WQD7005 SEM 1 2023/24

ALTERNATIVE ASSESSMENT (AA) 1: CASE STUDY

The GitHub link for all necessary documents and components is as follows:
https://github.com/unplayable98/7005AA1-22079565

Overview:
This study is to apply various tools to perform various operations to analyse a dataset regarding the customer behaviour and aim to generate insights based on the result of modelling and provide business suggestions. if possible

The applied tools and their roles are as follows:

Talend Data Integration (TDI): Perform data preprocessing such as joining of dataset.
Talend Data Preparation (TDP): Perform various data preprocessing to improve data quality such as data cleaning and data transformation.
SAS Enterprise Miner (SAS-EM): To perform decision tree-related modelling to analyse customer behaviours.

Dataset:

The dataset applied for this assessment is generated from an online data generator named Mockaroo (mockaroo.com). There are 1000 rows generated for this dataset, which is the maximum number of rows able to be generated on this website without a paid subscription.

While designing the data, the dataset “E-commerce Customer Behavior Dataset” by Laksika Tharmalingam on the Kaggle website is referred to. The designation of data is as follows:

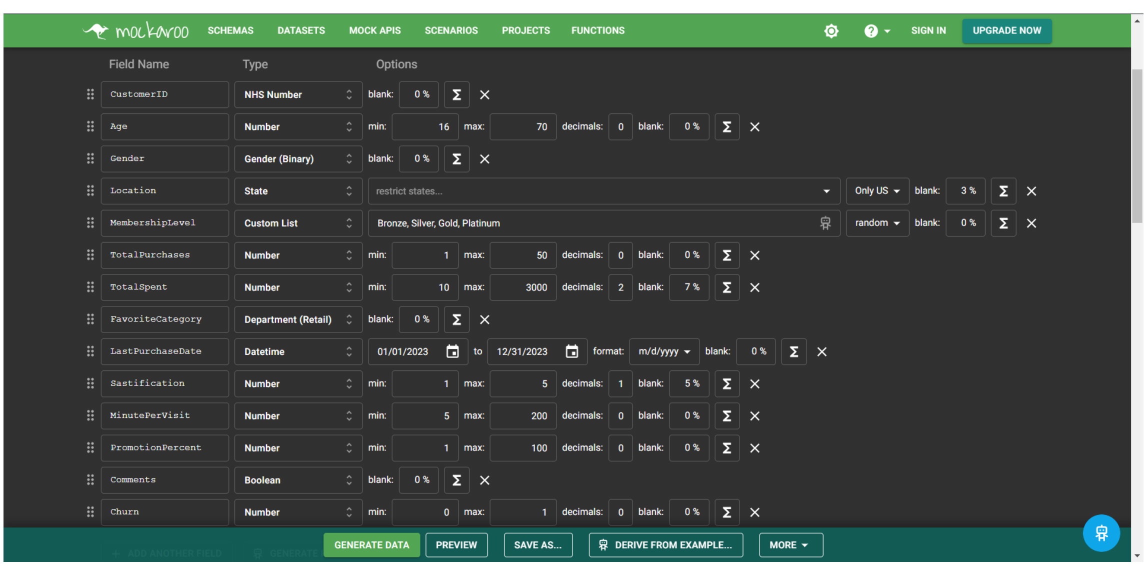


Figure 1: Dataset designation.

Besides the listed columns in the question, there are four columns created in this dataset, making the total column numbers to 14. The additional columns and their description are listed as follows:

Satisfaction: The average satisfaction of the customer for every purchase. The range is from 1 (Very unsatisfied) to 5 (Satisfied)
MinutePerVisit: The time stay on the website in minutes per visit.
PromotionPercent: The proportion of items purchased during promotion.
Comments: Whether the customer has given comments after purchase before.

Noted that the columns Location, TotalSpent and Satisfaction (after renaming) are designated to have missing values for 3 to 7% of the rows.

Dataset reference:
https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset

Dataset generated:
https://drive.google.com/file/d/1Q9DjruSQt2Ynp6NWz7IdTpDmym6oVMYJ/view?usp=sharing

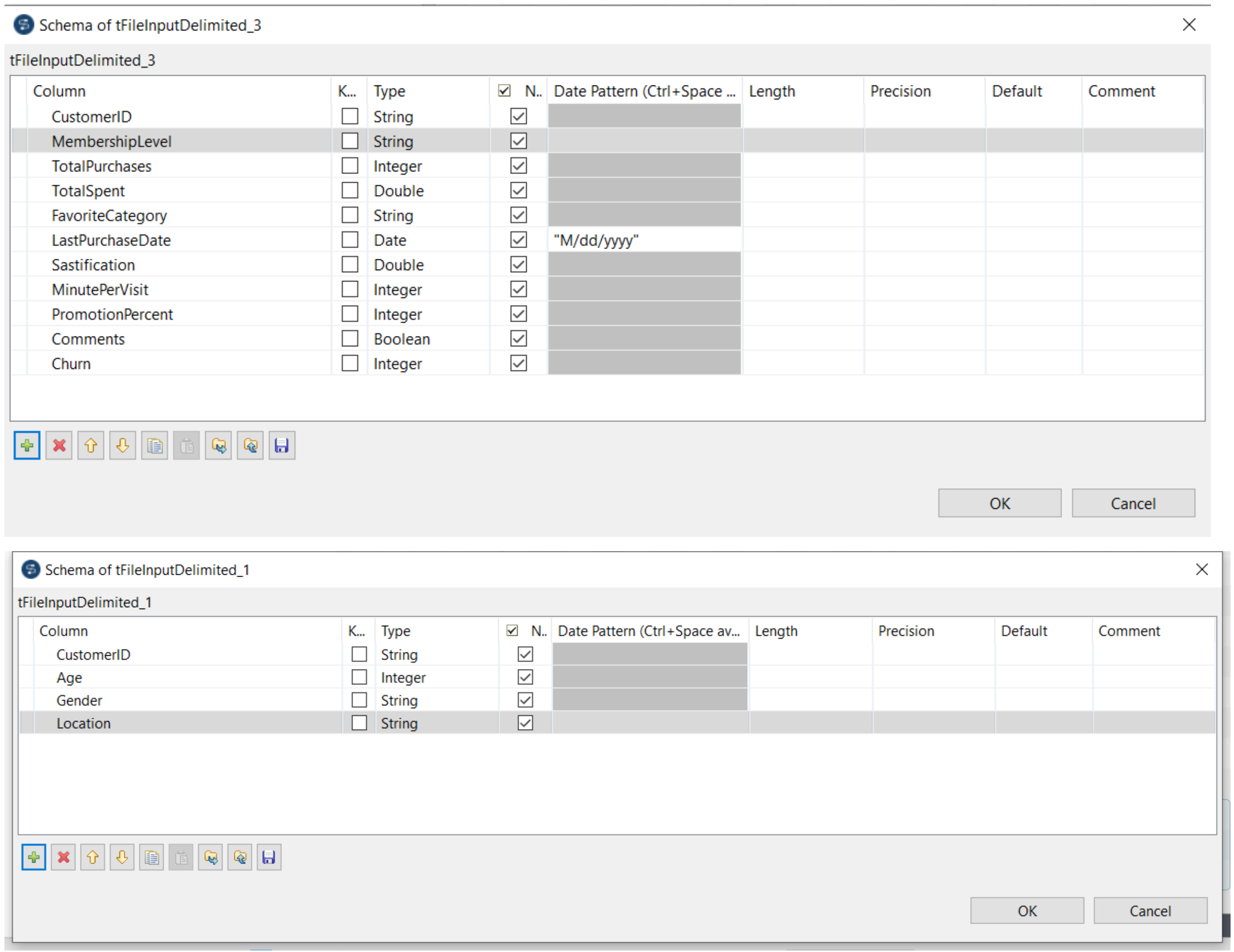
Before entering the data, I separate a few columns from the dataset manually into a reduced dataset. The list of columns for the dataset that needed to load into the Talend Data Integration is as follows:

D7005AA1_DataA	D7005AA1_DataB
CustomerID Age Gender Location	CustomerID MembershipLevel TotalPurchases TotalSpent FavouriteCategory LastPurchaseData Satisfaction MinutePerVisit PromotionPercent Comments Churn

Data Preparation Using Talend Data Integration and Talend Data Preparation:

Before loading both datasets (D7005AA1_DataA and D7005AA1_DataB) into SAS Enterprise Miner (SAS EM) for analysis, the preprocessing of data is performed on Talend Data Integration

(TDI) and Talend Data Preparation (TDP). First of all, the dataset is loaded into Talend Data Integration using the following schema:


Schema of tFileInputDelimited_3

tFileInputDelimited_3

Column	K...	Type	N..	Date Pattern (Ctrl+Space ...	Length	Precision	Default	Comment
CustomerID	☐	String	☑					
MembershipLevel	☐	String	☑					
TotalPurchases	☐	Integer	☑					
TotalSpent	☐	Double	☑					
FavoriteCategory	☐	String	☑					
LastPurchaseDate	☐	Date	☑	"M/dd/yyyy"				
Sastification	☐	Double	☑					
MinutePerVisit	☐	Integer	☑					
PromotionPercent	☐	Integer	☑					
Comments	☐	Boolean	☑					
Churn	☐	Integer	☑					

OK Cancel

Schema of tFileInputDelimited_1

tFileInputDelimited_1

Column	K...	Type	N..	Date Pattern (Ctrl+Space av...	Length	Precision	Default	Comment
CustomerID	☐	String	☑					
Age	☐	Integer	☑					
Gender	☐	String	☑					
Location	☐	String	☑					

OK Cancel

Figure 2: Schema when loading the dataset into Talend Data Integration.

