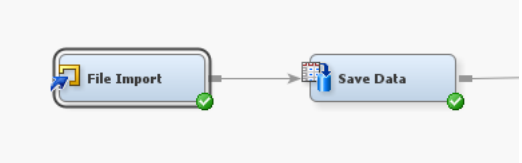
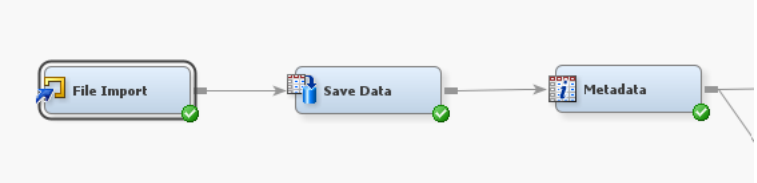
4.0 Data Import and Pre-processing using SAS Enterprise Miner

4.1 Importing Data

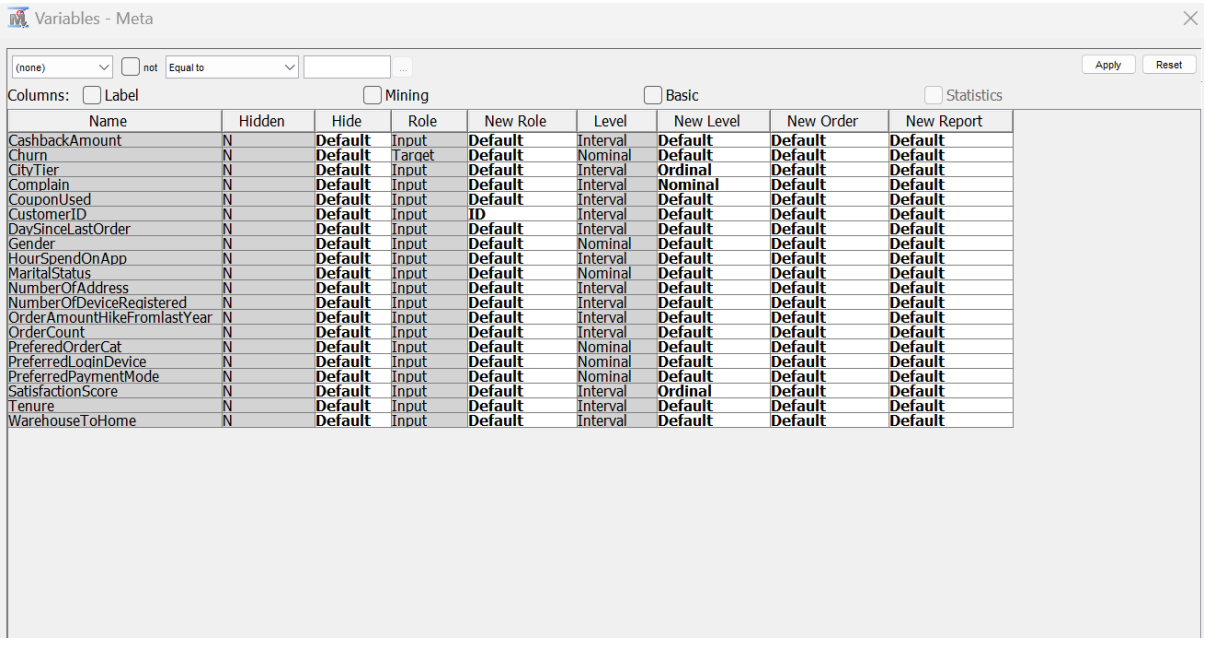
Import the CSV file using “File Import” node. Save it as a SAS file.



Specify the column metadata using “Metadata” node.



The roles and measurement levels of some variables were re-assigned to correctly define the column metadata. For instance, “customerID” should be assigned the role of ID instead of input because it serves as an identifier for individual customers rather than being used as an input feature for modelling. “CityTier” and “SatisfactionScore” should be considered as ordinal variables. “City Tier” ranks cities into different tiers, typically based on their economic development and other similar factors, making it an ordinal variable. For “SatisfactionScore”, the numbers represent a respondent’s level of satisfaction with a product or service, making it an ordinal variable. “Complain” taking values of 0 or 1 should be considered as nominal variable because it indicates the presence or absence of a complaint without any inherent order or ranking, hence it should be treated as a nominal variable.



