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.

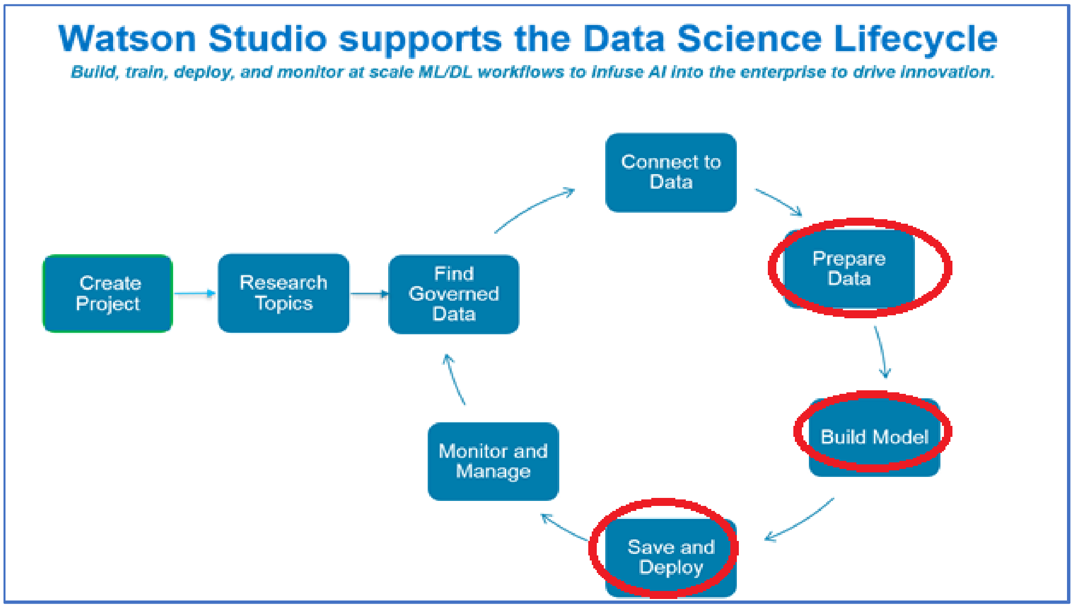

Watson Studio supports the Data Science Lifecycle
Build, train, deploy, and monitor at scale ML/DL workflows to infuse AI into the enterprise to drive innovation.
Create Project
Research Topics
Find Governed Data
Connect to Data
Prepare Data
Build Model
Save and Deploy
Monitor and Manage

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.

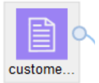
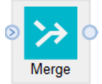
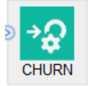
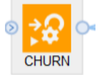
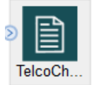
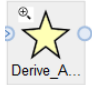
Background

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.

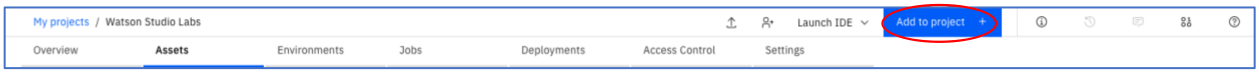
Other columns in the dataset are similarly very granular and could also be categorized for modeling purposes. This lab does not include steps to accomplish this, but it would be similar to what was done for the occupation column.

Lab Steps

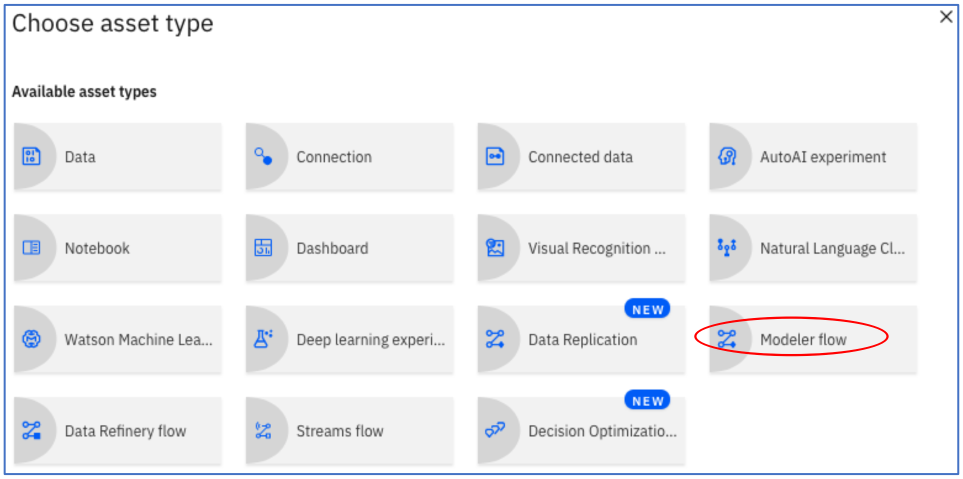
In this section, we will create a Machine Learning flow using SPSS nodes.

Step 1 - Create a New Flow

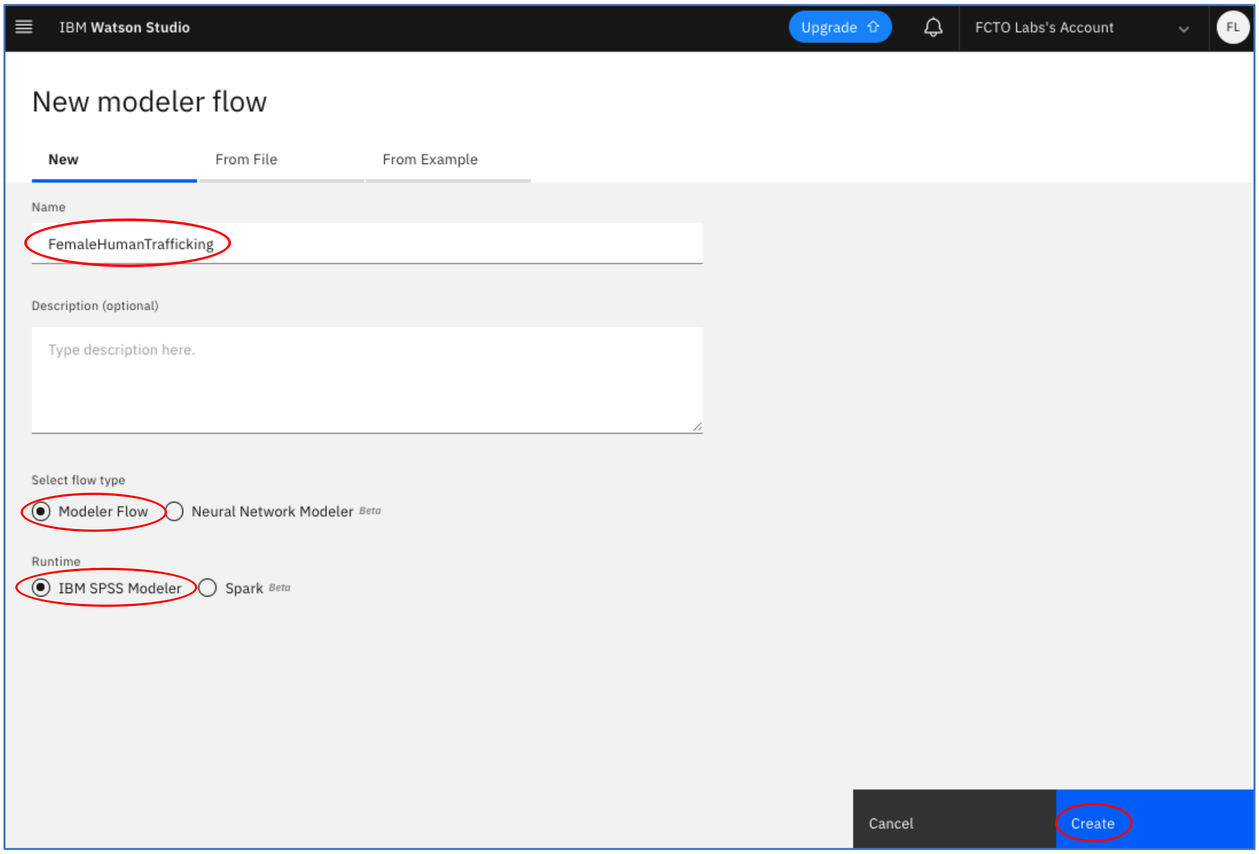
1. In the Watson Studio project, click on **Add to project**.



2. Select **Modeler Flow**.



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



New modeler flow

Name

FemaleHumanTrafficking

Description (optional)

Type description here.

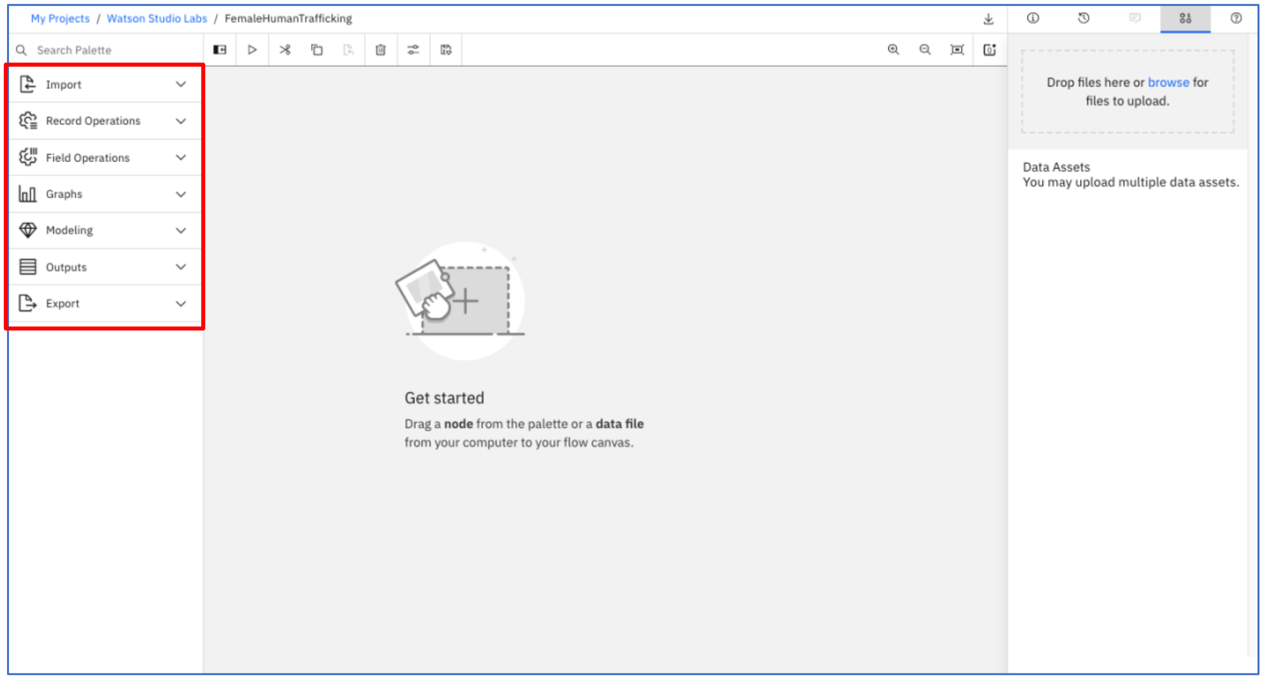
Select flow type

● Modeler Flow ○ Neural Network Modeler *Beta*

Runtime

● IBM SPSS Modeler ○ Spark *Beta*

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


My Projects / Watson Studio Labs / FemaleHumanTrafficking

Drop files here or browse for files to upload.

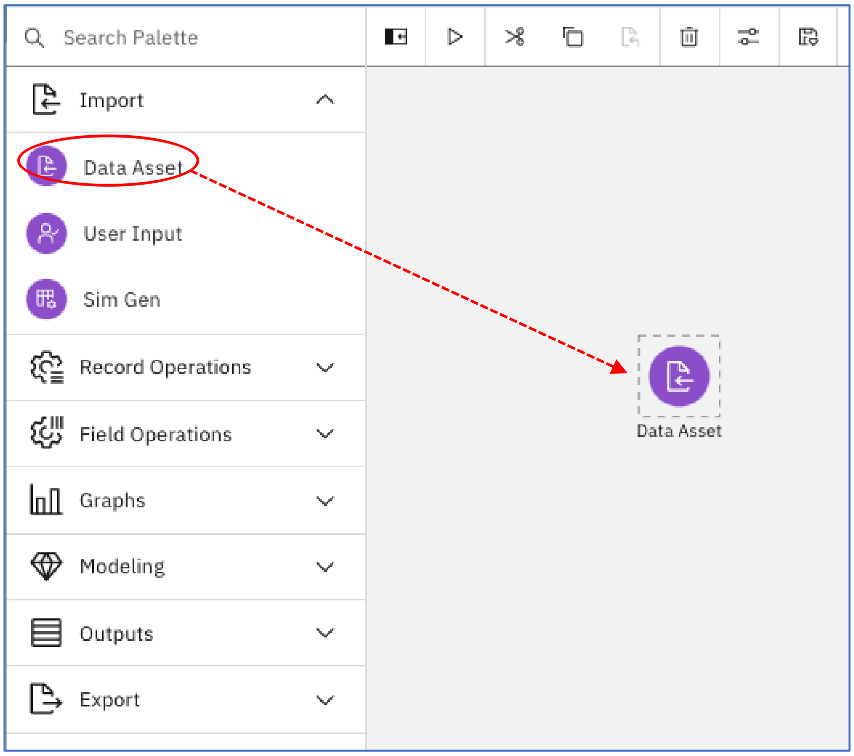
Data Assets
You may upload multiple data assets.

Get started

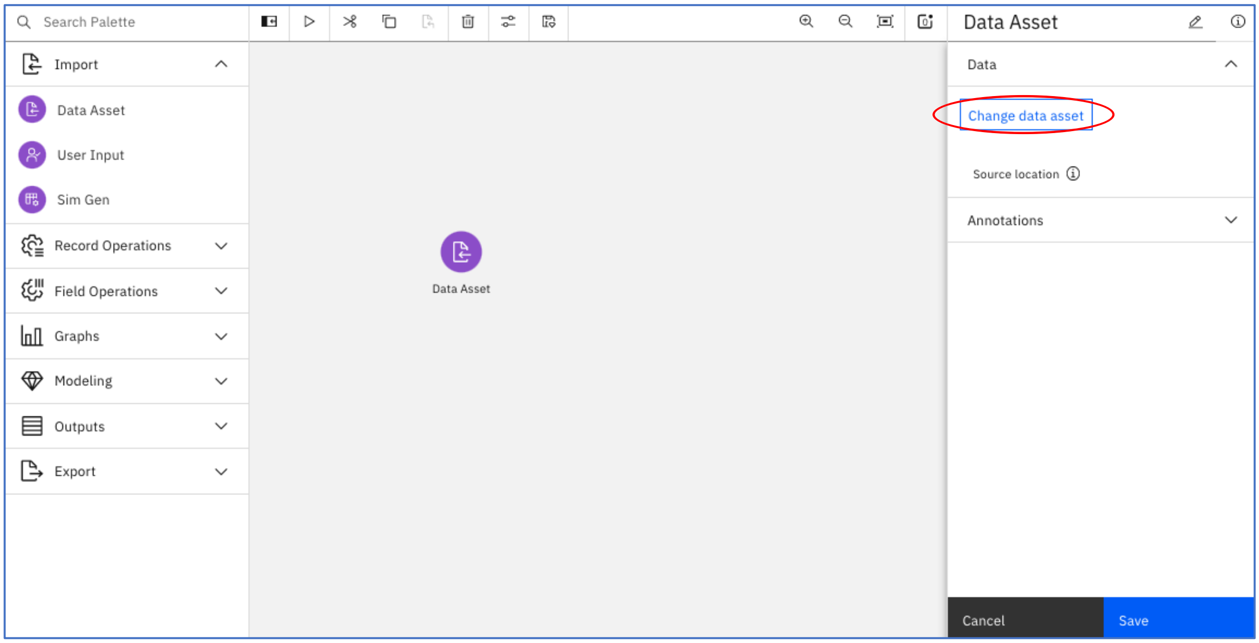
Drag a **node** from the palette or a **data file** from your computer to your flow canvas.

Step 2 - Load the Trafficking Datasets

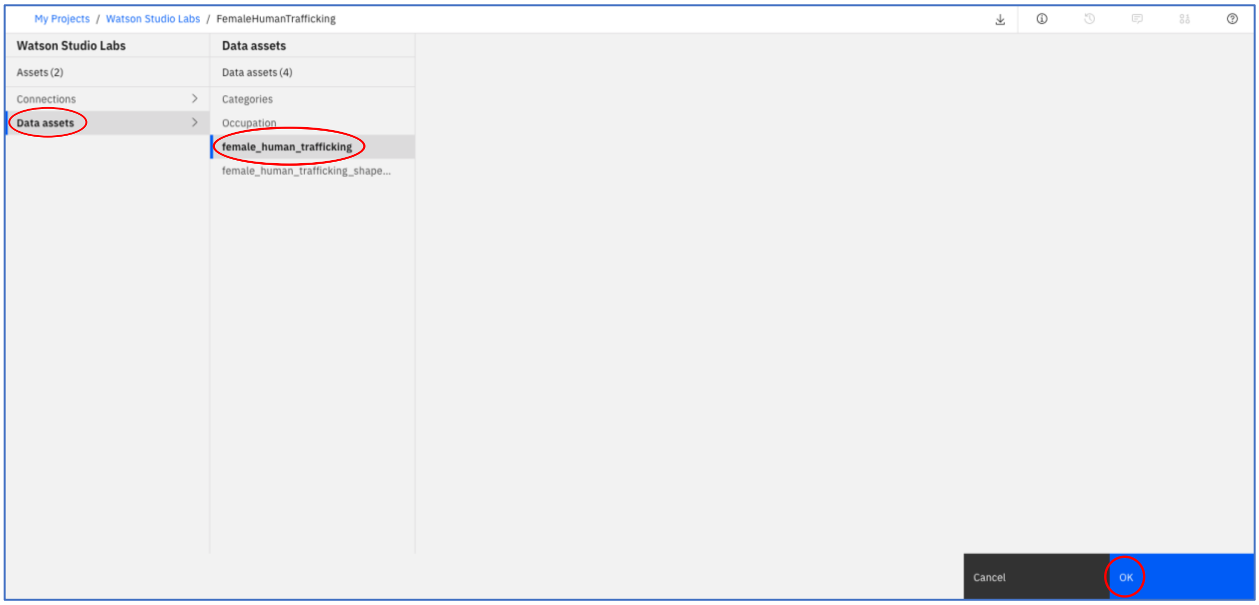
1. Click on **Import** and then **Data Asset** and hold the left mouse key on the Data Asset icon and **drag it onto the left side of the canvas**. Release the left mouse key.



