BIG DATA ANALYTICS
Data Preparation

Olga Klopp
klopp@essec.edu

Outline

Why prepare data?

Major Tasks in Data Preparation:

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation

Big Data Preparation

Outline

Why prepare data?

It's tempting to dive right into the modeling step without looking very hard at the dataset first, especially when you have a lot of data...

Resist the temptation!

Data in the real world is dirty:

- **incomplete**: lacking attribute values, lacking certain attributes of interest or containing only aggregate data
  - occupation=" "
- **noisy**: containing errors or outliers
  - Salary="-10", Age="222"
- **inconsistent**: containing discrepancies in codes or names
  - Age="42" Birthday="03/07/1997"
  - Was rating "1,2,3", now rating "A, B, C"
  - discrepancy between duplicate records

Why prepare data?

- If you don't take the time to examine the data before you start to model, you may find your-self redoing your work repeatedly as you discover bad data fields or variables that need to be transformed before modeling.
- In the worst case, you'll build a model that returns incorrect predictions - and you won't be sure why.

By addressing data issues early, you can save yourself some unnecessary work, and a lot of headaches!

Outline

Major Tasks in Data Preparation

- **Data cleaning**:
  - fill in missing values,
  - smooth noisy data,
  - identify or remove outliers,
  - resolve inconsistencies
- **Data integration**: integration of multiple databases or files
- **Data reduction**: obtains reduced representation in volume but produces the same or similar analytical results
- **Data transformation**

Data cleaning

- You can spot some problems just by using summary statistics
- Other problems are easier to find visually

Outliers

"An outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism"

- Outlier detection can be used for fraud detection
  - do nothing
- Data cleaning
  - enforce upper and lower bounds
  - let binning handle the problem

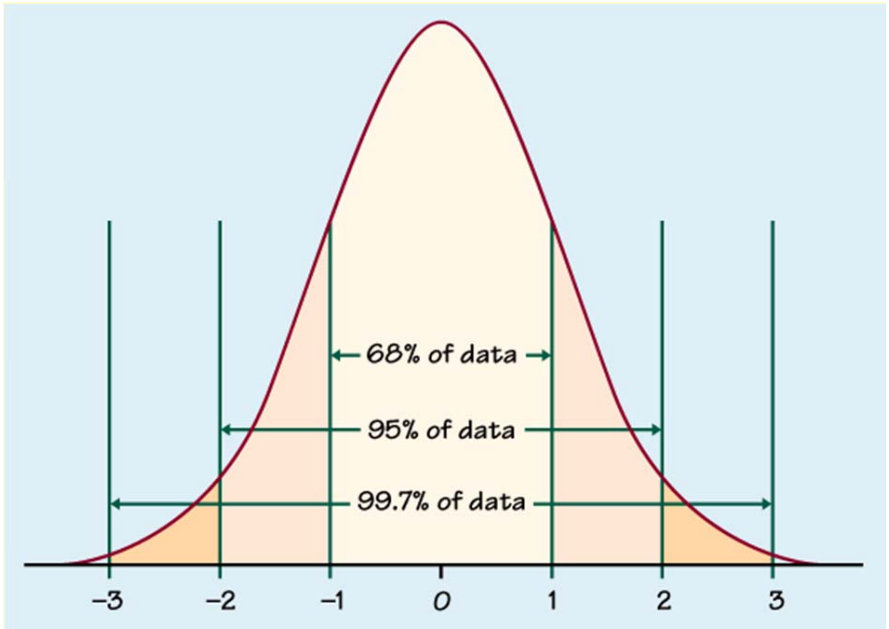
Outlier detection

For **Univariate** distribution:

- Compute mean $\bar{x}$ and std. deviation s (normal distribution assumed):

  for $k = 2$ or 3, x is an outlier if outside limits

$$(\bar{x} - ks, \bar{x} + ks)$$



Outlier detection

For **Univariate** distribution:

Boxplot:

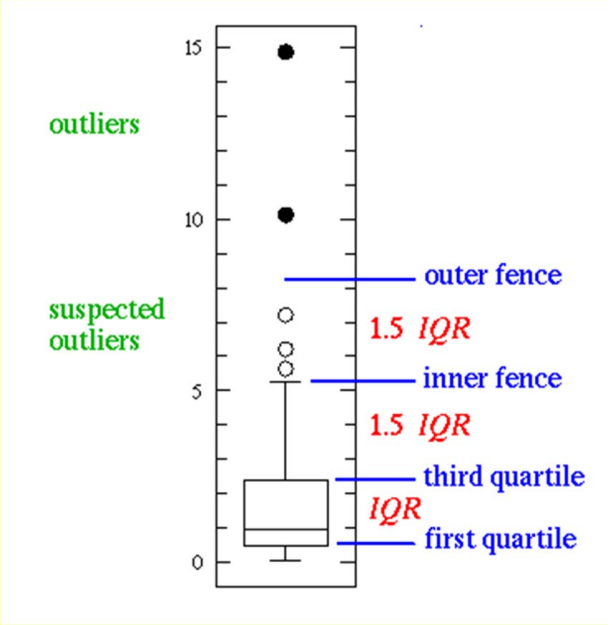
- An observation is an extreme outlier if it lies outside

  $$(Q1 - 3 \times \mathrm{IQR}, Q3 + 3 \times \mathrm{IQR})$$

  where $\mathrm{IQR} = Q3 - Q1$ (IQR = Inter Quartile Range)

- a mild outlier if it lies outside of the interval

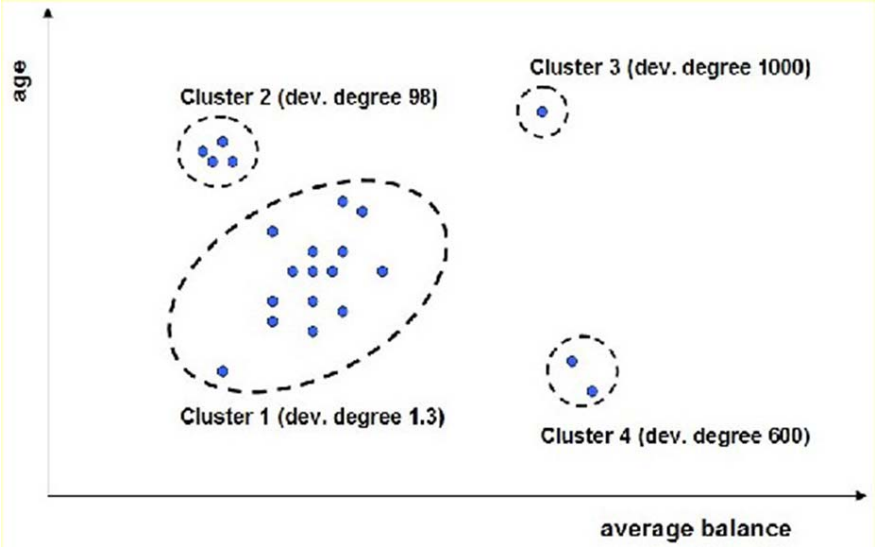
  $$(Q1 - 1.5 \times \mathrm{IQR}, Q3 + 1.5 \times \mathrm{IQR})$$



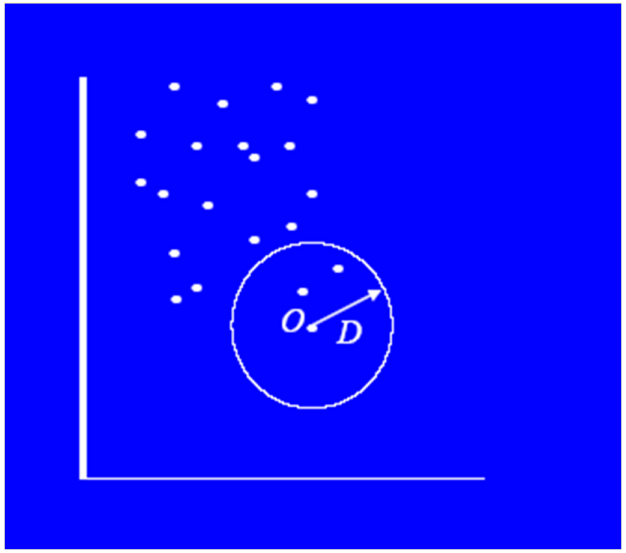
Outlier detection

For **Multivariate** distribution:

