# Team Essay 5

Sahil, Ariel, Ilya, Lekha, Anthony, Trang

4/20/2021

## Introduction

In this essay, our team utilizes K-means clustering analysis to partition a dataset of 62 animal species into sets of distinct groups. For this particular dataset, our model will focus on attempting to create groups of animals which share similarities in their sleeping characteristics. Our model will employ data on non dreaming sleep, dreaming sleep, total sleep in tandem with other attributes of each species to assist this process. Through the use of these variables we will create varying amounts of clusters that will provide us with an indication of which number of clusters works best. As a result, we plan to find the optimal amount of groups that maintain a high level of similarity between the sleep of species within a group and a low level of similarity between the sleep of species in different groups.

## Formula and Basics

K means clustering is an unsupervised algorithm which sorts a set of observations into clusters. The algorithm starts by indicating k, the number of clusters. Second, the algorithm randomly selects k observations which will serve as the initial centers of the clusters, or centroids. Third, the remaining observations are assigned to the closest centriod based of their distance from each cluster mean. To compute the distance, we will use the Euclidean distance formula: $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$. After, the new means of each centriod is computed and the previous steps are repeated in order to see if any observations are now closer to the newly calculated centriods. This process repeats until the clusters formed in the current step equal the clusters formed in the previous step.

## Loading Required R packages

```
library(tidyverse)  # data manipulation
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(cluster)    # clustering algorithms
library(factoextra) # clustering algorithms & visualization
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(readxl)
```

## Data Description

- BodyWt: body weight of the animal in kilograms
- BrainWt: brain weight of the animal in grams
- NonDreaming: slow wave ("nondreaming") sleep in hours per day
- Dreaming: paradoxical ("dreaming") sleep in hours per day
- TotalSleep: total sleep, sum of slow wave and paradoxical sleep in hours per day
- LifeSpan: maximum life span of the animal in years
- Gestation: gestation time of the animal in years
- Predation: predation index (1-5), 1 = minimum (least likely to be preyed upon); 5 = maximum (most likely to be preyed upon)
- Exposure sleep exposure index (1-5), 1 = least exposed (e.g. animal sleeps in a well-protected den); 5=most exposed
- Danger overall danger index (1-5) (based on the above two indices and other information) 1 = least danger (from other animals); 5 = most danger (from other animals)

## Examples of data and problem

```
animal_data <- read_excel("Animal.xlsx")
animal_data <- na.omit(animal_data) # To remove any missing value that might be present in the data
animal_data
```

```
## # A tibble: 62 x 11
##    Species  BodyWt BrainWt NonDreaming Dreaming TotalSleep LifeSpan Gestation
##    <chr>     <dbl>   <dbl>       <dbl>    <dbl>      <dbl>    <dbl>     <dbl>
##  1 Africa~ 6.65e+3  5712           0        0          3.3     38.6       645
##  2 Africa~ 1.00e+0     6.6         6.3      2          8.3      4.5        42
##  3 Arctic~ 3.38e+0    44.5         0        0         12.5     14          60
##  4 Arctic~ 9.20e-1     5.7         0        0         16.5      0          25
##  5 Asiane~ 2.55e+3  4603           2.1      1.8        3.9     69         624
##  6 Baboon  1.06e+1   180.          9.1      0.7        9.8     27         180
##  7 Bigbro~ 2.30e-2     0.3        15.8      3.9       19.7     19          35
##  8 Brazil~ 1.60e+2   169           5.2      1          6.2     30.4       392
##  9 Cat     3.30e+0    25.6        10.9      3.6       14.5     28          63
## 10 Chimpa~ 5.22e+1   440           8.3      1.4        9.7     50         230
## # ... with 52 more rows, and 3 more variables: Predation <dbl>, Exposure <dbl>,
## #   Danger <dbl>
```

## Visualization

```
animal_frame <- as.data.frame(animal_data)
rownames(animal_frame) <- animal_frame$Species
animal_frame <- subset(animal_frame, select = -Species)
animal_frame <- scale(animal_frame)
head(animal_frame)
```

```
##                            BodyWt    BrainWt NonDreaming   Dreaming  TotalSleep
## Africanelephant         7.1791720  5.8357398 -1.37857636 -1.0512000 -1.26975768
## Africangiantpouchedrat -0.2199724 -0.2972594 -0.08510548  0.2707959 -0.30095413
## ArcticFox              -0.2173200 -0.2565190 -1.37857636 -1.0512000  0.51284085
## Arcticgroundsquirrel   -0.2200614 -0.2982269 -1.37857636 -1.0512000  1.28788369
## Asianelephant           2.6115655  4.6436242 -0.94741940  0.1385964 -1.15350126
## Baboon                 -0.2093514 -0.1114012  0.48977046 -0.5885014 -0.01031307
##                          LifeSpan  Gestation  Predation   Exposure     Danger
```

