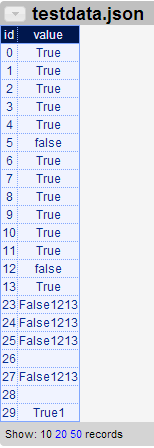
* **Objective:**

Initially Karma used to display all the raw data imported from the external sources such as JSON files or web service without processing any corrupt or inaccurate data. It does not identify the data type such as numbers or strings or date of any data imported. Therefore it cannot detect the inconsistency and distinguish individual record as valid data or invalid data. This further leads to absence of statistics and visualization of percentage of consistency in the imported data.

Figure 1 show the sample data imported which have most of the data is of type Boolean and few typos, inaccurate and missing data (from id 23-29).

This DR project is developed to overcome the problem discussed above as well as provide some useful statistics and metadata of imported data which will be used in future development. This project also provides bar chart display of accuracy of data for each of the column in the table.

* **Solution:**

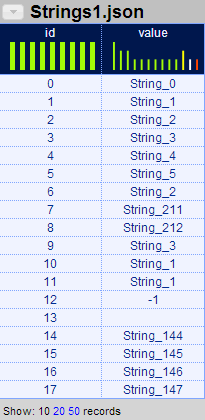
When the data is imported in Karma, each of the column data is passed in parallel to cleaning module. The cleaning module scans through the data to identify the majority of data type from Boolean, Date, Integer, Double, Weekday, String or Empty String in this sequence. The cleaning module also accepts (cluster) sampling parameters along with data to limit the number of rows to scan. After identifying data type, it calculates following parameters:- Valid/Invalid/ Missing data and their counts, data for drawing the D3 chart on UI in JSON form, Max token length (to find the longest word in the data), preferred length (recommended width length of the column to be used in styling in future) and average, max data length. Figure 2 shows a sample UI output after processing through the cleaning module.

Figure : Sample Data before cleaning module

Figure : Sample data after cleaning module

The part of output data represented stored in ‘histogram’ variable is also used to draw D3 charts on the Karma UI as shown in Figure 2 and Figure 3. Figure 3 appear when user clicks on smaller chart appeared below column heading.

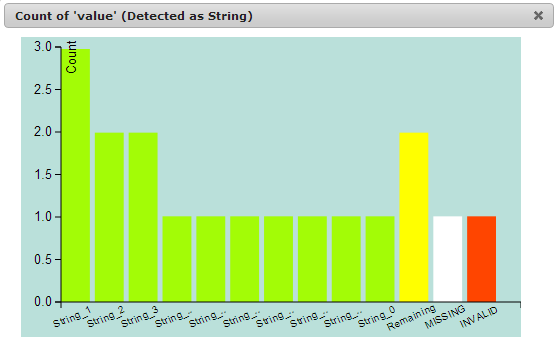
There are 2 types of methods are used to show the count of data on chart based on data types detected. First method is for String, Empty String, Boolean and Week days and Second method is for Numbers (Integer, Double) and Date. In first method count is displayed as a conventional frequency of occurrence of that particular data (ex. String or Weekday). In second method, the input range is divided multiple domains and count represents the number of input data observed in that domain. Due to this method, by looking at chart user can easily see any abnormality such as an unexpected numbers in the input. The Green bars represent the valid data of detected data type, orange bar represents invalid data (the data other than the detected data type), the White bar represents Missing /Empty data and the Yellow bar represents Remaining valid entries crunched into one single bar. 

Figure : Big Chart after clicking any smaller charts

* **Implementation:**

Cleaning module is implemented as an external rest web service in Karma which accepts JSON data via POST request and gives output as a JSON object string. Each row data is tried to parse into Boolean, Date, Integer, Double, Weekdays, String in sequence. After finding majority of data type in present in the input data, rest of the statistics are calculated such as Invalid, Missing data and their respective count, Max token length (Max word length), mean, Max total length. Preferred length for a particular column is calculated using *PERCENTILE* parameter (currently set to 80 percentile) on valid data. For String, it is 80 percentile of longest word in sentence, while for Numbers it is 80 percentile of longest digit numbers.

Upon data loading, multithreaded POST requests are sent to Cleaning web service and on receiving JSON output UI bar charts are drawn using client side processing technology such as D3 (Data driven Documents) JavaScript library, jQuery and CSS styling.

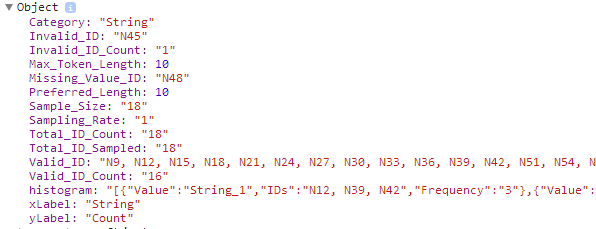


Figure : JSON Output from cleaning Web service