Assignment7.2

Lincoln Brown

First, I will load the suvey data into a variable and then print basic information about the dataset.

```
library(ggplot2)
library(purrr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(pastecs)
```

```
##
## Attaching package: 'pastecs'
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
library(ppcor)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(stringr)

survey <- read.csv("/media/x/disk/School/DSC520/git/dsc520/data/student-survey.csv")
print(colnames(survey))
```

```
## [1] "TimeReading" "TimeTV"      "Happiness"   "Gender"
```

```
print(dim(survey))
```

```
## [1] 11  4
```

```
print(str(survey))
```

```
## '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 ...
## NULL
```

```
print(head(survey))
```

```
##   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
```

```
survey |> stat.desc(norm=TRUE)
```

```
##                TimeReading       TimeTV   Happiness        Gender
## nbr.val       11.000000000  11.00000000  11.0000000 11.0000000000
## nbr.null       0.000000000   0.00000000   0.0000000  5.0000000000
## nbr.na         0.000000000   0.00000000   0.0000000  0.0000000000
## min            1.000000000  50.00000000  45.6700000  0.0000000000
## max            6.000000000  95.00000000  89.5200000  1.0000000000
## range          5.000000000  45.00000000  43.8500000  1.0000000000
## sum           40.000000000 815.00000000 806.3800000  6.0000000000
## median         4.000000000  75.00000000  75.9200000  1.0000000000
## mean           3.636363636  74.09090909  73.3072727  0.5454545455
## SE.mean        0.526959154   3.97824663   4.1059981  0.1574591643
## CI.mean.0.95   1.174138165   8.86408589   9.1487338  0.3508408816
## var            3.054545455 174.09090909 185.4514218  0.2727272727
## std.dev        1.747725795  13.19435141  13.6180550  0.5222329679
## coef.var       0.480624594   0.17808327   0.1857668  0.9574271078
## skewness      -0.002533230  -0.11848577  -0.5162276 -0.1582524145
## skew.2SE      -0.001917116  -0.08966855  -0.3906746 -0.1197634442
## kurtosis      -1.642178979  -1.03762883  -0.9143551 -2.1460055096
## kurt.2SE      -0.641769076  -0.40550884  -0.3573331 -0.8386661817
## normtest.W     0.920928865   0.98680678   0.9411966  0.6491717530
## normtest.p     0.326452517   0.99233227   0.5346664  0.0001051734
```

```
for(x in survey){
  print(shapiro.test(x))
}
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.92093, p-value = 0.3265
##
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.98681, p-value = 0.9923
##
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.9412, p-value = 0.5347
##
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.64917, p-value = 0.0001052
```

The results of normtest.W (Shapiro-Wilkes) and normtest.p (probability) indicate that these data are not normally distributed.

```
rt <- data.frame(survey$TimeReading, survey$TimeTV)
rh <- data.frame(survey$TimeReading, survey$Happiness)
rg <- data.frame(survey$TimeReading, survey$Gender)

th <- data.frame(survey$TimeTV, survey$Happiness)
tg <- data.frame(survey$TimeTV, survey$Gender)

hg <- data.frame(survey$Happiness, survey$Gender)


combinations <- list(rt, rh, rg, th, tg, hg)

cov_results <- map(combinations, cov)
```

All possible combinations of variables rt <- (TimeReading, TimeTV) (TimeReading, Happiness) (TimeReading, Gender)

(TimeTV, Happiness) (TimeTV, Gender)

(Happiness, Gender)

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.

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.

```
for(df in combinations){
  print(cov(df[1], df[2]))
  cat("\n\n")
}
```

```
##                     survey.TimeTV
## survey.TimeReading     -20.36364
##
##
##                     survey.Happiness
## survey.TimeReading        -10.35009
```

```
##
##
##                     survey.Gender
## survey.TimeReading   -0.08181818
##
##
##                survey.Happiness
## survey.TimeTV          114.3773
##
##
##                survey.Gender
## survey.TimeTV    0.04545455
##
##
##                   survey.Gender
## survey.Happiness      1.116636
```

Covariance helps us detect relationships between variables. Covariance is not effective at detecting the magnitude of the relationships and it is not always easy to understand because it is a product of the two variables, which is difficult to understand how it is measured.

