

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.

The screenshot shows the Project home page with the 'Assets' tab selected. The left pane displays a table of assets, including three CSV files: 'CSV billing.csv', 'CSV adult_person_info.csv', and 'CSV adult_org_info.csv'. The right pane has a 'Data' section with a 'Load' tab, which includes a dashed box for dragging and dropping files to upload.

Name	Type	Created by	Last modified
CSV billing.csv	Data Asset	CP4D	Apr 20, 2021, 2:14 PM
CSV adult_person_info.csv	Data Asset	CP4D	Apr 20, 2021, 12:33 PM
CSV adult_org_info.csv	Data Asset	CP4D	Apr 20, 2021, 12:33 PM

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

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

The screenshot shows the Data Refinery interface with a CSV file named 'billing.csv'. The file has 7 columns: customerID, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, and Churn. The 'Refine' button is highlighted in blue at the top right of the preview area.

5. Data Refinery should launch and open the data.

The screenshot shows the Data Refinery interface with the 'Refine' tab selected. The data preview is identical to the previous screenshot. On the right, there is a 'Steps' panel showing a single step: 'Convert column type AUTOMATIC'.

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.

The screenshot shows the Data Refinery interface for refining a CSV file named 'billing.csv'. On the left, a list of operations is shown, including 'filter', 'convert column type', and 'remove duplicates'. The main area displays a preview of the data with columns such as 'customerID', 'Month-to-month', 'Yes', 'Electronic check', and 'Credit card (automatic)'. A tooltip for the 'filter' operation is visible, providing syntax and purpose. The right side shows the 'Information' tab with details like 'Data Source: billing.csv', 'Data Refinery Flow Name: billing.csv_flow', and 'Steps: 1'. The 'Edit' button is highlighted.

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.

The screenshot shows the Data Refinery interface with the 'Operation' tab selected. The left panel lists frequently used operations such as 'Calculate', 'Convert column type', 'Filter', 'Math', etc. The 'MonthlyCharges' column is selected, showing its current type as 'String'. The right side shows the 'Information' tab with details like 'Data Source: billing.csv', 'Data Refinery Flow Name: billing.csv_flow', and 'Steps: 1'. The 'Edit' button is highlighted.

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**.

Projects / DataRefineryTraining / billing.csv / Refine data

Operation Code an operation to cleanse and shape your data

< Convert column type

Convert the data type of the columns to a different data type.

Automatically detect and convert data types

Auto-conversion did not detect that any conversion is needed. You can manually convert the data types.

Provide the columns and types to convert.

Select column

customerID	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges
7590-VHVEG	Month-to-month	Yes	Electronic check	29.85
5575-GNVDE	One year	No	Mailed check	56.95
3668-QPYBK	Month-to-month	Yes	Mailed check	53.85
7795-CFOCW	One year	No	Bank transfer (automatic)	42.3
9237-HQJTU	Month-to-month	Yes	Electronic check	70.7
9305-CDSKC	Month-to-month	Yes	Electronic check	99.65
1452-KIOVK	Month-to-month	Yes	Credit card (automatic)	89.1
6713-OKOMC	Month-to-month	No	Mailed check	29.75
7892-POOKP	Month-to-month	Yes	Electronic check	104.8
6388-TABGU	One year	No	Bank transfer (automatic)	56.15
9763-GRSKD	Month-to-month	Yes	Mailed check	49.95
7469-LKBCI	Two year	No	Credit card (automatic)	18.95
8091-TTVAX	One year	No	Credit card (automatic)	100.35
0280-XJGEX	Month-to-month	Yes	Bank transfer (automatic)	103.7
5129-JLPIS	Month-to-month	Yes	Electronic check	105.5
3655-SNQYZ	Two year	No	Credit card (automatic)	113.25
8191-XWSZG	One year	No	Mailed check	20.65
9959-WOFKT	Two year	No	Bank transfer (automatic)	106.7

SOURCE FILE: billing.csv SAMPLE SIZE: First 7043 rows

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

Projects / DataRefineryTraining / billing.csv / Refine data

Operation Code an operation to cleanse and shape your data

< Convert column type

Automatically detect and convert data types

Auto-conversion did not detect that any conversion is needed. You can manually convert the data types.

