```
#Copyright (c) 2018 Jouke Profijt.
#Licensed under GPLv3. See LICENCE
```

```
BirdBones <- read.csv("../data/bird.csv",header = T, sep = ",")
#respective collums for the lenght and diameter
length <- c(2,4,6,8,10)
diameter <- c(3,5,7,9,11)
```

Introduction

Research Question

What bone or group of bones that most birds have in common, is the most significant for the function in the diffrent ecological groups?

Data

```
Data recieved from:
```

Birds' Bones and Living Habits, Kaggle dataset

```
Bone measurements were measured from a skeleton collection of
Natural History Museum of Los Angeles County,
provided by Dr. D. Liu of beijing Museaum of Natural History
```

Exploratory Data Analyses

The data contains 420 bird samples where the bone lengths and diameters have been measured. The birds are separated in 6 diffrent groups:

- Swimming Birds, SW
- Wading Birds, W
- Terrestrial Birds, T
- Raptors, R
- Scansorial Birds, P
- Singing Birds, SO

Most samples have data for:

- Length and Diameter of the Humerus
- Length and Diameter of the Ulna
- Length and Diameter of the Femur
- Length and Diameter of the Tibiotarsus
- Length and Diameter of the Taesometatarsus

I'm creating a graph which displays the bonelengths on y axis an the Id on x colorcoded by their ecological group. by evaluating this we can see if some groups have overall larger or smaller bones and we see if there are big outliers.

```
# this omits several ggplot2 errors retaining to mising values
BirdBones.noNA <- BirdBones[complete.cases(BirdBones),]
```

```
# Displaing the data frame structure and a small summary
str(BirdBones)
```

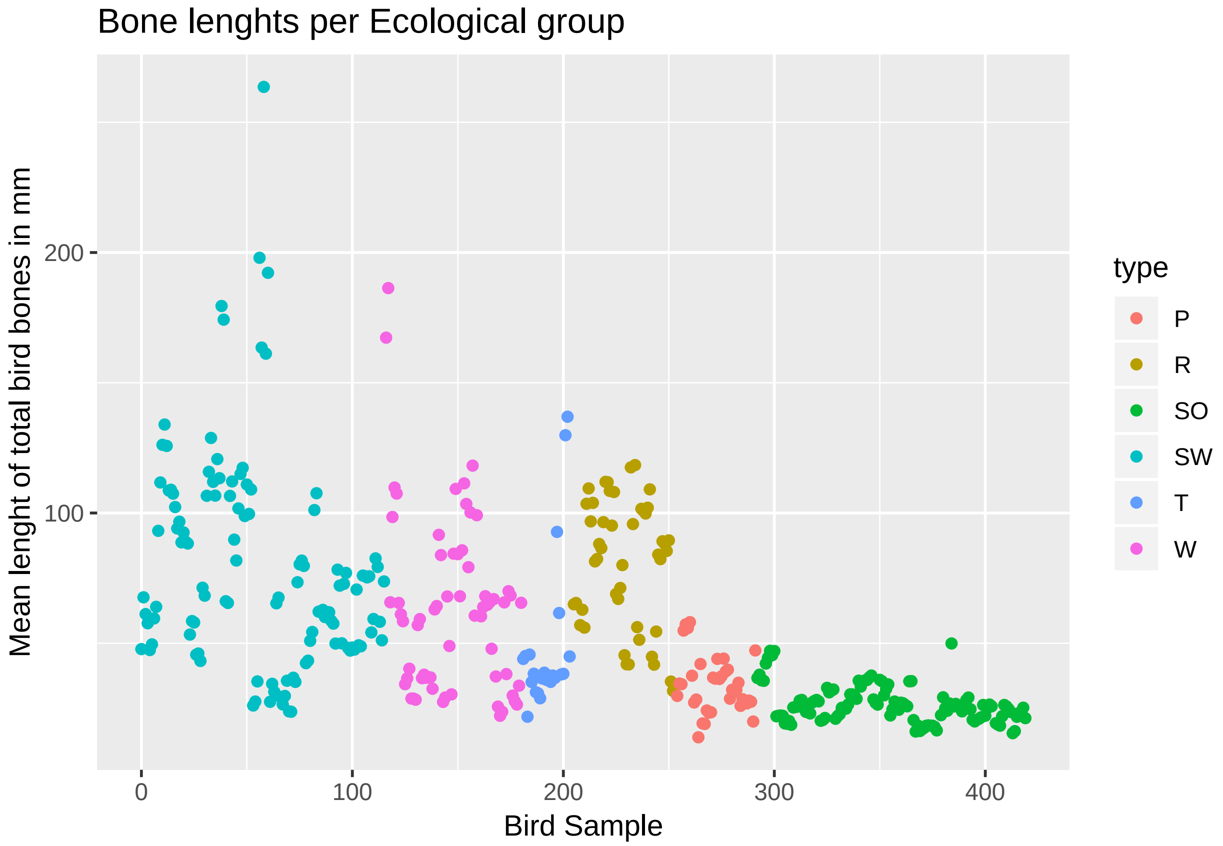
```
## 'data.frame':    420 obs. of  12 variables:
##  $ id   : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ huml : num  80.8 88.9 80 77.7 62.8 ...
##  $ humw : num  6.68 6.63 6.37 5.7 4.84 ...
##  $ ulnal: num  72 80.5 69.3 65.8 52.1 ...
##  $ ulnaw: num  4.88 5.59 5.28 4.77 3.73 3.47 4.5 4.55 6.13 7.05 ...
##  $ feml : num  41.8 47 43.1 40 34 ...
##  $ femw : num  3.7 4.3 3.9 3.52 2.72 4.41 3.41 3.78 5.45 7.44 ...
##  $ tibl : num  5.5 80.2 75.3 69.2 56.3 ...
##  $ tibw : num  4.03 4.51 4.04 3.4 2.96 2.73 3.56 3.81 5.58 7.31 ...
##  $ tarl : num  38.7 41.5 38.3 35.8 31.9 ...
##  $ tarw : num  3.84 4.01 3.34 3.41 3.13 2.83 3.64 3.81 4.37 6.34 ...
##  $ type : Factor w/ 6 levels "P","R","SO","SW",..: 4 4 4 4 4 4 4 4 4 4 ...
```

```
summary(BirdBones)
```

```
##        id             huml             humw            ulnal       
##  Min.   :  0.0   Min.   :  9.85   Min.   : 1.140   Min.   : 14.09  
##  1st Qu.:104.8   1st Qu.: 25.17   1st Qu.: 2.190   1st Qu.: 28.05  
##  Median :209.5   Median : 44.18   Median : 3.500   Median : 43.71  
##  Mean   :209.5   Mean   : 64.65   Mean   : 4.371   Mean   : 69.12  
##  3rd Qu.:314.2   3rd Qu.: 90.31   3rd Qu.: 5.810   3rd Qu.: 97.52  
##  Max.   :419.0   Max.   :420.00   Max.   :17.840   Max.   :422.00  
##                  NA's   :1        NA's   :1        NA's   :3       
##      ulnaw             feml             femw             tibl       
##  Min.   : 1.000   Min.   : 11.83   Min.   : 0.930   Min.   :  5.50  
##  1st Qu.: 1.870   1st Qu.: 21.30   1st Qu.: 1.715   1st Qu.: 36.42  
##  Median : 2.945   Median : 31.13   Median : 2.520   Median : 52.12  
##  Mean   : 3.597   Mean   : 36.87   Mean   : 3.221   Mean   : 64.66  
##  3rd Qu.: 4.770   3rd Qu.: 47.12   3rd Qu.: 4.135   3rd Qu.: 82.87  
##  Max.   :12.000   Max.   :117.07   Max.   :11.640   Max.   :240.00  
##  NA's   :2        NA's   :2        NA's   :1        NA's   :2       
##       tibw             tarl             tarw        type   
##  Min.   : 0.870   Min.   :  7.77   Min.   : 0.660   P : 38  
##  1st Qu.: 1.565   1st Qu.: 23.04   1st Qu.: 1.425   R : 50  
##  Median : 2.490   Median : 31.74   Median : 2.230   SO:128  
##  Mean   : 3.182   Mean   : 39.23   Mean   : 2.930   SW:116  
##  3rd Qu.: 4.255   3rd Qu.: 50.25   3rd Qu.: 3.500   T : 23  
##  Max.   :11.030   Max.   :175.00   Max.   :14.090   W : 65  
##  NA's   :1        NA's   :1        NA's   :1                
```

