Digestable: Condensed Views of Tabular Data

David Borland and David Gotz

ABSTRACT

Tables are a common, flexible, and effective method for displaying data, and are easily understood by a wide audience. However, their effectiveness suffers with larger datasets. We introduce Digestable, a visualization tool that generates condensed views of tabular data to enable the identification of high-level features, while supporting detailed exploration. Digestable uses a familiar tabular organization of the data, and the interface for producing condensed views is based on a standard interaction technique for tables: sorting. The visualizations produced by Digestable enable users to obtain a holistic view of the dataset and support the identification of features such as outliers and relationships between variables. We present the Digestable interface, including the methods used to produce condensed views and the visualizations and interactions supporting data exploration.

Keywords: Tabular visualization, visual summaries, simplification

1 INTRODUCTION

Tables are a widespread and effective method for displaying data across a wide range of disciplines and media. The familiar organization and display of data using rows and columns is easy to understand, and enables the examination of data in detail. However, tables are typically less effective for large datasets. Tufte suggests that "Tables usually outperform graphics in reporting on small data sets of 20 numbers or less" [4]. Many real-world datasets vastly outscale such dimensions, leading to the development of a wide range of visualization techniques to aid in the summarization, comprehension, and investigation of large, multivariate datasets.

Despite the proliferation of such techniques, tables and interactive spreadsheets remain common, and various tools utilize tablular formats (e.g., [2, 3]). Recent work shows that they are not just for "data cleanup at the initial stage of a linear analytic flow," but an integral part of the analytic process [1]. In order to bridge the gap between the effectiveness and familiarity of tables in organizing data and the need for methods to help summarize and provide overviews of data, we developed Digestable, a visualization tool that produces condensed views of tabular data, and provides interactions to enable data exploration (Figure 1).

The contributions of Digestable include (1) an interface for generating condensed tables that summarize tabular data, (2) data-type dependent methods for summarization and visualization, and (3) interactions to facilitate (a) the comparison of different columns and identification of relationships between columns, and (b) the detailed exploration of individual data entities.

2 METHODS

The design of Digestable begins with a familiar tabular model of rows and columns to organize the data, with sorting and reordering interactions as the means for ranking and comparison. This familiar basis is then enhanced with additional interactions and data-type-dependent visual representations to achieve the goals of improved summarization and exploration within that traditional tabular context.

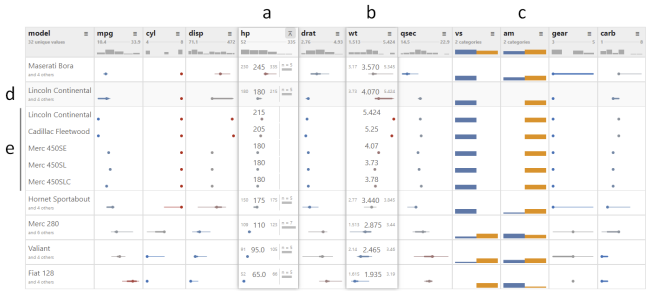

Figure 1: Example condensed table of automobile data in Digestable. A user-selected *active* column (a) is used for sorting and automatically grouping rows in the table. Different visual modes enable both text and visualization summaries of the distribution of values in each cell, and mouseover of any other column (b) enables a detailed comparison with the active column. Data-type dependent displays are used, including numeric (a, b) and categorical (c) columns. In addition, users can expand or collapse row groups (d) to inspect data in more detail (e).

The data are initially displayed as a standard table with each cell containing text or a number. Each column is automatically classified as type *ID*, *categorical*, or *numeric*, which affects both the simplification methods and visualizations available for that column. Distribution visualizations are included in the column headers. The user is then able to explore the data using simplification, different visual modes, and other interactive features.

2.1 Simplification

Simplification interactions enable the user to condense the rows of a table into a smaller number of *row groups*, based on the values from an individual column, enabling users to see an overview of the full dataset, and inspect for higher-level patterns across the columns.

To perform simplification the user selects a column to sort and group by, termed the *active* column. This column is first sorted in either ascending or descending order, then simplified by grouping adjacent rows based on the values in that column. The grouping process works differently based on the data type, and can be applied to either numeric or categorical columns.

2.1.1 Numeric Columns

Numeric columns are displayed using a combination of visual elements to provide users with both a view of the underlying data and access to a set of interactive controls. The interface design for numeric columns contains six primary elements (Figure 2). The column header includes the variable name (Figure 2a); a sort control (Figure 2b); and a range (Figure 2c) and compact histogram (Figure 2d) showing the distribution of values in the column. Beneath the header is a set of data cells, one for each row or row group. Each cell provides a view of the corresponding data value(s) (Figure 2e, see Section 2.2) and a bar representing the size of the row group represented by the cell (Figure 2f).

When a user clicks on a column's sort control, the individual non-grouped table rows are first sorted by that column's value, in increasing or decreasing order. An algorithm is then applied to simplify the column by grouping adjacent rows. Currently three such algorithms are supported: *quantiles*, *k-means*, and *gap* (Figure 2). For all algorithms, the user can specify n, the number of output rows for the table.

