



**UNIVERSITI  
MALAYA**

*Faculty of Computer Science  
and Information Technology*

**WQD7005  
DATA MINING**

**ALTERNATIVE ASSESSMENT 1**

**LECTURER NAME:  
PROF. DR. TEH YING WAH**

**STUDENT NAME:  
AMIRA HANEE BINTI SAIFULAZRI  
17202918/1**

**GitHub Link:**

**<https://github.com/amirahanee/WQD7005AltAsmnt1/tree/main>**

**Talend Data Integration**

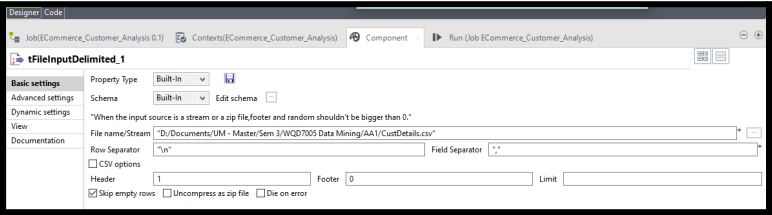
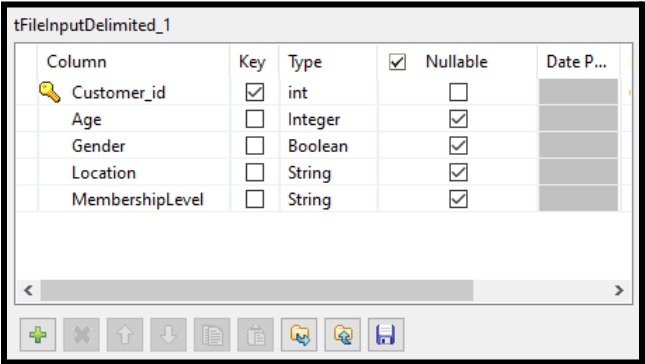
Talend Data Integration is a comprehensive open-source tool used for Extract, Transform, Load (ETL) processes. Once the data is imported into Talend, users can apply a variety of data integration tasks, including data cleansing, transformation, and loading into target systems. The importation of data into Talend Data Integration marks the initial phase of a workflow where the data will undergo various operations to meet specific business or analytical requirements.

I used multiple nodes to complete this integration, which are:

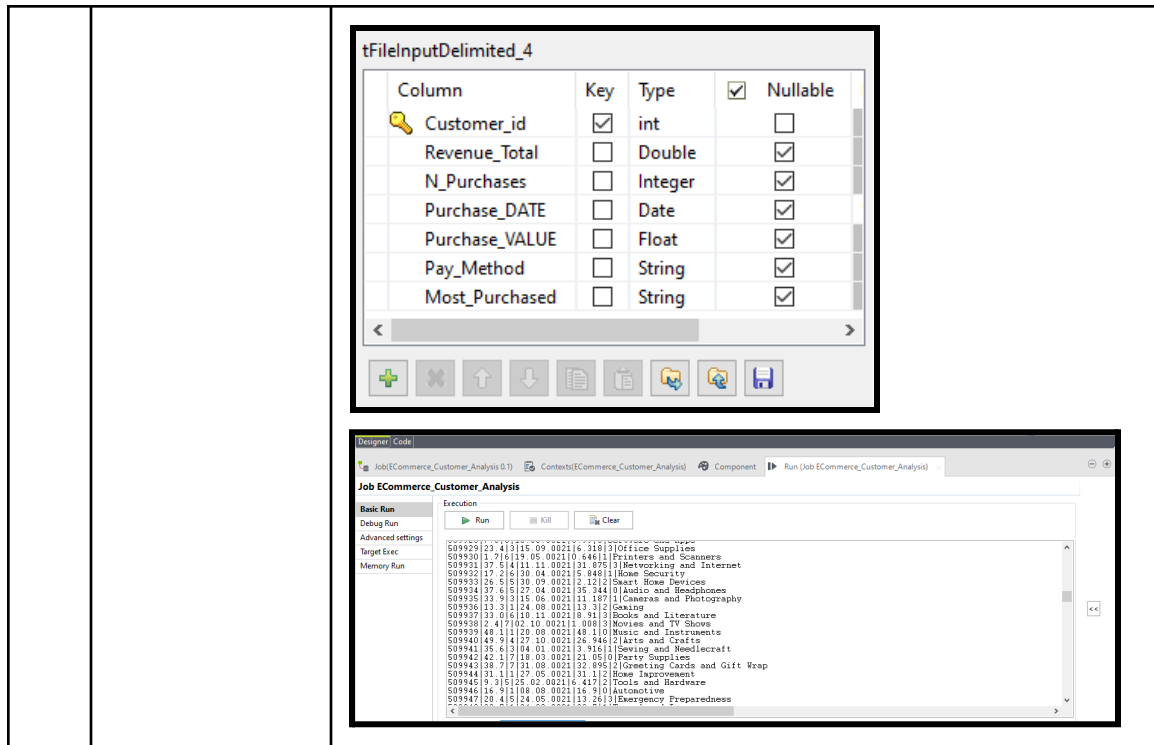
**1. tFileInputDelimited**

A component used for reading data from delimited text files. It is part of the input/output (I/O) family of components and is specifically designed to handle delimited files, where data is separated by a specified delimiter, such as a comma, tab, semicolon, etc.

It is being used three times to upload three different datasets.

No	Dataset Name	Evidence
1	CustDetails	<div></div> <div></div>

		<div><div>Designer   Code</div><div>Job forCustDetails 0.1</div><div>Contexts forCustDetails</div><div>Component</div><div>Run (Job forCustDetails)</div></div> <div><div>Job forCustDetails</div><div>Execution</div><div>Basic Run</div><div>Debug Run</div><div>Advanced settings</div><div>Target Exec</div><div>Memory Run</div><div>511221 27 true Wisconsin Nova</div><div>511223 41 false Maine Nova</div><div>511224 44 false Rhode_Sland Nova</div><div>511225 52 true Missouri Nova</div><div>511226 28 true Iowa Nova</div><div>511227 24 true New_Jersey Nova</div><div>511228 62 true Kentucky Nova</div><div>511229 60 false California Nova</div><div>511230 25 true Massachusetts Nova</div><div>511231 59 false Maine Nova</div><div>511232 59 true New_York Nova</div><div>511233 48 true Kansas Nova</div><div>511234 55 true Alaska Nova</div><div>511235 49 true Delaware Nova</div><div>511236 21 true Connecticut Nova</div><div>511237 57 true South_Carolina Nova</div><div>511238 54 true Wisconsin Nova</div><div>511239 59 true Utah Nova</div><div>511240 72 true Nevada Nova</div></div>
--	--	---



## 2. tMap1

The "tMap" component is a versatile transformation component that allows you to define mappings between input and output columns. It is commonly used for data transformations, lookups, and calculations within a Talend Job. The tMap component can have multiple input flows, and can apply various operations such as filtering, aggregating, and joining data in a visually intuitive way.