there are 420 total measurements, and by using complete cases i found that there are 413 measurements which are complete and do not contain missing values, aka > there are 7 measurements that contain missing values.

```
library(ggplot2)
library(reshape)
source("../scripts/BoneMeans.R")
BirdBones.noNA <- BoneMeans(data = BirdBones.noNA, length = length, diameter = diameter)
ggplot(data = BirdBones.noNA, aes(id, length.mean, colour = type)) +
  ggtitle("Bone lenghts per Ecological group")+
  ylab("Mean lenght of total bird bones in mm") +
  xlab("Bird Sample")+
  geom_point()
```

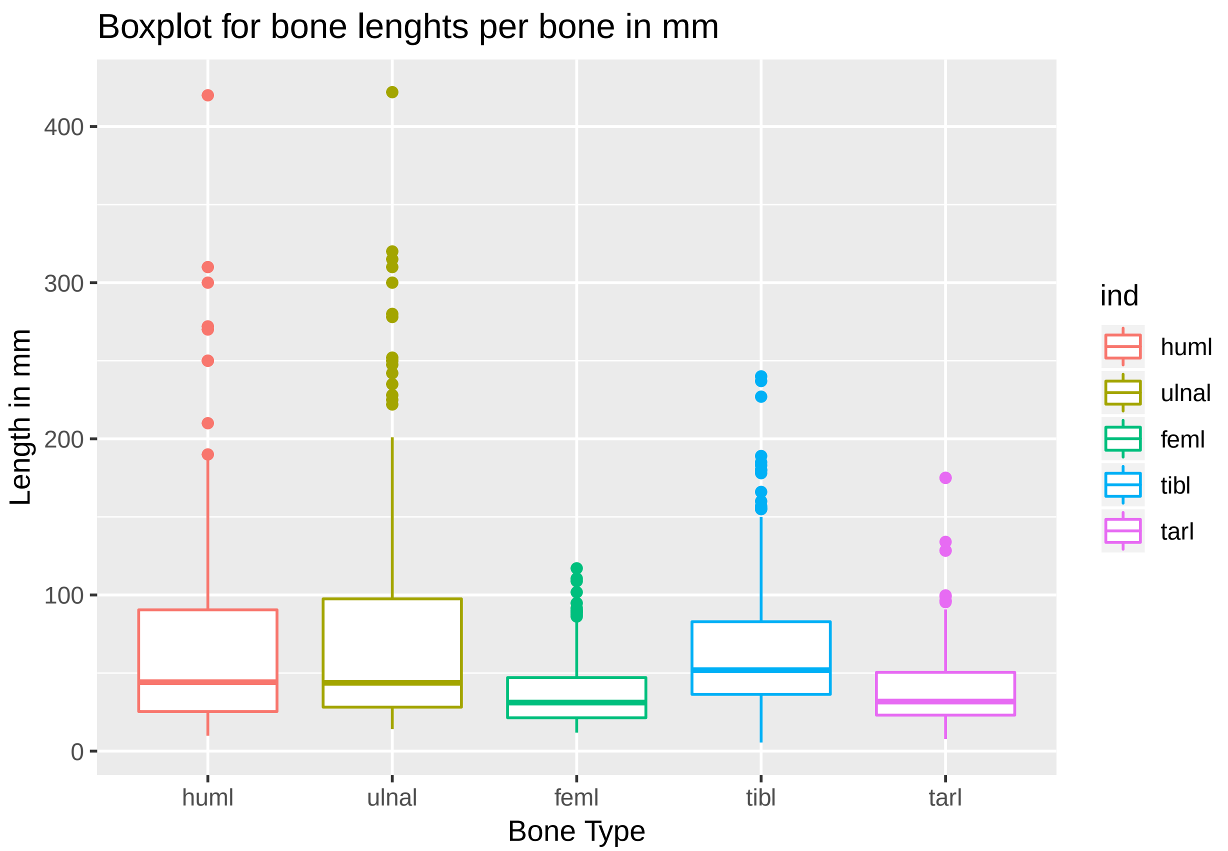


As seen above swimming birds have the biggest bones, but also shown is that there are a lot more samples in that group where there is a lot of variation. I can look into cleaning up the data and removing the biggest outliers in this group. Singing birds also have a lot of samples but there is much less variation and so more certanty.

