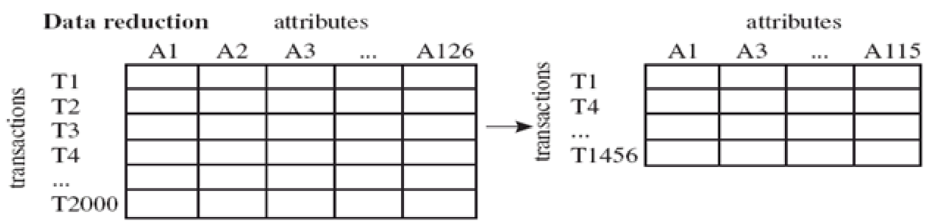
Understanding Data and Their Environment

Data Preprocessing

Dr Yu-wang Chen
Alliance Manchester Business School, The University of Manchester
Room: 3.74 AMBS, Tel: 0161 275 6345 (Ext: 56345)
Email: Yu-wang.Chen@manchester.ac.uk

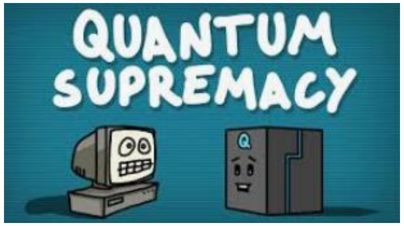
Outline

- Data integration
- Data description, summarisation and visualisation
- Data cleaning
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
- Data transformation



Data Reduction

- Why data reduction?
  - A database/data warehouse may store terabytes of data
  - Complex analysis may take a very long time to run on the complete data set (computational complexity)?

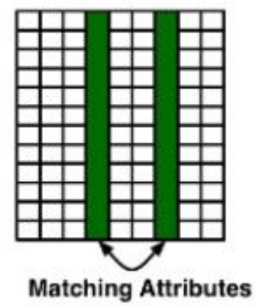


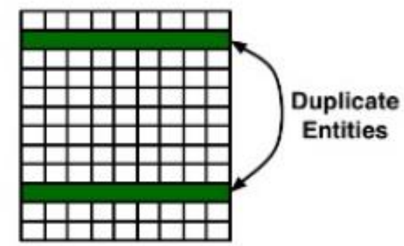
- Data reduction
  - Obtain a reduced representation of the data set - much smaller in volume but yet produces almost the same analytical results.

Data Reduction During Integration

- Redundant data occur often when integration of multiple databases

  - Column-oriented: the same attribute may have different names in different databases
  - Row-oriented: duplicate entities, etc.





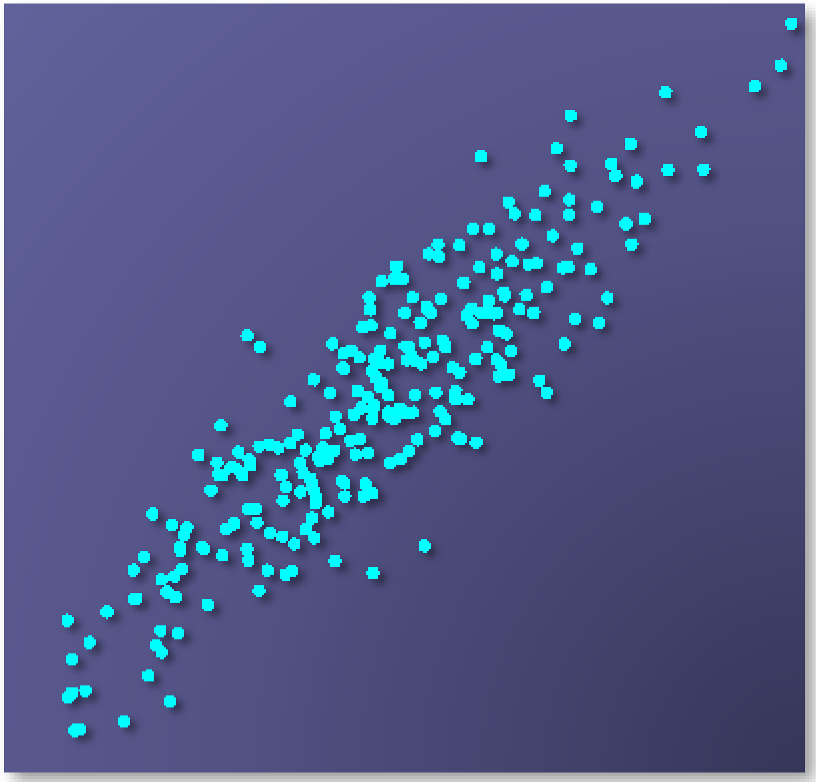
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve the efficiency and quality of data analytics.

Data Reduction Strategies

- Dimensionality reduction (variable reduction)
  - Remove data with poor quality (e.g., many missing values, large number of categories, zero-variance)
  - Remove redundant and irrelevant attributes or variables
  - Principal component analysis (PCA)
  - Variable clustering
  - Featuring engineering
- Data reduction (numerosity reduction)
  - Sampling techniques
  - Regression and log-linear models
  - Histogram analysis, clustering

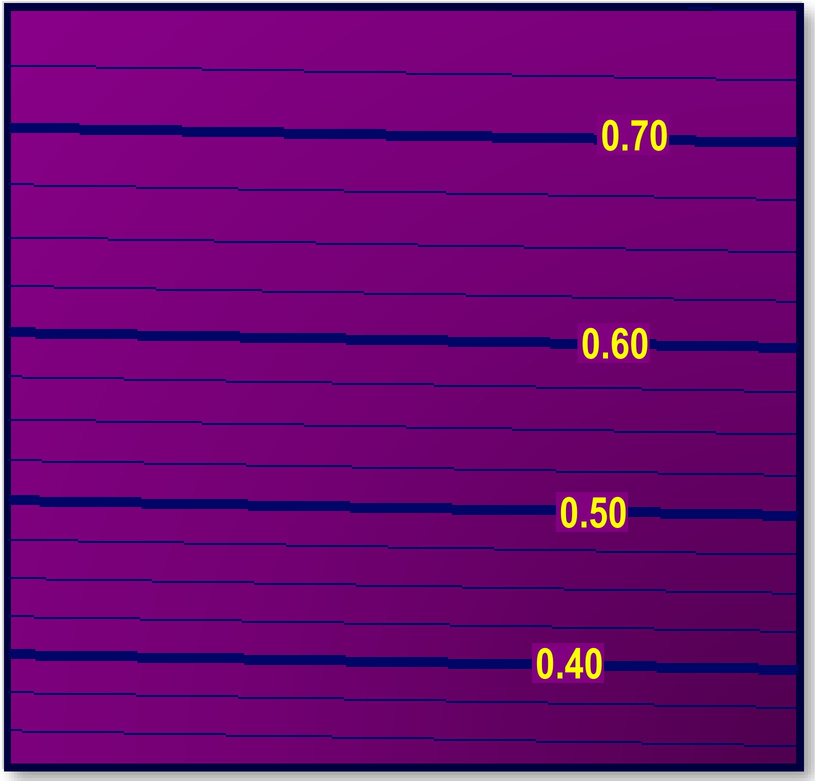
Variable Reduction – Correlation analysis

x_2



x_1

x_4



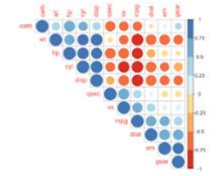
x_3

Redundancy:

Input x_2 has the identical information as input x_1.

Irrelevancy :

Outputs change with input x4 but much less with input x3.



Correlation Analysis – Numerical Variables

- **Correlation** between two variables x_1 and x_2 is the standard covariance, obtained by normalising the covariance with the standard deviation of each variable.

