Week7_Assignment_Part2_Guruprasad_VelikaduKrishnamoorthy

Guruprasad Velikadu Krishnamoorthy

2023-01-29

Loading Required Libraries

```
library(GGally, quietly = TRUE)
```

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

```
library(rstatix)
```

```
##
## Attaching package: 'rstatix'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(magrittr)
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(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
library(ggm)
```

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.

```r
# Reading from the student-survey file
student_survey <- read.csv("student-survey.csv")
# Calculating the Covariance of all fields in the DataFrame
cov(student_survey)
```

```
##               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
```

```r
# Calculating the covariance excluding Gender- the Categorical variable
cov(student_survey[, c(1:3)])
```

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

```r
# Solution: Covariance indicates the direction of linear relationship between the
# variables. Positive covariance indicates that as one variable deviates from mean, the
# other variable also deviates in the same direction and the Negative Covariance
# indicates that it occurs in the opposite direction.The Covariance values are not
# standardized and the values can range from Infinity to - Infinity.  The results of
# covariance function indicates, positive relation between TimeTV and Happiness. Whereas
# the results indicates negative relation between TimeTV vs TimeReading and TimeReading
# vs Happiness. In simple words, students were happier watching TV than Reading.Also Time
# spent watching TV and Reading had an inverse relationship.
```

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.

```r
# Solution: Examining the variables, it appears the Time spent Reading and Time spent
# watching TV does not seem to be measured in the same unit. TimeReading appears to be in
# Hours, while TimeTV seems like Minutes. This can have a big effect in the results of
# covariance as it is not standardized. As an experiment, I tried to convert the minutes
# into Hours as shown below :
student_survey1 <- student_survey
student_survey1$TimeTV <- student_survey1$TimeTV/60

# The results of covariance are significant as shown in the results between Covariance of
# TimeTV vs Happiness.
cov(student_survey$TimeTV, student_survey$Happiness)
```

```
## [1] 114.3773
```

```
cov(student_survey1$TimeTV, student_survey1$Happiness)
```

```
## [1] 1.906288
```

```
# To solve this problem, the results can be expressed in Standard deviation units as
# demonstrated below. This is also called as Correlation.
cov_by_sd <- round(cov(student_survey$TimeTV, student_survey$Happiness)/(sd(student_survey$TimeTV) *
    sd(student_survey$Happiness)), 2)
cov_by_sd1 <- round(cov(student_survey1$TimeTV, student_survey1$Happiness)/(sd(student_survey1$TimeTV) *
    sd(student_survey1$Happiness)), 2)
cov_by_sd
```

```
## [1] 0.64
```

```
cov_by_sd1
```

```
## [1] 0.64
```

