Data Pre-processing: Clean, Reduce and Transform

Anran Zhao

Research Associate at UK Data Service

Email: anran.zhao@manchester.ac.uk

Table of Content

- Definition, context, and collecting data
- Data integration (join tables)
- **Data cleaning (missing values, outliers, data types)**
- **Data reduction (correlation check, PCA, sampling)**
- **Data transformation (normalisation, one-hot encoding)**

Recap

- **Data pre-processing (a.k.a. data preparation)** is the process of manipulating or pre-processing raw data from one or more sources into a structured and clean data set for analysis. It is an important part of Data Analytics.

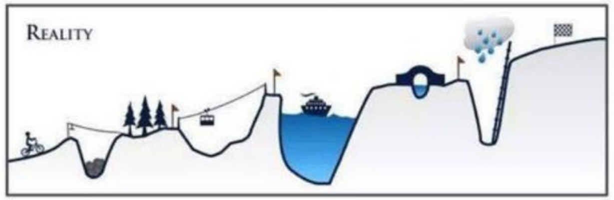


Table of Content

- Definition, context, and collecting data
- Data integration (join tables)
- **Data cleaning (missing values, outliers, data types)**
- Data reduction (correlation check, PCA, sampling)
- Data transformation (normalisation, one-hot encoding)

Data Cleaning - Quality Issues

- Data in the real world is dirty:
  - Incomplete or missing: lacking attribute values or certain attributes of interest, or containing only aggregate data,

    e.g., occupation=“ ” (missing data), Jan. 1 as everyone’s birthday? (disguised missing data)
  - Inaccurate or noisy: containing errors or outliers,

    e.g., salary=“-10” (an error)
  - Inconsistent: containing discrepancies in codes or names,

    e.g., age = “42” and birthday=“03/07/1997”

Dirty Data – Example

3. Inconsistency

1. Missing Values

Days On Market	Chain	House No.	Street	City	On Market Date	PostCode	Price
319	FALSE	40	Main Road	Manchester	08/03/2019	M19 2PE	£104,000
411	TRUE	198	Main Road	Edinburgh	08/02/2018	M19 2PF	£111,000
191	TRUE	58	Grange Road	Manchester	26/05/2018	M19 7YC	£96,000
247	TRUE	32	Green Lane	Manchester	20/02/2019	M19 3EN	
149	FALSE	35	The Drive	Manchester	29/04/2018	M19 9GI	£167,000
316	TRUE	147	Stanley Road	Manchester	04/02/2019	M19 2KB	£120,000
399	FALSE	19	Mill Lane	Manchester	26/05/2018		NULL
422	Unknown	145	Main Road	Manchester	16/07/2018	M19 3EC	POA
339	FALSE	194	The Grove	Manchester	08/06/2019	M19 5KH	£200,000
220	TRUE	175	The Green	Manchester	09/05/2018	M19 6AH	£155,000
116	TRUE	145	Grange Road	Manchester	26/05/2018	M19 3PF	£90,000
339	FALSE	194	The Grove	Manchester	08/06/2019	M88 5KH	£205,000
238	FALSE	61	Mill Road	Manchester	20/02/2019	M19 3RD	£197,000

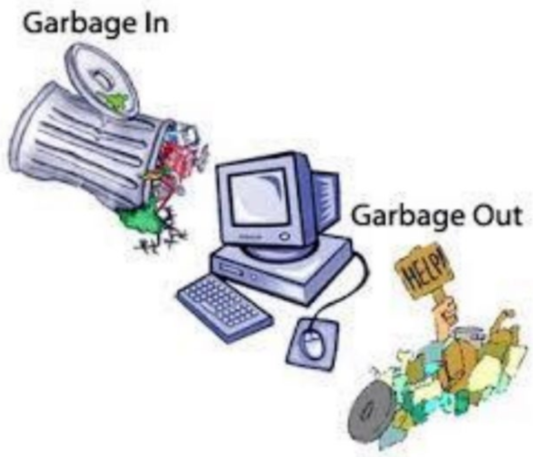
5. Duplicate records?

