Final Assessment. Dataset Analysis

Pau Blasco Roca. NIU: 1600959

28/5/2022

A statistical approach to the study of mammal sleep patterns in relation to body and brain weight

The intent of this paper is to study the relationship between the physical qualities of different species of mammals and their respective sleep patterns. These studies consist on an initial linear regression model, followed by a bootstrap approach. After that, further correlation and bootstrap studies are performed by separating mammals in several categories.

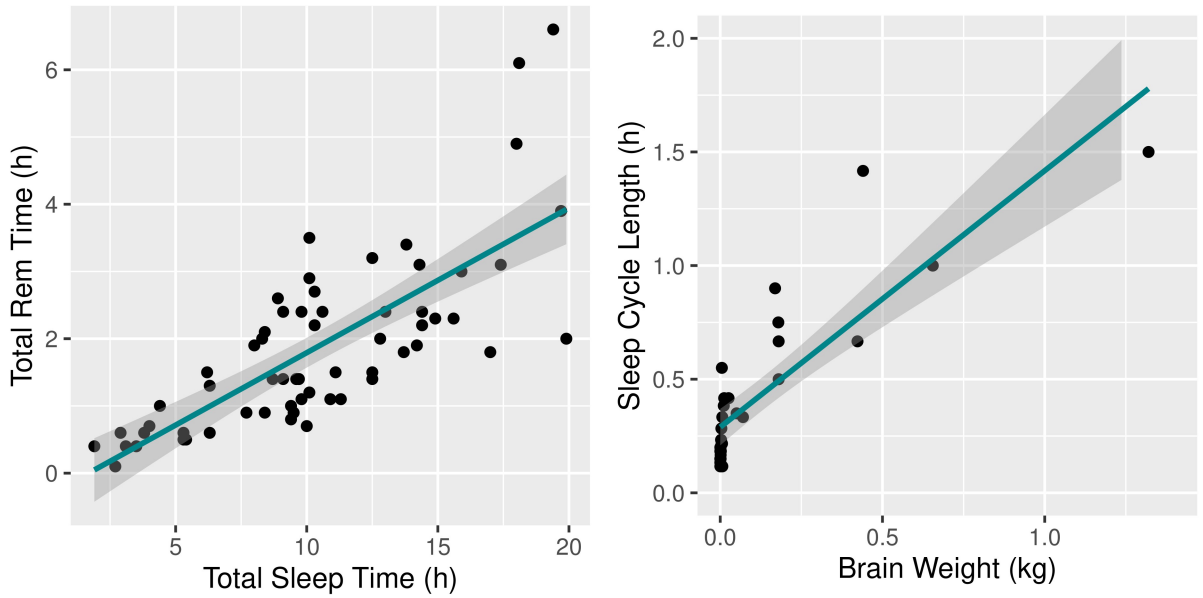
```
#Initial Linear Regression fittings for some of our data variables.
myData = msleep
myData
```

```
## # A tibble: 83 x 11
##    name   genus vore  order conservation sleep_total sleep_rem sleep_cycle awake
##    <chr>  <chr> <chr> <chr> <chr>              <dbl>     <dbl>       <dbl> <dbl>
##  1 Cheet~ Acin~ carni Carn~ lc                  12.1      NA        NA      11.9
##  2 Owl m~ Aotus omni  Prim~ <NA>                17         1.8      NA       7
##  3 Mount~ Aplo~ herbi Rode~ nt                  14.4       2.4      NA       9.6
##  4 Great~ Blar~ omni  Sori~ lc                  14.9       2.3       0.133   9.1
##  5 Cow    Bos   herbi Arti~ domesticated         4         0.7       0.667  20
##  6 Three~ Brad~ herbi Pilo~ <NA>                14.4       2.2       0.767   9.6
##  7 North~ Call~ carni Carn~ vu                   8.7       1.4       0.383  15.3
##  8 Vespe~ Calo~ <NA>  Rode~ <NA>                 7        NA        NA      17
##  9 Dog    Canis carni Carn~ domesticated        10.1       2.9       0.333  13.9
## 10 Roe d~ Capr~ herbi Arti~ lc                   3        NA        NA      21
## # ... with 73 more rows, and 2 more variables: brainwt <dbl>, bodywt <dbl>
```

```
par(mfrow=c(1,2))
P1<-ggplot(myData, aes(x=sleep_total, y=sleep_rem)) +
  geom_point() +
  labs(x='Total Sleep Time (h)', y='Total Rem Time (h)') +
  stat_smooth(method='lm', color = "turquoise4")+
  theme(plot.title = element_text(hjust=0.5, size=15, face='bold'), aspect.ratio = 1)
P2<-ggplot(myData, aes(x=brainwt, y=sleep_cycle)) +
  geom_point() +
  labs(x='Brain Weight (kg)', y='Sleep Cycle Length (h)') +
  ylim(0,2)+
  xlim(0,1.4)+
  stat_smooth(method='lm', color = "turquoise4")+
  theme(plot.title = element_text(hjust=0.5, size=15, face='bold'), aspect.ratio = 1)
P<-plot_grid(P1, P2)
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 22 rows containing non-finite values (stat_smooth).

## Warning: Removed 22 rows containing missing values (geom_point).

## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 53 rows containing non-finite values (stat_smooth).

## Warning: Removed 53 rows containing missing values (geom_point).
```

```
P
```




```
linearMod1 <- lm(sleep_total ~ sleep_rem, data=myData)  # build linear regression model on full data
summary(linearMod1)
```

```
##
## Call:
## lm(formula = sleep_total ~ sleep_rem, data = myData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.6159 -2.5782 -0.3274  2.5468  9.1845
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4660     0.6819   8.016 5.15e-11 ***
## sleep_rem     2.6248     0.2998   8.756 2.92e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.014 on 59 degrees of freedom
##   (22 observations deleted due to missingness)
## Multiple R-squared:  0.5651, Adjusted R-squared:  0.5578
## F-statistic: 76.67 on 1 and 59 DF,  p-value: 2.916e-12
```