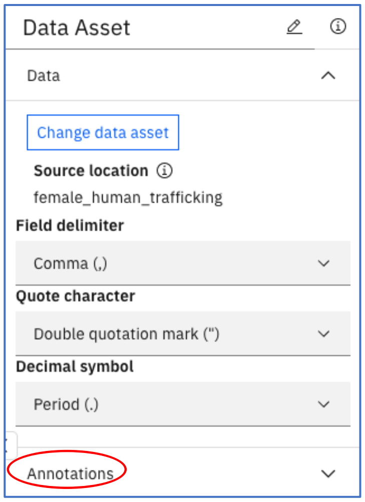
2. Double click on the **Data Asset**. In the window pane on the right-hand-side click on **Change data asset**.



3. Click on **Data Assets**, click on **female_human_trafficking**, then click **OK**.



4. Click on **ANNOTATIONS.**



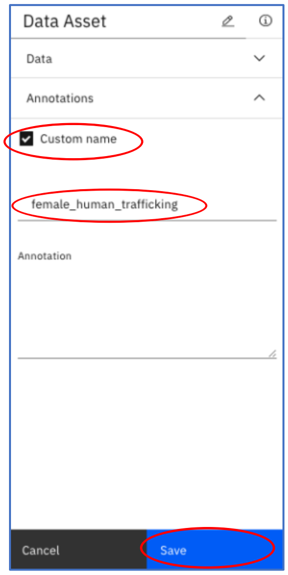
Source location
female_human_trafficking

Field delimiter
Comma (,)

Quote character
Double quotation mark (")

Decimal symbol
Period (.)

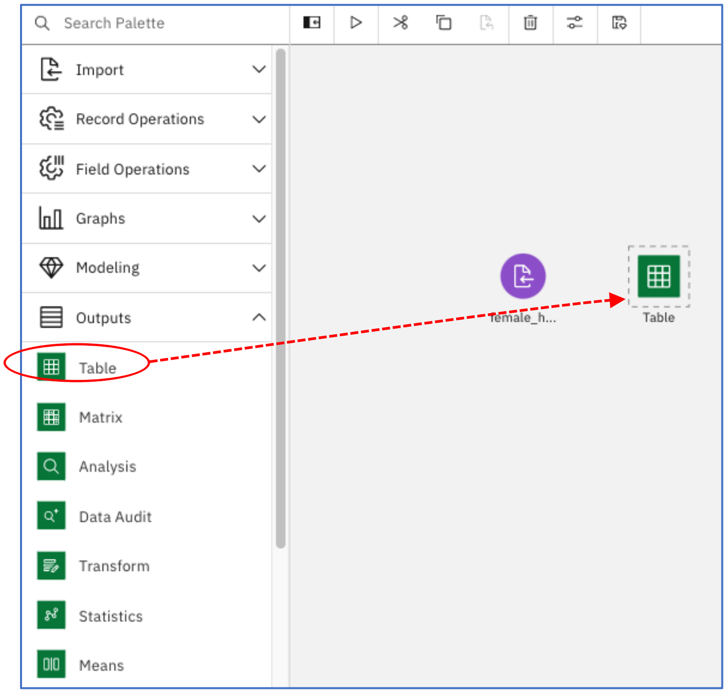
5. Click on **Custom name**, and type **female_human_trafficking**, and click on **Save**.



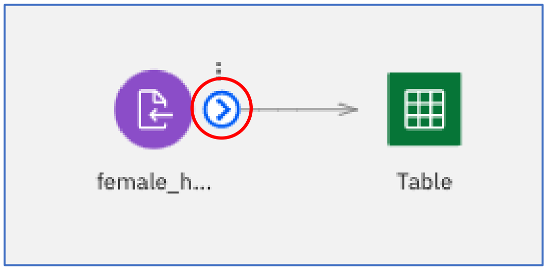
☑ Custom name

female_human_trafficking

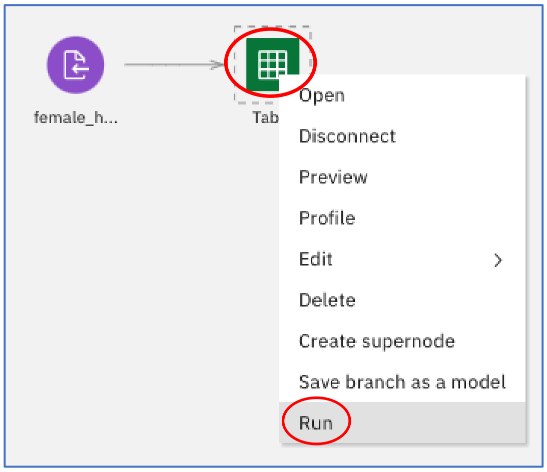
6. Click on the **Outputs** menu item in the Node Palette on the left and then click on the **Table** icon and drag the icon to the right of the female_human_trafficking to display its contents. If the Node Palette is not visible, click on the Node Palette icon ◧ .



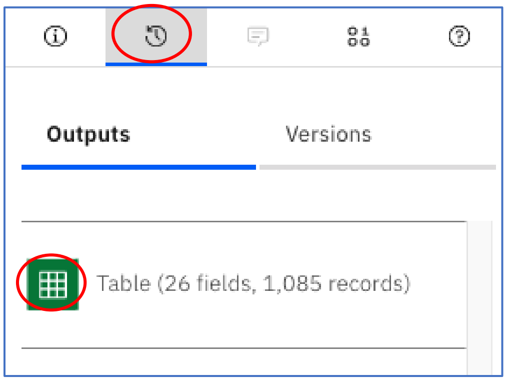
7. **Connect the right side of the female_human_trafficking icon to the left side of the Table icon**. This is accomplished by hovering over the data asset icon, clicking on the little blue arrow that appears the right side of the icon, holding the left mouse key, and dragging the mouse to the Table icon, and then releasing the left mouse key.



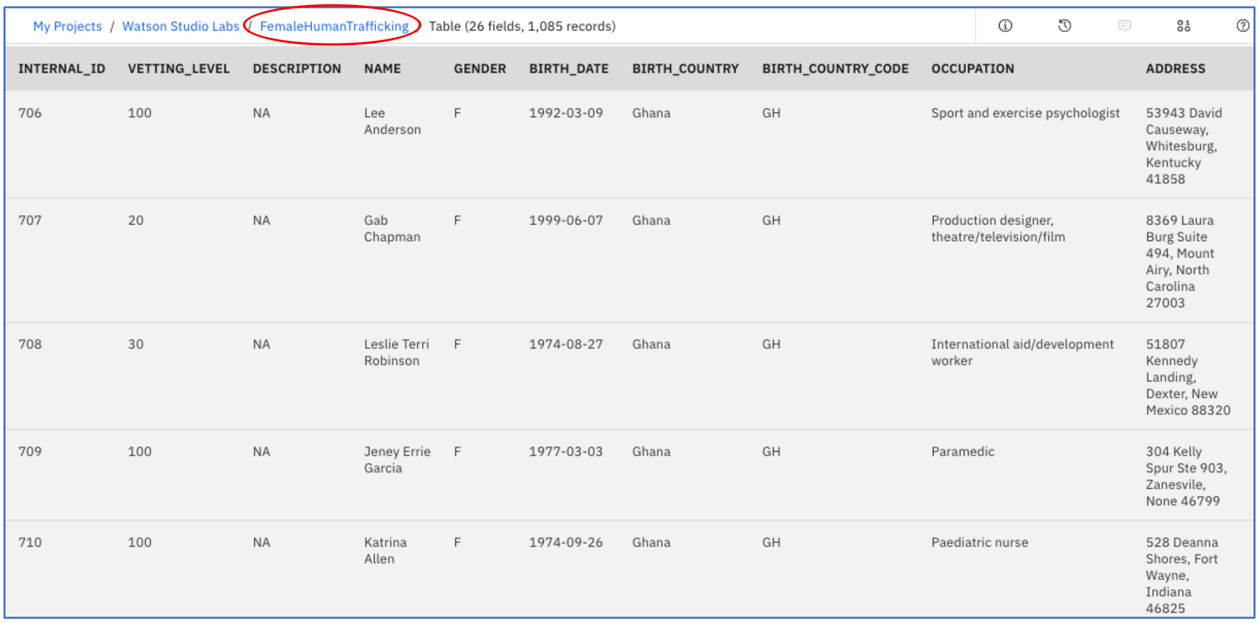
8. Right click on the **Table** icon and select **Run**.



9. 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. **Double click on the Table selection** and the contents of the female_human_trafficking is displayed.



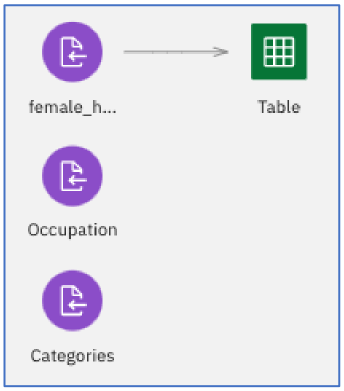
10. Each row contains travel information for a person. We will use this data to make predictions on trafficking risk. **Click on your flow name**.



My Projects / Watson Studio Labs / FemaleHumanTrafficking Table (26 fields, 1,085 records)

INTERNAL_ID	VETTING_LEVEL	DESCRIPTION	NAME	GENDER	BIRTH_DATE	BIRTH_COUNTRY	BIRTH_COUNTRY_CODE	OCCUPATION	ADDRESS
706	100	NA	Lee Anderson	F	1992-03-09	Ghana	GH	Sport and exercise psychologist	53943 David Causeway, Whitesburg, Kentucky 41858
707	20	NA	Gab Chapman	F	1999-06-07	Ghana	GH	Production designer, theatre/television/film	8369 Laura Burg Suite 494, Mount Airy, North Carolina 27003
708	30	NA	Leslie Terri Robinson	F	1974-08-27	Ghana	GH	International aid/development worker	51807 Kennedy Landing, Dexter, New Mexico 88320
709	100	NA	Jeney Errie Garcia	F	1977-03-03	Ghana	GH	Paramedic	304 Kelly Spur Ste 903, Zanesvile, None 46799
710	100	NA	Katrina Allen	F	1974-09-26	Ghana	GH	Paediatric nurse	528 Deanna Shores, Fort Wayne, Indiana 46825

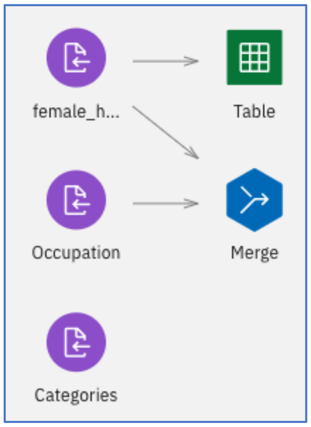
11. Repeat steps 1-5 for the occupation dataset and then repeat steps 1-5 for the Categories dataset. When complete, the canvas should appear as below.



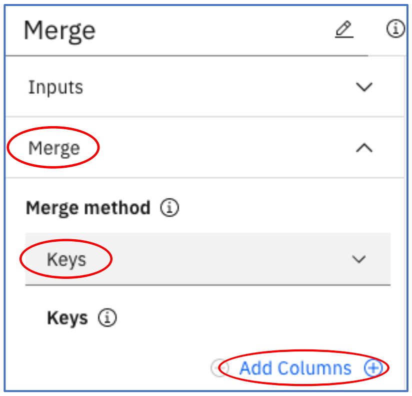
Step 3 - Join the Data Sources

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 **Occupations** 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 **Occupations** data source to the **Merge** node. The canvas should appear as below.



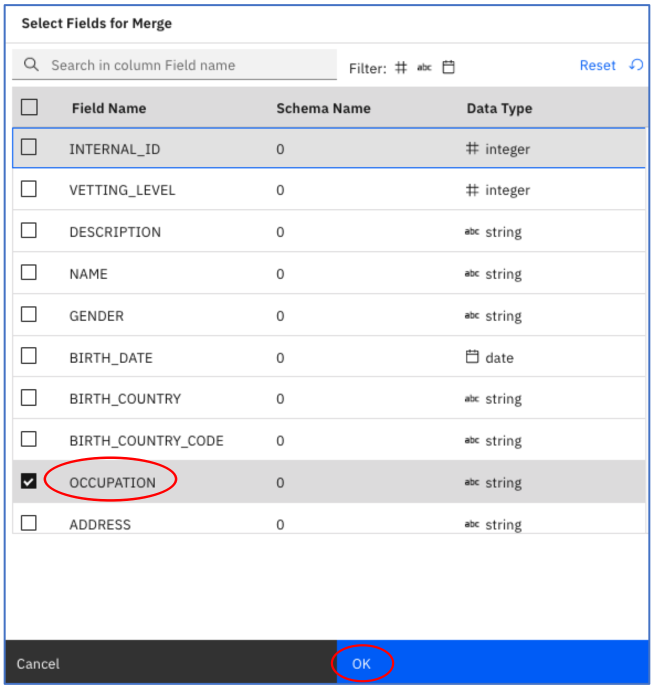
2. Double-click on the **Merge** Node. Click on **MERGE**, then click on **Keys** for the Merge method, and click on **Add Columns**.



Merge method

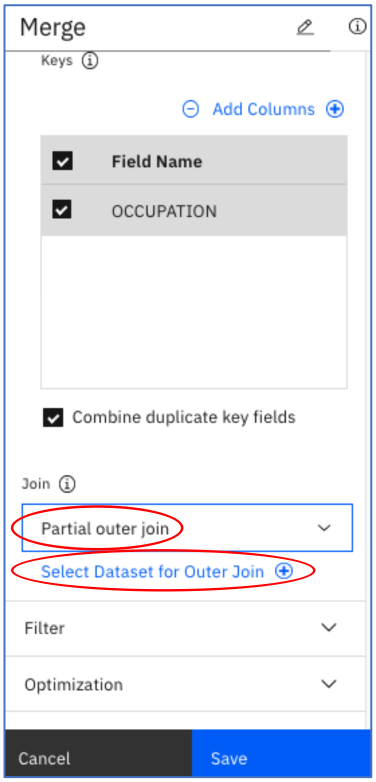
Keys

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



□	Field Name	Schema Name	Data Type
□	INTERNAL_ID	0	# integer
□	VETTING_LEVEL	0	# integer
□	DESCRIPTION	0	abc string
□	NAME	0	abc string
□	GENDER	0	abc string
□	BIRTH_DATE	0	date
□	BIRTH_COUNTRY	0	abc string
□	BIRTH_COUNTRY_CODE	0	abc string
☑	OCCUPATION	0	abc string
□	ADDRESS	0	abc string

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

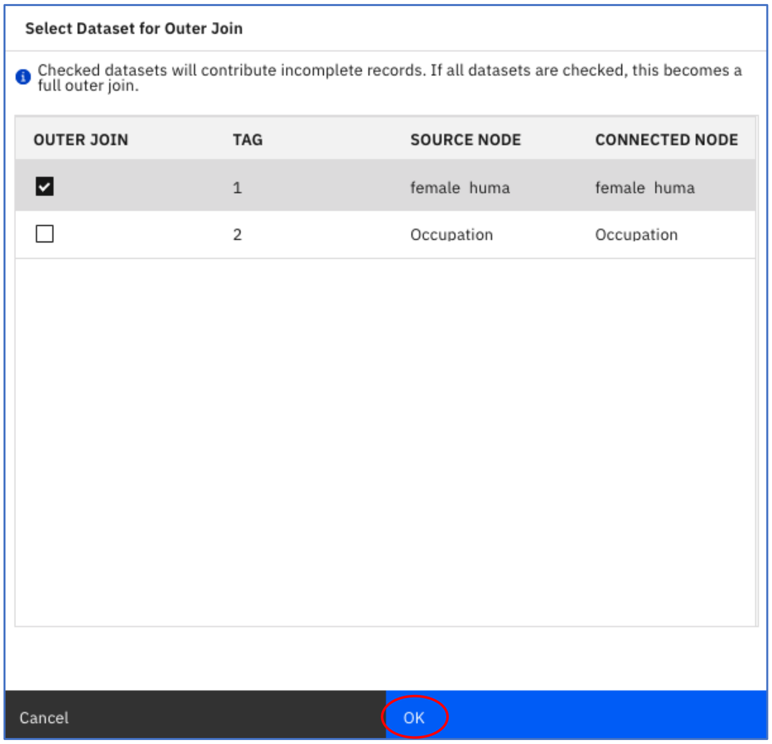
Keys

☑	Field Name
☑	OCCUPATION

☑ Combine duplicate key fields

Join: Partial outer join

5. Make sure the female_human_trafficking **SOURCE NODE** is checked and click **OK**.

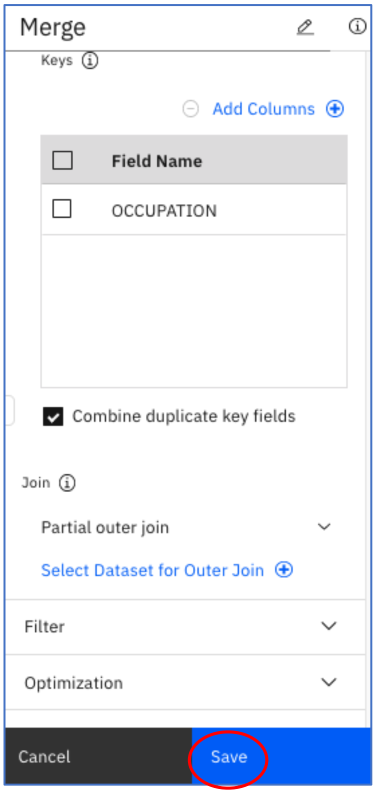


Select Dataset for Outer Join

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

OUTER JOIN	TAG	SOURCE NODE	CONNECTED NODE
☑	1	female_huma	female_huma
☐	2	Occupation	Occupation

6. Click on **Save**.



Merge

Keys

⊖ Add Columns ⊕

□	Field Name
□	OCCUPATION

☑ Combine duplicate key fields

Join

Partial outer join

Select Dataset for Outer Join ⊕

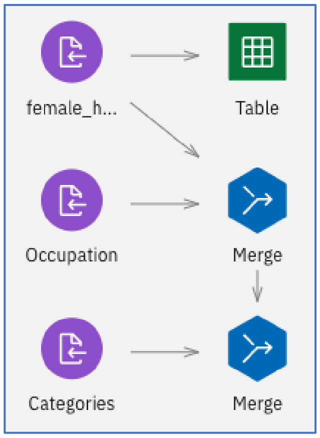
Filter

Optimization

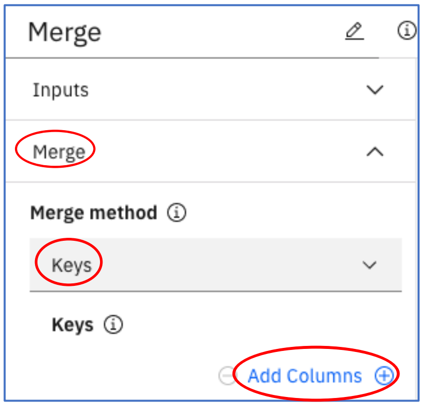
Cancel

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 **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.



