

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:

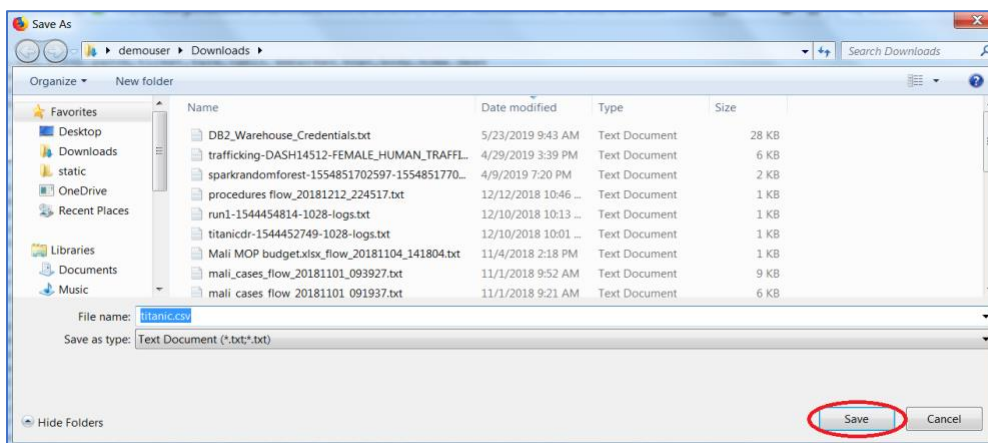
1. Use the Data Refinery Tool to:
 - a. Profile the data to help determine missing values
 - b. Visualize the data to gain a better understanding
 - c. Prepare the data for modeling
 - d. Run the sequence of data preparation operations on the entire data set.

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 ...



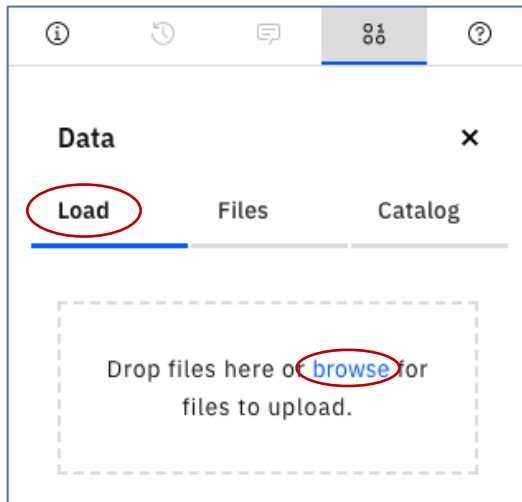
3. Click on **Save** to save the titanic.csv file (Note, if the file shown is titanic.csv.txt, remove the .txt).



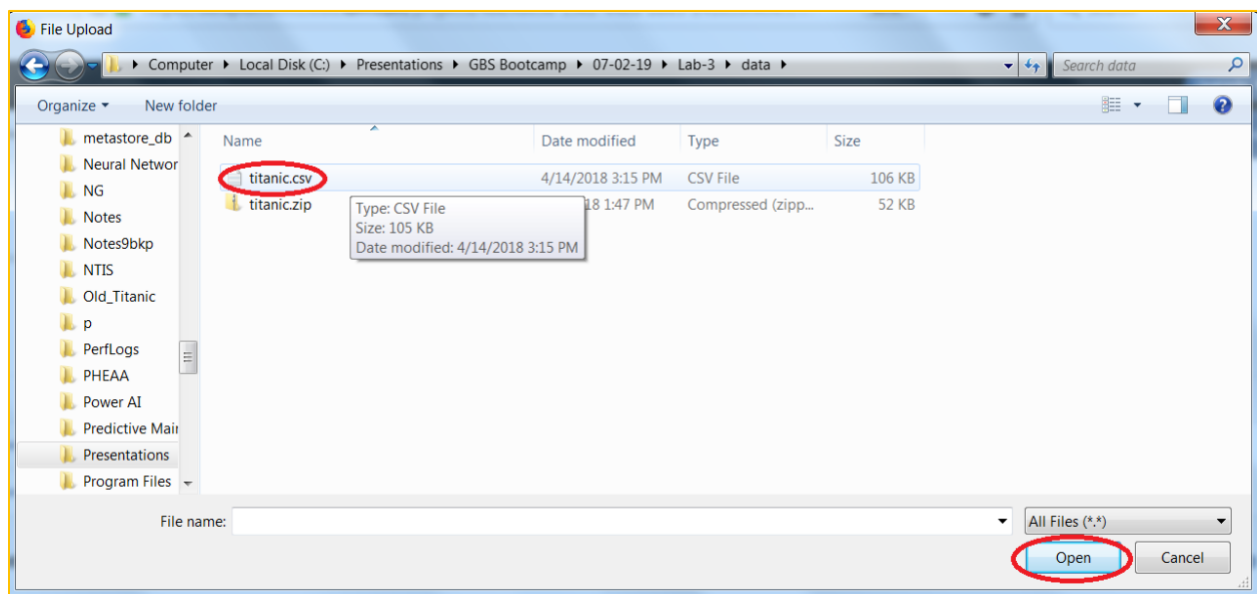
4. Go back to your Watson Studio Labs project. Click on the  icon.



