Lab Guide

Hands-on-Lab:
Data visualization with data refinery

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

Cloud Pak for Data Workshop 2020

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
- Quickly profile data
- Visualize the data with charts and graphs
- Use Data Flow steps to keep track of your work.
- Save the data refinery flow and create a job

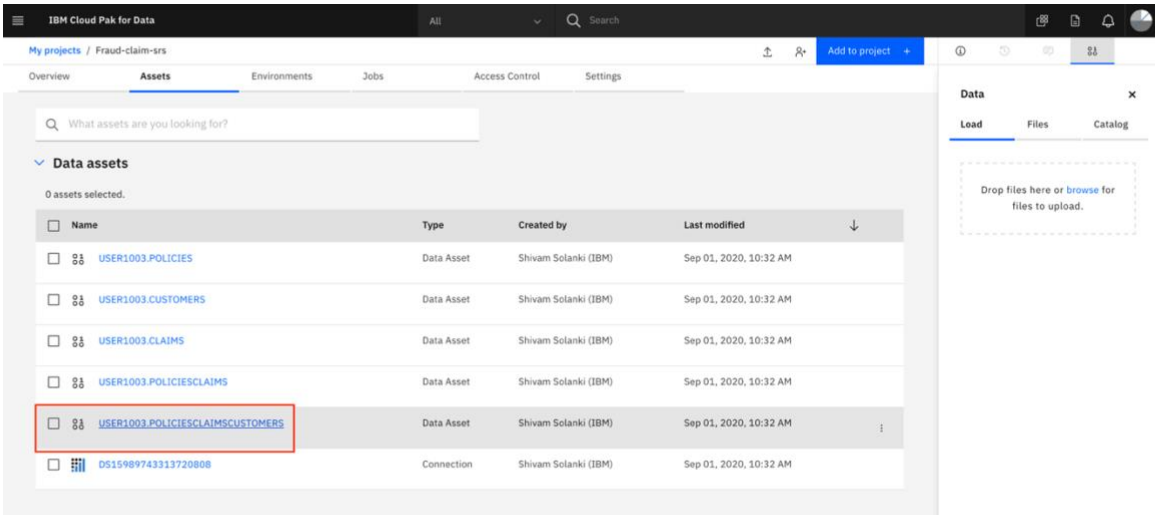
Steps

Step 1. Load the virtualized data into data refinery

1. If you are not already on the Project **Assets** tab from the last lab tutorial on Data Virtualization, open the 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 POLICIES, CLAIMS and CUSTOMERS created in the previous tutorial.

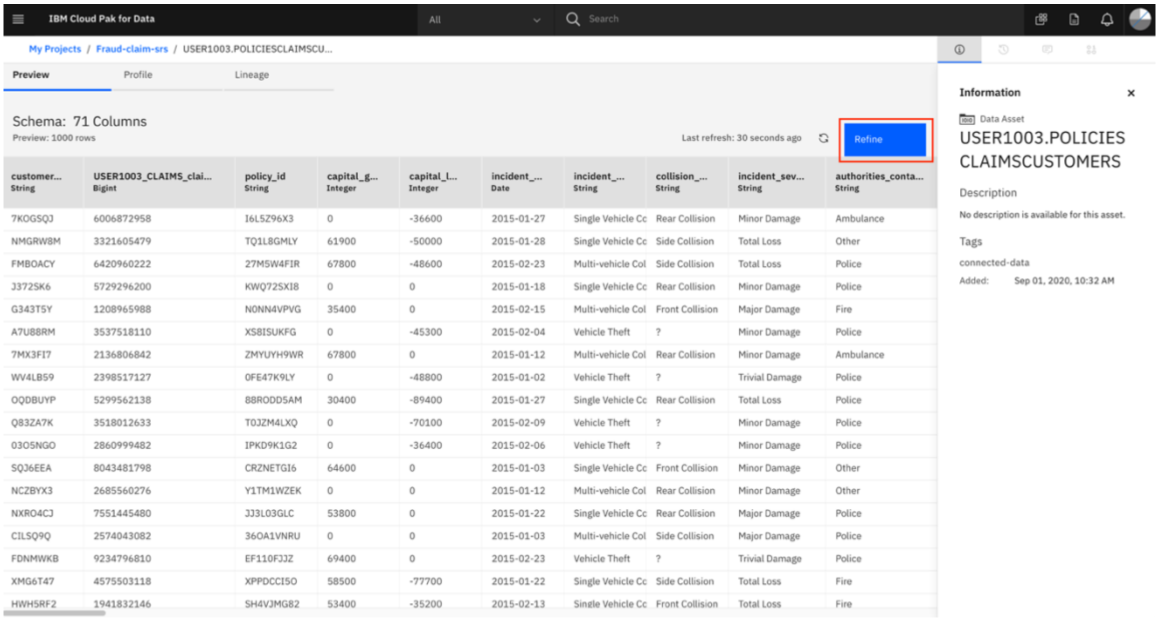


Data assets

0 assets selected.

Name	Type	Created by	Last modified
USER1003.POLICIES	Data Asset	Shivam Solanki (IBM)	Sep 01, 2020, 10:32 AM
USER1003.CUSTOMERS	Data Asset	Shivam Solanki (IBM)	Sep 01, 2020, 10:32 AM
USER1003.CLAIMS	Data Asset	Shivam Solanki (IBM)	Sep 01, 2020, 10:32 AM
USER1003.POLICIESCLAIMS	Data Asset	Shivam Solanki (IBM)	Sep 01, 2020, 10:32 AM
USER1003.POLICIESCLAIMSCUSTOMERS	Data Asset	Shivam Solanki (IBM)	Sep 01, 2020, 10:32 AM
DS15989743313720808	Connection	Shivam Solanki (IBM)	Sep 01, 2020, 10:32 AM

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



My Projects / Fraud-claim-srs / USER1003.POLICIESCLAIMSCU...

Schema: 71 Columns
Preview: 1000 rows

Last refresh: 30 seconds ago

customer... String	USER1003_CLAIMS_clai... Bigint	policy_id String	capital_g... Integer	capital_l... Integer	incident_... Date	incident_... String	collision_... String	incident_sev... String	authorities_conta... String
7KOGSQJ	6006872958	I6L5Z96X3	0	-36600	2015-01-27	Single Vehicle Cc	Rear Collision	Minor Damage	Ambulance
NMGRW8M	3321605479	TQ1L8GMLY	61900	-50000	2015-01-28	Single Vehicle Cc	Side Collision	Total Loss	Other
FMBOACY	6420960222	27M5W4FIR	67800	-48600	2015-02-23	Multi-vehicle Col	Side Collision	Total Loss	Police
J372SK6	5729296200	KWQ72SXI8	0	0	2015-01-18	Single Vehicle Cc	Rear Collision	Minor Damage	Police
G343T5Y	1208965988	N0NN4VPVG	35400	0	2015-02-15	Multi-vehicle Col	Front Collision	Major Damage	Fire
A7U88RM	3537518110	XS8ISUKFG	0	-45300	2015-02-04	Vehicle Theft	?	Minor Damage	Police
7MX3FI7	2136806842	ZMYUYH9WR	67800	0	2015-01-12	Multi-vehicle Col	Rear Collision	Minor Damage	Ambulance
WV4LB59	2398517127	0FE47K9LY	0	-48800	2015-01-02	Vehicle Theft	?	Trivial Damage	Police
OQDBUYP	5299562138	88RODD5AM	30400	-89400	2015-01-27	Single Vehicle Cc	Rear Collision	Total Loss	Police
Q83ZA7K	3518012633	T0JZM4LXQ	0	-70100	2015-02-09	Vehicle Theft	?	Minor Damage	Police
03O5NGO	2860999482	IPKD9K1G2	0	-36400	2015-02-06	Vehicle Theft	?	Minor Damage	Police
SQJ6EEA	8043481798	CRZNETGI6	64600	0	2015-01-03	Single Vehicle Cc	Front Collision	Minor Damage	Other
NCZBYX3	2685560276	Y1TM1WZEK	0	0	2015-01-12	Multi-vehicle Col	Rear Collision	Minor Damage	Other
NXRO4CJ	7551445480	JJ3L03GLC	53800	0	2015-01-22	Single Vehicle Cc	Rear Collision	Major Damage	Police
CILSQ9Q	2574043082	36OA1VNRU	0	0	2015-01-03	Multi-vehicle Col	Side Collision	Major Damage	Police
FDNMWKB	9234796810	EF110FJJZ	69400	0	2015-02-23	Vehicle Theft	?	Trivial Damage	Police
XMG6T47	4575503118	XPPDCCI5O	58500	-77700	2015-01-22	Single Vehicle Cc	Side Collision	Total Loss	Fire
HWH5RF2	1941832146	SH4VJMG82	53400	-35200	2015-02-13	Single Vehicle Cc	Front Collision	Total Loss	Fire

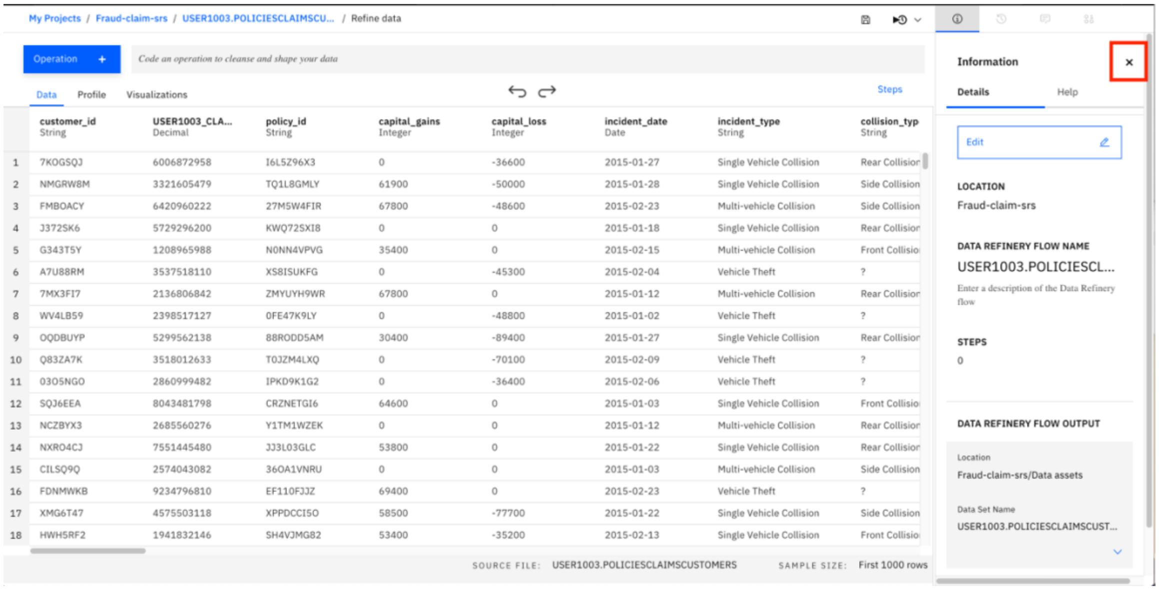
Information

Data Asset
USER1003.POLICIES
CLAIMSCUSTOMERS

Description
No description is available for this asset.

