Preparing Data With a Data Flow

A Data Flow extracts data from one or more data sources and loads it to one or more target sources. You can use a data flow to migrate data between database management systems and operating environments or to update existing targets.

Before you load data into your target data source, you can enhance it for your needs. This massaging of the data prior to loading is called *data preparation.* You can take advantage of the data preparation options to, for example, convert numeric codes to meaningful attributes, discard erroneous data, smooth out ragged data into manageable bins, and blend descriptive data from additional data sources.

The data preparation calculations that create new fields use some of the same tools that are available in the Synonym Editor. However, when the calculations are added in the Synonym Editor, they are added to the Master File. The data remains the same, but the calculations are performed every time the field is referenced in a request against the synonym. When you create new fields in a data flow and then run the flow, calculated values are loaded into the target data source as field values, not calculations.

Many of the data preparations examples in this topic use data from a station-based bike share system. Bikes are unlocked from one station and returned to any other station in the system. The data we will use has been extracted from the daily ridership and membership information publicly available from Citi Bike, a public-private partnership between New York City and Lyft Bikes.

Generating Sample Files for Data Preparation

Citi Bike provides data monthly as zipped comma-separated values (.csv) files that contain the following data values:

- Trip Duration (seconds)
- Start Time and Date
- Stop Time and Date
- Start Station Name
- End Station Name
- Station ID
- Station Latitude/Longitude
- Bike ID
- User Type (Customer = 24-hour pass or 3-day pass user. Subscriber = Annual Member.)
- Gender (Zero=unknown; 1=male; 2=female)
- Year of Birth

To download a ridership file;

1. Go to http://www.citibikenyc.com/system-data.
2. Click the link that says **Download Citi Bike trip history data.**
3. Click **201907-citibike-tripdata.csv.zip** to download the file.

As part of the data preparation, this data will be augmented to have:

- Trip duration in minutes.
- Age in years.
- Alphanumeric gender values.
- Additional date components.
- Start Station Zip Code, City and County.

Depending on which zip file you download, your results may vary from ones in this topic, which uses the data from July, 2019.

Once you have downloaded a file, you can upload it to the server.

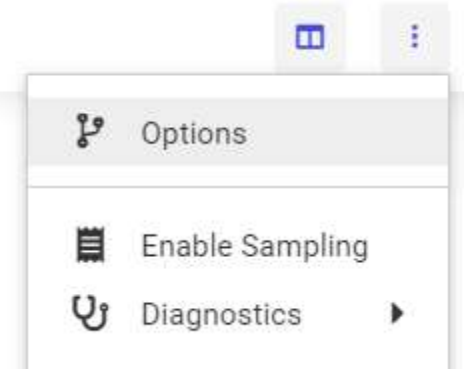
A supplementary file was created that has station zip codes and counties (station_zip.csv). You can download this file
from http://techsupport.informationbuilders.com/public/station_zip.csv after which you can upload it to the server.

Enabling Sampling for a Data Flow

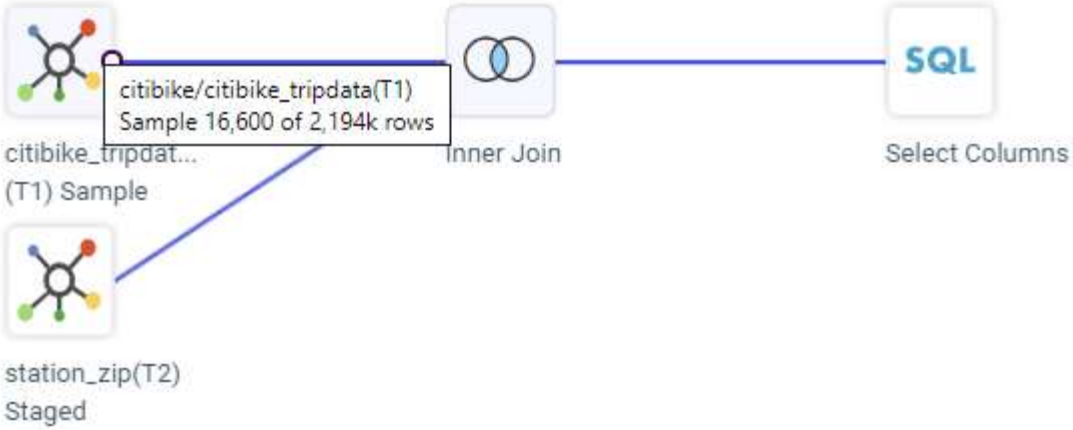
When a data source in a flow has a large volume of data, you can enable sampling for better response time. You can make decisions based on a sample, provided that sample is representative of the entire data set. Data Prep has a built-capability to automatically generate a random sample (with a 99% confidence level and +/- 1% margin of error).

To enable sampling:

1. Create a flow by right clicking an application, clicking **New,** then **Flow,** or by clicking **New** on the ribbon, then **Flow**.
2. Click **Advanced** on the ribbon, then click **Enable Sampling**, as shown in the following image.



Once you have enabled sampling, you can see if a file if a file was sampled, and the sample size, by hovering over the file, as shown in the following image.



If a data source in a flow is not large enough to require sampling, it will not be sampled, but will be staged in the same staging target as the large sources, to eliminate joining disparate data sources and, therefore, improve join performance. By default, the staging target data source is a DATREC file, but you can configure staging target to be the same relational data source as the load target by checking **Use ETL-TRG-DBMS for Sampling** in the Data Flow parameters of the Advanced Options dialog box that opens when you click **Options** on the data flow menu.

You can also enable sampling for all flows, so that a sample is taken automatically if needed when the flow is opened. Change the setting **Enable Sampling** to **On** in the Data Assist (Representative Sampling) section of the **Settings for Web Console Preferences** page available from the **Settings** menu of the Web Console Workspace page.

Editing Fields in a Data Flow

By default, all fields in a single-segment data source, or all fields from top segment in a multi-segment data source are automatically added to flow. You can turn off this option in the Advanced Options dialog box.

To edit the fields in the flow, right-click the SQL object, and click **Edit**. The Metadata and Query panes open.

You can add fields to the query or move them to a different part of the query by dragging them to the Order by, Columns, Order Across, and Filters and Variables categories in the Query panel. You can delete a field by right-clicking it and selecting **Delete**. In addition, you can add all of the fields from a segment ti the Columns or Order by the category by right-clicking the segment name, and clicking **Add to Query** and either **Column** or **Order by.**

A report of the field values opens.

You can also aggregate the data groups by one or more field values. To perform aggregation, click the menu icon for **Order By** and select **Switch to Group By.** Drag the columns you want to group by to the **Group By** field**.**

Each column in the order by field is assigned a default aggregation, Sum for numeric fields, and Max for others. You can change this by selecting the column and, from the context menu, **Aggregation,** then the type of aggregation to be performed, such as Sum, Count, or Average.

Showing Profiling in a Data Flow

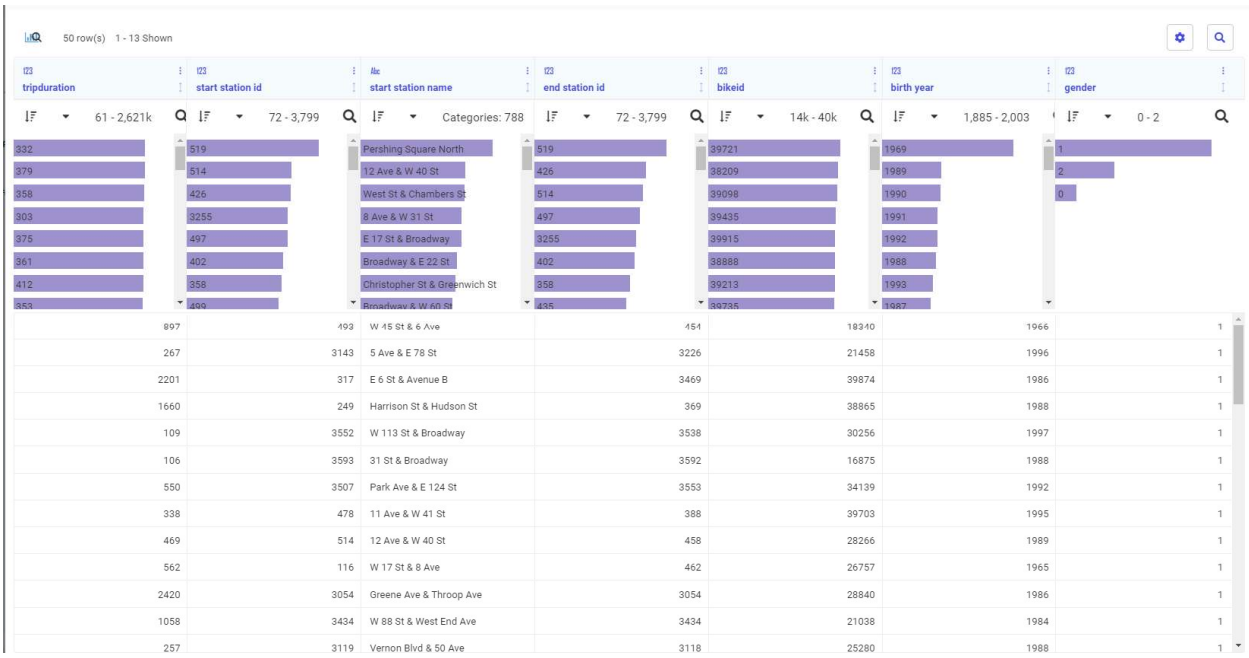
You can generate a distribution chart for each column in the flow, which shows a visual image of the field values. The profile chart displays below the field name and shows:

