Documentation for Each Tool

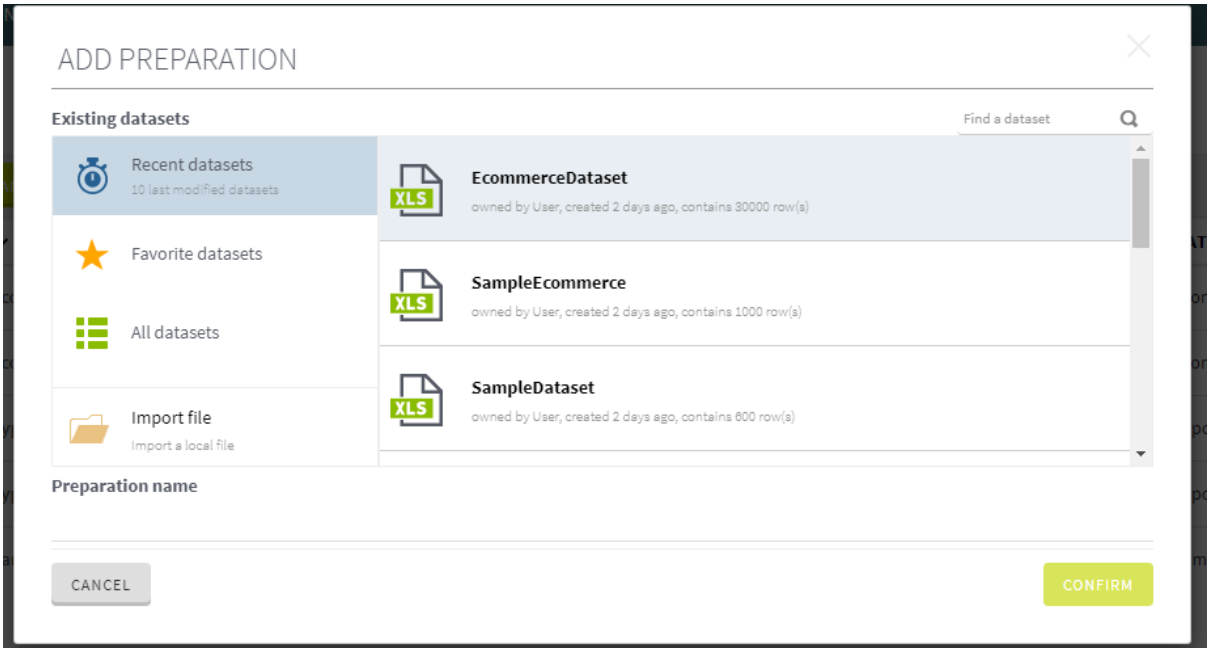
Nur Natisya Binti Abdul Yazid

S2195163

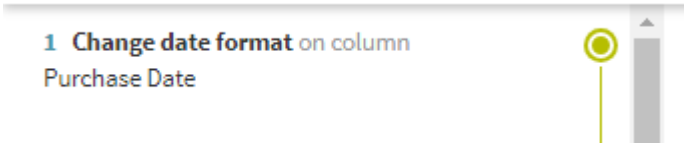
Github Link: https://github.com/s2195163/AA1.git

Talend Data Preparation

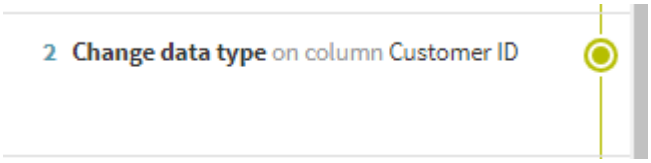
1. Upload Data (Add Preparation and choose the dataset)



