Ryan Torelli
CptS 483-04
Assignment 4
October 26, 2017

1.

```
# Load dplyr
library(dplyr)

# Read msleep_ggplot2.csv
msleep <- read.csv("msleep_ggplot2.csv", header=TRUE,
                           colClasses=c("name"="character",
                                        "genus"="character",
                                        "vore"="character",
                                        "order"="character",
                                        "conservation"="character"))
```

(a)

```
# Select variables containing “sleep” and print head
msleep %>%
   select(contains("sleep")) %>%
   head
```

```
  sleep_total sleep_rem sleep_cycle
1        12.1        NA          NA
2        17.0       1.8          NA
3        14.4       2.4          NA
4        14.9       2.3   0.1333333
5         4.0       0.7   0.6666667
6        14.4       2.2   0.7666667
```

(b)

```
# Filter for bodywt > 50 and sleep_total > 6 and count
msleep %>%
   filter(bodywt > 50, sleep_total > 6) %>%
   nrow
```

```
[1] 8
```

(c)
Select name, order, sleep_total, and bodywt
Arrange by descending order of sleep_total and print top 6 rows

```
msleep %>%
   select(name, order, sleep_total, bodywt) %>%
   arrange(desc(sleep_total)) %>%
   top_n(6, sleep_total)
```

```
                   name           order sleep_total bodywt
1       Little brown bat      Chiroptera        19.9  0.010
2          Big brown bat      Chiroptera        19.7  0.023
3   Thick-tailed opposum Didelphimorphia        19.4  0.370
4        Giant armadillo       Cingulata        18.1 60.000
5 North American Opossum Didelphimorphia        18.0  1.700
6   Long-nosed armadillo       Cingulata        17.4  3.500
```

(d)
Mutate wt_ratio from (brainwt/bodywt) and
rem_ratio from (sleep_rem/sleep_total)

```
msleep <- mutate(msleep, wt_ratio = brainwt / bodywt,
                         rem_ratio = sleep_rem / sleep_total)
head(msleep)
```

```
                      name      genus  vore        order conservation
1                  Cheetah   Acinonyx carni    Carnivora           lc
2              Owl monkey      Aotus  omni     Primates         <NA>
3          Mountain beaver Aplodontia herbi     Rodentia           nt
4 Greater short-tailed shrew    Blarina  omni Soricomorpha           lc
5                      Cow        Bos herbi Artiodactyla domesticated
6        Three-toed sloth   Bradypus herbi       Pilosa         <NA>

 sleep_total sleep_rem sleep_cycle awake brainwt  bodywt   wt_ratio rem_ratio
1        12.1        NA          NA  11.9      NA  50.000        NA       NA
2        17.0       1.8          NA   7.0 0.01550   0.480 0.032291670.1058824
3        14.4       2.4          NA   9.6      NA   1.350        NA 0.1666667
4        14.9       2.3   0.1333333   9.1 0.00029   0.019 0.015263160.1543624
5         4.0       0.7   0.6666667  20.0 0.42300 600.000 0.000705000.1750000
6        14.4       2.2   0.7666667   9.6      NA   3.850        NA 0.1527778
```

(e)
Group_by order and summarise the average, min and max of sleep_total
```
msleep %>%
  group_by(order) %>%
  summarise(mean(sleep_total),
            min(sleep_total),
            max(sleep_total)) %>%
  ungroup
```

	order	`mean(sleep_total)`	`min(sleep_total)`	`max(sleep_total)`
	<chr>	<dbl>	<dbl>	<dbl>
1	Afrosoricida	15.600000	15.6	15.6
2	Artiodactyla	4.516667	1.9	9.1
3	Carnivora	10.116667	3.5	15.8
4	Cetacea	4.500000	2.7	5.6
5	Chiroptera	19.800000	19.7	19.9
6	Cingulata	17.750000	17.4	18.1
7	Didelphimorphia	18.700000	18.0	19.4
8	Diprotodontia	12.400000	11.1	13.7
9	Erinaceomorpha	10.200000	10.1	10.3
10	Hyracoidea	5.666667	5.3	6.3
11	Lagomorpha	8.400000	8.4	8.4
12	Monotremata	8.600000	8.6	8.6
13	Perissodactyla	3.466667	2.9	4.4
14	Pilosa	14.400000	14.4	14.4
15	Primates	10.500000	8.0	17.0
16	Proboscidea	3.600000	3.3	3.9
17	Rodentia	12.468182	7.0	16.6
18	Scandentia	8.900000	8.9	8.9
19	Soricomorpha	11.100000	8.4	14.9

