Data Warehousing and Data Mining

Lecture 3
Data preprocessing

Outline

- What is ETL?
- Preprocessing data
- Introduction to SQL
- Application of SQL for simple data transformation and analytics

What is ETL?

- ETL stands extraction, transform and load
  - Extraction - retrieve raw data
  - Transform - manipulate data into our desired form
  - Load - push the data into the warehouse
- Describes the process of doing the above steps, manually or through tools

Why preprocess data?

- Raw data may contain noise, junk, null values
- Data may need to be integrated together from multiple sources
- Data may need to be normalised
- Curse of dimensionality

Data cleaning

- To process the data such that it ensures
  - Validity - conforms to **business rules** or constraints
  - Accuracy/Precision - conforms to a **known "true" value** (the data is real)
  - Consistency - measures are **equivalent across systems**
  - Completeness - **all required measures** are known
  - Uniformity - using the **same units** across all system

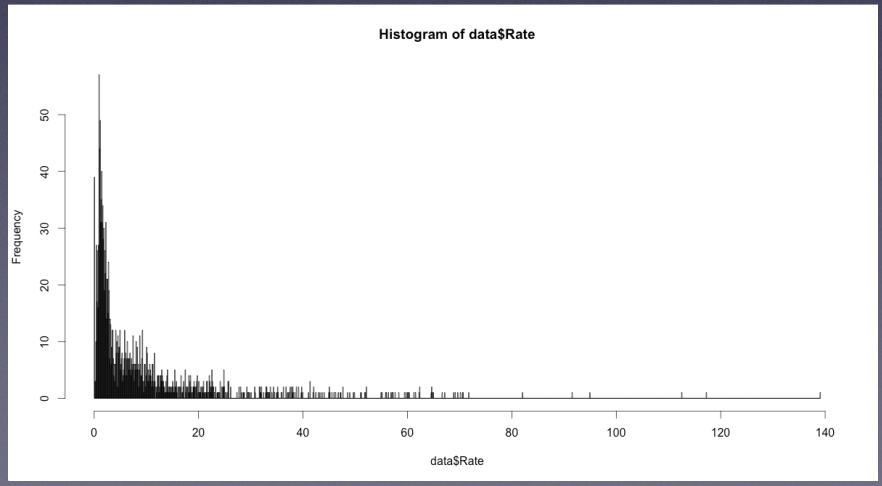
How to start cleaning data

- **Overview** the nature of your data
- Define and **determine error types**
- Search and **identify** error instances
- **Fix** the uncovered errors

Overview your data

- It is important to have big picture of your data
- Some metrics to get you acquainted with your data
  - Summary statistics
  - Simple plots

```
> summary(data)
          Country.or.Area       Year          Count             Rate        
 Dominican Republic:  17   Min.   :1995   Min.   :    0   Min.   :  0.000  
 El Salvador       :  17   1st Qu.:2000   1st Qu.:   46   1st Qu.:  1.600  
 Guatemala         :  17   Median :2004   Median :  240   Median :  4.200  
 Israel            :  17   Mean   :2003   Mean   : 2183   Mean   :  9.497  
 New Zealand       :  17   3rd Qu.:2007   3rd Qu.:  879   3rd Qu.: 11.100  
 Singapore         :  17   Max.   :2011   Max.   :45559   Max.   :139.100  
 (Other)           :1617                                                   
                Source     Source.Type
 CTS                :403   CJ:1605    
 National police    :219   PH: 114    
 CTS/Eurostat       :202              
 NSO                :153              
 CTS/Transmonee     : 93              
 CTS/National police: 77              
 (Other)            :572              
> 
```



