Data profiling

Tom Rochette <tom.rochette@coreteks.org>

April 9, 2020 — f39c5c8

1 Question

What is data profiling?

2 Answer

Data profiling is the process of extracting information about data. Given tabular data (think of an Excel spreadsheet), we commonly want to extract the following properties about each column:

- Number of rows
- Number of cells without data
- Number of cells with a value of zero
- Number of distinct/unique values
- Number of duplicate rows
- Minimum, mean, median, maximum, quantiles, range, standard deviation, variance, sum
- Values distribution
- Most common values
- Examples of values

The process of data profiling allows a data scientist or engineer to identify quickly potential sources of problems in the data such as:

- Negative numbers when numbers should all be positive
- Missing values which may need to be imputed or for which the row may have to be removed
- Issues with the distribution of values such as class imbalance if we plan to solve a classification problem

In an ideal situation, data profiling reports:

- No missing cells, this way you do not have to ask if data can be filled in or you don't need to impute the data using assumptions
- Proper normalization of the data (e.g., value separate from their unit), this way the data can be used as-is, otherwise you need to transform the column to extract the numeric value from the unit
- All the data in a column using the same unit, unless otherwise specified (e.g., you do not want data in meters, centimeters, feet or inches in the same column), this way your data is consistent, otherwise you need to identify the scales/units used and transform the data to use a common unit
- Little to no row duplication, this way you know that your data was collected without creating duplicate entries, which sometimes happen when databases are merged manually to create a data file, otherwise you may have to drop the duplicate rows or identify how many of the duplicates should be kept