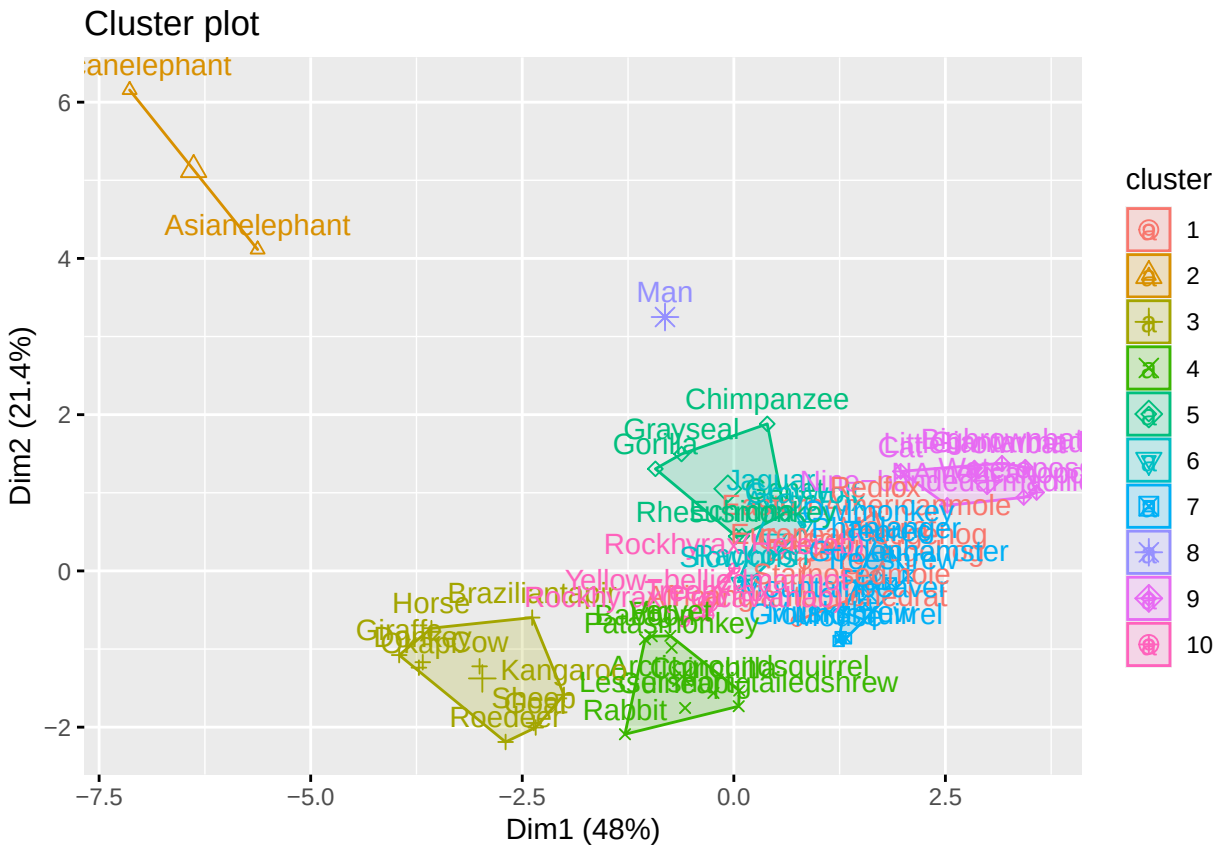
```
## Africanelephant         1.0946653  3.5003089  0.0873957  1.6080871  0.2685837
## Africangiantpouchedrat -0.7712876 -0.6234892  0.0873957 -0.8844479  0.2685837
## ArcticFox              -0.2514473 -0.5003908 -1.2672377 -0.8844479 -1.1190986
## Arcticgroundsquirrel   -1.0175277 -0.7397489  1.4420291 -0.2613142  0.2685837
## Asianelephant           2.7581540  3.3566940  0.0873957  1.6080871  0.9624248
## Baboon                  0.4599130  0.3202656  0.7647124  0.9849534  0.9624248
```

```
k10 <- kmeans(animal_frame, centers=10, nstart=25)
k10
```

```
## K-means clustering with 10 clusters of sizes 8, 2, 10, 9, 6, 5, 10, 1, 7, 4
##
## Cluster means:
##         BodyWt     BrainWt NonDreaming   Dreaming   TotalSleep   LifeSpan
## 1  -0.22010771 -0.29453176   0.1612699  0.5269327  0.006640992 -0.7596596
## 2   4.89536874  5.23968198  -1.1629979 -0.4563018 -1.211629471  1.9264097
## 3   0.02860344  0.04824984  -1.0295445 -0.7273110 -1.399577359  0.4670266
## 4  -0.19369724 -0.23718904   0.1932075 -0.4122353  0.116707840 -0.2064553
## 5  -0.15521371 -0.05487694  -0.1877619 -0.4563018 -0.223449851  1.2013693
## 6  -0.18874256 -0.22422758  -1.3785764 -1.0512000  0.408210063 -0.1529513
## 7  -0.22052563 -0.29922336   0.9702025  0.5351951  0.772480198 -0.7187564
## 8  -0.15213120  1.11457517  -0.1261681  0.2046961 -0.359082348  4.4544748
## 9  -0.20964008 -0.28467488   1.4811382  1.8666339  1.644956994 -0.2944416
## 10 -0.21819300 -0.28753117  -0.5829891 -0.7372260 -1.066308938 -0.5592475
##     Gestation   Predation    Exposure     Danger
## 1  -0.6303281 -0.58992098 -0.72866448 -0.5987177
## 2   3.4285015  0.08739570  1.60808713  0.6155042
## 3   0.9330223  1.23883406  1.60808713  1.5174976
## 4  -0.2021430  1.14099942  0.63876794  0.9624248
## 5   0.2108447 -1.04146543  0.05025272 -0.8878182
## 6  -0.3964410 -0.99631099 -0.63519442 -0.8415621
## 7  -0.6221215  0.15512737 -0.75982117 -0.0783369
## 8   0.9152414 -1.26723766 -0.88444792 -1.1190986
## 9  -0.6234892 -0.97695908 -0.79542881 -1.1190986
## 10  0.2655551 -0.08193347 -0.57288104 -0.2517972
##
## Clustering vector:
##         Africanelephant  Africangiantpouchedrat               ArcticFox
##                       2                       1                       6
##    Arcticgroundsquirrel           Asianelephant                  Baboon
##                       4                       2                       4
##             Bigbrownbat          Braziliantapir                     Cat
##                       9                       3                       9
##              Chimpanzee             Chinchilla                     Cow
##                       5                       4                       3
##          Deserthedgehog                  Donkey     EasternAmericanmole
##                       1                       3                       1
##                 Echidna        Europeanhedgehog                  Galago
##                       5                       1                       1
##                   Genet          Giantarmadillo                 Giraffe
##                       5                       9                       3
##                    Goat           Goldenhamster                 Gorilla
##                       3                       7                       5
##                Grayseal               Graywolf          Groundsquirrel
##                       5                       6                       7
```

```
##                Guineapig                    Horse                   Jaguar
##                        4                        3                        6
##                 Kangaroo Lessershort-tailedshrew          Littlebrownbat
##                        3                        4                        9
##                      Man                  Molerat          Mountainbeaver
##                        8                        1                        7
##                    Mouse                Muskshrew        NAmericanopossum
##                        7                        7                        9
##     Nine-bandedarmadillo                    Okapi               Owlmonkey
##                        9                        3                        7
##              Patasmonkey              Phanlanger                      Pig
##                        4                        7                        4
##                   Rabbit                  Raccoon                      Rat
##                        4                        6                        7
##                   Redfox             Rhesusmonkey      Rockhyrax(Heterob)
##                        1                        5                       10
##   Rockhyrax(Procaviahab)                  Roedeer                    Sheep
##                       10                        3                        3
##                Slowloris            Starnosedmole                   Tenrec
##                        6                        1                        7
##                Treehyrax                Treeshrew                   Vervet
##                       10                        7                        4
##             Wateropossum    Yellow-belliedmarmot
##                        9                       10
##
## Within cluster sum of squares by cluster:
##  [1]  7.564107 13.584216 17.943972 16.626587 12.822461  2.014048 11.028266
##  [8]  0.000000 13.368639  5.307656
##  (between_SS / total_SS =  83.6 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
fviz_cluster(k10, data = animal_frame)
```




