assignment_07_MunjewarSheetal-01

Sheetal M

2023-01-29

Install and Load required packages :

```
##
## 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

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v purrr   1.0.0
## v tidyr   1.2.1     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
##
##
## Attaching package: 'scales'
##
##
## The following object is masked from 'package:purrr':
##
##     discard
##
##
## The following object is masked from 'package:readr':
##
##     col_factor
##
##
##
```

```
## Attaching package: 'reshape'
##
##
## The following objects are masked from 'package:tidyr':
##
##     expand, smiths
##
##
## The following object is masked from 'package:dplyr':
##
##     rename
```

Set the working directory to the root of your DSC 520 directory setwd("E:\Data_Science_DSC510\DSC520-Statistics\dsc520")

Load the `data/student-survey.csv` to ssurvey_df <- read.csv("data/student-survey.csv")

Using `cor()` compute correlation coefficients for

```
setwd("E:\\Data_Science_DSC510\\DSC520-Statistics\\dsc520")
ssurvey_df <- read.csv("data/student-survey.csv")

ssurvey_df
```

```
##    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
## 7            4     75     81.46      0
## 8            5     60     75.92      1
## 9            5     65     69.37      0
## 10           6     50     45.67      0
## 11           6     70     77.56      1
```

```
ssurvey_df[,c(2,2:4)]
```

```
##    TimeTV TimeTV.1 Happiness Gender
## 1      90       90     86.20      1
## 2      95       95     88.70      0
## 3      85       85     70.17      0
## 4      80       80     61.31      1
## 5      75       75     89.52      1
## 6      70       70     60.50      1
## 7      75       75     81.46      0
## 8      60       60     75.92      1
## 9      65       65     69.37      0
## 10     50       50     45.67      0
## 11     70       70     77.56      1
```

```
#--------------------------------------#
#       **** Assignment-I *****        #
#--------------------------------------#
# Assignment-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

#-- Explanation :
#   Cor/Cov/Var function will compute variance of x or covariance or corelation
#   of x and y. Applying cor() function on survey variables, will produce
#   corelations matrix values betweem 1 amd -1, higeher positive number means
#   closer relationshi between the variables, and negative number mean inverse.
#   Give out for survey results indicate +ve corelation between TimeTV vs
#   Happiness (0.63) and -ve corelation between TimeTV and TimeReading (-0.88).
#   Results can be visualize using GGally::ggpairs.

library(GGally)
cor(ssurvey_df)
```

```
##             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
```

```
cov(ssurvey_df)
```

```
##              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
```

