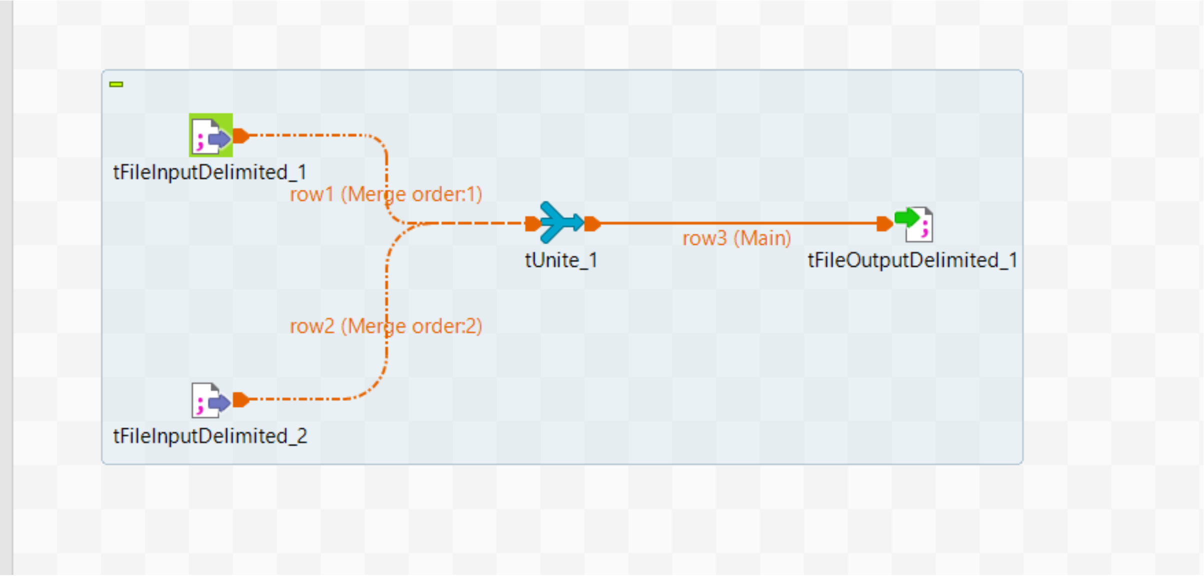
Documentation for Talend Data Integration

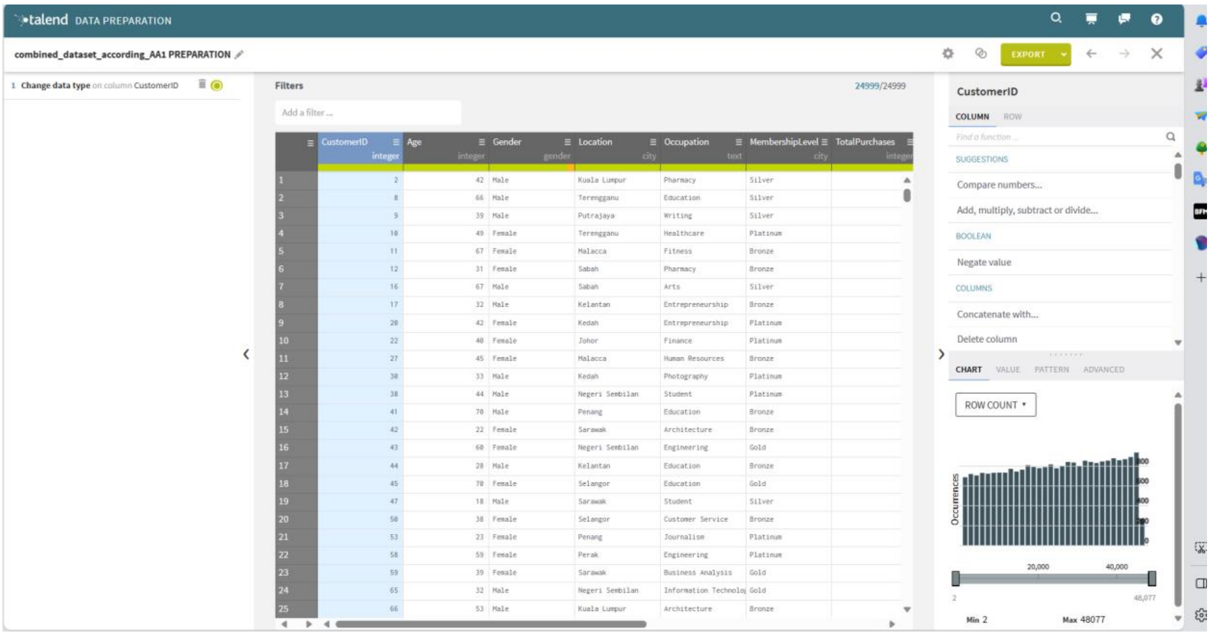
- Open Talend Data Integration
- Create new project named wqd7005-AA1
- Create a new job
- Add in the tFileInputDelimited node, tUnite_1 node and tFileOutputDelimited



