Homework 5

Darian-Florian Voda

2022-11-23

Loading packages

```
library(mosaic)
library(tidyverse)
library(hrbrthemes)
library(ggpubr)
library(dplyr)
library(ggplot2)
library(viridis)
library(gridExtra)
library(car)
library(MASS)
```

Exercise 53

- Use the data set 'ICM'.
- Without assuming the data to have normal distribution, decide at .05 significance level if the negative mood of students has identical data distributions depending on the social media use. [H_0: Has identical data distributions]

```
#### Exercise 53 ####

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


head(ICM,5)
```

```
##   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
##   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
##   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
##   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
##   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
```

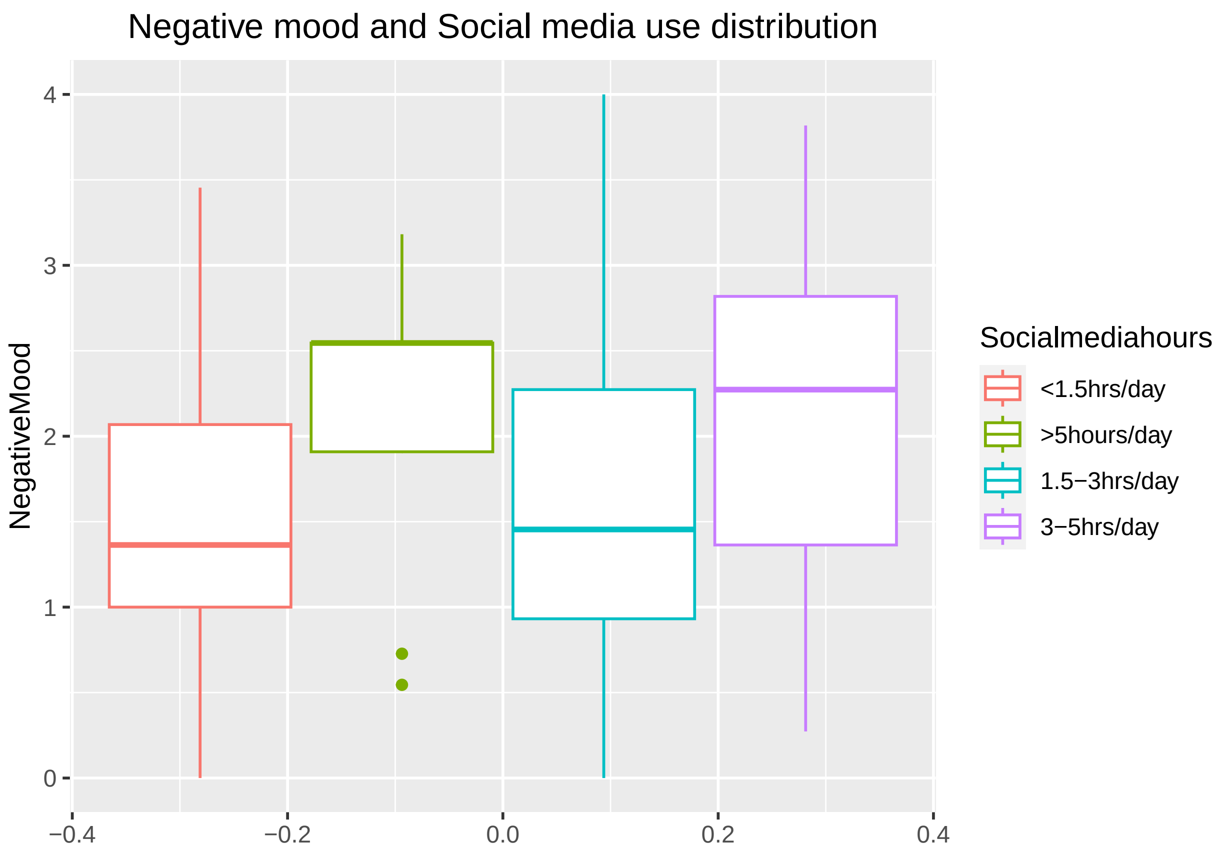
```
krus.res <- kruskal.test(NegativeMood ~ Socialmediahours, data =
                           ICM)
krus.res
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  NegativeMood by Socialmediahours
## Kruskal-Wallis chi-squared = 11.858, df = 3, p-value = 0.007884
```

```
paste("p-value is 0.007 which is less than 0.05, thus we reject H0")
```

```
## [1] "p-value is 0.007 which is less than 0.05, thus we reject H0"
```

```
ggplot(ICM, aes(group=Socialmediahours, y=NegativeMood, color=Socialmediahours)) +
  geom_boxplot()+
  labs(title="Negative mood and Social media use distribution")+
  theme(plot.title = element_text(hjust = 0.5))
```



Thus, we conclude that:

-H_0 is rejected, since p-value is less than 0.05 -The distribution of Negative mood data and Social media use is **not** identical

Exercise 54

- Use the data set 'ICM'.
- Without assuming the data to have normal distribution, decide at .05 significance level if the socialization of students has identical data distributions depending on the time spent with friends. [H_0: has identical data distribution]

```
#### Exercise 54 ####
ICM<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/ICM.txt",
                stringsAsFactors=F)

tail(ICM)
```

```
##     i..ID Gender Age Englishfluent Germanfluent       Transport
## 194   171 female  22           yes           no             Car
## 195   136 female  24           yes           no            Walk
## 196    52 female  18            no           no PublicTransport
## 197   170   male  25           yes           no             Car
## 198    65   male  28            no          yes PublicTransport
## 199    98   male  22           yes          yes PublicTransport
##     Highest_level_of_education Do_you_smoke Socialmediahours Timewithfriends
## 194                    College           No       3-5hrs/day    5-10hrs/week
## 195                 University           No       3-5hrs/day   10-20hrs/week
## 196                       none           No      <1.5hrs/day    5-10hrs/week
## 197                    College          Yes      <1.5hrs/day     2-5hrs/week
## 198                    College           No     1.5-3hrs/day     2-5hrs/week
## 199                 HighSchool           No      <1.5hrs/day    5-10hrs/week
##     Pet Siblings Children Relationshipstatus Activitieshours NegativeMood
## 194  No      Yes       No             Single              20   0.54545455
## 195  No      Yes       No       Relationship              20   0.36363636
## 196 Yes      Yes       No             Single              20   0.18181818
## 197  No      Yes      Yes           Divorced              20   0.09090909
## 198  No      Yes       No       Relationship              20   0.36363636
## 199  No      Yes       No            Married              20   0.00000000
##     PositiveMood Mentalhealth Socialization Activity SocialSupport
## 194     3.833333     3.166667      3.833333      4.0      3.666667
## 195     4.000000     3.666667      3.166667      3.4      3.666667
## 196     4.000000     4.000000      3.500000      3.8      4.000000
## 197     4.000000     4.000000      3.500000      4.0      3.000000
## 198     4.000000     3.666667      4.000000      3.6      3.666667
## 199     4.000000     4.000000      4.000000      4.0      3.666667
##     Communication_open_direct      OHS
## 194                        NA 5.586207
## 195                  4.384615 5.620690
## 196                  3.384615 5.482759
## 197                  4.000000 4.862069
## 198                  2.461538 4.379310
## 199                  4.384615 3.724138
```

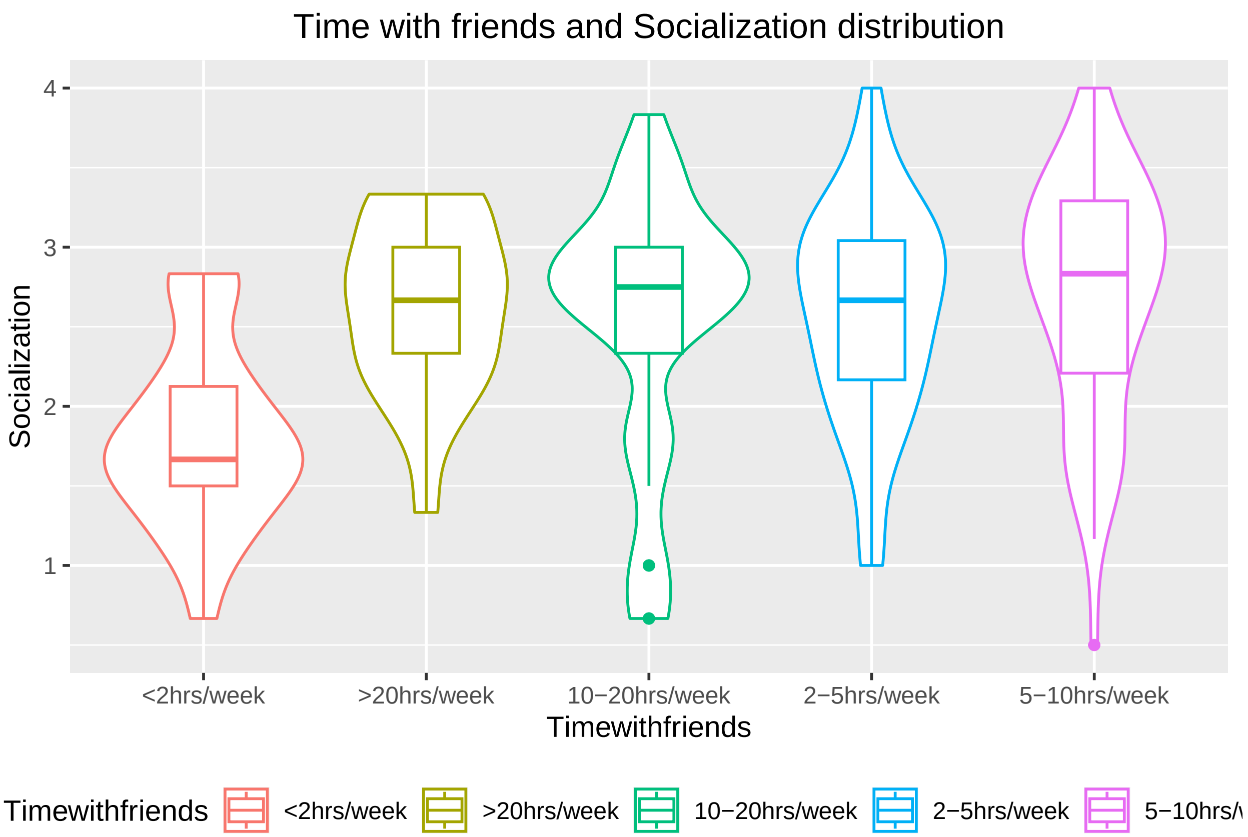
```
krus.res <- kruskal.test(Socialization ~ Timewithfriends, data =
                           ICM)
krus.res
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Socialization by Timewithfriends
## Kruskal-Wallis chi-squared = 28.087, df = 4, p-value = 1.198e-05
```

```
paste("p-value is 0.00001198 which is less than 0.05, thus we reject H0")
```

```
## [1] "p-value is 0.00001198 which is less than 0.05, thus we reject H0"
```

```
ggplot(ICM, aes(x=Timewithfriends, y=Socialization, color=Timewithfriends)) +
  geom_violin(width=0.9)+
  geom_boxplot(width=0.3)+
  scale_fill_viridis(discrete = TRUE, alpha=0.6)+
  labs(title="Time with friends and Socialization distribution")+
  theme(plot.title = element_text(hjust = 0.5), legend.position="bottom")
```