This cluster plot serves as a representation for our data given that K is equal to 10. As you can see by the 10 different colors, 10 groups of species which are the most similar were created. Later on we will see how our data changes given different K values and how it affects the total sum of squares within the clusters.

## Analysis

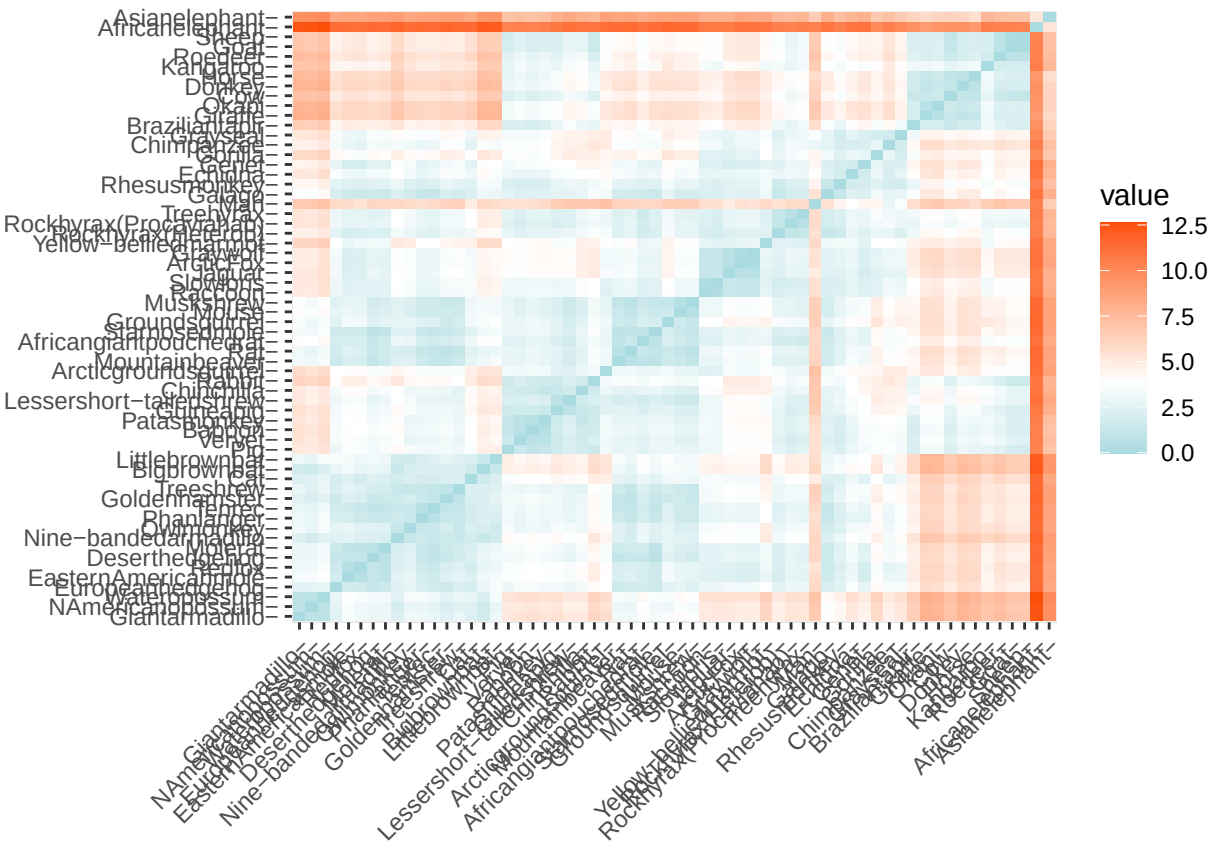
## Computation

```
animal_frame <- as.data.frame(animal_data)
rownames(animal_frame) <- animal_frame$Species
animal_frame <- subset(animal_frame, select = -Species)
animal_frame <- scale(animal_frame) # As we don't want the clustering algorithm to depend to an arbitra
head(animal_frame)
```

```
##                            BodyWt    BrainWt NonDreaming   Dreaming  TotalSleep
## Africanelephant         7.1791720  5.8357398 -1.37857636 -1.0512000 -1.26975768
## Africangiantpouchedrat -0.2199724 -0.2972594 -0.08510548  0.2707959 -0.30095413
## ArcticFox              -0.2173200 -0.2565190 -1.37857636 -1.0512000  0.51284085
## Arcticgroundsquirrel   -0.2200614 -0.2982269 -1.37857636 -1.0512000  1.28788369
## Asianelephant           2.6115655  4.6436242 -0.94741940  0.1385964 -1.15350126
## Baboon                 -0.2093514 -0.1114012  0.48977046 -0.5885014 -0.01031307
##                          LifeSpan  Gestation  Predation   Exposure     Danger
## Africanelephant         1.0946653  3.5003089  0.0873957  1.6080871  0.2685837
## Africangiantpouchedrat -0.7712876 -0.6234892  0.0873957 -0.8844479  0.2685837
## ArcticFox              -0.2514473 -0.5003908 -1.2672377 -0.8844479 -1.1190986
## Arcticgroundsquirrel   -1.0175277 -0.7397489  1.4420291 -0.2613142  0.2685837
## Asianelephant           2.7581540  3.3566940  0.0873957  1.6080871  0.9624248
```

```
## Baboon                    0.4599130  0.3202656  0.7647124  0.9849534  0.9624248
```

```
distance <- get_dist(animal_frame)
fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07")) # Visualizing a
```




