HW 8

Enter your name and EID here: Jongho Yoo (jy23294)

You will submit this homework assignment as 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.

We will use the packages `tidyverse`, `factoextra`, and `cluster` for this assignment.

```
# Load packages
library(tidyverse)
library(factoextra)
library(cluster)
```

Question 1: (1 pt)

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

Read the ***Abstract*** of the article and the section called *Results of Morphometric Analyses*. What was the goal of this study?

The goal of this study was to understand the evolutionary history of GKZ-N 00001 by comparing 39 morphometric measurements between 50 other mandibles of various species to ultimately determine which species GKZ-N 00001 most closely belongs to.

Question 2: (1 pt)

Under the supplemental information, I retrieved the data from a word document into a `.csv` document. Import the dataset from GitHub.

```
# upload data from github
sealions <- read_csv("https://raw.githubusercontent.com/laylaguyot/datasets/main//Sealions.csv")
```

How many rows and how many columns are in this dataset? What does a row represent? What does a column represent?

```
# find number of rows and columns
nrow(sealions)
```

```
## [1] 51
```

```
ncol(sealions)
```

```
## [1] 39
```

There are 51 rows and 39 columns. Each row represents a mandible of a fur seal, sea lion, or GKZ-N 00001. Each column represents the measurements of external morphologies with internal structures.

Question 3: (2 pts)

Before we can analyze the data, let's do some cleaning. When importing this dataset into RStudio, which variables were considered numeric? Why are some measurements not considered as numeric?

```
# Use `spec()` to retrieve full column specification
spec(sealions)
```

```
## cols(
##   ID = col_character(),
##   A = col_character(),
##   B = col_character(),
##   C = col_character(),
##   D = col_character(),
##   E = col_character(),
##   F = col_character(),
##   G = col_character(),
##   H = col_character(),
##   I = col_character(),
##   J = col_character(),
##   K = col_double(),
##   L = col_character(),
##   M = col_character(),
##   N = col_character(),
##   O = col_character(),
##   P = col_character(),
##   Q = col_character(),
##   R = col_character(),
##   S = col_character(),
##   T = col_character(),
##   U = col_character(),
##   V = col_character(),
##   W = col_character(),
##   X = col_character(),
##   Y = col_character(),
##   Z = col_character(),
##   AA = col_character(),
##   AB = col_character(),
##   AC = col_character(),
```

```
##   AD = col_double(),
##   AE = col_character(),
##   AF = col_character(),
##   AG = col_character(),
##   AH = col_character(),
##   AI = col_character(),
##   AJ = col_character(),
##   AK = col_character(),
##   AL = col_character()
## )
```

The 'K' and 'AD' variables are considered numeric. Looking at the data, only variables 'K' and 'AD' did not have any NA values (besides ID variable which is intended to be character). Therefore, in the other columns, since there are NA values, R coerces them as character values because NA is not a numeric value.

Fix this issue by making sure all measurements that should be considered as numeric variables are indeed defined as numeric. Once you know your code does what you want, save the dataset as `sealions_NA`.

```
# make all variables (except ID) as numeric
sealions_NA <- sealions %>%
  mutate_at(vars("A":"AL"), as.numeric)
```

What is the mean rostral tip of mandible `C`?

```
# find mean value of variable 'C' excluding NA
mean(sealions_NA$C, na.rm = TRUE)
```

```
## [1] 34.86622
```

The mean rostal tip of mandible C is 34.9 mm.

Question 4: (2 pts)

You are given the code in this question. But what does the code do? Write comments.

```
sealions_NA <- sealions_NA %>%
  # only select variables where the value for 'GKZ-N 00001' is not NA
  select_if(!(is.na(sealions_NA[sealions_NA$ID == "GKZ-N 00001",]))) %>%
  # remove observations where there is an NA value
  na.omit
```

How many columns and how many rows are remaining in this dataset?

```
# find number of columns and rows
ncol(sealions_NA)
```

```
## [1] 23
```

```
nrow(sealions_NA)
```

```
## [1] 42
```

There are 23 columns and 42 rows remaining in this dataset.