Fixing data : missing values

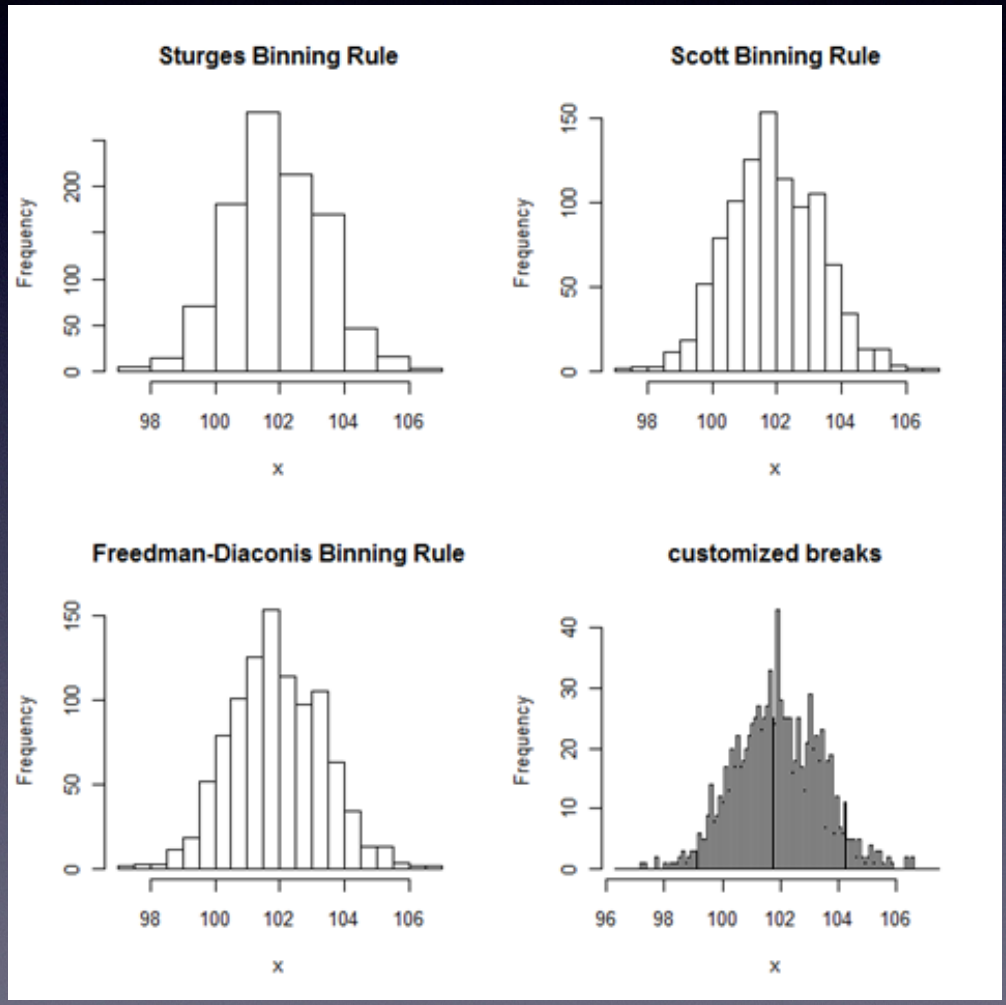
- **Ignore** it
- Fill in **manually**
- Fill in using a **fixed value**
- Fill in using the **mean value**
- Fill in using values from **similar data points**
- Fill in with the **most probable** values

Fixing data : noise

- Noise is random error or variance in the data
- Can be fixed by
  - Binning
  - Regression
  - Clustering

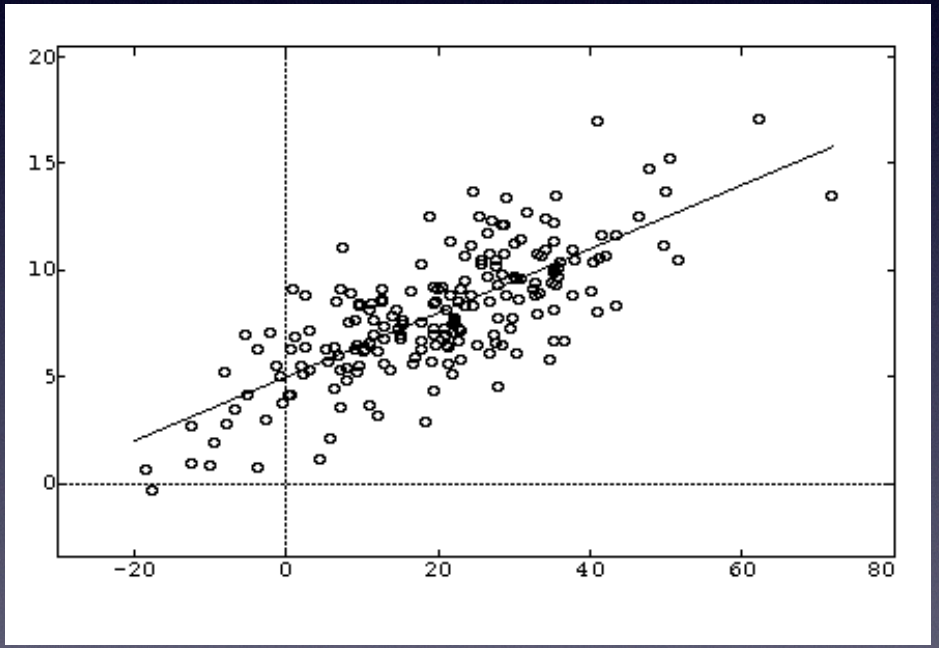
Data binning

- Grouping data points in a ‘bin’ (interval) together
- Replace all the values in the ‘bin’
- Aims to reduce noise