2. Change Date Format for better view and understanding

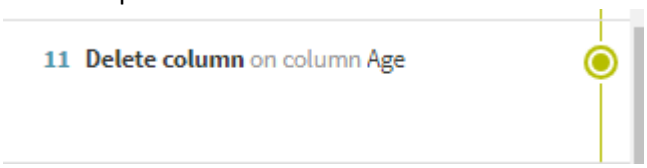


3. Change Data Type for CustomerID as it identifies it postal code

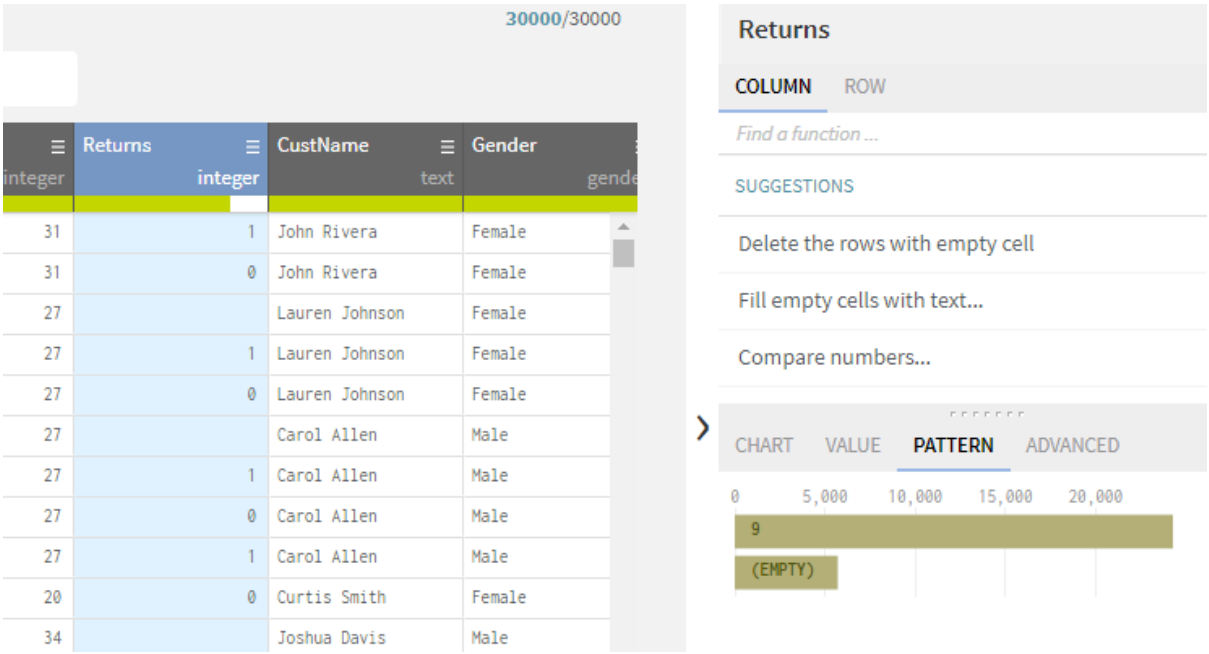


4. Rename column

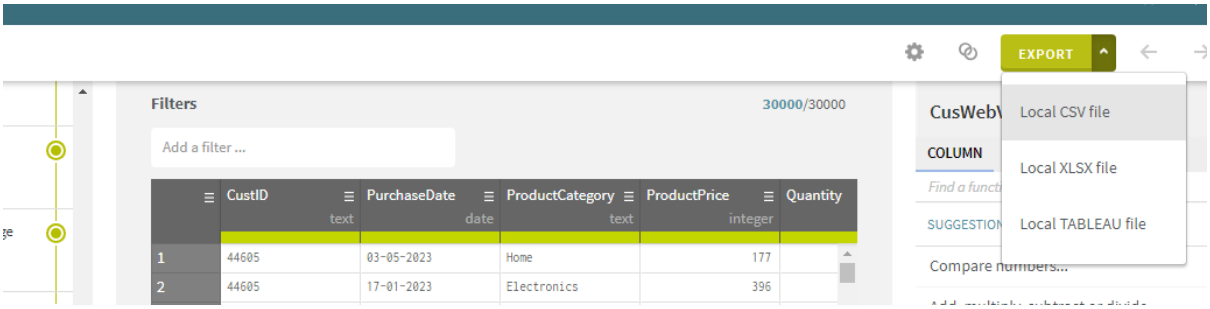
5. Delete duplicated column



6. Identified missing value

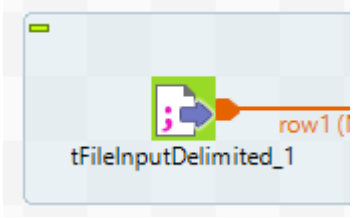


7. Export data (30000 observations)
   It doesn't import the whole data (47237 observations) as it can only import 30000 oservations.

