ALTERNATIVE ASSESSMENT 1

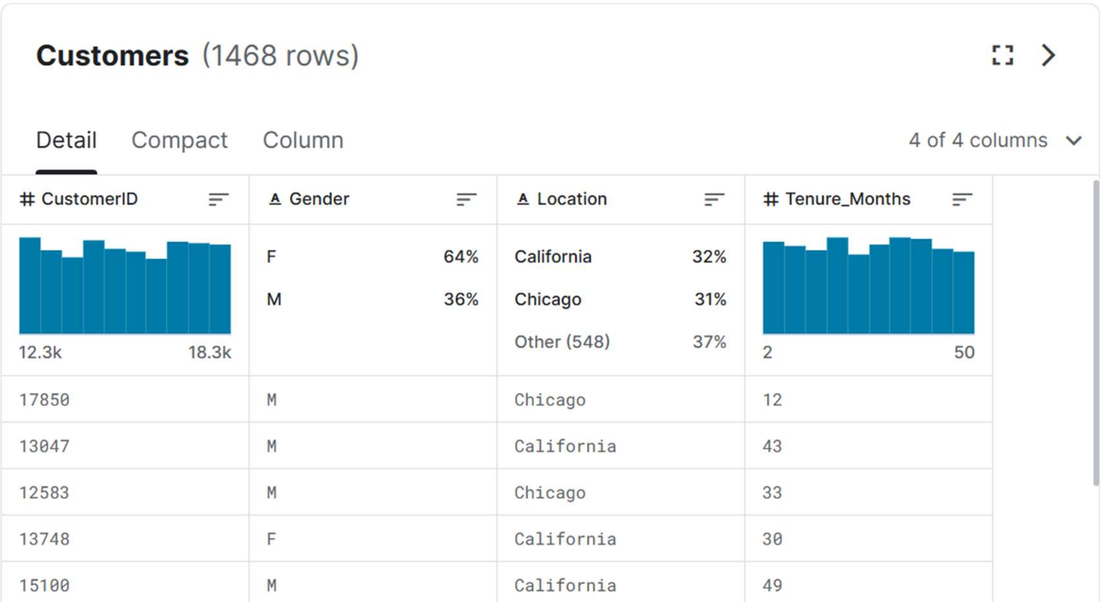
SIYU JIANG 22060253

1. Dataset Description

I found two datasets of customer transactions from an e-commerce company on Kaggle: CustomersData.xlsx and Online_Sales.csv.

[Marketing Insights for E-Commerce Company (kaggle.com)]

CustomersData.xlsx has 4 columns: CustomerID, Gender, Location, and Tenure_Months, and 1468 customers' information. The details are as follows:



Online_Sales.csv contains actual orders data (point of Sales data) at transaction level with below variables.

CustomerID: Customer unique ID

Transaction_ID: Transaction Unique ID

Transaction_Date: Date of Transaction

Product_SKU: SKU ID – Unique Id for product

Product_Description: Product Description

Product_Cateogry: Product Category

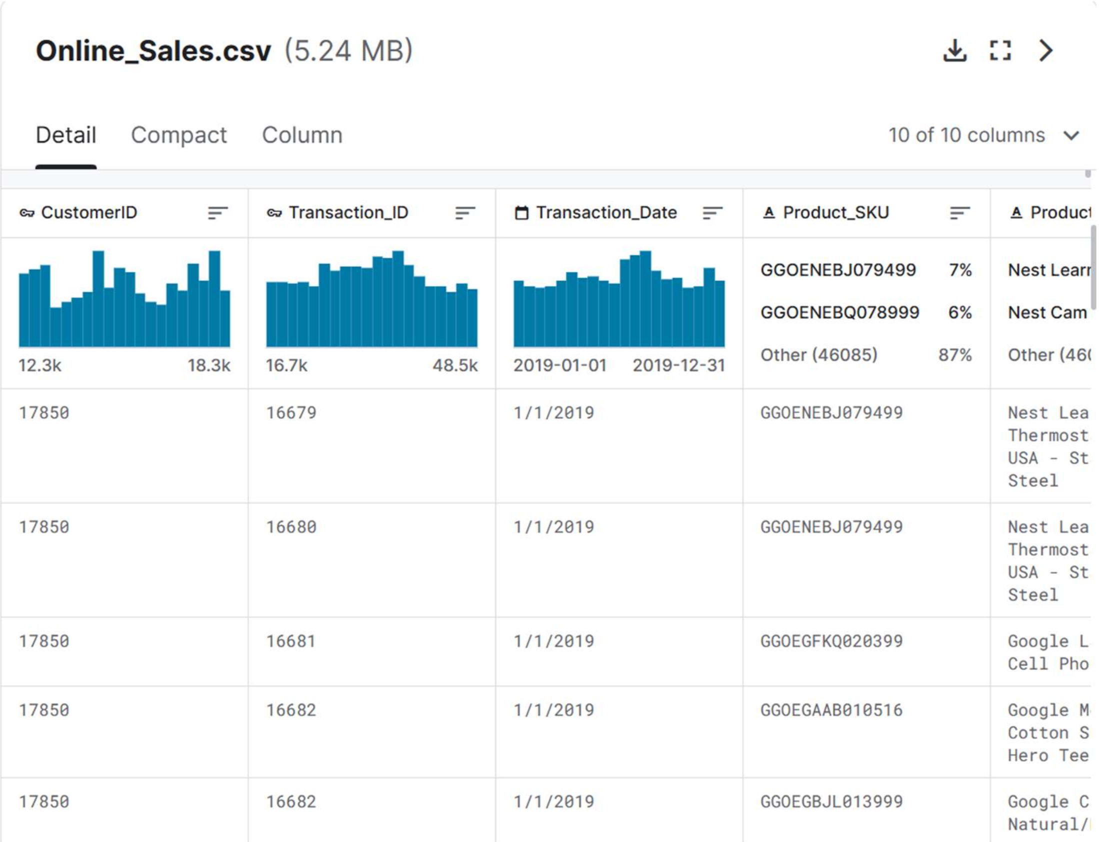
Quantity: Number of items ordered

Avg_Price: Price per one quantity

Delivery_Charges: Charges for delivery

Coupon_Status: Any discount coupon applied

This dataset has 52924 rows, the details are as follows:



Online_Sales.csv (5.24 MB)

Detail Compact Column 10 of 10 columns

CustomerID	Transaction_ID	Transaction_Date	Product_SKU	Product
12.3k 18.3k	16.7k 48.5k	2019-01-01 2019-12-31	GGOENEBJ079499 7% GGOENEBQ078999 6% Other (46085) 87%	Nest Learr Nest Cam Other (46(
17850	16679	1/1/2019	GGOENEBJ079499	Nest Lea Thermost USA - St Steel
17850	16680	1/1/2019	GGOENEBJ079499	Nest Lea Thermost USA - St Steel
17850	16681	1/1/2019	GGOEGFKQ020399	Google L Cell Pho
17850	16682	1/1/2019	GGOEGAAB010516	Google M Cotton S Hero Tee
17850	16682	1/1/2019	GGOEGBJL013999	Google C Natural/

2. Objectives

Customer value is the total value that a customer contributes to the company in the whole life cycle of its interaction with the company. Customer value is usually related to the number of purchases a customer makes, the purchase amount, tenure months, etc. And it is a common practice to assess the value of a customer by the amount he spends in a year.

Based on this business metric, the main objective is to analyze the total spent by each customer in 2019, which is divided into 3 tasks:

- To calculate total spent by each customer in 2019, and bin the data into different categories based on the distribution of total spent range in 1468 customers.
- Build classification model of customers' total spent range.
- Evaluate the result and compare the performance of different algorithms.

3. Data Preprocessing Using Talend Data Preparation

The original Online_Sales.csv dataset has 52924 rows while Talend Data Preparation can only access 30000 rows at most, so I split the dataset into two files, one with 30000 rows and the other with 22924 rows, as shown below:

talend DATA PREPARATION

Online_Sales Preparation (1)

Filters | 30000/30000

Add a filter ...