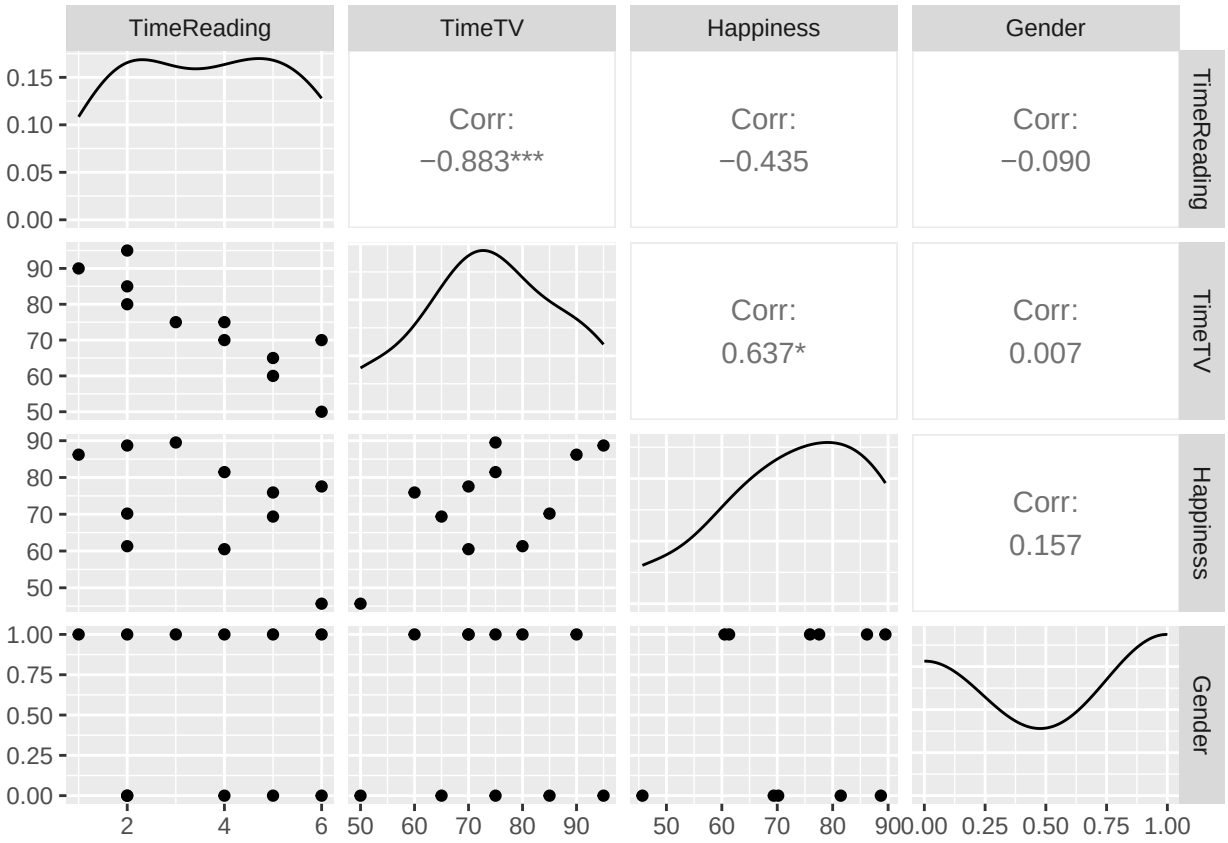
```
# Comparing the results after standardizing the values
identical(cov_by_sd, cov_by_sd1)
```

```
## [1] TRUE
```

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?

```
# Solution: The first steps to identify the correlation method of testing is to Plot the
# data (using ggally) to study the relationship. The results shows that there is a
# negative linear relationship between Time reading and Time TV, so I would pick Pearson
# for that.For the Gender variable correlations, I would use Pearson. However, as the
# dataset is small Kendall can also be a choice to perform correlation on the entire
# dataset.Also, the reason I wouldn't pick Spearman is that the data is not Ordinal.
# Hence picking Pearson method would be much ideal for the dataset provided.  Prediction:
# Based on the results of the plot, I would predict a negative correlation between
# TimeReading vs TimeTV.Also I predict a positive correlation between TimeTV vs Happiness
# based on the results of the plot.


GGally::ggpairs(student_survey)
```



