

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:

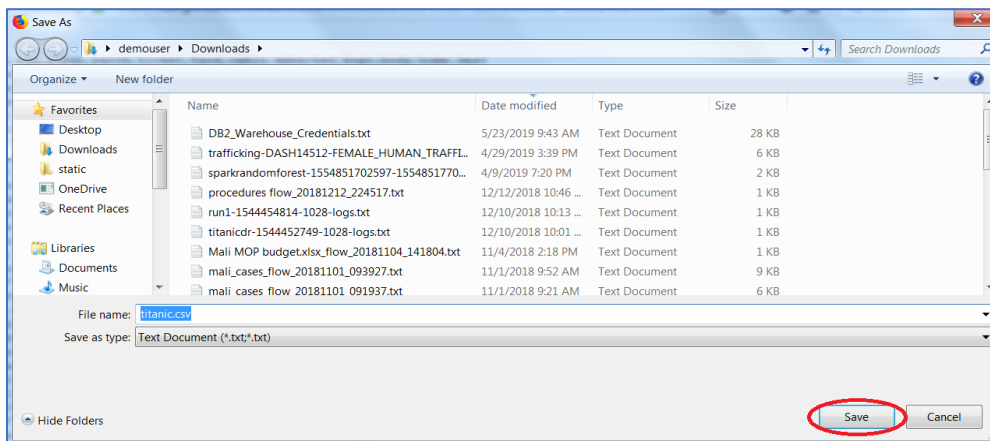
1. Use the Data Refinery Tool to:
 - a. Profile the data to help determine missing values
 - b. Visualize the data to gain a better understanding
 - c. Prepare the data for modeling
 - d. Run the sequence of data preparation operations on the entire data set.


Step 1: Adding a Data Asset to the Watson Studio Labs project

1. Download the Titanic data file from the following location by clicking [here](#).
2. Right-click on the screen and click on Save Page As ...



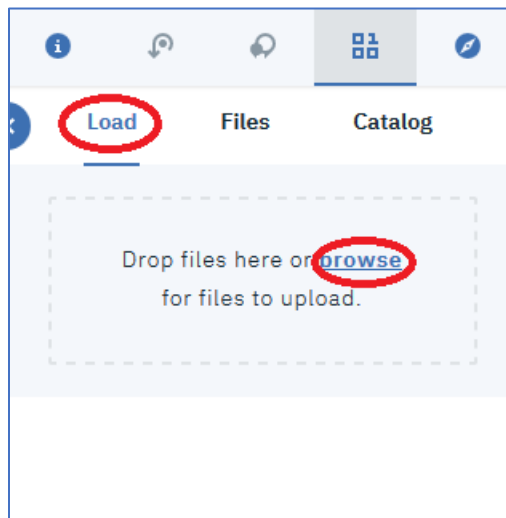
3. Click on **Save** to save the titanic.csv file (Note, if the file shown is titanic.csv.txt, remove the .txt).



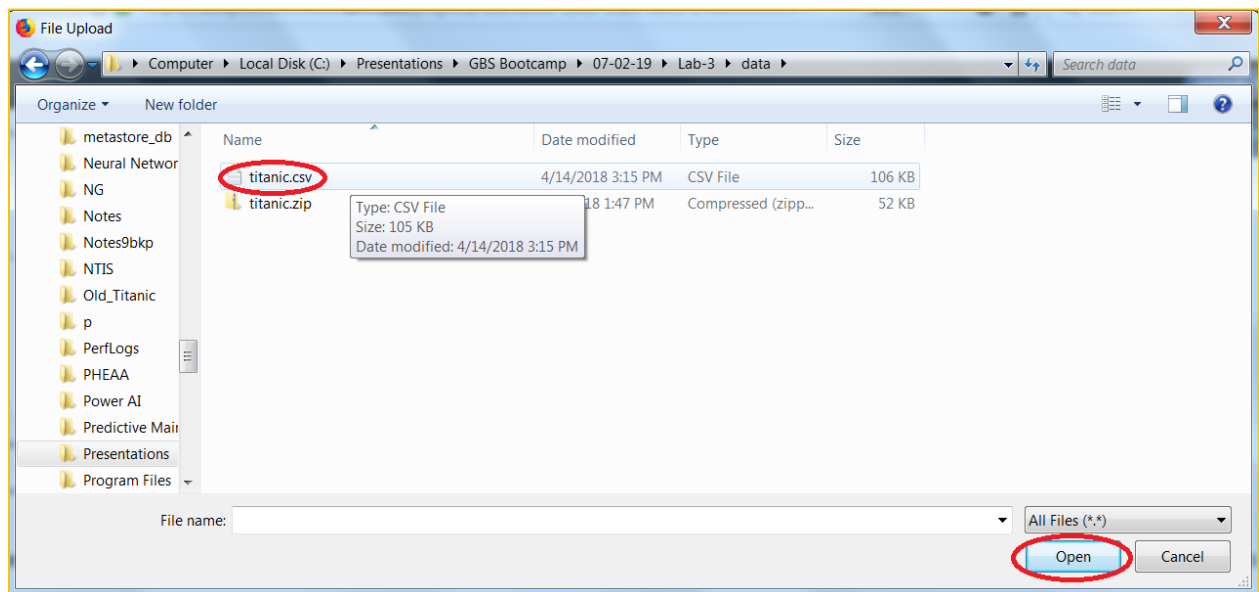
4. Go back to your Watson Studio Labs project. Click on the  icon.



5. Click on the **Load** tab and then click on **browse**. If you don't see the **Load** tab, click on the  icon again.



6. Go to the folder where the titanic_csv file is stored. Select the titanic.csv file and then click **Open**.

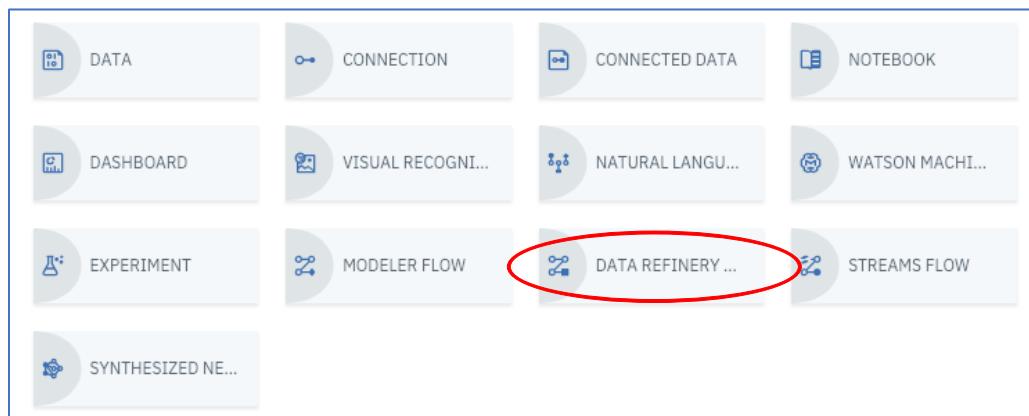
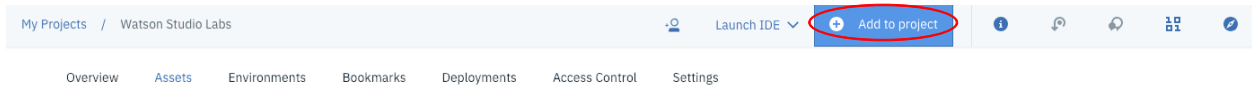