Tags
connected-data
Added: Sep 01, 2020, 10:32 AM

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



My Projects / Fraud-claim-srs / USER1003.POLICIESCLAIMSCU... / Refine data

Code an operation to cleanse and shape your data

	customer_id String	USER1003_CLA... Decimal	policy_id String	capital_gains Integer	capital_loss Integer	incident_date Date	incident_type String	collision_typ String
1	7KOGSQJ	6006872958	I6L5Z96X3	0	-36600	2015-01-27	Single Vehicle Collision	Rear Collision
2	NMGRW8M	3321605479	TQ1L8GMLY	61900	-50000	2015-01-28	Single Vehicle Collision	Side Collision
3	FMBOACY	6420960222	27M5W4FIR	67800	-48600	2015-02-23	Multi-vehicle Collision	Side Collision
4	J372SK6	5729296200	KWQ72SXI8	0	0	2015-01-18	Single Vehicle Collision	Rear Collision
5	G343T5Y	1208965988	N0NN4VPVG	35400	0	2015-02-15	Multi-vehicle Collision	Front Collisio
6	A7U88RM	3537518110	XS8ISUKFG	0	-45300	2015-02-04	Vehicle Theft	?
7	7MX3FI7	2136806842	ZMYUYH9WR	67800	0	2015-01-12	Multi-vehicle Collision	Rear Collision
8	WV4LB59	2398517127	0FE47K9LY	0	-48800	2015-01-02	Vehicle Theft	?
9	OQDBUYP	5299562138	88RODD5AM	30400	-89400	2015-01-27	Single Vehicle Collision	Rear Collision
10	Q83ZA7K	3518012633	T0JZM4LXQ	0	-70100	2015-02-09	Vehicle Theft	?
11	03O5NGO	2860999482	IPKD9K1G2	0	-36400	2015-02-06	Vehicle Theft	?
12	SQJ6EEA	8043481798	CRZNETGI6	64600	0	2015-01-03	Single Vehicle Collision	Front Collisio
13	NCZBYX3	2685560276	Y1TM1WZEK	0	0	2015-01-12	Multi-vehicle Collision	Rear Collision
14	NXRO4CJ	7551445480	JJ3L03GLC	53800	0	2015-01-22	Single Vehicle Collision	Rear Collision
15	CILSQ9Q	2574043082	36OA1VNRU	0	0	2015-01-03	Multi-vehicle Collision	Side Collision
16	FDNMWKB	9234796810	EF110FJJZ	69400	0	2015-02-23	Vehicle Theft	?
17	XMG6T47	4575503118	XPPDCCI5O	58500	-77700	2015-01-22	Single Vehicle Collision	Side Collision
18	HWH5RF2	1941832146	SH4VJMG82	53400	-35200	2015-02-13	Single Vehicle Collision	Front Collisio

SOURCE FILE: USER1003.POLICIESCLAIMSCUSTOMERS   SAMPLE SIZE: First 1000 rows

Information ×

Details   Help

LOCATION
Fraud-claim-srs

DATA REFINERY FLOW NAME
USER1003.POLICIESCL...
Enter a description of the Data Refinery flow

STEPS
0

DATA REFINERY FLOW OUTPUT
Location
Fraud-claim-srs/Data assets

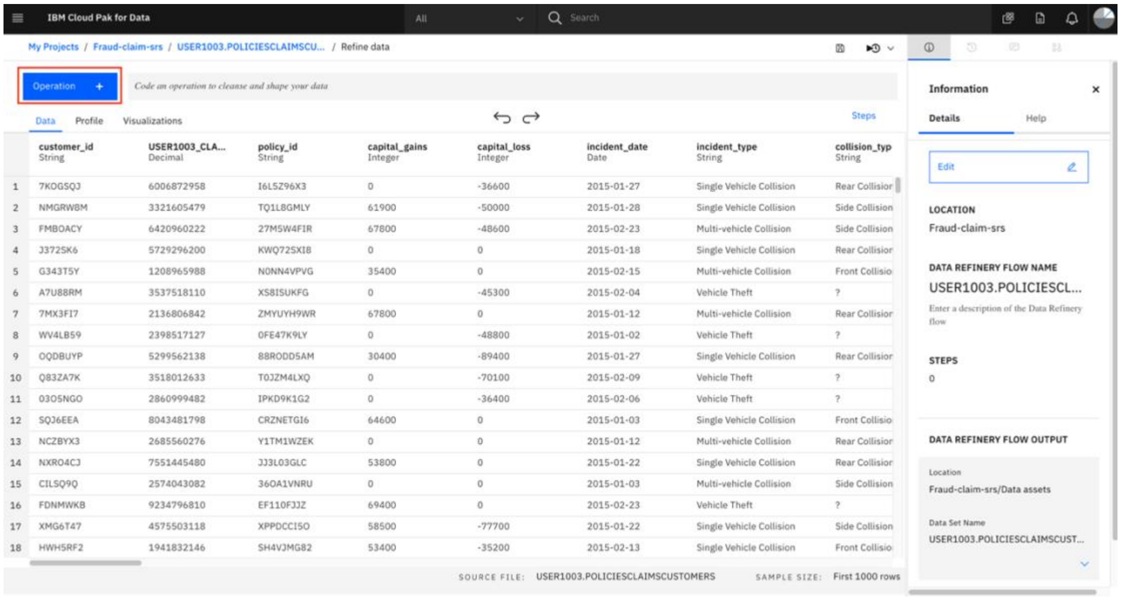
Data Set Name
USER1003.POLICIESCLAIMSCUST...

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

Step 2. Refine the data

We’ll start out on the Data tab.

1. Click the **Operation+** button.



IBM Cloud Pak for Data
All
Search
My Projects / Fraud-claim-srs / USER1003.POLICIESCLAIMSCU... / Refine data
Operation +
Code an operation to cleanse and shape your data
Data
Profile
Visualizations
Steps
Information
Details
Help
Edit

	customer_id String	USER1003_CLA... Decimal	policy_id String	capital_gains Integer	capital_loss Integer	incident_date Date	incident_type String	collision_typ String
1	7KOGSQJ	6006872958	I6L5Z96X3	0	-36600	2015-01-27	Single Vehicle Collision	Rear Collisior
2	NMGRW8M	3321605479	TQ1L8GMLY	61900	-50000	2015-01-28	Single Vehicle Collision	Side Collision
3	FMBOACY	6420960222	27M5W4FIR	67800	-48600	2015-02-23	Multi-vehicle Collision	Side Collision
4	J372SK6	5729296200	KWQ72SXI8	0	0	2015-01-18	Single Vehicle Collision	Rear Collisior
5	G343T5Y	1208965988	NONN4VPVG	35400	0	2015-02-15	Multi-vehicle Collision	Front Collisio
6	A7U88RM	3537518110	XS8ISUKFG	0	-45300	2015-02-04	Vehicle Theft	?
7	7MX3FI7	2136806842	ZMYUYH9WR	67800	0	2015-01-12	Multi-vehicle Collision	Rear Collisior
8	WV4LB59	2398517127	0FE47K9LY	0	-48800	2015-01-02	Vehicle Theft	?
9	OQDBUYP	5299562138	88RODD5AM	30400	-89400	2015-01-27	Single Vehicle Collision	Rear Collisior
10	Q83ZA7K	3518012633	T0JZM4LXQ	0	-70100	2015-02-09	Vehicle Theft	?
11	03O5NGO	2860999482	IPKD9K1G2	0	-36400	2015-02-06	Vehicle Theft	?
12	SQJ6EEA	8043481798	CRZNETGI6	64600	0	2015-01-03	Single Vehicle Collision	Front Collisio
13	NCZBYX3	2685560276	Y1TM1WZEK	0	0	2015-01-12	Multi-vehicle Collision	Rear Collisior
14	NXRO4CJ	7551445480	JJ3L03GLC	53800	0	2015-01-22	Single Vehicle Collision	Rear Collisior
15	CILSQ9Q	2574043082	36OA1VNRU	0	0	2015-01-03	Multi-vehicle Collision	Side Collision
16	FDNMWKB	9234796810	EF110FJJZ	69400	0	2015-02-23	Vehicle Theft	?
17	XMG6T47	4575503118	XPPDCCI5O	58500	-77700	2015-01-22	Single Vehicle Collision	Side Collision
18	HWH5RF2	1941832146	SH4VJMG82	53400	-35200	2015-02-13	Single Vehicle Collision	Front Collisio

SOURCE FILE: USER1003.POLICIESCLAIMSCUSTOMERS   SAMPLE SIZE: First 1000 rows

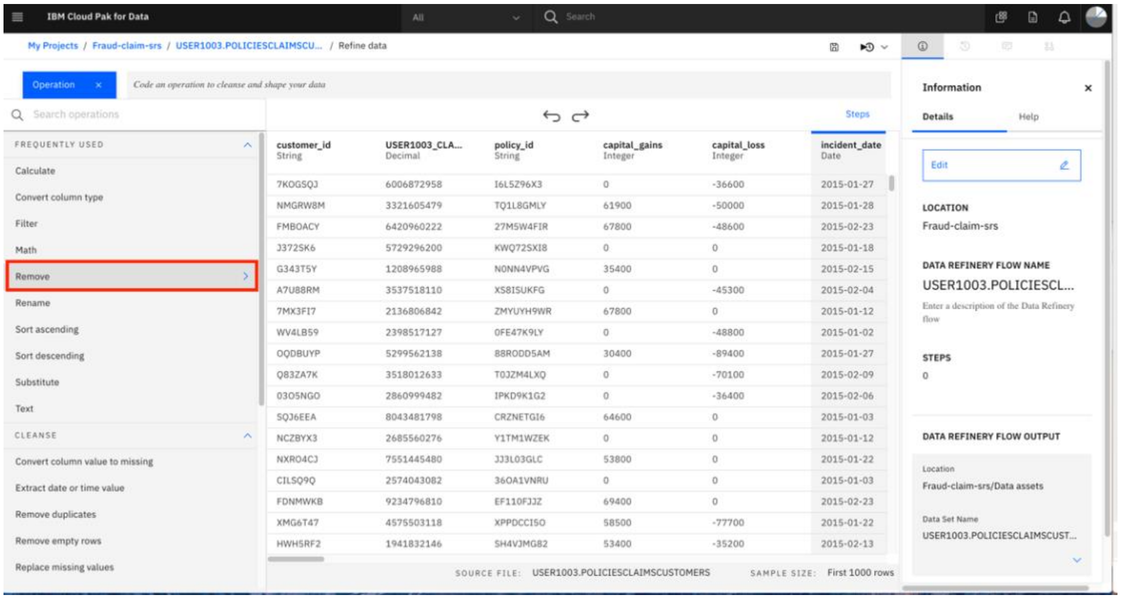
LOCATION
Fraud-claim-srs

DATA REFINERY FLOW NAME
USER1003.POLICIESCL...
Enter a description of the Data Refinery flow

STEPS
0

DATA REFINERY FLOW OUTPUT
Location
Fraud-claim-srs/Data assets
Data Set Name
USER1003.POLICIESCLAIMSCUST...

2. We notice that there are some attributes pertaining to personal information such as *name, phone_number, email_address, national_id, credit_card information* etc. These attributes are not required for building a machine learning fraud prediction model and contains sensitive information so we should drop them. Choose the Operator **Remove**.



customer_id String	USER1003_CLA... Decimal	policy_id String	capital_gains Integer	capital_loss Integer	incident_date Date
7KOGSQJ	6006872958	I6L5Z96X3	0	-36600	2015-01-27
NMGRW8M	3321605479	TQ1L8GMLY	61900	-50000	2015-01-28
FMBOACY	6420960222	27M5W4FIR	67800	-48600	2015-02-23
J372SK6	5729296200	KWQ72SXI8	0	0	2015-01-18
G343T5Y	1208965988	NONN4VPVG	35400	0	2015-02-15
A7U88RM	3537518110	XS8ISUKFG	0	-45300	2015-02-04
7MX3FI7	2136806842	ZMYUYH9WR	67800	0	2015-01-12
WV4LB59	2398517127	0FE47K9LY	0	-48800	2015-01-02
OQDBUYP	5299562138	88RODD5AM	30400	-89400	2015-01-27
Q83ZA7K	3518012633	T0JZM4LXQ	0	-70100	2015-02-09
03O5NGO	2860999482	IPKD9K1G2	0	-36400	2015-02-06
SQJ6EEA	8043481798	CRZNETGI6	64600	0	2015-01-03
NCZBYX3	2685560276	Y1TM1WZEK	0	0	2015-01-12
NXRO4CJ	7551445480	JJ3L03GLC	53800	0	2015-01-22
CILSQ9Q	2574043082	36OA1VNRU	0	0	2015-01-03
FDNMWKB	9234796810	EF110FJJZ	69400	0	2015-02-23
XMG6T47	4575503118	XPPDCCI5O	58500	-77700	2015-01-22
HWH5RF2	1941832146	SH4VJMG82	53400	-35200	2015-02-13

SOURCE FILE: USER1003.POLICIESCLAIMSCUSTOMERS   SAMPLE SIZE: First 1000 rows

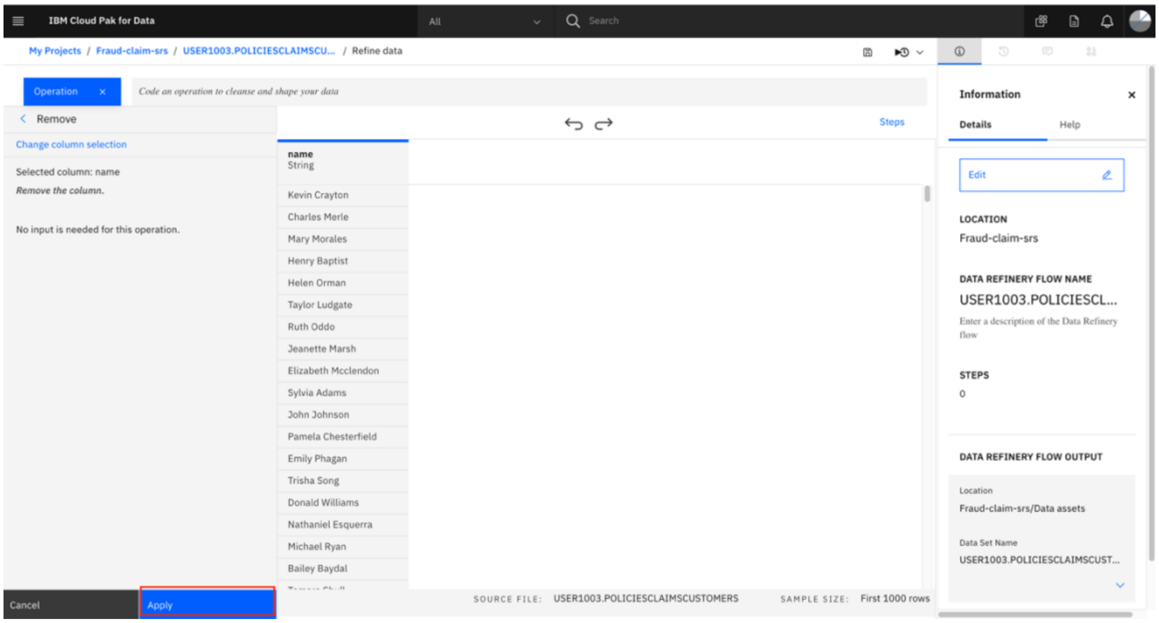
LOCATION
Fraud-claim-srs

DATA REFINERY FLOW NAME
USER1003.POLICIESCL...
Enter a description of the Data Refinery flow

STEPS
0

DATA REFINERY FLOW OUTPUT
Location
Fraud-claim-srs/Data assets
Data Set Name
USER1003.POLICIESCLAIMSCUST...

