Data Transformation and Exploration

Lecture 03.3: Data Exploration

Lauren Sullivan

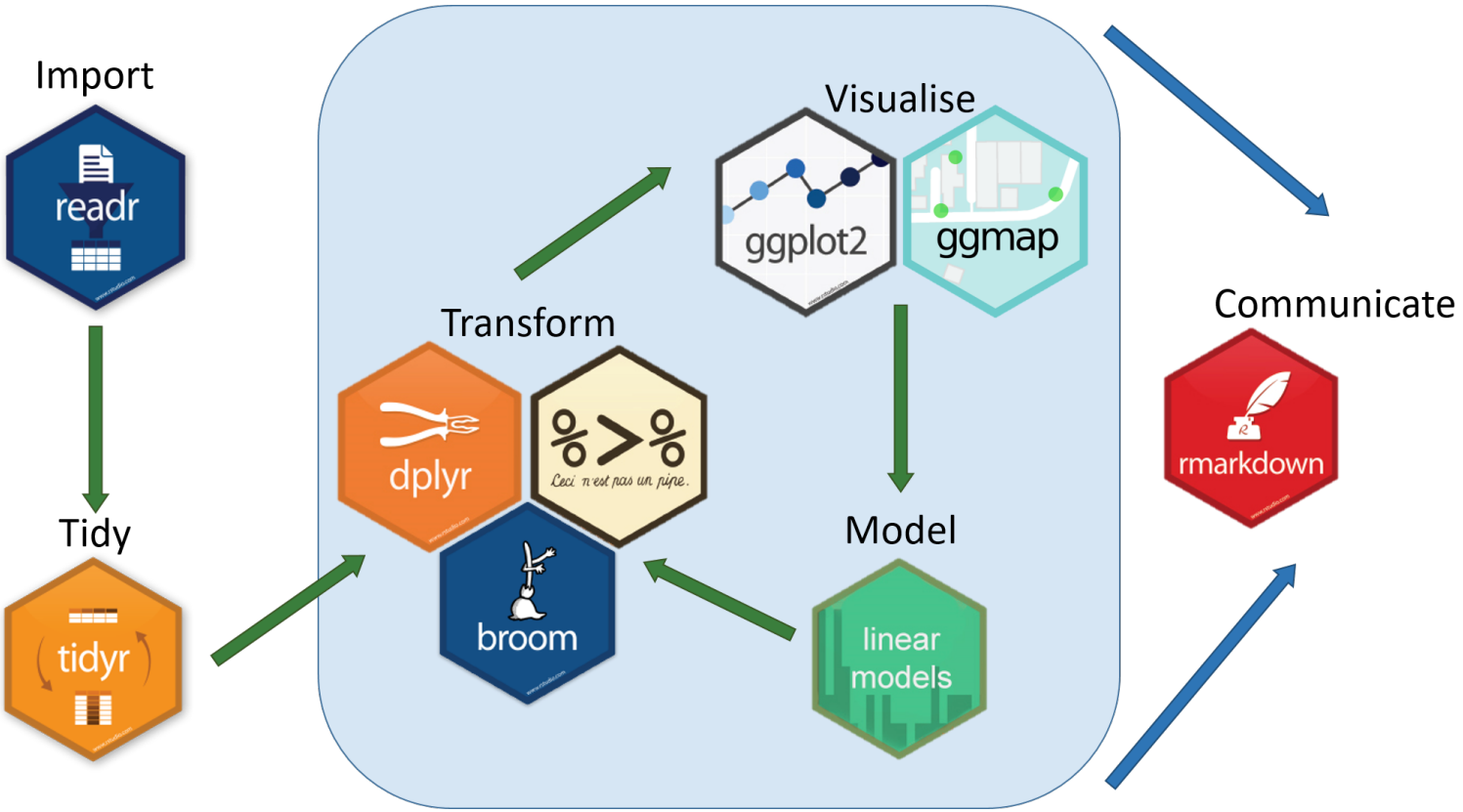
Module: Data Management, Visualization & Reproducibility

Exploratory Data Analysis

Visually exploring your data is often called Exploratory Data Analysis (EDA)

- Helps you figure out what is going on.
- Gives you a clearer picture of if your data entry and analyses are correct.
- Allows you to feel confident with analyses.
- Not a formal process, simply whatever you like to do to examine your data.

EDA with ggplot2



Let's do some EDA on our Bats Data

```r
bats <- read_csv("../data/bats.csv")
bats[1:10,]
```

```
## # A tibble: 10 x 7
##    age   sex   condition   RFA  mass moonlight avg_temp
##    <chr> <chr> <chr>     <dbl> <dbl>     <dbl>    <dbl>
##  1 A     M     NR         34.5   5          99     23.4
##  2 A     M     NR         41.1  11.3        77     25.4
##  3 A     F     L          42.9  13          77     25.4
##  4 A     M     NR         44.5  12          77     25.4
##  5 A     M     NR         35.7   8.8        85     25.0
##  6 A     F     L          46.6  22          85     25.0
##  7 A     F     L          43.9  12          77     26.5
##  8 A     F     L          40.2  11.5        58     26.3
##  9 A     F     P          33.4  10          58     26.3
## 10 A     F     L          44.6  13.3        58     26.3
```

