

Watson Studio SPSS Modeler Overview

Introduction

In this lab you will learn how to implement analytics in **SPSS Modeler**, a well-known visual data mining workbench which is part of **Watson Studio**. The lab will introduce the SPSS Modeler capability using the trafficking datasets. The lab will guide the development of an SPSS Modeler stream that will prepare the input data to train and evaluate a machine learning model for predicting the trafficking risk based on the travel itinerary.

End-to-End Data Science

The general flow of the End to End Data Science PoT will be guided by the activities shown in Figure 1- End to End Flow. The SPSS capability spans the Prepare Data, Build Model, and Save and Deploy activities.

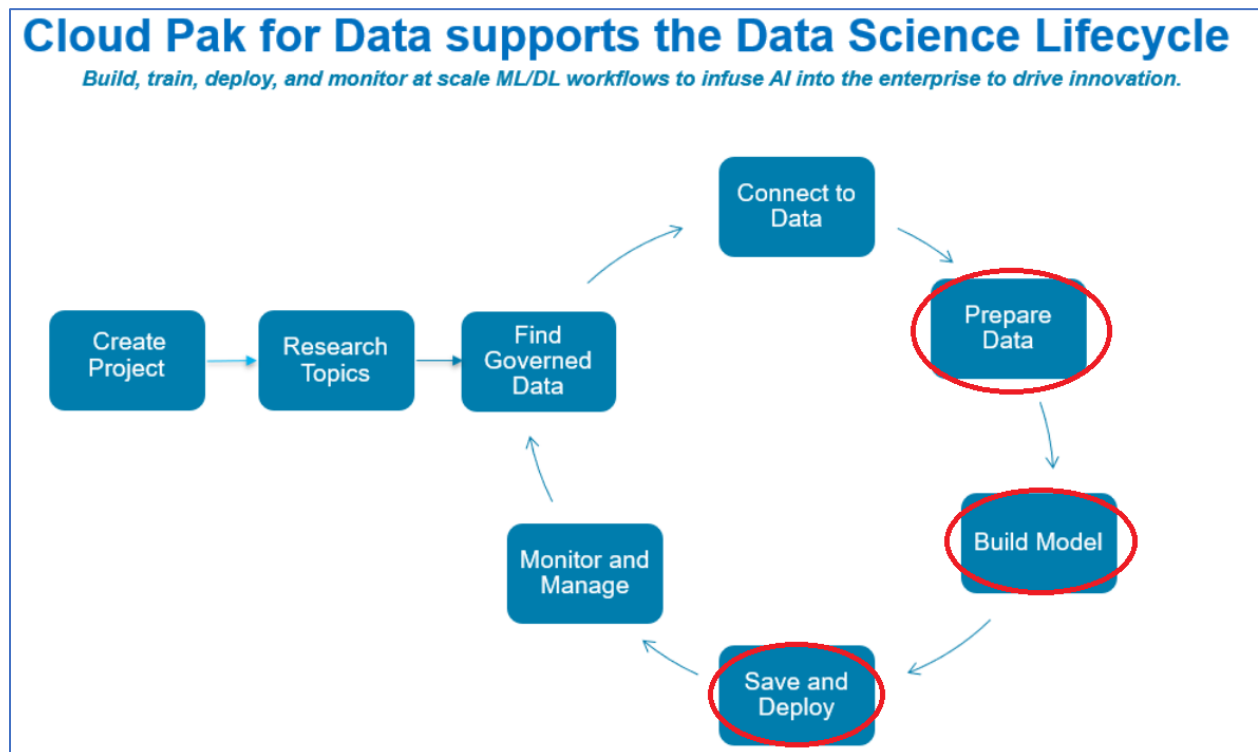


Figure 1- End to End Flow

Objectives

1. Become familiar with the Watson Studio SPSS Modeler capability
2. Load the trafficking data into SPSS Modeler
3. Join the datasets
4. Profile the trafficking data
5. Prepare the trafficking data
6. Train/Evaluate a machine learning model.
7. Save the model.







Background

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

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

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: **Import, 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.

Female Human Trafficking Data

The data sets used for this lab consist of **simulated** travel itinerary data. The use case corresponds to an analyst reviewing the travel data to assign a risk of trafficking. The risk is recorded as the VETTING_LEVEL column in the dataset. Some of the records have already been analyzed and have a VETTING_LEVEL of low (value is 30), medium (value is 20), or high risk (value is 10). Others have not yet been vetted (value is 100). We will use the data that has been vetted to train a model to predict the risk for the unvetted records. This can be used to automate the process and augment the analyst. For example, one option would be to send the predicted high-risk persons to the analyst for further investigation.

The OCCUPATION data included in the travel data is very granular. For modeling purposes, it was decided to categorize the OCCUPATION data. Two additional datasets are used for this purpose. The occupation.csv dataset maps the granular occupation data to a category code. The categories dataset maps a category code to a category description. These datasets will be joined to the main dataset to prepare the data for modeling.

Other columns in the dataset are similarly very granular and could also be categorized for modeling purposes. This lab does not include steps to accomplish this, but it would be similar to what was done for the occupation column.

Lab Steps

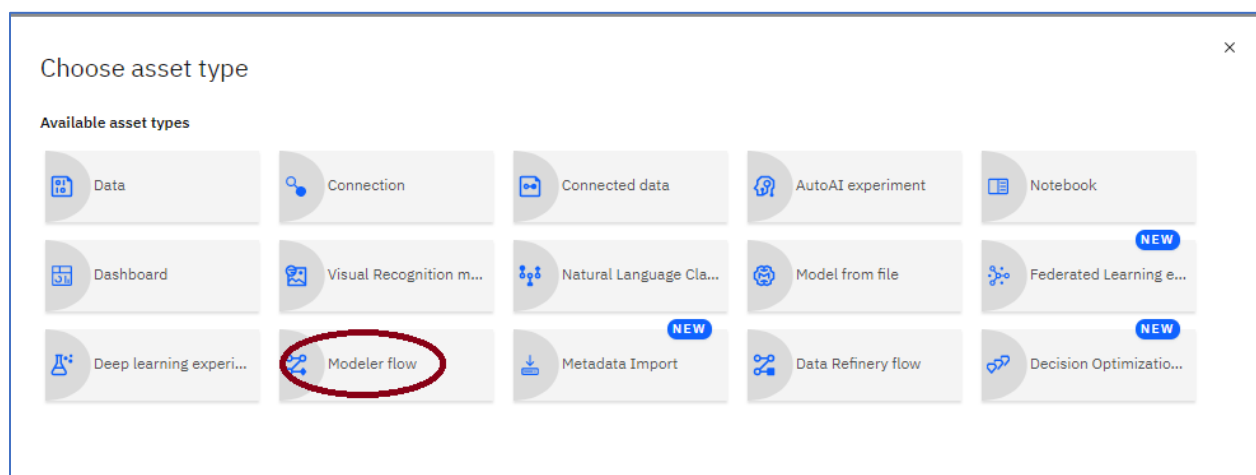
In this section, we will create a Machine Learning flow using SPSS nodes.

Step 1 - Create a New Flow

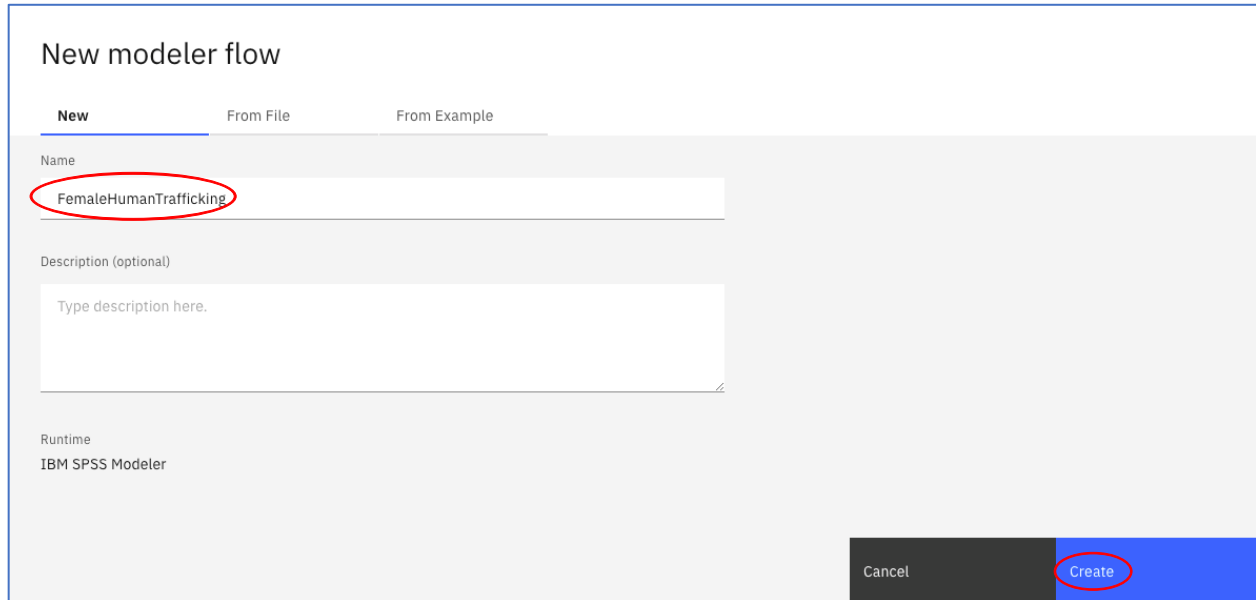
1. In the Watson Studio project, click on **Add to project**.



2. Select **Modeler Flow**.



3. Enter a **Name** for the flow, optionally enter a **Description**, and click on **Create**.



New modeler flow

New From File From Example

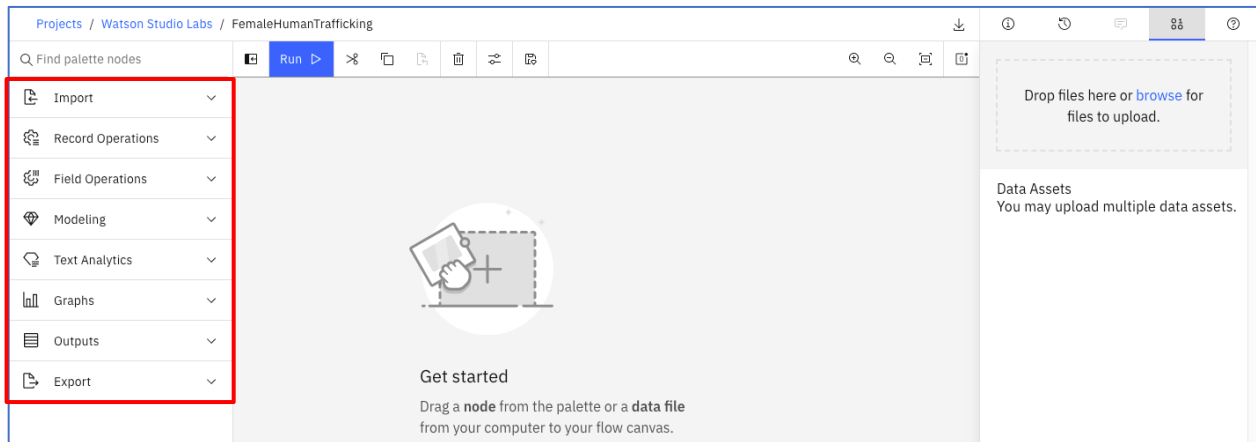
Name
FemaleHumanTrafficking

Description (optional)
Type description here.

Runtime
IBM SPSS Modeler

Cancel Create

4. This opens the Flow Editor. Note the palette of operations on the left-hand side.



Projects / Watson Studio Labs / FemaleHumanTrafficking

Find palette nodes

Run >

Import

Record Operations

Field Operations

Modeling

Text Analytics

Graphs

Outputs

Export

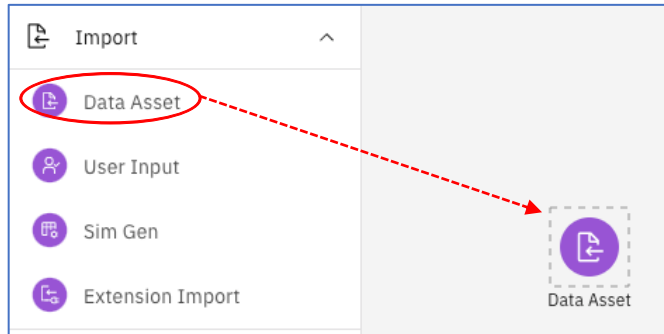
Drop files here or [browse](#) for files to upload.

Data Assets
You may upload multiple data assets.

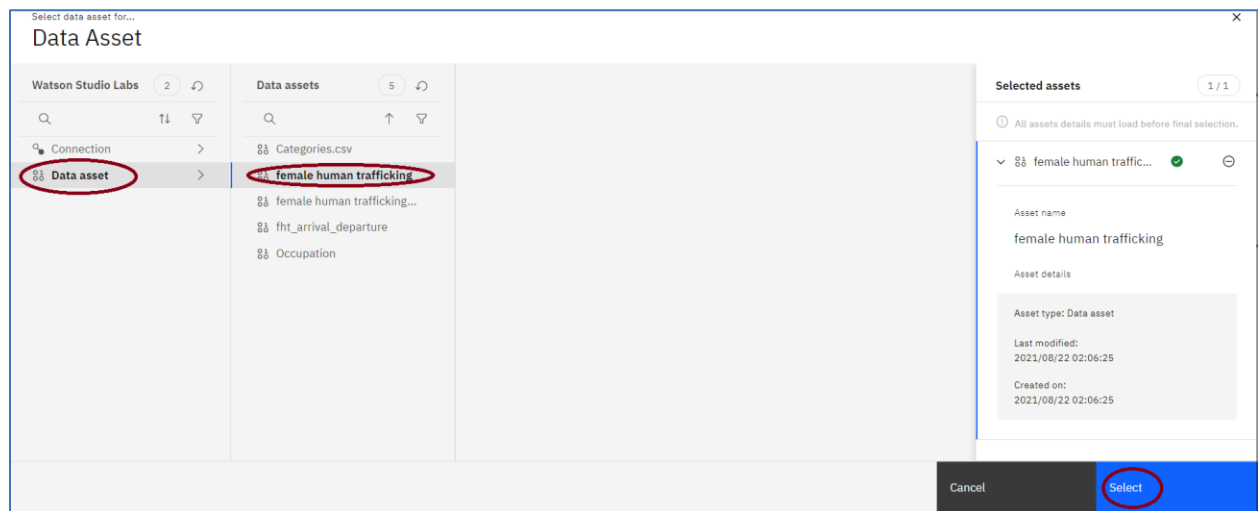
Get started
Drag a **node** from the palette or a **data file** from your computer to your flow canvas.

Step 2 - Load the Trafficking Datasets

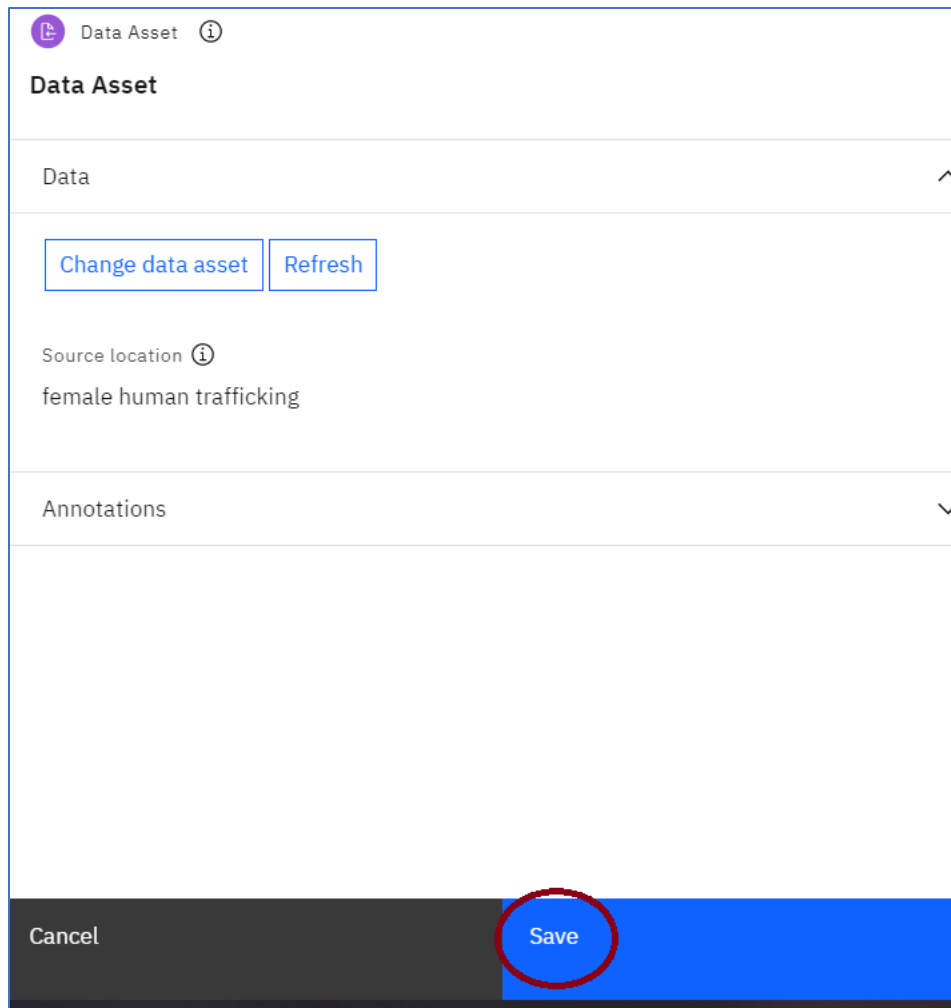
1. Click on **Import** and then **Data Asset** and hold the left mouse key on the Data Asset icon and **drag it onto the left side of the canvas**. Release the left mouse key.




