Master of Technology in Knowledge/Software Engineering

KE5107: Data Mining Methodology and Methods

Data Exploration and Visualisation

Fan Zhenzhen
Institute of Systems Science
National University of Singapore
E-mail: zhenzhen@nus.edu.sg

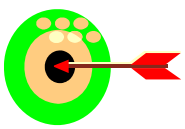
© 2016 NUS. The contents contained in this document may not be reproduced in any form or by any means, without the written permission of ISS, NUS, other than for the purpose for which it has been supplied.

Module Objectives

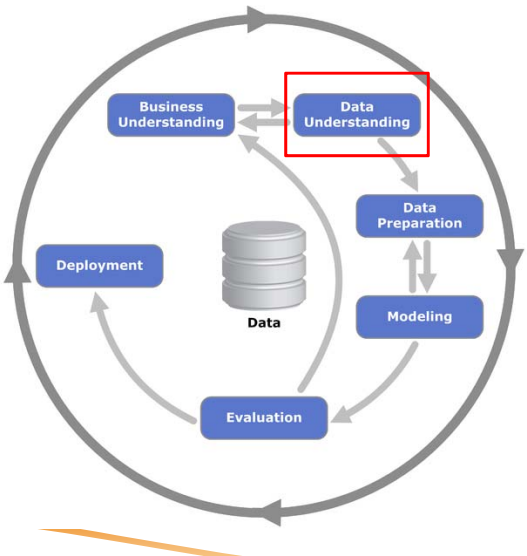
- To learn how to examine the underlying structure of the data.
- To learn how to examine data for any systematic relationship among the variables through workshops.
- To be familiar with some descriptive and graphic tools for exploring data visually in order to discover and confirm patterns and trends.

Agenda

- Data exploration goals
- Basic data visualisation methods
- Multi-dimensional visualisation
- Advanced visualisation techniques



CRISP Data Mining Methodology



© 2016 NUS. All rights reserved.

Data Exploration Goals

1. Data Quality Assessment

Before attempting to build a prediction model or do any data mining it is good practice to explore and profile the data to become familiar with it, determine its quality, identify missing values, inconsistencies, outliers etc

- Data errors should be fixed in the data cleaning and preparation phase.

© 2016 NUS. All rights reserved.

Goals for Exploratory Data Analysis

2. Knowledge Discovery

Visualisation provides a powerful means of finding patterns, trends, relationships, structure and exceptions in the data

- No specific goal is needed
- Very interactive and exploratory

© 2016 NUS. All rights reserved.

Goals for Exploratory Data Analysis

3. Find ways to improve the data set for modeling

Feature Selection

- Look for variables related to the target (useful for prediction)
- Look for variables that are "duplicates" of others (can delete)

Data Transformations

- Look for patterns and correlations that may suggest valuable transformations, e.g.

  chemical_ratio = Na/K

  weight_ratio = brain weight/body weight

© 2016 NUS. All rights reserved.

Goals for Exploratory Data Analysis

4. Data Profiling

Characterise a set of records in terms of their attributes. The characterisation will be a generalisation (i.e. need not be 100% correct)

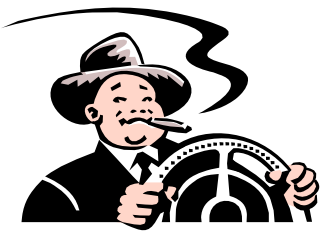
Common task is customer profiling, e.g.

- Most people who buy sports cars are young, rich & single!
- diabetes type2 patients are typically older & overweight
- Bank Y's investment account holders are usually married with children and a well paid job

© 2016 NUS. All rights reserved.

Data Exploration Principles

- Let the data speak for itself
  - Be open, make no prior assumptions
- Interpret what you see
  - Develop and explore hunches, then pursue this line of analysis until you confirm/disown it
  - Be interactive, pro-active



During data exploration you lead the way!

© 2016 NUS. All rights reserved.

Agenda

- Data Exploration Goals
- Basic Data Visualisation methods
- Advanced Visualisation Techniques

© 2016 NUS. All rights reserved.

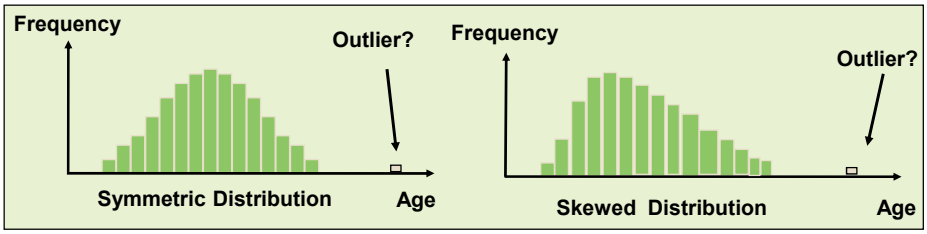
Data Exploration Basic Steps

1. Discover the "shape" & quality of the data
2. Look for outliers, extreme & unusual values
3. Look for relationships between variables
4. Build profiles of important subsets of the data

© 2016 NUS. All rights reserved.

Exploring Data "Shape"

- For numerical variable, histograms can tell us a lot about data spread, skewness, outliers

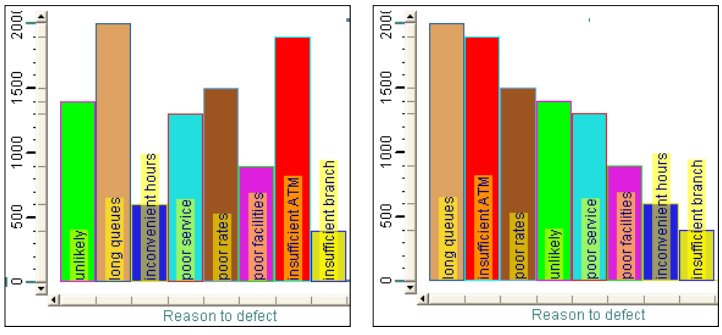


Does the spread look reasonable? Is the data flawed?
Are there outliers? Should we keep or discard them?
Should we attempt to remove skew, e.g. by a log transform?

© 2016 NUS. All rights reserved.

Exploring Data "Shape"

- For categorical variable use bar charts



Bars sorted by height

What are the most frequently and least frequently occurring categories? Is the distribution flat or unequal?

© 2016 NUS. All rights reserved.

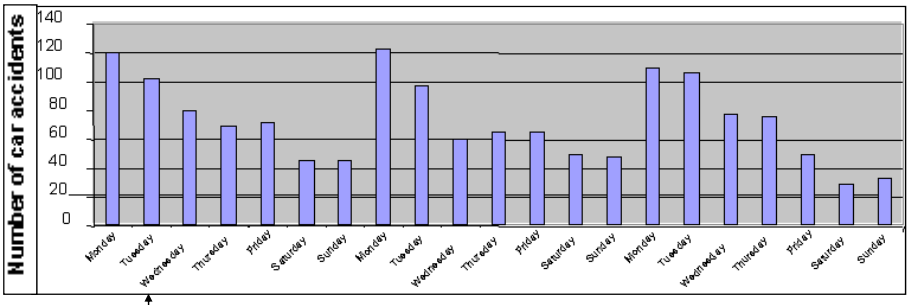
Pie Chart

- Visually appealing
- With limited features
- Hard to judge the size of pies
- Not recommended with a large number of categorical values

© 2016 NUS. All rights reserved.