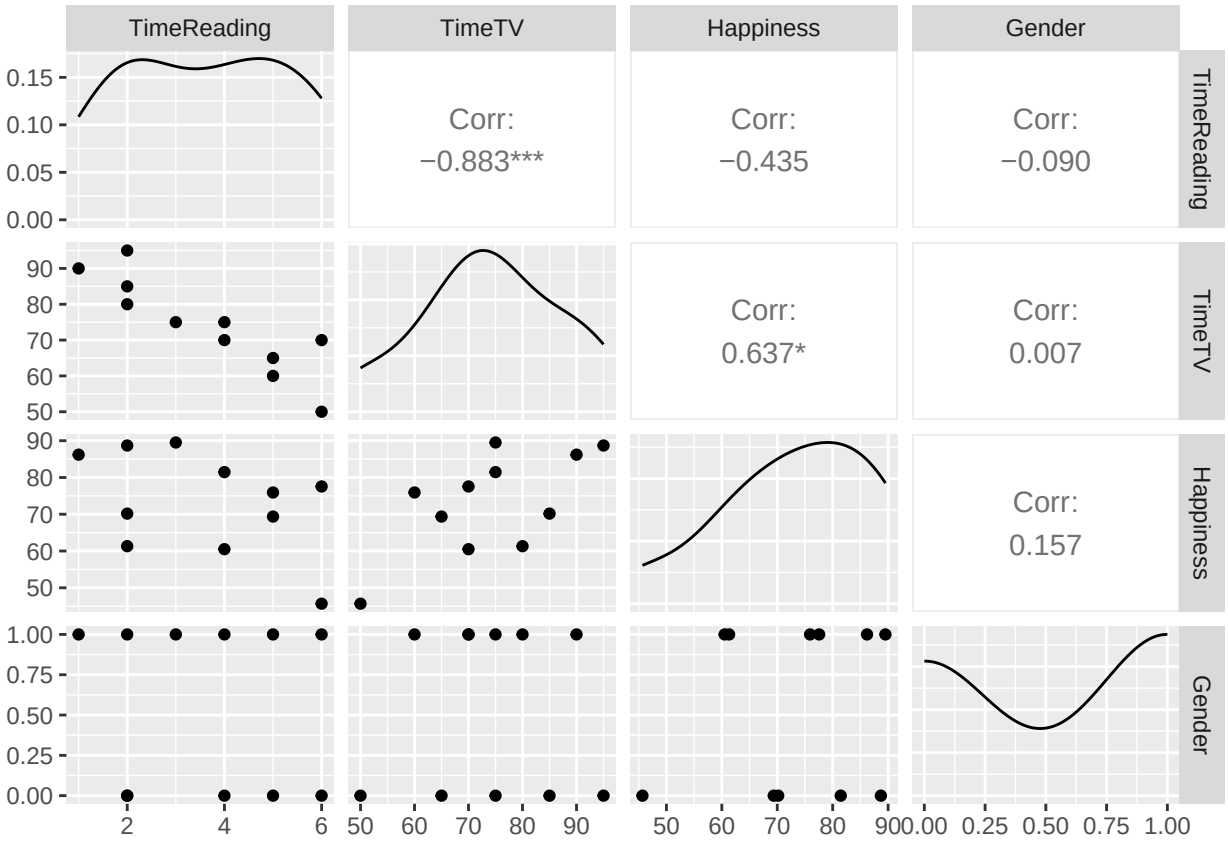
```
var(ssurvey_df)
```

```
##              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
```

```
cor(ssurvey_df, method = c("pearson", "kendall", "spearman"))
```

```
##             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
```

```
GGally::ggpairs(ssurvey_df)
```




```
#help -- ?cor()

#---------------------------------------#
#       **** Assignment-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.


#-- Explanation :

cor(ssurvey_df)
```

```
##             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
```

```
cov(ssurvey_df)
```

```
##              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
```

```
#cov(ssurvey_df)
#             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 diagonal elements 3,174,185 and 0.2 indicate the variance in data sets
#(lowest variance: 0.27 and Highest variance:185.451422), variance positive 174
#co-variance between TimeTV and Happiness indicates, happiness increases and
#TVtime goes up, however negative -20 variance indicates oppsite with TimeTv
#and TimeReading variance. positive 0.04 variance has minimal impact with Gender
#and TimeTV. Changing measures in Covariance unit will change the result/outcome.

# Problem is covariance -
# The main problem with covariance interpretation is that the wide range of
# results,it hard to interpret sometime. ( 0.2 to 185 in survey data frame.)

# Alternative : Correlation Coefficient method do have several advantages over
#               covariance for determining    strengths of relationships:
# Covariance can take on practically any number while a correlation is limited: -1 to +1.
# Because of it's numerical limitations, correlation is more useful for -
# Correlation does not have units. Covariance always has units
# Correlation isn't affected by changes in the center (i.e. mean) or scale of the variables

#-- References :
