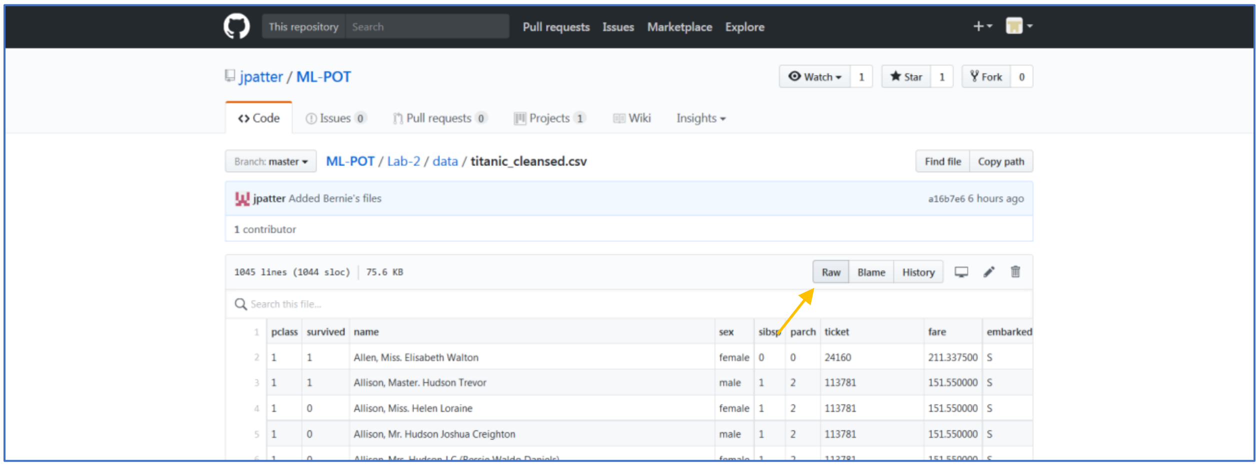
Watson Studio SPSS Modeler Overview

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 to run a machine learning algorithm predicting survivability of a passenger on the Titanic.

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

1. Download the Titanic data file by clicking on the link Titanic Data Set and following the instructions below.

   _Right click on Raw, and click on Save link as ….



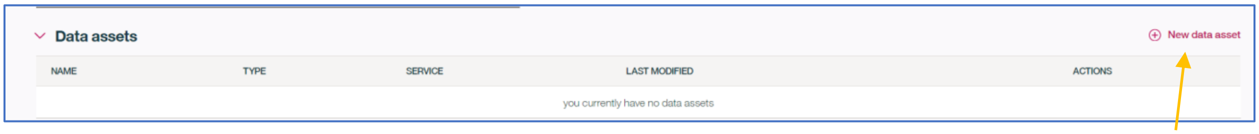
jpatter Added Bernie's files — a16b7e6 6 hours ago

1 contributor

1045 lines (1044 sloc) | 75.6 KB

pclass	survived	name	sex	sibsp	parch	ticket	fare	embarked
1	1	Allen, Miss. Elisabeth Walton	female	0	0	24160	211.337500	S
1	1	Allison, Master. Hudson Trevor	male	1	2	113781	151.550000	S
1	0	Allison, Miss. Helen Loraine	female	1	2	113781	151.550000	S
1	0	Allison, Mr. Hudson Joshua Creighton	male	1	2	113781	151.550000	S

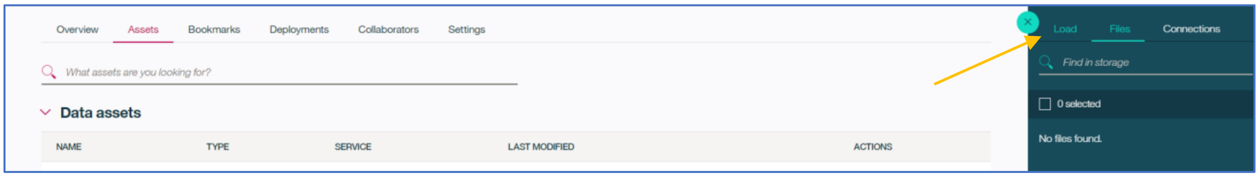
2. Go back to the Watson Studio project. Click on **New data asset**, or if you don't see **New data asset,** click on the icon.



NAME	TYPE	SERVICE	LAST MODIFIED	ACTIONS

you currently have no data assets

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

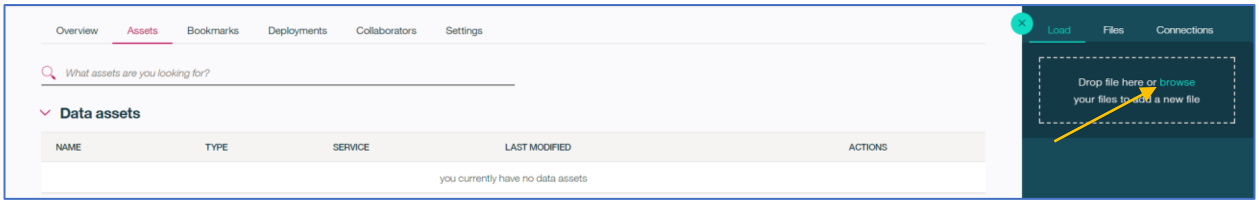


NAME	TYPE	SERVICE	LAST MODIFIED	ACTIONS

□ 0 selected

No files found.

4. Click on **browse**.

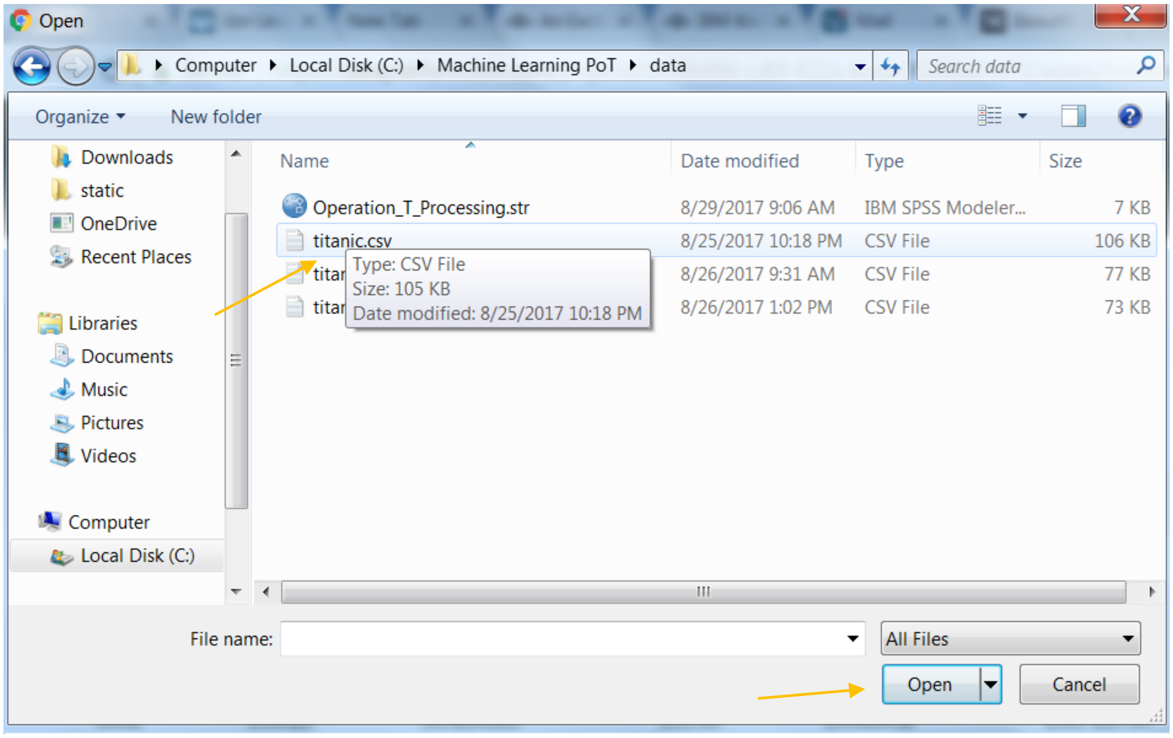


Drop file here or browse your files to add a new file

NAME	TYPE	SERVICE	LAST MODIFIED	ACTIONS

you currently have no data assets

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



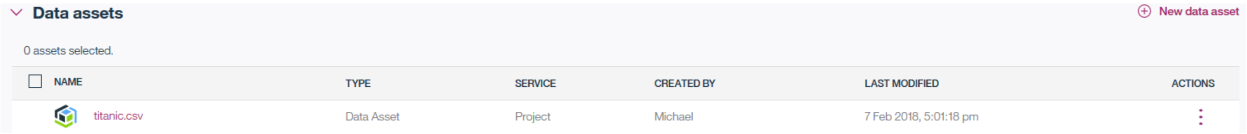
Name	Date modified	Type	Size
Operation_T_Processing.str	8/29/2017 9:06 AM	IBM SPSS Modeler...	7 KB
titanic.csv	8/25/2017 10:18 PM	CSV File	106 KB
titar	8/26/2017 9:31 AM	CSV File	77 KB
titar	8/26/2017 1:02 PM	CSV File	73 KB

Type: CSV File
Size: 105 KB
Date modified: 8/25/2017 10:18 PM

File name:

All Files

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



Data assets

0 assets selected.

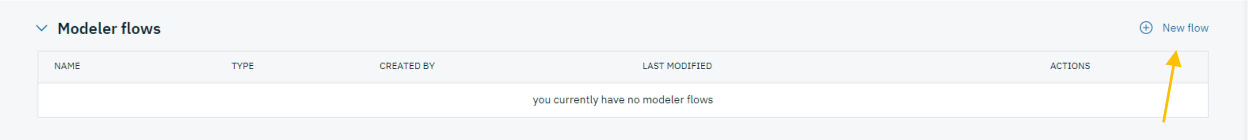
NAME	TYPE	SERVICE	CREATED BY	LAST MODIFIED	ACTIONS
titanic.csv	Data Asset	Project	Michael	7 Feb 2018, 5:01:18 pm	⋮

Step 2: 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.

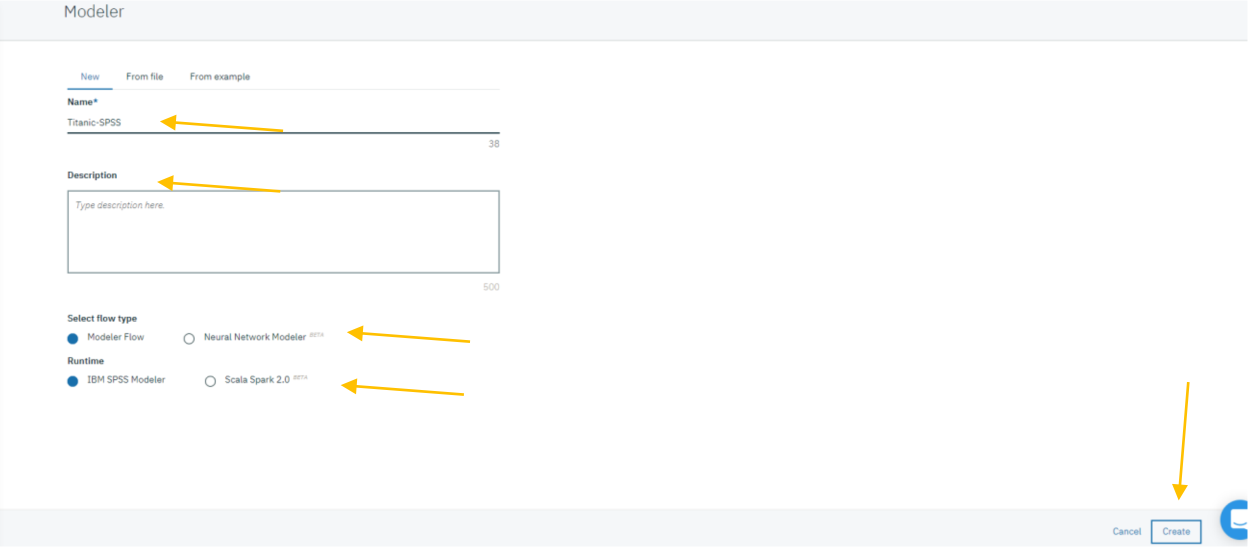
Step 2.1 Create a New Flow and Load the Data

1. In the Watson Studio project, click on **New flow** in the **Modeler flows** section.



you currently have no modeler flows

2. Enter a **Name** for the flow, optionally enter a **Description**, click on Modeler Flow for the **flow type** (should be the default), click on IBM SPSS Modeler for the **Runtime** (should be the default), and click on **Create.**



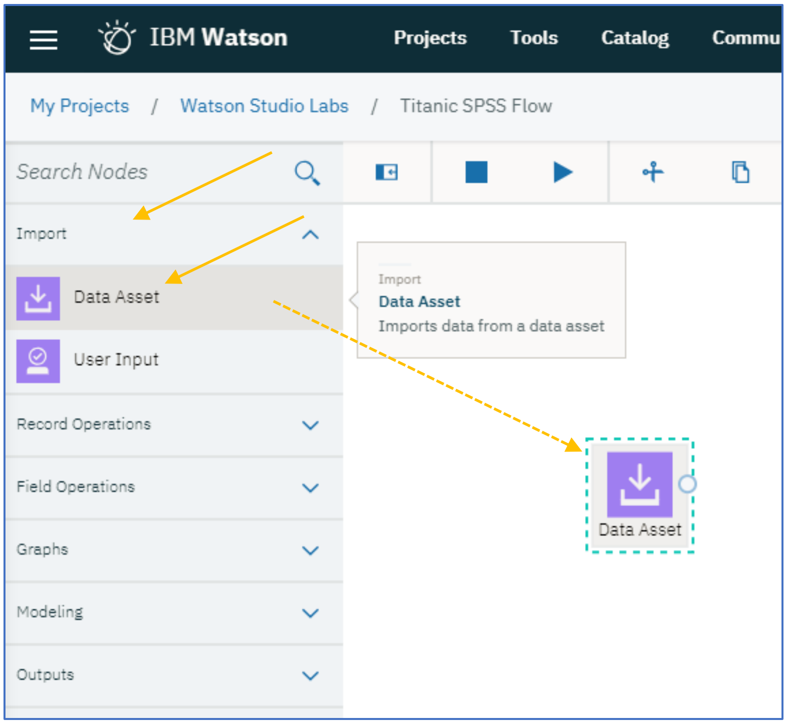
Name*: Titanic-SPSS

Description: Type description here.

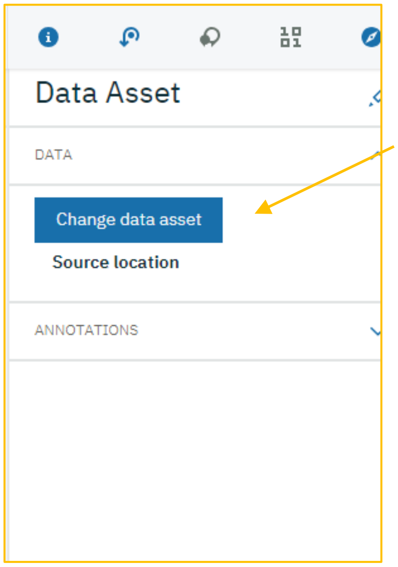
Select flow type: ● Modeler Flow ○ Neural Network Modeler BETA

Runtime: ● IBM SPSS Modeler ○ Scala Spark 2.0 BETA

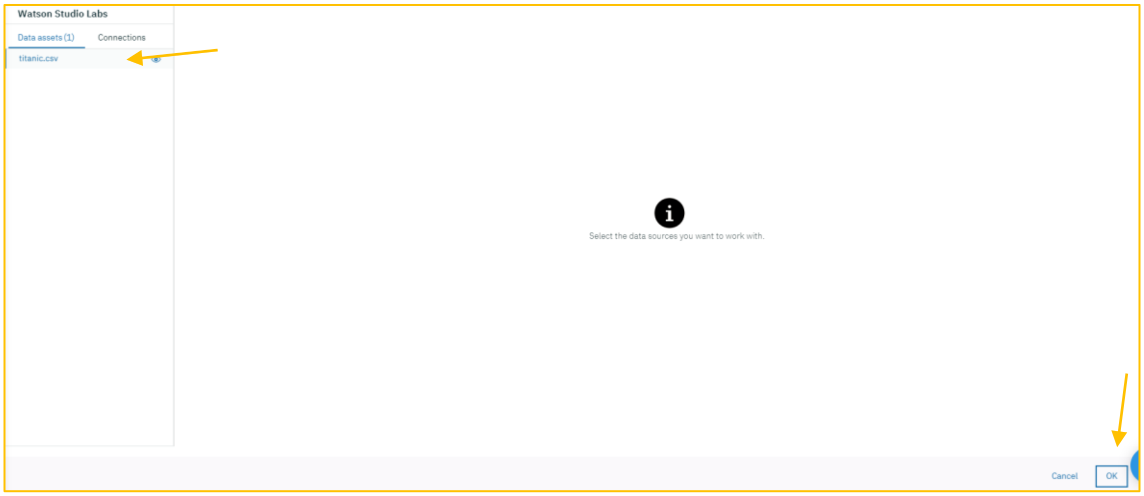
3. This opens the Flow Editor. Click on **Import** and then **Data Asset** and hold the left mouse key on the Data Asset icon and **drag it onto the left side of the canvas**. Release the left mouse key.



4. Double click on the **Data Asset**. In the window pane on the right-hand-side click on **Change data asset**.

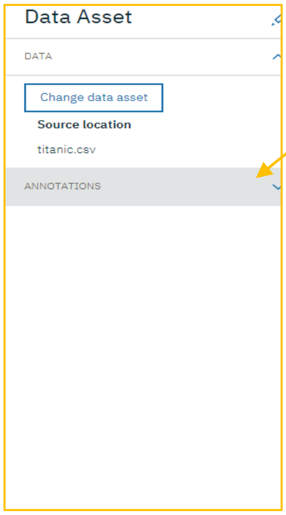


5. Select the **titanic.csv** option, and click **OK**.



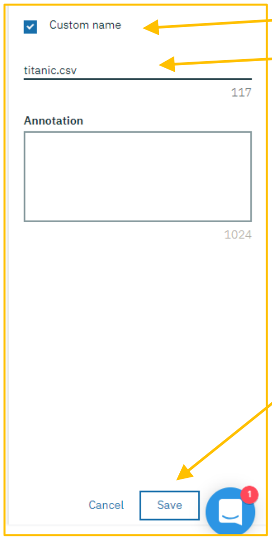
Select the data sources you want to work with.

6. Click on **Annotation.**



Source location
titanic.csv

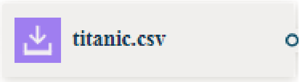
7. Click on **Custom name**, and type **titanic.csv**, and click on **Save**.



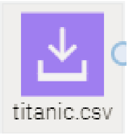
☑ Custom name
titanic.csv
117

8. Note, the depiction of the flow nodes in the user interface has slightly changed from what is shown in this document. The text in the UI is now below the icon, instead of to the right.

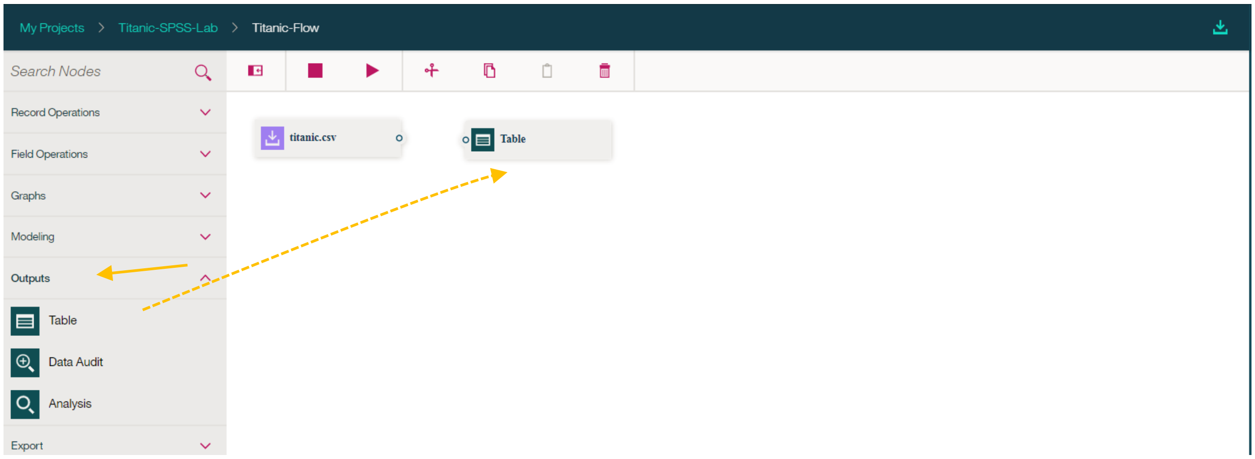
   Old icon --



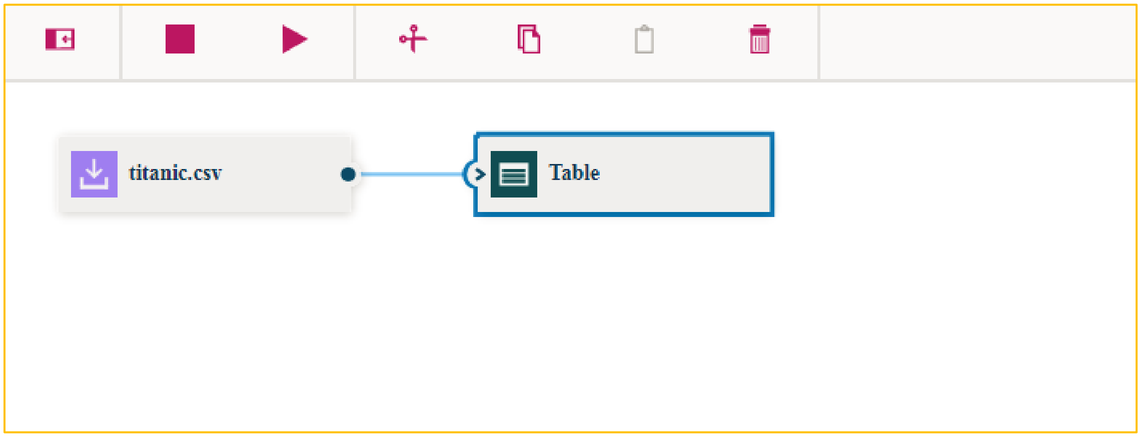
   New icon-



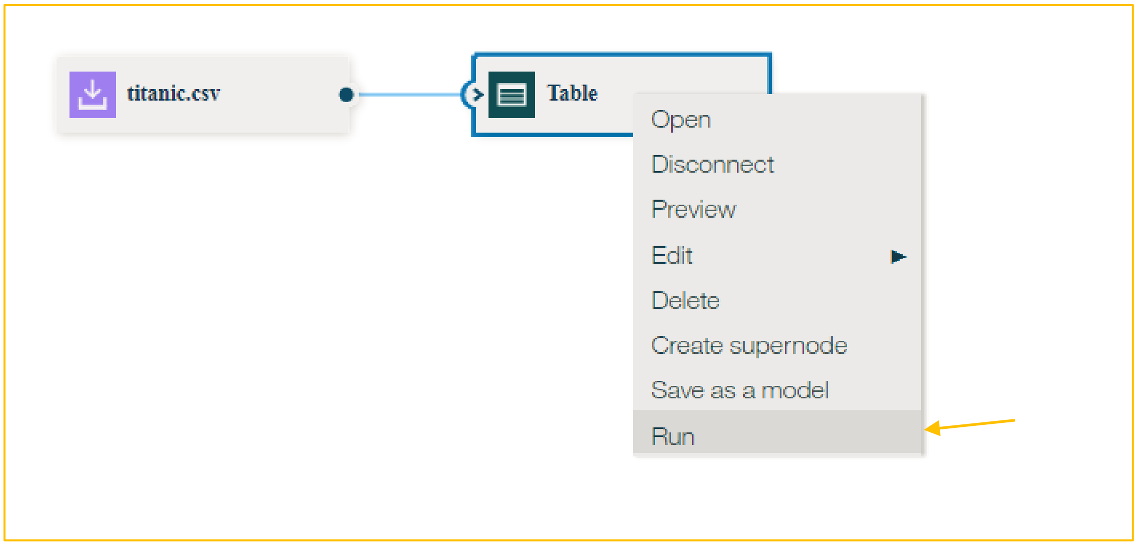
9. Click on the **Outputs** menu item in the Node Palette on the left and then click on the **Table** icon and drag the icon to the right of the titanic.csv icon. The SPSS Table node will display the contents of the csv file. If the Node Palette is not visible, click on the Node Palette icon



10. 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.