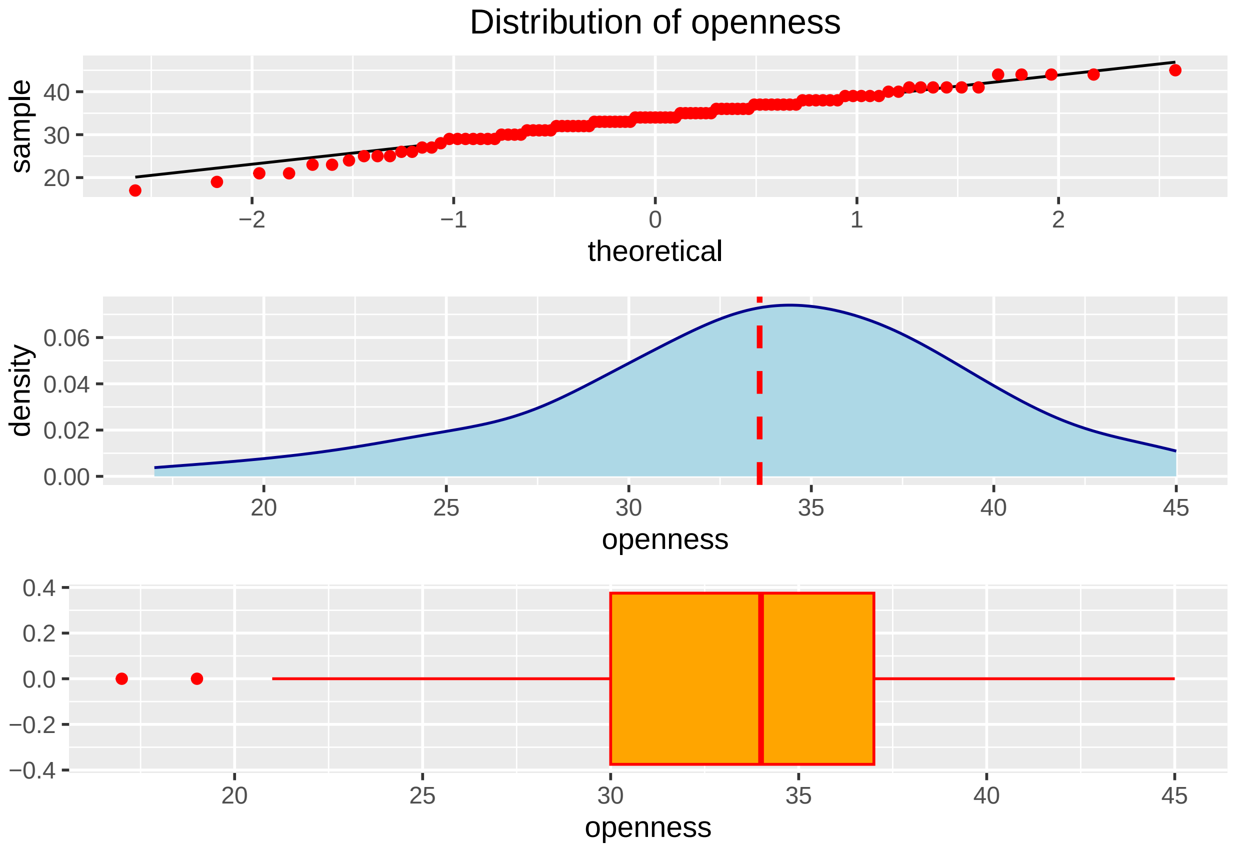
Thus, we can conclude that:

- H_0 is rejected, since p-value is less than 0.05
- The distribution of Time with friends data and Socialization is **not** identical

Exercise 56

- Use the data set "survey PCA".
- Assess the normality of the variable "openness". [H_0 : sample distribution is normal]

```
#### Exercise 56 #####

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

qqplot <- ggplot(survey, aes(sample = openness)) + geom_qq_line() + stat_qq(color="red") +
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Distribution of openness")

densityplot <- ggplot(survey, aes(openness)) + geom_density(color="darkblue", fill="lightblue") +
  geom_vline(aes(xintercept=mean(openness)), color="red", linetype="dashed", size=1)
bxplot <- ggplot(survey, aes(openness)) + geom_boxplot(color="red", fill="orange")
grid.arrange(qqplot, densityplot, bxplot, ncol = 1)
```



```
shapiro.test(survey$openness)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  survey$openness
## W = 0.97794, p-value = 0.08856
```

```
paste("p-value is greater than 0.5, so H0 is accepted")
```

```
## [1] "p-value is greater than 0.5, so H0 is accepted"
```

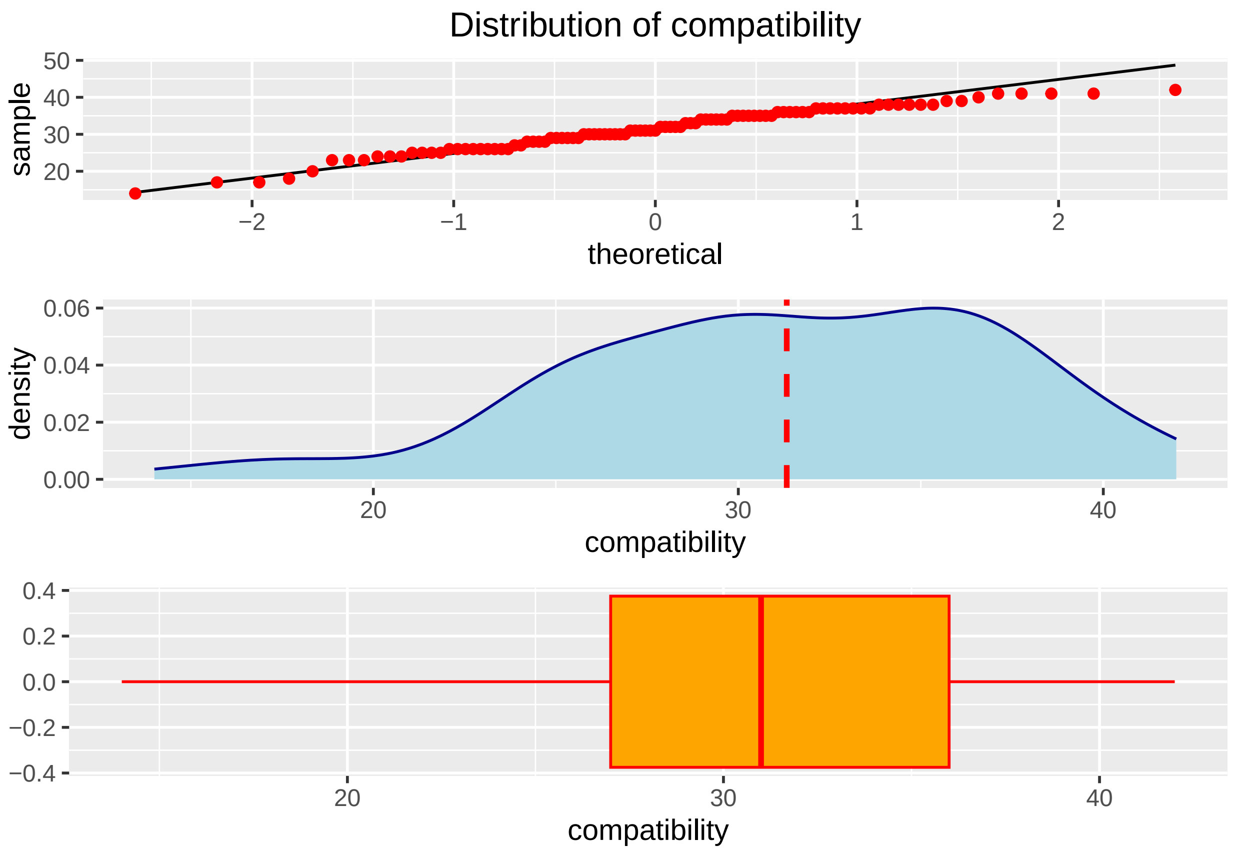
Thus, we can conclude that:

- H_0 is accepted, since p-value is greater than 0.05 confidence interval (0.08856)
- The distribution of openness is a **normal** distribution