Although the CustomerID is made up of 10 to 11-digit integers, however when the data type of CustomerID is set as Integer, some errors exist as the system identifies some of the numbers as strings. As the CustomerID is only used to check for duplicates, which is not involved much in analysis, the string data type is acceptable for Customer ID. The error is shown in Figure 3.

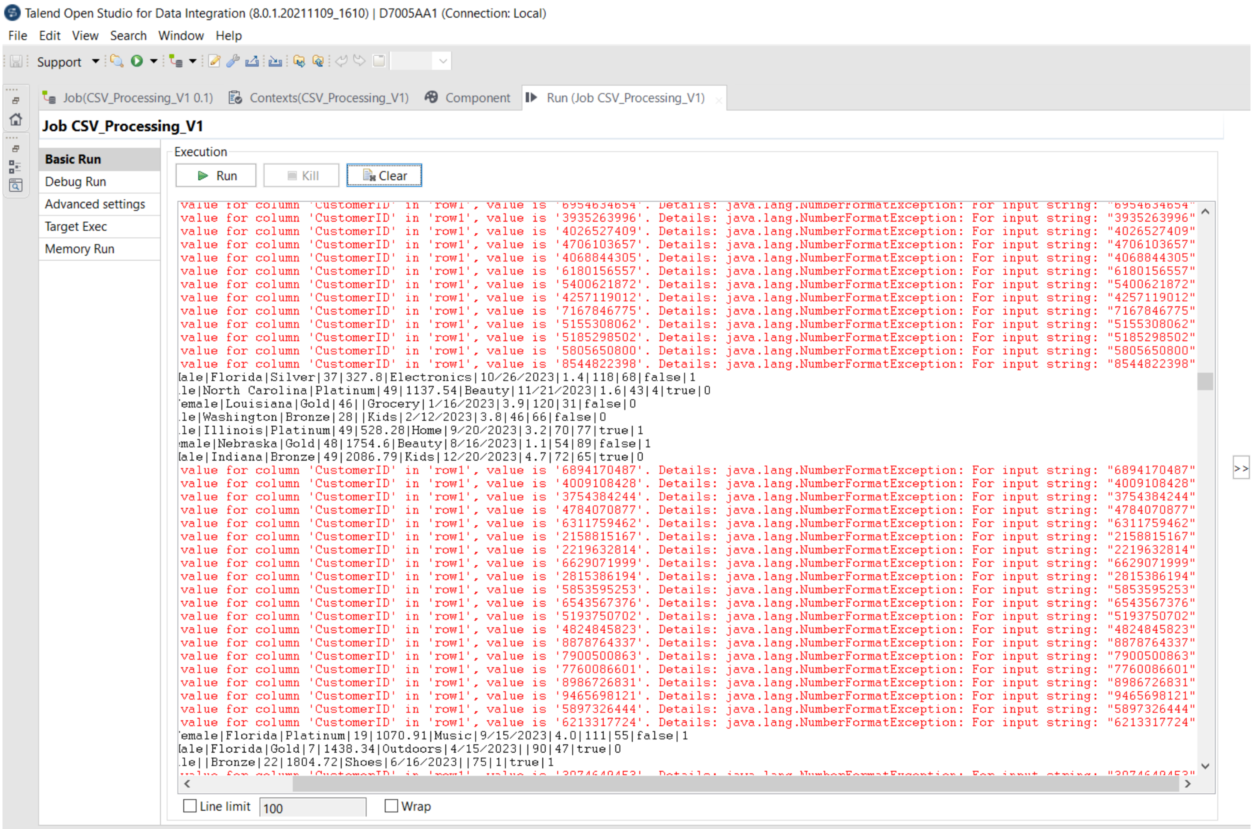


Figure 3: Error when loading the dataset to TDI when CustomerID type is set as integer.

First of all, both datasets are related so they needed to be joined together. The tLogRow component provides an output that enables us to determine whether the data under processing is correct. The tMap is used to join both datasets.

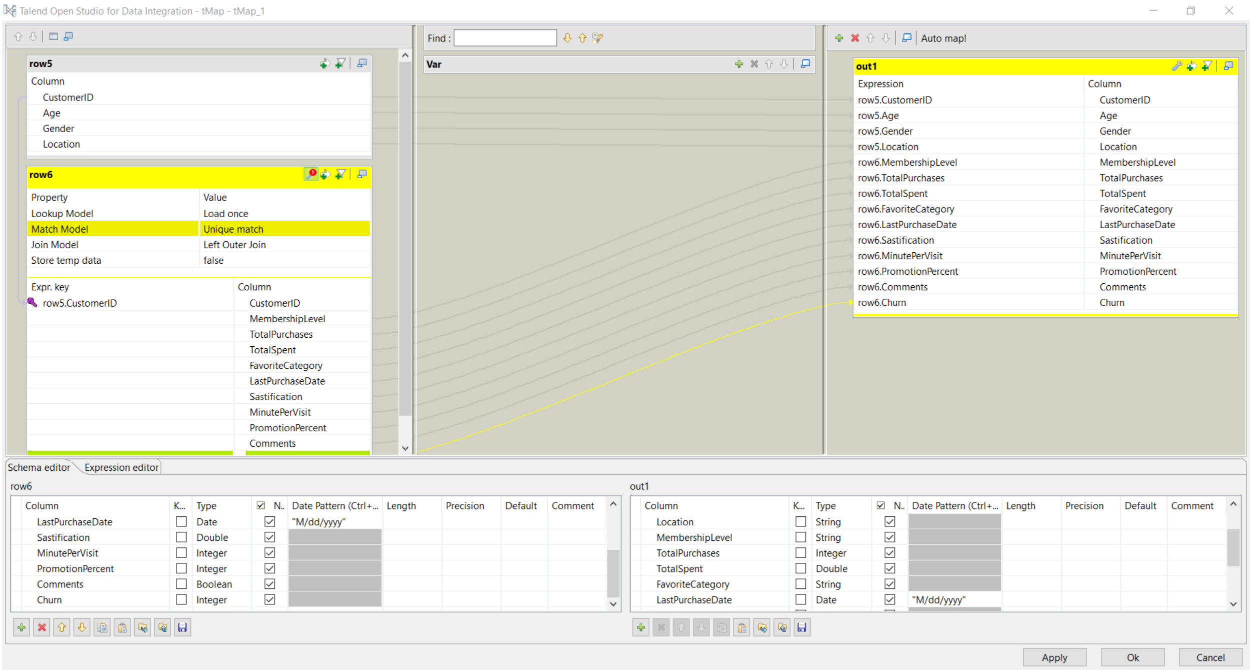


Figure 4: Setting of tMap.

Figure 4 shows both datasets are joined together based on the unique CustomerID match. After the output schema is determined, the input column is linked to the desired output column.