(f)
```
# Replace missing values of brainwt with the order's wt_ratio * mammal's bodywt in a copy
msleep2 <- msleep
msleep2 <- group_by(msleep2, order)
msleep2 <- mutate(msleep2, order_wt_ratio = mean(wt_ratio, na.rm=TRUE))
msleep2 <- ungroup(msleep2)
msleep2$brainwt <- ifelse(is.na(msleep2$brainwt), msleep2$order_wt_ratio*msleep2$bodywt,
                          msleep2$brainwt)

# Replace missing values of sleep_rem with the order's rem_ratio * mammal's sleep_total in a copy
msleep3 <- msleep
msleep3 <- group_by(msleep3, order)
msleep3 <- mutate(msleep3, order_rem_ratio = mean(rem_ratio, na.rm=TRUE))
msleep3 <- ungroup(msleep3)
msleep3$sleep_rem <- ifelse(is.na(msleep3$sleep_rem), msleep3$order_rem_ratio*msleep3$sleep_total,
                            msleep3$sleep_rem)
```

```
# Print head of msleep, msleep2, and msleep3
msleep %>%
  select(name, order, brainwt, bodywt, wt_ratio, sleep_rem, sleep_total, rem_ratio) %>%
  head
```

```
                        name        order brainwt  bodywt   wt_ratiosleep_rem
1                    Cheetah    Carnivora      NA  50.000         NA       NA
2                 Owl monkey     Primates 0.01550   0.480 0.03229167      1.8
3             Mountain beaver     Rodentia      NA   1.350         NA      2.4
4 Greater short-tailed shrew Soricomorpha 0.00029   0.019 0.01526316      2.3
5                        Cow Artiodactyla 0.42300 600.000 0.00070500      0.7
6           Three-toed sloth       Pilosa      NA   3.850         NA      2.2
  sleep_total rem_ratio
1        12.1        NA
2        17.0 0.1058824
3        14.4 0.1666667
4        14.9 0.1543624
5         4.0 0.1750000
6        14.4 0.1527778
```

```
msleep2 %>%
  select(name, order, brainwt, bodywt, wt_ratio, order_wt_ratio) %>%
  head
```

```
                        name        order    brainwt  bodywt   wt_ratio
                       <chr>        <chr>      <dbl>   <dbl>      <dbl>
1                    Cheetah    Carnivora 0.37129771  50.000         NA
2                 Owl monkey     Primates 0.01550000   0.480 0.03229167
3             Mountain beaver     Rodentia 0.01892814   1.350         NA
4 Greater short-tailed shrew Soricomorpha 0.00029000   0.019 0.01526316
5                        Cow Artiodactyla 0.42300000 600.000 0.00070500
6           Three-toed sloth       Pilosa        NaN   3.850         NA
```

```
msleep3 %>%
  select(name, order, sleep_rem, sleep_total, rem_ratio, order_rem_ratio) %>%
  head
```

