Homework 7

Sookja Kang, sk26949

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.

```
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.

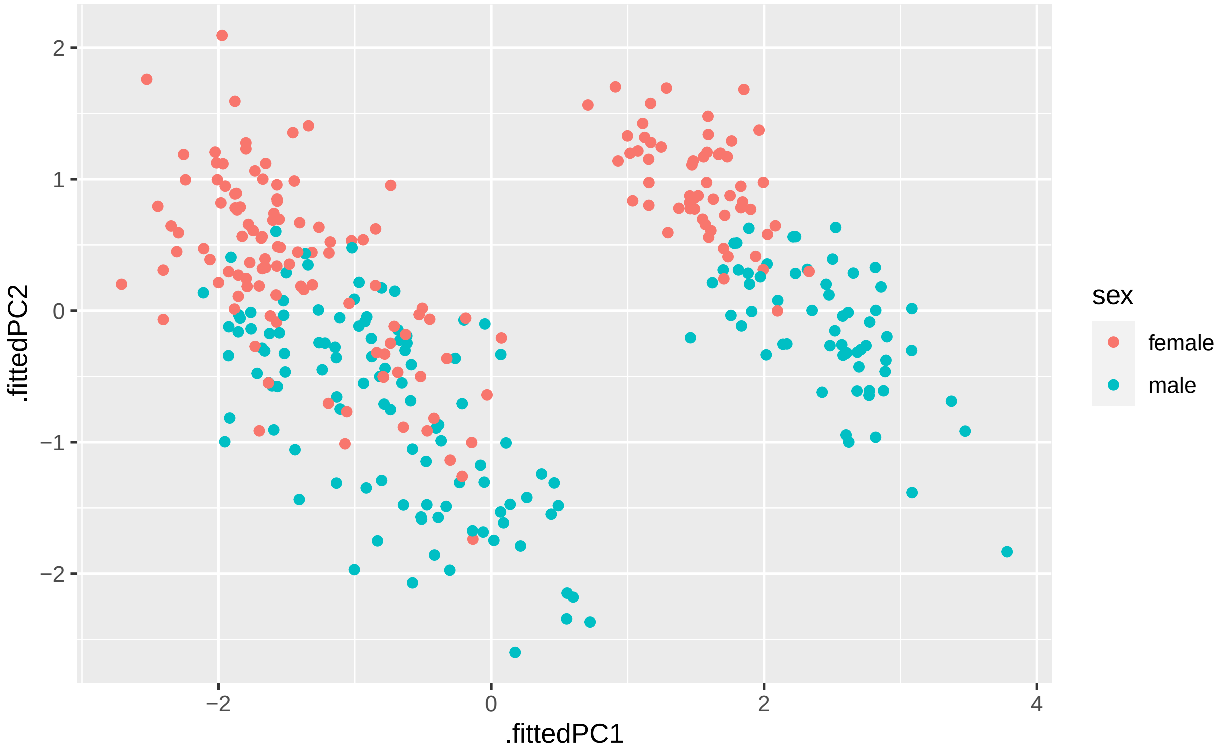
```
# your code goes here
#Plot1.
pca_fit <- penguins_clean %>%
  select(where(is.numeric)) %>%
  scale() %>%
  prcomp()

pca_fit
```

```
## Standard deviations (1, .., p=4):
## [1] 1.6569115 0.8821095 0.6071594 0.3284579
##
```

