Homework 2

Darian-Florian Voda

2022-10-20

Exercise 13: Contingency tables

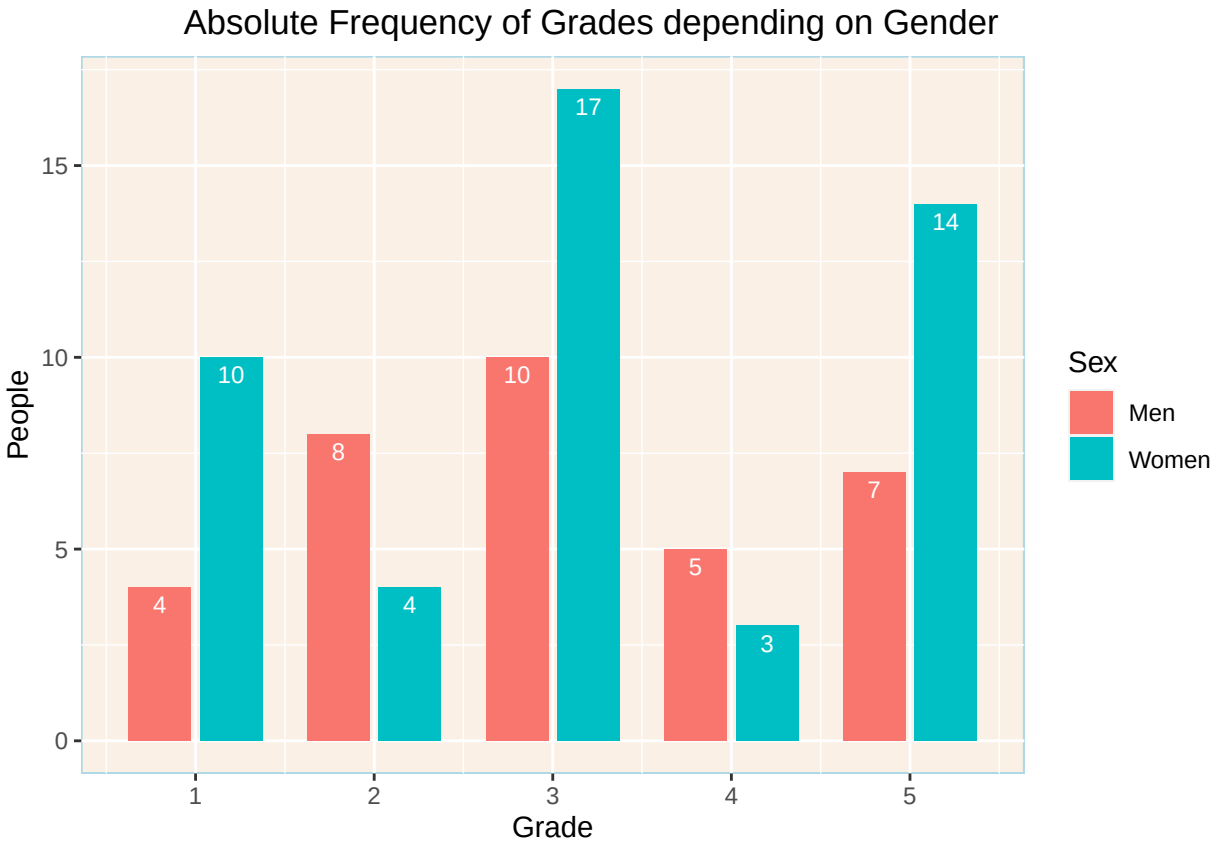
1. Use the data 'students.txt'
2. Determine the absolute frequencies of the ordinal feature "Grade" depending on gender.
3. Then, make a two-dimensional bar chart.

```
students<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/studer

library(mosaic)
library(ggplot2)
library(viridis)
library(hrbrthemes)
tally(~Grade | Sex, data = students)
```

```
##      Sex
## Grade  F  M
##     1 10  4
##     2  8  4
##     3 17 10
##     4  5  3
##     5 14  7
```

```
ex13 <- data.frame(Sex=rep(c("Men", "Women")),
                   Grade=rep(c(1, 2, 3, 4, 5),2),
                   people=c(4, 4, 10, 3, 7, 10, 8, 17, 5, 14))


ggplot(ex13, aes(x=Grade, y=people, fill=Sex)) +
  geom_bar(stat="identity", width=0.7, position=position_dodge(width=0.8)) +
  geom_text(
    aes(label = people),
    colour = "white", size = 3,
    vjust = 1.5, position = position_dodge(.8)) +
  ggtitle("Absolute Frequency of Grades depending on Gender") +
  theme(plot.title = element_text(hjust = 0.8),
        legend.position="right",
        panel.background = element_rect(fill = "linen",
                                        colour = "lightblue",
                                        size = 0.5, linetype = "solid"),
        panel.grid.major = element_line(size = 0.5, linetype = 'solid',
```

```
                                        colour = "white"),
      panel.grid.minor = element_line(size = 0.25, linetype = 'solid',
                                        colour = "white")) + labs(x="Grade", y="People")
```



Exercise 15: Median and arithmetic mean

In a company, the workers are paid on a daily basis following this salary structure:

```
##   Salary_Euro Workers
## 1         550       9
## 2         650      15
## 3         750      27
## 4         850      25
## 5         950      17
## 6        1050      10
## 7        1150       7
```

```
ex15 = data.frame(salary=c(550, 650, 750, 850, 950, 1050, 1150), workers=c(9, 15, 27, 25, 17, 10, 7))

summary(ex15)
```

```
##      salary        workers
```

```
##  Min.   : 550   Min.   : 7.00
##  1st Qu.: 700   1st Qu.: 9.50
##  Median : 850   Median :15.00
##  Mean   : 850   Mean   :15.71
##  3rd Qu.:1000   3rd Qu.:21.00
##  Max.   :1150   Max.   :27.00
```

THIS IS WRONG MEDIAN & MEAN (For me, personally)

```
nr_work = sum(ex15$workers)
sum_salaries = sum(ex15$salary*ex15$workers)

mean_val = sum_salaries/nr_work
mean_val
```

```
## [1] 826.3636
```

```
all_salaries = c(rep.int(ex15$salary,ex15$workers))
all_salaries
```

