HW 6

Enter your name and EID here: Harini Shanmugam

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: (2 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

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 look at fossil records of Otariidae sea lions to determine the origins of this family of sealions. The main finding was that there was almost no difference between the fossil GKZ-N 00001 and the currently alive and studied species *E. jubatus.*

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?

```
# count number of rows in dataset
nrow(sealions)
```

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

```r
# count number of columns in dataset
ncol(sealions)
```

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

There are 51 rows and 39 columns in this dataset. Each row represents a unique sea lion (its species, sex, and what number it is of that specific species and sex combination). Each column represents a different external feature.

Question 3: (1 pt)

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?

```r
# select the columns that are numeric
sealions %>%
  select_if(is.numeric)
```

```
## # A tibble: 51 x 2
##         K    AD
##     <dbl> <dbl>
##  1  57.8  69.3
##  2  64.6  76.9
##  3  63.6  74.4
##  4  45.2  69.1
##  5  41.4  70.6
##  6  43.6  67.1
##  7  43.6  64.8
##  8  41.8  64.2
##  9  68.0  68.2
## 10  44.0  63.3
## # ... with 41 more rows
```

```r
# select the columns that are not numeric
sealions %>%
  select_if(negate(is.numeric))
```

```
## # A tibble: 51 x 37
##    ID    A     B     C     D     E     F     G     H     I     J     L     M
##    <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr> <chr>
##  1 E. j~ 262   232   62.39 31.12 63.12 59.01 43.99 46.83 62.56 62.65 87.09 24.33
##  2 E. j~ 285   242   64.52 31.71 70.48 75.58 44.33 62.52 63.13 63.5  97.46 14.71
##  3 E. j~ 265.8 242.2 53.06 30.16 70.53 60.28 47.98 50.82 61.99 63.89 99.17 18.36
##  4 E. j~ 244   212   44.88 26.05 55.94 52.04 38.46 39.89 51.77 55.91 85.05 19.8
##  5 E. j~ 237   208.~ 39.38 26.09 51.21 49.44 37.25 37.93 45.6  49.02 83.41 24.12
##  6 E. j~ 228   201.~ 39.52 25.39 51.19 48.07 36.39 37.22 62.98 49.68 76    17.2
##  7 E. j~ 227   202.~ 48.39 24.85 48.46 49.25 39.05 39.12 48.61 52.58 81.2  18.94
```

```
##  8 E. j~ 226   190.~ 55.24 27.24 48.99 34.04 30.51 29.41 50.34 50.11 75.34 14.6
##  9 E. j~ 282.5 257.2 49.62 31.37 72.71 45.21 40.08 49.14 63.85 66.3  104.~ 17.66
## 10 E. j~ 237   215   50.53 16.15 50.37 46.99 38.65 37.59 50.2  54.06 80.81 19.97
## # ... with 41 more rows, and 24 more variables: N <chr>, O <chr>, P <chr>,
## #   Q <chr>, R <chr>, S <chr>, T <chr>, U <chr>, V <chr>, W <chr>, X <chr>,
## #   Y <chr>, Z <chr>, AA <chr>, AB <chr>, AC <chr>, AE <chr>, AF <chr>,
## #   AG <chr>, AH <chr>, AI <chr>, AJ <chr>, AK <chr>, AL <chr>
```

K and AD are the only numeric variables. All of the other variables are characters. They were not considered as numeric because those columns had missing values, which were registered as characters since they were denoted with the "-" symbol rather than empty values among numeric values.

Question 4: (1 pt)

Using `mutate_all()`, replace all `-` in the dataset by missing values *NA* then make sure all measurements are defined as numeric variables with `mutate_at()`. Overwrite the dataset `sealions`.

```
# overwrite dataset
sealions <- sealions %>%
  # replace all NA values with "-"
  mutate_all(na_if, "-") %>%
  # make all variables numeric
  mutate_at(2:39, as.numeric)
```

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

```
# mean of column "C" in sealions dataset excluding NA values
mean(sealions$C, na.rm=T)
```

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

The mean rostral tip of mandible C is 34.86622 millimeters.

