

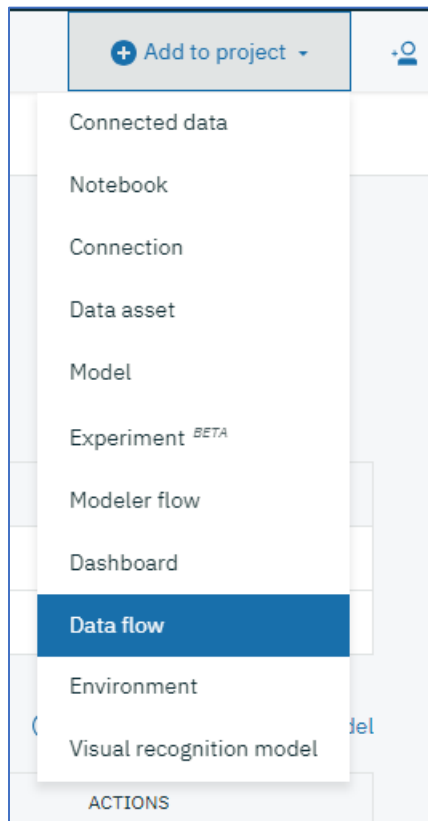
Data Refinery Lab

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:

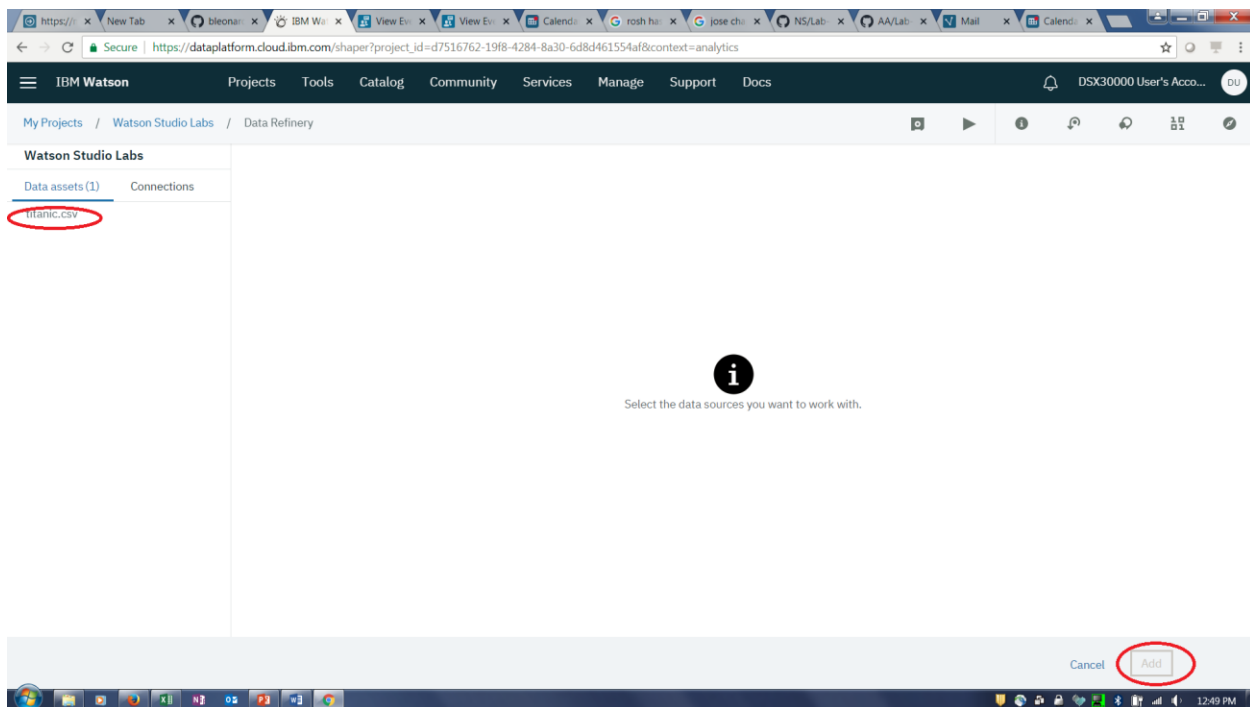
1. Use the Data Refinery Tool to:
 - a. Profile the data to help determine missing values
 - b. Visualize the data to gain a better understanding
 - c. Prepare the data for modeling
 - d. Run the sequence of data preparation operations on the entire data set.

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 Flow**.



