Homework 6

Darian-Florian Voda

2022-12-01

Loading packages

```
library(dplyr)
library(ggplot2)
library(car)
library(tidyverse)
library(ggpubr)
library(rstatix)
```

Exercise 70

- A fast food franchise is test marketing 3 new menu items.
- 18 franchisee restaurants are randomly chosen for participation in the study.
- 6 of the restaurants are randomly chosen to test market the first new menu item, another 6 for the second menu item, and the remaining 6 for the last menu item.
- At .05 level of significance, test whether the mean sales volume for the 3 new menu items are all equal.
- Dataset: fastfood.txt [H_0 is considered that the mean sales volume is equal for all 3 menu items]

```
#### Exercise 70 ####

data<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/fastfood.t
                 stringsAsFactors=F)

# EDA
head(data)
```

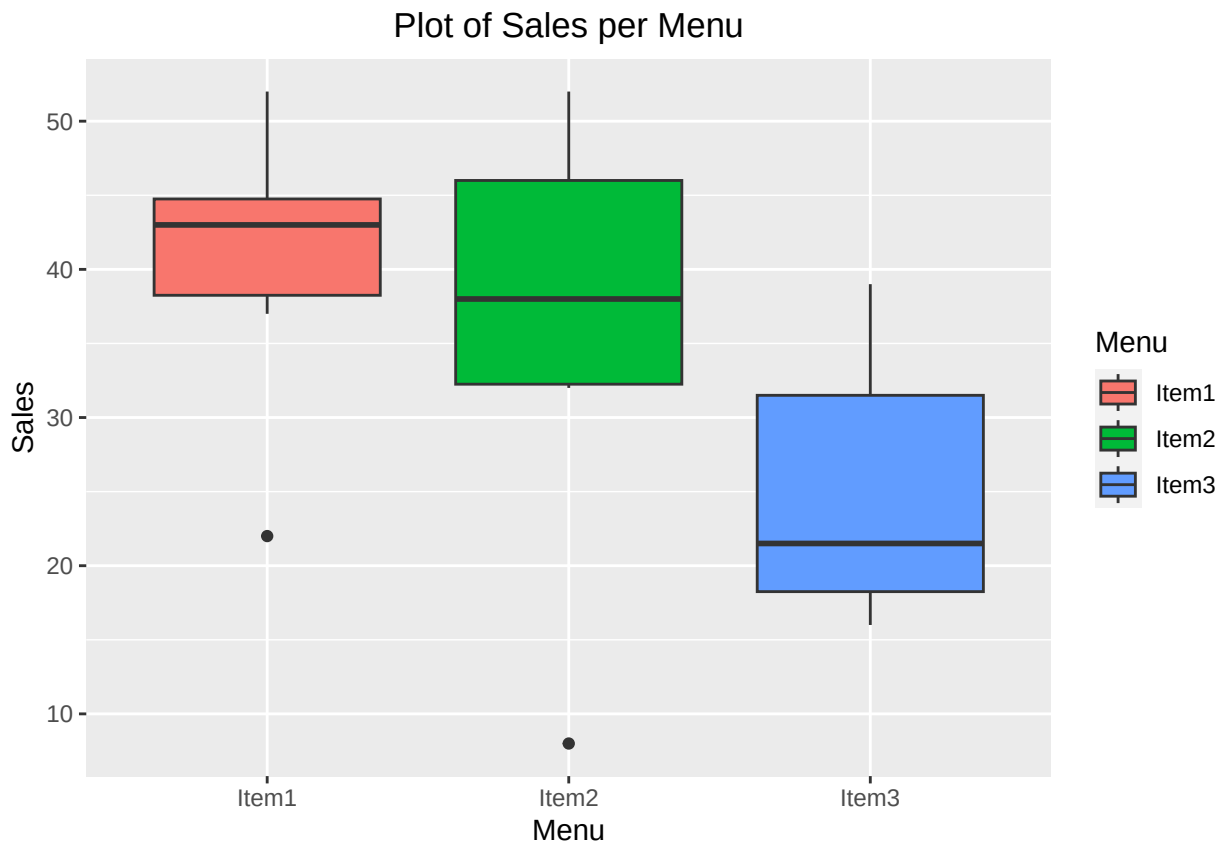
```
##    Menu Sales
## 1 Item1    22
## 2 Item1    42
## 3 Item1    44
## 4 Item1    52
## 5 Item1    45
## 6 Item1    37
```

```
# EDA
group_by(data, Menu) %>%
  summarise(
    count = n(),
```

```
    mean = mean(Sales, na.rm = TRUE),
    sd = sd(Sales, na.rm = TRUE)
    )
```

```
## # A tibble: 3 x 4
##   Menu  count  mean    sd
##   <chr> <int> <dbl> <dbl>
## 1 Item1     6  40.3 10.2
## 2 Item2     6  35.8 15.7
## 3 Item3     6  25    9.42
```

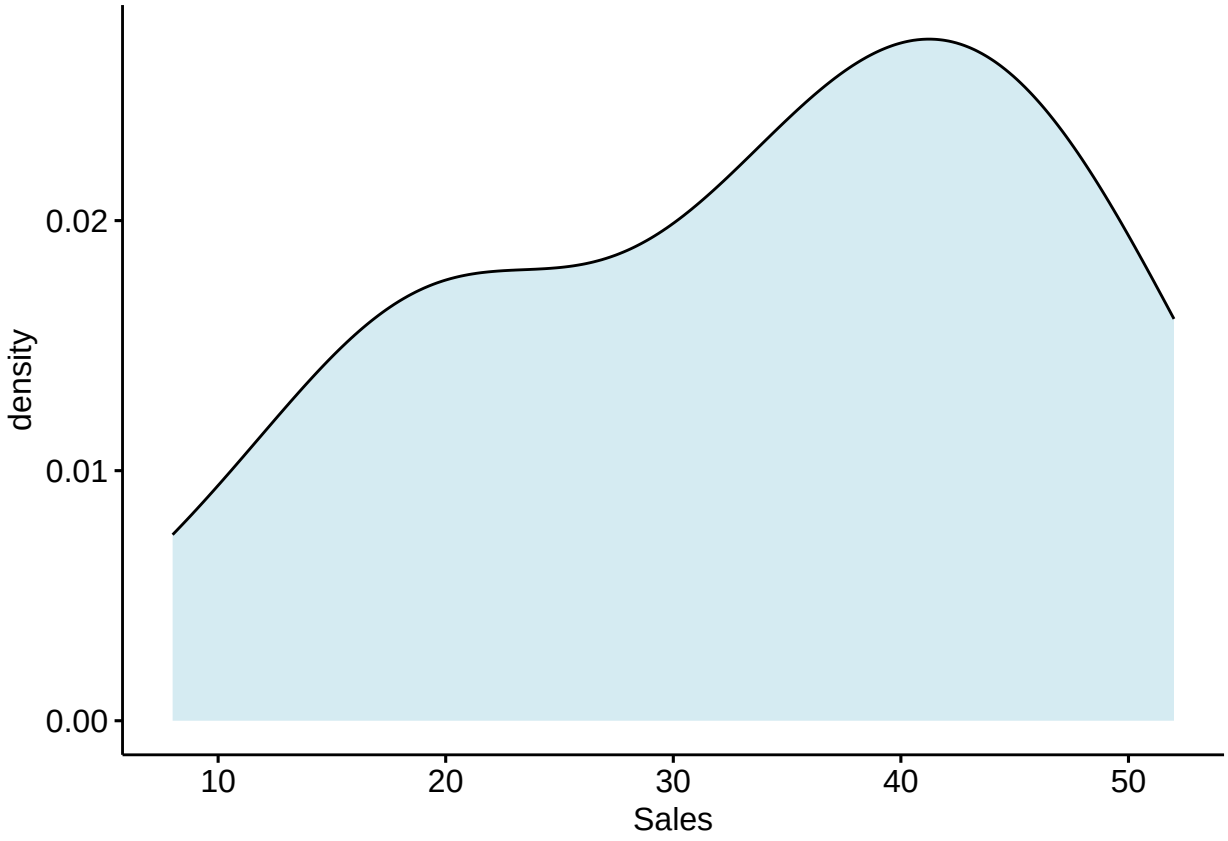
```
# Boxplot
ggplot(data, aes(x=Menu, y=Sales, fill=Menu)) +
  geom_boxplot()+
  labs(title="Plot of Sales per Menu",x="Menu", y = "Sales")+
  theme(plot.title = element_text(hjust = 0.5))
```



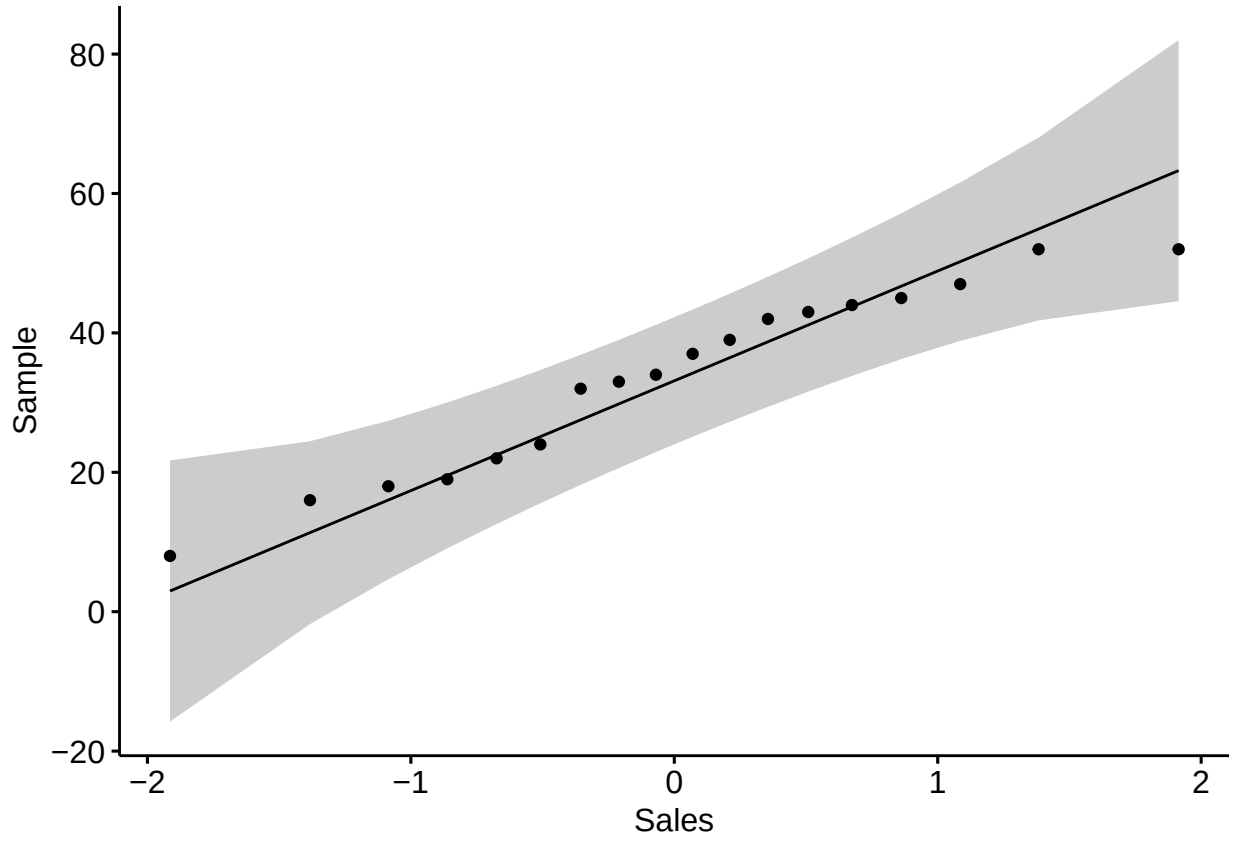

```
# Levene's Test for homogeneity
leveneTest(Sales ~ Menu, data = data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  0.6421   0.54
##       15
```

```r
# Density plot
ggdensity(data$Sales, fill = "lightblue") + labs(x="Sales")
```




```r
# QQ plot
ggqqplot(data$Sales) + labs(x="Sales")
```



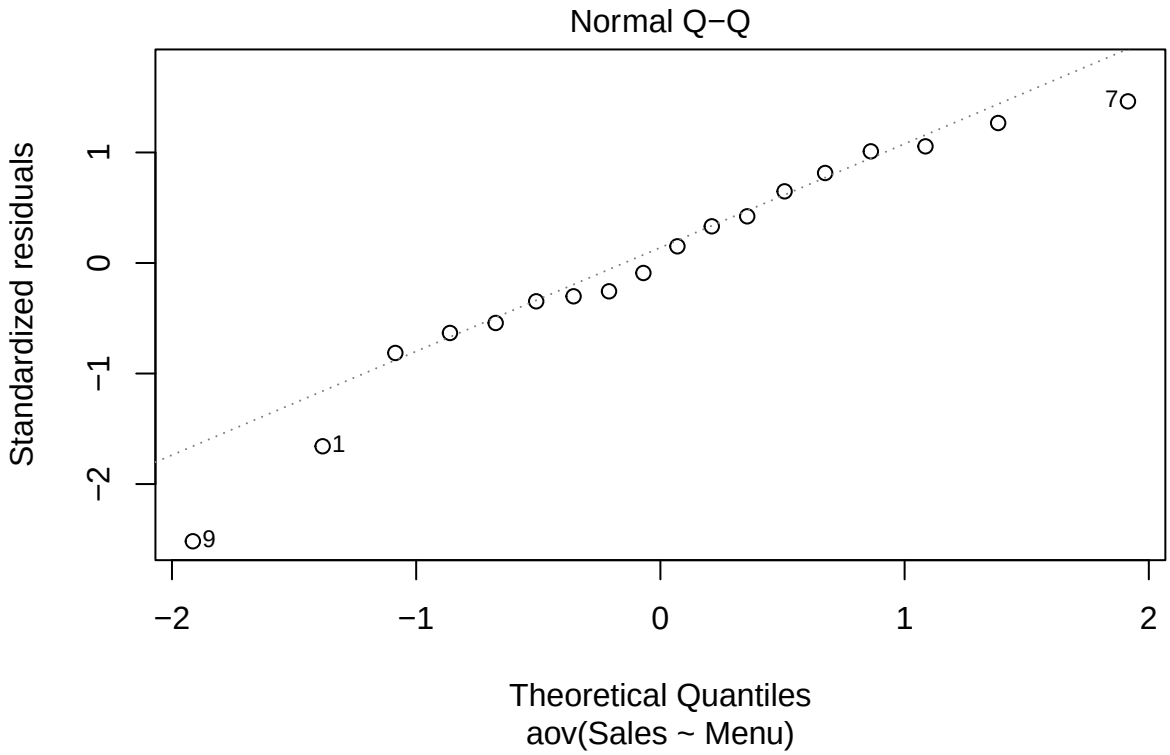

```
# ANOVA test
res.aov <- aov(Sales ~ Menu, data = data)

# Anova variable
res.aov
```

```
## Call:
##    aov(formula = Sales ~ Menu, data = data)
##
## Terms:
##                       Menu Residuals
## Sum of Squares   745.4444 2200.1667
## Deg. of Freedom         2        15
##
## Residual standard error: 12.11106
## Estimated effects may be unbalanced
```

```
# Extracting the p-value and F value
summary(res.aov)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Menu         2  745.4   372.7   2.541  0.112
## Residuals   15 2200.2   146.7
```

```r
# QQ plot of ANOVA
plot(res.aov, 2)
```




```r
# Extract residuals
aov_residuals <- residuals(object = res.aov)

# Normality test using Shapiro-Wilkins test
shapiro.test(x = aov_residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.94962, p-value = 0.4191
```

