previous:

The Cornell Lab of Ornithology

current:

Stanford

`dplyr`: a useful toolbox for manipulating data

Stepfanie M. Aguillon

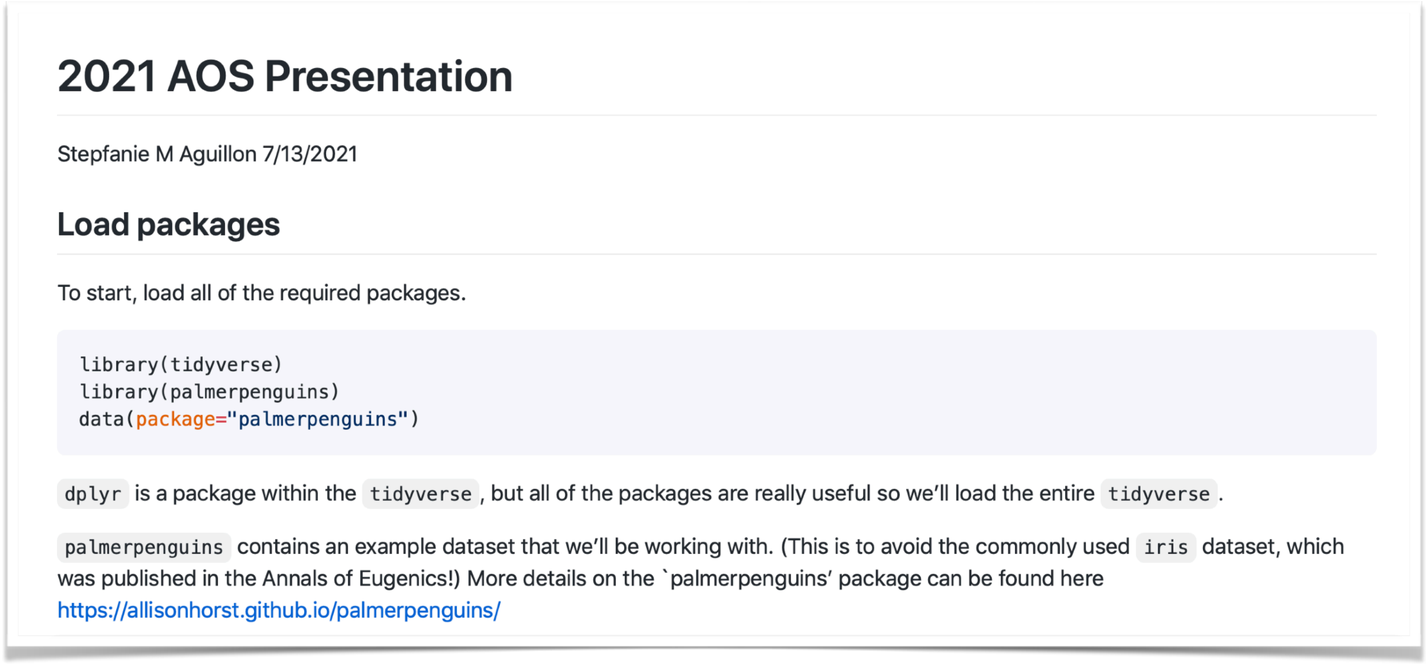
stepfanie.aguillon@gmail.com

@s_m_aguillon

Get your ornithological data into `dplyr` for a more reproducible workflow!

All code is available on my GitHub page!

- Find code for this presentation at: https://github.com/stepfanie-aguillon/AOS2021-dplyr



2021 AOS Presentation

Stepfanie M Aguillon 7/13/2021

Load packages

To start, load all of the required packages.

```
library(tidyverse)
library(palmerpenguins)
data(package="palmerpenguins")
```

`dplyr` is a package within the `tidyverse`, but all of the packages are really useful so we'll load the entire `tidyverse`.