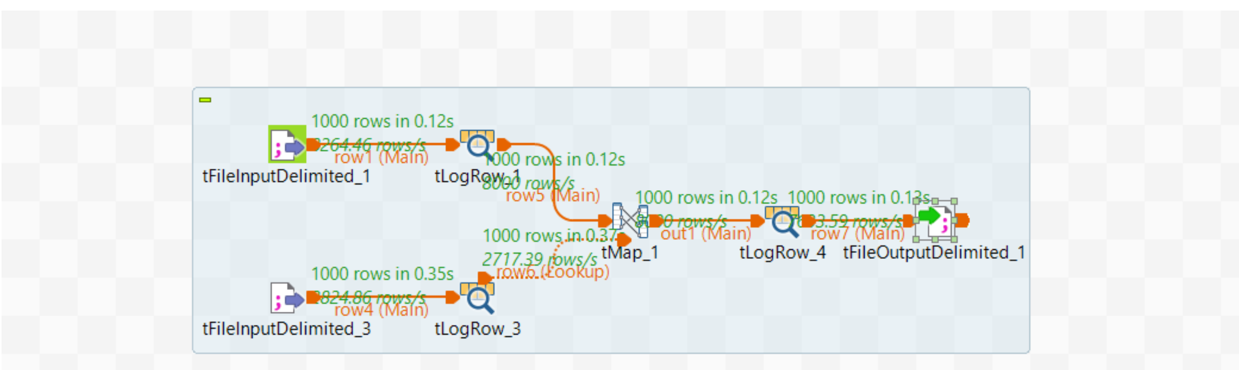


Figure 5: Workflow of producing combined CSV dataset.

As shown in Figure 5, after the datasets are joined, the combined dataset is now ready to be exported for further usage. To ensure smoother loading of data in the further step, the settings as shown in Figure 6 are applied.

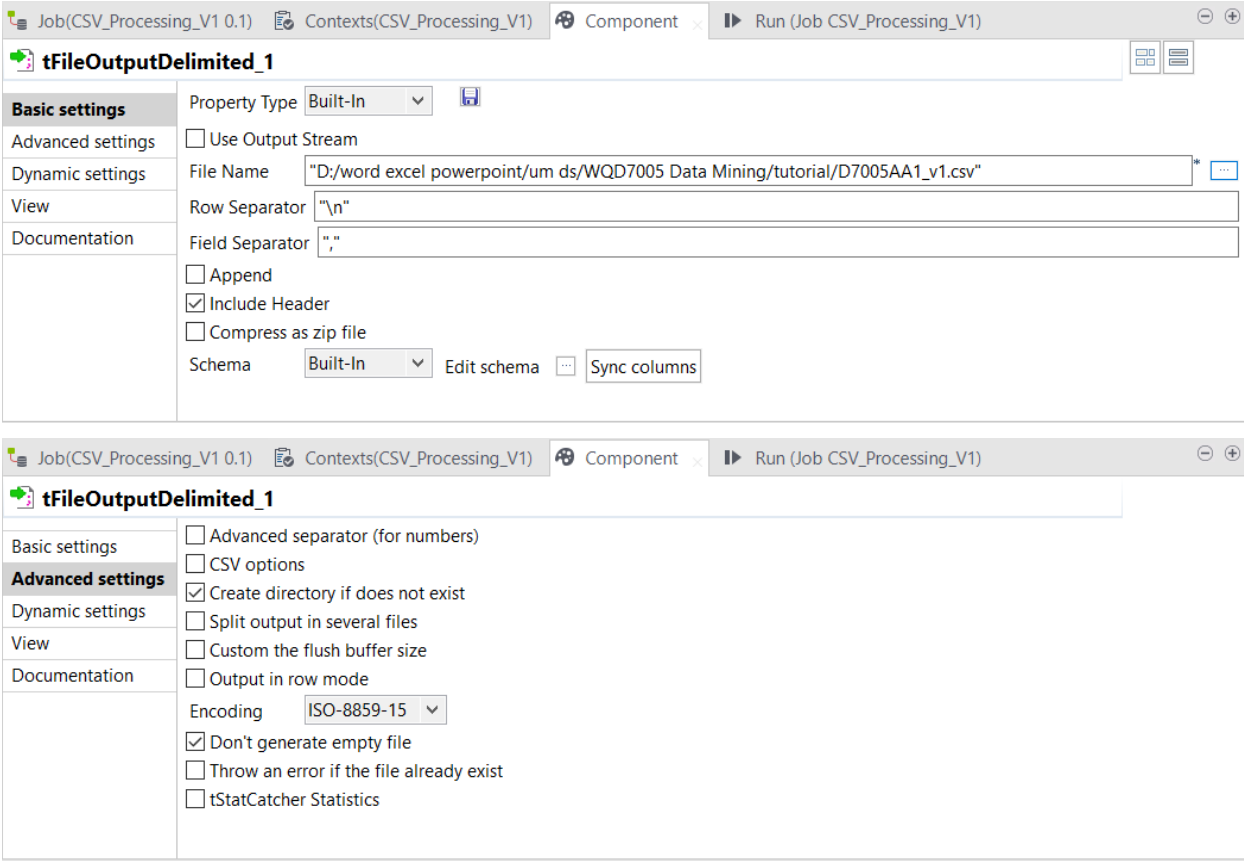


Figure 6: Setting of export of combined CSV dataset.

The newly exported dataset is now able to be used. After loading the combined dataset into TDI, the first attempt of handling missing values is operated, which is dropping all the rows with missing values. By using the tFilterRow component with the setting as shown in Figure 7, the remaining number of rows is shown.

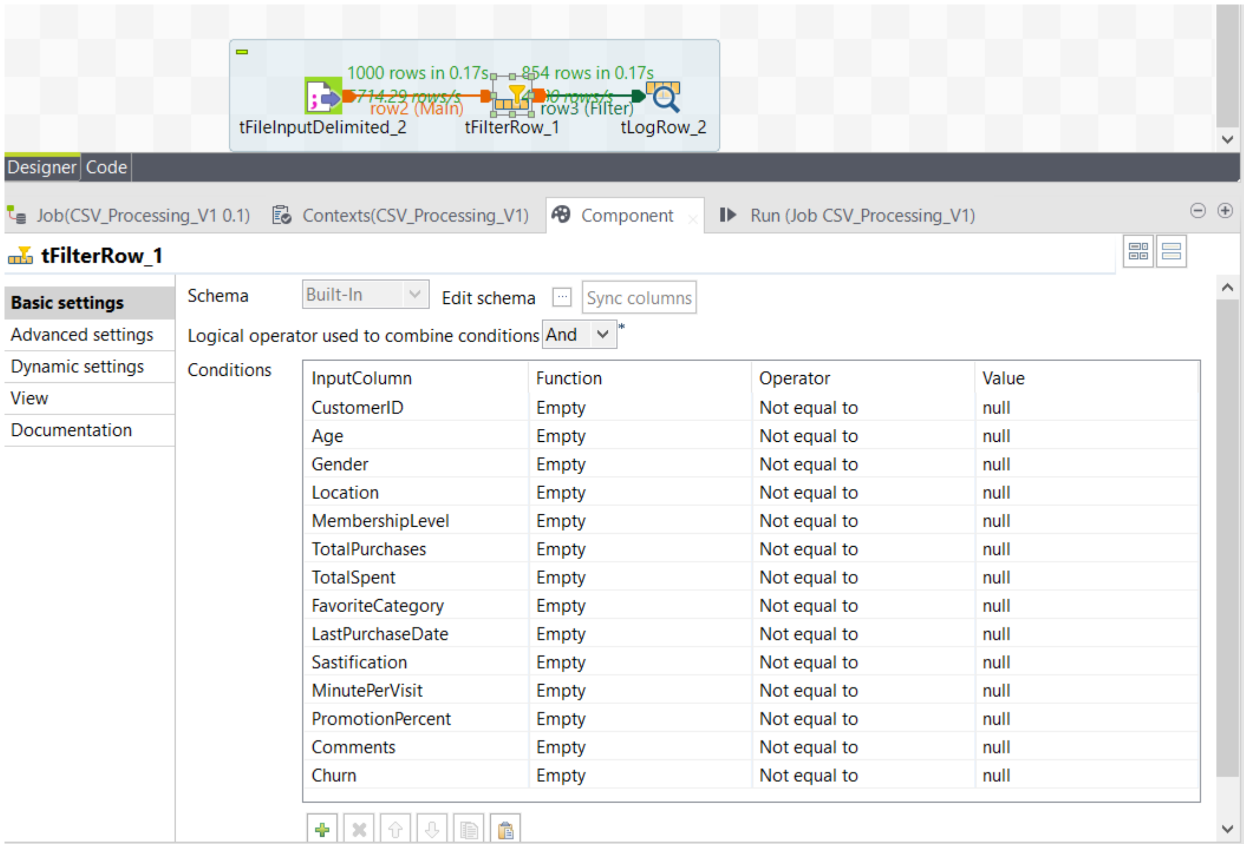


Figure 7: Filtering out all rows with missing values.