7. The file is now added as a Data Asset.

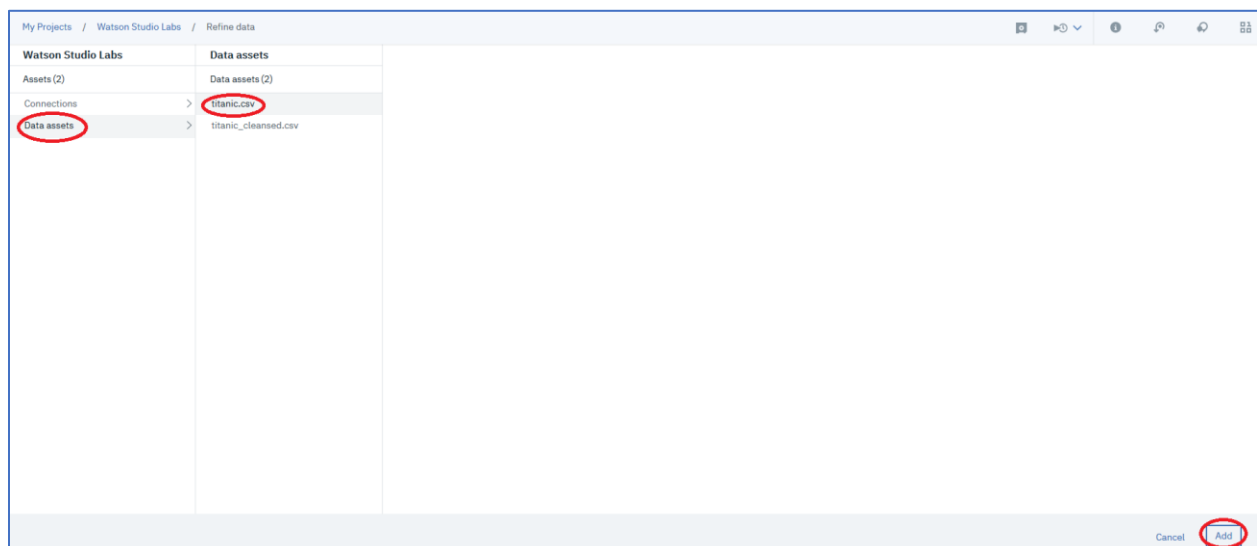
Data assets					New data asset
0 asset selected.					
<input type="checkbox"/>	NAME	TYPE	CREATED BY	LAST MODIFIED	ACTIONS
<input type="checkbox"/>	titanic.csv	Data Asset	John Doe	4 Nov 2018, 2:45:59 pm	

Step 2: 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.

My Projects / Watson Studio Labs / titanic.csv / Data Refinery

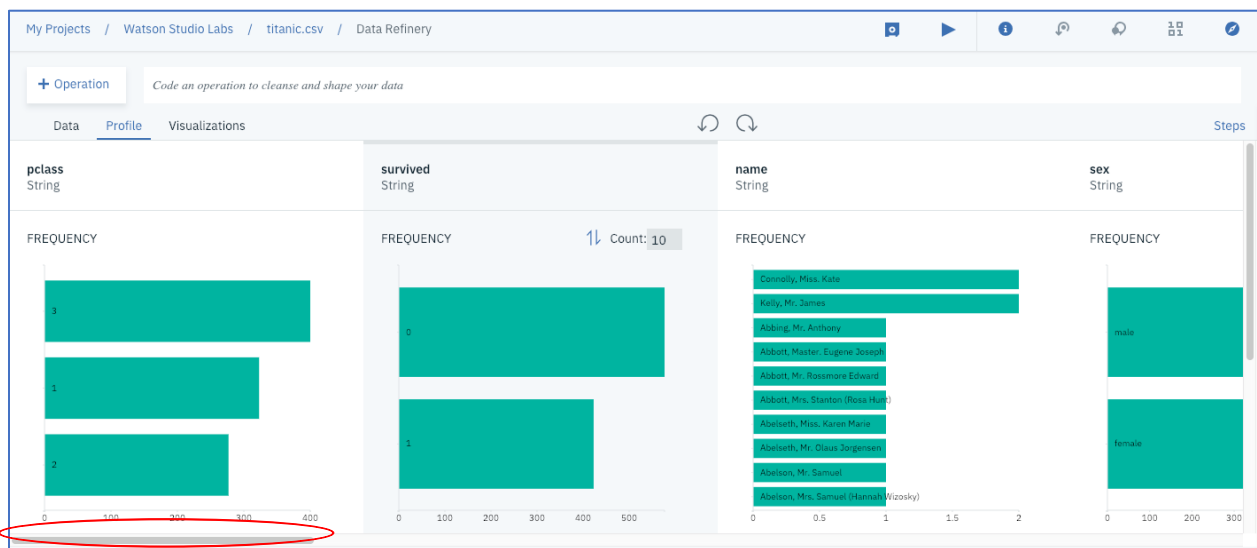
+ Operation Code an operation to cleanse and shape your data

Data **Profile** Visualizations Steps

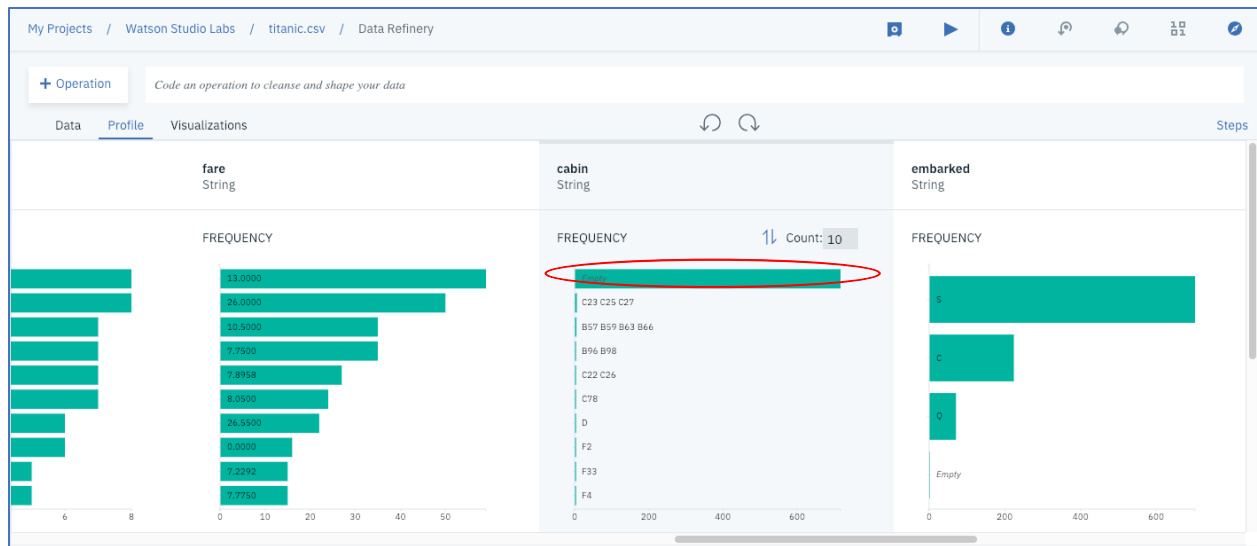
	pclass String	survived String	name String	sex String	age String	sibsp String
1	1	1	Allen, Miss. Elisabeth Walton	female	29	0
2	1	1	Allison, Master. Hudson Trevor	male	0.9167	1
3	1	0	Allison, Miss. Helen Loraine	female	2	1
4	1	0	Allison, Mr. Hudson Joshua Creighton	male	30	1
5	1	0	Allison, Mrs. Hudson 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

