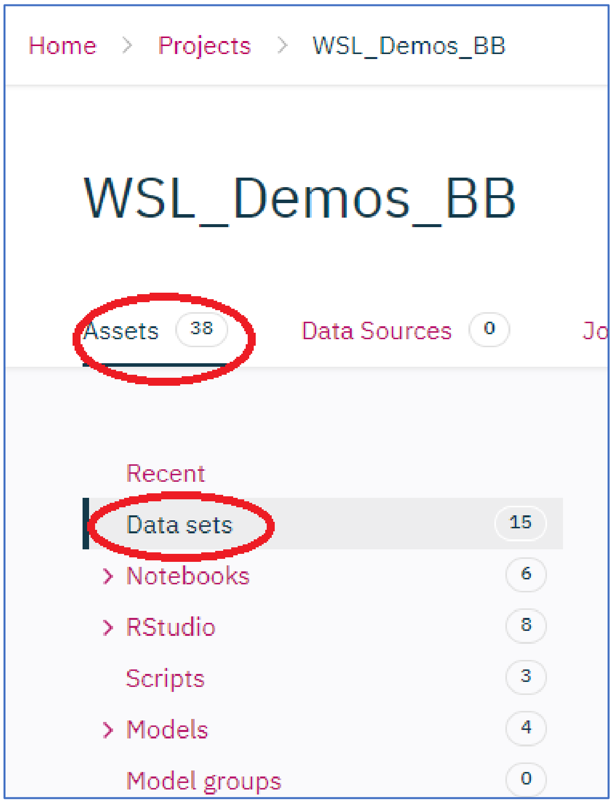
Data Refinery Lab

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:

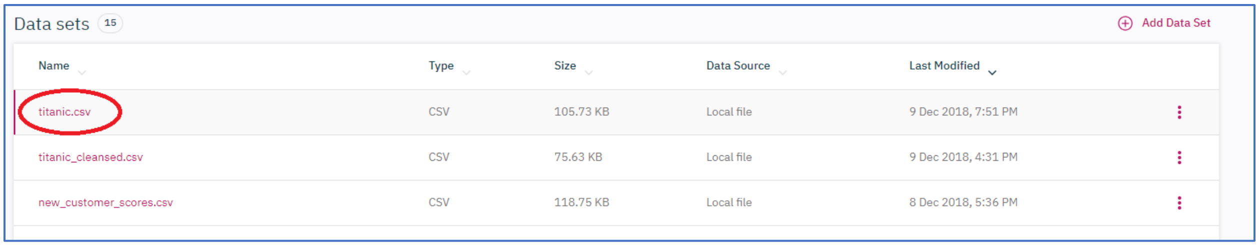
1. Use the Data Refinery Tool to:
   a. Profile the data to help determine missing values
   b. Visualize the data to gain a better understanding
   c. Prepare the data for modeling
   d. Run the sequence of data preparation operations on the entire data set.

Step 1: Add a Data Flow

1. In the Watson Studio Local project, click on **Assets** and click on **Data Sets**.

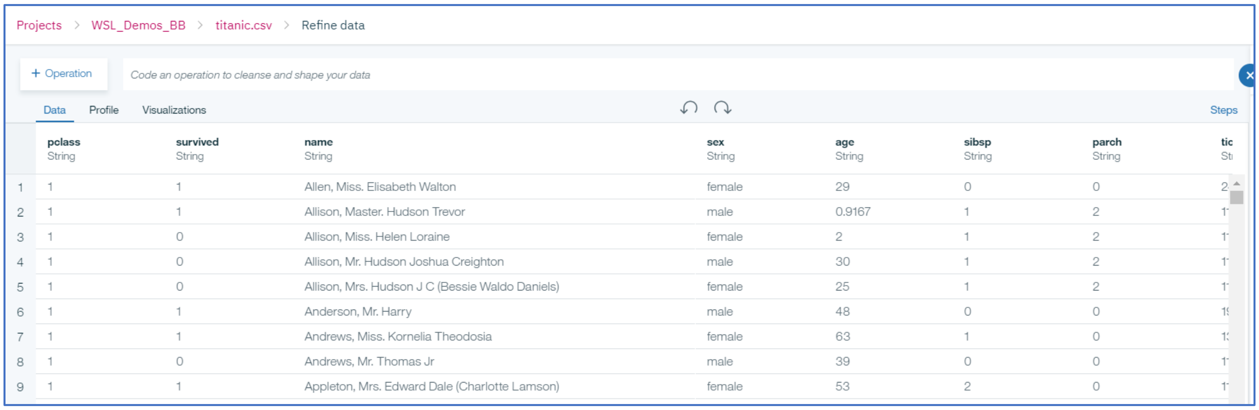


2. Click on titanic.csv.



Name	Type	Size	Data Source	Last Modified
titanic.csv	CSV	105.73 KB	Local file	9 Dec 2018, 7:51 PM
titanic_cleansed.csv	CSV	75.63 KB	Local file	9 Dec 2018, 4:31 PM
new_customer_scores.csv	CSV	118.75 KB	Local file	8 Dec 2018, 5:36 PM

3. The Data Refinery panel will display the data set.



Projects > WSL_Demos_BB > titanic.csv > Refine data

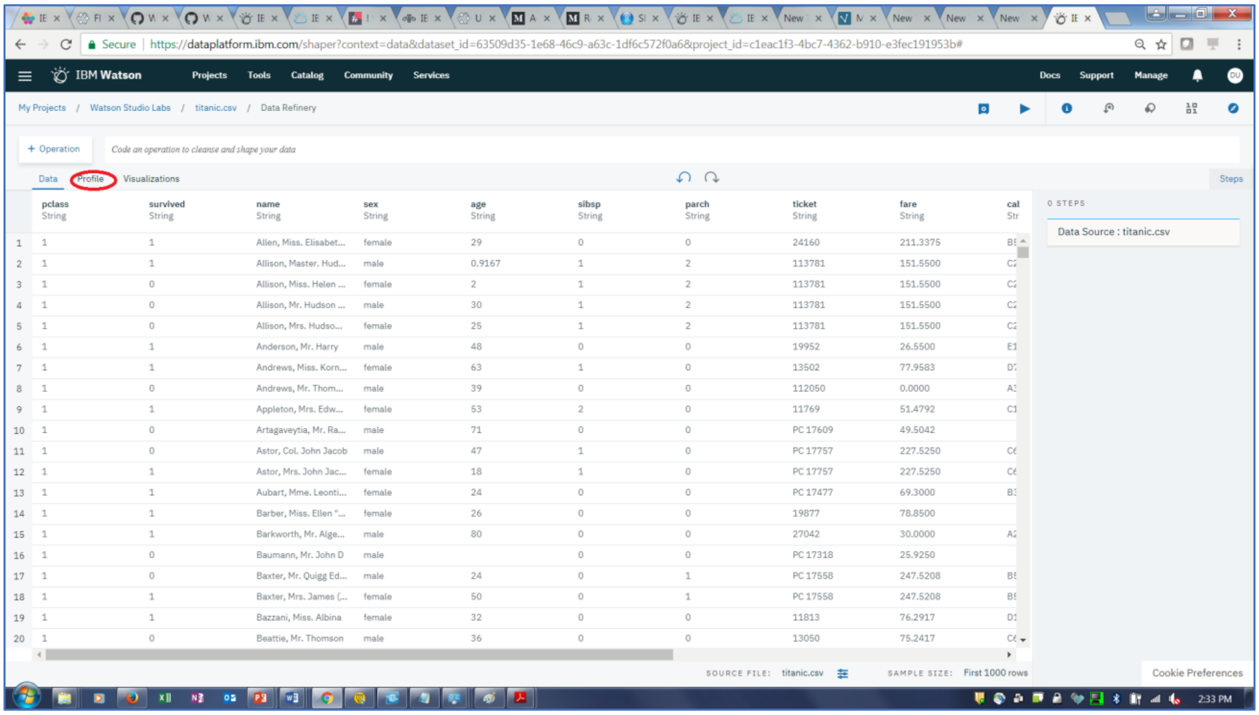
+ Operation | Code an operation to cleanse and shape your data

Data | Profile | Visualizations | Steps

	pclass String	survived String	name String	sex String	age String	sibsp String	parch String
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
5	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25	1	2
6	1	1	Anderson, Mr. Harry	male	48	0	0
7	1	1	Andrews, Miss. Kornelia Theodosia	female	63	1	0
8	1	0	Andrews, Mr. Thomas Jr	male	39	0	0
9	1	1	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53	2	0

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

1. Click on the **Profile** tab.



https://dataplatform.ibm.com/shaper?context=data&dataset_id=63509d35-1e68-46c9-a63c-1df6c572f0a6&project_id=c1eac1f3-4bc7-4362-b910-e3fec191953b#

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	parch String	ticket String	fare String
1	1	1	Allen, Miss. Elisabet...	female	29	0	0	24160	211.3375
2	1	1	Allison, Master. Hud...	male	0.9167	1	2	113781	151.5500
3	1	0	Allison, Miss. Helen ...	female	2	1	2	113781	151.5500
4	1	0	Allison, Mr. Hudson ...	male	30	1	2	113781	151.5500
5	1	0	Allison, Mrs. Hudso...	female	25	1	2	113781	151.5500
6	1	1	Anderson, Mr. Harry	male	48	0	0	19952	26.5500
7	1	1	Andrews, Miss. Korn...	female	63	1	0	13502	77.9583
8	1	0	Andrews, Mr. Thom...	male	39	0	0	112050	0.0000
9	1	1	Appleton, Mrs. Edw...	female	53	2	0	11769	51.4792
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
12	1	1	Astor, Mrs. John Jac...	female	18	1	0	PC 17757	227.5250
13	1	1	Aubart, Mme. Leonti...	female	24	0	0	PC 17477	69.3000
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
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
18	1	1	Baxter, Mrs. James (...	female	50	0	1	PC 17558	247.5208
19	1	1	Bazzani, Miss. Albina	female	32	0	0	11813	76.2917
20	1	0	Beattie, Mr. Thomson	male	36	0	0	13050	75.2417

