

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 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**.

The screenshot shows the IBM Data Science Experience dashboard. The top navigation bar has tabs for 'Projects' (which is highlighted with a teal underline), 'Tools', 'Data Services', and 'Community'. Below the navigation is a sidebar titled 'My Projects' with a search bar labeled 'Find project by name'. The main content area displays a table with one row, labeled 'NAME' under the column header. To the right of the table is a sidebar with sections for 'Recent Items' (containing 'Titanic') and 'Default Project'. A yellow arrow points from the text 'View All Projects' to the 'View All Projects' button in the sidebar.

2. Click on **New project**.

The screenshot shows the same IBM Data Science Experience interface as the previous one, but with a different view of the 'My Projects' section. It includes a dropdown menu above the search bar labeled 'All projects'. On the far right of the interface, there is a toolbar with various icons. One icon, specifically the 'New project' icon (a circular button with a plus sign), is highlighted with a yellow arrow.

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

New project

Define project details

Name
Titanic-Lab

Description
Project description

Choose project options

Restrict who can be a collaborator

Define storage

Select storage service
Target Cloud Object Storage Instance
cloud-object-storage-zz

Select Spark service
Spark-rj

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

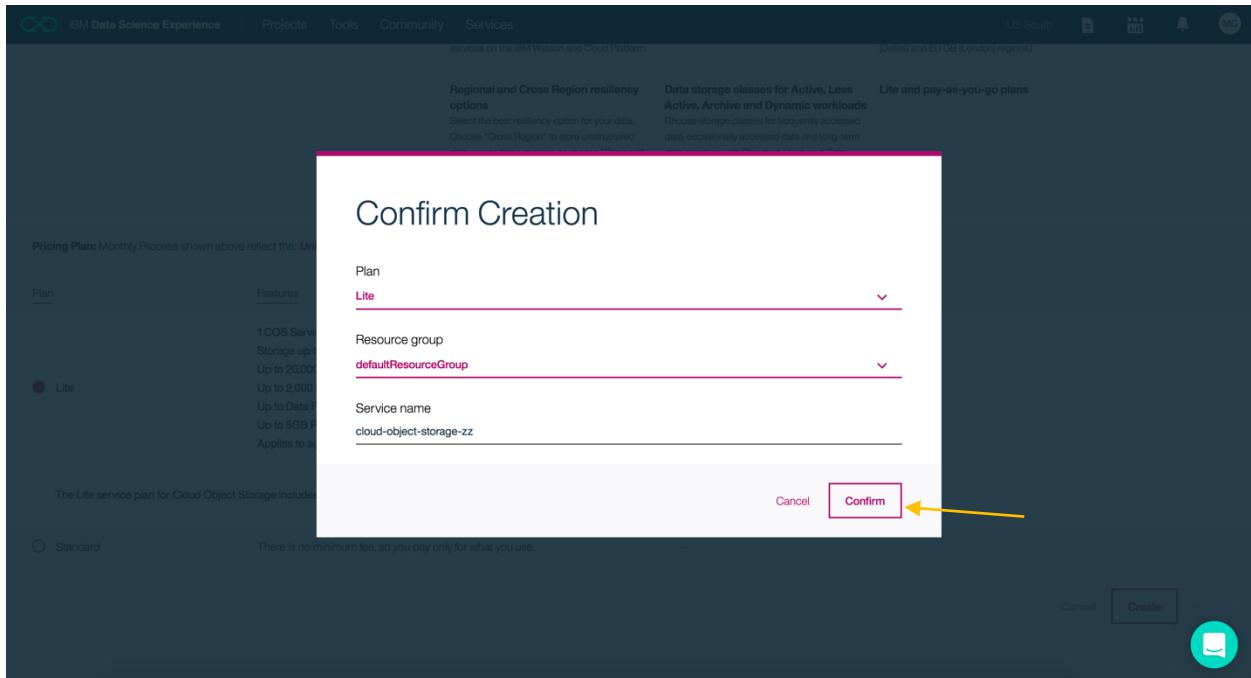
Cancel Create

If you were able to create a project, skip to step 4. If you were unable to select a Spark Service or a Storage type, move on to 3.A.

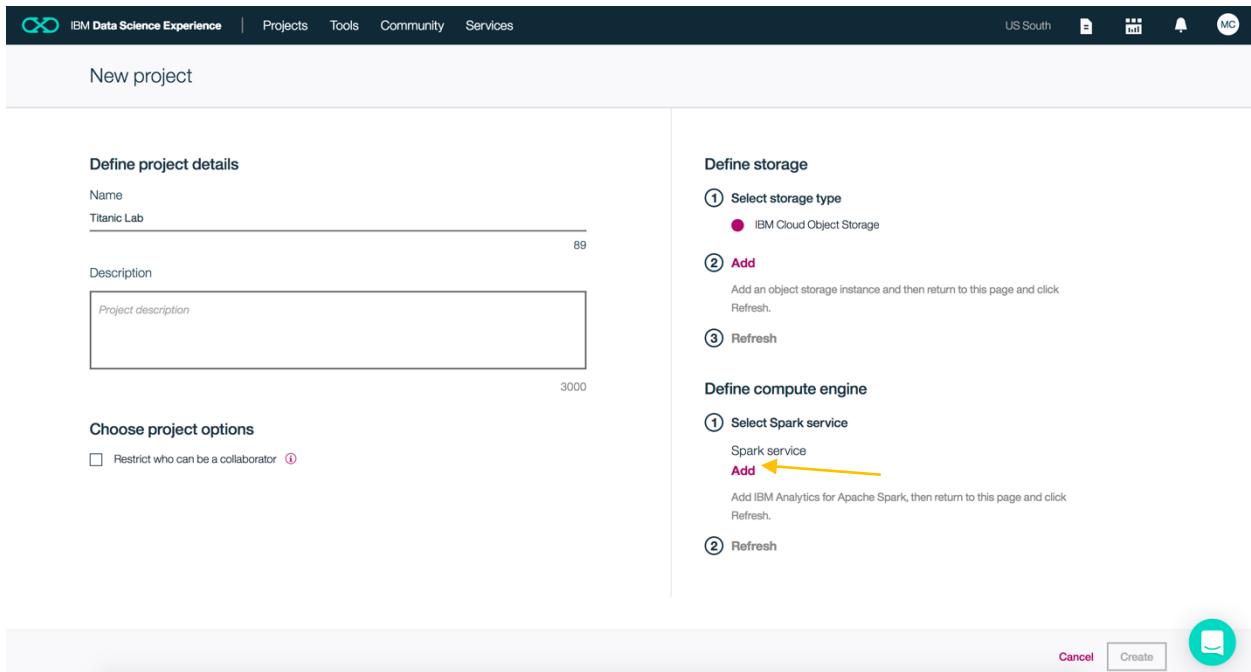
3.A. If a storage service is not available, click **Add**.

3.B. Click on **Lite** to choose the Lite plan and then click on **Create**.

3.C. When a “Confirm Creation” popup appears, click **Confirm.**



3.D. If a Spark service is not available, click **add.**



3.E. Click on **Lite to choose the Lite plan and then click on **Create**.**

IBM Data Science Experience | Projects Tools Community Services US South

Apache Spark

Existing New

Apache Spark

Apache Spark is an open source cluster computing framework optimized for extremely fast and large scale data processing, which you can access via the newly integrated notebook interface IBM Analytics for Apache Spark. You can connect to your existing data sources or take advantage of the on-demand big data optimization of Object Storage. Spark plans are based on the maximum number of executors available to process your analytic jobs. Executors exist only as long as they're needed for processing, so you're charged only for processing done.

Features	Pricing
Incredibly Fast Apache Spark delivers 100x the performance of Apache Hadoop for certain workloads because of its advanced in-memory computing engine.	Easy to Use and Powerful Apache Spark's Streaming and SQL programming models backed by MLlib and GraphX make it incredibly easy for developers and data scientists to build apps that exploit machine learning and graph analytics. Because the service is 100% compatible with Apache Spark, developers can build their apps and run them against the IBM managed service to benefit from operational, maintenance, and hardware excellence.
	Convenient Data Storage Object Storage enables a convenient way to upload your data from a file for immediate use by your Spark instance. You can set up Object Storage directly from the Spark service interface.

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

Plan	Features	Pricing
<input checked="" type="radio"/> Lite An entry level plan to run programs using up to 2 Spark executors	2 Spark Executors	Free

Terms Cancel Create

3.F. When a “Confirm Creation” popup appears, click **Confirm.**

IBM Data Science Experience | Projects Tools Community Services US South

Apache Spark

Existing New

Apache Spark

Apache Spark is an open source cluster computing framework optimized for extremely fast and large scale data processing, which you can access via the newly integrated notebook interface IBM Analytics for Apache Spark. You can connect to your existing data sources or take advantage of the on-demand big data optimization of Object Storage. Spark plans are based on the maximum number of executors available to process your analytic jobs. Executors exist only as long as they're needed for processing, so you're charged only for processing done.

Confirm Creation

Organization: michael.cronk_organization1

Plan: **Lite**

Space: **space1**

Service name: **Spark-rj**

Data Storage: enables a convenient way to upload your data from a file for immediate use by your Spark instance. You can set up Object Storage directly from the Spark service interface.

Cancel Confirm

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

Plan	Features
<input checked="" type="radio"/> Lite An entry level plan to run programs using up to 2 Spark executors	2 Spark Executors

Terms Cancel Create

3.G. Return to the “New Project” page and click **Refresh** under “Select Storage Type” and under “Define Compute Engine.” After the refresh completes, a selection should appear (see the Spark service in the image below).

The screenshot shows the 'New project' page in the IBM Data Science Experience. On the right side, under 'Define storage', there is a step-by-step process:

- ① Select storage type (radio button for IBM Cloud Object Storage)
- ② Add (instructions: Add an object storage instance and then return to this page and click Refresh.)
- ③ Refresh (highlighted with a yellow arrow)

Under 'Define compute engine', there is a 'Select Spark service' section with a dropdown menu showing 'Spark-rj'. A warning message states: "If you associate the same Spark service with multiple projects, the Spark history server will display job history information for all the projects."

3.H. Now that we have both a storage type and a Spark service, click **Create**.

The screenshot shows the 'New project' page in the IBM Data Science Experience. On the right side, under 'Define storage', the 'Select storage type' step is checked, and the 'Target Cloud Object Storage Instance' dropdown is set to 'cloud-object-storage-zz'.

Under 'Define compute engine', the 'Select Spark service' section shows 'Spark-rj' selected. A warning message states: "If you associate the same Spark service with multiple projects, the Spark history server will display job history information for all the projects."

At the bottom right of the page, there is a 'Create' button, which is highlighted with a yellow arrow.

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

The screenshot shows the 'Overview' tab selected in the navigation bar. The project name 'Titanic-SPSS-Lab' is displayed, along with the last update date: Nov 30 2017. Below this, there are sections for 'Date created', 'Description', 'Storage', 'Collaborators', and 'Bookmarks'. On the right side, there is a 'Recent activity' section which is currently empty. At the bottom, there are counts for 'Assets' (0), 'Bookmarks' (0), and 'Collaborators' (1). A yellow arrow points to the 'Overview' tab in the top navigation bar.

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

The screenshot shows a GitHub repository page for 'jpatter / ML-POT'. The file 'titanic_cleaned.csv' is listed under the 'ML-POT / Lab-2 / data' directory. A yellow arrow points to the 'Raw' button in the file preview header. The file preview shows the first few rows of the CSV data.

	pclass	survived	name	sex	sibsp	parch	ticket	fare	embarked
1	1	1	Allen, Miss. Elisabeth Walton	female	0	0	24160	211.337500	S
2	1	1	Allison, Master. Hudson Trevor	male	1	2	113781	151.550000	S
3	1	0	Allison, Miss. Helen Loraine	female	1	2	113781	151.550000	S
4	1	0	Allison, Mr. Hudson Joshua Creighton	male	1	2	113781	151.550000	S
5	1	0	Allison, Mrs. Hudson J C (Bettie Mabel Creighton)	female	1	2	113781	151.550000	S

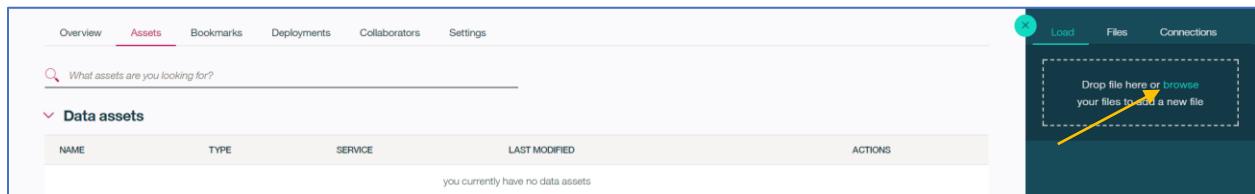
2. Go back to the Titanic project. Click on **New data asset**.



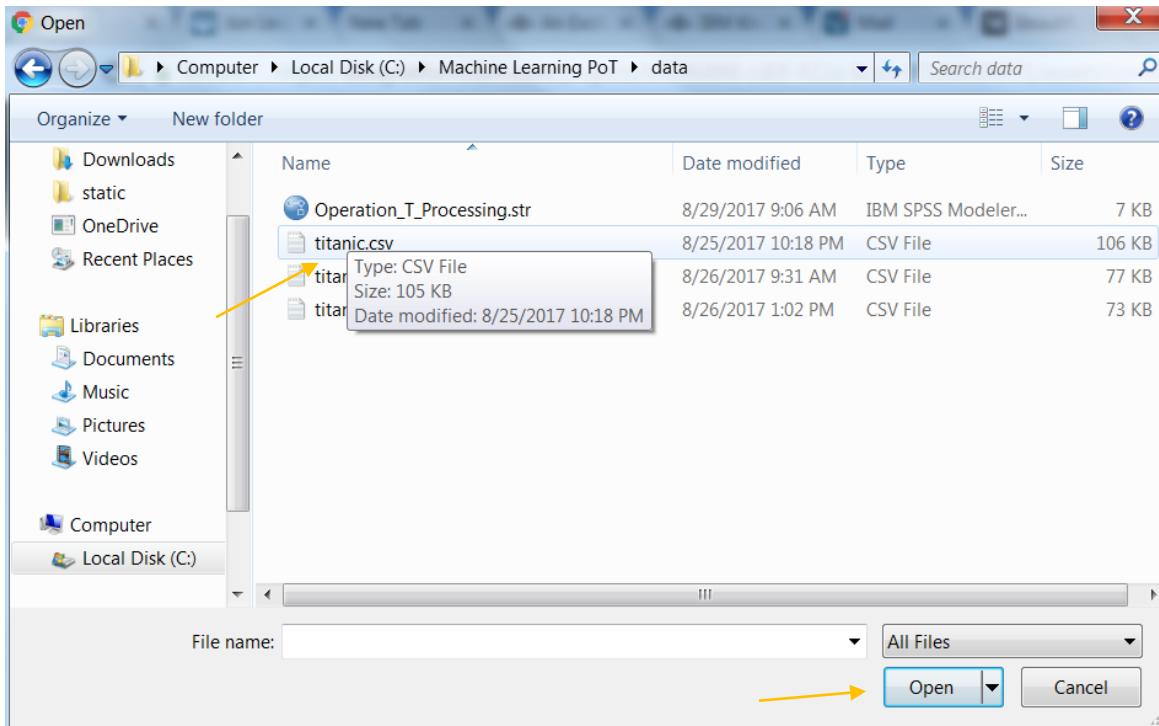
3. Click on the **Load** tab.



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 now added as a Data Asset.

Data assets					
0 assets selected.					
	NAME	TYPE	SERVICE	CREATED BY	LAST MODIFIED
<input type="checkbox"/>	 titanic.csv	Data Asset	Project	Michael	7 Feb 2018, 5:01:18 pm

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

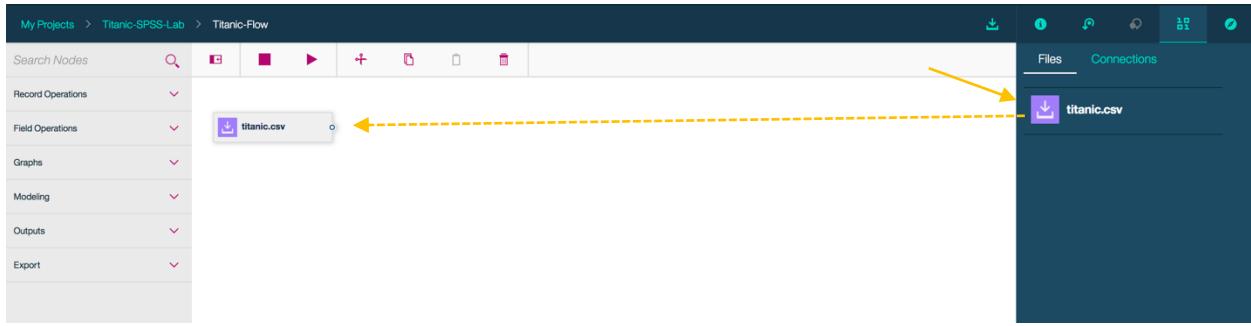
1. In project Titanic, click on **New flow** in the SPSS Modeler flows section.