	CustomerID fr_postal_code	Transaction_ID fr_postal_code	Transaction_Date date	Product_SKU text	Product_Descrip... text	Product_Category text	Quantity integer	Avg_Price decimal	Delivery_Charges decimal	Coupon_Status text
1	17850	16679	1/1/2019	GGOENEBJ079499	Nest Learning Thermo	Nest-USA	1	153.71	6.5	Used
2	17850	16680	1/1/2019	GGOENEBJ079499	Nest Learning Thermo	Nest-USA	1	153.71	6.5	Used
3	17850	16681	1/1/2019	GGOEGFKQ020399	Google Laptop and Ce	Office	1	2.05	6.5	Used
4	17850	16682	1/1/2019	GGOEGAAB010516	Google Men's 100% Co	Apparel	5	17.53	6.5	Not Used
5	17850	16682	1/1/2019	GGOEGBJL013999	Google Canvas Tote N	Bags	1	16.5	6.5	Used
6	17850	16682	1/1/2019	GGOEGBMJ013399	Sport Bag	Bags	15	5.15	6.5	Used
7	17850	16682	1/1/2019	GGOEGDHC018299	Google 22 oz Water B	Drinkware	15	3.08	6.5	Not Used
8	17850	16682	1/1/2019	GGOEGDHG014499	Google Infuser-Top W	Drinkware	15	10.31	6.5	Clicked
9	17850	16682	1/1/2019	GGOEGDWC020199	Engraved Ceramic Goo	Drinkware	5	9.27	6.5	Used
10	13047	16682	1/1/2019	GGOEGGOA017399	Maze Pen	Office	52	0.98	6.5	Used
11	13047	16682	1/1/2019	GGOEGOFH020299	Galaxy Screen Cleani	Office	31	1.99	6.5	Clicked
12	13047	16682	1/1/2019	GGOEGOXQ016399	Badge Holder	Office	31	1.99	6.5	Clicked
13	13047	16682	1/1/2019	GGOEYAAB031816	YouTube Men's Short	Apparel	5	17.53	6.5	Used
14	13047	16684	1/1/2019	GGOENEBQ078999	Nest Cam Outdoor Sec	Nest-USA	2	122.77	6.5	Clicked
15	13047	16684	1/1/2019	GGOENEBQ079199	Nest Protect Smoke +	Nest-USA	1	81.5	6.5	Used
16	13047	16685	1/1/2019	GGOEGAAR010714	Google Men's 100% Co	Apparel	1	14.02	6.5	Used
17	13047	16685	1/1/2019	GGOEGAEQ027913	Google Women's Short	Apparel	1	14.02	6.5	Clicked
18	13047	16685	1/1/2019	GGOEGDWR015799	Red Shine 15 oz Mug	Drinkware	1	10.72	6.5	Not Used
19	13047	16687	1/1/2019	GGOEGFQB013799	Compact Selfie Stick	Lifestyle	1	9.27	6.5	Clicked
20	13047	16687	1/1/2019	GGOEGGOA017399	Maze Pen	Office	3	1.02	6.5	Used
21	13047	16687	1/1/2019	GGOEGOAQ012899	Ballpoint LED Light	Office	1	2.58	6.5	Not Used
22	13047	16687	1/1/2019	GGOEGOAR021999	Color Changing Grip	Office	3	1.55	6.5	Clicked
23	13047	16687	1/1/2019	GGOEGOBG023599	Colored Pencil Set	Office	1	3.08	6.5	Used
24	13047	16687	1/1/2019	GGOEGOLC013299	Spiral Notebook and	Office	1	6.18	6.5	Clicked
25	13047	16688	1/1/2019	GGOENEBB078899	Nest Cam Indoor Secu	Nest-USA	1	122.77	6.5	Used
26	13047	16689	1/1/2019	GGOENEBJ079499	Nest Learning Thermo	Nest-USA	1	153.71	6.5	Used
27	12583	16692	1/1/2019	GGOEAFKQ020599	Android Sticker Shee	Office	1	2.47	102.79	Used
28	12583	16692	1/1/2019	GGOEGDHC015299	23 oz Wide Mouth Spo	Drinkware	26	8.72	102.79	Clicked
29	12583	16692	1/1/2019	GGOEGFKQ020399	Google Laptop and Ce	Office	1	1.64	102.79	Clicked
30	12583	16692	1/1/2019	GGOEYFKQ020699	YouTube Custom Decal	Office	1	1.64	102.79	Clicked
31	12583	16692	1/1/2019	GGOEYOCR078099	YouTube Spiral Journ	Notebooks & Journals	26	7.93	102.79	Not Used
32	12583	16693	1/1/2019	GGOEGATB060415	Google Women's Quilt	Apparel	1	61.89	6.5	Clicked

talend DATA PREPARATION

Online_Sales_2 Preparation (1)

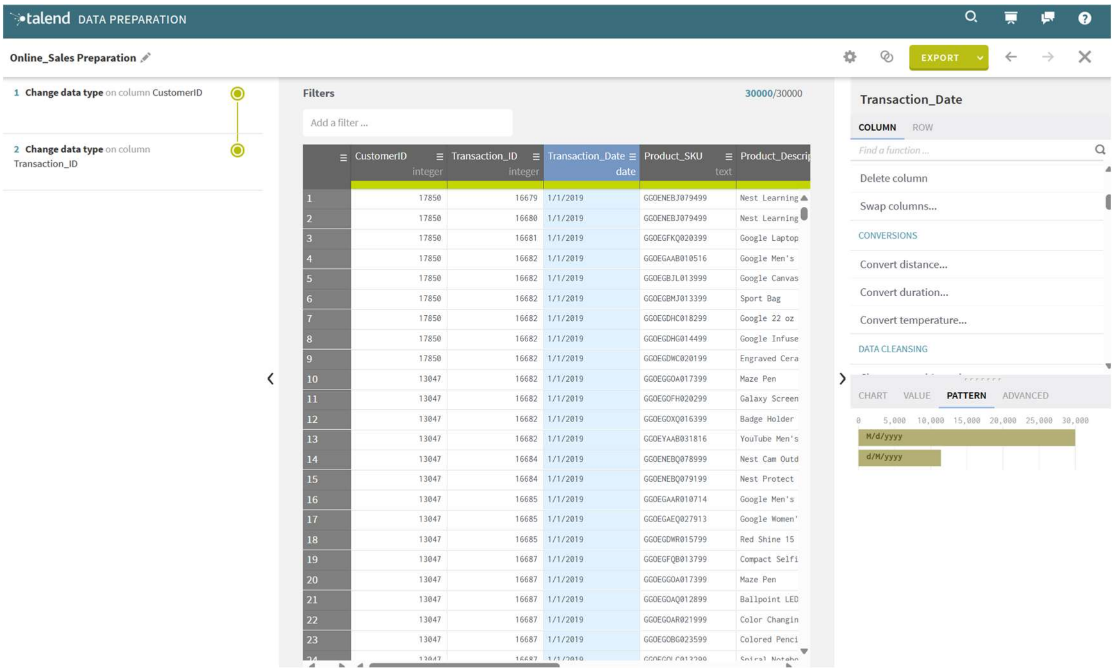
Filters | 22924/22924

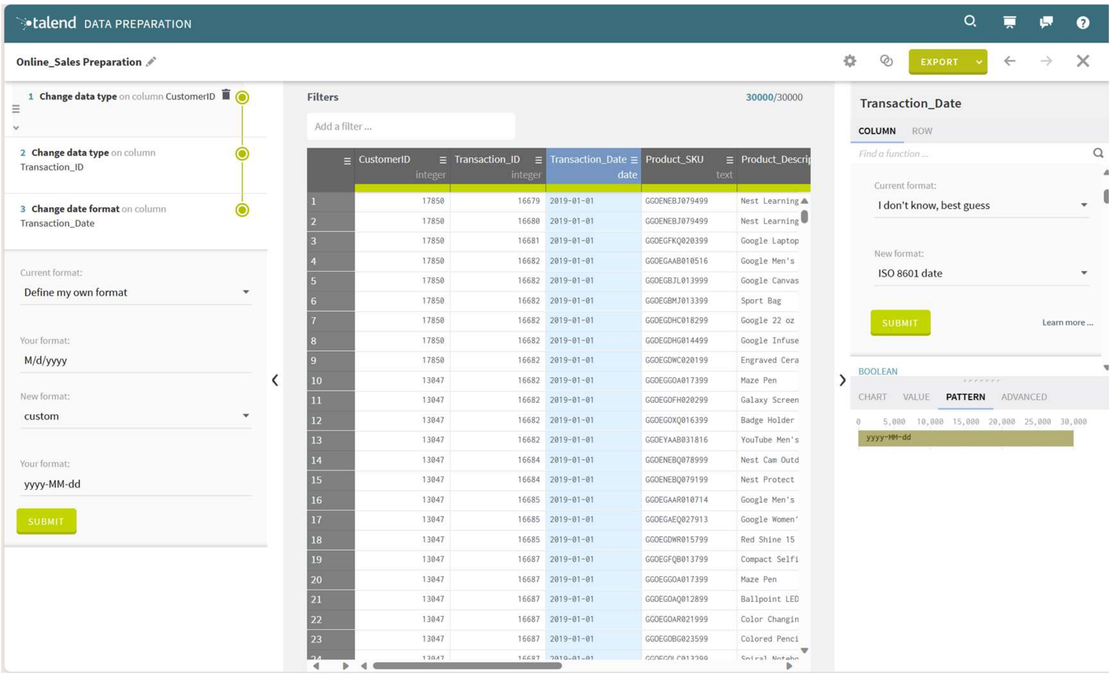
Add a filter ...