In this case study, those three datasets are being connected with Customer\_id as it is the primary keys for all three datasets. Refer to image below:

CustDetails	
Column	
Customer_id	
Age	
Gender	
Location	
MembershipLevel	

CustLog	
Expr. key	Column
CustDetails.Customer_id	Customer_id
	Time_Spent
	Browser
	Newsletter
	Voucher
	Churn

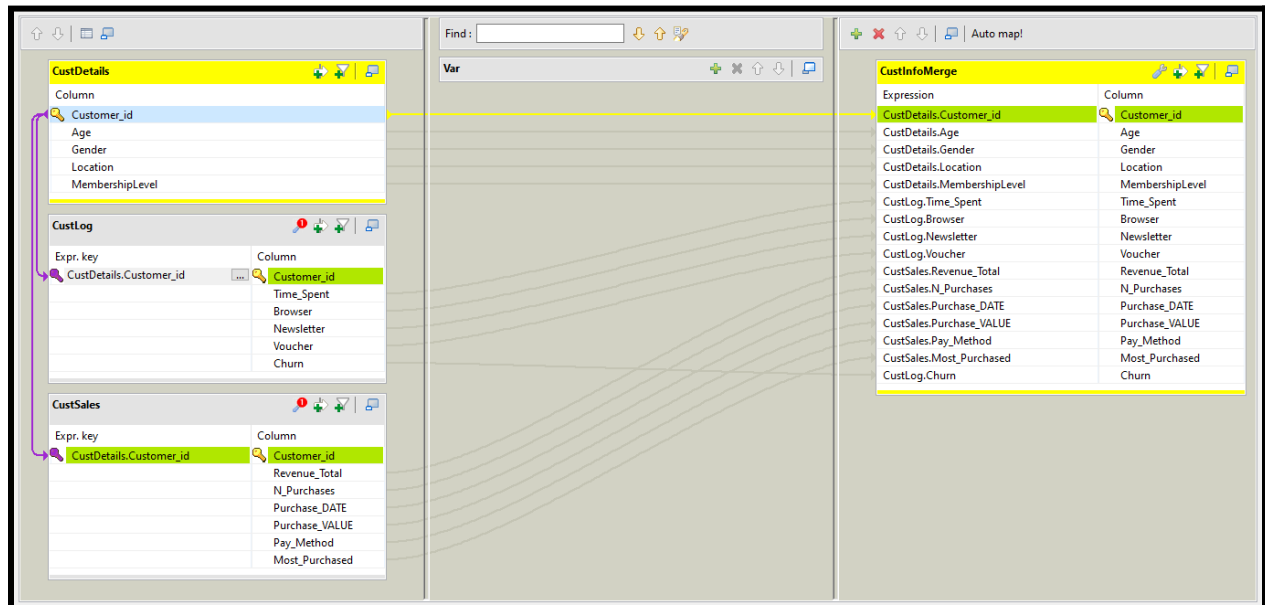
  

CustSales	
Expr. key	Column
CustDetails.Customer_id	Customer_id
	Revenue_Total
	N_Purchases
	Purchase_DATE
	Purchase_VALUE
	Pay_Method
	Most_Purchased

As result, one table is created with those variables, like image below:

CustInfoMerge	
Expression	Column
CustDetails.Customer_id	Customer_id
CustDetails.Age	Age
CustDetails.Gender	Gender
CustDetails.Location	Location
CustDetails.MembershipLevel	MembershipLevel
CustLog.Time_Spent	Time_Spent
CustLog.Browser	Browser
CustLog.Newsletter	Newsletter
CustLog.Voucher	Voucher
CustLog.Churn	Churn
CustSales.Revenue_Total	Revenue_Total
CustSales.N_Purchases	N_Purchases
CustSales.Purchase_DATE	Purchase_DATE
CustSales.Purchase_VALUE	Purchase_VALUE
CustSales.Pay_Method	Pay_Method
CustSales.Most_Purchased	Most_Purchased

This is how the connection looks like:



### 3. tFileOutputDelimited

In Talend Data Integration, tFileOutputDelimited is a component used for writing data to delimited text files. This component is part of the extensive set of tools provided by Talend for Extract, Transform, and Load (ETL) processes. It is primarily used for writing data to delimited text files, such as CSV (Comma-Separated Values) or other custom-delimited formats.

Based on the image below, it presents how the details in Component of one tFileOutputDelimited looks like where we can select the path and name of the CSV file that we want to export and save to local.

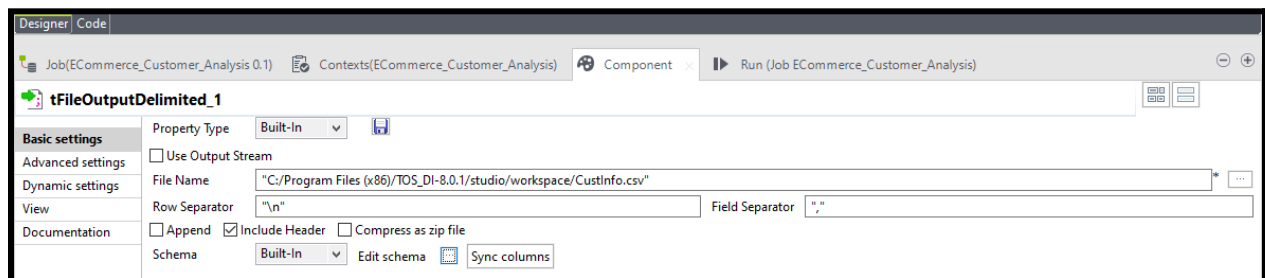
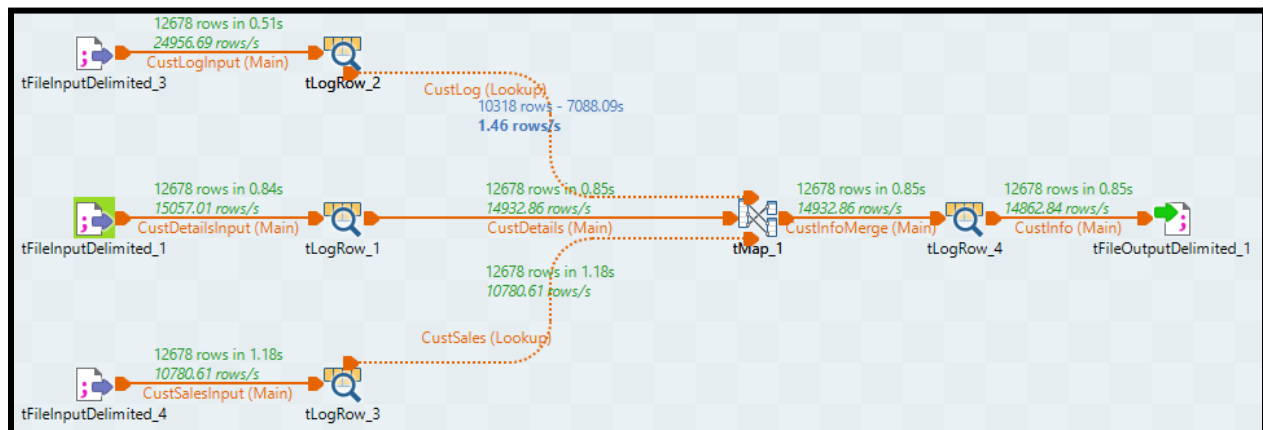