As shown in the figure above, only 854 rows are remaining in the dataset, which means 146 records (14.6%) are dropped. As the proportion of rows with missing values is not small enough to be removed, data imputation seems to be the better option to handle missing values.

The data preprocessing is now moved to Talend Data Preparation (TDP) as TDP provide an easier yet robust way to process the data.

After importing the dataset to TDP, the data health bar for every column is inspected. The first step is to change the data type of CustomerID to Integer from Phone Number to eliminate the invalid data issues questioned by the tool. Note that TDP accepts all CustomerID as integers, unlike TDI, indicating the CustomerID values from the original dataset might not have any issues with the data type.

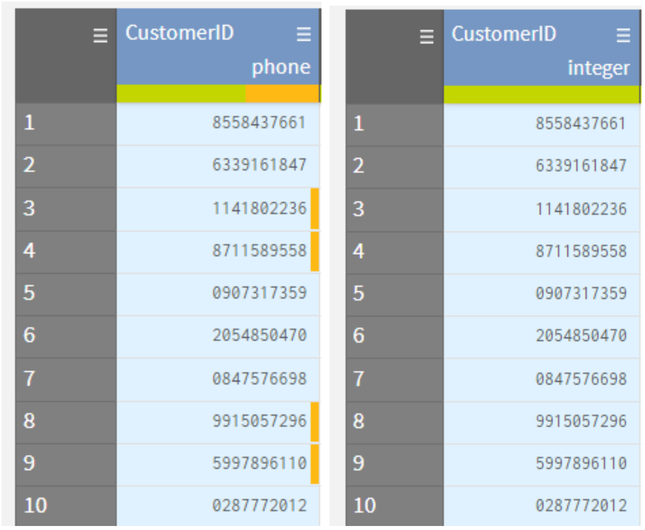


Figure 8: Data health increased to perfect green after changing the data type of CustomerID.

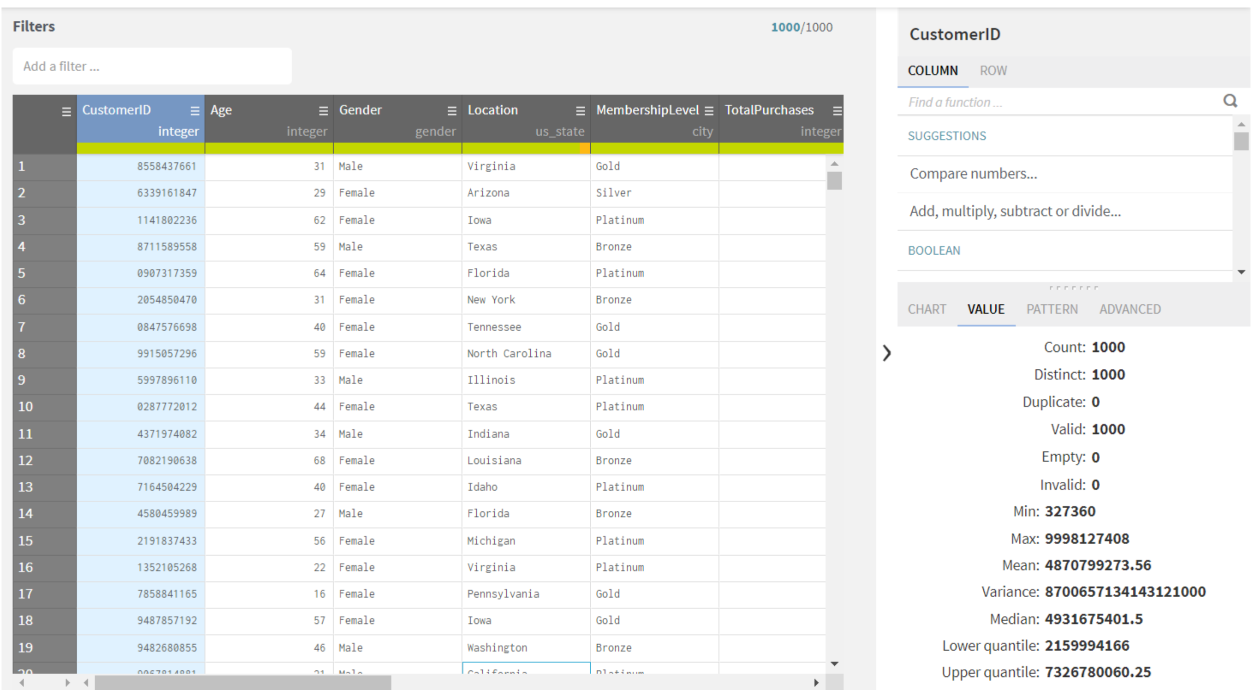


Figure 9: The right plane shows there is no duplicate CustomerID.

As shown in the output of Figure 9, as the duplicate check is performed, no duplicates are occurring at the CustomerID column which every row should have unique values. Hence, there is no action taken for the removal of duplicate data.

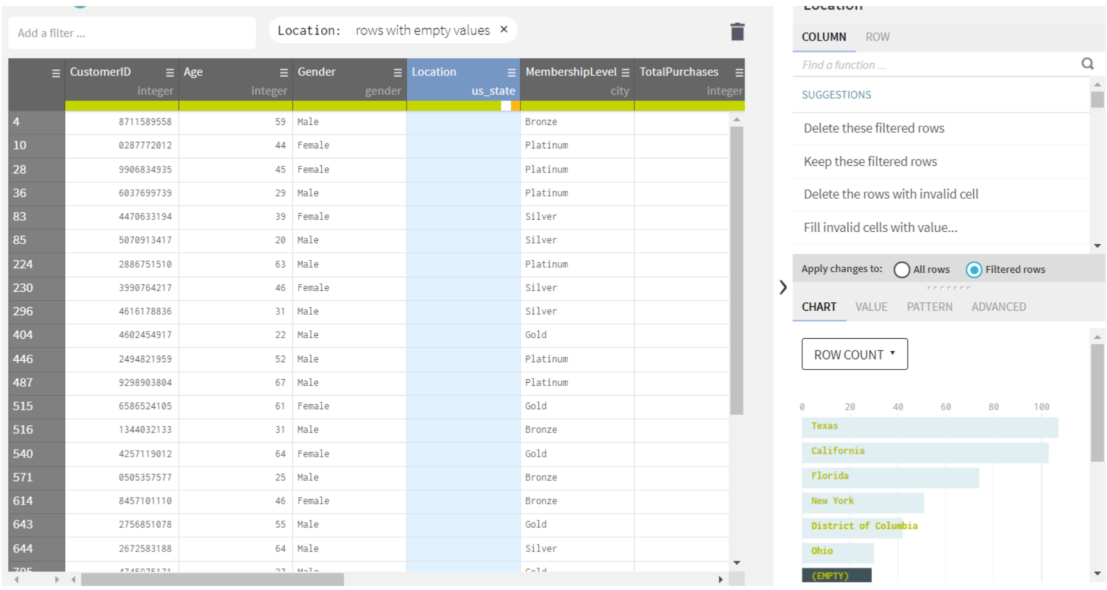


Figure 10: Handling missing values for Location.

Referring to Figure 10, there are missing values in location as designated during data generation. For this column, we select the state with the highest occurrence to perform data imputation for missing locations. The way to impute the data is shown in the left pane of Figure 8.

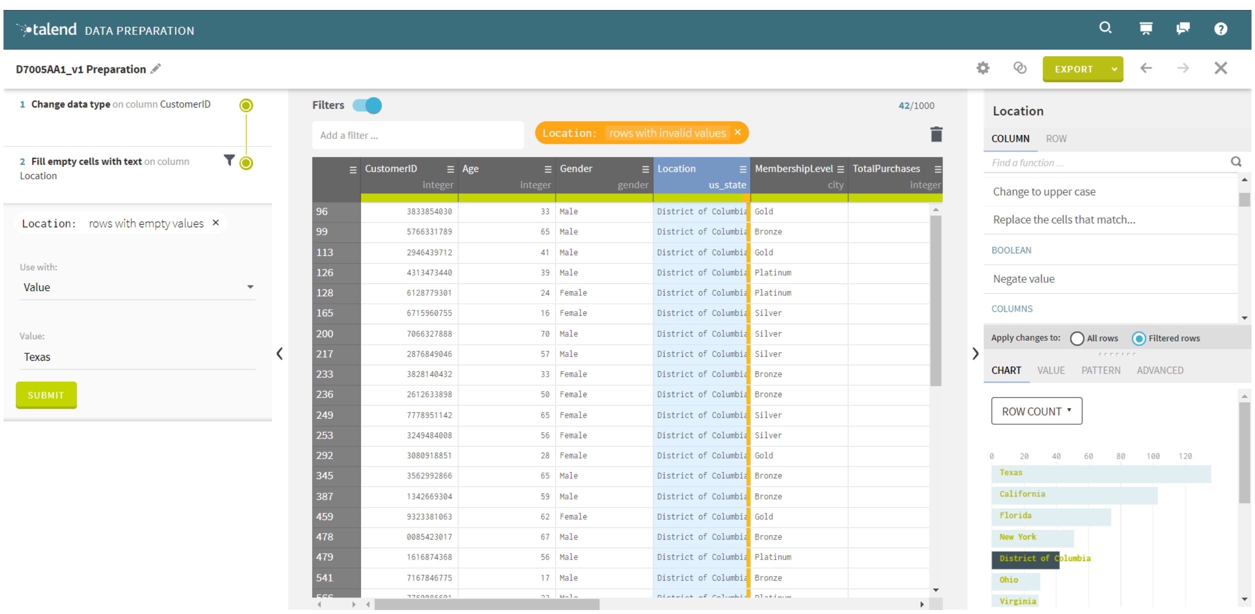


Figure 11: Location imputation and potential invalid location.

