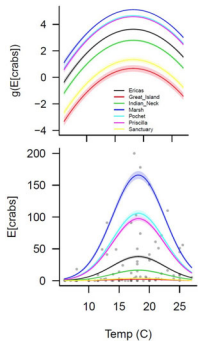
ECO 636
Applied Ecological Statistics

Week 1 – Data exploration – recorded lecture



Meg Graham MacLean, PhD

Department of Environmental Conservation

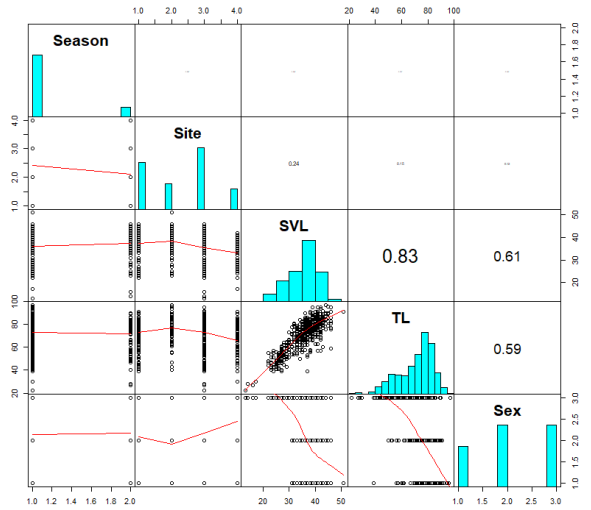
mgmaclean@umass.edu

2021 - Spring

Great!

Today

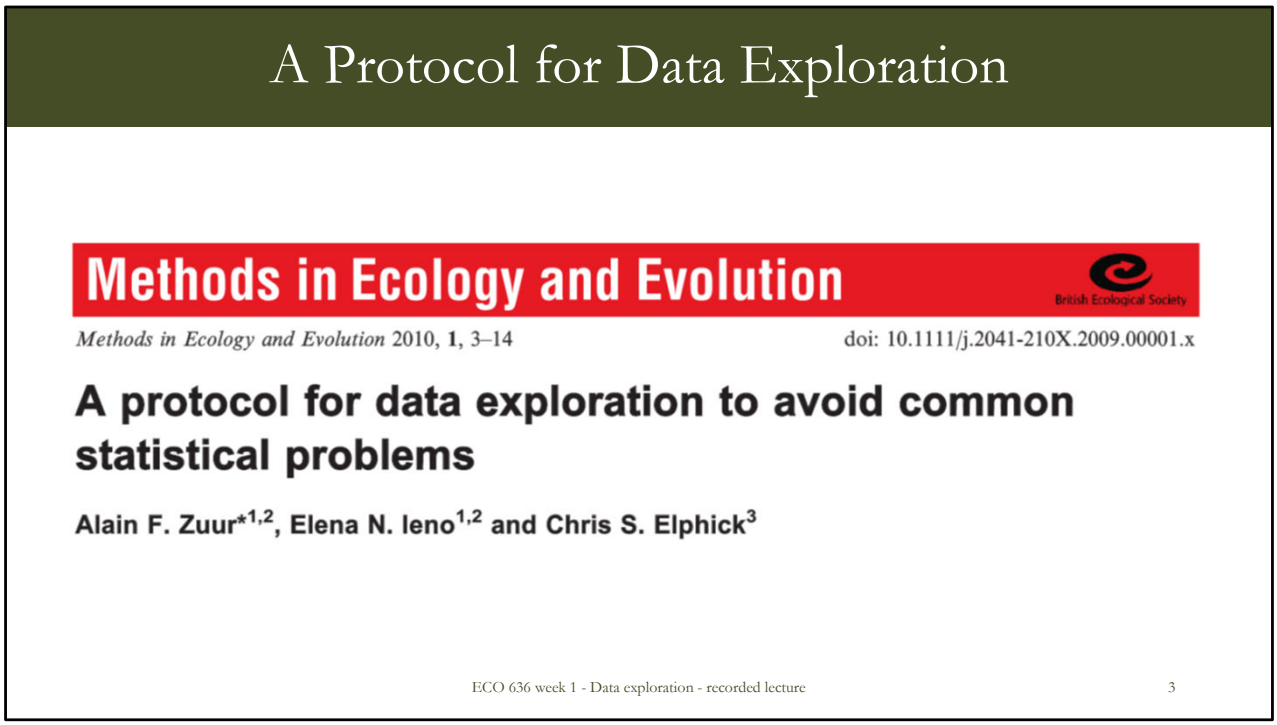
- Data exploration



Today we will be talking about data exploration – which should account for around half the time you spend analyzing your data!

