Labs 01 - Week1 Monday

Yunting Chiu

2021-05-25

Learning goals

- Get acquainted with R and RStudio, which we will be using throughout the course to analyze data as well as to learn the statistical concepts discussed in the course.
- Appreciate the value of visualization in exploring the relationship between variables.
- Start using R for building plots and calculating summary statistics.

Terminology

We've already thrown around a few new terms, so let's define them before we proceed.

- **R**: Name of the programming language we will be using throughout the course.
- **RStudio**: An integrated development environment for R. In other words, a convenient interface for writing and running R code.

I like to think of R as the engine of the car, and RStudio is the dashboard.

Starting slow

As the labs progress, you are encouraged to explore beyond what the labs dictate; a willingness to experiment will make you a much better programmer. Before we get to that stage, however, you need to build some basic fluency in R. Today we begin with the fundamental building blocks of R and RStudio: the interface, reading in data, and basic commands.

And to make versioning simpler, this is a solo lab. Additionally, we want to make sure everyone gets a significant amount of time at the steering wheel.

Getting started

Download R

If you don't have R installed

Go to the CRAN and download R, make sure you get the version that matches your operating system.

If you have R installed

If you have R installed run the following code - We can see my R Version is 4.0.2.

```
R.version
```

```
##                _
## platform       x86_64-apple-darwin17.0
```

```
## arch           x86_64
## os             darwin17.0
## system         x86_64, darwin17.0
## status
## major          4
## minor          0.2
## year           2020
## month          06
## day            22
## svn rev        78730
## language       R
## version.string R version 4.0.2 (2020-06-22)
## nickname       Taking Off Again
```

This should tell you what version of R you are currently using. If your R version is lower then 3.6.0 I would strongly recommend updating. In general it is a good idea to update your R version, unless you have a project right now that depend on a specific version of R.

Download RStudio

We recommend using RStudio as your IDE if you don't already have it installed. You can go to the RStudio website to download and install the software.

Launch RStudio

You can also open the RStudio application first and then create a project by going
`file -> new project...`

Create a new Rmarkdown file

`file -> new file -> R markdown...`

Hello RStudio!

RStudio is comprised of four panes.

- On the bottom left is the Console, this is where you can write code that will be evaluated. Try typing 2 + 2 here and hit enter, what do you get?
  Ans: I will get 4.

```
test <- 2 + 2
test
```

```
## [1] 4
```

- On the bottom right is the Files pane, as well as other panes that will come handy as we start our analysis.
- If you click on a file, it will open in the editor, on the top left pane.
- Finally, the top right pane shows your Environment. If you define a variable it would show up there. Try typing x <- 2 in the Console and hit enter, what do you get in the Environment pane? Ans: I will see the integer **2** in the global environment.

```
x <- 2
```

Packages

R is an open-source language, and developers contribute functionality to R via packages. In this lab we will work with three packages: `palmerpenguins` which contains the dataset, and `tidyverse` which is a collection of packages for doing data analysis in a "tidy" way.

Load these packages by running the following in the Console.
- Reference: https://github.com/tidymodels/parsnip

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 0.1.1 --
```

```
## v broom     0.7.6      v recipes   0.1.14
## v dials     0.0.9      v rsample   0.0.9
## v infer     0.5.3      v tune      0.1.1
## v modeldata 0.1.0      v workflows 0.2.1
## v parsnip   0.1.5      v yardstick 0.0.7
```

```
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
```

```
library(palmerpenguins)
# install.packages("devtools")
# devtools::install_github("tidymodels/parsnip")
# install.packages("parsnip")
```

If you haven't installed these packages yet and R complains, then you can install these packages by running the following command. (Note that R package names are case-sensitive)

