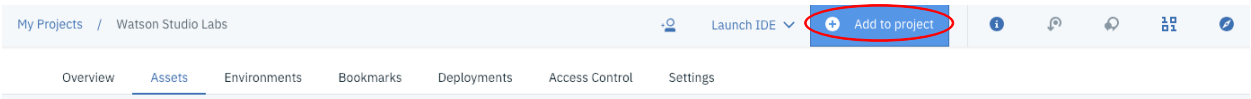
Data Refinery Lab

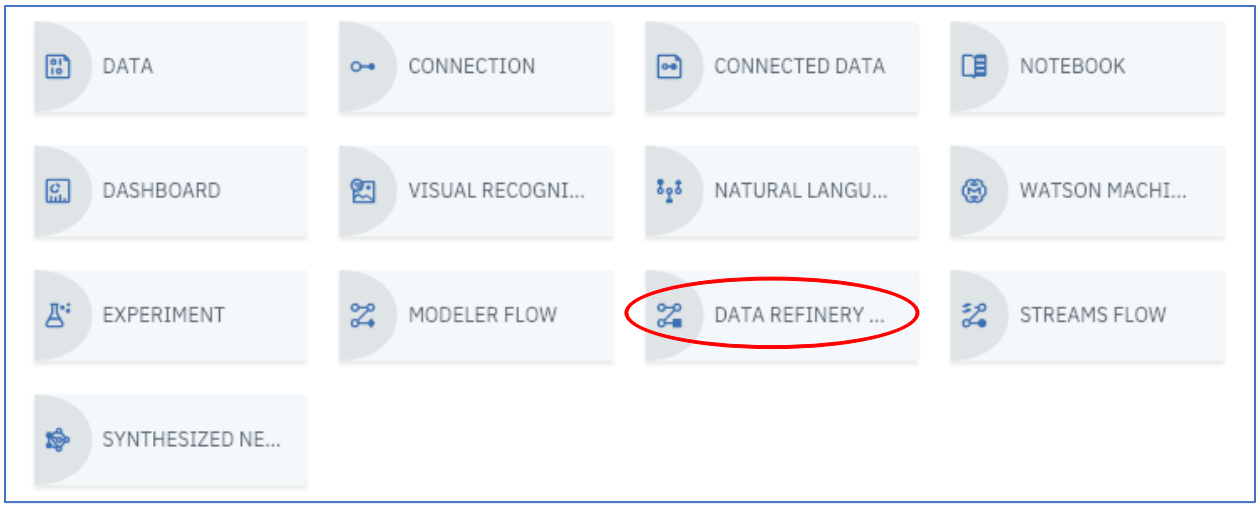
This lab will use the Titanic data set to demonstrate data profiling, data visualization, and data preparation capabilities of the Data Refinery tool. The lab consists of the following steps:

1. Use the Data Refinery Tool to:
    a. Profile the data to help determine missing values
    b. Visualize the data to gain a better understanding
    c. Prepare the data for modeling
    d. Run the sequence of data preparation operations on the entire data set.

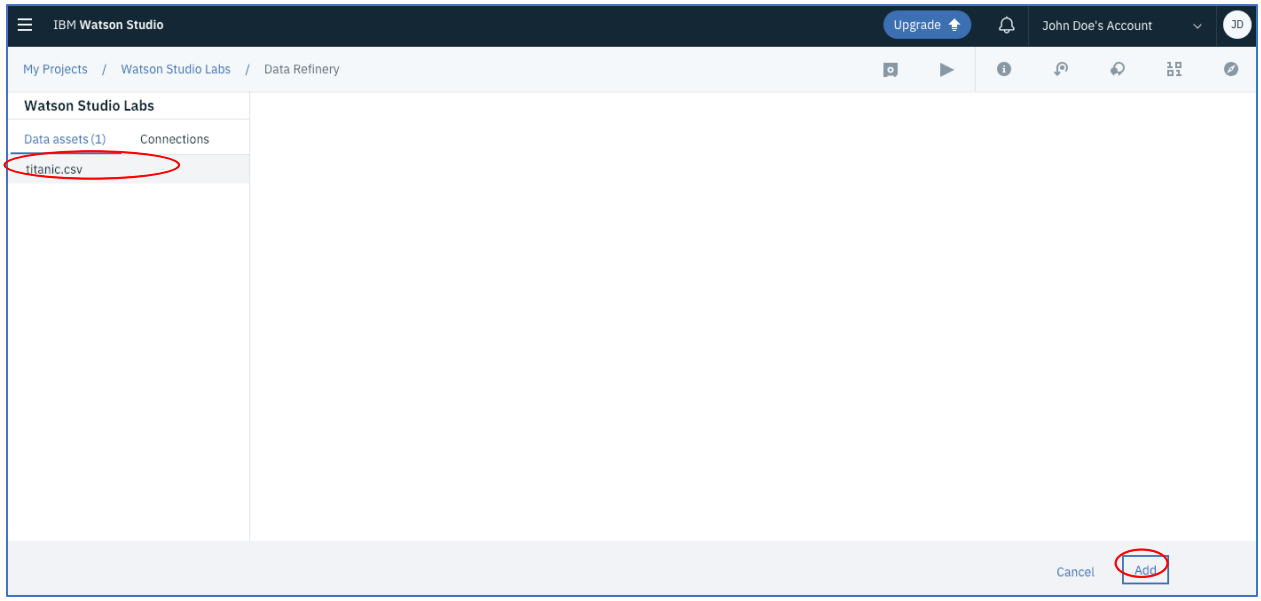
Step 1: Profile the data to help determine missing values.