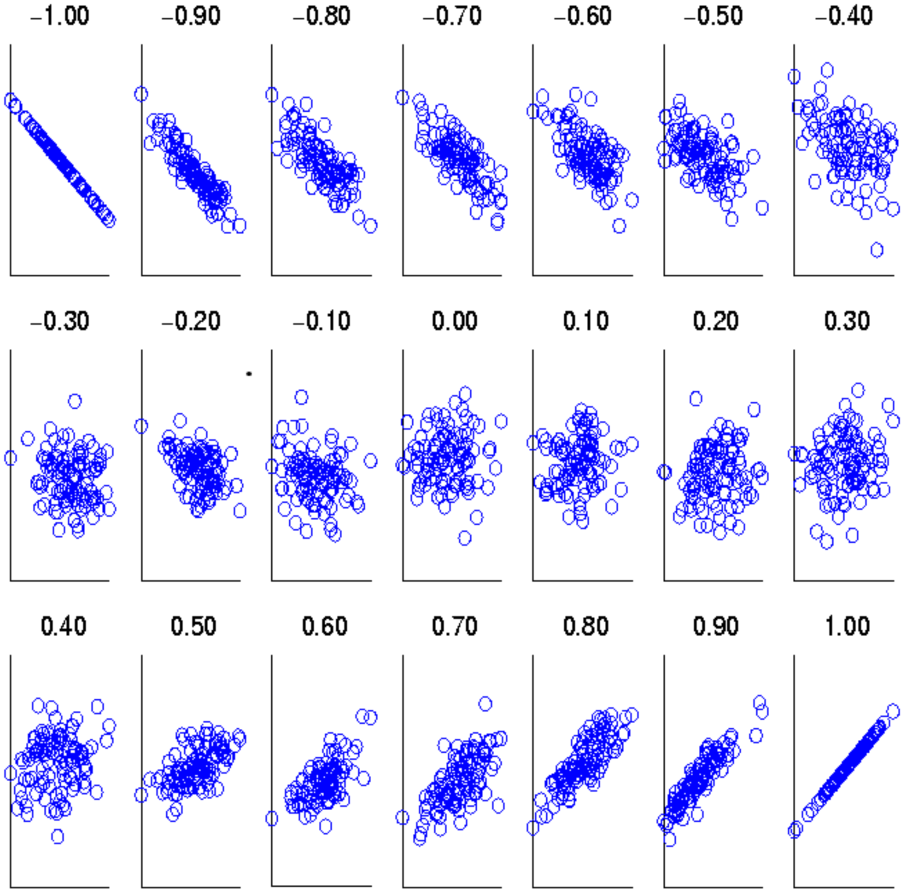
- **Sample correlation** for two attributes x_1 and x_2: where n is the number of tuples, μ_1 and μ_2 are the respective means, σ_1 and σ_2 are the respective standard deviation of x_1 and x_2

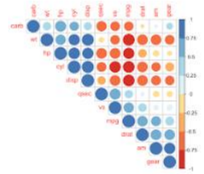
$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^{n} (x_{i2} - \hat{\mu}_2)^2}}$$

  - If $\rho_{12} > 0$: x_1 and x_2 are positively correlated (x_1's values increase as x_2's)
  - If $\rho_{12} = 0$: independent
  - If $\rho_{12} < 0$: negatively correlated

Visualising Correlation Coefficients

- Correlation coefficient value range: [–1, 1]
- A set of scatter plots shows sets of points and their correlation coefficients changing from –1 to 1





Correlation Analysis?

- Correlation (Pearson – conditions?)
  - Pearson *r* correlation: normal distribution + linearity?
  - Kendall rank correlation: ordinal data?
  - Spearman correlation: monotonic relationships?
- Correlation/ dependence/ association between independent and dependent variables?

Independent variable	Dependent variable: Continuous	Dependent variable: Categorical
Continuous	Correlation analysis	Linear discriminant analysis
Categorical	ANOVA	Chi-square test

Independence Analysis – Chi-square for Categorical Data

- X2 (chi-square) test: $$\chi^2 = \sum_{i}^{n} \frac{(O_i - E_i)^2}{E_i}$$ observed, expected
  - Null hypothesis: the two distributions are independent
  - A chi-square test for independence compares two variables in a contingency table to see if they are related.

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

How to derive the expected 90?

450/1500 * 300 = 90

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- The larger the X2 value, the more likely the variables are related.

Which Statistical Test?

- Choosing the correct statistical test
- https://stats.idre.ucla.edu/other/mult-pkg/whatstat/

idre UCLA Institute for Digital Research and Education

Search this website ...

HOME | SOFTWARE | RESOURCES | SERVICES | ABOUT US | DONATE

CHOOSING THE CORRECT STATISTICAL TEST IN SAS, STATA, SPSS AND R

The following table shows general guidelines for choosing a statistical analysis. We emphasize that these are general guidelines and should not be construed as hard and fast rules. Usually your data could be analyzed in multiple ways, each of which could yield legitimate answers. The table below covers a number of common analyses and helps you choose among them based on the number of dependent variables (sometimes referred to as outcome variables), the nature of your independent variables (sometimes referred to as predictors). You also want to consider the nature of your dependent variable, namely whether it is an interval variable, ordinal or categorical variable, and whether it is normally distributed (see What is the difference between categorical, ordinal and interval variables? for more information on this). The table then shows one or more statistical tests commonly used given these types of variables (but not necessarily the only type of test that could be used) and links showing how to do such tests using SAS, Stata and SPSS.

Number of Dependent Variables	Nature of Independent Variables	Nature of Dependent Variable(s)	Test(s)	How to SAS	How to Stata	How to SPSS	How to R

Variable Reduction Methods

- Some variable reduction methods use the original variables as inputs into subsequent models

  ⇨ Variable Selection.

- Some variable reduction methods use combinations of the original variables as inputs into subsequent models

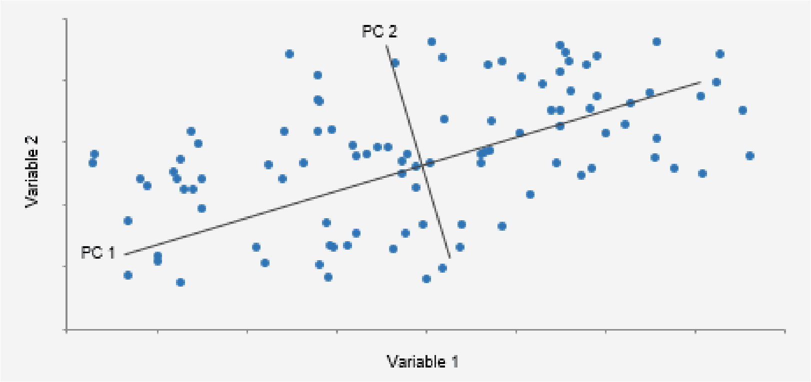
  ⇨ Variable/ Dimension Reduction

Variable Reduction – Principal component analysis

- Principal components are constructed as mathematical transformations of the input variables. Each is an uncorrelated, linear combination of original input variables.

$$pc_1 = a_1 x_1 + b_1 x_2 + c_1 x_3$$

- The coefficients of such a linear combination are the eigenvectors of the correlation or covariance matrix.
- The principal components are sorted by descending order of the eigenvalues.
- The eigenvalues represent the variances of the principal components.

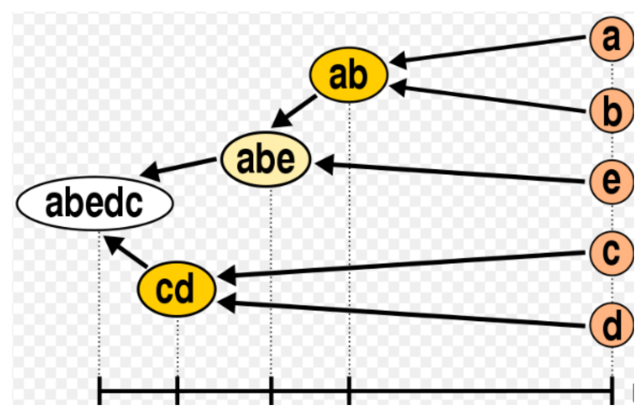


Principal Component Analysis

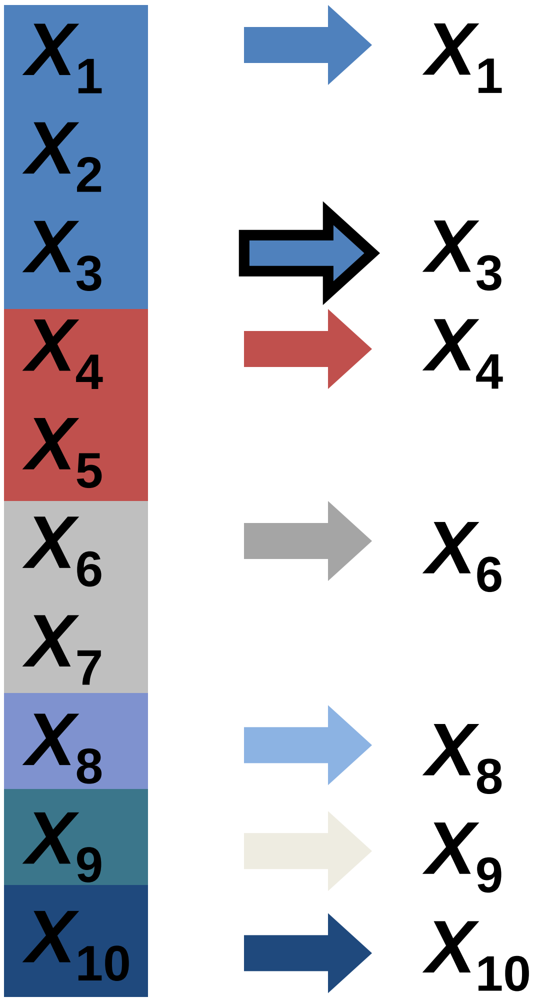
- Pros
  - Constructed variables are definitely uncorrelated.
  - The selection order of the principal components is automatically determined. The first principal component represents more of data variation than the second, and so on.
  - A small number of principal components can be kept to explain a lot of the variation in the data cloud.
- Cons
  - Difficult or impossible to interpret the constructed principal components (interpretability).
  - Difficult to know how many principal components should be selected as new input variables.
  - All original input variables still used, since they build the principal components.

