Programming in R: Solution for practical session 7a, Introduction toStatistical Inference using R

Juan Vanegas

2023-12-28

General Information

- The practical session is focus on statistical modeling and consists of 6 questions in which you are asked to conduct an analysis of a dataset.
- 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

R functions that are used in this practical session include: - t.test()). - geom_boxplot(). - geom_density(). - filter(). - table(). - chisq.test(). - function(). - drop_na()

Introduction Q1-Q6:

In questions Q1-Q6 we focus on the penguins dataset which available as a part of the palmerpenguins Rpackage. To assess the data, you need ti install the package. More information about the data is availablein https://allisonhorst.github.io/palmerpenguins/. For all the questions (Q1-Q6) we conduct a completecase analysis, i.e., all observations should not have missing values. In each question, start form the originalpenguins dataset after removing the missing values.

Question 1:

1. How many variables there are in the data ? How many male and female there are ? The first few lines of the data are shown below.

Solution

```
library(palmerpenguins)
data("penguins", package = "palmerpenguins")
penguins <- drop_na(penguins)
dim(penguins)
```

```
## [1] 333   8
```

```
table(penguins$sex)
```

```
##
## female   male
##    165    168
```

```
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##   <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
## 1 Adelie  Torgersen           39.1          18.7               181        3750
## 2 Adelie  Torgersen           39.5          17.4               186        3800
## 3 Adelie  Torgersen           40.3          18                 195        3250
## 4 Adelie  Torgersen           36.7          19.3               193        3450
## 5 Adelie  Torgersen           39.3          20.6               190        3650
## 6 Adelie  Torgersen           38.9          17.8               181        3625
## # i 2 more variables: sex <fct>, year <int>
```

2. In this question we focus on the variable flipper_length_mm for the female.

- Calculate the overall sample mean (for female).
- Test the hypothesis that the mean of the variable flipper_length_mm (for female) is equal to 198 against two sided alternative.
- The Boxplot and density estimate visualize the distribution of the variable flipper_length_mm. Produce these figures. Do you think that there is a problem to conduct the above test, taking into account the information that you observed in the figures?

Solution

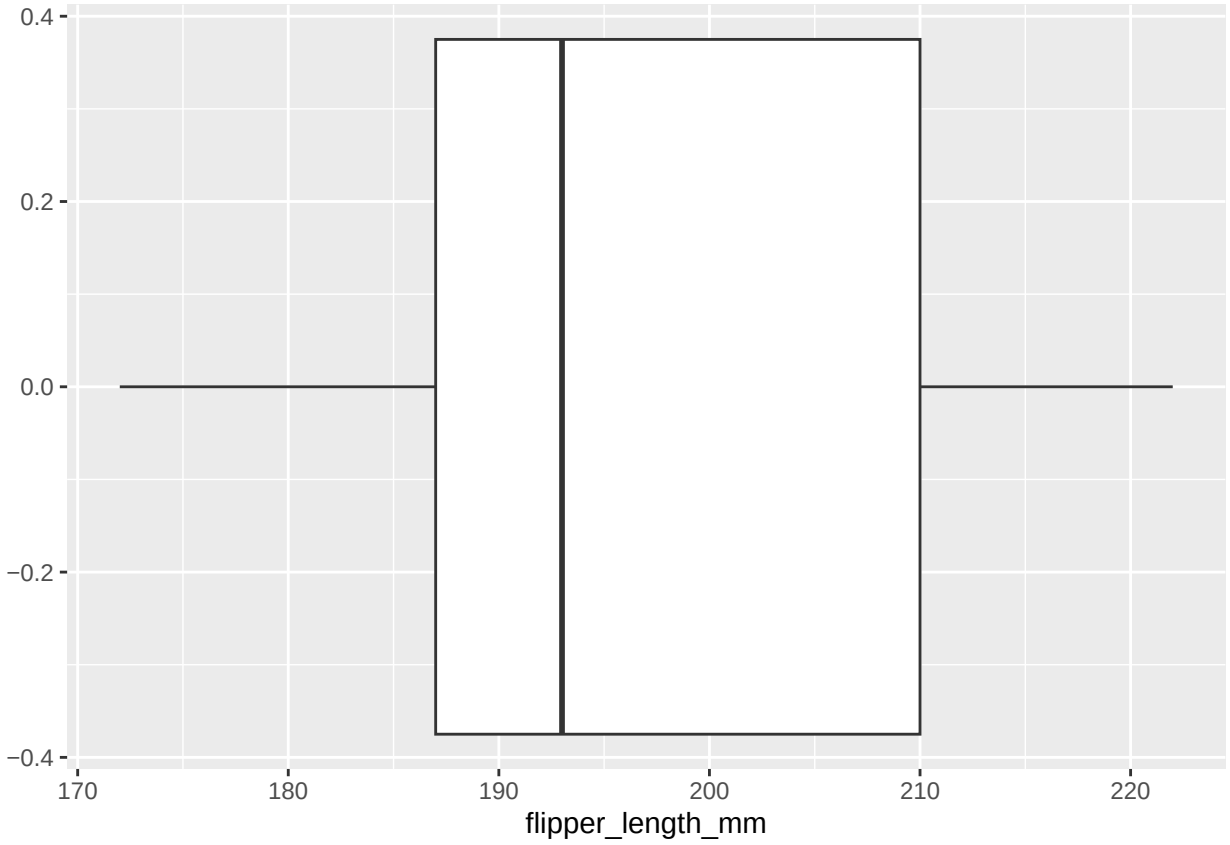
```
penguins.f<-penguins %>% filter(sex=="female")
#print(penguins.f)
mean(penguins.f$flipper_length_mm)
```

```
## [1] 197.3636
```