Provide the columns and types to convert.

CONVERSION 1

String most closely matches the column's data.

Column	Type
TotalCharges	Decimal

Decimal symbol

Thousands grouping symbol

Create new column for results

Select column

TotalCharges
29.85
1889.5
108.15
1840.75
151.65
820.5
1949.4
301.9
3046.05
3487.95
587.45
326.8
5681.1
5036.3
2686.05
7895.15
1022.95
7382.25

SOURCE FILE: billing.csv SAMPLE SIZE: First 7043 rows

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.

The screenshot shows the Data Refinery interface with a 'Filter' operation selected. The 'TotalCharges' column is being filtered for empty values. The results list 11 rows with empty values. The 'Steps' section shows two steps: 'Data Source' (billing.csv) and 'Convert column type' (AUTOMATIC). The 'Convert column type' step notes that it has automatically converted one or more columns to inferred data types, specifically mentioning strings converted to decimal.

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

The screenshot shows the Data Refinery interface with the filtered dataset. The 'Data' tab is selected, showing 11 rows. The 'Steps' section shows a 'Filter' step with the condition 'Filtered by: TotalCharges where value is empty'.

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.

The screenshot shows the Data Refinery interface with a table of data and its processing steps.

Data View:

	customerID	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
1	4472-LVYGI	Two year	Yes	Bank transfer (automatic)	52.55	NA	No
2	3115-CZMZD	Two year	No	Mailed check	20.25	NA	No
3	5709-LVOEQ	Two year	No	Mailed check	80.85	NA	No
4	4367-NUYAO	Two year	No	Mailed check	25.75	NA	No
5	1371-DWPAZ	Two year	No	Credit card (automatic)	56.05	NA	No
6	7644-OMVMY	Two year	No	Mailed check	19.85	NA	No
7	3213-VVOLG	Two year	No	Mailed check	25.35	NA	No
8	2520-SGTTA	Two year	No	Mailed check	20	NA	No
9	2923-ARZLG	One year	Yes	Mailed check	19.7	NA	No
10	4075-WKNIU	Two year	No	Mailed check	73.35	NA	No
11	2775-SEFEE	Two year	Yes	Bank transfer (automatic)	61.9	NA	No

Processing Steps:

- 3 Steps
- Data Source: billing.csv
- Convert column type: AUTOMATIC (Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol.)
- Convert column type: Manually converted data types for 1 column.
- Filter: JUST ADDED (Filtered by: TotalCharges where value is empty)

SOURCE FILE: billing.csv SAMPLE SIZE: First 11 rows

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

The screenshot shows the Data Refinery interface with the 'Change column selection' step selected.

Selected Column: TotalCharges

Processing Steps:

- 2 Steps
- Data Source: billing.csv
- Convert column type: AUTOMATIC (Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol.)
- Convert column type: Manually converted data types for 1 column.

SOURCE FILE: billing.csv SAMPLE SIZE: First 7043 rows

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**.

Projects / DataRefineryTraining / billing.csv / Refine data

Operation < Change column selection

SELECTED COLUMN: 1
TotalCharges

customerID	TotalCharges
	29.85
	1889.5
	108.15
	1840.75
	151.65
	820.5
	1949.4
	301.9
	3046.05
	3487.95
	587.45
	326.8
	5681.1
	5036.3
	2686.05
	7895.15
	1022.95
	7382.25

SOURCE FILE: billing.csv SAMPLE SIZE: First 7032 rows

Cancel Next

Steps

- 3 Steps
- Data Source
- billing.csv
- Convert column type AUTOMATIC
- Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol.
- Convert column type
- Manually converted data types for 1 column.
- Remove empty rows JUST ADDED
- Removed rows with blank or missing values in TotalCharges

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.