Thus, we can conclude that:

- H_0 is accepted, since the p-value from the ANOVA test is 0.112 which is greater than 0.05
- This is quite interesting, since the boxplot and density plot doesn't show equal mean values and a normal distribution
- Performed a Levene's test followed by a Shapiro-Wilk test to prove it's homogeneity and normality
- Each p-value for our test proves that our data can be trusted in an ANOVA test
- Thus, there **are** equal sales mean volumes between the Menus due to the fact that the sample size is quite small

Exercise 71

- Use the data set 'ICM'.
- At 0.05 level of significance, test whether the means of the negative mood of students are equal between the groups of social media use. [H_0 - the means of negative mood are equal between the social media use groups]

```
#### Exercise 71 ####

ICM<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/ICM.txt",
                stringsAsFactors=F)

# EDA
head(ICM)
```

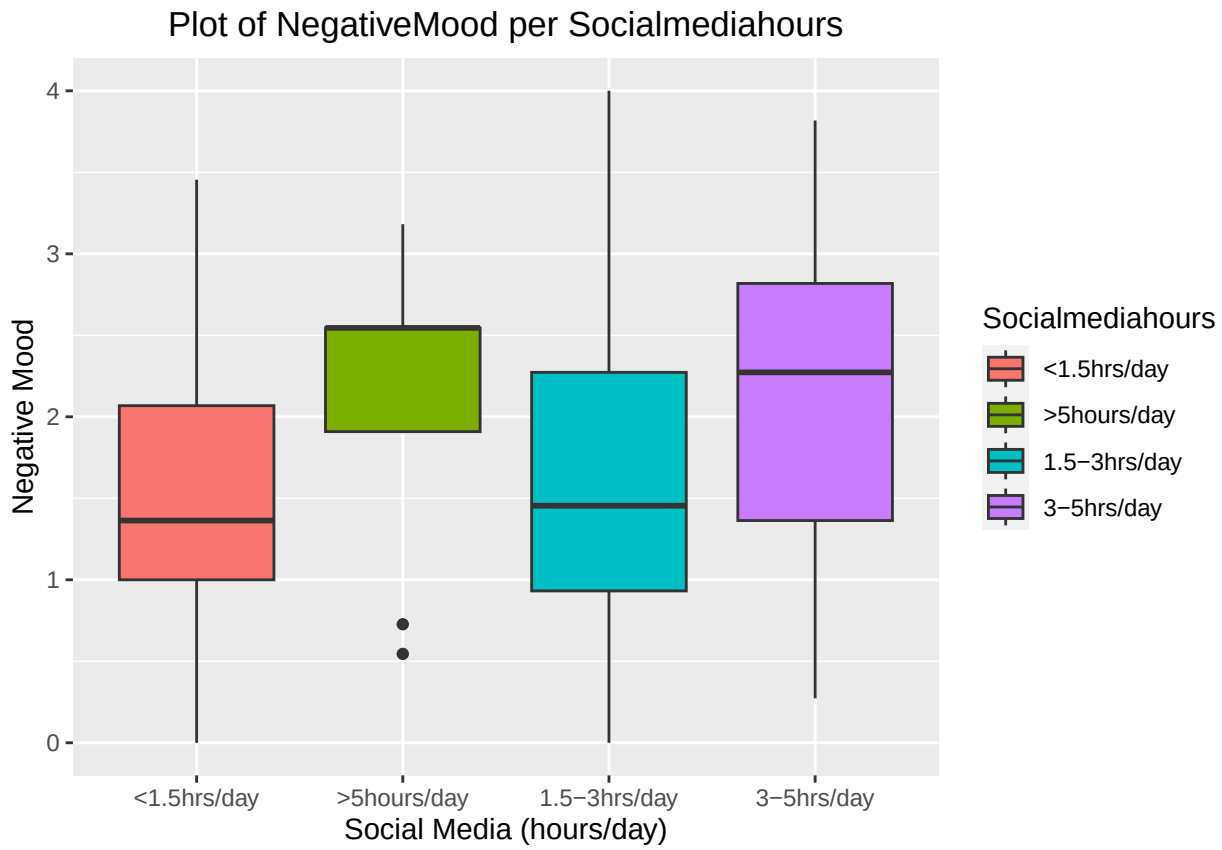
```
##   ï..ID Gender Age Englishfluent Germanfluent       Transport
## 1    75 female  22           yes           no PublicTransport
## 2    90 female  22           yes           no PublicTransport
## 3   173 female  37           yes          yes             Car
## 4   189 female  17           yes          yes             Car
## 5   100 female  19           yes          yes            Walk
## 6   155 female  16           yes           no            Walk
##   Highest_level_of_education Do_you_smoke Socialmediahours Timewithfriends Pet
## 1                    College           No     1.5-3hrs/day     2-5hrs/week  No
## 2                    College           No     1.5-3hrs/day     2-5hrs/week  No
## 3                 University           No      <1.5hrs/day    5-10hrs/week Yes
## 4                       none           No     1.5-3hrs/day   10-20hrs/week Yes
## 5                 HighSchool           No       3-5hrs/day     >20hrs/week  No
## 6                       none           No     1.5-3hrs/day   10-20hrs/week  No
##   Siblings Children Relationshipstatus Activitieshours NegativeMood
## 1      Yes       No       Relationship              10           NA
## 2      Yes       No       Relationship              10           NA
## 3       No      Yes       Relationship              20           NA
## 4      Yes       No             Single              40     4.000000
## 5      Yes       No             Single              20     2.818182
## 6      Yes       No             Single              10     2.454545
##   PositiveMood Mentalhealth Socialization Activity SocialSupport
## 1           NA    2.6666667            NA      2.8     4.0000000
## 2           NA    2.6666667            NA      2.8     4.0000000
## 3           NA    3.5000000            NA      3.4     2.3333333
## 4    0.0000000    1.0000000           1.0      3.2     0.6666667
## 5    0.3333333    0.8333333           2.5      1.2     2.3333333
## 6    0.3333333    1.6666667           2.5      2.6     1.3333333
##   Communication_open_direct      OHS
## 1                        NA 4.586207
## 2                        NA 4.586207
## 3                  3.384615 5.103448
## 4                  3.615385 3.137931
## 5                  3.153846 2.758621
## 6                  3.461538 3.586207
```

```
# EDA
group_by(ICM, Socialmediahours) %>%
```

```
  summarise(
    count = n(),
    mean = mean(NegativeMood, na.rm = TRUE),
    sd = sd(NegativeMood, na.rm = TRUE)
  )
```

```
## # A tibble: 4 x 4
##   Socialmediahours count  mean    sd
##   <chr>            <int> <dbl> <dbl>
## 1 <1.5hrs/day         64  1.52 0.817
## 2 >5hours/day         10  2.11 0.903
## 3 1.5-3hrs/day        88  1.58 0.855
## 4 3-5hrs/day          37  2.08 0.988
```

```
# Boxplot
ggplot(ICM, aes(x=Socialmediahours, y=NegativeMood, fill=Socialmediahours)) +
  geom_boxplot()+
  labs(title="Plot of NegativeMood per Socialmediahours",x="Social Media (hours/day)", y = "Negative Mo
  theme(plot.title = element_text(hjust = 0.5))
```

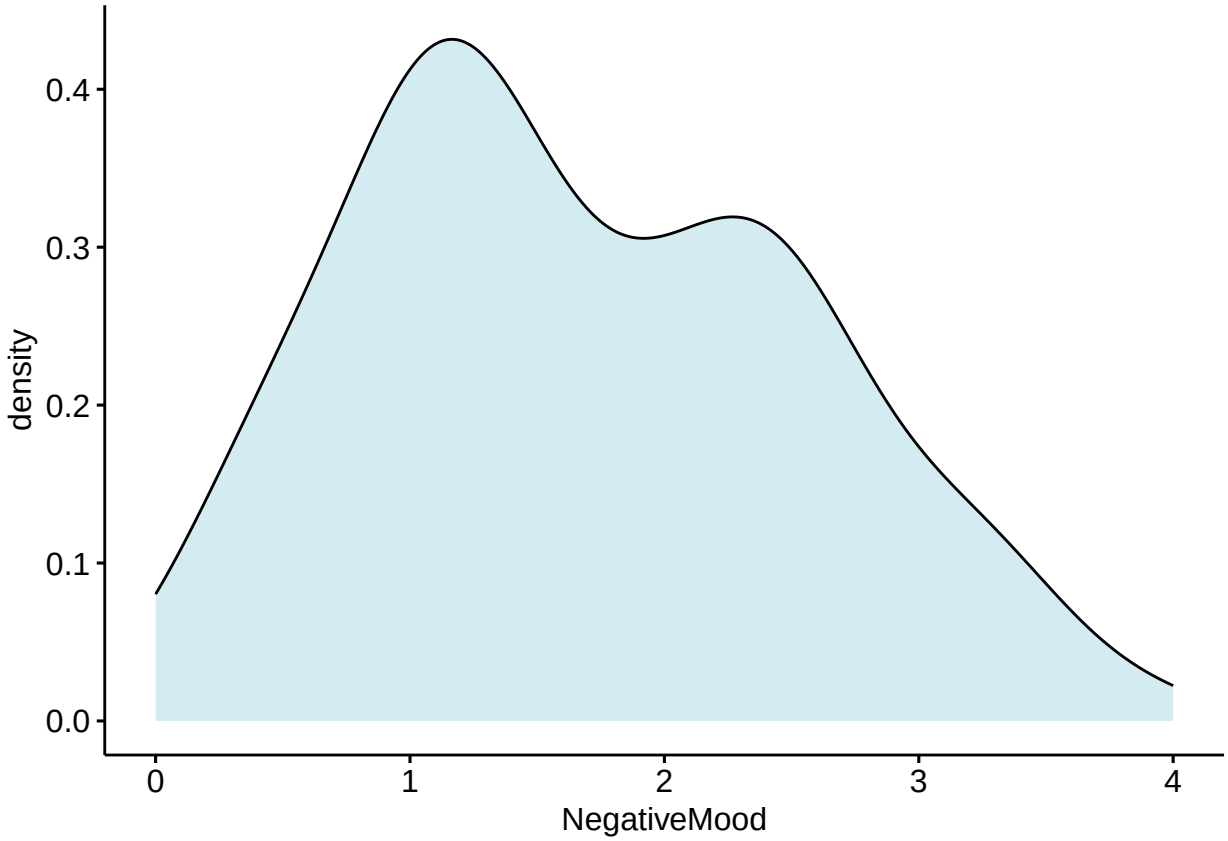



```
# Levene's Test for homogeneity
leveneTest(NegativeMood ~ Socialmediahours, data = ICM)
```

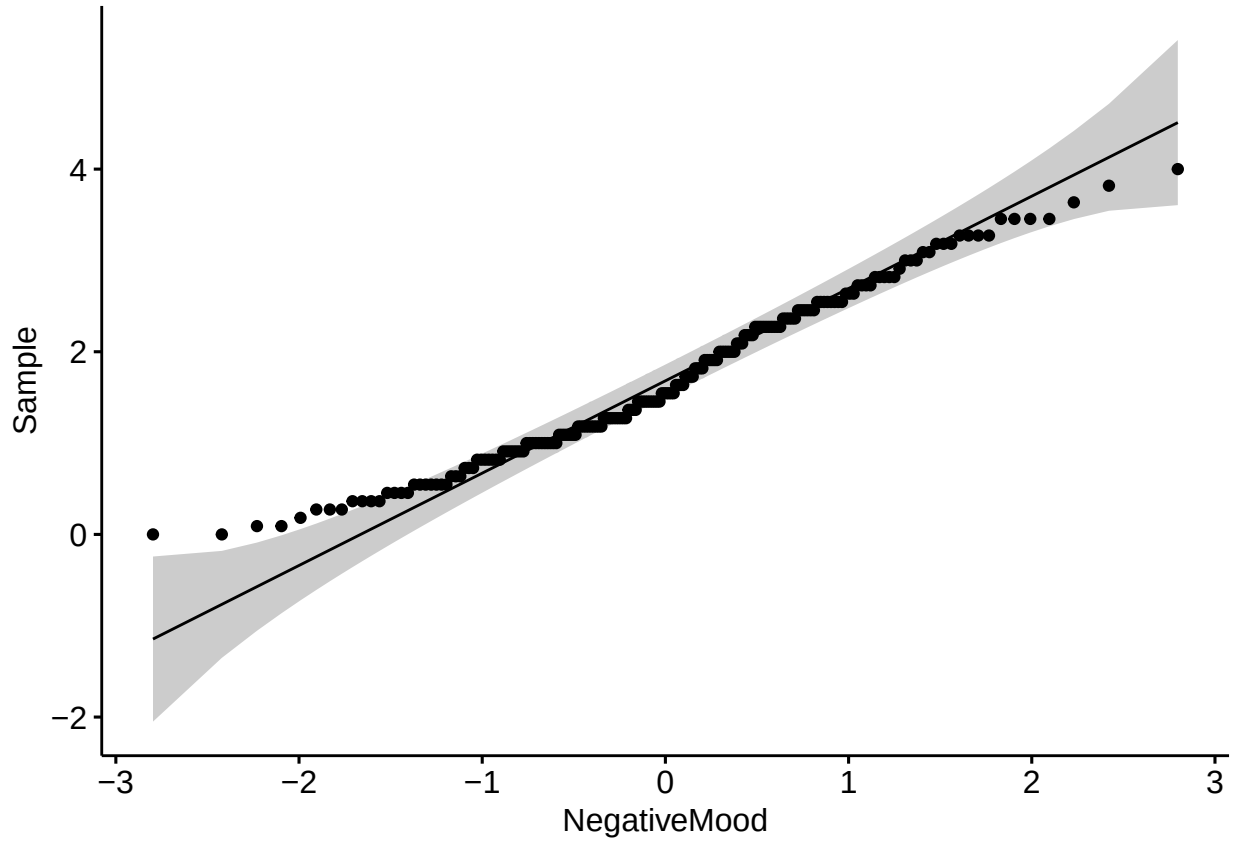
```
## Levene's Test for Homogeneity of Variance (center = median)
```

```
##         Df F value Pr(>F)
## group    3  0.6744 0.5687
##        190
```

```
# Density plot
ggdensity(ICM$NegativeMood, fill = "lightblue") + labs(x="NegativeMood")
```




```
# QQ plot
ggqqplot(ICM$NegativeMood) + labs(x="NegativeMood")
```




```
# ANOVA test
res.aov <- aov(NegativeMood ~ Socialmediahours, data = ICM)

# Anova variable
res.aov
```