#-- https://www.cuemath.com/algebra/covariance-matrix/
#-- https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/covariance/
#-- https://www.mygreatlearning.com/blog/covariance-vs-correlation/#variance

# Variance - Variance is the expectation of the squared deviation of a random
# variable from its mean
# Standard Deviation
# Standard deviation is a measure of the amount of variation or dispersion of a
# set of values. A low standard deviation indicates that the values tend to be
# close to the mean of the set, while a high standard deviation indicates that the
# values are spread out over a wider range. It essentially measures the absolute
# variability of   a random variable.
# Covariance and correlation are related to each other, in the sense that
# covariance determines the type of interaction between two variables, while
# correlation determines the direction as well as the strength of the
# relationship between two variables.

# To find coorelation,columns/df variables needs to be integer.
# str(ssurvey_df)
# summary(ssurvey_df)
# cor(ssurvey_df, use = "complete.obs")
# cov(ssurvey_df)
# cov(ssurvey_df$TimeTV,ssurvey_df$TimeReading)
```

```r
# cov(ssurvey_df$TimeTV,ssurvey_df$Happiness)
# cov(ssurvey_df$TimeReading,ssurvey_df$Happiness)

#---------------------------------------#
#       **** Assignment-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?

#-- Explanation :

#  Considering student survey dataset with no missing and NULL values and
#  skewness ratio, and positive and negative relationship between with variables
#  TimeTV/Happiness and TimeTV/TimeReading, I prefer to go with "Pearson" method.

#install.packages("moments")
library(moments)
skewness(ssurvey_df)
```

```
##  TimeReading       TimeTV    Happiness       Gender
## -0.002922561 -0.136695818 -0.595566474 -0.182574186
```

