ASSIGNMENT 7

Raghuwanshi, Prashant

2021-07-25

Question no 1 : Complete assignment05

```
library(ggplot2)
theme_set(theme_minimal())

## Set the working directory to the root of your DSC 520 directory

setwd("D:/MS_DataScience/DSC 520-Datastatiics.pdf/dsc520")

## Load the `data/r4ds/heights.csv` to

heights_df <- read.csv("data/r4ds/heights.csv")
head(heights_df)
```

```
##    earn   height    sex ed age  race
## 1 50000 74.42444   male 16  45 white
## 2 60000 65.53754 female 16  58 white
## 3 30000 63.62920 female 16  29 white
## 4 50000 63.10856 female 16  91 other
## 5 51000 63.40248 female 17  39 white
## 6  9000 64.39951 female 15  26 white
```

Using `cor()` compute correclation coefficients for

```
## height vs. earn
cor(heights_df$height, heights_df$earn)
```

```
## [1] 0.2418481
```

```
### age vs. earn
cor(heights_df$age, heights_df$earn)
```

```
## [1] 0.08100297
```

```
### ed vs. earn
cor(heights_df$ed, heights_df$earn)
```

```
## [1] 0.3399765
```

Spurious correlation

The following is data on US spending on science, space, and technology in millions of today's dollars

and Suicides by hanging strangulation and suffocation for the years 1999 to 2009

Compute the correlation between these variables

```
tech_spending <- c(18079, 18594, 19753, 20734, 20831, 23029, 23597, 23584, 25525, 27731, 29449)
suicides <- c(5427, 5688, 6198, 6462, 6635, 7336, 7248, 7491, 8161, 8578, 9000)
cor(tech_spending, suicides)
```

```
## [1] 0.9920817
```

Question no 2 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. ##The survey data is located in this StudentSurvey.csv file.

```
setwd("D:/MS_DataScience/DSC 520-Datastatiics.pdf/dsc520")
stdsry_df <- read.csv("data/student-survey.csv")
head(stdsry_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
```

###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????

```
cor(stdsry_df$TimeReading, stdsry_df$TimeTV)
```

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

Ans Negative correlationship shows no relationship between timespent in reading & watching tv

###You are also interested if there are other significant relationships that can be discovered?

```
library(GGally)
```

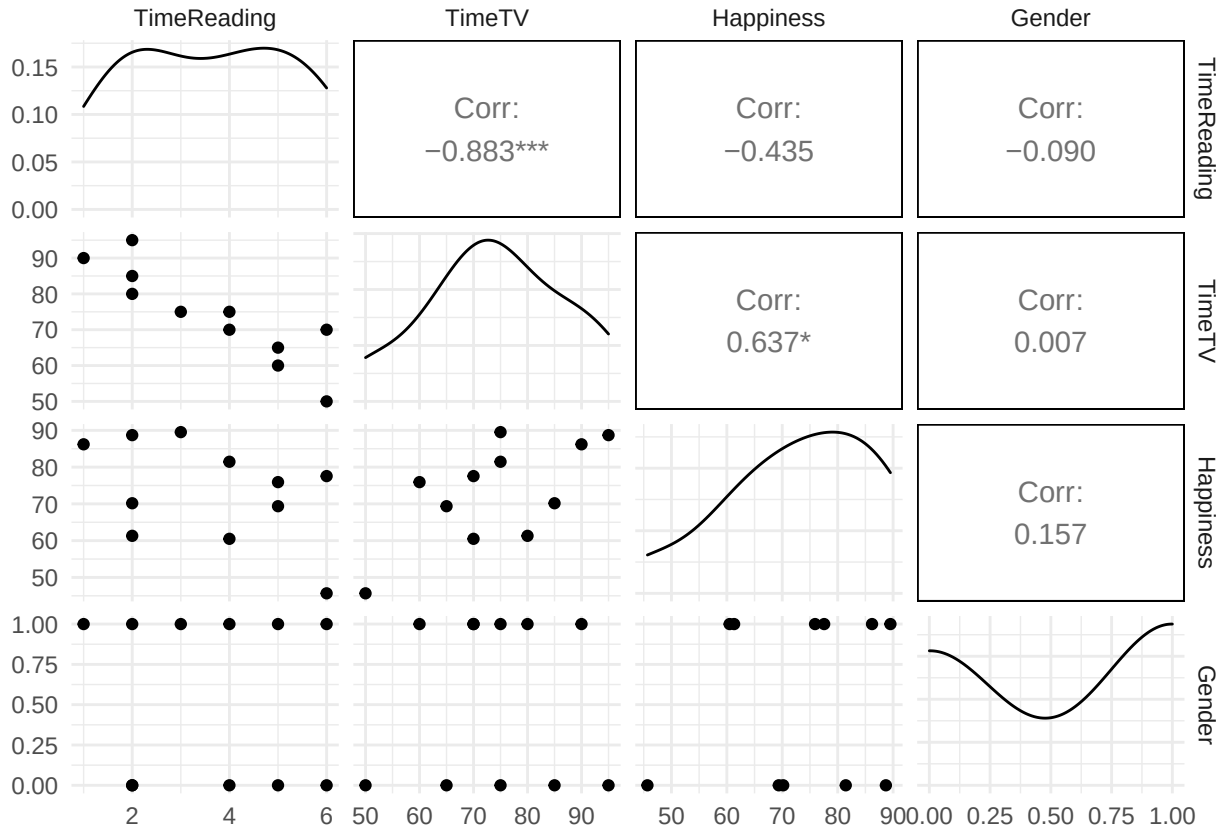
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
stdsry_mat = data.matrix(stdsry_df)
## Ans please find the other relationships between multiple variables
cor(stdsry_mat)
```

