Agenda

1. Introduction to R
2. Basic plots using ggplot2
3. Check moodle for reading assignments
4. Assignment #1 due Thursday (2/9) by noon.

Introduction

- R is a programming language specfically designed for statistical analysis. R is open-source, and is developed by a team of statisticians and programmers in both academia and industry. It is updated frequently and has become the industry standard. In the data science realm, alternatives to 'R' include Python with the Pandas library, and Julia. In the statistics realm, alternatives include SAS, Stata, and SPSS.
- RStudio is an integrated development environment for R. RStudio is also open-source software, and depends upon a valid R installation to function. Before RStudio, people used R through the command line directly, or through graphical user interfaces like RKwrd and Rcmdr, but RStudio is so vastly superior that these alternatives have few users left. RStudio employees are important drivers of R innovation, and currently maintain the `rmarkdown`, `knitr`, and `dplyr` packages, among others.
- R Markdown: is a syntax for composing relatively simple documents that combine R code and text. R Markdown is an extension of markdown (a general-purpose authoring format) that provides functionality for processing `R` code and output.

Basic data graphics in `R` There are three prominent graphics libraries in `R`:

- `graphics`: often called **base** graphics, these are the drawing methods that come pre-installed with `R`. These graphics are the most commonly-used, but often the least user-friendly. (e.g. `plot()`
- `lattice`: a nice-looking and powerful graphics library that is particularly adept at making multivariate comparisons. `lattice` graphics are very convenient and easy-to-learn for most common statistical plots, and are the default for most of the `mosaic` graphing functions. Customization of `lattice` graphics often involves writing `panel.functions` – which can be tricky, but powerful. (e.g. `xyplot()`)
- `ggplot2`: a very popular graphing library maintained by Hadley Wickham, based on his "Grammar of Graphics" paradigm. Unlike `lattice`, `ggplot2` uses an incremental philosophy towards building graphics.

Categorical vs. Quantitative Variables

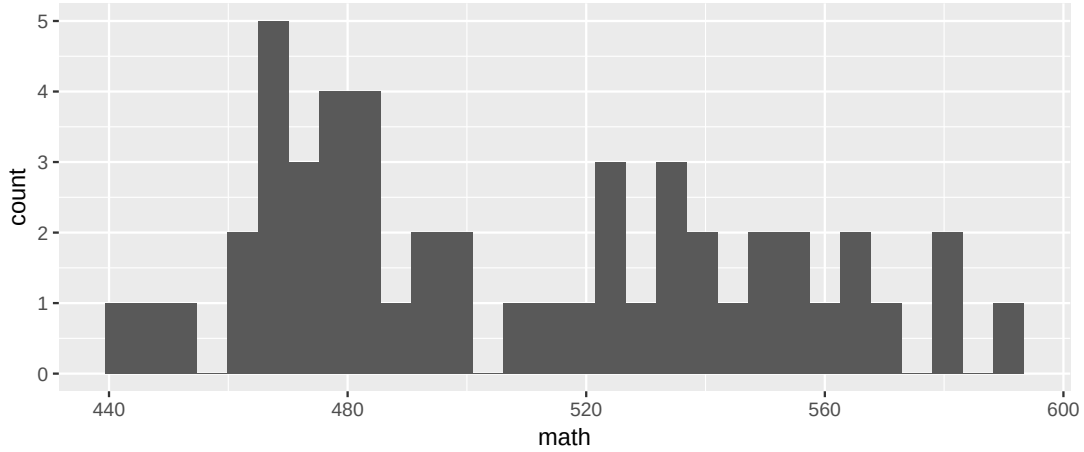
- It is important to be able to recognize the type of variables that you are working with so that adequate graphical displays can be created
- A quantitative variable is measured on a numeric scale, such that each numerical value represents an amount (length of time, distance, speed, temperature, number of children)
- A qualitative (or categorical) variable is a variable that can be classified into groups (employment status, race, religion)

Univariate displays

- A univariate distribution describes what values one paritcular variable takes and how often it takes them.
- Histograms are used to show the distribution of one quantitative variable.