Variable Reduction – Variable clustering

- The variable clustering algorithm divides the input variables into hierarchical clusters. The algorithm is divisive, at the start, all variables are in one single cluster.
  - Select one variable (or the cluster component) from each cluster as a cluster representative.
  - The representative variables (or components) are used as input variables, and the other input variables are rejected.
  - Each cluster can be described as a linear combination of the variables in the cluster.



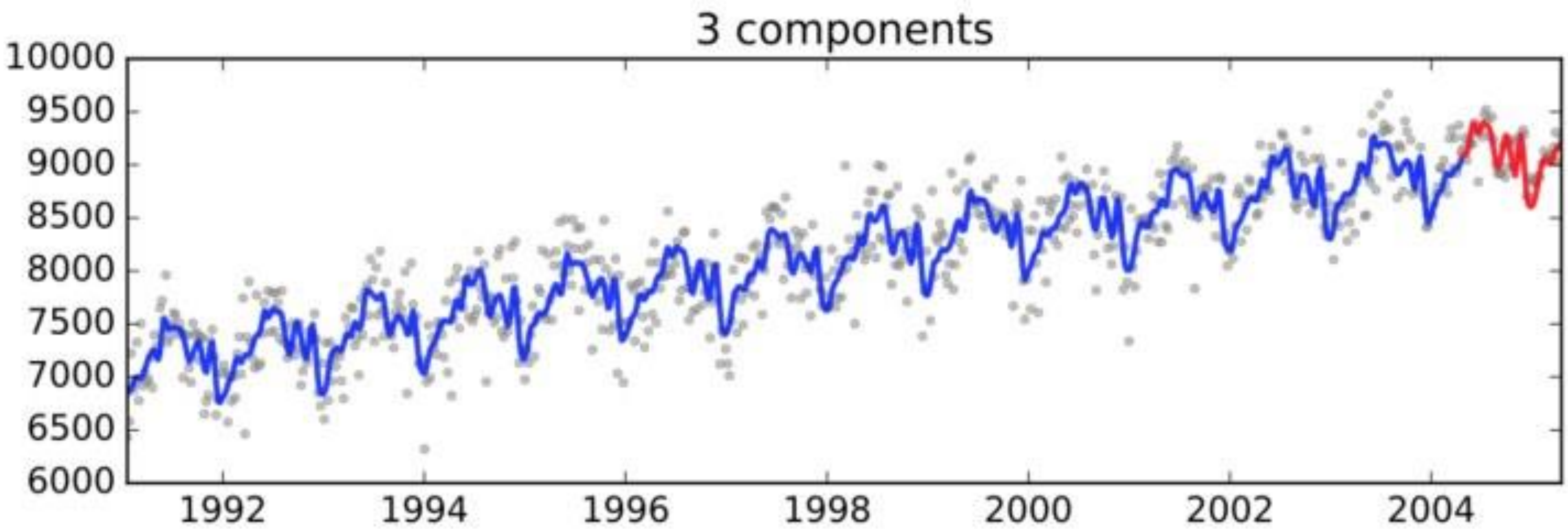
Variable Clustering



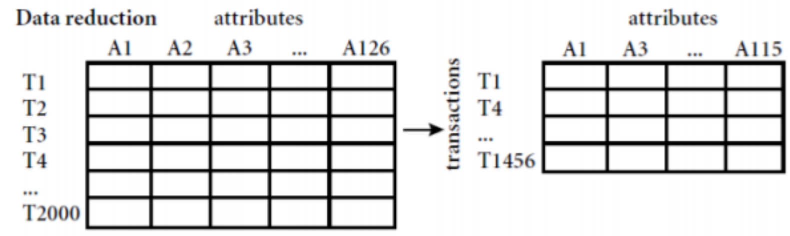
Inputs are selected by
- cluster representation
- expert opinion
- target correlation.

Feature Engineering

- Feature engineering: the process of using domain knowledge of the data to create features that make machine learning algorithms work (wiki).
  - Bucketing
  - Capture trends with ratios, differences, etc.
  - Time-series features



Numerosity Reduction



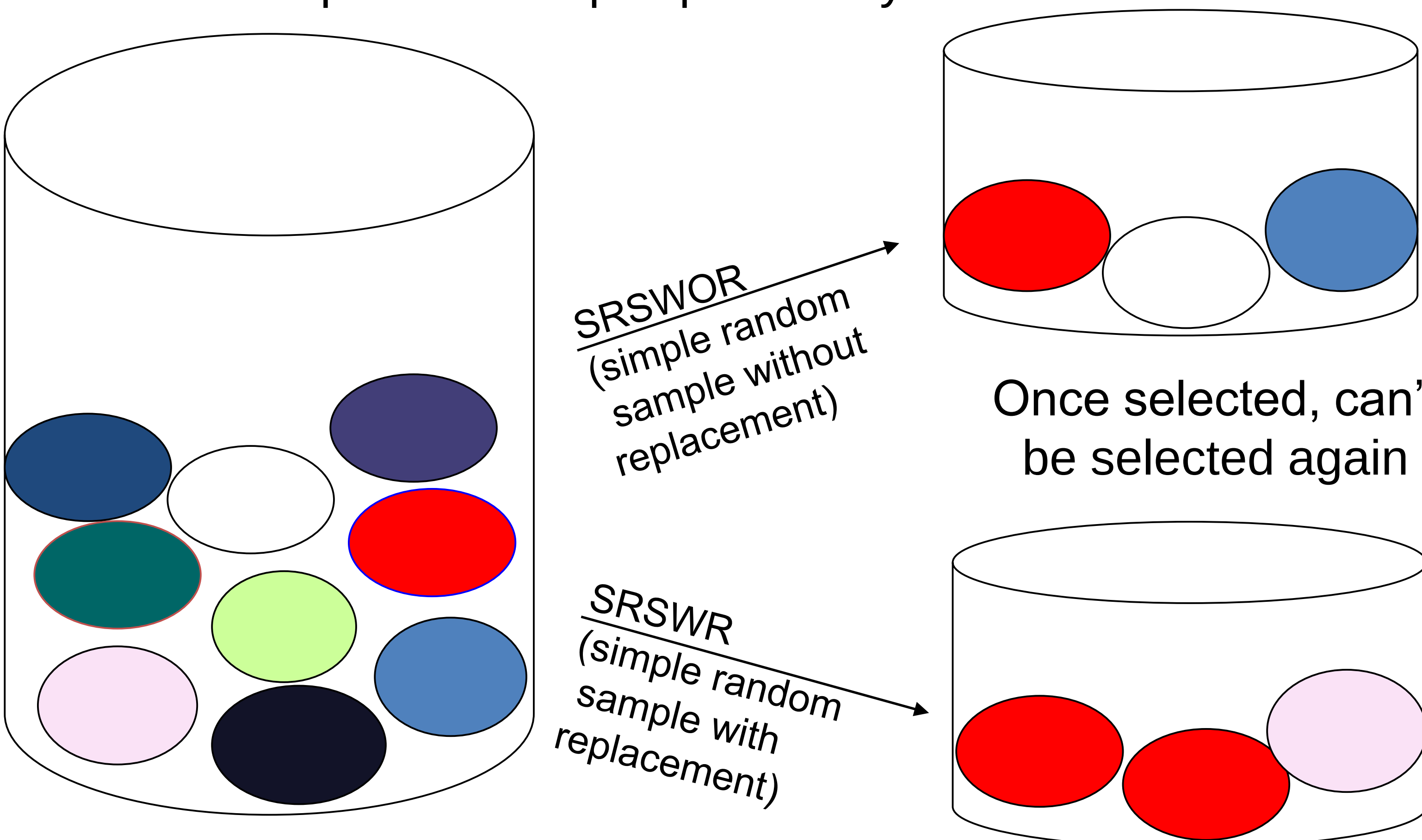
- Non-parametric methods
  - Do not assume models
  - Sampling, clustering, histograms, etc.
- Parametric methods
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers), e.g., regression, log-linear models

Sampling

- Sampling: obtaining a small set of samples to represent the whole data set (assuming the computational complexity is potentially sub-linear to the size of the data)
  - Simple random sampling
    - » There is an equal probability of selecting any particular object
  - Sampling without replacement
    - » Once an object is selected, it is removed from the population
  - Sampling with replacement
    - » A selected object is not removed from the population
  - Stratified sampling:
    - » Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - » Used in conjunction with skewed data