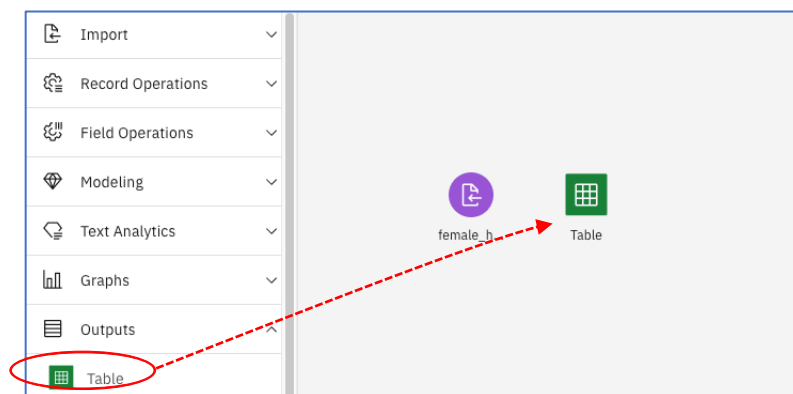
2. Double click on the **Data Asset**. Click on **Data Assets**, click on **female_human_trafficking** (make sure not to select on female_human_trafficking_shaped), then click **Select**.



3. Click **Save**.

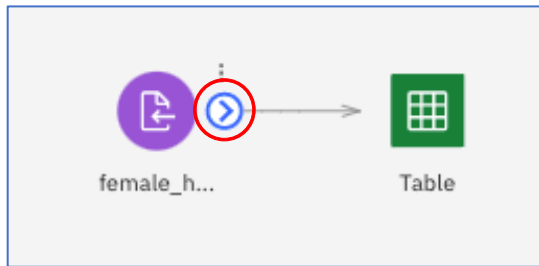


4. Click on the **Outputs** menu item in the Node Palette on the left and then click on the **Table** icon and drag the icon to the right of the female_human_trafficking to display its contents. If the Node Palette is not visible, click on the Node Palette icon .

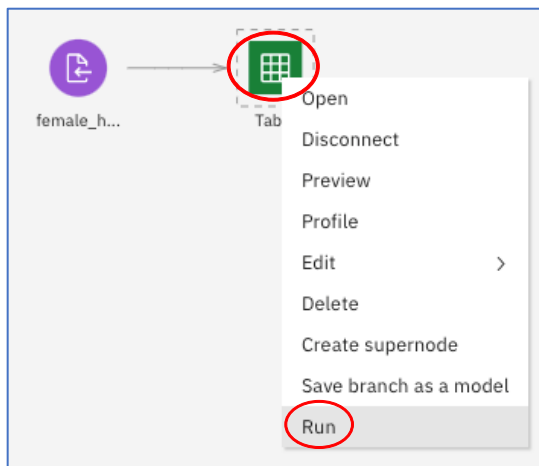


5. **Connect the right side of the female_human_trafficking icon to the left side of the Table icon.** This is accomplished by hovering over the data asset icon, clicking on the

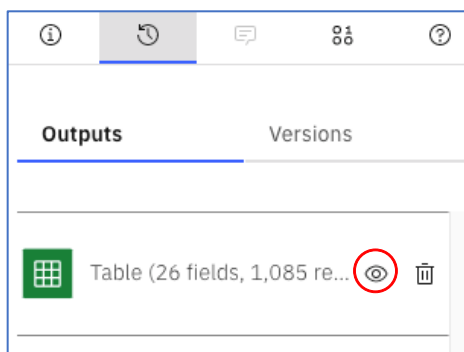
little blue arrow that appears the right side of the icon, holding the left mouse key, and dragging the mouse to the Table icon, and then releasing the left mouse key.



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. If the Table output selection does not appear, select the clock icon. **Click on the ⌚ icon.** and the contents of the female_human_trafficking is displayed.

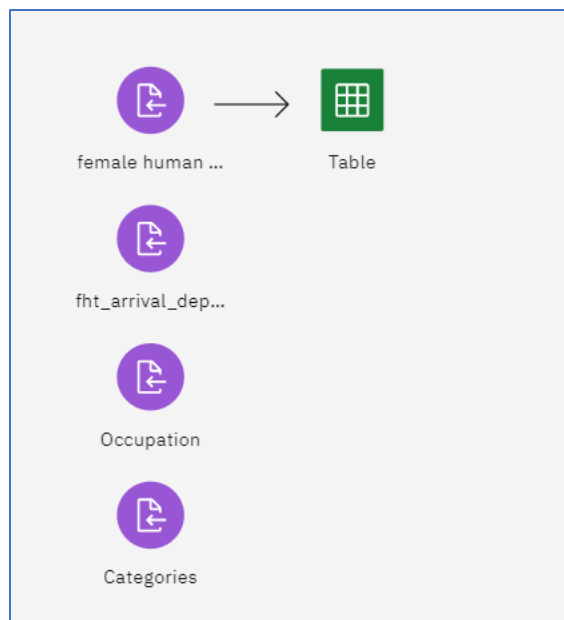


8. Each row contains travel information for a person. We will use this data to make predictions on trafficking risk. Click **x** to close the window.

View Output: Table (16 fields, 1,085 records)


INTERNAL_ID	VETTING_LEVEL	NAME	GENDER	BIRTH_DATE	BIRTH_COUNTRY	BIRTH_COUNTRY_CODE	OCCUPATION	ADDRESS	SSN	PASSPORT_NUM
215	30	Cristie Moore	F	1998-07-16	Ghana	GH	Osteopath	801 James Bypass, Palisades, Idaho 84337	284-95-7110	156564572
216	100	Susan Ashley Long	F	2000-05-09	Ghana	GH	Psychologist, clinical	49720 Houston Ramp Suite 059, Autaugaville, Alabama 36003	042-44-5637	470901453
217	100	Michelle Brown	F	1980-11-17	Ghana	GH	Occupational psychologist	0111 Nicole Port Suite 856, Lubbock, Texas 79405	741-14-7420	151443012
218	100	Jessica Cheryl Dunn	F	1973-05-06	Ghana	GH	Clinical research associate	0490 Nicholas Strm Apt. 079, Rock Falls, Iowa 50467	534-73-2450	800966716

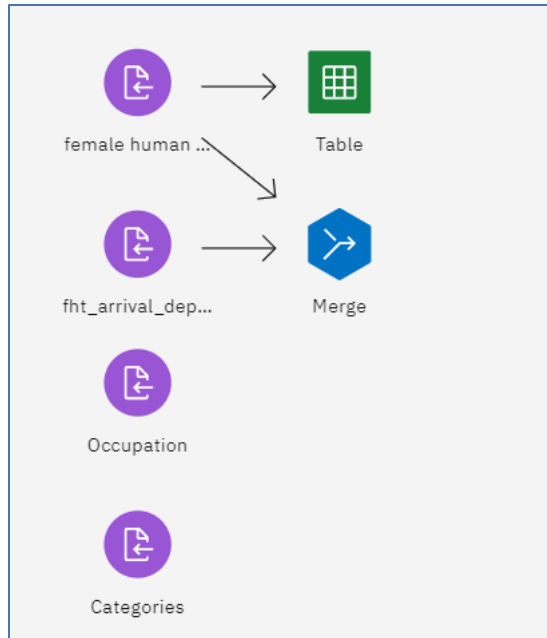
9. Repeat steps 1-3 for the **fht_arrival_departure**, **Occupation** dataset and the **Categories** dataset. When complete, the canvas should appear as below.



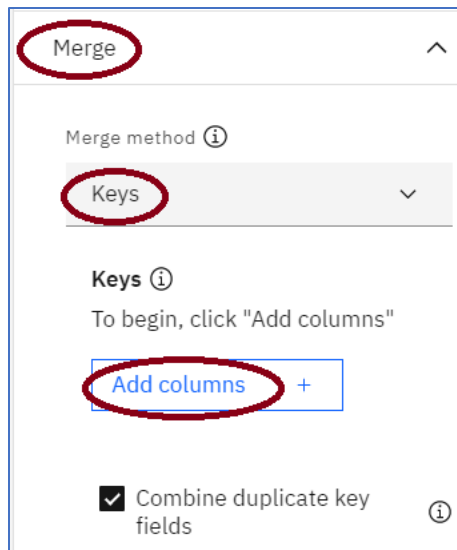
Step 3 - Join the Data Sources

In this step we will join the data sources using **Merge** Nodes.

1. Add a **Merge** node to the flow by clicking on the **Record Operations** menu in the Node Palette, and then dragging the **Merge** node to the right of the **fht_arrival_departure** data source. If the Node Palette is not visible, click on the Node Palette icon . Connect the **female_human_trafficking** data source to the Merge node. Connect the **fht_arrival_departure** data source to the **Merge** node. The canvas should appear as below.



2. Double-click on the **Merge** Node. Click on **MERGE**, then click on **Keys** for the Merge method, and click on **Add Columns**.



3. Click on the checkbox next to **INTERNAL_ID** and click **OK**.

Select Fields for Merge

Find in column

Field name

Filter:

#

abc

Reset

<div><input type="checkbox"/></div>	Field name	Schema name	Data type
<div><input checked="" type="checkbox"/></div>	INTERNAL_ID	female human trafficking	<div>#</div> integer
<div><input type="checkbox"/></div>	VETTING_LEVEL	female human trafficking	<div>#</div> integer
<div><input type="checkbox"/></div>	NAME	female human trafficking	<div>abc</div> string
<div><input type="checkbox"/></div>	GENDER	female human trafficking	<div>abc</div> string
<div><input type="checkbox"/></div>	BIRTH_DATE	female human trafficking	<div></div> date
<div><input type="checkbox"/></div>	BIRTH_COUNTRY	female human trafficking	<div>abc</div> string
<div><input type="checkbox"/></div>	BIRTH_COUNTRY_CODE	female human trafficking	<div>abc</div> string
<div><input type="checkbox"/></div>	OCCUPATION	female human trafficking	<div>abc</div> string
<div><input type="checkbox"/></div>	ADDRESS	female human trafficking	<div>abc</div> string
<div><input type="checkbox"/></div>	SSN	female human trafficking	<div>abc</div> string
<div><input type="checkbox"/></div>	PASSPORT_NUMBER	female human trafficking	<div>#</div> integer
<div><input type="checkbox"/></div>	PASSPORT_COUNTRY	female human trafficking	<div>abc</div> string
<div><input type="checkbox"/></div>	PASSPORT_COUNTRY_COC	female human trafficking	<div>abc</div> string

Cancel

OK

4. Scroll down in the **Merge** side panel that you have been working in. Select **Partial outer join** and then click on **Select Dataset for Outer Join**

Merge

Add columns +

☐Field name

☐INTERNAL_ID

☒Combine duplicate key fields ⓘ

Join ⓘ

Partial outer join

Select Dataset for Outer Join +


Filter

Optimization

Cancel

Save

5. Make sure the **female_human_trafficking** data source is checked. Click **OK**.

 Checked datasets will contribute incomplete records. If all datasets are checked, this becomes a full outer join.

OUTER JOIN	TAG	SOURCE NODE	CONNECTED NODE
<input checked="" type="checkbox"/>	1	female human traffi...	female human traffi...
<input type="checkbox"/>	2	fht_arrival_departure	fht_arrival_departure

Cancel

OK

6. Click **Save**.

Merge

Add columns +

<input type="checkbox"/>	Field name
<input type="checkbox"/>	INTERNAL_ID

☒ Combine duplicate key fields ⓘ

Join ⓘ


Partial outer join ▼

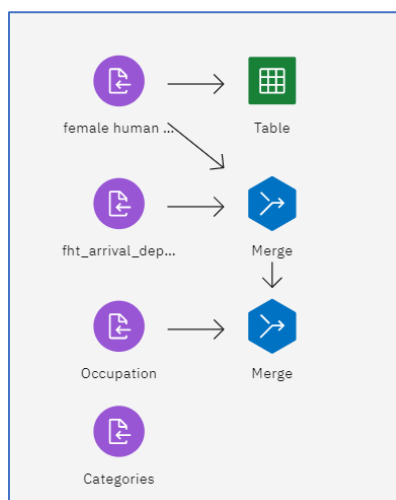
Select Dataset for Outer Join +

Filter ▼

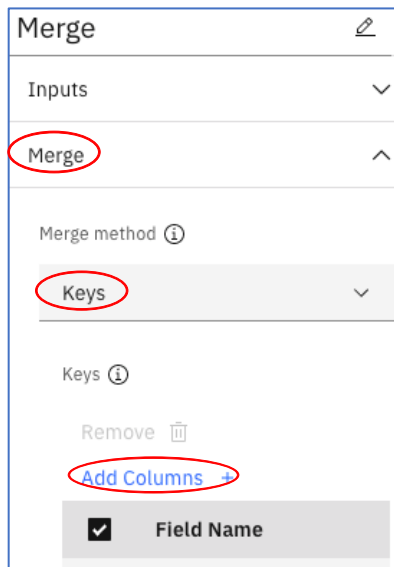
Optimization ▼

Cancel Save

7. Add a **Merge** node to the flow by clicking on the **Record Operations** menu in the Node Palette, and then dragging the **Merge** node to the right of the **Occupations** data source. If the Node Palette is not visible, click on the Node Palette icon . Connect the previous **Merge** node to this **Merge** node. Connect the **Occupations** data source to the **Merge** node. The canvas should appear as below.



8. Double-click on the **Merge** Node. Click on **Merge**, then click on **Keys** for the Merge method, and click on **Add Columns**.



Merge

Inputs

Merge

Merge method ⓘ

Keys

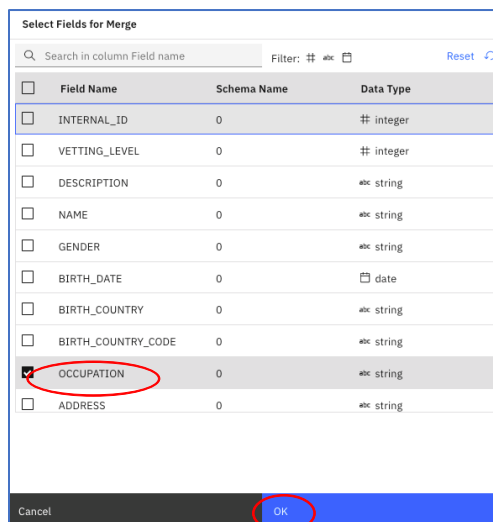
Keys ⓘ

Remove

Add Columns +

☒ Field Name

9. Click on **OCCUPATION** and then click on **Ok**. You may need to scroll down.




Select Fields for Merge

Search in column Field name Filter: # string Reset

<input type="checkbox"/>	Field Name	Schema Name	Data Type
<input type="checkbox"/>	INTERNAL_ID	0	# integer
<input type="checkbox"/>	VETTING_LEVEL	0	# integer
<input type="checkbox"/>	DESCRIPTION	0	# string
<input type="checkbox"/>	NAME	0	# string
<input type="checkbox"/>	GENDER	0	# string
<input type="checkbox"/>	BIRTH_DATE	0	date
<input type="checkbox"/>	BIRTH_COUNTRY	0	# string
<input type="checkbox"/>	BIRTH_COUNTRY_CODE	0	# string
<input checked="" type="checkbox"/>	OCCUPATION	0	# string
<input type="checkbox"/>	ADDRESS	0	# string

Cancel OK

10. Scroll down in the Merge side panel that you have been working in. Select **Partial outer join** and then click on **Select Dataset for Outer Join**.