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				
0 assets selected.				
<input type="checkbox"/>	Name	Type	Created by	Last modified
<input type="checkbox"/>	csv titanic.csv	Data Asset	FCTO Labs	Jul 19, 2020, 07:38 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**.

My projects / Watson Studio Labs				
Overview	Assets	Environments	Jobs	Deployments
		Access Control	Settings	

Choose asset type

Available asset types

Data

Connection

Connected data

AutoAI experiment

Notebook

Dashboard

Visual Recognition ...

Natural Language CL...

Watson Machine Lea...

Deep learning experi...

Modeler flow

Data Refinery flow

Streams flow

Decision Optimizatio...

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

My Projects / Watson Studio Labs / Refine data	
Watson Studio Labs	Data assets
Assets (2)	Data assets (1)
Connections	titanic.csv
Data assets	
Cancel Add	

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

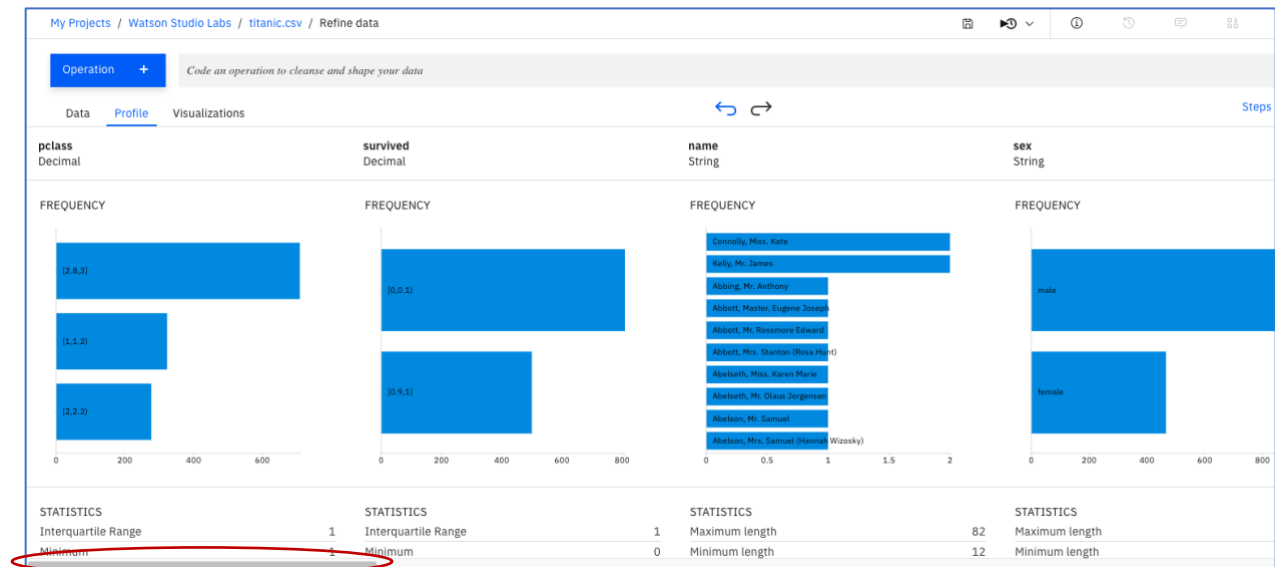
My Projects / Watson Studio Labs / titanic.csv / Refine data