8. Double click on the second **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

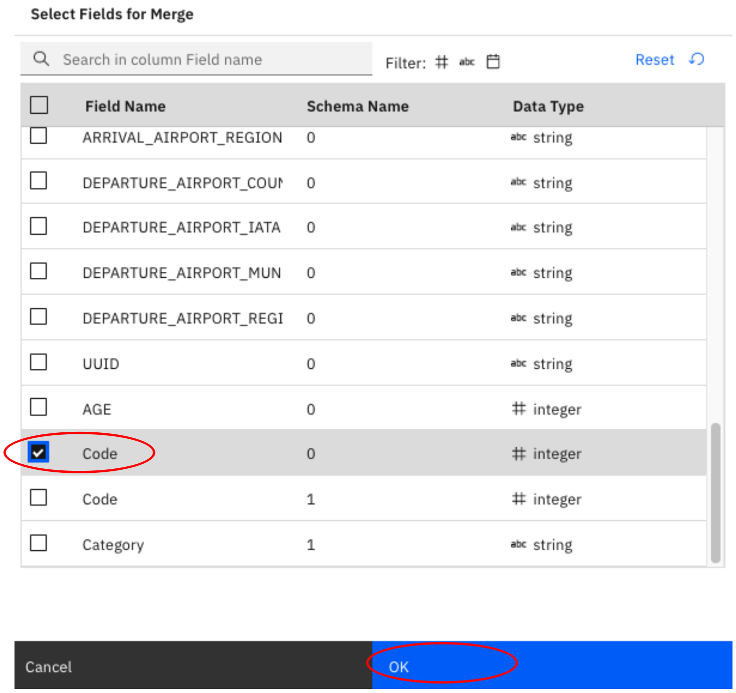
Merge method

Keys

Keys

Add Columns

9. Scroll down and click on the **Code** checkbox where the Schema Name column is 0. Click on **OK**.

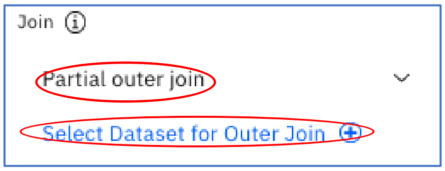

Select Fields for Merge

Search in column Field name — Filter: # abc — Reset

□	Field Name	Schema Name	Data Type
□	ARRIVAL_AIRPORT_REGION	0	abc string
□	DEPARTURE_AIRPORT_COUI	0	abc string
□	DEPARTURE_AIRPORT_IATA	0	abc string
□	DEPARTURE_AIRPORT_MUN	0	abc string
□	DEPARTURE_AIRPORT_REGI	0	abc string
□	UUID	0	abc string
□	AGE	0	# integer
☑	Code	0	# integer
□	Code	1	# integer
□	Category	1	abc string

Cancel — OK

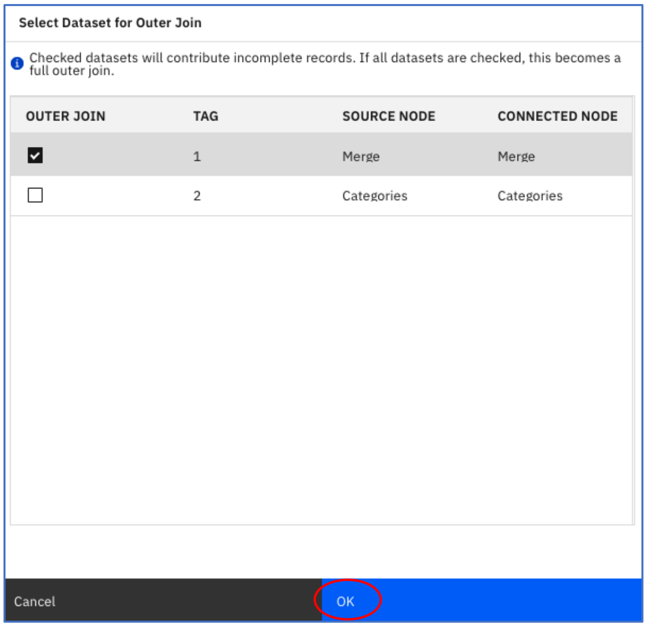
10. Scroll down in the side panel and click on **Partial Outer Join** and click on Select Dataset.


Join

Partial outer join

Select Dataset for Outer Join

11. Make sure the **Merge SOURCE NODE** is selected and click **Okay.**

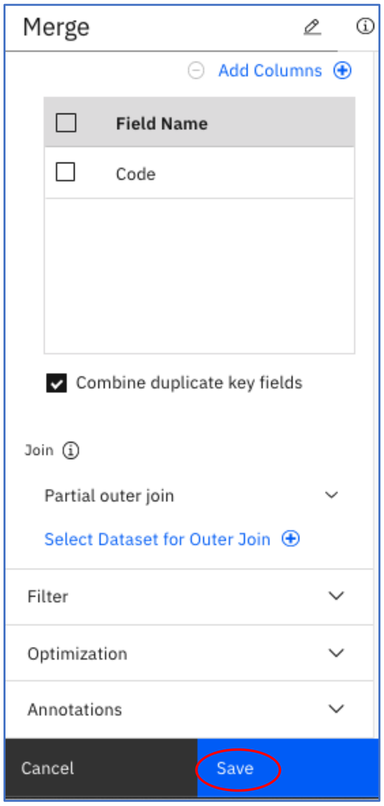


Select Dataset for Outer Join

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

OUTER JOIN	TAG	SOURCE NODE	CONNECTED NODE
☑	1	Merge	Merge
□	2	Categories	Categories

12. Click **Save**.



Merge

Add Columns

□	Field Name
□	Code

☑ Combine duplicate key fields

Join

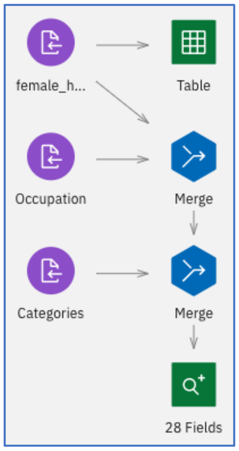
Partial outer join

Select Dataset for Outer Join

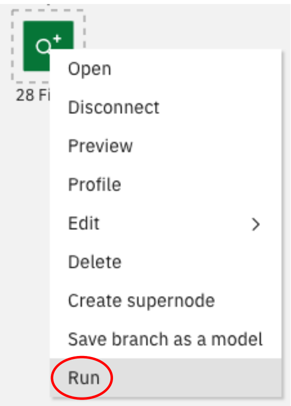
Step 4 - Explore the Data using the Data Audit Node

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.

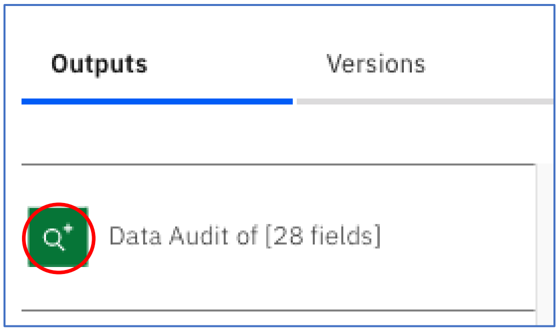
1. Add a **Data Audit** node to the flow clicking on the **Outputs** menu item in the Node Palette, and then dragging the **Data Audit** node to underneath the **Merge** node. If the Node Palette is not visible, click on the Node Palette icon ◨ . Connect the node to the Data Audit node. The canvas should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



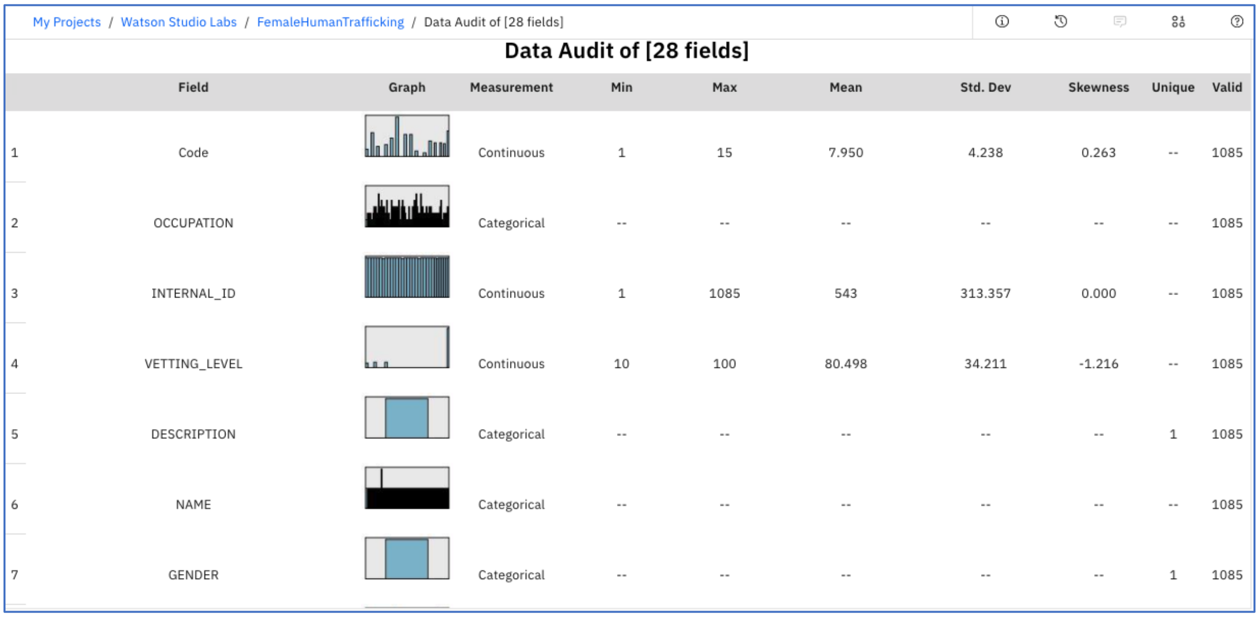
2. Right click on the **Data Audit** node and click **Run**.



3. The “Running Flow” prompt will appear and then when completed a Data Audit output selection will appear on the right side of the screen under the **Outputs** tab. If the **Outputs** tab doesn’t display, click on the clock icon.



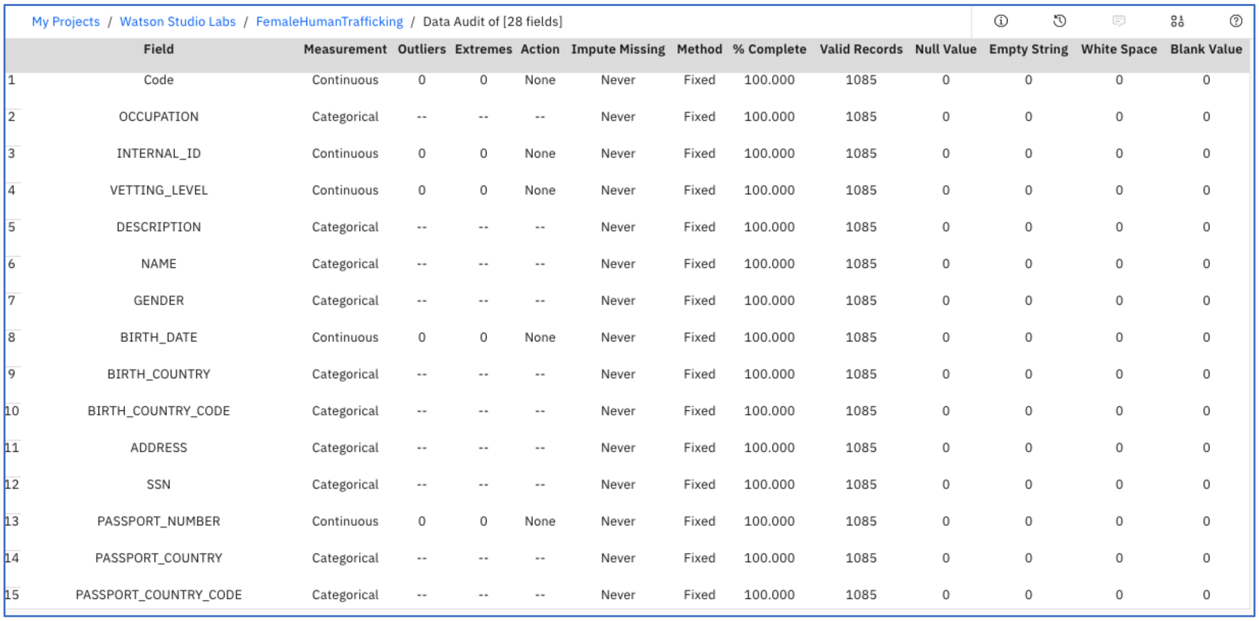
4. Double click on the **Data Audit of [28 fields]** to view the Data Audit output. The top section of the Data Audit report displays profiling information. For modeling purposes, fields that have only 1 unique value, or have many unique values should be eliminated. In addition, certain fields are directly 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.


My Projects / Watson Studio Labs / FemaleHumanTrafficking / Data Audit of [28 fields]

Data Audit of [28 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	313.357	0.000	--	1085
4	VETTING_LEVEL		Continuous	10	100	80.498	34.211	-1.216	--	1085
5	DESCRIPTION		Categorical	--	--	--	--	--	1	1085
6	NAME		Categorical	--	--	--	--	--	--	1085
7	GENDER		Categorical	--	--	--	--	--	1	1085

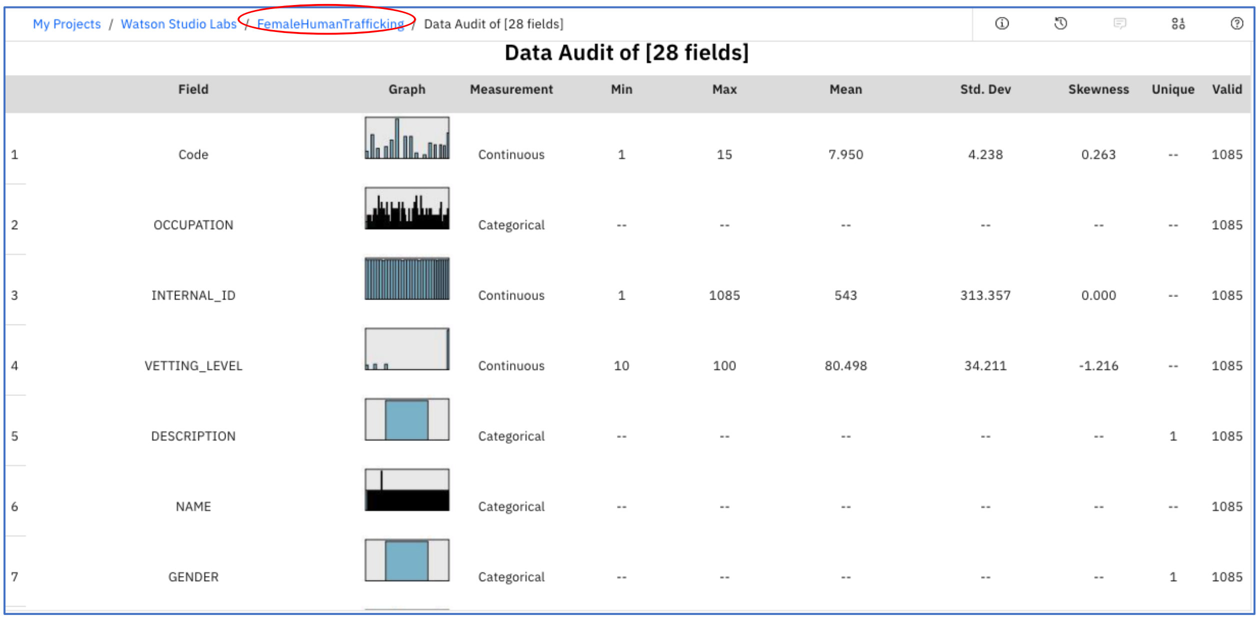
5. **Scroll down** to view the bottom section. It displays data quality checks in the form of missing values or anomalous values. In our travel data simulator, we didn’t simulate any of those type of values!



My Projects / Watson Studio Labs / FemaleHumanTrafficking / Data Audit of [28 fields]

	Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
1	Code	Continuous	0	0	None	Never	Fixed	100.000	1085	0	0	0	0
2	OCCUPATION	Categorical	--	--	--	Never	Fixed	100.000	1085	0	0	0	0
3	INTERNAL_ID	Continuous	0	0	None	Never	Fixed	100.000	1085	0	0	0	0
4	VETTING_LEVEL	Continuous	0	0	None	Never	Fixed	100.000	1085	0	0	0	0
5	DESCRIPTION	Categorical	--	--	--	Never	Fixed	100.000	1085	0	0	0	0
6	NAME	Categorical	--	--	--	Never	Fixed	100.000	1085	0	0	0	0
7	GENDER	Categorical	--	--	--	Never	Fixed	100.000	1085	0	0	0	0
8	BIRTH_DATE	Continuous	0	0	None	Never	Fixed	100.000	1085	0	0	0	0
9	BIRTH_COUNTRY	Categorical	--	--	--	Never	Fixed	100.000	1085	0	0	0	0
10	BIRTH_COUNTRY_CODE	Categorical	--	--	--	Never	Fixed	100.000	1085	0	0	0	0
11	ADDRESS	Categorical	--	--	--	Never	Fixed	100.000	1085	0	0	0	0
12	SSN	Categorical	--	--	--	Never	Fixed	100.000	1085	0	0	0	0
13	PASSPORT_NUMBER	Continuous	0	0	None	Never	Fixed	100.000	1085	0	0	0	0
14	PASSPORT_COUNTRY	Categorical	--	--	--	Never	Fixed	100.000	1085	0	0	0	0
15	PASSPORT_COUNTRY_CODE	Categorical	--	--	--	Never	Fixed	100.000	1085	0	0	0	0

6. Return to the flow by clicking on the **FemaleHumanTrafficking** breadcrumb at the top.



My Projects / Watson Studio Labs / FemaleHumanTrafficking / Data Audit of [28 fields]

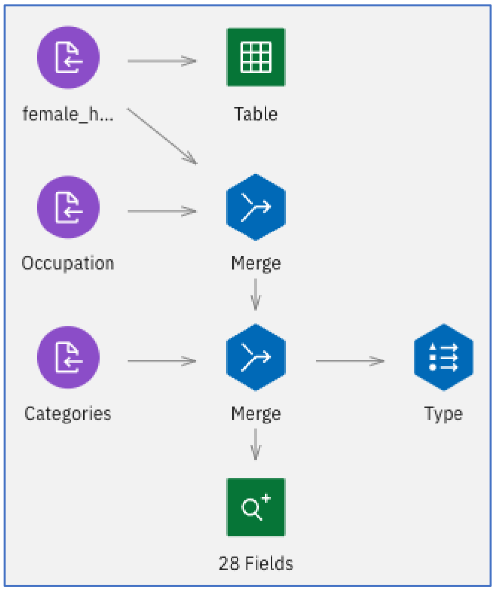
Data Audit of [28 fields]

	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	313.357	0.000	--	1085
4	VETTING_LEVEL		Continuous	10	100	80.498	34.211	-1.216	--	1085
5	DESCRIPTION		Categorical	--	--	--	--	--	1	1085
6	NAME		Categorical	--	--	--	--	--	--	1085
7	GENDER		Categorical	--	--	--	--	--	1	1085