2. Date data may not in desired format

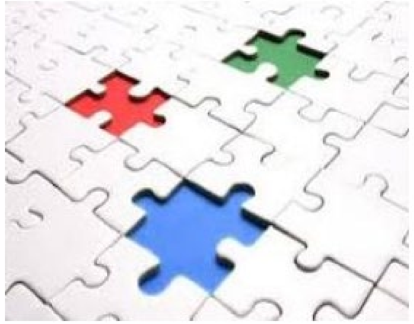
4. Incorrect (invalid) postcode?

Why Data Cleaning?

- “Data cleaning is one of the three biggest problems in data warehousing”— Ralph Kimball
- “Data cleaning is the number one problem in data warehousing”— DCI survey
- Quality data beats fancy data mining algorithms



Incomplete (Missing) Data



- Data is not always available
  - E.g., many rows have no recorded value for several attributes, such as customer income in sales data

- Missing data may be due to
  - Equipment malfunction
  - Inconsistent 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
  - No recorded history or changes of the data

No Easy Fix for Missing Values

Throw out the records with missing values?
- No? This creates a bias for the sample

Delete the column with missing values?
- No? Only if the column data is unnecessary

Replace missing values with a “special” value (e.g., -99)?
- No. This resembles any other value to data analytics.

Replace with some “typical” value? mean, median, or mode?
- Maybe. Possible changes to the distribution.

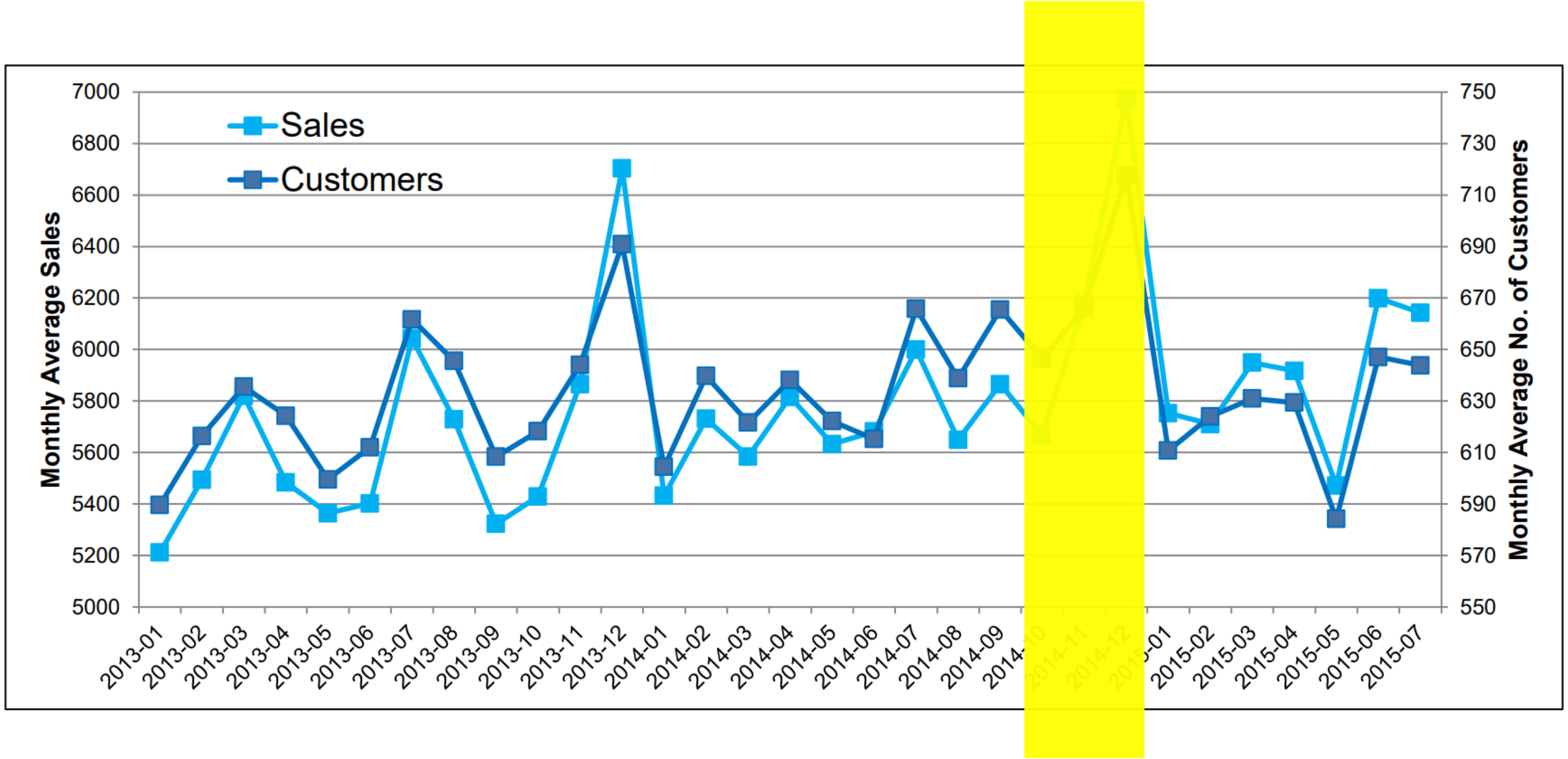
Impute a value? (Imputed values should be flagged.)
- Maybe. Use distribution of values to randomly choose a value.

