

Lab Guide

Hands-on-Lab: Data visualization with data refinery

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



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 go 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

The screenshot shows the IBM Cloud Pak for Data interface. The top navigation bar includes 'My projects / Telco-Churn-srs' and an 'Add to project' button. The 'Assets' tab is selected, showing a search bar and a list of data assets. The asset 'USER1003.BILLINGPRODUCTSCUSTOMER-SRS' is highlighted with a red box. A right-hand panel shows a 'Data' section with a 'Load' button and a 'Drop files here or browse for files to upload' message.

Name	Type	Created by	Last modified
USER1003.BILLING-TELCO-SRS	Data Asset	Shivam Solanki (IBM)	Sep 15, 2020, 02:57 PM
USER1003.PRODUCTS-TELCO-SRS	Data Asset	Shivam Solanki (IBM)	Sep 15, 2020, 02:57 PM
USER1003.CUSTOMERS-TELCO-SRS	Data Asset	Shivam Solanki (IBM)	Sep 15, 2020, 02:57 PM
USER1003.BILLINGPRODUCTS-SRS	Data Asset	Shivam Solanki (IBM)	Sep 15, 2020, 02:57 PM
USER1003.BILLINGPRODUCTSCUSTOMER-SRS	Data Asset	Shivam Solanki (IBM)	Sep 15, 2020, 02:57 PM
DS16001998422020483	Connection	Shivam Solanki (IBM)	Sep 15, 2020, 02:57 PM

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

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with 'IBM Cloud Pak for Data' and a search bar. Below it, the breadcrumb path is 'My Projects / Telco-Churn-srs / USER1003.BILLINGPRODUCTSCUSTOMER-SRS'. The main area is divided into 'Preview', 'Profile', and 'Lineage' tabs. The 'Preview' tab is active, showing a table with 25 columns and 1000 rows. The columns are: customerID, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingService, StreamingMedia, and Contract. The 'Refine' button is highlighted with a red box. On the right, an 'Information' panel shows details about the data asset, including its name, description, tags, and added date.

customerID	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingService	StreamingMedia	Contract
6857-TKDJV	Yes	Yes	No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service	Two year
2360-RDGRO	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes	Yes	Month-to-month
0584-BJQZ	Yes	Yes	DSL	Yes	Yes	Yes	Yes	Yes	No	Month-to-month
5134-IKDAY	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-month
1360-XFJMR	Yes	Yes	Fiber optic	No	Yes	Yes	Yes	Yes	Yes	One year
3070-DVEYC	Yes	Yes	Fiber optic	No	No	No	No	No	No	Month-to-month
5730-RIITO	No	No phone service	DSL	Yes	Yes	No	No	No	No	Month-to-month
9058-MJLZC	Yes	No	Fiber optic	Yes	No	No	No	Yes	Yes	Month-to-month
3186-AJIEK	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	Two year
9851-KIELU	Yes	No	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes	Month-to-month
3523-BRGUW	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service	One year
3908-BLSYF	Yes	No	Fiber optic	No	No	Yes	No	No	Yes	Month-to-month
3199-NPKCN	Yes	Yes	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	One year
5170-PTRKA	No	No phone service	DSL	Yes	No	No	Yes	No	No	One year
4661-NJEUX	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service	Month-to-month
2123-AGEEN	Yes	No	Fiber optic	No	No	Yes	No	Yes	No	Month-to-month

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

The screenshot shows the IBM Cloud Pak for Data interface with the 'Refine data' modal open. The modal has a 'Data' tab and a 'Profile' tab. The 'Data' tab is active, showing a table with 16 rows and 9 columns. The columns are: customerID, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, and StreamingService. The 'Refine data' button is highlighted with a red box. On the right, an 'Information' panel shows details about the data asset, including its name, location, data refinery flow name, and steps.

	customerID	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingService
1	6857-TKDJV	Yes	Yes	No	No internet service	No internet service	No internet service	No internet service	No internet service
2	2360-RDGRO	Yes	Yes	Fiber optic	Yes	No	Yes	No	No
3	0584-BJQZ	Yes	Yes	DSL	Yes	Yes	Yes	Yes	No
4	5134-IKDAY	Yes	No	Fiber optic	No	No	No	No	No
5	1360-XFJMR	Yes	Yes	Fiber optic	No	Yes	Yes	Yes	Yes
6	3070-DVEYC	Yes	Yes	Fiber optic	No	No	No	No	No
7	5730-RIITO	No	No phone service	DSL	Yes	Yes	No	No	No
8	9058-MJLZC	Yes	No	Fiber optic	Yes	No	No	No	No
9	3186-AJIEK	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes
10	9851-KIELU	Yes	No	Fiber optic	Yes	Yes	Yes	Yes	Yes
11	3523-BRGUW	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service
12	3908-BLSYF	Yes	No	Fiber optic	No	No	Yes	No	No
13	3199-NPKCN	Yes	Yes	Fiber optic	Yes	No	Yes	Yes	Yes
14	5170-PTRKA	No	No phone service	DSL	Yes	No	No	Yes	Yes
15	4661-NJEUX	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service
16	2123-AGEEN	Yes	No	Fiber optic	No	No	Yes	No	No

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.