This starts to illustrate which species of animals have large dissimilarities (red) between sleep characteristics versus those that appear to be fairly similar (teal).

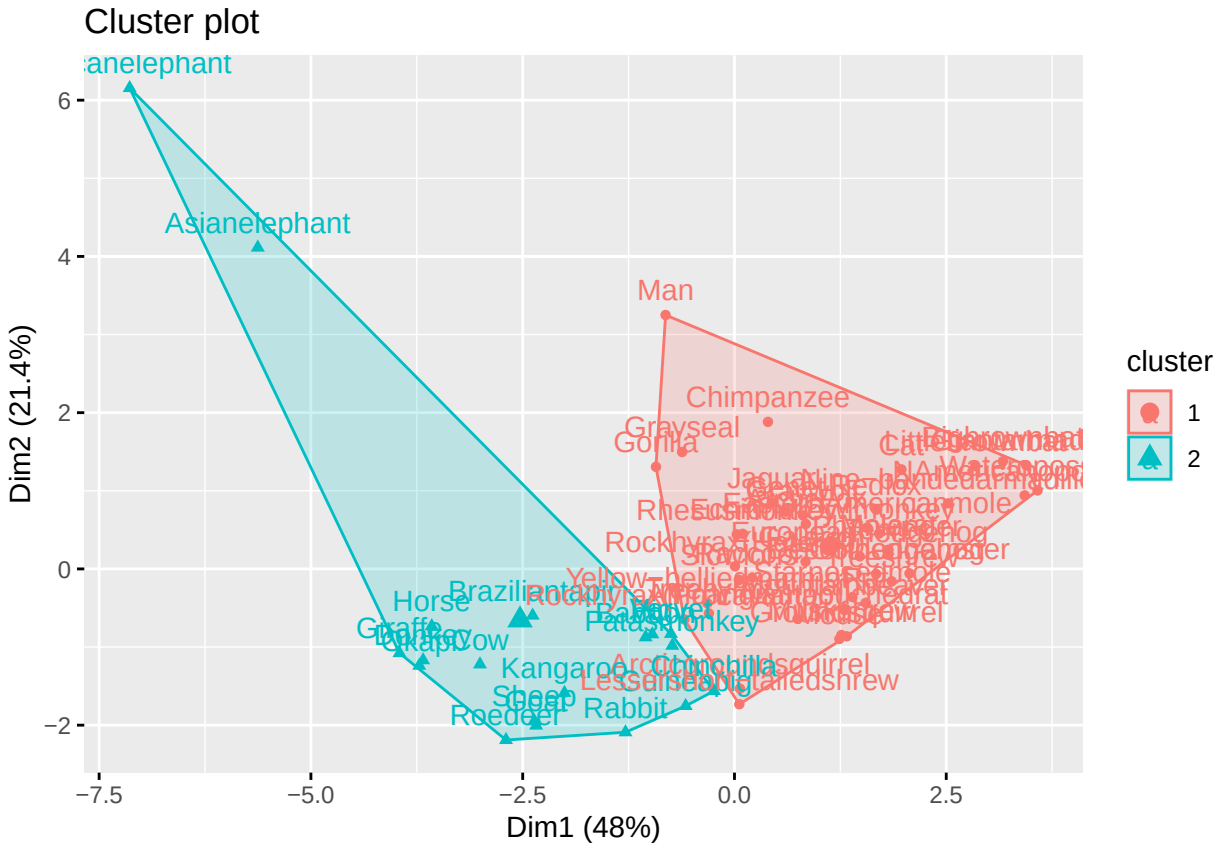
## Interpretation

```
k2 <- kmeans(animal_frame, centers = 2, nstart = 25) # The kmeans function also has an nstart option th
# str(k2)
k2
```

```
## K-means clustering with 2 clusters of sizes 43, 19
##
## Cluster means:
##       BodyWt    BrainWt NonDreaming   Dreaming TotalSleep   LifeSpan  Gestation
## 1 -0.2040613 -0.2192928   0.2257286  0.2492751  0.3839674 -0.1990179 -0.3691017
## 2  0.4618230  0.4962943  -0.5108595 -0.5641489 -0.8689790  0.4504090  0.8353354
##    Predation   Exposure     Danger
## 1 -0.4639086 -0.5946183 -0.5543442
## 2  1.0498984  1.3457150  1.2545684
##
## Clustering vector:
##         Africanelephant  Africangiantpouchedrat               ArcticFox
##                       2                       1                       1
##    Arcticgroundsquirrel           Asianelephant                  Baboon
```

```
##                       1                       2                       2
##             Bigbrownbat          Braziliantapir                     Cat
##                       1                       2                       1
##              Chimpanzee              Chinchilla                     Cow
##                       1                       2                       2
##          Deserthedgehog                  Donkey     EasternAmericanmole
##                       1                       2                       1
##                 Echidna        Europeanhedgehog                  Galago
##                       1                       1                       1
##                   Genet          Giantarmadillo                 Giraffe
##                       1                       1                       2
##                    Goat           Goldenhamster                 Gorilla
##                       2                       1                       1
##                Grayseal                Graywolf          Groundsquirrel
##                       1                       1                       1
##               Guineapig                   Horse                  Jaguar
##                       2                       2                       1
##                Kangaroo Lessershort-tailedshrew          Littlebrownbat
##                       2                       1                       1
##                     Man                 Molerat         Mountainbeaver
##                       1                       1                       1
##                   Mouse                Muskshrew        NAmericanopossum
##                       1                       1                       1
##    Nine-bandedarmadillo                   Okapi               Owlmonkey
##                       1                       2                       1
##             Patasmonkey              Phanlanger                     Pig
##                       2                       1                       2
##                  Rabbit                 Raccoon                     Rat
##                       2                       1                       1
##                  Redfox            Rhesusmonkey      Rockhyrax(Heterob)
##                       1                       1                       1
##  Rockhyrax(Procaviahab)                 Roedeer                   Sheep
##                       1                       2                       2
##               Slowloris           Starnosedmole                  Tenrec
##                       1                       1                       1
##               Treehyrax               Treeshrew                  Vervet
##                       1                       1                       2
##            Wateropossum    Yellow-belliedmarmot
##                       1                       1
##
## Within cluster sum of squares by cluster:
## [1] 231.0878 182.1647
##  (between_SS / total_SS =  32.3 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
fviz_cluster(k2, data = animal_frame)
```




