

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.

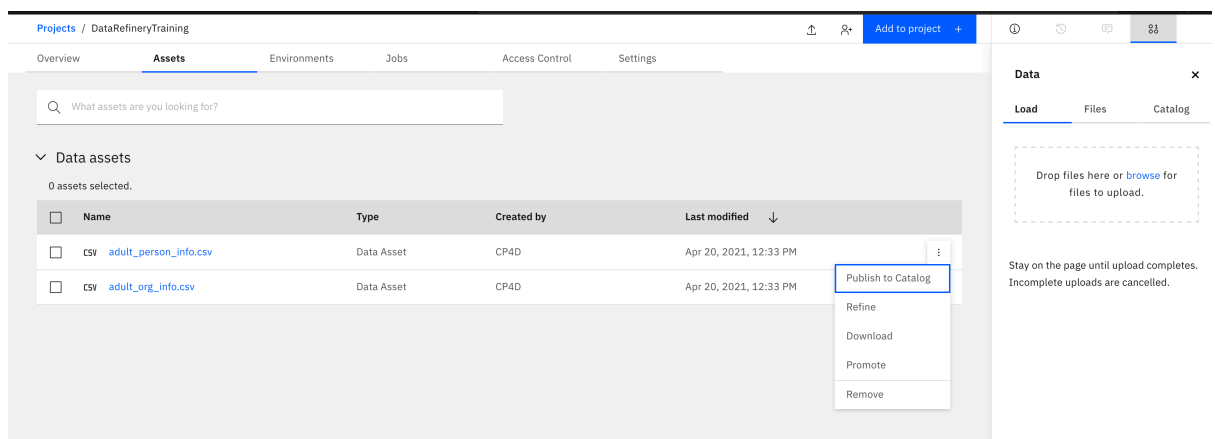
Estimated time

Completing this tutorial should take about 45 minutes.

Steps

Step 1. Add Datasets to the Project

- Go to the newly created **analytics** project and add the datasets to the project:
 - Click on **Assets** on the panel
 - At the top right of the page, click on the add data icon.
 - Click on **Load** and drag and drop the two files *adult_person_info.csv* and *adult_org_info.csv*.
 - You will notice that once the files are uploaded, they will be added under Data assets.



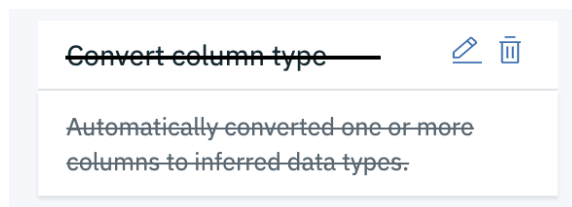
- Review Data Refinery UI:
 - Go to the ellipsis icon next to *adult_person_info.csv* under *Data assets* and select **Refine**. This will open a page that shows a sample of the content, where you can start cleaning and reshaping the data set.
 - On the panel on the right, you will find **Details** including the project the data asset belongs to, and description of the resulting data set we will get after the refining process. Close it for the time being.

- Click on **Steps**, which you can find right hand-side of the page. This is where you will see each operation you will define while transforming the data. It shows the data flow defining the operations to be done on the entire data set.

The screenshot shows the DataRefinery interface. The 'Data' tab displays a table with 18 rows of data. The columns are: UNIQUE_ID (Integer), AGE (String), EDUCATION (String), EDUCATION-NUM (Integer), MARITAL_STATUS (String), and OCCUPATION (String). The 'Steps' tab shows a single step named 'Convert column type' with a description: 'Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol.' The right sidebar shows 'Information' and 'Details' sections, including the 'Data Refinery Flow Name' and 'Data Refinery Flow Output'.