(1) Clustering: Very small clusters are outliers

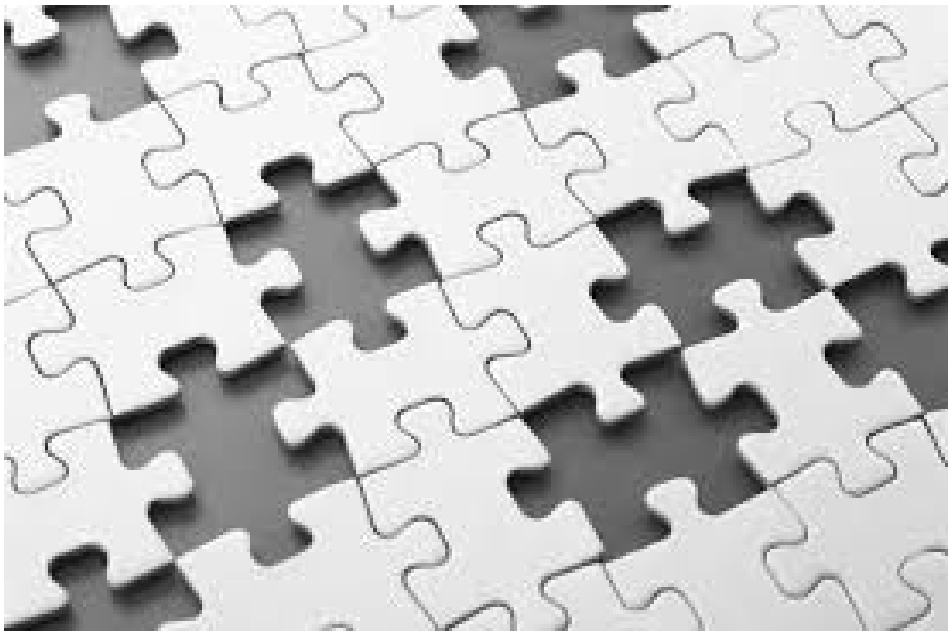


(2) Distance based: an instance with very few neighbors is an outlier



MISSING DATA

- There are always MVs in a real dataset
- MVs may have an impact on modeling, in fact, they can destroy it!



MISSING DATA

Missing data may be due to:

- equipment malfunction
- inconsistency with other recorded data and thus deleted
- data not entered due to misunderstanding
- certain data may not be considered important at the time of entry
- history or changes of the data not registered
- ...

MISSING DATA

- A few missing values may not be a real problem
- If a particular data field is largely unpopulated, it shouldn't be used as an input without some repair:
  - Some tools ignore missing values, others use some metric to fill in
  - In R, many modeling algorithms will, by default, quietly drop rows with missing values
- If a particular data field is largely unpopulated, its worth trying to determine why:

  Sometimes the fact that a value is missing is informative in and of itself

Missing Data

Whatever the reason for missing data, you must decide on the most appropriate action:

- Do you include a variable with missing values in your model, or not?
- If you decide to include it, do you drop all the rows where this field is missing, or do you convert the missing values to 0 or to an additional category?

Replacing missing values without elsewhere capturing that information removes information from the dataset!

How to Handle Missing Data?

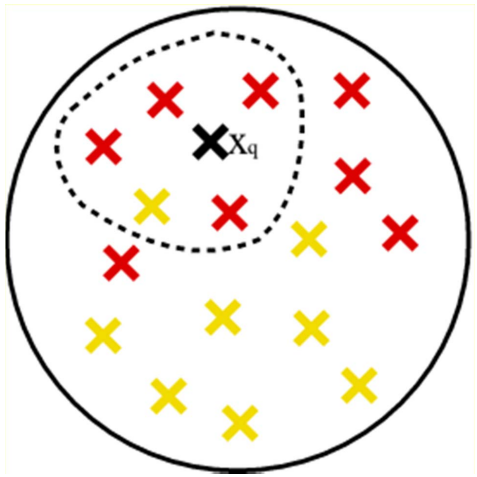
- **Ignore records** (use only cases with all values)
  - can lead to insufficient and/or biased sample sizes
- **Ignore attributes with missing values** (use only attributes with all values):
  - may leave out important features
- **Fill in the missing value**:
  - use a global constant (e.g., unknown)
  - use the attribute mean (it will do the least harm to the mean of the existing data)
  - use the attribute mean for all samples belonging to the same class to fill in the missing value

Fill in the missing value

- **Use the most probable value**:
  - Inference-based such as Bayesian formula or decision tree
- **Identify relationships among variables**:
  - Linear regression, Multiple linear regression, Nonlinear regression