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



12. 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. If the Table output selection does not appear, select the icon.

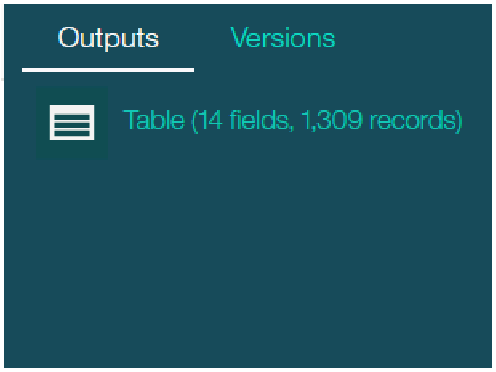
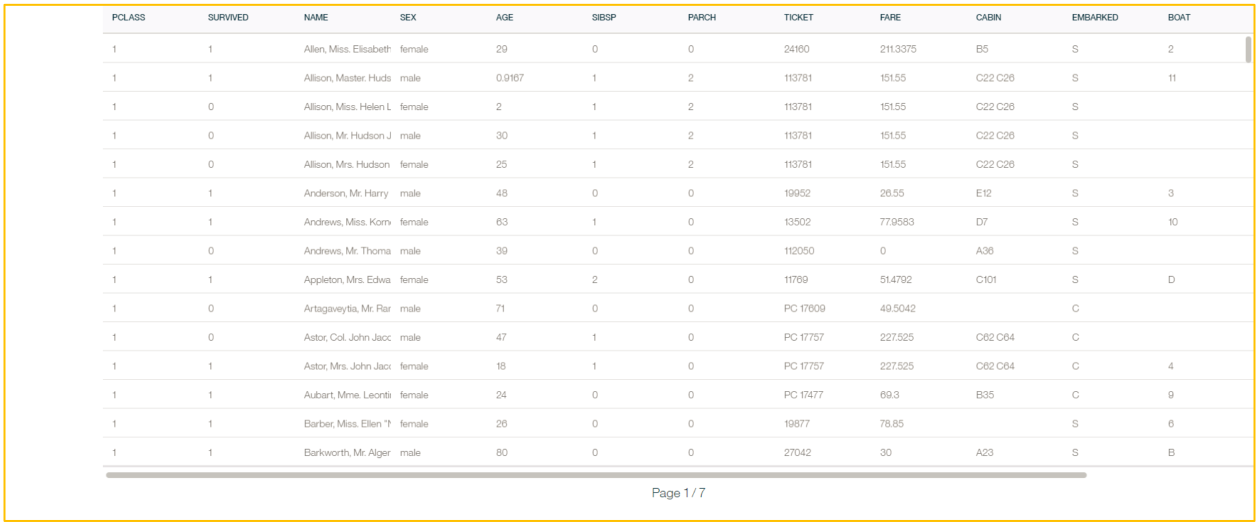


Table (14 fields, 1,309 records)

13. Double click on the Table selection and 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.



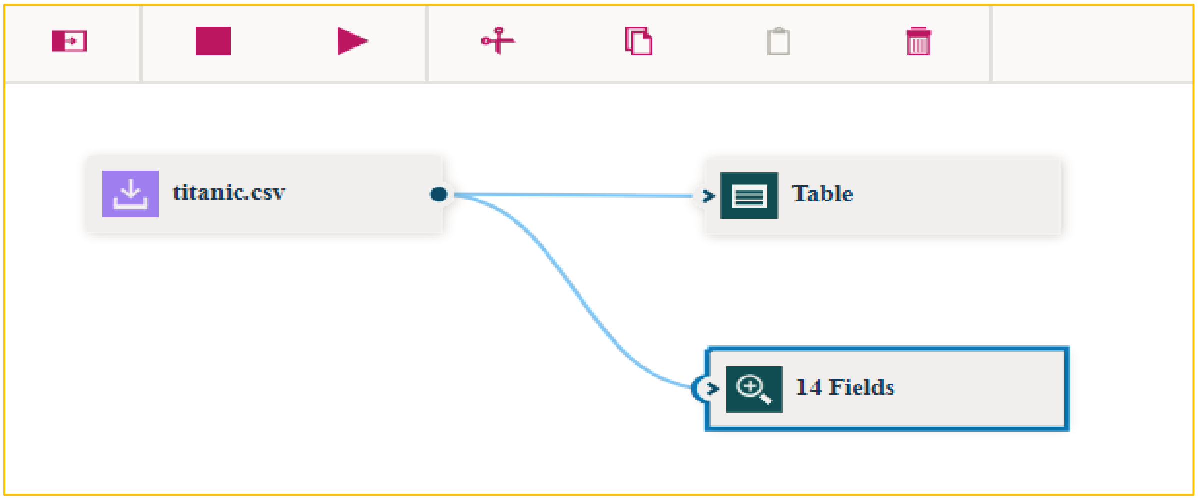
PCLASS	SURVIVED	NAME	SEX	AGE	SIBSP	PARCH	TICKET	FARE	CABIN	EMBARKED	BOAT
1	1	Allen, Miss. Elisabeth	female	29	0	0	24160	211.3375	B5	S	2
1	1	Allison, Master. Huds	male	0.9167	1	2	113781	151.55	C22 C26	S	11
1	0	Allison, Miss. Helen L	female	2	1	2	113781	151.55	C22 C26	S	
1	0	Allison, Mr. Hudson J	male	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	19952	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 17609	49.5042		C	
1	0	Astor, Col. John Jacc	male	47	1	0	PC 17757	227.525	C62 C64	C	
1	1	Astor, Mrs. John Jacc	female	18	1	0	PC 17757	227.525	C62 C64	C	4
1	1	Aubart, Mme. Leonti	female	24	0	0	PC 17477	69.3	B35	C	9
1	1	Barber, Miss. Ellen "N	female	26	0	0	19877	78.85		S	6
1	1	Barkworth, Mr. Alger	male	80	0	0	27042	30	A23	S	B

Page 1 / 7

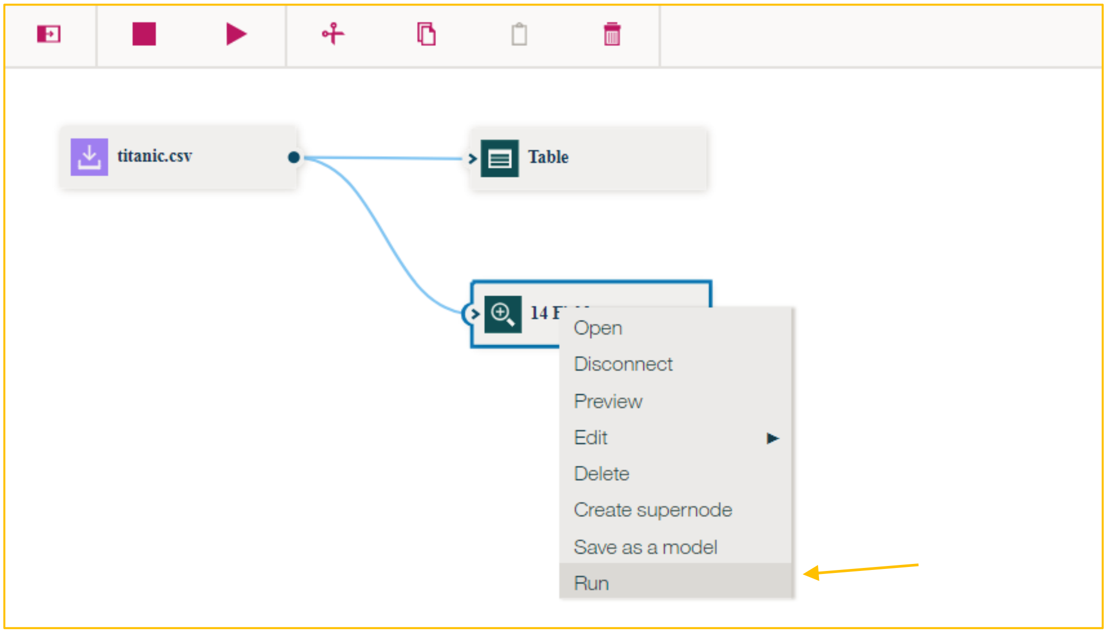
Step 2.2 Explore the Data using the Data Audit Node

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.

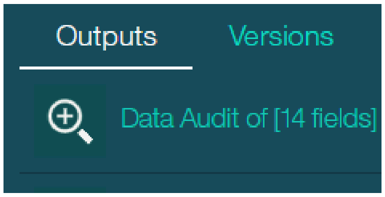
1. Add a **Data Audit** node to the flow clicking on the **Outputs** menu item in the Node Palette, and then dragging the **Data Audit** node to underneath the titanic.csv node. If the Node Palette is not visible, click on the Node Palette icon ⊡. Connect the titanic.csv node to the Data Audit node.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



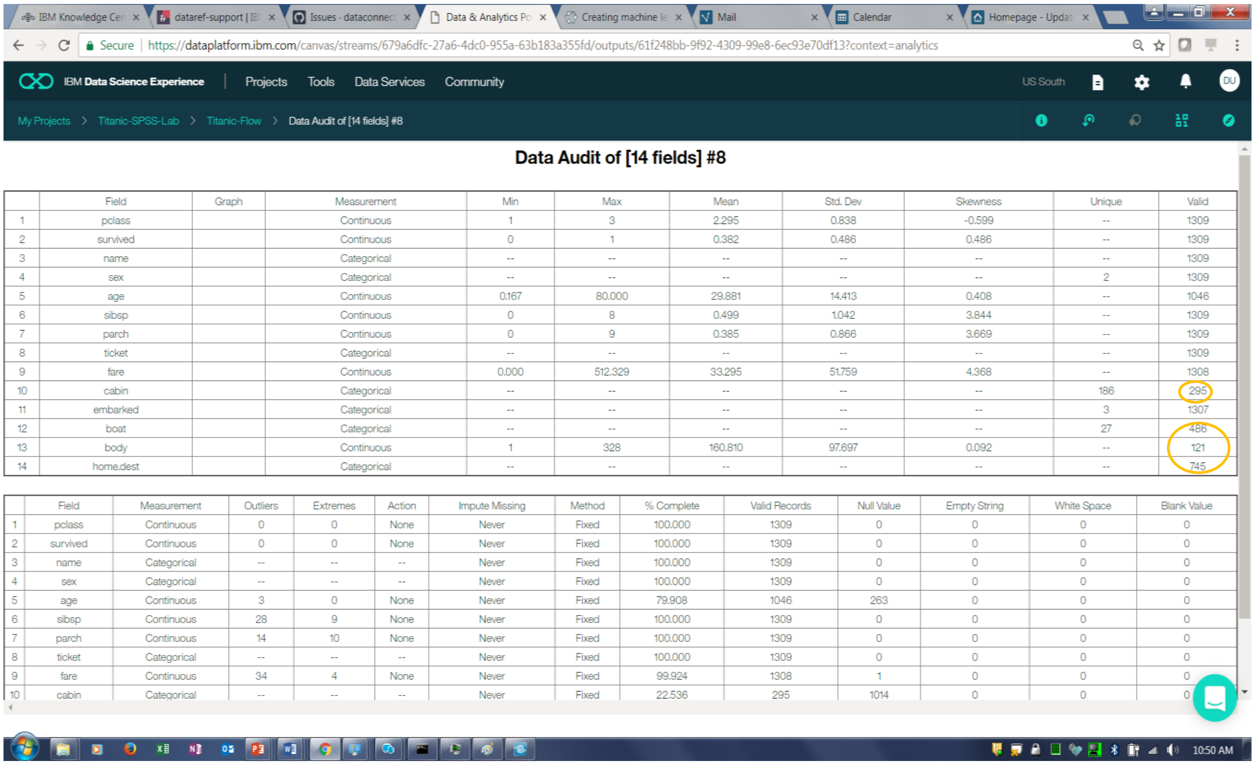
2. Right click on the **Data Audit** node and click **Run**.



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. If the **Outputs** tab doesn’t display, click on the icon.



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.


IBM Knowledge Cen
dataref-support | IB
Issues - dataconnect
Data & Analytics Po
Creating machine le
Mail
Calendar
Homepage - Updat
Secure | https://dataplatform.ibm.com/canvas/streams/679a6dfc-27a6-4dc0-955a-63b183a355fd/outputs/61f248bb-9f92-4309-99e8-6ec93e70df13?context=analytics
IBM Data Science Experience
Projects
Tools
Data Services
Community
US South
DU
My Projects > Titanic-SPSS-Lab > Titanic-Flow > Data Audit of [14 fields] #8
10:50 AM

Data Audit of [14 fields] #8

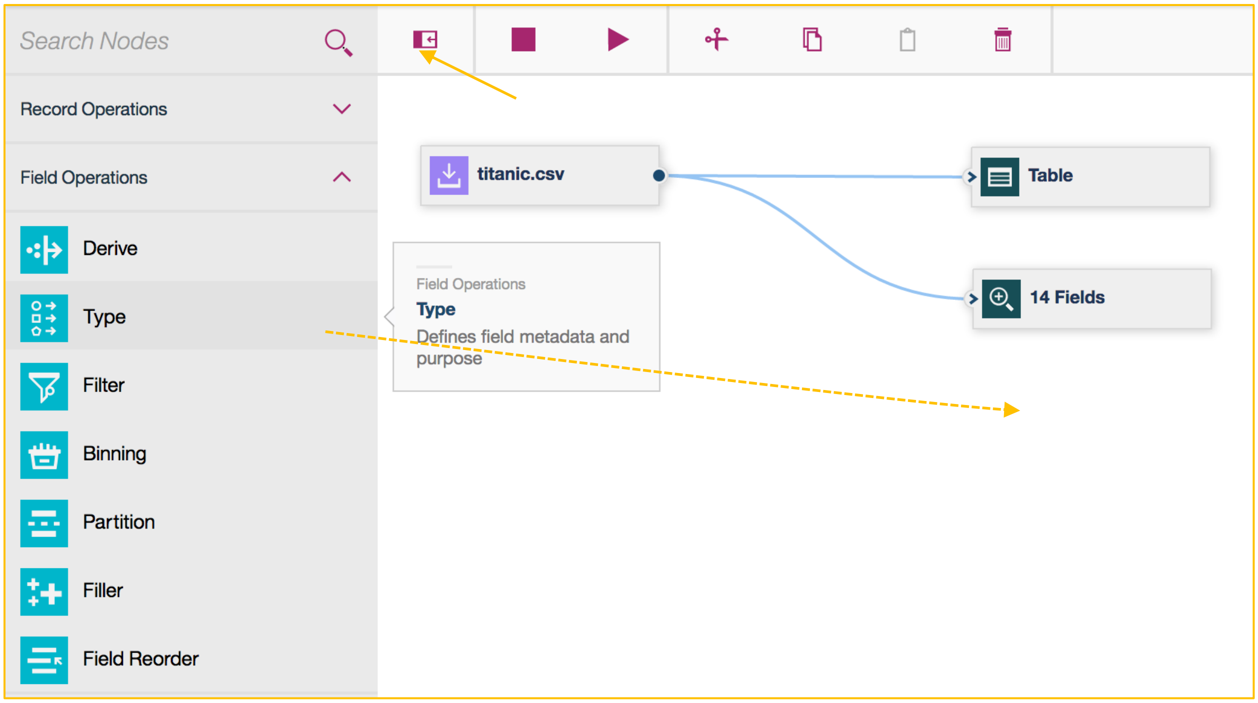
	Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
1	pclass		Continuous	1	3	2.295	0.838	-0.599	--	1309
2	survived		Continuous	0	1	0.382	0.486	0.486	--	1309
3	name		Categorical	--	--	--	--	--	--	1309
4	sex		Categorical	--	--	--	--	--	2	1309
5	age		Continuous	0.167	80.000	29.881	14.413	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

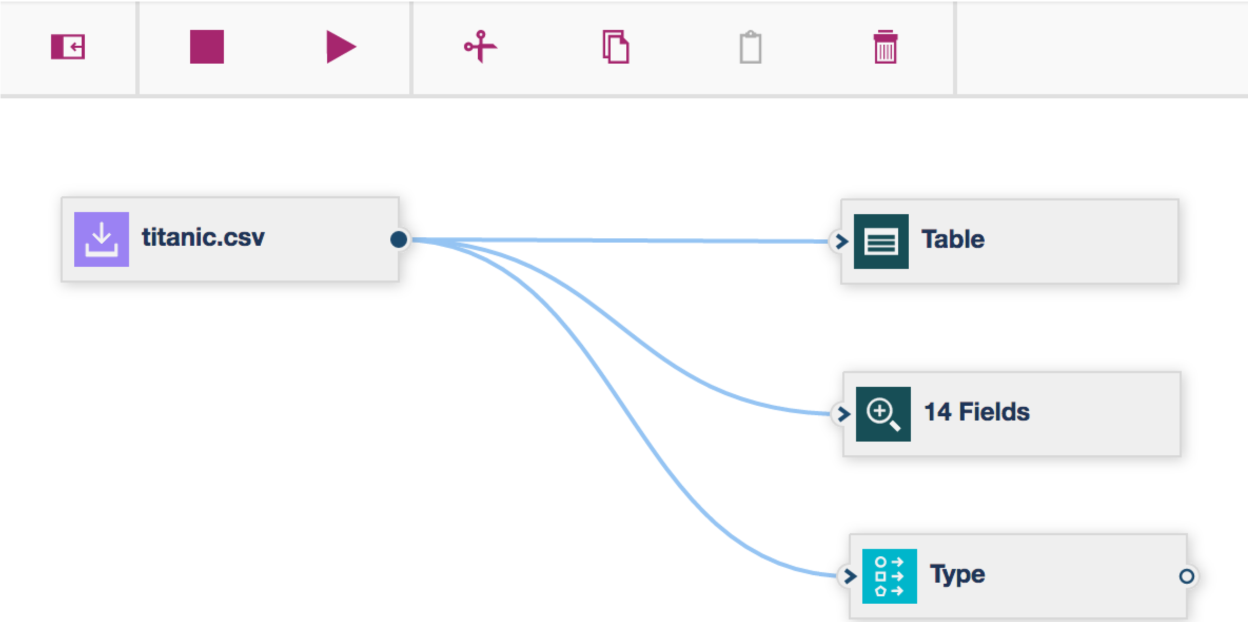
	Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
1	pclass	Continuous	0	0	None	Never	Fixed	100.000	1309	0	0	0	0
2	survived	Continuous	0	0	None	Never	Fixed	100.000	1309	0	0	0	0
3	name	Categorical	--	--	--	Never	Fixed	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 2.3 Explore the Data using Graph Nodes.

