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.

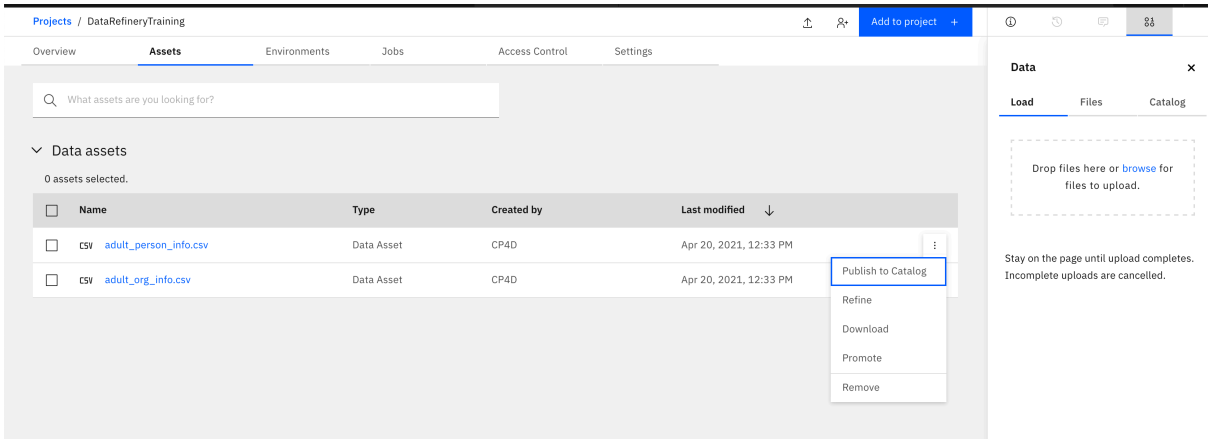
Estimated time

Completing this tutorial should take about 45 minutes.

Steps

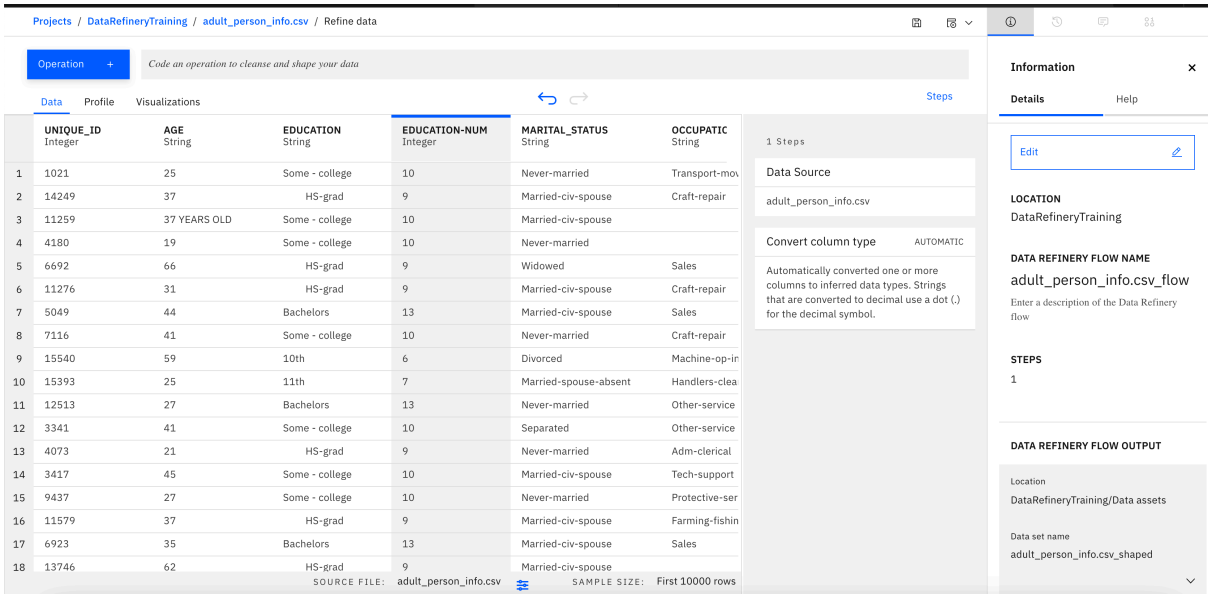
Step 1.Add Datasets to the Project

- 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.

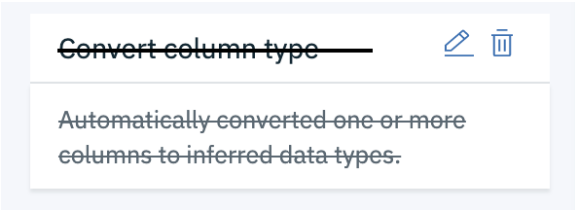


- Review Data Refinery UI:
    - Go to the ellipsis icon next to *adult_person_info.csv* under *Data assets* and select **Refine**. This will open a page that shows a sample of the content, where you can start cleaning and reshaping the data set.
    - On the panel on the right, you will find **Details** 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.