The screenshot shows a table titled "SPSS Modeler flows". It has columns for NAME, CREATED BY, LAST MODIFIED, and ACTIONS. A pink button labeled "+ New flow" is located in the ACTIONS column. A yellow arrow points to this button.

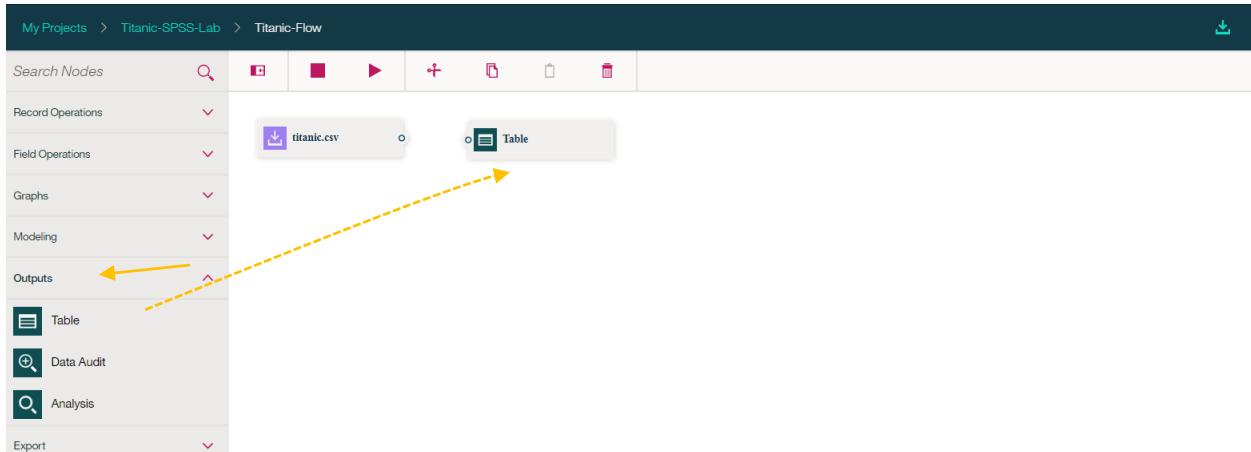
2. Enter a **Name** for the flow, optionally enter a **Description**, select IBM SPSS Modeler for the **Runtime**, and click on **Create**.

The screenshot shows the "New" dialog in SPSS Modeler. It includes fields for Name (set to "Titanic-Flow"), Description (with placeholder text "Type description here"), Runtime (set to "IBM SPSS Modeler"), and a "Create" button. A pink box highlights the "Create" button, and a yellow arrow points to it. Other buttons like "Cancel" are also visible.

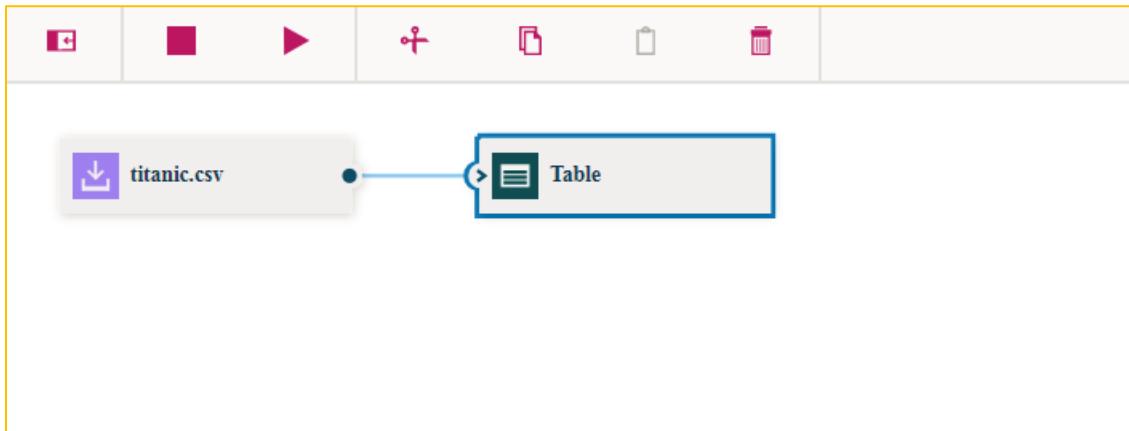
3. This opens the Flow Editor. Click on the titanic.csv file and hold the left mouse key and **drag the file onto the left side of the canvas**. Release the left mouse key.



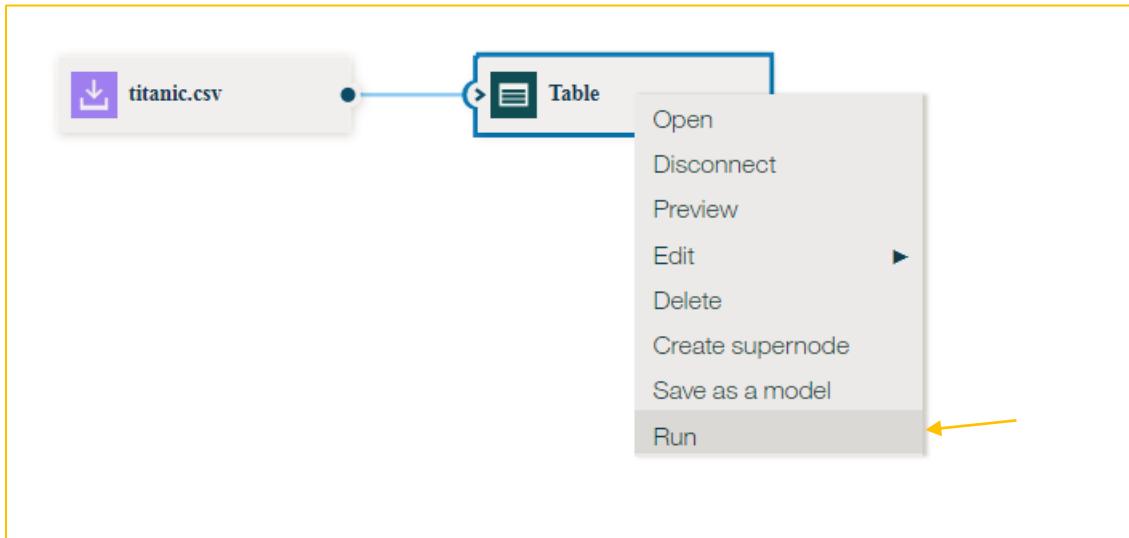
4. Click on the **Outputs** menu item in the Node Palette on the left and then click on the **Table** icon and drag the icon to the right of the titanic.csv icon. The SPSS Table node will display the contents of the csv file. If the Node Palette is not visible, click on the Node Palette icon 



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**.



7. 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.



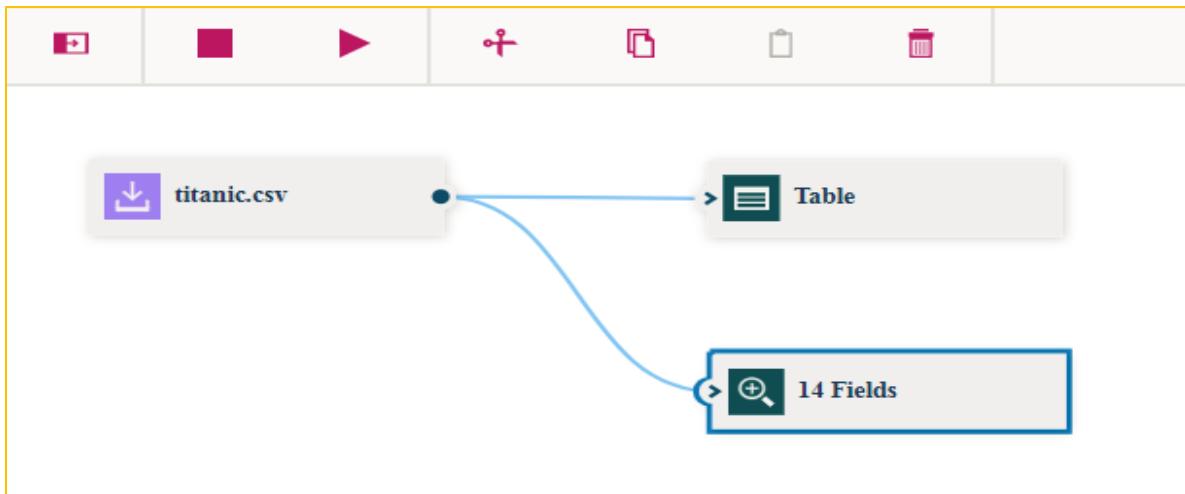
8. Double click on the Table selection above the contents of the titanic.csv will be displayed. Each row contains information on a passenger on the Titanic. We will use this data to make predictions on survivability.

PCCLASS	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. Hudson	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 J.	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. Thomas	male	39	0	0	112050	0	A36	S	
1	1	Appleton, Mrs. Edwina	female	53	2	0	11769	51.4792	C101	S	D
1	0	Artagaveyfia, Mr. Rar	male	71	0	0	PC 17609	49.5042	C		
1	0	Astor, Col. John Jacob	male	47	1	0	PC 17757	227.525	C62 C64	C	
1	1	Astor, Mrs. John Jacob	female	16	1	0	PC 17757	227.525	C62 C64	C	4
1	1	Aubart, Mme. Leonida	female	24	0	0	PC 17477	69.3	B35	C	9
1	1	Barber, Miss. Ellen "I"	female	26	0	0	19677	78.65		S	6
1	1	Barkworth, Mr. Alger	male	80	0	0	27042	30	A23	S	B

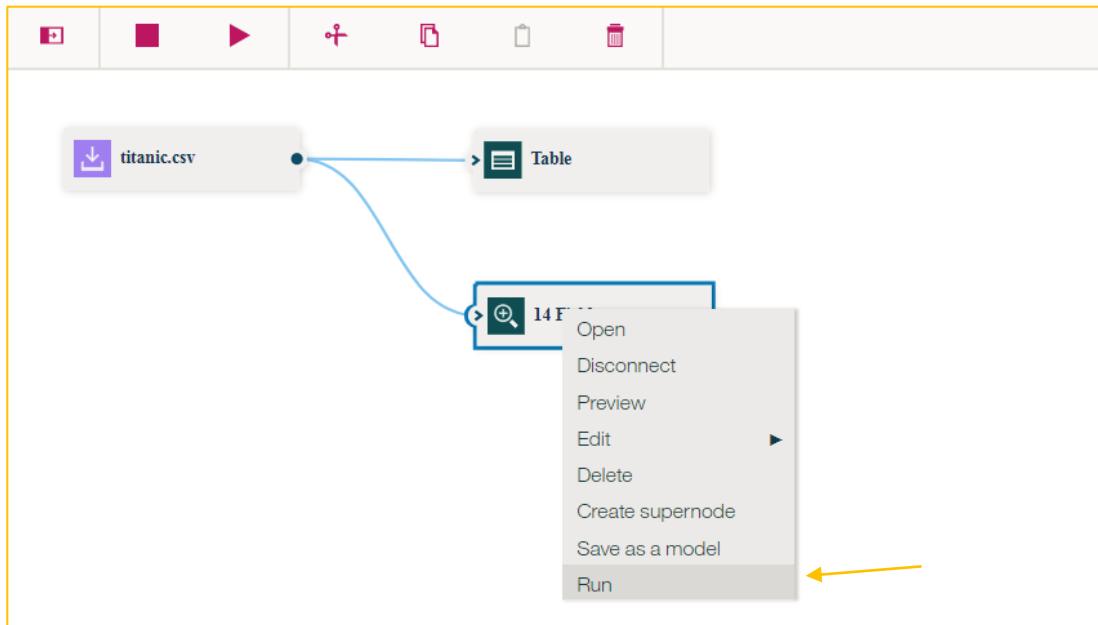
Step 3.2 Explore the Data using the Data Audit Node

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.

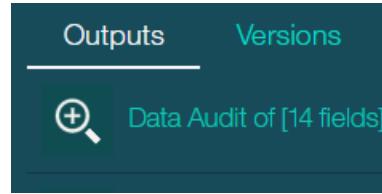
1. Add a **Data Audit** node to the flow clicking on the **Outputs** menu item in the Node Palette, and then dragging the **Data Audit** node to underneath the titanic.csv node. If the Node Palette is not visible, click on the Node Palette icon . Connect the titanic.csv node to the Data Audit node. The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

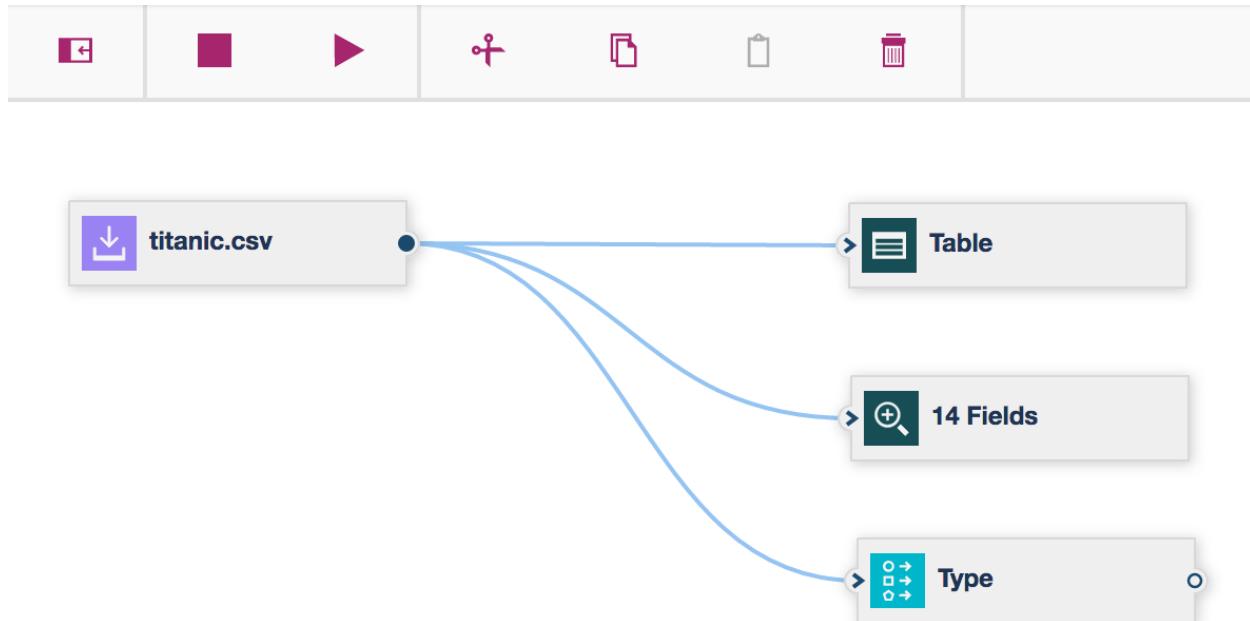
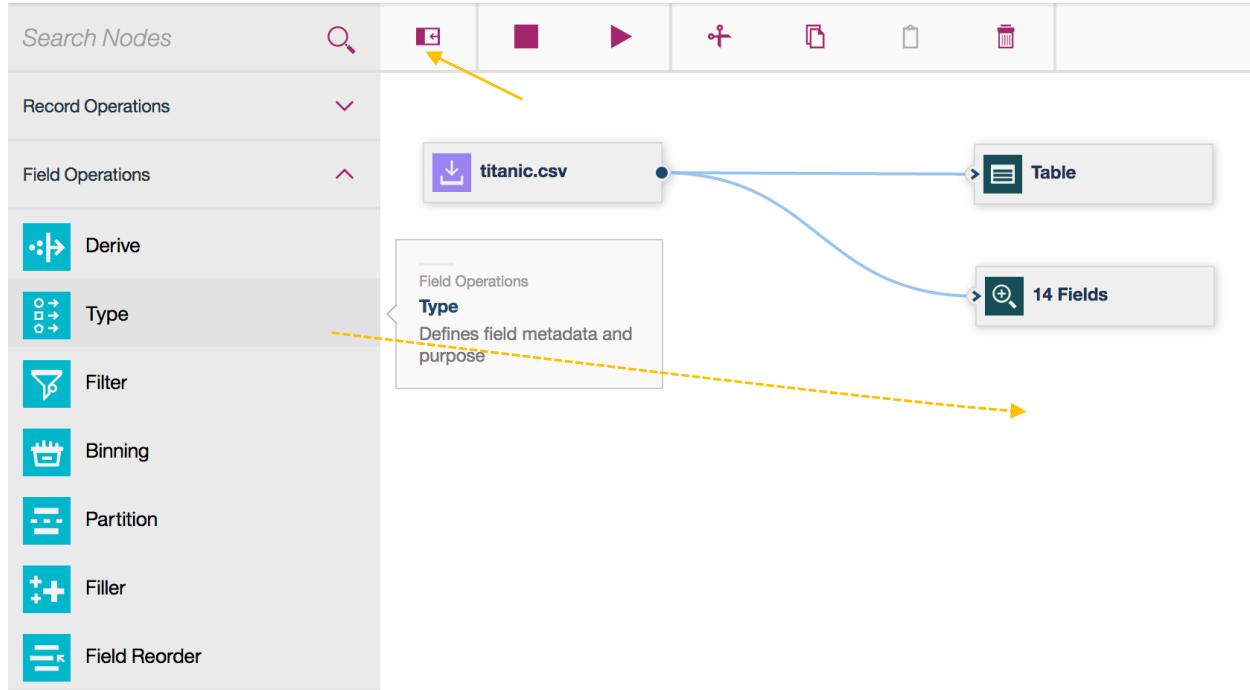
	Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
1	pclass		Continuous	1	3	2.295	0.838	-0.599	--	1309
2	survived		Continuous	0	1	0.382	0.486	0.486	--	1309
3	name		Categorical	--	--	--	--	--	--	1309
4	sex		Categorical	--	--	--	--	--	2	1309
5	age		Continuous	0.167	80.000	29.881	14.413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

