MATH 205: Statistical methods

Lab 3: Bivariate Data

Goal: Visualizing bivariate Data

- categorical vs categorical: bar plots
- categorical vs continuous: comparative box plots
- continuous vs continuous: scatter plots

Resources

- simpleR:

  https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf

- The R Graph Gallery

  https://www.r-graph-gallery.com/index.html

- Colors in R

  http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf

Handling bivariate categorical data

The `table` command will summarize bivariate data in a similar manner as it summarized univariate data. Suppose a student survey is done to evaluate if students who smoke study less. The data recorded is

Person	Smokes	amount of Studying
1	Y	less than 5 hours
2	N	5 - 10 hours
3	N	5 - 10 hours
4	Y	more than 10 hours
5	N	more than 10 hours
6	Y	less than 5 hours
7	Y	5 - 10 hours
8	Y	less than 5 hours
9	N	more than 5 hours
10	Y	5 - 10 hours

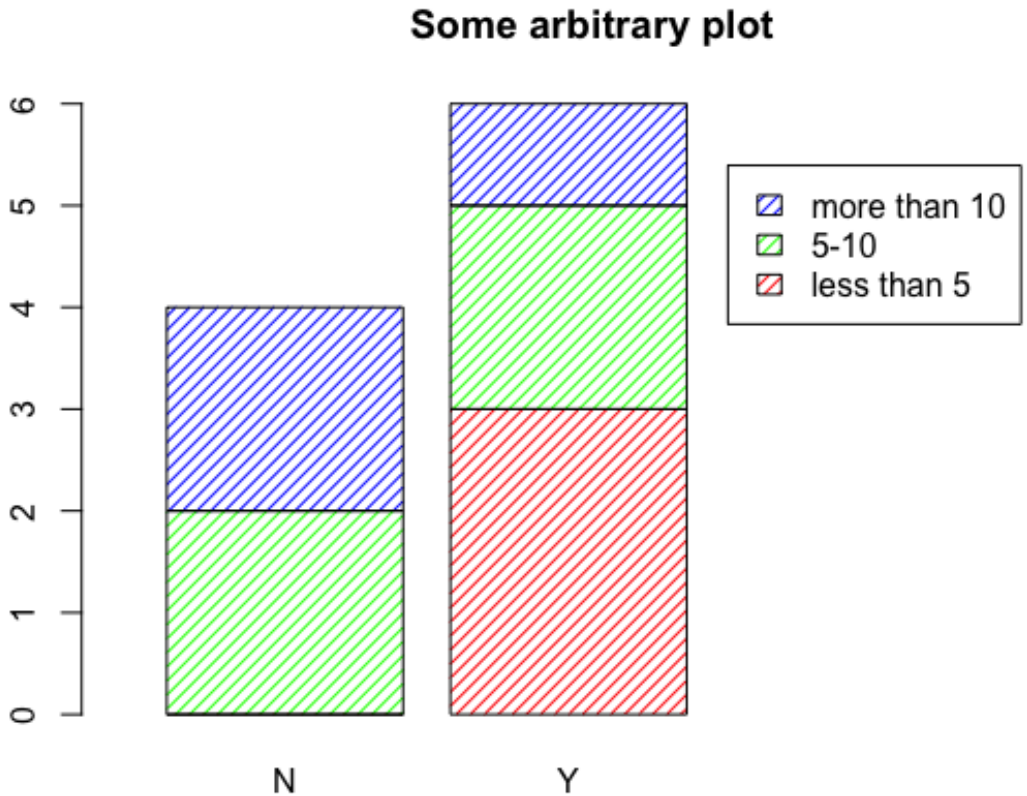
Handling bivariate categorical data

We can handle this in R by creating two vectors to hold our data, and then using the `table` command.