Question 5: (2 pts)

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

```
# overwrite data set
sealions <- sealions %>%
  # select the rows when its is not NA/ missing uptil the 51st row
  select_if(!(is.na(sealions[51,]))) %>%
  # remove na values
  na.omit
```

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

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

Question 6: (2 pts)

Use `dplyr` functions on `sealions` to split the `ID` variable into two variables `species` and `sex` with the function `separate()`. *Hint: in the `ID` variable, what symbol separates the species from sex?* 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: You could use the functions `mutate()` and `replace_na()`.* Save the resulting dataset as `sealions_clean`.

```
# overwrite object
sealions_clean <- sealions %>%
  # separate ID variable into two variables sepcies and sex
  separate(ID, into=c("species", "sex"), sep="\\[|\\]") %>%
  # replace na value for sex with m
  mutate(sex = replace_na(sex, "m"))
```

How many sealions are male/female?

```
# count of each type in sex varaible in sealions dataset
table(sealions_clean$sex)
```

```
##
##  f  m
## 23 19
```

There are 19 male and 23 female sealions.

Question 7: (1 pt)

Using `dplyr` functions, only keep numeric variables and scale each numeric variable. Save the resulting dataset as `sealions_num`. What should the mean of the scaled variable of the rostral tip of mandible C be?

```
# set changes to new object
sealions_num <- sealions %>%
  # only keep numeric variables
  select_if(is.numeric) %>%
  # scale the variables
  scale

# mean of the scaled variable C
df <- as.data.frame(sealions_num)
mean(df$C)
```

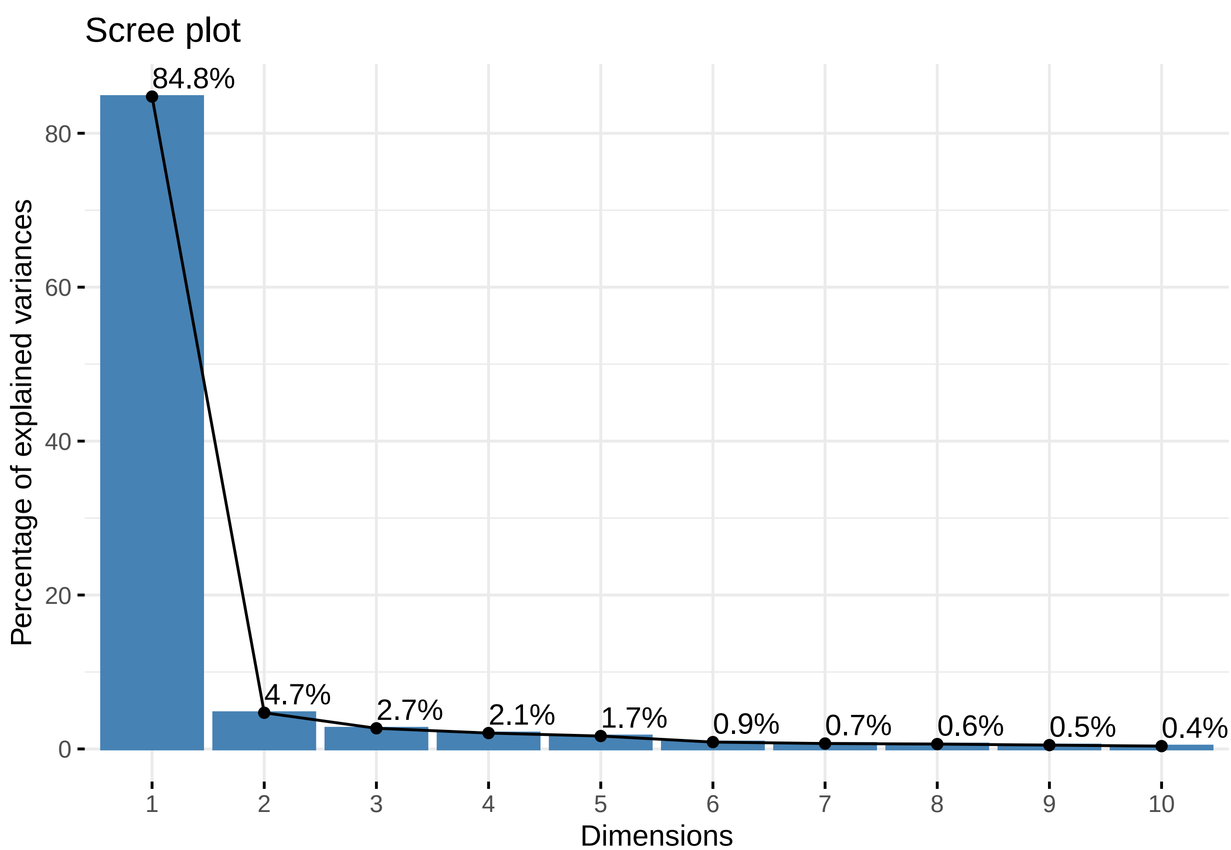
```
## [1] -9.410462e-16
```

The mean is -9.410462e^-16 which is negligible and can be rounded to 0.

