Homework 7

Jeremy Ulfohn // jau392

This homework is due on April 12, 2021 at 11:00pm. Please submit as a pdf file on Canvas.

For all problems in this homework, we will work with the `penguins_clean` dataset, which is a cleaned-up version of the `penguins` dataset from the **palmerpenguins** package.

Note: This homework is about the contents of the plots. Don't worry about styling. It's OK to use the default theme and plot labeling.

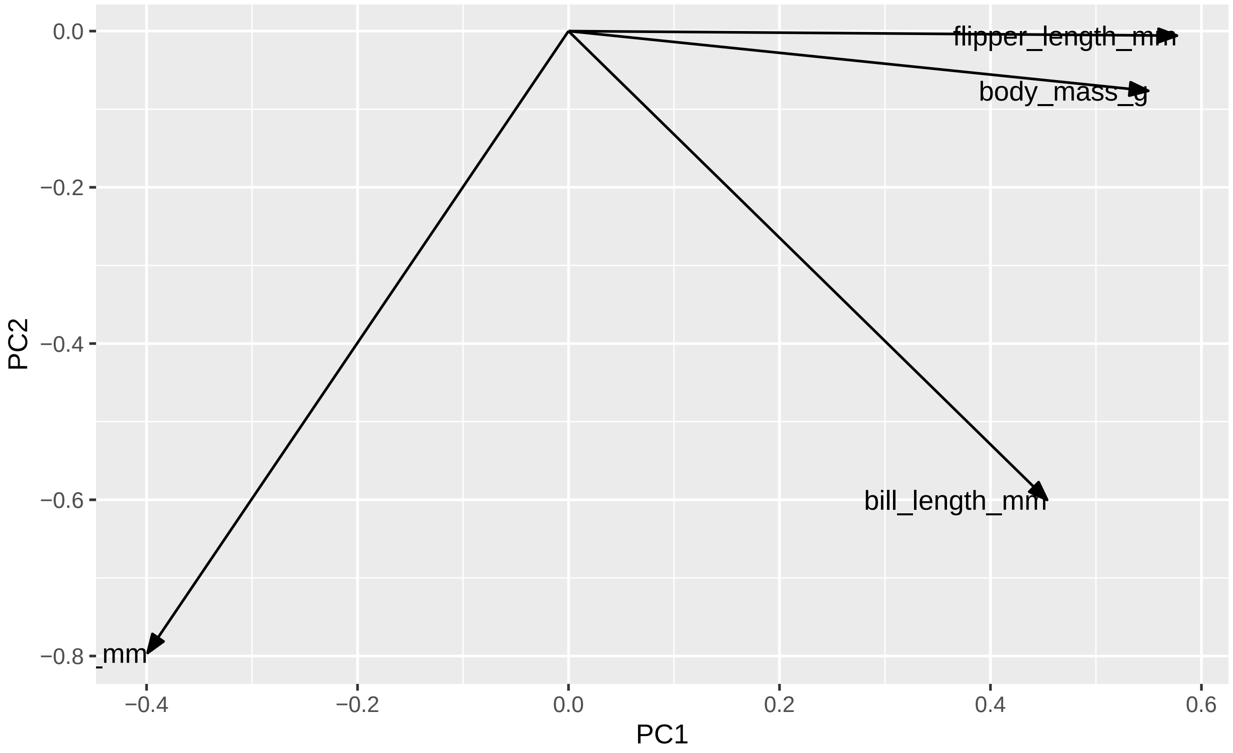
```r
library(palmerpenguins)

penguins_clean <- penguins %>%
  select(-year) %>% # remove the year column as it is distracting here
  na.omit()         # remove any rows with missing values!!

penguins_clean
```

```
## # A tibble: 333 x 7
##    species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g
##    <fct>   <fct>           <dbl>         <dbl>            <int>       <int>
##  1 Adelie  Torge~           39.1          18.7              181        3750
##  2 Adelie  Torge~           39.5          17.4              186        3800
##  3 Adelie  Torge~           40.3          18                195        3250
##  4 Adelie  Torge~           36.7          19.3              193        3450
##  5 Adelie  Torge~           39.3          20.6              190        3650
##  6 Adelie  Torge~           38.9          17.8              181        3625
##  7 Adelie  Torge~           39.2          19.6              195        4675
##  8 Adelie  Torge~           41.1          17.6              182        3200
##  9 Adelie  Torge~           38.6          21.2              191        3800
## 10 Adelie  Torge~           34.6          21.1              198        4400
## # ... with 323 more rows, and 1 more variable: sex <fct>
```

