DSX Local SPSS Modeler Overview

Overview

In this lab you will learn how to implement analytics in **SPSS Modeler**, a well-known visual data mining workbench which can be used in the **Data Science Experience** (DSX). 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. You can learn more about Modeler in DSX in official product documentation: *https://content-dsxlocal.mybluemix.net/docs/content/local-dev/spss-modeler.html*

Introduction

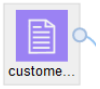
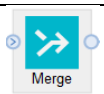
SPSS Modeler is a visual data mining workbench. Modeler can be used to complete all tasks of the analytic application development

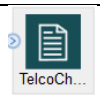
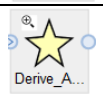
- Data understanding
- Data preparation
- Model building
- Model evaluation

Deployment is done in DSX **Deployment Manager**. We discuss flow deployment in more detail in the **DSX Deployment** hands-on lab.

Assets developed in Modeler are called "flows". Another frequently used term in Modeler documentation is "streams" (used in Modeler desktop documentation). A flow starts with one or several data sources. Using visual nodes, a user can apply different operations to data. Data "flows" from one node to another in the direction of the arrows.

Visual nodes in modeler are color-coded and organized by type of operation: **Record Operations**, **Field Operations, Graphs, Modeling, Output,** and **Export** (data sources). Most operations are well-known functions in data preparation and analytics, such as sampling, filtering, binning, etc.

The data sources are purple	
Data preparation operations are blue	
Algorithms are green	
The models that are created based on algorithms are orange	

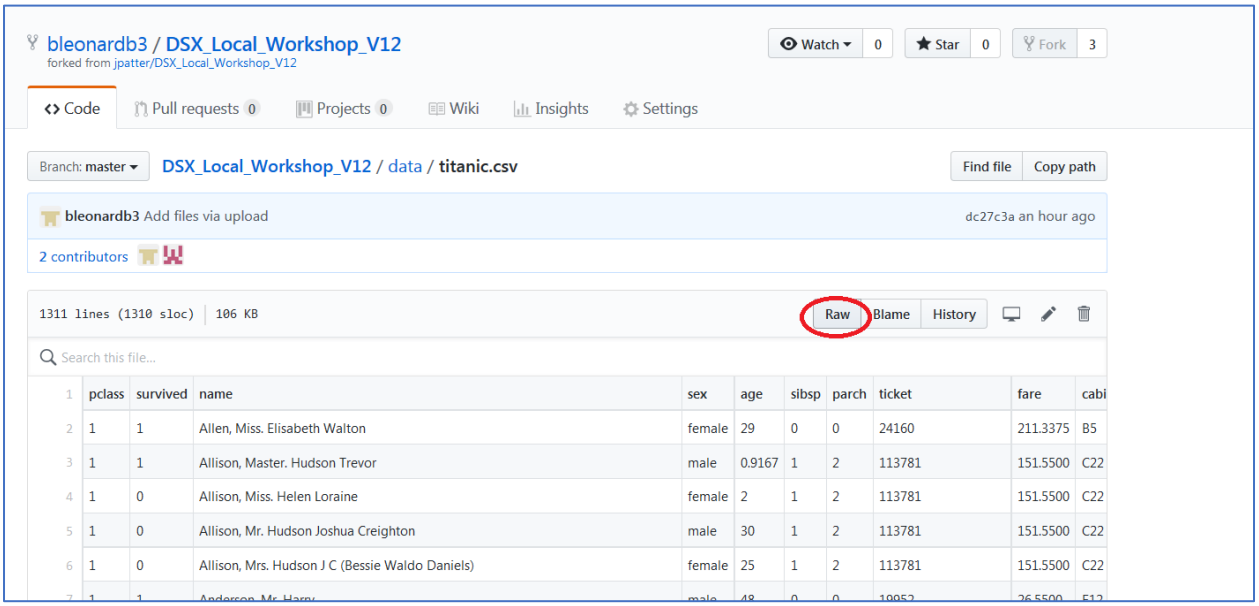
Different types of output (graphs, tables, external files) are black	
The nodes with a star icon are called "supernodes" because they contain several nodes. Supernodes are used for visual organization of the flow.	

If a user needs more information about a particular node, it can be looked up in Modeler documentation. SPSS also publishes the **Algorithms Guide** that explains how machine learning algorithms are implemented in Modeler (see **Reference** for more information).

Lab Steps

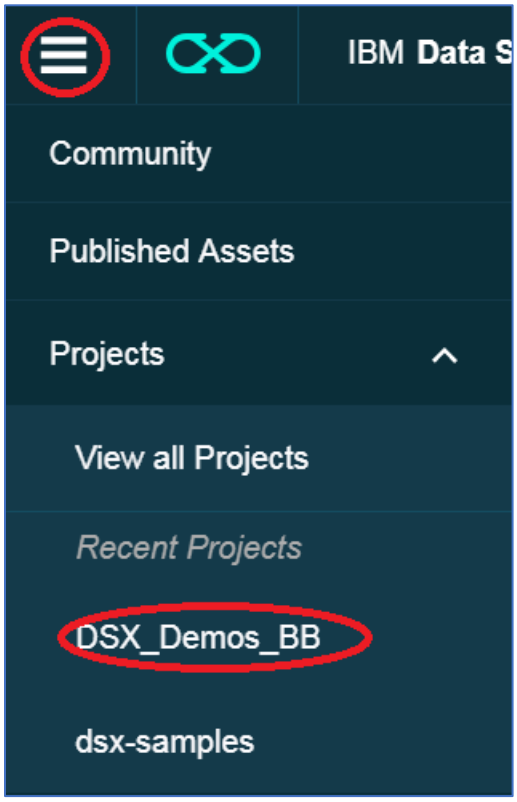
Step 1: Adding a Data Asset to the DSX Local Labs project

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



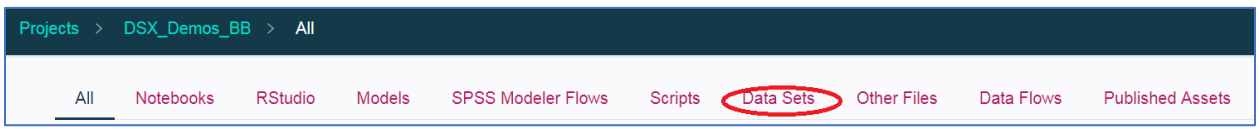
Right click on Raw, and click on Save link as ….

2. Go to the DSX Local project. Click on the hamburger icon , and then click on Projects and then the DSX_Demos_XX project.

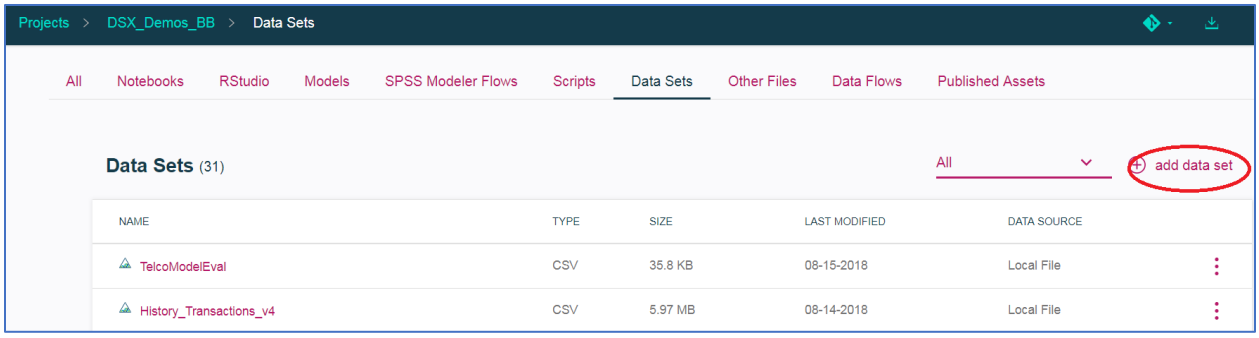


3. Click on **Assets.**

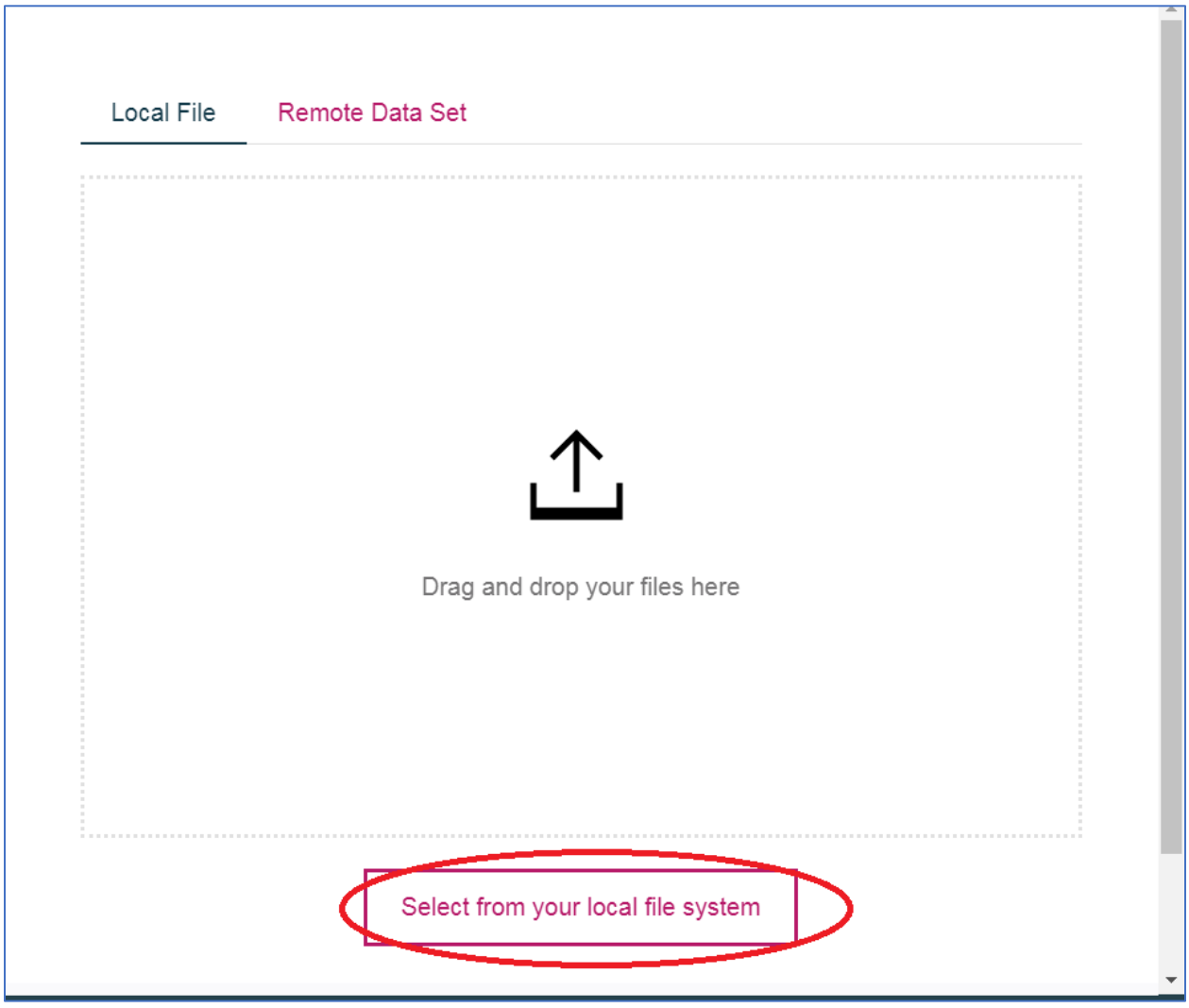
4. Click on **Data Sets**.



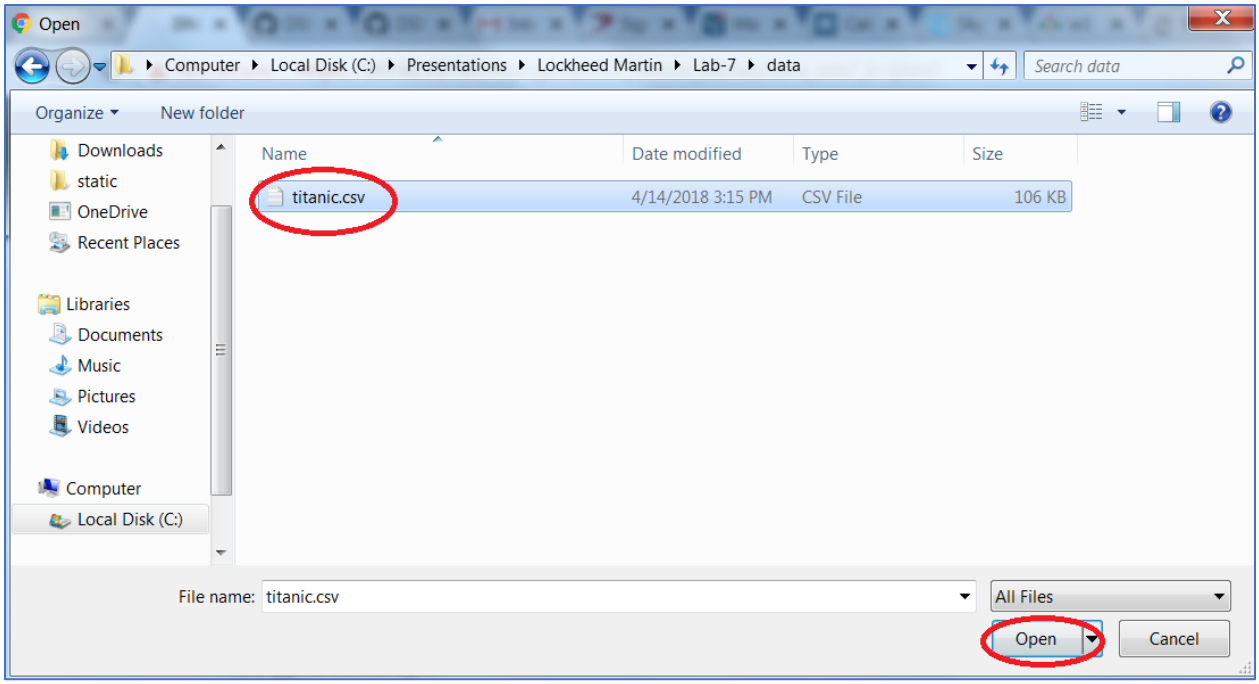
5. Click on **add data sets**.



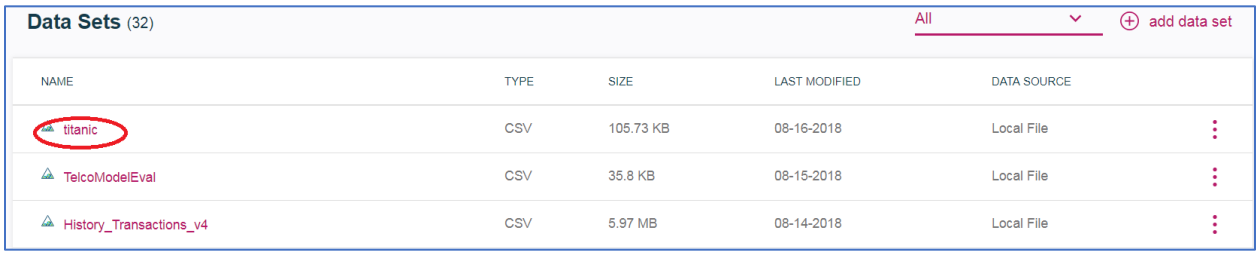
6. Click on **Select from your local file system**.



7. Navigate to the place where the **titanic.csv** was downloaded. Select the **titanic.csv** file and click **Open**.



8. The titanic.csv file is loaded into the project data sets.



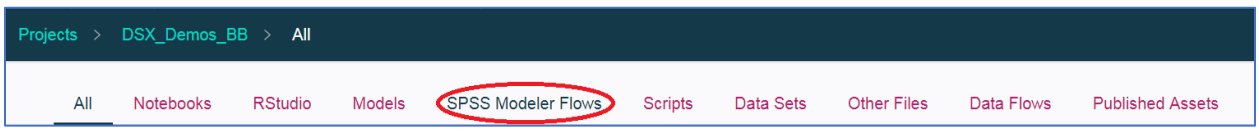
Step 2: Create a Model to predict survival

In this section, we will create a Machine Learning flow using SPSS nodes. Documentation describing the nodes is available *at https://content-dsxlocal.mybluemix.net/docs/content/local-dev/spss-modeler.html*.

