Programming in R: Solution for practical session 6a, Topics in Tidyverse and data analysis, (14/11/2023)

Ziv Shkedy, Rahmasari Nur Aazizah, Thi Huyen NGUYEN, Bernard Osangir, Rudradev Sengupta, Ewoud De Troyer and Marijke Van Moerbeke.

General information

- The practical session is focus on statistical modeling and Tidyverse and consists of 6 questions in which you are asked to conduct an analysis of a dataset. Dataset is available in R as a data frame as a part of the R package palmerpenguins.
- Your output should consists a PDF file which contains the results and R code.
- Solutions will be available online in BB in a later stage (you will receive an email about this via BB).

R functions

Some of the R functions that are used in this practical session are:

- lm(), aov().
- ggplot(), boxplot()....

Question 1:

In questions Q1-Q6 we focus on the penguins data which available as a part of the palmerpenguins R package. To assess the data, you need ti install the package. More information about the data is available in https://allisonhorst.github.io/palmerpenguins/.For all the questions (Q1-Q5) we conduct a complete case analysis, i.e., all observations should not have missing values. The first 6 lines in the data are shown below

```
library(palmerpenguins)
data("penguins", package = "palmerpenguins")
penguins <- drop_na(penguins)
#dim(penguins)
#table(penguins$sex)
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_l~1 body_~2 sex    year
##   <fct>   <fct>              <dbl>         <dbl>       <int>   <int> <fct> <int>
## 1 Adelie  Torgersen           39.1          18.7         181    3750 male   2007
## 2 Adelie  Torgersen           39.5          17.4         186    3800 fema~  2007
## 3 Adelie  Torgersen           40.3          18           195    3250 fema~  2007
## 4 Adelie  Torgersen           36.7          19.3         193    3450 fema~  2007
## 5 Adelie  Torgersen           39.3          20.6         190    3650 male   2007
## 6 Adelie  Torgersen           38.9          17.8         181    3625 fema~  2007
## # ... with abbreviated variable names 1: flipper_length_mm, 2: body_mass_g
```

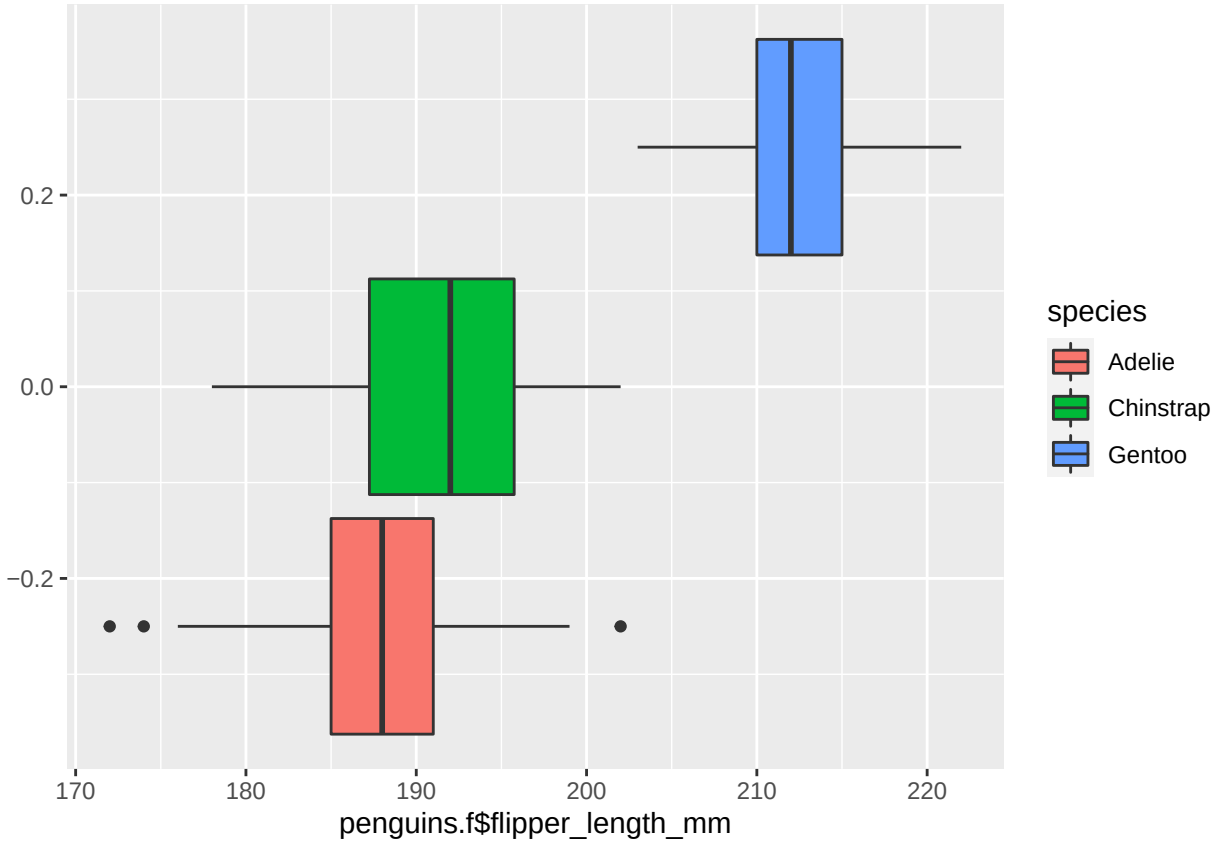
1. Create a new data frame contains only the female.
2. For the new data frame, calculate the mean, median and standard deviation for the variable flipper_length_mm by species and produce the table below.
3. Produce the figures below.