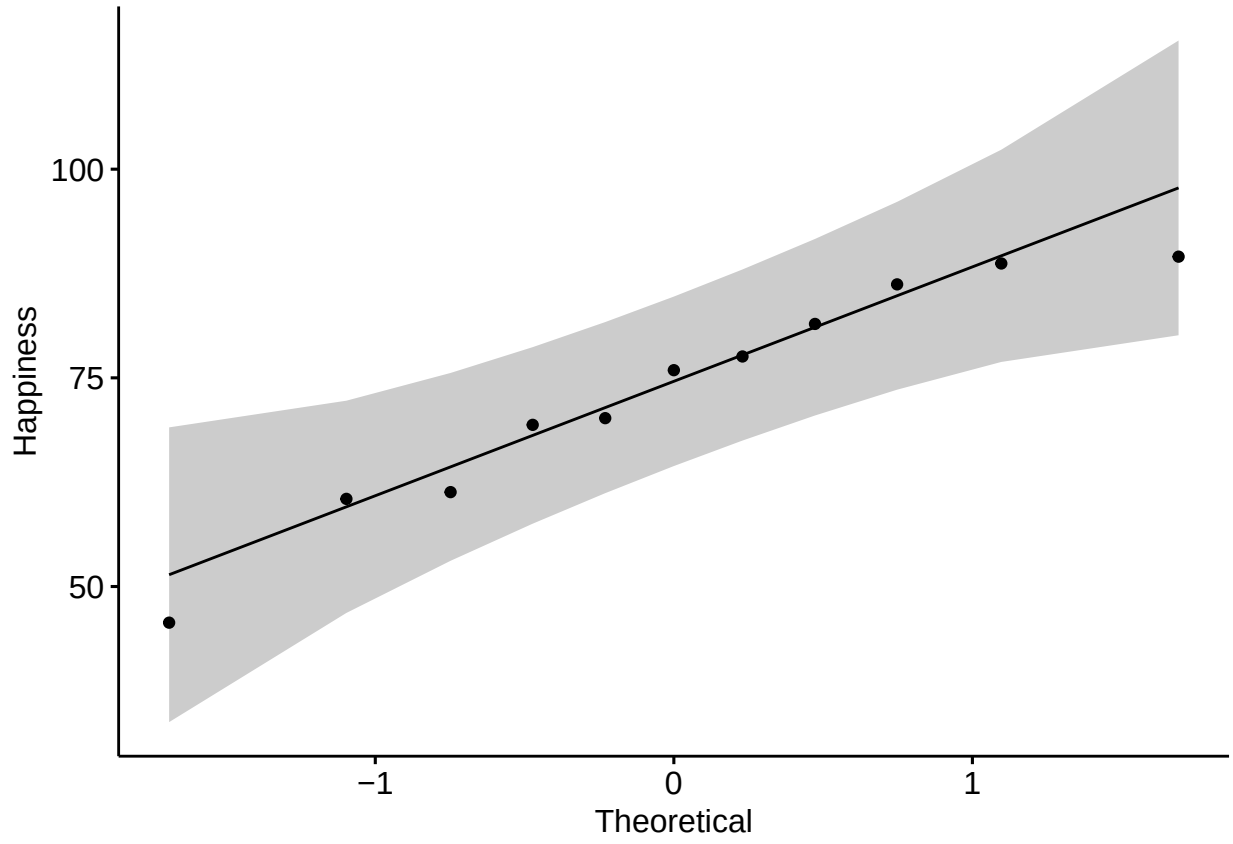
```r
cor(ssurvey_df, use = "complete.obs", method = c("pearson"))
```

```
##             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
```

```r
# cor(ssurvey_df, use = "complete.obs", method = c("pearson", "kendall", "spearman"))
#cor(heights_df$ed,heights_df$earn, method = 'kendall')
#cor(heights_df$ed,heights_df$earn, method = 'pearson')

#-  Visual inspection of the data normality using Q-Q plots (quantile-quantile
#   plots). Q-Q plot draws the correlation between a given sample and the normal
#   distribution.
#   http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r

#installed_packages("ggpubr")
library("ggpubr")
# Happiness
ggqqplot(ssurvey_df$Happiness, ylab = "Happiness")
```