Operation + Code an operation to cleanse and shape your data

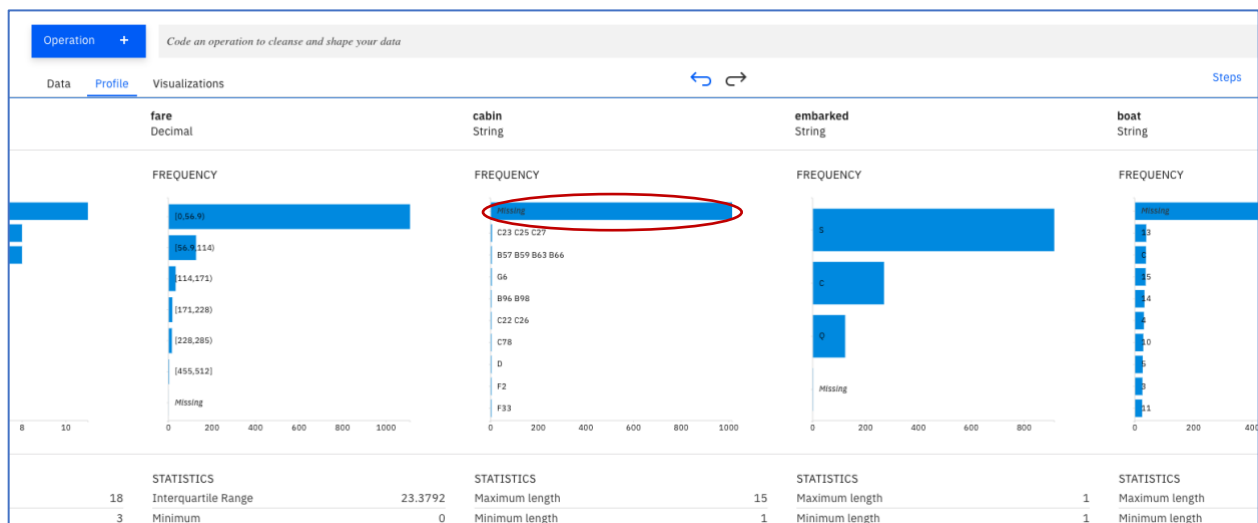
Data **Profile** Visualizations

	pclass Decimal	survived Decimal	name String	sex String	age Decimal	sibsp Decimal	parch Decimal
1	1	1	Allen, Miss. Elisabeth Walton	female	29	0	0
2	1	1	Allison, Master. Hudson Trevor	male	0.9167	1	2

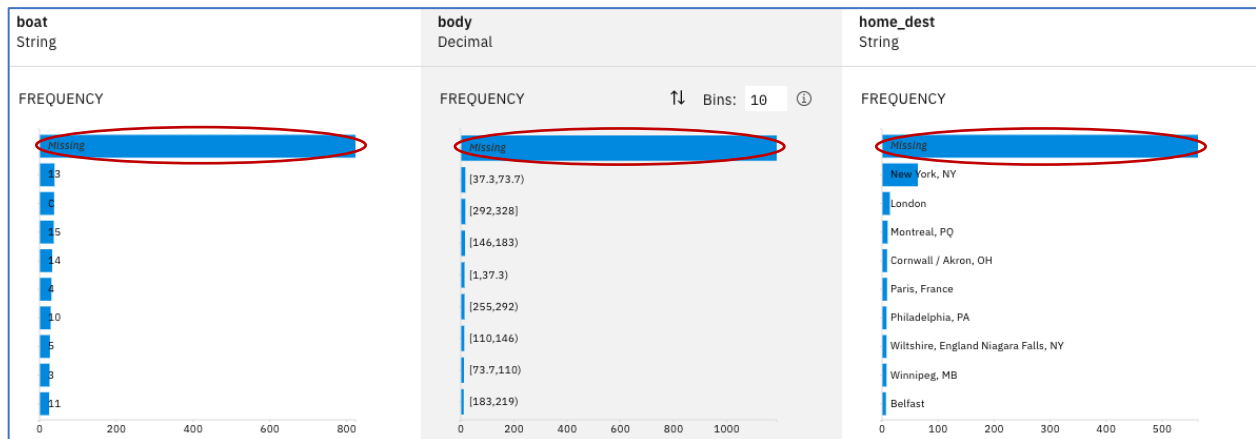
4. The Profile panel displays the counts of the top 10 count values for each column. Note that you can change 10 to another number if desired. You can also switch to the bottom 10 counts for a column. Scroll to the right to view the cabin column.



