# Learning objectives

In this tutorial, you will learn how to:

- Load data into the IBM Cloud Pak for Data platform for use with Data Refinery.
- Transform a sample data set, either by entering command-line R code or selecting menu operations.
- Use Data Flow steps to keep track of your work.
- Visualize data with charts and graphs.

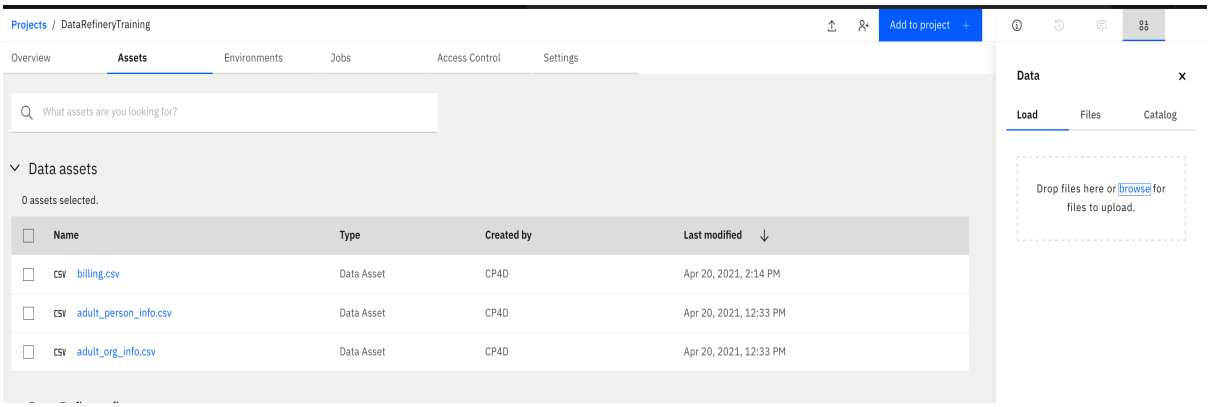
# Estimated time

Completing this tutorial should take about 45 minutes.

# Steps

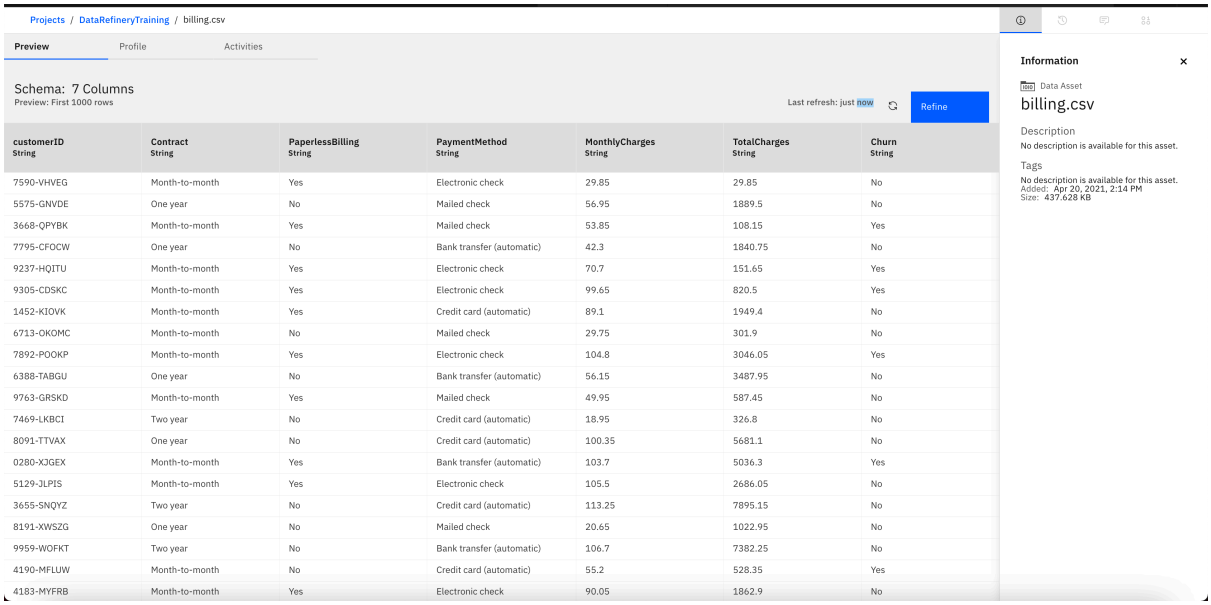
## Step 1. Load the billing.csv data into Data Refinery

1. Download the billing.csv file.
2. From the Project home, click on the **Assets** tab. Next, either drag and drop the downloaded billing.csv file to the right-hand side pane where it says **Drop files here** or browse for files to upload, or click on **browse** and choose the downloaded billing.csv file.



