

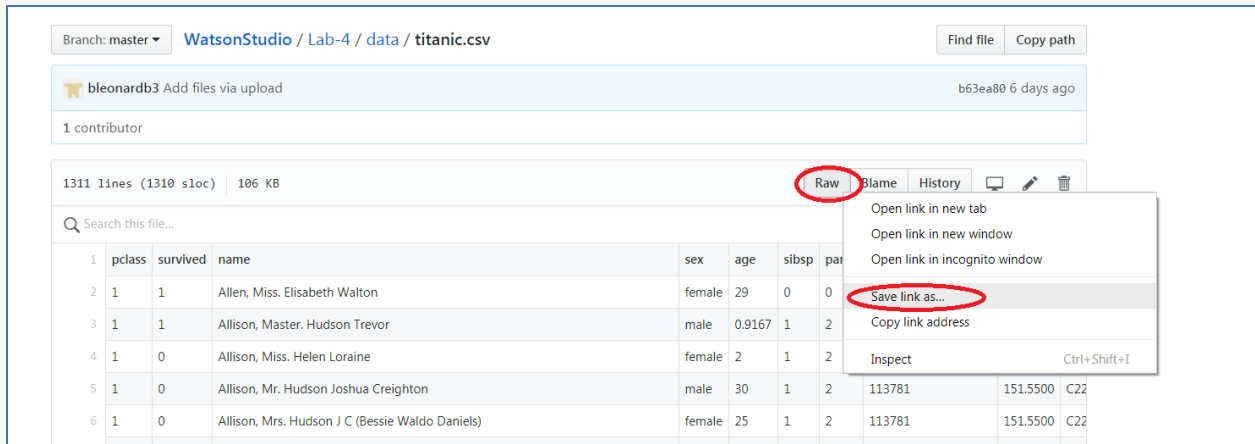
# Data Refinery Lab

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. Add a Data Asset to the Watson Studio Labs project (skip this step if you completed Lab 4-b)
2. 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: Add a Data Asset to the Watson Studio Labs project (skip this step if you completed Lab 4b)

1. Click [here](#)
2. Right-click on **Raw** and then click on **Save link as ...**



Branch: master | WatsonStudio / Lab-4 / data / titanic.csv | Find file | Copy path

bleonardb3 Add files via upload | b63ea80 6 days ago

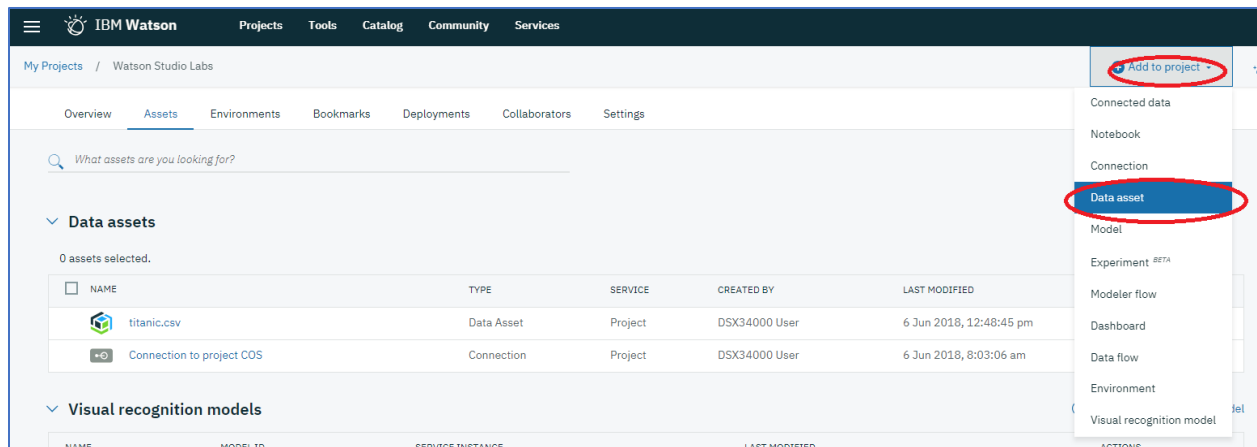
1 contributor

1311 lines (1310 sloc) | 106 KB

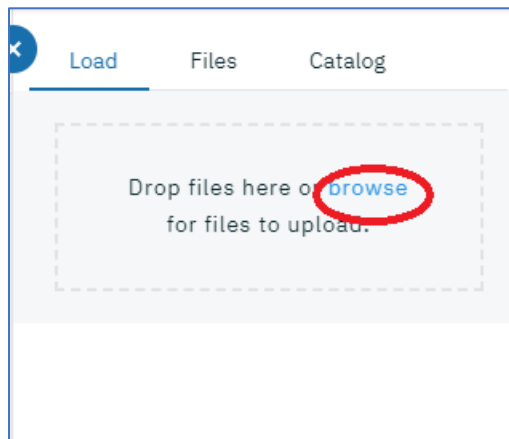
Search this file...

	pclass	survived	name	sex	age	sibsp	par			
1	1	1	Allen, Miss. Elisabeth Walton	female	29	0	0			
2	1	1	Allison, Master. Hudson Trevor	male	0.9167	1	2			
3	1	0	Allison, Miss. Helen Loraine	female	2	1	2			
4	1	0	Allison, Mr. Hudson Joshua Creighton	male	30	1	2	113781	151.5500	C22
5	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25	1	2	113781	151.5500	C22

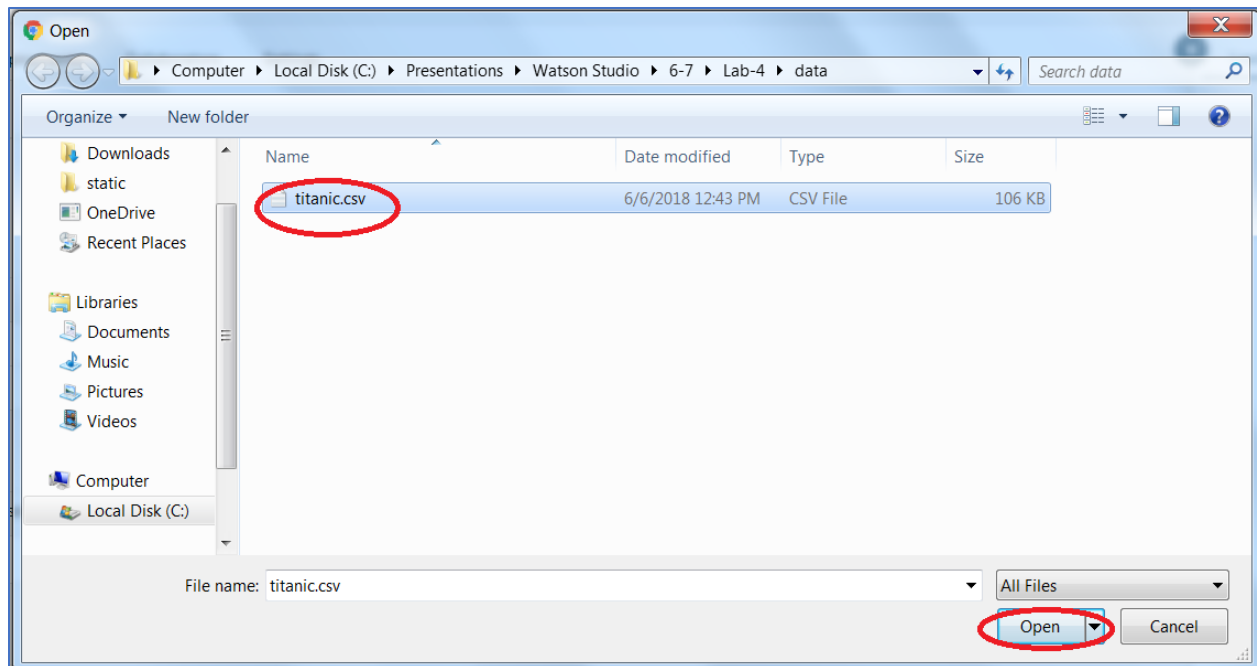
3. Go back to Watson Studio. Click on **Add to project** and then click **Data Assets**.