The screenshot shows the IBM Cloud Pak for Data interface. At the top, there's a navigation bar with 'My Projects / Fraud-claim-srs / USER1003.POLICIESCLAIMSCU...' and a 'Refine data' button. Below this, there's a 'Data' tab with a table of data. The table has columns: customer_id, USER1003_CLA..., policy_id, capital_gains, capital_loss, incident_date, incident_type, and collision_type. The table contains 18 rows of data. To the right of the table, there's an 'Information' panel with 'Details' and 'Help' tabs. The 'Details' tab is active, showing 'LOCATION: Fraud-claim-srs' and 'DATA REFINERY FLOW NAME: USER1003.POLICIESCL...'. There's also a 'STEPS' section showing '0' steps. At the bottom, there's a 'SOURCE FILE: USER1003.POLICIESCLAIMSCUSTOMERS' and 'SAMPLE SIZE: First 1000 rows'.

	customer_id String	USER1003_CLA... Decimal	policy_id String	capital_gains Integer	capital_loss Integer	incident_date Date	incident_type String	collision_type String
1	7KOGSQJ	6006872958	I6L5Z96X3	0	-36600	2015-01-27	Single Vehicle Collision	Rear Collision
2	NMGRWBM	3321605479	TQ1L8GMLY	61900	-50000	2015-01-28	Single Vehicle Collision	Side Collision
3	FMB0ACY	6420960222	27M5W4FIR	67800	-48600	2015-02-23	Multi-vehicle Collision	Side Collision
4	J372SK6	5729296200	KWQ725X18	0	0	2015-01-18	Single Vehicle Collision	Rear Collision
5	G343TSY	1208965988	NNN4VPVG	35400	0	2015-02-15	Multi-vehicle Collision	Front Collision
6	A7U88RM	3537518110	XS8ISUKFG	0	-45300	2015-02-04	Vehicle Theft	?
7	7MX3F17	2136806842	ZMYUYH9WR	67800	0	2015-01-12	Multi-vehicle Collision	Rear Collision
8	WV4LB59	2398517127	0FE47K9LY	0	-48800	2015-01-02	Vehicle Theft	?
9	OQDBUYV	5299562138	88RODD5AM	30400	-89400	2015-01-27	Single Vehicle Collision	Rear Collision
10	Q83ZA7K	3518012633	TOJZM4LXQ	0	-70100	2015-02-09	Vehicle Theft	?
11	O305NGO	2860999482	IPKD9K1G2	0	-36400	2015-02-06	Vehicle Theft	?
12	SQJ6EEA	8043481798	CRZNETG16	64600	0	2015-01-03	Single Vehicle Collision	Front Collision
13	NCZBYX3	2685562276	Y1TM1WZEK	0	0	2015-01-12	Multi-vehicle Collision	Rear Collision
14	NXRO4C3	7551445480	J3JL03GLC	53800	0	2015-01-22	Single Vehicle Collision	Rear Collision
15	CILSQ9Q	2574043082	360A1VNRU	0	0	2015-01-03	Multi-vehicle Collision	Side Collision
16	FDNMWKB	9234796810	EF110F3J2	69400	0	2015-02-23	Vehicle Theft	?
17	XMG6T47	4575503118	XPPOCC150	58500	-77700	2015-01-22	Single Vehicle Collision	Side Collision
18	HWHSRF2	1941832146	SH4VJMG82	53400	-35200	2015-02-13	Single Vehicle Collision	Front Collision

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 IBM Cloud Pak for Data interface with the 'Filter' operation selected. The 'Filter' panel is open, showing a list of conditions. The first condition is 'TotalCharges' with the operator 'Is empty'. The 'TotalCharges' column is highlighted in the list of columns. The 'Is empty' operator is selected in the dropdown menu. The 'Apply' button is highlighted in blue. The 'Data Source' panel on the right shows 'USER1003.BILLINGPRODUCTSCUSTO...'. At the bottom, there's a 'SOURCE FILE: USER1003.BILLINGPRODUCTSCUSTOMER-SRS' and 'SAMPLE SIZE: First 1000 rows'.