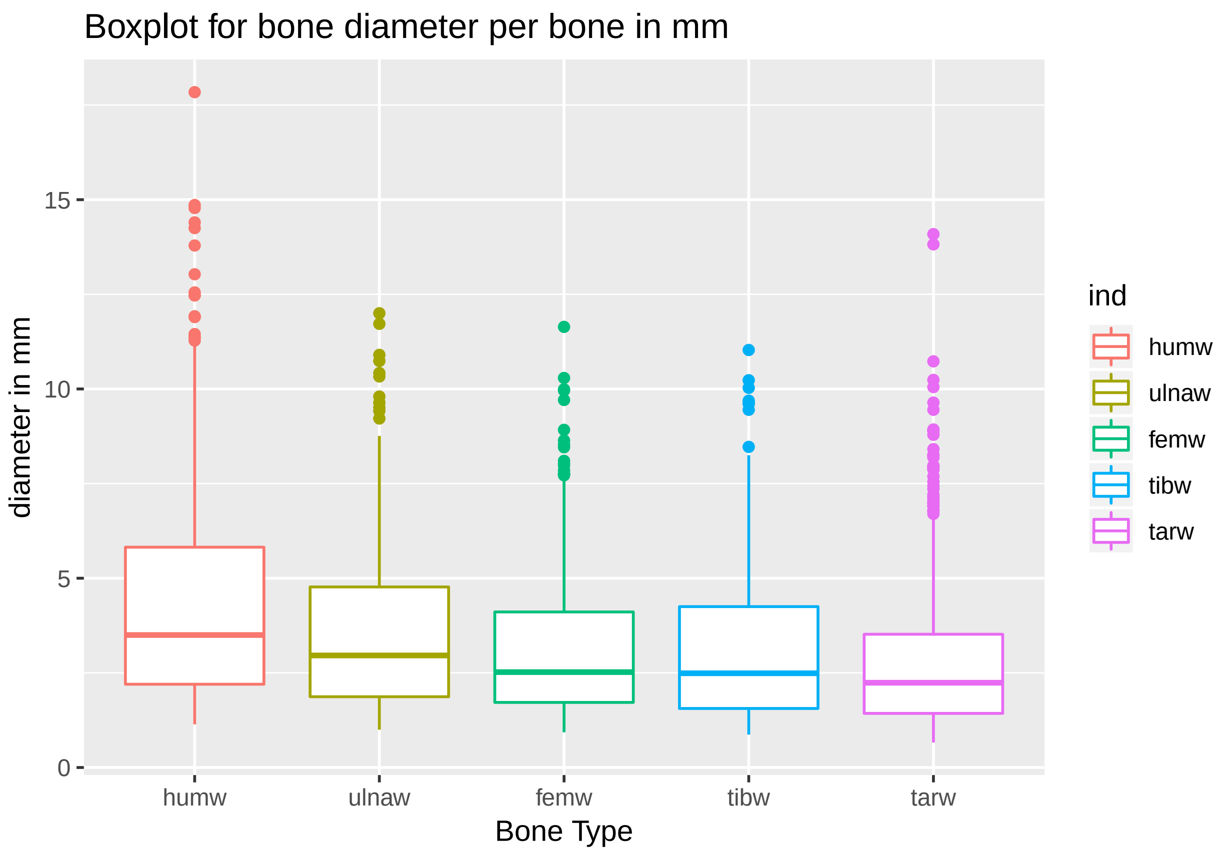
For the rest of the birds there are not a lot of sample so maby we could try and normalizing the data so there is an even amount of samples per group.

There are also 7 samples that contain missing values, we could just straight out not use these samples becouse 4 of these are part of the biggest group of samples. and the others are not part of the smallest groups.

```
library(ggplot2)

ggplot(stack(BirdBones.noNA[length]), aes(x = ind, y = values, color = ind)) +
  geom_boxplot()+
  ggtitle("Boxplot for bone lenghts per bone in mm")+
  xlab("Bone Type")+
  ylab("Length in mm")
```



```
ggplot(stack(BirdBones.noNA[diameter]), aes(x = ind, y = values, color = ind)) +
  geom_boxplot()+
  ggtitle("Boxplot for bone diameter per bone in mm")+
  xlab("Bone Type")+
  ylab("diameter in mm")
```

