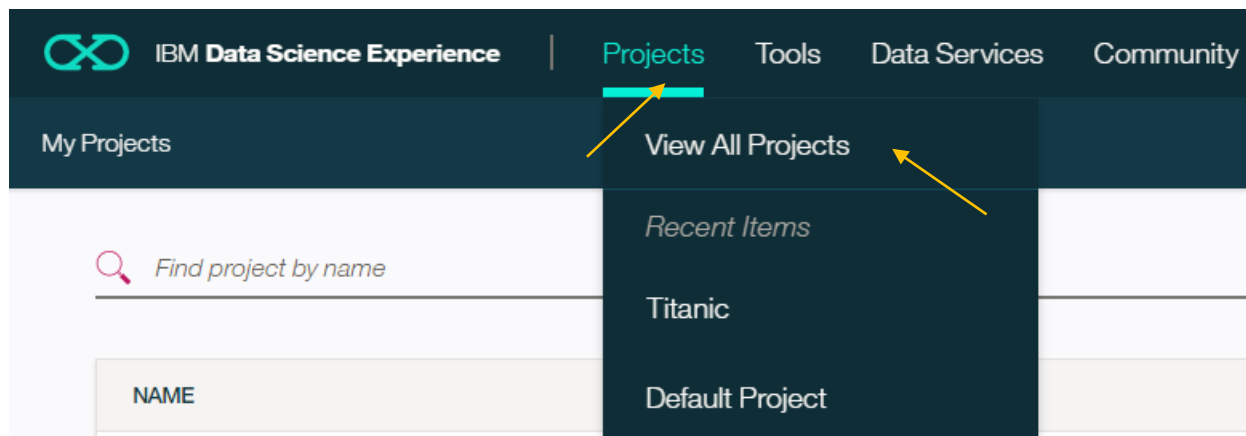


# Data Science Experience SPSS Modeler Overview

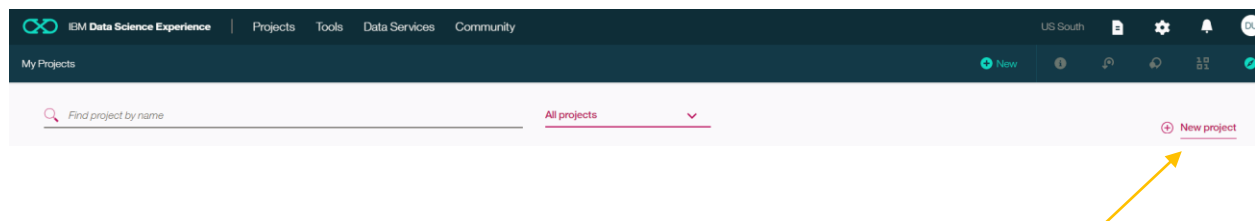
This lab will introduce the SPSS Modeler capability using the Titanic dataset. The lab will guide the development of an SPSS Modeler stream that will prepare the input data for modeling in order to run a machine learning algorithm predicting survivability of a passenger on the Titanic.

## Step 1: Create a new Project

1. Click on the **Projects** tab and then **View All Projects**.



2. Click on **New project**.



3. Enter a project **Name** (eg Titanic), optionally a **Description**, take the defaults for **Spark service**, and **Storage type**. Click on **Create**.

IBM Data Science Experience | Projects | Tools | Data Services | Community | US South

## New project

**Name**  
Titanic-SPSS-Lab 84

**Description**  
This lab demonstrates the SPSS Modeler capability 2951

☐ Restrict who can be a collaborator ⓘ

**Compute engine**  
Spark service  
DSX-Spark

⚠ If you associate the same Spark service with multiple projects, the Spark history server will display job history information for all the projects.

**Storage type**  
☒ Object Storage (Swift API) ☐ Cloud Object Storage (Beta)

**Target object storage instance**  
DSX-ObjectStorage

Cancel Create

4. You should be on the Project Overview screen. Click on the Asset tab.

IBM Data Science Experience | Projects | Tools | Data Services | Community | US South

My Projects > Titanic-SPSS-Lab

Overview Assets Bookmarks Deployments Collaborators Settings

### Titanic-SPSS-Lab

Last Updated: Nov 30 2017

0 Assets 0 Bookmarks 1 Collaborators

**Date created**  
Nov 30 2017

**Description**  
This lab demonstrates the SPSS Modeler capability

**Storage**  
0% of 5 GB used

**Collaborators**  
View all (1)  
DSX5000 User Admin

**Bookmarks**  
View all (0)  
You currently have 0 bookmarks

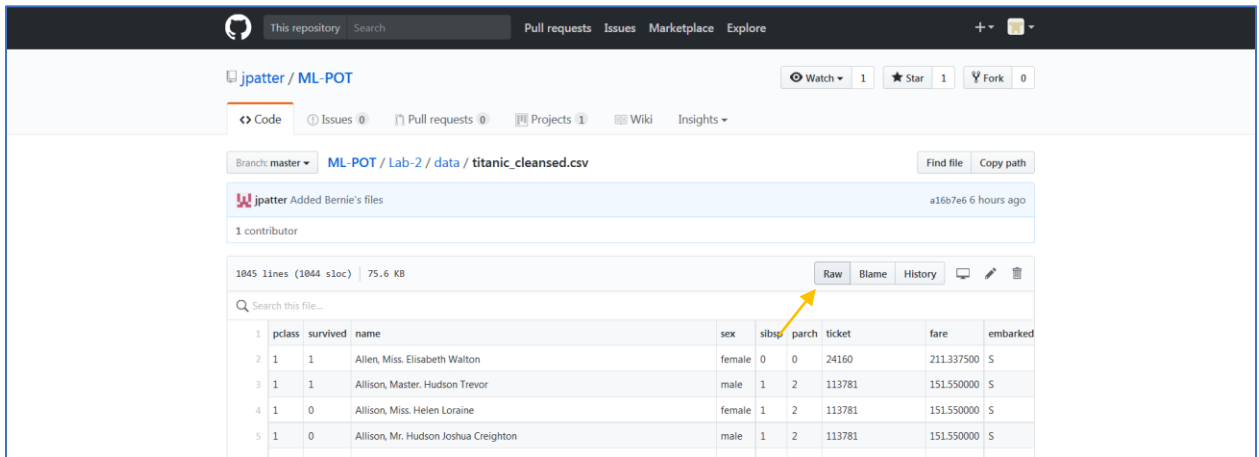
**Recent activity**

Alerts related to this project will show here when the project is active.

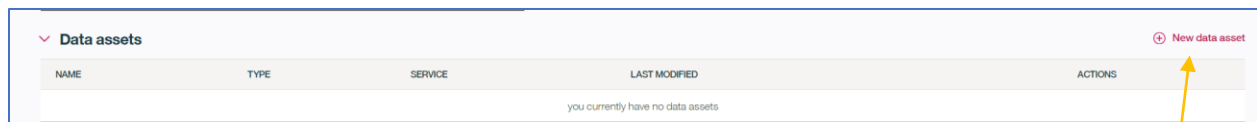
## Step 2: Adding a Data Asset to the Titanic project

1. Download the Titanic data file from [Titanic Data Set](#)

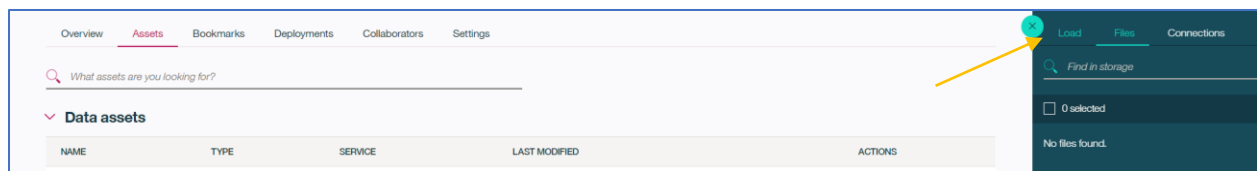
Right click on Raw, and click on Save link as ....



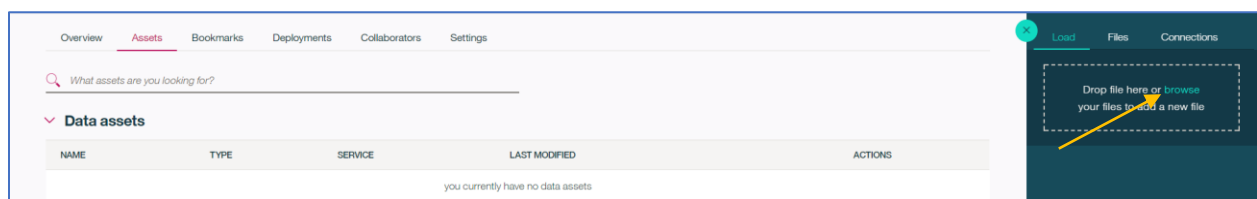
2. Go back to the Titanic project. Click on **New data asset**.



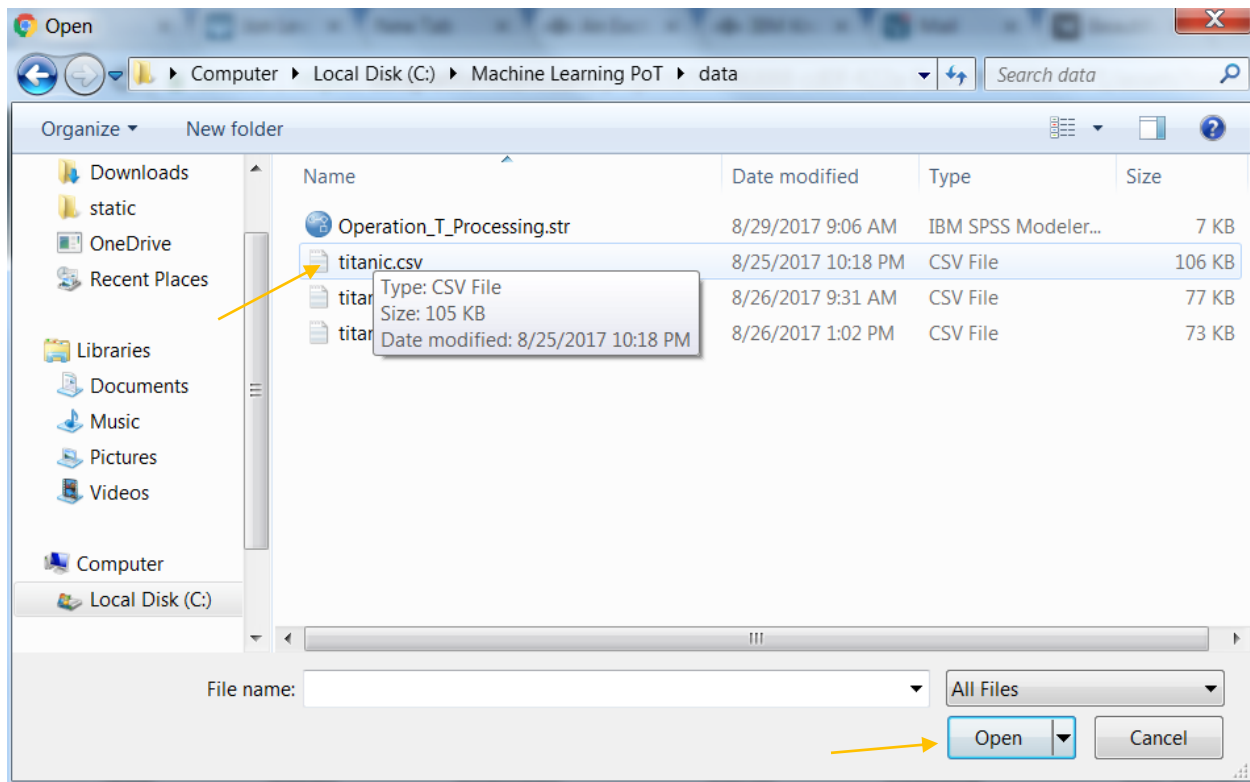
3. Click on the **Load** tab.