`palmerpenguins` contains an example dataset that we'll be working with. (This is to avoid the commonly used `iris` dataset, which was published in the Annals of Eugenics!) More details on the `palmerpenguins' package can be found here https://allisonhorst.github.io/palmerpenguins/

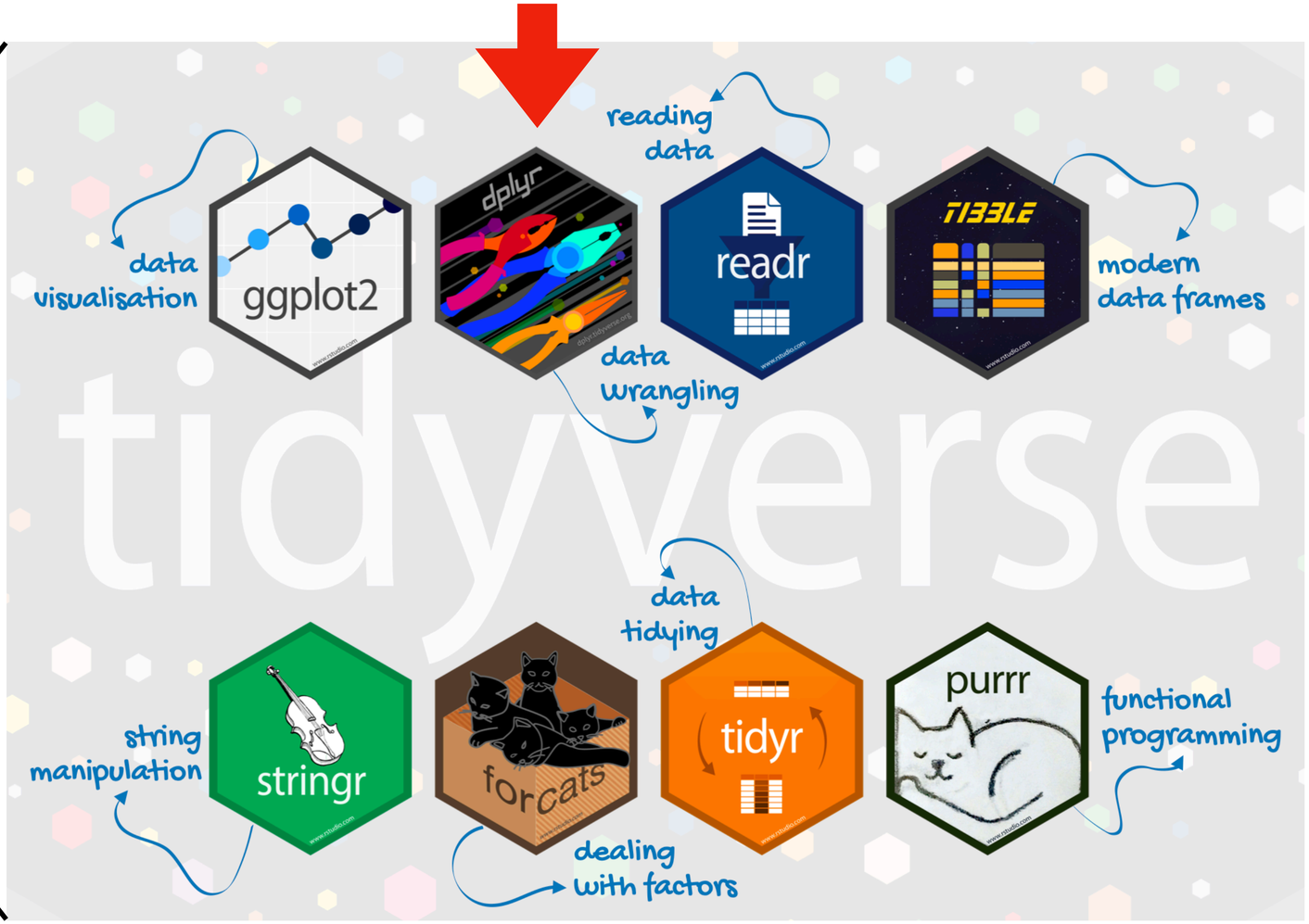
All code is available on my GitHub page!