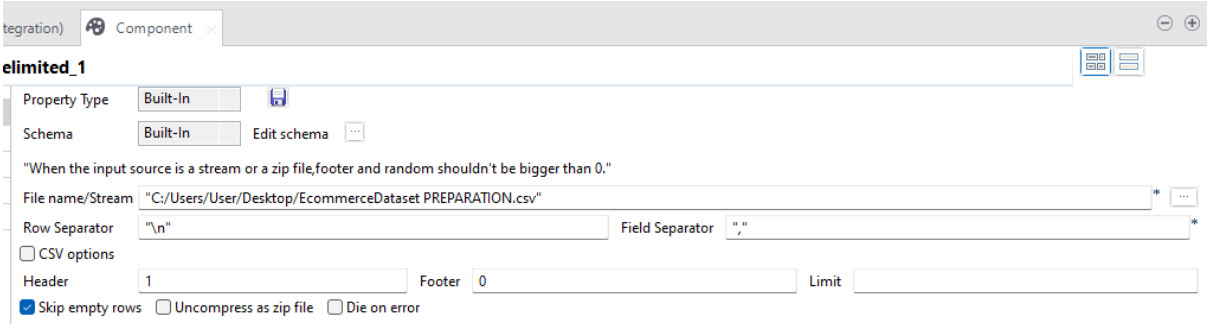


Talend Data Integration

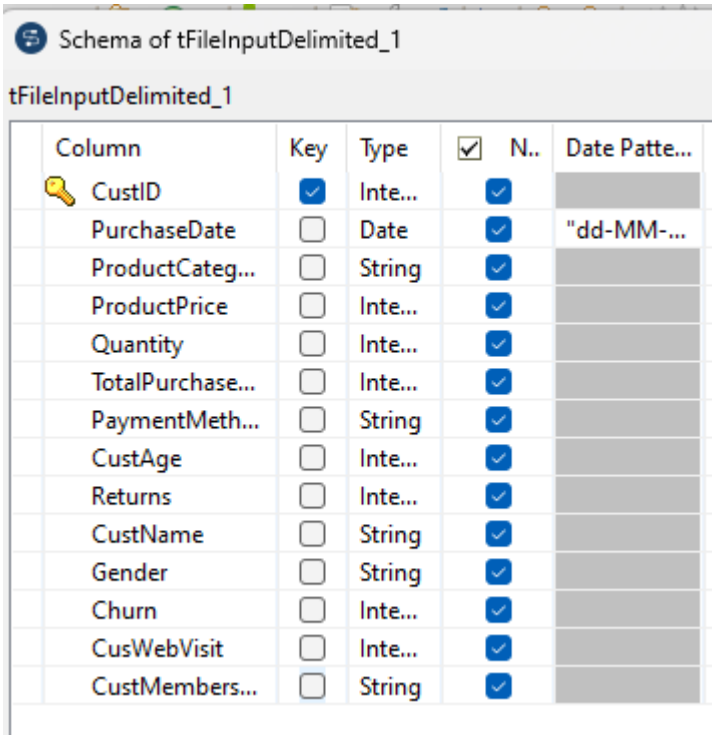
1. Drag and drop tFileInputDelimited



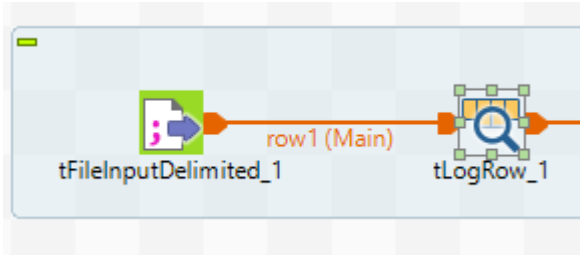
Browse the path of the file to allow import data.



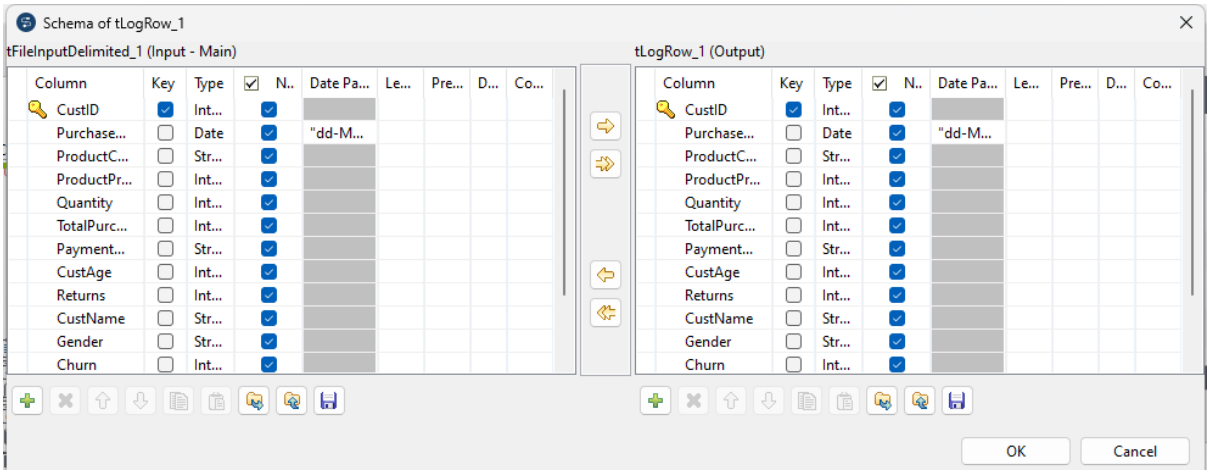
Edit Schema (Add columns, identify the key, selecting the right data type)