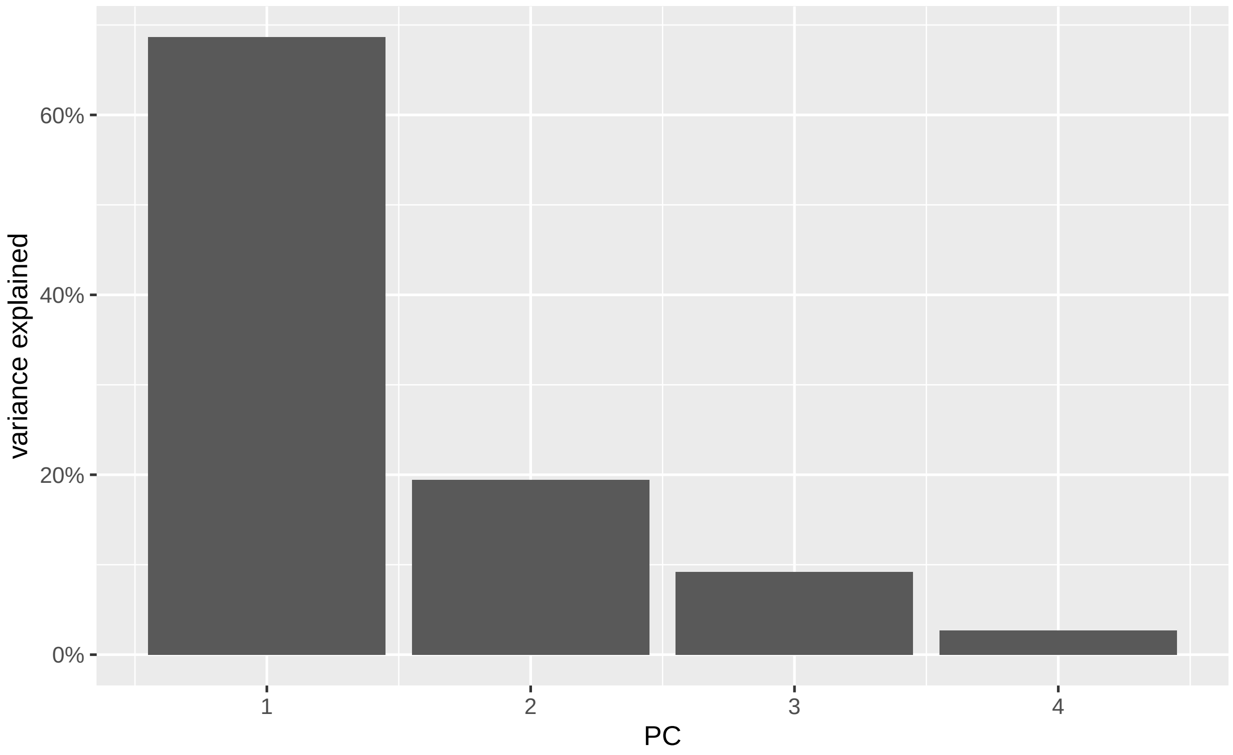
Problem 1: (2 pts)