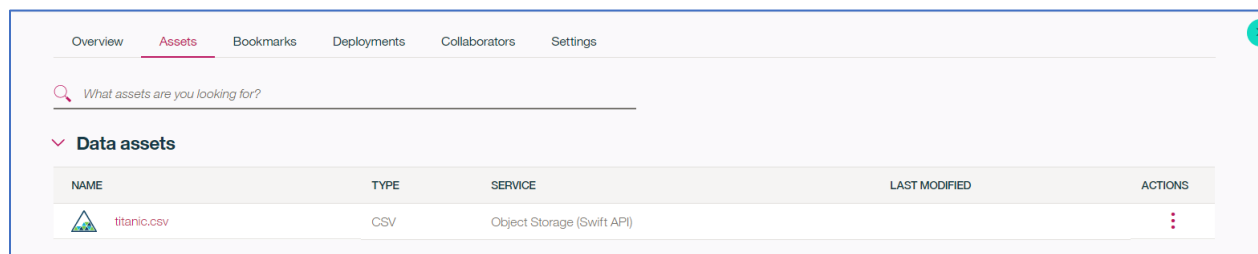
4. Click on **browse**.



5. Go to the folder where the titanic\_csv file is stored. Select the titanic.csv file and then click **Open**.



6. The file is added as a Data Asset.

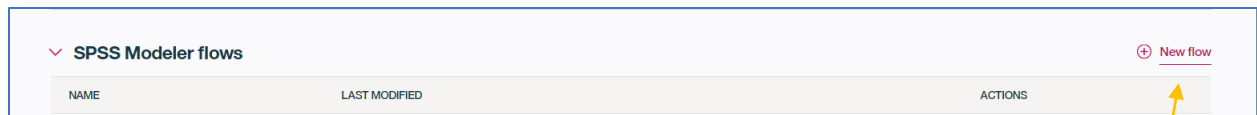


### Step 3: Create a Model to predict survival

In this section, we will create a Machine Learning flow using SPSS nodes. Documentation describing the nodes is available at <https://dataplatform.ibm.com/docs/content/analyze-data/ml-canvas-spss.html?context=analytics>.

#### Step 3.1 Create a New Flow and Load the Data

1. In project Titanic, click on **New flow** in the SPSS Modeler flows section.




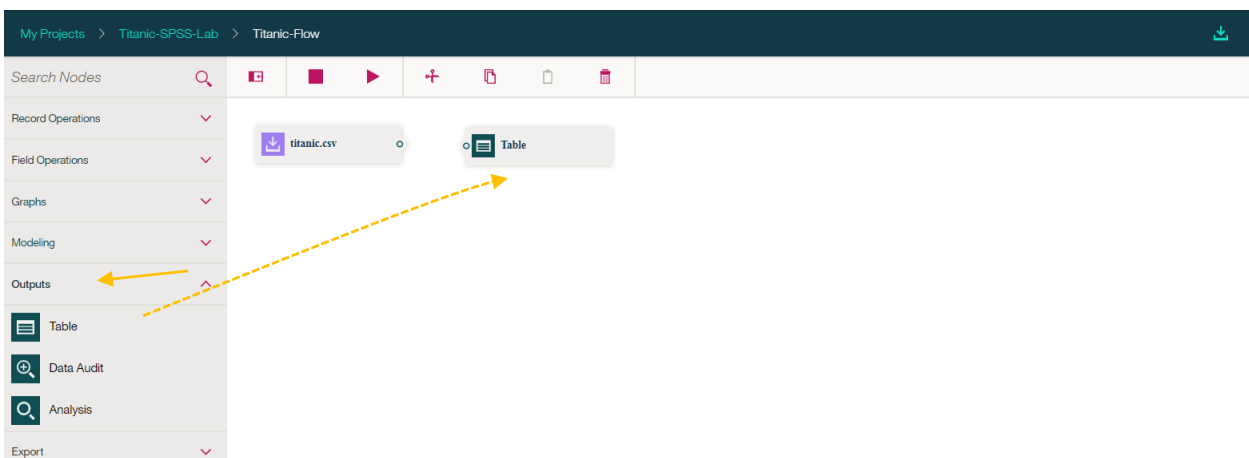
2. Enter a **Name** for the flow, optionally enter a **Description**, select IBM SPSS Modeler for the **Runtime**, and click on **Create**.

A screenshot of the 'SPSS Modeler BETA' 'New' form. The form has three tabs: 'New', 'From file', and 'From example'. The 'New' tab is selected. There are three main sections: 'Name\*', 'Description', and 'Runtime'. The 'Name\*' section has a text input field containing 'Titanic-Flow' with a yellow arrow pointing to it. The 'Description' section has a large text area with the placeholder text 'Type description here.' and a yellow arrow pointing to it. The 'Runtime' section has a dropdown menu with 'IBM SPSS Modeler' selected and a yellow arrow pointing to it. At the bottom, there are two buttons: 'Cancel' and 'Create', with a yellow arrow pointing to the 'Create' button.

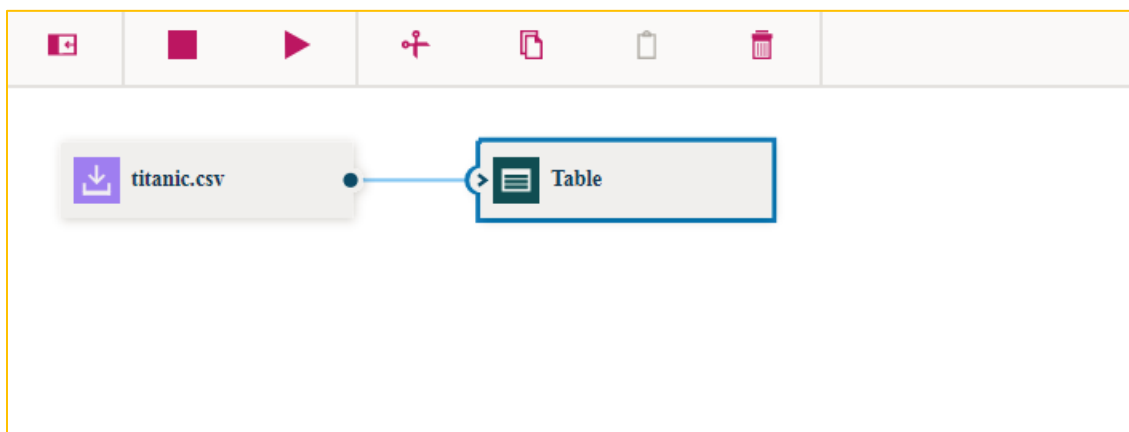
3. This opens the Flow Editor. Click on the titanic.csv file and hold the left mouse key and drag the file onto the left side of the canvas. Release the left mouse key.



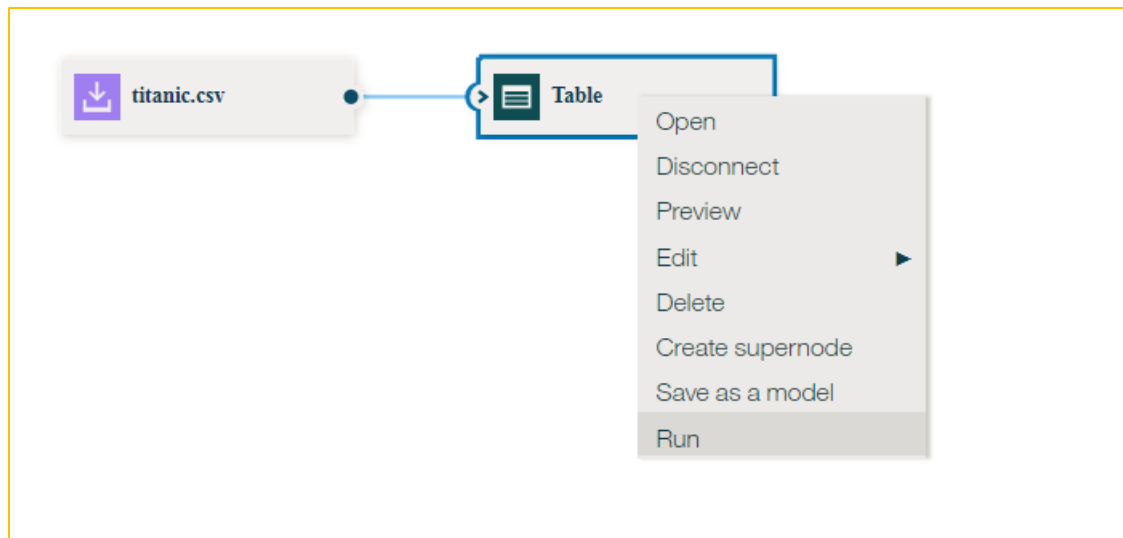
4. Click on the Outputs menu item in the Node Palette on the left and then click on the Table icon and drag the icon to the right of the titanic.csv icon. The SPSS Table node will display the contents of the csv file. If the Node Palette is not visible, click on the Node Palette icon 



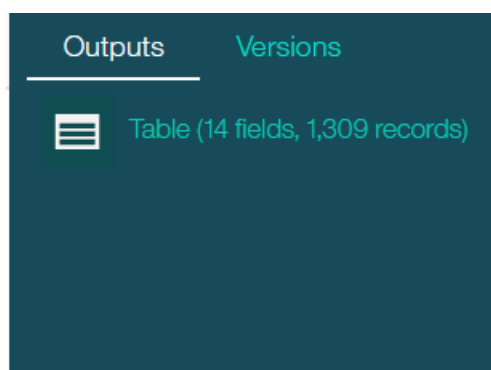
5. Connect the right side of the titanic.csv icon to the left side of the Table icon. This is accomplished by clicking on the little circle at the right side of the titanic.csv icon holding the left mouse key and dragging the mouse to the little circle on the left side of the Table icon, and then releasing the left mouse key.



6. Right click on the **Table** icon, and select **Run**.



- The “Running Flow” prompt will appear and then when completed a Table output selection will appear on the right side of the screen under the **Outputs** tab.




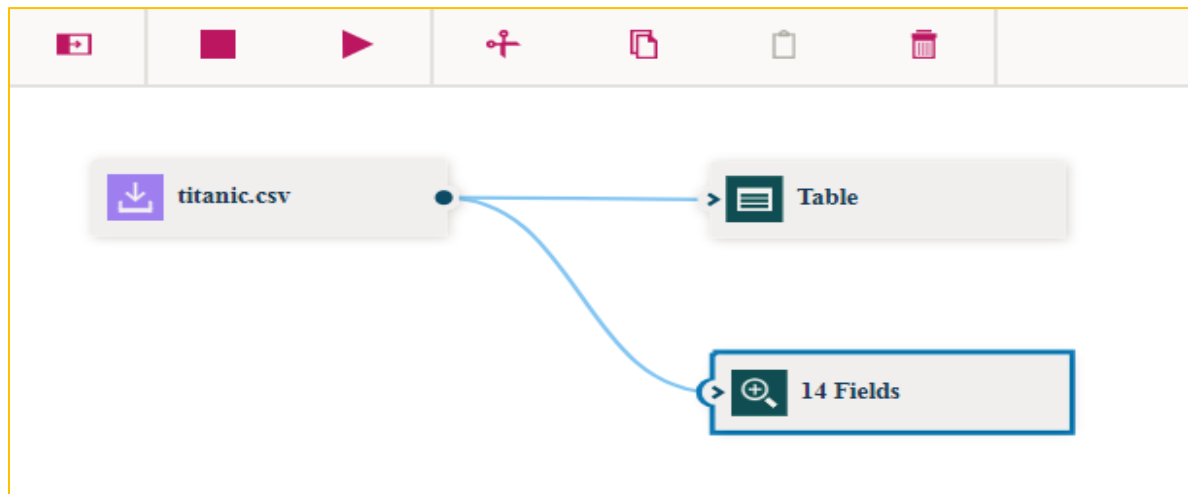
- Double click on the Table selection above the contents of the titanic.csv will be displayed. Each row contains information on a passenger on the Titanic. We will use this data to make predictions on survivability.