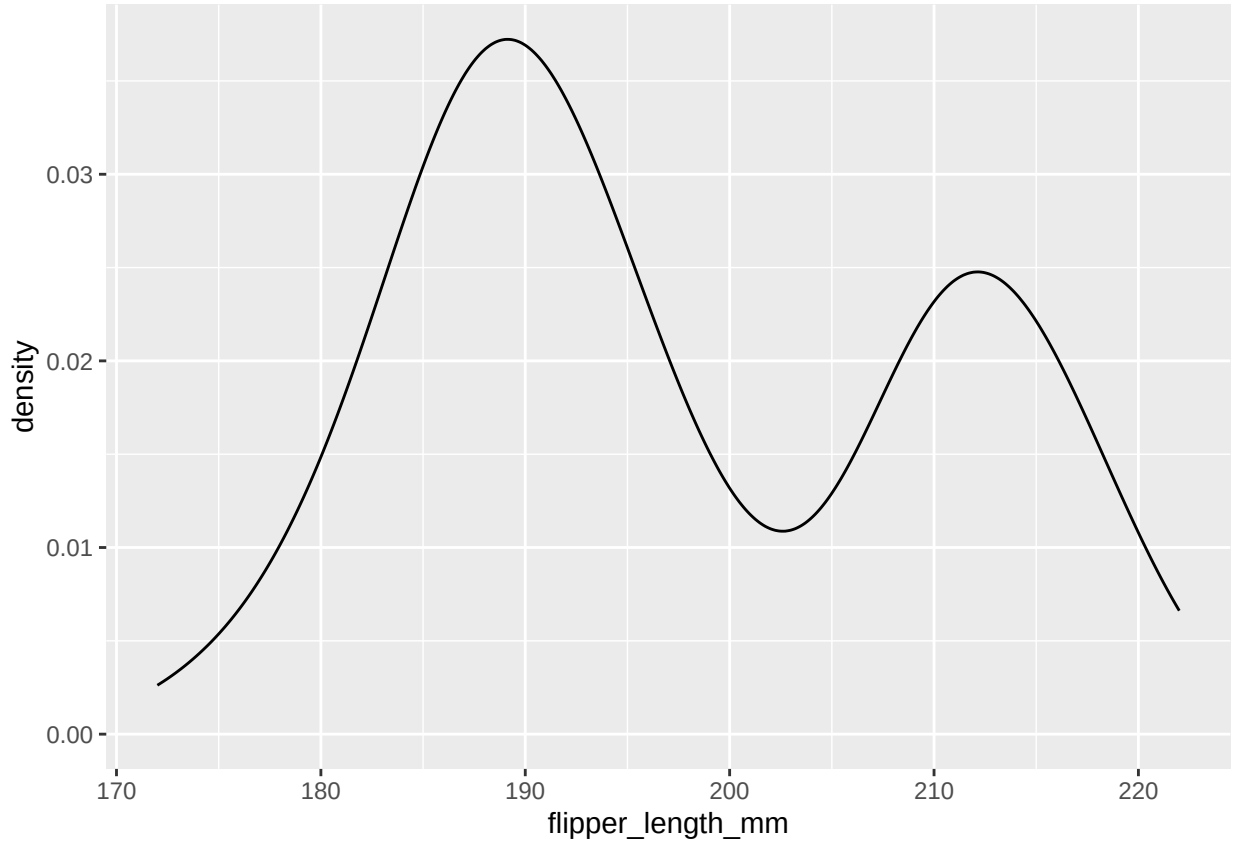
```r
t.test(penguins.f$flipper_length_mm,alternative = c("two.sided"),
mu = 198,conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  penguins.f$flipper_length_mm
## t = -0.6539, df = 164, p-value = 0.5141
## alternative hypothesis: true mean is not equal to 198
## 95 percent confidence interval:
##  195.4421 199.2852
## sample estimates:
## mean of x
##  197.3636
```

```r
ggplot(penguins.f,aes(flipper_length_mm))+geom_boxplot()
```



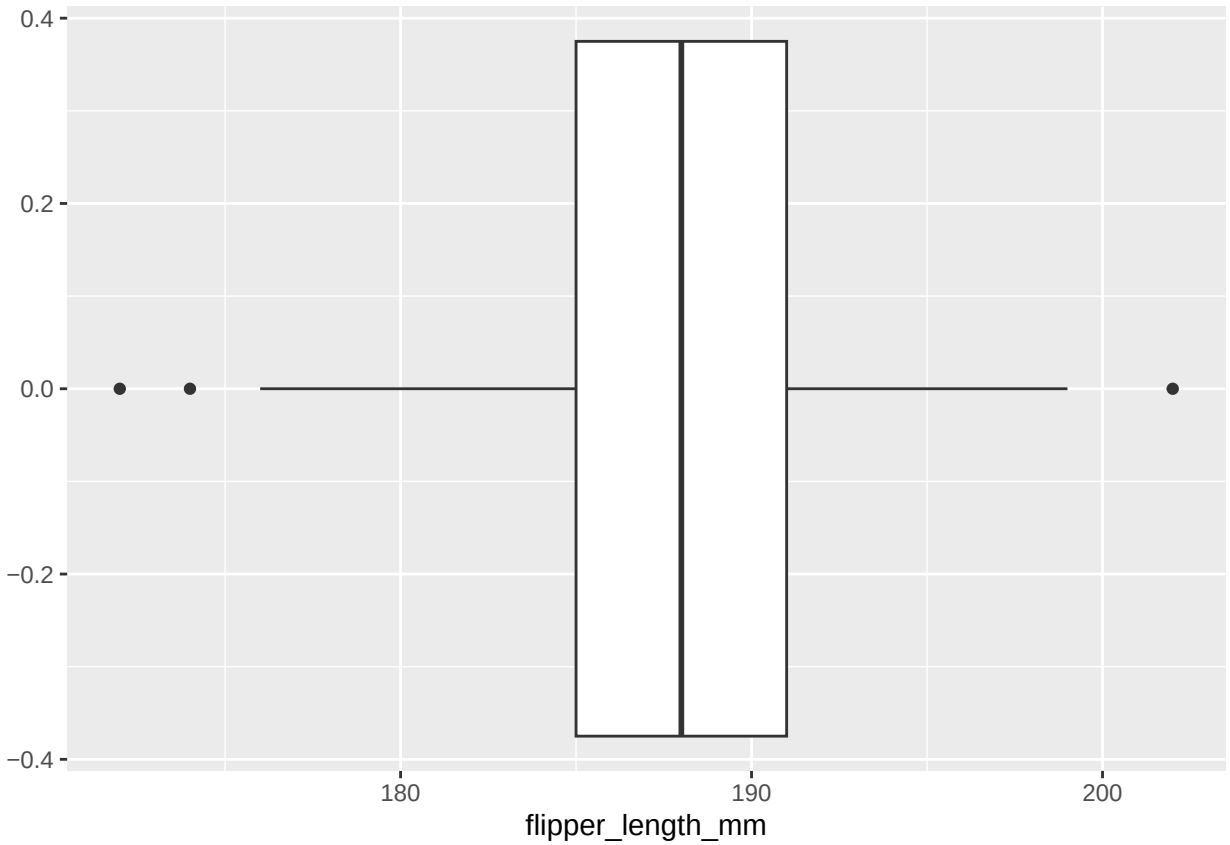

```r
ggplot(penguins.f,aes(flipper_length_mm))+geom_density()
```



