Data Preparation

The need for data preparation

- Accurate and effective results achievable only when the input data are reliable
- Data from primary sources gathered into a data mart may have several anomalies
- Techniques to create high quality datasets:
  - Data validation: identify and remove anomalies and inconsistencies
  - Data integration and transformation: improve accuracy and efficiency of learning algorithms
  - Data reduction and disretization: fewer attributes without losing information

Data validation

Quality of data may not be satisfactory due to:

- Incompleteness: not recorded, unavailable, malfunctioning recording devices, deliberate removal, failure in data transfer.
- Noise: error or anomalous values, outliers, malfunctioning devices for data measurement, recording and transmission. Error in conversion of measurement units.
- Inconsistency: changes in coding system used for data representation.

Incomplete data

Techniques to correct incomplete data:

- Elimination: discard all records with one or more missing attribute values. In a supervised learning, remove records with missing target values. There is a risk of losing substantial amount of information.
- Inspection: let an expert inspect and replace missing values. This could be subjective and arbitrary.
- Identification: encode and identify missing values. Example: a continuous attribute with only positive value, assign a value of -1 to all missing data. For a categorical attribute, assign a new value that is different from other valid values.
- Substitution: Automatic replacement of missing values, for example for a missing numerical attribute value, compute the mean of the other observations. Other ways of substitution can be used, for example, regression.

Data affected by noise

Techniques to identify and correct data affected by noise:

- Dispersion: if we assume that the data distribution is roughly normal, with a confidence of 100(1 – α)%, we consider as outliers those values of attribute $\mathbf{a_j}$ that fall outside the interval $(\overline{\mu_j} - z_{\alpha/2}\,\overline{\sigma_j}, \overline{\mu_j} + z_{\alpha/2}\overline{\sigma_j})$.

  For α = 0.05, we are about 95% confident.

  Outliers are then replaced by values that are more plausible, or remove the records entirely.
- Clustering method: Clusters are records having mutual distance that is less than the distance from the records in the other groups. Observations not in any clusters are considered outliers.
- A variant of the clustering method: An observation $\mathbf{x_i}$ is identified as an outlier if at least a percentage *p* of the observations in the dataset are found at a distance greater than *d* from $\mathbf{x_i}$.

Data transformation

Data transformation may be applied to improve the accuracy of the models.

- Standardization: also called normalization can be achieved as follows:

  - **Decimal scaling:**

$$x'_{ij} = x_{ij} / 10^h$$

where h is scaling intensity parameter: shift the decimal point to the left by h positions.

  - **Min-max:**

$$x'_{ij} = (x_{ij} - x_{min,j}) \times \delta + x'_{min,j}$$

where $\delta = (x'_{max,j} - x'_{min,j}) / (x_{max,j} - x_{min,j})$

E.g. transforming the values of attribute $\mathbf{a_j}$ into the interval $[x'_{min,j}, x'_{max,j}] = [0,1]$:

$$x'_{ij} = (x_{ij} - x_{min,j}) / (x_{max,j} - x_{min,j})$$

or into the interval [-1,+1]:

$$x'_{ij} = 2\,(x_{ij} - x_{min,j}) / (x_{max,j} - x_{min,j}) - 1$$

Data transformation

- **z-index:**

$$x'_{ij} = (x_{ij} - \bar{\mu}_j)/\bar{\sigma}_j$$

where $\bar{\mu}_j$ and $\bar{\sigma}_j$ are the sample mean and sample standard deviation of the attribute $\mathbf{a_j}$.

If the distribution of the values of $\mathbf{a_j}$ is approximately normal, the generated values of $\mathbf{x'_{ij}}$ should be within the range (-3,+3).

- Feature extraction:

- More complex data transformation to generate new data attributes.

- For example: a new attribute to capture trends in customer's spending computed as ratios or differences between spending amounts of contiguous periods.

- Other more complex transformations in Support Vector Machines, Neural Networks, Hybrid Machine Learning method.

Data reduction

- When dealing with large datasets, it may also be appropriate to reduce their size to improve the efficiency of the machine learning algorithms.
- Three criteria to consider before applying data reduction:
  - Efficiency: shorter computation time
  - Accuracy should not be compromised by data reduction. Data reduction based on attribute selection will lead to models with higher generalization capability on future data.
  - Simplicity: models can be translated into easy to understand rules. There is a trade-off between rule simplicity and accuracy.

Data reduction

Various possible ways to achieve data reduction:

1. Sampling: reducing the number of observations.

- Simple sampling: each record in the dataset has equal chances of being selected

- Stratified sampling: the population is divided into non-overlapping groups, and samples are selected from these groups proportionately. For example, if 60% of the population are males, 40% females. A sample size of 50 is selected, then there should be 30 males and 20 females.

Data reduction

2. Feature selection: subset of irrelevant/redundant features are eliminated before data mining begins resulting in reduction in the numbers of features/attributes.

a) Filter method:

- Relevant attributes are selected before moving on to the learning phase.
- Independent of learning methods used.
- Example: Compute the correlation between each input attribute and the target attribute. Those with correlations below a threshold are excluded from the learning phase.

b) Wrapper method:

- The selection of predictive attributes is based not only on the relevance of each single attribute, but also on the performance of the specific algorithm being applied.
- Example: SAS variable selection for multiple linear regression.

c) Embedded method:

- Attribute selection is inside the learning algorithm
- The selection is made as part of model generation
- Example: Decision tree methods for classification and regression.

Data reduction

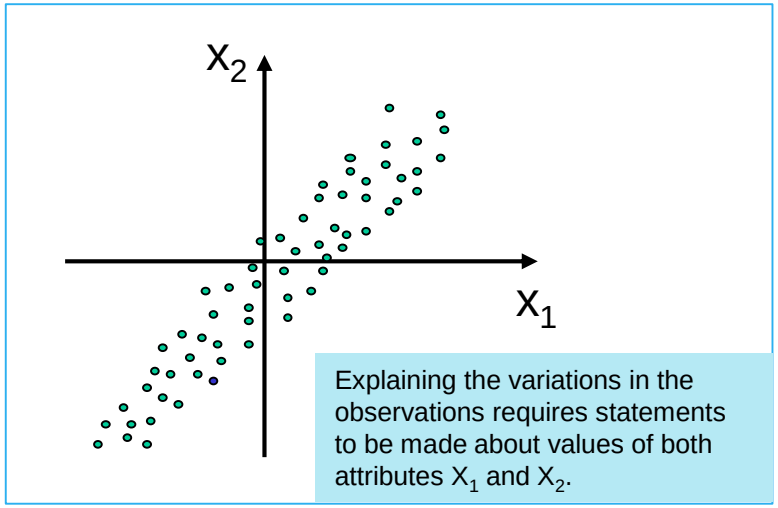
Which feature selection method to use?

- Filter methods: when dataset is large with large number of attributes.
- Wrapper and embedded methods: when dataset is not too large.
  - If there are n attributes in the data, there are 2^n possible subsets to search.
  - To reduce the search time, one of the following approaches can be followed:
    - **Forward or bottom-up**: start with no attribute, add one attribute at a time based on the ranking according to some relevance indicator (e.g. accuracy), stop when no more attributes meet the minimum threshold for inclusion.
    - **Backward or top-down**: start with all attributes, and eliminate one at a time, stop when no more attributes can be eliminated according to a pre-fixed threshold for removal.
    - **Forward-backward**: at each step the best attribute among those excluded is introduced and the worst attribute among those included is eliminated.

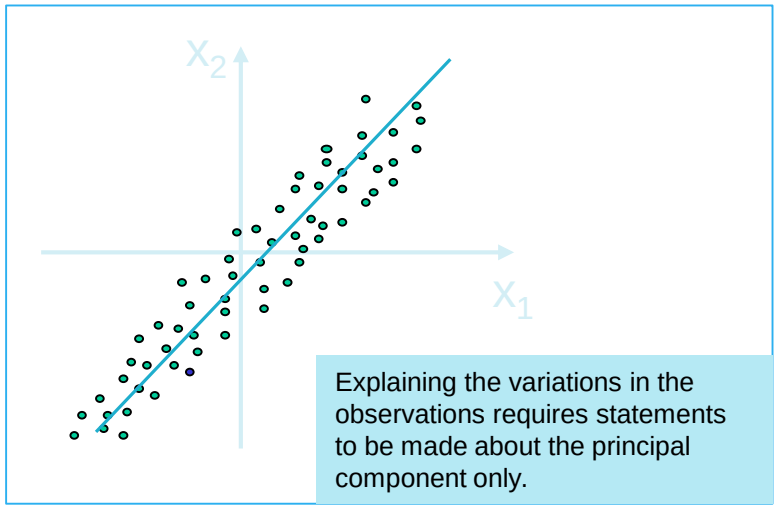
Data reduction

3. Principal Component Analysis

- Obtain a projective transformation that replaces a subset of the original numerical attributes with a lower number of new attributes obtained as their linear combination without causing loss of information.
- Principal components are generated to explain the variation in the data.
- Suppose the original data set is n-dimensional.
- A subset consisting of q principal components with q < n is expected to have an equivalent information content as the original n-dimensional data.



Explaining the variations in the observations requires statements to be made about values of both attributes X_1 and X_2.



Explaining the variations in the observations requires statements to be made about the principal component only.

Data reduction

Principal Component Analysis (PCA).