Image below shows overall view in Talend Data Integration:



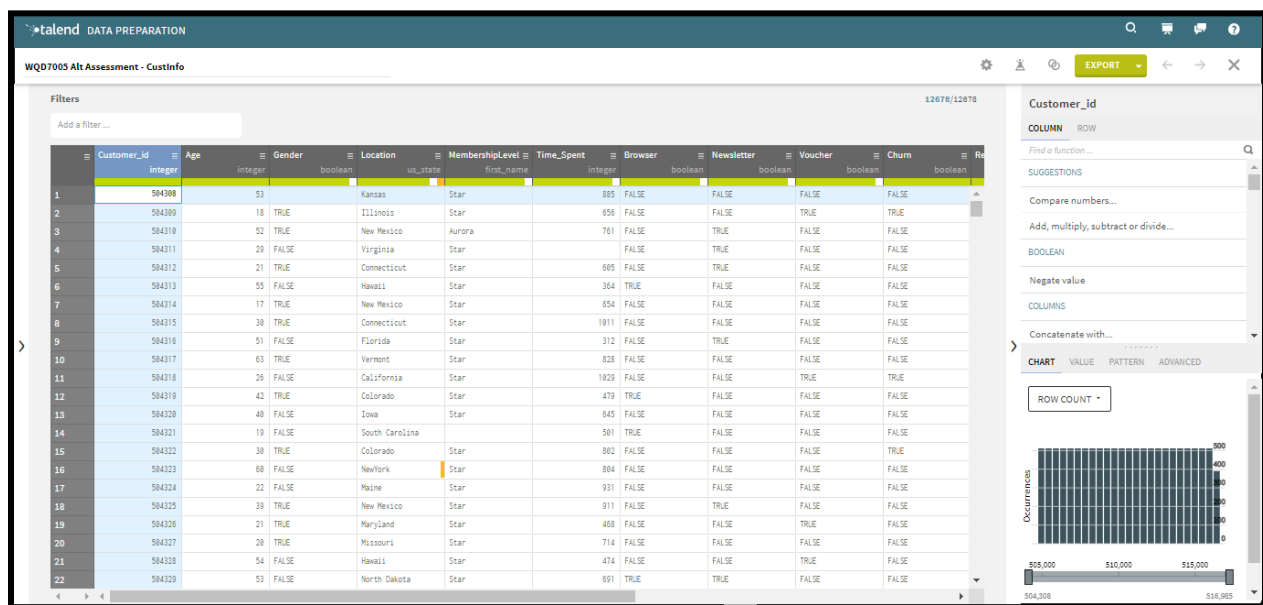
In this particular step, the "tFileOutputDelimited" component is applied by using CustDetails, CustLog and CustSales dataset, then go through "tMap" component to merge those three dataset and come out with one dataset as "customerinfo". This step involves configuring the component to write or transform data from the preceding stages and mapping it to the appropriate structure for the CSV file format. I define the file name, directory, and delimiter, ensuring that the output adheres to the desired specifications. The versatility of "tFileOutputDelimited" enables me to handle column mapping, include headers or footers, and manage various advanced settings to tailor the output file according to specific requirements. This step plays a crucial role in the ETL workflow, as it finalizes the process by persisting the transformed or processed data into the specified CSV file, ready for further analysis, reporting, or sharing. Image below whos the view of CustInfo dataset.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Customer	Age	Gender	Location	Members	Time_Spe	Browser	Newsletters	Voucher	Churn	Revenue	N_Purchase	Purchase	Purchase	Pay_Meth	Most_Purchased					
2	504308	53		Kansas	Star	885	FALSE	FALSE	FALSE	FALSE	45.3	2	22.06.0021	24.915		1	Fresh Fruits				
3	504309	18	TRUE	Illinois	Star	656	FALSE	FALSE	TRUE	TRUE	36.2	3	10.12.0021	2.896		2	Fresh Vegetables				
4	504310	52	TRUE	New Mexi	Aurora	761	FALSE	TRUE	FALSE	FALSE	10.6	1	14.03.0021	10.6		0	Dairy and Eggs				
5	504311	29	FALSE	Virginia	Star		FALSE	TRUE	FALSE	FALSE	54.1	5	25.10.0021	43.28		1	Meat and Poultry				
6	504312	21	TRUE	Connectic	Star	605	FALSE	TRUE	FALSE	FALSE	56.9	1	14.09.0021	56.9			Seafood				
7	504313	55	FALSE	Hawaii	Star	364	TRUE	FALSE	FALSE	FALSE		6	14.05.0021	12.467		1	Bread and Bakery				
8	504314	17	TRUE	New Mexi	Star	654	FALSE	FALSE	FALSE	FALSE	30.7	6	09.01.0021	2.456		0	Pantry Staples				
9	504315	30	TRUE	Connectic	Star	1011	FALSE	FALSE	FALSE	FALSE	8.1	7	28.03.0021	6.561		3	Beverages				
10	504316	51	FALSE	Florida	Star	312	FALSE	TRUE	FALSE	FALSE	18	4	04.08.0021	11.88		0	Snacks and Chips				
11	504317	63	TRUE	Vermont	Star	828	FALSE	FALSE	FALSE	FALSE	19.2	4	06.10.0021	11.904		3	Frozen Foods				
12	504318	26	FALSE	California	Star	1029	FALSE	FALSE	TRUE	TRUE	36.5	5	31.12.0021	31.39		2	Breakfast Foods				
13	504319	42	TRUE	Colorado	Star	479	TRUE	FALSE	FALSE	FALSE	14	4	22.11.0021	4.34		3	Canned Goods				
14	504320	40	FALSE	Iowa	Star	645	FALSE	FALSE	FALSE	FALSE	14.7	2	02.08.0021	2.94		3	Pasta and Grains				
15	504321	19	FALSE	South Carolina		501	TRUE	FALSE	FALSE	FALSE	37.4	4	07.05.0021	15.334		3	Cooking Oils and Condiments				
16	504322	30	TRUE	Colorado	Star	802	FALSE	FALSE	FALSE	TRUE	15.4	1	02.05.0021	15.4		3	Baking Ingredients				
17	504323	60	FALSE	New York	Star	804	FALSE	FALSE	FALSE	FALSE	28.7	7	04.06.0021	0.861		3	Spices and Seasonings				
18	504324	22	FALSE	Maine	Star	931	FALSE	FALSE	FALSE	FALSE	39.7	3	22.02.0021	30.569		2	International Foods				
19	504325	39	TRUE	New Mexi	Star	911	FALSE	TRUE	FALSE	FALSE	5.1	3	13.07.0021	1.53		3	Organic Products				
20	504326	21	TRUE	Maryland	Star	468	FALSE	FALSE	TRUE	FALSE	43.9	6	13.09.0021	19.755		1	Gluten-Free Products				
21	504327	20	TRUE	Missouri	Star	714	FALSE	FALSE	FALSE	FALSE	36.4	6	16.01.0021	14.56		2	Vegan and Plant-Based				
22	504328	54	FALSE	Hawaii	Star	474	FALSE	FALSE	TRUE	FALSE	23.2	1	03.07.0021	23.2		0	Ready-to-Eat Meals				
23	504329	53	FALSE	North Dak	Star	691	TRUE	TRUE	FALSE	FALSE	26.3	7	22.12.0021	15.517		3	Baby and Toddler				