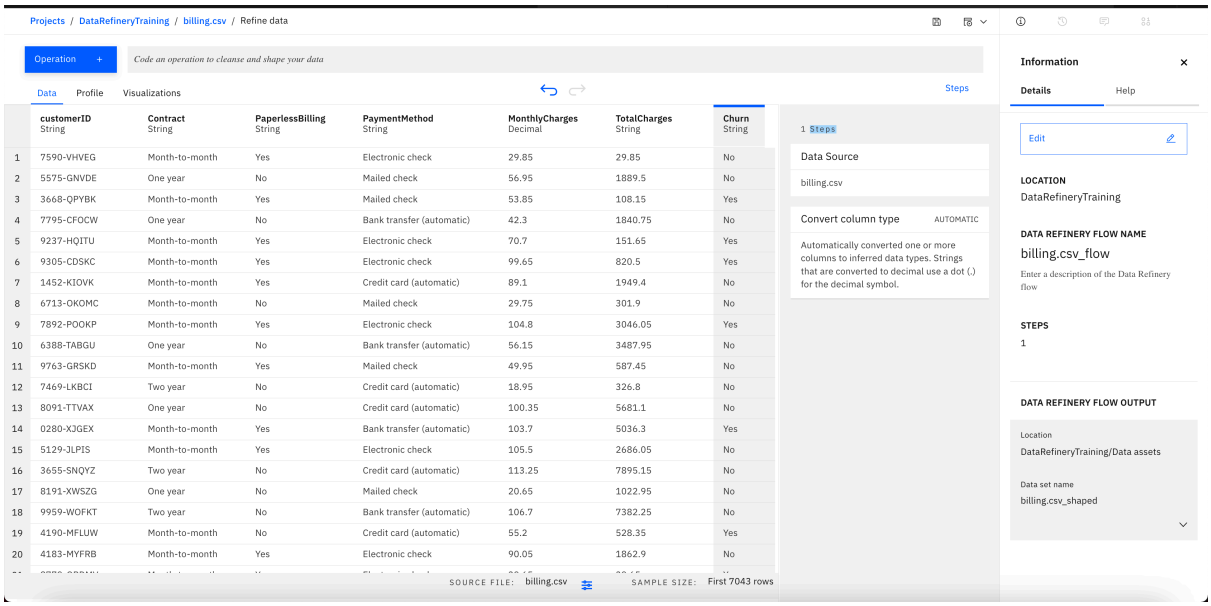

3. Click on the newly added billing.csv file.

4. You should be able to see the data as shown below. Click on **Refine**.




| customerID String | Contract String | PaperlessBilling String | PaymentMethod String | MonthlyCharges String | TotalCharges String | Churn String |
|---|---|---|---|---|---|---|
| 7590-VHVEG | Month-to-month | Yes | Electronic check | 29.85 | 29.85 | No |
| 5575-GNVDE | One year | No | Mailed check | 56.95 | 1889.5 | No |
| 3668-QPYBK | Month-to-month | Yes | Mailed check | 53.85 | 108.15 | Yes |
| 7795-CFOCW | One year | No | Bank transfer (automatic) | 42.3 | 1840.75 | No |
| 9237-HQITU | Month-to-month | Yes | Electronic check | 70.7 | 151.65 | Yes |
| 9305-CDSKC | Month-to-month | Yes | Electronic check | 99.65 | 820.5 | Yes |
| 1452-KIOVK | Month-to-month | Yes | Credit card (automatic) | 89.1 | 1949.4 | No |
| 6713-OKOMC | Month-to-month | No | Mailed check | 29.75 | 301.9 | No |
| 7892-POOKP | Month-to-month | Yes | Electronic check | 104.8 | 3046.05 | Yes |
| 6388-TABGU | One year | No | Bank transfer (automatic) | 56.15 | 3487.95 | No |
| 9763-GRSKD | Month-to-month | Yes | Mailed check | 49.95 | 587.45 | No |
| 7469-LKBCI | Two year | No | Credit card (automatic) | 18.95 | 326.8 | No |
| 8091-TTVAX | One year | No | Credit card (automatic) | 100.35 | 5681.1 | No |
| 0280-XJGEX | Month-to-month | Yes | Bank transfer (automatic) | 103.7 | 5036.3 | Yes |
| 5129-JLPIS | Month-to-month | Yes | Electronic check | 105.5 | 2686.05 | No |
| 3655-SNQYZ | Two year | No | Credit card (automatic) | 113.25 | 7895.15 | No |
| 8191-XWSZG | One year | No | Mailed check | 20.65 | 1022.95 | No |
| 9959-WOFKT | Two year | No | Bank transfer (automatic) | 106.7 | 7382.25 | No |
| 4190-MFLUW | Month-to-month | No | Credit card (automatic) | 55.2 | 528.35 | Yes |
| 4183-MYFRB | Month-to-month | Yes | Electronic check | 90.05 | 1862.9 | No |

5. Data Refinery should launch and open the data.

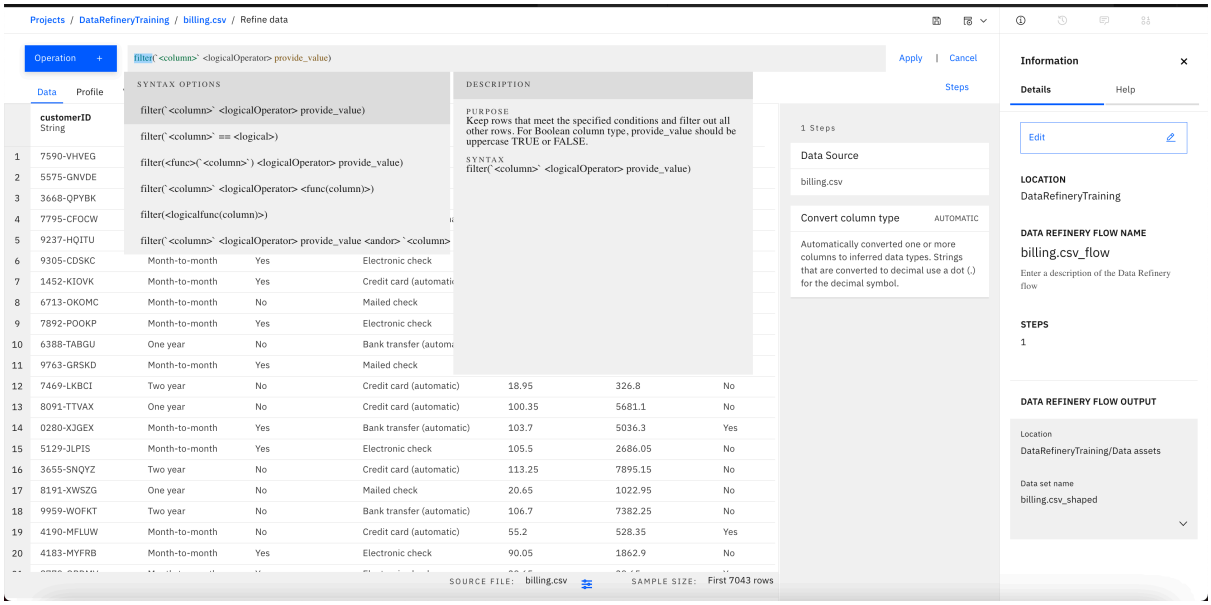



| | customerID String | Contract String | PaperlessBilling String | PaymentMethod String | MonthlyCharges Decimal | TotalCharges String | Churn String |
|---|---|---|---|---|---|---|---|
| 1 | 7590-VHVEG | Month-to-month | Yes | Electronic check | 29.85 | 29.85 | No |
| 2 | 5575-GNVDE | One year | No | Mailed check | 56.95 | 1889.5 | No |
| 3 | 3668-QPYBK | Month-to-month | Yes | Mailed check | 53.85 | 108.15 | Yes |
| 4 | 7795-CFOCW | One year | No | Bank transfer (automatic) | 42.3 | 1840.75 | No |
| 5 | 9237-HQITU | Month-to-month | Yes | Electronic check | 70.7 | 151.65 | Yes |
| 6 | 9305-CDSKC | Month-to-month | Yes | Electronic check | 99.65 | 820.5 | Yes |
| 7 | 1452-KIOVK | Month-to-month | Yes | Credit card (automatic) | 89.1 | 1949.4 | No |
| 8 | 6713-OKOMC | Month-to-month | No | Mailed check | 29.75 | 301.9 | No |
| 9 | 7892-POOKP | Month-to-month | Yes | Electronic check | 104.8 | 3046.05 | Yes |
| 10 | 6388-TABGU | One year | No | Bank transfer (automatic) | 56.15 | 3487.95 | No |
| 11 | 9763-GRSKD | Month-to-month | Yes | Mailed check | 49.95 | 587.45 | No |
| 12 | 7469-LKBCI | Two year | No | Credit card (automatic) | 18.95 | 326.8 | No |
| 13 | 8091-TTVAX | One year | No | Credit card (automatic) | 100.35 | 5681.1 | No |
| 14 | 0280-XJGEX | Month-to-month | Yes | Bank transfer (automatic) | 103.7 | 5036.3 | Yes |
| 15 | 5129-JLPIS | Month-to-month | Yes | Electronic check | 105.5 | 2686.05 | No |
| 16 | 3655-SNQYZ | Two year | No | Credit card (automatic) | 113.25 | 7895.15 | No |
| 17 | 8191-XWSZG | One year | No | Mailed check | 20.65 | 1022.95 | No |
| 18 | 9959-WOFKT | Two year | No | Bank transfer (automatic) | 106.7 | 7382.25 | No |
| 19 | 4190-MFLUW | Month-to-month | No | Credit card (automatic) | 55.2 | 528.35 | Yes |
| 20 | 4183-MYFRB | Month-to-month | Yes | Electronic check | 90.05 | 1862.9 | No |