PCLASS	SURVIVED	NAME	SEX	AGE	SIBSP	PARCH	TICKET	FARE	CABIN	EMBARKED	BOAT
1	1	Allen, Miss. Elisabeth	female	29	0	0	24160	211.3375	B5	S	2
1	1	Allison, Master. Huds	male	0.9167	1	2	113781	151.55	C22 C26	S	11
1	0	Allison, Miss. Helen L.	female	2	1	2	113781	151.55	C22 C26	S	
1	0	Allison, Mr. Hudson J.	male	30	1	2	113781	151.55	C22 C26	S	
1	0	Allison, Mrs. Hudson	female	25	1	2	113781	151.55	C22 C26	S	
1	1	Anderson, Mr. Harry	male	48	0	0	10952	26.55	E12	S	3
1	1	Andrews, Miss. Korn	female	63	1	0	13502	77.9583	D7	S	10
1	0	Andrews, Mr. Thoma	male	39	0	0	112050	0	A36	S	
1	1	Appleton, Mrs. Edwa	female	53	2	0	11769	51.4792	C101	S	D
1	0	Artagaveytia, Mr. Rar	male	71	0	0	PC 17809	49.5042		C	
1	0	Astor, Col. John Jacc	male	47	1	0	PC 17757	227.525	C82 C84	C	
1	1	Astor, Mrs. John Jac	female	18	1	0	PC 17757	227.525	C82 C84	C	4
1	1	Aubart, Mme. Leonth	female	24	0	0	PC 17477	69.3	B35	C	9
1	1	Barber, Miss. Ellen T.	female	26	0	0	19877	78.85		S	6
1	1	Barkworth, Mr. Alger	male	80	0	0	27042	30	A23	S	B

### Step 3.2 Explore the Data using the Data Audit Node

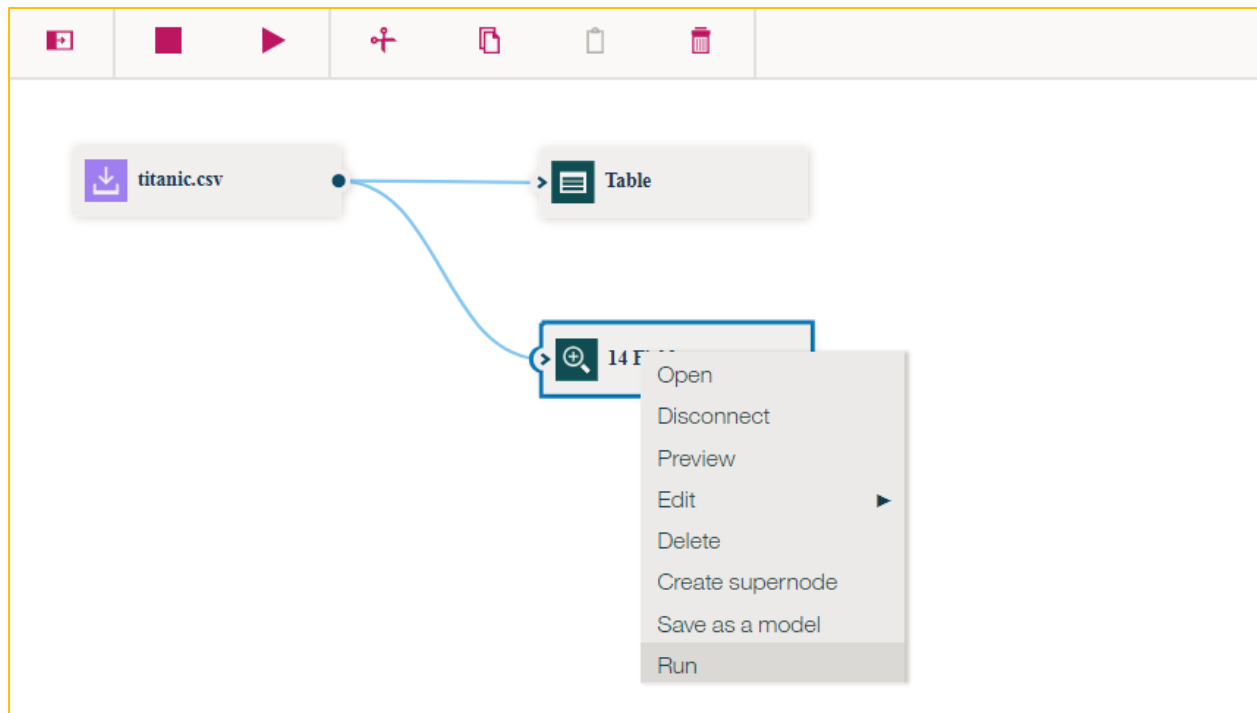
Perusing through the data in the table, we can see that there are missing values. The SPSS Modeler has a Data Audit node that provides profiling information on the input data that is useful for cleansing the data. It provides a comprehensive first look at the data, including summary statistics, as well as information about outliers, missing values, and extremes.

1. Add a **Data Audit** node to the flow clicking on the **Outputs** menu item in the Node Palette, and then dragging the **Data Audit** node to underneath the titanic.csv node. If the Node Palette is not visible, click on the Node Palette icon . Connect the titanic.csv node to the Data Audit node. The canvas should appear as below.

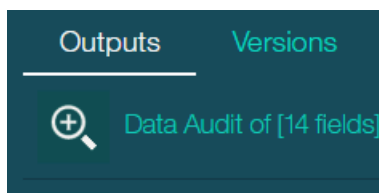


2. Right click on the **Data Audit** node and click **Run**.





3. The “Running Flow” prompt will appear and then when completed a Data Audit output selection will appear on the right side of the screen under the **Outputs** tab.



4. Double click on the **Data Audit of [14 fields]** to view the Data Audit output. We can see that several fields have many missing values (cabin, boat,body,home.dest). These fields will be removed using a **Filter** node below. Other fields have only a few missing values (fare, embarked, age). The rows containing the missing values will be removed using a **Select** node below.

**Data Audit of [14 fields] #8**


	Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
1	pclass		Continuous	1	3	2.295	0.838	-0.599	--	1309
2	survived		Continuous	0	1	0.382	0.486	0.486	--	1309
3	name		Categorical	--	--	--	--	--	--	1309
4	sex		Categorical	--	--	--	--	--	2	1309
5	age		Continuous	0.167	80.000	29.881	14.413	0.408	--	1046
6	sibsp		Continuous	0	8	0.499	1.042	3.844	--	1309
7	parch		Continuous	0	9	0.385	0.866	3.669	--	1309
8	ticket		Categorical	--	--	--	--	--	--	1309
9	fare		Continuous	0.000	512.329	33.295	51.759	4.368	--	1308
10	cabin		Categorical	--	--	--	--	--	186	295
11	embarked		Categorical	--	--	--	--	--	3	1307
12	boat		Categorical	--	--	--	--	--	27	486
13	body		Continuous	1	328	160.810	97.697	0.092	--	121
14	home.dest		Categorical	--	--	--	--	--	--	745

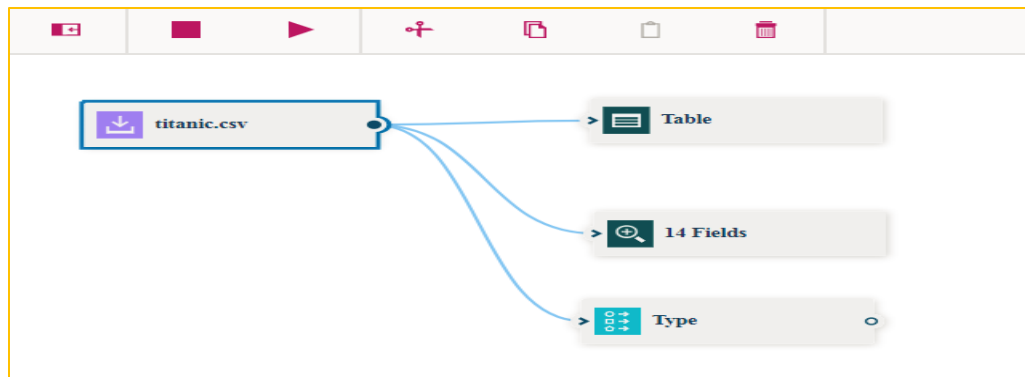
  

	Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
1	pclass	Continuous	0	0	None	Never	Fixed	100.000	1309	0	0	0	0
2	survived	Continuous	0	0	None	Never	Fixed	100.000	1309	0	0	0	0
3	name	Categorical	--	--	--	Never	Fixed	100.000	1309	0	0	0	0
4	sex	Categorical	--	--	--	Never	Fixed	100.000	1309	0	0	0	0
5	age	Continuous	3	0	None	Never	Fixed	79.908	1046	263	0	0	0
6	sibsp	Continuous	28	9	None	Never	Fixed	100.000	1309	0	0	0	0
7	parch	Continuous	14	10	None	Never	Fixed	100.000	1309	0	0	0	0
8	ticket	Categorical	--	--	--	Never	Fixed	100.000	1309	0	0	0	0
9	fare	Continuous	34	4	None	Never	Fixed	99.924	1308	1	0	0	0
10	cabin	Categorical	--	--	--	Never	Fixed	22.536	295	1014	0	0	0

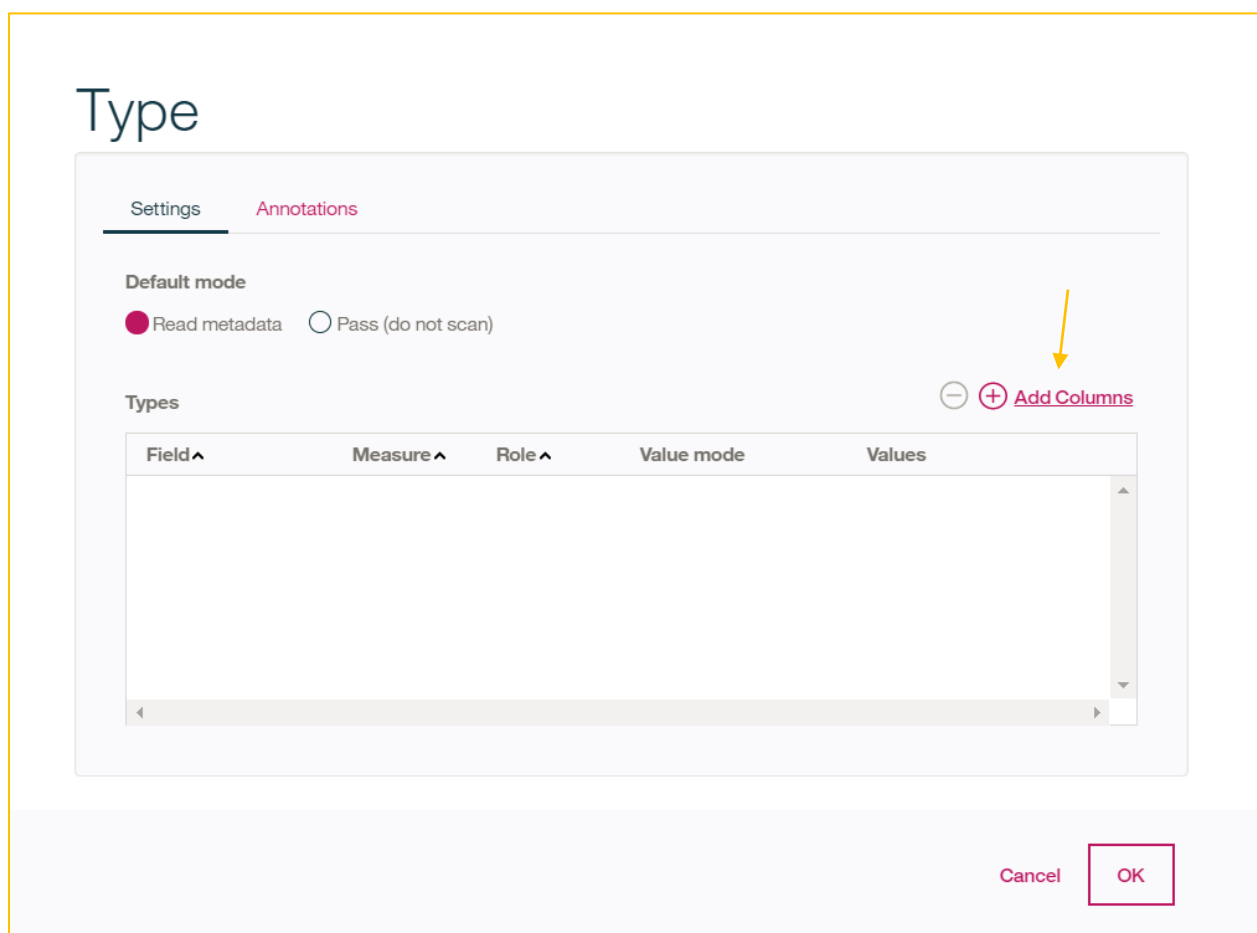
### Step 3.3 Explore the Data using Graph Nodes.