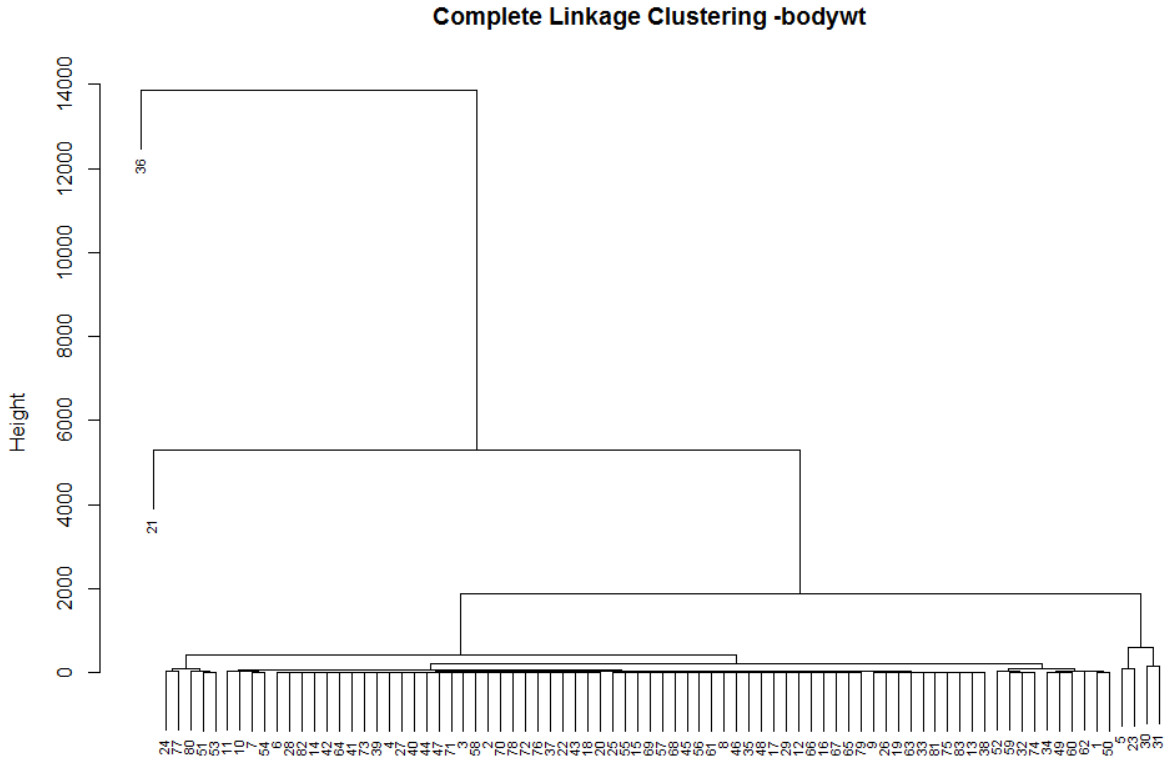
```
                        name        order sleep_rem sleep_total rem_ratio
                       <chr>        <chr>     <dbl>       <dbl>     <dbl>
1                    Cheetah    Carnivora  2.612764        12.1        NA
2                 Owl monkey     Primates  1.800000        17.0 0.1058824
3             Mountain beaver     Rodentia  2.400000        14.4 0.1666667
4 Greater short-tailed shrew Soricomorpha  2.300000        14.9 0.1543624
5                        Cow Artiodactyla  0.700000         4.0 0.1750000
6           Three-toed sloth       Pilosa  2.200000        14.4 0.1527778
```

There are several assumptions in imputing data by using an order's average ratio. There must be at least one member in an order with sufficient data to calculate an order's average ratio. Preferably, there are many members in an order from which to calculate a precise and accurate average ratio. Every mammal with a missing value must have sufficient data against which to multiply the order's average value. Most importantly, within a class, there must be a strong linear relationship between the two variables that form the ratio.

It is better to make a model and predict missing values than impute by an order's average ratio. This requires some data analysis to identify variables that correlate with the variable of interest. The candidate variables should then be culled to exclude collinearity and be expanded to include interaction terms. The resulting regression model should predict missing values with greater accuracy than use of an order's average ratio.