CONDITIONS (1)
CONDITION 1

Column: TotalCharges Operator: Is empty

1620.45
6812.95
1837.9
69.8
7344.45
545.15
1500.25
2283.15
6844.5
1043.3
504.2
497.55
7511.65
1782
20.05
609.65
2857.6

IBM Cloud Pak for Data

My Projects / Telco-Churn-srs / USER1003.BILLINGPRODUCTS... / Refine data

Operation + Code an operation to cleanse and shape your data

Data Profile Visualizations

	customerID String	PhoneService String	MultipleLines String	InternetService String	OnlineSecurity String	OnlineBackup String	DeviceProtection String	TechSupp String
1	5709-LVOEQ	Yes	No	DSL	Yes	Yes	Yes	No
2	1371-DWPAZ	No	No phone service	DSL	Yes	Yes	Yes	Yes
3	2923-ARZLG	Yes	No	No	No internet service	No internet service	No internet service	No internet s

1 Steps

Data Source

USER1003.BILLINGPRODUCTSCUSTO...

Filter JUST ADDED

Filtered by: TotalCharges where value is empty

SOURCE FILE: USER1003.BILLINGPRODUCTSCUSTOMER-SRS SAMPLE SIZE: First 3 rows

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.

IBM Cloud Pak for Data

My Projects / Telco-Churn-srs / USER1003.BILLINGPRODUCTS... / Refine data

Operation + Code an operation to cleanse and shape your data

Data Profile Visualizations

	customerID String	PhoneService String	MultipleLines String	InternetService String	OnlineSecurity String	OnlineBackup String	DeviceProtection String	TechSupp String
1	5709-LVOEQ	Yes	No	DSL	Yes	Yes	Yes	No
2	1371-DWPAZ	No	No phone service	DSL	Yes	Yes	Yes	Yes
3	2923-ARZLG	Yes	No	No	No internet service	No internet service	No internet service	No internet s

1 Steps

Data Source

USER1003.BILLINGPRODUCTSCUSTO...

Filter Filtered by: TotalCharges where value is empty

SOURCE FILE: USER1003.BILLINGPRODUCTSCUSTOMER-SRS SAMPLE SIZE: First 3 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 IBM Cloud Pak for Data interface. On the left, the 'Operation' menu is open, and 'Remove empty rows' is selected. The main area displays a data table with columns: customerID, PhoneService, MultipleLines, InternetService, OnlineSecurity, and OnlineBa. The table contains 15 rows of data. On the right, the 'Steps' panel shows '0 Steps' and 'Data Source: USER1003.BILLINGPRODUCTSCUSTO...'. At the bottom, it indicates 'SOURCE FILE: USER1003.BILLINGPRODUCTSCUSTOMER-SRS' and 'SAMPLE SIZE: First 1000 rows'.

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**.

The screenshot shows the IBM Cloud Pak for Data interface. On the left, the 'Operation' menu is open, and 'Remove' is selected. The 'Change column selection' section shows 'Selected column: customerID'. The main area displays a list of customer IDs. On the right, the 'Steps' panel shows '1 Steps' and 'Data Source: USER1003.BILLINGPRODUCTSCUSTO...'. At the bottom, it indicates 'SOURCE FILE: USER1003.BILLINGPRODUCTSCUSTOMER-SRS' and 'SAMPLE SIZE: First 997 rows'. The 'Apply' button is highlighted.

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.

IBM Cloud Pak for Data

My Projects / Telco-Churn-srs / USER1003.BILLINGPRODUCTS... / Refine data

Operation + Code an operation to cleanse and shape your data

Data Profile Visualizations

Undo Redo

	PhoneService String	MultipleLines String	InternetService String	OnlineSecurity String	OnlineBackup String	DeviceProtection String	TechSupport String	Streami String
1	Yes	Yes	No	No internet service	No internet service	No internet service	No internet service	No internet service
2	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes
3	Yes	Yes	DSL	Yes	Yes	Yes	Yes	Yes
4	Yes	No	Fiber optic	No	No	No	No	No
5	Yes	Yes	Fiber optic	No	Yes	Yes	Yes	Yes
6	Yes	Yes	Fiber optic	No	No	No	No	No
7	No	No phone service	DSL	Yes	Yes	No	No	No
8	Yes	No	Fiber optic	Yes	No	No	No	Yes
9	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes
10	Yes	No	Fiber optic	Yes	Yes	Yes	Yes	Yes
11	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service
12	Yes	No	Fiber optic	No	No	Yes	No	No
13	Yes	Yes	Fiber optic	Yes	No	Yes	Yes	Yes
14	No	No phone service	DSL	Yes	No	No	Yes	No
15	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service
16	Yes	No	Fiber optic	No	No	Yes	No	Yes

2 Steps

- Data Source
USER1003.BILLINGPRODUCTSCUSTO...
- Remove empty rows
Removed rows with blank or missing values in TotalCharges
- Remove JUST ADDED
Removed customerID

SOURCE FILE: USER1003.BILLINGPRODUCTSCUSTOMER-SRS SAMPLE SIZE: First 997 rows

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.