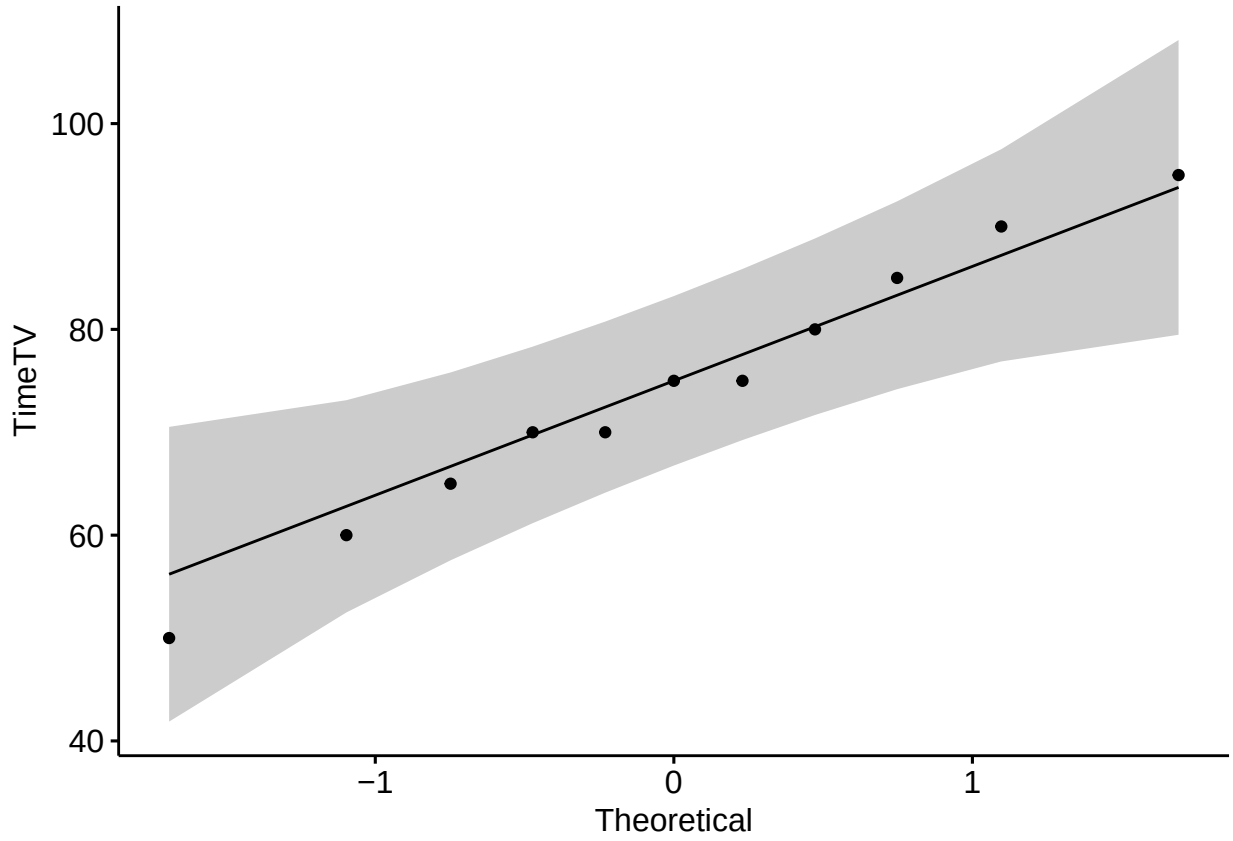
```
## Warning: The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
```

```
##   the data.
## i Did you forget to specify a 'group' aesthetic or to convert a numerical
##   variable into a factor?
```




```
# TimeTV
ggqqplot(ssurvey_df$TimeTV, ylab = "TimeTV")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a 'group' aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: sample
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a 'group' aesthetic or to convert a numerical
##   variable into a factor?
```




```
# References -
#   https://ademos.people.uic.edu/Chapter22.html

# The Pearson product-moment correlation is one of the most commonly used
# correlations in statistics. It's a measure of the strength and the direction
# of a linear relationship between two variables.
  # Your data is interval or ratio
  # Pearson only works with linear data. That means that your two correlated
    # factors have to approximate a line,and     not a curved or parabolic shape
  # Outliers in your data can really throw off a Pearson correlation

# Skewness interpretation :
#  As a general rule of thumb: If skewness is less than -1 or greater than 1,
# the distribution is highly skewed.If skewness is between -1 and -0.5 or
# between 0.5 and 1, the distribution is moderately skewed. If skewness is
# between -0.5 and 0.5, the distribution is approximately symmetric.
# he data you are analyzing needs to be normally distributed. This can be done
# in a couple of ways (Skewness, Kurtosis) but it can also be done in a
# quick and dirty manner through histograms

#---------------------------------------#
#       **** Assignment-IV *****        #
#---------------------------------------#
# Assignment-IV : 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.

#-- Explanation :

# Cor() function define corelation between all the variables with values between -1 to 1.
cor(ssurvey_df)
```

```
##             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
```

```
# - Corelation between thevariables ssurvey_df$TimeTV and ssurvey_df$Happiness,
#   using default method pearson.
cor.test(ssurvey_df$TimeTV,ssurvey_df$Happiness,method="pearson")
```