- Set the fil path for input .csv file and file path for output .csv file
- Set the schema and synchronize it
- Run the job

Documentation for Talend Data Preparation

- Open Talend Data Preparation
- Add datasets from local device to the Talend Data Preparation



combined_dataset_according_AA1 PREPARATION

1 Change data type on column CustomerID

Filters

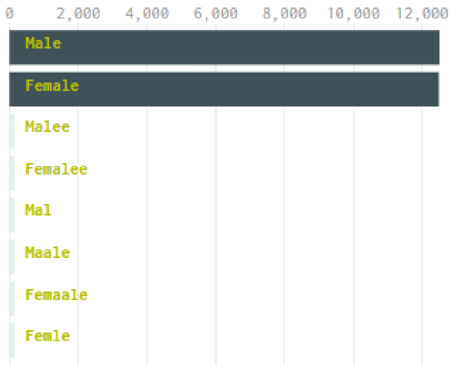
Add a filter ...

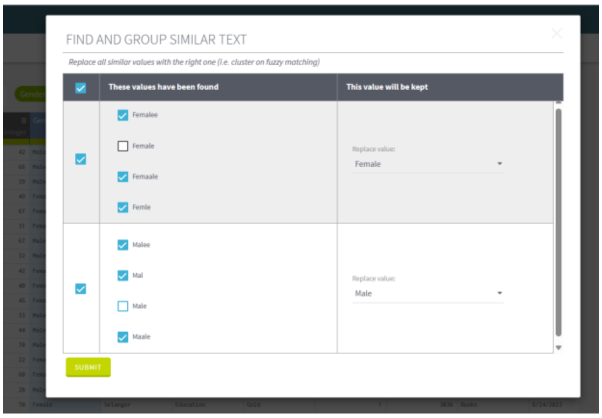
24999/24999

	CustomerID	Age	Gender	Location	Occupation	MembershipLevel	TotalPurchases
	integer	integer	gender	city	text	city	integer
1	2	42	Male	Kuala Lumpur	Pharmacy	Silver	
2	8	66	Male	Terengganu	Education	Silver	
3	9	39	Male	Putrajaya	Writing	Silver	
4	10	49	Female	Terengganu	Healthcare	Platinum	
5	11	67	Female	Malacca	Fitness	Bronze	
6	12	31	Female	Sabah	Pharmacy	Bronze	
7	16	67	Male	Sabah	Arts	Silver	
8	17	32	Male	Kelantan	Entrepreneurship	Bronze	
9	20	42	Female	Kedah	Entrepreneurship	Platinum	
10	22	40	Female	Johor	Finance	Platinum	
11	27	45	Female	Malacca	Human Resources	Bronze	
12	30	33	Male	Kedah	Photography	Platinum	
13	38	44	Male	Negeri Sembilan	Student	Platinum	
14	41	70	Male	Penang	Education	Bronze	
15	42	22	Female	Sarawak	Architecture	Bronze	
16	43	60	Female	Negeri Sembilan	Engineering	Gold	
17	44	28	Male	Kelantan	Education	Bronze	
18	45	70	Female	Selangor	Education	Gold	
19	47	18	Male	Sarawak	Student	Silver	
20	50	38	Female	Selangor	Customer Service	Bronze	
21	53	23	Female	Penang	Journalism	Platinum	
22	58	59	Female	Perak	Engineering	Platinum	
23	59	39	Female	Sarawak	Business Analysis	Gold	
24	65	32	Male	Negeri Sembilan	Information Technolo	Gold	
25	66	53	Male	Kuala Lumpur	Architecture	Bronze	