Perform a PCA of the `penguins_clean` dataset and make two plots: 1. A rotation plot of components 1 and 2; 2. A plot of the eigenvalues, showing the amount of variance explained by the various components.

```r
library(broom) # need the broom library for augment(), tidy()

# perform PCA with prcomp(), store in pca_fit variable
pca_fit <- penguins_clean %>%
  select(where(is.numeric)) %>%
  scale() %>%
  prcomp()

# specify arrow style, using specifications from class
arrow_style <- arrow(
  angle = 20, length = grid::unit(8, "pt"),
  ends = "first", type = "closed"
)

# to fix later: arrows are too long, bottom-left text unreadable
# BOTTOM LEFT == bill_depth_mm
pca_fit %>% # create rotation plot here
  tidy(matrix = "rotation") %>%
  pivot_wider( # first set summary table parametres
    names_from = "PC", values_from = "value", names_prefix = "PC"
  ) %>%
  ggplot(aes(PC1, PC2)) +
  geom_segment(
    xend = 0, yend = 0, arrow = arrow_style
  ) +
  geom_text(aes(label = column), hjust = 1) # ultimately use ggplot.geom_text()
```

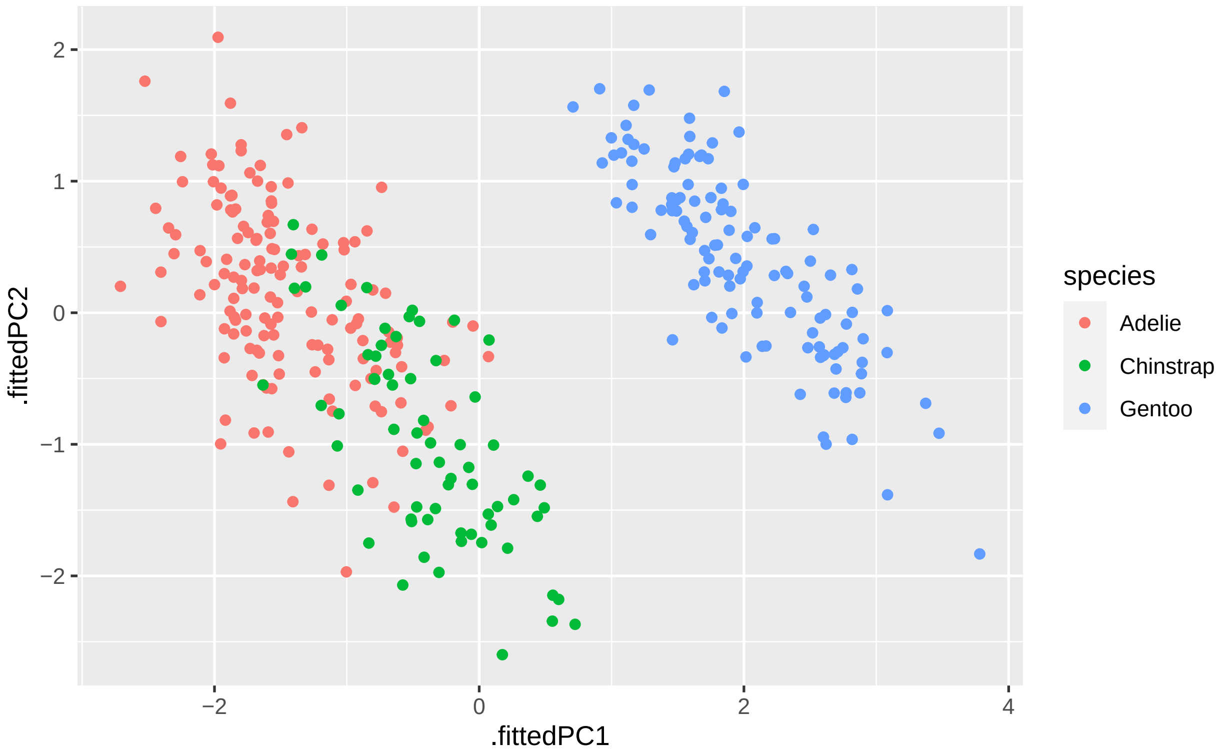


```
# now, create plot of Eigenvalues, showing explanation of variance
pca_fit %>%
  # obtain eigenvalues
  tidy(matrix = "eigenvalues") %>%
  ggplot(aes(PC, percent)) +
  geom_col() + # geom_col as usual to construct bar plot
  scale_x_continuous(breaks = 1:4) + # since we have 4 PC's as per rot.plot
  scale_y_continuous(
    name = "variance explained", label = scales::label_percent(accuracy = 1)
  )
```