4. Click on **browse**

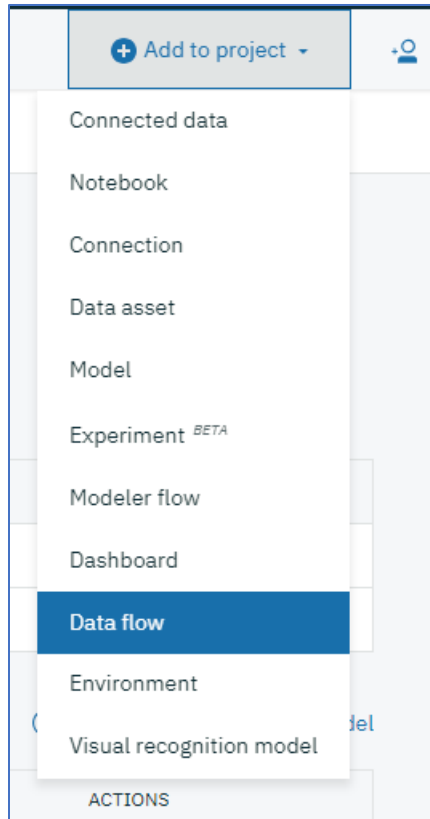


5. Navigate to the folder where you downloaded the titanic.csv data set. Select the titanic.csv file and click **Open**.



Step 2: Profile the data to help determine missing values.

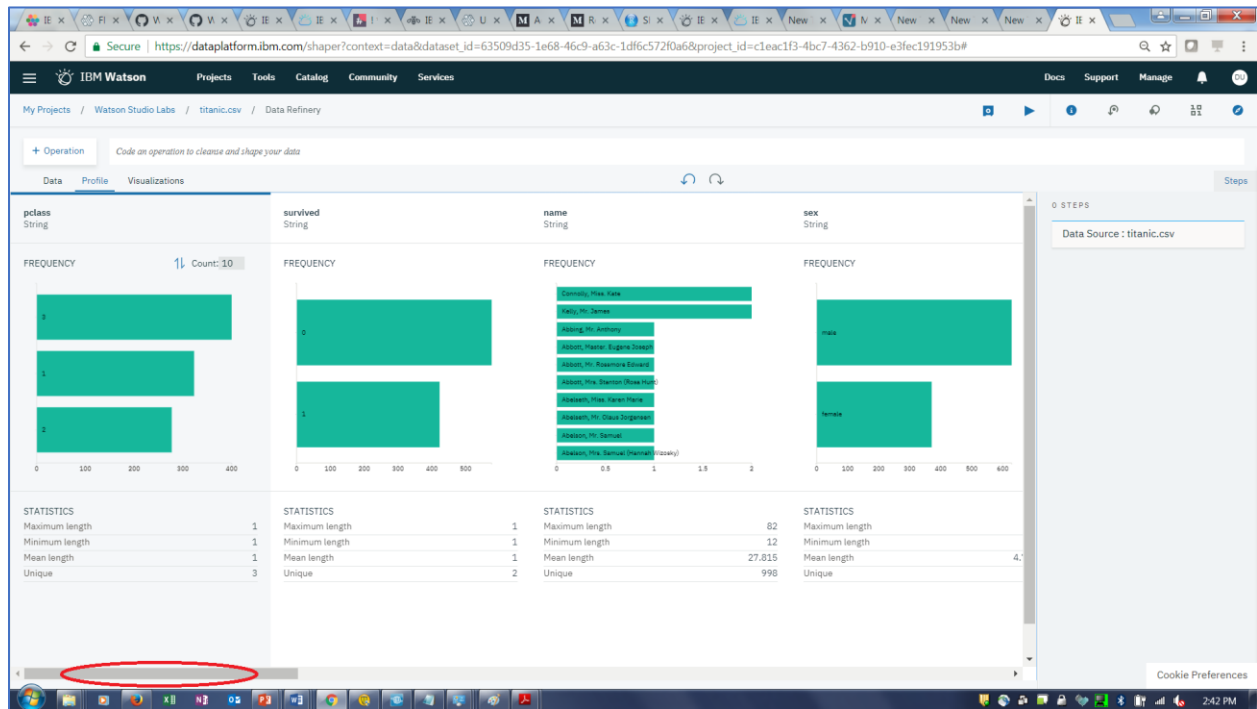
1. Add a Data Flow by clicking on **Add to project** and then click **Data Flow**.