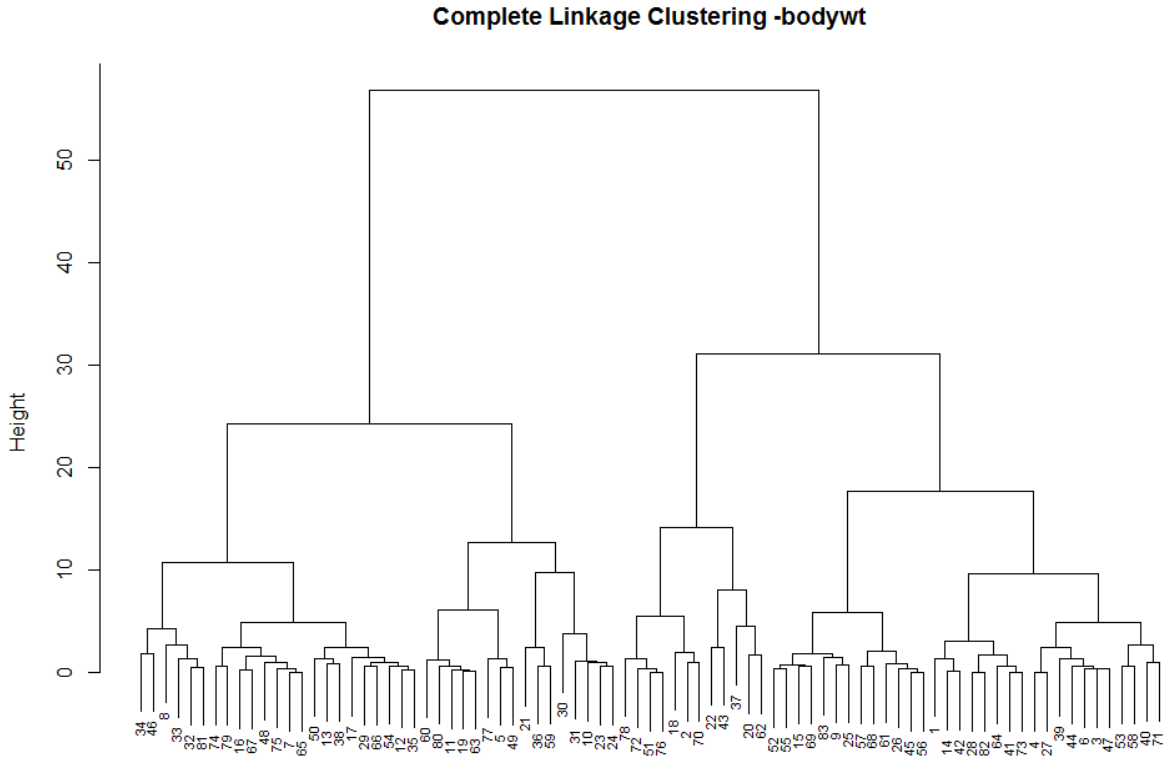
 # (g)

```
# Perform complete linkage clustering of msleep data with Euclidian distance as measure
hc <- hclust(dist(msleep), method="complete")

#Generate a dendrogram of your clustering
plot(hc, main="Complete Linkage Clustering", xlab="", sub="", cex=.7)
```



Complete linkage clustering reveals that the mammals are very similar, excepting two outliers. The mammals identified as #21 and #36 are most distant from other mammals. Mammal #21 is the Asian elephant, which sleeps 3.9h daily, has a 4.6kg brain, and weighs 2547kg. Mammal #36 is the African elephant, which sleeps 3.3h daily, has a 5.7kg brain, and weights 6654kg. In comparison to other mammals, these two elephants sleep very little and weigh very much. However, they are outliers due to body weight. When body weight is removed from the dataset and clustering repeated, the dendrogram does not have outliers.

```
hc <- hclust(dist(subset(msleep, select=-bodywt)), method="complete")
plot(hc, main="Complete Linkage Clustering -bodywt", xlab="", sub="", cex=.7)
```