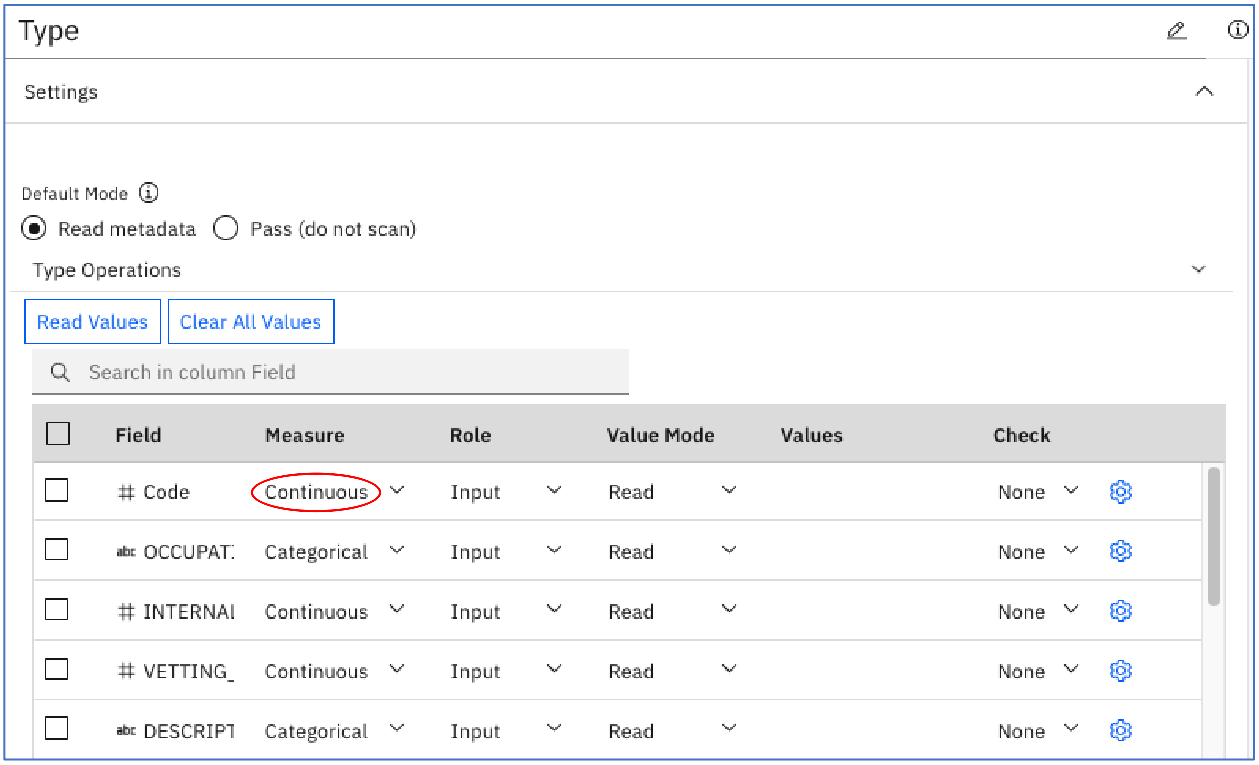
Step 5 - Explore the Data using Graph Nodes.

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 second **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

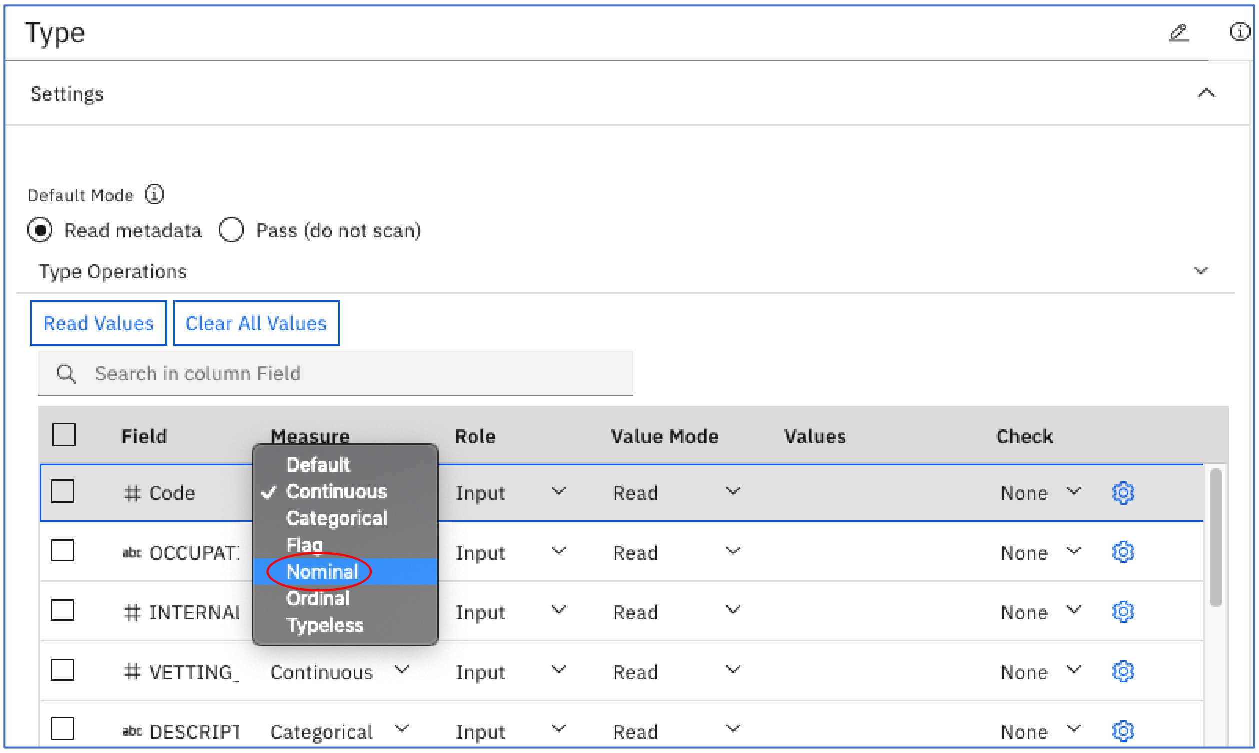
Settings

Default Mode ⓘ

● Read metadata ○ Pass (do not scan)

Type Operations

□	Field	Measure	Role	Value Mode	Values	Check
□	# Code	Continuous	Input	Read		None
□	abc OCCUPAT:	Categorical	Input	Read		None
□	# INTERNAI	Continuous	Input	Read		None
□	# VETTING_	Continuous	Input	Read		None
□	abc DESCRIPT	Categorical	Input	Read		None



Type

Settings

Default Mode

● Read metadata ○ Pass (do not scan)

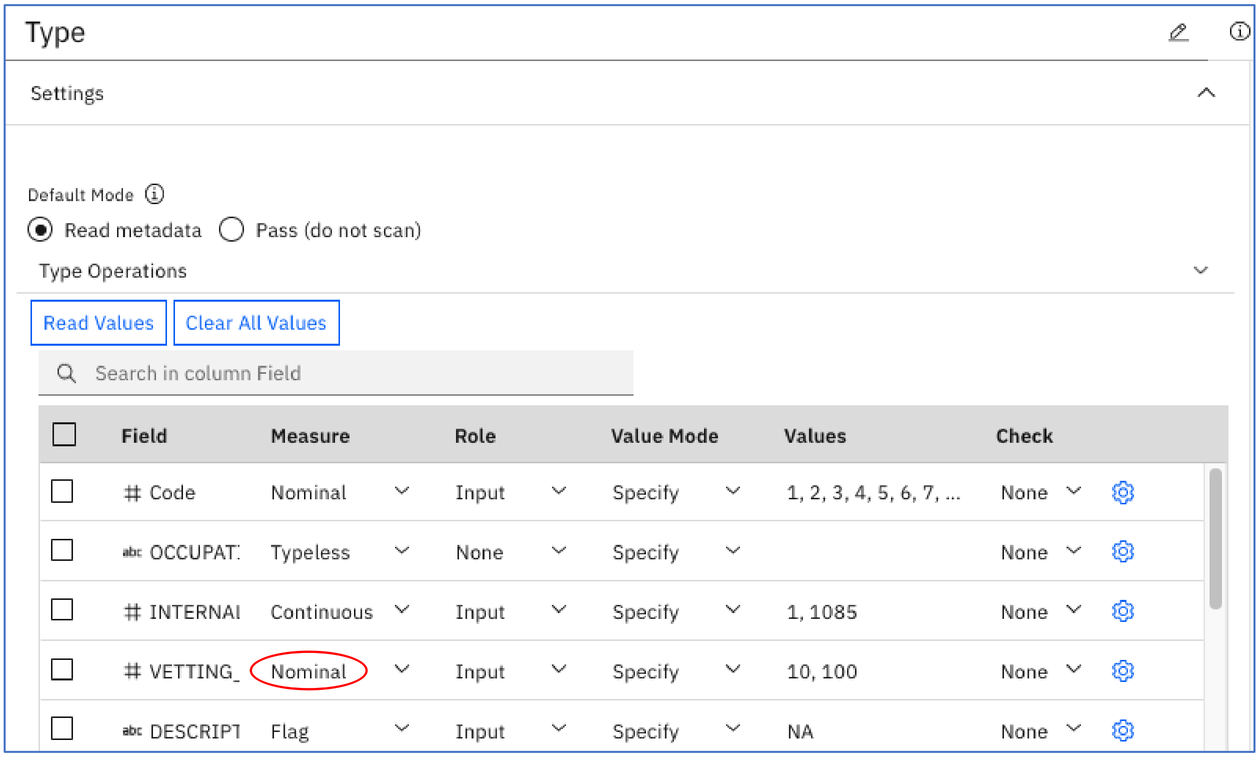
Type Operations

Read Values | Clear All Values

Search in column Field

□	Field	Measure	Role	Value Mode	Values	Check
□	# Code		Input	Read		None
□	abc OCCUPAT		Input	Read		None
□	# INTERNAI		Input	Read		None
□	# VETTING_	Continuous	Input	Read		None
□	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)

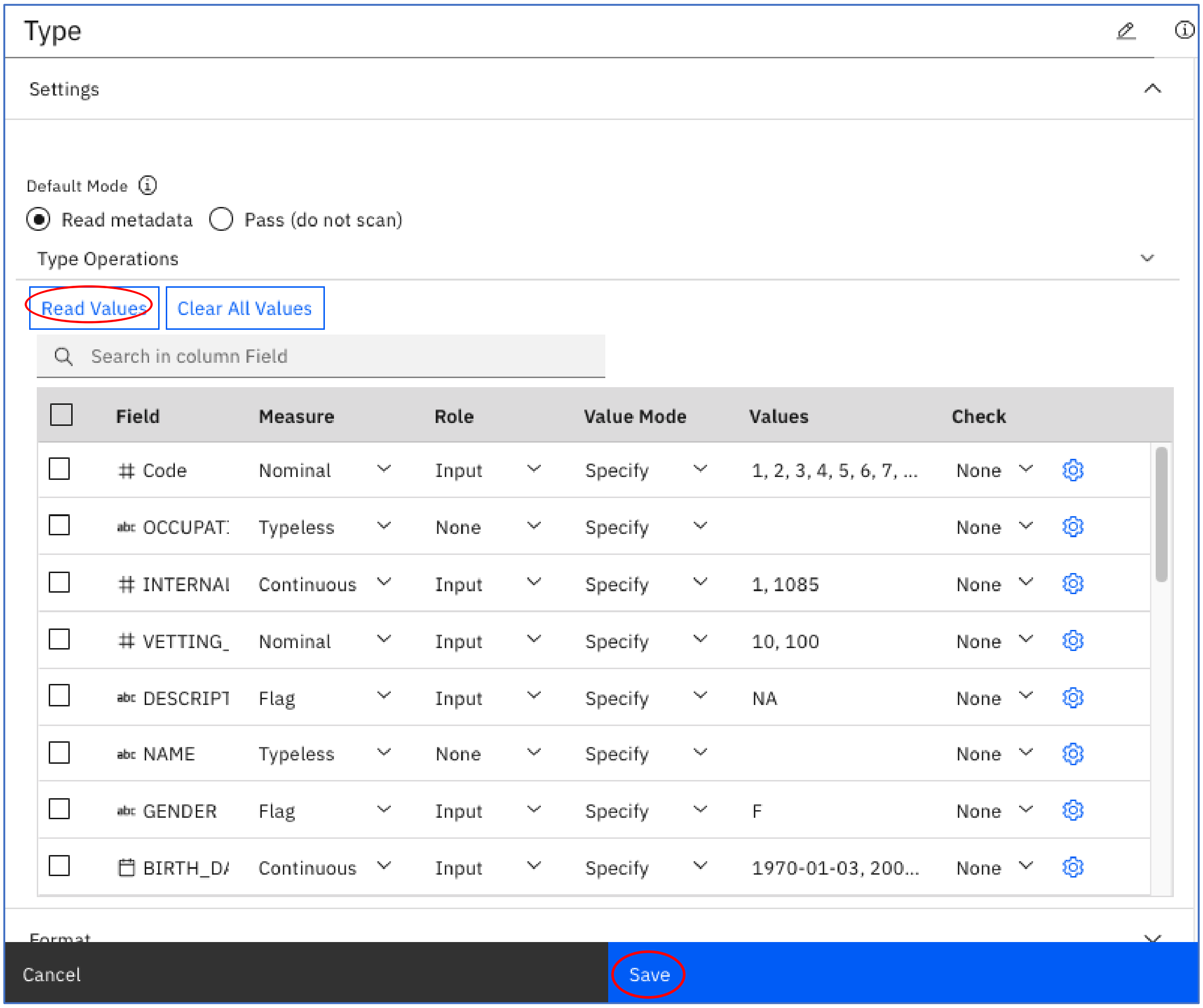
Type Operations

Read Values | Clear All Values

Search in column Field

□	Field	Measure	Role	Value Mode	Values	Check
□	# Code	Nominal	Input	Specify	1, 2, 3, 4, 5, 6, 7, ...	None
□	abc OCCUPAT	Typeless	None	Specify		None
□	# INTERNAI	Continuous	Input	Specify	1, 1085	None
□	# VETTING_	Nominal	Input	Specify	10, 100	None
□	abc DESCRIPT	Flag	Input	Specify	NA	None

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



Type

Settings

Default Mode ⓘ

● Read metadata ○ Pass (do not scan)

Type Operations

Read Values | Clear All Values

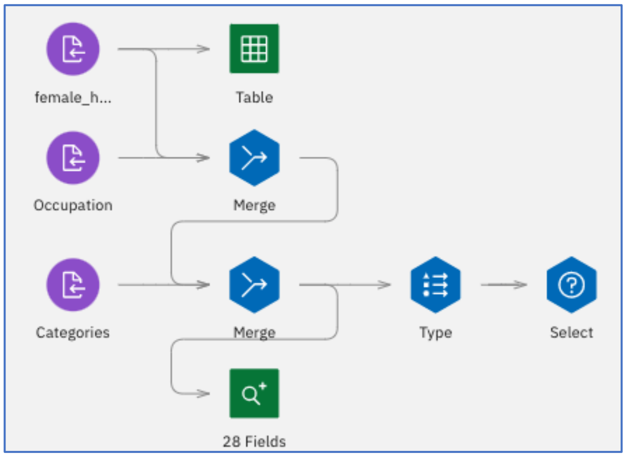
Search in column Field

□	Field	Measure	Role	Value Mode	Values	Check
□	# Code	Nominal	Input	Specify	1, 2, 3, 4, 5, 6, 7, ...	None
□	abc OCCUPAT	Typeless	None	Specify		None
□	# INTERNAI	Continuous	Input	Specify	1, 1085	None
□	# VETTING_	Nominal	Input	Specify	10, 100	None
□	abc DESCRIPT	Flag	Input	Specify	NA	None
□	abc NAME	Typeless	None	Specify		None
□	abc GENDER	Flag	Input	Specify	F	None
□	BIRTH_D/	Continuous	Input	Specify	1970-01-03, 200...	None

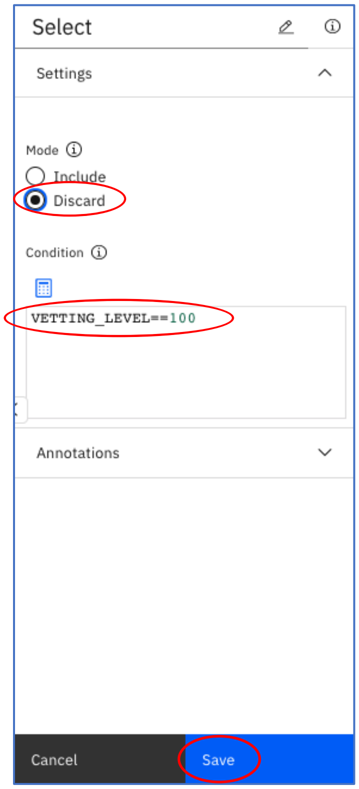
Format

Cancel | Save

5. 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.



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



Select

Settings

Mode

○ Include

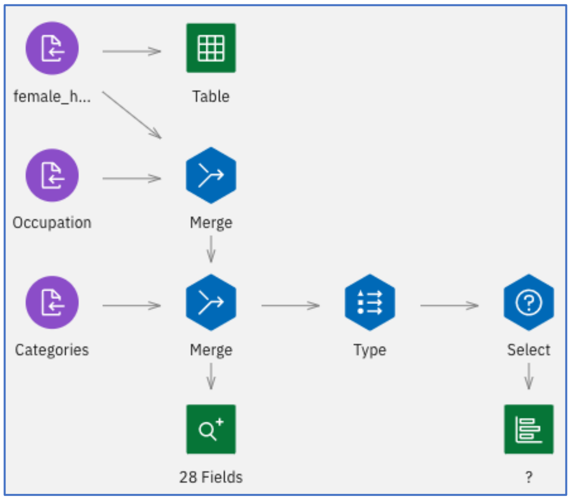
● Discard

Condition

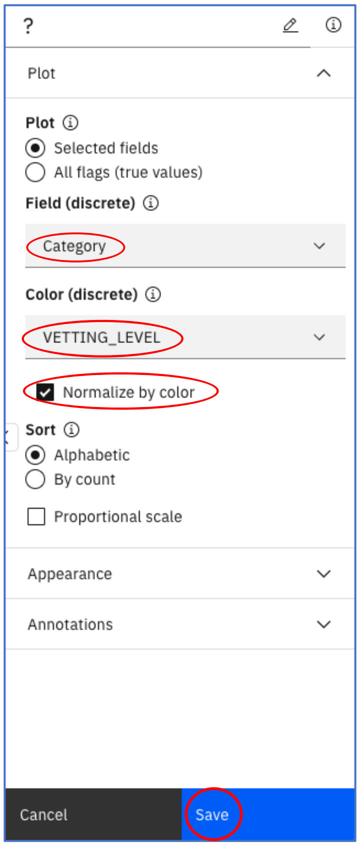
```
VETTING_LEVEL==100
```

Annotations

7. 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.



8. 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**.



?

