Data Collection and Modeling

Course 7 - Data Cleaning

Definition

“Data cleansing or data cleaning is the process of identifying and correcting corrupt, incomplete, duplicated, incorrect, and irrelevant data from a reference set, table, or database.

Data issues typically arise through user entry errors, incomplete data capture, non-standard formats, and data integration issues.” (Precisly)

“Data cleansing, also referred to as data cleaning or data scrubbing, is the process of fixing incorrect, incomplete, duplicate or otherwise erroneous data in a data set. It involves identifying data errors and then changing, updating or removing data to correct them. Data cleansing improves data quality and helps provide more accurate, consistent and reliable information for decision-making in an organization.” (Techtarget)

Motivation

Results of data processing are as good as the data they process.

Studies suggest that 29% of data is inaccurate and leads to loss of revenue and customers. Data cleaning improves accuracy and cand make data more structured and consistent.

- Organized data
- Improved productivity
- Improved mapping
- Avoid mistakes
- Reduce costs

Error Types

There may be various errors in a dataset: invalid, incompatible, inaccurate and corrupt data. The source can be human error, differences in representation, terminology or format.

- Duplicated data
- Typos, invalid, missing data
- Inconsistent data
- Irrelevant data

Overall Process

In a typical scenario, there are at least the following steps:

1. Profiling and inspection: statistics on data set to help identify missing values, erroneous ones, discrepancies
2. Data cleaning: fixing data issues
3. Verification: checks according to quality rules and necessary data standards
4. Reporting: results should be continuously reported with respect to numbers of issues found/fixed and generate data quality trends/visualizations

Data Profiling

Is the process of analyzing data and producing summaries to help identify quality issues and trends.

Typical methodology includes determination of mean, min, max, percentile and frequency distribution; it also includes detection of primary key candidates, functional dependencies and possible foreign keys.

Helps to identify eros like null values (the representation of missing or unknown values), values of of range/scale, values with questionable frequencies.

- Column profiling
- Cross-column profiling
- Cross-table profiling

Data Profiling

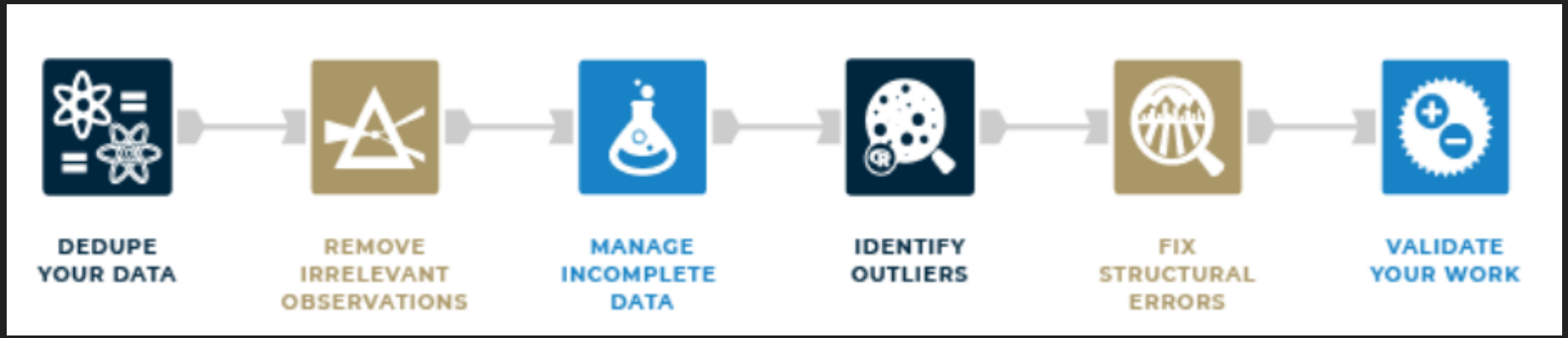
Can be:

- Structure discovery: checks for consistency and format (pattern matching)
- Content discovery: relates to data quality; data is processed to adapt to formatting and standardization to be integrated with existing data
- Relationship discovery: detect connections between datasets

Helps by producing:

- Data quality and credibility
- Proactive crisis management
- Predictive decision making
- Organized sorting

Cleaning Process



source: Alteryx

Duplicate removal

- Usually generated during data collection
- When data is combined from multiple sources
- One of the largest area in the cleaning process
- Approach depends on the situation
  - Deal with entire duplicates
  - Deal with partial duplicates

Irrelevant Observations Removal

- Identify relevant data; might need to check correlated values later in the analysis
- Exclude data from analysis, not from source
- Try to use 4 eyes principle when in doubt
- Remove elements like: Personal Identifiable data, URLs, blank spaces, HTML tags, tracking codes
- Dataset becomes more manageable and relevant

Dealing with Incomplete Data

- Can be found as NULL values, “not applicable”, “NA”, “0”, “none”
- Determine if they are plausible or generated by missing information
- Possible approaches:
  - Remove entries associated with it: can induce bias
  - Guess values based on similar data: linear regression, median, etc
  - Flag data as missing: can become information

Filter Outliers

- Data points that are extremely different from the rest in the dataset
- Some can represent true values (natural variation in the population), other may be produced by incorrect data entry or malfunctions/errors
- Can affect the result of analysis
- True outliers should be retained as they represent natural variations
- Can be identified by:
  - numeric techniques: z-scores
  - visual techniques: plots, histograms
  - sorting method
  - Interquartile range (the range of the middle half of the dataset)

Structural Errors

- Typos, capitalization, abbreviations, formatting
- Dependent on data type
- Can be handled using spell-checking, formatting (padding or trimming, adopting a consistent capitalization), using same measurement unit
- Can refer to columns (names) or values in columns

Data Validation

- Use to authenticate data, confirm quality, consistency and (formatting) uniformity
- Is data enough?
- Is uniformly formatted?
- Does data make sense?
- Use a sample to validate or invalidate working theory

Example

#	Id	Name	Birthday	Gender	IsTeacher?	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Missing values

Invalid values

Misfielded values

Misspellings

Uniqueness

Formats

Attribute dependencies

source: Quantdare

Data Quality Report

Based on data exploration and checks performed, a quality report can be produced by answering several questions:

- Are there missing attributes or values?
- Are there deviations? Do they represent outliers or not?
- Are there spelling inconsistencies?
- Was a plausibility check for values performed?
- Was there irrelevant data to be excluded?

Data Cleaning Report

Steps to be performed for data cleaning should be documented for future iterations or future projects:

- What types of noise occurred in data
- What techniques/methods have been used to remove the noise
- Was data removed? Note excluded data

Characteristics of Cleaned Data

The quality and cleanliness of the data set can be measured with respect to:

- Accuracy
- Validity
- Consistency
- Completeness
- Integrity
- Uniformity
- Timeliness

Benefits

Even though data cleaning takes time and effort, it brings benefits both for short and long term:

- Better decision making
- Minimise compliance risk
- Boost results and revenue
- Save time and increase productivity
- Protect reputation

Tools/Vendors

- Open source: DataCleaner, OpenRefine
- Data preparation: Altair, DataRobot, Tableau
- Data management: Ataccama, Informatica, SAP, SAs, Talend
- Data cleaning: Data Ladder, WinPure
- Data quality: Datactics, Experian, Precisely