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

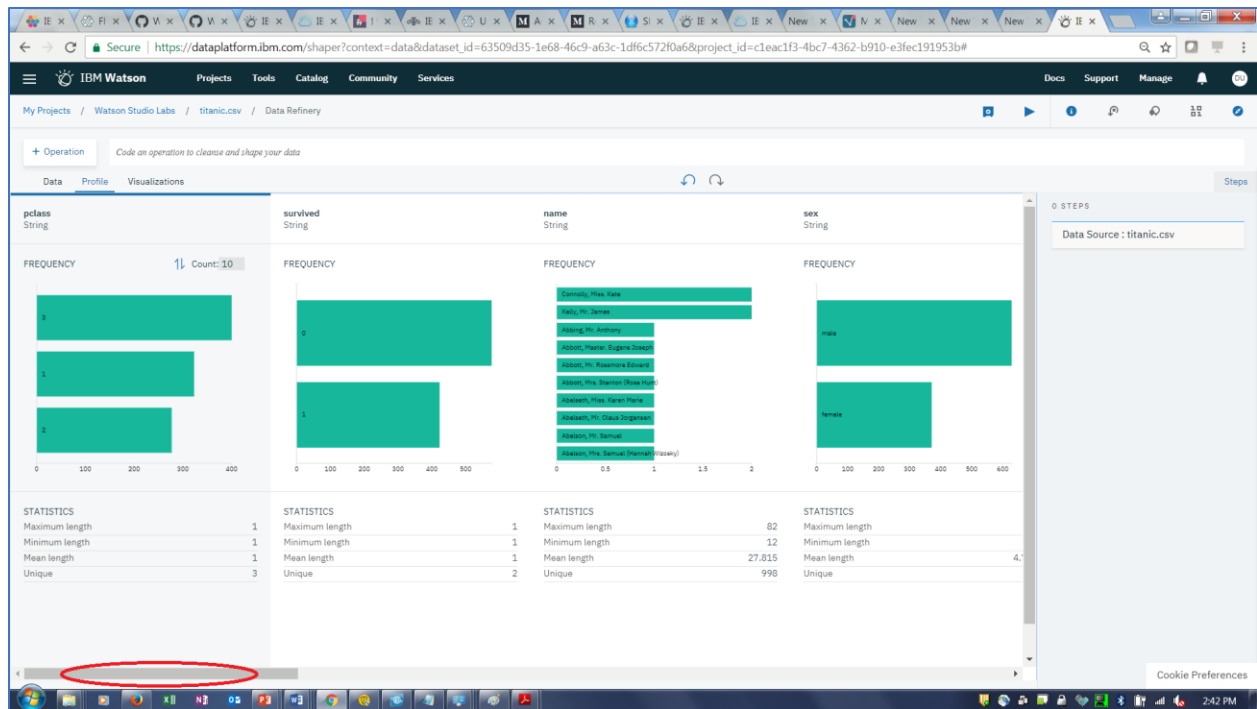


3. The Data Refinery panel will display the Titanic data set. Click on the **Profile** tab.

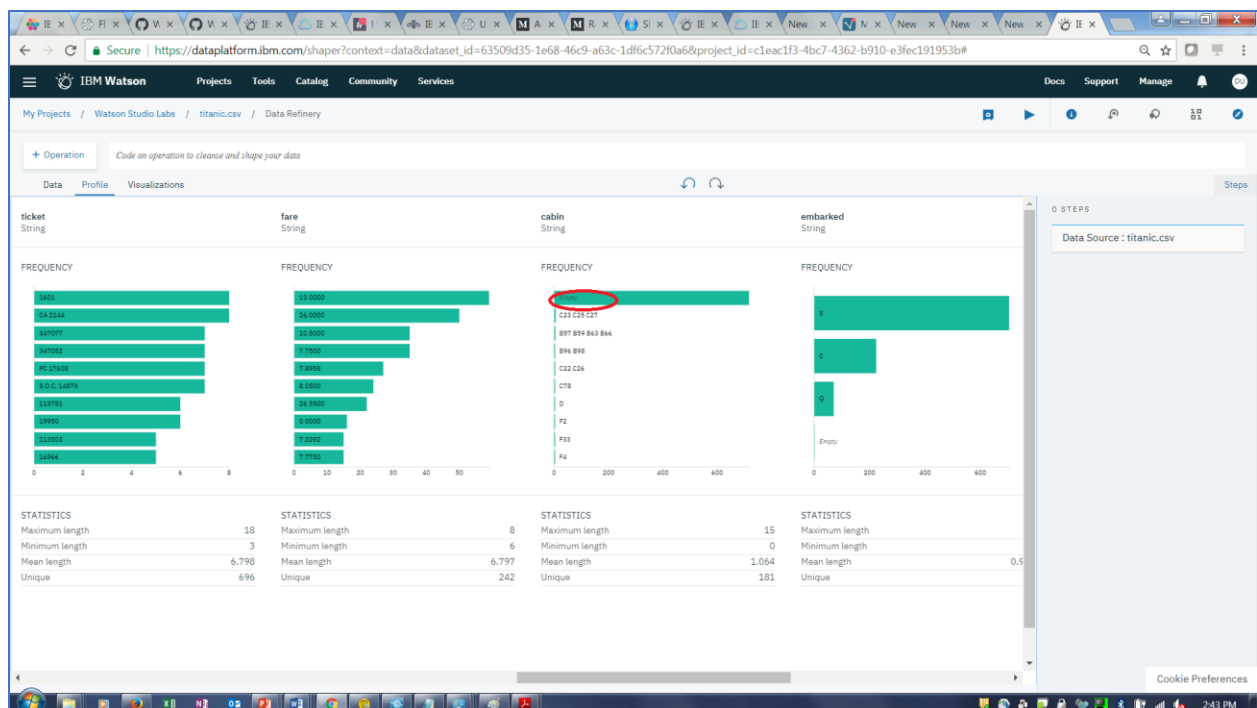
The screenshot shows the IBM Watson Data Platform interface with the 'Profile' tab selected. The table displays the top 10 count values for each column. The 'Data Source' is 'titanic.csv'. The table has columns for 'pclass', 'survived', 'name', 'sex', 'age', 'sibsp', 'parch', 'ticket', 'fare', and 'cabin'. The 'cabin' column is highlighted with a red circle. The bottom status bar shows the time as 2:33 PM.

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin
1	1	1	Allen, Miss. Elisabeth...	female	29	0	0	24160	211.3375	B1...
2	1	1	Allison, Master. Hud...	male	0.9167	1	2	113781	151.5500	C2...
3	1	0	Allison, Miss. Helen ...	female	2	1	2	113781	151.5500	C2...
4	1	0	Allison, Mr. Hudson ...	male	30	1	2	113781	151.5500	C2...
5	1	0	Allison, Mrs. Hudso...	female	25	1	2	113781	151.5500	C2...
6	1	1	Anderson, Mr. Harry	male	48	0	0	19952	26.5500	E1...
7	1	1	Andrews, Miss. Korn...	female	63	1	0	13502	77.9583	D1...
8	1	0	Andrews, Mr. Thom...	male	39	0	0	112050	0.0000	A3...
9	1	1	Appleton, Mrs. Edw...	female	53	2	0	11769	51.4792	C1...
10	1	0	Artagaveytia, Mr. Ra...	male	71	0	0	PC 17609	49.5042	
11	1	0	Astor, Col. John Jacob	male	47	1	0	PC 17757	227.5250	C6...
12	1	1	Astor, Mrs. John Jac...	female	18	1	0	PC 17757	227.5250	C6...
13	1	1	Aubart, Mme. LeontL...	female	24	0	0	PC 17477	69.3000	B1...
14	1	1	Barber, Miss. Ellen ...	female	26	0	0	19877	78.8500	
15	1	1	Barkworth, Mr. Alge...	male	80	0	0	27042	30.0000	A3...
16	1	0	Baumann, Mr. John D	male		0	0	PC 17318	25.9250	
17	1	0	Baxter, Mr. Quigg Ed...	male	24	0	1	PC 17558	247.5208	B1...
18	1	1	Baxter, Mrs. James (...)	female	50	0	1	PC 17558	247.5208	B1...
19	1	1	Bazzani, Miss. Albina	female	32	0	0	11813	76.2917	D1...
20	1	0	Beattie, Mr. Thomson	male	36	0	0	13050	75.2417	C6...

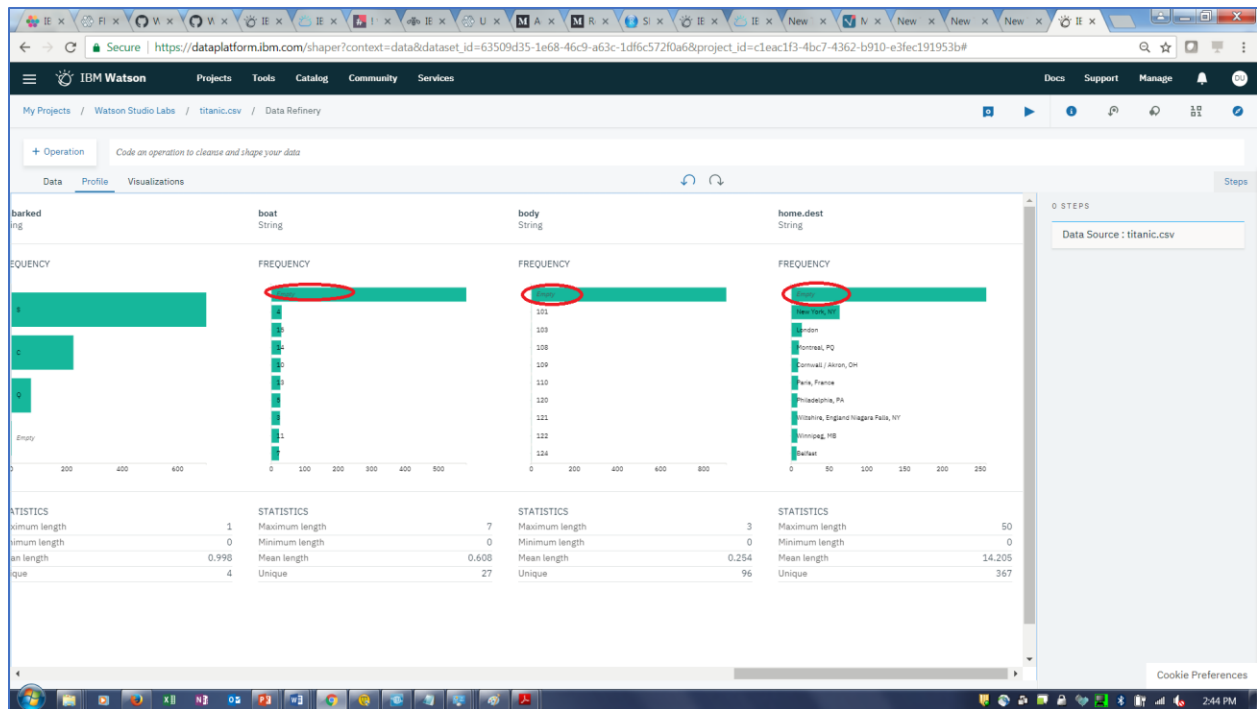
4. The Profile panel displays the counts of the top 10 count values for each column. Note that you can change 10 to another number if desired. You can also switch to the bottom 10 counts for a column. Scroll to the right to view the cabin column.