iv.a. Perform a correlation analysis of All variables

```
# Solution: All 3 methods of correlation tests are used to show the results. However as
# stated in the question above, Pearson would be my choice to perform the test.
cr1_pear <- cor(student_survey, use = "complete.obs", method = "pearson")
cr1_pear
```

```
##             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 Correlation using Kendall method
cr2_kend <- cor(student_survey, use = "complete.obs", method = "kendall")
cr2_kend
```

```
##             TimeReading      TimeTV   Happiness      Gender
## TimeReading  1.00000000 -0.80454045 -0.28894280 -0.07824608
## TimeTV      -0.80454045  1.00000000  0.46304237 -0.02507849
## Happiness   -0.28894280  0.46304237  1.00000000  0.09847319
## Gender      -0.07824608 -0.02507849  0.09847319  1.00000000
```

```r
# Calculating Correlation using Spearman method
cr3_spear <- cor(student_survey, use = "complete.obs", method = "spearman")
cr3_spear
```

```
##             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
```

```r
# The Dataset does not say what the symbol in Gender stands for. So I assumed 1 is for
# Male and 0 is for Female and derived below dataframes

student_survey_male <- student_survey %>%
    filter(Gender == 1) %>%
    select(1, 2, 3)
student_survey_female <- student_survey %>%
    filter(Gender == 0) %>%
    select(1, 2, 3)
student_survey_female
```

```
##   TimeReading TimeTV Happiness
## 1           2     95     88.70
## 2           2     85     70.17
## 3           4     75     81.46
## 4           5     65     69.37
## 5           6     50     45.67
```

```r
student_survey_male
```

```
##   TimeReading TimeTV Happiness
## 1           1     90     86.20
## 2           2     80     61.31
## 3           3     75     89.52
## 4           4     70     60.50
## 5           5     60     75.92
## 6           6     70     77.56
```

```r
# Another effective way of calculating correlation between all the variables. Results of
# correlation are arranged in order.
rstatix::cor_test(student_survey, method = "pearson") %>%
    select(var1, var2, cor) %>%
    arrange(cor)
```

```
## # A tibble: 16 x 3
##    var1        var2            cor
##    <chr>       <chr>         <dbl>
##  1 TimeReading TimeTV      -0.88
##  2 TimeTV      TimeReading -0.88
##  3 TimeReading Happiness   -0.43
##  4 Happiness   TimeReading -0.43
```

```
##  5 TimeReading Gender      -0.09
##  6 Gender      TimeReading -0.09
##  7 TimeTV      Gender       0.0066
##  8 Gender      TimeTV       0.0066
##  9 Happiness   Gender       0.16
## 10 Gender      Happiness    0.16
## 11 TimeTV      Happiness    0.64
## 12 Happiness   TimeTV       0.64
## 13 TimeReading TimeReading  1
## 14 TimeTV      TimeTV       1
## 15 Happiness   Happiness    1
## 16 Gender      Gender       1
```

```
# Correlation between all Variables for Male
rstatix::cor_test(student_survey_male, method = "pearson") %>%
    select(var1, var2, cor) %>%
    arrange(cor)
```

```
## # A tibble: 9 x 3
##   var1        var2          cor
##   <chr>       <chr>       <dbl>
## 1 TimeReading TimeTV      -0.86
## 2 TimeTV      TimeReading -0.86
## 3 TimeReading Happiness   -0.12
## 4 Happiness   TimeReading -0.12
## 5 TimeTV      Happiness    0.24
## 6 Happiness   TimeTV       0.24
## 7 TimeReading TimeReading  1
## 8 TimeTV      TimeTV       1
## 9 Happiness   Happiness    1
```

```
# Correlation between all Variables for Female
rstatix::cor_test(student_survey_female, method = "pearson") %>%
    select(var1, var2, cor) %>%
    arrange(cor)
```

```
## # A tibble: 9 x 3
##   var1        var2          cor
##   <chr>       <chr>       <dbl>
## 1 TimeReading TimeTV      -0.97
## 2 TimeTV      TimeReading -0.97
## 3 TimeReading Happiness   -0.74
## 4 Happiness   TimeReading -0.74
## 5 TimeTV      Happiness    0.87
## 6 Happiness   TimeTV       0.87
## 7 TimeReading TimeReading  1
## 8 TimeTV      TimeTV       1
## 9 Happiness   Happiness    1
```

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

```r
cor.test(student_survey$TimeTV, student_survey$Happiness, method = "pearson", conf.level = 0.95)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_survey$TimeTV and student_survey$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
```

```r
# The results indicates a p value of 0.035 which is less than 0.05 and hence we can
# reject the null hypothesis and consider the Alternative hypotheses that the Correlation
# does exist between the Time spent watching TV and Happiness.Also the ranges in the
# Confidence interval 0.05934031 & 0.89476238 is above zero which indicates that there is
# perhaps a correlation between the variables.

# The below examples show the correlation between two variables on the entire dataset and
# Male and Female datasets. The results indicate Female students did not enjoy Reading
# more than the Male students.
cor(student_survey$Happiness, student_survey$TimeReading)
```