(c) Meg Graham MacLean
mgmaclean@umass.edu

A Protocol for Data Exploration



Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2010, **1**, 3–14 doi: 10.1111/j.2041-210X.2009.00001.x

A protocol for data exploration to avoid common statistical problems

Alain F. Zuur*[1,2], Elena N. Ieno[1,2] and Chris S. Elphick[3]

ECO 636 week 1 - Data exploration - recorded lecture 3

Much of what I will cover today is directly from the posted Zuur paper, so I highly recommend giving that paper a good read before watching this lecture if you haven't already.

(c) Meg Graham MacLean
mgmaclean@umass.edu

A Protocol for Data Exploration

- Formulate a biological/ecological hypothesis & collect data
- Data Exploration
  1. Outliers (Y & X)
  2. Homogeneity (Y)
  3. Normality of errors (Y)
  4. Zero trouble (Y)
  5. Collinearity (X)
  6. Relationships (Y & X)
  7. Interactions (X)
  8. Independence (Y)
- Apply statistical model

Y = response variable
X = explanatory variable, covariates

Great – hopefully that means you have taken a look at that Zuur paper. In the paper, the authors present a useful protocol for data exploration, along with which parts of your data these exploration techniques apply to. I find this to be a great reference when thinking about data exploration! We'll walk through some of the visualizations associated with doing data exploration in this lecture.

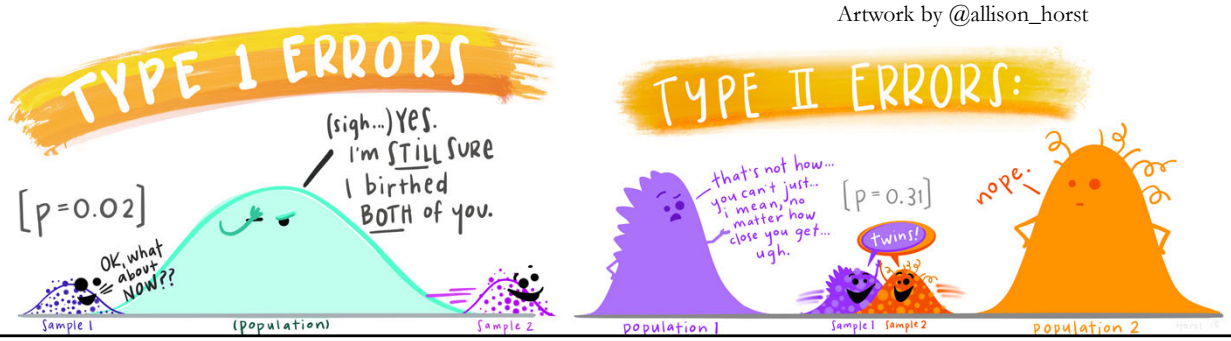
(c) Meg Graham MacLean
mgmaclean@umass.edu

Why Data Exploration?

Garbage In = Garbage Out!

- We are trying to avoid:
  - Type I and Type II errors
  - Results based on only a few influential points (outliers)
  - Ensure data meet model assumptions

Artwork by @allison_horst



But first – why do we do data exploration??

Well, primarily to avoid garbage results!

We want to avoid type I and type II errors – which are illustrated here by the very talented Allison Horst (which, if you haven't checked out some of her stats artwork, I highly recommend them!). Type I errors are when you find a significant difference between samples but in fact they came from the same population, and type II errors are the exact opposite, when you don't have a significant difference between samples when they actually come from different populations.

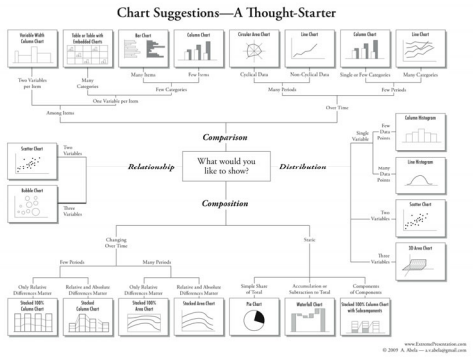
We also want to minimize the influence of outliers and

Ensure the data meet model assumptions when we are fitting a model. So, what how do we do data exploration?

Rapid fire visualizations!

Here are just a few of the many visualizations we can do:

- Boxplots
- Cleveland dotplot
- Scatterplots
- Pair plots
- Coplot
- Lattice graphs



*most of the graphing I do is using the ggplot library – but you can do a lot in base R

Well, one of the first things to do is to visualize the data to start looking for patterns – we will try a few here! We will go over Boxplots, dotplots, scatterplots, pair plots, coplots and lattice graphs today – but I really like this Chart Suggestions image as a way to start thinking about what data visualizations you might want to use in your data exploration. I have posted this image on Moodle too if you would like a closer look.