```
## Rotation (n x k) = (4 x 4):
##                          PC1         PC2        PC3        PC4
## bill_length_mm     0.4537532 -0.60019490 -0.6424951  0.1451695
## bill_depth_mm     -0.3990472 -0.79616951  0.4258004 -0.1599044
## flipper_length_mm  0.5768250 -0.00578817  0.2360952 -0.7819837
## body_mass_g        0.5496747 -0.07646366  0.5917374  0.5846861
```

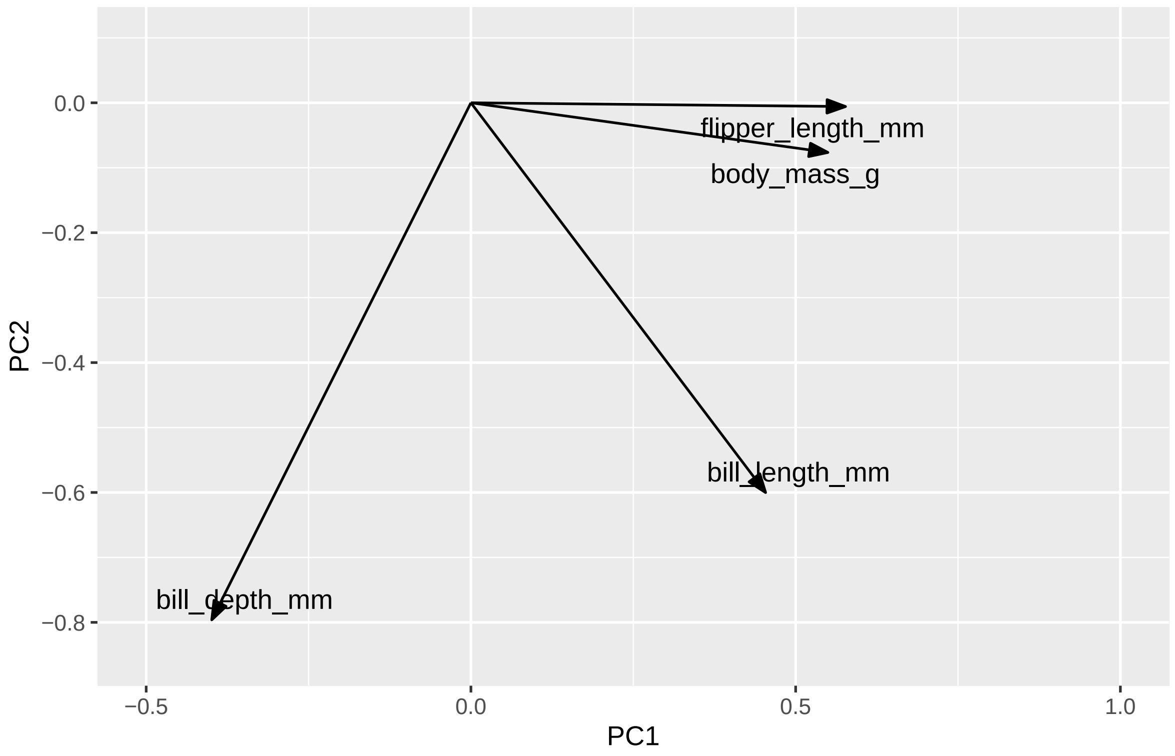
```
pca_fit %>%
  augment(penguins_clean) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  geom_point(aes(color = sex))
```



```
arrow_style <- arrow(
  angle = 20, length = grid::unit(8, "pt"),
  ends = "first", type = "closed"
)