The SPSS desktop version has a rich graphical icon set. Currently, the DSX version has only 4 graph nodes in the beta version. The Distribution node, and the Histogram node will be used to explore some of the characteristics of the Titanic Data Set. First, we will add a Type node to the canvas. The Type node specifies field metadata and properties. We will change the measurement property for the “pclass” and “survived fields” that was derived as “Continuous” by scanning the data values to “Ordered Set” and “Flag” respectively.

1. Add a **Type** node to the flow by clicking on the **Field Operations** menu item in the Node Palette and then drag the **Type** node underneath the **Data Audit** node. If the Node Palette is not visible, click on the Node Palette icon . Connect the titanic.csv node to the **Type** node. The canvas should appear as below.









2. Double click on the **Type** node. Click on **Add Columns**.








3. Click on the checkboxes adjacent to the pclass and survived fields, and then click on the left arrow next to **Select Fields for Type**.

# Type

 **Select Fields for Type** [Reset](#) 

Search in column Field name  Filter:   



<input type="checkbox"/>	Field name ^	Data type ^
<input type="checkbox"/>	pclass	 integer
<input type="checkbox"/>	survived	 integer
<input type="checkbox"/>	name	 string
<input type="checkbox"/>	sex	 string
<input type="checkbox"/>	age	 double
<input type="checkbox"/>	...	...



- Click on the measurement level field for pclass and select **Ordered Set**. Click on the measurement level field for survived and select **Flag**. Click on **OK**.

# Type


**Settings** **Annotations**

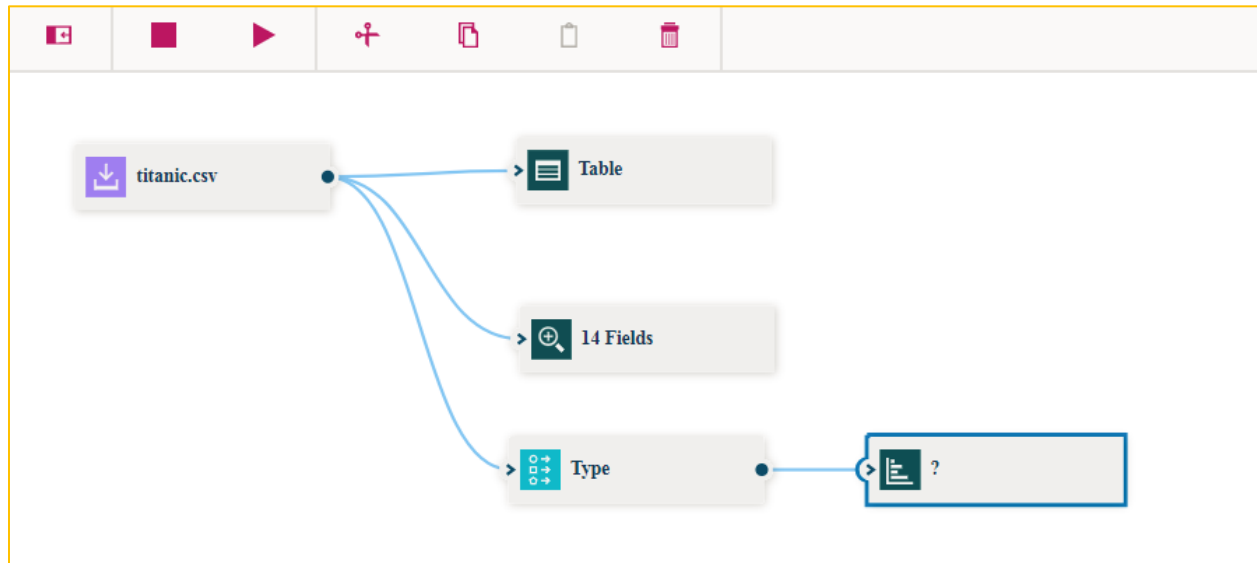
**Default mode**  
☒ Read metadata ☐ Pass (do not scan)

**Types**   [Add Columns](#)

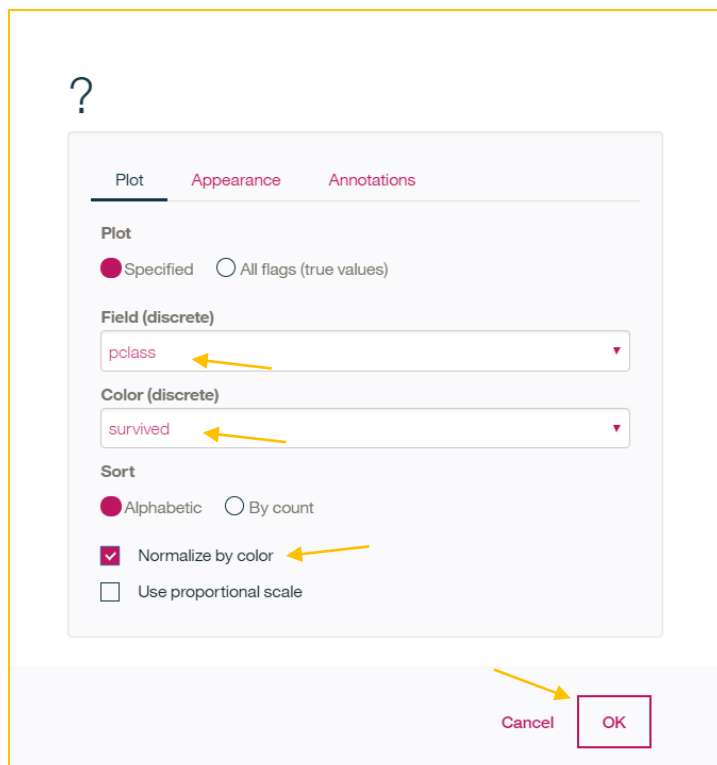
Field ^	Measure ^	Role ^	Value mode	Values
pclass	 <b>Ordered Set</b>	Input	Read	...
survived	 <b>Flag</b>	Input	Read	...
...	...	...	...	...

[Cancel](#) [OK](#)

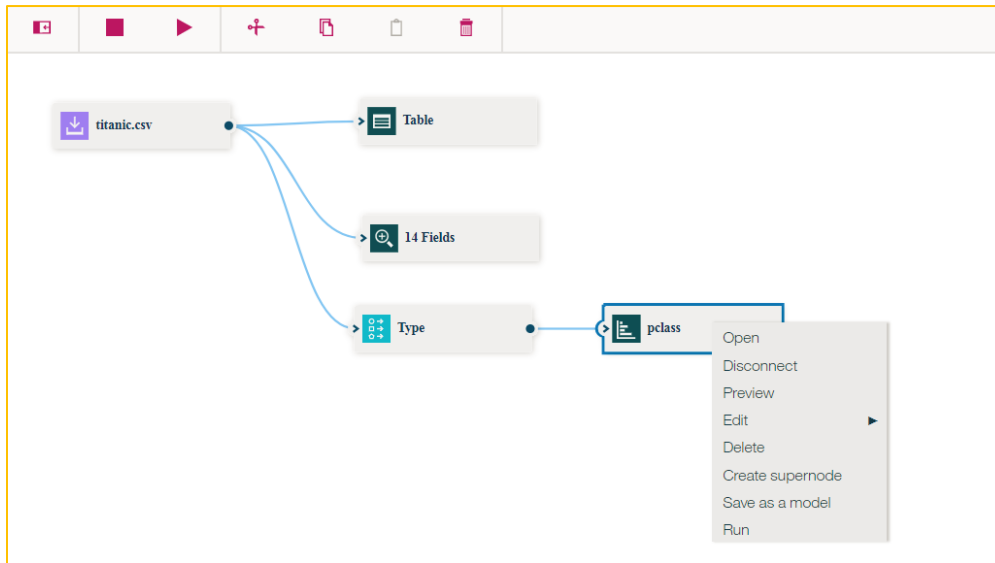
5. Add a **Distribution** node to the flow by clicking on the **Graph** menu item and then dragging the **Distribution** node to the canvas to the right of the **Type** node. If the Node Palette is not visible, click on the Node Palette icon . Connect the **Type** node to the **Distribution** node. The canvas should appear as below. The ? indicates that the fields to be plotted have not been identified.



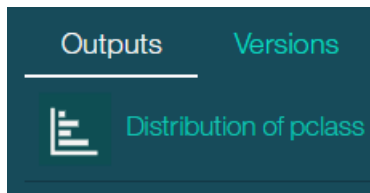
6. Double click on the Distribution Node. In the Field (discrete) dropdown, select pclass. In the Color (discrete) dropdown, select survived. Click on the normalize by color checkbox, and then click OK.



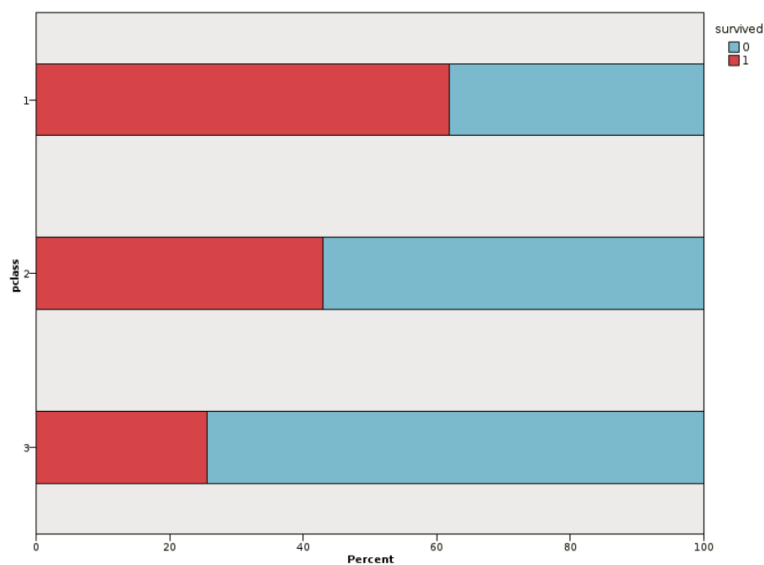
7. Right click on the Distribution node, and select Run.



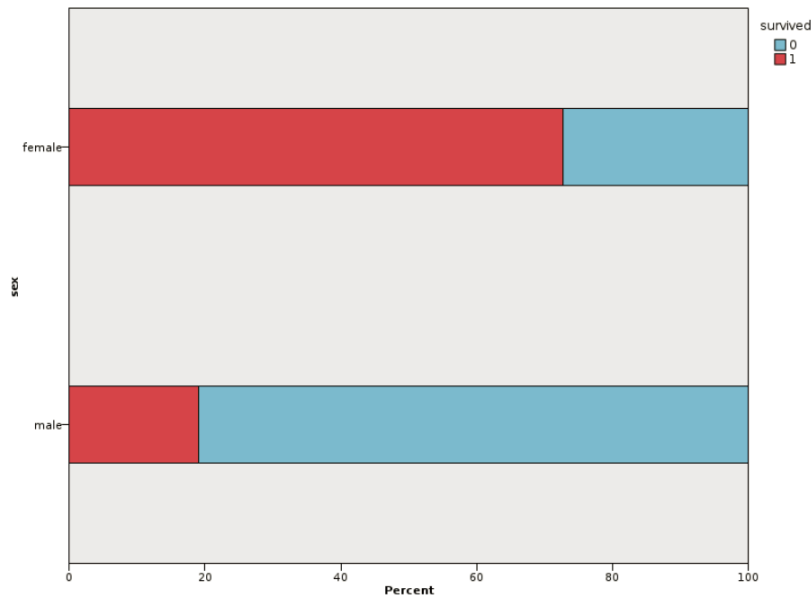
8. The Distribution of pclass output will appear under the **Outputs** tab on the right hand side of the screen.