Sampling – Without or With Replacement

All tuples have equal probability of selection



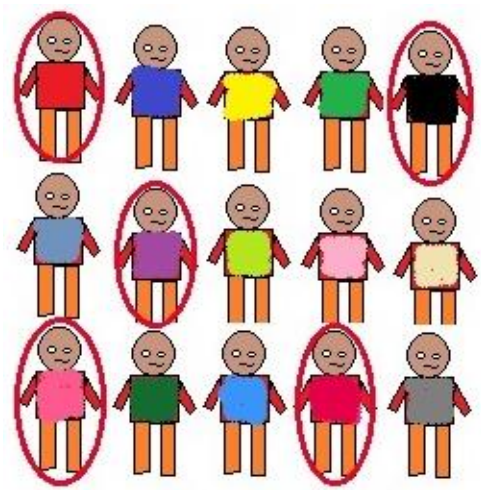
Raw Data

Once selected, can't be selected again

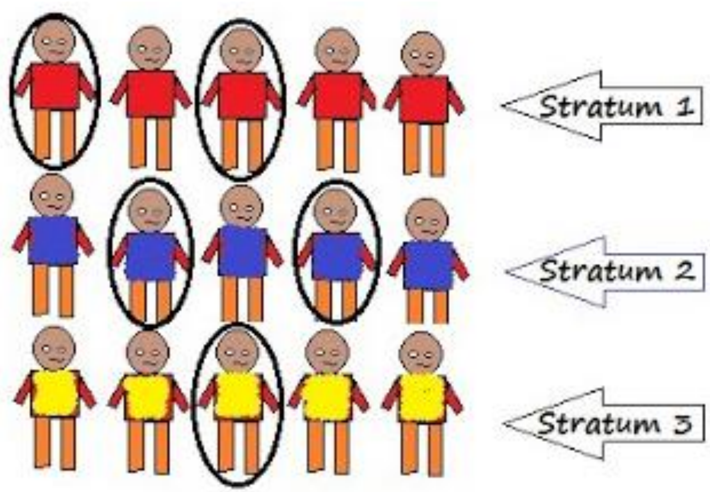
Once selected, can be selected again

Sampling – Stratified

Divides the objects of the population into small subgroups (strata) based on the similarity in such a way that the objects within the group are homogeneous and heterogeneous among the other subgroups



Simple random sampling

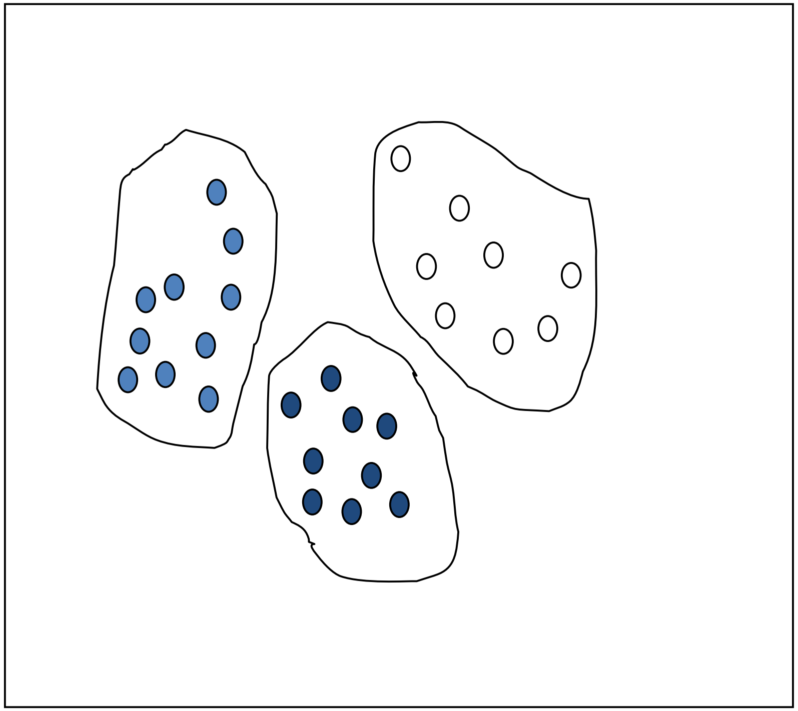


Stratified sampling

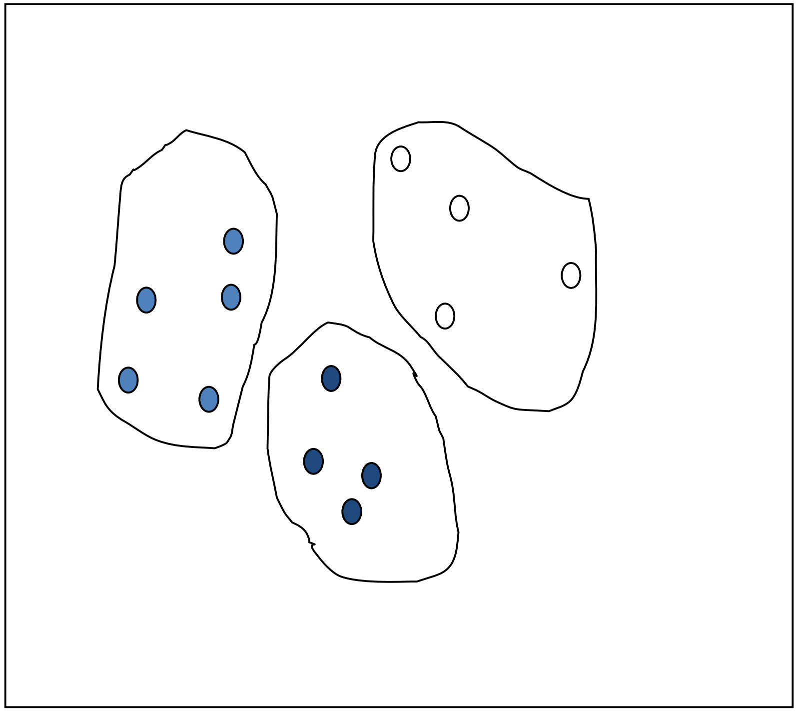
https://towardsdatascience.com/sampling-techniques-a4e34111d808

Sampling – Clustered

After data is clustered or stratified, perform a simple random sample (with or without replacement) in each cluster or strata



Raw Data



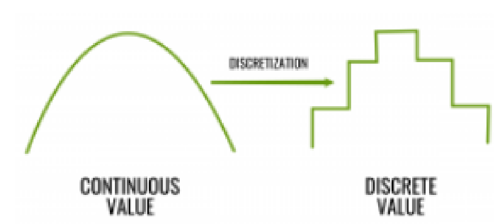
Cluster/Stratified Sample

Outline

- Data integration
- Data description, summarisation and visualisation
- Data cleaning
- Data reduction
- **Data transformation**

Data transformation $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

  - Normalisation/ Standardisation
  - Data discretization
  - Data generalisation



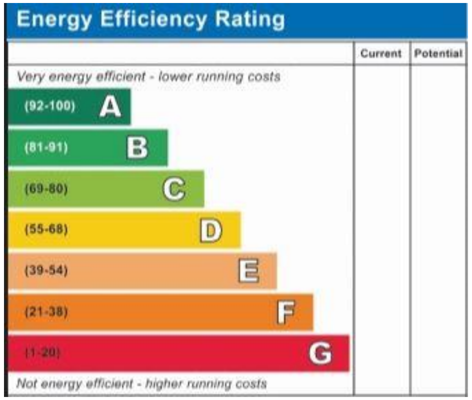
Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values, where each old value can be identified with one of the new values
- Relevant methods
  - Smoothing (for noisy data)
  - Aggregation and summarisation
  - Normalisation/ Standardisation: scaling to fall within a smaller, specified range
    - » min-max normalisation
    - » z-score normalization
  - Discretisation
  - Generalisation