## Talend Data Preprocessing

The initial step in the data processing journey using Talend Data Preparation involves importing the raw data into the platform. This step is crucial as it sets the foundation for subsequent transformations and analyses. Users typically leverage Talend's intuitive interface to connect to various data sources, including databases, spreadsheets, or flat files. Upon establishing the data connection, the imported dataset is visually presented within the Talend Data Preparation environment, providing users with an overview of the raw data's structure and content. This step lays the groundwork for the subsequent stages of data cleaning, enrichment, and transformation, empowering users to seamlessly transition from raw data to a refined and prepared dataset for more advanced analytics and decision-making processes.

Here is the view after we upload the CSV file from Talend Data Integration to Talend Data Preprocessing:





## Invalid - Location

The screenshot shows the Talend Data Preparation interface. A filter is applied to the 'Location' column, selecting rows with invalid values. The data table shows two rows with invalid values in the 'Location' column (rows 16 and 206). The right-hand panel displays the 'Location' column statistics:

Statistic	Value
Count	12678
Distinct	53
Duplicate	12625
Valid	12675
Empty	1
Invalid	2
Avg length	8.45
Min length	0
Max length	14

The screenshot shows the Talend Data Preparation interface with the 'FIND AND GROUP SIMILAR TEXT' dialog box open. The dialog box is used to find and group similar text values. The dialog box contains a table with the following data:

These values have been found	This value will be kept
<input checked="" type="checkbox"/> Arizona	Replace value: Arizona
<input type="checkbox"/> Arkansas	
<input checked="" type="checkbox"/> Arizona	
<input checked="" type="checkbox"/> North Carolina	North Carolina
<input checked="" type="checkbox"/> North Dakota	
<input checked="" type="checkbox"/> Washington	Washington
<input checked="" type="checkbox"/> Wisconsin	

The 'SUBMIT' button is visible at the bottom of the dialog box.

**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
<input checked="" type="checkbox"/> North Carolina	North Carolina
<input checked="" type="checkbox"/> North Dakota	
<input checked="" type="checkbox"/> Washington	Washington
<input checked="" type="checkbox"/> Wisconsin	
<input checked="" type="checkbox"/> New York	Replace value: New York
<input checked="" type="checkbox"/> New York	

**SUBMIT**

**CustInfo PREPARATION**

1. Find and group similar text on column Location

Location: rows with invalid values

Customer_Id	Age	Gender	Location	MembershipLevel	Time_Spent	Browser	Newsletter
1	504300	53	Kansas	Star	885	FALSE	FALSE
2	504300	18	TRUE	Star	656	FALSE	FALSE
3	504310	52	TRUE	Star	761	FALSE	TRUE
4	504311	20	FALSE	Star	605	FALSE	TRUE
5	504312	21	TRUE	Star	364	TRUE	FALSE
6	504313	55	FALSE	Star	654	FALSE	FALSE
7	504314	17	TRUE	Star	1011	FALSE	FALSE
8	504315	30	TRUE	Star	312	FALSE	TRUE
9	504316	51	FALSE	Star	828	FALSE	FALSE
10	504317	63	TRUE	Star	1029	FALSE	FALSE
11	504318	26	FALSE	Star	479	TRUE	FALSE
12	504319	42	TRUE	Star	645	FALSE	FALSE
13	504320	40	FALSE	Star	501	TRUE	FALSE
14	504321	19	FALSE	Star	802	FALSE	FALSE
15	504322	30	TRUE	Star	804	FALSE	FALSE
16	504323	60	FALSE	Star	931	FALSE	FALSE
17	504324	22	FALSE	Star	911	FALSE	TRUE
18	504325	30	TRUE	Star	408	FALSE	FALSE
19	504326	21	TRUE	Star	714	FALSE	FALSE
20	504327	20	TRUE	Star	474	FALSE	FALSE
21	504328	54	FALSE	Star	691	TRUE	TRUE
22	504329	53	FALSE	Star			

Statistics for Location: Count: 12678, Distinct: 51, Duplicate: 12627, Valid: 12677, Empty: 1, Invalid: 0

To rectify the problem of invalid location data, a systematic solution is applied by grouping and modifying the dataset within Talend. Initial identification of records with invalid location entries is followed by grouping the data based on relevant criteria, facilitating a focused approach. Talend's transformation components are then employed to cleanse and rectify the location data within each group. This involves operations such as standardizing formats, filling missing values, or applying geocoding techniques. Rigorous validation and testing are integral to ensuring the effectiveness of the cleansing process, and an iterative approach may be adopted for large or complex datasets. By documenting the changes made throughout this process, the dataset undergoes a structured transformation, resolving the invalid location issues and establishing a reliable foundation for subsequent data analyses.