SOURCE FILE: titanic.csv SAMPLE SIZE: First 1000 rows

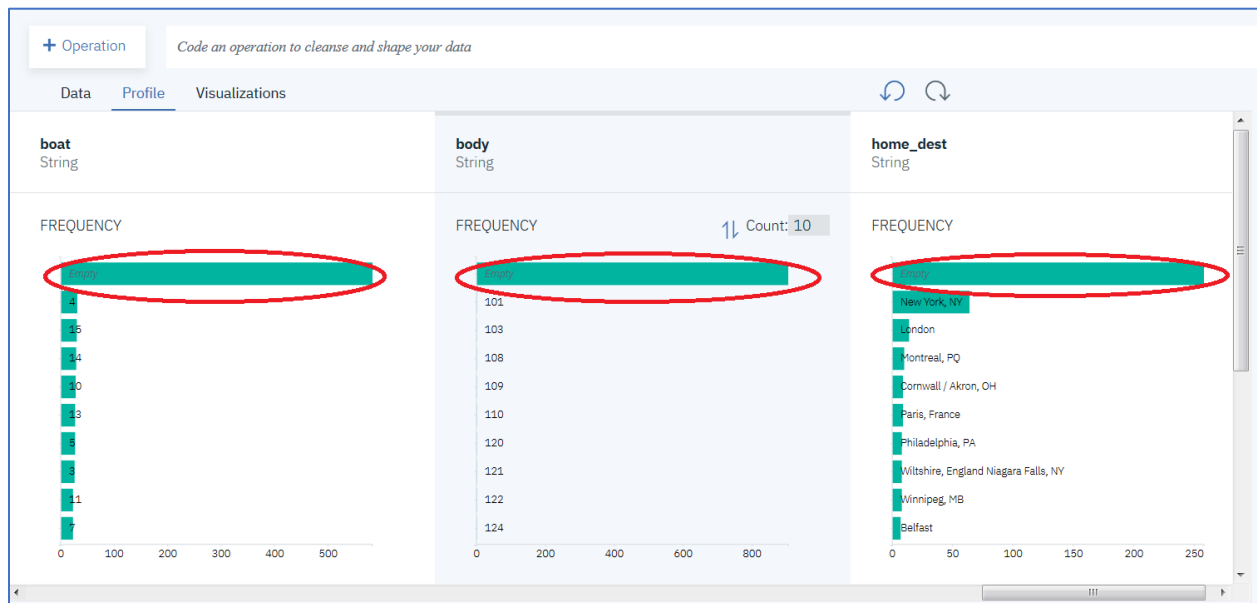
4. The Profile panel displays the counts of the top 10 count values for each column. Note that you can change 10 to another number if desired. You can also switch to the bottom 10 counts for a column. Scroll to the right to view the cabin column.



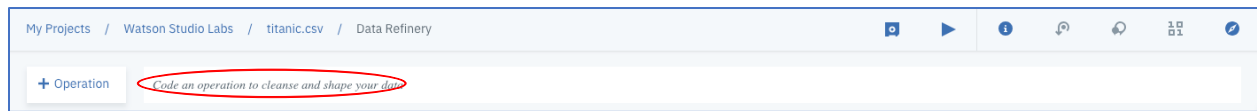
5. Note that the cabin column has many missing values and should be removed as part of the data preparation step.



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.

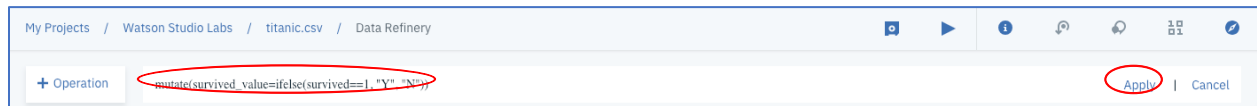


7. Age and Embarked also have missing values. Embarked has very few missing values. Age has over 100 missing values, but we will keep that column in the analysis. As part of data preparation, we will remove the rows that contain the missing age and embarked values.
8. Click on the **Data** tab. We will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

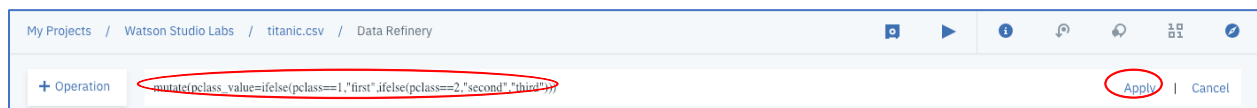


- Copy and paste the following:
`mutate(survived_value=ifelse(survived==1, "Y", "N"))`

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



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

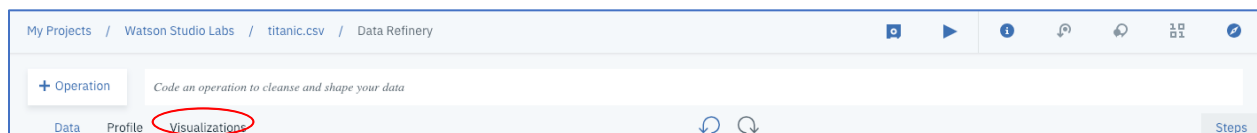


- The result is shown below. Notice that the right panel will contain a running list of the transformations.