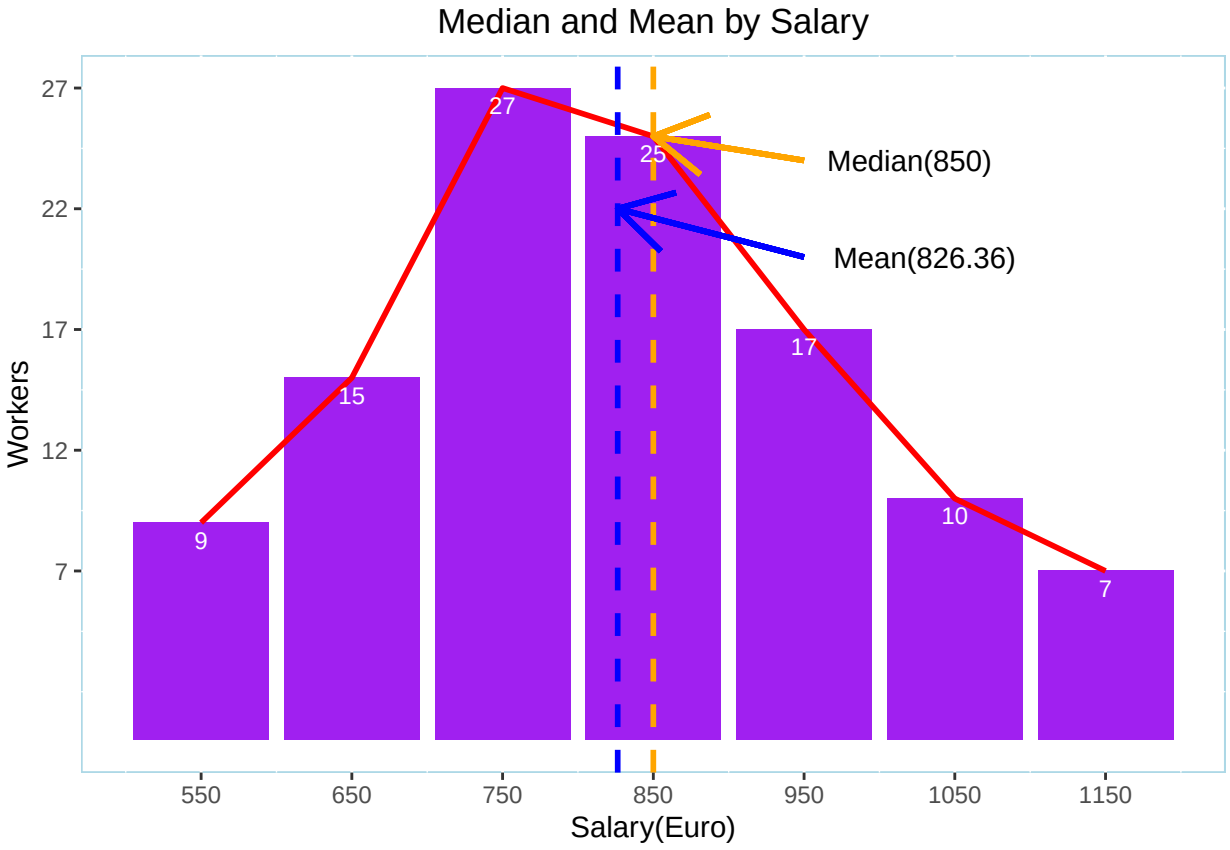
```
##   [1]  550  550  550  550  550  550  550  550  550  650  650  650  650  650  650
##  [16]  650  650  650  650  650  650  650  650  650  750  750  750  750  750  750
##  [31]  750  750  750  750  750  750  750  750  750  750  750  750  750  750  750
##  [46]  750  750  750  750  750  750  850  850  850  850  850  850  850  850  850
##  [61]  850  850  850  850  850  850  850  850  850  850  850  850  850  850  850
##  [76]  850  950  950  950  950  950  950  950  950  950  950  950  950  950  950
##  [91]  950  950  950 1050 1050 1050 1050 1050 1050 1050 1050 1050 1050 1150 1150
## [106] 1150 1150 1150 1150 1150
```

```
median_val = median(all_salaries)
median_val
```

```
## [1] 850
```

```
library(ggplot2)
ggplot(ex15, aes(x=salary, y=workers)) +
  geom_bar(stat="identity", fill="purple") +
  geom_line(aes(x=salary, y=workers),stat="identity",color="red",size=1)+
  geom_vline(aes(xintercept = mean_val), color='blue', lty='dashed', lwd=1) +
  geom_vline(aes(xintercept = median_val), color='orange', lty='dashed', lwd=1) +
  geom_text(
    aes(label = workers),
    colour = "white", size = 3,
    vjust = 1.5, position = position_dodge(.7)
    ) +
  ggtitle("Median and Mean by Salary") +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position="right",
        panel.background = element_rect(fill = "transparent",
                                        colour = "lightblue",
```

```
                                    size = 0.5, linetype = "solid")) +
      labs(x="Salary(Euro)", y="Workers") +
scale_x_continuous(breaks = round(seq(min(ex15$salary), max(ex15$salary), by = 100),1)) +
scale_y_continuous(breaks = round(seq(min(ex15$workers), max(ex15$workers), by = 5),1)) +
geom_segment(aes(x = 950, y = 20, xend = mean_val, yend = 22),
             arrow = arrow(length = unit(0.8, "cm")), lwd=1, color="blue") +
annotate("text", x=1030, y=20, label= "Mean(826.36)") +
geom_segment(aes(x = 950, y = 24, xend = median_val, yend = 25),
           arrow = arrow(length = unit(0.8, "cm")), lwd=1, color="orange") +
annotate("text", x=1020, y=24, label= "Median(850)")
```



Exercise 16: Median and arithmetic mean

In a company, the four employees receive the following salaries in Euro: 600 700 750 3200 1. Calculate the arithmetic mean of the salaries. 2. Is it a typical, representative value?

```
library(ggplot2)
ex16 = data.frame(salary = c(600, 700, 750, 3200), workers=c(1, 1, 1, 1))
mean(ex16$salary)
```

```
## [1] 1312.5
```

```
par(adj=0.3)
boxplot(ex16$salary ~ ex16$workers,
        ylab="Employees", xlab="Salary(EUR)",
        main = "Salaries of Employees",
        col = "orange",
        border = "brown",
        horizontal = TRUE
        )
```

Arithmetic mean is 1312.5 which is not typical due to the outlier 3200 that rises up the mean considerably.