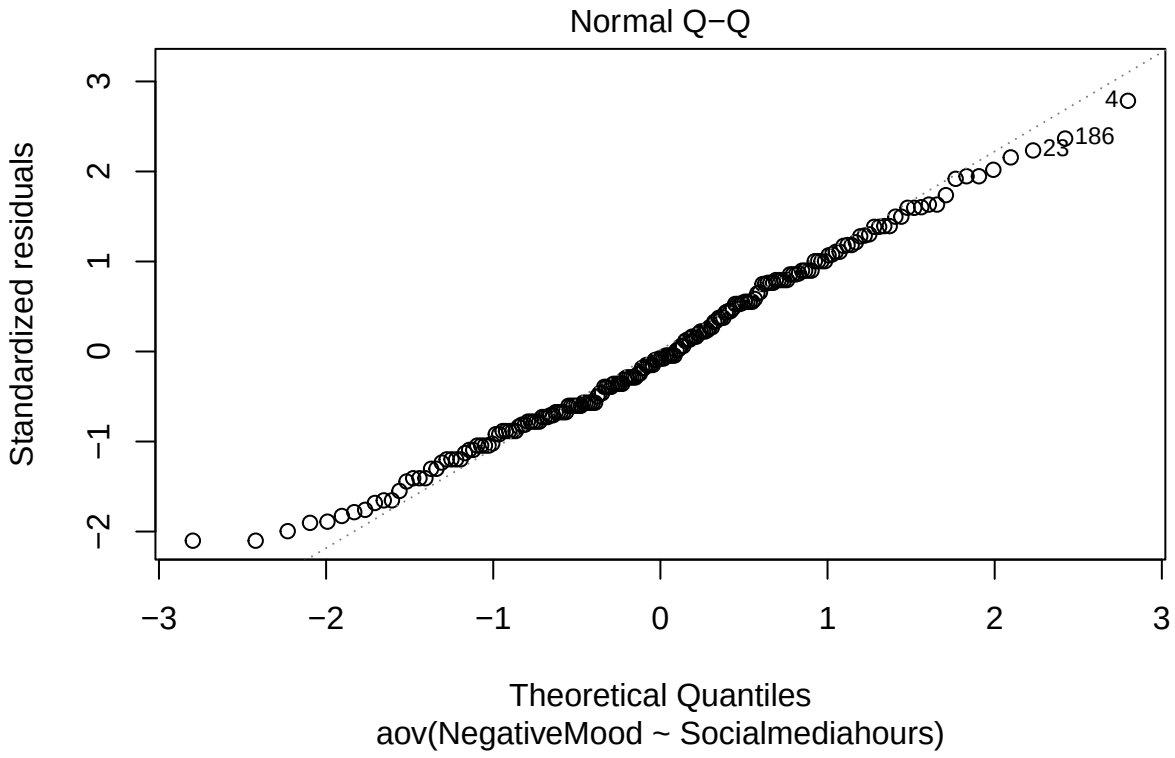
```
## Call:
##    aov(formula = NegativeMood ~ Socialmediahours, data = ICM)
##
## Terms:
##                 Socialmediahours Residuals
## Sum of Squares           9.95324 144.59567
## Deg. of Freedom                3       190
##
## Residual standard error: 0.8723702
## Estimated effects may be unbalanced
## 5 observations deleted due to missingness
```

```
# Extracting the p-value and F value
summary(res.aov)
```

```
##                   Df Sum Sq Mean Sq F value  Pr(>F)
## Socialmediahours   3   9.95   3.318    4.36 0.00538 **
## Residuals        190 144.60   0.761
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5 observations deleted due to missingness
```

```
# QQ plot of ANOVA
plot(res.aov, 2)
```




```
# Get rid of outliers
aov_residuals <- residuals(object = res.aov)

# Normality test using Shapiro-Wilkins test
shapiro.test(x = aov_residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.98989, p-value = 0.1883
```

Thus, we can conclude that:

- H_0 is rejected, since the p-value from the ANOVA test is 0.00538 which is less than 0.05
- Performed a Levene's test followed by a Shapiro-Wilk test to prove it's homogeneity and normality
- Each p-value for our test proves that our data can be trusted in an ANOVA test
- Thus, there **are no** equal Negative mood means between the Social media in hours per day group

Exercise 72

- Use the data set 'ICM'.
- At 0.05 level of significance, test whether the means of the socialization of students are equal between the groups of time spent with friends. [H_0 - the means of socialization are equal between the time spend with friends groups]

```
#### Exercise 72 ####

ICM<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/ICM.txt",
                stringsAsFactors=F)

# EDA
head(ICM)
```

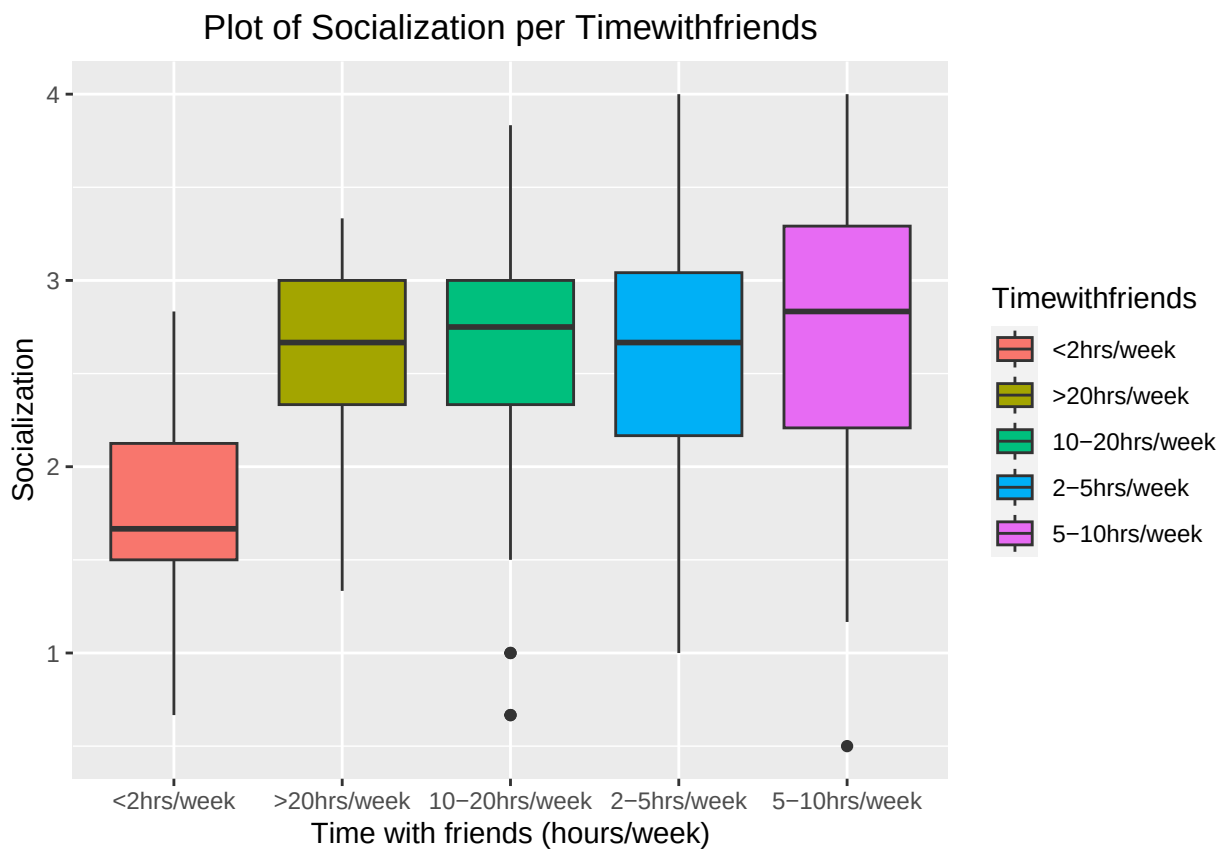
```
##   ï..ID Gender Age Englishfluent Germanfluent       Transport
## 1    75 female  22           yes           no PublicTransport
## 2    90 female  22           yes           no PublicTransport
## 3   173 female  37           yes          yes             Car
## 4   189 female  17           yes          yes             Car
## 5   100 female  19           yes          yes            Walk
## 6   155 female  16           yes           no            Walk
##   Highest_level_of_education Do_you_smoke Socialmediahours Timewithfriends Pet
## 1                    College           No     1.5-3hrs/day     2-5hrs/week  No
## 2                    College           No     1.5-3hrs/day     2-5hrs/week  No
## 3                 University           No      <1.5hrs/day    5-10hrs/week Yes
## 4                       none           No     1.5-3hrs/day   10-20hrs/week Yes
## 5                 HighSchool           No       3-5hrs/day     >20hrs/week  No
## 6                       none           No     1.5-3hrs/day   10-20hrs/week  No
##   Siblings Children Relationshipstatus Activitieshours NegativeMood
## 1      Yes       No       Relationship              10           NA
## 2      Yes       No       Relationship              10           NA
## 3       No      Yes       Relationship              20           NA
## 4      Yes       No             Single              40     4.000000
## 5      Yes       No             Single              20     2.818182
## 6      Yes       No             Single              10     2.454545
##   PositiveMood Mentalhealth Socialization Activity SocialSupport
## 1           NA    2.6666667            NA      2.8     4.0000000
## 2           NA    2.6666667            NA      2.8     4.0000000
## 3           NA    3.5000000            NA      3.4     2.3333333
## 4    0.0000000    1.0000000           1.0      3.2     0.6666667
## 5    0.3333333    0.8333333           2.5      1.2     2.3333333
## 6    0.3333333    1.6666667           2.5      2.6     1.3333333
##   Communication_open_direct      OHS
## 1                        NA 4.586207
## 2                        NA 4.586207
## 3                  3.384615 5.103448
## 4                  3.615385 3.137931
## 5                  3.153846 2.758621
## 6                  3.461538 3.586207
```

```
# EDA
group_by(ICM, Timewithfriends) %>%
```

```
  summarise(
    count = n(),
    mean = mean(Socialization, na.rm = TRUE),
    sd = sd(Socialization, na.rm = TRUE)
  )
```

```
## # A tibble: 5 x 4
##   Timewithfriends count  mean    sd
##   <chr>           <int> <dbl> <dbl>
## 1 <2hrs/week         26  1.80 0.570
## 2 >20hrs/week        21  2.63 0.547
## 3 10-20hrs/week      32  2.54 0.830
## 4 2-5hrs/week        60  2.59 0.678
## 5 5-10hrs/week       60  2.70 0.757
```

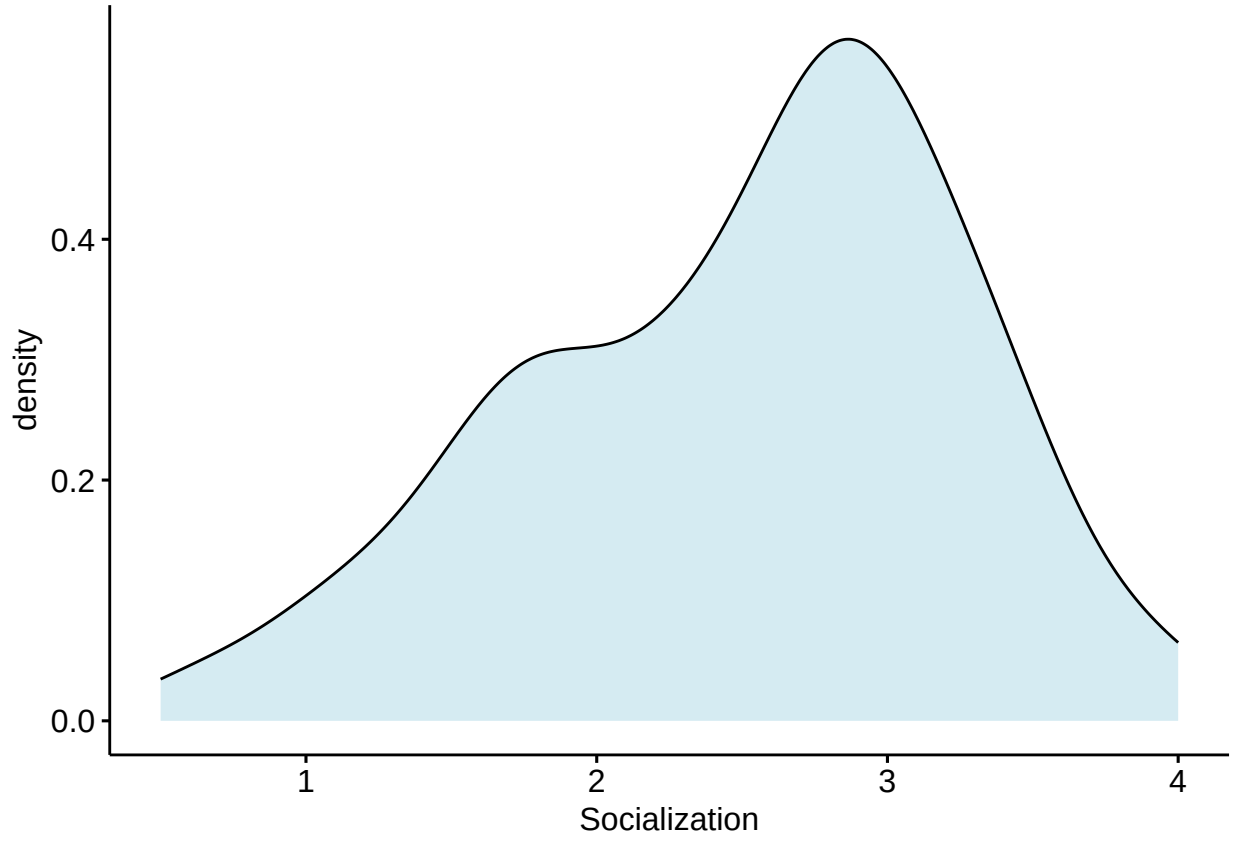
```
# Boxplot
ggplot(ICM, aes(x=Timewithfriends, y=Socialization, fill=Timewithfriends)) +
  geom_boxplot()+
  labs(title="Plot of Socialization per Timewithfriends",x="Time with friends (hours/week)",
       y = "Socialization")+
  theme(plot.title = element_text(hjust = 0.5))
```