Use data mining techniques that can handle missing values?
- Yes. For example, decision tree can be applicable.

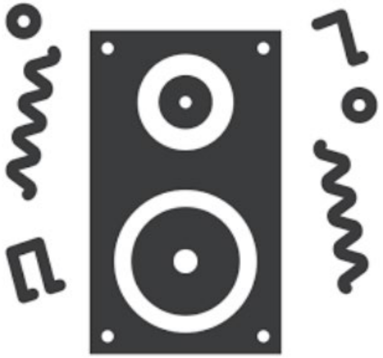
Partition records and build multiple models?
- Yes. This is possible when data isn’t insufficient.

No Easy Fix – Time Series Data

- How to find and impute these missing data?



Inaccurate (Noisy) Data



- Noise: random error or variance in a measured variable

- Incorrect attribute values may be due to
  - Faulty data collection instruments
  - Data entry problems
  - Data transmission problems
  - Technology limitation
  - Inconsistency in naming convention

How to Handle Noisy Data?

- Binning and smoothing
  - Sort data and partition into bins (equal-width, equal-depth)
  - Smooth by bin means, median, or boundaries, etc.
- Regression
  - Smooth by fitting the data into a function with regression
- Clustering
  - Detect and remove outliers that fall outside clusters
- Combined computer and human inspection
  - Detect suspicious values and check by human (e.g., deal with possible outliers)

Binning Methods for Data Smoothing

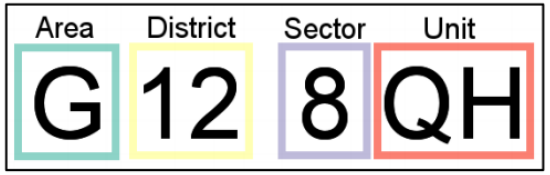
- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
  - Partition into 3 frequency (equal-depth) bins:

  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
  - Smoothing by bin means:

  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
  - Smoothing by bin boundaries:

  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Other relevant concepts



- Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
- Data auditing: analyse data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data validating: value range checks, regular expressions, uniqueness checks

Table of Content

- Definition, context, and collecting data
- Data integration (join tables)
- Data cleaning (missing values, outliers, data types)
- **Data reduction (correlation check, PCA, sampling)**
- Data transformation (normalisation, one-hot encoding)

