Cool Stats on Penguins

Melinda Higgins

3/24/2021

## Palmer Penguins Dataset



palperpenguins logo

For this report, we will be working with the “Palmer Penguins” dataset. This dataset is built into the palmerpenguins package.

This dataset contains measurements and observations of a sample of Palmer Archipelago (Antarctica) Penguins.

The data includes: Size measurements, clutch observations, and blood isotope ratios for adult foraging Adélie, Chinstrap, and Gentoo penguins observed on islands in the Palmer Archipelago near Palmer Station, Antarctica. Data were collected and made available by Dr. Kristen Gorman and the Palmer Station Long Term Ecological Research (LTER) Program.

**Before you knit this report**, be sure to install these packages - go to “Tools” and choose “Install Packages” in RStudio.

* palmerpenguins
* dplyr (or tidyverse)
* knitr
* ggplot2 (or tidyverse)
* tinytex (optional to knit to PDF)

You can learn more about this cool dataset at:

* CRAN packages, <https://cran.r-project.org/web/packages/palmerpenguins/index.html>
* Github documentation (by Allison Horst), <https://allisonhorst.github.io/palmerpenguins/>

library(palmerpenguins)  
  
# create a local dataset  
# that is a copy of the builtin penguins dataset  
ppdata <- penguins

## What is in this dataset?

Show a summary of the variables in this dataset using the names() function. You can also use the str() structure function to get a list of the variables and what type of variables they are.

names(ppdata)

## [1] "species" "island" "bill\_length\_mm"   
## [4] "bill\_depth\_mm" "flipper\_length\_mm" "body\_mass\_g"   
## [7] "sex" "year"

str(ppdata)

## tibble [344 x 8] (S3: tbl\_df/tbl/data.frame)  
## $ species : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ island : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ bill\_length\_mm : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...  
## $ bill\_depth\_mm : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...  
## $ flipper\_length\_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...  
## $ body\_mass\_g : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...  
## $ sex : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...  
## $ year : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...

## Get summary statistics of 4 size measurements of the Penguins

library(dplyr)  
ppdata %>%  
 select(bill\_length\_mm, bill\_depth\_mm,   
 flipper\_length\_mm, body\_mass\_g) %>%  
 summary()

## bill\_length\_mm bill\_depth\_mm flipper\_length\_mm body\_mass\_g   
## Min. :32.10 Min. :13.10 Min. :172.0 Min. :2700   
## 1st Qu.:39.23 1st Qu.:15.60 1st Qu.:190.0 1st Qu.:3550   
## Median :44.45 Median :17.30 Median :197.0 Median :4050   
## Mean :43.92 Mean :17.15 Mean :200.9 Mean :4202   
## 3rd Qu.:48.50 3rd Qu.:18.70 3rd Qu.:213.0 3rd Qu.:4750   
## Max. :59.60 Max. :21.50 Max. :231.0 Max. :6300   
## NA's :2 NA's :2 NA's :2 NA's :2

## Show categories for “Factor” variables: species, island and sex

library(knitr)  
ppdata %>%  
 pull(species) %>%  
 table(useNA = "ifany") %>%  
 knitr::kable(caption = "Penguin Species")

Penguin Species

|  |  |
| --- | --- |
| . | Freq |
| Adelie | 152 |
| Chinstrap | 68 |
| Gentoo | 124 |

ppdata %>%  
 pull(island) %>%  
 table(useNA = "ifany") %>%  
 knitr::kable(caption = "Penguin Island Location")

Penguin Island Location

|  |  |
| --- | --- |
| . | Freq |
| Biscoe | 168 |
| Dream | 124 |
| Torgersen | 52 |

ppdata %>%  
 pull(sex) %>%  
 table(useNA = "ifany") %>%  
 knitr::kable(caption = "Penguin Sex")

Penguin Sex

|  |  |
| --- | --- |
| . | Freq |
| female | 165 |
| male | 168 |
| NA | 11 |

## Get stats for only Adelie penguins

I added knitr::kable(caption = "Summary Stats for Adelie Penguins") to make a prettier table with a caption title.

