Case Study 2

Instructions:

Get into groups of 3 and answer the following questions. Submit your knitted report as a pdf/html file and include the original .Rmd file also. Submit one report per group with all members' names on it before the end of class. You currently have the fundamental R tools to complete this exercise, but you will may still have to explore new techniques and packages. **For each question, write the name of the team member who answered/coded it.**

1. *Text analysis.* Using the tweets.csv data that is available on the GitHub site, provide code to do the following:

a. Identify all tweets with the word 'flight' in them. Print them

```
tweets = read.csv("../Data/tweets.csv", header = FALSE)

tweet_vec = tweets$V1 %>% as.vector()

idx_a = str_detect(tweet_vec, pattern = "flight")
tweet_vec[idx_a]
```

```
## [1] "@VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing."
## [2] "@VirginAmerica So excited for my first cross country flight LAX to MCO I've heard nothing but gr
## [3] "@VirginAmerica amazing to me that we can't get any cold air from the vents. #VX358 #noair #worst
```

b. How many tweets end in a question mark? Print them.

```
idx_b = str_detect(tweet_vec, pattern = "\\?$")
sum(idx_b) #tweets end in a ?
```

```
## [1] 2
```

```
tweet_vec[idx_b]
```

```
## [1] "@VirginAmerica why are your first fares in May over three times more than other carriers when al
## [2] "@VirginAmerica will you be making BOS&gt;LAS non stop permanently anytime soon?"
```

c. How many tweets have airport codes in them (assume any three subsequent capital letters are airport codes). Print them.

```
idx_c = str_detect(tweet_vec, pattern = '[A-Z]{3}')
sum(idx_c) #tweets have airport codes in them.
```

```
## [1] 8
```

```
tweet_vec[idx_c]
```

```
## [1] "@VirginAmerica Really missed a prime opportunity for Men Without Hats parody, there. https://t.c
## [2] "@VirginAmerica SFO-PDX schedule is still MIA."
## [3] "@VirginAmerica So excited for my first cross country flight LAX to MCO I've heard nothing but gr
## [4] "@VirginAmerica  I flew from NYC to SFO last week and couldn't fully sit in my seat due to two la
## [5] "@VirginAmerica you know what would be amazingly awesome? BOS-FLL PLEASE!!!!!!! I want to fly wit
## [6] "@VirginAmerica will you be making BOS&gt;LAS non stop permanently anytime soon?"
## [7] "@VirginAmerica amazing to me that we can't get any cold air from the vents. #VX358 #noair #worst
## [8] "@VirginAmerica LAX to EWR - Middle seat on a red eye. Such a noob maneuver. #sendambien #andcher
```

d. Identify all tweets with URLs in them.

```
idx_d = str_detect(tweet_vec, pattern = 'http')

tweet_vec[idx_d]
```

```
## [1] "@VirginAmerica Really missed a prime opportunity for Men Without Hats parody, there. https://t.
## [2] "@VirginAmerica @virginmedia I'm flying your #fabulous #Seductive skies again! U take all the #s
## [3] "@VirginAmerica I love this graphic. http://t.co/UT5GrRwAaA"
```

e. Replace all instances of repeated exclamation points with a single exclamation point. Identify which tweet(s) had changes made and print them before and after.

```
idx_e = str_detect(tweet_vec, pattern = '!!!')
tweet_vec[idx_e]
```

```
## [1] "@VirginAmerica you know what would be amazingly awesome? BOS-FLL PLEASE!!!!!!! I want to fly wi
```

```
tweet_vec_replaced = str_replace(tweet_vec, "[!]+", "!")
tweet_vec_replaced[idx_e]
```

```
## [1] "@VirginAmerica you know what would be amazingly awesome? BOS-FLL PLEASE! I want to fly with only
```

f. Replace any consecutive exclamation points, periods, and question marks, with a single period. Print the changed tweets before and after.

```
#which ones need replacing?
idx_f = str_detect(tweet_vec, "(!!)|(\\.\\.)|(\\?\\?)")
tweet_vec[idx_f]
```

```
## [1] "@VirginAmerica plus you've added commercials to the experience... tacky."
## [2] "@VirginAmerica I didn't today... Must mean I need to take another trip!"
## [3] "@VirginAmerica you know what would be amazingly awesome? BOS-FLL PLEASE!!!!!!! I want to fly wi
## [4] "@VirginAmerica why are your first fares in May over three times more than other carriers when al
```

```
tweet_vec_periods = str_replace(tweet_vec, "![!]+", ".") #exc
tweet_vec_periods = str_replace(tweet_vec_periods, "\\?[\\?]+", ".") #questions
tweet_vec_periods = str_replace(tweet_vec_periods, "\\.[\\.]+", ".") #periods

tweet_vec_periods[idx_f]
```

```
## [1] "@VirginAmerica plus you've added commercials to the experience. tacky."
## [2] "@VirginAmerica I didn't today. Must mean I need to take another trip!"
## [3] "@VirginAmerica you know what would be amazingly awesome? BOS-FLL PLEASE. I want to fly with only
## [4] "@VirginAmerica why are your first fares in May over three times more than other carriers when al
```

g. Continue with the transformed tweets from f. Split the tweets on periods **excluding those in URLS**. Create a list where each element is a vector of the split strings from each tweet.

```
list = tweet_vec_periods %>% str_split("(\\. )")
```

h. Create a data.frame called `transformed_tweets` using all the tweets after performing the transformation in part f. Write this out as a .csv file similar to the original `tweets.csv` file. Name the csv `transformed_tweets.csv` and submit it along with your knitted .Rmd file.

```
transformed_tweets = data.frame(tweets = tweet_vec_periods)

write.csv(transformed_tweets, file = "../Data/transformed_tweets.csv")
```