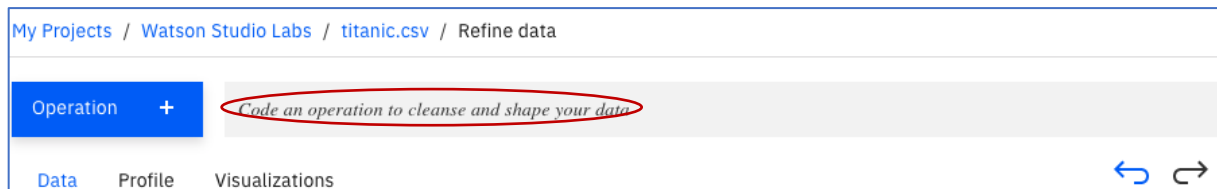
5. Note that the cabin column has many missing values and should be removed as part of the data preparation step.



- 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.



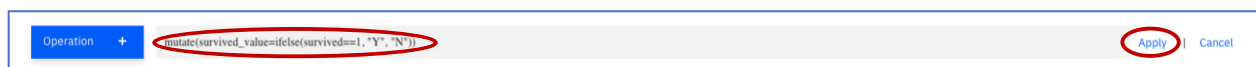
- Age and Embarked also have missing values. Embarked has very few missing values. Age has over 100 missing values, but we will keep that column in the analysis. As part of data preparation, we will remove the rows that contain the missing age and embarked values.
- Click on the **Data** tab. We will 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:

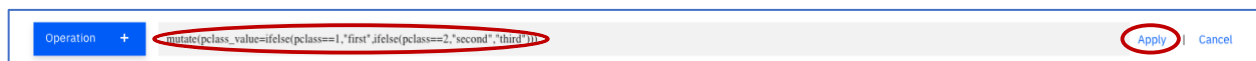
```
mutate(survived_value=ifelse(survived==1, "Y", "N"))
```

