HW 5

SDS348 Spring 2021

Rose Hedderman EID: rrh2298

This homework is due on Mar 12, 2021 at 8am. Submit a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

Question 1: (6 pts)

The dataset for this homework comes from the article:

Tsuzuku N, Kohno N. 2020. The oldest record of the Steller sea lion Eumetopias jubatus (Schreber, 1776) from the early Pleistocene of the North Pacific. https://doi.org/10.7717/peerj.9709 (https://doi.org/10.7717/peerj.9709)

Under the supplemental information, the data was retrieved from a word document into an excel document.

1.1 (4 pts) Read the ***Abstract*** of the article and the section called *Results of Morphometric Analyses*. What was the goal of this study and what was the main finding?

The goal of this study was to determine the genetic difference between the Steller sea lion and the earliest fossil to be identitfied in the same genus. The main finding of the article is that the after bivariate nalyses and PCA based on almost 40 measurements, there was almost no difference between the species.

1.2 (2 pts) Import the dataset from Excel. How many rows and how many columns are in this dataset? What does a row represent? What does a column represent?

```
library(readxl)
sealions <- read_excel("C:\\Users\\roseh\\OneDrive\\Desktop\\Spring 2021\\SDS 348\\HW5.xlsx")
#View(HW5)
```

There are 53 rows and 40 columns in the sealions dataset. A row represents the species of sea lion being studied and a column represents the identifier of each subject sea lion of a given species.

Question 2: (7 pts)

Before we can analyze the data, let's clean it.

2.1 (1 pt) When importing this dataset into R Studio, which variables were considered numeric? Why are some measurements not considered as numeric?

```
library(tidyverse)
```

No variables were considered numeric.

2.2 (2 pts) Using `dplyr` functions, replace all "-" in the dataset by missing values *NA* then make sure all measurements are defined as numeric variables. What is the mean rostral tip of mandible C?

```
sealion_clean <- sealions %>%
  na_if("-") %>%
  mutate_at(-c(1),as.numeric)
sealions <- sealion_clean %>%
  mutate_at(-c(1),as.numeric)
sealions
```

```
## # A tibble: 51 x 39
##    ID        A     B     C     D     E     F     G     H     I     J     K     L
##    <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 E. j~  262   232   62.4  31.1  63.1  59.0  44.0  46.8  62.6  62.6  57.8  87.1
##  2 E. j~  285   242   64.5  31.7  70.5  75.6  44.3  62.5  63.1  63.5  64.6  97.5
##  3 E. j~  266.  242.  53.1  30.2  70.5  60.3  48.0  50.8  62.0  63.9  63.6  99.2
##  4 E. j~  244   212   44.9  26.0  55.9  52.0  38.5  39.9  51.8  55.9  45.2  85.0
##  5 E. j~  237   209.  39.4  26.1  51.2  49.4  37.2  37.9  45.6  49.0  41.4  83.4
##  6 E. j~  228   201.  39.5  25.4  51.2  48.1  36.4  37.2  63.0  49.7  43.6  76
##  7 E. j~  227   202.  48.4  24.8  48.5  49.2  39.0  39.1  48.6  52.6  43.6  81.2
##  8 E. j~  226   190.  55.2  27.2  49.0  34.0  30.5  29.4  50.3  50.1  41.8  75.3
##  9 E. j~  282.  257.  49.6  31.4  72.7  45.2  40.1  49.1  63.8  66.3  68.0 104.
## 10 E. j~  237   215   50.5  16.2  50.4  47.0  38.6  37.6  50.2  54.1  44.0  80.8
## # ... with 41 more rows, and 26 more variables: M <dbl>, N <dbl>, O <dbl>,
## #   P <dbl>, Q <dbl>, R <dbl>, S <dbl>, T <dbl>, U <dbl>, V <dbl>, W <dbl>,
## #   X <dbl>, Y <dbl>, Z <dbl>, AA <dbl>, AB <dbl>, AC <dbl>, AD <dbl>,
## #   AE <dbl>, AF <dbl>, AG <dbl>, AH <dbl>, AI <dbl>, AJ <dbl>, AK <dbl>,
## #   AL <dbl>
```

```
sealions %>%
  summarize(mean(C, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `mean(C, na.rm = T)`
##                  <dbl>
## 1                 34.9
```

The mean rostral tip of mandible C is 34.866.