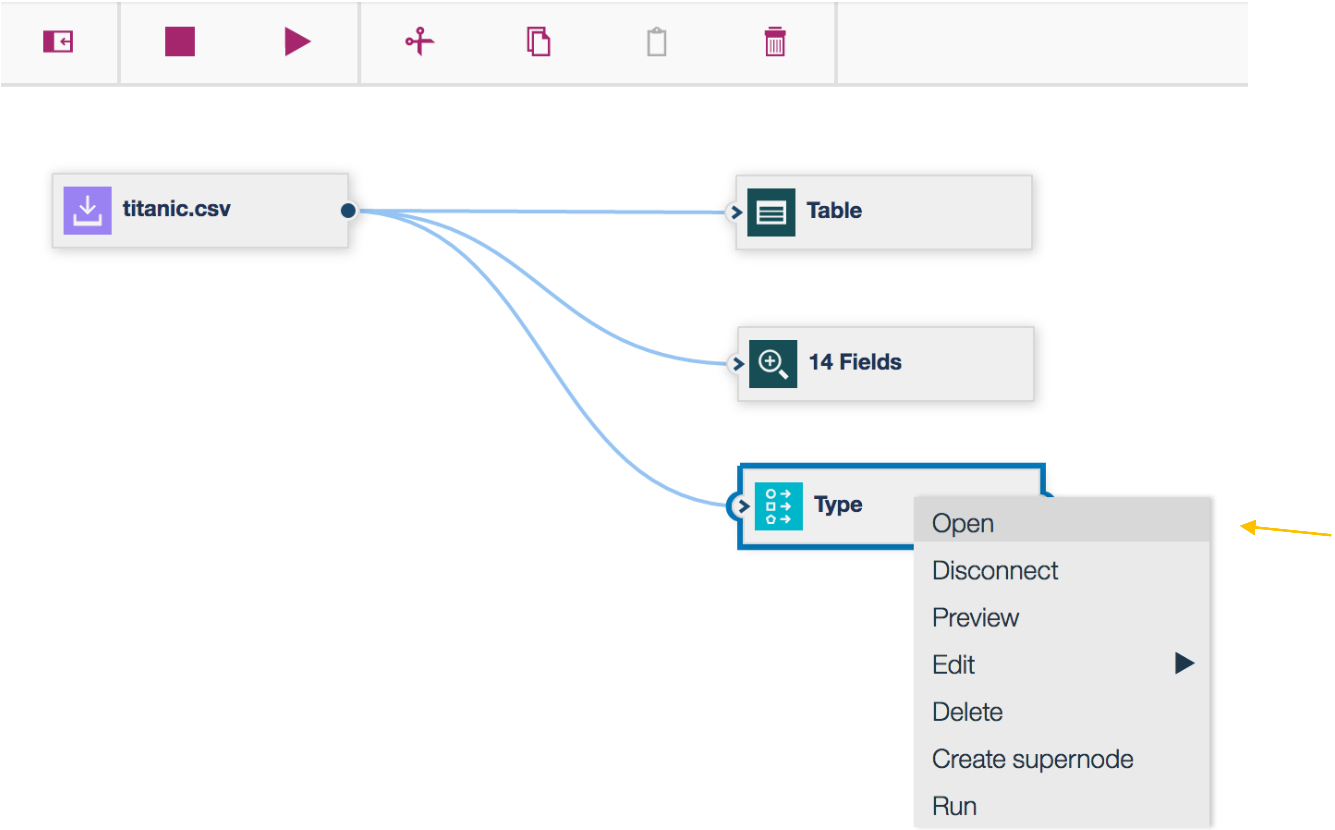
Let’s explore the data using Graph Nodes. The Distribution node, and the Histogram node will be used to explore some of the characteristics of the Titanic Data Set. First, we will add a Type node to the canvas. The Type node specifies field metadata and properties. We will change the measurement property 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.

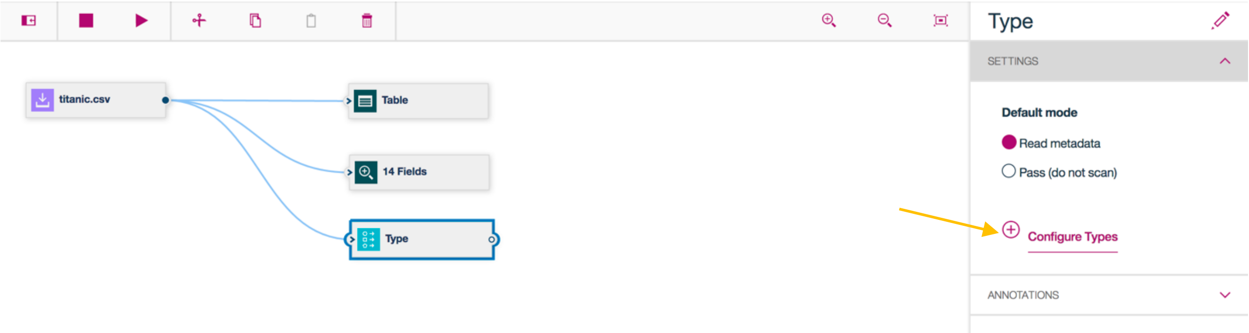




2. Double click on the **Type** node. This will open a **Type** menu pallet on the right side of the screen.



3. Click on the **Settings** dropdown. Select **Configure Types**.

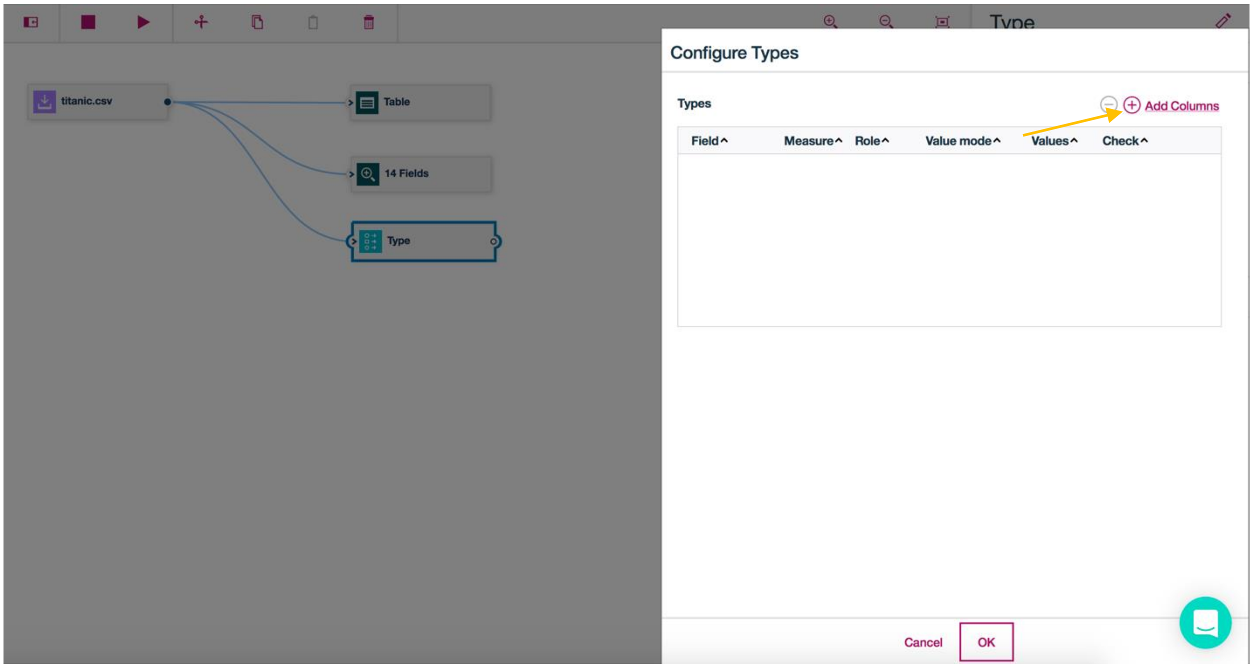


Default mode

● Read metadata

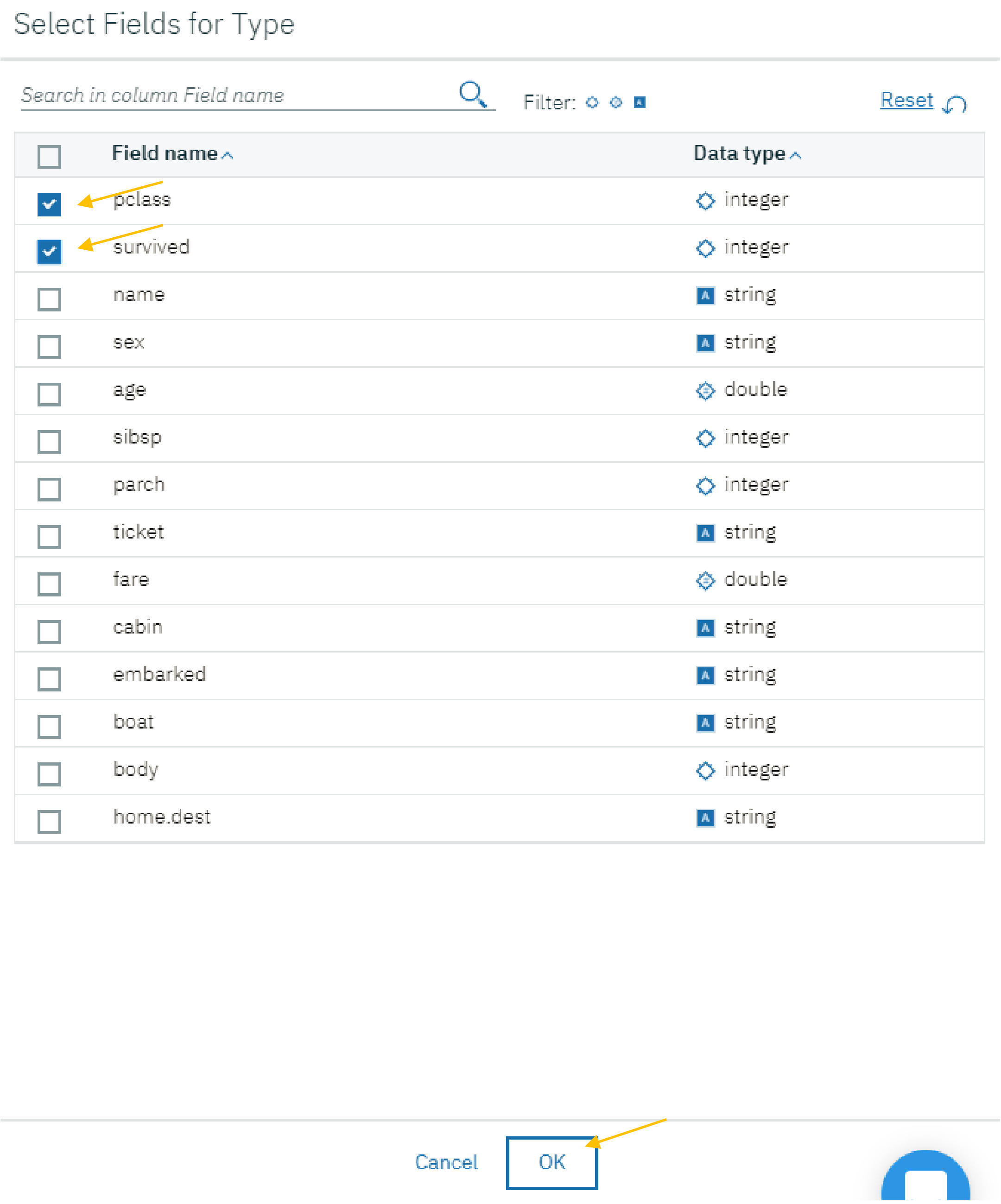
○ Pass (do not scan)

4. Select **Add Columns**.



Field	Measure	Role	Value mode	Values	Check

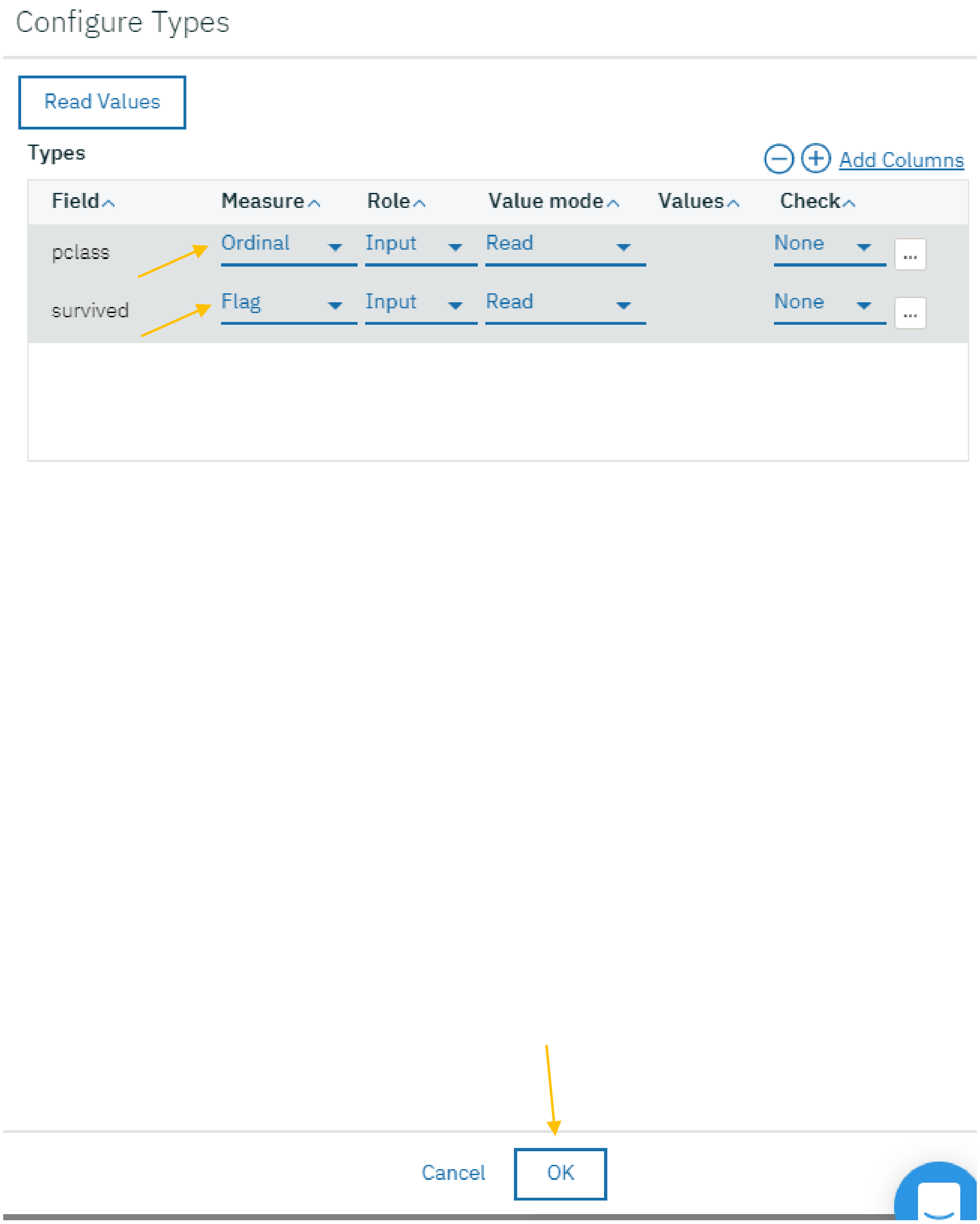
5. Click on the checkboxes adjacent to the **pclass** and survived **fields**, and then click on the left arrow next to **Select Fields for Type**.



Select Fields for Type

☐	Field name	Data type
☑	pclass	◇ integer
☑	survived	◇ integer
☐	name	🅰 string
☐	sex	🅰 string
☐	age	◇ double
☐	sibsp	◇ integer
☐	parch	◇ integer
☐	ticket	🅰 string
☐	fare	◇ double
☐	cabin	🅰 string
☐	embarked	🅰 string
☐	boat	🅰 string
☐	body	◇ integer
☐	home.dest	🅰 string

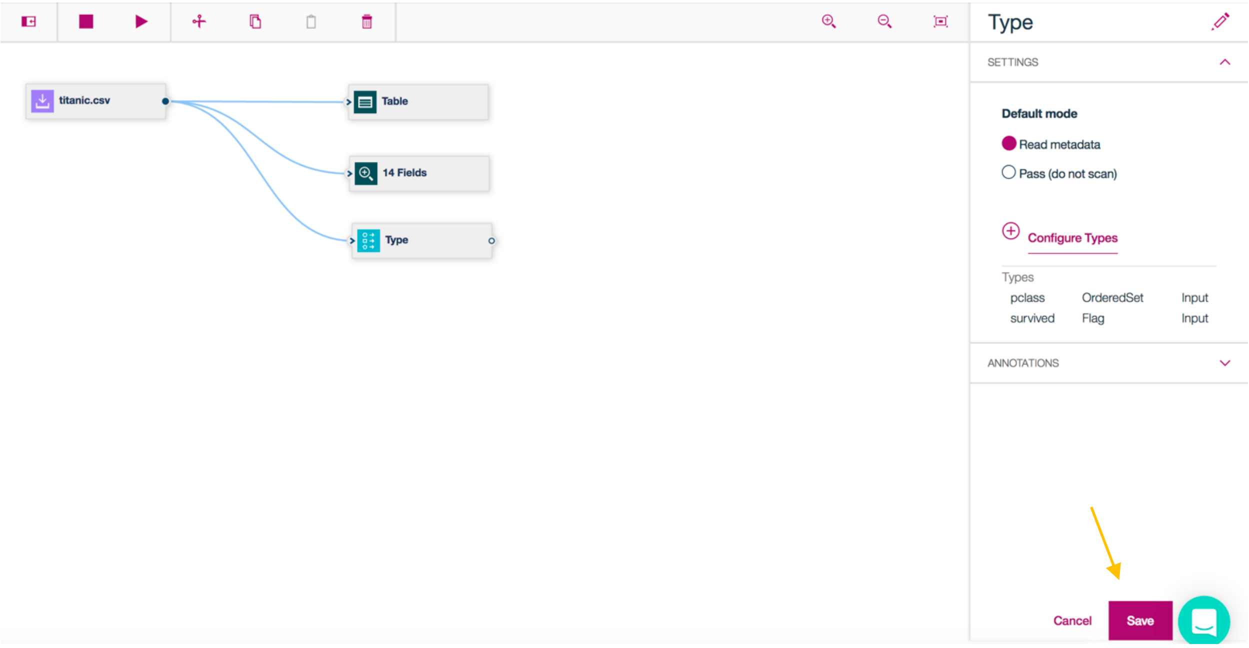
6. Click on the measurement level field for **pclass** and select **Ordinal**. Click on the measurement level field for **survived** and select **Flag**. Click on **OK**.



Configure Types

Field	Measure	Role	Value mode	Values	Check
pclass	Ordinal	Input	Read		None
survived	Flag	Input	Read		None

7. Click on **Save** in the bottom right of the Types pallet.



Type

SETTINGS

Default mode

● Read metadata

○ Pass (do not scan)

Configure Types

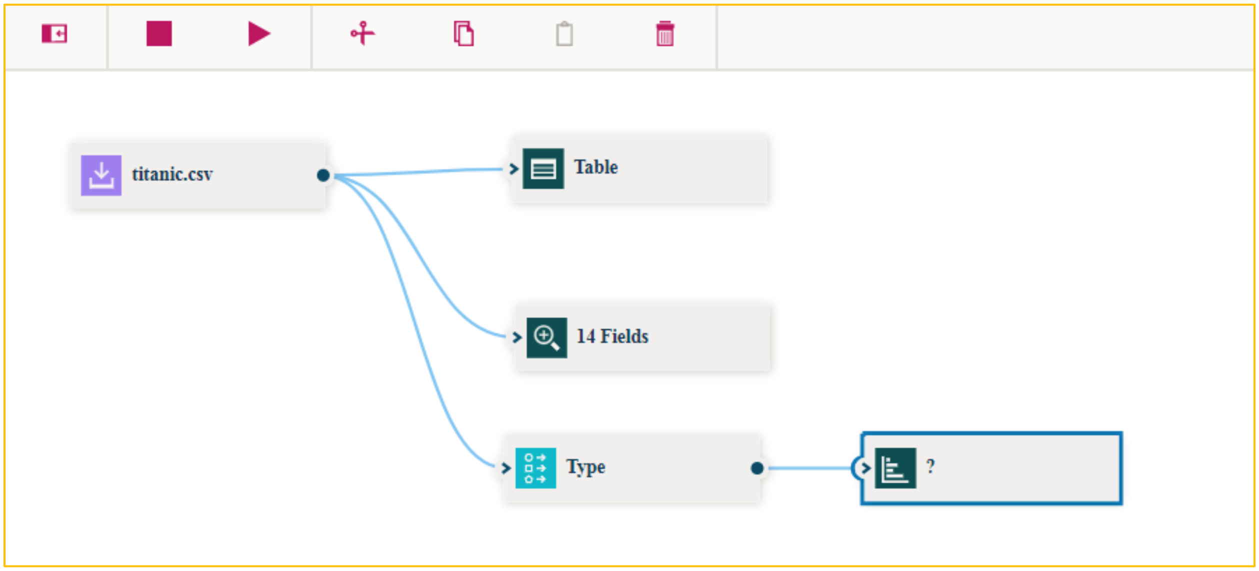
Types

pclass	OrderedSet	Input
survived	Flag	Input

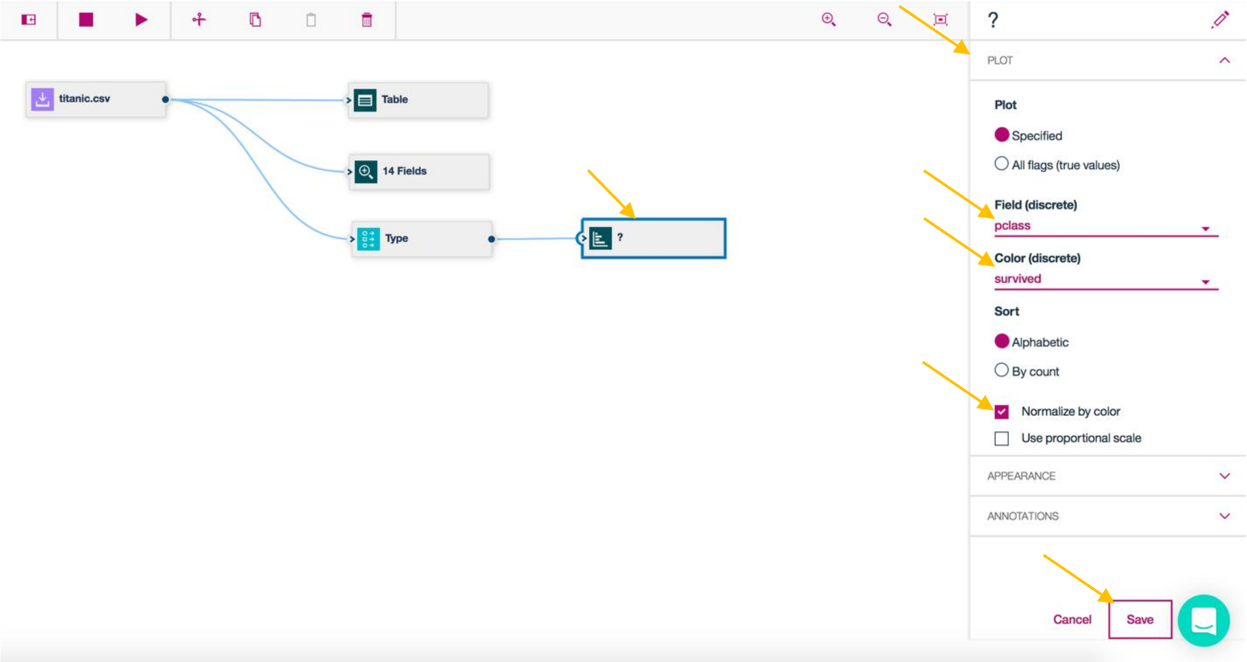
ANNOTATIONS

Cancel Save

8. 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.