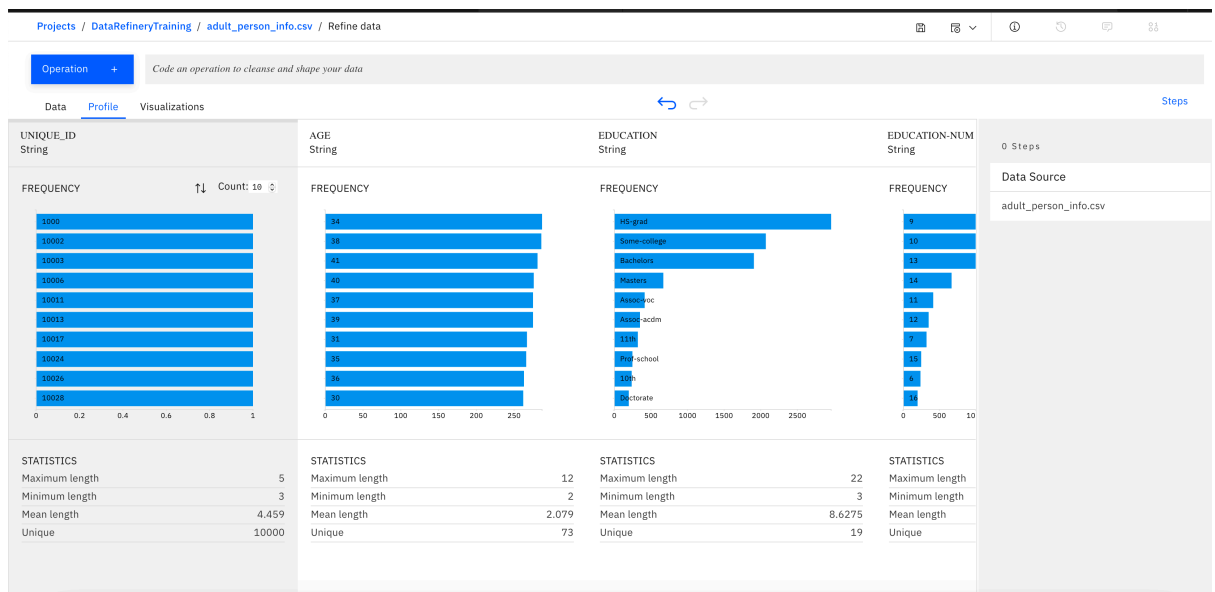
	UNIQUE_ID Integer	AGE String	EDUCATION String	EDUCATION-NUM Integer	MARITAL_STATUS String	OCCUPATION String
1	1021	25	Some - college	10	Never-married	Transport-mov
2	14249	37	HS-grad	9	Married-civ-spouse	Craft-repair
3	11259	37 YEARS OLD	Some - college	10	Married-civ-spouse	
4	4180	19	Some - college	10	Never-married	
5	6692	66	HS-grad	9	Widowed	Sales
6	11276	31	HS-grad	9	Married-civ-spouse	Craft-repair
7	5049	44	Bachelors	13	Married-civ-spouse	Sales
8	7116	41	Some - college	10	Never-married	Craft-repair
9	15540	59	10th	6	Divorced	Machine-op-in
10	15393	25	11th	7	Married-spouse-absent	Handlers-clea
11	12513	27	Bachelors	13	Never-married	Other-service
12	3341	41	Some - college	10	Separated	Other-service
13	4073	21	HS-grad	9	Never-married	Adm-clerical
14	3417	45	Some - college	10	Married-civ-spouse	Tech-support
15	9437	27	Some - college	10	Never-married	Protective-ser
16	11579	37	HS-grad	9	Married-civ-spouse	Farming-fishin
17	6923	35	Bachelors	13	Married-civ-spouse	Sales
18	13746	62	HS-grad	9	Married-civ-spouse	

- You may notice that an auto data type conversion step has been applied, in order to practice how to manually convert column type, we will remove this step for now by clicking on the trash bin icon.



Step 2. Review Data Profile

- Skim through data displayed in the **Data** tab and then click on the **Profile** tab and take a quick look at data summary and get a feel of the data. You will notice some weird formats under **FREQUENCY** for some fields. For example, you will notice that:
 - Some values under **AGE** contain additional string such as “years old”,
 - For **Education**, there are some additional values with extra spaces at the beginning and possibly the end of the string.
 - Empty cells in the **OCCUPATION** column,
 - There are multiple values under **GENDER** that seem to be meant to represent the value Male, etc.



Step 3.Data Harmonization: AGE

- Standardize the AGE field:
 - As mentioned earlier, you will notice some values with additional string such as years old. What we want is to just retain the numerical part, which can only be a two-digit number in our case (we know there are no additional characters that were added before the numerical part of the values or that the digits contain no weird characters).

AGE String
25
37
37 YEARS OLD
19

- Click on **+Operation** and select **Split column**, which you can find under **ORGANIZE**.
- Choose AGE as the **Selected column**.
- Under **POSITION** tab, type 2 in the **Positions** field and enter names in the **Names of new columns**. Make sure to unselect **Keep original column**
- Click **Apply**.

➔Keep in mind that this is not the best approach to handle this. This is just provide an example of how to use the **split column** operation.

× Operation Code an operation to cleanse and shape your data

< Split column

Change Column Selection

Selected column: AGE

Split the column by non-alphanumeric characters, position, pattern, or text.

DEFAULT TEXT PATTERN POSITION

Specify positions from the left as positive values (starting at 1) and positions from the right as negative values (starting at -1).

Positions*

2

Names of new columns*

AGE_num, AGE_str

☐ Keep original column ⓘ

Cancel Apply

AGE
String
25
37
37 YEARS OLD
19
66
31
44
41
59
25
27
41
21
45
27
37
35
62
57
50
20
30

- Go to the **Data** tab and remove the newly created column called *AGE_str*, which only contain the string part of the age.

AGE_str	EDUCATION
String	String
YEARS OLD	

Remove
Remove duplicates
Remove empty rows
Sort ascending

- Go to column called *AGE_num* and rename it to *AGE* by clicking on the pencil icon.
- Go to the **Profile** tab again to for a final check.