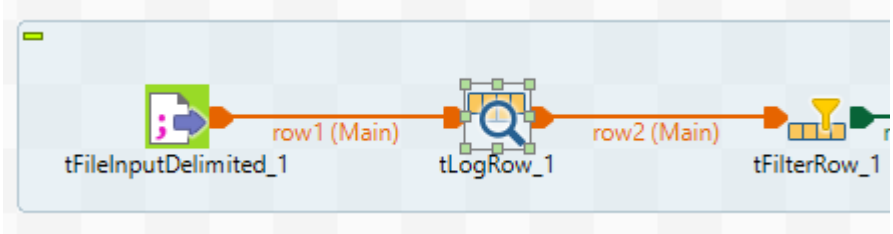
2. Drop tLogRow to display the dataset



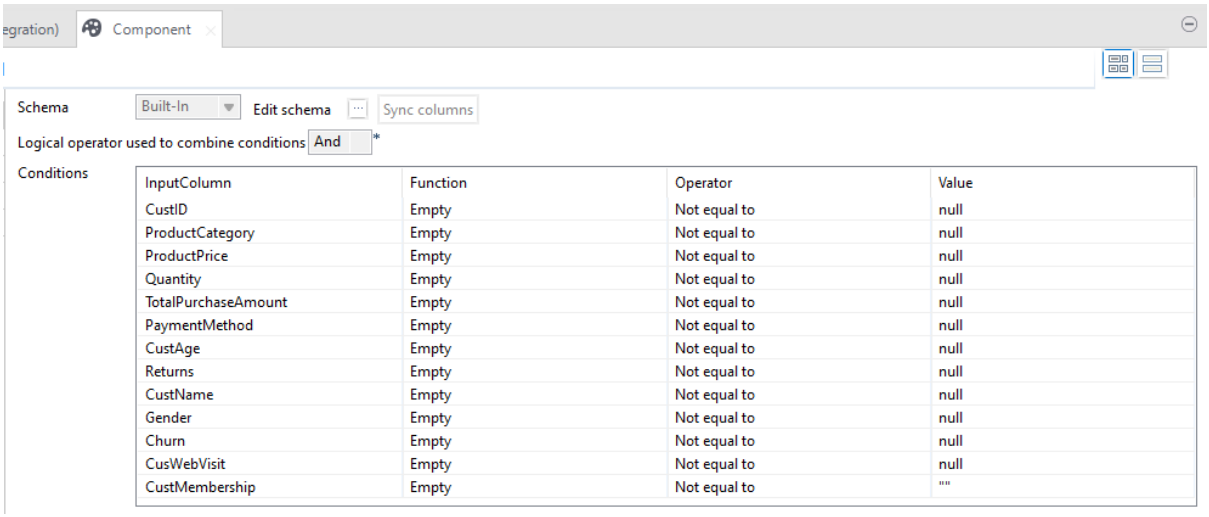
Check schema if it connects correctly



3. Drop tFilterRow to filter missing values

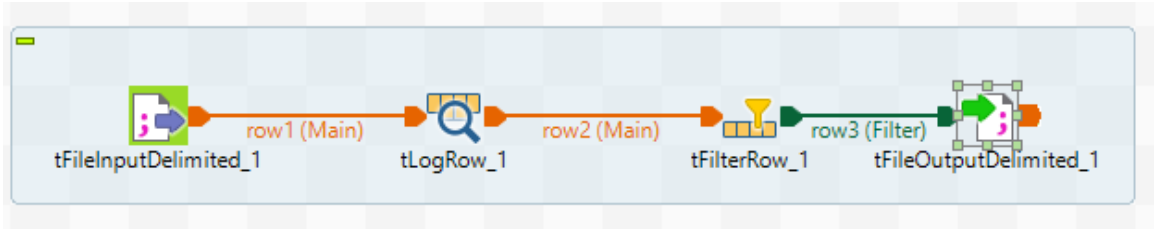


Edit the conditions under the component

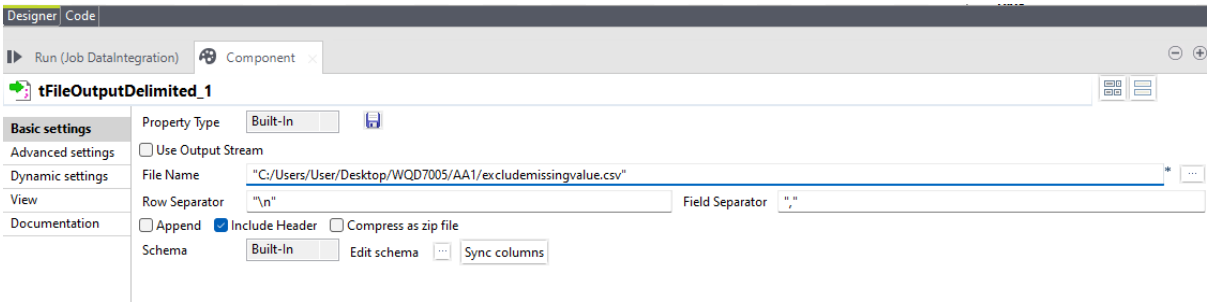


InputColumn	Function	Operator	Value
CustID	Empty	Not equal to	null
ProductCategory	Empty	Not equal to	null
ProductPrice	Empty	Not equal to	null