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



Plot
● Specified
○ All flags (true values)

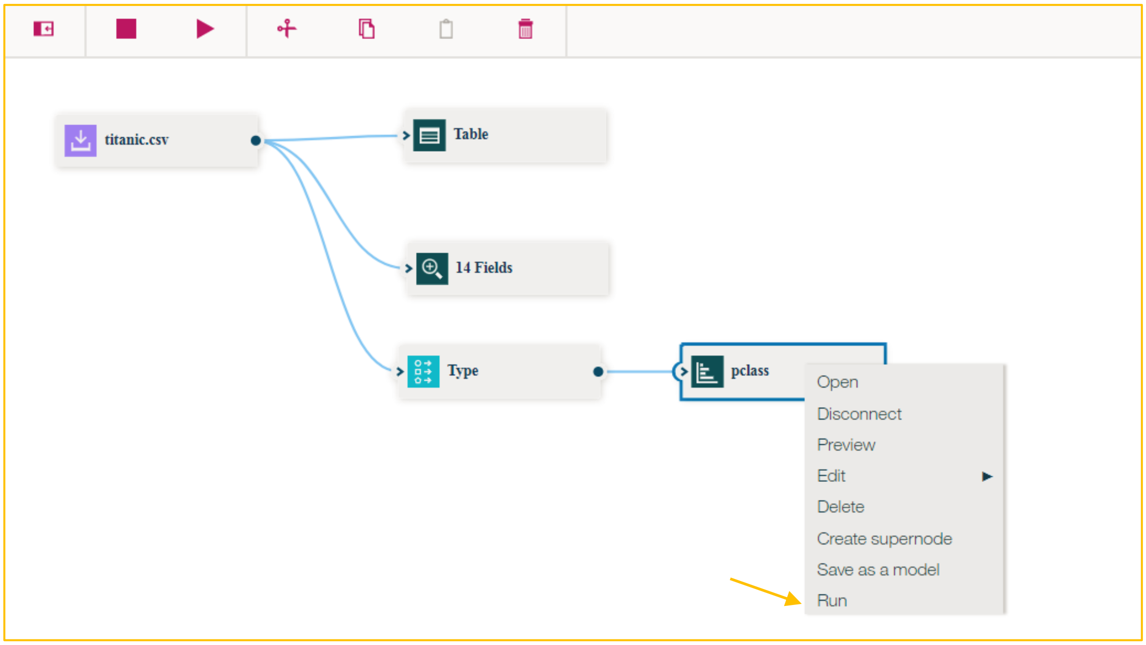
Field (discrete)
pclass

Color (discrete)
survived

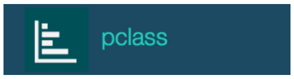
Sort
● Alphabetic
○ By count

☑ Normalize by color
☐ Use proportional scale

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

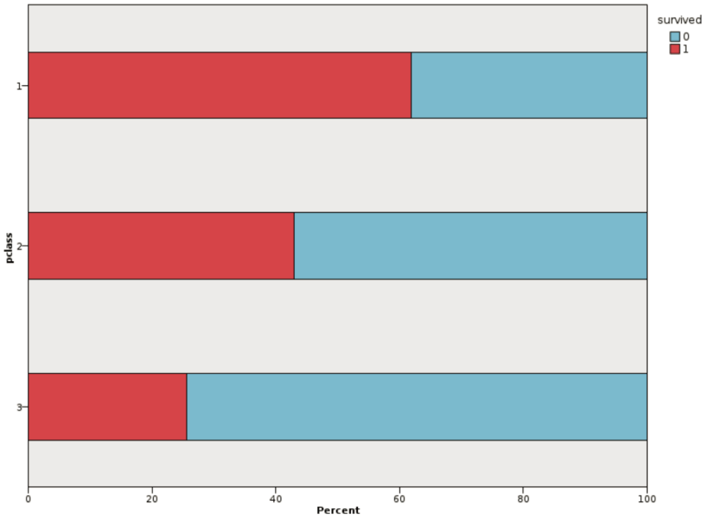


11. The Distribution of pclass output will appear under the **Outputs** tab.

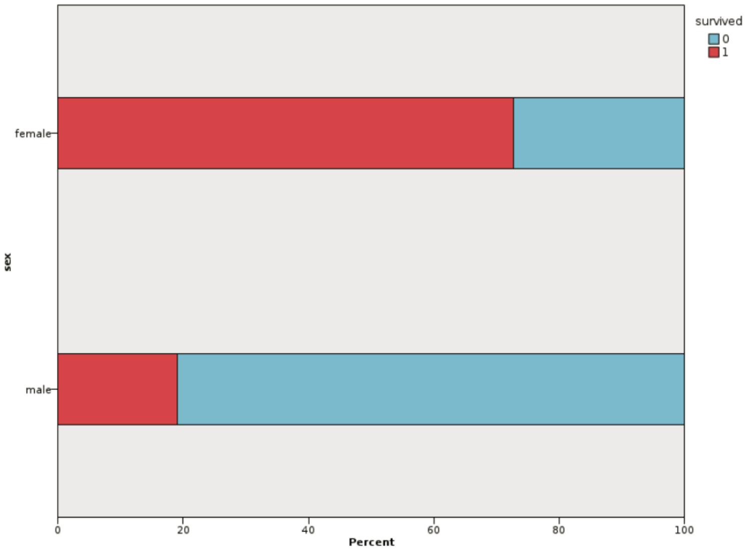


pclass

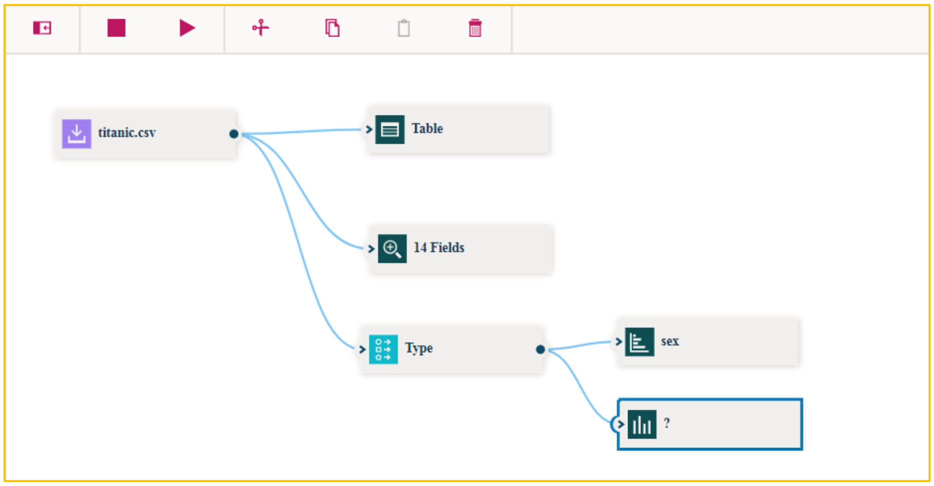
12. Double click on the **Distribution of pclass** to view the graph. We can see from the graph that the likelihood of surviving is correlated to the passenger class. The first class passengers have the highest rate of survivability. **Note if you see a graph with green bars, instead of the one below, redo Step 10 (this is a defect that has been reported).**



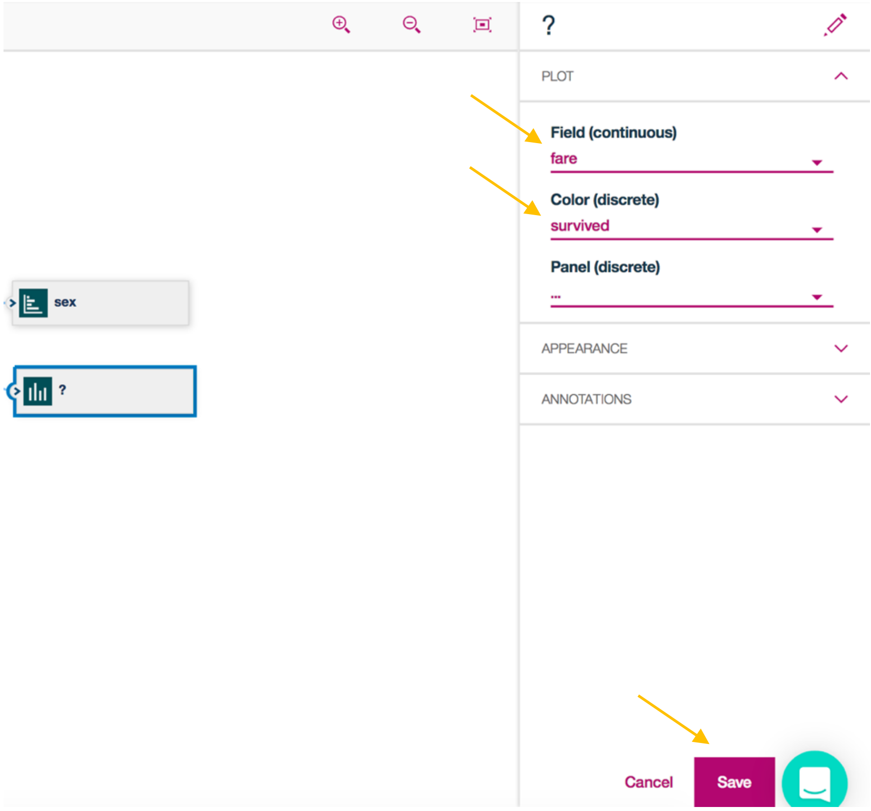
13. You can change the distribution graph to show the survivability by gender by double clicking on the Distribution node and replacing **pclass** with **sex** and clicking Save. Re-run the graph by right clicking on the Distribution node and selecting Run. Double click on the **Distribution of sex**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



14. 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.



15. Double click on the **Histogram** node. Click on the **Plot** dropdown. Select **fare** from the Field (continuous) dropdown. Select **survived** from the Color (discrete) dropdown. Click on **Save**.

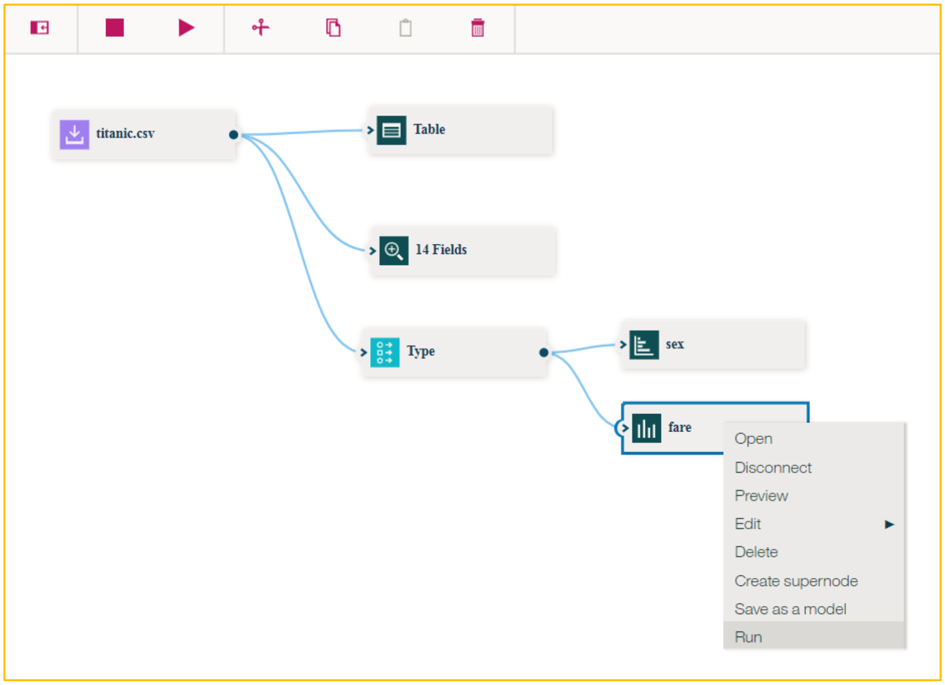


Field (continuous)
fare

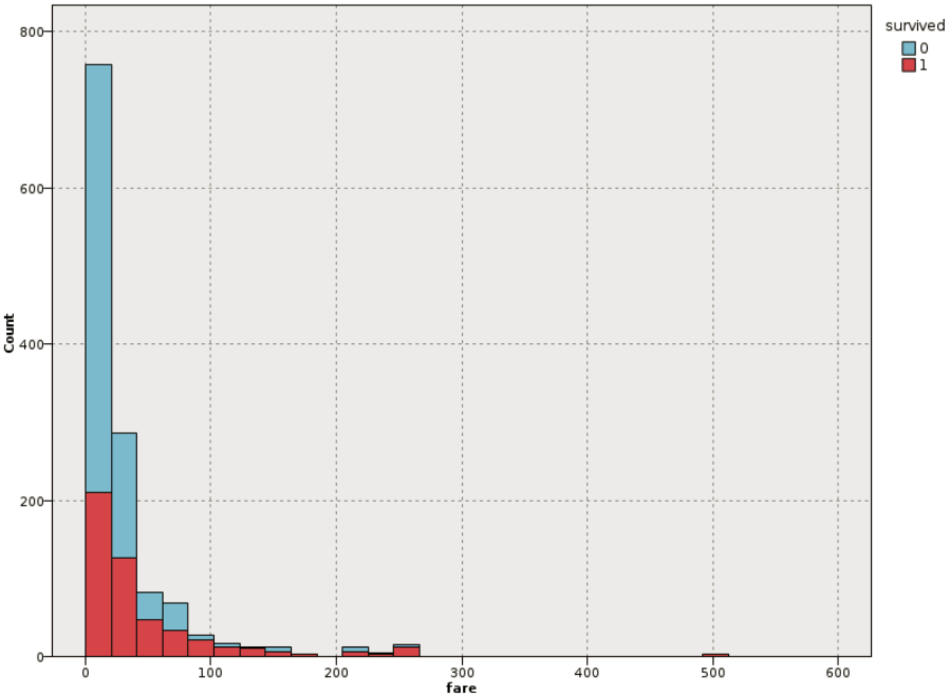
Color (discrete)
survived

Panel (discrete)
...

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



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



18. 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 2.4 Prepare the Data for Modeling

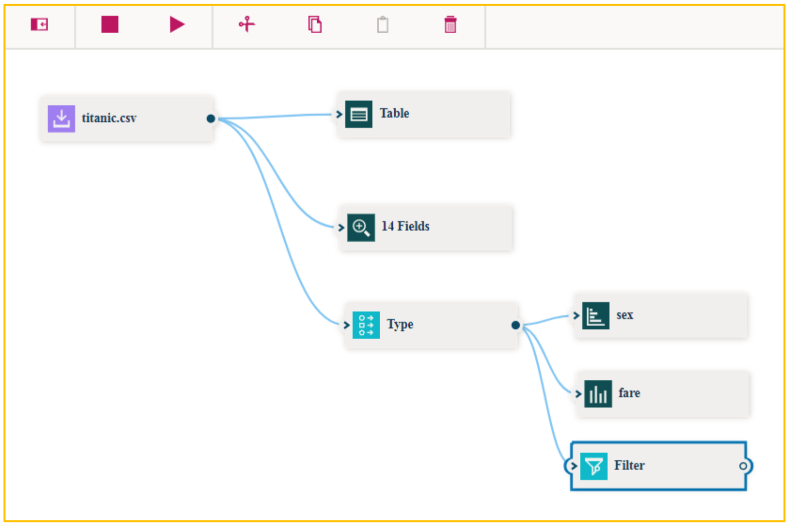
Based on our exploration of the data, there are several transformations that are needed to prepare the data for modeling. This section will introduce, the **Filter** node, the **Select** node, and the **Derive** node that will do the necessary transformations. The **Filter** and **Derive** nodes act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

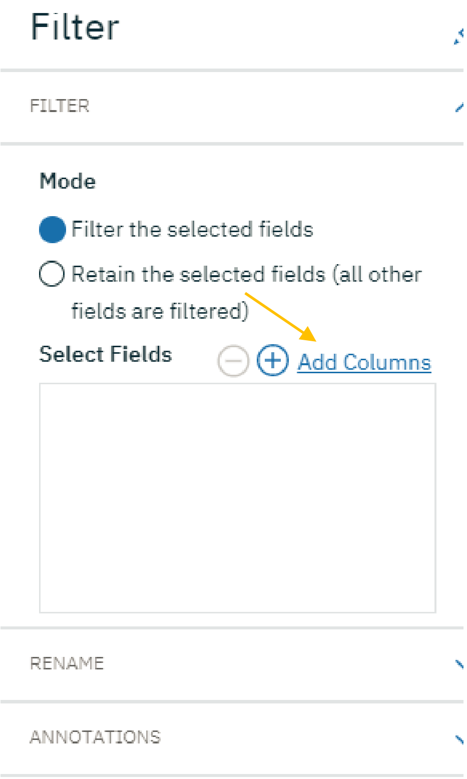
Derive node – The **Derive** node modifies data values or creates new fields from one or more existing fields. We will use the derive node to do a logarithmic transformation of the fare field. We will also use this node to bin the age and fare fields.

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. Connect the **Type** node to the **Filter** node. The canvas should appear as below.



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



Filter

FILTER

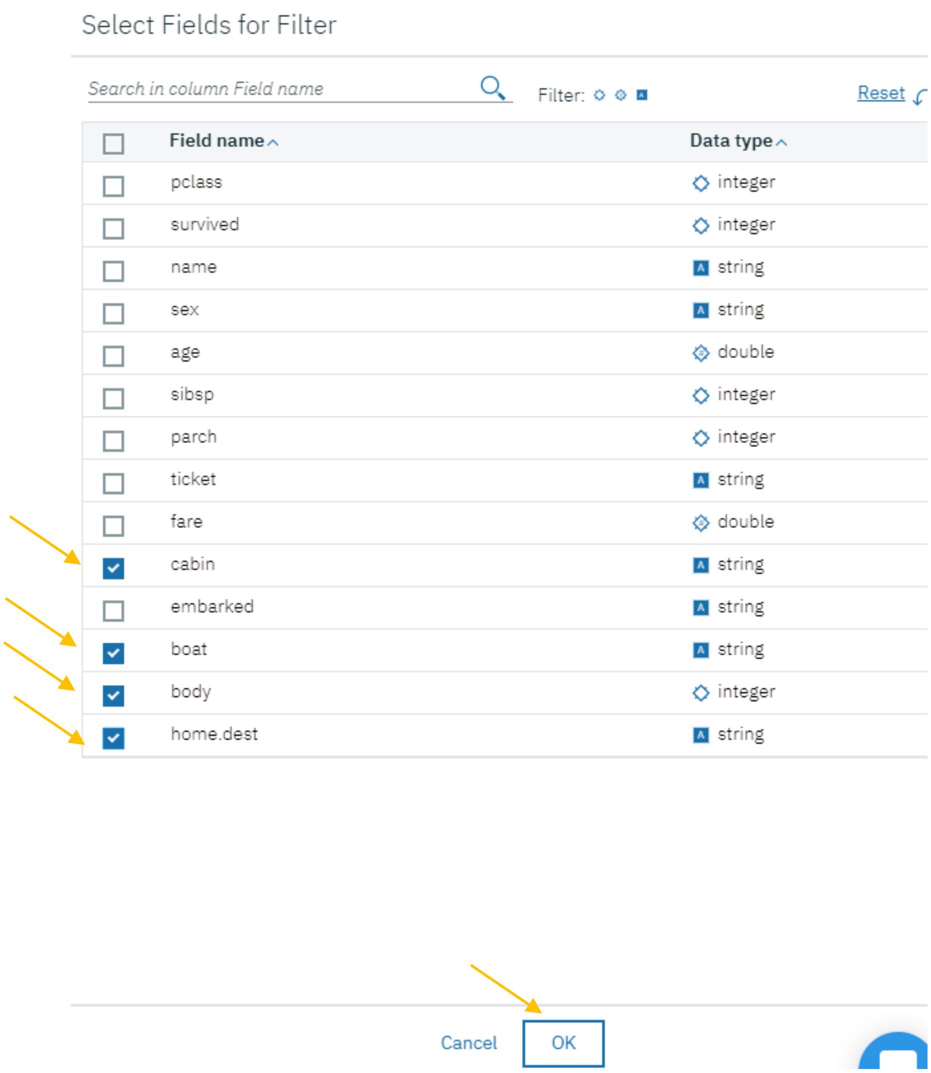
Mode

● Filter the selected fields

○ Retain the selected fields (all other fields are filtered)

Select Fields

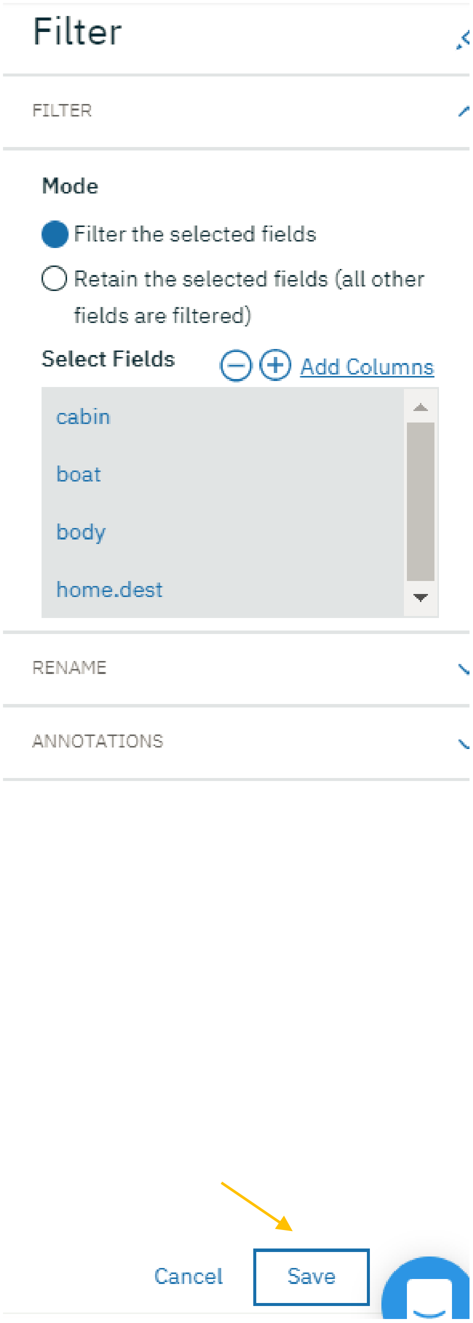
3. Click on the checkboxes adjacent to the **cabin**, **boat**, **body**, and **home.dest** fields, and then click on **OK**.



