Exploratory Data Analysis (EDA)

Exploratory Data Analysis with R

Odysseus listening to the song of the sirens. Late Archaic (500480 BCE)

Boris Steipe

UNIVERSITY OF TORONTO

Department of Biochemistry
Department of Molecular Genetics

Learning Objectives

Be able to load, transform and visualize data;

Begin generating ideas about how relationships in data can be explored;

Understand some principles of effective visualization;

Know where to go for help.

Exploratory Data Analysis (EDA)

Numerical analysis and graphical presentation of data without a particular model or underlying hypothesis in order to ...

- uncover underlying structure;
- define important variables;
- detect outliers and anomalies;
- detect trends;
- develop statistical models;
- test underlying assumptions.

The goal is hypothesis generation, not hypothesis testing.

Therefore EDA is often **the first step** of data analysis.

Exploratory Data Analysis (EDA)

The objectives of EDA include:

- uncovering underlying structure and identifying trends and patterns;
- extracting important variables;
- detecting outliers and anomalies;
- testing underlying assumptions;
- developing statistical models.

Exploratory Data Analysis (EDA)

The practice of EDA emphasizes looking at data in different ways through:

- computing and tabulating basic descriptors of data properties such as ranges, means, quantiles, and variances;
- generating graphics, such as boxplots, histograms, scatter plots;
- applying transformations, such as log or rank;
- comparing observations to statistical models, such as the QQ-plot, or linear and non-linear regression;
- simplifying data through dimension reduction;
- identifying underlying structure through clustering ...

... all with the final goal of defining which statistical model might be appropriate to use for hypothesis testing and prediction.

Why R for EDA ?

Full-featured programming language

"Statistical workbench"

Data manipulation is easy

Easy access to graphics

Sophisticated packages and libraries

Large community

Graphics

Good graphics are immensely valuable. Poor graphics are worse than none.

If you want to learn more about good graphics and information design, find a copy of Edward Tufte.s **The Visual Display of Quantitative Information**. You can also visit his Web site to get a sense of the field (www.edwardtufte.com).

Fundamentally, there is one simple rule.

Use less ink.

The rule has many corollaries.

Use Less Ink

Make sure that all elements on your graphics are necessary.

Make sure that all elements on your graphics are informative.

Make sure that all information in your data is displayed.

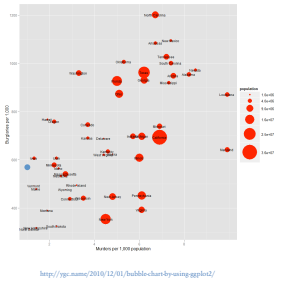
Not all of R's defaults use as little ink as possible, you can improve on them!

Refined Graphics

... for a popular alternative to R.s defaults, check out the ggplot2 package (http://www.ggplot2.org).

```
> install.packages("ggplot2")
...
> library(ggplot2)
...
```

Example: Bubble chart.

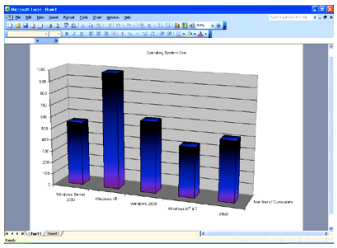


http://ygc.name/2010/12/01/bubble-chart-by-using-ggplot2/

```
crime <- read.csv("crimeRatesByState2008.csv", header=TRUE, sep="\t")
p <- ggplot(crime, aes(murder,burglary,size=population, label=state))
p <- p+geom_point(colour="red") +scale_area(to=c(1,20))+geom_text(size=3)
p + xlab("Murders per 100,000 population") + ylab("Burglaries per 100,000")
```


Graphics Examples

"... What if you could gussy up a report or pretty up a chart without much additional work? What if, using just one extra line of code, you could create a Microsoft Excel column chart that included a cool gradient fill like this one:"



(From Microsoft TechNet)

Graphics Example: Chartjunk

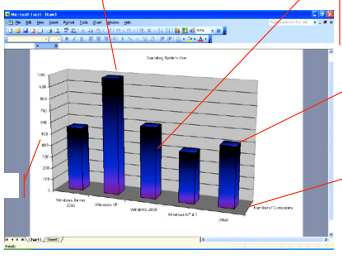
Only five numbers actually

Meaningless colors, no connection to actual scale. Note that the range of colors differs for each stack!

Values can't be easily retrieved from graph due to parallax.

No units

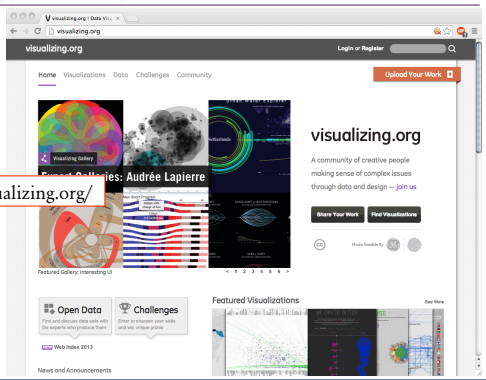
What does the third dimension even mean?