The output of kmeans is a list with several bits of information. The most important being:

- cluster: A vector of integers (from 1:k) indicating the cluster to which each point is allocated.
- centers: A matrix of cluster centers.
- totss: The total sum of squares.
- withinss: Vector of within-cluster sum of squares, one component per cluster.
- tot.withinss: Total within-cluster sum of squares, i.e. sum(withinss).
- betweenss: The between-cluster sum of squares, i.e. $totss - tot.withinss$.
- size: The number of points in each cluster.

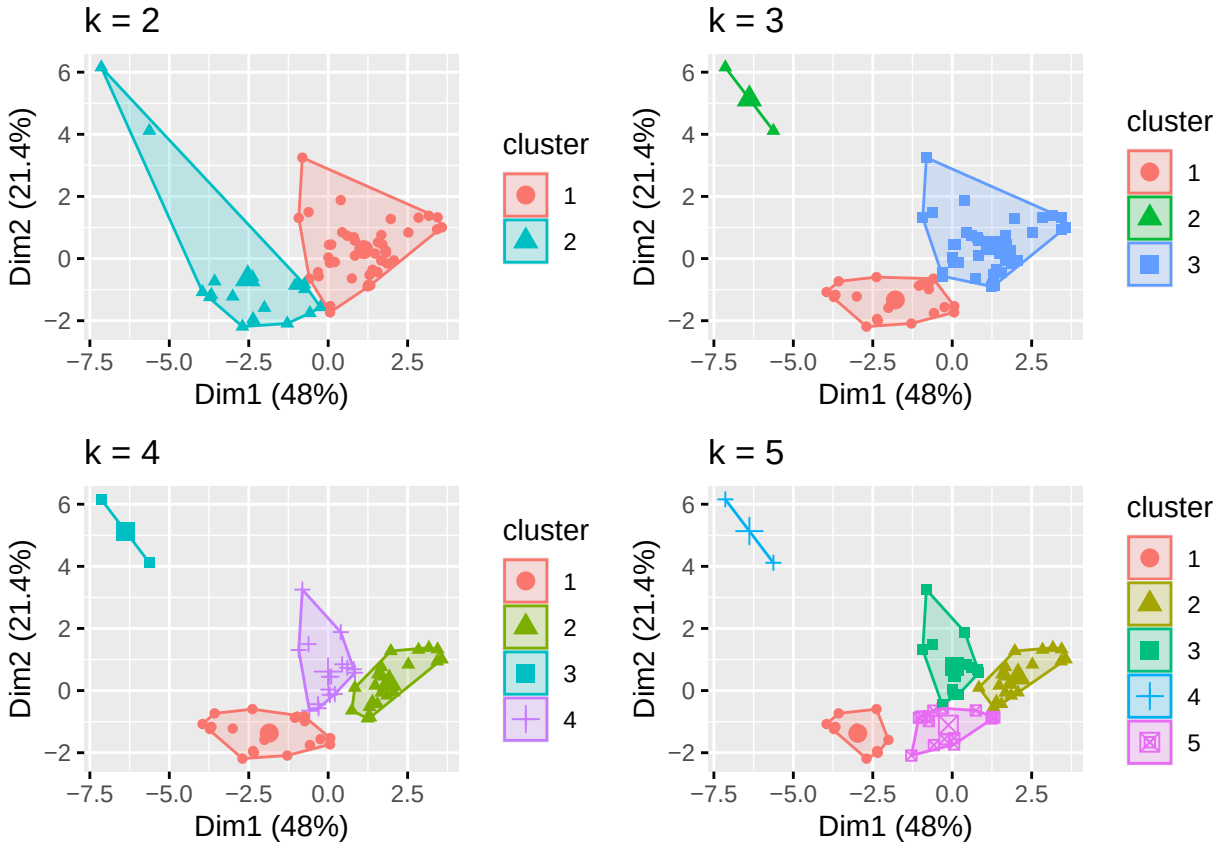
If we print the results we'll see that our groupings resulted in 2 cluster sizes of 19 and 43. We see the cluster centers (means) for the two groups across the nine variables (BodyWt, BrainWt, NonDreaming, Dreaming, TotalSleep, LifeSpan, Gestation, Predation, Exposure). We also get the cluster assignment for each observation (i.e. Rabbit was assigned to cluster 1, Raccoon was assigned to cluster 2, etc.).

```
# Differnet k-clusters
k3 <- kmeans(animal_frame, centers = 3, nstart = 25)
k4 <- kmeans(animal_frame, centers = 4, nstart = 25)
k5 <- kmeans(animal_frame, centers = 5, nstart = 25)

# plots to compare
p1 <- fviz_cluster(k2, geom = "point", data = animal_frame) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point",  data = animal_frame) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point",  data = animal_frame) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point",  data = animal_frame) + ggtitle("k = 5")

library(gridExtra)
```

```
##
```