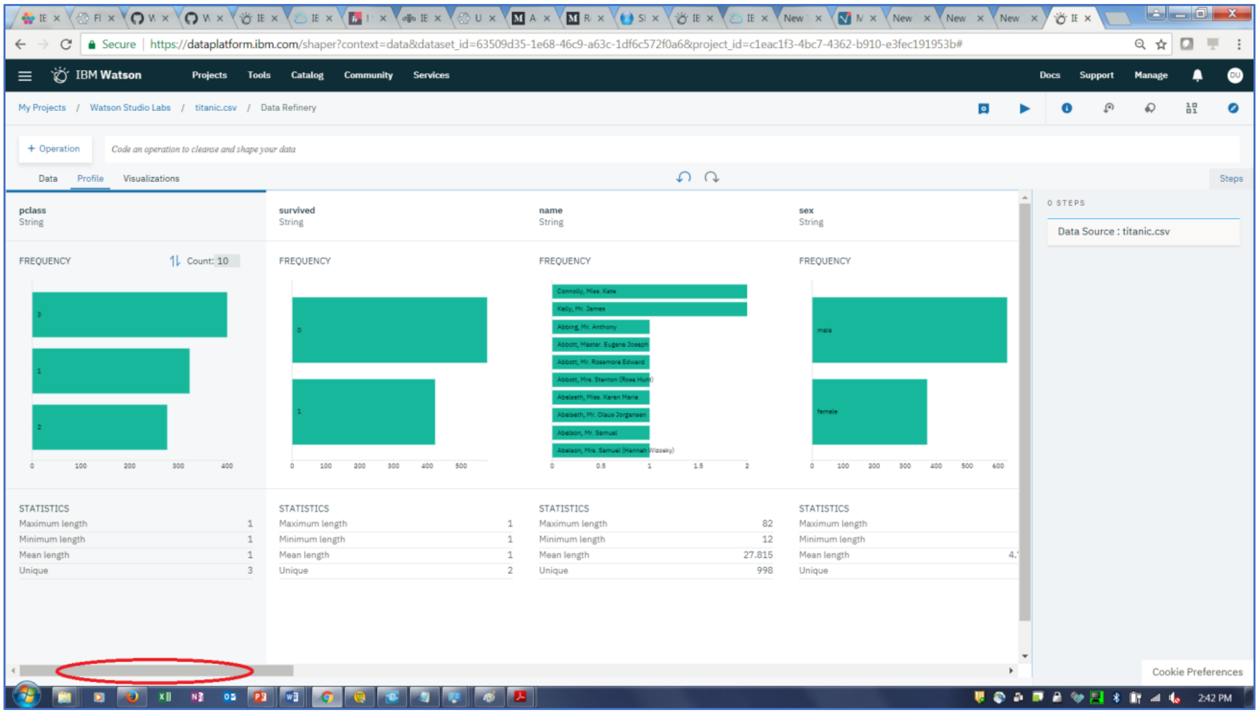
0 STEPS

Data Source : titanic.csv

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

Cookie Preferences

2. 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.


https://dataplatform.ibm.com/shaper?context=data&dataset_id=63509d35-1e68-46c9-a63c-1df6c572f0a6&project_id=c1eac1f3-4bc7-4362-b910-e3fec191953b#
IBM Watson
Projects
Tools
Catalog
Community
Services
Docs
Support
Manage
My Projects / Watson Studio Labs / titanic.csv / Data Refinery
+ Operation
Code an operation to cleanse and shape your data
Data
Profile
Visualizations
Steps
0 STEPS
Data Source : titanic.csv
pclass String
survived String
name String
sex String
FREQUENCY
Count: 10
Cookie Preferences
2:42 PM

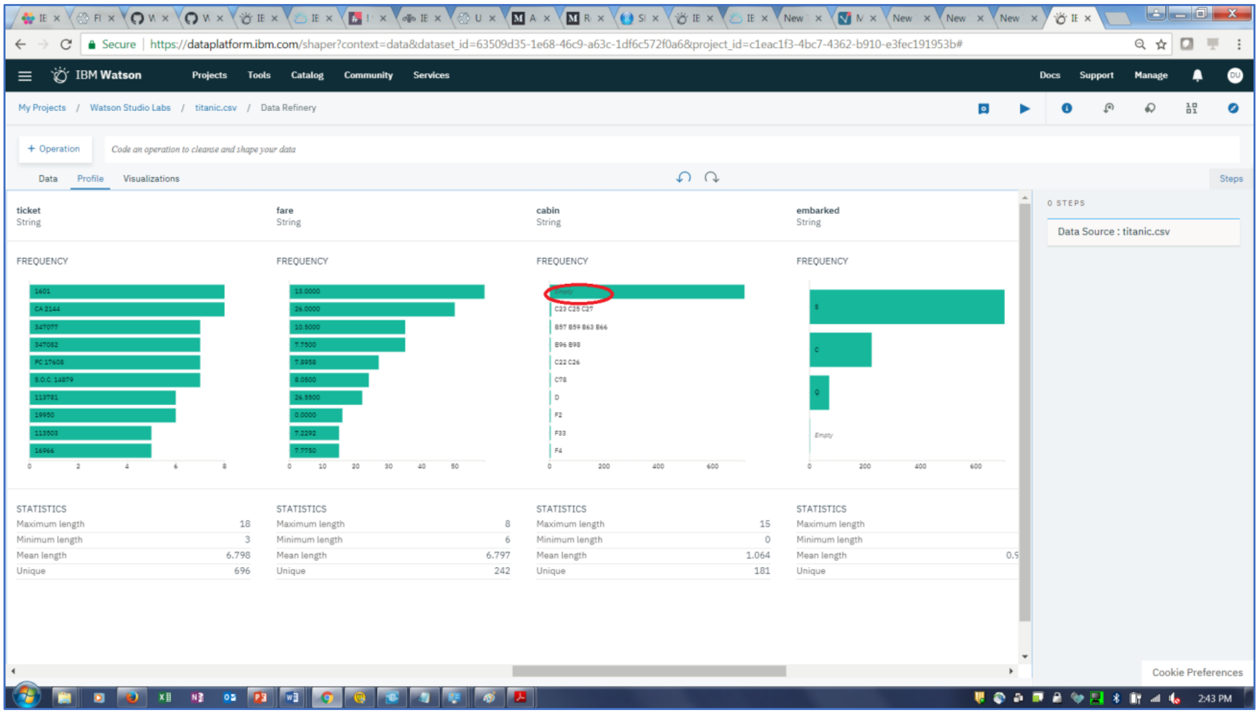
pclass STATISTICS	
Maximum length	1
Minimum length	1
Mean length	1
Unique	3

survived STATISTICS	
Maximum length	1
Minimum length	1
Mean length	1
Unique	2

name STATISTICS	
Maximum length	82
Minimum length	12
Mean length	27.815
Unique	998

sex STATISTICS	
Maximum length	
Minimum length	
Mean length	4.
Unique	

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


https://dataplatform.ibm.com/shaper?context=data&dataset_id=63509d35-1e68-46c9-a63c-1df6c572f0a6&project_id=c1eac1f3-4bc7-4362-b910-e3fec191953b#
IBM Watson
Projects
Tools
Catalog
Community
Services
Docs
Support
Manage
My Projects / Watson Studio Labs / titanic.csv / Data Refinery
+ Operation
Code an operation to cleanse and shape your data
Data
Profile
Visualizations
Steps
0 STEPS
Data Source : titanic.csv
ticket String
fare String
cabin String
embarked String
FREQUENCY
Empty
C23 C25 C27
B57 B59 B63 B66
B96 B98
C22 C26
C78
D
F2
F33
F4
S
C
Q
Cookie Preferences
2:43 PM

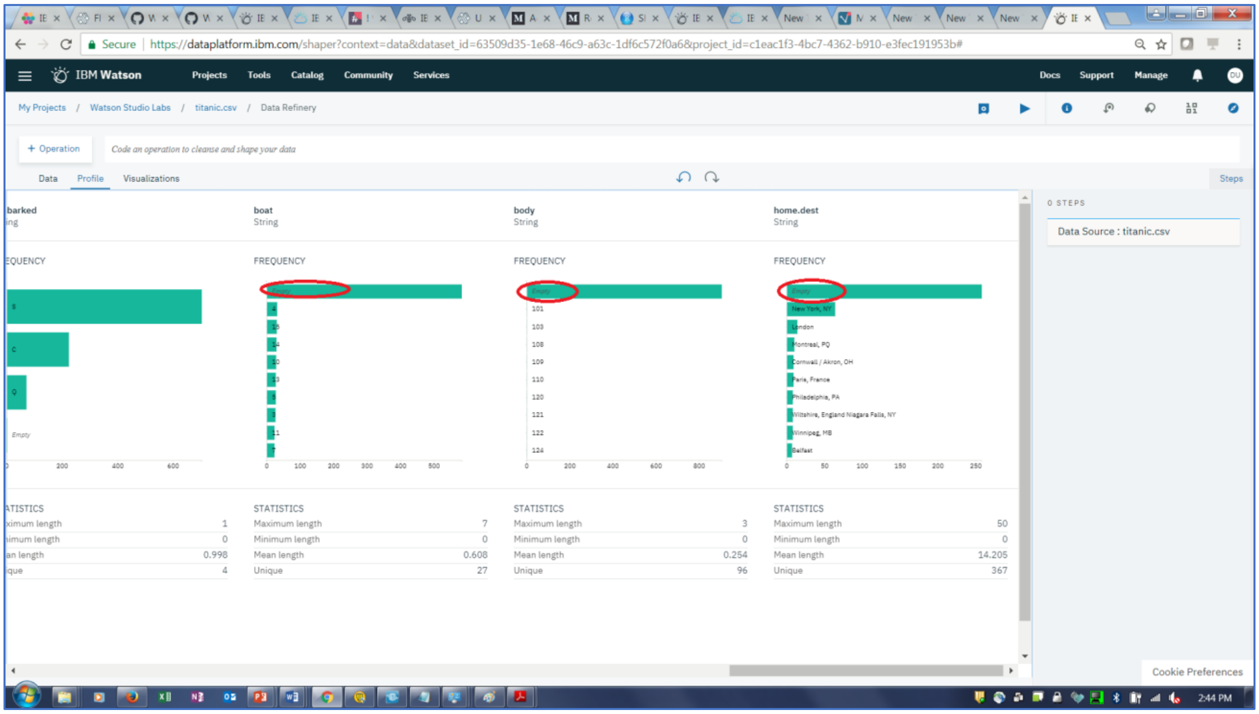
ticket STATISTICS	
Maximum length	18
Minimum length	3
Mean length	6.798
Unique	696

fare STATISTICS	
Maximum length	8
Minimum length	6
Mean length	6.797
Unique	242

cabin STATISTICS	
Maximum length	15
Minimum length	0
Mean length	1.064
Unique	181

embarked STATISTICS	
Maximum length	
Minimum length	
Mean length	0.5
Unique	

4. 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.



