Lab 02 - Data Summary

WILD 502 - Jay Rotella

Import and Organize the Data

First, import the data and take a quick look to see if all appears to have imported as expected.

```
library(dplyr)   # for data manipulation
library(ggplot2) # for plotting
library(GGally)  # for 'ggpairs' function

# read in the file
dat <- read.table("http://www.montana.edu/rotella/documents/502/FawnsData.txt", header = TRUE)

# check if things appear as expected
head(dat)
```

```
##    radio eh freq area sex mass length
## 1 191.08 10    1    1   1 34.2  125.5
## 2 191.19 11    1    0   0 27.5  112.0
## 3 191.35 10    1    0   0 35.5  128.0
## 4 191.36 11    1    0   0 33.0  122.0
## 5 191.37 10    1    0   0 34.3  123.0
## 6 191.40 11    1    0   0 29.6  117.5
```

```
summary(dat)
```

```
##      radio             eh             freq        area      
##  Min.   :191.1   Min.   :10.00   Min.   :1   Min.   :0.000  
##  1st Qu.:192.0   1st Qu.:10.00   1st Qu.:1   1st Qu.:0.000  
##  Median :192.4   Median :11.00   Median :1   Median :0.000  
##  Mean   :192.2   Mean   :10.59   Mean   :1   Mean   :0.487  
##  3rd Qu.:192.7   3rd Qu.:11.00   3rd Qu.:1   3rd Qu.:1.000  
##  Max.   :192.8   Max.   :11.00   Max.   :1   Max.   :1.000  
##       sex              mass           length     
##  Min.   :0.0000   Min.   :22.80   Min.   :108.0  
##  1st Qu.:0.0000   1st Qu.:31.90   1st Qu.:120.0  
##  Median :1.0000   Median :33.60   Median :124.0  
##  Mean   :0.5043   Mean   :34.38   Mean   :123.2  
##  3rd Qu.:1.0000   3rd Qu.:36.60   3rd Qu.:127.0  
##  Max.   :1.0000   Max.   :72.00   Max.   :135.5
```

Getting the Data Ready for Summary Work

We'll only work with those columns that are of interest for the modeling to be done. For example, we won't work with the 'radio' variable as that's the radio frequency of the animal's transmitter. Also, there's a 'freq' column that indicates how many animals are being represented in the row: as can be seen in the data summary above, the value is always 1, so we can drop it and simplify our work. Our primary interest is in evaluating the covariates (area, sex, mass, and length). But, for this known-fate problem, we might also be interested in looking at some summaries of animal fates, which are represented by 'eh' or encounter history (10 = lived, 11 = died). Finally, 'area' and 'sex' are stored as 0 or 1 and were imported as numeric rather than categorical variables. We can store them as factors and give them more informative names to make our data summaries a bit easier to interpret. Notice how the summary information has changed for the variables treated as factors and let's us quickly see that survival is modest and that the data are quite balanced across sexes and areas. We also see no 'NA' values, i.e., no missing data.

```
dat2 <- dat %>%
  select(-c(radio, freq)) %>%
  mutate(eh = factor(eh,
                       levels = c(11, 10),
                       labels = c("die", "live")),
         area = factor(area,
                       levels = c(0, 1),
                       labels = c("cntrl", "trmt")),
         sex = factor(sex,
                      levels = c(0, 1),
                      labels = c("female", "male")))
summary(dat2)
```

```
##     eh        area        sex          mass           length
##  die :68   cntrl:59   female:57   Min.   :22.80   Min.   :108.0
##  live:47   trmt :56   male  :58   1st Qu.:31.90   1st Qu.:120.0
##                                   Median :33.60   Median :124.0
##                                   Mean   :34.38   Mean   :123.2
##                                   3rd Qu.:36.60   3rd Qu.:127.0
##                                   Max.   :72.00   Max.   :135.5
```

Summary Information

You might be interested in looking at summary information about the live-dead status of the fawns overall or broken out by area or sex. We would want to be clear about whether we were doing this to generate hypotheses (i.e., looking for patterns in the data), or to check our data summaries given that we have *a priori* hypotheses about how survival might vary with area and/or sex. The code first provides a table of values broken out by levels of categorical variables used in each call. Next, the values in the table are converted to proportions of values in the 2nd margin (columns) of the table, i.e., the values sum to 1 within each column, which is desired here given that 'live' and 'die' are on separate rows.

```
table(dat2$eh)
```

```
##
##  die live
##   68   47
```

```
round(prop.table(table(dat2$eh)), 3)
```

```
##
##   die  live
## 0.591 0.409
```

```
table(dat2$eh, dat2$area)
```

```
##
##        cntrl trmt
##   die     36   32
##   live    23   24
```

```
round(prop.table(table(dat2$eh, dat2$area), margin = 2), 3)
```

```
##
```

```
##        cntrl  trmt
##   die  0.610 0.571
##   live 0.390 0.429
```

```
table(dat2$eh, dat2$sex)
```

```
##
##        female male
##   die      28   40
##   live     29   18
```