What it does tell us is that positive covariance values indicate a positive relationship (both variables move in the same direction) and negative covariance values indicate a negative relationship where the variables move in opposing directions (one increases when the other decreases or vice versa)

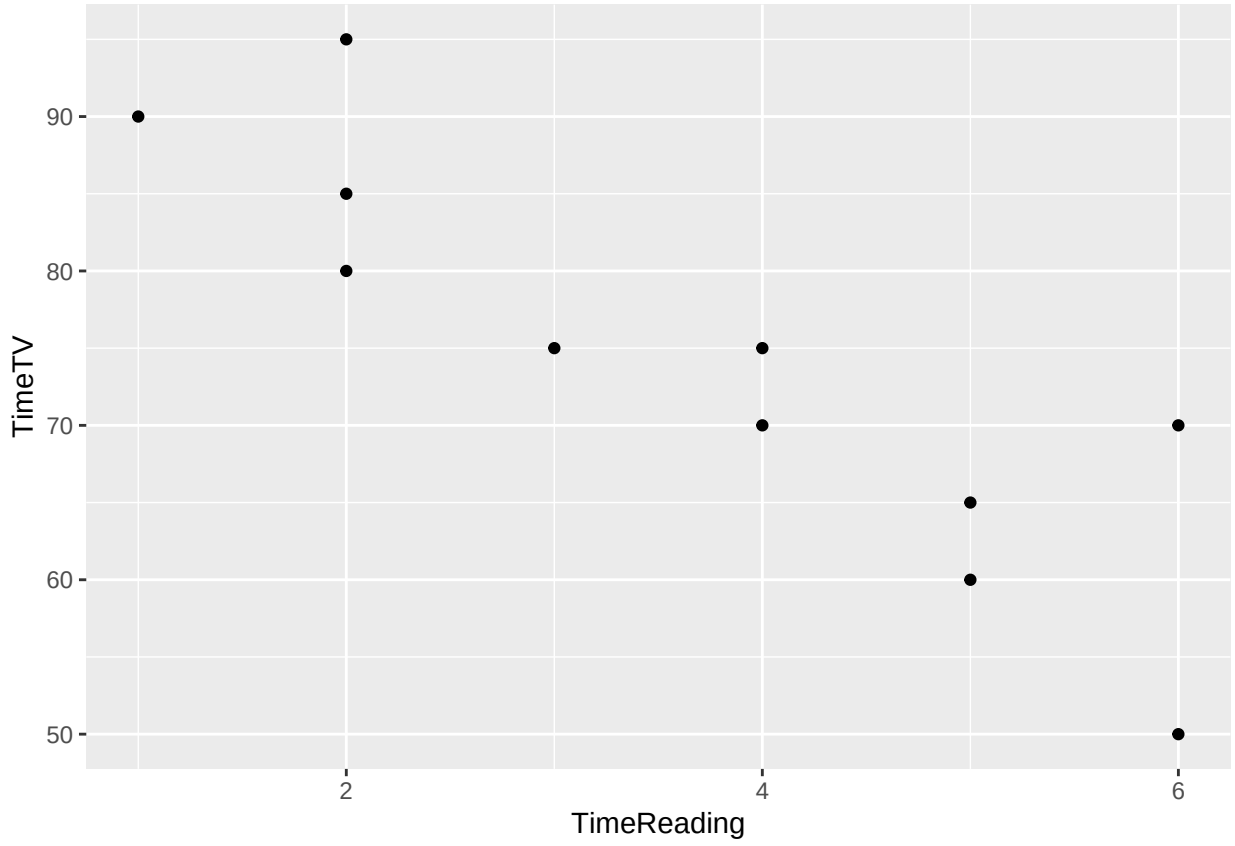
2. Examine the Survey data variables. What measurement is being used for the variables? Explain what effect changing the measurement being used for the variables would have on the covariance calculation. Would this be a problem? Explain and provide a better alternative if needed.

The Survey data variables appear to all have different measurement variables. The time spent reading appears to be measured in hours, the time spent watching tv in minutes, the happiness score seems to range from 0-100, although the min score is 45.67 and the max is 89.52, so we cannot be sure what the boundaries are. Gender is measured as a binary, either 1 for male or female and 0 for the opposite gender. Without a codebook, we cannot be certain which gender is represented by 1.