ATISTICS	
ximum length	1
imum length	0
an length	0.998
ique	4

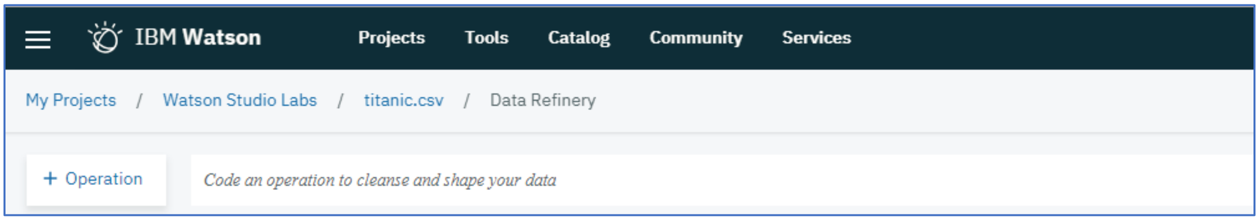
STATISTICS	
Maximum length	7
Minimum length	0
Mean length	0.608
Unique	27

STATISTICS	
Maximum length	3
Minimum length	0
Mean length	0.254
Unique	96

STATISTICS	
Maximum length	50
Minimum length	0
Mean length	14.205
Unique	367

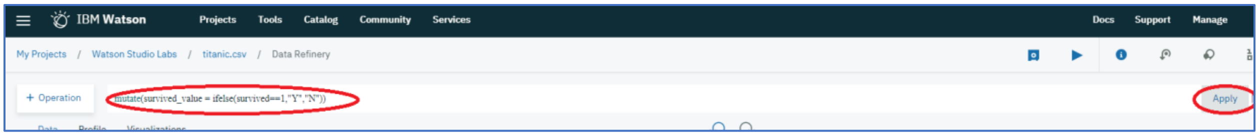
5. Age and Embarked also have missing values. Embarked has only 2 missing values. Age has over 100 missing values, but we will keep that column in the analysis. As part of data preparation, we will remove the rows that contain the missing age and embarked values.

6. 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.**


IBM Watson
Projects
Tools
Catalog
Community
Services
My Projects / Watson Studio Labs / titanic.csv / Data Refinery
+ Operation
Code an operation to cleanse and shape your data

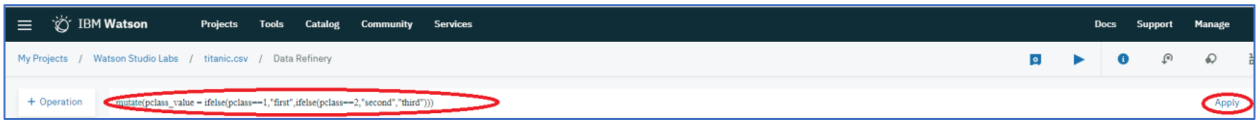
7. Type the following:
 mutate(survived_value=ifelse(survived==1, "Y", "N"))

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



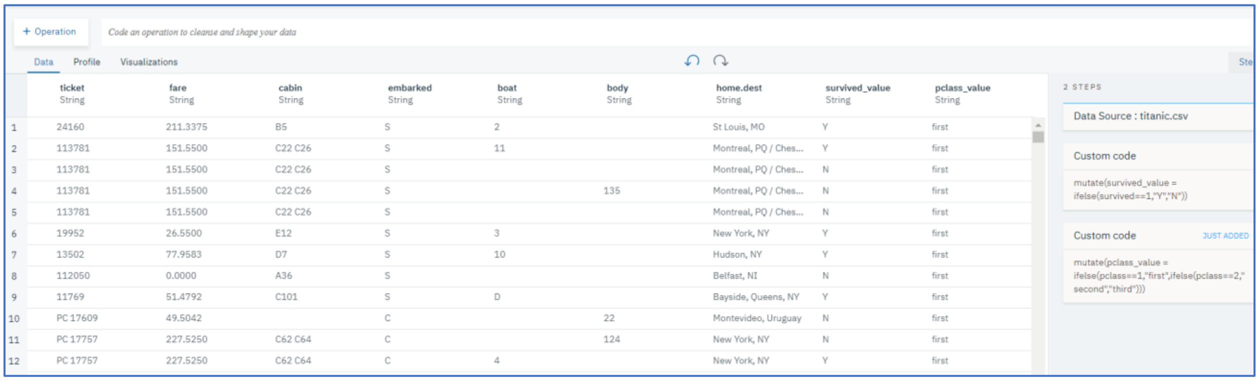
```
mutate(survived_value = ifelse(survived==1,"Y","N"))
```

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



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

9. 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	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

2 STEPS

Data Source : titanic.csv

Custom code

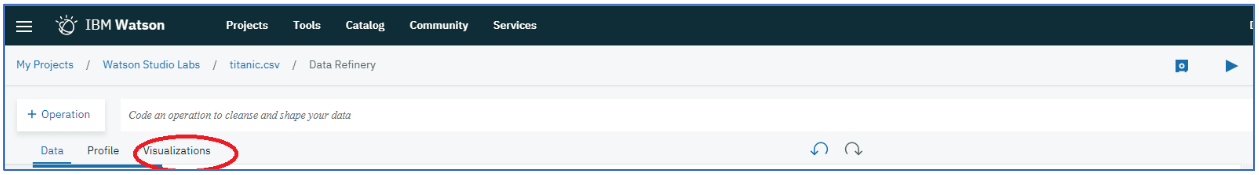
```
mutate(survived_value =
ifelse(survived==1,"Y","N"))
```

Custom code JUST ADDED

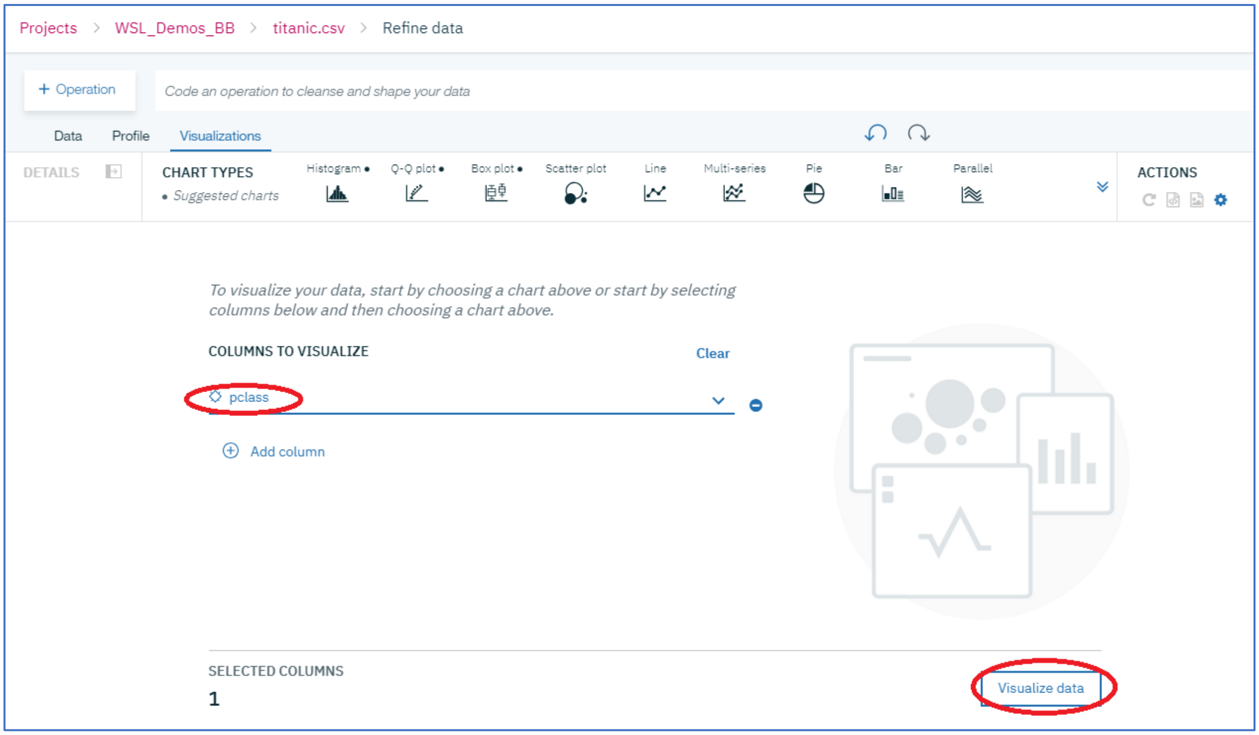
```
mutate(pclass_value =
ifelse(pclass==1,"first",ifelse(pclass==2,"
second","third")))
```

Step 3: Visualize the data to get a better understanding

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


IBM Watson
Projects
Tools
Catalog
Community
Services
My Projects / Watson Studio Labs / titanic.csv / Data Refinery
+ Operation
Code an operation to cleanse and shape your data
Data
Profile
Visualizations

2. 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**



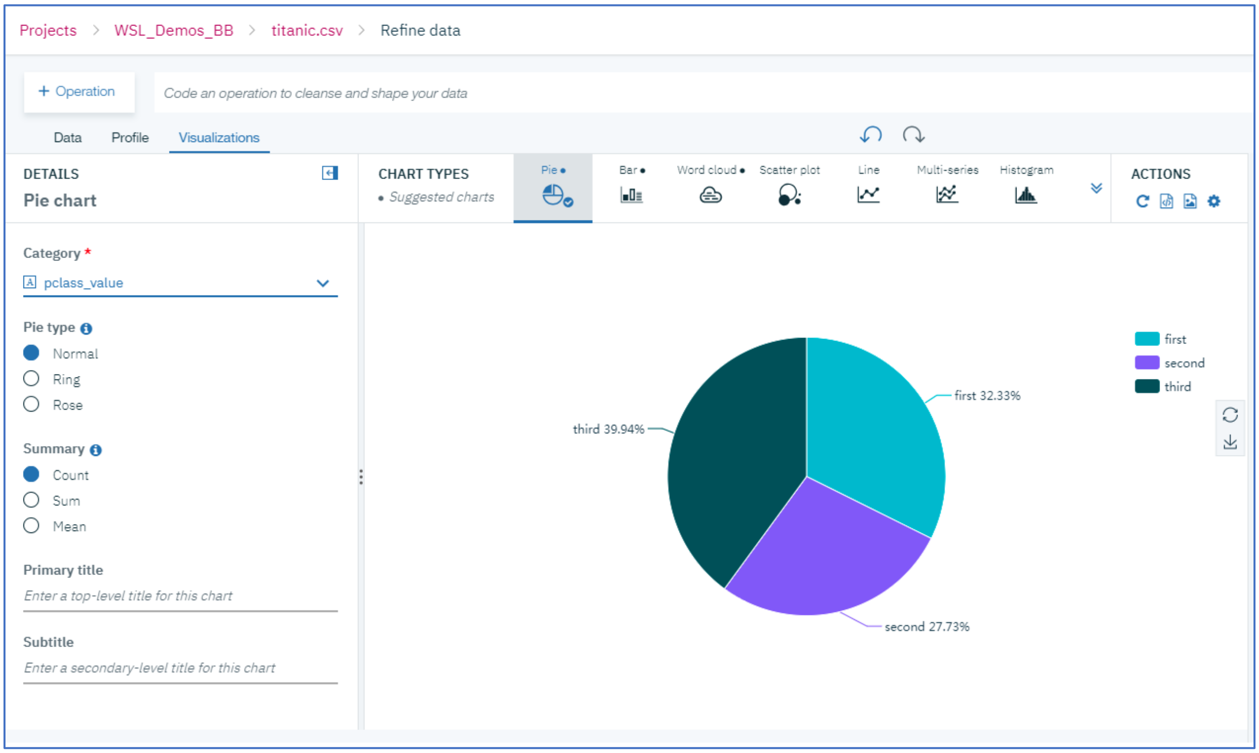
Projects > WSL_Demos_BB > titanic.csv > Refine data

To visualize your data, start by choosing a chart above or start by selecting columns below and then choosing a chart above.