Example: Crime in the United States

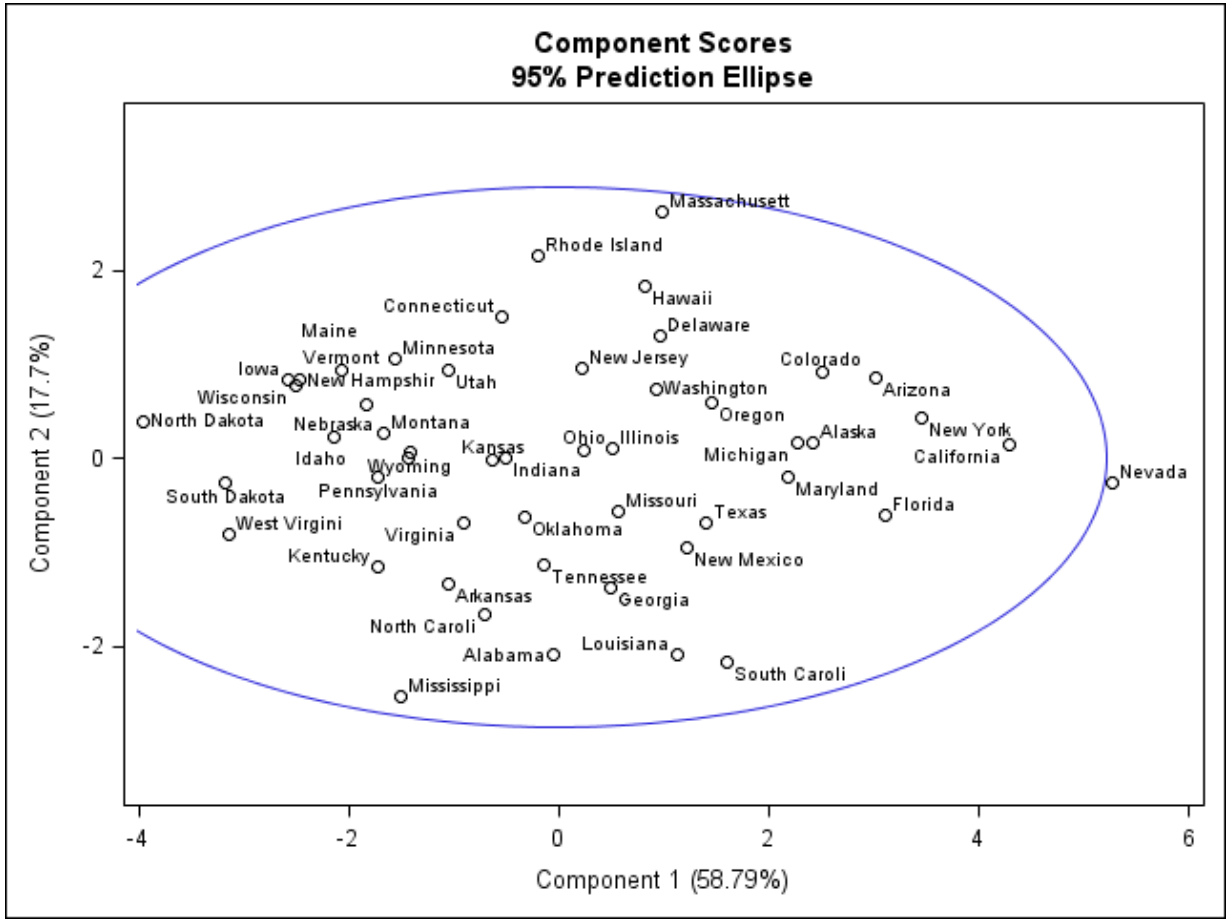
```
data Crime;
    input State $1-20 Murder Rape Robbery Assault
        Burglary Larceny Auto_Theft;
    datalines;
 Alabama       14.2  25.2  96.8  278.3 1135.5 1881.9 280.7
 Alaska        10.8  51.6  96.8  284.0 1331.7 3369.8  753.3
 Arizona        9.5 34.2 138.2 312.3  2346.1 4467.4 439.5
 Arkansas       8.8 27.6   83.2 203.4   972.6 1862.1 183.4
    .....................
 Washington     4.3  39.6 106.2 224.8 1605.6 3386.9 360.3
 West Virginia  6.0  13.2  42.2  90.9    597.4 1341.7 163.3
 Wisconsin      2.8  12.9   52.2  63.7    846.9 2614.2 220.7
 Wyoming        5.4  21.9  39.7 173.9    811.6 2772.2 282.0
 ;

 ods graphics on;
 title 'Crime Rates per 100,000 Population by State';
 proc princomp out=Crime_Components plots= score(ellipse ncomp=3);
   id State;
 run;
 ods graphics off;
```

ods = output delivery system

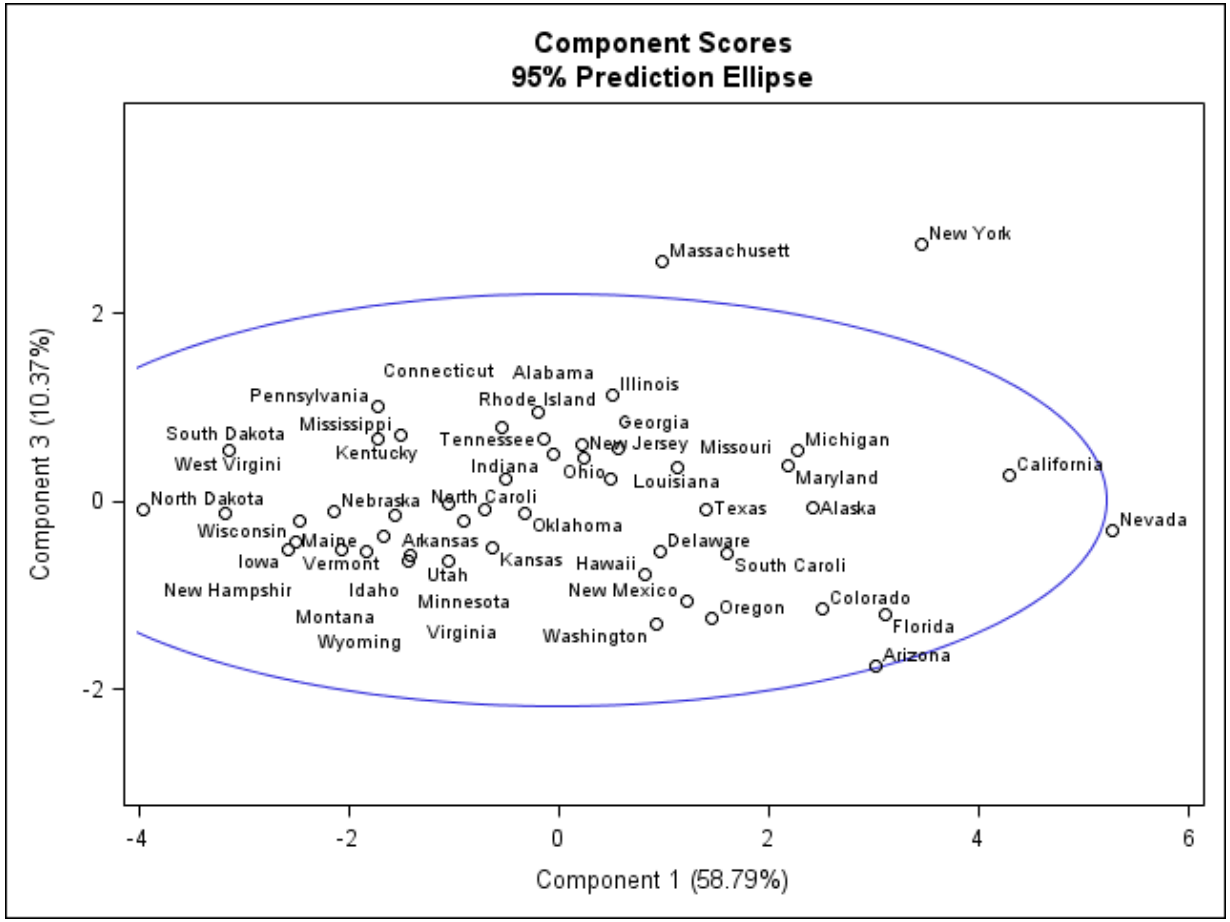
(This example is taken from
http://support.sas.com/documentation/cdl/en/statug/66859/HTML/default/viewer.htm#statug_princomp_gettingstarted.htm **)**

Data reduction



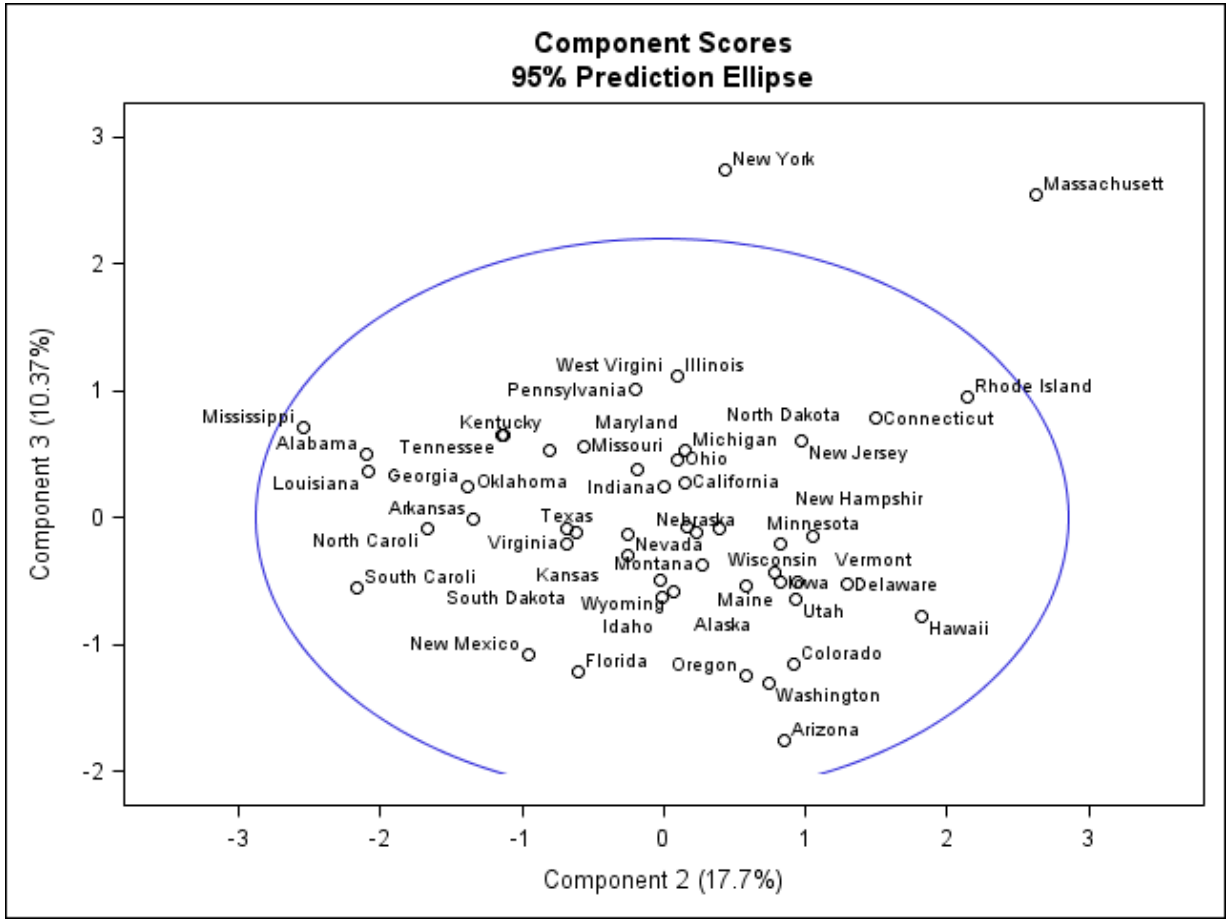
- The first principal component explains 58.79% of the variations in the data.
- The second explains an additional 17.7%.