Merge 

[Add Columns](#) +

☒ Field Name

☒ OCCUPATION


☒ Combine duplicate key fields ⓘ

Join ⓘ

Partial outer join ▼

[Select Dataset for Outer Join](#) ▶

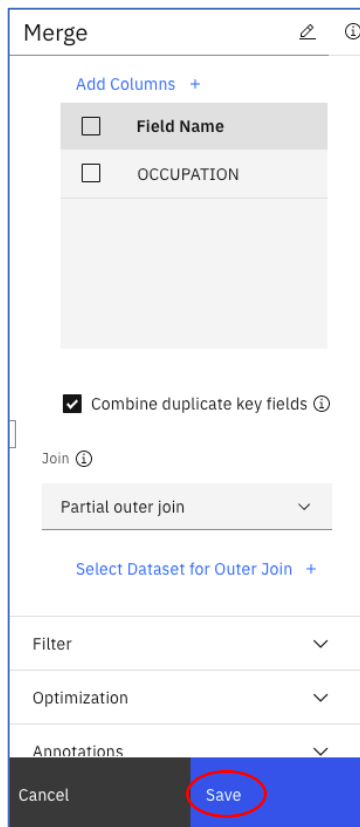
11. Make sure the Merge node is checked and click **OK**.

 Checked datasets will contribute incomplete records. If all datasets are checked, this becomes a full outer join.


OUTER JOIN	TAG	SOURCE NODE	CONNECTED NODE
<input checked="" type="checkbox"/>	1	Merge	Merge
<input type="checkbox"/>	2	Occupation	Occupation

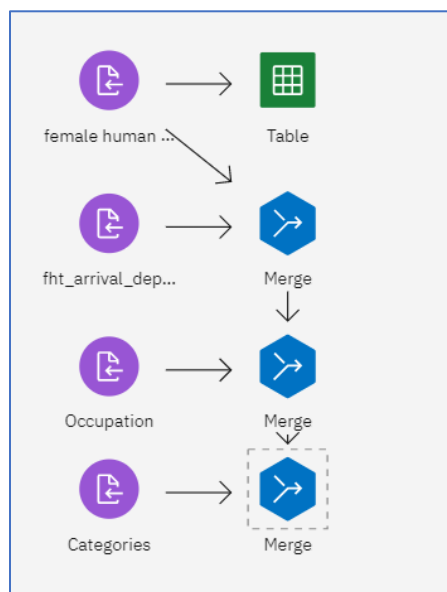
Cancel **OK**

12. Click on **Save**.



The screenshot shows the 'Merge' configuration window. At the top, there is a section 'Add Columns +' with a list of fields: 'Field Name' and 'OCCUPATION', each with an unchecked checkbox. Below this is a checked checkbox labeled 'Combine duplicate key fields'. Under the 'Join' section, 'Partial outer join' is selected from a dropdown menu. There is a link 'Select Dataset for Outer Join +'. At the bottom, there are expandable sections for 'Filter', 'Optimization', and 'Annotations'. The bottom bar contains a 'Cancel' button and a 'Save' button, which is circled in red.

13. Add a **Merge** node to the flow by clicking on the **Record Operations** menu in the Node Palette, and then dragging the **Merge** node to the right of the **Categories** data source. If the Node Palette is not visible, click on the Node Palette icon . Connect the prior **Merge** node to this **Merge** node. Connect the **Categories** data source to the **Merge** node. The canvas should appear as below.



14. Double click on the third **Merge** node to set the merge options. Click on **Merge**, click on **Keys** for the Merge method, and then click on **Add Columns** to add the key columns.

Merge

Inputs

Merge

Merge method ⓘ

Keys

Keys ⓘ

Remove

Add Columns +

☒ Field Name

15. Scroll down and click on the first **Code** checkbox. Click on **OK**.

Find in column Field name

Filter: # abc

Reset

<input type="checkbox"/>	Field name	Schema name	Data type
<input type="checkbox"/>	COUNTRIES_VISITED_COL	Merge	# integer
<input type="checkbox"/>	AGE	Merge	# integer
<input type="checkbox"/>	ARRIVAL_AIRPORT_REGI	Merge	abc string
<input type="checkbox"/>	DEPARTURE_AIRPORT_IA	Merge	abc string
<input type="checkbox"/>	ARRIVAL_AIRPORT_MUNI	Merge	abc string
<input type="checkbox"/>	ARRIVAL_AIRPORT_IATA	Merge	abc string
<input type="checkbox"/>	_ID	Merge	abc string
<input type="checkbox"/>	ARRIVAL_AIRPORT_COUN	Merge	abc string
<input type="checkbox"/>	DEPARTURE_AIRPORT_RE	Merge	abc string
<input type="checkbox"/>	DEPARTURE_AIRPORT_CC	Merge	abc string
<input type="checkbox"/>	DEPARTURE_AIRPORT_MI	Merge	abc string
<input checked="" type="checkbox"/>	Code	Merge	# integer
<input type="checkbox"/>	Code	Categories	# integer

Cancel

OK

16. Scroll down in the side panel and click on **Partial Outer Join** and click on **Select Dataset**.

Join ⓘ

Partial outer join ▼

Select Dataset for Outer Join +

17. Make sure the **Merge SOURCE NODE** is selected and click **Ok**.

Select Dataset for Outer Join

ⓘ Checked datasets will contribute incomplete records. If all datasets are checked, this becomes a full outer join.

OUTER JOIN	TAG	SOURCE NODE	CONNECTED NODE
<input checked="" type="checkbox"/>	1	Merge	Merge
<input type="checkbox"/>	2	Categories	Categories

Cancel OK

18. Click **Save**.

Merge ⓘ

Remove [X]

Add Columns +

☐ Field Name

☐ Code

☒ Combine duplicate key fields ⓘ

Join ⓘ

Partial outer join ▼

Select Dataset for Outer Join +


Filter ▼

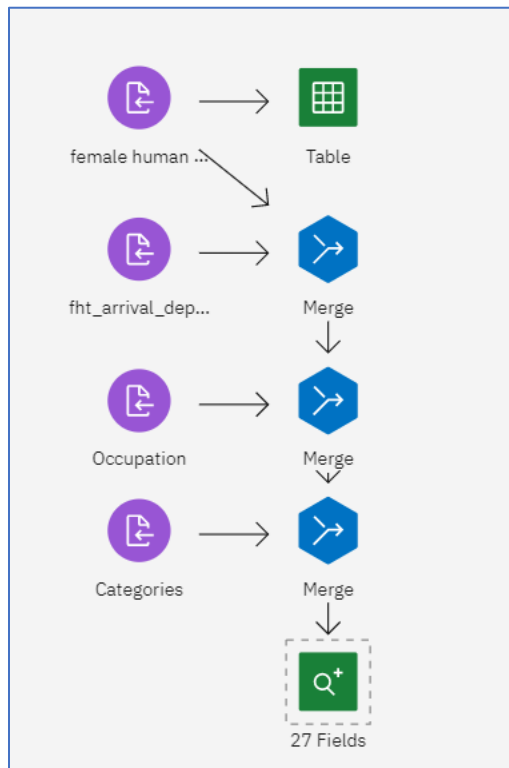
Optimization ▼

Cancel Save

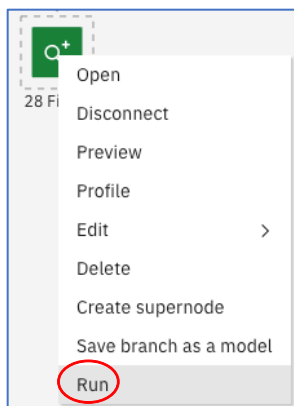
Step 4 - Explore the Data using the Data Audit Node


The SPSS Modeler has a Data Audit node that provides profiling information on the input data that is useful for cleansing and preparing the data. It provides a comprehensive first look at the data, including summary statistics, as well as information about outliers, missing values, and extremes.

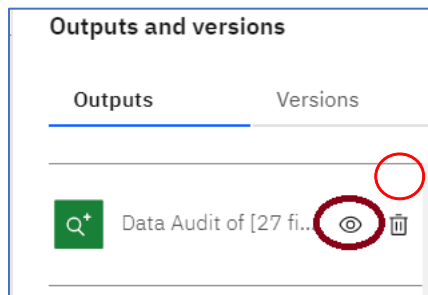
1. Add a **Data Audit** node to the flow by clicking on the **Outputs** menu item in the Node Palette, and then dragging the **Data Audit** node to underneath the final **Merge** node. If the Node Palette is not visible, click on the Node Palette icon . Connect the node to the Data Audit node. The canvas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



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









- 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 clock icon. **Click on the  icon to display the contents.**



- The top section of the Data Audit report displays profiling information. For modeling purposes, fields that have only 1 unique value, or have many unique values should be eliminated. In addition, certain fields are directly related such as PASSPORT_COUNTRY, PASSPORT_COUNTRY_CODE, BIRTH_COUNTRY, and BIRTH_COUNTRY_CODE. Only one of these fields need to be retained. The fields that we will keep for modeling purposes are VETTING_LEVEL, Category, AGE, COUNTRIES_VISITED_COUNT, ARRIVAL_AIRPORT_REGION, DEPARTURE_AIRPORT_COUNTRY, PASSPORT_COUNTRY. Later in the lab we will apply a filter operation to retain these fields.

View Output: Data Audit of [27 fields]

	Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
1	Code		Continuous	1	15	7.950	4.238	0.263	--	1085
2	OCCUPATION		Categorical	--	--	--	--	--	--	1085
3	INTERNAL_ID		Continuous	1	1085	543.000	313.357	-0.000	--	1085
4	VETTING_LEVEL		Continuous	10	100	80.498	34.211	-1.216	--	1085
5	NAME		Categorical	--	--	--	--	--	--	1085
6	GENDER		Categorical	--	--	--	--	--	1	1085


- Scroll down** to view the bottom section. It displays data quality checks in the form of missing values or anomalous values. In our travel data simulator, we didn’t simulate any of those type of values! Click the **x** to close the window and return to the flow.

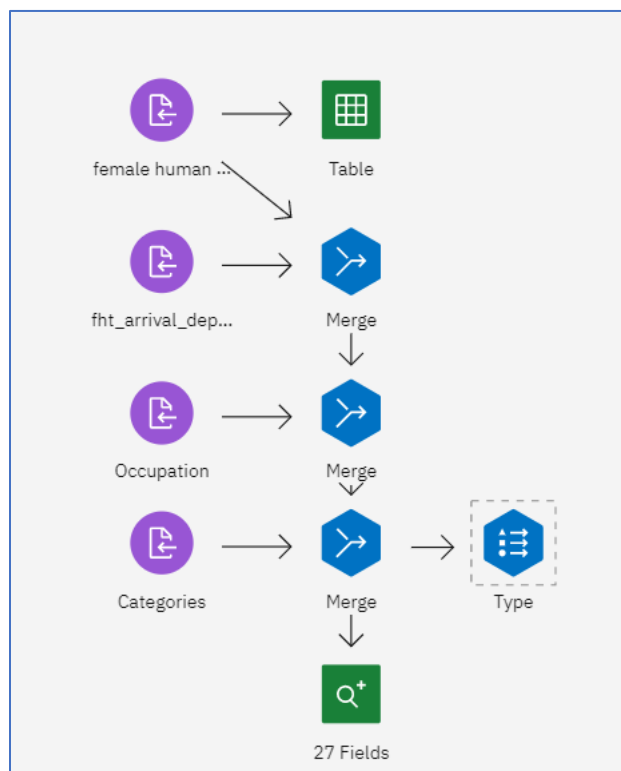
View Output: Data Audit of [27 fields]

	Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String
1	Code	Continuous	0	0	None	Never	Fixed	100.000	1085	0	0
2	OCCUPATION	Categorical	--	--	--	Never	Fixed	100.000	1085	0	0
3	INTERNAL_ID	Continuous	0	0	None	Never	Fixed	100.000	1085	0	0
4	VETTING_LEVEL	Continuous	0	0	None	Never	Fixed	100.000	1085	0	0
5	NAME	Categorical	--	--	--	Never	Fixed	100.000	1085	0	0
6	GENDER	Categorical	--	--	--	Never	Fixed	100.000	1085	0	0
7	BIRTH_DATE	Continuous	0	0	None	Never	Fixed	100.000	1085	0	0

Step 5 - 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 trafficking data. 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 "Code" and "VETTING_LEVEL" fields that were derived as "Continuous" (by scanning the data values) to "Nominal".

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 to the right of the third **Merge** node. If the Node Palette is not visible, click on the Node Palette icon . Connect the **Merge** 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. Select the dropdown in the **Measure** column next to **Code**. Change the **Measure** to **Nominal**.

Type

Settings

Default Mode ⓘ
☒ Read metadata ☐ Pass (do not scan)
Type Operations

Read Values

Clear All Values

Search in column Field

<input type="checkbox"/>	Field	Measure	Role	Value Mode	Values	Check
<input type="checkbox"/>	# Code	Continuous	Input	Read		None
<input type="checkbox"/>	abc OCCUPAT	Categorical	Input	Read		None
<input type="checkbox"/>	# INTERNAL	Continuous	Input	Read		None
<input type="checkbox"/>	# VETTING_	Continuous	Input	Read		None
<input type="checkbox"/>	abc DESCRIPT	Categorical	Input	Read		None

Type

Settings

Default Mode ⓘ
☒ Read metadata ☐ Pass (do not scan)
Type Operations

Read Values

Clear All Values

Search in column Field

<input type="checkbox"/>	Field	Measure	Role	Value Mode	Values	Check
<input type="checkbox"/>	# Code	Continuous	Input	Read		None
<input type="checkbox"/>	abc OCCUPAT	Categorical	Input	Read		None
<input type="checkbox"/>	# INTERNAL	Continuous	Input	Read		None
<input type="checkbox"/>	# VETTING_	Continuous	Input	Read		None
<input type="checkbox"/>	abc DESCRIPT	Categorical	Input	Read		None

3. Following the same process, change the **Measure** of VETTING_LEVEL to **Nominal**.

Type

Settings

Default Mode ⓘ

☒ Read metadata ☐ Pass (do not scan)

Type Operations

Read Values

Clear All Values

Search in column Field

<input type="checkbox"/>	Field	Measure		Role		Value Mode		Values		Check
<input type="checkbox"/>	# Code	Nominal	▼	Input	▼	Specify	▼	1, 2, 3, 4, 5, 6, 7, ...		None ▼ ⚙
<input type="checkbox"/>	abc OCCUPAT:	Typeless	▼	None	▼	Specify	▼			None ▼ ⚙
<input type="checkbox"/>	# INTERNAL	Continuous	▼	Input	▼	Specify	▼	1, 1085		None ▼ ⚙
<input type="checkbox"/>	# VETTING_	Nominal	▼	Input	▼	Specify	▼	10, 100		None ▼ ⚙

4. Click **Read Values** and click **Save**.

Type ✎ ⓘ

Settings ^

Default Mode ⓘ
☒ Read metadata ☐ Pass (do not scan)

Type Operations v

Read Values Clear All Values

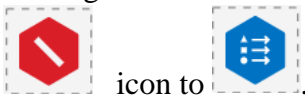
🔍 Search in column Field







<input type="checkbox"/>	Field	Measure		Role		Value Mode		Values	Check	
<input type="checkbox"/>	# Code	Nominal	v	Input	v	Specify	v	1, 2, 3, 4, 5, 6, 7, ...	None	v ⚙
<input type="checkbox"/>	abc OCCUPAT	Typeless	v	None	v	Specify	v		None	v ⚙
<input type="checkbox"/>	# INTERNAL	Continuous	v	Input	v	Specify	v	1, 1085	None	v ⚙
<input type="checkbox"/>	# VETTING_	Nominal	v	Input	v	Specify	v	10, 100	None	v ⚙
<input type="checkbox"/>	abc DESCRIP1	Flag	v	Input	v	Specify	v	NA	None	v ⚙
<input type="checkbox"/>	abc NAME	Typeless	v	None	v	Specify	v		None	v ⚙
<input type="checkbox"/>	abc GENDER	Flag	v	Input	v	Specify	v	F	None	v ⚙
<input type="checkbox"/>	📅 BIRTH_D/	Continuous	v	Input	v	Specify	v	1970-01-03, 200...	None	v ⚙


Format v

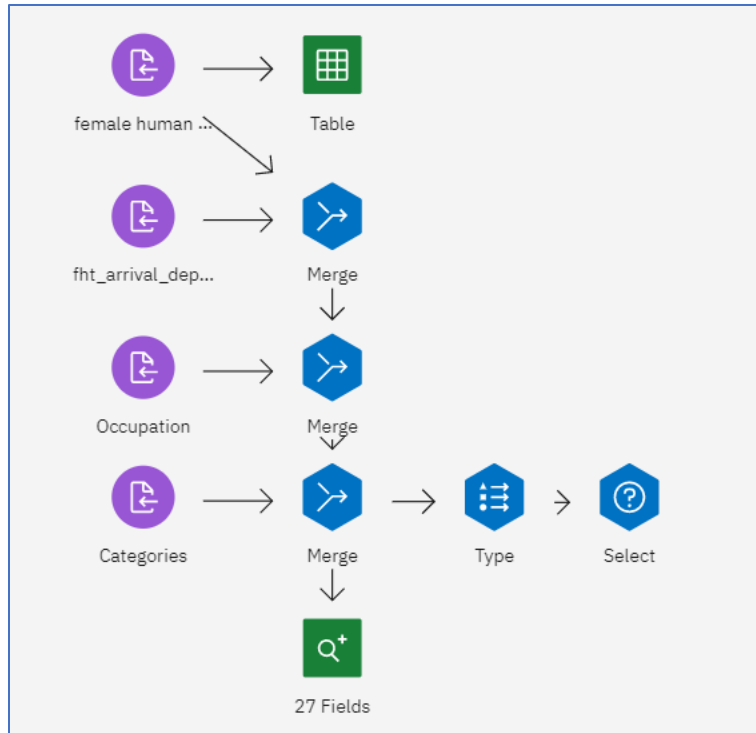
Cancel Save

5. Informational messages are displayed as some fields are set to Typeless measure given the large number of distinct values. Click **Clear all**. The Type node will change from



Messages	
Last run was now	
	Node: Type Large set type field 'OCCUPATION' has changed to typeless
	Node: Type Large set type field 'NAME' has changed to typeless
	Node: Type Field 'GENDER' has only one value
	Node: Type Large set type field 'ADDRESS' has changed to typeless
	Node: Type Large set type field 'SSN' has changed to typeless
	Node: Type Large set type field 'COUNTRIES_VISITED' has changed to typeless
Clear all	

6. We will now discard the unvetted records. Add a **Select** node to the flow 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 **Type** node. If the Node Palette is not visible, click on the Node Palette icon . The canvas should appear as below.