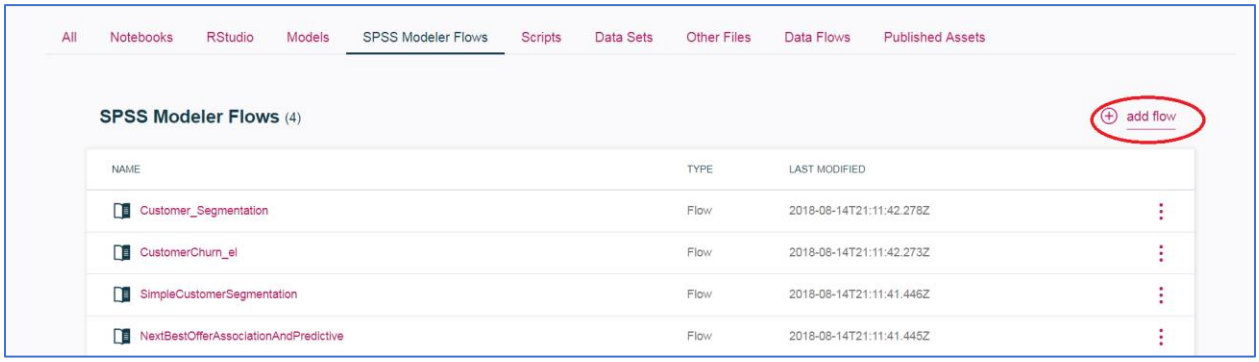
.

Step 2.1 Create a New Flow and Load the Data

1. In the DSX Local project, click on **SPSS Modeler Flows**

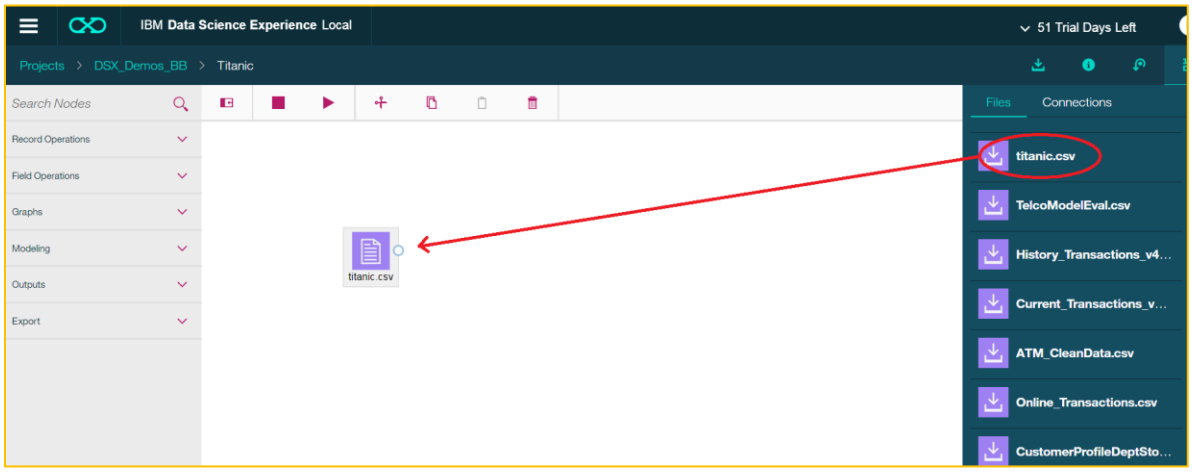


2. Click on **add flow**.

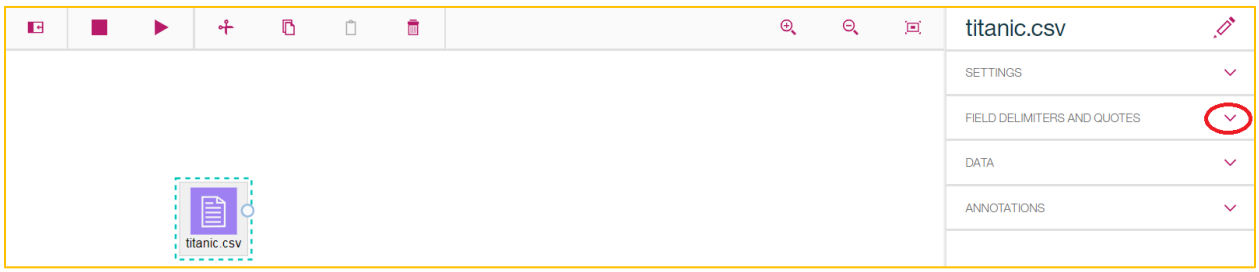


3. Click on the **Blank** tab, enter Titanic for the **Name,** optionally enter a **Description**, and click **Create**.

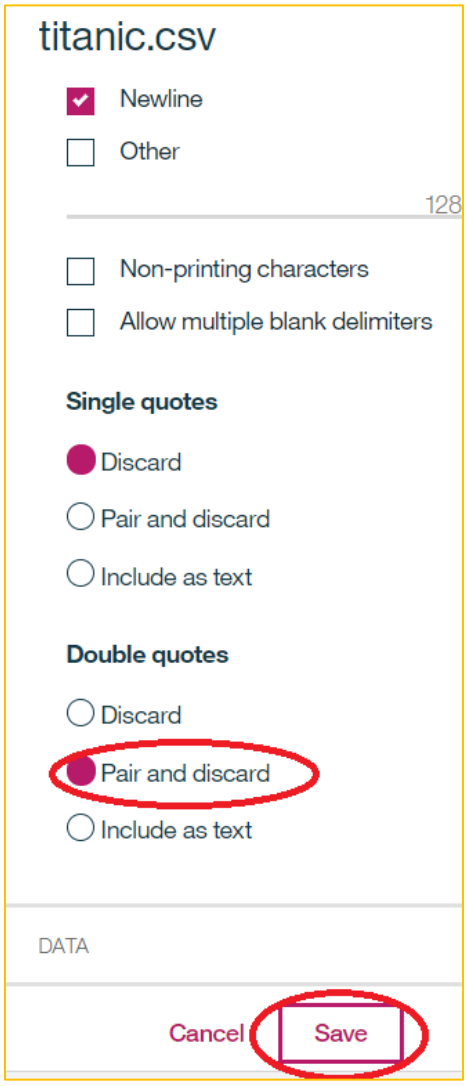
4. This opens the Flow Editor. Click on the  , and then click on the **titanic.csv** entry, and hold down the left mouse key to drag it onto the canvas. Release the left mouse key.



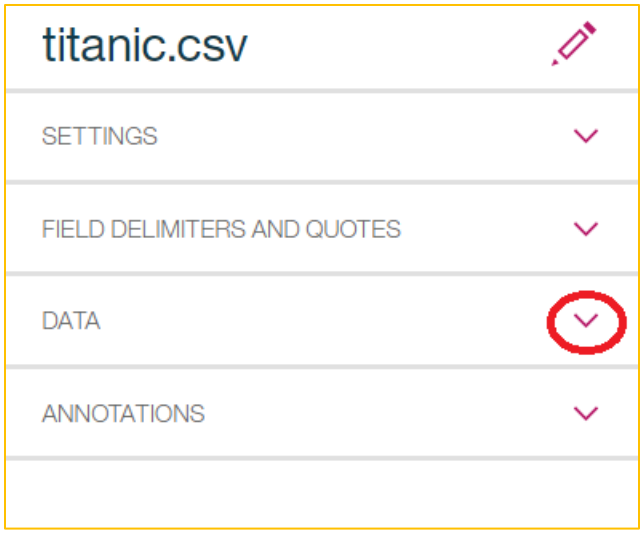
5. Double click on the titanic.csv icon. This will bring up the settings panel on the right hand side. Click on the down arrow next to **FIELD DELIMITERS AND QUOTES**.



6. Scroll down and click on **Pair and discard** under Double quotes, and then click **Save**.

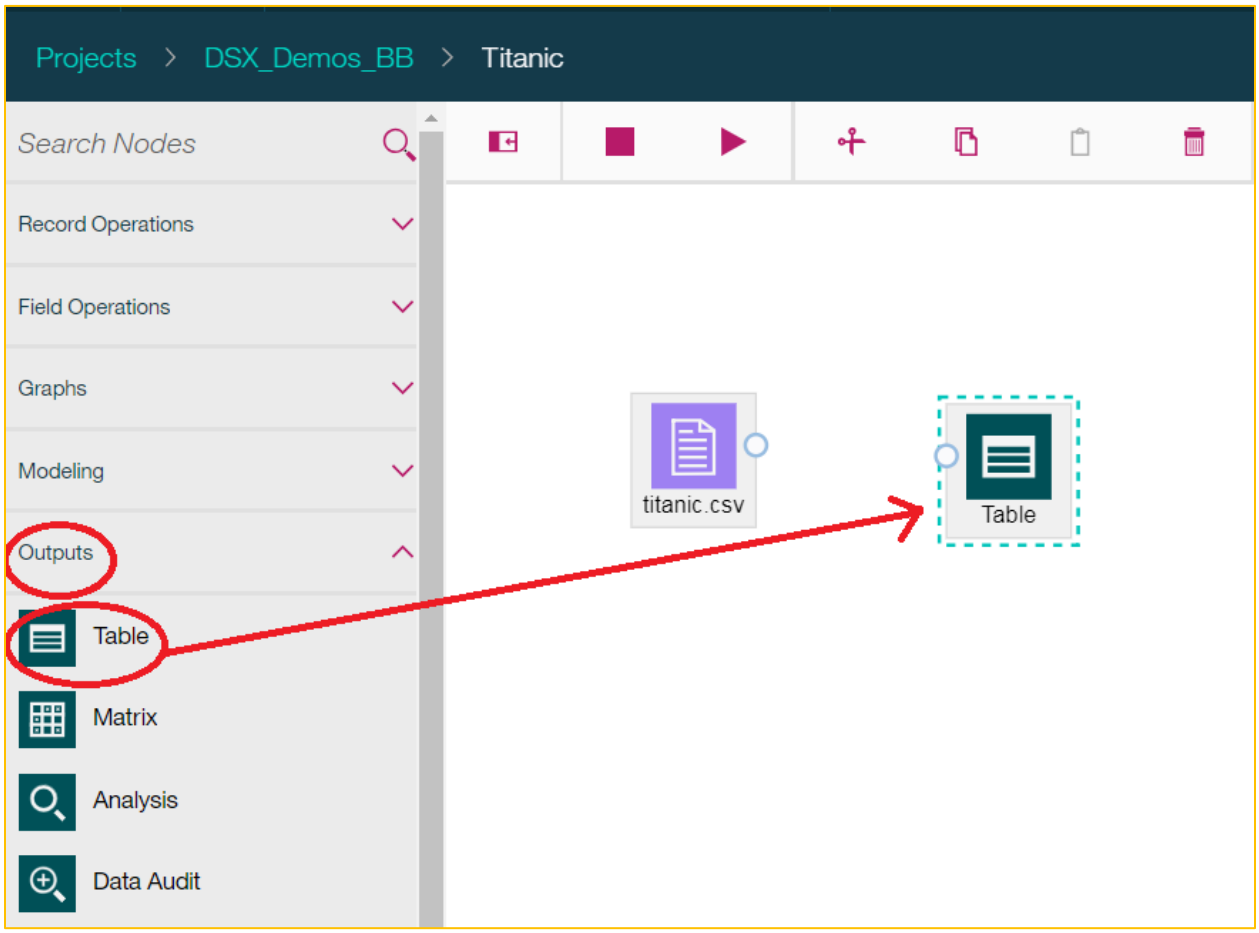


7. Double Click on the **titanic.csv** icon again. This time select the down arrow next to **Data**.

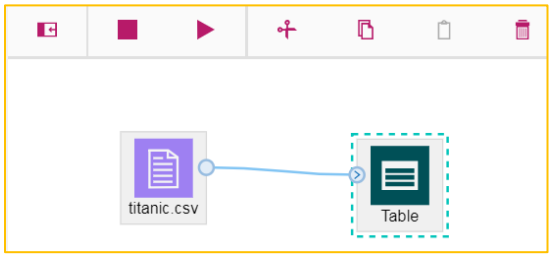


8. Override the pclass and survived storage classes, by clicking in the check box and changing the storage class from Integer to String.

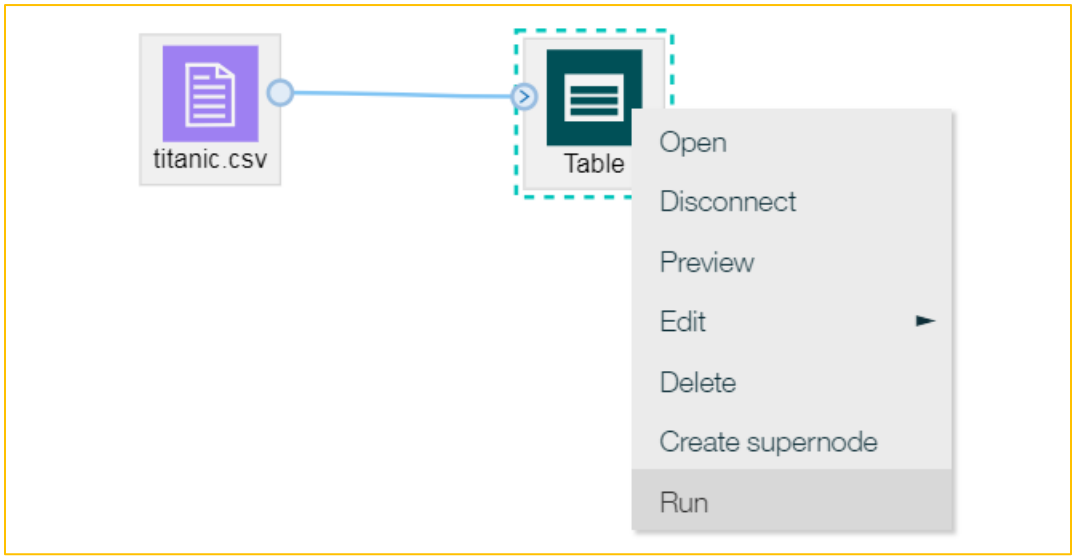
9. Click on the **Outputs** menu item in the Node Palette on the left and then click on the **Table** icon and drag the icon to the right of the titanic.csv icon. The SPSS Table node will display the contents of the csv file. If the Node Palette is not visible, click on the Node Palette icon 



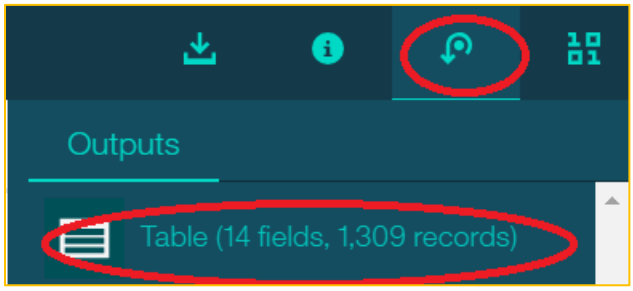
10. Connect the right side of the titanic.csv icon to the left side of the Table icon. This is accomplished by clicking on the little circle at the right side of the titanic.csv icon holding the left mouse key and dragging the mouse to the little circle on the left side of the Table icon, and then releasing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



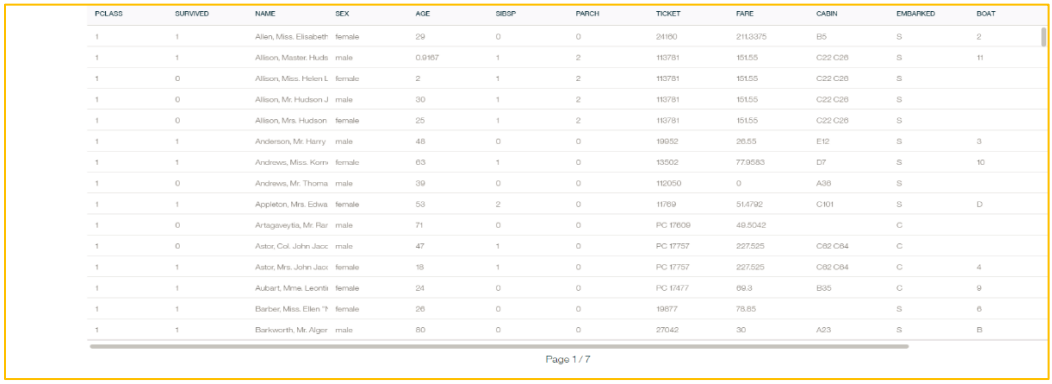
11. Right click on the **Table** icon, and select **Run**.