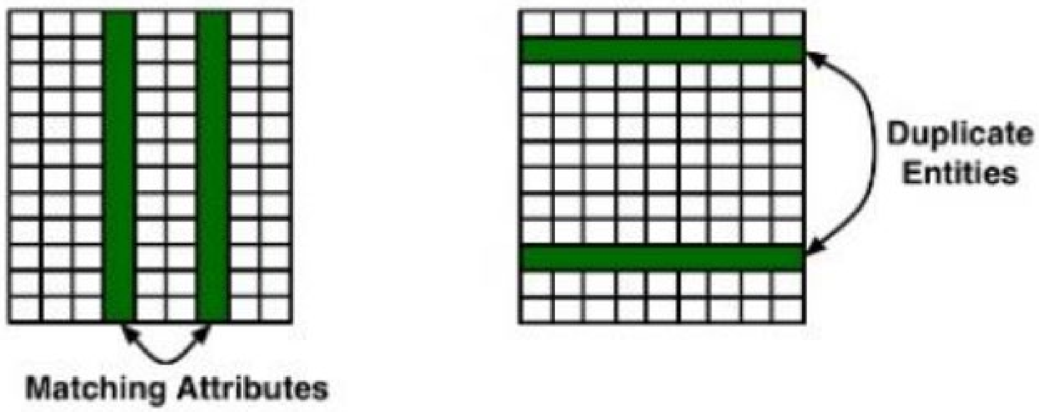
Data Reduction

- **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

- **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 is often created when integrating multiple databases
  - Column-oriented: the same attribute may have different names in different databases
  - Row-oriented: duplicate entities, etc.

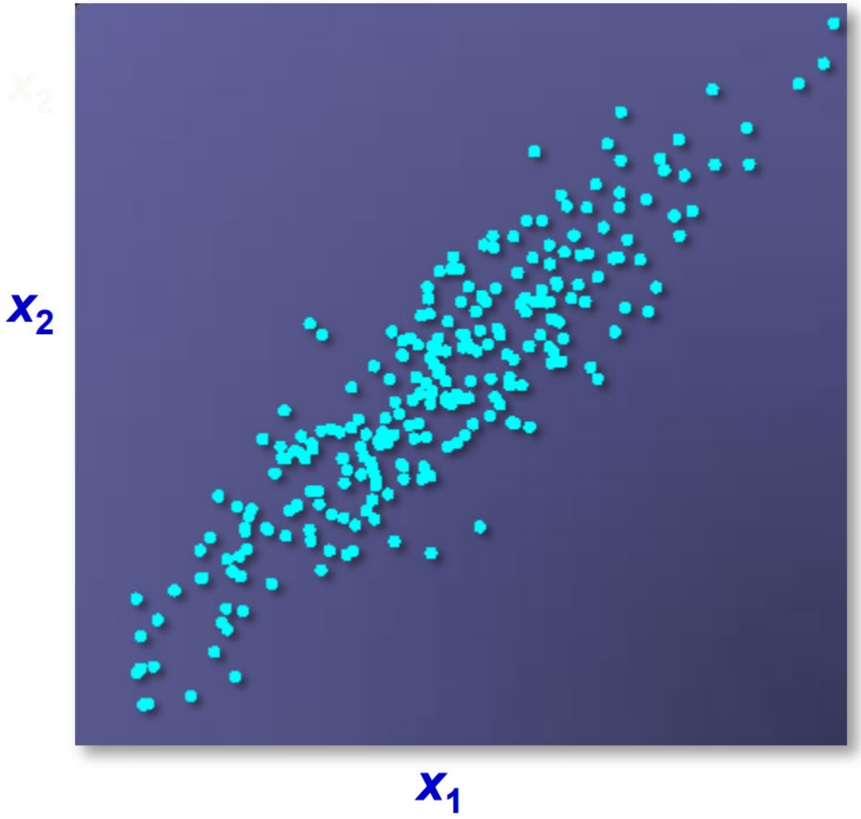


Data Reduction Strategies

- **Dimensionality reduction**
  - Remove redundant and irrelevant attributes
  - Principal component analysis (PCA)
  - Variable clustering
  - Featuring engineering

- **Numerosity reduction**
  - Sampling techniques
  - Regression and log-linear models
  - Histograms, clustering

Variable Reduction – Correlation analysis



Redundancy:

Input x_2 has the same information as input x_1.

Correlation Analysis – Numerical Variables

- **Correlation** between two variables *x1* and *x2* is the standard covariance, obtained by normalising the covariance with the standard deviation of each variable.