```
# install.packages(c("tidyverse", "palmerpenguins"))
```

Note that the packages are also loaded with the same commands in your R Markdown document.

Warm up

Before we introduce the data, let's warm up with some simple exercises.

The top portion of your R Markdown file (between the three dashed lines) is called YAML. It stands for "YAML Ain't Markup Language". It is a human friendly data serialization standard for all programming

languages. All you need to know is that this area is called the YAML (we will refer to it as such) and that it contains meta information about your document.

YAML

Open the R Markdown (Rmd) file in your project, change the author name to your name, and knit the document.

Data

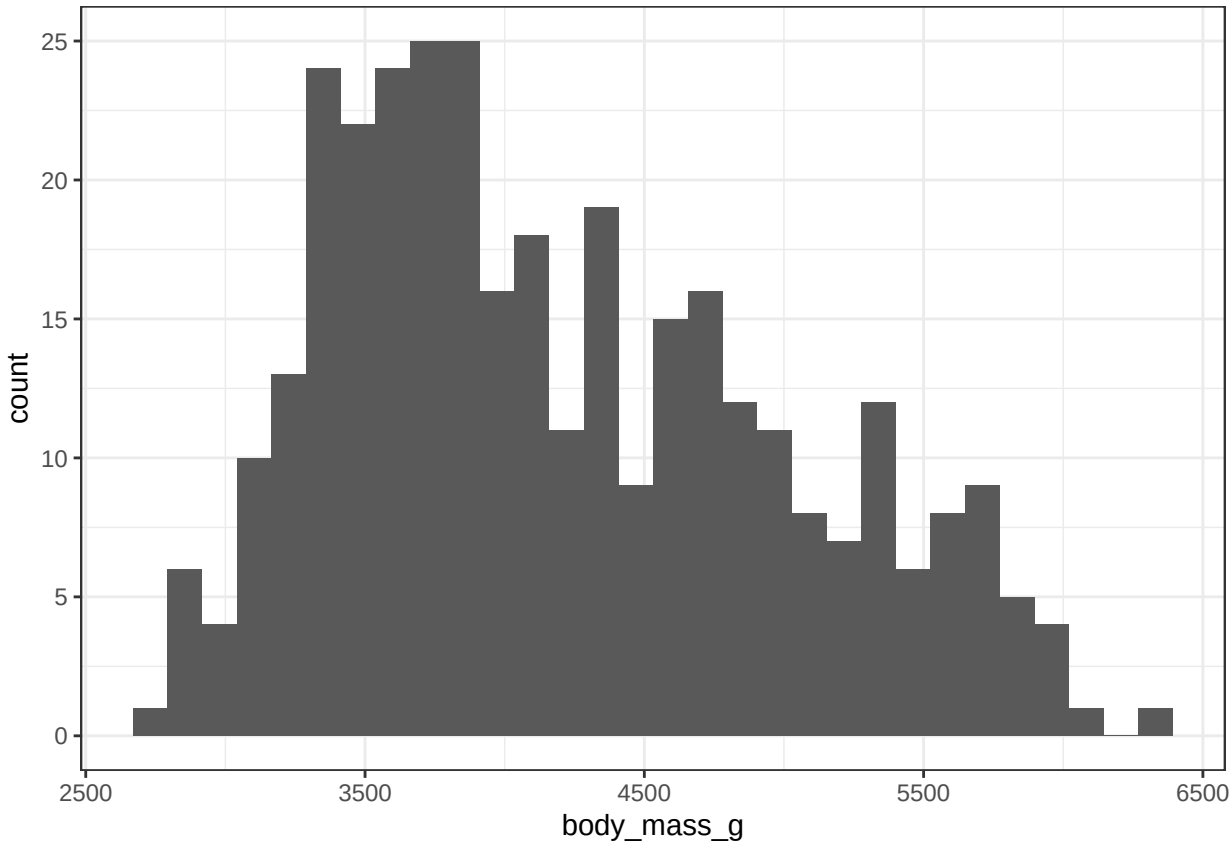
The data frame we will be working with today is called penguins and it's in the palmerpenguins package.
- count the number of `species` and `islands` with `dplyr::count()`

```
penguins %>%
  count(species, island) %>%
  rename(count = n)
```

```
## # A tibble: 5 x 3
##   species   island    count
##   <fct>     <fct>     <int>
## 1 Adelie    Biscoe       44
## 2 Adelie    Dream        56
## 3 Adelie    Torgersen    52
## 4 Chinstrap Dream        68
## 5 Gentoo    Biscoe      124
```

Visualize the distribution of `body_mass_g` with ggplot