Plot

Plot

● Selected fields

○ All flags (true values)

Field (discrete)

Category

Color (discrete)

VETTING_LEVEL

☑ Normalize by color

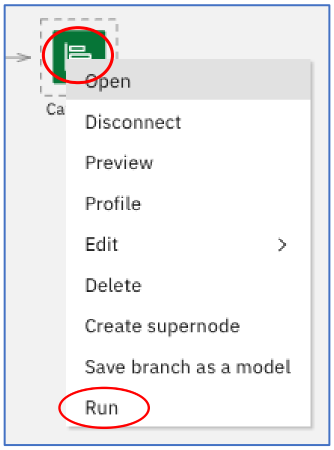
Sort

● Alphabetic

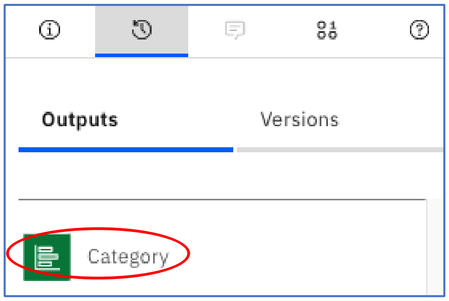
○ By count

☐ Proportional scale

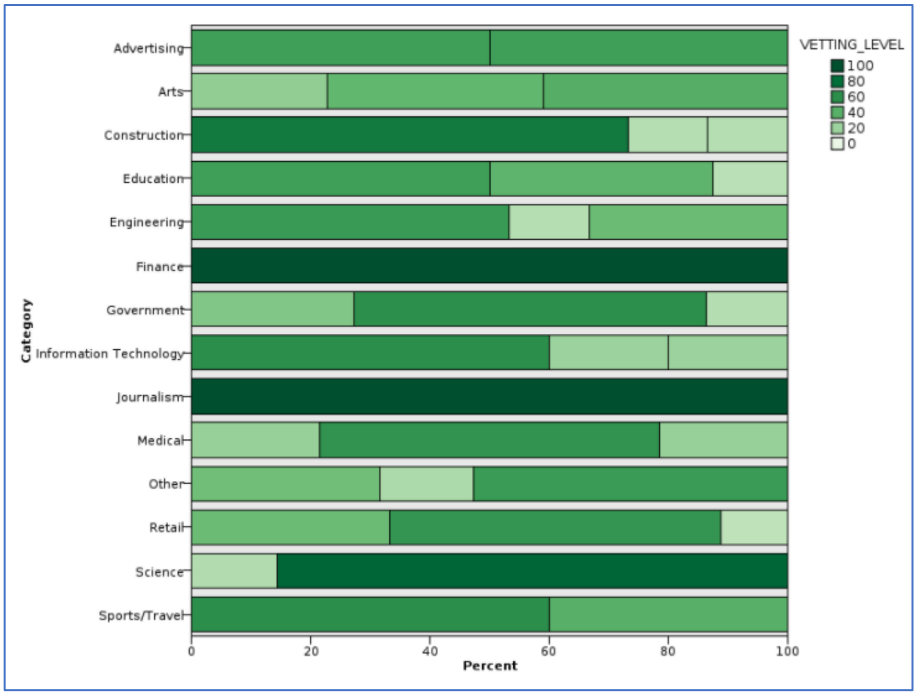
9. Right click on the Distribution node and select **Run**.



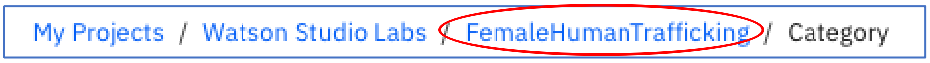
10. The Distribution output will appear under the **Outputs** tab. Double-click on Categories to view the graph. If you don't see the Outputs, click on the Outputs ("clock") icon.



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

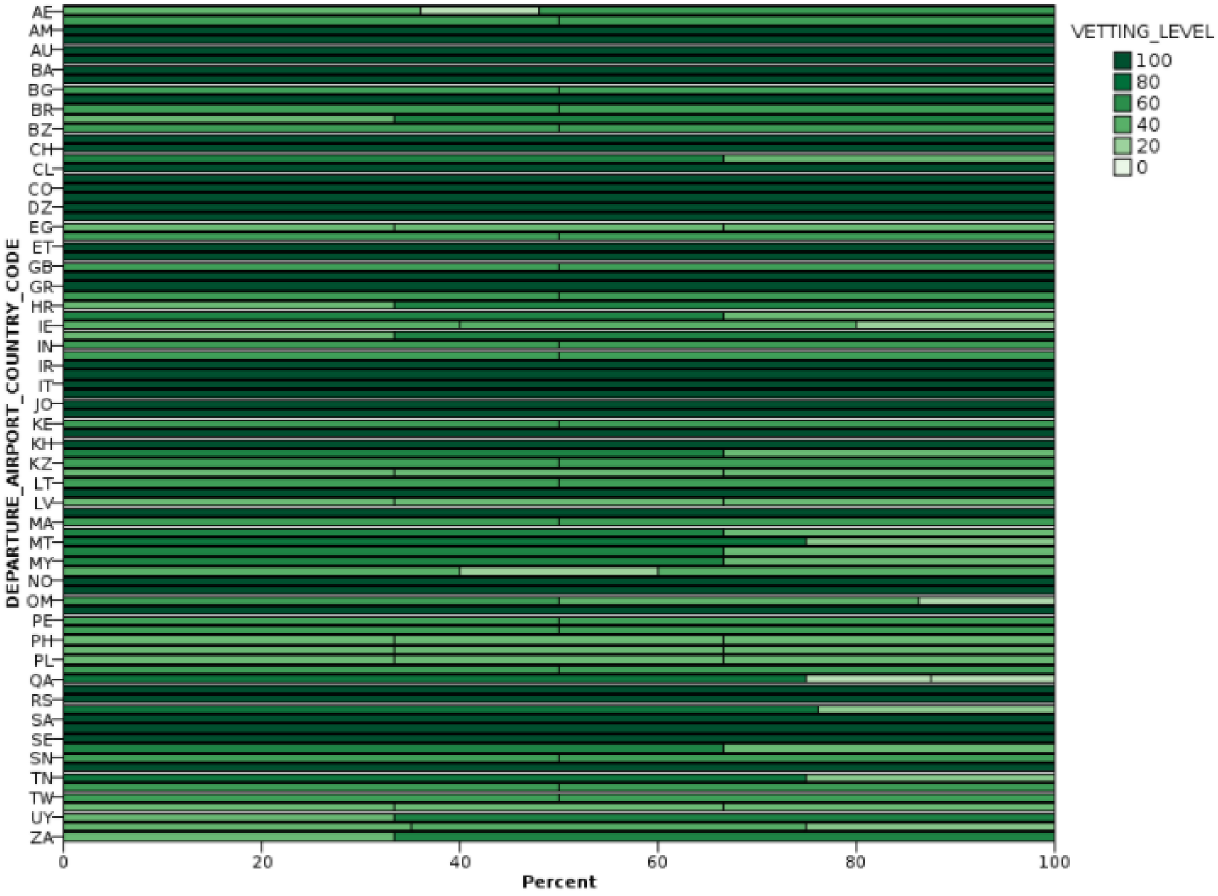


12. Return to the flow by clicking on the FemaleHumanTrafficking breadcrumb at the top.

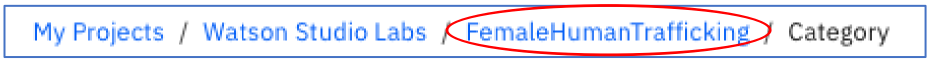


My Projects / Watson Studio Labs / FemaleHumanTrafficking / Category

13. 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. Double click on the **DEPARTURE_AIRPORT_COUNTRY_CODE** in the **Outputs** pane to display the graph.

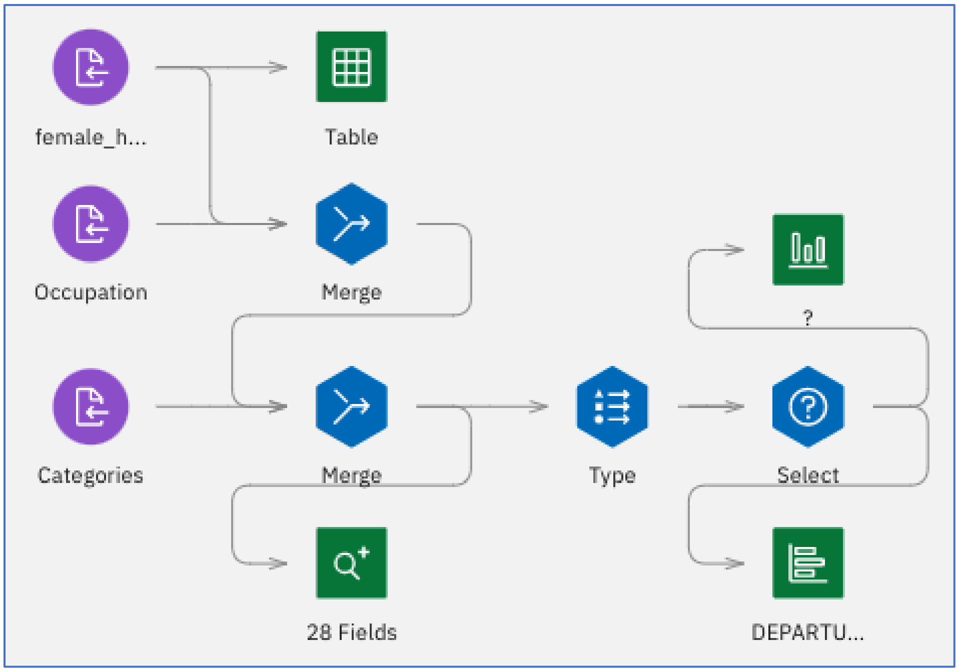


14. Return to the flow by clicking on the FemaleHumanTrafficking breadcrumb at the top.

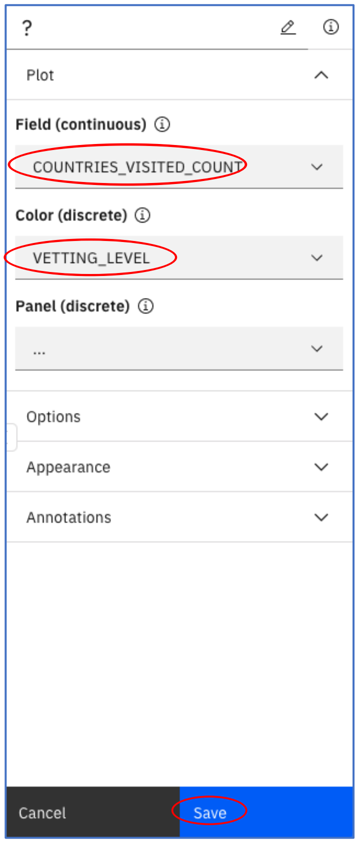


My Projects / Watson Studio Labs / FemaleHumanTrafficking / Category

15. 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.



16. 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**.



?

Plot

Field (continuous)

COUNTRIES_VISITED_COUNT

Color (discrete)

VETTING_LEVEL

Panel (discrete)

...

Options

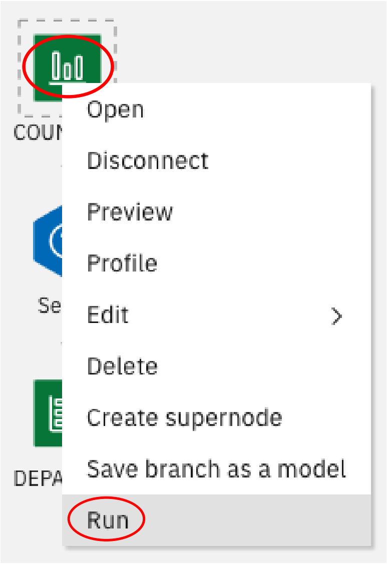
Appearance

Annotations

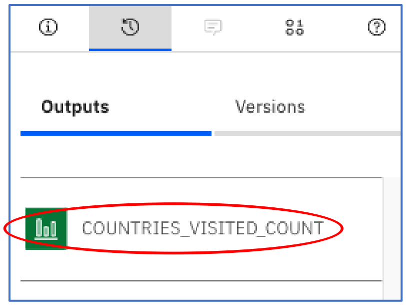
Cancel

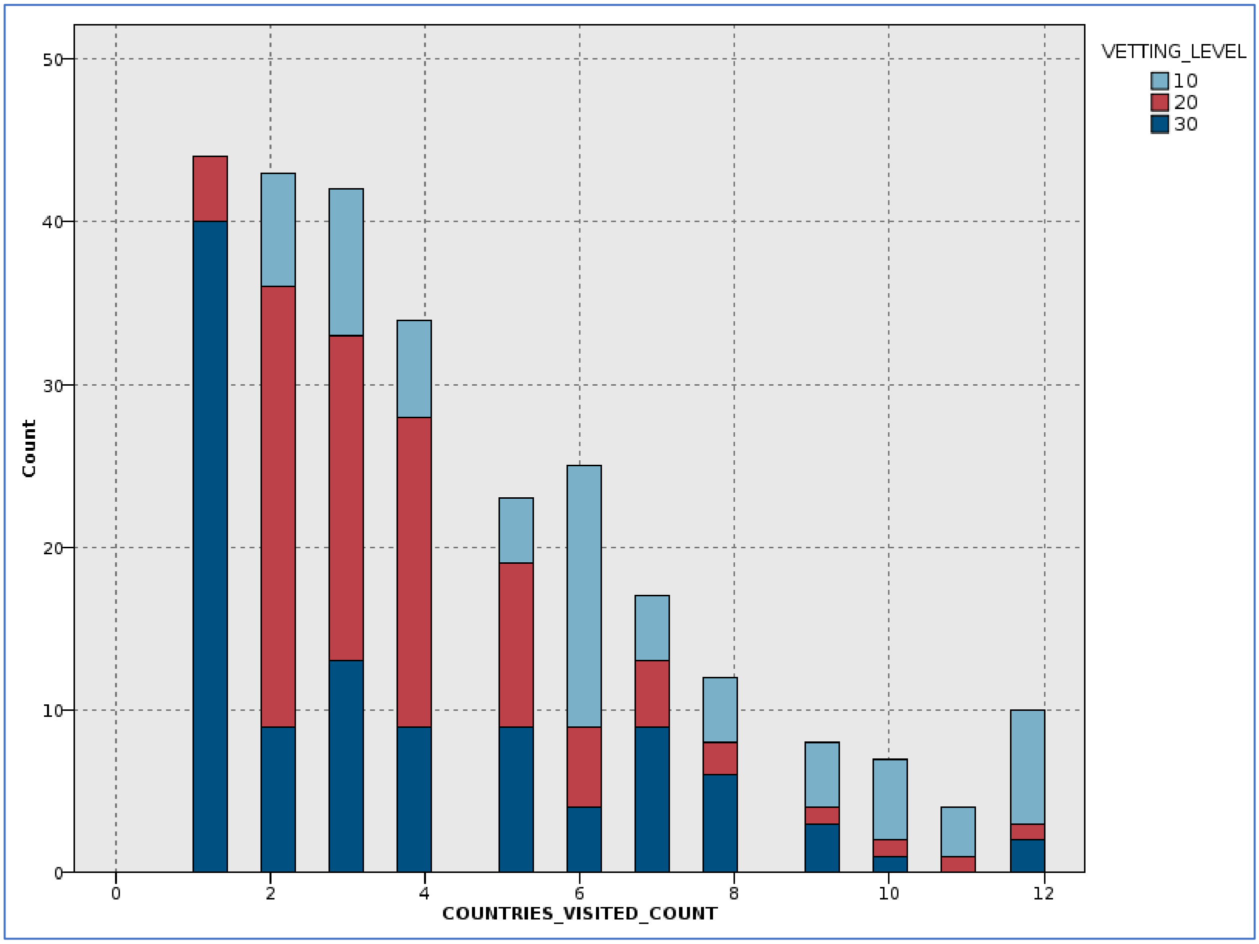
Save

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

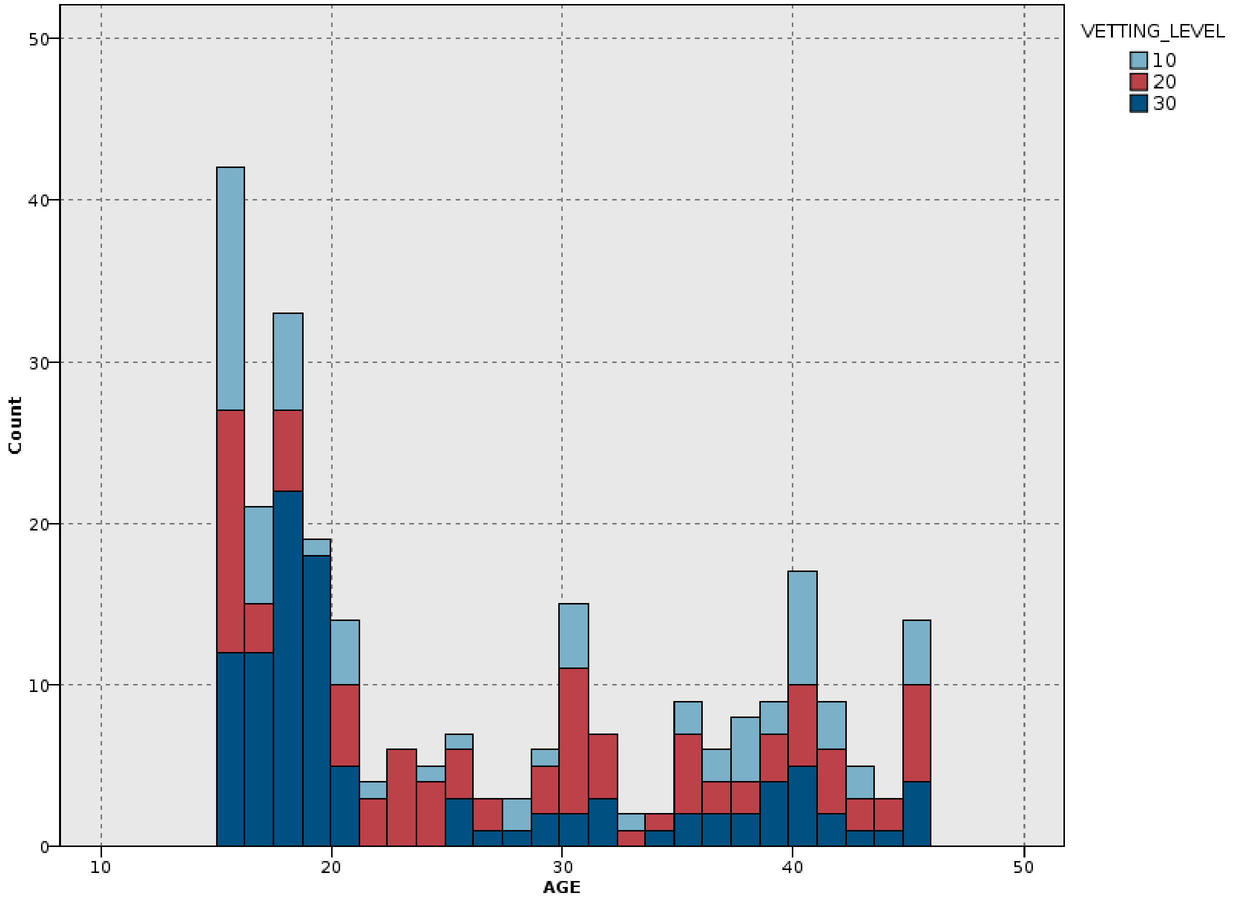


18. Double click on the **COUNTRIES_VISITED_COUNT** under the **Outputs** tab at the right of the screen.


Outputs
Versions
COUNTRIES_VISITED_COUNT


Count
0
10
20
30
40
50
0
2
4
6
8
10
12
COUNTRIES_VISITED_COUNT
VETTING_LEVEL
10
20
30

19. The general trend appears to be that the more countries visited, the higher likelihood to be a “High Risk”. 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**. Double click on the **AGE** in the **Outputs** pane to display the graph.



