HW 6

Enter your name and EID here: Erik Mercado, emm4376

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 main goal of the study was to find out which sea lion is the closest to the new found mandible and the result was that the male E. jubatus was the closest to the sample mandible.

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.

```
# download 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 and columns
nrow(sealions)
```

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

```
ncol(sealions)
```

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

There are 51 rows and 39 39 columns. A row represents mandibles of fur seals and sea lions with GKZ-N 00001. A column represents a measurement for external morphologies with internal structures by CT scan data.

Question 3: (1 pt)

Before we can analyze the data, let's do some cleaning. Using a combination of the `select()`, `where()`, and a predicate function like `is.character()` we can scan through all the columns of the dataset and see which columns are of character type.

```
## Select all of the columns that have character type
sealions |>
    select(where(is.character))
```

```
## # 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>
```

When importing this dataset into RStudio, which variables were considered numeric? Why are some measurements not considered as numeric? Use the `is.numeric()` predicate function here.

```
## Select all of the columns that have numeric type
 sealions |>
  select(where(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
```

Variables K and AD are considered numeric. Some measurements are not considered numeric because they have missing data/have a non-numeric value for their missing values.

Question 4: (1 pt)

The functions `mutate()` and `across()`, when used together, can make changes across a range of columns in a data frame.

Using `mutate()` and `across()`, replace all `-` in the dataset with the missing values *NA* and then then make sure all measurements are defined as numeric variables with. The first part of the code replaces the "-" with NA values. Write the second part to coerce all of the columns (except for the ID column!) to be numeric. Make sure to overwrite the dataset `sealions`.

NOTE: Look at the examples in `?across` to get a sense of how to use the `across()` function.

```
# overwrite original dataframe
sealions <- sealions |>
    ## Replace all "-" with NA
    mutate(across(where(is.character), ~ na_if(.x, "-"))) |>
    ## Coerce all columns (except for ID) to be numeric
    mutate(across(c(2:39), ~ as.numeric(.x)))
```

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

```
# filter out rows with na values for C and then find the mean of C
sealions |>
  filter(!is.na(C)) |>
  summarise(mean_C = mean(C))
```

```
## # A tibble: 1 x 1
##   mean_C
##    <dbl>
## 1   34.9
```

The mean rostral tip of mandible C is 34.86622 mm.

Question 5: (2 pts)

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

```
sealions <- sealions %>%
  ## only includes columns with no na values in the 51st row
  select_if(!is.na(sealions[51,])) %>%
  ## get rids of any rows with missing values
  na.omit
```

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

There 42 rows and 23 columns remaining.

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`.

```
# separate ID into sex and species
sealions_clean <- sealions |>
  separate(ID, into = c("species", "sex"), sep = "\\[|\\]") |>
  mutate(sex = replace_na(sex, "m"))
```

How many sealions are male/female?

```
# count different sexes from dataset
sealions_clean |>
  group_by(sex) |>
  summarise(n = n())
```

```
## # A tibble: 2 x 2
##   sex       n
##   <chr> <int>
## 1 f        23
## 2 m        19
```

There are 23 females and 28 males.

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?

```
# only selsct numeric columns and scaling them. then take mean of C
sealions_num <- sealions_clean |>
  select(where(is.numeric)) |>
  mutate(across(everything(), scale))
mean(sealions_num$C)
```

