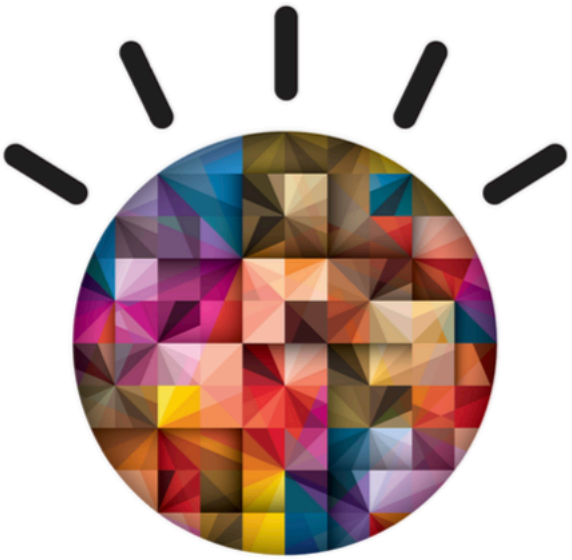
IBM Analytics



Data Science Workshop

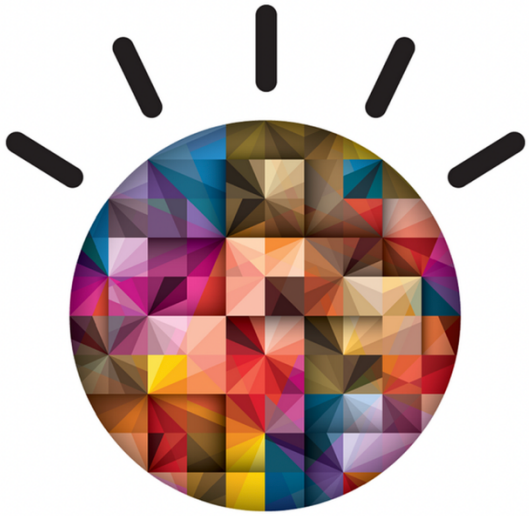
SPSS Modeler Guide

Contents

SPSS Modeler

IBM SPSS Modeler is a comprehensive data science platform, designed to bring **predictive intelligence** to everyday business problems, enabling front-line employees to make more **effective decisions** and **improve outcomes.**

This enables organizations to improve business processes and help people or systems **consistently make the right decisions** by delivering recommended actions at the point of impact. The result is a **rapid return on investment** (ROI) and the ability to **proactively** and **repeatedly reduce costs** and **increase productivity**.



This hands-on Data Science workshop is an instructor led session using IBM's data mining and predictive modeling software and is designed for those who are familiar with predictive analytics, as well as for beginners. Through this workshop you will experience firsthand how IBM SPSS Modeler works and how easy it is to implement predictive analytics.

Exercise 1: Predict in 20 minutes

Use Case

Goal: Approach:

Identify who has high probability to cancel the contract (Churn) to help marketing team to create a list of customer targets to receive specific marketing campaign

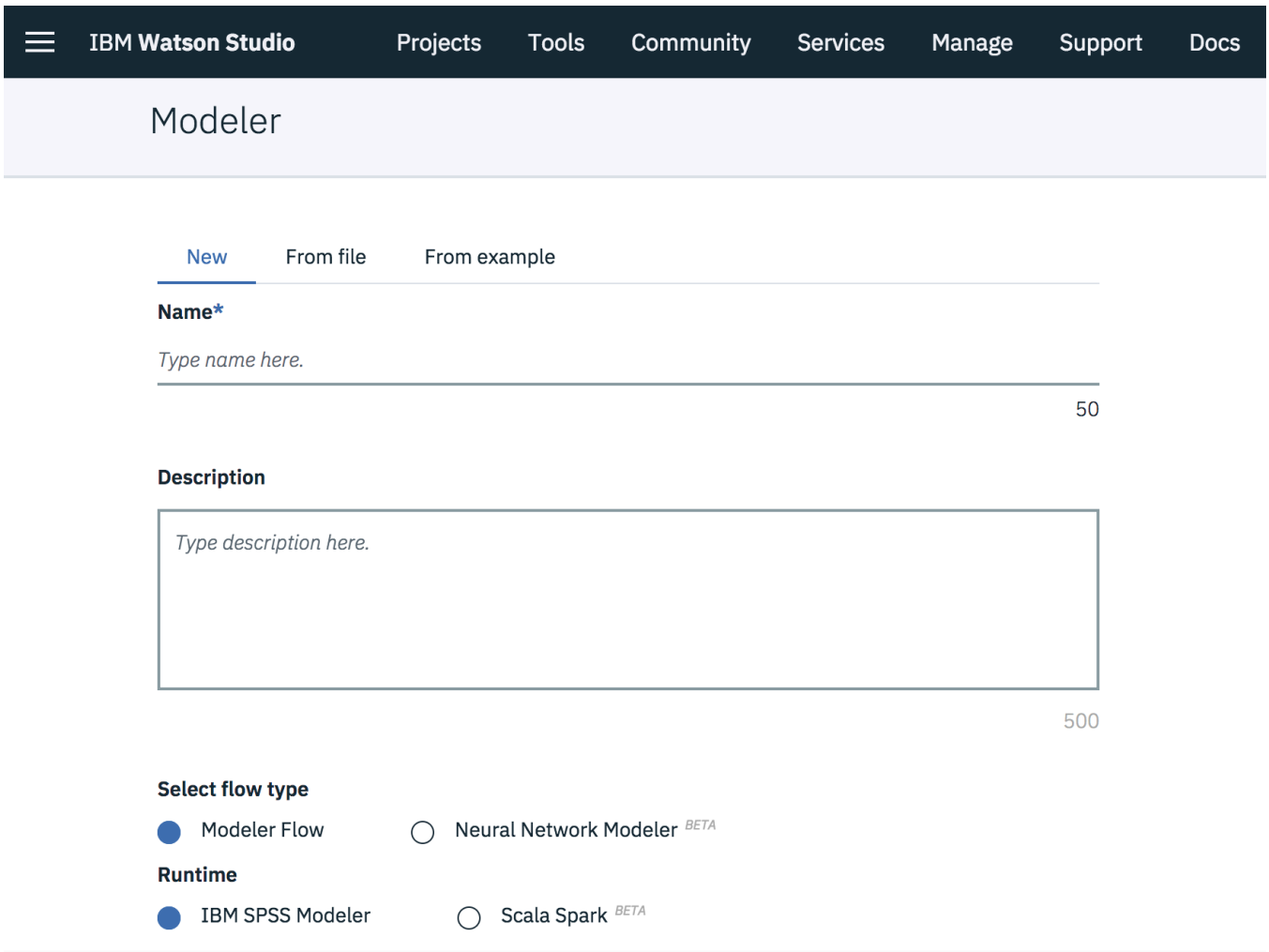
- Prepare data for modeling
- Define which fields to use
- Choose the modeling technique
- Automatically generate a model to identify who has cancel
- Review results

Why?

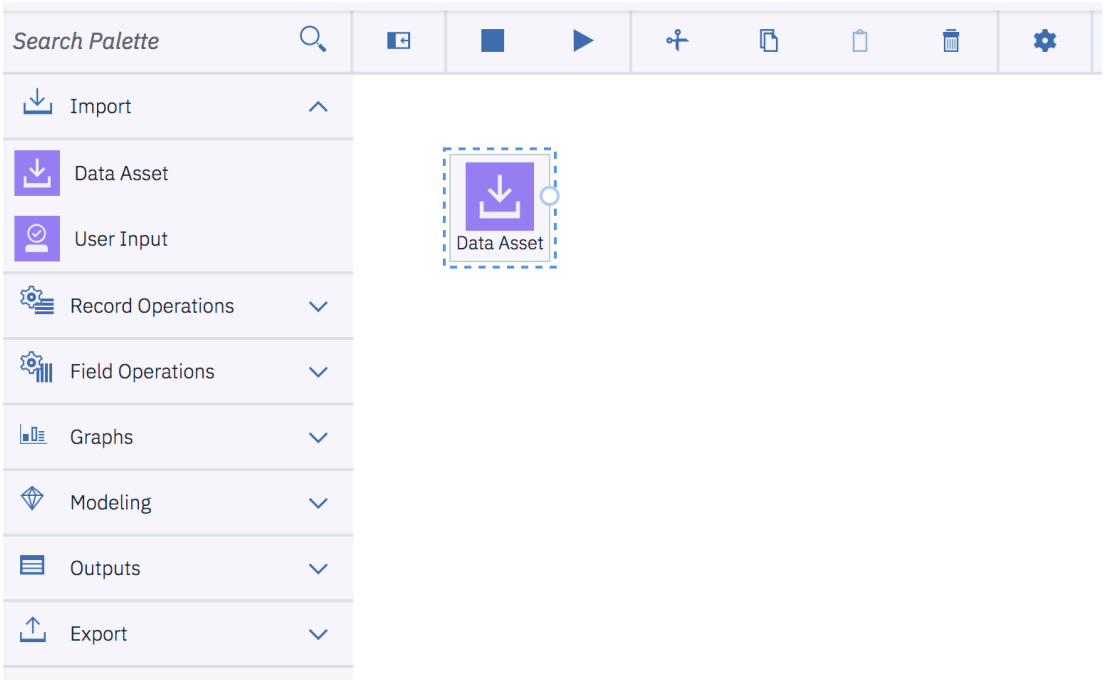
To save marketing cost and reduce the customer churn, identify those likely to churn and focus marketing efforts on those customers.

Predictive in 20 Minutes

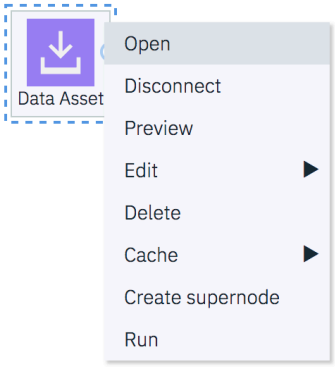
1. Open our Project and add ***Customer.csv*** file to **Data assets**.
2. Create a new model. Click on **+ New Modeler Flow**.
3. Type a name and description for our model and click **Create**.