7. Double-click the **Select** node. Click on **Discard** for **Mode**. In the Condition, enter **VETTING_LEVEL==100**, click **Save**.

Select

Settings

Mode

☐ Include


☒ Discard

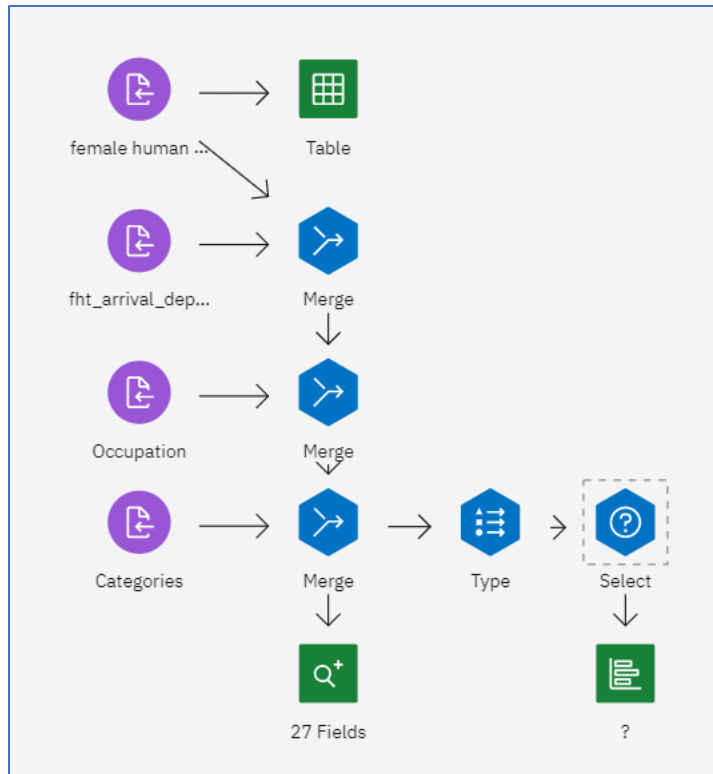
Condition

VETTING_LEVEL==100

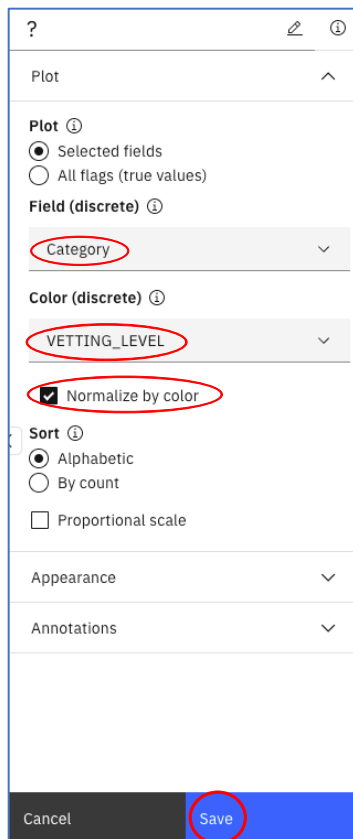
Annotations

Cancel Save

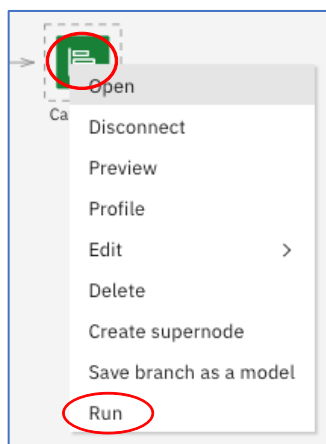
8. Add a **Distribution** node to the flow by clicking on the **Graph** menu item and then dragging the **Distribution** node to the canvas underneath the **Select** node. If the Node Palette is not visible, click on the Node Palette icon . Connect the **Select** node to the **Distribution** node. The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





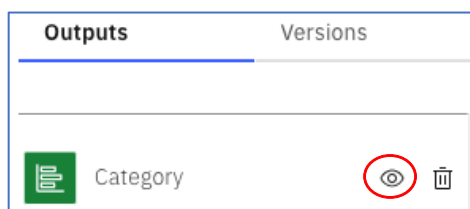
9. Double click on the Distribution Node. In the **Field (discrete)** dropdown, select **Category**. In the **Color (discrete)** dropdown, select **VETTING_LEVEL**. Click on the **normalize by color** checkbox, and then click **Save**.



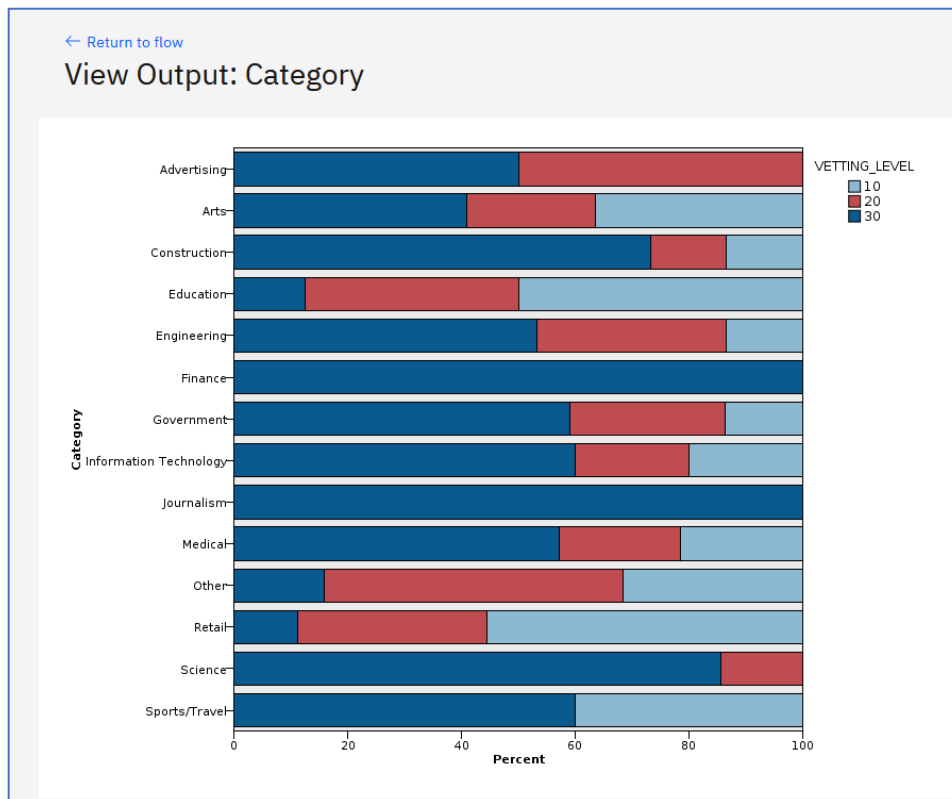
10. Right click on the Distribution node and select **Run**.



11. The Distribution output will appear under the **Outputs** tab. Click on the  icon. If you don't see the Outputs, click on the Outputs ("clock")  icon.




12. We can see from the graph that the VETTING_LEVEL does differ based on Category.



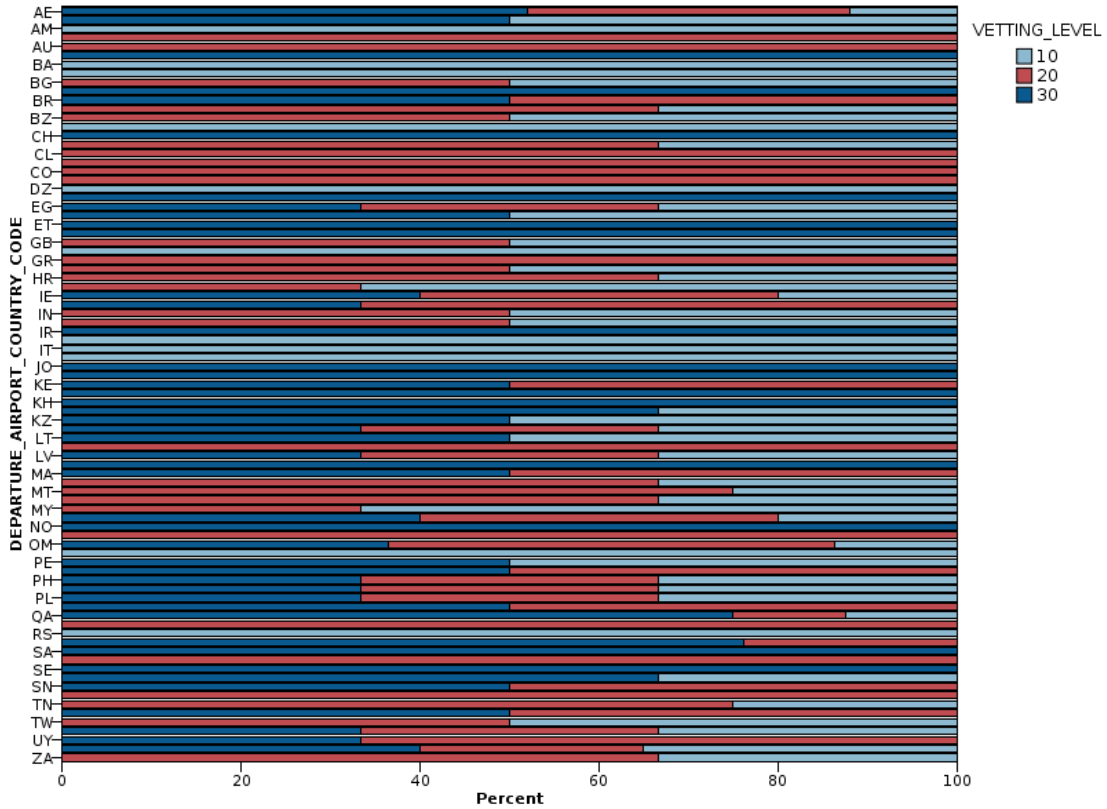
13. Return to the flow by clicking on the FemaleHumanTrafficking breadcrumb at the top or click the close (“x”) icon.

[My Projects](#) / [Watson Studio Labs](#) / [FemaleHumanTrafficking](#) / [Category](#)

14. You can change the distribution graph to show the **VETTING_LEVEL** by **DEPARTURE_AIRPORT_COUNTRY_CODE** by double clicking on the Distribution node and replacing **Category** with **DEPARTURE_AIRPORT_COUNTRY_CODE** and clicking Save. Re-run the graph by right clicking on the Distribution node and selecting **Run**. Click on the  icon adjacent to the **DEPARTURE_AIRPORT_COUNTRY_CODE** in the **Outputs** pane to display the graph.


[← Return to flow](#)

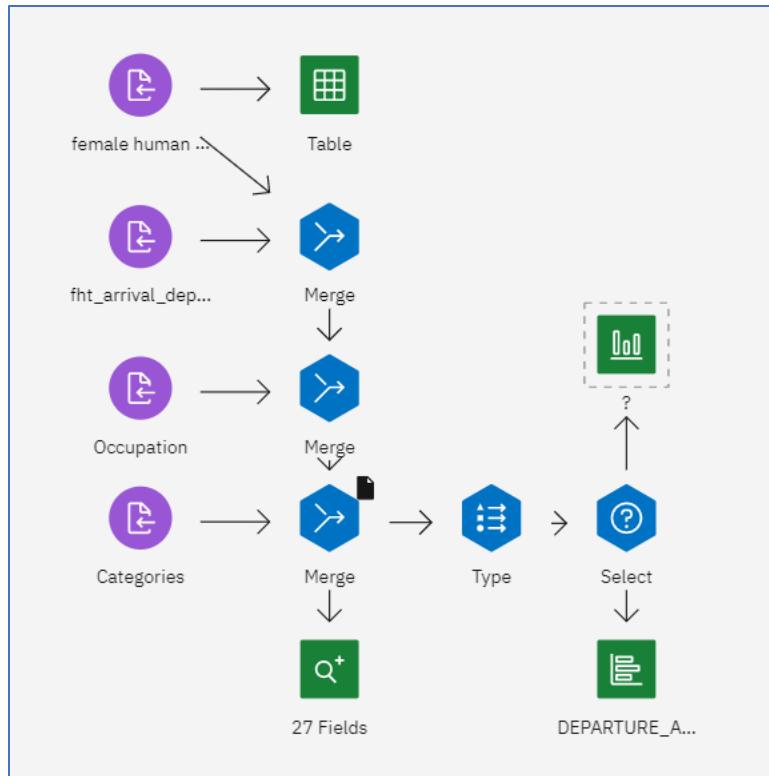
View Output: DEPARTURE_AIRPORT_COUNTRY_CODE



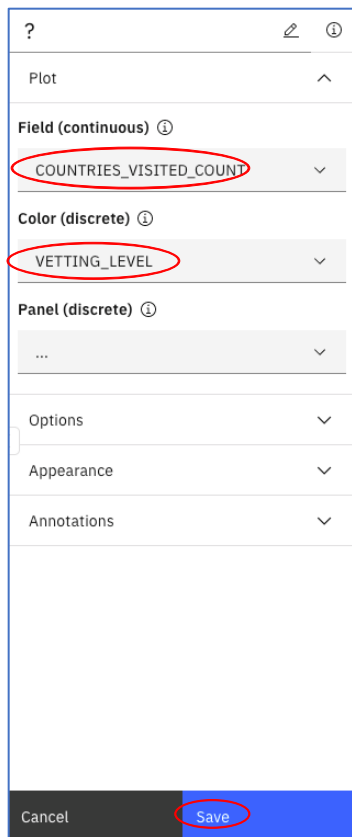
15. Return to the flow by clicking on the FemaleHumanTrafficking breadcrumb at the top or click on the close (“x”) icon.

[My Projects](#) / [Watson Studio Labs](#) / [FemaleHumanTrafficking](#)

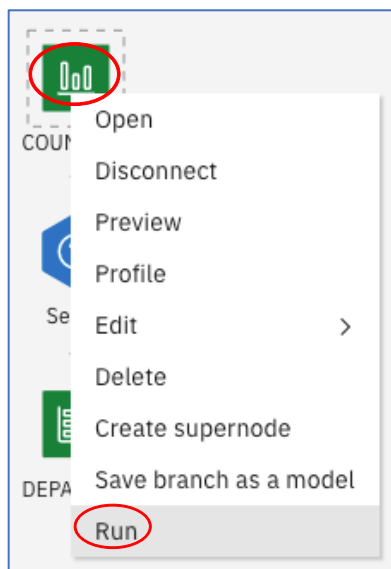
16. Add a **Histogram** node to the flow by clicking on the **Graphs** menu item and then dragging the **Histogram** node to the canvas above the **Select** node. If the Node Palette is not visible, click on the Node Palette icon . Connect the **Select** node to the **Histogram** node. The canvas should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




17. Double click on the **Histogram** node. Select **COUNTRIES_VISITED_COUNT** from the Field (continuous) dropdown. Select **VETTING_LEVEL** from the Color (discrete) dropdown. Click on **Save**.