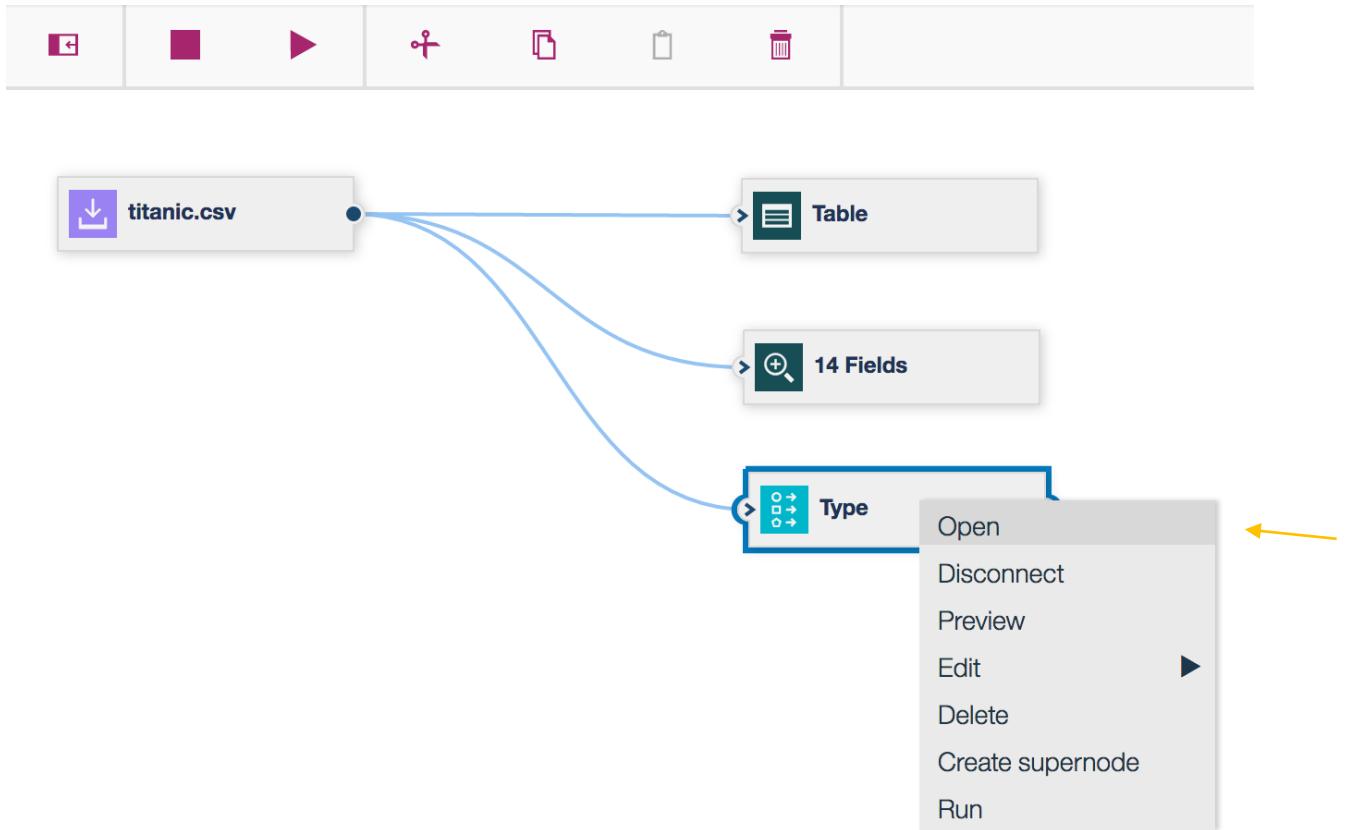
Step 3.3 Explore the Data using Graph Nodes.

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.

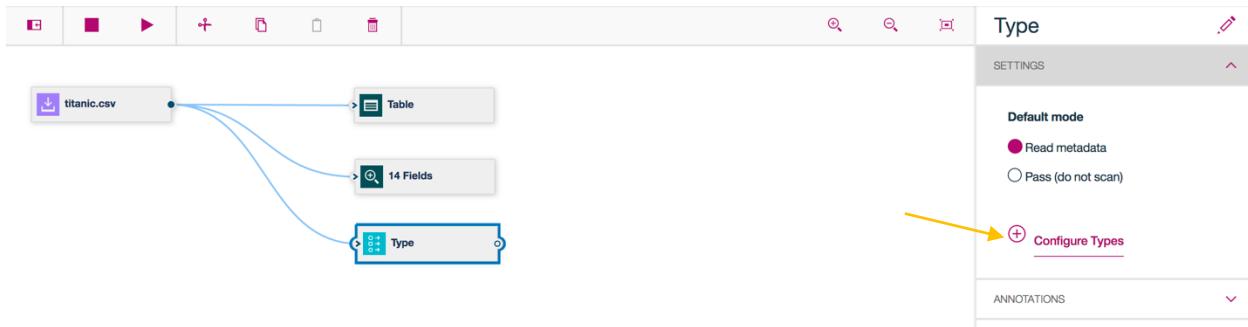
1. Add a **Type** node to the flow by clicking on the **Field Operations** menu item in the Node Palette and then drag the **Type** node underneath the **Data Audit** node. If the Node Palette is not visible, click on the Node Palette icon . Connect the titanic.csv node to the **Type** node. The canvas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



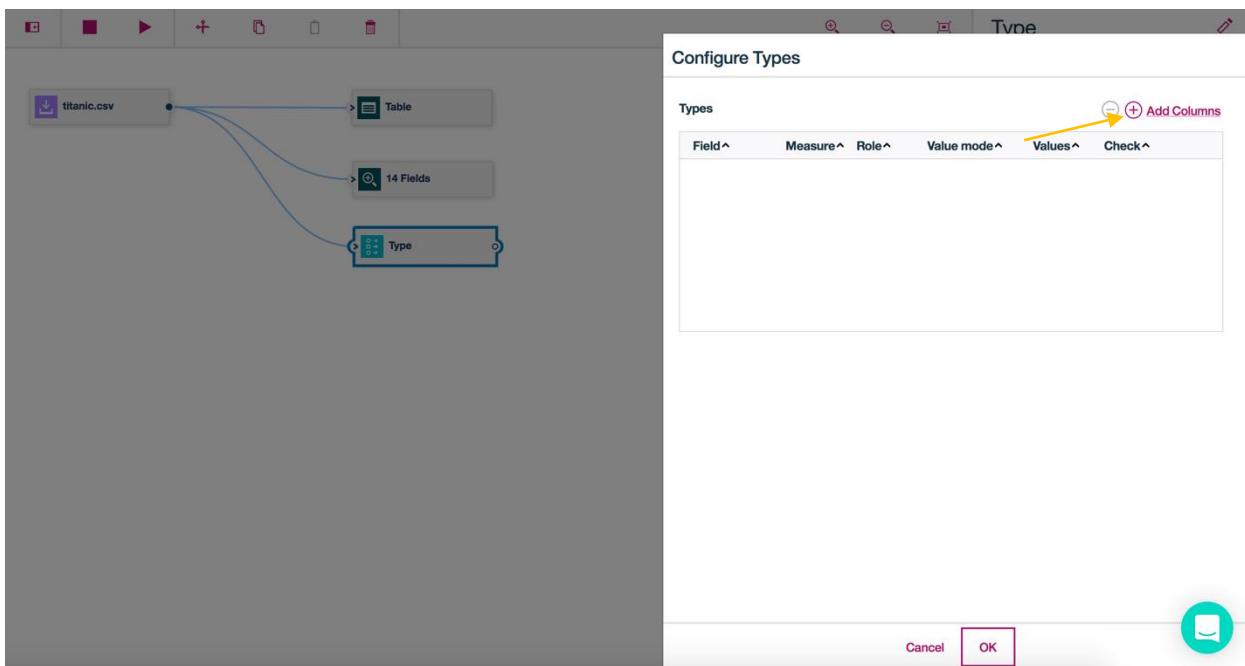
2. Double click on the **Type** node. This will open a **Type** menu pallet on the right side of the screen.



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



4. Select Add Columns.



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

Search in column Field name Filter:

	Field name^	Data type^
<input checked="" type="checkbox"/>	pclass	integer
<input checked="" type="checkbox"/>	survived	integer
<input type="checkbox"/>	name	string

Reset ↻

Yellow arrows point to the checkboxes next to 'pclass' and 'survived' in the list.

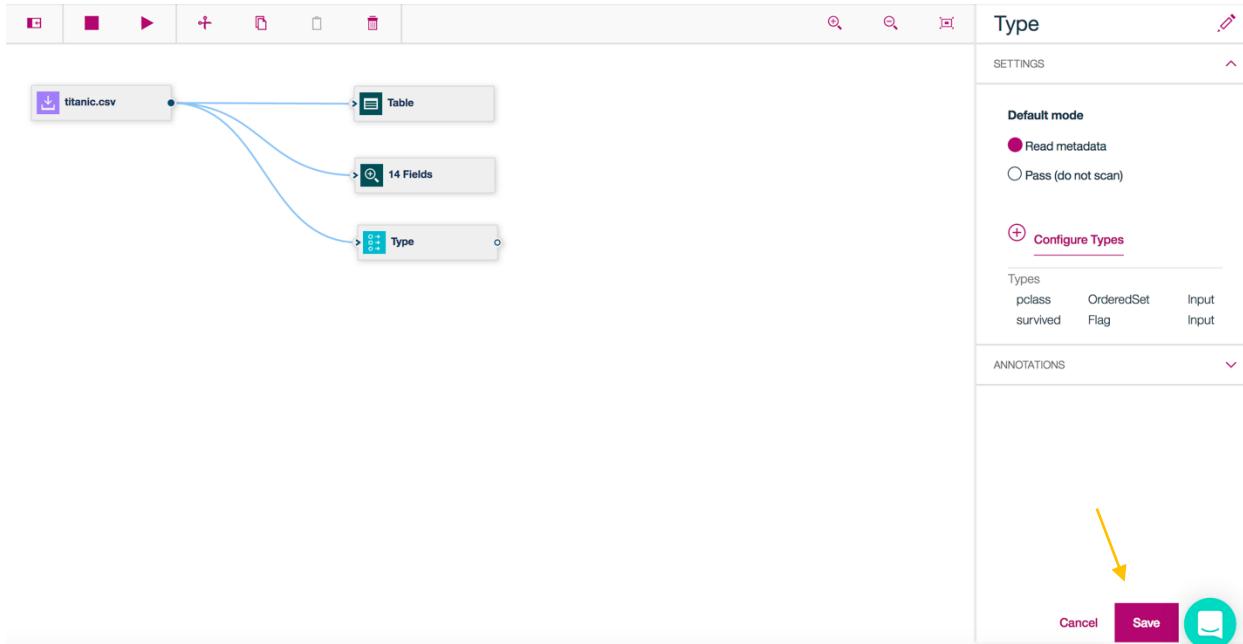
6. Click on the measurement level field for **pclass** and select **Ordered Set**. Click on the measurement level field for **survived** and select **Flag**. Click on **OK**.

Configure Types

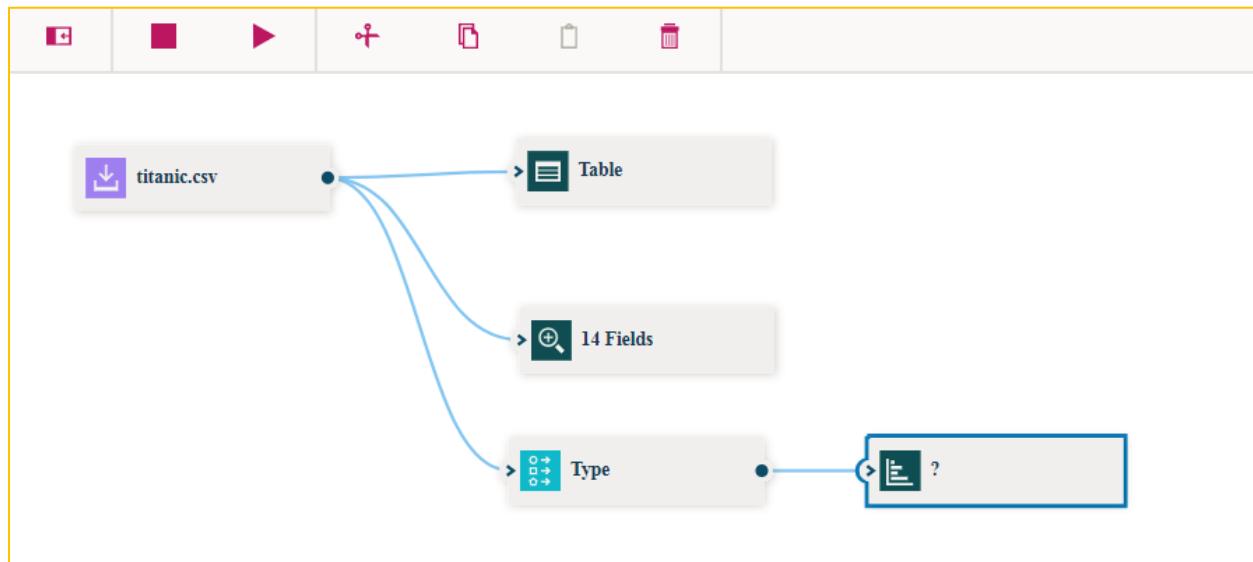
Field ^	Measure ^	Role ^	Value mode ^	Values ^	Check ^
pclass	Ordered Set	Input	Read	None	...
survived	Flag	Input	Read	None	...

Cancel OK 

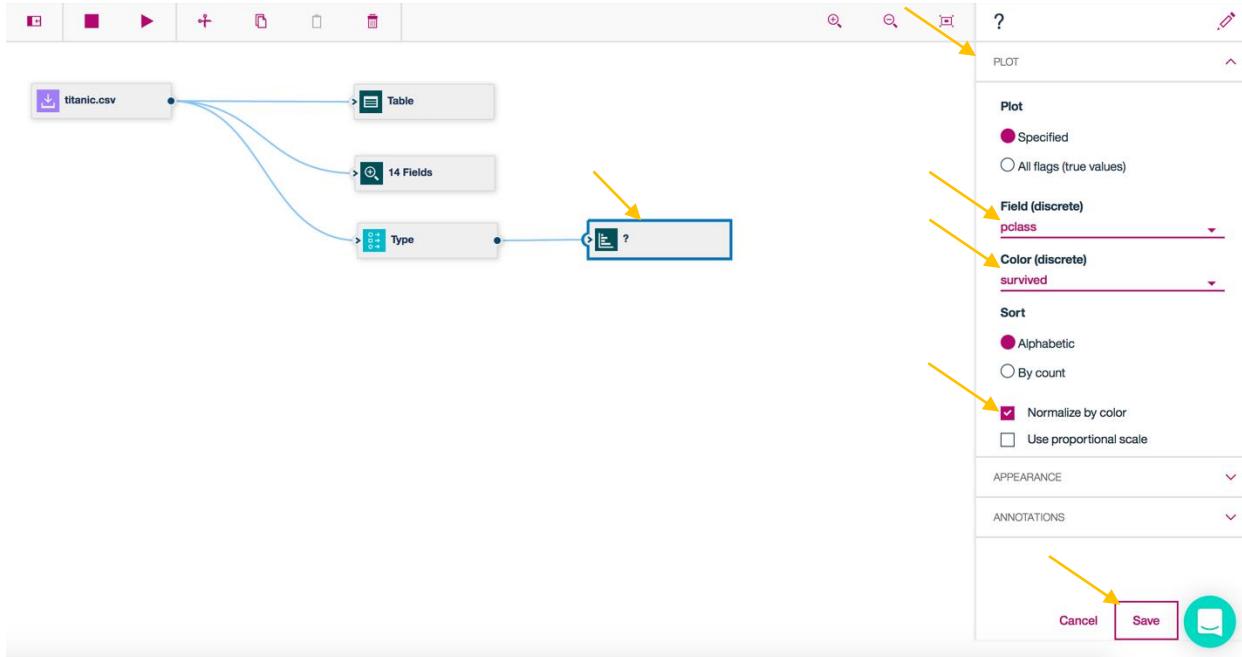
7. Click on **Save** in the bottom right of the Types pallet.



