Overview of Visual Profiling

Contents:

Overview

In Trifacta® Self-Managed Enterprise Edition, **visual profiling** provides real-time interactive visualizations of your dataset to assist in the discovery, cleansing, and transformation of your data. Visual representations are required for interpreting large volumes of data, and the platform's innovative profiling techniques visualize key statistical information in a dynamic, easy-to-consume format for faster transformation.

- At the individual column level, visual profiles provide interactive statistical information visualized in an appropriate manner for the data type. For example, columns of Zip Code data type can be represented on a geographical map of the United States.
- All visual profiles are interactive, so you can dig into the details of the data. Select one or more elements in a profile, and you can take immediate action on the data, either through steps you define or through transform recommendations provided by the platform.
- The Transformer page displays a set of recommended actions to take based on the values, rows, or columns that you select in the data grid. These recommendations are motivated by platform logic and prior usage information. For more information, see *Overview of Predictive Transformation*.

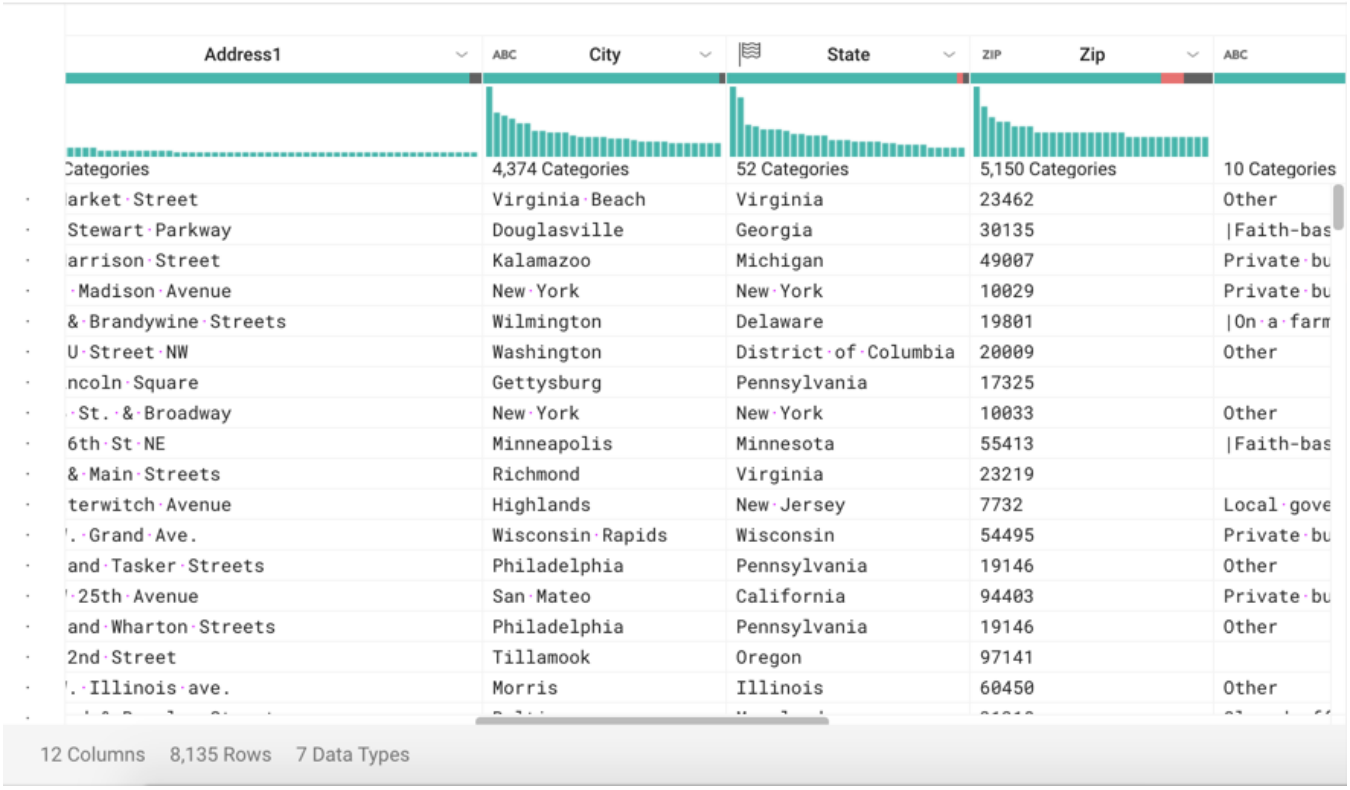
Visual profiles are available while you transform your data in the Transformer page, when you dig into the detail of individual columns, and after you execute your job at scale. Each of these interfaces has different usage patterns designed to accelerate and simplify data transformation for that specific area of the process.

