Pre-mining data before analysis

Statistical Data Mining I

Rachael Hageman Blair

Why not dive in ? …..

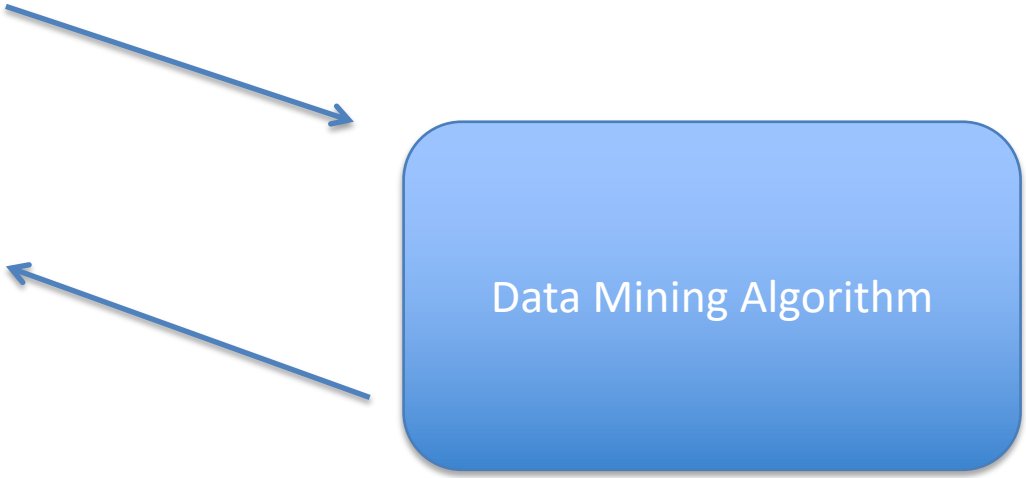
Data should “speak for itself”

prior to any formal analysis.

- Data “exploration” can draw attention to:
  - quirks and errors in the data.
  - obvious trends.
  - technical artifacts.
  - the types of analysis that should be performed.

Why not dive in ? …..

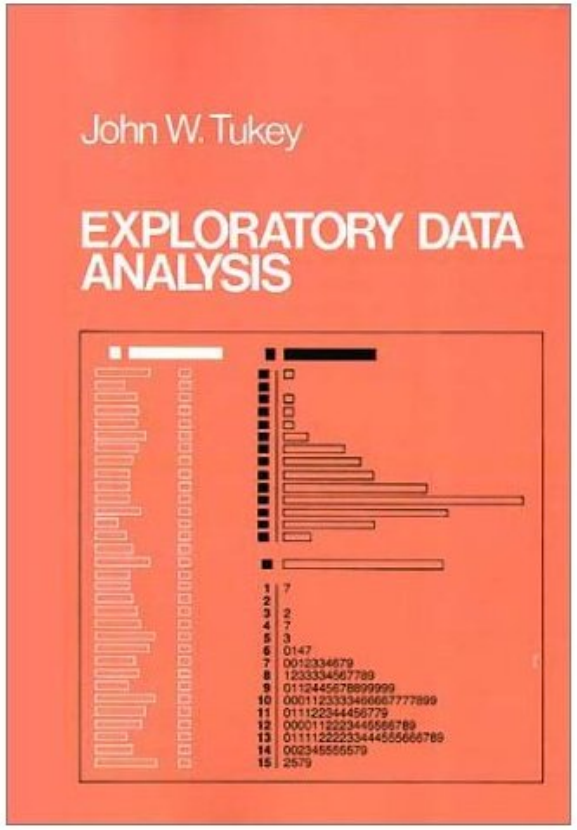
"Garbage in equals Garbage out"



Revealing views of the data …..

- John Tukey coined the name **Exploratory Data Analysis** (EDA) to describe the use of graphs to display and help understand data.

- Centers around quick and dirty Methods (pencil and paper) to visualize and examine small data sets.

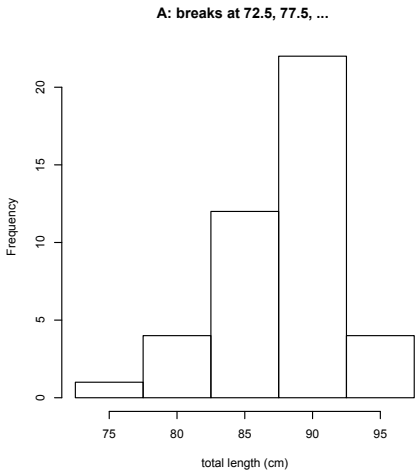


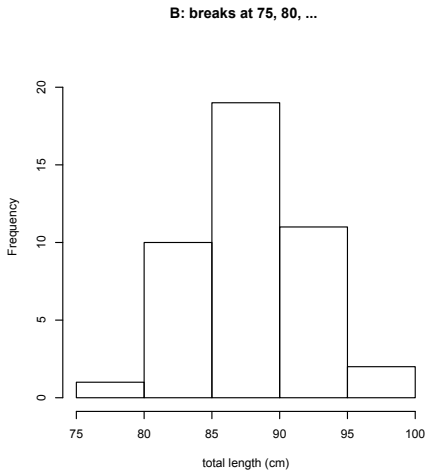
Times have changed …..