Select Fields for Filter

☐	Field name	Data type
☐	pclass	◇ integer
☐	survived	◇ integer
☐	name	A string
☐	sex	A string
☐	age	◇ double
☐	sibsp	◇ integer
☐	parch	◇ integer
☐	ticket	A string
☐	fare	◇ double
☑	cabin	A string
☐	embarked	A string
☑	boat	A string
☑	body	◇ integer
☑	home.dest	A string

4. Click **Save** on the Filter panel.



Filter

FILTER

Mode

● Filter the selected fields

○ Retain the selected fields (all other fields are filtered)

Select Fields ⊖ ⊕ Add Columns

cabin

boat

body

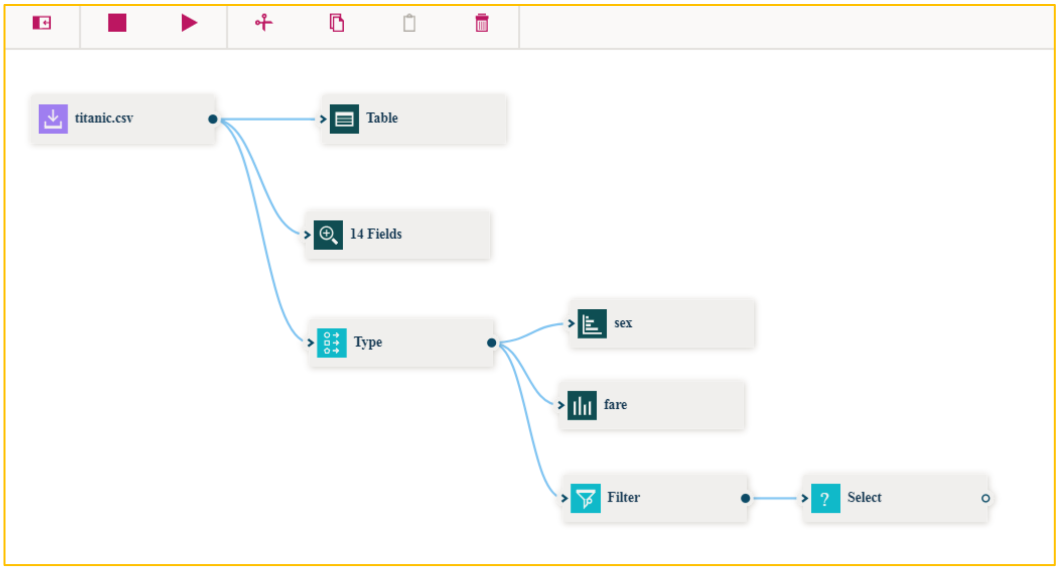
home.dest

RENAME

ANNOTATIONS

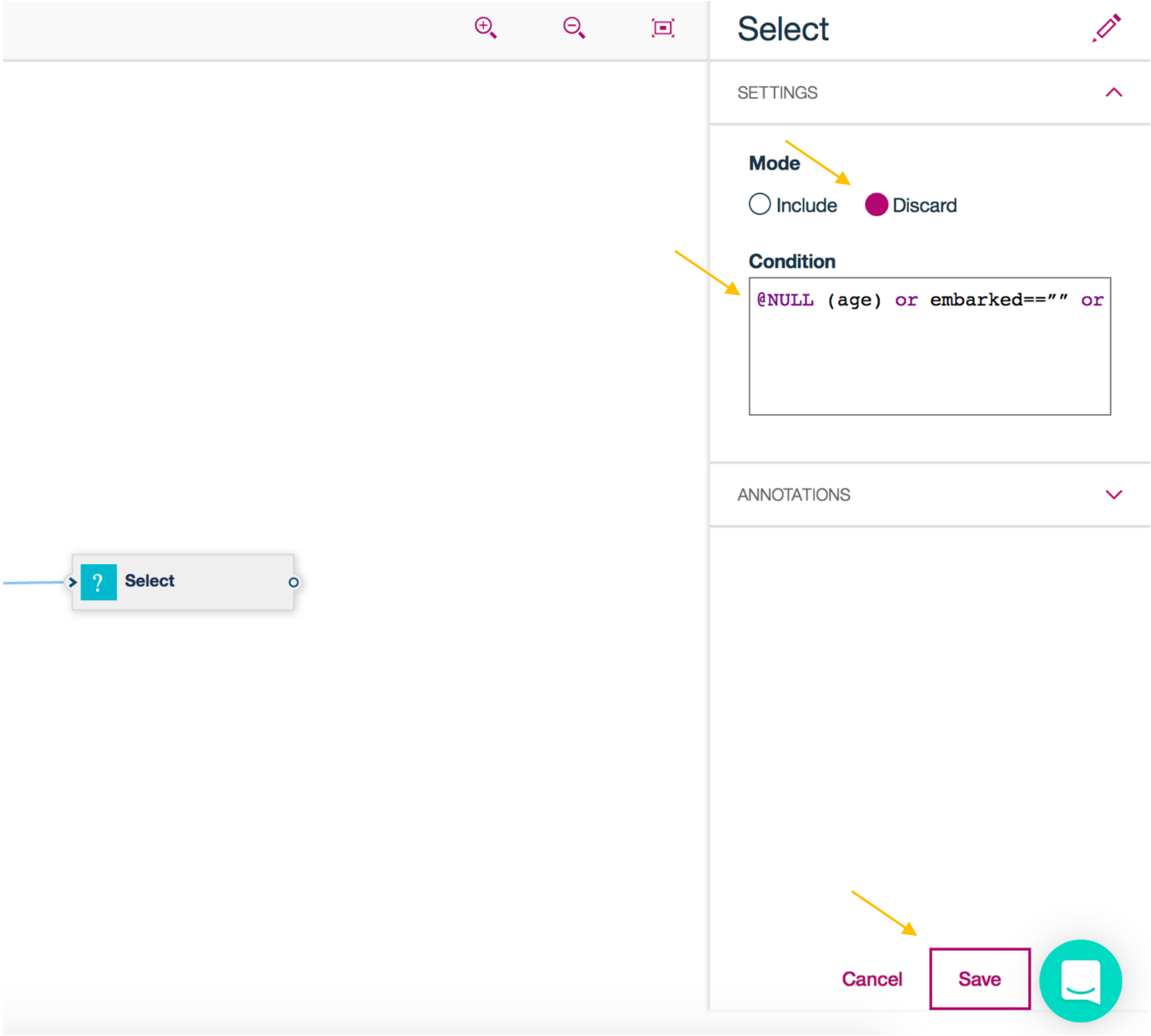
Cancel Save

5. 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. Click on the **Settings** dropdown. 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 **Save**.

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



Select

SETTINGS

Mode

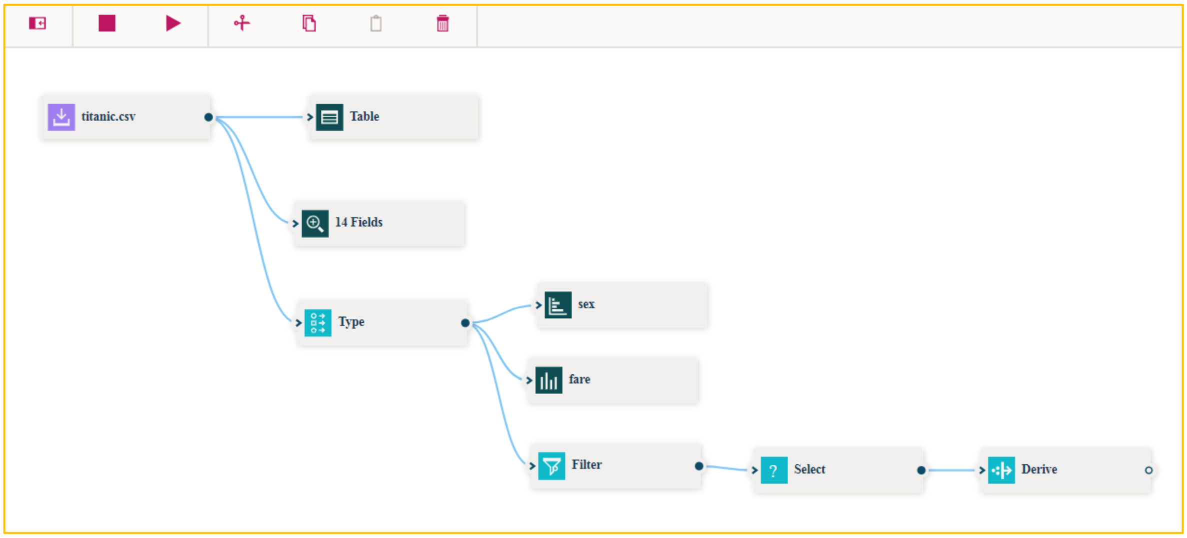
○ Include ● Discard

Condition

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

ANNOTATIONS

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.

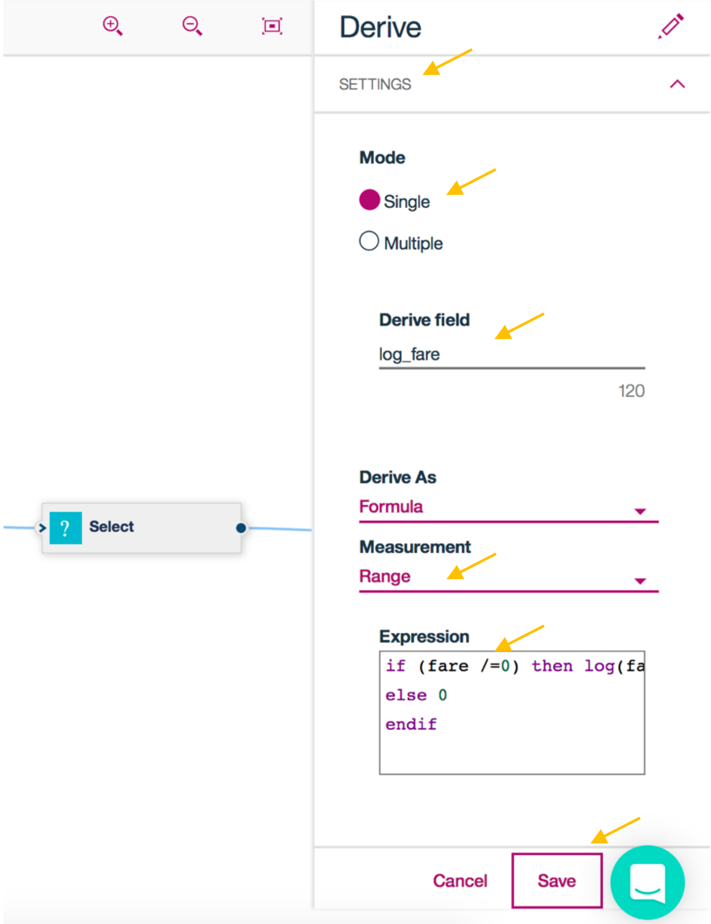

titanic.csv
Table
14 Fields
Type
sex
fare
Filter
Select
Derive

8. Double click on the **Derive** node. Click on the **Settings** Dropdown. Click on the **Single** radio button, enter log_fare for the **Derive** field, select **Range** for the measurement, enter the following code in the **Expression** text box, and click Save.

```
if (fare /=0) then log(fare)

else 0

endif
```