Projects / DataRefineryTraining / billing.csv / Refine data

Operation + Code an operation to cleanse and shape your data

Data Profile Visualizations

Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
String	String	String	Decimal	Decimal	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

SOURCE FILE: billing.csv SAMPLE SIZE: First 7032 rows

Steps

- 4 Steps
- Data Source
- billing.csv
- Convert column type AUTOMATIC
- Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol.
- Convert column type
- Manually converted data types for 1 column.
- Remove JUST ADDED
- Removed customerID

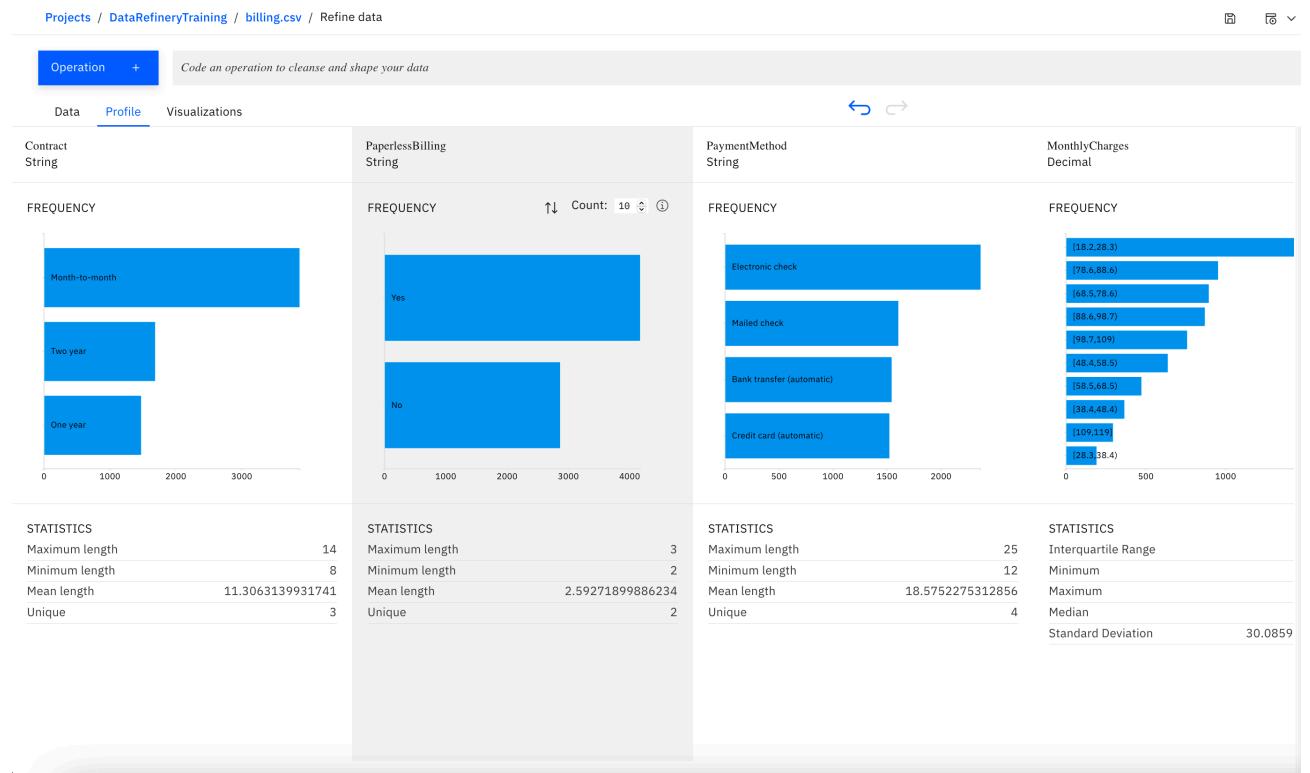
As you refine your data, the IBM Data Refinery keeps track of the steps in your data flow. You can modify them and even select a step to return to a particular moment in your data's transformation.