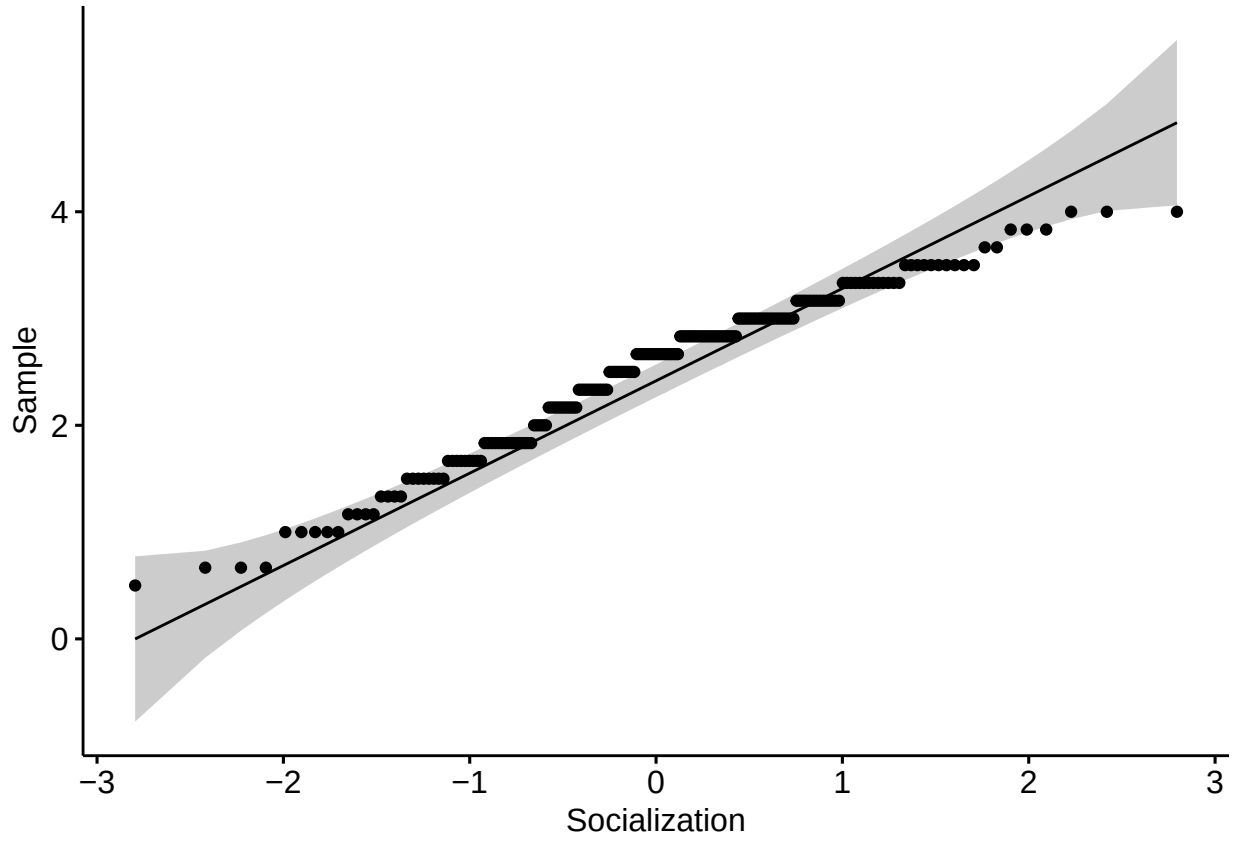

```
# Levene's Test for homogeneity
leveneTest(Socialization ~ Timewithfriends, data = ICM)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  0.8679 0.4843
##       188
```

```
# Density plot
ggdensity(ICM$Socialization, fill = "lightblue") + labs(x="Socialization")
```




```
# QQ plot
ggqqplot(ICM$Socialization) + labs(x="Socialization")
```




```
# ANOVA test
res.aov <- aov(Socialization ~ Timewithfriends, data = ICM)

# Anova variable
res.aov
```

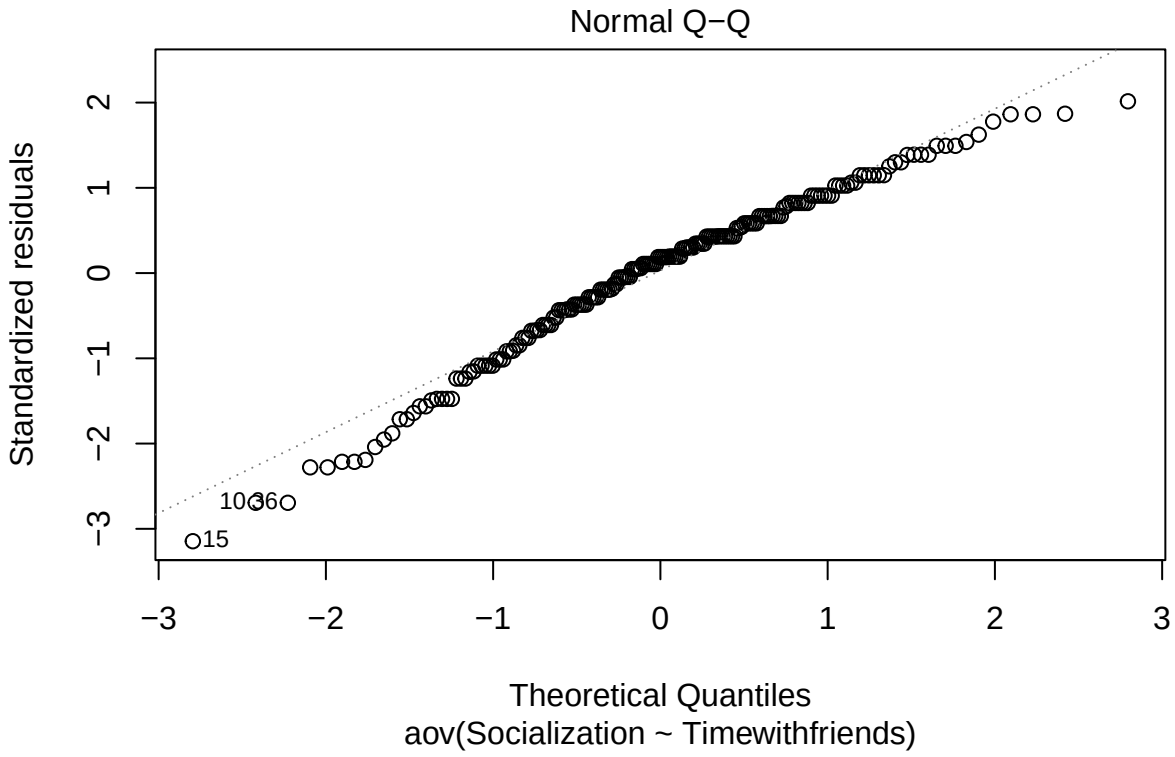
```
## Call:
##    aov(formula = Socialization ~ Timewithfriends, data = ICM)
##
## Terms:
##                 Timewithfriends Residuals
## Sum of Squares         15.80318  93.44638
## Deg. of Freedom               4       188
##
## Residual standard error: 0.7050214
## Estimated effects may be unbalanced
## 6 observations deleted due to missingness
```

```
# Extracting the p-value and F value
summary(res.aov)
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## Timewithfriends   4  15.80   3.951   7.948 6.11e-06 ***
## Residuals       188  93.45   0.497
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 6 observations deleted due to missingness
```

```
# QQ plot of ANOVA
plot(res.aov, 2)
```




```
# Get rid of outliers
aov_residuals <- residuals(object = res.aov)

# Normality test using Shapiro-Wilkins test
shapiro.test(x = aov_residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.97375, p-value = 0.00109
```

Thus, we can conclude that:

- H_0 is rejected, since the p-value from the ANOVA test is 0.00000611 which is less than 0.05
- Performed a Levene's test followed by a Shapiro-Wilk test to prove it's homogeneity and normality
- Each p-value for our test proves that our data can be trusted in an ANOVA test
- Thus, there **are no** equal Socialization means between the Time spent with friends in hours per week group

Exercise 75

- Use the dataset mtcars and apply a simple linear regression model to estimate the miles per gallon if the weight of the automobile is 3 (in 1000 lbs).
- Are the the assumptions met for linear regression?
- Find the coefficient of determination.
- Is there a significant relationship between the variables? [H_0 - there is no significant relationship]
- Develop a 95% confidence interval of the mean miles per gallon for the weight of 3.
- Plot the residual of the simple linear regression model against the independent variable.
- Normal probability plot for the standardized residual.

```
#### Exercise 75 ####
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
# linear model
lm(mpg ~ wt, data=mtcars)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Coefficients:
## (Intercept)           wt
##      37.285       -5.344
```

```
mpg.lm = lm(mpg ~ wt, data=mtcars)

summary(mpg.lm)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

```
coeffs = coefficients(mpg.lm)
coeffs
```

```
## (Intercept)          wt
##   37.285126   -5.344472
```

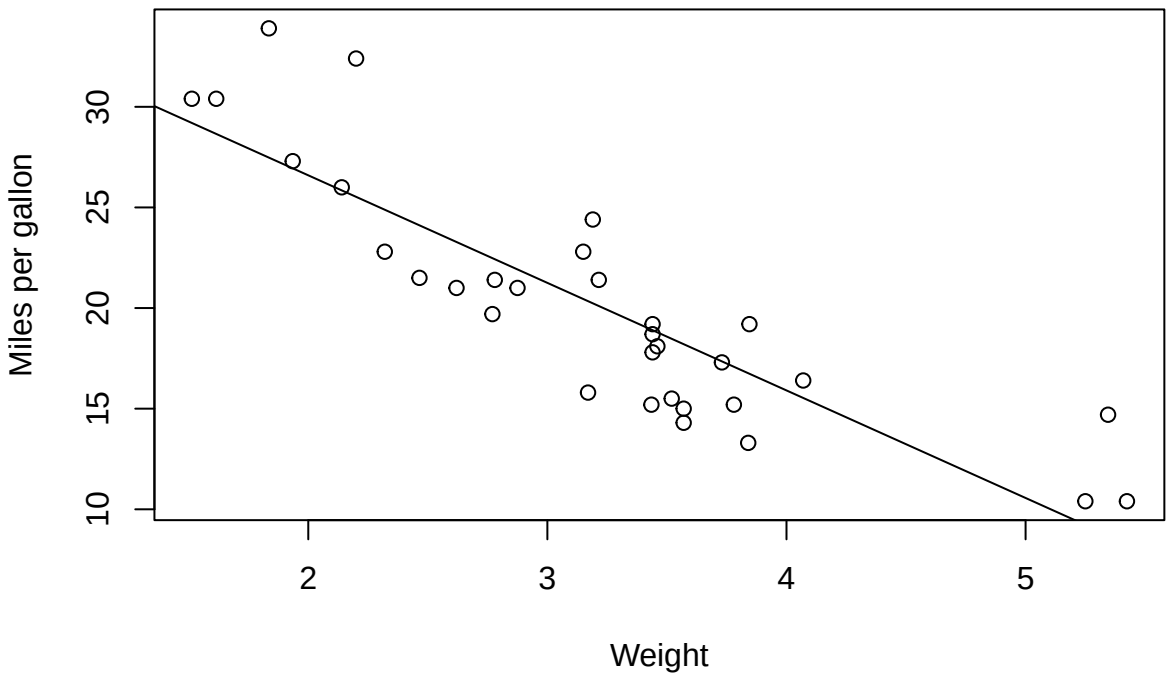
```
# weight (in 1000 lbs)
weight.auto = 3.00
duration = coeffs[1] + coeffs[2]*weight.auto

# mpg with respect to weight=3
duration
```

```
## (Intercept)
##    21.25171
```

```
paste("Based on the simple linear regression model,
      if the weight of the cars has been 3.00 (in 1000 lbs),
      we expect to consume 1 gallon of gas at 21.25 miles")
```

```
## [1] "Based on the simple linear regression model, \n      if the weight of the cars has been 3.00 (i
```

```
# Plot
plot(mtcars$wt, mtcars$mpg, xlab="Weight", ylab="Miles per gallon")
abline(lm(mtcars$mpg ~ mtcars$wt))
```




```
# Coefficient determination
mpg.lm = lm(mpg ~ wt, data=mtcars)
summary(mpg.lm)$r.squared
```

```
## [1] 0.7528328
```

```
paste("The results suggests that 75% of the dependent variable is predicted by the independent variable
```

```
## [1] "The results suggests that 75% of the dependent variable is predicted by the independent variable
```

```
# Significant relationship between variables
mpg.lm = lm(mpg ~ wt, data=mtcars)
summary(mpg.lm)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

```
paste("As the p-value is 0.000000000129,
      which is much less than 0.05,
      we reject the null hypothesis that beta = 0.")
```

```
## [1] "As the p-value is 0.000000000129, \n      which is much less than 0.05, \n      we reject the nu
```

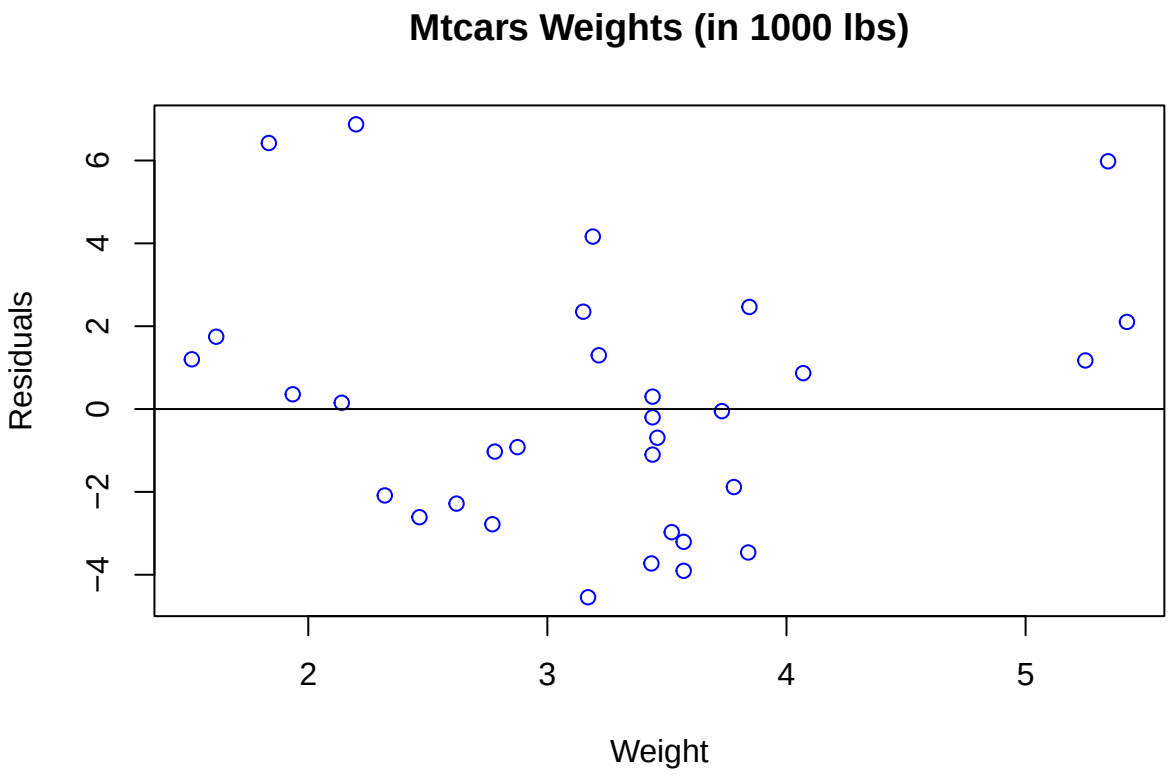
```
# Confidence Interval for weight = 3
mpg.lm = lm(mpg ~ wt, data=mtcars)
newdata=data.frame(wt=3.00)
predict(mpg.lm, newdata, interval="confidence")
```

```
##        fit      lwr      upr
## 1 21.25171 20.12444 22.37899
```

