Talend Data Integration:Merge two data sets into one.

Step:

Open Talend Data Integration Studio:

Launch the Talend Data Integration tool and open your project.

Create a New Job:

Right-click in the Repository panel on the left under Job Designs.

Choose Create job.

Provide a name for the job and a description, if desired. Click Finish.

Add a tFileInputDelimited Component:

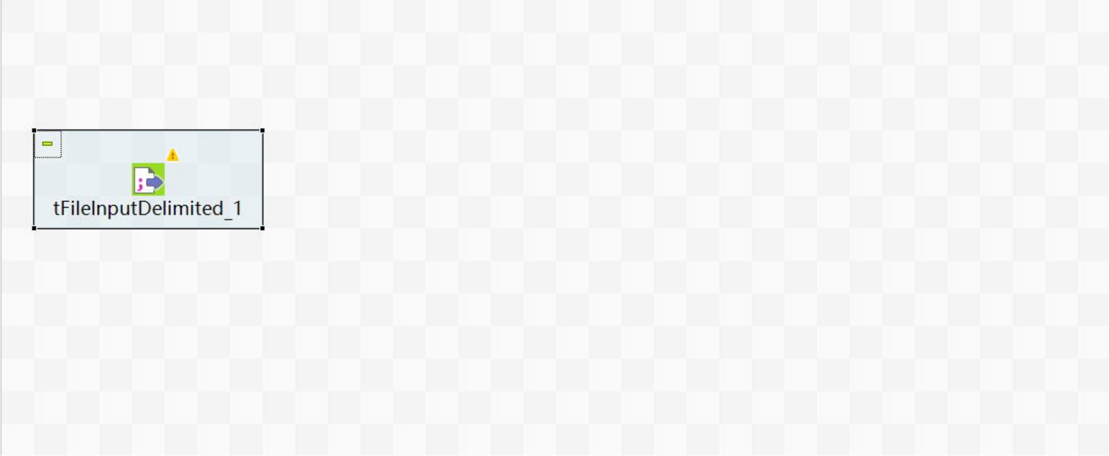
This component is used to read delimited files like CSV.

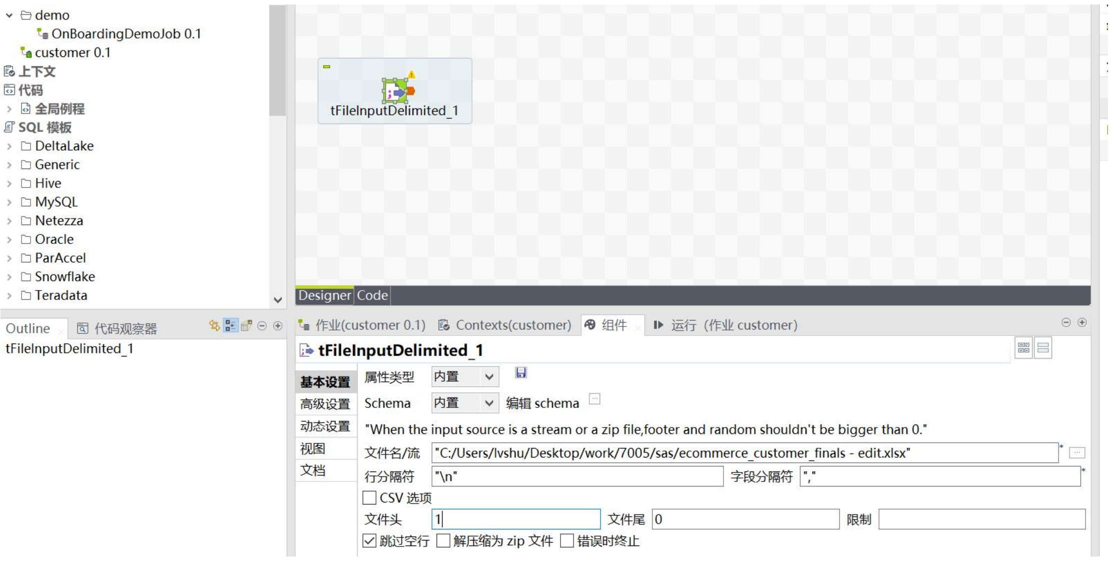
From the Palette panel on the right, type "tFileInputDelimited" into the search bar.

Drag the tFileInputDelimited component to the design workspace.

Configure the tFileInputDelimited Component:

Click on the tFileInputDelimited component to select it.





tFileInputDelimited_1

属性类型 内置

Schema 内置 编辑 schema

"When the input source is a stream or a zip file,footer and random shouldn't be bigger than 0."