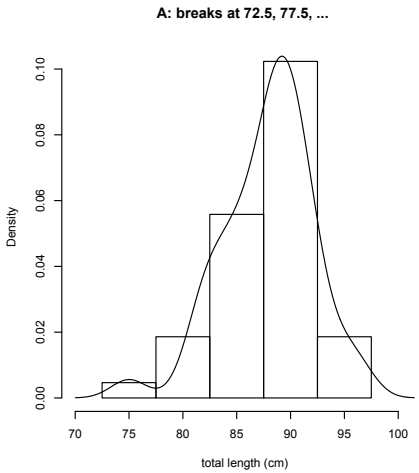
- Billions of transactions/ year in financial sectors.
- Social network billions of nodes with billions of connections, and covariates.
- Human genome project gigabytes of genetic information.
- Astronomy ~ terabytes and petabytes of data.

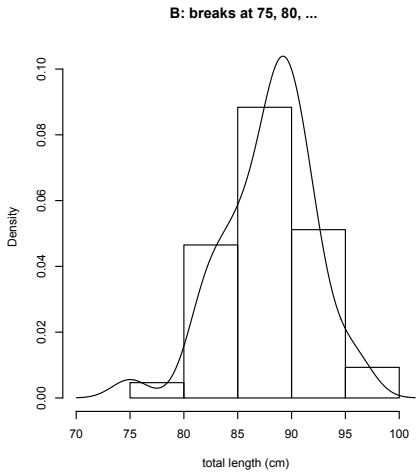
Views of a single sample …..

- Histograms and density plots









Views of a single sample …..

- Stem and Leaf Plots

  heights for 37 rowers

```
> stem(rowers$ht)

  The decimal point is 1 digit(s) to the right of the |

  15 | 6
  16 |
  16 | 5
  17 | 4
  17 | 5678899
  18 | 00000011223
  18 | 55666668899
  19 | 123
  19 | 58
```

Lower quartile is 179 (10th largest)

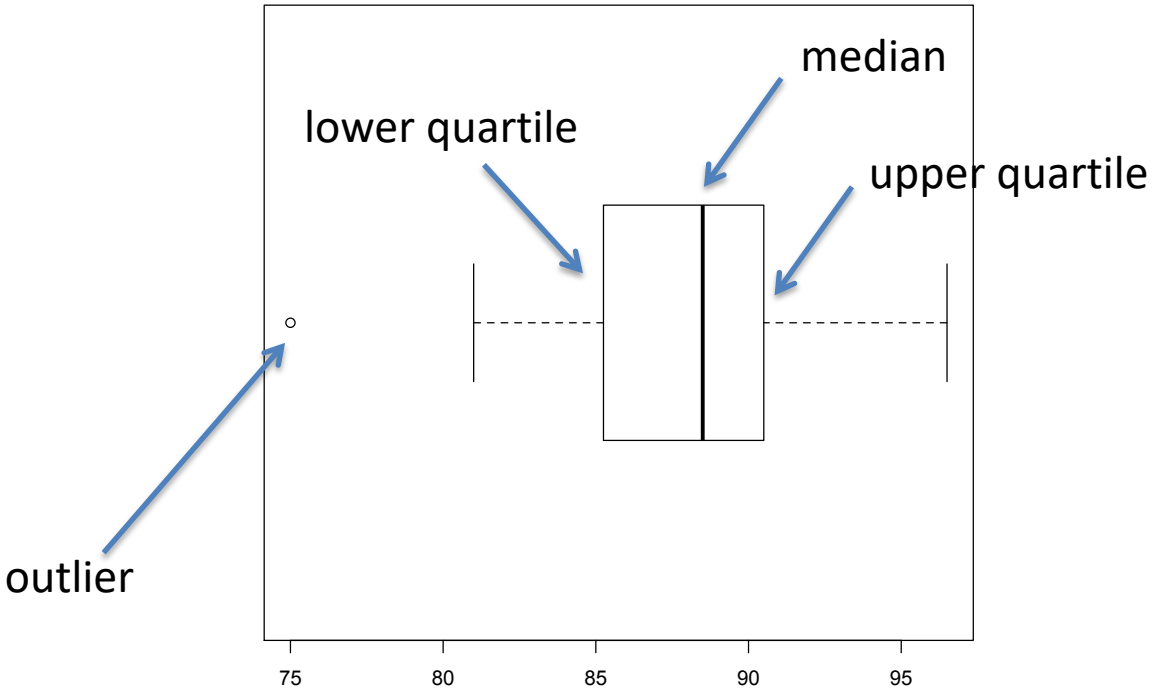
Median is 182 (19th largest)