Most of the data visualization I do is using the ggplot library because it is what I am used to – however, I have some friends who are wizards with base R. In this class, use whichever you are more comfortable with! You will see code for both sprinkled throughout the exercises.

Salamander data

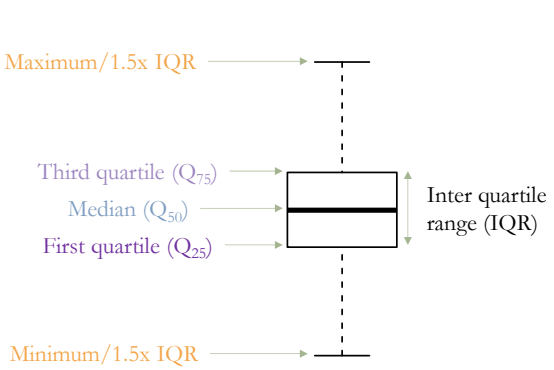
	Season	Site	Occ1	Occ2	SVL	TL	Sex	Cap	Ind
1	Spring	P1A	1	1	43	86	U	N	xxBBP1A
2	Spring	P1A	1	1	33	66	U	N	xYxBP1A
3	Spring	P1A	1	1	42	84	M	N	xYBxP1A
4	Spring	P1A	1	1	36	76	U	N	xYYxP1A
5	Spring	P1A	1	1	44	76	M	N	xxBYP1A
6	Spring	P1A	1	1	42	74	U	N	xBxYP1A
7	Spring	P1A	1	1	40	76	U	N	xYxYP1A
8	Spring	P1B	1	1	37	77	U	N	xxYBP1B
9	Spring	P1B	1	1	23	40	U	N	xxxYP1B
10	Spring	P1B	1	1	39	76	U	N	xxYYP1B
11	Spring	P2A	1	2	43	82	M	N	x(OB)B(OB)P2A
12	Spring	P2A	1	2	36	73	U	N	BBYxP2A
13	Spring	P2A	1	2	33	59	U	N	xBYxP2A
14	Spring	P2B	1	2	37	72	U	N	xBxBP2B

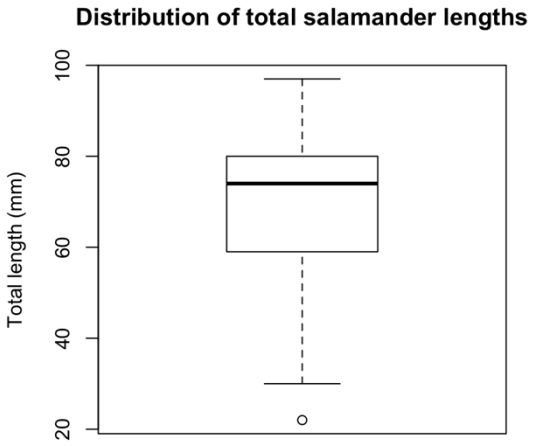
So, let's try some data exploration with some real data – these data were collected by Chris Sutherland, one of the faculty members that used to teach this course. I have uploaded the data to this week's Moodle section so you can try to recreate all of the plots yourself as practice. If you feel like you have R nailed and are confident in your ability to recreate all of the following graphs, no need to practice, but if you feel like you could use a bit more time doing data visualization in R, I highly recommend downloading the data and trying this for yourself. The code is all posted on these slides.

The posted salamander data includes information about the season the observations were collected, the site at which the salamanders were found along with other identifying features, but for today the important variables of note are SVL or snout to vent length, TL or total length, and sex. Alright – so lets do some data exploration!

(c) Meg Graham MacLean
mgmaclean@umass.edu

Boxplot






```
boxplot(manders$TL, ylab = "Total length (mm)")
title("Distribution of total salamander lengths")
```

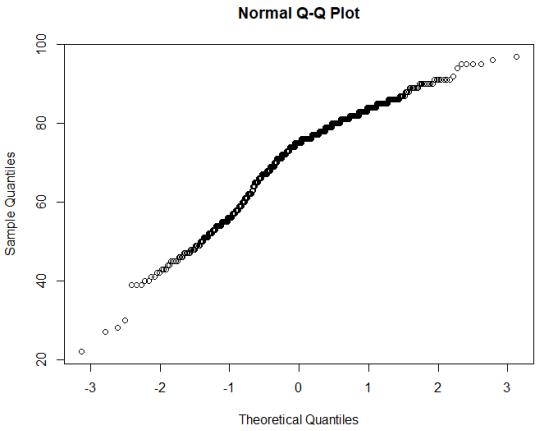
ECO 636 week 1 - Data explor

First, lets check out a boxplot. Boxplots are great for looking at the distribution of the data, as well as whether or not there are observations worth checking as potential outliers. As a reminder – the box of the boxplot represents the distance from the first to the third quartile, with the line in the middle representing the median. When the data are relatively normally distributed – the median is close to center in the box.