- **Sample correlation** for two attributes *x1* and *x2*: where *n* is the number of samples, *μ1* and *μ2* are the respective means, *σ1* and *σ2* are the respective standard deviation of *x1* and *x2*

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

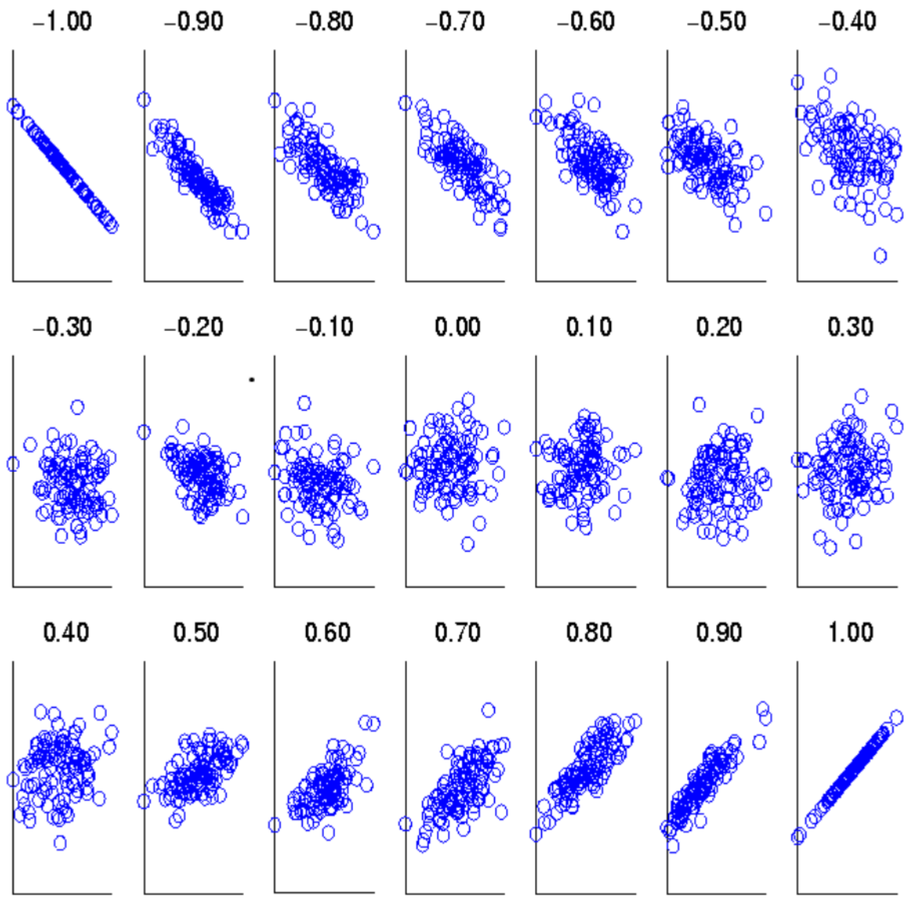
Correlation Analysis – Numerical Variables

- **Sample correlation** for two attributes *x1* and *x2*: where *n* is the number of tuples, *μ1* and *μ2* are the respective means, *σ1* and *σ2* are the respective standard deviation of *x1* and *x2*
  - If ρ12 > 0: x1 and x2 are positively correlated (x1 's values increase as x2 's increase)
  - If ρ12 = 0: independent
  - If ρ12 < 0: negatively correlated

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

Visualising Correlation Coefficients

- Correlation coefficient value range: [–1, 1]



Correlation Analysis

- Methods for testing 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

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.

Numerosity Reduction

- **Non-parametric methods**
- Do not assume models
- E.g. Sampling, clustering, histograms, etc.

- **Parametric methods**
- Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data
- E.g. regression, log-linear models

Sampling

- Sampling: obtaining a small set of samples to represent the whole data set
  - Simple random sampling
  - Sampling without replacement
  - Sampling with replacement
  - Stratified sampling

Table of Content

- Definition, context, and collecting data
- Data integration (join tables)
- Data cleaning (missing values, outliers, data types)
- Data reduction (correlation check, PCA, sampling)
- **Data transformation (normalisation and one-hot encoding)**