3. Type **name** in the Select Column dropdown and click on **Apply.**

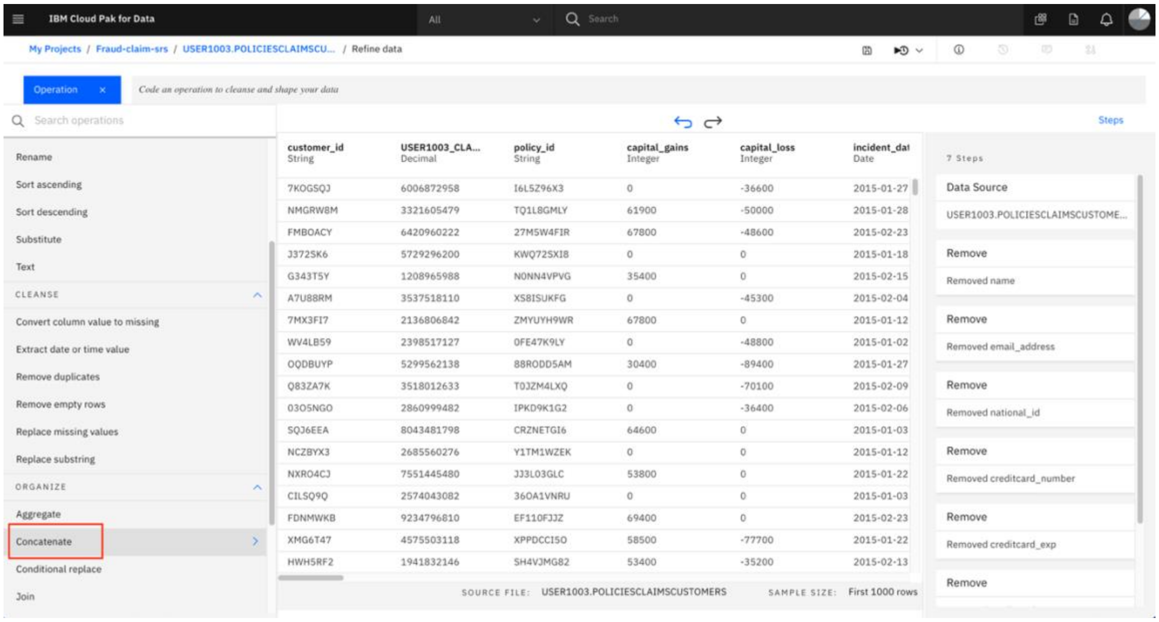


4. Repeat the process to remove the columns **email_address, phone_number, national_id, creditcard_number, creditcard_exp**, and **creditcard_type.**

   After removing these columns, you can see the steps on the right-hand side where you can track the data refinery flow. Click on Steps to hide it.

5. Scroll to the right and observe the columns **country, street_address, city, state, state_code**, and **insured_zip.** We are going to concatenate all these attributes to create a single column named **address**.

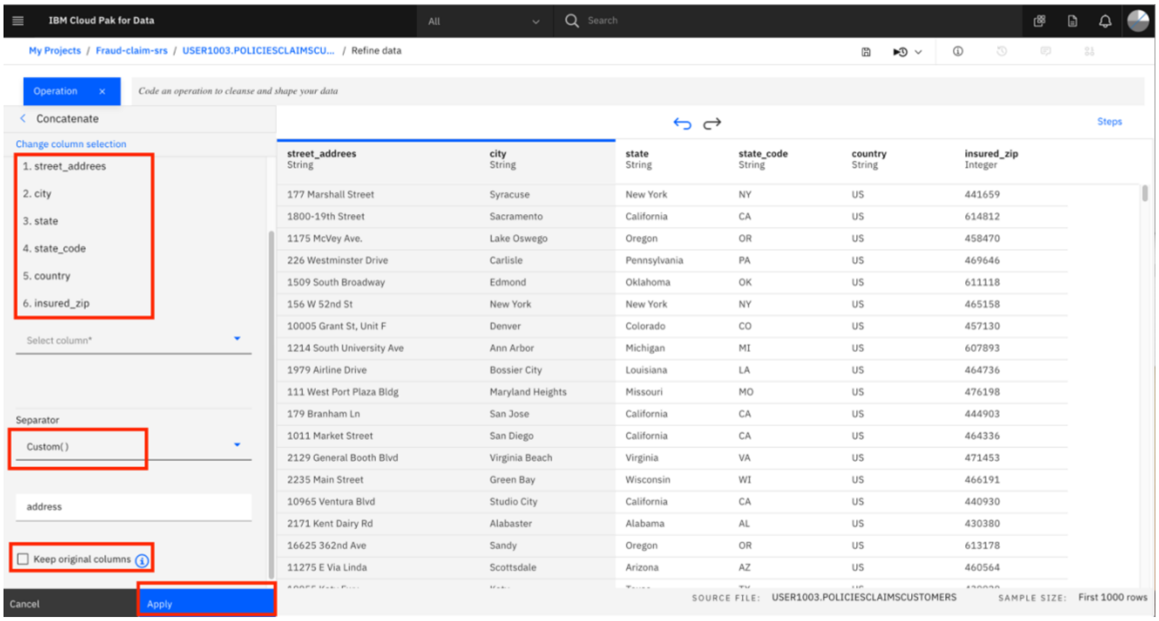
6. Click the **Operation+** button and select **Concatenate**.



customer_id String	USER1003_CLA... Decimal	policy_id String	capital_gains Integer	capital_loss Integer	incident_dat Date
7KOGSQJ	6006872958	I6L5Z96X3	0	-36600	2015-01-27
NMGRW8M	3321605479	TQ1L8GMLY	61900	-50000	2015-01-28
FMBOACY	6420960222	27M5W4FIR	67800	-48600	2015-02-23
J372SK6	5729296200	KWQ72SXI8	0	0	2015-01-18
G343T5Y	1208965988	NONN4VPVG	35400	0	2015-02-15
A7U88RM	3537518110	XS8ISUKFG	0	-45300	2015-02-04
7MX3FI7	2136806842	ZMYUYH9WR	67800	0	2015-01-12
WV4LB59	2398517127	0FE47K9LY	0	-48800	2015-01-02
OQDBUYP	5299562138	88RODD5AM	30400	-89400	2015-01-27
Q83ZA7K	3518012633	T0JZM4LXQ	0	-70100	2015-02-09
03O5NGO	2860999482	IPKD9K1G2	0	-36400	2015-02-06
SQJ6EEA	8043481798	CRZNETGI6	64600	0	2015-01-03
NCZBYX3	2685560276	Y1TM1WZEK	0	0	2015-01-12
NXRO4CJ	7551445480	JJ3L03GLC	53800	0	2015-01-22
CILSQ9Q	2574043082	36OA1VNRU	0	0	2015-01-03
FDNMWKB	9234796810	EF110FJJZ	69400	0	2015-02-23
XMG6T47	4575503118	XPPDCCI5O	58500	-77700	2015-01-22
HWH5RF2	1941832146	SH4VJMG82	53400	-35200	2015-02-13

SOURCE FILE: USER1003.POLICIESCLAIMSCUSTOMERS SAMPLE SIZE: First 1000 rows

7. Select the column **street_address** and then click on **Next.** Then select **city, state, state_code, insured_zip**, and **country.**
8. Enter **Space** in the Separator and **address** in the Name of the concatenated column. Uncheck the **Keep original columns** and click on **Apply.**


IBM Cloud Pak for Data
All
Search
My Projects / Fraud-claim-srs / USER1003.POLICIESCLAIMSCU... / Refine data
Operation ×
Code an operation to cleanse and shape your data
Steps
Cancel
Apply

< Concatenate

Change column selection

1. street_addrees
2. city
3. state
4. state_code
5. country
6. insured_zip

Select column*

Separator: Custom( )

address

☐ Keep original columns

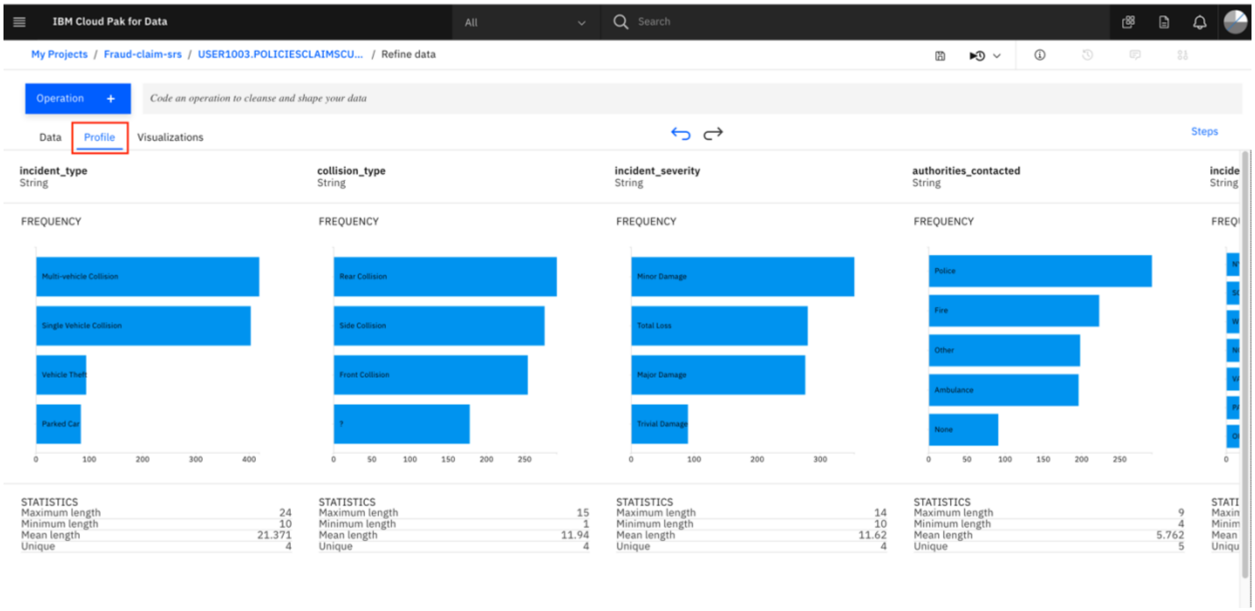
street_addrees String	city String	state String	state_code String	country String	insured_zip Integer
177 Marshall Street	Syracuse	New York	NY	US	441659
1800-19th Street	Sacramento	California	CA	US	614812
1175 McVey Ave.	Lake Oswego	Oregon	OR	US	458470
226 Westminster Drive	Carlisle	Pennsylvania	PA	US	469646
1509 South Broadway	Edmond	Oklahoma	OK	US	611118
156 W 52nd St	New York	New York	NY	US	465158
10005 Grant St, Unit F	Denver	Colorado	CO	US	457130
1214 South University Ave	Ann Arbor	Michigan	MI	US	607893
1979 Airline Drive	Bossier City	Louisiana	LA	US	464736
111 West Port Plaza Bldg	Maryland Heights	Missouri	MO	US	476198
179 Branham Ln	San Jose	California	CA	US	444903
1011 Market Street	San Diego	California	CA	US	464336
2129 General Booth Blvd	Virginia Beach	Virginia	VA	US	471453
2235 Main Street	Green Bay	Wisconsin	WI	US	466191
10965 Ventura Blvd	Studio City	California	CA	US	440930
2171 Kent Dairy Rd	Alabaster	Alabama	AL	US	430380
16625 362nd Ave	Sandy	Oregon	OR	US	613178
11275 E Via Linda	Scottsdale	Arizona	AZ	US	460564

SOURCE FILE: USER1003.POLICIESCLAIMSCUSTOMERS SAMPLE SIZE: First 1000 rows

9. Click on the save button to save the progress.

Step 3. Profile the data

The **Profile** tab displays a quick view of several histograms about the data.