Fill in the missing value

- **Nearest-Neighbour estimator**:
  - Finding the k neighbours nearest to the point and fill in the most frequent value or the average value
  - Finding neighbours in a large dataset may be slow

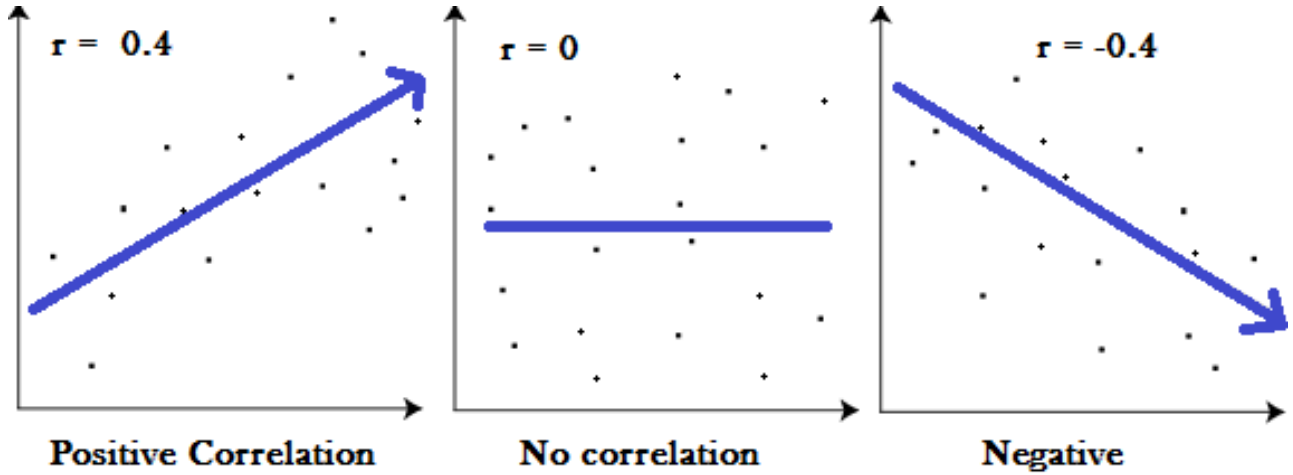


Handling Redundancy in Data Integration

- Redundant data occurs often when integrating databases:
  - The same attribute may have different names in different databases
  - One attribute may be a derived attribute in another table:
    - may be detected by correlation analysis:

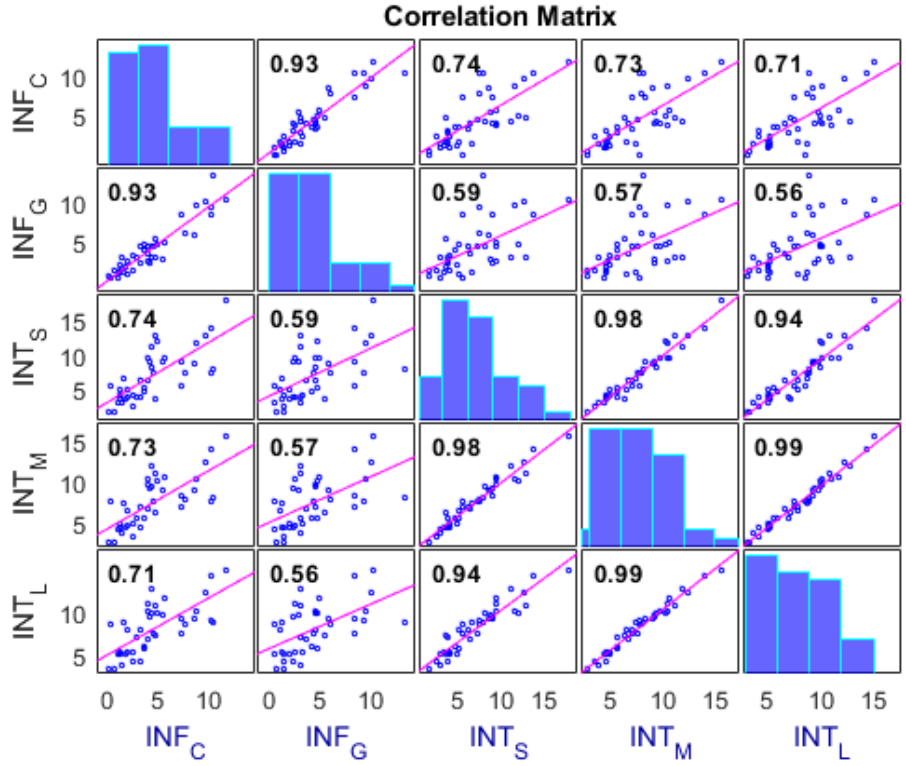
$$r_{XY} = \frac{\frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2}\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(y_n - \bar{y})^2}}$$

Recall: Correlation coefficient



- Correlation coefficient is used to find how strong a relationship is between data
- The formula returns a value between -1 and 1, where:
  - 1 indicates a strong positive relationship
  - -1 indicates a strong negative relationship.
  - A result of zero indicates no relationship at all.