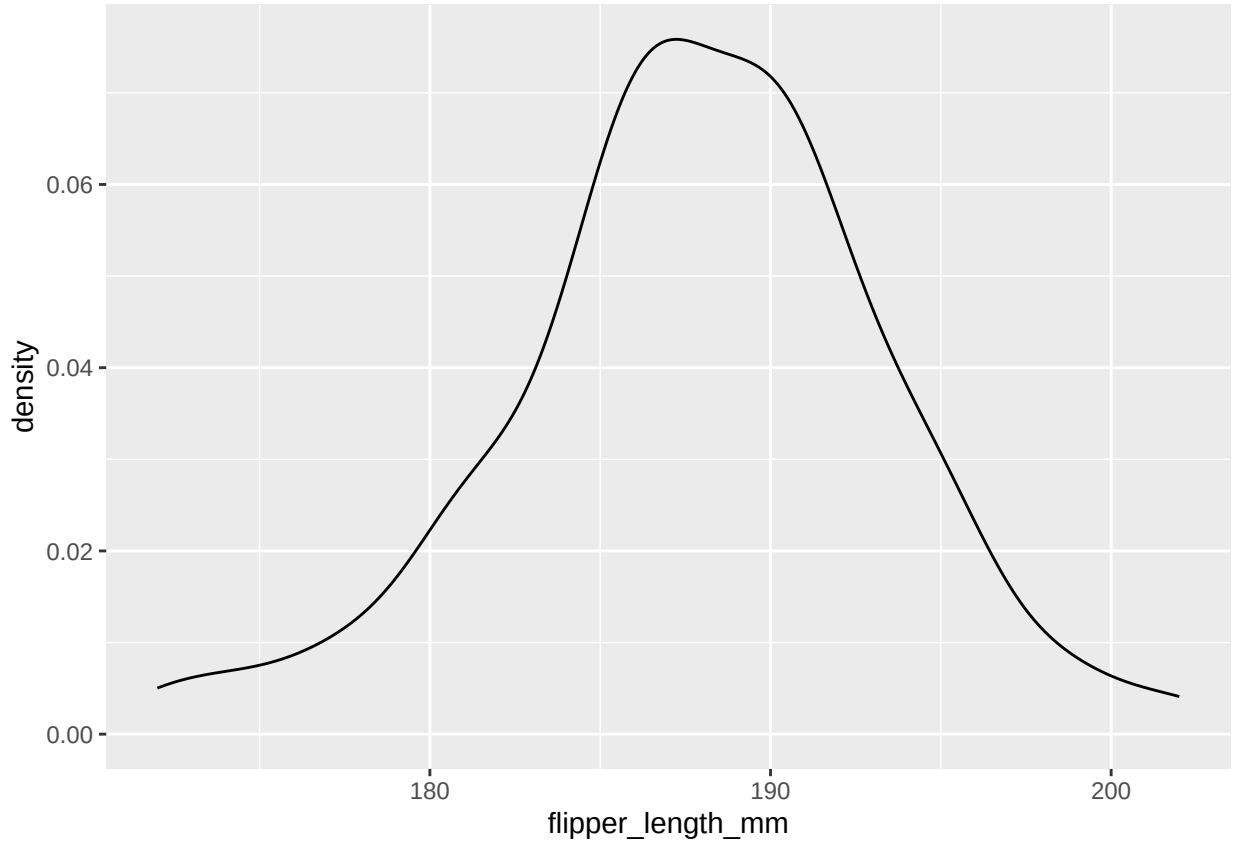
Question 2: Taking into account the problem that was observed in Q1, in this question we focus on the variable flipper_length_mm for female of the species Adelie. 1. Produce the boxplot and density estimate, shown below, for the variable flipper_length_mm (for female from the species Adelie). 2. How many observations are included in the analysis ? 3. Test the hypothesis that the mean of the variable flipper_length_mm (for female from the species Adelie) is equal to 190 against a two sided alternative and construct a 95% confidence interval for the mean.

Solution

```
library(palmerpenguins)
data("penguins", package = "palmerpenguins")
penguins <- drop_na(penguins)
#dim(penguins)
#head(penguins)
penguins.f<-penguins %>% filter(sex=="female")
penguins.fa<-penguins.f %>% filter(species=="Adelie")
#dim(penguins.fa)
ggplot(penguins.fa,aes(flipper_length_mm))+geom_boxplot()
```




```
ggplot(penguins.fa,aes(flipper_length_mm))+geom_density()
```




```
t.test(penguins.fa$flipper_length_mm,alternative = c("two.sided"),
mu = 190,conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  penguins.fa$flipper_length_mm
## t = -3.3679, df = 72, p-value = 0.001219
## alternative hypothesis: true mean is not equal to 190
## 95 percent confidence interval:
##  186.4891 189.0999
## sample estimates:
## mean of x
##  187.7945
```

Question 3:

In this question we focus on the variable bill_length_mm for the species Adelie and Chinstrap for the female population. 1. Test the hypothesis that the mean bill_length_mm for Adelie female is equal to the mean of the Chinstrap female. 2. Construct a 95% confidence interval for the mean difference. 3. Calculate the mean bill_length_mm by species. 4. Produce the figures below.

Solution

```
library(palmerpenguins)
data("penguins", package = "palmerpenguins")
```

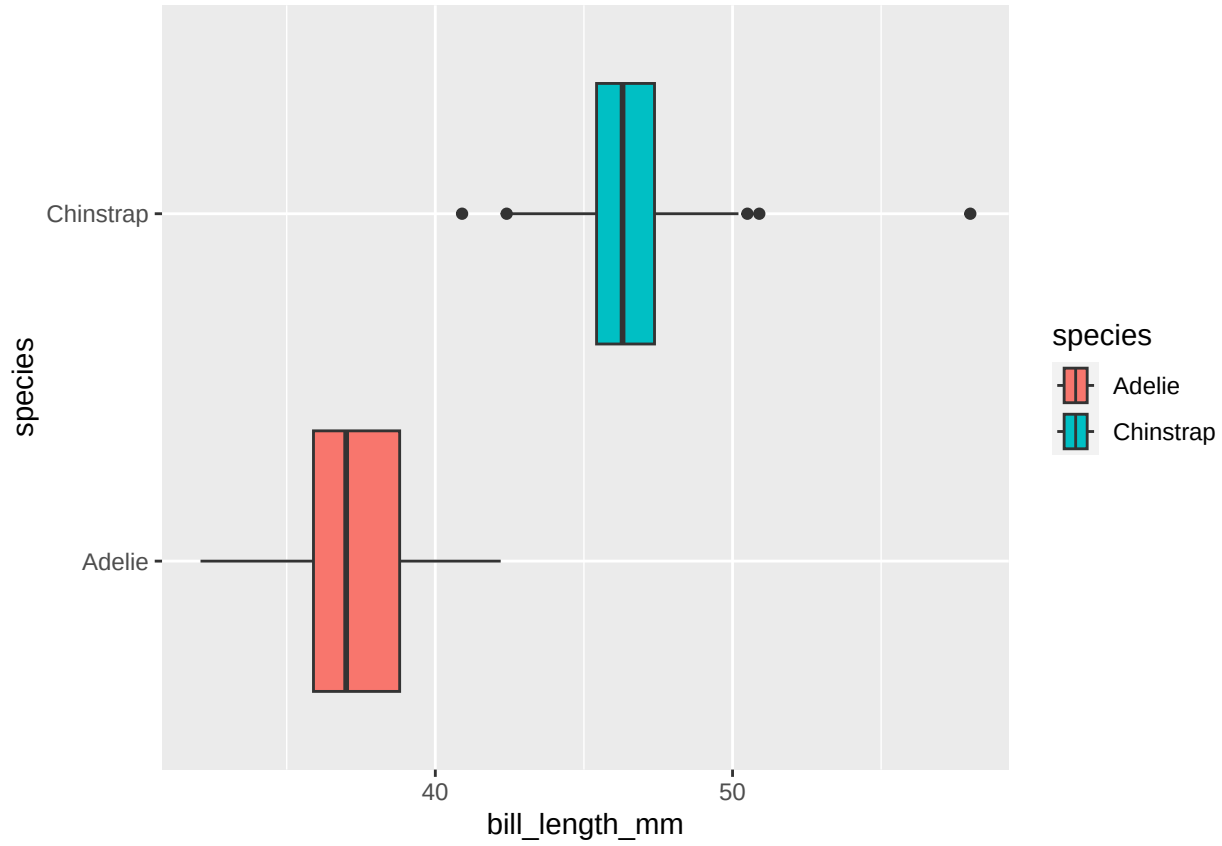
```r
penguins <- drop_na(penguins)
#dim(penguins)
#head(penguins)
#print(penguins$species)
penguins.f<-penguins %>% filter(sex=="female")
#print(penguins.f)
penguins.fac<-penguins.f %>% filter(species%in%c("Adelie","Chinstrap"))
dim(penguins.fac)
```