2. *Data Visualization.* The R package `palmerpenguins` has a dataset called `penguins` in which you will find measurements of 344 penguins, collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica. 3 different species of penguins were observed. We will visualize this data and learn about these species of penguins using ggplot. For each plot, make sure to include labels for axes, legends, and plot titles in plain language (not code speak).

a. Install the R package `palmerpenguins`. Check for missing data. Remove any rows which have missing data using the `complete.cases` function.

```
#install.packages("palmerpenguins")
library(palmerpenguins)

idx = complete.cases(penguins)
data = penguins[idx,]
```

b. Make a violin plot of the bill depth for each species of penguins, giving each violin a different color per species. Set the opacity of each violin to 0.5. Describe the distribution for each species; what do you find?

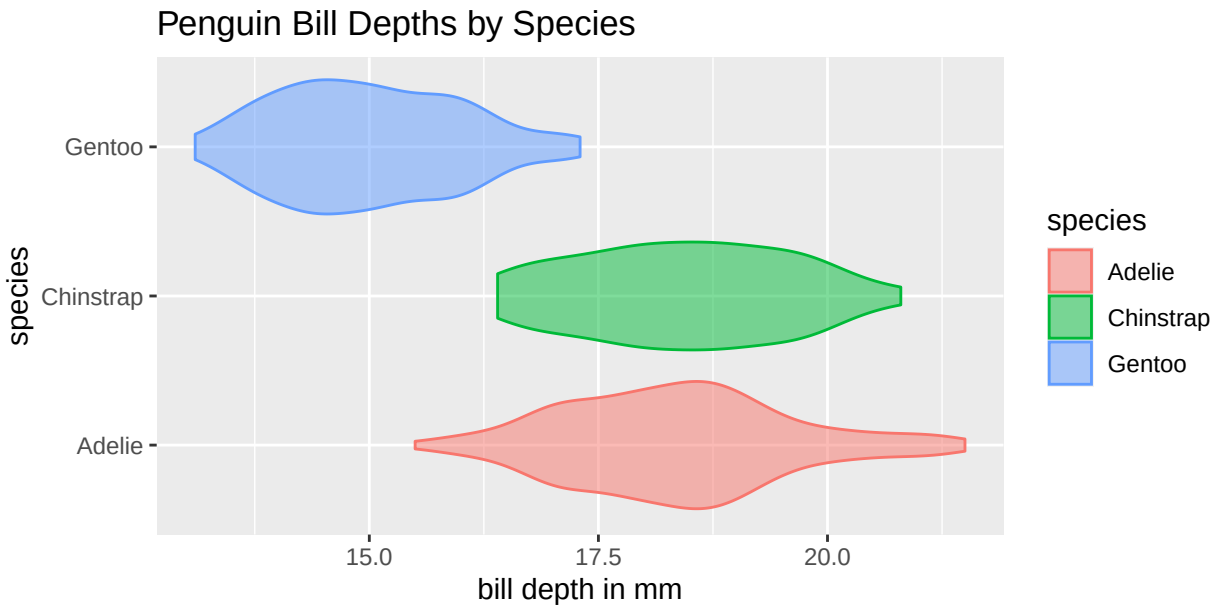
Repeat this for another violin plot of body mass.

(Hint: you can use `{r, fig.asp=1}` and change the value of `1` to play with the printed aspect ratio for figures in .Rmd)

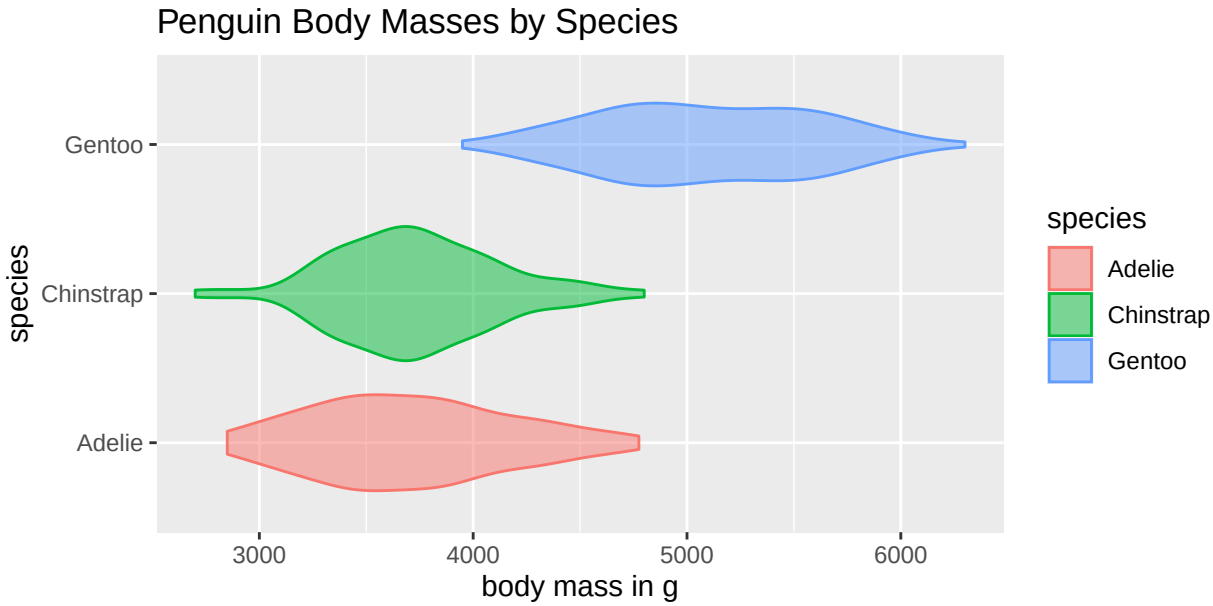
```
gg1 = ggplot(data, aes(x = species,
                       y = bill_depth_mm,
                       fill = species,
                       colour = species)) +
  geom_violin(alpha = 0.5) +
  guides(alpha = FALSE) +
  coord_flip() +
  ylab("bill depth in mm") +
  ggtitle("Penguin Bill Depths by Species")
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
gg2 = ggplot(data, aes(x = species,
                       y = body_mass_g,
                       fill = species,
                       colour = species)) +
  geom_violin(alpha = 0.5) +
  guides(alpha = FALSE) +
  coord_flip() +
  ylab("body mass in g")+
  ggtitle("Penguin Body Masses by Species")

gg1
```




```
gg2
```