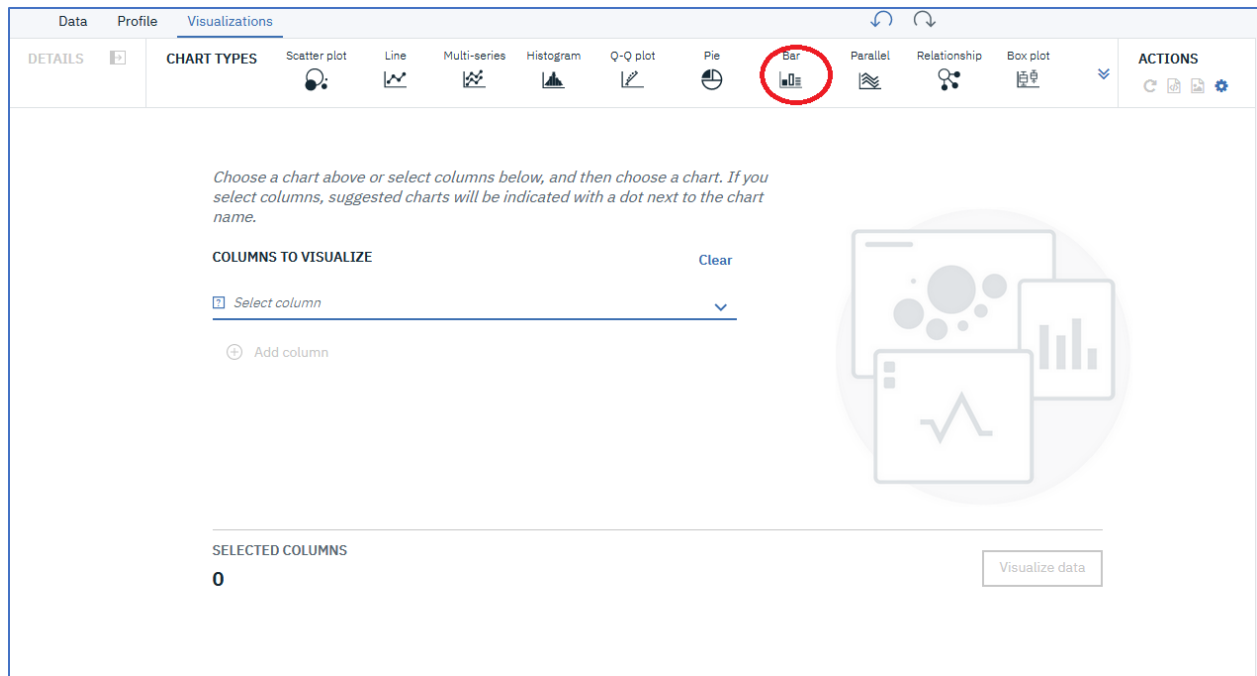
	ticket String	fare String	cabin String	embarked String	boat String	body String	home.dest String	survived_value String	pclass_value String
1	24160	211.3375	B5	S	2		St Louis, MO	Y	first
2	113781	151.5500	C22 C26	S	11		Montreal, PQ / Ches...	Y	first
3	113781	151.5500	C22 C26	S			Montreal, PQ / Ches...	N	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	0		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

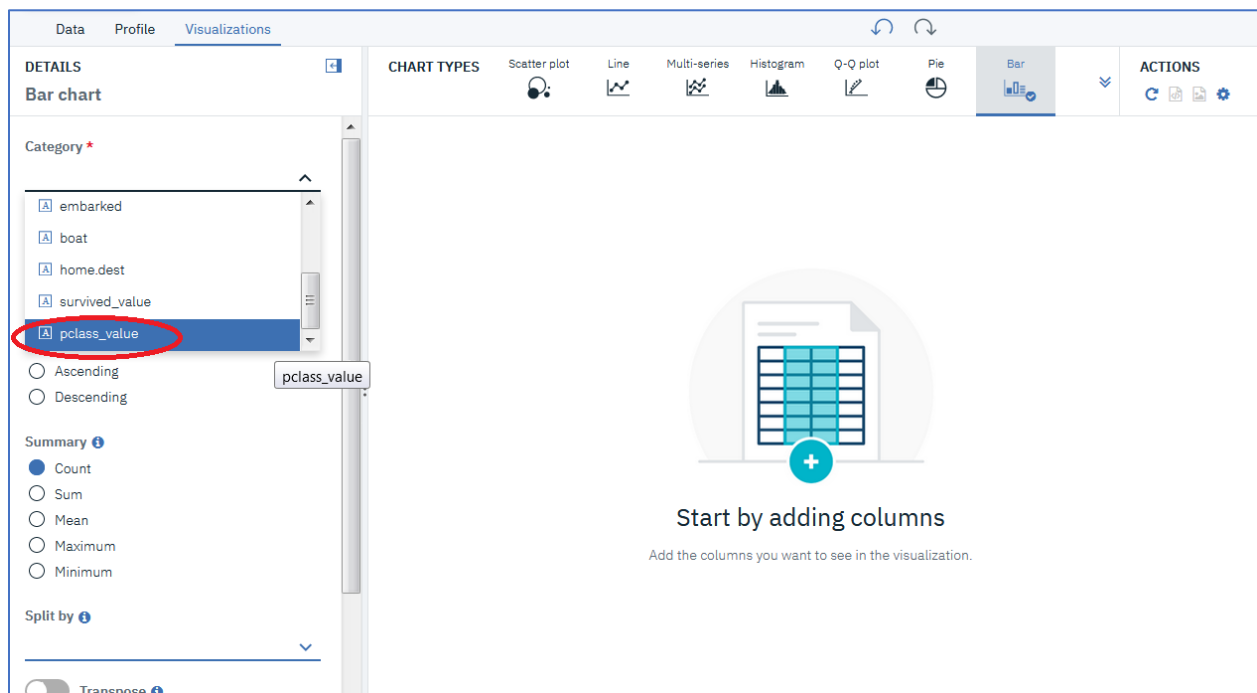
- Click on the **Visualizations** tab.



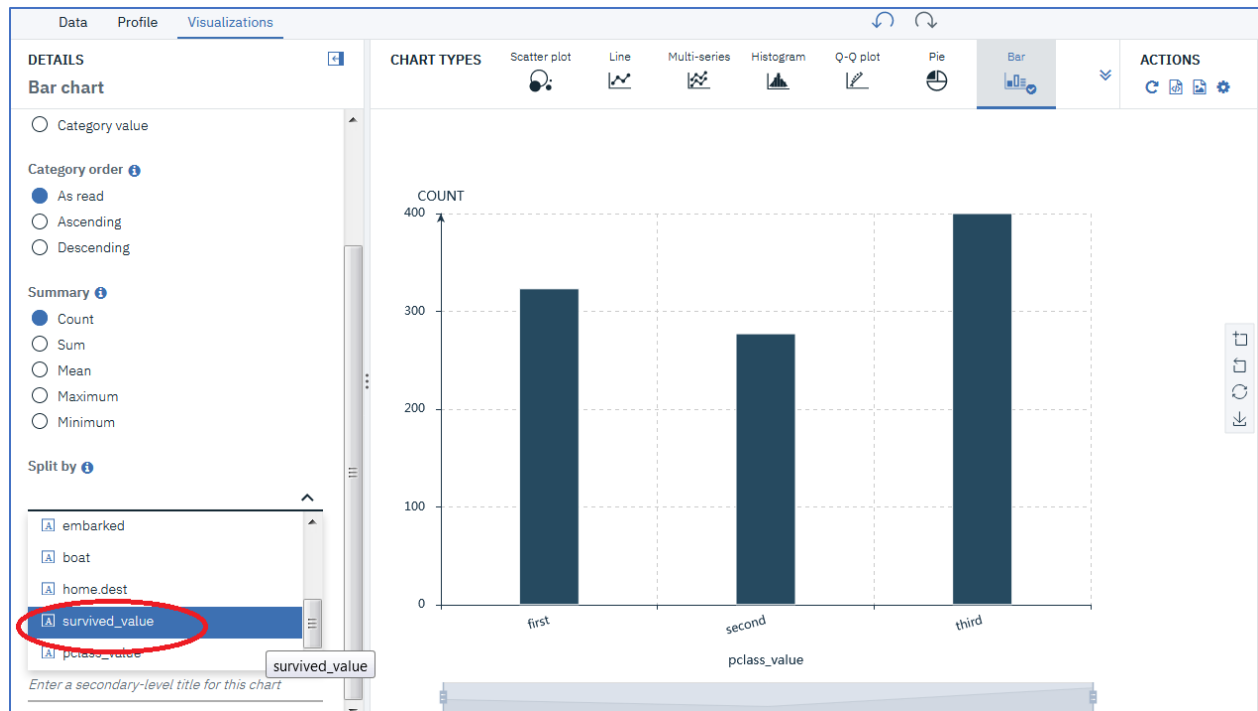
- Let's take a look at the breakdown of passengers by passenger class. We will use our new pclass_value field. Select the **Bar** Chart Type.



