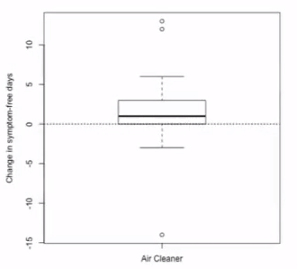
Principles of Analytic Graphs

This lecture is about basic principles necessary for building analytic graphics. The basic goal is to provide some general rules that one can follow when we're building analytic graphics from data, and we're trying to tell a story about the data.

**Principles of Analytic Graphics**

Principle 1: Show comparisons

* Evidence for a hypothesis is always relative to another competing hypothesis
* Always ask "Compared to What?"

**Show Comparisons**

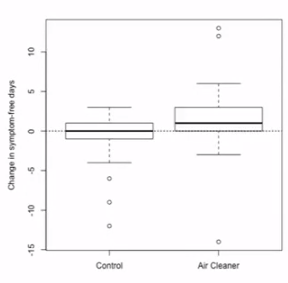
A Basic example

Affect of an air cleaner on the asthma of children with asthma.

Median increase was about 1 more symptom free day.

We need to know what we're comparing it to.

We instead add a control set where the houses have no air cleaner.



It's always to show a comparison in a plot so you can have a basis to compare what you're showing; so you can compare the evidence between two different hypotheses.

\

**Principles of Analytic Graphics**

Principle 1: Show comparisons

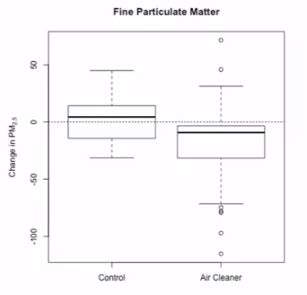
Principle 2: Show causality, mechanism, explanation, systematic structure

* What is your causal framework for thinking about a question?

For Principle 2, you need to show how you believe the world works. What is the causal framework that fits into the question you're interested in?

**Show causality, mechanism**

In our example, our hypothesis is that the air cleaner removes particulates in the air, and this matter that's being removed is no longer going into the child's lungs and triggering their asthma symptoms.

This graph shows the amount of particulates in the air, for the contol group and the group of homes that got the air cleaner.

Now we can see that not only did a child's symptom -free days increase, but also the fine particulate matter decreased on average.

**Principles of Analytic Graphics**

Principle 1: Show comparisons

Principle 2: Show causality, mechanism, explanation, systematic structure

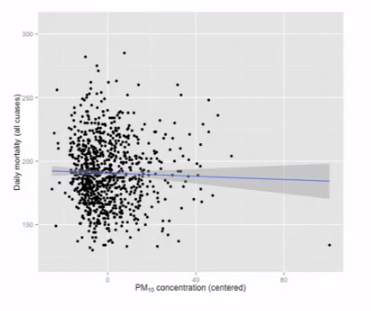
Principle 3: Show multivariate data

* Multivariate = more than 2 variables
* The real world is multivariate
* Need to "escape flatland"

Show as much data on one plot as you can, the reason is because the world has a lot of things going on. Having a lot of data on a plot gives a much richer story.

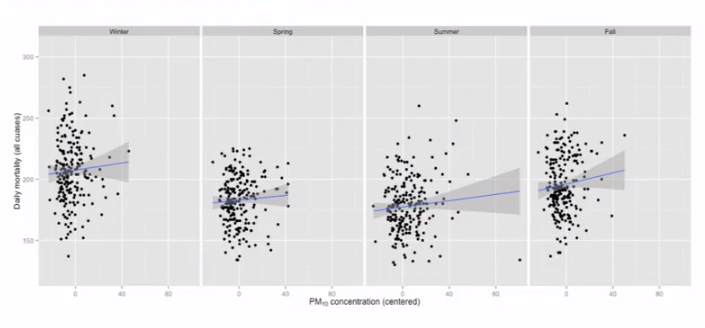
**Show Multivariate Data**

This example comes from outdoor air pollution.