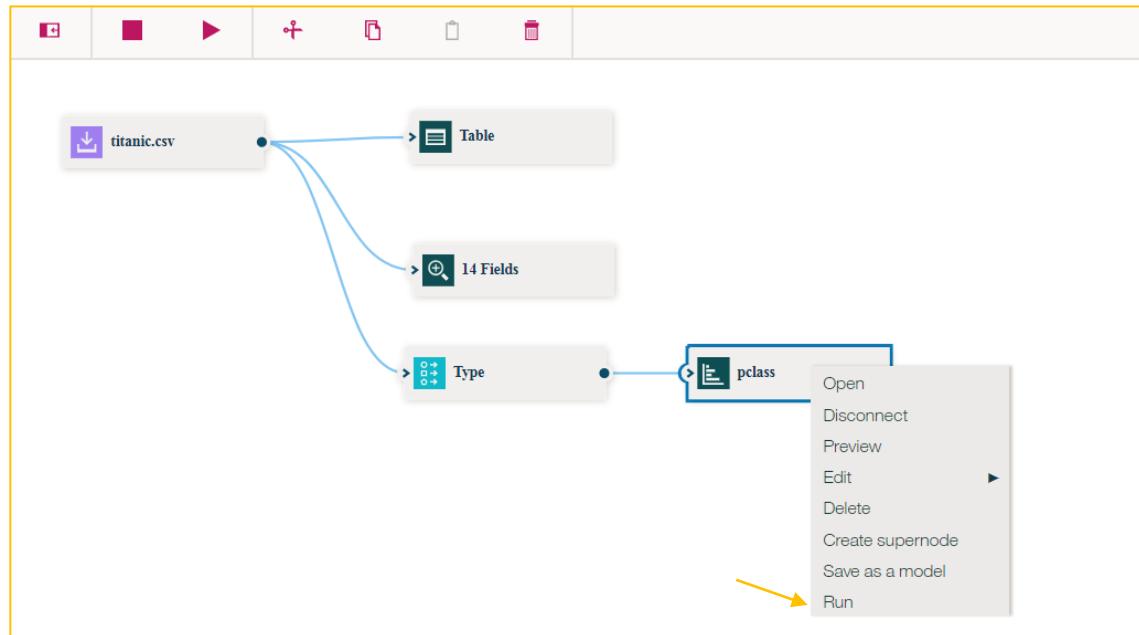
8. Add a **Distribution** node to the flow by clicking on the **Graph** menu item and then dragging the **Distribution** node to the canvas to the right of the **Type** node. If the Node Palette is not visible, click on the Node Palette icon . Connect the **Type** node to the **Distribution** node. The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



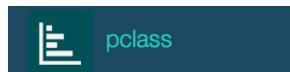
9. Double click on the Distribution Node. Click on the **Plot** dropdown. In the Field (discrete) dropdown, select **pclass**. In the Color (discrete) dropdown, select **survived**. Click on the **normalize by color** checkbox, and then click **Save**.



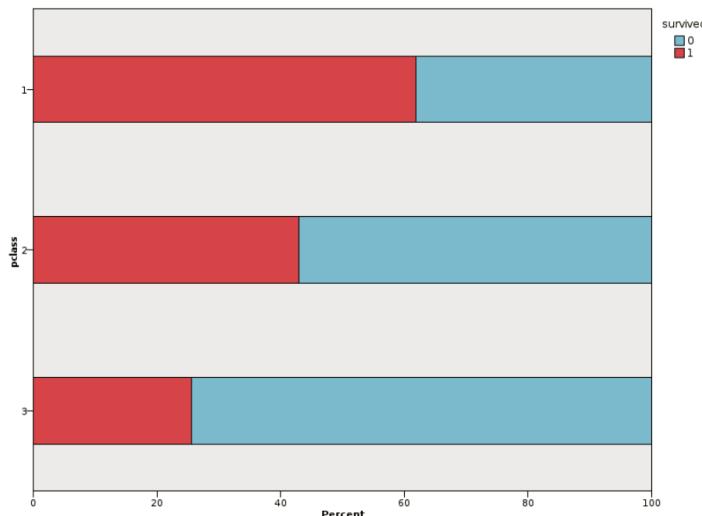
10. Right click on the Distribution node, and select Run.



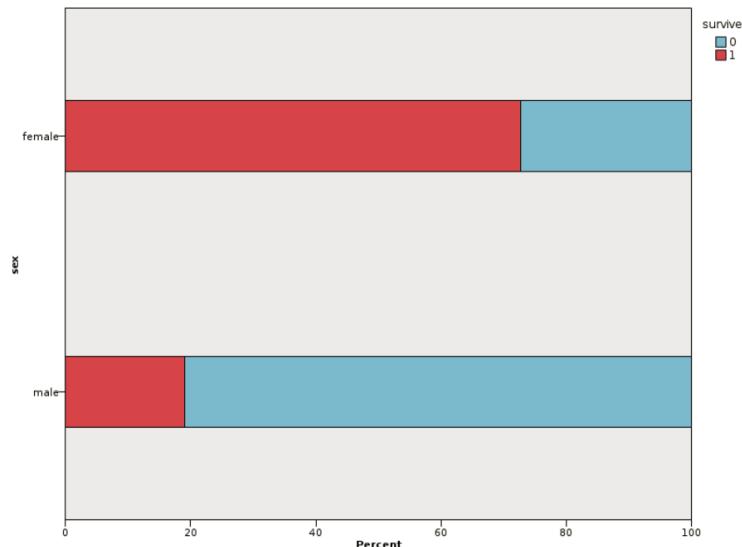
11. The Distribution of pclass output will appear under the **Outputs** tab.



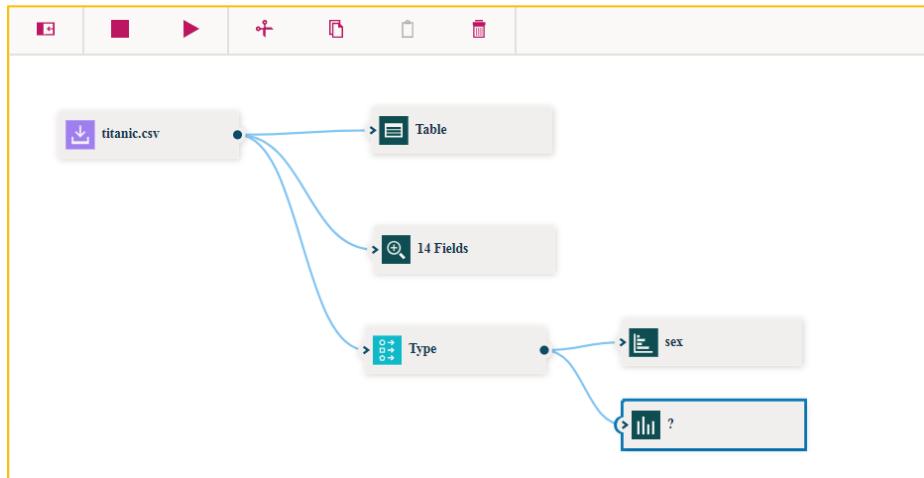
12. Double click on the **Distribution of pclass #1** to view the graph. We can see from the graph that the likelihood of surviving is correlated to the passenger class. The first class passengers have the highest rate of survivability.



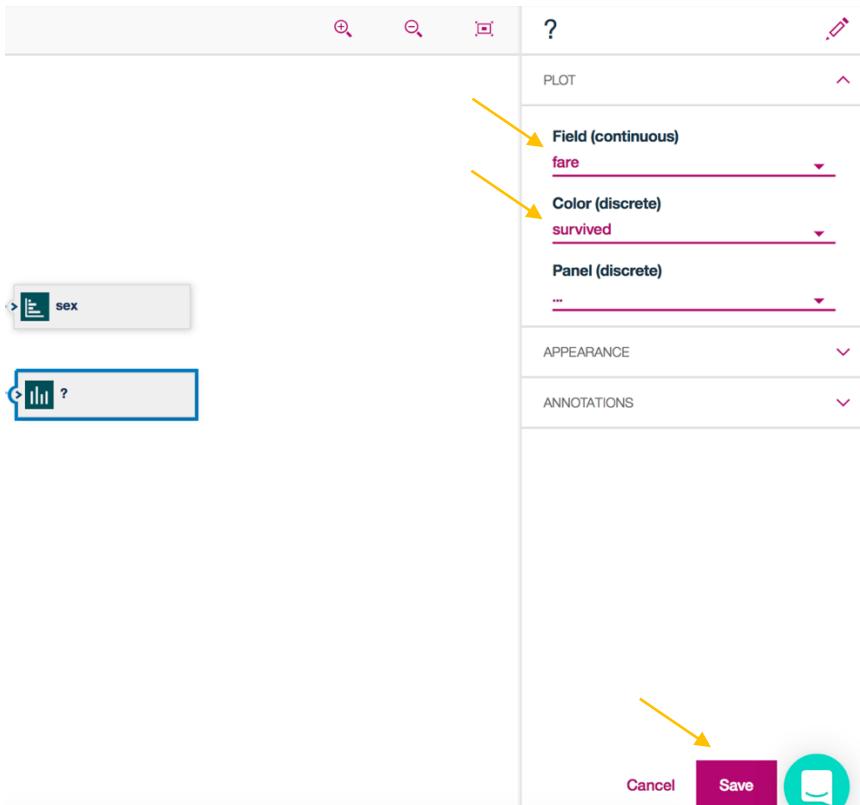
13. You can change the distribution graph to show the survivability by gender by double clicking on the Distribution node and replacing pclass with sex and clicking Save. Re-run the graph by right clicking on the Distribution node and selecting Run. Double click on the Distribution of sex #1 to display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



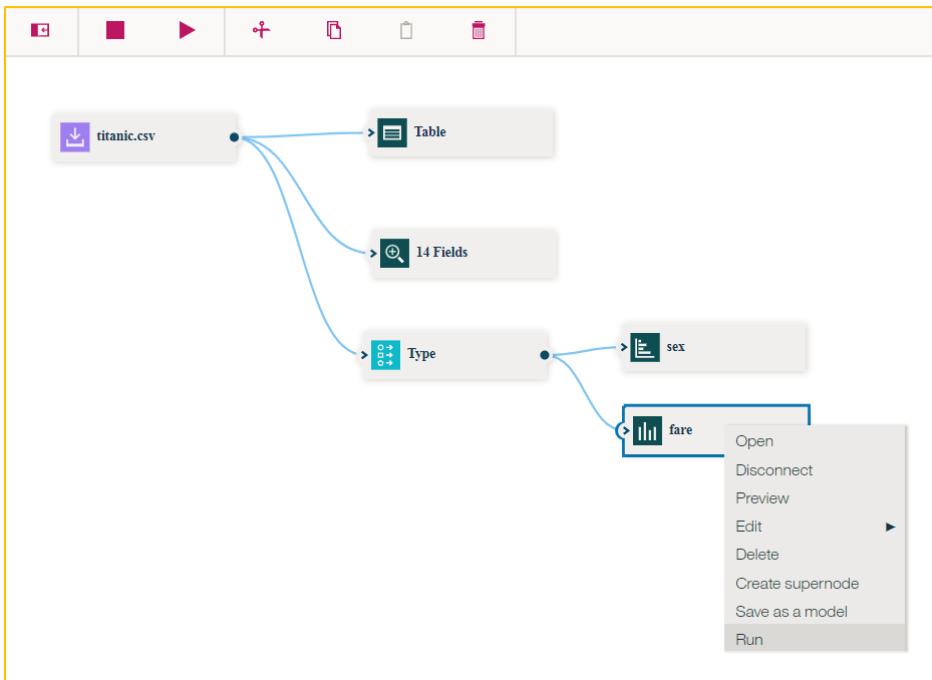
14. Add a **Histogram** node to the flow by clicking on the **Graphs** menu item and then dragging the **Histogram** node to the canvas underneath the **Distribution** node. If the Node Palette is not visible, click on the Node Palette icon . Connect the **Type** node to the **Histogram** node. The canvas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



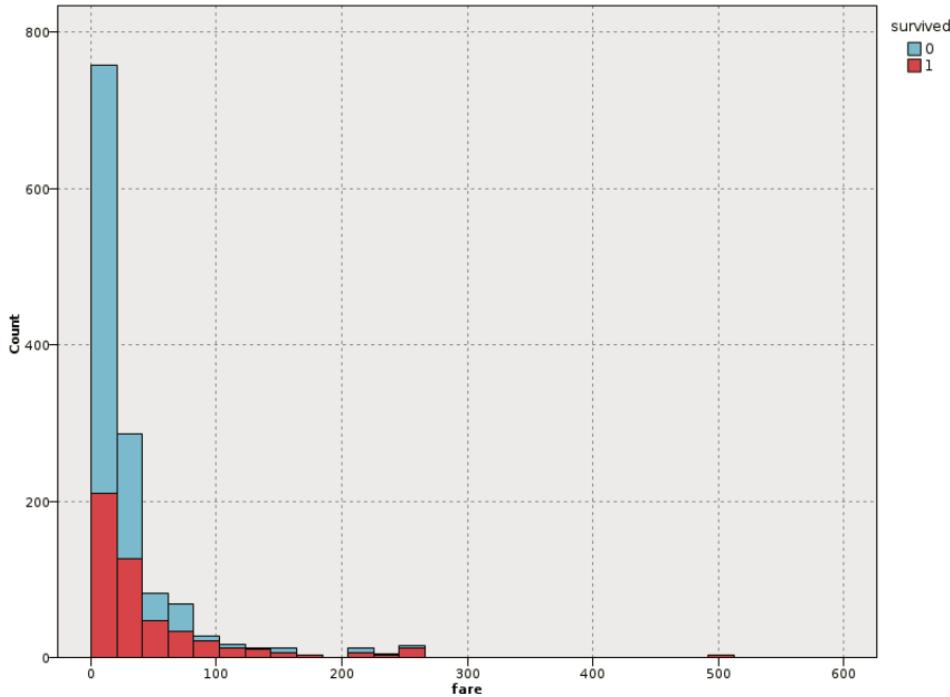
15. Double click on the **Histogram** node. Click on the **Plot** dropdown. Select **fare** from the Field (continuous) dropdown. Select **survived** from the Color (discrete) dropdown. Click on **Save**.



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 3.4 Prepare the Data for Modeling

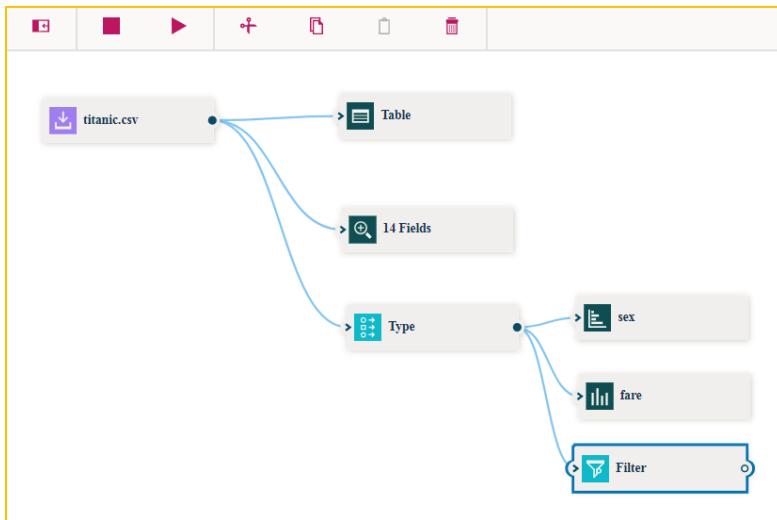
Based on our exploration of the data, there are several transformations that are needed to prepare the data for modeling. This section will introduce, the Filter node, the Select node, and the Derive node that will do the necessary transformations. The Filter and Derive nodes act on a field level, whereas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

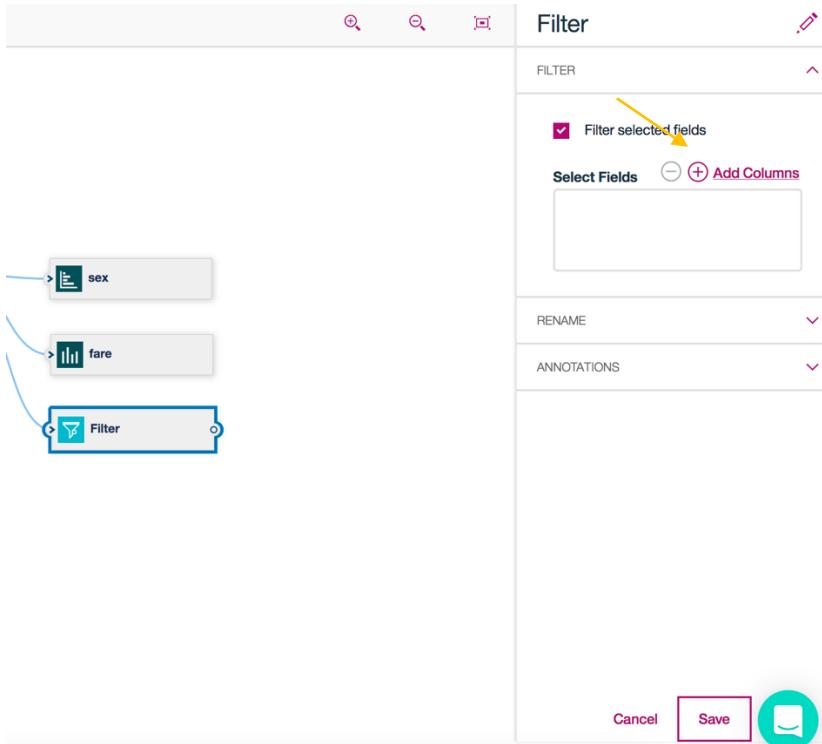
Derive node – The Derive node modifies data values or creates new fields from one or more existing fields. We will use the derive node to do a logarithmic transformation of the fare field. We will also use this node to bin the age and fare fields.