pca_fit %>%
  tidy(matrix = "rotation") %>%  # extract rotation matrix
  pivot_wider(
    names_from = "PC", values_from = "value",
    names_prefix = "PC"
  ) %>%
  ggplot(aes(PC1, PC2)) +
  geom_segment(
    xend = 0, yend = 0,
    arrow = arrow_style
  ) +
  geom_text_repel(aes(label = column)) +
```

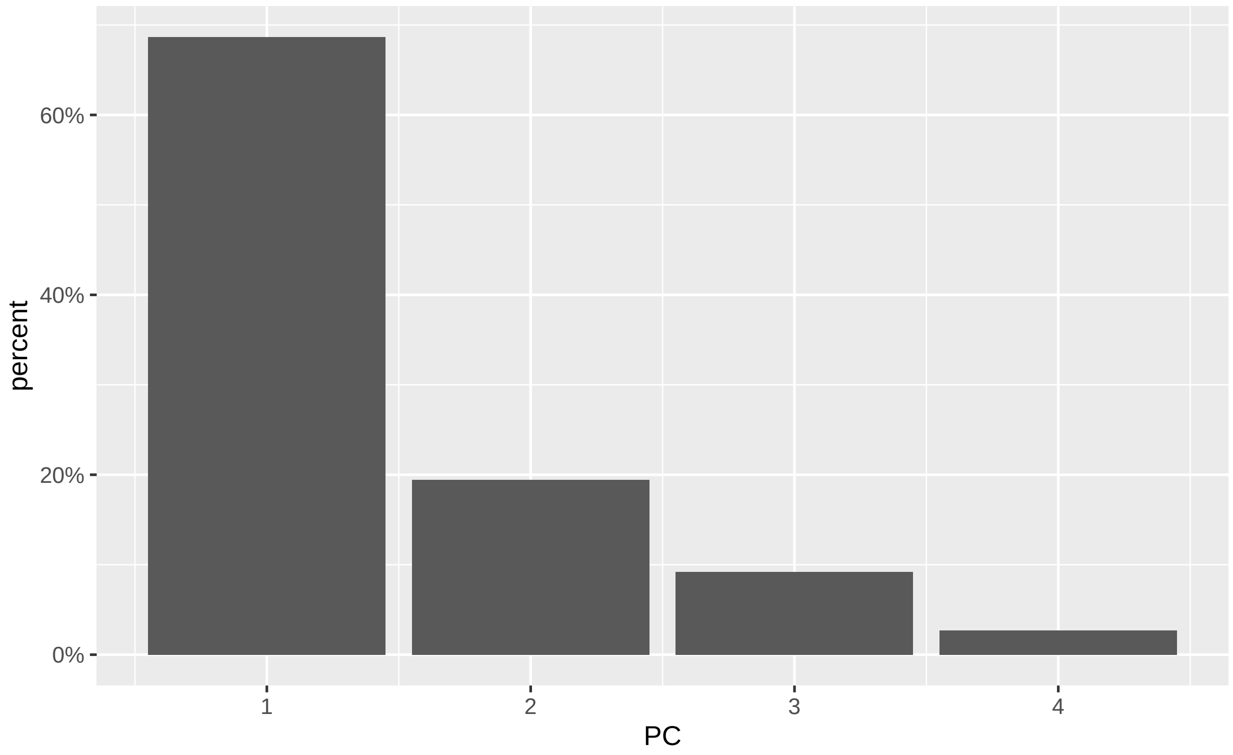
```r
  xlim(-0.50, 1) + ylim(-0.85, .1) +
  coord_fixed()
```



```r
#Plot2.
pca_fit %>%
  tidy(matrix = "eigenvalues")
```

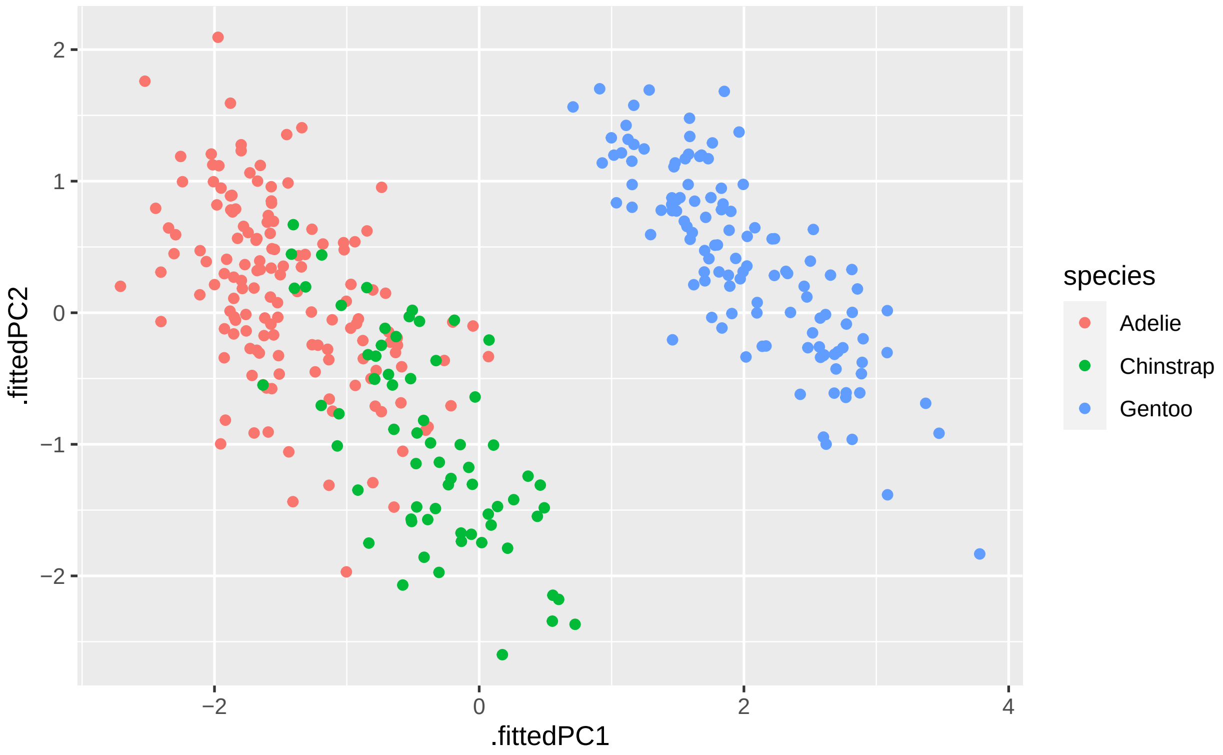
```
## # A tibble: 4 x 4
##      PC std.dev percent cumulative
##   <dbl>   <dbl>   <dbl>      <dbl>
## 1     1   1.66   0.686       0.686
## 2     2   0.882  0.195       0.881
## 3     3   0.607  0.0922      0.973
## 4     4   0.328  0.0270      1
```