Upper quartile is 186 (28th largest)

Views of a single sample …..

- Box Pots

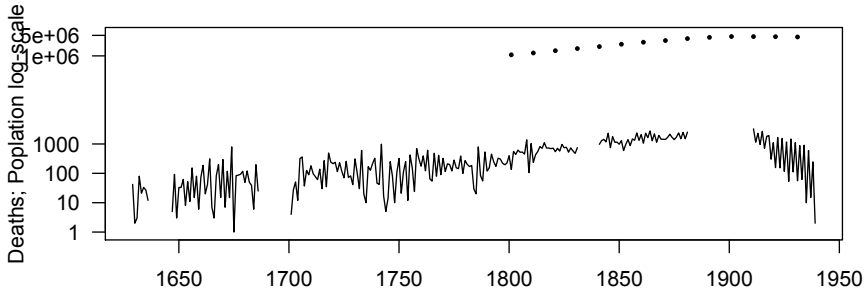
  Female possum length

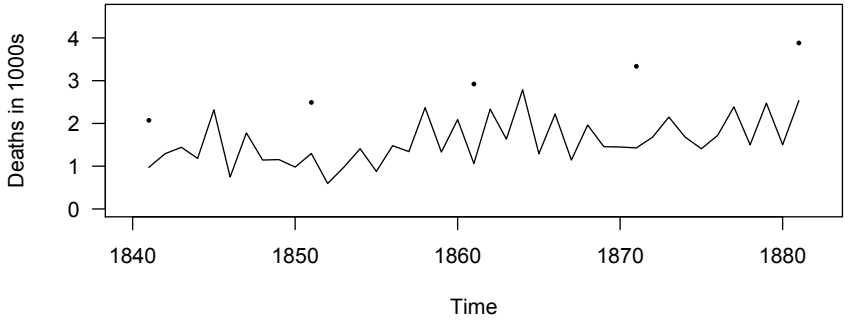


Views of a single sample …..