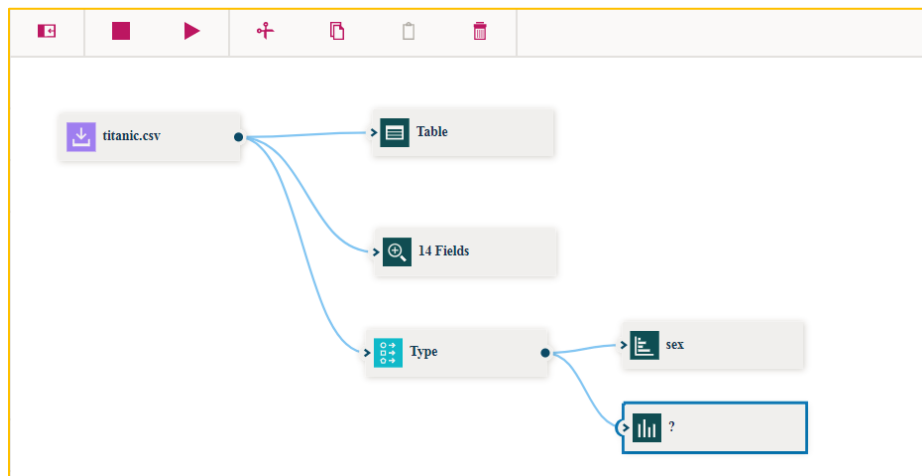
9. Double click on the **Distribution of pclass #1** to view the graph. We can see from the graph that the likelihood of surviving is correlated to the passenger class. The first class passengers have the highest rate of survivability.



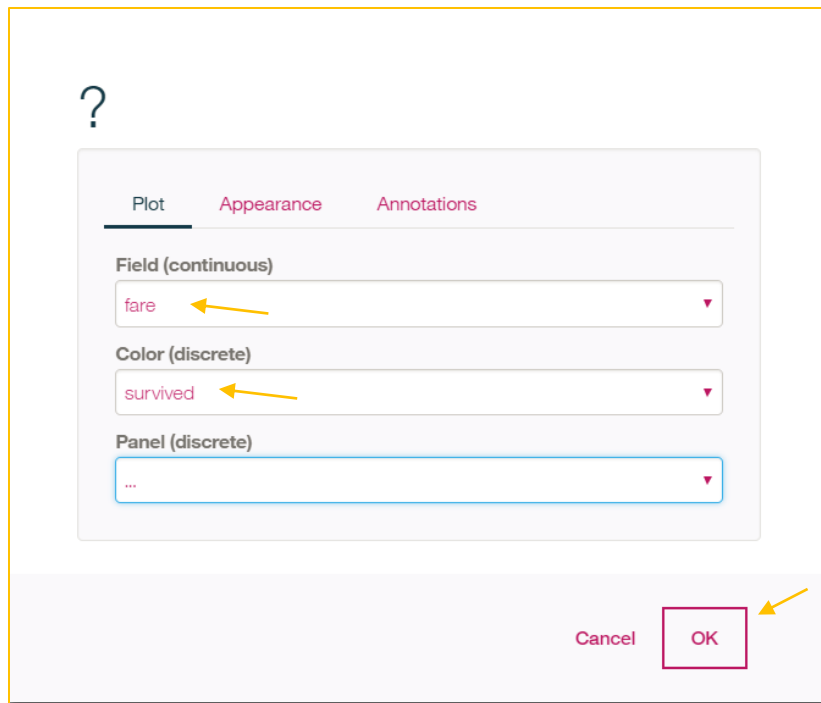
10. You can change the distribution graph to show the survivability by gender by double clicking on the Distribution node and replacing pclass with sex and clicking OK. Re-run the graph by right clicking on the Distribution node and selecting Run. Double click on the Distribution of sex #1 to display the graph.



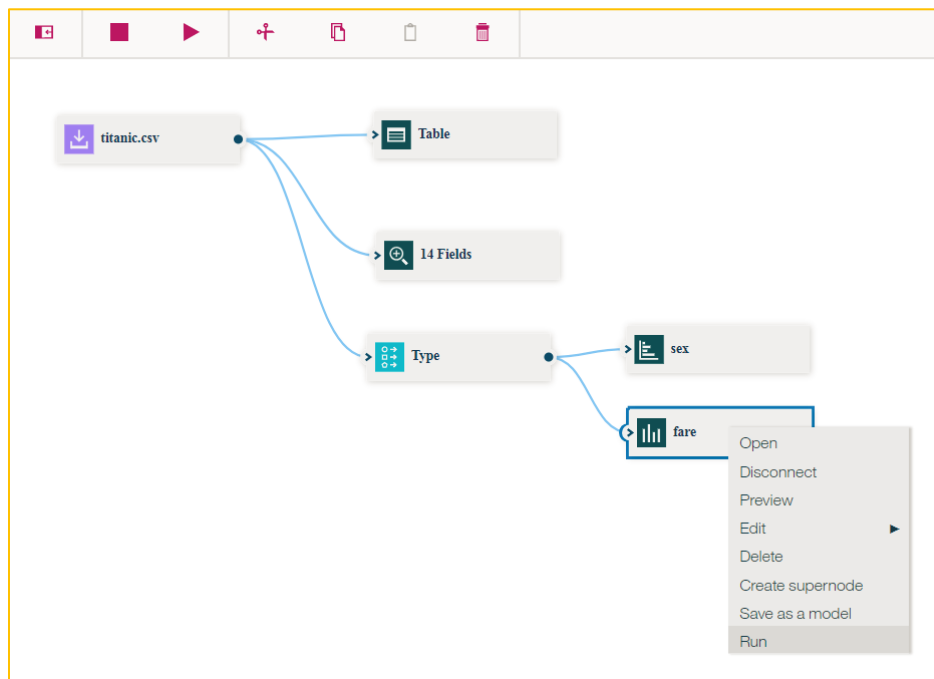
11. Add a **Histogram** node to the flow by clicking on the **Graphs** menu item and then dragging the **Histogram** node to the canvas underneath the **Distribution** node. If the Node Palette is not visible, click on the Node Palette icon . Connect the **Type** node to the **Histogram** node. The canvas should appear as below. The ? indicates that the fields to be plotted have not been identified.



12. Select fare from the Field (continuous) dropdown. Select survived from the Color (discrete) dropdown. Click on OK.

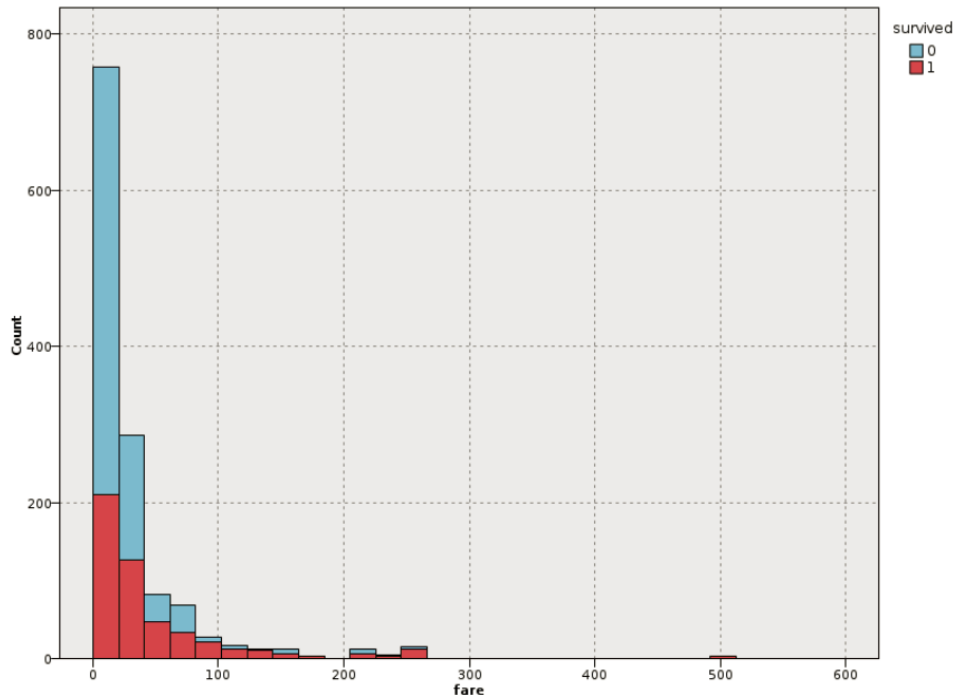


13. Right click on the **Histogram** node and select **Run**.



14. Double click on the Histogram of fare  **Histogram of fare** under the Outputs tab at the right of the screen.





15. We can see that the histogram is skewed. Skewness will impact the effectiveness of some machine learning techniques. One way to deal with skewness is to do a logarithmic transformation of the data. We will do this transformation in the preparing the data for modeling section below.

### Step 3.4 Prepare the Data for Modeling

Based on our exploration of the data, there are several transformations that are needed to prepare the data for modeling. This section will introduce, the Filter node, the Select node, and the Derive node that will do the necessary transformations. The Filter and Derive nodes act on a field level, whereas the Select node acts on a record level.

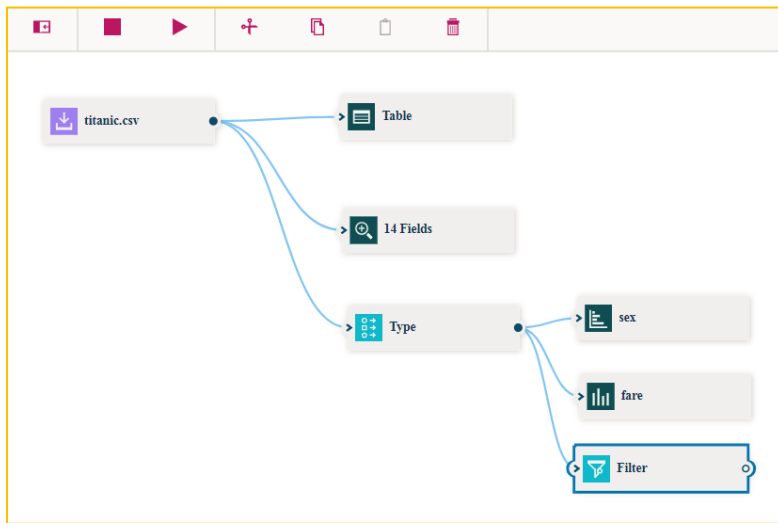
**Filter node** – The Filter node performs two functions. It specifies fields that can be dropped. It also allows fields to be renamed. We will drop the fields cabin,boat,body, and home.dest.

**Derive node** – The Derive node modifies data values or creates new fields from one or more existing fields. We will use the derive node to do a logarithmic transformation of the fare field. We will also use this node to bin the age and fare fields.

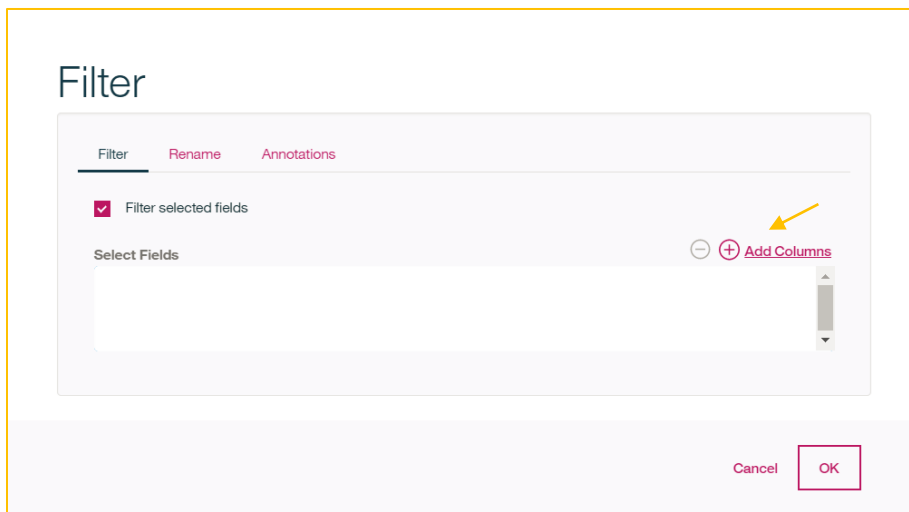
**Select node** – The Select node is used to select or discard a subset of records from the data stream based on a specific condition. We will remove the rows where there are missing information in the fare, age, or embarked fields.