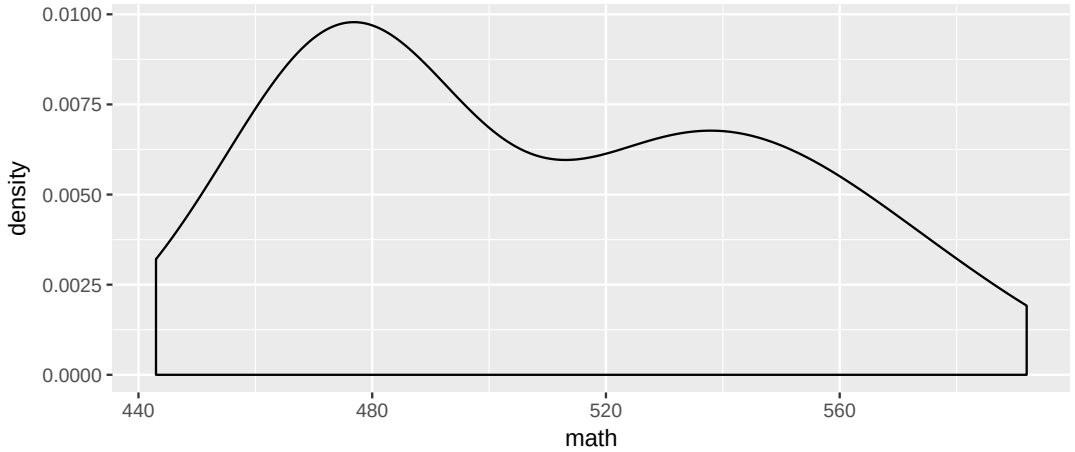
```
require(ggplot2)
require(mosaicData)

ggplot(data=SAT)+
   geom_histogram(aes(x=math))
```



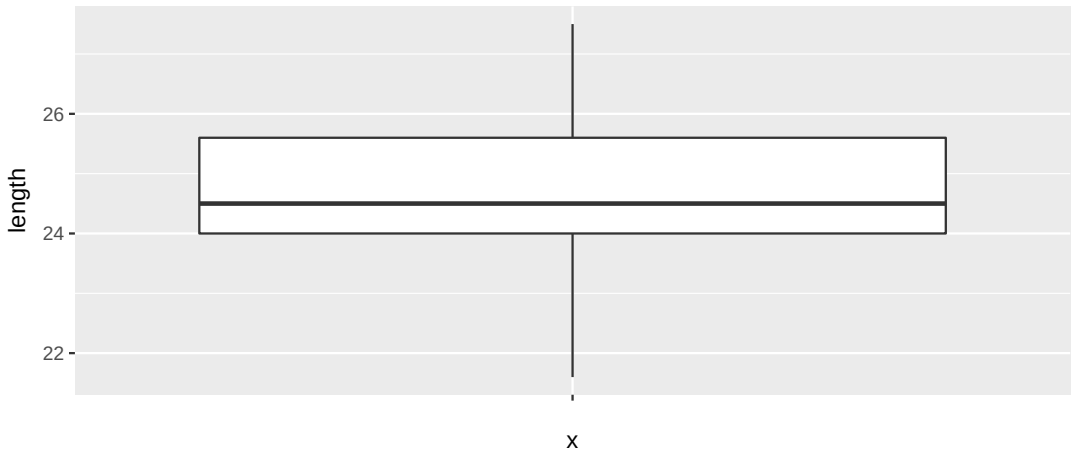
- Density plots are very similar to histograms, except they show the smoothed distribution of one quantitative variable.

```
ggplot(data=SAT)+
  geom_density(aes(x=math))
```