On the X Axis we have particulate matter less than 10 microns in aerodynamic diameter. The concentrations of those from day to day (each dot is a day). On the Y axis we have the daily mortality in NY City. There seems to be a slightly negative correlation between negative PM levels and mortality.

There are other variables that might be of interest; so how does this relationship change across seasons.



This plot shows PM10 to mortality. We can see now that the relationship is now slightly positive.

So it is important to show as many variables as is reasonable so you can get a clear picture of the relationships in your data.

**Principles of Analytic Graphics**

Principle 4: Integration of evidence

* Completely integrate words, numbers, images, diagrams
* Data graphics should make use of many modes of data presentation
* Don't let the tool drive the analysis

The tools in R are flexible so you won't be forced to use only a specific type of data graphic.

Principle 5: Describe and document the evidence with appropriate labels, scales, sources, etc.

* A data graphic should tell a complete story that is credible

Principle 6: Content is king

* Analytical presentations ultimately stand or fall depending on the quality, relevance, and integrity of their content.
* You need a story that you're trying to tell. Get the purpose first, then you can decide what's the best way to present that and how am I going to present it and what is it going to look like

**Summary**

1) Show Comparisons (ie. Compare test group with control group)

2) Show causality, mechanism, explanation (How you think the world works)

3) Show multivariate data (Real life is complex need many variables shown)

4) Integrate multiple modes of evidence (tables, plots, texts; different plots)

5) Describe and document the evidence (source code to your plot)

6) Content is King. (A Purpose)

Exploratory Graphs

This lecture is abuot constructing exploratory graphs; graphs you make for yourself so you can look at the data and explore what's going on in the data sets you're looking at.

**Why do we use graphs and data analysis?**

* To understand data properties
* To find patterns in data
* To suggest modeling strategies
* To "debug" analyses
* To communicate results

Exploratory graphs are really about the first four things on this list above.

**Characteristics of exploratory graphs**

* They are made quickly
* A large number are made
* The goal is for personal understanding
* Axes/legends are generally cleaned up (later)
* Color/size are primarily used for information (to separate information, but not presentable)

**Air Pollution in the United States**

The U.S. Environmental Protection Agency (EPA) sets national ambient air quality standards for outdoor air polution

For fine particle pollution (PM2.5), the "annual mean, averaged over 3 years" cannot exceed 12 ug/m3.

Data on daily PM2.5 are available from the U.S. EPA web site

**Question:** Are there any counties in the U.S. that exceed that national standard for fine particle pollution?

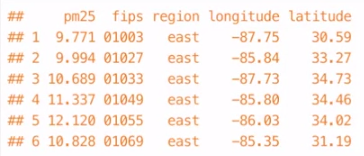
**Data**

Annual average PM2.5 averaged over the period 2008 through 2010

pollution <- read.csv("data/avgpm25.csv", colClasses = c("numeric", "character",

"factor", "numeric", "numeric"))

head(pollution)



At the start you still want to have an underlying question, even if it is vague. Your question will drive the way you see the data. We want to see if the counties exceed this national air quality standard.

**Simple Summaries of Data**

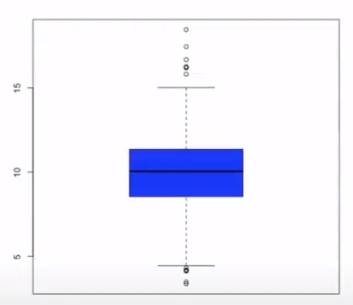
One dimension

* Five-number summary
* Boxplots
* Histograms
* Density plot
* Barplot

**Five Number Summary**

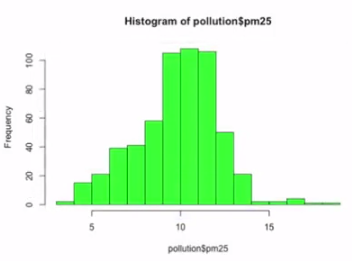
summary(pollution$pm25) ## 6 numbers because mean is included.