Exercise 17: Arithmetic mean

In a company, the employees receive an average salary of 2000 Euro. Male employees receive an average salary of 2080 Euro, while female employees receive an average salary of 1680 Euro 1. Determine the percentage of male and female employees in this company.

```
mean_salary = 2000
ex17 = data.frame(salaries = c(2080, 1680), sex=c("M", "F"))
summary(ex17)
```

```
##     salaries        sex
##  Min.   :1680   Length:2
##  1st Qu.:1780   Class :character
```

```
##  Median :1880   Mode  :character
##  Mean   :1880
##  3rd Qu.:1980
##  Max.   :2080
```

In this exercise, we won't use any Data Visualization, but we will use **MATH** Since $mean\_salary$ is 2000, we will use the mean formula in order to find the percentages:

$$\begin{aligned} mean &= 2000 \\ mean_male &= 2080 \\ mean_female &= 1680 \\ \frac{x \cdot mean_male + y \cdot mean_female}{x+y} &= mean \\ 2080x + 1680y &= 2000x + 2000y \\ 80x - 320y &= 0 \\ x = \frac{320y}{80} x &= 4y \end{aligned}$$

Since x is 4 times bigger than y, we need to think which values for y gives us a percent of 100 assigning values for both x and y

$$\begin{aligned} y &= 20; \\ \rightarrow x &= 80 \\ Verifying &: \\ 0.8 \cdot 2080 + 0.2 \cdot 1680 &= 2000 \end{aligned}$$

Thus, there are 80% males and 20% females in the company.

Exercise 18: Median and arithmetic mean

In a fitness studio, the athletes showed the following body masses in kg:

body_mass	athletes
61	5
63	18
65	42
67	27
69	8
71	2

1. Calculate the median.
2. Calculate the arithmetic mean.

```
ex18 = data.frame(body_mass = c(61, 63,65, 67, 69, 71), athletes=c(5, 18, 42, 27, 8, 2))


nr_athletes = sum(ex18$athletes)
```

```r
sum_body = sum(ex18$body_mass*ex18$athletes)

mean_val = sum_body/nr_athletes
mean_val
```

```
## [1] 65.41176
```

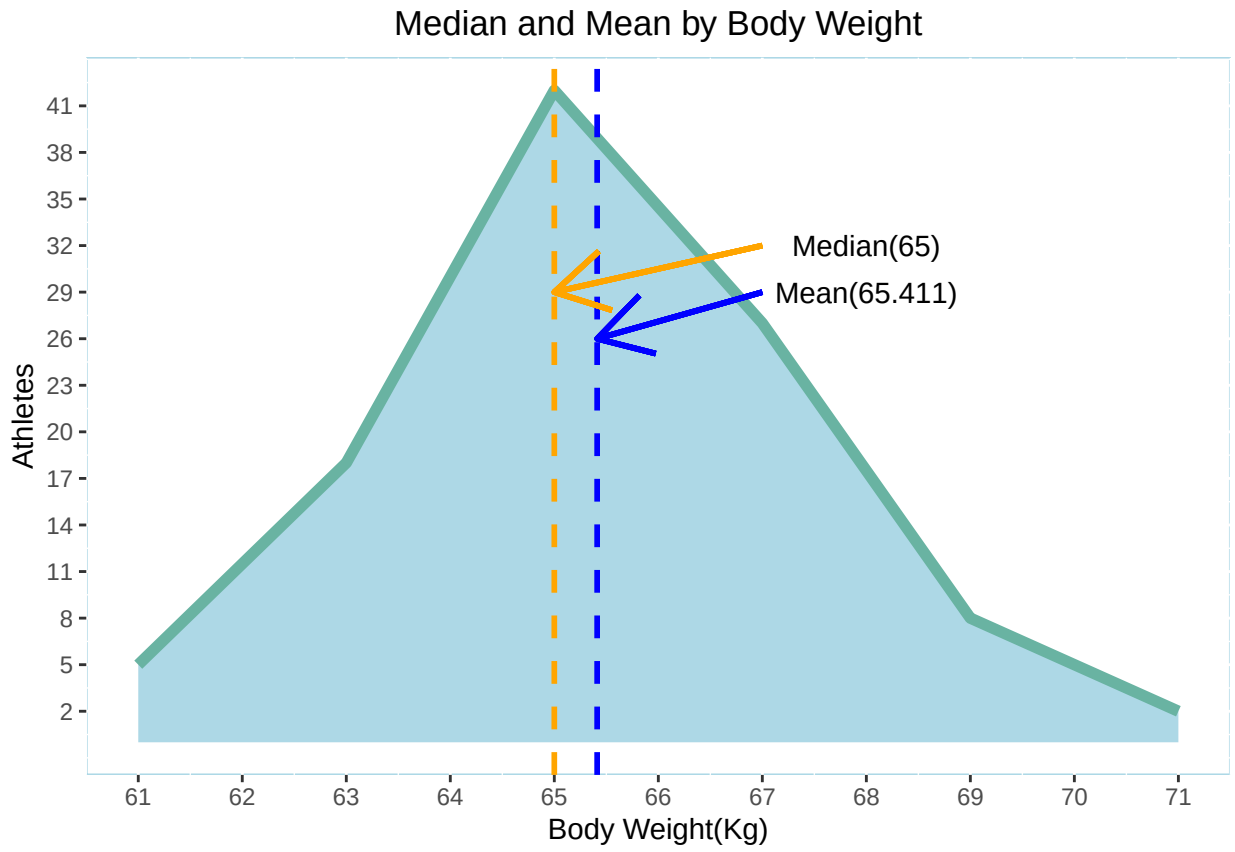
```r
all_body = c(rep.int(ex18$body_mass,ex18$athletes))
all_body
```

```
##   [1] 61 61 61 61 61 63 63 63 63 63 63 63 63 63 63 63 63 63 63 63 63 63 63 65 65
##  [26] 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65
##  [51] 65 65 65 65 65 65 65 65 65 65 65 65 65 65 65 67 67 67 67 67 67 67 67 67 67
##  [76] 67 67 67 67 67 67 67 67 67 67 67 67 67 67 67 67 67 69 69 69 69 69 69 69 69
## [101] 71 71
```

```r
median_val = median(all_body)
median_val
```

```
## [1] 65
```

```r
library(ggplot2)
library(hrbrthemes)
ggplot(ex18, aes(x=body_mass, y=athletes)) +
  geom_area(stat="identity", fill="lightblue") +
  geom_line(aes(x=body_mass, y=athletes),stat="identity",color="#69b3a2",size=2)+
  geom_vline(aes(xintercept = mean_val), color='blue', lty='dashed', lwd=1) +
  geom_vline(aes(xintercept = median_val), color='orange', lty='dashed', lwd=1) +
  #geom_text(
    #aes(label = athletes), colour = "red", lwd=4, vjust = 1.5, position = position_dodge(.7)) +
  ggtitle("Median and Mean by Body Weight") +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position="right",
        panel.background = element_rect(fill = "transparent",
                                        colour = "lightblue",
                                        size = 0.5, linetype = "solid")) +
  labs(x="Body Weight(Kg)", y="Athletes") +
  scale_x_continuous(breaks = round(seq(min(ex18$body_mass), max(ex18$body_mass), by = 1),1)) +
  scale_y_continuous(breaks = round(seq(min(ex18$athletes), max(ex18$athletes), by = 3),1)) +
  geom_segment(aes(x = 67, y = 29, xend = mean_val, yend = 26),
               arrow = arrow(length = unit(0.8, "cm")), lwd=1, color="blue") +
  annotate("text", x=68, y=29, label= "Mean(65.411)") +
  geom_segment(aes(x = 67, y = 32, xend = median_val, yend = 29),
              arrow = arrow(length = unit(0.8, "cm")), lwd=1, color="orange") +
  annotate("text", x=68, y=32, label= "Median(65)")
```