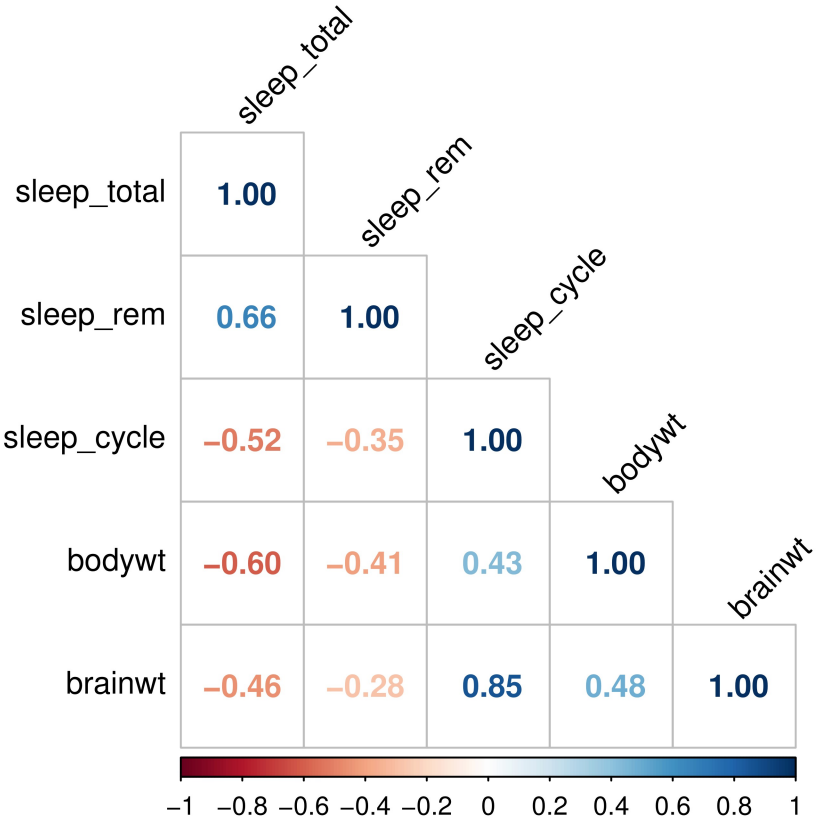
```
linearMod2 <- lm(brainwt ~ sleep_cycle, data=myData)  # build linear regression model on full data
summary(linearMod2)
```

```
##
## Call:
## lm(formula = brainwt ~ sleep_cycle, data = myData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31902 -0.07711  0.01706  0.05289  0.50741
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.15158    0.04193  -3.615  0.00117 **
## sleep_cycle  0.64278    0.07477   8.597 2.42e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1472 on 28 degrees of freedom
##   (53 observations deleted due to missingness)
## Multiple R-squared:  0.7253, Adjusted R-squared:  0.7154
## F-statistic: 73.91 on 1 and 28 DF,  p-value: 2.42e-09
```

```
#Studying only numeric variables, without distinguishing by mammal category
myData = msleep %>% select(sleep_total,sleep_rem,sleep_cycle,bodywt,brainwt)
res <- cor(myData, use = "complete.obs")
round(res, 2)
```

```
##             sleep_total sleep_rem sleep_cycle bodywt brainwt
## sleep_total        1.00      0.66       -0.52  -0.60   -0.46
## sleep_rem          0.66      1.00       -0.35  -0.41   -0.28
## sleep_cycle       -0.52     -0.35        1.00   0.43    0.85
## bodywt            -0.60     -0.41        0.43   1.00    0.48
## brainwt           -0.46     -0.28        0.85   0.48    1.00
```

```
#we calculate the correlation matrix between all the variables in the dataset
corrplot(res, method="number", type = "lower", order = "original",
         tl.col = "black", tl.srt = 45)
```




```
#Parametric bootstrap to prove H0="body weight influences total sleep time" (the slope is not 0)
set.seed(1111)
myData = msleep %>% select(bodywt,sleep_total)
N=10000
siz=35
x <- 1:N;
for (i in 1:N){
  boot  = na.omit(myData[sample(na.omit(siz),(siz),replace = TRUE),])
  bootX = as.matrix(boot[,c(2)])
  bootY = as.matrix(boot[,c(1)])
  coefs = lm(bootX ~ bootY)$coefficients
  x[i]= coefs["bootY"]
}

#We find the standard errors without having to derive an equation manually
x = sort(x)
SE_slope = sd(x)
mean_slope = mean(x)
#outputting the estimate and the Standard Error
SE_slope
```

```
## [1] 0.005741366
```