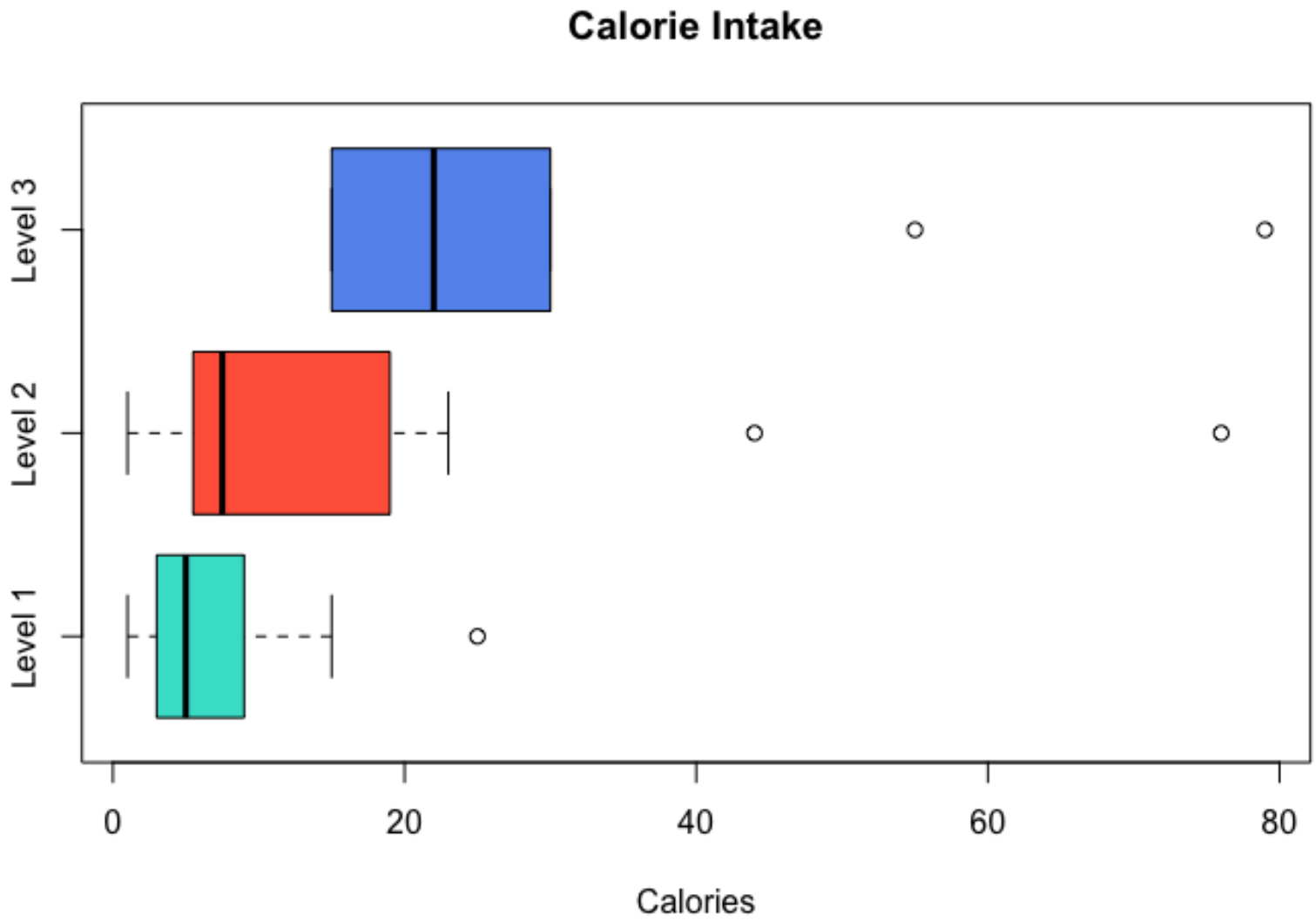
```
> smokes = c("Y","N","N","Y","N","Y","Y","Y","N","Y")
> amount = c(1,2,2,3,3,1,2,1,3,2)
> table(smokes,amount)
       amount
smokes 1 2 3
     N 0 2 2
     Y 3 2 1
```