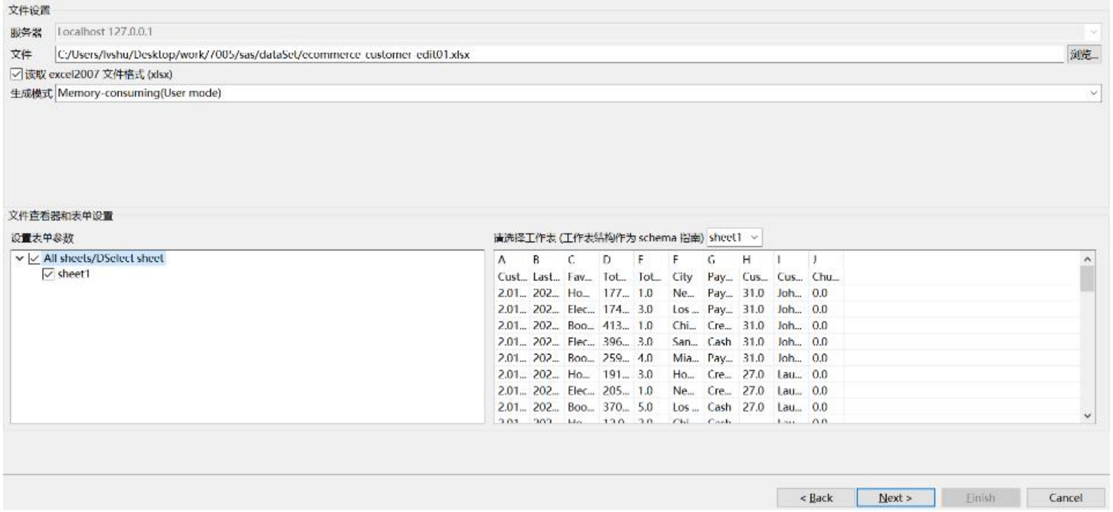
文件名/流 "C:/Users/lvshu/Desktop/work/7005/sas/ecommerce_customer_finals - edit.xlsx"

行分隔符 "\n" 字段分隔符 ","

□ CSV 选项

文件头 1 文件尾 0 限制

☑ 跳过空行 □ 解压缩为 zip 文件 □ 错误时终止



文件设置

服务器 Localhost 127.0.0.1

文件 C:/Users/lvshu/Desktop/work/7005/sas/dataSet/ecommerce_customer_edit01.xlsx

☑ 读取 excel2007 文件格式 (xlsx)

生成模式 Memory-consuming(User mode)

文件查看器和表单设置

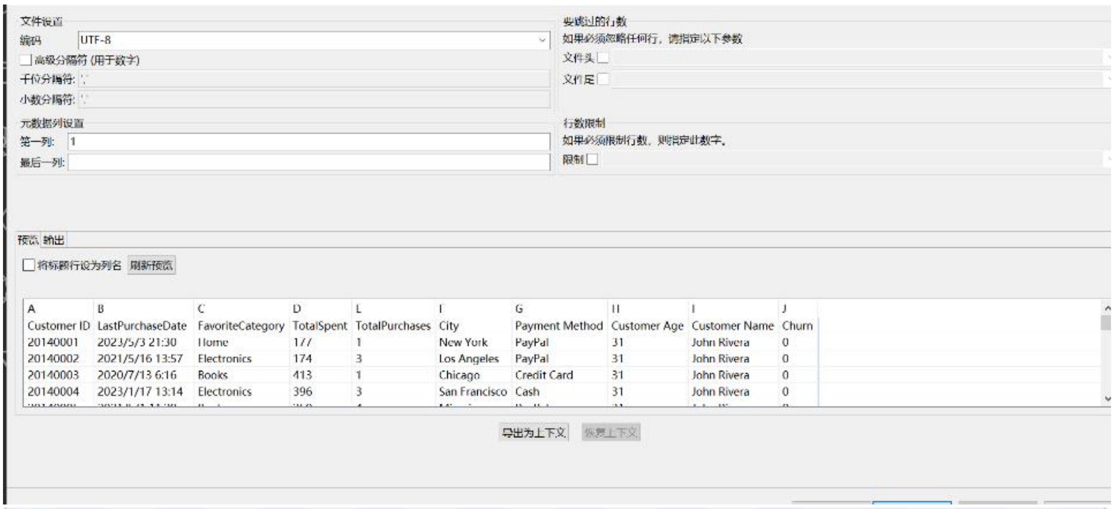
设置表单参数

☑ All sheets/DSelect sheet

☑ sheet1

请选择工作表 (工作表结构作为 schema 指南) sheet1

A	B	C	D	E	F	G	H	I	J
Cust...	Last...	Fav...	Tot...	Tot...	City	Pay...	Cus...	Cus...	Chu...
2.01...	202...	Ho...	177...	1.0	Ne...	Pay...	31.0	Joh...	0.0
2.01...	202...	Elec...	174...	3.0	Los ...	Pay...	31.0	Joh...	0.0
2.01...	202...	Boo...	413...	1.0	Chi...	Cre...	31.0	Joh...	0.0
2.01...	202...	Elec...	396...	3.0	San...	Cash	31.0	Joh...	0.0
2.01...	202...	Boo...	259...	4.0	Mia...	Pay...	31.0	Joh...	0.0
2.01...	202...	Ho...	191...	3.0	Ho...	Cre...	27.0	Lau...	0.0
2.01...	202...	Elec...	205...	1.0	Ne...	Cre...	27.0	Lau...	0.0
2.01...	202...	Boo...	370...	5.0	Los ...	Cash	27.0	Lau...	0.0



文件设置

编码 UTF-8

☐ 高级分隔符 (用于数字)

千位分隔符: ','

小数分隔符: '.'