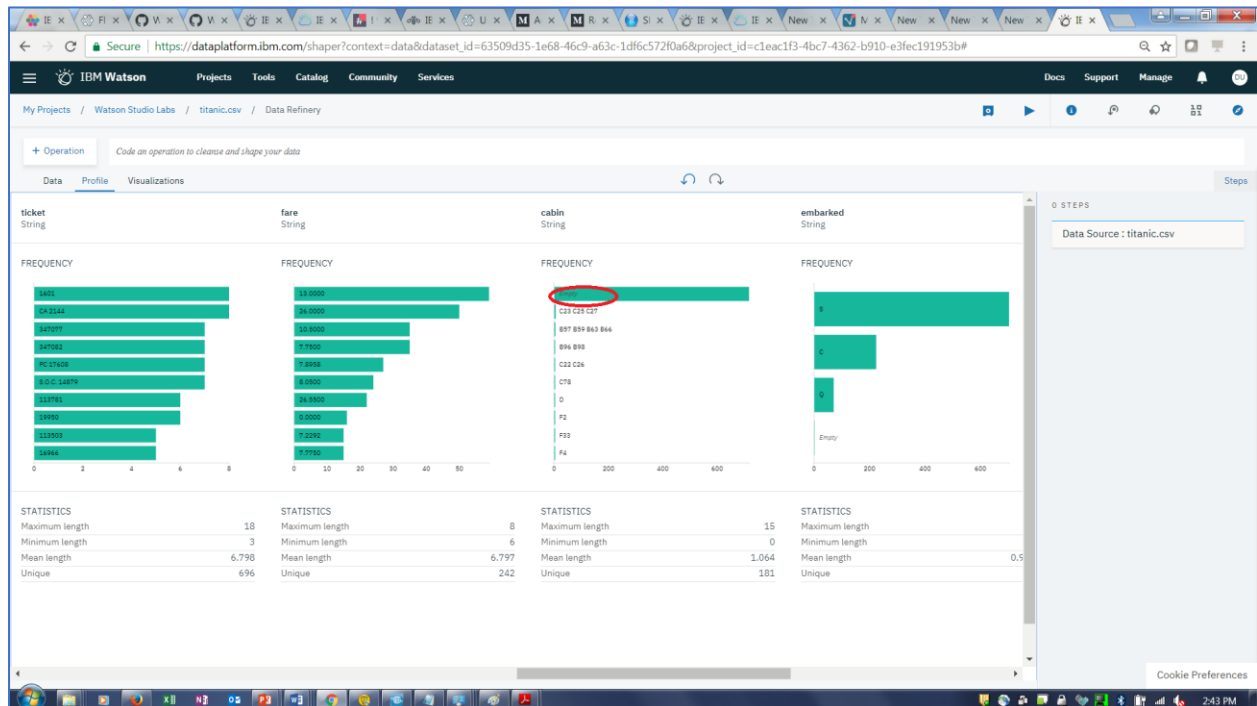
2. The Data Refinery panel will display the Titanic data set. Click on the **Profile** tab.

	pclass String	survived String	name String	sex String	age String	sibsp String	parch String	ticket String	fare String	cal String
1	1	1	Allen, Miss. Elisabet...	female	29	0	0	24160	211.3375	BF
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	A2
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	B2
11	1	0	Astor, Col. John Jacob	male	47	1	0	PC 17757	227.5250	C4
12	1	1	Astor, Mrs. John Jac...	female	18	1	0	PC 17757	227.5250	C4
13	1	1	Aubart, Mme. Leont...	female	24	0	0	PC 17477	69.3000	B3
14	1	1	Barber, Miss. Ellen ...	female	26	0	0	19877	78.8500	A2
15	1	1	Barkworth, Mr. Alge...	male	80	0	0	27042	30.0000	A2
16	1	0	Baumann, Mr. John D	male		0	0	PC 17318	25.9250	B2
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	C4

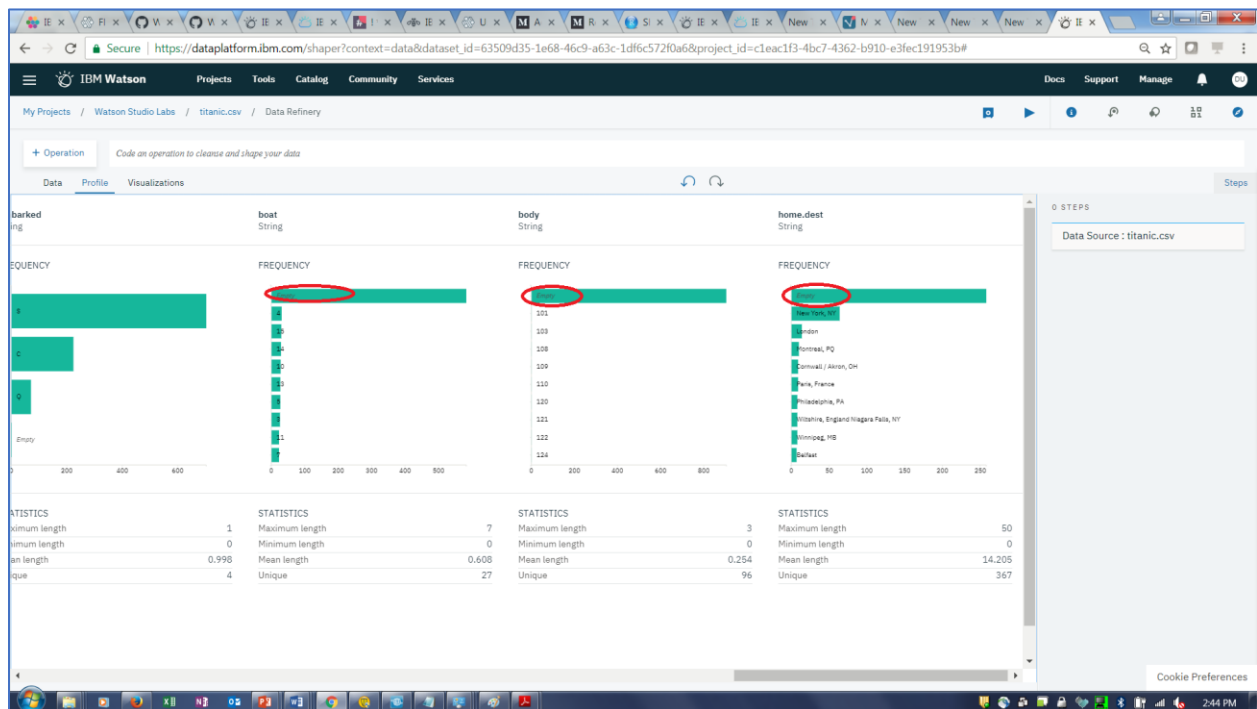
- The Profile panel displays the counts of the top 10 count values for each column. Note that you can change 10 to another number if desired. You can also switch to the bottom 10 counts for a column. Scroll to the right to view the cabin column.



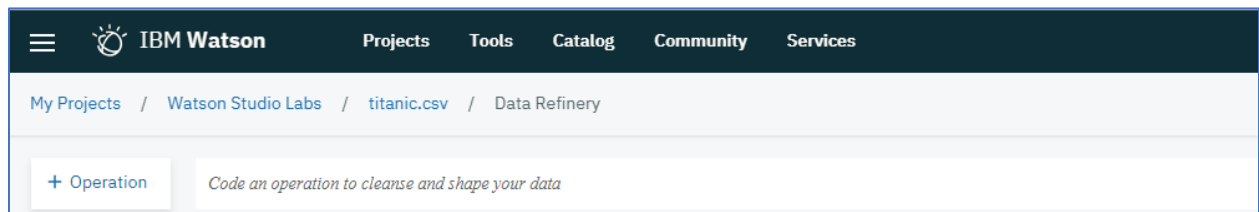
- Note that the cabin column has many missing values and should be removed as part of the data preparation step.