barplot

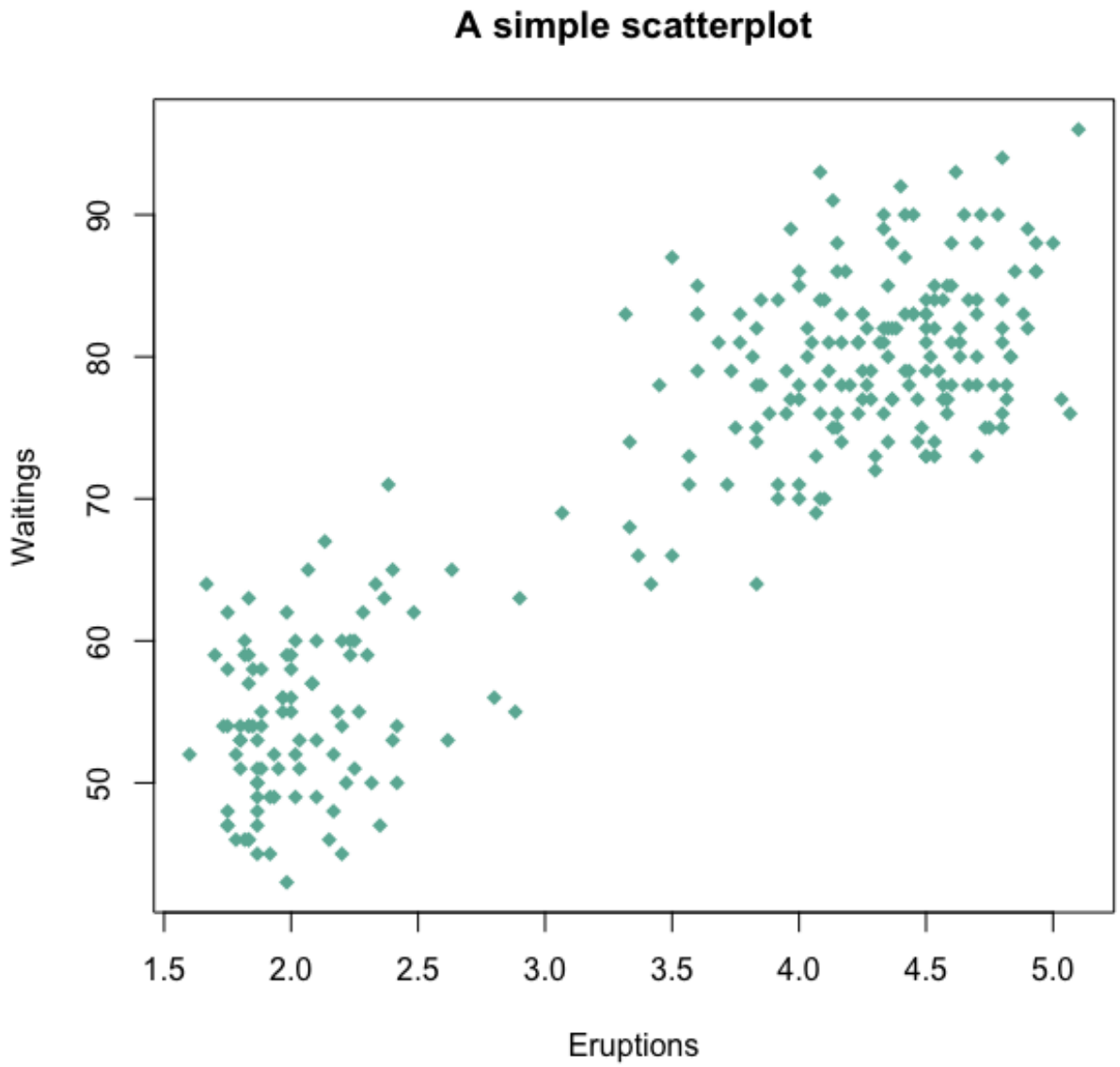
- Essentially, **barplot** plots each colums of data of a table
- **barplot** can be stacked (default), or besides (by setting the option *beside=TRUE*)