Since covariance is a product, the covariance of cov(TimeReading, TimeTV) would be -20.36364 hourMinutes. Which makes no sense. Perhaps even more confusing would be cov(Happiness, Gender) which would be measured as something like scoreGender.

Changing the measurement variable for TimeReading to minutes it would be the same measuring factor as time spent watching TV. This change would not be a problem.

```
ggplot(survey, aes(x=TimeReading, y=TimeTV)) + geom_point()
```




```
convert_df <- survey
convert_df$TimeReading <- survey$TimeReading * 60

m <- cov(convert_df$TimeReading, convert_df$TimeTV)
class(m)
```

```
## [1] "numeric"
```

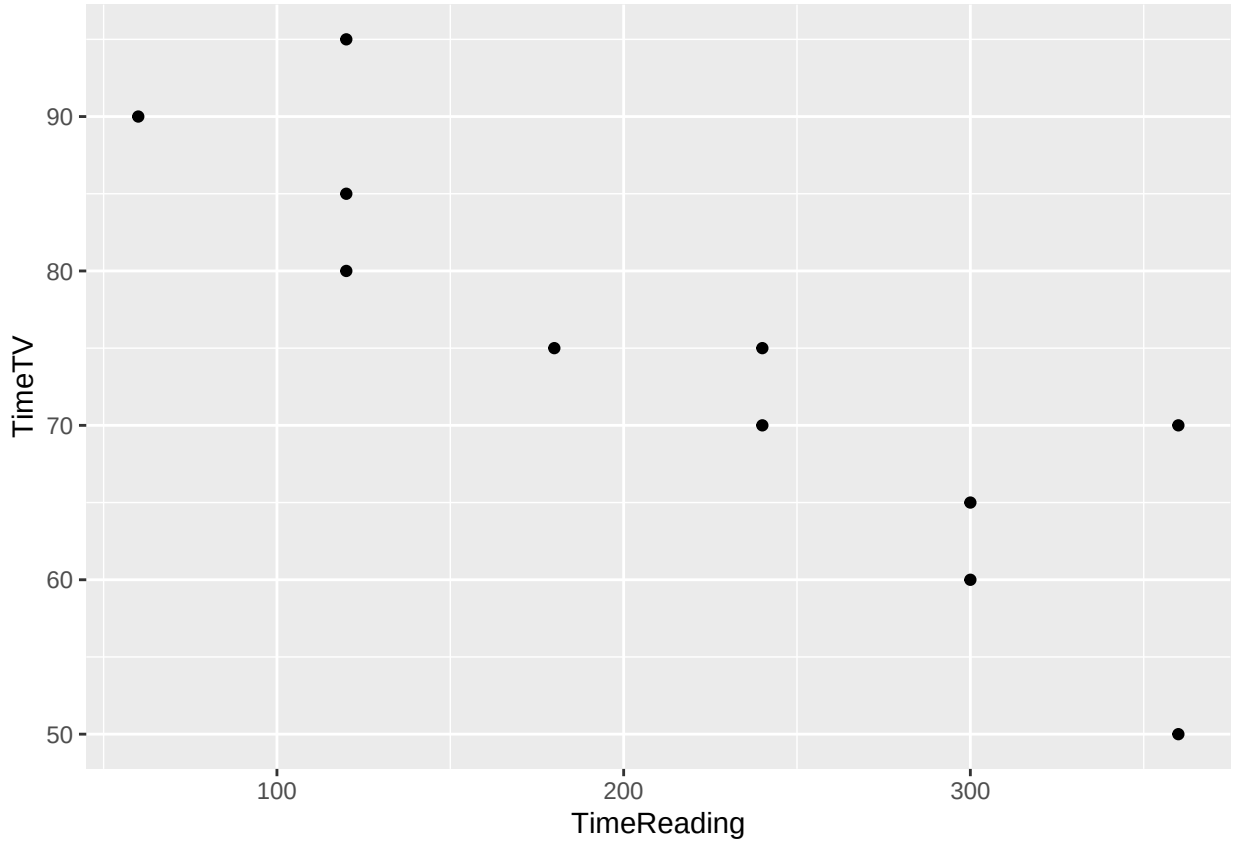
```
sqrt(abs(m))
```

```
## [1] 34.95452
```

```
print(convert_df)
```

```
##    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
```

```
ggplot(convert_df, aes(x=TimeReading, y=TimeTV)) + geom_point()
```



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? Since the data are not normally distributed, the Spearman correlation coefficient is a better choice.