Scatter Matrix



- A scatter matrix consists of several pair-wise scatter plots of variables presented in a matrix format
- It can be used to determine whether the variables are correlated and whether the correlation is positive or negative
- Histograms of the variables appear along the matrix diagonal.

Data Reduction

The reduction of the number of dimensions of the problem:

- the reduction of the number of individuals
- the number of variables
- the number of categories of the variables.

Reduction of the number of individuals: Sampling

Sampling is the process of selecting a subset of a population to represent the whole, during analysis and modeling

DON'T BE TOO PROUD TO SAMPLE

Sampling

- Many data scientists spend too much time adapting algorithms to work directly with big data
- Often this is wasted effort, as for many model types you would get almost exactly the same results on a reasonably sized data sample!
- You may need to work with all of your data when what you're modeling isn't well served by sampling:
  - characterizing rare events
  - performing bulk calculations over social networks
  - ...

Sampling

Sampling is a necessary task also:

- When you're in the middle of developing or refining a modeling procedure:
  - it's easier to test and debug the code on small subsamples before training the model on the entire dataset
- Visualization can be easier with a sub-sample of the data:
  - e.g. `ggplot` runs faster on smaller datasets
  - too much data will often obscure the patterns in a graph.

Sampling

- The cost of sampling is proportional to the sample size and not to the original dataset size
- Choose a representative subset of the data:
  - Simple random sampling (SRS) (with or without reposition)
  - Stratified sampling: approximate the percentage of each class (or subpopulation of interest) in the overall database

Population segmentation

It may be necessary to segment the population into groups that are homogeneous in relation to the aims of the study, in order to construct a specific model for each segment: **stratification of models**

- It can only be used where the volume of data is large enough for each segment
- The pre-segmentation of the population can often improve the results significantly
- It can be based on rules drawn up by experts or by the statistician
- It can also be carried out more or less automatically:
  - e.g. using a statistical algorithm ("clustering" algorithm).

Population segmentation

There are many ways of segmenting a population:

- The population is segmented according to general characteristics which have no direct relationship with the dependent variable:
  - In the case of a business, it may relate to its size, its legal status or its sector of activity
  - At the Banque de France, the financial health of businesses is modeled by sector:
    - industry, commerce, transport, hotels, cafes and restaurants, construction, and business services
  - For a physical person, we may use characteristics such as age or occupation:
    - specific marketing offers aimed at certain customer seg "youth", "senior", and etc

Segmentation according to general characteristics

- Is usually applied according to rules provided by experts rather than by statistical clustering
- From the statistical viewpoint, this method is not always the best because behavior is not always related to general characteristics

Population segmentation

- Another type of pre-segmentation is carried out on behavioral data linked to the dependent variable
  - e.g. the product to which the scoring relates
- It produces segments that are more homogeneous in terms of the dependent variable
- Generally more effective because it has some of the discriminant power required in the model
- Implemented according to expert rules, or simply on a common-sense basis, or by a statistical method such as a decision tree with one or two levels of depth
  - e.g. a common-sense rule could be to establish a consumer credit propensity score for two segments: those customers who already have this kind of credit, and the rest

Population segmentation

- A third type of pre-segmentation can be required because of the nature of the available data:
  - e.g. they will be much less rich for a prospect than for a customer
  - two populations must be separated into two segments for which specific models will be constructed.