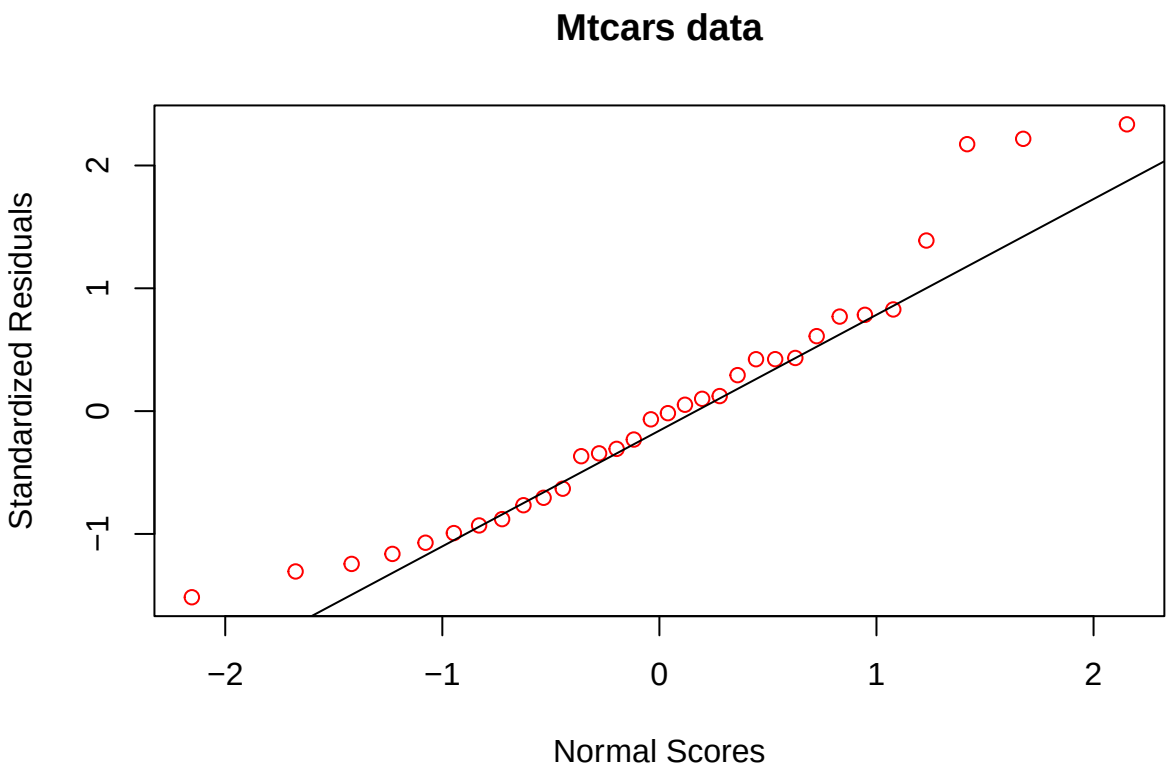
```
paste("The 95% confidence interval of the mean miles per gallon for the weight of 3.00,
      is between 20.12444 and 22.37899 miles per gallon.")
```

```
## [1] "The 95% confidence interval of the mean miles per gallon for the weight of 3.00,\n      is betwe
```

```
# Residual Plot
mpg.lm = lm(mpg ~ wt, data=mtcars)
mpg.res=resid(mpg.lm)
plot(mtcars$wt, mpg.res, ylab="Residuals", xlab="Weight",
     main="Mtcars Weights (in 1000 lbs)", col="blue")
abline(0, 0)
```




```
# Normal Probability Plot of Residuals
mpg.lm = lm(mpg ~ wt, data=mtcars)
mpg.stdres = rstandard(mpg.lm)
qqnorm(mpg.stdres, ylab="Standardized Residuals", xlab="Normal Scores", main="Mtcars data", col="red")
qqline(mpg.stdres)
```



Thus, we can conclude that:

- All the assumptions for a linear regression are met (Homogeneity of variance, Independence of observations, Normality, linear relationship)
- The coefficient of determination is **0.7528328**
- H_0 is rejected due to low p-value, thus, there is a high significance between miles per gallon and weight of the car

Exercise 76

- Use the dataset incomehappy.txt and apply a simple linear regression model to estimate the happiness if the income is 6 (in 1000 Euro per month).
- Are the the assumptions met for linear regression?
- Find the coefficient of determination.
- Is there a significant relationship between the variables? [H_0 - there is no significant relationship]
- Develop a 95% confidence interval of the mean happiness for the income of 6.
- Plot the residual of the simple linear regression model against the independent variable.
- Normal probability plot for the standardized residual.

```
#### Exercise 76 ####
incomehappy<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/in
                stringsAsFactors=F)


head(incomehappy)
```

```
##   ID   income happiness
## 1  1 3.862647  2.314489
## 2  2 4.979381  3.433490
## 3  3 4.923957  4.599373
## 4  4 3.214372  2.791114
## 5  5 7.196409  5.596398
## 6  6 3.729643  2.458556
```

```
# linear model
lm(happiness ~ income, data=incomehappy)
```

```
##
## Call:
## lm(formula = happiness ~ income, data = incomehappy)
##
## Coefficients:
## (Intercept)       income
##      0.2043       0.7138
```

```
happiness.lm = lm(happiness ~ income, data=incomehappy)

summary(happiness.lm)
```

```
##
## Call:
## lm(formula = happiness ~ income, data = incomehappy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02479 -0.48526  0.04078  0.45898  2.37805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.20427    0.08884   2.299   0.0219 *
## income       0.71383    0.01854  38.505   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7181 on 496 degrees of freedom
## Multiple R-squared:  0.7493, Adjusted R-squared:  0.7488
## F-statistic:  1483 on 1 and 496 DF,  p-value: < 2.2e-16
```

```
coeffs = coefficients(happiness.lm)
coeffs
```

```
## (Intercept)      income
##   0.2042704   0.7138255
```

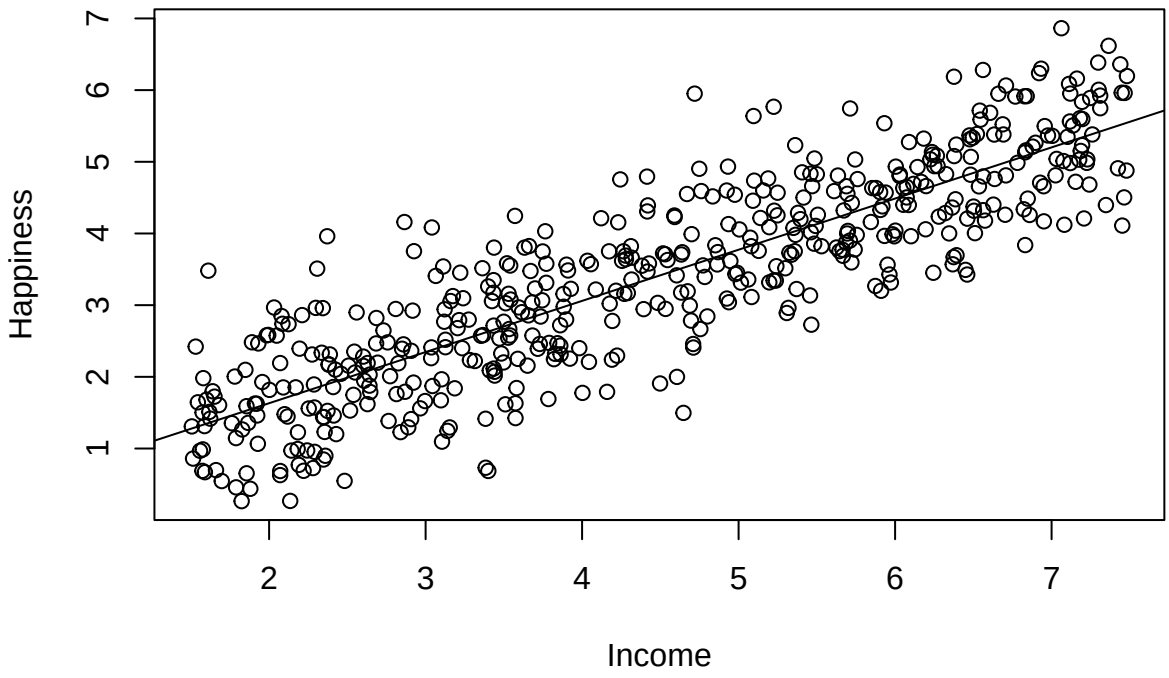
```
# income
income.auto = 6.00
duration = coeffs[1] + coeffs[2]*income.auto

# happiness with respect to income=3
duration
```

```
## (Intercept)
##    4.487223
```

```
paste("Based on the simple linear regression model,
      if the income has been 6.00 (in 1000 Euro),
      we expect to have a happiness index of 4.487223")
```

```
## [1] "Based on the simple linear regression model, \n      if the income has been 6.00 (in 1000 Euro)
```

```
# Plot
plot(incomehappy$income, incomehappy$happiness, xlab="Income", ylab="Happiness")
abline(lm(incomehappy$happiness ~ incomehappy$income))
```




```
# Coefficient determination
happiness.lm = lm(happiness ~ income, data=incomehappy)
summary(happiness.lm)$r.squared
```

```
## [1] 0.7493218
```

```
paste("The results suggests that 74% of the dependent variable is predicted by the independent variable
```

```
## [1] "The results suggests that 74% of the dependent variable is predicted by the independent variable
```

```r
# Significant relationship between variables
happiness.lm = lm(happiness ~ income, data=incomehappy)
summary(happiness.lm)
```

```
##
## Call:
## lm(formula = happiness ~ income, data = incomehappy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02479 -0.48526  0.04078  0.45898  2.37805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.20427    0.08884   2.299   0.0219 *
## income       0.71383    0.01854  38.505   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7181 on 496 degrees of freedom
## Multiple R-squared:  0.7493, Adjusted R-squared:  0.7488
## F-statistic:  1483 on 1 and 496 DF,  p-value: < 2.2e-16
```

```r
paste("As the p-value is very very low,
      which is much less than 0.05,
      we reject the null hypothesis that beta = 0.")
```

```
## [1] "As the p-value is very very low, \n      which is much less than 0.05, \n      we reject the nul
```

```r
# Confidence Interval for income = 6
happiness.lm = lm(happiness ~ income, data=incomehappy)
newdata=data.frame(income=6.00)
predict(happiness.lm, newdata, interval="confidence")
```

```
##        fit     lwr      upr
## 1 4.487223 4.40287 4.571577
```

```r
paste("The 95% confidence interval of the mean happiness index for the income of 6.00,
      is between 4.40287 and 4.571577.")
```

```
## [1] "The 95% confidence interval of the mean happiness index for the income of 6.00,\n      is betwe
```

```r
# Residual Plot
happiness.lm = lm(happiness ~ income, data=incomehappy)
happiness.res=resid(happiness.lm)
plot(incomehappy$income, happiness.res, ylab="Residuals", xlab="Weight",
     main="Happiness by Income", col="blue")
abline(0, 0)
```


```
# Normal Probability Plot of Residuals
happiness.lm = lm(happiness ~ income, data=incomehappy)
happiness.stdres = rstandard(happiness.lm)
qqnorm(happiness.stdres, ylab="Standardized Residuals", xlab="Normal Scores", main="incomehappy data",
qqline(happiness.stdres)
```

Thus, we can conclude that:

- All the assumptions for a linear regression are met (Homogeneity of variance, Independence of observations, Normality, linear relationship)
- The coefficient of determination is **0.7493218**
- H_0 is rejected due to low p-value, thus, there is a high significance between happiness index and income [of course it is]

Exercise 79

- Find the Pearson correlation coefficient of body weight and body height in the data set students.
- Is there any linear relationship between the variables? [H_0 - there is no linear relationship between the variables]
- Test for significance of the correlation. [H_0 - the variables correlation coefficient is 0]

```
#### Exercise 79 ####
students<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/stude
                     stringsAsFactors=F)


head(students)
```

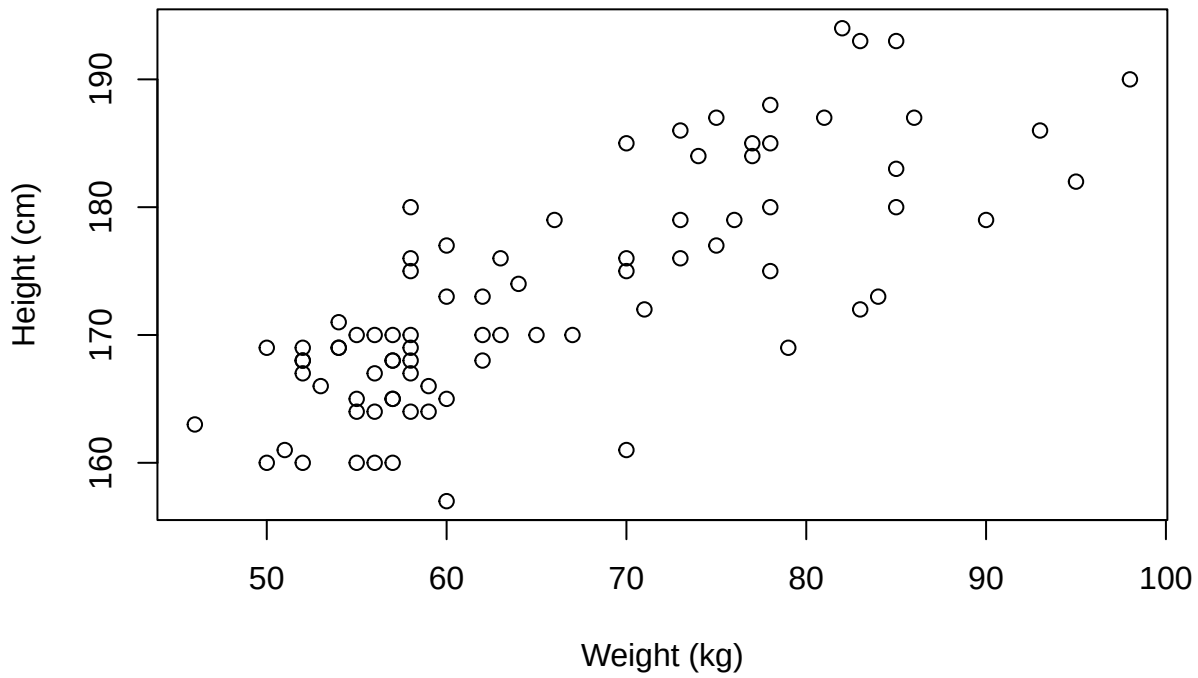
```
##   ID Sex Sex_coded Blood_group Blood_group_coded Rhesus_factor
## 1 24   M         0           0                 0             +
## 2  5   M         0           0                 0             +
```

```
## 3 54   F         1           A                 1            +
## 4  9   M         0           0                 0            +
## 5 34   F         1           A                 1            +
## 6 52   F         1           0                 0            +
##   Rhesus_factor_coded Smoking Smoking_coded Size_cm Weight_kg Points_exam Grade
## 1                   1      no             0     190        98           1     5
## 2                   1      no             0     187        81           2     5
## 3                   1      no             0     171        54           2     5
## 4                   1      no             0     185        70           3     5
## 5                   1      no             0     166        53           3     5
## 6                   1     yes             1     164        55           3     5
```