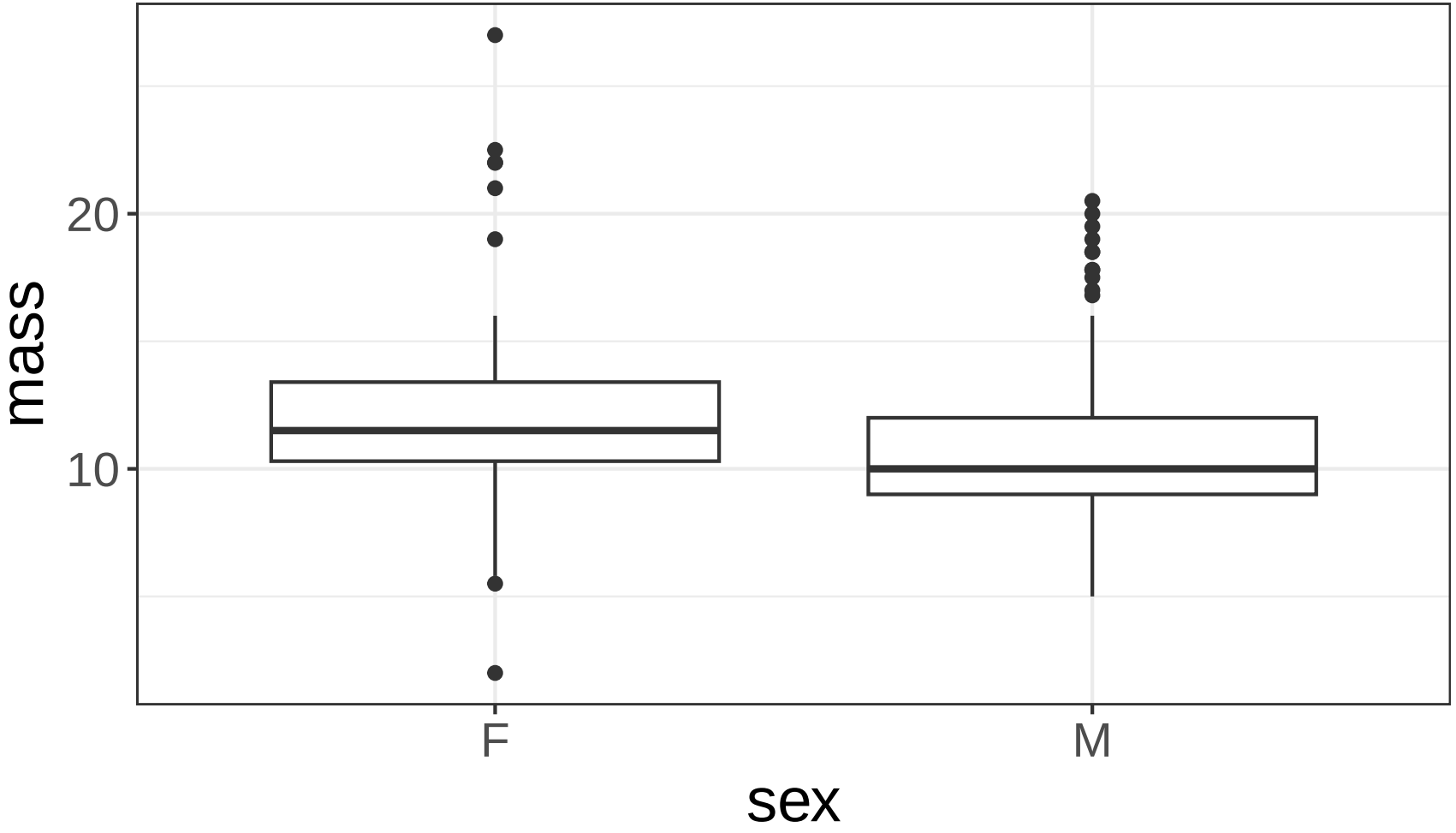
Summary Tables

```r
summary(bats)
```

```
##      age                sex             condition              RFA
##  Length:168         Length:168         Length:168         Min.   :28.90
##  Class :character   Class :character   Class :character   1st Qu.:38.05
##  Mode  :character   Mode  :character   Mode  :character   Median :40.60
##                                                           Mean   :40.49
##                                                           3rd Qu.:42.95
##                                                           Max.   :51.90
##                                                           NA's   :1
##       mass          moonlight         avg_temp
##  Min.   : 2.00   Min.   :  0.00   Min.   :17.16
##  1st Qu.: 9.50   1st Qu.:  7.00   1st Qu.:22.83
##  Median :11.00   Median : 50.00   Median :25.36
##  Mean   :11.72   Mean   : 48.95   Mean   :24.34
##  3rd Qu.:13.00   3rd Qu.: 85.00   3rd Qu.:26.68
##  Max.   :27.00   Max.   :100.00   Max.   :28.12
##
```

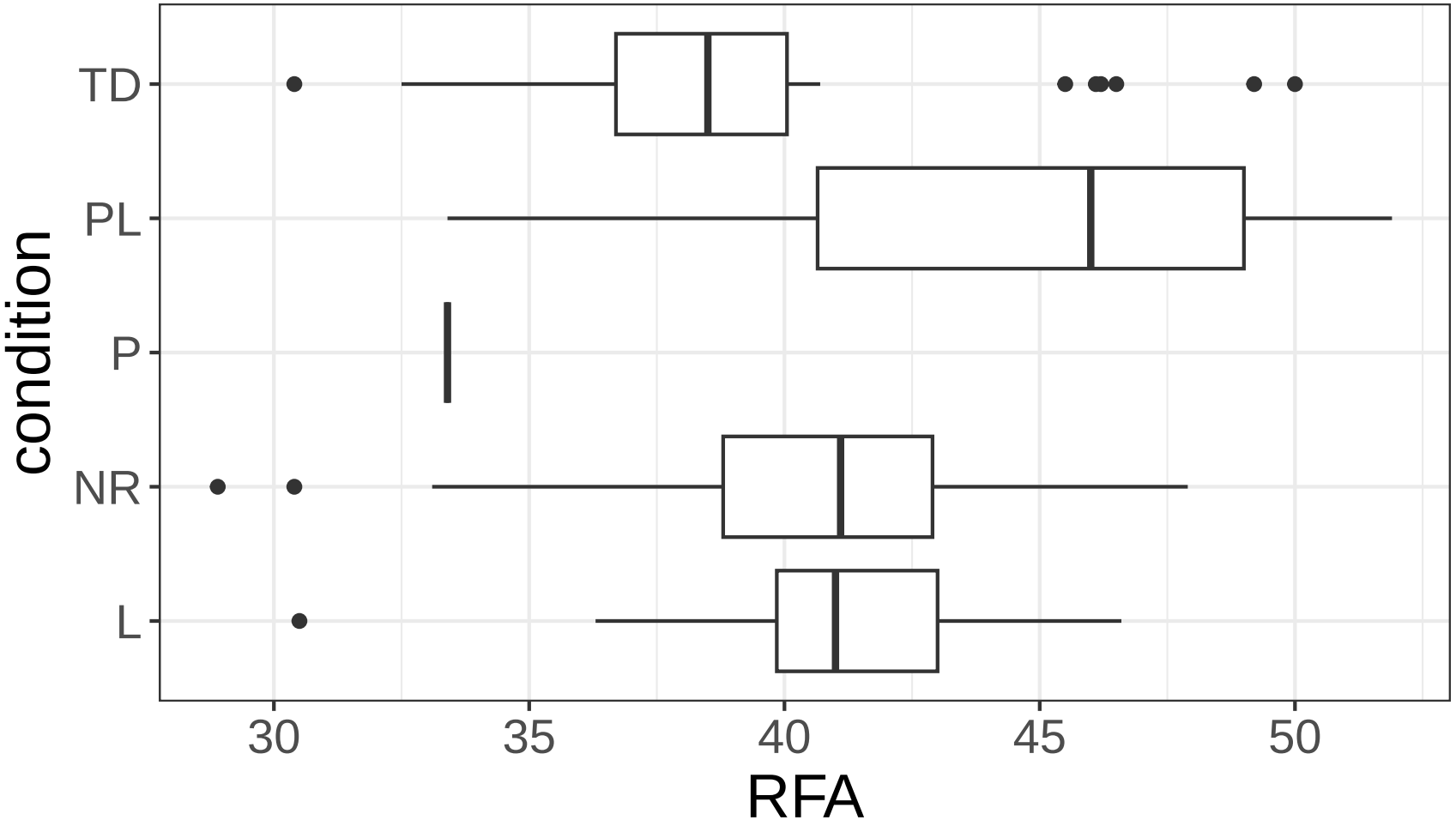
Look at Your Data - Boxplots

```
ggplot(data = bats)+
  geom_boxplot(aes(x = sex, y = mass))+
  theme_bw()+
  theme(text = element_text(size=18))
```