STATISTICS	incident_type	collision_type	incident_severity	authorities_contacted
Maximum length	24	15	14	9
Minimum length	10	1	10	4
Mean length	21.371	11.94	11.62	5.762
Unique	4	4	4	5

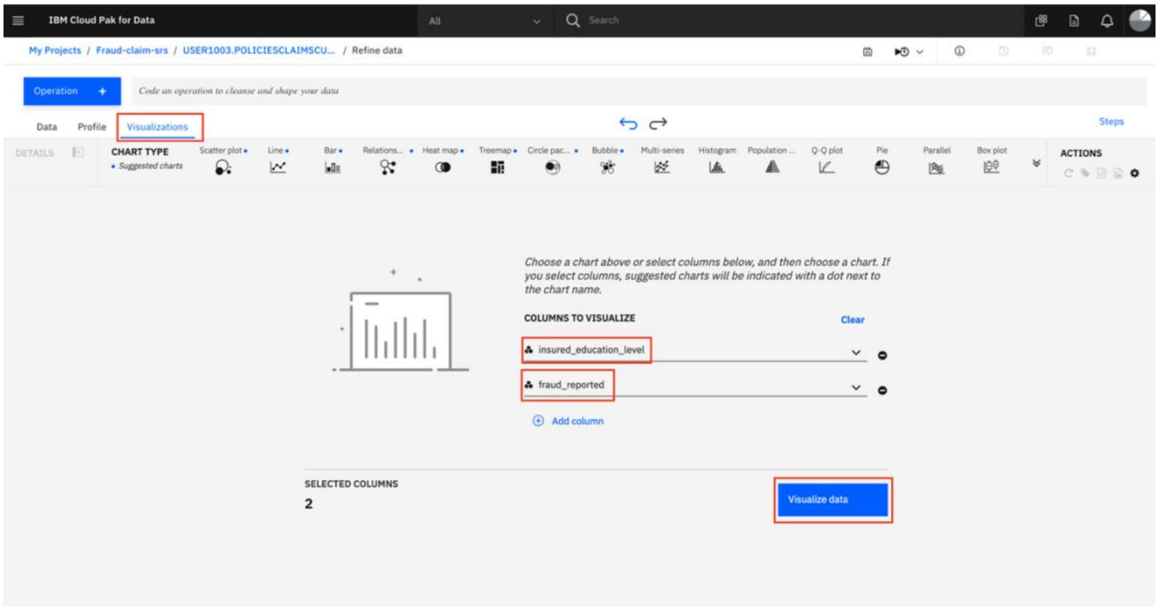
You can get insights into the data from the histograms:

- Highest number of incidents are collisions
- Rear collision dominates the collision type. There is a missing type denoted by "**?". **We will take care of this missing value using the **Data** tab in the next step
- You can see the distribution of incident severity
- Scroll right to observe the distribution of **fraud_reported.** Only 1/4th of the auto insurance claims were fraudulent

Step 4. Visualize the data

1. Select the **Visualizations** tab.

2. Under the **Columns to Visualize,** select **insured_education_level.** Click on **Add column** and select **fraud_reported.** Click on **Visualize data.**



Choose a chart above or select columns below, and then choose a chart. If you select columns, suggested charts will be indicated with a dot next to the chart name.

COLUMNS TO VISUALIZE

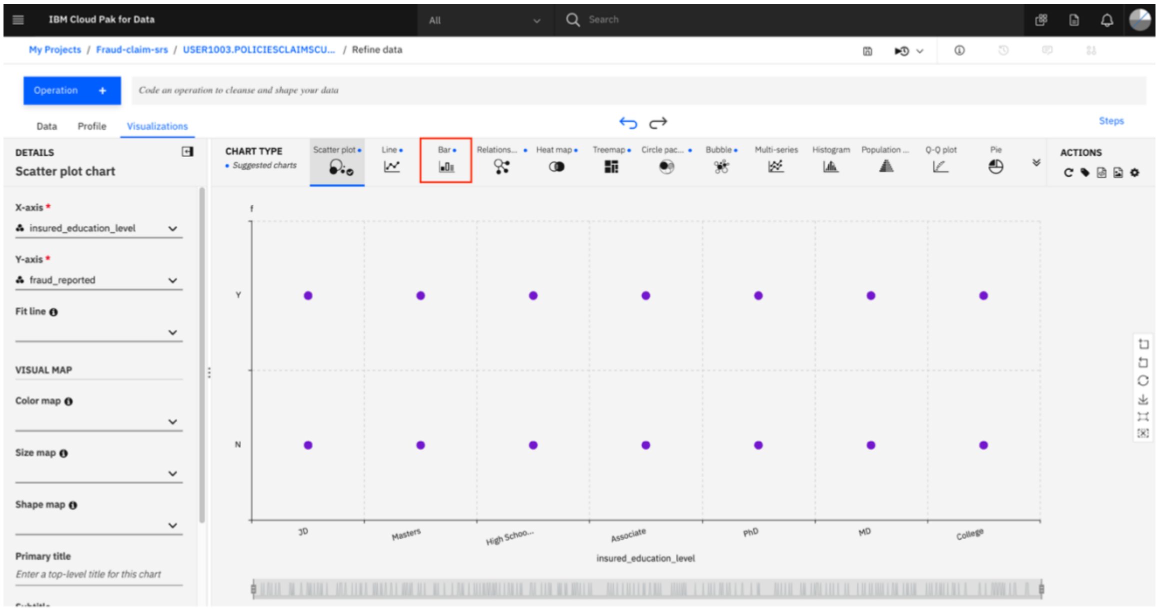
insured_education_level

fraud_reported

SELECTED COLUMNS

2

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



DETAILS

Scatter plot chart

X-axis *: insured_education_level

Y-axis *: fraud_reported

Fit line

VISUAL MAP

Color map

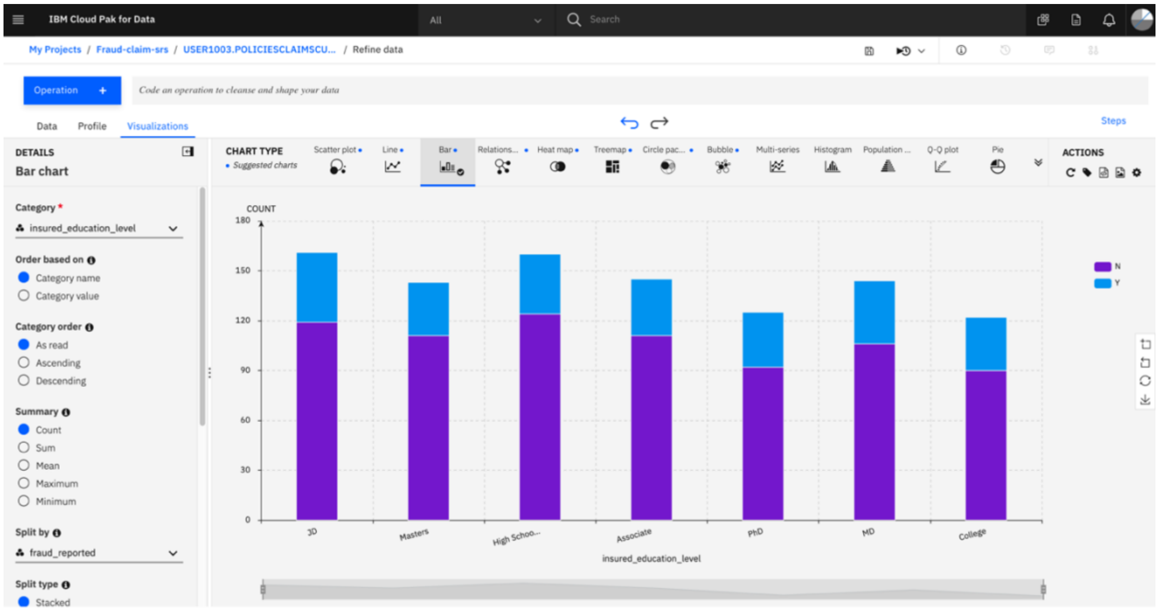
Size map

Shape map

Primary title

Enter a top-level title for this chart

Subtitle



DETAILS

Bar chart

Category *

insured_education_level

Order based on

● Category name
○ Category value

Category order

● As read
○ Ascending
○ Descending

Summary

● Count
○ Sum
○ Mean
○ Maximum
○ Minimum

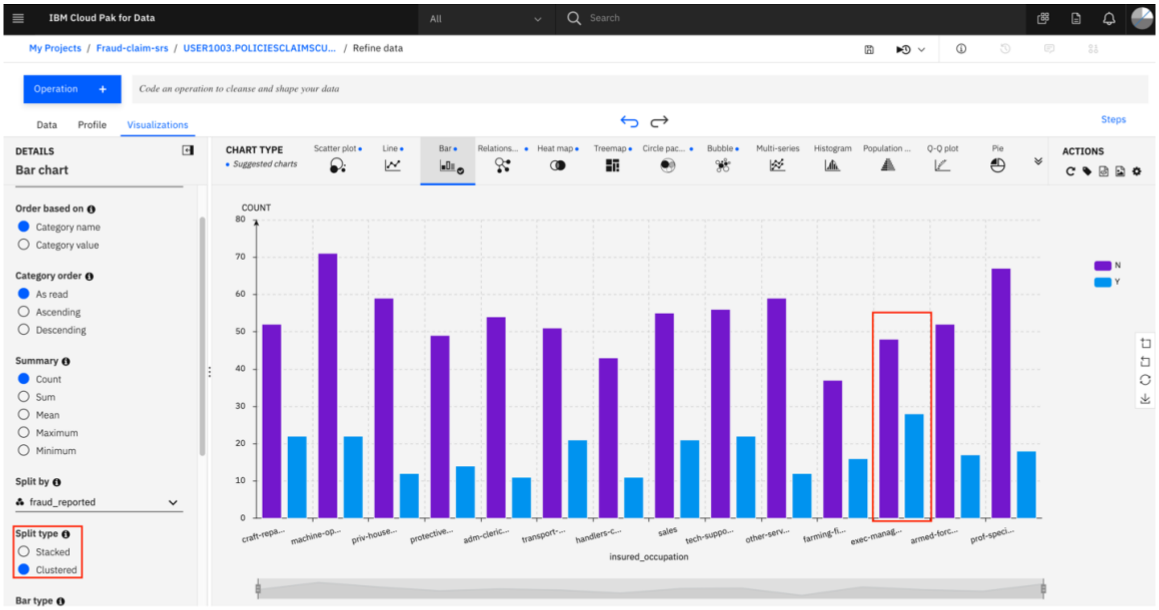
Split by

fraud_reported

Split type

● Stacked

4. We will now visualize other columns. Select **insured_occupation** in the **Category**. Choose the **Split type** as **Clustered**.