Question 8: (2 pts)

Let's perform PCA on the measurements available for the fossil specimen GKZ-N 00001. Using the function `prcomp()`, calculate the principal components (PCs) for the scaled data, `sealions_num`, obtained in the previous question. Construct a scree plot with the function `fviz_eig()` from the package `factoextra`. What is the cumulative percentage of explained variance for PC1 and PC2?

```
# apple prcomp to sealions_num dataset and save to new obbject
pca <- sealions_num %>%
  prcomp

# scree plot
fviz_eig(pca, addlabels = TRUE)
```



The cumulative percentage of explained variance for PC1 and PC2 is 89.5%.

Question 9: (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
# count number of unique/ known species
sealions_clean %>%
  group_by(species) %>%
  summarize(n_distinct(species))
```

```
## # A tibble: 4 x 2
##   species          `n_distinct(species)`
##   <chr>                            <int>
## 1 "C. ursinus "                        1
## 2 "E. jubatus "                        1
## 3 "GKZ-N 00001"                        1
## 4 "Z. japonicus "                      1
```

There are 3 known species in sealions_clean. So we should look for 3 clusters to identify what species GKZ-N 00001 most likely belongs to.

Perform the PAM clustering algorithm on `sealions_num`, run the PAM clustering algorithm.

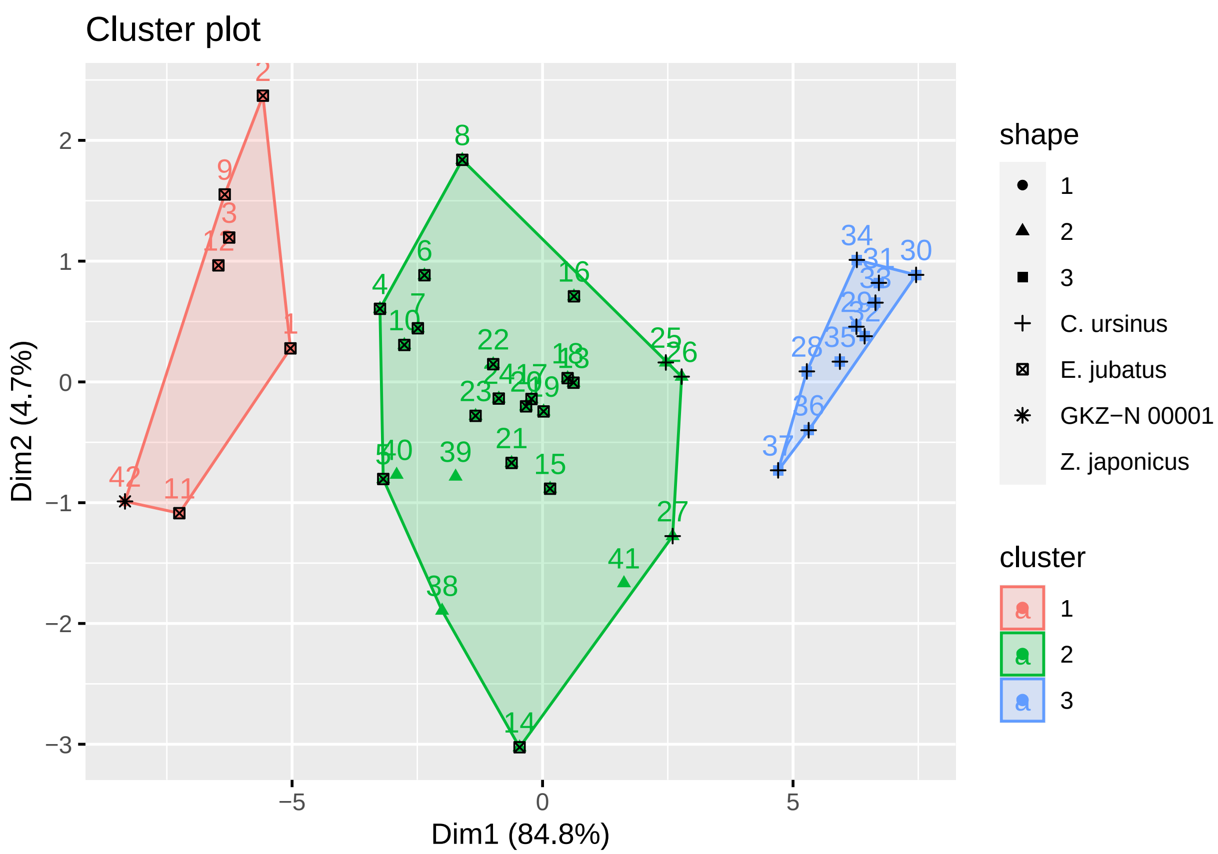
```r
# pam clustering
pam_results <- sealions_num %>%
  # k=3 clusters
  pam(k = 3)