Uses

- **Locate anomalies.** Visual profiling surfaces missing or invalid data in individual columns. These values can then be selected and transformed as needed.
- **Identify distributions.** In the data grid, you can review value distribution for each column in your dataset. When exploring the column details, you can also identify and select statistical outliers among your column data.
- **Identify areas for further refinement.** After a job has completed, you can review its visual profile through the application and then take action on problematic data.

Example

In the following example, a dataset containing address information has been loaded in the Transformer page:



Address1	City	State	Zip	
Categories	4,374 Categories	52 Categories	5,150 Categories	10 Categories
arket Street	Virginia Beach	Virginia	23462	Other
Stewart Parkway	Douglasville	Georgia	30135	|Faith-bas
arrison Street	Kalamazoo	Michigan	49007	Private bu
Madison Avenue	New York	New York	10029	Private bu
& Brandywine Streets	Wilmington	Delaware	19801	|On a farm
U Street NW	Washington	District of Columbia	20009	Other
ncoln Square	Gettysburg	Pennsylvania	17325	
St. & Broadway	New York	New York	10033	Other
6th St NE	Minneapolis	Minnesota	55413	|Faith-bas
& Main Streets	Richmond	Virginia	23219	
terwitch Avenue	Highlands	New Jersey	7732	Local gove
. Grand Ave.	Wisconsin Rapids	Wisconsin	54495	Private bu
and Tasker Streets	Philadelphia	Pennsylvania	19146	Other
25th Avenue	San Mateo	California	94403	Private bu
and Wharton Streets	Philadelphia	Pennsylvania	19146	Other
2nd Street	Tillamook	Oregon	97141	
. Illinois ave.	Morris	Illinois	60450	Other

12 Columns 8,135 Rows 7 Data Types

Figure: Example dataset

In this example, we are interested in exploring geographic information. From the column drop-down for the `Zip` column, you select **Column Details**.

Explore detail on demand. Generate visual profiles from the column drop-down.

When you explore the column details of the new column, you can see the following representation of the data:

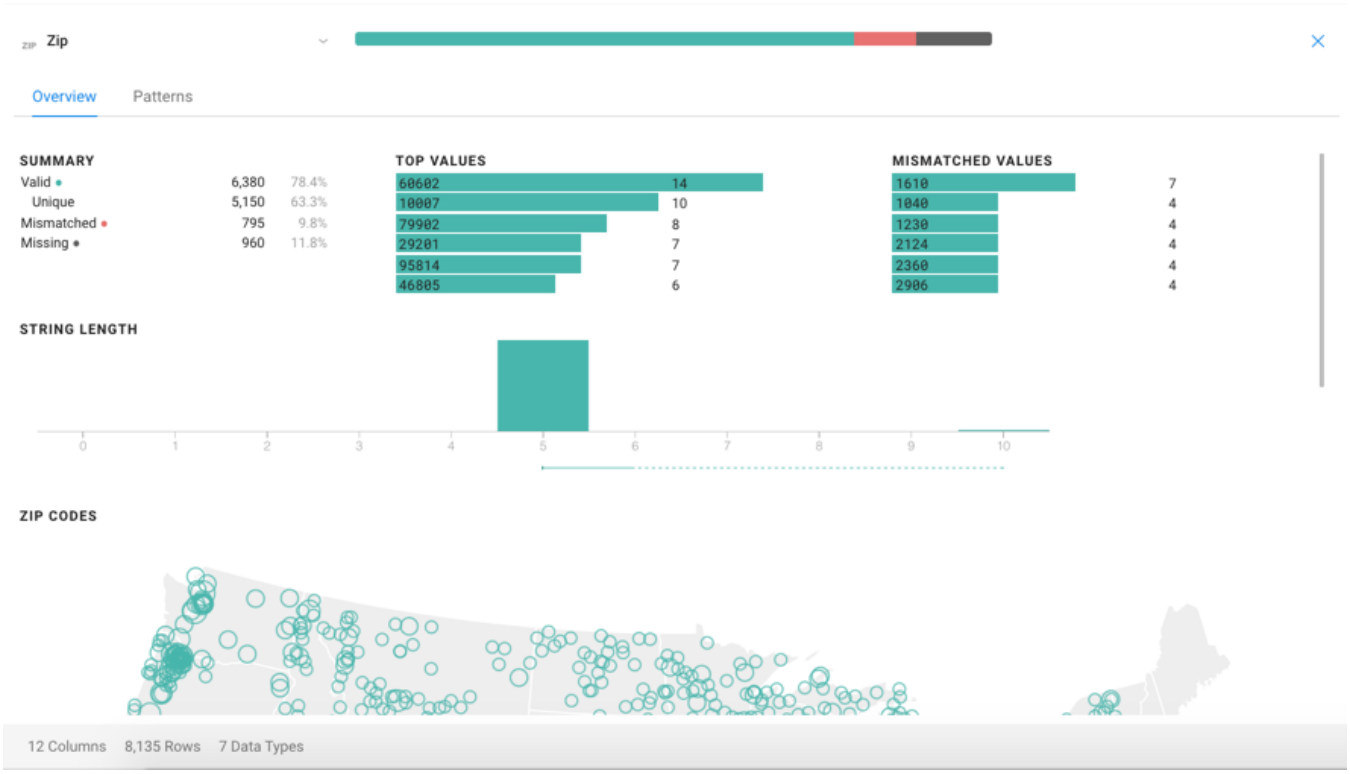


Figure: Zip Code data type represented as a U.S. map

In this case, the values in your Zip column are recognized as being of Zipcode data type. The application then represents these values as a U.S. map, which quickly renders numeric data into a format that's much easier to read and analyze.