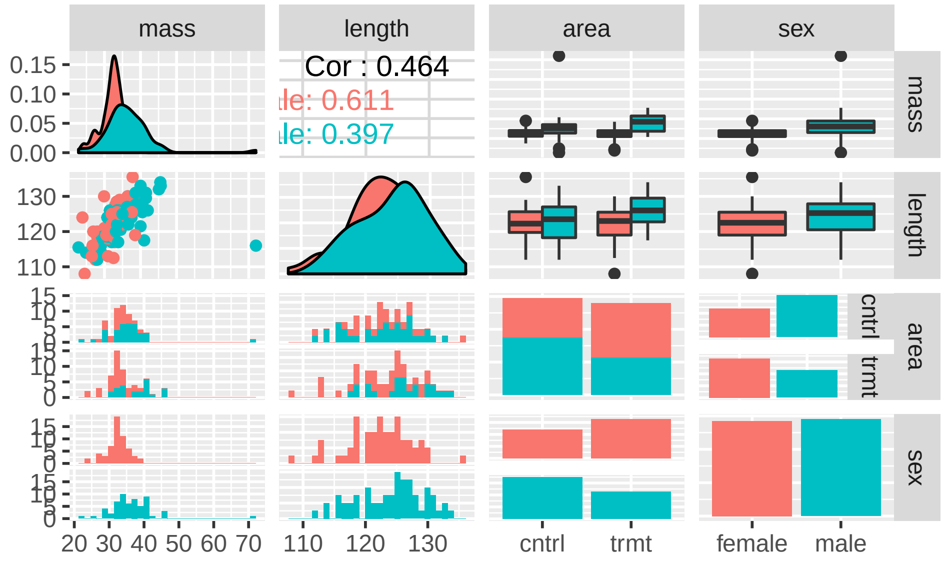
```
round(prop.table(table(dat2$eh, dat2$sex), margin = 2), 3)
```

```
##
##        female  male
##   die   0.491 0.690
##   live  0.509 0.310
```

Evaluate Information about the Covariates

Next, we'll quickly look at (1) how the data for each covariate are distributed (e.g., widely, narrowly, skewed, uniform, normal, other), (2) possible oddities in the data, (3) pair-wise relations. We do this using the 'ggpairs' function of the GGally package, which works in conjunction with the ggplot2 package. Note: the call to 'ggpairs' selects the continuous variables first, which makes the resulting plots easier to read (you can put area and sex first to see what happens if you don't do this; it still works but seems harder to interpret).

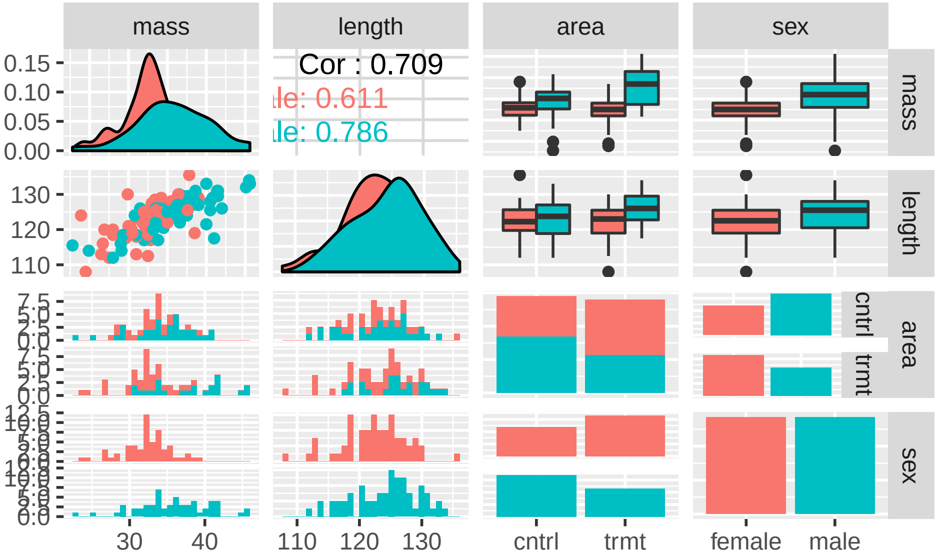
```
dat2 %>%
  select(mass, length, area, sex) %>%
  ggpairs(aes(color = sex))
```



Take a look at the plots and see if you notice anything that looks a bit unusual. Specifically, there appears to be 1 very high value of mass that's just above 70 and far away from other points. It appears to be for a male on the control area. Further, if you look at the plot of points for mass vs. length, it appears that the high mass value is for a relatively short animal. If these were your data, you would definitely want to look back at your datasheets and field notes to see if this was a valid point, a data-entry error that could be fixed, or an issue that you're just discovering. We'll drop the row as it looks to be an error and because such a strong outlier could influence our results.