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

```
# Corelation with confidence level 0.99
cor.test(ssurvey_df$TimeTV,ssurvey_df$Happiness,method="pearson",conf.level = 0.99 )
```

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

```
#- Matix values interpretation :
  # - Values 0.63 represent positive relationship between variables
  #   TimeTV and Happiness.
  # - Values -0.88 represent negative relationship between variables
  #   TimeTV and TimeReading.
```

```r
#----------------------------------------#
#       **** Assignment-V  *****         #
#----------------------------------------#

# Assignment-V : Calculate the correlation coefficient and the coefficient of
# determination, describe what you conclude about the results.

#-- Explanation :

# Objective is to find the co-relation between predictor variables TimeTV and
# TimeReading, Positive corelation coefficients (0.63) positive colinear
# relationships between them, however coefficient of detemination prediction
# (0.47), means a 47% variation in the Happiness can be explained by the time
# spend on watching TV and reading time.

# correlation coefficient
cor.test(ssurvey_df$TimeTV,ssurvey_df$Happiness,method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  ssurvey_df$TimeTV and ssurvey_df$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
# coefficient of determination
ss_model <- lm(ssurvey_df$Happiness ~ ssurvey_df$TimeTV + ssurvey_df$TimeReading, data=ssurvey_df)

#view model summary
summary(ss_model)
```

```
##
## Call:
## lm(formula = ssurvey_df$Happiness ~ ssurvey_df$TimeTV + ssurvey_df$TimeReading,
##     data = ssurvey_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.623  -9.142  -1.549   5.686  18.002
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -30.7722    55.7788  -0.552    0.596
## ssurvey_df$TimeTV         1.1837     0.5614   2.108    0.068 .
## ssurvey_df$TimeReading    4.5032     4.2386   1.062    0.319
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 10.99 on 8 degrees of freedom
## Multiple R-squared:  0.4787, Adjusted R-squared:  0.3484
## F-statistic: 3.674 on 2 and 8 DF,  p-value: 0.07382
```

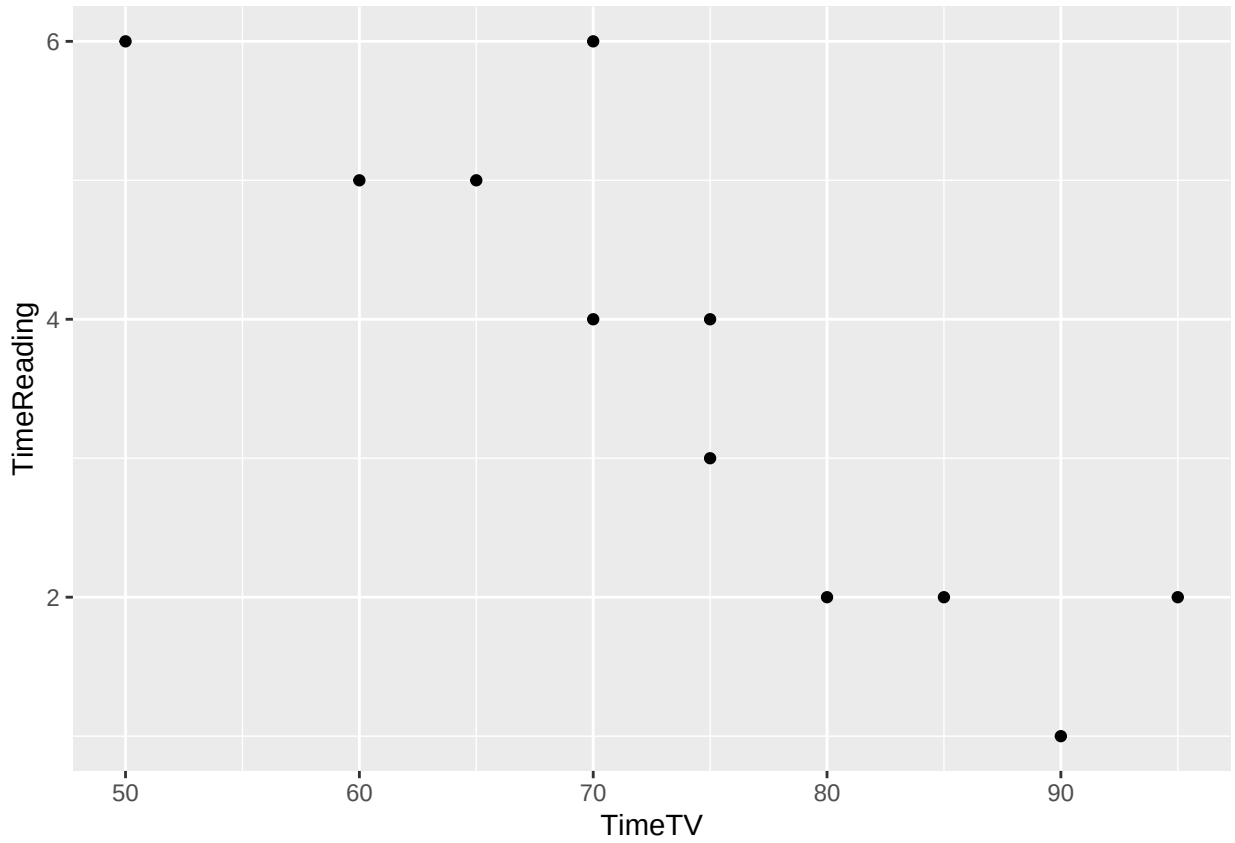
```
summary(ss_model)$r.squared
```

```
## [1] 0.4787487
```