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

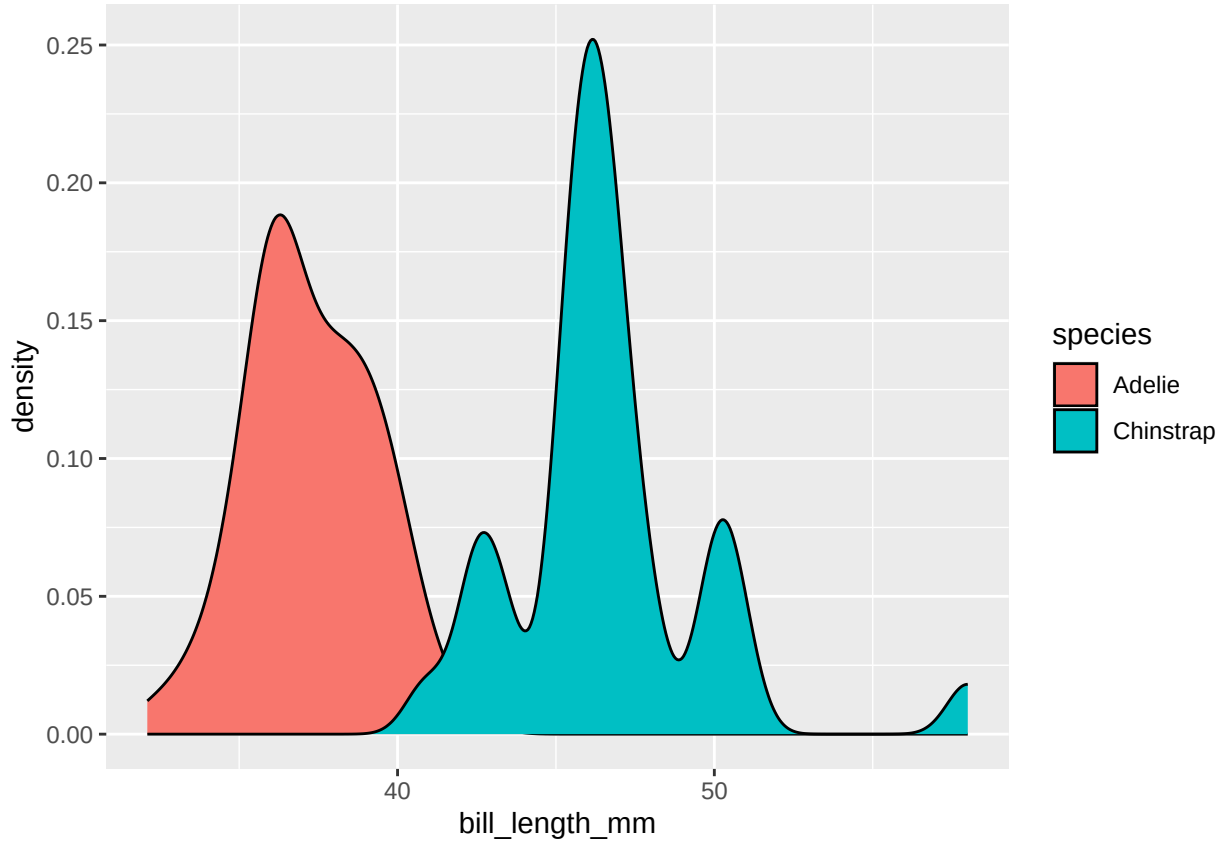
```r
tapply(penguins.fac$bill_length_mm,penguins.fac$species,mean)
```

```
##    Adelie Chinstrap    Gentoo
##  37.25753  46.57353        NA
```

```r
ggplot(penguins.fac,aes(bill_length_mm,species,fill=species))+geom_boxplot()
```




```r
ggplot(penguins.fac,aes(bill_length_mm,fill=species))+geom_density()
```




```
t.test(penguins.fac$bill_length_mm~penguins.fac$species,alternative = c("two.sided"),
mu = 0,conf.level = 0.95,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  penguins.fac$bill_length_mm by penguins.fac$species
## t = -18.535, df = 105, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Adelie and group Chinstrap is not equa
## 95 percent confidence interval:
##  -10.312585  -8.319405
## sample estimates:
##    mean in group Adelie mean in group Chinstrap
##                37.25753                46.57353
```

Question 4:

In this question we focus on the variable bill_length_mm for all the species (for both male and female). 1. Define an indicator variable which takes the value of 1 if bill_depth_mm < 18 and zero otherwise. Note that for all subjects (male and female) for which the bill depth is higher than 18 mm, the new variable is equal to 1. We define these subjects as subjects with large bill depth. If you define the vector correctly, you should have the zero/one vector shown below.

Solution

```
library(palmerpenguins)
data("penguins", package = "palmerpenguins")
penguins <- drop_na(penguins)
#quantile(penguins$bill_depth_mm)
#print(penguins$bill_depth_mm)
bill_depth_1<-penguins$bill_depth_mm*0
bill_depth_1[penguins$bill_depth_mm < 18]<-1
print(bill_depth_1)
```

```
##   [1] 0 1 0 0 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0
##  [38] 0 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 0 0 1 0 1 0 1 0 1 0
##  [75] 1 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 0 0 1 0 1 0 1 0 0 0 1 0 1 0 1 0 1 0 0 0 1
## [112] 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 0 0 1 1 1 0 1 1 1 1 1 1 1 0 0 0 1 0 1 0 1 1
## [149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [186] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [223] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [260] 1 1 1 1 1 1 1 0 0 0 0 1 0 0 0 0 1 0 1 0 1 0 0 1 0 0 1 1 1 0 1 0 0 0 1 0 1
## [297] 0 1 0 0 1 0 0 1 0 0 1 0 1 1 0 1 0 0 1 1 0 1 0 1 0 1 0 0 0 0 1 1 0 0 0 0 0
```

2. Produce a 2×2 table of bill depth (small/large) by gender (see below).
3. Conduct a chi-square test to test the hypothesis that indicator for bill depth is independent on gender.