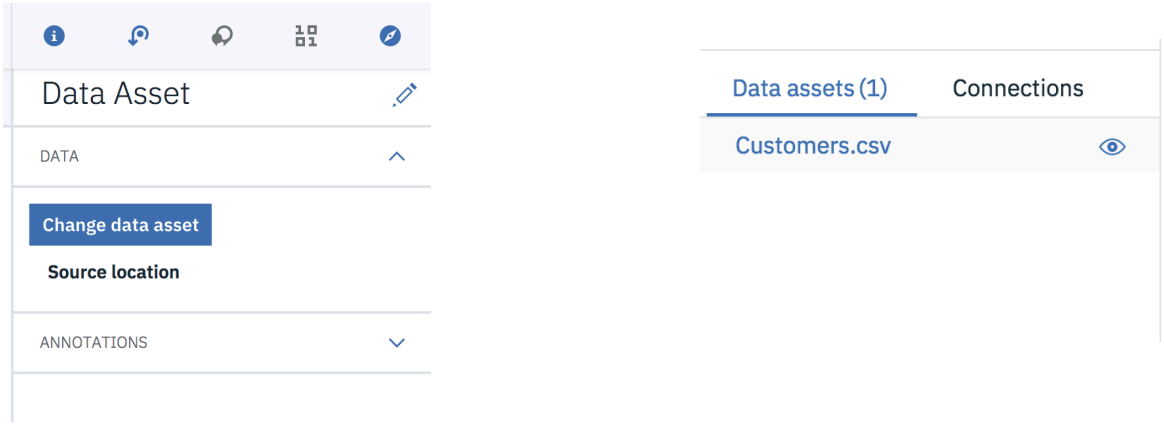
4. From Import, add the **Data Asset** node to the workflow.



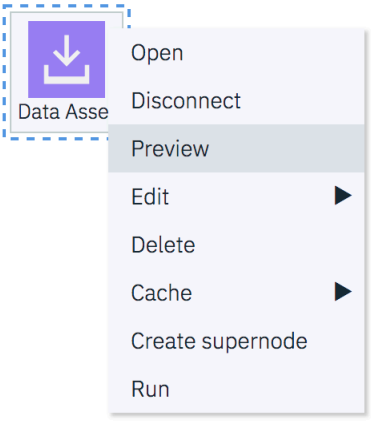
5. Access the Data Asset properties. Right click on Data Asset and select **Open**



6. Click on **Change data asset** to select the **Customer.csv** file already load in the project. Click **OK** and **SAVE**.



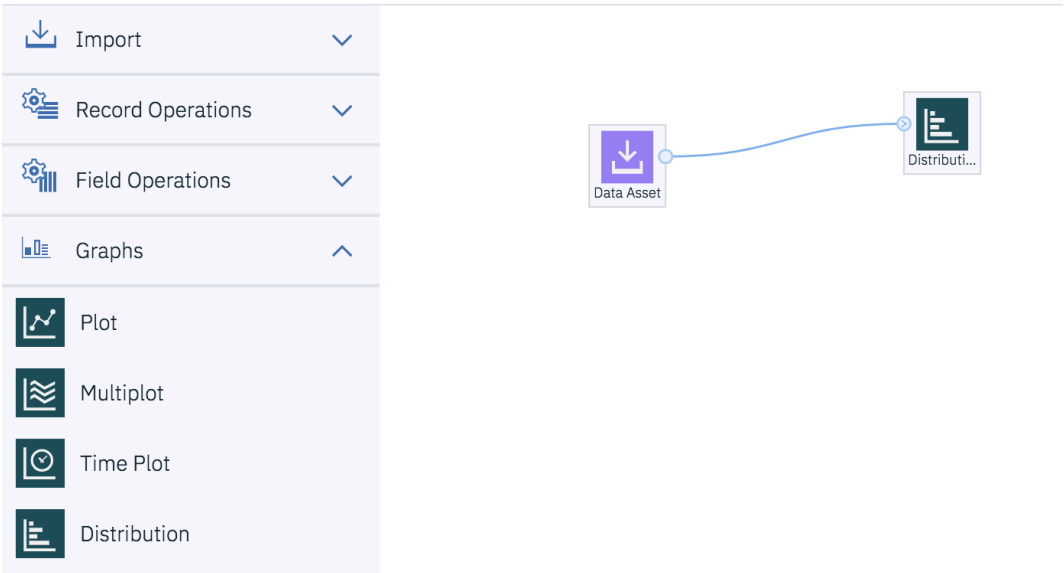
7. Rigth click on Data Asset and select **Preview**, to visualize the data available on file.



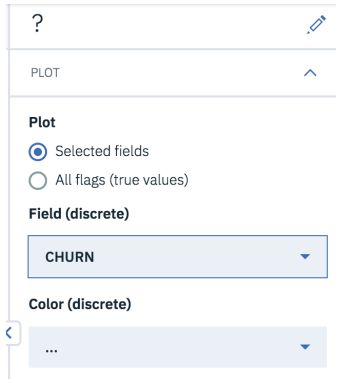
8. To go back to the workflow, click on the workflow name available on the top: My Projects / Project Name / **Workflow Name** / Preview.

	ID Decimal	Sex String	Status String	Children Decimal	Est_Income Decimal	Car_Owner String	Usage Decimal	Age Decimal
2	6	M	M	2	29616	N	75.29	49.426667
3	8	M	M	0	19732.8	N	47.25	50.673333
4	11	M	S	2	96.33	N	59.01	56.473333
5	14	F	M	2	52004.8	N	28.14	25.14
6	17	M	M	2	53010.8	N	58.87	18.84
7	18	M	M	1	75004.5	N	58.72	64.8

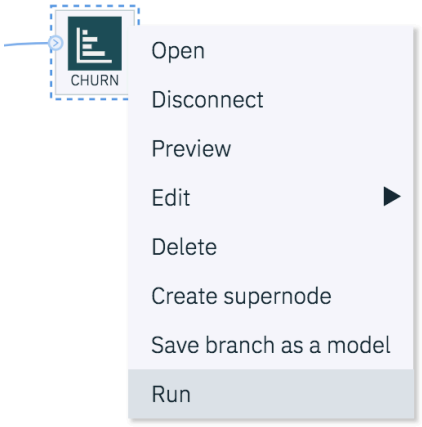
9. From Graphs options, add a Distribution node. Connect it to the data souce. Click and hold the button on the first node, move the cursor to the second node and release when the cursor is on the mark of second node.



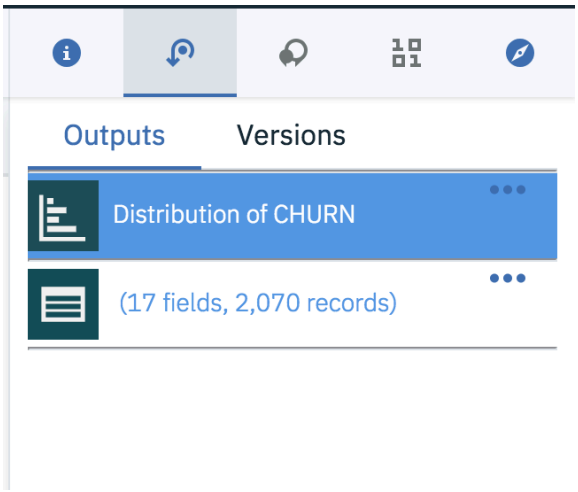
10. Double-click on the Distribution node to access the properties. At **Field (discrete)**, select **CHURN** and click save.



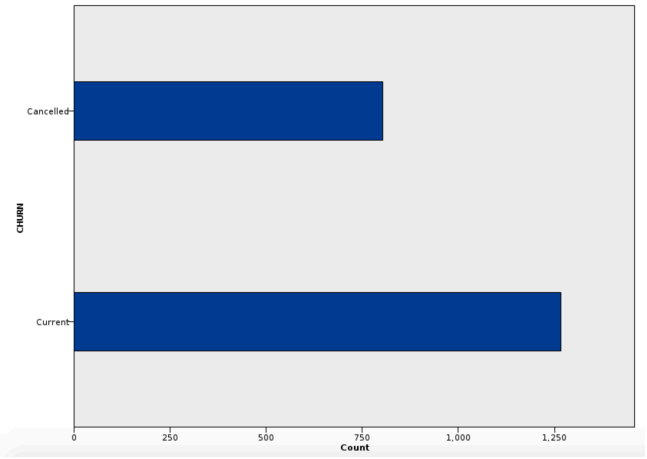
11. Note that distributin node was rename to CHURN. Right click on it and select **Run**.



12. Double-click on the graph Distribution of CHURN to open it.

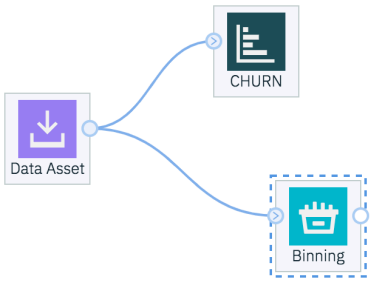


13. The resulting graph shows that of the 2070 customers in this data set, around 40% of customers have cancelled their contract. The task is to build a model to understand the relationships within the data that led to the canceled their contracts.

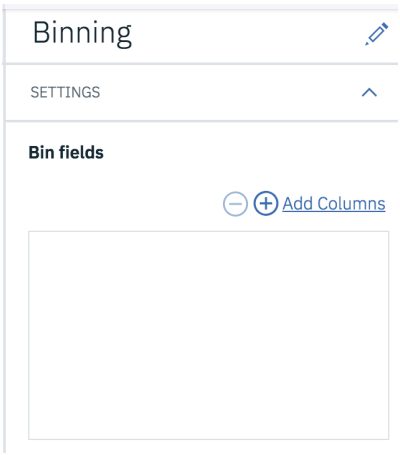


14. From the Fild Operations, add a Binning node to the workflow and connect it to the data source. The Binning node allows you to automatically generate bins (categories) using several techniques. In this case, we will be creating categories from the continuous variable Age.

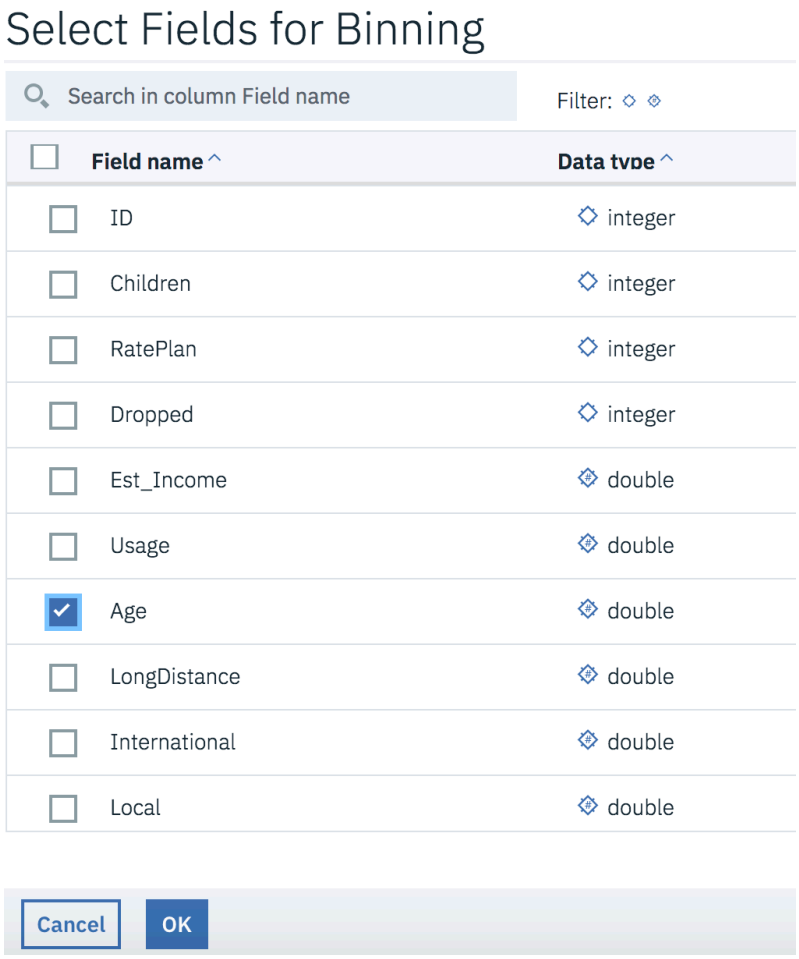
15. Click and hold the button on the Data Asset node, move the cursor to the Binning node and release when the cursor is on the mark of second node.