COLUMNS TO VISUALIZE: pclass

SELECTED COLUMNS: 1

3. The result is shown below.



Projects > WSL_Demos_BB > titanic.csv > Refine data

DETAILS: Pie chart

Category *: pclass_value

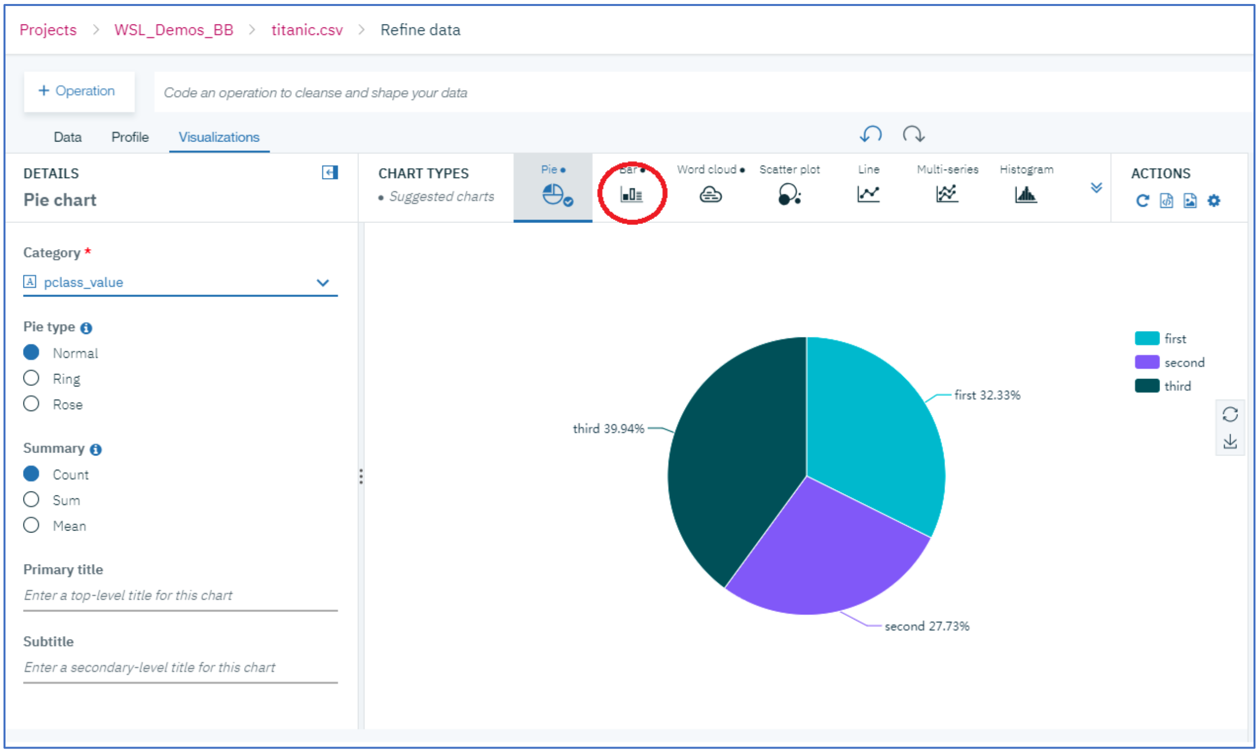
Pie type: ● Normal ○ Ring ○ Rose

Summary: ● Count ○ Sum ○ Mean

Primary title: *Enter a top-level title for this chart*

Subtitle: *Enter a secondary-level title for this chart*

4. We can switch this to a bar chart, by clicking on **Bar**.



Projects > WSL_Demos_BB > titanic.csv > Refine data

+ Operation

Code an operation to cleanse and shape your data

Data Profile Visualizations

DETAILS

Pie chart

Category *

pclass_value

Pie type

● Normal

○ Ring

○ Rose

Summary

● Count

○ Sum

○ Mean

Primary title

Enter a top-level title for this chart

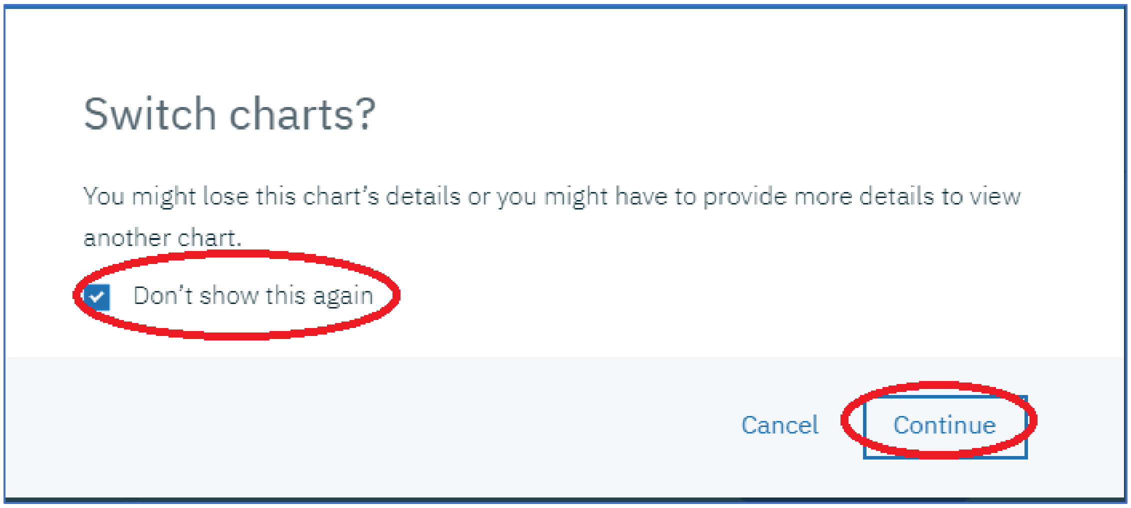
Subtitle

Enter a secondary-level title for this chart

CHART TYPES

• *Suggested charts*

5. A pop-up warning is displayed. Click the **Don't show this again checkbox** and click **Continue**.

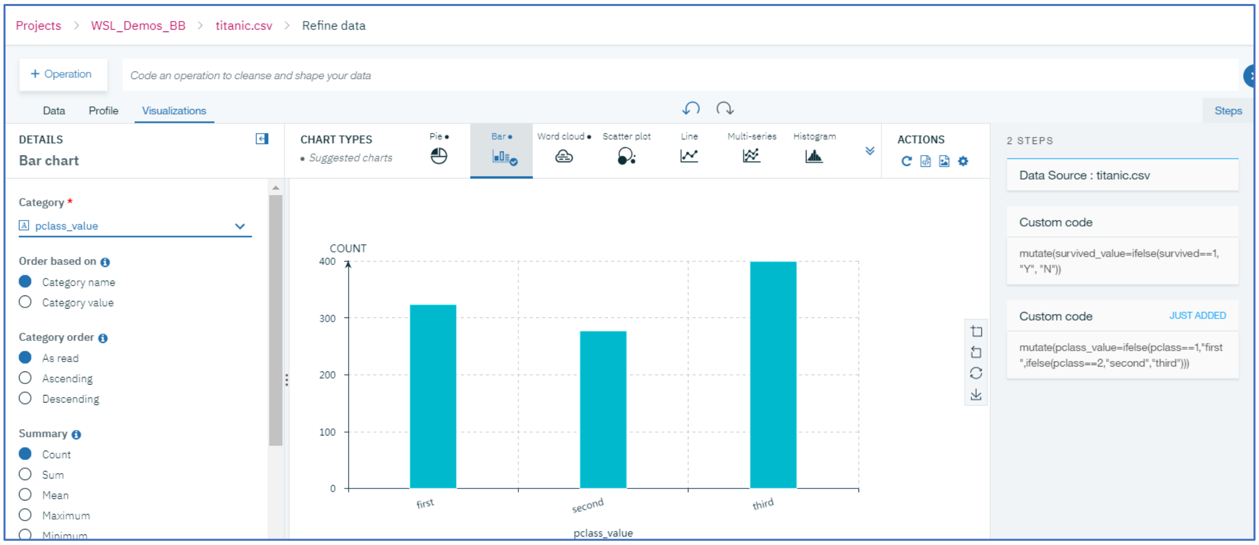


Switch charts?