2.3 (2 pts) Using `dplyr` functions, only keep numeric variables that are not missing for the fossil specimen GKZ-N 00001 (hint: you can use `select_if()` on the condition that `HW5_clean[51,]` has *no* missing value with `is.na()`). Then remove the rest of the missing values. How many columns and how many rows are remaining in this dataset?

```
# sealions 2 contains the sealions dataset and only includes values that are numeric and no NA values
sealions2 <- sealions %>%
  select_if(!is.na(sealion_clean[51,])) %>%
  drop_na()
sealions2
```

```
## # A tibble: 42 x 23
##    ID        C     D     I     J     K     L     M     X     Y     Z    AA    AB
##    <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 E. j~  62.4  31.1  62.6  62.6  57.8  87.1  24.3  65.6  48.0  79.1  55.5  8.89
##  2 E. j~  64.5  31.7  63.1  63.5  64.6  97.5  14.7  73.1  51.6  68.2  60.4 10.9
##  3 E. j~  53.1  30.2  62.0  63.9  63.6  99.2  18.4  72.9  46.8  84.6  63.7 11.0
##  4 E. j~  44.9  26.0  51.8  55.9  45.2  85.0  19.8  63.5  41.3  70.2  42.4  7.04
##  5 E. j~  39.4  26.1  45.6  49.0  41.4  83.4  24.1  56.1  38.8  63.4  44.0  7
##  6 E. j~  39.5  25.4  63.0  49.7  43.6  76    17.2  54.2  35.8  55.3  37.1  7.48
##  7 E. j~  48.4  24.8  48.6  52.6  43.6  81.2  18.9  57.5  35.7  57.8  40.5  6.87
##  8 E. j~  55.2  27.2  50.3  50.1  41.8  75.3  14.6  54.2  32.7  45.7  38.7  8.54
##  9 E. j~  49.6  31.4  63.8  66.3  68.0 104.   17.7  76.4  49.7  95.3  50.6 11.8
## 10 E. j~  50.5  16.2  50.2  54.1  44.0  80.8  20.0  58.5  38.8  59.0  40.3  8.16
## # ... with 32 more rows, and 10 more variables: AC <dbl>, AD <dbl>, AE <dbl>,
## #   AF <dbl>, AG <dbl>, AH <dbl>, AI <dbl>, AJ <dbl>, AK <dbl>, AL <dbl>
```

There are 42 rows and 23 columns in the new dataset.

2.4 (2 pts) Using `dplyr` functions, only keep numeric variables and scale (also called standardize) each numeric variable. What should the mean of the scaled variable of the rostral tip of mandible C be?

```
# This removes all variables (such as ID) that are not numeric and then scales the values
sealions3 <- sealions2 %>%
  select_if(is.numeric) %>%
  mutate_if(is.numeric,scale)
sealions3
```

```
## # A tibble: 42 x 22
##    C[,1]  D[,1] I[,1] J[,1] K[,1] L[,1]  M[,1] X[,1] Y[,1]  Z[,1] AA[,1]  AB[,1]
##    <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl>   <dbl>
##  1 1.91   1.33  1.53  1.32  1.26  0.840  0.876 1.11  1.55   1.49   1.57   0.542
##  2 2.06   1.40  1.56  1.37  1.69  1.30  -1.30  1.52  1.84   0.857  1.95   1.25
##  3 1.26   1.22  1.50  1.39  1.63  1.38  -0.472 1.51  1.46   1.81   2.20   1.30
##  4 0.694  0.742 0.891 0.928 0.455 0.748 -0.147 0.994 1.03   0.974  0.549 -0.108
##  5 0.312  0.746 0.526 0.526 0.213 0.675  0.828 0.586 0.827  0.579  0.680 -0.122
##  6 0.321  0.665 1.55  0.564 0.353 0.343 -0.734 0.485 0.585  0.105  0.142  0.0465
##  7 0.937  0.602 0.704 0.734 0.355 0.576 -0.341 0.661 0.581  0.254  0.406 -0.168
##  8 1.41   0.881 0.807 0.589 0.241 0.313 -1.32  0.482 0.344 -0.447  0.265  0.419
##  9 1.02   1.36  1.61  1.53  1.90  1.61  -0.630 1.70  1.69   2.43   1.19   1.57
## 10 1.09  -0.415 0.798 0.820 0.381 0.558 -0.108 0.718 0.829  0.325  0.389  0.285
## # ... with 32 more rows, and 10 more variables: AC[,1] <dbl>, AD[,1] <dbl>,
## #   AE[,1] <dbl>, AF[,1] <dbl>, AG[,1] <dbl>, AH[,1] <dbl>, AI[,1] <dbl>,
## #   AJ[,1] <dbl>, AK[,1] <dbl>, AL[,1] <dbl>
```

```
# this finds the mean of the scaled variable of the rostral tip of mandible C using dplyr functions
sealions3 %>%
  summarize(mean(C, na.rm = T))
```

```
## # A tibble: 1 x 1
##   `mean(C, na.rm = T)`
##                  <dbl>
## 1 0.000000000000000149
```