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

```r
cor(student_survey_male$Happiness, student_survey_male$TimeReading)
```

```
## [1] -0.1246403
```

```r
cor(student_survey_female$Happiness, student_survey_female$TimeReading)
```

```
## [1] -0.735294
```

```r
# The below results indicate students in general were happy watching TV. Female students
# enjoyed watching TV more than the male students.
cor(student_survey$Happiness, student_survey$TimeTV)
```

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

```r
cor(student_survey_male$Happiness, student_survey_male$TimeTV)
```

```
## [1] 0.2354574
```

```r
cor(student_survey_female$Happiness, student_survey_female$TimeTV)
```

```
## [1] 0.8723756
```

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

```
cor.test(student_survey$TimeTV, student_survey$Happiness, method = "pearson", conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_survey$TimeTV and student_survey$Happiness
## t = 2.4761, df = 9, p-value = 0.03521
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.1570212  0.9306275
## sample estimates:
##      cor
## 0.636556
```

```
# The results indicates a p value of 0.0352 which is less than 0.05 and hence we can
# reject the null hypothesis and consider the Alternative hypotheses that the Correlation
# does exist between the Time spent watching TV and Happiness.The actual correlation
# value is 0.636556

cor.test(student_survey$Happiness, student_survey$TimeReading, method = "pearson", conf.level = 0.99)
```

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

```
cor.test(student_survey_male$Happiness, student_survey_male$TimeReading, method = "pearson",
    conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_survey_male$Happiness and student_survey_male$TimeReading
## t = -0.25124, df = 4, p-value = 0.814
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.9235209  0.8768247
## sample estimates:
##       cor
## -0.1246403
```

```r
cor.test(student_survey_female$Happiness, student_survey_female$TimeReading, method = "pearson",
    conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_survey_female$Happiness and student_survey_female$TimeReading
## t = -1.8791, df = 3, p-value = 0.1568
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.9920448  0.7070350
## sample estimates:
##       cor
## -0.735294
```

```r
cor.test(student_survey$Happiness, student_survey$TimeTV, method = "pearson", conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_survey$Happiness and student_survey$TimeTV
## t = 2.4761, df = 9, p-value = 0.03521
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.1570212  0.9306275
## sample estimates:
##      cor
## 0.636556
```

```r
cor.test(student_survey_male$Happiness, student_survey_male$TimeTV, method = "pearson", conf.level = 0.9
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_survey_male$Happiness and student_survey_male$TimeTV
## t = 0.48454, df = 4, p-value = 0.6533
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.8474956  0.9387141
## sample estimates:
##       cor
## 0.2354574
```

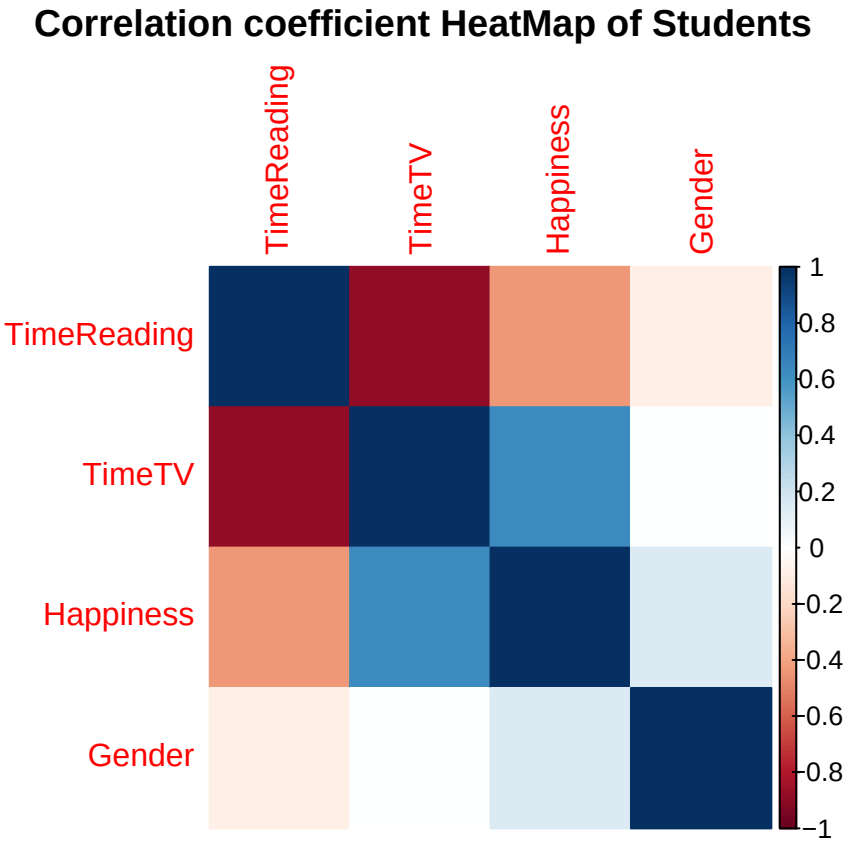
```r
cor.test(student_survey_female$Happiness, student_survey_female$TimeTV, method = "pearson",
    conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_survey_female$Happiness and student_survey_female$TimeTV
## t = 3.091, df = 3, p-value = 0.05367
```

```
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.4450017  0.9964375
## sample estimates:
##       cor
## 0.8723756
```