DETAILS

Bar chart

Order based on

● Category name
○ Category value

Category order

● As read
○ Ascending
○ Descending

Summary

● Count
○ Sum
○ Mean
○ Maximum
○ Minimum

Split by

fraud_reported

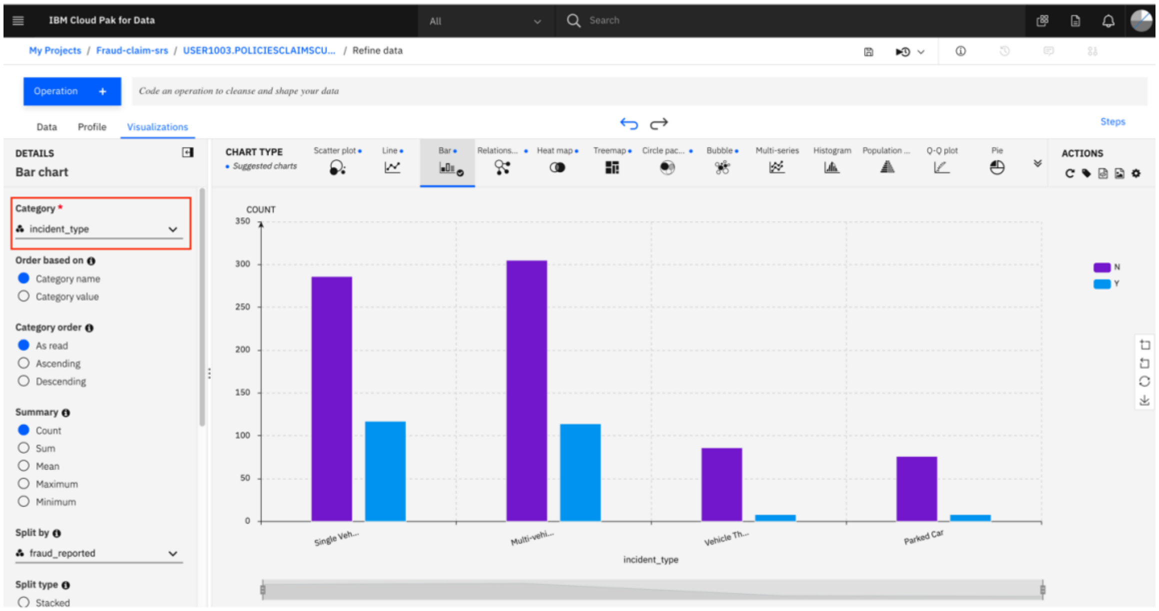
Split type

○ Stacked
● Clustered

Bar type

It can be observed from the data that people in exec-managerial positions have a greater number of frauds compared to other occupations.

5. Select **incident_type** in the **Category.**



Category *
incident_type

Order based on
● Category name
○ Category value

Category order
● As read
○ Ascending
○ Descending

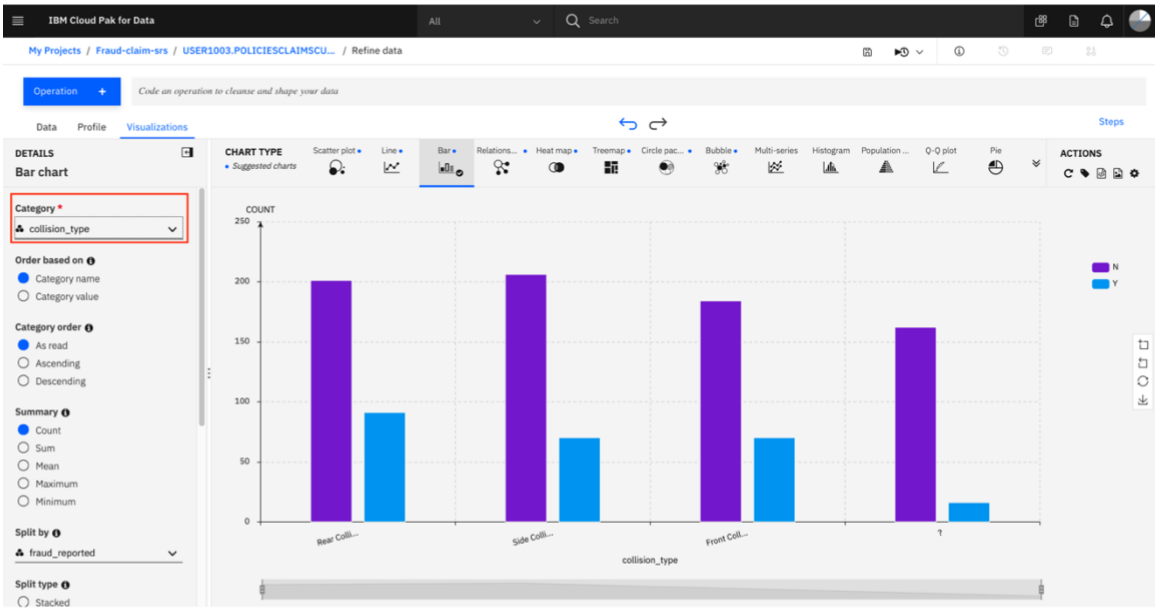
Summary
● Count
○ Sum
○ Mean
○ Maximum
○ Minimum

Split by
fraud_reported

Split type
○ Stacked

Multi-vehicle and single vehicle collisions have a greater number of frauds compared to parked and vehicle theft. One of the reasons could be that in a collision, there is a high possibility of more damage to a car, as well as the passengers, and hence the need to file false insurance claims.

6. Select **collision_type** in the **Category.**



Category *
collision_type

Order based on
● Category name
○ Category value

Category order
● As read
○ Ascending
○ Descending

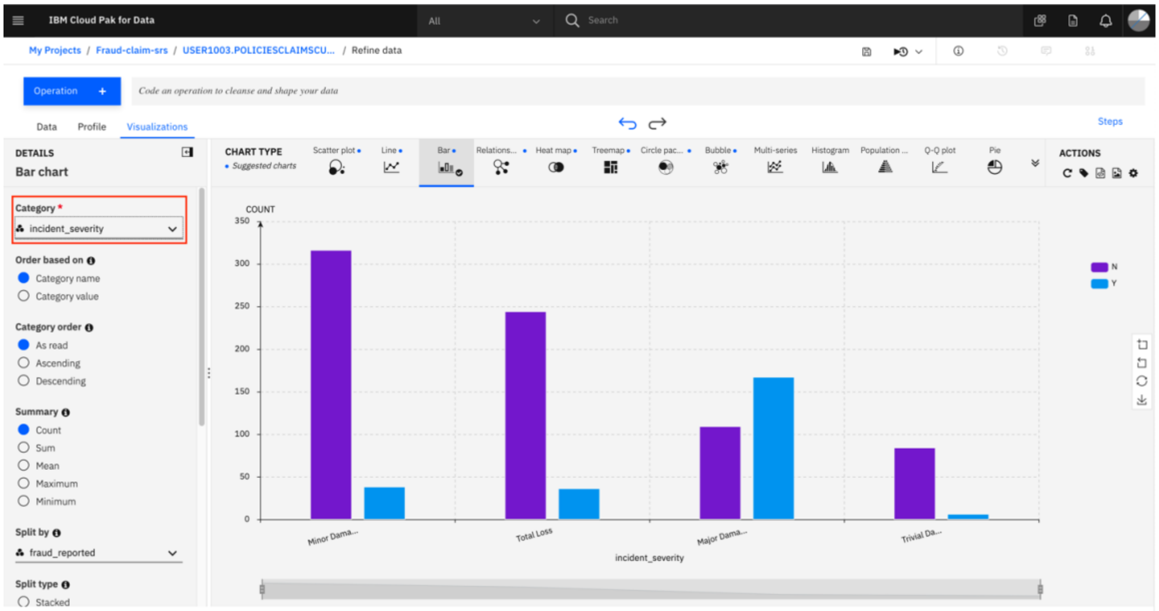
Summary
● Count
○ Sum
○ Mean
○ Maximum
○ Minimum

Split by
fraud_reported

Split type
○ Stacked

While there are significant numbers of false claims in front and side collisions, rear collisions are the highest. This data is for the US and there, many people use dash cams while driving to record whatever is happening in front of their vehicle. In rear collisions, the footage from dash cams is not very helpful to exclusively prove whose mistake it was (insurance owner or the other car owner). Maybe that is the reason for more fraudulent claims in rear collisions.

7. Select **incident_severity** in the **Category.**



Category *
incident_severity

Order based on
● Category name
○ Category value

Category order
● As read
○ Ascending
○ Descending

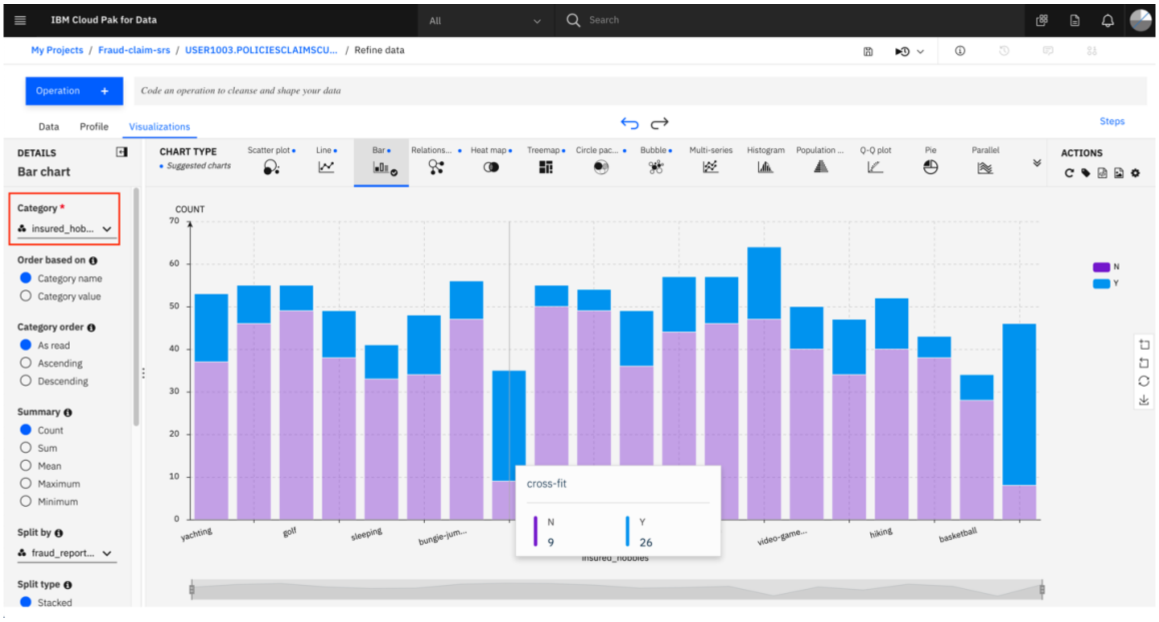
Summary
● Count
○ Sum
○ Mean
○ Maximum
○ Minimum

Split by
fraud_reported

Split type
○ Stacked

Here, compared to minor damage, total loss and trivial damage, fraudulent claims are highest in major damage. One reason could be that the high amount of repair cost which will be incurred by the insurer due to major damage.

8. Select **insured_hobbies** in the **Category.**



Category *
insured_hob...

Order based on
● Category name
○ Category value

Category order
● As read
○ Ascending
○ Descending

Summary
● Count
○ Sum
○ Mean
○ Maximum
○ Minimum

Split by
fraud_report...

Split type
● Stacked

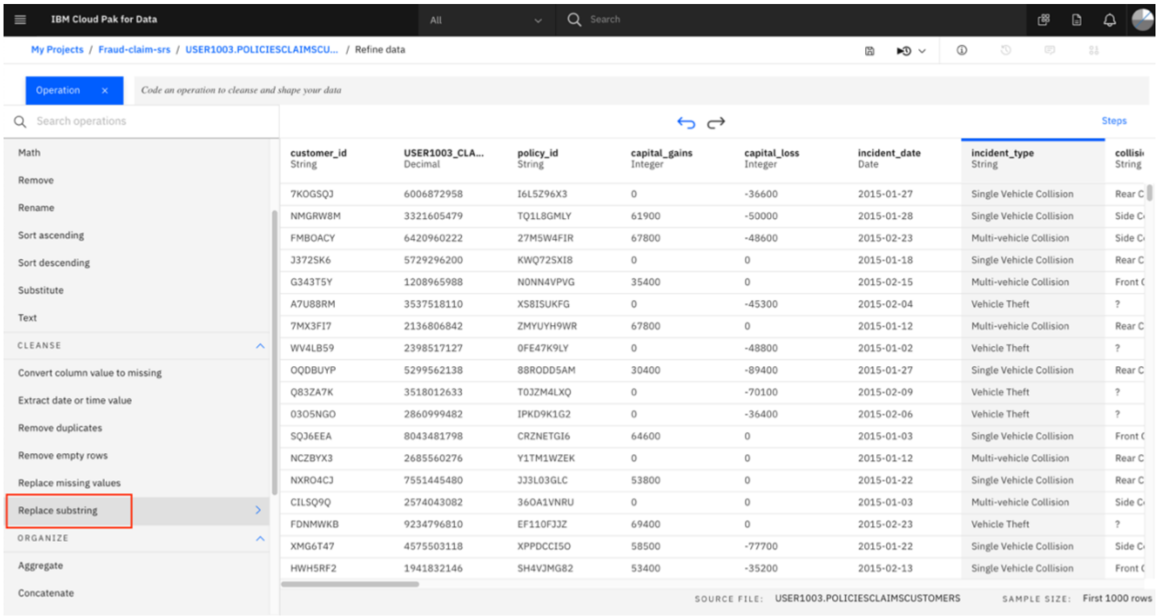
cross-fit
N 9
Y 26

One thing which is striking in this graph is that people with **chess** and **cross-fit** as hobby have extremely high number of fraudulent claims.