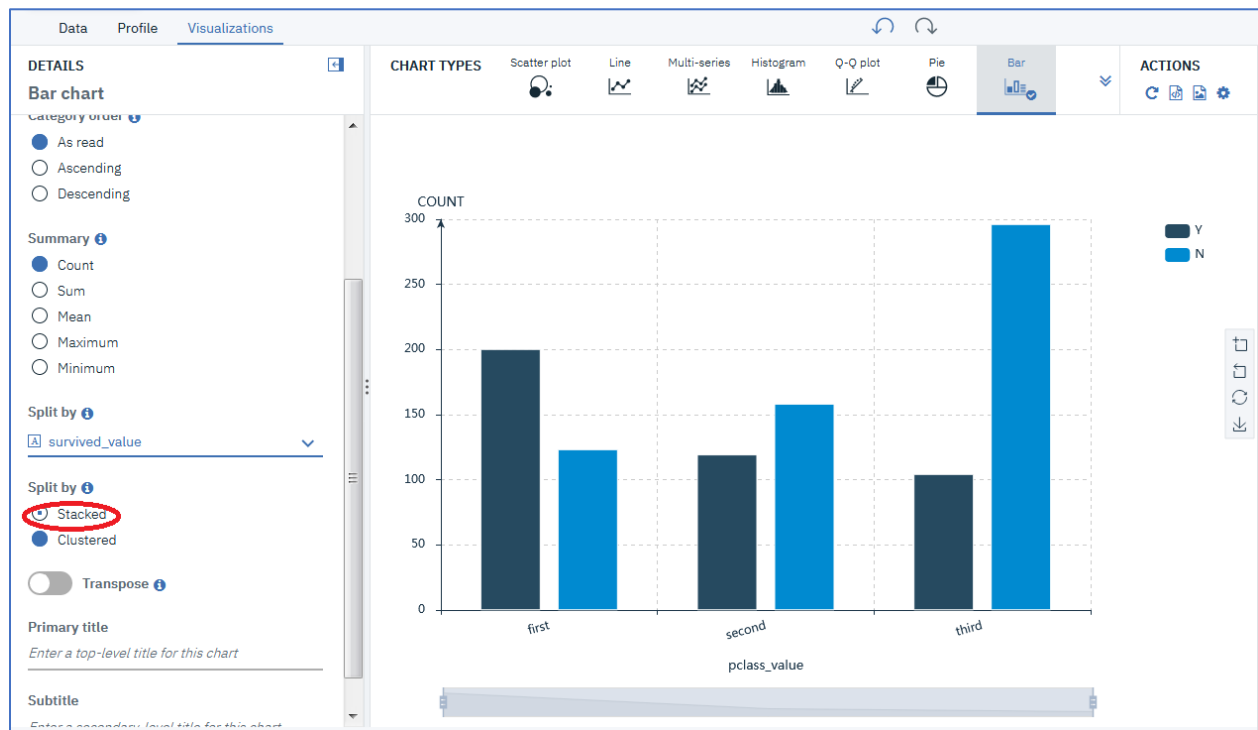
3. In the **Category** required field, select **pclass_value**.



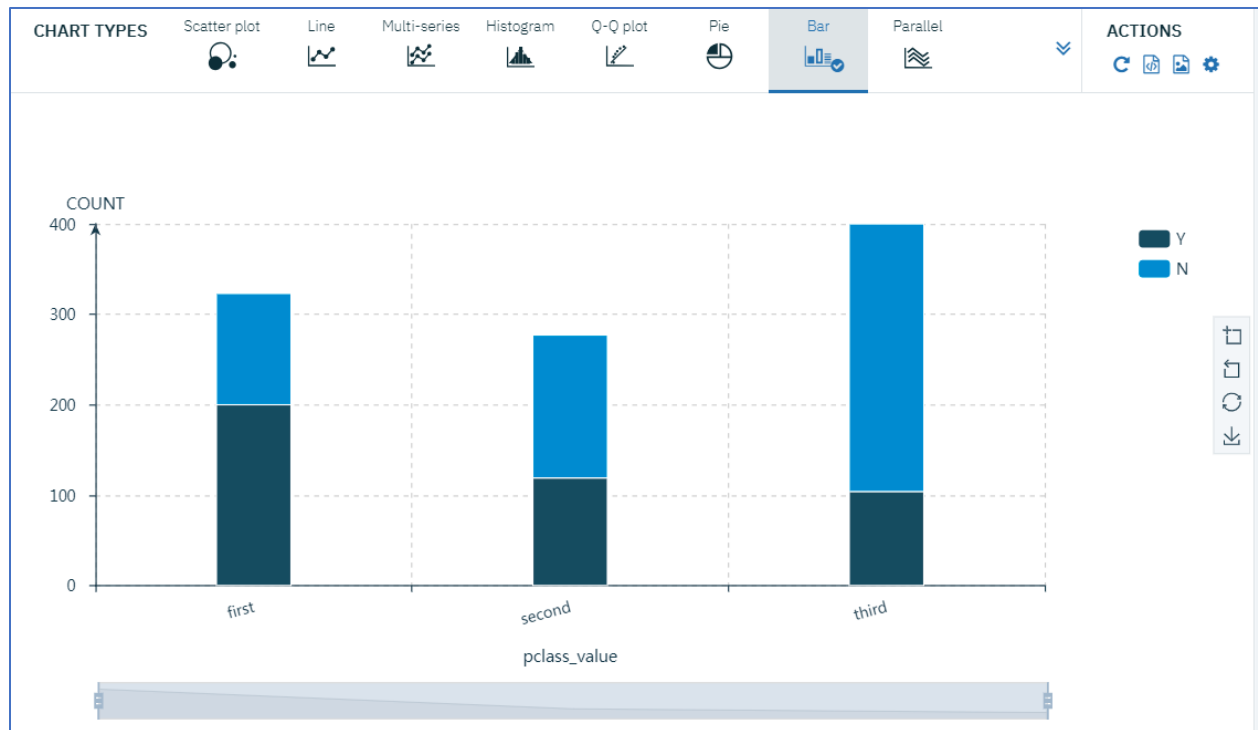
4. In the **Split by** field, select **survived_value**.