 # (h)

# Cut into 4 or 8 clusters and print cluster membership by order
cut <- cutree(hc, 8)
table(msleep$order, cut)

	1	2	3	4	5	6	7	8
Afrosoricida	0	1	0	0	0	0	0	0
Artiodactyla	0	0	3	1	0	0	2	0
Carnivora	4	1	0	1	2	3	1	0
Cetacea	0	0	2	0	0	0	1	0
Chiroptera	0	0	0	0	0	0	0	2
Cingulata	0	1	0	0	0	0	0	1
Didelphimorphia	0	0	0	0	0	0	0	2
Diprotodontia	1	0	0	0	0	1	0	0
Erinaceomorpha	0	0	0	0	0	2	0	0
Hyracoidea	0	0	2	0	1	0	0	0
Lagomorpha	0	0	0	1	0	0	0	0
Monotremata	0	0	0	1	0	0	0	0
Perissodactyla	0	0	1	0	0	0	2	0
Pilosa	1	0	0	0	0	0	0	0
Primates	0	1	0	7	1	3	0	0
Proboscidea	0	0	0	0	0	0	2	0
Rodentia	11	3	0	3	2	3	0	0
Scandentia	0	0	0	1	0	0	0	0
Soricomorpha	2	0	0	2	0	1	0	0

cut <- cutree(hc, 4)
table(msleep$order, cut)

	1	2	3	4
Afrosoricida	0	1	0	0
Artiodactyla	0	0	5	1
Carnivora	7	1	1	3
Cetacea	0	0	3	0
Chiroptera	0	2	0	0
Cingulata	0	2	0	0
Didelphimorphia	0	2	0	0
Diprotodontia	2	0	0	0
Erinaceomorpha	2	0	0	0
Hyracoidea	0	0	2	1
Lagomorpha	0	0	0	1
Monotremata	0	0	0	1
Perissodactyla	0	0	3	0
Pilosa	1	0	0	0
Primates	3	1	0	8
Proboscidea	0	0	2	0
Rodentia	14	3	0	5
Scandentia	0	0	0	1
Soricomorpha	3	0	0	2

Mammals of order Carnivora are more spread out among clusters than mammals of order Primate or Rodentia. In the k=8 table, four of 12 carnivores cluster in one bin and the remaining carnivores spread among five other bins. Seven of 12 primates cluster in one bin and the remaining spread among three other bins. Eleven of 22 rodents cluster in one bin and the remaining spread among four other bins. Carnivores and rodents cluster the greatest number of

members in the same bin; they are more similar to each other than to primates with respect to this dataset.

 # (i)

```
# Select sleep_total, sleep_rem, and brainwt, using calculated missing values
msleep4 <- msleep
msleep4$brainwt <- msleep2$brainwt
msleep4$sleep_rem <- msleep3$sleep_rem
msleep4 <- subset(msleep4, select=c(sleep_total, sleep_rem, brainwt))

# Omit a few entries with NaN values
msleep4 <- na.omit(msleep4)

# View variances and means
apply(msleep4, 2, mean)
```

```
sleep_total   sleep_rem     brainwt
 10.8197368   1.9523600   0.1587237
```

```
apply(msleep4, 2, var)
```

```
sleep_total   sleep_rem     brainwt
 18.5266719   1.4318227   0.1723787
```