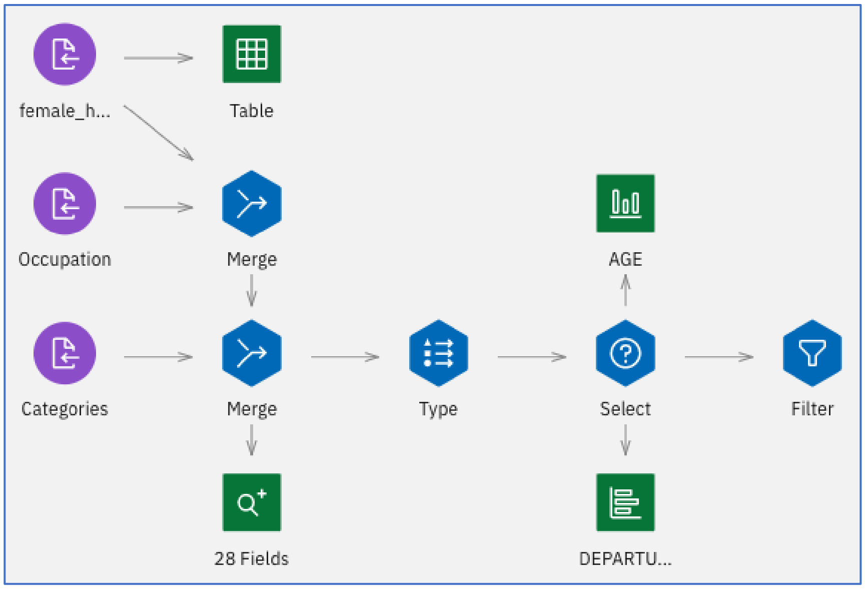
Step 6 - Prepare the Data for Modeling

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.

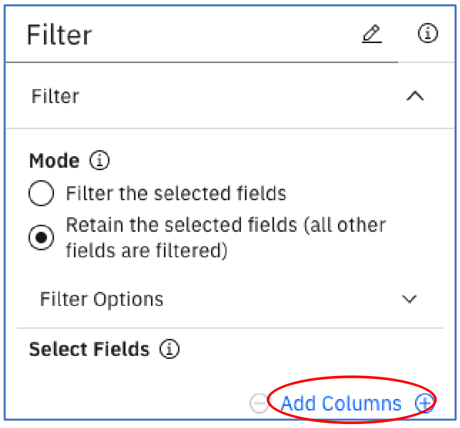
Filter node – The **Filter** node performs two functions. It specifies fields that can be dropped or the fields that should be retained. It also allows fields to be renamed. We will retain the following fields – VETTING_LEVEL, COUNTRIES_VISITED_COUNT, ARRIVAL_AIRPORT_REGION, DEPARTURE_AIRPORT_COUNTRY_CODE,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

Filter

Mode

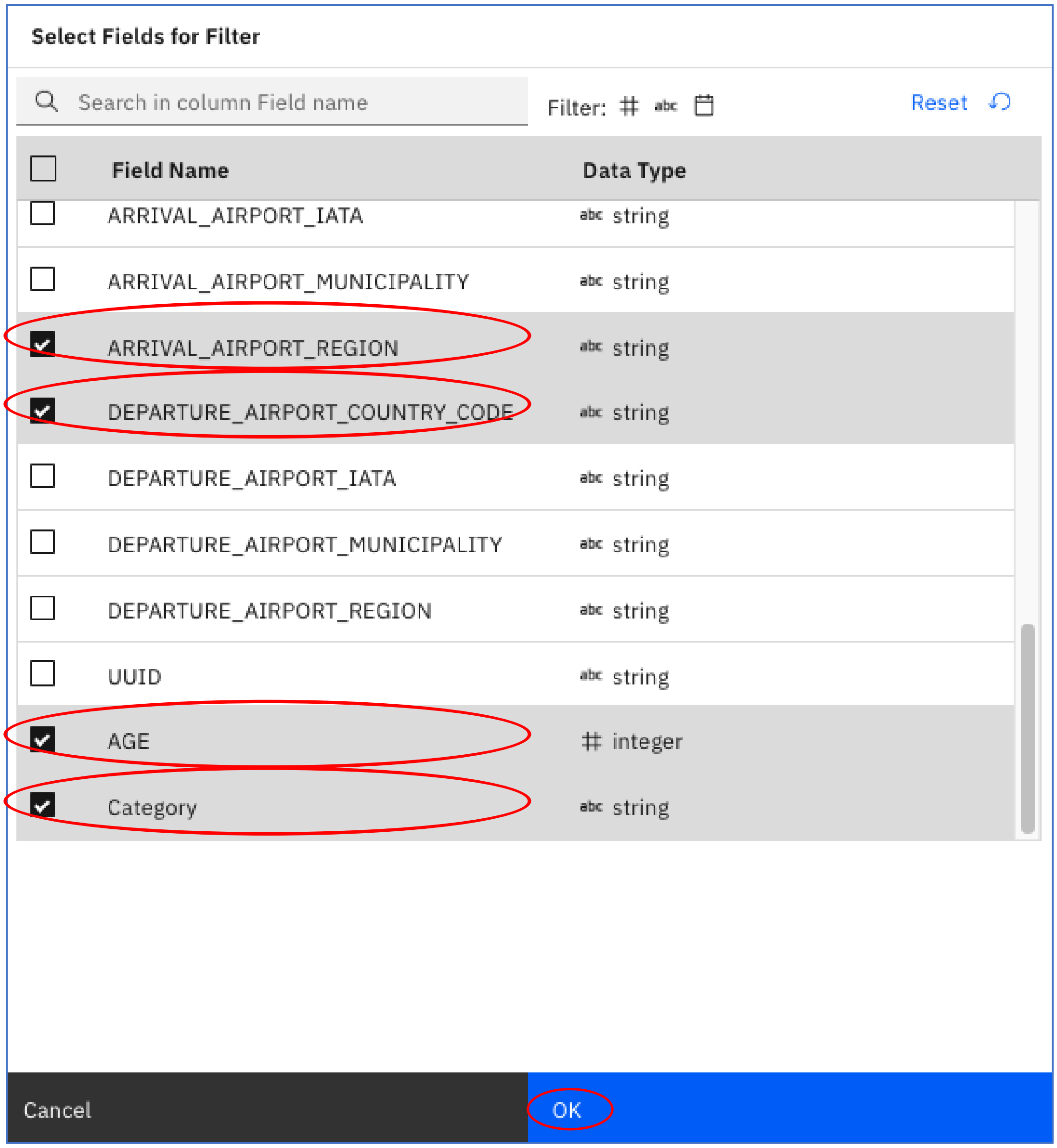
○ Filter the selected fields

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

Filter Options

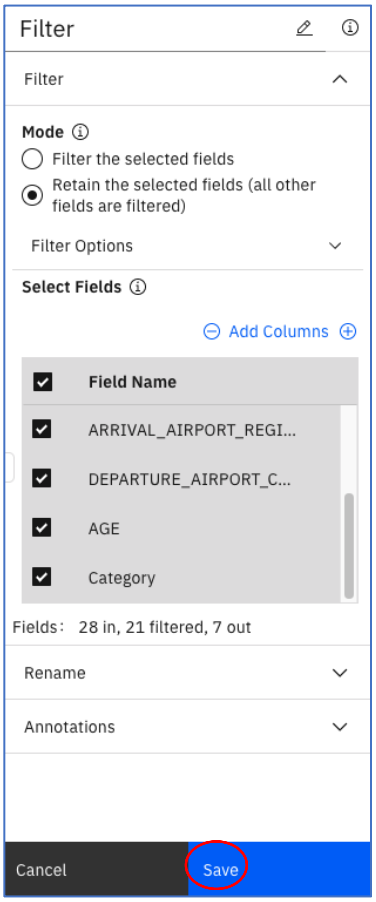
Select Fields

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.



☐	Field Name	Data Type
☐	ARRIVAL_AIRPORT_IATA	abc string
☐	ARRIVAL_AIRPORT_MUNICIPALITY	abc string
☑	ARRIVAL_AIRPORT_REGION	abc string
☑	DEPARTURE_AIRPORT_COUNTRY_CODE	abc string
☐	DEPARTURE_AIRPORT_IATA	abc string
☐	DEPARTURE_AIRPORT_MUNICIPALITY	abc string
☐	DEPARTURE_AIRPORT_REGION	abc string
☐	UUID	abc string
☑	AGE	# integer
☑	Category	abc string

4. Click **Save**.



Filter

Filter

Mode

○ Filter the selected fields

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

Filter Options

Select Fields

Add Columns

☑	**Field Name**
☑	ARRIVAL_AIRPORT_REGI...
☑	DEPARTURE_AIRPORT_C...
☑	AGE
☑	Category

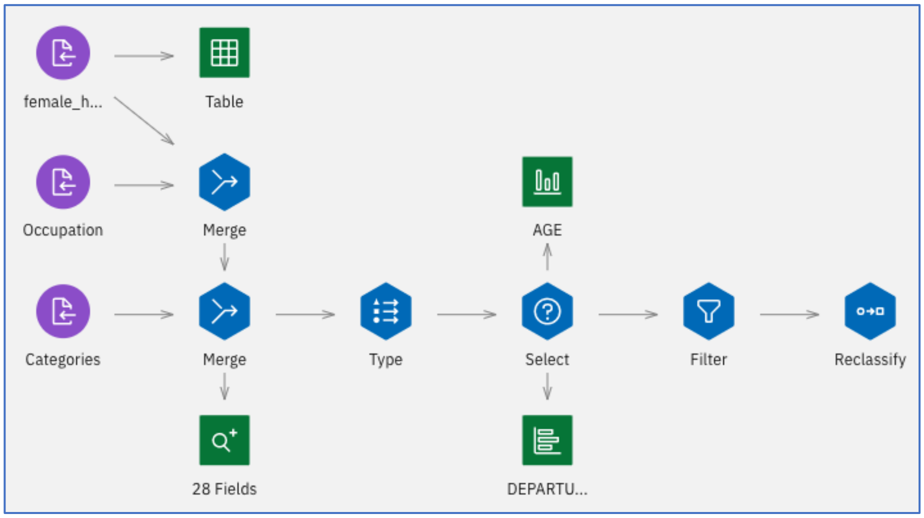
Fields: 28 in, 21 filtered, 7 out

Rename

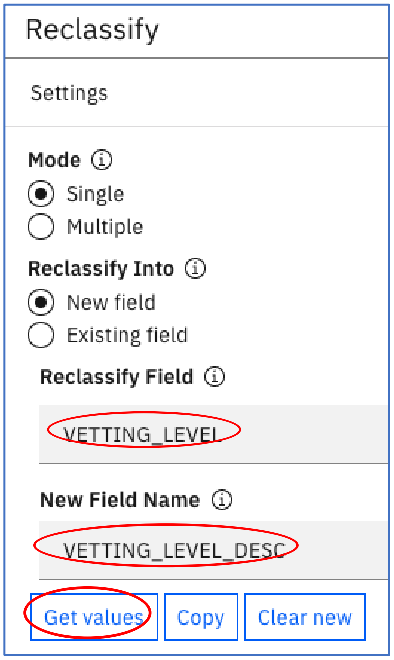
Annotations

Cancel 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.



6. 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

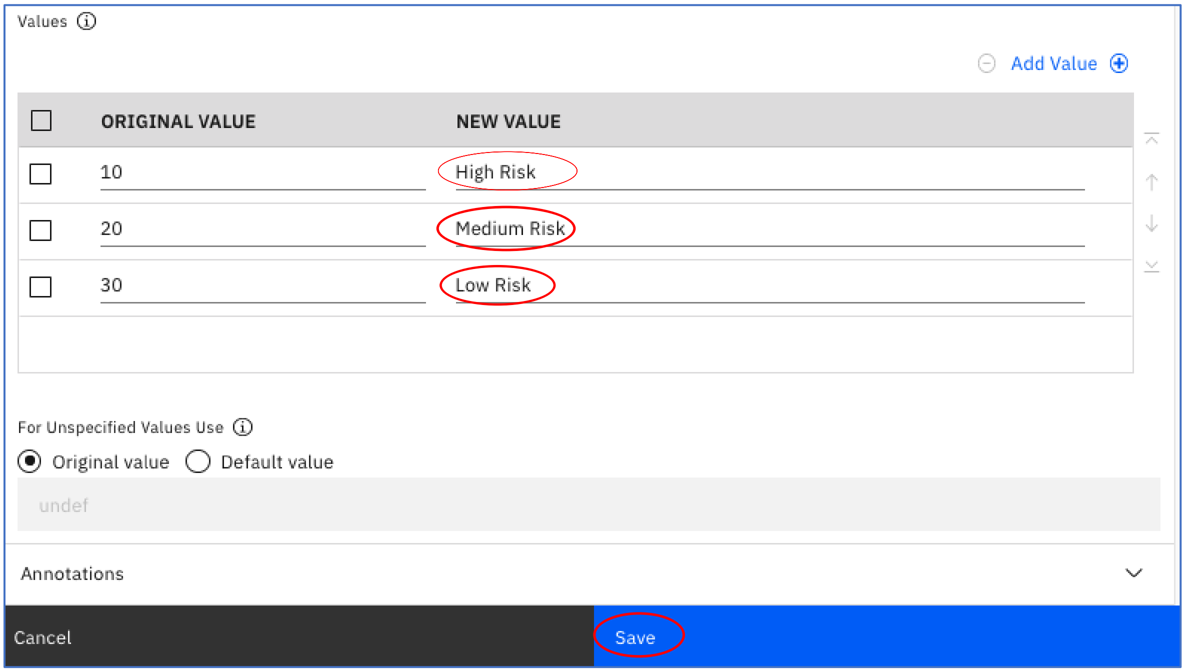
Reclassify Field ⓘ

VETTING_LEVEL

New Field Name ⓘ

VETTING_LEVEL_DESC

7. 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 ⓘ

□	ORIGINAL VALUE	NEW VALUE
□	10	High Risk
□	20	Medium Risk
□	30	Low Risk

For Unspecified Values Use ⓘ

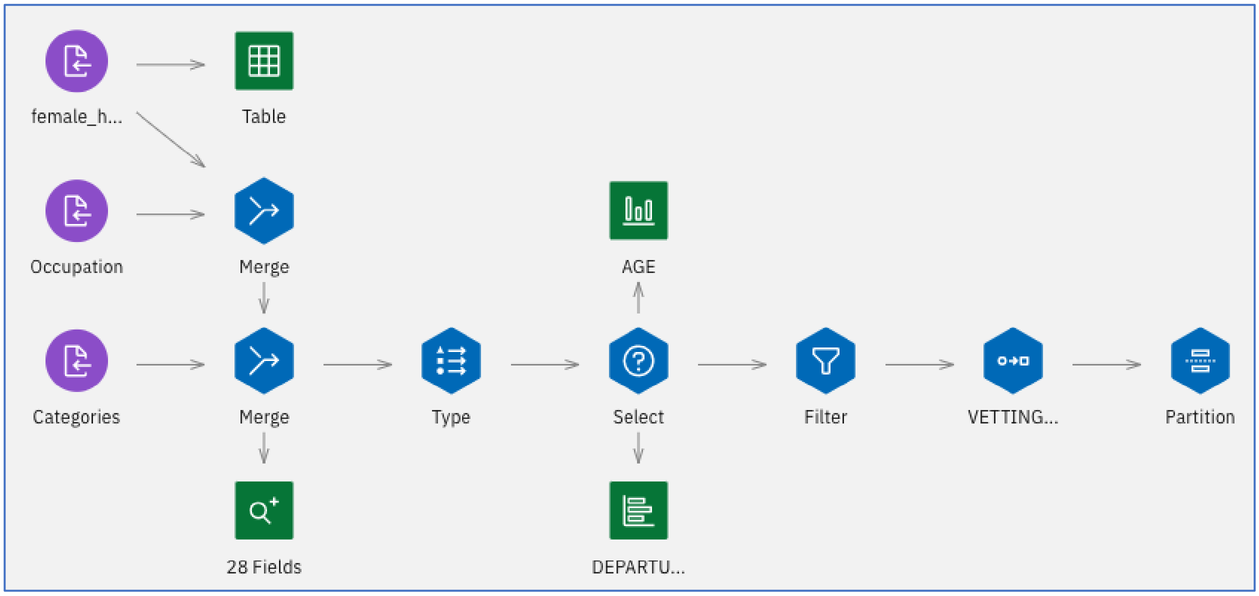
● Original value ○ Default value

undef

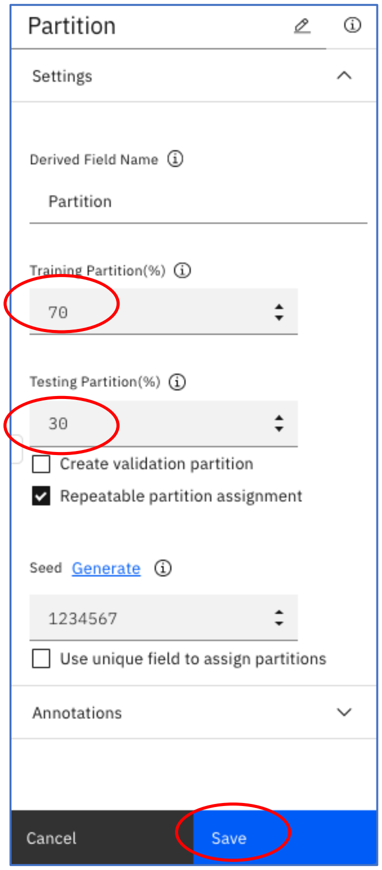
Step 7 - Modeling and Evaluation

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**.