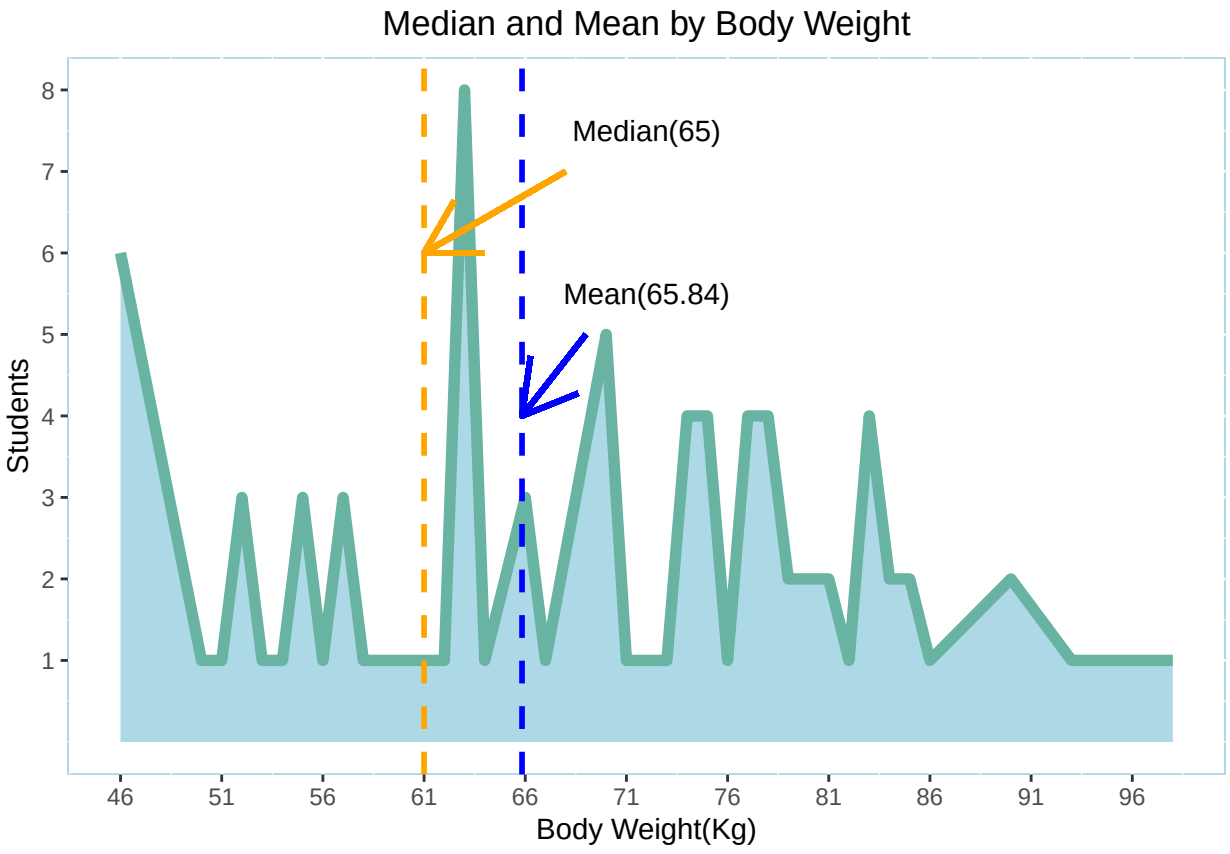
Exercise 20: Arithmetic means and medians with R

1. Use the data 'students.txt'
2. Determine the arithmetic mean and median for the variable body weight.
3. Determine the arithmetic means and median for the variable body weight depending on smoking behaviour.

Subpoint 2

```
students<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/studen
mean_val = mean(students$Weight_kg)
median_val = median(students$Weight_kg)

d = table(unlist(students$Weight_kg))

ex20 = data.frame(weight = c(unique(students$Weight_kg)),  nr_stud=c(d))

library(ggplot2)
library(hrbrthemes)
ggplot(ex20, aes(x=weight, y=nr_stud)) +
  geom_area(stat="identity", fill="lightblue") +
  geom_line(aes(x=weight, y=nr_stud),stat="identity",color="#69b3a2",size=2)+
```

```
geom_vline(aes(xintercept = mean_val), color='blue', lty='dashed', lwd=1) +
geom_vline(aes(xintercept = median_val), color='orange', lty='dashed', lwd=1) +
#geom_text(
#aes(label = athletes), colour = "red", lwd=4, vjust = 1.5, position = position_dodge(.7)) +
ggtitle("Median and Mean by Body Weight") +
theme(plot.title = element_text(hjust = 0.5),
      legend.position="right",
      panel.background = element_rect(fill = "transparent",
                                      colour = "lightblue",
                                      size = 0.5, linetype = "solid")) +
labs(x="Body Weight(Kg)", y="Students") +
scale_x_continuous(breaks = round(seq(min(ex20$weight), max(ex20$weight), by = 5),1)) +
scale_y_continuous(breaks = round(seq(min(ex20$nr_stud), max(ex20$nr_stud), by = 1),1)) +
geom_segment(aes(x = 69, y = 5, xend = mean_val, yend = 4),
             arrow = arrow(length = unit(0.8, "cm")), lwd=1, color="blue") +
annotate("text", x=72, y=5.5, label= "Mean(65.84)") +
geom_segment(aes(x = 68, y = 7, xend = median_val, yend = 6),
             arrow = arrow(length = unit(0.8, "cm")), lwd=1, color="orange") +
annotate("text", x=72, y=7.5, label= "Median(65)")
```



Subpoint 3

```
library(mosaic)
students<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/studen
mean_val1 = mean(students$Weight_kg[students$Smoking=="yes"])
mean_val1
```

```
## [1] 64.96154
```

```
mean_val0 = mean(students$Weight_kg[students$Smoking=="no"])
mean_val0
```

```
## [1] 66.25
```

```
median_val1 = median(students$Weight_kg[students$Smoking=="yes"])
median_val1
```