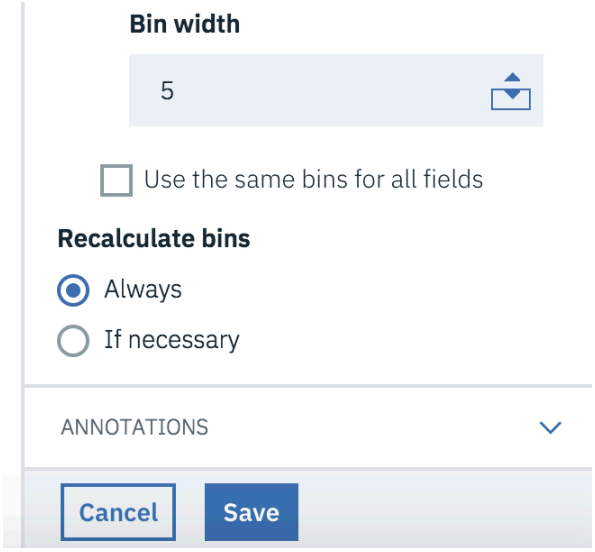
16. Double-click on **Binning** to open properties.



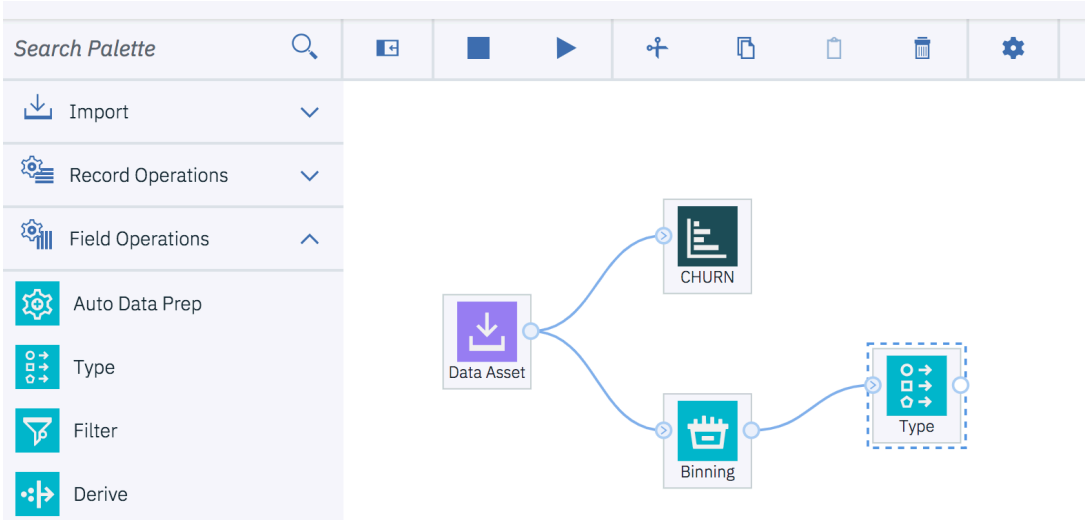
17. Click on **Add Columns** and select **Ages**. Click OK.



18. Scroll down in the **Binning** properties, change the option **Bin width** to **5**. Click Save.



19. From the Field Operations, add a Type node to the workflow and connect it to the Binning node.



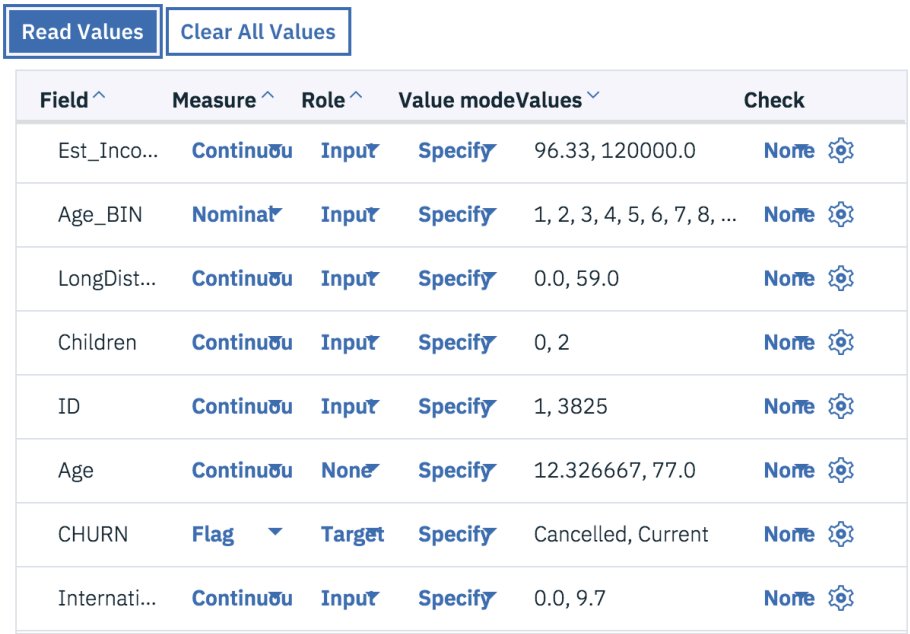
20. Open the Type node and click the Read Values button to scan the data as well as to display and update the range of values.

Using the drop-down box under Role, modify the following Fields:

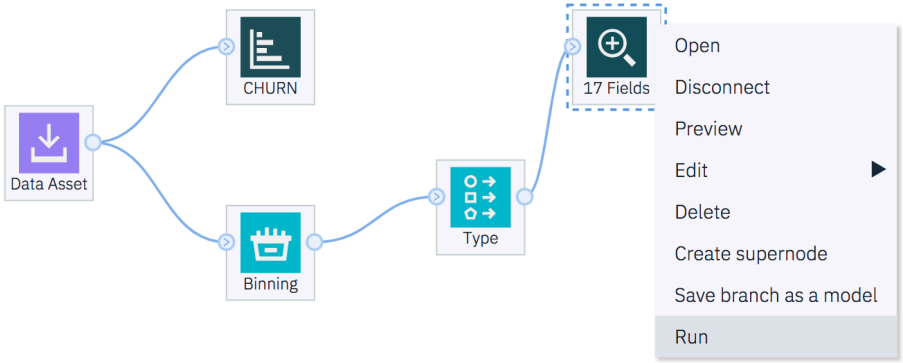
- ID = Record ID
- Age = None
- CHURN = Target

The Measurement of our Target should be set to Flag, which reflects two potential Values: Cancelled or Current. The remaining, including our new Age_Bin, will remain as Inputs in our analysis.

Save

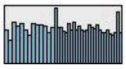
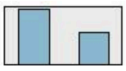
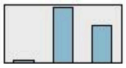


21. From the Output, add the Data Audit node and connect it to the Type node. Right-click to select Run

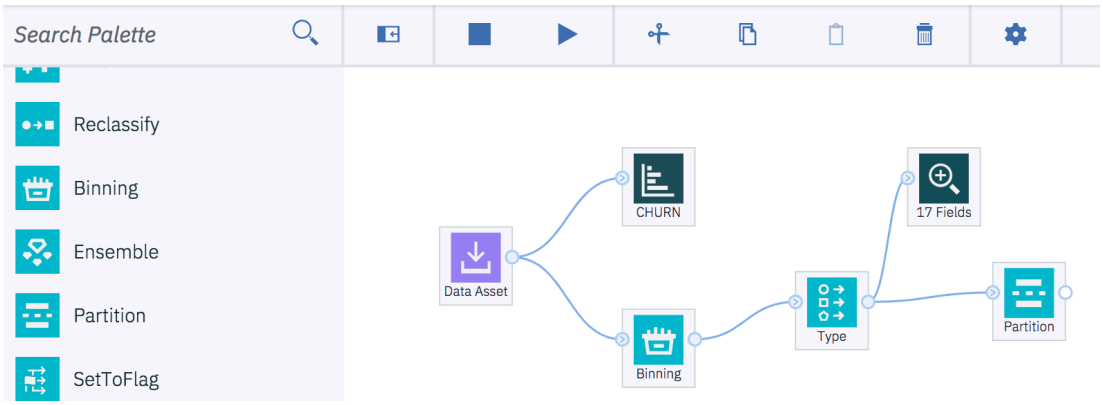


22. Double-click in the Data Audit (outputs) In the Data Audit results output, thumbnail graphs, storage icons, and summary statistics for all fields can be found. It also provides information about outliers, extremes and missing values.

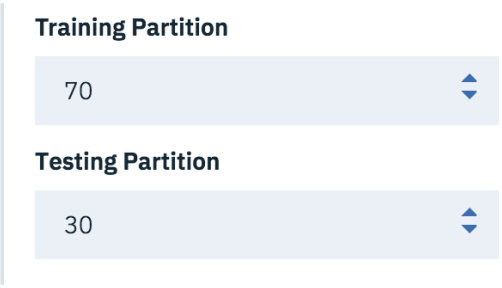
Data Audit of [17 fields]

	Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev
1	ID		Continuous	1	3825	1901.152	1094.709
2	Sex		Flag	--	--	--	--
3	Status		Nominal	--	--	--	--
4	Children		Continuous	0	2	1.148	0.843

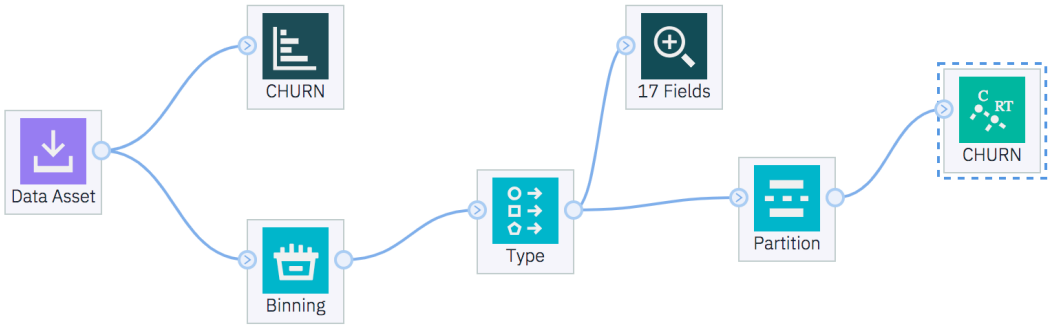
23. Now that we have explored our data, we can build a model to uncover the key drivers resulting in cancelled contract. Go back to the workflow and from the Field Operations, add a Partition node to the workflow and connect it to the Type node.