Bill Depths: Adelie and Chinstrap penguins have bill depths centered around ~18mm and have roughly equal spread. Gentoo penguins' bills are smaller on average with bill depths centered around 14-15mm and a similar, if slightly smaller, spread than the other species.

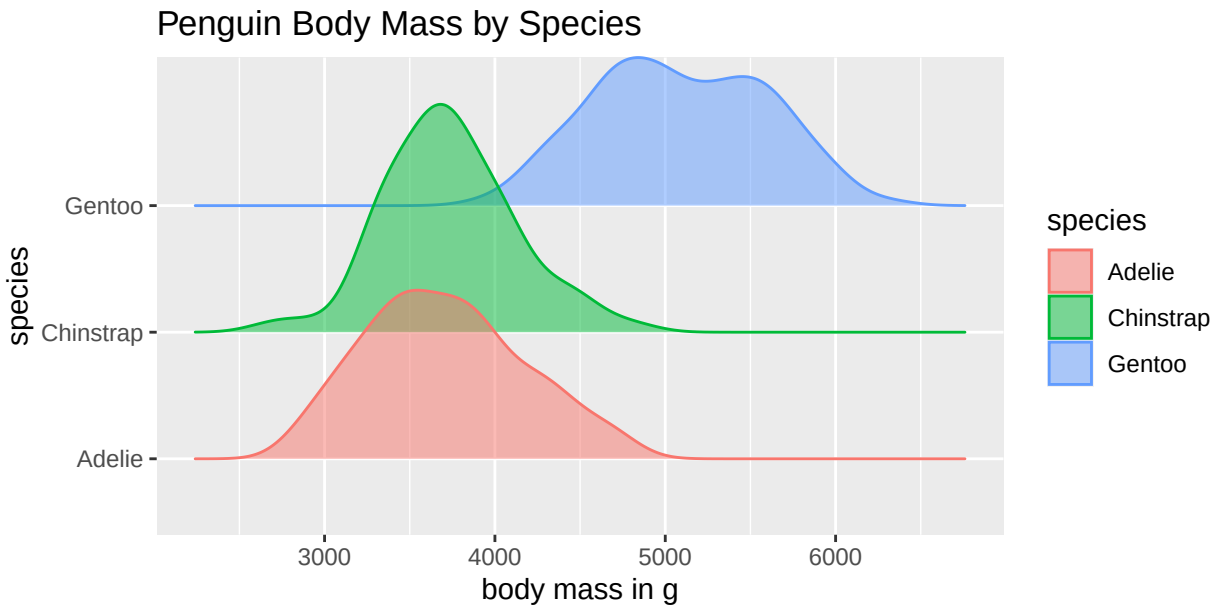
Body Masses: Adelie and Chinstrap penguins have body masses centered around 3,600g with roughly equal spread. Gentoo penguins on average are much larger, with body masses centered around 5200g and slightly more variability than the other species.

c. Create a density plot of the body mass for each species of penguin and visualize them in two different ways: 1. with a ridge plot, 2. with a facet. What does your plot tell you about the sizes of the penguin species? Which plot do you think helps illustrate this message best?