Look at Your Data - Boxplots

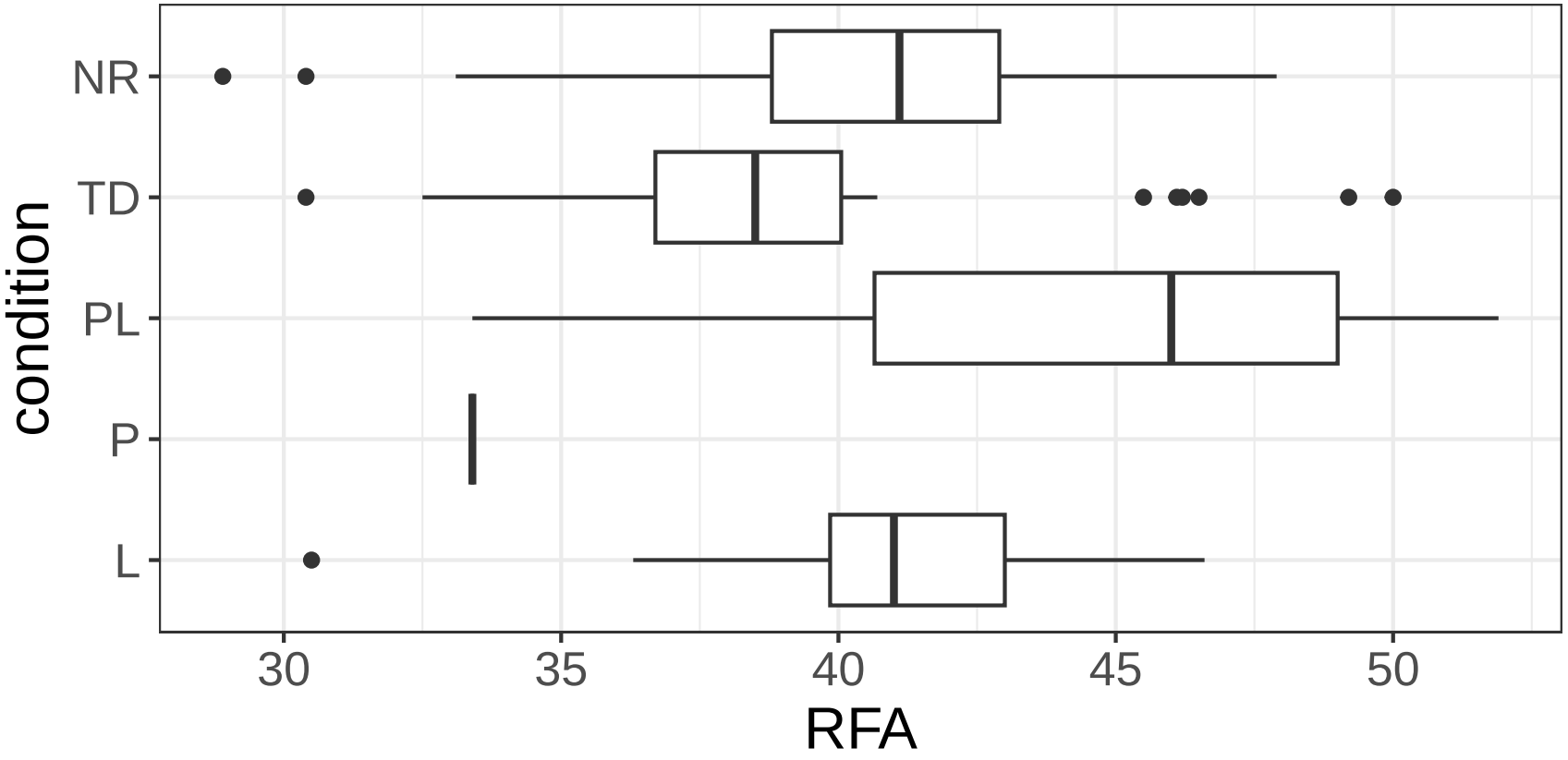
```
ggplot(bats)+
  geom_boxplot(aes(x = condition, y = RFA))+
  theme_bw()+
  theme(text = element_text(size=18))+
  coord_flip()
```



Look at Your Data - Boxplots (re-ordered)