Partition

Settings

Derived Field Name: Partition

Training Partition(%): 70

Testing Partition(%): 30

□ Create validation partition

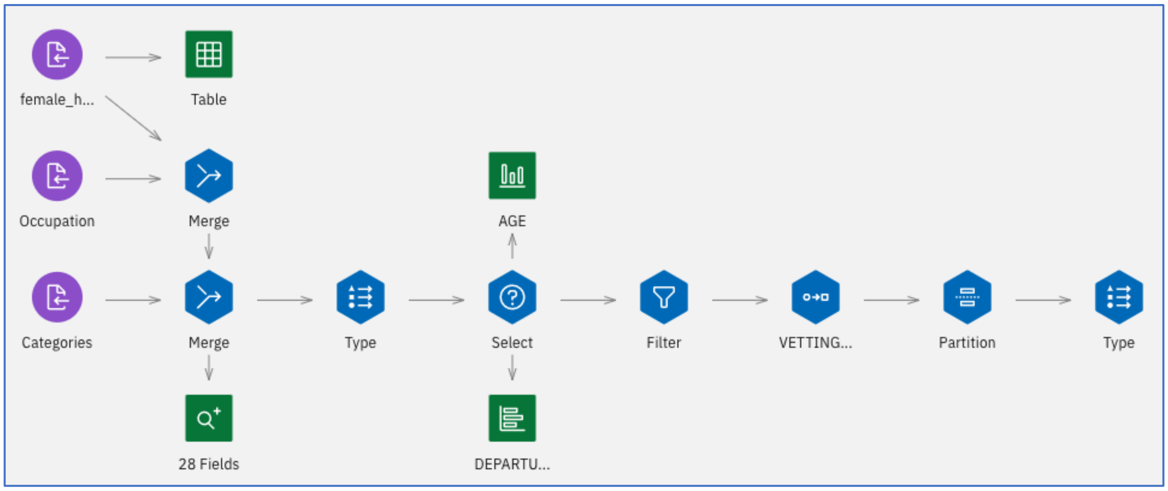
☑ Repeatable partition assignment

Seed Generate: 1234567

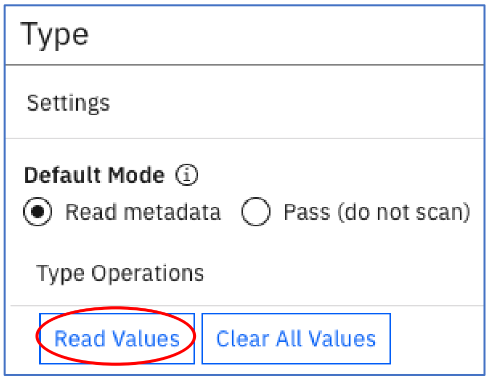
□ Use unique field to assign partitions

Annotations

3. Add a **Type** node 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

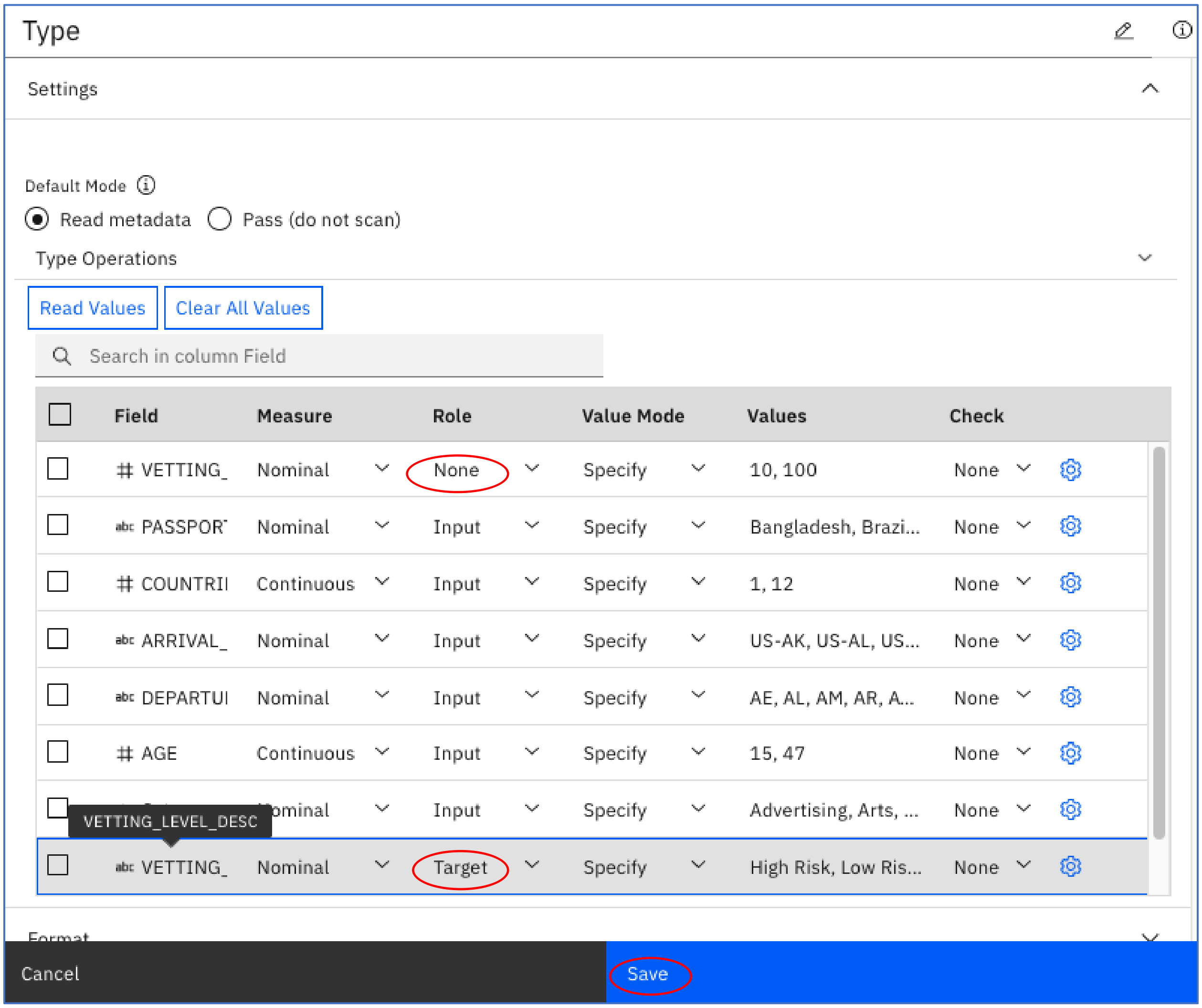
Settings

Default Mode ⓘ

● Read metadata ○ Pass (do not scan)

Type Operations

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

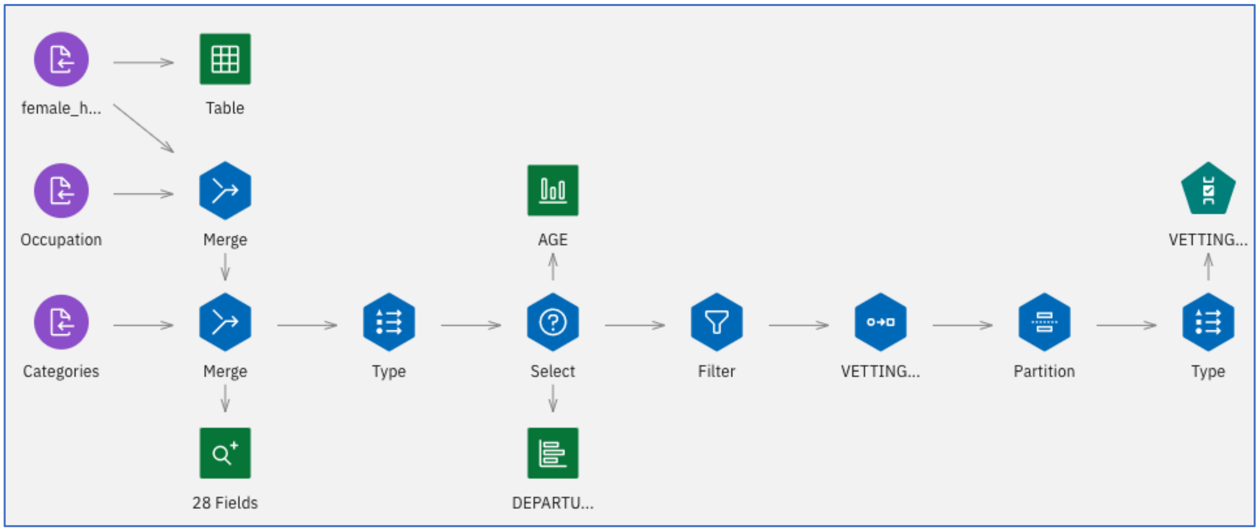
Default Mode ⓘ

● Read metadata ○ Pass (do not scan)

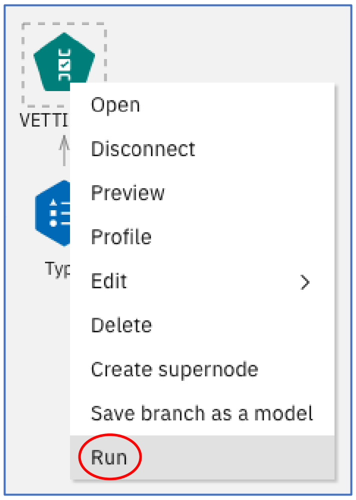
Type Operations

□	Field	Measure	Role	Value Mode	Values	Check
□	# VETTING_	Nominal	None	Specify	10, 100	None
□	abc PASSPOR	Nominal	Input	Specify	Bangladesh, Brazi...	None
□	# COUNTRI	Continuous	Input	Specify	1, 12	None
□	abc ARRIVAL_	Nominal	Input	Specify	US-AK, US-AL, US...	None
□	abc DEPARTU	Nominal	Input	Specify	AE, AL, AM, AR, A...	None
□	# AGE	Continuous	Input	Specify	15, 47	None
□		Nominal	Input	Specify	Advertising, Arts, ...	None
□	abc VETTING_	Nominal	Target	Specify	High Risk, Low Ris...	None

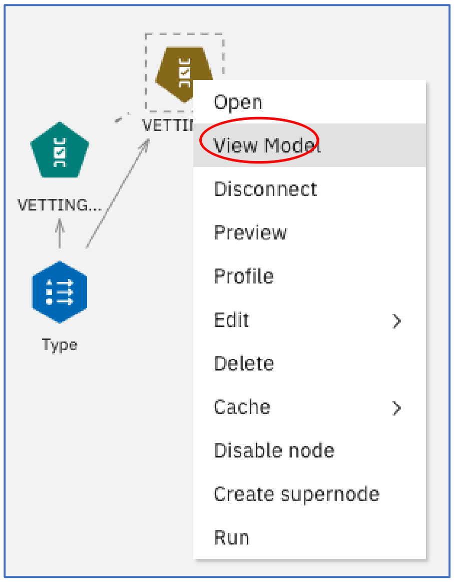
6. Add a **Feature Selection** node by clicking on the **Modeling** menu item in the Node palette and dragging the **Feature Selection** node onto the canvas to the right of the **Type** node. Connect the **Type** node to the **Feature Selection** node. The canvas should appear as below. The Feature Selection node provides the correlation of each of the input features to the target field. It gives an indication of the Importance of each feature.



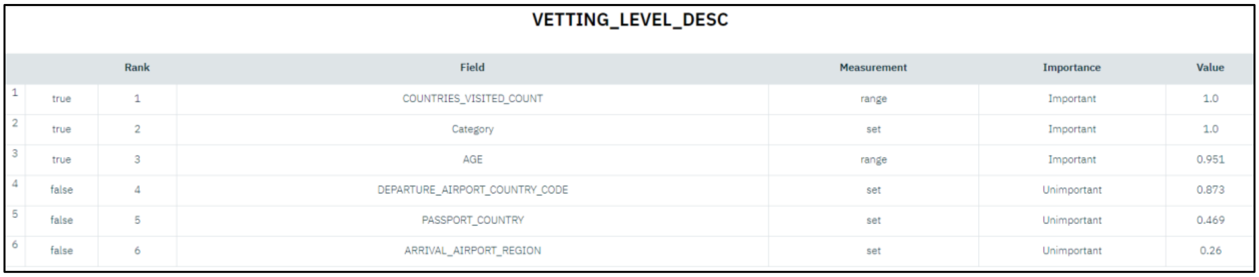
7. Right-click on **Feature Selection** and click **Run**.



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



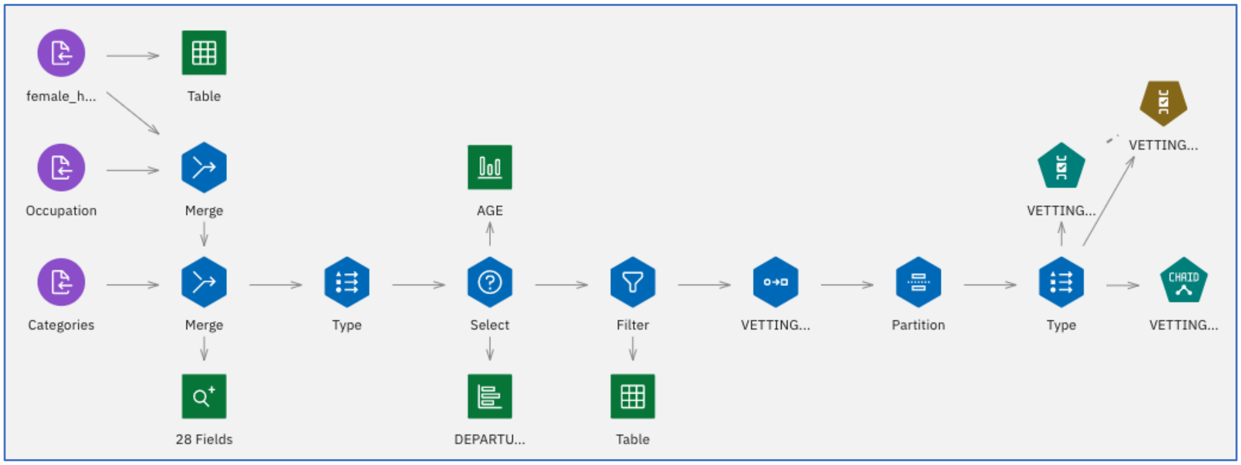
9. The Feature Selection output is displayed. Note that the ranges for what is Important can be changed in the modeling options. According to the default criteria, the COUNTRIES_VISITED_COUNT, Category, and AGE are the most important features.



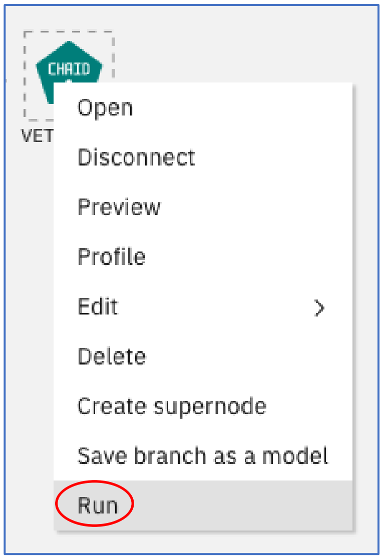
VETTING_LEVEL_DESC

		Rank	Field	Measurement	Importance	Value
1	true	1	COUNTRIES_VISITED_COUNT	range	Important	1.0
2	true	2	Category	set	Important	1.0
3	true	3	AGE	range	Important	0.951
4	false	4	DEPARTURE_AIRPORT_COUNTRY_CODE	set	Unimportant	0.873
5	false	5	PASSPORT_COUNTRY	set	Unimportant	0.469
6	false	6	ARRIVAL_AIRPORT_REGION	set	Unimportant	0.26

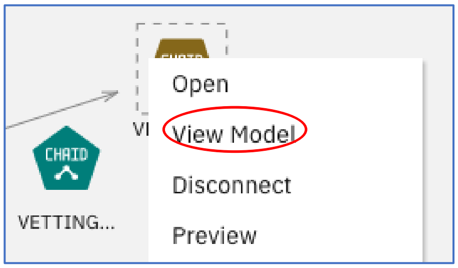
7. Add a **CHAID** node by clicking on the **Modeling** menu item in the Node palette and dragging the **CHAID** node onto the canvas to the right of the **Type** node. Connect the **Type** node to the **CHAID** node. The canvas should appear as below.



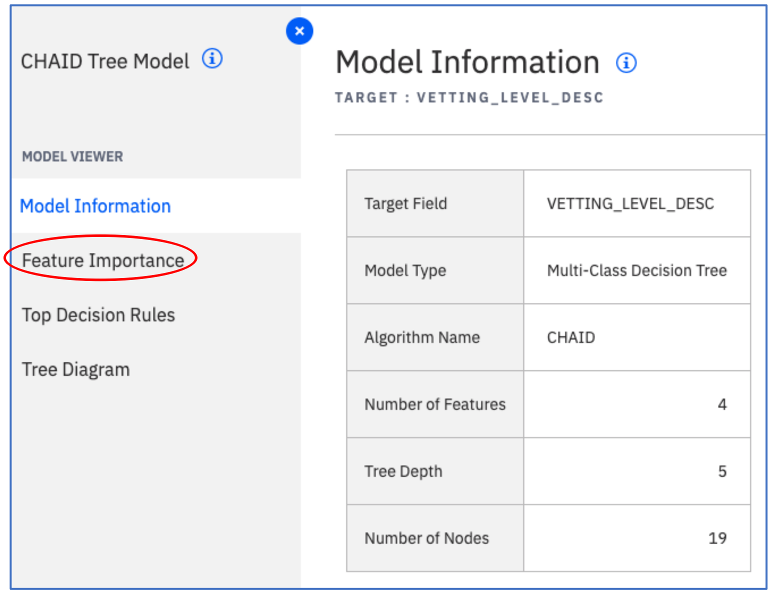
8. Right-click on the CHAID node and click Run.



9. 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**.