The mean of the scaled variable of the rostral tip of mandible C is now 1.487009e-16.

Question 3: (6 pts)

Let's now perform PCA on the measurements available for the fossil specimen GKZ-N 00001.

3.1 (2 pts) Using the function `prcomp()`, calculate the principal components (PCs) for the dataset obtained in Question 2.4. Find the percentage of explained variance for each PC. What is the cumulative percentage of explained variance for PC1 and PC2?

```
sealion_pca <- sealions3 %>%
  prcomp
sealion_pca
```

```
## Standard deviations (1, .., p=22):
##  [1] 4.31832285 1.01733661 0.76853474 0.67401193 0.60753548 0.44301161
##  [7] 0.39383982 0.37325326 0.33111979 0.28423200 0.20952585 0.20712107
## [13] 0.18760632 0.14793722 0.14002759 0.11973326 0.11555864 0.10014573
## [19] 0.09387754 0.08251348 0.05608239 0.04456930
##
## Rotation (n x k) = (22 x 22):
##            PC1         PC2          PC3          PC4         PC5          PC6
## C  -0.21463902  0.17716482 -0.068648839 -0.104309121  0.34037552 -0.277165766
## D  -0.21783961  0.04150373 -0.326127619 -0.179509304 -0.06812348  0.002952071
## I  -0.21809115  0.18788014 -0.112780104 -0.121624534  0.11559285  0.057531864
## J  -0.22693934  0.09209169 -0.144692837 -0.098151145  0.09222458 -0.008471288
##             PC7         PC8         PC9         PC10         PC11         PC12
## C   0.386892436 -0.15254562  0.11363486 -0.242091937  0.244255542 -0.237377343
## D   0.118426330  0.20202944 -0.12287517 -0.171646780 -0.560805715 -0.102614757
## I  -0.149419814  0.22977537  0.52833044  0.162937097  0.160480122 -0.332291560
## J  -0.004421999 -0.01506705  0.03970697  0.235791971  0.052245397  0.079502913
##           PC13         PC14         PC15         PC16         PC17        PC18
## C   0.15014894 -0.128630323  0.116502416  0.483923490 -0.032992192 -0.11576725
## D   0.21261661 -0.483057516  0.049730777 -0.220589907 -0.133260497 -0.01007973
## I  -0.37676851 -0.002281316 -0.048522963 -0.318358286  0.065293496 -0.17406962
## J   0.27930999  0.027705817  0.009480918 -0.008472822  0.193201777  0.26937044
##            PC19         PC20         PC21         PC22
## C   0.033583099  0.176710250  0.063694453  0.126649664
## D   0.167489643 -0.068578241 -0.001986879  0.053417649
## I   0.170828256  0.070883296 -0.139137958  0.148935998
## J   0.028734438  0.263503963 -0.443080792 -0.615620161
##  [ reached getOption("max.print") -- omitted 18 rows ]
```