We have observed some of the important factors that can be used in the creating a machine learning model to predict fraud claims. But before moving on to the next step of modeling, we will use the insights gained from the visualization for further data cleaning and feature engineering.

Step 5. Refine the data based on Visualization steps

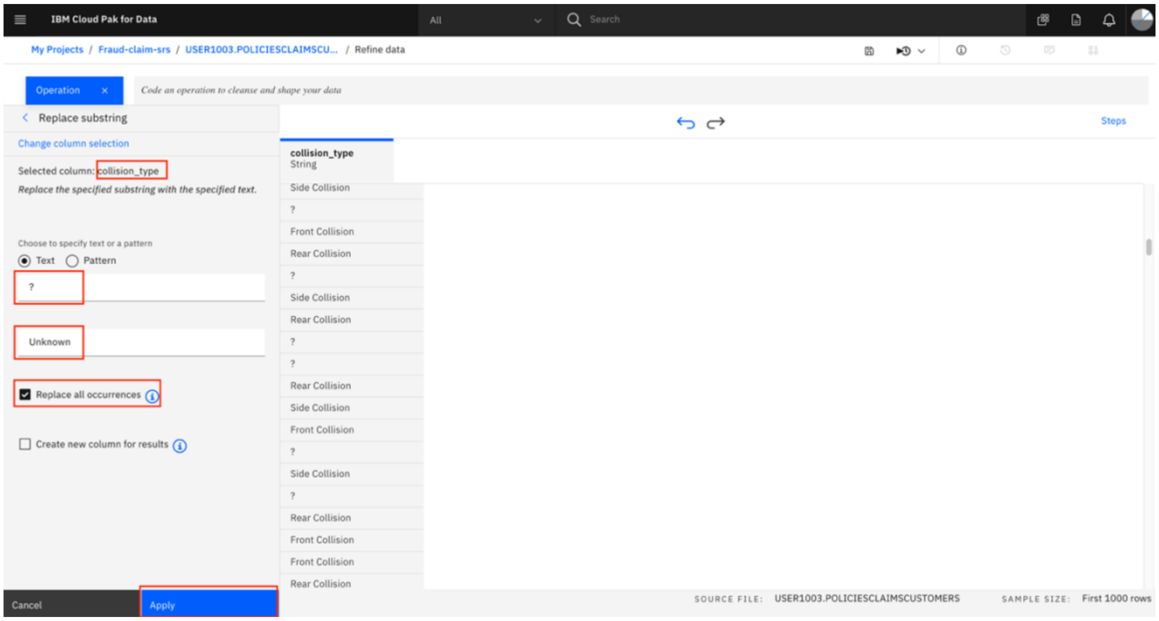
1. Click on the **Data** tab.
2. Click the **Operation+** button.
3. Select **Replace substring**.



customer_id String	USER1003_CLA... Decimal	policy_id String	capital_gains Integer	capital_loss Integer	incident_date Date	incident_type String	collisi String
7KOGSQJ	6006872958	I6L5Z96X3	0	-36600	2015-01-27	Single Vehicle Collision	Rear C
NMGRW8M	3321605479	TQ1L8GMLY	61900	-50000	2015-01-28	Single Vehicle Collision	Side C
FMBOACY	6420960222	27M5W4FIR	67800	-48600	2015-02-23	Multi-vehicle Collision	Side C
J372SK6	5729296200	KWQ72SXI8	0	0	2015-01-18	Single Vehicle Collision	Rear C
G343T5Y	1208965988	NONN4VPVG	35400	0	2015-02-15	Multi-vehicle Collision	Front C
A7U88RM	3537518110	XS8ISUKFG	0	-45300	2015-02-04	Vehicle Theft	?
7MX3FI7	2136806842	ZMYUYH9WR	67800	0	2015-01-12	Multi-vehicle Collision	Rear C
WV4LB59	2398517127	0FE47K9LY	0	-48800	2015-01-02	Vehicle Theft	?
OQDBUYP	5299562138	88RODD5AM	30400	-89400	2015-01-27	Single Vehicle Collision	Rear C
Q83ZA7K	3518012633	T0JZM4LXQ	0	-70100	2015-02-09	Vehicle Theft	?
03O5NGO	2860999482	IPKD9K1G2	0	-36400	2015-02-06	Vehicle Theft	?
SQJ6EEA	8043481798	CRZNETGI6	64600	0	2015-01-03	Single Vehicle Collision	Front C
NCZBYX3	2685560276	Y1TM1WZEK	0	0	2015-01-12	Multi-vehicle Collision	Rear C
NXRO4CJ	7551445480	JJ3L03GLC	53800	0	2015-01-22	Single Vehicle Collision	Rear C
CILSQ9Q	2574043082	36OA1VNRU	0	0	2015-01-03	Multi-vehicle Collision	Side C
FDNMWKB	9234796810	EF110FJJZ	69400	0	2015-02-23	Vehicle Theft	?
XMG6T47	4575503118	XPPDCCI5O	58500	-77700	2015-01-22	Single Vehicle Collision	Side C
HWH5RF2	1941832146	SH4VJMG82	53400	-35200	2015-02-13	Single Vehicle Collision	Front C

SOURCE FILE: USER1003.POLICIESCLAIMSCUSTOMERS SAMPLE SIZE: First 1000 rows

4. Select **collision_type** in the column and click **Next**.
5. Enter "**?**" in the Text to replace, and **Unknown** in the replacement text. Check **Replace all occurrences** if not selected already and click **Apply.**


IBM Cloud Pak for Data
All
Search
My Projects / Fraud-claim-srs / USER1003.POLICIESCLAIMSCU... / Refine data
Operation ×
Code an operation to cleanse and shape your data
Replace substring
Change column selection
Steps
Cancel
Apply

Selected column: collision_type

Replace the specified substring with the specified text.

Choose to specify text or a pattern

● Text ○ Pattern

?

Unknown

☑ Replace all occurrences

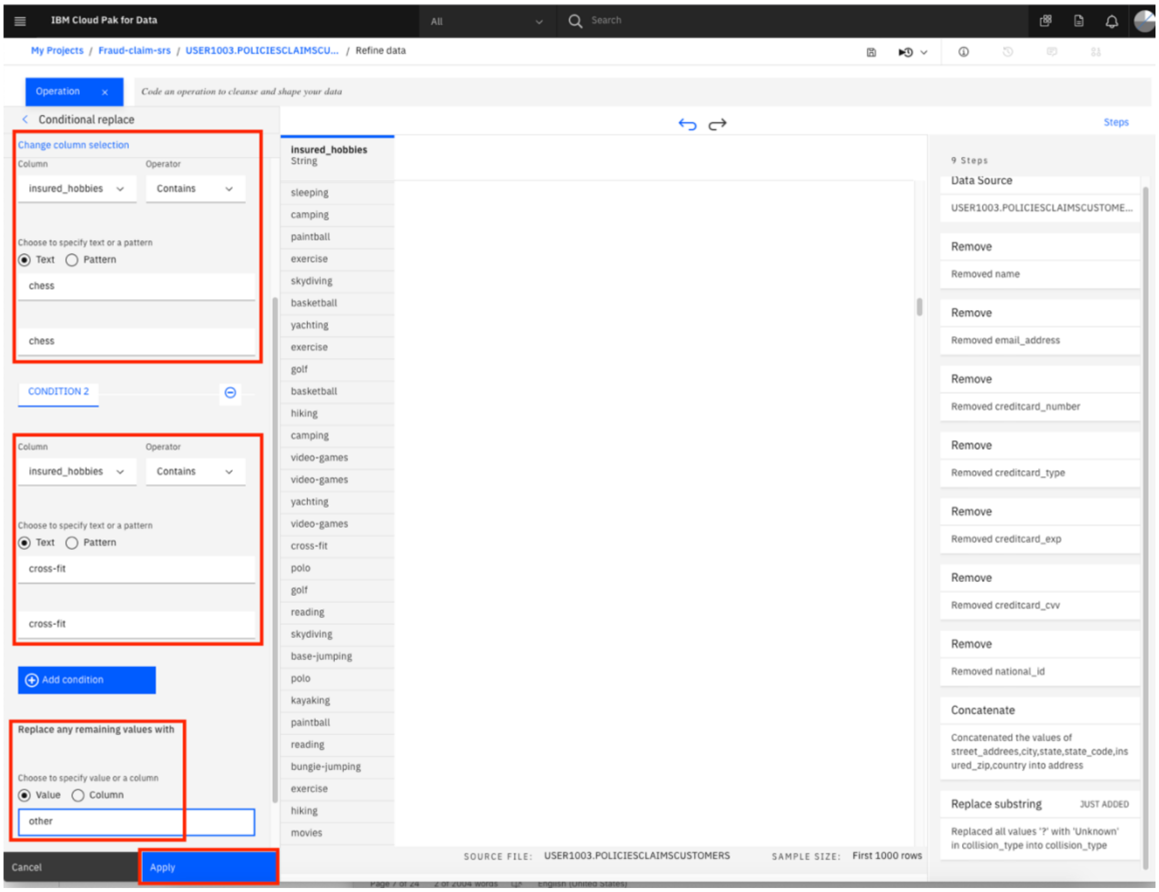
☐ Create new column for results

collision_type String
Side Collision
?
Front Collision
Rear Collision
?
Side Collision
Rear Collision
?
?
Rear Collision
Side Collision
Front Collision
?
Side Collision
?
Rear Collision
Front Collision
Front Collision
Rear Collision

SOURCE FILE: USER1003.POLICIESCLAIMSCUSTOMERS SAMPLE SIZE: First 1000 rows

6. Now we are going to create a new feature based on the **insured_hobbies.** As we observed in the visualization step, the number of fraud claims were remarkably high for the customers with **chess** and **cross-fit** as a hobby. So, we will keep them and rename all other hobbies as **other** in this step.

   Click the **Operation+** button.
7. Select **Conditional replace**.
8. Select **insured_hobbies** in the column selection.
9. Specify condition 1 if the **insured_hobbies** Contains **chess,** then replace it with **chess**.
10. Add another condition. If the **insured_hobbies** Contains **cross-fit,** then replace it with **cross-fit**.
11. Replace any remaining value with **other**.
12. Click **Apply**.



My Projects / Fraud-claim-srs / USER1003.POLICIESCLAIMSCU... / Refine data

Conditional replace

Change column selection

Column: insured_hobbies
Operator: Contains

Choose to specify text or a pattern
● Text ○ Pattern

chess

chess

CONDITION 2

Column: insured_hobbies
Operator: Contains

Choose to specify text or a pattern
● Text ○ Pattern

cross-fit

cross-fit

Replace any remaining values with

Choose to specify value or a column
● Value ○ Column

other

insured_hobbies
String

sleeping
camping
paintball
exercise
skydiving
basketball
yachting
exercise
golf
basketball
hiking
camping
video-games
video-games
yachting
video-games
cross-fit
polo
golf
reading
skydiving
base-jumping
polo
kayaking
paintball
reading
bungie-jumping
exercise
hiking
movies

SOURCE FILE: USER1003.POLICIESCLAIMSCUSTOMERS SAMPLE SIZE: First 1000 rows

9 Steps

Data Source
USER1003.POLICIESCLAIMSCUSTOME...