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

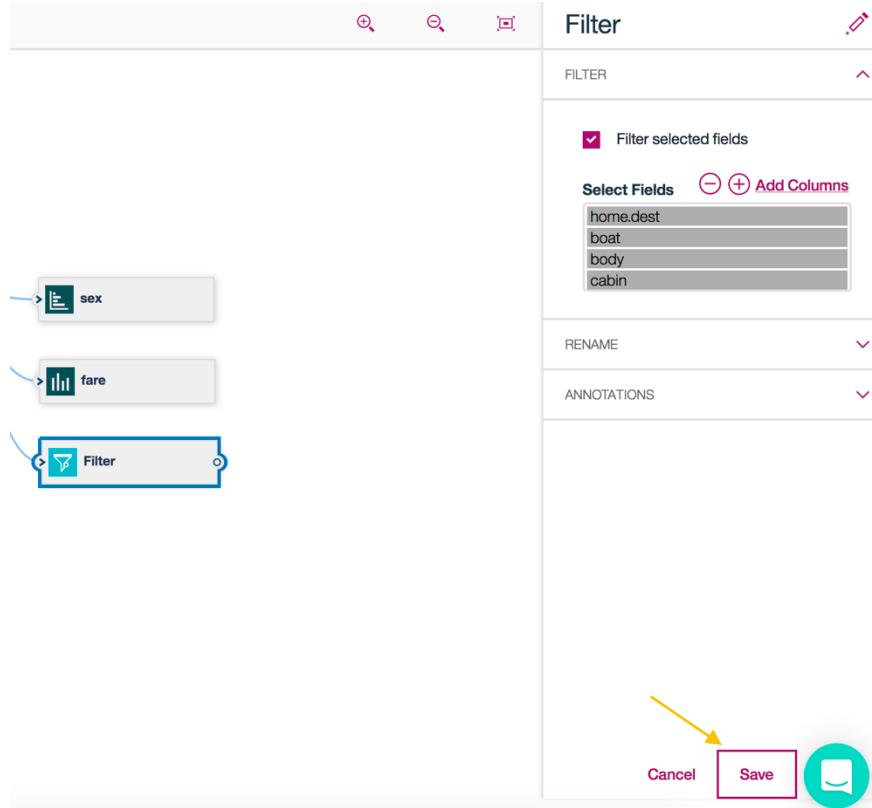


3. Click on the checkboxes adjacent to the **cabin**, **boat**, **body**, and **home.dest** fields, and then click on **Select Fields for Filter**.

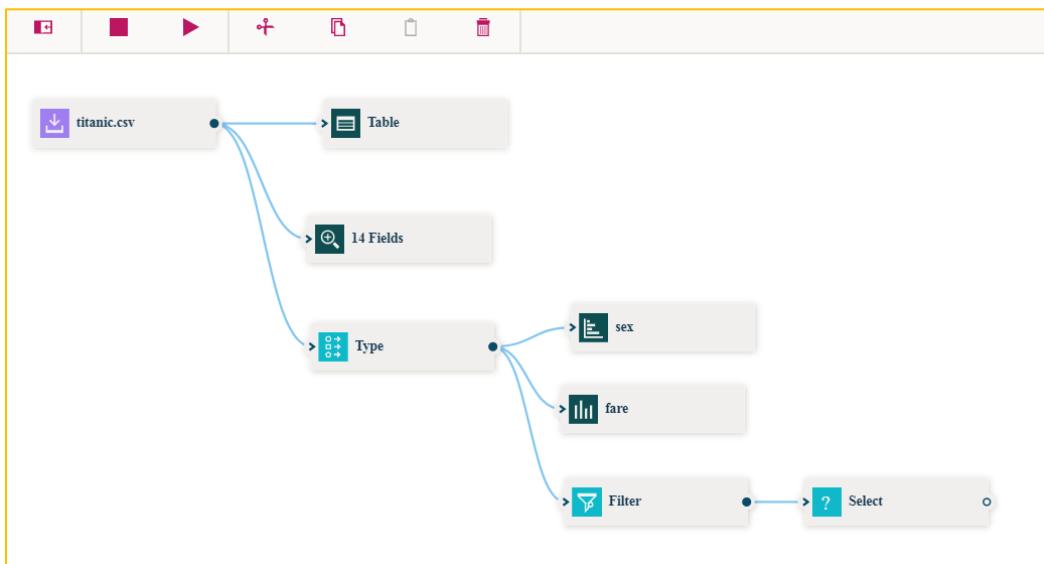
The screenshot shows the 'Select Fields for Filter' dialog. At the top left is a back arrow and the title 'Select Fields for Filter'. At the top right are 'Reset' and a circular 'Apply' button. Below is a search bar with placeholder 'Search in column Field name' and a magnifying glass icon. To the right is a 'Filter' section with three small icons. The main area is a table with two columns: 'Field name' and 'Data type'. The 'Field name' column contains checkboxes. Yellow arrows point to the checkboxes for 'cabin', 'boat', 'body', and 'home.dest'. The 'Data type' column lists the corresponding data types for each field.

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"/> sibsp	integer
<input type="checkbox"/> parch	integer
<input type="checkbox"/> ticket	string
<input type="checkbox"/> fare	double
<input checked="" type="checkbox"/> cabin	string
<input type="checkbox"/> embarked	string
<input checked="" type="checkbox"/> boat	string
<input checked="" type="checkbox"/> body	integer
<input checked="" type="checkbox"/> home.dest	string

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

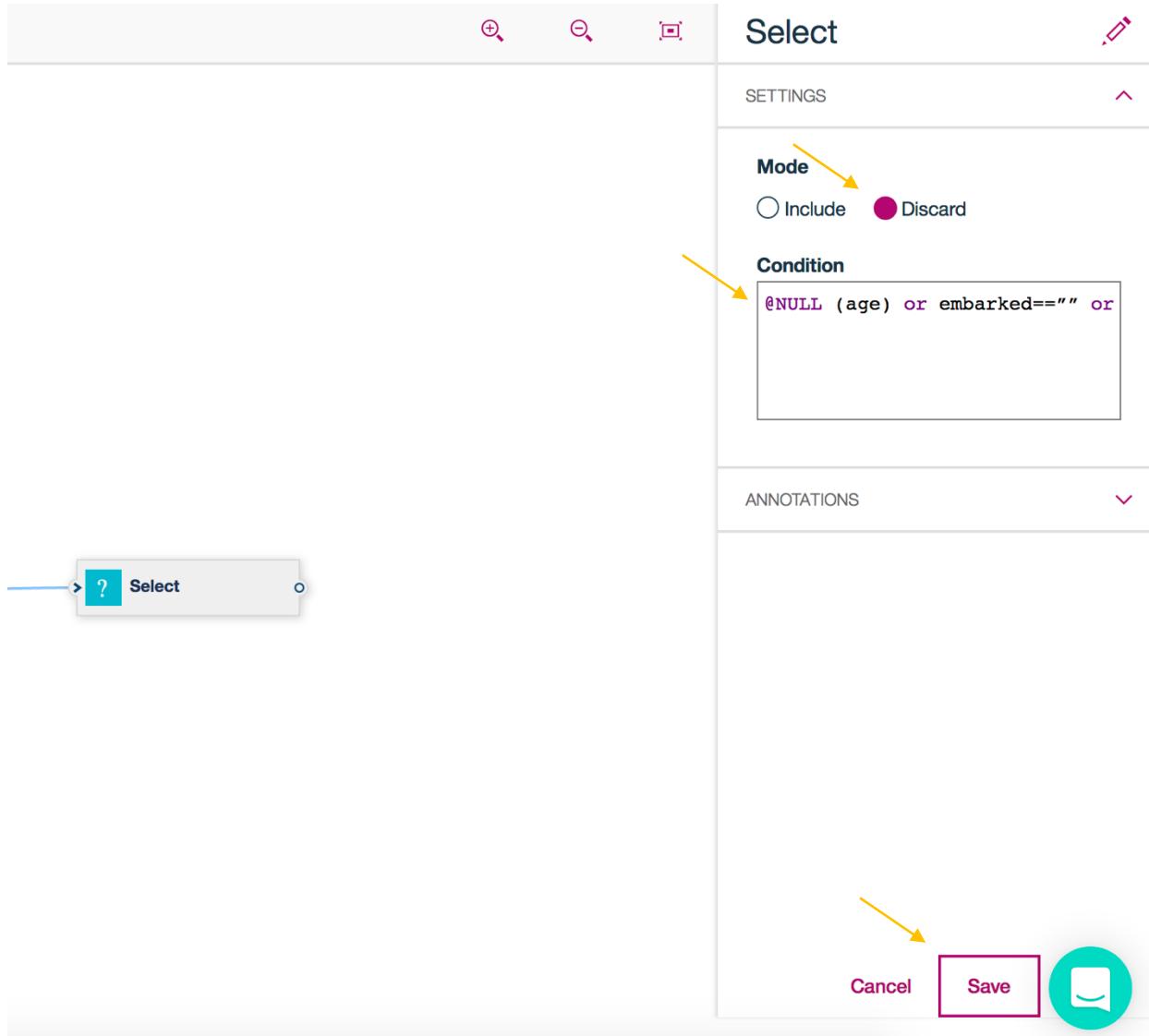


5. Add a **Select** node by clicking on the **Record Operations** menu item in the Node palette, and then dragging the **Select** node to the canvas to the right of the **Filter** node. Connect the **Filter** node to the **Select** node. If the Node Palette is not visible, click on the Node Palette icon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

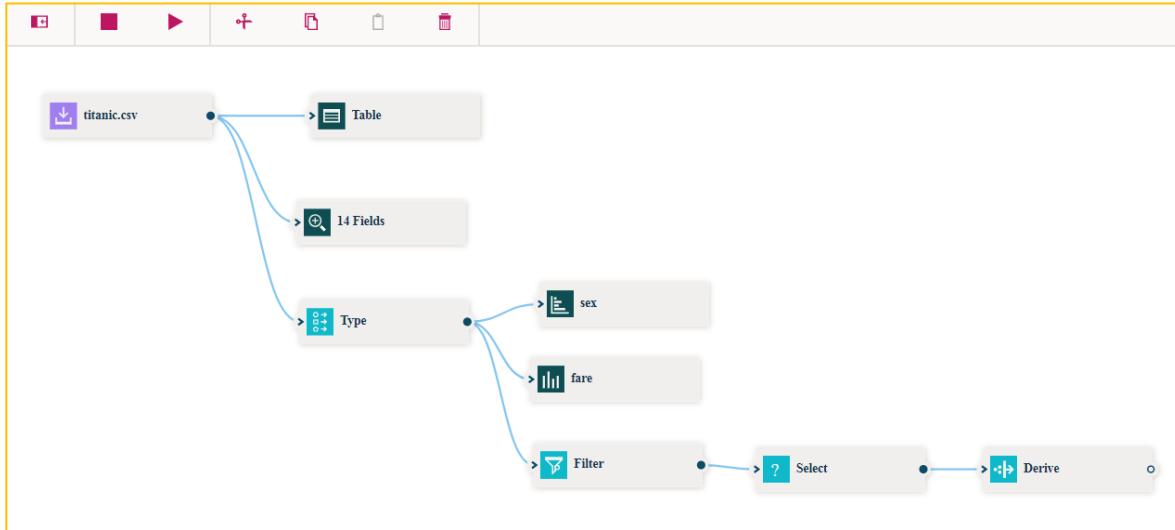


6. Double click on the **Select** node. Click on the **Settings** dropdown. In the **Select** panel, click on the **Discard** radio button, and re-type in the code shown below in the **Condition text box**, and then click **Save**.

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



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.

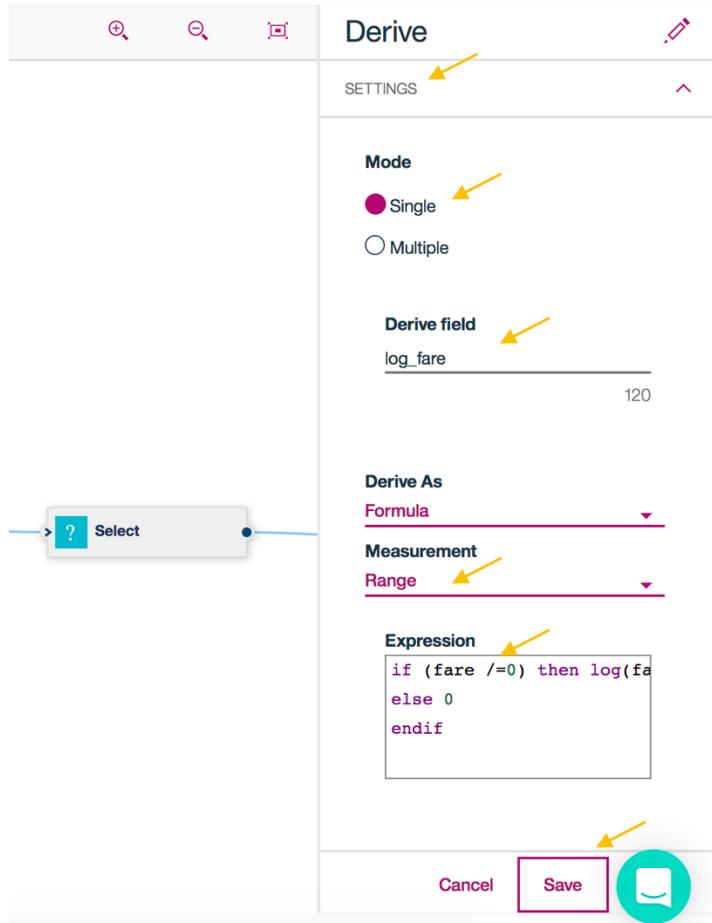


8. Double click on the **Derive** node. Click on the **Settings** Dropdown. Click on the **Single** radio button, enter log_fare for the **Derive field**, select **Range** for the measurement, enter the following code in the **Expression** text box, and click Save.

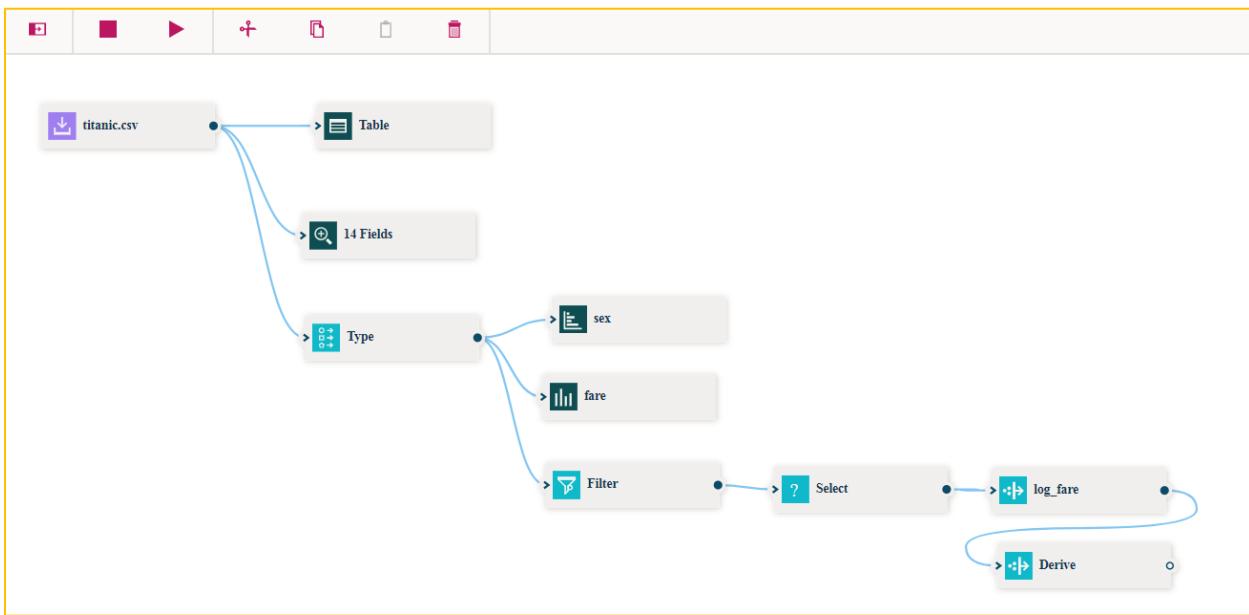
```
if (fare /=0) then log(fare)
```

```
else 0
```

```
endif
```



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.