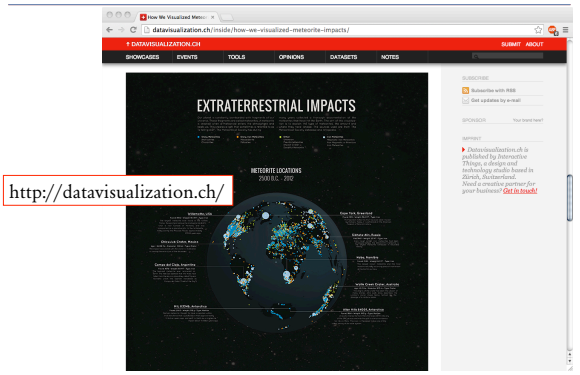
(From Microsoft TechNet)

Graphics Examples

http://visualizing.org/

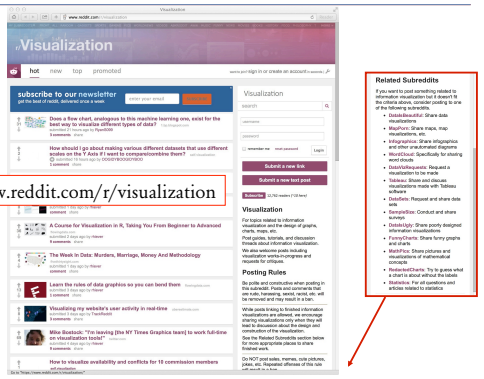


Graphics Examples



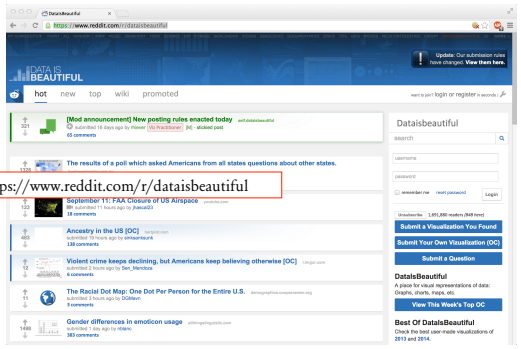
http://datavisualization.ch/

Graphics Inspiration and Advice



https://www.reddit.com/r/visualization

Graphics Inspiration and Advice



https://www.reddit.com/r/datisbeautiful

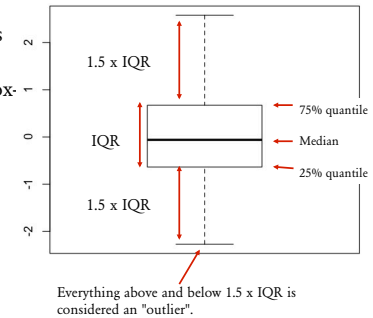
Graphics for Descriptive Statistics

Statistics can be vaguely divided into *descriptive* and *inferential* domains. Both are supported by graphics. Let us discuss basic *summary statistics* (mean, median, variance, standard deviation) and how to visualize them...

Box Plot

Descriptive statistics can be intuitively summarized in a Box-plot.

```
> boxplot(x)
```

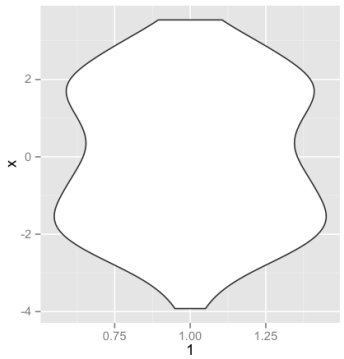


IQR = Inter Quantile Range = 75% quantile to 25% quantile

Violinplot

Internal structure of a data-vector can be made visible in a violin plot. The principle is the same as for a boxplot, but a width is calculated from a smoothed histogram.

```
p <- ggplot(X, aes(1,x))
p + geom_violin()
```



QQPLOT

One of the first things we may ask about data is whether it deviates from an expectation e.g. to be normally distributed.

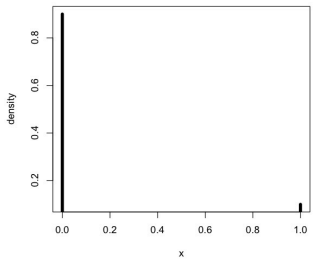
The quantile-quantile plot provides a way to visually verify this.

The QQ-plot shows the theoretical quantiles versus the empirical quantiles. If the distribution assumed (theoretical one) is indeed the correct one, we should observe a straight line.

