Print

# Practice: Checking Data Quality with Summary Statistics and Graphs

In this practice, you use the file **Messy Data 2.jmp** to identify data quality issues using summary statistics and graphs. This data table has seven variables: **acceptable?**, **lot number**, **unit number**, **location**, **humidity**, **voltage**, and **density**.

1. Use the Columns Viewer under the **Cols** menu and the **Distribution platform** to identify evidence of the following issues:
    - incorrect formatting
    - incomplete data,
    - missing data, and
    - dirty or messy data.
2. For each variable, what do you learn?

**acceptable?**: This variable is missing eight values.
**lot number**: This variable is coded as numeric/continuous data. It should be nominal.
**unit number**: There are six units, and the most frequently used are units 2 and 9.
**location**: There are nine locations and four missing values. This variable is messy. The names need to be cleaned up.
**humidity**: This variable is missing four values.
**voltage**: There are 88 missing values for voltage.
**density**: All but five of the values are the same (40).

Hide Solution

*Statistical Thinking for Industrial Problem Solving*

Close