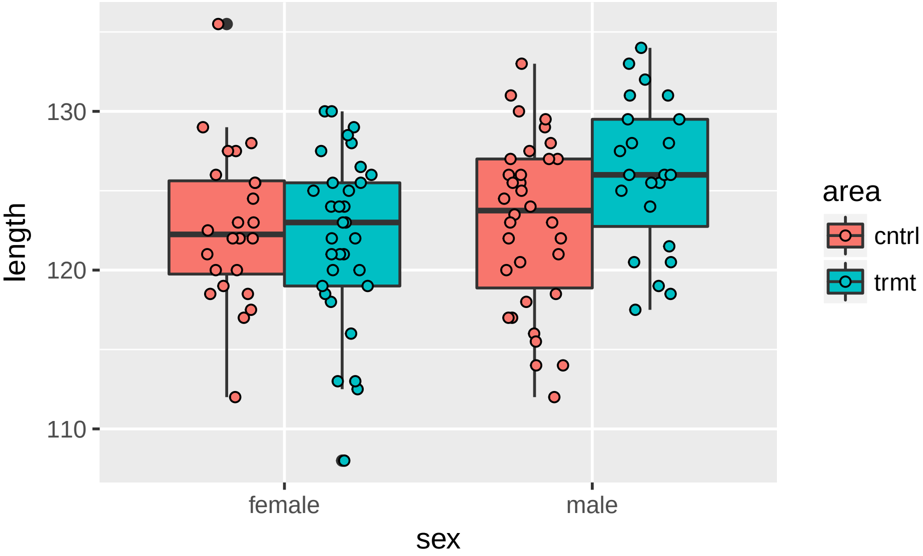
```
dat3 <- dat2 %>%
  subset(mass < 70)

dat3 %>%
  select(mass, length, area, sex) %>%
  ggpairs(aes(color = sex))
```

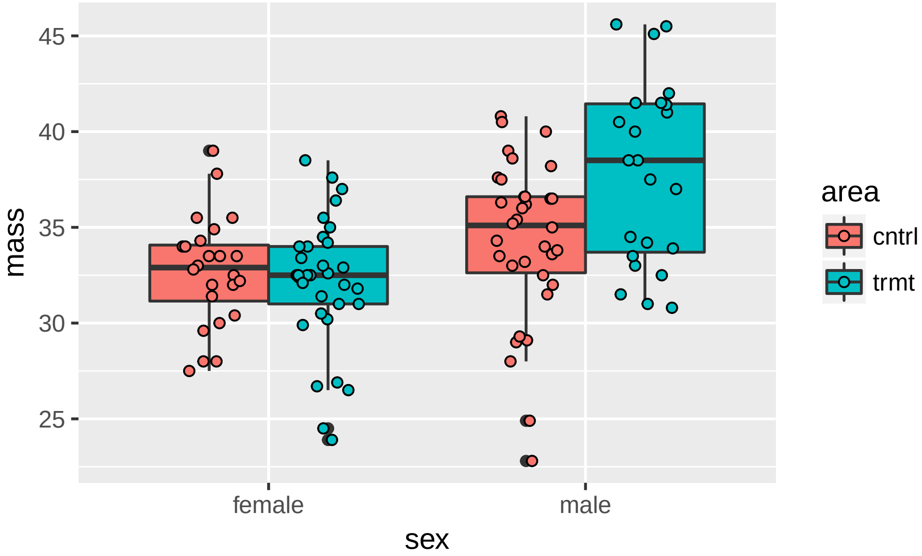


The plots now indicate several interesting features. It appears that there's a pretty good spread of mass and length values for each sex. However, the correlation between mass and length is fairly strong such that you'd want to keep that in mind when building models and interpreting coefficients, i.e., knowing an animal's mass tells you something about it's length. That correlation was much weaker when the animal with the mass of 72 was in the dataset (see earlier plot). We also see some evidence of a tendency for males to perhaps be heavier and a bit longer than the females.

```
ggplot(dat3, aes(x = sex, y = length, fill = area)) +
  geom_boxplot() +
  geom_point(pch = 21, position = position_jitterdodge())
```



```r
ggplot(dat3, aes(x = sex, y = mass, fill = area)) +
  geom_boxplot() +
  geom_point(pch = 21, position = position_jitterdodge())
```



You could certainly do more, but I hope the exercises above reinforce the importance of studying your covariate data before running your models.