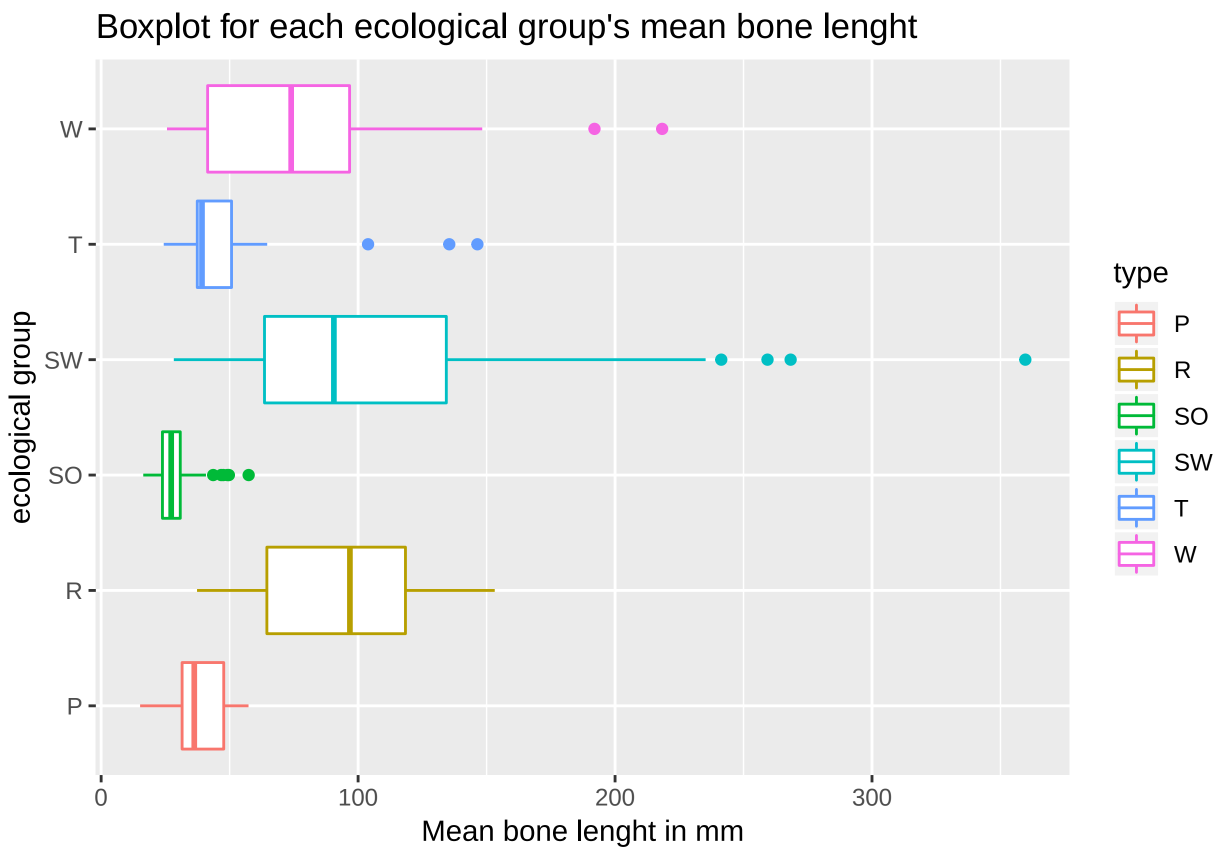


What we see above is that there are a considerable amount of outliers between the bones themselves, but this was expected as they are from diffrent groups and the diffrent groups dont have the same amount of measurements. below i will do a comparison between the group bone mean lengths which will show outliers in their respective group. using the above boxplots we can maby see which bones are not very important > see if they don't differ at all wich means we dont need them that much for classification.

As we can see the femur lenght and taesometatarsus length do not contain a lot of variation and maby are candidates for exclution from analysis.

```r
# diameter & lenght indexes for only the longer bones.
length.long <- c(2, 4, 8)
diameter.long <- c(3, 5, 9)
BirdBones.noNA.long <- BoneMeans(BirdBones.noNA, length.long, diameter.long)
```

```r
library(ggplot2)

ggplot(BirdBones.noNA.long, aes(x = type, y = length.mean, color = type)) +
  geom_boxplot()+
  coord_flip()+
  ggtitle("Boxplot for each ecological group's mean bone lenght")+
  ylab("Mean bone lenght in mm")+
  xlab("ecological group")
```

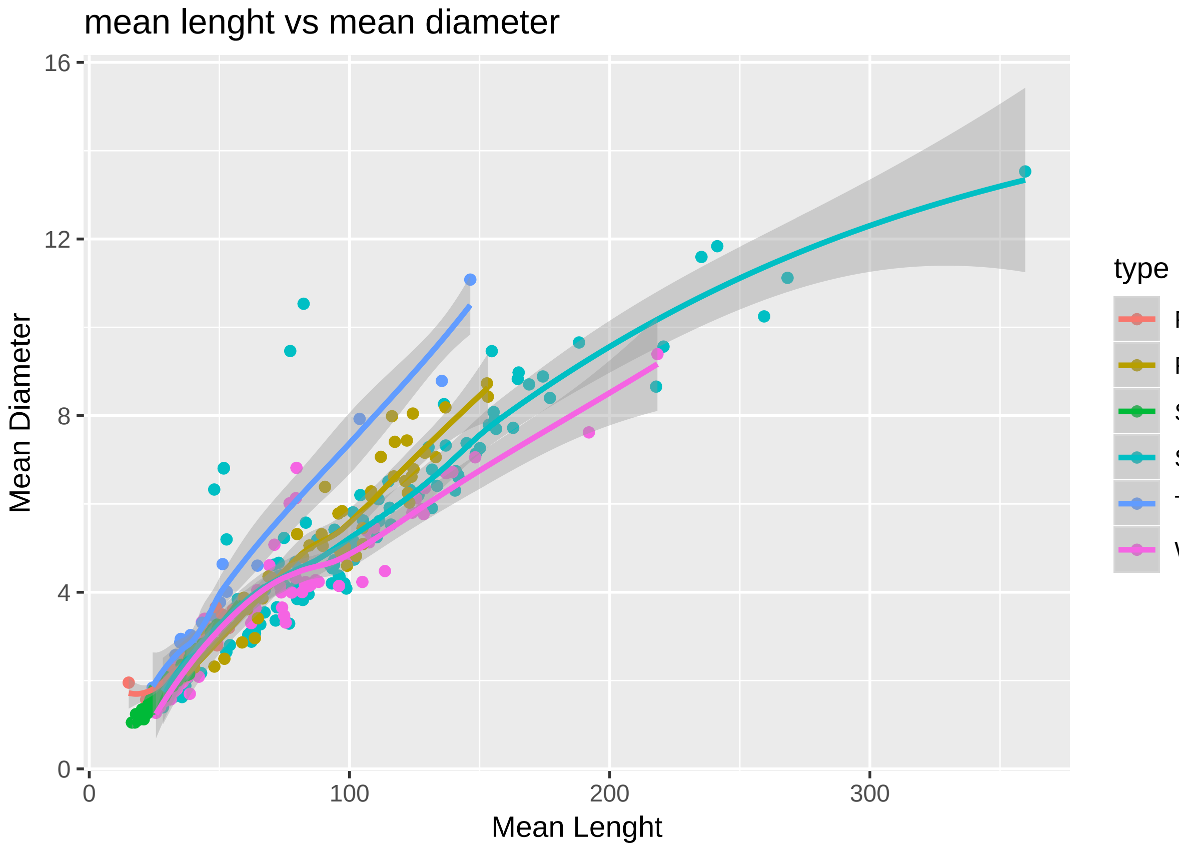


As you can see there are quite a few outliers in all groups except in group R, The raptors. but we saw in the above boxplot that there were loads of outliers between all bones, yet here that is significantly reduced. so if we are going to inspect the date we have to look at them per group and NOT by bone type.