Data reduction

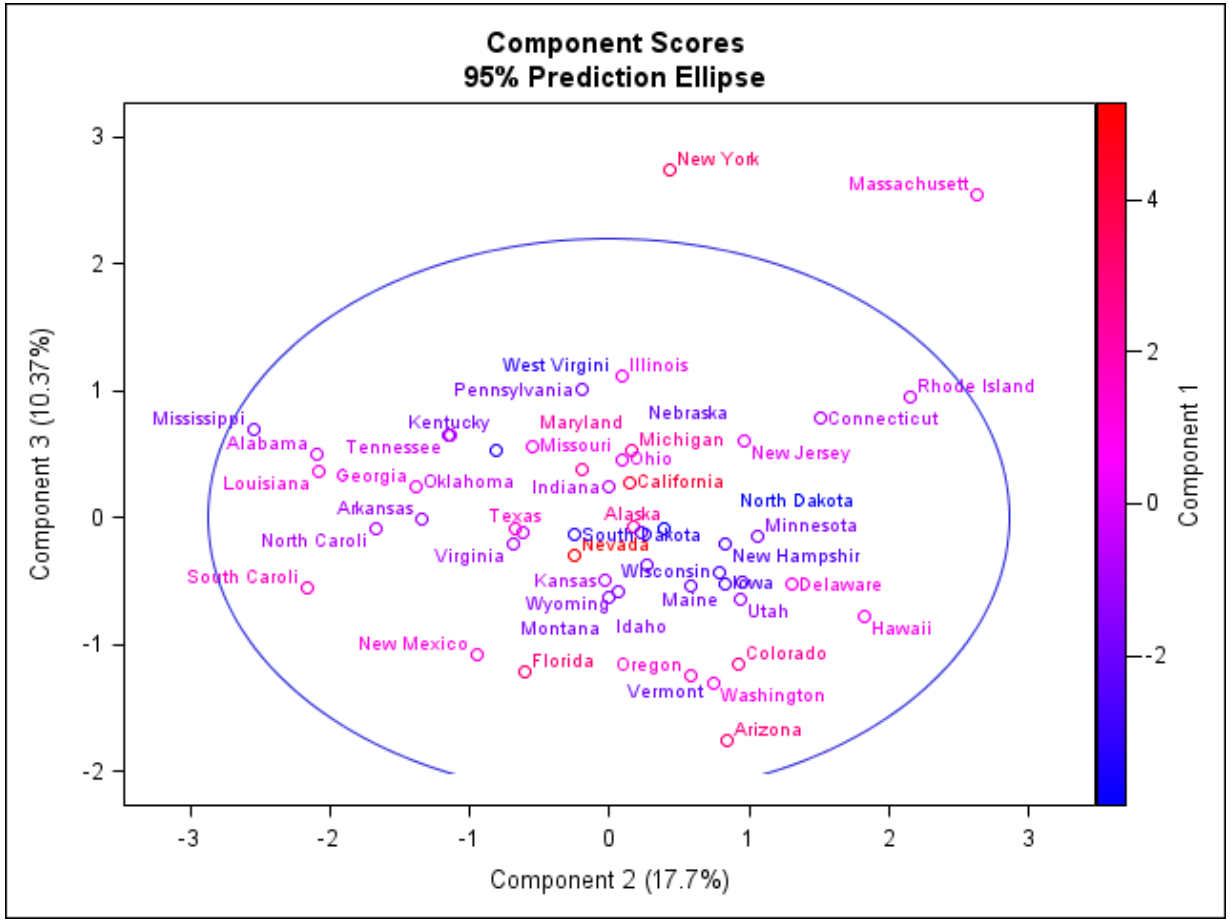


Assuming bivariate normal distribution of components 1 and 3, the plot identifies Massachusetts and New York as outliers.

Data reduction



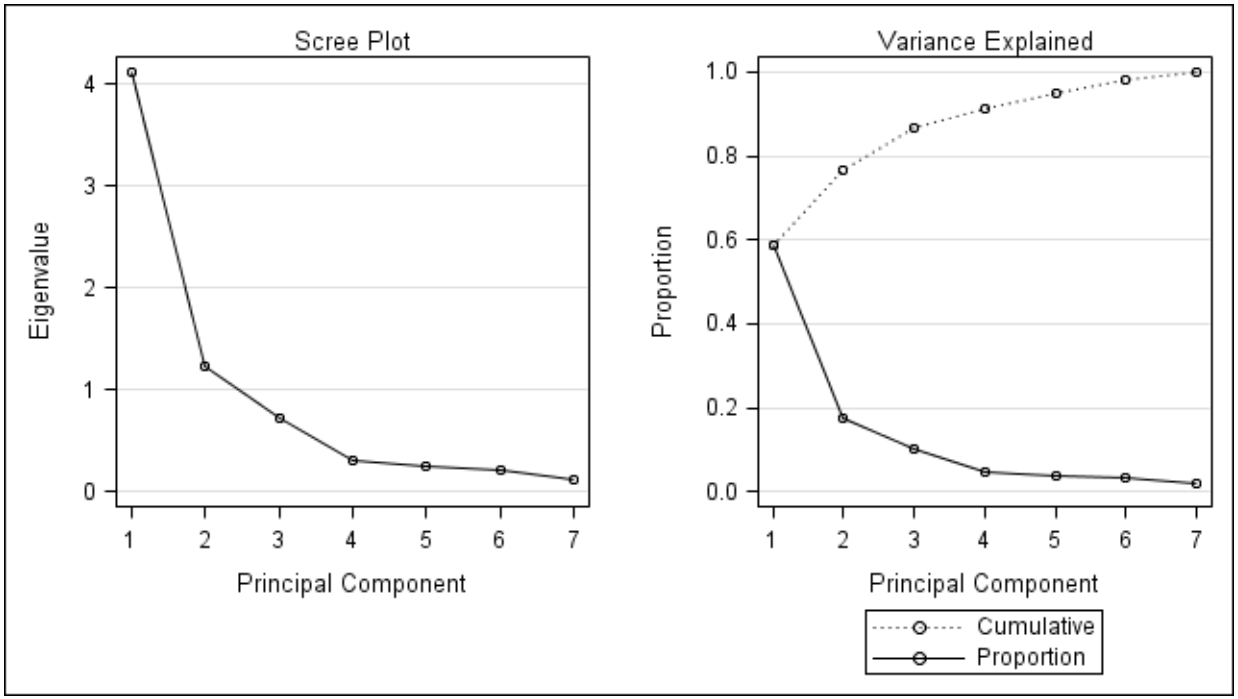
Data reduction



Data reduction

- Why are Nevada, New York, Massachusetts considered outliers?
- Principal components are linear combinations of the original attributes:
  - Prin1 = 0.300279×Murder + 0.431759×Rape + 0.396875×Robbery + 0.396652×Assault + 0.440157×Burglary + 0.357360×Larceny + 0.295177×Auto_Theft
  - Prin3 = 0.178245×Murder - 0.244198×Rape + 0.495861×Robbery - 0.069510×Assault - 0.209895×Burglary - 0.539231×Larceny + 0.568384×Auto_Theft
- Instead of checking the values of their pricipal components and try to relate them to the original attribute values, other data mining/classification tools could be used.

Data reduction



The scree plot can be used to identify the relevant components to be included in the transformed dataset.

Data reduction

4. Data discretization

- Data discretization reduces continuous attributes to categorical attributes with small number of distinct values.
- It also aims to reduce the number of distinct values assumed by the categorical variables.
- For example: Weekly spending of a mobile customer is discretized into five classes:
  - Low: [0,10) euros
  - Medium low: [10,20) euros
  - Medium: [20,30) euros
  - Medium high: [30,40) euros
  - High: 40 euros or more.
- A second example: Instead of checking a student's major, only her faculty/school is recorded.

Data reduction

Popular methods for data discretization:

- Subjective subdivision: classes are defined based on experience and judgments of experts in the application domain.
- Subdivision into classes based on equal size or equal width.

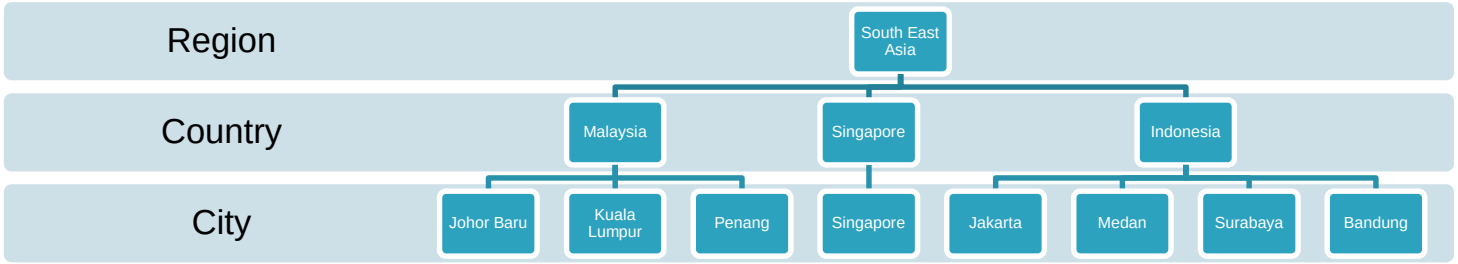
  Example: Original values of $\mathbf{a}_j$ are: 3,4,4,7,12,15,21,23,27

  - Equal size (of 3):
    - Class 1: 3,4,4
    - Class 2: 7,12,15
    - Class 3: 21,23,27
  - Equal width (of 8):
    - Class 1 – interval [3,11): 3,4,4,7
    - Class 2 – interval [11,19): 12,15
    - Class 3 – interval [19,27]: 21,23,27

Data reduction

Hierarchical discretization:

- Based on hierarchical relationships between concepts.
- Given a hierarchical relationship of the one-to-many kind, it is possible to replace each value of an attribute with the corresponding value found at a higher level in the hierarchy of concepts.