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

☑ Don't show this again

6. The result is shown below.



DETAILS
Bar chart

Category *
pclass_value

Order based on
● Category name
○ Category value

Category order
● As read
○ Ascending
○ Descending

Summary
● Count
○ Sum
○ Mean
○ Maximum
○ Minimum

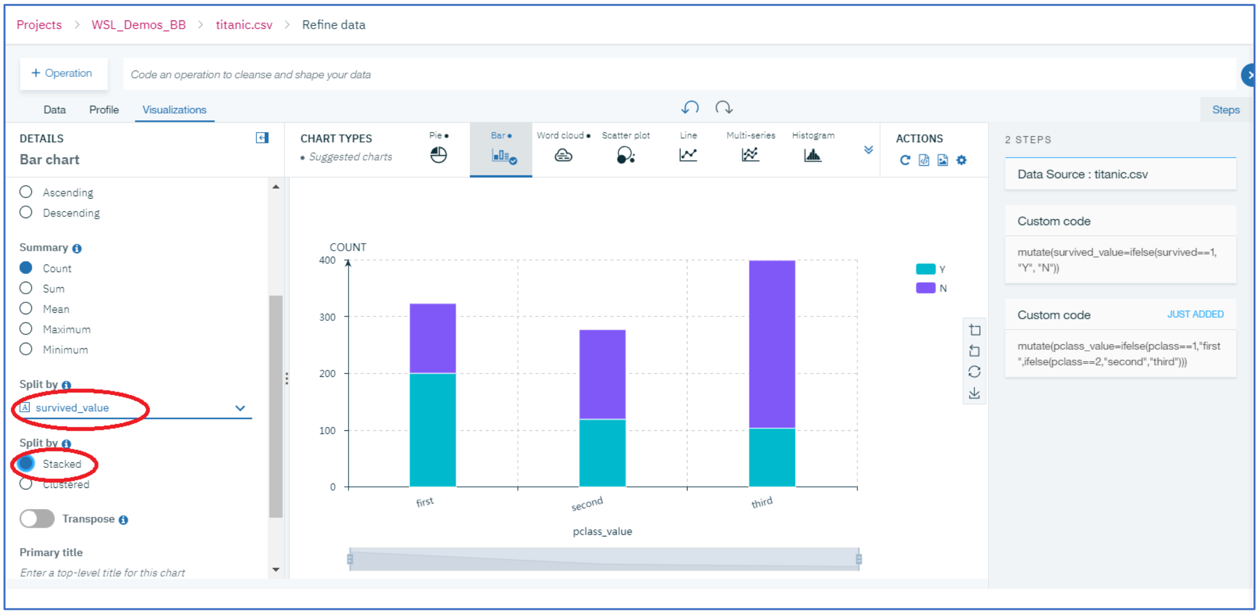
2 STEPS

Data Source : titanic.csv

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

Custom code JUST ADDED
mutate(pclass_value=ifelse(pclass==1,"first ",ifelse(pclass==2,"second","third")))

7. Let’s examine the relationship between survival and the passenger class. We will add the survived_value as the **Split by** column and change to **Stacked** view. The result is shown below. We can see that survival probability for first class customers is significantly better.



DETAILS
Bar chart

○ Ascending
○ Descending

Summary
● Count
○ Sum
○ Mean
○ Maximum
○ Minimum

Split by
survived_value

Split by
● Stacked
○ Clustered

Transpose

Primary title
Enter a top-level title for this chart

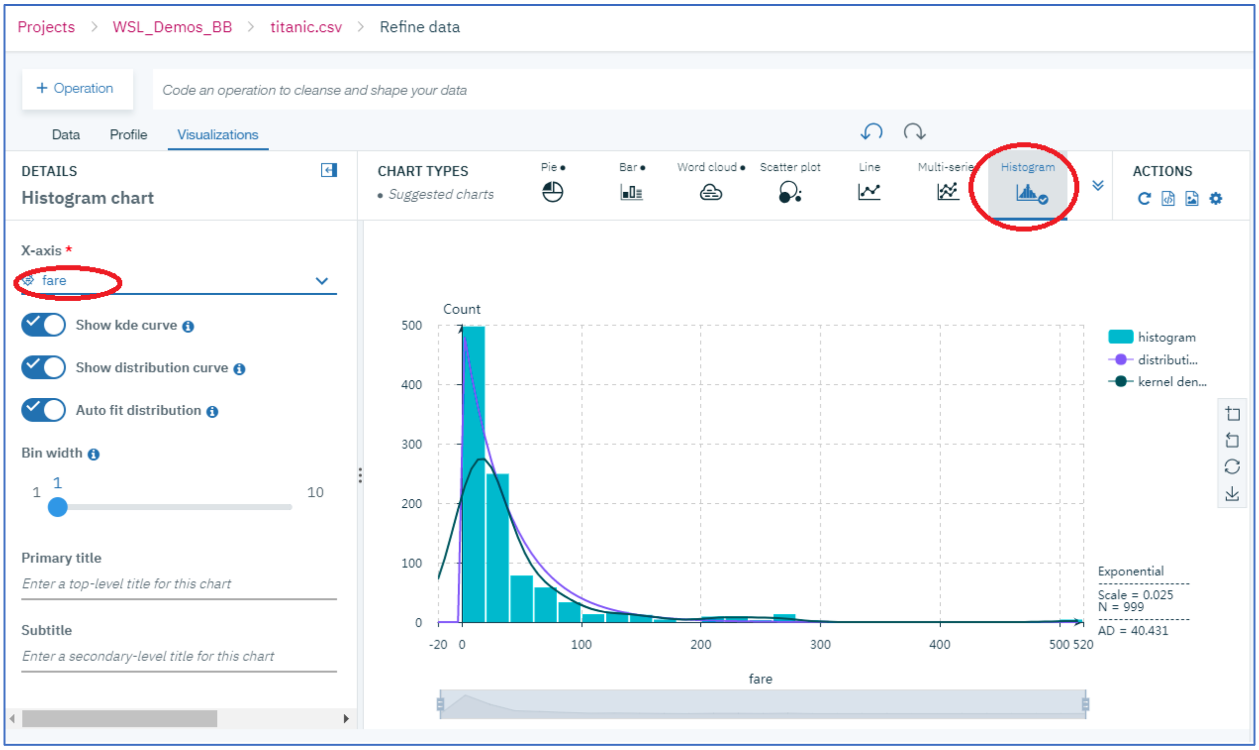
2 STEPS

Data Source : titanic.csv

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

Custom code JUST ADDED
mutate(pclass_value=ifelse(pclass==1,"first ",ifelse(pclass==2,"second","third")))

8. 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.



X-axis *: fare

Show kde curve: on

Show distribution curve: on

Auto fit distribution: on

Bin width: 1 (1–10)

Primary title: *Enter a top-level title for this chart*

Subtitle: *Enter a secondary-level title for this chart*

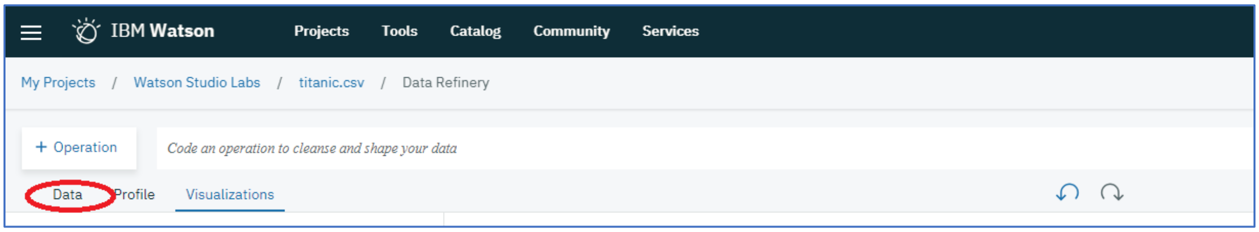
Step 4: Prepare the data for modeling

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

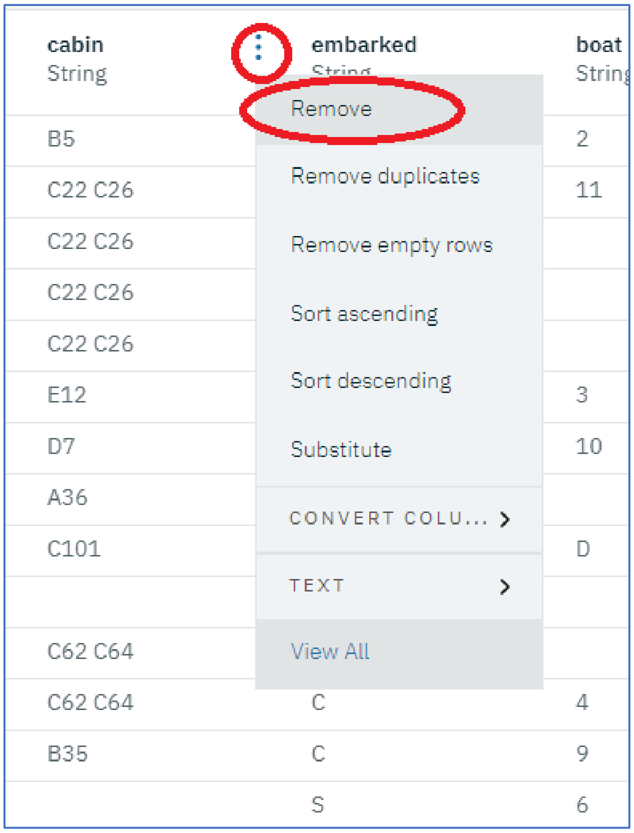
1. Remove columns cabin, boat, body, home.dest
2. Remove rows with missing values of age, and embarked.
3. Create a new column(log_fare) that is the logarithm of the fare column

We will also bin the age, and log_fare fields.

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

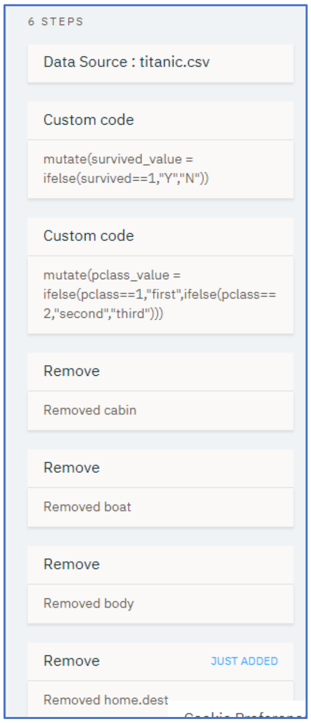

IBM Watson
Projects
Tools
Catalog
Community
Services
My Projects / Watson Studio Labs / titanic.csv / Data Refinery
+ Operation
Code an operation to cleanse and shape your data
Data
Profile
Visualizations