Exploring Data "Shape"

- Correct bin size can be essential
- Trends and periodicity can depend on time scale



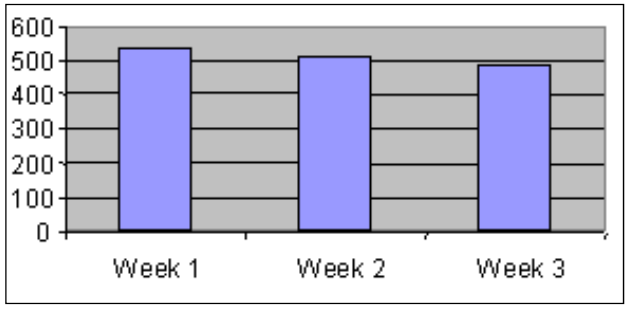
This bin size shows that a similar pattern repeated is every week

© 2016 NUS. All rights reserved.

Exploring Data "shape"

- A different bin size shows a different trend….

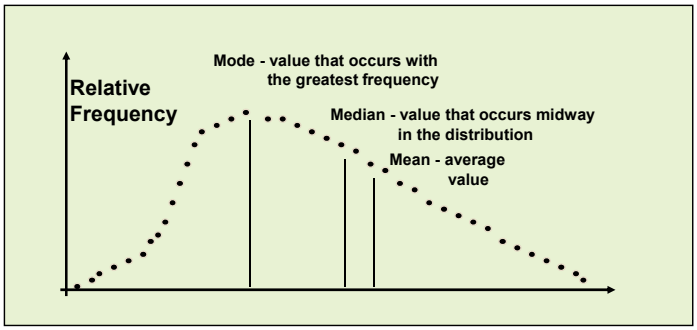
Weekly accidents decreasing



© 2016 NUS. All rights reserved.

Identifying Skew and Outliers

- Simple statistical measures such as *maxima*, *minima*, *mean*, *mode*, *median*, *variance* are informative and show up skews, outliers, noise (for numerical data)



© 2016 NUS. All rights reserved.

Data Summary

- Summary of these simple statistical measures can be presented as numbers , which are not easy to interpret.

```
   MinTemp          MaxTemp          Rainfall        Evaporation
Min.   :-5.300   Min.   : 8.40   Min.   : 0.000   Min.   : 0.600
1st Qu.: 1.925   1st Qu.:14.70   1st Qu.: 0.000   1st Qu.: 2.200
Median : 7.100   Median :19.20   Median : 0.000   Median : 4.000
Mean   : 7.011   Mean   :20.29   Mean   : 1.386   Mean   : 4.509
3rd Qu.:12.400   3rd Qu.:24.95   3rd Qu.: 0.200   3rd Qu.: 6.400
Max.   :20.900   Max.   :35.80   Max.   :25.800   Max.   :13.800
```

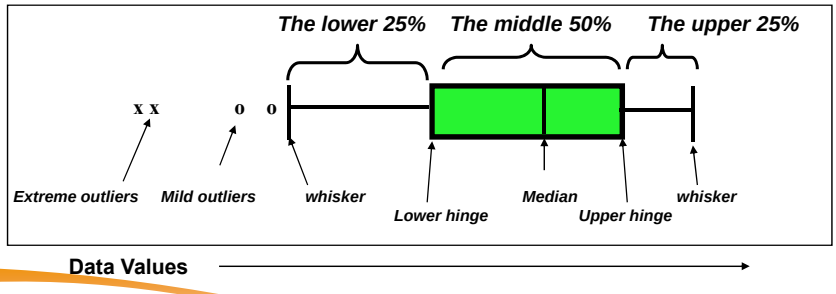
Skewness:

```
    MinTemp      MaxTemp      Rainfall   Evaporation
0.05347228   0.38465167   3.63725482   0.70497219
```

© 2016 NUS. All rights reserved.

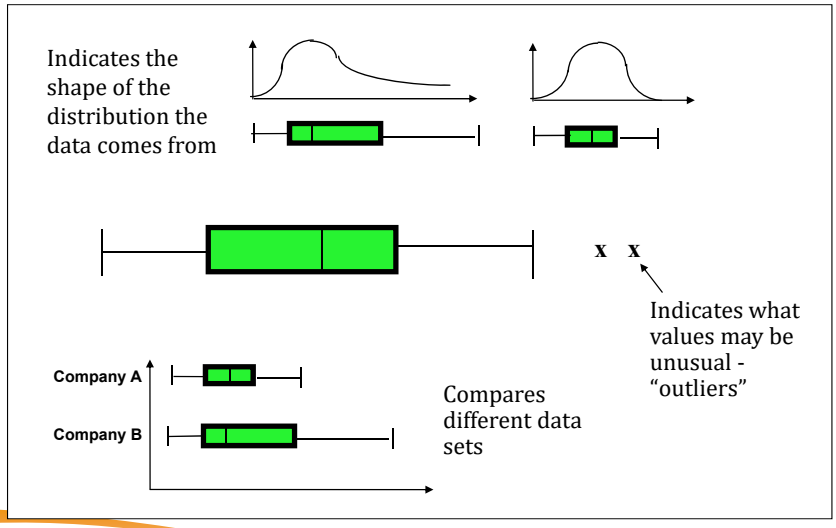
Box Plots (Box & Whisker)

- Show key statistical measures in a single picture!
  - Box shows the middle 50% of the data
  - The whiskers show the end points (excluding outliers)
  - Mild outliers lie between 1.5 to 3 times the box length on either side of the box, extreme outliers are beyond this



© 2016 NUS. All rights reserved.

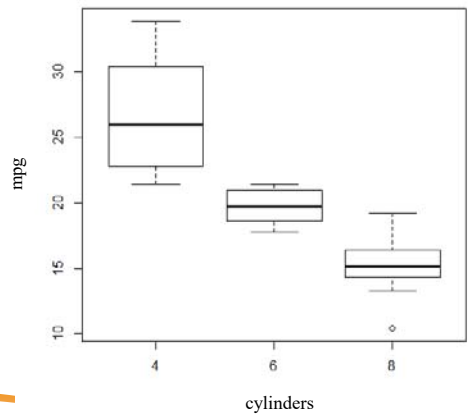
Using Box Plots



© 2016 NUS. All rights reserved.

Box Plot Examples

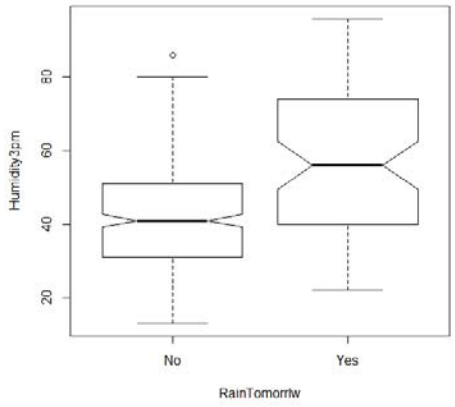
- Good for comparing spread and shape across different categories



© 2016 NUS. All rights reserved.

Box Plot Examples

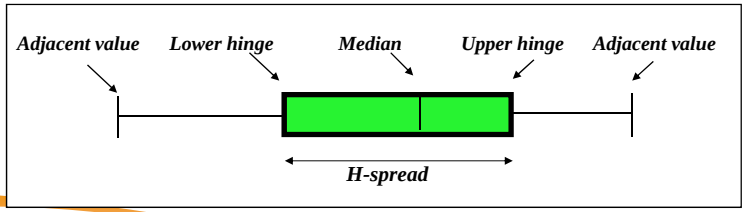
• The notches around the median indicate an approximate 95% confidence level for the differences between the medians



© 2016 NUS. All rights reserved.

How Box-Plots Work

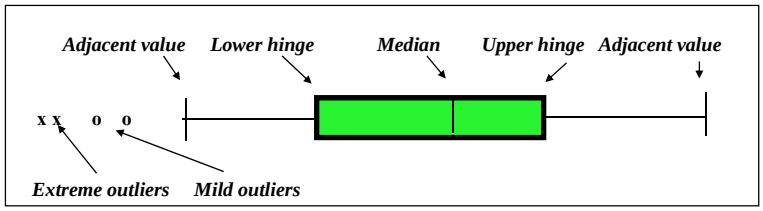
1. Calculate the Median and "Hinges"
   - The median splits the data sample into two parts
   - The hinges are the middle values of each part
   - The Upper and lower Hinges form the sides of the box
   - The difference between the hinges called the H-spread
2. Define inner fences
   - Lower inner fence = lower hinge - 1.5* (H-spread)
   - Upper inner fence = Upper hinge + 1.5* (H-spread)
3. The data values that are closest to, but still inside the inner fences are called *adjacent values*, these form the whiskers



© 2016 NUS. All rights reserved.