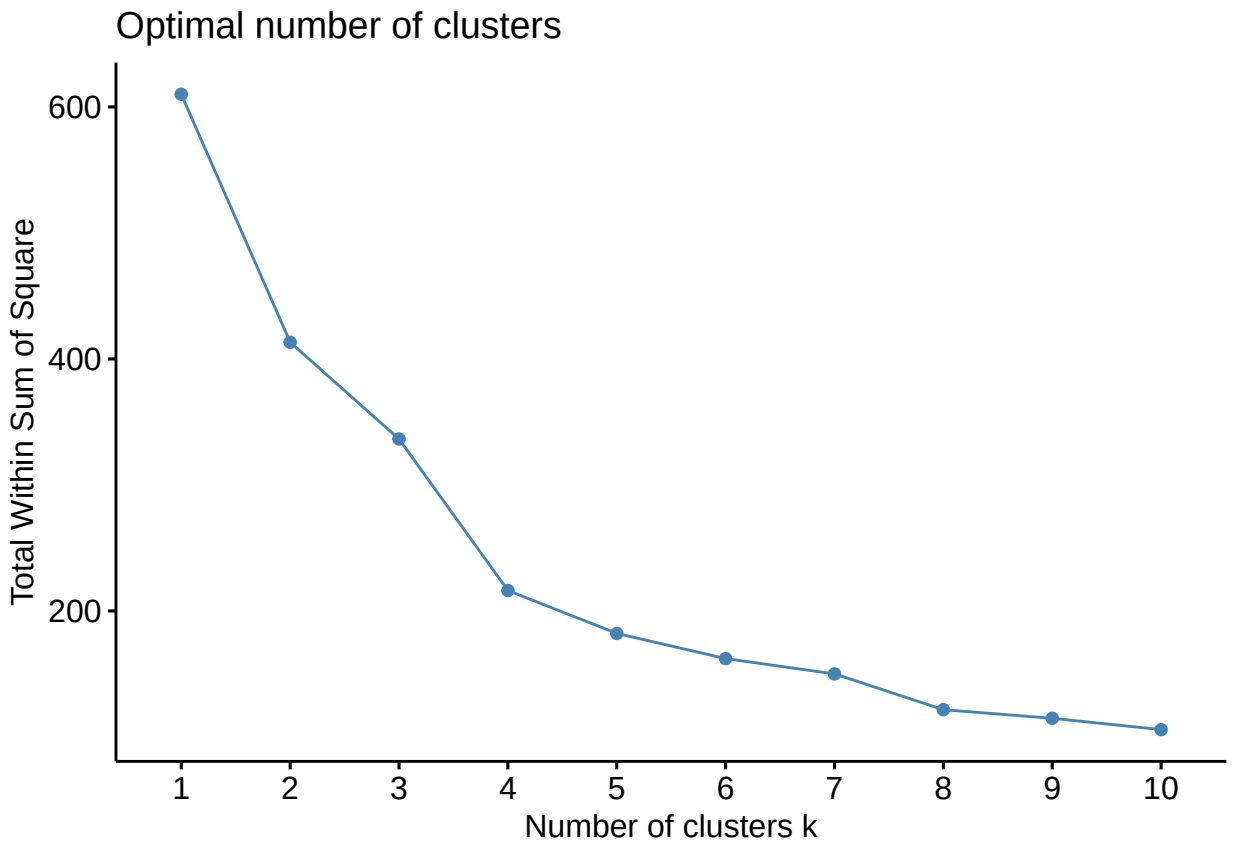
```
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
grid.arrange(p1, p2, p3, p4, nrow = 2)
```