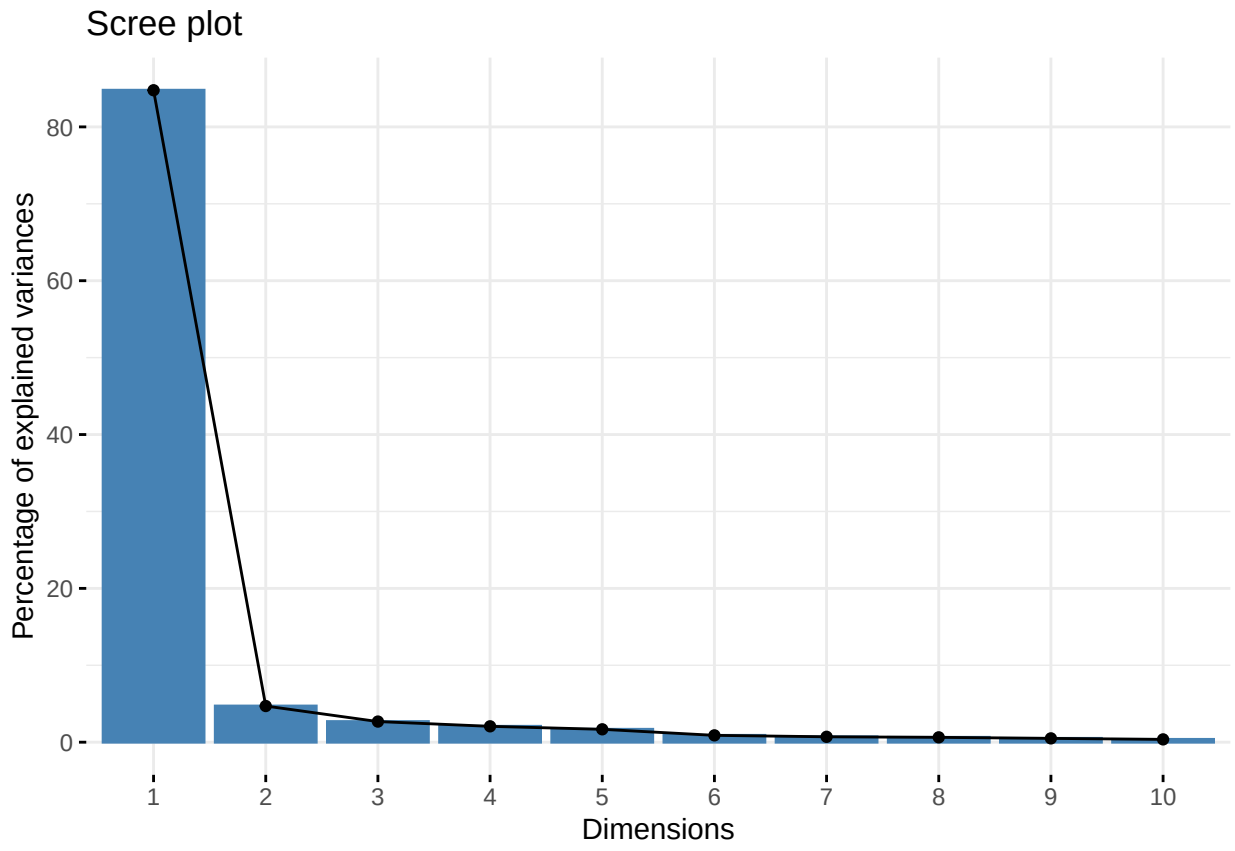
```
## [1] 1.487009e-16
```

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

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?

```
# find the cumulative percentage of variance from pc1 and pc2
library(factoextra)
sealions_pca <- sealions_num |>
  prcomp()
fviz_eig(sealions_pca)
```




```
summary(sealions_pca)
```

```
## Importance of components:
##                           PC1      PC2      PC3      PC4      PC5      PC6      PC7
```

```
## Standard deviation     4.3183 1.01734 0.76853 0.67401 0.60754 0.44301 0.39384
## Proportion of Variance 0.8476 0.04704 0.02685 0.02065 0.01678 0.00892 0.00705
## Cumulative Proportion  0.8476 0.89468 0.92152 0.94217 0.95895 0.96787 0.97492
##                            PC8     PC9    PC10   PC11    PC12   PC13    PC14
## Standard deviation     0.37325 0.33112 0.28423 0.2095 0.20712 0.1876 0.14794
## Proportion of Variance 0.00633 0.00498 0.00367 0.0020 0.00195 0.0016 0.00099
## Cumulative Proportion  0.98126 0.98624 0.98991 0.9919 0.99386 0.9955 0.99645
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.14003 0.11973 0.11556 0.10015 0.09388 0.08251 0.05608
## Proportion of Variance 0.00089 0.00065 0.00061 0.00046 0.00040 0.00031 0.00014
## Cumulative Proportion  0.99734 0.99799 0.99860 0.99906 0.99946 0.99977 0.99991
##                           PC22
## Standard deviation     0.04457
## Proportion of Variance 0.00009
## Cumulative Proportion  1.00000
```

The cumulative percentage of explained variance for PC1 and PC2 is 0.89468.

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?

```
# get species count
 sealions_clean |>
  group_by(species) |>
  summarize(n = n())
```

```
## # A tibble: 4 x 2
##   species             n
##   <chr>           <int>
## 1 "C. ursinus "      13
## 2 "E. jubatus "      24
## 3 "GKZ-N 00001"       1
## 4 "Z. japonicus "     4
```

There are 3 known species.