- The range of values for numeric and date fields.
- The count of unique values (categories) for character fields.

To show the profiling distribution charts:

1. Create a flow and select columns
2. Click the **Show Profiling** button () above the sample data on the Edit Selection page or the flow page.

The profiling distribution charts display below the field names and above the field values, as shown in the following image.



Some the interesting facts about the data shown by the distribution charts that are that the number of start stations is 788 (the number of categories displayed above the distribution chart for the Start Station Name alphanumeric field), the Pershing Square North station is the most popular starting station (which makes sense as it is near a major transportation hub), and that the trip durations range from 61 seconds to over 2 million seconds (a range is shown for numeric fields).

- An icon that identifies the data type of the field displays above the field name in each column. If the column was calculated, an equal sign is added before the icon.

Data Type	
Alphanumeric	Abc
Integer	123
Decimal	12
Date-Time	
Date	
Location	;

- Below the field name, there is a summary of the distribution of the data values. For numeric and date fields, it shows the range of values, and for character fields, it shows a count of categories (unique values).
- By default, the distribution charts are sorted by frequency in descending order. You can click the Sort Type button to sort item by value, and the arrow to change the sort order.

Description	
Sort Type	
Sort Order	

- A search icon (magnifying glass) in each column enables you to search for values in the column. Enter the text you want to find. As you type and search the finds the values, it highlights any matching values and scrolls the chart so that the first bar with the highlighted values is at the type. To remove the search, either delete the characters or click the X to the right of the search box. To close the search box, click the X when the search box is empty.

Adding and Replacing Fields in a Flow

You can add new fields or replace existing fields by creating expressions.

Each field has a menu icon ( )

The menu provides the following options:

- **Format.** Opens the **Edit Display Format** dialog box.
- **Rename.** Enables you to change the title of the field.
- **Replace with Expression.** Enables you to replace a field with values derived using a calculation.
- **Add new Expression.** Enables you to create a new field with values derived using a calculation.
- **Delete.** Removes the field from the flow.

Calculated fields also have:

- **Properties.** Opens the Properties dialog box.
- **Edit Define (Advanced).** Opens the Advanced Expressions dialog box.

When aggregated, fields also have :

- **Edit Compute (Advanced).** Opens the Advanced Expressions dialog box.
- **Aggregations.** Enables you to select an aggregation operator.
- **Add to Filter Aggregated.** Opens a filter card for the field.

Order by fields also have:

- **Sort.** Options are Ascending or Descending. The default is Descending.
- **Rank.** Options are Yes (ranks the values in addition to sorting) or No. The default is No.
- **Limit.** Options are No Limit, 1, 5, 10, or custom. The default is No.
- **Visibility.** Options are Show or Hide. The default is Show.
- **Add to Filter.** Opens a filter card for the field.

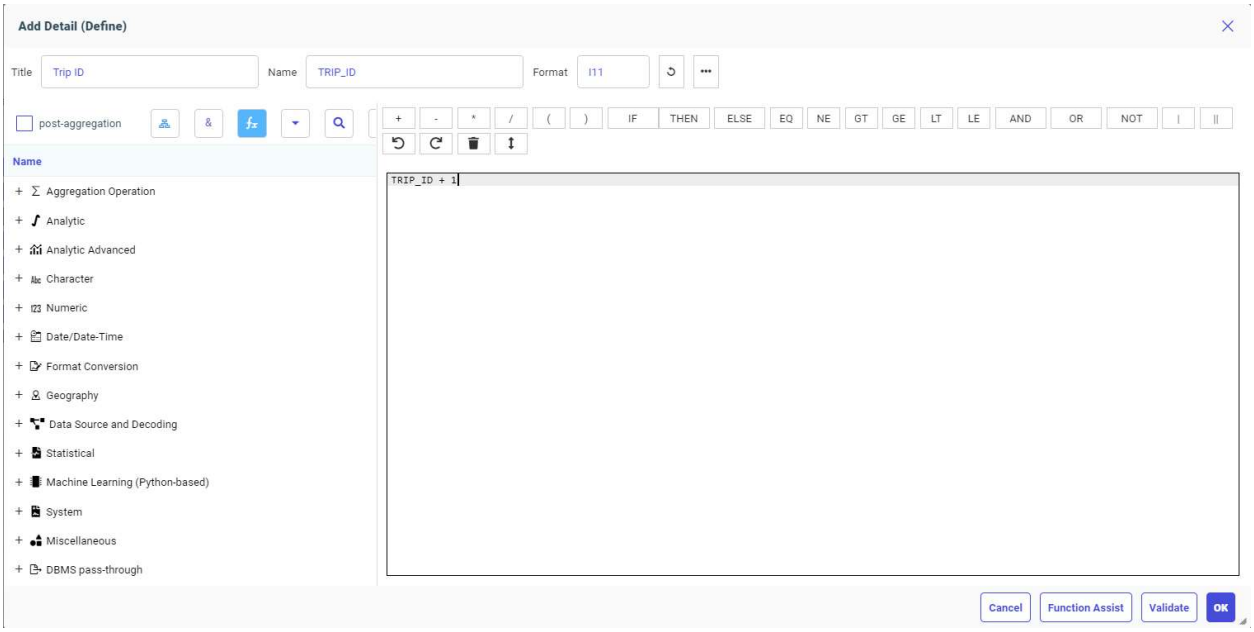
When you create an advanced expression, you use the expression calculator, which is shown in the following example.

Example: Creating a New Trip ID Field

To create the TRIP_ID field, click the menu icon in the tripduration column and click **Add New Expression**, then **Advanced Expressions.**

The Add Detail (Define) Expression Calculator opens. Enter the value Trip ID for the Title and Name will be filled automatically as TRIP_ID. For the expression enter TRIP_ID + 1. This creates a counter, where the value for each row increases by one.

Optionally, you can click **Validate** to make sure that the expression is valid. As the expression is arithmetic, the format changes from alphanumeric to integer, as show in the following image.



Click **OK** to add this field to the flow.

Example: Creating a New Field with Trip Duration in Minutes

Trip duration is more meaningful when expressed in minutes instead of seconds.

To create the TRIP_DURATION_MINUTES field, click the menu icon in the tripduration column and click **Replace with Expression,** then **Advanced Expression.**

The Add Detail (Define) Expression calculator opens. Perform the following steps:

1. Enter the title *Trip Duration,Minutes.* The field name becomes TRIP_DURATION_MINUTES.
2. After the text *tripduration* add a division sign and the number 60 (tripduration/60) to calculate minutes.
3. Click **OK.**

The tripduration field is replaced by the Trip Duration Minutes field. The profiling distribution chart for this field shows that the most common trip durations are in the four to seven minute range:

Example: Creating New Fields that Calculate Time Limits and Overages

In this example, we will calculate trip limits and overages.

Trip Limit

Allowed minutes are conditional based on their user type. The allowed minutes for Subscriber is 45, and for Customers 30.