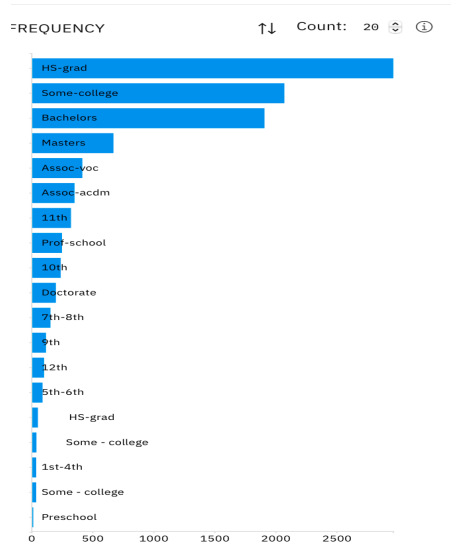
Step 4. Convert Data Type

- Change **AGE** data type to **Integer** using CONVERT COLUMN...> Integer. Data Refinery will put a dot in front of the recommended data type.

AGE	EDUCATION	EDUCATION-NUM	MA
String	String	String	String
25		10	Nev
37		9	Mai
37		10	Mai
19		10	Nev
66		9	Wic
31		9	Mai
44		13	Mai
41			Nev
59			Div
25			Mai
27			Nev
41	Some - college		Se
21	HS-grad		Nev
45	Some - college		Mai
27	Some - college		Nev
37	HS-grad		Mai

Step 5.Data Harmonization: EDUCATION

- Standardize the *EDUCATION* field:
 - Click on the **Profile** tab and take a closer look at the column *EDUCATION*. You notice there are some additional values with extra spaces at the beginning and possibly the end of the string.



- Click on **+Operation** and select **Text**, which you can find under **FREQUENTLY USED**.
- Choose *EDUCATION* as the **Selected column**, **Collapse spaces** as the **Text Operation**.
- Click **Apply** and go to the **Profile** tab again to check if all the additional values have been removed. You will notice the we still have *Some - college* as an additional value, which we want to harmonize and change to *Some-college*.

- Click on **+Operation** and select **Replace substring**, which you can find under **CLEANSE**.
- Choose **EDUCATION** as the **Selected column**.
- Under **TEXT** tab, type Some - college in **Value** field and Some-college in the *Enter the replacement string*. Make sure to select **Replace all occurrences**
- Click **Apply**.

Operation x Code an operation to cleanse and shape your data

< Replace substring

Change column selection

Selected column: EDUCATION

Replace the specified substring with the specified text.

Choose to specify text or a pattern

☒ Text ☐ Pattern

Some - college

Some-college

☒ Replace all occurrences ⓘ

☐ Create new column for results ⓘ

EDUCATION String

FREQUENCY ↑↓ Count: 20 ⓘ

Education Level	Frequency
HS-grad	20
Some-college	18
Bachelors	15
Masters	10
Assoc-voc	8
Assoc-acdm	7
11th	6
Prof-school	5
10th	4
Doctorate	3
7th-8th	2
9th	2
12th	2
5th-6th	1
Some - college	1
1st-4th	1
Preschool	1

- We also want to convert all values in the EDUCATION column to lower case. So, click on **+Operation** and select **Text**, which you can find under **FREQUENTLY USED**.
- Choose EDUCATION as the **Selected column**, **Lower case** as the **Text Operation**.
- Click **Apply** and go to the **Profile** tab again to for a final check.

Operation x Code an operation to cleanse and shape your data

< Text

Change column selection

Selected column: EDUCATION

Text operation

Convert the text to lower case.

Lower case

☐ Create new column for results ⓘ

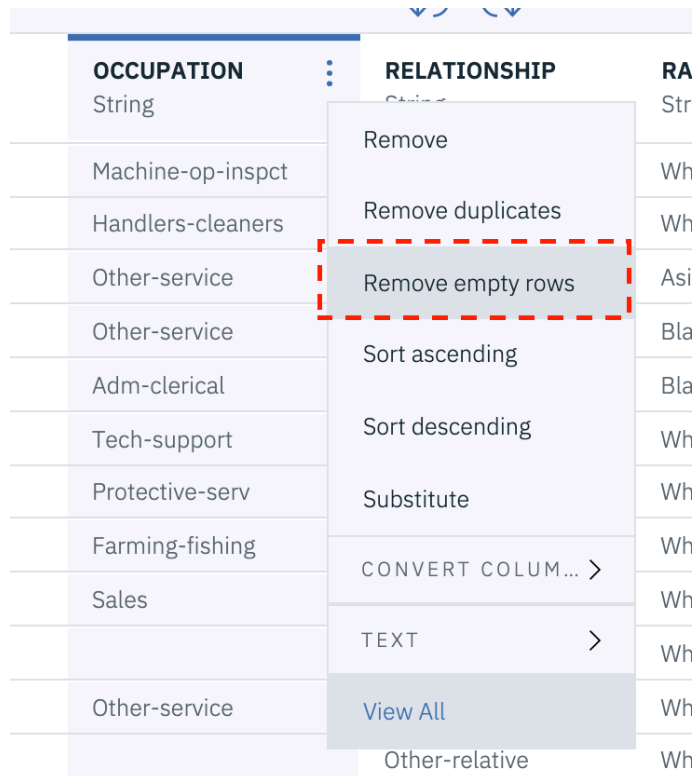
EDUCATION String

FREQUENCY ↑↓ Count: 20 ⓘ

Education Level	Frequency
Some-college	20
HS-grad	18
Bachelors	15
Masters	10
Assoc-voc	8
Assoc-acdm	7
11th	6
Prof-school	5
10th	4
Doctorate	3
7th-8th	2
9th	2
12th	2
5th-6th	1
Some - college	1
1st-4th	1
Preschool	1