Population segmentation

Rules for classifying:

- Simplicity of pre-segmentation (there must not be too many rules)
- A limited number of segments
- Segment sizes generally of the same order of magnitude
- Uniformity of the segments in terms of the independent variables
- Uniformity of the segments in terms of the dependent variable
  - e.g. avoid combining high-risk and low-risk individuals in the same segment

Test and training splits:

Building a model to make predictions

- You need data to construct the model: **the training set**
  - *The training set is the data that you feed to the model - building algorithm, so that the algorithm can set the correct parameters to best predict the outcome variable*
- You also need data to test whether the model makes correct predictions on new data: **the test set**
  - *The test set is the data that you feed into the resulting model, to verify that the models predictions are accurate*

Unbalanced Target Distribution

Sometimes, classes have very unequal frequency:

- medical diagnosis: 90% healthy, 10% disease
- eCommerce: 99% dont buy, 1% buy
- Security: $> 99.99\%$ of Americans are not terrorists
- Majority class classifier can be 97% correct, but useless

Handling Unbalanced Data

With two classes: let positive targets be a minority

- Separate raw test set (e.g. 30% of data) and raw train set
- Put aside raw test and don't use it till the final model
- Select remaining positive targets from raw train set
- Join with equal number of negative targets from raw train set, and randomly sort it
- Separate randomized balanced set into balanced train and balanced test sets

Reduction of the number of variables:

- Disregarding some variables which are too closely correlated with each other
  - if taken into account simultaneously, would violate the commonly required assumption of non-collinearity of the independent variables (e. g. in linear regression, linear discriminant analysis or logistic regression)
- Disregarding certain variables that are not at all relevant or not discriminant with respect to the specified objective or to the phenomenon to be detected

Reduction of the number of variables:

- Combining several variables into a single variable
  - e.g. the number of products purchased with a two-year guarantee and the number of products purchased with a five-year guarantee can be added together
- Using factor analysis to convert some of the initial variables into a smaller number of variables.

Reduction of the number of variables

- 100 to 200 or more variables may be investigated during the development of a classification or prediction model
- The number of ultimately discriminant variables selected for the calculation of the model is much smaller:
  - 3 or 4 for a simple model
  - 5 - 10 for a model of normal quality
  - 11 - 20 for a very fine model

Types of statistical data:

- **Numerical** (discrete and continuous): person's height, IQ scores ...
- **Categorical** data represent characteristics such as a persons gender, marital status, hometown, or the types of movies they like
  - Categorical data can take on numerical values (such as 1 indicating male and 2 indicating female), but those numbers dont have mathematical meaning (you couldn't add them together, for example)
- **Ordinal** data mixes numerical and categorical data: the data fall into categories but the numbers placed on the categories have meaning
  - e.g. rating a restaurant on a scale from 0 (lowest) to 4 (highest) stars gives ordinal data

Data conversions

Some tools can deal with categorical values but other need fields to be numeric:

- Convert ordinal fields to numeric to be able to use ” $>$ ” and ” $<$ ” comparisons on such fields:

$$(A+) \rightarrow 4.0, (A) \rightarrow 3.7, (B+) \rightarrow 3.3, (B) \rightarrow 3.0$$

- Multi-valued, unordered attributes with small no. of values:
  - e.g. Color=Red, Orange, Yellow, ..., Violet
  - for each value v create a binary flag variable C_v , which is 1 if Color= v, 0 otherwise

Reduction of the number of categories

- Categorical data, many values:
  - e.g. US State Code (50 values) or Profession Code (7,000 values, but only few frequent)
  - Group values naturally: 50 US States $\rightarrow$ 3 or 5 regions
  - Create binary flag-fields for selected values

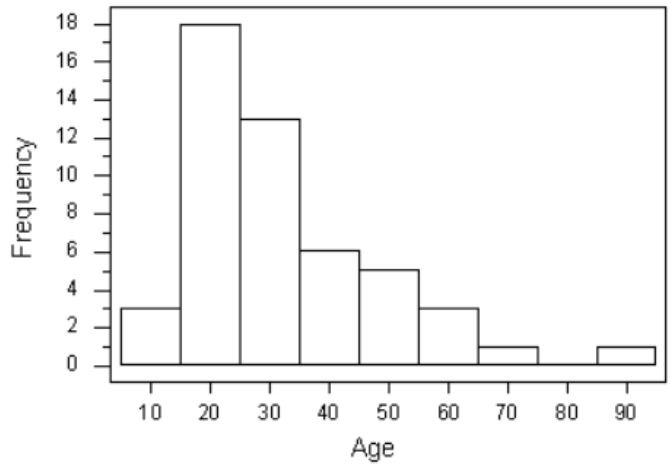
Discretization of continuous variables

Discretization (binning) divides the range of a continuous attribute into intervals