12. The "Running Flow" prompt may appear and 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.



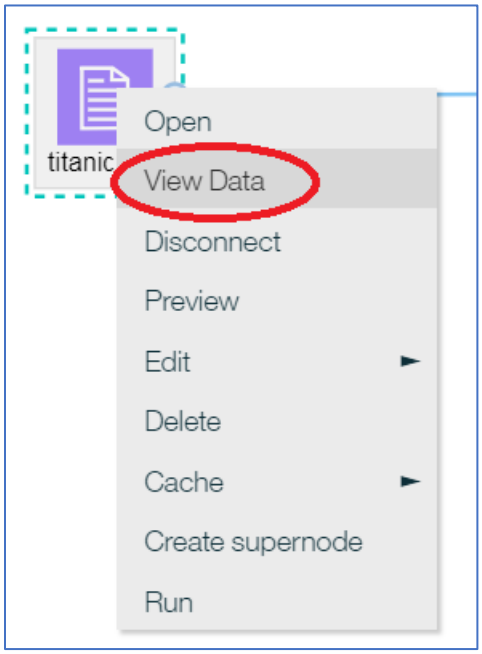
13. Double click on the Table selection and the contents of the titanic.csv will be displayed. Each row contains information on a passenger on the Titanic. We will use this data to make predictions on survivability.



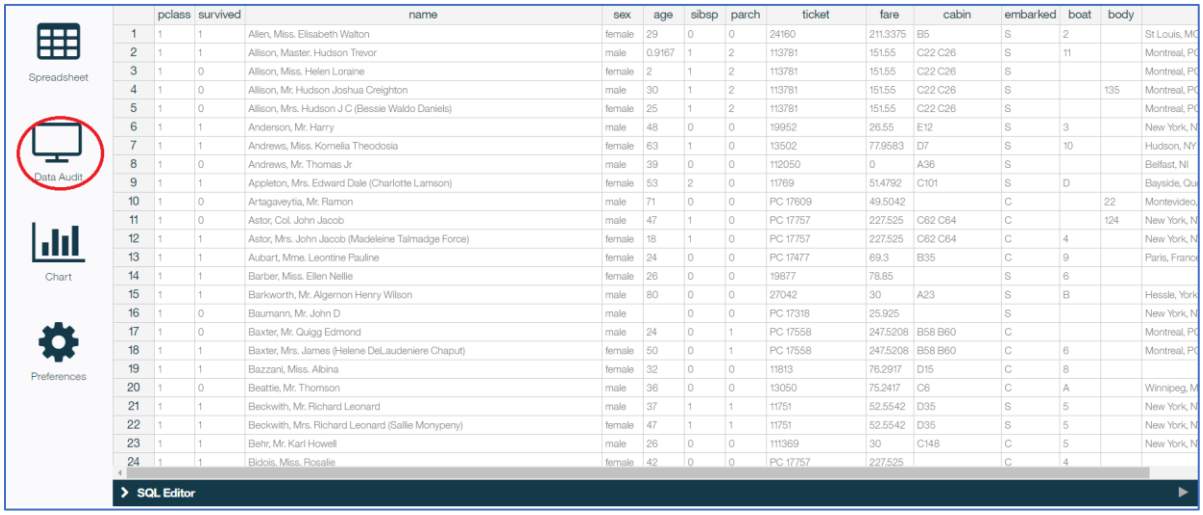
Step 2.2 Explore the Data using the View Data – Data Audit option.

Perusing through the data in the table, we can see that there are missing values. The SPSS Modeler has a View Data option that provides multiple ways to explore the data. The View Data option combines the capabilities of several SPSS modeler nodes (Table node, Data Audit node, and Visualization nodes) in one place.

1. Right-click on the **titanic.csv** node and click on **View Data**.

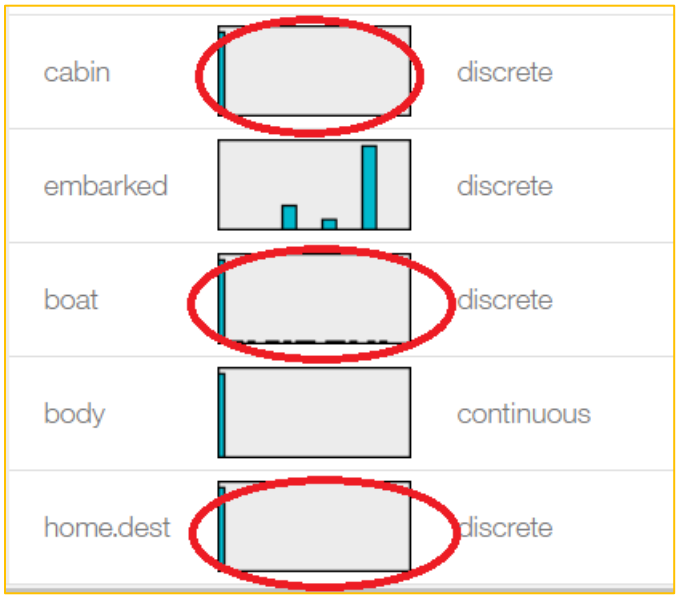


2. The **View Data** results are shown below. Four options appear at the left. The Spreadsheet option is similar to the Table node, the Data Audit option is similar to the Data Audit node, and the Chart option provides similar capabilities to the Graph nodes. The Preferences option provides different charting themes. Click on the **Data Audit** option.

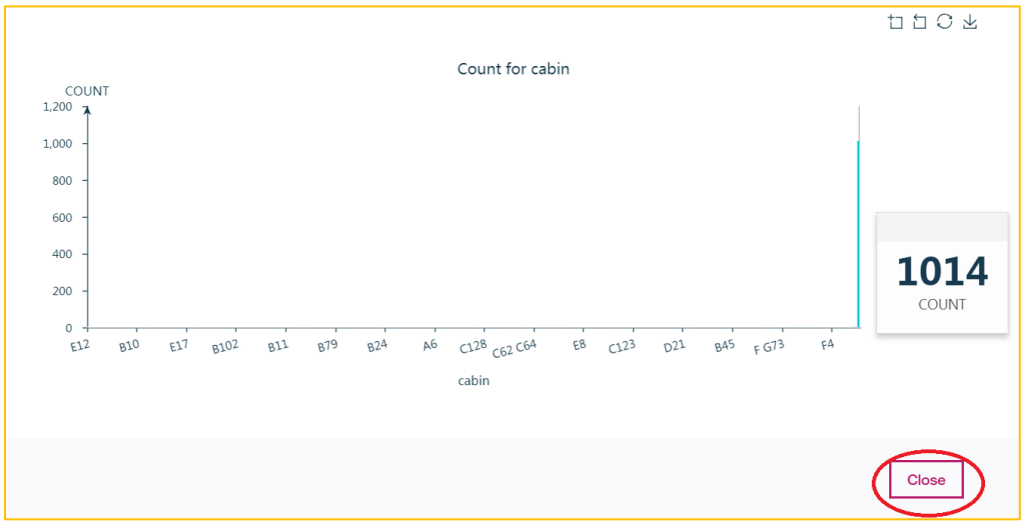


	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	
1	1	1	Allen, Miss. Elisabeth Walton	female	29	0	0	24160	211.3375	B5	S	2		St Louis, MO
2	1	1	Allison, Master. Hudson Trevor	male	0.9167	1	2	113781	151.55	C22 C26	S	11		Montreal, PQ
3	1	0	Allison, Miss. Helen Loraine	female	2	1	2	113781	151.55	C22 C26	S			Montreal, PQ
4	1	0	Allison, Mr. Hudson Joshua Creighton	male	30	1	2	113781	151.55	C22 C26	S		135	Montreal, PQ
5	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25	1	2	113781	151.55	C22 C26	S			Montreal, PQ
6	1	1	Anderson, Mr. Harry	male	48	0	0	19952	26.55	E12	S	3		New York, NY
7	1	1	Andrews, Miss. Kornelia Theodosia	female	63	1	0	13502	77.9583	D7	S	10		Hudson, NY
8	1	0	Andrews, Mr. Thomas Jr	male	39	0	0	112050	0	A36	S			Belfast, NI
9	1	1	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53	2	0	11769	51.4792	C101	S	D		Bayside, Qu
10	1	0	Artagaveytia, Mr. Ramon	male	71	0	0	PC 17609	49.5042		C		22	Montevideo,
11	1	0	Astor, Col. John Jacob	male	47	1	0	PC 17757	227.525	C62 C64	C		124	New York, NY
12	1	1	Astor, Mrs. John Jacob (Madeleine Talmadge Force)	female	18	1	0	PC 17757	227.525	C62 C64	C	4		New York, NY
13	1	1	Aubart, Mme. Leontine Pauline	female	24	0	0	PC 17477	69.3	B35	C	9		Paris, France
14	1	1	Barber, Miss. Ellen Nellie	female	26	0	0	19877	78.85		S	6		
15	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80	0	0	27042	30	A23	S	B		Hessle, York
16	1	0	Baumann, Mr. John D	male		0	0	PC 17318	25.925		S			New York, NY
17	1	0	Baxter, Mr. Quigg Edmond	male	24	0	1	PC 17558	247.5208	B58 B60	C			Montreal, PQ
18	1	1	Baxter, Mrs. James (Helene DeLaudeniere Chaput)	female	50	0	1	PC 17558	247.5208	B58 B60	C	6		Montreal, PQ
19	1	1	Bazzani, Miss. Albina	female	32	0	0	11813	76.2917	D15	C	8		
20	1	0	Beattie, Mr. Thomson	male	36	0	0	13050	75.2417	C6	C	A		Winnipeg, MB
21	1	1	Beckwith, Mr. Richard Leonard	male	37	1	1	11751	52.5542	D35	S	5		New York, NY
22	1	1	Beckwith, Mrs. Richard Leonard (Sallie Monypeny)	female	47	1	1	11751	52.5542	D35	S	5		New York, NY
23	1	1	Behr, Mr. Karl Howell	male	26	0	0	111369	30	C148	C	5		New York, NY
24	1	1	Bidois, Miss. Rosalie	female	42	0	0	PC 17757	227.525		C	4		

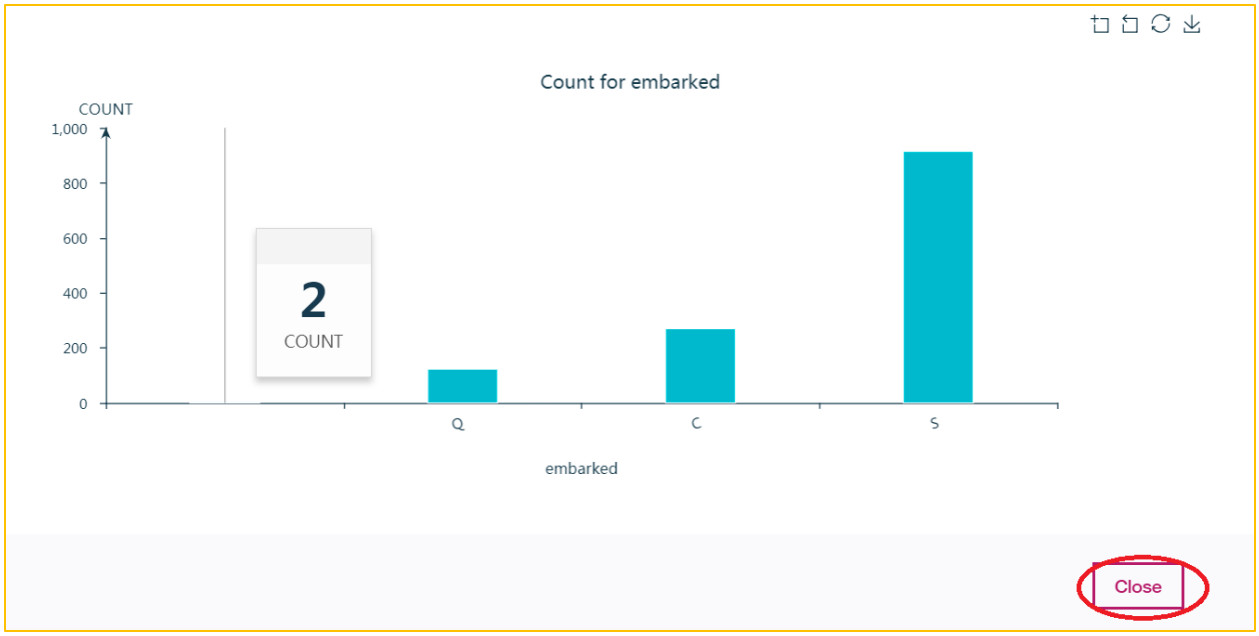
3. Data Audit provides profiling information on the input data that is useful for cleansing the data. It provides a comprehensive first look at the data, including summary statistics, as well as information about outliers, missing values, and extremes. You can click on the thumbnail images to display a larger image of the thumbnail with values. Click on the cabin, boat, and home.dest thumbnails. We can see that each of these contains a significant number of blank or empty string values. These fields will be removed using a **Filter** node below.



4. Clicking on the cabin thumbnail displays a larger image. Hovering the cursor over the spike in the graph, we can see 1014 values out of 1309 are empty or blank strings. Click on **Close** to return to the Audit view.



5. Clicking on the embarked thumbnail icon shows that there are two missing values for this field. Click on **Close** to return to the **Audit** panel.



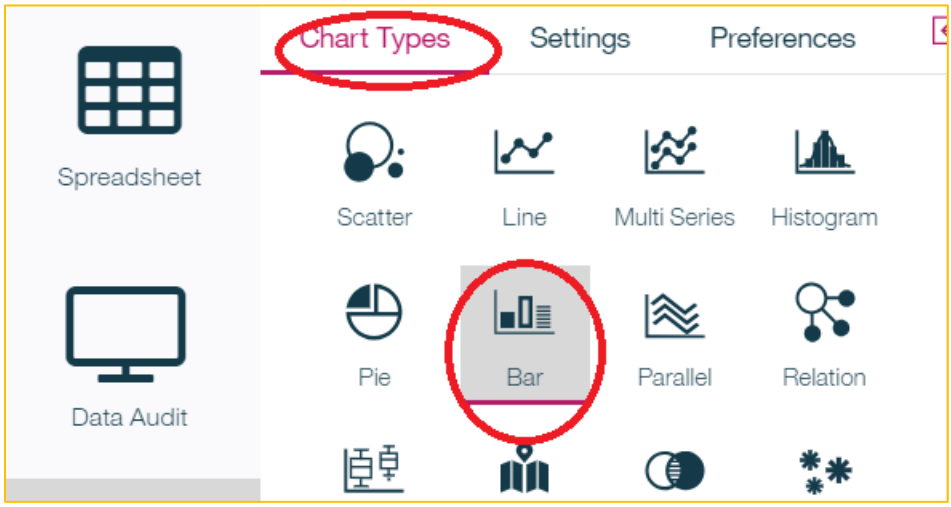
6. Click on the **Quality** tab. The Quality tab provides information on null values for continuous variables and displays outlier and extreme values. Note that quality information on discrete (string) fields is missing from this panel (this information does appear in the Quality information provided by the Data Audit node, so you may consider adding a Data Audit node as well). We see that there are quite a few missing values from the age field, 1 missing value from the fare field, and many missing values in the body field. We will remove the rows containing the missing values from the fare, embarked, and age fields using a **Select** node below. We will drop the body field in addition to the other fields mentioned above using a **Filter** node.

Step 2.3 Explore the Data using the View Data – Chart Option

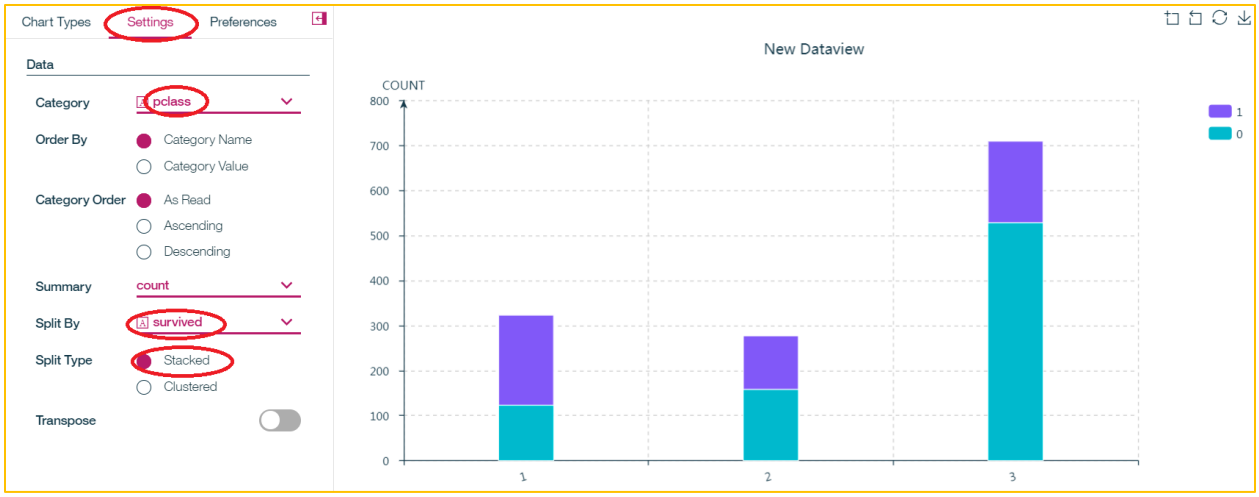
Let's explore the characteristics of the data using the Chart option.