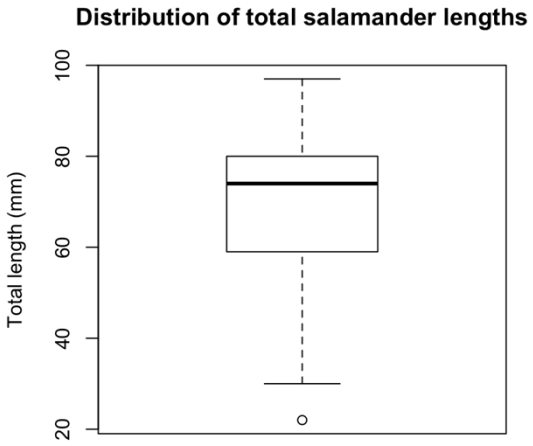
The whiskers show the difference from the box to 1.5x the interquartile range, or the min or max, whichever comes first. The dots plotted outside of the whiskers are the observations that fall outside of that 1.5x range from the box and are often the observations that should be checked as potential outliers.

Here I plotted the total length of the salamanders, which for these examples I'll pretend is the response variable, or the thing I am interested in predicting. They don't look super normally distributed in this boxplot – with at least one observation worth checking as an outlier.

Boxplot




```
qqnorm(manders$TL)
```

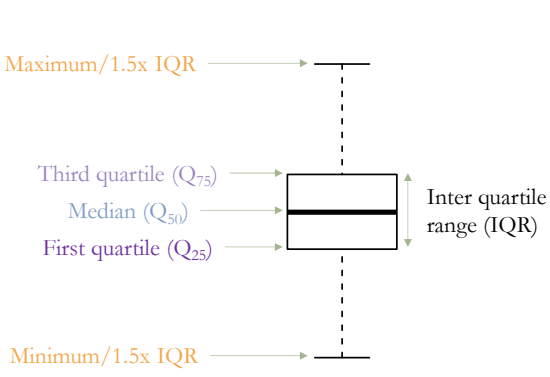


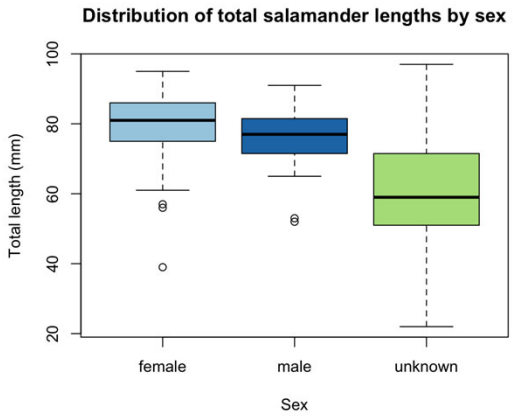

```
boxplot(manders$TL, ylab = "Total length (mm)")
title("Distribution of total salamander lengths")
```

ECO 636 week 1 - Data explor

Another, and more common method for checking the distribution of the residuals of the response variable, is to use a Q-Q plot, which we will go over in lots more detail later in the semester. For a quick definition, the Q-Q plot plots the sample or actual observed quantiles of the response data on the y, with theoretical quantiles from a normal distribution on the x. Data that are perfectly normally distributed would form a straight diagonal line from the lower left of the plot to the upper right. In this plot of total length you can see that it has a bit of a hump in the middle, again showing that these data are likely not normally distributed.

Boxplot



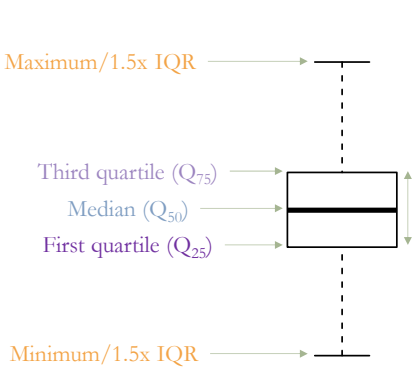


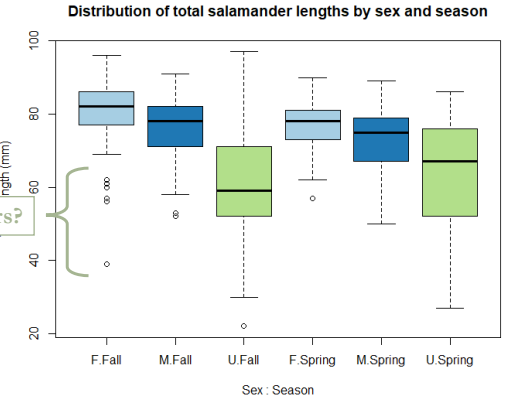

```
boxplot(TL ~ Sex, data = manders,
        ylab = "Total length (mm)",
        col = c("#a6cee3", "#1f78b4", "#b2df8a"))
title("Distribution of total salamander lengths by sex")
```