What we can also see in these plots are which birds are most likely the largest, as seen above color cyan or SW or Swimming Birds are the biggest of them all closely followed by W or Wading Birds

```
ggplot(BirdBones.noNA.long,aes(x=length.mean,y=diameter.mean,color=type))+
  geom_point()+
  geom_smooth()+
  ggtitle("mean lenght vs mean diameter")+
  xlab("Mean Lenght")+
  ylab("Mean Diameter")
```

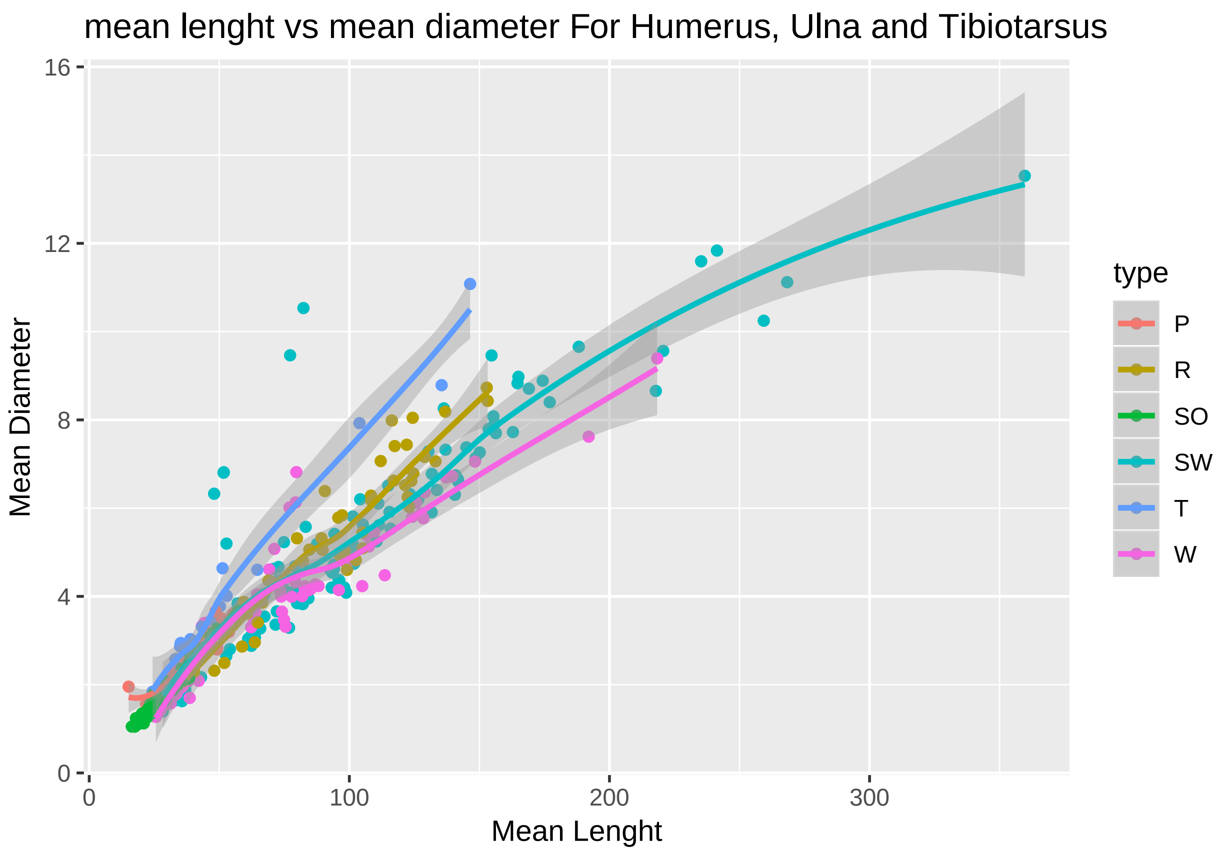
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



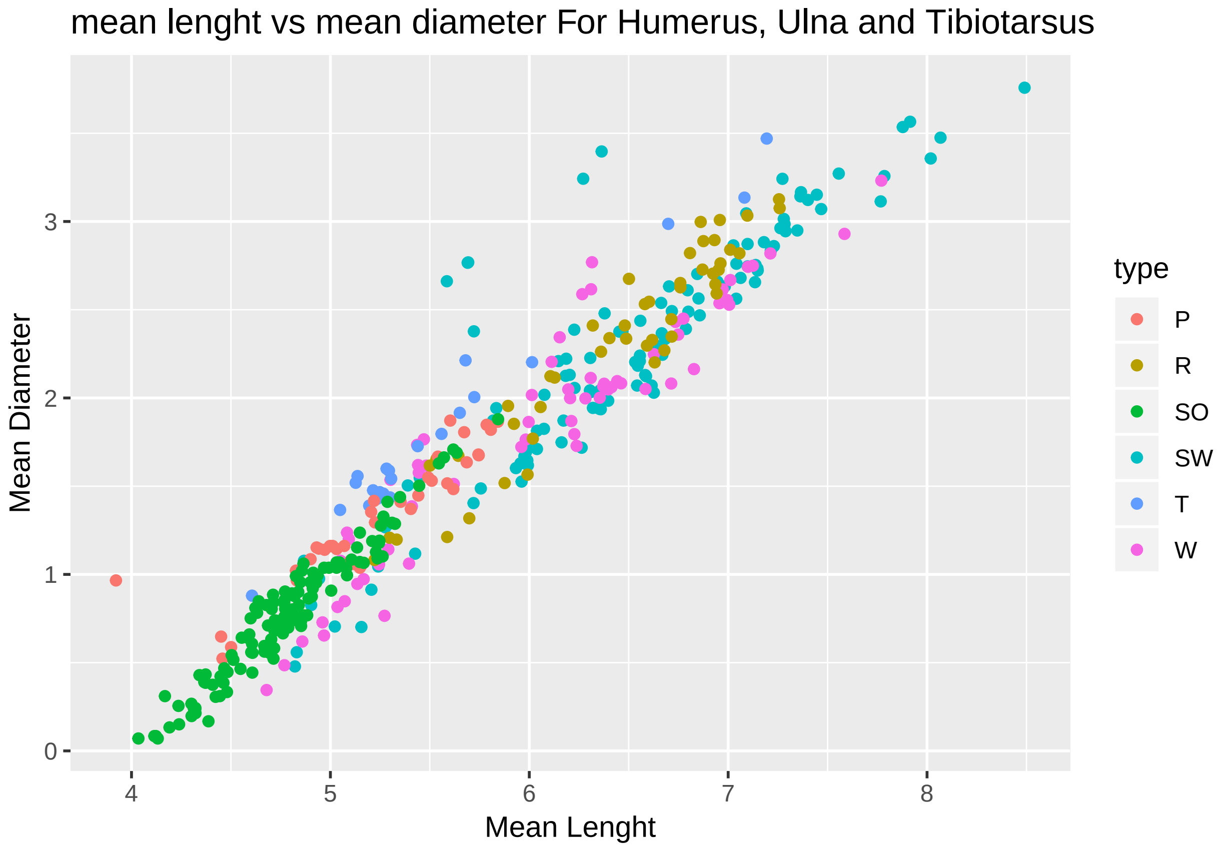
Untransformed datapoints separated by goup, again here we can see which birds are the biggest, but for smaller birds this plot is not very readable. we do see something odd, where T has a climbing line around lenght 50, other birds have a decreasing line. also Swimming Birds have some results that are very diffrent form their mean line.

```
ggplot(BirdBones.noNA.long,aes(x=length.mean,y=diameter.mean,color=type))+
  geom_point()+
  geom_smooth()+
  ggtitle("mean lenght vs mean diameter For Humerus, Ulna and Tibiotarsus")+
  xlab("Mean Lenght")+
  ylab("Mean Diameter")
```