Interested in other methods for discretization? An old paper and its implementation and R package.

Data encoding for discrete attributes

Dummy variables for nominal discrete attribute:

- Example the attribute City has the following N = 8 possible values:

Johor Baru, Kuala Lumpur, Penang, Singapore, Jakarta, Medan, Surabaya and Bandung

- Use N-1 = 8-1 = 7 dummy binary variables to encode:

City	I_1	I_2	I_3	I_4	I_5	I_6	I_7
Johor Baru	0	0	0	0	0	0	0
Kuala Lumpur	0	0	0	0	0	0	1
Penang	0	0	0	0	0	1	0
Singapore	0	0	0	0	1	0	0
Jakarta	0	0	0	1	0	0	0
Medan	0	0	1	0	0	0	0
Surabaya	0	1	0	0	0	0	0
Bandung	1	0	0	0	0	0	0

Data encoding for discrete attributes

Dummy variables for nominal discrete attribute:

- It does not matter which city has a string of all zero.
- The city with all zero input is used as a base for comparison using the model.
- The other city has exactly one input with value equals to 1 – it does not matter which one.

City	I_1	I_2	I_3	I_4	I_5	I_6	I_7
Johor Baru	0	0	0	0	0	0	0
Kuala Lumpur	0	0	0	0	0	0	1
Penang	0	0	0	0	0	1	0
Singapore	0	0	0	0	1	0	0
Jakarta	0	0	0	1	0	0	0
Medan	0	0	1	0	0	0	0
Surabaya	0	1	0	0	0	0	0
Bandung	1	0	0	0	0	0	0

Data encoding for discrete attributes

Thermometer encoding for ordinal discrete attribute:

- Example. Weekly spending of a mobile customer is discretized into N= 5 sub-intervals:
  - Low: [0,10) euros
  - Medium low: [10,20) euros
  - Medium: [20,30) euros
  - Medium high: [30,40) euros
  - High: 40 euros or more.
- Use N-1 = 5-1 = 4 dummy binary variables and apply binary encoding:

Weekly spending	I_1	I_2	I_3	I_4
Low: [0,10) €	0	0	0	0
Medium low: [10,20) €	0	0	0	1
Medium: [20,30) €	0	0	1	1
Medium high: [30,40) €	0	1	1	1
High: 40 € or more	1	1	1	1

Data encoding for discrete attributes

Thermometer encoding for ordinal discrete attribute:

- Interpretation:
  - Weekly spending is high if and only if $I_1 = 1$
  - Weekly spending is medium high or lower if $I_1 = 0$
  - ✓ Weekly spending is at least 30 € if and only if $I_2 = 1$
  - ✓ Weekly spending is less than 30 € if and only if $I_2 = 0$
- …… etc.

Weekly spending	I_1	I_2	I_3	I_4
Low: [0,10) €	0	0	0	0
Medium low: [10,20) €	0	0	0	1
Medium: [20,30) €	0	0	1	1
Medium high: [30,40) €	0	1	1	1
High: 40 € or more	1	1	1	1

Data imbalance

- Imbalanced class distribution: When the number of observations belonging to one class is significantly different from those belonging to other classes.
- To build a balanced training data set, one may pick one of these approaches:
  - ✓ **Undersampling**: reduce the number of observations from majority class
  - ✓ **Oversampling**: increase the number of observations from minority class.

  SMOTE (Synthetic Minority Oversampling Technique):

  1. For each minority class sample $\mathbf{x}_i$, find its k-nearest minority class neighbors, N_1, N_2, … N_k
  2. Use a subset of the k-nearest neighbors to generate synthetic samples:

$$F_{new} = \mathbf{x}_i + rand(0,1)\,[\mathbf{x}_i - N_i]$$

  - ✓ **Cost sensitive learning:**

Classified as –>	Cost matrix 1	
	R	NR
Restatement (R)	0	1
Non-restatement (NR)	5	0

Data Exploration

Three main phases of exploratory data analysis

- **Univariate analysis**: properties of each single attribute of a data set are investigated
- **Bivariate analysis**: pairs of attributes are considered to measure the intensity of the relationship existing between them.

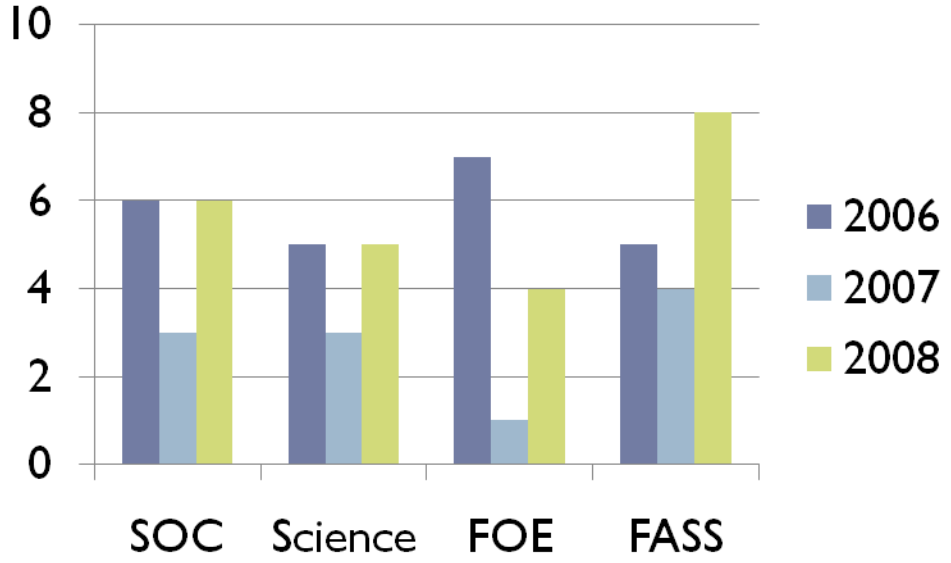
  For supervised learning models: between each explanatory variable and the target variable.
- **Multivariate analysis**: the relationships within a subset of attributes are investigated.

Graphical analysis of categorical attributes

Example of a vertical bar chart: the number of NUS students who spent one year as exchange students in Timbuktu.

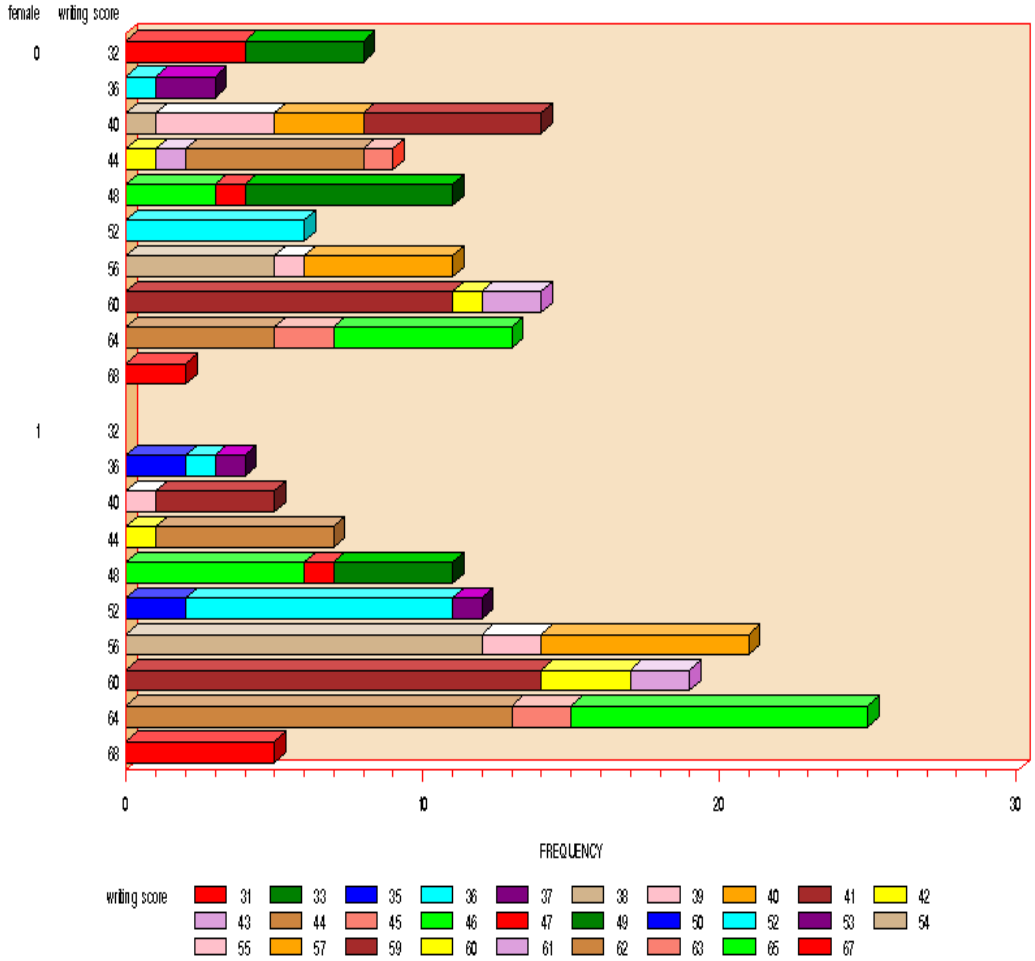
Categorical variables:

- Faculty: SOC, Science, FOE, FASS
- Acad. Year: 2006, 2007, 2008



Graphical analysis of categorical attributes

Example of a horizontal bar chart:



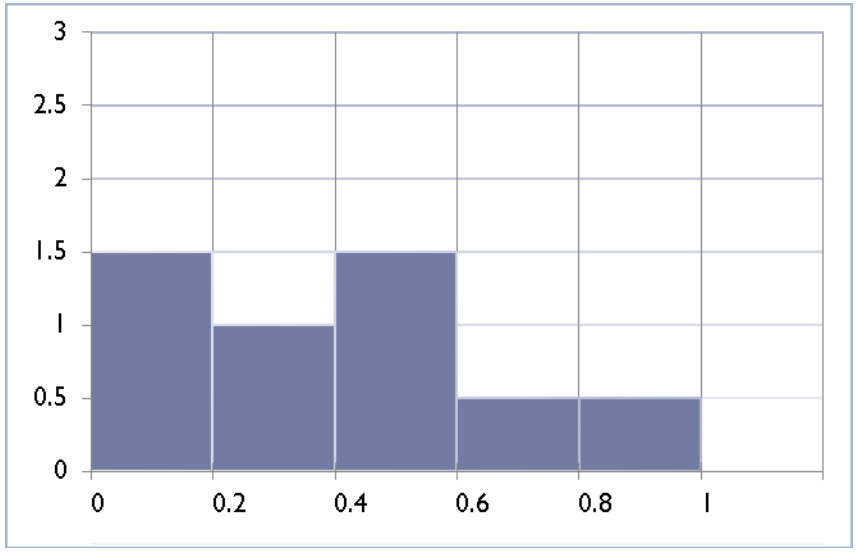
- Attributes:
  - Gender (male or female)
  - Writing scores (31 to 67)
- Drawn using SAS Analyst
- SAS file automatically created
- length of bar indicates frequency

Graphical analysis of numerical attributes

• if the numerical attribute has discrete and limited number of values, it is possible to display it using a bar chart representation.