Solution

```
table(bill_depth_1,penguins$sex)
```

```
##
## bill_depth_1 female male
##            0     32  100
##            1    133   68
```

```
chisq.test(bill_depth_1,penguins$sex)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  bill_depth_1 and penguins$sex
## X-squared = 54.363, df = 1, p-value = 1.667e-13
```

Question 5:

In this question we focus on the indicator variable for bill depth defined in Q4. 1. Estimate the proportion of female with large bill depth (for which the indicator is equal to 1). 2. Estimate the proportion of male with large bill depth (for which the indicator is equal to 1). 3. Test the hypothesis that the proportion of female with large bill depth (indicator equal to 1 is) equal to the proportion of male with large bill depth. You need to program the procedure by yourself and to produce the output below. Note that the output should contains the sample difference between the proportions, the Z score and the P value (see the data frame below).

Solution

```
library(palmerpenguins)
data("penguins", package = "palmerpenguins")
penguins <- drop_na(penguins)
bill_depth_1<-penguins$bill_depth_mm*0
bill_depth_1[penguins$bill_depth_mm < 18]<-1
```

```r
bill_depth_1.f<-bill_depth_1[penguins$sex=="female"]
nf<-length(bill_depth_1.f)
pf<-sum(bill_depth_1.f)/nf
pf #p(female)
```

```
## [1] 0.8060606
```

```r
bill_depth_1.m<-bill_depth_1[penguins$sex=="male"]
nm<-length(bill_depth_1.m)
pm<-sum(bill_depth_1.m)/nm
pm #p(female)
```

```
## [1] 0.4047619
```

```r
pool.sd<-sqrt((pm*(1-pm)/nm)+(pf*(1-pf)/nf))
z<-(pf-pm)/pool.sd
z #z-score
```

```
## [1] 8.223159
```

```r
p.val<-(1-pnorm(z,0,1))*2
diff<-pf-pm
stat<-c("Difference","Z-score","P.value")
value<-c(diff,z,p.val)
data.frame(stat,value)
```

```
##         stat        value
## 1 Difference 4.012987e-01
## 2    Z-score 8.223159e+00
## 3    P.value 2.220446e-16
```

Question 6:

In this question you need to write a R function that conduct the analysis that you program in Q5. The input of the function are two vectors: the indicator (x) and the gender (y). The output of the function is shown below.

Solution

```r
library(palmerpenguins)
data("penguins", package = "palmerpenguins")
penguins <- drop_na(penguins)
bill_depth_1<-penguins$bill_depth_mm*0
bill_depth_1[penguins$bill_depth_mm < 18]<-1

my.test<-function(x,y)
{
  xx<-table(x,y)
  print(xx)
  x.1<-x[y=="female"]
  nf<-length(x.1)
  pf<-sum(x.1)/nf
  x.2<-x[y=="male"]
  nm<-length(x.2)
  pm<-sum(x.2)/nm#pm
  pool.sd<-sqrt((pm*(1-pm)/nm)+(pf*(1-pf)/nf))
```

```
  z<-(pf-pm)/pool.sd
  p.val<-(1-pnorm(z,0,1))*2
  diff<-pf-pm
  pall<-(sum(x.1)+sum(x.2))/(nf+nm)
  stat<-c("P(Overall)","P(Group1)","P(Group2)","Difference","Z-score","P.value")
  value<-c(pall,pf,pm,diff,z,p.val)
  print(data.frame(stat,value))
}

my.test(bill_depth_1,penguins$sex)
```

```
##    y
## x   female male
##   0     32  100
##   1    133   68
##         stat        value
## 1 P(Overall) 6.036036e-01
## 2  P(Group1) 8.060606e-01
## 3  P(Group2) 4.047619e-01
## 4 Difference 4.012987e-01
## 5    Z-score 8.223159e+00
## 6    P.value 2.220446e-16
```

Practical Session 7b

Question 2:

In this question we focus on the decathlon data is a part of the FactoMineR R package. More information can be found in https://rdrr.io/cran/FactoMineR/man/decathlon.html (https://rdrr.io/cran/FactoMineR/man/decathlon.html).

```
library(FactoMineR)
data(decathlon)
names(decathlon)
```

```
##  [1] "100m"         "Long.jump"    "Shot.put"     "High.jump"    "400m"
##  [6] "110m.hurdle" "Discus"       "Pole.vault"  "Javeline"     "1500m"
## [11] "Rank"         "Points"       "Competition"
```

We focus on the results of long and high jump which were obtained in the Olympics games, use the variable Competition to select the data. Explore and visualize the data. In your analysis focus on the following aspects: 1. The dimension of the reduced data. 2. The correlation between the the results in long and high jump (use both numerical and graphical displays). 3. A table with the median, minimum and maximum by event. - Produce a density plot for the distribution of long and high jump in the Olympics. - Test the hypothesis that the mean long jump is equal to 7 against a two sided hypothesis (use significant levels of 5%) and construct a 95% confidence interval for the mean. **In your analysis use table captions and figure captions.**

Solution

```
# 1
decathlon.olympic<-decathlon[decathlon$Competition=="OlympicG",]
dim(decathlon.olympic)
```

```
## [1] 28 13
```