Remove
Removed name

Remove
Removed email_address

Remove
Removed creditcard_number

Remove
Removed creditcard_type

Remove
Removed creditcard_exp

Remove
Removed creditcard_cvv

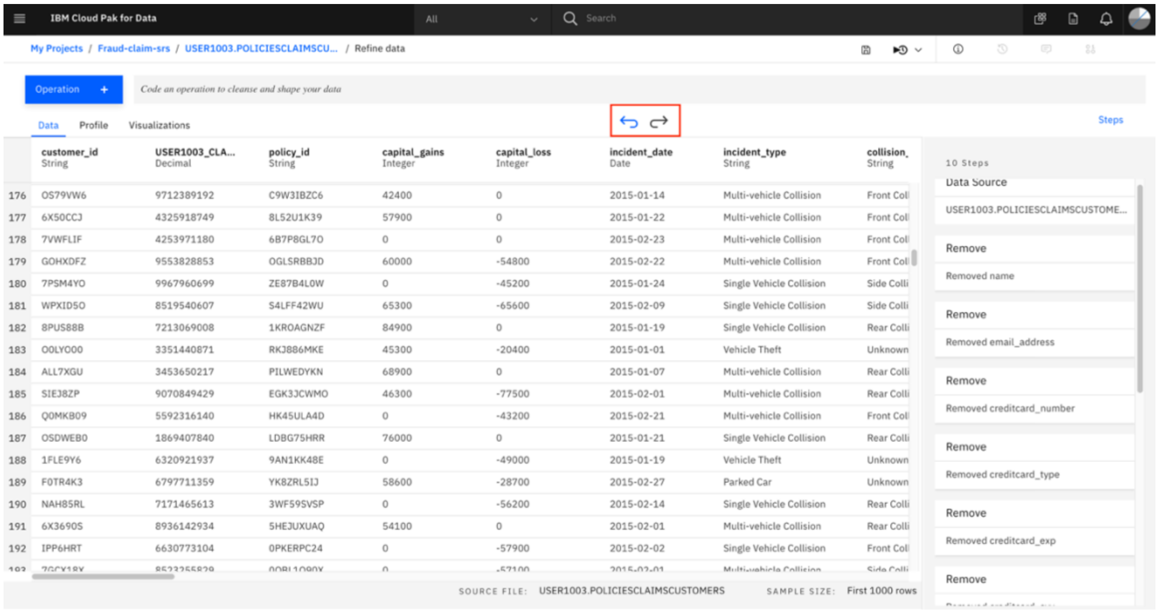
Remove
Removed national_id

Concatenate
Concatenated the values of street_addrees,city,state,state_code,insured_zip,country into address

Replace substring JUST ADDED
Replaced all values '?' with 'Unknown' in collision_type into collision_type

Step 6. Use data flow steps to keep track of your work

What if we accidentally change something? Data refinery keeps track of the steps and we can undo (or redo) an action using the circular arrows.


IBM Cloud Pak for Data
All
Search
My Projects / Fraud-claim-srs / USER1003.POLICIESCLAIMSCU... / Refine data
Operation +
Code an operation to cleanse and shape your data
Data
Profile
Visualizations
Steps

	customer_id String	USER1003_CLA... Decimal	policy_id String	capital_gains Integer	capital_loss Integer	incident_date Date	incident_type String	collision_ String
176	OS79VW6	9712389192	C9W3IBZC6	42400	0	2015-01-14	Multi-vehicle Collision	Front Coll
177	6X50CCJ	4325918749	8L52U1K39	57900	0	2015-01-22	Multi-vehicle Collision	Front Coll
178	7VWFLIF	4253971180	6B7P8GL7O	0	0	2015-02-23	Multi-vehicle Collision	Front Coll
179	GOHXDFZ	9553828853	OGLSRBBJD	60000	-54800	2015-02-22	Multi-vehicle Collision	Front Coll
180	7PSM4YO	9967960699	ZE87B4L0W	0	-45200	2015-01-24	Single Vehicle Collision	Side Colli
181	WPXID5O	8519540607	S4LFF42WU	65300	-65600	2015-02-09	Single Vehicle Collision	Side Colli
182	8PUS88B	7213069008	1KROAGNZF	84900	0	2015-01-19	Single Vehicle Collision	Rear Colli
183	O0LYO00	3351440871	RKJ886MKE	45300	-20400	2015-01-01	Vehicle Theft	Unknown
184	ALL7XGU	3453650217	PILWEDYKN	68900	0	2015-01-07	Multi-vehicle Collision	Rear Colli
185	SIEJ8ZP	9070849429	EGK3JCWMO	46300	-77500	2015-02-01	Multi-vehicle Collision	Rear Colli
186	Q0MKB09	5592316140	HK45ULA4D	0	-43200	2015-02-21	Multi-vehicle Collision	Front Coll
187	OSDWEB0	1869407840	LDBG75HRR	76000	0	2015-01-21	Single Vehicle Collision	Rear Colli
188	1FLE9Y6	6320921937	9AN1KK48E	0	-49000	2015-01-19	Vehicle Theft	Unknown
189	F0TR4K3	6797711359	YK8ZRL5IJ	58600	-28700	2015-02-27	Parked Car	Unknown
190	NAH85RL	7171465613	3WF59SVSP	0	-56200	2015-02-14	Single Vehicle Collision	Rear Colli
191	6X3690S	8936142934	5HEJUXUAQ	54100	0	2015-02-01	Multi-vehicle Collision	Rear Colli
192	IPP6HRT	6630773104	0PKERPC24	0	-57900	2015-02-02	Single Vehicle Collision	Front Coll
193	7GCX18X	8523255829	0OBL1Q90X	0	-57100	2015-02-01	Multi-vehicle Collision	Side Colli

SOURCE FILE: USER1003.POLICIESCLAIMSCUSTOMERS SAMPLE SIZE: First 1000 rows

10 Steps

Data Source

USER1003.POLICIESCLAIMSCUSTOME...

Remove

Removed name

Remove

Removed email_address

Remove

Removed creditcard_number

Remove

Removed creditcard_type

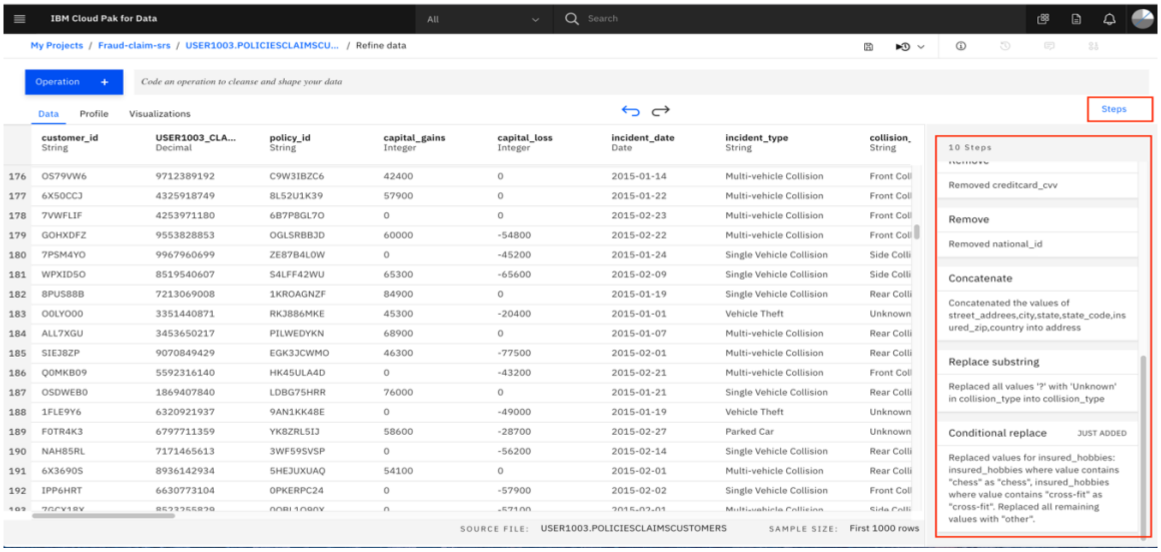
Remove

Removed creditcard_exp

Remove

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
All
Search
My Projects / Fraud-claim-srs / USER1003.POLICIESCLAIMSCU... / Refine data
Operation +
Code an operation to cleanse and shape your data
Data
Profile
Visualizations
Steps

	customer_id String	USER1003_CLA... Decimal	policy_id String	capital_gains Integer	capital_loss Integer	incident_date Date	incident_type String	collision_ String
176	OS79VW6	9712389192	C9W3IBZC6	42400	0	2015-01-14	Multi-vehicle Collision	Front Coll
177	6X50CCJ	4325918749	8L52U1K39	57900	0	2015-01-22	Multi-vehicle Collision	Front Coll
178	7VWFLIF	4253971180	6B7P8GL7O	0	0	2015-02-23	Multi-vehicle Collision	Front Coll
179	GOHXDFZ	9553828853	OGLSRBBJD	60000	-54800	2015-02-22	Multi-vehicle Collision	Front Coll
180	7PSM4YO	9967960699	ZE87B4L0W	0	-45200	2015-01-24	Single Vehicle Collision	Side Colli
181	WPXID5O	8519540607	S4LFF42WU	65300	-65600	2015-02-09	Single Vehicle Collision	Side Colli
182	8PUS88B	7213069008	1KROAGNZF	84900	0	2015-01-19	Single Vehicle Collision	Rear Colli
183	O0LYO00	3351440871	RKJ886MKE	45300	-20400	2015-01-01	Vehicle Theft	Unknown
184	ALL7XGU	3453650217	PILWEDYKN	68900	0	2015-01-07	Multi-vehicle Collision	Rear Colli
185	SIEJ8ZP	9070849429	EGK3JCWMO	46300	-77500	2015-02-01	Multi-vehicle Collision	Rear Colli
186	Q0MKB09	5592316140	HK45ULA4D	0	-43200	2015-02-21	Multi-vehicle Collision	Front Coll
187	OSDWEB0	1869407840	LDBG75HRR	76000	0	2015-01-21	Single Vehicle Collision	Rear Colli
188	1FLE9Y6	6320921937	9AN1KK48E	0	-49000	2015-01-19	Vehicle Theft	Unknown
189	F0TR4K3	6797711359	YK8ZRL5IJ	58600	-28700	2015-02-27	Parked Car	Unknown
190	NAH85RL	7171465613	3WF59SVSP	0	-56200	2015-02-14	Single Vehicle Collision	Rear Colli
191	6X3690S	8936142934	5HEJUXUAQ	54100	0	2015-02-01	Multi-vehicle Collision	Rear Colli
192	IPP6HRT	6630773104	0PKERPC24	0	-57900	2015-02-02	Single Vehicle Collision	Front Coll
193	7GCX18X	8523255829	0OBL1Q90X	0	-57100	2015-02-01	Multi-vehicle Collision	Side Colli

SOURCE FILE: USER1003.POLICIESCLAIMSCUSTOMERS SAMPLE SIZE: First 1000 rows

10 Steps

Removed creditcard_cvv

Remove

Removed national_id

Concatenate

Concatenated the values of street_addrees,city,state,state_code,insured_zip,country into address

Replace substring

Replaced all values '?' with 'Unknown' in collision_type into collision_type

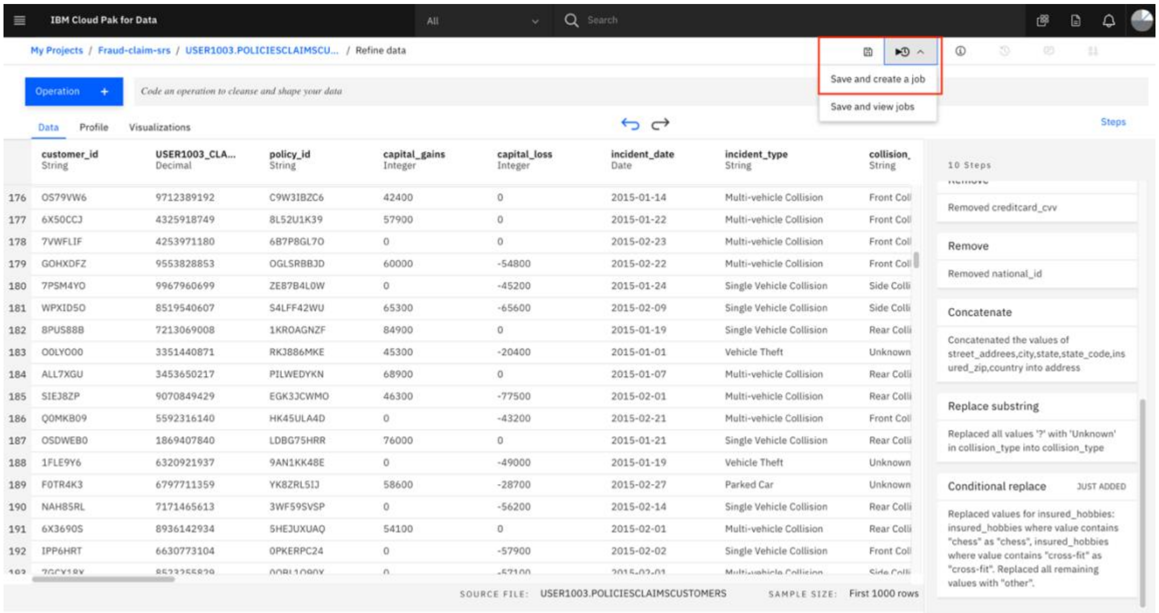
Conditional replace JUST ADDED

Replaced values for insured_hobbies: insured_hobbies where value contains "chess" as "chess", insured_hobbies where value contains "cross-fit" as "cross-fit". Replaced all remaining values with "other".