Name	Hidden	Hide	Role	New Role	Level	New Level	New Order	New Report
CashbackAmount	N	Default	Input	Default	Interval	Default	Default	Default
Churn	N	Default	Target	Default	Nominal	Default	Default	Default
CityTier	N	Default	Input	Default	Interval	Ordinal	Default	Default
Complain	N	Default	Input	Default	Interval	Nominal	Default	Default
CouponUsed	N	Default	Input	Default	Interval	Default	Default	Default
CustomerID	N	Default	Input	ID	Interval	Default	Default	Default
DaySinceLastOrder	N	Default	Input	Default	Interval	Default	Default	Default
Gender	N	Default	Input	Default	Nominal	Default	Default	Default
HourSpendOnApp	N	Default	Input	Default	Interval	Default	Default	Default
MaritalStatus	N	Default	Input	Default	Nominal	Default	Default	Default
NumberOfAddress	N	Default	Input	Default	Interval	Default	Default	Default
NumberOfDeviceRegistered	N	Default	Input	Default	Interval	Default	Default	Default
OrderAmountHikeFromlastYear	N	Default	Input	Default	Interval	Default	Default	Default
OrderCount	N	Default	Input	Default	Interval	Default	Default	Default
PreferedOrderCat	N	Default	Input	Default	Nominal	Default	Default	Default
PreferredLoginDevice	N	Default	Input	Default	Nominal	Default	Default	Default
PreferredPaymentMode	N	Default	Input	Default	Nominal	Default	Default	Default
SatisfactionScore	N	Default	Input	Default	Interval	Ordinal	Default	Default
Tenure	N	Default	Input	Default	Interval	Default	Default	Default
WarehouseToHome	N	Default	Input	Default	Interval	Default	Default	Default

4.2 Handling Missing Data

Check for missing values using StatExplore Node.

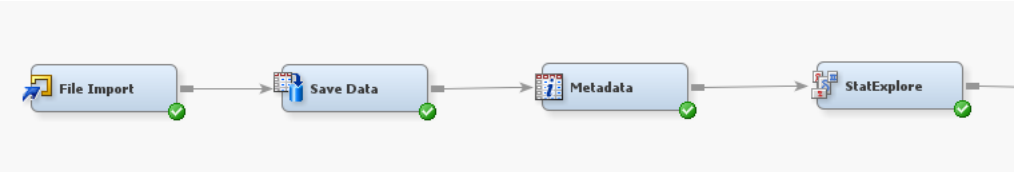


Figure below shows variables summary after specifying the metadata.

Variable Summary

Role	Measurement Level	Frequency Count
ID	INTERVAL	1
INPUT	INTERVAL	10
INPUT	NOMINAL	6
INPUT	ORDINAL	2
TARGET	NOMINAL	1

Figure below shows class variables summary. None of the class variables have missing values.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	CityTier	INPUT	3	0	1	65.12	3	30.59
TRAIN	Complain	INPUT	2	0	0	71.51	1	28.49
TRAIN	Gender	INPUT	2	0	Male	60.11	Female	39.89
TRAIN	MaritalStatus	INPUT	3	0	Married	53.04	Single	31.90
TRAIN	PreferedOrderCat	INPUT	5	0	Mobile	36.94	Laptop & Accessory	36.41
TRAIN	PreferredLoginDevice	INPUT	2	0	Mobile Phone	70.98	Computer	29.02
TRAIN	PreferredPaymentMode	INPUT	5	0	Debit Card	41.10	Credit Card	31.51
TRAIN	SatisfactionScore	INPUT	5	0	3	30.16	1	20.67
TRAIN	Churn	TARGET	2	0	0	83.16	1	16.84

Figure below shows interval variables summary. Seven variables namely “CouponUsed”, “DaySinceaLastOrder”, “HourSpendOnApp”, “OrderAmountHikeFromLastYear”, “OrderCount”, “Tenure” and “WarehouseToHome” have missing values.

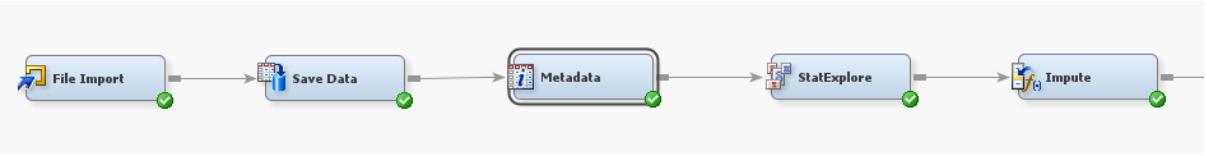
Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
CashbackAmount	INPUT	177.2215	49.19387	5630	0	0	163	325	1.149595	0.973546
CouponUsed	INPUT	1.751023	1.894621	5374	256	0	1	16	2.545653	9.132281
DaySinceLastOrder	INPUT	4.543491	3.654433	5323	307	0	3	46	1.191	4.023964
HourSpendOnApp	INPUT	2.931535	0.721926	5375	255	0	3	5	-0.02721	-0.66708
NumberOfAddress	INPUT	4.214032	2.583586	5630	0	1	3	22	1.088639	0.959229
NumberOfDeviceRegistered	INPUT	3.688988	1.023999	5630	0	1	4	6	-0.39697	0.582849
OrderAmountHikeFromlastYear	INPUT	15.70792	3.675485	5365	265	11	15	26	0.790785	-0.28038
OrderCount	INPUT	3.008004	2.93968	5372	258	1	2	16	2.196414	4.718466
Tenure	INPUT	10.1899	8.557241	5366	264	0	9	61	0.736513	-0.00737
WarehouseToHome	INPUT	15.60271	8.261845	5379	251	5	14	36	0.898406	-0.28639

4.3 Imputing Missing Data

Impute missing values using “Impute” node



The missing values of interval variables were imputed using the mean values. Imputing missing values with the mean assumes that the missing data is missing completely at random (MCAR) or missing at random (MAR) and does not introduce bias.

