21500 Lab 6 HW: Correlations and Scatterplots

Saifeldin Fathelbab

## HOMEWORK

We will use the Palmer Penguins dataset for the following items:

In a new R Markdown file:

1. Load your necessary libraries (tidyverse and corrplot)

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.7 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(corrplot)

## corrplot 0.92 loaded

1. Load the penguin dataset by running this code chunk:

penguins <- read.csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-07-28/penguins.csv")  
head(penguins)

## species island bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g  
## 1 Adelie Torgersen 39.1 18.7 181 3750  
## 2 Adelie Torgersen 39.5 17.4 186 3800  
## 3 Adelie Torgersen 40.3 18.0 195 3250  
## 4 Adelie Torgersen NA NA NA NA  
## 5 Adelie Torgersen 36.7 19.3 193 3450  
## 6 Adelie Torgersen 39.3 20.6 190 3650  
## sex year  
## 1 male 2007  
## 2 female 2007  
## 3 female 2007  
## 4 <NA> 2007  
## 5 female 2007  
## 6 male 2007

1. Run the following code to remove rows with missing data:

penguins <-  
 penguins %>%   
 drop\_na()  
  
head(penguins)

## species island bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g  
## 1 Adelie Torgersen 39.1 18.7 181 3750  
## 2 Adelie Torgersen 39.5 17.4 186 3800  
## 3 Adelie Torgersen 40.3 18.0 195 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  
## sex year  
## 1 male 2007  
## 2 female 2007  
## 3 female 2007  
## 4 female 2007  
## 5 male 2007  
## 6 female 2007

1. Make a scatterplot of bill length and flipper length? Describe the relationship between bill length and flipper length in our penguin data.

plot(penguins$bill\_length\_mm,penguins$flipper\_length\_mm,data=penguins,xlab="Bill Length (mm)",ylab="Flipper Length (mm)")

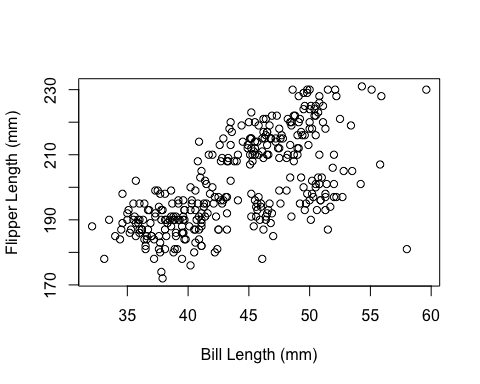
## Warning in plot.window(...): "data" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a  
## graphical parameter  
  
## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a  
## graphical parameter

## Warning in box(...): "data" is not a graphical parameter

## Warning in title(...): "data" is not a graphical parameter



## Observation

* With the increase in bill length, flipper length is also increasing.
* In the given dataset, there is very small percentage of penguins who have bill length greater than 55 mm and less than 35 mm.
* Overall for all species, the dataset has wide range for both bill and flipper lengths and data is not concentrated at one point.

1. What is the correlation between bill length and flipper length?

cor(penguins$bill\_length\_mm,penguins$flipper\_length\_mm)

## [1] 0.6530956

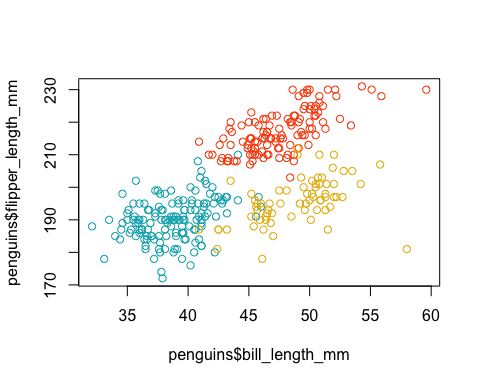
1. What is the PERCENT of variance accounted between ( as a percent) for bill length and flipper length ? Round your final answer to 2 decimal places.

r2 <- cor(penguins$bill\_length\_mm,penguins$flipper\_length\_mm)^2  
  
r2 <- r2\*100  
  
round(r2,2)