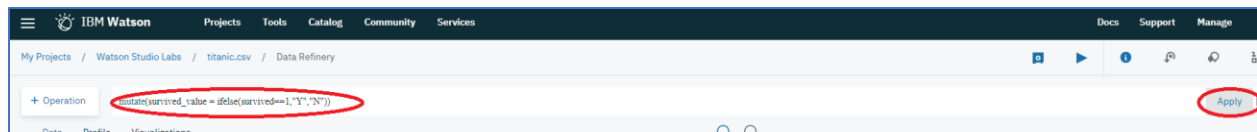
5. In a similar fashion, scroll to the right to examine the boat, body, and home.dest columns. These also have many missing values and should be removed as part of the data preparation step.



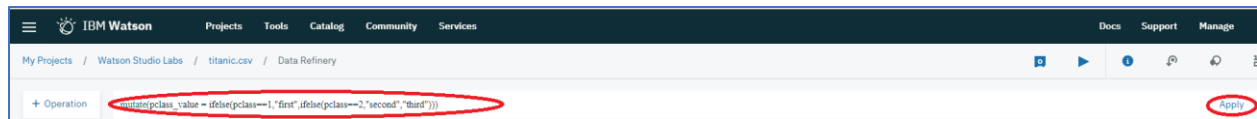
6. Age and Embarked also have missing values. Embarked has very few missing values. Age has over 100 missing values, but we will keep that column in the analysis. As part of data preparation, we will remove the rows that contain the missing age and embarked values.
7. Click on the **Data** tab. We will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



8. 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"



- Type the following to create pclass\_value,  
mutate(pclass\_value=ifelse(pclass==1,"first",ifelse(pclass==2,"second","third")))

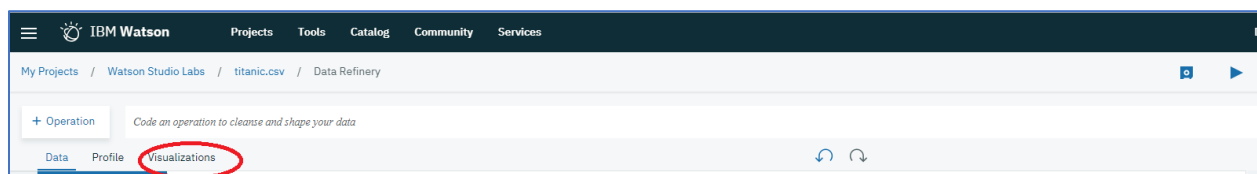


- The result is shown below. Notice that the right panel will contain a running list of the transformations.

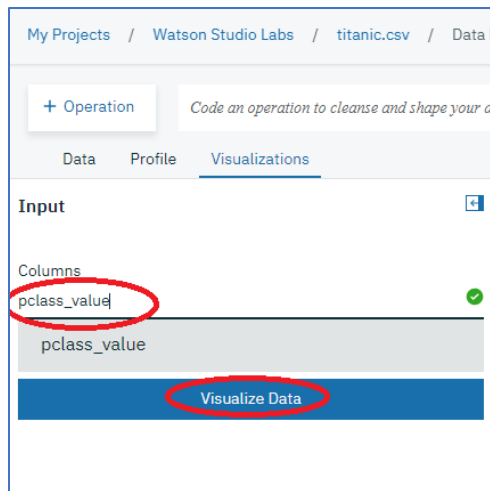
	ticket	fare	cabin	embarked	boat	body	home.dest	survived_value	pclass_value
	String	String	String	String	String	String	String	String	String
1	24160	211.3375	B5	S	2		St Louis, MO	Y	first
2	113781	151.5500	C22 C26	S	11		Montreal, PQ / Ches...	Y	first
3	113781	151.5500	C22 C26	S			Montreal, PQ / Ches...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

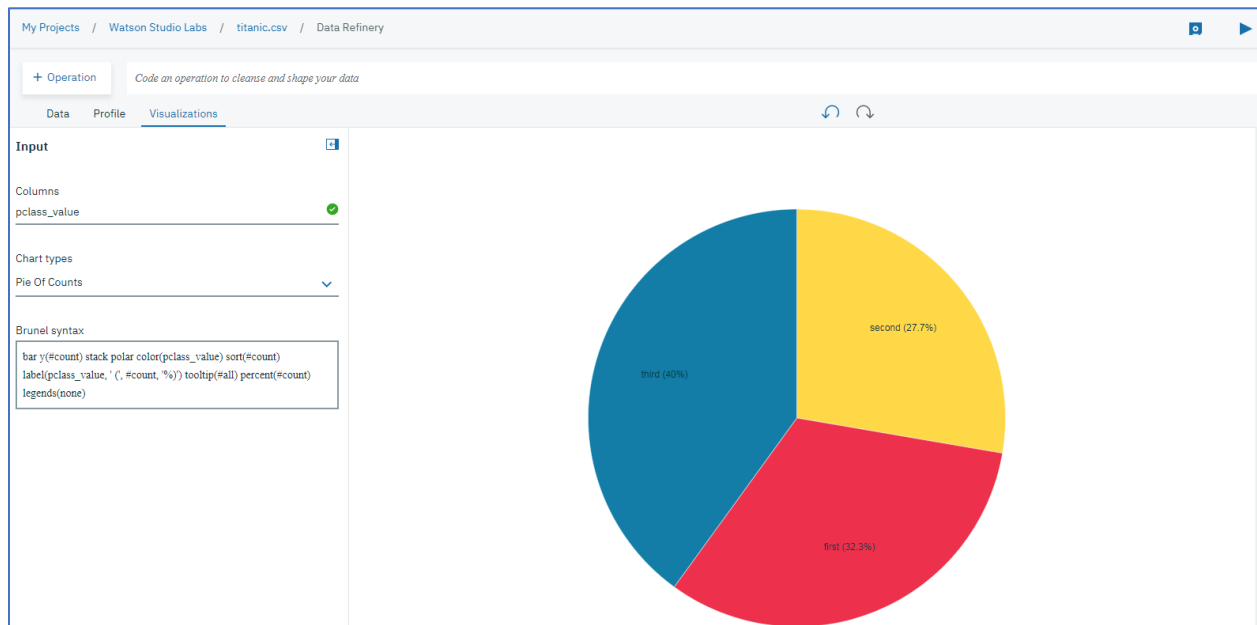
- Click on the **Visualizations** tab.



