Assignment: ASSIGNMENT 5

Name: Anjale, Jiteshwar

Date: 2021-04-28

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.

# import required packages
library(ggplot2)
library(ggm)

## Warning: package 'ggm' was built under R version 4.0.5

## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/anjal/OneDrive/Desktop/MS/DSC520/dsc520")

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

head(student_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

## 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.

cov(student_df)

##              TimeReading       TimeTV  Happiness      Gender
## TimeReading   3.05454545 -20.36363636 -10.350091 -0.08181818
```

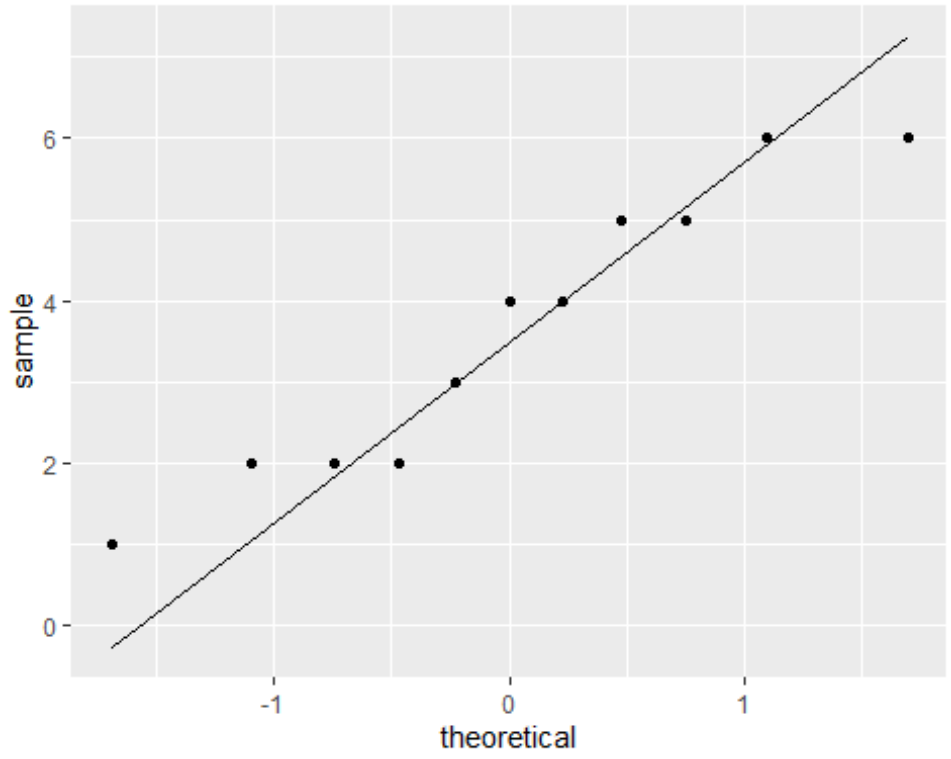
```
## TimeTV      -20.36363636 174.09090909 114.377273  0.04545455
## Happiness   -10.35009091 114.37727273 185.451422  1.11663636
## Gender       -0.08181818   0.04545455   1.116636  0.27272727

# Conclusions from the  covariance
# 1.Time of reading is negatively related to Time of watching TV,
# 2.Time of reading is negatively related to Happiness.
# 3.Time of watching TV is positively related to Happiness.
# 4.As gender is represented as integer, we can ignore the covariance
associated with gender.



# 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.

str(student_df)

## '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 ...

summary(student_df)

##   TimeReading        TimeTV        Happiness         Gender
##  Min.   :1.000   Min.   :50.00   Min.   :45.67   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:67.50   1st Qu.:65.34   1st Qu.:0.0000
##  Median :4.000   Median :75.00   Median :75.92   Median :1.0000
##  Mean   :3.636   Mean   :74.09   Mean   :73.31   Mean   :0.5455
##  3rd Qu.:5.000   3rd Qu.:82.50   3rd Qu.:83.83   3rd Qu.:1.0000
##  Max.   :6.000   Max.   :95.00   Max.   :89.52   Max.   :1.0000