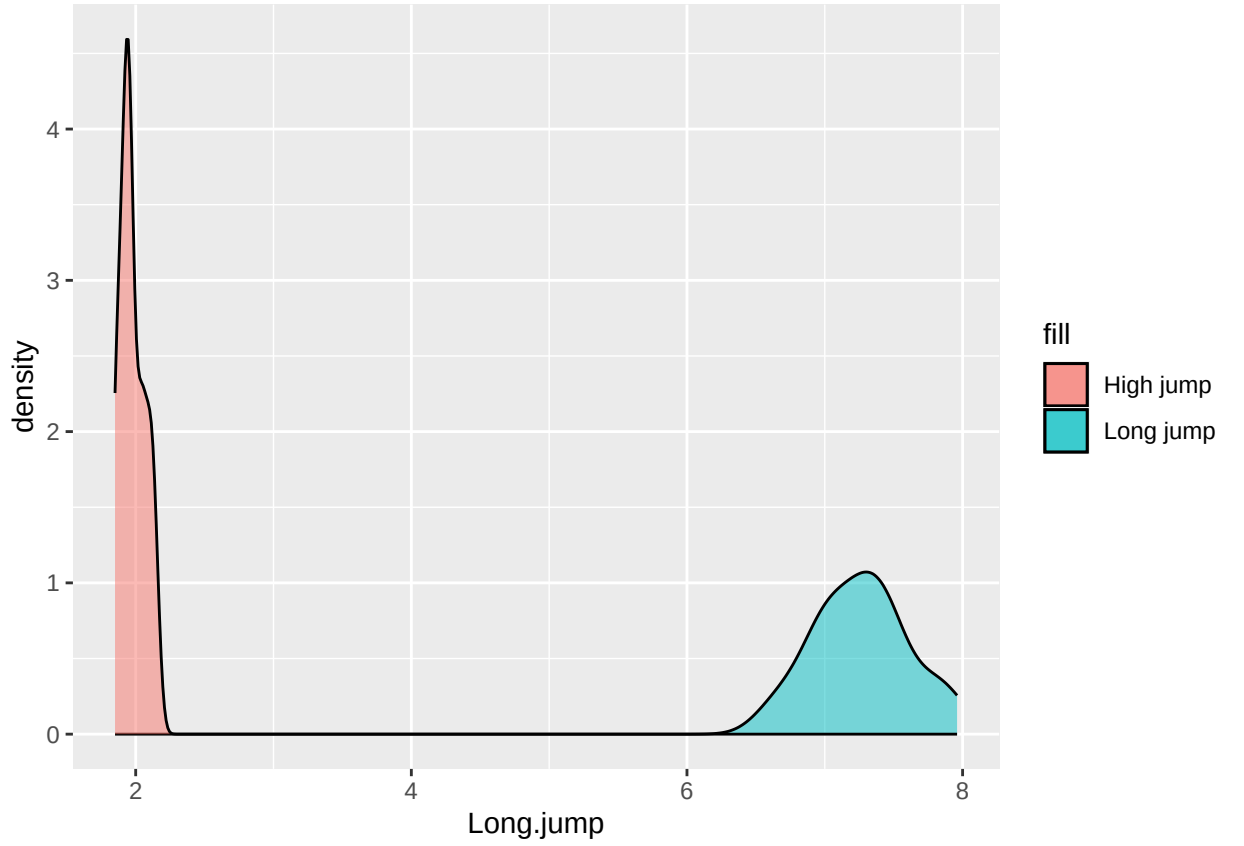
```r
# 2
cor(decathlon.olympic$Long.jump,decathlon.olympic$High.jump)
```

```
## [1] 0.3456621
```

```r
# 3
decathlon.olympic.dt <- as.data.table(decathlon.olympic, keep.rownames = TRUE) %>%
  select(-Competition, -Rank, -Points)
table.stats <- decathlon.olympic.dt %>%
  melt(id.vars = c("rn"), variable.name = "Event", value.name = "Time") %>%
  group_by(Event) %>%
  summarise(Median = median(Time), Min = min(Time), Max = max(Time))
table.stats
```

```
## # A tibble: 10 x 4
##    Event       Median    Min    Max
##    <fct>        <dbl>  <dbl>  <dbl>
##  1 100m          10.9   10.4   11.4
##  2 Long.jump      7.28   6.61   7.96
##  3 Shot.put      14.8   13.1   16.4
##  4 High.jump      1.94   1.85   2.15
##  5 400m          49.4   46.8   53.2
##  6 110m.hurdle   14.4   14.0   15.4
##  7 Discus        44.5   39.8   51.6
##  8 Pole.vault     4.7    4.2    5.4
##  9 Javeline      58.9   50.6   70.5
## 10 1500m        276.   263.   317
```

```r
# 4
library(ggplot2)
ggplot(decathlon.olympic) +
  geom_density(aes(x = Long.jump, fill = "Long jump"), alpha = 0.5) +
  geom_density(aes(x = High.jump, fill = "High jump"), alpha = 0.5)
```




```
# 5
t.test(decathlon.olympic$Long.jump, mu = 7, alternative = "two.sided", conf.level = 0.95)
```

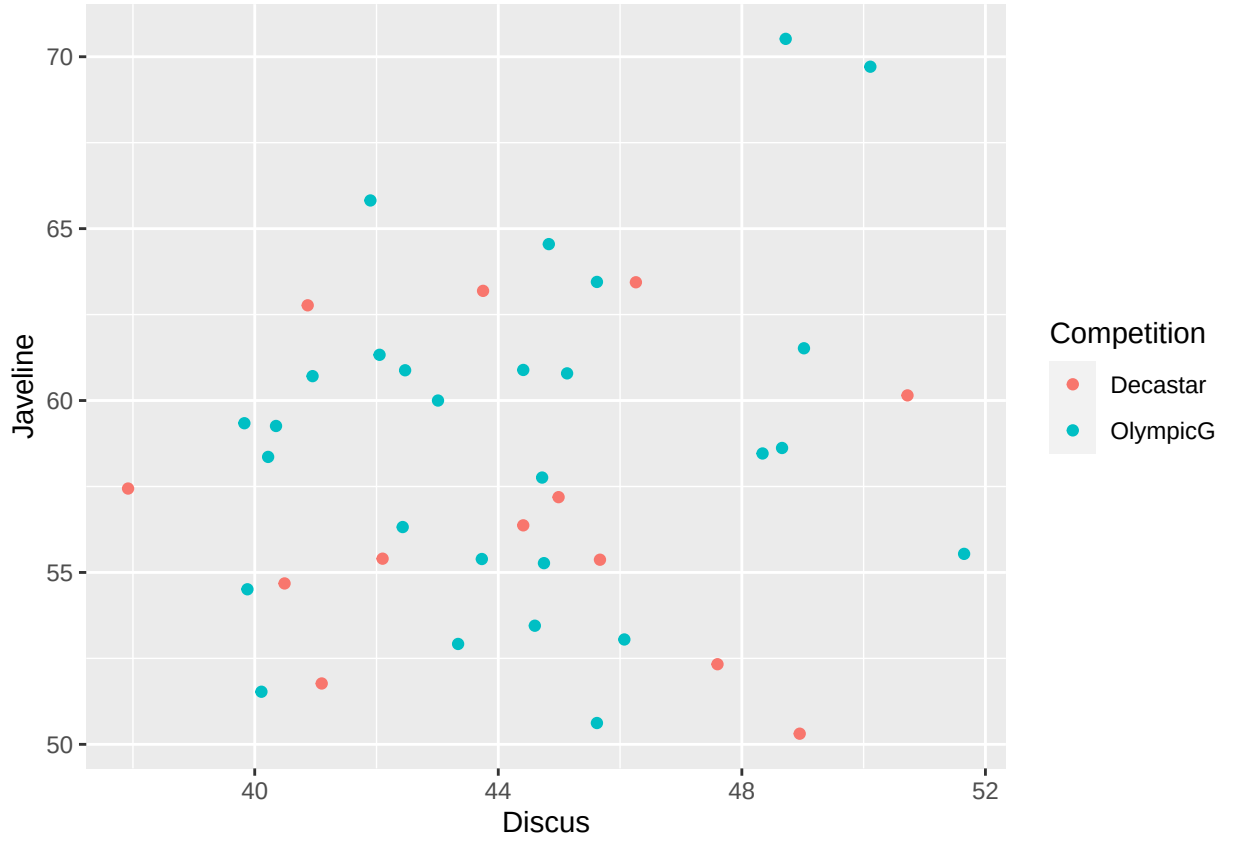
```
##
##  One Sample t-test
##
## data:  decathlon.olympic$Long.jump
## t = 4.1216, df = 27, p-value = 0.0003206
## alternative hypothesis: true mean is not equal to 7
## 95 percent confidence interval:
##  7.133436 7.397993
## sample estimates:
## mean of x
##  7.265714
```