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




19. Click on the  icon adjacent to the **COUNTRIES_VISITED_COUNT** under the **Outputs** tab at the right of the screen.


Outputs

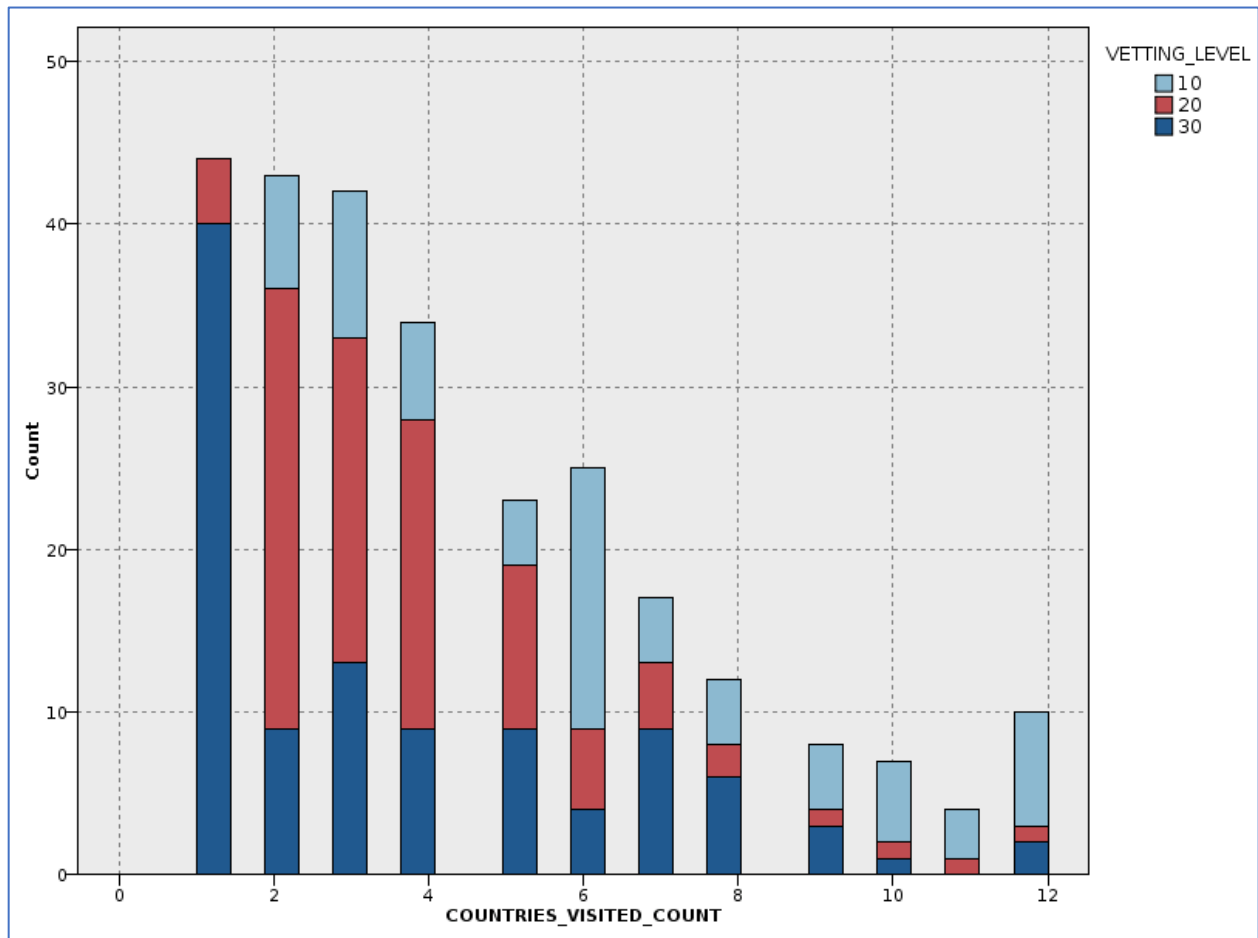
Versions




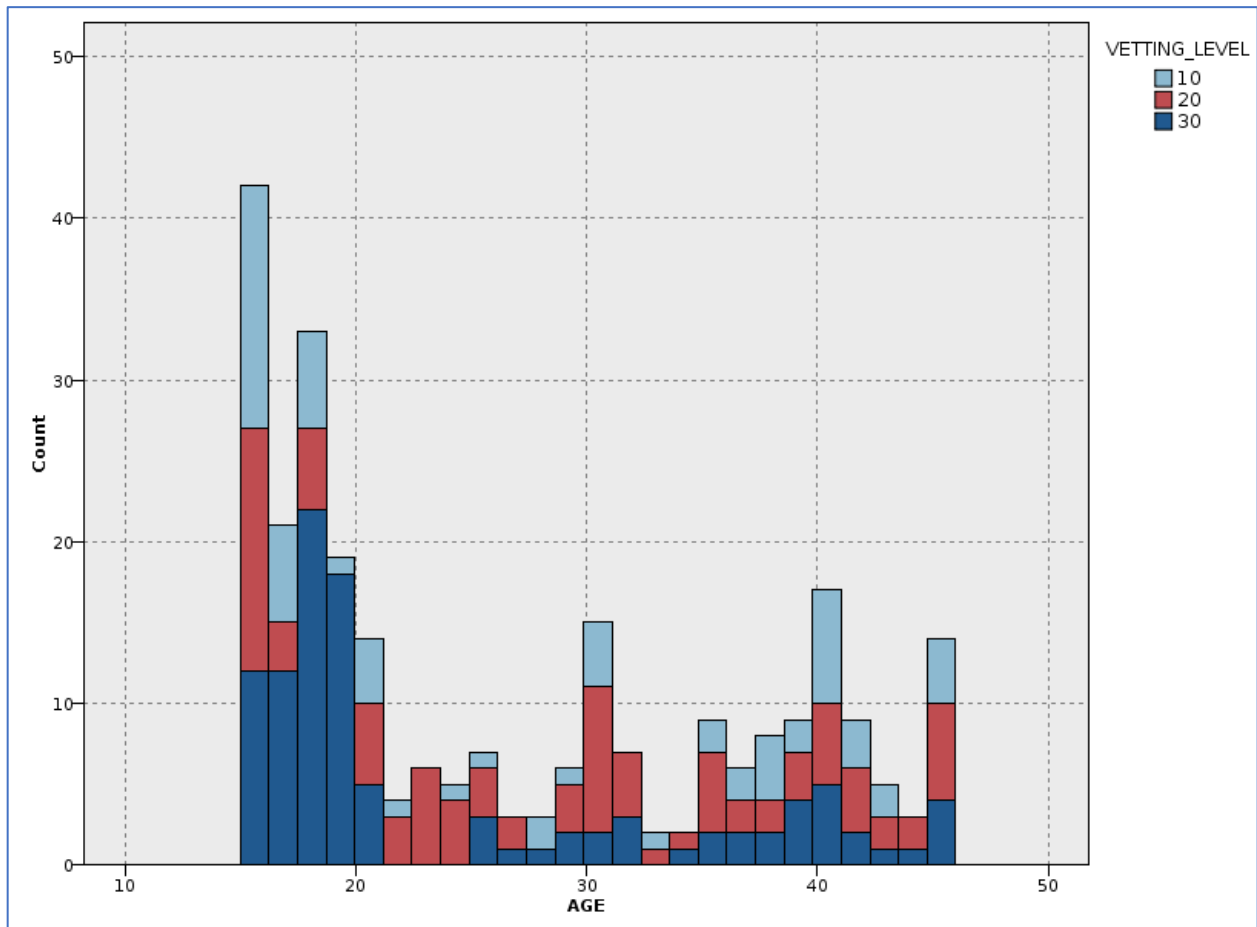
COUNTRIES_VISITED_C...








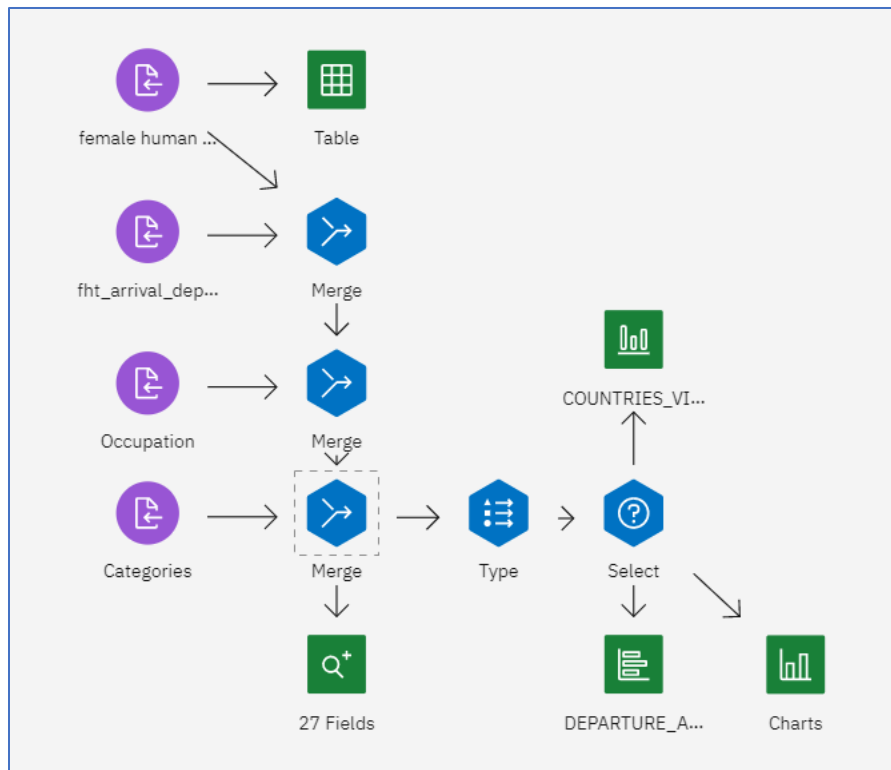
20. The general trend appears to be that the more countries visited, the higher likelihood to be a “High Risk”. Close the display by clicking on the close (“x”) icon. You can change the histogram to show the **AGE** by **VETTING_LEVEL** by double clicking on the Histogram node and replacing **COUNTRIES_VISITED_COUNT** with **AGE** and clicking **Save**. Re-run the graph by right clicking on the **Histogram** node and selecting **Run**. Click on the  icon adjacent to the **AGE** in the **Outputs** pane to display the graph. Close the display by clicking on the close (“x”) icon.



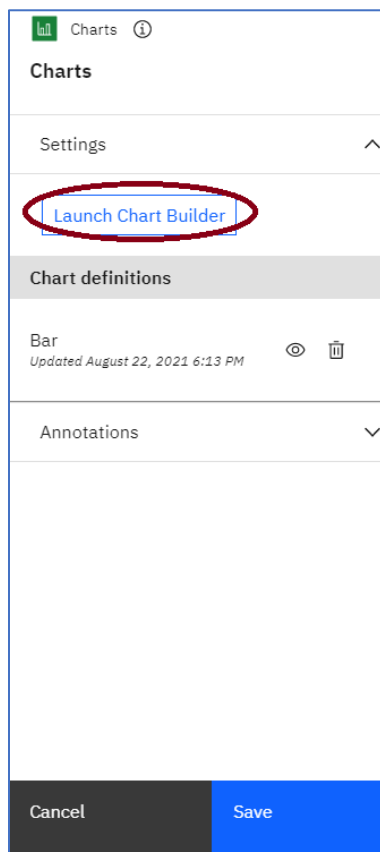
21. A relatively recent addition to the Graphs palette has been the Charts component



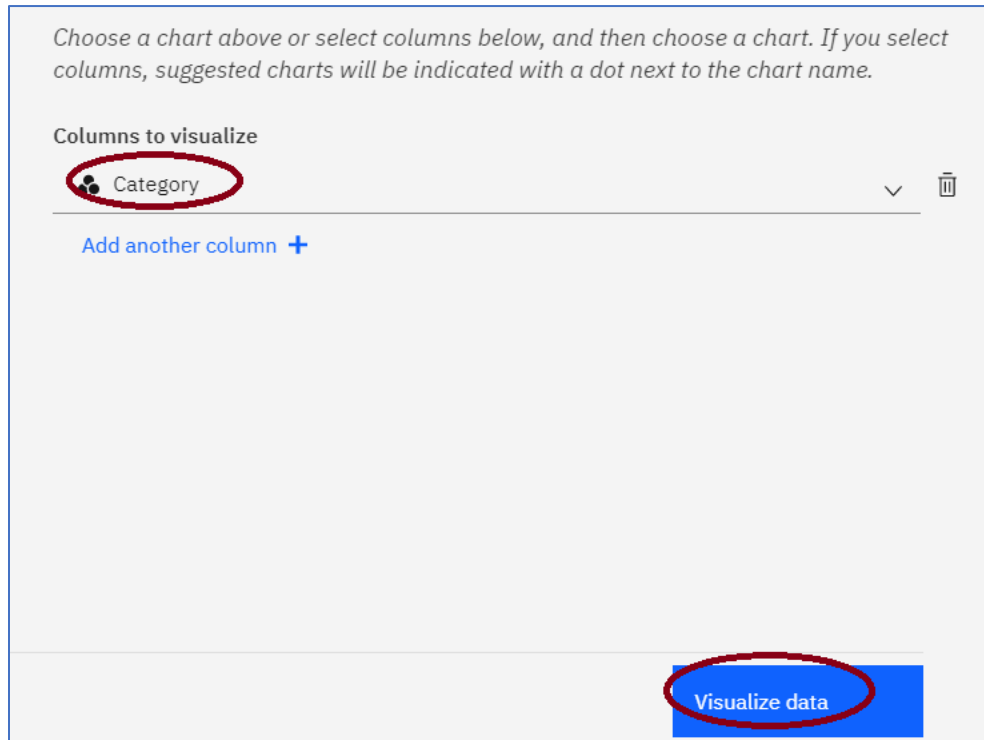
This facility has been architected to be embedded in tooling applications. It is the same facility that drives the Visualization in the Data Refinery. Add a **Chart** node to the flow by clicking on the **Graphs** menu item and then dragging the **Chart** node to the canvas to the right of the **Distribution** node. If the Node Palette is not visible, click on the Node Palette icon . Connect the **Select** node to the **Chart** node. The canvas should appear as below.



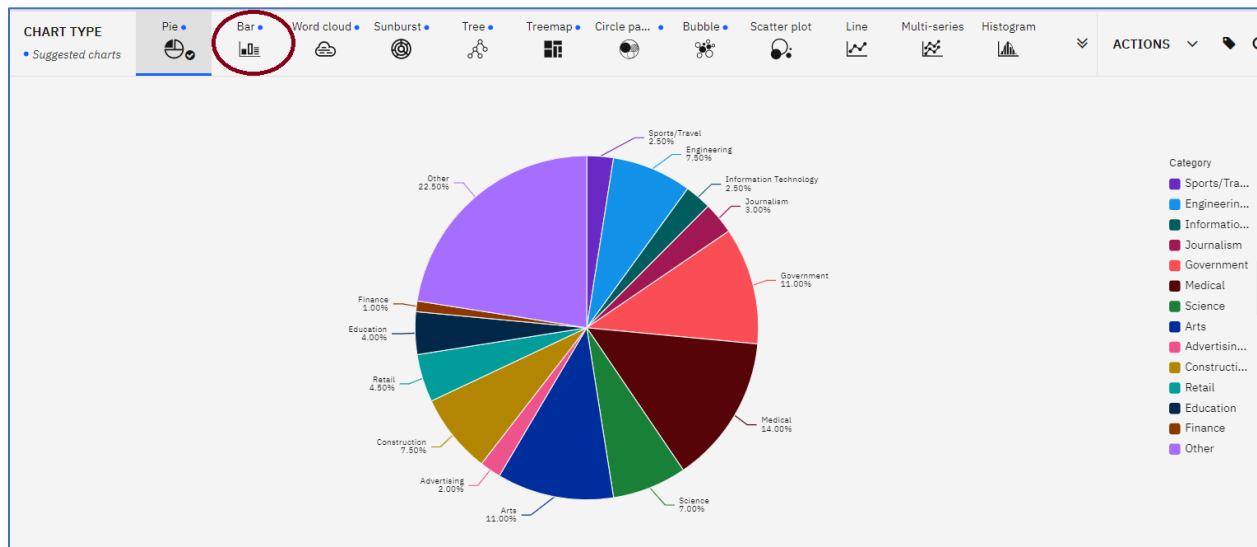
22. Double-click on the **Charts** node. Click on **Launch Chart Builder**.