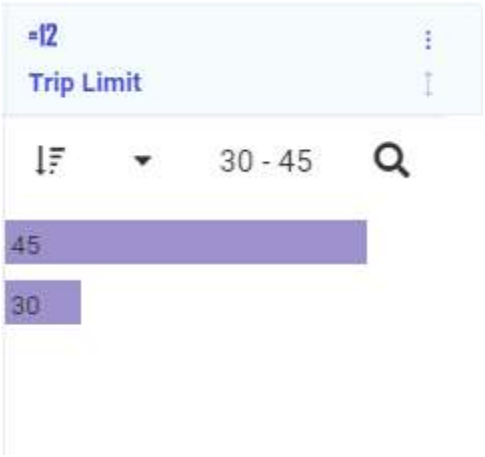
1. Click the *usertype* column, and click **Add new expression,** then **Advanced Expression.**

   The Add Detail (Define) Expression Calculator. Enter the value *Trip Limit* for the Title, and Name will be filled in automatically as TRIP_LIMIT. For the expression, enter the following conditional expression:

```
IF usertype EQ 'Subscriber' THEN 45 ELSE 30
```

2. Click **OK.**

   The field is added, as shown in the following image.



Base Minutes

Next, we will create a field with the number of base (included) minutes for each trip. This value is the trip duration in minutes, if that is not greater than the trip limit. If it is greater, we will use the trip limit as the value, and then calculate how many minutes of overage the trip used. We will generate the base minutes field by applying the MIN function, which returns the minimum value of its arguments.

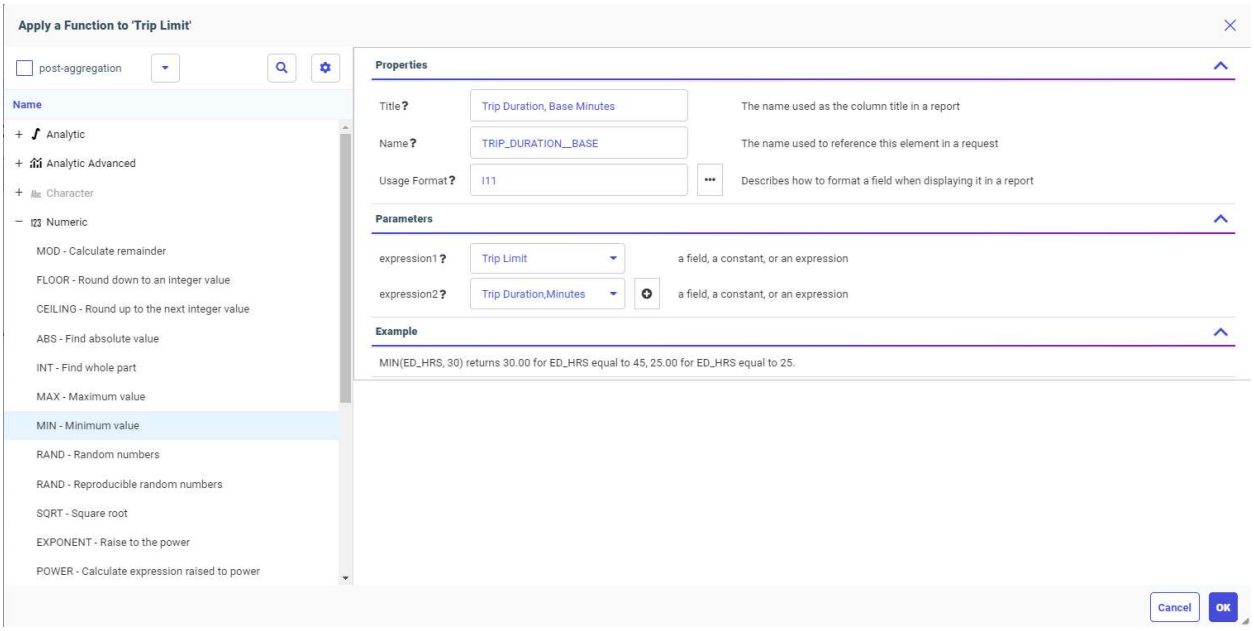
1. Click the menu icon in the Trip Limit field, and click **Add new Expression,** then **Apply Function.**
2. The Apply Function dialog box opens and the Numeric folder opens.
3. Click the function **MIN - Minimum Value.**
4. The properties for the MIN function open, with default values for the Title, Name, Usage Format, and first expression fields.
5. Change the default entries by entering the following values.
   - **Title:** Trip Duration, Base Minutes.

     Name: The default name becomes TRIP_DURATION_BASE_MINUTES. Change it to TRIP_DURATION_BASE

   - **Usage Format:** I11, or click the menu button to get a properties panel where you can select Integer and type in the length.

- **Expression1:** Trip Limit (this was already there, by default)
- **Expression2:** Trip Duration,Minutes (select this field from the drop-down list).

The completed dialog box is shown in the following image.



6. Click **OK.**

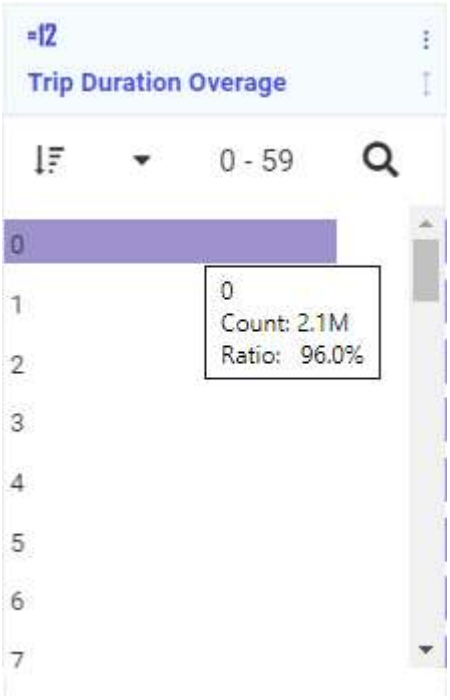
The field is added, as shown in the following image.

Overage Minutes

Next, we will create a field with the number of minutes over the allowed amount (if any).

1. Click the menu icon in the Trip Limit field, click **Add new expression,** then **Apply Function.**
2. The Apply function dialog box opens and the Numeric folder opens.
3. Click the function **MAX - Maximum Value.**
4. The properties for the MAX function open, with default values for the Title, Name, Usage Format, and first expression fields.
5. Change the default entries by entering the following values.
    - **Title:** Trip Duration,Overage. The default name becomes TRIP_DURATION_OVERAGE
    - **Usage Format: I**11
    - **Expression1:** 0 (zero)
    - **Expression2:** TRIP_DURATION_MINUTES - TRIP_LIMIT. This calculates the overage. If it is greater than zero, this value will be returned. Otherwise, zero is returned.
6. Click **OK.**

The field is added. Since we omitted trips over 90 minutes, the overages are from 0 to 59 minutes. Hover over the zero bar and you can see that approximately 96% of the trips have no overage, as shown in the following image.