**Boxplot**

boxplot(pollution$pm25, col = "blue")

**Histogram**

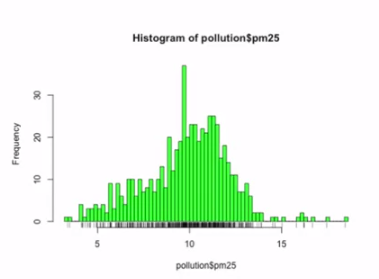
hist(pollution$pm25, col = "green")



This is the same data as a histogram.

hist(pollution$pm25, col = "green", breaks = 100)

rug(pollution$pm25)



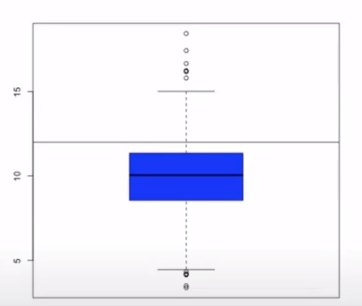
We've added a rug showing the exact spot of each data point below the graph.

Also it is now split up more due to an increase in the "breaks".

**Overlaying Features**

boxplot(pollution$pm25, col = "blue")

abline(h = 12)

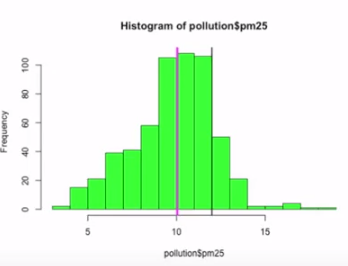


This is the same boxplot as before, but now there is a line at 12 for the nationalair quality standard. This gives an idea of who is above and below the line.

hist(pollution$pm25, col = "green")

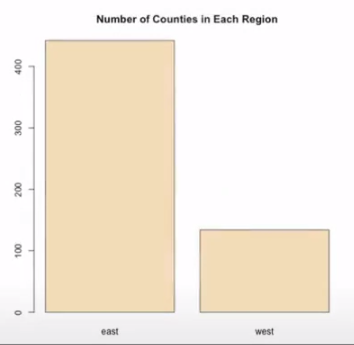
abline(v = 12, lwd = 2)

abline(v = median(polution$pm25), col = "magenta", lwd = 4)



**Barplot**

barplot(table(pollution$region), col = "wheat", main = "number of Counties in Each Region")



Bar plots are great for categorical data. The majority of counties are in east US.

**Simple Summaries of Data**

Two dimensions

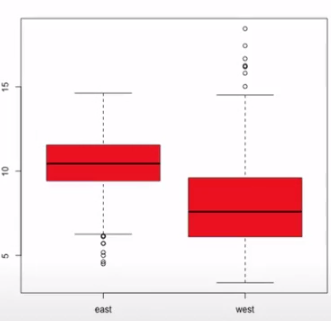
* Multiple/overlayed 1-D plots (Lattice/ggplot2)
* Scatterplots
* Smooth scatterplots

2 dimensions

* Overlayed/multiple 2-D plots; coplots
* Use color, size, shape to add dimensions
* Spinning plots
* Actual 3-D plots (not that useful)

**Multiple Boxplots**

boxplot(pm25 ~ region, data = pollution, col = "red")



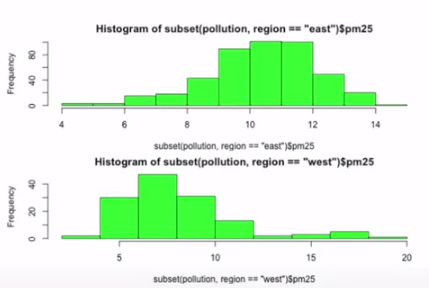
Here is a way to look at two dimensional data; in one dimension, the east/west categories, and the other is the pm25 variable.

**Multiple Histograms**

par(mfrow = c(2, 1), mar = c(4, 4, 2, 1))

hist(subset(pollution, region == "east")$pm25, col = "green")

hist(subset(pollution, region == "west")$pm25, col = "green")



**Scatterplot**

with(pollution, plot(latitude, pm25))

abline(h = 12, lwd = 2, lty = 2)