1. Add a Data Flow by clicking on **Add to project** and then click **Data Refinery flow.**

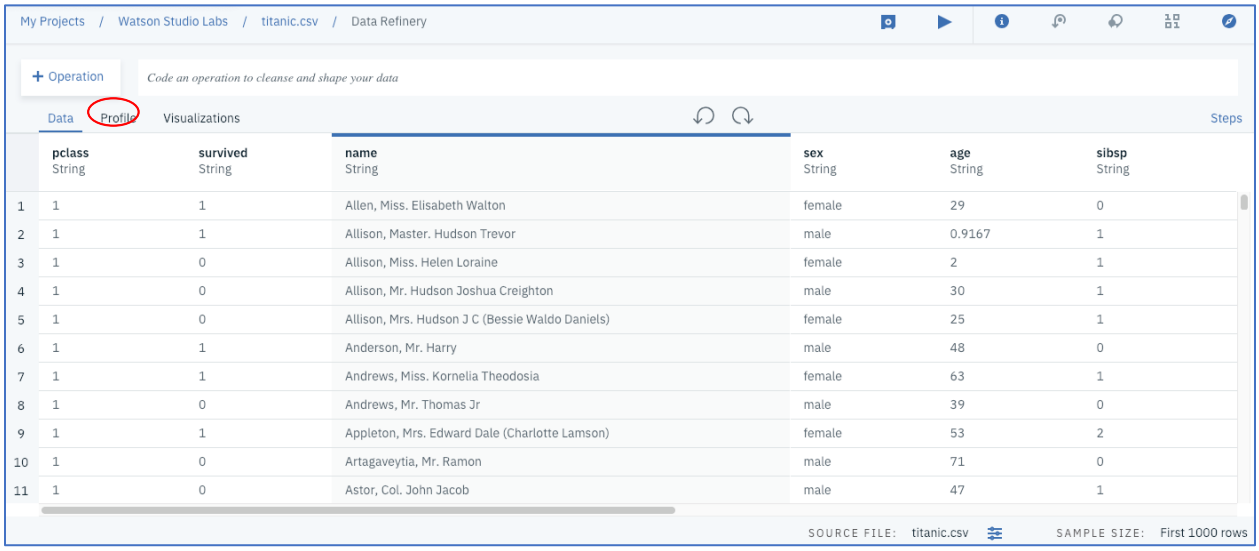




2. Select **titanic.csv** and then click on **Add**.

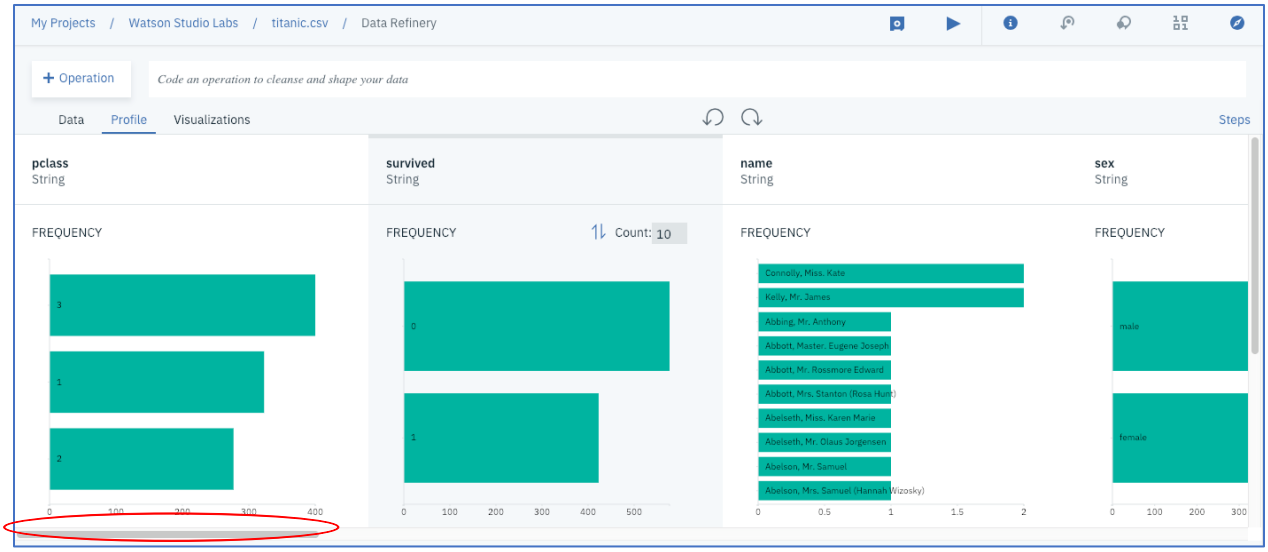


3. The Data Refinery panel will display the Titanic data set. Click on the **Profile** tab.

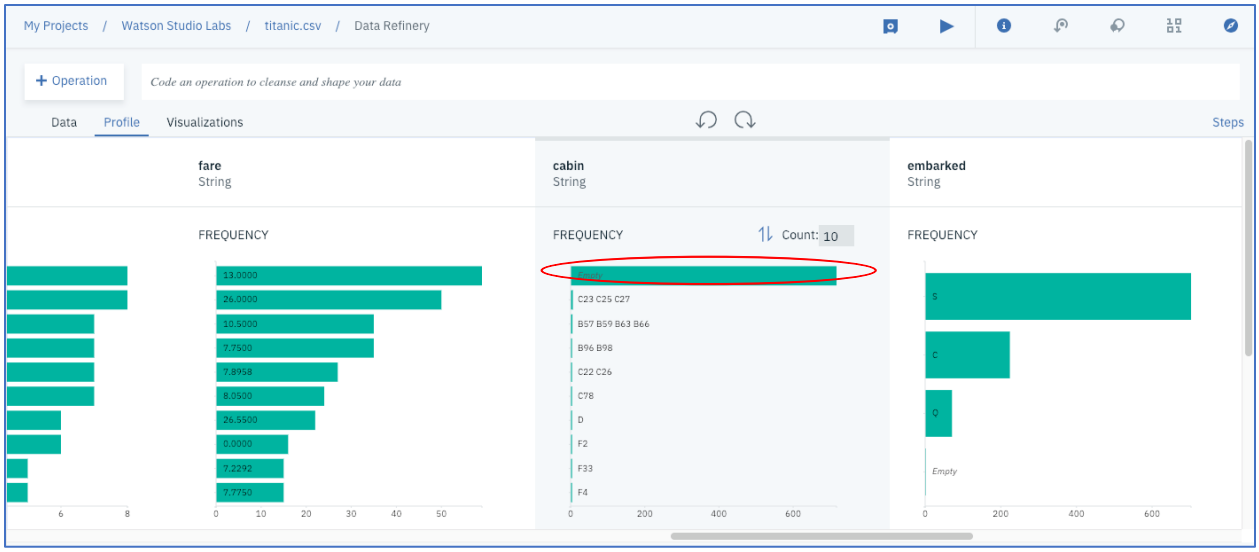


	pclass String	survived String	name String	sex String	age String	sibsp String
1	1	1	Allen, Miss. Elisabeth Walton	female	29	0
2	1	1	Allison, Master. Hudson Trevor	male	0.9167	1
3	1	0	Allison, Miss. Helen Loraine	female	2	1
4	1	0	Allison, Mr. Hudson Joshua Creighton	male	30	1
5	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25	1
6	1	1	Anderson, Mr. Harry	male	48	0
7	1	1	Andrews, Miss. Kornelia Theodosia	female	63	1
8	1	0	Andrews, Mr. Thomas Jr	male	39	0
9	1	1	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53	2
10	1	0	Artagaveytia, Mr. Ramon	male	71	0
11	1	0	Astor, Col. John Jacob	male	47	1

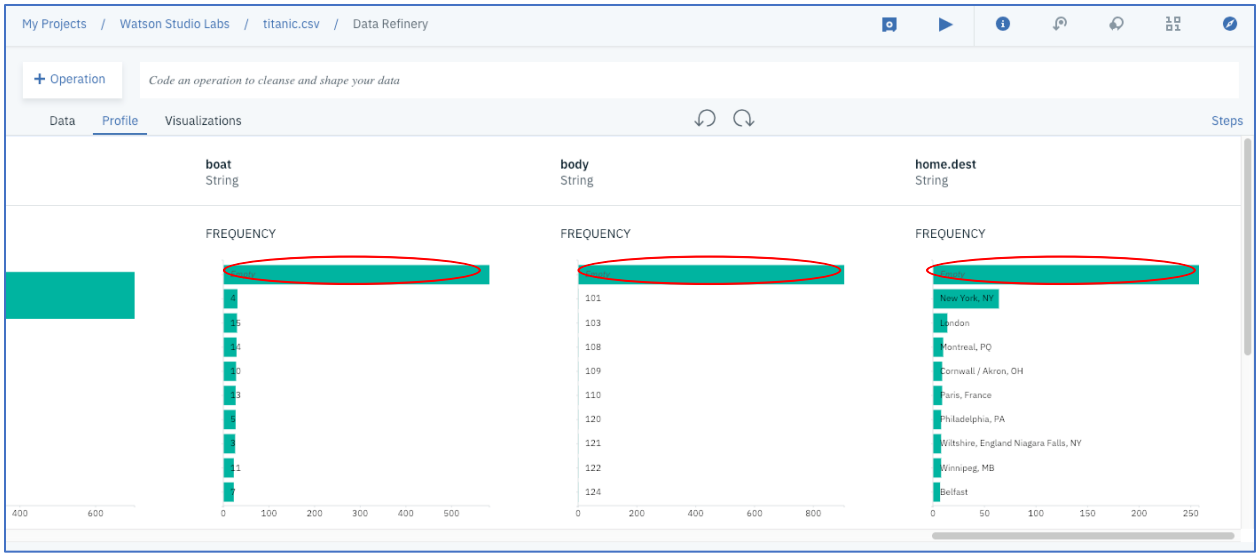
4. The Profile panel displays the counts of the top 10 count values for each column. Note that you can change 10 to another number if desired. You can also switch to the bottom 10 counts for a column. Scroll to the right to view the cabin column.



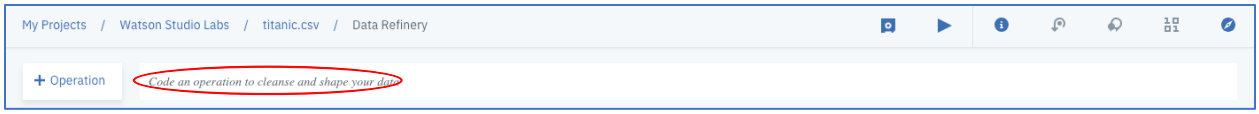
5. Note that the cabin column has many missing values and should be removed as part of the data preparation step.



6. In a similar fashion, scroll to the right to examine the boat, body, and home.dest columns. These also have many missing values and should be removed as part of the data preparation step.

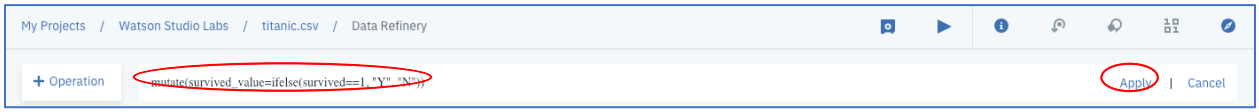


7. Age and Embarked also have missing values. Embarked has very few missing values. Age has over 100 missing values, but we will keep that column in the analysis. As part of data preparation, we will remove the rows that contain the missing age and embarked values.
8. Click on the **Data** tab. We will add columns that contain more readable values for the survived and pclass columns. The column survived_value will contain a "Y" or "N". The pclass_value column will contain "first", "second", or "third". We will use the mutate (R dpylr function) and ifelse functions to do the conversion. Click on the **Code an operation to cleanse and shape your data.**

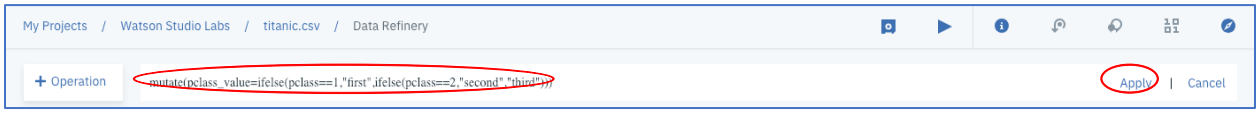


9. Copy and paste the following:
   mutate(survived_value=ifelse(survived==1, "Y", "N"))

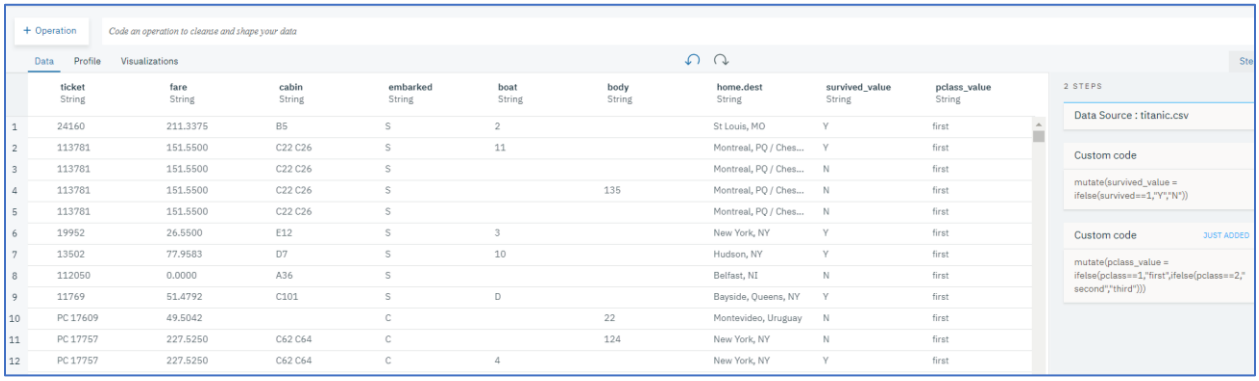
   and then click Apply. If you scroll to the right you should see the new column "survived_value".