1. Click on the **Chart** option, then click on the palette icon.

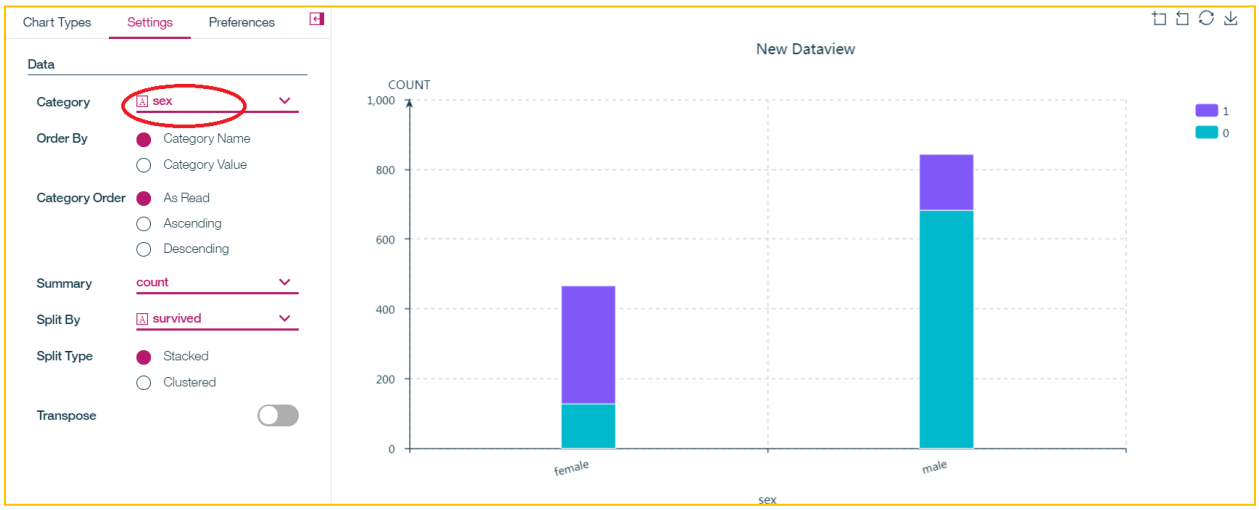
2. Click on **Chart Type**, then click on **Bar.**



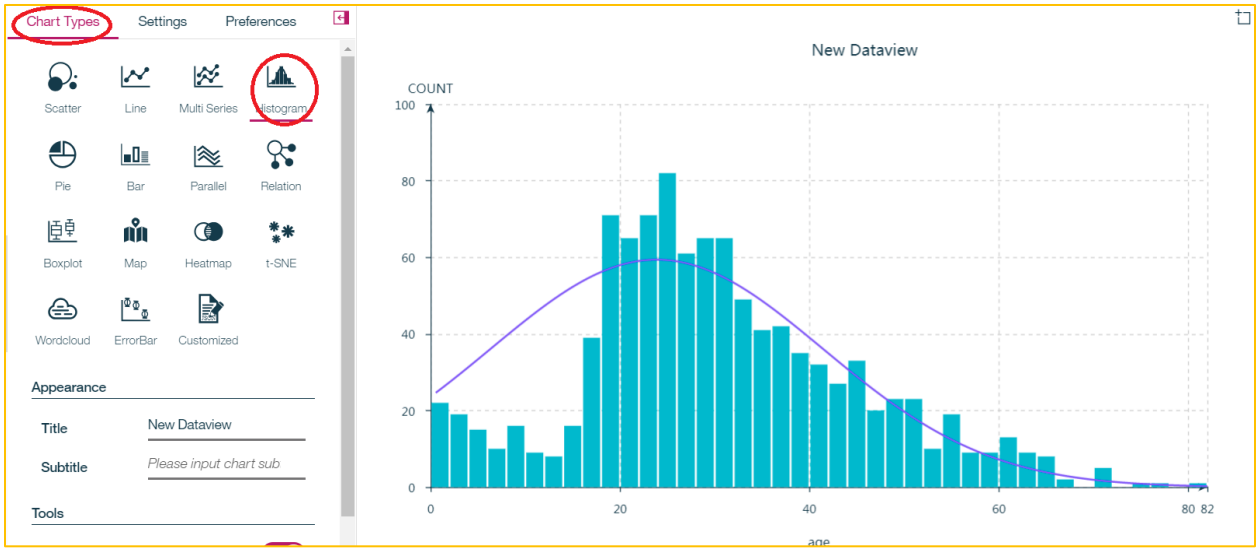
3. Select **Settings**, select pclass for the **Category**, select survived for the **Split By**, and click on **Stacked**. We can see from the graph that the likelihood of surviving is correlated to the passenger class. The first-class passengers have the highest rate of survivability.



4. Change the **Category** to sex. The survivability rate is strongly influenced by the passenger sex.



5. Let's examine the histogram of the age and fare fields. Click on the **Chart Type** and click on **Histogram**. The graph defaults to using age for the histogram. A normal curve is superimposed on the age histogram.



6. Click on **Settings** and select fare in the **Value field**. We can see that the histogram is highly 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.
7. Note the above visualizations can also be done using **Graph** nodes. The **Distribution** node can be used to generate the bar charts above, and the **Histogram** node can be used to generate the histograms. The View Data option is very convenient in that these visualizations can be done in one place.

Step 2.4 Prepare the Data for Modeling

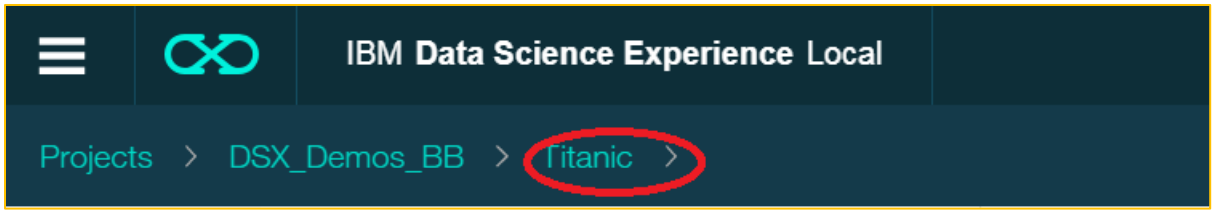
Based on our exploration of the data, there are several transformations that are needed to prepare the data for modeling. This section will introduce, the **Filter** node, the **Select** node, and the **Derive** node that will do the necessary transformations. The **Filter** and **Derive** nodes act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

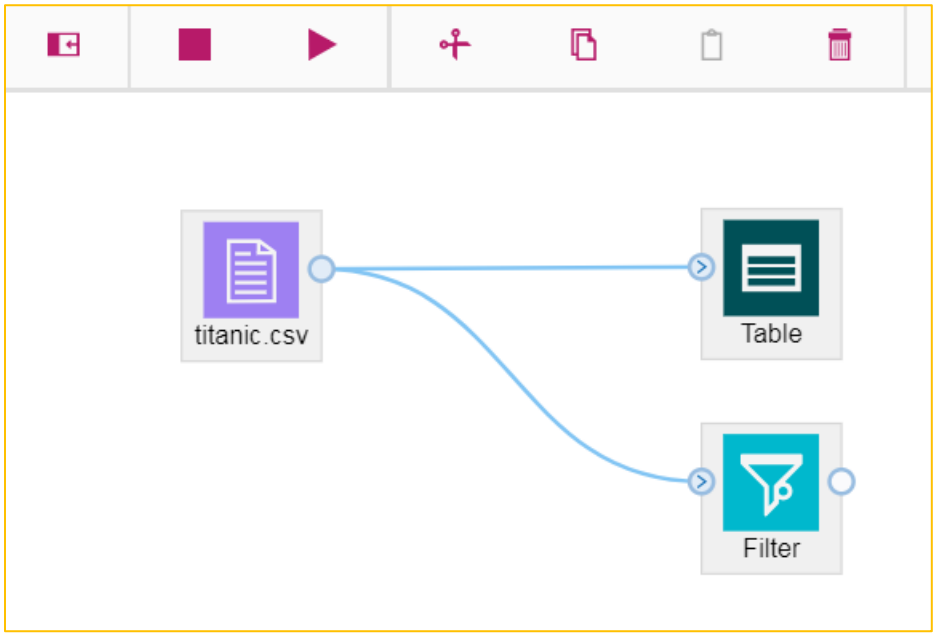
Derive node – The **Derive** node modifies data values or creates new fields from one or more existing fields. We will use the derive node to do a logarithmic transformation of the fare field. We will also use this node to bin the age and fare fields.

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 is missing information in the fare, age, or embarked fields.

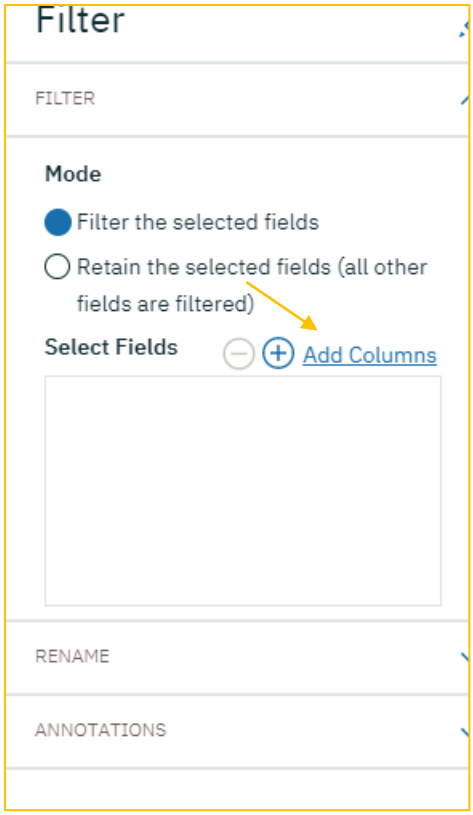
1. Click **Titanic** to return to the canvas.



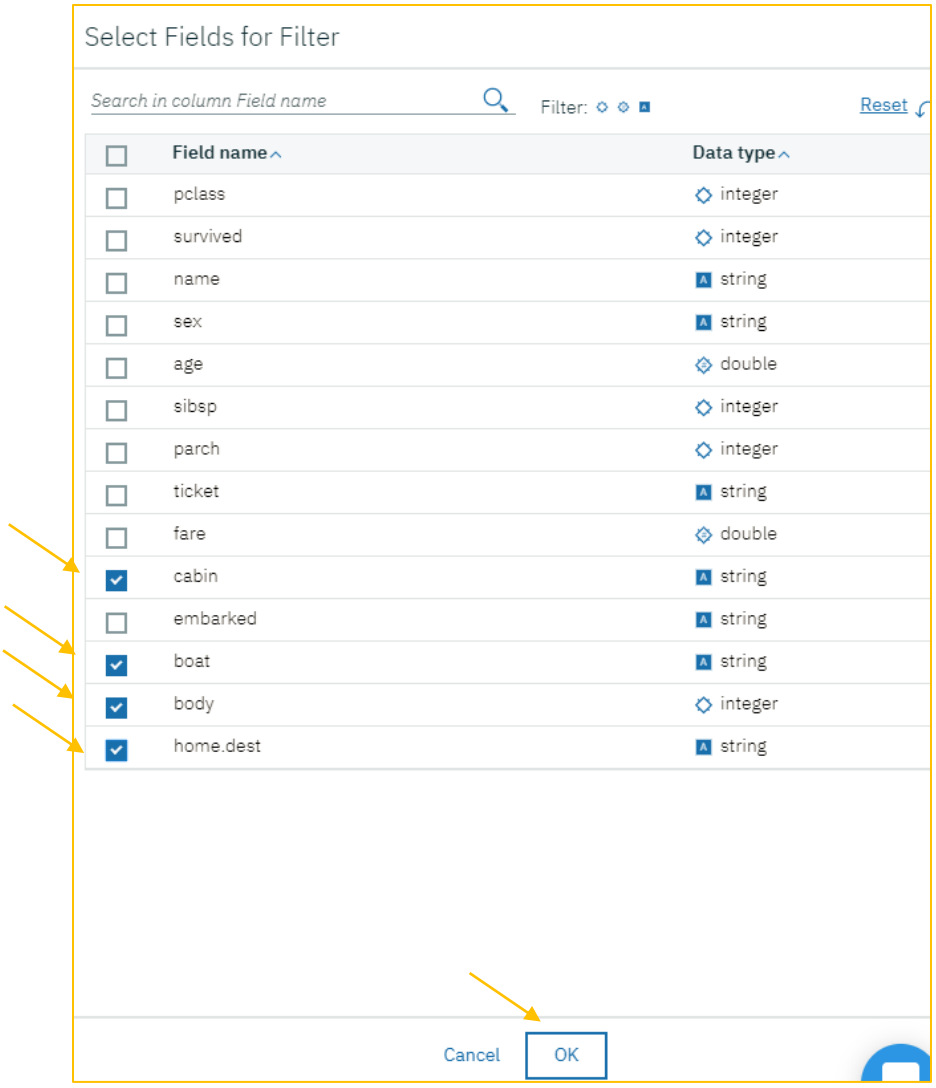
2. Add a **Filter** node to drop fields with many missing values. Add the **Filter** node by clicking on the **Field Operations** menu item in the Node palette and dragging the **Filter** node onto the canvas underneath the **Table** node. If the Node Palette is not visible, click on the Node Palette icon first. Connect the **Titanic.csv** node to the **Filter** node. The canvas should appear as below.



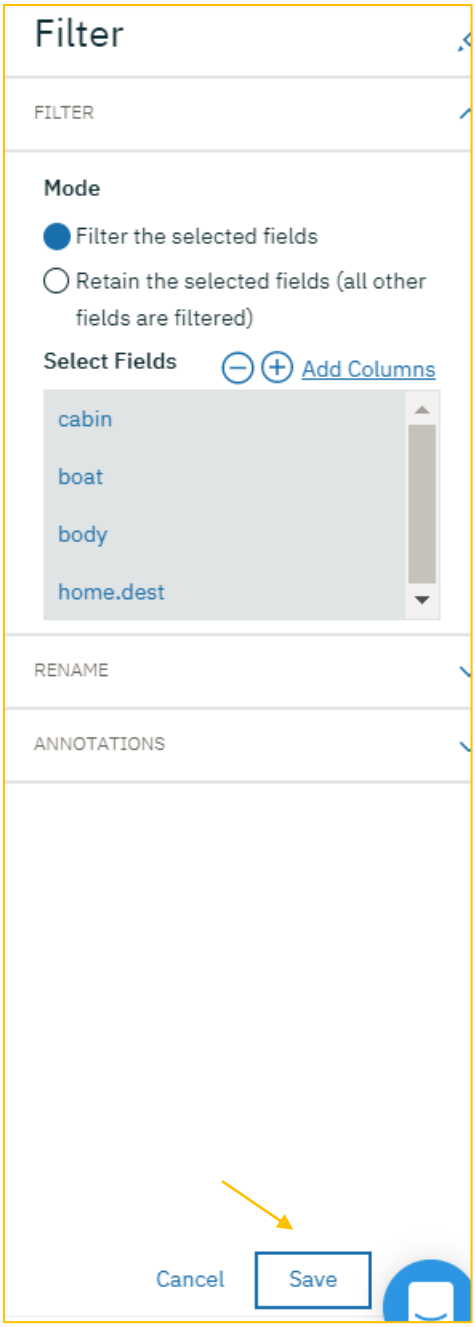
3. Double click on the **Filter** node. Click on the **Filter** dropdown. In the Filter panel, click on **Add Columns**.



