Lab Guide

Hands-on-Lab:
Data visualization with data refinery

Shivam R Solanki
Data Scientist
Shivam.raj.solanki@ibm.com

IBM

Data refinery is part of IBM Watson® and comes with IBM Watson Studio on the IBM Public Cloud, and IBM Watson Knowledge Catalog running on-premises using IBM Cloud Pak® for Data. It's a self-service data-preparation client for data scientists, data engineers, and business analysts. With it, you can quickly transform large amounts of raw data into quality consumable information that's ready for analytics. Data refinery makes it easy to explore, prepare, and deliver data that people across your organization can trust.

Learning objectives

In this lab tutorial, you will learn how to:

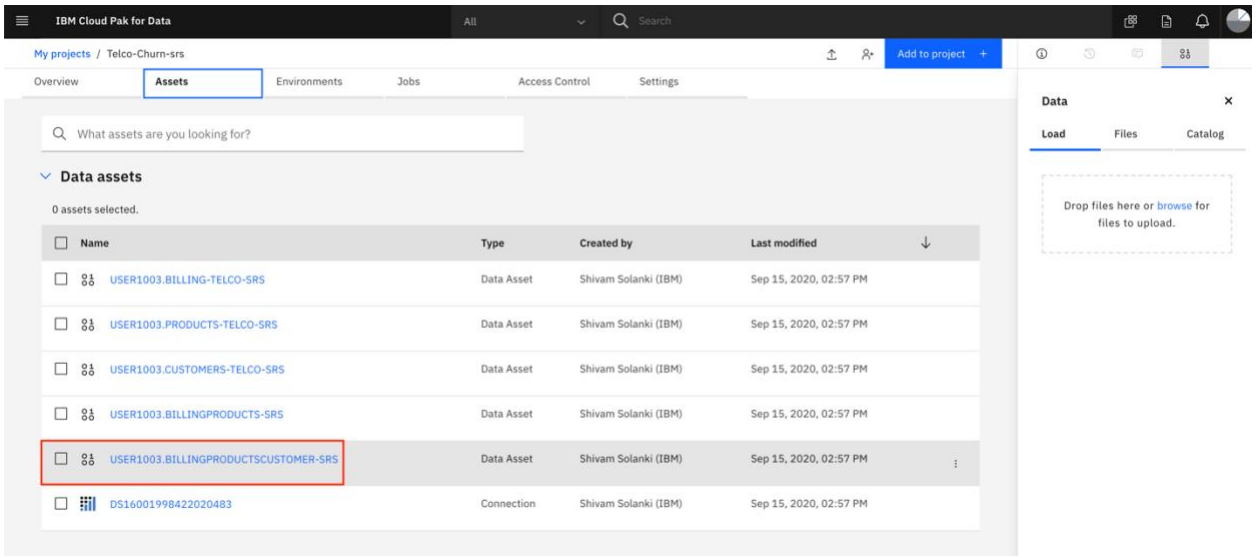
- Load data into the IBM Cloud Pak for Data platform for use with data refinery.
- Transform a sample data set
- Use Data Flow steps to keep track of your work.
- Quickly profile data
- Visualize the data with charts and graphs
- Save the data refinery flow and create a job

Steps

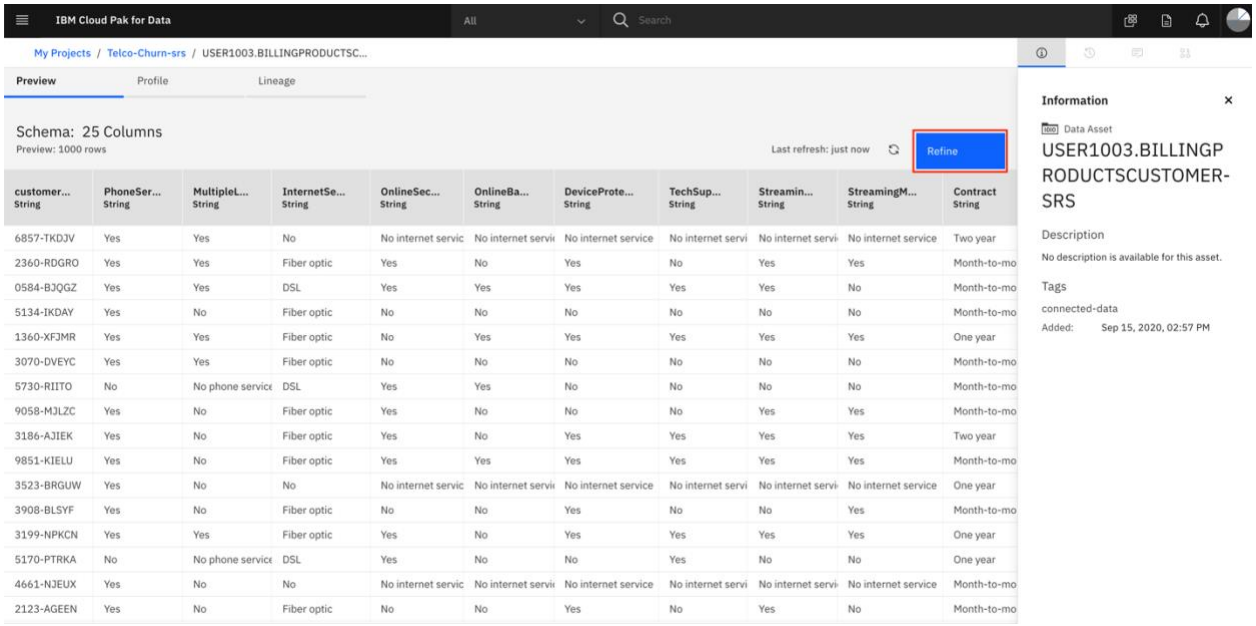
Step 1. Load the virtualized data into data refinery

1. If you are not already on your Project **Assets** tab from the last lab tutorial on Data Virtualization, please to your Project that you created earlier and then click on the **Assets** tab.

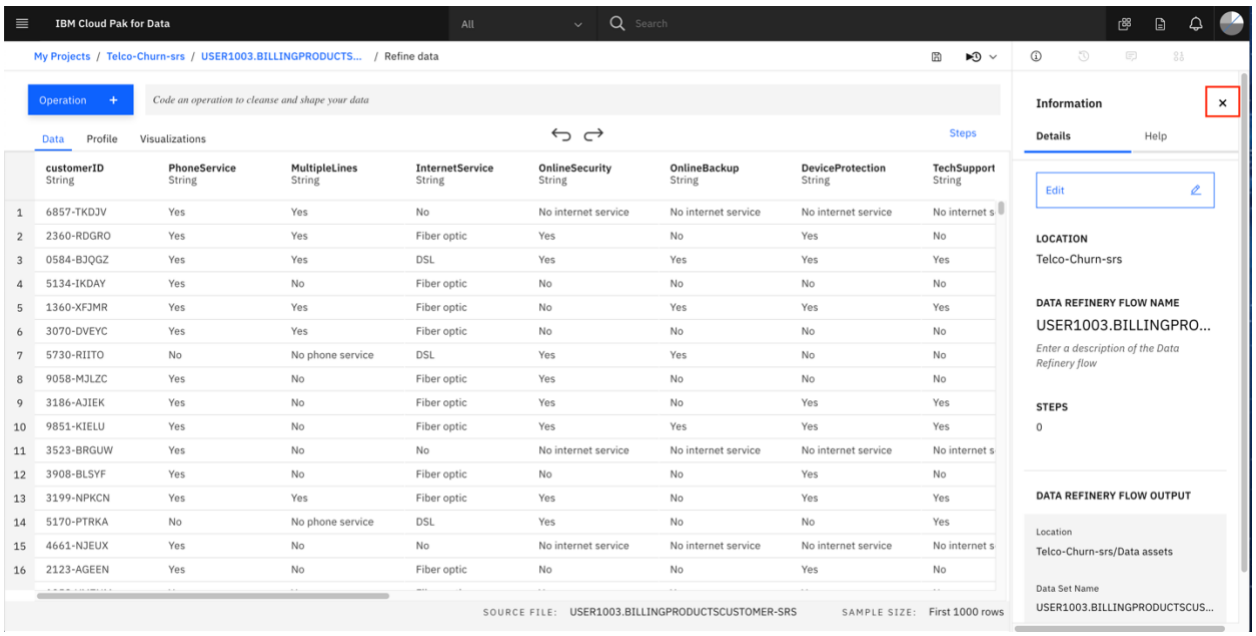
2. From the **Assets** tab, select the Data Asset that contains the combined table BILLING, PRODUCTS and CUSTOMERS created in the previous tutorial



