Announcements

1. HW #2 due on Wednesday in class
2. R Quiz on 2/15
3. Exam 1 handed out on Friday, 2/16

Agenda

1. Numerial data visualization (cont.)
2. Bivariate Relationships

Motivating Example: College Tuition The data set shows the tuitions and fees charged by the 50 colleges in Massachusetts from 2016-2017.

```
library(mosaic)
library(rvest)
library(readr)

url <- "http://www.collegecalc.org/colleges/new-england/"

Tuition <- read_html(url) %>%
  html_nodes("table") %>%
  html_table(fill=TRUE)

Tuition <- Tuition[[1]] %>%
  mutate(tuition = parse_number(Tuition)) %>%
  select(-Tuition) %>%
  arrange(desc(tuition))
```

```
head(Tuition, 7)

##                School                      Location tuition
## 1    Tufts University       Medford, Massachusetts   51304
## 2      Boston College Chestnut Hill, Massachusetts   50480
## 3    Brown University     Providence, Rhode Island   50224
## 4   Dartmouth College       Hanover, New Hampshire   49998
## 5 Brandeis University       Waltham, Massachusetts   49586
## 6     Yale University       New Haven, Connecticut   49480
## 7   Boston University        Boston, Massachusetts   49176
```

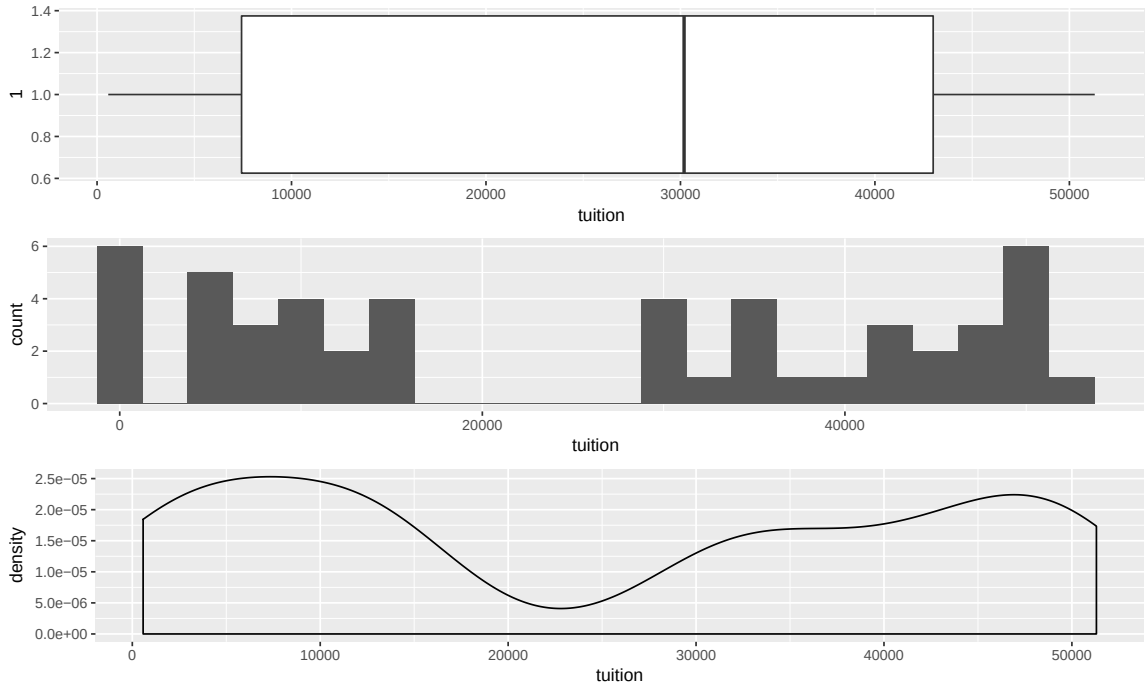
Next, we can calculate some measures of center and spread for tuition.

```
favstats(~ tuition, data = Tuition)

##  min   Q1 median    Q3   max     mean       sd  n missing
##  576 7431  30186 42996 51304 25217.86 18344.91 50       0
```

A box plot, histogram, and density plot reveal different features of the distribution.

```
library(gridExtra)
grid.arrange(
  qplot(data = Tuition, y = tuition, geom = "boxplot", x = 1) + coord_flip(),
  qplot(data = Tuition, x = tuition, geom = "histogram", binwidth = 2500),
  qplot(data = Tuition, x = tuition, geom = "density", adjust = 0.6))
```



1. What information can you glean from the histogram or density plot that is not revealed by the numerical table or the box plot?

2. What do you know about college tuition that might explain the features of this distribution?

Thought Experiment Consider the following variable:

- The annual `income` of all working adults in the United States

Think about the distribution of the variable, and discuss the following questions with a neighbor.

1. Sketch a density plot for the distribution. What features does it have? Is it symmetric? It is unimodal?

2. How would you summarize the distribution numerically? Which measures are most appropriate?

3. Suppose that the government issued a tax rebate in the amount of $2000 to each American taxpayer. How would the distribution of `income` change? What would happen to your measures of center and spread?