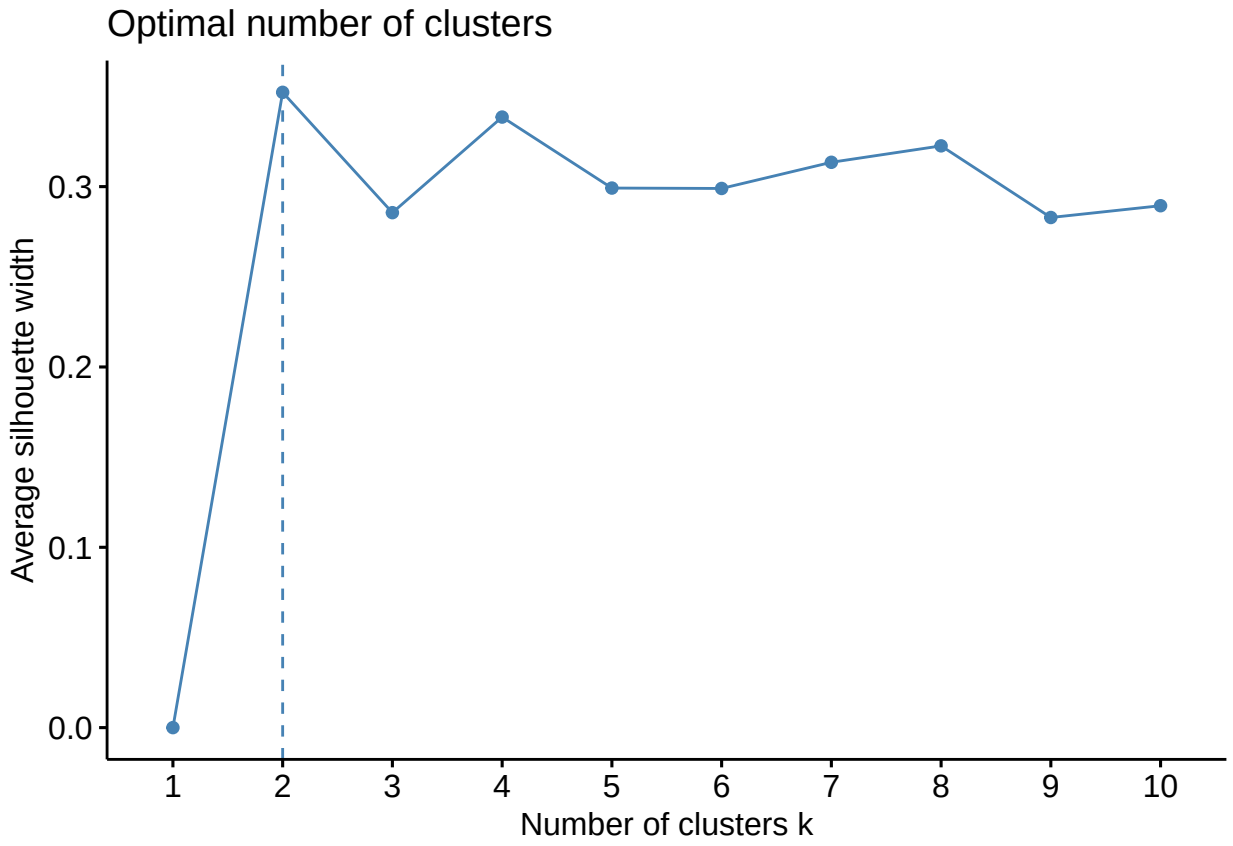



## Model Assessment

```
set.seed(123)

fviz_nbclust(animal_frame, kmeans, method = "wss") # Elbow method
```




```
fviz_nbclust(animal_frame, kmeans, method = "silhouette") # Average Silhouette
```