Exercise 57

- Use the data set "survey PCA".
- Assess the normality of the variable "compatibility". [H_0 : sample distribution is normal]

```
#### Exercise 57 #####

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

qqplot <- ggplot(survey, aes(sample = compatibility)) + geom_qq_line() + stat_qq(color="red") +
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Distribution of compatibility")

densityplot <- ggplot(survey, aes(compatibility)) + geom_density(color="darkblue", fill="lightblue") +
  geom_vline(aes(xintercept=mean(compatibility)), color="red", linetype="dashed", size=1)
bxplot <- ggplot(survey, aes(compatibility)) + geom_boxplot(color="red", fill="orange")
grid.arrange(qqplot, densityplot, bxplot, ncol = 1)
```



```
shapiro.test(survey$compatibility)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  survey$compatibility
## W = 0.97105, p-value = 0.02543
```

```
paste("p-value is less than 0.5, so H0 is rejected")
```

```
## [1] "p-value is less than 0.5, so H0 is rejected"
```

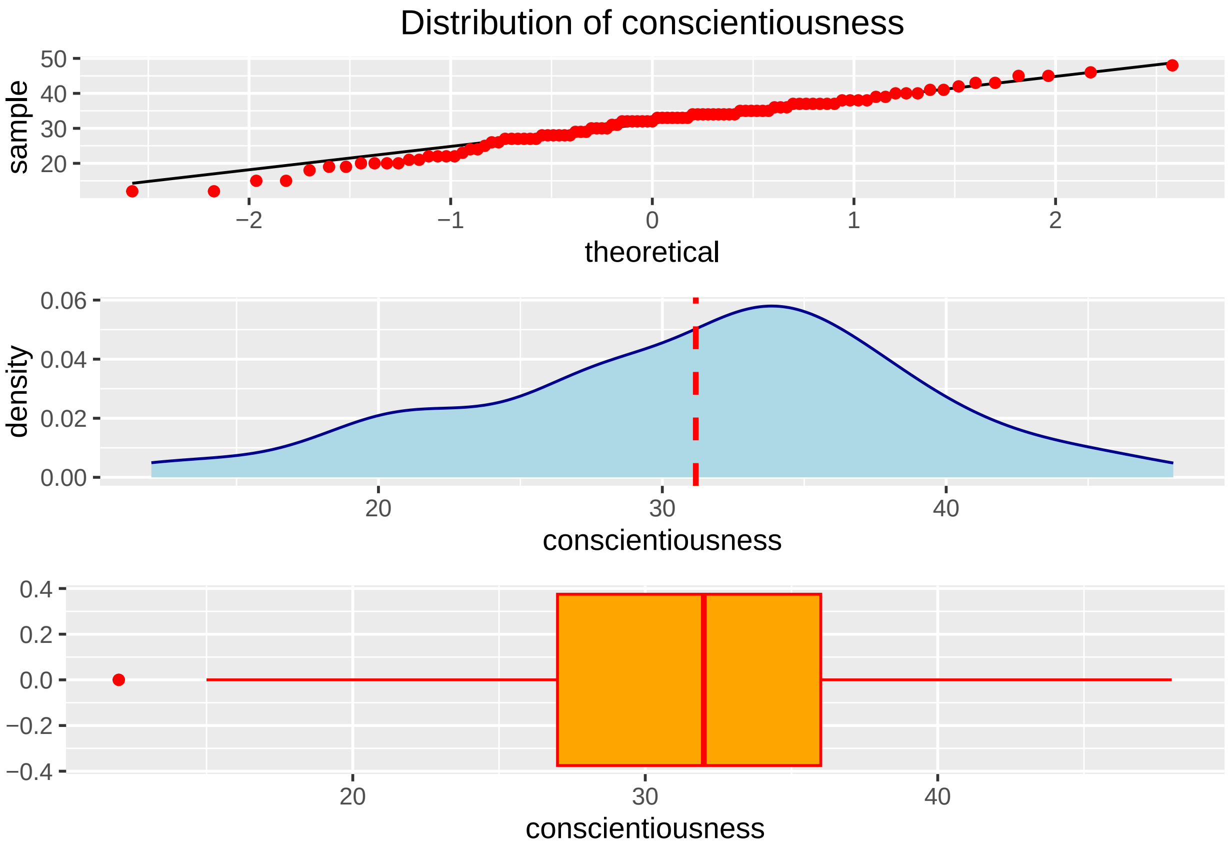
Thus, we can conclude that:

- H_0 is rejected, since p-value is less than 0.05 confidence interval (0.02543)
- The distribution of compatibility is *not* a **normal** distribution

Exercise 58

- Use the data set "survey PCA".
- Assess the normality of the variable "conscientiousness". [H_0 : sample distribution is normal]

```
#### Exercise 58 ####

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

qqplot <- ggplot(survey, aes(sample = conscientiousness)) + geom_qq_line() + stat_qq(color="red") +
  theme(plot.title = element_text(hjust = 0.5))+
  labs(title="Distribution of conscientiousness")

densityplot <- ggplot(survey, aes(conscientiousness)) + geom_density(color="darkblue", fill="lightblue")
  geom_vline(aes(xintercept=mean(conscientiousness)), color="red", linetype="dashed", size=1)
bxplot <- ggplot(survey, aes(conscientiousness)) + geom_boxplot(color="red", fill="orange")
grid.arrange(qqplot, densityplot, bxplot, ncol = 1)
```



```
shapiro.test(survey$conscientiousness)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  survey$conscientiousness
## W = 0.98133, p-value = 0.1638
```

```
paste("p-value is greater than 0.5, so H0 is accepted")
```

```
## [1] "p-value is greater than 0.5, so H0 is accepted"
```

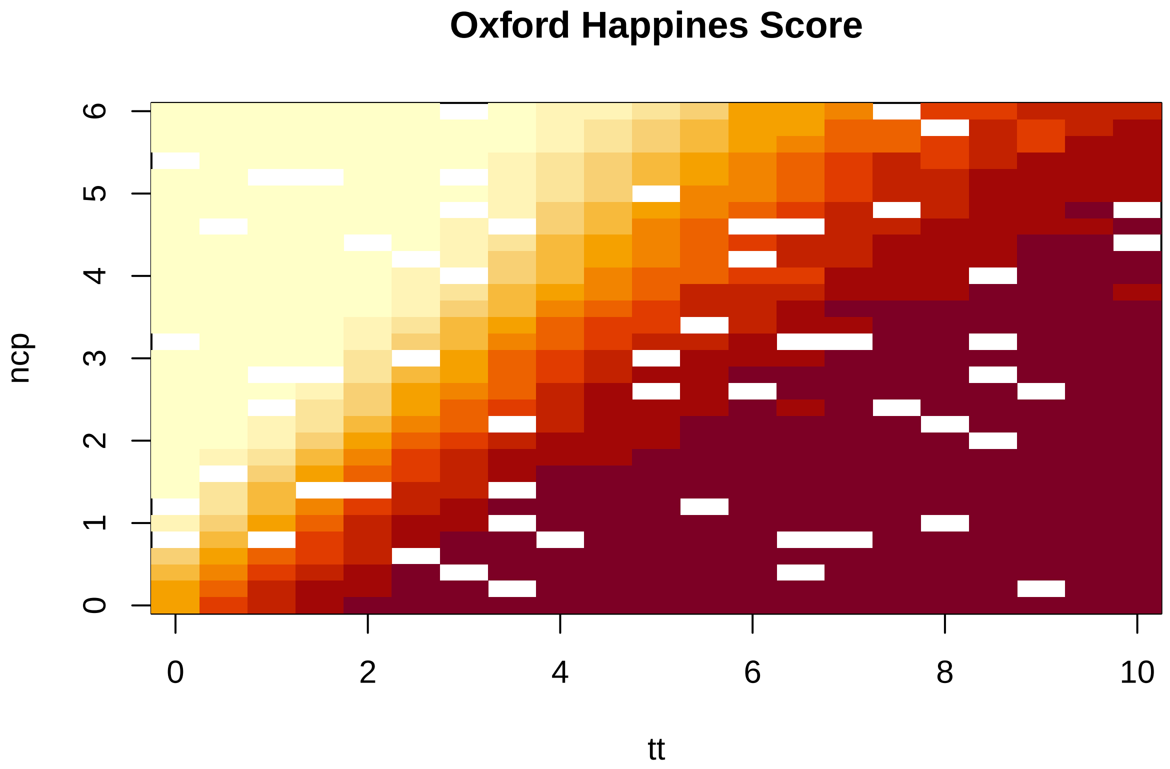
Thus, we can conclude that:

- H_0 is accepted, since p-value is greater than 0.05 confidence interval (0.1638)
- The distribution of conscientiousness is a **normal** distribution