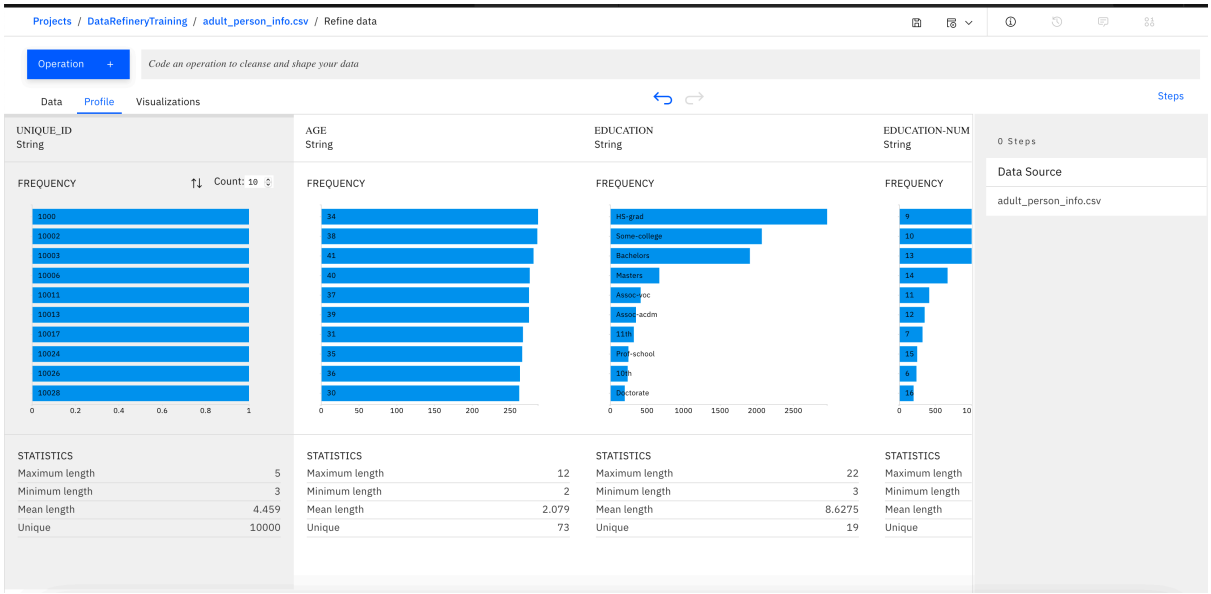


- You may notice that an auto data type conversion step has been applied, in order to practice how to manually convert column type, we will remove this step for now by clicking on the trash bin icon.



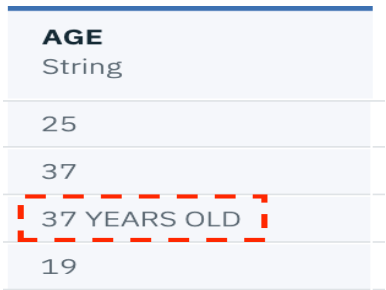
Step 2.Review Data Profile

- 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.



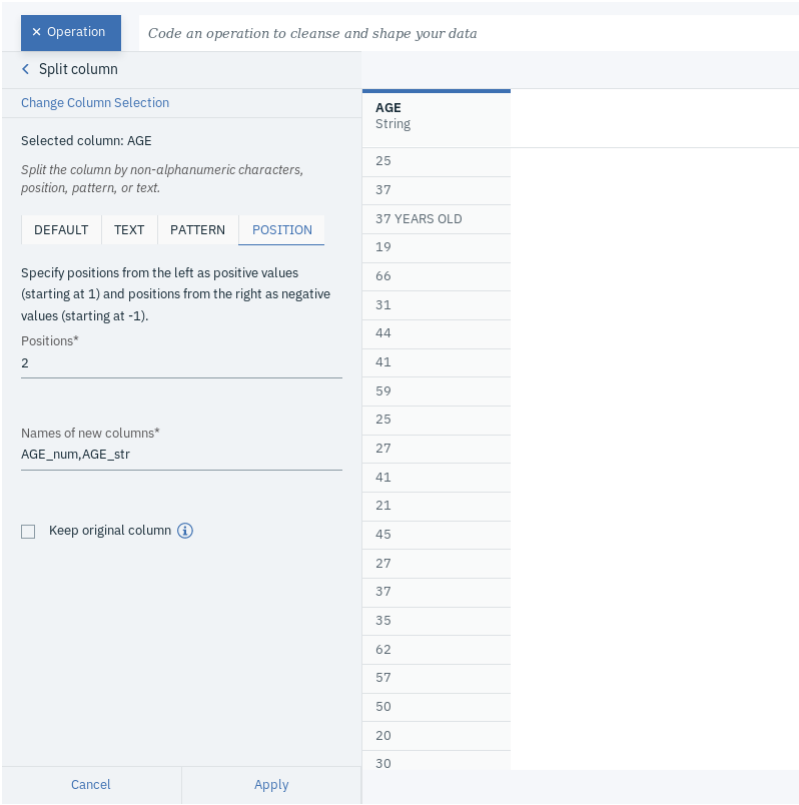
Step 3.Data Harmonization: AGE

- 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).

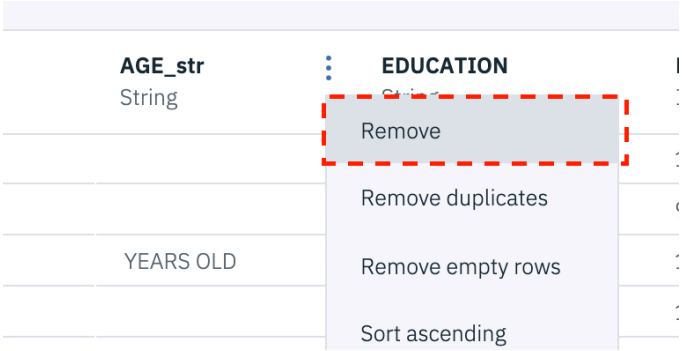


  - Click on **+Operation** and select **Split column**, which you can find under **ORGANIZE**.
  - Choose AGE as the **Selected column**.
  - Under **POSITION** tab, type 2 in the **Positions** field and enter names in the **Names of new columns**. Make sure to unselect **Keep original column**
  - Click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



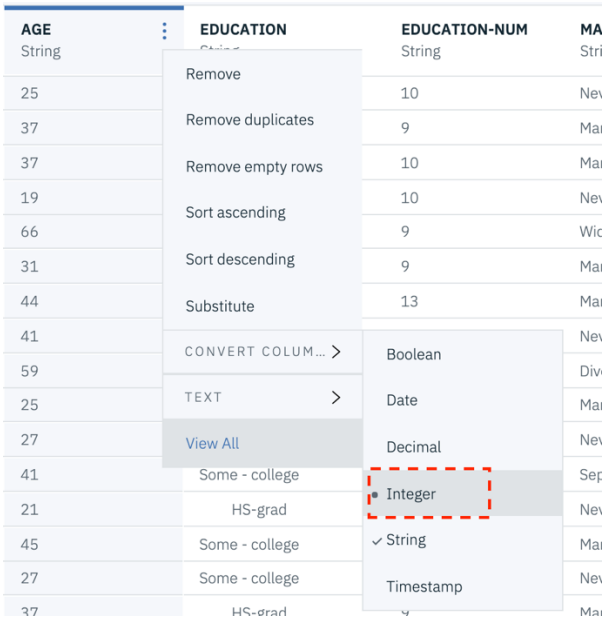
- Go to the **Data** tab and remove the newly created column called *AGE_str*, which only contain the string part of the age.