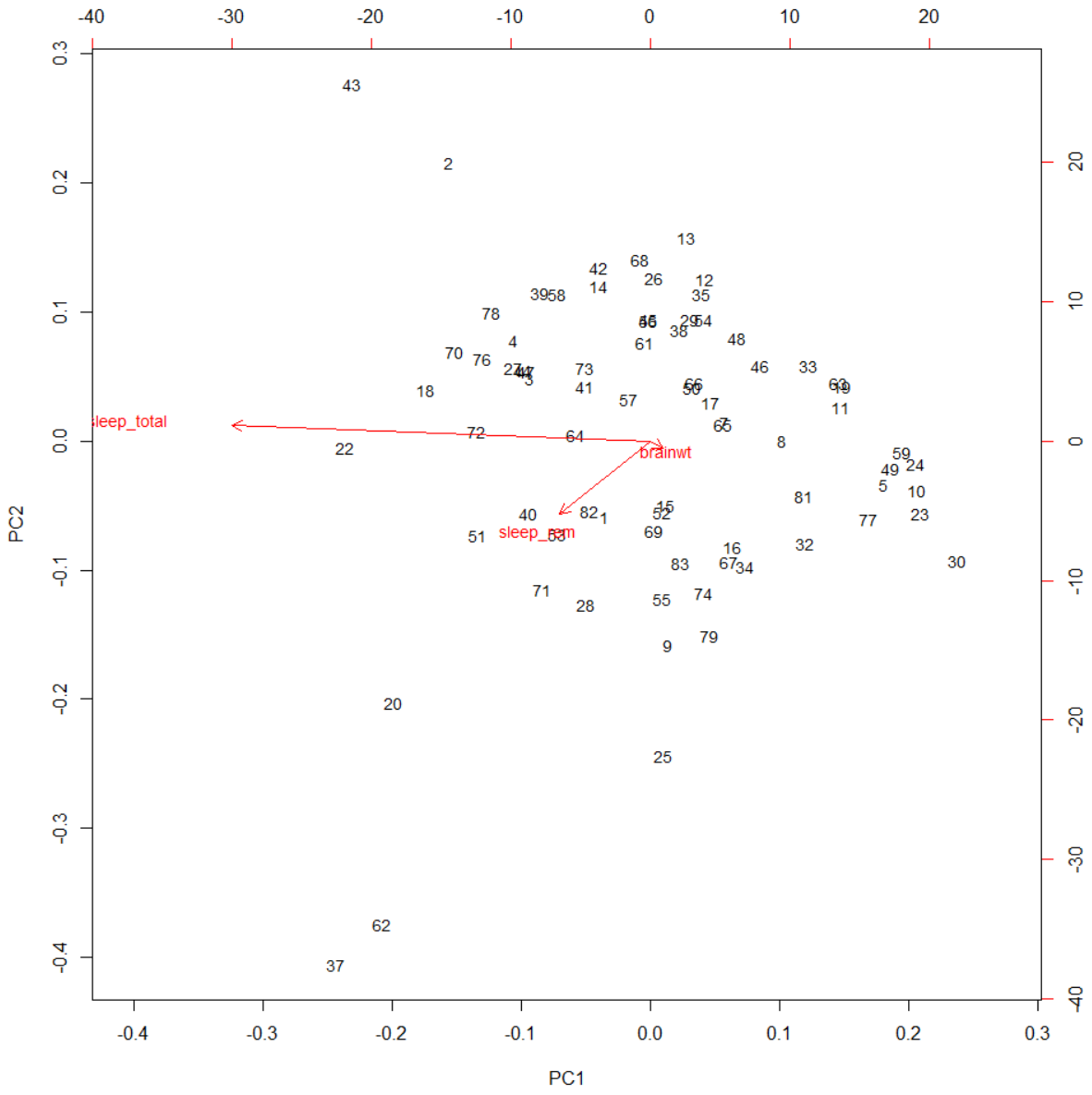
```
# Perform PCA
pca <- prcomp(msleep, scale=FALSE)

# View principal component loading vectors by column
pca$rotation
```

```
                    PC1         PC2          PC3
sleep_total -0.9769144  0.2076280  0.05028728
sleep_rem   -0.2116157 -0.9727629 -0.09461009
brainwt      0.0292739 -0.1030675  0.99424349
```

```
# View proportion of variance explained by each principal component
pca_var <- pca$sdev^2
output <- pca_var / sum(pca_var)
output
```

```
[1] 0.962967825 0.029521889 0.007510286
```