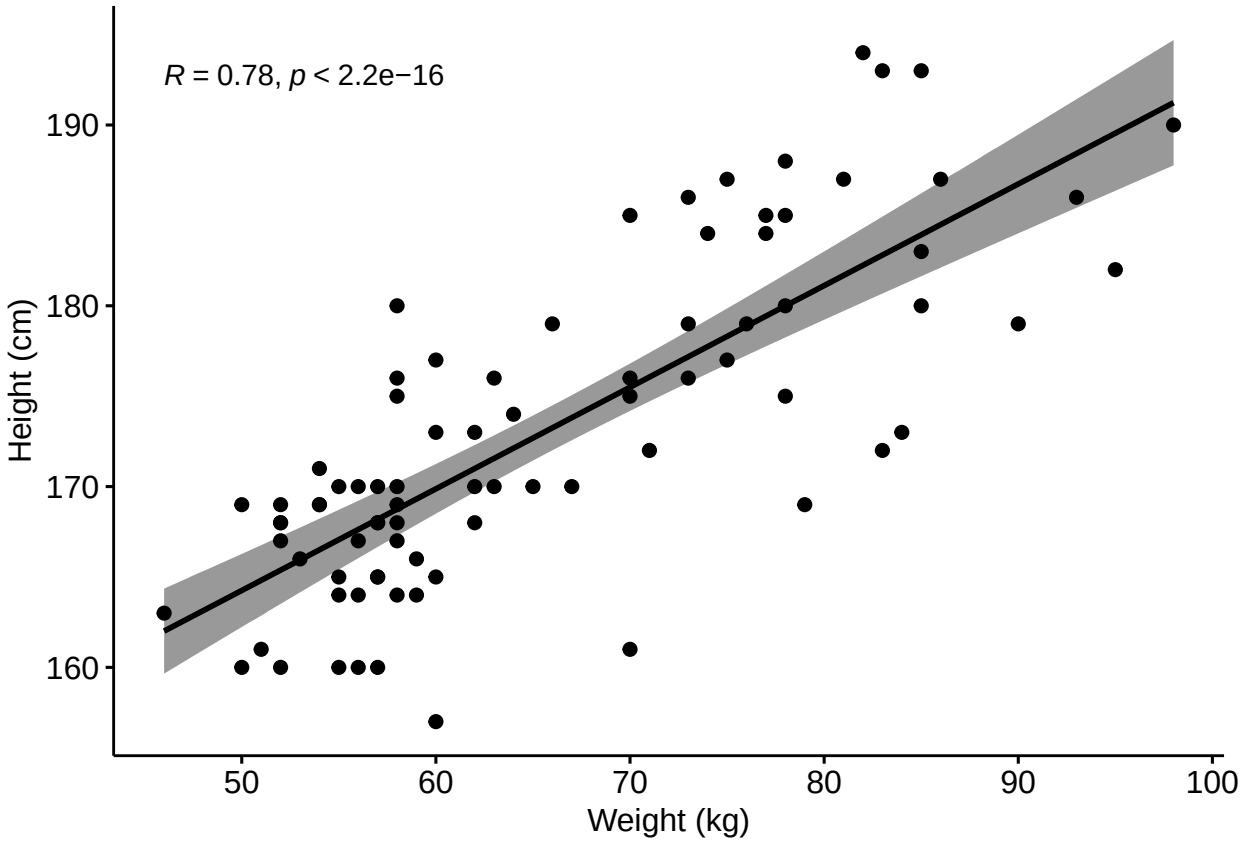
```
# Computation of the correlation coefficient
cor(students$Weight_kg, students$Size_cm)
```

```
## [1] 0.7790491
```

```
# Simple plot + Scatter plot
plot(students$Weight_kg, students$Size_cm, xlab="Weight (kg)", ylab="Height (cm)")
```




```
ggscatter(students, x = "Weight_kg", y = "Size_cm",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Weight (kg)", ylab = "Height (cm)")
```




```
# Shapiro-Wilk normality tests
shapiro.test(students$Weight_kg)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  students$Weight_kg
## W = 0.91953, p-value = 7.405e-05
```

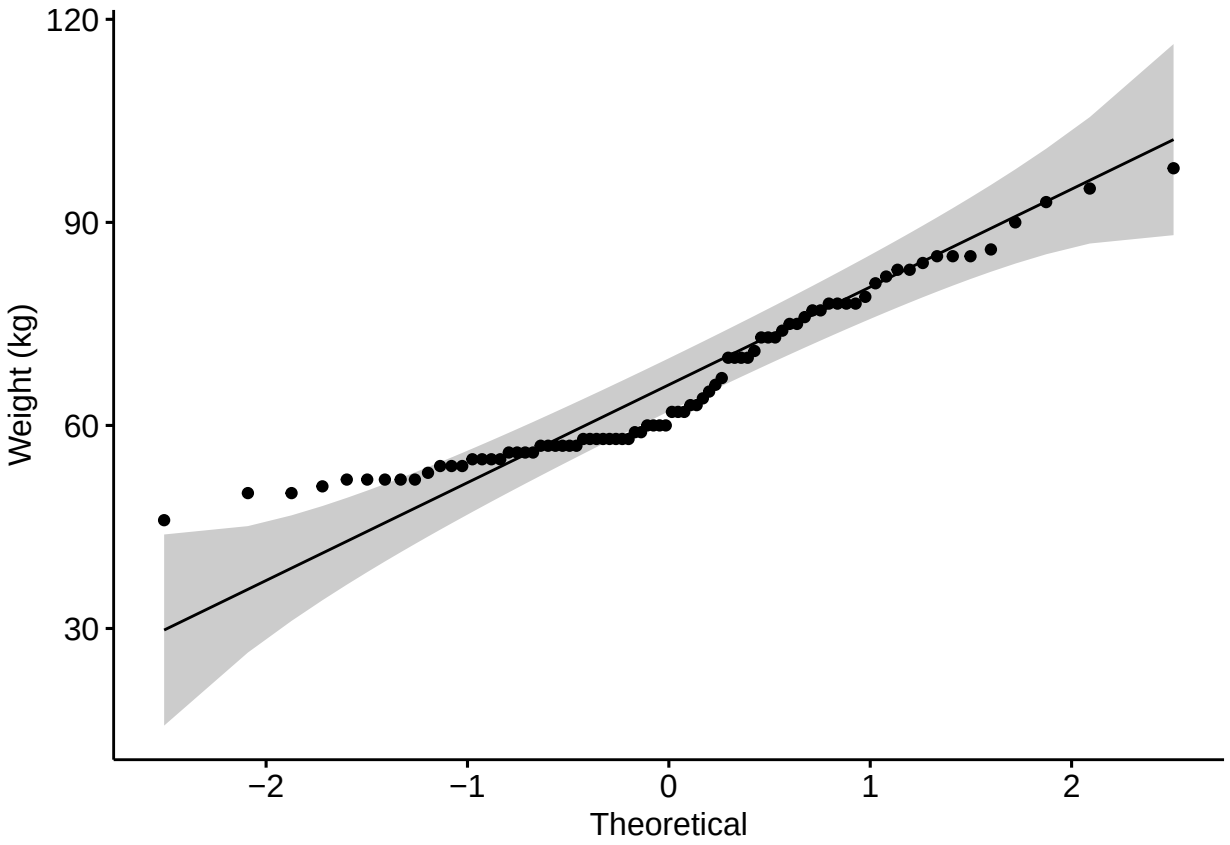
```
shapiro.test(students$Size_cm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  students$Size_cm
## W = 0.9582, p-value = 0.009213
```

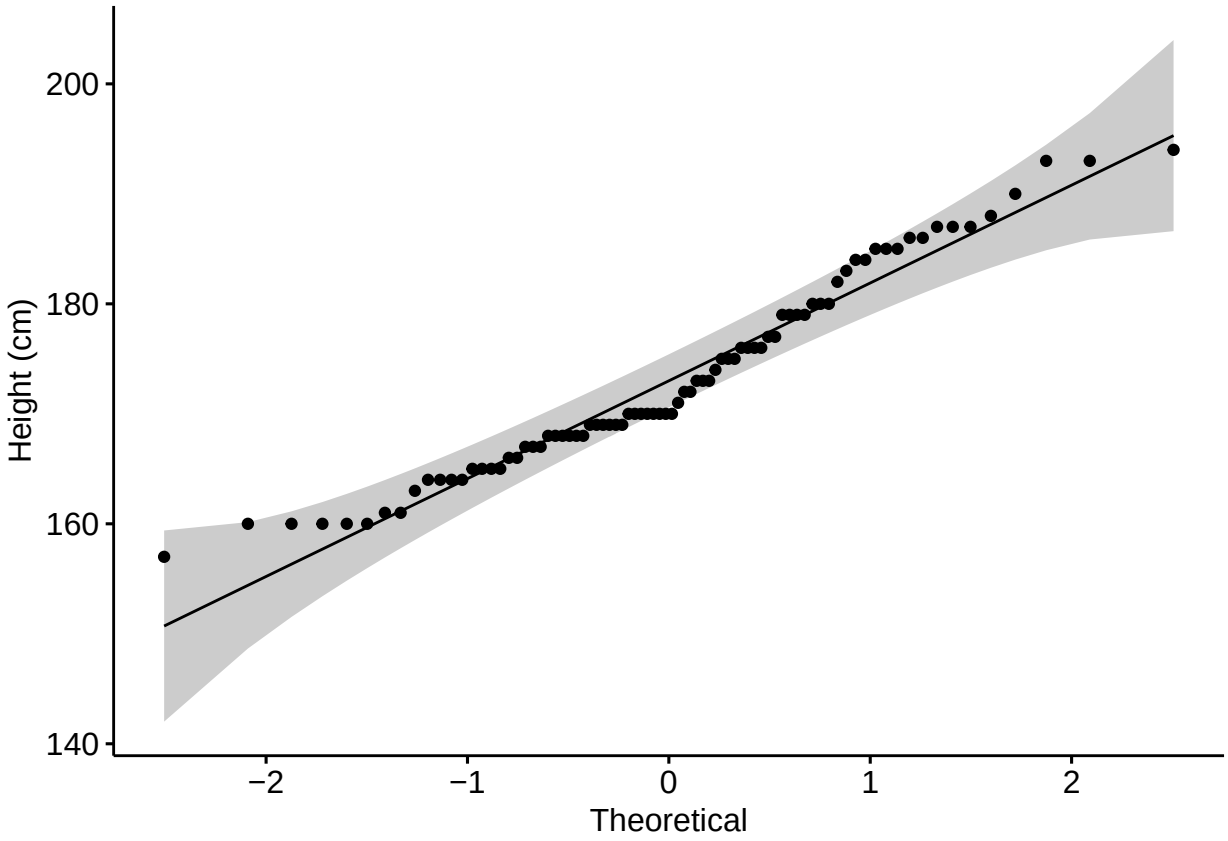
```
paste("Both p-values < 0.05, but, however, our sample is 82, so data is normally distributed (CLT)")
```

```
## [1] "Both p-values < 0.05, but, however, our sample is 82, so data is normally distributed (CLT)"
```

```
# QQ Plots of the variables
ggqqplot(students$Weight_kg, ylab = "Weight (kg)")
```




```
ggqqplot(students$Size_cm, ylab = "Height (cm)")
```




```
# Significance level (p-value) of the correlation
cor.test(students$Weight_kg, students$Size_cm,
         method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  students$Weight_kg and students$Size_cm
## t = 11.114, df = 80, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6763923 0.8520152
## sample estimates:
##       cor
## 0.7790491
```

Thus, we can conclude that:

- P-values for Shapiro-Wilk normality tests are both less than 0.05, but since we have 82 observations, we accept the significance due to the Central Limit Theorem
- There is a linear relationship between the variables because the scatter plot does not show a curved pattern.
- The test for significance is rejected, since p-value is less than 0.05, thus, we have a highly positive correlation of **0.7790491**

Exercise 80

- Find the Pearson correlation coefficient of negative mood and positive mood in the data set ICM.
- Is there any linear relationship between the variables?
- Test for significance of the

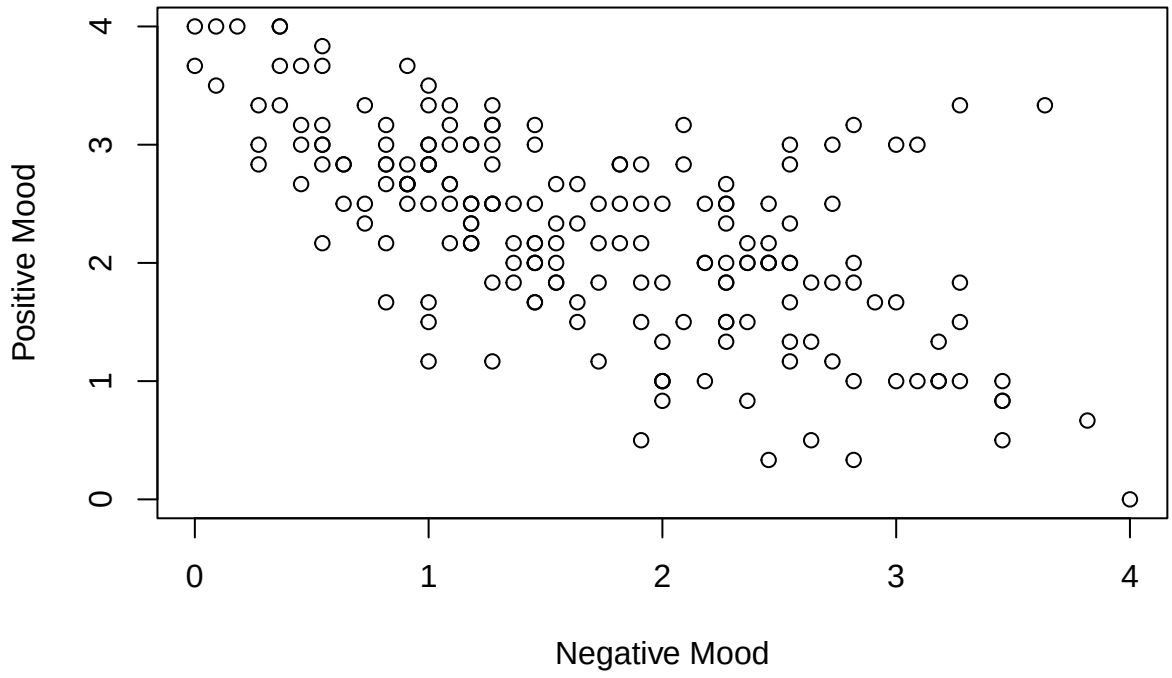
```
#### Exercise 80 ####
ICM<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/ICM.txt",
                stringsAsFactors=F)


head(ICM)
```

```
##   ï..ID Gender Age Englishfluent Germanfluent       Transport
## 1    75 female  22           yes           no PublicTransport
## 2    90 female  22           yes           no PublicTransport
## 3   173 female  37           yes          yes             Car
## 4   189 female  17           yes          yes             Car
## 5   100 female  19           yes          yes            Walk
## 6   155 female  16           yes           no            Walk
##   Highest_level_of_education Do_you_smoke Socialmediahours Timewithfriends Pet
## 1                    College           No     1.5-3hrs/day     2-5hrs/week  No
## 2                    College           No     1.5-3hrs/day     2-5hrs/week  No
## 3                 University           No      <1.5hrs/day    5-10hrs/week Yes
## 4                       none           No     1.5-3hrs/day   10-20hrs/week Yes
## 5                 HighSchool           No       3-5hrs/day     >20hrs/week  No
## 6                       none           No     1.5-3hrs/day   10-20hrs/week  No
##   Siblings Children Relationshipstatus Activitieshours NegativeMood
## 1      Yes       No       Relationship              10           NA
## 2      Yes       No       Relationship              10           NA
## 3       No      Yes       Relationship              20           NA
## 4      Yes       No             Single              40     4.000000
## 5      Yes       No             Single              20     2.818182
## 6      Yes       No             Single              10     2.454545
##   PositiveMood Mentalhealth Socialization Activity SocialSupport
## 1           NA    2.6666667            NA      2.8     4.0000000
## 2           NA    2.6666667            NA      2.8     4.0000000
## 3           NA    3.5000000            NA      3.4     2.3333333
## 4    0.0000000    1.0000000           1.0      3.2     0.6666667
## 5    0.3333333    0.8333333           2.5      1.2     2.3333333
## 6    0.3333333    1.6666667           2.5      2.6     1.3333333
##   Communication_open_direct      OHS
## 1                        NA 4.586207
## 2                        NA 4.586207
## 3                  3.384615 5.103448
## 4                  3.615385 3.137931
## 5                  3.153846 2.758621
## 6                  3.461538 3.586207
```