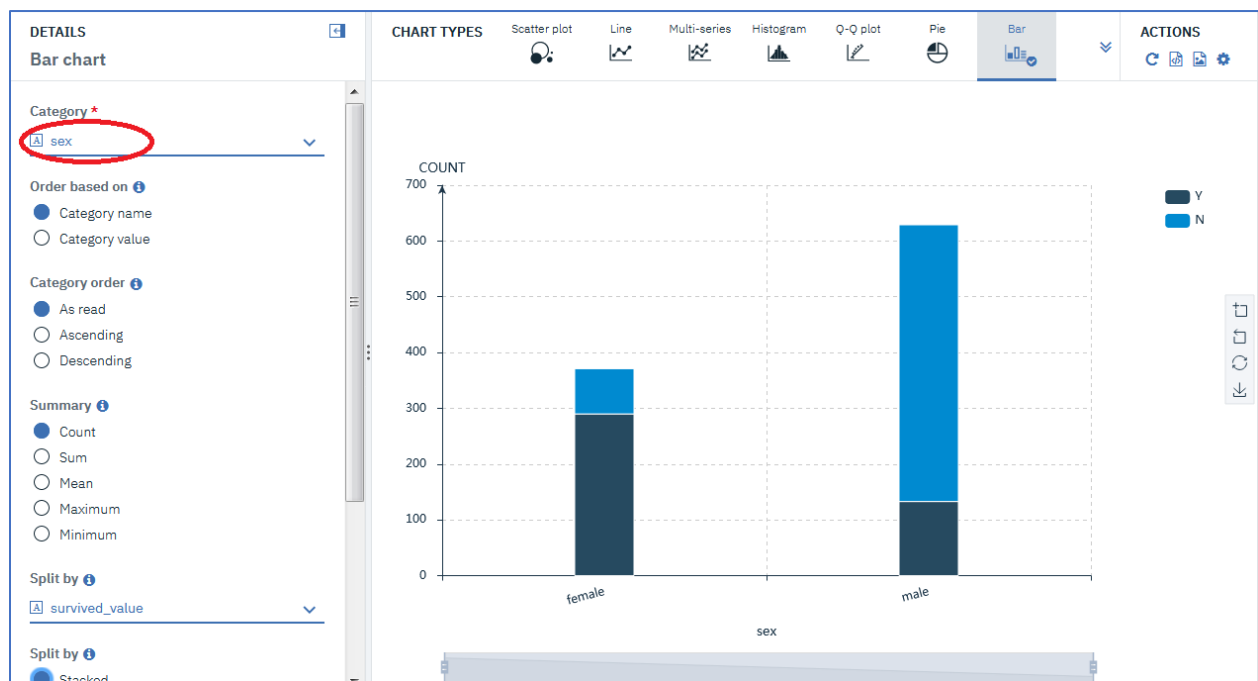
5. Select **Stacked**.



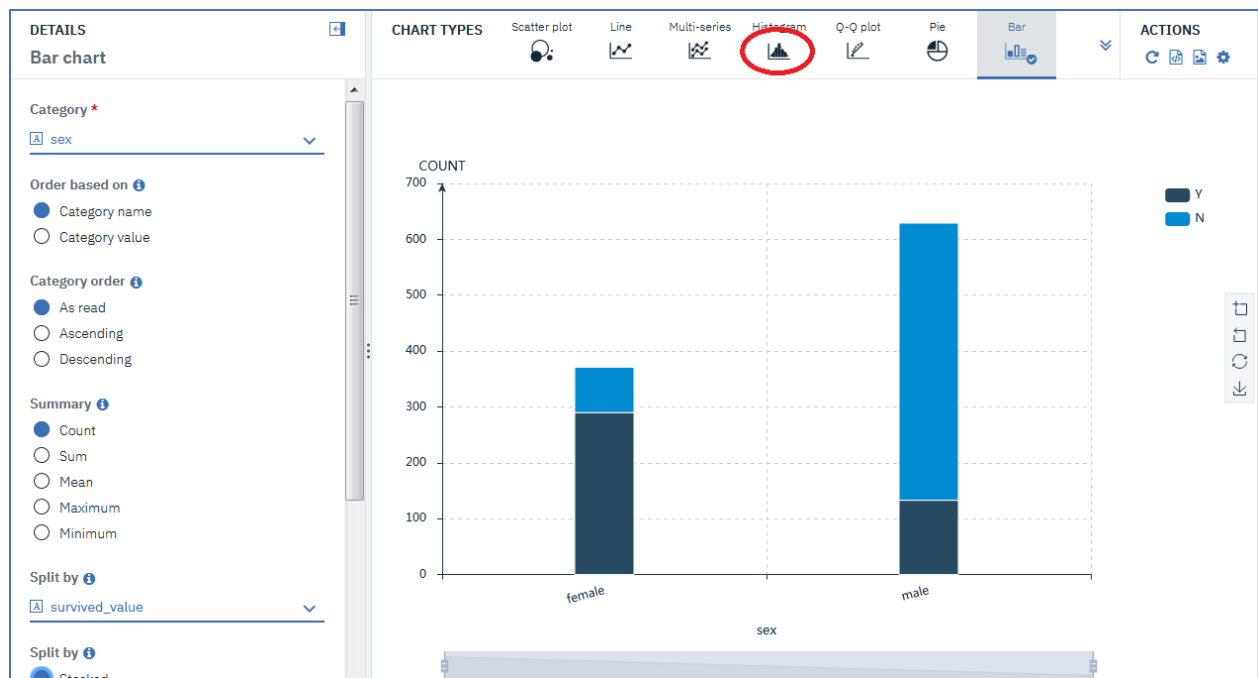
6. The result is shown below. The percentage of survivors is the greatest in first class, followed by second class, and then third-class passengers.



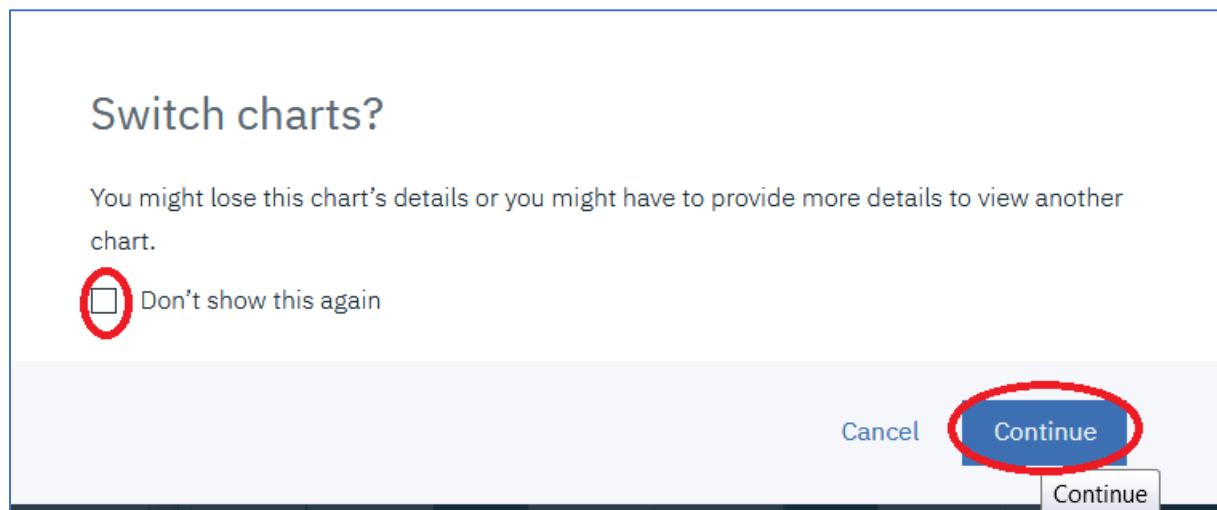
7. Change the **Category** to **sex**. We can see that survivorship for females is significantly greater than for males.