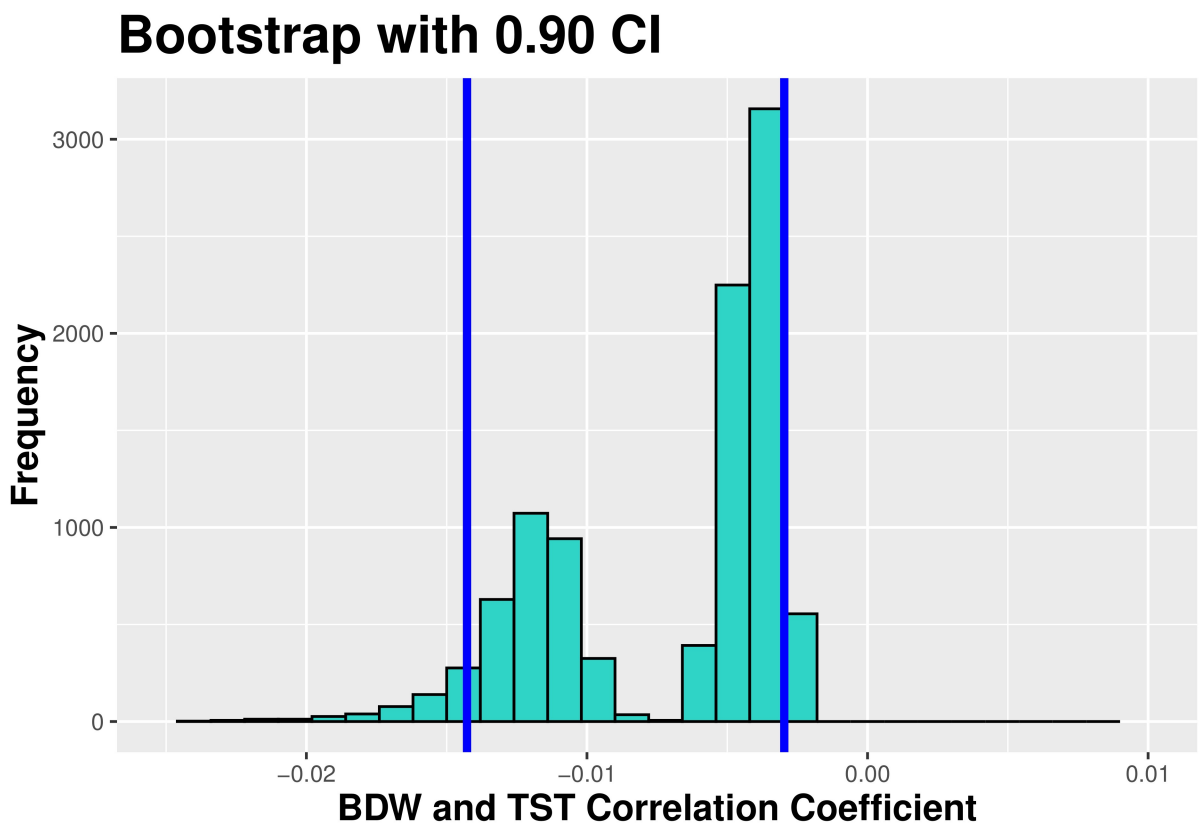
```
mean_slope
```

```
## [1] -0.007267462
```

```r
#Our CI is determined by the 500th and 9500th value of the ordered set of results
p<- ggplot() + aes(x)+ geom_histogram(binwidth=0.0012, colour="black", fill="#30d5c8")  + xlim(-0.025,0
  ggtitle("Bootstrap with 0.90 CI") + xlab("BDW and TST Correlation Coefficient") + ylab("Frequency") +
  theme(plot.title = element_text(color="black", size=20, face="bold"), axis.title.x = element_text(col
  geom_vline(xintercept = x[500], color = "blue", size=1.5) +
  geom_vline(xintercept = x[9500], color = "blue", size=1.5)
p
```

```
## Warning: Removed 47 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```