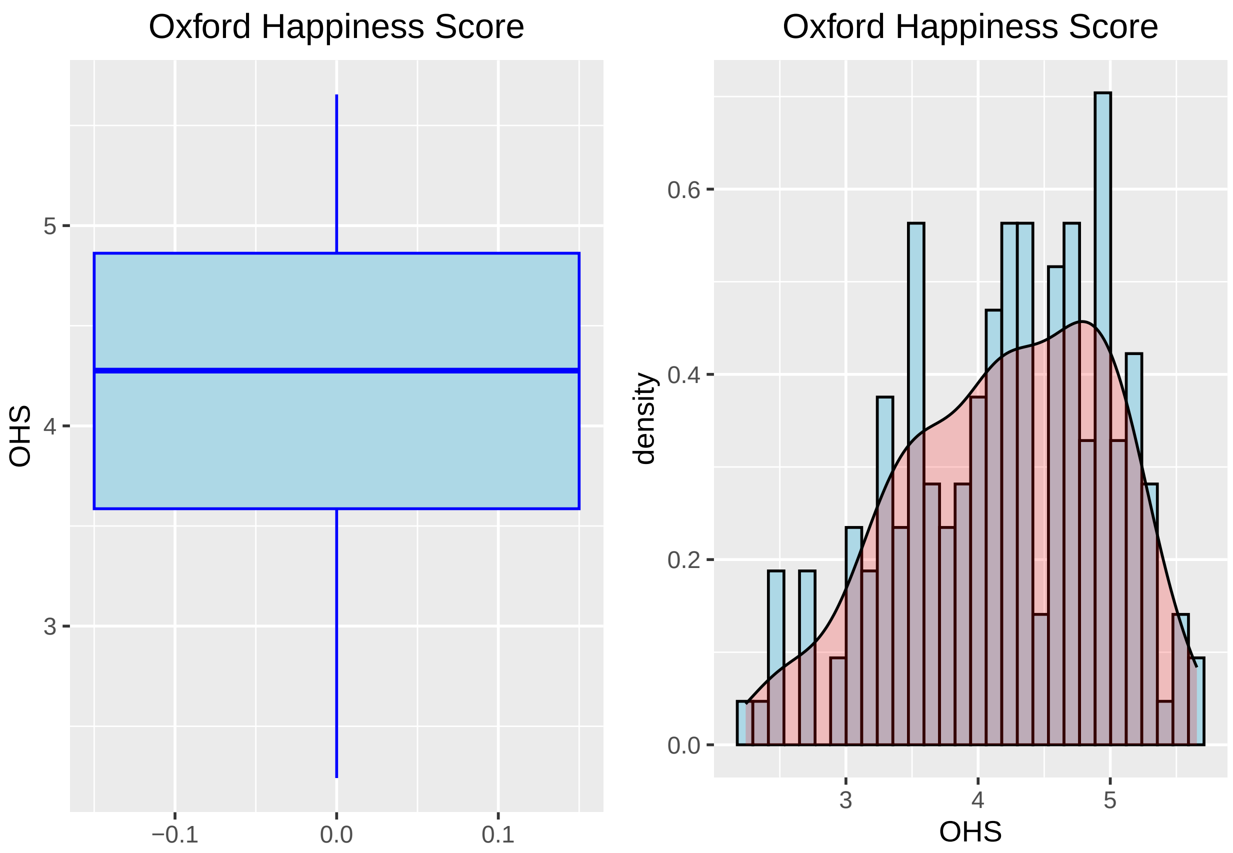
Exercise 60

- Use the data set "ICM".
- Does the OHS (Oxford Happiness Score) of the students differ from the average score of 4? [H_0: OHS is the same with average score of 4]

```
#### Exercise 60 ####

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


box = ggplot(ICM, aes(y=OHS)) +
  geom_boxplot(width=0.3, color="blue", fill="lightblue")+
  scale_fill_viridis(discrete = TRUE, alpha=0.6)+
  labs(title="Oxford Happiness Score")+
  theme(plot.title = element_text(hjust = 0.5), legend.position="bottom")

tt <- seq(0, 10, length.out = 21)
ncp <- seq(0, 6, length.out = 31)
ptn <- outer(tt, ncp, function(t, d) pt(t, df = ICM$OHS, ncp = d))
t.tit <- "Oxford Happines Score"
matrixx = image(tt, ncp, ptn, zlim = c(0,1), main = t.tit)
```



```
hist = ggplot(ICM, aes(x=OHS)) +
  geom_histogram(aes(y=..density..), colour="black", fill="lightblue")+
  geom_density(alpha=.2, fill="red")+
  labs(title="Oxford Happiness Score")+
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(box, hist, ncol=2)
```



```r
res <- t.test(ICM$OHS, mu = 4)
res
```

```
##
##  One Sample t-test
##
## data:  ICM$OHS
## t = 3.5485, df = 180, p-value = 0.0004943
## alternative hypothesis: true mean is not equal to 4
## 95 percent confidence interval:
##  4.090915 4.318687
## sample estimates:
## mean of x
##  4.204801
```

```r
paste("p-value less than 0.05, thus it is rejected")
```

```
## [1] "p-value less than 0.05, thus it is rejected"
```

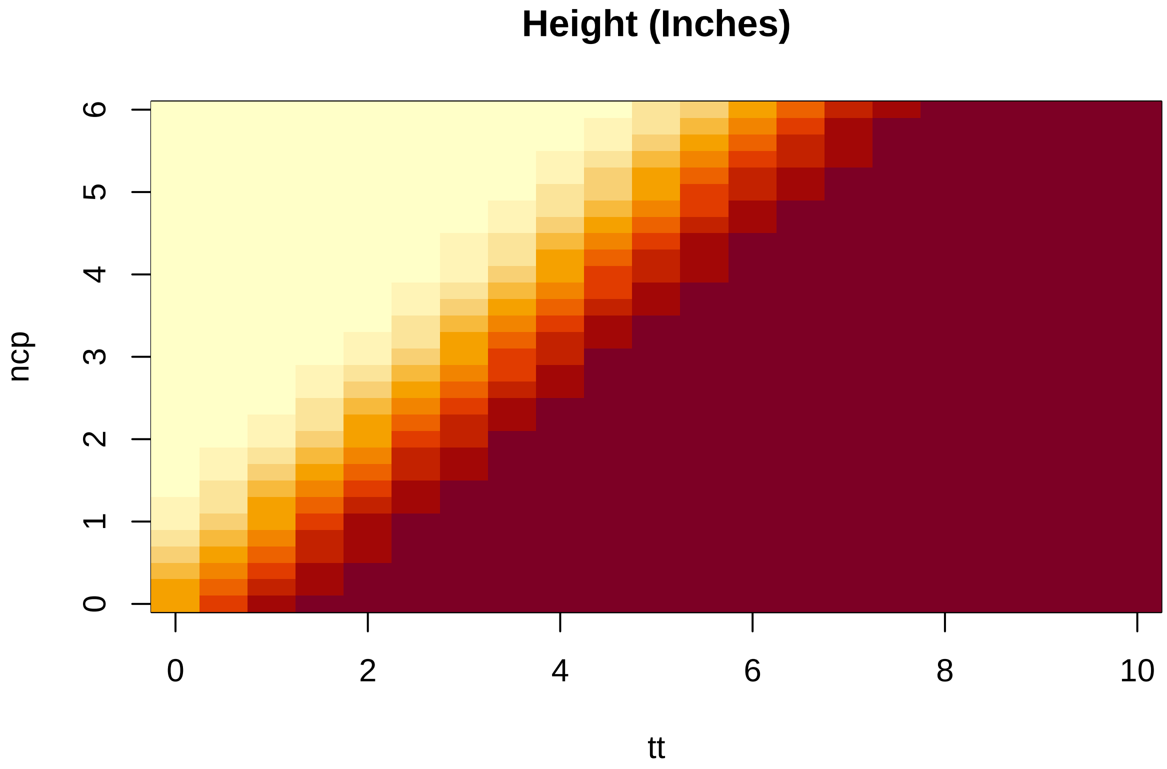
Thus, we can conclude that:

- H_0 is rejected, since p-value is less than 0.05 confidence interval (0.0004)
- The distribution of happiness score is **not** as the average OHS of 4 (it is actually higher, 4.204)