元数据列设置

第一列: 1

最后一列:

要跳过的行数

如果必须忽略任何行，请指定以下参数

文件头 ☐

文件尾 ☐

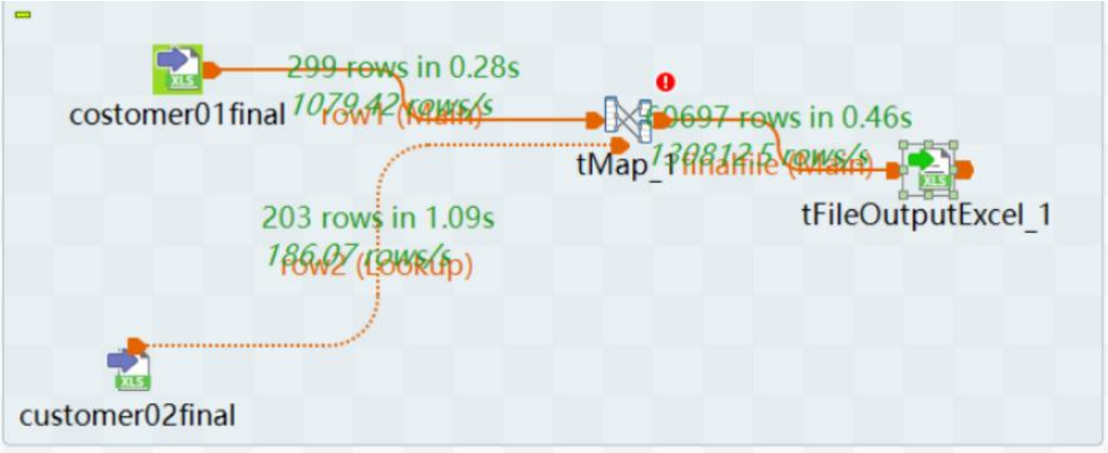
行数限制

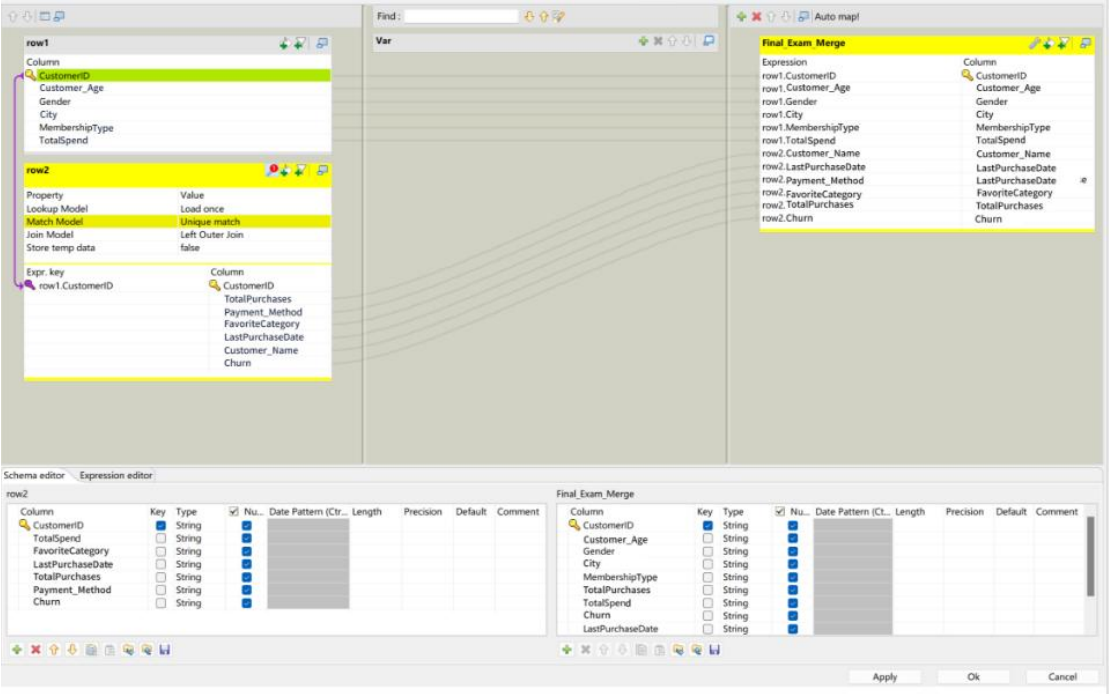
如果必须限制行数，则指定此数字。

限制 ☐

☐ 将标题行设为列名

A	B	C	D	E	F	G	H	I	J
Customer ID	LastPurchaseDate	FavoriteCategory	TotalSpent	TotalPurchases	City	Payment Method	Customer Age	Customer Name	Churn
20140001	2023/5/3 21:30	Home	177	1	New York	PayPal	31	John Rivera	0
20140002	2021/5/16 13:57	Electronics	174	3	Los Angeles	PayPal	31	John Rivera	0
20140003	2020/7/13 6:16	Books	413	1	Chicago	Credit Card	31	John Rivera	0
20140004	2023/1/17 13:14	Electronics	396	3	San Francisco	Cash	31	John Rivera	0





row1

Column
CustomerID
Customer_Age
Gender
City
MembershipType
TotalSpend

row2

Property	Value
Lookup Model	Load once
Match Model	Unique match
Join Model	Left Outer Join
Store temp data	false