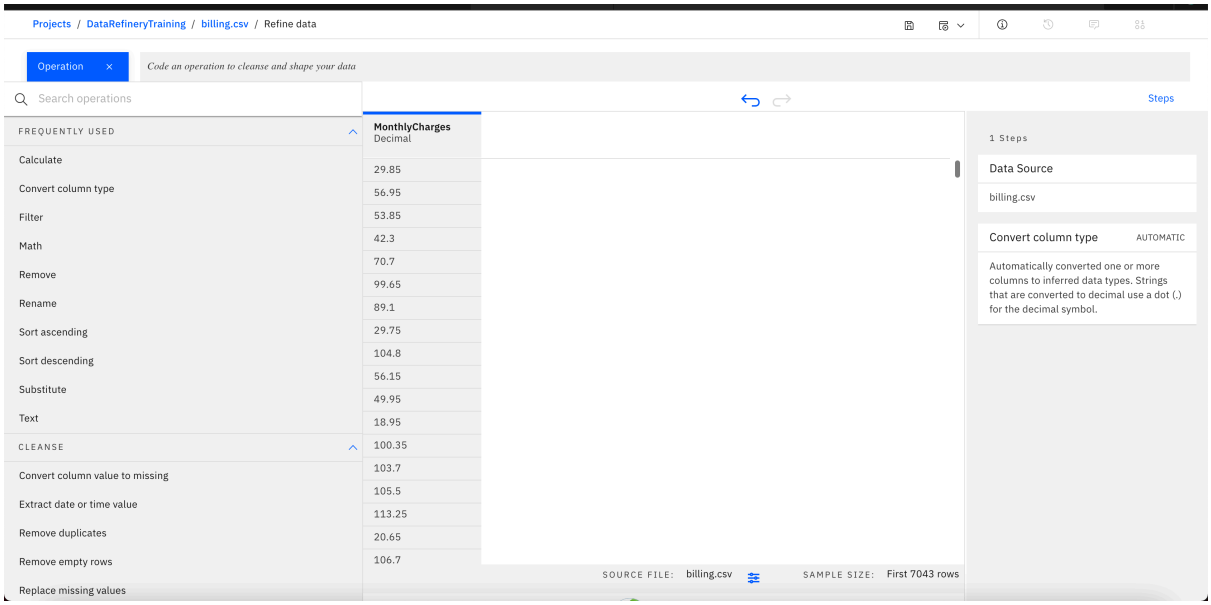
6. Click the **X** on the **Information** page to close it.

## Step 2. Refine your data

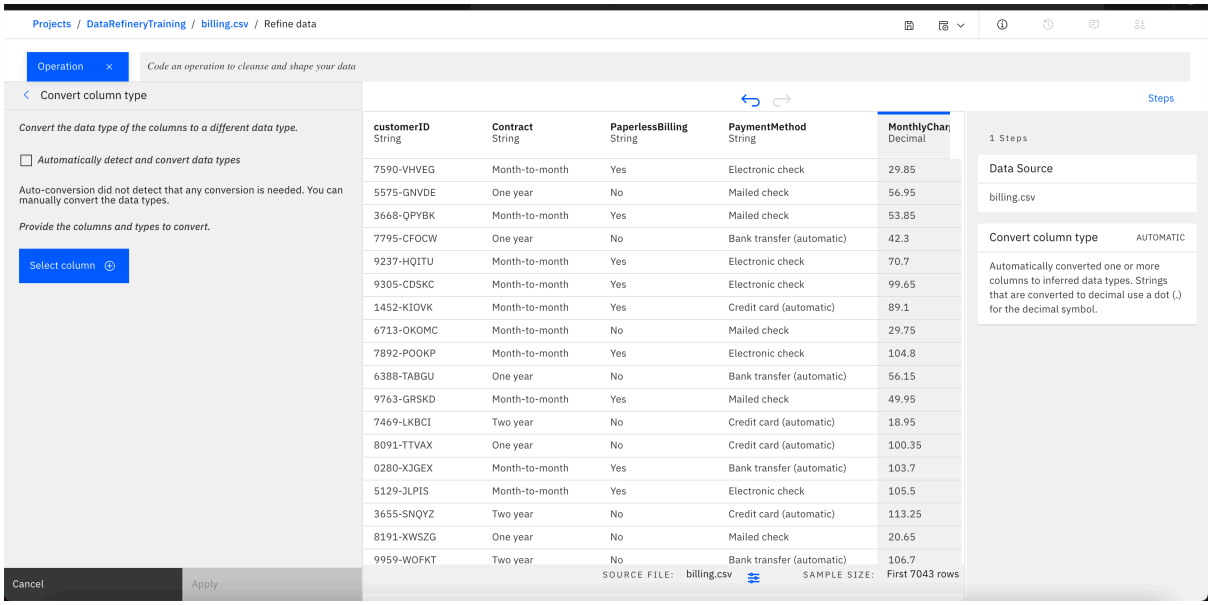
We'll start out on the Data tab. Transform your sample data set by entering R code in the command line or selecting operations from the menu. For example, type filter on the command line and observe that autocomplete will give hints on the syntax and how to use the command.



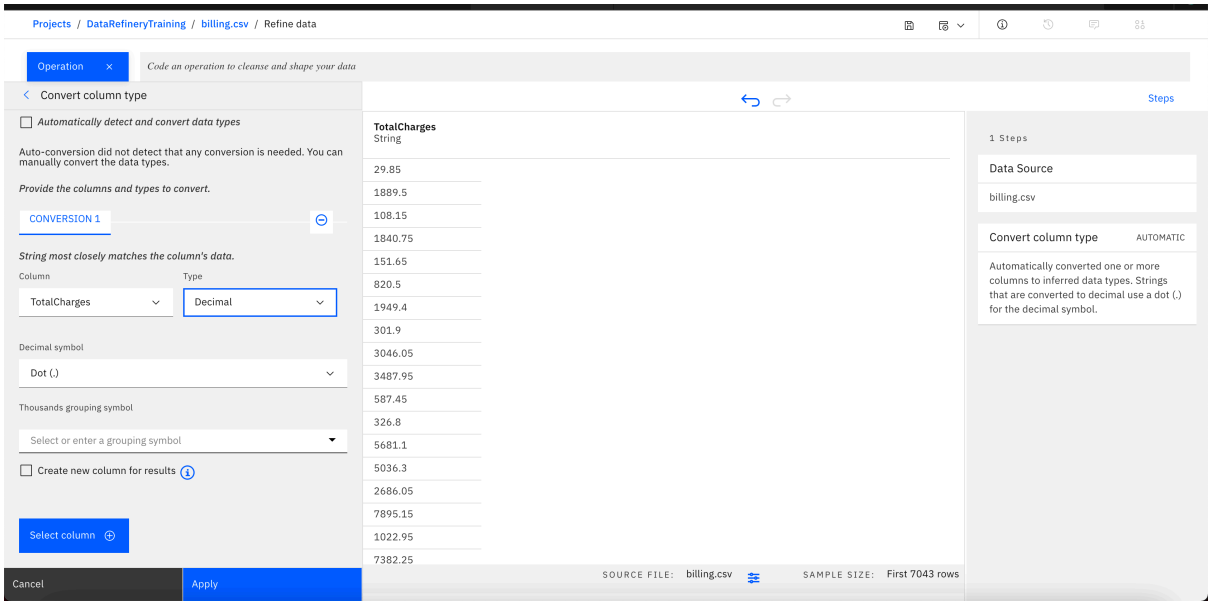

Alternatively, hover over an operation or function name to see a description and detailed information for completing the command. When you're ready, click **Apply** to apply the operation to your data set, then click the **+Operation** button.