- Let's take a look at the breakdown of passengers by passenger class. We will use our new pclass\_value field. Enter or select pclass\_value and then click **Visualize Data**

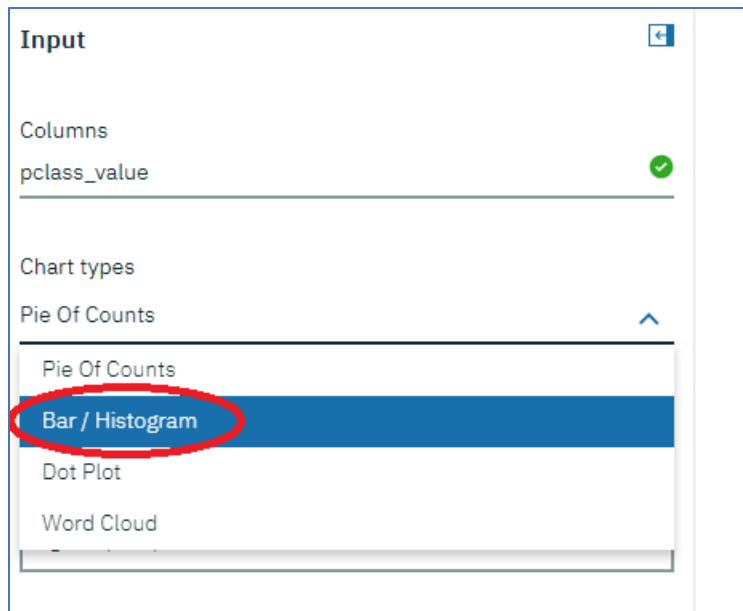


3. The result is shown below.

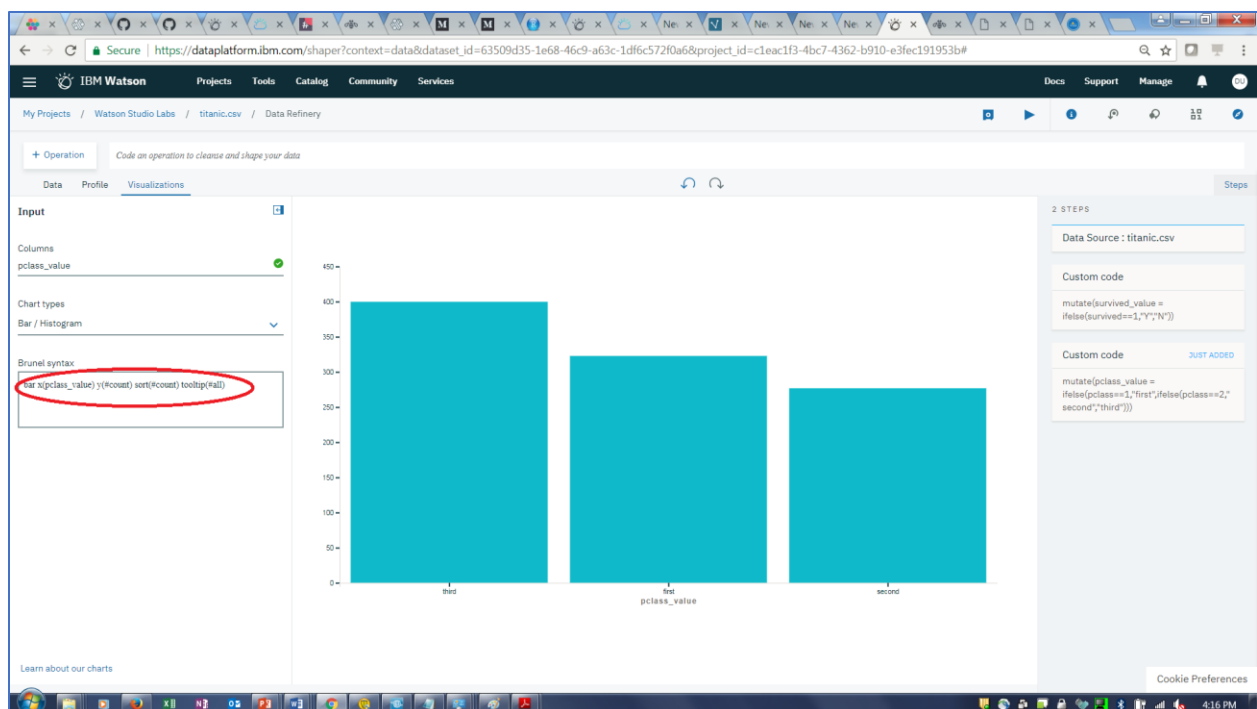


4. We can switch this to a bar chart, by switching the Chart type.

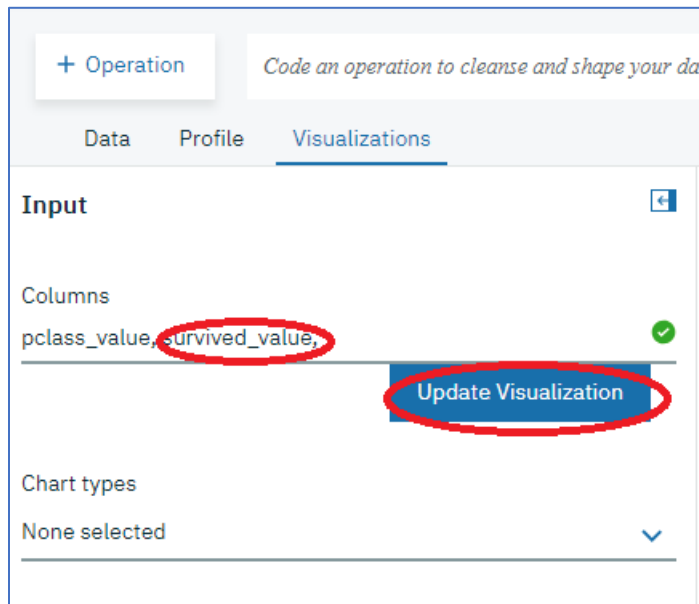




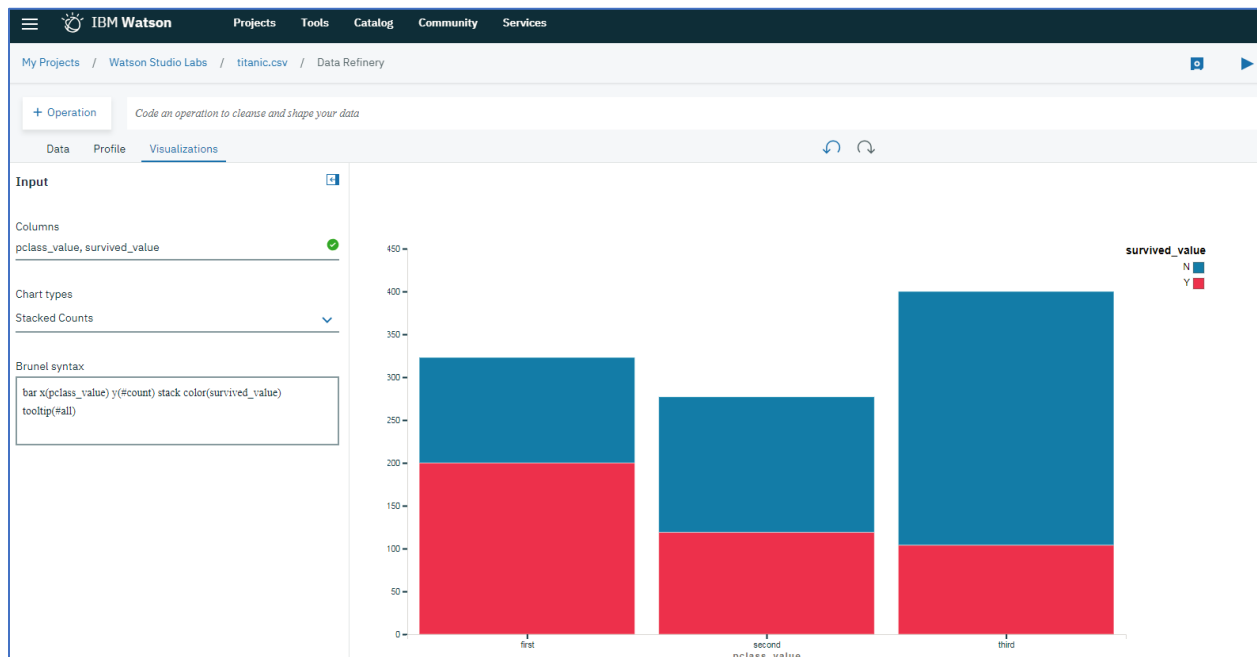
- The result is shown below. Note the Brunel coding syntax. According to the Brunel github repo, *Brunel defines a highly succinct and novel language that defines interactive data visualizations based on tabular data. The language is well suited for both data scientists and more aggressive business users. The system interprets the language