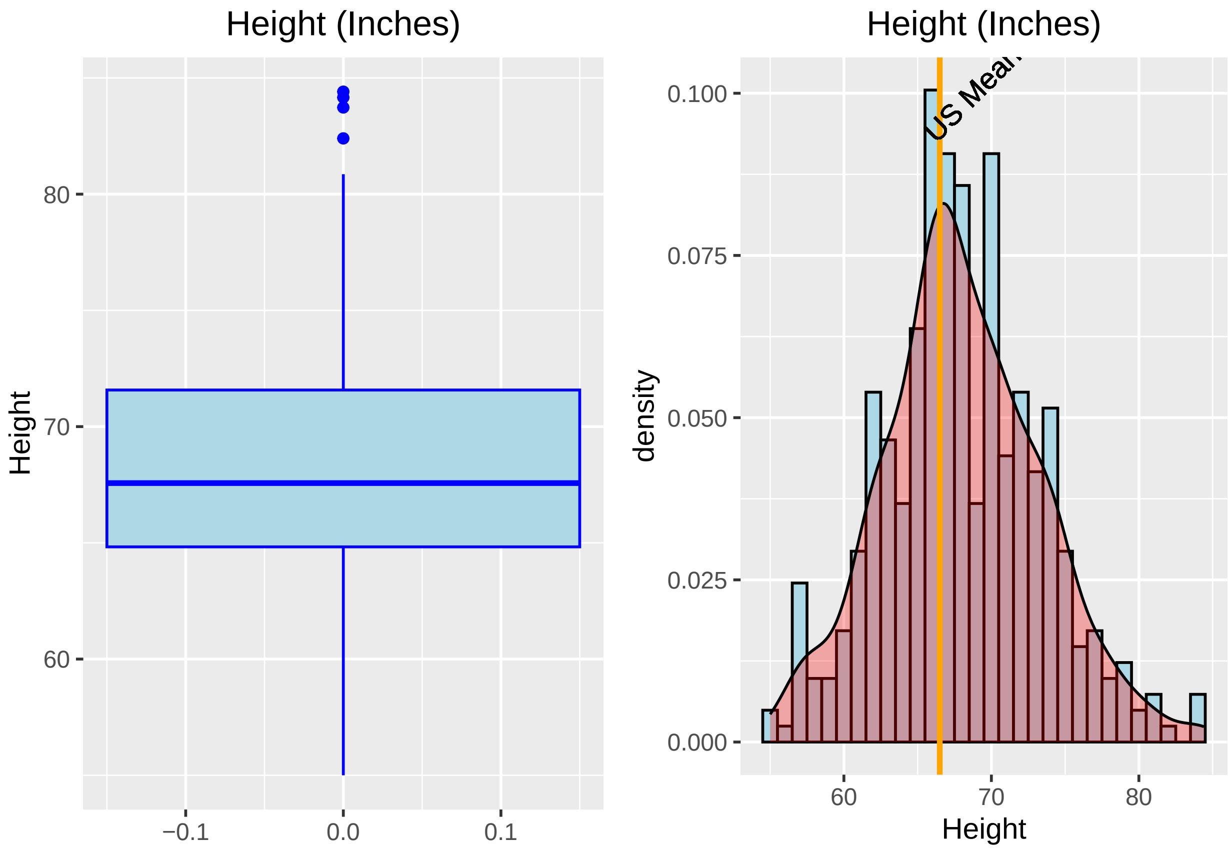
Exercise 61

- Use the data set "height".
- According to the CDC, the mean height of U.S. adults ages 20 and older is about 66.5 inches.
- We have a sample of 408 college students from a single college. Let's test if the mean height of students at this college is significantly different than 66.5 inches using a one-sample t test. H_0: the mean height is the same as the value of 66.5 inches

```
#### Exercise 61 ####

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


box = ggplot(data, aes(y=Height)) +
  geom_boxplot(width=0.3, color="blue", fill="lightblue")+
  scale_fill_viridis(discrete = TRUE, alpha=0.6)+
  labs(title="Height (Inches)")+
  theme(plot.title = element_text(hjust = 0.5), legend.position="bottom")


tt <- seq(0, 10, length.out = 21)
ncp <- seq(0, 6, length.out = 31)
ptn <- outer(tt, ncp, function(t, d) pt(t, df = data$Height, ncp = d))
t.tit <- "Height (Inches)"
matrixx = image(tt, ncp, ptn, zlim = c(0,1), main = t.tit)
```



```
hist = ggplot(data, aes(x=Height)) +
  geom_histogram(aes(y=..density..), colour="black", binwidth=1, fill="lightblue")+
  geom_density(alpha=0.3, fill="red")+
  geom_vline(xintercept=66.5, color="orange", size=1)+
  geom_text(label="US Mean", aes(x=68.7, y=0.1), angle=45, size=4)+
  labs(title="Height (Inches)")+
  theme(plot.title = element_text(hjust = 0.5))


grid.arrange(box, hist, ncol=2)
```



```
res <- t.test(data$Height, mu = 66.5)
res
```

```
##
##  One Sample t-test
##
## data:  data$Height
## t = 5.8096, df = 407, p-value = 1.264e-08
## alternative hypothesis: true mean is not equal to 66.5
## 95 percent confidence interval:
##  67.51346 68.55007
## sample estimates:
## mean of x
##  68.03176
```

```r
paste("p-value is 0.00000001264 which is less than 0.05, thus H0 is rejected")
```

```
## [1] "p-value is 0.00000001264 which is less than 0.05, thus H0 is rejected"
```

Thus, we can conclude that:

- H_0 is rejected, since p-value is less than 0.05 confidence interval (0.00000001264)
- The distribution of height in inches is **not** as the US average of 66.5 (it is actually higher, 68.03)

Exercise 63

- Use the dataset 'diet paired'.
- Is there a statistically significant difference between the body weight of the patients before the diet and after the diet? [H_0: we assume that there is no difference between the body weights]

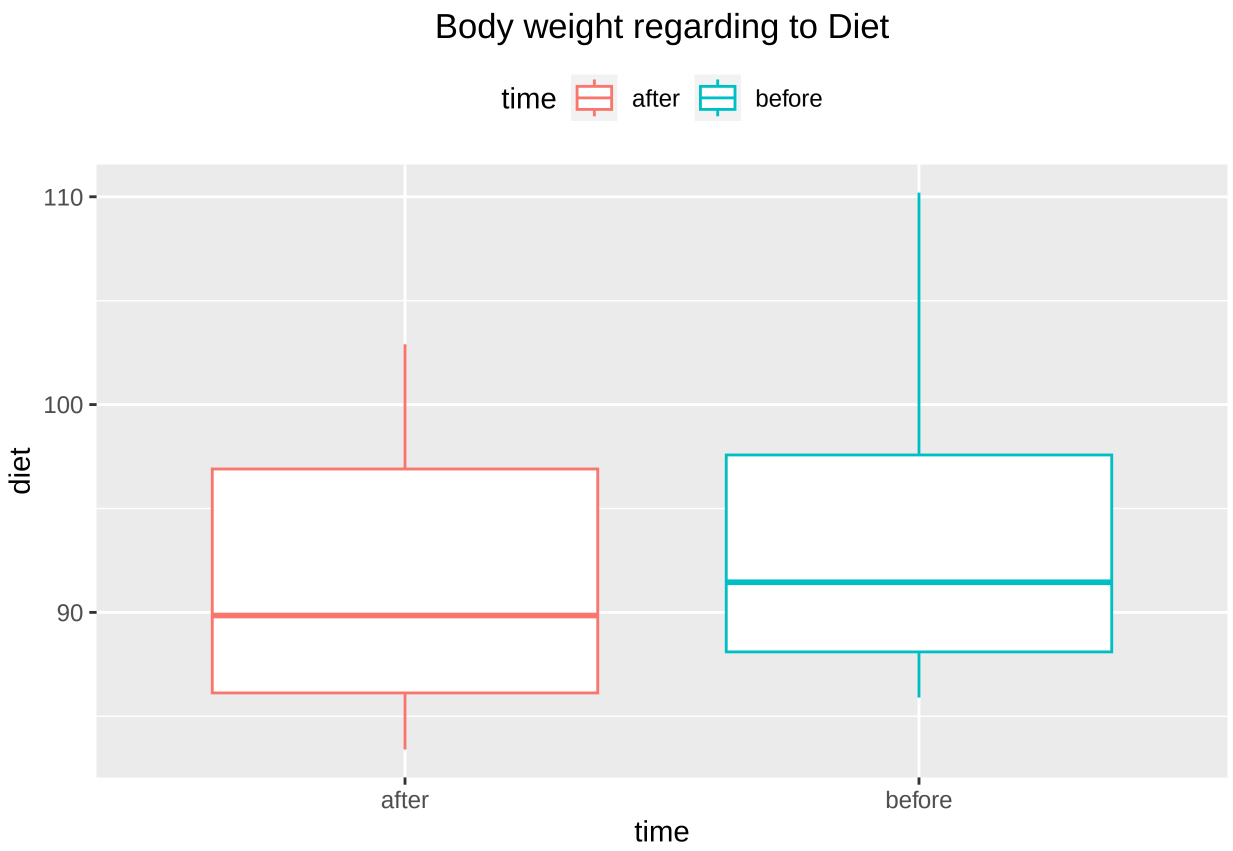
```r
#### Exercise 63 ####

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

head(data)
```

```
##   ï..Patient before_diet after_diet
## 1          1        86.2       83.4
## 2          2        92.7       85.8
## 3          3       102.1       98.3
## 4          4        85.9       83.6
## 5          5        96.3       91.1
## 6          6        90.2       92.7
```

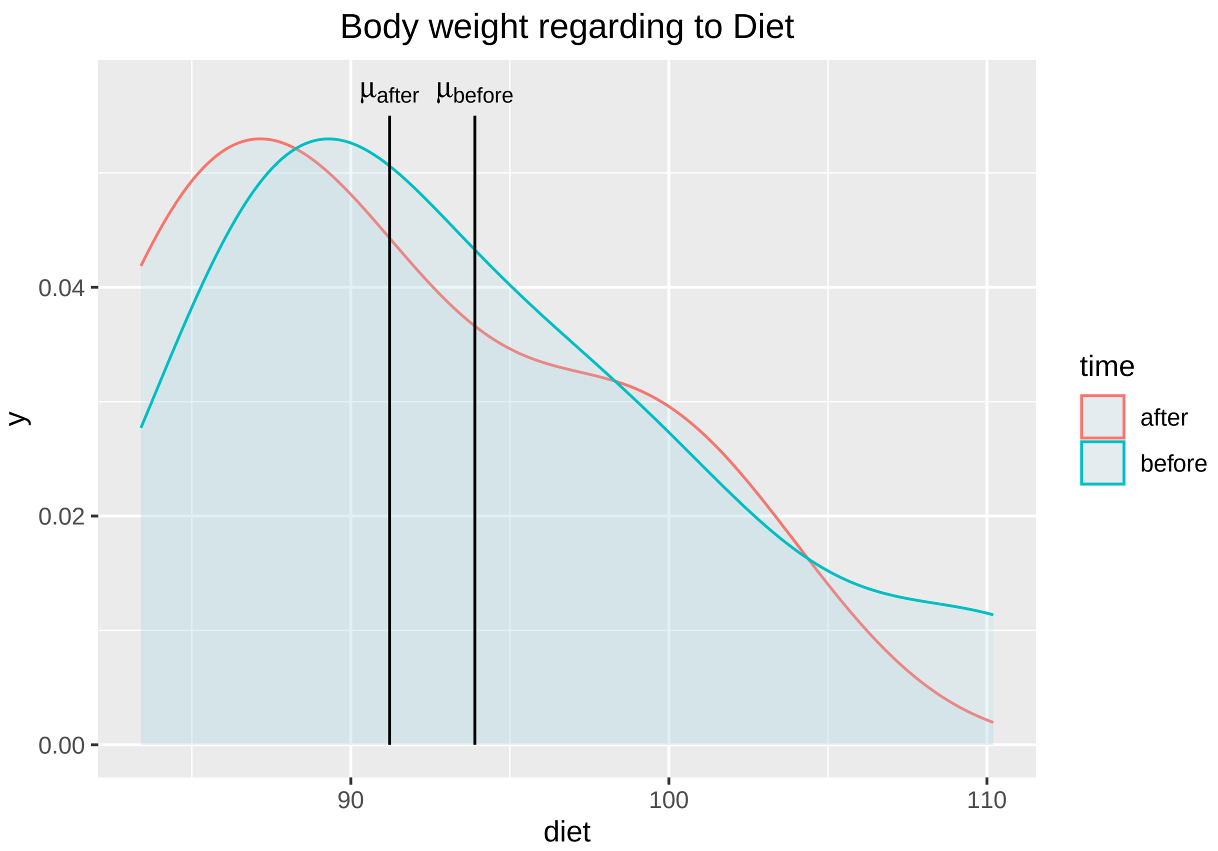
```r
summary(data)
```

```
##    ï..Patient     before_diet       after_diet
##  Min.   : 1.00   Min.   : 85.90   Min.   : 83.40
##  1st Qu.: 3.25   1st Qu.: 88.10   1st Qu.: 86.12
##  Median : 5.50   Median : 91.45   Median : 89.85
##  Mean   : 5.50   Mean   : 93.90   Mean   : 91.22
##  3rd Qu.: 7.75   3rd Qu.: 97.58   3rd Qu.: 96.90
##  Max.   :10.00   Max.   :110.20   Max.   :102.90
```

```r
data_transformed <- data.frame(
  diet = c(data$before_diet, data$after_diet),
  time = c(
    rep("before", length(data$before_diet)),
    rep("after", length(data$after_diet))))

ggplot(data_transformed, aes(x = time, y = diet,
                             color = time)) + geom_boxplot() + labs(title="Body weight regarding to Di
  theme(plot.title = element_text(hjust = 0.5), legend.position="top")
```



```
ggplot(data_transformed, aes(diet, color = time)) +
  geom_density(fill="lightblue", alpha=0.2) + annotate("segment",
                              x = mean(data$before_diet),
                              xend = mean(data$before_diet),
                              y = 0, yend = 0.055, color = "black") +
  annotate("text",
             x = mean(data$before_diet),
             y = 0.057,
             label = expression(mu[before])) +
  annotate("segment",
              x = mean(data$after),
              xend = mean(data$after_diet),
              y = 0, yend = 0.055, color = "black") +
  annotate("text",
              x = mean(data$after_diet),
              y = 0.057,
              label = expression(mu[after])) + labs(title="Body weight regarding to Diet")+
  theme(plot.title = element_text(hjust = 0.5))
```



```
result <-t.test(data$before_diet, data$after_diet, paired=TRUE)
result
```

```
##
##  Paired t-test
##
## data:  data$before_diet and data$after_diet
## t = 2.5492, df = 9, p-value = 0.03124
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3017326 5.0582674
## sample estimates:
## mean of the differences
##                    2.68
```

```
paste("p-value is 0.03124 which is less than 0.05, thus H0 is rejected")
```

```
## [1] "p-value is 0.03124 which is less than 0.05, thus H0 is rejected"
```