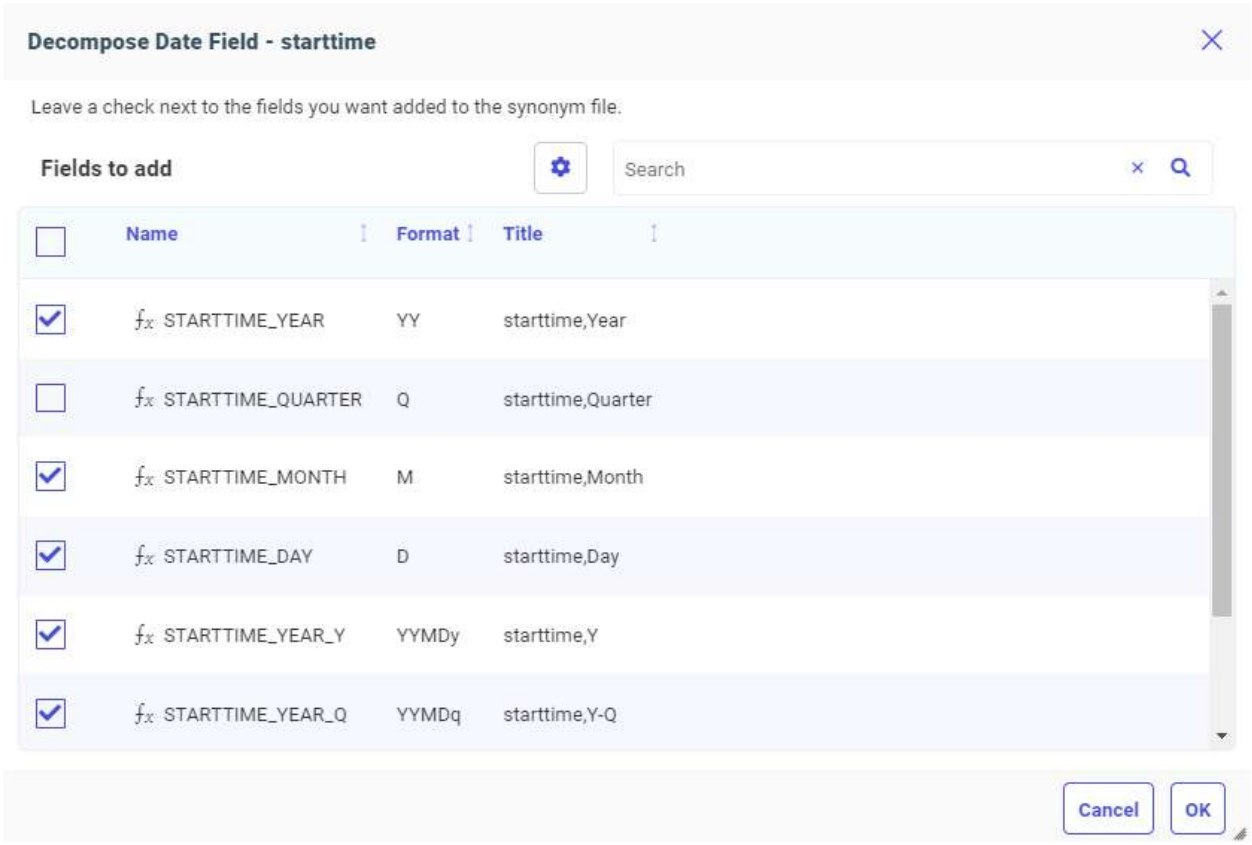
Example: Creating New Fields with Date Components

This trip data includes a time stamp, to the millisecond, that each ride started. For visualization, we want to break these files into their component value, Year, Month, and Day of Month.

Start Year, Start Month, Start Day

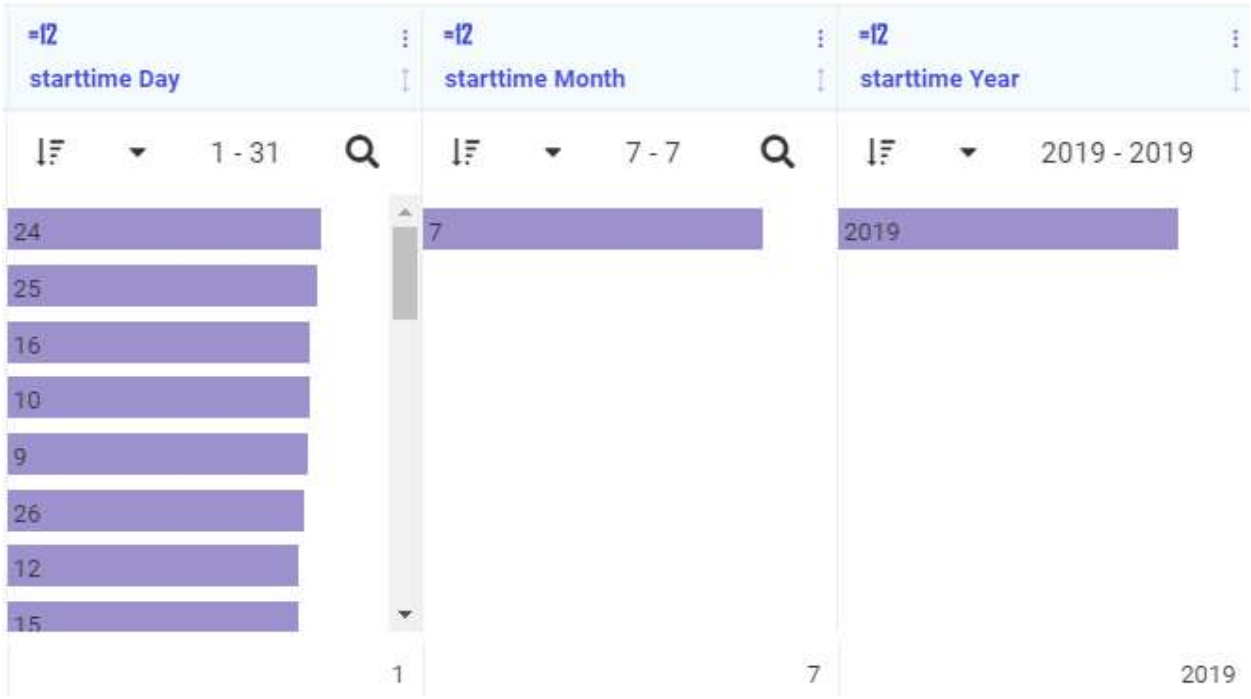
1. Click the menu icon in the starttime column, click **Add new Expression** then, **Decompose Date...**
2. The **Decompose Date** dialog box opens.
3. Check the following components.
    - STARTTIME_YEAR
    - STARTTIME_MONTH
    - STARTTIME_DAY

    Leave the default selection as is, as shown in the following image.



4. Click **OK.**

The fields are added, as shown in the following image.



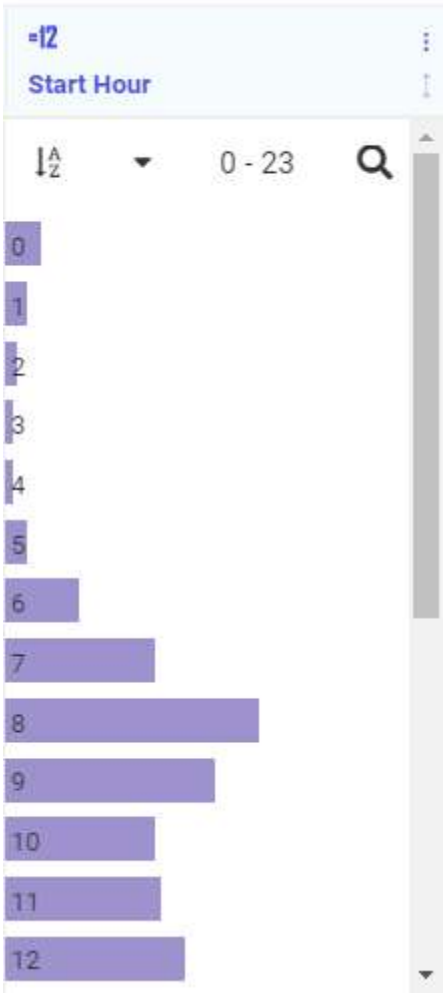
Hour of the Day

The date function DPART returns components from a date or date-time field. Extracting the hour of the day will show how usage varies during the day.

1. Click the menu icon in the starttime column, click **Add new Expression,** then **Apply Function.**
2. The **Apply Function** dialog box opens, with the Date/Date-Time functions open.
3. Double-click DTPART to select it.
4. The properties panel for DTPART opens.
5. Enter or select the following values.
    - **Title:** Start.Hour. The Name becomes START_HOUR.
    - **Component:** HOUR (select this from the drop-down list)
6. Click **OK.**

The field is added, as shown in the following image.

To get a better idea of usage throughout the day, click the down pointing triangle and click **Sort by Value.** The distribution chart is now sorted by hour of the day, as shown in the following image.



Example: Creating a New Field With Age

The trip data includes the birth year for each subscriber, but age would be more meaningful. The range of values for birth year starts at 1885, which is very unlikely. Also, the most frequent birth year is 1969.

Click on the bar that represents 1969. This highlights the percentage of each value in the other columns that correspond to birth year 1969. Most of them have gender 0 (not identified), as shown in the following image.

In addition, if you scroll over to the usertype column, you see that most of them are Customers (day pass buyers), not Subscribers. This indicates that the 1969 is a default birth year for customers.

You can remove the selection for 1969 by clicking the red X in the column header.

To add a field for age:

1. Click the menu icon for the birth year column and click **Add new Expression,** then **Advanced Expression.**

   The Add Detail (Define) dialog box opens.