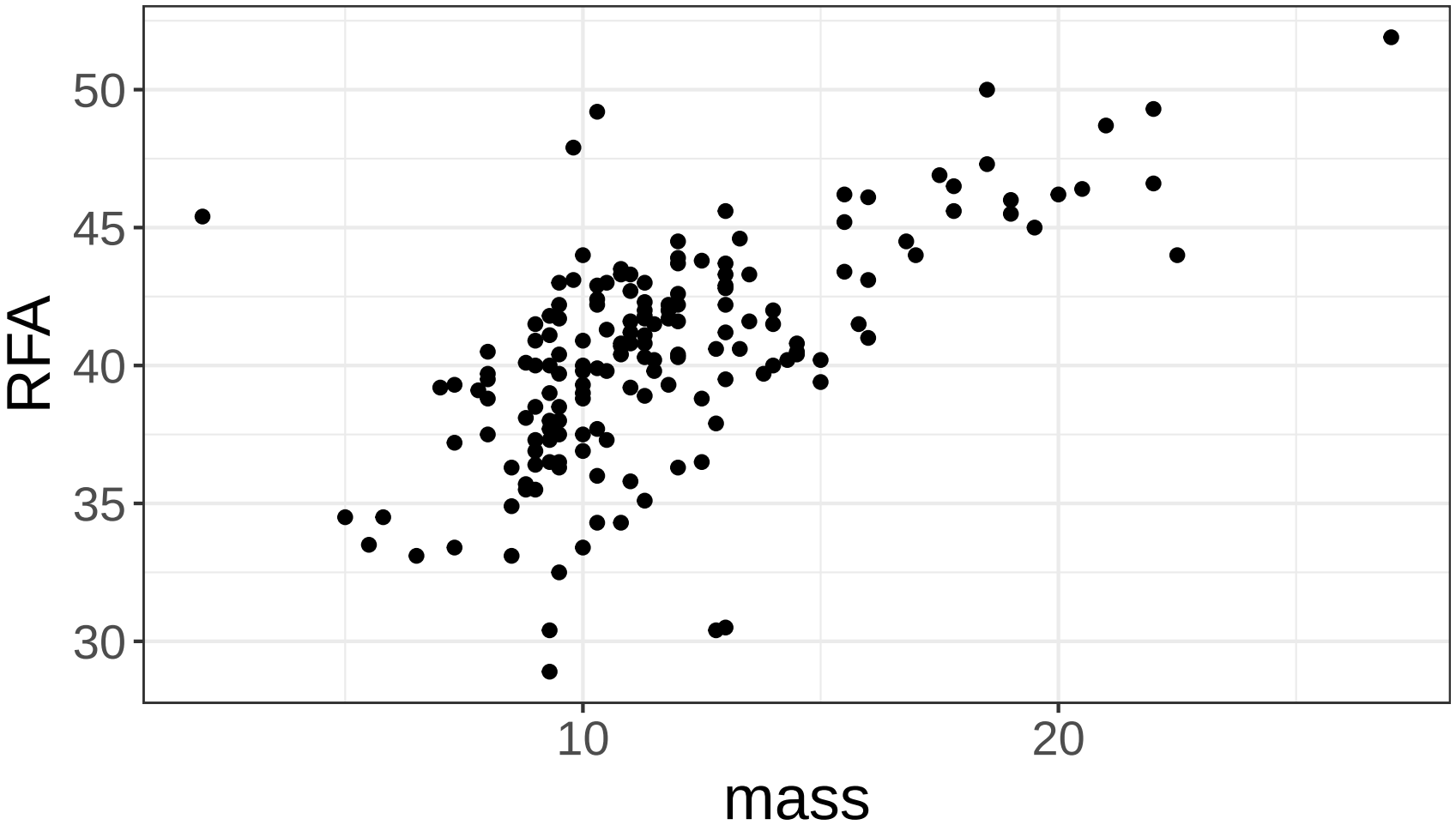
```
bats$condition<-factor(bats$condition, levels =
                        c("L", "P","PL", "TD", "NR"))

ggplot(bats)+
  geom_boxplot(aes(x = condition, y = RFA))+
  theme_bw()+
  theme(text = element_text(size=16))+
  coord_flip()
```



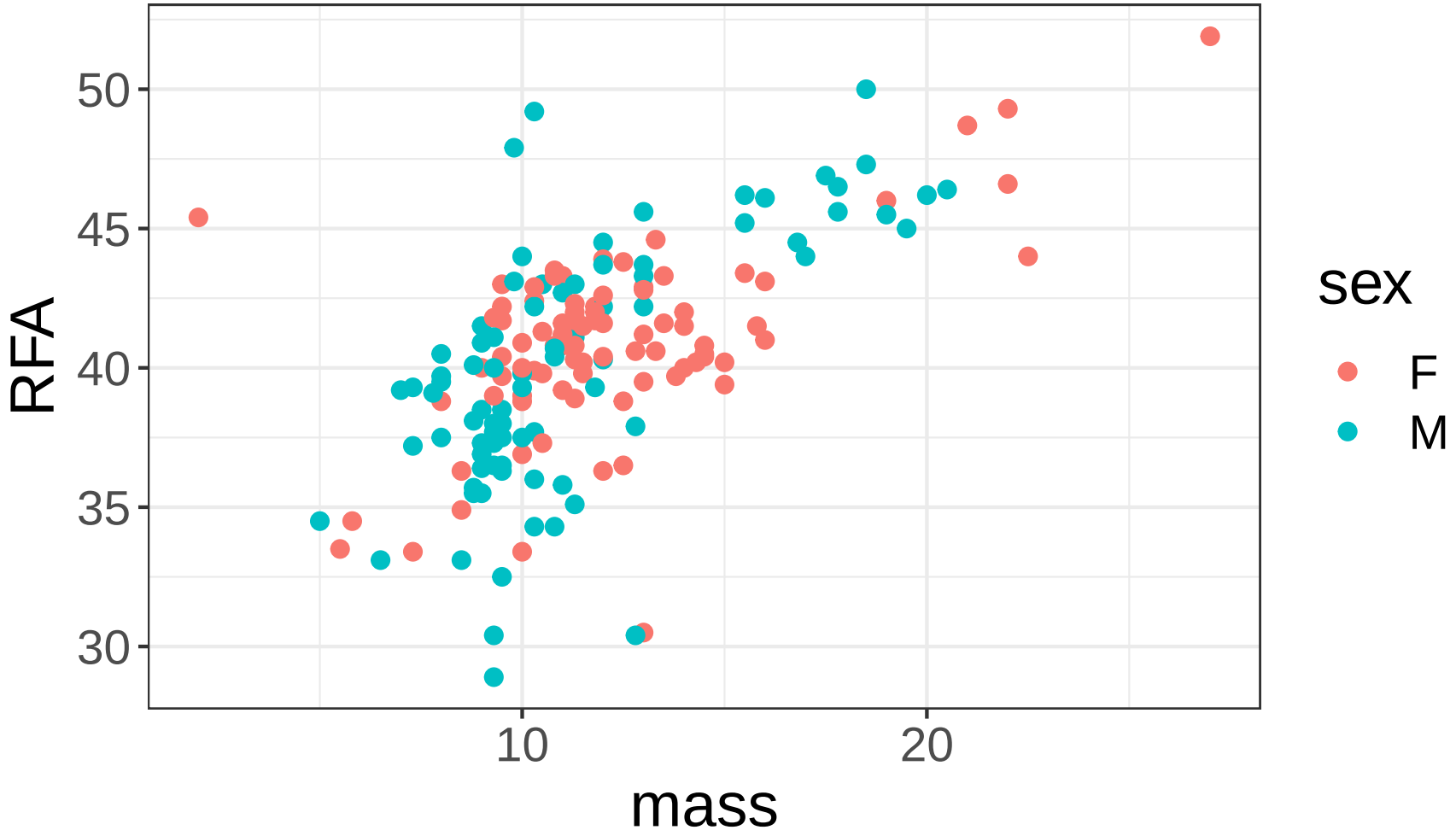
Look at Your Data - Scatterplots

```
ggplot(bats)+
  geom_point(aes(x = mass, y = RFA))+
  theme_bw()+
  theme(text = element_text(size=18))
```