Type-specific visualizations. The profile of the column values is represented in a type-specific visualization to assist in rapid analyzing and taking action on some or all values in the column.

Visual Profiling Interfaces

Wherever you can interact with data, visual profiling simplifies the process.

Customized visualizations. Each interface has been optimized for the scope of the data it is visualizing, whether the data is a single column, the entire sample of a dataset, or generated results.

Data Grid

In the Transformer page, the **data grid** is a tabular representation of a sample of your dataset. It is the primary interface through which you build your transformation recipes. Profiling tools:

- **Data Quality Bar:** At the top of each column, you can see graphs counting the missing, invalid, and valid values for the column's current data type. Select one of the categories, and you can take immediate action on all of the category's values in the column.
- **Column Histogram:** Individual values in the column are represented in a histogram at the top of the column. You can select one or more of these values, review relevant data, and take action.
- See *Data Grid Panel*.

Whenever a transform is selected or specified, a preview of its effects is displayed in the data grid, including any changes to the data quality bar and column histogram of affected columns. See *Transform Preview*.

For additional details on visual transformation, see *Transform Basics*.

Column Details

Through the Transformer page, you can explore statistical details about individual columns, visually represented based on the column's data type. From the drop-down for any column, select **Column Details**.

In this interface, you can review the range of values in the column and can optionally select one or more values from other columns to see which values in the current column apply. The visualizations for a column depend on the data type.

See *Column Details Panel*.

Pattern Profiling

In the Column Details panel, you can review profiling of patterns detected in the values for the selected column. These patterns can be selected, which identifies the relevant values in the column that match the pattern. You can then use these selections as the basis for building transforms that apply to the matching values.

For more information, see *Column Details Panel*.

Job Results

After the application has successfully executed a job for which profiling is enabled, you can explore a visualization of the generated dataset in the Job Results page. See *Job Results Page*.

Profiling Engine

Decoupled from the user interface, the profiling engine performs the calculations required to power the visualizations before job execution and after the job results have been generated.

- In the Transformer page, the profile engine is called for incremental changes whenever a step is added to your recipe, so that you can see immediate updates to the visual profile for each column. It utilizes separate algorithms for generating the data quality bars, column histograms, value counts, frequency distributions, and other relevant statistics. When you dig into the column details, the visual profile is up-to-date and can be updated again based on your selections in that interface.
- During job execution, it is queried as a separate job when profiling is executed across the entire dataset.

NOTE: When you choose to profile your results, you are creating two distinct tasks: 1) run your transform recipe against your source and 2) profile the results. Due to the computational complexity of generating the interactive results, a profiling task often takes longer to complete than a transformation task and is therefore an optional element of a job run.

Exact vs. Approximate Metrics in Visual Profiles

Generally, visual profiles represented in the user interface, in places like column histograms and column details, are exact measurements against the current sample.

On generated results, visual profiles tend favor approximations.

NOTE: The computational cost of generating exact visual profiling measurements on large datasets in interactive visual profiles severely impacts performance. Depending on the environment, you may choose to run profiling jobs on generated results as separate jobs. For more information on enabling this feature, see *Profiling Options*.

Below, you can review details on how metrics are calculated in visual profiling performed in different areas of the platform.

User Interface

The UI leverages the Photon running environment when displaying visual profiles on sampled data.

> **NOTE:** Profiles are executed on the currently sampled data. Results may vary when the full transformation job is executed.

Metric Type	Measurement
Frequency (top-k)	Exact
Unique value counts	Exact
Numerical histograms	Exact
Simple statistics (mean, stdev, min, max)	Exact
Quartiles	Exact

Photon Running Environment

When profiling jobs are executed on Trifacta Server, they leverage the server-side version of the Photon running environment.

Metric Type	Measurement
Frequency (top-k)	Approximate
Numerical histograms	Approximate
Simple statistics (mean, stdev, min, max)	Exact
Quartiles	Exact

Spark Running Environment

For profiling jobs, the Spark running environment is used for Spark transformation jobs and optionally for Photon jobs. For more information on enabling the execution of visual profiling on Spark for Photon jobs, see *Profiling Options*.

Metric Type	Measurement
Frequency (top-k)	Approximate
Numerical histograms	Approximate
Simple statistics (mean, stdev, min, max)	Exact
Quartiles	Approximate