Data Transformation – Examples

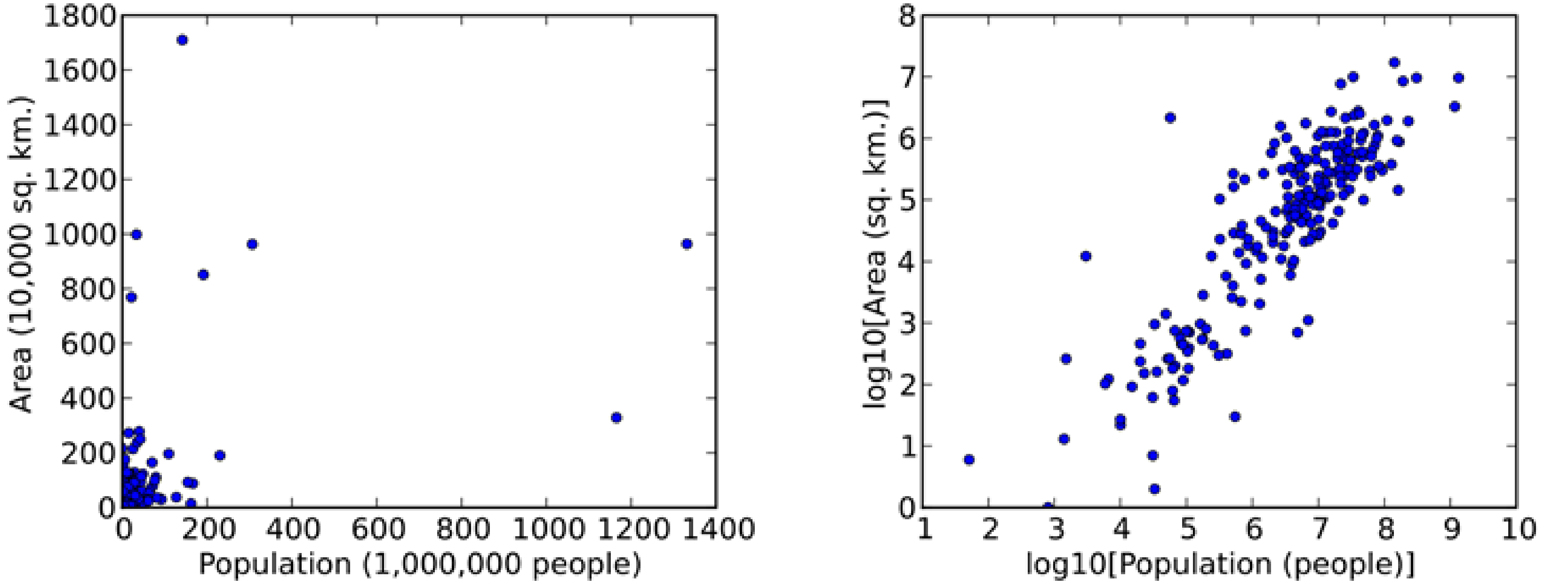
- Standardise numeric values: e.g., all numeric values are replaced by the notion of "how far is this value from the average?"
  - Standardisation is useful, although it sometimes has no effect on the results (such as for decision trees and regression)
- Change counts into percentages.
- Translate dates to durations.
- Capture trends with ratios, differences, etc.
- Replace categorical values with appropriate numeric values (many techniques work better with numeric values than with categorical values)

Year		Age
1881	- - ->	139
2011	- - ->	9

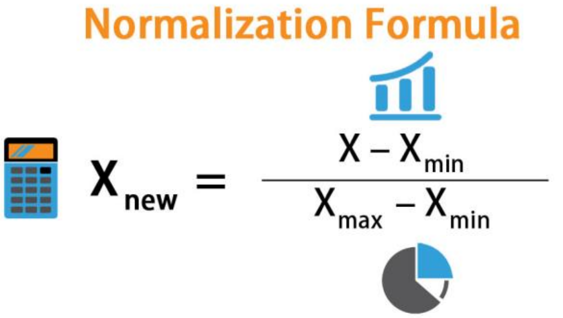


Data Transformation – Examples

- Transform variables to bring information to the surface.
- Transform using mathematical functions, such as logs, reciprocal, or square root, for "stretching" and "squishing"



Data Normalisation



- min-max normalisation
  - Different range -> similar range for variables

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- Example – income, min £12,000, max £98,000 – map to 0.0 – 1.0
- £73,600 is transformed to :

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

Data Normalisation

$$Z = \frac{\overset{\text{Score}}{x} - \overset{\text{Mean}}{\mu}}{\underset{\text{SD}}{\sigma}}$$

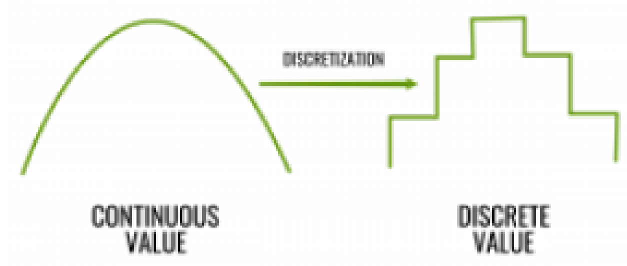
- z-score normalisation (μ: mean, σ: standard deviation)

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- Z-score: The distance between the raw score and the population mean in the unit of the standard deviation
- Let μ = 54,000, σ = 16,000.

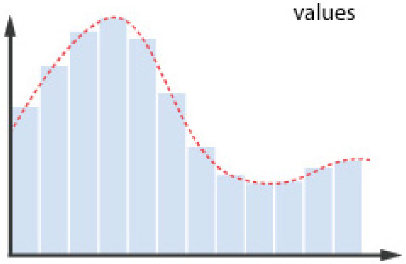
$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

Data Discretisation - Numeric



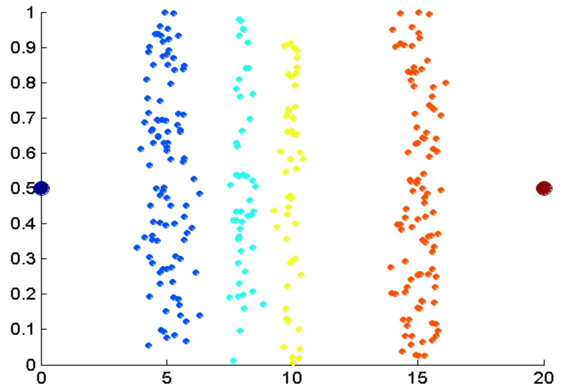
- Three typical types of attributes
  - Nominal—values from an unordered set, e.g., colour, profession
  - Ordinal—values from an ordered set, e.g., military or academic rank
  - Numeric—real numbers, e.g., integer or real numbers
- Discretisation: divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretisation
  - Discretisation can be performed recursively
  - Prepare for further analysis, e.g., classification

Data Discretisation Methods

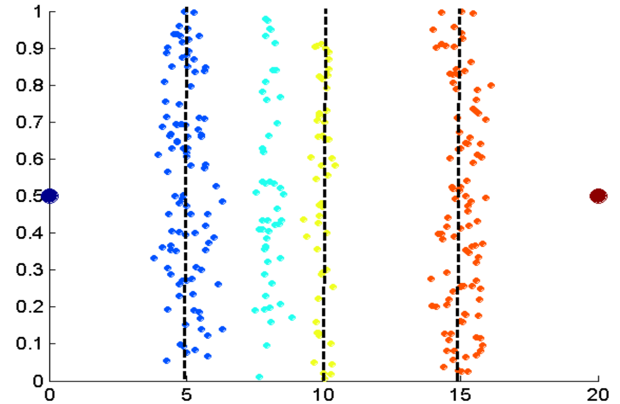


- Binning and histogram analysis
  - Top-down split, unsupervised
- Clustering analysis
  - Unsupervised, top-down split or bottom-up merge
- Decision-tree analysis
  - Supervised, top-down split
- Correlation (e.g., χ^2) analysis
  - Unsupervised, bottom-up merge

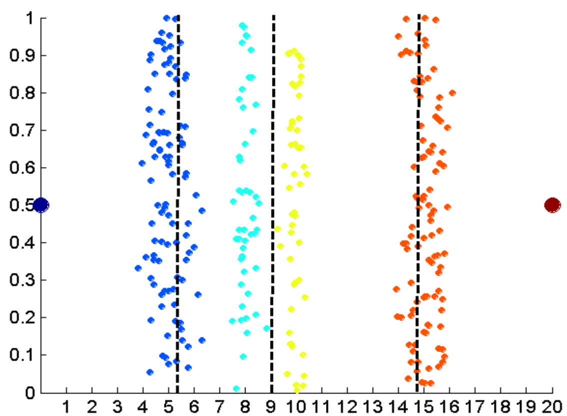
Unsupervised Discretisation - Binning vs. Clustering