Question 3

Our aim in this question is to compare between the results for Discus and Javeline obtained in the Decastar and the Olympics competitions. 1. Produce a scatterplot of Discus versus Javeline (colored by competition). 2. Add a regression line to the scatterplot above (competition). 3. Produce a plot with two panels in which data (Discus versus Javeline) obtained in the Decastar competition is presented in one panel and the results obtained in the Olympics in the second panel 4. Produce a table in which the mean and S.D (for each event) are presented by competition. produce a stripplot for the discuss and boxplot for Javeline. 5 Explore the correlation between Long jump and High jump by competition.

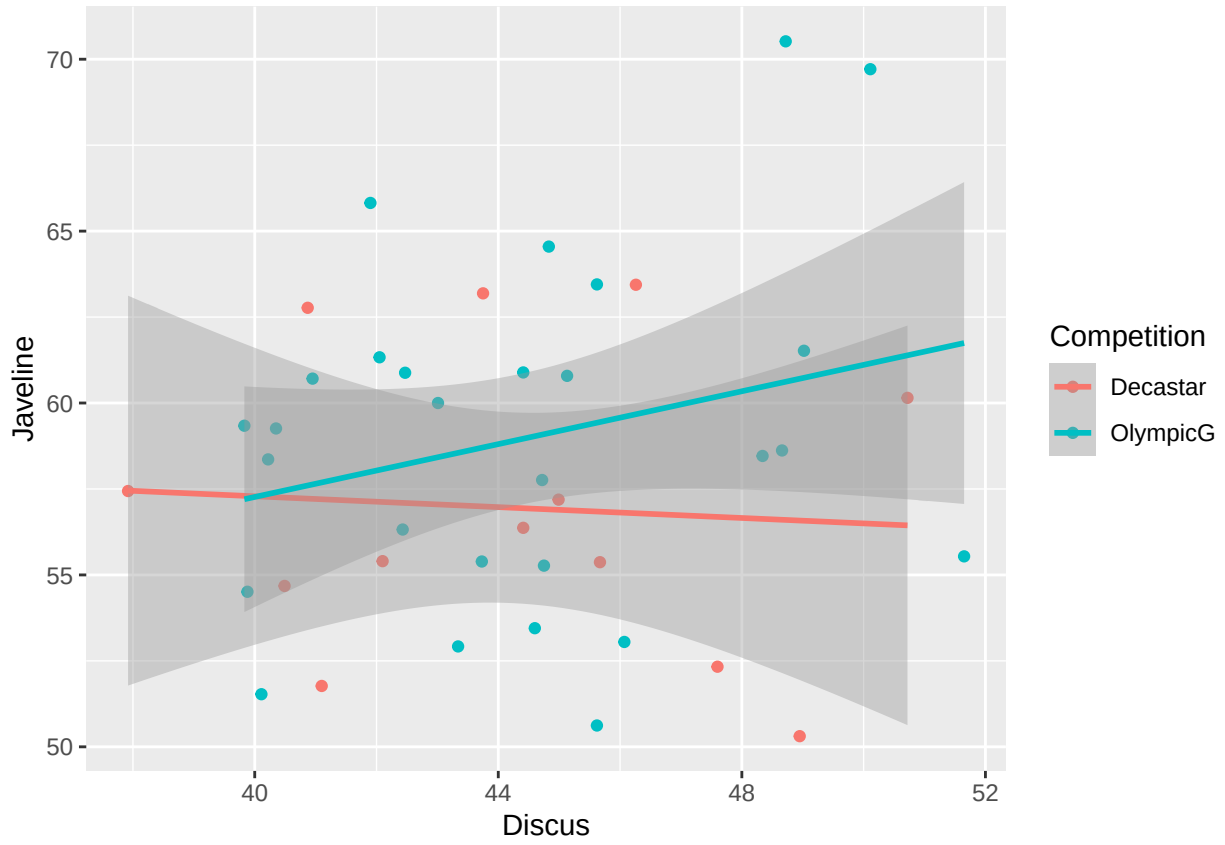
```
# 1
ggplot(decathlon) +
```

```
  geom_point(aes(x = Discus, y = Javeline, color = Competition))
```