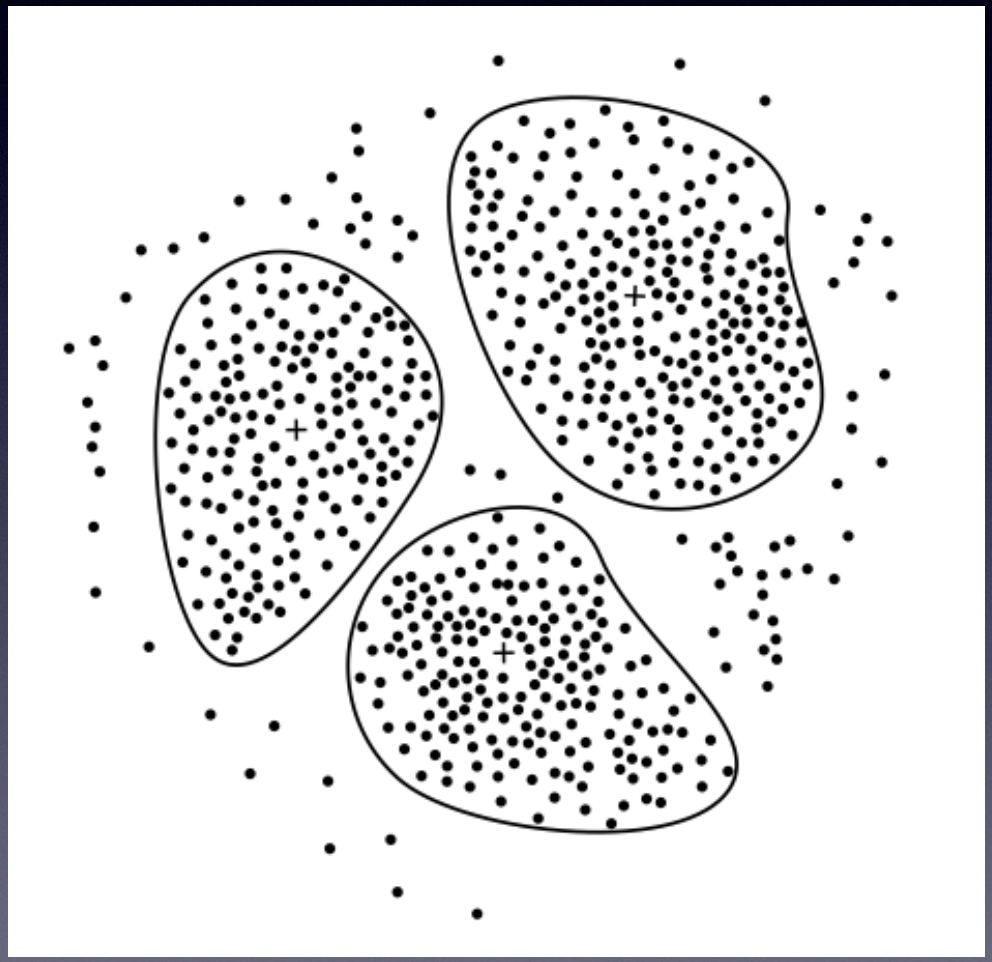
Regression

- Fit a function to the data points
- Replace data points with fitted values



Clustering

- Groups can be discovered through clustering
- Noisy data are more likely to be thrown out of the cluster (outlier)
- These may be removed



Does noise need to be removed?

- As you may have noticed, a lot of the proposed methods use the same techniques as actual knowledge discovery
- Modern techniques will quantify noise for you

Advanced techniques in data cleansing

- Statistical - used to detect anomalies
- Pattern-based - ensures consistency
- Association rule - ensure basic business rules

Tools to help clean data

- Data scrubbing tools - helps by cross checking against “standard” values
- Data audit tools - helps by allowing users to define business logic and run it through data
- ETL GUI frontends - visualised user interface for running data through the ETL process

Data integration

- Combine data from multiple sources together
- Things to be careful of
  - **Entity identification problem** - how can you be sure two fields from different files **mean the same thing**?
  - **Redundancy** - it is not useful to have data that can be **represented by another column** (near perfect correlation)