1. Add a **Filter** node to drop fields with many missing values. Add the **Filter** node by clicking on the **Field Operations** menu item in the Node palette and dragging the **Filter**

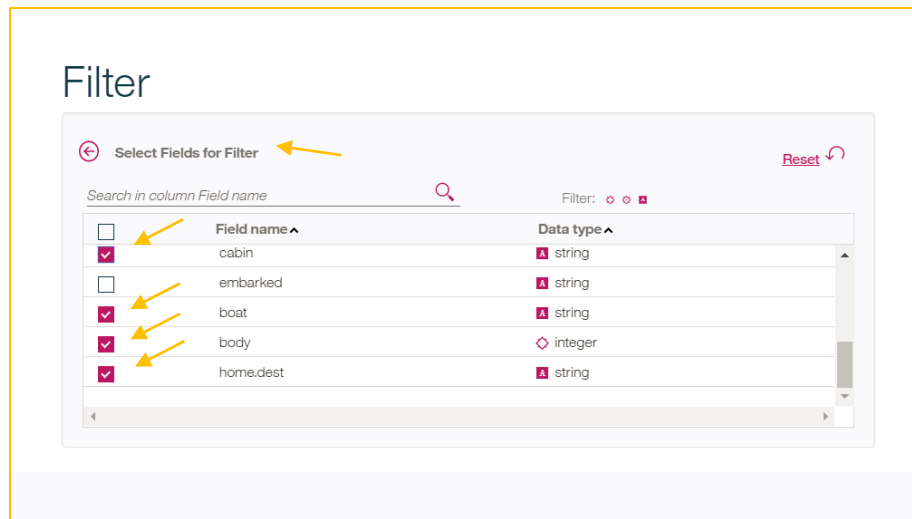
node onto the canvas underneath the fare **Histogram** node. If the Node Palette is not visible, click on the Node Palette icon  first. The canvas should appear as below.



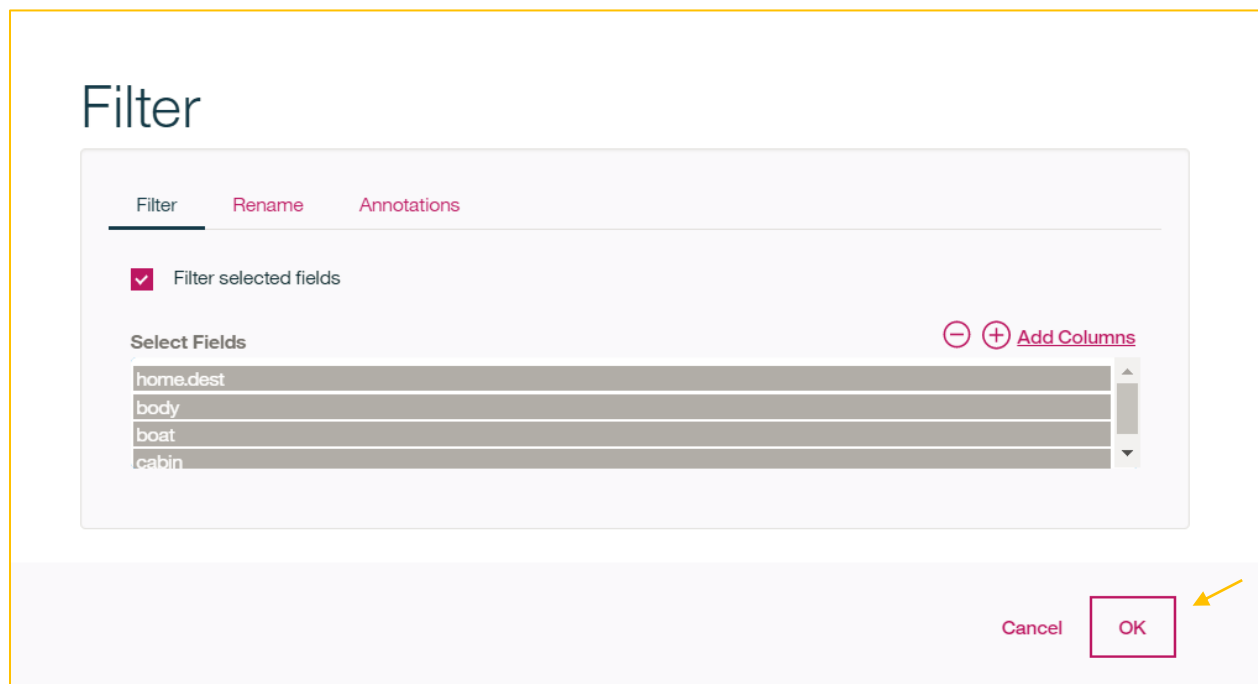
2. Double click on the **Filter** node. In the Filter panel, click on the **Add Columns**.




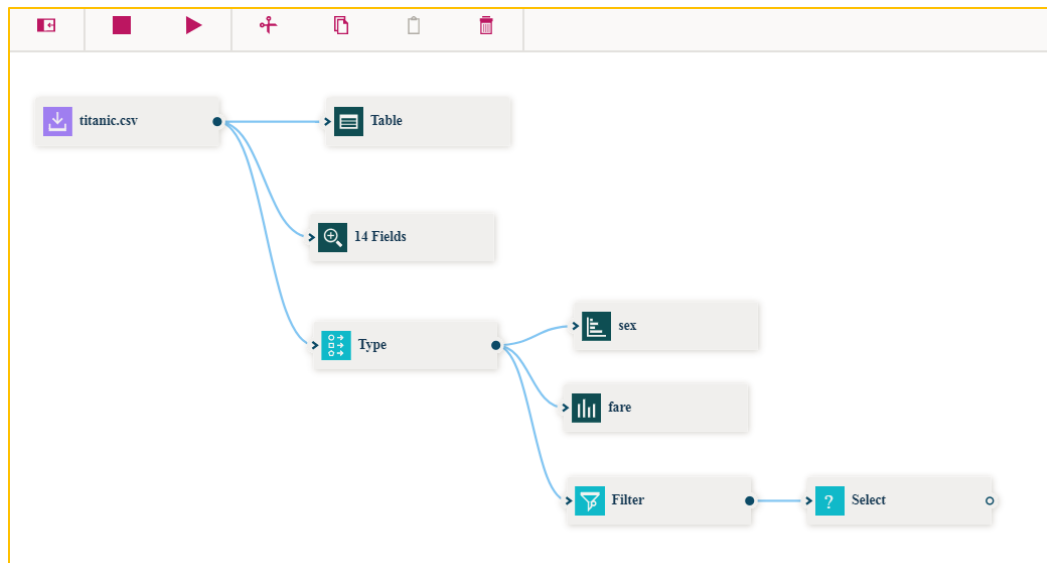
3. Click on the checkboxes adjacent to the **cabin**, **boat**, **body**, and **home.dest** fields, and then click on **Select Fields for Filter**.



- Click **OK** on the Filter panel.



- Add a **Select** node by clicking on the **Record Operations** menu item in the Node palette, and then dragging the **Select** node to the canvas to the right of the **Filter** node. Connect the **Filter** node to the **Select** node. If the Node Palette is not visible, click on the Node Palette icon  first. The canvas should appear as below.



6. Double click on the **Select** node. In the **Select** panel, click on the **Discard** radio button, and re-type in the code shown below in the **Condition** text box, and then click **OK**.

## Select

Settings Annotations

Mode


☐ Include ☒ Discard

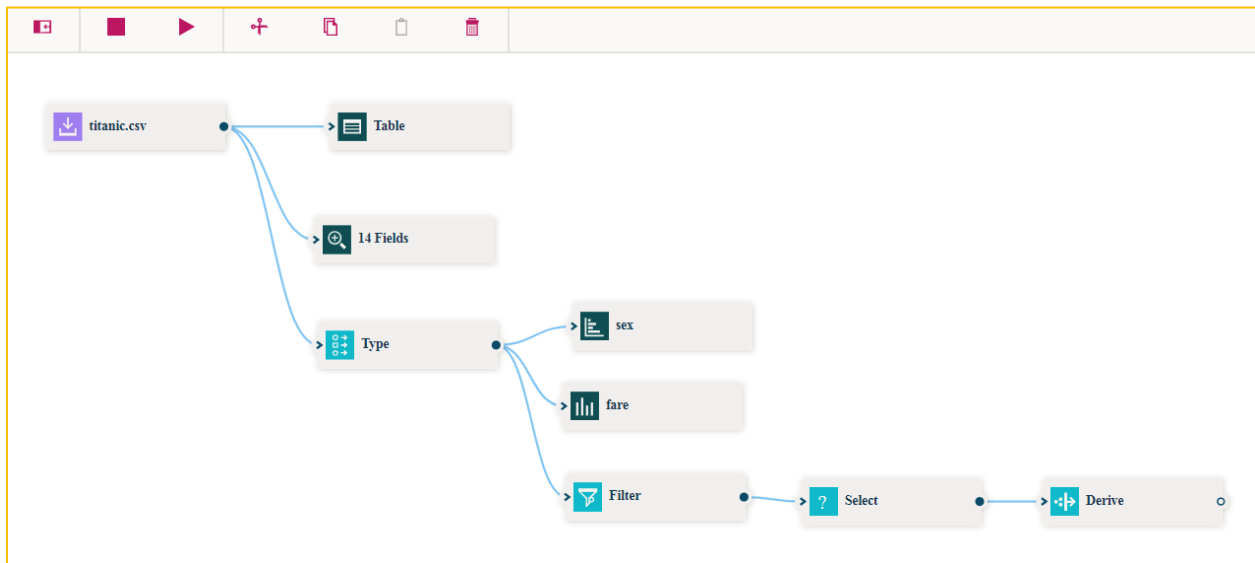
Condition

```
@NULL(age) or embarked==" " or @NULL(fare)
```

Cancel OK


7. Add a **Derive** node to the canvas by clicking on the **Field Operations** menu item in the Node palette, and then dragging the **Derive** node onto the canvas to the right of the

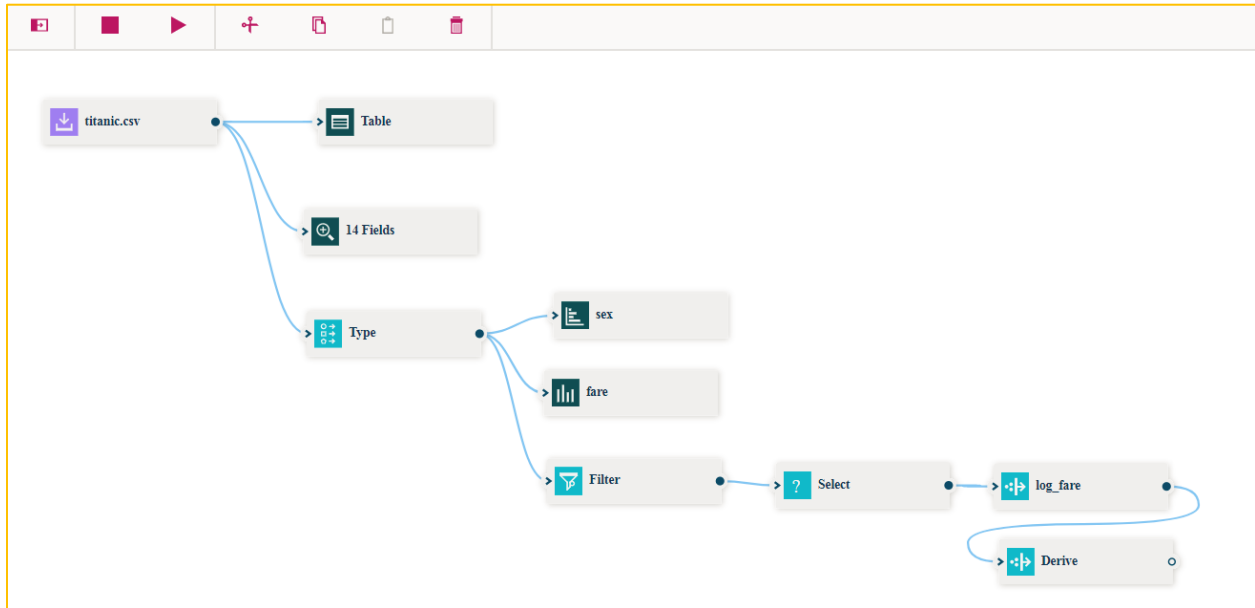
**Select** node. If the Node Palette is not visible, click on the Node Palette icon  first. Connect the **Select** node to the **Derive** node. The canvas should appear as below.