- Some methods require discrete values, e.g. most versions of Naive Bayes
- Reduce data size by discretization
- Prepare for further analysis
- Discretization is very useful for generating a summary of data

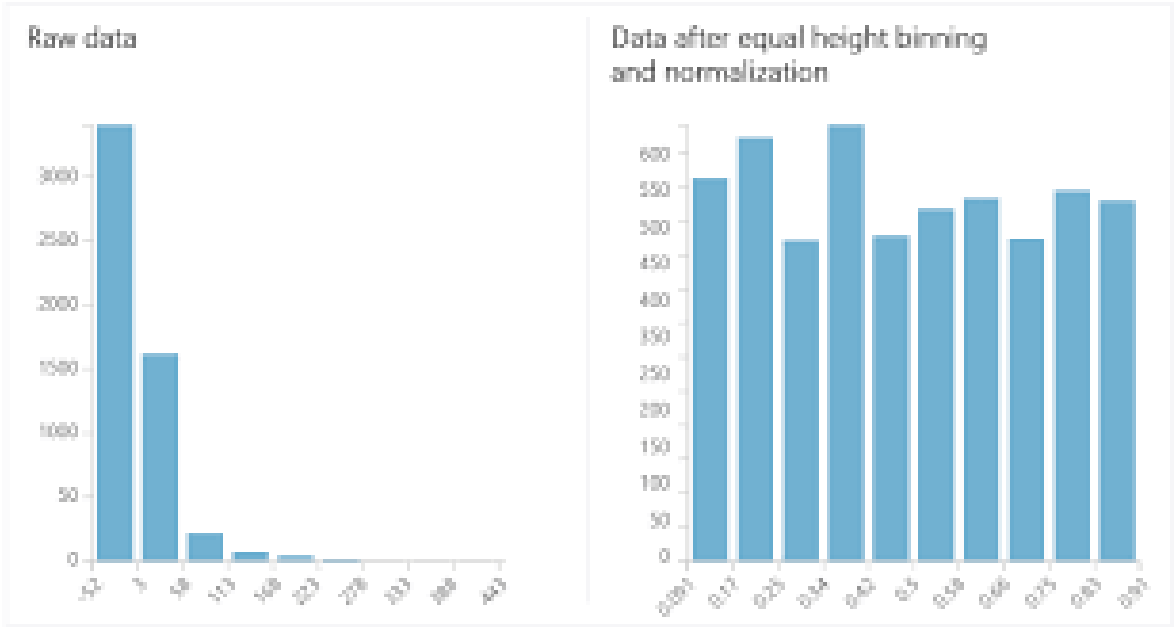
Equal-width Binning

- It divides the range into N intervals of equal size: uniform grid
- Advantage:
  - Simple and easy to implement
  - Produce a reasonable abstraction of data
- Disadvantage:
  - Unsupervised
  - How we choose N?
  - Sensitive to outliers



Equal-depth Binning:

- It divides the range into N intervals, each containing approximately the same number of samples
  - Generally preferred because avoids clumping
  - Give more intuitive breakpoints
- Many other methods exist: 1R Classifier, Entropy Based Discretization...



Normalization

- For distance-based methods, normalization helps to prevent that attributes with large ranges out-weight attributes with small ranges:
  - **min-max normalization**:

$$x' = \frac{x - \min_x}{\max_x - \min_x}(\text{new max} - \text{new min}) + \text{new min}$$

  - **z-score normalization**:

$$x' = \frac{x - \bar{x}}{\sigma_x}$$

  - **normalization by decimal scaling**:

$$x' = \frac{x}{10^j}$$

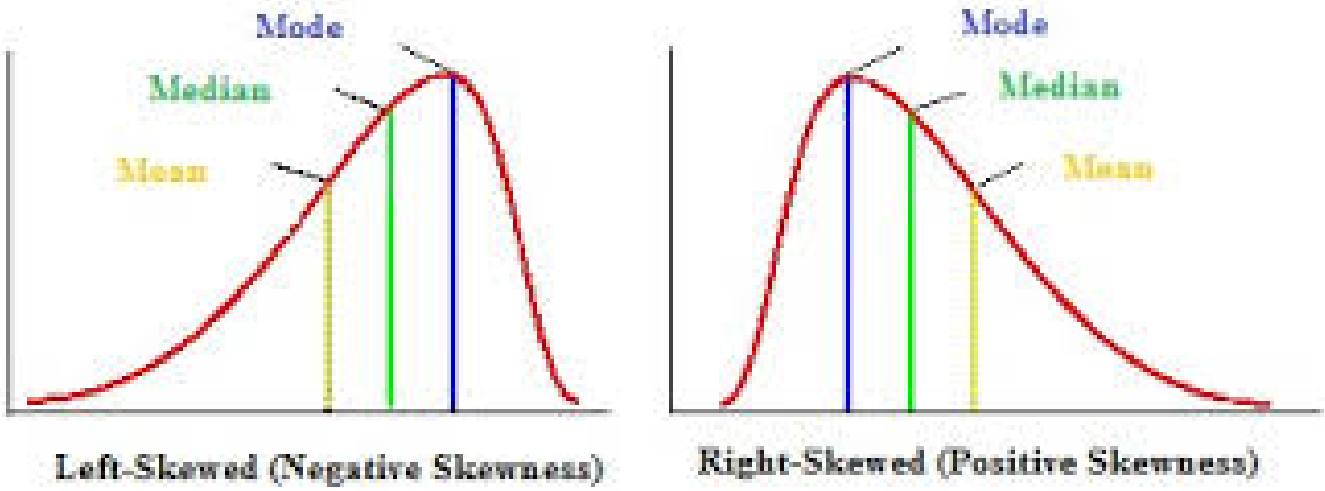
where j is the smallest integer such that $\max |x'| < 1$

Creation of relevant indicators from the raw data

- Replacing absolute values with ratios, often the most relevant ones:
  - e.g. by calculating the changes of variables over time (for example the mean for the recent period, divided by the mean for the previous period)
- Making linear combinations of variables

Creation of relevant indicators from the raw data

- Composing variables with other functions (such as logarithms or square roots of continuous variables):
  - to smooth their distribution and compress a highly right-skewed distribution (commonly found in financial or reliability studies)



Outline

Big data preparation

- Big data is often very "raw": considerable preprocessing is required before it can be considered in an analysis
- Volume poses a challenge to data cleaning:
  - manual or interactive editing of a dataset are usually preferred for reasonably sized data
  - it quickly becomes impractical when the data grows
- In big data applications, data preparation is largely an automated task

Big data preparation

- Algorithms can speed things up:
  - By examining data fields, automatically filling in blank values
  - Renaming certain fields to ensure consistency when data files are being joined
  - Doing format conversions as needed
  - Data preparation software may create a new field in the data file that aggregates counts from preexisting fields
  - Apply a statistical formula – such as a linear or logistic regression model – to the data
  - After going through, data is output into a finalized file that can be loaded into a database or other data store, where it is available to be analyzed.

Big data preparation

- Even though data preparation has become highly automated, it can still take up significant amount of time
- Data scientists spend a majority of their time locating and cleansing data rather than actually analyzing it
- There has been an increase in the number of software vendors attempting to tackle the data preparation problem
- Many organizations are putting more resources toward automating the process of preparing data
- The data preparation phase requires relatively extensive functionality, not provided in all softwares.

Software: Data preparation functions

- The following functions should be presented:
  - file handling (merging, aggregation, transposition, etc.)
  - data display, coloring of individuals according to a criterion;
  - detection, filtering and Winsorization of outliers;
  - analysis and imputation of missing values;
  - transformation of variables (recoding, standardization, automatic normalization, discretization, etc.);
  - creation of new variables;
  - selection of the best independent variables;

Summary

Every real world data set needs some kind of data pre-processing:

- Deal with missing values
- Correct erroneous values
- Select relevant attributes
- Adapt data set format to the software tool to be used

In general, data pre-processing consumes more than 60% of a data mining project effort

Key takeaways

- What you do with **missing values** depends on how many there are, and whether they're missing randomly or systematically
- When in doubt, assume that missing values are missing systematically
- Appropriate **data transformations** can make the data easier to understand and easier to model
- **Normalization and rescaling** are important when relative changes are more important than absolute ones.

References

- *Data Mining: Practical Machine Learning Tools and Techniques*, second edition, Ian H. Witten and Eibe Frank, (2005)
- *DM: Introduction: Machine Learning and Data Mining*, Gregory Piatetsky-Shapiro and Gary Parker
- *Data Mining and Statistics for Decision Making*, Stéphane Tuffréy, Wiley (2011)
- *Practical Data Science with R*, Nina Zumel, John Mount, Manning (2014)