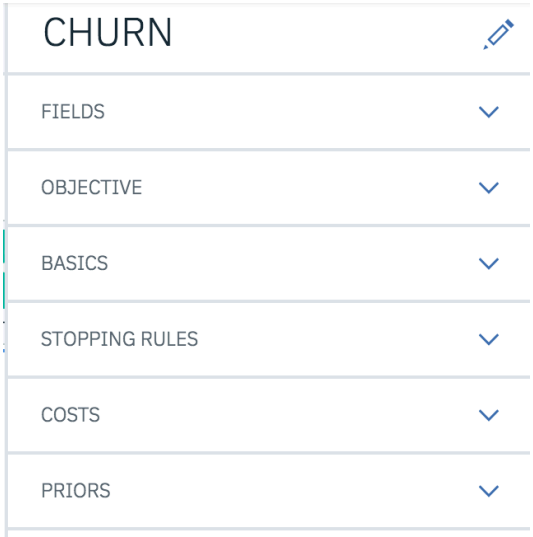
24. Double-click on the Partition node and change Training Partition to 70 and Testing Partition to 30. Click Save.



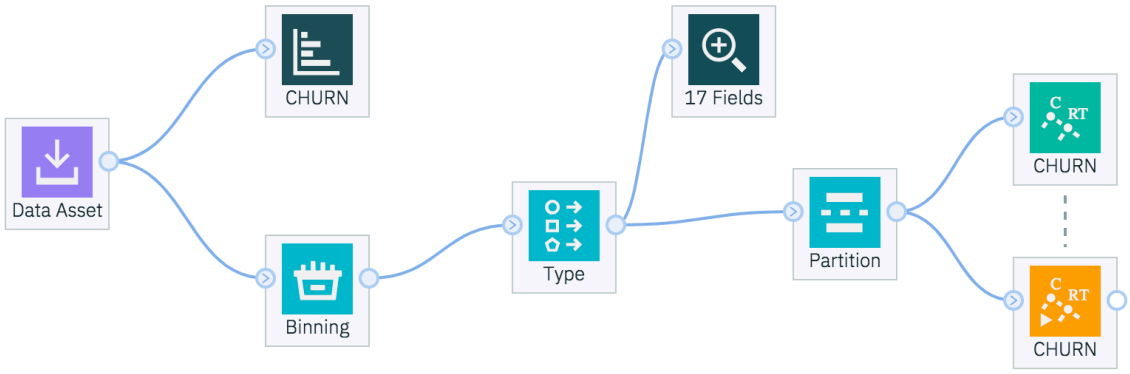
25. From the Modeling, add the C&R Tree node to the workflow and connect it to the Partition node. Note that the C&R Tree node name changes to CHURN when it is connected to the Partition node. This is because we defined CHURN as the Target Role in Step 20.



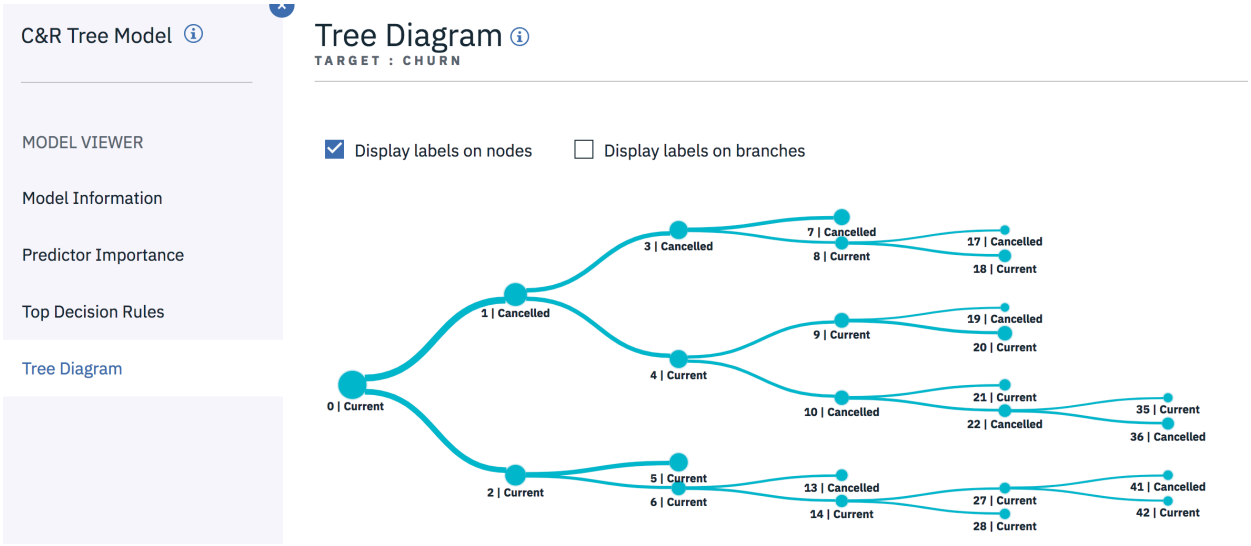
26. Open the C&R Tree node to view the settings before running the model. This example uses the defaults so closed properties and click Run.



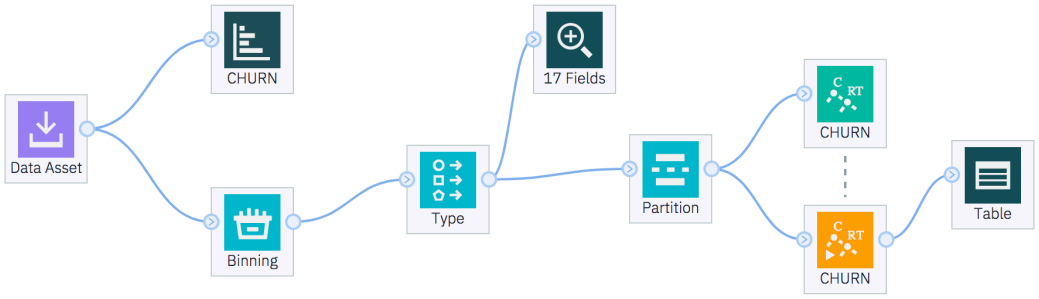
27. The C&R Tree model is generated, added to the workflow and named CHURN.



28. Right-click on the generated model and select View Model to see the outputs. The Model outputs contains Model Information, Predictor Importance, Top Decision Rules and the Tree Diagram. Click the Viewer each one.



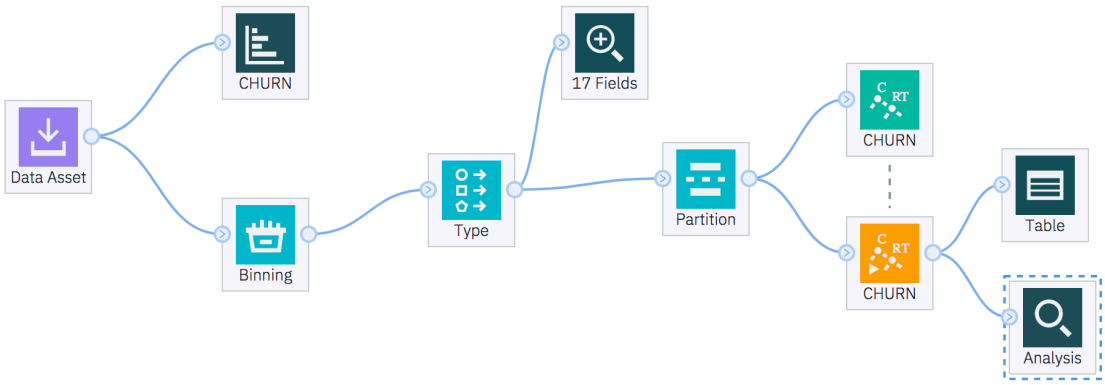
29. To view model output, from the Outputs add a Table node to the generated CHURN model and run it.



30. Look at the last five columns in the table. The fifth from last column is the actual outcome, whether the customer cancelled or is current; the second to last column is the prediction from the model; and the last column is the confidence in the prediction. Note that the model predicted that the customer would cancelled with 78.1% confidence.

LocalBilltype	LongDistanceBillty	CHURN	Age_BIN	Partition	$R-CHURN	$RC-CHURN
Budget	Intnl_discount	Cancelled	3.000	1_Training	Cancelled	0.781
FreeLocal	Standard	Current	8.000	1_Training	Current	0.896
FreeLocal	Standard	Current	9.000	1_Training	Cancelled	0.781
Budget	Standard	Current	10.000	2_Testing	Current	0.810

31. To see the overall accuracy of the model, from the Output, add an Analysis node, join it to the generated model and run it.



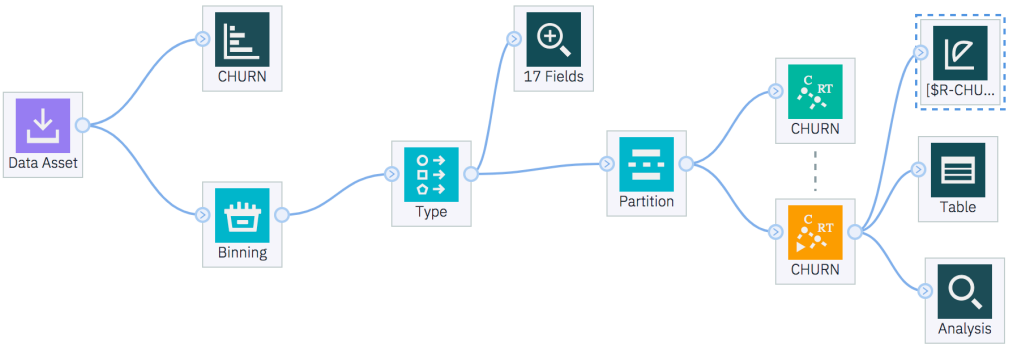
32. The resulting output indicates an overall accuracy of 77.93%. That is, the model predicted with 77.93% accuracy which customers Cancelled or Current the contract.