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



4. Data refinery should launch and open the data. Click on **Maybe Later** and close the modal.

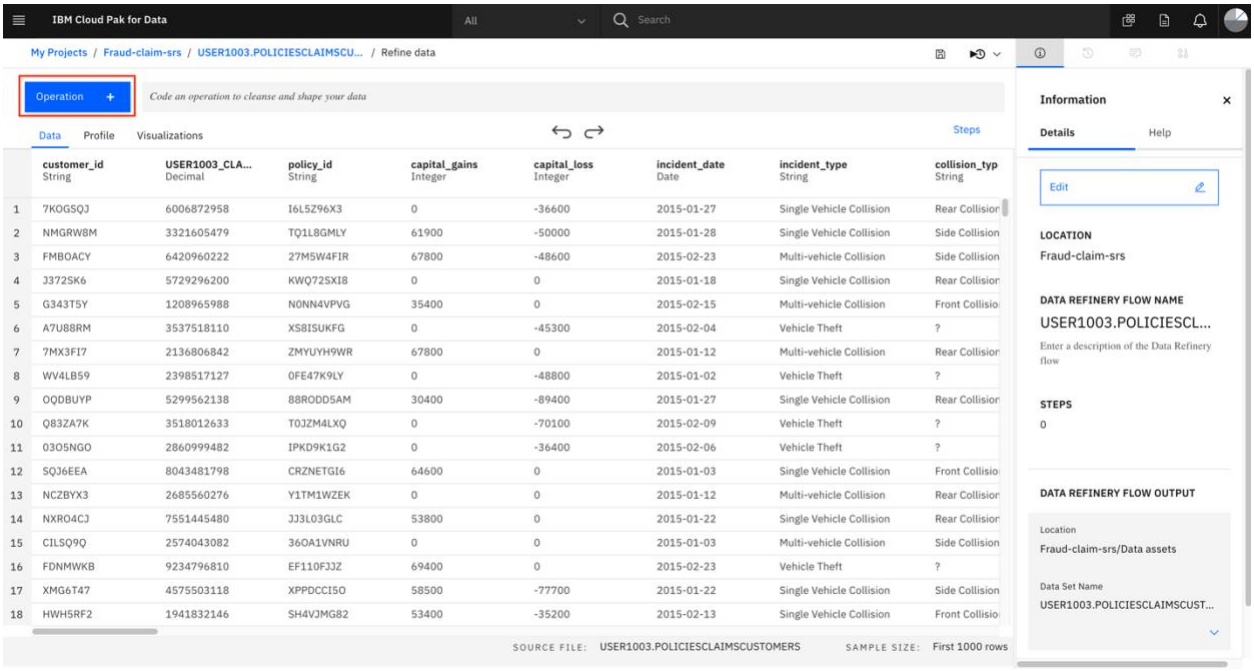


5. Click the **X** by the **Details** button to close it.

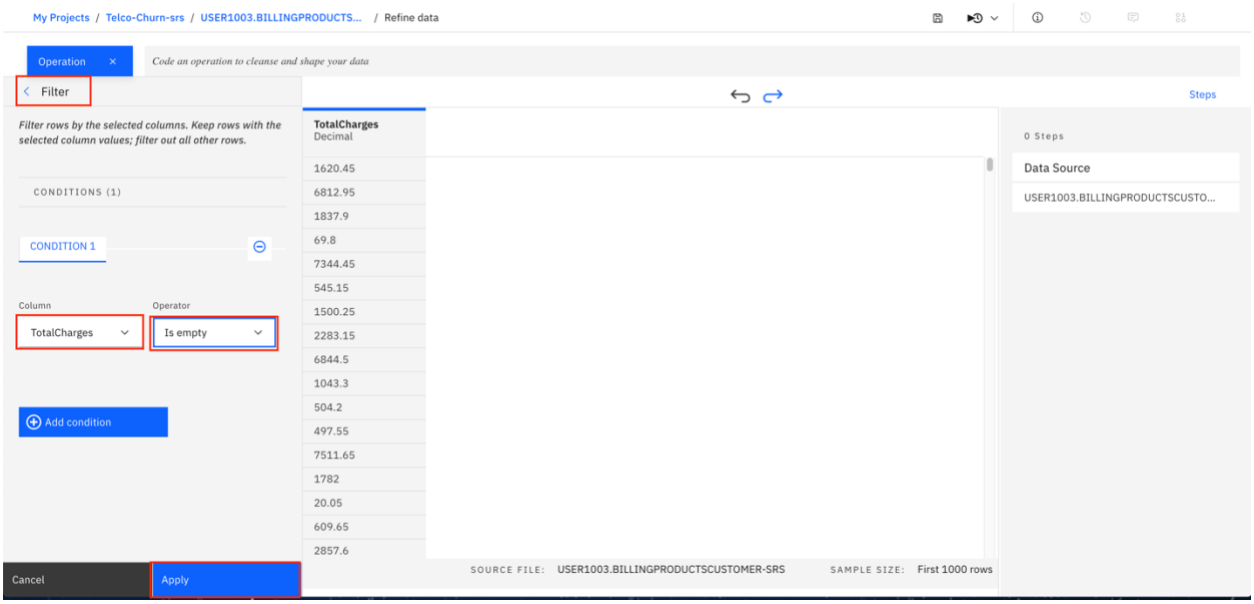
Step 2. Refine your data

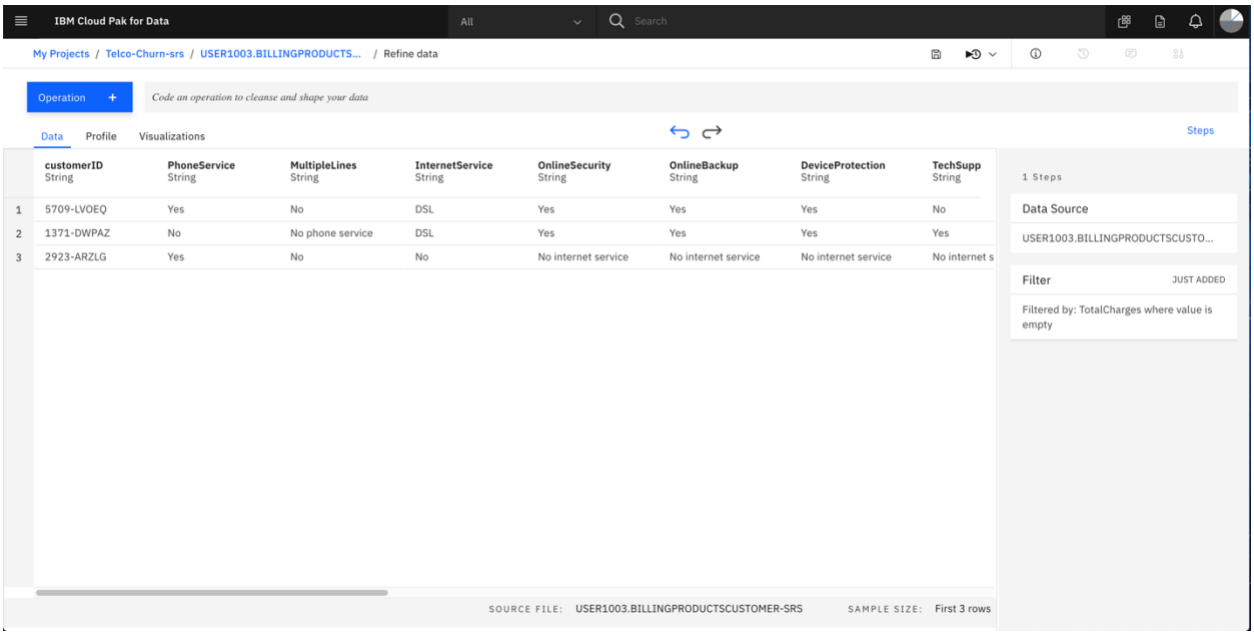