# TimeReading - By looking at the values, I assumed that the TimeReading is
measure in minutes.
#               It looks like TimeReading varies from 1 minutes to 6 minutes.
# TimeTV - By looking at the values, I assumed that the TimeTV is measure in
minutes.
#         It looks like TimeTV varies from 50 minutes to 95 minutes.
# Happiness - By looking at the values, I assumed that the Happiness is
measure in percentages.
#             It looks like Happiness index varies from 45.67% to 89.52%
# Gender - By looking at the values, I assumed that the Gender is measure in
boolean.
#          It is not specified that 0 or 1 mean male/female. Need so more
info on the variable.
#
# covariance calculated for the variables have different units. I feel that
```

```
we need to use correlation coefficient to determine the relationship between these variable.
```

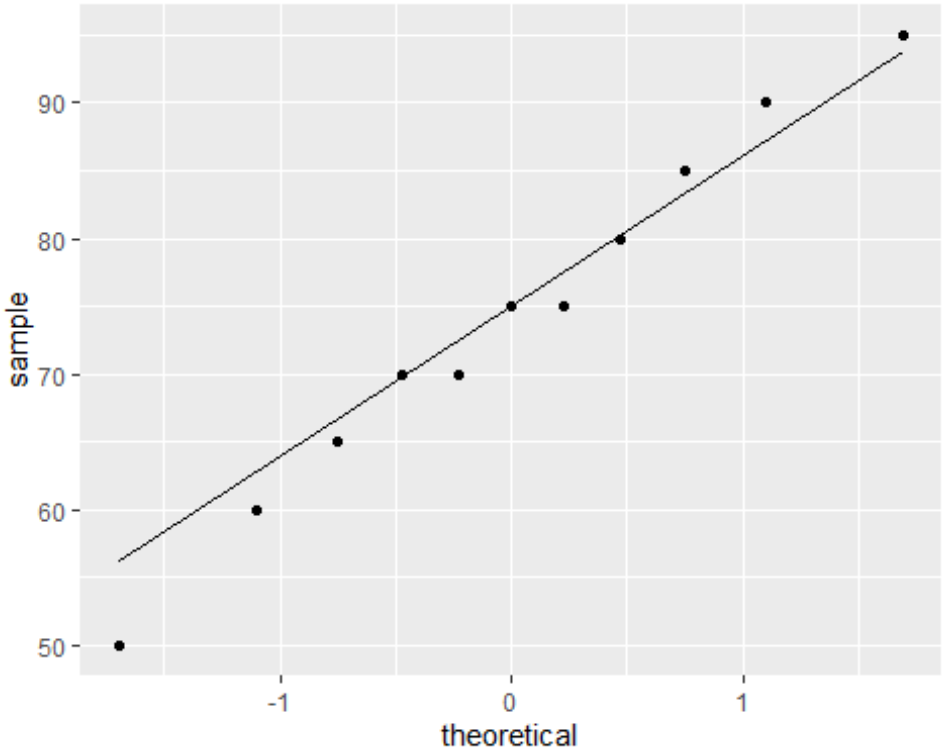
```
cor(student_df)
```

```
##             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
```

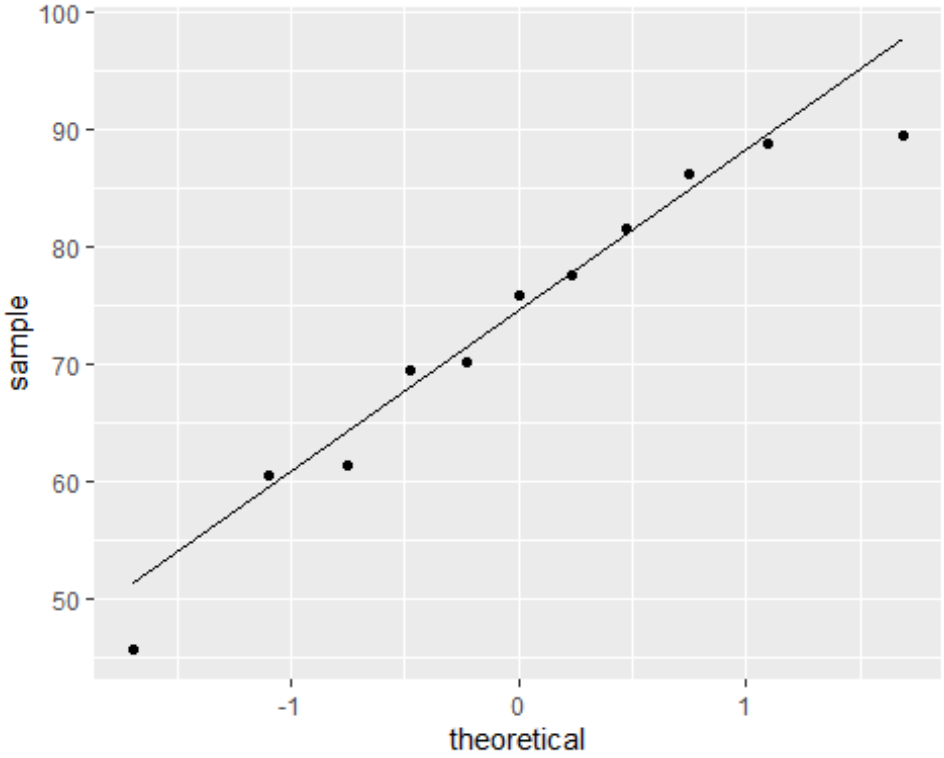
```
# iii.Choose the type of correlation test to perform, explain why you chose this test, and make a prediction if the test yields a positive or negative correlation?