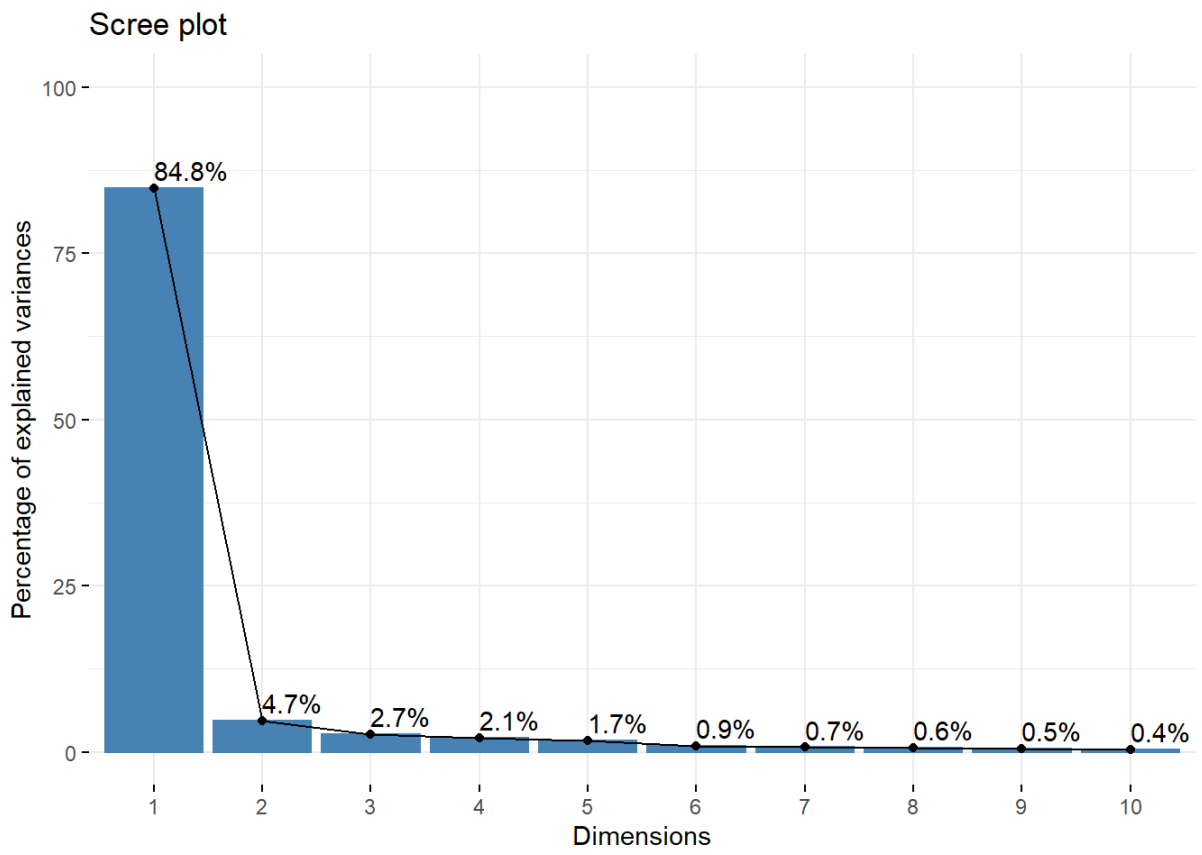
```
#perc_sealionpca <- sealion_pca %>%
#  select(c("PC1", "PC2"))
percent <- 100* (sealion_pca$sdev^2 / sum(sealion_pca$sdev^2))
percent
```

```
##  [1] 84.763237599  4.704426229  2.684752919  2.064964013  1.677724352
##  [6]  0.892087650  0.705044571  0.633263627  0.498365075  0.367217403
## [11]  0.199550372  0.194996080  0.159982417  0.099479186  0.089126030
## [16]  0.065163874  0.060699091  0.045587127  0.040059054  0.030947613
## [21]  0.014296522  0.009029193
```

The cumulative percentage of explained variance for PC1 is 84.918%, PC2 is 4.696%, PC3 is 2.71%, PC4 is 2.03%, PC5 is 1.59%, PC6 is 0.84%, PC7 is 0.685%, PC8 is 0.645%, PC9 is 0.489%, PC10 is 0.360%, PC11 is 0.199%, PC12 is 0.196%, PC13 is 0.158%, PC14 is 0.10%, PC15 is 0.089%, PC16 is 0.065%, PC17 is 0.0626% , PC18 is 0.0456%, PC19 is 0.040%, PC20 is 0.031%, PC21 is 0.0145%, and PC22 is 0.009%.