We'll start out on the Data tab.

Click the **+Operation** button.



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**.



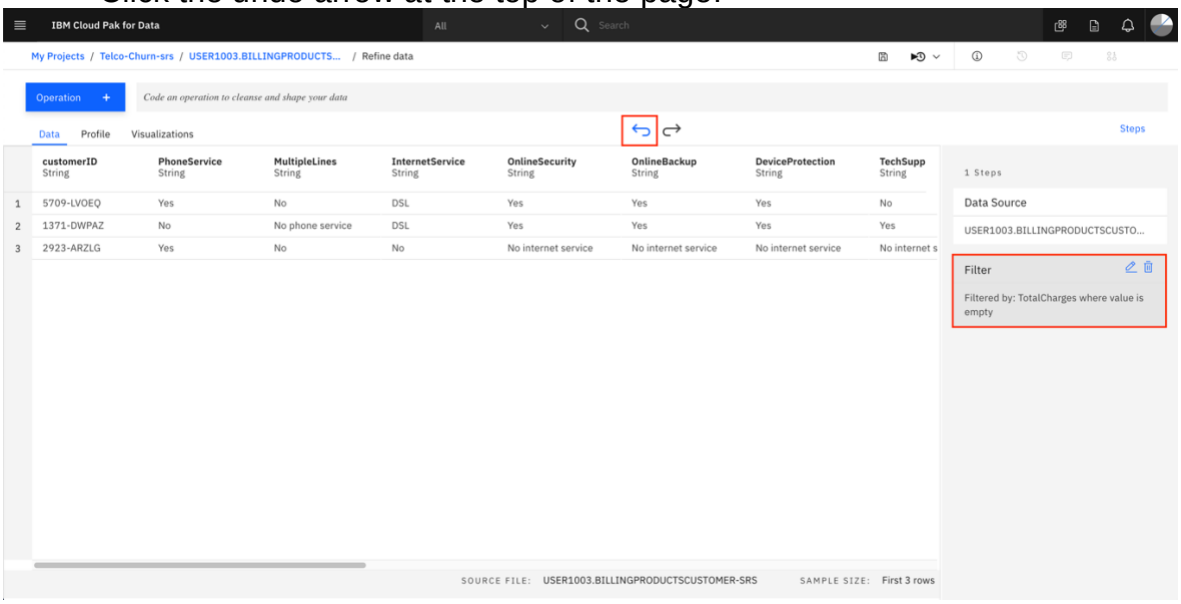


We can see that there are only three rows with an empty value for TotalCharges.

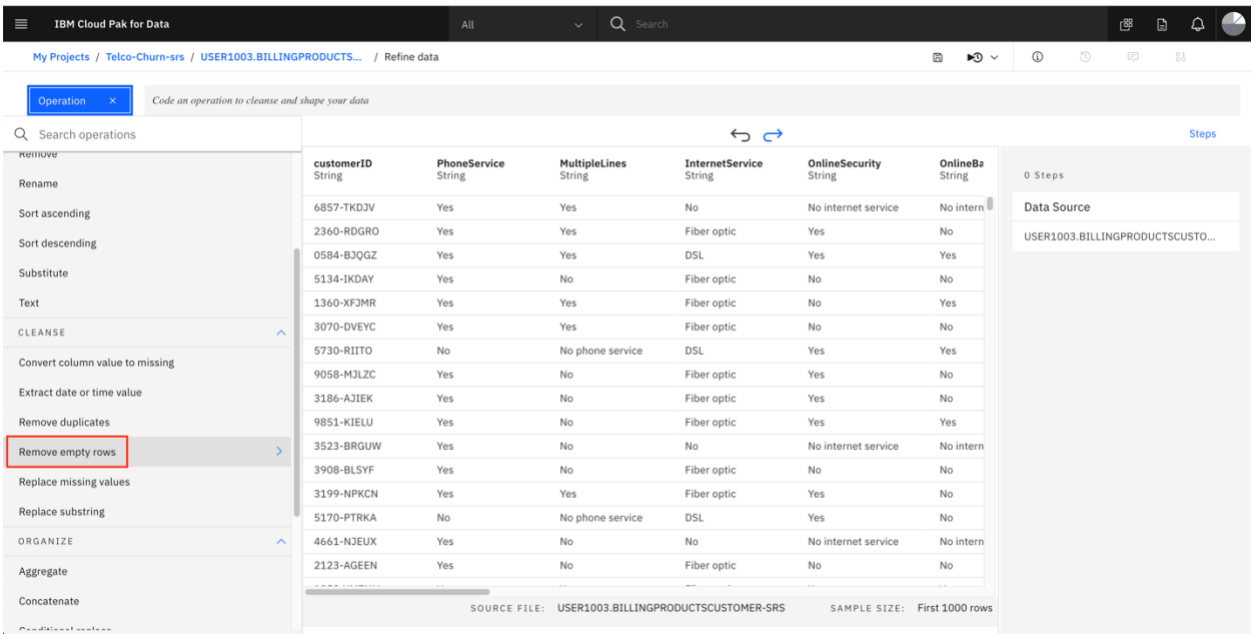
It should be safe to just drop these rows from the data set, so let's do that.