Thus, we can conclude the followings:

- H_0 is rejected, since p-value is less than 0.05 confidence interval (0.03124)
- The distribution of body weight before diet is **not** as the same after diet

Exercise 64

- Use the dataset 'OHS 2020 paired'.
- Is there a statistically significant difference between the happiness of the students between the three time points? [H_0: we assume that there is no difference between the three time points]

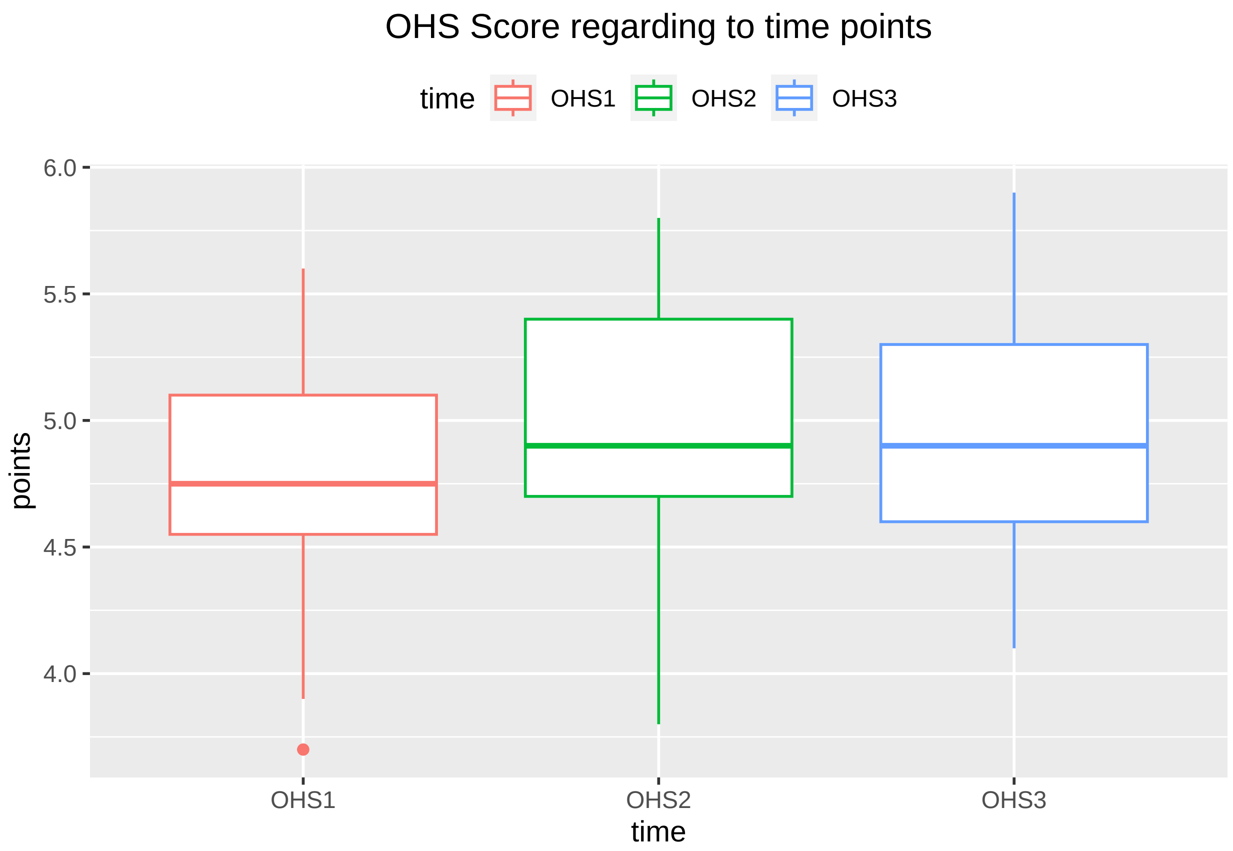
```
#### Exercise 64 ####

data = read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/OHS_2020_
                  stringsAsFactors=F)

head(data)
```

```
##    ï..Name OHS_1 OHS_2 OHS_3
## 1 Jennifer    NA   4.8   5.2
## 2    Tanja   4.6   4.8    NA
## 3    Heike   3.7   3.8   4.5
## 4    David   4.6   5.0   4.9
## 5  Florian   4.2   4.6   4.6
## 6   Denise   4.6   5.4   5.3
```