# Print a biplot
biplot(pca, cex=.9)



 # (j)

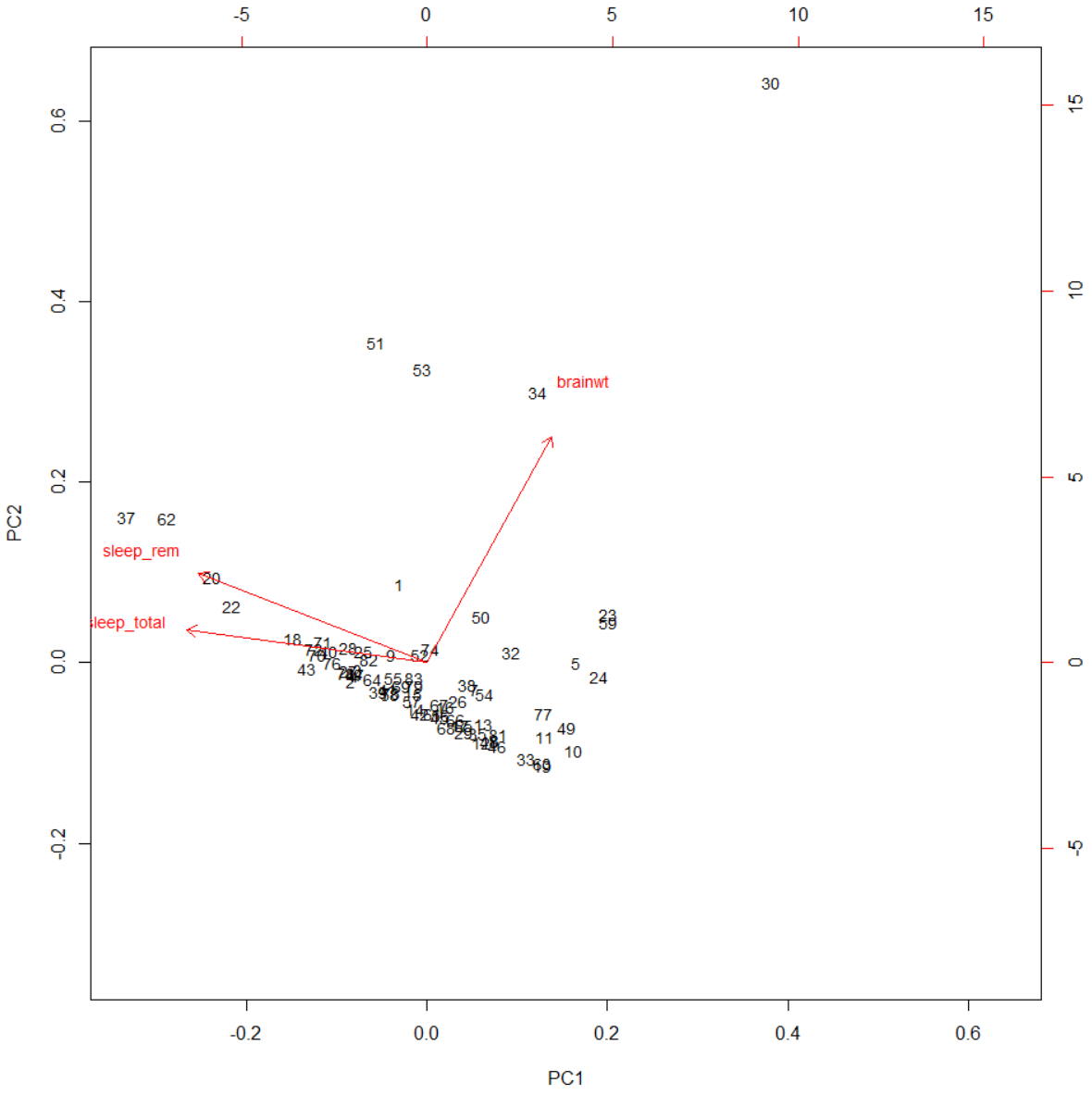
```
# Perform PCA with scaling
pca <- prcomp(msleep4, scale=TRUE)

#Principal component loading vectors by column
pca$rotation
```

```
                    PC1       PC2        PC3
sleep_total -0.6790635 0.1332548 -0.7218836
sleep_rem   -0.6439466 0.3639864  0.6729389
brainwt      0.3524282 0.9218227 -0.1613609
```

```
# proportion of variance explained by each principal component
pca_var <- pca$sdev^2
output <- pca_var / sum(pca_var)
output
```

```
[1] 0.62604312 0.29844912 0.07550776
```

#Plot first two principal components
biplot(pca, cex=.9)



There is a significant difference in biplots when variables are scaled. The total sleep time drives principal components in the unscaled analysis. It ranges from 2 to 20 with mean of 11, whereas REM sleep time and brain weight have small variances centered on means of 1.9 and 0.15, respectively. The large absolute values and large variance of total sleep time impact the unscaled analysis greatly. In the analysis of this dataset, it is very important to scale since only the scaled analysis shows variance without bias from significantly different absolute values among variables.