- Check for invalid values and data formatting

- Correcting the Gender using “find and group similar text…”



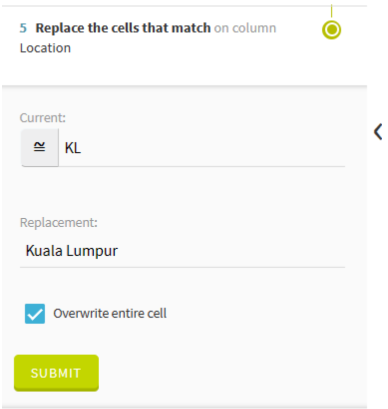

FIND AND GROUP SIMILAR TEXT

Replace all similar values with the right one (i.e. cluster on fuzzy matching)

☑	These values have been found	This value will be kept
☑	☑ Femalee □ Female ☑ Femaale ☑ Femle	Replace value: Female
☑	☑ Malee ☑ Mal □ Male ☑ Maale	Replace value: Male

SUBMIT

- Correcting the wrongly labelled Kuala Lumpur at Location attribute using “replace the cells that match…”


5 **Replace the cells that match** on column Location

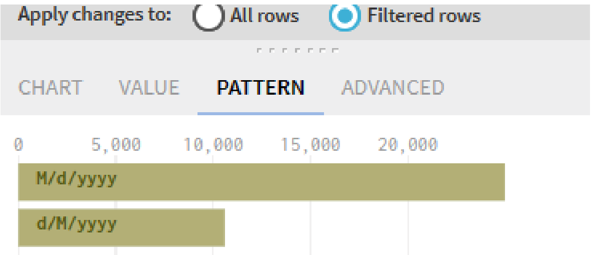
Current:
≅ KL

Replacement:
Kuala Lumpur

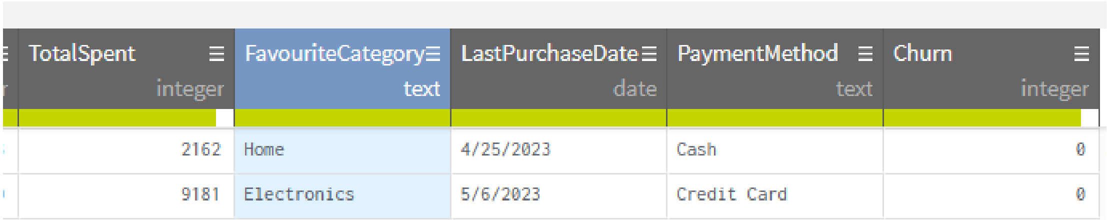
☑ Overwrite entire cell

SUBMIT

- Check for date format whether it is standardized.


Apply changes to: ○ All rows ● Filtered rows

- Missing data is found in both TotalSpent and Churn Attribute. However, it is ignored as the missing value will be handled by SAS Enterprise Miner.

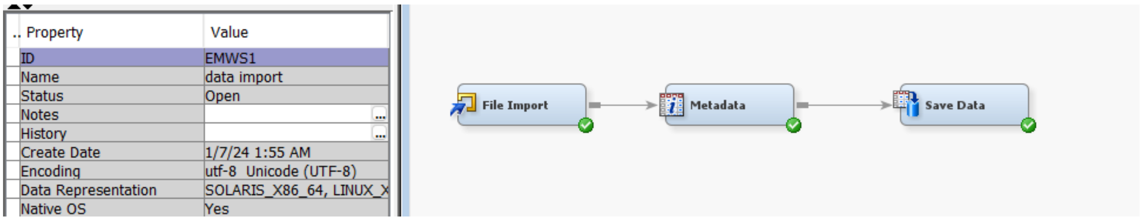


TotalSpent integer	FavouriteCategory text	LastPurchaseDate date	PaymentMethod text	Churn integer
2162	Home	4/25/2023	Cash	0
9181	Electronics	5/6/2023	Credit Card	0