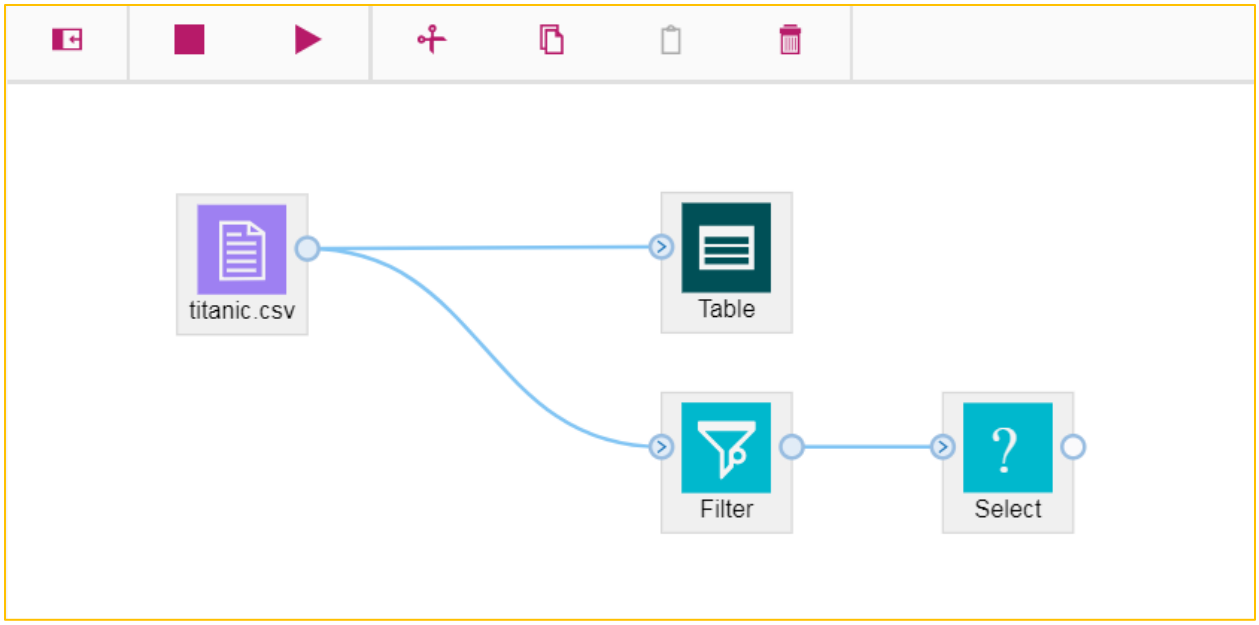
4. Click on the checkboxes adjacent to the **cabin**, **boat**, **body**, and **home.dest** fields, and then click on **OK**.



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

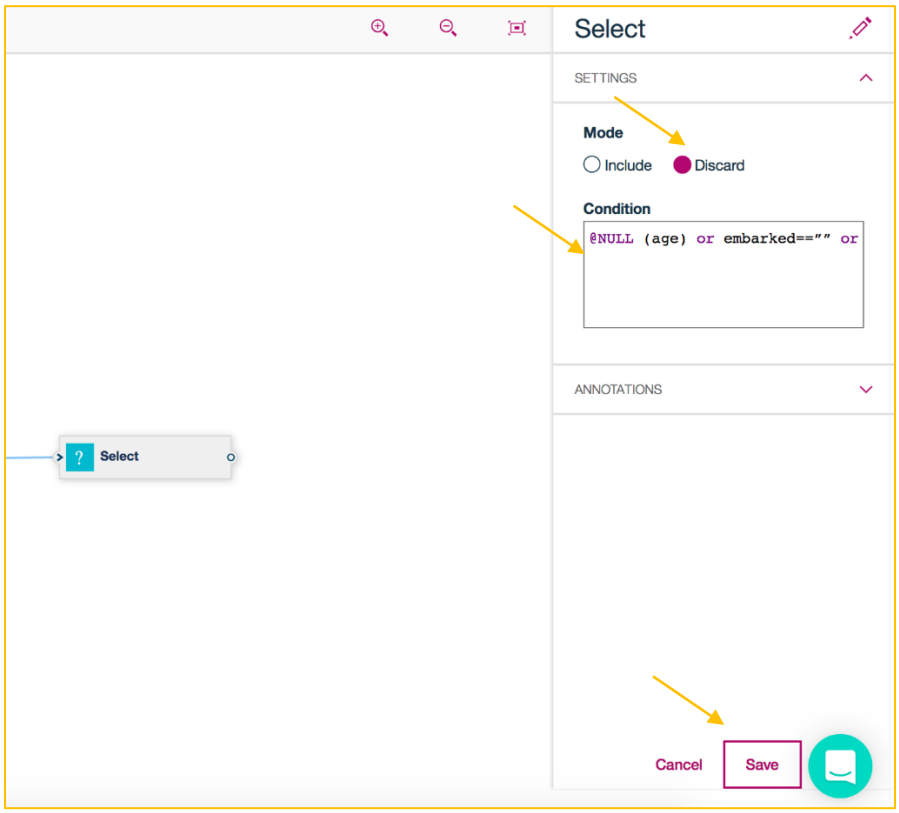


6. Add a **Select** node by clicking on the **Record Operations** menu item in the Node palette, and then dragging the **Select** node to the canvas to the right of the **Filter** node. Connect the **Filter** node to the **Select** node. If the Node Palette is not visible, click on the Node Palette icon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

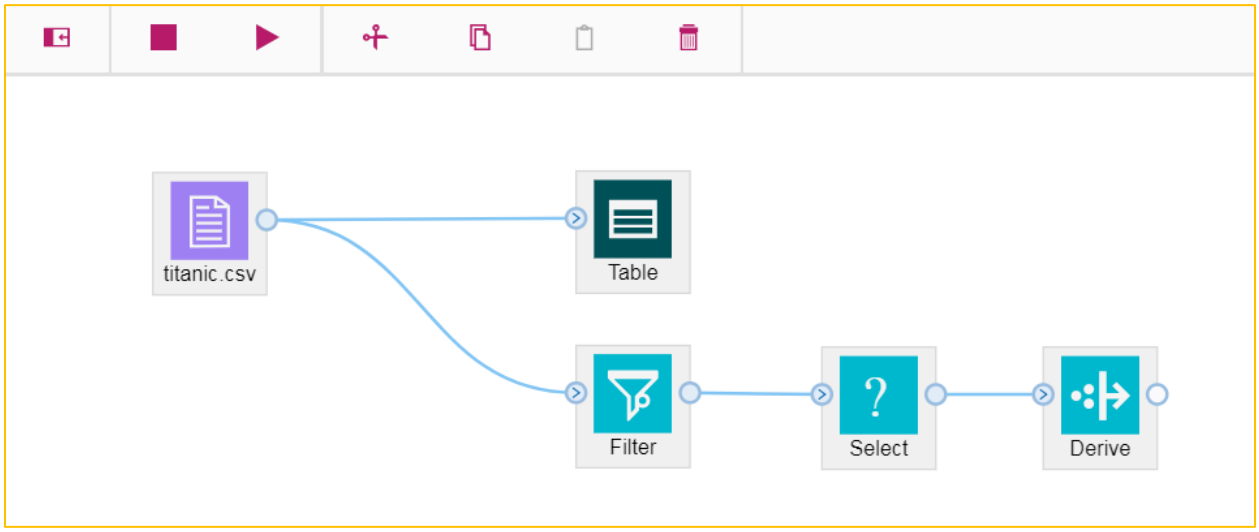


7. Double click on the **Select** node. Click on the **Settings** dropdown. In the **Select** panel, click on the **Discard** radio button, and re-type (or cut and paste) in the code shown below in the **Condition text box**, and then click **Save**.

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

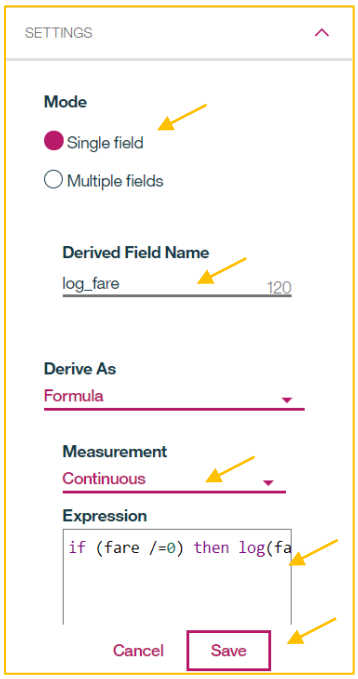


8. Add a **Derive** node to the canvas by clicking on the **Field Operations** menu item in the Node palette, and then dragging the **Derive node** onto the canvas to the right of the **Select** node. If the Node Palette is not visible, click on the Node Palette icon first. Connect the **Select** node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



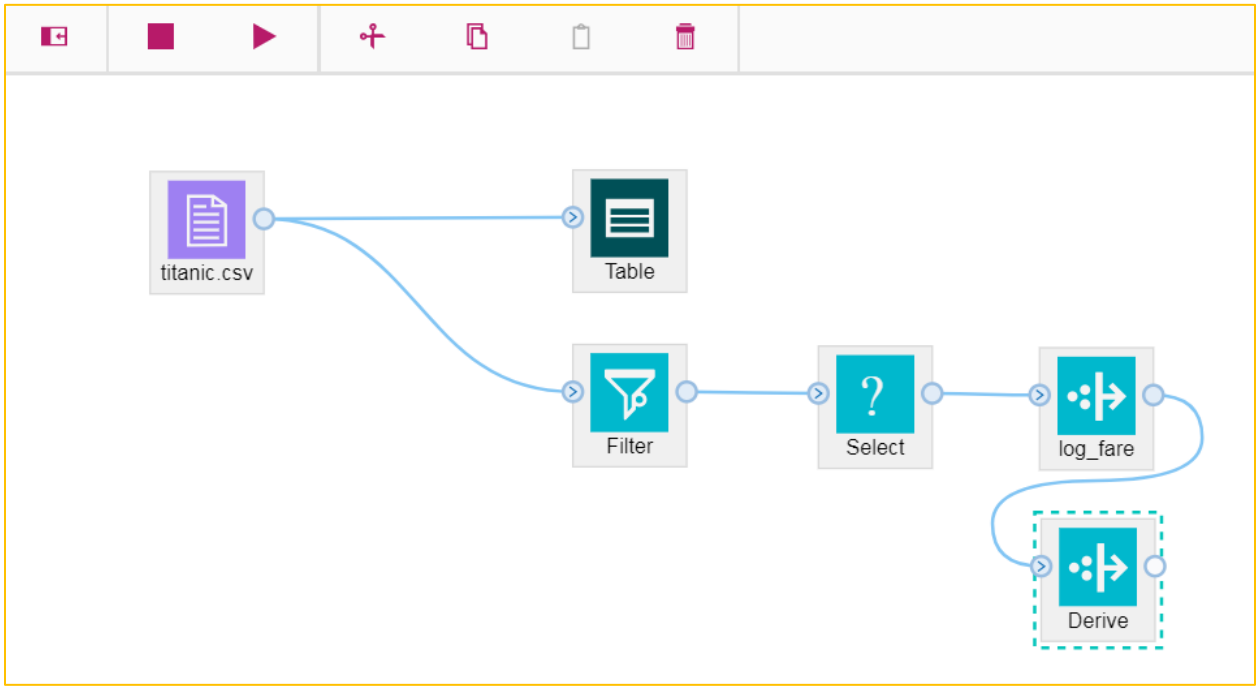
9. Double click on the **Derive** node. Click on the **Settings** Dropdown. Click on the **Single field** radio button, enter log_fare for the **Derived Field Name**, select **Continuous** for the **Measurement**, enter the following code in the **Expression** text box, and click Save.

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



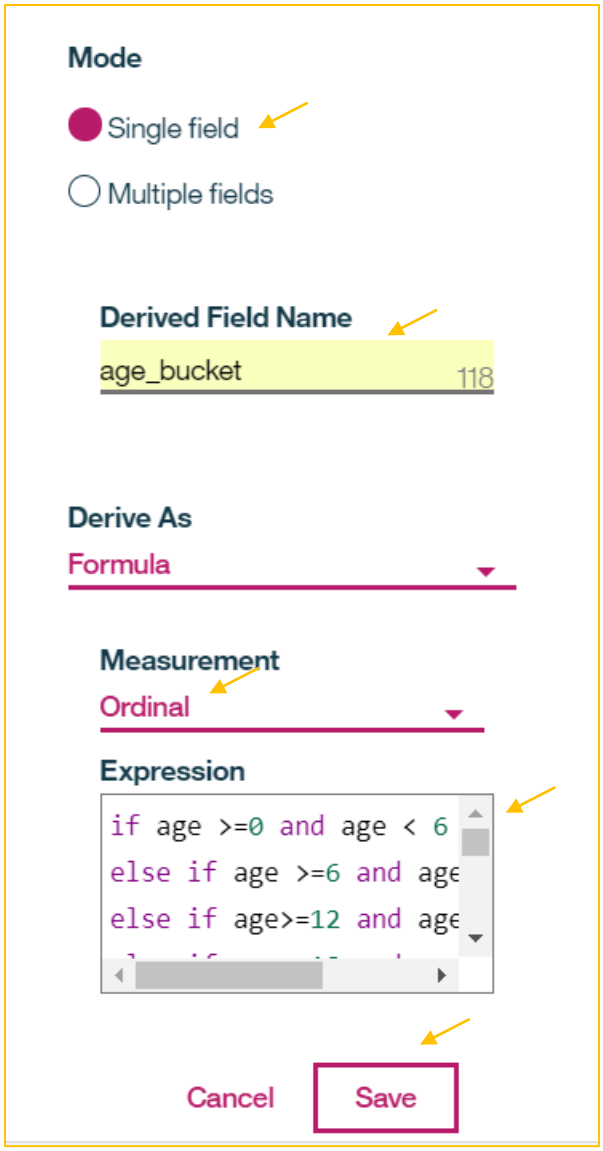
10. Binning of continuous fields is a technique sometimes used in preparing data for modeling. We will bin the age field, and the log_fare field. Add a **Derive** node by clicking on the **Field Operations** menu item in the Node palette and dragging the **Derive** node on the canvas underneath the log_fare **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.