10. Copy and paste the following to create pclass_value,
    mutate(pclass_value=ifelse(pclass==1,"first",ifelse(pclass==2,"second","third")))

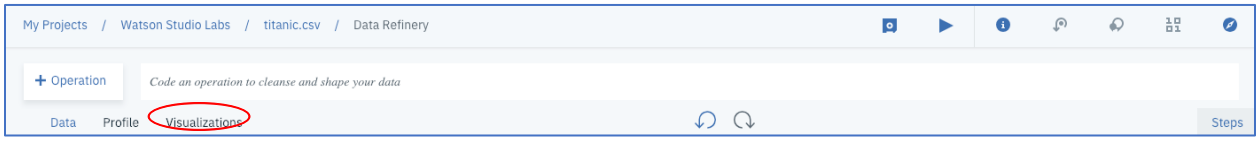


11. The result is shown below. Notice that the right panel will contain a running list of the transformations.

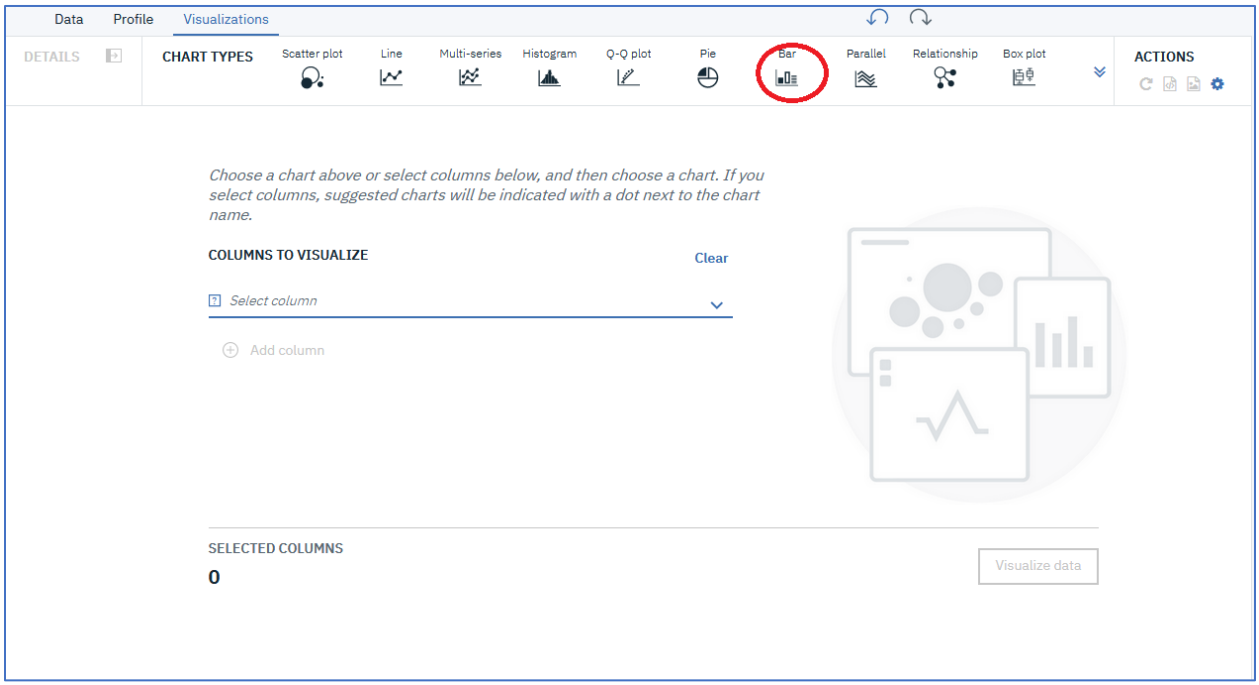


Step 3: Visualize the data to get a better understanding

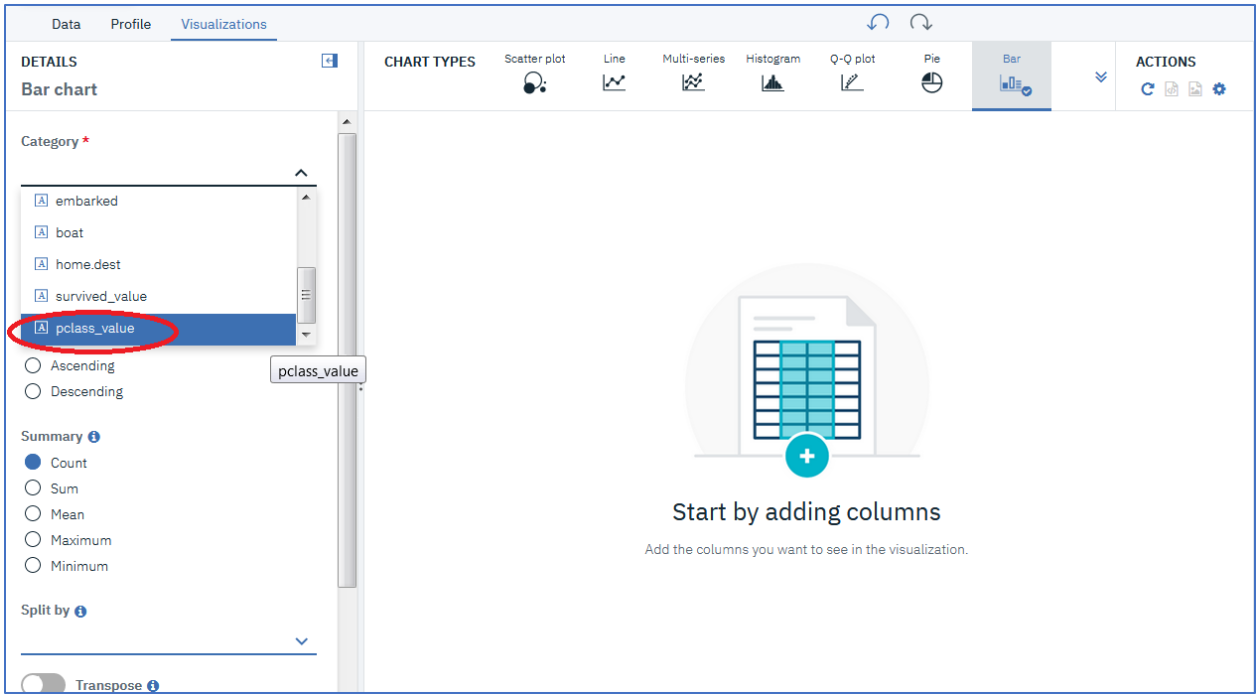
1. Click on the **Visualizations** tab.



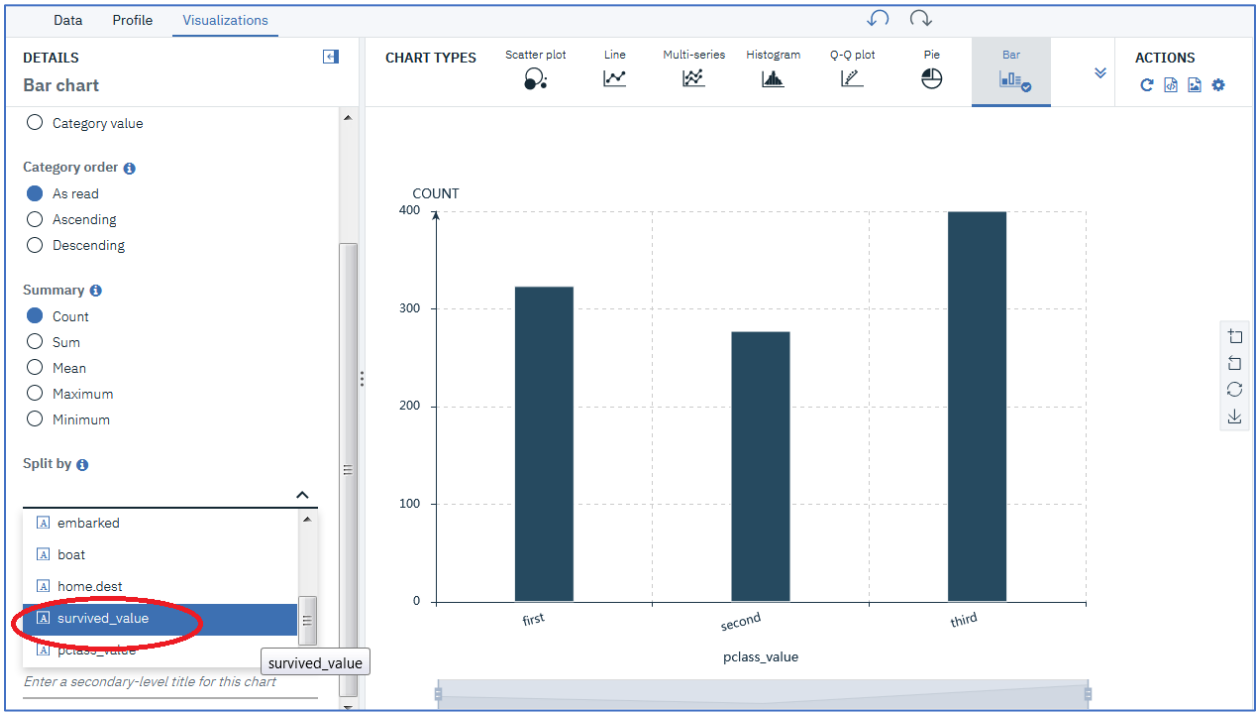
2. Let's take a look at the breakdown of passengers by passenger class. We will use our new pclass_value field. Select the **Bar** Chart Type.