I predict the variables with negative covariance will have negative correlation.

```
test_fun <- function(df) {
  names <- colnames(df)
  result <- cor.test(df[,1], df[,2], mode="spearman")
  return(result)
}

for(df in combinations){
  names <- colnames(df)
  print(names)
  print(test_fun(df))
}
```

```
## [1] "survey.TimeReading" "survey.TimeTV"
##
##  Pearson's product-moment correlation
##
## data:  df[, 1] and df[, 2]
## 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
##
## [1] "survey.TimeReading" "survey.Happiness"
##
##  Pearson's product-moment correlation
##
## data:  df[, 1] and df[, 2]
## 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
##
## [1] "survey.TimeReading" "survey.Gender"
##
##  Pearson's product-moment correlation
##
## data:  df[, 1] and df[, 2]
## t = -0.27001, df = 9, p-value = 0.7932
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6543311  0.5392294
## sample estimates:
##         cor
## -0.08964215
##
## [1] "survey.TimeTV"    "survey.Happiness"
##
##  Pearson's product-moment correlation
##
## data:  df[, 1] and df[, 2]
## 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
##
## [1] "survey.TimeTV" "survey.Gender"
##
##  Pearson's product-moment correlation
##
## data:  df[, 1] and df[, 2]
## t = 0.01979, df = 9, p-value = 0.9846
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5956354  0.6040812
## sample estimates:
##         cor
```

```
## 0.006596673
##
## [1] "survey.Happiness" "survey.Gender"
##
##  Pearson's product-moment correlation
##
## data:  df[, 1] and df[, 2]
## t = 0.47695, df = 9, p-value = 0.6448
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4889126  0.6917342
## sample estimates:
##       cor
## 0.1570118
```

4. Perform a correlation analysis of:
   a. All variables

```
for(df in combinations){
  print(cor(df[1], df[2], method="spearman"))
  cat("\n\n")
}
```

```
##                     survey.TimeTV
## survey.TimeReading    -0.9072536
##
##
##                     survey.Happiness
## survey.TimeReading       -0.4065196
##
##
##                     survey.Gender
## survey.TimeReading   -0.08801408
##
##
##                survey.Happiness
## survey.TimeTV         0.5662159
##
##
##                survey.Gender
## survey.TimeTV   -0.02899963
##
##
##                   survey.Gender
## survey.Happiness     0.1154701
```

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

```
cor(survey$TimeReading, survey$TimeTV, method="spearman")
```

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

```
pcor(survey, method="spearman")
```

```
## $estimate
##             TimeReading     TimeTV Happiness     Gender
## TimeReading   1.0000000 -0.9113002 0.3713077 -0.3424044
## TimeTV       -0.9113002  1.0000000 0.5576799 -0.3566278
```

```
## Happiness     0.3713077  0.5576799 1.0000000  0.2667871
## Gender       -0.3424044 -0.3566278 0.2667871  1.0000000
##
## $p.value
##              TimeReading       TimeTV Happiness    Gender
## TimeReading 0.0000000000 0.0006269124 0.3251823 0.3670649
## TimeTV      0.0006269124 0.0000000000 0.1187160 0.3461451
## Happiness   0.3251822825 0.1187159880 0.0000000 0.4877183
## Gender      0.3670649001 0.3461451404 0.4877183 0.0000000
##
## $statistic
##             TimeReading    TimeTV Happiness     Gender
## TimeReading   0.0000000 -5.855781 1.0580263 -0.9642003
## TimeTV       -5.8557807  0.000000 1.7775693 -1.0099565
## Happiness     1.0580263  1.777569 0.0000000  0.7323977
## Gender       -0.9642003 -1.009957 0.7323977  0.0000000
##
## $n
## [1] 11
##
## $gp
## [1] 2
##
## $method
## [1] "spearman"
```

The above results indicate a significant negative relationship between time spent reading versus time spent watching TV.