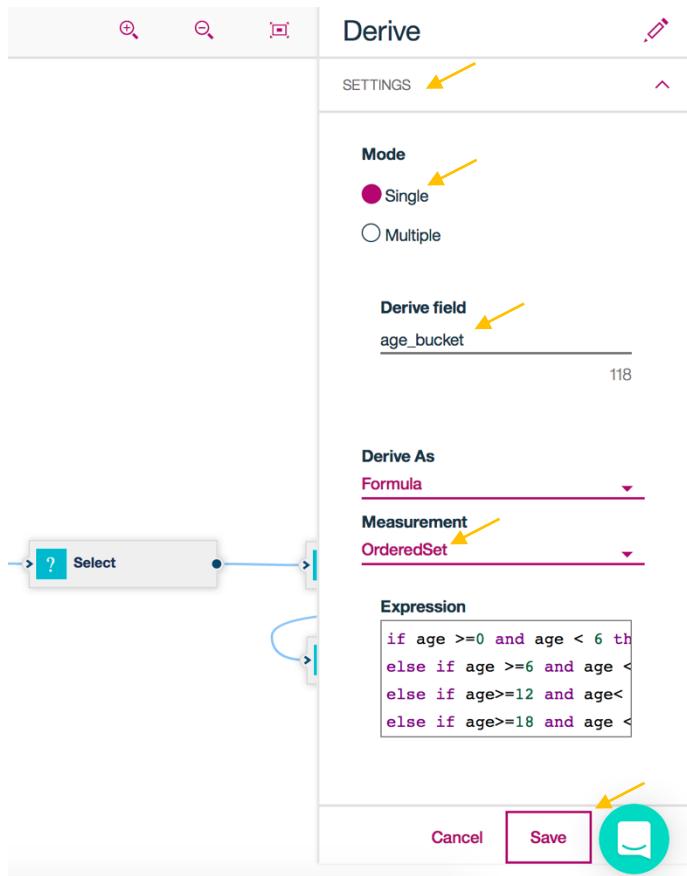


10. Double click on the **Derive** node. Click on the **Settings** dropdown. Click on the **Single** radio button, enter age_bucket for the **Derive field**, select **OrderedSet** for the **Measurement**, enter the following code in the **Expression** text box, and the click **Save**.

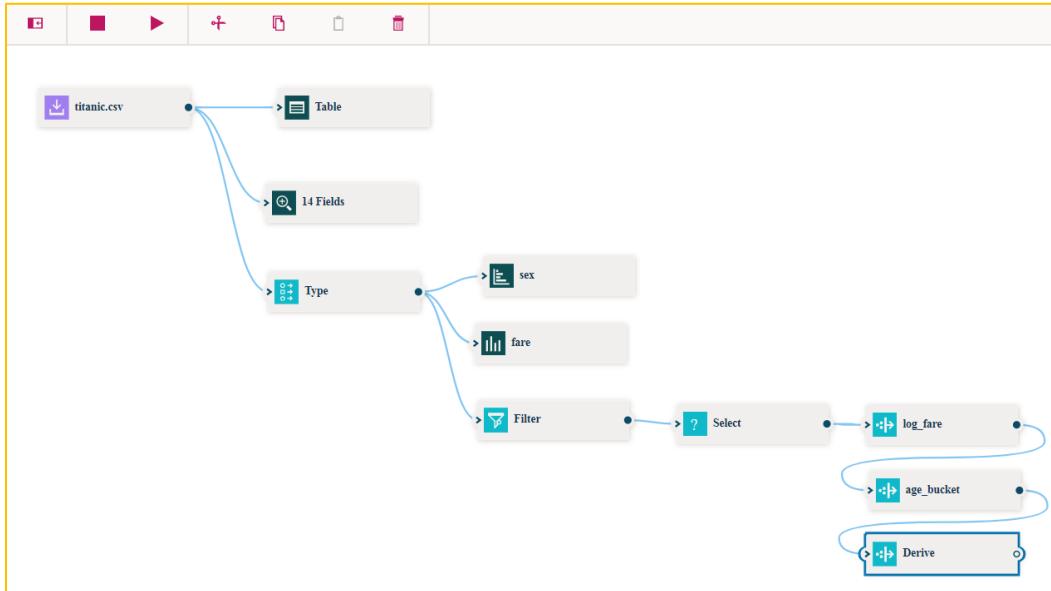
```

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
endif

```

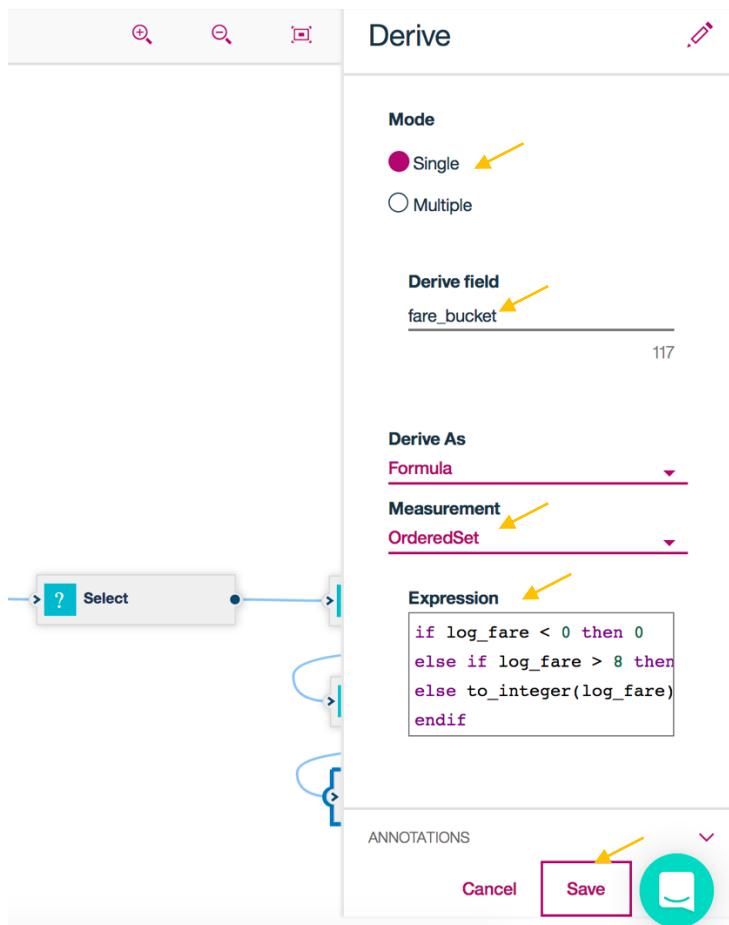


11. Add a **Derive** node by clicking on the Field Operations menu item in the Node palette and dragging the Derive node onto the canvas underneath the age_bucket Derive node. Connect the age_bucket Derive node to the newly created Derive Node. The canvas should appear as below.



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**.

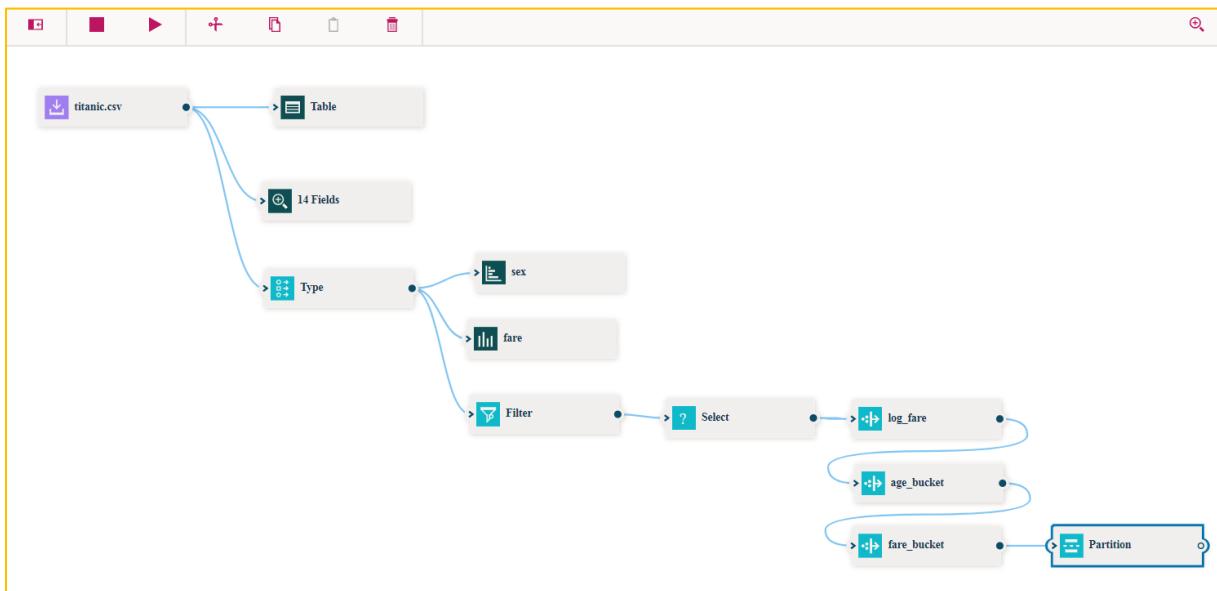
```
if log_fare < 0 then 0
else if log_fare > 8 then 9
else to_integer(log_fare)+1
endif
endif
```



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.

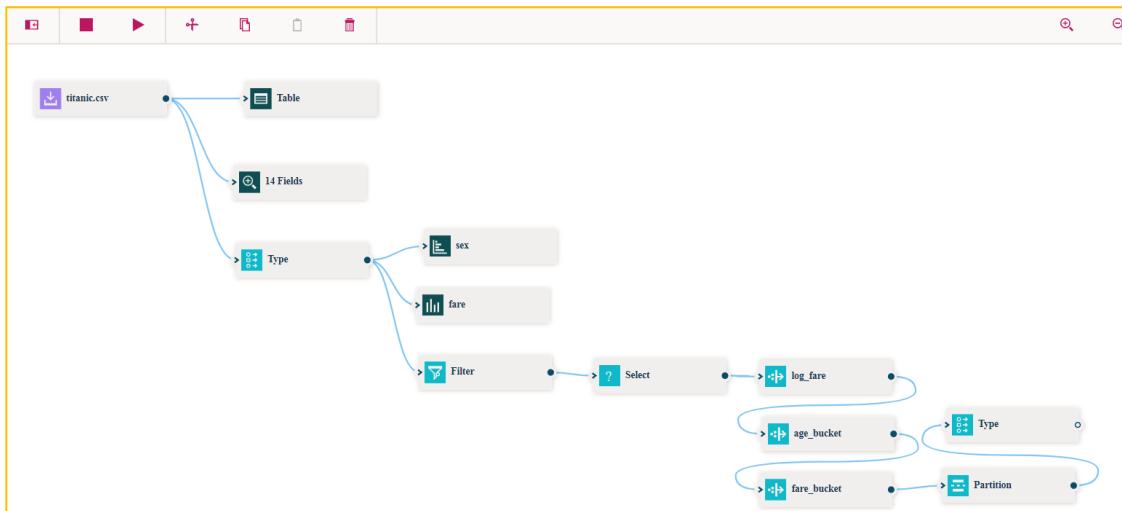
1. Add a **Partition** node by clicking on the Field Operations menu item in the Node palette and dragging the **Partition** node onto the canvas to the right of the fare_bucket **Derive** node. Connect the fare_bucket **Derive** node to the **Partition** node. The canvas should appear as below.



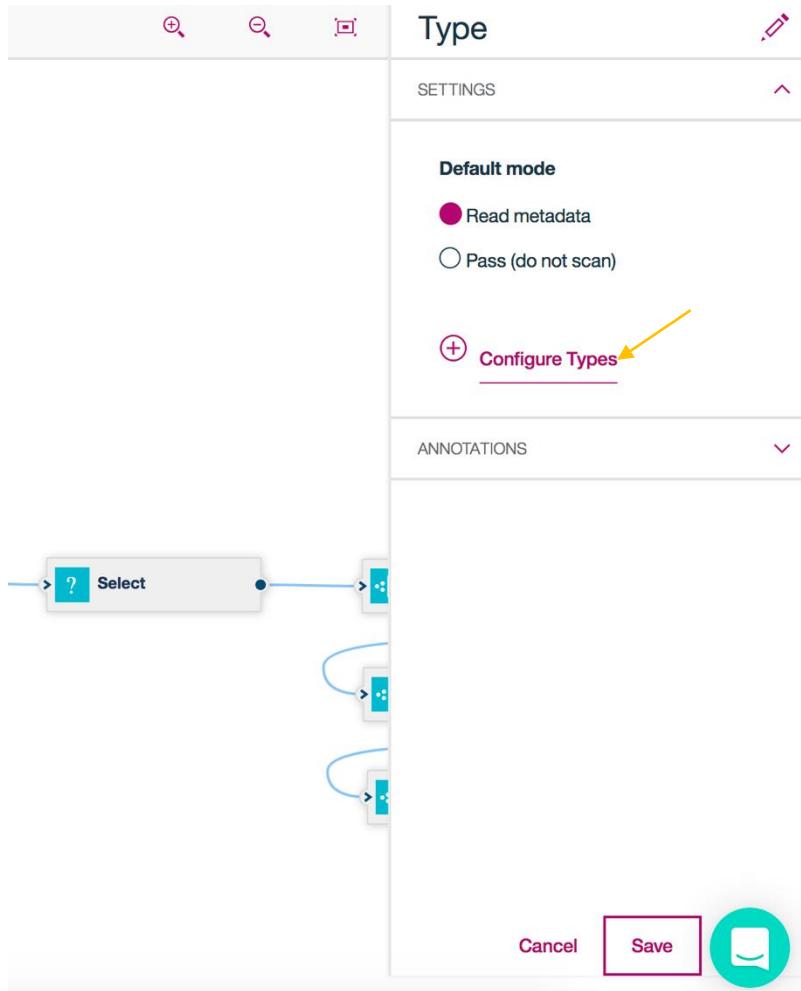
2. Double click on the Partition node. Set the **Training Partition** to 70 and the **Test Partition** to 30. Leave the other defaults, and click on **Save**.



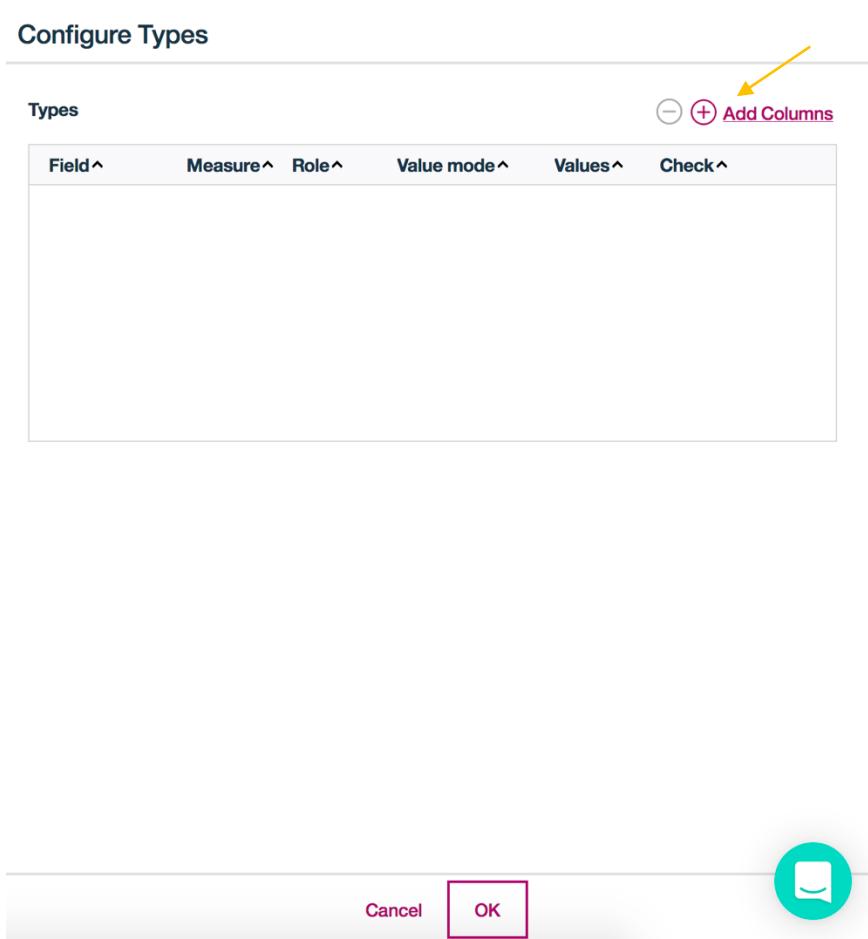
3. Add a **Type** node by clicking on the **Field Operations** in the Node palette and dragging the **Type** node onto the canvas above the **Partition** node. Connect the **Partition** node to the **Type** node. The canvas should appear as below.