3.2 (1 pt) Construct a scree plot using the package `factoextra` with the function `fviz_screeplot` and determine how many principal components should be considered.

```
library(factoextra)
fviz_screeplot(sealion_pca, addlabels = TRUE, ylim = c(0, 100))
```



From the plot, only one of the principal components should be considered.

3.3 (2 pts) Consider the matrix, x, of new data provided by the PCA, save it as a data frame, and add the ID variable from the dataset created in Question 2.3. Next, use the `ID` variable to create two variables `species` and `sex` by using the function `separate()` (hint: in the `ID` variable, what symbol separates the species from sex?). Finally, the article states that the fossil specimen has to be male. Replace the missing value of sex for the fossil specimen GKZ-N 00001 (hint: use the functions `mutate()` and `replace_na()`).

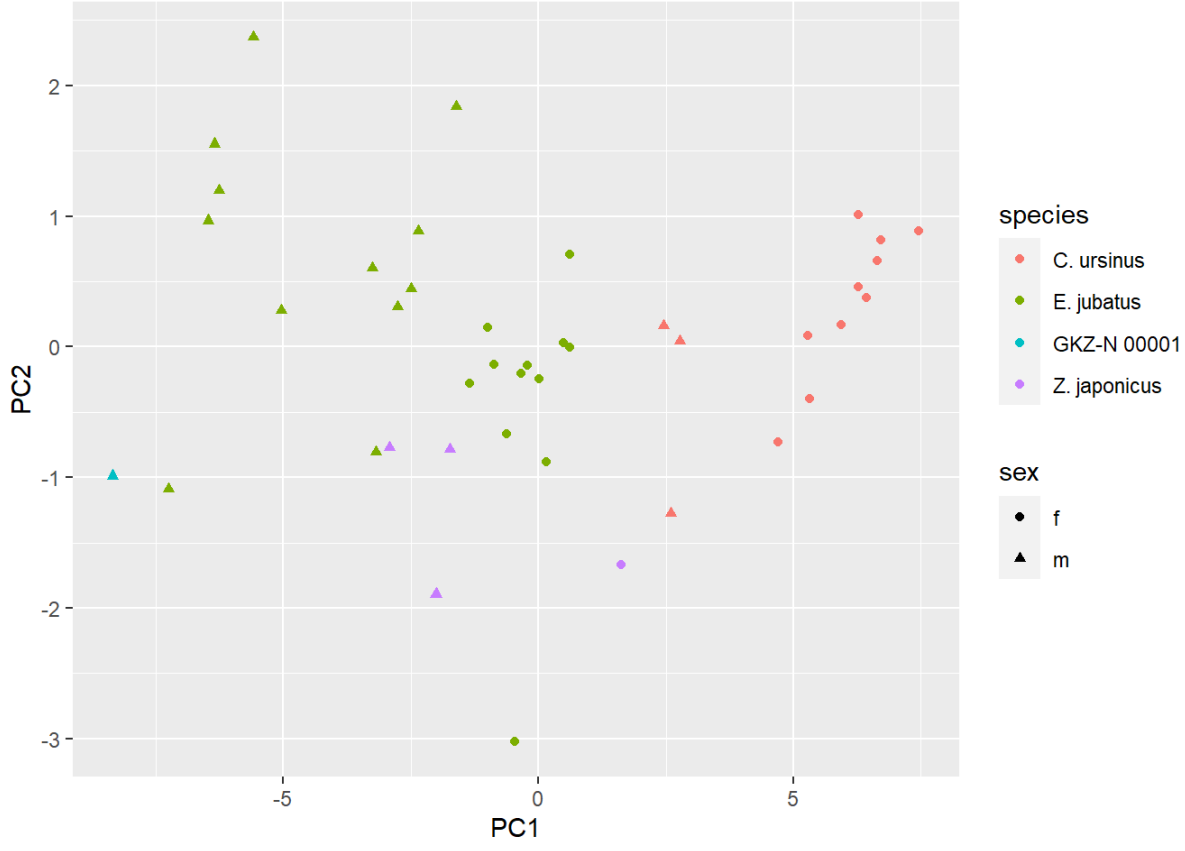
```
x <- data.frame(sealion_pca$x, ID = sealions2$ID) %>%
  separate(ID, into = c("species", "sex1"), sep = "([[])") %>%
  separate(sex1, into = c("sex", "other"), sep = "(])") %>%
  select(-other) %>%
  replace_na(list(sex = "m"))

x
```

```