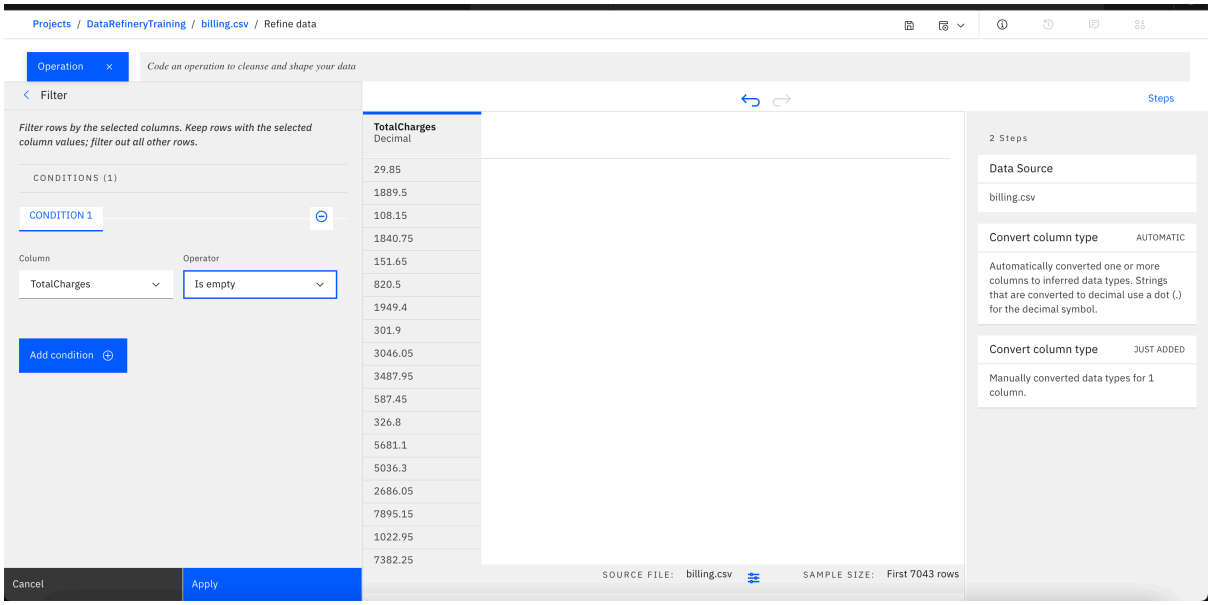

We notice that TotalCharges is a string, but since it represents a decimal number, let's convert the values to decimal. Choose the Operator **Convert Column Type**.



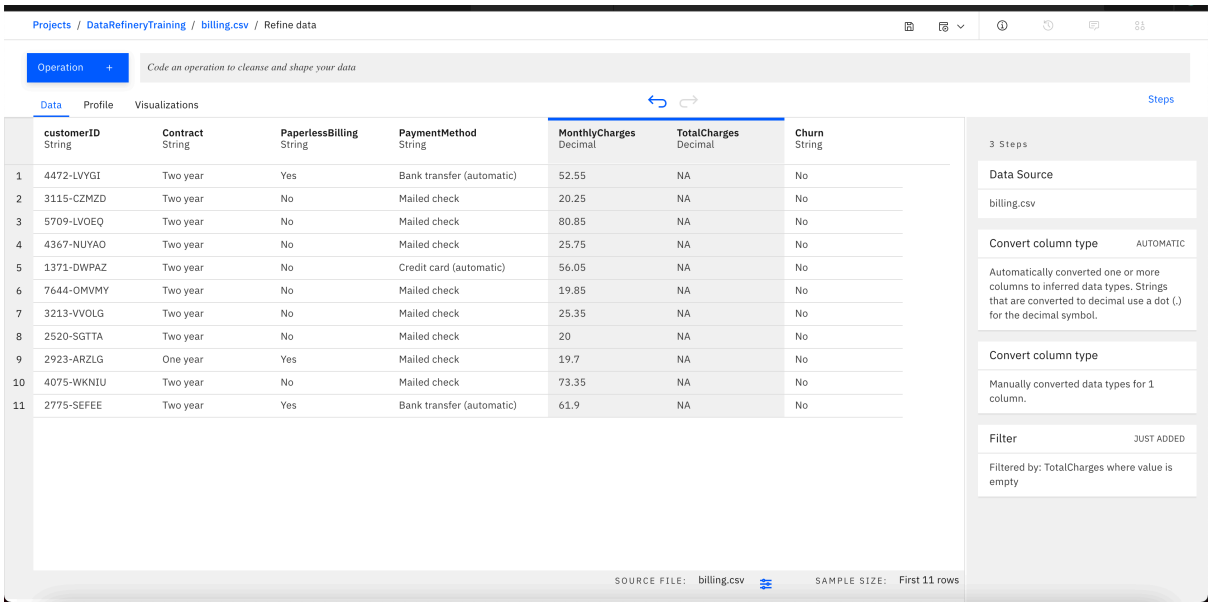

Click **+ Select column**, then pick **Column > TotalCharges** and **Type > Decimal**, then click **Apply**.




We want to make sure that there are no empty values, and there happen to be some for the TotalCharges column, so let's fix that. Click on the operation **Filter** and choose the **TotalCharges** column from the drop-down, operator **Is empty**, then **Apply**.



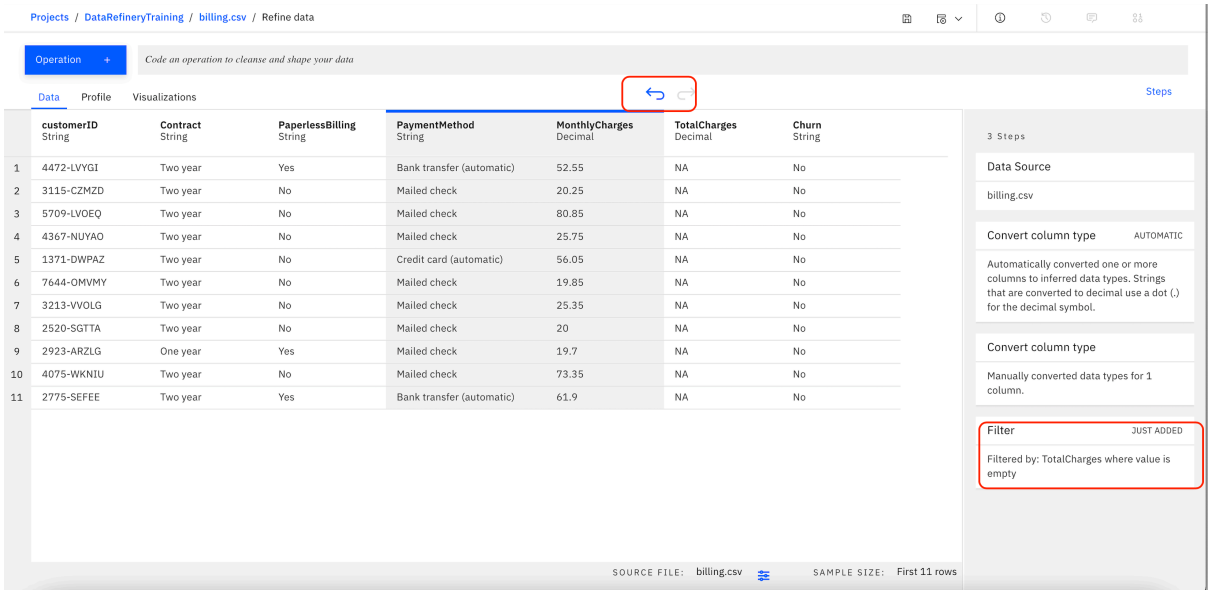

We can see that there are only 11 rows with an empty value for TotalCharges.