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")))
```



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

Operation + Code an operation to cleanse and shape your data

Data Profile Visualizations

	pclass Decimal	survived Decimal	name String	sex String	age Decimal
1	1	1	Allen, Miss. Elisabeth Walton	female	29
2	1	1	Allison, Master. Hudson Trevor	male	0.9167
3	1	0	Allison, Miss. Helen Loraine	female	2
4	1	0	Allison, Mr. Hudson Joshua Creighton	male	30
5	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25
6	1	1	Anderson, Mr. Harry	male	48
7	1	1	Andrews, Miss. Kornelia Theodosia	female	63
8	1	0	Andrews, Mr. Thomas Jr	male	39
9	1	1	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53
10	1	0	Artagaveytia, Mr. Ramon	male	71
11	1	0	Astor, Col. John Jacob	male	47
12	1	1	Astor, Mrs. John Jacob (Madeleine Talmadge Force)	female	18
13	1	1	Aubart, Mme. Leontine Pauline	female	24
14	1	1	Barber, Miss. Ellen "Nellie"	female	26

SOURCE FILE: titanic.csv SAMPLE SIZE: First 1309 rows

3 Steps

Data Source

titanic.csv

Convert column type AUTOMATIC

Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol.

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.

My Projects / Watson Studio Labs / titanic.csv / Refine data

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. Select the **Bar** Chart Type.

Data Profile Visualizations

DETAILS CHART TYPE

Scatter plot Line Multi-series Histogram Population ... Q-Q plot Pie **Bar** Parallel Relationship

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

Data Profile Visualizations

DETAILS Bar chart

Category * **pclass_value**

Order based on Category name

CHART TYPE

Scatter plot Line Multi-series Histogram Population ... Q-Q plot Pie **Bar**

COUNT

800

700

600

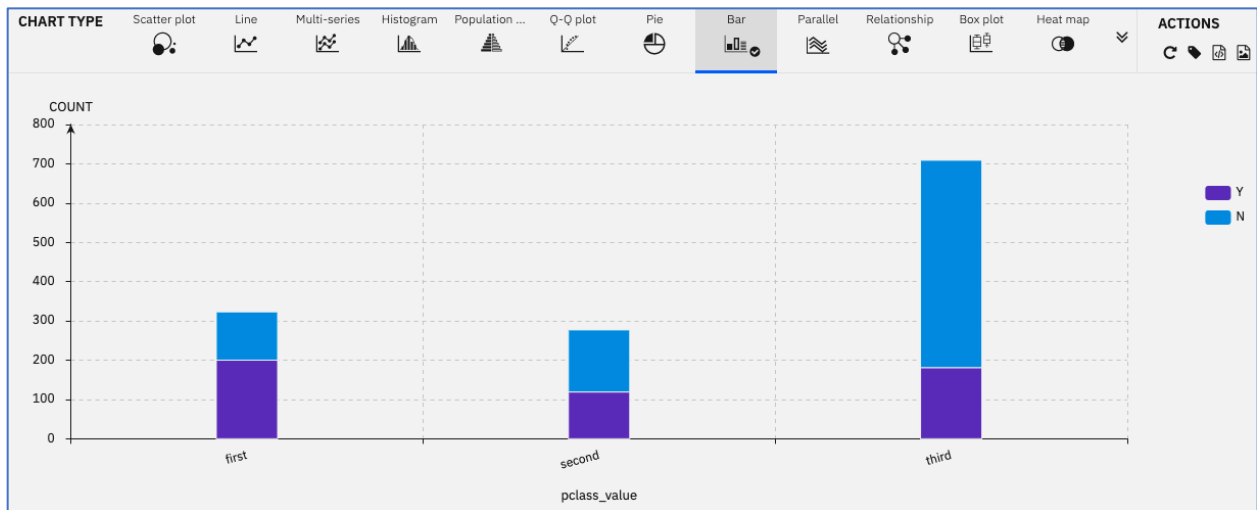
500

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

The screenshot shows the 'Visualizations' tab with the following settings:

- Split by:** survived_value
- Split type:** Stacked

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



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

The screenshot shows the 'Visualizations' tab with the following settings:

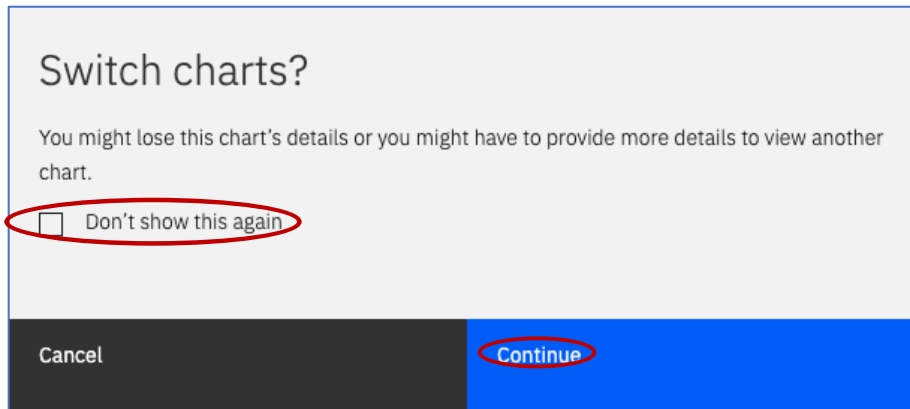
- Category:** sex

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

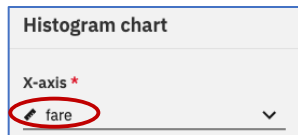
The screenshot shows the 'CHART TYPE' menu with the following options:

- Scatter plot
- Line
- Multi-series
- Histogram**
- Population ...
- Q-Q plot
- Pie
- Bar

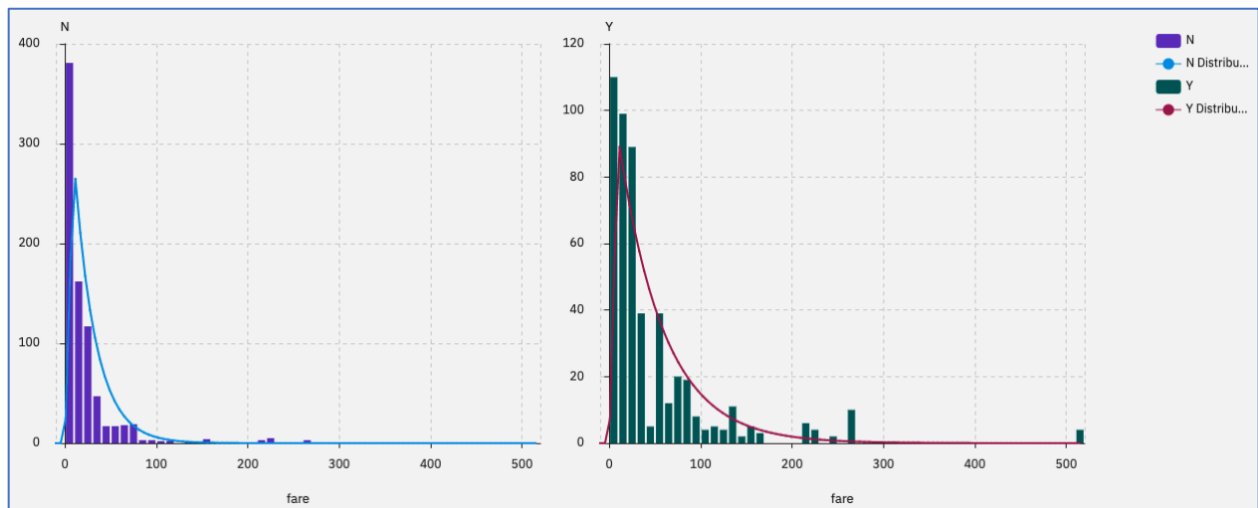
8. Click on the **Don't show this again** check box and click **Continue**.