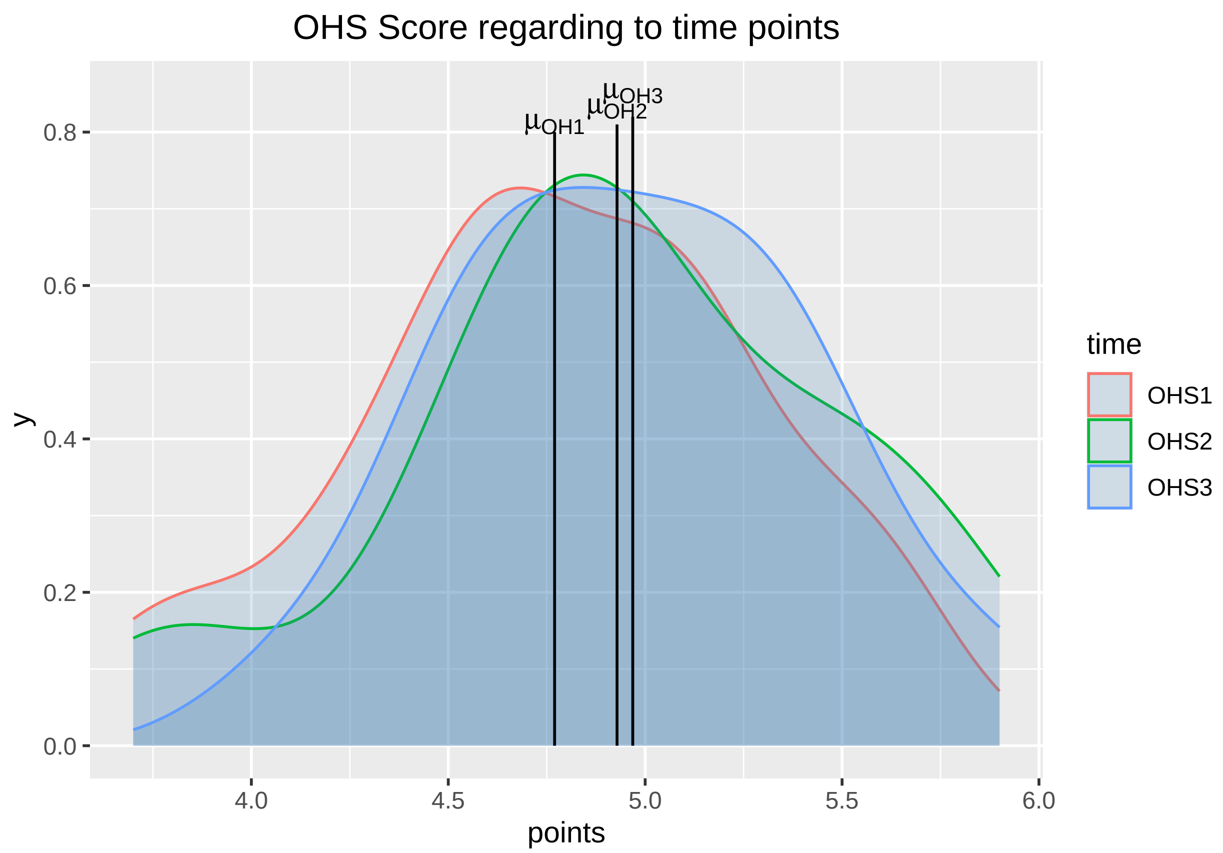
```
summary(data)
```

```
##    ï..Name              OHS_1           OHS_2           OHS_3      
##  Length:21          Min.   :3.70   Min.   :3.800   Min.   :4.100  
##  Class :character   1st Qu.:4.55   1st Qu.:4.700   1st Qu.:4.600  
##  Mode  :character   Median :4.75   Median :4.900   Median :4.900  
##                     Mean   :4.77   Mean   :4.929   Mean   :4.968  
##                     3rd Qu.:5.10   3rd Qu.:5.400   3rd Qu.:5.300  
##                     Max.   :5.60   Max.   :5.800   Max.   :5.900  
##                     NA's   :1                      NA's   :2      
```

```
data_transformed <- data.frame(
  points = c(data$OHS_1, data$OHS_2, data$OHS_3),
  time = c(
    rep("OHS1", length(data$OHS_1)),
    rep("OHS2", length(data$OHS_2)),
    rep("OHS3", length(data$OHS_3))))

ggplot(data_transformed, aes(x = time, y = points,
                             color = time)) + geom_boxplot() + labs(title="OHS Score regarding to time p
  theme(plot.title = element_text(hjust = 0.5), legend.position="top")
```



```
ggplot(data_transformed, aes(points, color = time)) +
  geom_density(fill="steelblue", alpha=0.2) + annotate("segment",
                                                        x = mean(data$OHS_1, na.rm=TRUE),
                                                        xend = mean(data$OHS_1, na.rm=TRUE),
                                                        y = 0, yend = 0.8, color = "black") +
  annotate("text",
           x = mean(data$OHS_1, na.rm=TRUE),
           y = 0.81,
           label = expression(mu[OH1])) +
  annotate("segment",
           x = mean(data$OHS_2),
           xend = mean(data$OHS_2),
           y = 0, yend = 0.81, color = "black") +
  annotate("text",
           x = mean(data$OHS_2),
           y = 0.83,
           label = expression(mu[OH2])) +
  annotate("segment",
           x = mean(data$OHS_3, na.rm=TRUE),
           xend = mean(data$OHS_3, na.rm=TRUE),
           y = 0, yend = 0.82, color = "black") +
  annotate("text",
           x = mean(data$OHS_3, na.rm=TRUE),
           y = 0.85,
           label = expression(mu[OH3])) +labs(title="OHS Score regarding to time points")+
  theme(plot.title = element_text(hjust = 0.5))
```



```
result1 <-t.test(data$OHS_1, data$OHS_2, paired=TRUE)
result1
```

```
##
##  Paired t-test
##
## data:  data$OHS_1 and data$OHS_2
## t = -1.8311, df = 19, p-value = 0.08281
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.35360152  0.02360152
## sample estimates:
## mean of the differences
##                  -0.165
```

```
paste("p-value for OHS_1 and OHS_2 is 0.08281 which is greather than 0.05, thus H0 is accepted")
```

```
## [1] "p-value for OHS_1 and OHS_2 is 0.08281 which is greather than 0.05, thus H0 is accepted"
```

```
result2 <-t.test(data$OHS_1, data$OHS_3, paired=TRUE)
result2
```

```
##
##  Paired t-test
##
## data:  data$OHS_1 and data$OHS_3
## t = -1.9266, df = 17, p-value = 0.07092
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.46557721  0.02113277
## sample estimates:
## mean of the differences
##              -0.2222222
```

```
paste("p-value for OHS_1 and OHS_3 is 0.07092 which is greather than 0.05, thus H0 is accepted")
```

```
## [1] "p-value for OHS_1 and OHS_3 is 0.07092 which is greather than 0.05, thus H0 is accepted"
```

```
result3 <-t.test(data$OHS_2, data$OHS_3, paired=TRUE)
result3
```

```
##
##  Paired t-test
##
## data:  data$OHS_2 and data$OHS_3
## t = -1.026, df = 18, p-value = 0.3185
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.20853101  0.07168891
## sample estimates:
## mean of the differences
##             -0.06842105
```

```
paste("p-value for OHS_2 and OHS_3 is 0.3185 which is greather than 0.05, thus H0 is accepted")
```

```
## [1] "p-value for OHS_2 and OHS_3 is 0.3185 which is greather than 0.05, thus H0 is accepted"
```

- H_0 is accepted, since **ALL** p-values are greater than 0.05 confidence interval (0.08281, 0.07092, 0.3185)
- The distribution of OHS scores between the three time points **is** as the same

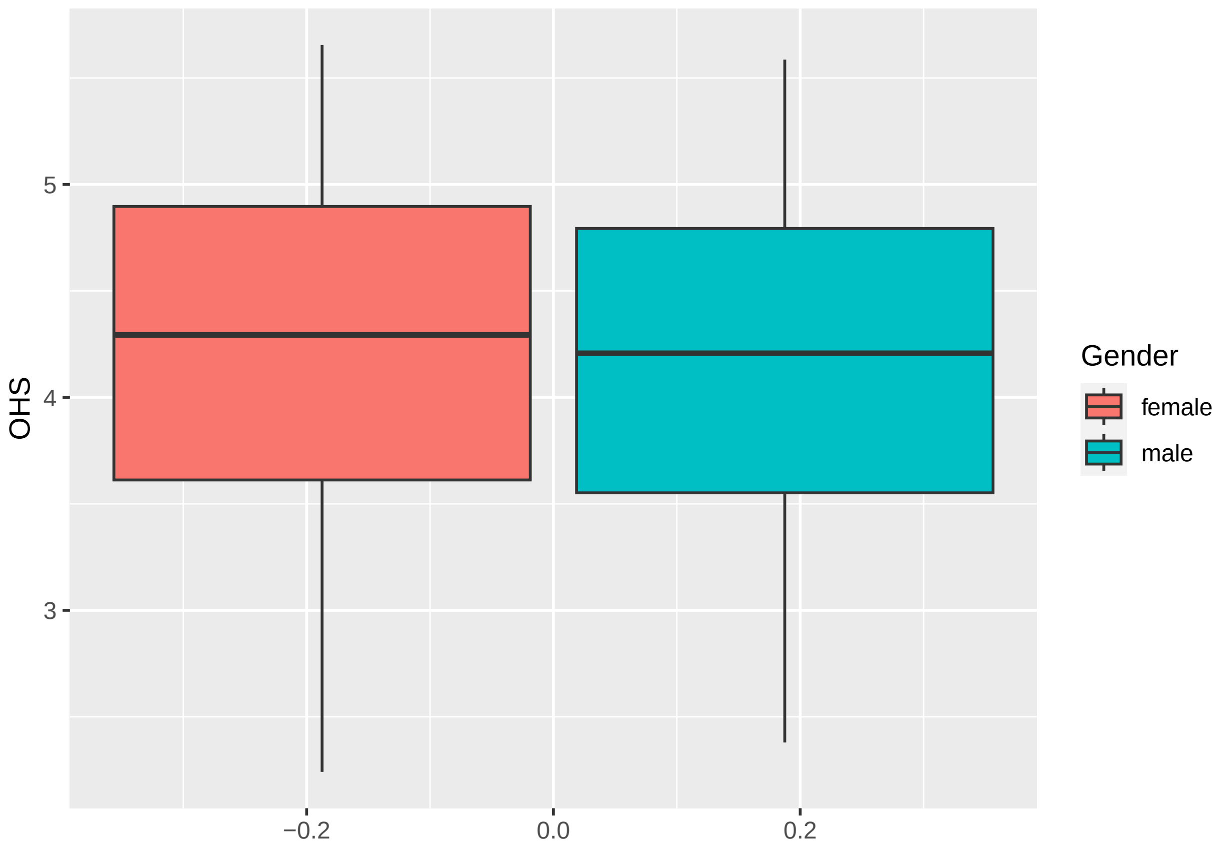
Exercise 66

- Assuming that the data in ICM follows a normal distribution, find the 95% confidence interval estimate of the difference between the Oxford Happiness Score of male and female students. [H_0: the difference between the two genders is 0]

```
#### Exercise 66 ####
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
```

```
ggplot(ICM, aes(group = Gender, y = OHS, fill=Gender))+ geom_boxplot(alpha=1)
```



```
result_t_test <- t.test(OHS ~ Gender, data=ICM)
result_t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  OHS by Gender
## t = 0.34808, df = 111.98, p-value = 0.7284
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
##  -0.2011127  0.2868317
## sample estimates:
## mean in group female   mean in group male
##             4.218298             4.175439
```

```
paste("p-value is 0.7284 which is greater than 0.05, thus H0 is accepted")
```

```
## [1] "p-value is 0.7284 which is greater than 0.05, thus H0 is accepted"
```

```
mean_female <- mean(ICM$OHS[ICM$Gender == "female"],
                       na.rm = T)
mean_female
```

```
## [1] 4.218298
```

```
mean_male <- mean(ICM$OHS[ICM$Gender == "male"],
                  na.rm = T)
mean_male
```

```
## [1] 4.175439
```

Thus, we can conclude the followings:

- H_0 is accepted, since p-value is greater than 0.05 confidence interval (0.7284)
- The distribution of OHS score between the genders **is quite** the same