- Export the dataset as basic_clean_dataset_2023.csv

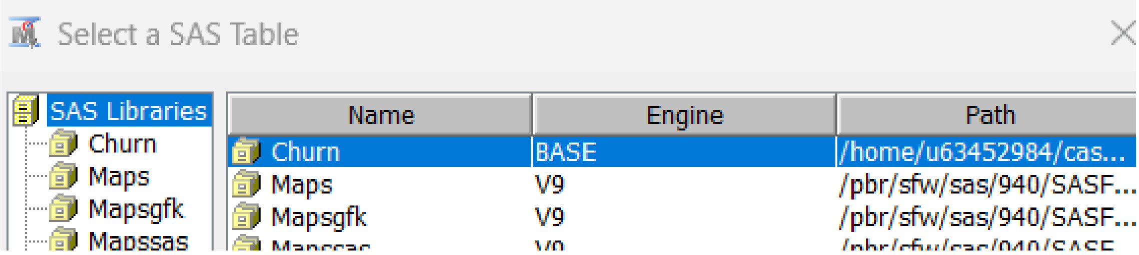
Documentation for SAS Enteprise Miner

- Open SAS Enterprise Miner 15.2
- Create a new project
- Create a new diagram, add in the “File Import”, “Metadata” and “Save Data” node



.. Property	Value
ID	EMWS1
Name	data import
Status	Open
Notes	
History	
Create Date	1/7/24 1:55 AM
Encoding	utf-8 Unicode (UTF-8)
Data Representation	SOLARIS_X86_64, LINUX_X
Native OS	Yes

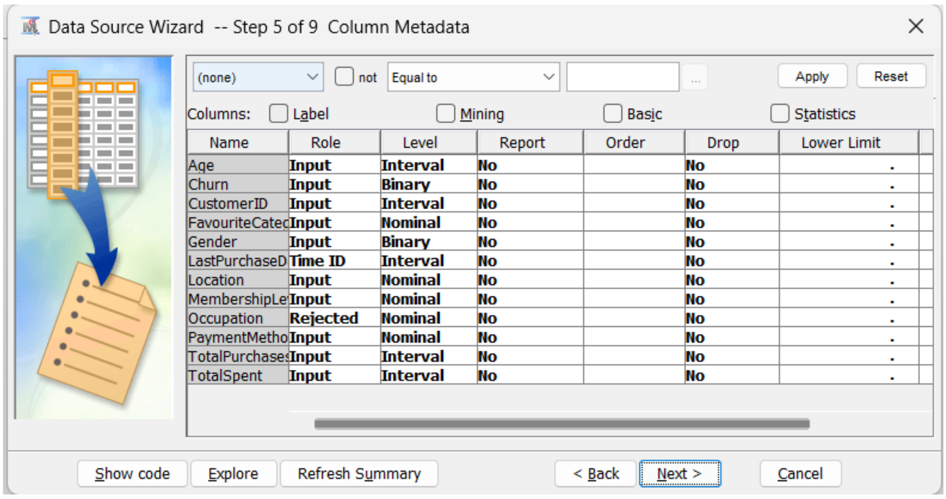
- Create a new library from the save data
- Create a new data source from the created table



Select a SAS Table

Name	Engine	Path
Churn	BASE	/home/u63452984/cas...
Maps	V9	/pbr/sfw/sas/940/SASF...
Mapsgfk	V9	/pbr/sfw/sas/940/SASF...
Mapssas	V9	/pbr/sfw/sas/940/SASF