Quantity	Empty	Not equal to	null
TotalPurchaseAmount	Empty	Not equal to	null
PaymentMethod	Empty	Not equal to	null
CustAge	Empty	Not equal to	null
Returns	Empty	Not equal to	null
CustName	Empty	Not equal to	null
Gender	Empty	Not equal to	null
Churn	Empty	Not equal to	null
CusWebVisit	Empty	Not equal to	null
CustMembership	Empty	Not equal to	""

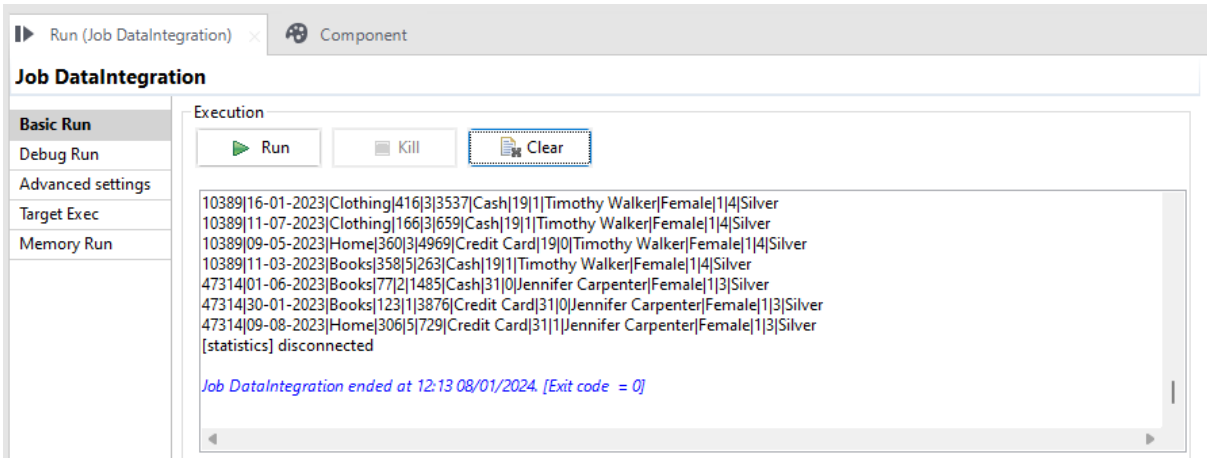
4. Drop FileOutputDelimited to generate the new dataset (Non-missing value dataset)