Step 7. Save data flow and create a job

Once you have refined your data, you would want to save and 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.



IBM Cloud Pak for Data
All
Search
My Projects / Fraud-claim-srs / USER1003.POLICIESCLAIMSCU... / Refine data
Save and create a job
Save and view jobs
Operation +
Code an operation to cleanse and shape your data
Data
Profile
Visualizations
Steps

	customer_id String	USER1003_CLA... Decimal	policy_id String	capital_gains Integer	capital_loss Integer	incident_date Date	incident_type String	collision_ String
176	OS79VW6	9712389192	C9W3IBZC6	42400	0	2015-01-14	Multi-vehicle Collision	Front Coll
177	6X50CCJ	4325918749	8L52U1K39	57900	0	2015-01-22	Multi-vehicle Collision	Front Coll
178	7VWFLIF	4253971180	6B7P8GL7O	0	0	2015-02-23	Multi-vehicle Collision	Front Coll
179	GOHXDFZ	9553828853	OGLSRBBJD	60000	-54800	2015-02-22	Multi-vehicle Collision	Front Coll
180	7PSM4YO	9967960699	ZE87B4L0W	0	-45200	2015-01-24	Single Vehicle Collision	Side Colli
181	WPXID5O	8519540607	S4LFF42WU	65300	-65600	2015-02-09	Single Vehicle Collision	Side Colli
182	8PUS88B	7213069008	1KROAGNZF	84900	0	2015-01-19	Single Vehicle Collision	Rear Colli
183	O0LYO00	3351440871	RKJ886MKE	45300	-20400	2015-01-01	Vehicle Theft	Unknown
184	ALL7XGU	3453650217	PILWEDYKN	68900	0	2015-01-07	Multi-vehicle Collision	Rear Colli
185	SIEJ8ZP	9070849429	EGK3JCWMO	46300	-77500	2015-02-01	Multi-vehicle Collision	Rear Colli
186	Q0MKB09	5592316140	HK45ULA4D	0	-43200	2015-02-21	Multi-vehicle Collision	Front Coll
187	OSDWEB0	1869407840	LDBG75HRR	76000	0	2015-01-21	Single Vehicle Collision	Rear Colli
188	1FLE9Y6	6320921937	9AN1KK48E	0	-49000	2015-01-19	Vehicle Theft	Unknown
189	F0TR4K3	6797711359	YK8ZRL5IJ	58600	-28700	2015-02-27	Parked Car	Unknown
190	NAH85RL	7171465613	3WF59SVSP	0	-56200	2015-02-14	Single Vehicle Collision	Rear Colli
191	6X3690S	8936142934	5HEJUXUAQ	54100	0	2015-02-01	Multi-vehicle Collision	Rear Colli
192	IPP6HRT	6630773104	0PKERPC24	0	-57900	2015-02-02	Single Vehicle Collision	Front Coll

SOURCE FILE: USER1003.POLICIESCLAIMSCUSTOMERS SAMPLE SIZE: First 1000 rows

10 Steps

Removed creditcard_cvv

Remove

Removed national_id

Concatenate

Concatenated the values of street_addrees,city,state,state_code,insured_zip,country into address

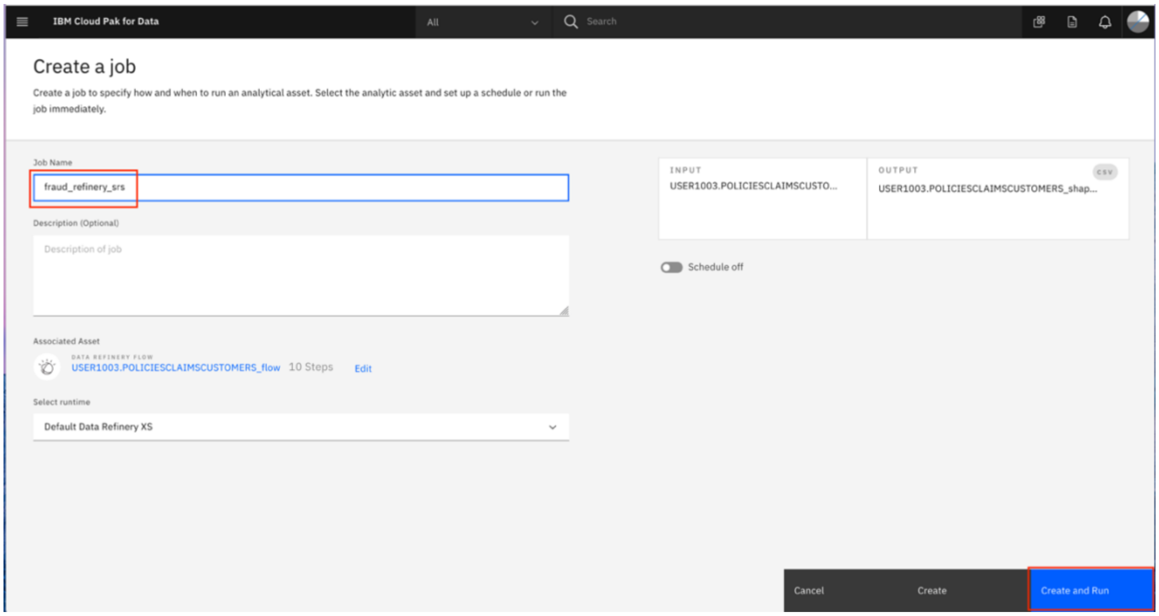
Replace substring

Replaced all values '?' with 'Unknown' in collision_type into collision_type

Conditional replace JUST ADDED

Replaced values for insured_hobbies: insured_hobbies where value contains "chess" as "chess", insured_hobbies where value contains "cross-fit" as "cross-fit". Replaced all remaining values with "other".

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. 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: fraud_refinery_srs

Description (Optional): Description of job

INPUT: USER1003.POLICIESCLAIMSCUSTO...

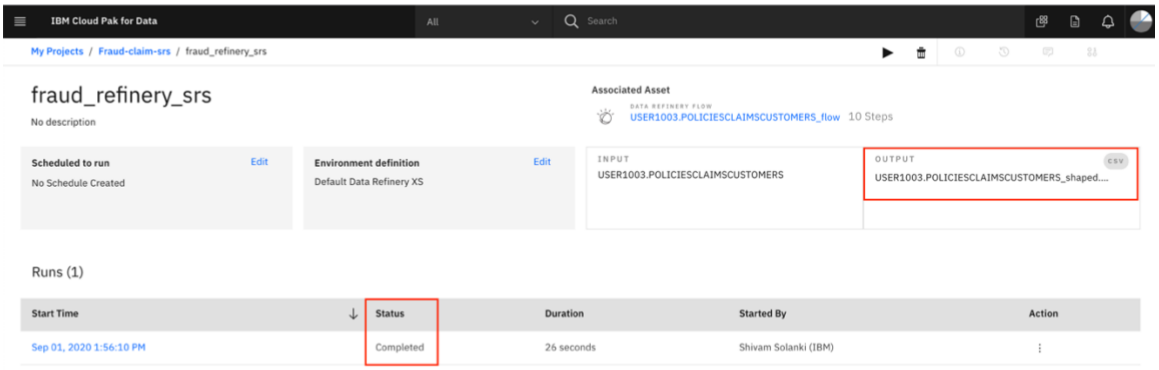
OUTPUT: USER1003.POLICIESCLAIMSCUSTOMERS_shap... CSV

Schedule off

Associated Asset: DATA REFINERY FLOW USER1003.POLICIESCLAIMSCUSTOMERS_flow 10 Steps Edit

Select runtime: Default Data Refinery XS

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



My Projects / Fraud-claim-srs / fraud_refinery_srs

fraud_refinery_srs

No description

Associated Asset

DATA REFINERY FLOW

USER1003.POLICIESCLAIMSCUSTOMERS_flow 10 Steps

Scheduled to run — Edit

No Schedule Created

Environment definition — Edit

Default Data Refinery XS

INPUT

USER1003.POLICIESCLAIMSCUSTOMERS

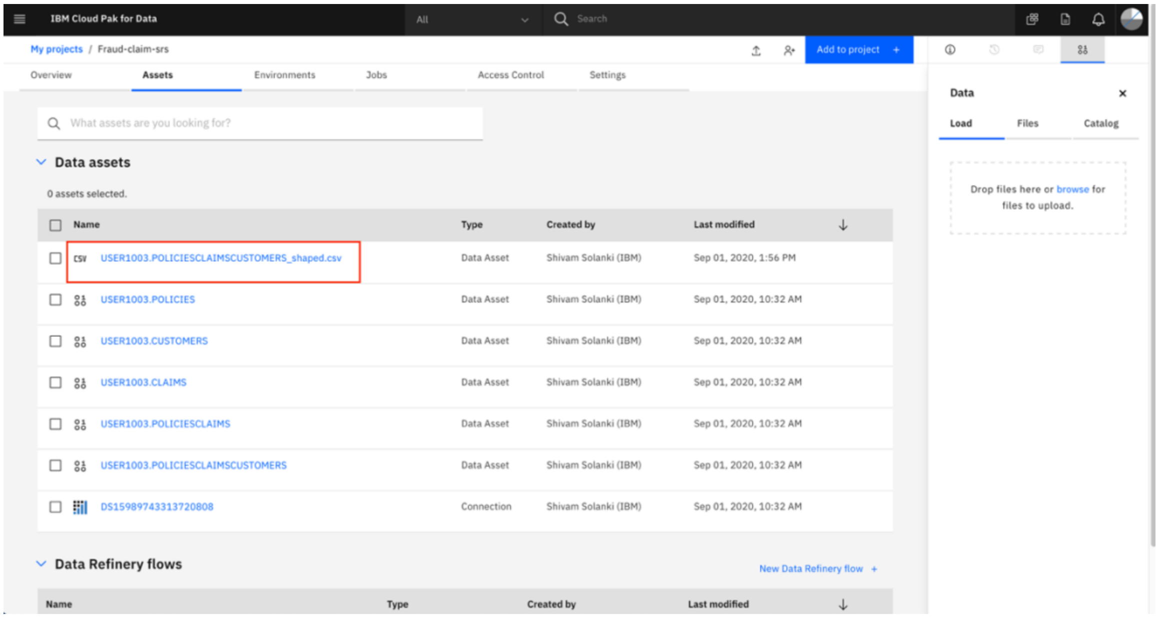
OUTPUT — CSV

USER1003.POLICIESCLAIMSCUSTOMERS_shaped....

Runs (1)

Start Time ↓	Status	Duration	Started By	Action
Sep 01, 2020 1:56:10 PM	Completed	26 seconds	Shivam Solanki (IBM)	⋮

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.



My projects / Fraud-claim-srs

Data assets

0 assets selected.

□	Name	Type	Created by	Last modified ↓
□	CSV USER1003.POLICIESCLAIMSCUSTOMERS_shaped.csv	Data Asset	Shivam Solanki (IBM)	Sep 01, 2020, 1:56 PM
□	USER1003.POLICIES	Data Asset	Shivam Solanki (IBM)	Sep 01, 2020, 10:32 AM
□	USER1003.CUSTOMERS	Data Asset	Shivam Solanki (IBM)	Sep 01, 2020, 10:32 AM
□	USER1003.CLAIMS	Data Asset	Shivam Solanki (IBM)	Sep 01, 2020, 10:32 AM
□	USER1003.POLICIESCLAIMS	Data Asset	Shivam Solanki (IBM)	Sep 01, 2020, 10:32 AM
□	USER1003.POLICIESCLAIMSCUSTOMERS	Data Asset	Shivam Solanki (IBM)	Sep 01, 2020, 10:32 AM
□	DS15989743313720808	Connection	Shivam Solanki (IBM)	Sep 01, 2020, 10:32 AM

Data Refinery flows

Name	Type	Created by	Last modified ↓

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 lab tutorial showed you a small sampling of the power of the IBM data refinery on IBM Cloud Pak for Data. You learned how to refine and visualize the virtualized data which was not even moved or copied from the Db2 Warehouse on IBM Cloud. The tutorial also explained how you can transform data using various operations on the columns such as removing columns, concatenating columns, replacing missing values with a string or creating a new value. The lab 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. The lab tutorial showed how you can quickly profile the data to see histograms and statistics for each column. And finally, it explained how you can create more in-depth visualizations and create a bar chart of various features against the fraud column to gain insights about the importance of columns in detecting fraud.