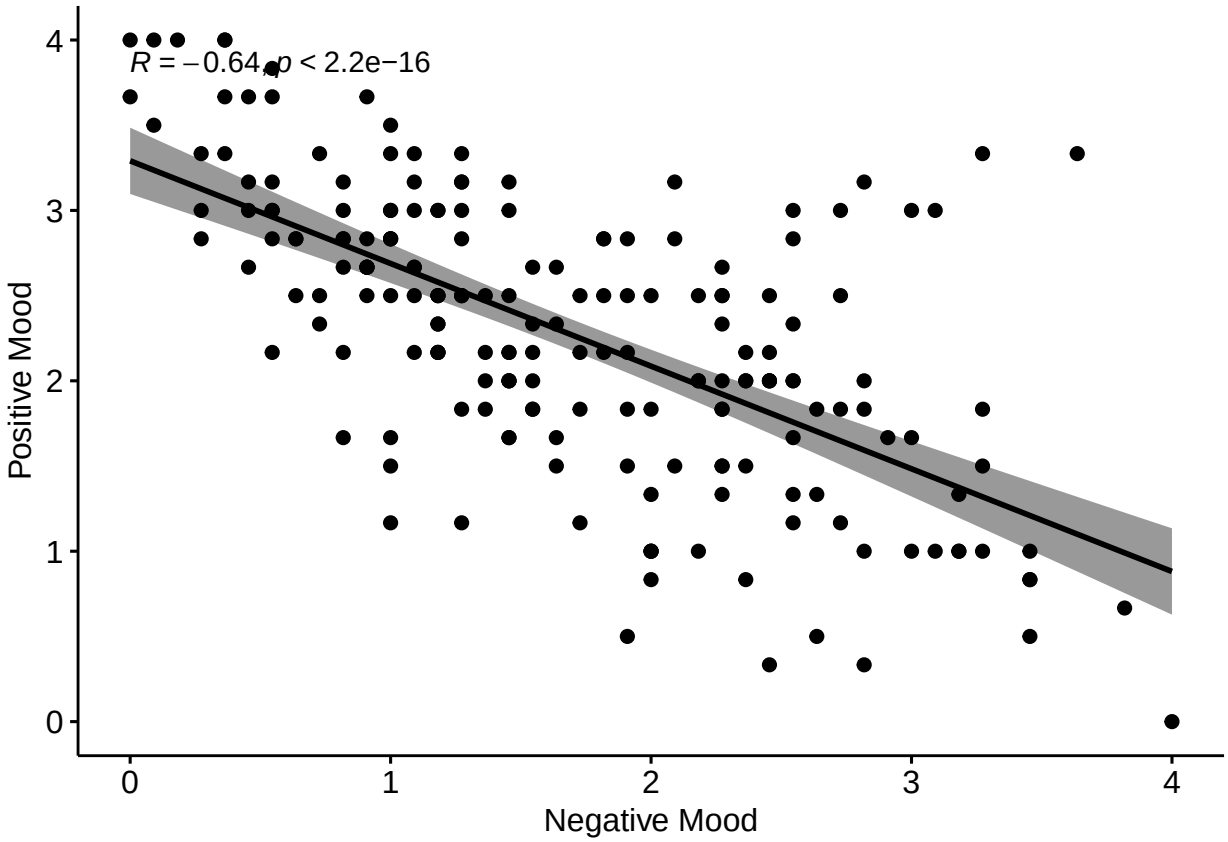
```
# Computation of the correlation coefficient
cor(ICM$NegativeMood, ICM$PositiveMood, use="complete.obs")
```

```
## [1] -0.6433565
```

```r
# Simple plot + Scatter plot
plot(ICM$NegativeMood, ICM$PositiveMood, xlab="Negative Mood", ylab="Positive Mood")
```




```r
ggscatter(ICM, x = "NegativeMood", y = "PositiveMood",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Negative Mood", ylab = "Positive Mood")
```




```
# Shapiro-Wilk normality tests
shapiro.test(ICM$NegativeMood)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ICM$NegativeMood
## W = 0.97664, p-value = 0.002498
```

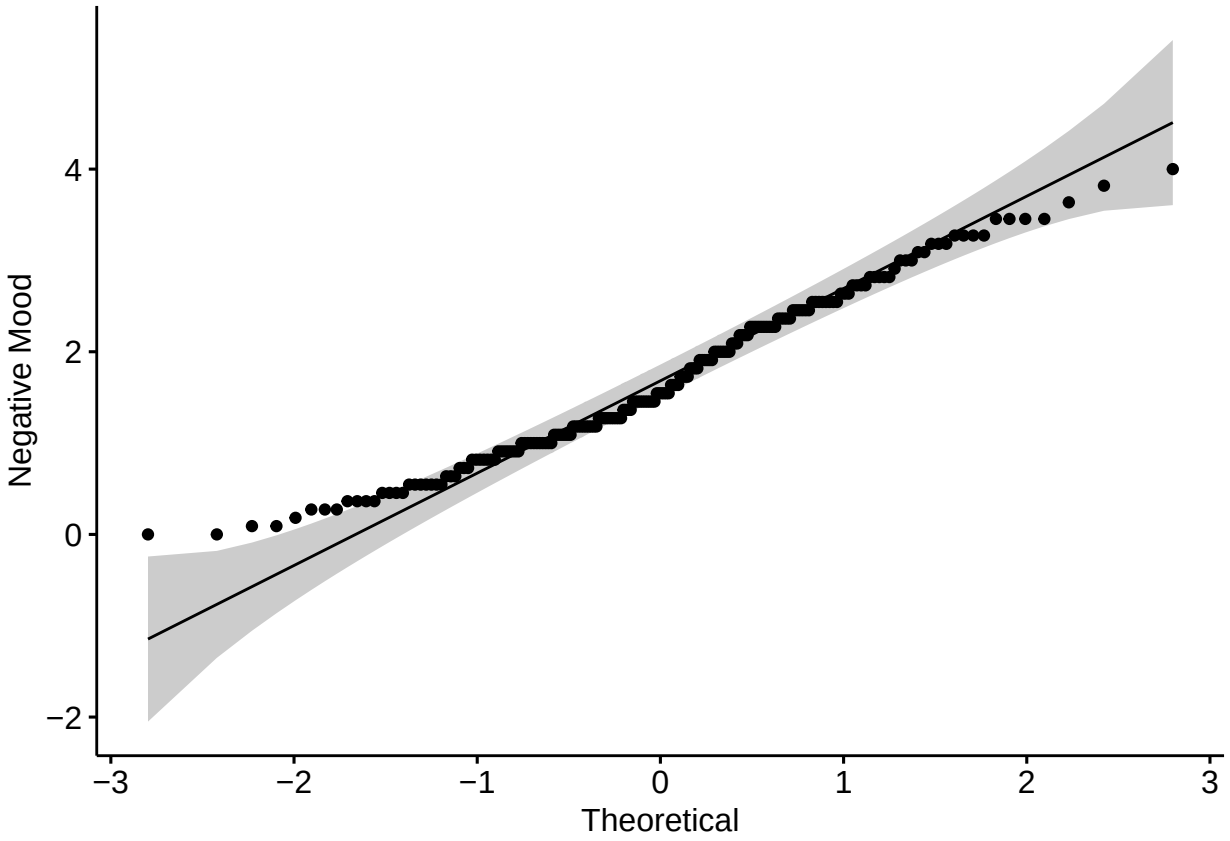
```
shapiro.test(ICM$PositiveMood)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ICM$PositiveMood
## W = 0.9851, p-value = 0.03631
```

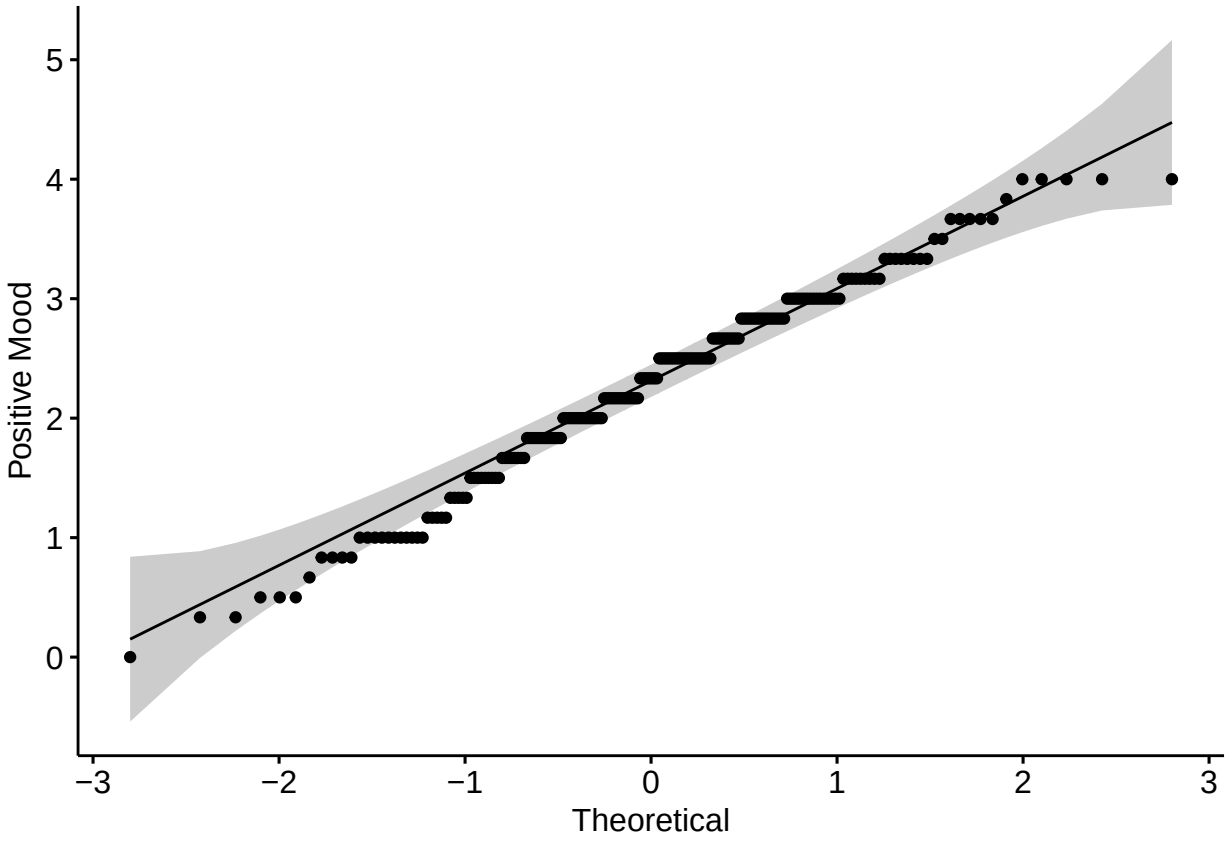
```
paste("Both p-values < 0.05, but, however, our sample is 199, so data is normally distributed (CLT)")
```

```
## [1] "Both p-values < 0.05, but, however, our sample is 199, so data is normally distributed (CLT)"
```

```
# QQ Plots of the variables
ggqqplot(ICM$NegativeMood, ylab = "Negative Mood")
```




```
ggqqplot(ICM$PositiveMood, ylab = "Positive Mood")
```




```
# Significance level (p-value) of the correlation
cor.test(ICM$NegativeMood, ICM$PositiveMood,
         method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  ICM$NegativeMood and ICM$PositiveMood
## t = -11.644, df = 192, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7190609 -0.5525618
## sample estimates:
##        cor
## -0.6433565
```

Thus, we can conclude that:

- P-values for Shapiro-Wilk normality tests are both less than 0.05, but since we have 199 observations, we accept the significance due to the Central Limit Theorem
- There is a linear relationship between the variables because the scatter plot does not show a curved pattern.
- The test for significance is rejected, since p-value is less than 0.05, thus, we have a highly negative correlation of **-0.6433565**

Exercise 83

- Calculate Spearman's rho as correlation coefficient for the variables body weight and body height in the data set students.
- Test for significance of the correlation.

```
#### Exercise 83 ####
students<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/studer
                     stringsAsFactors=F)

# Significance level (p-value) of the correlation
cor.test(students$Weight_kg, students$Size_cm,
         method = "spearman", exact=FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  students$Weight_kg and students$Size_cm
## S = 20764, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.7740172
```

Thus, we can conclude that:

- The rho correlation coefficient between weight and height is 0.7740172 and the p-value is lower than 0.05
- There is a statistically highly significant positive correlation between weight and height.