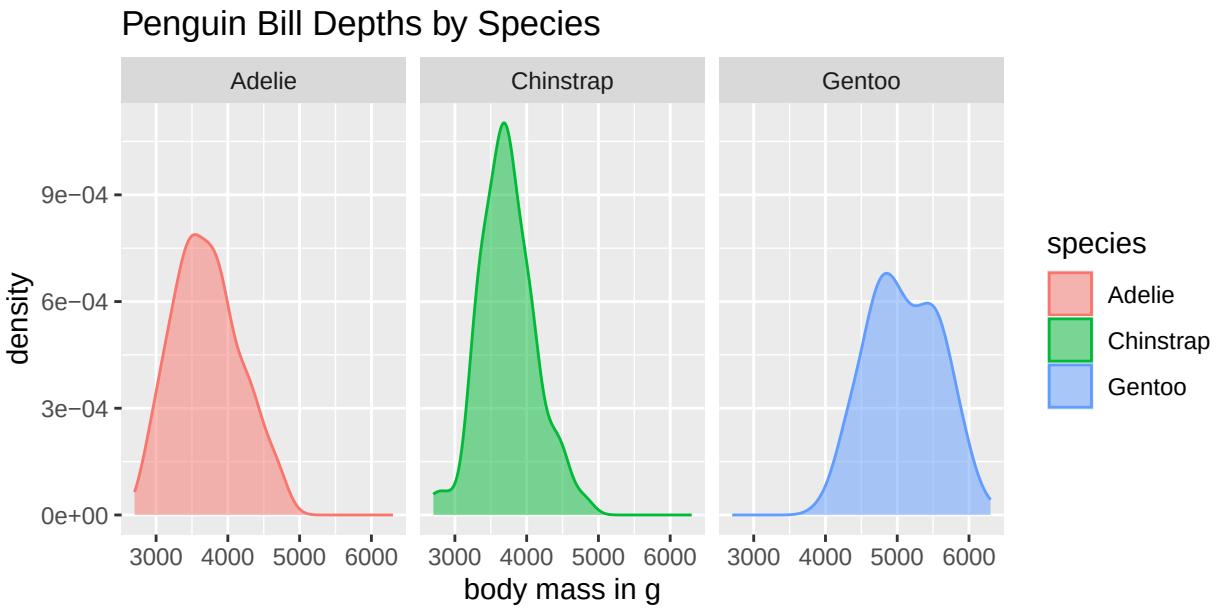
```
library(ggridges)

ridges = ggplot(data,
```

```
        aes(x = body_mass_g,
            y = species,
            fill = species,
            colour = species)) +
  geom_density_ridges(alpha = 0.5) +
  coord_cartesian(clip = "off")+
  xlab("body mass in g")+
  ggtitle("Penguin Body Mass by Species")

ridges
```

```
## Picking joint bandwidth of 153
```



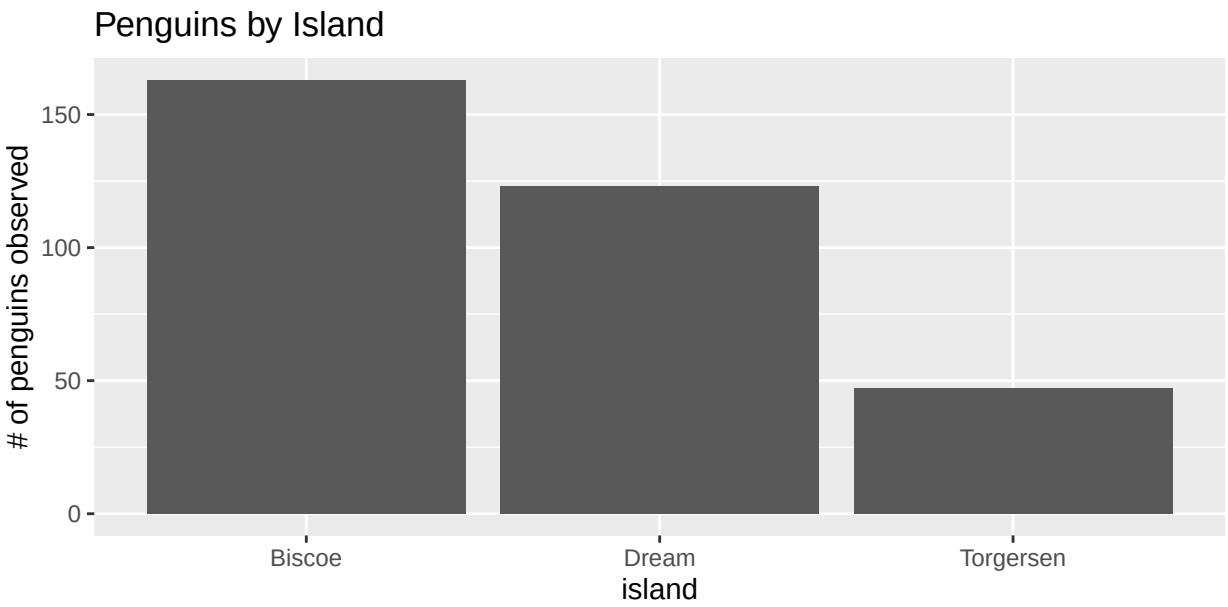

```
facet = ggplot(data,
               aes(x = body_mass_g,
                   fill = species,
                   colour = species))+
        geom_density(alpha = 0.5) +
        facet_wrap(~species) +
        xlab("body mass in g")+
        ggtitle("Penguin Bill Depths by Species")

facet
```



These plots reinforce the interpretation that Gentoo penguins are larger on average- they show the same information as a violin plot. I prefer the ridge plot as the x-axis is shared across the 3 ridges allowing for easy comparison of the distribution across species. I also don't think a facet is needed here since we don't have an overplotted figure with the ridge plot.

d. There are 3 islands on which the penguins can be found: Biscoe, Dream, and Torgesen. Create a bar plot to show the counts of the penguins on each island. Interpret the plot. Which island had the most penguin observations? The least?

```
ggplot(data) +
  geom_bar(aes(x = island)) +
  ggtitle("Penguins by Island") +
  ylab("# of penguins observed")
```

