visual data in R

# Making plots with ggplot package

To do this part of the homework you can copy and past the code in the ‘code blocks’ into R and run it. Then when you figure out how to answer the questions just copy and paste your code from R into this document!

Rather than a normal handout/lab this one tells a story and has you learning ggplot while you do so. Be sure to check out the reading here: <https://r4ds.had.co.nz/data-visualisation.html> before doing this project. If you need more background on how ggplot works check out chpt 3 of <https://socviz.co/makeplot.html#makeplot> “Data Visualization”

YOu might not know how to do the things being asked for right away. that is fine and expected. What i want you to learn is how to *find the answers* rather than what the answer is. In other words, I am more concerned with the process than the result. But if you are getting frustrated reach out and I’m always happy to help

## Story

You are working on your Ph.D with a professor of anthropology at Fancypants University. She has just learned about a new species of monkey that she wants you to collect data on. So you are off to the island of Naboombu, where you will be given the task of watching these primates.



## Collecting your data, part 1

When you arrive on the western part of the island you notice that there are a whole lot of monkeys here and you need to find a way to get an idea of what they look like. So you and your team decide to start by measuring their tails (they aren’t apes!). You manage to measure 500 of these fellas (ok, this is a bit of a reach but lets say that you are *really* quick at measuring and the monkeys like having their tails measured….)

below is a very quick reminder of how R works

#first we make an object and assign something to it...  
  
the\_numbers <- c(4, 8, 15, 16, 23, 42)  
#once we have an object we can do things to that object.  
  
mean(the\_numbers)

## [1] 18

sum(the\_numbers)

## [1] 108

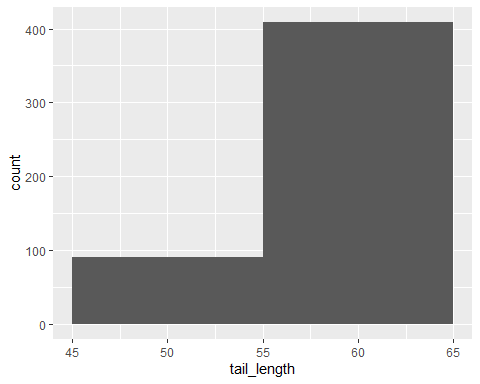
sd(the\_numbers)

## [1] 13.49074

We are going to be simulating the data you collected!

Run the following R code in a new session to get the first monkey data! **Be sure to run the first line of code that loads in the tidyverse library.** I made some comments so you can follow along with what the code is doing:

library(tidyverse)  
set.seed(2020) #this sets the randomizer in R  
tail\_length <-rnorm(500,57,2) # create a set of 500 random numbers with a mean of 57 and a standard deviation of 2   
sample\_number <-1:500 # create a vector of numbers from 1 to 500  
monkey\_pop <- rep(letters[1], 500) # this makes a vector of 500 'a's  
monkey\_data\_1 <- tibble(sample\_number, tail\_length, monkey\_pop) #create a data table (a tibble is a function that makes nice data tables with a modern flair...)  
  
monkey\_data\_1 %>% ggplot(aes(tail\_length)) + geom\_histogram(binwidth = 10) # this takes the monkey data and plots a histogram of the tail length



### Question 1: Changing the bin length:

You note that the plot you made with the code above doesn’t look very nice. After thinking for a bit you realize the bin length is too large and you can’t see enough of the data. What code would you use to change the binwidth to be 1. copy the code from above and change it so the figure looks better. Then copy that code into the box below:

### 

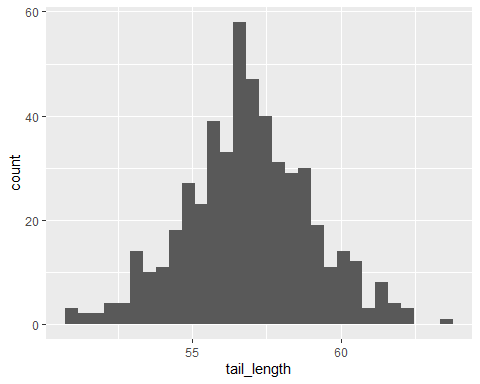
#copy your answer to question 1 here:

### Question 2: Ok, but you know your professor wants the histogram to be blue. How can you change the code below to make it blue?