How Box-Plots Work

4. Define outer fences
   - Lower outer fence = Lower hinge - 3* (H-spread)
   - Upper outer fence = Upper hinge + 3* (H-spread)
5. Data values that are between the inner and outer fences are called *mild outliers*
6. Data values that are beyond the outer fence are called *extreme outliers*



© 2016 NUS. All rights reserved.

How Box-Plots Work: Example

53	158	32	40	71	34	125	37
62	30	44	40	39	35	86	64
34	31	52	32	26	56		

Arranged Data:

1	26
2	30
3	31
4	32
5	32
6	34
7	34
8	35
9	37
10	39
11	40
12	40
13	44
14	52
15	53
16	56
17	62
18	64
19	71
20	86
21	125
22	158

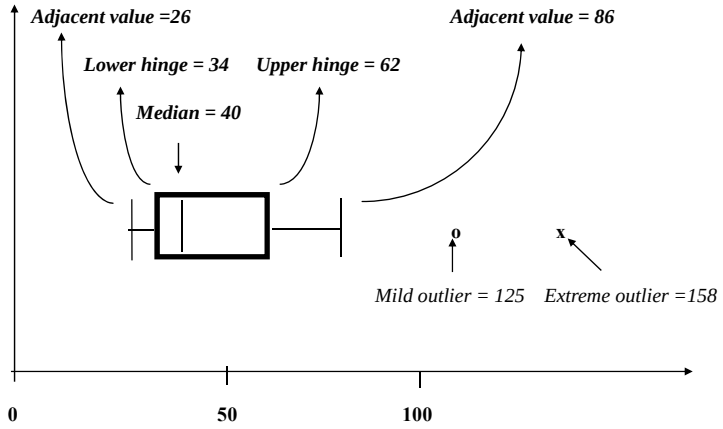
Median = (40+40)/2 = 40
Lower Hinge = 34
Upper Hinge = 62

H-spread = 62 - 34 = 28
Lower inner fence = 34 - 1.5*28 = -8 Adjacent value = 26
Upper inner fence = 62 + 1.5*28 = 104 Adjacent value = 86

Lower outer fence = 34 - 3*28 = -50
Upper outer fence = 62 + 3*28 = 146

© 2016 NUS. All rights reserved.

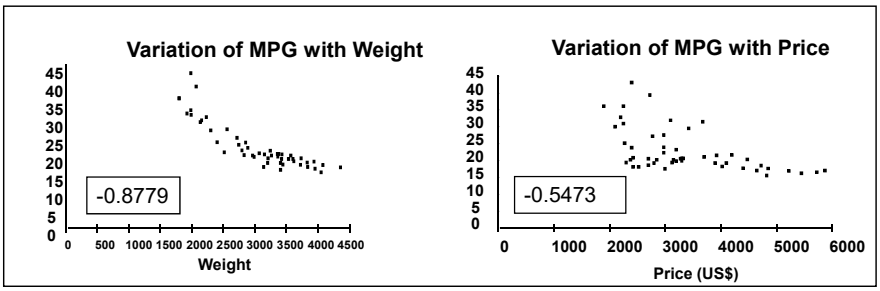
How Box-Plots Work: Example



© 2016 NUS. All rights reserved.

Finding Relationships

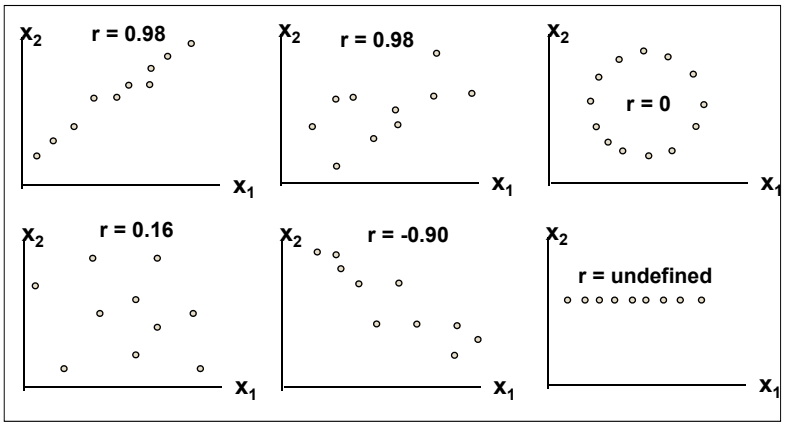
- Statistical correlation coefficient finds relationships between 2 numerical variables. Coeff ~ [-1,+1]



Beware: Correlation coefficient indicates straight line relationships only. Correlation does not always imply causality!

© 2016 NUS. All rights reserved.

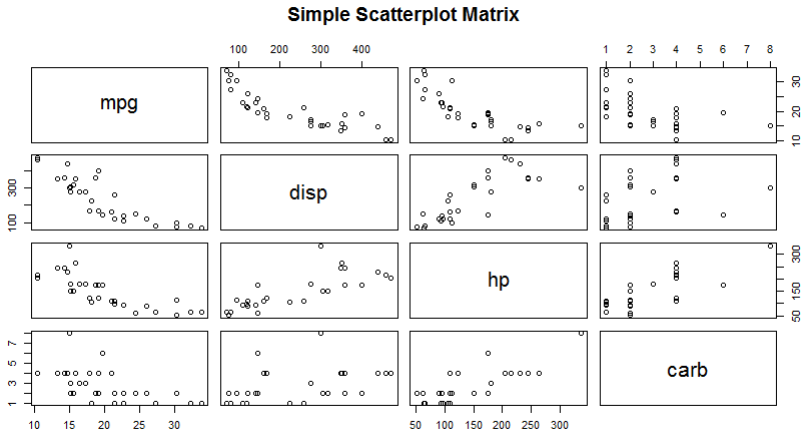
Correlation Coefficient



© 2016 NUS. All rights reserved.

Scatterplot Matrices

Use scatterplot matrix to examine pairwise correlation among a few numerical variables



© 2016 NUS. All rights reserved.

Multi-Dimensional Visualisation

How can we find relationships between 3+ variables?

Basic

- Scatter plots
- Overlays
- Co-Plots
- 3-D Graphics
- Many more….

More advanced

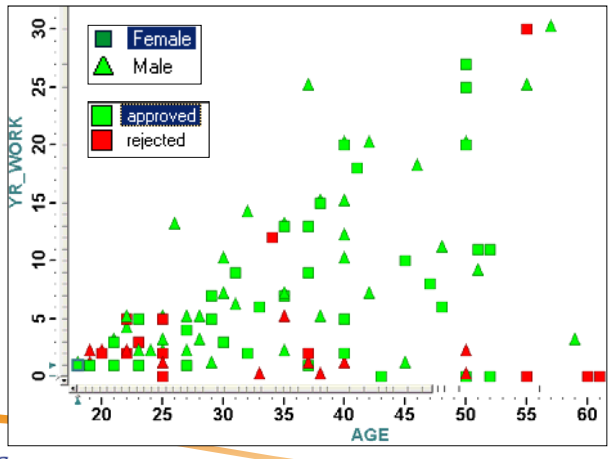
- Principle Component Analysis
- Multi-Dimensional Scaling
- Parallel Coordinates
- Star Plots
- Chernoff faces
- Many more….

© 2016 NUS. All rights reserved.