Derive

SETTINGS

Mode

● Single

○ Multiple

Derive field

log_fare

120

Derive As

Formula

Measurement

Range

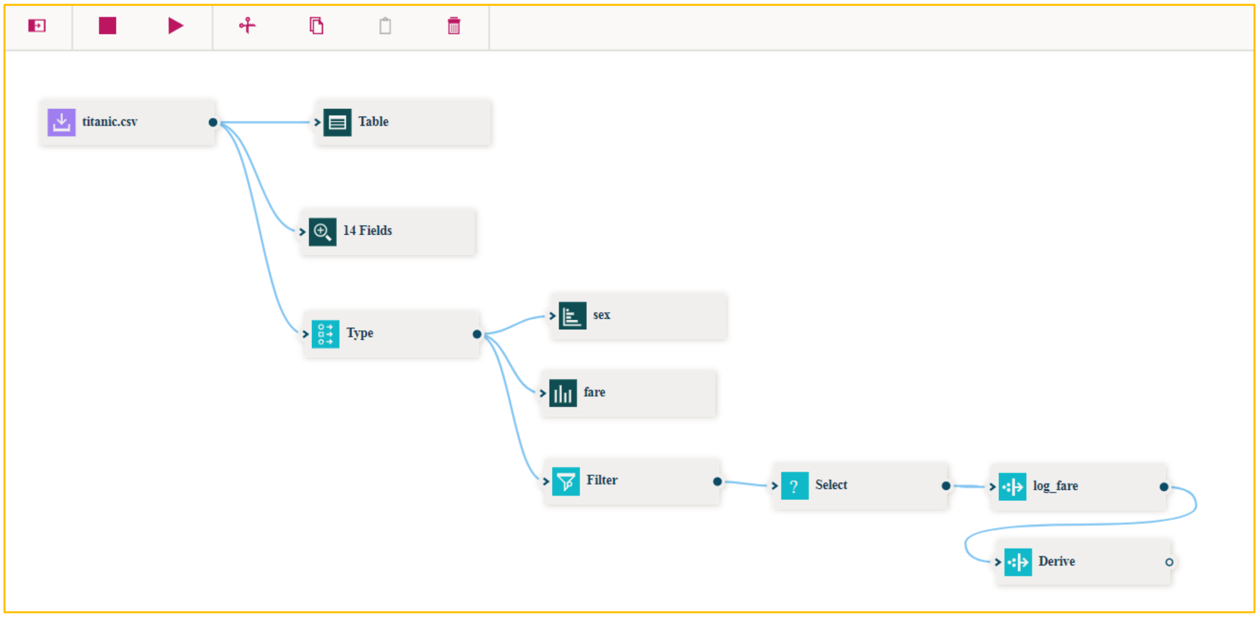
Expression

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

Cancel Save

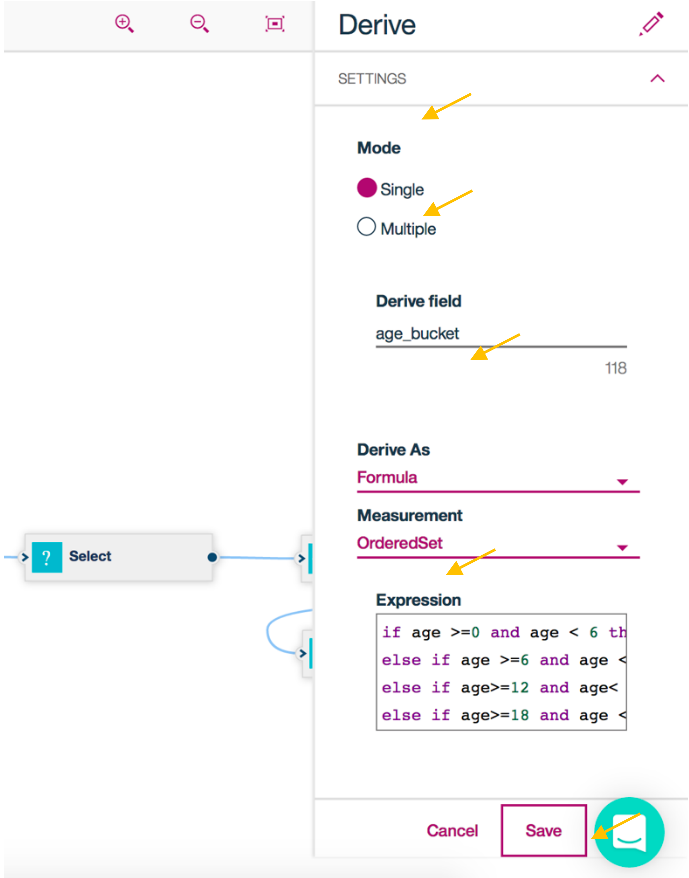
9. Binning of continuous fields is a technique sometimes used in preparing data for modeling. We will bin the age field, and the log_fare field. Add a **Derive** node by clicking on the **Field Operations** menu item in the Node palette and dragging the **Derive** node on the canvas underneath the log_fare **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 **Settings** dropdown. Click on the **Single** radio button, enter age_bucket for the **Derive** field, select OrderedSet for the **Measurement,** enter the following code in the **Expression** text box, and the click **Save**.

```
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

Mode

● Single

○ Multiple

Derive field

age_bucket

118

Derive As

Formula

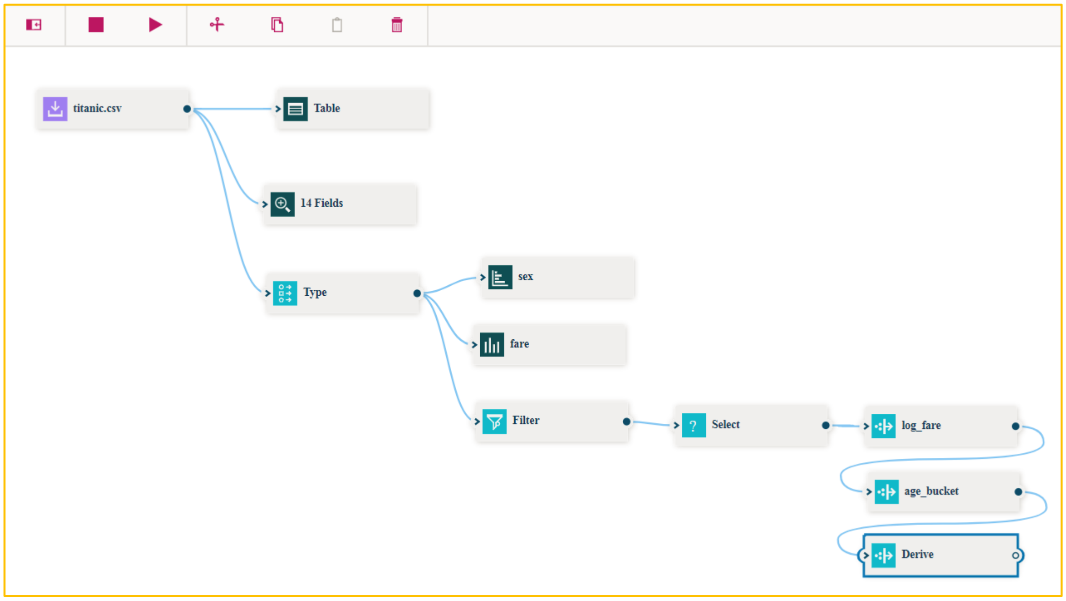
Measurement

OrderedSet

Expression

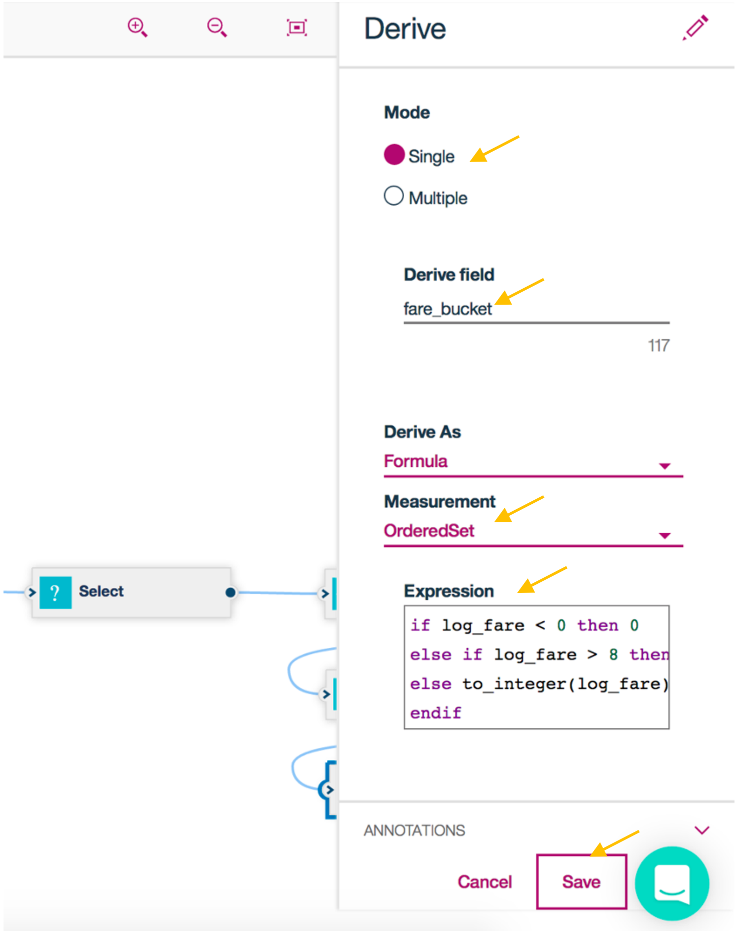
```
if age >=0 and age < 6 th
else if age >=6 and age <
else if age>=12 and age<
else if age>=18 and age <
```

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, click on OrderedSet for the **Measurement**, enter the following code in the **Expression** text box, and click on **Save**.

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



Derive

Mode

● Single

○ Multiple

Derive field

fare_bucket

117

Derive As

Formula

Measurement

OrderedSet

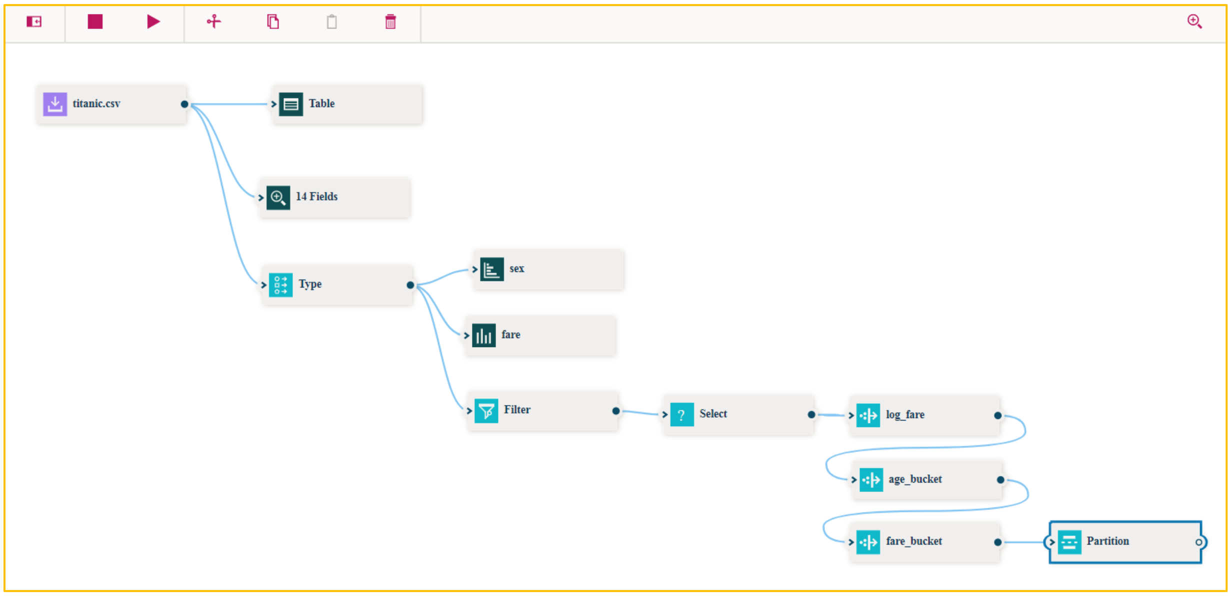
Expression

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

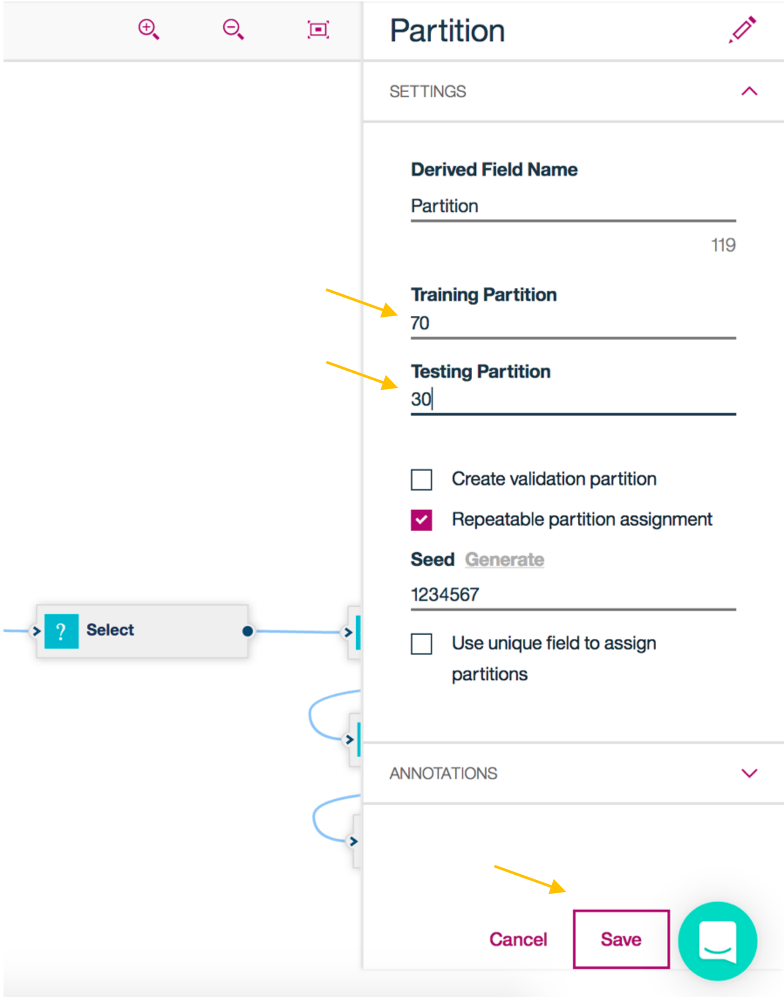
Step 2.5 Modeling and Evaluation

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 **Save**.



Partition

SETTINGS

Derived Field Name

Partition

119

Training Partition

70

Testing Partition

30

□ Create validation partition

☑ Repeatable partition assignment

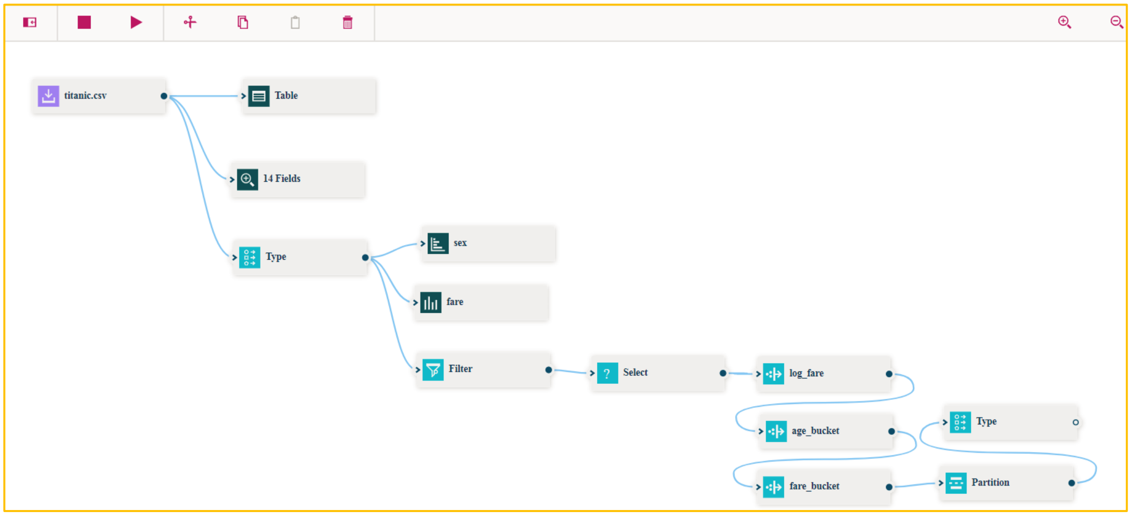
Seed Generate

1234567

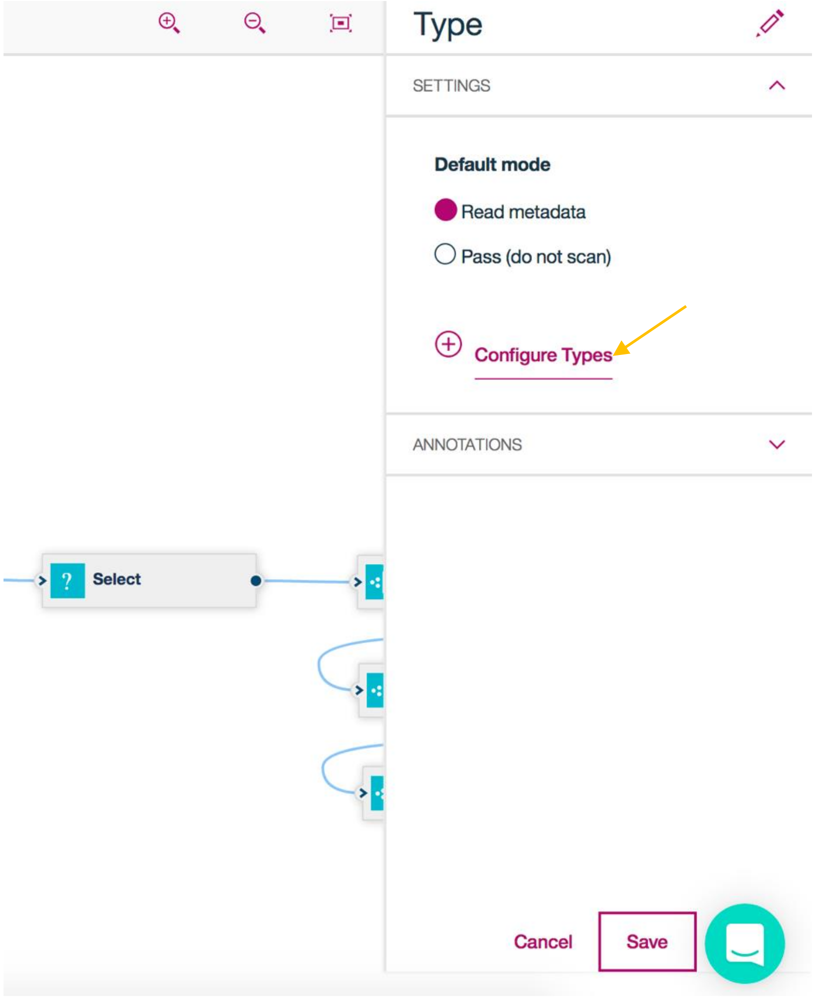
□ Use unique field to assign partitions

ANNOTATIONS

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 **Configure Types**.



Type

SETTINGS

Default mode

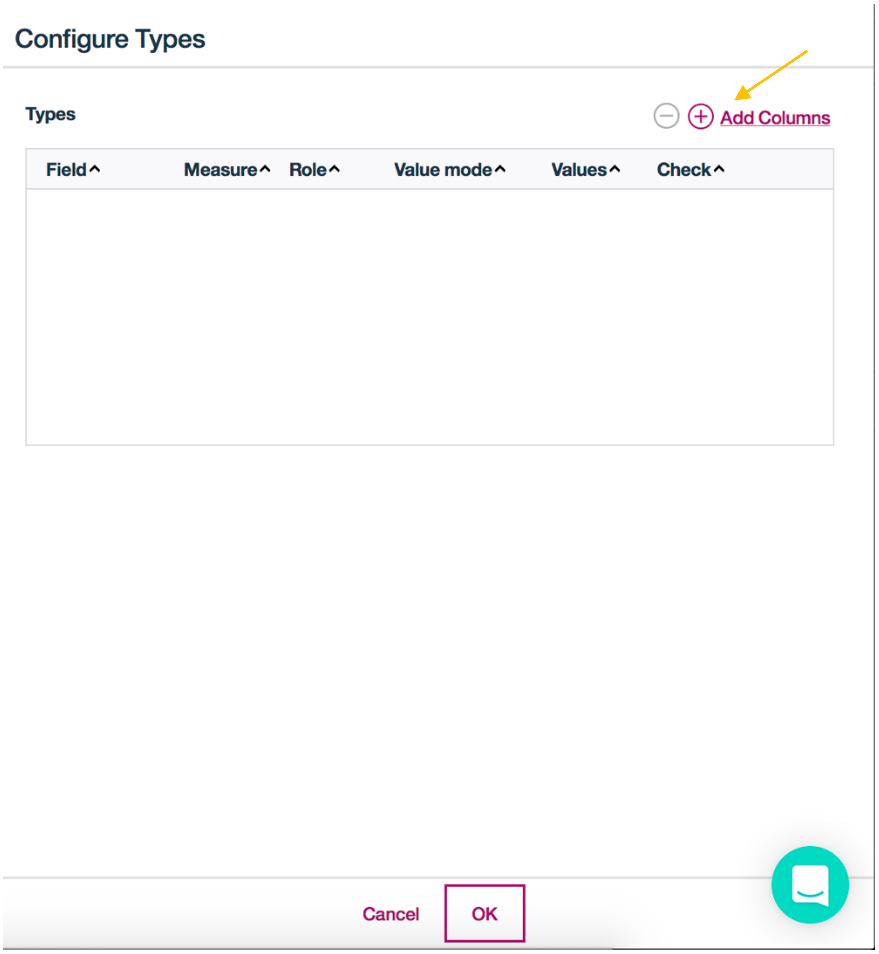
● Read metadata

○ Pass (do not scan)

⊕ Configure Types

ANNOTATIONS

5. Click on **Add Columns.**

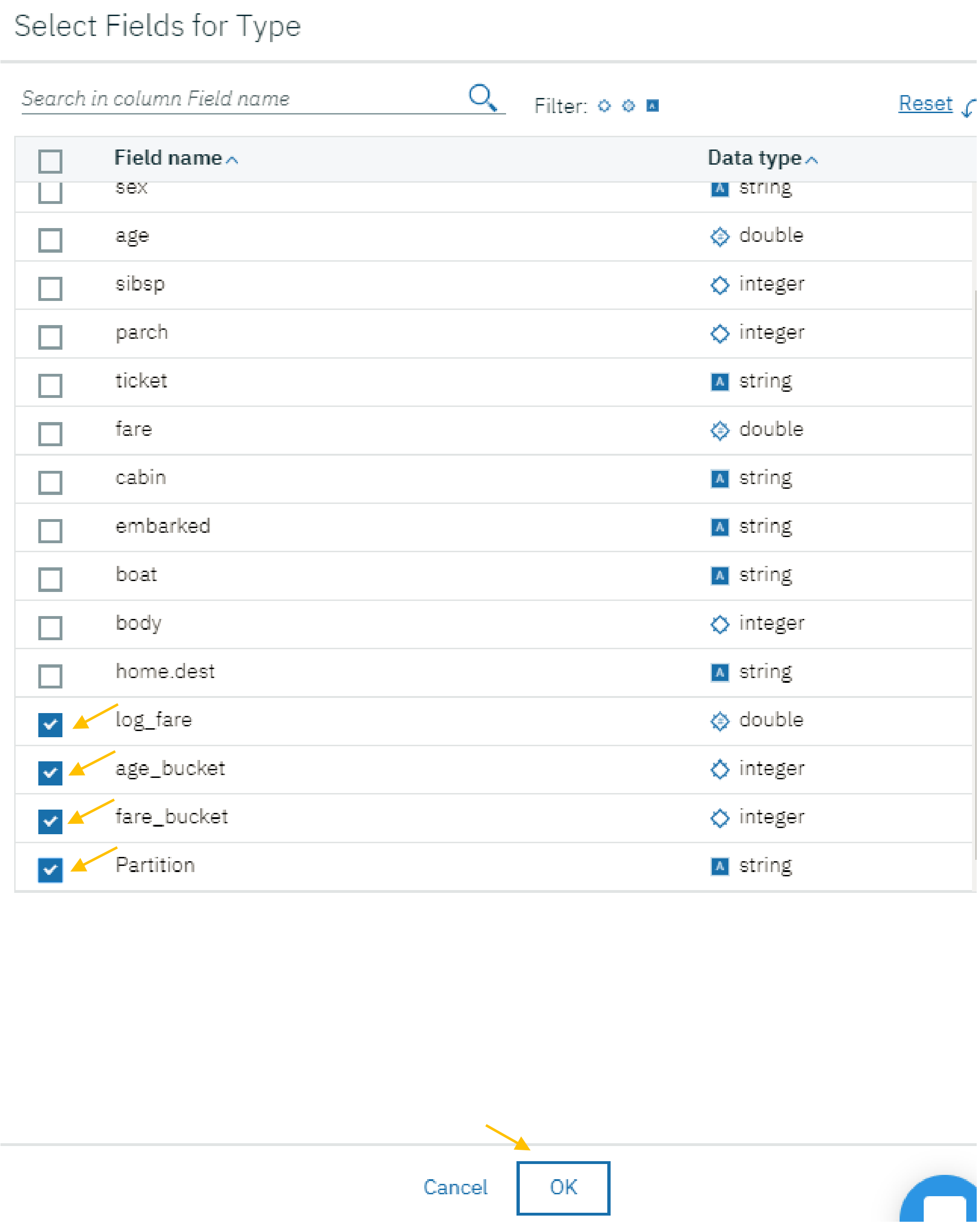


Configure Types

Types

Field ^	Measure ^	Role ^	Value mode ^	Values ^	Check ^

6. Click on checkboxes adjacent to the **log_fare**, **age_bucket**, **fare_bucket**, and **Partition** fields (You may need to scroll down). Click on **OK.**

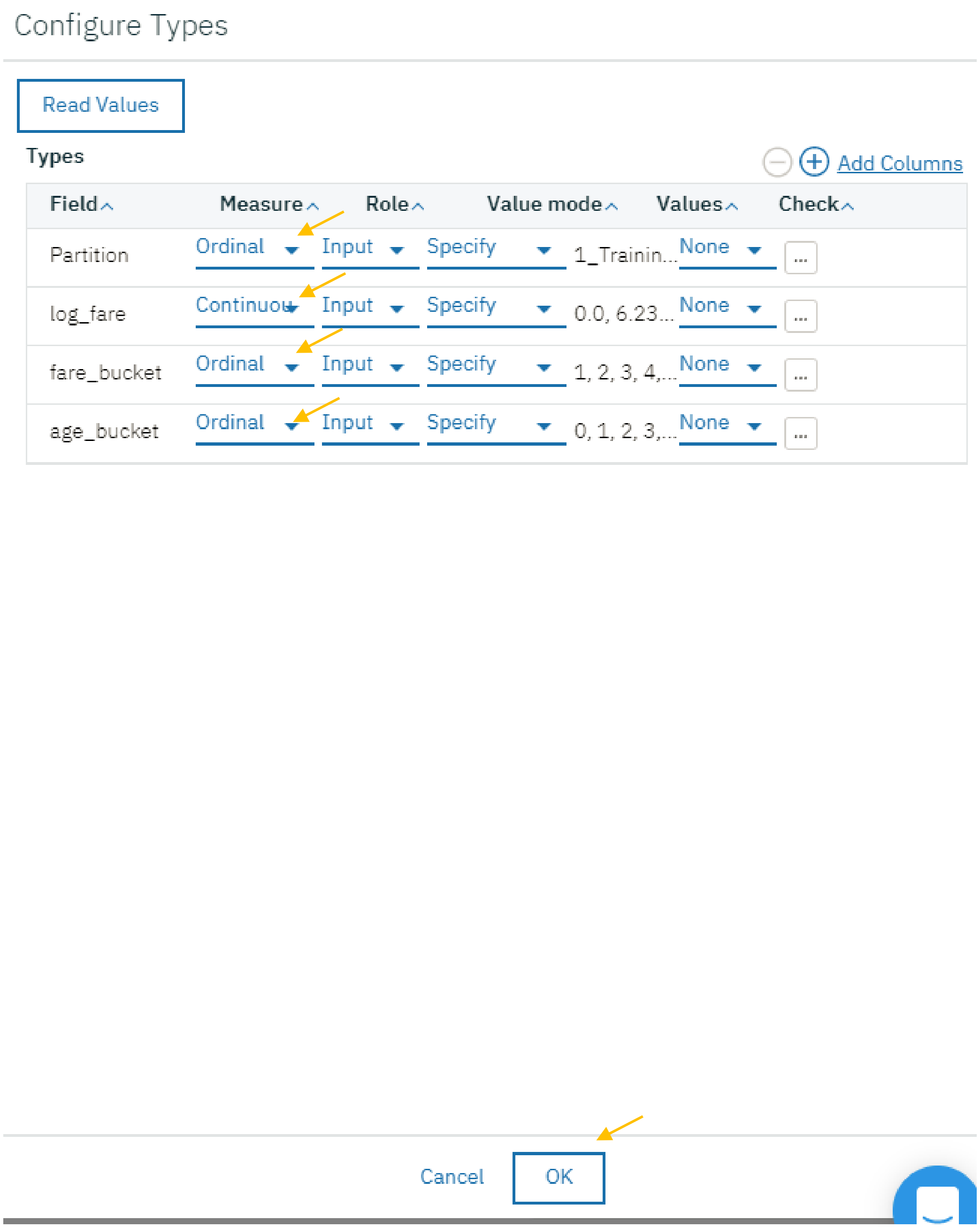


Select Fields for Type

Search in column Field name Filter: ◇ ◇ 🅰 Reset

□	Field name	Data type
□	sex	string
□	age	double
□	sibsp	integer
□	parch	integer
□	ticket	string
□	fare	double
□	cabin	string
□	embarked	string
□	boat	string
□	body	integer
□	home.dest	string
☑	log_fare	double
☑	age_bucket	integer
☑	fare_bucket	integer
☑	Partition	string