Question 5: (2 pts)

In `sealions_NA`, split the `ID` variable into two variables `species` and `sex` using the brackets `[` or `]` as separators. *Hint:* *`sep = "\\[|\\]"`*. Save the resulting dataset as `sealions_clean`.

```
# separate ID column into species separate name and sex variables
sealions_clean <- sealions_NA %>%
 separate(ID, into = c("species","sex"), sep = "\\[|\\]")
```

REFERENCE SOURCE LINK

How many sea lions are male/female?

```
# first group by sex to find number of female and male
sealions_clean %>%
  group_by(sex) %>%
  summarize(count = n())
```

```
## # A tibble: 3 x 2
##   sex   count
##   <chr> <int>
## 1 f        23
## 2 m        18
## 3 <NA>      1
```

There are 23 females and 18 males

Question 6: (2 pts)

Only keep numeric variables and scale each numeric variable in `sealions_clean`. Save the resulting dataset as `sealions_num`.

```
# only keep numeric variables and then scale
sealions_num <- sealions_clean %>%
  select_if(is.numeric) %>%
  mutate_all(scale)
```

What should the mean of the scaled variable of the rostral tip of mandible C be?

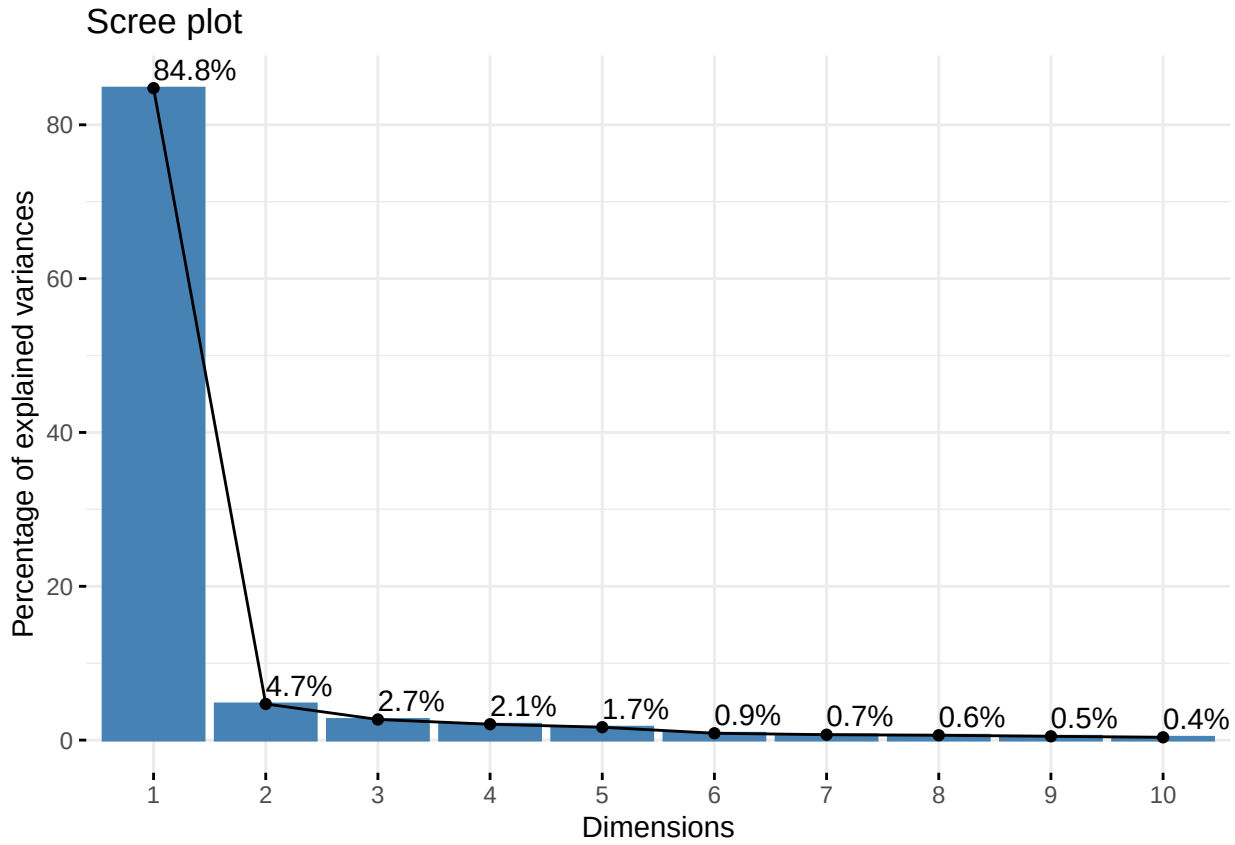
The mean of the scaled variable should be 0. When scaling, you are normalizing the variables so that the values are centered at about 0, and assigned a value depending on the standard deviation from the mean. Therefore, the scaled variable of mandible C should be around 0.