Scatter Plots

Show relationships between **numerical** variables

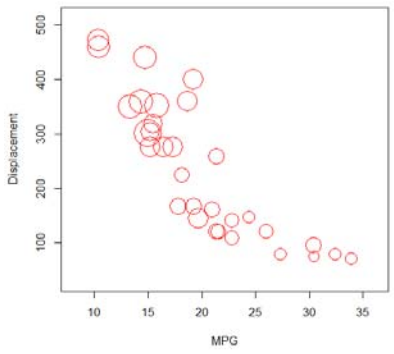
Can map additional variables to **point size**, **color**, **shape**



© 2016 NUS. All rights reserved.

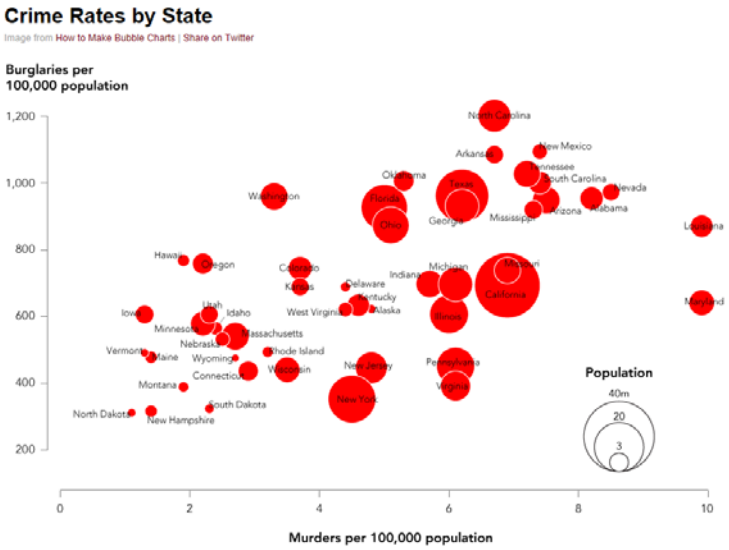
Bubble Plots

- Another name for scatter plots with a data variable assigned to **point size** (e.g. *hp*)



© 2016 NUS. All rights reserved.

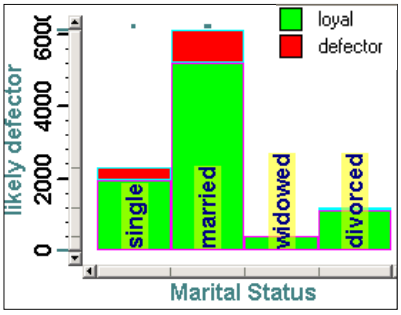
Bubble plot: another example



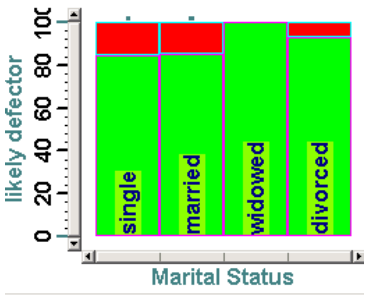
© 2016 NUS. All rights reserved.

Overlays

Stacked Bar Plot: Marital Status 2D Bar with defector overlaid as color



Making all bars the same height allows the relative proportions in each bar to be compared

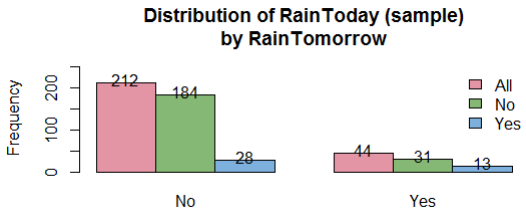


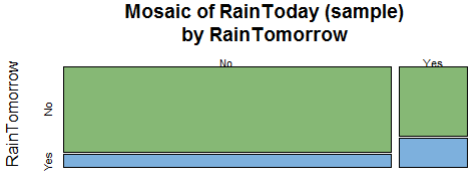
What Patterns can you see?

© 2016 NUS. All rights reserved.

Mosaic Plot

- A variation of normalized bar plot with overlaid categorical variable.

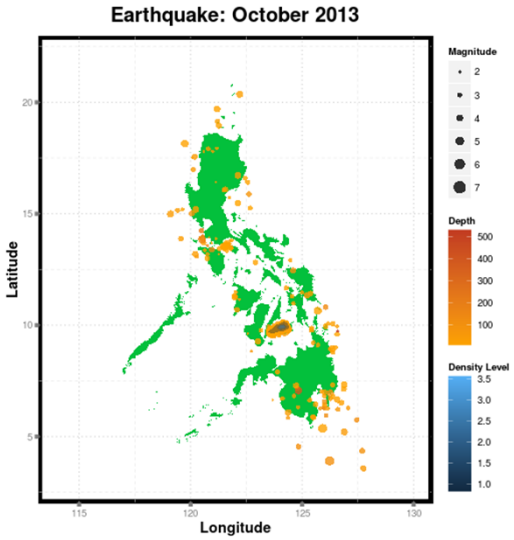




© 2016 NUS. All rights reserved.

Overlay on Maps

- Using longitude and latitude coordinates to overlay selected variables on geographic maps

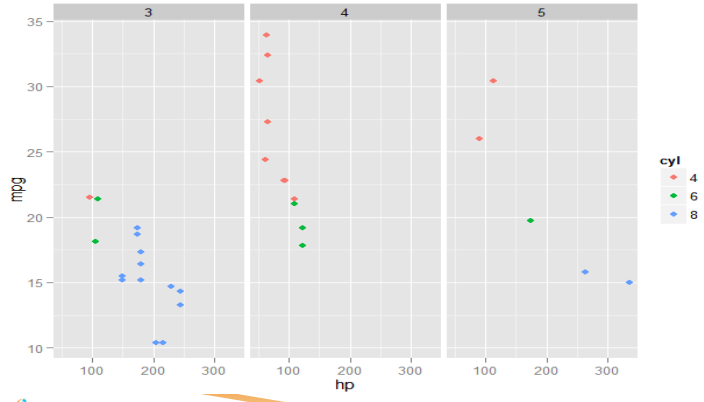


http://alstatr.blogspot.sg/2013/11/r-mapping-philippine-earthquakes.html

© 2016 NUS. All rights reserved.

Co-Plots

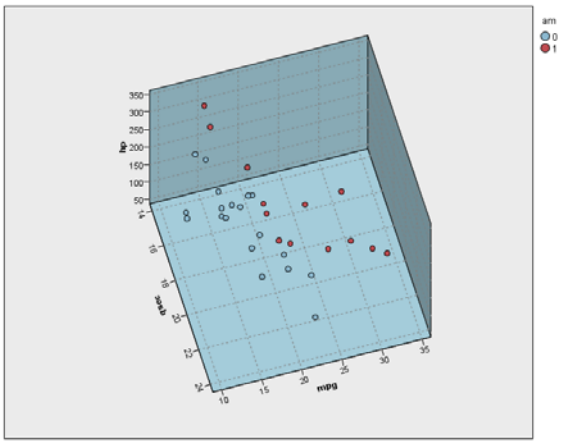
- Plots of two variables conditioned on the values of one or two other variables
  - Example: the variation of car MPG with horse power, conditional on the number of gears, with color showing the number of cylanders.



© 2016 NUS. All rights reserved.