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

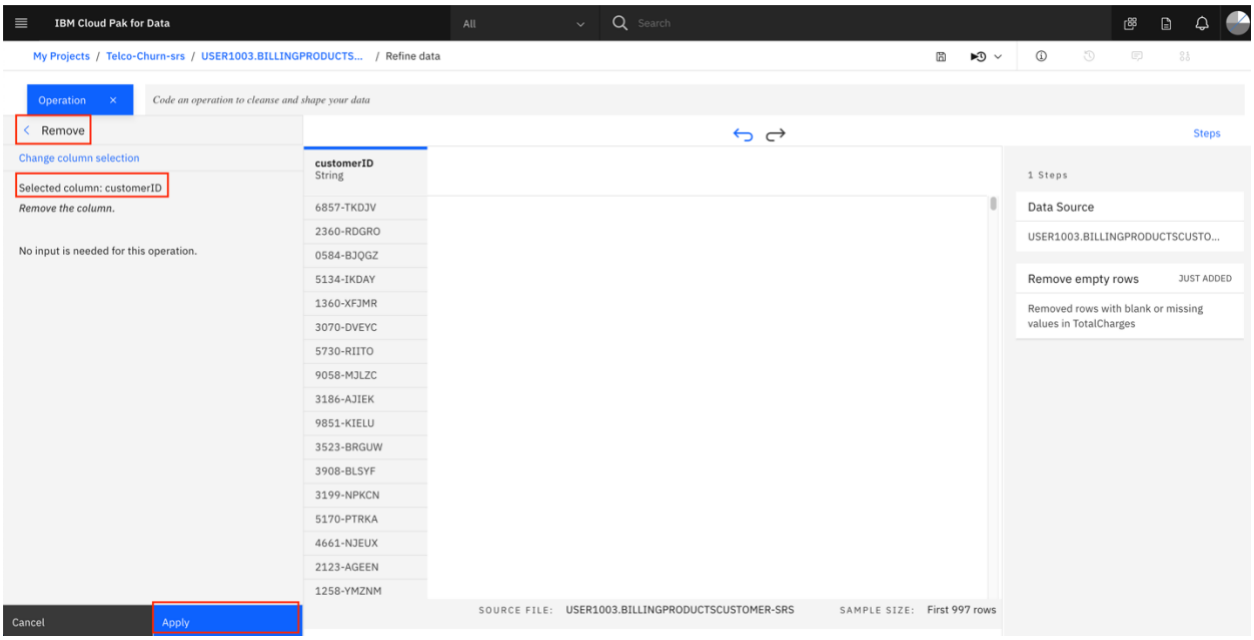
- Hover over the corresponding step in the Steps section and the delete icon (trash can) will appear. Click on this icon to remove the filter.
- Click the undo arrow at the top of the page.



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



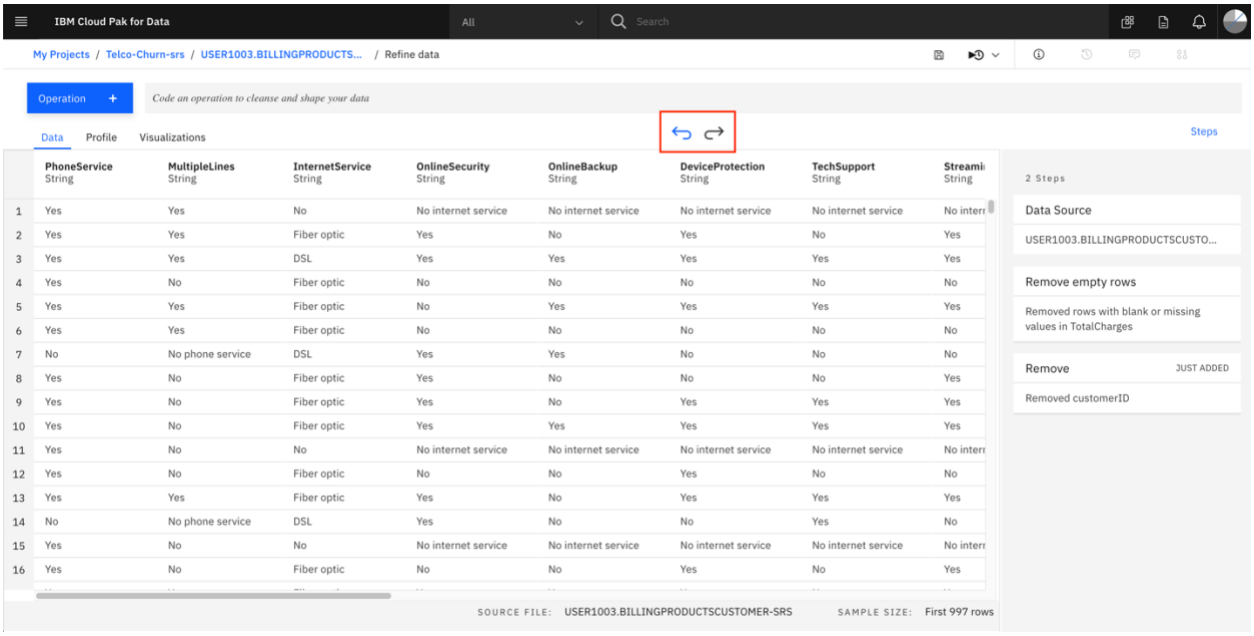
Finally, we can remove the CustomerID column, since that won't be useful for training a machine learning model in the next exercise. Choose the **Remove** operator, then **Change column selection**. Under **Select a column**, pick **CustomerID**, then **Next**, then **Apply**.



Click on the save button to save the progress.

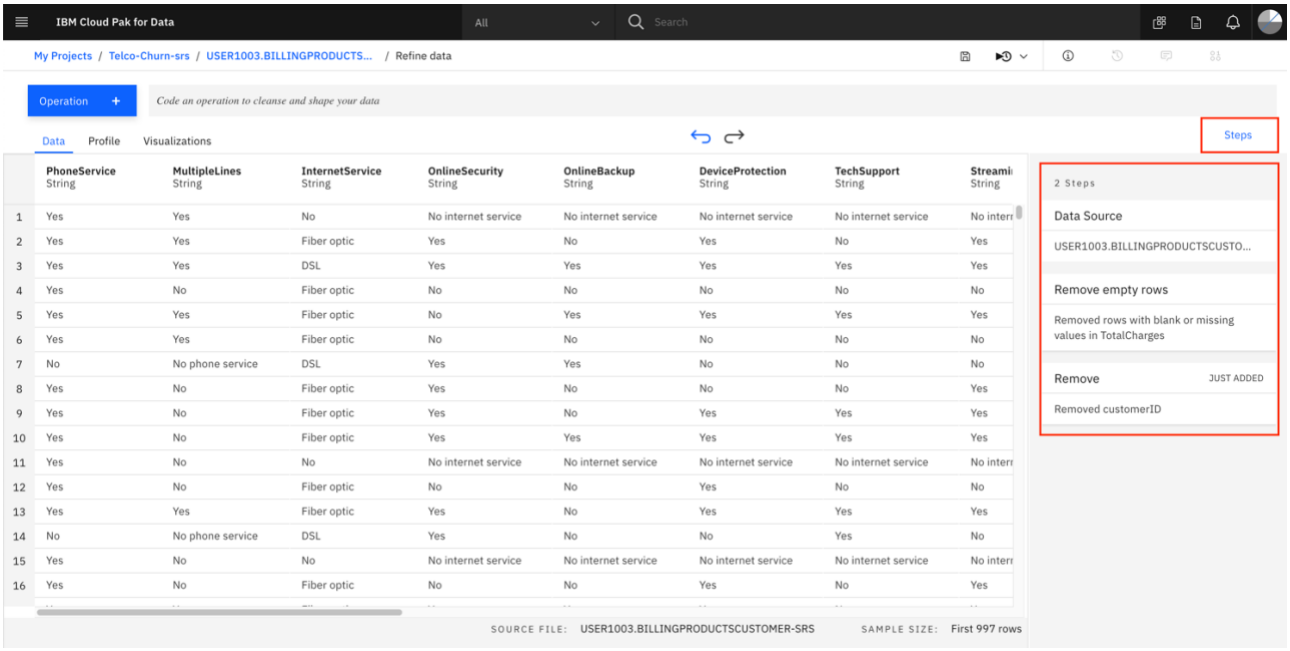
Step 3. Use data flow steps to keep track of your work

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 circular arrows.