10. The Model Information is displayed. Click on **Feature Importance**.

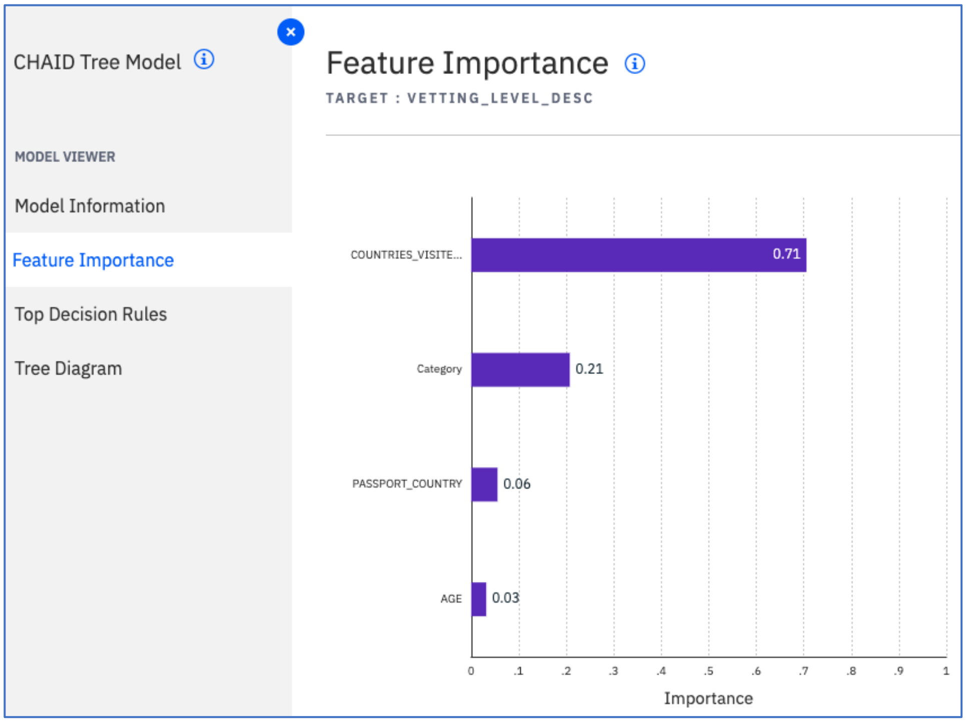


Model Information

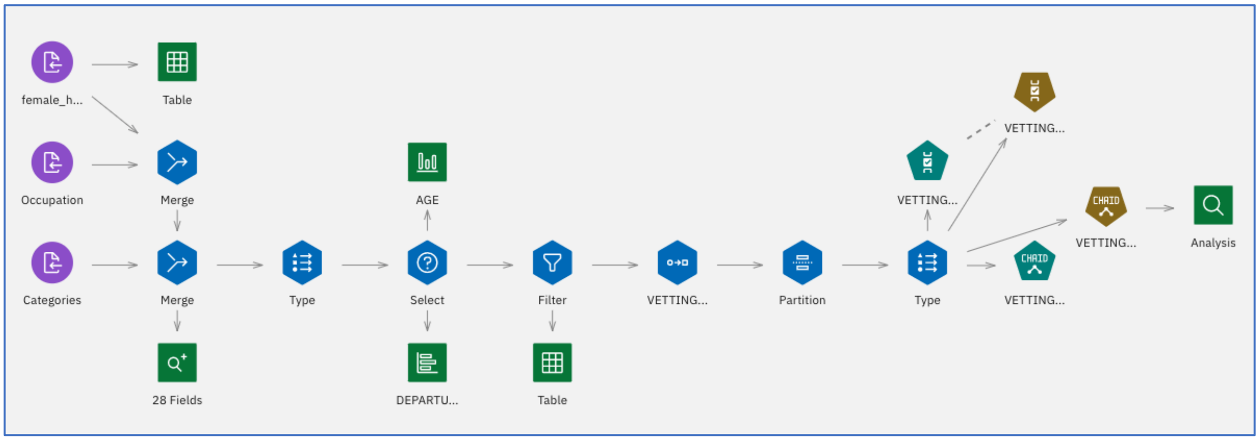
TARGET : VETTING_LEVEL_DESC

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

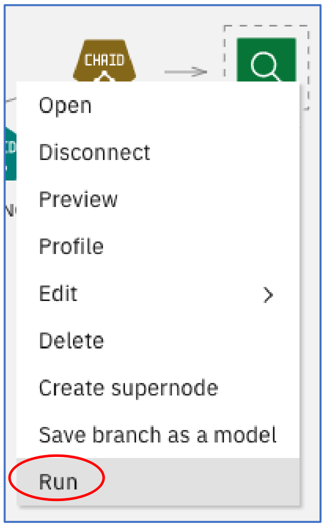
11. Feature Importance is displayed with similar results to the Feature Selection output. Click on **Tree Diagram and/or Top Decision Rules** to see the algorithm output.



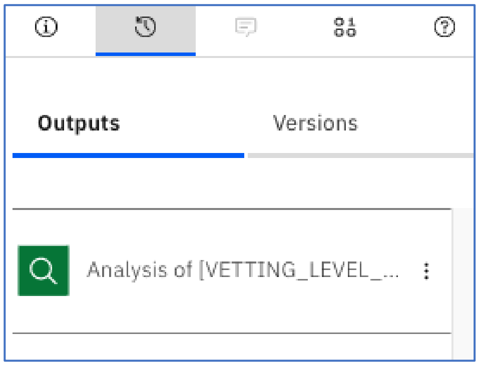
7. Add an Analysis node by clicking on the Output menu item in the Node palette and dragging the Analysis node onto the canvas to the right of the CHAID Model node. Connect the CHAID Model node to the Analysis node. The canvas should appear as below. The canvas should appear as below.



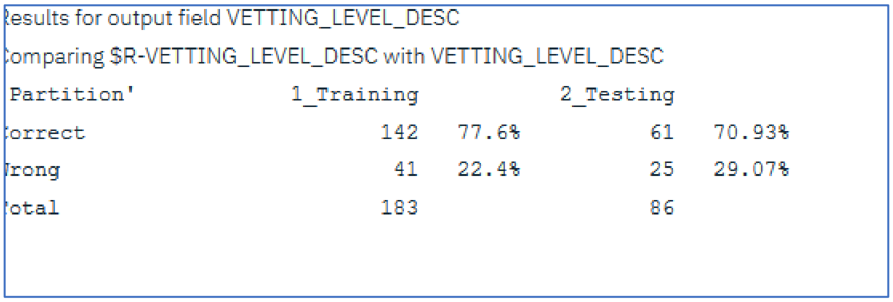
8. Right-click the **Analysis** node and click **Run**.



9. Double-click on the Analysis results in the Output area.

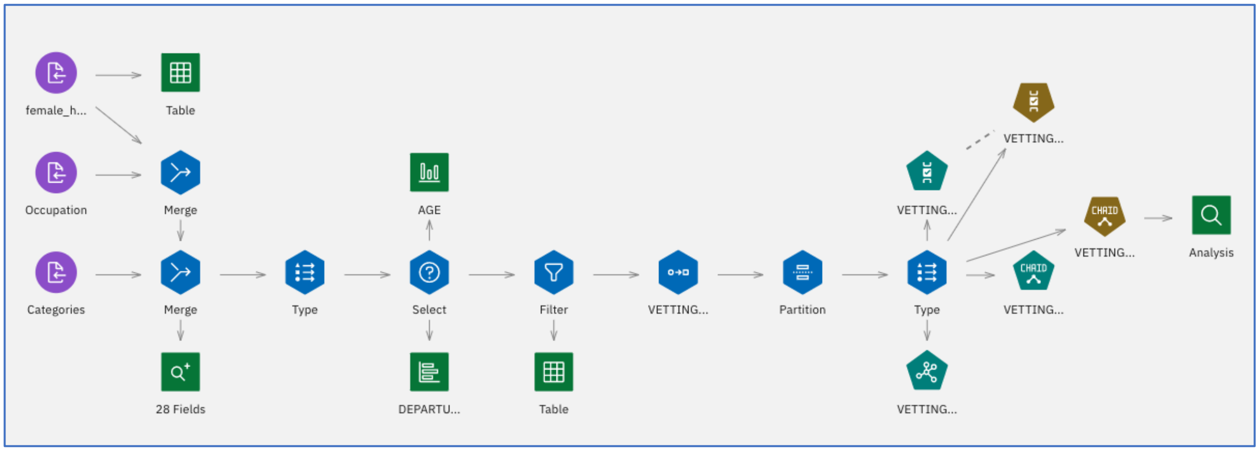


10. Accuracy results are displayed for the CHAID algorithm.

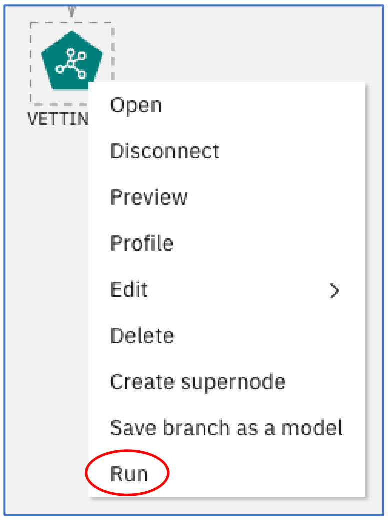


```
Results for output field VETTING_LEVEL_DESC
Comparing $R-VETTING_LEVEL_DESC with VETTING_LEVEL_DESC
 Partition'            1_Training           2_Testing
Correct                      142    77.6%          61    70.93%
Wrong                         41    22.4%          25    29.07%
Total                        183                   86
```

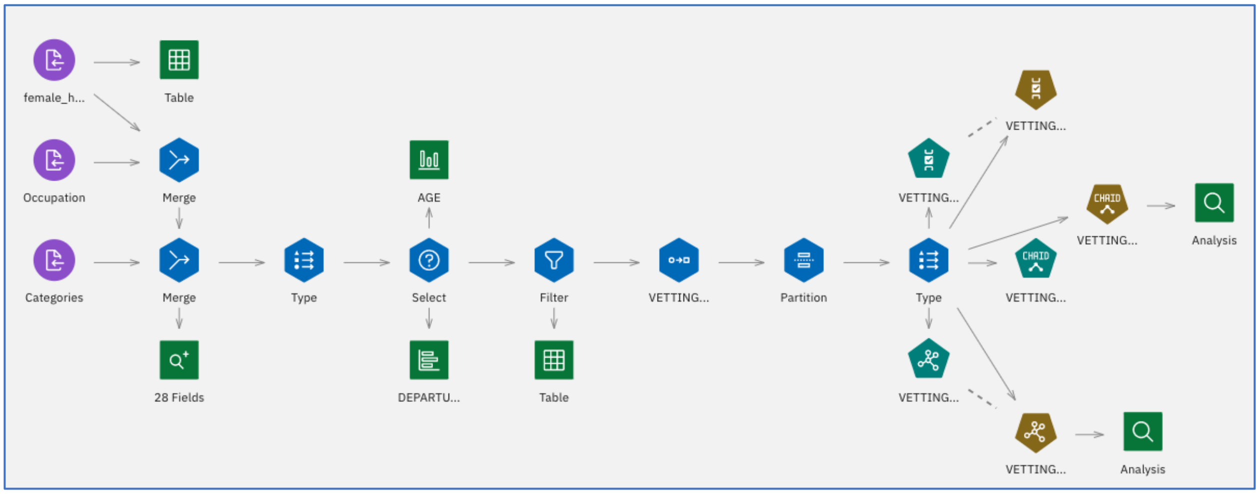
10. Add a **Random Forest** node by clicking on the **Modeling** menu item in the Node palette and dragging the **Random Forest** node onto the canvas underneath the **CHAID** node. Connect the **Type** node to the **Random Forest** node. The canvas should appear as below.



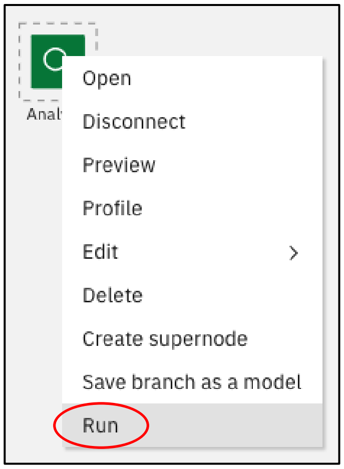
11. Right-click the **Random Forest** node and click **Run**.



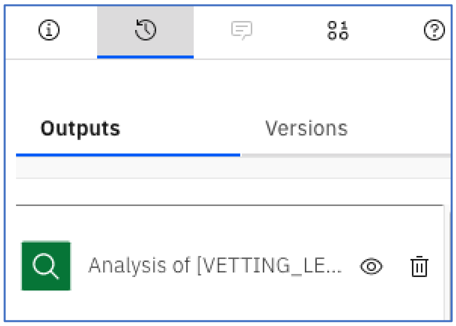
12. A **Random Forest Model** node is created. The **Random Forest Model** node does not have a **View Model** option. Add an **Analysis** node to the right of the **Random Forest Model** node by clicking on the **Outputs** menu of the Node Palette. Connect the **Analysis** node to the **Random Forest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



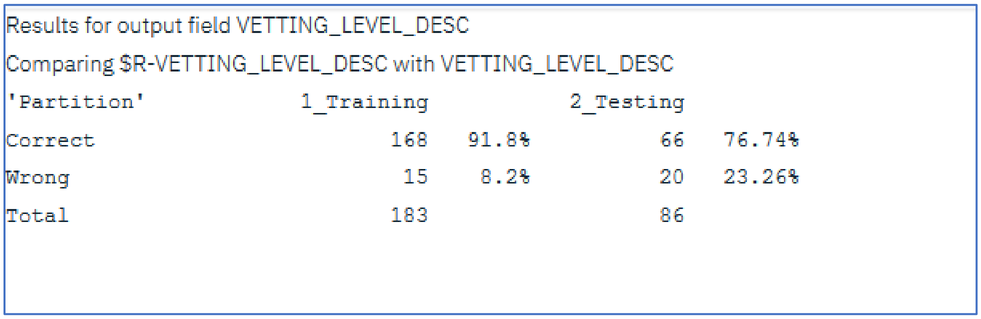
13. Right-click on the **Analysis** node and click **Run**.



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



15. The results appear below. Based on the results, it appears the Random Forest model is overfitting given the disparity between training and testing results.

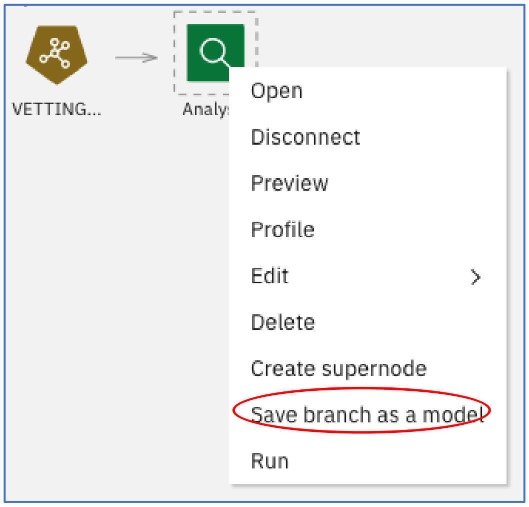


```
Results for output field VETTING_LEVEL_DESC
Comparing $R-VETTING_LEVEL_DESC with VETTING_LEVEL_DESC
'Partition'             1_Training           2_Testing
Correct                        168    91.8%           66    76.74%
Wrong                           15     8.2%           20    23.26%
Total                          183                    86
```

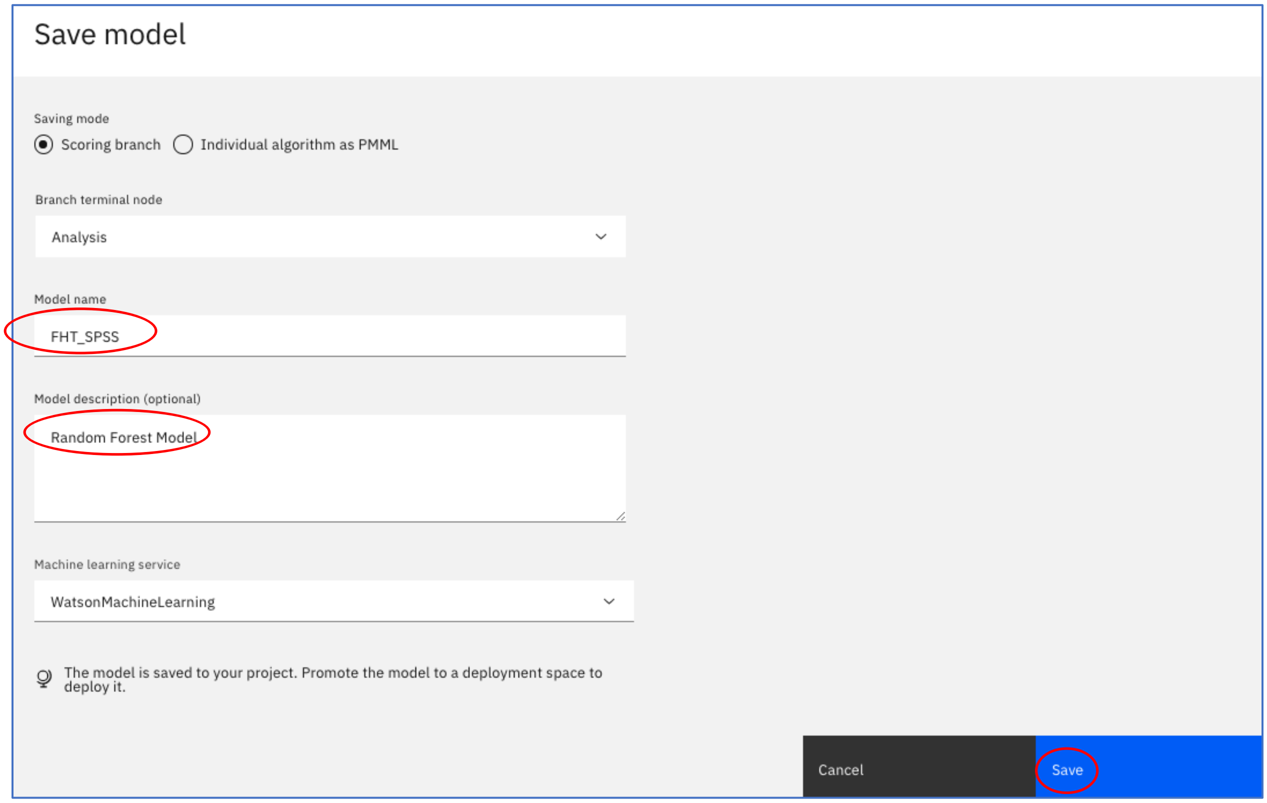
Step 8 - Saving a Model

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.

1. Right click on the Random Forest 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

Model name

FHT_SPSS

Model description (optional)

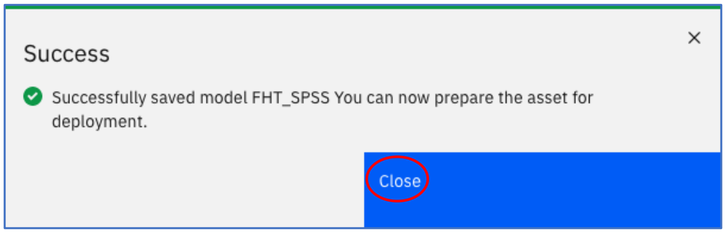
Random Forest Model

Machine learning service

WatsonMachineLearning

The model is saved to your project. Promote the model to a deployment space to deploy it.

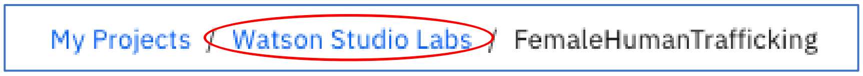
3. Click **Close**.



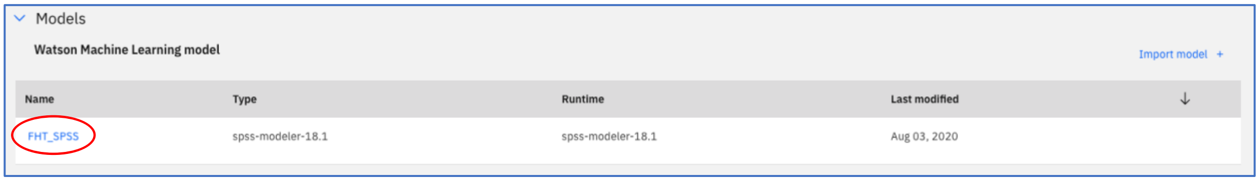
Success

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

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



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



Name	Type	Runtime	Last modified
FHT_SPSS	spss-modeler-18.1	spss-modeler-18.1	Aug 03, 2020

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.