IBM Cloud Pak for Data

My Projects / Telco-Churn-srs / USER1003.BILLINGPRODUCTS... / Refine data

Operation + Code an operation to cleanse and shape your data

Data Profile Visualizations

	PhoneService String	MultipleLines String	InternetService String	OnlineSecurity String	OnlineBackup String	DeviceProtection String	TechSupport String	Streami String
1	Yes	Yes	No	No internet service	No internet service	No internet service	No internet service	No interr
2	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes
3	Yes	Yes	DSL	Yes	Yes	Yes	Yes	Yes
4	Yes	No	Fiber optic	No	No	No	No	No
5	Yes	Yes	Fiber optic	No	Yes	Yes	Yes	Yes
6	Yes	Yes	Fiber optic	No	No	No	No	No
7	No	No phone service	DSL	Yes	Yes	No	No	No
8	Yes	No	Fiber optic	Yes	No	No	No	Yes
9	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes
10	Yes	No	Fiber optic	Yes	Yes	Yes	Yes	Yes
11	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No interr
12	Yes	No	Fiber optic	No	No	Yes	No	No
13	Yes	Yes	Fiber optic	Yes	No	Yes	Yes	Yes
14	No	No phone service	DSL	Yes	No	No	Yes	No
15	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No interr
16	Yes	No	Fiber optic	No	No	Yes	No	Yes

SOURCE FILE: USER1003.BILLINGPRODUCTSCUSTOMER-SRS SAMPLE SIZE: First 997 rows

2 Steps

Data Source

USER1003.BILLINGPRODUCTSCUSTO...

Remove empty rows

Removed rows with blank or missing values in TotalCharges

Remove JUST ADDED

Removed customerID

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.

IBM Cloud Pak for Data

My Projects / Telco-Churn-srs / USER1003.BILLINGPRODUCTS... / Refine data

Operation + Code an operation to cleanse and shape your data

Data Profile Visualizations

Contract String

PaperlessBilling String

PaymentMethod String

MonthlyCharges Decimal

TotalC Decim

FREQUENCY

Month-to-month

Two year

One year

Yes

No

Electronic check

Bank transfer (automatic)

Credit card (automatic)