Expr. key	Column
row1.CustomerID	CustomerID
	TotalPurchases
	Payment_Method
	FavoriteCategory
	LastPurchaseDate
	Customer_Name
	Churn

Final_Exam_Merge

Expression	Column
row1.CustomerID	CustomerID
row1.Customer_Age	Customer_Age
row1.Gender	Gender
row1.City	City
row1.MembershipType	MembershipType
row1.TotalSpend	TotalSpend
row2.Customer_Name	Customer_Name
row2.LastPurchaseDate	LastPurchaseDate
row2.Payment_Method	LastPurchaseDate
row2.FavoriteCategory	FavoriteCategory
row2.TotalPurchases	TotalPurchases
row2.Churn	Churn

row2

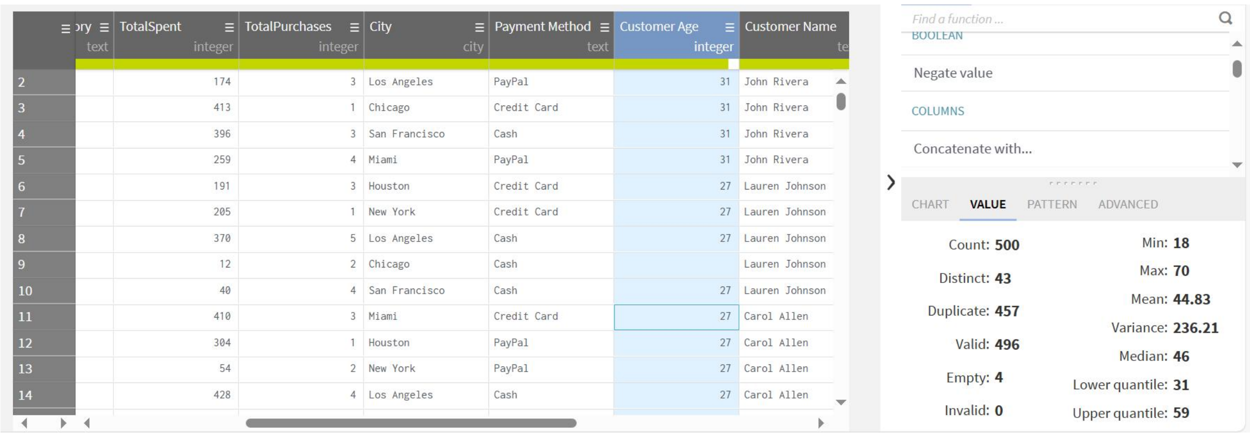
Column	Key	Type	Nu...	Date Pattern (Ctr...	Length	Precision	Default	Comment
CustomerID	☑	String	☑					
TotalSpend	□	String	☑					
FavoriteCategory	□	String	☑					
LastPurchaseDate	□	String	☑					
TotalPurchases	□	String	☑					
Payment_Method	□	String	☑					
Churn	□	String	☑					

Final_Exam_Merge

Column	Key	Type	Nu...	Date Pattern (Ct...	Length	Precision	Default	Comment
CustomerID	☑	String	☑					
Customer_Age	□	String	☑					
Gender	□	String	☑					
City	□	String	☑					
MembershipType	□	String	☑					
TotalPurchases	□	String	☑					
TotalSpend	□	String	☑					
Churn	□	String	☑					
LastPurchaseDate	□	String	☑					

Talend Data Prep: Remove missing values

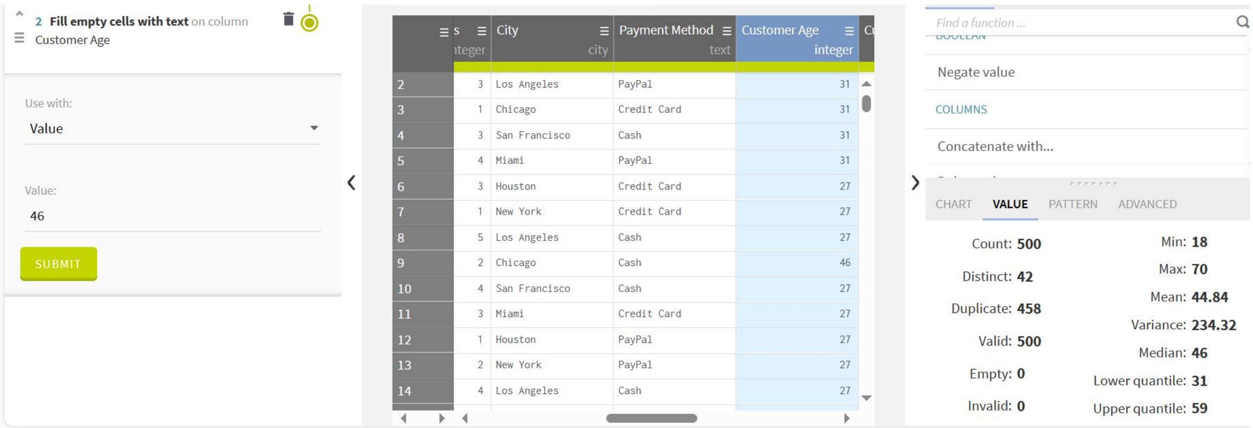
When clicking on each column you can see the number of missing values in that column.