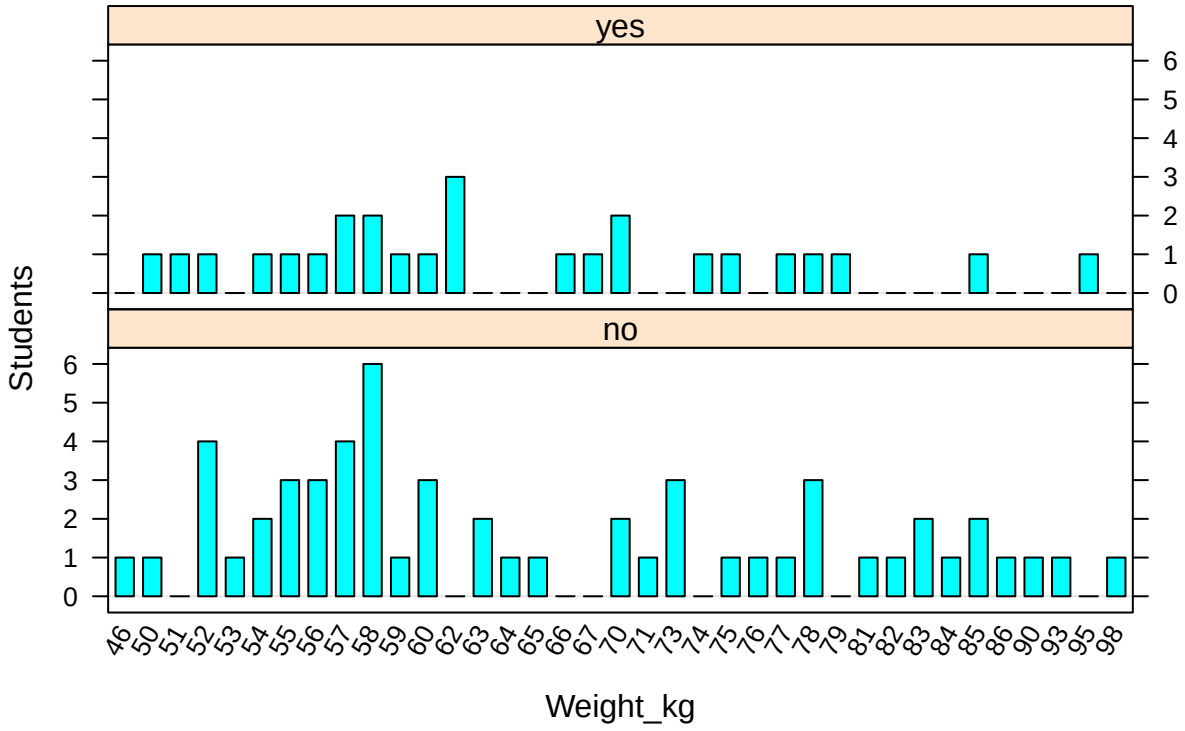
```
## [1] 62
```

```
median_val0 = median(students$Weight_kg[students$Smoking=="yes"])
median_val0
```

```
## [1] 62
```

```
df = tally(~Weight_kg | Smoking, data = students)

ex20_2 = data.frame(Weight_kg = c(unique(students$Weight_kg)), Smoking = c(df))

bargraph(~Weight_kg | Smoking, data = students, ylab="Students",
main="Means and medians for body weight depending on smoking behaviour",
scales=list(x=list(rot=60)), layout=c(1,2))
```

Means and medians for body weight depending on smoking behaviour



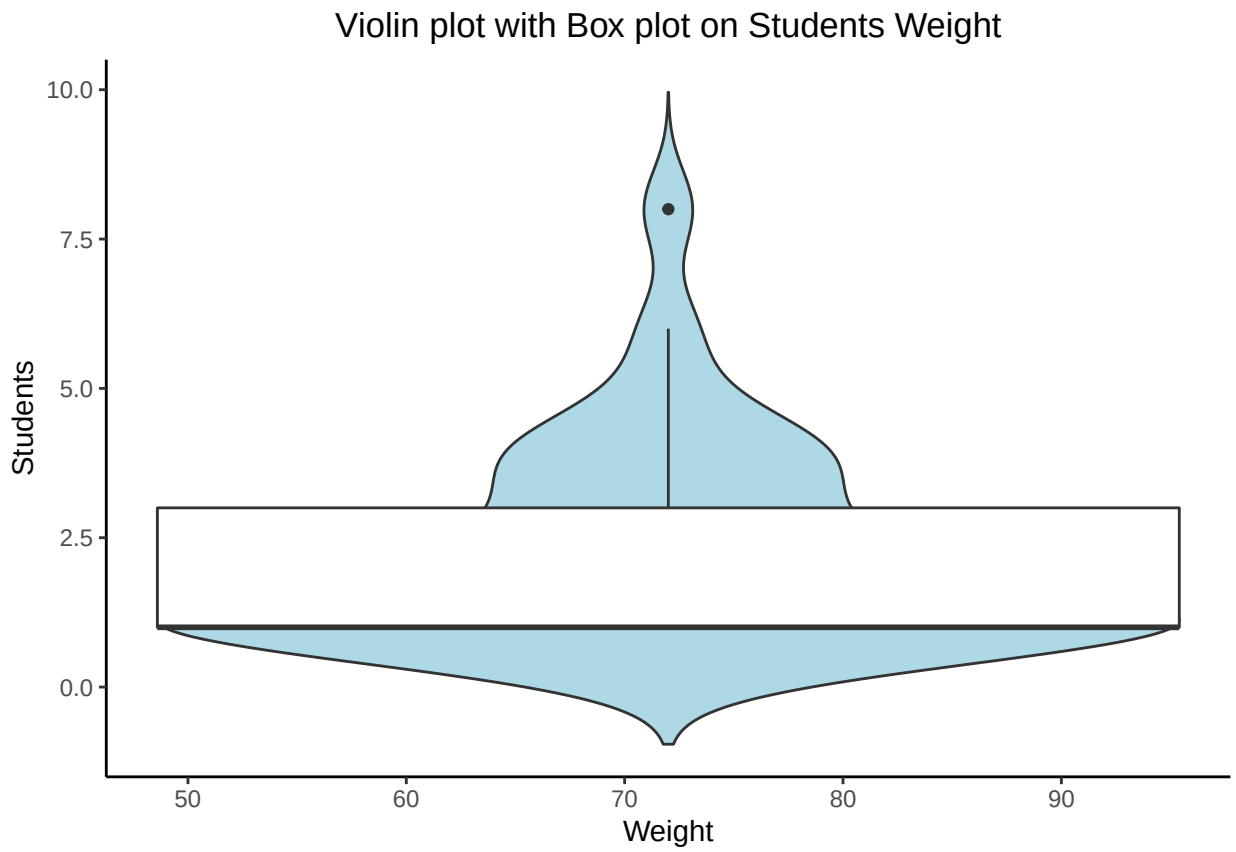
Exercise 22: Boxplot with R

1. Use the data 'students.txt'
2. Create a boxplot for the variable body weight.
3. Create a boxplot for the variable body weight depending on smoking behaviour.