Step 6.Remove Missing Values in OCCUPATION

- Removing empty rows (List-wise deletion):
 - Go to the **Data** tab.
 - Go to the column called OCCUPATION and remove rows with any empty values by clicking menu next to the column name and selecting **Remove empty rows**.



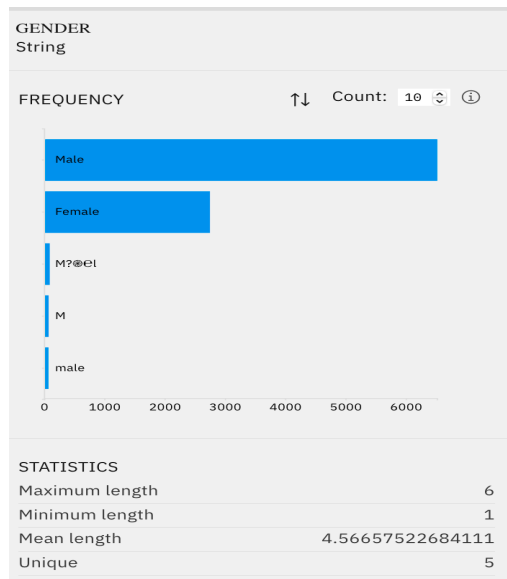
The screenshot shows a data table with columns OCCUPATION, RELATIONSHIP, and RA. The OCCUPATION column has a dropdown menu open, showing options like 'Remove', 'Remove duplicates', 'Remove empty rows', 'Sort ascending', 'Sort descending', 'Substitute', 'CONVERT COLUMN...', 'TEXT', and 'View All'. The 'Remove empty rows' option is highlighted with a red dashed box. The table data is as follows:

OCCUPATION	RELATIONSHIP	RA
String	String	Str
Machine-op-inspct		Wh
Handlers-cleaners		Wh
Other-service		Asi
Other-service		Bla
Adm-clerical		Bla
Tech-support		Wh
Protective-serv		Wh
Farming-fishing		Wh
Sales		Wh
		Wh
Other-service		Wh
	Other-relative	Wh

- Go to the **Profile** tab to check if all empty values have been remove for OCCUPATION.

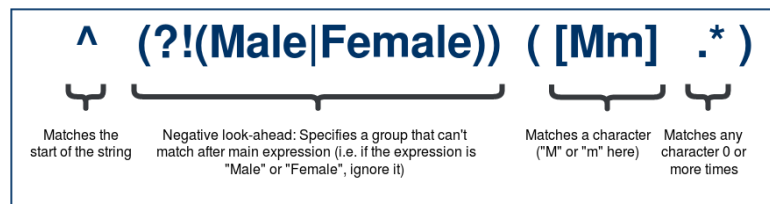
Step 7.Data Harmonization: GENDER

- Standardize the GENDER field:
 - Click on the **Profile** tab and take a closer look at the column *GENDER*. You will notice some additional values other than Male and Female, mainly ones that we want to change to Male.



- Click on **+Operation** and select **Replace substring**, which you can find under **CLEANSE**.
- Choose **GENDER** as the *Selected column*.
- Under **PATTERN** tab, type `^(?!((Male|Female)))([Mm].*)` in the **Regular expression** field and **Male** under **Enter the replacement string**. Make sure to select **Replace all occurrences**.

What is meant by `^(?!((Male|Female)))([Mm].*)` is to find any expression that doesn't start with Male or Female and starts with the letter M or m, which could be followed by any character.



- Click **Apply** and go to the **Profile** tab again to for a final check.

Operation × Code an operation to cleanse and shape your data

< **Replace substring** EDIT MODE

Change column selection

Selected column: GENDER

Replace the specified substring with the specified text.

Choose to specify text or a pattern

☐ Text ☒ Pattern

Regular expression: `^(?!((Male|Female)))([Mm].*)`

Replacement string: Male

☒ Replace all occurrences ⓘ

☐ Create new column for results ⓘ

GENDER String

FREQUENCY ↑↓ Count: 10 ⓘ

Category	Frequency
Male	~5800
Female	~2800
M?@e!	~100
M	~100
male	~100

STATISTICS

Maximum length	6
----------------	---

Step 8.Remove Duplicates

- Remove duplicate values based on the **UNIQUE_ID**:
 - Go to the **Data** tab.
 - Go to the column called **UNIQUE_ID** and remove rows with any duplicate **UNIQUE_ID** values by menu next to the column name and selecting **Remove duplicates**.