ppdata %>%  
 filter(species == "Adelie") %>%  
 select(bill\_length\_mm, bill\_depth\_mm,   
 flipper\_length\_mm, body\_mass\_g) %>%  
 summary() %>%  
 knitr::kable(caption = "Summary Stats for Adelie Penguins")

Summary Stats for Adelie Penguins

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | bill\_length\_mm | bill\_depth\_mm | flipper\_length\_mm | body\_mass\_g |
|  | Min. :32.10 | Min. :15.50 | Min. :172 | Min. :2850 |
|  | 1st Qu.:36.75 | 1st Qu.:17.50 | 1st Qu.:186 | 1st Qu.:3350 |
|  | Median :38.80 | Median :18.40 | Median :190 | Median :3700 |
|  | Mean :38.79 | Mean :18.35 | Mean :190 | Mean :3701 |
|  | 3rd Qu.:40.75 | 3rd Qu.:19.00 | 3rd Qu.:195 | 3rd Qu.:4000 |
|  | Max. :46.00 | Max. :21.50 | Max. :210 | Max. :4775 |
|  | NA’s :1 | NA’s :1 | NA’s :1 | NA’s :1 |

## Get stats for the **Chinstrap** species penguins

* Change the species name in the filter.
* Remember to update the caption title.

# Change filter(species = "Chinstrap")  
# and change   
# knitr::kable(caption = "Summary Stats for Chinstrap Penguins")  
ppdata %>%  
 filter(species == "Chinstrap") %>%  
 select(bill\_length\_mm, bill\_depth\_mm,   
 flipper\_length\_mm, body\_mass\_g) %>%  
 summary() %>%  
 knitr::kable(caption = "Summary Stats for Chinstrap Penguins")

Summary Stats for Chinstrap Penguins

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | bill\_length\_mm | bill\_depth\_mm | flipper\_length\_mm | body\_mass\_g |
|  | Min. :40.90 | Min. :16.40 | Min. :178.0 | Min. :2700 |
|  | 1st Qu.:46.35 | 1st Qu.:17.50 | 1st Qu.:191.0 | 1st Qu.:3488 |
|  | Median :49.55 | Median :18.45 | Median :196.0 | Median :3700 |
|  | Mean :48.83 | Mean :18.42 | Mean :195.8 | Mean :3733 |
|  | 3rd Qu.:51.08 | 3rd Qu.:19.40 | 3rd Qu.:201.0 | 3rd Qu.:3950 |
|  | Max. :58.00 | Max. :20.80 | Max. :212.0 | Max. :4800 |

## Get stats for the penguins on the **Dream island**

* Change the filter for island instead of species and specify the “Dream” island.
* Remember to update the caption title.

# Change filter(island = "Dream")  
# and change   
# knitr::kable(caption = "Summary Stats for Penguins on Dream Island")  
ppdata %>%  
 filter(island == "Dream") %>%  
 select(bill\_length\_mm, bill\_depth\_mm,   
 flipper\_length\_mm, body\_mass\_g) %>%  
 summary() %>%  
 knitr::kable(caption = "Summary Stats for Penguins on Dream Island")

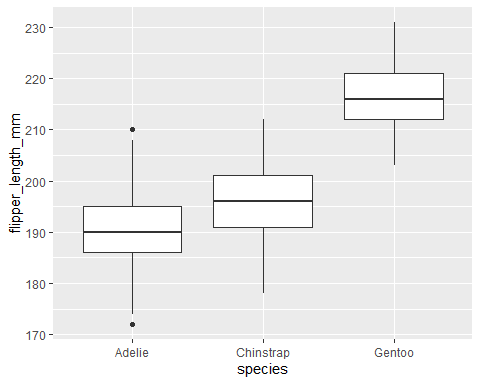
Summary Stats for Penguins on Dream Island