- Univariate time series

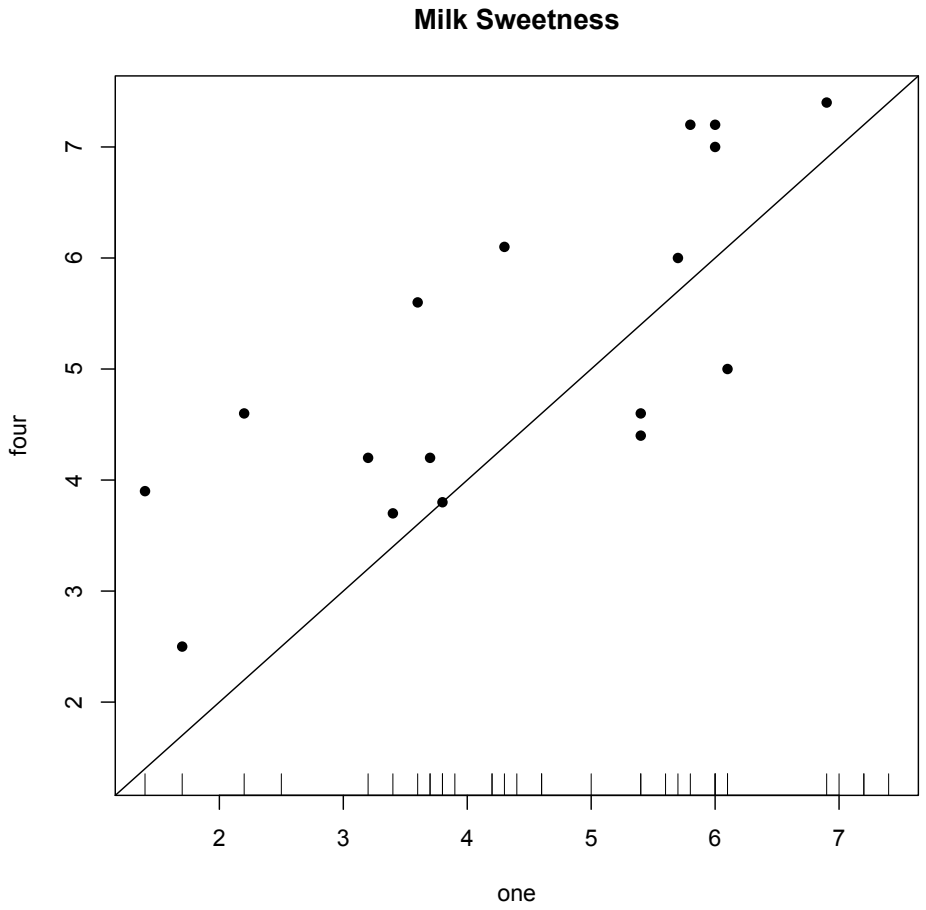
  London Measles Data





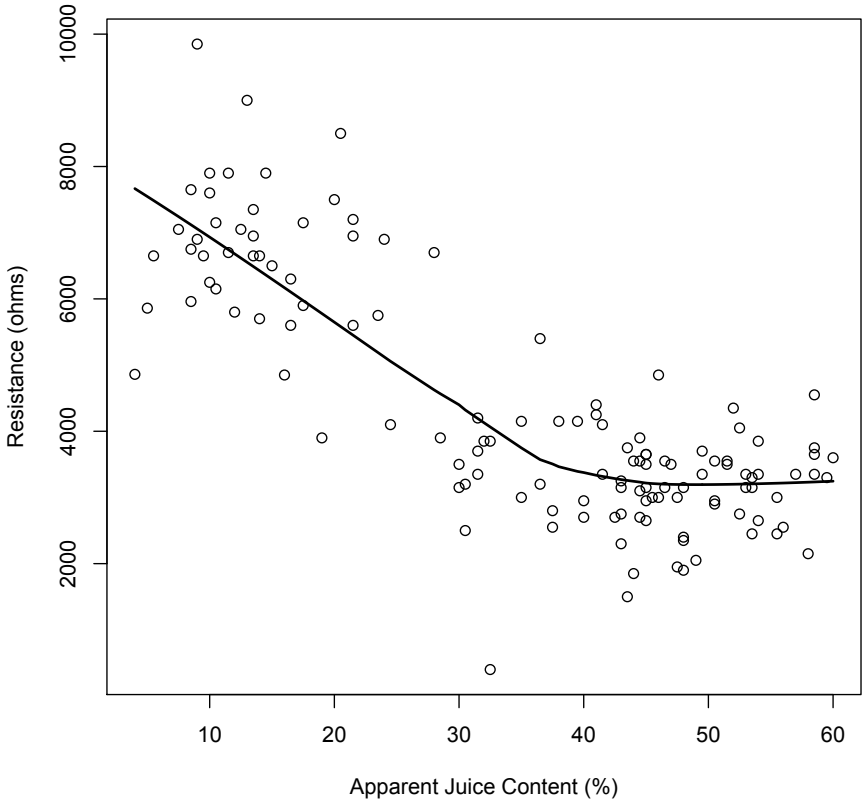
Views of groups of data …..

- Patterns in Bivariate Data



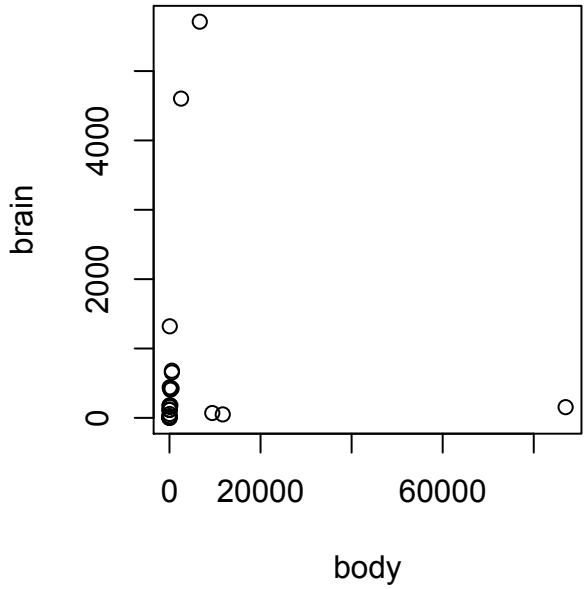
Views of groups of data …..