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.

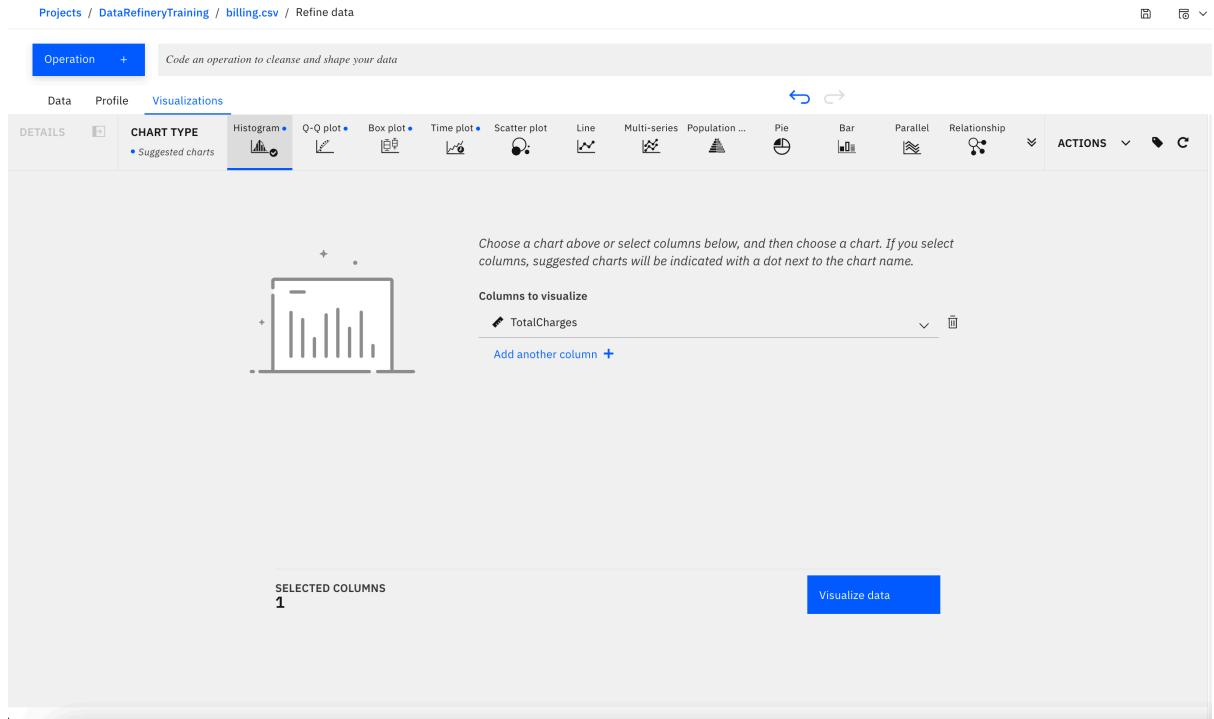


You can get insights into the data from the histograms:

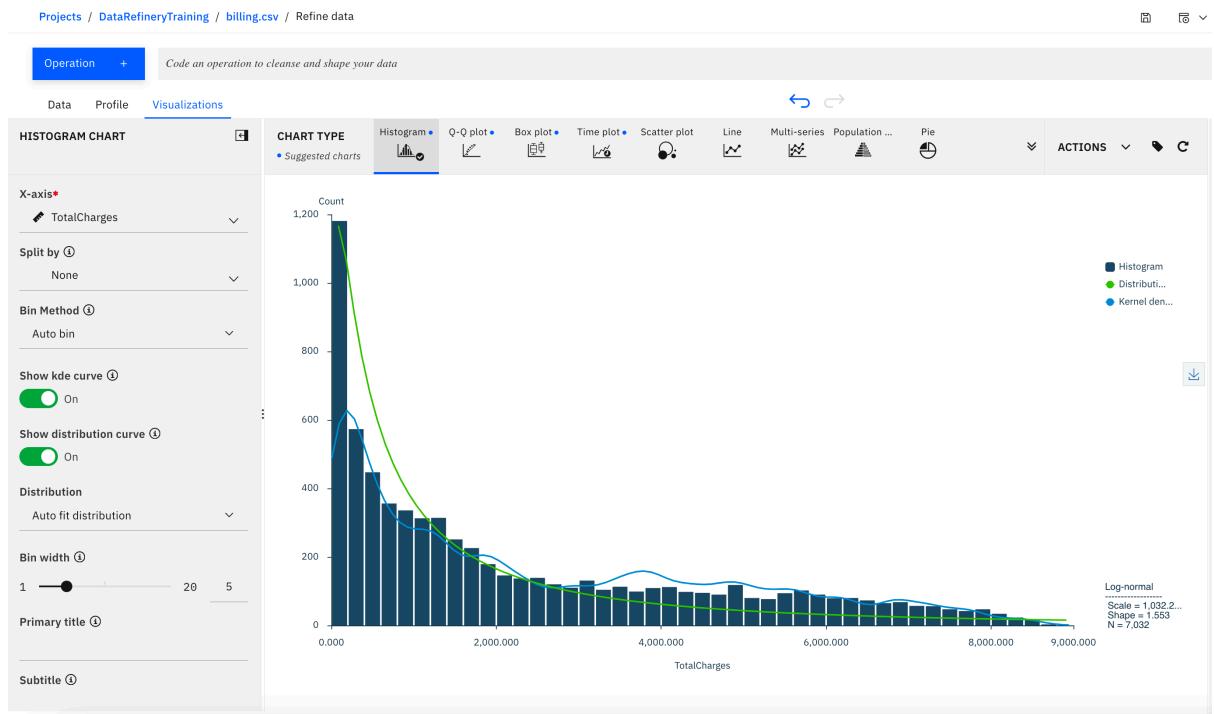
- Twice as many customers are month to month as are a one- or two-year contract.
- More choose paperless billing, but around 40 percent still prefer a paper bill sent to them.
- You can see the distribution of MonthlyCharges and TotalCharges.
- From the Churn column, you can see that a significant number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