#checking normality of data


#Probability Plot of the TimeReading variable.
ggplot(student_df, aes(sample=TimeReading)) + stat_qq() + stat_qq_line()
```




```
#Probability Plot of the TimeTV variable.
ggplot(student_df, aes(sample=TimeTV)) + stat_qq() + stat_qq_line()
```




```
#Probability Plot of the Happiness variable.
ggplot(student_df, aes(sample=Happiness)) + stat_qq() + stat_qq_line()
```



```r
#By looking at plots, I can confirm that data is normally distributed. we can used Perason's correlation coefficient to check the correlation between variables.

cor.test(student_df$TimeReading,student_df$TimeTV)

##
##  Pearson's product-moment correlation
##
## data:  student_df$TimeReading and student_df$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

cor.test(student_df$TimeReading,student_df$Happiness)

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

cor.test(student_df$Happiness,student_df$TimeTV)

##
##  Pearson's product-moment correlation
##
## data:  student_df$Happiness and student_df$TimeTV
## 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

# iv. Perform a correlation analysis of:
# 1.All variables
cor(student_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
```

```
# 2.A single correlation between two a pair of the variables
cor(student_df$TimeReading, student_df$TimeTV)
```

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

```
# 3.Repeat your correlation test in step 2 but set the confidence interval at
99%
cor.test(student_df$TimeReading, student_df$TimeTV, conf.level = 0.99)
```

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

```
# Describe what the calculations in the correlation matrix suggest about the
relationship between the variables. Be specific with your explanation.
cor(student_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
```

```
#Correlation coefficient between a variable and itself is 1 i.e. completely
positively correlated. Correlation coefficient is < 0 that would signify
negative correlation.
# TimeReading and TimeTV have negative correlation.
# TimeReading and Happiness have negative correlation.
# TimeTV and Happiness have positive correlation.

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

# calculating correlation coefficient (r) between variables
cor(student_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
```

```
# calculating coefficient of determination (r^2) between two variables
cor(student_df)^2
```

```
##             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
```

```
# The coefficient of determination is a measurement used to explain how much
variability of one factor can be caused by its relationship to another
related factor. This correlation, known as the "goodness of fit," is
# represented as a value between 0.0 and 1.0.

# Looking at coeffiecient of determination between TimeTV and Happiness
shared variability is
# about 40% which would imply that TV time variability effects Happiness upto
40% only, while
# remaining 60% variability in Happiness must be caused by some other
variable.



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

cor(student_df$TimeReading, student_df$TimeTV)^2
```

```
## [1] 0.7798085
```

```
# Looking at coefficient of determination (r^2) we can say that variability
in TimeReading can cause upto 77% variability in TimeTV
# There could be other variables that may cause 23% variability in TimeTV.

# 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.

student_df2 <- student_df[,1:3]

# Run partial correlation between TimeTV and Happiness while controlling
TimeReading

pcor(c("TimeTV","Happiness","TimeReading"), var(student_df2))
```

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

```
pcor(c("TimeTV","Happiness","TimeReading"), var(student_df2))^2
```

```
## [1] 0.3571871
```

```
#If we keep TimeReading controlling , the correlation coefficient between TV
time and happiness decrease to 0.59
# and coefficient of determination has decreased to 35%. This decrease
suggests that variation in Happiness was also effected positively by
TimeReading by about 5%.
```