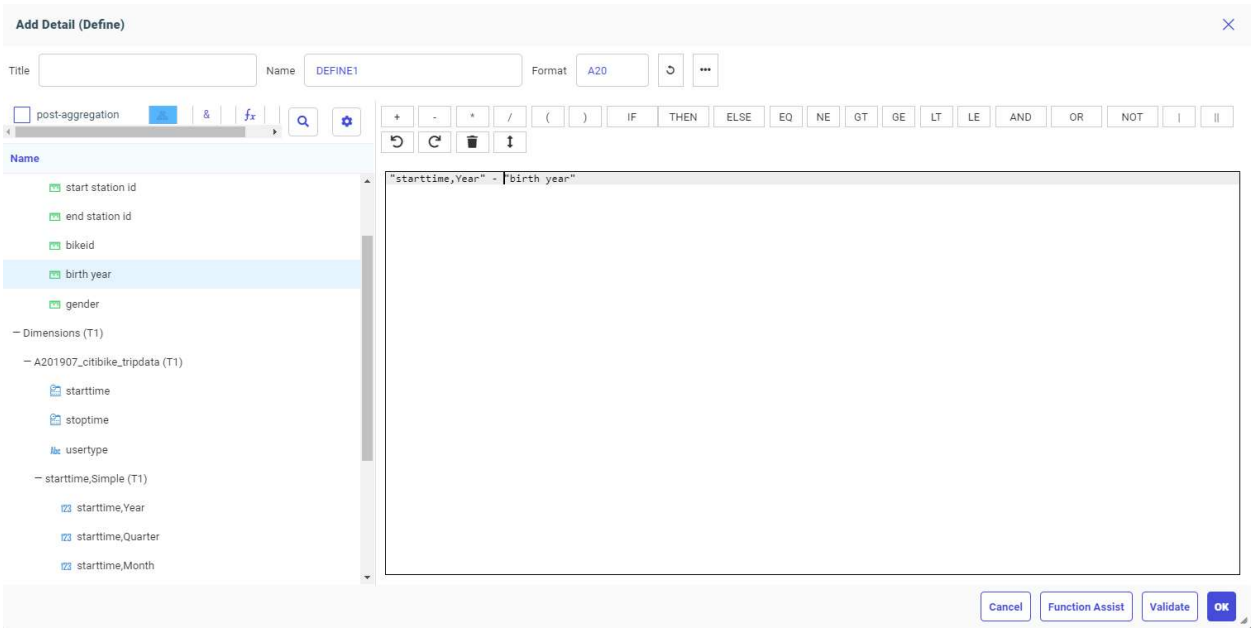
2. Enter the title Age. The field name becomes AGE.

3. Click the Columns button ( )

   The folder tree opens.

4. Scroll down and drag starttime,Year to the input area. enter a minus sign (-) after this. and drag birth year after the minus sign.

   The following image shows this expression.



   This will be the calculation for the age of those riders that fall within the age parameters we want to see.

   However, we want to omit riders who are Customers with the birth year 1969, or any rider born prior to 1939, so we will set their value to Missing.

5. Edit the expression to be the following

```
6. IF "birth year" LT 1939 OR ("birth year" EQ 1969 AND usertype EQ 'Cus
   tomer' )
7.      THEN MISSING
```

```
    ELSE "starttime,Year" - "birth year"
```

8. Click the **Validate** button to make sure the expression is valid. This also refreshes the format. You can also click the **Refresh** button to update the format.
9. Click **OK.**

The age range is 16 through 79, and the most common age is 30

Example: Creating a New Field with Gender as Text

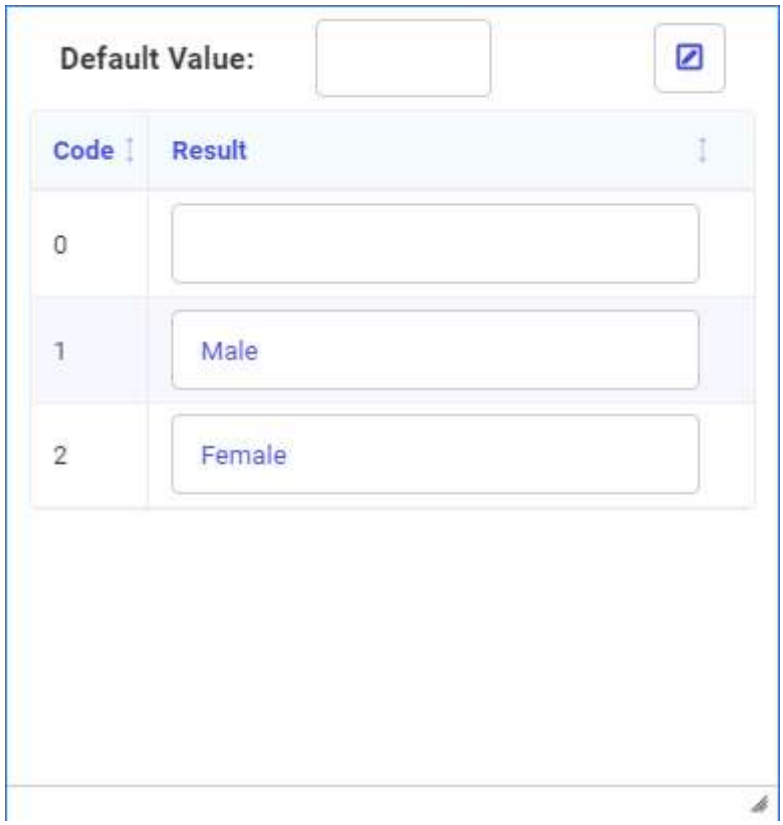
The trip data includes the gender of each subscriber as code 1 or 2 (zero for unknown). For our visualizations, we want to use the labels Male or Female, and when those are not supplied, we will use NULL or Missing.

To create the GENDER_TEXT field:

1. Click the menu icon in the gender column, and click **Add new Expression**, then **Create Decode.**

   A Decode card options.

2. Enter Male for 1, Female for 2, as shown in the following image.



3. Click the Edit button () and enter the following values.

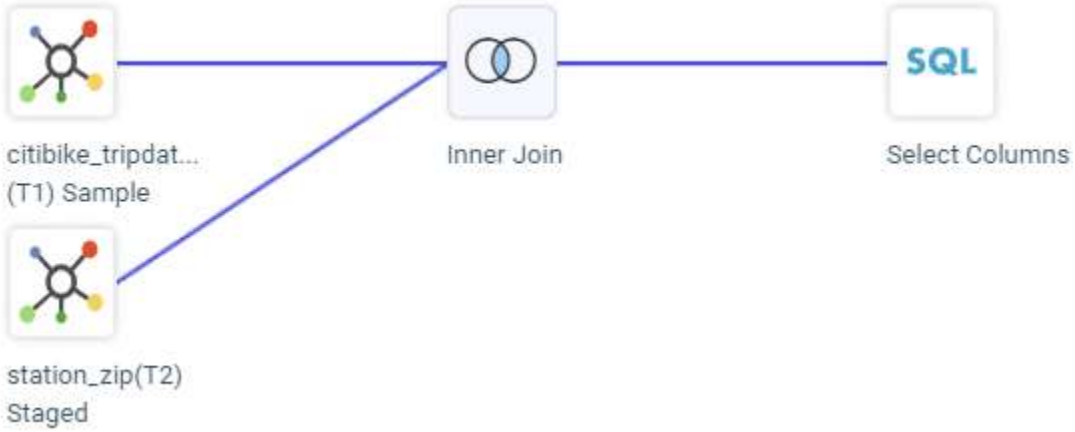
   Name: GENDER_TEXT.

   Title: Gender.