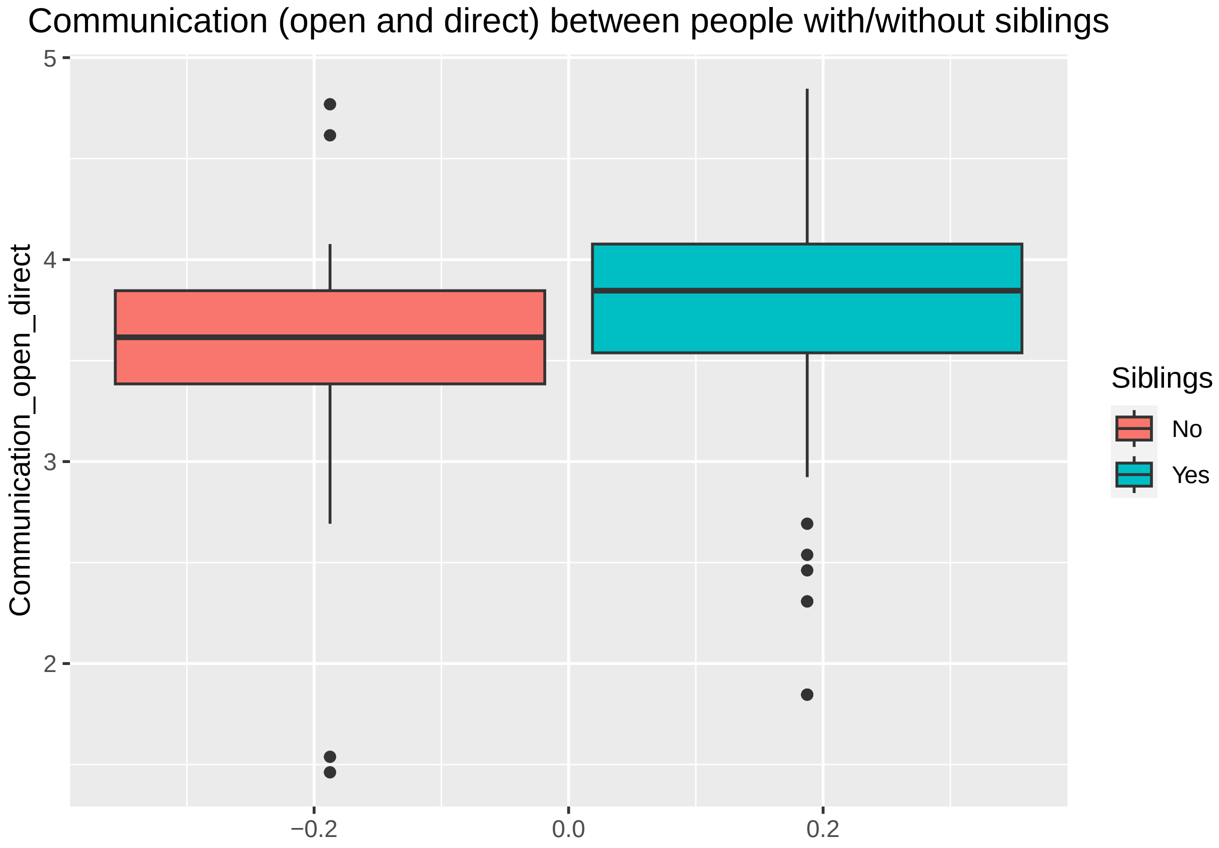
Exercise 67

- Assuming that the data in ICM follows a normal distribution, find the 95% confidence interval estimate of the difference between the Communication style (open and direct) of students with siblings and students without siblings. [H_0: the difference between the students with/without siblings is close to 0]

```
#### Exercise 67 ####

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
```

```
ggplot(ICM, aes(group = Siblings, y = Communication_open_direct, fill=Siblings)) +
  geom_boxplot(alpha=1) +
  labs(title="Communication (open and direct) between people with/without siblings")+
  theme(plot.title = element_text(hjust = 0.5))
```



```
result_t_test <- t.test(Communication_open_direct ~ Siblings, data=ICM)
result_t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  Communication_open_direct by Siblings
## t = -1.7155, df = 24.719, p-value = 0.09877
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.62731413  0.05735326
## sample estimates:
##  mean in group No mean in group Yes
##          3.498328          3.783308
```

```
paste("p-value is 0.09877 which is greater than 0.05, thus H0 is accepted")
```

```
## [1] "p-value is 0.09877 which is greater than 0.05, thus H0 is accepted"
```

```
mean_siblings <- mean(ICM$Communication_open_direct[ICM$Siblings == "Yes"],
                      na.rm = T)
mean_siblings
```

```
## [1] 3.783308
```

```
mean_no_siblings <- mean(ICM$Communication_open_direct[ICM$Siblings == "No"],
                    na.rm = T)
mean_no_siblings
```

```
## [1] 3.498328
```

Thus, we can conclude the followings:

- H_0 is accepted, since p-value is greater than 0.05 confidence interval (0.09877)
- The distribution of Communication (open and direct) score between the people with/without siblings **is quite** the same

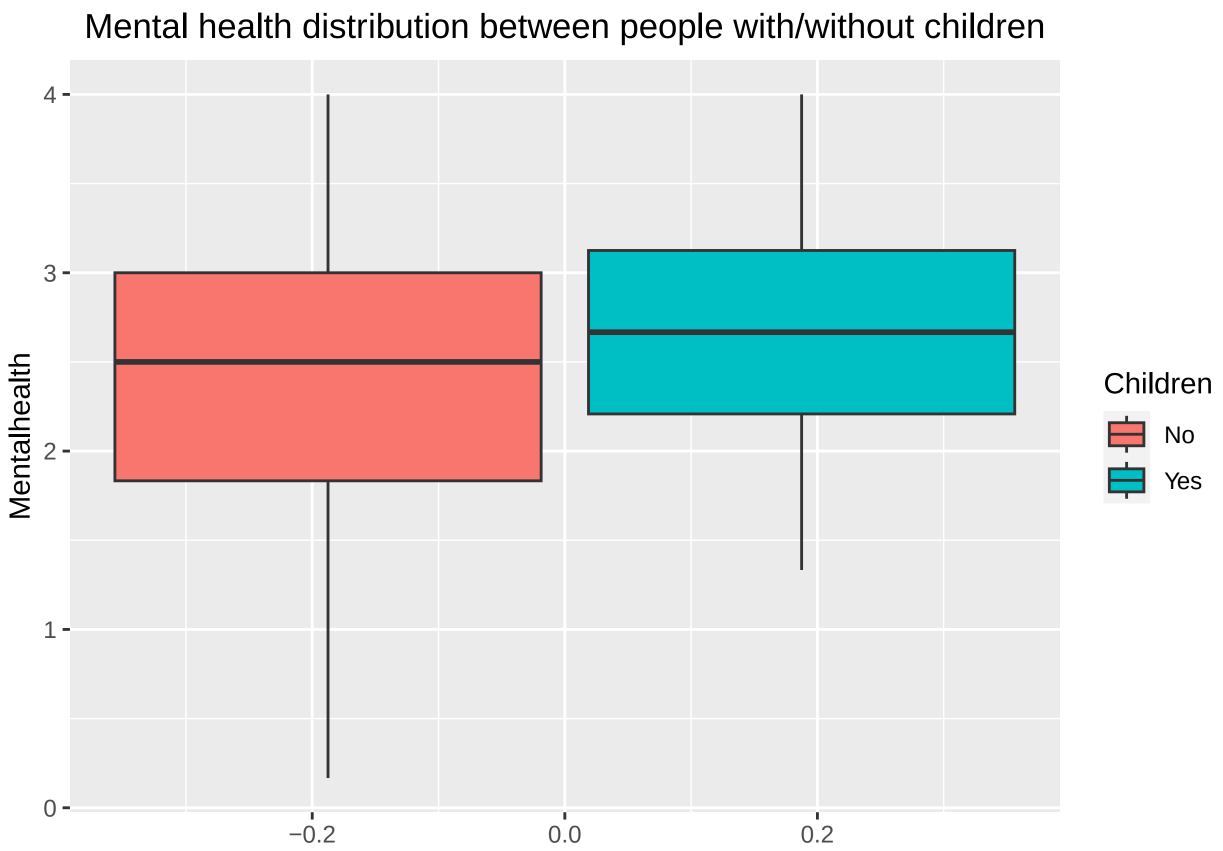
Exercise 68

- Assuming that the data in ICM follows a normal distribution, find the 95% confidence interval estimate of the difference between the mental health of students with children and students without children [H_0: the difference between the students with/without children is close to 0]

```
#### Exercise 68 ####

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
```

```
ggplot(ICM, aes(group = Children, y = Mentalhealth, fill=Children)) +
  geom_boxplot(alpha=1) +
  labs(title="Mental health distribution between people with/without children")+
  theme(plot.title = element_text(hjust = 0.5))
```



```
result_t_test <- t.test(Mentalhealth ~ Children, data=ICM)
result_t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  Mentalhealth by Children
## t = -2.253, df = 44.366, p-value = 0.02925
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.60023943 -0.03349073
## sample estimates:
##  mean in group No mean in group Yes
##          2.399802          2.716667
```

```
paste("p-value is 0.02925 which is less than 0.05, thus H0 is rejected")
```

```
## [1] "p-value is 0.02925 which is less than 0.05, thus H0 is rejected"
```

```
mean_children <- mean(ICM$Mentalhealth[ICM$Children == "Yes"],
                      na.rm = T)
mean_children
```

```
## [1] 2.716667
```

```
mean_no_children <- mean(ICM$Mentalhealth[ICM$Children == "No"],
                         na.rm = T)
mean_no_children
```

```
## [1] 2.399802
```

Thus, we can conclude the followings:

- H_0 is rejected, since p-value is lower than 0.05 confidence interval (0.02925)
- The distribution of mental health score between the people with/without children is **not** the same