```
##             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
```

```
## Ans please find the other relationships between multiple variables in plots
GGally::ggpairs(stdsry_df)
```



####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(stdsry_mat)
```

```
##              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
```

```
## Ans after analyzing the cov data, it seems most of covariance is showing towards negative directions
## in most of the comparison the variance is showing higher negative value , seems this value is due to
## used, however the variance for gender Vs others or time reading & time tv are showing some relevant
```

####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. ##Ans we are having 4 different measures, seems time reading & time tv have same scale, ####Would this be a problem? Explain and provide a better alternative if needed. ##Ans difference in measurement of scale of variable is a problem to determin the covariance between the variables ## to over come this issue usually we are performing standarization process

####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?

```r
##ANS : Correlation test is used to evaluate the association between two or more variables.
## here first i am testing the mean between variables to verify the noremal distributions
##T-TEST  (two sample test T test and paired two sample t-test)
## two sample t-test is can be performed between gender & happiness variable
## paired two sample test test can be perfromed between time reading & timetv
##variance of each group
aggregate(Happiness ~ Gender, data = stdsry_df, var)
```

```
##   Gender Happiness
## 1      0  266.9072
## 2      1  148.2333
```

```r
## normality of happyness distributions
##value of the Shapiro-Wilk Test is greater than 0.05, the data is normal
shapiro.test(stdsry_df$Happiness)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  stdsry_df$Happiness
## W = 0.9412, p-value = 0.5347
```

```r
shapiro.test(stdsry_df$Happiness[stdsry_df$Gender==1])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  stdsry_df$Happiness[stdsry_df$Gender == 1]
## W = 0.89745, p-value = 0.3591
```

```r
shapiro.test(stdsry_df$Happiness[stdsry_df$Gender==0])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  stdsry_df$Happiness[stdsry_df$Gender == 0]
## W = 0.93043, p-value = 0.5993
```

```r
t.test(Happiness ~ Gender, stdsry_df, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Happiness by Gender
## t = -0.47695, df = 9, p-value = 0.6448
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -23.51356  15.32489
## sample estimates:
## mean in group 0 mean in group 1
##        71.07400        75.16833
```

```r
# here iam using Pearson correlation (r),
##which measures a linear dependence between two variables (x and y).
## It???s also known as a parametric correlation test because it depends to the distribution of the data
## t can be used only when x and y are from normal distribution.
## he plot of y = f(x) is named the linear regression curve
```

```
cor.test(stdsry_df$TimeReading, stdsry_df$TimeTV, method=c("pearson"))
```

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

```
## no Relation ship, negative cor
cor.test(stdsry_df$TimeReading, stdsry_df$Gender, method=c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  stdsry_df$TimeReading and stdsry_df$Gender
## 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
```

```
## no Relation ship, negative cor
cor.test(stdsry_df$Happiness, stdsry_df$Gender, method=c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  stdsry_df$Happiness and stdsry_df$Gender
## 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
```

```
## good Relation ship, positive cor
```

####Perform a correlation analysis of: #### All variables

```
cor(stdsry_mat)
```

```
##             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
```