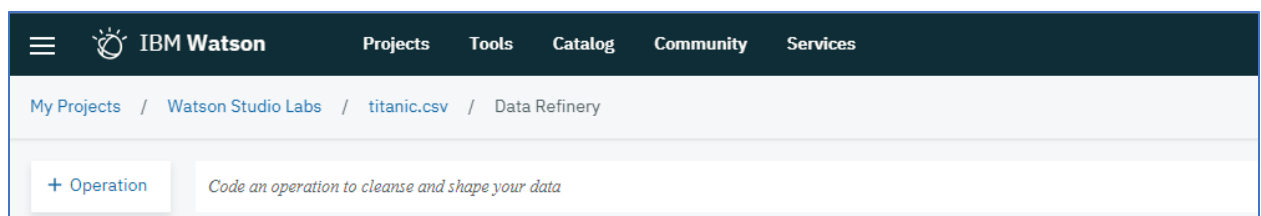
- Note that the cabin column has many missing values and should be removed as part of the data preparation step.



- 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.

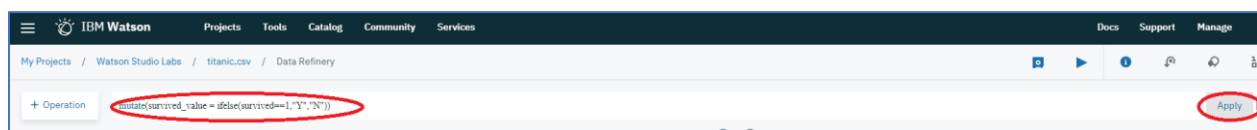


7. Age and Embarked also have missing values. Embarked has very few missing values. Age has over 100 missing values, but we will keep that column in the analysis. As part of data preparation, we will remove the rows that contain the missing age and embarked values.
8. Click on the **Data** tab. We will 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 dplyr function) and ifelse functions to do the conversion. Click on the **Code an operation to cleanse and shape your data**.

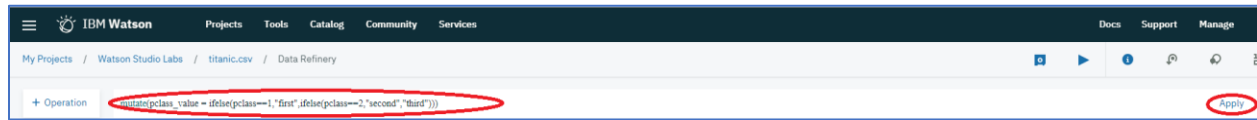


9. Type the following:
`mutate(survived_value=ifelse(survived==1, "Y", "N"))`

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



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



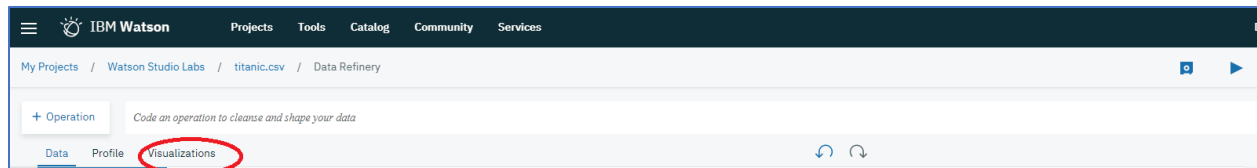
11. The result is shown below. Notice that the right panel will contain a running list of the transformations.

The screenshot shows the IBM Watson Data Refinery interface with the 'Data' tab selected. The table displays the results of the transformation, with columns for ticket, fare, cabin, embarked, boat, body, home.dest, survived_value, and pclass_value. The right panel shows the 'Steps' section with the 'Data Source' and 'Custom code'.

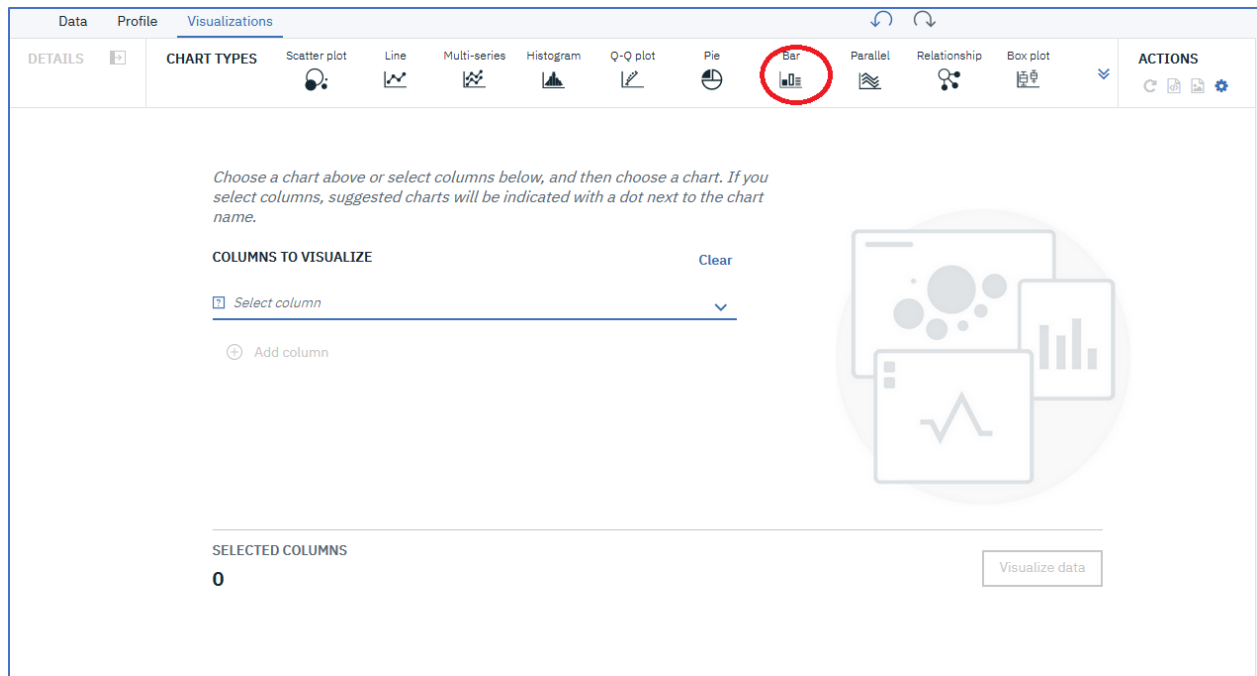
	ticket String	fare String	cabin String	embarked String	boat String	body String	home.dest String	survived_value String	pclass_value String
1	24160	211.3375	B5	S	2		St Louis, MO	Y	first
2	113781	151.5500	C22 C26	S	11		Montreal, PQ / Ches...	Y	first
3	113781	151.5500	C22 C26	S			Montreal, PQ / Ches...	N	first
4	113781	151.5500	C22 C26	S		135	Montreal, PQ / Ches...	N	first
5	113781	151.5500	C22 C26	S			Montreal, PQ / Ches...	N	first
6	19952	26.5500	E12	S	3		New York, NY	Y	first
7	13502	77.9583	D7	S	10		Hudson, NY	Y	first
8	112050	0.0000	A36	S			Belfast, NI	N	first
9	11769	51.4792	C101	S	D		Bayside, Queens, NY	Y	first
10	PC 17609	49.5042		C		22	Montevideo, Uruguay	N	first
11	PC 17757	227.5250	C62 C64	C		124	New York, NY	N	first
12	PC 17757	227.5250	C62 C64	C	4		New York, NY	Y	first