#### (n.b. for help see section **3.5 “Mapping aesthetics vs setting them” of** [**https://socviz.co/makeplot.html#makeplot**](https://socviz.co/makeplot.html#makeplot)**)**

monkey\_data\_1 %>% ggplot(aes(tail\_length)) + geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



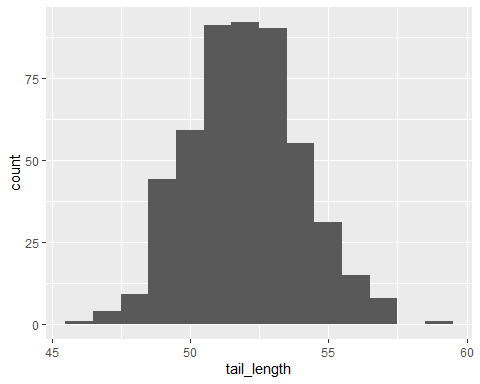
#copy your answer to question 2 here:

You send this lovely blue histogram off to your professor and go and take a nap (measuring 500 tails *plus* data entry is serious work!)

## Collecting your data, part 2

The next day you are talking to some local people and hear about another group of monkeys on a different part of the island. **FIrst run the code below to get that data and make a histogram!**

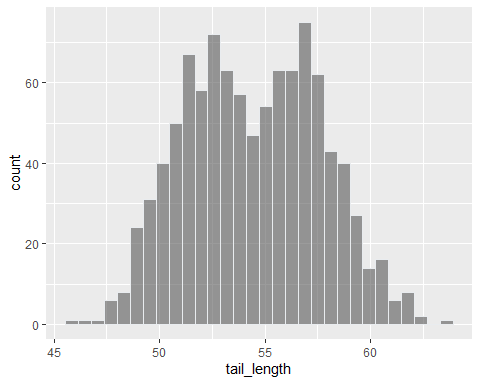
tail\_length <-rnorm(500,52,2)  
sample\_number <-501:1000  
monkey\_pop <- rep(letters[2], 500)  
monkey\_data\_2 <- tibble(sample\_number, tail\_length, monkey\_pop)  
  
monkey\_data\_2 %>% ggplot(aes(tail\_length)) + geom\_histogram(binwidth = 1)



Now you want to look at both of the populations at the same time! Thankfully you have taken this class and have some R code ready to do this. Run the code below in your R session to make a histogram of both datasets

all\_data <- bind\_rows(monkey\_data\_1, monkey\_data\_2) # here we are using the function bindrows to add the two datasets together!  
  
all\_data %>% ggplot(aes(tail\_length )) + geom\_histogram(color="#e9ecef", alpha=0.6, position = 'identity')

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



hmmm…that doesn’t look right since you can’t tell the difference groups. But you think back to your stats class in college and remember that we need to set an aes(...) argument that specifies that a variable will be expressed by one of the available visual elements, such as size and color. you sorta did this above when we made the histogram blue….

### Question 3: How can you change the code below to make the different monkey populations have different colors (note: you just need to add one argument to the aes() part. i put “\*\*\*\*\*\*\*\*\*\*\*\*\*” to show where to put this

#edit the below code to color by the variable monkey\_pop  
  
#all\_data %>% ggplot(aes(tail\_length, fill=monkey\_pop, \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* )) + geom\_histogram(color="#e9ecef", alpha=0.6, position = 'identity')

So you look at this and think to yourself “Hey, they overlap a bit but there are difference at the tails….😀. I am going to send this to my adviser and she will be soooooo happy!”

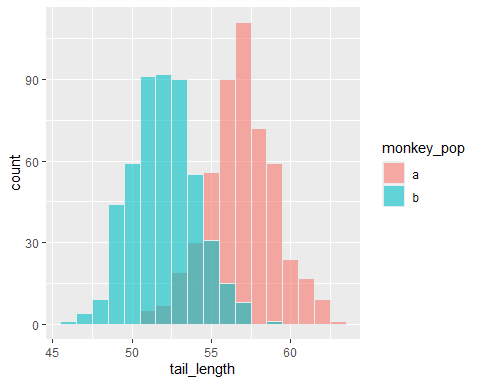
A few hours later you get an email with these requests. Your adviser wants you to do all of the following to the figure you have made