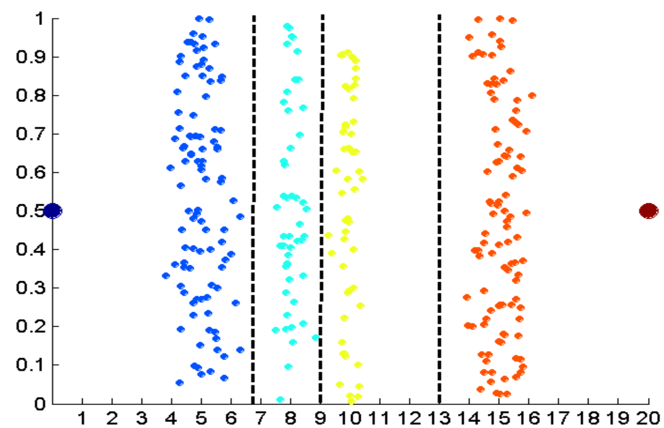
Data



Equal width (distance) binning



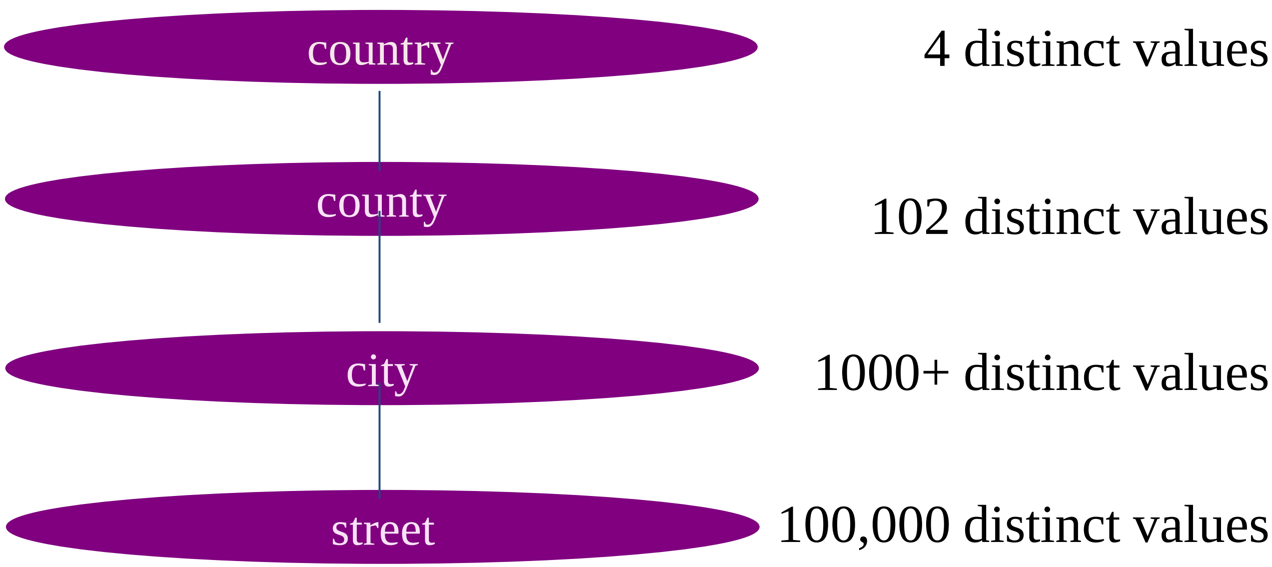
Equal depth (frequency) (binning)



K-means clustering leads to better results

Data Generalisation - Categorical

- Generalisation: generalise/replace low level concepts (such as age ranges) by higher level concepts (such as young, middle-aged, or senior)
  - Specification or automatic generation of hierarchies (or attribute levels) by analysing the number of distinct values, e.g., {street, city, county, country}



Case Study 3 – Data Preprocessing

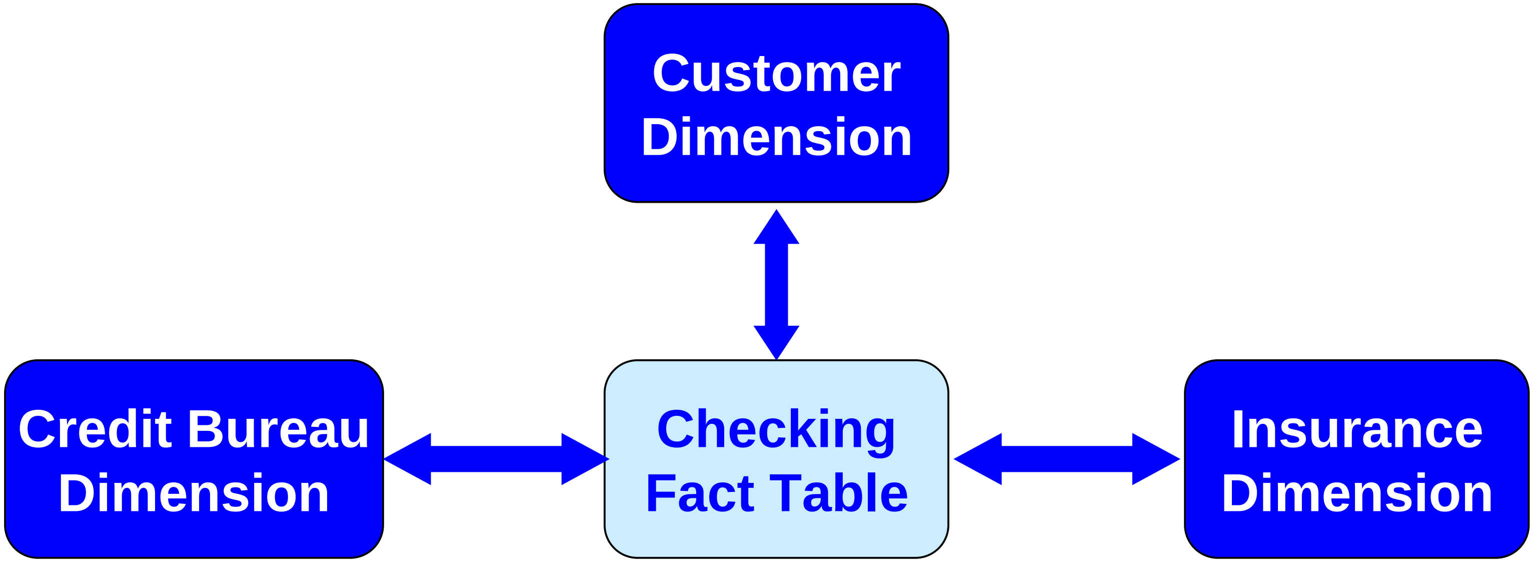
- Business: Large financial institution
- Objective: from a population of existing clients with sufficient tenure and other qualifications, identify a subset of clients who are most likely to have interest in an insurance investment product (INS).

Analytic Objective Example

- The financial institution has highly detailed data that is challenging to transform into a structure suitable for predictive modelling. As is the case with most organisations, the financial institution has a large amount of data about its customers, products, and employees in transactional systems.
- This transactional information can be extracted, transformed, and loaded into a data mart for the Marketing Department.

Financial Institution Target Star Schema

- The analyst can produce, from the financial institution's source data, a dimensional data model that is a star schema.



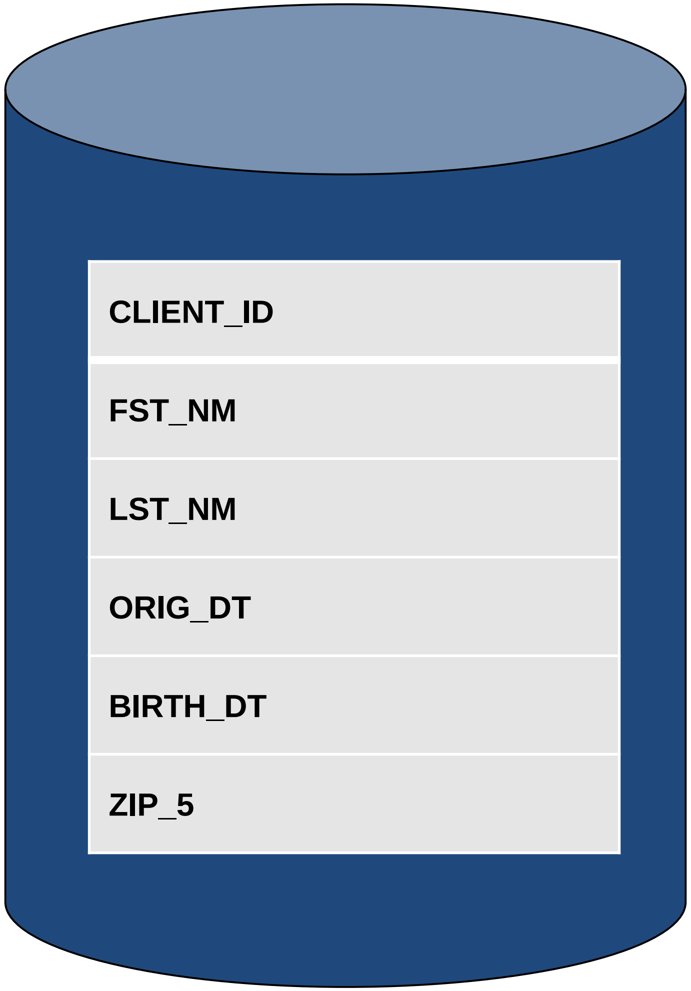
Checking_transactions Table

- The **checking_transactions** table contains the following attributes, one per a record fact.
- This fact contains some measured or observed variables.
- The fact table contains the data, and the dimensions identify each tuple in the data.