Question 7: (2 pts)

Let's perform PCA on the measurements available for the fossil specimen GKZ-N 00001. Find the principal components (PCs) for the scaled data, `sealions_num`, and save the results as `sealions_pca`. Construct a scree plot. What is the percentage of explained variance for the first principal component?

```
# perform PCA and save as new object
sealions_pca <- sealions_num %>%
  prcomp

# Visualize percentage of variance explained for each PC in a scree plot
fviz_eig(sealions_pca, addlabels = TRUE)
```



The first principal component explains 84.8% of the variance.

Question 8: (2 pts)

How many *known species* are there in `sealions_clean`? Therefore, how many clusters should we look for to identify what species GKZ-N 00001 most likely belongs to?

```r
# determine distinct species and how many of each, excluding GKZ-N 00001
sealions_clean %>%
  filter(species != "GKZ-N 00001") %>%
  group_by(species) %>%
  summarize(count = n())
```

```
## # A tibble: 3 x 2
##   species         count
##   <chr>           <int>
## 1 "C. ursinus "      13
## 2 "E. jubatus "      24
## 3 "Z. japonicus "     4
```

```r
# determine distinct number of species, excluding GKZ-N 00001
sealions_clean %>%
  filter(species != "GKZ-N 00001") %>%
  distinct(species) %>%
  nrow()
```

```
## [1] 3
```

There are 3 known species in the 'sealions_clean' dataset (excluding GKZ-N 00001). Therefore, we should look for 3 clusters to determine what species GKZ-N 00001 most likely belongs to.

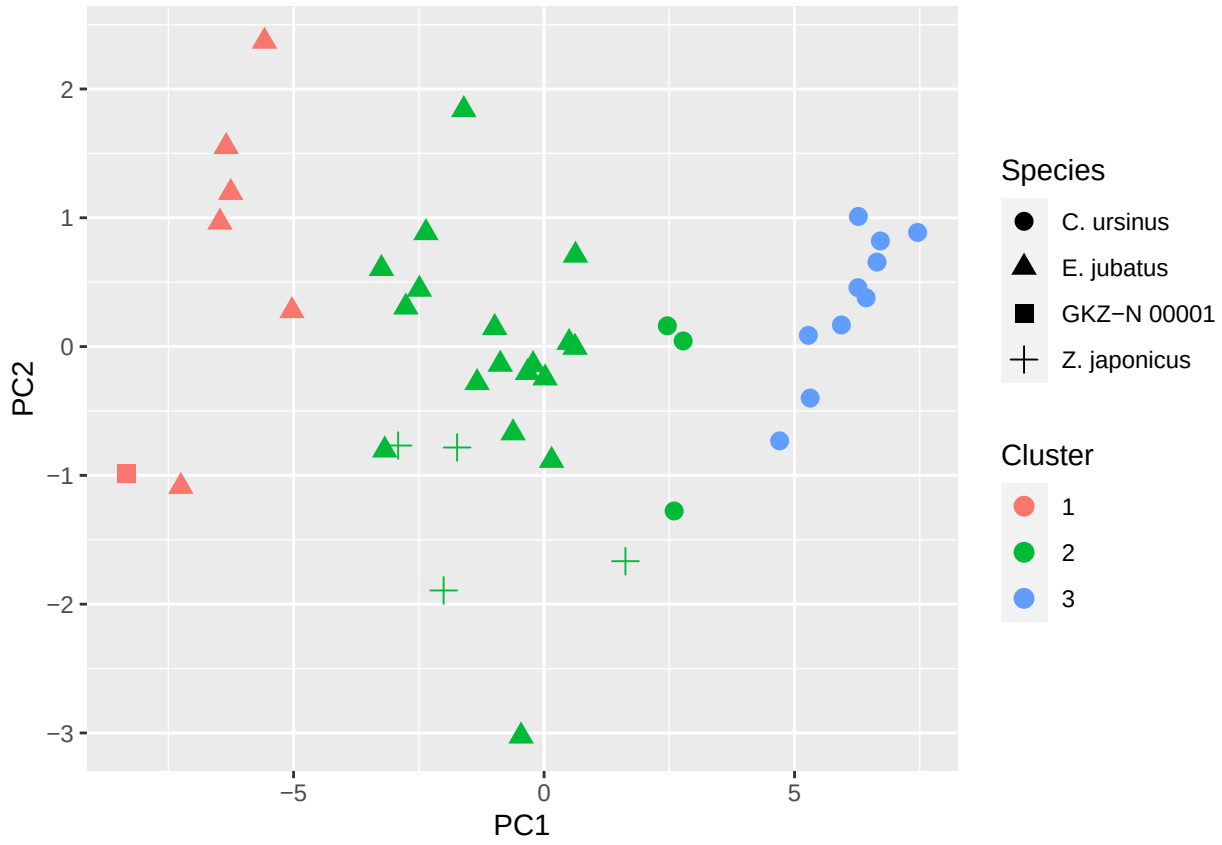
Perform the PAM clustering algorithm on `sealions_num` with the appropriate number of clusters.

```r
# pam clustering with 3 clusters
sealions_pam <- sealions_num %>%
  pam(k = 3)
```