3-D Data Visualisation

- 3-D offers more scope for viewing multiple dimensions
- Need interactivity (rotating and zooming) to make best use of 3D visualisation



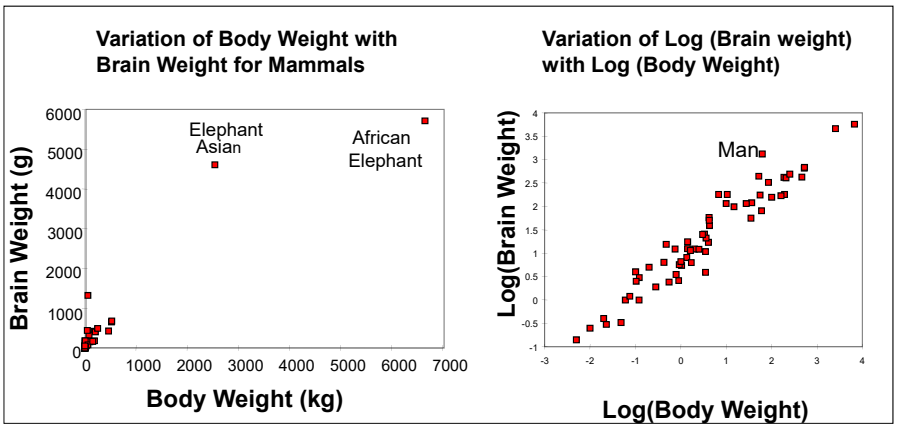
© 2016 NUS. All rights reserved.

Data Transformations for More Visible Patterns

- Can make patterns easier to see
- Examples
  - *Rescaling*
    Log() ~ reduce effects of large variables
    Square() ~ exaggerate effect of large variables
  - *Derived variables*
    profit = units sold * unit cost
    int.call rate = number international calls/total calls
    activity rate = average monthly transactions/ avg. balance
    Odds of an event = p / (1-p)

© 2016 NUS. All rights reserved.

Re-Scaling Example



Log Transformation ⟶

© 2016 NUS. All rights reserved.

Derived Variable Example

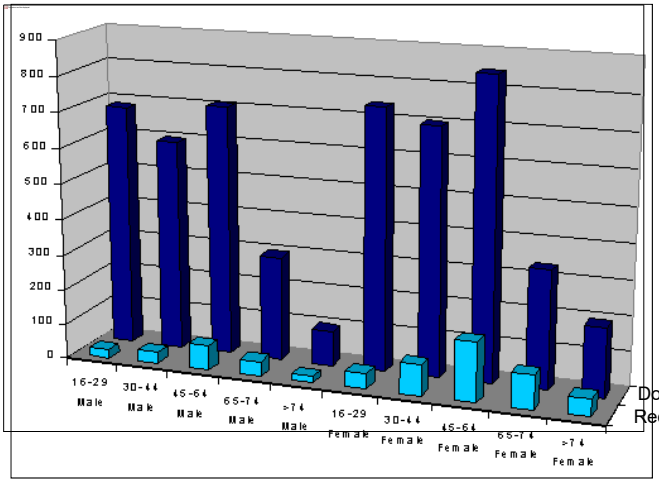
The following data describes the number of people who take psychotropic drugs from randomly taken samples classed by age and sex

Sex	age group	Mean age	Receiving drugs	Not receiving drugs
Male	16-29	23.2	21	683
Male	30-44	36.5	32	596
Male	45-64	54.3	70	705
Male	65-74	69.2	43	295
Male	>74	79.5	19	99
Female	16-29	23.2	46	738
Female	30-44	36.5	89	700
Female	45-64	54.3	169	847
Female	65-74	69.2	98	336
Female	>74	79.5	51	196

© 2016 NUS. All rights reserved.

Derived Variable Example

Is there a relationship between taking drugs and sex or age?



© 2016 NUS. All rights reserved.

Derived Variable Example

Use Log Odds transformation

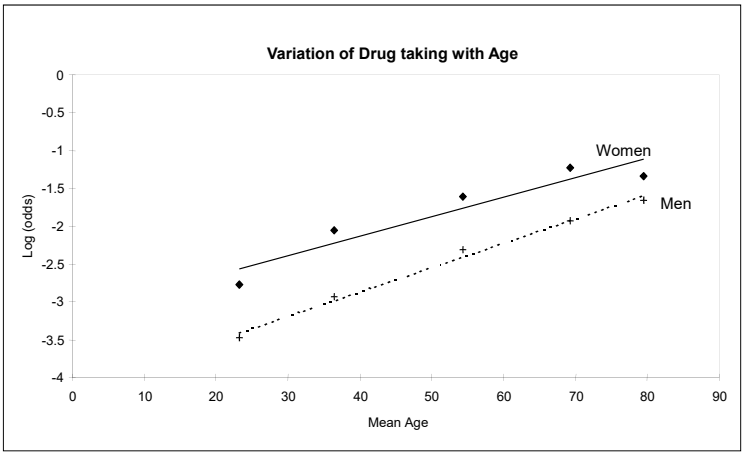
– Log (odds) = Log(No. receive/No. not receiving)

[Odds of an event = p / (1-p)]

Sex	Age group	Mean age	Receiving drugs	Not receiving drugs	Odds	Log(odds)
Male	16-29	23.2	21	683	0.0307	-3.48197
Male	30-44	36.5	32	596	0.0537	-2.9245
Male	45-64	54.3	70	705	0.0993	-2.3097
Male	65-74	69.2	43	295	0.1458	-1.92578
Male	>74	79.5	19	99	0.1919	-1.65068
Female	16-29	23.2	46	738	0.0623	-2.7753
Female	30-44	36.5	89	700	0.1271	-2.06244
Female	45-64	54.3	169	847	0.1995	-1.6118
Female	65-74	69.2	98	336	0.2917	-1.23214
Female	>74	79.5	51	196	0.2602	-1.34629

© 2016 NUS. All rights reserved.

Derived Variable Example



© 2016 NUS. All rights reserved.

Agenda

- Data Exploration Goals
- Basic Data Visualisation methods
- Advanced Visualisation Techniques

© 2016 NUS. All rights reserved.

Visualisation for Multivariate Data

- Visualising data with a large number of variables can be very challenging.
- One major approach is to simplify the problem by
  - Reducing the number of variables describing the data – Principal Component Analysis
  - Reducing the number of dimensions of the problem – Multi-dimensional Scaling
- Or find innovative ways to visualise many variables in one graph by
  - Showing links between many different categories – Link Analysis
  - Using parallel vertical axes - Parallel Coordinates
  - Using axes starting from one central point – Radar Chart
  - Using icons – e.g. Star Plot, Chernoff Faces

© 2016 NUS. All rights reserved.

Principle Component Analysis

- Used when some of the variables in a data set are correlated (there is some redundancy)
- We can find a smaller set of uncorrelated variables which are linear combinations of the original variables

$$Z_1 = a_{11} * X_1 + a_{12} * X_2 + \ldots\ldots$$

$$Z_i = a_{i1} * X_1 + a_{i2} * X_2 + \ldots\ldots$$

- These are the Principle Components (sometimes called factors), derived in decreasing order of importance - the earlier ones account for greater variations in the data
- PCA is performed to help:
  - Visually analyse and explore the data
  - Identify natural groupings or summaries of data
  - Create a set of factors that can later be used to create predictive models

© 2016 NUS. All rights reserved.

PCA Example

- Data describing the % of people employed in nine different industries in countries in Europe:
- Industries are
  - AGR (agriculture)
  - MIN (Mining)
  - MAN (Manufacture)
  - PS (Power Supplies)
  - CON (Construction)
  - SER (Service Industries)
  - FIN (Finance)
  - SPS (Social and Personal services)
  - TC (Transport & Communications)

© 2016 NUS. All rights reserved.