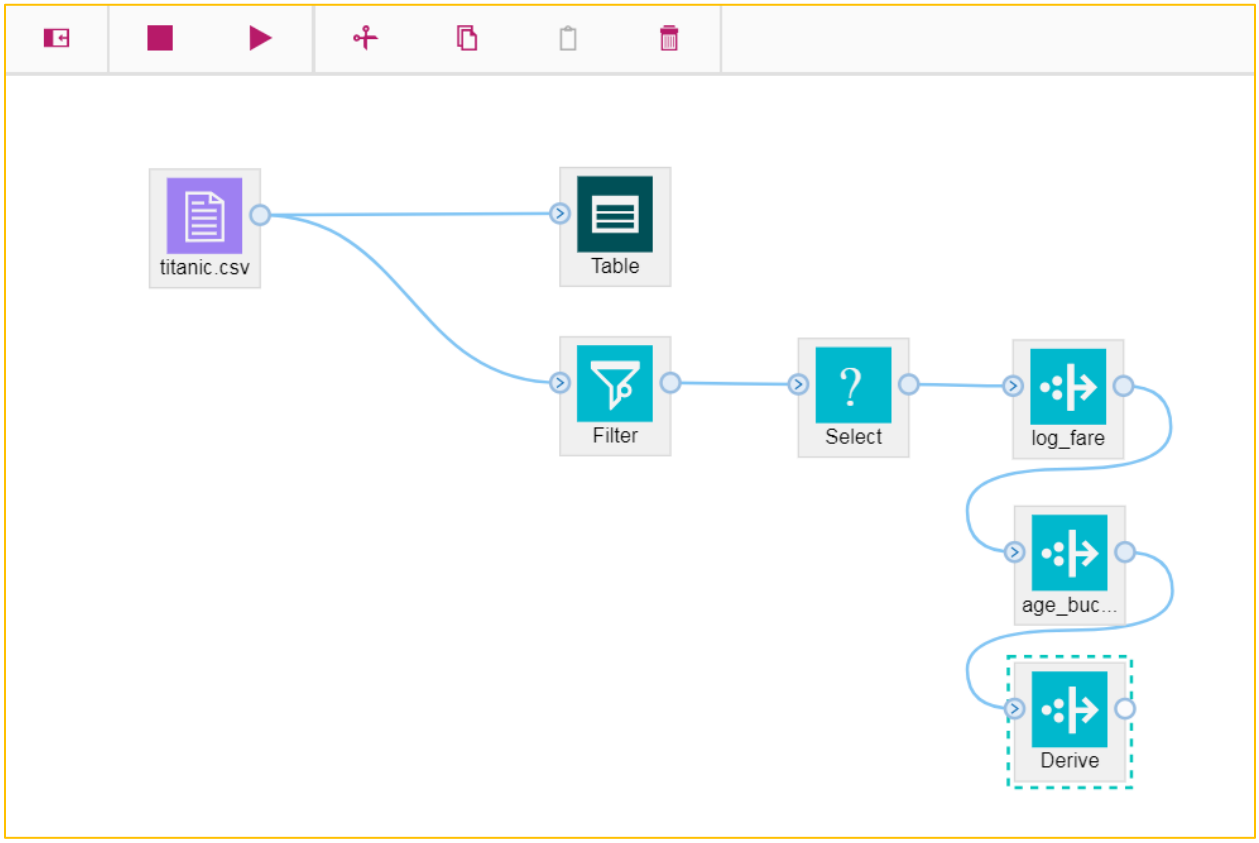


11. Double click on the **Derive** node. Click on the **Settings** dropdown. Click on the **Single field** radio button, enter age_bucket for the **Derived Field Name**, select Ordinal for the **Measurement,** enter the following code in the **Expression** text box, and the click **Save**.

```
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
```

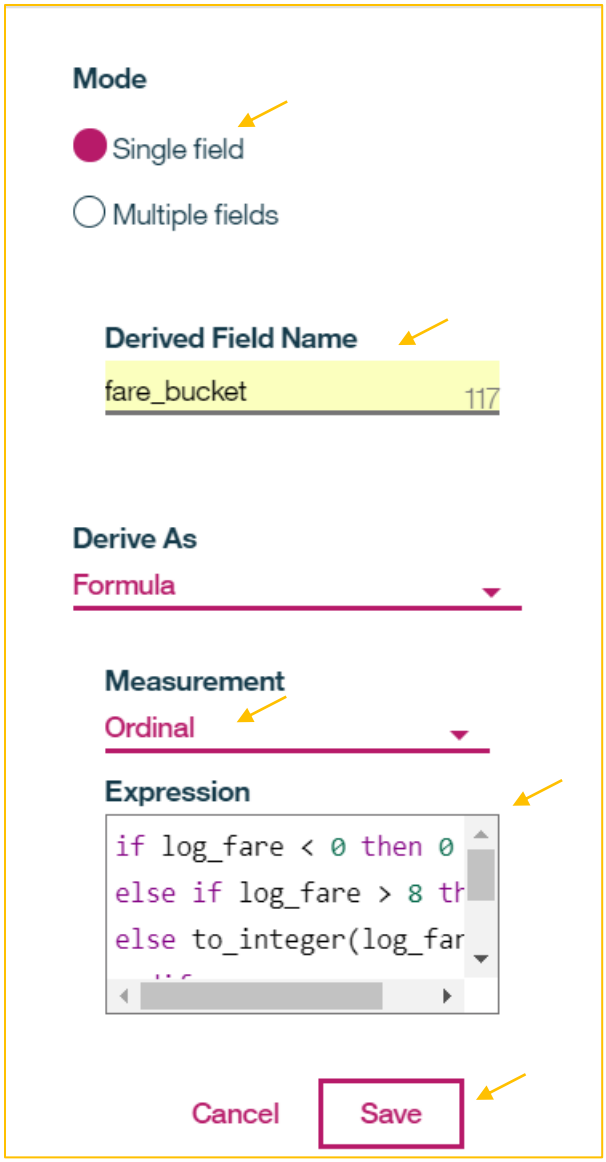

Mode
Single field
Multiple fields
Derived Field Name
age_bucket
118
Derive As
Formula
Measurement
Ordinal
Expression
if age >=0 and age < 6
else if age >=6 and age
else if age>=12 and age
Cancel
Save

12. 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.



13. Double click the **Derive** node. In the **Derive** panel, click on the **Single field** radio button, enter fare_bucket in the **Derived Field Name**, click on Ordinal for the **Measurement**, enter the following code in the **Expression** text box, and click on **Save**.

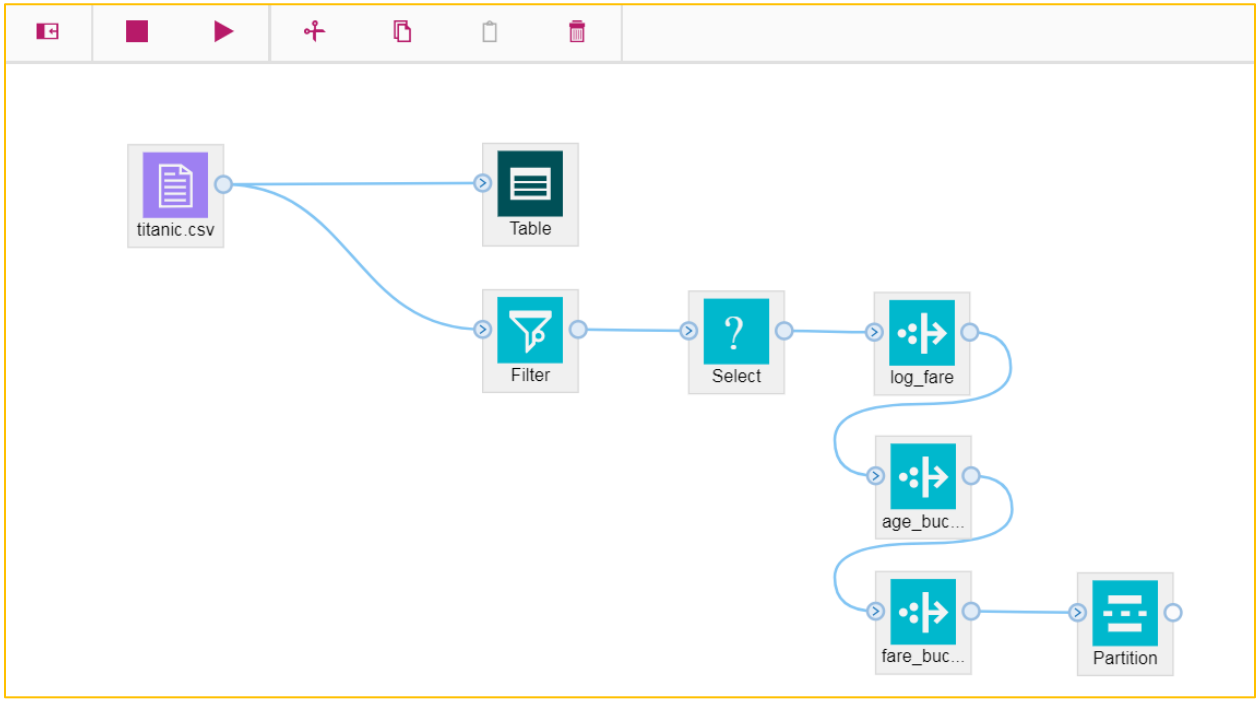
```
if log_fare < 0 then 0
else if log_fare > 8 then 9
else to_integer(log_fare)+1
endif
endif
```


Mode
Single field
Multiple fields
Derived Field Name
fare_bucket
117
Derive As
Formula
Measurement
Ordinal
Expression
if log_fare < 0 then 0
else if log_fare > 8 th
else to_integer(log_far
Cancel
Save

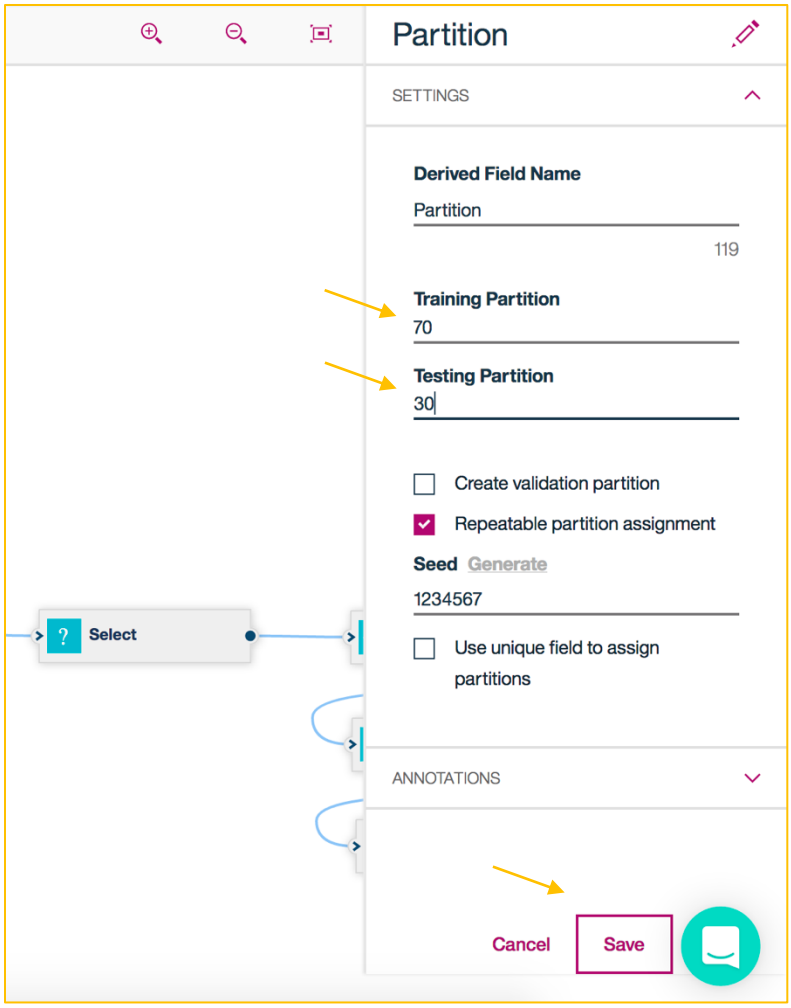
Step 2.5 Modeling and Evaluation

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, and to assign roles to each field. 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.

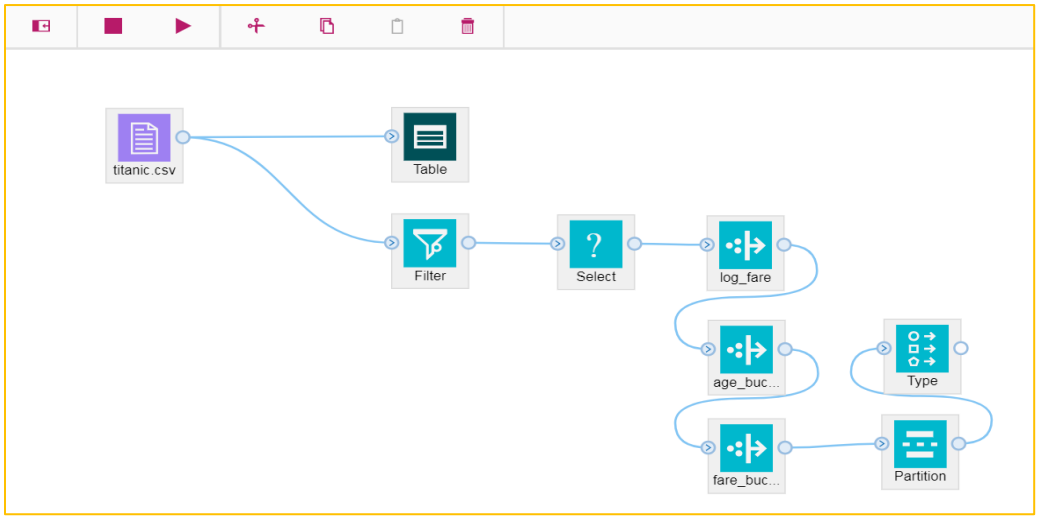
1. Add a **Partition** node by clicking on the Field Operations menu item in the Node palette and dragging the **Partition** node onto the canvas to the right of the fare_bucket **Derive** node. Connect the fare_bucket **Derive** node to the **Partition** node. The canvas should appear as below.



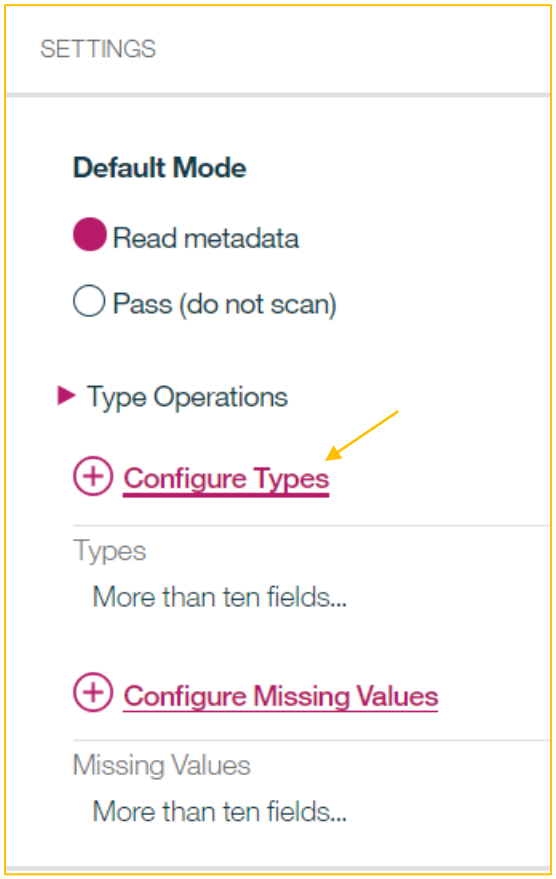
2. Double click on the Partition node. Set the **Training Partition** to 70 and the **Test Partition** to 30. Leave the other defaults, and click on **Save**.



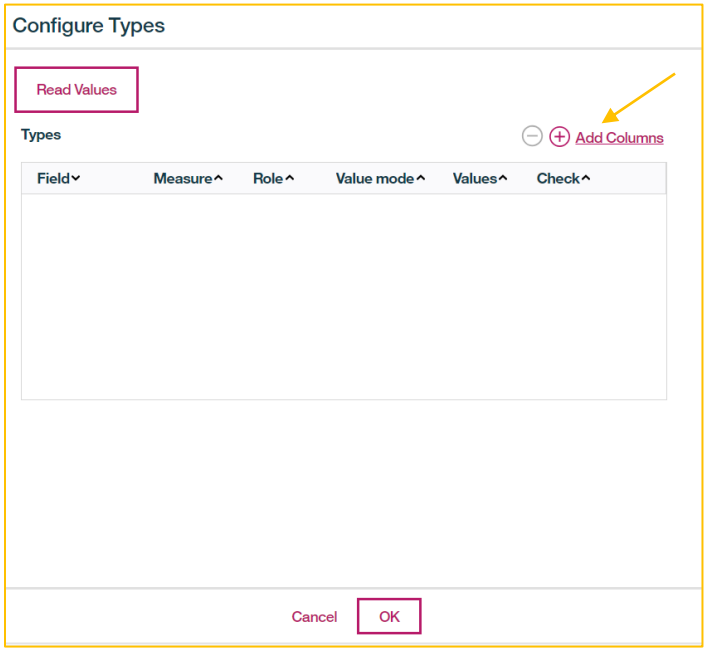
3. Add a **Type** node by clicking on the **Field Operations** in the Node palette and dragging the **Type** node onto the canvas above the **Partition** node. Connect the **Partition** node to the **Type** node. The canvas should appear as below.