## Date

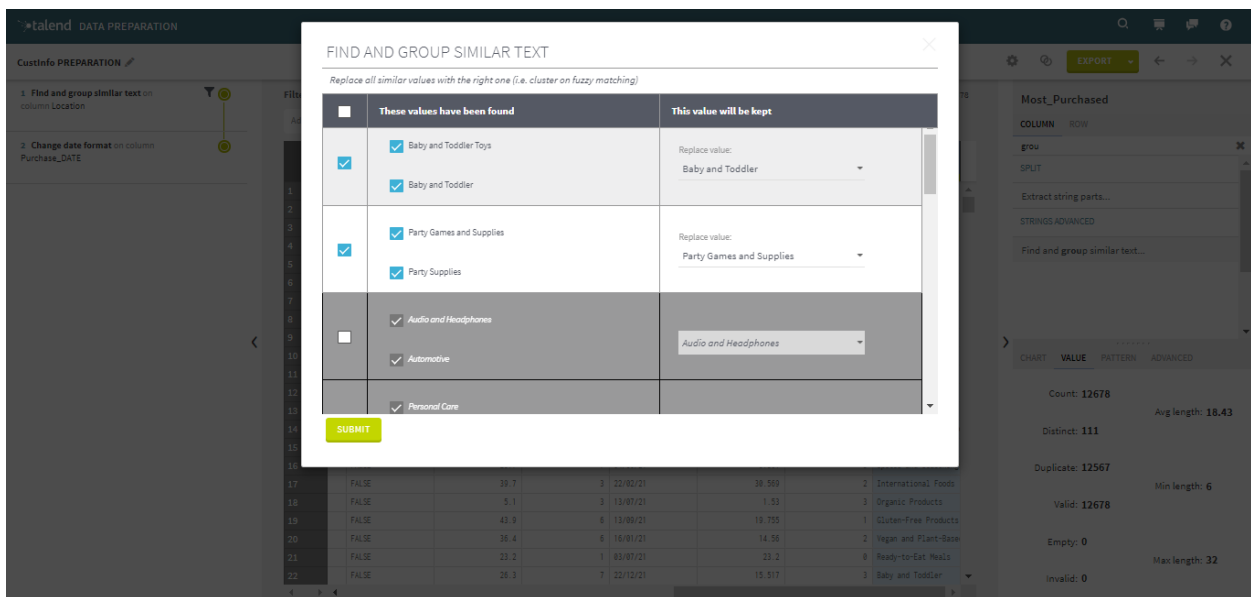
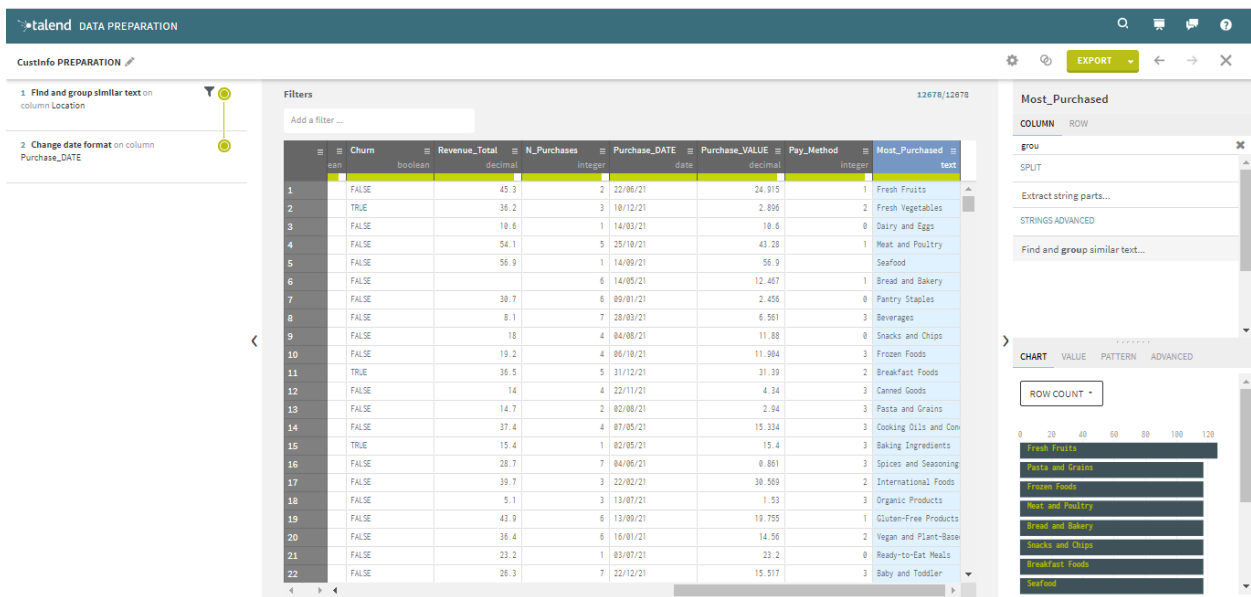
The image displays two screenshots of the Talend Data Preparation interface, illustrating the process of standardizing date formats in a dataset.

**Top Screenshot:** The interface shows a data table with columns: Browser, Newsletter, Voucher, Churn, Revenue\_Total, N\_Purchases, Purchase\_DATE, Purchase\_VALUE, Pay\_Method, and Most\_Purchased. The 'Purchase\_DATE' column contains dates in various formats (e.g., 22.06.0021, 10.12.0021, 14.03.0021). The right sidebar shows the 'Purchase\_DATE' column selected, with a 'Find a function...' search bar and a list of suggestions including 'Calculate time until...', 'Extract date parts...', and 'Change date format...'. The 'ROW COUNT' is displayed as 12678/12678.

**Bottom Screenshot:** The interface shows the same data table, but the 'Purchase\_DATE' column is now standardized to a consistent format (e.g., 22/06/21, 10/12/21, 14/03/21). The right sidebar shows the 'Purchase\_DATE' column selected, with a 'Find a function...' search bar and a list of suggestions including 'Concatenate with...', 'Delete column', and 'Swap columns...'. The 'ROW COUNT' is displayed as 12678/12678.

Following the resolution of invalid location data, the subsequent step involves standardizing the format of the date column within the dataset using Talend. This crucial data preparation step ensures consistency and coherence in date representations. Utilizing these components, users can define the desired date format, convert data types if necessary, and handle any anomalies in the date column. This process guarantees uniformity, making it easier for downstream analyses, reporting, and integration with other systems.

Most\_Purchased



talend DATA PREPARATION

CustInfo PREPARATION

1 Find and group similar text on column Location

2 Change date format on column Purchase\_DATE

**FIND AND GROUP SIMILAR TEXT**

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

<input type="checkbox"/>	These values have been found	This value will be kept
<input type="checkbox"/>	<input checked="" type="checkbox"/> Personal Care <input checked="" type="checkbox"/> Personalized Gifts	Personal Care
<input type="checkbox"/>	<input checked="" type="checkbox"/> Fitness and Exercise <input checked="" type="checkbox"/> Fitness Accessories	Fitness Accessories
<input type="checkbox"/>	<input checked="" type="checkbox"/> Arts and Crafts <input checked="" type="checkbox"/> Arts and Crafts Kits for Kids	Arts and Crafts
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Home Security	Replace value:

SUBMIT

Most\_Purchased

COLUMN ROW

group

SPLIT

Extract string parts...

STRINGS ADVANCED

Find and group similar text...

CHART VALUE PATTERN ADVANCED

Count: 12678 Avg length: 18.43

Distinct: 111

Duplicate: 12567 Min length: 6

Valid: 12678

Empty: 0

Invalid: 0 Max length: 32

talend DATA PREPARATION

CustInfo PREPARATION

1 Find and group similar text on column Location

2 Change date format on column Purchase\_DATE

**FIND AND GROUP SIMILAR TEXT**

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

<input type="checkbox"/>	These values have been found	This value will be kept
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Home Security Systems	Home Security
<input type="checkbox"/>	<input checked="" type="checkbox"/> Beauty and Skincare <input checked="" type="checkbox"/> Bath and Body	Bath and Body
<input type="checkbox"/>	<input checked="" type="checkbox"/> Sewing and Needlecraft <input checked="" type="checkbox"/> Snacks and Chips	Snacks and Chips
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Cleaning Tools and Supplies	Replace value: Cleaning Supplies

SUBMIT

Most\_Purchased

COLUMN ROW

group

SPLIT

Extract string parts...

STRINGS ADVANCED

Find and group similar text...