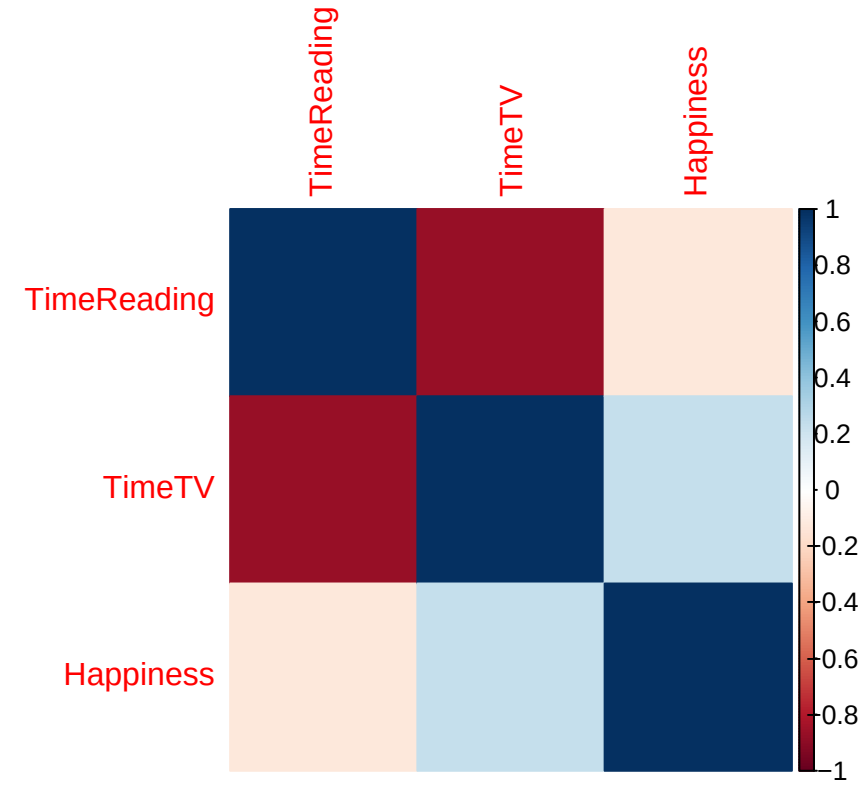
iv. d) Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.

```
corrplot(cr1_pear, method = "color", title = "Correlation coefficient HeatMap of Students",
    mar = c(0, 0, 2, 0))
```