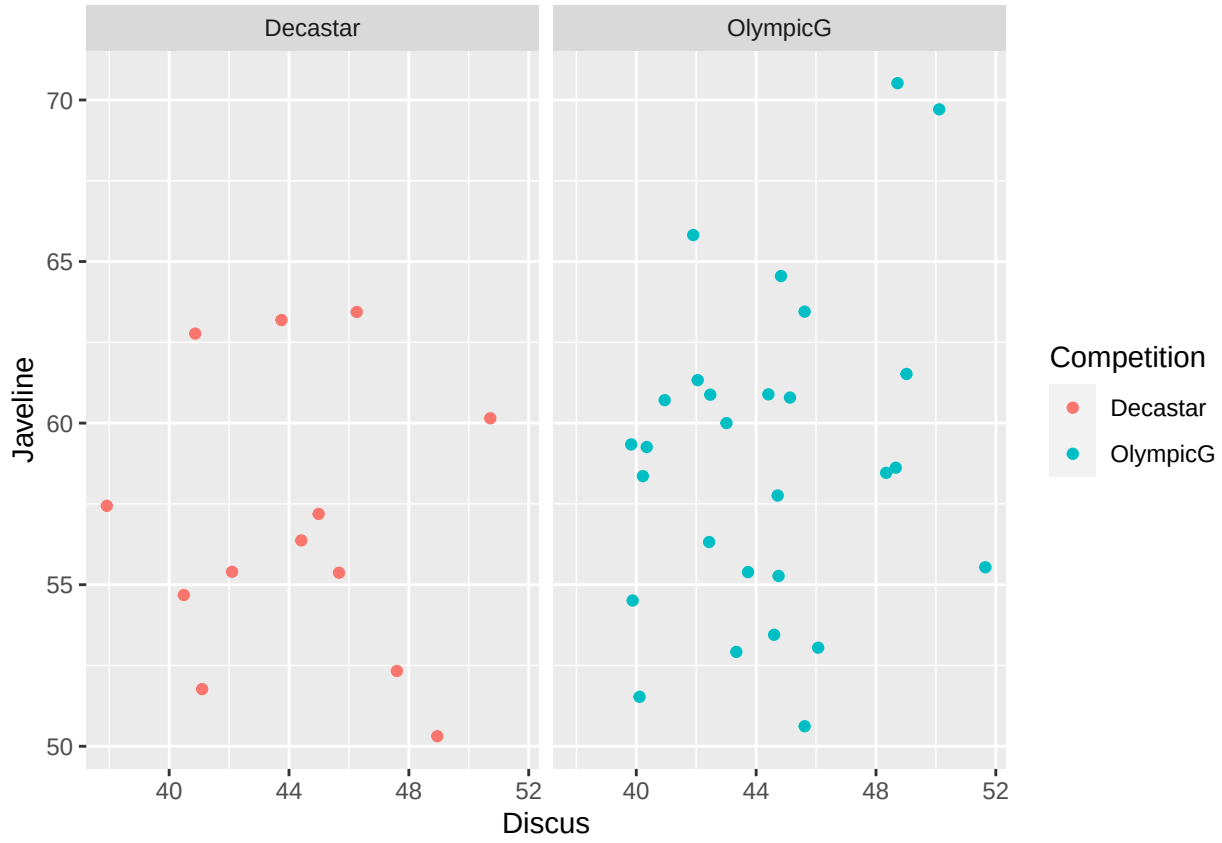



```
# 2
ggplot(decathlon) +
  geom_point(aes(x = Discus, y = Javeline, color = Competition)) +
  geom_smooth(aes(x = Discus, y = Javeline, color = Competition), method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



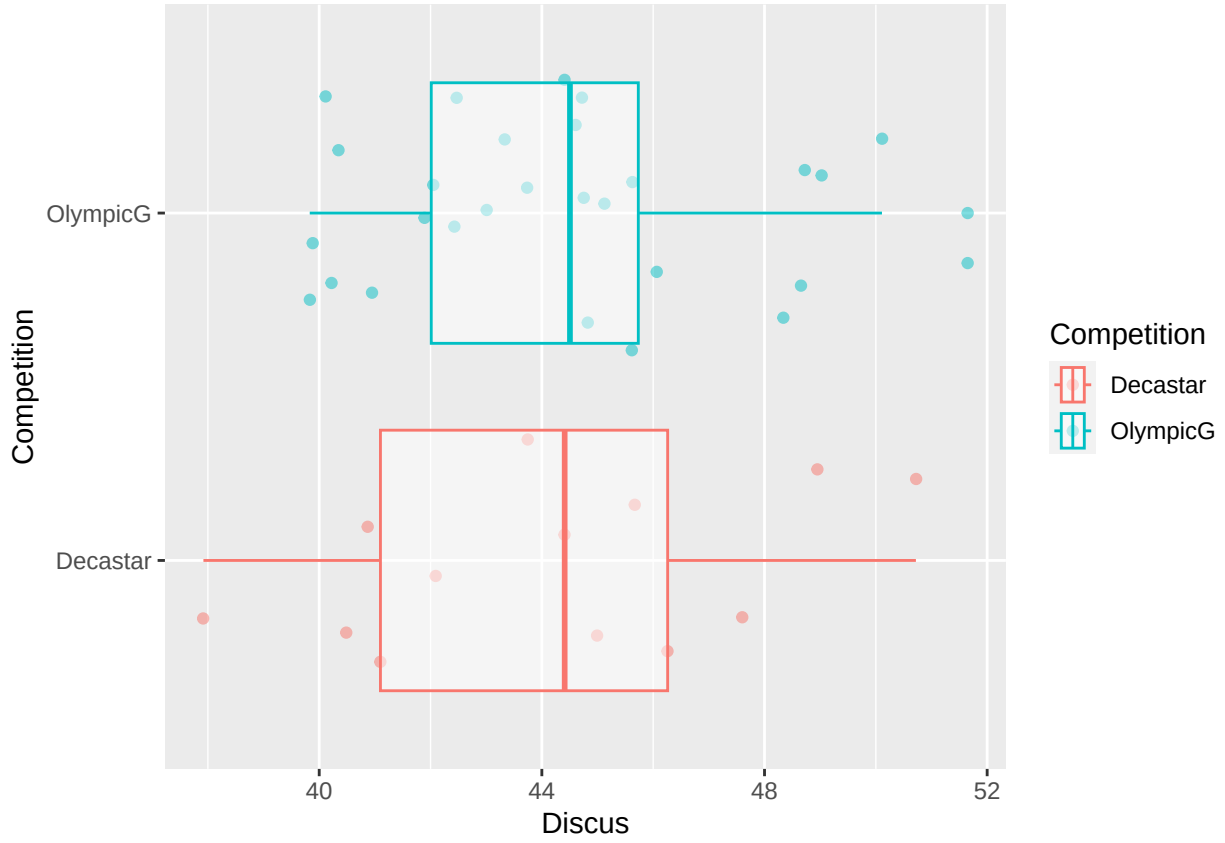

```
# 3
ggplot(decathlon) +
  geom_point(aes(x = Discus, y = Javeline, color = Competition)) +
  facet_wrap(~Competition)
```




```
# 4
decathlon.dt <- as.data.table(decathlon, keep.rownames = TRUE) %>%
  select(-Competition, -Rank, -Points)
table.stats <- decathlon.dt %>%
  melt(id.vars = c("rn"), variable.name = "Event", value.name = "Time") %>%
  group_by(Event) %>%
  summarise(Mean = mean(Time), SD = sd(Time))
table.stats
```

```
## # A tibble: 10 x 3
##    Event          Mean      SD
##    <fct>         <dbl>   <dbl>
##  1 100m          11.0   0.263
##  2 Long.jump      7.26  0.316
##  3 Shot.put      14.5   0.824
##  4 High.jump      1.98  0.0890
##  5 400m          49.6   1.15
##  6 110m.hurdle   14.6   0.472
##  7 Discus        44.3   3.38
##  8 Pole.vault     4.76  0.278
##  9 Javeline      58.3   4.83
## 10 1500m        279.    11.7
```

```
ggplot(decathlon) +
  geom_jitter(aes(x = Discus, y = Competition, color = Competition), alpha = 0.5) +
  geom_boxplot(aes(x = Discus, y = Competition, color = Competition), alpha = 0.5)
```




```
# 5
Corr <- decathlon %>%
  group_by(Competition) %>%
  summarise(Correlation = cor(Long.jump, High.jump))
Corr
```

```
## # A tibble: 2 x 2
##   Competition Correlation
##   <fct>             <dbl>
## 1 Decastar          0.157
## 2 OlympicG          0.346
```