Solution

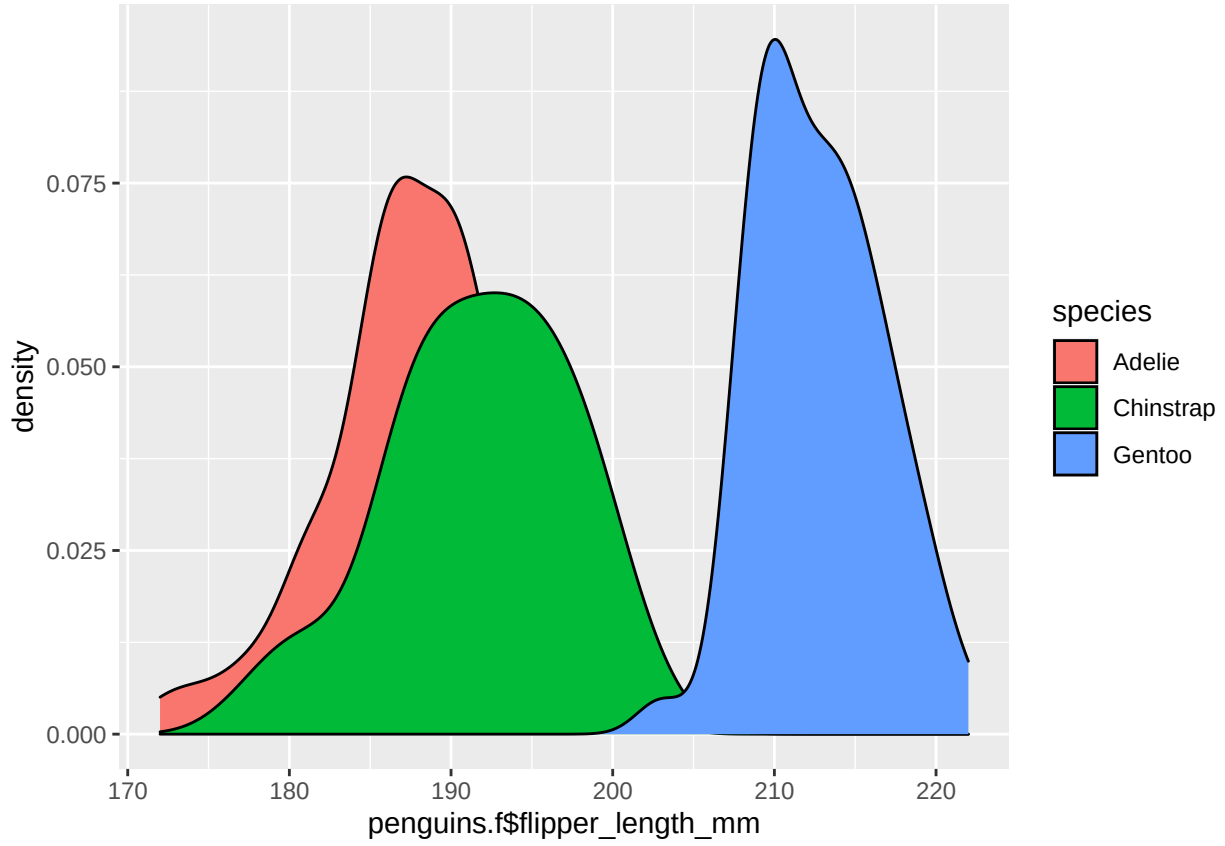
```
penguins.f<-penguins %>% filter(sex=="female")
XXX<-penguins.f %>% group_by(species) %>% summarize(average = mean(flipper_length_mm), median = median(
print(XXX)
```

```
## # A tibble: 3 x 4
##   species   average median standard_deviation
##   <fct>       <dbl>  <dbl>              <dbl>
## 1 Adelie       188.    188               5.60
## 2 Chinstrap    192.    192               5.75
## 3 Gentoo       213.    212               3.90
```

```
ggplot(penguins.f,aes(penguins.f$flipper_length_mm,fill=species))+geom_boxplot()
```




```
ggplot(penguins.f,aes(penguins.f$flipper_length_mm,fill=species))+geom_density()
```