	ory text	TotalSpent integer	TotalPurchases integer	City city	Payment Method text	Customer Age integer	Customer Name te
2		174	3	Los Angeles	PayPal	31	John Rivera
3		413	1	Chicago	Credit Card	31	John Rivera
4		396	3	San Francisco	Cash	31	John Rivera
5		259	4	Miami	PayPal	31	John Rivera
6		191	3	Houston	Credit Card	27	Lauren Johnson
7		205	1	New York	Credit Card	27	Lauren Johnson
8		370	5	Los Angeles	Cash	27	Lauren Johnson
9		12	2	Chicago	Cash		Lauren Johnson
10		40	4	San Francisco	Cash	27	Lauren Johnson
11		410	3	Miami	Credit Card	27	Carol Allen
12		304	1	Houston	PayPal	27	Carol Allen
13		54	2	New York	PayPal	27	Carol Allen
14		428	4	Los Angeles	Cash	27	Carol Allen

Count: **500**
Distinct: **43**
Duplicate: **457**
Valid: **496**
Empty: **4**
Invalid: **0**

Min: **18**
Max: **70**
Mean: **44.83**
Variance: **236.21**
Median: **46**
Lower quantile: **31**
Upper quantile: **59**



2 **Fill empty cells with text** on column
Customer Age

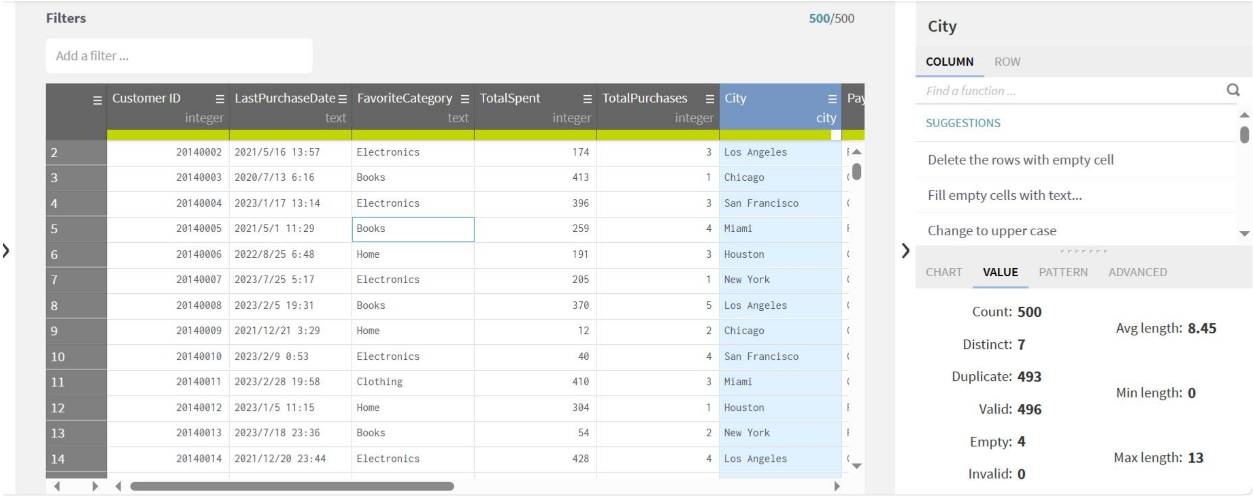
Use with:
Value

Value:
46

	s integer	City city	Payment Method text	Customer Age integer	Cu
2	3	Los Angeles	PayPal	31	
3	1	Chicago	Credit Card	31	
4	3	San Francisco	Cash	31	
5	4	Miami	PayPal	31	
6	3	Houston	Credit Card	27	
7	1	New York	Credit Card	27	
8	5	Los Angeles	Cash	27	
9	2	Chicago	Cash	46	
10	4	San Francisco	Cash	27	
11	3	Miami	Credit Card	27	
12	1	Houston	PayPal	27	
13	2	New York	PayPal	27	
14	4	Los Angeles	Cash	27	

Count: **500**
Distinct: **42**
Duplicate: **458**
Valid: **500**
Empty: **0**
Invalid: **0**

Min: **18**
Max: **70**
Mean: **44.84**
Variance: **234.32**
Median: **46**
Lower quantile: **31**
Upper quantile: **59**