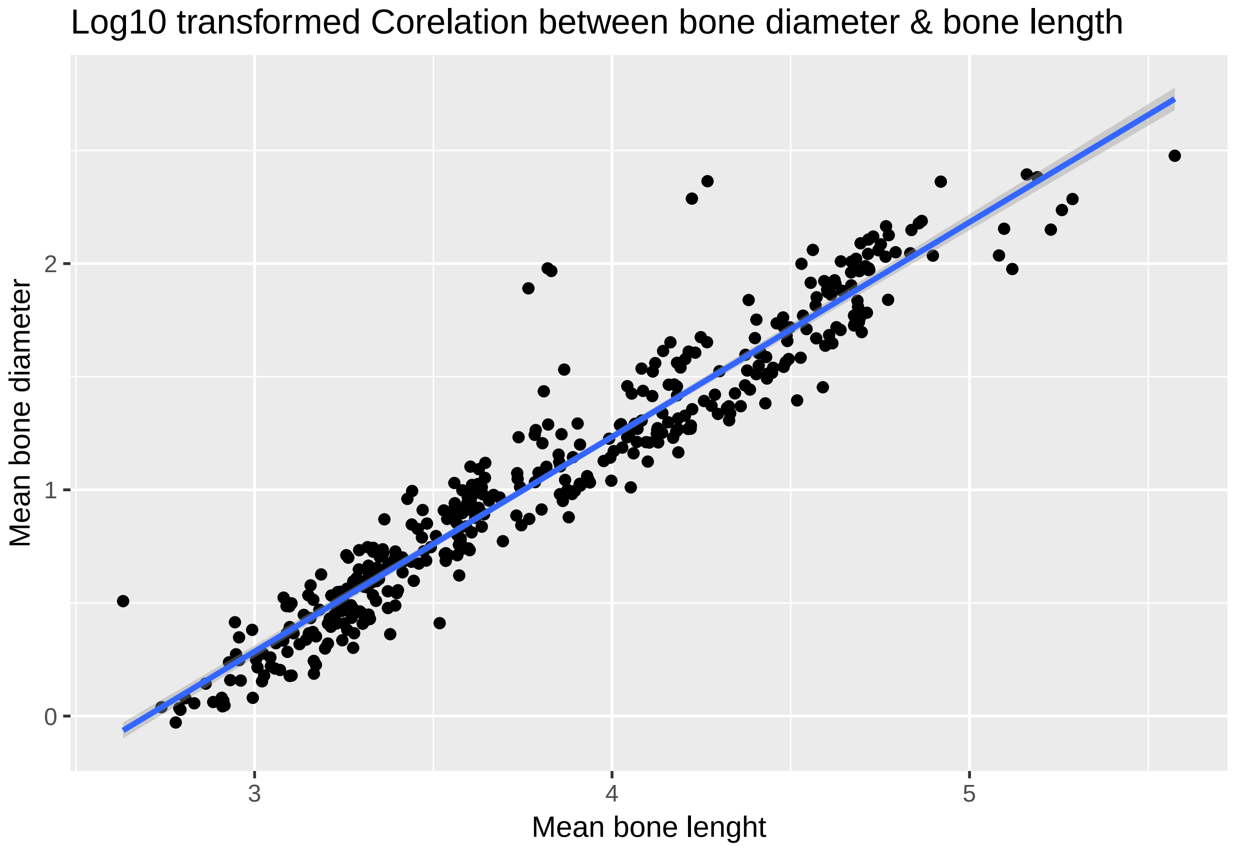
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
BirdBones.noNA.long$log2length <- log2(BirdBones.noNA.long$length.mean)
BirdBones.noNA.long$log2diameter <- log2(BirdBones.noNA.long$diameter.mean)

ggplot(BirdBones.noNA.long,aes(x=log2length,y=log2diameter,color=type))+
  geom_point()+
  ggtitle("mean lenght vs mean diameter For Humerus, Ulna and Tibiotarsus")+
  xlab("Mean Lenght")+
  ylab("Mean Diameter")
```