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



10. 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.



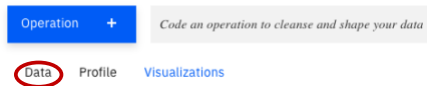
Step 4: Prepare the data for modeling

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 table of data columns. The 'cabin' column is selected, indicated by a red circle around its vertical ellipsis icon. A context menu is open for the 'cabin' column, showing options: 'Remove' (circled in red) and 'Remove duplicates'. The table also shows other columns like 'embarked' and data rows with values like 'B5' and 'C22 C26'.

cabin String	embarked String
B5	
C22 C26	

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.

The screenshot shows the 'Steps' panel on the right side of the interface. It is titled 'Steps' (circled in red) and shows a list of 7 steps. The first step is 'Data Source' with the value 'titanic.csv'. The second step is 'Convert column type' with the value 'AUTOMATIC'. Below this, there is a description: 'Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol.' The third step is 'Custom code' with the following code: `mutate(survived_value=ifelse(survived==1, "Y", "N"))`. The fourth step is another 'Custom code' block with the following code: `mutate(pclass_value=ifelse(pclass==1,"first",ifelse(pclass==2,"second","third")))`.

Steps
7 Steps
Data Source
titanic.csv
Convert column type
AUTOMATIC
Automatically converted one or more columns to inferred data types. Strings that are converted to decimal use a dot (.) for the decimal symbol.
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>

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

embarked	survived_value
String	String
S	Remove
S	Remove duplicates
S	Remove empty rows
S	Sort ascending

- 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	survived_value
Decimal	String	String
211.3375	Remove	Y
151.55	Remove duplicates	Y
151.55	Remove empty rows	N
151.55	Sort ascending	N
151.55	Sort descending	N
26.55	Substitute	Y
77.9583	CONVERT COLU...>	Decimal
0	View All	Integer
51.4792		

- 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 Cancel
-------------	------------------------------	----------------

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 Decimal		sibsp Decimal	parch Decimal
29		0	
0.9167		2	
2		2	
30		2	
25		2	
48		0	
63		✓ Decimal	
39		Integer	
53			

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

Operation +

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)))))))

Apply Cancel

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)))
```