CHART VALUE PATTERN ADVANCED

Count: 12678 Avg length: 18.43

Distinct: 111

Duplicate: 12567 Min length: 6

Valid: 12678

Empty: 0

Invalid: 0 Max length: 32

**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
<input type="checkbox"/>	<input checked="" type="checkbox"/> Snacks and Chips	Snacks and Chips
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Cleaning Tools and Supplies <input checked="" type="checkbox"/> Cleaning Supplies	Replace value: Cleaning Supplies
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Electronics Accessories <input checked="" type="checkbox"/> Electronics and Appliances	Replace value: Electronics Accessories
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Baby Gear and Nursery <input checked="" type="checkbox"/> Baby Care Products	Replace value: Baby Care Products

**SUBMIT**

**Most\_Purchased**  
 COLUMN ROW  
 SPLIT  
 Extract string parts...  
 STRINGS ADVANCED  
 Find and group similar text...  
 CHART VALUE PATTERN ADVANCED  
 Count: 12678  
 Distinct: 111  
 Duplicate: 12567  
 Valid: 12678  
 Empty: 0  
 Invalid: 0  
 Avg length: 18.43  
 Min length: 6  
 Max length: 32

**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
<input type="checkbox"/>	<input checked="" type="checkbox"/> Baby Care Products	
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Individual Sports <input checked="" type="checkbox"/> Individual Sports	Replace value: Individual Sports
<input type="checkbox"/>	<input checked="" type="checkbox"/> Health and Wellness <input checked="" type="checkbox"/> Health Monitors	Health Monitors
<input type="checkbox"/>	<input checked="" type="checkbox"/> Fresh Vegetables <input checked="" type="checkbox"/> Fresh Fruits	Fresh Fruits

**SUBMIT**

**Most\_Purchased**  
 COLUMN ROW  
 SPLIT  
 Extract string parts...  
 STRINGS ADVANCED  
 Find and group similar text...  
 CHART VALUE PATTERN ADVANCED  
 Count: 12678  
 Distinct: 111  
 Duplicate: 12567  
 Valid: 12678  
 Empty: 0  
 Invalid: 0  
 Avg length: 18.43  
 Min length: 6  
 Max length: 32

The subsequent step involves grouping and modifying the data within the "most\_purchased" column to ensure uniformity in format throughout the dataset using Talend. This process addresses any inconsistencies, variations, or irregularities in the "most\_purchased" column, fostering standardized representations for ease of analysis and interpretation. Whether it involves categorization, renaming, or standardizing values, the objective is to bring cohesion to the "most\_purchased" column. This harmonization facilitates a more seamless integration of the dataset for downstream processes, such as reporting or machine learning, where consistency in data formats is paramount. Through careful grouping and data modification, Talend ensures that the "most\_purchased" column adheres to a standardized format, enhancing the overall quality and coherence of the dataset.

# Gender

talend DATA PREPARATION

1 Find and group similar text on column Location

2 Change date format on column Purchase\_DATE

3 Find and group similar text on column Most\_Purchased

Filters

12678/12678

	Customer_Id	Age	Gender	Location	MembershipLevel	Time_Spent	Browser	Newsletter
	integer	integer	boolean	varchar	first_name	integer	boolean	boolean
1	504300	53		Kansas	Star	885	FALSE	FALSE
2	504309	18	TRUE	Illinois	Star	656	FALSE	FALSE
3	504310	52	TRUE	New Mexico	Aurora	761	FALSE	TRUE
4	504311	29	FALSE	Virginia	Star		FALSE	TRUE
5	504312	21	TRUE	Connecticut	Star	605	FALSE	TRUE
6	504313	55	FALSE	Hawaii	Star	364	TRUE	FALSE
7	504314	17	TRUE	New Mexico	Star	654	FALSE	FALSE
8	504315	38	TRUE	Connecticut	Star	1011	FALSE	FALSE
9	504316	51	FALSE	Florida	Star	312	FALSE	TRUE
10	504317	63	TRUE	Vermont	Star	828	FALSE	FALSE
11	504318	26	FALSE	California	Star	1029	FALSE	FALSE
12	504319	42	TRUE	Colorado	Star	479	TRUE	FALSE
13	504320	40	FALSE	Iowa	Star	645	FALSE	FALSE
14	504321	19	FALSE	South Carolina		501	TRUE	FALSE
15	504322	38	TRUE	Colorado	Star	802	FALSE	FALSE
16	504323	60	FALSE	New York	Star	804	FALSE	FALSE
17	504324	22	FALSE	Maine	Star	931	FALSE	FALSE
18	504325	39	TRUE	New Mexico	Star	911	FALSE	TRUE
19	504326	21	TRUE	Maryland	Star	468	FALSE	FALSE
20	504327	20	TRUE	Missouri	Star	714	FALSE	FALSE
21	504328	54	FALSE	Hawaii	Star	474	FALSE	FALSE
22	504329	53	FALSE	North Dakota	Star	691	TRUE	TRUE

Gender

COLUMN ROW

Find a function...

Delete column

Swap columns...

CONVERSIONS

Convert distance...

Convert duration...

Convert temperature...

DATA CLEANSING

CHART VALUE PATTERN ADVANCED

Count: 12678

Distinct: 3

Duplicate: 12675

Valid: 12677

Empty: 1

Invalid: 0

talend DATA PREPARATION

1 Find and group similar text on column Location

2 Change date format on column Purchase\_DATE

3 Find and group similar text on column Most\_Purchased

4 Replace the cells that match on column Gender

5 Replace the cells that match on column Gender

6 Fill empty cells with text on column Gender

Filters

12678/12678

	Customer_Id	Age	Gender	Location	MembershipLevel	Time_Spent	Browser	Newsletter
	integer	integer	gender	varchar	first_name	integer	boolean	boolean
1	504300	53	Men	Kansas	Star	885	FALSE	FALSE
2	504309	18	Men	Illinois	Star	656	FALSE	FALSE
3	504310	52	Men	New Mexico	Aurora	761	FALSE	TRUE
4	504311	29	Woman	Virginia	Star		FALSE	TRUE
5	504312	21	Men	Connecticut	Star	605	FALSE	TRUE
6	504313	55	Woman	Hawaii	Star	364	TRUE	FALSE
7	504314	17	Men	New Mexico	Star	654	FALSE	FALSE
8	504315	38	Men	Connecticut	Star	1011	FALSE	FALSE
9	504316	51	Woman	Florida	Star	312	FALSE	TRUE
10	504317	63	Men	Vermont	Star	828	FALSE	FALSE
11	504318	26	Woman	California	Star	1029	FALSE	FALSE
12	504319	42	Men	Colorado	Star	479	TRUE	FALSE
13	504320	40	Woman	Iowa	Star	645	FALSE	FALSE
14	504321	19	Woman	South Carolina		501	TRUE	FALSE
15	504322	38	Men	Colorado	Star	802	FALSE	FALSE
16	504323	60	Woman	New York	Star	804	FALSE	FALSE
17	504324	22	Woman	Maine	Star	931	FALSE	FALSE
18	504325	39	Men	New Mexico	Star	911	FALSE	TRUE
19	504326	21	Men	Maryland	Star	468	FALSE	FALSE
20	504327	20	Men	Missouri	Star	714	FALSE	FALSE
21	504328	54	Woman	Hawaii	Star	474	FALSE	FALSE
22	504329	53	Woman	North Dakota	Star	691	TRUE	TRUE

Gender

COLUMN ROW

Find a function...