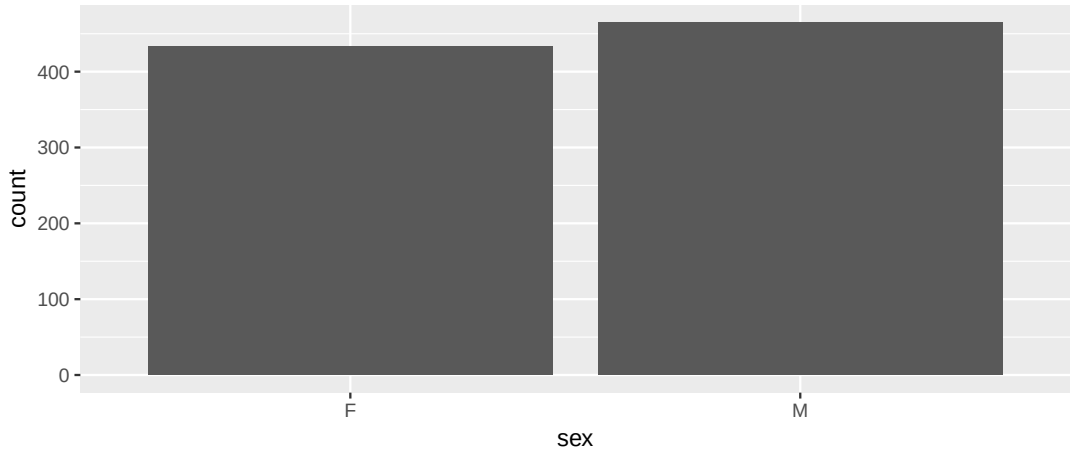
- Box plots (or box-and-whisker plots are also used to describe the distribution of one quantitative variable. It gives summarized information about the variable. It allows us to see the min, max, and median of a particular variable (along with the first and third quartiles).

```
ggplot(data=KidsFeet)+
  geom_boxplot(aes(x="", y=length))
```



- Traditional bar graphs are used to summarize a single qualitative (categorical) variable.

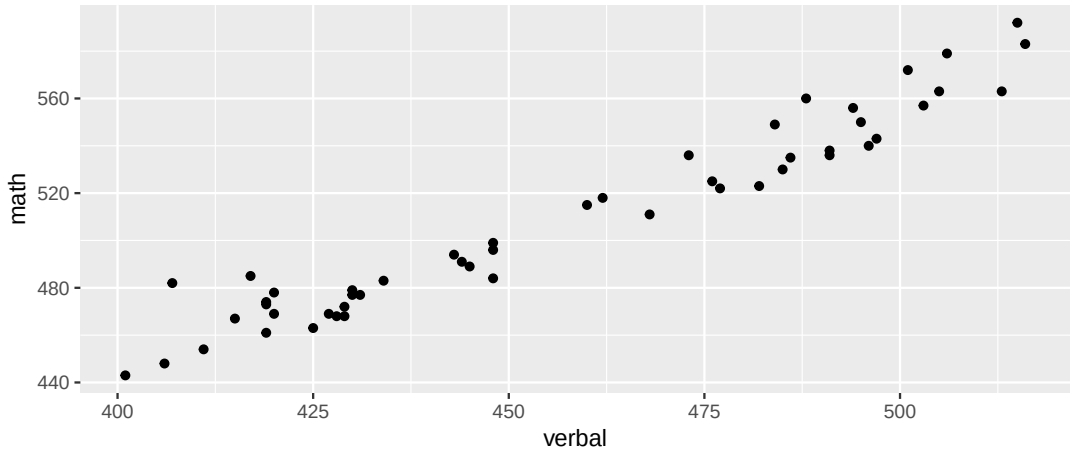
```
ggplot(data=Galton)+
  geom_bar(aes(x=sex))
```



Bivariante and Multivariate displays

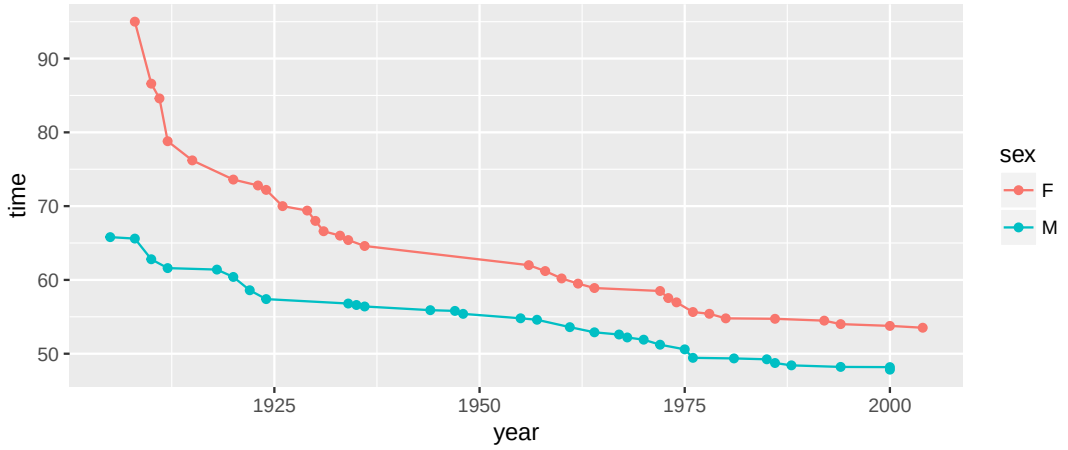
- Scatterplots are appropriate for two quantitative variables.

```
ggplot(data=SAT)+
  geom_point(aes(x=verbal, y=math))
```