and produces visualizations using the user's choice of existing lower-level visualization technologies typically used by application engineers such as RAVE or D3. It can operate stand-alone and integrated into Jupyter (IPython) notebooks with further integrations as well as other low-level rendering support depending on the desires of the community. If you understand the syntax, you can make changes and update the visualization.*



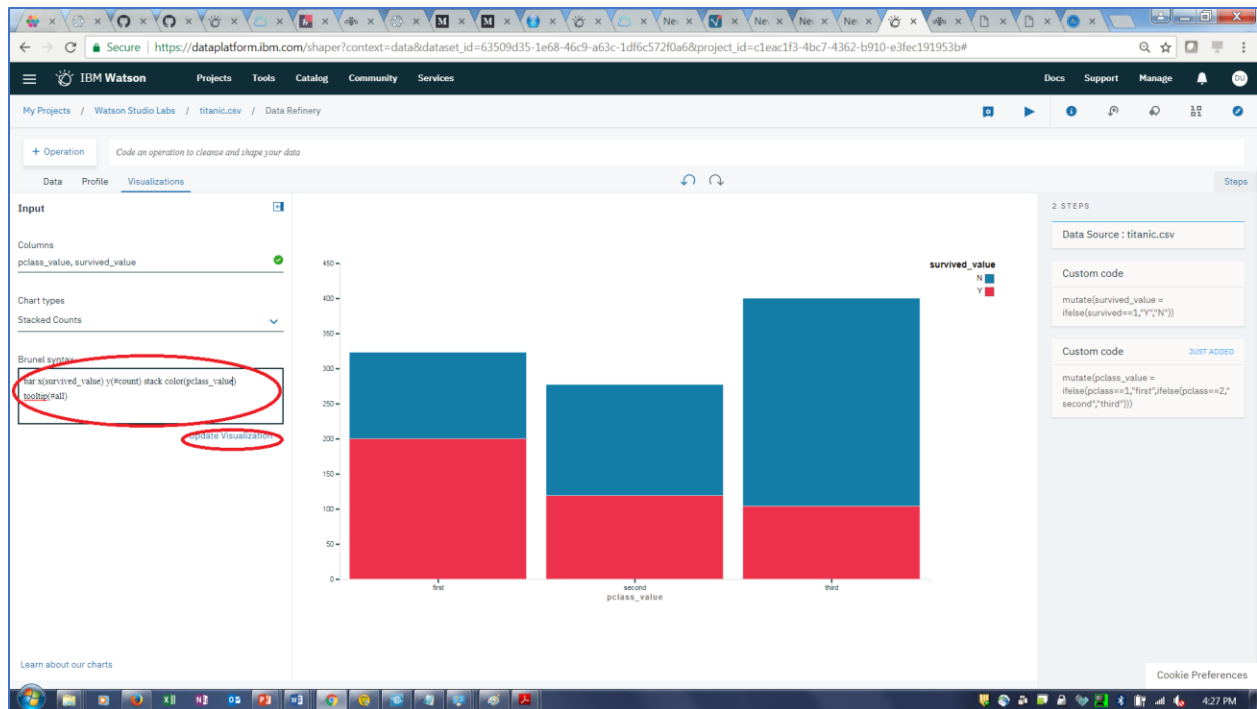
6. Let's examine the relationship between survival and the passenger class. We will add the `survived_value` and click **Update Visualization**.



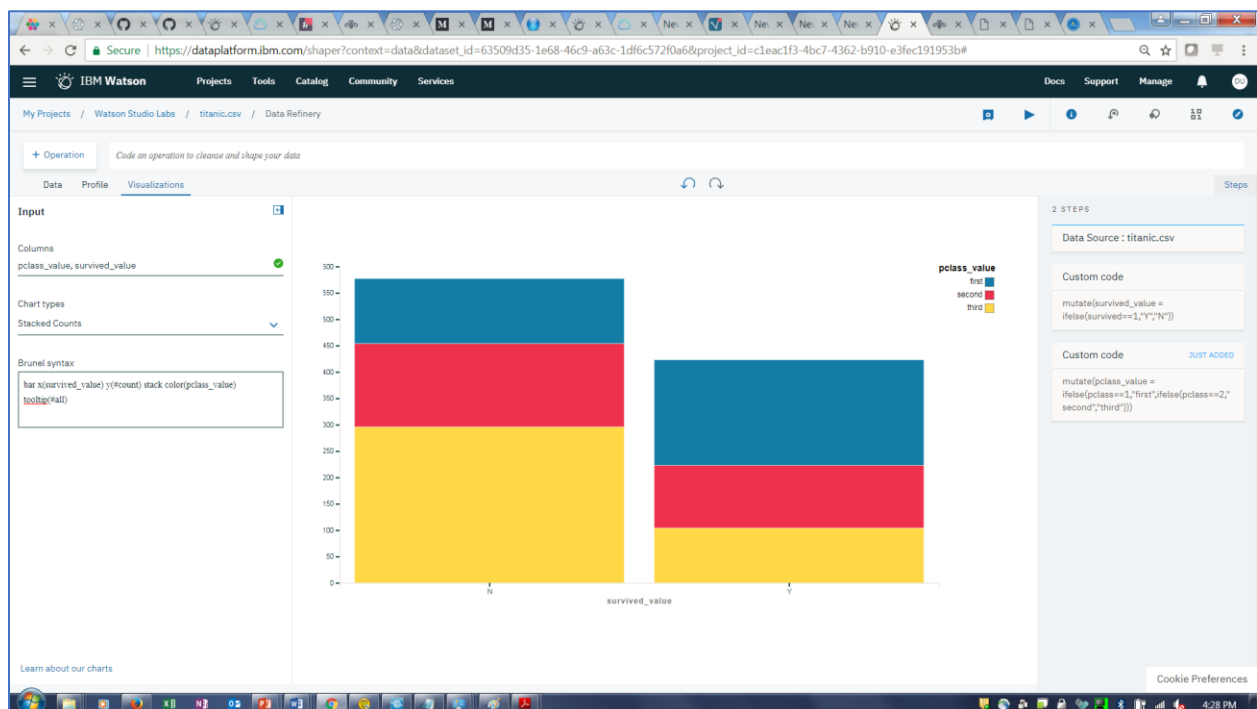
7. The result is shown below. We can see that survival probability for first class customers is significantly better.