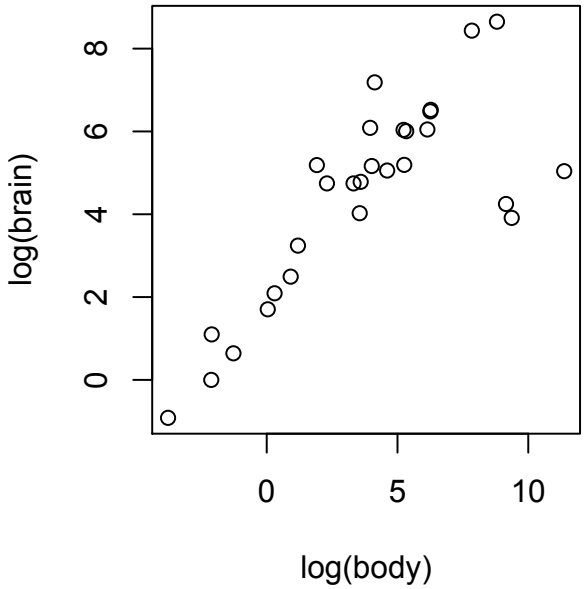
- Fitting of a smooth trend curve



Views of groups of data …..

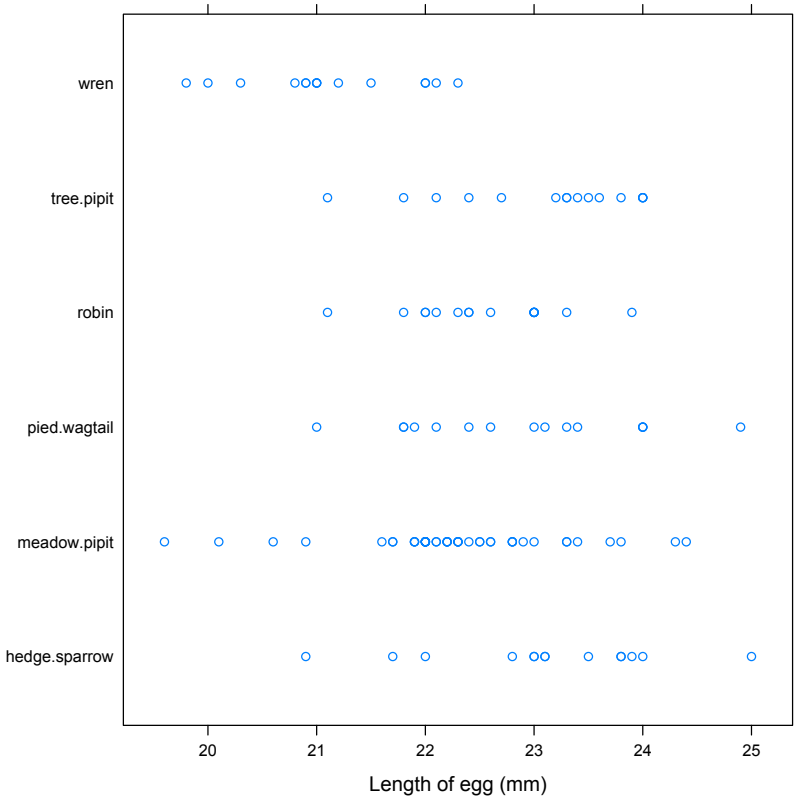
- What is the appropriate scale?

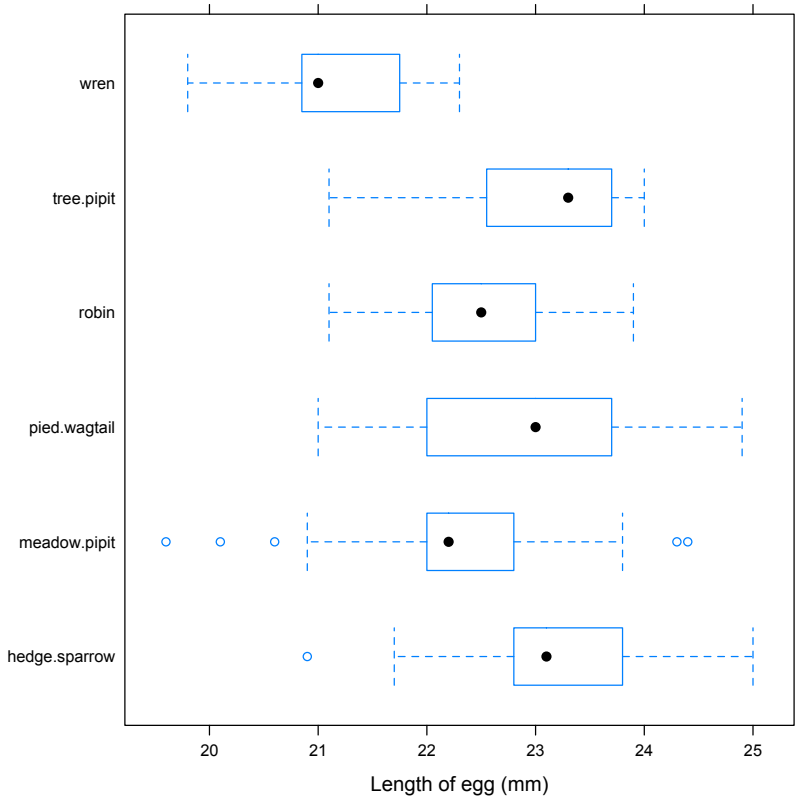




Views of groups of data …..