- 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.

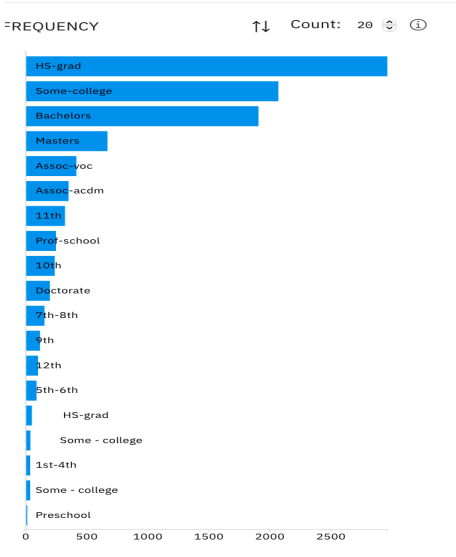
Step 4. Convert Data Type

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



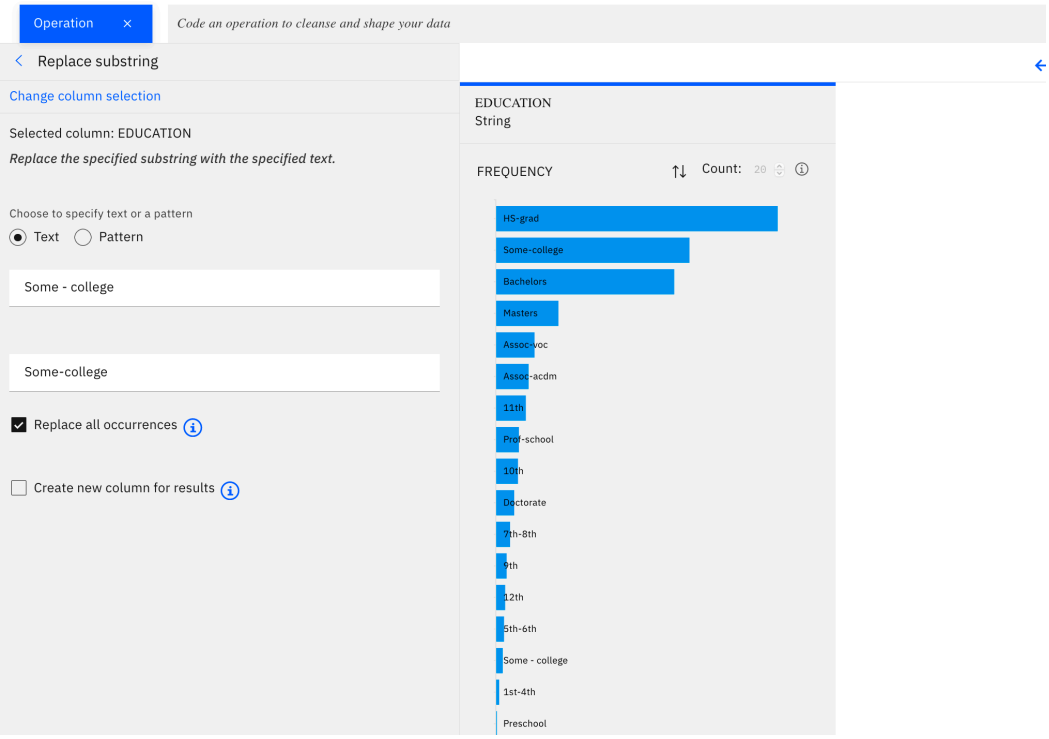
Step 5.Data Harmonization: EDUCATION

- 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.

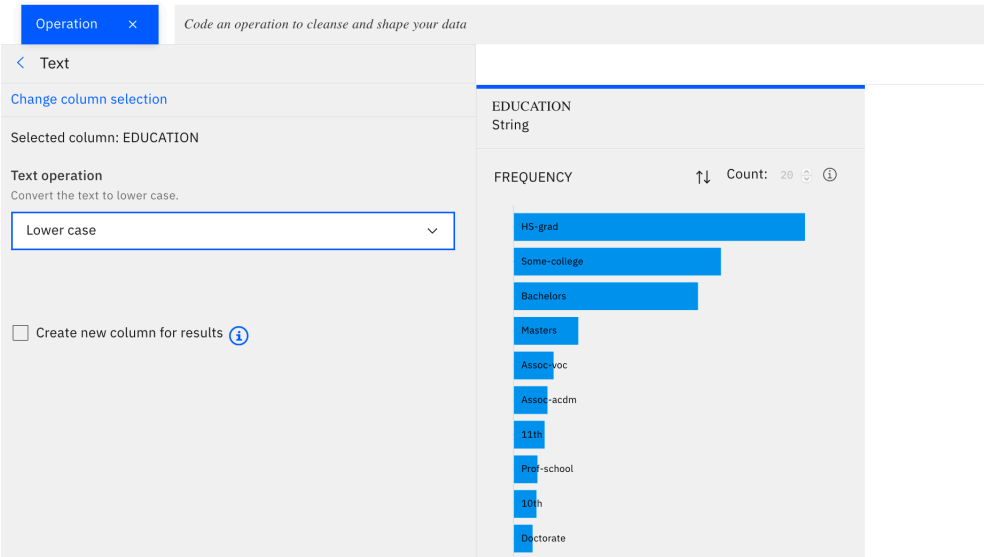


- Click on **+Operation** and select **Text**, which you can find under **FREQUENTLY USED**.
- Choose *EDUCATION* as the **Selected column**, **Collapse spaces** as the **Text Operation**.
- Click **Apply** and go to the **Profile** tab again to check if all the additional values have been removed. You 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**.