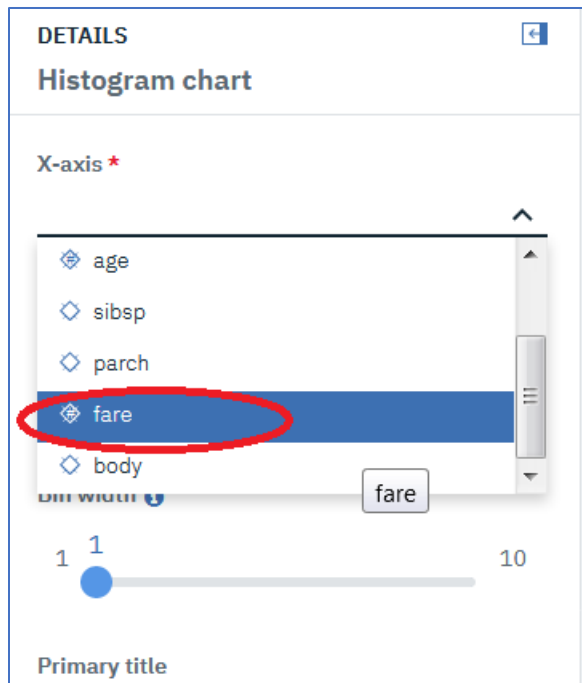
8. Click on the **Histogram** Chart Type.



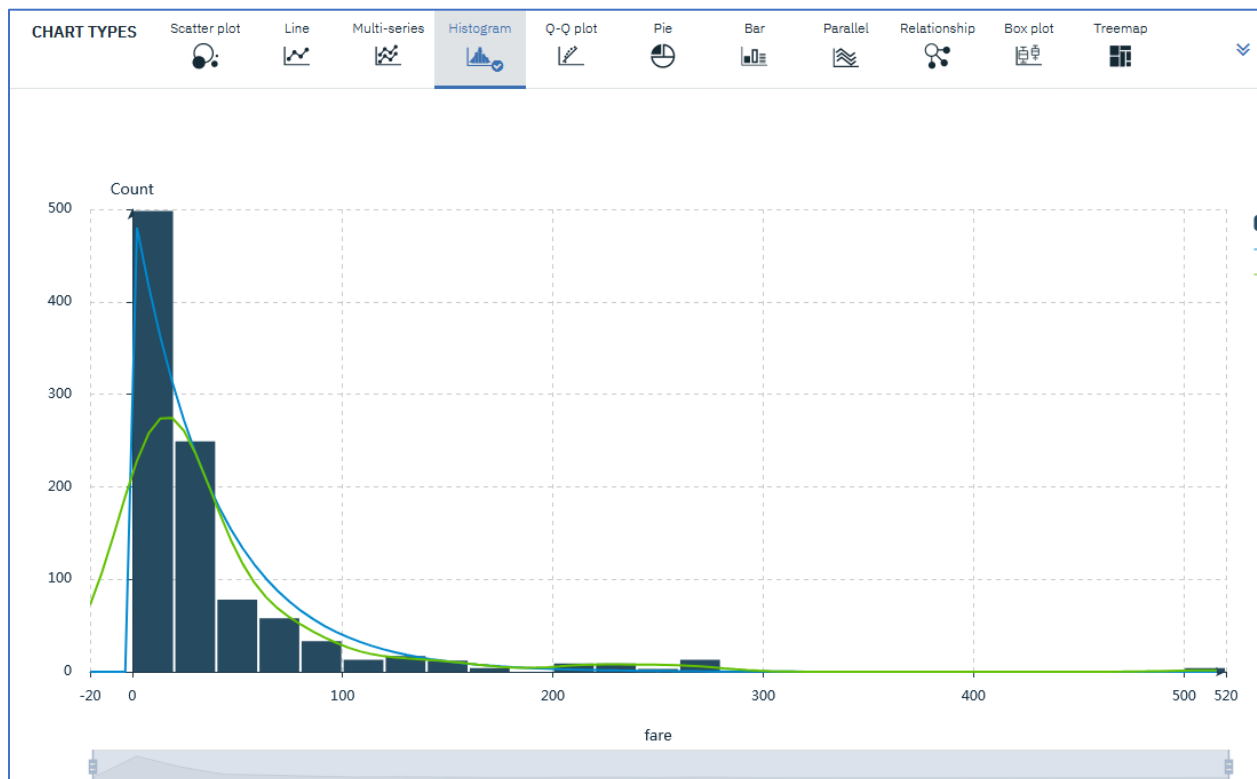
9. Click on the **Don't show this again** check box and click **Continue**.



10. Select **fare** for the X-axis. Select **None** for the Split by.



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.



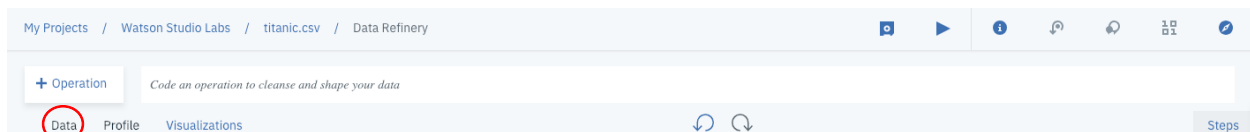
Step 4: Prepare the data for modeling

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 a data table with columns 'cabin', 'embarked', and 'boat'. The 'cabin' column is highlighted with a red circle, and a vertical ellipse next to it is also circled in red. A context menu is open over the 'cabin' column, with the 'Remove' option highlighted. The table contains several rows of data, including 'B5', 'C22 C26', 'E12', 'D7', 'A36', 'C101', 'C62 C64', and 'B35'.

cabin	embarked	boat
B5	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

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

embarked	survived_value	pclass
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		first
S		first

- If the fare column is String, 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	Custom co
26.5500		Y	mutate(pcl
77.9583		Y	ifelse(pclas
0.0000		N	d"
51.4792			a
49.5042			a
227.5250			d c
227.5250			d c
69.3000			a
78.8500			a
30.0000			d b

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**.

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

7. Convert the **age** from Decimal 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 176
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 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**.

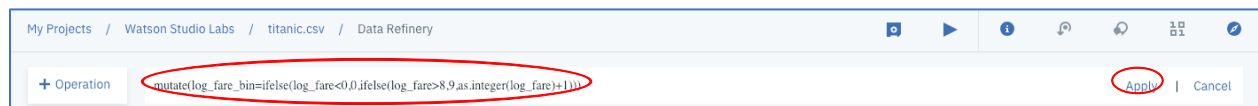
Bin	Age Range
0	0-5
1	6-11
2	12-17
3	18-39
4	40-64
5	65-79
6	Over 79



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**




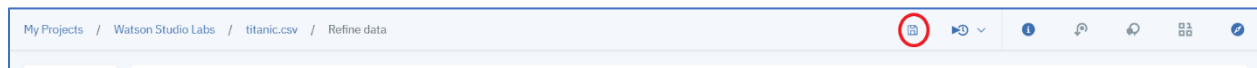
10. Now we will drop the **age**, **fare**, and **log_fare** columns as they are no longer needed for modeling purposes. Select the vertical ellipse adjacent to the column and click on **Remove** as shown below.