## [1] 42.65

1. Run the following code? Describe the resulting plot. Why are there three color codes given in the code? What does this graph tell you about the relationship between the length of the bills and the length of the flippers in these 3 species of penguins?

penguins$species <- as.factor(penguins$species)  
  
my\_cols <- c("#00AFBB", "#E7B800", "#FC4E07")   
  
plot(penguins$bill\_length\_mm, penguins$flipper\_length\_mm,   
 col = my\_cols[penguins$species])



## Observation and explanation of relationship:

* There is very small number of species which are intermixed with each other in terms of bill length and flipper.
* There is higher chance that penguins from same species have similar bill and flipper lengths.
* One of the species have higher flipper length as compared to others.
* Only 1 penguin has bill length of approx 60 mm and flipper length of 230mm which is highest among all species.

Q: Why three colors codes?

Answer: It is because we have 3 species in the dataset so we want to give the col option in the plot function a three elements vector in order to show three colors. It is required that the number of elements of the vector my\_cols are equal to the number of categories in the col command otherwise we get an error.

1. Refer to lab 4. Make a filter for each species of penguin (you did this for lab 4 HW so you can copy and paste your code). Then, using each of the created filters, plot the relationship between bill length and flipper length.

df <- subset(penguins, penguins$species == "Chinstrap")  
  
plot(penguins$bill\_length\_mm,penguins$flipper\_length\_mm,data=df, xlab="Bill Length (mm)",ylab="Flipper Length (mm)")

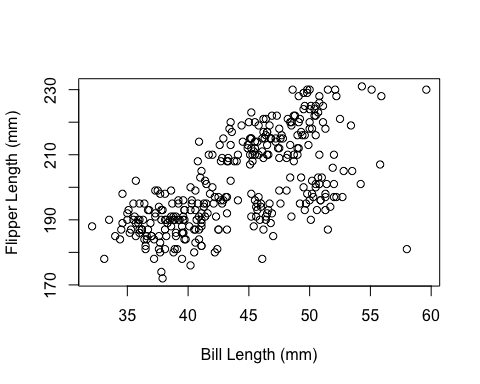
## Warning in plot.window(...): "data" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a  
## graphical parameter  
  
## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a  
## graphical parameter

## Warning in box(...): "data" is not a graphical parameter

## Warning in title(...): "data" is not a graphical parameter



df <- subset(penguins, penguins$species == "Gentoo")  
  
plot(penguins$bill\_length\_mm,penguins$flipper\_length\_mm,data=df, xlab="Bill Length (mm)",ylab="Flipper Length (mm)")

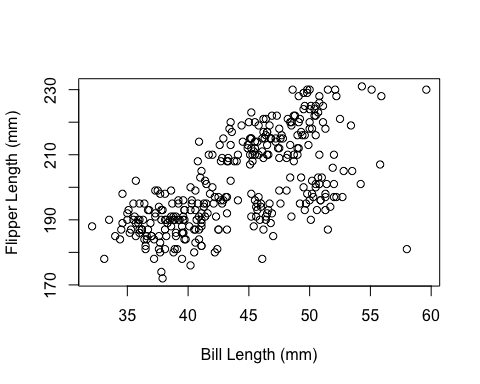
## Warning in plot.window(...): "data" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a  
## graphical parameter  
  
## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a  
## graphical parameter

## Warning in box(...): "data" is not a graphical parameter

## Warning in title(...): "data" is not a graphical parameter



df <- subset(penguins, penguins$species == "Adelie")  
plot(penguins$bill\_length\_mm,penguins$flipper\_length\_mm,data=df, xlab="Bill Length (mm)",ylab="Flipper Length (mm)")

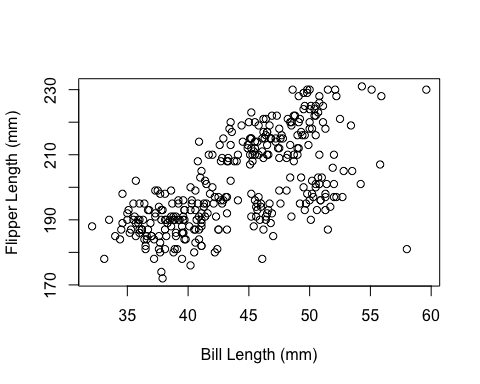
## Warning in plot.window(...): "data" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a  
## graphical parameter  
  
## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not a  
## graphical parameter

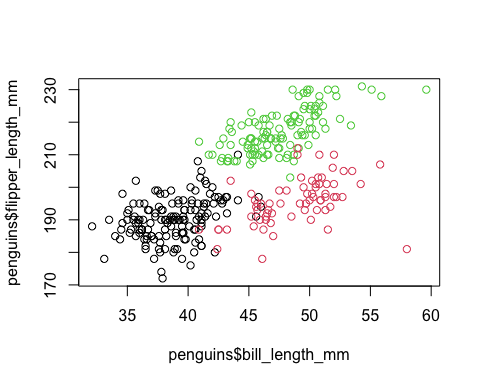
## Warning in box(...): "data" is not a graphical parameter

## Warning in title(...): "data" is not a graphical parameter



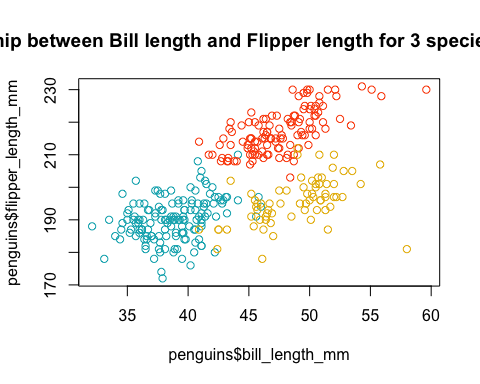
1. Use your answers in #7 to match the colors in the graph to the species of penguin.

plot(penguins$bill\_length\_mm, penguins$flipper\_length\_mm,   
 col = penguins$species)



1. Modify the code in #7 to add the appropriate title to the graph.

my\_cols <- c("#00AFBB", "#E7B800", "#FC4E07")   
plot(penguins$bill\_length\_mm, penguins$flipper\_length\_mm,   
 col = my\_cols[penguins$species],main="Relationship between Bill length and Flipper length for 3 species of Penguins")



1. Run a correlation matrix on the penguin\_matrix data. Round the values to 3 decimal places. What is the strongest relationship you see in this correlation matrix? Does it match with your prediction from the scatterplot matrix?

numeric\_columns <- penguins[,c(3,4,5,6,8)]   
  
cor\_matrix <- cor(numeric\_columns)  
  
round(cor\_matrix,3)

## bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g  
## bill\_length\_mm 1.000 -0.229 0.653 0.589  
## bill\_depth\_mm -0.229 1.000 -0.578 -0.472  
## flipper\_length\_mm 0.653 -0.578 1.000 0.873  
## body\_mass\_g 0.589 -0.472 0.873 1.000  
## year 0.033 -0.048 0.151 0.022  
## year  
## bill\_length\_mm 0.033  
## bill\_depth\_mm -0.048  
## flipper\_length\_mm 0.151  
## body\_mass\_g 0.022  
## year 1.000

## Observation:

* Strongest relationship exists between body\_mass\_g and flipper\_length\_mm for all the penguins because the corelation coefficient value is 0.873 as compared to all other values in the corelation matrix given above.
* It does not match with our scatterplot matrix since we have used bill\_length\_mm and flipper\_length\_mm before. The value we got for these two variables is lower than body\_mass\_g and flipper\_length\_mm.