Step 3: Visualize the data to get a better understanding

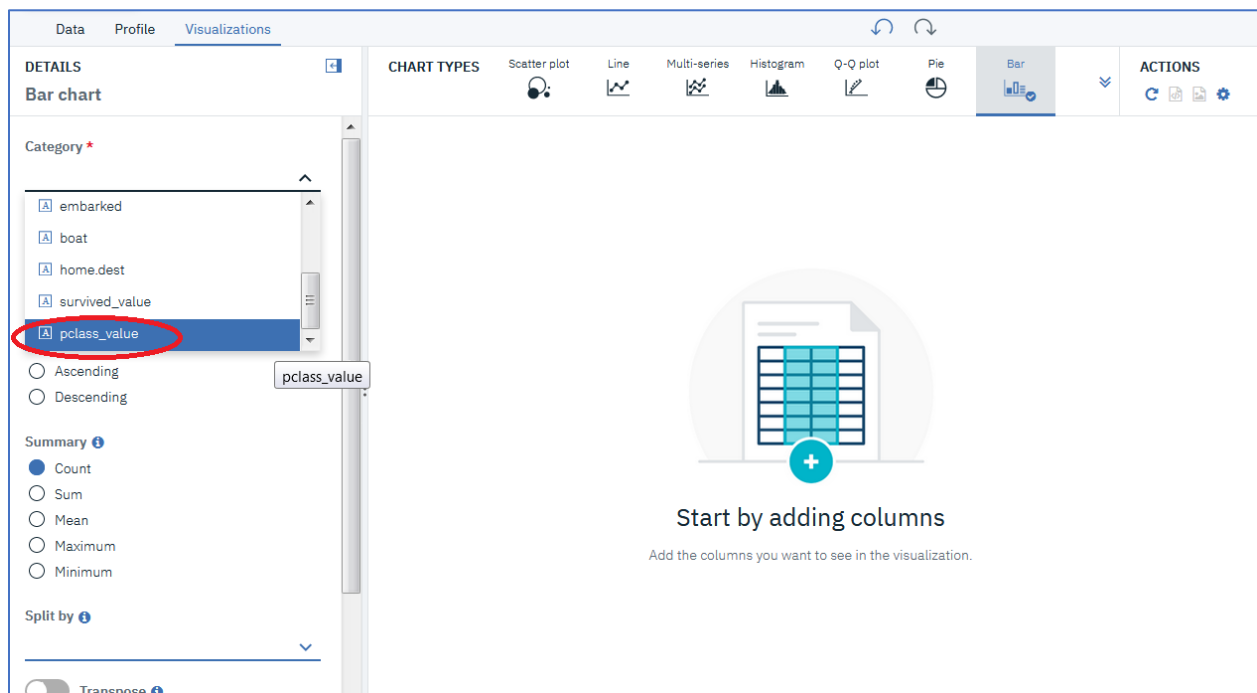
1. Click on the **Visualizations** tab.



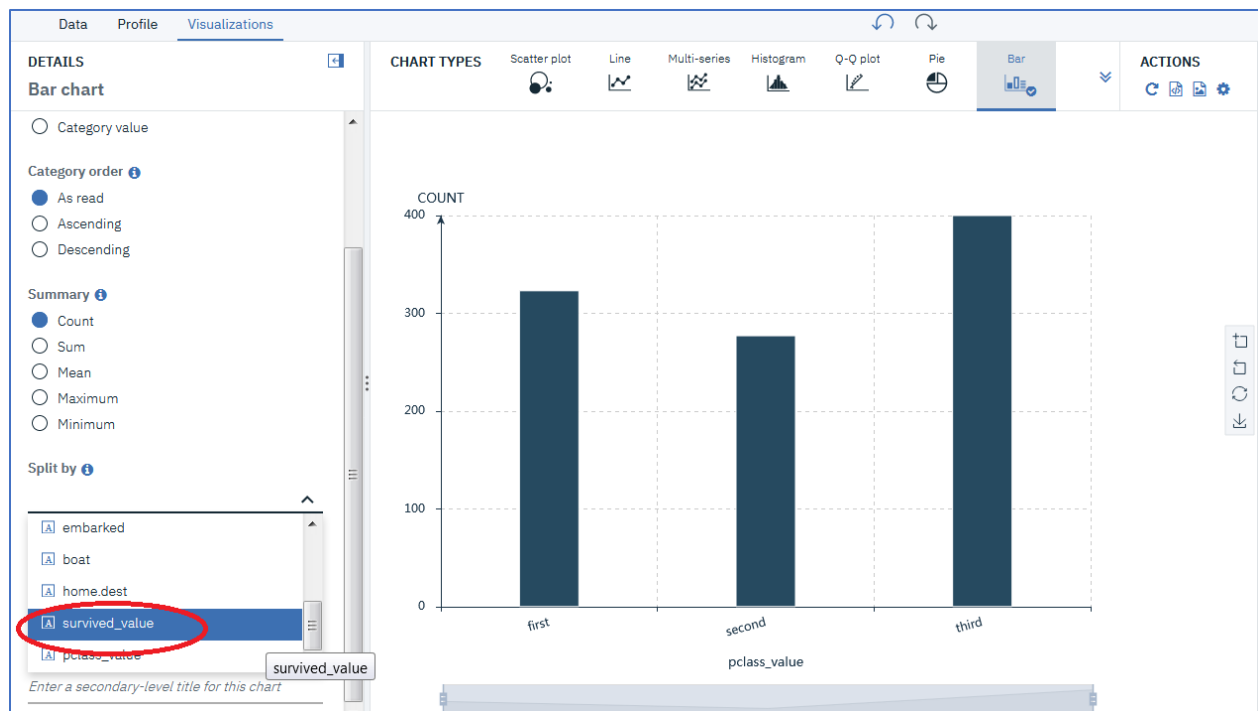
2. Let's take a look at the breakdown of passengers by passenger class. We will use our new pclass_value field. Select the **Bar Chart** Type.