```
penguins %>%
  ggplot(aes(body_mass_g)) +
  geom_histogram(bins = 30) +
  theme_bw()
```

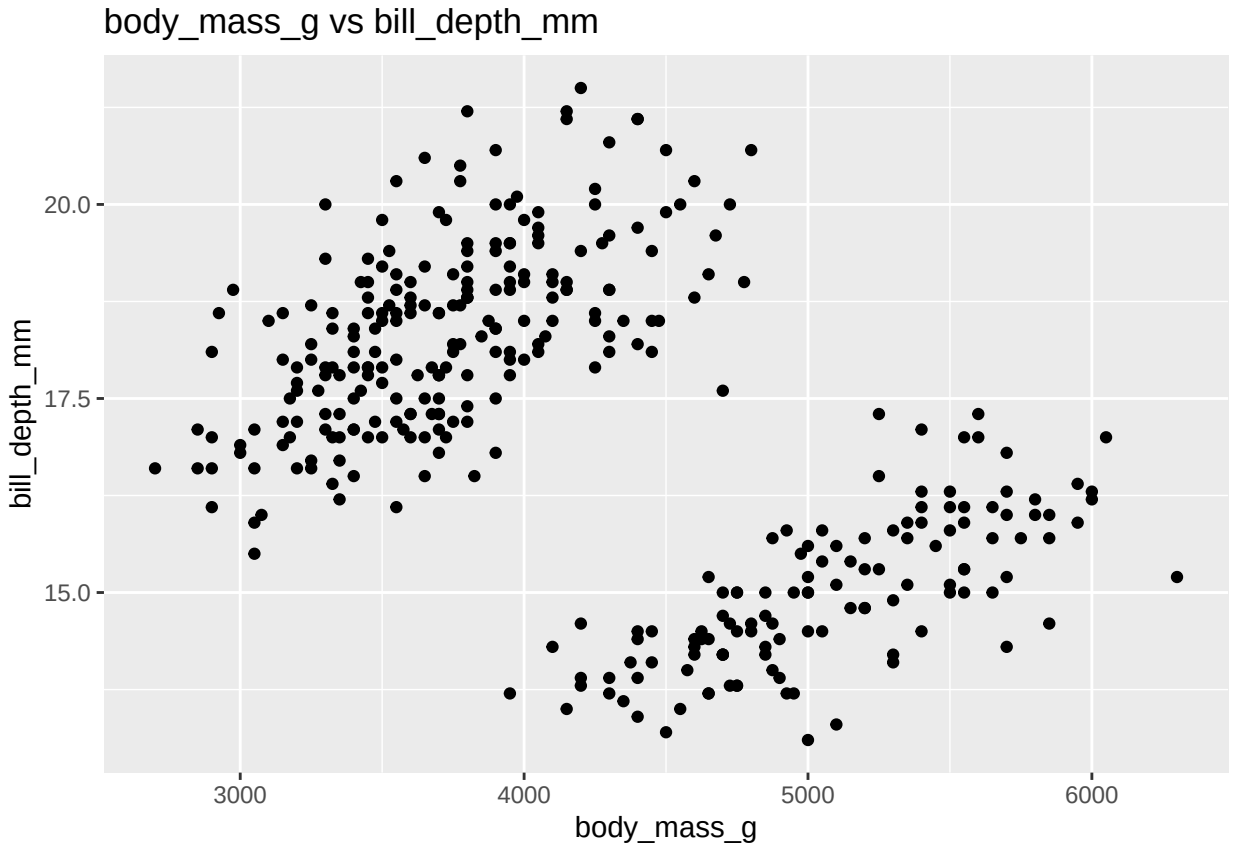


Look at the correlation between `body_mass_g` and some of the other variables

```
ggplot(penguins, aes(body_mass_g, flipper_length_mm)) +
  geom_point() +
  ggtitle("body_mass_g vs flipper_length_mm")
```


```
ggplot(penguins, aes(body_mass_g, bill_depth_mm)) +
  geom_point() +
  ggtitle("body_mass_g vs bill_depth_mm")
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



Modeling

Fit a linear model using `parsnip` to model `body_mass_g`

```
lm_spec <- linear_reg() %>%
  set_engine("lm")

lm_fit <- lm_spec %>%
  fit(body_mass_g ~ species + island + bill_length_mm + bill_depth_mm + flipper_length_mm,
      data = penguins)
lm_fit
```

```
## parsnip model object
##
## Fit time:  6ms
##
## Call:
## stats::lm(formula = body_mass_g ~ species + island + bill_length_mm +
##     bill_depth_mm + flipper_length_mm, data = data)
##
## Coefficients:
##       (Intercept)   speciesChinstrap      speciesGentoo        islandDream
##         -4353.133           -530.857            906.293             -4.306
##   islandTorgersen     bill_length_mm      bill_depth_mm  flipper_length_mm
##           -59.852             41.310            140.252             20.531
```

Get parameter estimates:

```r
tidy(lm_fit)
```

```
## # A tibble: 8 x 5
##   term              estimate std.error statistic  p.value
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       -4353.      497.     -8.77   9.45e-17
## 2 speciesChinstrap   -531.       89.7    -5.92   8.19e- 9
## 3 speciesGentoo       906.      149.      6.09   3.06e- 9
## 4 islandDream          -4.31     64.0    -0.0673 9.46e- 1
## 5 islandTorgersen     -59.9      65.5    -0.914  3.61e- 1
## 6 bill_length_mm       41.3       7.18    5.75   2.01e- 8
## 7 bill_depth_mm       140.       19.0     7.38   1.28e-12
## 8 flipper_length_mm    20.5       3.13    6.56   2.03e-10
```