PCA Example

	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
Belgium	3.3	0.9	27.6	0.9	8.2	19.1	6.2	26.6	7.2
Denmark	9.2	0.1	21.8	0.6	8.3	14.6	6.5	32.2	7.1
France	10.8	0.8	27.5	0.9	8.9	16.8	6	22.6	5.7
W.Germa	6.7	1.3	35.8	0.9	7.3	14.4	5	22.3	6.1
Ireland	23.2	1	20.7	1.3	7.5	16.8	2.8	20.8	6.1
Italy	15.9	0.6	27.6	0.5	10	18.1	1.6	20.1	5.7
Luxembur	7.7	3.1	30.8	0.8	9.2	18.5	4.6	19.2	6.2
Netherlan	6.3	0.1	22.5	1	9.9	18	6.8	28.5	6.8
UK	2.7	1.4	30.2	1.4	6.9	16.9	5.7	28.3	6.4
Austria	12.7	1.1	30.2	1.4	9	16.8	4.9	16.8	7
Finland	13	0.4	25.9	1.3	7.4	14.7	5.5	24.3	7.6
Greece	41.4	0.6	17.6	0.6	8.1	11.5	2.4	11	6.7
Norway	9	0.5	22.4	0.8	8.6	16.9	4.7	27.6	9.4
Portugal	27.8	0.3	24.5	0.6	8.4	13.3	2.7	16.7	5.7
Spain	22.9	0.8	28.5	0.7	11.5	9.7	8.5	11.8	5.5
Sweden	6.1	0.4	25.9	0.8	7.2	14.4	6	32.4	6.8
Switzerlan	7.7	0.2	37.8	0.8	9.5	17.5	5.3	15.4	5.7
Turkey	66.8	0.7	7.9	0.1	2.8	5.2	1.1	11.9	3.2
Bulgaria	23.6	1.9	32.3	0.6	7.9	8	0.7	18.2	6.7
Czechosk	16.5	2.9	35.5	1.2	8.7	9.2	0.9	17.9	7
E.German	4.2	2.9	41.2	1.3	7.6	11.2	1.2	22.1	8.4
Hungary	21.7	3.1	29.6	1.9	8.2	9.4	0.9	17.2	8
Poland	31.1	2.5	25.7	0.9	8.4	7.5	0.9	16.1	6.9
Romania	34.7	2.1	30.1	0.6	8.7	5.9	1.3	11.7	5
USSR	23.7	1.4	25.8	0.6	9.2	6.1	0.5	23.6	9.3
Yugoslavia	48.7	1.5	16.8	1.1	4.9	6.4	11.3	5.3	4

© 2016 NUS. All rights reserved.

PCA Example

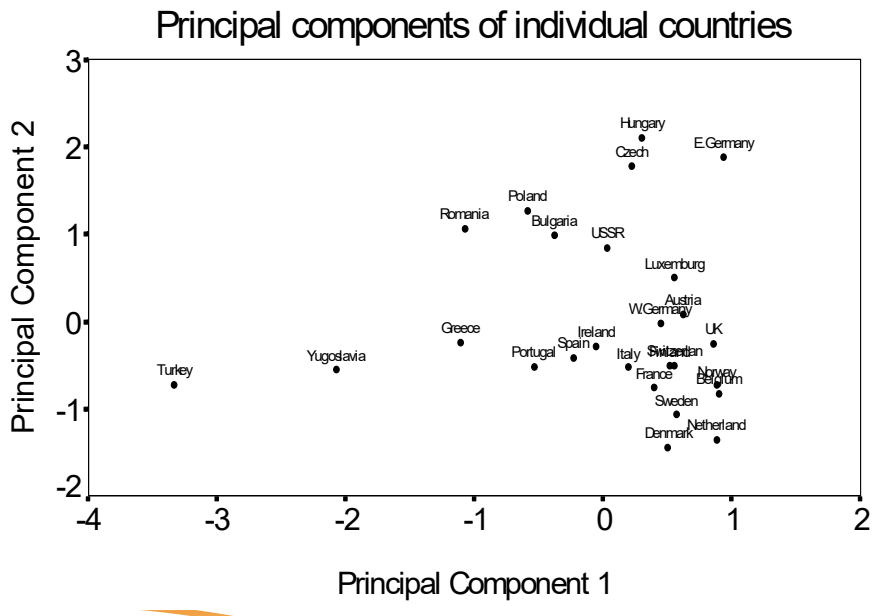
- Looking at the first two PC's can show groupings in the data

Coefficients of Principal components:

	Princ.Comp.1	Princ.Comp. 2
AGR	-0.97812	0.07822
MIN	-0.00247	0.9017
MAN	0.64891	0.5182
PS	0.47752	0.38107
CON	0.60724	0.07486
SER	0.70759	-0.51108
FIN	0.13888	-0.66218
SPS	0.72344	-0.32331
TC	0.685	0.29569

- But we must plot them in a scatter plot first!

© 2016 NUS. All rights reserved.



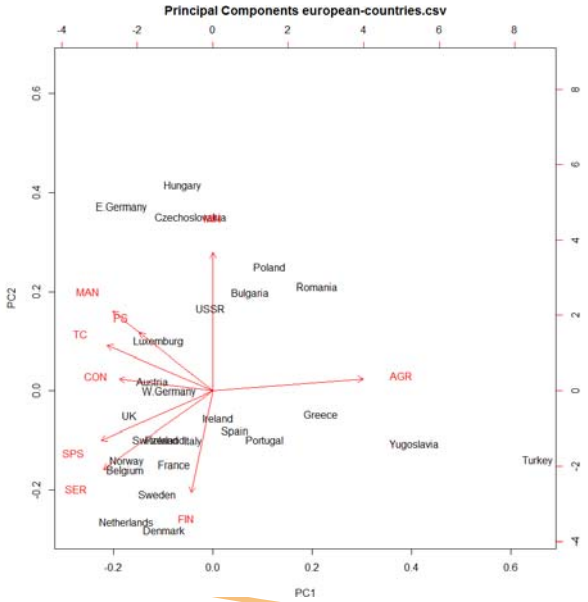
© 2016 NUS. All rights reserved.

Bi-Plots

- To allow display of both the observations and variables of a matrix of multivariate data on the same plot.
- Associated with PCA and MDS, or other dimension reduction techniques
- Observations represented as points
- Variables represented as vectors
- Interpretation:
  - Angle between vectors ~ correlation between fields
  - Cosine of the angle between a vector and an axis ~ its importance contribution to the axis dimension
  - Points close to each other ~ observations with similar values

© 2016 NUS. All rights reserved.

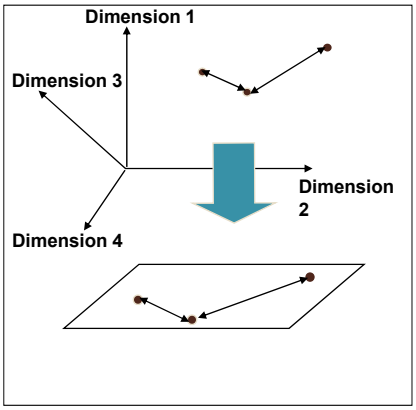
Bi-Plot Example



© 2016 NUS. All rights reserved.

Multi-Dimensional Scaling

- Alternative to PCA
- Maps objects from N-dimensional space to M-dimensional space while preserving distances between them (M<N)
- Not an exact procedure - moves objects around in the smaller M-dimensional space and checks how well the real distances between them are reproduced in the new configuration (maximizes the goodness-of-fit)
- Good when relationships between objects (variables) are not known but a distance matrix can be estimated from similarities / differences between objects (PCA relies on the correlation matrix to indicate distances between objects)



© 2016 NUS. All rights reserved.