• discretize if the variable is continuous or discrete with infinite number of possible values: subdivide the horizontal axis into a finite, moderate number of intervals.

• An empirical density histogram plotted using Excel:



Measures of central tendency for numerical attributes

- Mean: $\bar{\mu} = \sum_{i=1}^{m} x_i / m = (x_1 + x_2 + \ldots + x_m)/m$

- The sum of deviations/spread is zero

$$\sum_{i=1}^{m} (x_i - \bar{\mu}) = 0$$

- The value of c that minimizes the sum of squared deviations:

$$\min \sum_{i=1}^{m} (x_i - c)^2$$

is $c = \bar{\mu}$

- Weighted sample mean = $\sum_{i=1}^{m} w_i x_i / \sum_{i=1}^{m} w_i$

Measures of central tendency for numerical attributes

Median:

- Suppose $x_1, x_2, \ldots x_m$ are m observations arranged in a non-decreasing way
- If m is an odd member, the median is the observation occupying position (m+1)/2

$$x_{med} = x_{(m+1)/2}$$

- If m is an even number, the median is the middle point in the interval between the observations of position m/2 and (m+2)/2

$$x_{med} = (x_{m/2} + x_{(m+2)/2})/2$$

- Example 1: -2, 0, 5, 7, 10

The median is 5.

- Example 2: 50, 85, 100, 102, 110, 200

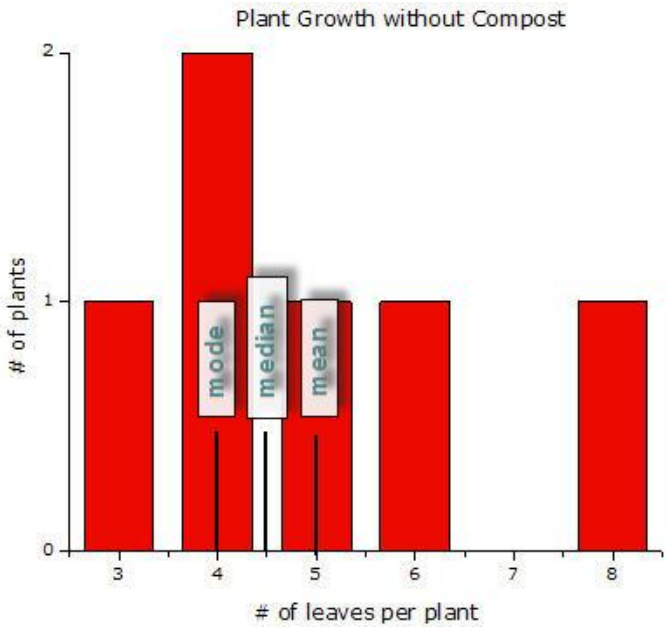
The median is (100 + 102)/2 = 101

Measures of central tendency for numerical attributes

• Mode: the value that corresponds to the peak of the empirical density curve.

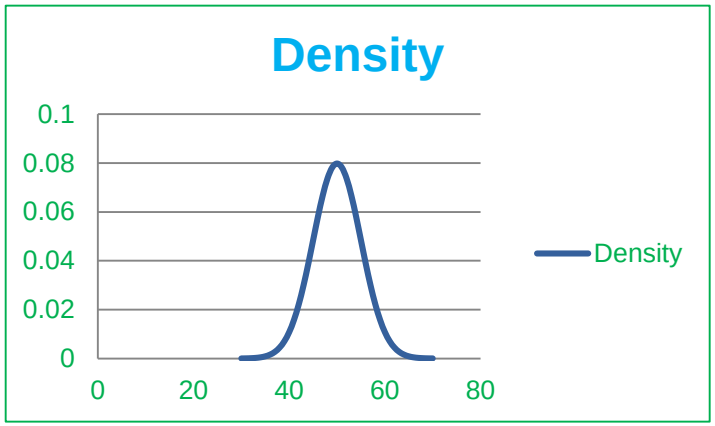
○ If the empirical density curve has been calculated by partition into intervals, each value of the interval that corresponds to the maximum empirical frequency is the mode.

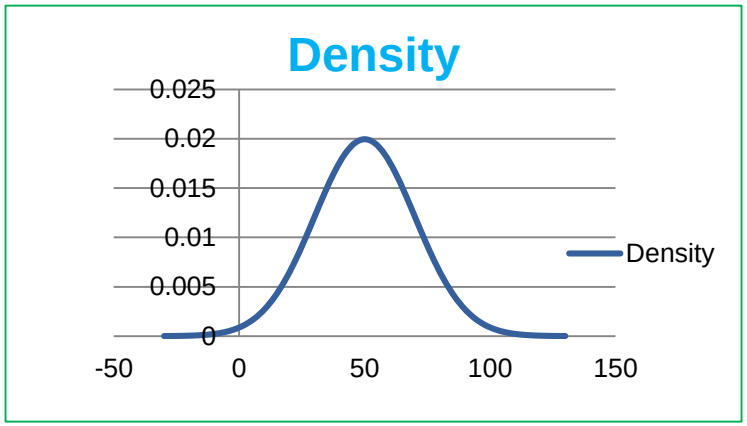
○ Graph taken from ScienceBuddies:



Measures of dispersion for numerical attributes

- Normal distribution with μ = 50 and σ =5 (left) and μ = 50 and σ = 20 (right)





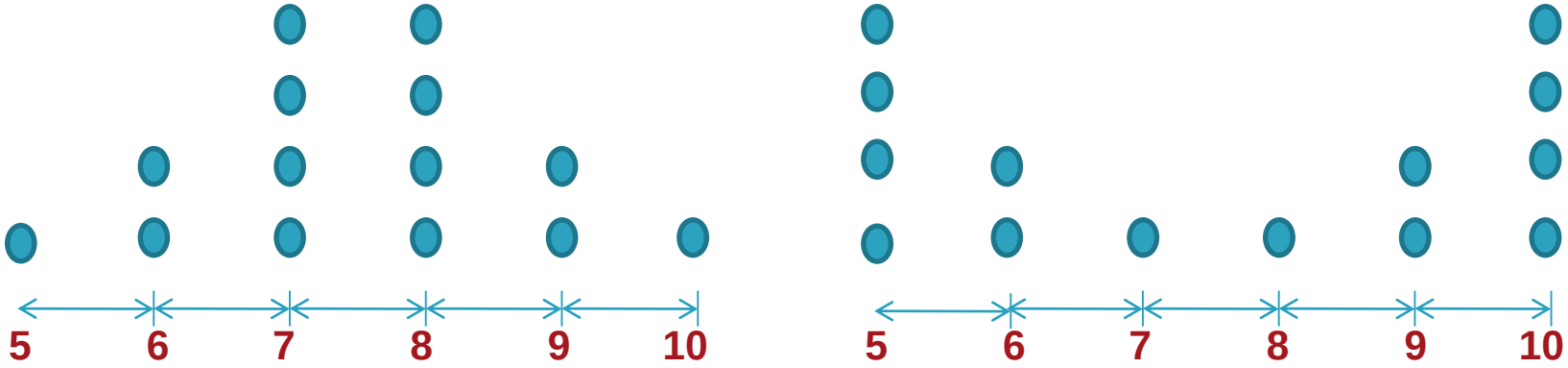
- In most applications, it is desirable to have data with small dispersion (left).
- In some applications, a higher dispersion may be desired, for example, for the purpose of classification/discriminating between classes.

Measures of dispersion for numerical attributes

- Range is the simplest measure of dispersion

$$x^{range} = x^{max} - x^{min}$$

  - It cannot catch the actual dispersion of the data, only useful in identifying the interval in which the values of a data attribute fall



  - Mean absolute deviation (MAD):

$$MAD = \sum_{i=1}^{m} |x_i - \bar{\mu}| / m$$

Measures of dispersion for numerical attributes

• Sample variance:

$$\bar{\sigma}^2 = \left(\sum_{i=1}^{m} s_i^2\right)/(m-1) = \sum_{i=1}^{m} (x_i - \bar{\mu})^2/(m-1)$$

○ A lower sample variance implies a lower dispersion of the values around the sample mean.

○ As the size of the sample increases, the sample mean $\bar{\mu}$ approximates the population mean μ, and the sample variance approximates the population variance σ^2 of the distribution from which the values of the attribute are drawn.

○ To have the measure of dispersion back to the original scale in which the observations are expressed, the sample standard of deviation is defined as:

$$\bar{\sigma} = \text{sqrt}(\bar{\sigma}^2)$$

Measures of dispersion for numerical attributes

- Normal distribution:
  - the interval ($\bar{\mu} \pm \bar{\sigma}$) contains approximately 68% of the observed values.
  - the interval ($\bar{\mu} \pm 2\bar{\sigma}$) contains approximately 95% of the observed values.
  - the interval ($\bar{\mu} \pm 3\bar{\sigma}$) contains approximately 100% of the observed values.

- **Coefficient of variation (CV)** is defined as the ratio between the sample standard deviation and the sample mean

$$CV = 100 \times \bar{\sigma} / \bar{\mu}$$

CV is often used to compare two or more groups of data.