Transformation

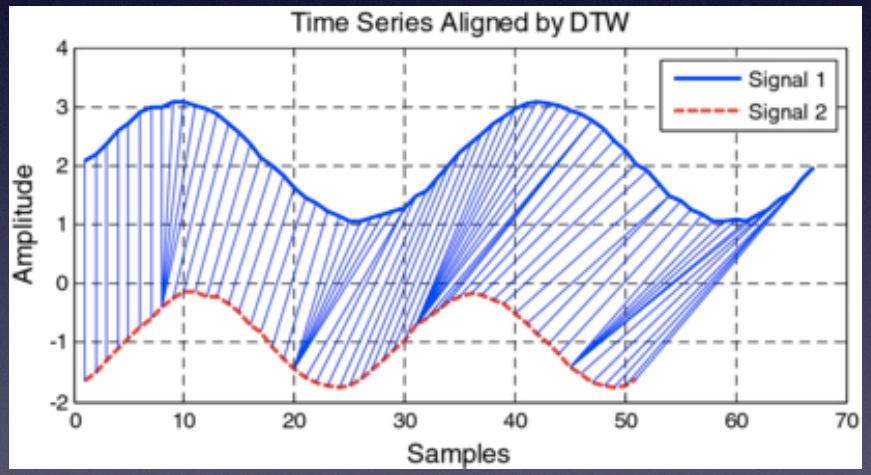
- Convert the data we have into appropriate forms
- Some conversion methods are
  - Smoothing (Data binning, regression, clustering)
  - Aggregation (Count, sum)
  - Generalisation (Move up the concept hierarchy)
  - Normalisation (Scale to fall into a range)
  - Attribute construction (New features are created)

Data reduction

- Curse of dimensionality!
- Methods of data reduction may include
  - Aggregation
  - Attribute subset selection
  - Dimensionality reduction
  - Concept hierarchy generation
  - Numerosity reduction

Numerosity Reduction

- Reduce dimensionality by reducing many attributes into function parameters
- One clear example is the use of Dynamic Time Warping (DTW) to reduce many time series



Data discretisation

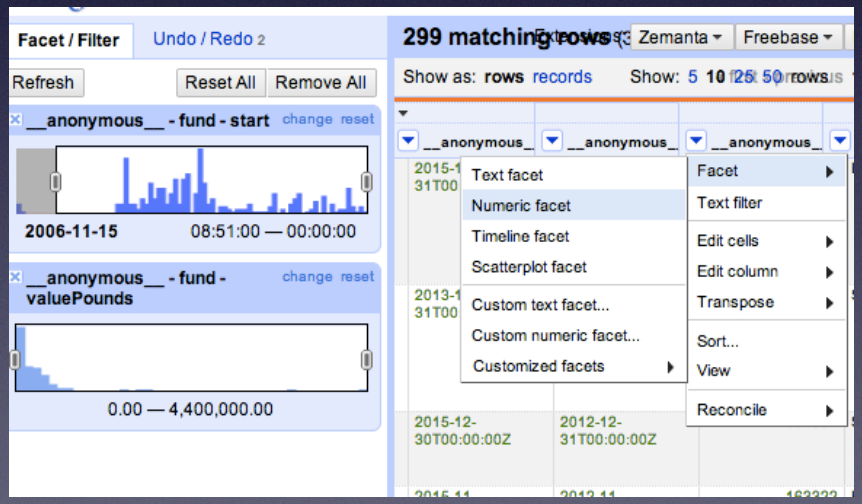
- Convert continuous data into discrete
- Methods include
  - Data binning
  - Generating concept hierarchy
  - Other methods

How to apply all this?

1. Get your dataset
2. Understand your dataset
3. Open it in your favourite editor, ETL or tool for cleaning data
4. Select the features you want
5. Perform any transformation you may need immediately
6. Push it into a database or data warehouse, ready for analysis

Notable tools

- Previously called “Google Refine”
- Platform for data cleansing
- Memory efficient
- Can handle large datasets



What is SQL

- SQL stands for “Structured Query Language”
- Standard language for accessing databases
- Variations exist, but main functions are standard

What is SQLite

- A vendor of a very simple “database” system
- Designed to replace “fopen()”
- Uses flat file to store data
- We will use this to practice our SQL querying skills!

Data types in SQL

- Dependent for each database vendor
- The following is for SQLite

- **NULL**. The value is a NULL value.
- **INTEGER**. The value is a signed integer, stored in 1, 2, 3, 4, 6, or 8 bytes depending on the magnitude of the value.
- **REAL**. The value is a floating point value, stored as an 8-byte IEEE floating point number.
- **TEXT**. The value is a text string, stored using the database encoding (UTF-8, UTF-16BE or UTF-16LE).
- **BLOB**. The value is a blob of data, stored exactly as it was input.