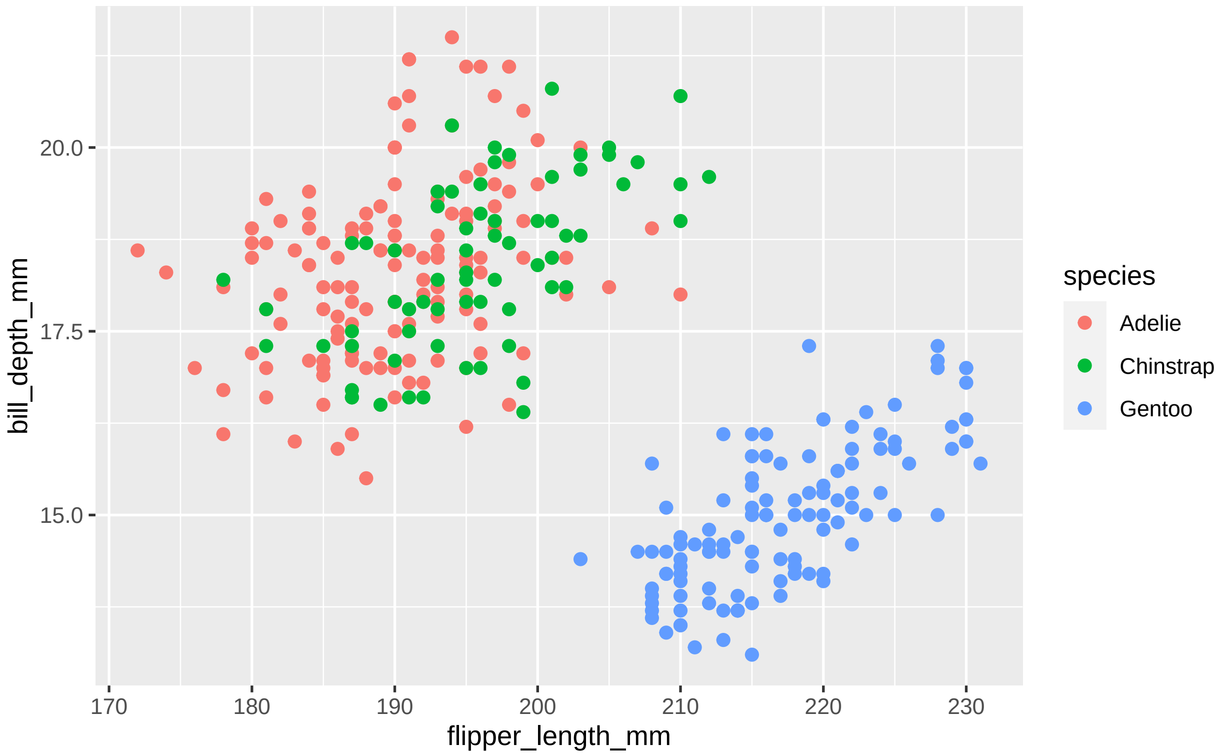


Problem 2: (4 pts) Make a scatter plot of PC 2 versus PC 1 and color by penguin species. Then use the rotation plot from Problem 1 to describe the physical characteristics by which the different penguin species differ. Finally, make one more scatter plot of the raw data that can support your interpretation of the PC analysis.

```r
pca_fit %>% # makes scatterplot of pca data
  augment(penguins_clean) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  geom_point(aes(color = species)) # split color by species, not sex
```



```
# scatterplot of raw data, possibly of bill lenmm, depmm, filpperlenmm, bodymassg
penguins_clean %>%
  ggplot(aes(x=flipper_length_mm, y = bill_depth_mm, color = species)) +
  geom_point(size = 2)
```



PC_Scatter and Rotation_Plot: PC1 more than likely represents measures of the overall size of the bird, given the nature of the dataset. We see that as flipper length, body mass, and bill length increase, we can generally expect an increased overall bird size. On the other hand, PC2 more than likely deals with the difference between flipper length and bill depth. We see that body mass only weakly predicts this difference, yet that increased bill length has a much stronger relationship with flipper length-bill depth difference. This is not too surprising, since both bill depth and bill length are measures of the same anatomical structure.