Besides missing values, there exist invalid values too. However, these values are the same, which is the District of Columbia, which is not a US state, but a federal district of the United States which not belong to any state. Hence, it is safe to ignore this flag as it is still an area of the United States.

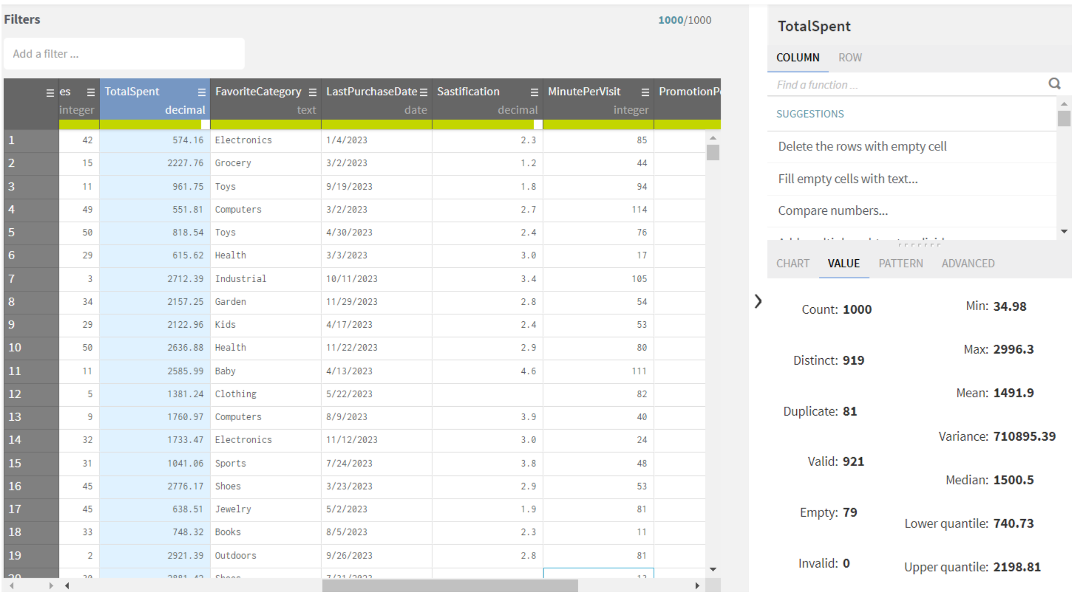


Figure 12: Deciding the imputation method of missing TotalSpent values.

For the TotalSpent, since the range of the value is large (34.98 to 2996.3), it is safer to select the median as the imputation of missing values to minimise the effect and bias caused by the potential outliers although the boxplot does not significantly skew to either side.

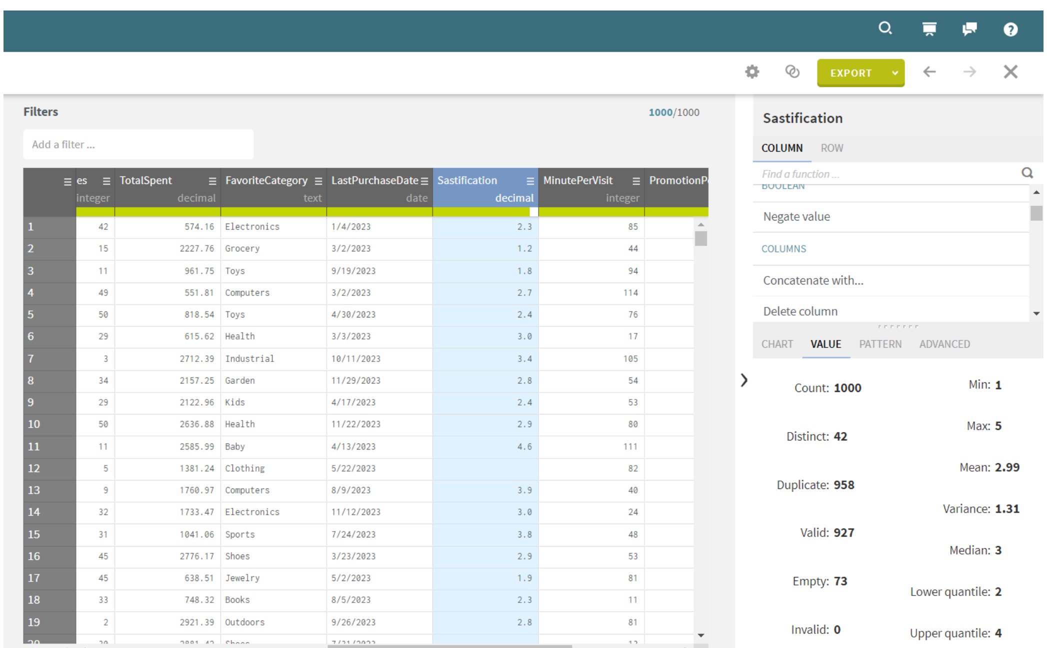


Figure 13: Deciding the imputation method of missing Satisfaction values.

Regarding missing values of Satisfaction, the mean and median values are nearly identical which indicates that the distribution of values is possibly evenly distributed and both values are located at the middle of the range. Hence, the mean can be selected to be imputed. However, to ensure data formalisation which allows only one decimal place for input, the mean value of 2.99 is rounded off to 3.0 for imputation of missing values.

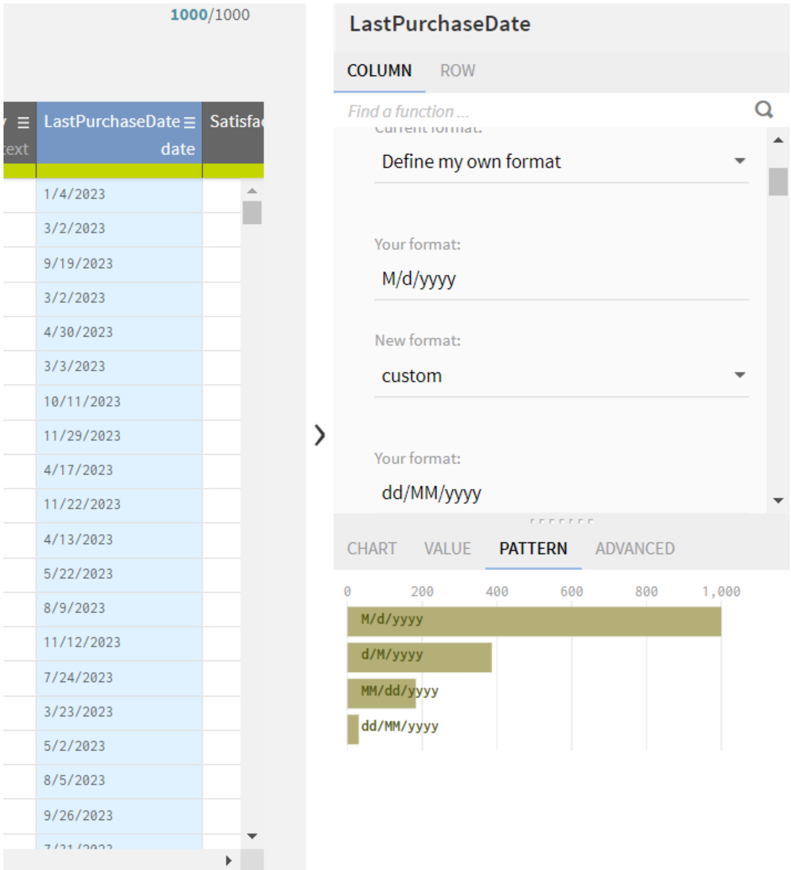


Figure 14: Transforming data format.