+ Operation Code an operation to cleanse and shape your data				
Data Profile Visualizations				
	UNIQUE_ID String	AGE String	EDUCATION String	MARITAL_STA... String
		Remove		
1	1021	Remove duplicates	some-college	Never-married
2	14249	Remove empty rows	hs-grad	Married-civ-spouse
3	11259	Sort ascending	some-college	Married-civ-spouse
4	4180	Sort descending	some-college	Never-married
5	6692	Substitute	hs-grad	Widowed
6	11276	CONVERT COL... >	hs-grad	Married-civ-spouse
7	5049	TEXT >	bachelors	Married-civ-spouse
8	7116	View All	some-college	Never-married
9	15540		10th	Divorced
10	15393		11th	Married-spouse-ab...
11	12513		bachelors	Never-married
12	3341	41	some-college	Separated
13	4073	21	hs-grad	Never-married
14	3417	45	some-college	Married-civ-spouse
15	9437	27	some-college	Never-married
16	11579	37	hs-grad	Married-civ-spouse
17	6923	35	bachelors	Married-civ-spouse
18	13746	62	hs-grad	Married-civ-spouse
19	14244	57	assoc-acdm	Married-civ-spouse
20	906	50	masters	Married-spouse-ab...


Step 9.Join Datasets

- Now we will join two datasets:
 - Click on the **Data** tab to see a sample of your data.
 - Click on **+Operation** and then select **Join**, which you can find under **ORGANIZE**. This is to join both the data assets we added namely the one we are currently refining, *adult_person_info.csv*, and *adult_org_info.csv*.

Operation ×		Code an operation to cleanse and shape your data		
Q Search operations				
Text		UNIQUE_ID String	AGE Integer	EDUCATION String
CLEANSE	^			EDUCATION-NUM String
Convert column value to missing		1021	25	some-college
Extract date or time value		14249	37	hs-grad
Remove duplicates		6692	66	hs-grad
Remove empty rows		11276	31	hs-grad
Replace missing values		5049	44	bachelors
Replace substring		7116	41	some-college
ORGANIZE	^	12513	27	bachelors
Aggregate		3341	41	some-college
Concatenate		4073	21	hs-grad
Conditional replace		3417	45	some-college
Join	>	9437	27	some-college
Sample		11579	37	hs-grad
Split column		6923	35	bachelors
Union		14244	57	assoc-acdm
NATURAL LANGUAGE	^	7860	20	hs-grad
		15399	30	hs-grad
		4273	24	bachelors

PS: Make sure UNIQUE_ID for both datasets are either String or Integer format to sync

- Select **Inner join** as the method of how we want our data to be combined (Inner Join selects records that have matching column value(s) in both tables). By default, the **Source** is selected as the current data asset (*adult_person_info.csv*).
- Choose *adult_org_info.csv* as the **Data set to join**. Click on the little eye icon to preview data, you will find it has a column **UNIQUE_ID** which is our join key. Click **Apply**.

Data set to join with adult_person_info.csv		
DataRefineryTraining	Data assets	
Assets (2)	Data assets (2)	
Connections	>	adult_org_info.csv 
Data assets	>	adult_person_info.csv

- Use the default values for the **Suffix** field, which is just a way for you to differentiate any duplicate fields resulted during the joining process. You can also modify it if you want.
- For the **JOIN KEYS**, select UNIQUE_ID representing the employee ID, as the join key for both data sets and click **Next**

Operation ✕ Code an operation to cleanse and shape your data

< Join

Combine data from two data sets based on a comparison of the values in specified key columns.

Inner join

Returns only the rows in each data set that match rows in the other data set. Returns one row in the original data set for each matching row in the joining data set.

The default suffix for each data set will be used to differentiate any duplicate column names in the resulting data set.

Source Data set to join

adult_person_info.csv adult_org_info.csv

*Suffix *Suffix

_X _Y

JOIN KEYS

adult_person_info.csv adult_org_info.csv

UNIQUE_ID UNIQUE_ID

+ Add Join Key

UNIQUE_ID	AGE	EDUCATION	EDUCATION-NUM	MARITAL_STATUS	OCCUPATION
1021	25	some-college	10	Never-married	Transport-moving
14249	37	hs-grad	9	Married-civ-spouse	Craft-repair
6692	66	hs-grad	9	Widowed	Sales
11276	31	hs-grad	9	Married-civ-spouse	Craft-repair
5049	44	bachelors	13	Married-civ-spouse	Sales
7116	41	some-college	10	Never-married	Craft-repair
15540	59	10th	6	Divorced	Machine-op-inspct
15393	25	11th	7	Married-spouse-absent	Handlers-cleaners
12513	27	bachelors	13	Never-married	Other-service
3341	41	some-college	10	Separated	Other-service
4073	21	hs-grad	9	Never-married	Adm-clerical
3417	45	some-college	10	Married-civ-spouse	Tech-support
9437	27	some-college	10	Never-married	Protective-serv
11579	37	hs-grad	9	Married-civ-spouse	Farming-fishing
6923	35	bachelors	13	Married-civ-spouse	Sales
14244	57	assoc-acdm	12	Married-civ-spouse	Other-service
7860	20	hs-grad	9	Never-married	Other-service
15399	30	hs-grad	9	Married-civ-spouse	Adm-clerical
4273	24	bachelors	13	Never-married	Farming-fishing
5818	39	some-college	10	Married-civ-spouse	Farming-fishing

- Keep all columns and select **Apply**.