pam_results
```

```
## Medoids:
##      ID          C           D           I          J           K          L
## [1,]  3  1.2618249  1.22177839  1.49625052  1.3937596  1.62575587  1.3810056
## [2,] 21  0.1374137 -0.04546734  0.02459524  0.0139526  0.05209785  0.4070515
## [3,] 32 -1.4412068 -1.18306858 -1.24933966 -1.3699401 -1.34409429 -1.4418484
##               M           X           Y          Z          AA          AB
## [1,] -0.4719187  1.50976094  1.45967808  1.8052423  2.2041335  1.2971166
## [2,]  0.7177680  0.07288501 -0.02465227  0.5409734  0.2814031  0.3872614
## [3,] -0.7179830 -1.53716132 -1.31613034 -1.2753982 -1.2446719 -1.5870190
##              AC          AD          AE          AF          AG          AH
## [1,]  1.2693312  1.0631112  1.7248343  1.6149063  1.40394950  1.2428992
## [2,]  0.3353378  0.5456277 -0.4141844 -0.2474853 -0.01103809  0.2779867
## [3,] -1.5148513 -1.7135154 -1.2594119 -1.2465560 -1.45827611 -1.4588559
##              AI          AJ          AK          AL
## [1,]  0.8797091  0.7661937  1.0232476  1.2766766
## [2,]  0.3988148  0.1415360 -0.1517907 -0.1773895
## [3,] -1.5443337 -1.6092652 -1.3911168 -1.0982980
## Clustering vector:
##  [1] 1 1 1 2 2 2 2 2 1 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 2
## [39] 2 2 2 1
## Objective function:
##    build     swap
## 2.214681 2.194480
##
## Available components:
##  [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"
##  [6] "clusinfo"   "silinfo"    "diss"       "call"       "data"
```

Question 10: (2 pts)

Represent the clusters along the first two principal components and specify to shape the observations by their `species` in the aesthetics. *Note: you can either use* `ggplot` *or* `fviz_cluster()`*.*

```
# cluster plot
fviz_cluster(pam_results, data = sealions,
             shape=sealions$species) +
  # specify shapes of points
  geom_point(aes(shape=sealions_clean$species)) +
  guides(shape = guide_legend(title = "shape"))
```



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

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

Question 11: (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).

We took data on the measurements of different parts of current known sealion species to identify the species of an unknown sealion fossil using some tidying and PCA and identified it to belong to the E. jubatus species, just as told by the research article, "there is almost no difference between GKZ-N 00001 and extant male individuals of *E. jubatus*".

Formatting: (2 pts)

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

```
##                                                                                              sysna
##                                                                                             "Darwi
##                                                                                              relea
##                                                                                             "21.6
##                                                                                              versi
## "Darwin Kernel Version 21.6.0: Wed Aug 10 14:28:35 PDT 2022; root:xnu-8020.141.5~2/RELEASE_ARM64_T81
##                                                                                             nodena
##                                                                               "Harinis-Air.attlocal.n
##                                                                                              machi
##                                                                                              "arm
##                                                                                                lo
##                                                                                               "ro
##                                                                                                 u
##                                                                                     "harinishanmuga
##                                                                                         effective_u
##                                                                                     "harinishanmuga
```