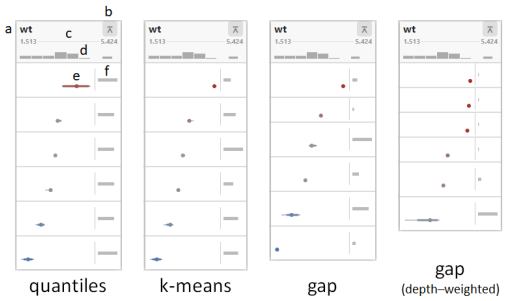

Figure 2: Numeric simplification methods. Three algorithms—*quantiles*, *k-means*, and *gap*—are implemented for grouping rows based on numeric column values. Here they are applied to the same *wt* (weight) column from an automobile dataset, each with six output rows. The size of each row group is indicated by a horizontal bar incorporated into the active column. For the gap method, an optional depth weighting can be applied, showing more detail at the top of the table, and greater summarization at the bottom.

2.1.2 Categorical Columns

Categorical columns are displayed similarly to numeric columns. The header differs only in the use of a color-coded bar chart showing the relative frequencies of each unique value. The column cells also adopt a categorical bar chart design.

Simplification is accomplished via a process similar to that of numerical columns: rows are sorted first, then grouped. As there is no inherent ordering for categorical data, clicking the sort control will apply an alphanumeric sort to the rows based on the category label. Rows with the same value are then grouped and represented via a single cell.

To provide more information about the column and its relation to other columns, the user can then choose to sort the table by (a) the category count in the current column or (b) a second column, via a secondary sort button that appears for each column when the active columns is categorical. E.g., the user may select a categorical column to group by, and then a numeric column to sort by. Row groups will then be sorted by their median value, which can reveal relationships between categories and numeric data.

2.2 Visual Modes

To aid in interpreting the simplified table, different visual modes are available incorporating both text and visualizations (Figure 3).

2.2.1 Text

Tables typically display textual and numeric data. To support a similar familiar display of data when simplification is applied, a data-type-dependent display of the distribution of values in the row group is shown in text mode. For numeric columns, the minimum, median, and maximum values in the group are shown (Figure 3a). For categorical columns, the category with the largest count in the group is shown, along with the number of other categories (Figure 3b). For ID columns, the id value at the top of the sort order in the group is shown, along with the number of other values in the group (Figure 3c). In all cases the size of the group is indicated in the active column next to the distribution text.

2.2.2 Visualizations

To support understanding of the distribution of values within each group in each column, different visualizations can be shown in place of the textual display. For numeric columns, a glyph indicates the distribution range, interquartile range (IQR), and median value (Figure 3d). A blue-grey-red color map redundantly encodes the median value to aid in identifying high and low values and relationships between columns.

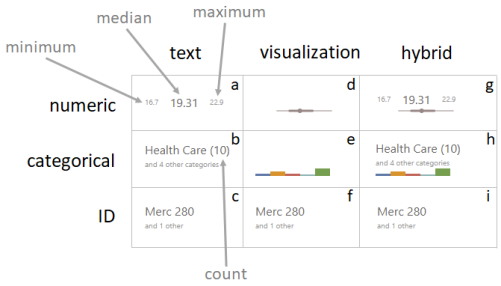

Figure 3: Visual modes for each column type, indicating the distribution of values in each cell.

For categorical columns, a bar chart indicating category count is shown. Bar chart height can be scaled by the maximum value in the cell, or the maximum value across all cells in the column (Figure 3e). Due to the difficulty in summarizing unique labels graphically, ID columns are displayed via the same method used for text mode. In all cases the relative size of the group is indicated in the active column with a horizontal bar next to the distribution visualization (e.g. Figure 2).

2.2.3 Text + Visualizations

The user can also select a hybrid mode showing both text and visualizations at the same time (Figure 3g-i). While this provides additional information, it may be overwhelming to interpret. We have therefore implemented an interactive mode, in which by default each column is displayed using the visualization mode; mouse over of any column shows hybrid displays for that column and the active column (Figure 1).

2.3 Additional Features

Additional features include the ability to expand groups to inspect each row in more detail (Figure 1d-e), and pin individual rows such that they are always visible. To support tables with many columns the active column moves to stay in view as the user scrolls the table left and right.

3 Future Work

Future work includes comparative evaluations against standard tables and other multivariate visualization techniques. In addition, various features, such as methods for clustering and condensing columns to handle datasets with many variables could be explored, as could grouping rows based on multiple columns instead of one.

Acknowledgments

The research reported in this article was supported in part by a grant from the National Science Foundation (1704018).

References

[1] L. Bartram, M. Correll, and M. Tory. Untidy Data: The Unreasonable Effectiveness of Tables. *IEEE Transactions on Visualization and Computer Graphics*, 28(01):686–696, Jan. 2022. Publisher: IEEE Computer Society. doi: 10.1109/TVCG.2021.3114830

[2] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit. LineUp: Visual Analysis of Multi-Attribute Rankings. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2277–2286, Dec. 2013. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2013.173

[3] C. Nobre, N. Gehlenborg, H. Coon, and A. Lex. Lineage: Visualizing Multivariate Clinical Data in Genealogy Graphs. *IEEE Transactions on Visualization and Computer Graphics*, 25(3):1543–1558, Mar. 2019. doi: 10.1109/TVCG.2018.2811488

[4] E. R. Tufte. *The visual display of quantitative information*. Graphics Press, USA, 1986.