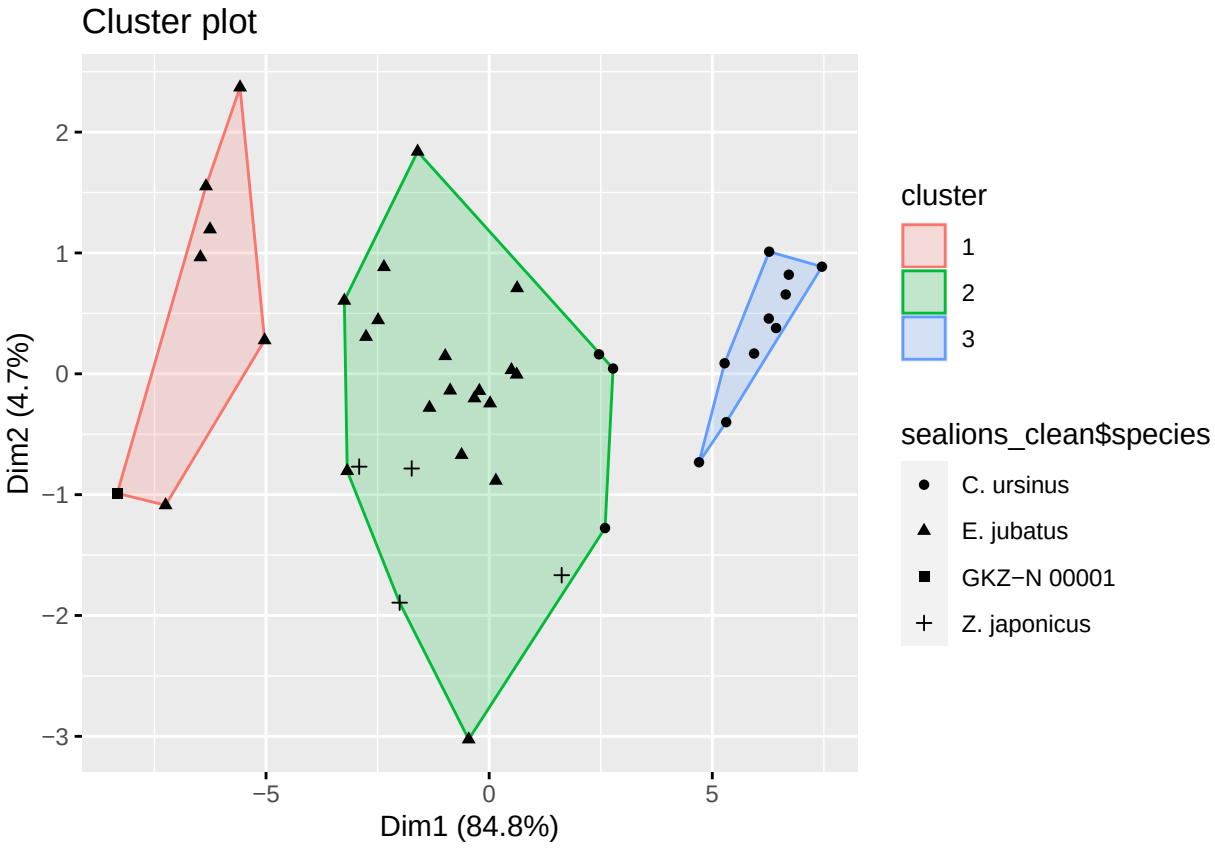
Try using the PAM clustering algorithm using the `pam()` function in R (it is similar to K-means). Perform the PAM clustering algorithm on `sealions_num`, run the PAM clustering algorithm.

```
# apply pam algorithm
 pam_results <- sealions_num |>
  pam(k=3)
```

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()`.*

```
# plot clusters
fviz_cluster(pam_results, geom = NULL) +
  geom_point(aes(shape = sealions_clean$species))## Add geoms to complete the plot
```



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

The fossil specimen seems to be close to E. jubatus.

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).

The first steps were preparing the data to be how we need it. Then by measuring and comparing the different rostral tip of mandible of different species, we were able to

---

Formatting: (2 pts)

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

```
##                                                                                                  sysname
##                                                                                                 "Darwin"
##                                                                                                  release
##                                                                                                 "21.6.0"
##                                                                                                  version
## "Darwin Kernel Version 21.6.0: Sun Nov  6 23:31:16 PST 2022; root:xnu-8020.240.14~1/RELEASE_X86_64"
##                                                                                                 nodename
##                                                                                    "Eriks-MBP-2423.lan"
##                                                                                                  machine
##                                                                                                 "x86_64"
##                                                                                                    login
##                                                                                                  "root"
##                                                                                                     user
##                                                                                                  "erik"
##                                                                                          effective_user
##                                                                                                  "erik"
```