```r
#Non parametric bootstrap to prove H0="body weight influences total sleep time" (the slope is not 0)
set.seed(2222)
myData = msleep %>% select(bodywt,sleep_total)
correlator = function(base,i){return(cor(base[i,]$bodywt, base[i,]$sleep_total))}
crtl_boot = boot(myData,correlator,R=10000)
#Our CI is determined following normal distribution quantiles.
boot.ci(boot.out = crtl_boot, type = c("norm"), conf = 0.90)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
```

```
## boot.ci(boot.out = crtl_boot, conf = 0.9, type = c("norm"))
##
## Intervals :
## Level      Normal
## 90%   (-0.3798, -0.1458 )
## Calculations and Intervals on Original Scale
```

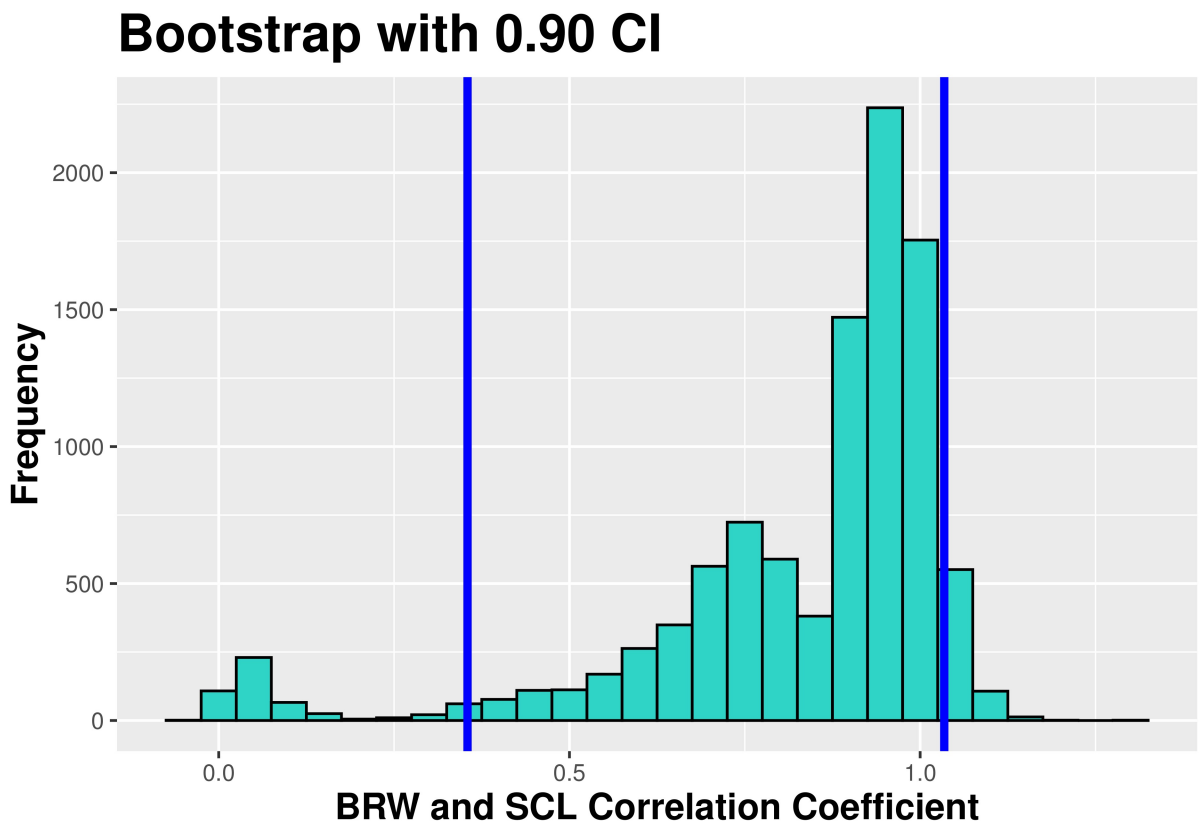
```
#Parametric Bootstrap to prove H0="brain weight influences sleep cycle length" (the slope is not 0)
set.seed(5555)
myData = msleep %>% select(sleep_cycle,brainwt)
N=10000
siz=35
x <- 1:N;
for (i in 1:N){
  boot  = na.omit(myData[sample(na.omit(siz),(siz),replace = TRUE),])
  bootX = as.matrix(boot[,c(2)])
  bootY = as.matrix(boot[,c(1)])
  coefs = lm(bootX ~ bootY)$coefficients
  x[i]= coefs["bootY"]
}

#We find the standard errors without having to derive an equation manually
x = sort(x)
SE_slope = sd(x)
mean_slope = mean(x)
#outputting the estimate and the Standard Error