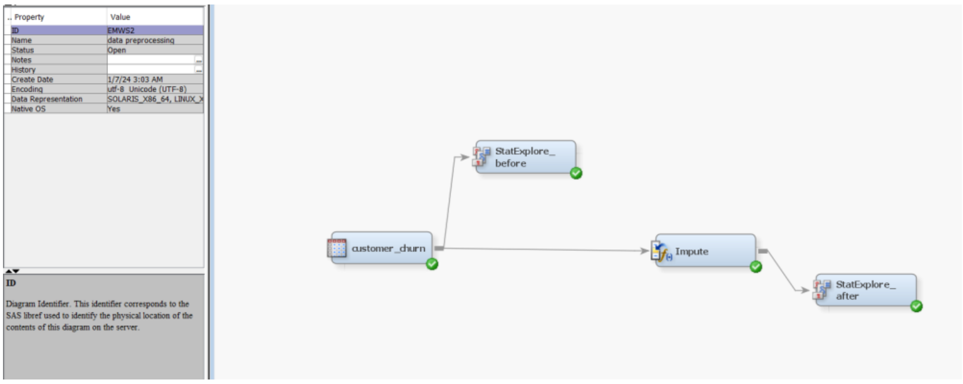
- Specifying the variable role and level



Data Source Wizard -- Step 5 of 9 Column Metadata

Name	Role	Level	Report	Order	Drop	Lower Limit
Age	Input	Interval	No		No	.
Churn	Input	Binary	No		No	.
CustomerID	Input	Interval	No		No	.
FavouriteCatec	Input	Nominal	No		No	.
Gender	Input	Binary	No		No	.
LastPurchaseD	Time ID	Interval	No		No	.
Location	Input	Nominal	No		No	.
MembershipLe	Input	Nominal	No		No	.
Occupation	Rejected	Nominal	No		No	.
PaymentMetho	Input	Nominal	No		No	.
TotalPurchase	Input	Interval	No		No	.
TotalSpent	Input	Interval	No		No	.

- Use StatExplore to check for missing values and “Impute” to fill up the missing values

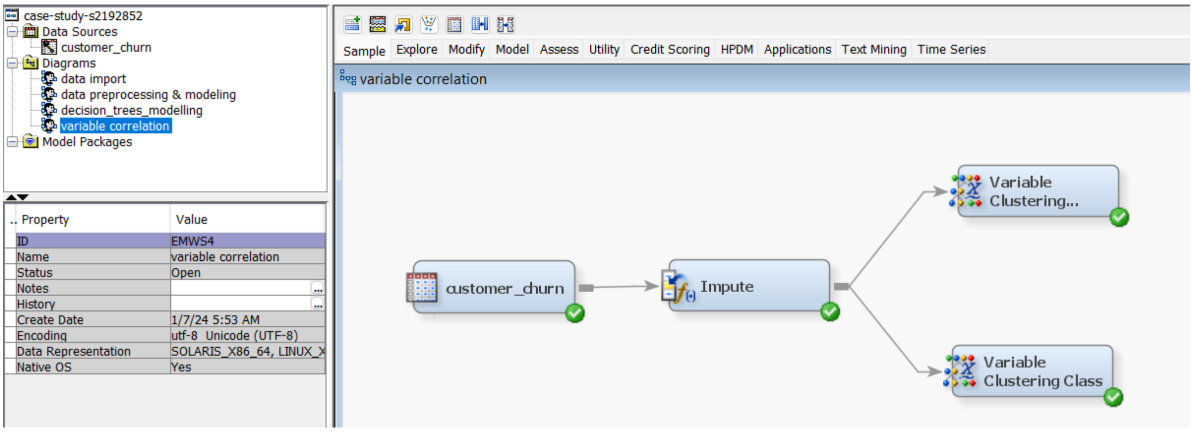


.. Property	Value
ID	EMWS2
Name	data preprocessing
Status	Open
Notes	
History	
Create Date	1/7/24 3:03 AM
Encoding	utf-8 Unicode (UTF-8)
Data Representation	SOLARIS_X86_64, LINUX_X
Native OS	Yes

ID

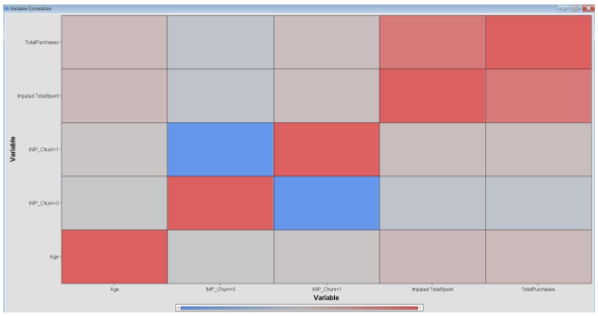
Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