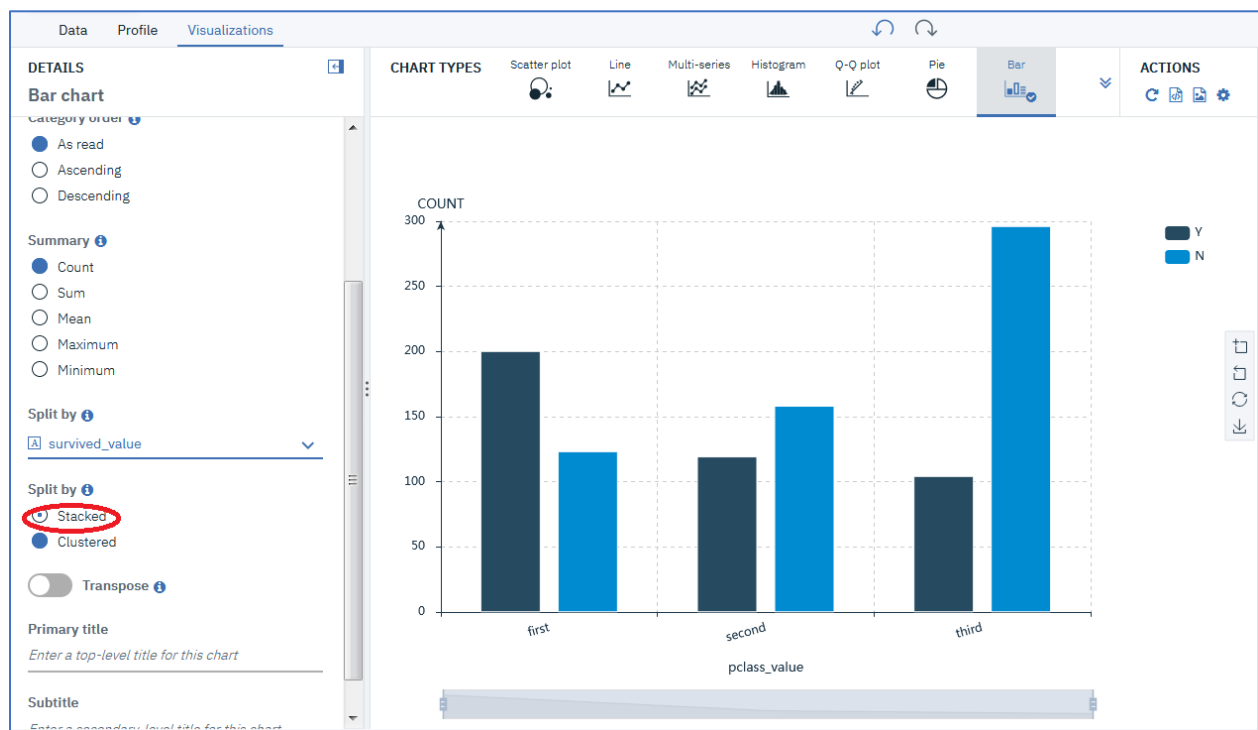
3. In the **Category** required field, select **pclass_value**.



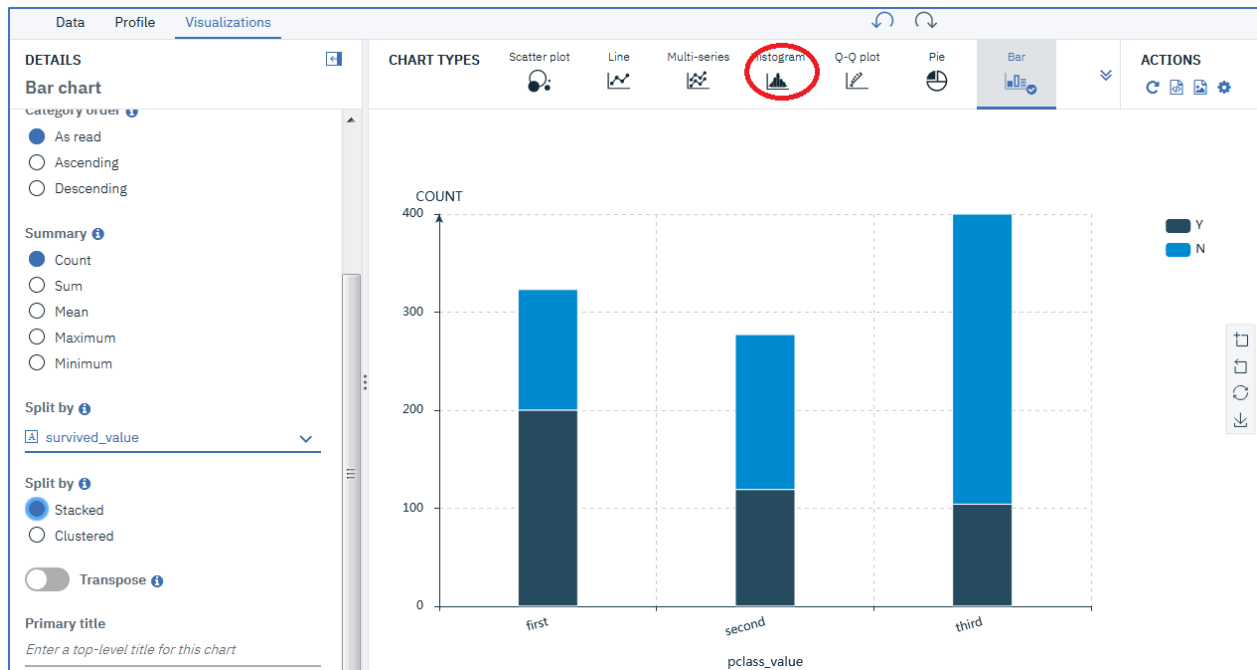
4. In the **Split by** field, select **survived_value**.



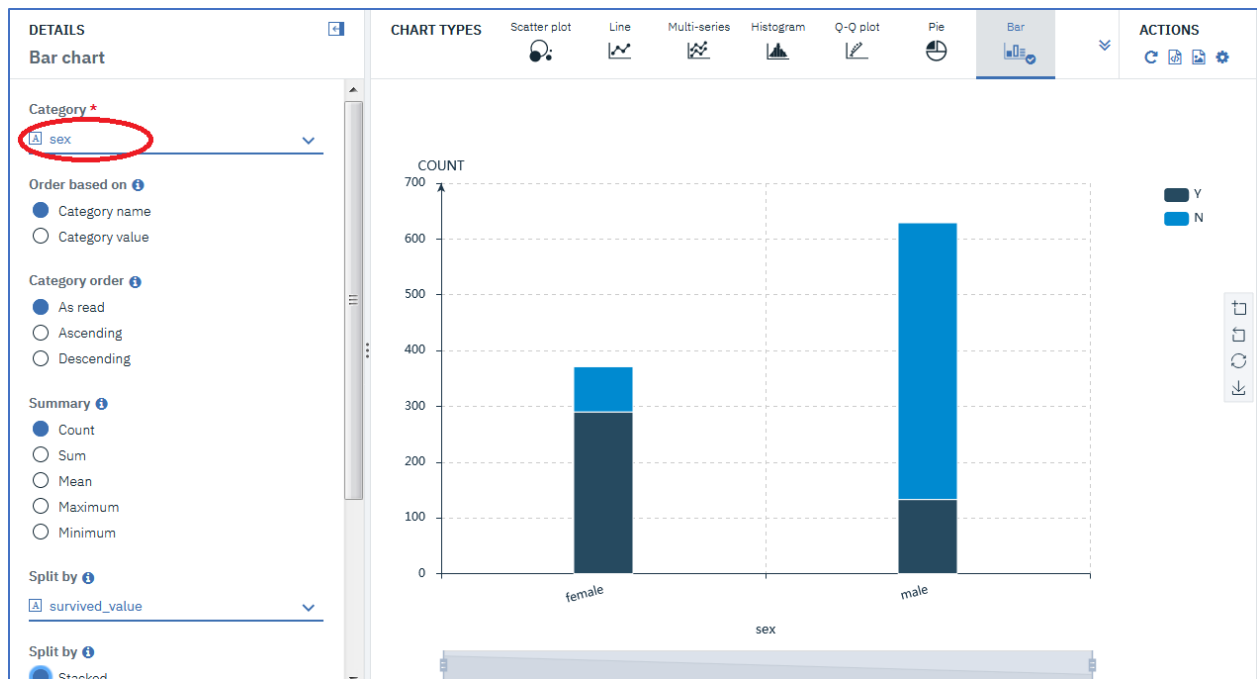
5. Select **Stacked**.