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | bill\_length\_mm | bill\_depth\_mm | flipper\_length\_mm | body\_mass\_g |
|  | Min. :32.10 | Min. :15.50 | Min. :178.0 | Min. :2700 |
|  | 1st Qu.:39.15 | 1st Qu.:17.50 | 1st Qu.:187.8 | 1st Qu.:3400 |
|  | Median :44.65 | Median :18.40 | Median :193.0 | Median :3688 |
|  | Mean :44.17 | Mean :18.34 | Mean :193.1 | Mean :3713 |
|  | 3rd Qu.:49.85 | 3rd Qu.:19.00 | 3rd Qu.:198.0 | 3rd Qu.:3956 |
|  | Max. :58.00 | Max. :21.20 | Max. :212.0 | Max. :4800 |

## Let’s make some plots - boxplot of flipper\_length\_mm

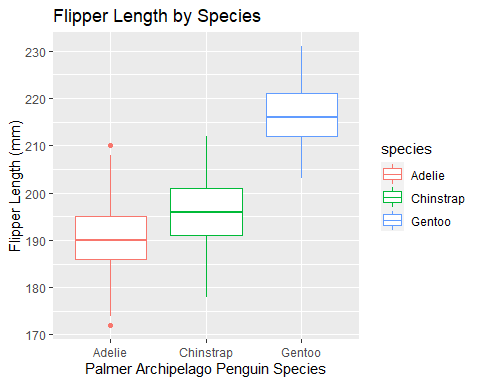
Here is a boxplot of the flipper lengths of the penguins by species.

library(ggplot2)  
ggplot(data = ppdata,  
 aes(x = species, y = flipper\_length\_mm)) +  
 geom\_boxplot()



We can make the plot a little nicer by updating the axis labels, adding a title and adding some color using the aes aesthetic.

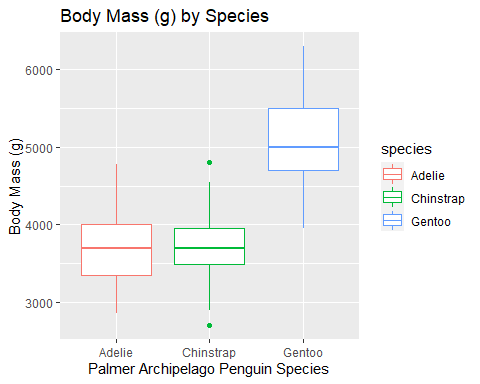
ggplot(data = ppdata,  
 aes(x = species, y = flipper\_length\_mm)) +  
 geom\_boxplot(aes(color = species)) +  
 xlab("Palmer Archipelago Penguin Species") +  
 ylab("Flipper Length (mm)") +  
 ggtitle("Flipper Length by Species")



## Make a boxplot of body\_mass\_g by Species

Use the code above as your guide to make another boxplot of the Body Mass (in grams) for the 3 species of penguins. Set y = body\_mass\_g. Remember to update your y-axis label.

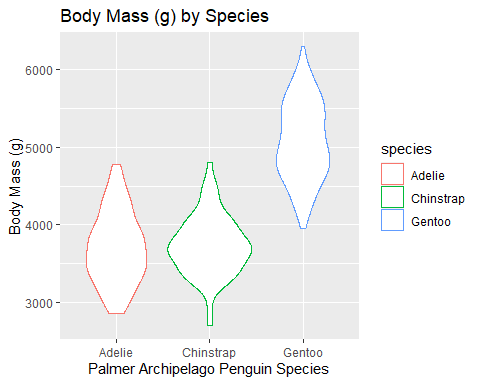
# Update y = body\_mass\_g  
# and update ggtitle("Body Mass (g) by Species")  
ggplot(data = ppdata,  
 aes(x = species, y = body\_mass\_g)) +  
 geom\_boxplot(aes(color = species)) +  
 xlab("Palmer Archipelago Penguin Species") +  
 ylab("Body Mass (g)") +  
 ggtitle("Body Mass (g) by Species")