As this is a US-related dataset, the date format is recorded in American style. To change the date format to a format that is used more frequently in Malaysia, choose the "Change date format" at the right pane of the interface and apply the setting in Figure 14.

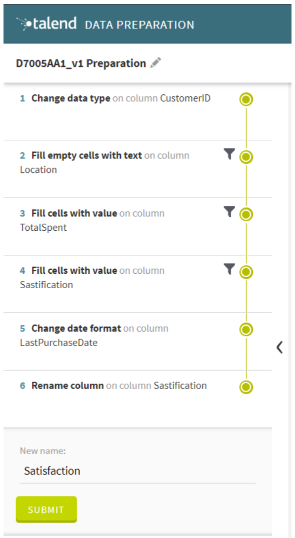


Figure 15: Data processing workflow on Talend Data Preparation.

Finally, rename the unsatisfied column name (for example, wrong spelling) during the final check. As the data quality of all columns is satisfied, the processed dataset is exported into CSV format to the local environment as shown in Figure 13.

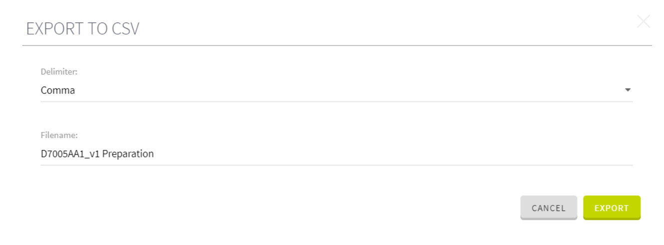


Figure 16: Exporting processed dataset.

Modelling using SAS Enterprise Miner

After a new diagram in a new project is created, the File Import node is placed into the diagram. This node is to import the processed dataset exported from the Talend Data Preparation.

Before that, open the CSV file with Notepad, notice that all values are covered with quotation marks. Hence, we can remove the quotation marks by using the Replace function as shown in Figure 17 below. This step is to ensure the data can be identified as Interval level when setting the schema of the dataset later in Figure 18.

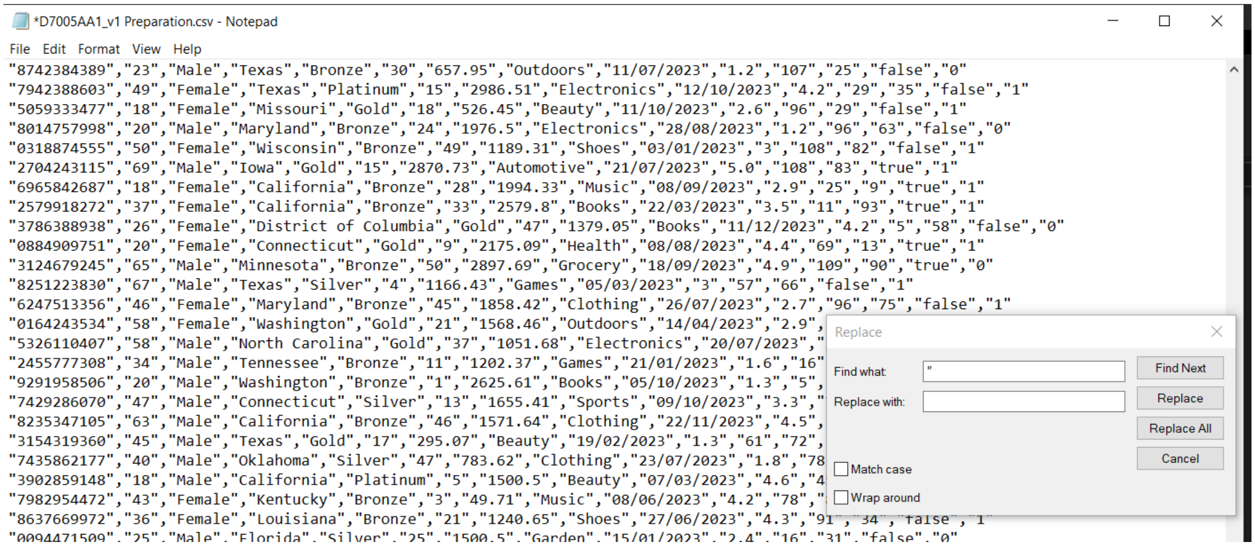
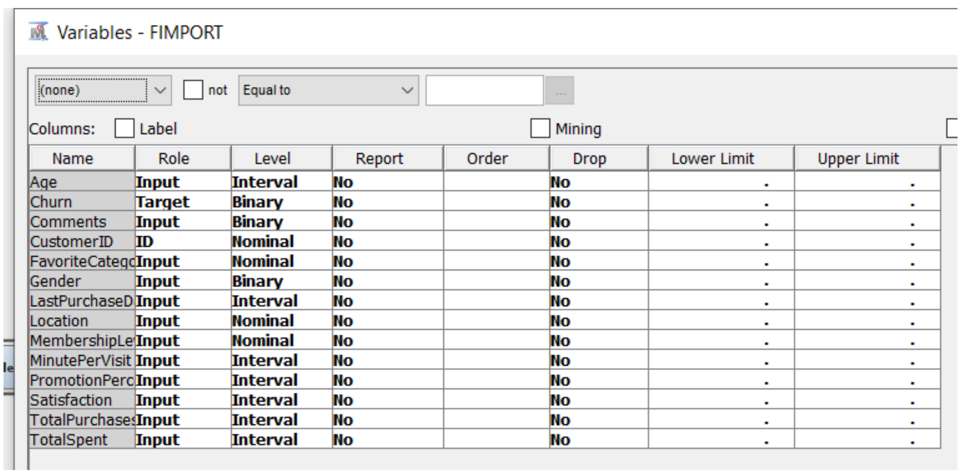


Figure 17: Removing quotation marks in NotePad.

The roles of each variable can be assigned during or after data import by selecting “Edit Variables” on the File Import Node. The setting for each variable is shown in Figure 18.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Churn	Target	Binary	No		No	.	.
Comments	Input	Binary	No		No	.	.
CustomerID	ID	Nominal	No		No	.	.
FavoriteCategc	Input	Nominal	No		No	.	.
Gender	Input	Binary	No		No	.	.
LastPurchaseD	Input	Interval	No		No	.	.
Location	Input	Nominal	No		No	.	.
MembershipLe	Input	Nominal	No		No	.	.
MinutePerVisit	Input	Interval	No		No	.	.
PromotionPerc	Input	Interval	No		No	.	.
Satisfaction	Input	Interval	No		No	.	.
TotalPurchase	Input	Interval	No		No	.	.
TotalSpent	Input	Interval	No		No	.	.

Figure 18: Data Schema settings.

The most important setting in Figure 18 is to set “Churn” as the target variable and make “CustomerID” as ID type to avoid the inclusion of this column during modelling. Next is data partitioning. A “Data Partition” node can added to the diagram and linked with “File Import”. In this study, the partition of data is set as 70% testing and 30% visualization.

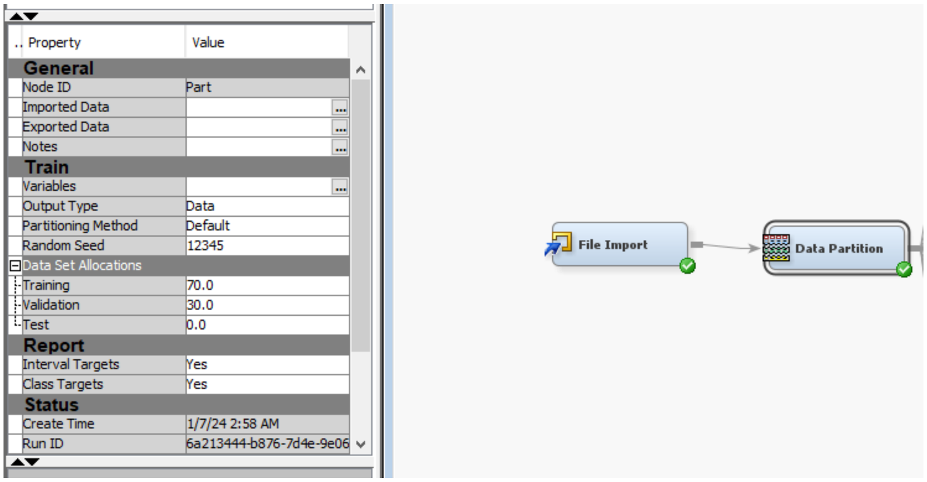


Figure 19: Setting up data partition.