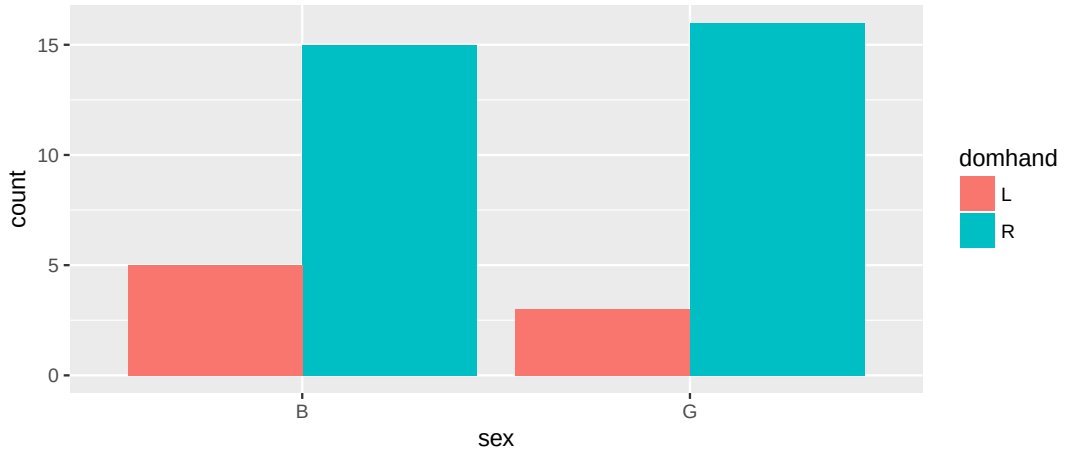
- Time series: just a scatterplot with time on x-axis and points (usually) connected by lines.

```
ggplot(data=SwimRecords)+
  geom_point(aes(x=year, y=time, color=sex))+
  geom_line(aes(x=year, y=time, color=sex))
```