Analysis of Breakfast Cereals (1)

Consider twenty three breakfast cereals:

- All Bran (AllB)
- All Bran with extra fibre (AllF)
- AppleJacks(AppJ)
- Cornflakes(CorF)
- Corn Pops (CorP)
- Cracklin' Oat Bran (Crac)
- Crispix (Cris)
- Froot Loops (Froo)
- Frosted Flakes (FroF)
- Frosted Mini-Wheats (FrMW)
- Fruitful Bran (FruB)
- Just Right Crunch Nuggets (JRCN)
- Just Right Fruit & Nut (JRFN)
- Meusliz Crispy blend (MuCB)
- Nut & Honey Crunch (Nut&)
- Nutri Grain Almond Raisin (NGAR)
- Nutri Grain Wheat (NutW)
- Product 18 (prod)
- Raisin Bran (RaBr)
- Raisin Squares (RaiS)
- Rice Krispies (RiKr)
- Sugar Smacks (Smac)
- Special K (Spec)

Data collected:

- Number of calories
- Amount of protein (g)
- Amount of fat (g)
- Amount of sodium (mg)
- Amount of dietary fibre (g)
- Amount of complex carbohydrates (g)
- Amount of sugars (g)
- Display shelf position (shelf 1, 2, or 3)
- Vitamin & mineral content (0, 25 or 100)
- Type of cereal (Hot or cold)

© 2016 NUS. All rights reserved.

Analysis of Breakfast Cereals (2)

The raw data….

AllB	0.1818	0.6	0.3333	0.8125	0.6429	0	0.3333	3	0.9677	0
AllF	0	0.6	0	0.4375	1	0.0667	0	3	1	0
AppJ	0.5455	0.2	0	0.3906	0.0714	0.2667	0.9333	2	0.0323	0
CorF	0.4545	0.2	0	0.9063	0.0714	0.9333	0.1333	1	0.0484	0
CorP	0.5455	0	0	0.2813	0.0714	0.4	0.8	2	0	0
Crac	0.5455	0.4	1	0.4375	0.2857	0.2	0.4667	3	0.4516	0
Cris	0.5455	0.2	0	0.6875	0.0714	0.9333	0.2	3	0.0323	0
Froo	0.5455	0.2	0.3333	0.3906	0.0714	0.2667	0.8667	2	0.0323	0
FroF	0.5455	0	0	0.625	0.0714	0.4667	0.7333	2	0.0161	0
FrMW	0.4545	0.4	0	0	0.2143	0.4667	0.4667	2	0.2581	0
FruB	0.6364	0.4	0	0.75	0.3571	0.4667	0.8	3	0.5484	0
JRCN	0.5455	0.2	0.3333	0.5313	0.0714	0.6667	0.4	3	0.129	1
JRFN	0.8182	0.4	0.3333	0.5313	0.1429	0.8667	0.6	3	0.2419	1
MuCB	1	0.4	0.6667	0.4688	0.2143	0.6667	0.8667	3	0.4516	0
Nut&	0.6364	0.2	0.3333	0.5938	0	0.5333	0.6	2	0.0645	0
NGAR	0.8182	0.4	0.6667	0.6875	0.2143	0.9333	0.4667	3	0.3548	0
NutW	0.3636	0.4	0	0.5313	0.2143	0.7333	0.1333	3	0.2258	0
Prod	0.4545	0.4	0	1	0.0714	0.8667	0.2	3	0.0806	1
RaBr	0.6364	0.4	0.3333	0.6563	0.3571	0.4667	0.8	2	0.7097	0
Rais	0.3636	0.2	0	0	0.1429	0.5333	0.4	3	0.2903	0
RiKr	0.5455	0.2	0	0.9063	0	1	0.2	1	0.0484	0
Smac	0.5455	0.2	0.3333	0.2188	0.0714	0.1333	1	2	0.0645	0
Spec	0.5455	1	0	0.7188	0.0714	0.6	0.2	1	0.1129	0

© 2016 NUS. All rights reserved.

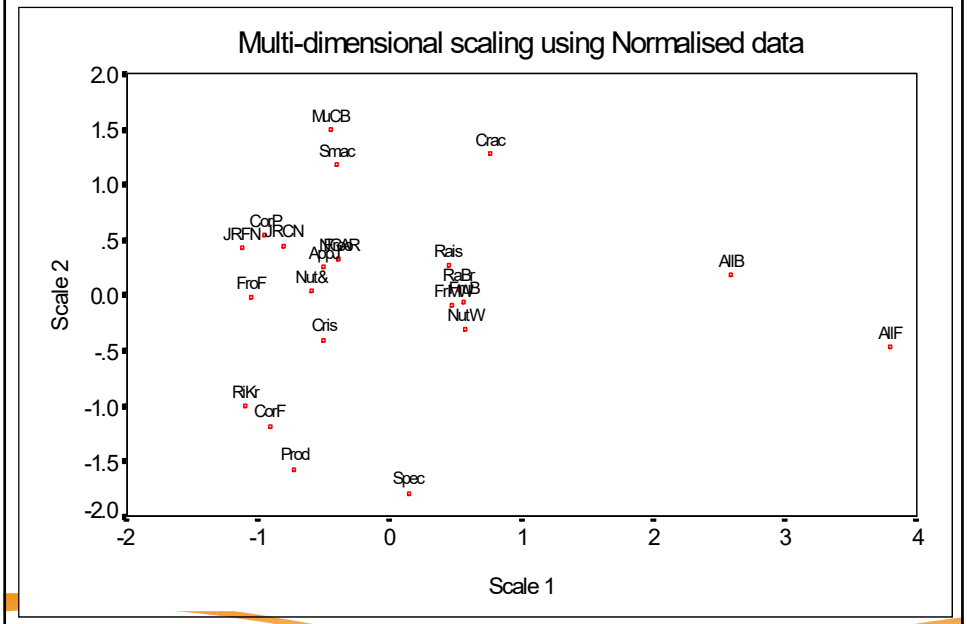
Analysis of Breakfast Cereals (3)

After MDS….

Cereal	x	y
AllB	-1.94950	.96880
AllF	-2.54750	1.89990
AppJ	.50930	.52080
CorF	1.74630	-.06280
CorP	.62080	.43560
Crac	-1.50940	.26850
Cris	.19300	-.76100
Froo	.36430	.39810
FroF	1.48990	.23990
FrMW	.06770	.64070
FruB	-.65360	.08110
JRCN	.00790	-1.61720

Cereal	x	y
JRFN	-.09340	-1.85250
MuCB	-.75230	-.57030
Nut&	.36060	.00250
NGAR	-.49820	-.79920
NutW	-.26060	-.12870
Prod	.29640	-2.10090
RaBr	-.25810	.52940
Rais	-.55240	.15500
RiKr	1.83320	-.29600
Smac	.25140	1.04770
Spec	1.33440	1.00080

© 2016 NUS. All rights reserved.



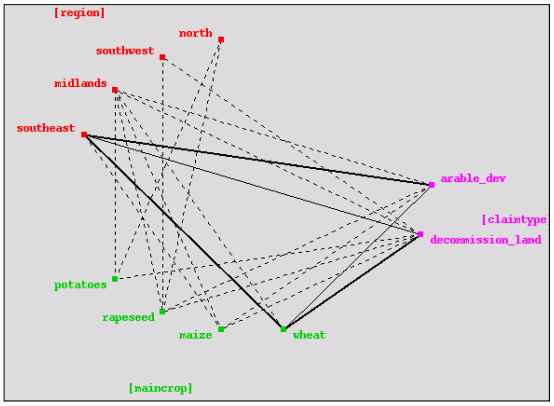
© 2016 NUS. All rights reserved.

Link Analysis

For relationships between the different categories in categorical fields.

Strong links are plotted with thick, dark lines

Very useful for Market Basket Analysis to discover which products tend to be purchased together

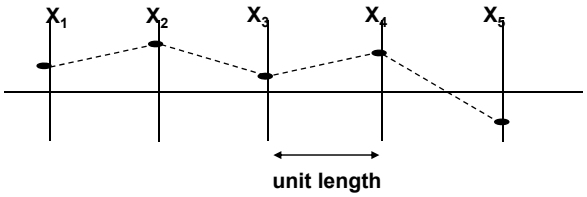


IBM SPSS Modeler Web Chart

© 2016 NUS. All rights reserved.