8. Double click on the **Derive** node. Click on the **Single** radio button, enter log\_fare for the **Derive field**, enter the following code in the **Expression** text box, select **Range** for the measurement and click OK.

if (fare /=0) then log(fare) else 0 endif

9. Binning of continuous fields is a technique sometimes used in preparing data for modeling. We will bin the age field, and the log\_fare field. Add a **Derive** node by clicking on the **Field Operations** menu item in the Node palette and dragging the **Derive** node on the canvas underneath the log\_fare **Derive** node. If the Node Palette is not visible, click on the Node Palette icon  first. Connect the log\_fare **Derive** node to the newly added **Derive** node. The canvas should appear as below.



10. Double click on the **Derive** node. Click on the **Single** radio button, enter age\_bucket for the **Derive** field, enter the following code in the **Expression** text box, select OrderedSet for the **Measurement**, and then click **OK**.

```
if age >=0 and age < 6 then 0
else if age >=6 and age < 12 then 1
else if age >=12 and age < 18 then 2
else if age >=18 and age < 40 then 3
else if age >=40 and age < 65 then 4
else if age >=65 and age < 80 then 5
else 6
endif
endif
endif
endif
endif
endif
```

# Derive

Settings

Annotations

Mode

☒ Single ☐ Multiple

Derive field

age\_bucket

118

Expression

```
if age >=0 and age < 6 then 0
else if age >=6 and age < 12 then 1
else if age >=12 and age < 18 then 2
else if age >=18 and age < 40 then 3
```

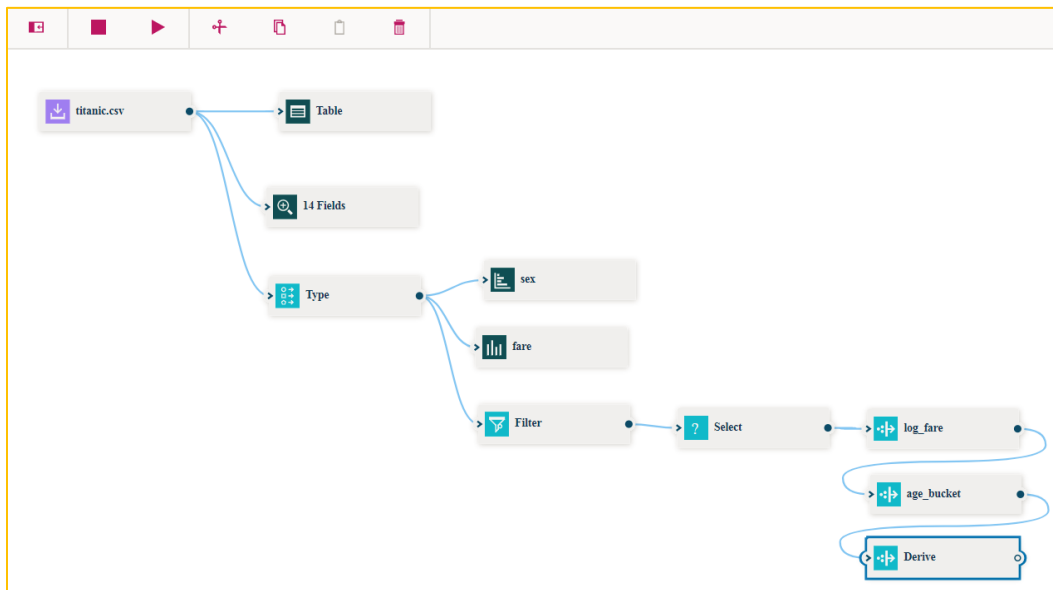
Measurement

OrderedSet

Cancel

OK

11. Add a **Derive** node by clicking on the **Field Operations** menu item in the Node palette and dragging the **Derive** node onto the canvas underneath the age\_bucket **Derive** node. Connect the age\_bucket **Derive** node to the newly created **Derive** Node. The canvas should appear as below.



12. Double click the **Derive** node. In the **Derive** panel, click on the **Single** radio button, enter fare\_bucket in the **Derive field**, enter the following code in the **Expression** text box, click on OrderedSet for the **Measurement**, and click on **OK**.

```
if log_fare < 0 then 0
else if log_fare > 8 then 9
else to_integer(log_fare)+1
endif
endif
```

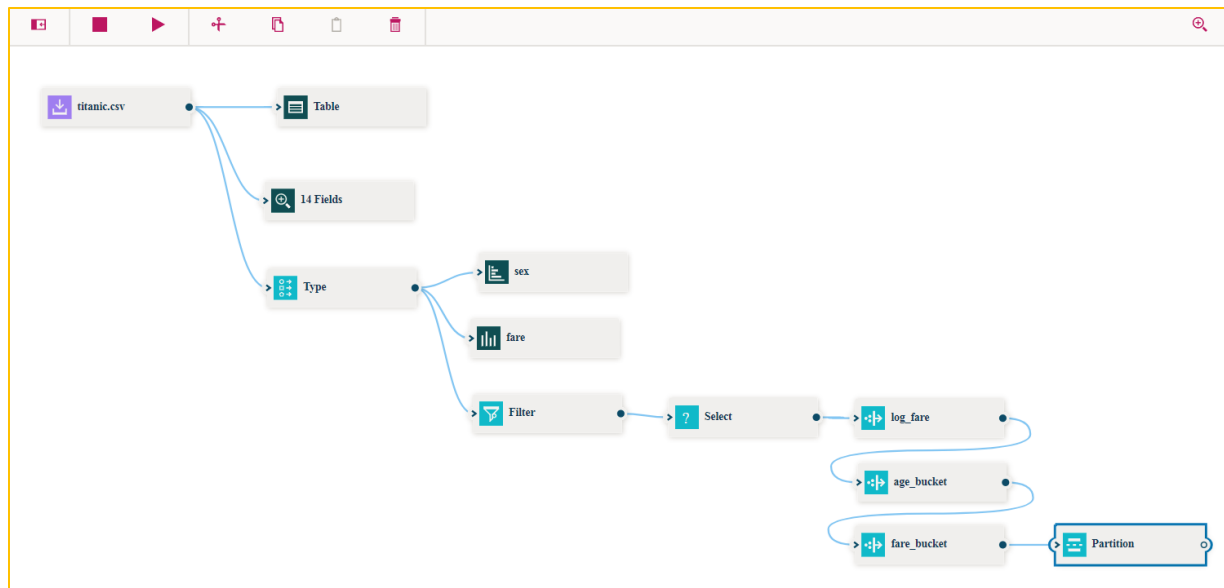


## Step 3.5 Modeling and Evaluation

Now that the data is prepared, we can start the modeling effort. First, we will add a **Partition** node to divide the data set into Training and Testing sets. In addition, a **Type** node is needed prior to modeling to type the new data fields that were created. Then we will add a **Logistic Regression** node, and use the Training set to train the model. Finally, we will add an **Analysis** node to evaluate the results.

1. Add a **Partition** node by clicking on the Field Operations menu item in the Node palette and dragging the **Partition** node onto the canvas to the right of the fare\_bucket **Derive** node. Connect the fare\_bucket **Derive** node to the **Partition** node. The canvas should appear as below.





2. Double click on the Partition node. Set the Training Partition to 70 and the Test Partition to 30. Leave the other defaults, and click on OK.

## Partition

SETTINGS

Annotations

Derived Field Name

Partition

119

Training Partition

70

Testing Partition

30

☐ Create validation partition

☒ Repeatable partition assignment

Seed

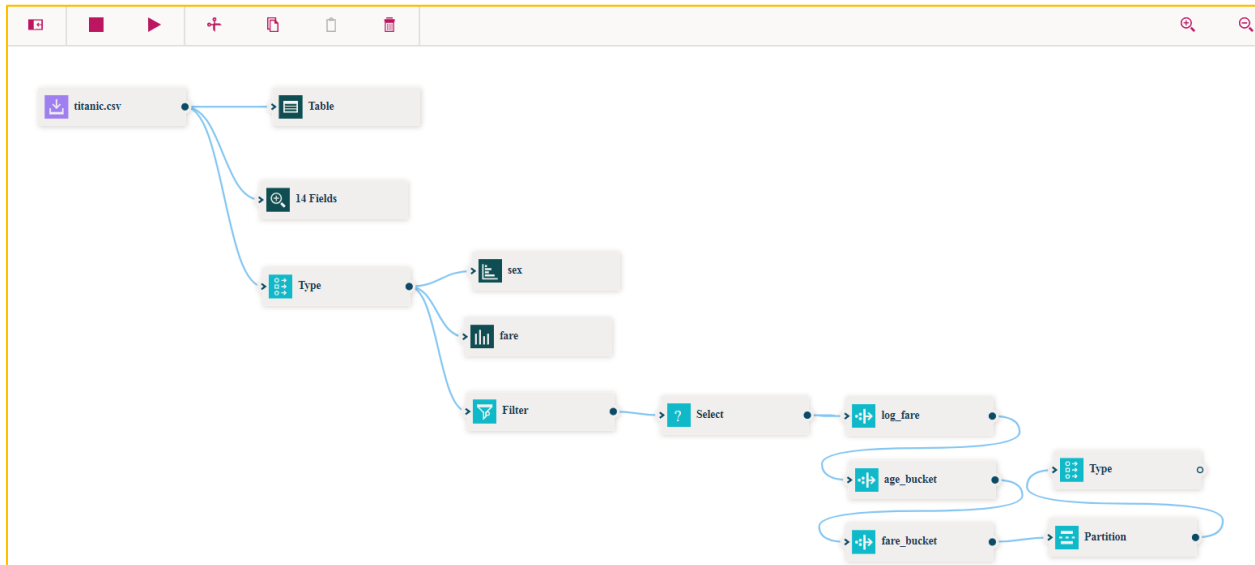
1234567

☐ Use unique field to assign partitions

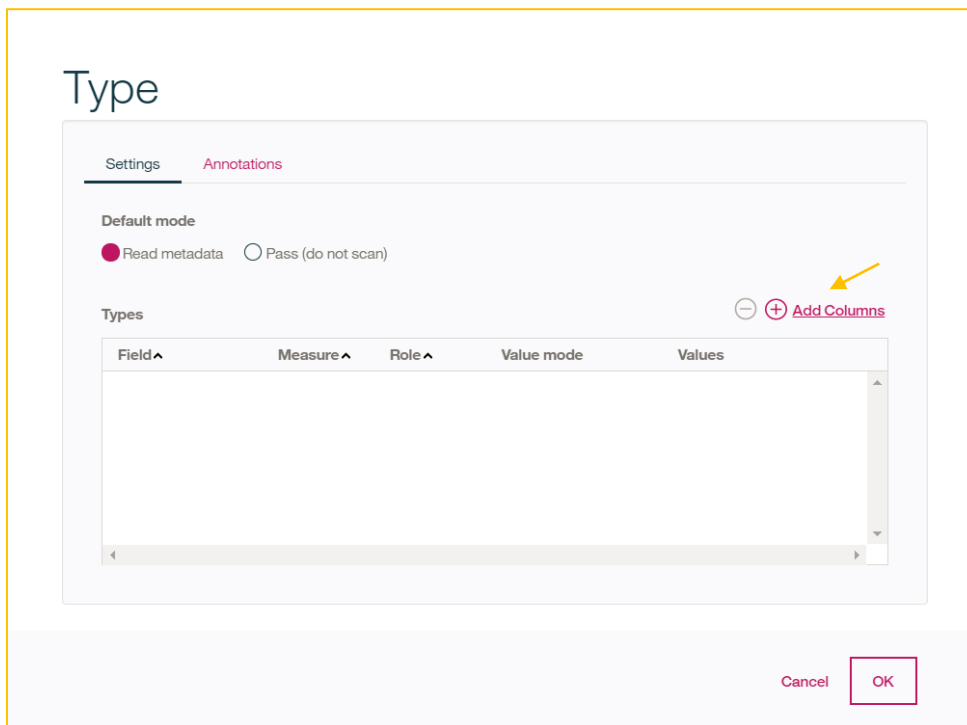
Cancel

OK

3. Add a **Type** node by clicking on the **Field Operations** in the Node palette and dragging the **Type** node onto the canvas above the **Partition** node. Connect the **Partition** node to the **Type** node. The canvas should appear as below.









4. Double click on the **Type** node. Click on **Add Columns**.








5. Click on checkboxes adjacent to the log\_fare, age\_bucket, fare\_bucket, and Partition fields. Click on Select Fields for Type.