SE_slope
```

```
## [1] 0.225093
```

```
mean_slope
```

```
## [1] 0.8317667
```

```
#Our CI is determined by the 500th and 9500th value of the ordered set of results
p<- ggplot() + aes(x)+ geom_histogram(binwidth=0.05, colour="black", fill="#30d5c8") +
  geom_vline(xintercept = x[500], color = "blue", size=1.5) +
  geom_vline(xintercept = x[9500], color = "blue", size=1.5) +
  ggtitle("Bootstrap with 0.90 CI") + xlab("BRW and SCL Correlation Coefficient") + ylab("Frequency") +
  theme(plot.title = element_text(color="black", size=20, face="bold"), axis.title.x = element_text(col
p
```




```
#Non parametric bootstrap to prove HO="brain weight influences sleep cycle length" (the slope is not 0)
set.seed(2222)
myDataInc = msleep %>% select(sleep_cycle,brainwt)
myData = myDataInc[complete.cases(myDataInc), ]
correlator = function(base,i){return(cor(base[i,]$sleep_cycle, base[i,]$brainwt))}
crtl_boot = boot(myData,correlator,R=10000)
#Our CI is determined following normal distribution quantiles.
boot.ci(boot.out = crtl_boot, type = c("norm"), conf=0.90)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = crtl_boot, conf = 0.9, type = c("norm"))
##
## Intervals :
## Level      Normal
## 90%   ( 0.7631,  0.9168 )
## Calculations and Intervals on Original Scale
```