Parallel Coordinates

- 2-D Visualisation of multi-dimensional data, *Inselberg (IBM, early '80s)*
- Map N-dimensions to a series of parallel vertical axes, one per variable
- A line joining a point on each axis ~ a point in the N-dimensional space



- Good for generally numerical datasets with a moderate number of dimensions and no more than a few thousand records.

© 2016 NUS. All rights reserved.

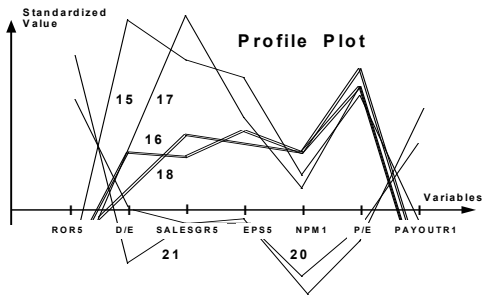
Parallel Coordinates

Similar to profile plots

Can help identify similar records, useful variables

TABLE 2: STANDARDIZED VALUES

ID#	ROR5	D/E	SALESGR5	EPS5	NPM1	P/E	PAYOUTR1
15	-0.47	3.70	2.91	2.54	0.67	2.22	-0.14
16	-0.58	0.37	1.39	1.23	1.07	2.42	-1.99
17	-0.77	0.93	2.77	1.36	0.27	1.81	-0.85
18	-0.55	0.74	0.67	1.04	0.76	1.81	-0.84
20	1.79	0.00	-0.19	-0.19	-1.24	-0.44	1.54
21	3.00	-0.93	-0.23	-0.19	-1.20	-0.24	1.37

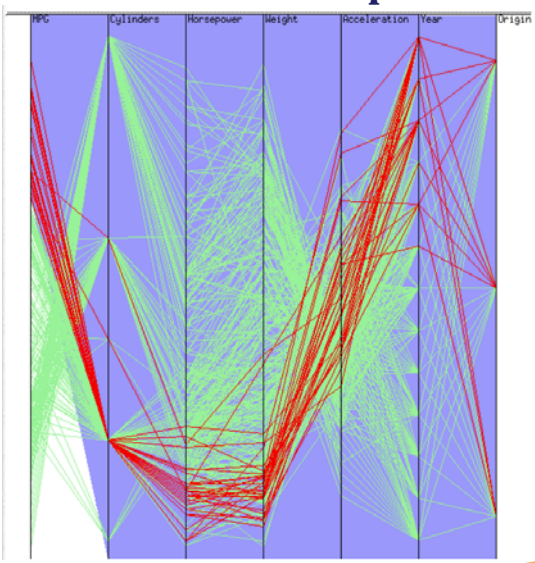


© 2016 NUS. All rights reserved.

Parallel Coordinates Example

XmdvTool: public-domain software based on parallel co-ords (http://davis.wpi.edu/~xmdv)

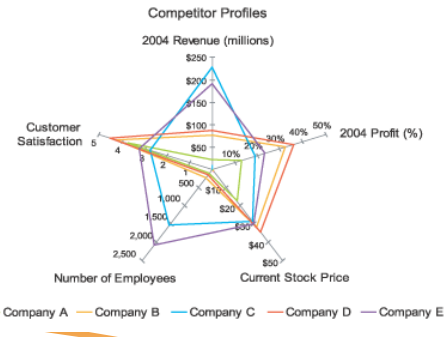
Use brushing to highlight a subset of observations to examine their profiles



© 2016 NUS. All rights reserved.

Radar Chart

- Displaying multivariate data in the form of a two-dimensional chart of three or more quantitative variables represented on axes starting from the same point
- Also known as spider chart, cobweb chart, etc.
- Used to discover similar observations and outliers in multivariate data.

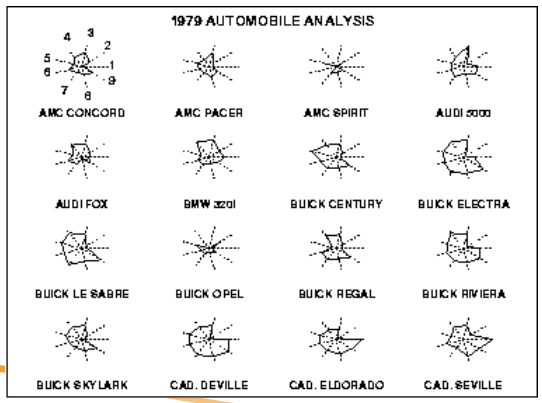


http://www.scottlogic.com/blog/2011/09/23/a-critique-of-radar-charts.html

© 2016 NUS. All rights reserved.

Icon-based Multi-Dimensional Visualization

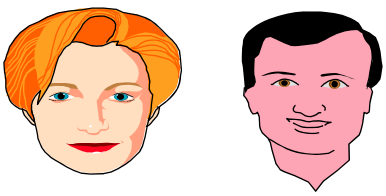
- Using icons like stars, stick figures, shape coding, color icons, etc. to discover similar or dissimilar observations.
- **Star Plot**: Each star represents a single observation.
- Helpful for small dataset



© 2016 NUS. All rights reserved.

Icon-based Multi-Dimensional Visualization

- Chernoff Faces: Each multi-dimensional observation is represented as a face with facial features determined by the component values. These features are
  - Upper hair
  - Chin curve
  - Lower hair
  - Eye size
  - Lip size
  - Eye space
  - Eye slant
  - Lip curve
  - Face size (eyes to mouth)



© 2016 NUS. All rights reserved.

Chernoff Faces Example

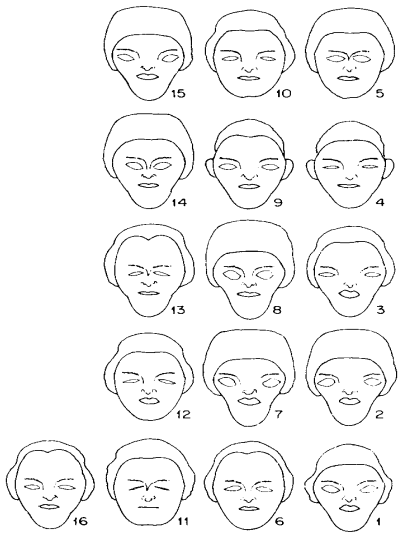
- Data below gives the number of crimes per 100,000 population in several American cities

	Murder Man-slaughter	Rape	Robbery	Assault	Burglary	Largency	Auto Theft
Atlanta	16.5	24.8	106	147	1112	905	494
Boston	4.2	13.3	122	90	982	669	954
Chicago	11.6	24.7	340	242	808	609	645
Dallas	18.1	34.2	184	293	1668	901	602
Denver	6.9	41.5	173	191	1534	1368	780
Detroit	13	35.7	477	220	1566	1183	788
Hartford	2.5	8.8	68	103	1017	724	468
Honolulu	3.6	12.7	42	28	1457	1102	637
Houston	16.8	26.6	289	186	1509	787	697
Kansas City	10.8	43.2	255	226	1494	955	765
Los Angeles	9.7	51.8	286	355	1902	1386	862
New Orleans	10.3	39.7	266	283	1056	1036	776

© 2016 NUS. All rights reserved.

Chernoff Faces Example

1 Atlanta
2 Boston
3 Chicago
4 Dallas
5 Denver
6 Detroit
7 Hartford
8 Honolulu
9 Houston
10 Kansas City
11 Los Angeles
12 New Orleans
13 New York
14 Portland
15 Tucson
16 Washington



- In this example, the lips indicate the number of burglaries: Thin lips = high, Full lips = low

© 2016 NUS. All rights reserved.

Summary: Understanding your Data

- **Establish Data Quality**
  - Rubbish in – rubbish out!
- **Knowledge Discovery**
  - Look for useful patterns, relationships, trends
  - Be open, don't be limited by preconceived ideas!
- **Identify Data Enhancements**
  - Identify what has to be done in data preparation

© 2016 NUS. All rights reserved.