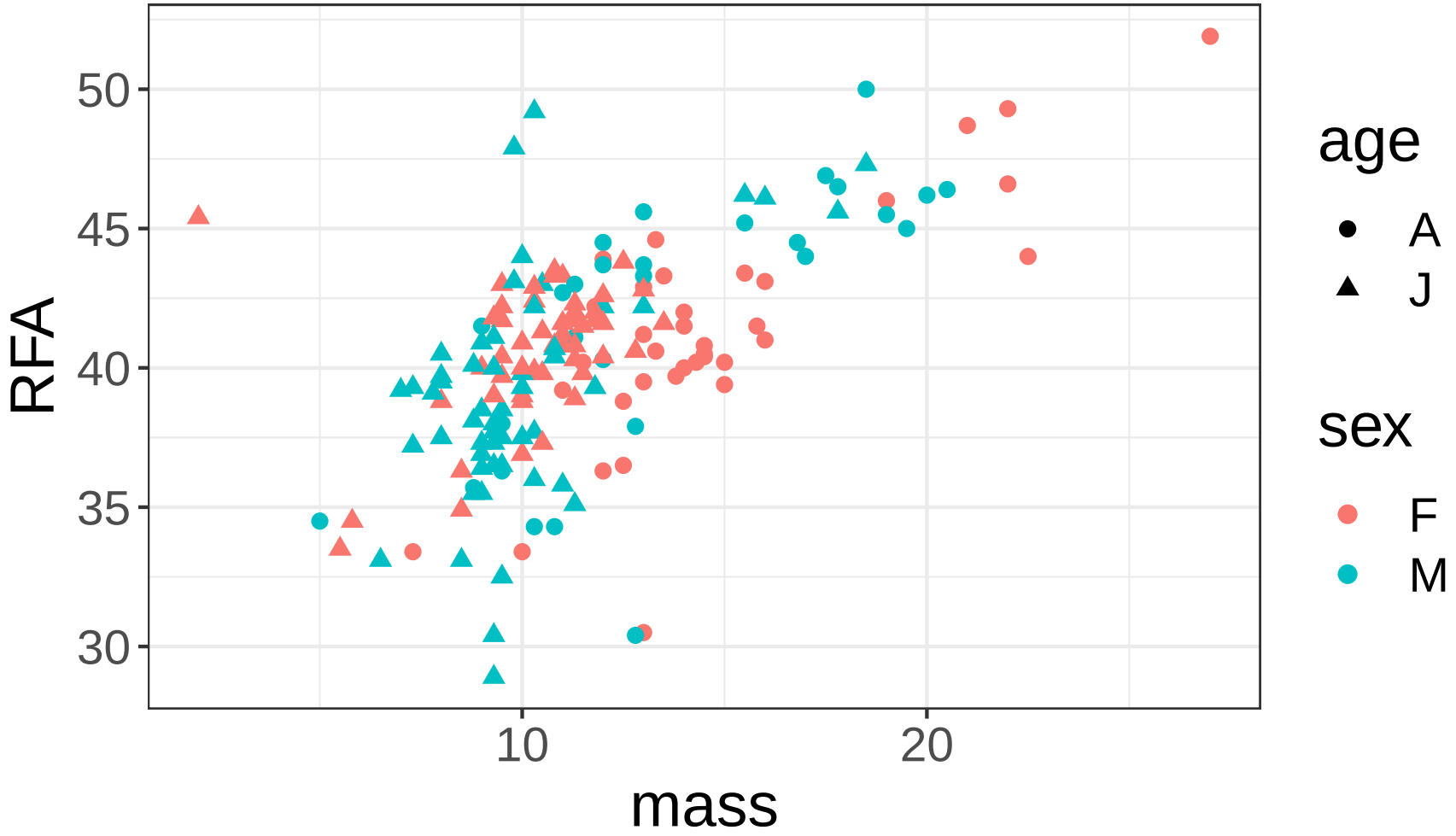
Look at Your Data - Scatterplots

```
ggplot(bats)+
  geom_point(aes(x = mass, y = RFA, color = sex), size = 2)+
  theme_bw()+
  theme(text = element_text(size=18))
```

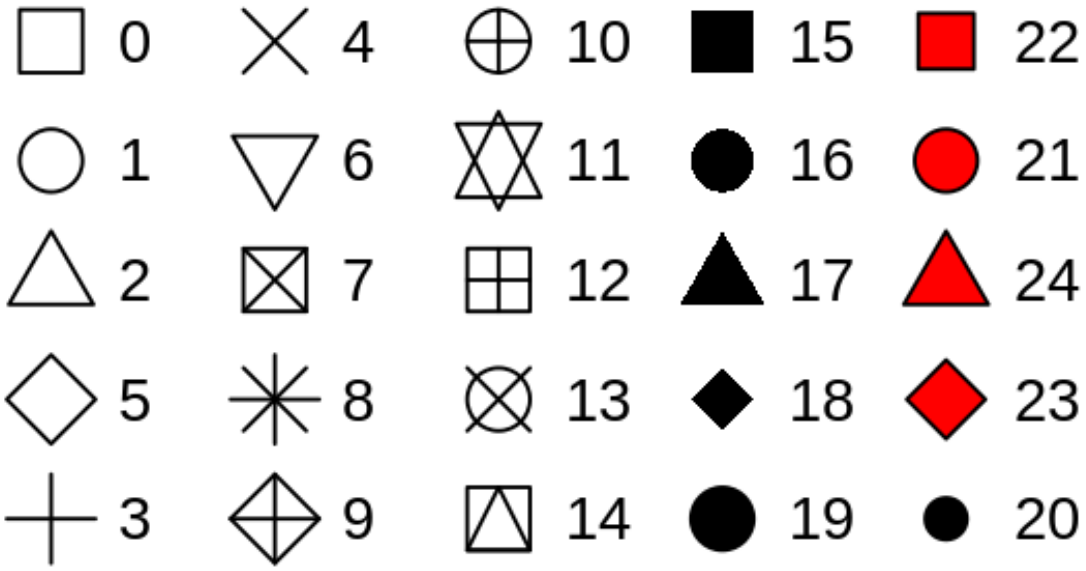


Look at Your Data - Scatterplots

```
ggplot(bats)+
  geom_point(aes(x = mass, y = RFA, color = sex, shape = age), size = 2)+
  theme_bw()+
  theme(text = element_text(size=18))
```