CHECKING_ID
CHKING_TRANS_DT
CHKING_TRANS_AMT
CHKING_TRANS_CHANNEL_CD
CHKING_TRANS_METHOD_CD
CHKING_TRANS_TYPE_CD

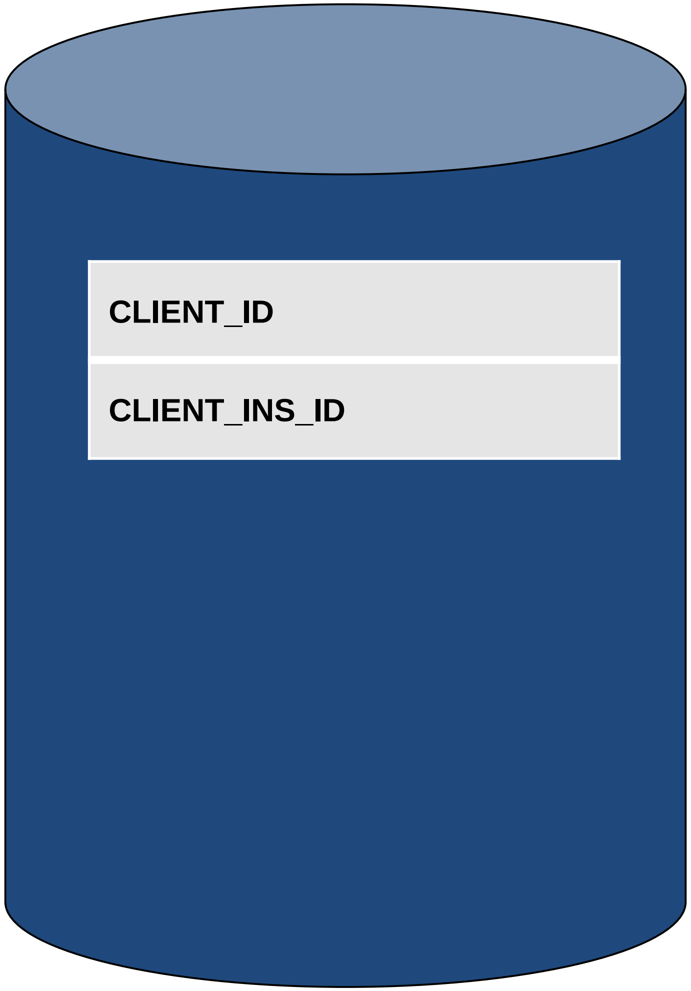
Client Table

- The **client** table contains client information.
- In practice, this data set could also contain address and other information.
- For this demonstration, only **CLIENT_ID**, **FST_NM**, **LST_NM**, **ORIG_DT**, **BIRTH_DT**, and **ZIP_5** are used.



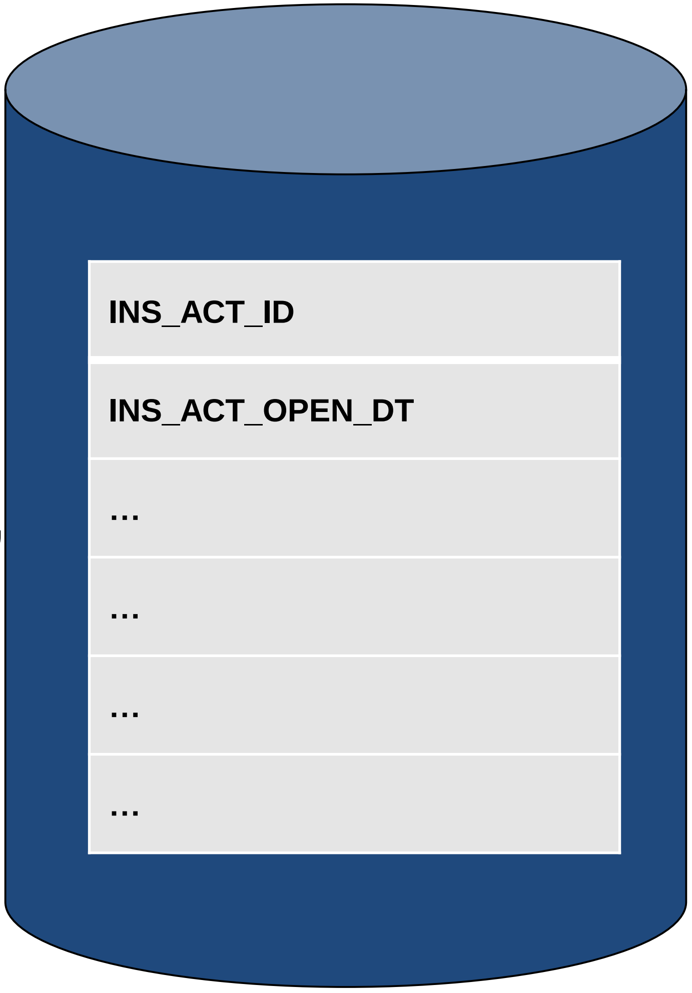
Client_ins_account Table

- The **client_ins_account** table matches client IDs to INS account IDs.



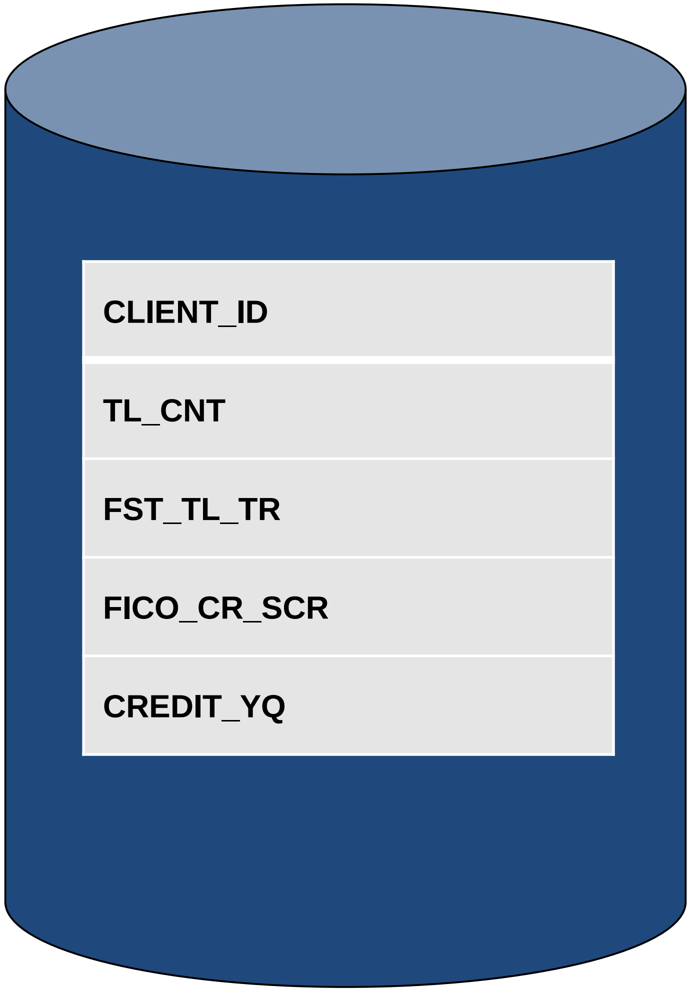
Ins_account Table

- The **ins_account** table contains the insurance account information.
- In practice, this data set would contain other fields such as rates, maturity dates, and initial deposit amount.
- For this demonstration, only **INS_ACT_ID** and **INS_ACT_OPEN_DT** are used.



Credit_bureau Table

- The **credit_bureau** table contains credit bureau information.
- In practice, this data set could contain credit scores from more than one credit bureau and also a history of credit scores.



SAS Enterprise Guide – Data Management

- SAS Enterprise Guide can be used for data management, as well as a wide variety of other tasks:
  - Data exploration
  - Querying and reporting
  - Graphical analysis
  - Statistical analysis
  - Scoring
  - …

Business Scenario 1

- The head of Marketing wants to know which customers have the highest propensity for buying insurance products from the institution.
- This could present a cross-selling opportunity.
- Create part of an analytical data mart by combining information from many tables: checking account data, customer records, insurance data, and credit bureau information.