Results for output field CHURN

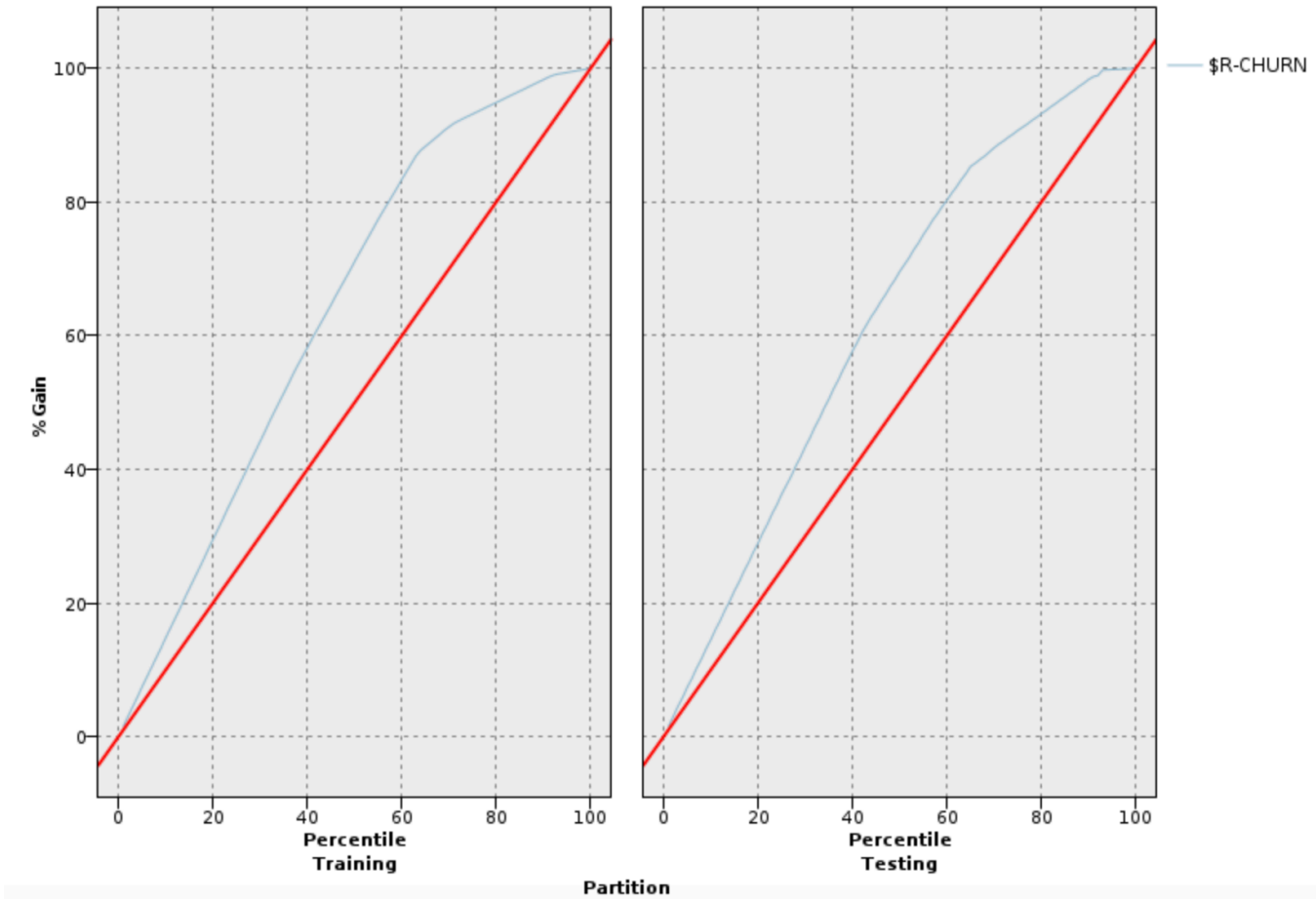
Comparing $R-CHURN with CHURN

'Partition'	1_Training		2_Testing	
Correct	1,173	82.49%	505	77.93%
Wrong	249	17.51%	143	22.07%
Total	1,422		648	

33. To further evaluate the model, select an Evaluation node from the Graphs, connect it to the model node and select Run.



34. In the resulting gains chart, the red line reflects what you could expect without Predictive Analytics. The blue line; however, reflects the lift in response you could achieve utilizing Predictive Analytics. Therefore, if you were to randomly select 50% of your client base, you could expect to have captured 50% of those likely cancelled. By using Predictive Analytics, you can more effectively target those 50% of clients and capture almost 80% of those likely to cancel.



35. Finally, to see the relationships between fields, select a Matrix node from the Output and connect it to the model node. Using the drop-down menu, choose “$R-CHURN” for Rows and “CHURN” for Columns. Select Run.

Matrix of $R-CHURN by CHURN

$R-CHURN	Cancelled	Current
Cancelled	580	168
Current	224	1098

Cells contain: cross-tabulation of fields (including missing values)
Chi-square = 738.419, df = 1, probability = 0

Summary

- Prepare data for modeling
- Define which fields to use
- Choose the modeling technique
- Automatically generate a model to identify who has cancelled
- Review results

Over the course of the last 20 minutes, we were able to successfully train a model by exploring SPSS Modeler’s ability to read in data, create new fields via data preparation techniques, choose and run a predictive modeling algorithm, and evaluate the results to accurately identify customer response.

Use Case

Goal: Create segments of customers

Approach:
- Merge disparate data sources, including customer data
- Define which fields to use
- Automatically generate a model to group customers
- Apply business terms to new customer groups
- Export newly created groups to a database

Why?
- Better customer understanding (demographics, socio-economic, etc.)
- Tailored messages for each group/segment
- Personal and more relevant for consumers

Customer Reference

A US cable television network turns on the insights with an analytics solution that predicts success of new shows six weeks in advance.

Business challenge: This cable television network faces the challenge of managing huge volumes of information. Previously the network's research team spent a significant amount of time processing data on spreadsheets rather than analyzing it, and based decisions on a combination of experience and instinct. The company needed a large-scale analytics solution to organize this wealth of data, make sense of it, and provide answers and actionable insights.

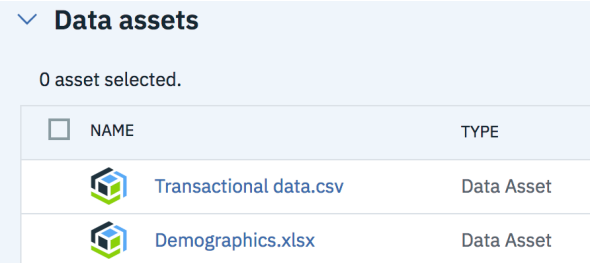
The transformation: The solution combines television ratings data with information gathered minute by minute and viewer by viewer from a variety of channels and other sources to determine who's watching and why. Then it centralizes the data and makes it available for in-depth, predictive analytics. With insights into audience preferences gained from sophisticated statistical models, including intelligent segmentation, the network can optimize advertising revenue and viewership like never before.

IBM's implemented solution:
Accelerates analytics by extracting insights from billions of rows of audience data in seconds, instead of days.
Triples views of video-on-demand service through data-driven marketing.
Predicts success of a new show six weeks in advance of its release and adjusts marketing accordingly.

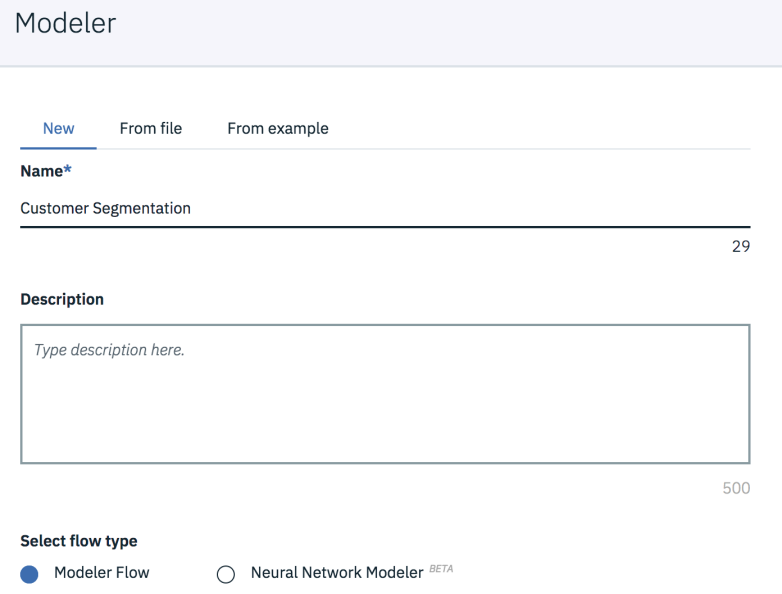
"A single day of analysis work enabled the network to design a campaign that increased the consumption of its video on demand service. Previously, that analysis would have taken weeks."

Finding Patterns and Groups

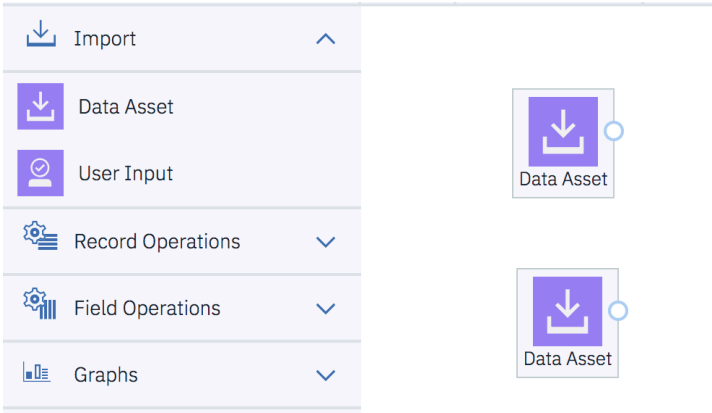
1. Create a new project and add the files Demographics.xlsx and Transactional data.csv to the Data Assets.



2. Create a new Modeler flows on the project. Type a name and description and click Create.

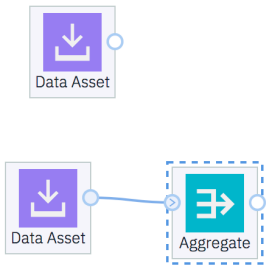


3. From the Import, add two Data Assets and connect one to the Demographics data and the other to the Transactional data.

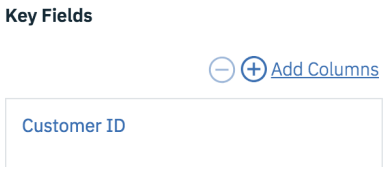


We will be joining two data files together; one file contains customer transaction data, and exists in a database, while the other file is an Excel file and contains customer demographics. Though the files are in different formats, the Merge node in IBM SPSS Modeler can very easily join the files without first requiring the analyst to translate one file and/or the other into the same format.