Measures of relative location for numerical attributes

- Quantiles:
  - Suppose we arrange the m values $\{x_1, x_2, \ldots. x_m\}$ in a non-decreasing order.
  - Given any value p such that $0 \leq p \leq 1$, the p-order quantile is the value q_p such that pm observations will fall on the left of q_p and the remaining (1-p)m on its right.
  - Sometimes p-quantiles are called 100 p^{th} percentiles.
  - 0.5-order quantile coincides with the median.
  - q_L = 0.25-order quantile, also called lower quartile.
  - q_U = 0.75-order quantile, also called upper quartile.
  - Interquartile range is defined as the difference between the upper and lower quartiles

$$D_q = q_U - q_L = q_{0.75} - q_{0.25}$$

Identification of outliers for numerical attributes

- z-index expresses the signed distance between a value and the sample mean by using the sample standard deviation as a measurement unit:

$$z_i^{ind} = (x_i - \overline{\mu})/\overline{\sigma}$$

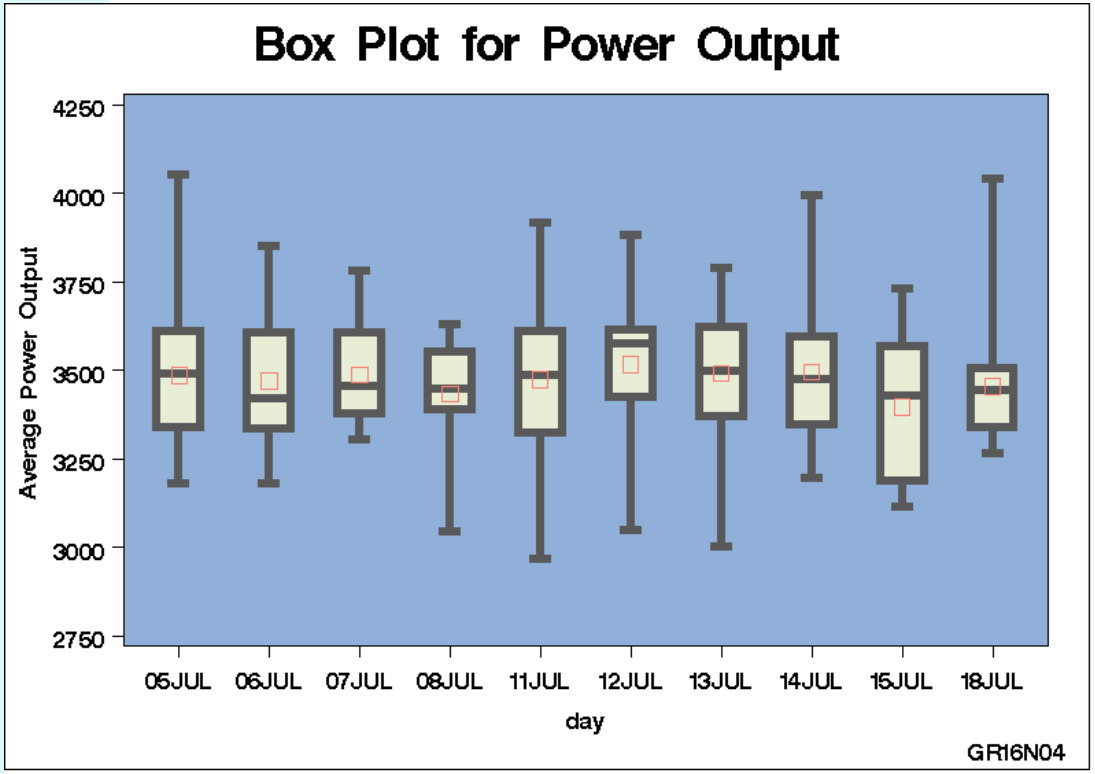
- The z-index can be used to identify outliers in most cases: data values which are outside 3-standard deviations from the sample mean are considered to be suspicious:

$|z_i^{ind}| > 3$: suspicious

$|z_i^{ind}| >> 3$: highly suspicious

Identification of outliers for numerical attributes

Boxplot is another way to identify outliers.



- the length of the box represents the interquartile range (the distance between the 25th and the 75th percentiles)
- the dot in the box interior represents the mean
- the horizontal line in the box interior represents the median
- the vertical lines issuing from the box extend to the minimum and maximum values of the analysis variable. At the end of these lines are the whiskers.

Measures of heterogeneity for categorical attributes

• Measures such as central tendency, dispersion and relative location cannot be used for categorical attributes.

• It is preferable to define some measures that express the regularity of the arrangement of the data $\{x_1, x_2, x_3 \ldots, x_m\}$ within the set of H distinct values taken by the attribute.

• For example: {medium, large, large, medium, small, extra large, small} ; m = 7, H = 4.

• The highest heterogeneity is obtained when the relative empirical frequencies are equal for all classes.

• For example: {approved, approved, not approved, not approved}; 50% of the applications are approved, the other 50% are not approved; m = 4, H = 2.

• In contrast, the lowest heterogeneity occurs when the relative empirical frequency is 1 for one of the classes.

• For example, all (= 100%) of the applications are approved.

• Two measures of heterogeneity: Gini index and entropy index.

Measures of heterogeneity for categorical attributes

Gini index:

$$G = 1 - \sum_{h=1}^{H} f_h^2$$

- When one of the H classes has relative frequency $f_h = 1$, then G has the lowest value of 0.
- When all the classes have the same relative frequency, then G has the highest value of (H-1)/H.
- It is possible to normalize Gini index so that the values lie in [0,1]

$$G_{rel} = G/[\ (H-1)/H)\ ]$$

Measures of heterogeneity for categorical attributes

Entropy index:

$$E = -\sum_{h=1}^{H} f_h \log_2 f_h$$

- When one of the H classes has relative frequency $f_h = 1$, then E has the lowest value of 0.
- When all the classes have the same relative frequency, then E has the highest value of $\log_2 H$.
- It is possible to normalize entropy index so that the values lie in [0,1]

$$E_{rel} = E / \log_2 H$$

Bivariate analysis

Measure of correlation for numerical attributes.

- Summary indicators are useful to express the nature and intensity of the relationship between numerical attributes.
- The **sample covariance** for the pair of attributes $\mathbf{a_j}$ and $\mathbf{a_k}$:

$$\nu_{jk} = cov(\mathbf{a_j},\mathbf{a_k}) = (1/m\text{-}2) \sum_{i=1}^{m} (x_{ij} - \overline{\mu}_j)(x_{ik} - \overline{\mu}_k)$$

- Concordance: if values of attribute $\mathbf{a_j}$ lower/*higher* than the mean $\overline{\mu}_j$ are associated with values of attribute $\mathbf{a_k}$ lower/*higher* than the mean $\overline{\mu}_k$.
- Discordance: if values of attribute $\mathbf{a_j}$ lower/*higher* than the mean $\overline{\mu}_j$ are associated with values of attribute $\mathbf{a_k}$ higher/*lower* than the mean $\overline{\mu}_k$.
- Positive covariance: the attributes $\mathbf{a_j}$ and $\mathbf{a_k}$ are concordant.
- Negative covariance: the attributes $\mathbf{a_j}$ and $\mathbf{a_k}$ are discordant.

Bivariate analysis

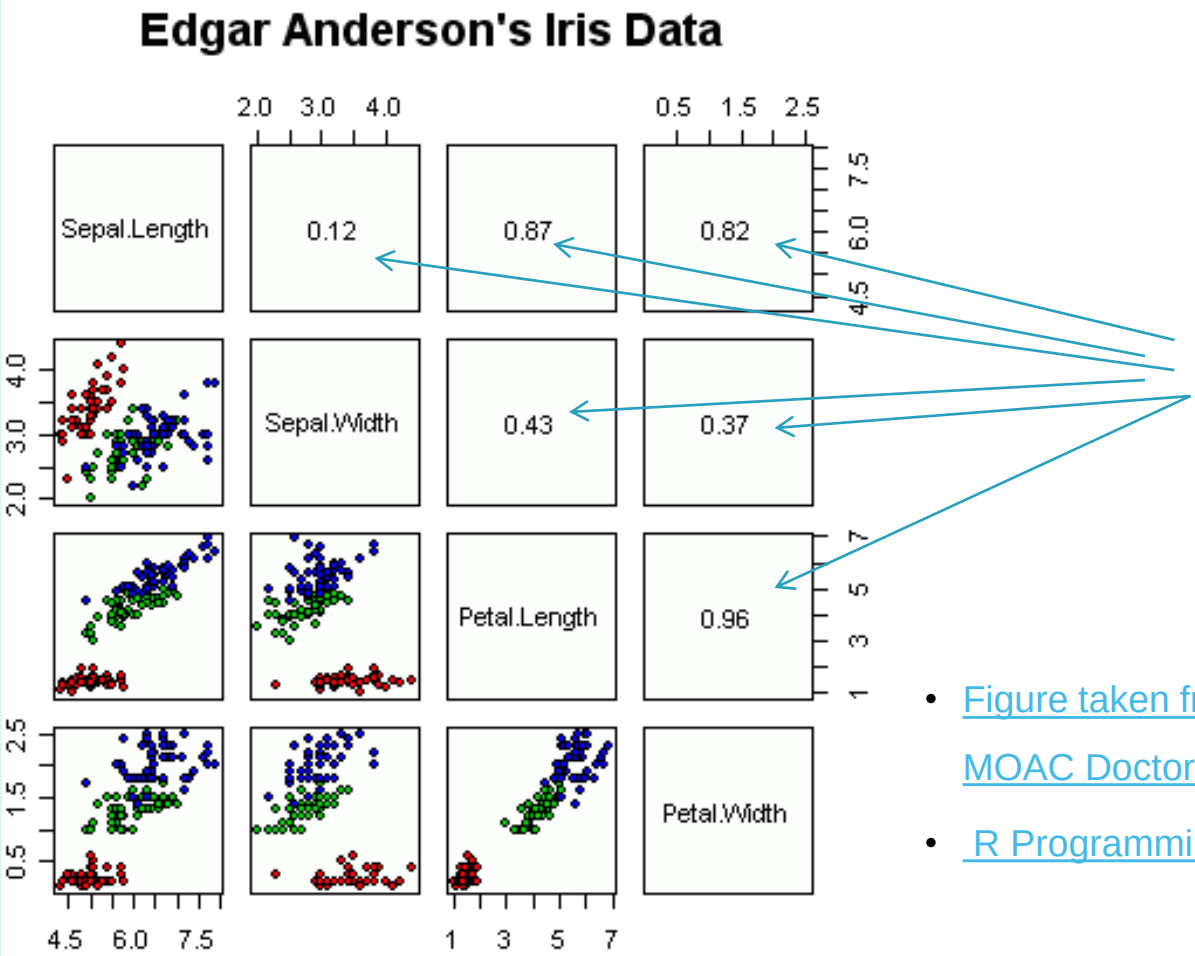
Measure of correlation for numerical attributes.