Mailed check

Count: 10

STATISTICS

Maximum length 14

Minimum length 8

Mean length 11.4062186559679

Unique 3

STATISTICS

Maximum length 3

Minimum length 2

Mean length 2.59779338014042

Unique 2

STATISTICS

Maximum length 25

Minimum length 12

Mean length 18.7352056168506

Unique 4

STATISTICS

Interquartile Range 55.85

Minimum 18.4

Maximum 116.5

Median 70.2

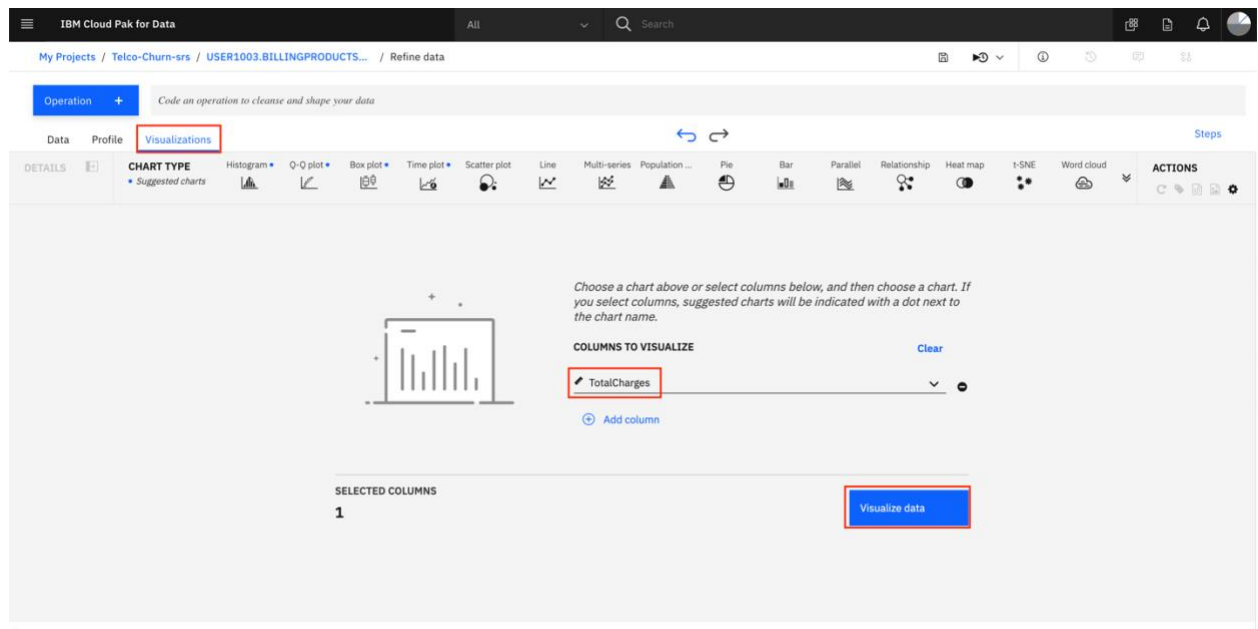
Standard Deviation 30.204208489938

You can get insights into the data from the histograms:

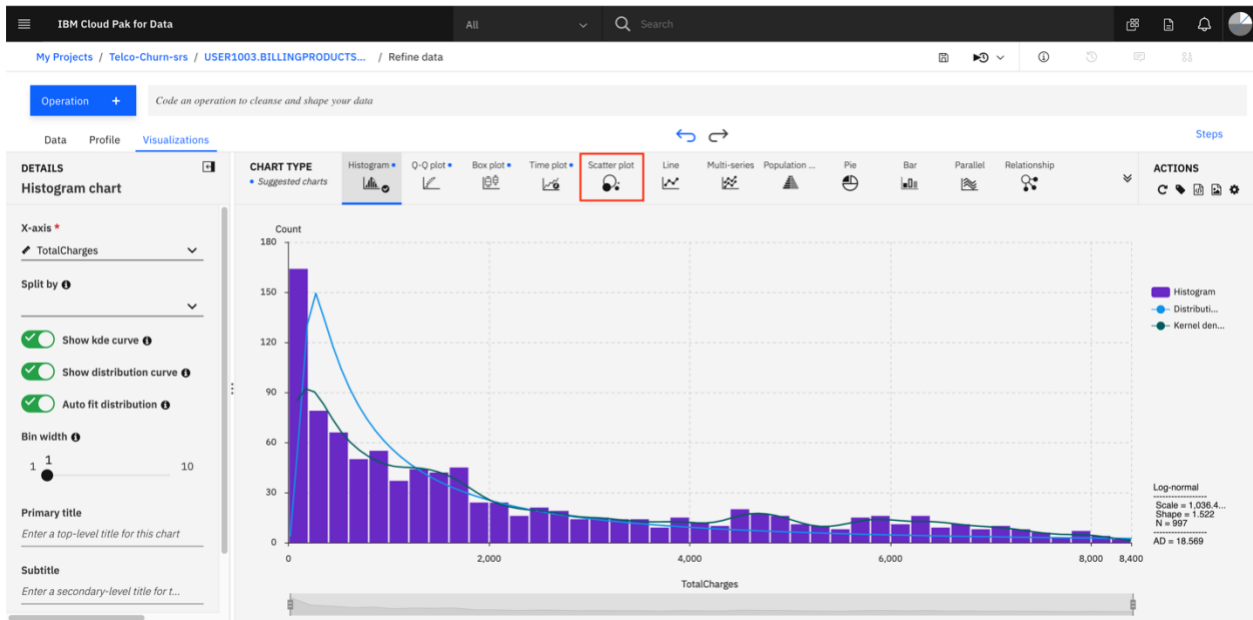
- Twice as many customers are month to month as are a one- or two-year contract.
- More choose paperless billing, but around 40 percent still prefer a paper bill sent to them.
- You can see the distribution of MonthlyCharges and TotalCharges.
- From the Churn column, you can see that a significant number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