- Find code for this presentation at: https://github.com/stepfanie-aguillon/AOS2021-dplyr

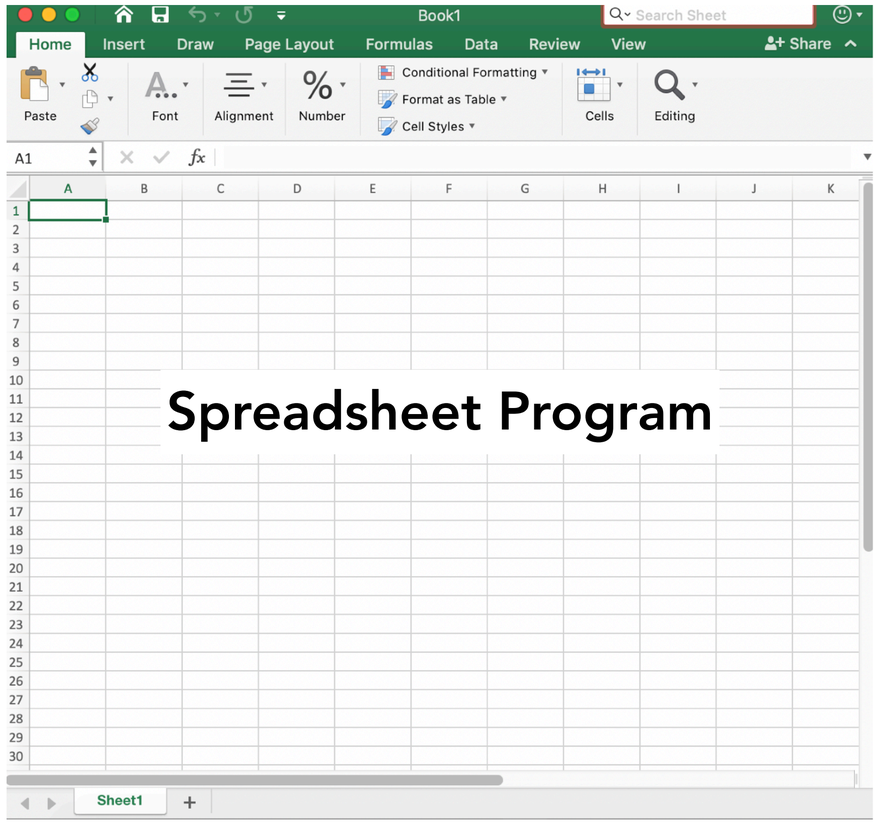
- Examples throughout use data from the `palmerpenguins` R package

Horst, AM, Hill AP, Gorman KB. 2020. `palmerpenguins`: Palmer Archipelago (Antarctica) penguin data. https://allisonhorst.github.io/palmerpenguins/

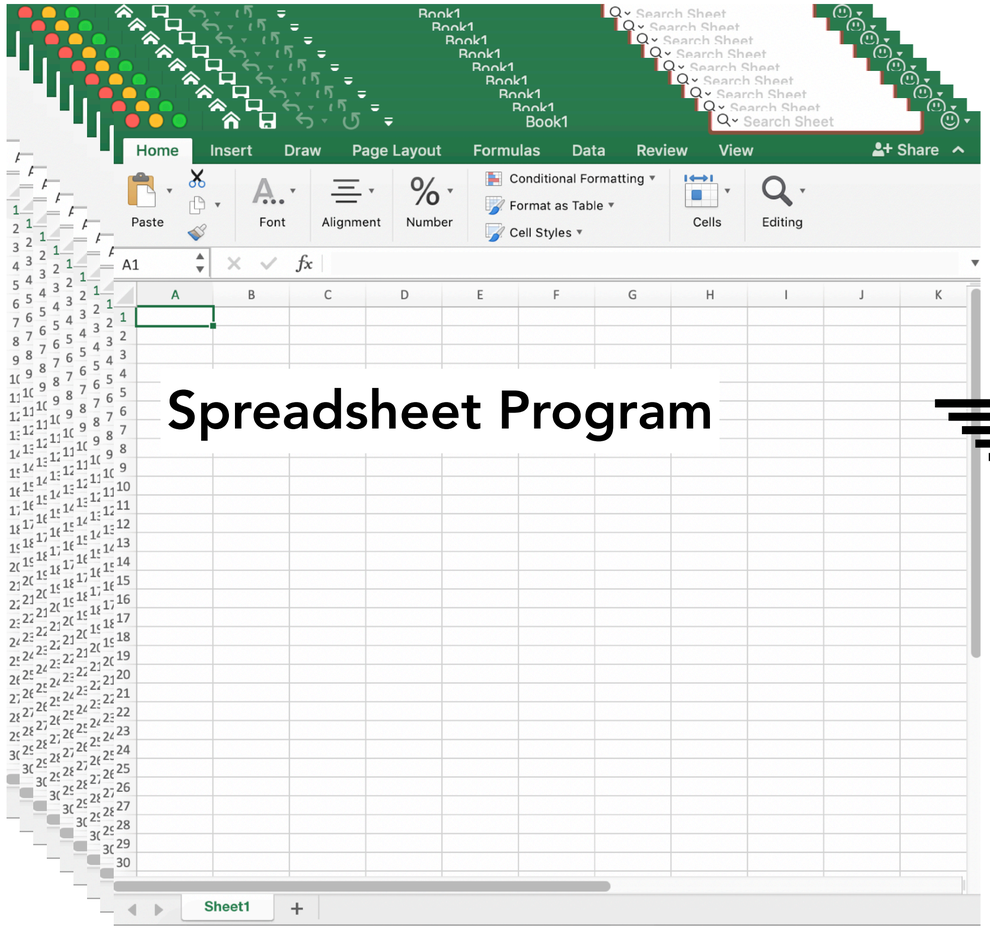
If you are new to R, this presentation is for you!