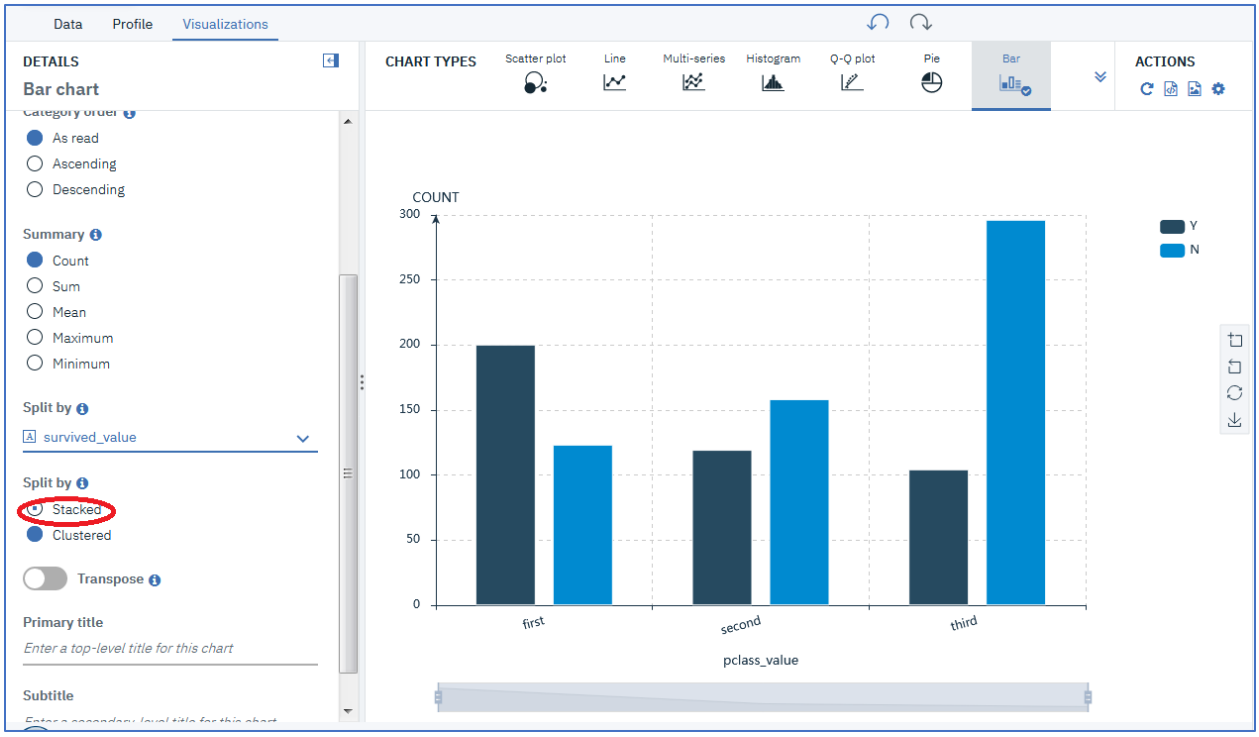
3. In the **Category** required field, select **pclass_value**.



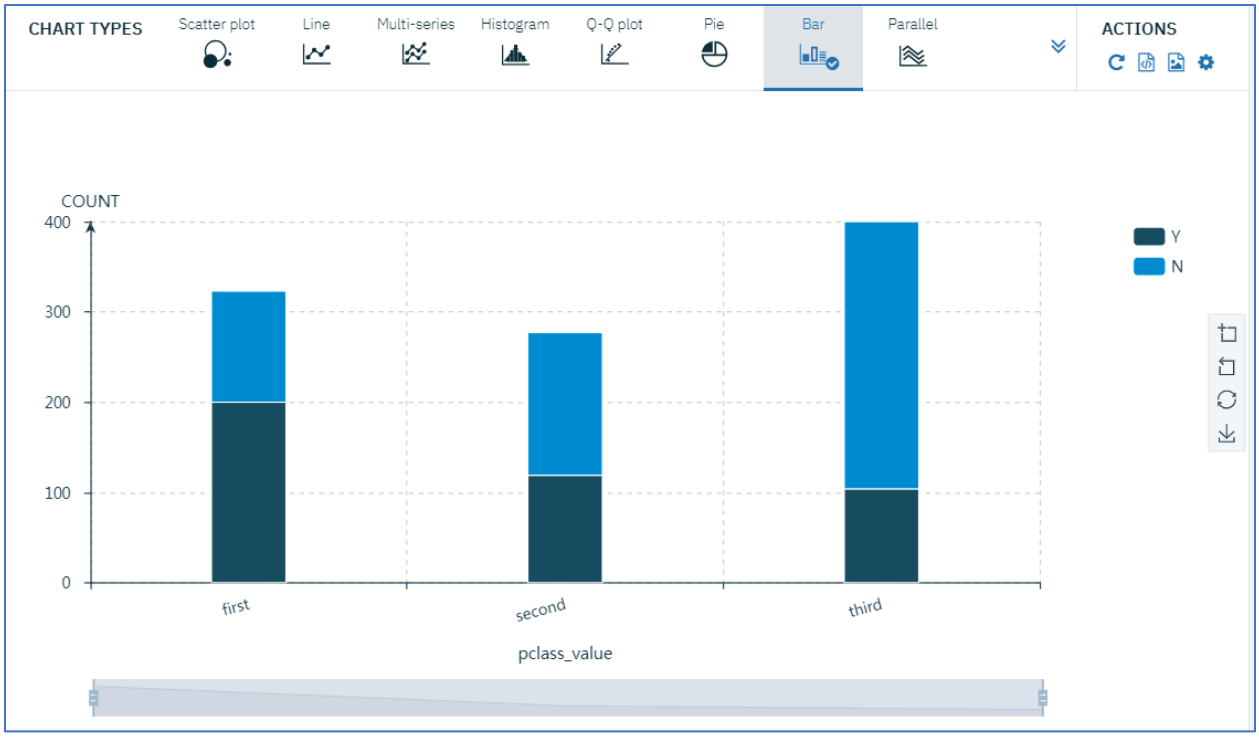
4. In the **Split by** field, select **survived_value**.



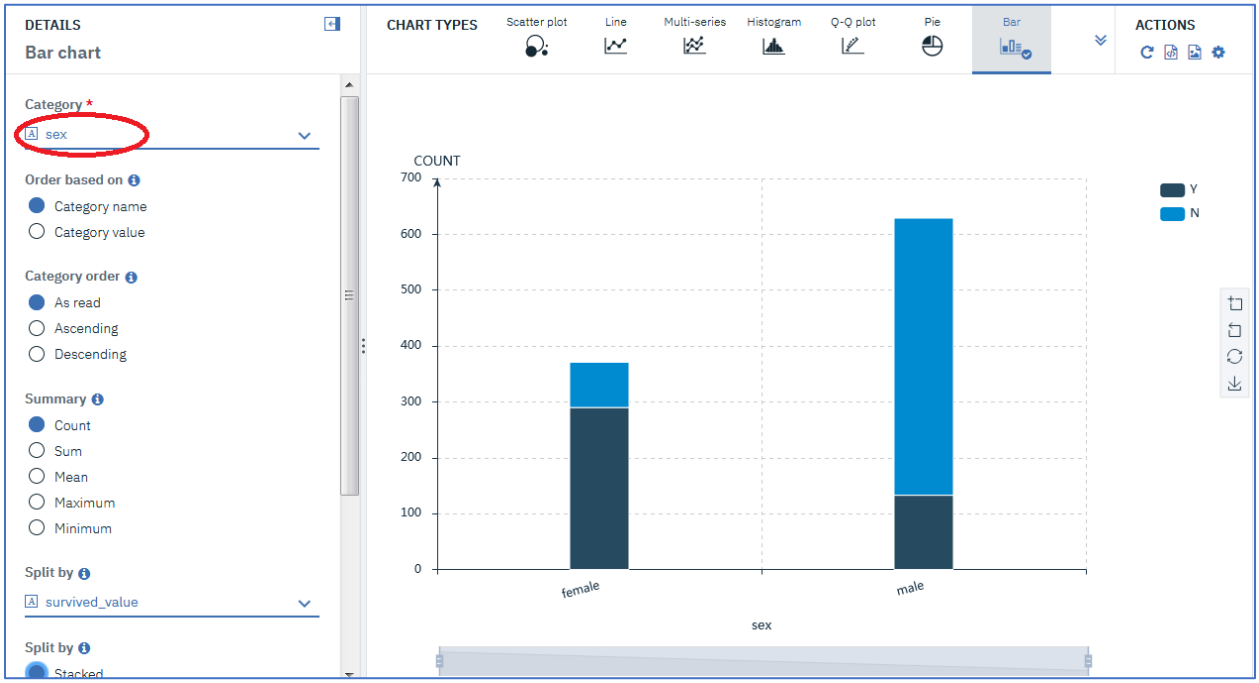
5. Select **Stacked**.