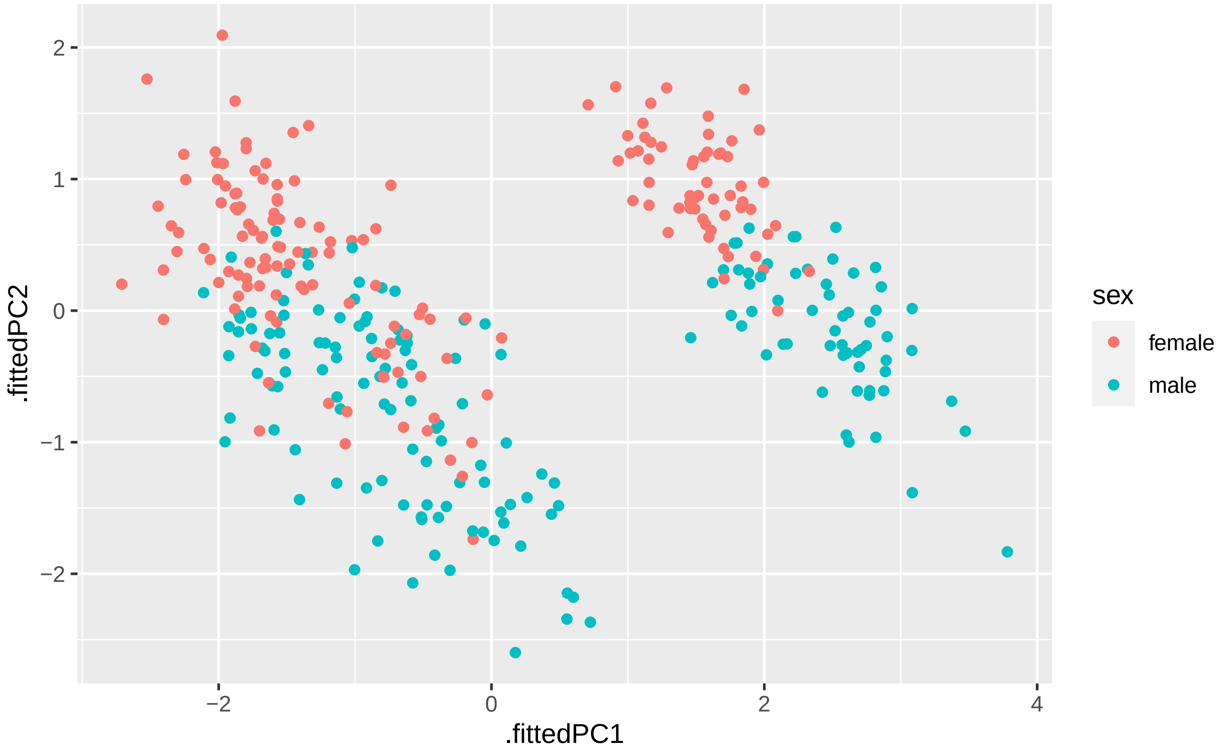
We can use the raw data to look further into these characteristics of the PCA. Looking into the raw flipper length versus bill depth scatterplot (whose units are conveniently both mm), we see that: (1) part of the reason for poor correlation is the difference in Gentoo penguins' measurements from those of Adelie and Chinstrap, with Gentoo having shallower bills and longer flippers in general. (2) Even within the same species, flipper length only relatively weakly correlates with bill depth. From the naked eye, this correlation is tightest for Gentoo penguins and by far the lowest for Adelie, where the plotted points are very far from linear and look almost arbitrary.

Finally, the graph of PC1 and PC2 colored by species corroborates these same findings, and also shows the negative PC1 and PC2 correlation for all 3 species which is, again, tightest for Gentoos. This negative correlation illustrates part of how variation in bill depth does not follow the same relation with body size that the other 3 given measures do when we did NOT control for species; Gentoos in general have significantly greater body masses and smaller bill depths than do Adelies or Chinstraps.