Exercise 84

- Calculate Spearman's rho as correlation coefficient for the variables negative mood and OHS in the data set ICM.
- Is there any linear relationship between the variables?
- Test for significance of the correlation.

```
#### Exercise 84 ####

# Significance level (p-value) of the correlation
ICM<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/ICM.txt",
                stringsAsFactors=F)


head(ICM)
```

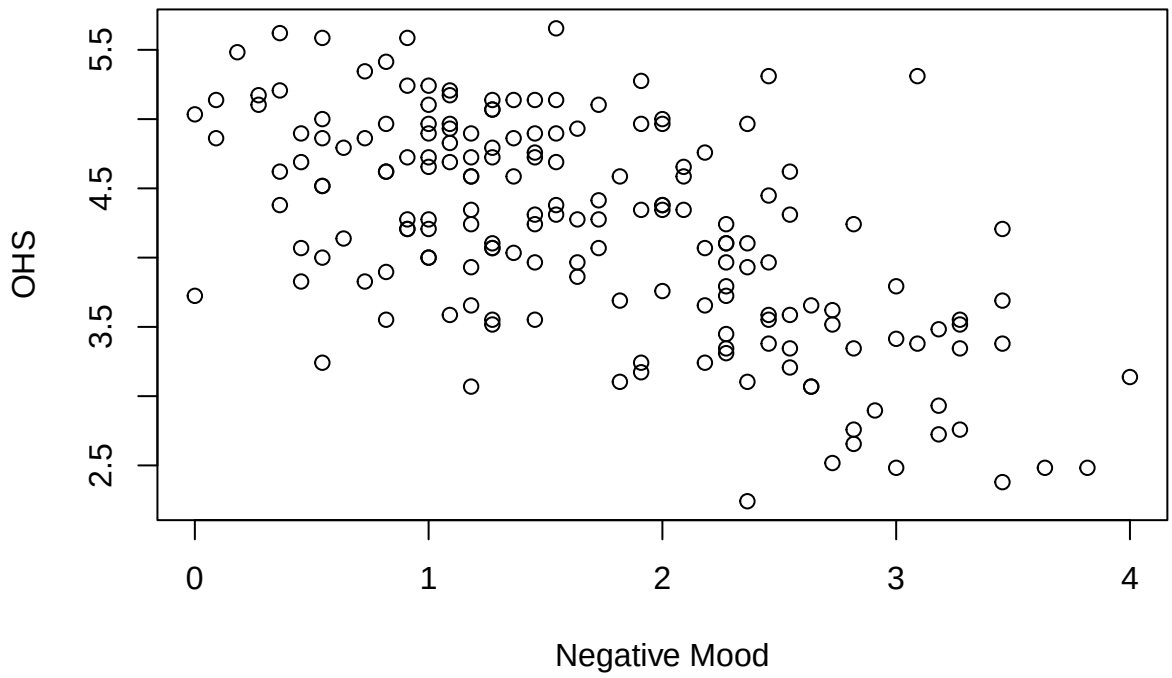
```
##   ï..ID Gender Age Englishfluent Germanfluent       Transport
## 1    75 female  22           yes           no PublicTransport
## 2    90 female  22           yes           no PublicTransport
## 3   173 female  37           yes          yes             Car
## 4   189 female  17           yes          yes             Car
```

```
## 5   100 female  19           yes          yes            Walk
## 6   155 female  16           yes           no            Walk
##   Highest_level_of_education Do_you_smoke Socialmediahours Timewithfriends Pet
## 1                    College           No     1.5-3hrs/day     2-5hrs/week  No
## 2                    College           No     1.5-3hrs/day     2-5hrs/week  No
## 3                 University           No      <1.5hrs/day    5-10hrs/week Yes
## 4                       none           No     1.5-3hrs/day   10-20hrs/week Yes
## 5                 HighSchool           No       3-5hrs/day     >20hrs/week  No
## 6                       none           No     1.5-3hrs/day   10-20hrs/week  No
##   Siblings Children Relationshipstatus Activitieshours NegativeMood
## 1      Yes       No       Relationship              10           NA
## 2      Yes       No       Relationship              10           NA
## 3       No      Yes       Relationship              20           NA
## 4      Yes       No             Single              40     4.000000
## 5      Yes       No             Single              20     2.818182
## 6      Yes       No             Single              10     2.454545
##   PositiveMood Mentalhealth Socialization Activity SocialSupport
## 1           NA    2.6666667            NA      2.8     4.0000000
## 2           NA    2.6666667            NA      2.8     4.0000000
## 3           NA    3.5000000            NA      3.4     2.3333333
## 4    0.0000000    1.0000000           1.0      3.2     0.6666667
## 5    0.3333333    0.8333333           2.5      1.2     2.3333333
## 6    0.3333333    1.6666667           2.5      2.6     1.3333333
##   Communication_open_direct      OHS
## 1                        NA 4.586207
## 2                        NA 4.586207
## 3                  3.384615 5.103448
## 4                  3.615385 3.137931
## 5                  3.153846 2.758621
## 6                  3.461538 3.586207
```

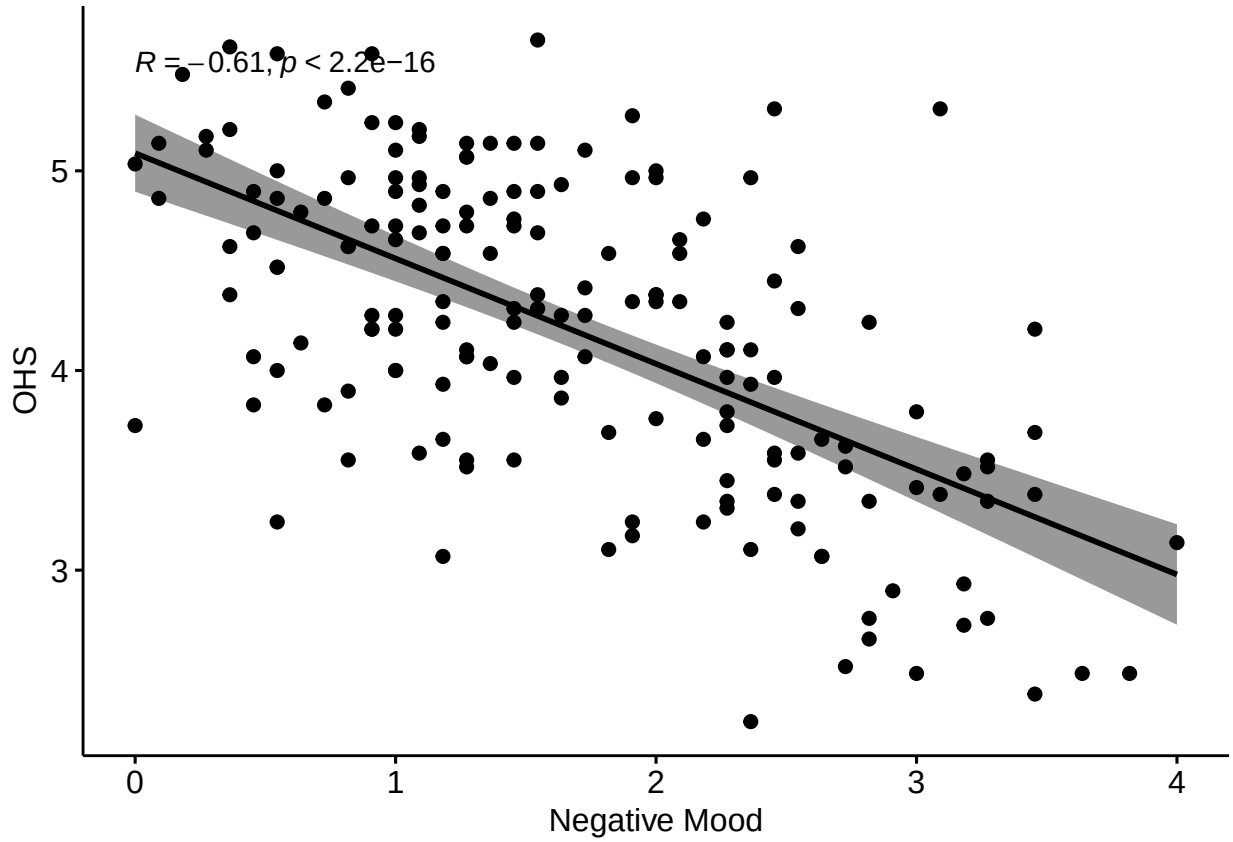
```
# Computation of the correlation coefficient
cor(ICM$NegativeMood, ICM$OHS, use="complete.obs")
```

```
## [1] -0.6140032
```

```
# Simple plot + Scatter plot
plot(ICM$NegativeMood, ICM$OHS, xlab="Negative Mood", ylab="OHS")
```




```
ggscatter(ICM, x = "NegativeMood", y = "OHS",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Negative Mood", ylab = "OHS")
```




```
# Shapiro-Wilk normality tests
shapiro.test(ICM$NegativeMood)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ICM$NegativeMood
## W = 0.97664, p-value = 0.002498
```

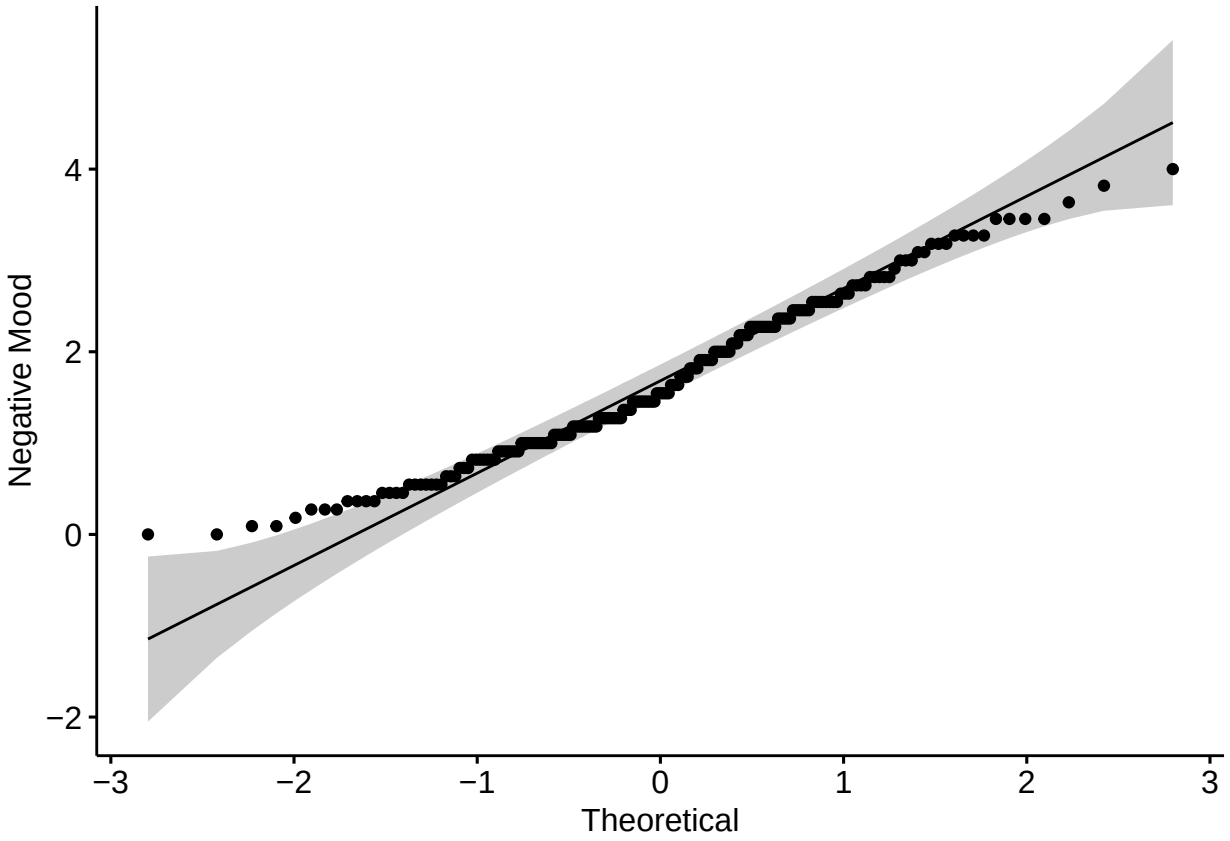
```
shapiro.test(ICM$OHS)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ICM$OHS
## W = 0.97477, p-value = 0.002283
```

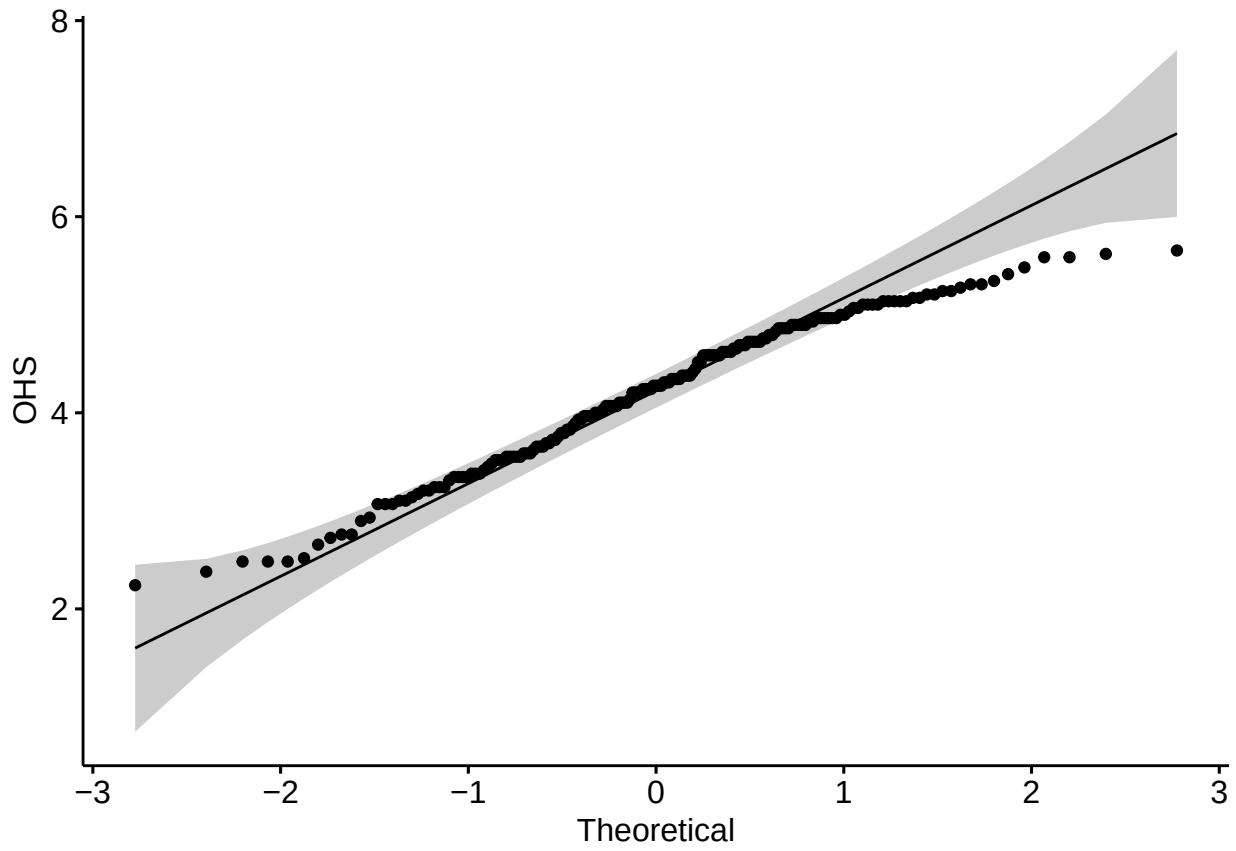
```
paste("Both p-values < 0.05, but, however, our sample is 199, so data is normally distributed (CLT)")
```

```
## [1] "Both p-values < 0.05, but, however, our sample is 199, so data is normally distributed (CLT)"
```

```
# QQ Plots of the variables
ggqqplot(ICM$NegativeMood, ylab = "Negative Mood")
```




```
ggqqplot(ICM$OHS, ylab = "OHS")
```




```
# Significance level (p-value) of the correlation
cor.test(ICM$NegativeMood, ICM$OHS,
         method = "spearman", exact=FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  ICM$NegativeMood and ICM$OHS
## S = 1453320, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.5725575
```

Thus, we can conclude that:

- The rho correlation coefficient between weight and height is 0.7740172 and the p-value is lower than 0.05
- There is a statistically highly significant positive correlation between weight and height.
- There is a linear relationship between the variables because the scatter plot does not show a curved pattern.