Question 2:

1. Create a new data frame contains all the female with bill_depth_mm < 18. How many observations are
   included in the new data frame.
2. Sort the data frame according to the value of bill depth.
3. Calculate the quantiles of the bill depth.
4. Define an indicator variable that takes the value of 1 for $16 <$ bill depth < 18 and zero otherwise. Produce a
   2×3 contingency table (see below) of the indicator and species. Use a chi-square test to test the hypothesis that the indicator and the species are independent.
5. Create a new data frame contains observations of female with bill depth between 16 to 18 mm. How many
   observations are included in the new data frame.

Solution

```
library(palmerpenguins)
data("penguins", package = "palmerpenguins")
penguins <- drop_na(penguins)
penguins.f1<-penguins %>% filter(sex=="female")%>% filter(bill_depth_mm < 18)
dim(penguins.f1)
```

```
## [1] 133   8
```

```
penguins.f1 %>% arrange(bill_depth_mm) %>% head()
```

```
## # A tibble: 6 x 8
##   species island bill_length_mm bill_depth_mm flipper_leng~1 body_~2 sex    year
##   <fct>   <fct>           <dbl>         <dbl>          <int>   <int> <fct> <int>
## 1 Gentoo  Biscoe           42.9          13.1            215    5000 fema~  2007
## 2 Gentoo  Biscoe           46.1          13.2            211    4500 fema~  2007
## 3 Gentoo  Biscoe           44.9          13.3            213    5100 fema~  2008
## 4 Gentoo  Biscoe           43.3          13.4            209    4400 fema~  2007
## 5 Gentoo  Biscoe           46.5          13.5            210    4550 fema~  2007
## 6 Gentoo  Biscoe           42            13.5            210    4150 fema~  2007
## # ... with abbreviated variable names 1: flipper_length_mm, 2: body_mass_g
```

```
quantile(penguins.f1$bill_depth_mm)
```

```
##   0%  25%  50%  75% 100%
## 13.1 14.4 16.5 17.3 17.9
```

```
#ggplot(penguins.f1,aes(penguins.f1$bill_depth_mm,fill=species))+geom_density()
penguins.f2<-penguins %>% filter(sex=="female")%>% filter(bill_depth_mm >16 & bill_depth_mm < 18)
dim(penguins.f2)
```

```
## [1] 72  8
```

```
penguins.f1<-penguins %>% filter(sex=="female")
index<-penguins.f1$bill_depth_mm*0
index[penguins.f1$bill_depth_mm >16 & penguins.f1$bill_depth_mm < 18]<-1
table(index,penguins.f1$species)
```

```
##
## index Adelie Chinstrap Gentoo
##     0     26         9     58
##     1     47        25      0
```

```
chisq.test(index,penguins.f1$species)
```

```
##
##  Pearson's Chi-squared test
##
## data:  index and penguins.f1$species
## X-squared = 70.032, df = 2, p-value = 6.204e-16
```

Question 3:

In this question we focus on the variable bill_length_mm for the species Adelie and Chinstrap.