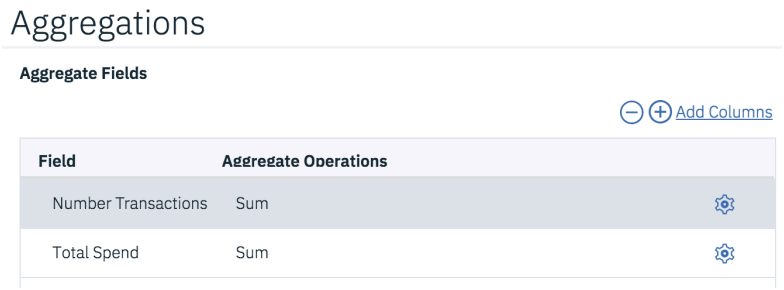
4. First, we need to aggregate the transactional data to Customer ID. From the Record Operations, add a Aggregate node to the workflow and connect it to the Transactional data Data Asset node.



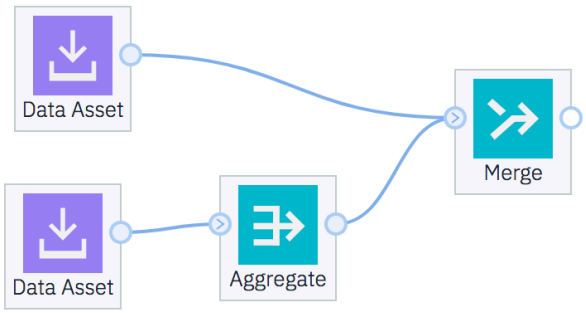
5. Double-click to the Aggregate node, in Key Fields, click Add Columns and select Customer ID.



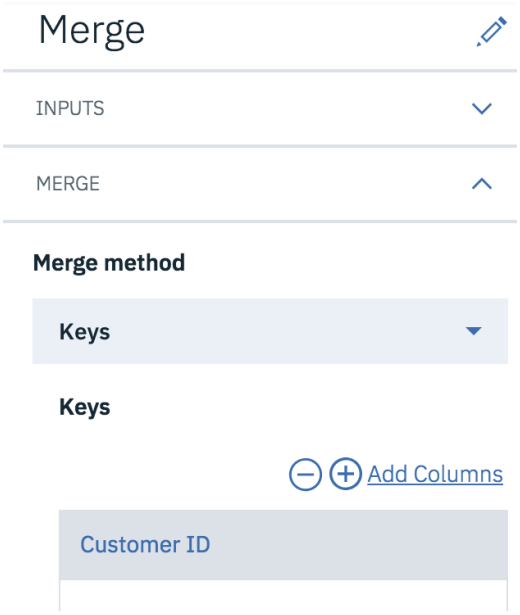
6. Click Aggregations, click Add Columns and select Total Spend and Number Transactions and click OK. Click on the edit icon and uncheck the Mean operate, click OK and Save



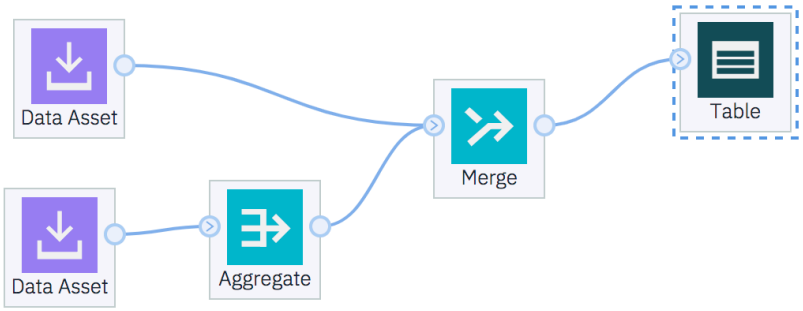
7. From the Record Operations, add a Merge node to the workflow and connect the Data Assets node (Demographics) and Aggregate node to the Merge node.



8. Open the Merge node and change the Merge method to Keys. Click on Add Columns and add Customer ID. Click Save.



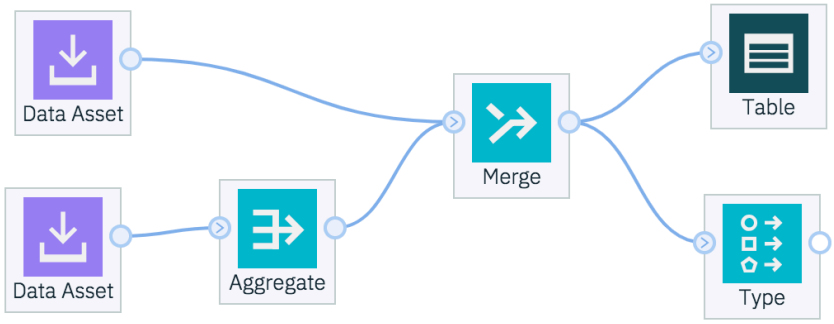
9. From the Output, add a Table node to the workflow and connect it to the Merge node.



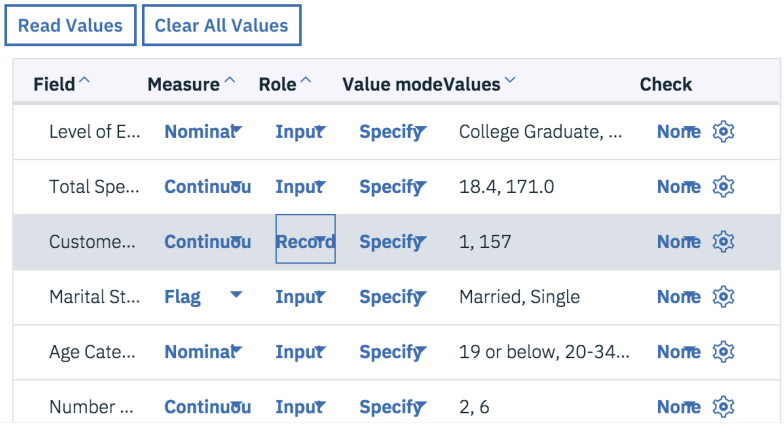
10. Once connected, right click on the Table node and select Run to review the data after having been merged.

Customer ID	Marital Status	Income in Thousa	Number of Childre	Level of Education	Age Category_	Total Spend_Sum	Number Transacti
1.000	Married	140.000	3.000	Some High School	35 - 49	43.000	26.000
2.000	Married	225.000	4.000	High School Gradu	35 - 49	56.800	34.000
3.000	Married	225.000	3.000	High School Gradu	35 - 49	56.800	34.000
4.000	Married	210.000	4.000	High School Gradu	50 - 64	84.000	36.000
5.000	Married	150.000	3.000	Some High School	35 - 49	48.000	32.000

11. From the Field Operations, select a Type node and attach it to the Merge node.

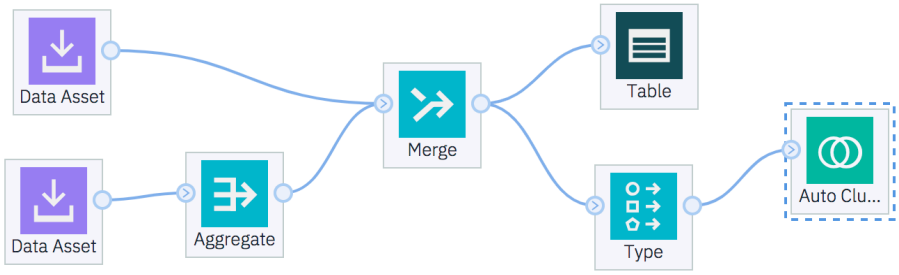


12. Double-click on the Type node to edit it. Click on the Read Values button and change the role of "Customer ID" to "Record ID". Once done, click Save.



At this point we are ready to cluster the cases into segments. For this we will use the 'Auto Cluster' node. The Auto Cluster node allows you to try all of the clustering algorithms and, at your discretion, any or all of their parameters. It builds all of the models you specify and shows you the best models (3 is the default) to use with your data. So, in one step, you will have the best model(s) without having to know or guess which might work for you. This is also a nice way to see how other modeling algorithms will perform. The same holds for the other auto modeling nodes that address classification and numeric modeling techniques.

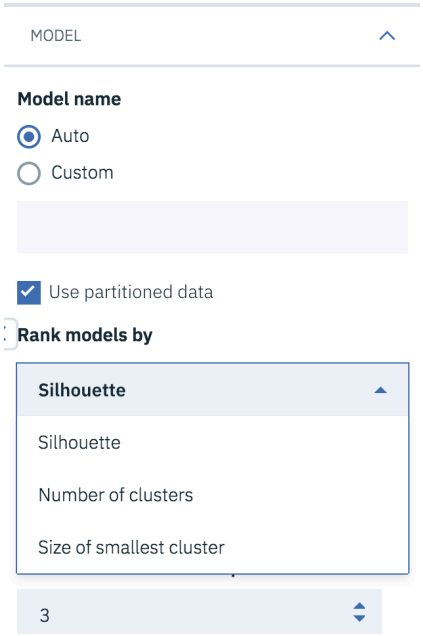
13. From the Modeling select the Auto Cluster modeling node and attach it to the Type node.



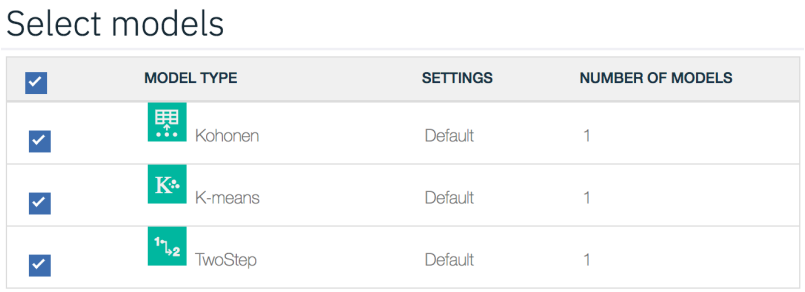
14. Double-click on the Auto Cluster node to edit it.

On the ‘Model’ options, there are various ways of ranking the quality of the models that are built. Keep the default method, ‘Silhouette’.

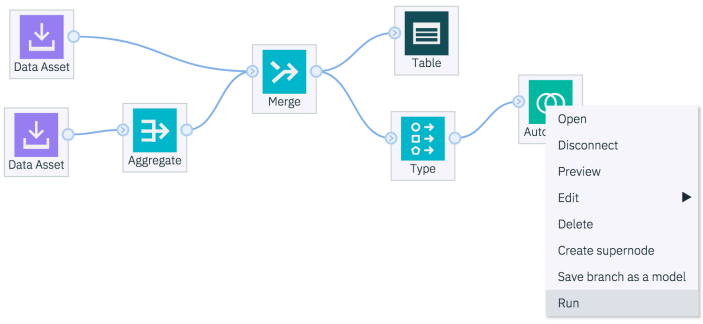
Also keep the default number of models to keep, 3. This means that the 3 best models, based on their silhouette measure, will be retained for our use.