Step 5. Visualize with charts and graphs

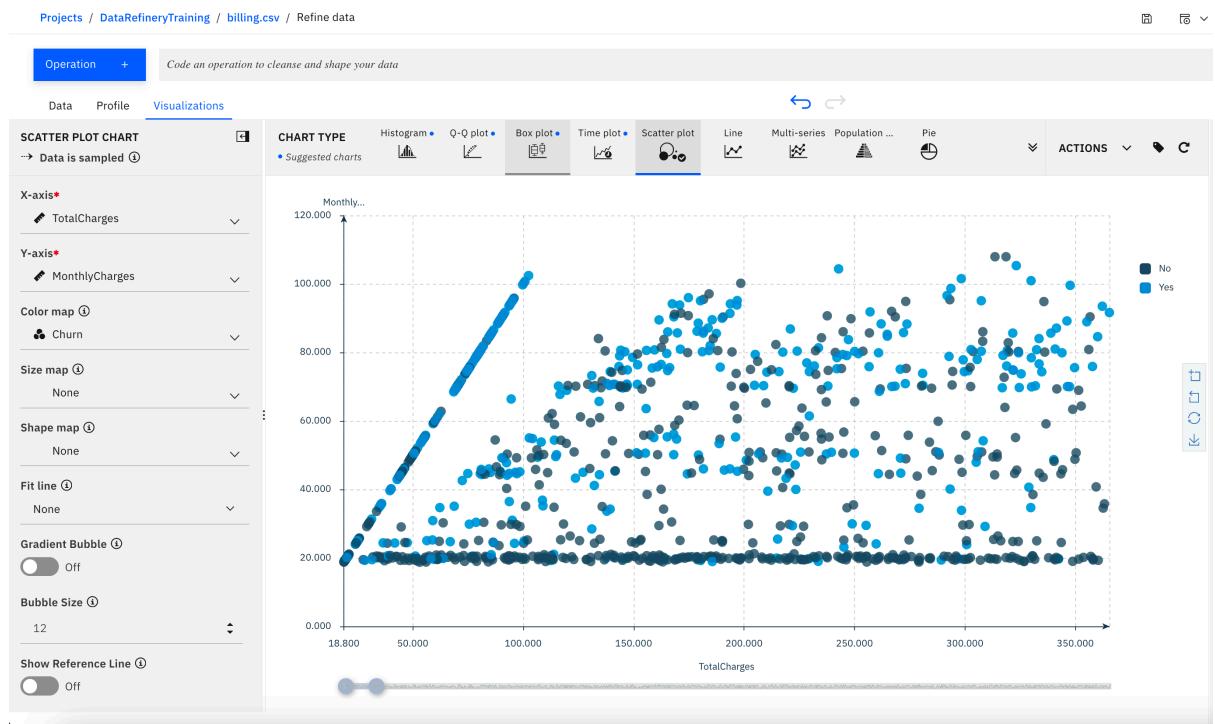
1. Choose the **Visualizations** tab to bring up an option to choose which columns to visualize. Under **Columns to Visualize**, choose **TotalCharges** and click **Visualize data**.



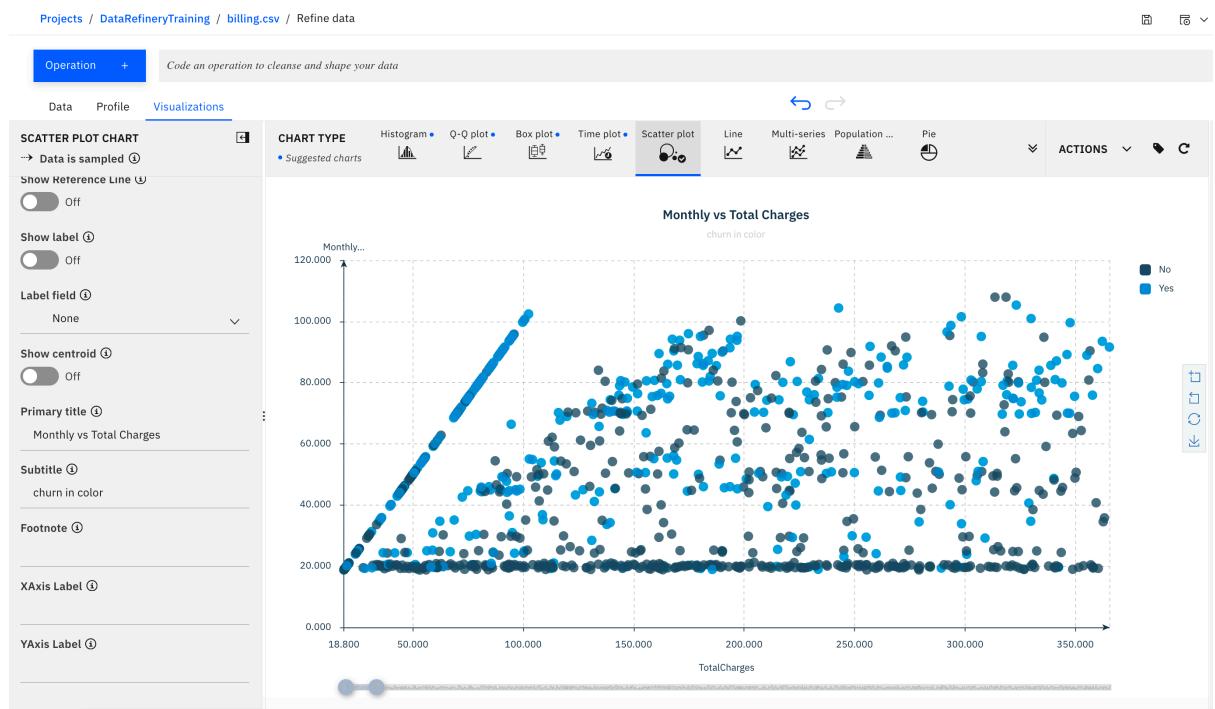
2. We first see the data in a histogram by default. You can choose other chart types. We'll pick Scatter plot next by clicking on it.