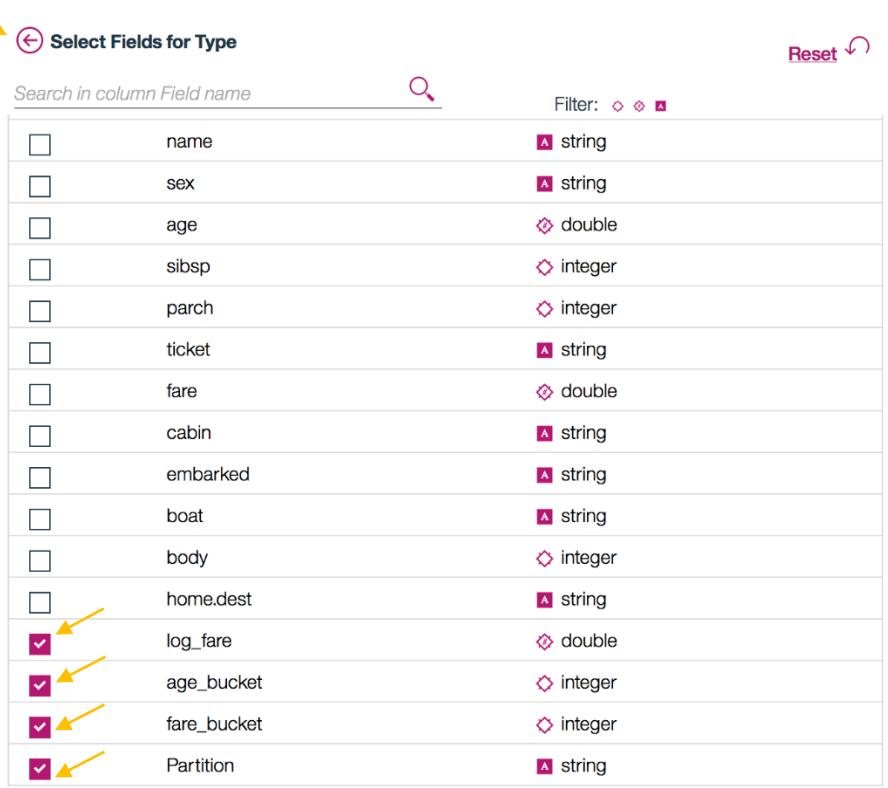
4. Double click on the **Type** node. Click on **Configure Types**.



5. Click on **Add Columns**.



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

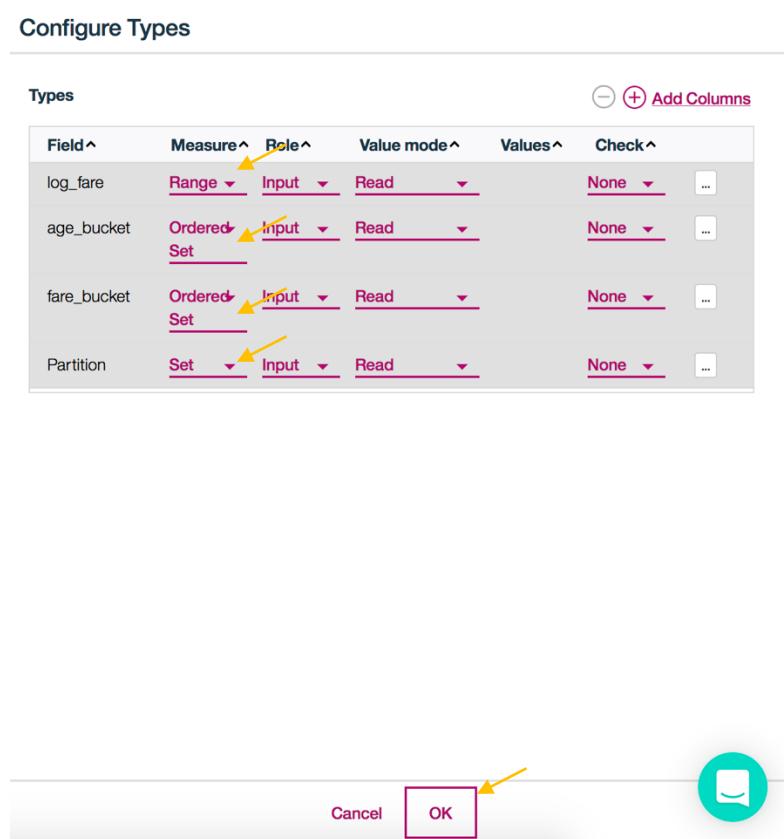


The screenshot shows a table titled "Select Fields for Type". The columns are "Field name" and "Type". A search bar at the top left and a "Reset" button at the top right are also visible.

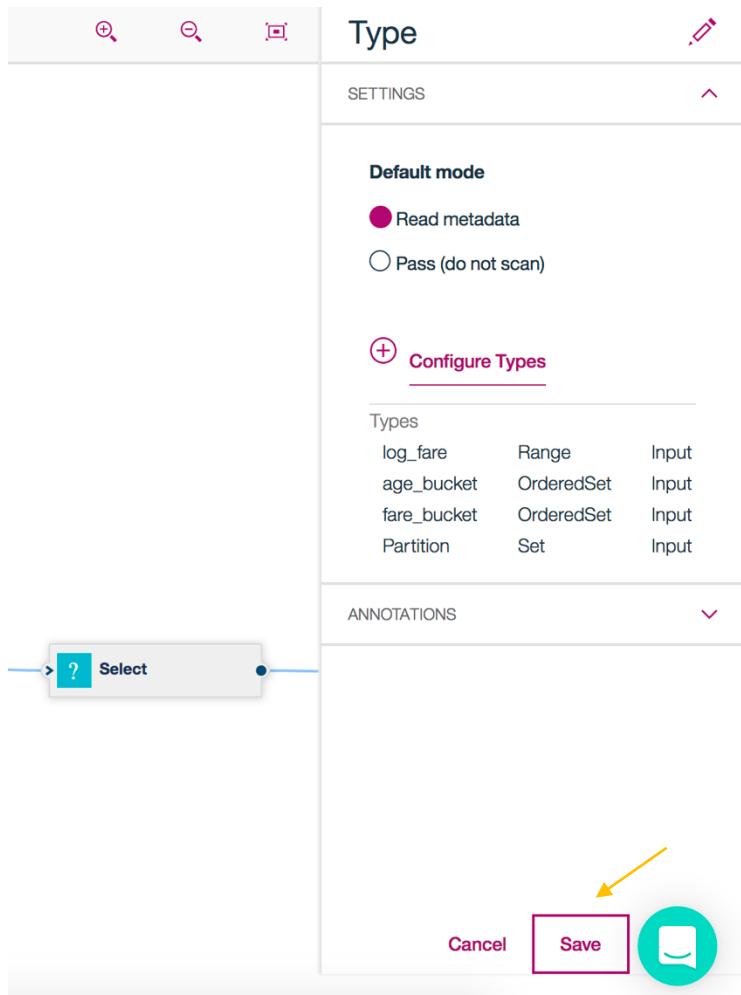
Field name	Type
<input type="checkbox"/> name	string
<input type="checkbox"/> sex	string
<input type="checkbox"/> age	double
<input type="checkbox"/> sibsp	integer
<input type="checkbox"/> parch	integer
<input type="checkbox"/> ticket	string
<input type="checkbox"/> fare	double
<input type="checkbox"/> cabin	string
<input type="checkbox"/> embarked	string
<input type="checkbox"/> boat	string
<input type="checkbox"/> body	integer
<input type="checkbox"/> home.dest	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

Yellow arrows point to the checkboxes for "log_fare", "age_bucket", "fare_bucket", and "Partition".

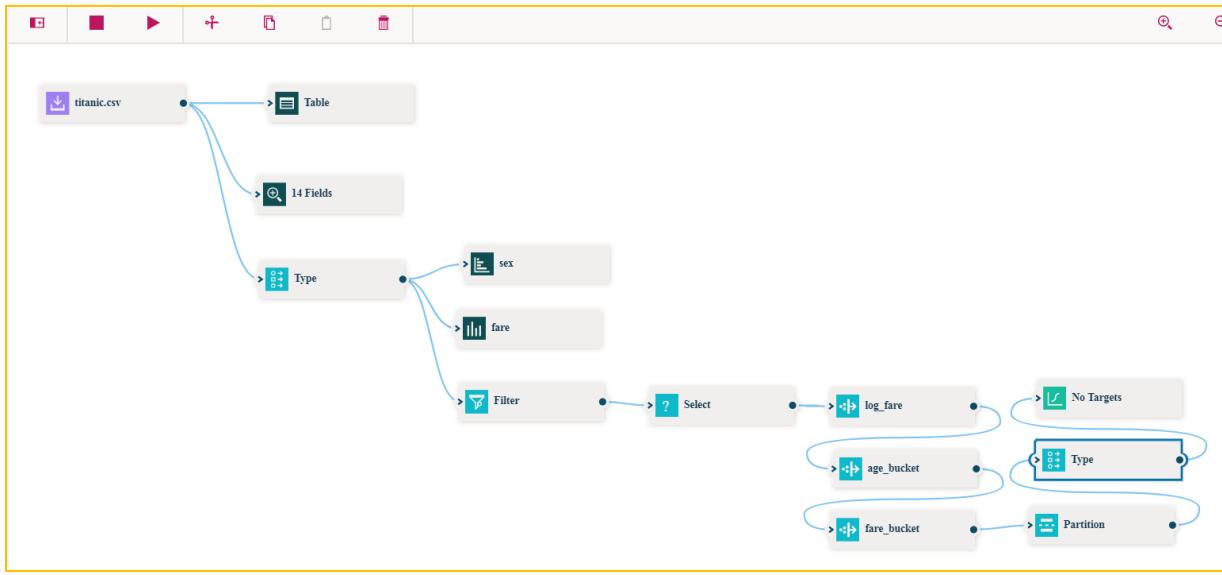
7. For the **Partition** field, select **Set** for the **Measurement**. For the `log_fare`, select **Range** for the **Measurement**. For the `fare_bucket` field, select **OrderedSet** for the **Measurement**, and for the `age_bucket`, select **OrderedSet** for the **Measurement**, and click **OK**.



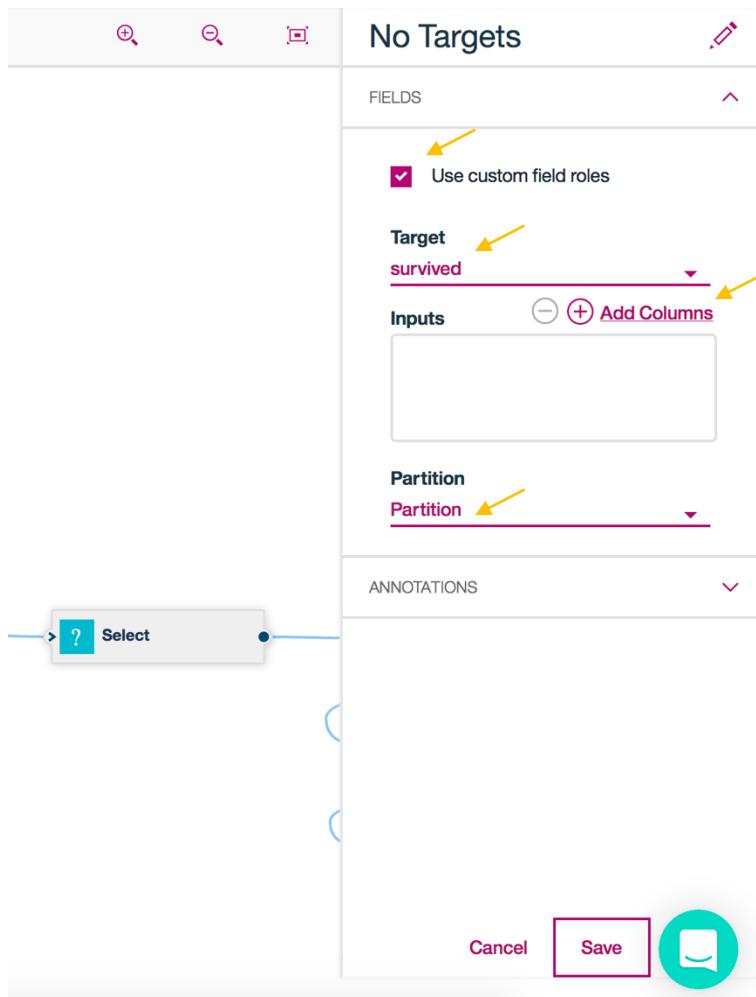
8. Click on **Save**



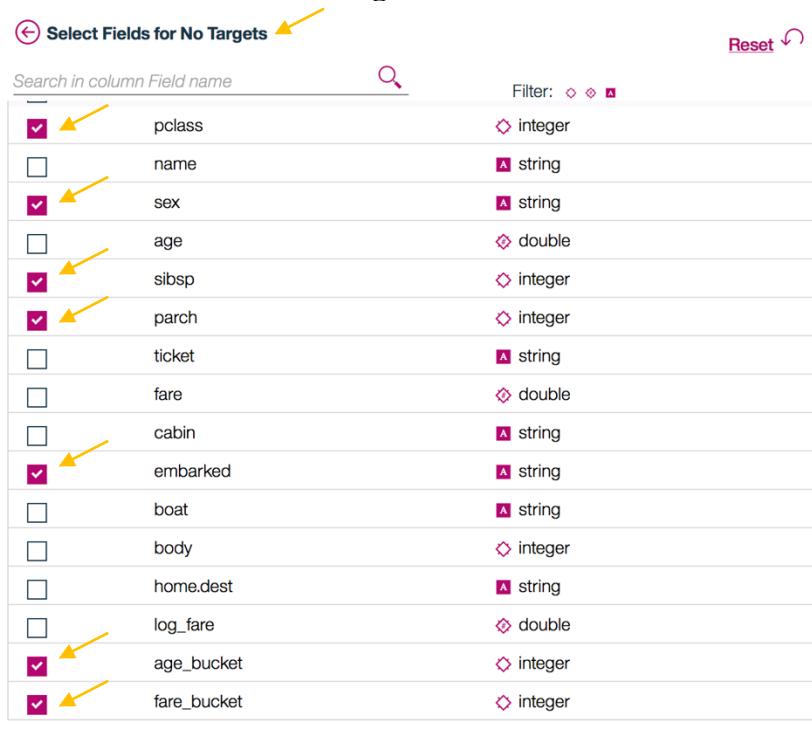
9. Add a **Logistic Regression** node by clicking on the **Modeling** menu item in the Node palette and dragging the **Logistic** node onto the canvas above the **Type** node. Connect the **Type** node to the **Logistic Regression** node. The canvas should appear as below.



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.



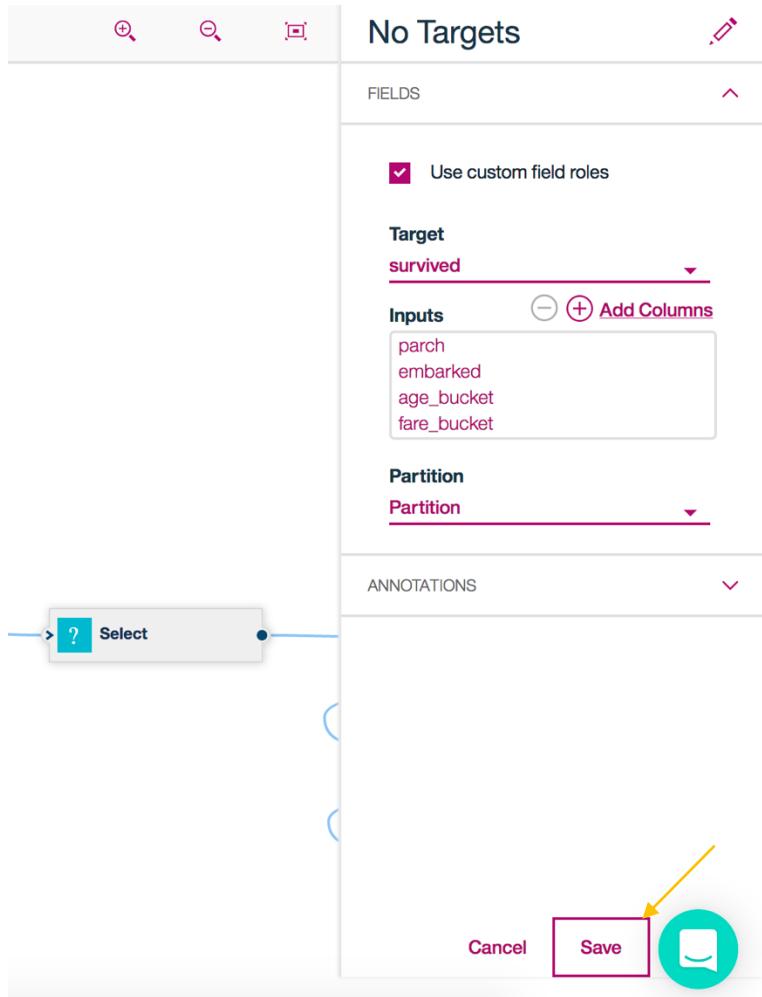
11. Click on the checkboxes next to pclass, sex, sibsp, parch, embarked, age_bucket, fare_bucket fields (you may have to scroll down), and then click the arrow to the left of the **Select Fields for No Targets**.