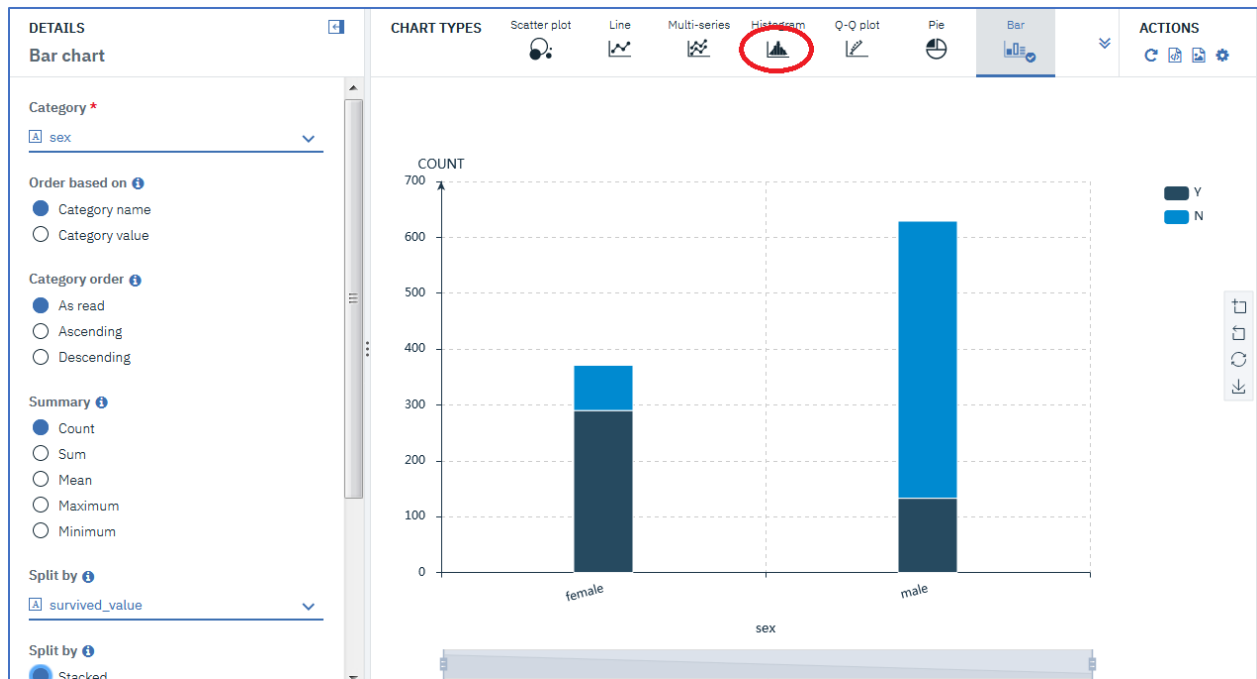
6. The result is shown below. The percentage of survivors is the greatest in first class, followed by second class, and then third-class passengers.



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**.

Switch charts?

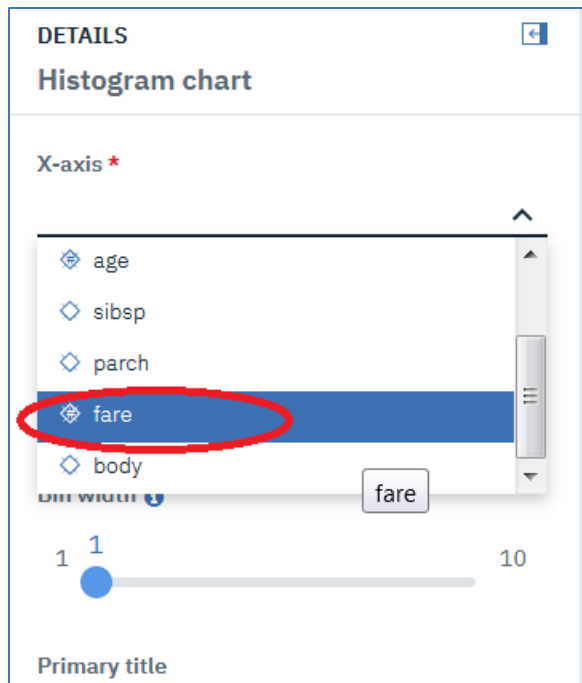
You might lose this chart's details or you might have to provide more details to view another chart.

☐ Don't show this again

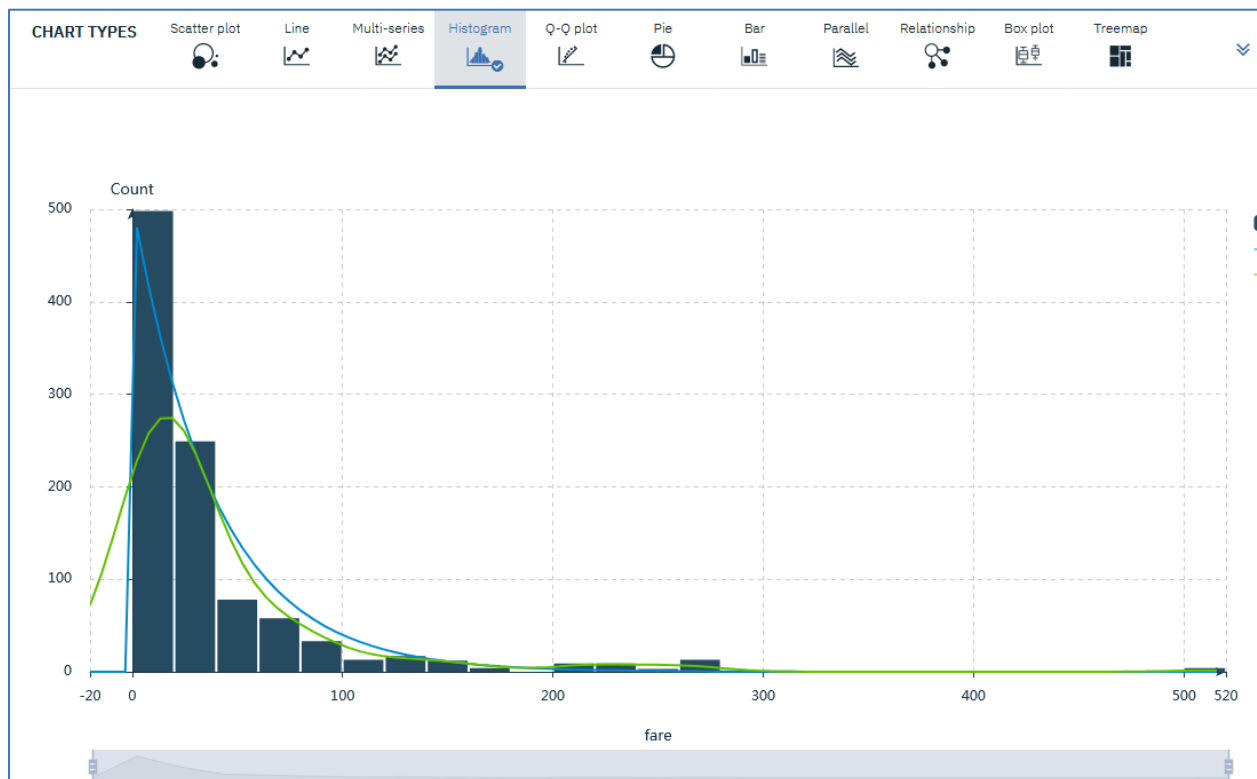
Cancel Continue

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.