```r
pca_fit %>%
  tidy(matrix = "eigenvalues") %>%
  ggplot(aes(PC, percent)) +
  geom_col() +
  scale_x_continuous(breaks = 1:4) +
  scale_y_continuous(labels = scales::label_percent())
```

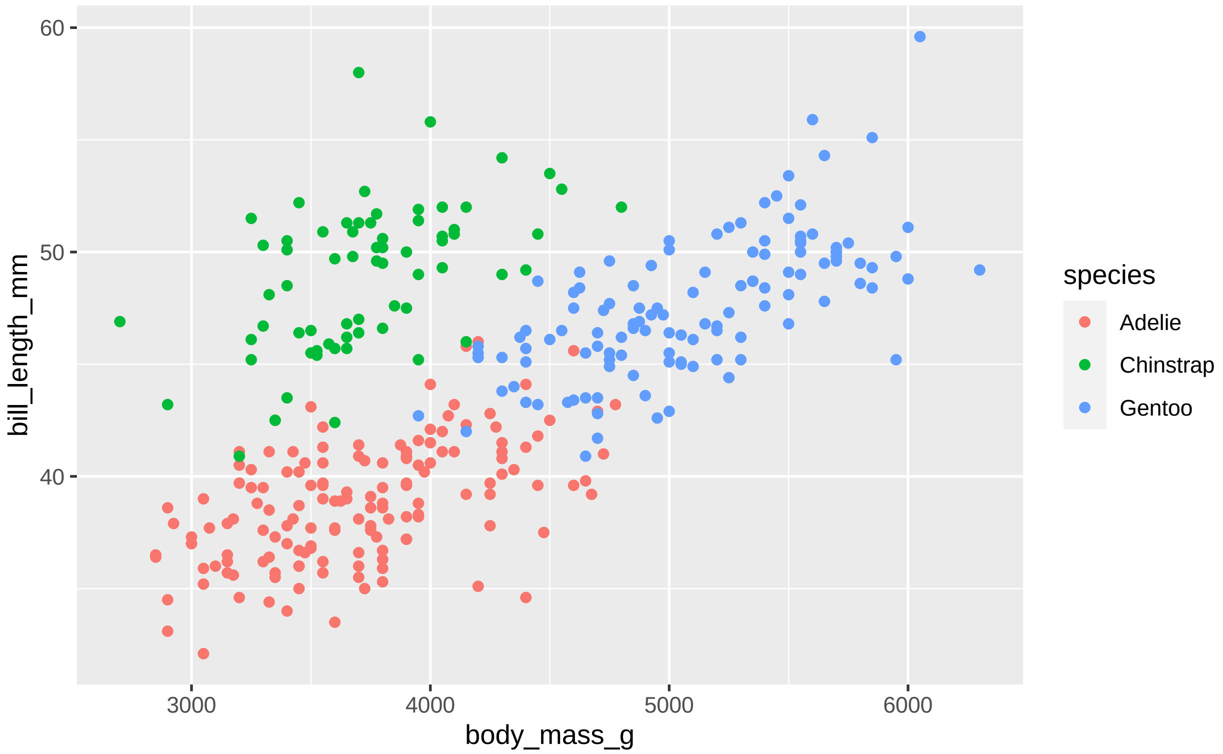


Problem 2: (4 pts) Make a scatter plot of PC 2 versus PC 1 and color by penguin species. Then use the rotation plot from Problem 1 to describe the physical characteristics by which the different penguin species differ. Finally, make one more scatter plot of the raw data that can support your interpretation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