Question 9: (2 pts)

Let's represent the sea lions colored by clusters, shaped by their `species`, and along the first two principal components to take into account all of their measurements. *Big hint: use the object* ***x*** *from the results of the PCA (those are the new coordinates of the sea lions) which you should consider as a data frame. Then add the information about the* ***species*** *from* ***sealions_clean*** *and the* ***clustering*** *from the PAM results.* Create the visualization with a `ggplot()`, add the appropriate variables to the aesthetics, include labels on the axes and legend, and tada!*

```r
# represent the sea lions into clusters
sealions_pca$x %>%
  as.data.frame %>%
  mutate(species = sealions_clean$species,
  cluster = as.factor(sealions_pam$clustering)) %>%
  ggplot(aes(x = PC1, y = PC2, color = cluster, shape = species)) +
  geom_point(size = 3) +
  labs(x = "PC1", y = "PC2", color = "Cluster", shape = "Species")
```



The fossil specimen GKZ-N 00001 appears to be close to which species?

The fossil specimen GKZ-N 00001 appears to be closest to the E. jubatus species. The square data point represents GKZ-N 00001, which is near the cluster of triangle data points on the left representing E. jubatus.

Question 10: (2 pts)

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

In this assignment, we first cleaned the data, such as making the variables numeric and scaling variables to use for clustering anaylsis. Then, we utilized PCA and PAM analysis to create a clustering plot and determine which species GKZ-N 00001 belongs to. The clustering plot in question 9 indicated that GKZ-N 00001 belongs in the E. jubatus species, as it is closest to that cluster. This conclusion is supported by the findings discussed by the researchers: "The mandibular fossil (GKZ-N 00001) from the lower Pleistocene Omma Formation (0.8 Ma) is specifically identified as E. jubatus based on the morphometric analyses"

Formatting: (2 pts)

Comment your code, write full sentences, and knit your file!

```
##                                                                                                                       sysn
##                                                                                                                      "Darw
##                                                                                                                       rele
##                                                                                                                      "21.3
##                                                                                                                       vers
## "Darwin Kernel Version 21.3.0: Wed Jan  5 21:37:58 PST 2022; root:xnu-8019.80.24~20/RELEASE_ARM64_T81
##                                                                                                                      noden
##                                                                                                  "Stevens-MacBook-Pro-2.loc
##                                                                                                                       mach
##                                                                                                                       "arm
##                                                                                                                         lo
##                                                                                                                        "ro
##                                                                                                                          u
##                                                                                                                   "steveny
##                                                                                                                 effective_u
##                                                                                                                   "steveny
```