Step 5. Visualize the data

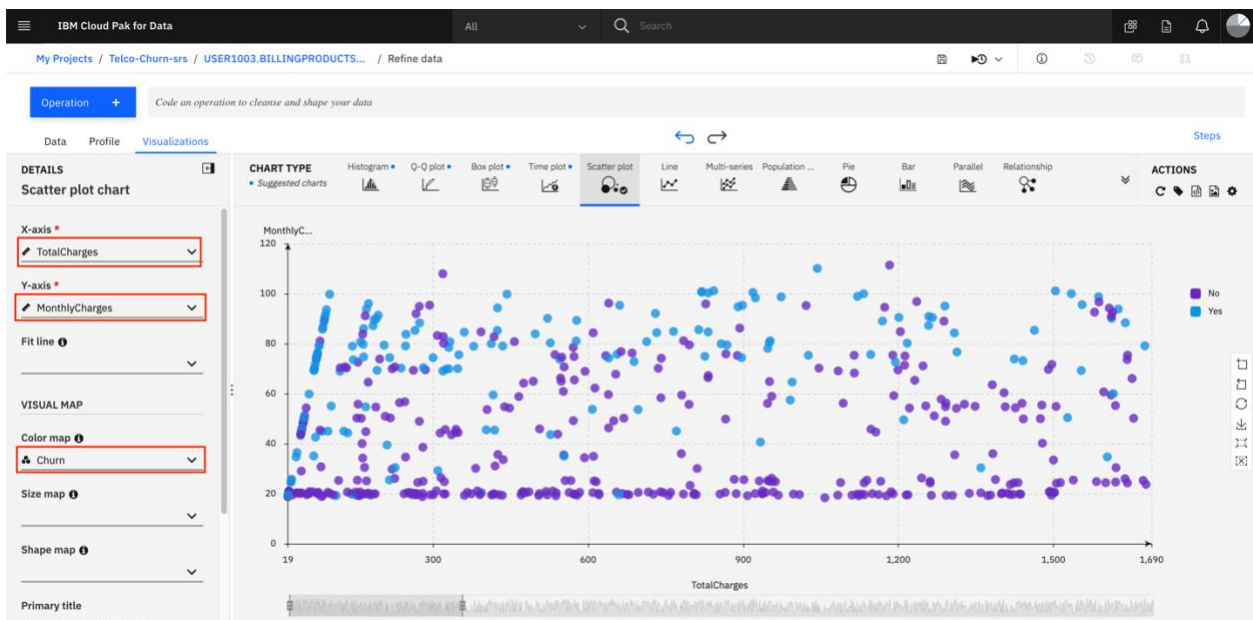
1. Choose the **Visualizations** tab to bring up an option to choose which columns to visualize. Under **Columns to Visualize**, choose **TotalCharges** and click **Visualize data**.



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.

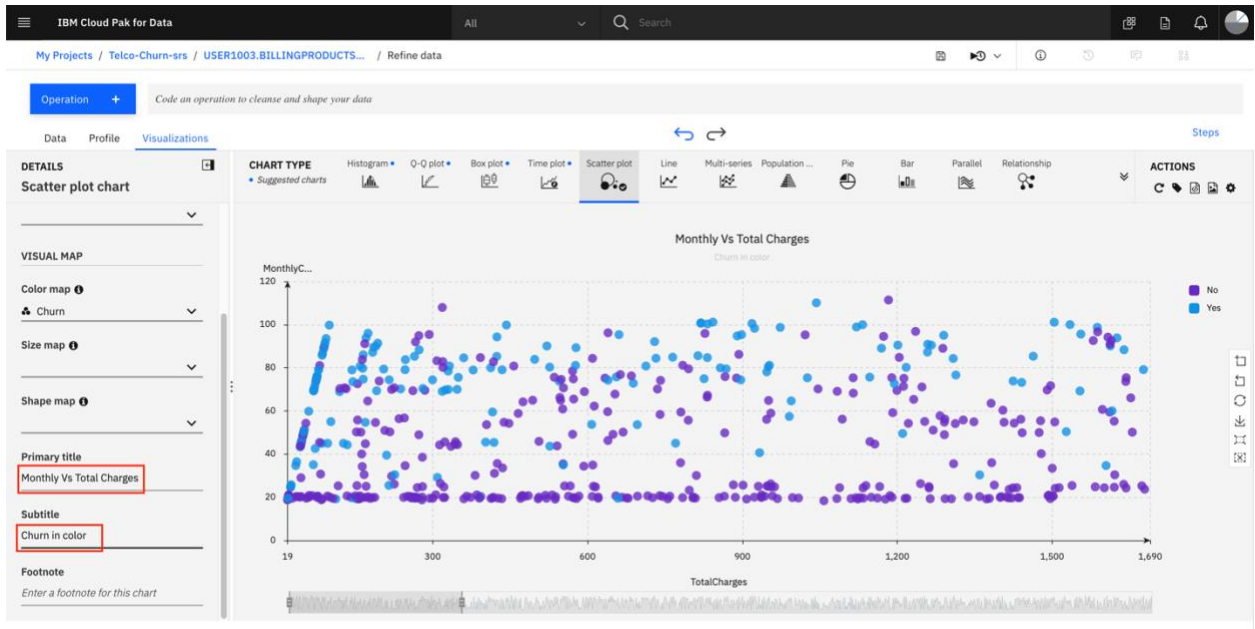


3. In the scatter plot, choose **TotalCharges** for the x-axis, **MonthlyCharges** for the y-axis, and **Churn** for the color map.