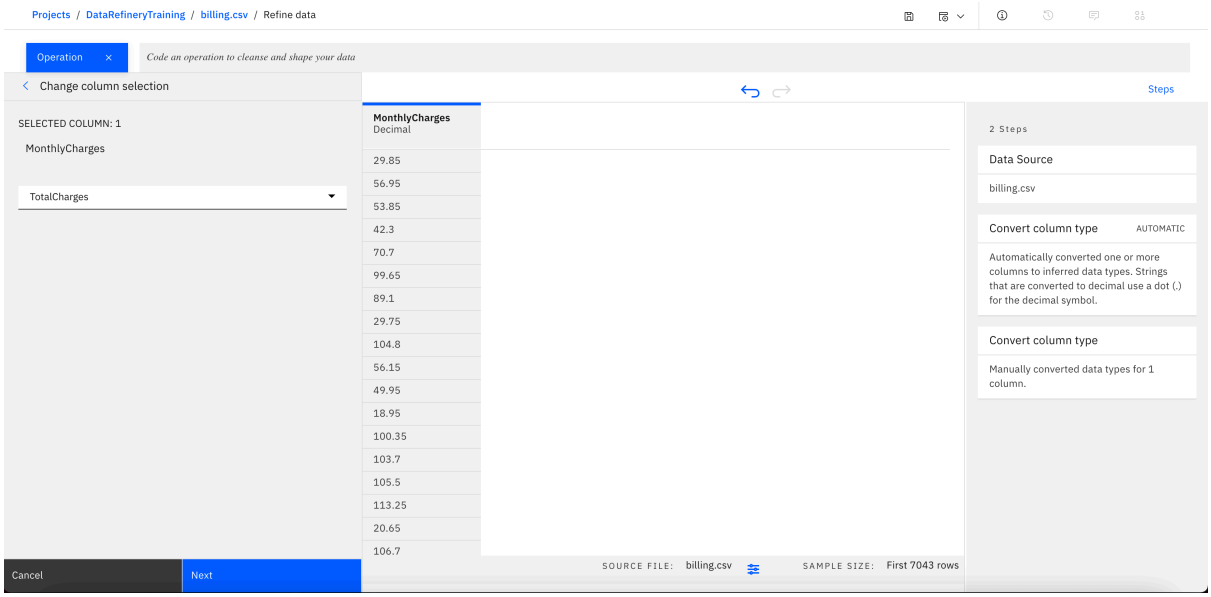

It should be safe to just drop these rows from the data set, so let's do that.

Remove the filter you just added. You can delete it using one of the following methods:

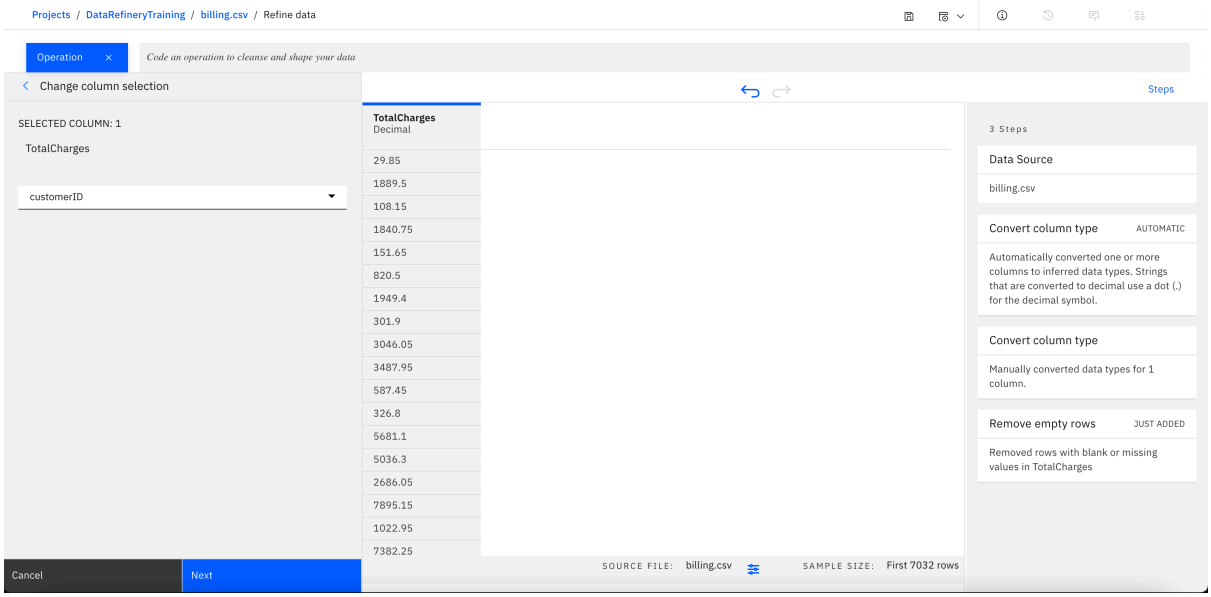
- Hover over the corresponding step in the Steps section and the delete icon (trash can) will appear. Click on this icon to remove the filter.
- Click the undo arrow at the top of the page.




Next, choose the operation **Remove empty rows**, select the TotalCharges column, click **Next**, then click **Apply** on the next screen.

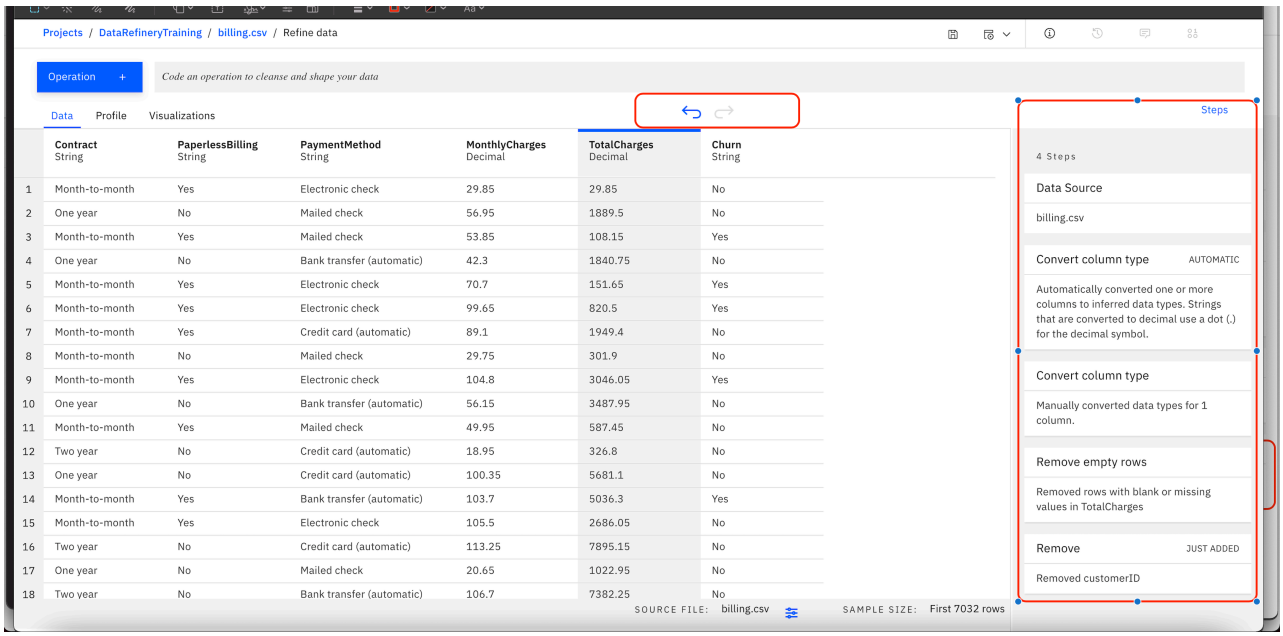



Also, we can remove the CustomerID column, since that may not be useful for training a machine learning model. Choose the **Remove** operator, then **Change column selection**. Under **Select a column**, pick **CustomerID**, then **Next**, then **Apply**.