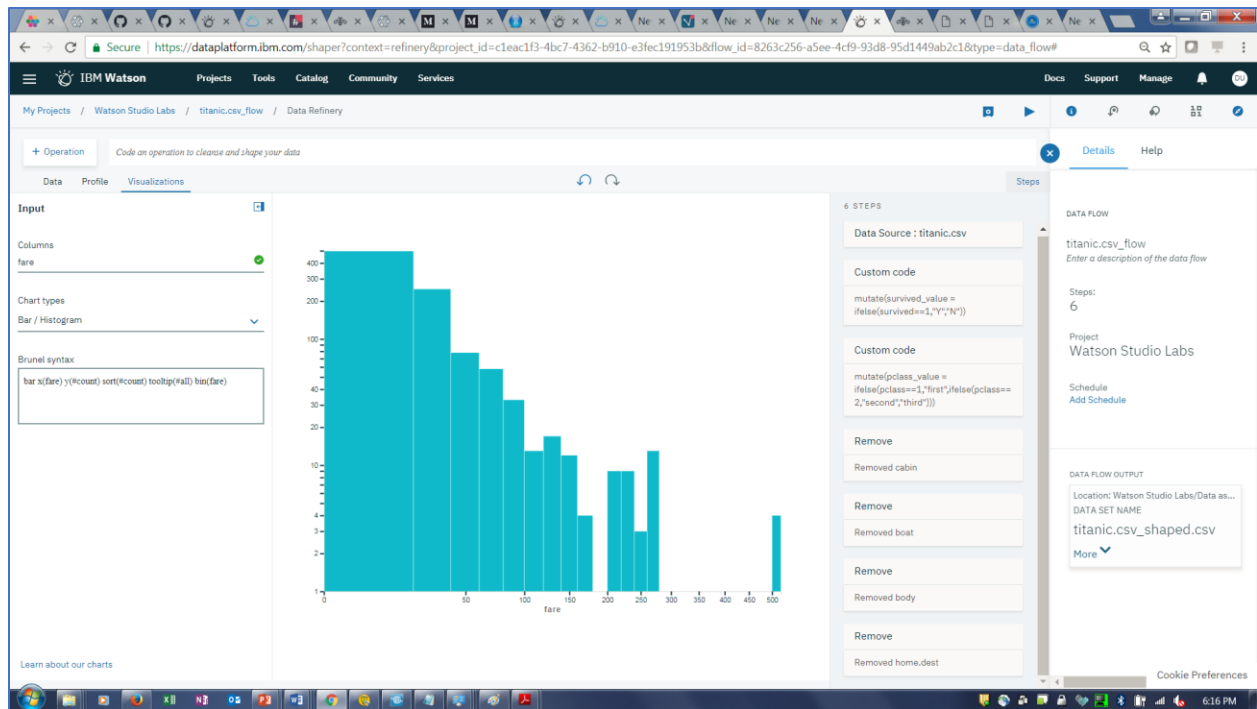
8. If we want to change `pclass_value` to be on the y-axis, and `survived_value` to be on the x-axis, we can edit the Brunel syntax, and then click **Update Visualization**.



9. The result is shown below.



10. Plot the fare values. 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.



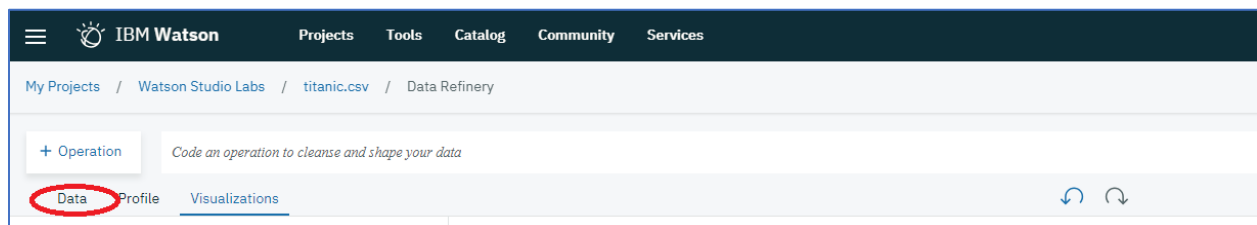
## Step 4: Prepare the data for modeling

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



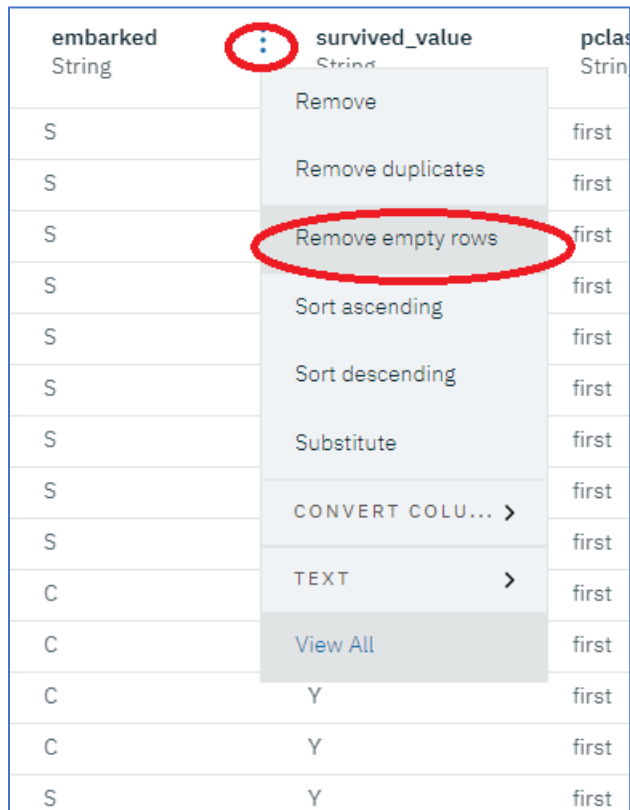
2. Remove the cabin column by selecting on the vertical ellipse and then clicking on **Remove**.

cabin String	embarked String	boat 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.