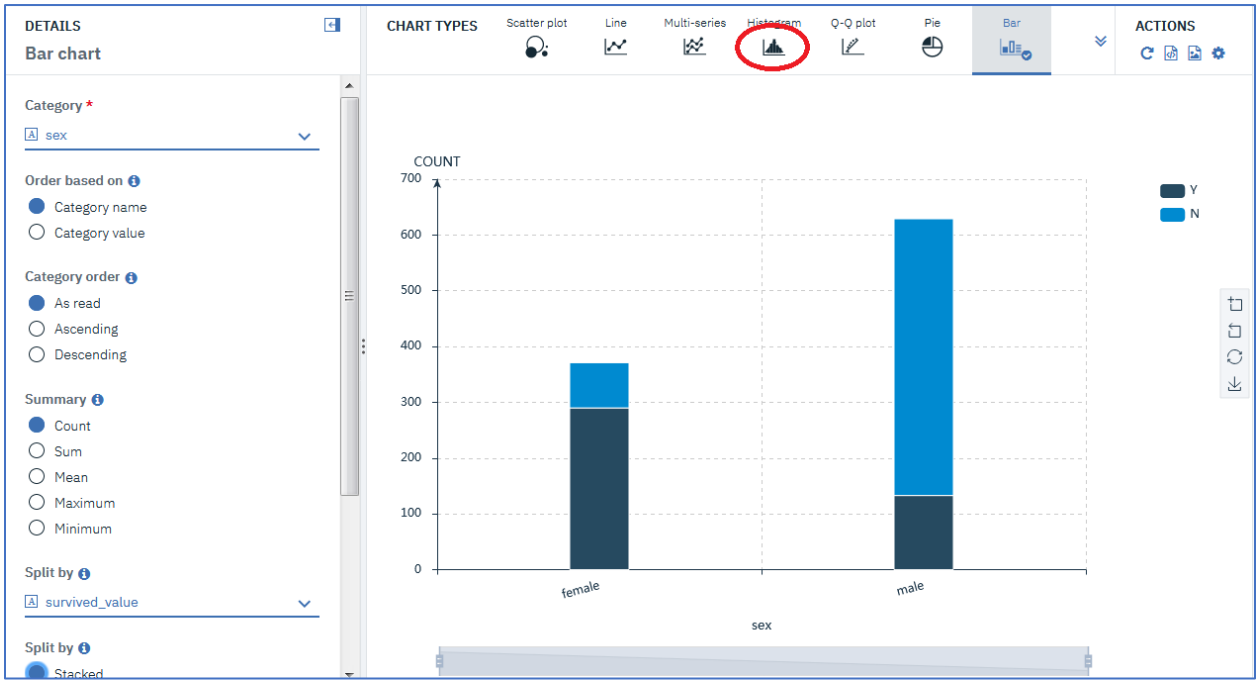
6. The result is shown below. The percentage of survivors is the greatest in first class, followed by second class, and then third-class passengers.



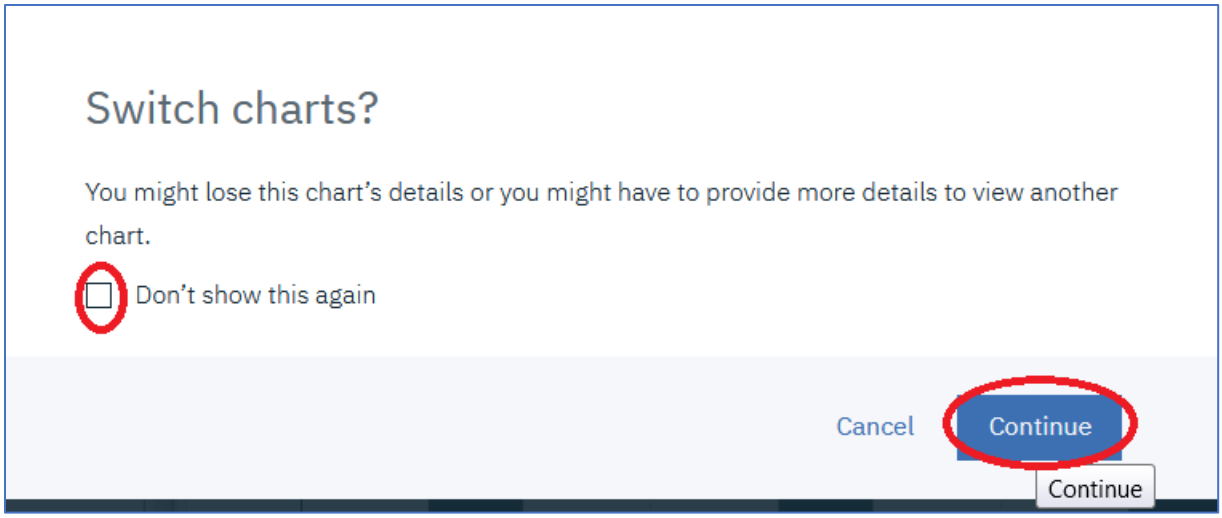
7. Change the **Category** to **sex**. We can see that survivorship for females is significantly greater than for males.



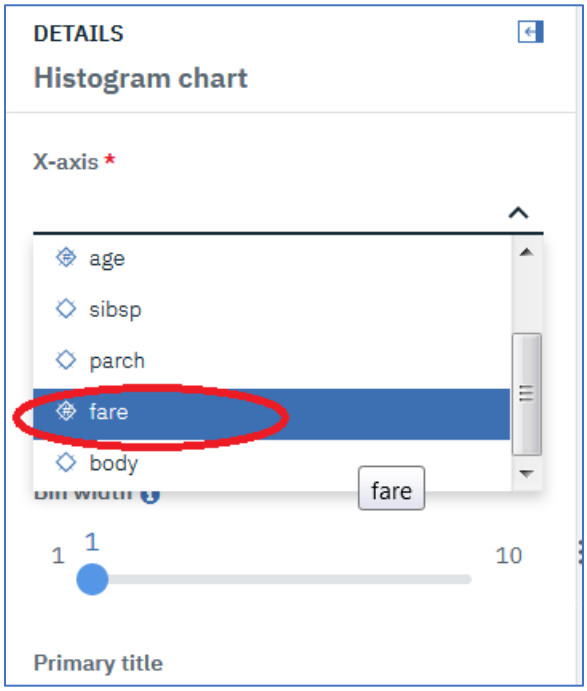
8. Click on the **Histogram** Chart Type.



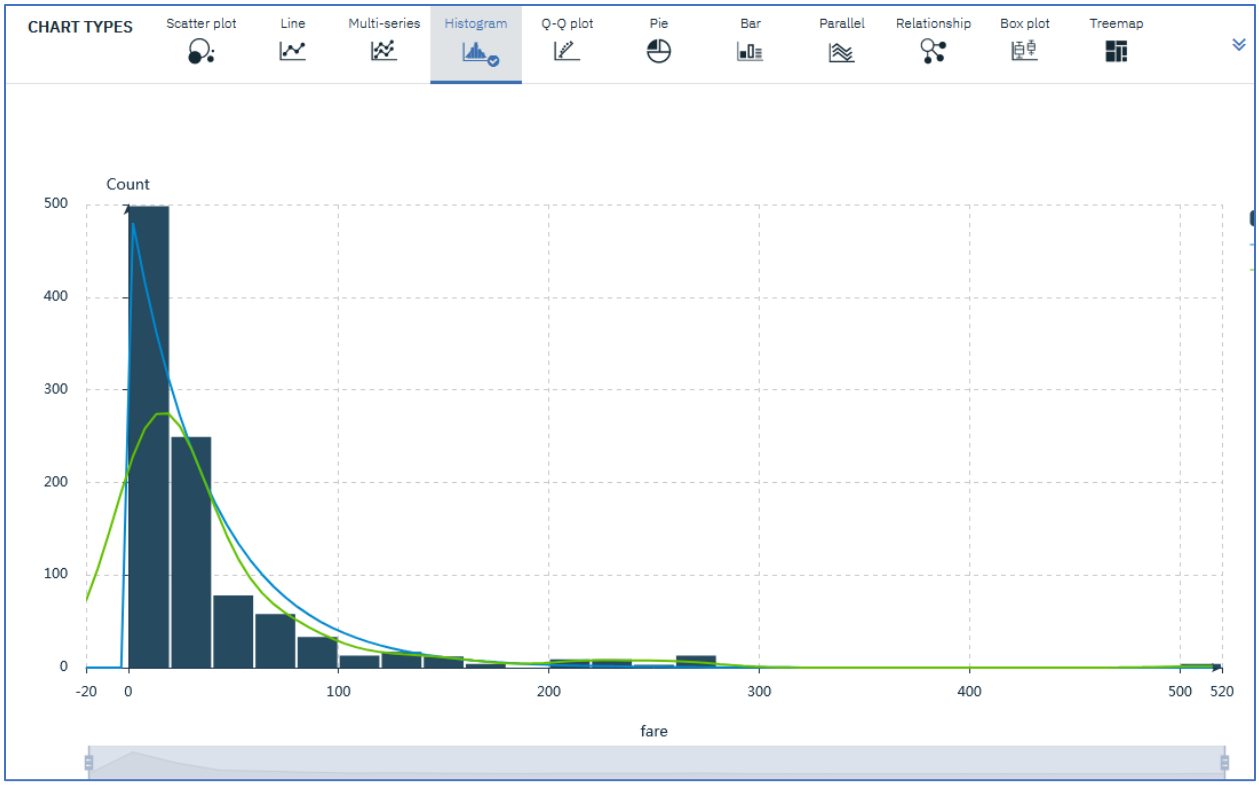
9. Click on the **Don't show this again** check box and click **Continue**.



10. Select **fare** for the X-axis.



11. The result is shown below. Note that it is highly skewed which affects the performance of some machine learning algorithms. One way to deal with this is to apply a logarithmic transformation. We will do that as part of data preparation.



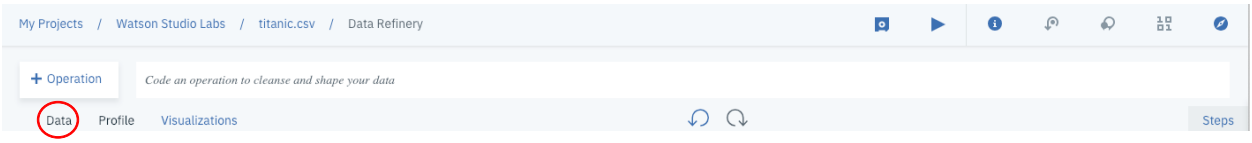
Step 4: Prepare the data for modeling

Based on the data analysis, we need to do the following to prepare the data for modeling.