Browse the path,



5. Run it



The saved file: csv format

excludemissingvalue

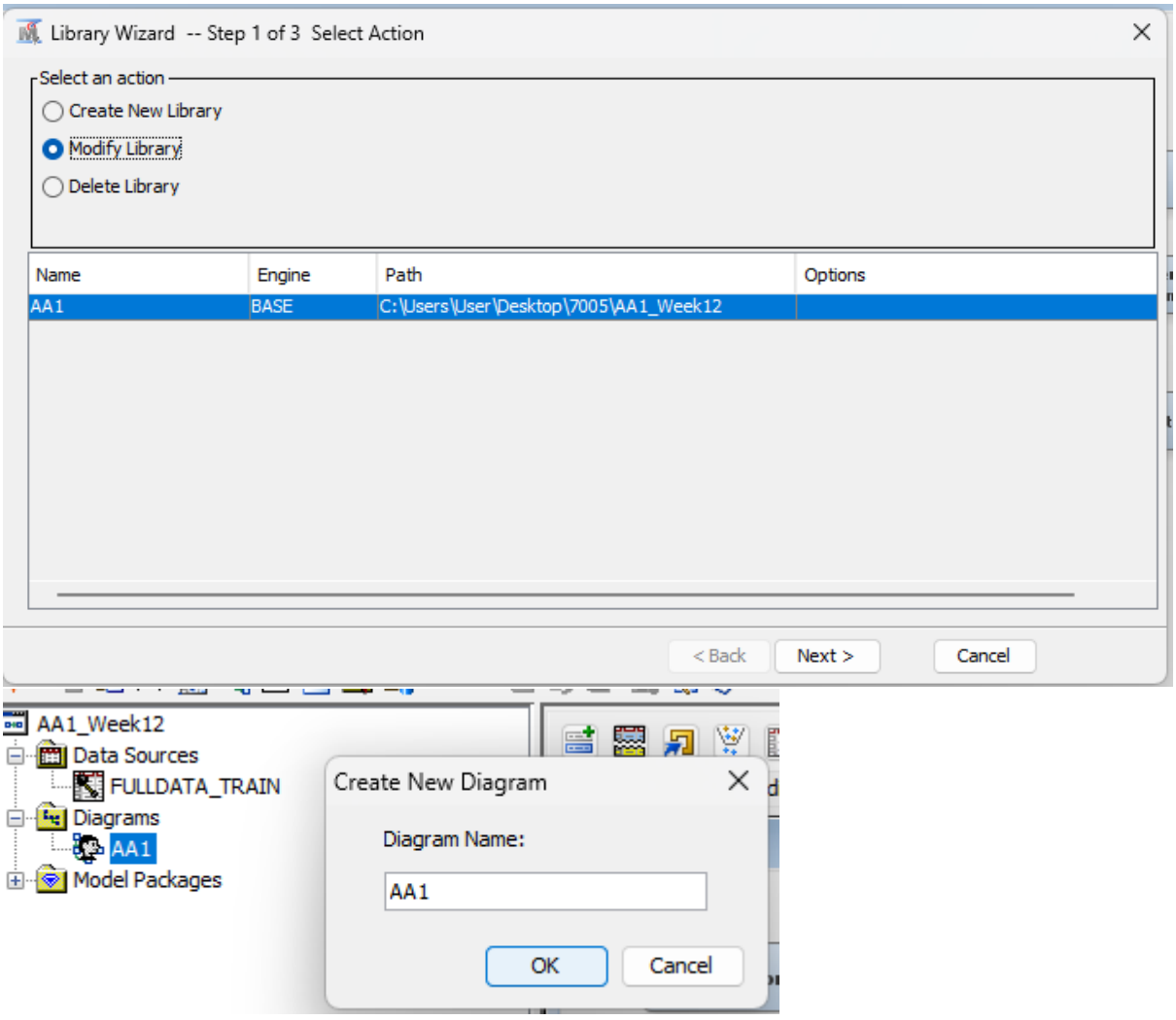
File Edit View

```
CustID,PurchaseDate,ProductCategory,ProductPrice,Quantity,TotalPurchaseAmount,PaymentMethod,CustAge,Returns,CustName,Gende
r,Churn,CusWebVisit,CustMembership
44605,03-05-2023,Home,177,1,2427,PayPal,31,1,John Rivera,Female,0,2,Bronze
44605,17-01-2023,Electronics,396,3,937,Cash,31,0,John Rivera,Female,0,2,Bronze
13738,05-02-2023,Books,370,5,1486,Cash,27,1,Lauren Johnson,Female,0,3,Silver
13738,09-02-2023,Electronics,40,4,4327,Cash,27,0,Lauren Johnson,Female,0,3,Silver
33969,05-01-2023,Home,304,1,3883,PayPal,27,1,Carol Allen,Male,0,4,Silver
33969,18-07-2023,Books,54,2,4187,PayPal,27,0,Carol Allen,Male,0,4,Silver
33969,05-07-2023,Clothing,473,3,2881,Credit Card,27,1,Carol Allen,Male,0,4,Silver
42650,29-04-2023,Home,43,1,2312,Cash,20,0,Curtis Smith,Female,0,1,Bronze
16921,24-01-2023,Books,51,3,1881,PayPal,54,1,Cheyenne James,Male,0,1,Bronze
21035,09-09-2023,Home,237,2,1088,Cash,50,1,Peter Watson,Female,1,1,Bronze
1254,15-08-2023,Home,476,1,1687,Cash,70,1,Mr. David Morgan,Male,0,2,Bronze
1254,21-03-2023,Home,413,5,3742,Cash,70,0,Mr. David Morgan,Male,0,2,Bronze
13389,10-02-2023,Clothing,312,2,4477,PayPal,51,1,Monica Ramos,Male,0,4,Silver
13389,02-06-2023,Books,40,2,2341,Credit Card,51,1,Monica Ramos,Male,0,4,Silver
13389,17-04-2023,Home,495,4,1176,Credit Card,51,0,Monica Ramos,Male,0,4,Silver
24473,22-05-2023,Home,436,3,1755,PayPal,38,1,Nicole Lewis,Female,0,1,Bronze
16825,18-03-2023,Books,227,2,3400,Cash,37,0,Melissa Cabrera,Male,1,1,Bronze
18467,28-02-2023,Clothing,50,1,551,PayPal,23,1,Terri Carter,Male,0,2,Bronze
[illegible]09-05-2023,Electronics,92,4,3660,PayPal,23,0,Terri Carter,Male,0,2,Bronze
47314,30-01-2023,Books,123,1,3876,Credit Card,31,0,Jennifer Carpenter,Female,1,3,Silver
47314,09-08-2023,Home,306,5,729,Credit Card,31,1,Jennifer Carpenter,Female,1,3,Silver
```