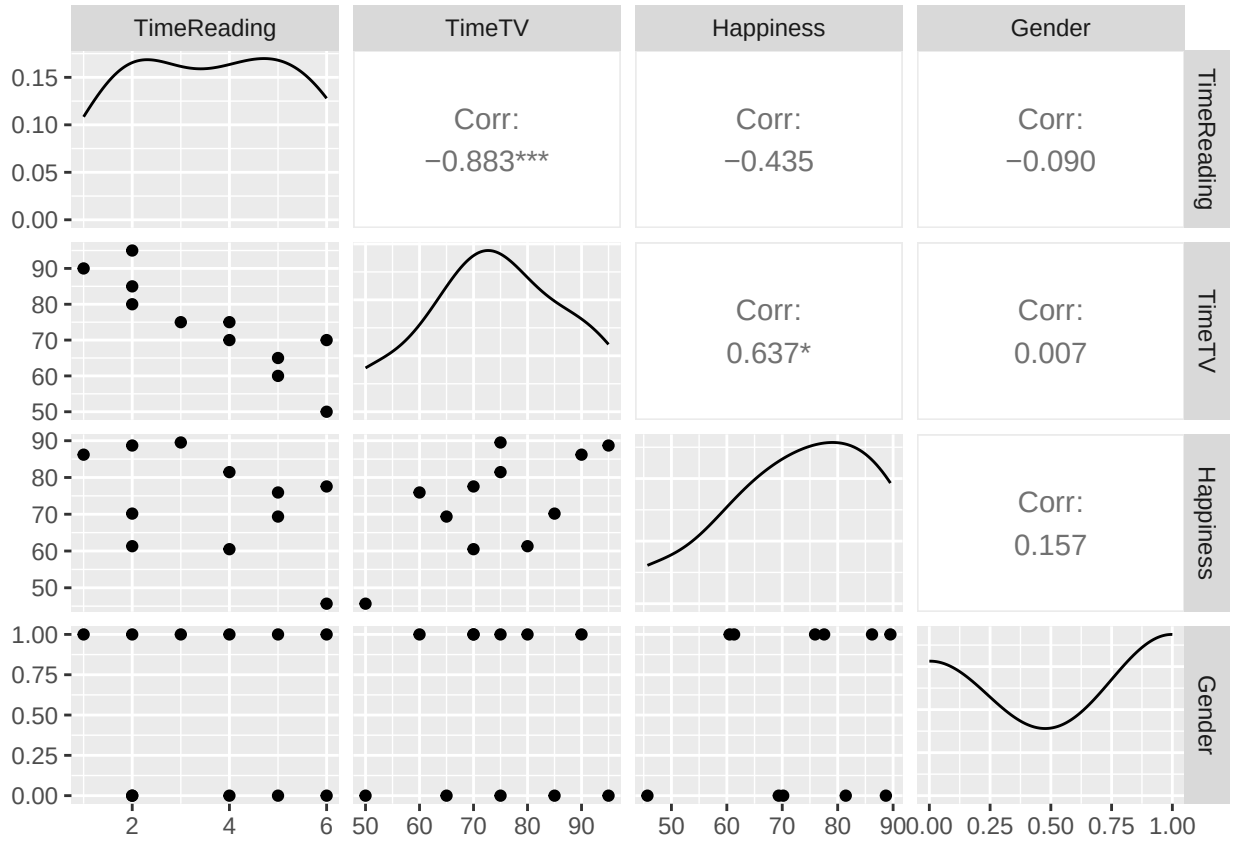
```
# References -
# https://www.statology.org/good-r-squared-value/

#---------------------------------------#
#       **** Assignment-VI  *****       #
#---------------------------------------#

# Assignment-VI : Based on your analysis can you say that watching more TV caused
# students to read less? Explain.

#-- Explanation :
# - Negative co-relation between the variables TimeTV and TimeReading,
#   and correlation coefficient (-0.88)
# - indicates student who spend more time watching TV will spend less
#   hours on reading and vise versa.

#?ggplot()
ggplot(ssurvey_df, aes(x=ssurvey_df$TimeTV, y=ssurvey_df$TimeReading)) +
  geom_point() +
  xlab("TimeTV") +
  ylab("TimeReading")
```

```
## Warning: Use of ‘ssurvey_df$TimeTV‘ is discouraged.
## i Use ‘TimeTV‘ instead.
```

```
## Warning: Use of ‘ssurvey_df$TimeReading‘ is discouraged.
## i Use ‘TimeReading‘ instead.
```




```
GGally::ggpairs(ssurvey_df)
```




```
#---------------------------------------#
#       **** Assignment-VII  *****      #
#---------------------------------------#

# Assignment-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.

#-- Explanation :

#  With vector V1, Partial correlation value between variables TimeTV and Happiness
#  is 0.63, which signifies that both variables highly consistent and they increase
#  with each other.

#  With vector V2, partial correlation value between variables TimeTV and
#  Happiness changed, TimeTV and Happiness vector is still the same because the
#  vector TimeReading affecting them. So now the correlation value dropped to
#  0.63 to 0.59 because TimeTV and TimeReading are inconsist with the
#  value of -0.8729450.


# install.packages("ppcor")
# install.packages("dplyr")
library(ppcor)
```