```r
library(ggplot2)
ggplot(BirdBones.noNA, aes(x = log(length.mean), y = log(diameter.mean))) +
  geom_point()+
  geom_smooth(method = lm)+
  ggtitle("Log10 transformed Corelation between bone diameter & bone length")+
  xlab("Mean bone lenght")+
  ylab("Mean bone diameter")
```

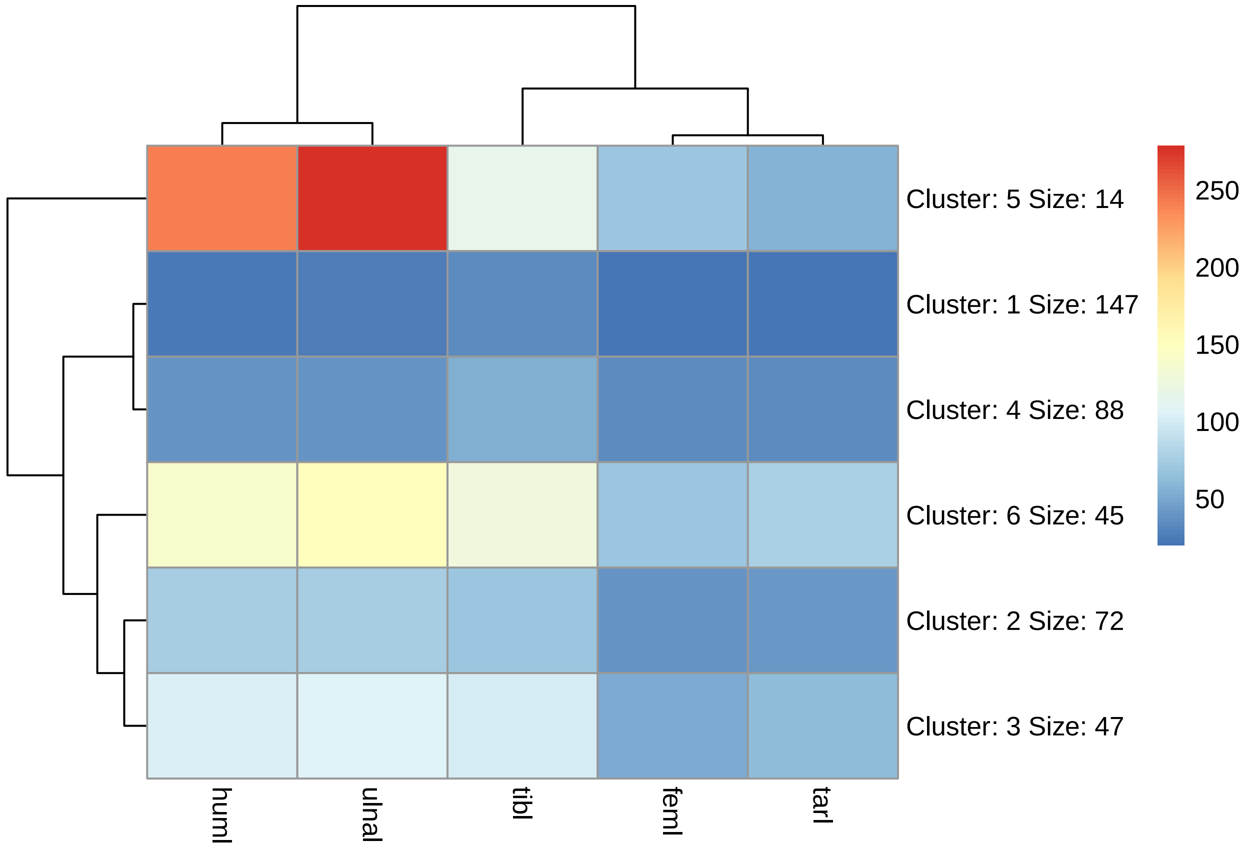


As expected there is a coralation between the bone lenght and bone diameter, you can see this because the plot gives a liniar line. it does make a lot of sense if you have longer bones there you will most likely also have thicker bones(bigger diameters)

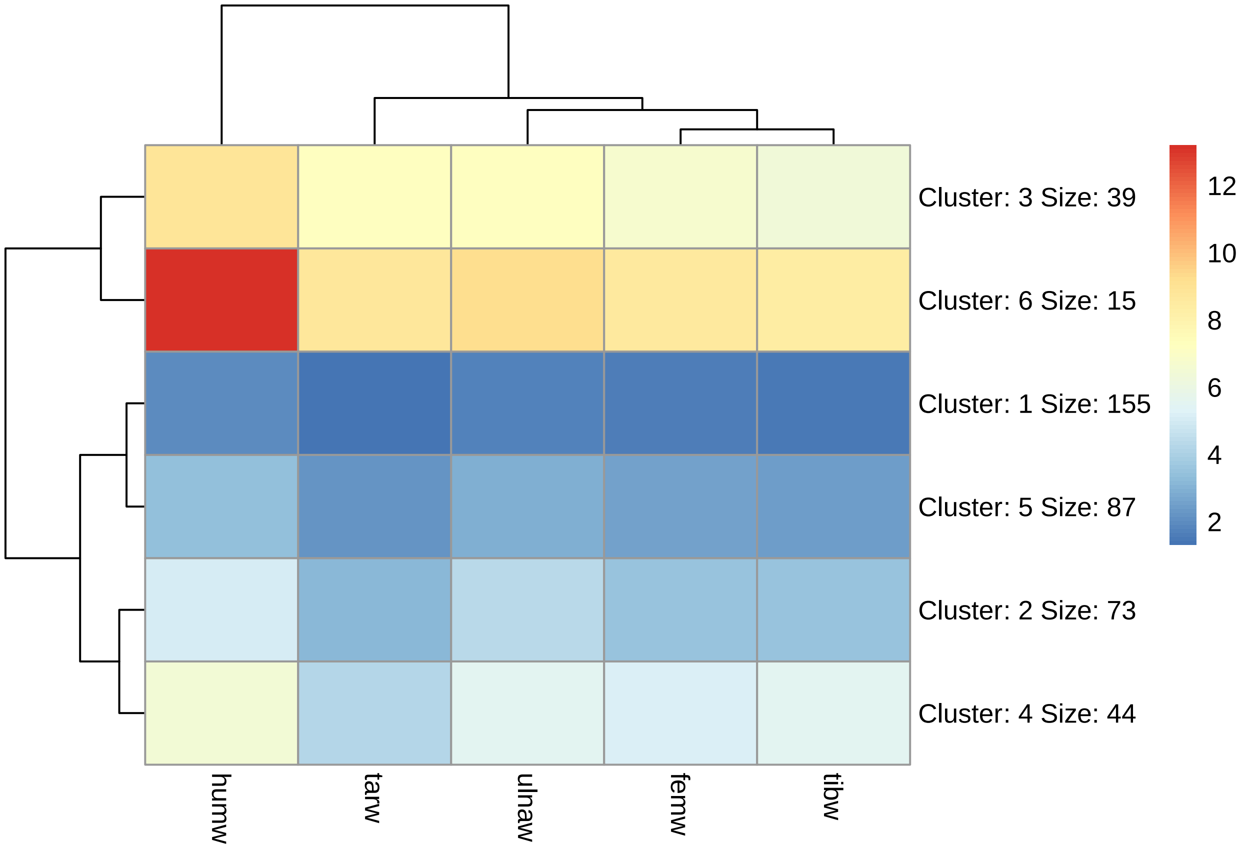
```
# m <- as.matrix(BirdBones.noNA$length.mean, ncol=2)
# 6 groups so 6 clusters is assumed
# cl <- kmeans(m, 6)
#
# ```
# ```{r}
# BirdBones.noNA$cluster <- factor(cl$cluster)
# centers <- as.data.frame((cl$centers))
# ```
# ```{r}
# library(ggplot2)
#
#
# ggplot(data=BirdBones.noNA, aes(x=length.me43an, y=id, color=type )) +
#  geom_point() +
#  geom_point(data=centers, aes(x=V1,y=V2, color='Center')) +
#  geom_point(data=centers, aes(x=V1,y=V2, color='Center'), size=50, alpha=.4, legend=FALSE)
```