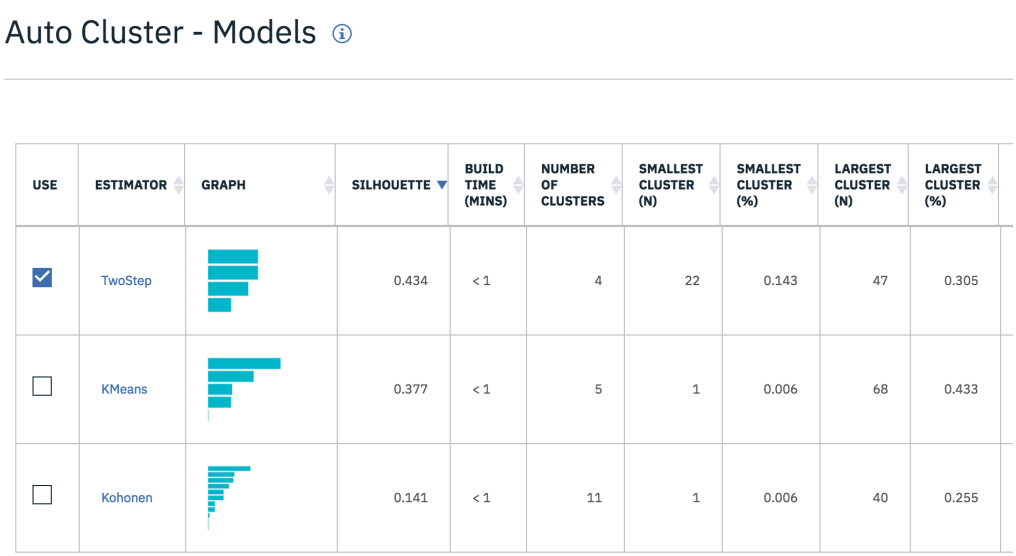
15. Scroll down on the ‘Expert’ options and click on Select models. You notice that there are 3 clustering algorithms in the list.



16. Go back to the workflow and Click Run in the Auto Cluster node.



17. Right-click on the model results and select View Model to view the results of the auto-clustering analysis.



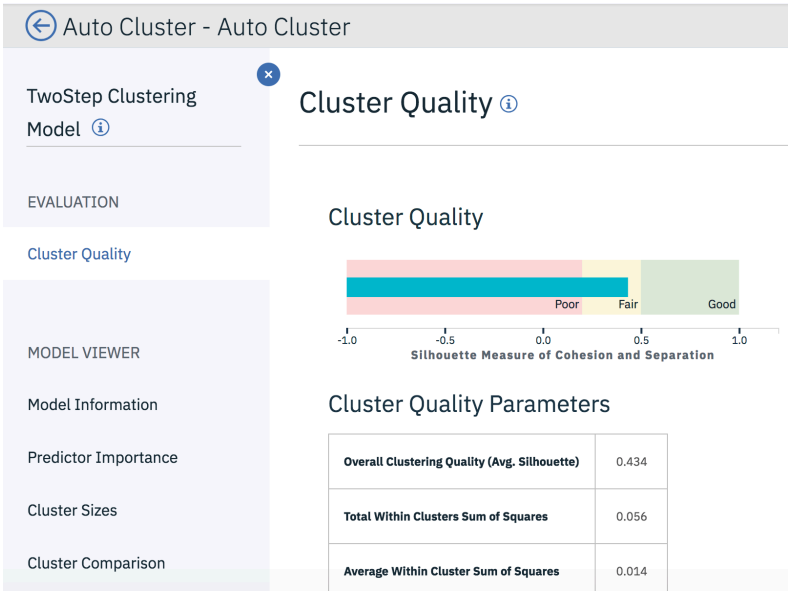
USE	ESTIMATOR	GRAPH	SILHOUETTE	BUILD TIME (MINS)	NUMBER OF CLUSTERS	SMALLEST CLUSTER (N)	SMALLEST CLUSTER (%)	LARGEST CLUSTER (N)	LARGEST CLUSTER (%)
☑	TwoStep		0.434	< 1	4	22	0.143	47	0.305
☐	KMeans		0.377	< 1	5	1	0.006	68	0.433
☐	Kohonen		0.141	< 1	11	1	0.006	40	0.255

This is where you see the 3 ‘best’ models for segmenting the data. The list is in descending order by Silhouette measure like we specified. There are other important statistics about each model in the table.

The results show that the TwoStep algorithm has the best Silhouette measure followed by the K-means and Kohonen models, which were discarded.

The check boxes to the left indicate that the TwoStep model will be used since it was ranked highest according to our ranking criterion. With cluster models, only one can be selected at a time; but you could choose to use any of the others by clicking the check box. For our exercise we will stay with the TwoStep model.

18. Click the first model name labeled ‘TwoStep’ to see the results of the TwoStep cluster analysis.



Looking at the Cluster Quality measure in the left panel, we see that the Silhouette measure (which is a measure of the clusters' internal cohesion AND how well they exclude dissimilar cases) is fair, with a value of just under 0.5. Such results are common but may also suggest that fewer and/or other variables might be needed to increase the Silhouette value.

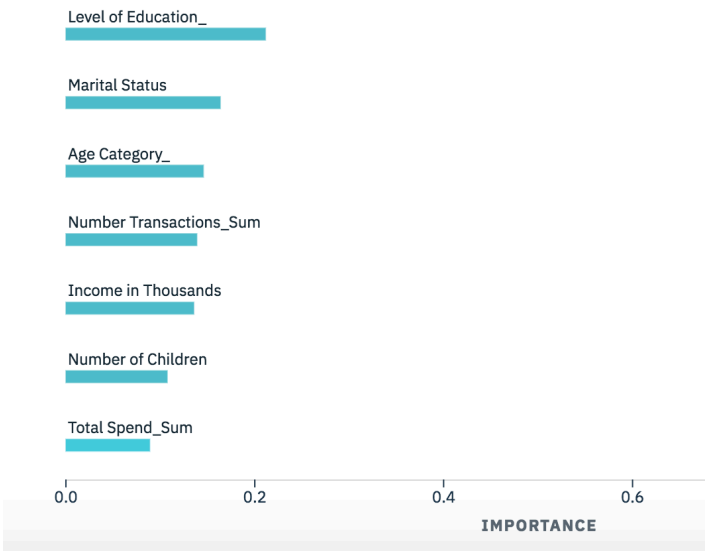
19. From the menu in the right viewer select Model Information.

We specified a range of 2 to 4 clusters; and the Two Step clustering engine resolved into 4 clusters.

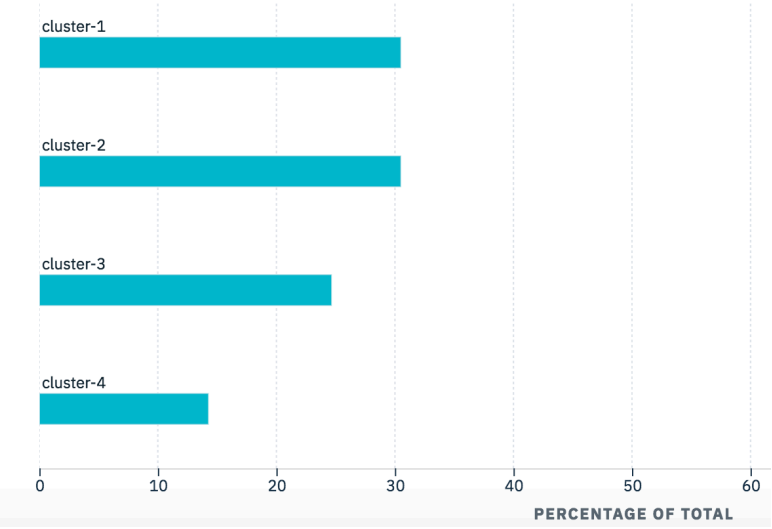
Algorithm		TwoStep
Model Class		Distribution Based
Number of Features		7
Distance Measure		Log Likelihood
Number of Clusters		4
Number of instances in each cluster	**cluster-1**	47 (30.52%)
	cluster-2	47 (30.52%)

20. From the menu in the right viewer select Predictor Importance.

Now the viewer displays a graph with the variables ranked in order of importance for cluster definition. We can see that Level of Education is the most important variable, followed by Marital Status.

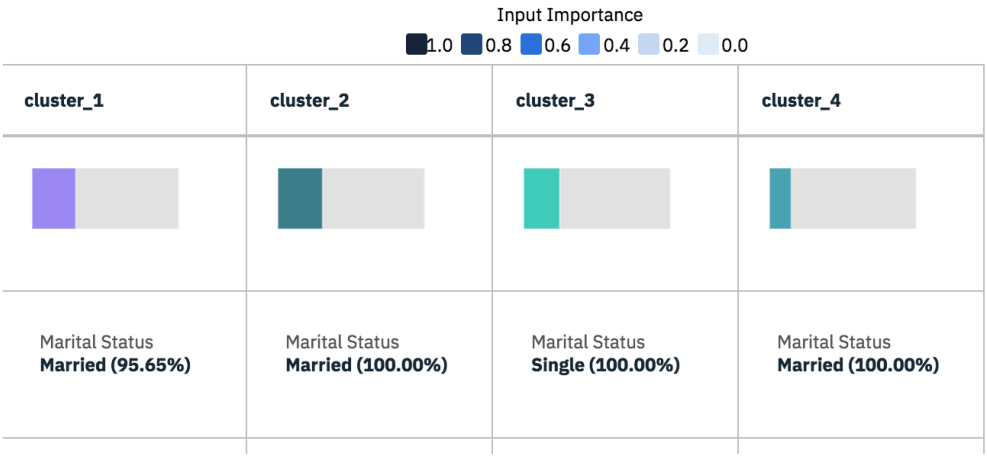


21. From the menu in the right viewer select Cluster Sizes.



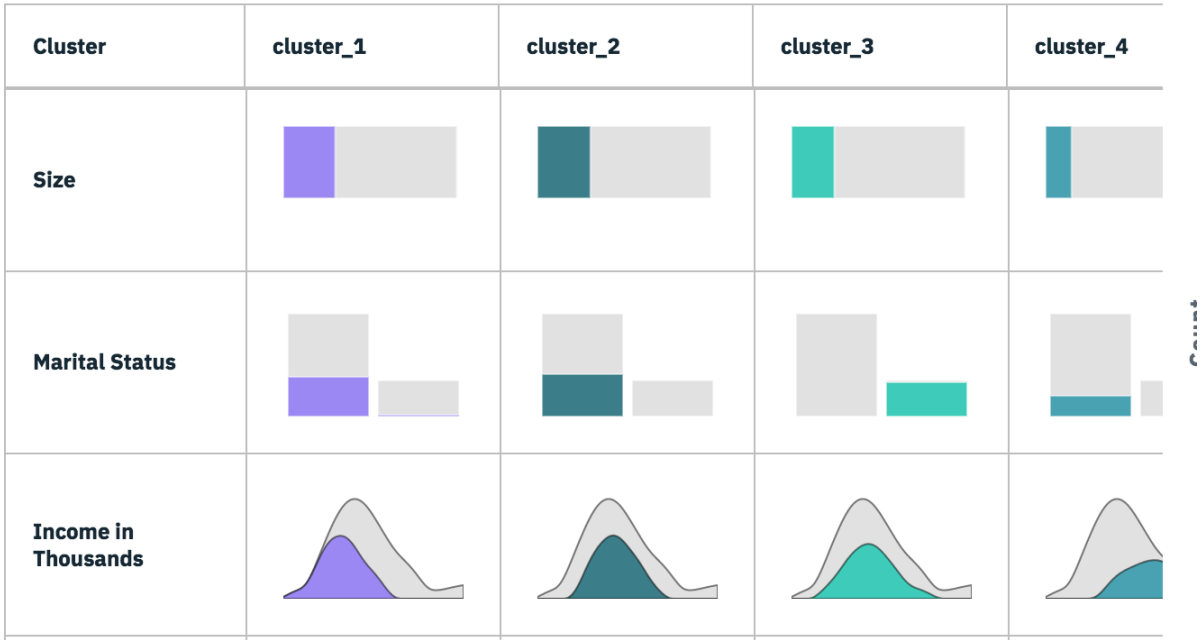
22. From the menu in the right viewer select Cluster Comparison.