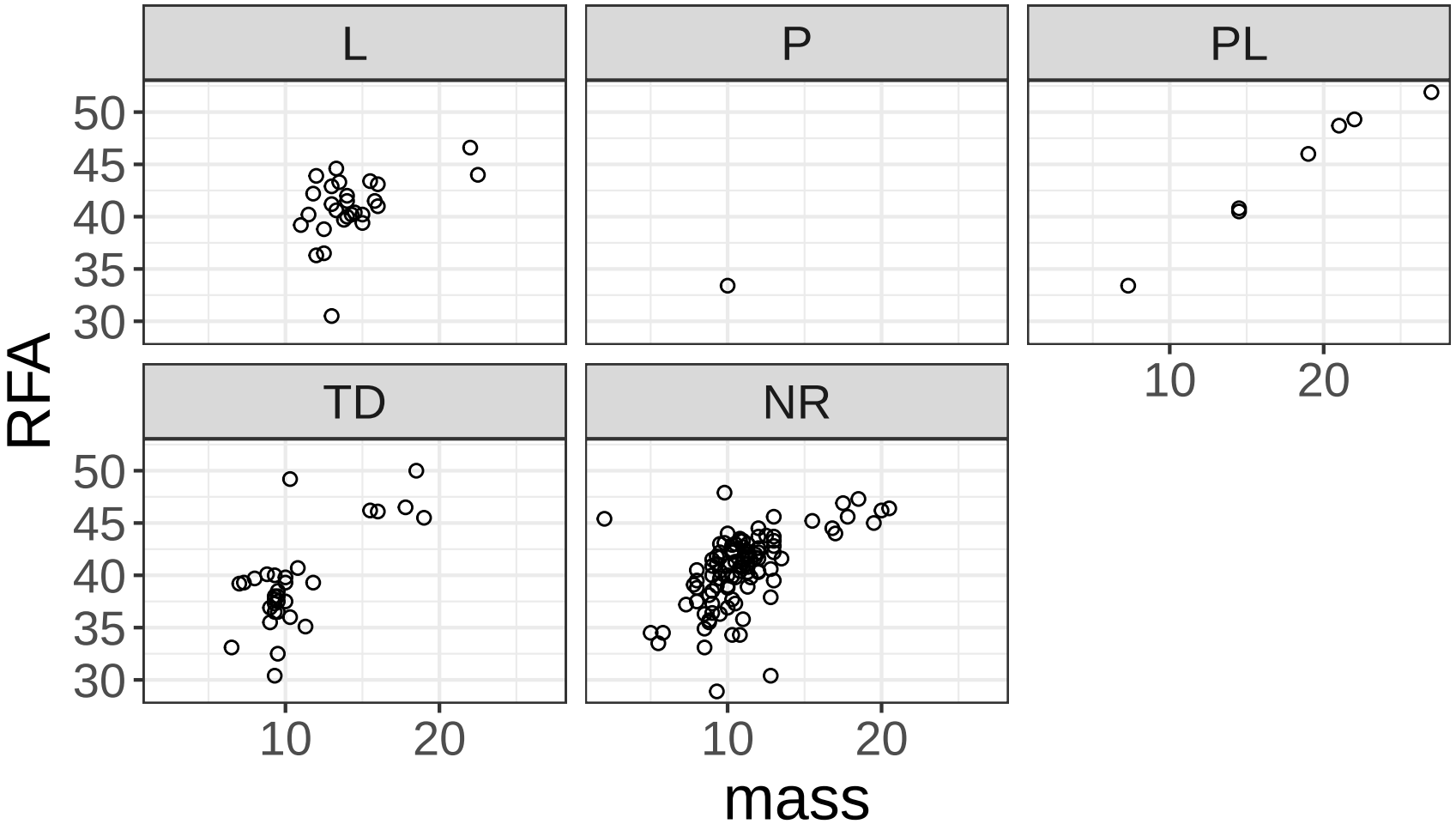


Point Shapes



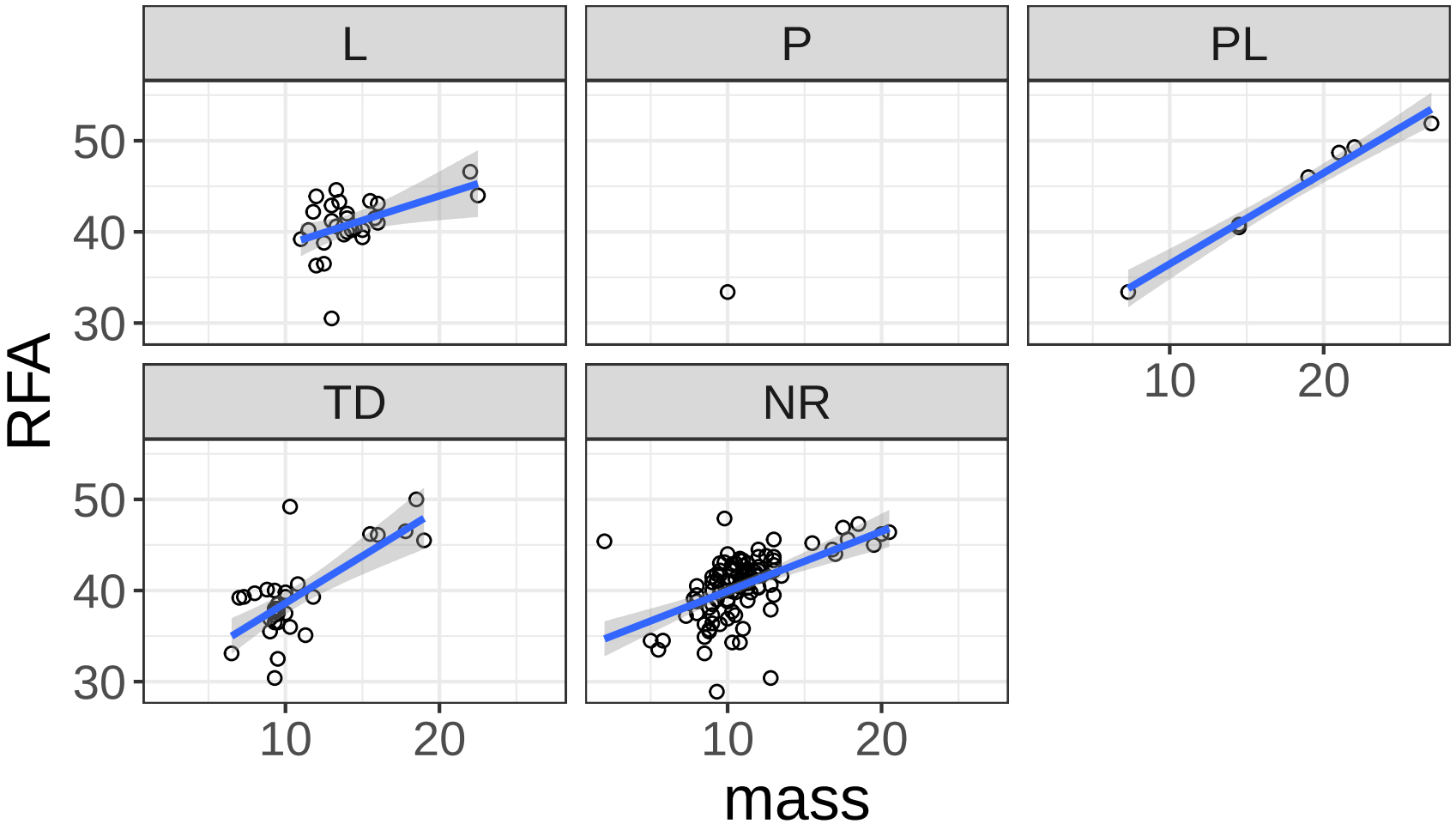
Look at Your Data - Scatterplots + Facets

```
ggplot(bats, aes(x = mass, y = RFA))+
  geom_point(shape = 1)+
  facet_wrap(~condition, nrow=2)+
  theme_bw()+
  theme(text = element_text(size=18))