## Change the “geom” and make a new plot

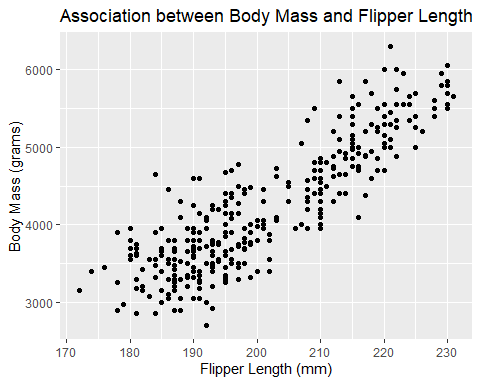
Now take the code you wrote above and change geom\_boxplot to geom\_violin and see what happens.

# change geom\_boxplot()  
# to geom\_violin(aes(color = species))  
ggplot(data = ppdata,  
 aes(x = species, y = body\_mass\_g)) +  
 geom\_violin(aes(color = species)) +  
 xlab("Palmer Archipelago Penguin Species") +  
 ylab("Body Mass (g)") +  
 ggtitle("Body Mass (g) by Species")



## Make a scatterplot of body mass by flipper length

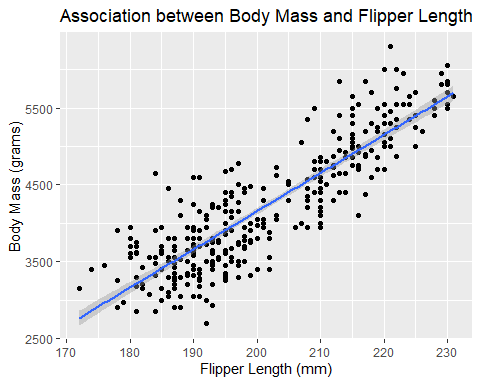
ggplot(data = ppdata,  
 aes(x = flipper\_length\_mm, y = body\_mass\_g)) +  
 geom\_point() +  
 xlab("Flipper Length (mm)") +  
 ylab("Body Mass (grams)") +  
 ggtitle("Association between Body Mass and Flipper Length")



## Add best fit line

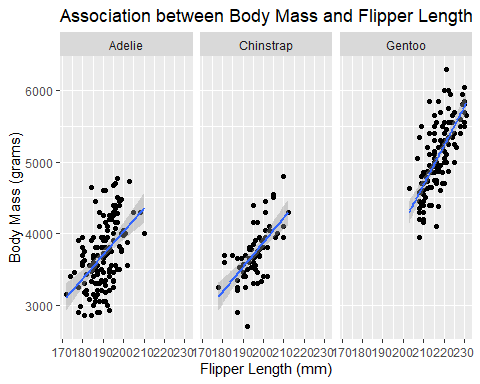
Add a best fit line by adding geom\_smooth(method = lm) which adds the “linear model” simple linear regression line.

ggplot(data = ppdata,  
 aes(x = flipper\_length\_mm, y = body\_mass\_g)) +  
 geom\_point() +  
 geom\_smooth(method = lm) +  
 xlab("Flipper Length (mm)") +  
 ylab("Body Mass (grams)") +  
 ggtitle("Association between Body Mass and Flipper Length")



## Add panels or facets by species

ggplot(data = ppdata,  
 aes(x = flipper\_length\_mm, y = body\_mass\_g)) +  
 geom\_point() +  
 geom\_smooth(method = lm) +  
 xlab("Flipper Length (mm)") +  
 ylab("Body Mass (grams)") +  
 ggtitle("Association between Body Mass and Flipper Length") +  
 facet\_wrap(~ species)



## or color points and lines by species

ggplot(data = ppdata,  
 aes(x = flipper\_length\_mm, y = body\_mass\_g,  
 color = species)) +  
 geom\_point() +  
 geom\_smooth(method = lm) +  
 xlab("Flipper Length (mm)") +  
 ylab("Body Mass (grams)") +  
 ggtitle("Association between Body Mass and Flipper Length")