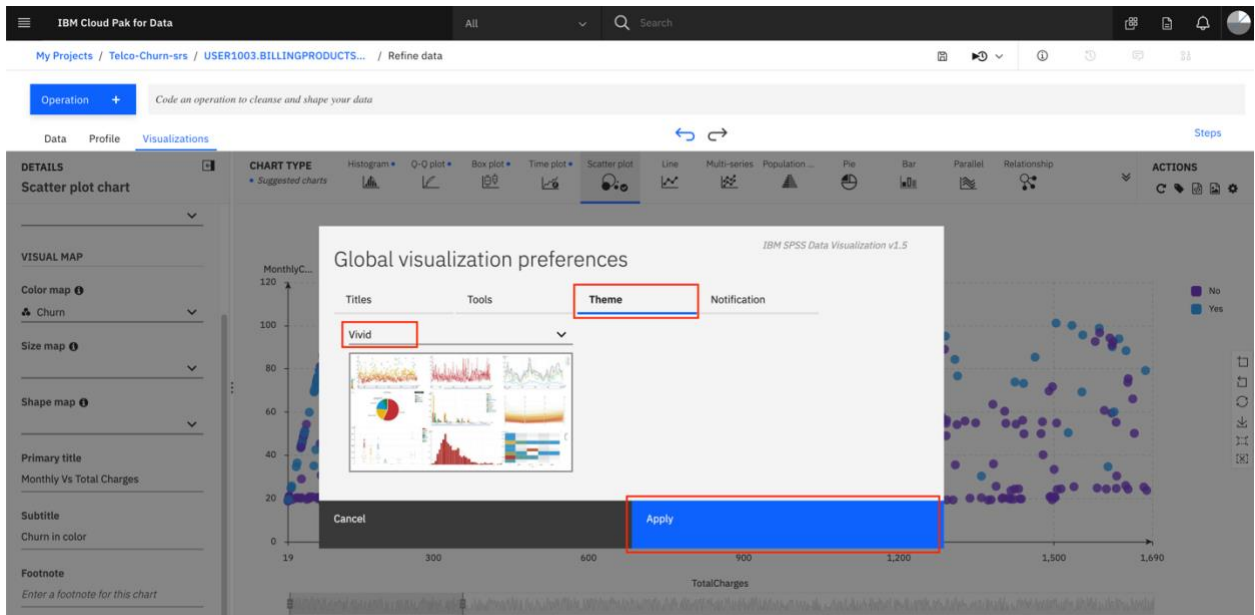


4. Scroll down and give the scatter plot a title and sub-title if you wish. Under the **Actions** panel, notice that you can perform tasks such as start over, download chart details, display data label in chart, download chart image, or set global

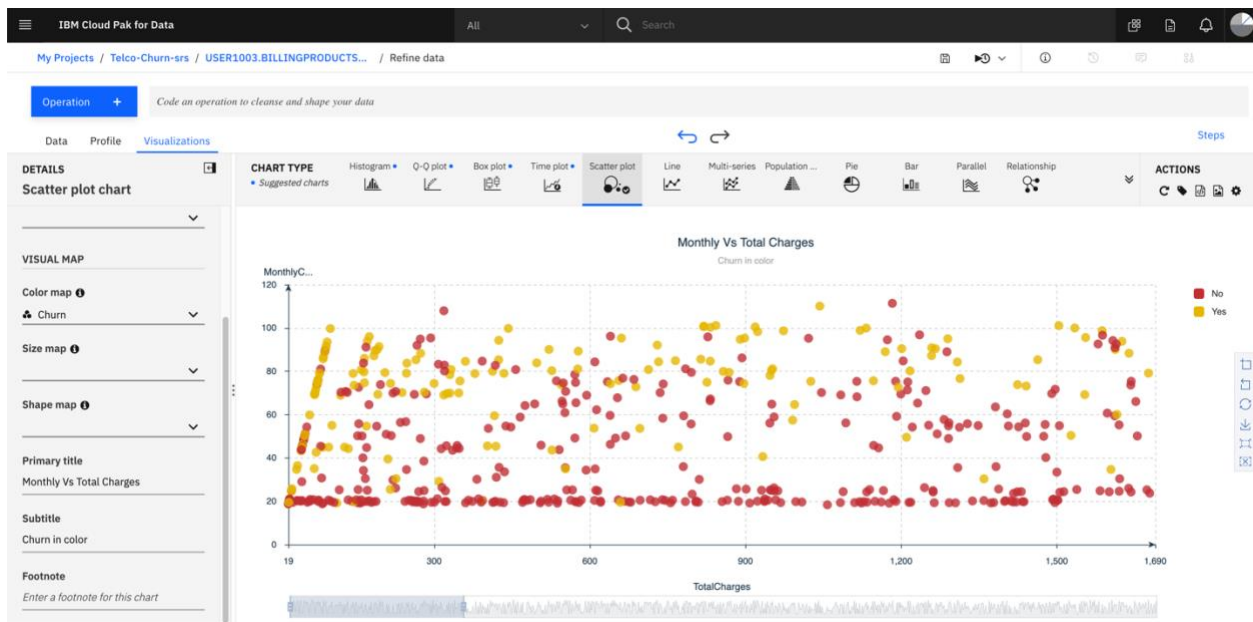
visualization preferences (hover over the icons to see the names). Click on the gear icon in the **Actions** panel.