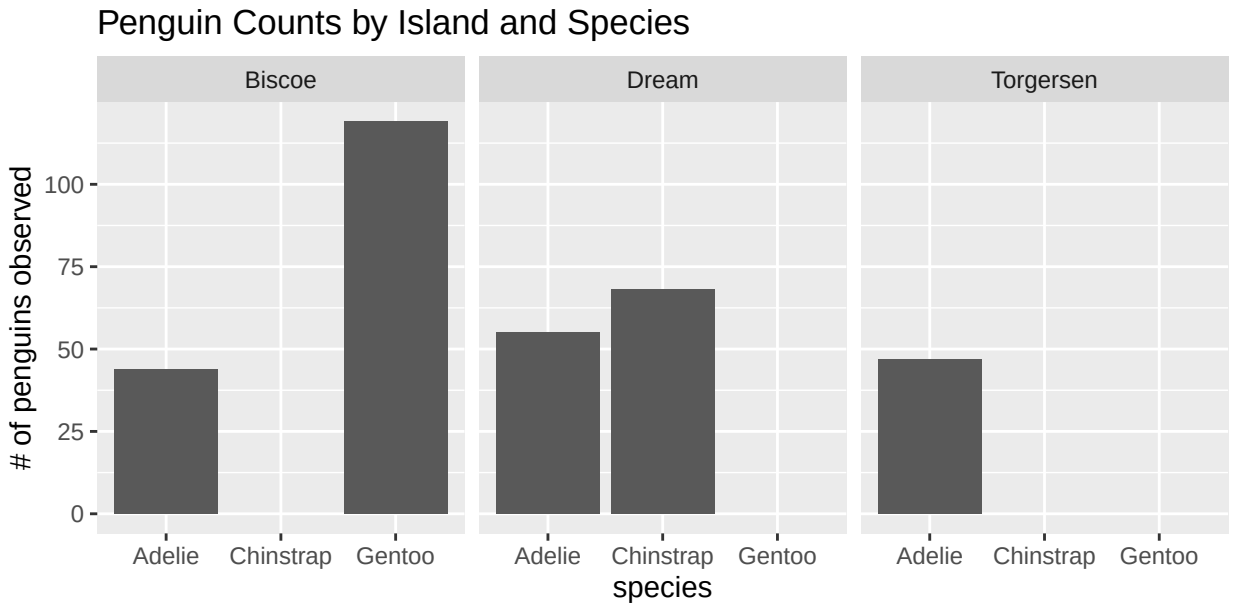


Biscoe island had the most penguin observations (>150) while Torgersen had the fewest (<50).

e. We want to examine the biodiversity of each island. Modify your bar plot with a facet create a bar

plot of species counts where each island gets its own subplot. Interpret the plot. Which island is least biodiverse? Which species of penguins are restricted to living on a single island?

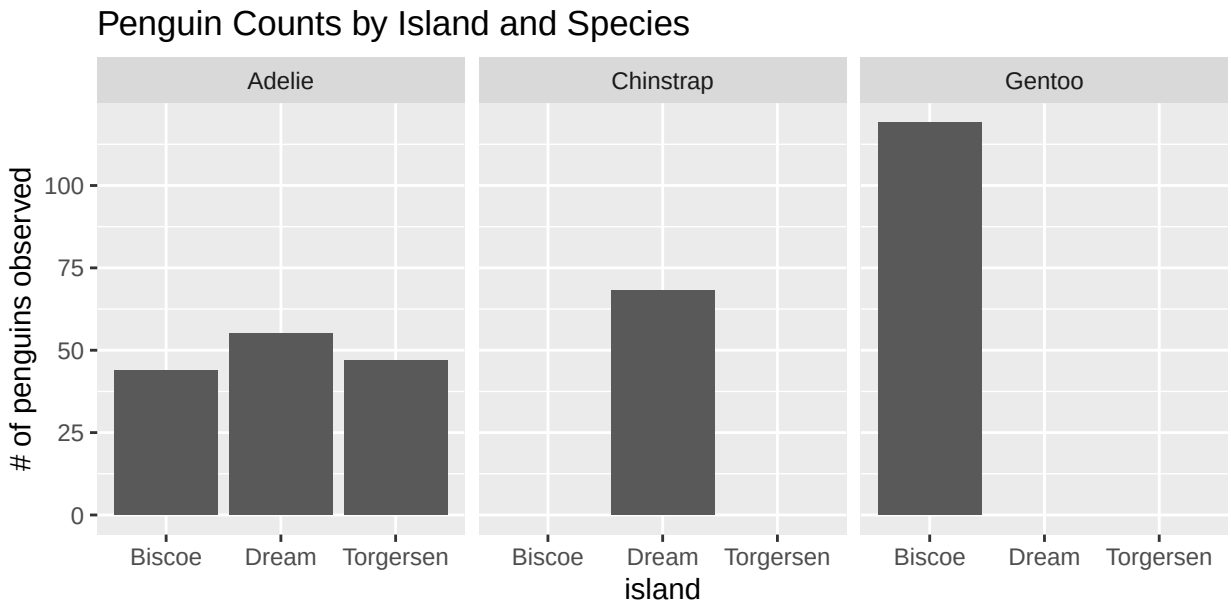
```
ggplot(data) +
  geom_bar(aes(x = species)) +
  facet_grid(~island) +
  ggtitle("Penguin Counts by Island and Species") +
  ylab("# of penguins observed")
```



If we facet on island, we see that Torgersen Island is the least biodiverse island (for penguins at least)- only Adelie penguins are observed here. Gentoos only live on Biscoe Island and Chinstraps only live on Dream Island, according to our observations. To be sure of this claim, we would want to inquire about how the data were collected to make sure we had an accurate sample of penguins on each island.