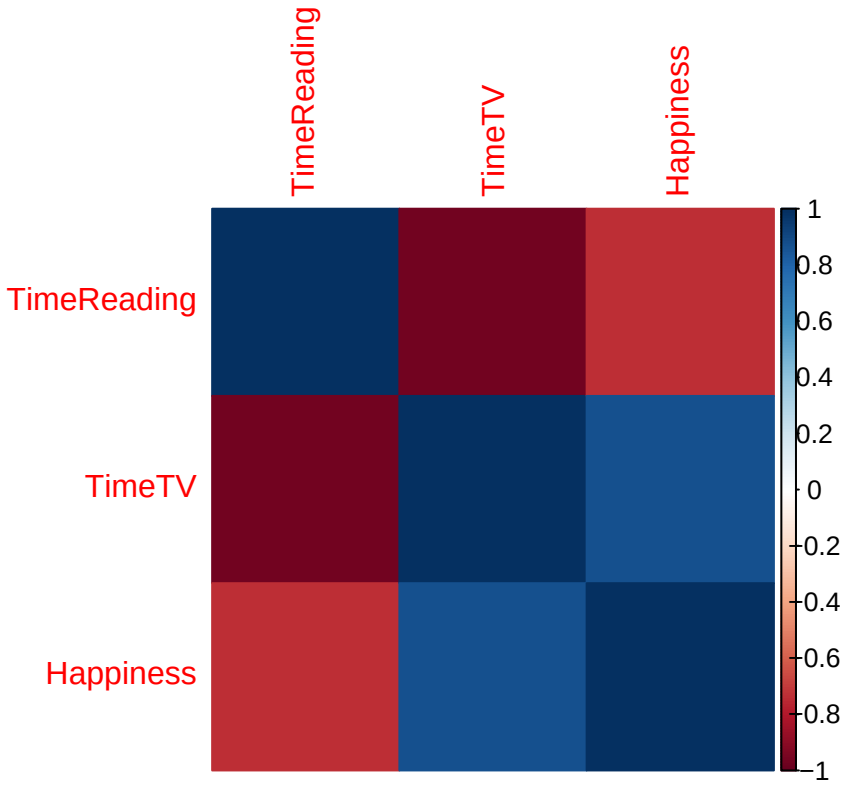



```
# The Correlation Plot suggests the below results: There is a strong Negative Correlation
# between Time spent watching TV versus time spent reading.  There is Negative
# correlation(Not strong though, since the color is not Dark Red) between Time spent
# reading and Happiness.  There is a positive correlation between Time spent watching TV
# and Happiness(indicated by Blue color in the plot) There is a slight positive
# correlation between the Gender and Happiness(indicated by light blue).  There is a
# slight negative correlation between the Gender and Time reading(indicated by light Red)
# There is no correlation between Gender and Time spent watching TV(Indicated by white
# color)
```

```r
cr_pear_male <- cor(student_survey_male, use = "complete.obs", method = "pearson")
corrplot(cr_pear_male, method = "color", title = "Correlation coefficient HeatMap of Male Students",
    mar = c(0, 0, 2, 0))
```




```r
# The below plot indicates that Female students in general did not enjoy Reading. They
# enjoyed reading lesser than Male and also enjoyed watching TV more than Male students.
cr_pear_female <- cor(student_survey_female, use = "complete.obs", method = "pearson")
corrplot(cr_pear_female, method = "color", title = "Correlation coefficient HeatMap of Female Students"
    mar = c(0, 0, 2, 0))
```



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

```
# Correlation coefficient
corr_coeff <- cor(student_survey, method = "pearson")
corr_coeff
```

```
##             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
```

```
# Coefficient of Determination - R^2
coeff_of_det <- corr_coeff^2
coeff_of_det
```

```
##             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
```

```r
# Coefficient of Determination - Shown in Percentage
coeff_of_det * 100
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading 100.0000000  77.98085292  18.910873   0.80357143
## TimeTV       77.9808529 100.00000000  40.520352   0.00435161
## Happiness    18.9108726  40.52035234 100.000000   2.46527174
## Gender        0.8035714   0.00435161   2.465272 100.00000000
```

```r
# Solution: The Correlation coefficient indicates that: Time spent watching TV accounts
# for 77% variability in the Time spent Reading Happiness accounts for 18% of the
# variability in the Time spent reading and 40% in the Time spent Watching TV and only 2%
# of Gender choices.  Gender has the least share in the variability of any other other
# variables in the dataset.
```