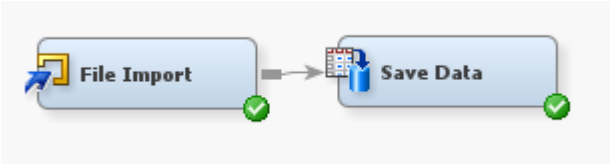
Ln 24296, Col 3 100% Unix (LF

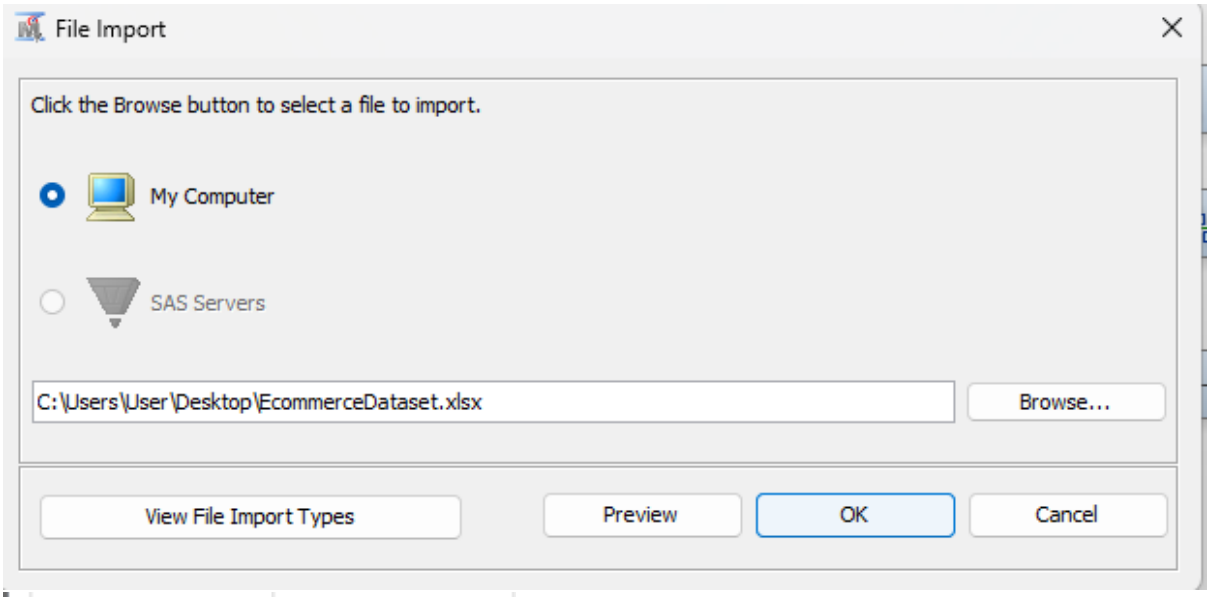
SAS Enterprise Miner

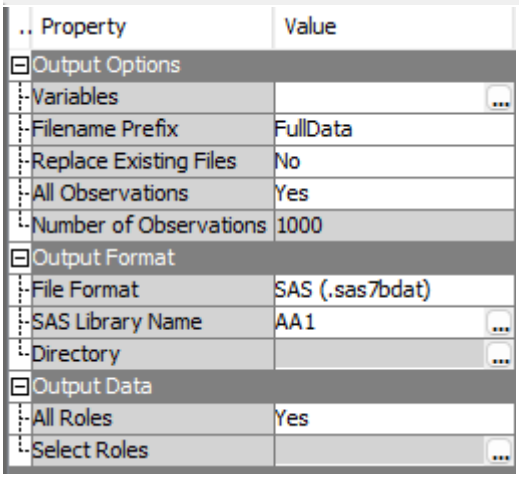
1. Create library, then create diagram



2. Import data and save it as SAS data

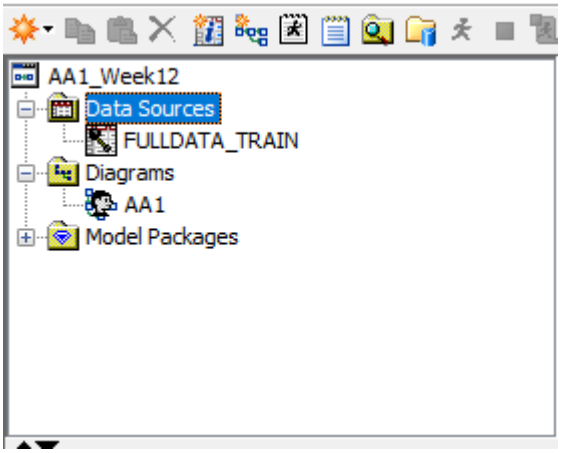




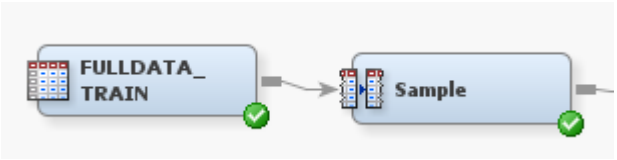


Property	Value
Output Options	
Variables	
Filename Prefix	FullData
Replace Existing Files	No
All Observations	Yes
Number of Observations	1000
Output Format	
File Format	SAS (.sas7bdat)
SAS Library Name	AA1
Directory	
Output Data	
All Roles	Yes
Select Roles	