Concatenate with...

Delete column

Swap columns...

CONVERSIONS

Convert distance...

Convert duration...

Convert temperature...

CHART VALUE PATTERN ADVANCED

ROW COUNT

0 2,000 4,000 6,000 8,000

Men

Woman

To uniformly represent gender in the dataset, Talend's tMap component is utilized. A new output column, "Gender," is created using a ternary expression, assigning "Men" for true and "Women" for false in the specified boolean column. Executing the Talend job applies this transformation across all rows, ensuring consistent gender values by replacing "true" with "Men" and "false" with "Women" in the designated column.

## Fill Empty Cells

The screenshot displays the Talend Data Preparation interface for a job named "CustInfo PREPARATION". The job is configured with 12 steps, including replacing cells that match on column Gender and filling empty cells with text on columns Gender, Location, MembershipLevel, Browser, Newsletter, Pay\_Method, and Voucher. The central table displays 22 rows of customer data with columns: Customer\_id, Age, Gender, Location, MembershipLevel, Time\_Spent, Browser, Newsletter, and Pay\_Method. The right sidebar shows a "Pay\_Method" chart with a bar graph and a "ROW COUNT" table.

Customer_id	Age	Gender	Location	MembershipLevel	Time_Spent	Browser	Newsletter	Pay_Method
1	504300	Men	Kansas	Star	885	FALSE	FALSE	
2	504309	Men	Illinois	Star	656	FALSE	FALSE	
3	504310	Men	New Mexico	Aurora	761	FALSE	TRUE	
4	504311	Woman	Virginia	Star		FALSE	TRUE	
5	504312	Men	Connecticut	Star	685	FALSE	TRUE	
6	504313	Woman	Hawaii	Star	364	TRUE	FALSE	
7	504314	Men	New Mexico	Star	654	FALSE	FALSE	
8	504315	Men	Connecticut	Star	1011	FALSE	FALSE	
9	504316	Woman	Florida	Star	312	FALSE	TRUE	
10	504317	Men	Vermont	Star	828	FALSE	FALSE	
11	504318	Woman	California	Star	1029	FALSE	FALSE	
12	504319	Men	Colorado	Star	479	TRUE	FALSE	
13	504320	Woman	Iowa	Star	645	FALSE	FALSE	
14	504321	Woman	South Carolina	Star	501	TRUE	FALSE	
15	504322	Men	Colorado	Star	802	FALSE	FALSE	
16	504323	Woman	New York	Star	804	FALSE	FALSE	
17	504324	Woman	Maine	Star	931	FALSE	FALSE	
18	504325	Men	New Mexico	Star	911	FALSE	TRUE	
19	504326	Men	Maryland	Star	468	FALSE	FALSE	
20	504327	Men	Missouri	Star	714	FALSE	FALSE	
21	504328	Woman	Hawaii	Star	474	FALSE	FALSE	
22	504329	Woman	North Dakota	Star	691	TRUE	TRUE	



talend DATA PREPARATION																
CustInfo PREPARATION																
Filters																
Add a filter...																
	customer_id	Age	Gender	Location	MembershipLevel	Time_Spent	Browser	Newsletter	Voucher	Churn	Revenue_Total	Num_Purchases	Purchase_DATE	Purchase_VALUE	Pay_Method	Most_Purchased
	integer	integer	gender	us_state	first_name	integer	boolean	boolean	boolean	boolean	decimal	integer	date	decimal		text
1	504008	53	Men	Kansas	Star	855	FALSE	FALSE	FALSE	FALSE	45.3	2	23/06/21	24.915		1 Fresh Fruits
2	504009	18	Men	Illinois	Star	850	FALSE	FALSE	TRUE	TRUE	36.2	3	18/12/21	2.890		2 Fresh Vegetable
3	504010	52	Men	New Mexico	Aurora	791	FALSE	TRUE	FALSE	FALSE	10.0	1	14/03/21	18.0		0 Dairy and Eggs
4	504011	29	Women	Virginia	Star		FALSE	TRUE	FALSE	FALSE	54.1	5	25/08/21	43.28		1 Meat and Poultry
5	504012	21	Men	Connecticut	Star	685	FALSE	TRUE	FALSE	FALSE	56.9	1	14/09/21	56.5		1 Seafood
6	504013	55	Women	Hawaii	Star	354	TRUE	FALSE	FALSE	FALSE		6	14/09/21	12.487		1 Bread and Baker
7	504014	17	Men	New Mexico	Star	654	FALSE	FALSE	FALSE	FALSE	39.7	6	09/01/21	2.456		0 Pantry Staples
8	504015	30	Men	Connecticut	Star	1011	FALSE	FALSE	FALSE	FALSE	8.1	7	28/09/21	6.581		3 Beverages
9	504016	51	Women	Florida	Star	312	FALSE	TRUE	FALSE	FALSE	18	4	04/08/21	11.88		0 Snacks and Chips
10	504017	63	Men	Vermont	Star	828	FALSE	FALSE	FALSE	FALSE	19.2	4	05/08/21	11.904		0 Frozen Foods
11	504018	26	Women	California	Star	1029	FALSE	FALSE	TRUE	TRUE	36.5	5	31/12/21	31.39		2 Breakfast Foods
12	504019	42	Men	Colorado	Star	470	TRUE	FALSE	FALSE	FALSE	14	4	23/11/21	4.54		3 Canned Goods
13	504020	40	Women	Iowa	Star	645	FALSE	FALSE	FALSE	FALSE	14.7	2	02/09/21	2.94		5 Pasta and Grains
14	504021	19	Women	South Carolina	Star	991	TRUE	FALSE	FALSE	FALSE	37.4	4	01/02/21	15.334		3 Cooking Oils and
15	504022	30	Men	Colorado	Star	882	FALSE	FALSE	FALSE	TRUE	15.4	1	02/09/21	15.4		9 Baking Ingredients
16	504023	60	Women	New York	Star	884	FALSE	FALSE	FALSE	FALSE	20.7	7	04/06/21	9.807		0 Spices and Season
17	504024	22	Women	Maine	Star	931	FALSE	FALSE	FALSE	FALSE	39.7	3	23/02/21	39.569		2 International A
18	504025	35	Men	New Mexico	Star	911	FALSE	TRUE	FALSE	FALSE	5.1	3	13/09/21	1.53		3 Organic Product
19	504026	21	Men	Maryland	Star	488	FALSE	FALSE	TRUE	FALSE	43.9	6	13/09/21	19.755		1 Gluten-Free Pro
20	504027	20	Men	Missouri	Star	714	FALSE	FALSE	FALSE	FALSE	36.4	6	16/01/21	14.56		2 Veget and Plant
21	504028	54	Women	Hawaii	Star	474	FALSE	FALSE	TRUE	FALSE	23.2	1	03/01/21	23.2		0 Ready-to-Eat Me
22	504029	53	Women	North Dakota	Star	691	TRUE	TRUE	FALSE	FALSE	26.3	7	23/12/21	15.517		3 Baby and Toddler
23	504030	22	Men	Ohio	Star	394	FALSE	FALSE	TRUE	FALSE	33.1	2	18/01/21	0.862		0 Pet Supplies
24	504031	32	Men	Nebraska	Star	1013	TRUE	FALSE	TRUE	TRUE	1.2	4	20/01/21	0.204		0 Cleaning Supplie
25	504032	30	Men	Montana	Star	105	FALSE	FALSE	FALSE	FALSE	41.3	6	16/04/21	24.367		1 Personal Care
26	504033	24	Men	Indiana	Star	710	FALSE	FALSE	FALSE	FALSE	25.5	3	13/05/21	4.98		0 Health and Welln
27	504034	57	Men	Wisconsin	Star	654	FALSE	FALSE	TRUE	FALSE	34.1	5	01/09/21	18.873		3 Vitamins and Su
28	504035	27	Men	Alabama	Star	888	TRUE	FALSE	FALSE	FALSE	39.8	4	21/01/21	26.18		2 Home and Kitchen
29	504036	21	Women	Alabama	Star	992	FALSE	FALSE	FALSE	FALSE	31.1	2	11/01/21	9.819		0 Office and Scho
30	504037	52	Men	Indiana	Star	311	FALSE	FALSE	TRUE	TRUE	40.1	6	19/08/21	34.151		3 Electronics Acco
31	504038	52	Men	Maine	Star	283	TRUE	TRUE	FALSE	FALSE	42.7	2	04/02/21	22.204		1 Outdoor and Gard
32	504039	52	Men	Montana	Star	941	FALSE	FALSE	FALSE	FALSE	53.5	4	13/12/21	14.99		0 Home Decor
33	504040	55	Women	Arkansas	Star		FALSE	FALSE	FALSE	FALSE	39	2	03/02/21	23.79		1 Clothing and Acc
34	504041	61	Men	Missouri	Star	489	FALSE	FALSE	FALSE	TRUE	39.3	5	16/09/21	5.404		0 Shoes and Footwe
35	504042	17	Women	Pennsylvania	Star	330	TRUE	FALSE	FALSE	TRUE	34.7	1	11/09/21	34.7		1 Beauty and Skin
36	504043	41	Men	New Hampshire	Star	723	FALSE	FALSE	FALSE	FALSE	41.3	6	11/08/21	9.895		1 Hair Care

