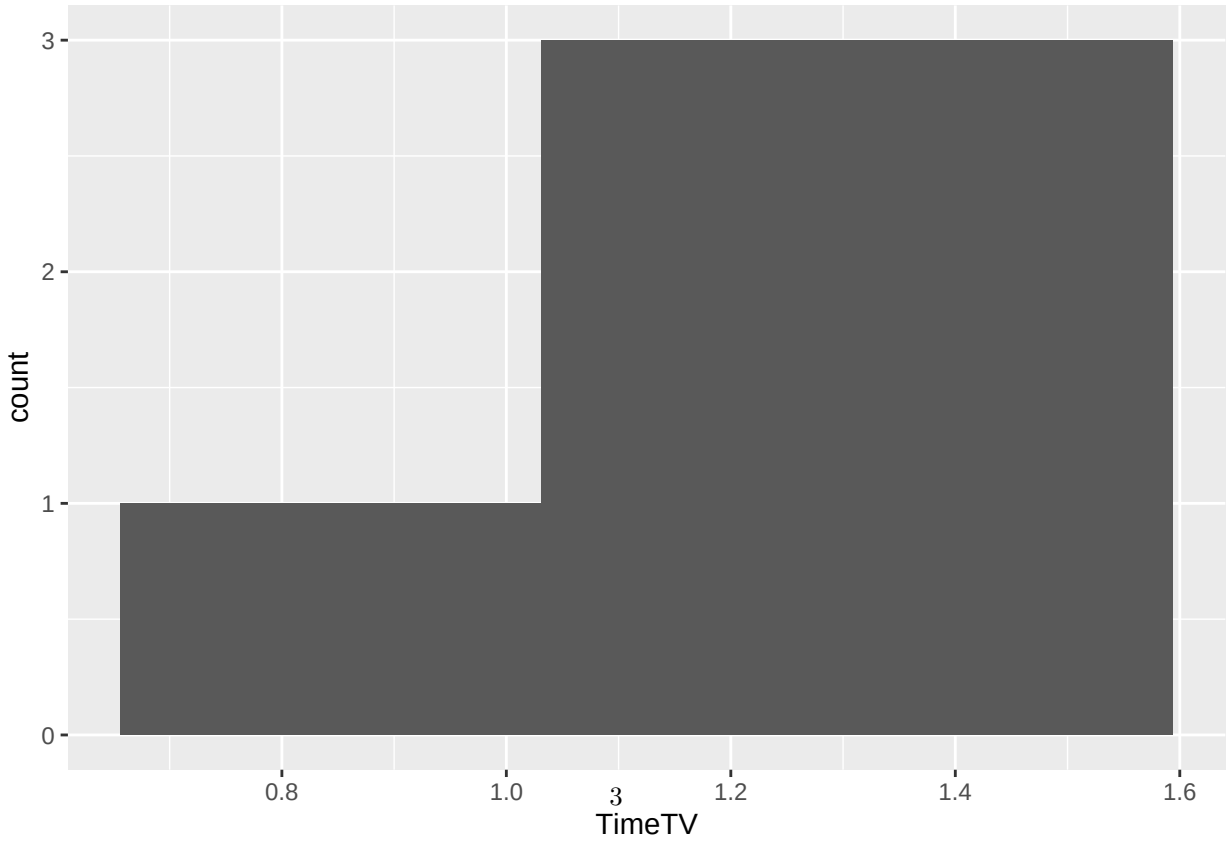
Luckily, in terms of the time variables from this dataset, they seem to both be in hours of measurement. However, happiness is given as a score out of 100, which does not really have an unit of measurement.Therefore, if we want to assess the covariance between any of the time variables and happiness, it would be difficult to compare in an objective way since the measures are not standardized.