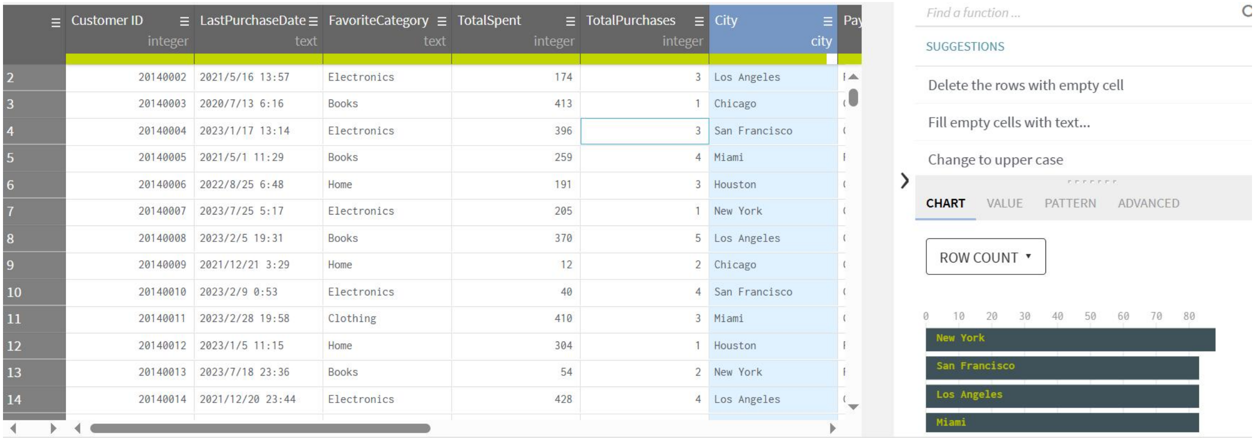


500/500

	Customer ID integer	LastPurchaseDate text	FavoriteCategory text	TotalSpent integer	TotalPurchases integer	City city
2	20140002	2021/5/16 13:57	Electronics	174	3	Los Angeles
3	20140003	2020/7/13 6:16	Books	413	1	Chicago
4	20140004	2023/1/17 13:14	Electronics	396	3	San Francisco
5	20140005	2021/5/1 11:29	Books	259	4	Miami
6	20140006	2022/8/25 6:48	Home	191	3	Houston
7	20140007	2023/7/25 5:17	Electronics	205	1	New York
8	20140008	2023/2/5 19:31	Books	370	5	Los Angeles
9	20140009	2021/12/21 3:29	Home	12	2	Chicago
10	20140010	2023/2/9 0:53	Electronics	40	4	San Francisco
11	20140011	2023/2/28 19:58	Clothing	410	3	Miami
12	20140012	2023/1/5 11:15	Home	304	1	Houston
13	20140013	2023/7/18 23:36	Books	54	2	New York
14	20140014	2021/12/20 23:44	Electronics	428	4	Los Angeles

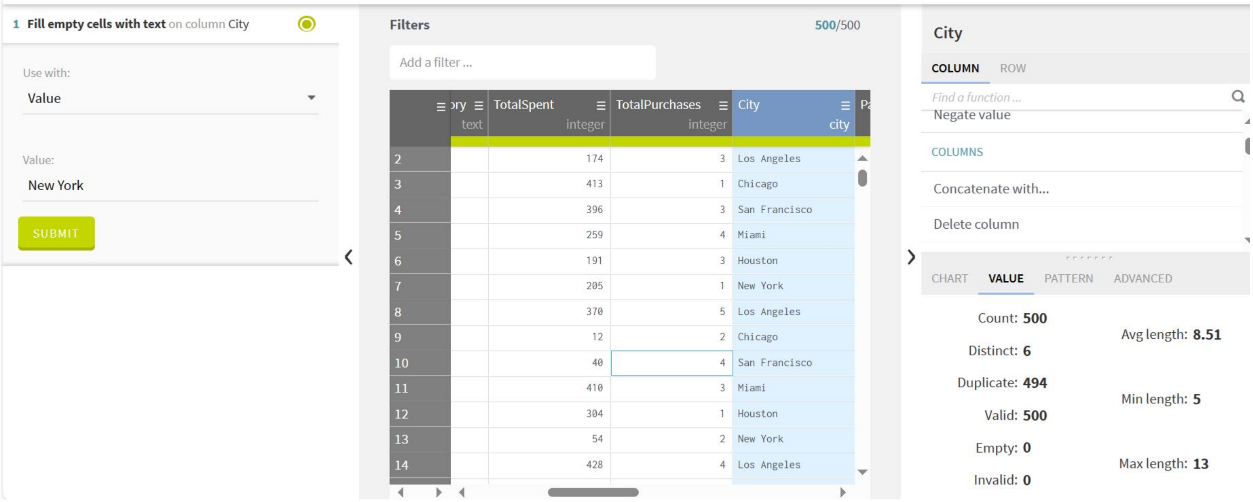
Count: 500
Distinct: 7
Duplicate: 493
Valid: 496
Empty: 4
Invalid: 0
Avg length: 8.45
Min length: 0
Max length: 13

The city names with the highest frequency of occurrence usually represent the major cities in the dataset, and populating such city names helps to maintain the consistency of the overall data distribution. For missing addresses, this experiment uses taking the city name with the highest frequency of occurrence to fill it.