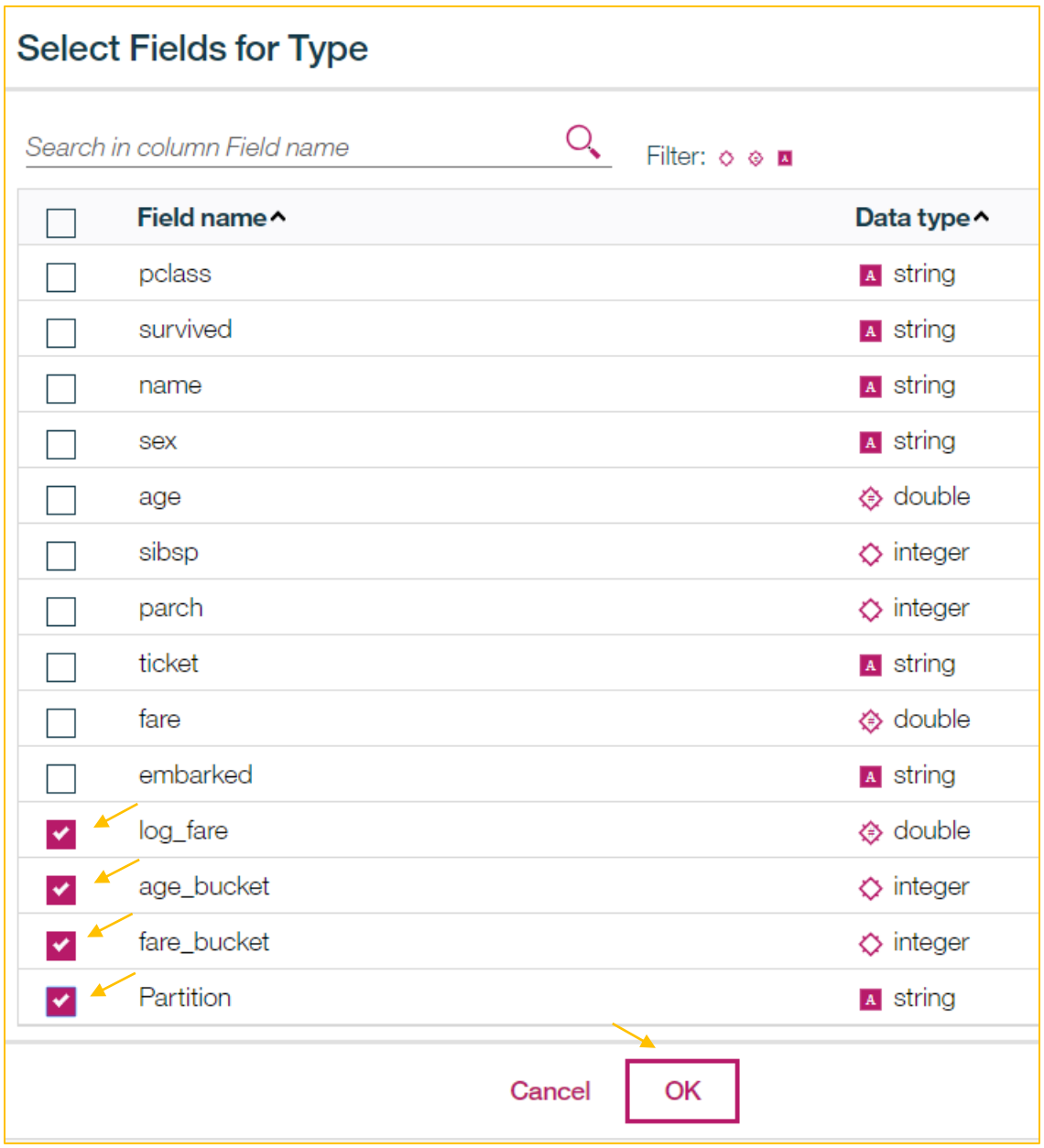
4. Double click on the **Type** node. Click on **Configure Types**.



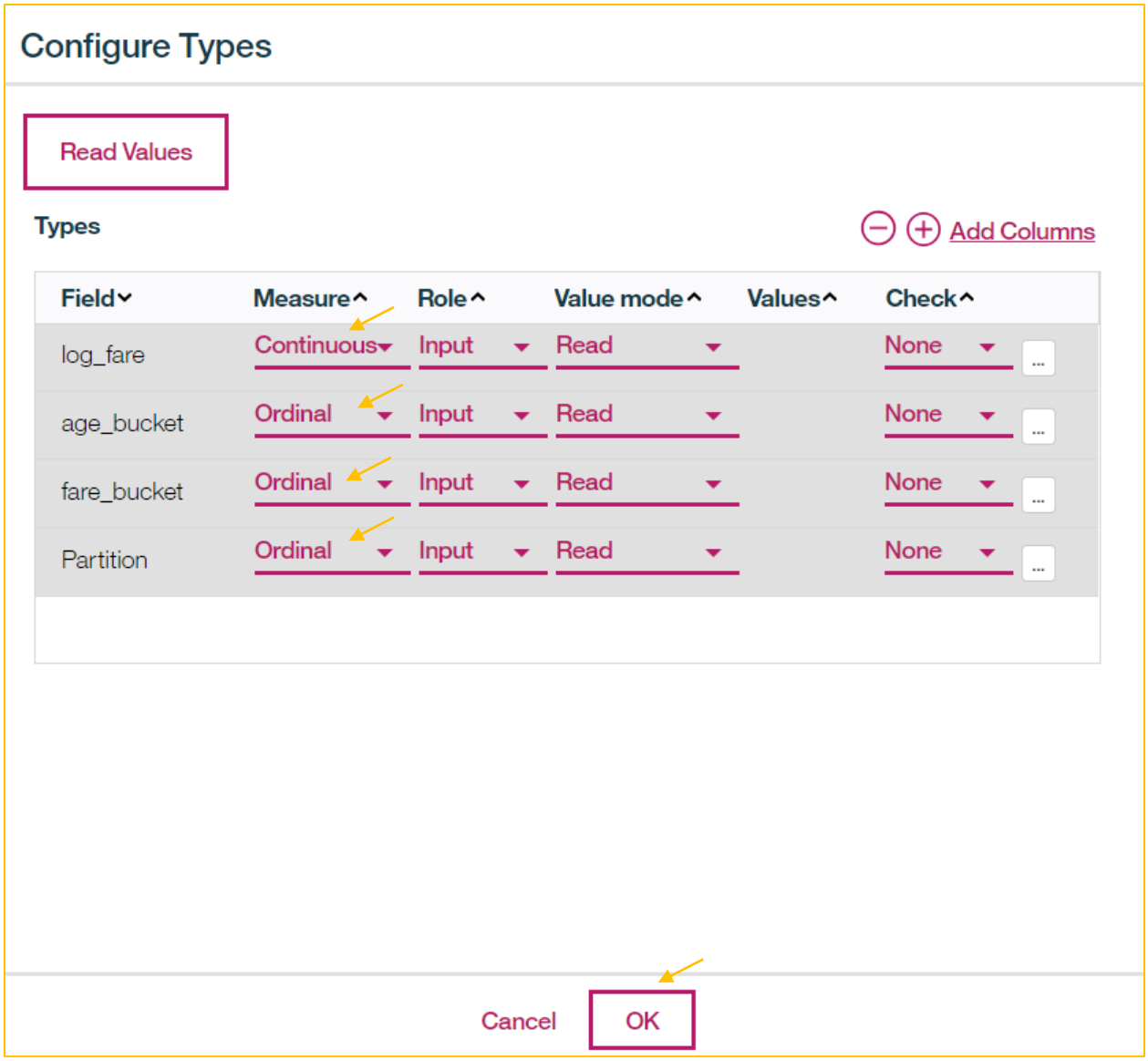
5. Click on **Add Columns.**



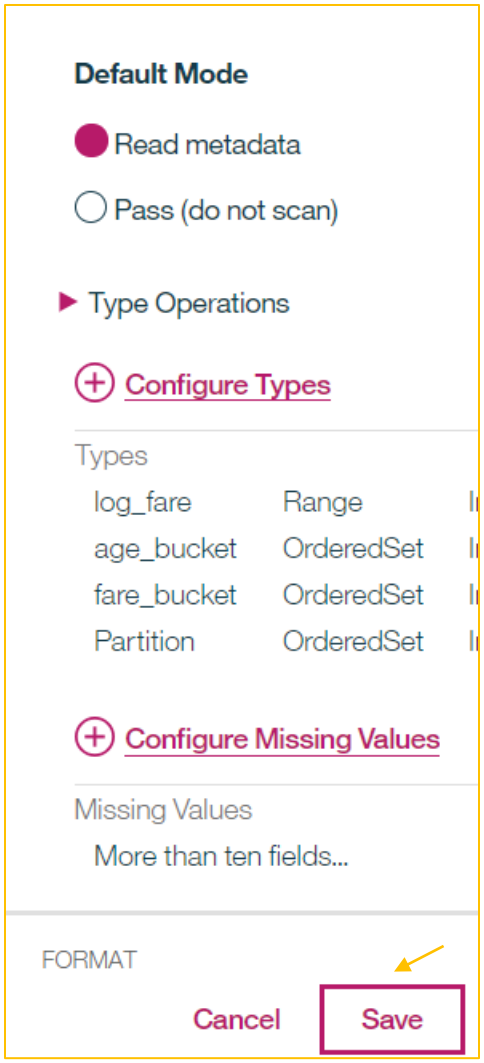
6. Click on checkboxes adjacent to the **log_fare**, **age_bucket**, **fare_bucket**, and **Partition** fields (You may need to scroll down). Click on **OK.**



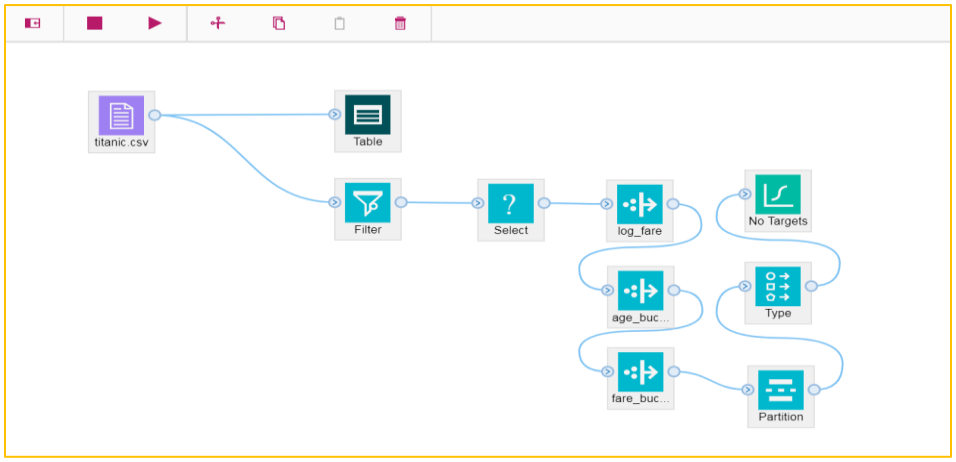
7. For the **Partition** field, select **Ordinal** for the **Measurement**. For the log_fare, select **Continuous** for the **Measurement**. For the fare_bucket field, select **Ordinal** for the **Measurement**, and for the age_bucket, select **Ordinal** for the **Measurement**, and click **OK**.



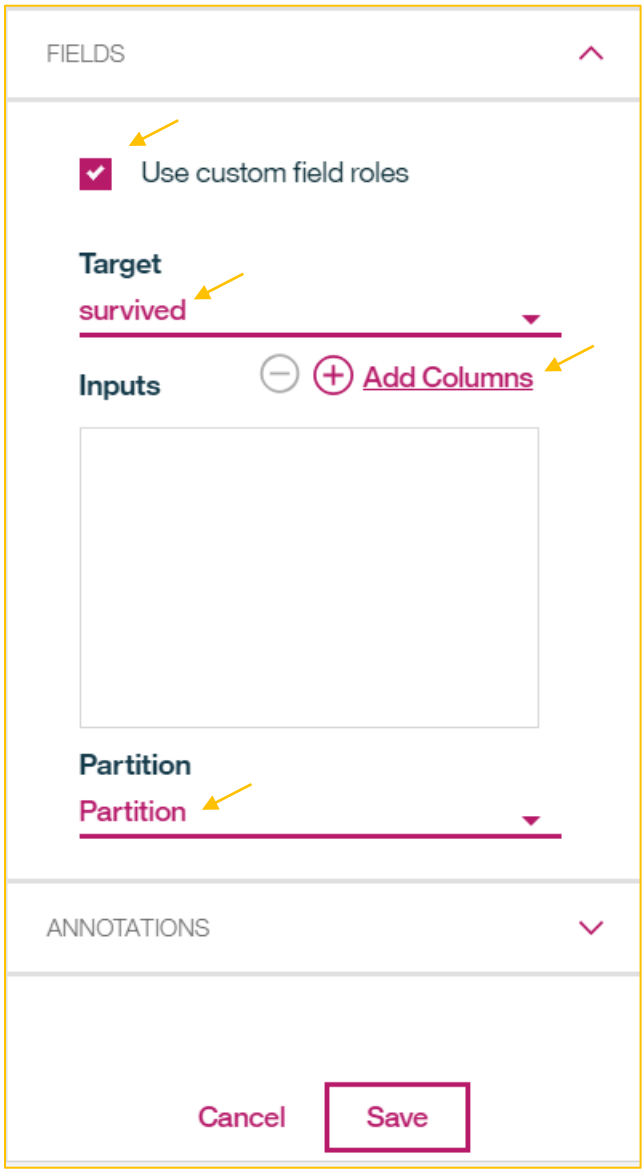
8. Click on **Save**



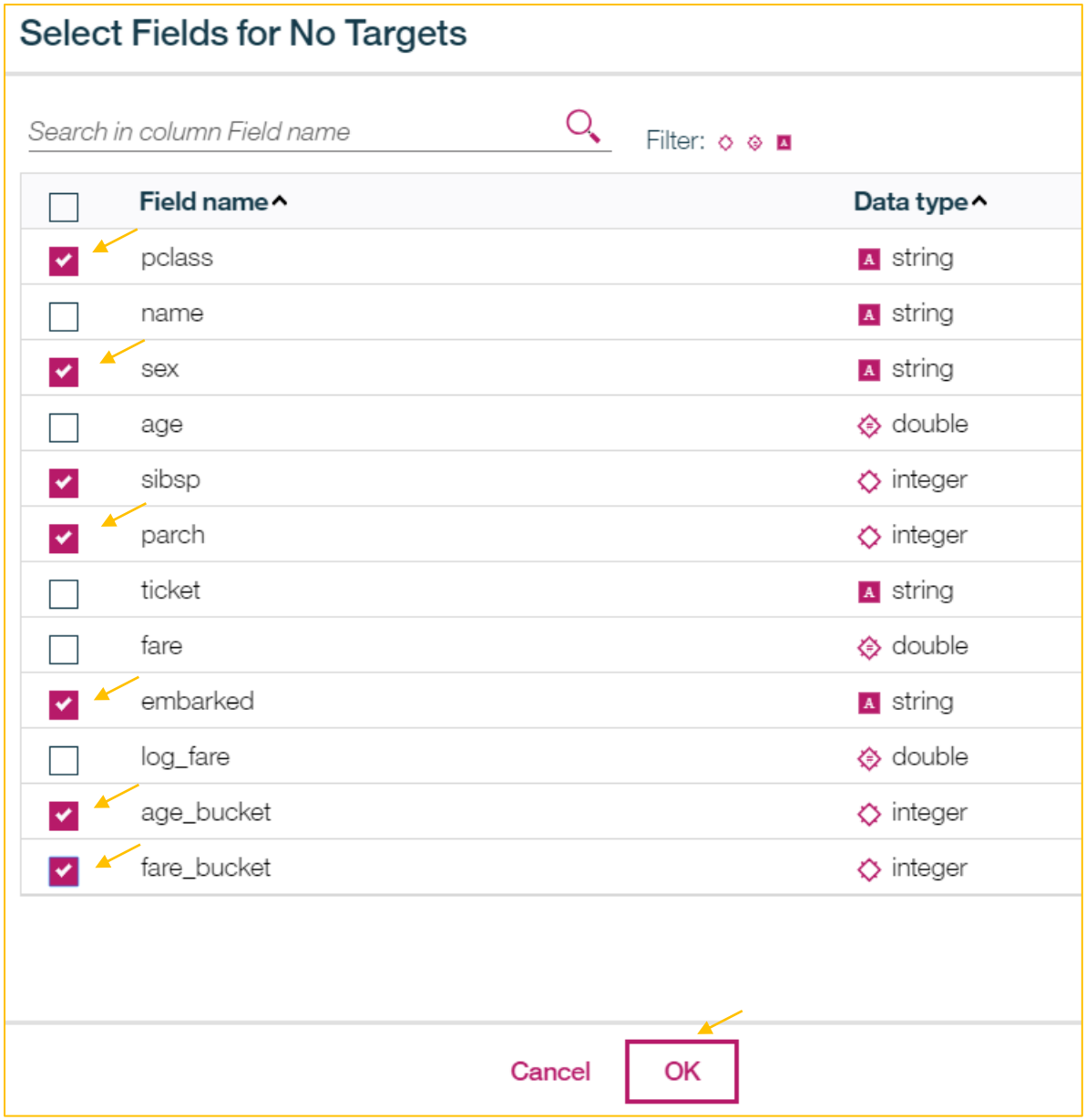
9. Add a **Logistic Regression** node by clicking on the **Modeling** menu item in the Node palette and dragging the **Logistic** node onto the canvas above the **Type** node. Connect the **Type** node to the **Logistic Regression** node. The canvas should appear as below.