✕ Operation Code an operation to cleanse and shape your data

< Join

Select the columns in the resulting data set

☒ Clear all selections

☒ UNIQUE_ID

☒ AGE

☒ EDUCATION

☒ EDUCATION-NUM

☒ MARITAL_STATUS

☒ OCCUPATION_x

☒ RELATIONSHIP

☒ RACE

☒ GENDER

☒ NATIVE_COUNTRY

☒ EMPLOYER_TYPE

☒ OCCUPATION_y

☒ CAPITAL_GAIN

☒ CAPITAL_LOSS

☒ HOURS_PER_WEEK

☒ INCOME

- We will notice that there are 2 columns representing **OCCUPATION**, one coming from each of the data sets. Let's check to see if they contain the exact same values.
 - Click on **+Operation** and select **Calculate**, which you can find under FREQUENTLY USED.
 - Choose OCCUPATION_x as the Selected column, Is equal to as the Operation and OCCUPATION_y as the COLUMN.

- Select to Create new column for the results and enter **"OCCUPATION_CHECK"** as the New column name.
- Click **Apply**. You will see the resulting column added at the right end of the table.

×

Operation

Code an operation to cleanse and shape your data

<

Calculate

Change column selection

Selected column: OCCUPATION_x

Perform a calculation with another column or with a specified value.

Is equal to

Choose to specify value or a column

Value

Column

OCCUPATION_y

✓

Create new column for results

i

OCCUPATION_CHECK

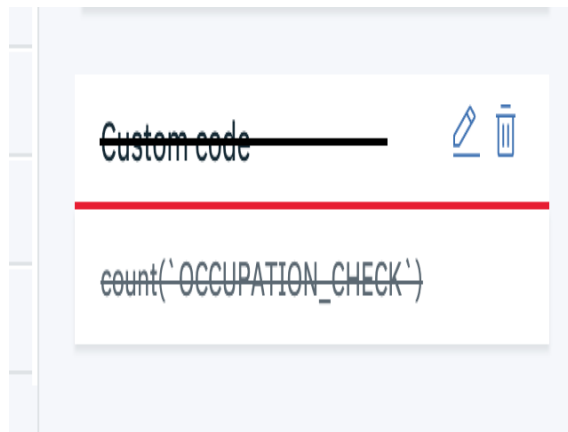
- Count OCCUPATION_CHECK:
 - In the space next to the **+Operation**, place the cursor and select **count**.
 - Click on the count that was added to the box and select count().
 - Click on and choose the newly created column (we called it OCCUPATION_CHECK).
 - Click **Apply**.
- The result shows that OCCUPATION_x and OCCUPATION_y have identical values.

<div> <div>Operation</div> <div>+</div> <div>Code an operation to cleanse and shape your data</div> </div>		
<div> <div>Data</div> <div>Profile</div> <div>Visualizations</div> </div>		
<div> <div>OCCUPATION_C...</div> <div>Boolean</div> <div>n</div> <div>Integer</div> </div>		
1	true	9478

Step 10.Undo Steps

- So we can just keep one of them and keep one OCCUPATION column:
 - Go back **2 steps** by either clicking on the Undo button found at the top middle of the page or by going to the step added under **Steps** and clicking on the bin icon. Whichever way you select, you will need to do it **twice**.

Delete steps



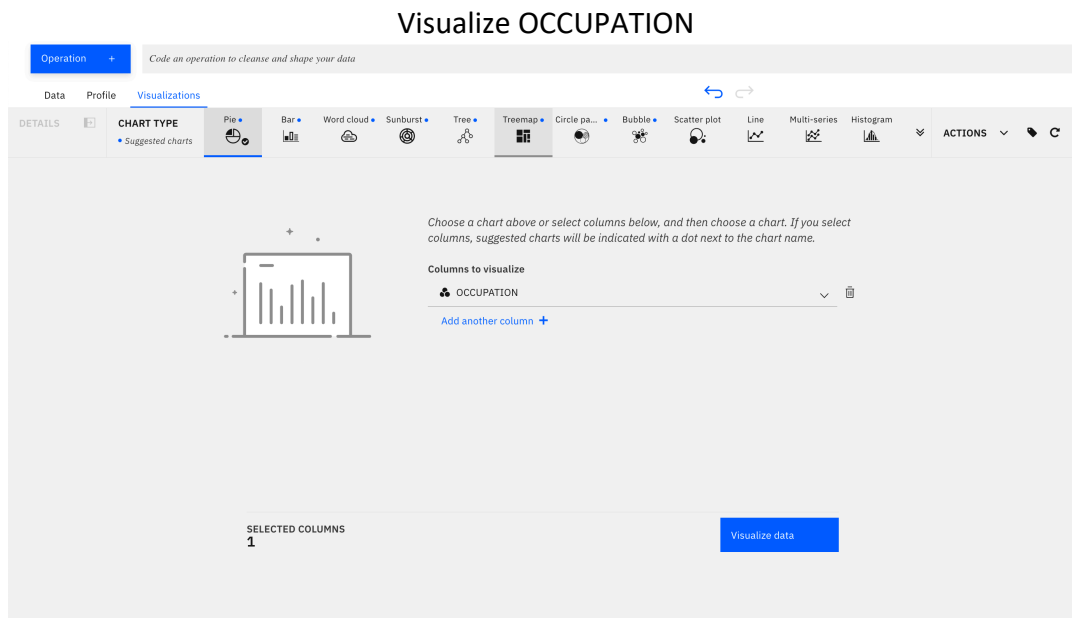
- Go to column called *OCCUPATION_x* and rename it to *OCCUPATION*.
- Go to column called *OCCUPATION_y* and remove