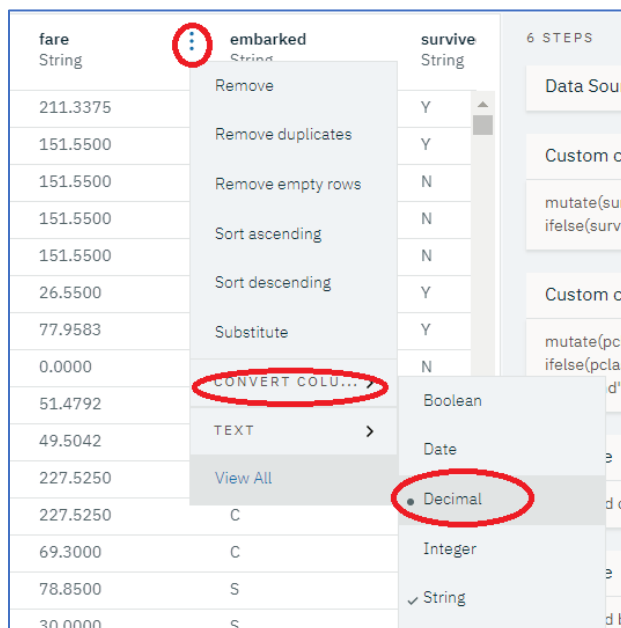
6 STEPS
Data Source : titanic.csv
Custom code
<code>mutate(survived_value = ifelse(survived==1,"Y","N"))</code>
Custom code
<code>mutate(pclass_value = ifelse(pclass==1,"first",ifelse(pclass== 2,"second","third")))</code>
Remove
Removed cabin
Remove
Removed boat
Remove
Removed body
Remove JUST ADDED
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 String	survived_value String	pclass 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

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**.

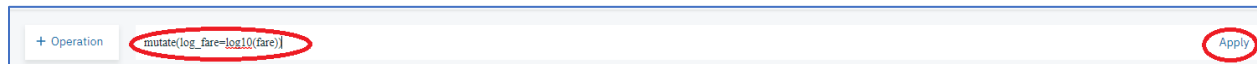


fare String	embarked String	survived_value String	pclass String
211.3375	S	Y	first
151.5500	S	Y	first
151.5500	S	N	first
151.5500	S	N	first
151.5500	S	N	first
26.5500	S	Y	first
77.9583	S	Y	first
0.0000	S	N	first
51.4792	C		first
49.5042	C		first
227.5250	S		first
227.5250	C		first
69.3000	S		first
78.8500	S		first
30.0000	S		first

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**.

age Integer	sibsp String	parch String	ticket 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

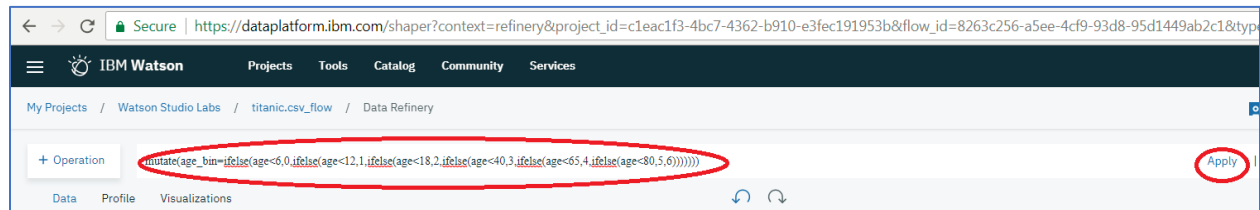
8. Bin the age column into the following bins by clicking into the **Code an operation to cleanse and shape your data**, and entering

```
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

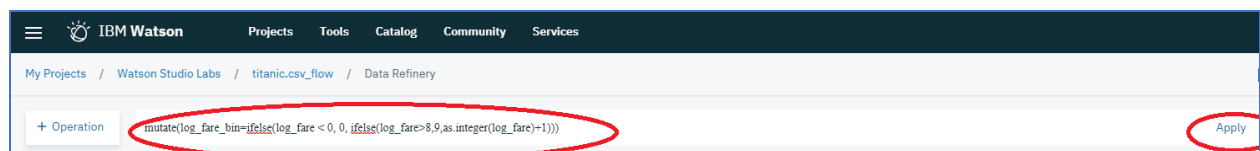
4	40-64
5	65-79
6	Over 79




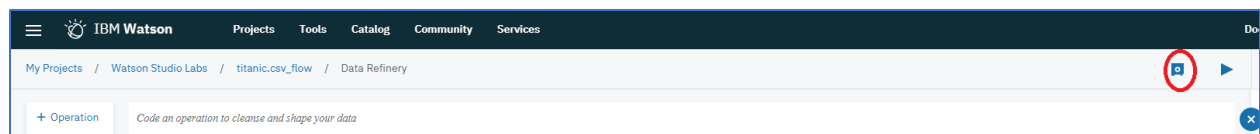
9. Bin the log\_fare column, by clicking into the **Code an operation to cleanse and shape your data**, and entering

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

and then clicking **Apply**




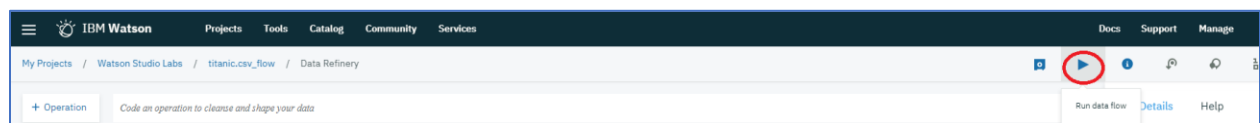
10. Save the Data Flow by clicking on the Save Data Flow icon .



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. Optionally change the name of the output file by clicking on the edit option. Click **Save and Run**.



### Data flow details

Name\*  
titanic.csv\_flow  
Enter a description of the data flow

Steps:  
11

Project  
Watson Studio Labs

Schedule  
[Add Schedule](#)

Required Fields\*

Location: Watson Studio Labs/Data assets

DATA SET NAME  
titanic.csv\_shaped.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**

3. You can continue to work on other items, or monitor the Data Flow run status.

## 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](#)

4. The completed flow is shown below.

IBM Watson

Projects

Tools

Catalog

Community

Services

Docs

Support

Manage

My Projects / Watson Studio Labs / titanic.csv\_flow

Refine

Summary

Source