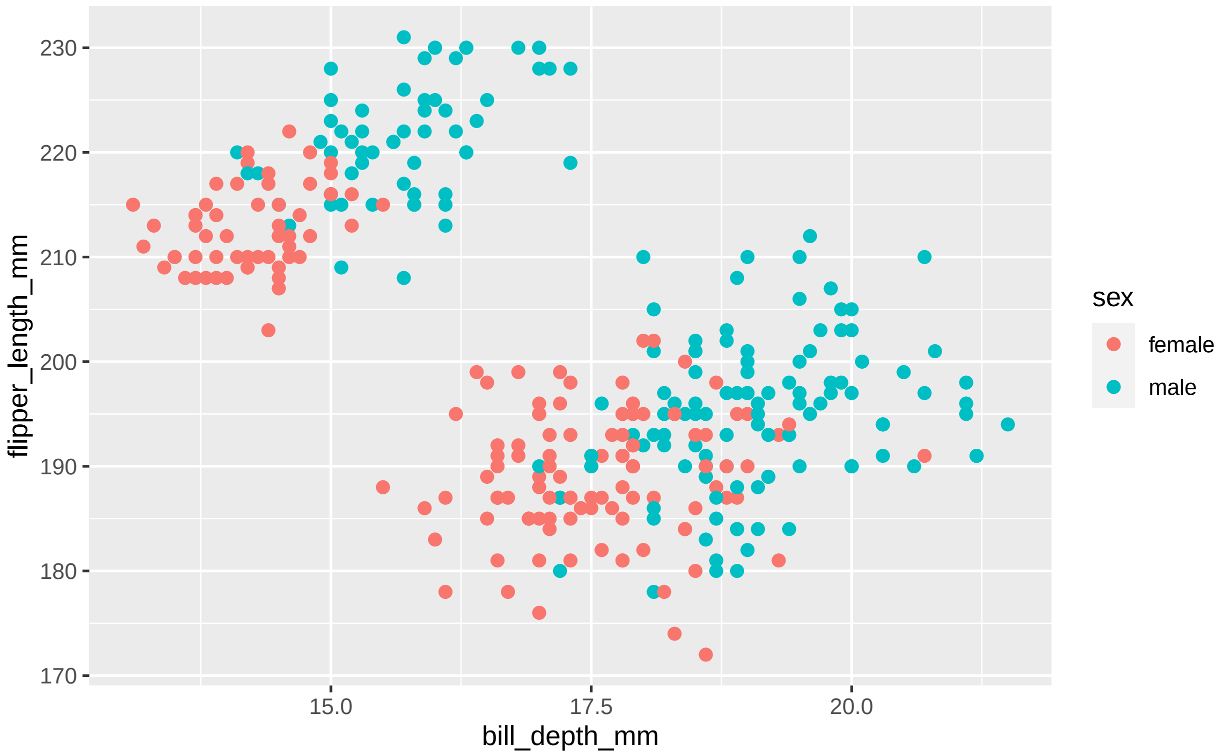
Problem 3: **(4 pts)** Again make a scatter plot of PC 2 versus PC 1, but now color by sex. Then use the rotation plot from Problem 1 to describe the physical characteristics by which the different penguin sexes differ. Finally, make one more scatter plot of the raw data that can support your interpretation of the PC analysis.

Hint: It helps to facet by penguin species.

```
pca_fit %>% # makes scatterplot of pca data
  augment(penguins_clean) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  geom_point(aes(color = sex)) # split color by sex this time, NOT species
```



```
penguins_clean %>% # same thing with the scatterplot from (2) # FIXME!
  ggplot(aes(x = bill_depth_mm, y = flipper_length_mm, color = sex)) +
  geom_point(size = 2)
```



From the PC1, PC2, sex scatterplot, we can see the same negative correlation as we did for species. However, we now have the added information of how sex informs body size and flipper length-bill depth difference. We see that generally speaking, females have lower body masses and *higher* flipper-bill depth differences than do males. This gives yet another reason why the rotation plot's arrow for PC2 goes left rather than right; females generally have a significantly greater difference than do males. The raw data shows that this is to a significant degree because females' flipper lengths are significantly shorter in general, yet their bill depths are similar to those of males.