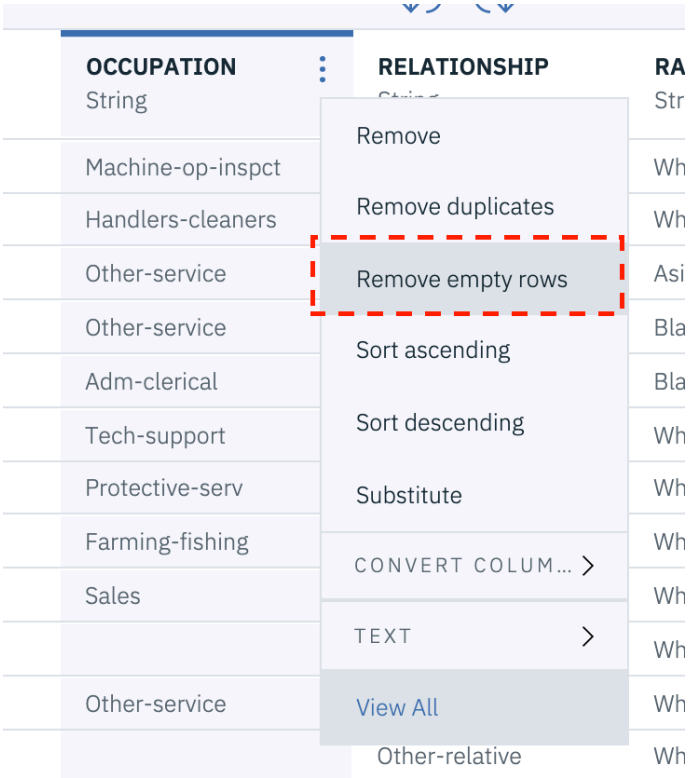


- We also want to convert all values in the EDUCATION column to lower case. So, click on **+Operation** and select **Text**, which you can find under **FREQUENTLY USED**.
- Choose EDUCATION as the **Selected column**, **Lower case** as the **Text Operation**.
- Click **Apply** and go to the **Profile**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



Step 6.Remove Missing Values in OCCUPATION

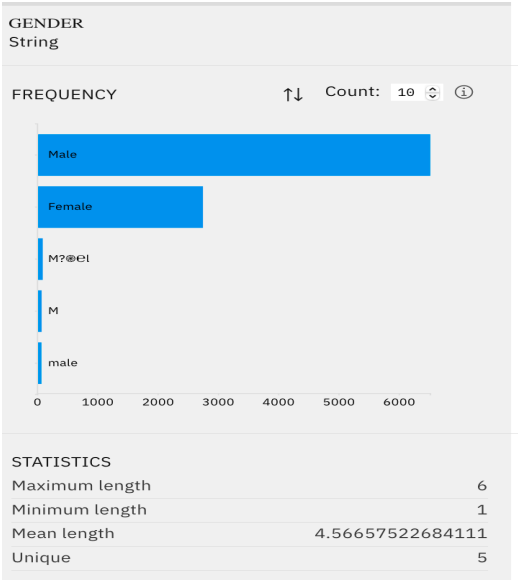
- Removing empty rows (List-wise deletion):
  - Go to the **Data** tab.
  - Go to the column called OCCUPATION and remove rows with any empty values by clicking menu next to the column name and selecting **Remove empty rows**.



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

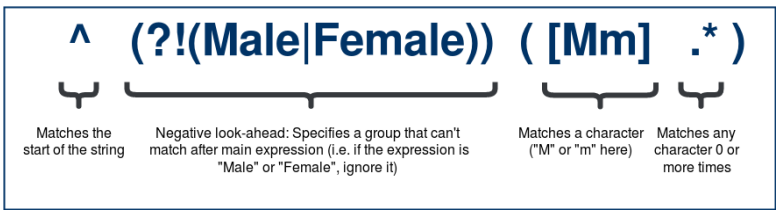
Step 7.Data Harmonization: GENDER

- 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.

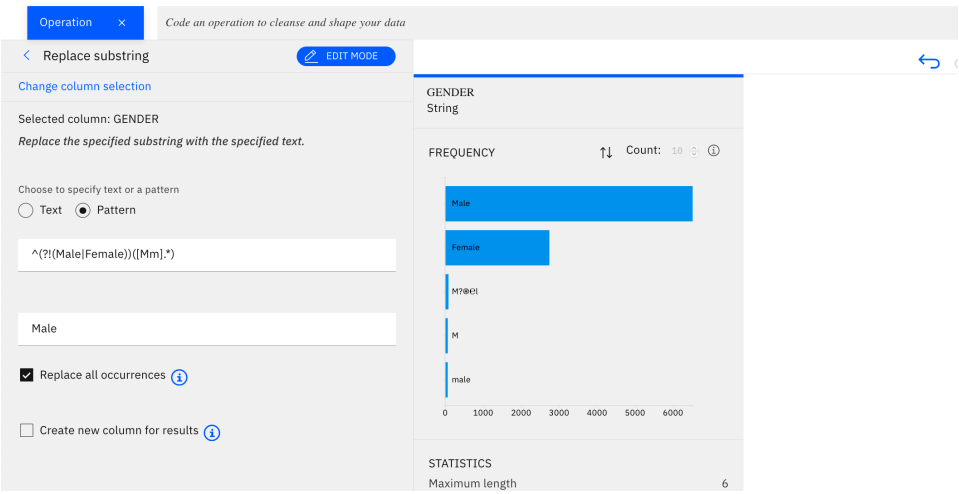


- Click on **+Operation** and select **Replace substring**, which you can find under **CLEANSE**.
- Choose **GENDER** as the *Selected column*.
- Under **PATTERN** tab, type ^(?!(Male|Female))([Mm].*) in the **Regular expression** field and Male under **Enter the replacement string**. Make sure to select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