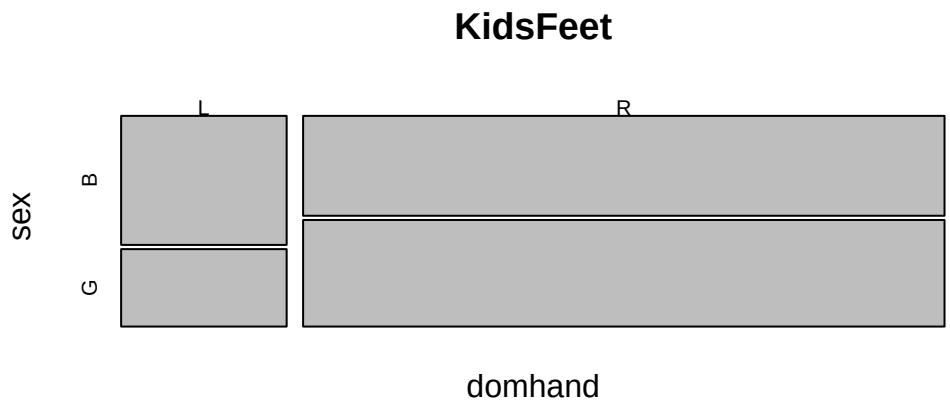
Bivariate Relationships

- Response variable (aka dependent variable): the variable that you are trying to understand
- Explanatory variable (aka independent variable, aka predictor): the variable that you think might be related to the response variable

1. Visualize: Put response variable on y-axis and explanatory variable on x-axis
   - Two numerical variables: scatterplot [`qplot()`]
     - Overall patterns and deviations from those patterns
     - Form (e.g. linear, quadratic, etc.), direction (positive or negative), and strength (how much scatter?)
     - Outliers
   - Two categorical variables: mosaic plot [`mosaicplot()`]
   - Numerical response and a categorical explanatory variable:
     - Side-by-side box plots [`geom = "boxplot"`]
     - Multiple density plots [`geom = "density"` with `color` aesthetic or facets]
   - Multivariate relationships:
     - For a third variable that is categorical, use the `color` aesthetic or facets
     - For a third variable that is numerical, consider using the `cuts` option, or 3d effects!
2. Numerical Summary: Correlation (r)—a numerical measure of direction and strength of a *linear* relationship!

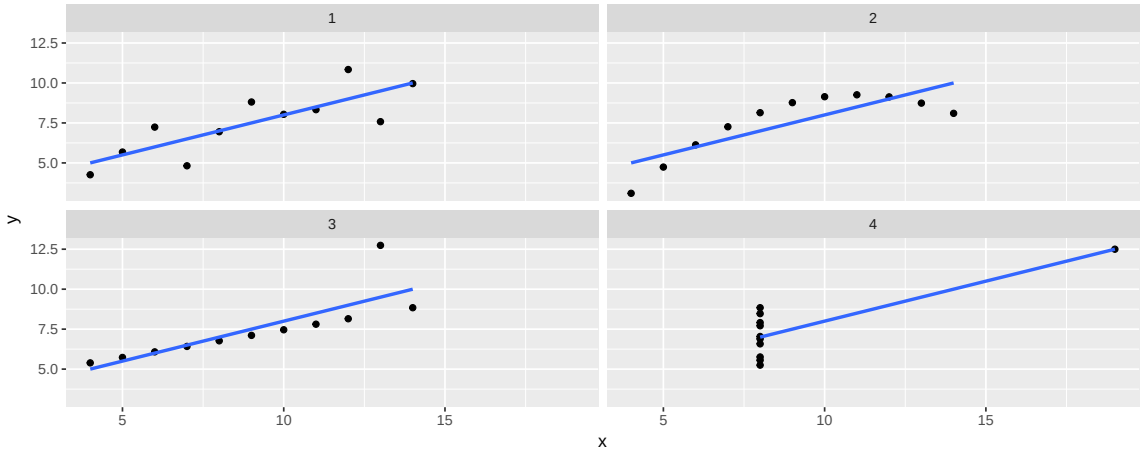
```
library(mosaic)
qplot(data = KidsFeet, y = length, x = width)
qplot(data = KidsFeet, y = length, x = width, color = sex)
qplot(data = KidsFeet, y = length, x = sex, geom = "boxplot")
qplot(data = KidsFeet, x = length, color = sex, geom = "density")
qplot(data = KidsFeet, x = length, facets = ~sex, geom = "density")
```

```
mosaicplot(domhand ~ sex, data = KidsFeet)
```



Correlation The (Pearson Product-Moment) correlation coefficient [`cor()`] is a measure of the strength and direction of the *linear* relationship between two numerical variables. It is usually denoted r and is measured on the scale of $[-1, 1]$.

```
## Warning:  package 'tidyr' was built under R version 3.4.2
```



Note that correlation only measures the strength of a *linear* relationship. In each of the four very different (Anscombe) data sets shown above, the correlation coefficient is the same (up to three digits)!

Examples Get a feel for the value of the correlation coefficient in different scatterplots.

1. Do a Google Image search for "scatterplot" and describe the form, direction, and strength of three different-looking patterns. Sketch each plot.

   (a) :

   (b) :

   (c) :