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

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

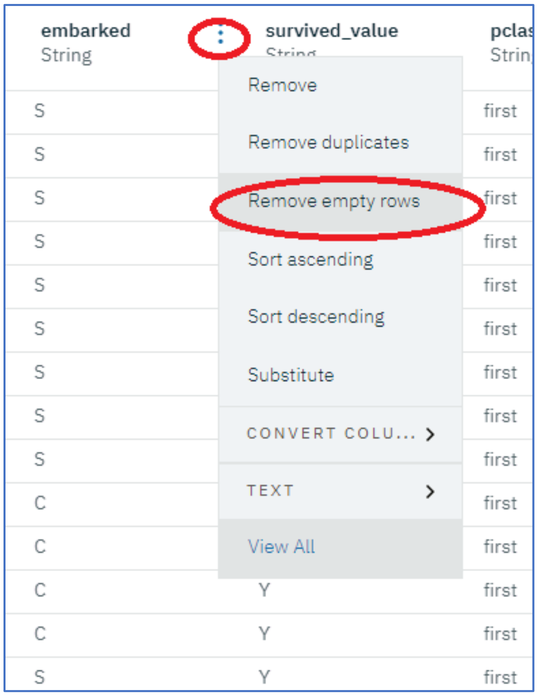
Remove

Removed body

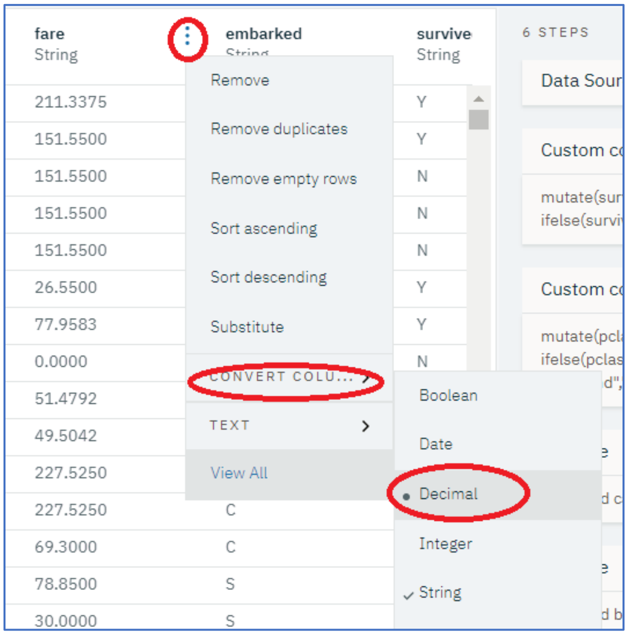
Remove JUST ADDED

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**.

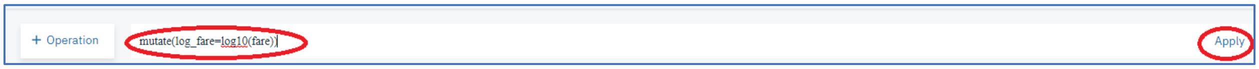


5. Convert the fare column from a String to a Decimal by clicking on the vertical ellipse adjacent to the column, click on **Convert Column**, and then click on **Decimal**.



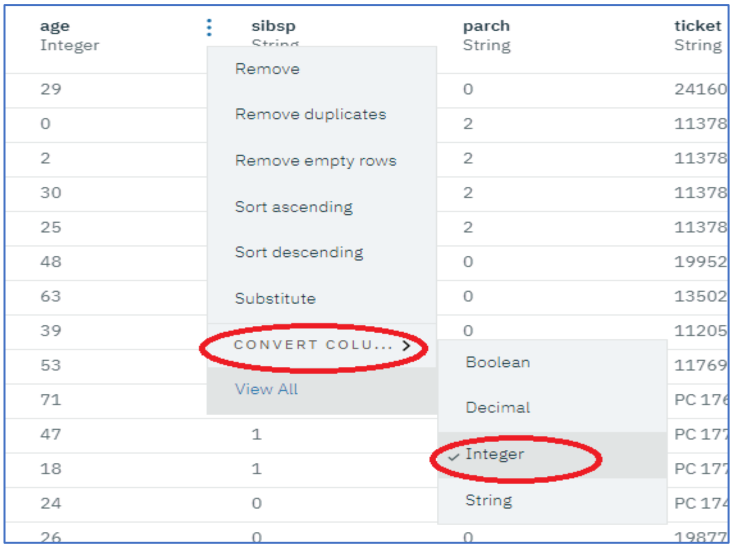
6. Create a new column that is the log to the base 10 of the fare by clicking into the **Code an operation to cleanse and shape your data,** and entering mutate(log_fare=log10(fare))

then click **Apply**.



```
mutate(log_fare=log10(fare))
```

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**.



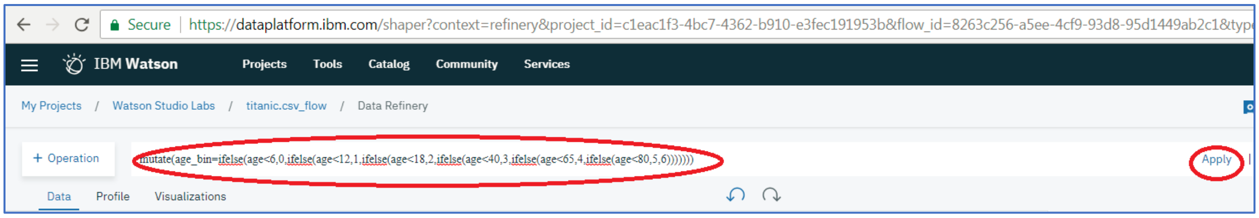
age Integer	sibsp String	parch String	ticket String
29		0	24160
0		2	11378
2		2	11378
30		2	11378
25		2	11378
48		0	19952
63		0	13502
39		0	11205
53			11769
71			PC 176
47	1		PC 177
18	1		PC 177
24	0		PC 174
26	0	0	19877

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

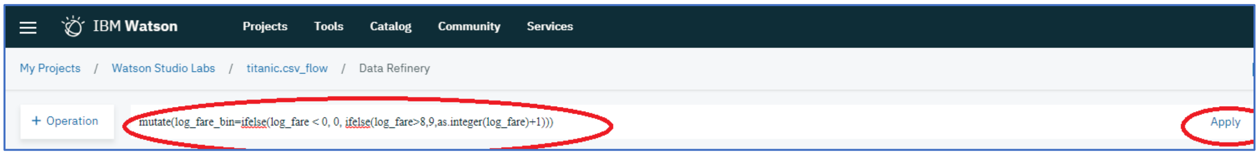

Secure
https://dataplatform.ibm.com/shaper?context=refinery&project_id=c1eac1f3-4bc7-4362-b910-e3fec191953b&flow_id=8263c256-a5ee-4cf9-93d8-95d1449ab2c1&typ
IBM Watson
Projects
Tools
Catalog
Community
Services
My Projects / Watson Studio Labs / titanic.csv_flow / Data Refinery
+ Operation
Apply
Data
Profile
Visualizations

```
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)))))))
```

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

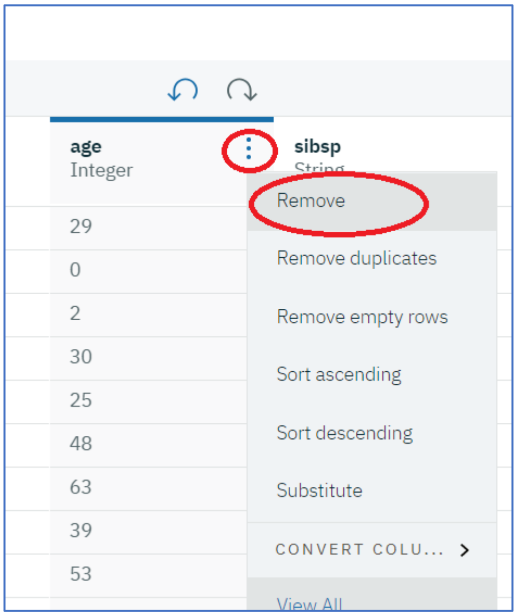
   mutate(log_fare_bin=ifelse(log_fare<0,0,ifelse(log_fare>8,9,as.integer(log_fare)+1)))

   and then clicking **Apply**

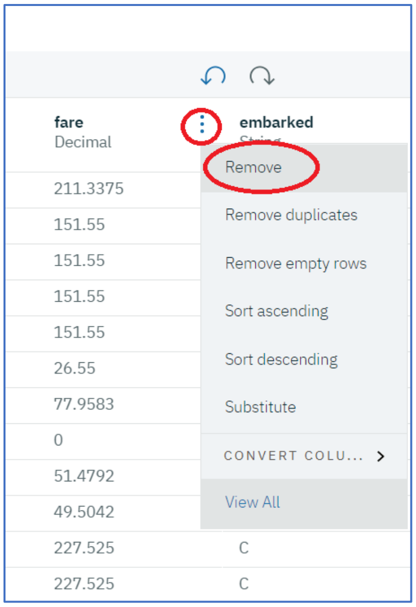

IBM Watson
Projects
Tools
Catalog
Community
Services
My Projects / Watson Studio Labs / titanic.csv_flow / Data Refinery
+ Operation
Apply

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

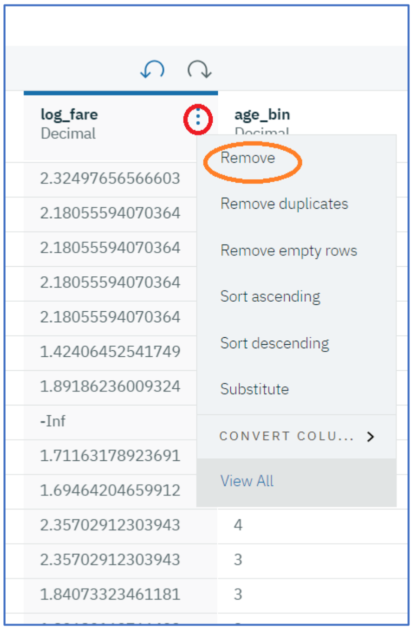
10. Now we will drop the age, fare, and log_fare columns as they are no longer needed for modeling purposes. Select the vertical ellipse adjacent to the column and click on **Remove** as shown below.