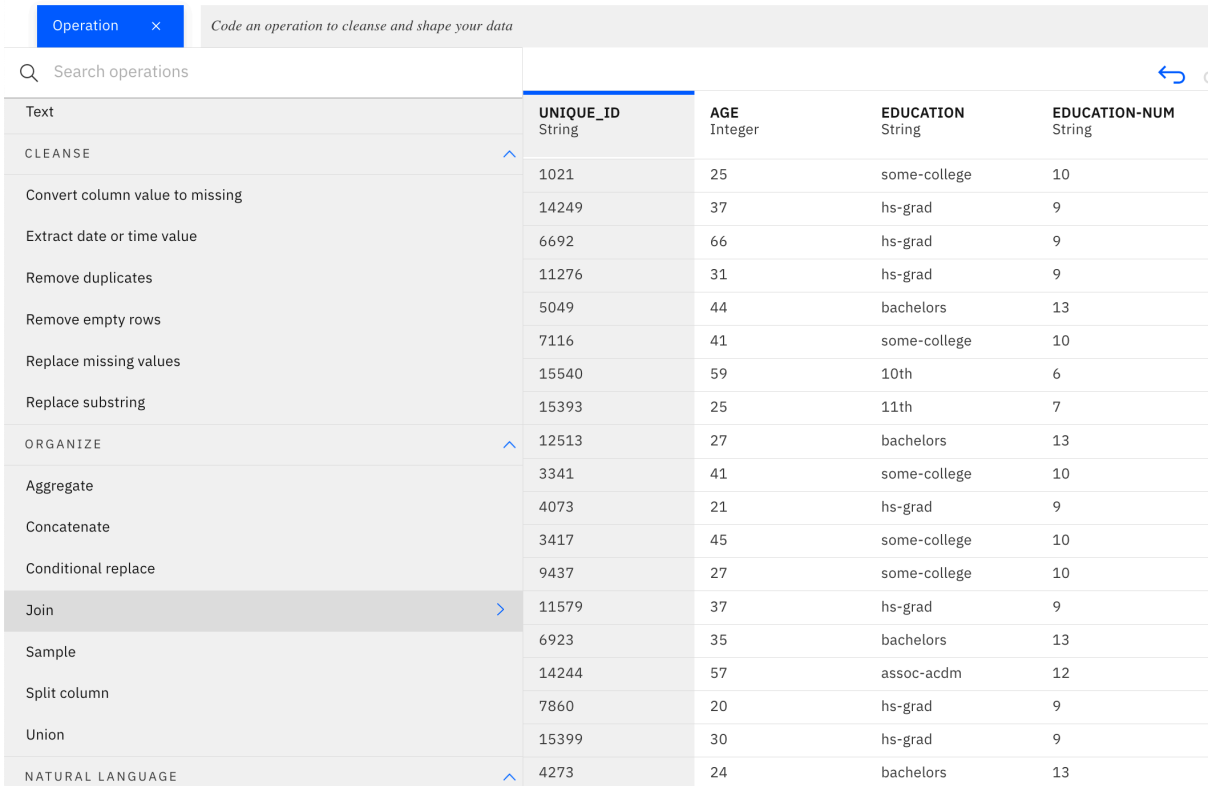


Step 8.Remove Duplicates

- 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**.

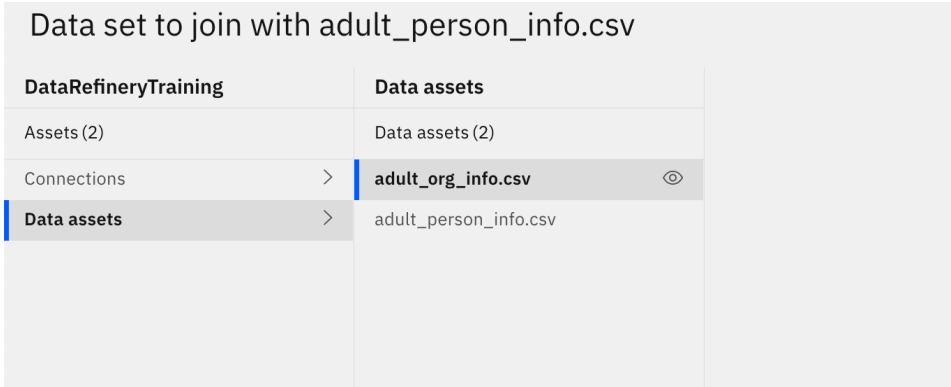
Step 9.Join Datasets

- 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*.