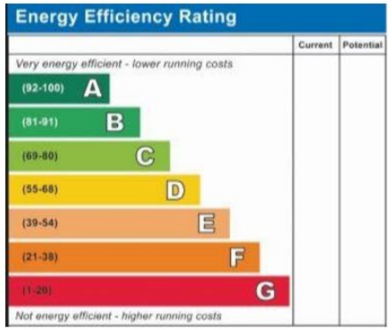
Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values, s.t., each old value can be identified with one of the new values

- Relevant methods:
- Normalisation/ Standardisation: scale data to fall within a smaller, specified range
  - » min-max normalisation
  - » z-score normalisation
  - » normalisation by decimal scaling

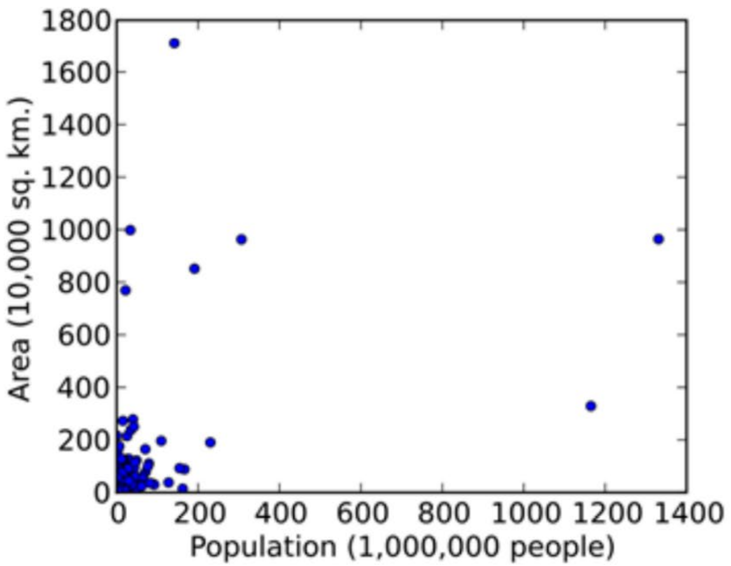
Data Transformation Examples

- Standardise numeric values
- Change counts into percentages.
- Translate dates to durations.
- Capture trends with ratios, differences, etc.
- Replace categorical values with appropriate numeric values



Data Transformation – Examples cont.

- Transform variables to bring information to the surface.



- Transform using mathematical functions, such as logs, reciprocal, or square root, for “stretching” and “squishing”

One-hot Encoding

- Use binary variables to replace a categorical feature.

Human-Readable

Pet
Cat
Dog
Turtle
Fish
Cat

Machine-Readable

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
1	0	0	0

Min-Max Normalisation

- min-max normalisation

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

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

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

Z-score Normalisation

- 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.

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

Normalisation by Decimal Scaling

- Normalisation by decimal scaling

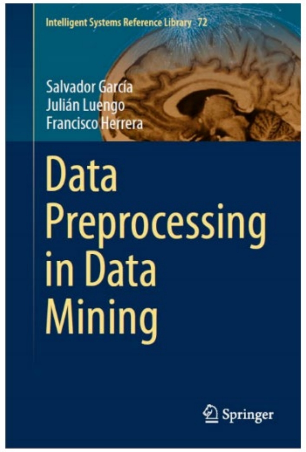
$$v' = \frac{v}{10^j}$$

- where *j* is the smallest integer such that Max(| *v'*|) <1

- Example — recorded values from -722 to 821
- Divide each value by 1000
  - -28 normalised to -.028
  - 444 normalised to 0.444

Acknowledgement

- Some of the content is based on …



García, S., Luengo, J. and Herrera, F., 2015. Data preprocessing in data mining New York: Springer.

Yu-wang Chen. “Understanding Data and Their Environment- Data Preprocessing” (2019)

You might be interested in...

Upcoming events:

- **Online workshop: Data Pre-processing Methods in Python**, on 1pm Jan28
- **Online workshop: Techniques and Methods of Analysis for Social Network Data**, on 2pm Jan 27
- **UK Data Service Computational Social Science Drop-in**, on 1pm Feb 9

- Recent events:
- Text-mining series
- Social Network Analysis series
- Data in the spotlight: UK and cross-national surveys