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

```r
cor.test(student_survey$TimeTV, student_survey$TimeReading, method = "pearson")
```

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

```r
# Solution: Yes watching more TV caused students to read less because there is a strong
# negative correlation between the 2 variables of -0.883 and also Time spent watching TV
# accounts for 77% variability in the Time spent Reading. These results are evident from
# the Correlation plot above which indicates a negative correlation.

cor(student_survey$TimeTV, student_survey$TimeReading, method = "pearson")
```

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

```r
cor(student_survey_male$TimeTV, student_survey_male$TimeReading, method = "pearson")
```

```
## [1] -0.8641428
```

```r
cor(student_survey_female$TimeTV, student_survey_female$TimeReading, method = "pearson")
```

```
## [1] -0.9682789
```

```
# The above results indicate that the Female students(assuming Gender=0 is female) did
# not enjoy studying compared to watching TV than the Male students.
```

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.

```
# TimeTV vs Happiness
cor(student_survey$TimeTV, student_survey$Happiness, method = "pearson")
```

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

```
# TimeTV vs Happiness controlling TimeReading
pcor(c("TimeTV", "Happiness", "TimeReading"), var(student_survey))
```

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

```
# Explanation: The correlation coefficient between Time spent watching TV and Happiness
# is 0.6365. However while controlling the time spent Reading, the correlation drops to
# 0.597

# TimeTV vs Time Reading
cor(student_survey$TimeTV, student_survey$TimeReading, method = "pearson")
```

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

```
# TimeTV vs Time Reading controlling Happiness
pcor(c("TimeTV", "TimeReading", "Happiness"), var(student_survey))
```

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

```
# Explanation: The Correlation coefficient between Time spent watching TV and Time spent
# reading is -0.883. However controlling the factor of Happiness, the correlation
# increases slightly to -0.87

# Gender vs Happiness
cor(student_survey$Gender, student_survey$Happiness, method = "pearson")
```

```
## [1] 0.1570118
```

```
# Gender vs Happiness controlling Time spent on TV and Reading
pcor(c("Gender", "Happiness", "TimeTV", "TimeReading"), var(student_survey))
```

```
## [1] 0.2833152
```

```
# Explanation: The correlation coefficient between Happiness and the Gender is 0.157.
# However controlling the Time spent reading and time spent watching TV, the correlation
# increases to 0.283
```

Session Info

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 22621)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggm_2.5         corrplot_0.92  dplyr_1.0.10   magrittr_2.0.3 rstatix_0.7.1
## [6] GGally_2.1.2    ggplot2_3.4.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.9          highr_0.9           plyr_1.8.8          pillar_1.8.1
##  [5] compiler_4.2.2      formatR_1.12        RColorBrewer_1.1-3  tools_4.2.2
##  [9] digest_0.6.30       evaluate_0.18       lifecycle_1.0.3     tibble_3.1.8
## [13] gtable_0.3.1        pkgconfig_2.0.3     rlang_1.0.6         igraph_1.3.5
## [17] cli_3.4.1           DBI_1.1.3           rstudioapi_0.14     yaml_2.3.6
## [21] xfun_0.34           fastmap_1.1.0       withr_2.5.0         stringr_1.4.1
## [25] knitr_1.41          generics_0.1.3      vctrs_0.5.0         grid_4.2.2
## [29] tidyselect_1.2.0    reshape_0.8.9       glue_1.6.2          R6_2.5.1
## [33] fansi_1.0.3         rmarkdown_2.18      carData_3.0-5       farver_2.1.1
## [37] car_3.1-1           tidyr_1.2.1         purrr_0.3.5         backports_1.4.1
## [41] scales_1.2.1        htmltools_0.5.3     abind_1.4-5         assertthat_0.2.1
## [45] colorspace_2.0-3    labeling_0.4.2      utf8_1.2.2          stringi_1.7.8
## [49] munsell_0.5.0       broom_1.0.1
```