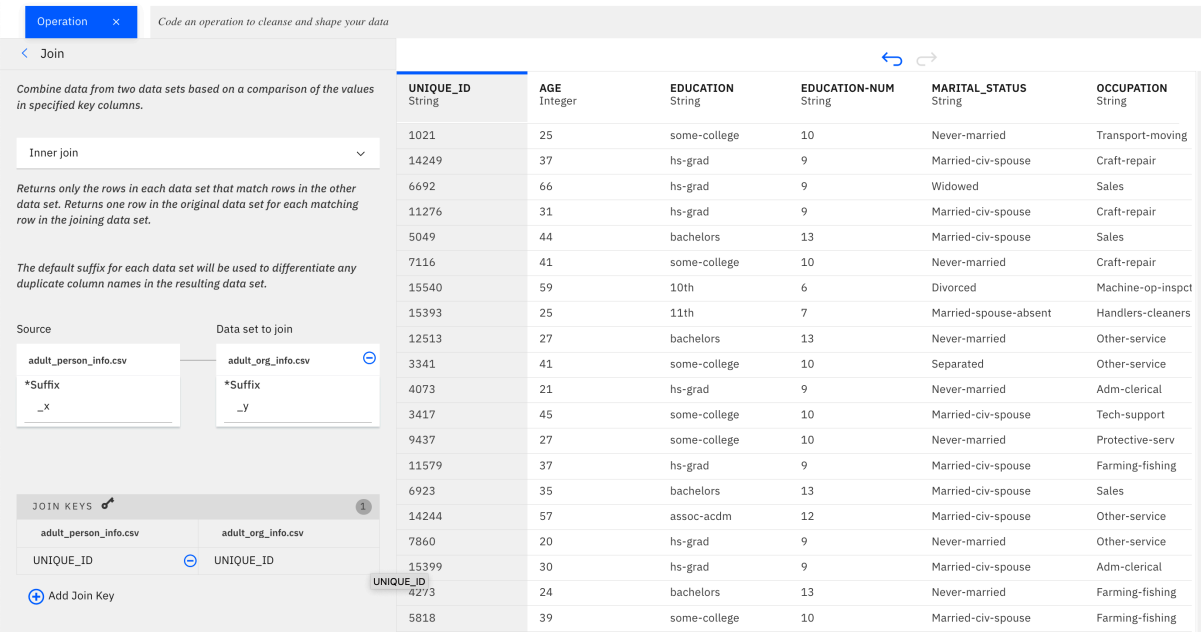


PS: Make sure UNIQUE_ID for both datasets are either String or Integer format to sync

- 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**.



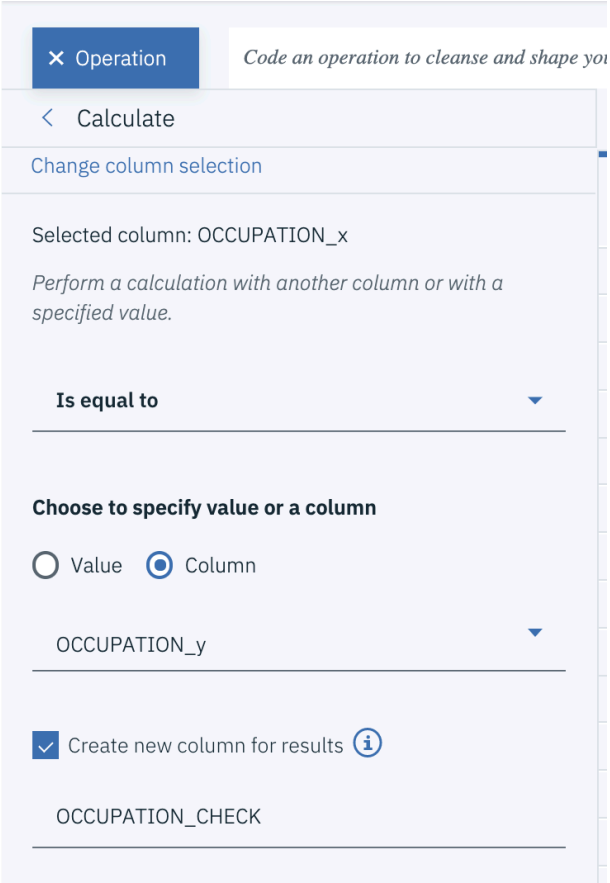
- Use the default values for the **Suffix** field, which is just a way for you to differentiate any duplicate fields resulted during the joining process. You can also modify it if you want.
- For the **JOIN KEYS**, select UNIQUE_ID representing the employee ID, as the join key for both data sets and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



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

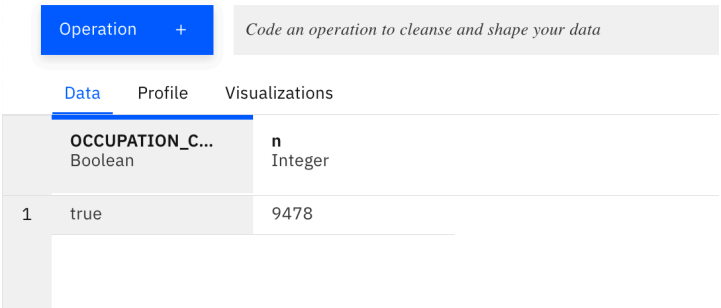
- We will notice that there are 2 columns representing **OCCUPATION**, one coming from each of the data sets. Let's check to see if they contain the exact same values.
    - Click on **+Operation** and select **Calculate**, which you can find under FREQUENTLY USED.
    - Choose OCCUPATION_x as the Selected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



- 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**.

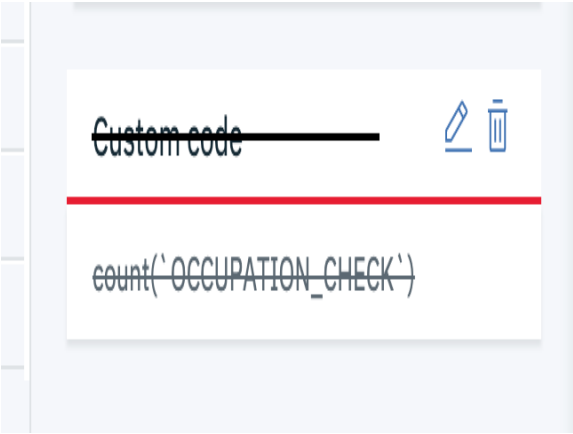
- The result shows that OCCUPATION_x and OCCUPATION_y have identical values.