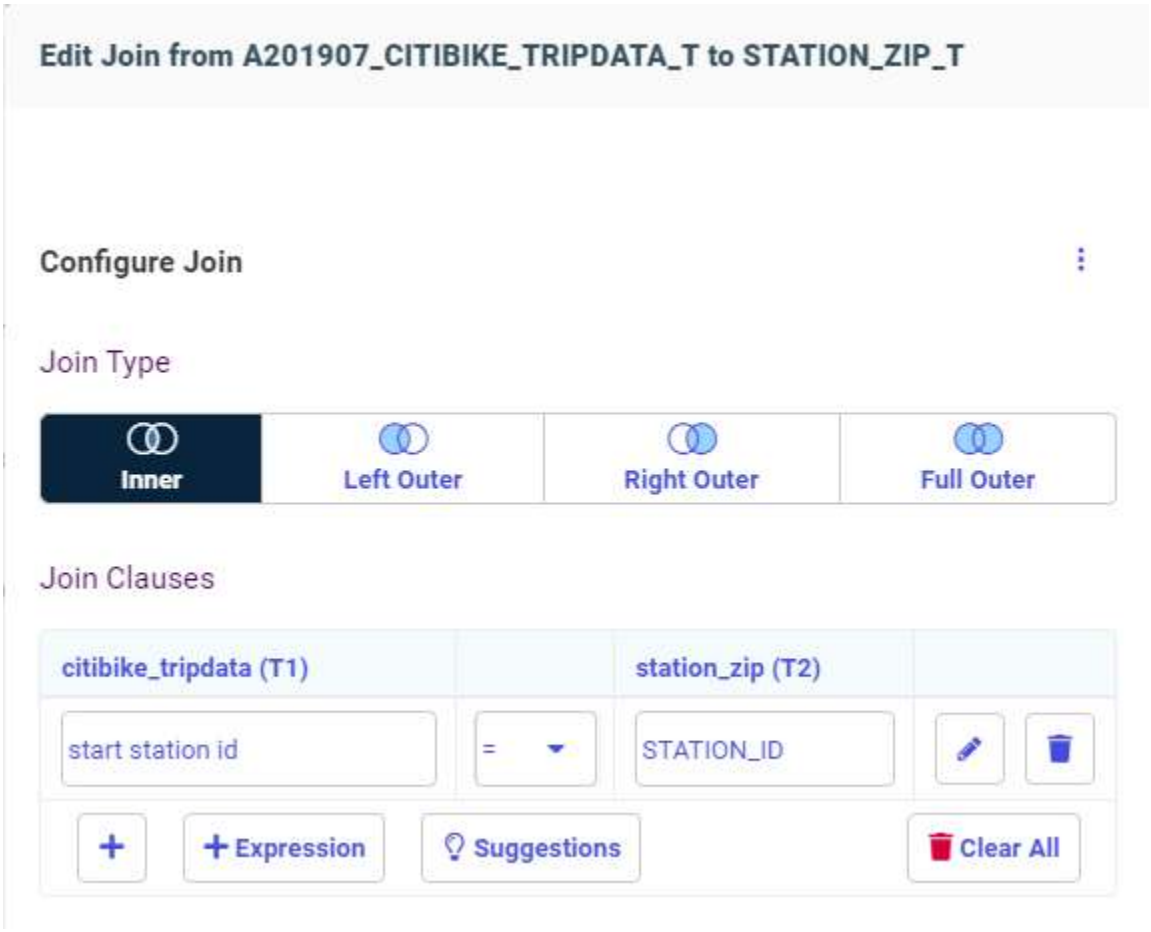
4. Click **OK**
5. Click outside the Decode card to close it.

Join Editing in a Data Flow

The trip data contains station names, IDs, latitudes, and longitudes, but no zip code, city, or county information. To see the zip code and county information, you can download the station station_zip.csv file from http:/techsupport.informationbuilders.com/public/station_zip.csv. You can then upload this file to the server using the instructions on Uploading Files in the *Server Administration* manual, and join it to the trip data file by dragging it onto the flow canvas, as shown in the following image

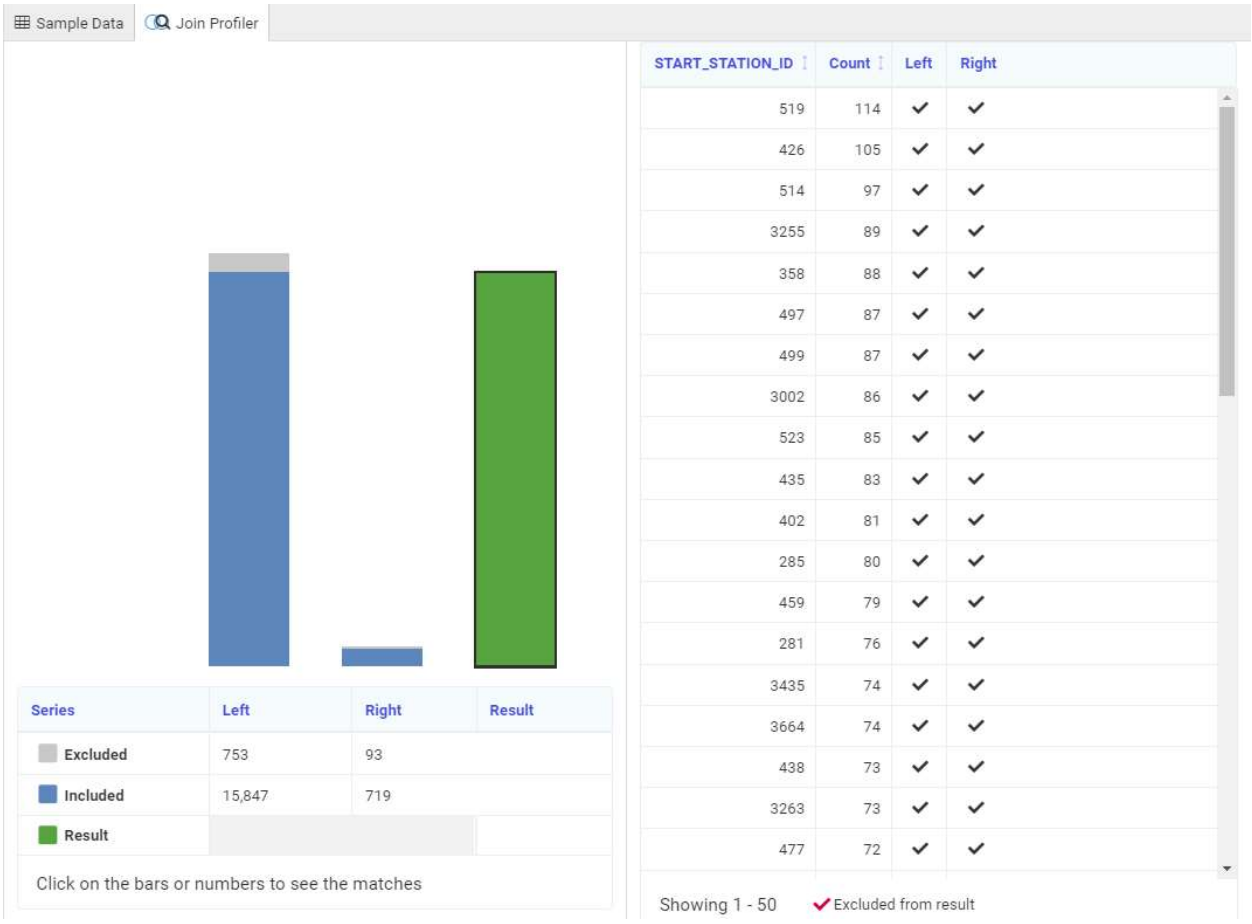


Right-click the Join object and click **Join Editor.** The Join Configuration panel shows the join properties and enables you to edit them, as shown in the following image.



An inner join was automatically created based on the similar field names start_station _id in the trip data file and STATION_ID in the station zip file. You can edit the join condition, add additional join conditions, and change the type of join.

The Join Profiler tab shows the effects of the join. There was not a perfect match between the two files, as shown in the following image.