Another type of boxplot that is super helpful is the conditional boxplot, where we can compare the distributions of the observations within groups. This is a way we can quickly look for homogeneity of variance across groups in the response variable. For example, here we are looking at total lengths by sex and we can see that the variance is not similar across groups, with the distribution of the total lengths for the salamanders without a known sex covering a much larger range.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Boxplot






```
boxplot(TL ~ Sex + Season, data = manders,
        ylab = "Total length (mm)",
        col = c("#a6cee3", "#1f78b4", "#b2df8a"))
title("Distribution of total salamander lengths by sex and season")
```


Similarly, we can check the distribution of total length by sex and season and start to see that there might be some interesting observations to check as potential outliers in some of these groups. So, how do we check them?

(c) Meg Graham MacLean
mgmaclean@umass.edu

Outliers

Know your data!

- Data points that can be considered extreme relative to the data
  - Can disproportionately influence analyses
  - Can be errors (human or otherwise) or naturally extreme values
  - Identify outliers using plots!
- You can use Cook's distance to *help* identify outliers (doesn't mean they are!)
  - Just tells you how influential each point is

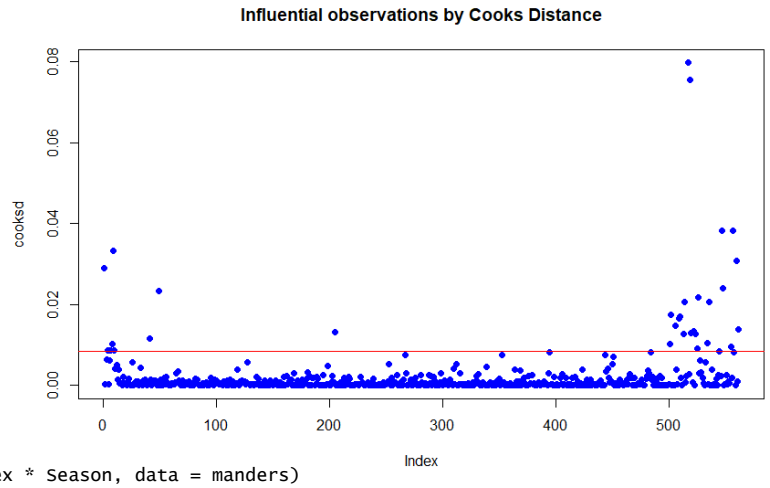
$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_i - \hat{Y}_{j(i)}\right)^2}{p \times MSE}$$

Most importantly, you really need to know and understand your data! This will help you identify if a point is an outlier because of an error somewhere in the data collection, or may actually reflect something of interest happening in your system that you happened to record (sometimes can be a super awesome discovery!).

Outliers by definition are observations that are considered extreme relative to the rest of the data, and these points can disproportionately influence your analysis. Often you can identify these points using graphs – some of which we will explore in a moment.

You can also use Cook's distance to help identify outliers based on how influential they are to your model. However, just because a point is influential doesn't mean it is an outlier, but if a point is identified as a potential outlier through data visualization and then is also quite influential, it is probably a good idea to figure out if this point should stay in your data by determining if it reflects a true phenomena that is part of what you are trying to model.

Outliers



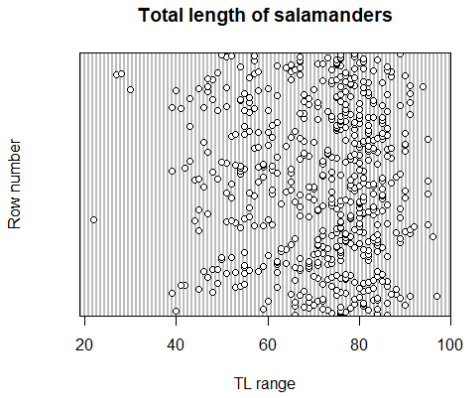

```
mod <- lm(TL ~ Sex * Season, data = manders)
cooksd <- cooks.distance(mod)
plot(cooksd)
abline(h = 4*mean(cooksd, na.rm=T), col="red")
```

1 - Data exploration - recorded lecture 13

You can fairly simply compute the Cook's distance for your data (I did here for the total length data) using R, however, you must have defined your model to determine how influential each point is! Here is a great example of how data exploration can be very iterative with model building. In this example, I created a linear model where I tried to predict total length based on the sex of the salamander and the season in which it was caught. I then plotted the Cook's D or the measure of the influence of each observation on that model with the red line represented 4 times the mean Cook's D, a widely used cut-off. You'll notice that many of the observations are above the red line, but not all of these observations are considered potential outliers.

