Data Quality

Dataforståelse 4

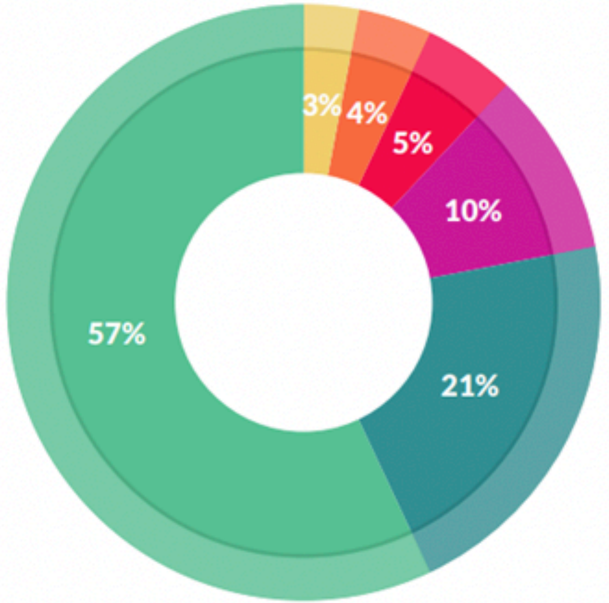
NIFR 2022

Dagsorden

Dataforståelse 4

- Data Quality Concepts
- Edit Data with MySQL
  - Update
- Identifying bad data

According to a survey conducted by Figure Eight in 2016, almost 60% of Data Scientists' time is spent on cleaning and organizing data. You can find the survey results at here.



Multi Dimensional Nature of Data Quality

Parameters

- Accuracy
- Completeness
- Time-Related Dimensions
  - Currency, Timeliness & Volatility
- Consistency

Accuracy

Data Quality Dimension

For single data values, accuracy measures the distance between a value *v* and a value *v'* which is considered correct. Two kinds of accuracy can be identified, namely a *syntactic accuracy* and a *semantic accuracy*.

Accuracy

Data Quality Dimension

the distance between a value *v* and a value *v'* which is considered correct

V: Britney Spars

V':Britney Spears

Accuracy

Data Quality Dimension

the distance between a value *v* and a value *v'* which is considered correct

V: Britney Spars

V': Britney Spears

V: BriasdtneySpars

Semantic Accuracy

Data Quality Dimension

Song	Artist	Album
We Will Rock You	Robin	Neonlys

Syntactic Accuracy

Data Quality Dimension

Song	Artist	Album
We Will Rock You	Qrueen	News of the world

Completeness can be generically defined as »the extent to which data are of sufficient breadth, depth and scope for the task at hand«

Completeness

Data Quality Dimension

- Column Completeness
  - Function of missing values in a column of a table
- Population Completeness
  - Missing Rows

Time Related Dimensions

Data Quality Dimension

- Stable Data
- Time Variable Data

ID	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead Poets Society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	NULL	1964	0	1985

Fig. 1: A relation *Movies* with data quality problems

Time Related Dimensions

Data Quality Dimension

- Currency
  - How often is data updated?
- Volatility
  - Frequency according to which data vary in time
- Timeliness
  - How current is data relative to usage?

Time Related Dimensions

Data Quality Dimension

- Stable Data
- Time Variable Data

ID	Title	Director	Year	#Remakes	LastRemakeYear
1	Casablanca	Weir	1942	3	1940
2	Dead Poets Society	Curtiz	1989	0	NULL
3	Rman Holiday	Wylder	1953	0	NULL
4	Sabrina	NULL	1964	0	1985

Fig. 1: A relation *Movies* with data quality problems

Consistency

Data Quality Dimension

The consistency dimension captures the violation of semantic rules defined over (a set of) data items.

Multiple violations of dimensions

Consistency & Completeness

Song	Artist	Album
We Will Rock you	Queen	NULL

UFO Exercise

Hvilke kvalitetskriterier overskrider datasettet?

Hvordan kan dette observeres?

SQL Update

```
UPDATE employees
SET
    email = 'mary'
WHERE
    employeeNumber = 1056;
```

Table name

```
UPDATE employees
SET
    email = 'mary'
WHERE
    employeeNumber = 1056;
```

Column name

```
UPDATE employees
SET
    email = 'mary'
WHERE
    employeeNumber = 1056;
```

Constraint

```
UPDATE employees
SET
    email = 'mary'
WHERE
    employeeNumber = 1056;
```

Identifying bad data

Data exploration

Correcting the values

Data cleaning

IS NULL & DISTINCT