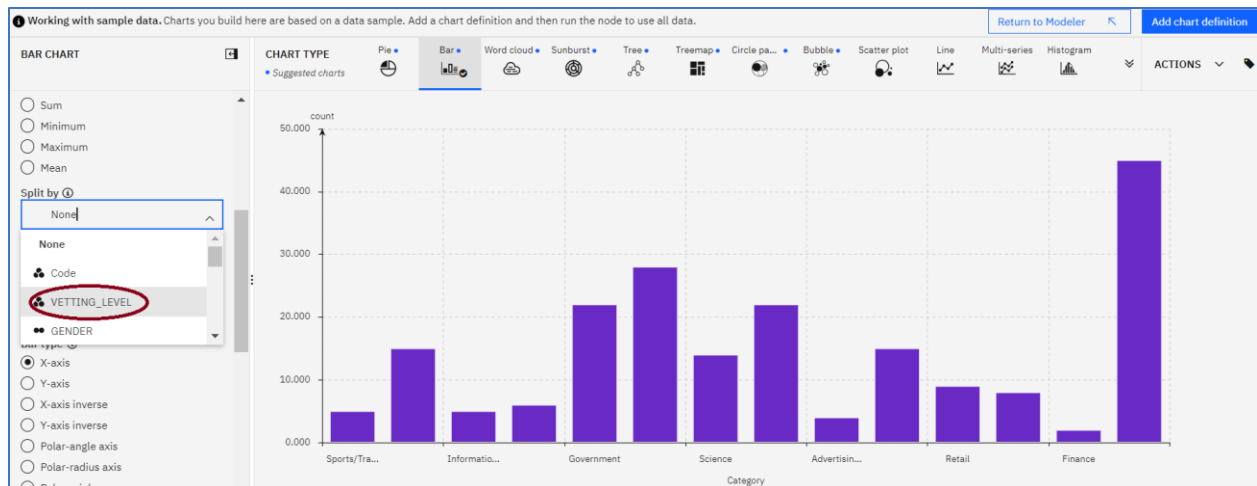
23. You are now in the Charts facility. Select **Category** as the **Columns to visualize**.



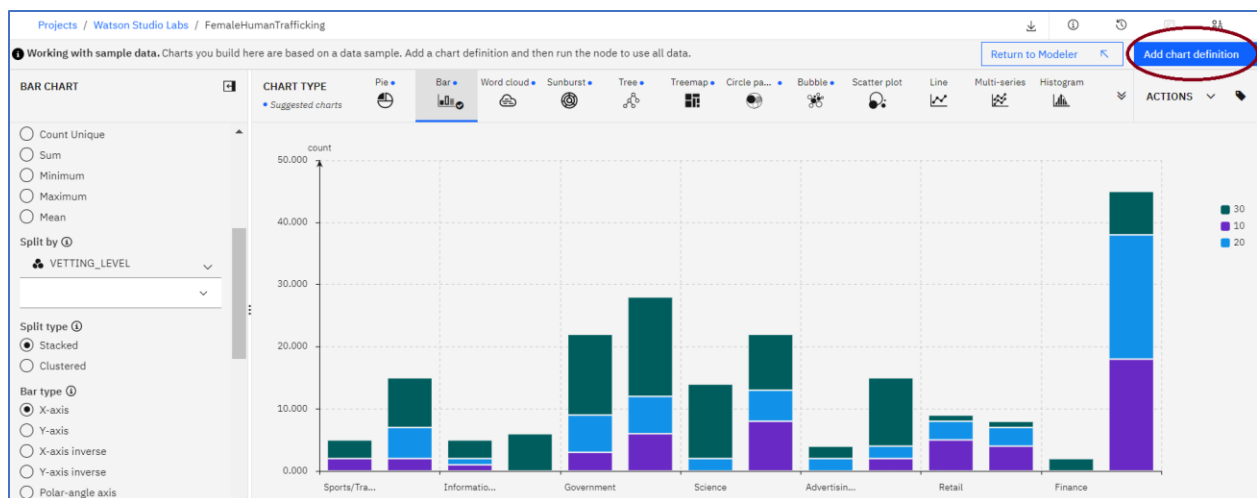
24. A pie chart is chosen by the tool to display the Category data. Click on **Bar**.



25. Click on **VETTING_LEVEL** for the **Split by**.



26. You can add the chart definition to be run in the modeler flow. Click on **Add chart definition**.



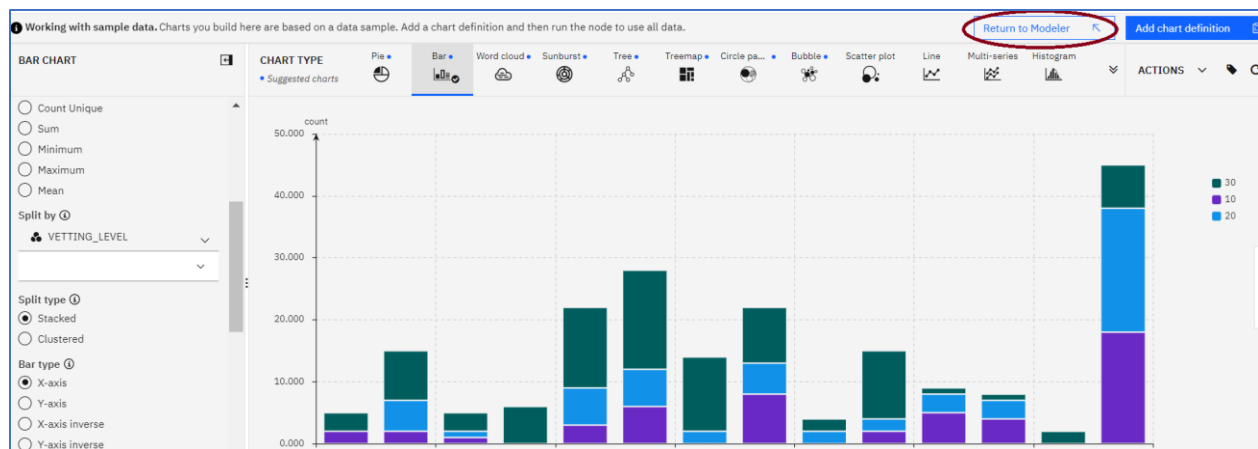
27. Click **Yes** to add the chart definition to the chart node.

Add chart definition

Do you want to add this chart definition to the node?

No Yes

28. You can continue to visualize the data and add charts to the modeler flow. We will Return to Modeler now. Click **Return to Modeler**.



29. Click **Yes**.

Return to Modeler

Do you want to exit Chart Builder and return to your flow?

No Yes

30. We can see one bar chart has been retained. Click **Save**.

Charts

Charts

Settings

Launch Chart Builder

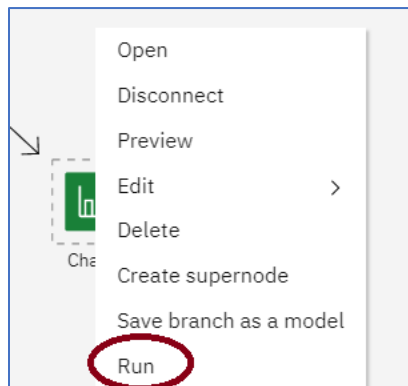
Chart definitions


Bar
Updated August 22, 2021 6:52 PM

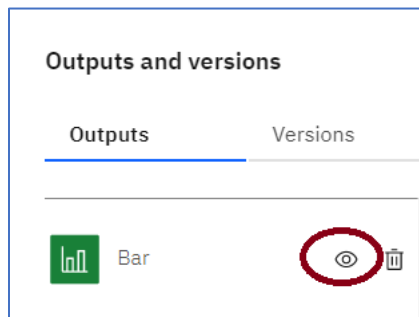
Annotations

Cancel Save

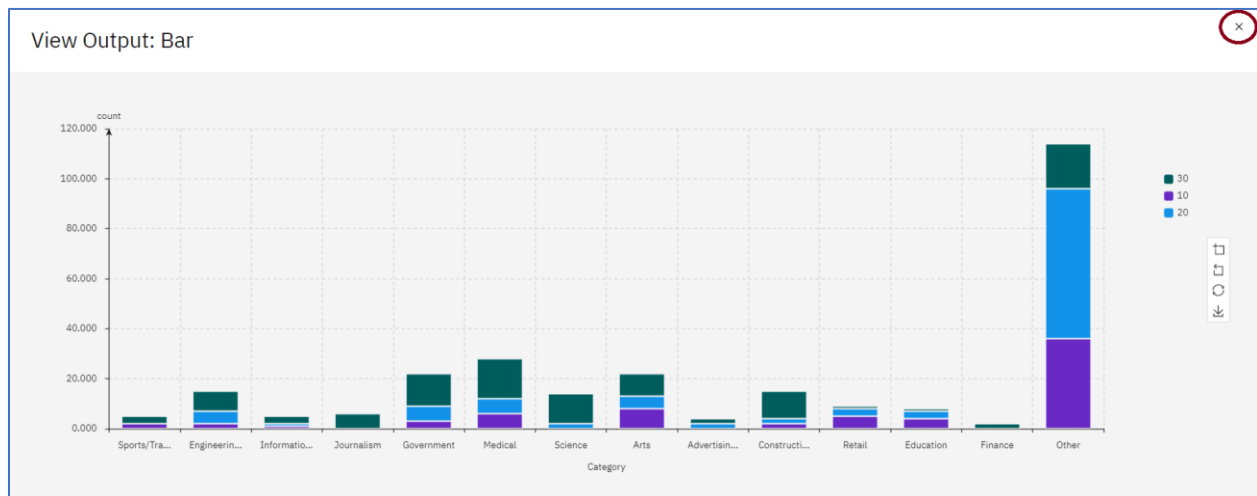
31. Right-click on the **Chart** node, and click **Run**.



32. Click on the  icon adjacent to the **Bar** under the **Outputs** tab at the right of the screen



33. The visualization is displayed. Note, when running the visualization through the modeler flow it will be using all the data, not just a sample as was used in the Charts facility. In this case it didn't make a difference since the sample size is larger than the rows in the dataset.




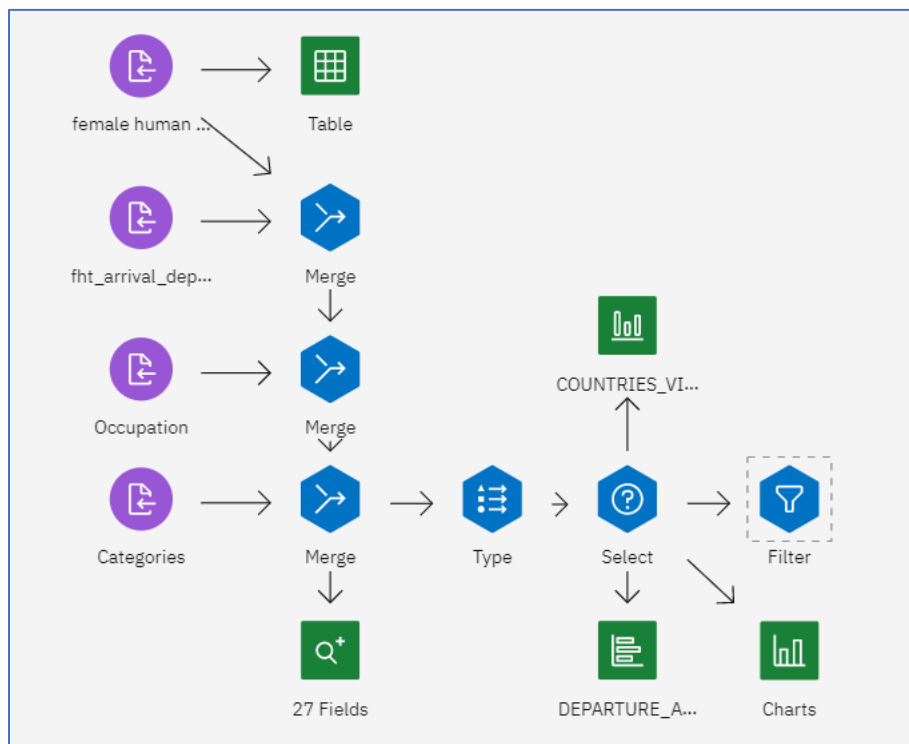
Step 6 - Prepare the Data for Modeling

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 and the **Reclassify** node that will do the necessary transformations. The **Filter** and **Reclassify** nodes act on a field level.

Filter node – The **Filter** node performs two functions. It specifies fields that can be dropped or the fields that should be retained. It also allows fields to be renamed. We will retain the following fields – VETTING_LEVEL, COUNTRIES_VISITED_COUNT, ARRIVAL_AIRPORT_REGION, DEPARTURE_AIRPORT_COUNTRY_CODE, AGE, and Category.

Reclassify node – The **Reclassify** node allows us to map input values to output values. We will use this node to map the VETTING_LEVEL values of 10, 20, 30, and 100 to “High Risk”, “Medium Risk”, “Low Risk”, and “Unvetted” respectively.

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



2. Double-click on the **Filter** node. Click **Retain the selected ...**, and click **Add Column**.

Filter

Mode ⓘ
☐ Filter the selected fields
☒ Retain the selected fields (all other fields are filtered)

Filter Options

Select Fields ⓘ

Remove Add Columns

3. Scroll down and click on VETTING_LEVEL, PASSPORT_COUNTRY, COUNTRIES_VISITED_COUNT, ARRIVAL_AIRPORT_REGION, DEPARTURE_AIRPORT_COUNTRY_CODE, AGE, and CATEGORY, then click **OK**. Scroll as required to check all of the above fields.

Select Fields for Filter

Find in column Field name

Filter: # abc

Reset

<input type="checkbox"/>	Field name	Data type
<input type="checkbox"/>	Code	# integer
<input type="checkbox"/>	OCCUPATION	abc string
<input type="checkbox"/>	INTERNAL_ID	# integer
<input checked="" type="checkbox"/>	VETTING_LEVEL	# integer
<input type="checkbox"/>	NAME	abc string
<input type="checkbox"/>	GENDER	abc string
<input type="checkbox"/>	BIRTH_DATE	date
<input type="checkbox"/>	BIRTH_COUNTRY	abc string
<input type="checkbox"/>	BIRTH_COUNTRY_CODE	abc string
<input type="checkbox"/>	ADDRESS	abc string
<input type="checkbox"/>	SSN	abc string
<input type="checkbox"/>	PASSPORT_NUMBER	# integer
<input checked="" type="checkbox"/>	PASSPORT_COUNTRY	abc string

Find in column

Field name

Filter: # abc

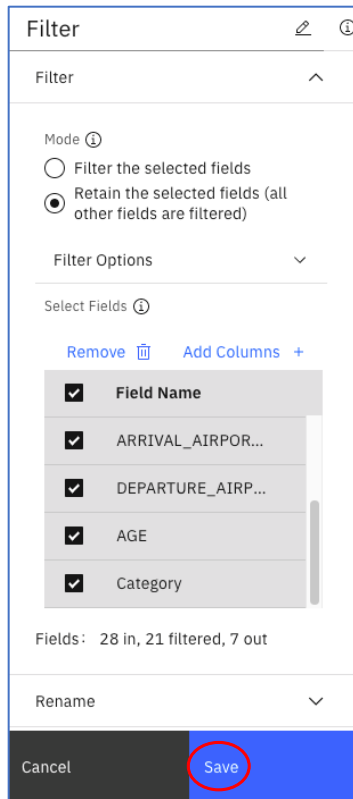
Reset


<input type="checkbox"/>	Field name	Data type
<input type="checkbox"/>	COUNTRIES_VISITED	abc string
<input checked="" type="checkbox"/>	COUNTRIES_VISITED_COUNT	# integer
<input checked="" type="checkbox"/>	AGE	# integer
<input checked="" type="checkbox"/>	ARRIVAL_AIRPORT_REGION	abc string
<input type="checkbox"/>	DEPARTURE_AIRPORT_IATA	abc string
<input type="checkbox"/>	ARRIVAL_AIRPORT_MUNICIPALITY	abc string
<input type="checkbox"/>	ARRIVAL_AIRPORT_IATA	abc string
<input type="checkbox"/>	_ID	abc string
<input type="checkbox"/>	ARRIVAL_AIRPORT_COUNTRY_CODE	abc string
<input type="checkbox"/>	DEPARTURE_AIRPORT_REGION	abc string
<input checked="" type="checkbox"/>	DEPARTURE_AIRPORT_COUNTRY_CODE	abc string
<input type="checkbox"/>	DEPARTURE_AIRPORT_MUNICIPALITY	abc string
<input checked="" type="checkbox"/>	Category	abc string