### Question 4

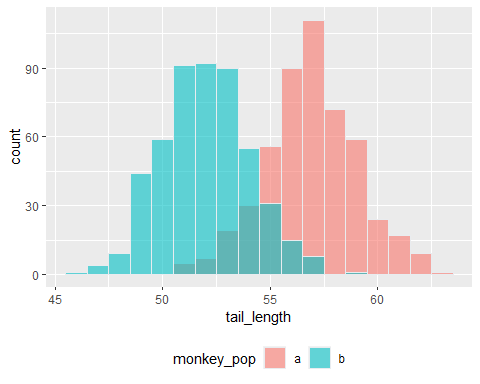
* Put the legend on the bottom rather than on the right
* Label the x axis “tail length (cm)”
* Label the y axis “number of observations”
* Give it a title that is meaningful
* Put a credit somewhere that says who made the figure
* Change the legend so that it says “Monkey groups”
* Change the labels from “a” and “b” to “Group one” and “Group 2”

Lets take this step by step. Copy the below R code and use each step to build the new graph. If you need help use the class readings or ask on Discord! Then copy the code you used below. For example, to get the first part of this we would update the code like this

#old code  
all\_data %>% ggplot(aes(tail\_length, fill=monkey\_pop )) + geom\_histogram(color="#e9ecef", alpha=0.6, position = 'identity', binwidth = 1)



# new code  
  
all\_data %>% ggplot(aes(tail\_length, fill=monkey\_pop )) + geom\_histogram(color="#e9ecef", alpha=0.6, position = 'identity', binwidth = 1) + theme(legend.position = "bottom")



To figure out how to do this we are going to need to check online. One good source is <https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf>. For example if you look at the right hand column of page 2 you will see a box that talks about how to change the labels…another good source is this site: <https://ggplot2.tidyverse.org/reference/> which i use quite a bit to rejigger figures.

#copy your answer to question 4 here:  
  
  
#plot a  
  
  
#plot b  
  
  
  
#plot c  
  
  
  
#plot d  
  
  
  
#plot e  
  
  
#plot f  
  
  
#plot g

## Part 3: New visualizations with new data

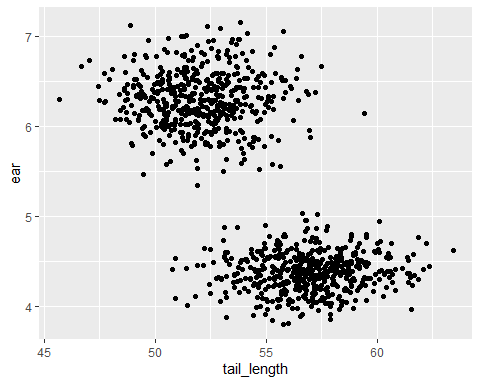
You and your professor submit your data paper to a journal. After a few months you get your peer review back and the reviewers say “wow, this is cool! But ya know what, we think you need more measurements before we can conclude these are 2 different populations. Can you go and measure ear length?”

copy and Run the following code to get the new\_data

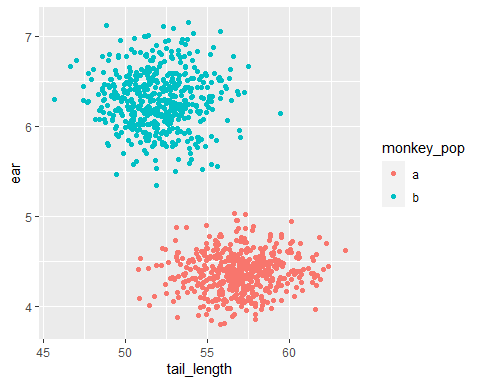
ear1 <- rnorm(500, 4.4, .22)  
ear2 <- rnorm(500, 6.3, .31)  
  
ear1 <-tibble(ear =rnorm(500, 4.4, .22), monkey\_pop = rep(letters[1], 500), sample\_number = 1:500)  
  
  
ear2 <-tibble(ear =rnorm(500, 6.3, .31), monkey\_pop = rep(letters[2], 500), sample\_number = 501:1000)  
  
ear\_all <- bind\_rows(ear1, ear2)  
  
new\_data <- all\_data %>% left\_join(ear\_all, by= c("monkey\_pop", "sample\_number"))

Since you know have two measurements on each monkey you need a new way to visualize these data. Remembering what you learned at AppState you go and make a scatterplot of these data. Copy the below code into R and see what happens!

new\_data %>% ggplot(aes(tail\_length, ear)) + geom\_point() # this makes the scatterplot geom



new\_data %>% ggplot(aes(tail\_length, ear, color = monkey\_pop)) + geom\_point()



You are thrilled and send it off to your adviser.