Property	Value
General	
Node ID	Impt
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Count
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Mean
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	.
Method Options	
Random Seed	12345
Tuning Parameters	
Tree Imputation	
Score	
Hide Original Variables	Yes
Indicator Variables	
Type	None
Source	Imputed Variables
Role	Rejected
Report	
Validation and Test Data	No
Distribution of Missing	No
Status	
Create Time	1/6/24 4:06 PM
Run ID	4eea0a9d-1763-f148-8f23-
Last Error	
Last Status	Complete
Last Run Time	1/7/24 2:46 AM
Run Duration	0 Hr. 0 Min. 3.20 Sec.
Grid Host	
User-Added Node	No

4.4 Assessing Impact of Imputation

Assess the impact of imputation using “StatExplore” node.

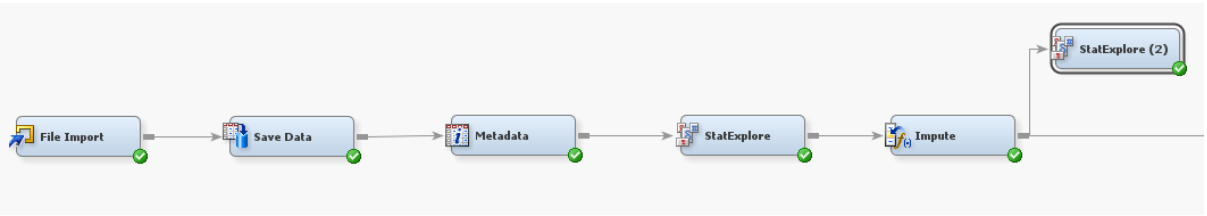


Figure below shows interval variables summary before imputation.

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
CashbackAmount	INPUT	177.2215	49.19387	5630	0	0	163	325	1.149595	0.973546
CouponUsed	INPUT	1.751023	1.894621	5374	256	0	1	16	2.545653	9.132281
DaySinceLastOrder	INPUT	4.543491	3.654433	5323	307	0	3	46	1.191	4.023964
HourSpendOnApp	INPUT	2.931535	0.721926	5375	255	0	3	5	-0.02721	-0.66708
NumberOfAddress	INPUT	4.214032	2.583586	5630	0	1	3	22	1.088639	0.959229
NumberOfDeviceRegistered	INPUT	3.688988	1.023999	5630	0	1	4	6	-0.39697	0.582849
OrderAmountHikeFromlastYear	INPUT	15.70792	3.675485	5365	265	11	15	26	0.790785	-0.28038
OrderCount	INPUT	3.008004	2.93968	5372	258	1	2	16	2.196414	4.718466
Tenure	INPUT	10.1899	8.557241	5366	264	0	9	61	0.736513	-0.00737
WarehouseToHome	INPUT	15.60271	8.261845	5379	251	5	14	36	0.898406	-0.28639

Figure below shows interval variables summary after imputation. The imputation did not significantly alter the distribution or central tendencies of the variables.

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
CashbackAmount	INPUT	177.2215	49.19387	5630	0	0	163	325	1.149595	0.973546
IMP_CouponUsed	INPUT	1.751023	1.851038	5630	0	0	1	16	2.605547	9.709842
IMP_DaySinceLastOrder	INPUT	4.543491	3.553382	5630	0	0	4	46	1.224844	4.428875
IMP_HourSpendOnApp	INPUT	2.931535	0.705384	5630	0	0	3	5	-0.02785	-0.55635
IMP_OrderAmountHikeFromlastYear	INPUT	15.70792	3.587926	5630	0	11	15	26	0.810069	-0.14601
IMP_OrderCount	INPUT	3.008004	2.871521	5630	0	1	2	16	2.24851	5.088971
IMP_Tenure	INPUT	10.1899	8.354164	5630	0	0	9	61	0.754404	0.139888
IMP_WarehouseToHome	INPUT	15.60271	8.075545	5630	0	5	14	36	0.919117	-0.15973
NumberOfAddress	INPUT	4.214032	2.583586	5630	0	1	3	22	1.088639	0.959229
NumberOfDeviceRegistered	INPUT	3.688988	1.023999	5630	0	1	4	6	-0.39697	0.582849