Now, the data is ready to undergo modelling. In this study, decision tree-related models are used to model customer behaviour. Various decision tree models are constructed based on step-by-step tuning. The first two models are constructed as below:

Model Name	Nominal / Ordinal Criterion	Max Branch	Max Depth	Min Leaf Size	Validation Misclass Rate	Training Misclass Rate
DT1	Entropy	2	6	5	0.475	0.352
DT2	Gini	2	6	5	0.449	0.336

The misclassification results can be view via construction and running of Model Comparison node with the node linked to the model node. DT1 and DT2 are using fully default parameters except the nominal and ordinal target criterion, which is a part of parameter tuning. Observing the misclassification rate for validation, both models seem to have a large room for improvement available. As there are 51 locations and multiple options in FavouriteCategory available, the maximum branch is now increased to 3 as did for models DT3 and DT4.

Model Name	Nominal / Ordinal Criterion	Max Branch	Max Depth	Min Leaf Size	Validation Misclass Rate	Training Misclass Rate
DT3	Entropy	3	6	5	0.385	0.250
DT4	Gini	3	6	5	0.422	0.332

Compared to model DT1, DT3 with extra maximum branch brings a huge improvement in reducing misclassification rate for both validation and training set. For the Gini tune, DT4 does improve in classification rate although it is not as large as the Entropy tune. Now, the maximum allowed branches is increasing to 4 to observe what will happen.

Model Name	Nominal / Ordinal Criterion	Max Branch	Max Depth	Min Leaf Size	Validation Misclass Rate	Training Misclass Rate
DT5	Entropy	4	6	5	0.445	0.360
DT6	Gini	4	6	5	0.425	0.299

DT5 suffering a huge drop in performance of validation set classification tasks when compared to DT3 which has the same Entropy tune. The Gini-tuned DT6 is better than DT5 but not DT4. This situation indicates that is overfitting. Next, the maximum depth is adjusted based on models DT3 and DT4. The maximum depth is increasing by 1.

Model Name	Nominal / Ordinal Criterion	Max Branch	Max Depth	Min Leaf Size	Validation Misclass Rate	Training Misclass Rate
DT3	Entropy	3	6	5	0.385	0.250
DT4	Gini	3	6	5	0.422	0.332
DT7	Entropy	3	7	5	0.385	0.250
DT8	Gini	3	7	5	0.418	0.332

Increasing maximum depth do not improve the classification rate for both tuned criterion. That means increasing the maximum depth to 7 is causing overfitting. The DT3 and DT4 are still selected as the best models for each tune. Now, the maximum depth is decreased to 5 to observe whether this action can reduce the misclassification rate for validation set.

Model Name	Nominal / Ordinal Criterion	Max Branch	Max Depth	Min Leaf Size	Validation Misclass Rate	Training Misclass Rate
DT3	Entropy	3	6	5	0.385	0.250
DT4	Gini	3	6	5	0.422	0.332
DT9	Entropy	3	5	5	0.408	0.316

DT10	Gini	3	5	5	0.425	0.346

Both reduced maximum depth models (DT9 and DT10) are performing worse than before with increasing of misclassification rate for both validation and training set, indicating occurrence of underfitting. Hence the maximum branch of six is the best depth for this model. The next tuning is based on minimum leaf size.

Model Name	Nominal / Ordinal Criterion	Max Branch	Max Depth	Min Leaf Size	Validation Misclass Rate	Training Misclass Rate
DT3	Entropy	3	6	5	0.385	0.250
DT4	Gini	3	6	5	0.422	0.332
DT11	Entropy	3	6	4	0.365	0.306
DT12	Gini	3	6	4	0.412	0.322

Both reduced minimum leaf size models (DT11 and DT12) are improving the performance in reducing misclassification rate for validation rate. For now, DT11 is the best Entropy-tuned model while DT12 is the best Gini-tuned model. Next, the minimum leaf size is further reduced to 3.

Model Name	Nominal / Ordinal Criterion	Max Branch	Max Depth	Min Leaf Size	Validation Misclass Rate	Training Misclass Rate
DT11	Entropy	3	6	4	0.365	0.306
DT12	Gini	3	6	4	0.412	0.322
DT13	Entropy	3	6	3	0.372	0.231
DT14	Gini	3	6	3	0.412	0.340

Both DT13 and DT14 does not further improve the classification result of validation set. For DT13, the increasing difference between the validation and training set misclassification rate indicates the model is overfitting at this stage. There is no further decision tree tuning so the model DT11 with Entropy-tuned, maximum 3 branches, maximum 6 layer depth and minimum leaf size of 4 is the best decision tree model for this study. The overall performance ranking is tabulated in the output of Figure 20.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate
Y	Tree11	Tree11	DT11 (E B3 D6 L4)	Churn		0.365449	699	0.306152
	Tree13	Tree13	DT13 (E B3 D6 L3)	Churn		0.372093	699	0.23176
	Tree3	Tree3	DT3 (E B3 D6)	Churn		0.385382	699	0.250358
	Tree7	Tree7	DT7 (E B3 D7)	Churn		0.385382	699	0.250358
	Tree9	Tree9	DT9 (E B3 D5)	Churn		0.408638	699	0.316166
	Tree14	Tree14	DT14 (E B3 D6 L3)	Churn		0.41196	699	0.340486
	Tree12	Tree12	DT12 (G B3 D6 L4)	Churn		0.41196	699	0.321888
	Tree8	Tree8	DT8 (G B3 D7)	Churn		0.418605	699	0.331903
	Tree4	Tree4	DT4 (G B3 D6)	Churn		0.421927	699	0.331903
	Tree10	Tree10	DT10 (G B3 D5)	Churn		0.425249	699	0.346209
	Tree6	Tree6	DT6 (G B4 D6)	Churn		0.425249	699	0.360515
	Tree5	Tree5	DT5 (E B4 D6)	Churn		0.445183	699	0.298999
	Tree2	Tree2	DT2 (G B2 D6)	Churn		0.448505	699	0.336195
	Tree	Tree	DT1 (E B2 D6)	Churn		0.475083	699	0.351931

Figure 20: Overall decision tree model ranking.

Notice that the best Gini-tuned model (DT12) is only can be ranked six out of fourteen models, indicating that Gini-tuning is not quite suitable in this study.

Several models are included in the modelling such as Bagging Ensemble Method (Random Forest), Gradient Boosting HP Tree and HP Forest. The Ensemble model 1 is the ensemble of the Top 2 best decision tree models (DT11 and DT13) while the Ensemble model 2 is the ensemble of the best Entrophy-tuned model (DT11) and the best Gini-tuned model (DT12). Both HP environment tree and forest and the Gradient Boosting use the default parameters. The overall workflow is shown in Figure 21 and the overall results are in the output of Figure 22.