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.

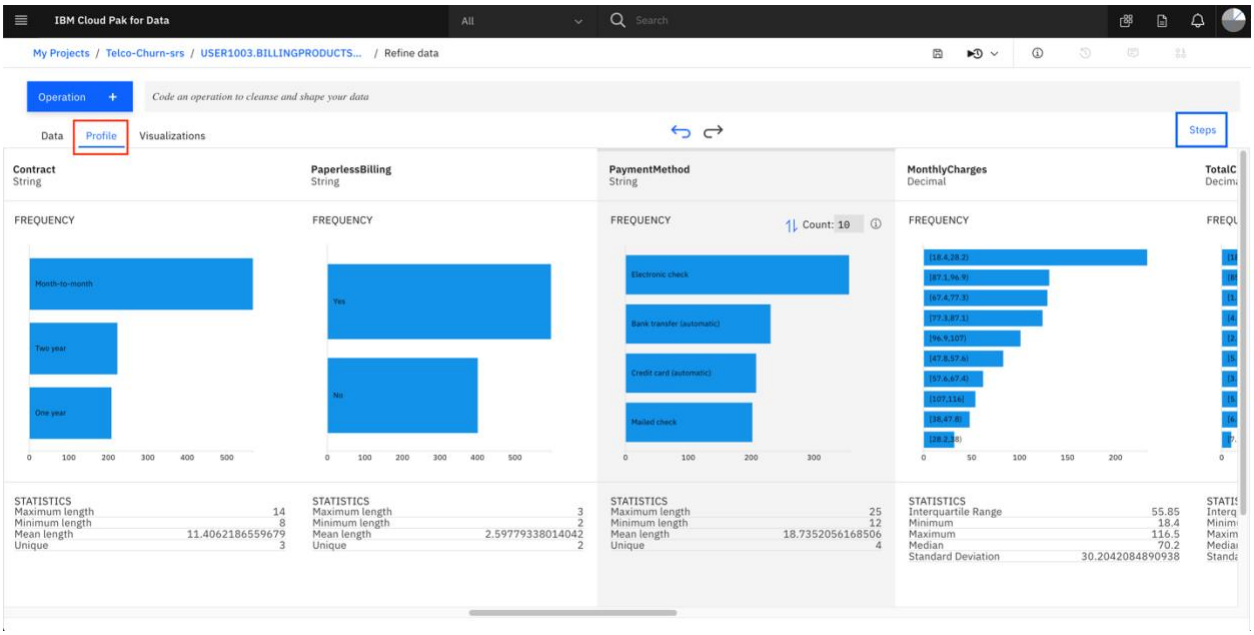
To see the steps in the data flow that you have performed, click the **Steps** button. The operations you have performed on the data will be shown.



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.

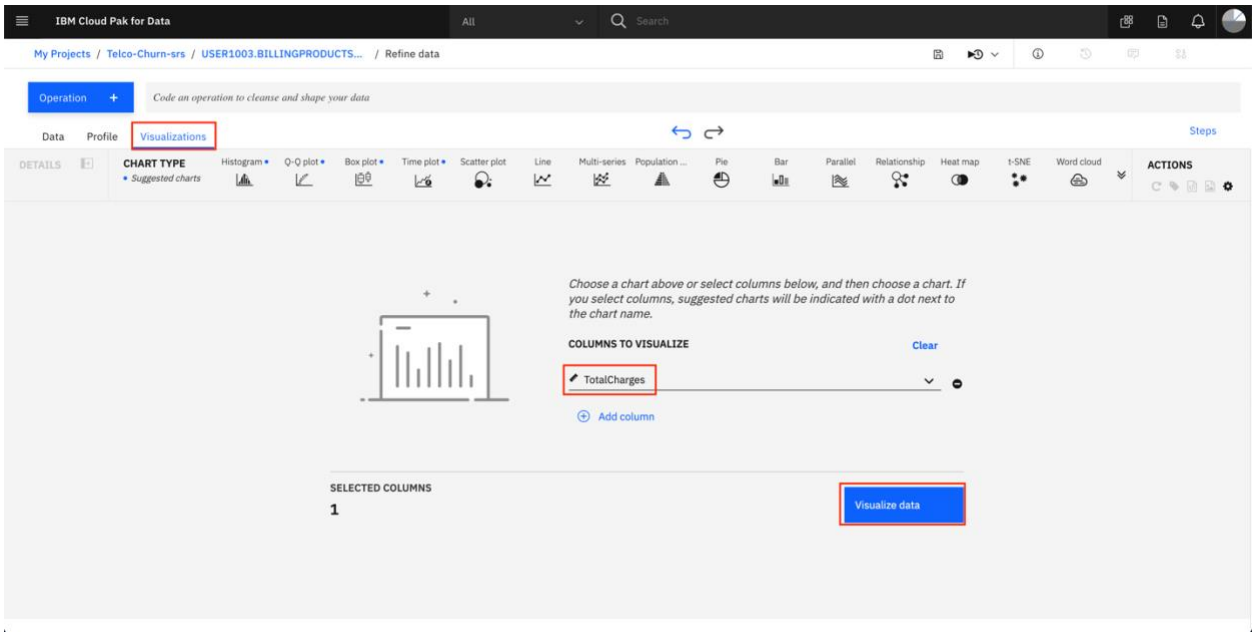


You can get insights into the data from the histograms:

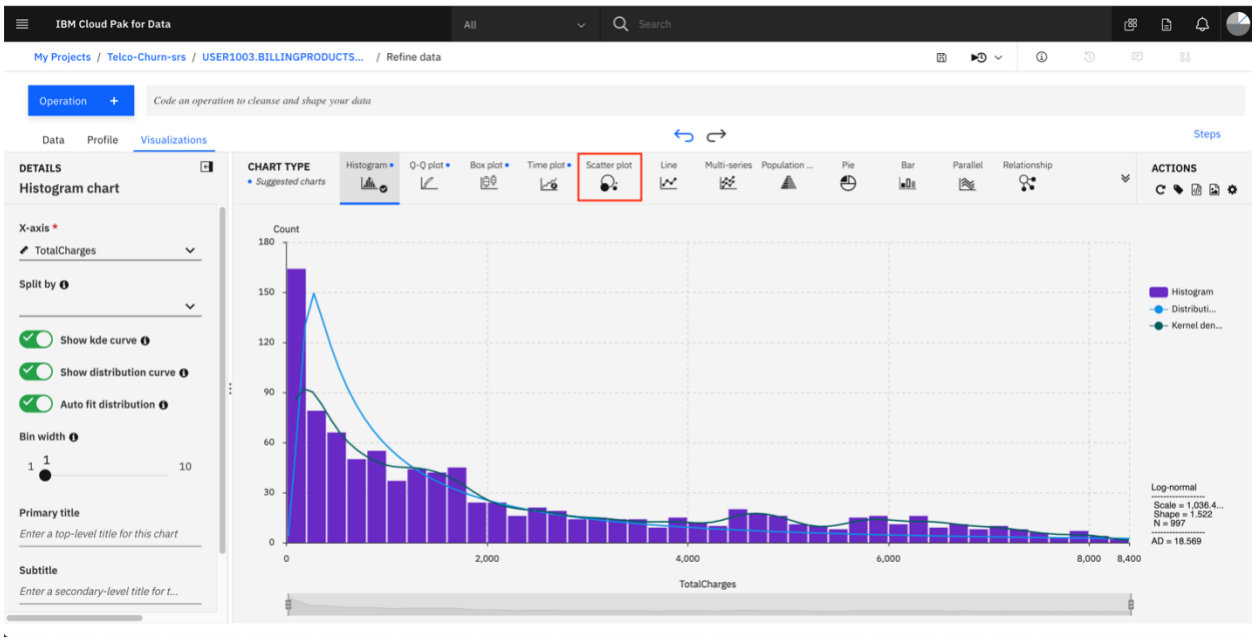
- Twice as many customers are month to month as are a one- or two-year contract.
- More choose paperless billing, but around 40 percent still prefer a paper bill sent to them.
- You can see the distribution of MonthlyCharges and TotalCharges.
- From the Churn column, you can see that a significant number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