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

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

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

```
library(dplyr)
library(purrr)

V1 <- ssurvey_df %>% dplyr::select(TimeTV,Happiness)
V2 <- ssurvey_df %>% dplyr::select(TimeTV,Happiness,TimeReading)
ppcor::pcor(V1)
```

```
## $estimate
##             TimeTV Happiness
## TimeTV    1.000000  0.636556
## Happiness 0.636556  1.000000
##
## $p.value
##               TimeTV  Happiness
## TimeTV    0.00000000 0.03521425
## Happiness 0.03521425 0.00000000
##
## $statistic
##             TimeTV Happiness
## TimeTV    0.000000  2.476131
## Happiness 2.476131  0.000000
##
## $n
## [1] 11
##
## $gp
## [1] 0
##
## $method
## [1] "pearson"
```

```
ppcor::pcor(V2)
```

```
## $estimate
##                 TimeTV Happiness TimeReading
## TimeTV       1.0000000 0.5976513  -0.8729450
## Happiness    0.5976513 1.0000000   0.3516355
## TimeReading -0.8729450 0.3516355   1.0000000
##
## $p.value
##                   TimeTV  Happiness  TimeReading
## TimeTV      0.0000000000 0.06804372 0.0009753126
## Happiness   0.0680437248 0.00000000 0.3190589526
## TimeReading 0.0009753126 0.31905895 0.0000000000
##
```

```
## $statistic
##                TimeTV Happiness TimeReading
## TimeTV       0.000000  2.108388   -5.061434
## Happiness    2.108388  0.000000    1.062425
## TimeReading -5.061434  1.062425    0.000000
##
## $n
## [1] 11
##
## $gp
## [1] 1
##
## $method
## [1] "pearson"
```

```
#pcor(ssurvey_df)

#- Reference - https://www.statology.org/partial-correlation-r/
#             https://www.geeksforgeeks.org/how-to-calculate-partial-correlation-in-r/
```