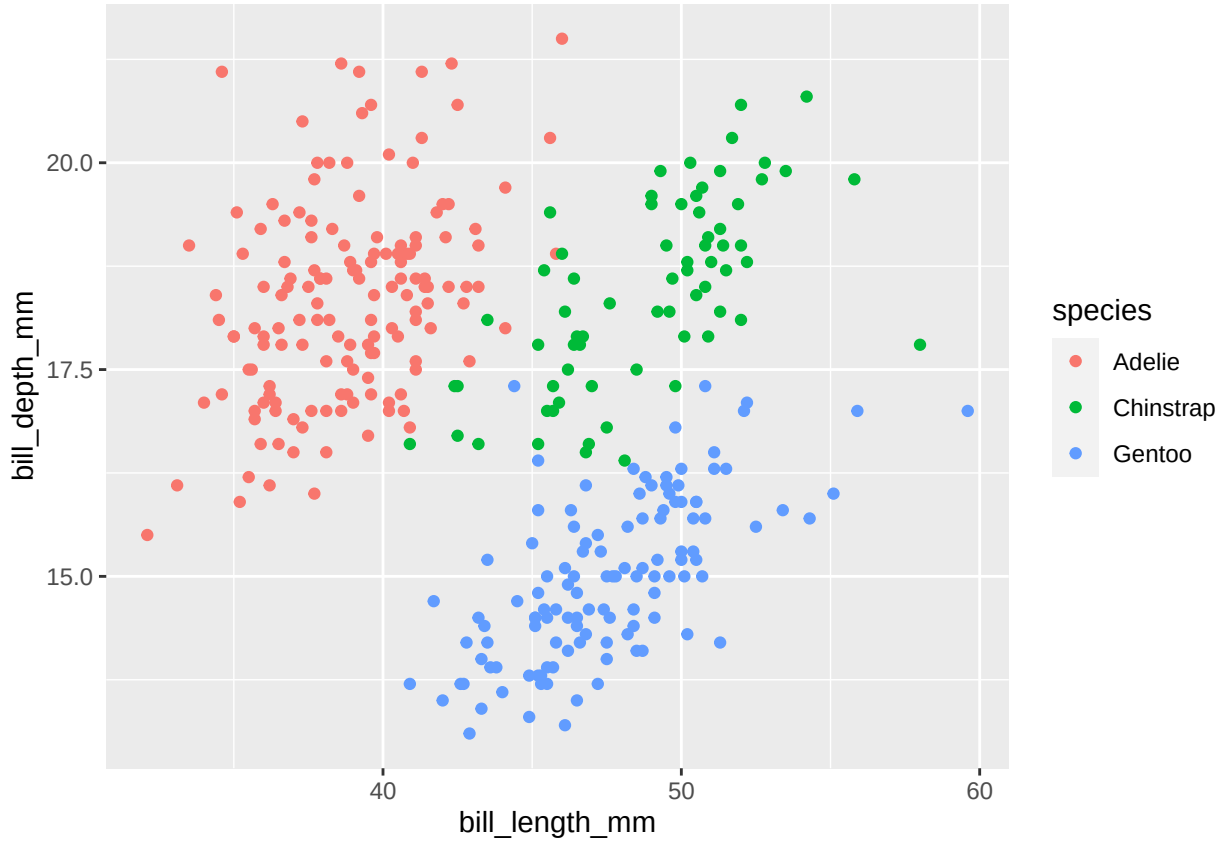
1. Produce the figures below and fit a regression model based on your figure.

Solution

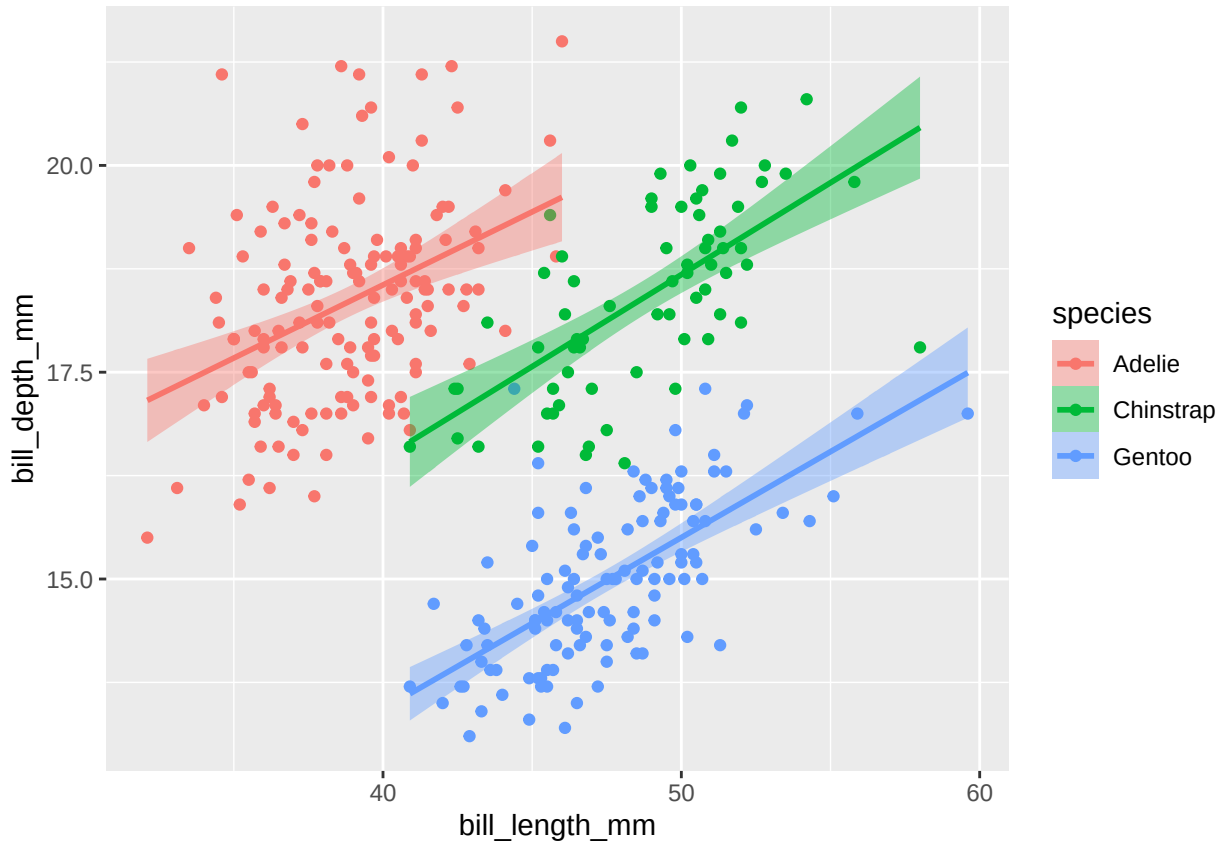
```
library(palmerpenguins)
data("penguins", package = "palmerpenguins")
penguins <- drop_na(penguins)
penguins.f<-penguins %>% filter(sex=="female")
penguins.fac<-penguins.f %>% filter(species%in%c("Adelie","Chinstrap"))
dim(penguins.fac)
```

```
## [1] 107   8
```

```
#print(penguins.fac$species)
#tapply(penguins.fac$bill_length_mm,penguins.fac$species,mean)
ggplot(penguins,aes(bill_length_mm,bill_depth_mm,fill=species,color=species))+geom_point()
```




```
ggplot(penguins,aes(bill_length_mm,bill_depth_mm,fill=species,color=species))+geom_point()+geom_smooth(
```




```
fit.1<-lm(bill_depth_mm~bill_length_mm+species,data=penguins)
summary(fit.1)
```