Step 10.Undo Steps

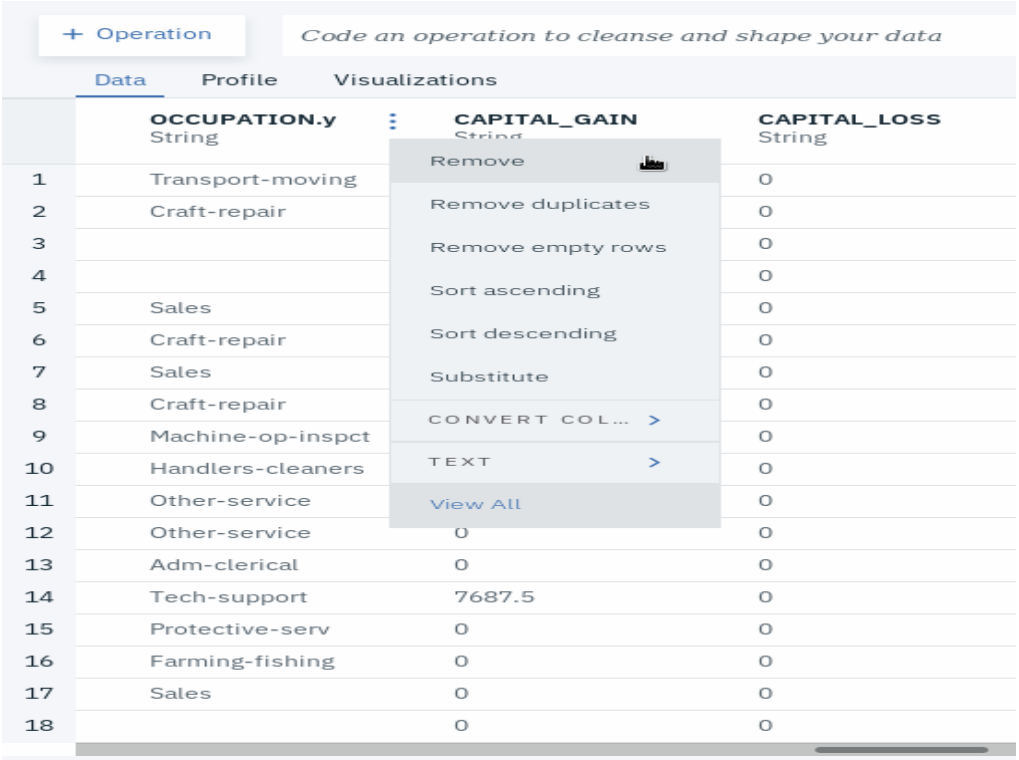
- So we can just keep one of them and keep one OCCUPATION column:
  - Go back **2 steps** by either clicking on the Undo button found at the top middle of the page or by going to the step added under **Steps** and clicking on the bin icon. Whichever way you select, you will need to do it **twice**.

Delete steps



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

- Go to column called *OCCUPATION_y* and remove

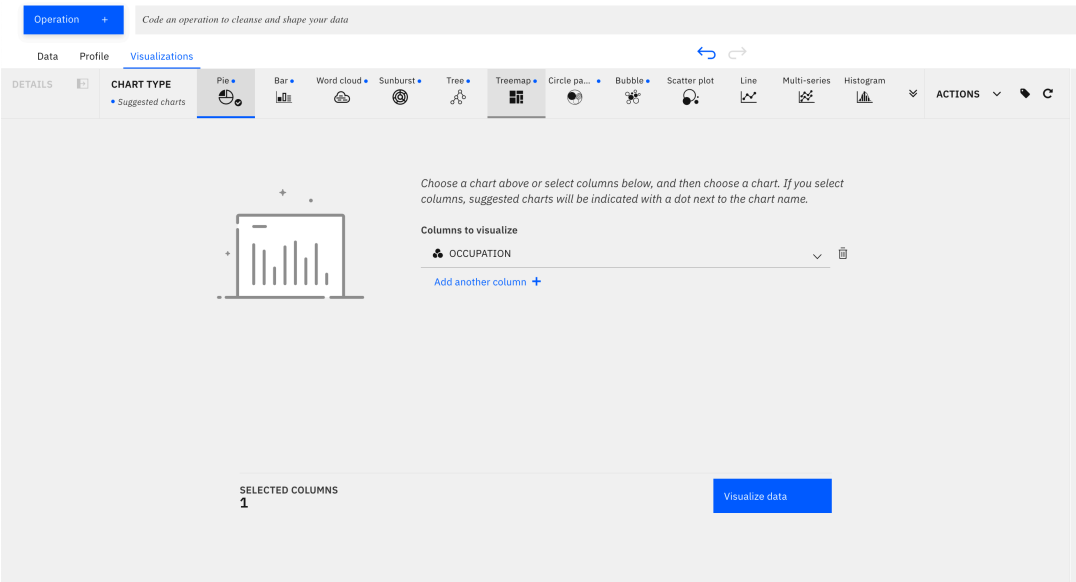


Step 11.Data Visualization

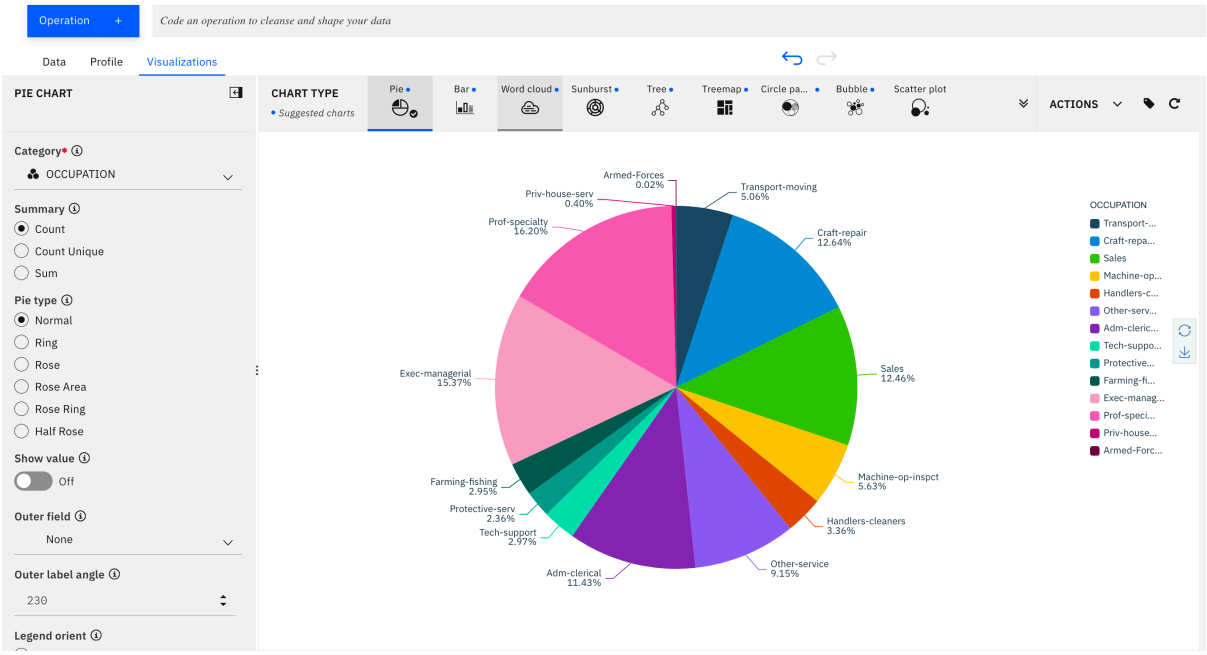
Data Refinery does not only include powerful shaping operations to clean, organize, fix, and validate data but also has built-in visualization capability to derive insights from data.