Cleveland dotplot

Good for looking for outliers and homogeneity of variance

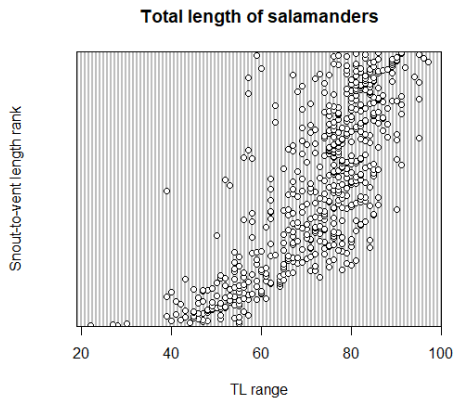



```
dotchart(manders$TL,
    ylab = "Row number",
    xlab = "TL range")
title("Total length of salamanders")
```


Another method for trying to identify potential outliers is to use a Clearvland dotplot and look for values that are far away from the main mass of observations.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Cleveland dotplot

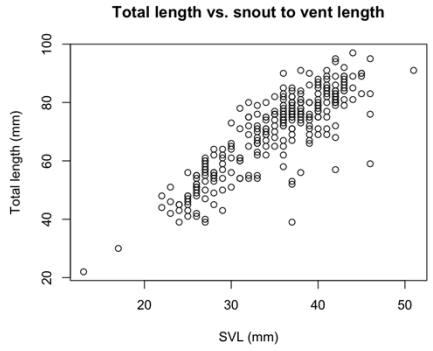



```
dotchart(manders$TL[order(manders$SVL)],
    ylab = "Snout-to-vent length rank",
    xlab = "TL range")
title("Total length of salamanders")
```

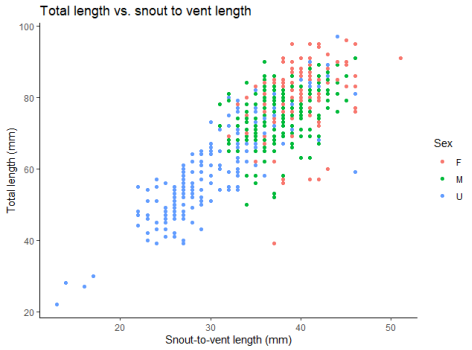

Ordering the dotplot by some explanatory variable can also help make potential outliers more obvious.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Scatterplots



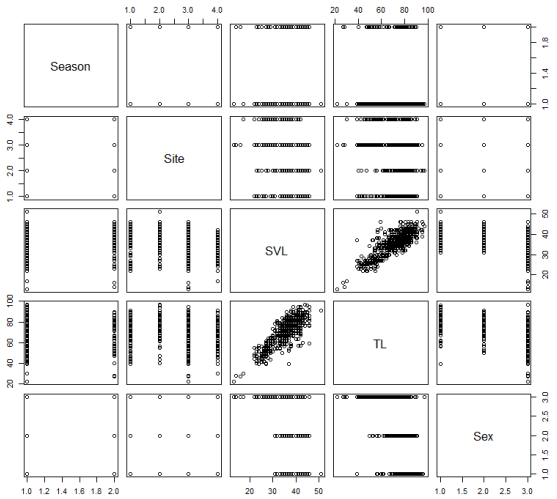

```
plot(manders$SVL, manders$TL,
     main = "Total length vs. snout to vent length",
     ylab = "Total length (mm)",
     xlab = "SVL (mm)")
```


```
ggplot(data = manders, aes(x = SVL, y = TL, color = Sex)) +
  geom_point() +
  labs(title = "Total length vs. snout to vent length",
       x = "Snout-to-vent length (mm)",
       y = "Total length (mm)") +
  scale_fill_manual(values = c("#a6cee3", "#1f78b4", "#b2df8a")) +
  theme_classic()
```

Another great data exploration tool, especially for exploring the collinearity between variables, is the scatterplot. You can also add fun colors or shapes to help plot different groups or categorical variables along with the continuous variables.