Structure of SQL

- Contains a SELECT clause
- Contains a FROM clause
- Contains constraints (WHERE,LIMIT), joins (JOIN), aggregations (GROUP_BY) and others (ORDER)

Common SQL syntax and functions

- The following are essential to get started
  - SELECT
  - FROM
  - WHERE
  - GROUP_BY
  - LIMIT
  - JOIN…ON
  - Aggregation functions (COUNT, SUM, etc.)

Example: UN Data on crime rates and education

Country or A	Year	Count	Rate	Source	Source Type
Afghanistan	2008	712	2.4	WHO	PH
Albania	2010	127	4	CTS/Transmc	CJ
Albania	2009	85	2.7	CTS/Transmc	CJ
Albania	2008	93	2.9	CTS/Transmc	CJ
Albania	2007	105	3.3	CTS/Transmc	CJ
Albania	2006	87	2.8	CTS/Transmc	CJ
Albania	2005	131	4.2	CTS/Transmc	CJ
Albania	2004	119	3.8	CTS/Transmc	CJ
Albania	2003	144	4.6	CTS/Transmc	CJ
Albania	2002	231	7.5	CTS/Transmc	CJ
Albania	2001	208	6.8	CTS/Transmc	CJ
Albania	2000	275	9	CTS/Transmc	CJ
Albania	1999	496	16.1	CTS/Transmc	CJ
Albania	1998	573	18.6	CTS/Transmc	CJ
Albania	1997	1542	49.9	CTS/Transmc	CJ
Albania	1996	248	8	CTS/Transmc	CJ
Albania	1995	210	6.7	CTS/Transmc	CJ
Algeria	2008	516	1.5	CTS	CJ
Algeria	2007	438	1.3	CTS	CJ
Andorra	2004	1	1.3	Interpol	CJ
Angola	2008	3426	19	WHO	PH
Anguilla	2008	1	6.8	NSO	CJ
Anguilla	2007	4	27.8	NSO	CJ
Anguilla	2006	4	28.6	NSO	CJ
Anguilla	2005	1	7.4	NSO	CJ
Anguilla	2004	1	7.6	NSO	CJ

Reference Ar	Time Period	Sex	Age group	Units of measurement	Observation Value
Afghanistan	2012	All genders	Not applicable	Number	867551
Afghanistan	2012	Female	Not applicable	Number	424123
Afghanistan	2012	Male	Not applicable	Number	443428
Afghanistan	2011	Male	Not applicable	Number	431278
Afghanistan	2011	All genders	Not applicable	Number	844028
Afghanistan	2011	Female	Not applicable	Number	412750
Afghanistan	2010	Female	Not applicable	Number	400150
Afghanistan	2010	Male	Not applicable	Number	417844
Afghanistan	2010	All genders	Not applicable	Number	817994
Afghanistan	2009	All genders	Not applicable	Number	782319
Afghanistan	2009	Male	Not applicable	Number	399593
Afghanistan	2009	Female	Not applicable	Number	382726
Afghanistan	2008	Female	Not applicable	Number	364876
Afghanistan	2008	All genders	Not applicable	Number	745702
Afghanistan	2008	Male	Not applicable	Number	380826
Afghanistan	2007	Male	Not applicable	Number	362980
Afghanistan	2007	All genders	Not applicable	Number	711133
Afghanistan	2007	Female	Not applicable	Number	348153
Afghanistan	2006	Male	Not applicable	Number	345132
Afghanistan	2006	Female	Not applicable	Number	331787
Afghanistan	2006	All genders	Not applicable	Number	676919
Afghanistan	2005	Female	Not applicable	Number	317457
Afghanistan	2005	Male	Not applicable	Number	329163
Afghanistan	2005	All genders	Not applicable	Number	646620
Afghanistan	2004	All genders	Not applicable	Number	620239
Afghanistan	2004	Female	Not applicable	Number	304453
Afghanistan	2004	Male	Not applicable	Number	315786
Afghanistan	2003	Male	Not applicable	Number	302841
Afghanistan	2003	Female	Not applicable	Number	291868
Afghanistan	2003	All genders	Not applicable	Number	594709
Afghanistan	2002	All genders	Not applicable	Number	568671

Live demonstration

Quick references

- To import data from a simple csv file
  .mode list
  .separator ,
  .import <filename> <tablename>
- .schema - show schema
- .tables - list tables
- .quit - exit the shell

Assignment 1

1. Clean and prepare the file “Marital status of men and women.xls”, and document your actions.
2. Create an SQLite database with your data and the CrimeRates data (you can use tools, but use wisely!)
3. Using SQL, query the data that may tell you
   - **The effects of crime rates on the percentage of married men and women aged 20-24 and 25-29**