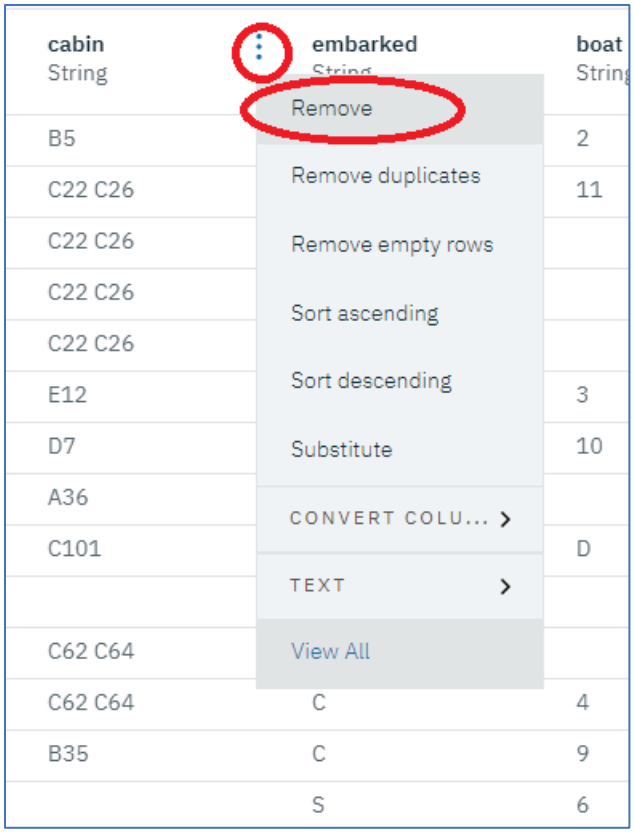
1. Remove columns cabin, boat, body, home.dest
2. Remove rows with missing values of age, and embarked.
3. Create a new column(log_fare) that is the logarithm of the fare column

We will also bin the age, and log_fare fields.

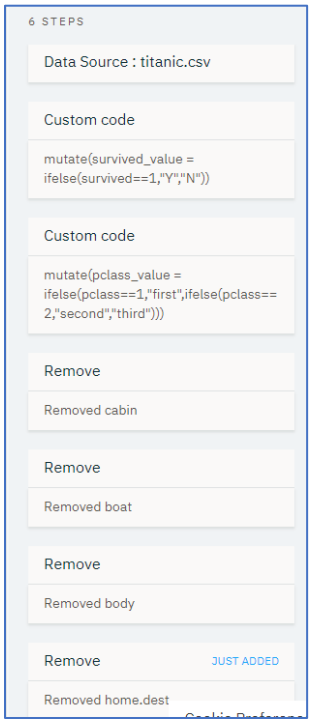
1. Return to the Data panel by clicking on the **Data** tab



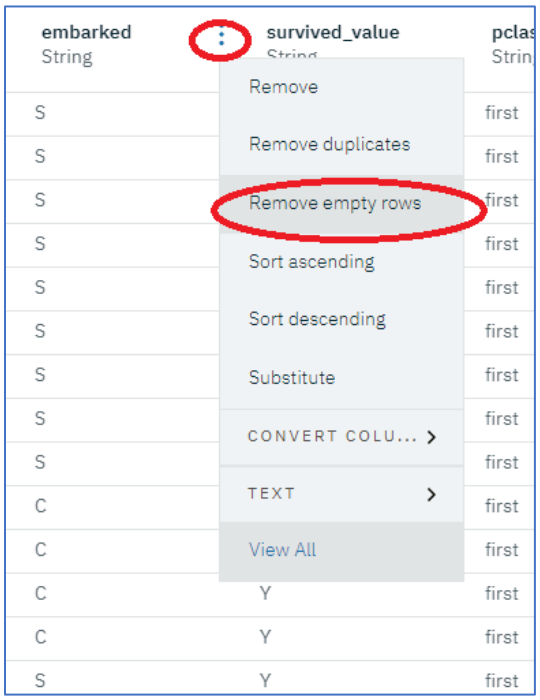
2. Remove the **cabin** column by selecting on the vertical ellipse and then clicking on **Remove**.



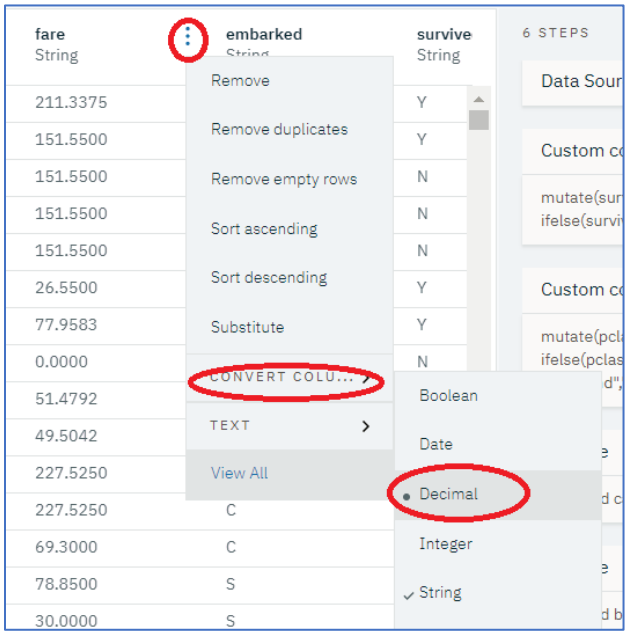
3. Remove the **boat**, **body**, and **home.dest** columns in a similar manner by selecting on the vertical ellipse adjacent to the column and clicking on **Remove**. Notice the STEPS panel on the right-hand side that provides a running list of the data operations.



4. For the **age** and **embarked** columns, click on the vertical ellipse adjacent to the columns, and click on **Remove empty rows**.



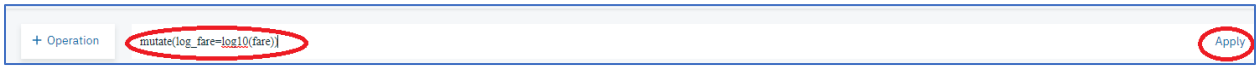
5. Convert the **fare** column from a String to a Decimal by clicking on the vertical ellipse adjacent to the column, click on **Convert Column**, and then click on **Decimal**.



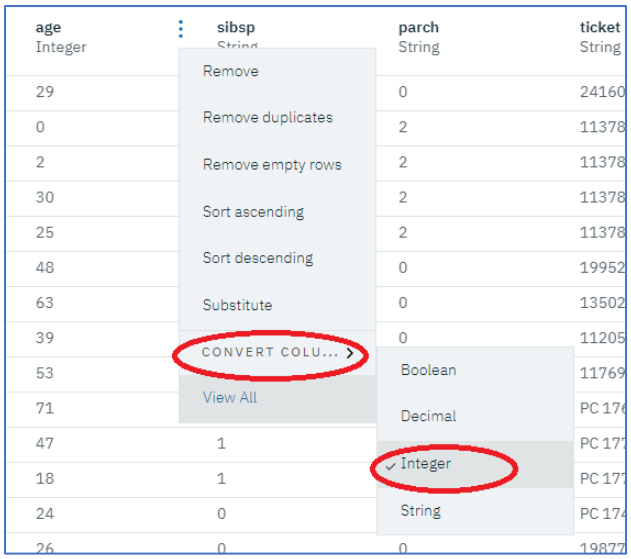
6. Create a new column that is the log to the base 10 of the fare by clicking into the **Code an operation to cleanse and shape your data,** and entering

mutate(log_fare=log10(fare))

then click **Apply**.



7. Convert the **age** from String to Integer by clicking on the vertical ellipse adjacent to the age column, clicking on **Convert Column**, and clicking on **Integer**.

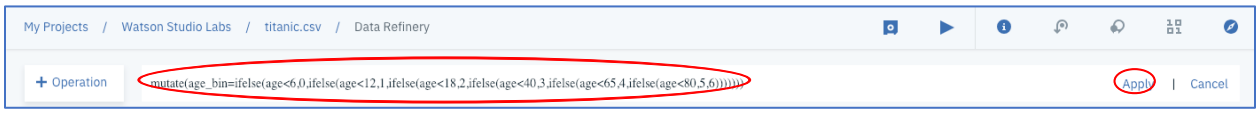


8. Bin the **age** column into the following bins by clicking into the **Code an operation to cleanse and shape your data,** and copying and pasting the following

   mutate(age_bin=ifelse(age<6,0,ifelse(age<12,1,ifelse(age<18,2,ifelse(age<40,3,ifelse(age<65,4,ifelse(age<80,5,6)))))))

   and then click **Apply**.