## Step 3. Use data flow steps to keep track of your work

What if we do something we don't want? Data Refinery keeps track of the steps and we can undo (or redo) an action using the arrows.




| | Contract String | PaperlessBilling String | PaymentMethod String | MonthlyCharges Decimal | TotalCharges Decimal | Churn String |
|---|---|---|---|---|---|---|
| 1 | Month-to-month | Yes | Electronic check | 29.85 | 29.85 | No |
| 2 | One year | No | Mailed check | 56.95 | 1889.5 | No |
| 3 | Month-to-month | Yes | Mailed check | 53.85 | 108.15 | Yes |
| 4 | One year | No | Bank transfer (automatic) | 42.3 | 1840.75 | No |
| 5 | Month-to-month | Yes | Electronic check | 70.7 | 151.65 | Yes |
| 6 | Month-to-month | Yes | Electronic check | 99.65 | 820.5 | Yes |
| 7 | Month-to-month | Yes | Credit card (automatic) | 89.1 | 1949.4 | No |
| 8 | Month-to-month | No | Mailed check | 29.75 | 301.9 | No |
| 9 | Month-to-month | Yes | Electronic check | 104.8 | 3046.05 | Yes |
| 10 | One year | No | Bank transfer (automatic) | 56.15 | 3487.95 | No |
| 11 | Month-to-month | Yes | Mailed check | 49.95 | 587.45 | No |
| 12 | Two year | No | Credit card (automatic) | 18.95 | 326.8 | No |
| 13 | One year | No | Credit card (automatic) | 100.35 | 5681.1 | No |
| 14 | Month-to-month | Yes | Bank transfer (automatic) | 103.7 | 5036.3 | Yes |
| 15 | Month-to-month | Yes | Electronic check | 105.5 | 2686.05 | No |
| 16 | Two year | No | Credit card (automatic) | 113.25 | 7895.15 | No |
| 17 | One year | No | Mailed check | 20.65 | 1022.95 | No |
| 18 | Two year | No | Bank transfer (automatic) | 106.7 | 7382.25 | No |

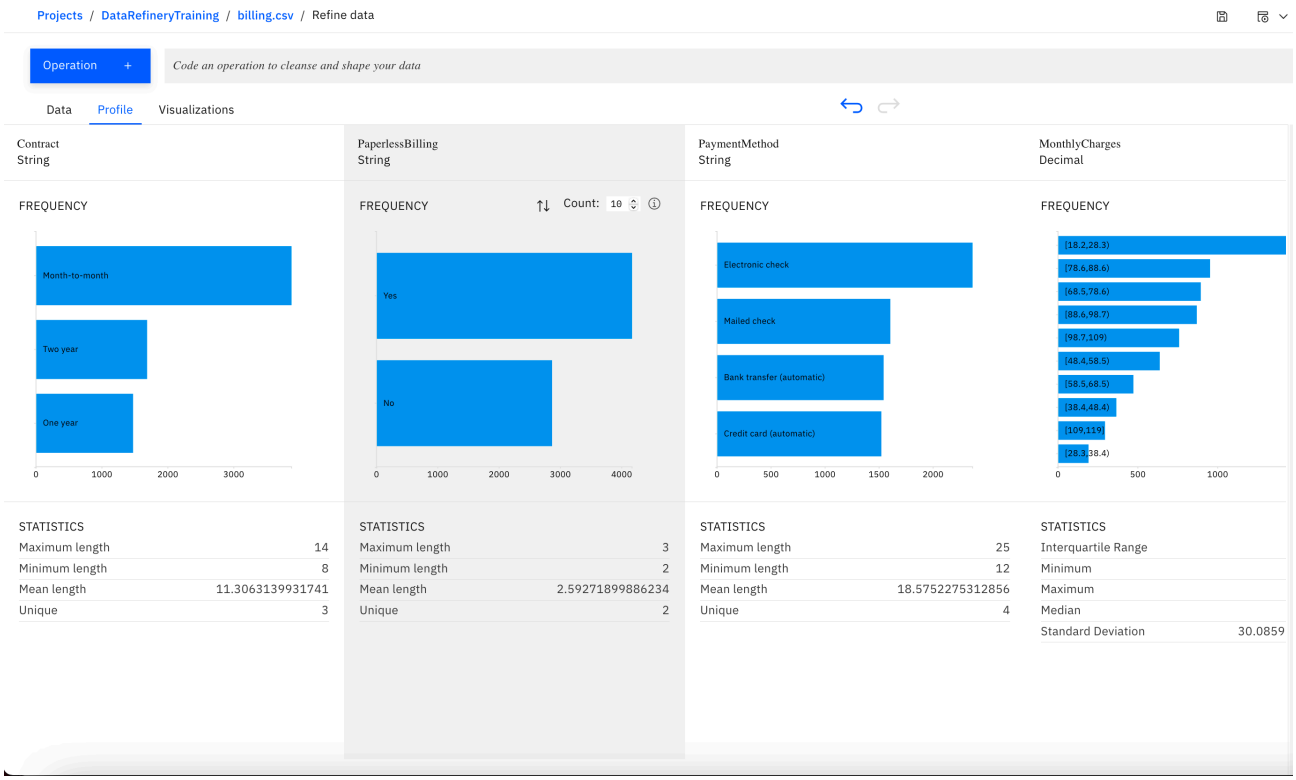
As you refine your data, the IBM Data Refinery keeps track of the steps in your data flow. You can modify them and even select a step to return to a particular moment in your data's transformation.

To see the steps in the data flow that you have performed, click the **Steps** button. The operations you have performed on the data will be shown.

You can modify these steps in real time and save for future use.

## Step 4. Profile the data

Clicking on the **Profile** tab will bring up a quick view of several histograms about the data.