Step 5. Visualize the data

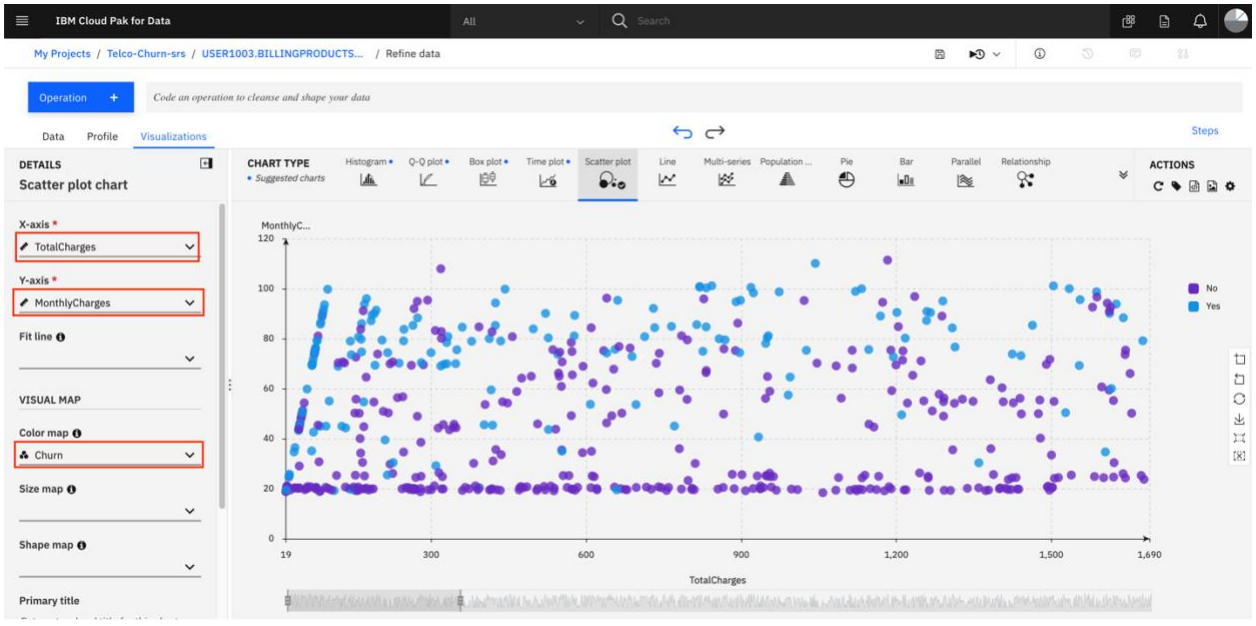
1. Choose the **Visualizations** tab to bring up an option to choose which columns to visualize. Under **Columns to Visualize**, choose **TotalCharges** and click **Visualize data**.



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.

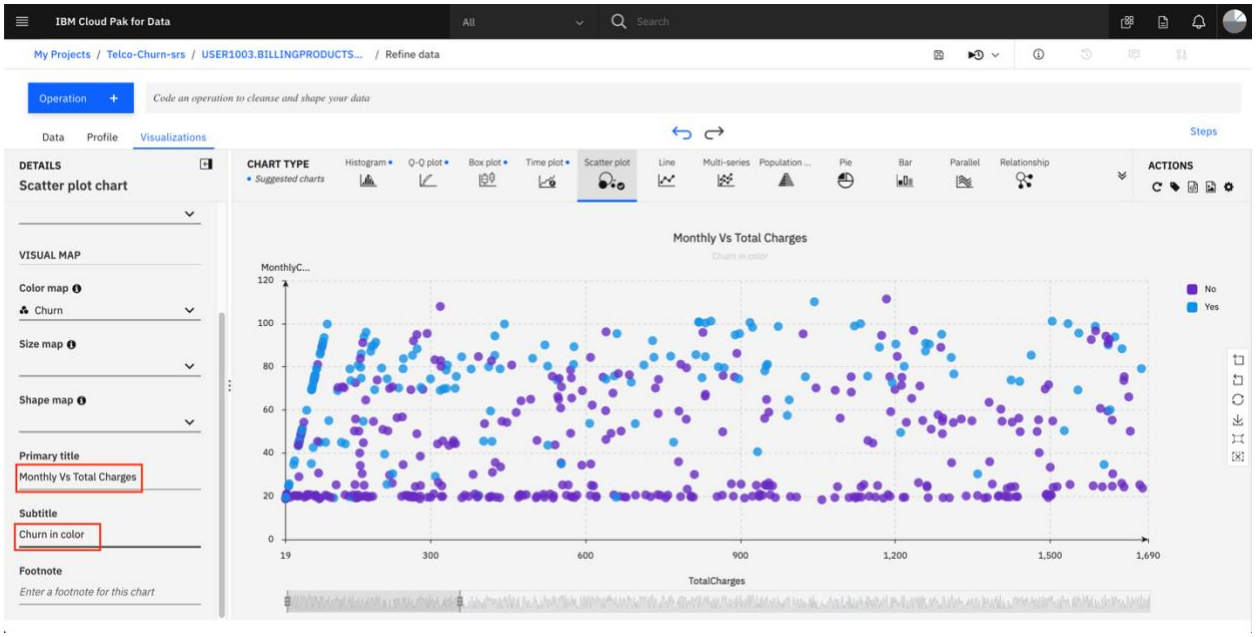


3. In the scatter plot, choose **TotalCharges** for the x-axis, **MonthlyCharges** for the y-axis, and **Churn** for the color map.