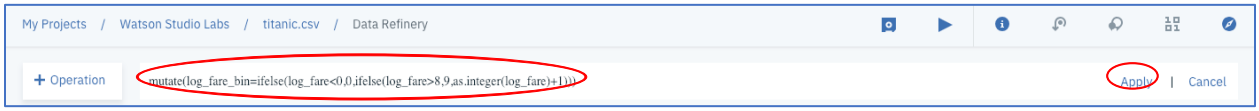
Bin	Age Range
0	0-5
1	6-11
2	12-17
3	18-39
4	40-64
5	65-79
6	Over 79



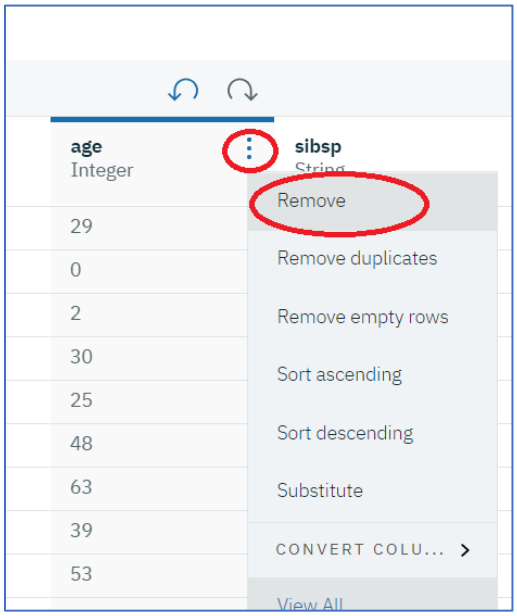
9. Bin the **log_fare** column, by clicking into the **Code an operation to cleanse and shape your data,** and copying and pasting the following

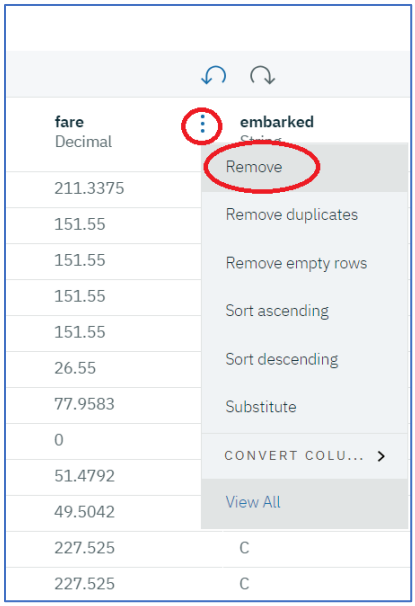
   mutate(log_fare_bin=ifelse(log_fare<0,0,ifelse(log_fare>8,9,as.integer(log_fare)+1)))

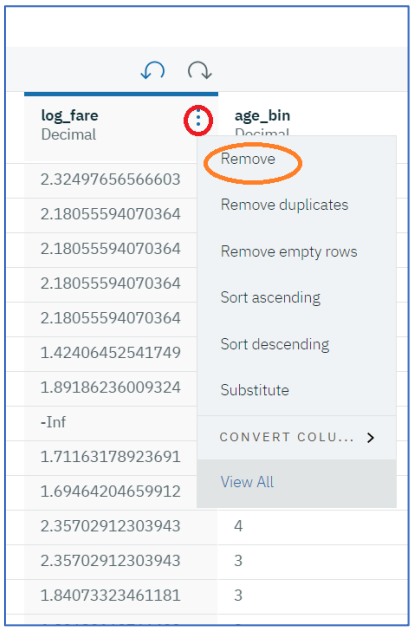
   and then clicking **Apply**



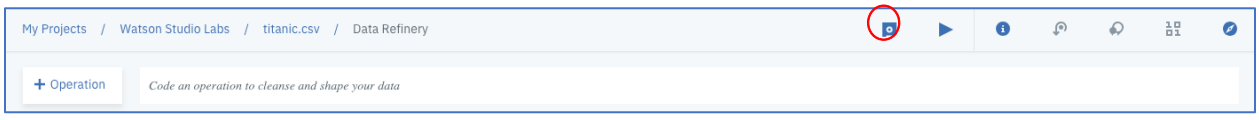
10. Now we will drop the **age**, **fare**, and **log_fare** columns as they are no longer needed for modeling purposes. Select the vertical ellipse adjacent to the column and click on **Remove** as shown below.







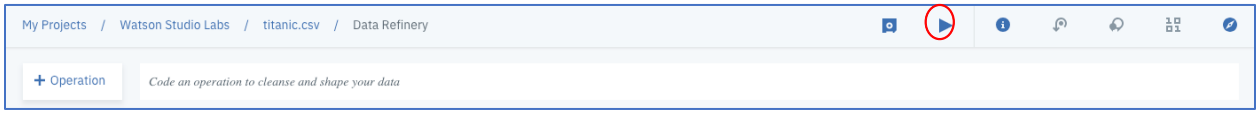
11. Save the Data Flow by clicking on the Save Data Flow icon .



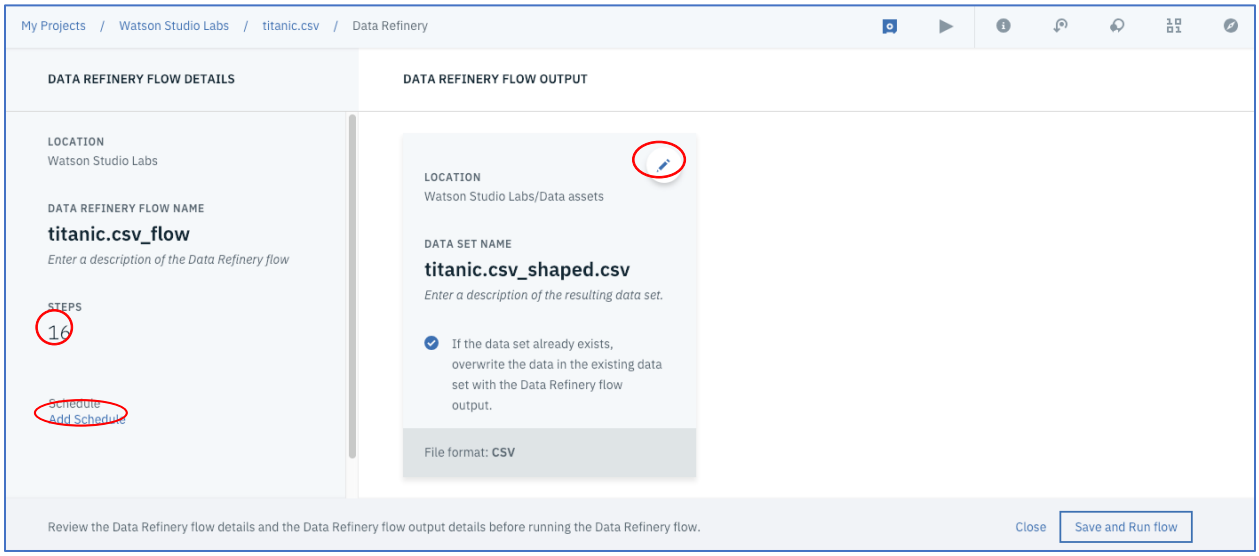
Step 5: Run the sequence of Data Flow operations on the entire data set.

When users are interacting with the Data Refinery tool, the operations are applied to a subset of the data set to facilitate faster response times. To run the data operations on the entire data set, the user selects the run option.

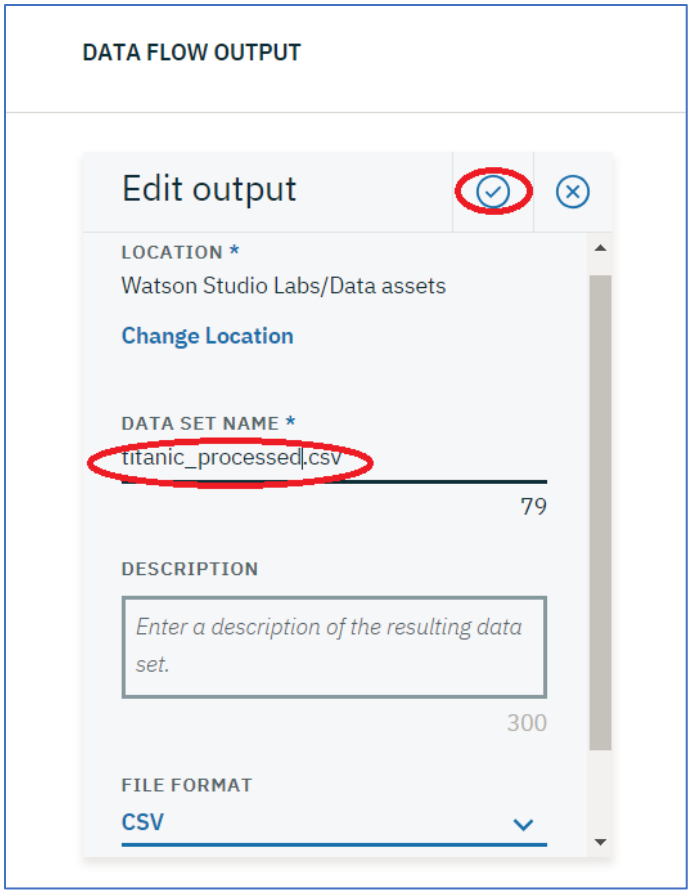
1. Click on run icon 



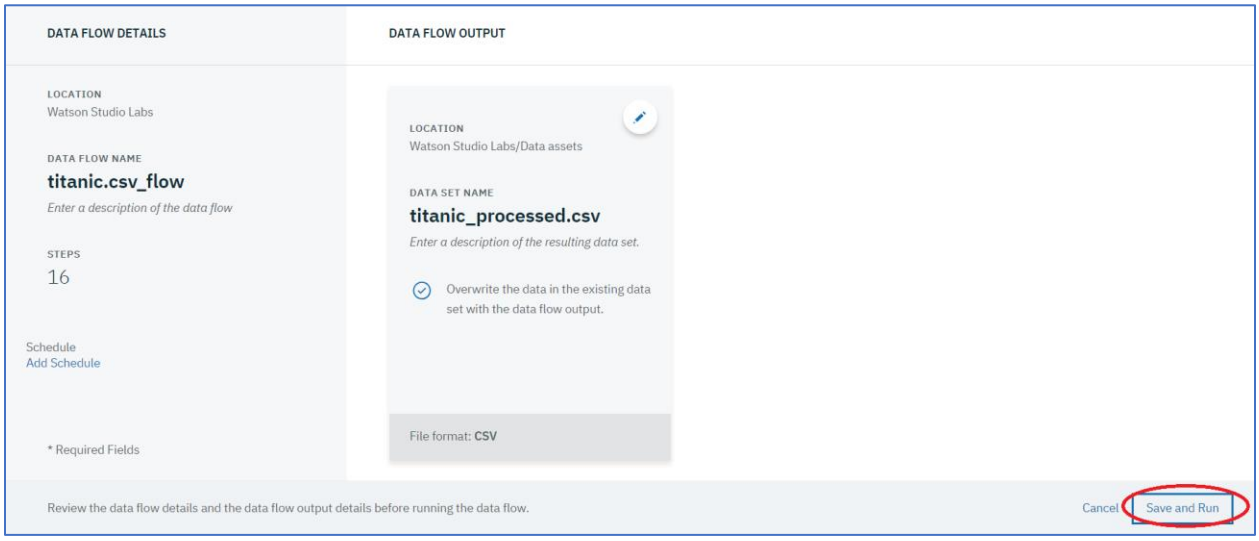
2. Note the number of steps used to transform the data. It should be 16. Also, a schedule can be set up if the transformation process needs to run on a scheduled basis. We are just going to do a one-time run. Change the name of the output file by clicking on the edit option (pencil icon).