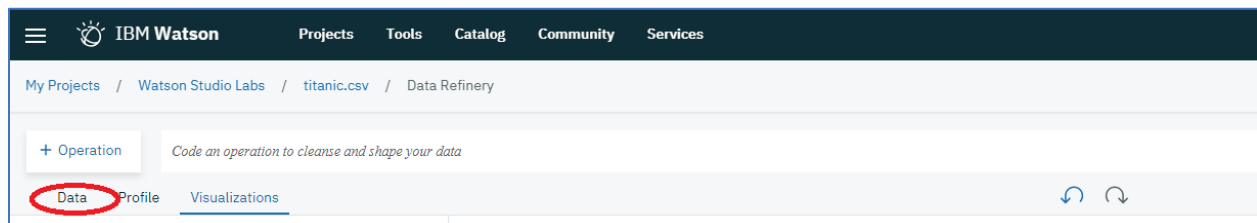
Step 4: Prepare the data for modeling

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**.

The screenshot shows the IBM Watson Data Refinery interface. The top navigation bar includes 'My Projects', 'Watson Studio Labs', 'titanic.csv', and 'Data Refinery'. Below this, there's a '+ Operation' button and a text input field. The 'Data' tab is highlighted with a red circle, and the 'Profile' and 'Visualizations' tabs are also visible. The 'Data' tab is currently active, showing a table of data.

cabin	embarked	boat
String	String	String
B5		2
C22 C26		11
C22 C26		
C22 C26		
C22 C26		
E12		3
D7		10
A36		
C101		D
C62 C64		
C62 C64	C	4
B35	C	9
	S	6

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.

6 STEPS

Data Source : titanic.csv

Custom code

```
mutate(survived_value =  
ifelse(survived==1,"Y","N"))
```

Custom code

```
mutate(pclass_value =  
ifelse(pclass==1,"first",ifelse(pclass==  
2,"second","third")))
```

Remove

Removed cabin

Remove

Removed boat

Remove

Removed body

Remove JUST ADDED

Removed home.dest

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

embarked	survived_value	pclas
String	String	String
S		first
S		first
S		first
S		first
S		first
S		first
S		first
S		first
S		first
C		first
C		first
C	Y	first
C	Y	first
S	Y	first

- 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**.

fare	embarked	survive	6 STEPS
String	String	String	
211.3375		Y	Data Sour
151.5500		Y	Custom co
151.5500		N	mutate(sur
151.5500		N	ifelse(survi
151.5500		N	
26.5500		Y	Custom co
77.9583		Y	mutate(pcl
0.0000		N	ifelse(pclas
51.4792			d",
49.5042			a
227.5250			d c
227.5250			d c
69.3000			a
78.8500			d b
30.0000			

- 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**.

+ Operation	mutate(log_fare=log10(fare))	Apply
-------------	------------------------------	-------

- 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**.

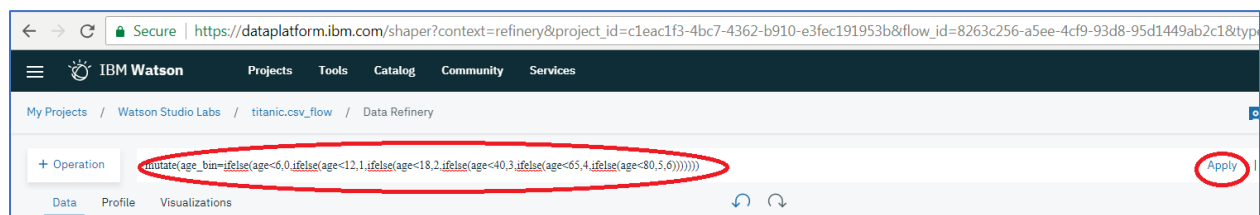
age	sibsp	parch	ticket
Integer	String	String	String
29		0	24160
0		2	11378
2		2	11378
30		2	11378
25		2	11378
48		0	19952
63		0	13502
39		0	11205
53			11769
71			PC 174
47	1		PC 177
18	1		PC 177
24	0		PC 174
26	0	0	19877