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

Visualize OCCUPATION

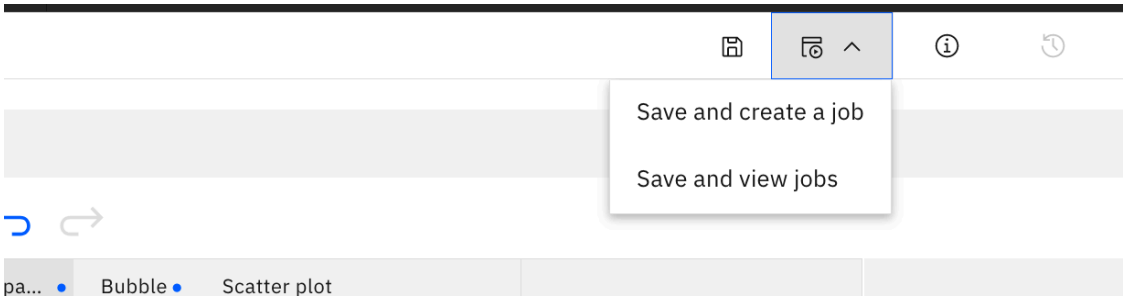


- Data Refinery will suggest the best techniques to visualize the data. For OCCUPATION, it has suggested the pie chart showing you the distribution of OCCUPATION. Click on other **CHART TYPES** to see other visualization outputs if interested.

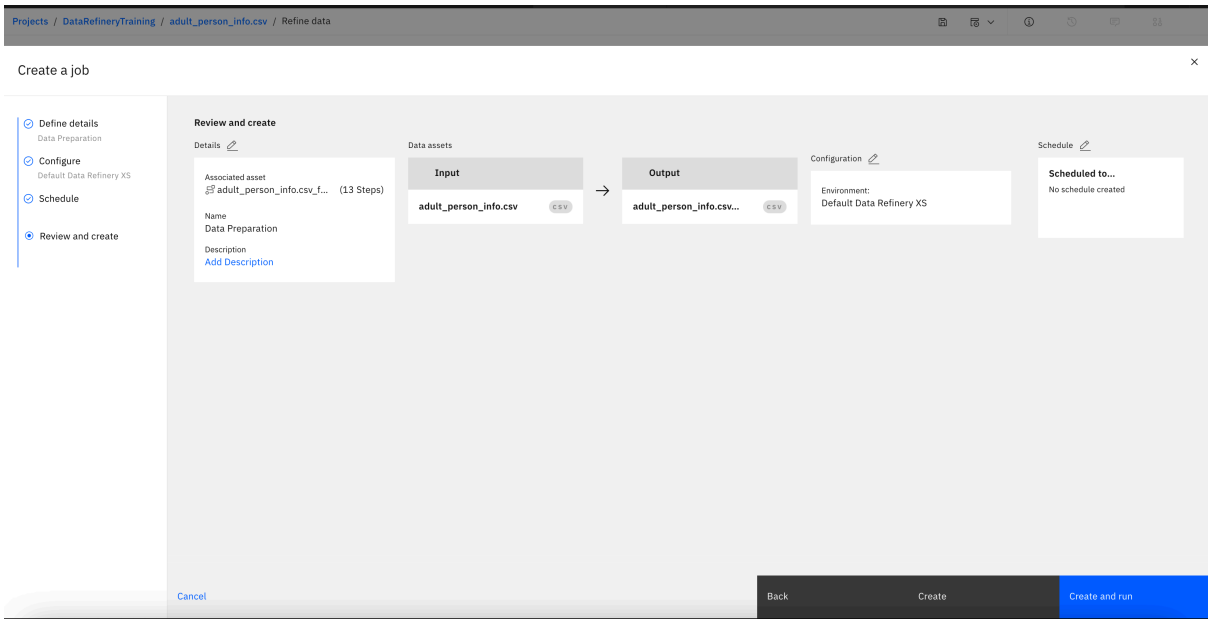


Step 12.Save and Run the Data Flow

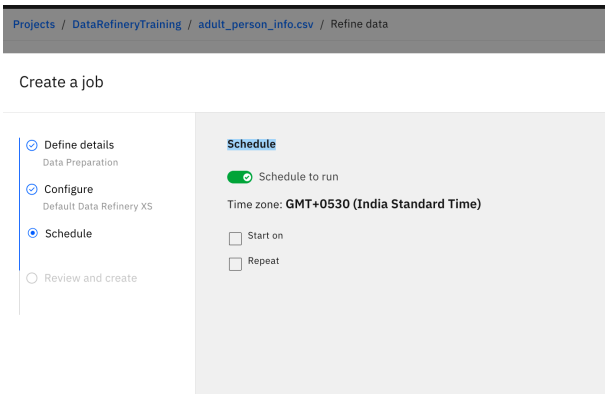
- Let's say our data preparation effort is complete and you want to run the data flow. Click on the Jobs on the top right corner, choose **Save and create job**.



- 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.