##         PC1       PC2        PC3        PC4       PC5           PC6        PC7
## 1 -5.032235 0.2779438 -1.1718231  0.3775059 0.8550514 -0.1673534096 -0.2754529
## 2 -5.582133 2.3699275 -0.2417352  0.3246010 0.6925580 -0.4404413270  0.6291360
## 3 -6.255439 1.1956981 -0.2735351  0.3009795 0.1530323 -0.0002656182 -0.6928507
## 4 -3.246004 0.6052780  0.2322060 -0.2175704 0.7076523  0.4362657824 -0.4825607
##           PC8         PC9        PC10        PC11       PC12       PC13
## 1  0.04959747  0.30707650 -0.48388236  0.26033761 -0.5219201  0.2158761
## 2  0.11546493 -0.07666667 -0.39146522  0.09906975  0.1714887 -0.1309448
## 3  0.03508570 -0.78652761 -0.12059685  0.17038023 -0.1263589 -0.3673894
## 4 -0.31434608  0.58035787 -0.05914516 -0.37188186  0.5434417  0.2387662
##          PC14        PC15        PC16        PC17        PC18          PC19
## 1 -0.09929164 -0.09811109  0.27979899 -0.02437626 0.028946270  0.0635297777
## 2  0.05502252 -0.32657352 -0.21953657 -0.23380003 0.197491907 -0.0001832198
## 3  0.21730143  0.07854900  0.02090820  0.11311273 0.002486631  0.1162798417
## 4 -0.11015870  0.03071431 -0.01629767 -0.08599059 0.011268507 -0.0030663753
##          PC20        PC21       PC22     species sex
## 1  0.03782925 -0.05068427 0.03453319 E. jubatus    m
## 2 -0.10155040  0.07488273 0.02520626 E. jubatus    m
## 3 -0.05293213  0.02071424 0.01388521 E. jubatus    m
## 4 -0.02988570 -0.01231091 0.00235269 E. jubatus    m
##  [ reached 'max' / getOption("max.print") -- omitted 38 rows ]
```