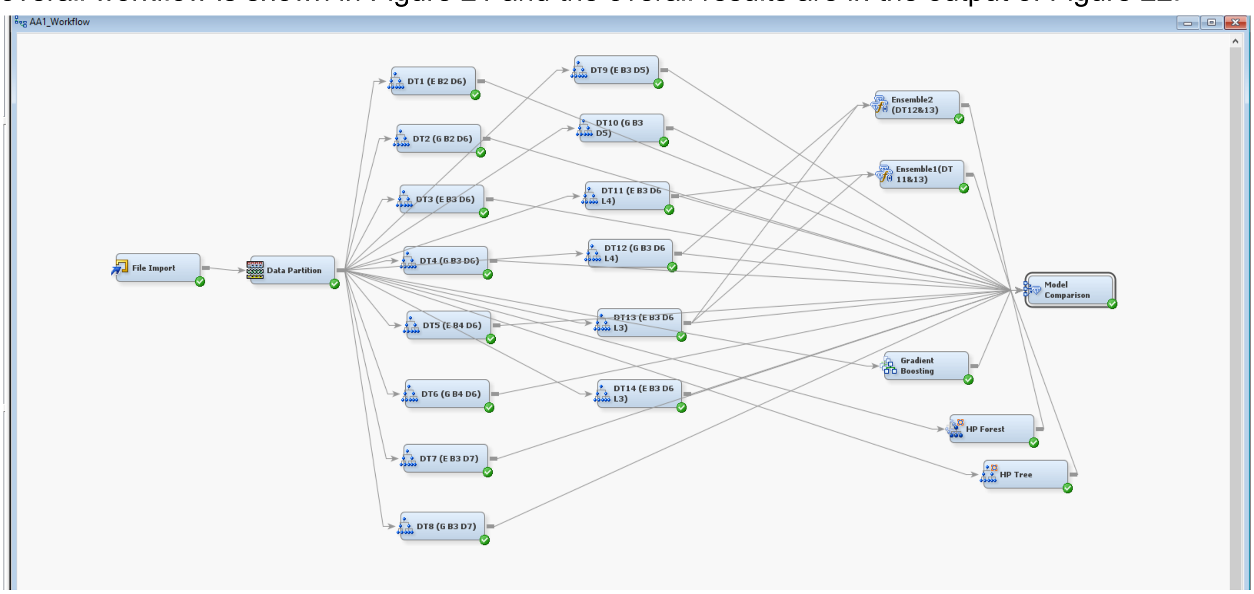
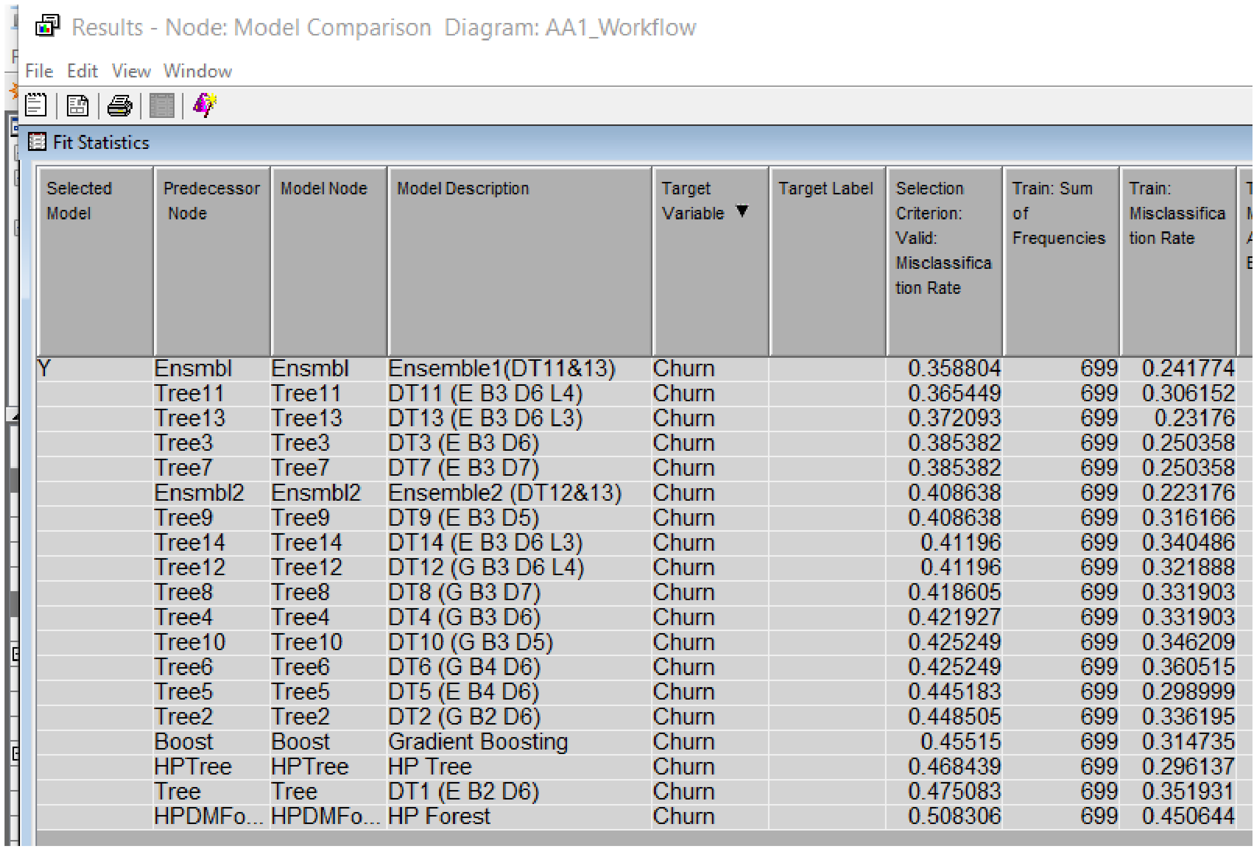


Figure 21: Overall Modelling Diagram in SAS EM.


Results - Node: Model Comparison Diagram: AA1_Workflow

File Edit View Window

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate
Y	Ensmbl	Ensmbl	Ensemble1(DT11&13)	Churn		0.358804	699	0.241774
	Tree11	Tree11	DT11 (E B3 D6 L4)	Churn		0.365449	699	0.306152
	Tree13	Tree13	DT13 (E B3 D6 L3)	Churn		0.372093	699	0.23176
	Tree3	Tree3	DT3 (E B3 D6)	Churn		0.385382	699	0.250358
	Tree7	Tree7	DT7 (E B3 D7)	Churn		0.385382	699	0.250358
	Ensmbl2	Ensmbl2	Ensemble2 (DT12&13)	Churn		0.408638	699	0.223176
	Tree9	Tree9	DT9 (E B3 D5)	Churn		0.408638	699	0.316166
	Tree14	Tree14	DT14 (E B3 D6 L3)	Churn		0.41196	699	0.340486
	Tree12	Tree12	DT12 (G B3 D6 L4)	Churn		0.41196	699	0.321888
	Tree8	Tree8	DT8 (G B3 D7)	Churn		0.418605	699	0.331903
	Tree4	Tree4	DT4 (G B3 D6)	Churn		0.421927	699	0.331903
	Tree10	Tree10	DT10 (G B3 D5)	Churn		0.425249	699	0.346209
	Tree6	Tree6	DT6 (G B4 D6)	Churn		0.425249	699	0.360515
	Tree5	Tree5	DT5 (E B4 D6)	Churn		0.445183	699	0.298999
	Tree2	Tree2	DT2 (G B2 D6)	Churn		0.448505	699	0.336195
	Boost	Boost	Gradient Boosting	Churn		0.45515	699	0.314735
	HPTree	HPTree	HP Tree	Churn		0.468439	699	0.296137
	Tree	Tree	DT1 (E B2 D6)	Churn		0.475083	699	0.351931
	HPDMFo...	HPDMFo...	HP Forest	Churn		0.508306	699	0.450644

Figure 22: Overall modelling result.

From the output in Figure 22, all three models with the default setting (Gradient Boosting, HP Tree and HP Forest) are among the weakest models in classification. Future tuning with better understanding is recommended so that the fined-tuned models have the opportunity to surpass the best decision tree model. The ensemble model of DT11 and DT13 is the best-performed model in validation set churn classification. Although the bagged ensemble model is the best model, for easier interpretation, the second best model which is DT11 is used to generate insight regarding customer behaviour.

Although the misclassification rate is 36% which is considered high, by observing the decision tree diagram, the Location and the favoured Content are the top two layers for every branch. However, by observing the variable importance in Figure 23, the most important variable for validation is age. Hence, I would like to suggest the e-commerce owner put more focus on location, age and favoured category as these factors are likely to influence the churn of the customer.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
FavoriteCategory		3	1.0000	0.0000	0.0000
Satisfaction		4	0.8091	0.0000	0.0000
TotalSpent		3	0.7997	0.7545	0.9435
MinutePerVisit		4	0.7112	0.7278	1.0232
Location		1	0.7105	0.0000	0.0000
Age		3	0.6137	1.0000	1.6295
TotalPurchases		1	0.4191	0.0000	0.0000
LastPurchaseDate		1	0.3999	0.0000	0.0000
PromotionPercent		1	0.3548	0.2862	0.8068

Figure 23: Result output of model DT11.

Summary and Difficulties

The random forest which is the bagged ensemble method of the decision trees is the best type of model for classifying the churn of customers. More detailed insights are subject to be discovered as at the moment there is no much analysis that can be explained due to a quite high misclassification rate for the best model available. Other models other than decision trees are advised to be applied in this study. The high rate of misclassification may be due to the data is generated highly randomly with limited limitations applied. Besides, data from 50 United States and 1 federal district (District of Columbia) is insufficient for 1000 rows, which increases the difficulty of insight generations. With the existence of this limitation, the extension of this study can include the tuning of HP environment model and Gradient Boosting in attempt of generate more valuable insights from this dataset.