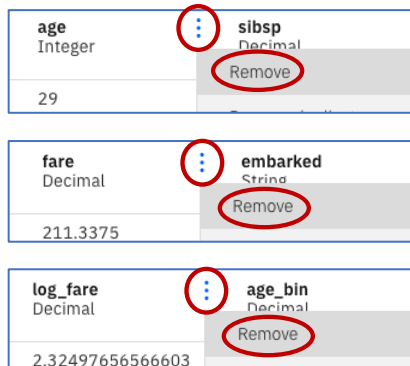
and then clicking **Apply**


Operation +

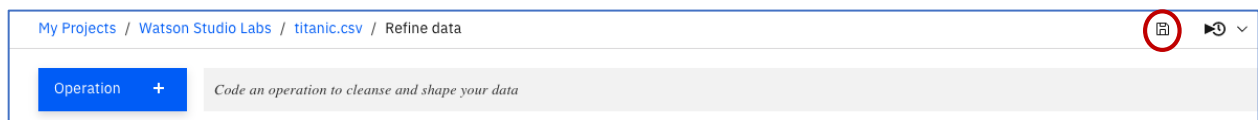
mutate(log_fare_bin=ifelse(log_fare<0,0,ifelse(log_fare>8,9,as.integer(log_fare)+1)))

Apply Cancel


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.

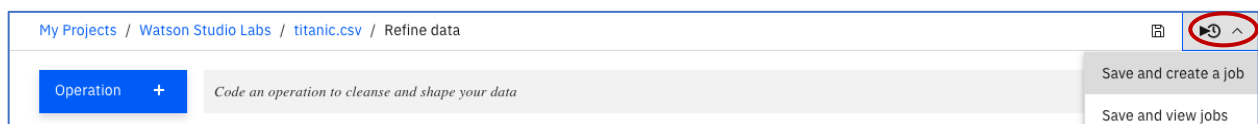


11. Save the Data Flow by clicking on the Save Data Flow icon .

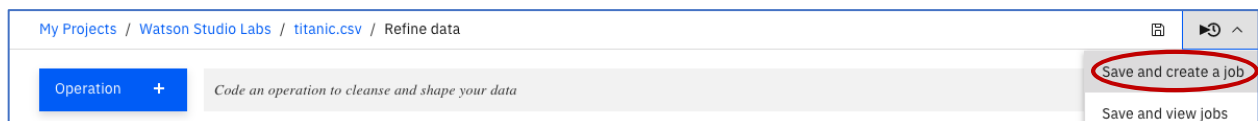


Step 5: Run the sequence of Data Flow operations on the entire data set.

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**



- This action results in the following page display. Fill in the **Job Name**, for example **titanic_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

titanic_flow_job

Description (Optional)

Description of job

Associated Asset

DATA REFINERY FLOW

titanic_flow 16 Steps Edit

Select runtime

Default Data Refinery XS

INPUT

titanic.csv CSV

OUTPUT

titanic_shaped.csv CSV

Schedule off

Cancel

Create

Create and Run

- Note the number of steps used to transform the data. It should be 16 or 17 (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

Scheduled to run

Edit

No Schedule Created

Environment definition

Edit

Default Data Refinery XS

Associated Asset

DATA REFINERY FLOW

titanic_flow 16 Steps

INPUT

titanic.csv CSV

OUTPUT

titanic_shaped.csv CSV

Runs (1)

Start Time	Status	Duration	Started By	Action
Jul 19, 2020 8:39:32 PM	Running	---	FCTO Labs	!