```
#brief correlation study dividing the dataset by the mamals' diet.
#NOTE: this does NOT appear in the paper, but it was useful to the investigation.
myDataCarni = msleep %>% filter(vore=="carni") %>% select(sleep_total,sleep_rem,sleep_cycle,bodywt,brai
myDataHerbi = msleep %>% filter(vore=="herbi") %>% select(sleep_total,sleep_rem,sleep_cycle,bodywt,brai
myDataOmniv = msleep %>% filter(vore=="omni")  %>% select(sleep_total,sleep_rem,sleep_cycle,bodywt,brai
```

```r
par(mfrow=c(1,3))
res <- cor(myDataCarni, use = "complete.obs")
round(res, 2)
```

```
##             sleep_total sleep_rem sleep_cycle bodywt brainwt
## sleep_total        1.00      0.62        0.53  -0.49   -0.88
## sleep_rem          0.62      1.00        0.68  -0.08   -0.55
## sleep_cycle        0.53      0.68        1.00  -0.73   -0.81
## bodywt            -0.49     -0.08       -0.73   1.00    0.82
## brainwt           -0.88     -0.55       -0.81   0.82    1.00
```

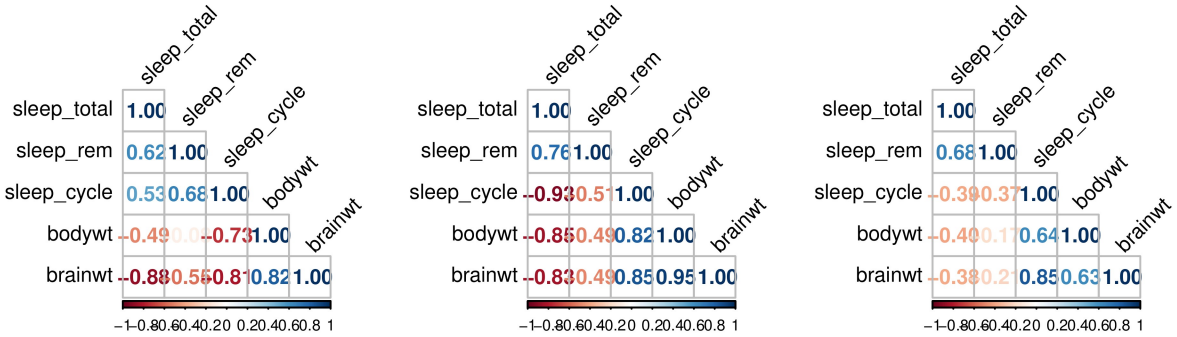
```r
corrplot(res, method="number", type = "lower", order = "original",
         tl.col = "black", tl.srt = 45)
res <- cor(myDataHerbi, use = "complete.obs")
round(res, 2)
```