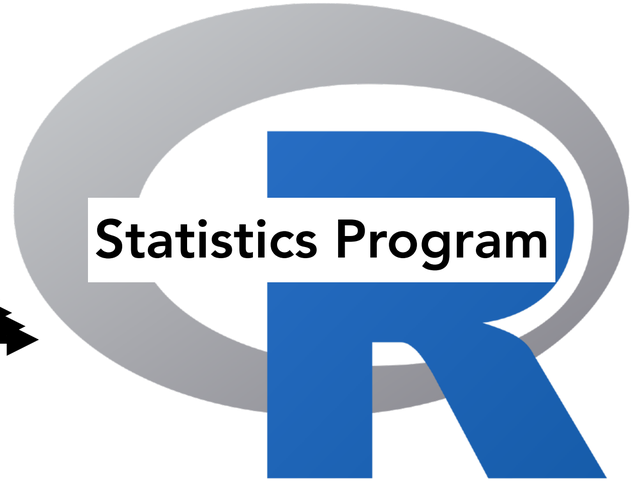
Making a reproducible workflow for your data processing



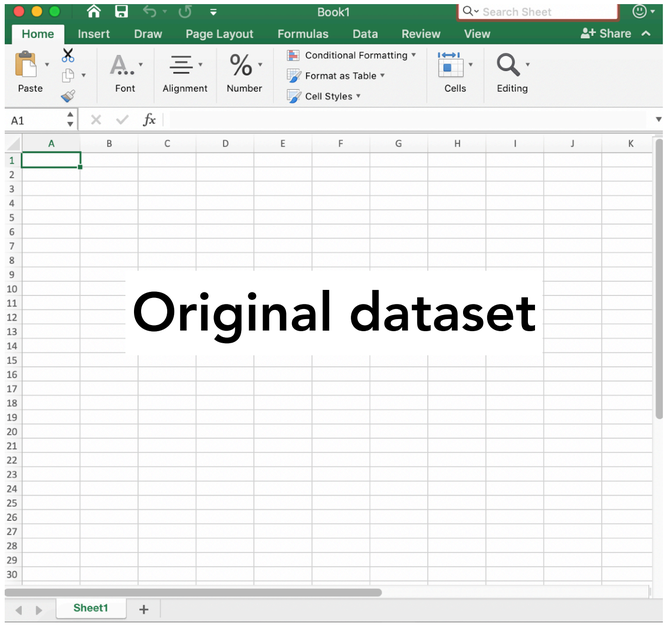
Making a reproducible workflow for your data processing