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

```
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**.

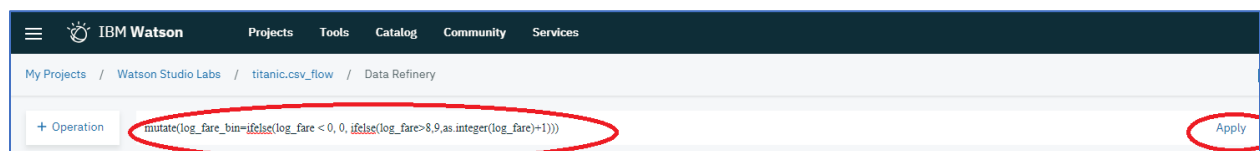
Bin	Age Range
0	0-5
1	6-11
2	12-17
3	18-39
4	40-64
5	65-79
6	Over 79



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

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

and then clicking **Apply**




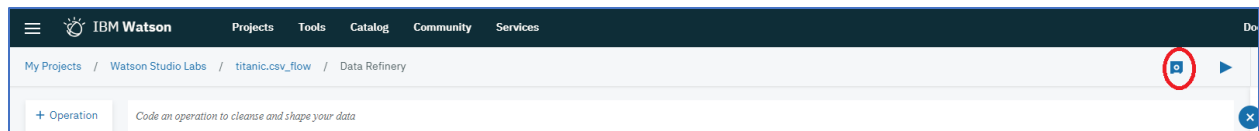
10. Now we will drop the age, fare, and log_fare columns as they are no longer needed for modeling purposes. Select the vertical ellipse adjacent to the column and click on **Remove** as shown below.

age	sibsp		
Integer	String		
29		Remove	
0		Remove duplicates	
2		Remove empty rows	
30		Sort ascending	
25		Sort descending	
48		Substitute	
63		CONVERT COLU... >	
39			
53			
		View All	

fare	embarked		
Decimal	String		
211.3375		Remove	
151.55		Remove duplicates	
151.55		Remove empty rows	
151.55		Sort ascending	
151.55		Sort descending	
26.55		Substitute	
77.9583		CONVERT COLU... >	
0			
51.4792		View All	
49.5042			
227.525	C		
227.525	C		


log_fare	age_bin
Decimal	Decimal
2.32497656566603	
2.18055594070364	
2.18055594070364	
2.18055594070364	
2.18055594070364	
1.42406452541749	
1.89186236009324	
-Inf	
1.71163178923691	
1.69464204659912	
2.35702912303943	4
2.35702912303943	3
1.84073323461181	3

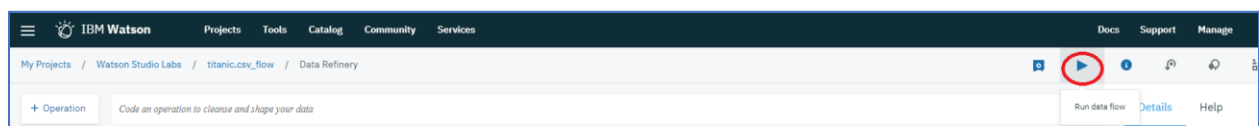
11. Save the Data Flow by clicking on the Save Data Flow icon .



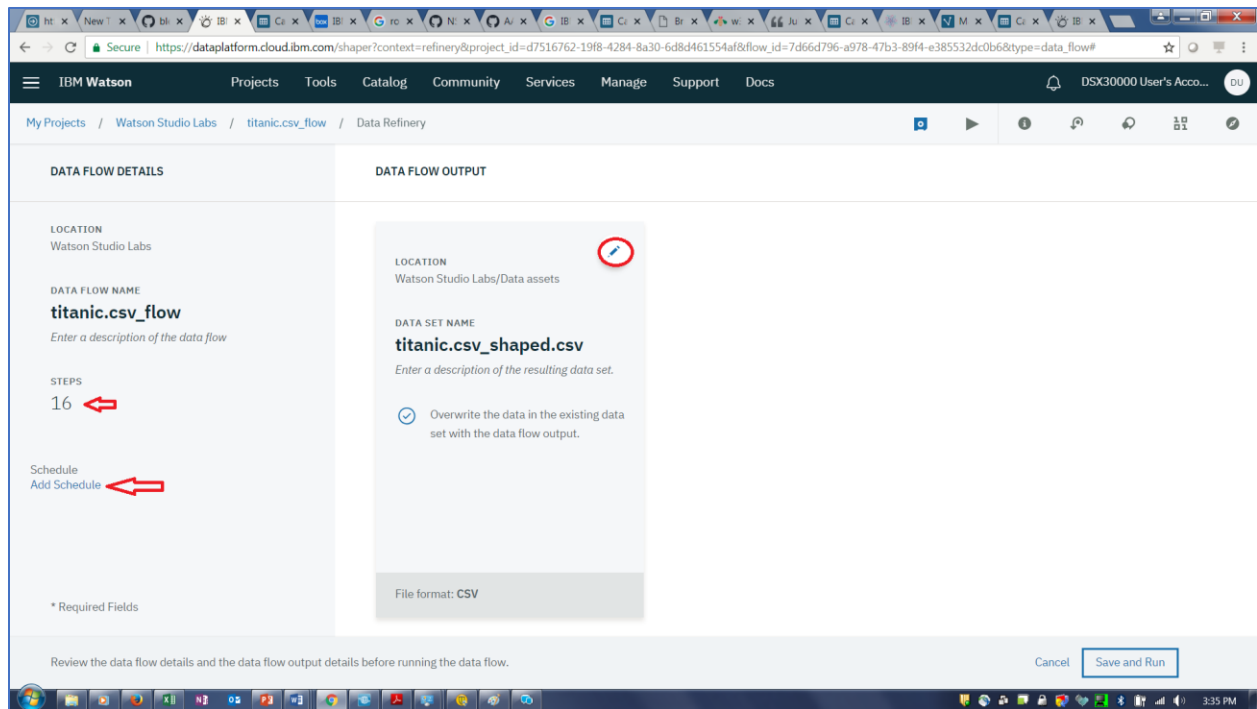
Step 5: Run the sequence of Data Flow operations on the entire data set.

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.

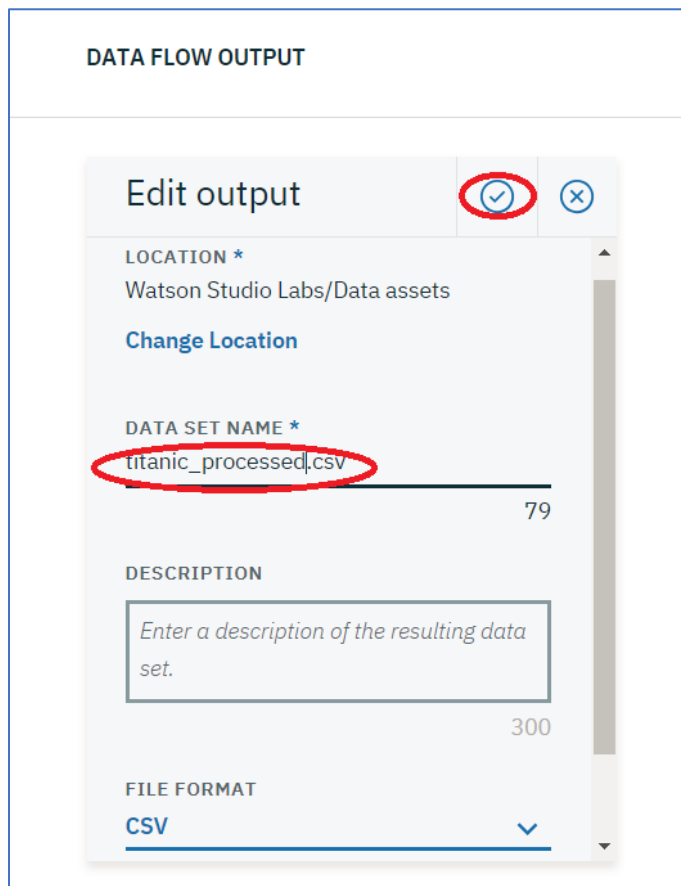
1. Click on run icon 



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**.

DATA FLOW DETAILS

LOCATION
Watson Studio Labs

DATA FLOW NAME
titanic.csv_flow
Enter a description of the data flow

STEPS
16

Schedule
Add Schedule

* Required Fields

DATA FLOW OUTPUT

LOCATION
Watson Studio Labs/Data assets

DATA SET NAME
titanic_processed.csv
Enter a description of the resulting data set.

☒ Overwrite the data in the existing data set with the data flow output.

File format: **CSV**

Review the data flow details and the data flow output details before running the data flow.

Cancel **Save and Run**

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

What's next?

Your data flow is currently running. You can view its progress on the Summary and Runs page. When the flow completes, you can view its output from there too.

[Continue Working](#) **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.

My Projects / **Watson Studio Labs** / titanic.csv_flow

Refine ▶ ⓘ 🔊 🔇 10 01

Summary

Source ⓘ ↻

	Data flow	Output
titanic.csv	16 Steps	titanic_processed.csv

Runs

History Schedule

TIMESTAMP	STATUS	DURATION	ROWS READ / WRITTEN	SIZE	INITIATED BY
23 Jul 2018 - 11:36 pm	Completed	10 sec	1309 / 1044	0.116 MB	DSX30000 User

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

My Projects / Watson Studio Labs + Add to project 👤

Overview **Assets** Environments Bookmarks Deployments Access Control Settings

🔍 What assets are you looking for?

▼ **Data assets**

0 asset selected.

<input type="checkbox"/>	NAME	TYPE	SERVICE	CREATED BY	LAST MODIFIED	ACTIONS
	titanic_processed.csv	Data Asset	Project	DSX30000 User	23 Jul 2018, 11:37:03 pm	⋮
	titanic_cleansed.csv	Data Asset	Project	DSX30000 User	22 Jul 2018, 11:17:09 am	⋮
	titanic.csv	Data Asset	Project	DSX30000 User	19 Jul 2018, 12:47:01 pm	⋮

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 <small>Type: String</small>	SURVIVED <small>Type: String</small>	NAME <small>Type: String</small>	SEX <small>Type: String</small>	SIBSP <small>Type: String</small>	PARCH <small>Type: String</small>	TICKET <small>Type: String</small>	EMBARKED <small>Type: String</small>	SURVIVED_VALUE <small>Type: String</small>	PCLASS_VALUE <small>Type: String</small>	AGE_BIN <small>Type: Decimal</small>	LOG_FARE_BIN <small>Type: Decimal</small>
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. Helr	female	1	2	113781	S	N	first	0.0	3.0
1	0	Allison, Mr. Hudsr	male	1	2	113781	S	N	first	3.0	3.0
1	0	Allison, Mrs. Hudr	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. Ki	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