None of the other relationships have significant p-values.

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

```
test_fun <- function(df) {
  names <- colnames(df)
  result <- cor.test(df[,1], df[,2], method="spearman", exact=FALSE, conf.level = .99)
  return(result)
}

for(df in combinations){
  names <- colnames(df)
  print(names)
  print(test_fun(df))
}
```

```
## [1] "survey.TimeReading" "survey.TimeTV"
##
##  Spearman's rank correlation rho
##
## data:  df[, 1] and df[, 2]
## S = 419.6, p-value = 0.0001152
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.9072536
##
```

```
## [1] "survey.TimeReading" "survey.Happiness"
##
##  Spearman's rank correlation rho
##
## data:  df[, 1] and df[, 2]
## S = 309.43, p-value = 0.2147
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.4065196
##
## [1] "survey.TimeReading" "survey.Gender"
##
##  Spearman's rank correlation rho
##
## data:  df[, 1] and df[, 2]
## S = 239.36, p-value = 0.7969
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.08801408
##
## [1] "survey.TimeTV"    "survey.Happiness"
##
##  Spearman's rank correlation rho
##
## data:  df[, 1] and df[, 2]
## S = 95.432, p-value = 0.06939
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5662159
##
## [1] "survey.TimeTV" "survey.Gender"
##
##  Spearman's rank correlation rho
##
## data:  df[, 1] and df[, 2]
## S = 226.38, p-value = 0.9325
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.02899963
##
## [1] "survey.Happiness" "survey.Gender"
##
##  Spearman's rank correlation rho
##
## data:  df[, 1] and df[, 2]
## S = 194.6, p-value = 0.7353
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1154701
```

d. Describe what the calculations in the correlation matrix suggest about the relationship between the v
The correlation matrix indicates a negative relationship between TimeReading and TimeTV, the p-value of
It also appears that TimeTV and Happiness have positive correlation, with a significant p-value of 0.035

```
cor(survey, method="spearman")
```

```
##             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
```

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

```
cor_coef <- cor(survey$TimeReading, survey$TimeTV, method="spearman")
coef_deter <- cor_coef^2

output1 <- str_interp("Correlation Coefficient: ${cor_coef}")

output2 <- str_interp("Coefficient of Determination: ${coef_deter}")
print(output1)
```

```
## [1] "Correlation Coefficient: -0.907253627251832"
```

```
print(output2)
```

```
## [1] "Coefficient of Determination: 0.823109144161606"
```

The correlation coefficient -.90 indicates a strong positive correlation.

The coefficient of determination indicates that TimeReading shares 82% of the variability in TimeTV.

6. Based on your analysis can you say that watching more TV caused students to read less? Explain. Given the negative strong negative correlation account for 82% of the variability of TimeReading and TimeTV. I am 99% confident that watching TV negatively affects time spent reading.

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.

```
pcor.test(survey$TimeReading, survey$TimeTV, survey$Gender, method="spearman")
```

```
##     estimate      p.value statistic  n gp   Method
## 1 -0.9137348 0.0002180915 -6.360723 11  1 spearman
```

```
pcor.test(survey$TimeReading, survey$TimeTV, survey$Happiness, method="spearman")
```

```
##     estimate      p.value statistic  n gp   Method
## 1 -0.8990805 0.0004011345 -5.808771 11  1 spearman
```

```
pcor.test(survey$TimeTV, survey$Happiness, survey$TimeReading, method="spearman")
```

```
##    estimate   p.value statistic  n gp   Method
## 1 0.5137092 0.1288075  1.693531 11  1 spearman
```

```
pcor.test(survey$TimeReading, survey$Happiness, survey$TimeTV, method="spearman")
```

```
##    estimate   p.value statistic  n gp   Method
## 1 0.3091761 0.3847034 0.9195348 11  1 spearman
```

Comparing TimeReading and TimeTV, controlling for Gender results in little change. Comparing TimeReading and TimeTV, controlling for Happiness results in a higher p-value, but does not alter the significance.

Comparing TimeTV and Happiness, controlling for TimeReaading produces no significant findings. Comparing TimeReading and Happiness, controlling for TimeTV produces weak correlation with an insignificant pvalue.