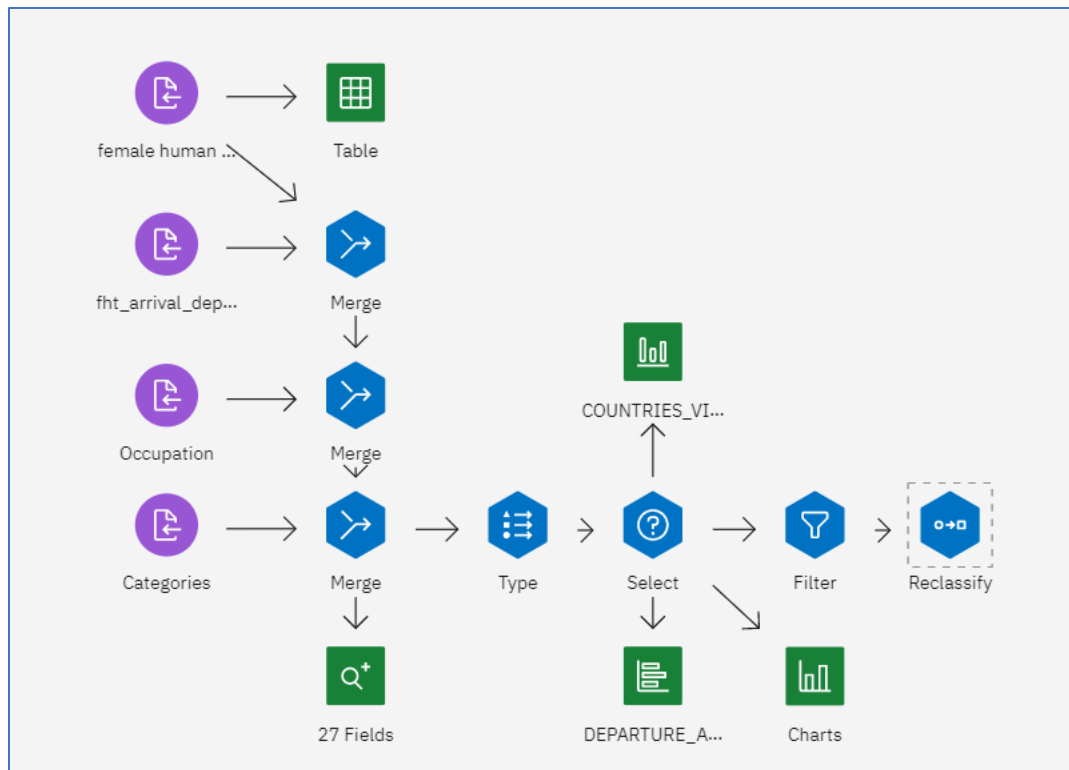
Cancel

OK

4. Click **Save**.



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



- Double-click on the **Reclassify** node. Configure the **Reclassify** node as follows. Select **VETTING_LEVEL** for the **Reclassify** field. Enter **VETTING_LEVEL_DESC** for the **New Field Name**. Click **Get values**.

Reclassify

Settings

Mode ⓘ
☒ Single
☐ Multiple

Reclassify Into ⓘ
☒ New field
☐ Existing field

Reclassify Field ⓘ
VETTING_LEVEL

New Field Name ⓘ
VETTING_LEVEL_DESC

Get values Copy Clear new

- Scroll down. Enter in “High Risk” as the new value for “10”, “Medium Risk” as the new value for “20”, “Low Risk” as the new value for “30”, and “Unvetted” as the new value for “100”. Click on **Save**.

Values ⓘ

Remove Add Value +

<input type="checkbox"/>	ORIGINAL VALUE	NEW VALUE
<input type="checkbox"/>	10	High Risk
<input type="checkbox"/>	20	Medium Risk
<input type="checkbox"/>	30	Low Risk
<input type="checkbox"/>	100	Unvetted


For Unspecified Values Use ⓘ
☒ Original value ☐ Default value
undef

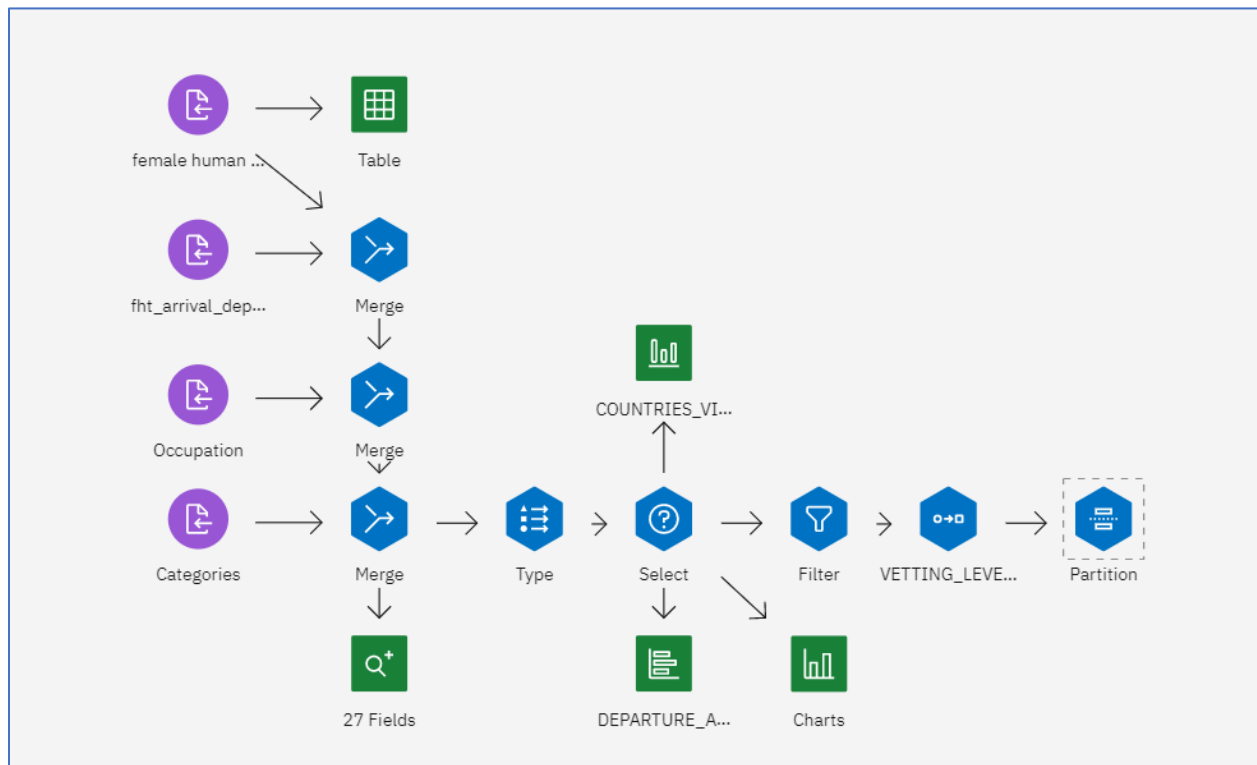
Annotations

Cancel Save

Step 7 - 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 set the roles of the data fields. Then we will add several modeling nodes and use the Training set to train the model. Finally, we will add **Analysis** nodes to evaluate the results.

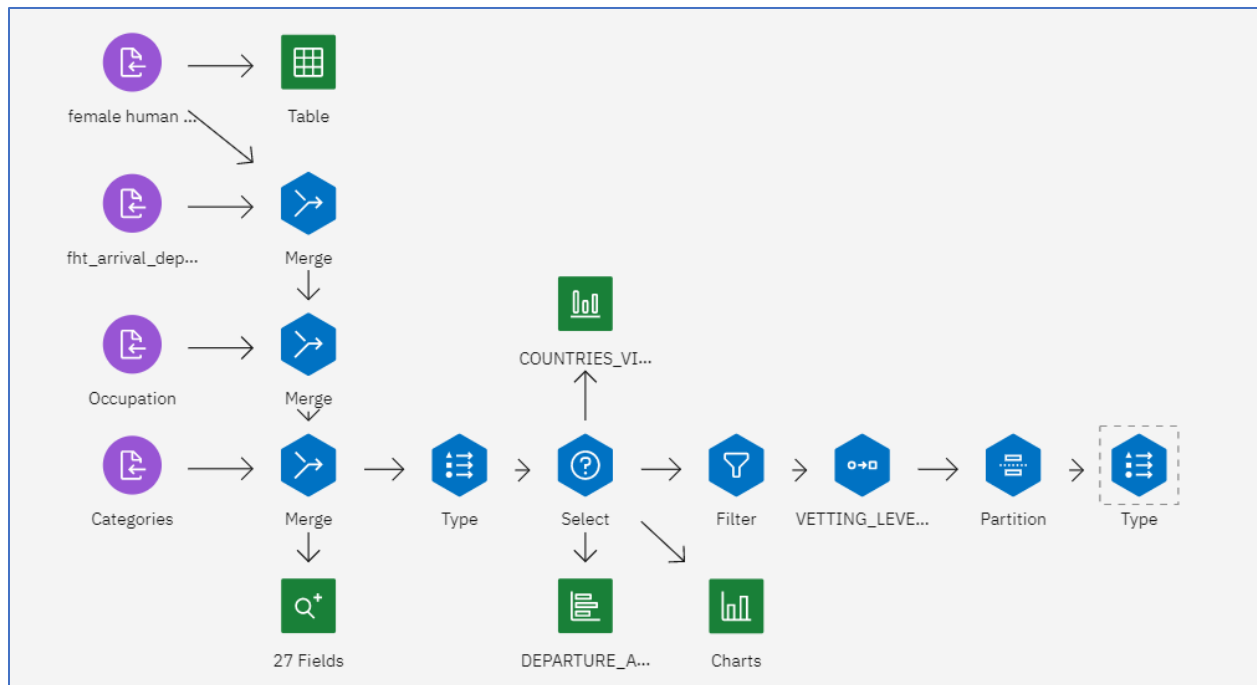
1. Add a **Partition** node to the canvas by clicking on the **Field Operations** menu item in the Node Palette, and then dragging the **Partition** node onto the canvas to the right of the **Reclassify** node. If the Node Palette is not visible, click on the Node Palette icon . Connect the **Reclassify**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



2. Double-click on the **Partition** node. Use a 70-30 breakdown between training and testing. Leave the other defaults and click **Save**.

The image shows the 'Partition' settings dialog box. At the top, there's a title bar with 'Partition' and icons for edit and help. Below is a 'Settings' section with an expand/collapse arrow. The 'Derived Field Name' is set to 'Partition'. The 'Training Partition(%)' is set to 70, and the 'Testing Partition(%)' is set to 30. There are checkboxes for 'Create validation partition' (unchecked) and 'Repeatable partition assignment' (checked). A 'Seed' field is set to 1234567, with a 'Generate' link and a help icon. There's also a checkbox for 'Use unique field to assign partitions' (unchecked). At the bottom, there's an 'Annotations' section with a collapse arrow. The bottom bar has 'Cancel' and 'Save' buttons, with 'Save' highlighted by a red circle.

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



4. Double-click on the **Type** Node. Click on **Read Values**.

Type
Settings
Default Mode ⓘ <input checked="" type="radio"/> Read metadata <input type="radio"/> Pass (do not scan)
Type Operations
<div>Read Values</div> <div>Clear All Values</div>

5. Hover over the Field name to see the full name. Change the role of **VETTING_LEVEL** to None. Change role of **VETTING_LEVEL_DESC** to **Target**. Click **Save**.

Type

Settings

Default Mode ⓘ
☒ Read metadata ☐ Pass (do not scan)

Type Operations

Read Values Clear All Values

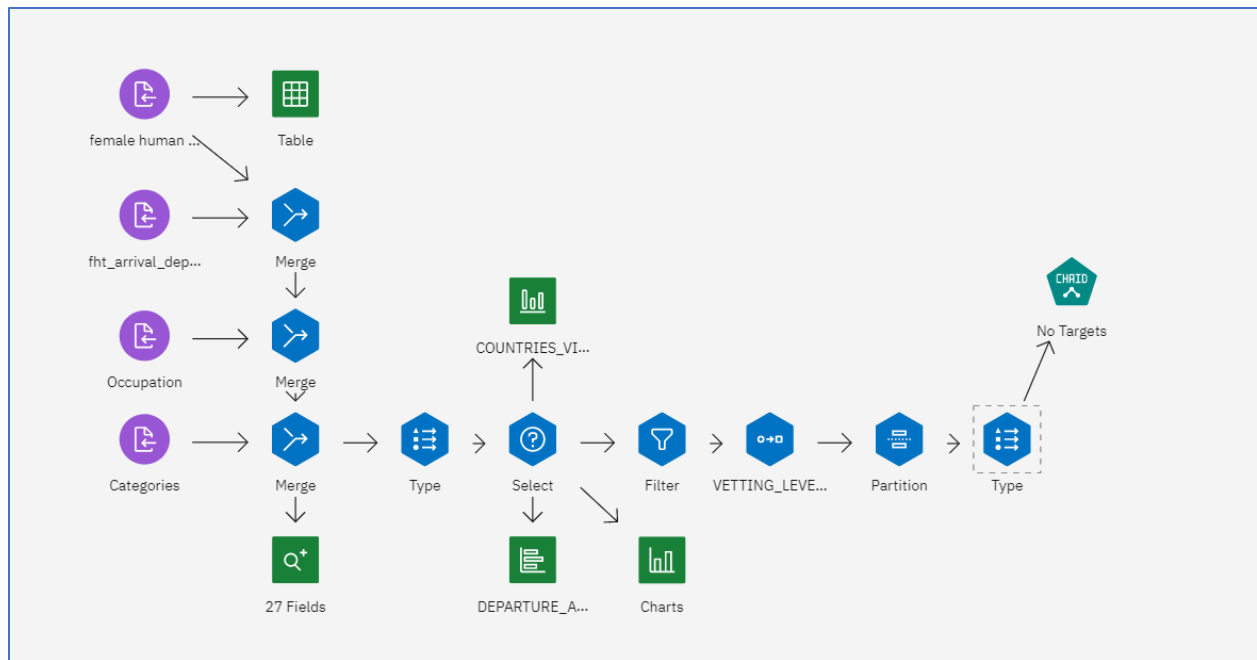
Search in column Field

<input type="checkbox"/>	Field	Measure	Role	Value Mode	Values	Check
<input type="checkbox"/>	# VETTING_	Nominal	None	Specify	10, 100	None
<input type="checkbox"/>	abc PASSPOR	Nominal	Input	Specify	Bangladesh, Brazi...	None
<input type="checkbox"/>	# COUNTRII	Continuous	Input	Specify	1, 12	None
<input type="checkbox"/>	abc ARRIVAL_	Nominal	Input	Specify	US-AK, US-AL, US...	None
<input type="checkbox"/>	abc DEPARTUI	Nominal	Input	Specify	AE, AL, AM, AR, A...	None
<input type="checkbox"/>	# AGE	Continuous	Input	Specify	15, 47	None
<input type="checkbox"/>	VETTING_LEVEL_DESC	Nominal	Input	Specify	Advertising, Arts, ...	None
<input type="checkbox"/>	abc VETTING_	Nominal	Target	Specify	High Risk, Low Ris...	None

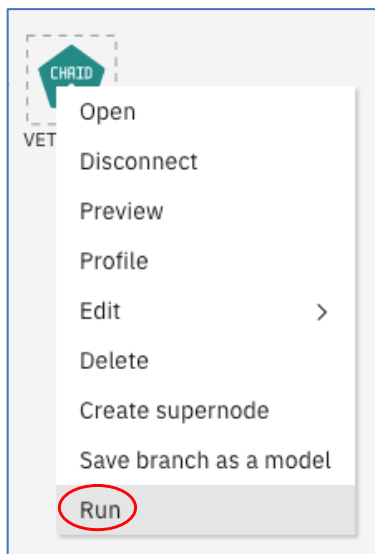
Format

Cancel Save

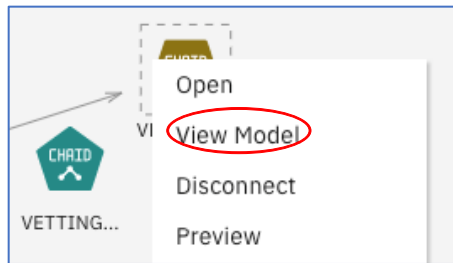
6. Add a **CHAID** node by clicking on the **Modeling** menu item in the Node palette and dragging the **CHAID** node onto the canvas to the right of the **Type** node. Connect the **Type** node to the **CHAID** node. The canvas should appear as below.



7. Right-click on the CHAID node and click Run.



8. A **Model** node is created. Drag the **Model** node to the right of the **CHAID** node. Right-click on the **Model** node and click **View Model**.