####A single correlation between two a pair of the variables

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

```
##
##  Pearson's product-moment correlation
##
## data:  stdsry_df$TimeReading and stdsry_df$TimeTV
## 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
## no Relation ship, negative cor
cor.test(stdsry_df$TimeReading, stdsry_df$Gender, method=c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  stdsry_df$TimeReading and stdsry_df$Gender
## 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
```

```r
## no Relation ship, negative cor
cor.test(stdsry_df$Happiness, stdsry_df$Gender, method=c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  stdsry_df$Happiness and stdsry_df$Gender
## 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
```

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

```r
cor.test(stdsry_df$Happiness, stdsry_df$Gender, method=c("pearson"),conf.level = 0.99 )
```

```
##
##  Pearson's product-moment correlation
##
## data:  stdsry_df$Happiness and stdsry_df$Gender
## t = 0.47695, df = 9, p-value = 0.6448
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.6365617  0.7890897
```

```
## sample estimates:
##       cor
## 0.1570118
```

```
## good Relation ship, positive cor
cor.test(stdsry_df$TimeReading, stdsry_df$TimeTV, method=c("pearson"),conf.level = 0.99 )
```

```
##
##  Pearson's product-moment correlation
##
## data:  stdsry_df$TimeReading and stdsry_df$TimeTV
## t = -5.6457, df = 9, p-value = 0.0003153
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.9801052 -0.4453124
## sample estimates:
##        cor
## -0.8830677
```

```
## no Relation ship, negative cor
cor.test(stdsry_df$TimeReading, stdsry_df$Gender, method=c("pearson"), conf.level = 0.99 )
```

```
##
##  Pearson's product-moment correlation
##
## data:  stdsry_df$TimeReading and stdsry_df$Gender
## t = -0.27001, df = 9, p-value = 0.7932
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.7618362  0.6755104
## sample estimates:
##         cor
## -0.08964215
```

####Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation. ####Calculate the correlation coefficient and the coefficient of determination, describe what you conclude about the results.

```
cov(stdsry_mat)
```

```
##              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
```

```
### Ans after analyzing the cov data, it seems most of covariance is showing towards negative direction
### in most of the comparison the variance is showing higher negative value , seems this value is due t
### used, however the variance for gender Vs others or time reading & time tv are showing some relevant
###Based on your analysis can you say that watching more TV caused students to read less? Explain.
### Ans : based on negative correlation , seems their is no relationship between reading and watching t
```

####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.

```
library(ppcor)
```

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

```
pcor(stdsry_mat)
```

```
## $estimate
##             TimeReading     TimeTV Happiness     Gender
## TimeReading   1.0000000 -0.8827973 0.4013124 -0.2706036
## TimeTV       -0.8827973  1.0000000 0.6311611 -0.2943135
## Happiness     0.4013124  0.6311611 1.0000000  0.2833152
## Gender       -0.2706036 -0.2943135 0.2833152  1.0000000
##
## $p.value
##             TimeReading      TimeTV  Happiness    Gender
## TimeReading 0.000000000 0.001615344 0.28437887 0.4812716
## TimeTV      0.001615344 0.000000000 0.06832112 0.4420392
## Happiness   0.284378868 0.068321119 0.00000000 0.4600603
## Gender      0.481271572 0.442039185 0.46006033 0.0000000
##
## $statistic
##             TimeReading     TimeTV Happiness     Gender
## TimeReading   0.0000000 -4.9720962 1.1592148 -0.7436966
## TimeTV       -4.9720962  0.0000000 2.1528933 -0.8147673
## Happiness     1.1592148  2.1528933 0.0000000  0.7816064
## Gender       -0.7436966 -0.8147673 0.7816064  0.0000000
##
## $n
## [1] 11
##
## $gp
## [1] 2
##
## $method
## [1] "pearson"
```

```
pcor.test(stdsry_df$TimeReading, stdsry_df$TimeT, stdsry_df$Happiness)
```

```
##    estimate      p.value statistic  n gp  Method
## 1 -0.872945 0.0009753126 -5.061434 11  1 pearson
```

```
pcor.test(stdsry_df$TimeReading, stdsry_df$TimeT, stdsry_df$Gender)
```

```
##     estimate      p.value statistic  n gp  Method
## 1 -0.8860628 0.0006411949 -5.406281 11  1 pearson
```

```
### Ans time reading and time telivison is having partial correlation while controlling for the effect
```