Pair plots



Should be done for every analysis for:

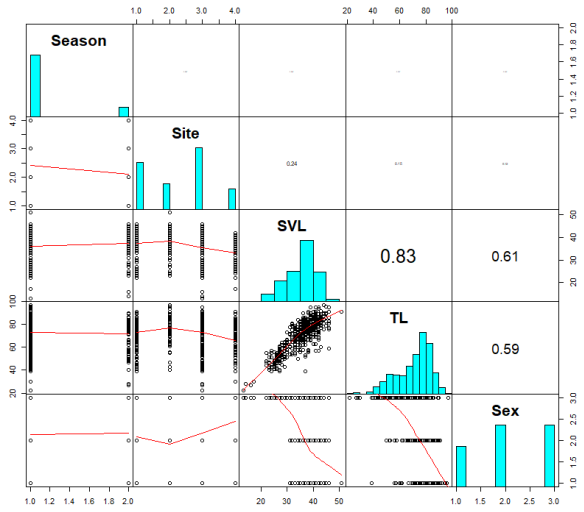
- All variables >1
  - Especially important for multivariate analyses (more than one response variable)
  - For identifying collinearity issues
  - For exploring relationships between response and explanatory variables

```
pairs(manders[, c(1,2, 5:7)])
```


A fancy, and I find more useful method for creating scatterplots is to do a pair plot, where you plot the relationship between all variables in one quick and simple R line. This should be done for every analysis since it is so useful for giving you insights into your data!

(c) Meg Graham MacLean
mgmaclean@umass.edu

Pair plots



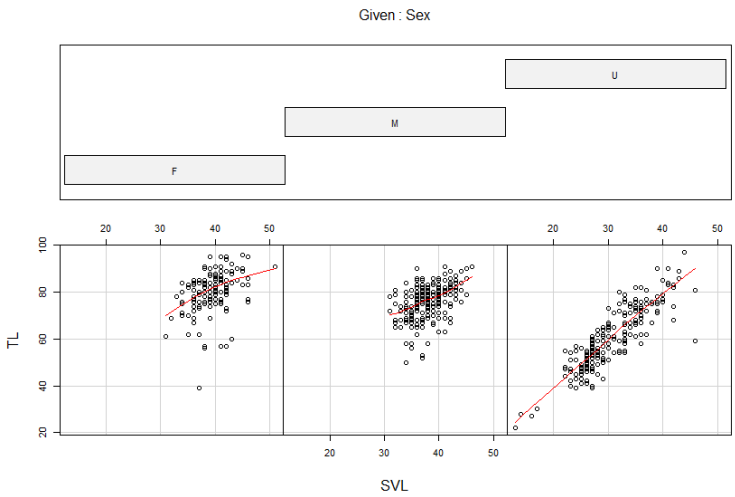

```
## put histograms on the diagonal
panel.hist <- function(x, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
  h <- hist(x, plot = FALSE)
  breaks <- h$breaks; nB <- length(breaks)
  y <- h$counts; y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}

## put (absolute) correlations on the upper panels,
## with size proportional to the correlations.
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste0(prefix, txt)
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}
pairs(manders[, c(1,2, 5:7)], diag.panel = panel.hist,
      cex.labels = 2, font.labels = 2,
      lower.panel = panel.smooth, upper.panel = panel.cor,
      gap=0, row1attop=TRUE)
```


You can also make sure fancy pair plots that integrate histograms of the data in each of the variables, as well as trend lines and correlation coefficients to help you identify issues of collinearity between explanatory variables. If you can, use this code to give this plot a try! I hope you save the code somewhere, since I find it useful for most studies.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Coplot

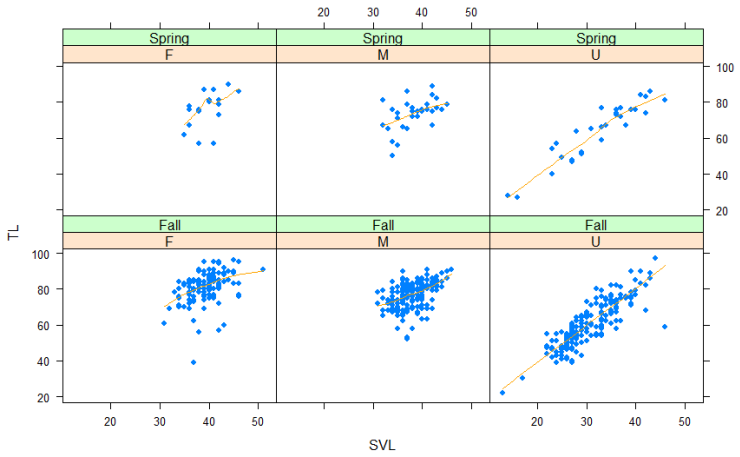



```
coplot(TL ~ SVL | Sex, data = manders, rows = 1, panel = panel.smooth)
```


Co-plots are a simpler separated version of the colored scatterplot we saw earlier, where we can look at the different variables within defined groups and the interactions between these groups.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Lattice graphs




```
library(lattice)
xyplot(TL ~ SVL | Sex * Season, data = manders, pch = 16,
       type = c("p", "smooth"), col.line = "orange")
```


An even fancier version of the co-plot is lattice graphs, where we can divide the data into even further groups. For example, in this plot we can see the observations and relationships between SVL and TL for each sex in each season.