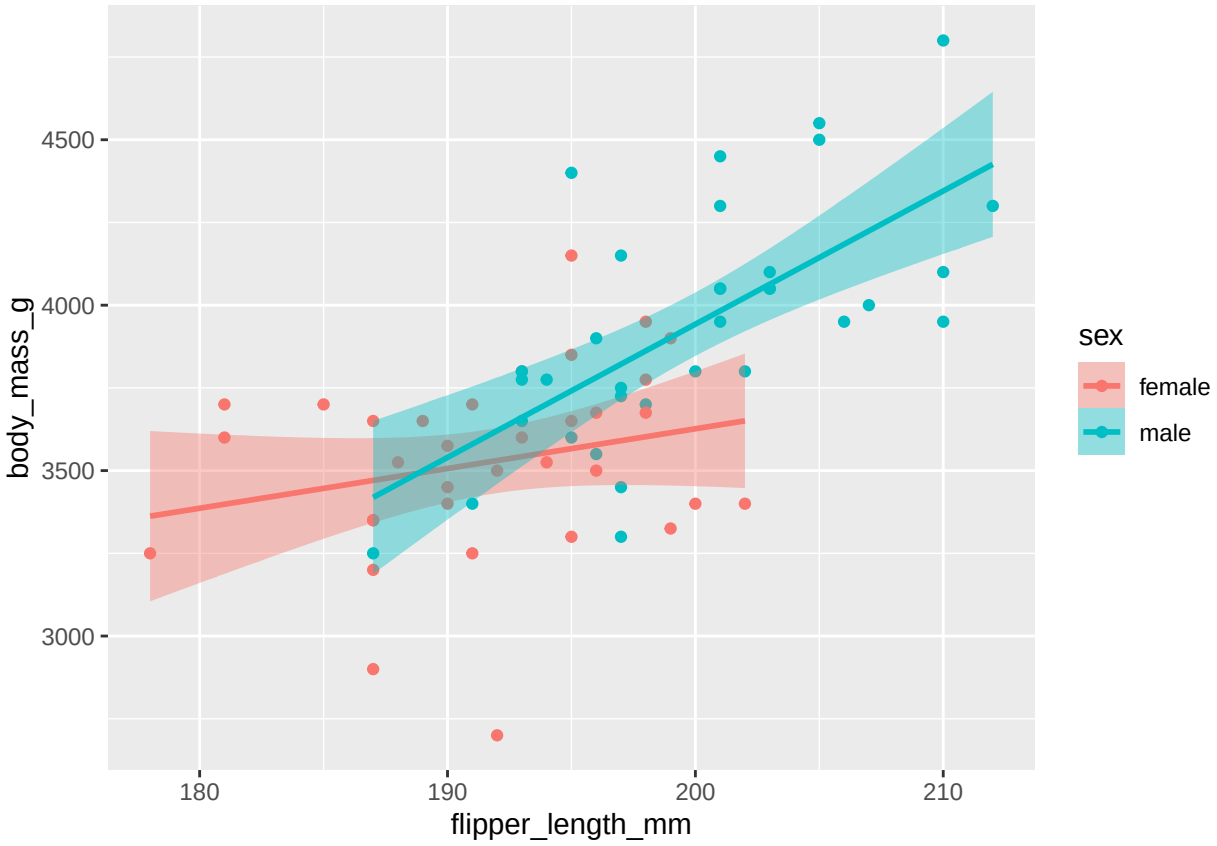
```
##
## Call:
## lm(formula = bill_depth_mm ~ bill_length_mm + species, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4579 -0.6814 -0.0431  0.5441  3.5994
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      10.56526    0.69093  15.291  < 2e-16 ***
## bill_length_mm    0.20044    0.01768  11.337  < 2e-16 ***
## speciesChinstrap -1.93308    0.22572  -8.564 4.26e-16 ***
## speciesGentoo    -5.10332    0.19440 -26.252  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9543 on 329 degrees of freedom
## Multiple R-squared:  0.7673, Adjusted R-squared:  0.7652
## F-statistic: 361.6 on 3 and 329 DF,  p-value: < 2.2e-16
```