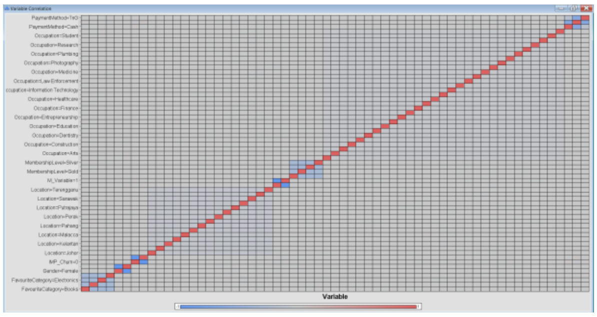
- Perform variable selection via correlation analysis



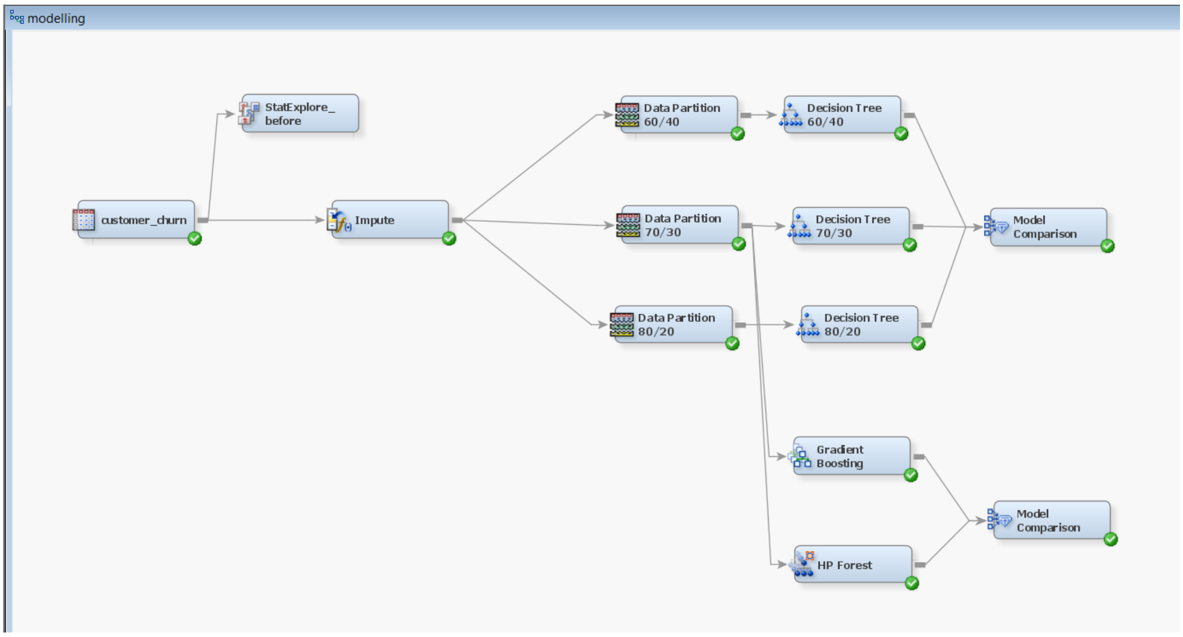
case-study-s2192852
- Data Sources
  - customer_churn
- Diagrams
  - data import
  - data preprocessing & modeling
  - decision_trees_modelling
  - variable correlation
- Model Packages

.. Property	Value
ID	EMWS4
Name	variable correlation
Status	Open
Notes	
History	
Create Date	1/7/24 5:53 AM
Encoding	utf-8 Unicode (UTF-8)
Data Representation	SOLARIS_X86_64, LINUX_X
Native OS	Yes





- Identify the data splitting ratio
- Perform Decision Tree modelling using Decision Tree node, Bagging modelling using HP Forest node and Boosting modelling using Gradient Boosting node



- Evaluate the results