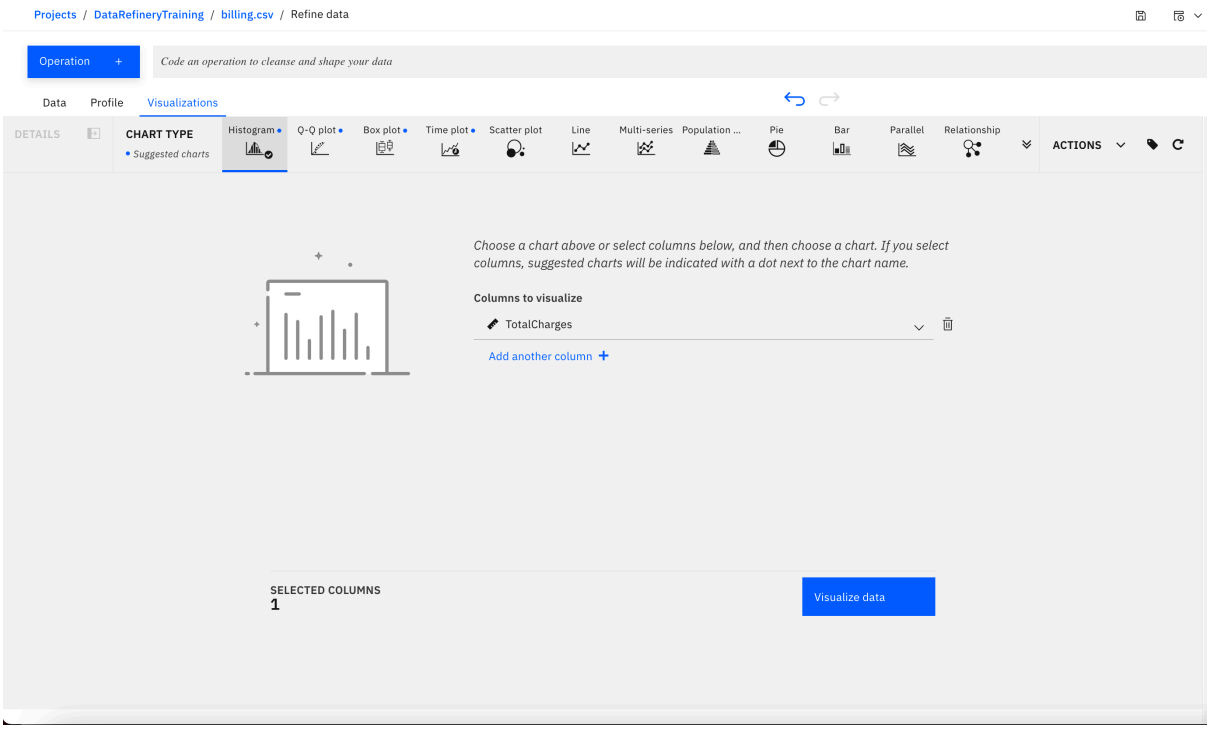



You can get insights into the data from the histograms:

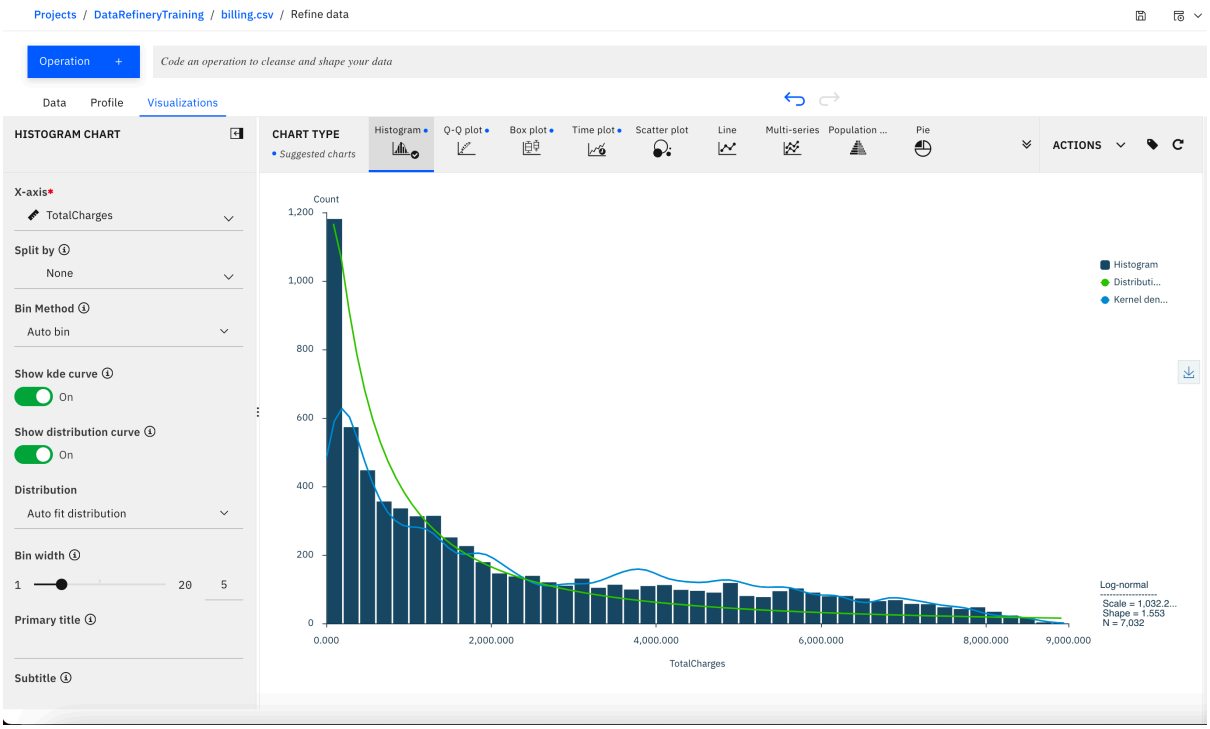
- Twice as many customers are month to month as are a one- or two-year contract.
- More choose paperless billing, but around 40 percent still prefer a paper bill sent to them.
- You can see the distribution of MonthlyCharges and TotalCharges.
- From the Churn column, you can see that a significant number of customers will cancel their service.

## Step 5. Visualize with charts and graphs

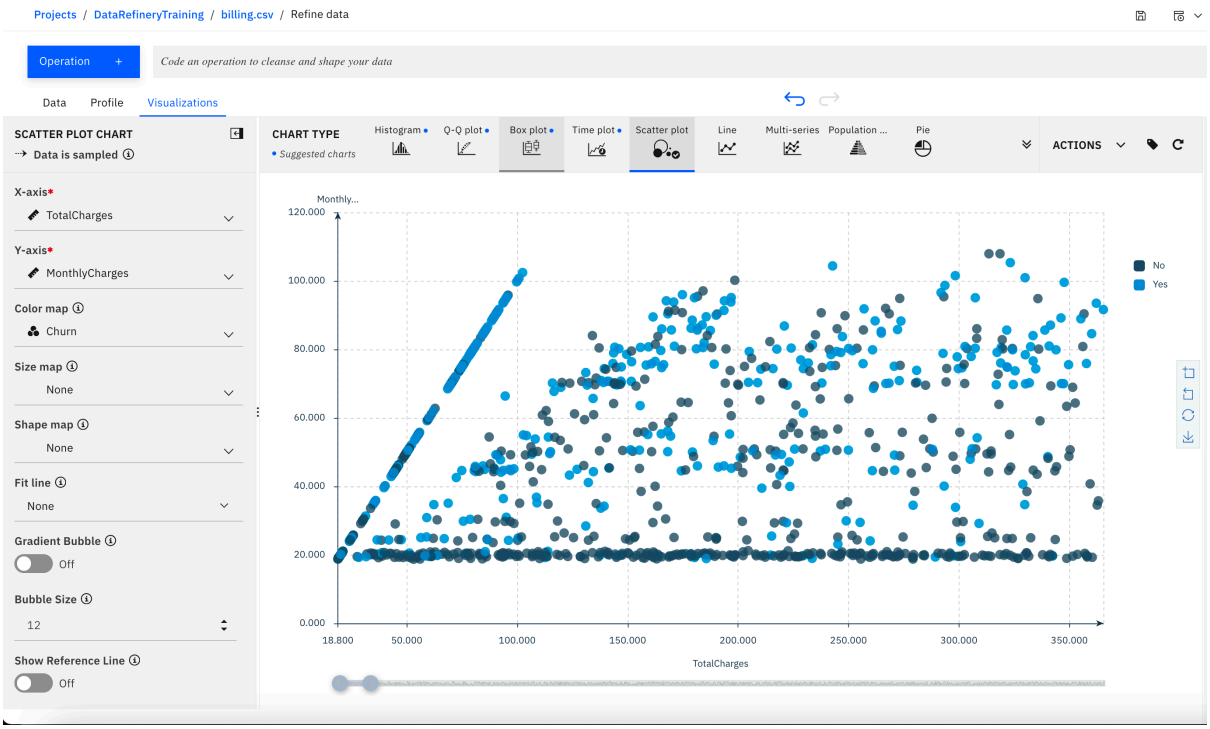
1. Choose the **Visualizations** tab to bring up an option to choose which columns to visualize. Under **Columns to Visualize**, choose **TotalCharges** and click **Visualize data**.