```r
# your code goes here
pca_fit %>%
  augment(penguins_clean) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  geom_point(aes(color = species))
```



```
#Plot2.
ggplot(penguins_clean, aes(body_mass_g, bill_length_mm, color = species)) +
geom_point()
```

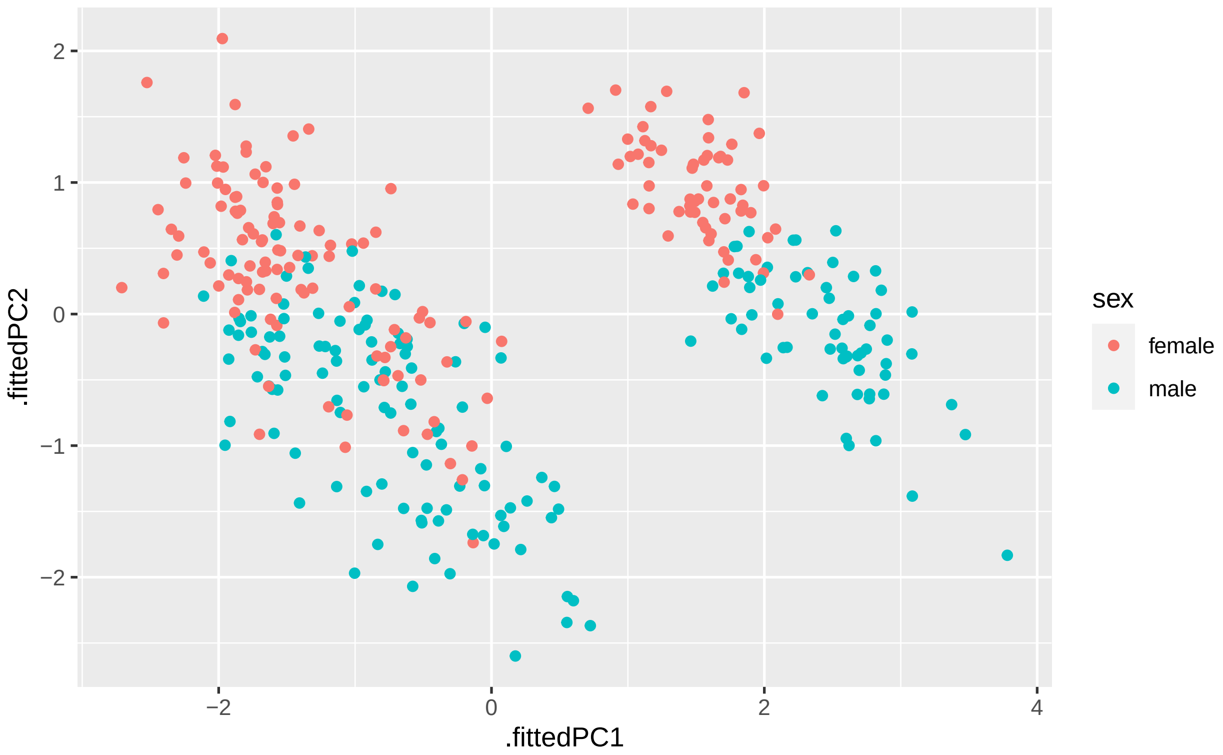


The rotation matrix shows that flipper_length, body_mass, and bill_length contribute positively to PC1 while bill_depth contributes negatively to PC1. Both bill_depth and bill_length contributes negatively to PC2. Considering this information, PC1 represents overall size of penguins while PC2 represents the size of beak. From the rotation matrix and the scatter plot for PC2 against PC1, this shows penguins with high body mass have relatively long bill lengths. Also, the scatter plot (PC2 against PC1) shows that Gentoo has relatively bigger body size compared to Chinstrap and Adelie. Chinstrap has bigger body size than Adelie. In terms of bill length, Adelie has shorter bill length than Chinstrap and Gentoo. This also found in the scatter plot using the raw data of bill length against body mass. In terms of body mass, this plot shows: Adelie < Chinstrap < Gentoo. Also, this plot shows that penguins with large body mass relatively have longer bill lengths.

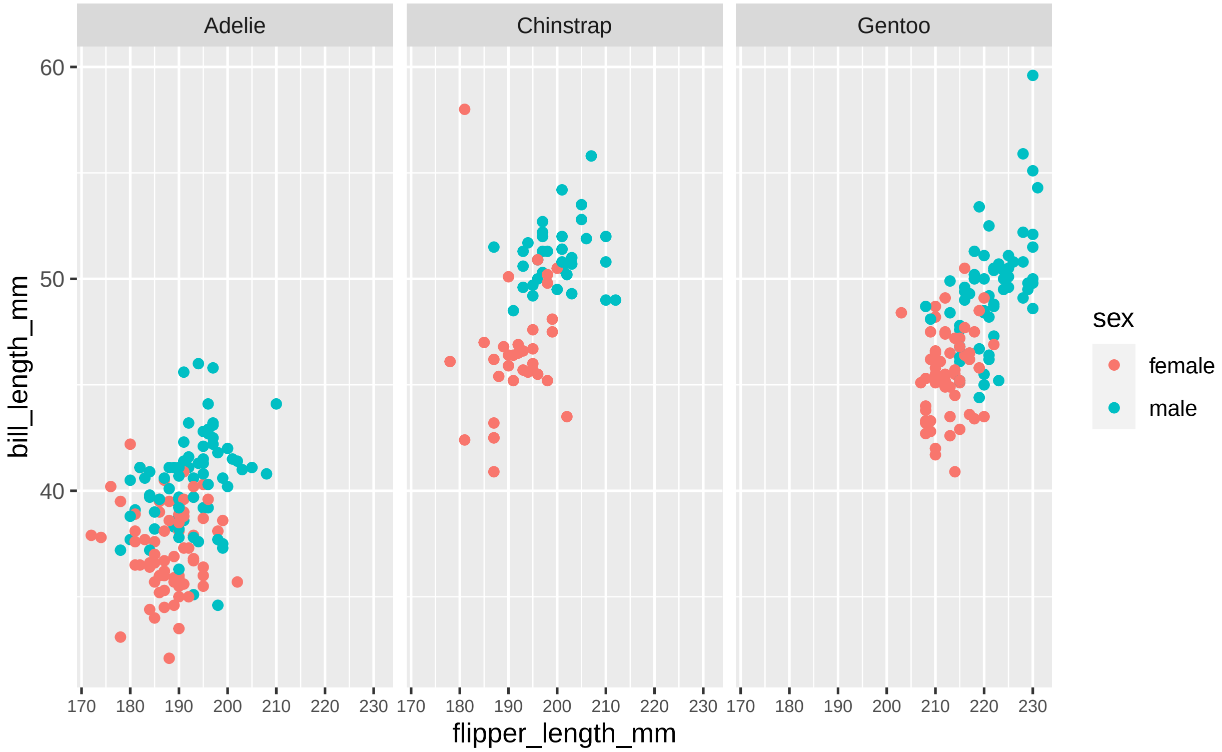
Problem 3: (4 pts) Again make a scatter plot of PC 2 versus PC 1, but now color by sex. Then use the rotation plot from Problem 1 to describe the physical characteristics by which the different penguin sexes differ. Finally, make one more scatter plot of the raw data that can support your interpretation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

Hint: It helps to facet by penguin species.

```
# your code goes here
#Plot1.
pca_fit %>%
  augment(penguins_clean) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  geom_point(aes(color = sex))
```



```
#Plot2.
ggplot(penguins_clean, aes(flipper_length_mm, bill_length_mm, color = sex)) +
  geom_point() +
  theme(axis.text.x = element_text(size = 7)) +
  facet_wrap(~species)
```



The rotation matrix shows that the flipper_length contributes on PC1; the bill length contributes on both PC1 and PC2. The scatter plot of PC 2 against PC 1 by sex shows male penguins have bigger body size, flipper length, and bill length than female penguins when we look at the different scatter group. The scatter plot using the raw data for bill length against flipper length by sex shows male penguins are bigger than female penguins for flipper length and bill length. In terms of species, Adelie has the shortest lengths for both flipper and bill. Chinstrap and Gentoo have similar range for bill length but Gentoo has longer flipper than Chinstrap.