1. Is there a significant relationship between bill length and flipper length? Run a hypothesis test on this correlation and give an APA summary of your findings.

t.test(penguins$bill\_length\_mm,penguins$flipper\_length\_mm)

##   
## Welch Two Sample t-test  
##   
## data: penguins$bill\_length\_mm and penguins$flipper\_length\_mm  
## t = -190.4, df = 430.8, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -158.5946 -155.3537  
## sample estimates:  
## mean of x mean of y   
## 43.99279 200.96697

## Observation:

In order to test the significant relationship between bill\_length\_mm and flipper\_length\_mm we conduct t-test. It is requirement for t-test that we define 2 hypothesis which are given as

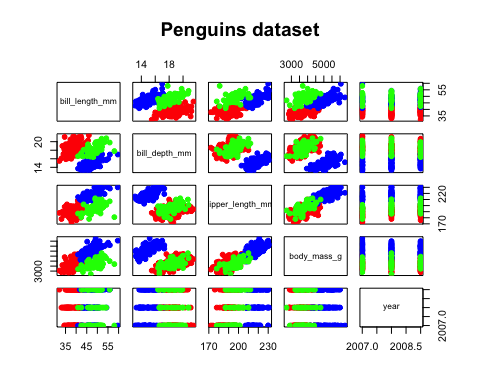
*Null hypothesis :* There is no significant difference between mean value of bill length and flipper length for penguins. *Alternate Hypothesis :* There is significant difference between mean value of bill length and flipper length for penguins.

The acceptance of null or alternate hypothesis depends upon the p-value which is based upon the significance level . So at 5% significance level, our p-value from the t-test is <0.05 which means that we reject our null hypothesis and accept the alternate hypothesis. If it was other way around i.e p-value >0.05, the null hypothesis would have been accepted.

In other words we can say that there is significant difference between the means of bill length and flipper length. Remember that our observation is based upon 95% Confidence interval.

1. Visualize the penguin\_matrix data using the corrgram of your choice. Do you find the corrgram to be an effective way to visualize the data? Explain.

pairs(numeric\_columns,   
 labels = colnames(numeric\_columns),   
 pch = 21,   
 bg = rainbow(3)[penguins$species],   
 col = rainbow(3)[penguins$species],   
 main = "Penguins dataset",  
 row1attop = TRUE,   
 gap = 1,   
 cex.labels = NULL,   
 font.labels = 1)



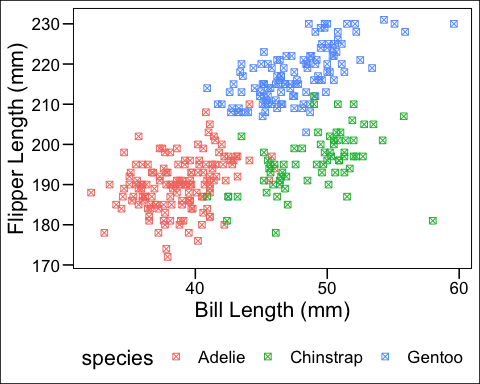
## Observation

Yes corrgram is an effective way of finding relationship between numerical data in a given dataset. It can help us visualize the corelation matrix from question 11 in graphical form. From the above graph we can deduce that there are several variables which have positive corelation with each other since there is an increasing trend of scatterplot points between variables. For example the variables which have positive corelation with each other are body\_mass\_g and flipper\_length\_mm for all species. Similarly the variables bill\_length\_mm and body\_mass\_g have almost positive corelation.. It can also help us to determine which numeric variables have negative or no corelation with each other. In the above graph we can see that year numeric variable is not linked to any of the other numeric varibles and has a flat line coressponding to y-axis.

### EXTRA CREDIT:

1. Use ggplot2 to create a scatterplot of the relationship between bill length and flipper length, colored by species

library(ggplot2)  
  
ggplot(penguins) +  
 aes(x = bill\_length\_mm, y = flipper\_length\_mm, colour = species) +  
 geom\_point(shape = "square cross",   
 size = 2L) +  
 scale\_color\_hue(direction = 1) +  
 labs(x = "Bill Length (mm)", y = "Flipper Length (mm)") +  
 ggthemes::theme\_base() +  
 theme(legend.position = "bottom")



1. Use ggplot2 to create a scatterplot of the relationship between bill length and flipper length, faceted by species

ggplot(penguins) +  
 aes(x = bill\_length\_mm, y = flipper\_length\_mm) +  
 geom\_point(shape = "square cross",   
 size = 1.5, colour = "#112446") +  
 labs(x = "Bill Length (mm)", y = "Flipper Length (mm)") +  
 ggthemes::theme\_base() +  
 facet\_wrap(vars(species))