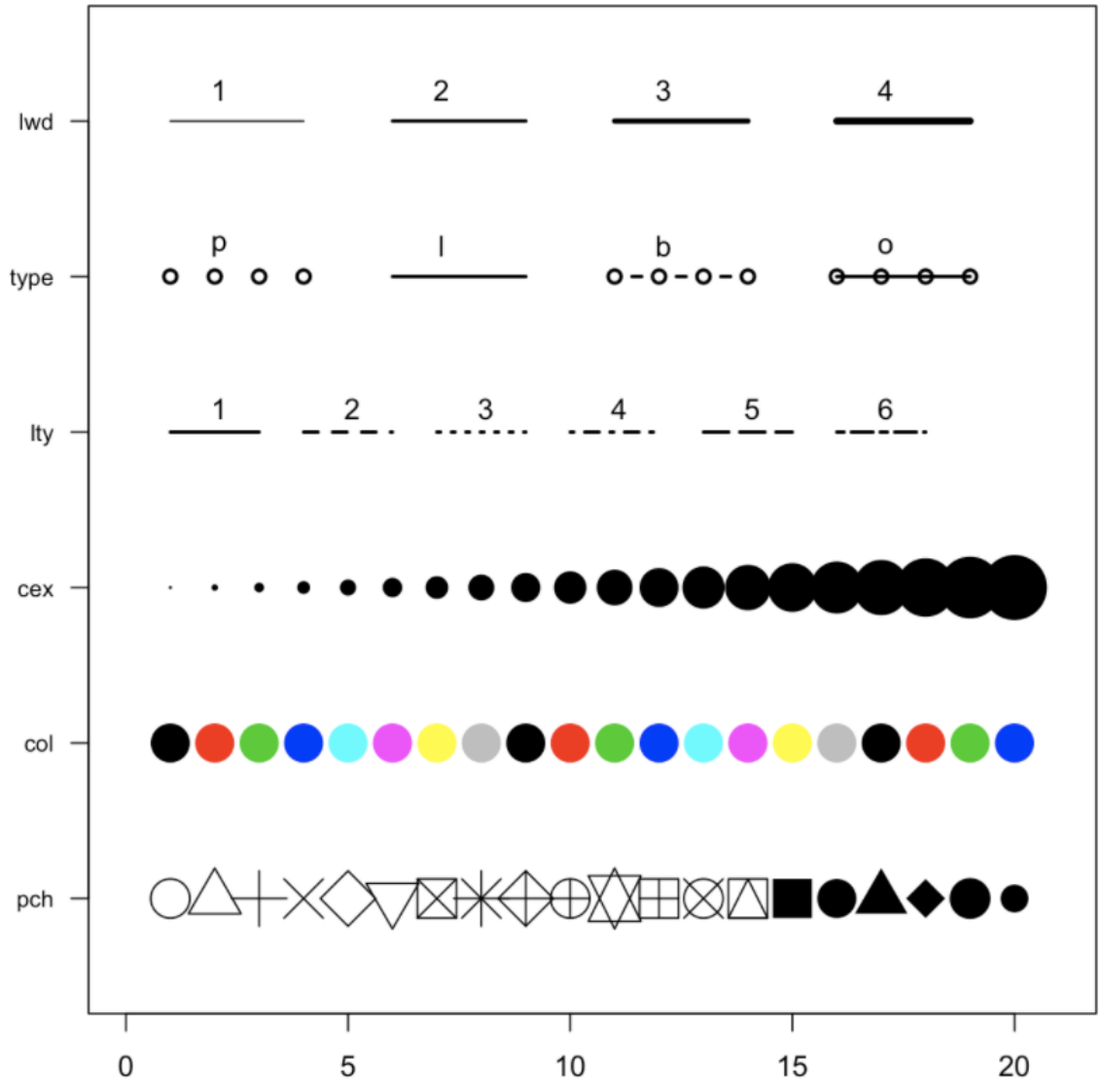
Categorical vs. continuous: comparative box plots



Continuous vs. continuous: scatter plots



Some options



Practice problem

- Choose and load one built-in dataset with two features
- Construct a plot to visualize the relation between two features
- Configure at least 3 options of the plot