29
0
2
30
25
48
63
39
53

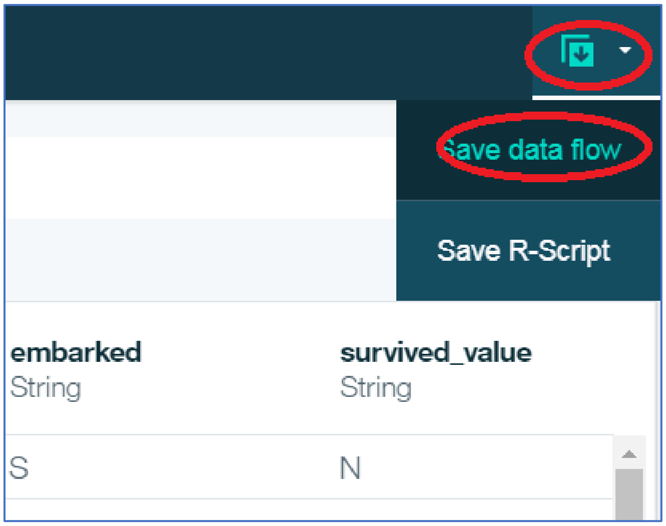


fare Decimal	embarked String
211.3375	
151.55	
151.55	
151.55	
151.55	
26.55	
77.9583	
0	
51.4792	
49.5042	
227.525	C
227.525	C



log_fare Decimal	age_bin 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

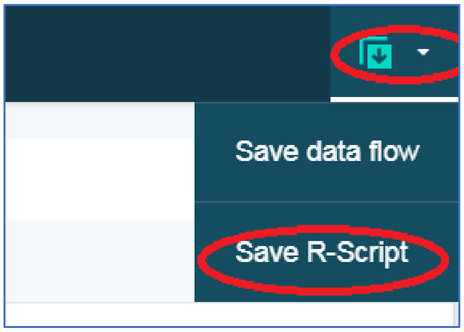
11. Save the Data Flow by clicking on the Save Data Flow icon .



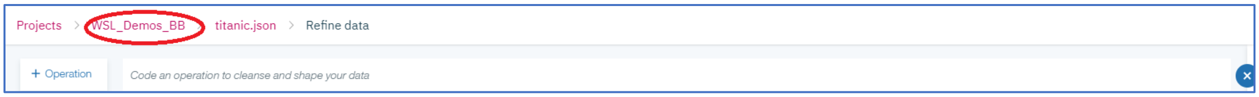
Step 5: Run the sequence of Data Flow operations on the entire data set.

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 (titanic.csv), the user will generate an **R-Script** and run a **Job**.

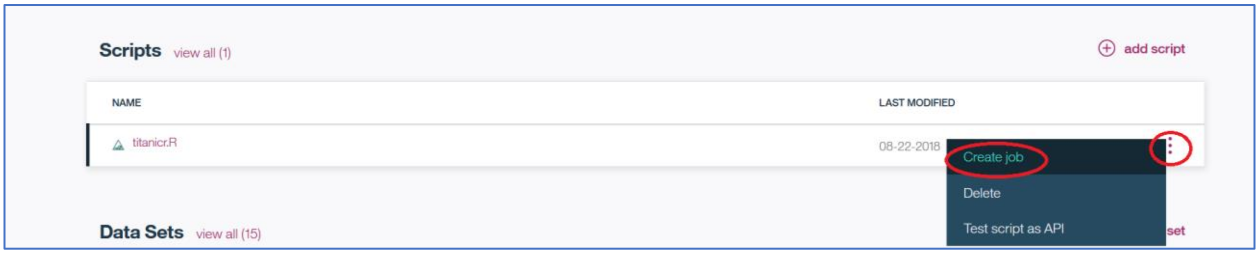
1. Click on the ⬇ icon to **Save R-Script**.



2. Navigate to the project **Scripts** by clicking on the WSL_Demos_XX link and then clicking on **Scripts**.

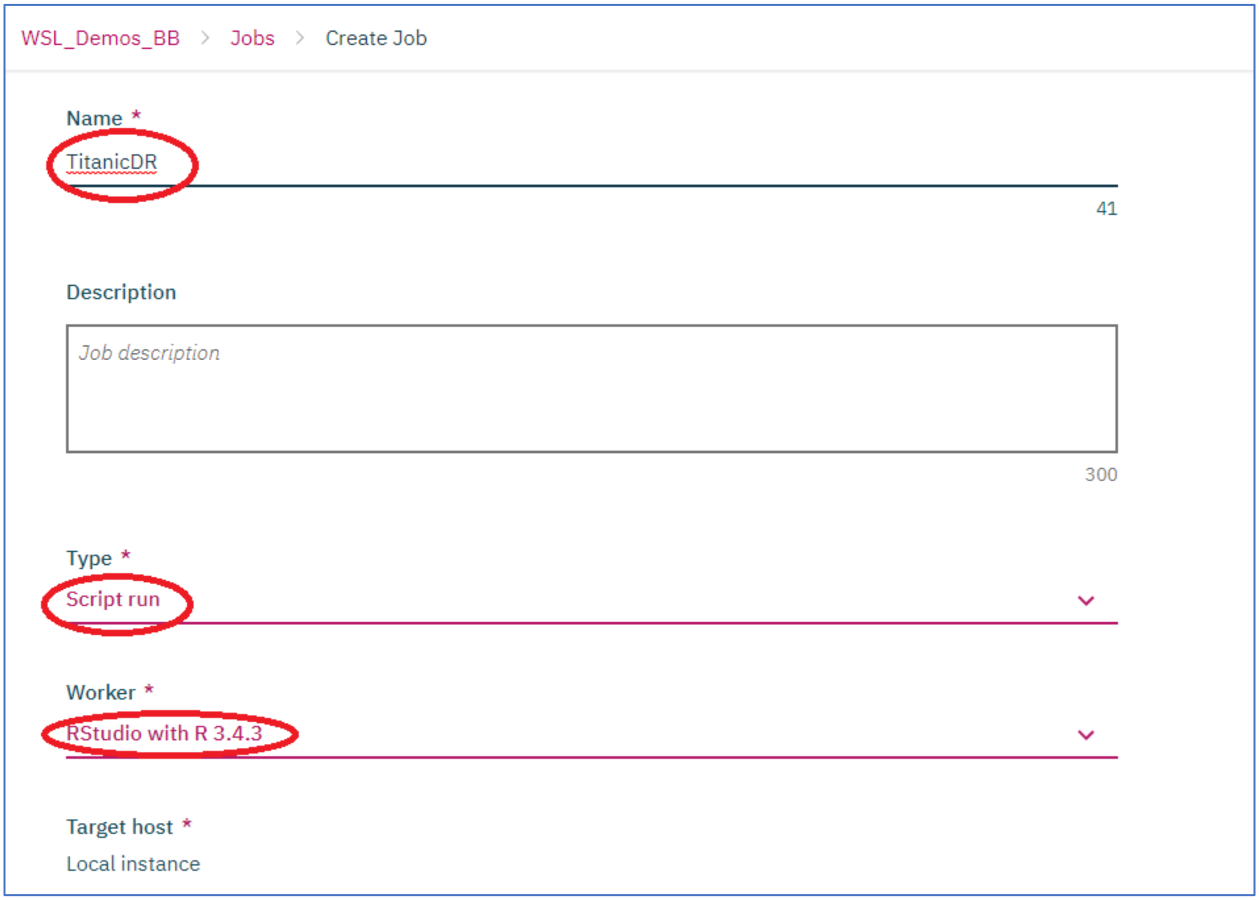

Projects > WSL_Demos_BB titanic.json > Refine data
+ Operation
Code an operation to cleanse and shape your data

3. Click on the vertical ellipse for **titanic.R** script, and click on **Create Job**.



NAME	LAST MODIFIED
titanicr.R	08-22-2018

4. Enter a **Name**, select **Script Run** for **Type,** select **RStudio with R 3.4.3** for **Worker**, select the **Titanic.R** script for **Source asset**, enter for **arg1** the following **input=datasets/titanic.csv**, click + and enter for **arg2** the following **output=datasets/titanic_shaped_out.csv,** and then click **Create**.



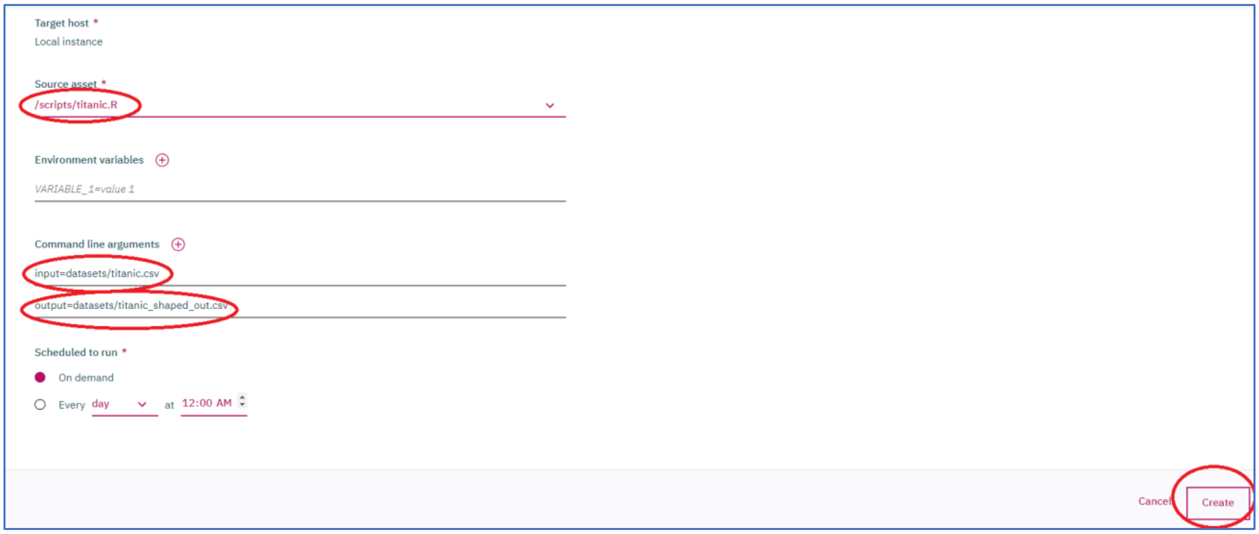
Name *
TitanicDR
41

Description
Job description
300

Type *
Script run

Worker *
RStudio with R 3.4.3

Target host *
Local instance



Target host *
Local instance

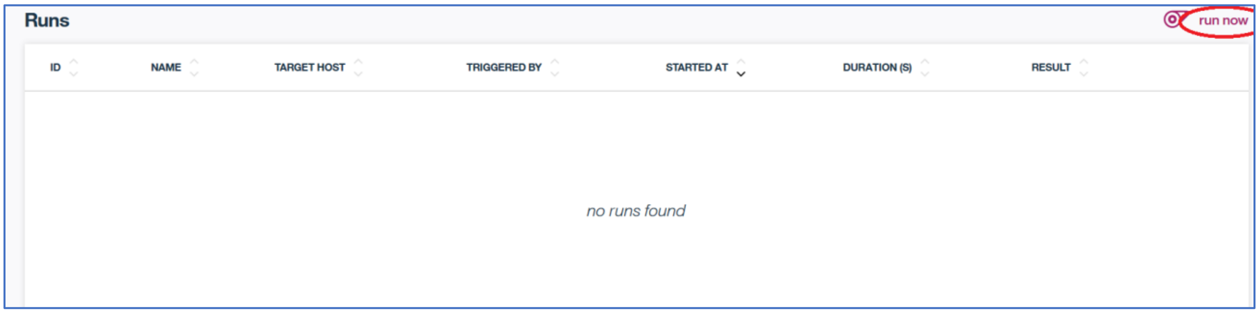
Source asset *
/scripts/titanic.R

Environment variables
VARIABLE_1=value 1

Command line arguments
input=datasets/titanic.csv
output=datasets/titanic_shaped_out.csv