	CustomerID fr_postal_code	Transaction_ID fr_postal_code	Transaction_Date date	Product_SKU text	Product_Descrip... text	Product_Category text	Quantity integer	Avg_Price decimal	Delivery_Charges integer	Coupon_Status text
1	14702	34345	8/1/2019	GGOEGBJL013999	Google Canvas Tote N	Bags	3	15.99	6	Used
2	14702	34345	8/1/2019	GGOEGDHQ015399	26 oz Double Wall In	Drinkware	8	24.99	6	Not Used
3	14702	34345	8/1/2019	GGOEGHPJ080310	Google Blackout Cap	Headgear	8	13.29	6	Not Used
4	14702	34345	8/1/2019	GGOEGOAB021499	Metal Texture Roller	Office	8	8.7	6	Clicked
5	14702	34345	8/1/2019	GGOEGOCB078299	Google Leather Journ	Notebooks & Journals	8	16.5	6	Used
6	14031	34346	8/1/2019	GGOEAAAJ080814	Android Men's Engine	Apparel	1	15.99	6	Used
7	14031	34346	8/1/2019	GGOEAAAQ032013	Android Men's Short	Apparel	2	6.8	6	Used
8	14031	34346	8/1/2019	GGOEGAPB057813	Women's Performance	Apparel	1	67.19	6	Clicked
9	14031	34347	8/1/2019	GGOEAAAP081214	Android Men's Take C	Apparel	1	8	6	Used
10	14031	34347	8/1/2019	GGOEAAEJ035714	Android Men's Vintag	Apparel	1	8.4	6	Not Used
11	14031	34347	8/1/2019	GGOEGAAB010514	Google Men's 100% Co	Apparel	1	13.59	6	Clicked
12	14031	34347	8/1/2019	GGOEGAAL010616	Google Men's 100% Co	Apparel	1	13.59	6	Not Used
13	14031	34347	8/1/2019	GGOEGAAQ010414	Google Men's 100% Co	Apparel	1	13.59	6	Clicked
14	14031	34347	8/1/2019	GGOEGAEJ031315	Google Tri-blend Hoo	Apparel	1	31.99	6	Used
15	14031	34347	8/1/2019	GGOEGAFB035814	Google Men's Zip Ho	Apparel	1	44.79	6	Used
16	13187	34347	8/1/2019	GGOEGALQ034215	Google Women's Vinta	Apparel	1	10.63	6	Clicked
17	13187	34347	8/1/2019	GGOEGBJL013999	Google Canvas Tote N	Bags	1	12.79	6	Used
18	13187	34347	8/1/2019	GGOEGDHC018299	Google 22 oz Water B	Drinkware	1	2.39	6	Used
19	13187	34347	8/1/2019	GGOEGHPJ080310	Google Blackout Cap	Headgear	1	10.63	6	Not Used
20	13187	34348	8/1/2019	GGOEGAAB033815	Google Men's Vintage	Apparel	1	7.6	6	Clicked
21	13187	34348	8/1/2019	GGOEGAEJ028015	Google Women's Short	Apparel	1	4.08	6	Used
22	13187	34348	8/1/2019	GGOEGALB034113	Google Women's Vinta	Apparel	1	4.56	6	Clicked
23	13187	34348	8/1/2019	GGOEGALB034114	Google Women's Vinta	Apparel	1	4.56	6	Used
24	13187	34348	8/1/2019	GGOEGALL074614	Google Women's Short	Apparel	2	6.8	6	Clicked
25	13187	34349	8/1/2019	GGOEAAAQ032013	Android Men's Short	Apparel	1	6.8	12.99	Used
26	13187	34349	8/1/2019	GGOEAAEH035213	Android Men's Vintag	Apparel	1	7.2	12.99	Clicked
27	13187	34349	8/1/2019	GGOEADHH073999	Android 17oz Stainle	Drinkware	1	10.63	12.99	Not Used
28	13187	34349	8/1/2019	GGOEAHPJ004213	Android Stretch Fit	Headgear	1	8.8	12.99	Clicked
29	13187	34349	8/1/2019	GGOEGALB034114	Google Women's Vinta	Apparel	2	4.56	12.99	Clicked
30	13187	34350	8/1/2019	GGOEGALB036516	Google Women's Scoop	Apparel	2	6	6	Not Used
31	13187	34350	8/1/2019	GGOEGALQ036616	Google Women's Scoop	Apparel	2	6	6	Not Used
32	13187	34350	8/1/2019	GGOEGATJ060516	Google Women's Quilt	Apparel	1	37.49	6	Not Used

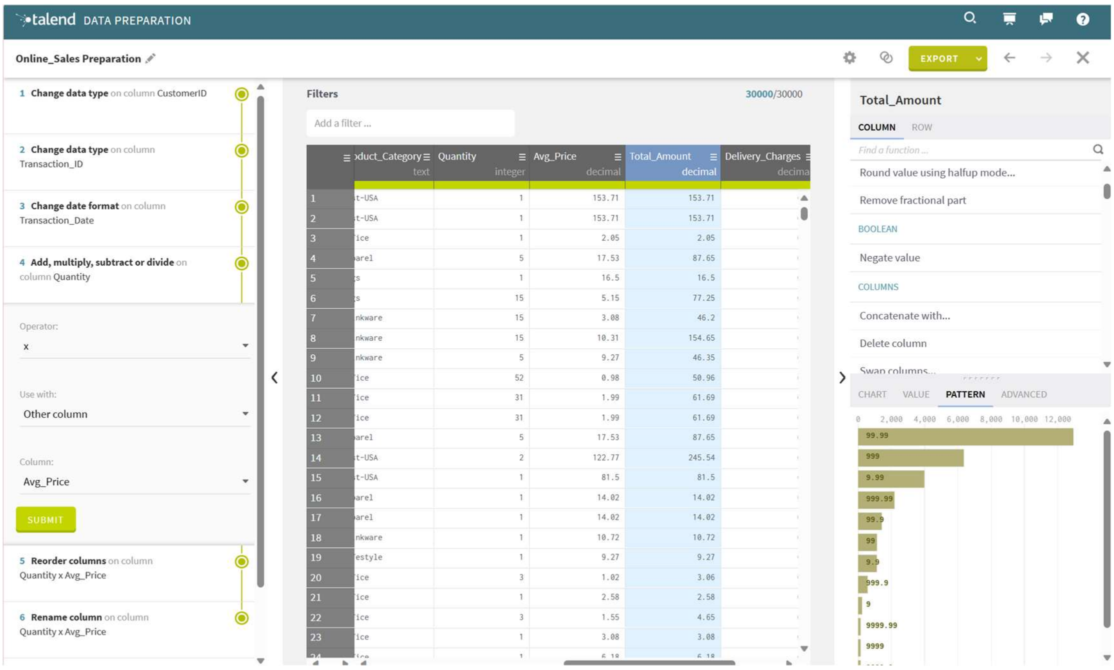
To make the data type more accurate and convenient for subsequent processing, I change the data type of CustomerID, Transaction_ID into integer, and Delivery_Charges into decimal.

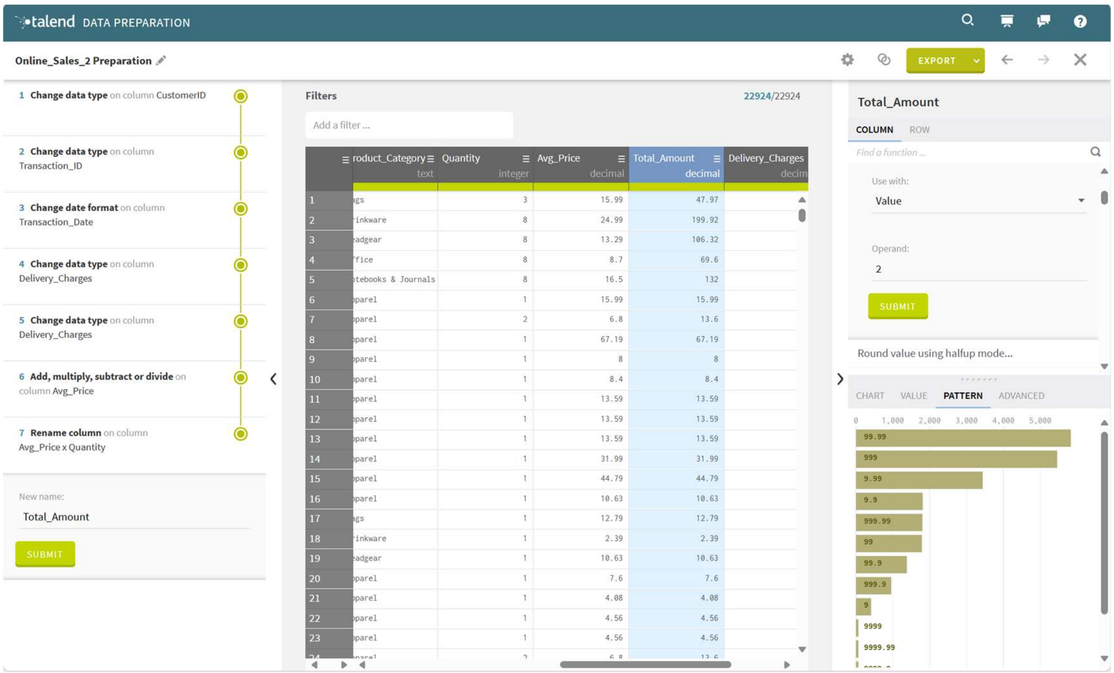
In addition, the format of the Transaction_Date is not uniform, some shows ‘M/d/yyyy’, the others show ‘d/M/yyyy’, so I change the date format into ‘yyyy-MM-dd’ for subsequent processing:





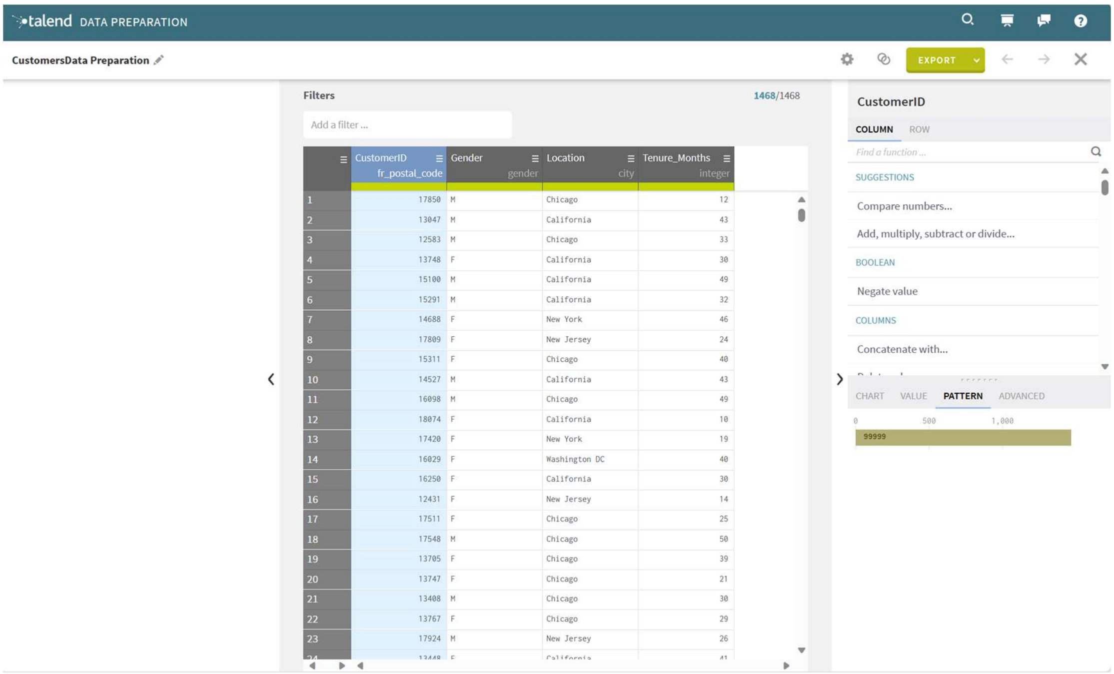
For subsequent analysis, I calculate the ‘Total_Amount’ spent on each transaction using ‘Quantity’ mutyplies ‘Avg_Price’, and rename the new column as ‘Total_Amount’.





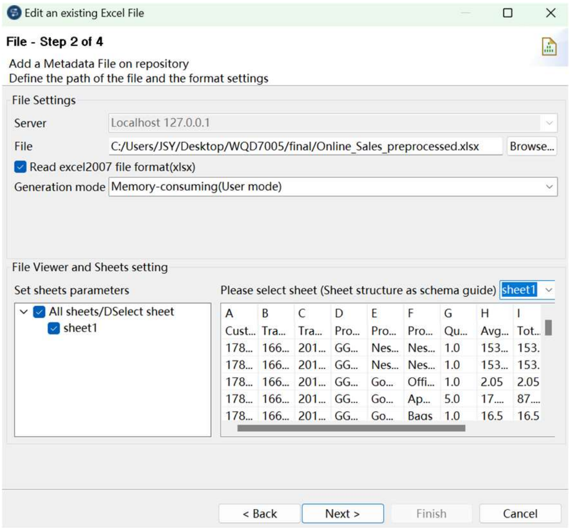
After preprocessing, the two splitted sets of Online_Sales.csv file are cleaned and then combined as Online_Sales_preprocessed.xlsx.

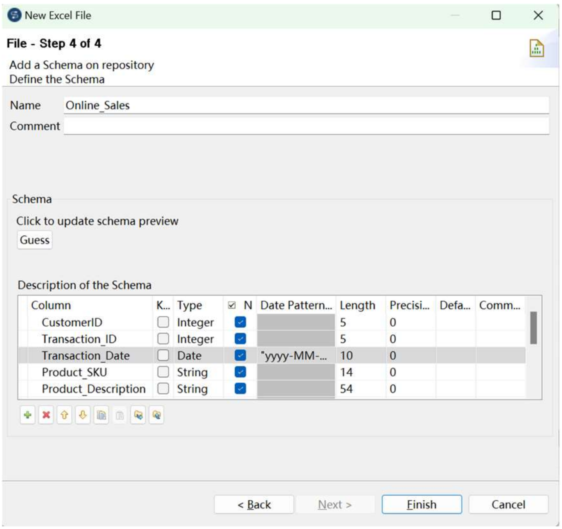
The original CustomerData.xlsx file is as follows, there are no missing value or invalid value at all, and it need no more preprocessing:

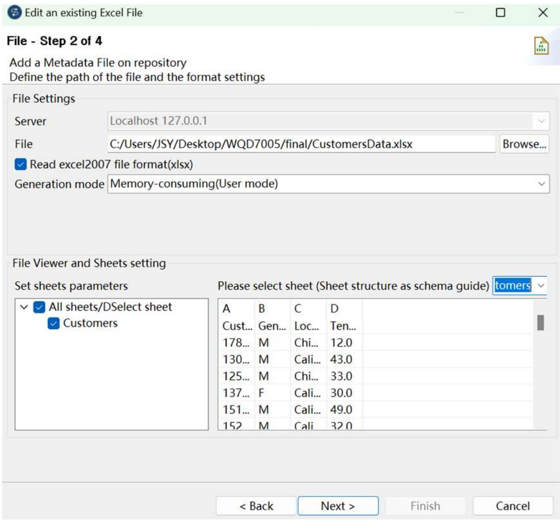


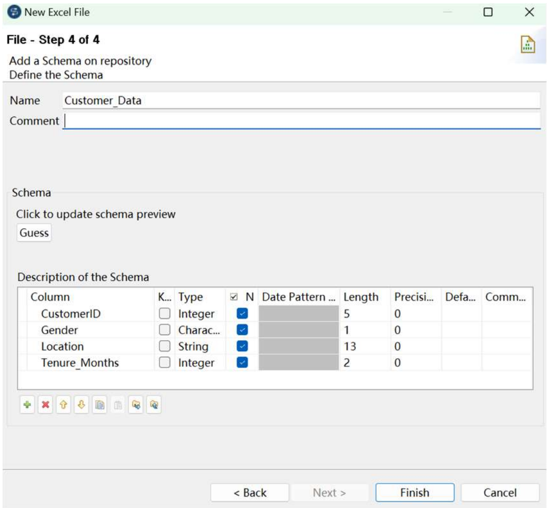
4. Data Integration Using Talend Open Studio for Data Integration:

Import CustomerData.xlsx and Online_Sales_preprocessed.xlsx, and justify the data pattern:

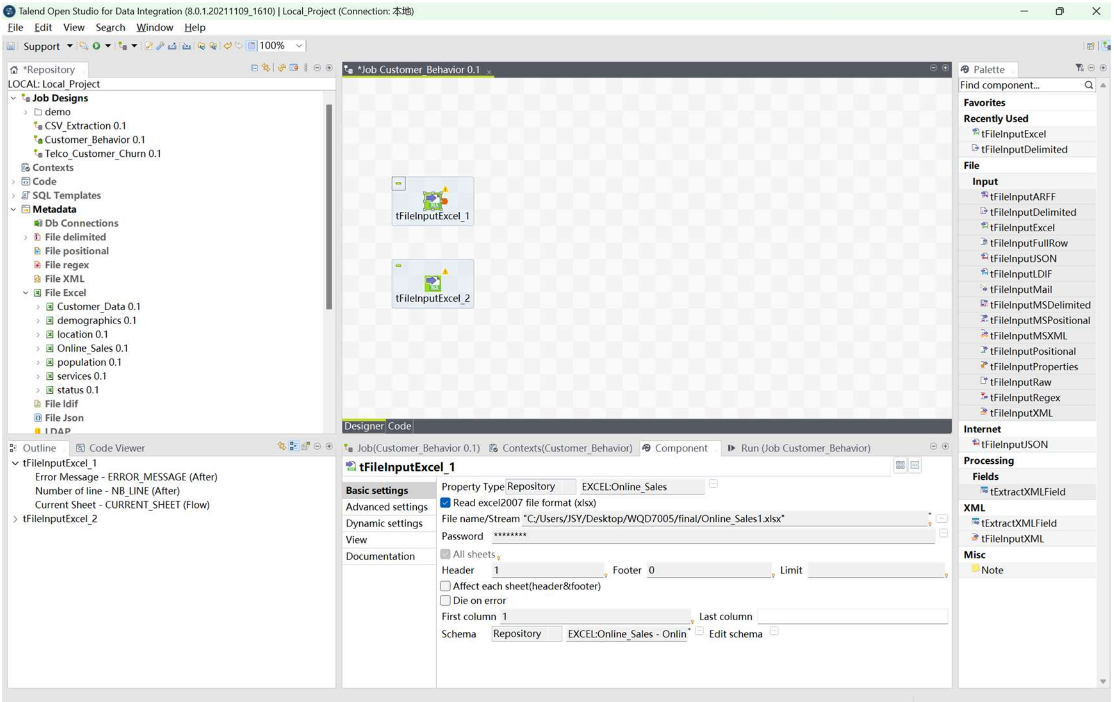




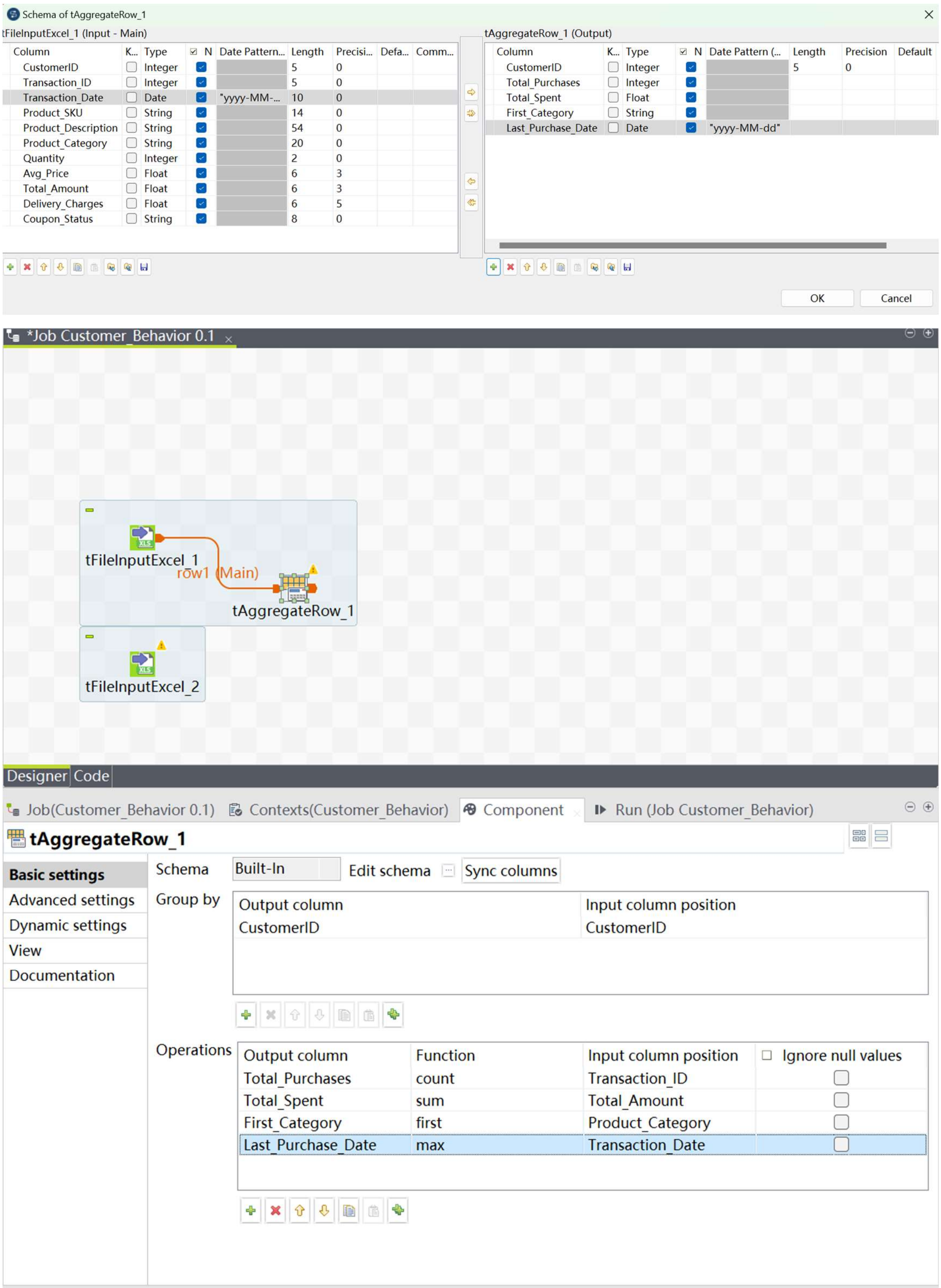




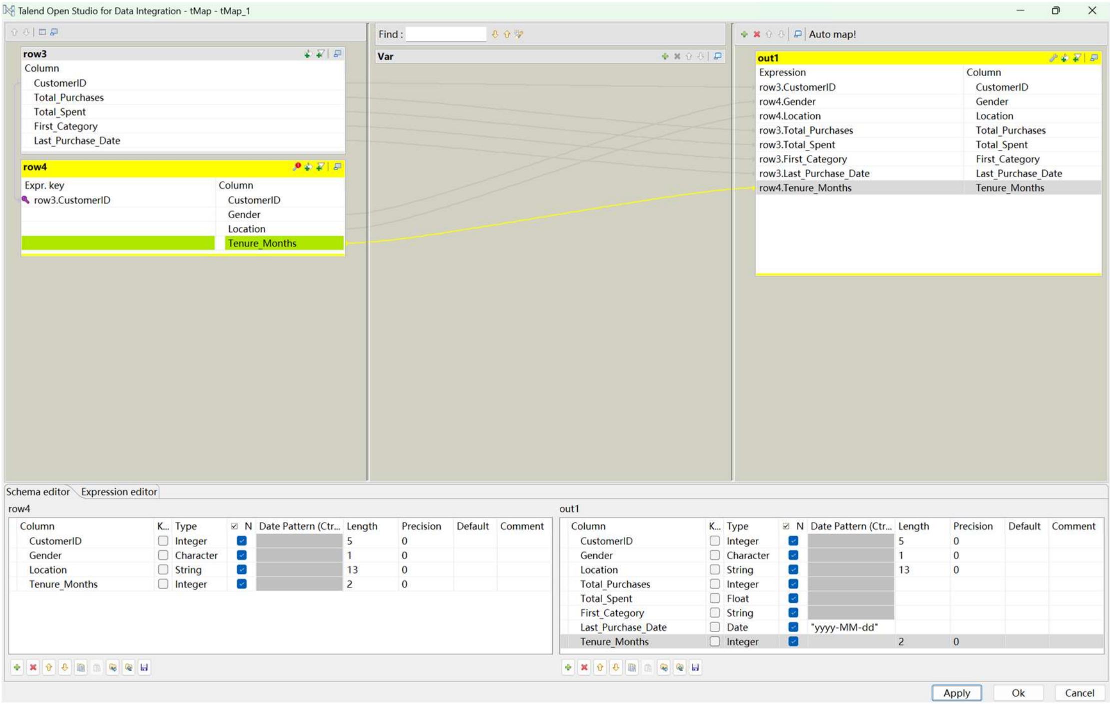
Create a new job named 'Customer_Behavior' and load Online_Sales and Customer_Data files in the tFileInputExcel components:

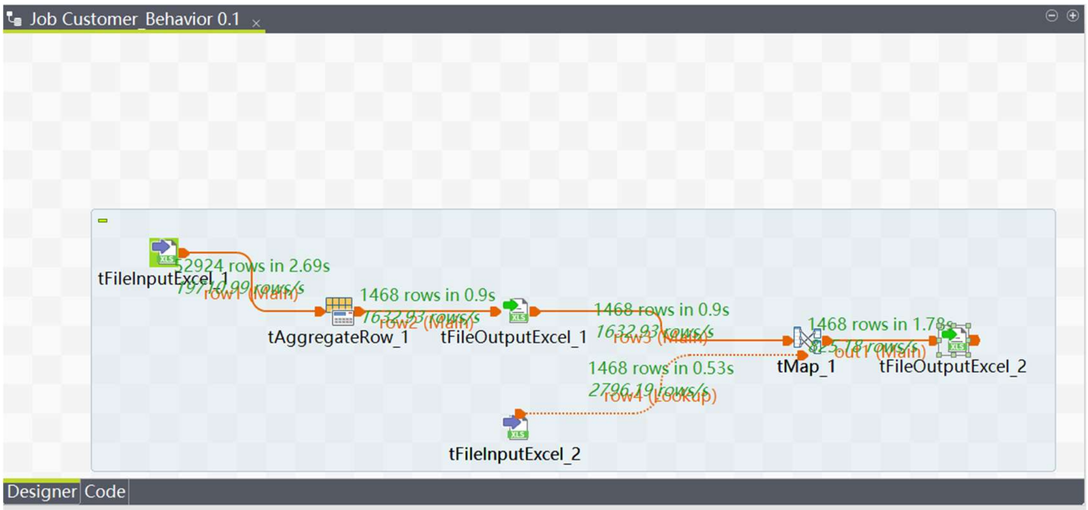


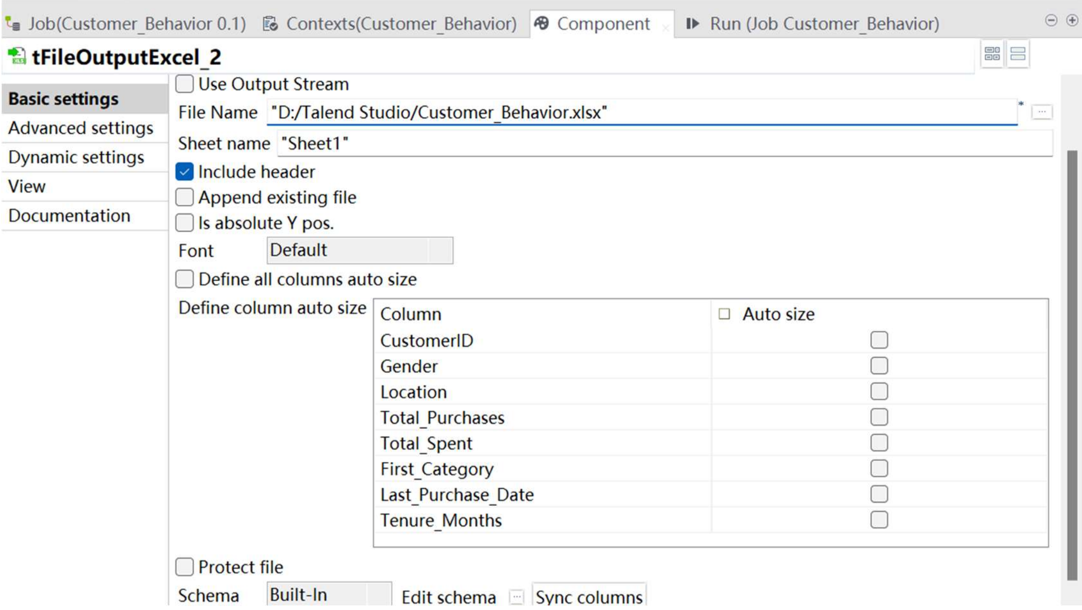
Add columns of Total_Purchases, Total_Spent, First_Category, Last_Purchased_Date, and get these data from count of TransactionID, sum of Total_Amount, first of Product_Category, and max of Transaction_Data, separately using Operations in tAggregateRow component:



Join the output of Aggregation and the CustomerData file using tMap component, and run the whole working process, then we get one combined dataset named Customer_Behavior.xlsx, having 8 columns of CustomerID, Gender, Location, Total_Purchases, Total_Spent, First_Category, Last_Purchased_Date, and Tenure_Months, with 1468 customers' rows.


Talend Open Studio for Data Integration - tMap - tMap_1
row3
Column
CustomerID
Total_Purchases
Total_Spent
First_Category
Last_Purchase_Date
row4
Expr. key
Column
row3.CustomerID
CustomerID
Gender
Location
Tenure_Months
Find :
Var
Auto map!
out1
Expression
Column
row3.CustomerID
CustomerID
row4.Gender
Gender
row4.Location
Location
row3.Total_Purchases
Total_Purchases
row3.Total_Spent
Total_Spent
row3.First_Category
First_Category
row3.Last_Purchase_Date
Last_Purchase_Date
row4.Tenure_Months
Tenure_Months
Schema editor
Expression editor
row4
Column K.. Type N Date Pattern (Ctr... Length Precision Default Comment
CustomerID Integer 5 0
Gender Character 1 0
Location String 13 0
Tenure_Months Integer 2 0
out1
Column K.. Type N Date Pattern (Ctr... Length Precision Default Comment
CustomerID Integer 5 0
Gender Character 1 0
Location String 13 0
Total_Purchases Integer
Total_Spent Float
First_Category String
Last_Purchase_Date Date "yyyy-MM-dd"
Tenure_Months Integer 2 0
Apply
Ok
Cancel


Job Customer_Behavior 0.1
tFileInputExcel_1
2924 rows in 2.69s
1797.99 rows/s
row1 (Main)
tAggregateRow_1
1468 rows in 0.9s
1632.93 rows/s
row2 (Main)
tFileOutputExcel_1
1468 rows in 0.9s
1632.93 rows/s
row3 (Main)
tMap_1
1468 rows in 1.78s
825.78 rows/s
out1 (Main)
tFileOutputExcel_2
1468 rows in 0.53s
2796.19 rows/s
row4 (Lookup)
tFileInputExcel_2
Designer
Code


Job(Customer_Behavior 0.1)
Contexts(Customer_Behavior)
Component
Run (Job Customer_Behavior)
tFileOutputExcel_2
Basic settings
Advanced settings
Dynamic settings
View
Documentation
Use Output Stream
File Name "D:/Talend Studio/Customer_Behavior.xlsx"
Sheet name "Sheet1"
Include header
Append existing file
Is absolute Y pos.
Font Default
Define all columns auto size
Define column auto size
Column
Auto size
CustomerID
Gender
Location
Total_Purchases
Total_Spent
First_Category
Last_Purchase_Date
Tenure_Months
Protect file
Schema Built-In
Edit schema
Sync columns

5. Data Analyticis Using SAS Enterprise Miner

5.1 Data preparation

Before data analytics, ‘Total_spent’ by customers need to be binned into different categories. Here I use Excel as it’s simple and clear for binning.

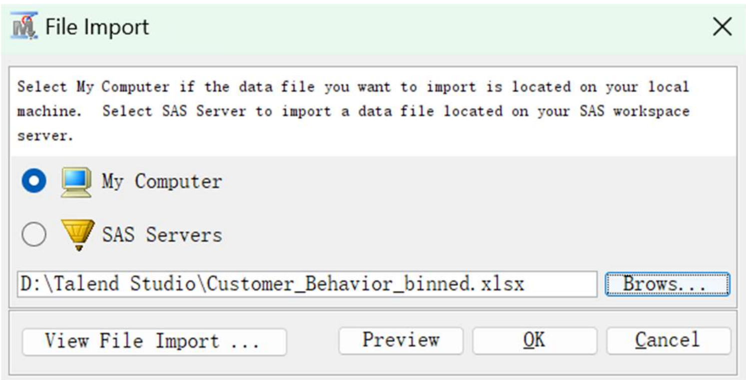
The Total_Spent of 1468 rows are distributed as follows, so I divide them into three categories based on the range:

Total Spent Range	Number of customers	Category
⩽ 1000	507	1
1000 < , ⩽ 3000	470	2
> 3000	491	3