	Customer ID integer	LastPurchaseDate text	FavoriteCategory text	TotalSpent integer	TotalPurchases integer	City city
2	20140002	2021/5/16 13:57	Electronics	174	3	Los Angeles
3	20140003	2020/7/13 6:16	Books	413	1	Chicago
4	20140004	2023/1/17 13:14	Electronics	396	3	San Francisco
5	20140005	2021/5/1 11:29	Books	259	4	Miami
6	20140006	2022/8/25 6:48	Home	191	3	Houston
7	20140007	2023/7/25 5:17	Electronics	205	1	New York
8	20140008	2023/2/5 19:31	Books	370	5	Los Angeles
9	20140009	2021/12/21 3:29	Home	12	2	Chicago
10	20140010	2023/2/9 0:53	Electronics	40	4	San Francisco
11	20140011	2023/2/28 19:58	Clothing	410	3	Miami
12	20140012	2023/1/5 11:15	Home	304	1	Houston
13	20140013	2023/7/18 23:36	Books	54	2	New York
14	20140014	2021/12/20 23:44	Electronics	428	4	Los Angeles

As shown, New York is the most frequent. So fill in the four missing city names as New York.



1 Fill empty cells with text on column City

Use with: Value

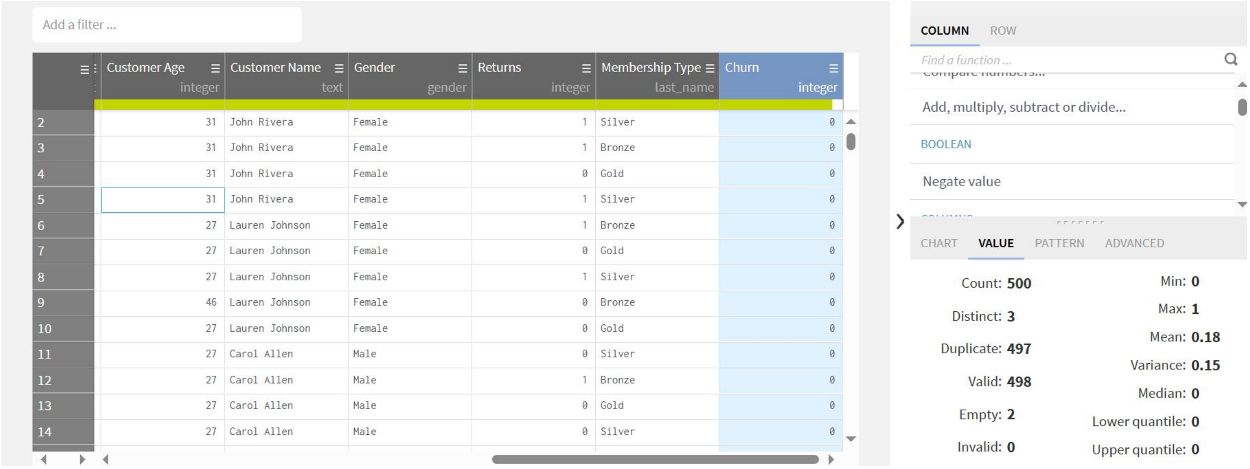
Value: New York

500/500

	ory text	TotalSpent integer	TotalPurchases integer	City city
2		174	3	Los Angeles
3		413	1	Chicago
4		396	3	San Francisco
5		259	4	Miami
6		191	3	Houston
7		205	1	New York
8		370	5	Los Angeles
9		12	2	Chicago
10		40	4	San Francisco
11		410	3	Miami
12		304	1	Houston
13		54	2	New York
14		428	4	Los Angeles

Count: 500
Distinct: 6
Duplicate: 494
Valid: 500
Empty: 0
Invalid: 0
Avg length: 8.51
Min length: 5
Max length: 13

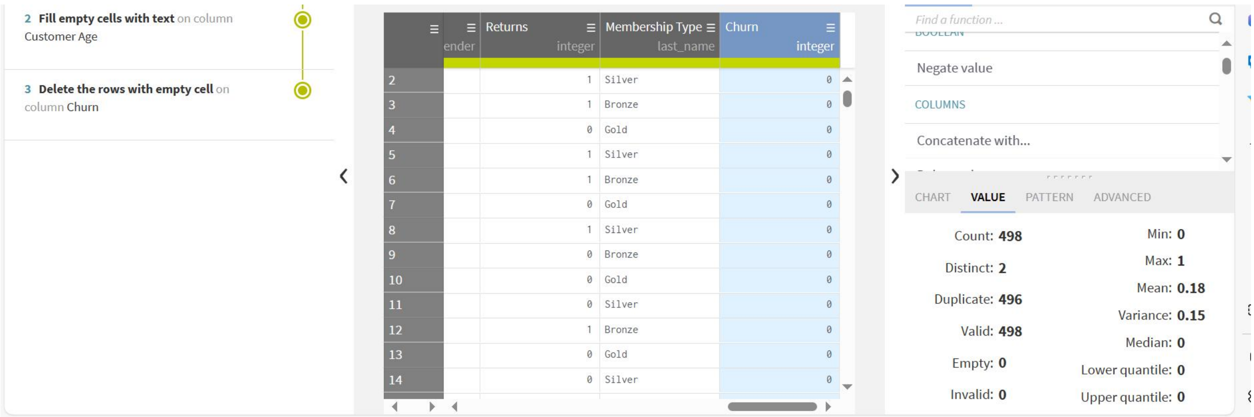
For critical attributes such as missing churn value, where there are few missing values, the deletion of missing values is used to deal with it.Churn value is usually a critical attribute in user churn prediction as it directly reflects whether a user is churned or not. In this case, it is important to ensure the accuracy of this attribute as it is the target variable for model training. Removing missing values avoids introducing uncertainty about the accuracy of user churn prediction during the modeling process.