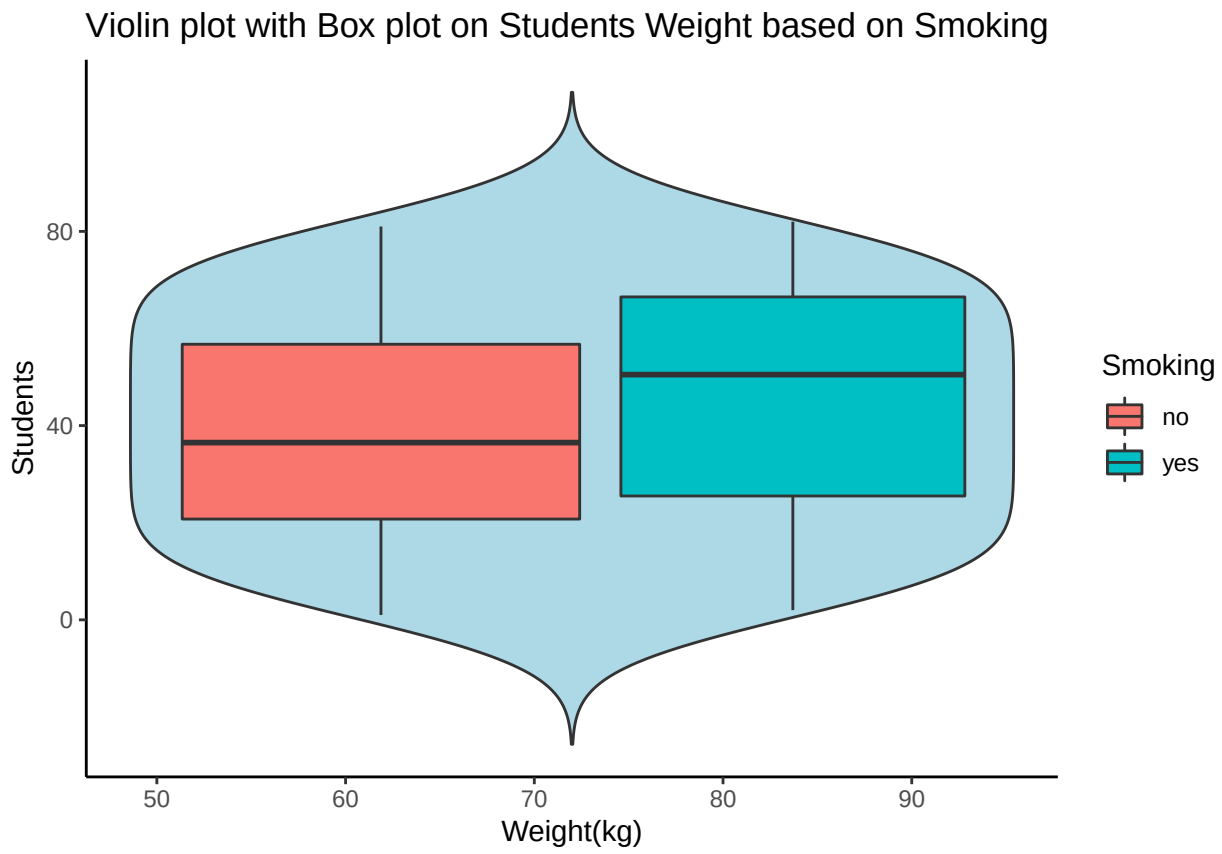
```
# Exercise 22
library(ggplot2)
library(dplyr)
library(hrbrthemes)
library(viridis)
students<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/studer


d = table(unlist(students$Weight_kg))
d
```

```
##
## 46 50 51 52 53 54 55 56 57 58 59 60 62 63 64 65 66 67 70 71 73 74 75 76 77 78
##  1  2  1  5  1  3  4  4  6  8  2  4  3  2  1  1  1  1  4  1  3  1  2  1  2  4
## 79 81 82 83 84 85 86 90 93 95 98
##  1  1  1  2  1  3  1  1  1  1  1
```

```
ex22 = data.frame(weight = c(unique(students$Weight_kg)),  nr_stud=c(d))

ggplot(ex22, aes(x=weight, y=nr_stud)) +
  geom_violin(trim=FALSE, fill="lightblue")+
  labs(title="Violin plot with Box plot on Students Weight", x="Weight", y = "Students")+
  geom_boxplot(width=0.1)+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```



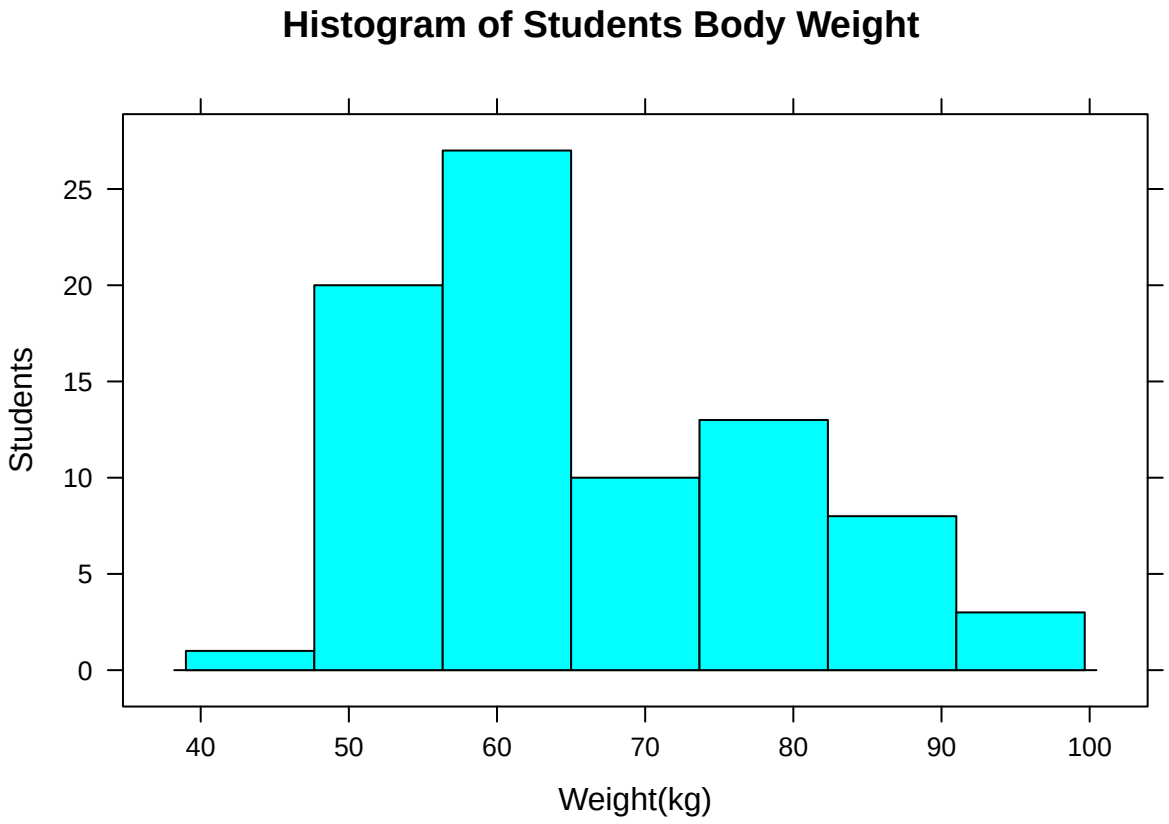