She comes back and asks for the following updates:

### Question 5: updating the scatteplot

* Get rid of the legend
* Add a regression line for each group
* Add a transparency or a jittter to deal with over plotting
* Make the 2 groups have a different shape
* Add an ellipse to the groups ( *see “stat\_ellipse()” for a hint*)

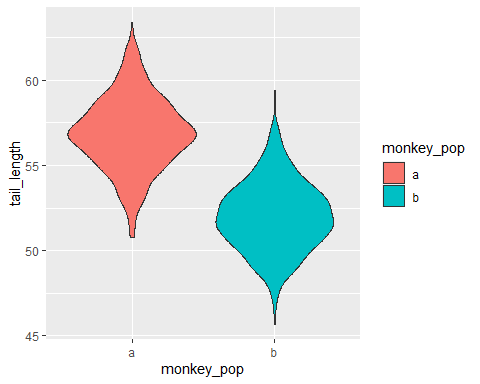
#write you answers to question 4 here  
  
  
#plot a  
  
  
#plot b  
  
  
  
#plot c  
  
  
# plot d

### Question 6: taking it to 11…

Your adviser now asks if you can make the plots with the new data the following different geoms.

* make a violin plot
* make a boxplot
* make a ridgeline plot (n.b. to do this one we need to install a new package called “ggridges” (<https://www.r-graph-gallery.com/294-basic-ridgeline-plot.html>)

# write you answers to question 6 here. i put the first one here to help you out  
  
#plot a  
new\_data %>% ggplot(aes( monkey\_pop, tail\_length, fill = monkey\_pop)) + geom\_violin()



# plot b  
  
  
  
#plot c

## export the data

she wants the whole dataset sent to her so she can take a look at it herself. you know that you can use write\_csv to do a basic export.

### Question 7

* change the name of the csv file so it is called “new\_primate\_data” rather than “my\_monkeydata.csv”
* change the format of the exported file to an Excel sheet (this is hard)

#write you answers to question 7 here. as a tip look at this code: write\_csv(new\_data, "my\_monkeydata.csv")  
  
  
#part a  
  
  
  
  
  
#part b

# What else can you do with R visualizations

Now that you have learned a lot about how to work with visualization data i want you to take a moment and think about how to approach another dataset

The palmerpenguins package contains two datasets. to install it copy the following code into R

install.library("palmerpenguins")

One is called penguins, and is a simplified version of the raw data; see [?penguins](https://allisonhorst.github.io/palmerpenguins/reference/penguins.html) for more info

The second dataset is penguins\_raw, and contains all the variables and original names as downloaded; see [?penguins\_raw](https://allisonhorst.github.io/palmerpenguins/reference/penguins_raw.html) for more info.

Both datasets contain data for 344 penguins. There are 3 different species of penguins in this dataset, collected from 3 islands in the Palmer Archipelago, Antarctica. The culmen is the upper ridge of a bird’s bill. In the simplified penguins data, culmen length and depth are renamed as variables bill\_length\_mm and bill\_depth\_mm to be more intuitive.

think about what the below code is doing. Once again we are using the pipe (%>%) to take the dataset and apply a function to it (in this case, the count function)

library(palmerpenguins)

## Warning: package 'palmerpenguins' was built under R version 4.0.3

glimpse(penguins)

## Rows: 344  
## Columns: 8  
## $ species <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, A...  
## $ island <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torge...  
## $ bill\_length\_mm <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34....  
## $ bill\_depth\_mm <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18....  
## $ flipper\_length\_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, ...  
## $ body\_mass\_g <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 347...  
## $ sex <fct> male, female, female, NA, female, male, female, m...  
## $ year <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2...

penguins %>%  
 count(species)

## # A tibble: 3 x 2  
## species n  
## <fct> <int>  
## 1 Adelie 152  
## 2 Chinstrap 68  
## 3 Gentoo 124

penguins %>%  
 group\_by(species) %>%  
 summarize(across(where(is.numeric), mean, na.rm = TRUE))

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 3 x 6  
## species bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g year  
## <fct> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Adelie 38.8 18.3 190. 3701. 2008.  
## 2 Chinstrap 48.8 18.4 196. 3733. 2008.  
## 3 Gentoo 47.5 15.0 217. 5076. 2008.

Take a look at these data. what sort of figures might you want to make? Think about how you could tell a story about these penguins via an image. Then write some code that will let you do this. When you are done, take a moment to explain in prose why you made this figure and how you used R to do it