+ Operation				Code an operation to cleanse and shape your data			
Data				Profile			
				Visualizations			
	OCCUPATION.y		CAPITAL_GAIN		CAPITAL_LOSS		
	String		String		String		
1	Transport-moving				0		
2	Craft-repair				0		
3					0		
4					0		
5	Sales				0		
6	Craft-repair				0		
7	Sales				0		
8	Craft-repair				0		
9	Machine-op-inspct				0		
10	Handlers-cleaners				0		
11	Other-service				0		
12	Other-service		0		0		
13	Adm-clerical		0		0		
14	Tech-support		7687.5		0		
15	Protective-serv		0		0		
16	Farming-fishing		0		0		
17	Sales		0		0		
18			0		0		

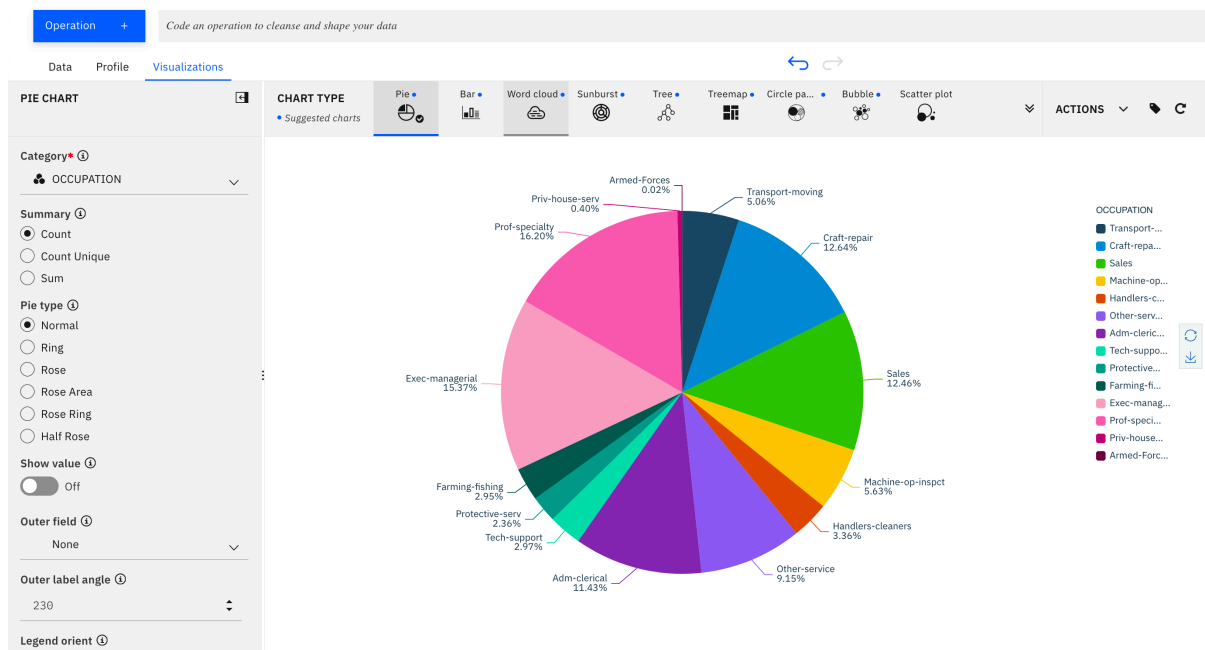
Step 11.Data Visualization

Data Refinery does not only include powerful shaping operations to clean, organize, fix, and validate data but also has built-in visualization capability to derive insights from data.

- Click on the **Visualization Tab**, choose **OCCUPATION** from the dropdown menu, then click **Visualize Data**.