```
##             sleep_total sleep_rem sleep_cycle bodywt brainwt
## sleep_total        1.00      0.76       -0.93  -0.85   -0.83
## sleep_rem          0.76      1.00       -0.51  -0.49   -0.49
## sleep_cycle       -0.93     -0.51        1.00   0.82    0.85
## bodywt            -0.85     -0.49        0.82   1.00    0.95
## brainwt           -0.83     -0.49        0.85   0.95    1.00
```

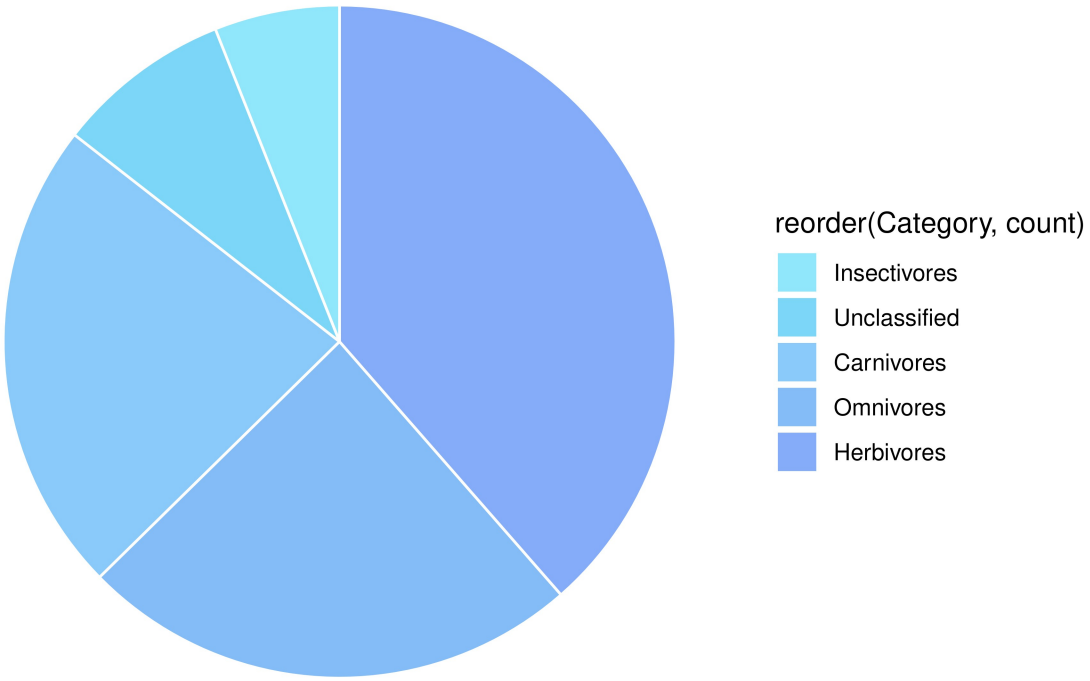
```r
corrplot(res, method="number", type = "lower", order = "original",
         tl.col = "black", tl.srt = 45)
res <- cor(myDataOmniv, use = "complete.obs")
round(res, 2)
```

```
##             sleep_total sleep_rem sleep_cycle bodywt brainwt
## sleep_total        1.00      0.68       -0.39  -0.40   -0.38
## sleep_rem          0.68      1.00       -0.37  -0.17   -0.21
## sleep_cycle       -0.39     -0.37        1.00   0.64    0.85
## bodywt            -0.40     -0.17        0.64   1.00    0.63
## brainwt           -0.38     -0.21        0.85   0.63    1.00
```

```r
corrplot(res, method="number", type = "lower", order = "original",
         tl.col = "black", tl.srt = 45)
```




```r
#Pie plot of the distribution of the dataset based on the mammals' diet
data <- data.frame(
  Category=c("Carnivores", "Omnivores", "Herbivores", "Insectivores", "Unclassified"),
  count=c(19, 20, 32, 5, 7)
)
ggplot(data, aes(x="", y=count, fill=reorder(Category,count))) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0) +
  #good looking color palette
  scale_fill_manual(values=c("#90E6FC", "#7CD7F7", "#8BCBFC", "#83BDF7", "#84ACFA")) +
  theme_void() # remove background, grid, numeric labels
```




```
#Data filtering based on their diet
myDataCH = msleep %>% filter(vore=="carni" | vore=="herbi")
par(mfrow=c(1,2))
#Two plots with their respective regression lines
P1<-ggplot(myDataCH, aes(x=sleep_total, y=sleep_cycle, col=vore)) +
  geom_point() +
  xlim(4,14)+
  labs(x='Total Sleep Time (h)', y='Sleep Cycle Length (h)', col='Diet') +
  stat_smooth(method='lm')+
  theme(plot.title = element_text(hjust=0.5, size=15, face='bold'), aspect.ratio = 1.2)
P2<-ggplot(myDataCH, aes(x=brainwt, y=sleep_cycle, col=vore)) +
  geom_point() +
  xlim(0,0.7)+
  labs(x='Brain Weight (kg)', y='Sleep Cycle Length (h)', col='Diet') +
  stat_smooth(method='lm')+
  theme(plot.title = element_text(hjust=0.5, size=15, face='bold'), aspect.ratio = 1.2)
P<-plot_grid(P1, P2)
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 38 rows containing non-finite values (stat_smooth).

## Warning: Removed 38 rows containing missing values (geom_point).

## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 36 rows containing non-finite values (stat_smooth).

## Warning: Removed 36 rows containing missing values (geom_point).
```

```
P
```