7. For the **Partition** field, select **Ordinal** for the **Measurement**. For the log_fare, select **Continuous** for the **Measurement**. For the fare_bucket field, select **Ordinal** for the **Measurement**, and for the age_bucket, select **Ordinal** for the **Measurement**, and click **OK**.



Configure Types

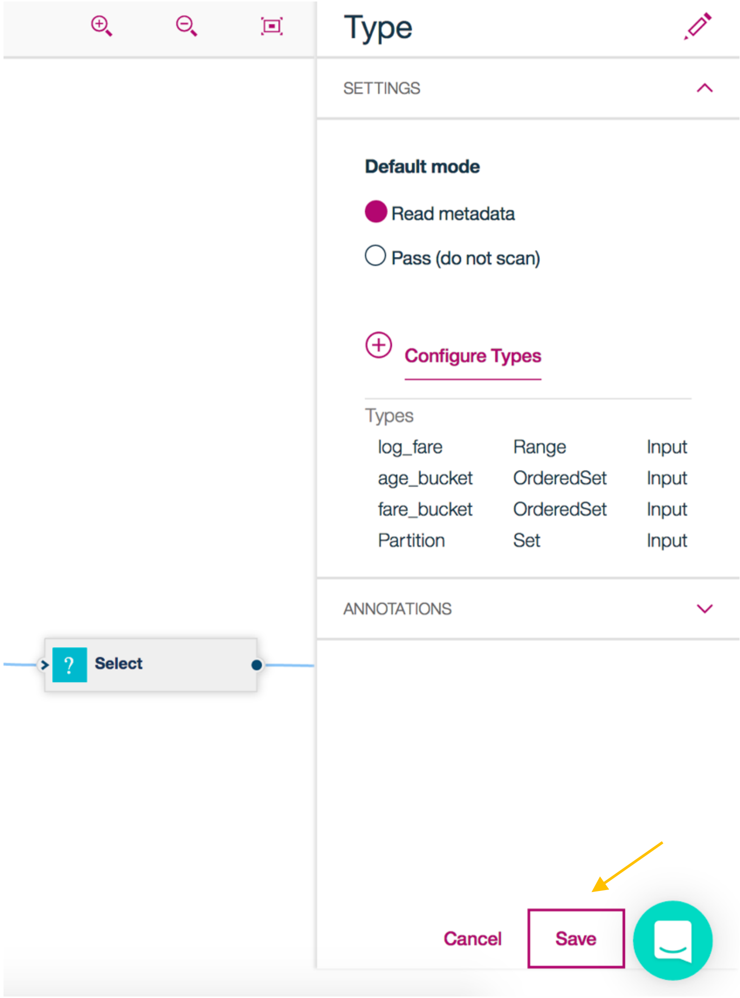
Read Values

Types

Add Columns

Field	Measure	Role	Value mode	Values	Check
Partition	Ordinal	Input	Specify	1_Trainin...	None
log_fare	Continuou	Input	Specify	0.0, 6.23...	None
fare_bucket	Ordinal	Input	Specify	1, 2, 3, 4,...	None
age_bucket	Ordinal	Input	Specify	0, 1, 2, 3,...	None

8. Click on **Save**



Default mode

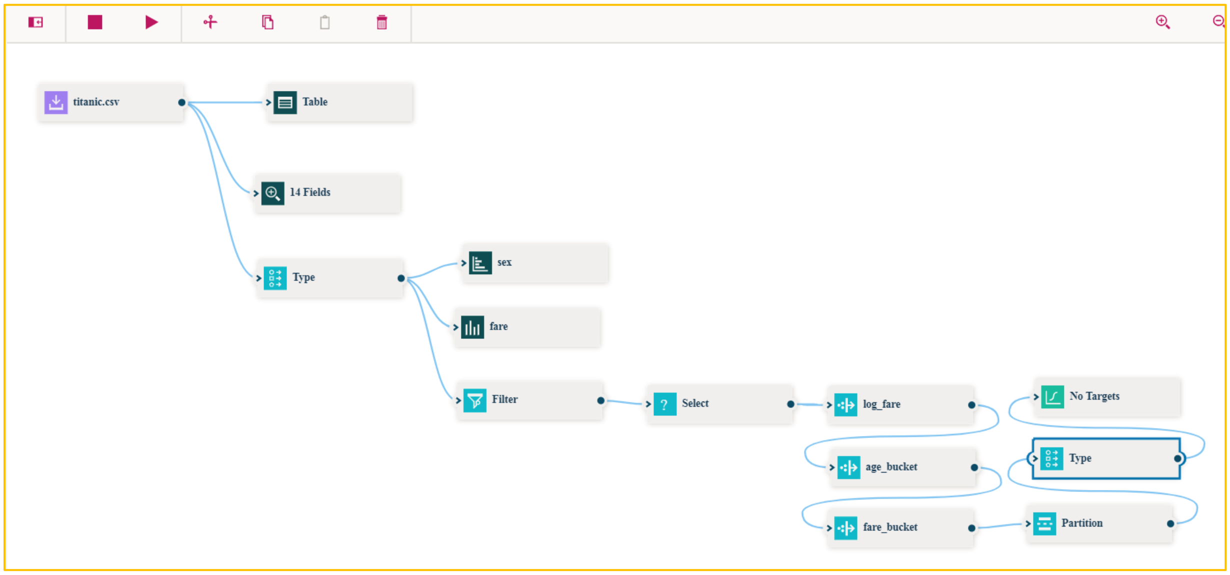
● Read metadata

○ Pass (do not scan)

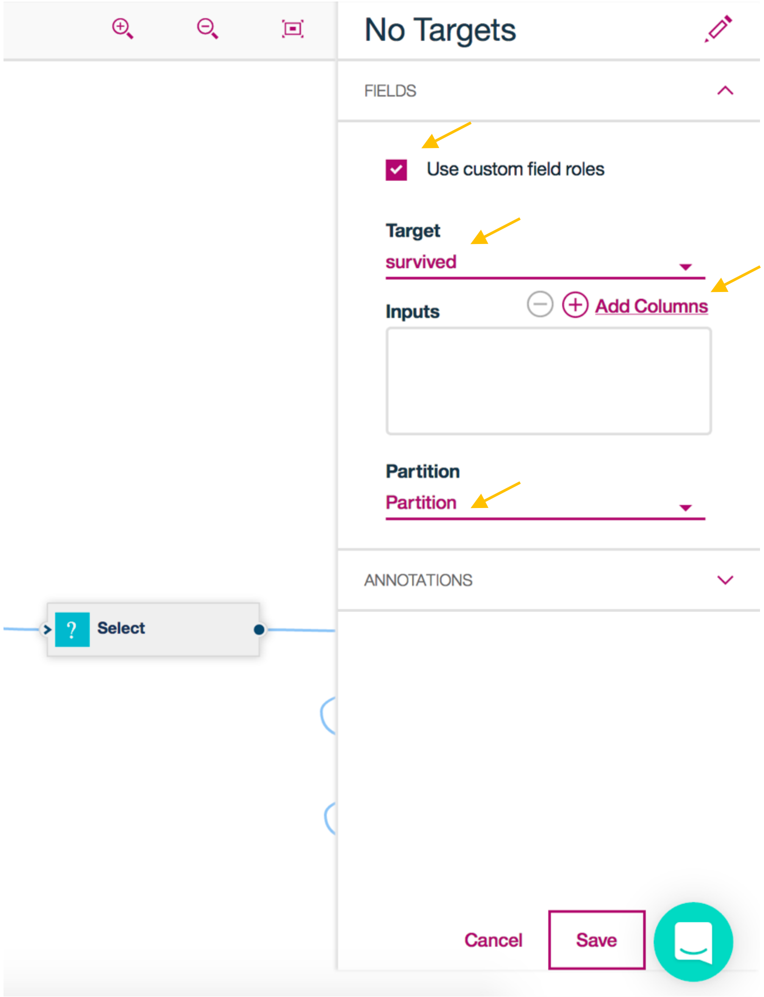
Types

log_fare	Range	Input
age_bucket	OrderedSet	Input
fare_bucket	OrderedSet	Input
Partition	Set	Input

9. 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.



10. 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

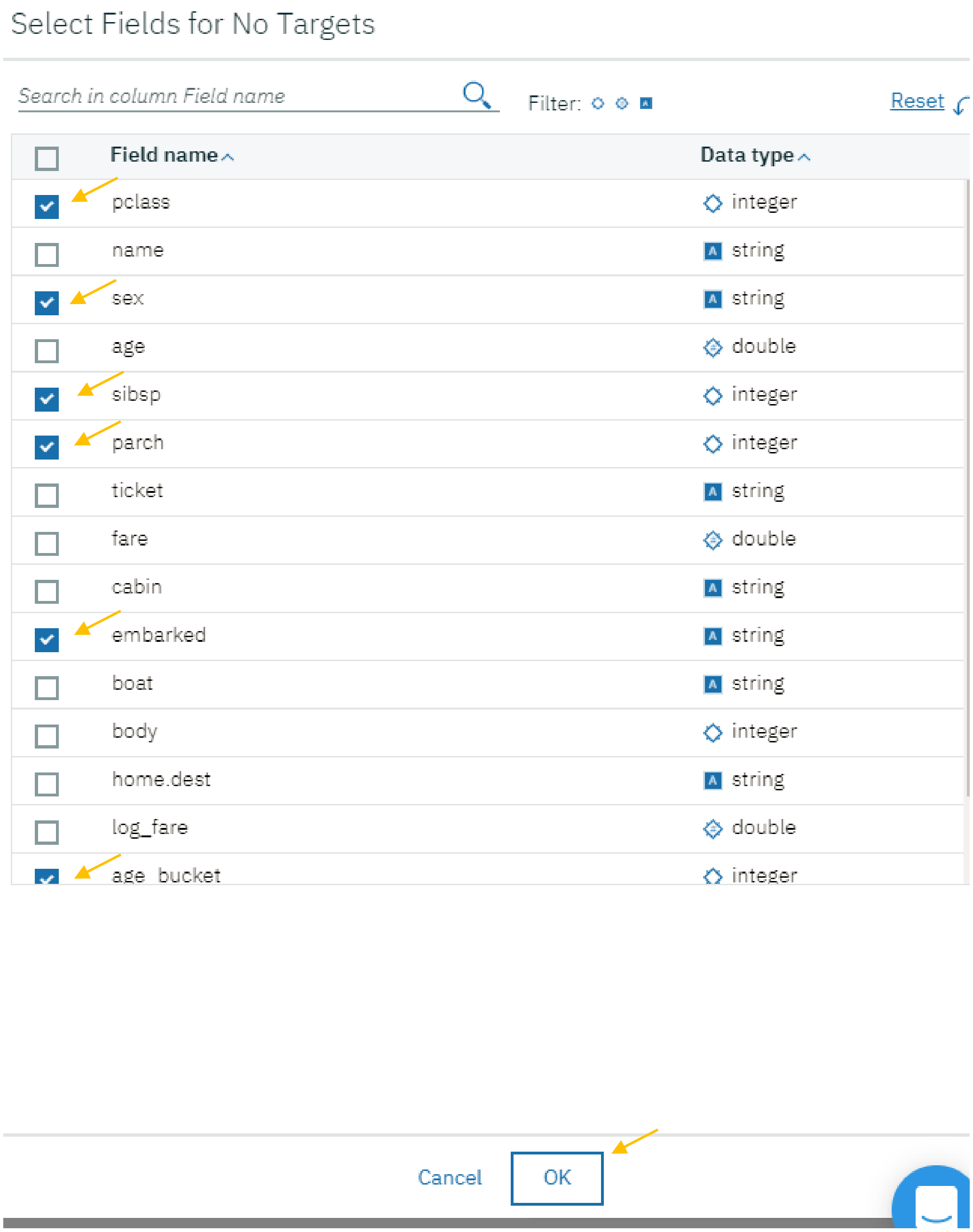
☑ Use custom field roles

Target: survived

Inputs

Partition: Partition

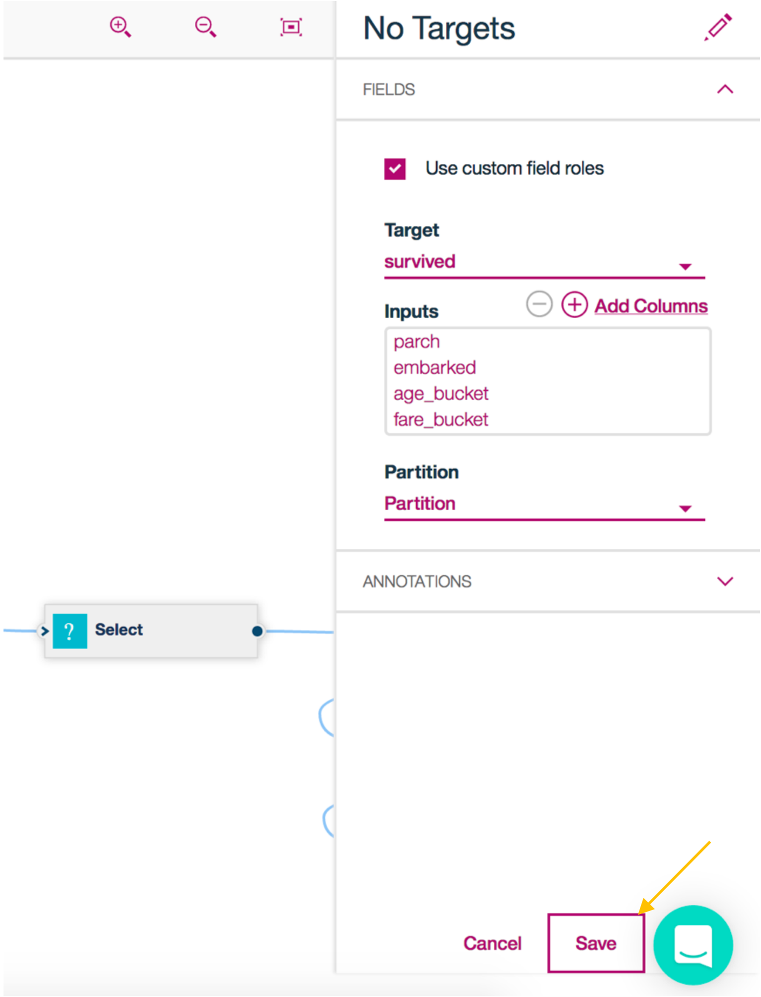
11. Click on the checkboxes next to pclass, sex, sibsp, parch, embarked, age_bucket, fare_bucket fields (you have to scroll down), and then click **OK**.



Select Fields for No Targets

□	Field name	Data type
☑	pclass	integer
□	name	string
☑	sex	string
□	age	double
☑	sibsp	integer
☑	parch	integer
□	ticket	string
□	fare	double
□	cabin	string
☑	embarked	string
□	boat	string
□	body	integer
□	home.dest	string
□	log_fare	double
☑	age_bucket	integer

12. Click **Save**.



No Targets

FIELDS

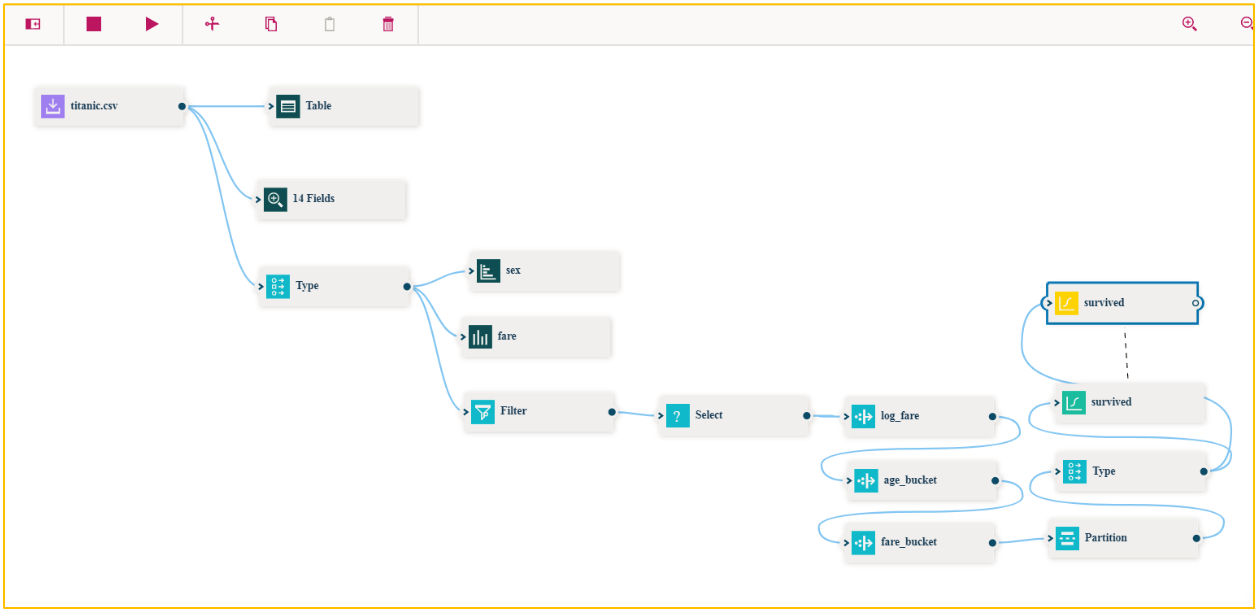
☑ Use custom field roles

Target: survived

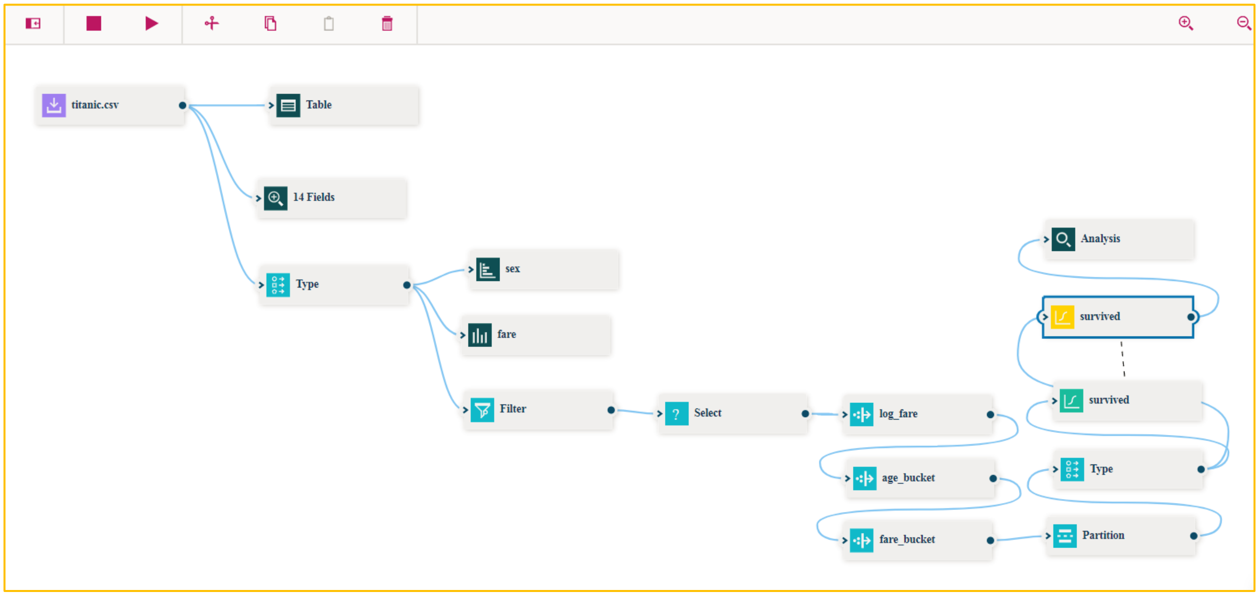
Inputs: parch, embarked, age_bucket, fare_bucket

Partition: Partition

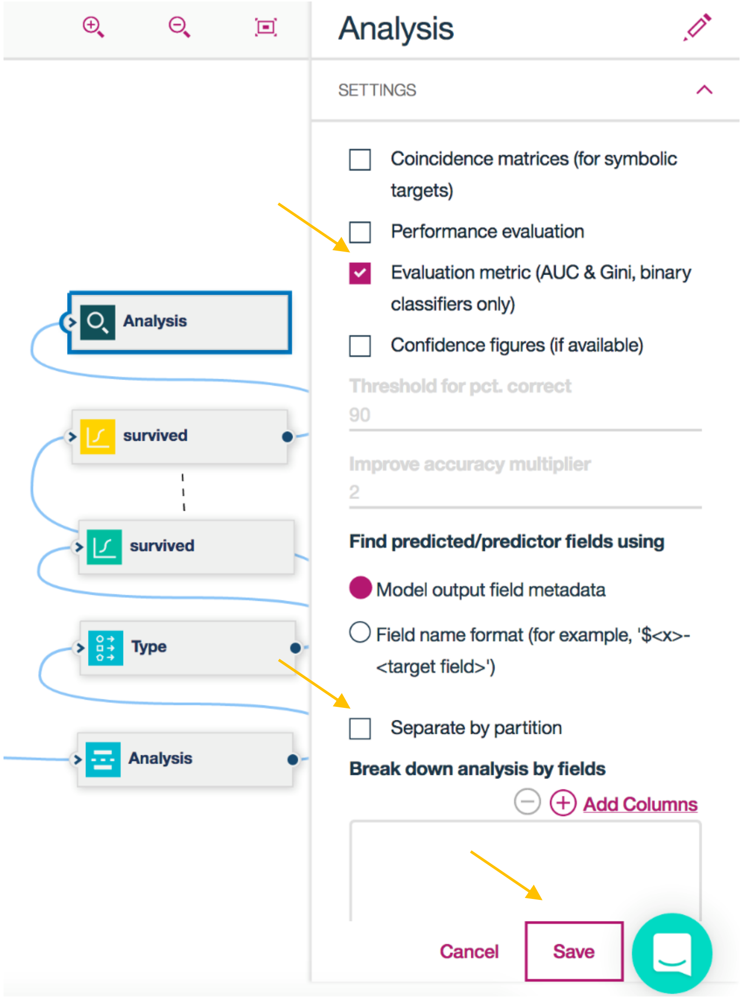
13. Right click on the **Logistic Regression** node and then click **Run**. A **Logistic Regression** "nugget will be created" connected by a dotted line to the **Logistic Regression** node. Drag the nugget and place it above the **Logistic Regression** node. The canvas should appear as below.



14. 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.



15. Double click on the Analysis node. Click on the **Settings** dropdown. Click on the **Evaluation metric** checkbox, uncheck **Separate by partition**, and click on **Save**.



Analysis

SETTINGS

□ Coincidence matrices (for symbolic targets)

□ Performance evaluation

☑ Evaluation metric (AUC & Gini, binary classifiers only)

□ Confidence figures (if available)

Threshold for pct. correct

90

Improve accuracy multiplier

2

Find predicted/predictor fields using

● Model output field metadata

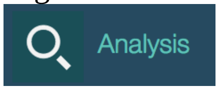
○ Field name format (for example, '$<x>-<target field>')

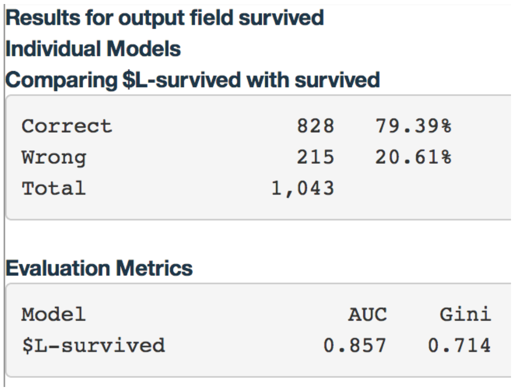
□ Separate by partition

Break down analysis by fields

⊖ ⊕ Add Columns

Cancel Save

16. Right click on the Analysis node, and select Run. After completion, double click on the  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