```
ggplot(students, aes(x=Weight_kg, y=ID, fill=Smoking)) +
  geom_violin(trim=FALSE, fill="lightblue")+
  labs(title="Violin plot with Box plot on Students Weight based on Smoking ", x="Weight", y = "Students
  geom_boxplot(width=0.1)+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5)) +
  xlab("Weight(kg)") + ylab("Students")
```



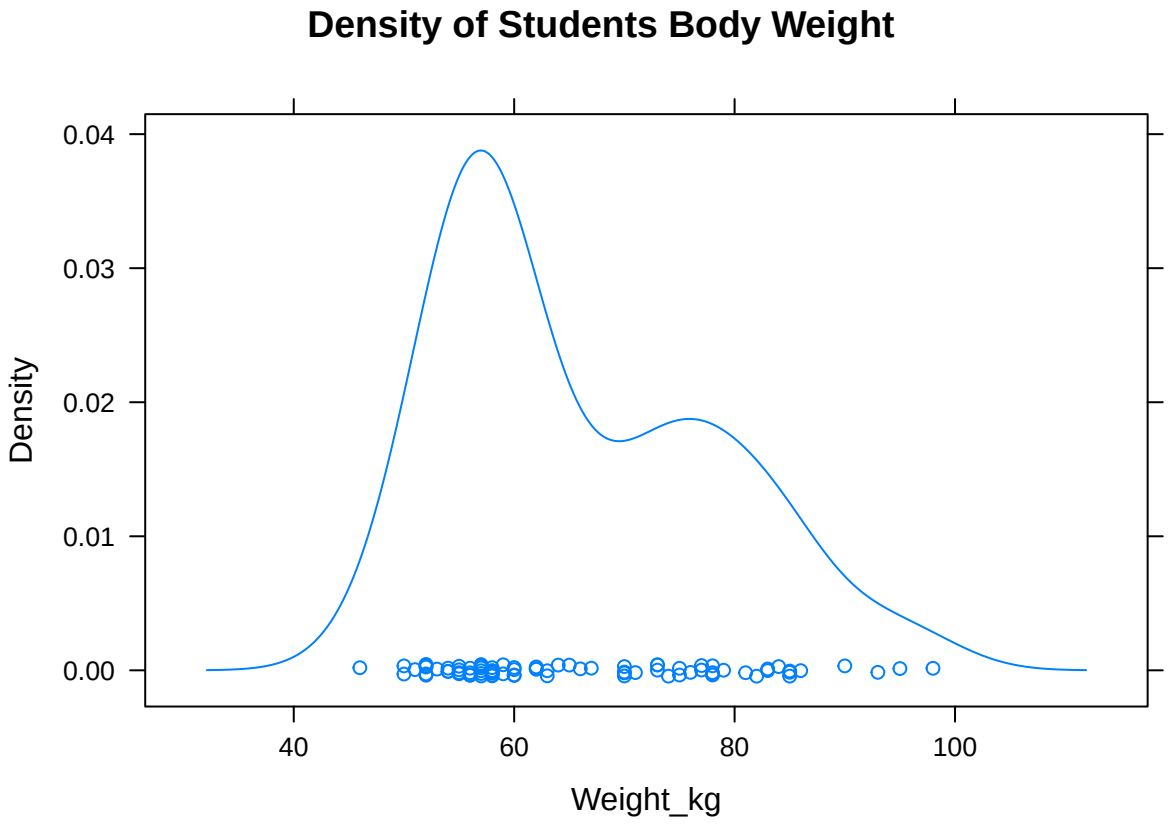
Exercise 24: Skewness and kurtosis with R

1. Use the data 'students.txt'
2. Create a histogram for the variable body weight.
3. Determine the skewness for the variable body weight.
4. Determine the kurtosis for the variable body weight.

```
library(ggplot2)
library(moments)

students<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/stude

graph1 <- histogram(students$Weight_kg, ylab="Students",
                    main="Histogram of Students Body Weight",
                    xlab="Weight(kg)",
                    type="count"
)

graph2 <- densityplot(~Weight_kg, data = students, main="Density of Students Body Weight")
plot(graph1, position=c(0, 0, 1, 1))
```