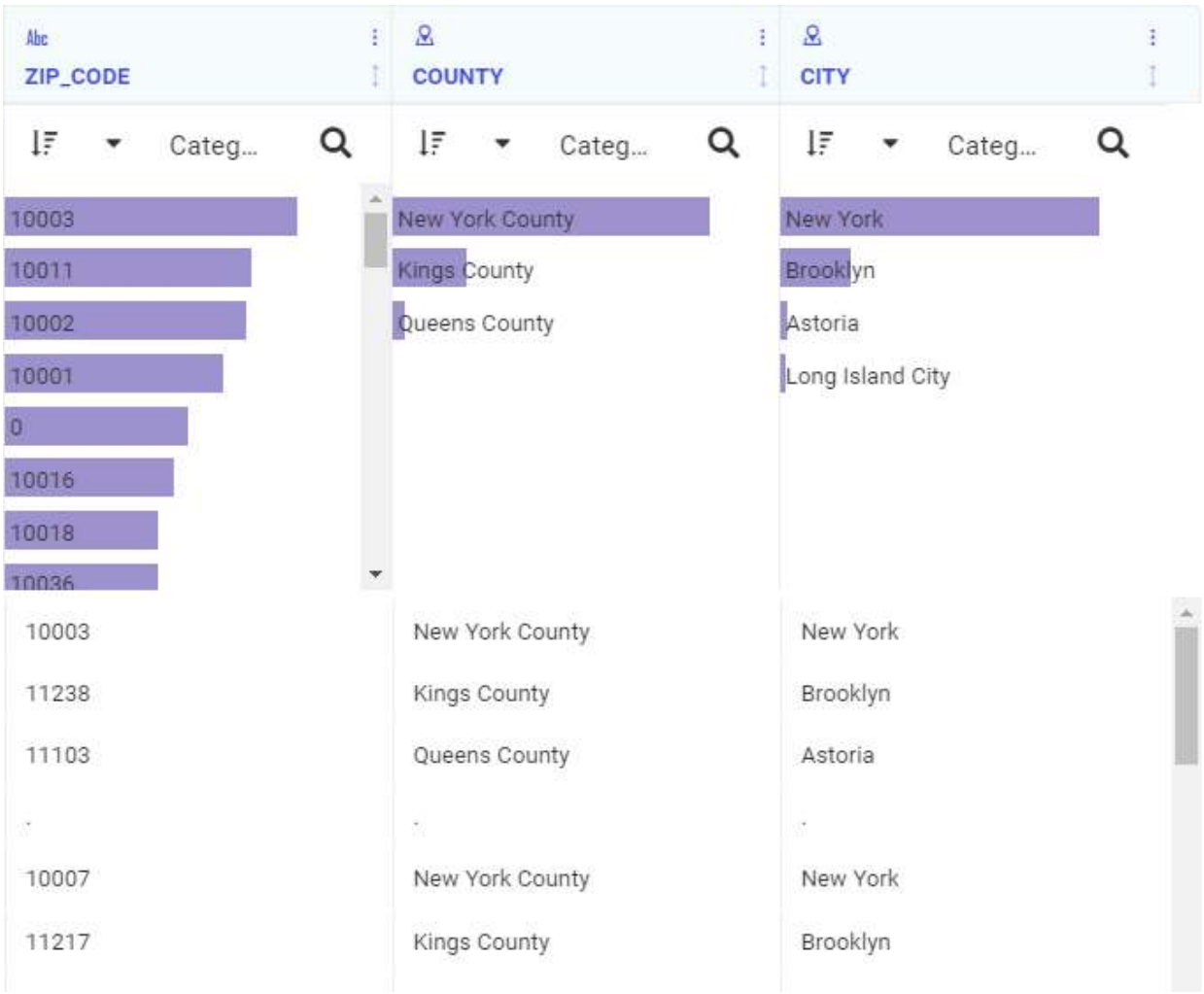
The trip data file was sampled, but the station zip file is small, so the entire file was used for the profiling. In addition, stations may have been opened or closed or have been unused during the time period reflected in the data.

The gray area at the top of the first bar, and the excluded count, represent start stations that were not found in the station_zip file. For the second bar, the excluded count represents stations in the station_zip file that did not have any rides. Because of the inner join, these records will be excluded.

Click the Close button (X) at the top right of the window to return to the data flow view. Right-click the SQL object, and click **Edit.**

Multi-select the fields ZIP_CODE, COUNTY and CITY. Right-click and select **Add to Query,** then **Column.**

Click **Show Profiling,** then scroll over to see the distribution charts and values. You will see different values because of the random sampling, as shown in the following image.

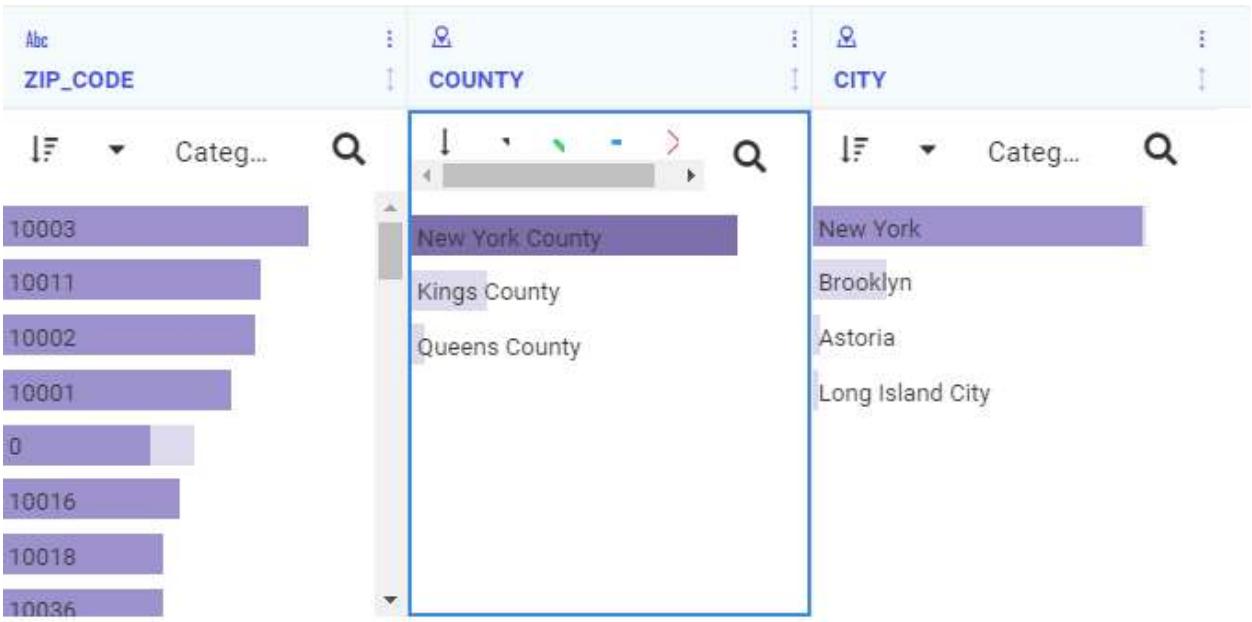


Applying a Filter in a Data Flow

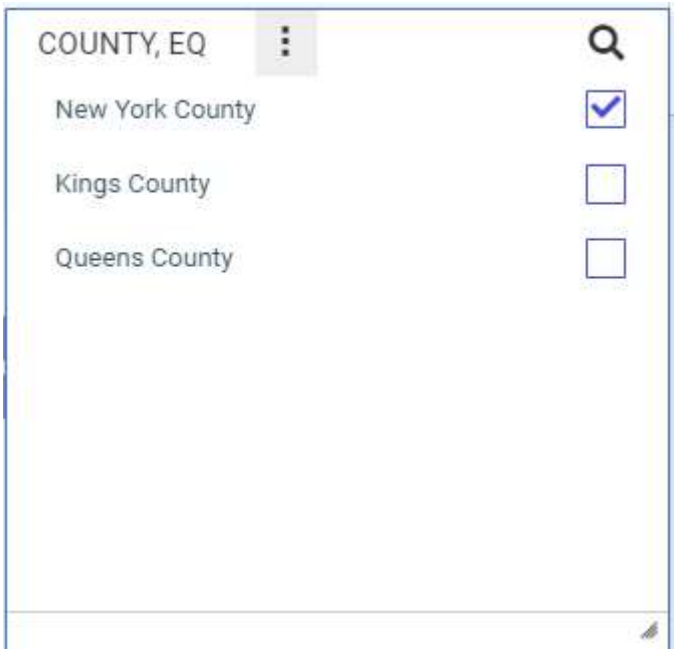
Using Citi Bike trip data, we will restrict our analysis to rides that start in Manhattan, so we can limit the data loaded to just that borough. The geographic data identifies the county, so we will just load the data in New York County.

1. In the COUNTY field, click the bar for New York County.

   The display changes to reflect the selection. The dark portions of the bars in each column show the proportion of rows that are selected, as shown in the following image.