```
Correct          828    79.39%
Wrong            215    20.61%
Total          1,043
```

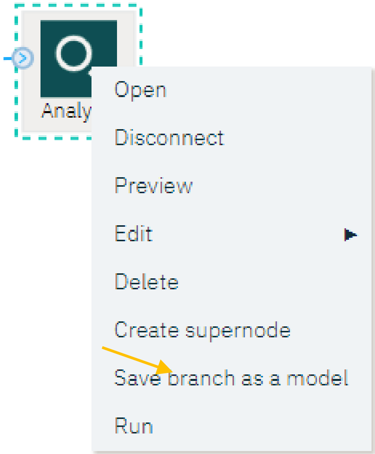
Evaluation Metrics

```
Model                    AUC     Gini
$L-survived            0.857    0.714
```

Step 2.6 Saving a Model

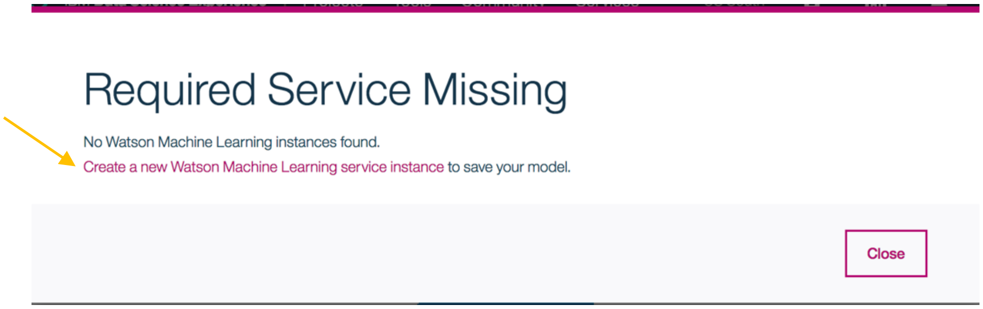
Now that we have created and evaluated a model, we will save the model as an asset. This saved model can be deployed at a future date, removing the need to recreate the same model from scratch.

1. Right click on the Analysis node and then click on **Save branch as a model**.



If you receive a "Required Service Missing" page, proceed to step 1.A. If you do not receive this page, proceed to step 2.

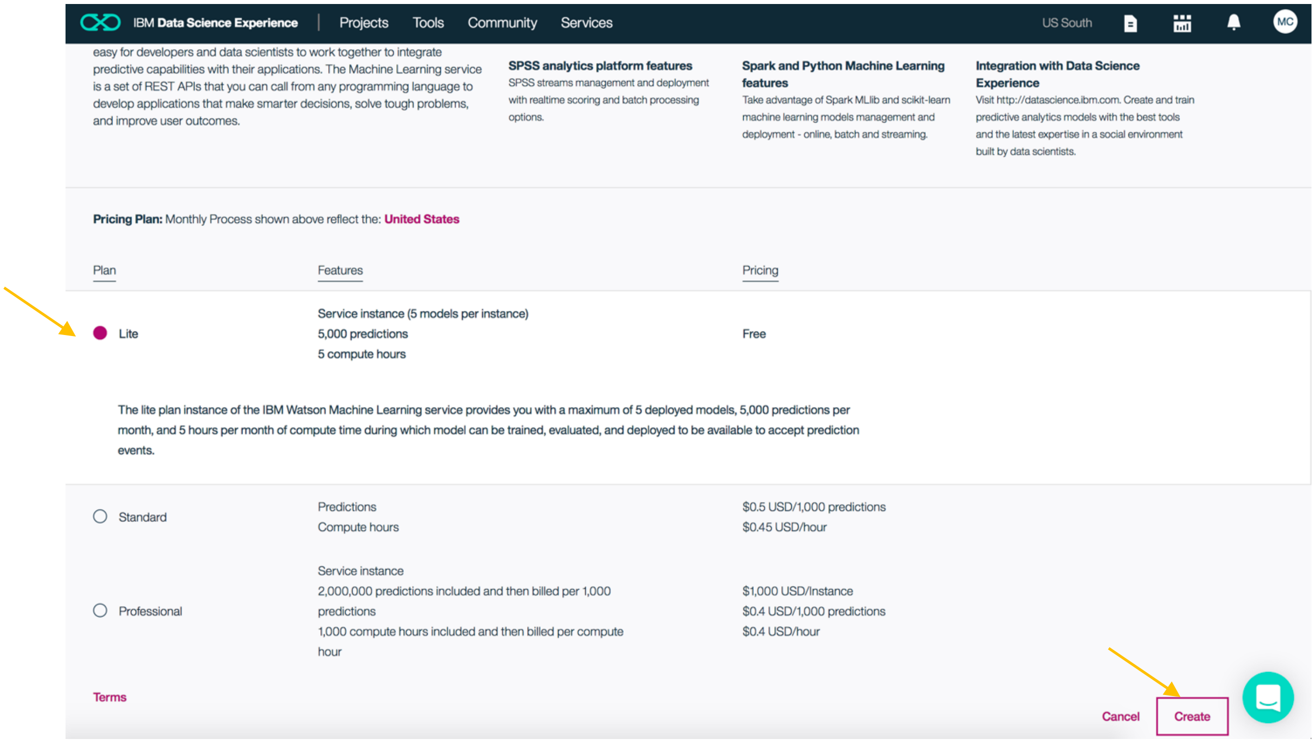
1.A. Click on "**Create a new Watson Machine Learning service instance**."



Required Service Missing

No Watson Machine Learning instances found.
Create a new Watson Machine Learning service instance to save your model.

1.B. Scroll down and click on "**Lite**" to select the Lite plan and then click on **Create**.



easy for developers and data scientists to work together to integrate predictive capabilities with their applications. The Machine Learning service is a set of REST APIs that you can call from any programming language to develop applications that make smarter decisions, solve tough problems, and improve user outcomes.

SPSS analytics platform features
SPSS streams management and deployment with realtime scoring and batch processing options.

Spark and Python Machine Learning features
Take advantage of Spark MLlib and scikit-learn machine learning models management and deployment - online, batch and streaming.

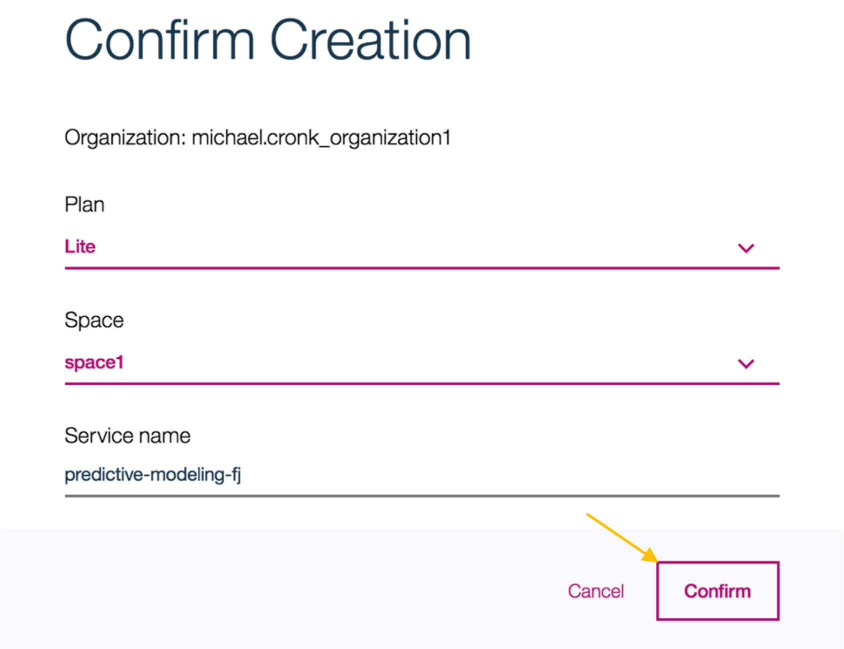
Integration with Data Science Experience
Visit http://datascience.ibm.com. Create and train predictive analytics models with the best tools and the latest expertise in a social environment built by data scientists.

Pricing Plan: Monthly Process shown above reflect the: **United States**

Plan	Features	Pricing
● Lite	Service instance (5 models per instance) 5,000 predictions 5 compute hours	Free
○ Standard	Predictions Compute hours	$0.5 USD/1,000 predictions $0.45 USD/hour
○ Professional	Service instance 2,000,000 predictions included and then billed per 1,000 predictions 1,000 compute hours included and then billed per compute hour	$1,000 USD/Instance $0.4 USD/1,000 predictions $0.4 USD/hour

The lite plan instance of the IBM Watson Machine Learning service provides you with a maximum of 5 deployed models, 5,000 predictions per month, and 5 hours per month of compute time during which model can be trained, evaluated, and deployed to be available to accept prediction events.

1.C. Click on **Confirm**.



Confirm Creation

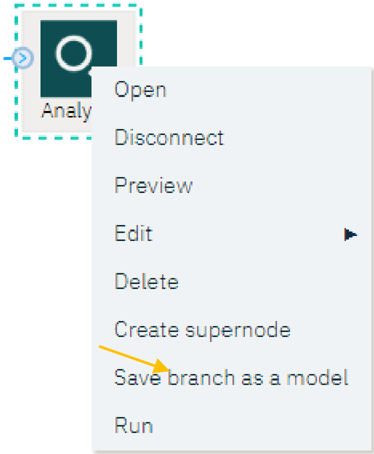
Organization: michael.cronk_organization1

Plan
Lite

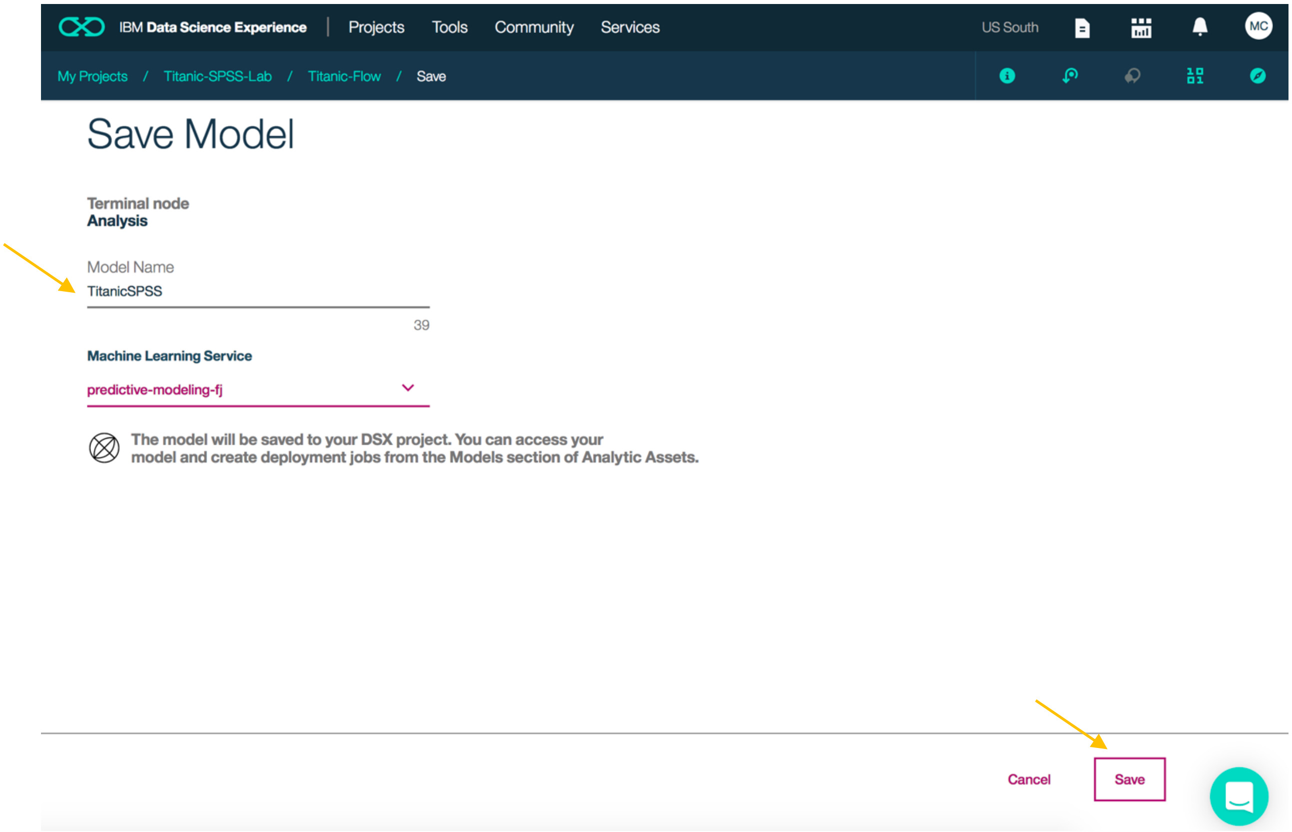
Space
space1

Service name
predictive-modeling-fj

1.D. Return to your SPSS flow. Right click on the Analysis node and then click on **Save as a model**.



2. Type in "**TitanicSPSS**" as the Model Name and click **Save**.


IBM Data Science Experience
Projects
Tools
Community
Services
US South
MC
My Projects / Titanic-SPSS-Lab / Titanic-Flow / Save
Cancel
Save

Save Model

Terminal node
Analysis

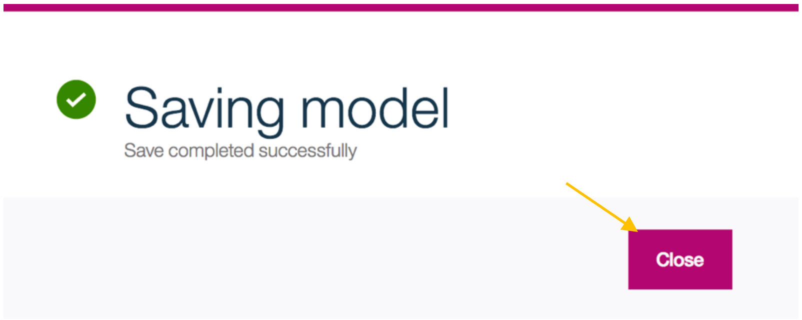
Model Name
TitanicSPSS

39

Machine Learning Service
predictive-modeling-fj

The model will be saved to your DSX project. You can access your model and create deployment jobs from the Models section of Analytic Assets.

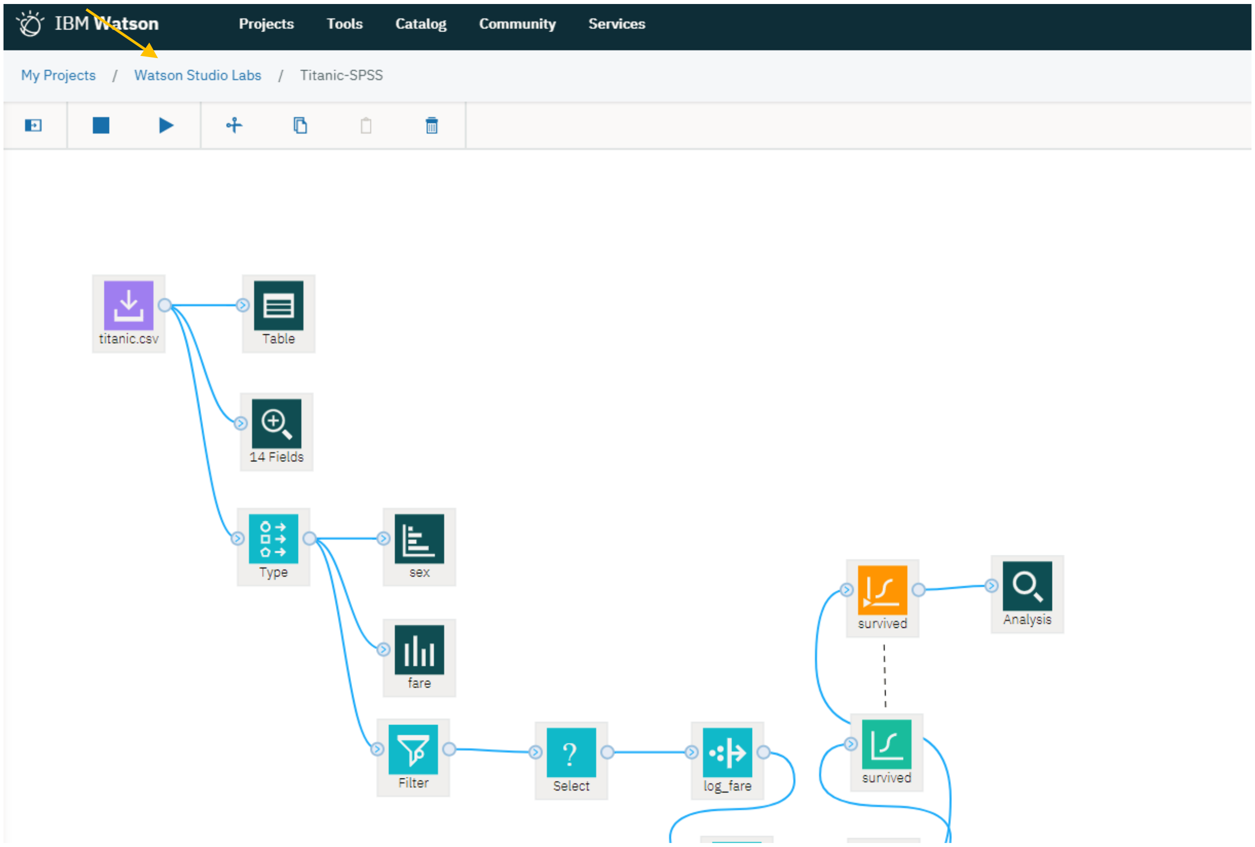
3. Click **Close**.



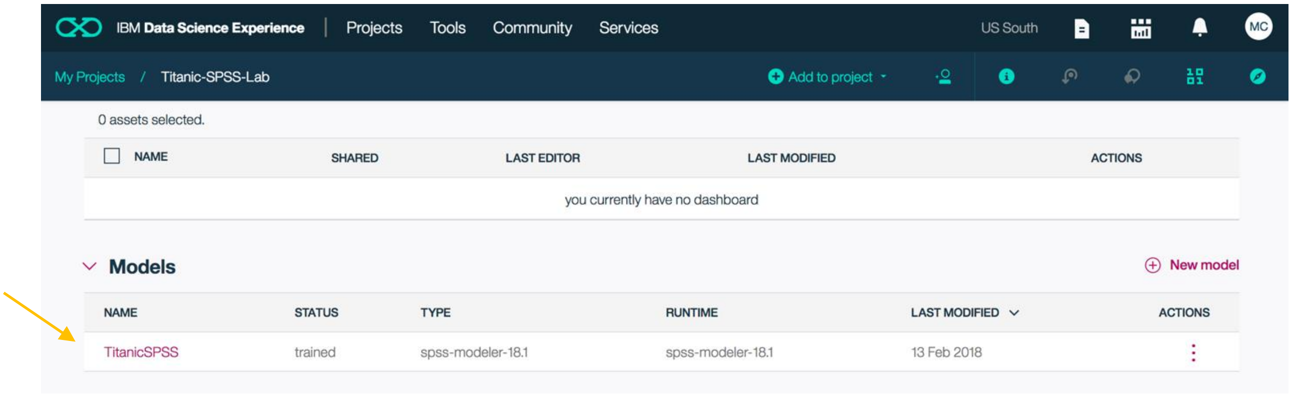
Saving model

Save completed successfully

4. Navigate to your project “assets” page. In this example, click on **Watson Studio Labs**.



5. Note that the model you built is now saved as an asset and the work you have completed can be easily reused in the future.



0 assets selected.

□ NAME	SHARED	LAST EDITOR	LAST MODIFIED	ACTIONS
you currently have no dashboard				

Models

NAME	STATUS	TYPE	RUNTIME	LAST MODIFIED	ACTIONS
TitanicSPSS	trained	spss-modeler-18.1	spss-modeler-18.1	13 Feb 2018	⋮