```



Look at Your Data - Scatterplots + Trends

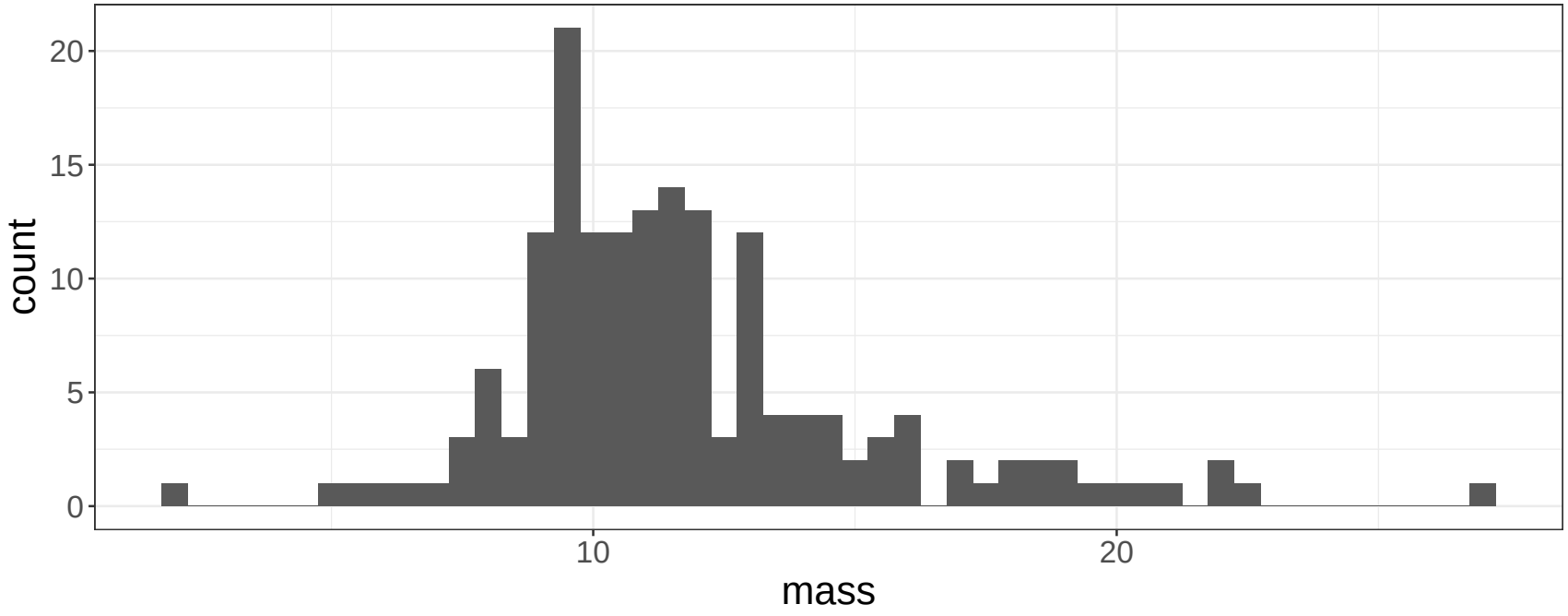
```
ggplot(bats, aes(x = mass, y = RFA))+
  geom_point(shape = 1)+
  geom_smooth(method = lm, se = TRUE)+
  facet_wrap(~condition, nrow=2)+
  theme_bw()+
  theme(text = element_text(size=18))
```