We could also visualize the range of each penguin species by faceting on the species level.

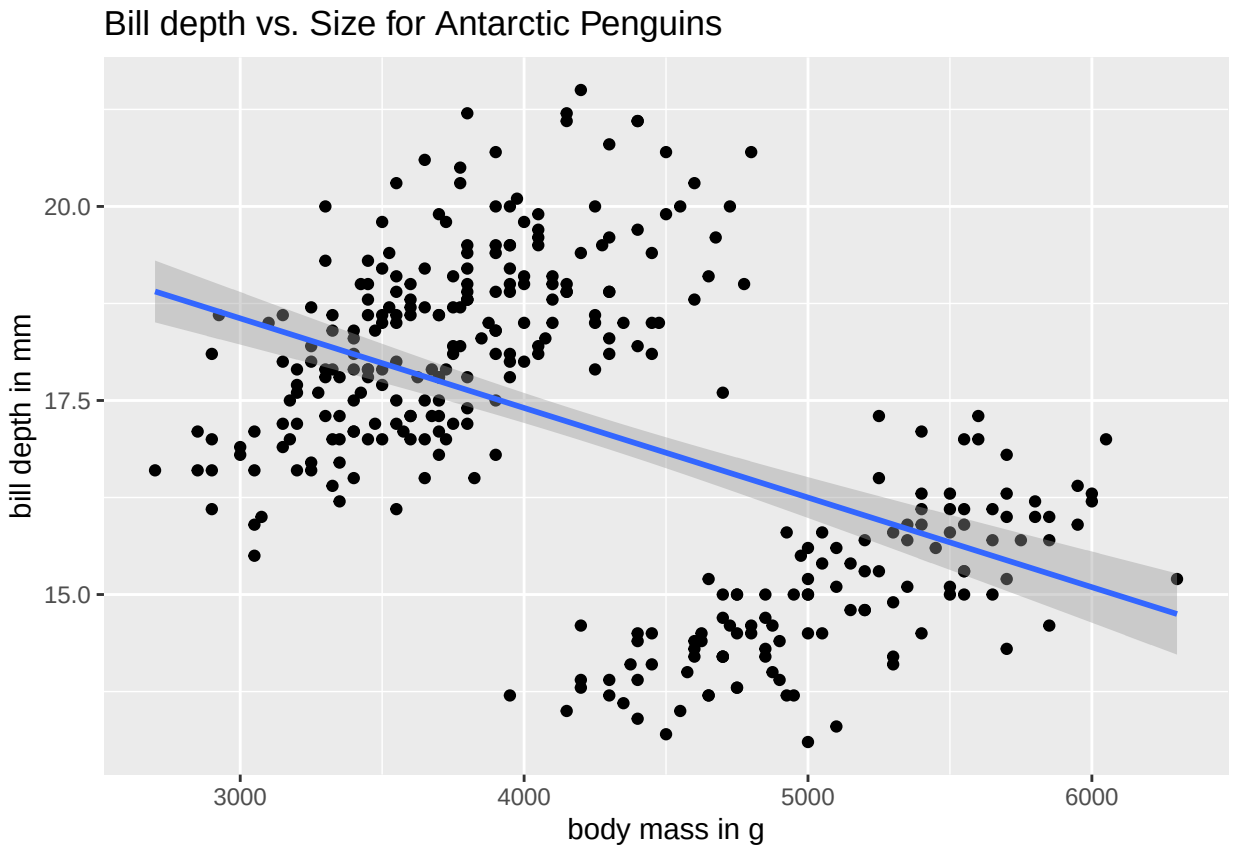
```
ggplot(data) +
  geom_bar(aes(x = island)) +
  facet_grid(~species) +
  ggtitle("Penguin Counts by Island and Species") +
  ylab("# of penguins observed")
```