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		

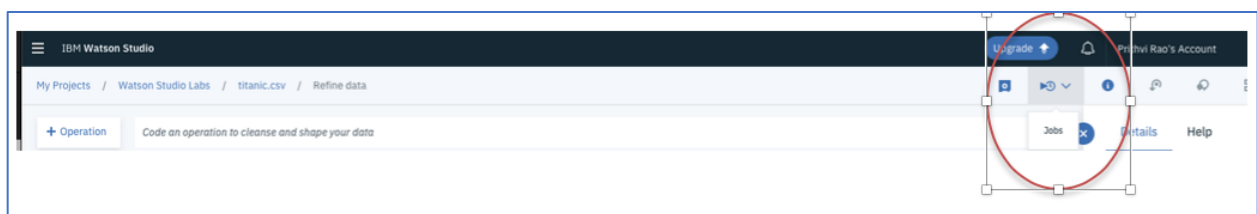
log_fare	age_bin
Decimal	Decimal
2.32497656566603	Remove
2.18055594070364	Remove duplicates
2.18055594070364	Remove empty rows
2.18055594070364	Sort ascending
2.18055594070364	Sort descending
1.42406452541749	Substitute
1.89186236009324	CONVERT COLU... >
-Inf	View All
1.71163178923691	
1.69464204659912	
2.35702912303943	4
2.35702912303943	3
1.84073323461181	3

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

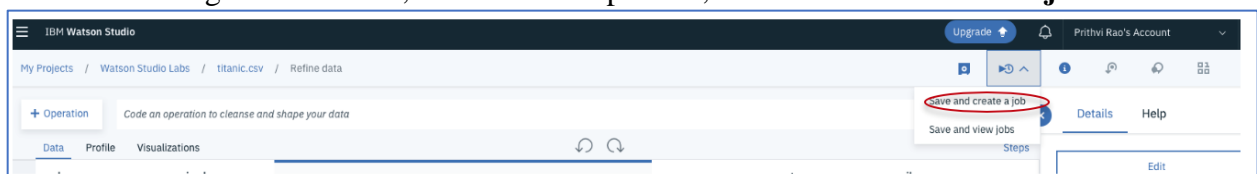


Step 5: Run the sequence of Data Flow operations on the entire data set.

1. 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 **Jobs** icon .



2. Selecting the **Jobs** icon, results in a drop down, select **Save and create a job**



3. This action results in the following page display. Fill in the **Job Name**, for example **titantic_flow_job**, leave the default for runtime, and click on the **Create and Run** button to run the job.

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

Description (Optional)

Associated asset

titanic.csv_flow

16 Steps

Edit

Select runtime

Default Data Refinery XS

INPUT

titanic.csv

OUTPUT

titanic.csv_shaped.csv

Schedule to run

No schedule enabled

Cancel Create **Create and Run**

- Note the number of steps used to transform the data. It should be 17 or 18 (depending on whether you needed to convert the Fare column). A schedule can be set up if the transformation process needs to run on a scheduled basis.

titanic_flow_job

No description

Associated Asset

titanic.csv_flow

16 Steps

Scheduled to run

No Schedule Created

Environment definition

Default Data Refinery XS

INPUT

titanic.csv

OUTPUT

titanic.csv_shaped.csv

Runs

Start Time	Status	Duration	Started By	Action
Jul 17, 2019, 6:38:57 PM	Running	---	Prithvi Rao	

- After some time, the job is completed and the status is displayed as shown in the figure below. If it is taking more than a minute, refresh the browser to see if the status changes to Completed.

titanic_flow_job

No description

Associated Asset

DATA REFINERY FLOW

titanic.csv_flow 16 Steps

Scheduled to run

Edit

No Schedule Created

Environment definition

Edit

Default Data Refinery XS

INPUT

CSV

titanic.csv

OUTPUT

CSV

titanic.csv_shaped.csv

Runs

Start Time ▲	Status	Duration	Started By	Action
Jul 17, 2019, 6:38:57 PM	Completed	20 seconds	Prithvi Rao	

6. The output of the Data Refinery process is listed in the Data Assets. Click on **Watson Studio Labs**

IBM Watson Studio

My Projects

/

Watson Studio Labs

/

titanic_flow_job

7. Click on **titanic.csv_shaped.csv** to view the asset contents.

My Projects / Watson Studio Labs

Launch IDE

Add to project

Overview

Assets

Environments

Jobs

Bookmarks

Deployments

Access Control

Settings

What assets are you looking for?

▼ Data assets

0 asset selected.

<input type="checkbox"/>	NAME	TYPE	CREATED BY	LAST MODIFIED ▼	ACTIONS
<input type="checkbox"/>	CSV titanic.csv_shaped.csv	Data Asset	Prithvi Rao	17 Jul 2019, 6:39:51 pm	
<input type="checkbox"/>	CSV titanic.csv	Data Asset	Prithvi Rao	16 Jul 2019, 7:14:59 am	
<input type="checkbox"/>	CSV titanic_cleansed.csv	Data Asset	Prithvi Rao	15 Jul 2019, 6:27:27 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.csv_shaped.csv

Preview **Lineage**

Schema: 12 Columns
 Preview: 1000 rows | Last refresh: 14 minutes ago | [Refresh](#) Refine

pclass	survived	name	sex	sibsp	parch	ticket	embarked	survived_v...	pclass_val...
Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String	Type: String
1	0	Allison, Miss. Hel	female	1	2	113781	S	N	first
1	1	Anderson, Mr. H	male	0	0	19952	S	Y	first
1	1	Appleton, Mrs. E	female	2	0	11769	S	Y	first
1	1	Astor, Mrs. John	female	1	0	PC 17757	C	Y	first
1	1	Barkworth, Mr. A	male	0	0	27042	S	Y	first
1	1	Baxter, Mrs. Jam	female	0	1	PC 17558	C	Y	first
1	1	Beckwith, Mr. Ric	male	1	1	11751	S	Y	first
1	1	Bidois, Miss. Ros	female	0	0	PC 17757	C	Y	first
1	1	Bishop, Mr. Dicki	male	1	0	11967	C	Y	first
1	1	Bjornstrom-Steff	male	0	0	110564	S	Y	first
1	1	Bonnell, Miss. Ca	female	0	0	36928	S	Y	first
1	1	Bowen, Miss. Gra	female	0	0	PC 17608	C	Y	first
1	0	Brady, Mr. John	male	0	0	113054	S	N	first
1	1	Brown, Mrs. Jam	female	0	0	PC 17610	C	Y	first
1	1	Burns, Miss. Eliza	female	0	0	16966	C	Y	first
1	1	Calderhead, Mr. I	male	0	0	PC 17476	S	Y	first
1	1	Cardeza, Mrs. Jai	female	0	1	PC 17755	C	Y	first
1	0	Carrau, Mr. Jose	male	0	0	113059	S	N	first

You have completed the Lab !!!

- ✓ Profiled the data to help determine missing values
- ✓ Visualized the data to gain a better understanding
- ✓ Prepared the data for modeling
- ✓ Ran the sequence of data preparation operations on the entire data set.
- ✓ Verified the output data asset.