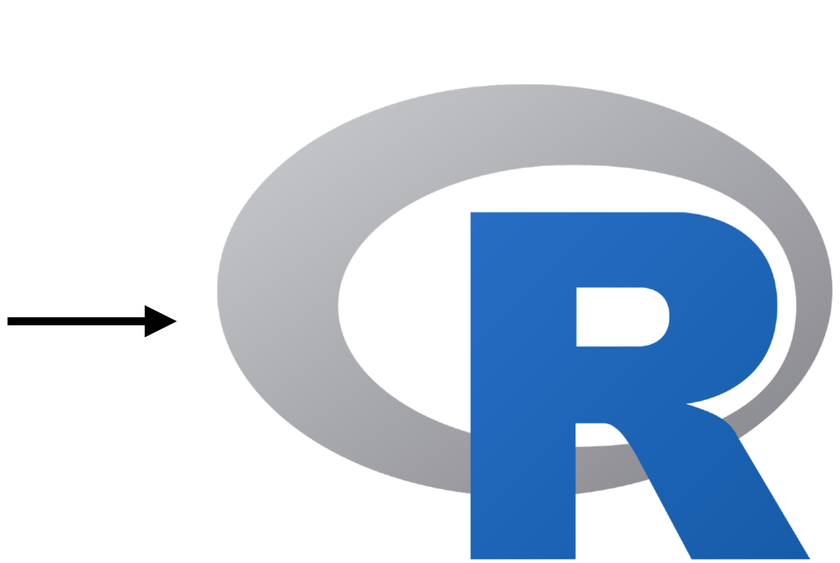


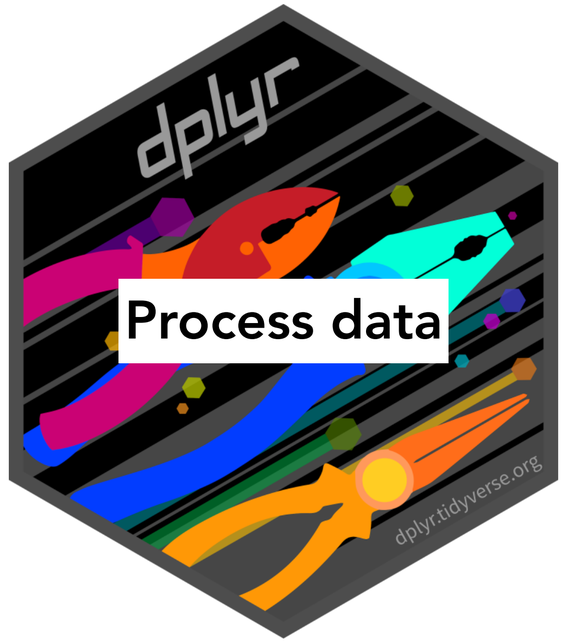
???



Making a reproducible workflow for your data processing







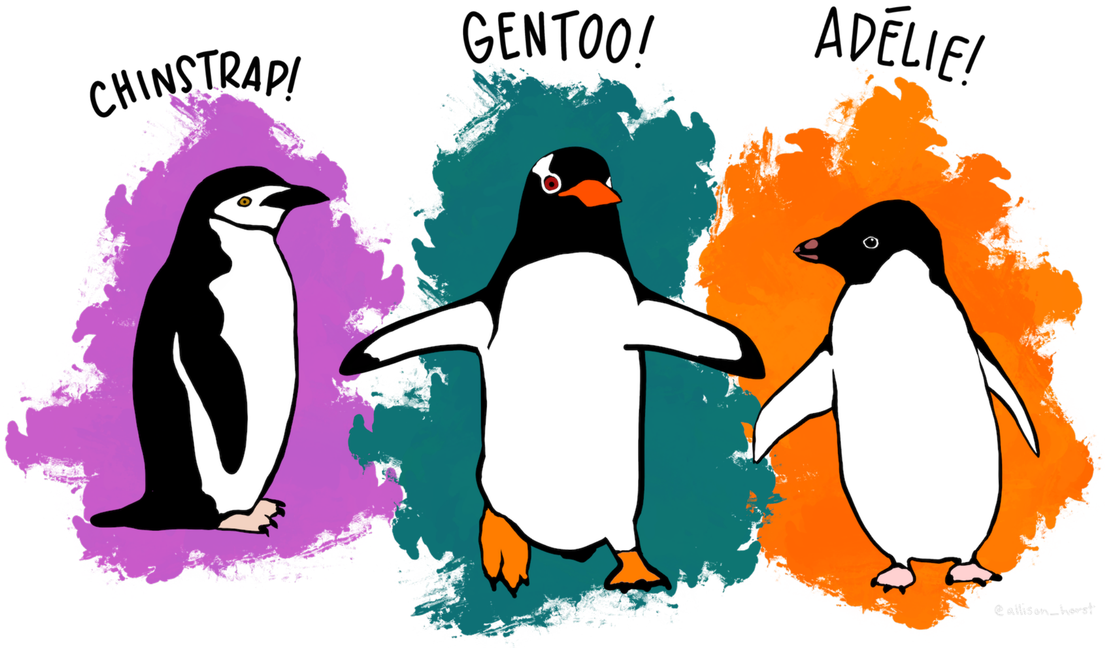
Useful data processing functions from `dplyr`

1. Working with observations (data rows)
2. Working with variables (data columns)
3. Summarizing the dataset

*But there's much, much more on my GitHub if you're interested!