3.4 (1 pt) Using `ggplot` and the dataset created in the previous question, represent the observations along the new variables PC1 and PC2. In the aesthetics, color the observations by their species and shape the observations by their sex. The fossil specimen GKZ-N 00001 appears to be close to which species?

```
ggplot(x, aes(x = PC1, y = PC2, color = species, shape = sex)) +
  geom_point()
```



The fossil specimen GKZ-N 00001 appears to be close to the E. jubatus species.

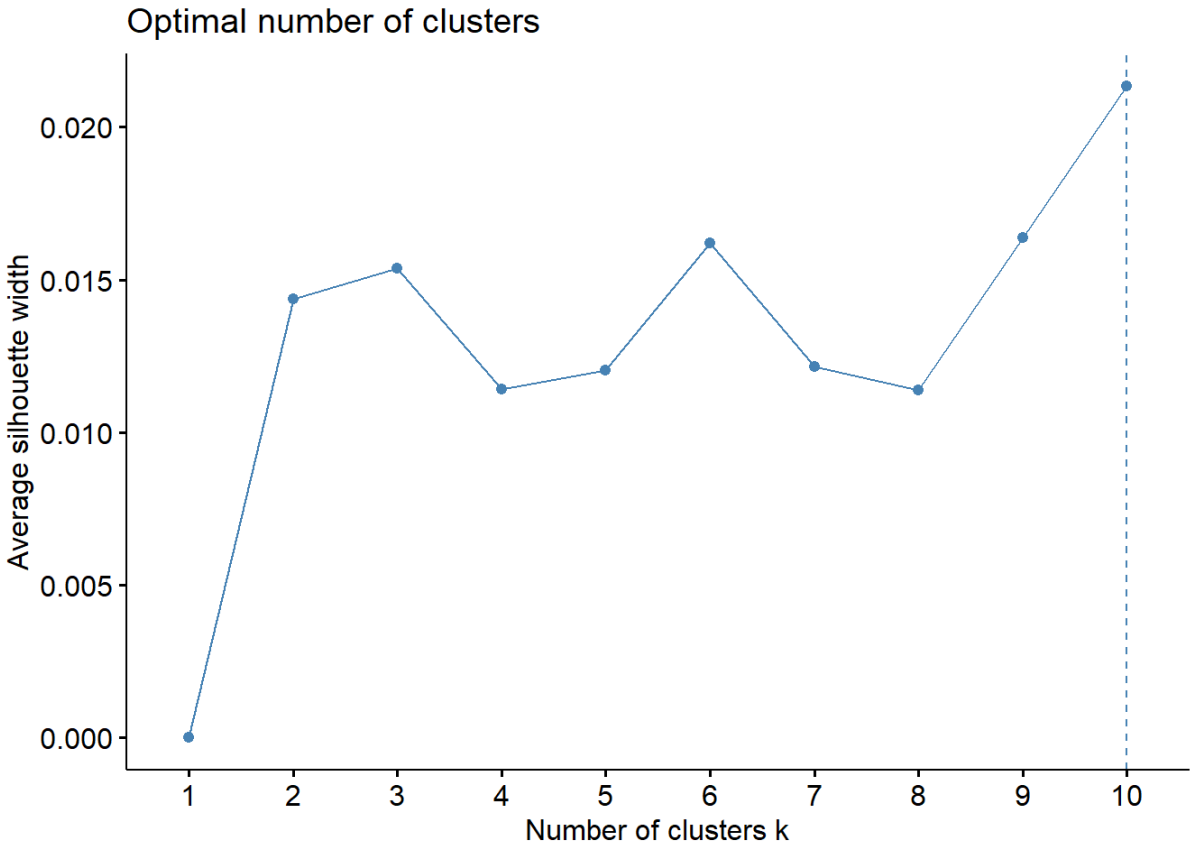
Question 4: (6 pts)

Let's now perform the partition around medoids (PAM) algorithm on the new variables to identify clusters of sea lions and determine which cluster the fossil specimen GKZ-N 00001 is likely to belong to.

4.1 (2 pts) Using the function `pam()` from the library `cluster`, perform the PAM algorithm on the dataset obtained in Question 3.3. Make sure to only select the variables `PC1` and `PC2`. Add the identification of the cluster number to the dataset from Question 3.3 (hint: use one of the elements created through the PAM algorithm). How many clusters are we looking for if the goal is to (hopefully) recover the different species?

```
 library(cluster)
sealion_num<- x %>%
  select_if(is.numeric) %>%
  scale


# Choose the number of clusters
fviz_nbclust(sealion_num, FUNcluster = pam, method = "s")
```