- **Linear correlation coefficient (or Pearson coefficient) is defined as**

$$r_{jk} = corr(\mathbf{a_j}, \mathbf{a_k}) = \nu_{jk} / (\bar{\sigma}_j \bar{\sigma}_k)$$

- If $r_{jk} > 0$, the attributes are concordant.
- If $r_{jk} < 0$, the attributes are discordant.
- If $r_{jk} = 0$ or $r_{jk} \approx 0$, no linear relationship exists between the two attributes.
- r_{jk} always lies in the interval $[-1, 1]$

Bivariate analysis

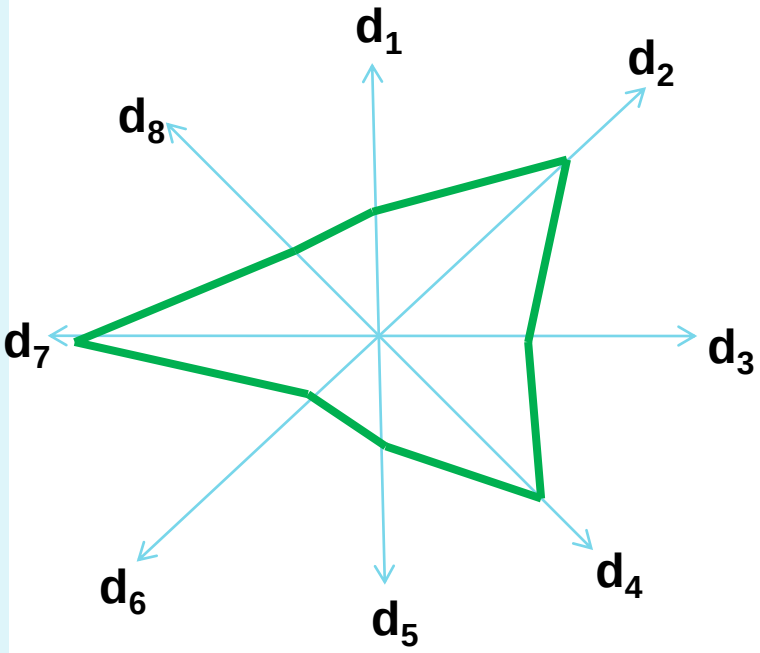


Correlation coefficients

- Figure taken from Warwick University MOAC Doctoral Training Centre.
- R Programming Language Website

Multivariate analysis

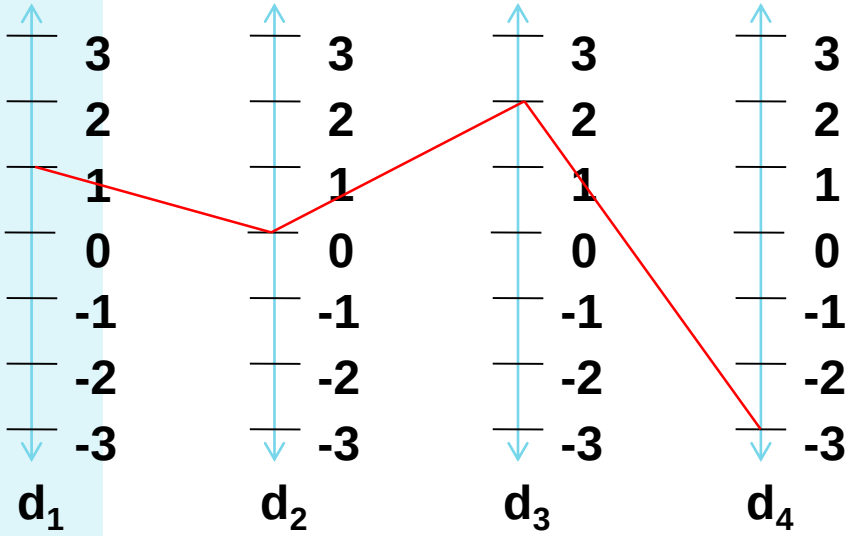
- **Star plot** is a designed to show the differences among values of the attributes for the records in the data set.
- It is effective when the number of observations is not large.
- Below is an example of a data observation describes by 8 attributes (dimensions).

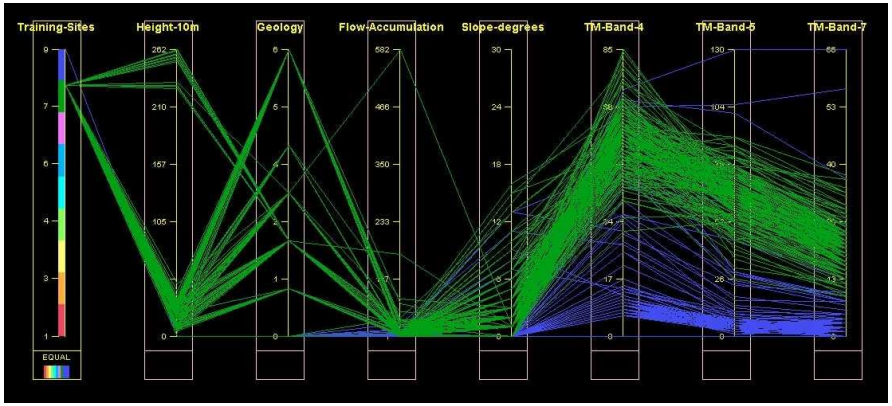


- **Space out all dimensions with equal angles centered at the origin**
- **Each spoke encodes a variable's value**

Multivariate analysis

Parallel coordinates

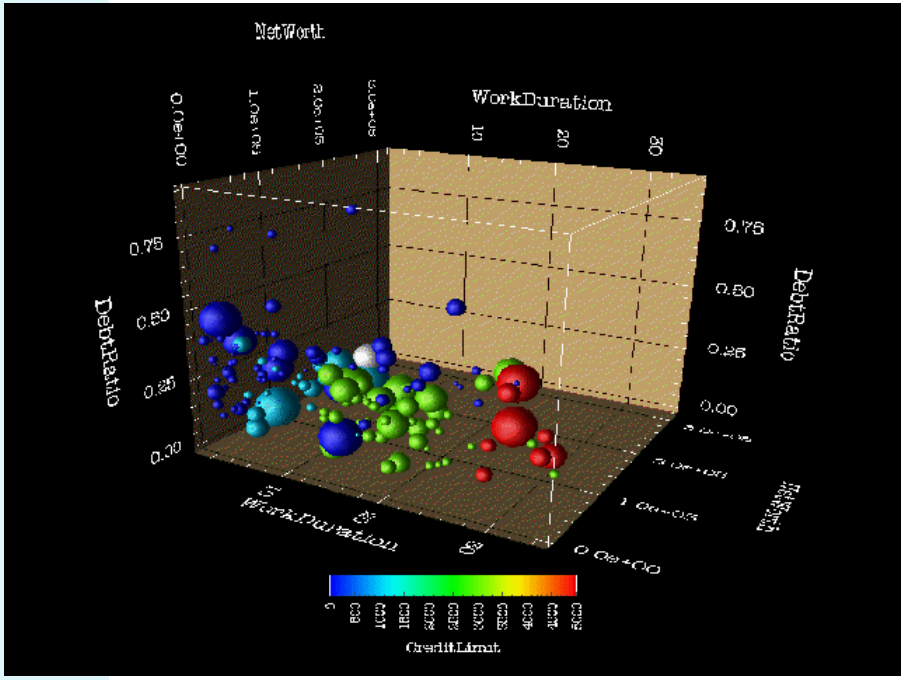




- d parallel lines, equally space, one for each dimension
- the position of the vertex on the i-th axis corresponds to the i-th coordinate of the point variable's value
- Colored figure taken from "GeoVISTA *Studio*: a geocomputational workbench" by Mark Gahegan, Masahiro Takatsuka, Mike Wheeler and Frank Hardisty, GeoVista Center, Penn State University.

Multivariate analysis: other visualization tools

Result obtained from Credit Card Application by IBM Open DX.



- Colored spheres show approved card holders.
- In the original application, historical information on each card holder could be obtained by clicking on each of the spheres.
- The three axes show applicant information: Net Worth, WorkDuration, DebtRatio.
- The size of the sphere is the salary, and the color is the credit limit for that card holder.
- A white sphere is used to depict a new applicant.
- The applicant's credit limit can be determined by the 3D position of that sphere relative to the historical data base information

Multivariate analysis: SOM

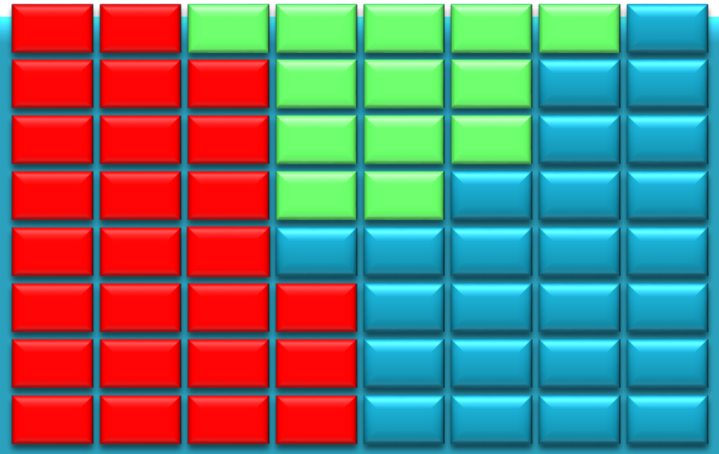
	Dove	Hen	Duck	Geese	Owl	Hawk	Eagle	Fox	Dog	Wolf	Cat	Tiger	Lion	Horse	Zebra	Cow
Small	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0
Medium	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
Big	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
2 legs	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
4 legs	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Hair	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Hoves	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
Mane	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0
Feathers	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
Hunt	0	0	0	0	1	1	1	1	0	1	1	1	1	0	0	0
Run	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0
Fly	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0
Swim	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0