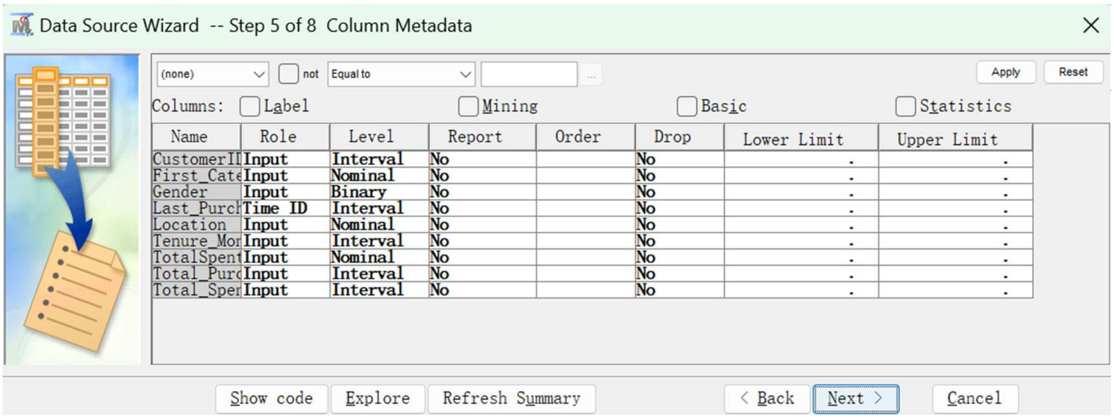
A	B	C	D	E	F	G	H	I	J	K
CustomerID	Gender	Location	Total_Purchases	Total_Spent	First_Category	Last_Purchase_Date	Tenure_Months	TotalSpent_Categories		
16353	M	Washington DC	13	843.8	Nest-USA	2019-12-20	6	=IF(E2<=1000,1,IF(E2<=3000,2,IF(E2>3000,3,0)))		
14307	F	California	79	7622.34	Apparel	2019-09-23	28	IF(logical_test, [value_if_true], [value_if_false])		
14309	F	New Jersey	22	817.73	Bags	2019-12-15	49	1		
16359	F	New York	5	62.07	Headgear	2019-09-04	49	1		
14312	M	Chicago	23	3213.7	Apparel	2019-12-12	27	3		
16365	F	California	7	484.09	Office	2019-06-09	49	1		
16367	M	California	37	1293.83	Apparel	2019-05-06	20	2		
14320	F	Chicago	34	2823.66	Apparel	2019-10-01	29	2		
14321	F	Washington DC	29	2182.59	Apparel	2019-08-14	24	2		
14329	F	California	76	4147.01	Nest-USA	2019-09-02	22	3		
16378	F	Chicago	10	1766.29	Nest-USA	2019-12-12	4	2		
14334	M	California	42	2980.63	Google	2019-10-13	41	2		
16385	M	California	21	788.36	Office	2019-04-04	19	1		
16387	F	California	4	27.97	Drinkware	2019-10-02	20	1		
14341	F	Chicago	24	1586.26	Lifestyle	2019-11-03	3	2		
14344	F	New York	14	1328.59	Nest-USA	2019-06-22	15	2		
16393	F	Chicago	22	1314.52	Apparel	2019-05-06	20	2		
16395	M	Washington DC	60	6562.08	Drinkware	2019-09-26	18	3		
16401	M	Chicago	14	728.72	Apparel	2019-09-05	40	1		
16402	M	Chicago	57	5175.15	Lifestyle	2019-02-24	31	3		
14355	F	New York	11	1270.43	Drinkware	2019-04-13	7	2		
16403	F	California	10	765.38	Apparel	2019-11-30	26	1		
16405	F	California	3	38.22	Apparel	2019-12-15	20	1		
16407	M	California	20	1203.76	Apparel	2019-11-30	48	2		
16409	F	Chicago	48	5066.26	Apparel	2019-11-03	45	3		
16411	F	Chicago	21	862.27	Drinkware	2019-02-23	40	1		
14368	F	New York	8	1553.99	Nest-USA	2019-10-12	4	2		
16419	F	California	17	946.56	Lifestyle	2019-12-06	50	1		
14373	F	Chicago	4	148.46	Apparel	2019-06-20	40	1		
16422	F	California	54	4336.18	Nest-USA	2019-12-07	35	3		

The binned dataset is saved as Customer_Behavior_binned.xlsx.

Then I import the file in SAS Enterprise Miner:



The column metadata auto-classified by SAS is as follows:

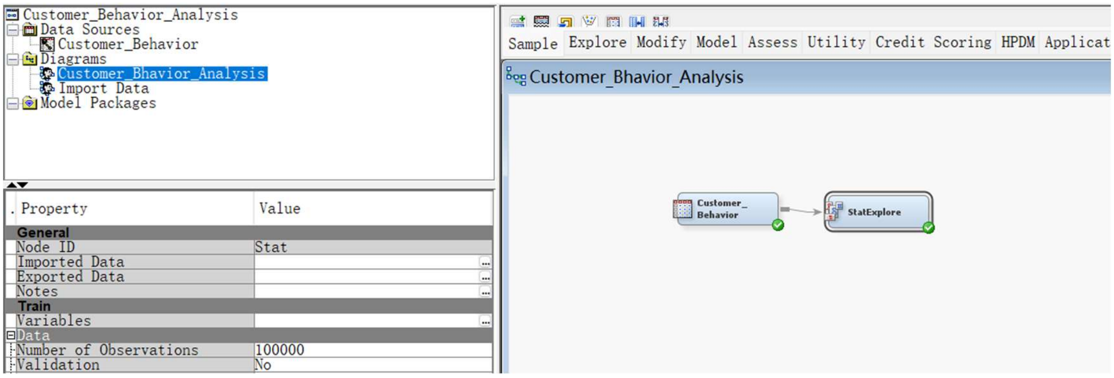


Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
CustomerII	Input	Interval	No		No	.	.
First_Cate	Input	Nominal	No		No	.	.
Gender	Input	Binary	No		No	.	.
Last_Purcl	Time ID	Interval	No		No	.	.
Location	Input	Nominal	No		No	.	.
Tenure_Mor	Input	Interval	No		No	.	.
TotalSpent	Input	Nominal	No		No	.	.
Total_Purc	Input	Interval	No		No	.	.
Total_Sper	Input	Interval	No		No	.	.

After manual reclassification of roles and levels, column metadata is as follows:

Name	Role	Level	Report	Order	Drop	Lower Limit
CustomerID	ID	Interval	No		No	
First_Category	Input	Nominal	No		No	
Gender	Input	Binary	No		No	
Last_Purchase_Date	Input	Interval	No		No	
Location	Input	Nominal	No		No	
Tenure_Months	Input	Interval	No		No	
Total_Purchases	Input	Interval	No		No	
Total_Spent	Rejected	Interval	No		No	
TotalSpent_Categories	Target	Nominal	No		No	

Identify missing value using StatExplore, and there is no missing value at all as shown below, so imputation is not needed.



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	First_Category	INPUT	19	0	Apparel	34.47	West-USA	27.11
TRAIN	Gender	INPUT	2	0	F	63.62	M	36.38
TRAIN	Location	INPUT	5	0	California	31.61	Chicago	31.06
TRAIN	TotalSpent_Categories	TARGET	3	0	1	34.54	3	33.45

Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	TotalSpent_Categories	TARGET	1	507	34.5368
TRAIN	TotalSpent_Categories	TARGET	3	491	33.4469
TRAIN	TotalSpent_Categories	TARGET	2	470	32.0163

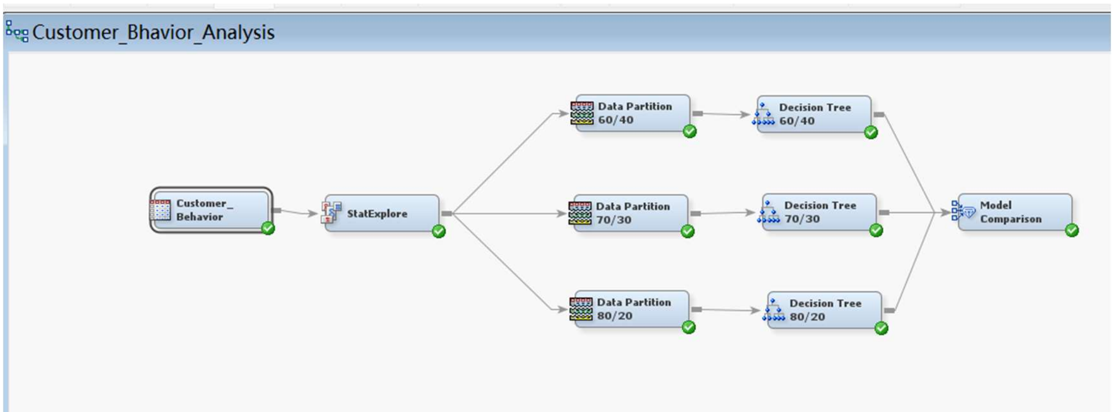
Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Last_Purchase_Date	INPUT	21769.71	101.937	1468	0	21550	21783	21914	-0.45435	-0.87708
Tenure_Months	INPUT	25.91213	13.95967	1468	0	2	26	50	-0.00265	-1.16852
Total_Purchases	INPUT	36.05177	50.88568	1468	0	1	21	695	5.784595	53.62543

5.2 Decision Tree Analysis

Split dataset using Data Partition tool into 60% train set, 40% validation set; 70% train set, 30% validation set; and 80% train set, 20% validation set, separately. Then connect them to decision tree and follow with model comparison, as shown below.