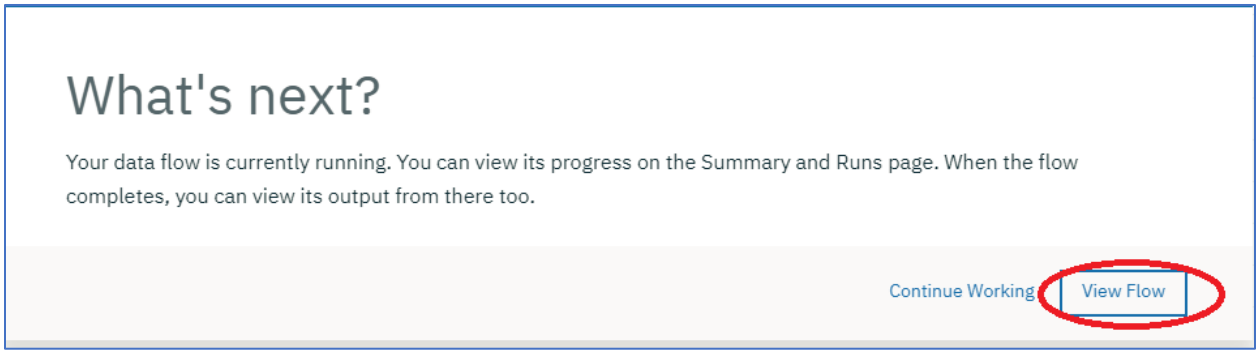
3. Type in **titanic_processed.csv** as the new file name, and click on the check mark.



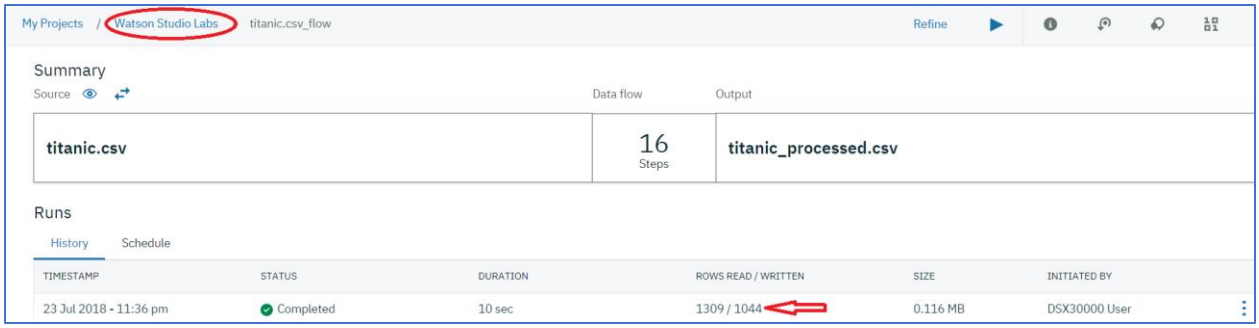
4. Click **Save and Run**.



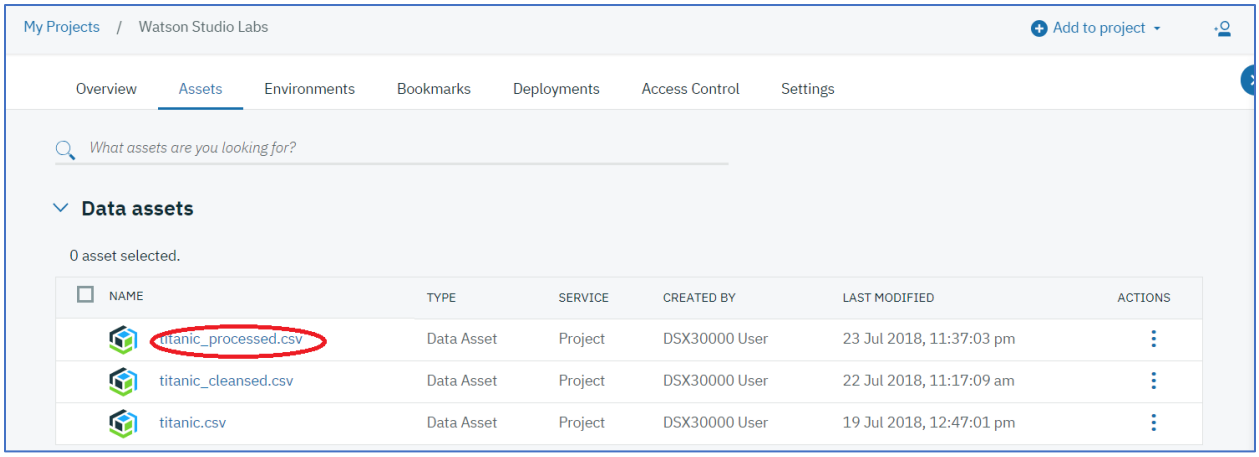
5. You can continue to work on other items or monitor the Data Flow run status. Click on **View Flow**.



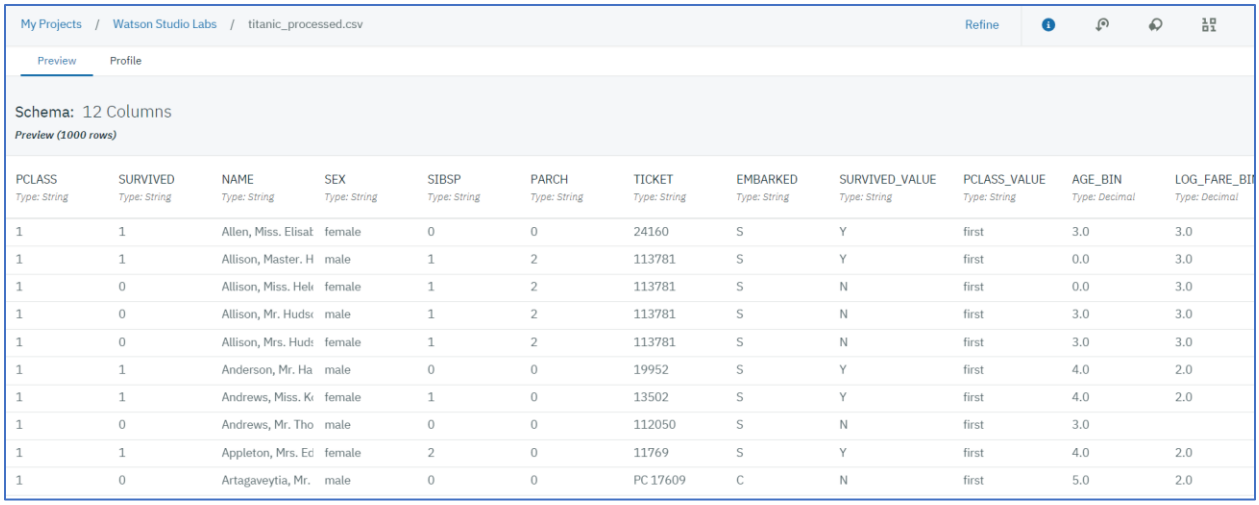
6. The completed flow is shown below. Note that 1044 records were written to the output file. Click on Watson Studio Labs to go back to the project Assets page.



7. The output of the Data Refinery process should be listed in the Data Assets. Click on the asset to view the contents.



8. The asset contents are displayed below. Review to confirm that the data transformations specified have been applied to all the data.



My Projects / Watson Studio Labs / titanic_processed.csv

Refine

Preview | Profile

Schema: 12 Columns

Preview (1000 rows)

PCLASS *Type: String*	SURVIVED *Type: String*	NAME *Type: String*	SEX *Type: String*	SIBSP *Type: String*	PARCH *Type: String*	TICKET *Type: String*	EMBARKED *Type: String*	SURVIVED_VALUE *Type: String*	PCLASS_VALUE *Type: String*	AGE_BIN *Type: Decimal*	LOG_FARE_BI *Type: Decimal*
1	1	Allen, Miss. Elisat	female	0	0	24160	S	Y	first	3.0	3.0
1	1	Allison, Master. H	male	1	2	113781	S	Y	first	0.0	3.0
1	0	Allison, Miss. Hel	female	1	2	113781	S	N	first	0.0	3.0
1	0	Allison, Mr. Hudsc	male	1	2	113781	S	N	first	3.0	3.0
1	0	Allison, Mrs. Hud	female	1	2	113781	S	N	first	3.0	3.0
1	1	Anderson, Mr. Ha	male	0	0	19952	S	Y	first	4.0	2.0
1	1	Andrews, Miss. K	female	1	0	13502	S	Y	first	4.0	2.0
1	0	Andrews, Mr. Tho	male	0	0	112050	S	N	first	3.0	
1	1	Appleton, Mrs. Ed	female	2	0	11769	S	Y	first	4.0	2.0
1	0	Artagaveytia, Mr.	male	0	0	PC 17609	C	N	first	5.0	2.0