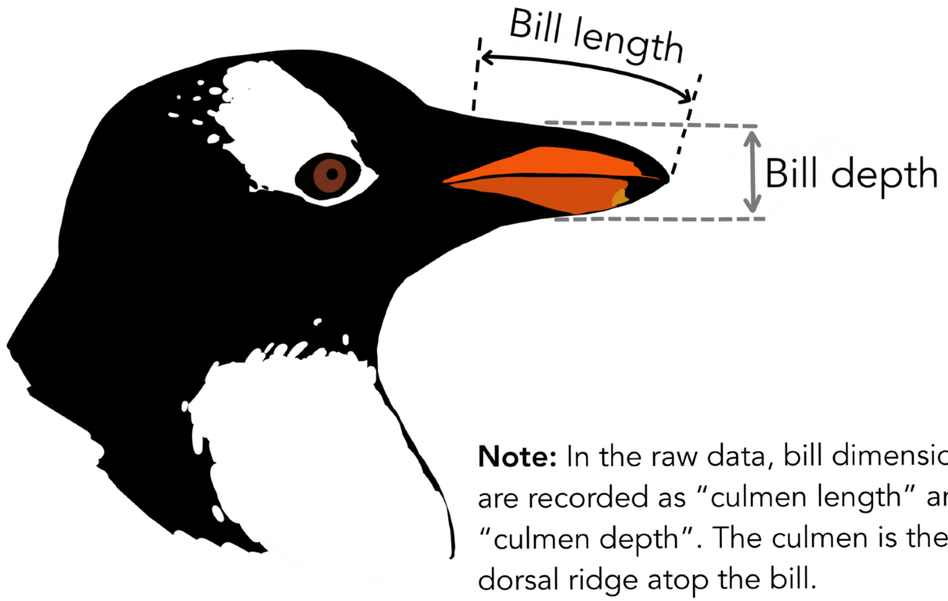
The `penguins` dataset

3 penguin species



Artwork by @allison_horst

Morphological measures



Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

Artwork by @allison_horst

Horst, AM, Hill AP, Gorman KB. 2020. `palmerpenguins`: Palmer Archipelago (Antarctica) penguin data. https://allisonhorst.github.io/palmerpenguins/

The `penguins` dataset

Snippet of the dataset:

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	NA	NA	NA	NA	NA	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007

Horst, AM, Hill AP, Gorman KB. 2020. `palmerpenguins`: Palmer Archipelago (Antarctica) penguin data.
https://allisonhorst.github.io/palmerpenguins/

1. Working with observations (data rows)

Separate out a particular species using **`filter()`**

```
gentoo <- filter(penguins, species==“Gentoo”)
```

Take a random sample of 25 observations using **`sample_n()`**

```
random_sample <- sample_n(penguins, 25, replace=FALSE)
```

2. Working with variables (data columns)

Select variables that describe the bill for each species using **`select()`**

```
bill_variables <- select(penguins, species, bill_length_mm,
     bill_depth_mm)
```

Create a new variable to describe bill shape using **`mutate()`**

```
new_penguins <- mutate(penguins, overall_bill=bill_depth_mm/
     bill_length_mm)
```

3. Summarizing the dataset

Summarize the average body mass for each species using **`group_by()`** and **`summarize()`**

```
summarize <- group_by(penguins, species) %>%
   summarize(avg_body_mass = mean(na.omit(body_mass_g)))
```

*More details on the pipe operator (%>%) on GitHub!

There's so much more to explore within the `tidyverse` packages!

stepfanie.aguillon@gmail.com @s_m_aguillon

www.stepfanieaguillon.com

More details and examples can be found on my GitHub page!

Please reach out with any questions!!

GitHub

https://github.com/stepfanie-aguillon/AOS2021-dplyr