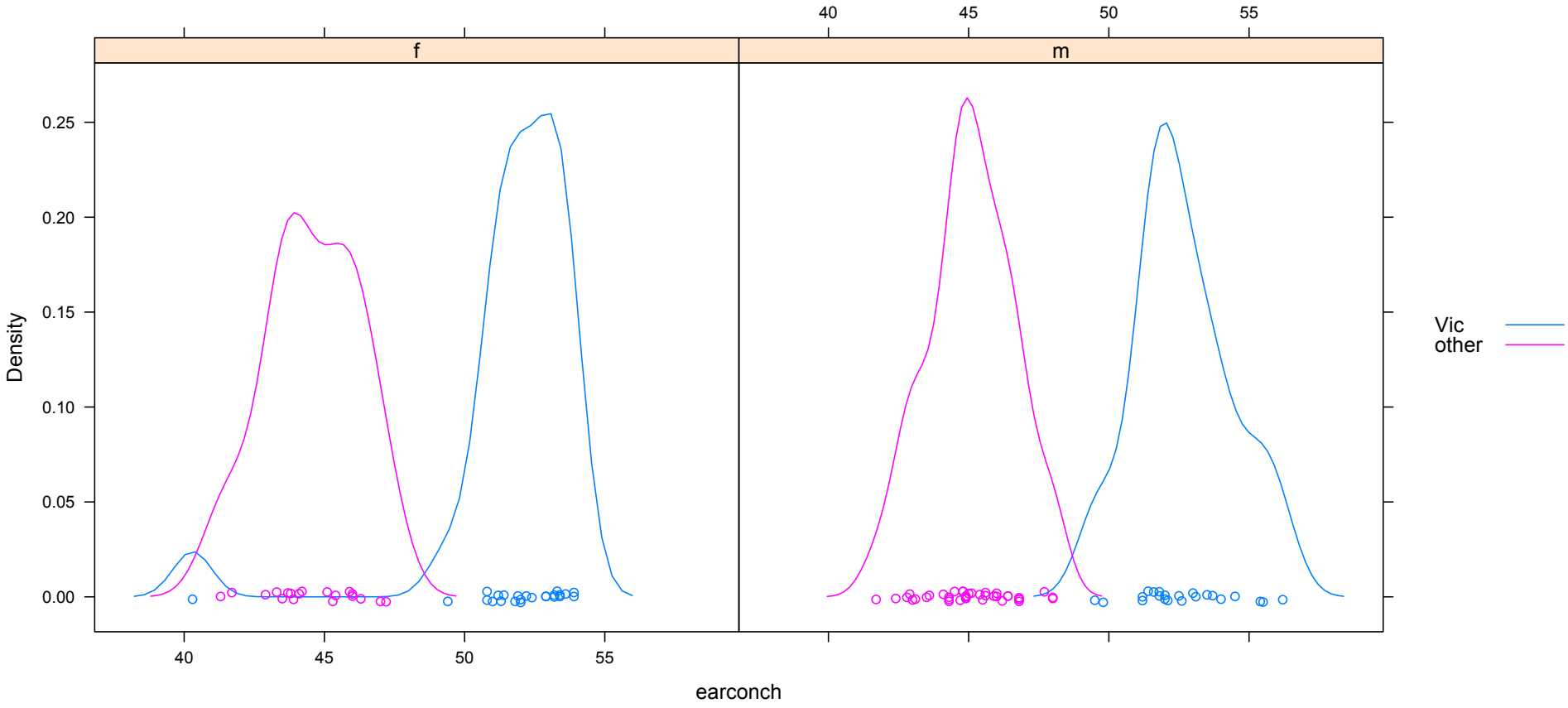
- Strip plots and box plots





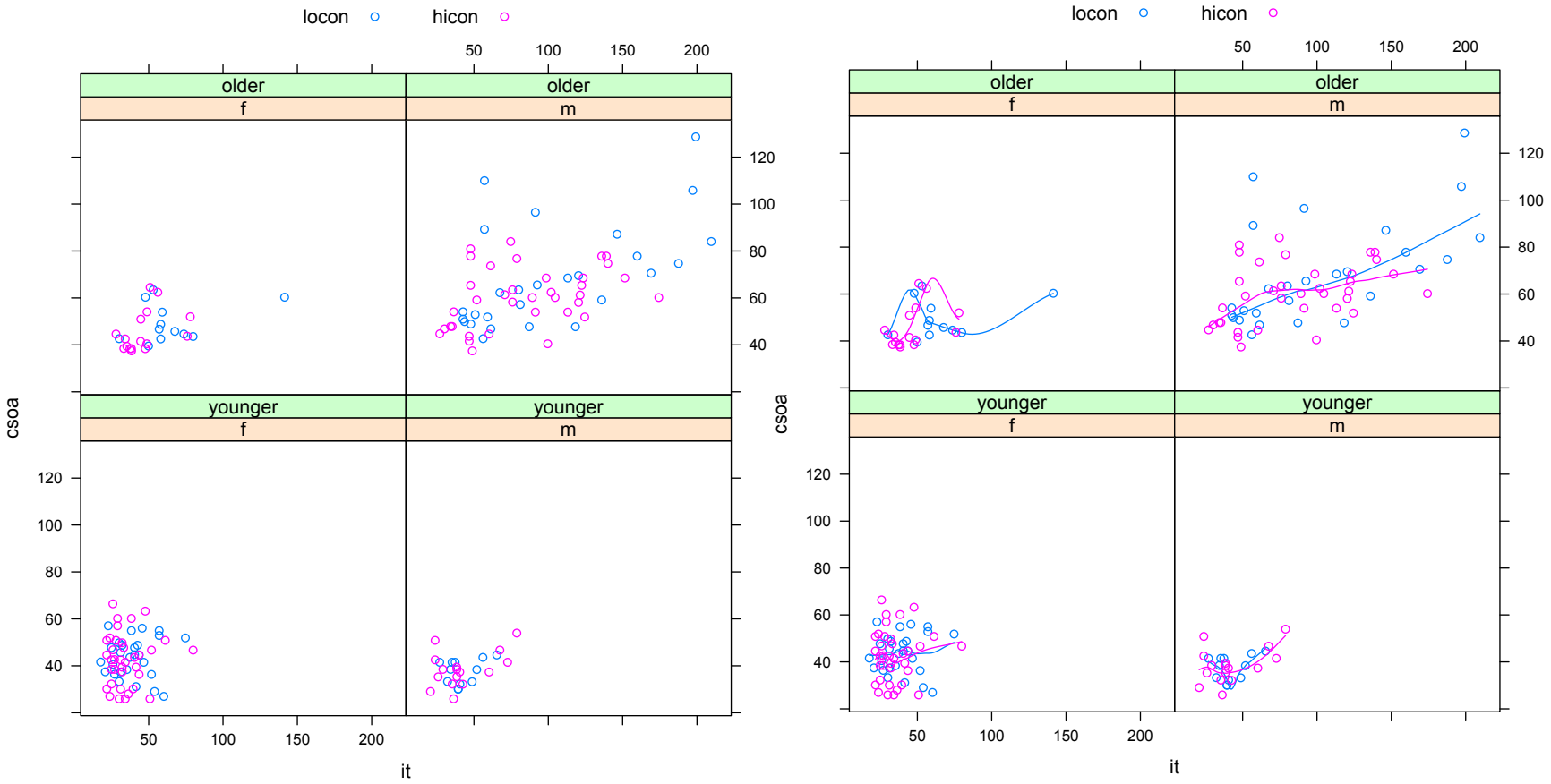
Views of groups of data …..

- Comparing density plots



Views of groups of data …..

- Scatterplots broken down by multiple factors



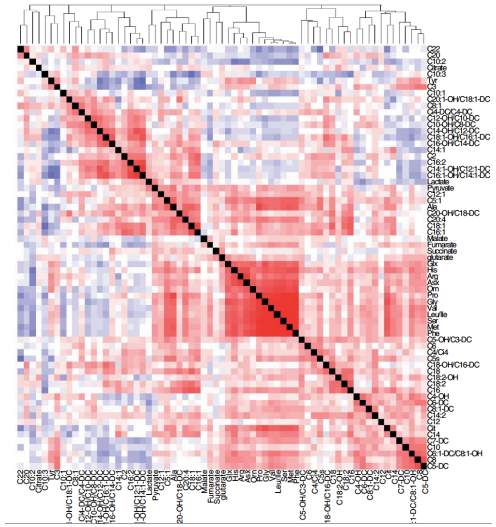
The analyst should look for …..