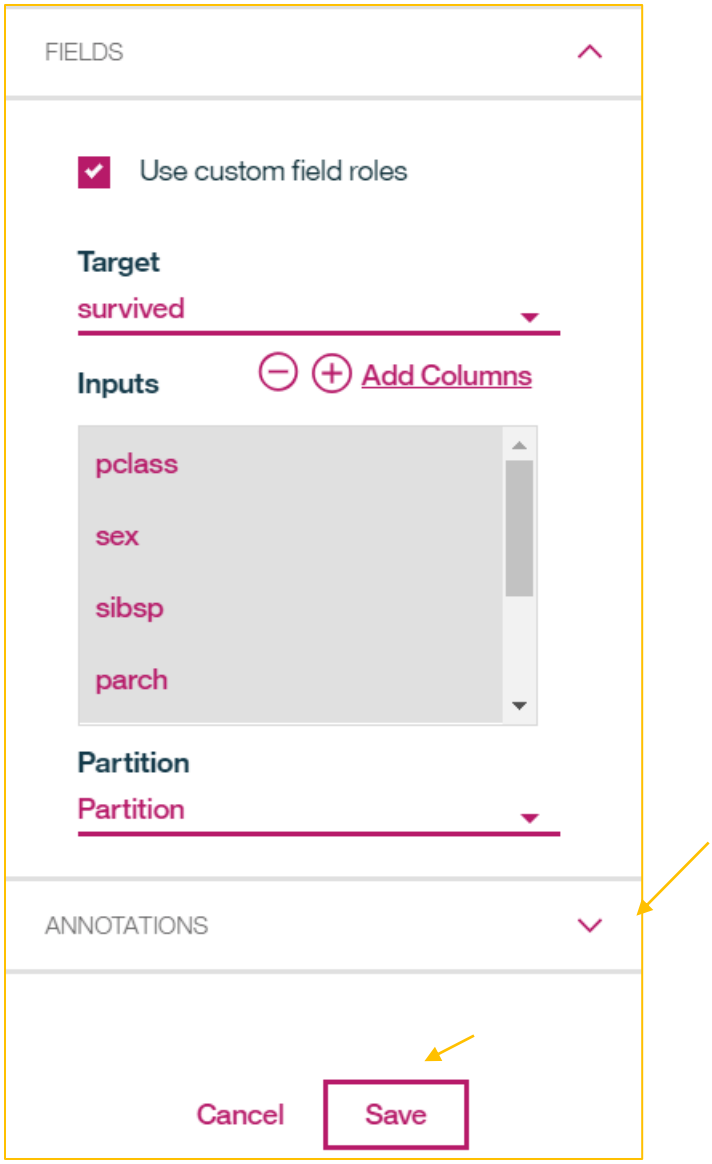
10. Double click on the **Logistic Regression** node. Click on the checkbox next to **Use custom field roles**, select **survived** for the **Target**, select **Partition** for the **Partition**, and click on **Add Columns** to add the input fields.



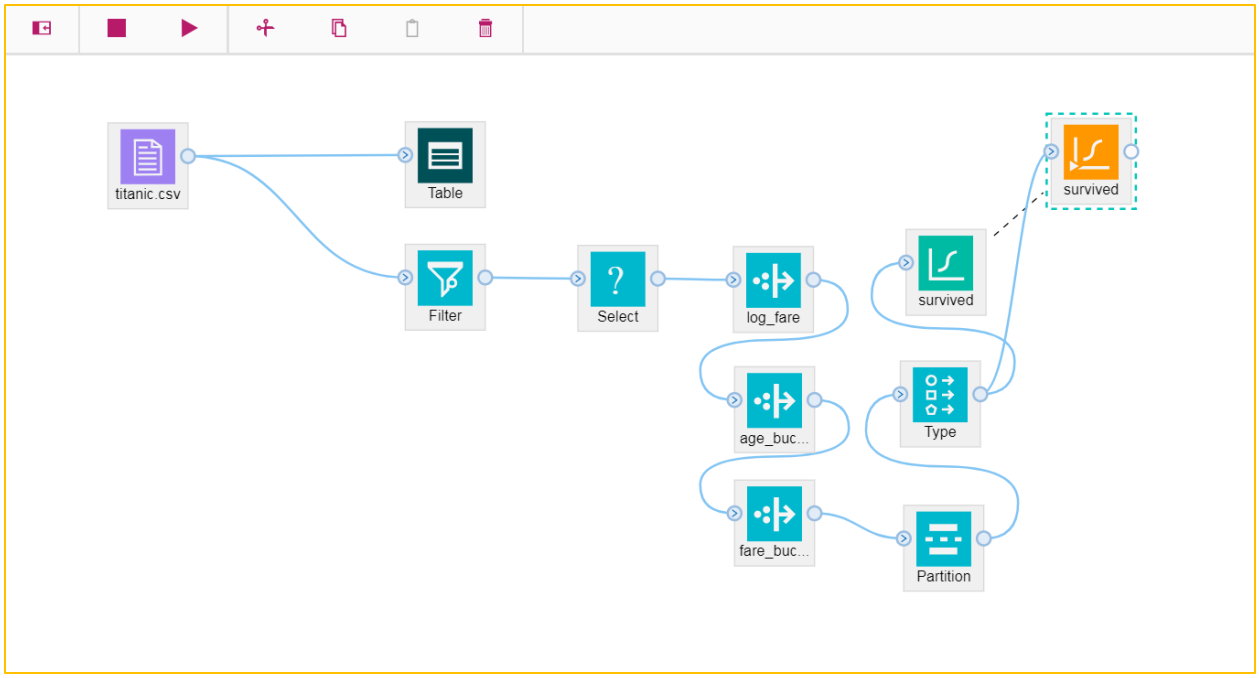
11. Click on the checkboxes next to pclass, sex, sibsp, parch, embarked, age_bucket, fare_bucket fields (you may have to scroll down), and then click OK.



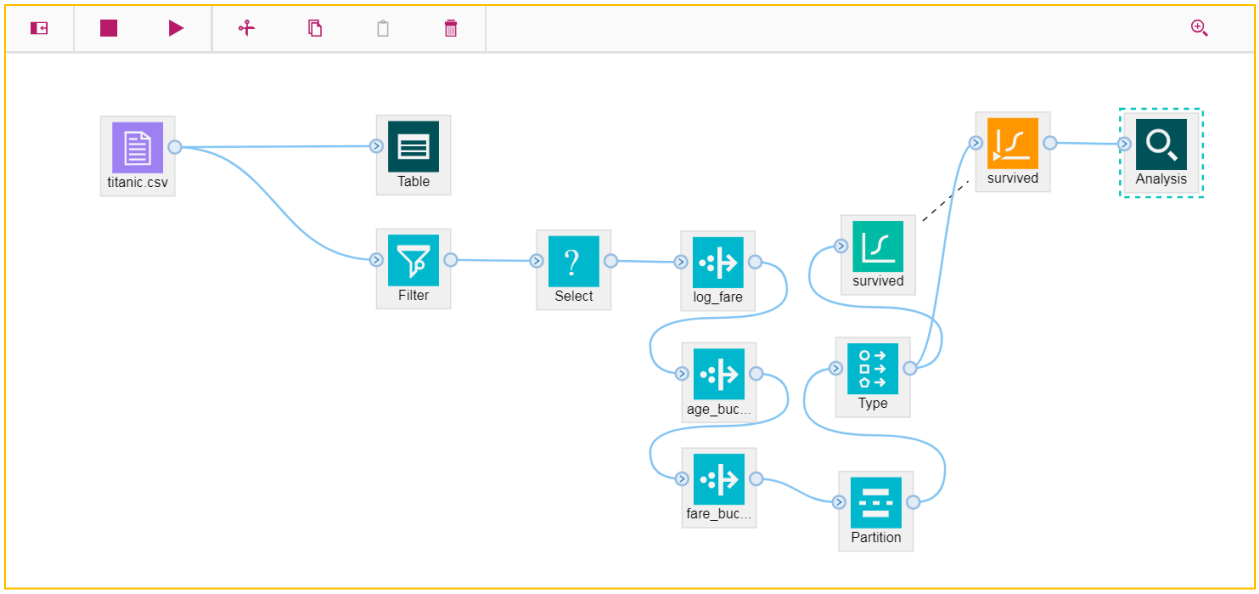
12. Click **Save**.



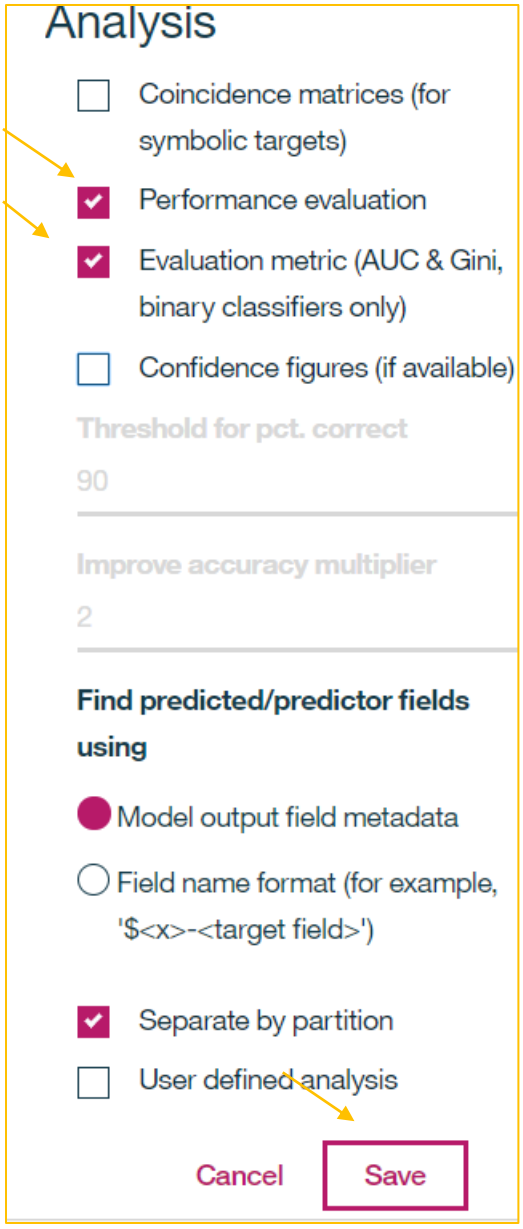
13. Right click on the **Logistic Regression** node and then click **Run**. A **Logistic Regression** "nugget will be created" connected by a dotted line to the **Logistic Regression** node. Drag the nugget and place it above the **Logistic Regression** node. The canvas should appear as below.



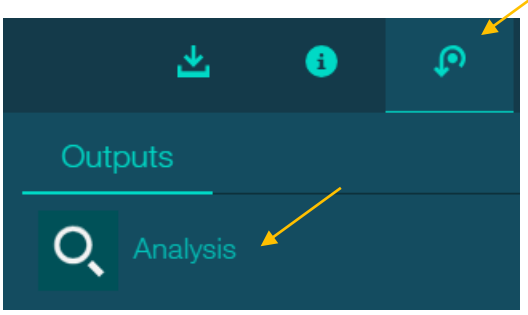
14. Add an **Analysis** node by clicking on the **Outputs** menu item in the Node palette and dragging the **Analysis** node onto the canvas above the nugget icon. Connect the nugget icon to the **Analysis** node. The canvas should appear as below.



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, click on the Output icon and then double click on the Analysis link.



17. 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	

Performance Evaluation

0	0.317
1	0.628

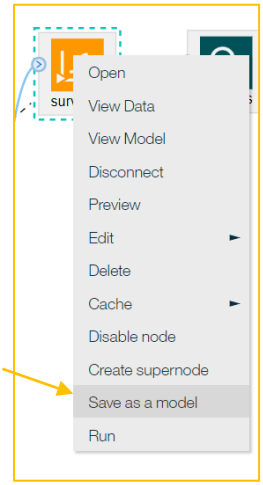
Evaluation Metrics

Model	AUC	Gini
$L-survived	0.857	0.714

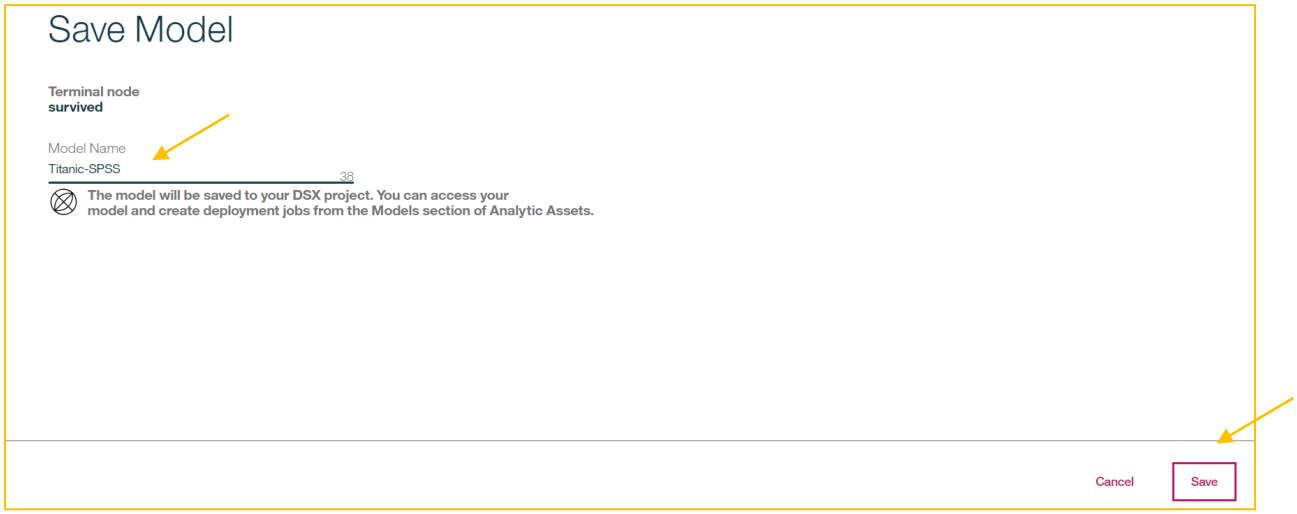
Step 2.6 Saving a Model

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.

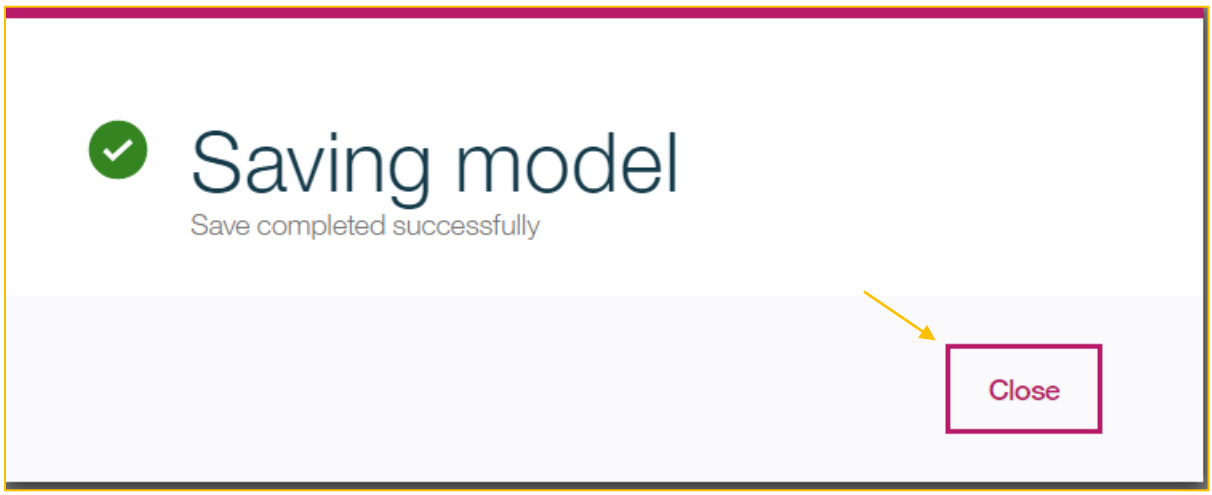
1. Right click on the Generated Model and then click on **Save as a model**.



2. Type in "**Titanic-SPSS**" as the Model Name and click **Save**.



3. Click **Close**.



4. Navigate to your project's **Models** page to see the saved Titanic-SPSS model. This model can then be deployed.