- 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

Scheduled to run

Edit

No Schedule Created

Environment definition

Edit

Default Data Refinery XS

Associated Asset

DATA REFINERY FLOW

titanic_flow 16 Steps

INPUT

titanic.csv CSV

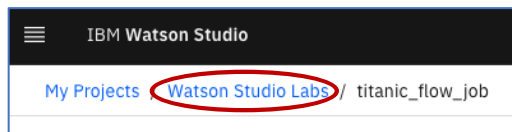
OUTPUT

titanic_shaped.csv CSV

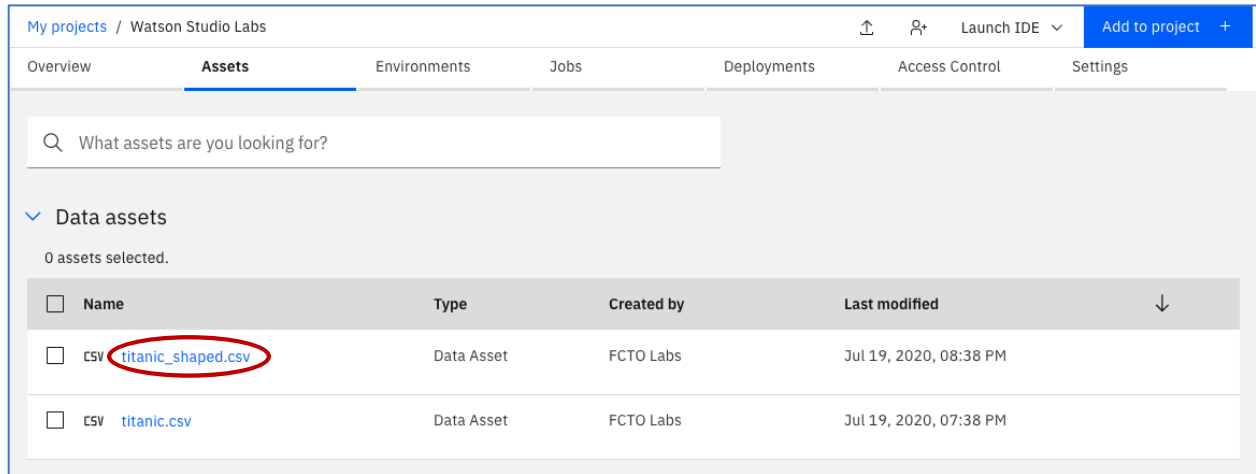
Runs (1)

Start Time	Status	Duration	Started By	Action
Jul 19, 2020 8:37:14 PM	Completed	24 seconds	FCTO Labs	!

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



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



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

Schema: 12 Columns
Preview: First 1000 rows

pclass String	survived String	name String	sex String	sibsp String	parch String	ticket String	embarked String	survived_v... String	pclass_val... String	age_bin String	log_fare_... String
1.0	0.0	Allison, Miss. Helé	female	1.0	2.0	113781	S	N	first	0.0	3.0
1.0	1.0	Anderson, Mr. Har	male	0.0	0.0	19952	S	Y	first	4.0	2.0
1.0	1.0	Appleton, Mrs. Ed	female	2.0	0.0	11769	S	Y	first	4.0	2.0
1.0	1.0	Astor, Mrs. John J	female	1.0	0.0	PC 17757	C	Y	first	3.0	3.0
1.0	1.0	Barkworth, Mr. Alj	male	0.0	0.0	27042	S	Y	first	6.0	2.0
1.0	1.0	Baxter, Mrs. Jame	female	0.0	1.0	PC 17558	C	Y	first	4.0	3.0
1.0	1.0	Beckwith, Mr. Ricl	male	1.0	1.0	11751	S	Y	first	3.0	2.0
1.0	1.0	Bidois, Miss. Rosa	female	0.0	0.0	PC 17757	C	Y	first	4.0	3.0
1.0	1.0	Bishop, Mr. Dickin	male	1.0	0.0	11967	C	Y	first	3.0	2.0

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.