2. Click the green check mark to include the selected rows.
3. A filter card opens, as shown in the following image.



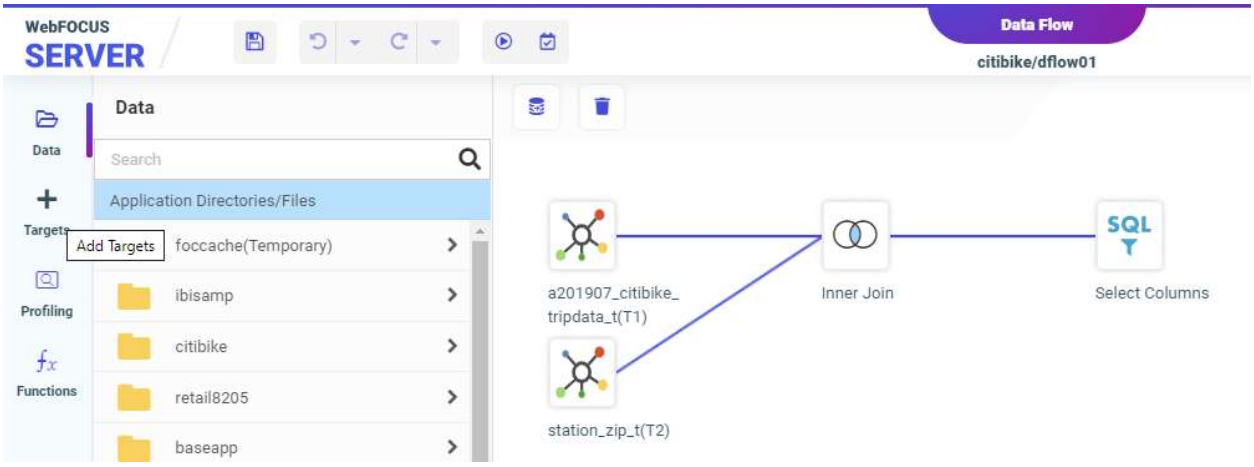
4. Click outside the filter card to close it.

Adding a Target in a Data Flow

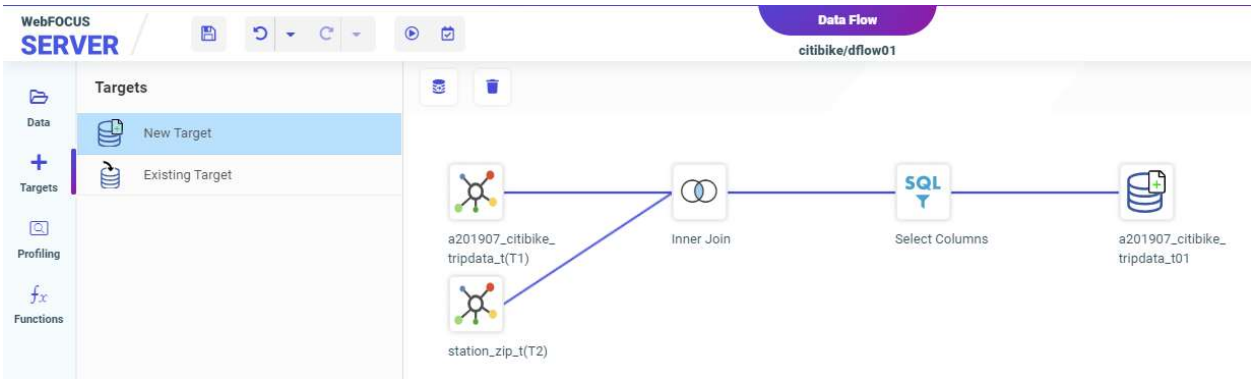
Once a data flow contains all of the fields you need, you can load it to a target data store.

To add a new target:

1. Click **Targets** on the sidebar of the data flow interface, as shown in the following image.

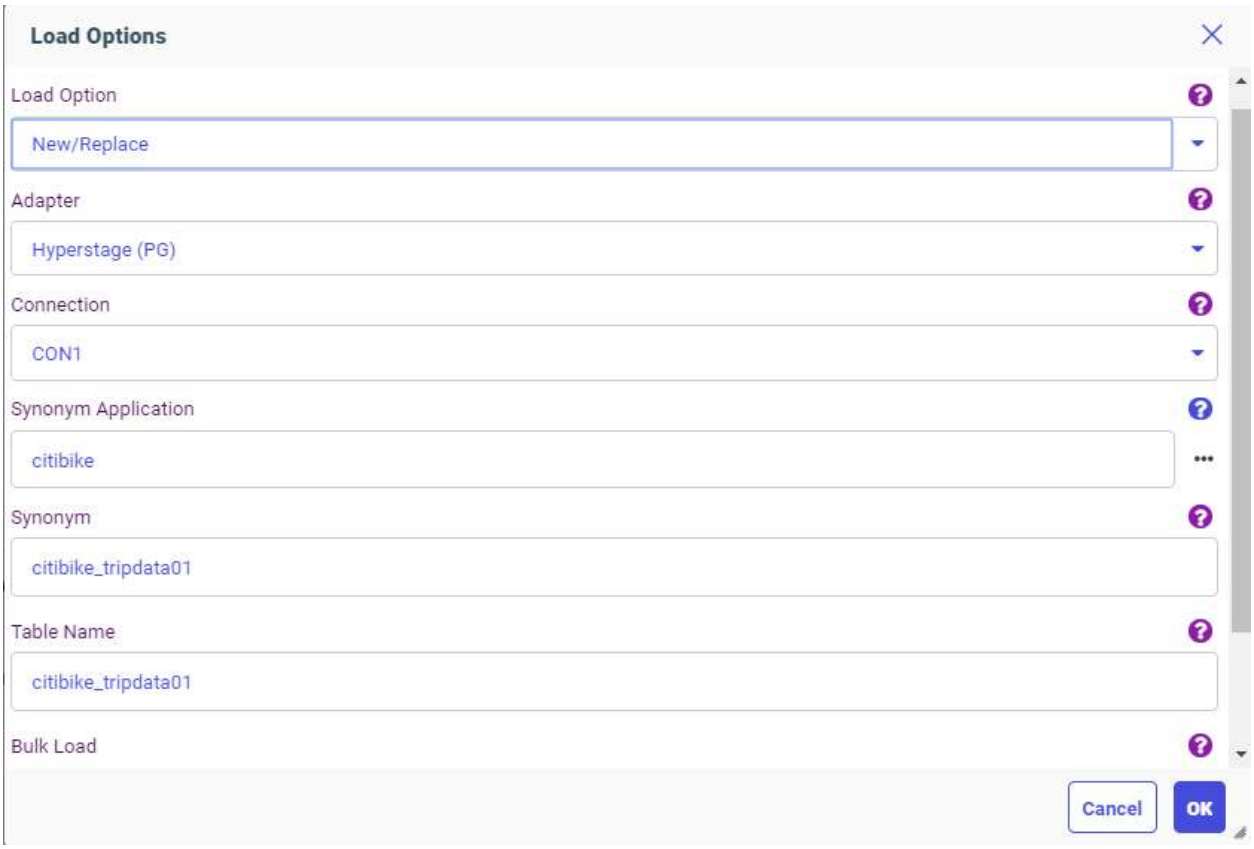


2. Click **New Target** to add a new target to the flow, as shown in the following image.



3. Right-click the target, and select **Load Options** to select or enter the target DBMS, application, synonym and table name.

   For example, the following load options load the data into a Hyperstage table, to an application named citibike, with synonym and table name bike_share_nyc.



4. Click **OK.**

Running a Data Flow

When you have made all of the data preparation transformations for a flow, and have added your target, on the toolbar click the **Run** button ( ).

When the data is loaded, a report similar to the following displays.

```
(ICM18122) Request - citibike/dflow02.fex (Owner: USER1) submitted.
(ICM18741) citibike/citibike_tripdata01 type Hyperstage (PG) New target
(FOC2662) BULK LOAD PROCESS STARTED AT 12.05.04
(FOC2661) TARGET FILE CITIBIKE/CITIBIKE_TRIPDATA01
(ICM18950) Command Invoked to Load Data: C:\ibi\srv82\ibi\srv82\home\hs\bin
\dlp.exe
(ICM18745) Commit forced at: 1000000 for 1000000 row(s)
(ICM18745) Commit forced at: 2000000 for 1000000 row(s)
```

```
(ICM18745) Commit forced at: 2077758 for  77758 row(s)
(FOC2663) BULK LOAD PROCESS ENDED AT 12.07.29, ELAPSED TIME = 00:02:25.219
(FOC2661) TARGET FILE CITIBIKE/CITIBIKE_TRIPDATA01
CITIBIKE/citibike_tripdata01 HELD AS SQLHYPG TABLE
1
0 NUMBER OF RECORDS IN TABLE=  2077758  LINES=2077758
0
(ICM18744) Ending Load
(ICM18040) Return Code = 0
(ICM18076) Request: citibike/dflow02.fex - finished processing
(ICM18007) CPU Time : 106125
```