Scheduled to run *
● On demand
○ Every day at 12:00 AM

5. Scroll down to the **Runs** section and click **run now.**

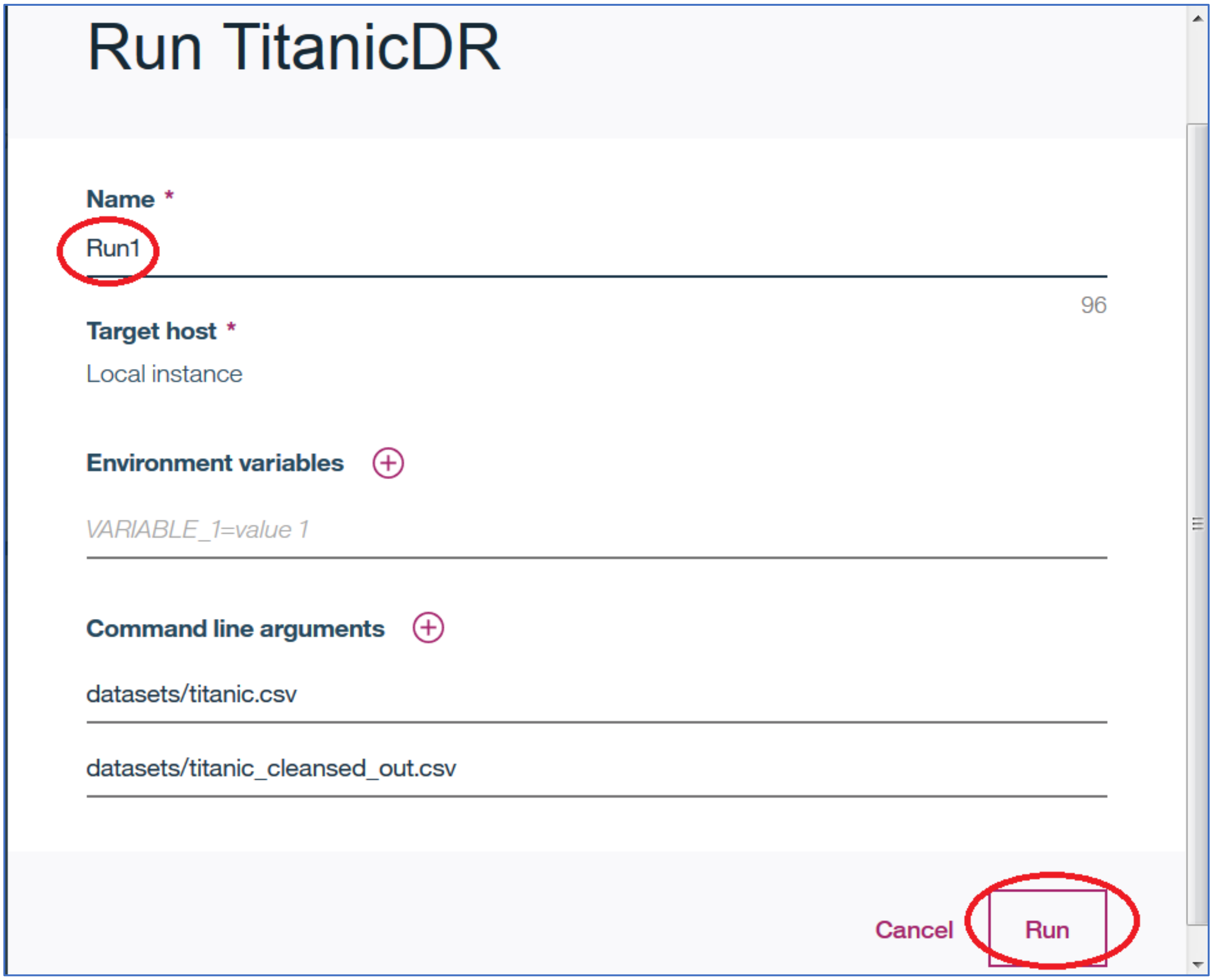


Runs

ID	NAME	TARGET HOST	TRIGGERED BY	STARTED AT	DURATION (S)	RESULT

no runs found

6. Enter a Name for the Run, and then click **Run**.



Run TitanicDR

Name *

Run1

96

Target host *

Local instance

Environment variables

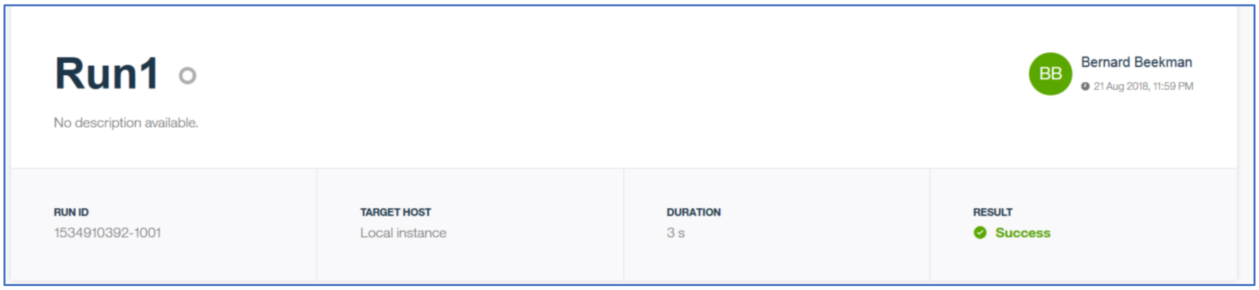
VARIABLE_1=value 1

Command line arguments

datasets/titanic.csv

datasets/titanic_cleansed_out.csv

7. Wait for the script to complete.



Run1

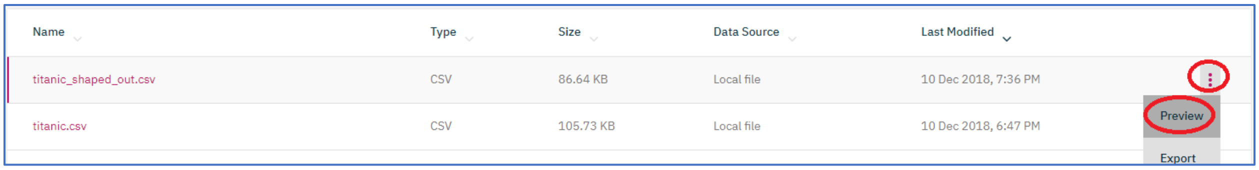
No description available.

BB Bernard Beekman

21 Aug 2018, 11:59 PM

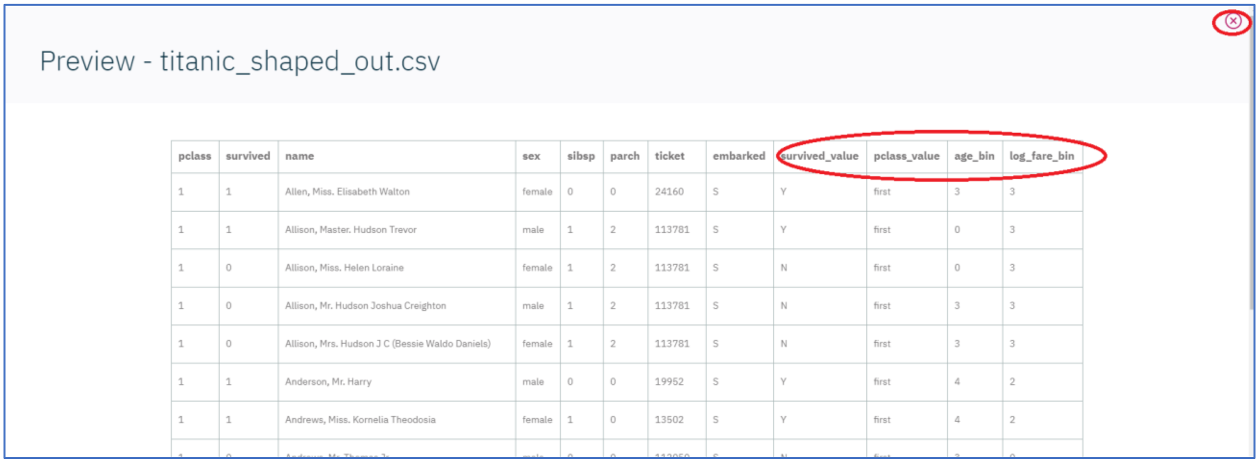
RUN ID	TARGET HOST	DURATION	RESULT
1534910392-1001	Local instance	3 s	Success

8. When the script completes navigate to Data Sets by clicking on the **WSL_Demos_XX** project link and then click on **Data Sets**. Notice that the **titanic_shaped_out.csv** data set has been created. Click on the vertical ellipse on the right side and click on **Preview**.



Name	Type	Size	Data Source	Last Modified
titanic_shaped_out.csv	CSV	86.64 KB	Local file	10 Dec 2018, 7:36 PM
titanic.csv	CSV	105.73 KB	Local file	10 Dec 2018, 6:47 PM

9. We can see that the new fields defined in the Data Refinery flow have been created. Click on the “x” to close the window.



Preview - titanic_shaped_out.csv

pclass	survived	name	sex	sibsp	parch	ticket	embarked	survived_value	pclass_value	age_bin	log_fare_bin
1	1	Allen, Miss. Elisabeth Walton	female	0	0	24160	S	Y	first	3	3
1	1	Allison, Master. Hudson Trevor	male	1	2	113781	S	Y	first	0	3
1	0	Allison, Miss. Helen Loraine	female	1	2	113781	S	N	first	0	3
1	0	Allison, Mr. Hudson Joshua Creighton	male	1	2	113781	S	N	first	3	3
1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	1	2	113781	S	N	first	3	3
1	1	Anderson, Mr. Harry	male	0	0	19952	S	Y	first	4	2
1	1	Andrews, Miss. Kornelia Theodosia	female	1	0	13502	S	Y	first	4	2
1	0	Andrews, Mr. Thomas Jr	male	0	0	112050	S	N	first	3	0