Add a filter ...

	Customer Age integer	Customer Name text	Gender gender	Returns integer	Membership Type last_name	Churn integer
2	31	John Rivera	Female	1	Silver	0
3	31	John Rivera	Female	1	Bronze	0
4	31	John Rivera	Female	0	Gold	0
5	31	John Rivera	Female	1	Silver	0
6	27	Lauren Johnson	Female	1	Bronze	0
7	27	Lauren Johnson	Female	0	Gold	0
8	27	Lauren Johnson	Female	1	Silver	0
9	46	Lauren Johnson	Female	0	Bronze	0
10	27	Lauren Johnson	Female	0	Gold	0
11	27	Carol Allen	Male	0	Silver	0
12	27	Carol Allen	Male	1	Bronze	0
13	27	Carol Allen	Male	0	Gold	0
14	27	Carol Allen	Male	0	Silver	0

Count: 500
Distinct: 3
Duplicate: 497
Valid: 498
Empty: 2
Invalid: 0
Min: 0
Max: 1
Mean: 0.18
Variance: 0.15
Median: 0
Lower quantile: 0
Upper quantile: 0



2 **Fill empty cells with text** on column Customer Age

3 **Delete the rows with empty cell** on column Churn

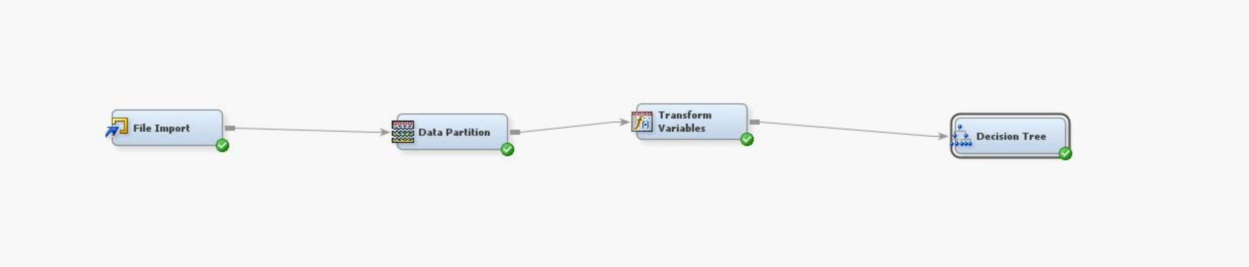
	Returns integer	Membership Type last_name	Churn integer
2	1	Silver	0
3	1	Bronze	0
4	0	Gold	0
5	1	Silver	0
6	1	Bronze	0
7	0	Gold	0
8	1	Silver	0
9	0	Bronze	0
10	0	Gold	0
11	0	Silver	0
12	1	Bronze	0
13	0	Gold	0
14	0	Silver	0

Count: 498
Distinct: 2
Duplicate: 496
Valid: 498
Empty: 0
Invalid: 0
Min: 0
Max: 1
Mean: 0.18
Variance: 0.15
Median: 0
Lower quantile: 0
Upper quantile: 0

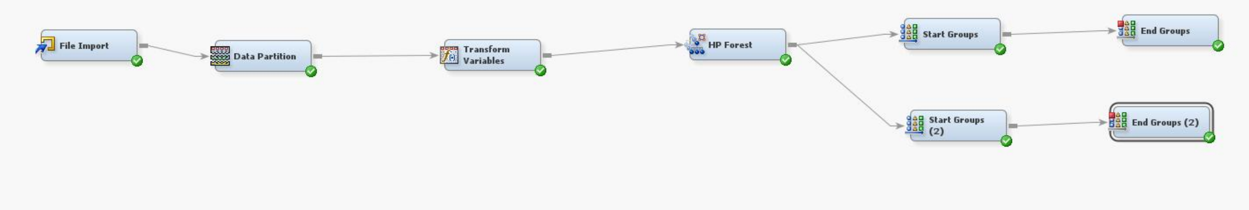
SAS e-Miner:

decision trees：

1. To create a project: Select the "File" menu and then select "New Project". Name the project and set the properties of the project.
2. Create a flowchart: In the project, select the "Diagram" menu, and then select "Create Diagram".
3. Import data: In the project, import the dataset that contains the data you want to analyze.
4. Set the ratio of test set to validation set
5. Configure the decision tree nodes.
6. Run the entire flowchart. Select the "Run" menu and then select "Run Diagram".


File Import
Data Partition
Transform Variables
Decision Tree

Cluser:


File Import
Data Partition
Transform Variables
HP Forest
Start Groups
End Groups
Start Groups (2)
End Groups (2)