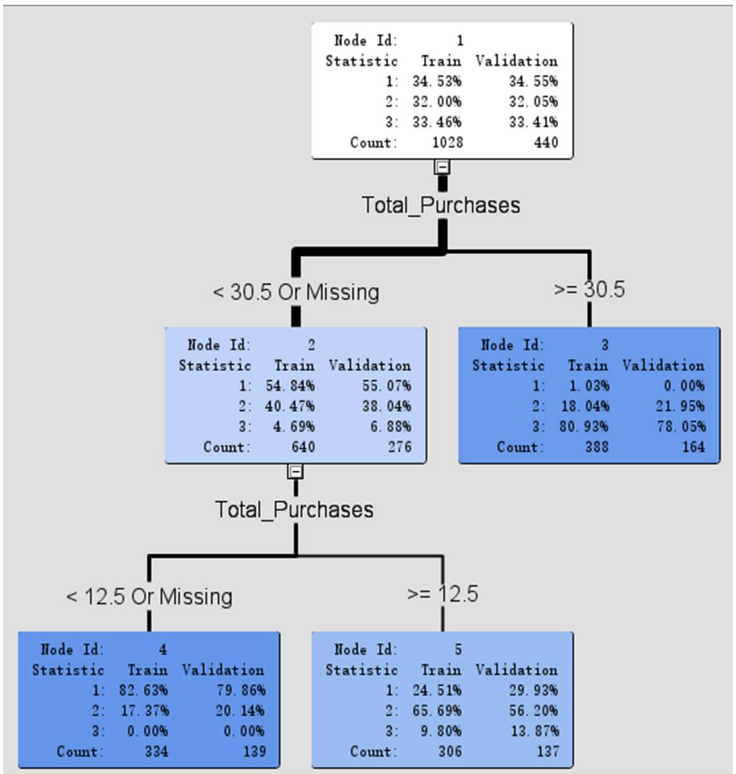
Comparison result is as shown below, the best model is using 70% train set and 30%

validation set, with the lowest Misclassification Rate of 0.28182 in validation.

Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Tree2	Decision Tree 70/30	0.28182	0.11988	0.23054	0.13804
	Tree3	Decision Tree 80/20	0.28669	0.10241	0.24340	0.12910
	Tree	Decision Tree 60/40	0.29302	0.11336	0.21453	0.14455

We select the data splitting ratio of 70:30 for further analysis, and the result is as follows. Based on the dicision tree diagram, the prediction of TotalSpent_Categories is mainly governed by the attributes Total_Purchases after pruning unnecessary branches or attributes that do not provide significant value to the prediction.



Event Classification Table

Data Role=TRAIN Target=TotalSpent_Categories Target Label=TotalSpent_Categories

False Negative	True Negative	False Positive	True Positive
30	610	74	314

Data Role=VALIDATE Target=TotalSpent_Categories Target Label=TotalSpent_Categories

False Negative	True Negative	False Positive	True Positive
19	257	36	128

Based on the confusion matrix, we can calculate the precision, recall, F1-score, accuracy and specificity as follows:

- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $F1 = \frac{2 \times precision \times recall}{precision + recall}$
- $Accuracy = \frac{TP+TN}{TP+FN+TN+FP}$
- $Specificity = \frac{TN}{TN+FP}$

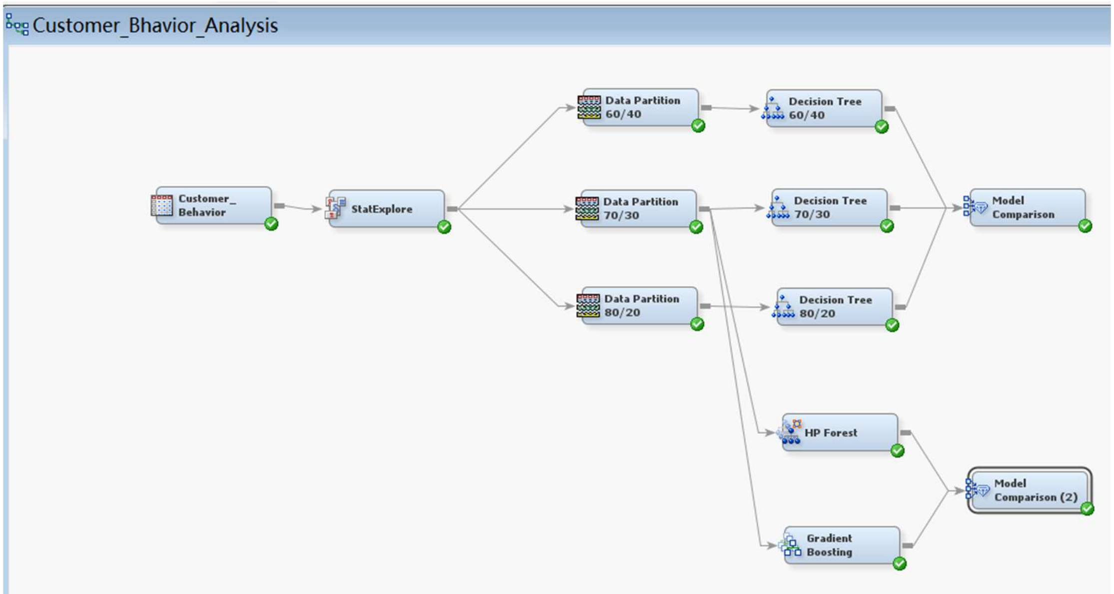
Metrics	**Decision Tree 70:30**	
	Train	Validate
Precision	0.809	0.780
Recall	0.913	0.871
F1-Score	0.858	0.823
Accuracy	0.899	0.875
Specificity	0.892	0.877

- Discussion:

Precision indicates how many of the samples predicted as positive by the model are true positives. A higher precision indicates that the model is more accurate in predicting positive examples. Recall represents the proportion of actual positive examples successfully captured by the model. A higher recall indicates that the model is better able to identify positive examples. F1-Score combines Precision and Recall and is a measure of the overall performance of the model. It has a trade-off between Precision and Recall. Accuracy represents the overall proportion of correct predictions by the model. A higher accuracy indicates a better overall model performance. Specificity represents the proportion of negative examples that the model successfully predicts. Higher specificity indicates that the model is better able to avoid misclassifying negative examples as positive examples. Taken together, the model performs relatively well on the training and validation sets, with high accuracy, precision, and recall.

5.3 Ensemble Methods

Using HP Forest as the bagging modelling technique and Gradient Boosting as the boosting modelling technique to predict TotalSpent_Category as follows:



Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
HPDMForest	HP Forest	TRAIN	TotalSpent_Categories	TotalSpent_Categories	28	622	62	316
HPDMForest	HP Forest	VALIDATE	TotalSpent_Categories	TotalSpent_Categories	19	257	36	128
Boost	Gradient Boosting	TRAIN	TotalSpent_Categories	TotalSpent_Categories	55	648	36	289
Boost	Gradient Boosting	VALIDATE	TotalSpent_Categories	TotalSpent_Categories	28	274	19	119

Based on the confusion matrix, we can calculate the precision, recall, F1-score, accuracy and specificity as follows:

Metrics	HP Forest		Gradient Boosting	
	Train	Validate	Train	Validate
Precision	0.836	0.780	0.889	0.862
Recall	0.919	0.871	0.840	0.810
F1-Score	0.875	0.823	0.864	0.835
Accuracy	0.912	0.875	0.911	0.893
Specificity	0.909	0.877	0.947	0.935

- Discussion:

  HP Forest and Gradient Boosting has similar performance with high precision, recall, F1-Score, Accuracy, and Specificity. Both perform better than Decision Tree.

6. Business Suggestions

In this case study, the total spent by customers in 2019 can be used to evaluate customer value, which can help companies screen out high-value customers and focus on arranging after-sales and other services, thereby reducing customer churn, and increasing profit margins.

In addition, as shown in the decision tree analysis results, the total spent range classification of customers is mainly related to the number of purchases, so revenue can be improved through the following methods:

Increase the number of promotions, but limit the promotion time, thereby stimulating customers' desire to purchase, increasing the number of customer purchases, and increasing total spent.