# Type

 **Select Fields for Type** **Reset** 

Search in column Field name  Filter:   

<input type="checkbox"/>	Field name ^	Data type ^
<input type="checkbox"/>	embarked	 string
<input checked="" type="checkbox"/>	log_fare	 double
<input checked="" type="checkbox"/>	age_bucket	 integer
<input checked="" type="checkbox"/>	fare_bucket	 integer
<input checked="" type="checkbox"/>	Partition	 string

6. For the **Partition** field, select **Set** for the **Measurement**. For the log\_fare, select **Range** for the **Measurement**. For the fare\_bucket field, select **OrderedSet** for the **Measurement**, and for the age\_bucket, select **OrderedSet** for the **Measurement**, and click **OK**.

# Type

**Settings** **Annotations**

**Default mode**  
☒ Read metadata ☐ Pass (do not scan)

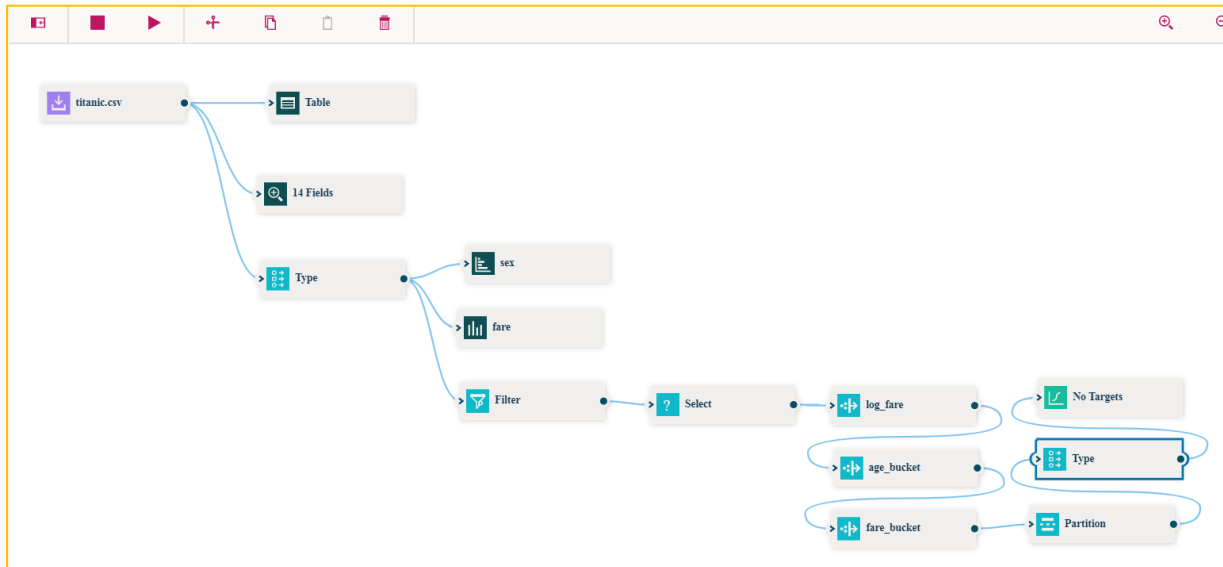
**Types** - + Add Columns

Field ^	Measure ^	Role ^	Value mode	Values
Partition	<b>Set</b>	Input	Specify	1_Training,2 ...
log_fare	<b>Range</b>	Input	Specify	0,0,6.2389€ ...
fare_bucket	<b>Ordered Set</b>	Input	Specify	1,7 ...
age_bucket	<b>Ordered Set</b>	Input	Specify	0,6 ...

Cancel

**OK**

7. Add a **Logistic Regression** node by clicking on the **Modeling** menu item in the Node palette and dragging the **Logistic** node onto the canvas above the **Type** node. Connect the **Type** node to the **Logistic Regression** node. The canvas should appear as below.



8. Double click on the **Logistic Regression** node. Click on the checkbox next to **Use custom field roles**, select **survived** for the **Target**, select **Partition** for the **Partition**, and click on **Add Columns** to add the input fields.

## No Targets

Fields

Annotations

☒ Use custom field roles

Target  
survived

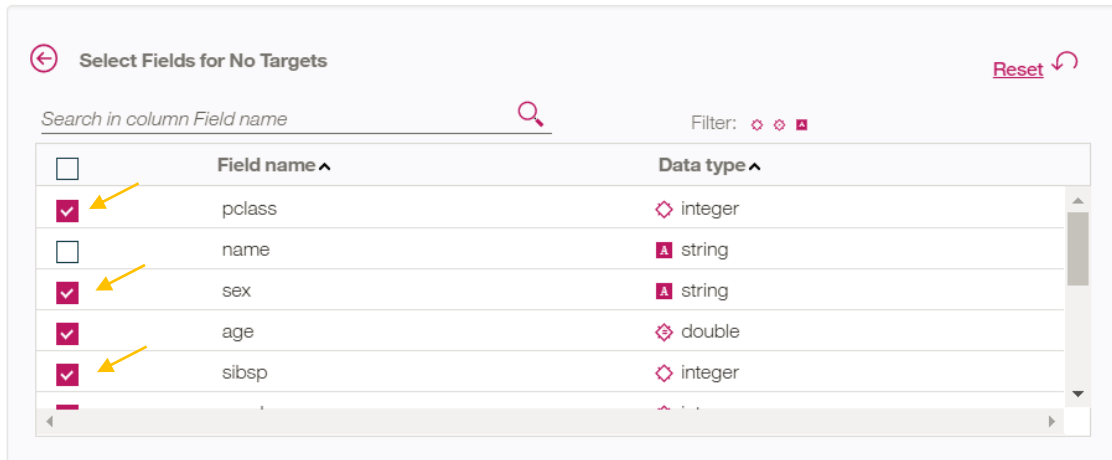
Inputs

Partition  
Partition

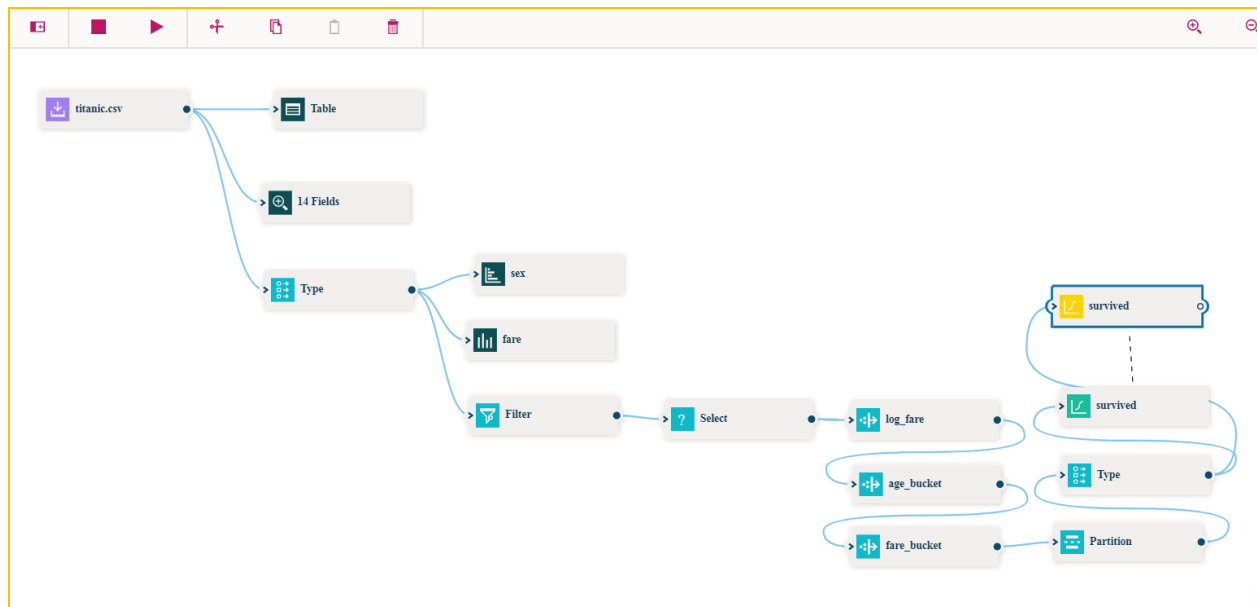
Cancel OK

- Click on the checkboxes next to pclass, sex, sibsp, parch, embarked, age\_bucket, fare\_bucket fields, and then click the arrow to the left of the **Select Fields for No Targets**. Note that both the age field and the fare field appear to get automatically checked when the selection is made for age\_bucket and fare\_bucket. This appears to be a bug.

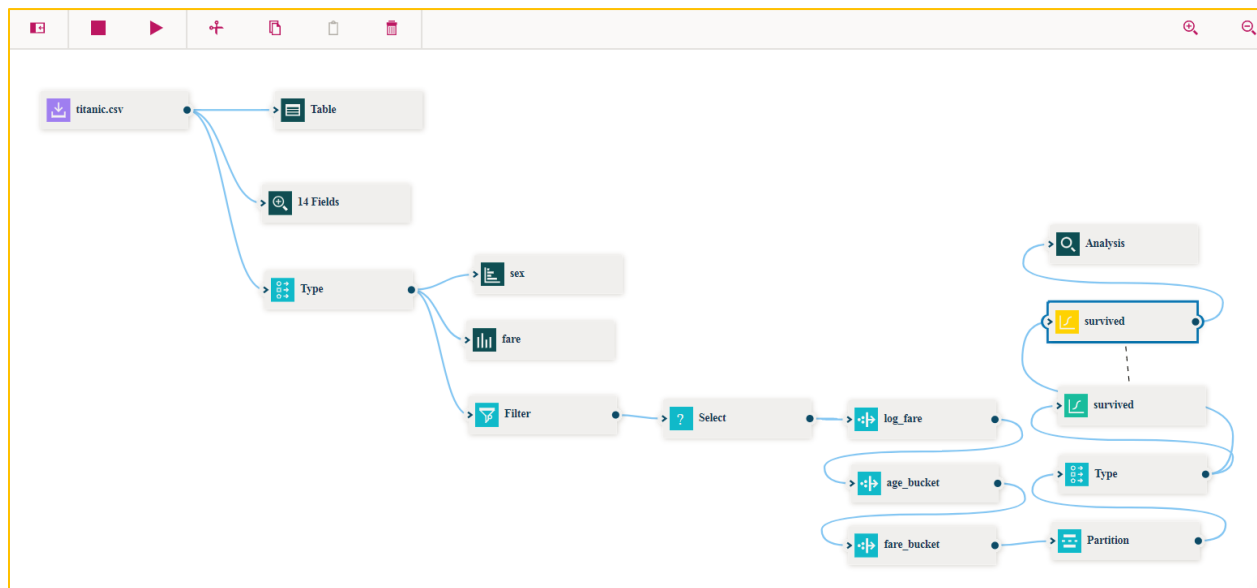
## No Targets



- Right click on the **Logistic Regression** node and then click **Run**. A **Logistic Regression** “nugget will be created” connected by a dotted line to the **Logistic Regression** node. Drag the nugget and place it above the **Logistic Regression** node. The canvas should appear as below.



11. Add an **Analysis** node by clicking on the **Outputs** menu item in the Node palette and dragging the **Analysis** node onto the canvas above the nugget icon. Connect the nugget icon to the **Analysis** node. The canvas should appear as below.



12. Double click on the Analysis node. Click on the Coincidence matrices checkbox, click on the Evaluation metric checkbox, click on the Separation by partition checkbox and click on OK.

## Analysis

Settings

Annotations

☒ Coincidence matrices (categorical targets)  
☐ Performance evaluation  
☒ Evaluation metric (binary classifiers)  
☐ Confidence figures (if available)

Threshold for pct. correct  
90


Improve accuracy multiplier  
2

☒ Separate by partition

Break down by  
pclass  
survived  
name  
sex

Cancel

OK

13. Right click on the Analysis node, and select Run. After completion, double click on the  Analysis of [survived] link in the Outputs tab on the right side of the screen. The results should be similar to those shown below.

Results for output field survived

Individual Models

Comparing \$L-survived with survived

'Partition'	1_Training		2_Testing	
Correct	587	80.08%	251	80.97%
Wrong	146	19.92%	59	19.03%
Total	733		310	

Coincidence Matrix for \$L-survived (rows show actuals)

'Partition' = 1_Training		0	1
0		369	62
1		84	218

'Partition' = 2_Testing		0	1
0		159	28
1		31	92

Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
\$L-survived	0.855	0.709	0.857	0.714