Question 4:

1. Create a new data frame contains only the Chinstrap species.
2. Fit a linear regression model in which the body mass is the response and flipper length and gender are the predictors.
3. Visualize your model.

Solution

```
library(palmerpenguins)
data("penguins", package = "palmerpenguins")
penguins <- drop_na(penguins)
penguins.c<-penguins %>% filter(species%in%c("Chinstrap"))
ggplot(penguins.c,aes(flipper_length_mm,body_mass_g,fill=sex,color=sex))+geom_point()+geom_smooth(metho
```




```
fit.2<-lm(body_mass_g~sex+flipper_length_mm:sex,data=penguins.c)
summary(fit.2)
```

```
##
## Call:
## lm(formula = body_mass_g ~ sex + flipper_length_mm:sex, data = penguins.c)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -830.38 -171.79  -13.33  174.34  658.82
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1224.901   1613.134   0.759   0.4504
## sexmale                     -5336.249   2285.691  -2.335   0.0227 *
## sexfemale:flipper_length_mm    12.008      8.410   1.428   0.1582
## sexmale:flipper_length_mm      40.269      8.097   4.974 5.22e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 278 on 64 degrees of freedom
## Multiple R-squared:  0.5003, Adjusted R-squared:  0.4769
## F-statistic: 21.36 on 3 and 64 DF,  p-value: 1.059e-09
```

Question 5:

1. Create a new data frame contains all observations of the Biscoe island.
2. Sort the new data by body_mass_g.
3. Sort the new data by gender and body_mass_g.
4. For the new data:

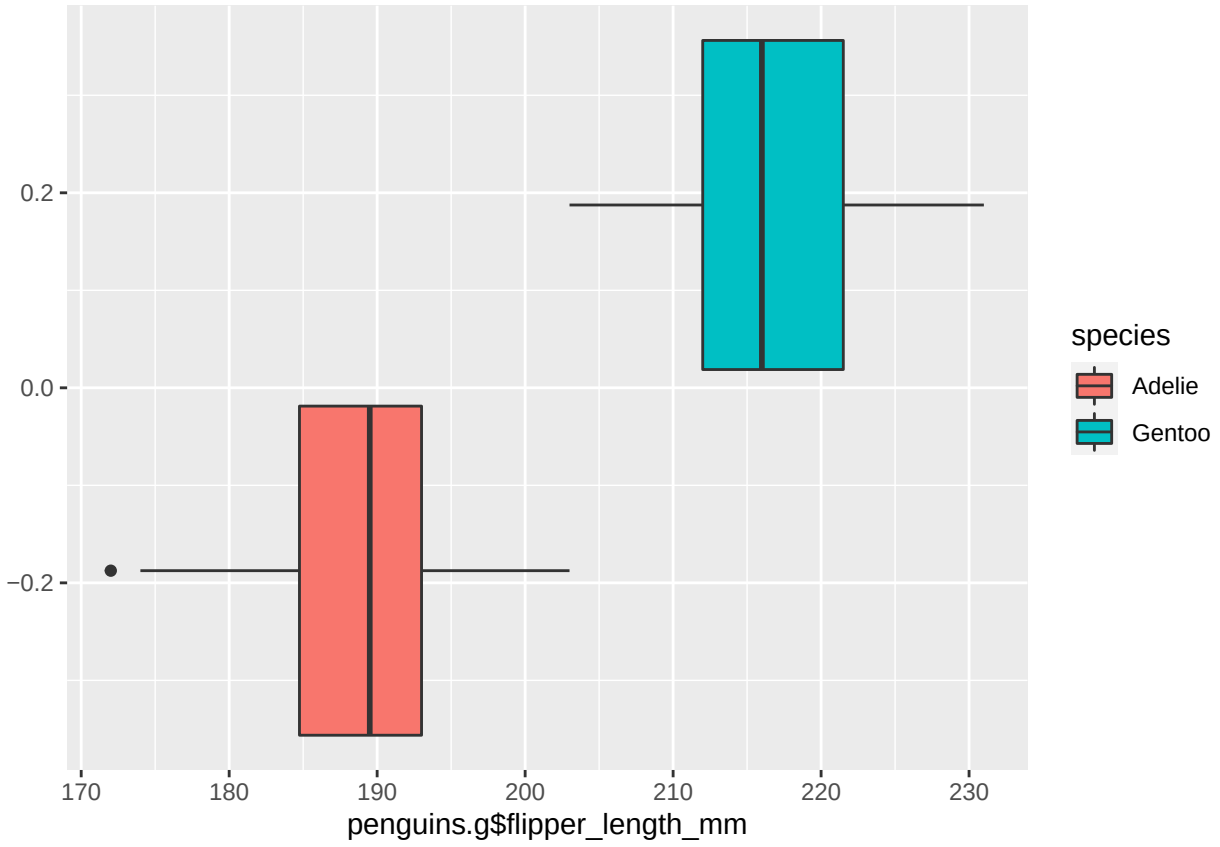
(a) Test that the mean flipper length is equal cross the species levels.
(b) Produce the figure below.
(c) Use the chi-square test to test the hypothesis that species and gender are independent and produce the table below.