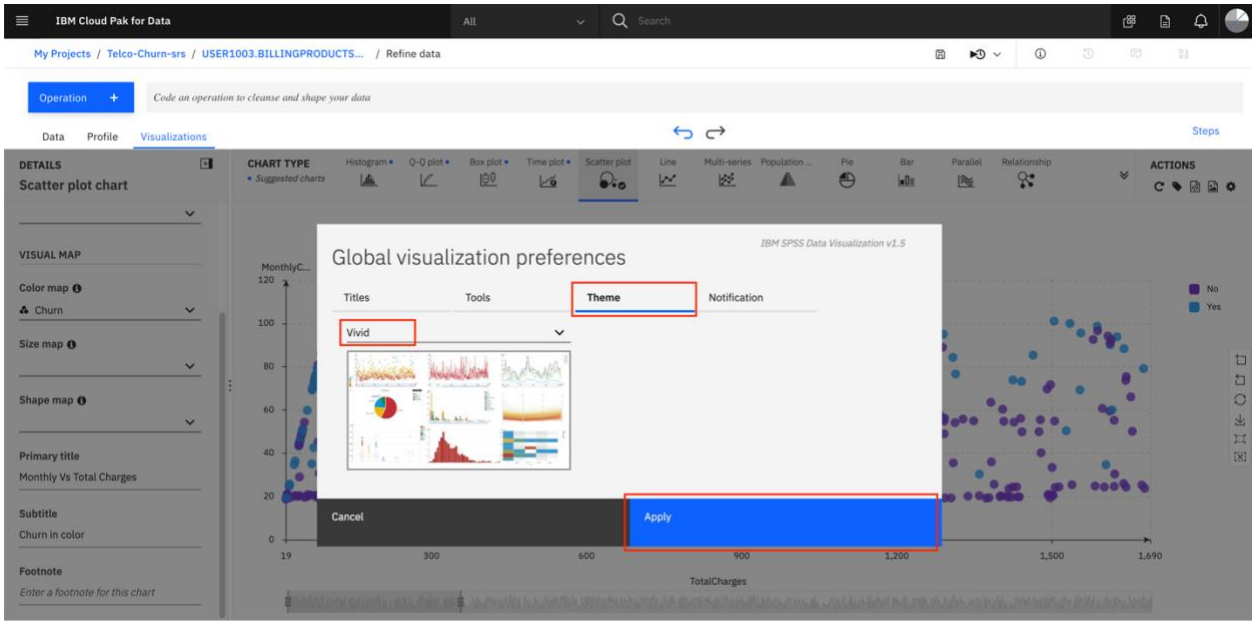


4. Scroll down and give the scatter plot a title and sub-title if you wish. Under the **Actions** panel, notice that you can perform tasks such as start over, download chart details, display data label in chart, download chart image, or set global

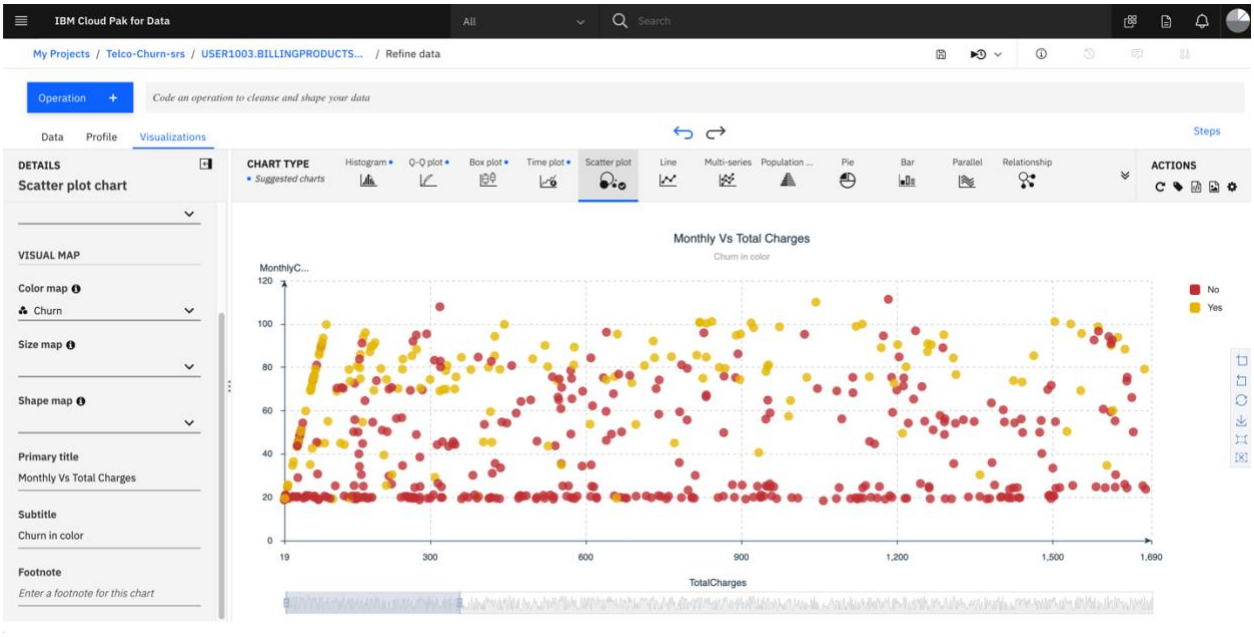
visualization preferences (hover over the icons to see the names). Click on the gear icon in the **Actions** panel.



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**.



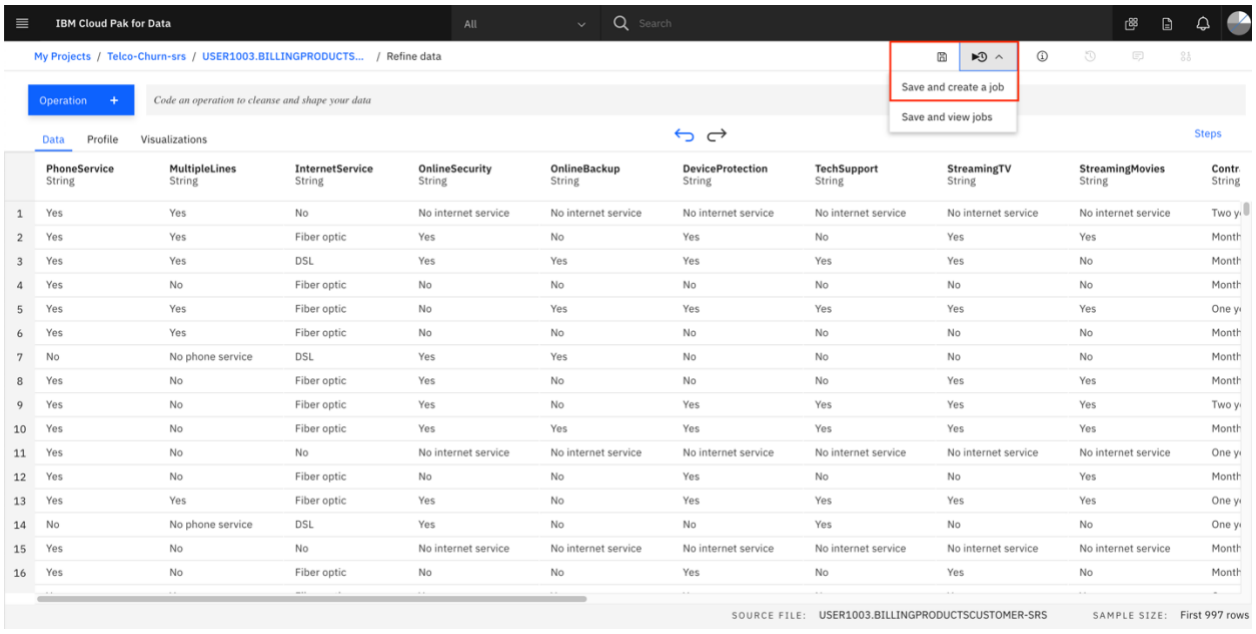
Now the colors for all of our charts will be reflected.