f. Create a scatterplot of body mass (x-axis) vs. bill depth (y-axis). Leave the color of the points as black (i.e. do not use `colour` aesthetic). Add a line of best fit to the data points using the geom_smooth with the `lm` method. Interpret the plot. What does the scatterplot and best fit line suggest about the relationship between body mass and bill depth?

```
ggplot(data, aes(x = body_mass_g,
                 y = bill_depth_mm)) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab('body mass in g') +
  ylab('bill depth in mm') +
  ggtitle('Bill depth vs. Size for Antarctic Penguins')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

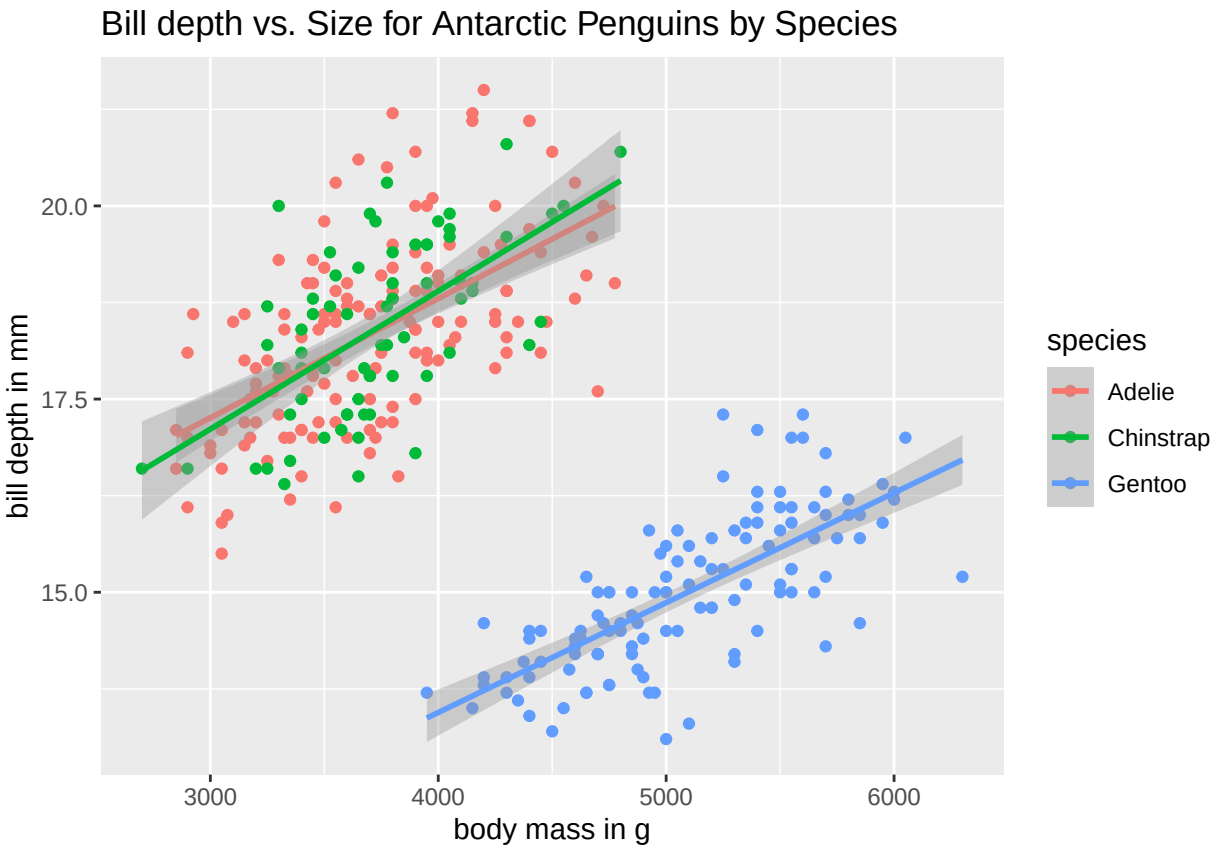


The scatterplot shows a weak negative association between body mass and bill depth overall. The line of best fit suggests that as body mass increases, bill depth tends to decrease. This claim is somewhat strange – generally we think that larger animals will have bigger anatomical features like bill depth. We further question the negative association by observing that the points form 2 clusters, and the best fit line does not really capture their trend very well. We are suspicious of the claim that as body mass increases, bill depth tends to decrease.

g. Modify your scatterplot to color the points based on the different species of the penguin. Draw a best fit line for each species individually. Interpret the plot. Does the figure's message regarding the relationship between body mass and bill depth change if we look at each species separately? If yes, argue why this might have happened. (Hint: reflect on part b. and c.'s results.)

```
ggplot(data, aes(x = body_mass_g,
                 y = bill_depth_mm,
                 colour = species)) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab('body mass in g') +
  ylab('bill depth in mm') +
  ggtitle('Bill depth vs. Size for Antarctic Penguins by Species')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



When we separate by on the species level, each individual species shows a positive association between body mass and bill depth. This is counter to our finding before. This suggests that if we condition on the species level, we observe that bill depths grow as penguins get larger and larger. This makes more biological sense. In addition, the conditional lines of best fit per species seem to better capture the trend in each of the two clusters.

h. Between the 2 figures you made in parts f. and g., which plot do you believe gives a more truthful and accurate description of the relationship between body mass and bill depth. Argue why you think so.