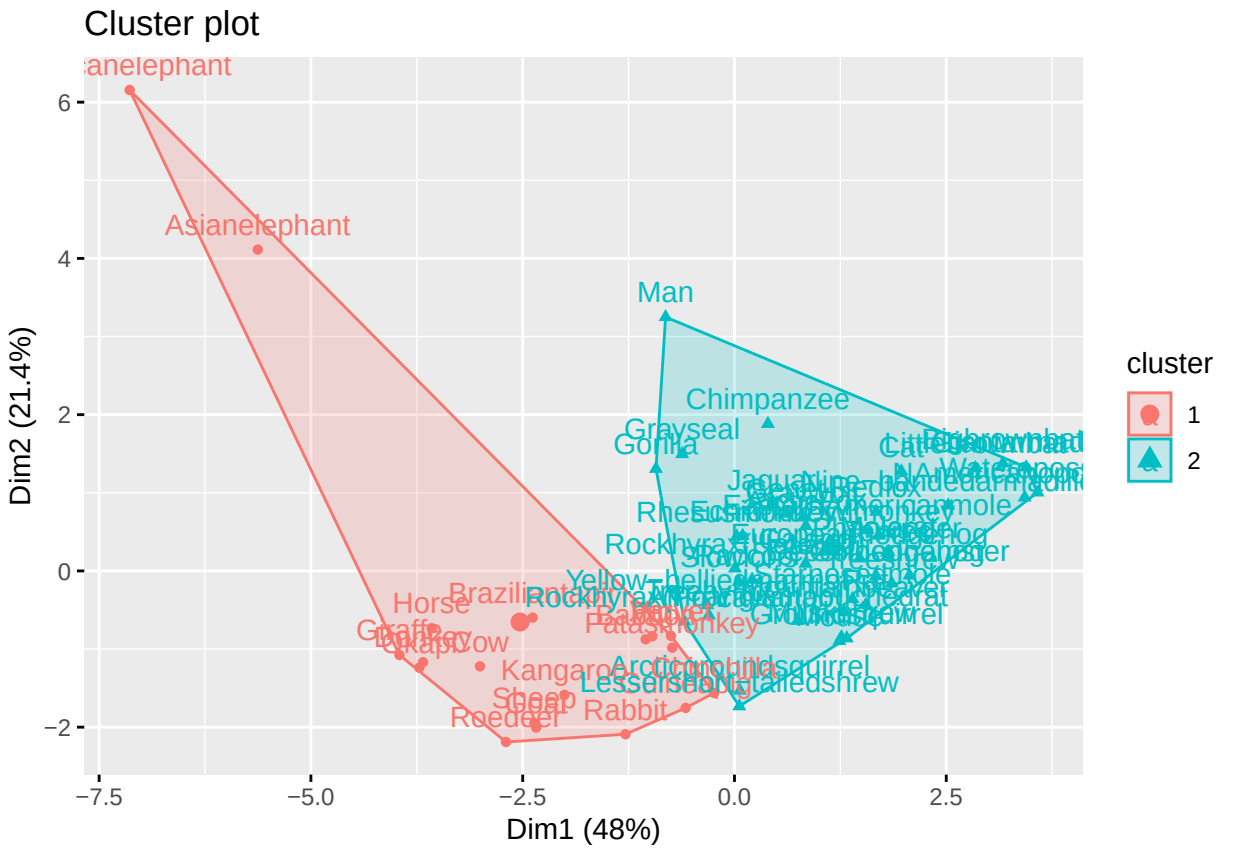

## Prediction and Model accuracy

```
# Compute k-means clustering with k = 2
set.seed(123)
final <- kmeans(animal_frame, 2, nstart = 25)
print(final)
```

```
## K-means clustering with 2 clusters of sizes 19, 43
##
## Cluster means:
##       BodyWt    BrainWt NonDreaming   Dreaming TotalSleep   LifeSpan  Gestation
## 1  0.4618230  0.4962943  -0.5108595 -0.5641489 -0.8689790  0.4504090  0.8353354
## 2 -0.2040613 -0.2192928   0.2257286  0.2492751  0.3839674 -0.1990179 -0.3691017
##    Predation   Exposure     Danger
## 1  1.0498984  1.3457150  1.2545684
## 2 -0.4639086 -0.5946183 -0.5543442
##
## Clustering vector:
##         Africanelephant  Africangiantpouchedrat               ArcticFox
##                       1                       2                       2
##    Arcticgroundsquirrel           Asianelephant                  Baboon
##                       2                       1                       1
##             Bigbrownbat          Braziliantapir                     Cat
##                       2                       1                       2
##              Chimpanzee             Chinchilla                     Cow
##                       2                       1                       1
##          Deserthedgehog                  Donkey     EasternAmericanmole
##                       2                       1                       2
```

```
##                   Echidna         Europeanhedgehog                   Galago
##                         2                        2                        2
##                     Genet           Giantarmadillo                  Giraffe
##                         2                        2                        1
##                      Goat            Goldenhamster                  Gorilla
##                         1                        2                        2
##                  Grayseal                Graywolf           Groundsquirrel
##                         2                        2                        2
##                 Guineapig                   Horse                   Jaguar
##                         1                        1                        2
##                  Kangaroo Lessershort-tailedshrew           Littlebrownbat
##                         1                        2                        2
##                       Man                  Molerat           Mountainbeaver
##                         2                        2                        2
##                     Mouse                Muskshrew         NAmericanopossum
##                         2                        2                        2
##     Nine-bandedarmadillo                    Okapi                 Owlmonkey
##                         2                        1                        2
##              Patasmonkey              Phanlanger                      Pig
##                         1                        2                        1
##                   Rabbit                  Raccoon                      Rat
##                         1                        2                        2
##                   Redfox             Rhesusmonkey      Rockhyrax(Heterob)
##                         2                        2                        2
##   Rockhyrax(Procaviahab)                  Roedeer                    Sheep
##                         2                        1                        1
##                Slowloris            Starnosedmole                   Tenrec
##                         2                        2                        2
##                Treehyrax                Treeshrew                   Vervet
##                         2                        2                        1
##             Wateropossum    Yellow-belliedmarmot
##                         2                        2
##
## Within cluster sum of squares by cluster:
## [1] 182.1647 231.0878
##  (between_SS / total_SS =  32.3 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
fviz_cluster(final, data = animal_frame)
```




```
# K-means clustering with 2 clusters of sizes 19, 43
```

We prefer to use k = 2 (Elbow Method) as that is when we get the total within-cluster sum of square = 32.3% which is better instead of total within-cluster sum of square = 64.5% when k = 4. Hence, we come to the conclusion that 2-means-clustering model works best for our dataset.

## Conclusion

Overall, our group has built a model that takes a data set and divides it into a particular number of subsets optimized such that each of the data points in each subset shares a close proximity to its group members while holding an ideal distance from data points of a different subset, known as clustering. In the absence of a response variable, our model uses k-means clustering, an unsupervised machine learning method, in which the "k" is the number of groups the data should be split into for optimal categorization of clusters; in our model, we found the value from implementing the Elbow and average-silhouette method, and computing the Euclidean distances between clusters and data points from which distance matrix is put together for sustaining dissimilarity between clusters.

We carried out our model construction using a dataset that was collected from observing 62 animals of different species and associating each with a set of attributes: body weight, brain weight, non-dreaming nature, dreaming nature, total sleep time, life span, gestation time, predation index, exposure index, and danger index (1 to 5). Using R code to carry it out, we first installed the relevant R packages required for us to analyze our data and did the required modifications for applicability. As part of the steps of the algorithm, we then randomly chose 10 for value of k and visually depicted our data grouped into ten different clusters with the computed total sum of squares being 83.7% within each of the clusters.

Following this, we now had to find the ideal k value for which the total sum of squares value would be best minimized. After standardizing the data and evaluating the distance matrix, it was easier to make out which animal species (data points) were more similar and dissimilar from one another. We executed the k-means

algorithm again with k = 2 and viewed the results, finding the number of data points in both groups to be 19 and 43, and then repeated the same for k-values of 3, 4 and 5. We ran both the Elbow method and Average-silhouette method on our data, for k values from 1 to 10 and graphically displayed the results for both. We found that the k value, under which the total sum of squares was best minimized and the data was relatively better-clustered, was 2 at a total sum of squares percentage of 32.3%.

## Summary

In summary, our group has implemented a k-means clustering approach to separate 62 animal species into distinct groups of animals with different sleeping characteristics. We utilized all the attributes available in the data-set, and Euclidean distance in the process of grouping the data. We tested out our model with varying values of parameters k, and evaluated them on the total within sum of square metric. Based on the above evaluation, we reached a best model with k = 2, and total within sum of square = 32.3%.

## References

https://uc-r.github.io/kmeans_clustering#replication