Update 중

Data Quality Score matrix

05.July, 2019
Enzo

Table score

- Data Quality Score

  - Qi = Data quality issues identified for a column
  - Ti = Total row number
  - Ci = Number of columns in table
  - Pn = if the column required Y then have weight. The weight is propotion of completeness

$$QS = \frac{\left(1 - \sum_{i=1}^{n} \frac{Qi}{Ti}\right) * \left(1 - \sum_{i=1}^{n} \frac{Pi}{Ti}\right)}{N}$$

(출처: IBM knowledge center)

예시)

- Data Quality Score

Table 1. Data quality score with rules calculation

Column 1*	Column 1 quality dimensions	Column 2	Column 2 quality dimensions	Column 3	Column 3 quality dimensions
555-123-4567		17 King Way		555-123-4567	Duplicate value
555-456-1234		22 B Street		555-123-4568	
4567	Rule violation	45 H Lane		555-123-4569	
555-0236-8565		6 Villa Drive		555-123-4520	
555-236-8596		4 Parker Road		555-123-4567	Duplicate value
555-897-5632		NULL	Missing value	555-123-4530	
Renee	• Suspect value • Rule violation	NULL	Missing value	NULL	Missing value
NULL	• Missing value • Rule violation		Missing value	555-123-4545	
555-897-8523		555-123-4567	Suspect value	NULL	Mission value
3	Rule violation	09876	Suspect value	555-123-4555	

- **Column 1**:4/10 data quality issues identified for a column quality score of 60 percent.
- **Column 2**: 5/10 data quality issues identified for a column quality score of 50 percent.
- **Column 3**: 4/10 data quality issues identified for a column quality score of 60 percent.
- **Data set quality score:** (60+50+60) / 3 = 57 percent.

- Column 1: 4/10 data quality issues identified for a column score of 60%
- Column 2: 5/10 data quality issues identified for a column score of 50% and Penalty value 3/10
- Column 3: 4/10 data quality issues identified for a column score of 60%
- Data set quality score: (1- 4/10)+((1- 3/10)*(2/10))+(1- 4/10) / 3 = 0.45

(기본 수식 출처: IBM knowledge center)