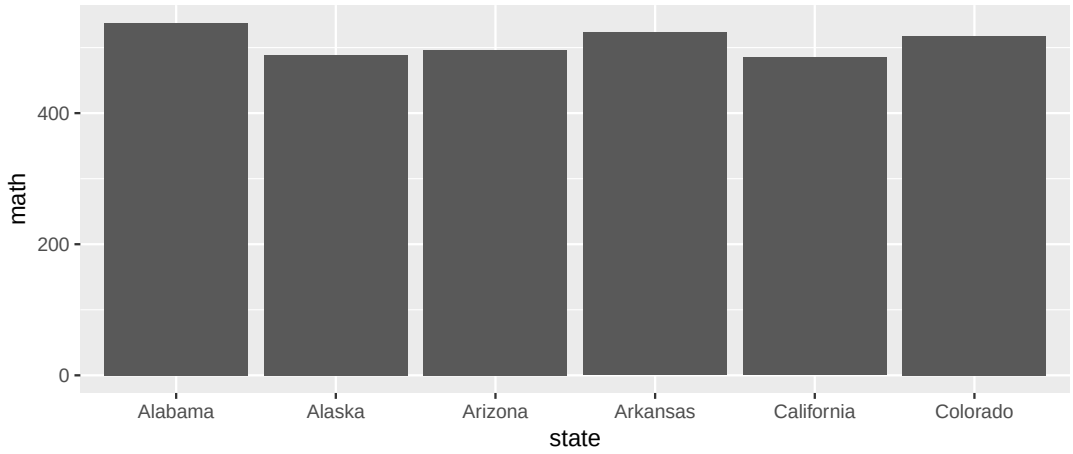


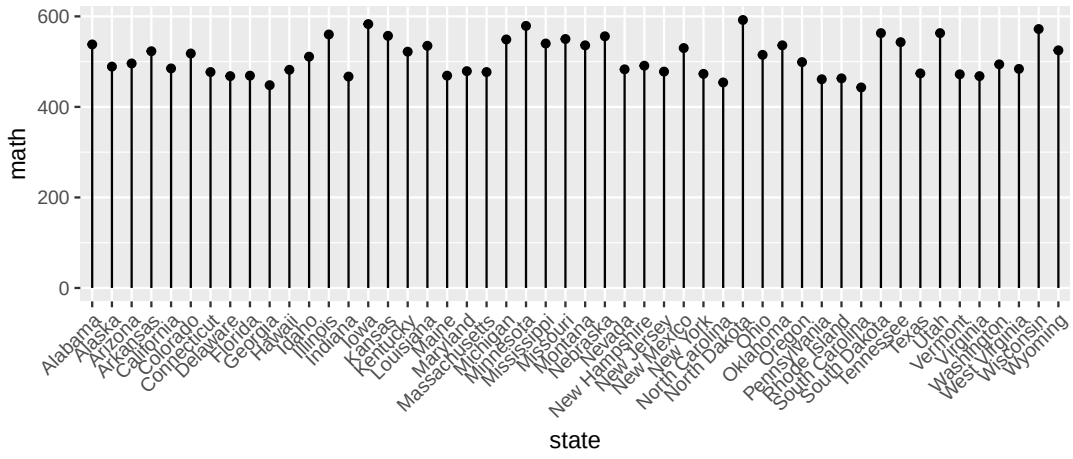
- Clustered bar graphs are traditionally used to summarize two categorical variables.

```
ggplot(data=KidsFeet)+
  geom_bar(aes(x=sex, fill=domhand), position="dodge")
```



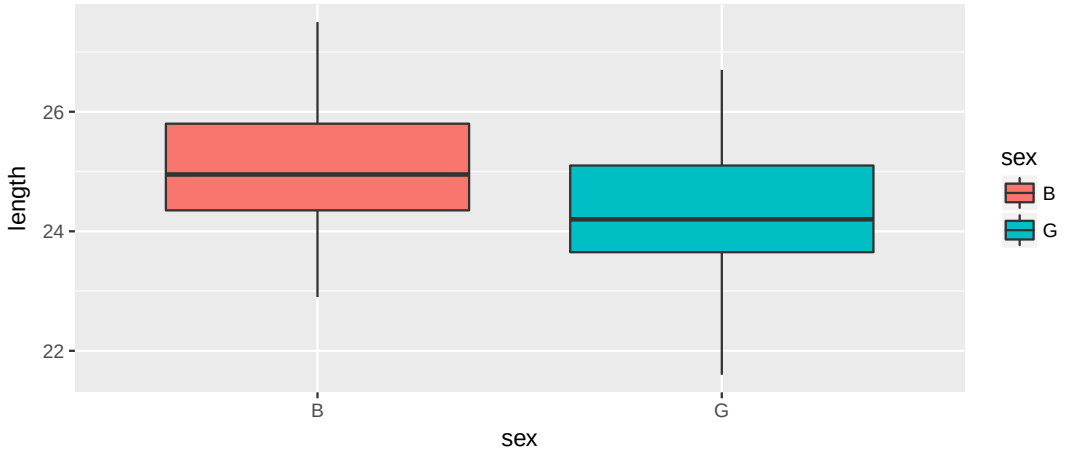
- Bar graphs can also be used when we want to compare a quantitative variable among different observational units. Or, when we wish to plot pre-tabulated data. A lollipop plot is another alternative. (We will come back to how we would code this later on – the code is a little trickier than the ones we have looked at so far.)





- Box-and-whisker plot can also be used to breakdown a quantitative response by a categorical explanatory variable.

```
ggplot(data=KidsFeet)+
  geom_boxplot(aes(x=sex, y=length, fill=sex))
```



Bells and Whistles Most R plotting functions can take many arguments that will modify their behavior. You can read the documentation for more information.

```
help(ggplot2)
```