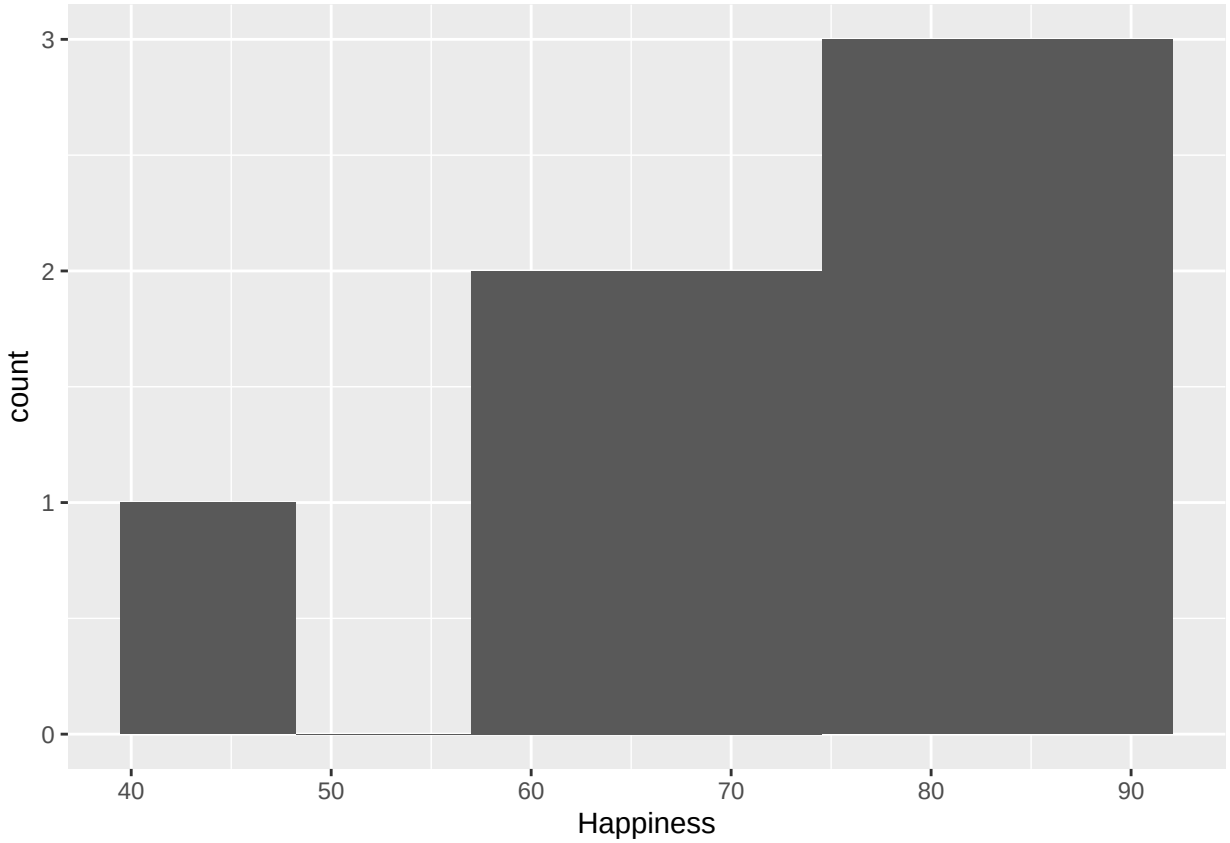
If we changed the measurements of our variables, we would not be able to compare the covariance values among the variables. Everything would have to be the measured in the same units to allow for this accurate comparison. Since it seems like both Time variables are measured in hours, it might not be a huge problem, since our question pertains to these variables and they are being measured in the same units. However, if we want to do any comparisons with the time variables and happiness, the comparison will not be objective, since the datasets have different units.

Type of correlation to perform

Testing normality







I have decided to use the Spearman's correlation coefficient. Not only are the variables in this sample not normally distributed, but the happiness variable has a type of score ranking to it, which is very optimal for use of the Spearman's correlation coefficient.

I predict that the Spearman's test will be negative!

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%

All variables

```
##             TimeReading      TimeTV  Happiness      Gender
## TimeReading  1.00000000 -0.90725363 -0.4065196 -0.08801408
## TimeTV      -0.90725363  1.00000000  0.5662159 -0.02899963
## Happiness   -0.40651964  0.56621595  1.0000000  0.11547005
## Gender      -0.08801408 -0.02899963  0.1154701  1.00000000
```

Correlation between TimeReading & TimeTV

```
## [1] -0.9072536
```

Correlation between TimeReading & TimeTV with 99% confidence interval

```
## [1] -0.9888904 -0.4051469
```

From the correlation matrix, we can assess the following variable relationships: 1. TimeTV/TimeReading – strong (high) negative/inverse relationship. As the time watching TV increases, the time spent reading decreases. 2. TimeTV/Happiness – relatively high positive relationship. As the time spent watching TV increases, the happiness of an individual increases. 3. Happiness/TimeReading – moderate negative correlation. As the time spent reading increases, the happiness of an individual decreases.

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

```
## "-0.907254" "0.823109"
```

The coefficient of determination is a measure of the amount of variability in one variable that is shared by the other.

The variables of TimeReading & TimeTV have a strong negative correlation of -0.907254. The value of R^2 or coefficient of determination is then 0.823109. From converting this into a percentage (82.3), we can say that the time spent reading shares 82.3% of the variability in time spent watching TV. This only leaves 17.7% of the variability to be accounted for by other variables, which is definitely not the majority.

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

We can not say that watching more TV "caused" students to read less as correlation does not imply causation. There can be other variables/factors involved in influencing this relationship between TV time and reading time that we have not fully explored yet.

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.

```
## "-0.872945" "0.762033"
```

I am going to be controlling for happiness score to truly investigate just the relationship between TV time and reading time!

The partial correlation is equal to -0.873, and the partial coefficient of determination is 0.762.

```
## $tval
## [1] -5.061434
##
## $df
## [1] 8
##
## $pvalue
## [1] 0.0009753126
```

The partial correlation between TV time and reading time is less than the correlation when the effect of happiness is not controlled for. However, it is not a huge or considerable difference between -0.873 and -0.907. The partial correlation is statistically significant, with a p-value of 0.001, which is $p < 0.01 < 0.05 < 0.1$.

In terms of variance, the value of R2 is also slightly less when controlling for happiness than when not (0.762 vs. 0.823). The value of R2 for the partial correlation is 0.762, which shows that TV time can still account for 76.2% of the variation in reading time, even when the happiness of the student is taken into account.

The inclusion of the Happiness control variable did not drastically impact my interpretation of the results, and if anything, it makes me feel a bit more confident about the relationship between TV time and reading time, and their inverse relationship!