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

	PhoneService String	MultipleLines String	InternetService String	OnlineSecurity String	OnlineBackup String	DeviceProtection String	TechSupport String	StreamingTV String	StreamingMovies String	Contr. String
1	Yes	Yes	No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service	Two y
2	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes	Yes	Month
3	Yes	Yes	DSL	Yes	Yes	Yes	Yes	Yes	No	Month
4	Yes	No	Fiber optic	No	No	No	No	No	No	Month
5	Yes	Yes	Fiber optic	No	Yes	Yes	Yes	Yes	Yes	One y
6	Yes	Yes	Fiber optic	No	No	No	No	No	No	Month
7	No	No phone service	DSL	Yes	Yes	No	No	No	No	Month
8	Yes	No	Fiber optic	Yes	No	No	No	Yes	Yes	Month
9	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	Two y
10	Yes	No	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes	Month
11	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service	One y
12	Yes	No	Fiber optic	No	No	Yes	No	No	Yes	Month
13	Yes	Yes	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	One y
14	No	No phone service	DSL	Yes	No	No	Yes	No	No	One y
15	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service	Month
16	Yes	No	Fiber optic	No	No	Yes	No	Yes	No	Month

SOURCE FILE: USER1003.BILLINGPRODUCTSCUSTOMER-SRS SAMPLE SIZE: First 997 rows

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**

Create a job

Create a job to specify how and when to run an analytical asset. Select the analytic asset and set up a schedule or run the job immediately.

Job Name:

Description (Optional):

Associated Asset: DATA REFINERY FLOW, USER1003.BILLINGPRODUCTSCUSTOMER-SRS... 2 Steps [Edit](#)

Select runtime: Default Data Refinery XS

INPUT: USER1003.BILLINGPRODUCTSCUSTOMER-SRS... OUTPUT: USER1003.BILLINGPRODUCTSCUSTOMER-SRS... [CSV](#)

☐ Schedule off

[Cancel](#) [Create](#) [Create and Run](#)

- 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**

churn_refinery_srs

No description

Scheduled to run: No Schedule Created [Edit](#)

Environment definition: Default Data Refinery XS [Edit](#)

Associated Asset: DATA REFINERY FLOW, USER1003.BILLINGPRODUCTSCUSTOMER-SRS... 2 Steps

INPUT: USER1003.BILLINGPRODUCTSCUSTOMER-SRS OUTPUT: USER1003.BILLINGPRODUCTSCUSTOMER-SRS_sh... [CSV](#)

Runs (1)

Start Time	Status	Duration	Started By	Action
Sep 15, 2020 6:56:57 PM	Completed	47 seconds	Shivam Solanki (IBM)	...

- 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.

The screenshot shows the IBM Cloud Pak for Data interface. The top navigation bar includes the IBM Cloud Pak for Data logo, a search bar, and a 'Add to project' button. The main content area is divided into tabs: Overview, Assets, Environments, Jobs, Access Control, and Settings. The 'Assets' tab is active, displaying a list of data assets. The first asset, 'USER1003.BILLINGPRODUCTSCUSTOMER-SRS_shaped.csv', is highlighted with a red box. The table lists the following assets:

Name	Type	Created by	Last modified
USER1003.BILLINGPRODUCTSCUSTOMER-SRS_shaped.csv	Data Asset	Shivam Solanki (IBM)	Sep 15, 2020, 06:58 PM
USER1003.BILLING-TELCO-SRS	Data Asset	Shivam Solanki (IBM)	Sep 15, 2020, 02:57 PM
USER1003.PRODUCTS-TELCO-SRS	Data Asset	Shivam Solanki (IBM)	Sep 15, 2020, 02:57 PM
USER1003.CUSTOMERS-TELCO-SRS	Data Asset	Shivam Solanki (IBM)	Sep 15, 2020, 02:57 PM
USER1003.BILLINGPRODUCTS-SRS	Data Asset	Shivam Solanki (IBM)	Sep 15, 2020, 02:57 PM
USER1003.BILLINGPRODUCTSCUSTOMER-SRS	Data Asset	Shivam Solanki (IBM)	Sep 15, 2020, 02:57 PM
DS16001998422020483	Connection	Shivam Solanki (IBM)	Sep 15, 2020, 02:57 PM

The sidebar on the right shows a 'Data' section with tabs for Load, Files, and Catalog. A message indicates: 'Drop files here or browse for files to upload.'

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.