3. Create Data Source



4. Drop SAS data into diagram, then do sampling (10% of the population).

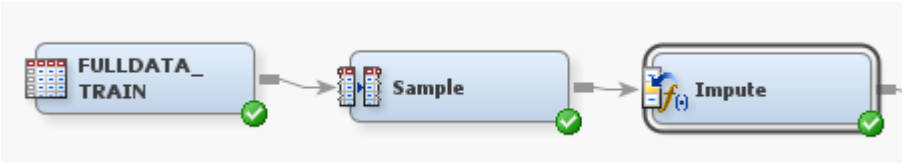


Train	
Variables	
Output Type	Data
Sample Method	Default
Random Seed	12345
Size	
Type	Percentage
Observations	.
Percentage	10.0
Alpha	0.01
PValue	0.01
Cluster Method	Random

Sampling Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Ids4_DATA	47237
SAMPLE	EMWS1.Smpl_DATA	4724

5. Drop imputation node for manipulating the missing values. Impute using the MODE method.



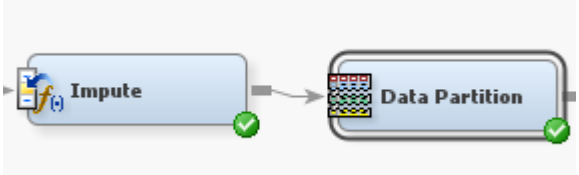
Missing Cutoff	50.0
Class Variables	
Default Input Method	Count
Default Target Method	Count
Normalize Values	Yes

Identify the number of missing values

Imputation Summary

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Returns	COUNT	IMP_Retur...	0	INPUT	BINARY	Returns	871

6. Data Splitting 50 for train and 50 for valid



Property	Value
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	1/7/24 10:56 AM
Run ID	e7d50cf2-8c5d-481e-abf

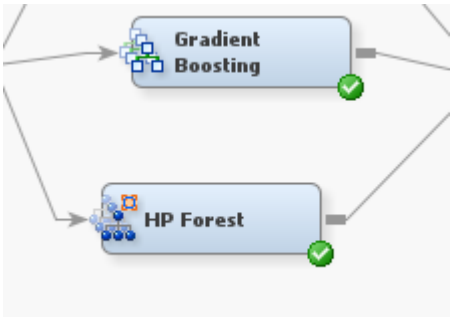
7. Modeling using decision tree

Edit the splitting rule

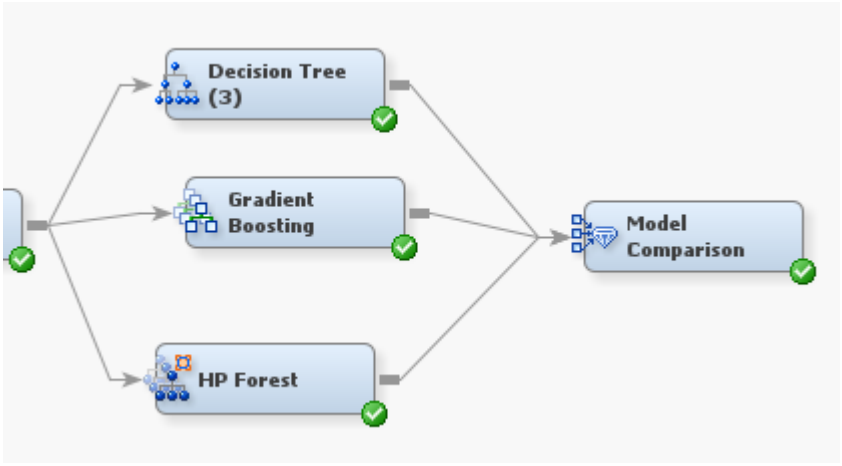
Splitting Rule	
Interval Target Criterion	Variance
Nominal Target Criterion	Entropy
Ordinal Target Criterion	Gini
Significance Level	0.2
Missing Values	Use in sea
Use Input Once	No
Maximum Branch	3
Maximum Depth	7
Minimum Categorical Size	5

Gini

8. Ensemble Method: drop gradient boosting and random forest nodes



9. Model Comparison: drop model comparison node


Decision Tree
(3)
Gradient
Boosting
HP Forest
Model
Comparison