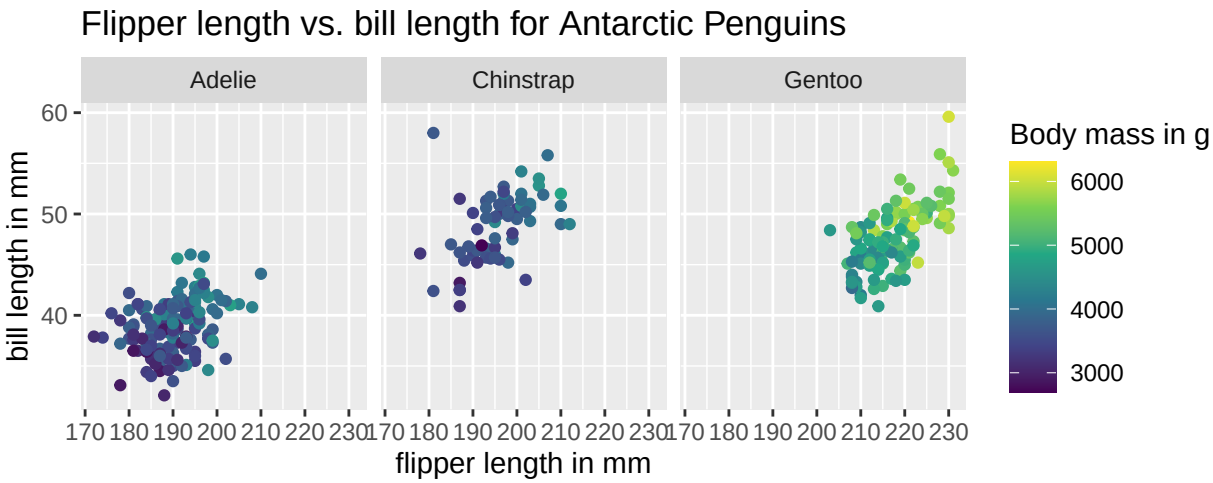
Notice that the lower cluster is all Gentoo penguins, which tend to have larger body masses and smaller bill depths compared to other species. If we group all species together, this creates the illusion that bill depths decrease as penguins get larger in mass, when really this effect is caused by making a jump across species (no Adelies or Chinstraps are ever observed to have body mass greater than 5000g, for example). We say that species is a **confounder** variable and the overall effect is called Simpson's paradox or the reversal paradox.

I believe that it makes more sense biologically and statistically to first condition on species and then examine the relationship between body mass and bill depth. We've controlled for our confounder variable, we arrive at a biologically consistent interpretation, and our best fit lines fit the data clouds better.

i. Create a scatterplot of flipper length (x-axis) vs. bill length (y-axis). Facet the plot by species and change the point color to correspond to the *body mass* of the penguin. Use the viridis color scale. Interpret the plot. How could you "double up" on an aesthetic to reinforce the visualization of body mass?

```
ggplot(data, aes(x = flipper_length_mm,
                 y = bill_length_mm,
                 colour = body_mass_g)) +
  geom_point() +
  facet_grid(~species)+
```

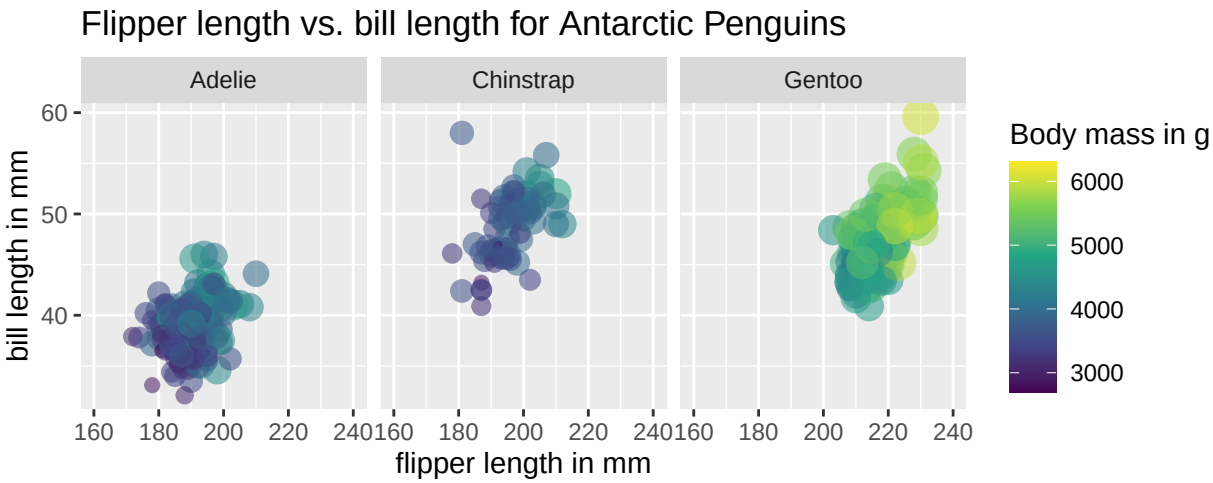
```
  scale_color_viridis_c()+
  xlab('flipper length in mm') +
  ylab('bill length in mm') +
  ggtitle('Flipper length vs. bill length for Antarctic Penguins') +
  labs(colour = "Body mass in g")
```



As flipper length increases, so does bill length, on average. This is true for all species. We also notice that body mass increases as each of these morphological features increases in length. This makes intuitive sense– longer flippers and longer bills make up more mass!

To reinforce the body mass message, we could double up by setting both `size` and `colour` aesthetics to be based on body mass. Then we arrive at the following plot:

```
ggplot(data, aes(x = flipper_length_mm,
                 y = bill_length_mm,
                 colour = body_mass_g,
                 size = body_mass_g,
                 alpha = 0.5)) +
  geom_point() +
  facet_grid(~species) +
  scale_color_viridis_c() +
  expand_limits(x = c(160, 240)) +
  xlab('flipper length in mm') +
  ylab('bill length in mm') +
  ggtitle('Flipper length vs. bill length for Antarctic Penguins') +
  guides(size = FALSE, alpha = FALSE) +
  labs(colour = "Body mass in g")
```


Flipper length vs. bill length for Antarctic Penguins
Adelie
Chinstrap
Gentoo
bill length in mm
60
50
40
160
180
200
220
240
flipper length in mm
Body mass in g
6000
5000
4000
3000