Look at Your Data - Distributions

`geom_histogram()` - allows you to see the distribution of your continuous data

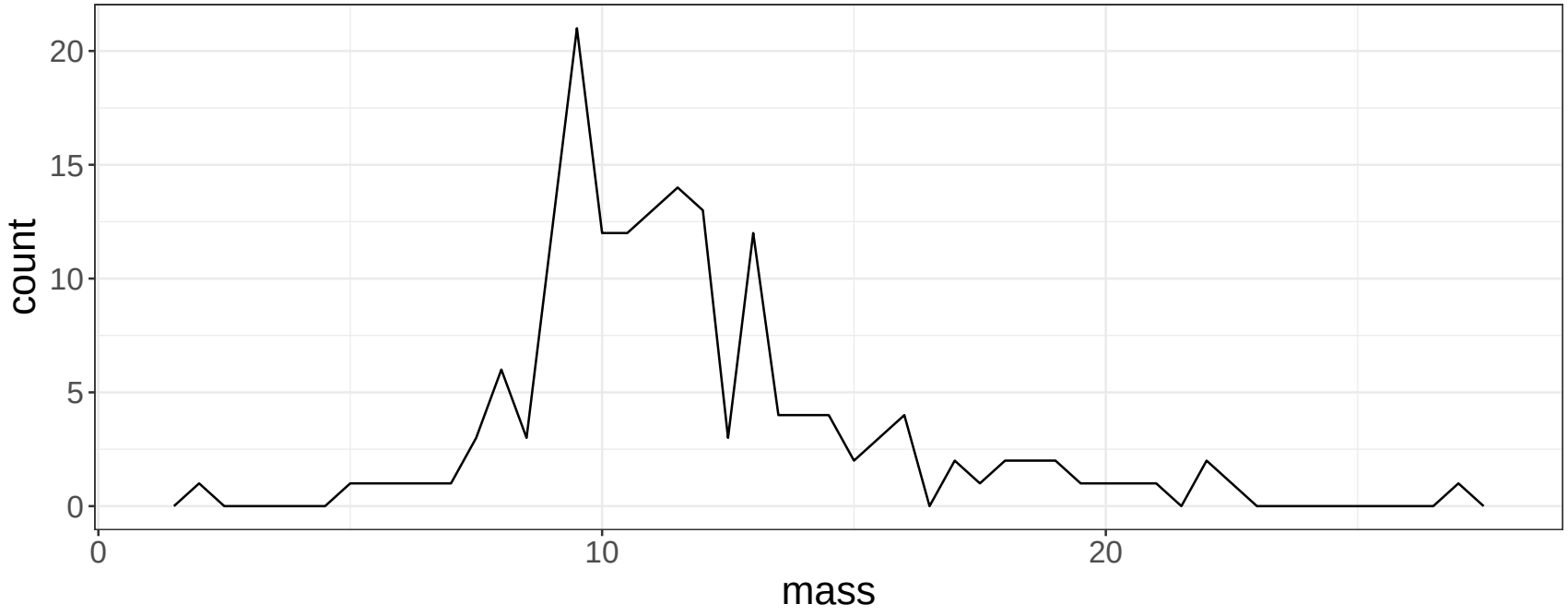
```
ggplot(bats)+
  geom_histogram(aes(x = mass), binwidth=.5)+
  theme_bw()+
  theme(text = element_text(size=18))
```



Look at Your Data - Distributions

`geom_freqpoly()` - same calculations as `geom_histogram()` but displays as lines instead of bars.

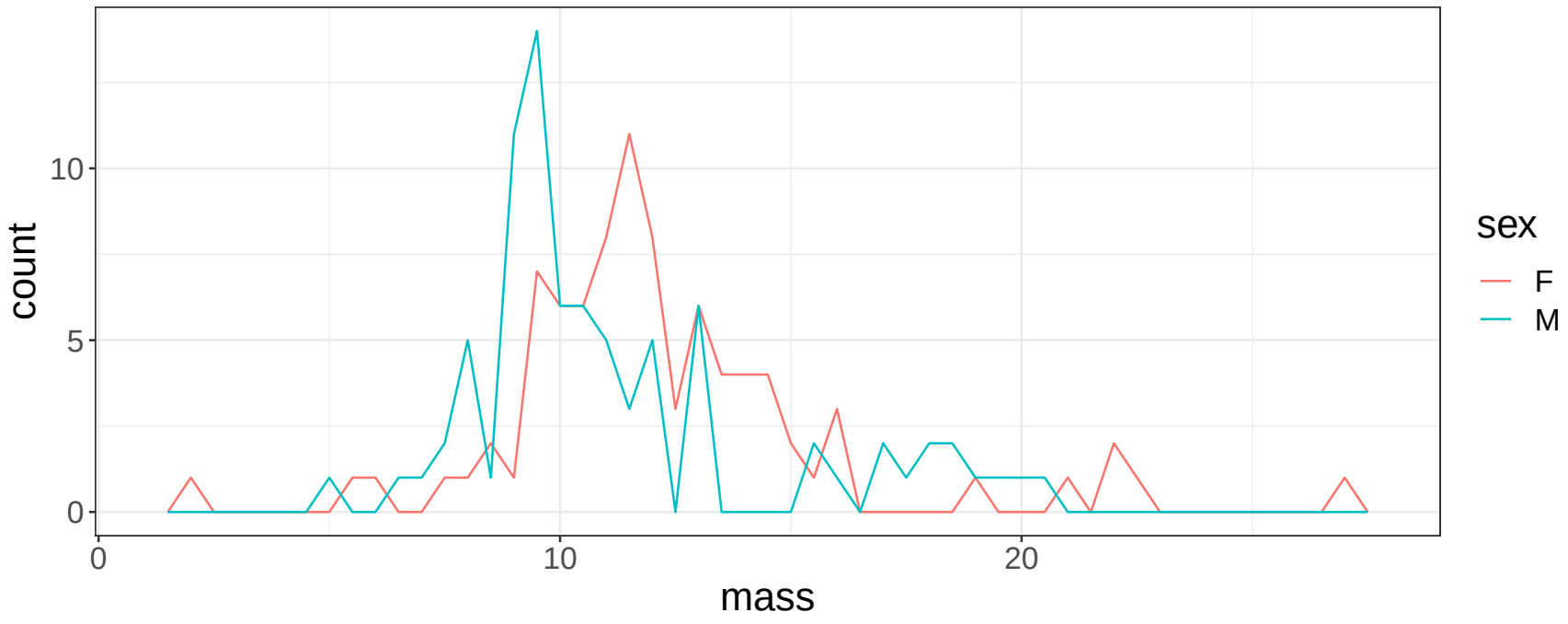
```
ggplot(bats)+
  geom_freqpoly(aes(x = mass), binwidth = 0.5)+
  theme_bw()+
  theme(text = element_text(size=18))
```



Look at Your Data - Distributions

`geom_freqpoly()` - helpful if you want to look at multiple distributions at once.

```
ggplot(bats)+
  geom_freqpoly(aes(x = mass, color = sex), binwidth = 0.5)+
  theme_bw()+
  theme(text = element_text(size=18))
```



Look at Your Data - Distributions

`geom_density()` - normalizes your data instead of counts.

```
ggplot(bats)+
  geom_density(aes(x = mass, fill = sex), alpha = 0.5)+
  theme_bw()+
  theme(text = element_text(size=18))
```