Solution

```
library(palmerpenguins)
data("penguins", package = "palmerpenguins")
penguins <- drop_na(penguins)
#head(penguins)
penguins.g<- penguins %>% filter(island%in%c("Biscoe"))
#penguins.g$species
#penguins.g$island
#quantile(penguins.g)
t.test(penguins.g$flipper_length_mm~penguins.g$species)
```

```
##
##  Welch Two Sample t-test
##
## data:  penguins.g$flipper_length_mm by penguins.g$species
## t = -24.091, df = 75.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Adelie and group Gentoo is not equal t
## 95 percent confidence interval:
##  -30.79132 -26.08836
## sample estimates:
## mean in group Adelie mean in group Gentoo
##             188.7955             217.2353
```

```
ggplot(penguins.g,aes(penguins.g$flipper_length_mm,fill=species))+geom_boxplot()
```




```
xx<-table(penguins.g$sex,penguins.g$species)
xx1<-xx[,-c(2)]
xx1
```

```
##
##          Adelie Gentoo
##   female     22     58
##   male       22     61
```

```
chisq.test(penguins.g$sex,penguins.g$species)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  penguins.g$sex and penguins.g$species
## X-squared = 1.1478e-30, df = 1, p-value = 1
```

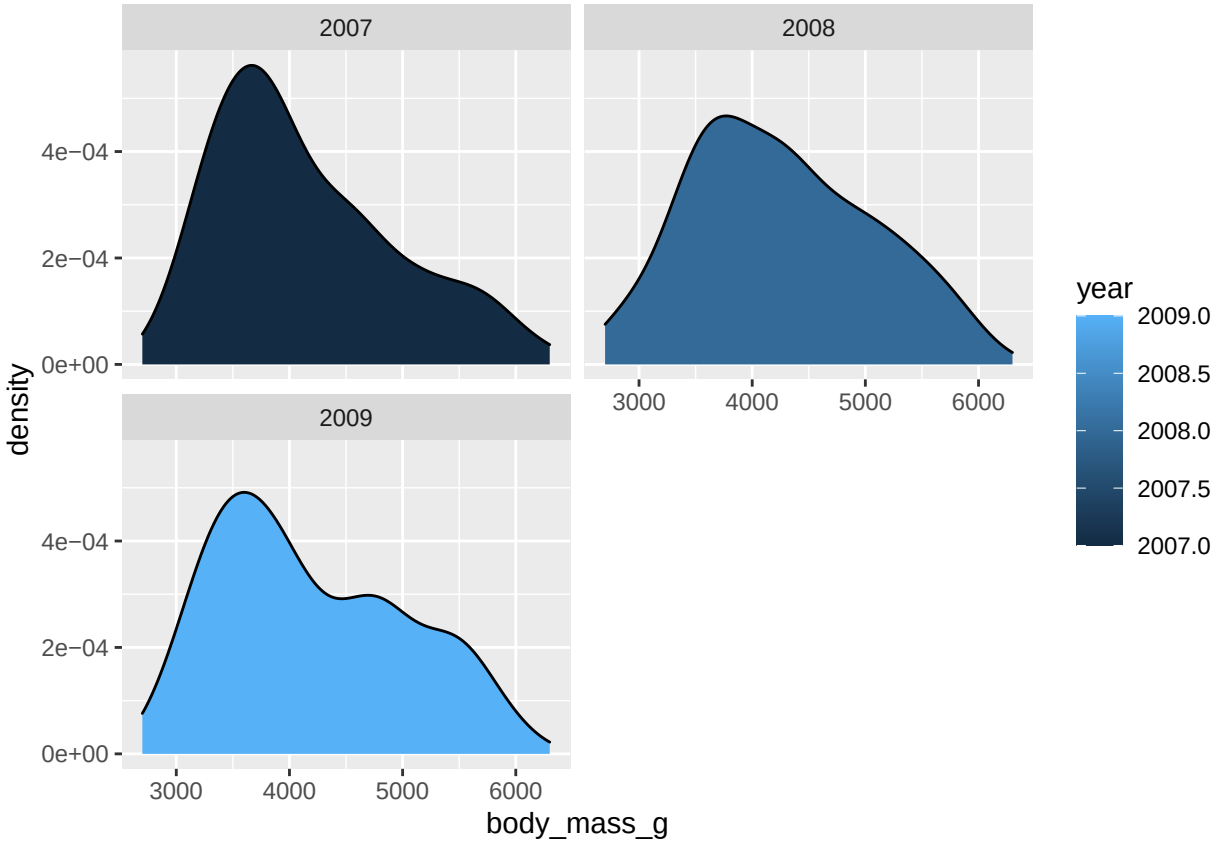
Question 6:

1. In the penguins data, how many observations there are at each year.
2. Produce the multi-way density and histogram below.

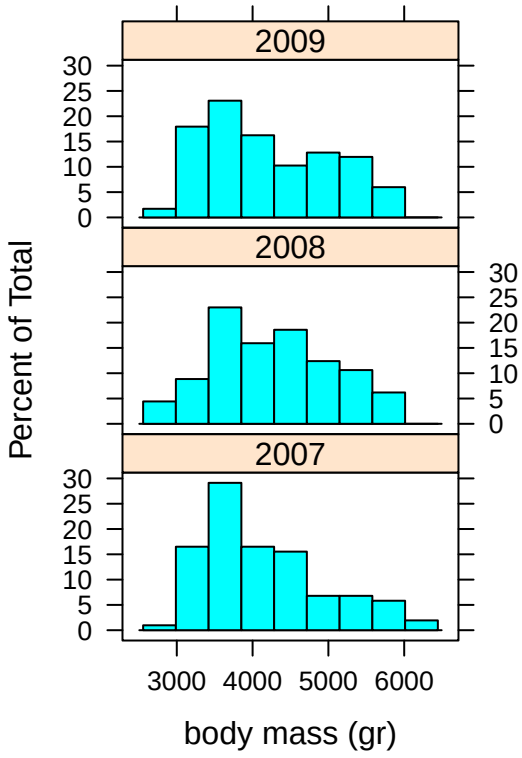
3. Create a new dataset with observations before 2009 and include in the data the year, sex, body_mass_g.
4. Define a new indicator variable which take the value of 1 for observations with body mass > 4000 and zero otherwise.
5. Produce a 2×2 table (year X indicator) for the overall sample, for male and for female (separately).

Solution

```
library(palmerpenguins)
data("penguins", package = "palmerpenguins")
penguins <- drop_na(penguins)
ggplot(penguins, aes(body_mass_g,fill = year)) +
geom_density() +
facet_wrap(~year,ncol = 2)
```




```
histogram(~ penguins$body_mass_g | as.factor(penguins$year),
        data=penguins, layout = c(1, 3),
          aspect = 0.5,xlab = "body mass (gr)")
```




```
#table(penguins$year)
#quantile(penguins$body_mass_g)
penguins.g<- penguins %>% filter(year < 2009)
body_m<-penguins.g$body_mass_g*0+1
body_m[penguins.g$body_mass_g <4001]<-0
table(body_m,penguins.g$year) # all sample
```

```
##
## body_m 2007 2008
##      0   56   49
##      1   47   64
```

```
penguins.gf<- penguins.g %>% filter(sex=="female")
#dim(penguins.gf)
body_mf<-body_m[penguins.g$sex=="female"]
#length(body_mf)
table(body_mf,penguins.gf$year) #female
```

```
##
## body_mf 2007 2008
##       0   34   35
##       1   17   21
```

```r
penguins.gm<- penguins.g %>% filter(sex=="male")
#dim(penguins.gm)
body_mm<-body_m[penguins.g$sex=="male"]
#length(body_mm)
table(body_mm,penguins.gm$year) #male
```

```
##
## body_mm 2007 2008
##       0   22   14
##       1   30   43
```