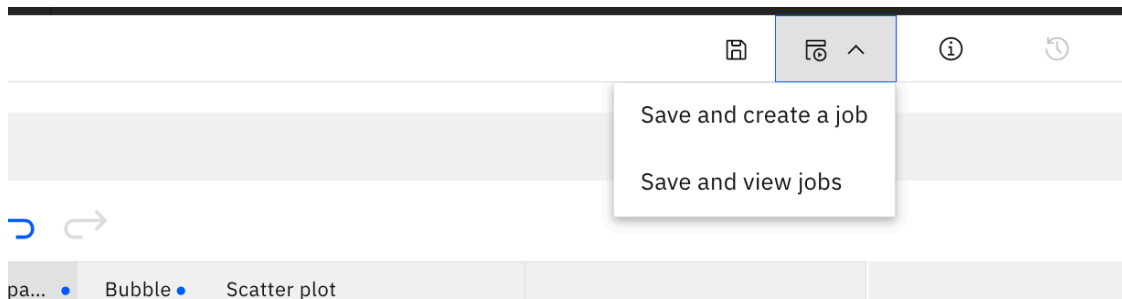


- Data Refinery will suggest the best techniques to visualize the data. For OCCUPATION, it has suggested the pie chart showing you the distribution of OCCUPATION. Click on other **CHART TYPES** to see other visualization outputs if interested.

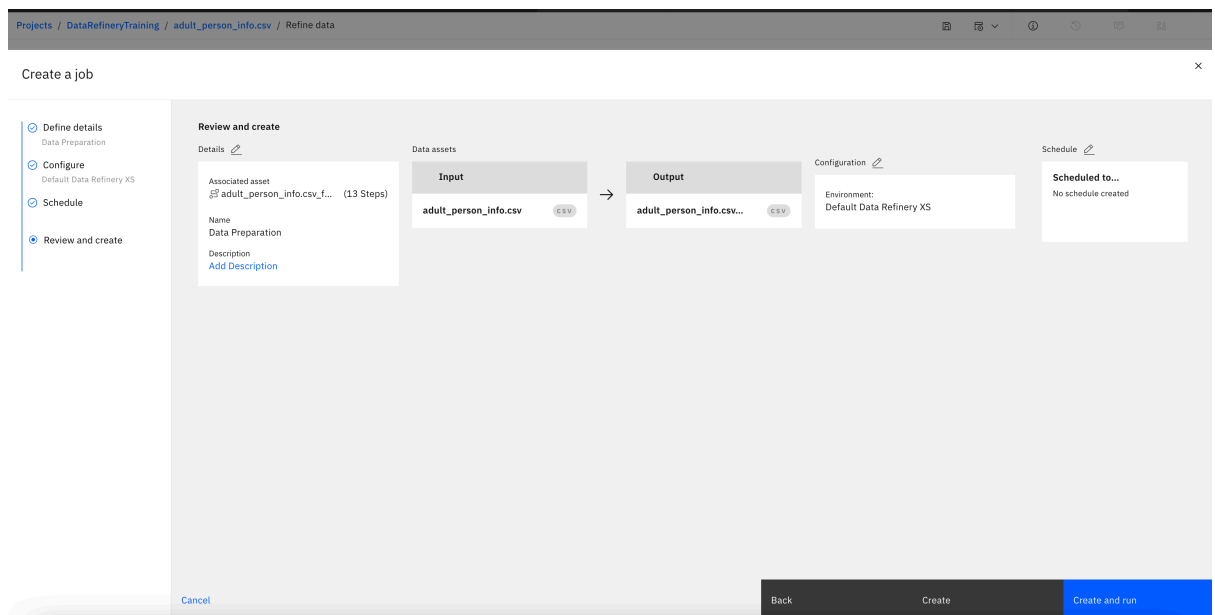


Step 12. Save and Run the Data Flow

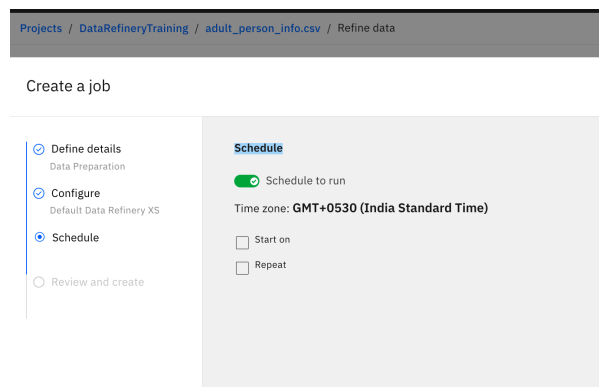
- Let's say our data preparation effort is complete and you want to run the data flow. Click on the Jobs on the top right corner, choose **Save and create job**.



- This will take you to a page where you will need to configure the Data Refinery flow details (the stream) and the Data Refinery Flow output (a file). Give the job a name (We named it Data Preparation) and keep the rest as it is. Click **Create** and **Run**.



- Note that if you click on the **Schedule** button, you will have the ability to schedule your data preparation, which is very useful if you do these steps on dynamic data.



- At this point, you should see a status on your data preparation job.
- If you go back to the **Assets** page of your project (by clicking on My Projects > <ProjectName>), you will notice that the new csv file has been added as a new data asset. You can also find the data flow you have created if you scroll down to the end of the same page. Note that you can refine your data flow at any time by clicking on the menu next to the data flow name.