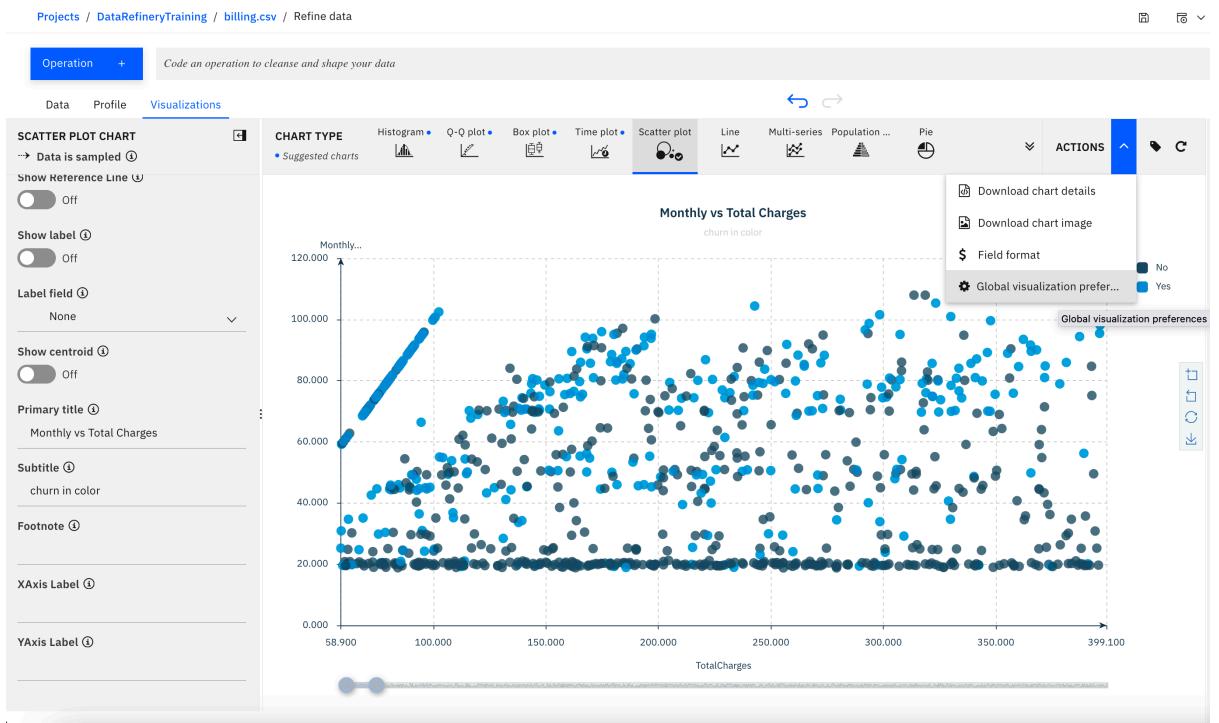


3. In the scatter plot, choose **TotalCharges** for the x-axis, **MonthlyCharges** for the y-axis, and **Churn** for the color map.



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.





- 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.