```
plot(graph2, position=c(0, 0, 1, 1))
```




```
skewness(students$Weight_kg)
```

```
## [1] 0.6754061
```

```
kurtosis(students$Weight_kg)
```

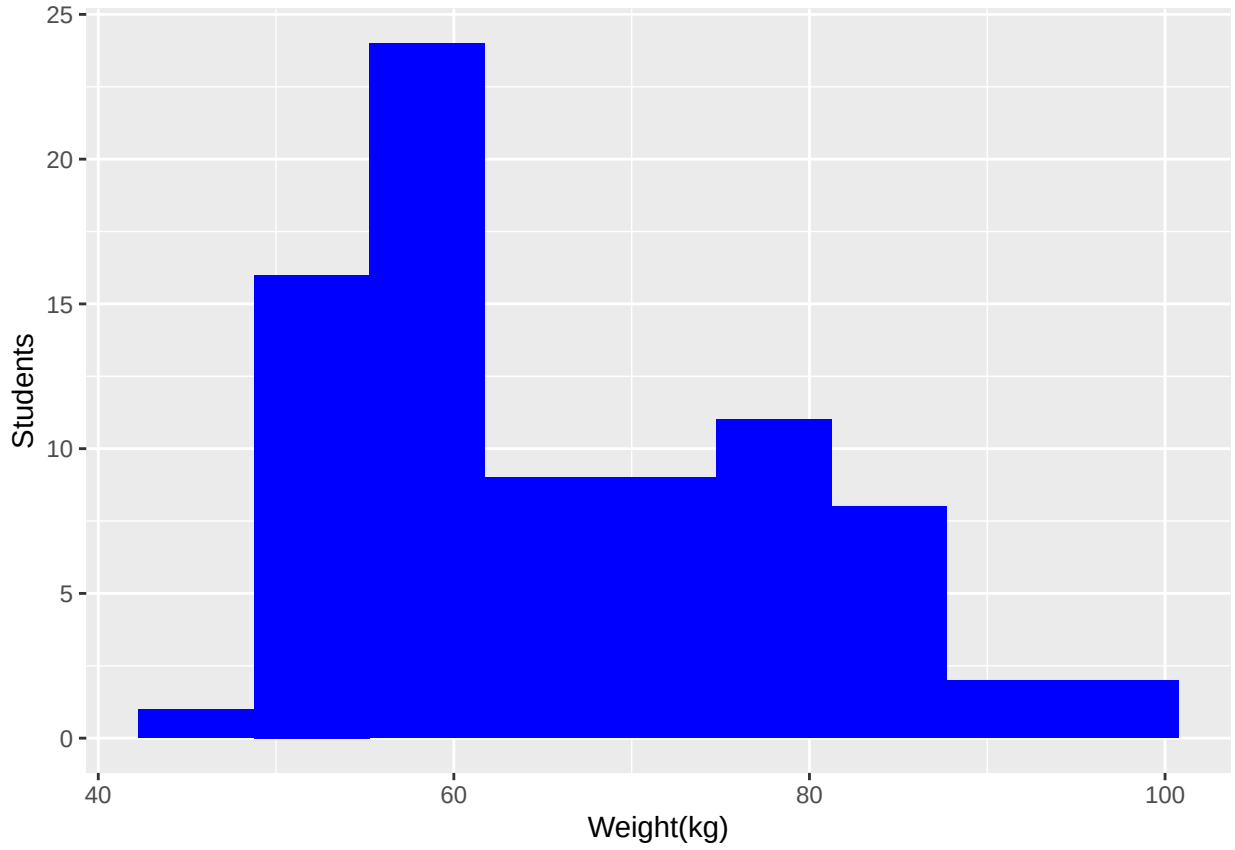
```
## [1] 2.418121
```

Exercise 26: Density function of body weight

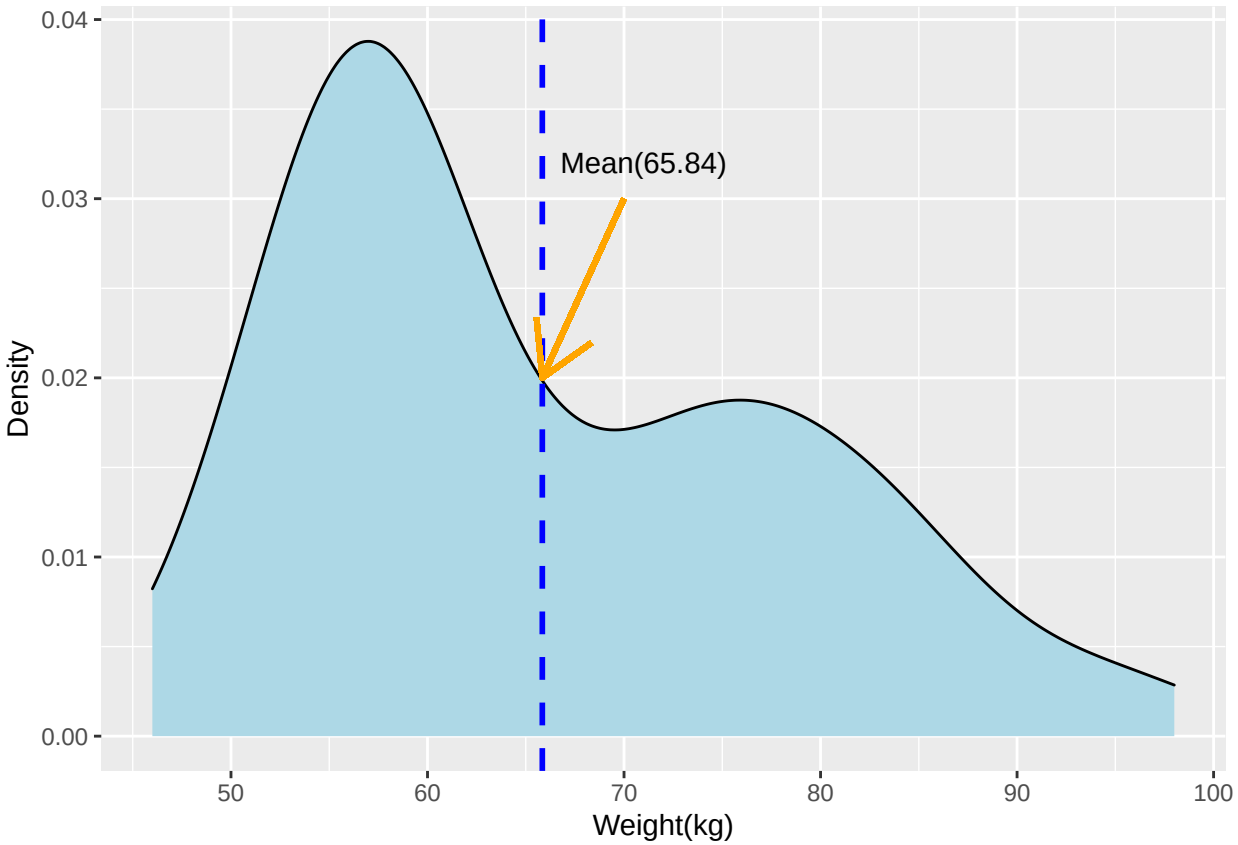
1. Use the data 'students.txt'
2. Create a histogram for the variable body weight.
3. Create a density function plot for the variable body weight.

```
library(ggplot2)
library(mosaic)
students<-read.delim("C:/Users/daria/OneDrive/Desktop/Master - AppDS/Statistics/Datasets-20221007/stude
```

```
ggplot(students, aes(Weight_kg)) +
geom_histogram(bins=sqrt(length(students$Weight_kg)), fill="blue") +
xlab("Weight(kg)") + ylab("Students")
```




```
ggplot(students, aes(x=Weight_kg)) +
      geom_density(fill="lightblue") +
      geom_vline(aes(xintercept=mean(Weight_kg)),
                 color="blue", linetype="dashed", size=1) +
  xlab("Weight(kg)") + ylab("Density") +
  geom_segment(aes(x = 70, y = 0.03, xend = mean(students$Weight_kg), yend = 0.02),
                arrow = arrow(length = unit(0.8, "cm")), lwd=1, color="orange") +
  annotate("text", x=71, y=0.032, label= "Mean(65.84)")
```


Mean(65.84)
Density
0.04
0.03
0.02
0.01
0.00
50
60
70
80
90
100
Weight(kg)