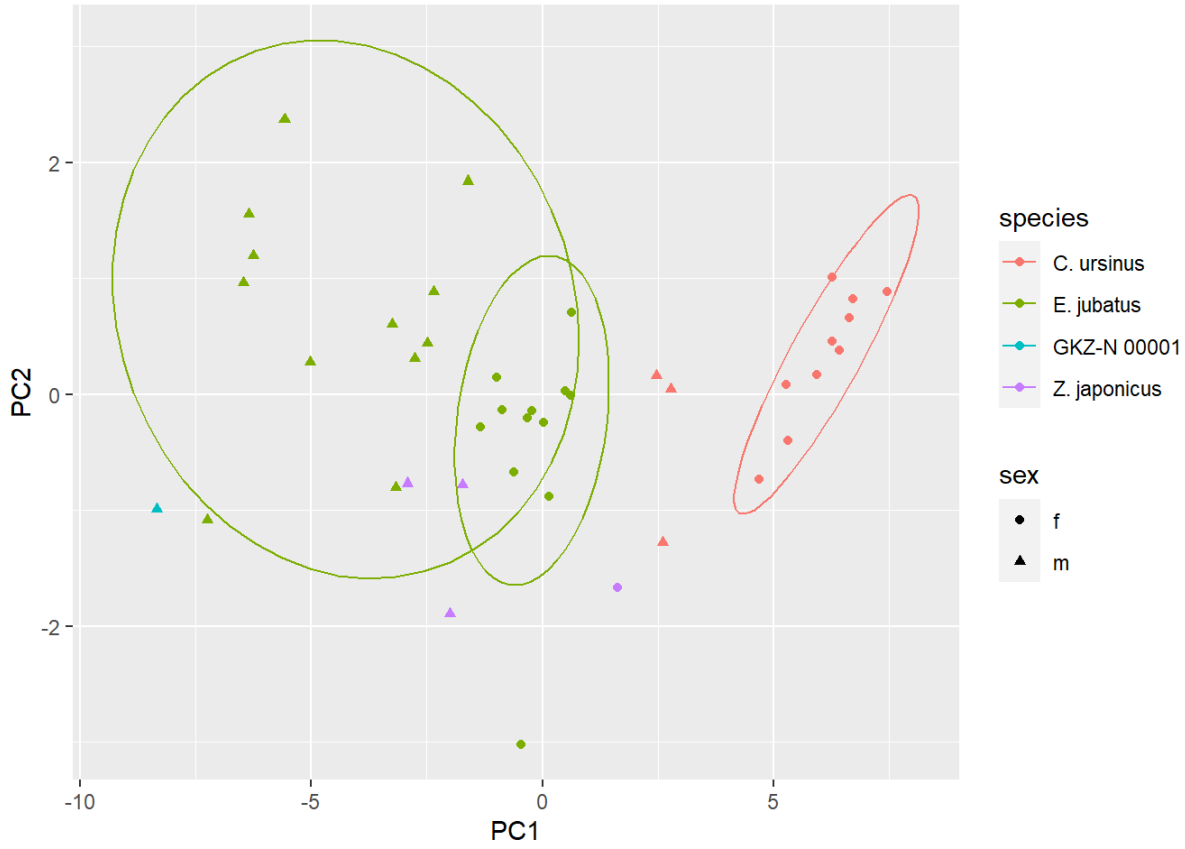



```
# Perform PAM on the data
sealion_pam <- sealion_num %>%
  pam(k=10)
```

From the PAM algorithm, 10 clusters appears to be the correct number to recover the species.

4.2 (2 pts) Using `ggplot` and the dataset created in the previous question, let's create a scatterplot with the variables PC1 and PC2 to visualize the groups of species/sex and the clusters. In `geom_point`, specify the aesthetics of coloring by `species` and shaping by `sex`. Then add a layer called `stat_ellipse()` with the aesthetic of group by `cluster`. In the cluster containing the fossil specimen GKZ-N 00001, what species and sex are the other sea lions? What can you conclude about the species and sex of the fossil specimen GKZ-N 00001?

```
ggplot(x, aes(x = PC1, y = PC2, color = species, shape = sex)) +
  geom_point() +
  stat_ellipse()
```



Your answer goes here. 1-2 sentences.

4.3 (2 pts) Putting it all together. Reflect on and summarize in 1-2 sentences the different steps taken through this assignment. Compare your conclusions to the findings discussed by the researchers in the article (cite their findings).

In this assignment, I took the sealions dataset from the supplemental information of the article given and organized that data by taking out all the NA values careful not to lose inforation on the fossil specimen GKZ-N 00001 and then ran a PC analysis and graphed results. I concluded that .

```
##       sysname       release       version      nodename       machine
##     "Windows"      "10 x64"  "build 18363"    "ROSE-XPS"      "x86-64"
##         login          user effective_user
##       "roseh"       "roseh"       "roseh"
```