23. From the menu in the right viewer select Clusters.

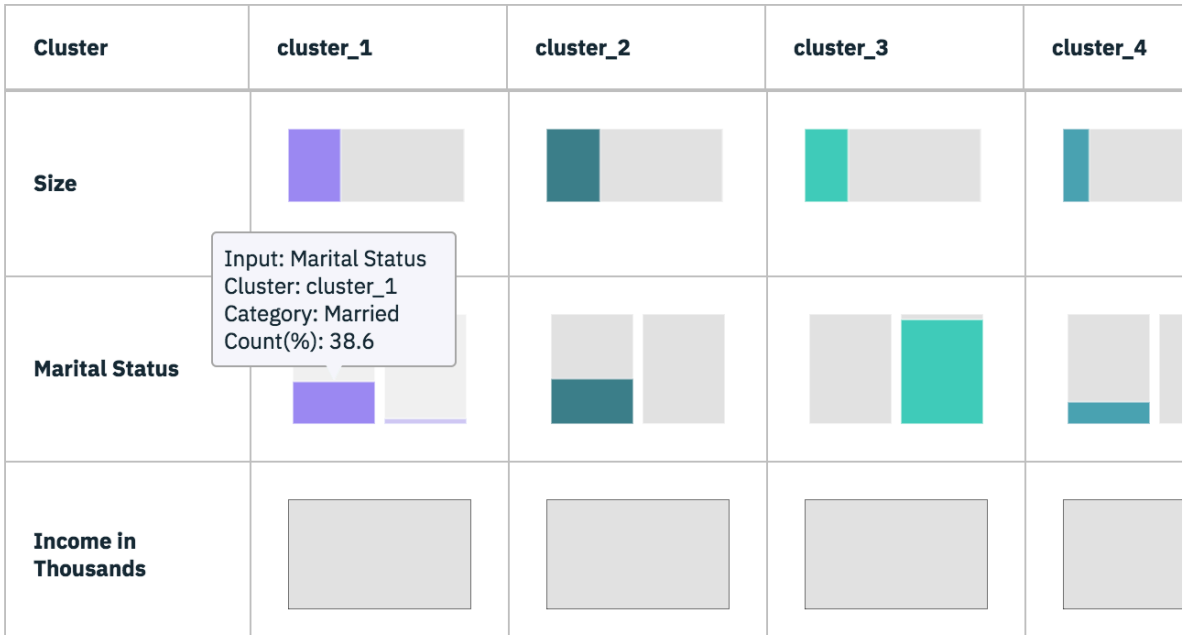


24. From the menu in the right viewer select Cell Distributions (Absolute).

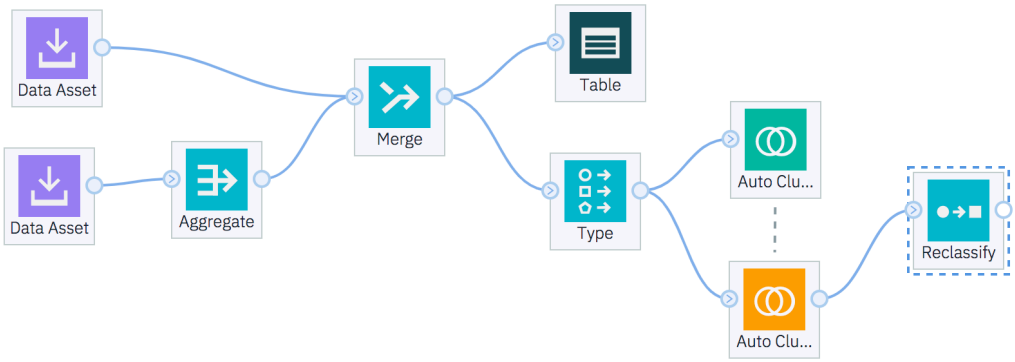
The panel of the Viewer displays the clusters in order of their size, left to right. The darkness of the shading of each variable indicates its importance in cluster definition; the lighter the shading, the less important is the variable in defining the clusters.



25. From the menu in the right viewer select Cell Distributions (Relative).



26. Go back to the workflow. From the Field Operations, drag a Reclassify node onto the workflow and connect it to the model node.

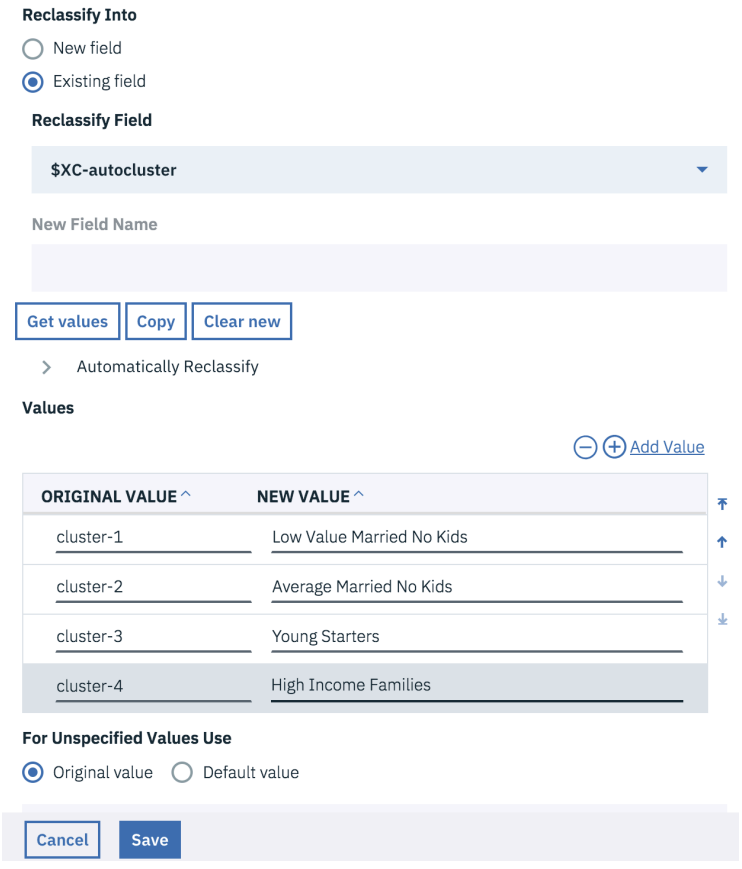


27. Double-click on the Reclassify node to edit it.

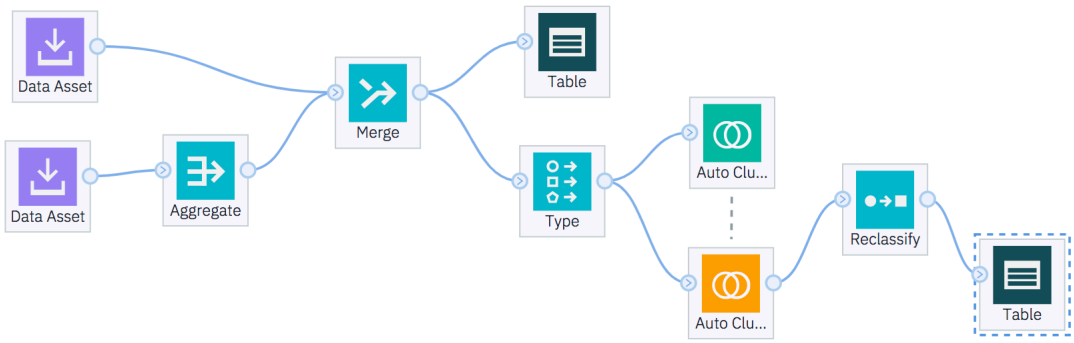
In the Reclassify Field, select the variable **$XC-autocluster**

Click “Existing Field” just above so that it doesn’t make a new field.

Click the “**Get values**” button to populate the ‘Original value’ column for you. Enter the new values on the right, which better describe the clusters. An example is shown below, but you can enter your own labels as desired. Once completed, click Save.



28. From the Output, drag a Table node onto the workflow and connect it to the Reclassify node. Once connected, right-click on the Table node and Select Run.



29. The resulting table now includes a new column with the cluster assignments.

Customer ID	Marital Status	Income in Thousai	Number of Childre	Level of Education	Age Category_	Total Spend_Sum	Number Transacti	$XC-autocluster
1.000	Married	140.000	3.000	Some High School	35 - 49	43.000	26.000	Low Value Married
2.000	Married	225.000	4.000	High School Gradua	35 - 49	56.800	34.000	Average Married Nc
3.000	Married	225.000	3.000	High School Gradua	35 - 49	56.800	34.000	Average Married Nc
4.000	Married	210.000	4.000	High School Gradua	50 - 64	84.000	36.000	Average Married Nc
5.000	Married	150.000	3.000	Some High School	35 - 49	48.000	32.000	Low Value Married
6.000	Married	200.000	4.000	High School Gradua	50 - 64	68.000	36.000	Average Married Nc
7.000	Married	310.000	4.000	College Graduate	50 - 64	124.000	46.000	High Income Famil
8.000	Married	170.000	3.000	High School Gradua	35 - 49	54.000	32.000	Average Married Nc
9.000	Married	193.000	3.000	High School Gradua	50 - 64	66.800	32.000	Average Married Nc

Summary

- Merge disparate data sources, including customer data from a database or CRM
- Define which fields to use
- Automatically generate a model to group customers
- Apply business terms to grouped customers
- Send new groups to file

For this exercise, we merged two data sources together in order to identify groupings within our data, we used an automated clustering technique, specifying desired parameters. The resulting clusters were reclassified into business terms and could be exported back to the file or database.

IBM Corporation
Software Group
Route 100
Somers, NY 10589

Produced in the United States of America
February 2017