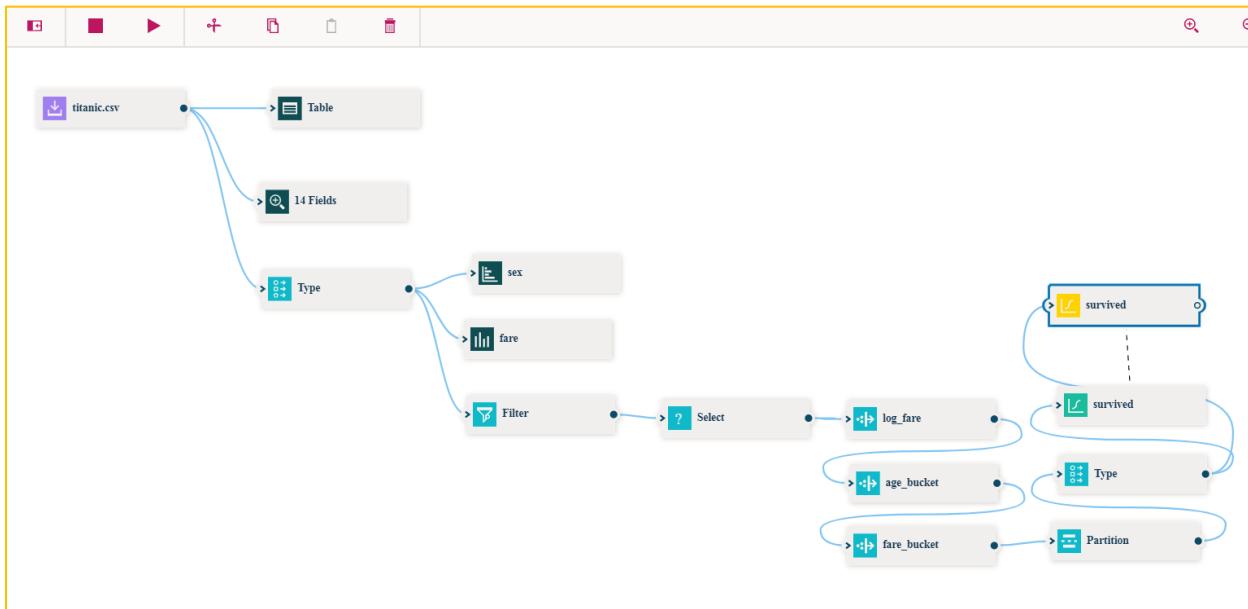
The screenshot shows a table titled "Select Fields for No Targets". The table has two columns: "Field name" and "Type". A search bar at the top left and a "Reset" button at the top right are also visible.

Field name	Type
<input checked="" type="checkbox"/> pclass	integer
<input type="checkbox"/> name	string
<input checked="" type="checkbox"/> sex	string
<input type="checkbox"/> age	double
<input checked="" type="checkbox"/> sibsp	integer
<input checked="" type="checkbox"/> parch	integer
<input type="checkbox"/> ticket	string
<input type="checkbox"/> fare	double
<input type="checkbox"/> cabin	string
<input checked="" type="checkbox"/> embarked	string
<input type="checkbox"/> boat	string
<input type="checkbox"/> body	integer
<input type="checkbox"/> home.dest	string
<input type="checkbox"/> log_fare	double
<input checked="" type="checkbox"/> age_bucket	integer
<input checked="" type="checkbox"/> fare_bucket	integer

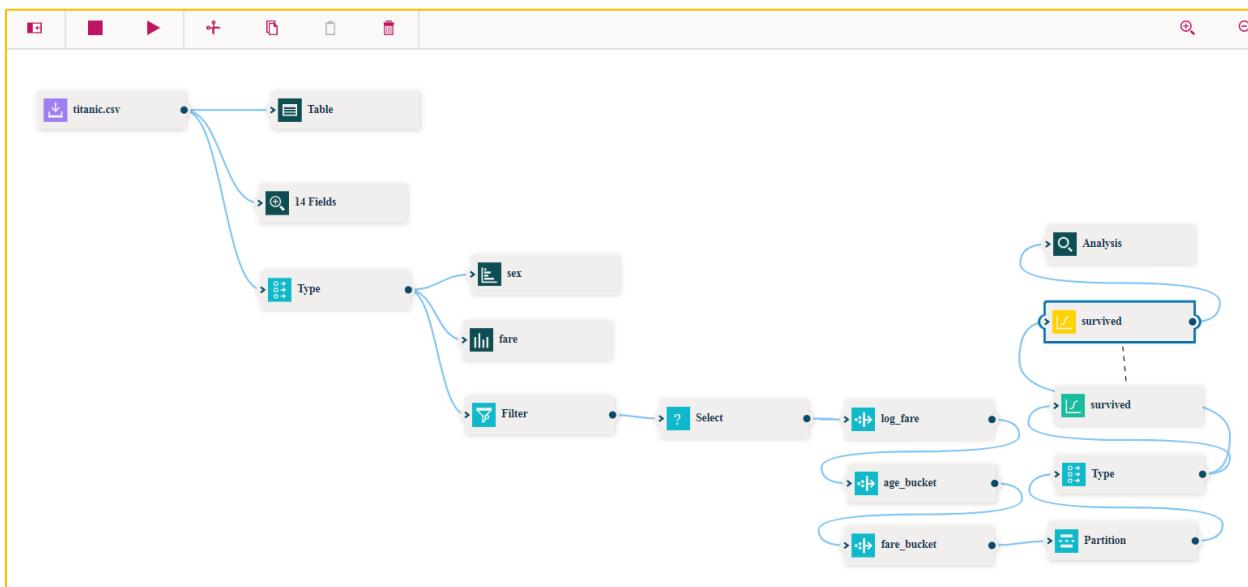
12. Click Save.



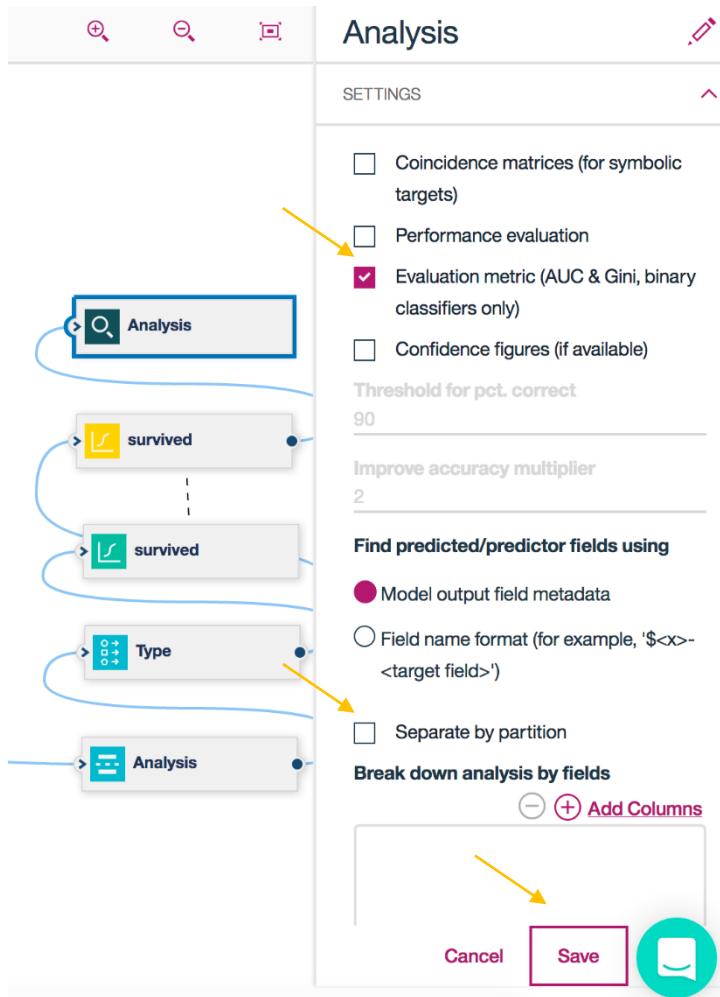
13. Right click on the **Logistic Regression** node and then click **Run**. A **Logistic Regression** “nugget will be created” connected by a dotted line to the **Logistic Regression** node. Drag the nugget and place it above the **Logistic Regression** node. The canvas should appear as below.



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**.



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.

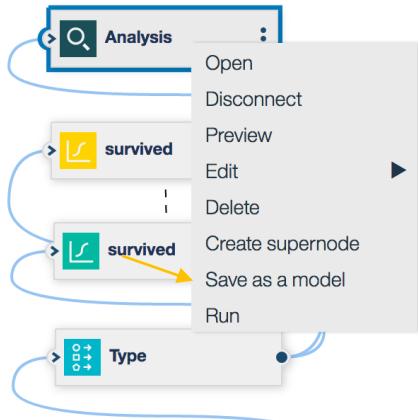
Results for output field survived			
Individual Models			
Comparing \$L-survived with survived			
Correct	828	79.39%	
Wrong	215	20.61%	
Total	1,043		

Evaluation Metrics			
Model	AUC	Gini	
\$L-survived	0.857	0.714	

Step 3.6 Saving a Model

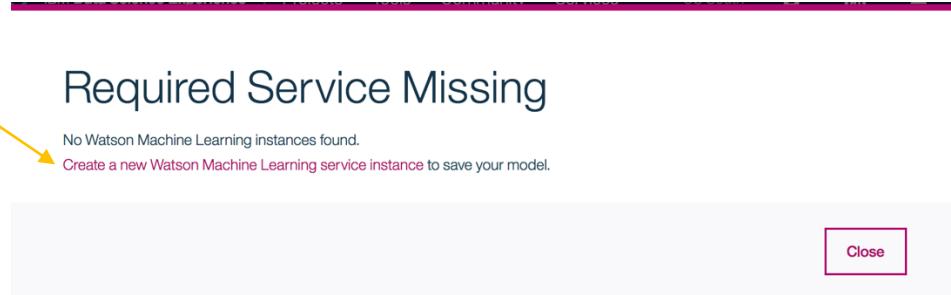
Now that we have created and evaluated a model, we will save the model as an asset. This saved model can be deployed at a future date, removing the need to recreate the same model from scratch.

1. Right click on the Analysis node and then click on **Save 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.”



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

The screenshot shows the IBM Data Science Experience interface. At the top, there are sections for SPSS analytics platform features, Spark and Python Machine Learning features, and Integration with Data Science Experience. Below this, a table lists three plans: Lite, Standard, and Professional. The Lite plan is selected, indicated by a yellow arrow pointing to the radio button. The table columns are Plan, Features, and Pricing. The Lite plan details are: Service instance (5 models per instance), 5,000 predictions, 5 compute hours, and Free. A note below states: "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." The Standard and Professional plans are also listed with their respective details. At the bottom right, there are buttons for Cancel and Create, with a yellow arrow pointing to the Create button.

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

Confirm Creation

Organization: michael.cronk_organization1

Plan

Lite

Space

space1

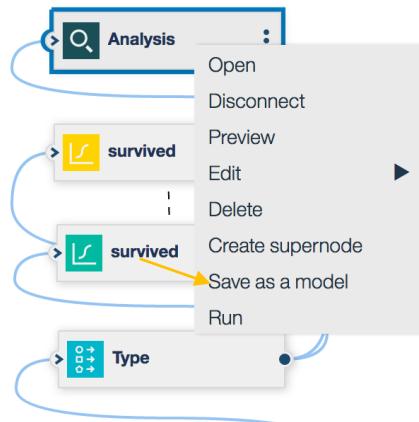
Service name

predictive-modeling-fj

Cancel

Confirm

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**.

The screenshot shows the 'Save Model' dialog box from IBM Data Science Experience. The 'Model Name' field is filled with 'TitanicSPSS'. The 'Machine Learning Service' dropdown is set to 'predictive-modeling-fj'. A note at the bottom states: '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.' The 'Save' button is highlighted with a yellow arrow.

IBM Data Science Experience | Projects Tools Community Services

My Projects / Titanic-SPSS-Lab / Titanic-Flow / Save

US South

Save Model

Terminal node
Analysis

Model Name
TitanicSPSS

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.

Cancel Save

3. Click **Close**.

Saving model

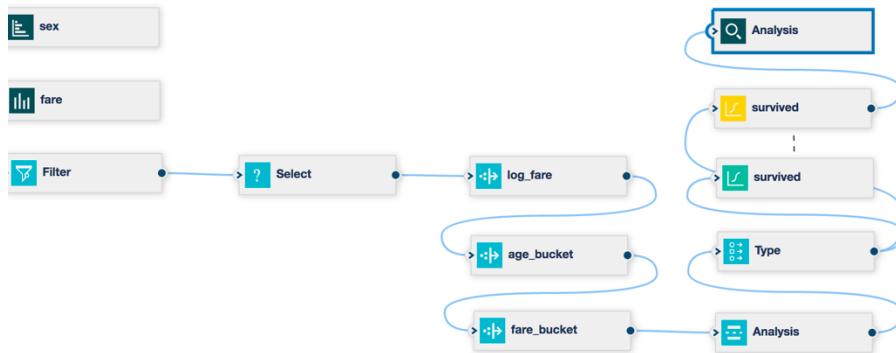
Save completed successfully

Close

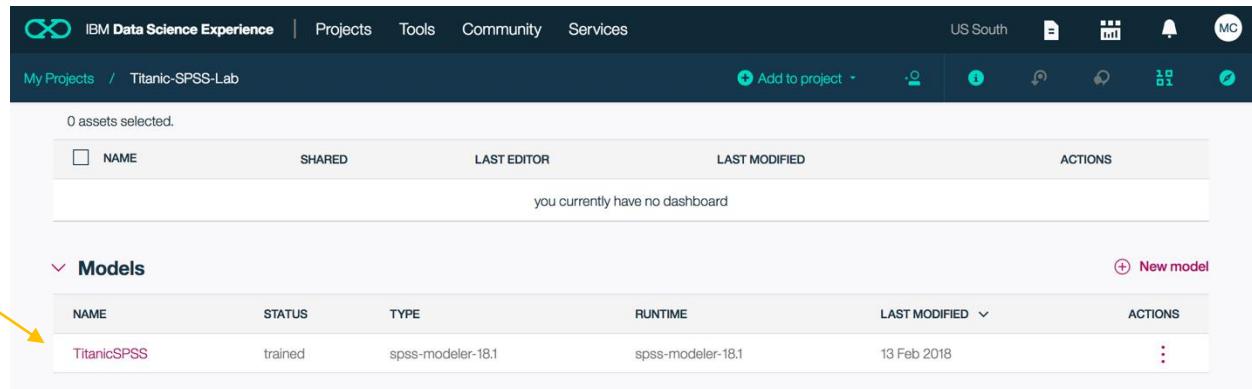


4. Navigate to your project “assets” page. In this example, click on **Titanic-SPSS-Lab**.

The screenshot shows the IBM Data Science Experience interface. At the top, there's a navigation bar with the IBM logo, 'Data Science Experience', 'Projects', 'Tools', 'Community', and 'Services'. Below the navigation bar, the 'US South' region is selected. On the left, a sidebar shows 'My Projects' with 'Titanic-SPSS-Lab' highlighted. The main area displays a dashboard with various icons and a search bar.



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.



The screenshot shows the IBM Data Science Experience interface. At the top, there is a navigation bar with links for Projects, Tools, Community, and Services. On the right side of the header, there are icons for US South, a file, a dashboard, a bell, and a user profile. Below the header, the page title is "My Projects / Titanic-SPSS-Lab". A message indicates "0 assets selected." There is a table with columns: NAME, SHARED, LAST EDITOR, LAST MODIFIED, and ACTIONS. A note below the table says "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	⋮

A yellow arrow points from the left towards the "Models" section of the table.