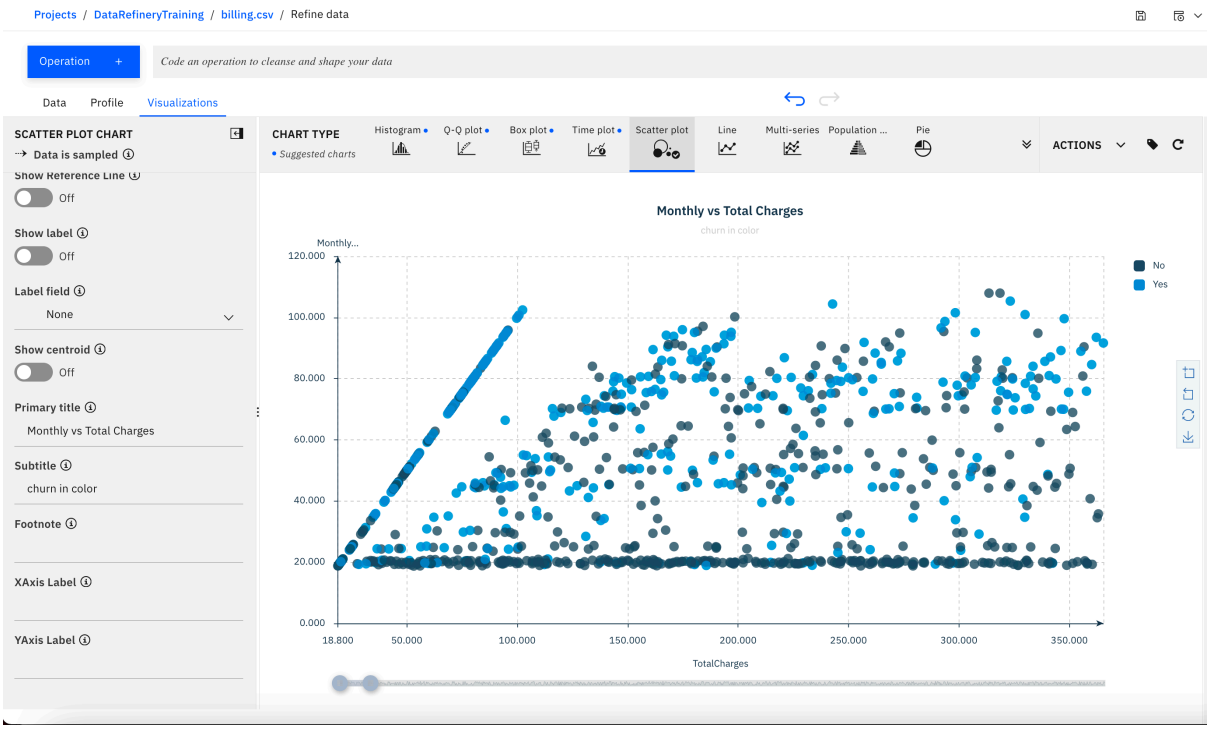

2. We first see the data in a histogram by default. You can choose other chart types. We'll pick Scatter plot next by clicking on it.

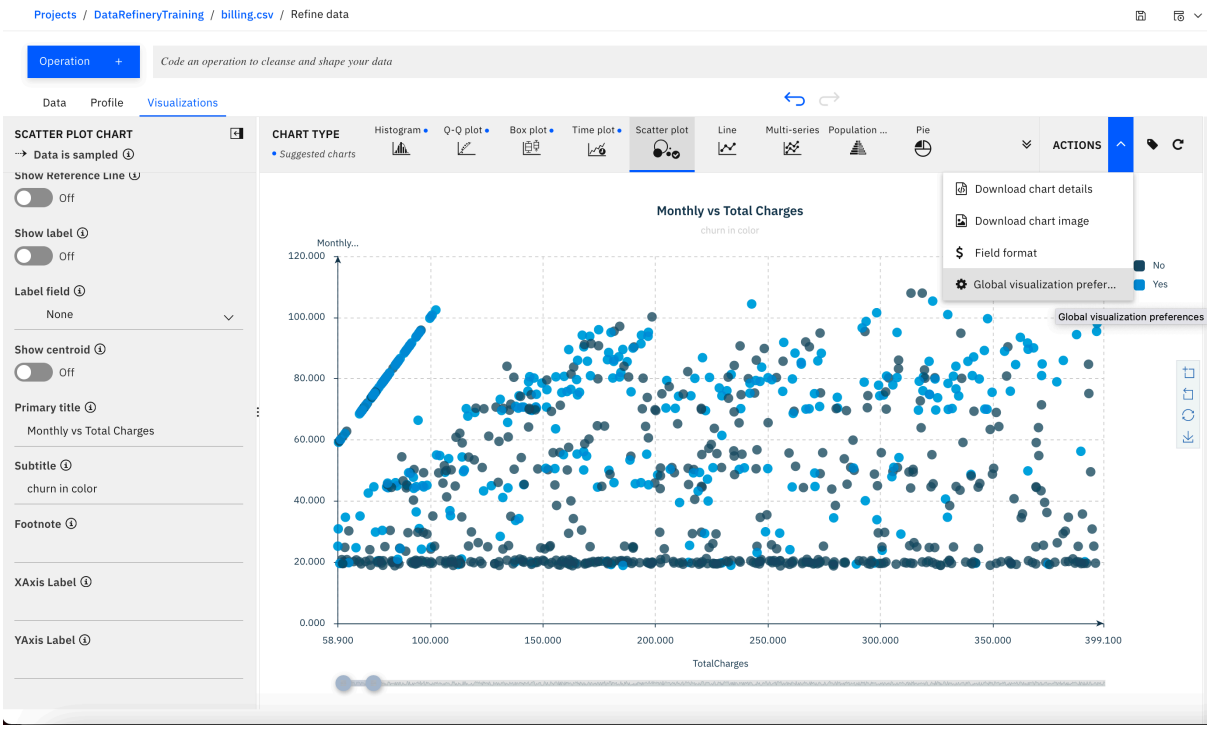


3. In the scatter plot, choose **TotalCharges** for the x-axis, **MonthlyCharges** for the y-axis, and **Churn** for the color map.




4. Scroll down and give the scatter plot a title and sub-title if you wish. Under the **Actions** panel, notice that you can perform tasks such as start over, download chart details, display data label in chart, download chart image, or set global visualization preferences (hover over the icons to see the names). Click on the gear icon in the **Actions** panel.






5. We see that we can do things in the global visualization preferences for titles, tools, color schemes, and notifications. Click on the **Theme** tab, update the color scheme to **Vivid**, then click **Apply**.

   Now the colours for all of our charts will be reflected.