Input Files

- client_ins_account.sas7bdat
- credit_bureau.sas7bdat
- ins_account.sas7dbat
- client.sas7bdat

CLIENT_INS_ACCOUNT

Filter and Sort | Query Builder | Data

	CLIENT_ID	INS_ACT_ID
1	0960116451	000006791438
2	0934829685	000006651763
3	0429866676	000005015721
		000006222459
		000006249488
		000001520515

CREDIT_BUREAU

Filter and Sort | Query Builder | Data | Describe | Graph | Analyze | Export | Send To

	CLIENT_ID	TL_CNT	FST_TL_YR	FICO_CR_SCR	CREDIT_YQ
1	0996257578	15	1995	723	20023
2	0603759831	14	1993		
3	0640025008	12	1969		
4	0724761848	11	1972		
5	0713012963	10	1997		
6	0311961267	12	1978		
7	0279674316	4	1979		
8	0023159720	19	1983		
			1986		
			1981		

INS_ACCOUNT

Filter and Sort | Query Builder | Data | Desc

	INS_ACT_ID	INS_ACT_OPEN_DT
1	000008429518	30DEC2002
2	000006791438	12JAN2003
3	000001537872	17SEP2001
4	000006651763	03OCT2002
5	000004446880	10JUL2001
6	000005828507	26APR2002
7	000005015721	04APR2003
8	000003783089	10AUG2001
9	000004339495	20JUL2001
10	000006222459	10MAR2003
11	000008180235	01JAN2002
12	000001344091	22JUL2001
13	000006249488	28JAN2003
14	000001520515	08JUN2002
15	000008797409	08FEB2002

CLIENT

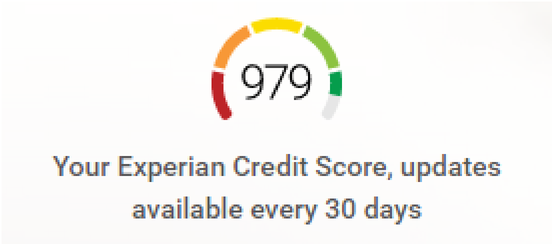
Filter and Sort | Query Builder | Data | Describe | Graph | Analyze | Export | Send To

	CLIENT_ID	FST_NM	LST_NM	ORIG_DT	BIRTH_DT	ZIP_5
1	0996257578	JENNIE	JOHNSON	25AUG1998	19OCT1973	91320
2	0603759831	LORETTA	MITCHELL	15JUN2001	19DEC1969	92129
3	0640025008	WALTER	THAMMAVONG	27DEC1999	12MAY1947	32766
4	0724761848	ANDRE	DAHLEM	07FEB1997	29APR1950	60102
5	0221655391	AMADA	SNYDER	01FEB2001	.	85044
6	0713012963	MARION	BURROUGHS	12SEP2000	26AUG1977	94928
7	0311961267	DAVE	PINKSTON	16DEC2001	15AUG1954	80904
8	0341857444	BORIS	GRISHAM	08JAN1990	.	30076
9	0023159720	FRANCIS	MCMILLAN	10AUG1995	28OCT1956	90255
10	0365287752	SONIA	SWITZER	30MAY1999	21JAN1962	95123

The University of Manchester
Alliance Manchester Business School

Business Scenario 2

- Investigate the distribution of credit scores.
  - Create a report of credit scores by customers without insurance and customers with insurance.
- Does age have an influence on credit scores? Which customers have the highest credit scores, young customers or older customers?
  - Create a graph of credit scores by age.



http://www.experian.co.uk/

Exploratory Analysis

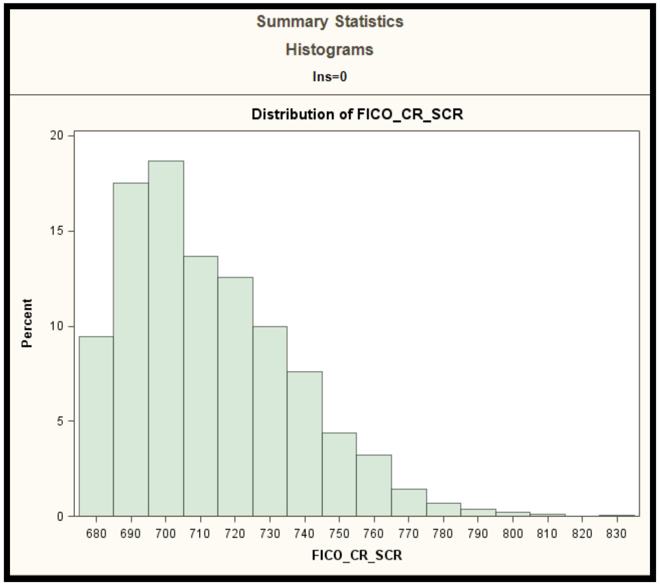
Summary Statistics
Results

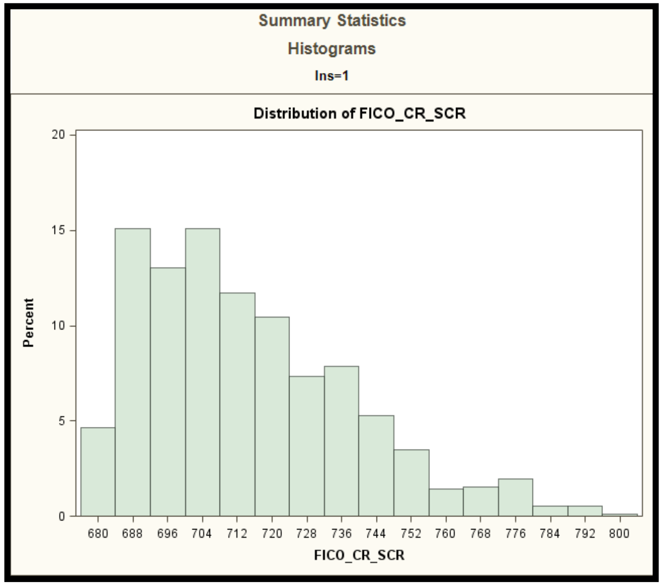
Ins=0

Analysis Variable : FICO_CR_SCR				
Mean	Std Dev	Minimum	Maximum	N
712.4001289	23.8778122	681.0000000	833.0000000	1552

Ins=1

Analysis Variable : FICO_CR_SCR				
Mean	Std Dev	Minimum	Maximum	N
713.8015464	24.1675719	681.0000000	800.0000000	776





Case Study 4 – Data Preprocessing to Analytics

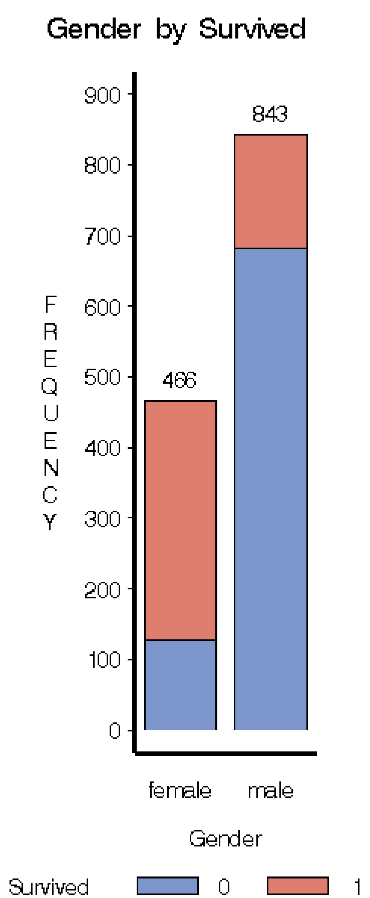
- Case: TITANIC Passenger's Survival Analysis
  - The TITANIC data set consists of several variables that describe the passengers on board the ill-fated RMS Titanic, which sank on its maiden voyage in April 1912.
- Objectives:
  - Predict the Passenger's survival status by relevant variables

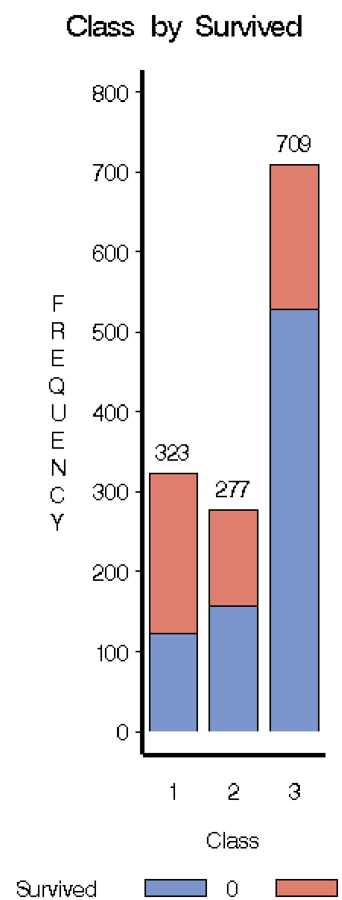
TITANIC Data Set

Variable	Description	Role	Level
Age	Age of the passenger in years	Input	Interval
Class	Ticket class (1, 2, or 3)	Input	Ordinal
Fare	Ticket fare	Input	Interval
Gender	Gender of the passenger (male, female)	Input	Binary
Name	Passenger’s name	ID	Nominal
Survived	Passenger’s survival status (1=survived, 0= died)	Target	Binary

Preliminary Analysis

- A simple analysis shows men and 3rd class passengers much more likely to die





Summary

- Introduction to data preprocessing/ preparation
- Data integration - integration of multiple data files, databases, or sources
- Data description, summarisation and visualisation
- Data cleaning - fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies
- Data reduction - obtain reduced representation in volume but produces the same or similar analytical results
- Data transformation - normalisation, discretisation, and generalisation