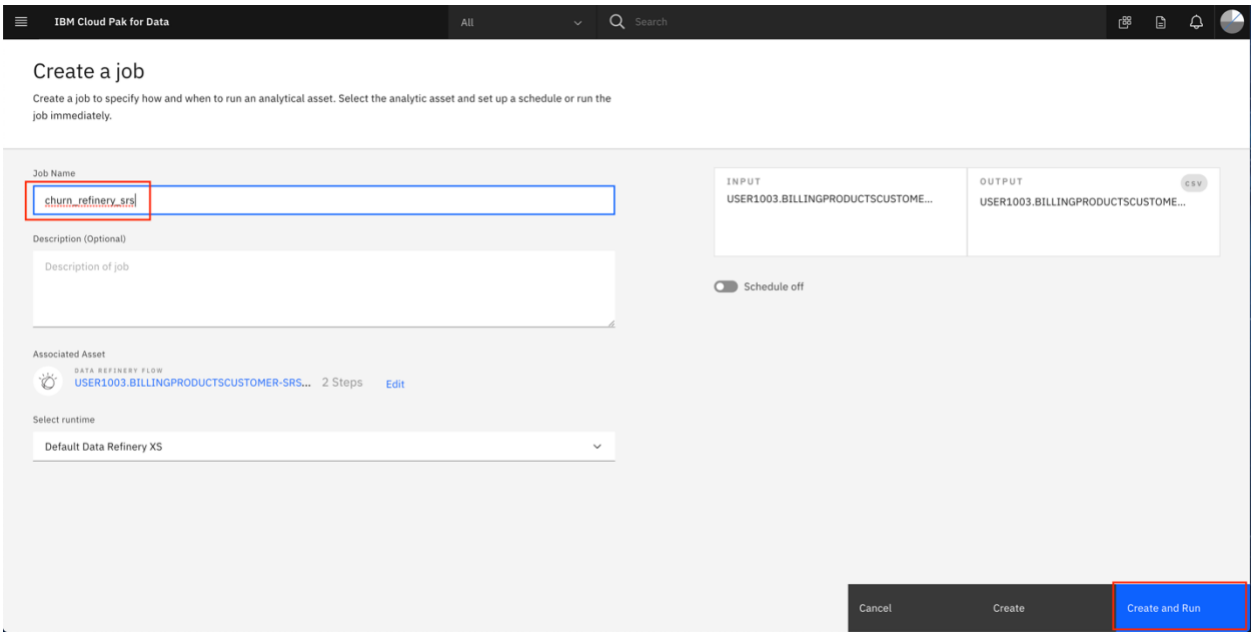
Step 6. Save data flow and create a job

Once you have refined your data, you would want to save create a job that can run the data refinery flow and return the refined and pre-processed data as its output.

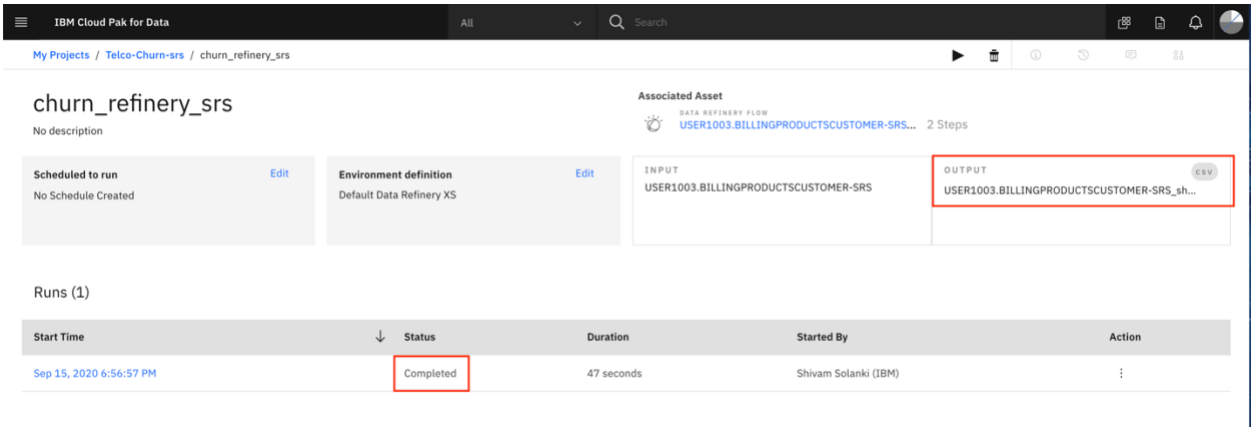
1. Click on **Save and create a job** from the Play dropdown button shown below



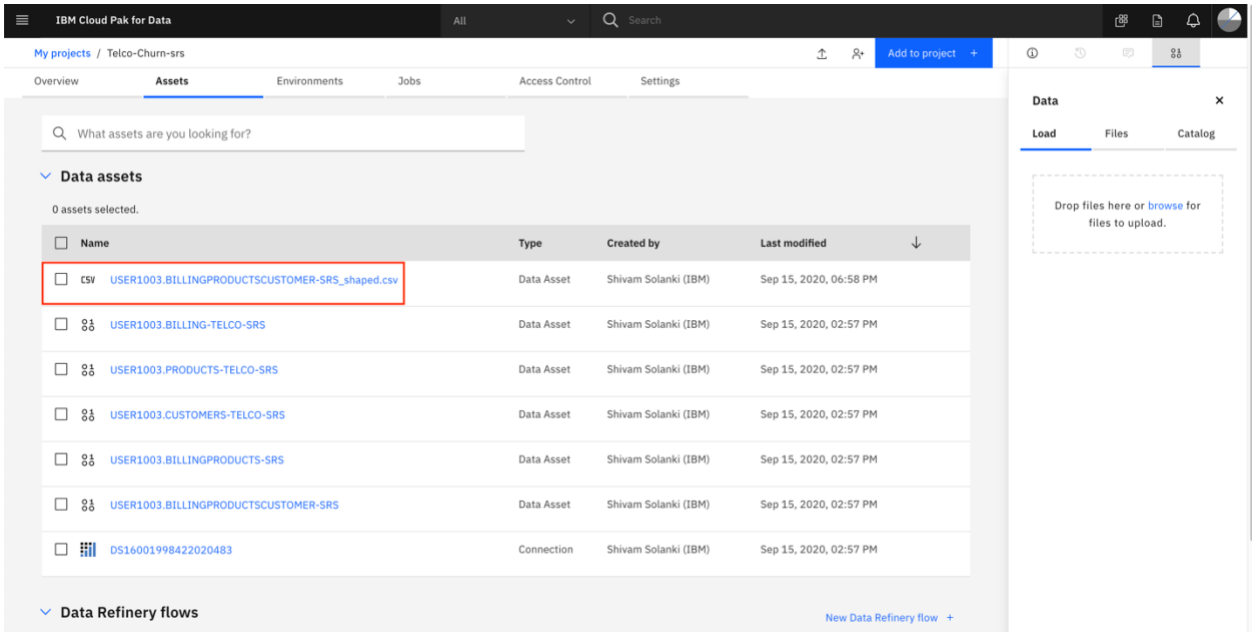
2. Enter the name of the job with your initials at the end to avoid conflict with other data refinery flows running in the same environment. Then click on **Create and Run**



3. You will see a similar window with status: running as shown below. When the data refinery process has ran successfully, the status will update to **Completed**



4. Click on the Project name to confirm the output of the data refinery flow as a csv file with the file name shown in the above image.



You have successfully completed the data processing and visualization step. We will be using this shaped data in the next step so make sure that you have complete this task before moving on to the modeling step.

Conclusion

This tutorial showed you a small sampling of the power of Data Refinery on IBM Cloud Pak for Data. The tutorial also explained how you can transform data using various operations on the columns, such as removing empty rows, or deleting columns altogether. The tutorial also explained that all the steps in our data flow are recorded, so you can remove steps, repeat them, or edit an individual step. It showed how you can quickly profile data to see histograms and statistics for each column. And finally, it explained how you can create more in-depth visualizations and create a scatter-plot mapping total charges vs. monthly charges, with the churn results highlighted in color.