A Protocol for Data Exploration

- Formulate a biological/ecological hypothesis & collect data
- Data Exploration

1. Outliers (Y & X)	boxplot & Cleveland dotplot
2. Homogeneity (Y)	conditional boxplot
3. Normality of errors (Y)	histogram or QQ-plot
4. Zero trouble (Y)*	
5. Collinearity (X)	scatterplots, correlations, VIF*
6. Relationships (Y & X)	pair plots
7. Interactions (X)	coplots
8. Independence (Y)*	

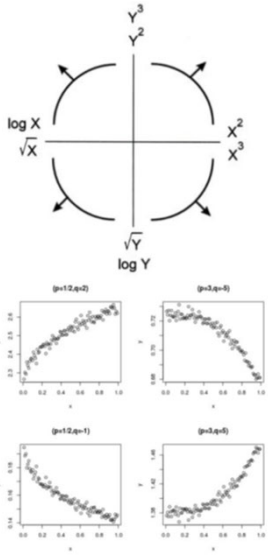
- Apply statistical model

*we will chat about these more later

So, we have just very quickly gone through a whole bunch of data visualizations, and each of those visualizations can be used to do different sorts of data exploration. While many can serve more than one purpose, I put the names of different visualizations next to the exploration type it is most often used for. One note is that we didn't look at zero trouble, independence, or what VIF or variance inflation factor is, all of which are mentioned in the Zuur paper. But not to worry, we will get to those later on in the semester!

What happens when you find an issue?

- Fix any issues where you can (data entry, etc.)
- Remove data points that are truly unreasonable (and you can't fix)
  - Report any data that has been removed!
- Apply a transformation (to reduce the effect of outliers or improve linearity or normality)
  - NOT recommended if you can use another method! (we will get to some later)
  - Usually through a trial and error approach
    - Mosteller and Tukey's bulging rule



Finally, what happens if you actually find an issue after all of your data exploration?!

Well – first, fix any issues that are easy to fix, like fixing data entry issues. If you just transcribed the data wrong from a data sheet to a table – hooray – easy fix!

Second, remove any data points that are truly unreasonable, but you can't fix, and keep a record of the removal and report the removal in any publication or other use of the "cleaned up" data

Last, you can apply a transformation or standardization to the data – but only do this if you have no other options. We will talk about potential other methods later on in the semester! So, don't transform unless you *have* to. Transformations are usually applied through trial and error, again why they aren't always recommended, but you can use the Mosteller and Tukey bulging rule, pictured here at the right, to help you get started on which transformation might work best for your data. The bulge of the data can help you figure out how to transform your data using these rules.

(c) Meg Graham MacLean
mgmaclean@umass.edu

What happens when you find an issue?

Standardization can sometimes help to ensure all variables have equal contributions to the model when you have variables of different scales

$$y_i^{new} = \frac{(y_i - \bar{y})}{S_y}$$

- y_i is the ith observation
- y_i^{new} is the centered value for the ith term
- $\bar{y}$ is the mean
- S_y is the sample standard deviation

Standardization methods can also help ensure that all of your variables have equal contribution to the model if they are in vastly different units or scales between variables. Traditional standardizations subtracts the mean from each observation and then divides by the standard deviation. In this way each explanatory variable will all be unitless and in the same ranges.

(c) Meg Graham MacLean
mgmaclean@umass.edu

Summary

Know your data!

- Identify any issues
  - Outliers
  - Non-normal distributions
  - Non-linear relationships
- Resolve any issues
  - Fix/edit data as needed
  - Transformations (only if there aren't other good options)
  - Standardizations
- Data exploration should take >50% of the time of your analysis!

Awesome – that's the end of the first recorded lecture and it was a bit of a doozy! Really, the moral of the story is know your data! Data exploration should take at least half of your analysis time, because it is super important to make sure you know what you are looking at and then interpreting from your results. I hope you will use the posted Salamander data to try some of the data visualizations on your own. Good luck and let me know if you run into any trouble!

(c) Meg Graham MacLean
mgmaclean@umass.edu

For this week:

For Wednesday:

1) Please familiarize yourself with our Moodle page
2) Start the "Introduction to R" workbook (on Moodle and contains installation instructions for R and R Studio!)
3) Amanda will be at lab to help if you need it!

For Thursday:

1) Read the Zuur et al. 2010 paper posted in the Week 1 section

For Tuesday (next week):

1) Read chapter 5.1 from Zuur et al. 2007 book (posted on Moodle)
2) Answer the individual evaluation questions on Moodle

Thanks and see you on Tuesday!

(c) Meg Graham MacLean
mgmaclean@umass.edu