Multivariate analysis: SOM

Data

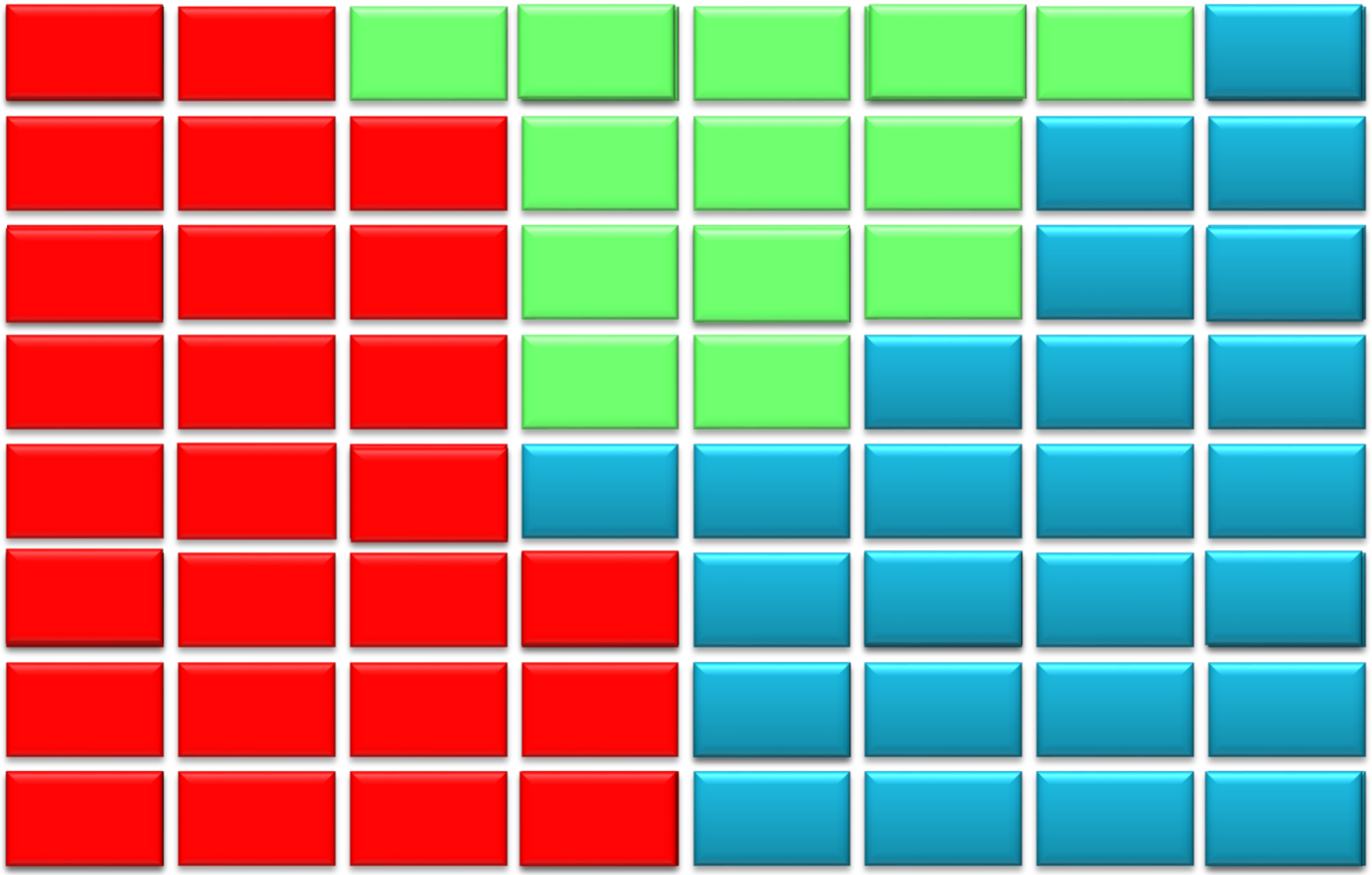
1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	1	1	1	0	0	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	0	1	1	1	1	0	0	0
0	0	0	0	0	0	0	0	1	1	0	1	1	1	0	0
1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0

Intelligent data analysis tool
(a self organizing map)



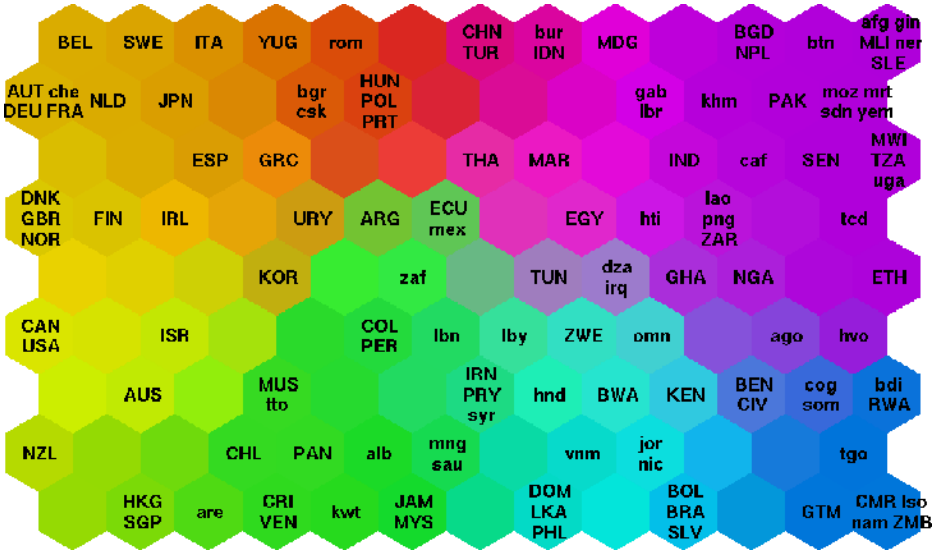
Multivariate analysis: other visualization tools

What kind of information can we get from the data?



Multivariate analysis: SOM

World poverty map as generated by neural network approach SOM (Self Organizing Maps):



• Figure taken from: The Laboratory of Computer and Information Science (CIS) of the Department of Computer Science and Engineering at the Helsinki University of Technology.

•The data consisted of World Bank statistics of countries in 1992. Altogether 39 indicators describing various quality-of-life factors, such as state of health, nutrition, educational services, etc, were used.

Data preparation & exploration: bankruptcy predition

The article “An investigation of bankruptcy prediction in imbalanced datasets” by D. Veganzones and E. Severin, Decision Support Systems 112 (2018), 111-124 discusses the following issues:

- data imbalance
- sensitivity of machine learning method to data imbalance
- feature selection

Data:

Samples	Training set proportion						Test set proportion
	50/50	60/40	70/30	80/20	90/10	95/5	95/5
Bankrupt	750	600	450	300	150	75	75
Non-bankrupt	750	900	1050	1200	1350	1425	1425
Total	1500	1500	1500	1500	1500	1500	1500

Data preparation & exploration: bankruptcy predition

Features:

- 50 financial ratios
- Two-step variable selection process:

  1. compute correlation between each variable, select (filter) variables with correlation values lower than 0.65.

  2. Four wrapper methods used to reduce variables further: a variable must be selected by at least 2 of the 4 methods

Data preparation & exploration: bankruptcy predition

Selected features:

- Variables selected by sample.

Service	Construction	Retail	All
C/TA	(C + MS)/CL	(C + MS)/TS	(C + MS)/CL
CL/TA	CA/CL	CA/CL	C/CA
WC/TA	FE/VA	CL/TA	QA/TA
FE/TA	EBITDA/TA	FE/TA	EBIT/TA
EBITDA/TA	LTD/SF	EBITDA/TA	LTD/SF
SF/PE	SF/TA	LTD/SF	TD/TA
TD/TA	EBIT/VA	SF/PE	CF/VA
NI/TS	NI/TS	NI/TS	NI/TS
AC/TS	TS/TA	NOWC/TS	NOWC/TS
NOWC/TS			
See Table 2			

Eg.

C/TA: Cash/total assets

NOWC/TS: Net op. work. Capital/total sales

Data preparation & exploration: bankruptcy predition

Classification methods:

- Linear discriminant analysis, Logistic regression, Neural networks, Support vector machines

Sampling methods:

Over-sampling approach:

- Random over-sampling approach: duplicate minority class
- SMOTE: Synthetic minority oversampling technique

Under-sampling approach:

- Random under-sampling: remove samples from majority class to generate balanced subset
- Easy ensemble: extract N subsets from majority class with the same number of samples as the minority class. Train N classifiers to form an ensemble.

Data preparation & exploration: bankruptcy predition

Evaluation metric:

$$Sensitivity = \frac{TP}{TP + FN} \text{ is the percentage of bankrupt samples correctly classified.}$$

$$Specificity = \frac{TN}{TN + FP} \text{ is the percentage of non} - \text{bankrupt samples correctly classified.}$$

$$G - mean = \sqrt{\frac{TP}{TP + FN} * \frac{TN}{TN + FP}},$$

AUC: Area under the curve

Data preparation & exploration: bankruptcy predition

Determining the cut-off point to decide bankrupt vs non-bankrupt:

$$\text{Expected cost of misclassification} = c_1 \frac{e_1}{n_1} p_1 + c_2 \frac{e_2}{n_2} p_2$$

where c_1 and c_2 are the respective costs of misclassification for bankrupt and non-bankrupt firms; e_1 and e_2 are the type-I and type-II error respectively; n_1 represents the number of bankrupt firms, while n_2 is that of non-bankrupt firms; and p_1 and p_2 are the prior probabilities of bankrupt and non-bankrupt firms respectively.

- $c_1 = c_2 = 1$
- p_1 and p_2 are prior probabilities of bankrupt and non-bankrupt firms, respectively.

Data preparation & exploration: bankruptcy predition

Results:

- Lots, lots of graphs.
- Imbalanced distribution in which the minority class represents 20% (or less) of the samples in the dataset significantly disturbs prediction performance.
- Support vector machine method is less sensitive than other prediction method to imbalanced distributions.
- Sampling methods can recover a satisfactory portion of performance losses.
- Random over-sampling leads to better performance for LDA and LR.
- Under-sampling is sub-optimal in almost all scenarios.

Reference

Business Intelligence: Data Mining and Optimization for Decision Making by Carlo Vercellis, 2009, Wiley. Chapters 6 and 7.

Also available in RBR Section Central Library.