```
library(ggplot2)
library(pheatmap)
df.hum <- data.frame(BirdBones.noNA$huml, BirdBones.noNA$humw)
kmeans.hum <- kmeans((df.hum), 6)
```

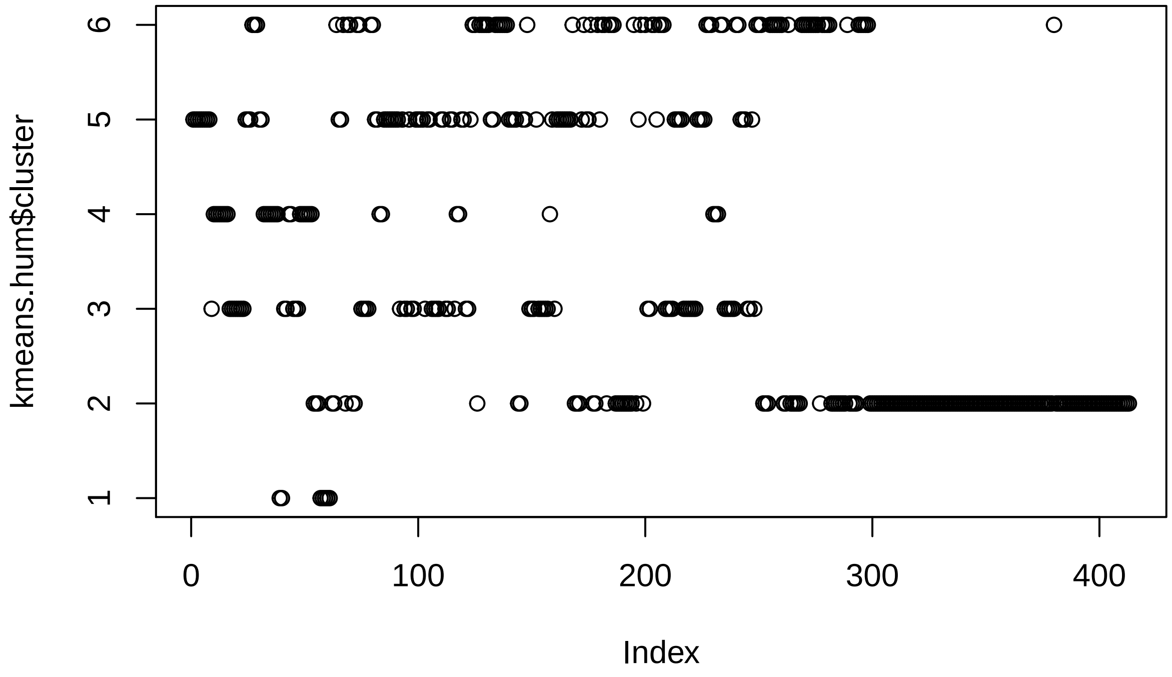
```r
dm.len <- data.matrix(BirdBones.noNA[length])
dm.dia <- data.matrix(BirdBones.noNA[diameter])

pheatmap(dm.len, kmeans_k = 6)
```



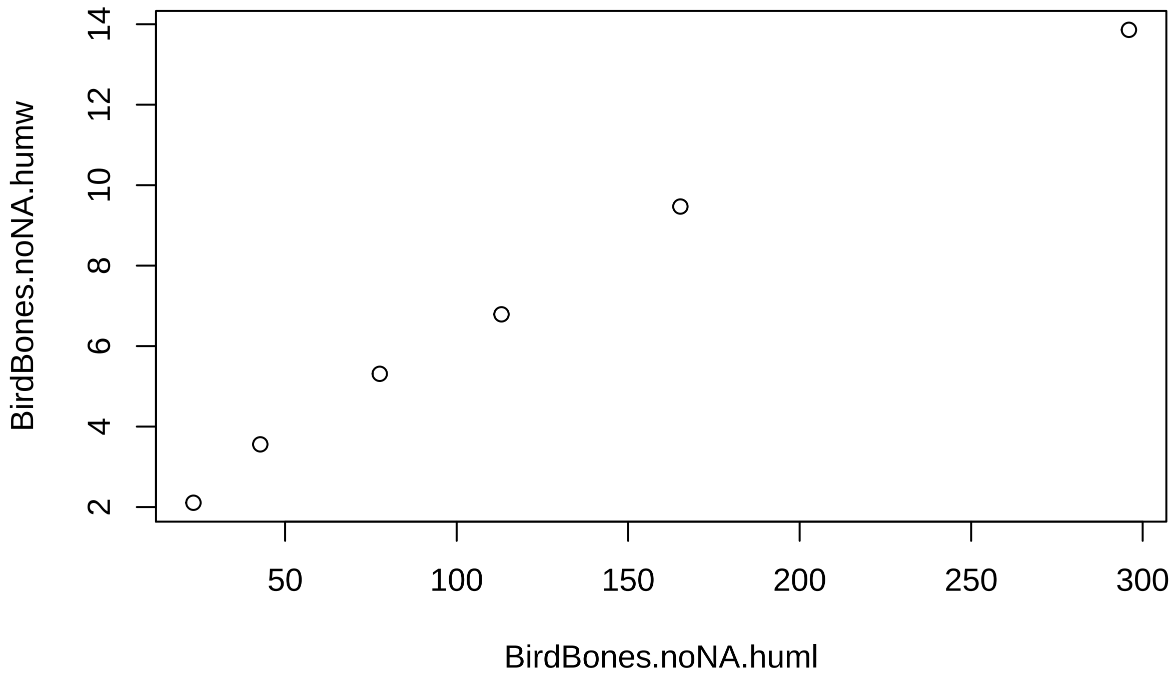
```r
pheatmap(dm.dia, kmeans_k = 6)
```



```
plot(kmeans.hum$cluster)
```



```
plot(kmeans.hum$centers)
```


BirdBones.noNA.humw
2
4
6
8
10
12
14
50
100
150
200
250
300
BirdBones.noNA.huml