R provides `qqnorm()` and `qqplot()`.

probability distributions

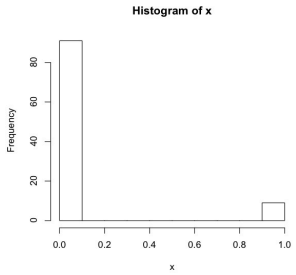
Can be either discrete or continuous (uniform, Bernoulli, normal, etc)
Defined by a density function, p(x) or f(x)
Bernoulli distribution Be(p): flip a coin (T=0, H=1). Assume probability of H is 0.1 ...




```
x <- 0:1
f <- dbinom(x, size=1, prob=0.1)
plot(x, f, xlab="x", ylab="density", type="h", lwd=5)
```


probability distributions

Random sampling: Generate 100 observations from a Be(0.1)

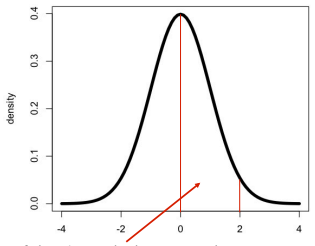



```
set.seed(100)
x <- rbinom(100, size=1, prob=0.1)
hist(x)
```


probability distributions

Normal distribution N(μ,^{2})
μ is the mean and ^{2} is the variance.

Extremely important because of the Central Limit Theorem: if a random variable is the sum of a large number of small random variables, it will be normally distributed.



The area under the curve is the probability of observing a value between 0 and 2.

```
x <- seq(-4, 4, 0.1)
f <- dnorm(x, mean=0, sd=1)
plot(x, f, xlab="x", ylab="density", lwd=5, type="l")
```


exploring data

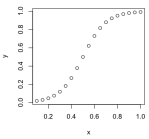
When teaching (or learning new procedures) I prefer to work with **synthetic data**.
Synthetic data has the advantage that I know what the outcome of the analysis should be.
Typically one would create values acording to a function and then add noise.

R has several functions to create sequences of values or you can write your own ...

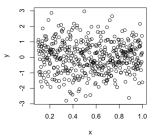
```
0:10
seq(0, pi, 5*pi/180)
rep(1:3, each=3, times=2)
for (i in 1:10) { print(i*i) }
```


synthetic data

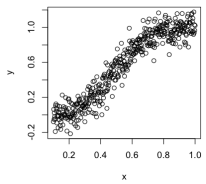
Function ...



Noise ...



Noisy Function ...



Explore functions and noise.

arbitrary probabilities

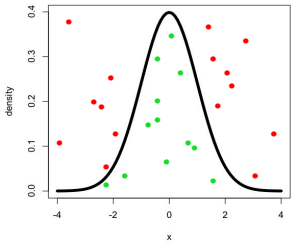
Map the probabilities onto the unit interval, then use runif() to choose.

arbitrary probability distributions

Calculate the function value at a point in the domain of interest. Then choose a uniform value between the minimum and maximum in the domain. If the random value is smaller than the function value, accept, otherwise reject.

Think of this as randomly placing dots on a plot of the function, then accepting those that lie **under** the curve.

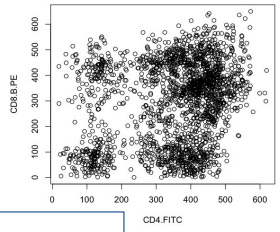
Try this.



Multivariate data - scatter plots

Biological data sets often contain several variables, they are multivariate.

Scatter plots allow us to look at two variables at a time.



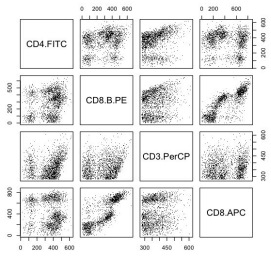

```
# GvHD flow cytometry data
gvhd <- read.table("GvHD.txt", header=TRUE)
# Only extract the CD3 positive cells
gvhdCD3p <- as.data.frame(gvhd[gvhd[, 5]>280, 3:6])
plot(gvhdCD3p[, 1:2])
```

This can be used to assess independence and identify subgroups.

trellis graphics

Trellis Graphics is a family of techniques for viewing complex, multi-variable data sets.

```
plot(gvhdCD3p, pch=".")
```



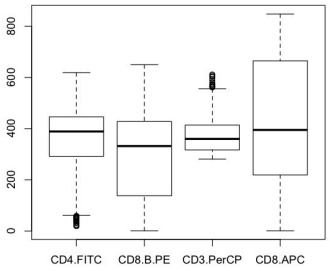
Tip:

Many more possibilities are available in the "lattice" package. See `?Lattice`

Boxplots

The boxplot function can be used to display several variables at a time.

```
boxplot(gvhdCD3p)
```



Exercise: Interpret this plot.

Summary

EDA should be the first step in any statistical analysis!

Good modeling starts and ends with EDA.

R provides a great framework for EDA.

boris.steipe@utoronto.ca