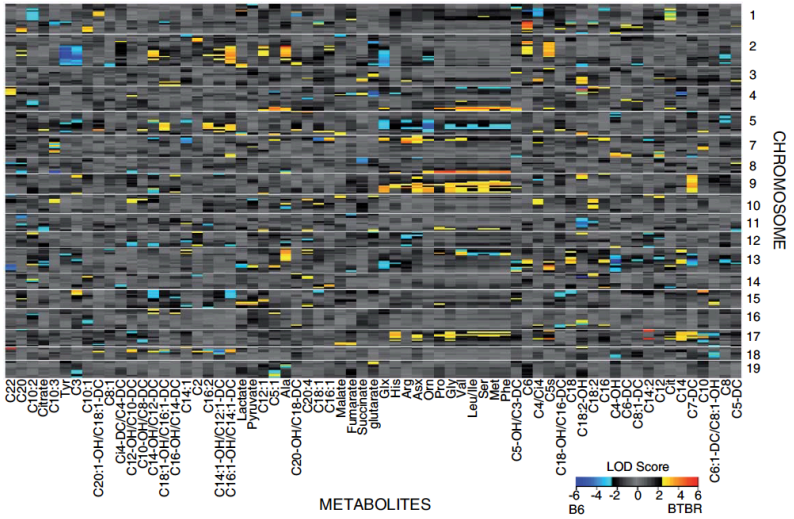
- Outliers.
- Missing Data.
- Clusters in the data.
- Unexpected patterns within groups.
- Between-groups differences in the scatter of data.
- Whether there are unanticipated trends associated, e.g., with the order of data collection, confounding of treatments etc.

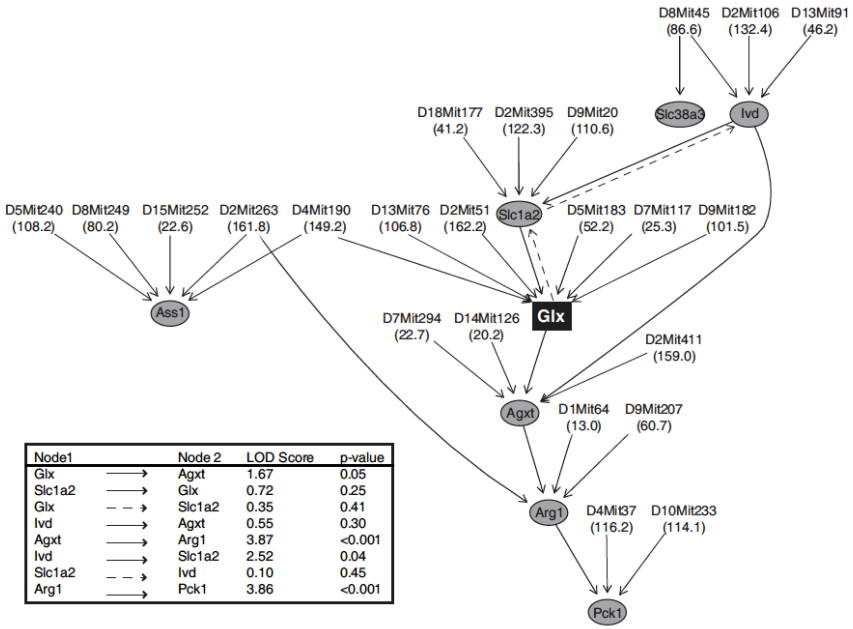
Knowledge Discovery in Databases (KDD)

1. Select the target data set - Which data or which variables and cases are to be used.
2. Data Cleaning – removal of noise, identification of potential outliers, impute missing data etc.
3. Preprocessing the data – decide upon the data transformation, track time-dependent information.
4. Decide which data mining technique is appropriate.
5. Analyze the cleaned data using data mining techniques.
6. Interpret and assess the knowledge derived from data mining results.

Be Skeptical …







Node1		Node 2	LOD Score	p-value
Glx	→	Agxt	1.67	0.05
Slc1a2	→	Glx	0.72	0.25
Glx	- - →	Slc1a2	0.35	0.41
Ivd	→	Agxt	0.55	0.30
Agxt	→	Arg1	3.87	<0.001
Ivd	→	Slc1a2	2.52	0.04
Slc1a2	- - →	Ivd	0.10	0.45
Arg1	→	Pck1	3.86	<0.001

Ferrara et al. Genetic Networks of Liver Metabolism Revealed Integration Of Metabolic and Transcriptional Profiling. *PLoS Genetics* 4(3) e1000034.