9. Note your results could differ. The Model Information is displayed. Click on **Feature Importance**.

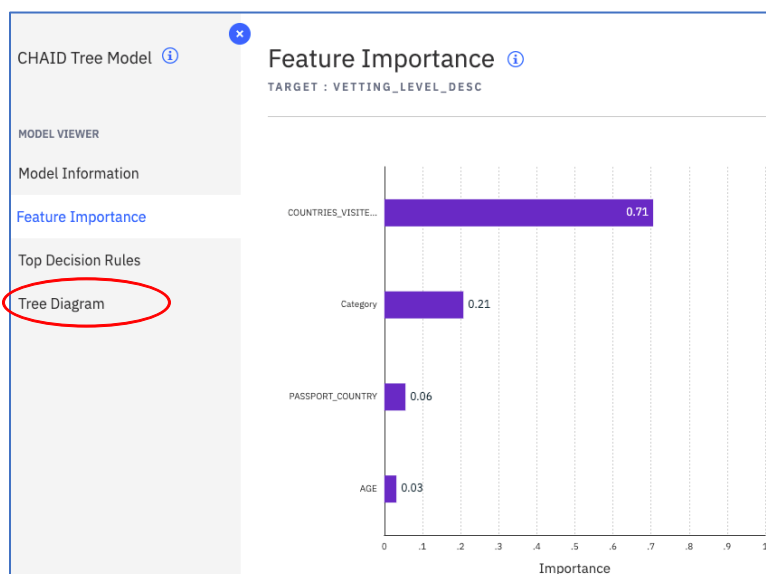
CHAID Tree Model ⓘ
MODEL VIEWER
Model Information
Feature Importance
Top Decision Rules
Tree Diagram

Model Information ⓘ

TARGET : VETTING_LEVEL_DESC

Target Field	VETTING_LEVEL_DESC
Model Type	Multi-Class Decision Tree
Algorithm Name	CHAID
Number of Features	4
Tree Depth	5
Number of Nodes	19

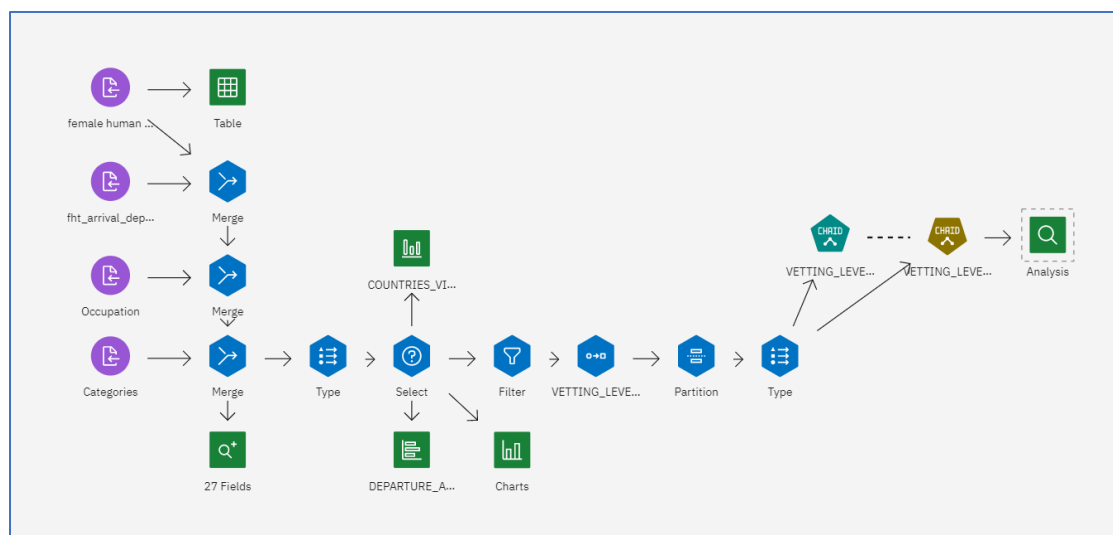
10. Feature Importance is displayed. Click on **Tree Diagram** and/or **Top Decision Rules** to see the algorithm output.



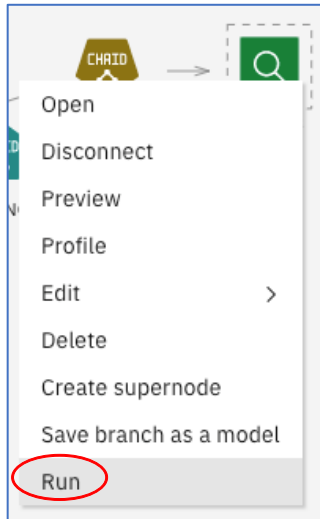
11. The Top Decision Rules are shown below.


Top Decision Rules ⓘ					
TARGET : VETTING_LEVEL_DESC					
Rule ID	Rule	Mode category	Record count	Record percentage	Rule confidence
12	((COUNTRIES_VISITED_COUNT > 2.0 and COUNTRIES_VISITED_COUNT <= 5.0)) and ((Category = Arts or Category = Construction or Category = Engineering or Category = Government or Category = Information Technology or Category = Journalism or Category = Medical or Category = Retail or Category = Science or Category = Sports/Travel))	Low Risk	38	20.765	47.368
7	((COUNTRIES_VISITED_COUNT <= 1.0)) and ((DEPARTURE_AIRPORT_COUNTRY_CODE = AE or DEPARTURE_AIRPORT_COUNTRY_CODE = QA or DEPARTURE_AIRPORT_COUNTRY_CODE = RU))	Low Risk	24	13.115	100.000
22	((COUNTRIES_VISITED_COUNT > 2.0 and COUNTRIES_VISITED_COUNT <= 5.0)) and ((Category = Other)) and ((AGE > 21.0))	Medium Risk	24	13.115	100.000
	((COUNTRIES_VISITED_COUNT > 1.0 and COUNTRIES_VISITED_COUNT <= 2.0)) and ((Category = Advertising or Category = Education or Category = Entertainment or Category = Food or Category = Health or Category = Home or Category = Information Technology or Category = Insurance or Category = Law or Category = Life or Category = Medical or Category = Music or Category = News or Category = Real Estate or Category = Retail or Category = Science or Category = Sports/Travel or Category = Technology or Category = Travel or Category = Transportation or Category = Utilities or Category = Other))				

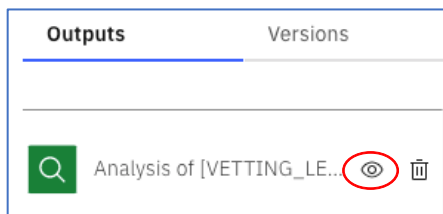
- Close the display by clicking on the close (“x”) icon. Add an Analysis node by clicking on the Output menu item in the Node palette and dragging the Analysis node onto the canvas to the right of the CHAID Model node. Connect the CHAID Model node to the Analysis node. The canvas should appear as below. The canvas should appear as below.



8. Right-click the **Analysis** node and click **Run**.



9. Click on the  icon adjacent to the Analysis results in the Output area.



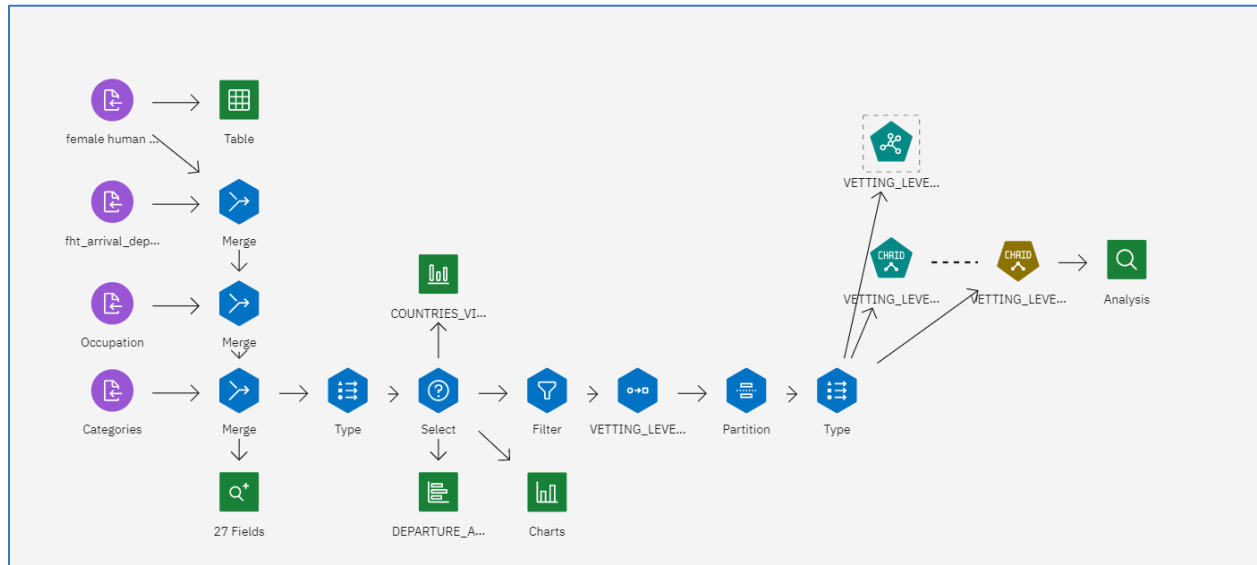
10. Accuracy results are displayed for the CHAID algorithm. Close the display by clicking on the close (“x”) icon.

Results for output field VETTING_LEVEL_DESC

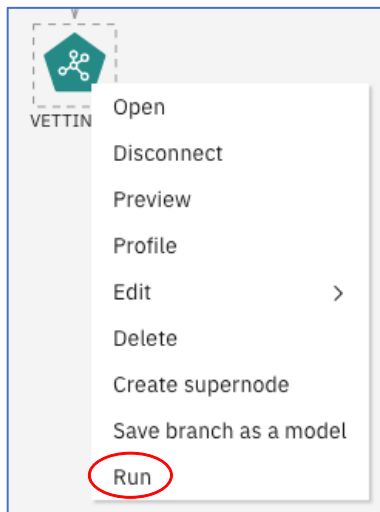
Comparing \$R-VETTING_LEVEL_DESC with VETTING_LEVEL_DESC

Partition'	1_Training		2_Testing	
Correct	142	77.6%	61	70.93%
Wrong	41	22.4%	25	29.07%
Total	183		86	

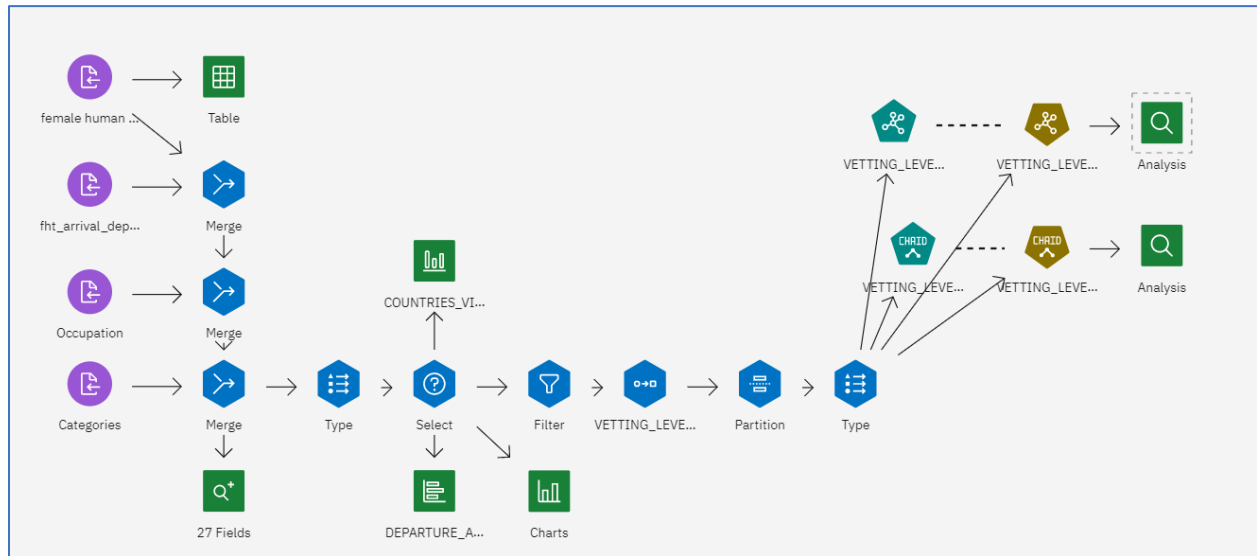
6. Add a **Random Forest** node by clicking on the **Modeling** menu item in the Node palette and dragging the **Random Forest** node onto the canvas above the **CHAID** node. Connect the **Type** node to the **Random Forest** node. The canvas should appear as below.



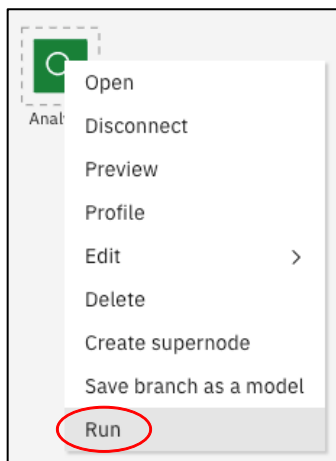
7. Right-click the **Random Forest** node and click **Run**.



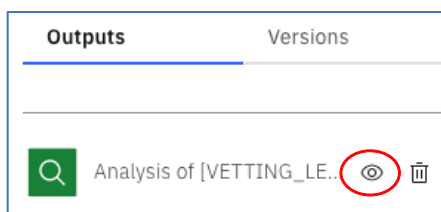
8. A **Random Forest Model** node is created. Clear the messages if displayed. Drag the generated **Random Forest Model** node to the right of the **Random Forest** node. The **Random Forest Model** node does not have a **View Model** option. Add an **Analysis** node to the right of the **Random Forest Model** node by clicking on the **Outputs** menu of the Node Palette. Connect the **Analysis** node to the **Random Forest Model** node. The canvas should appear as shown below.



9. Right-click on the **Analysis** node and click **Run**.



10. The **Analysis** node output appears in the **Outputs** area. Double-click **Analysis of ...**



11. The results appear below. Based on the results, the Random Forest model is overfitting given the disparity between training and testing results. Click on the close (“x”) icon to remove the display.

Results for output field VETTING_LEVEL_DESC

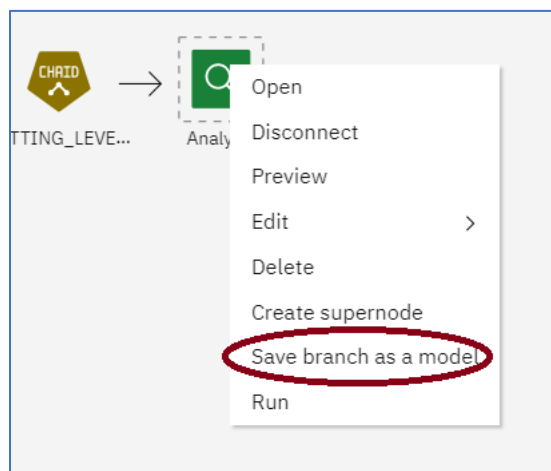
Comparing \$R-VETTING_LEVEL_DESC with VETTING_LEVEL_DESC

'Partition'	1_Training		2_Testing	
Correct	168	91.8%	66	76.74%
Wrong	15	8.2%	20	23.26%
Total	183		86	

Step 8 - 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. Since the Random Forest model was badly overfitting the data, we choose to save the CHAID model (also overfitted but not as badly). Note, that we need to collect more data to avoid the overfitting.

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



2. Type in “**FHT_SPSS**” as the Model Name, optionally add a **Description**, and click **Save**.

Save model

Saving mode
☒ Scoring branch ☐ Individual algorithm as PMML

Branch terminal node
Analysis

Multiple data sources are supported for batch deployments, but not for real-time deployments.

Model name
FHT_SPSS

Model description (optional)
Chaid Model

Cancel Save

3. Click **Close**.

Success

✓ Successfully saved model FHT_SPSS You can now prepare the asset for deployment.

Close

4. Navigate to your project “assets” page. Click on **Watson Studio Labs**.

My Projects Watson Studio Labs FemaleHumanTrafficking

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.

✓ Models				
Watson Machine Learning models				
Name	Type	Software specification	Last modified	↓
FHT_SPSS	spss-modeler_18.2	spss-modeler_18.2	Jan 10, 2021	

You have completed Lab-4!

- ✓ Became familiar with the Watson Studio SPSS Modeler capability
- ✓ Loaded the trafficking data into SPSS Modeler
- ✓ Joined the datasets
- ✓ Profiled the trafficking data
- ✓ Prepared the trafficking data
- ✓ Trained/Evaluated a machine learning model.
- ✓ Saved the model.