**SAS Enterprise Miner**

Enterprise Miner - WQD7005AltAsmnt1

File Edit View Actions Options Window Help

WQD7005AltAsmnt1

- Data Sources
- Diagrams
- CaseStudy
- Model Packages

Property Value

Property	Value
<b>General</b>	
Node ID	FIMPORT
Imported Data	
Exported Data	
Notes	
<b>Train</b>	
Variables	
Import File	
Maximum Rows to Import	1000000
Maximum Columns to Import	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0
Guessing Rows	500
File Location	Local
File Type	csv
Advanced Advisor	No
Run	No
<b>Score</b>	
Role	Train
<b>Report</b>	
Summarize	No
<b>Status</b>	
Create Time	1/7/24 9:28 AM
Run ID	

General

Diagram CaseStudy opened

wie190006@siswa.um.edu.my as u62569858 Connected to SASApp - Logical Workspace Server (odaws01-apse1.oda.sas.com)

File Import

Select My Computer if the data file you want to import is located on your local machine. Select SAS Server to import a data file located on your SAS workspace server.

☒ My Computer

☐ SAS Servers

D:\Download D\CustInfo PREPARATION.csv

Browse...

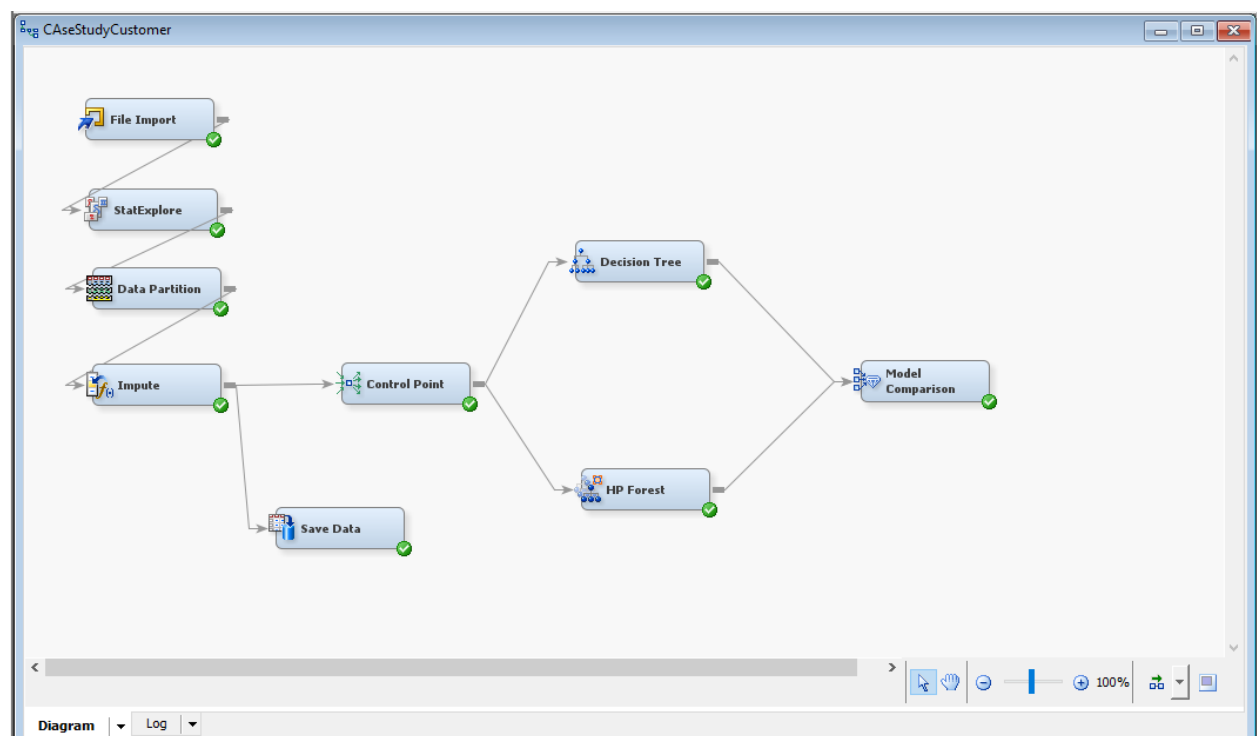
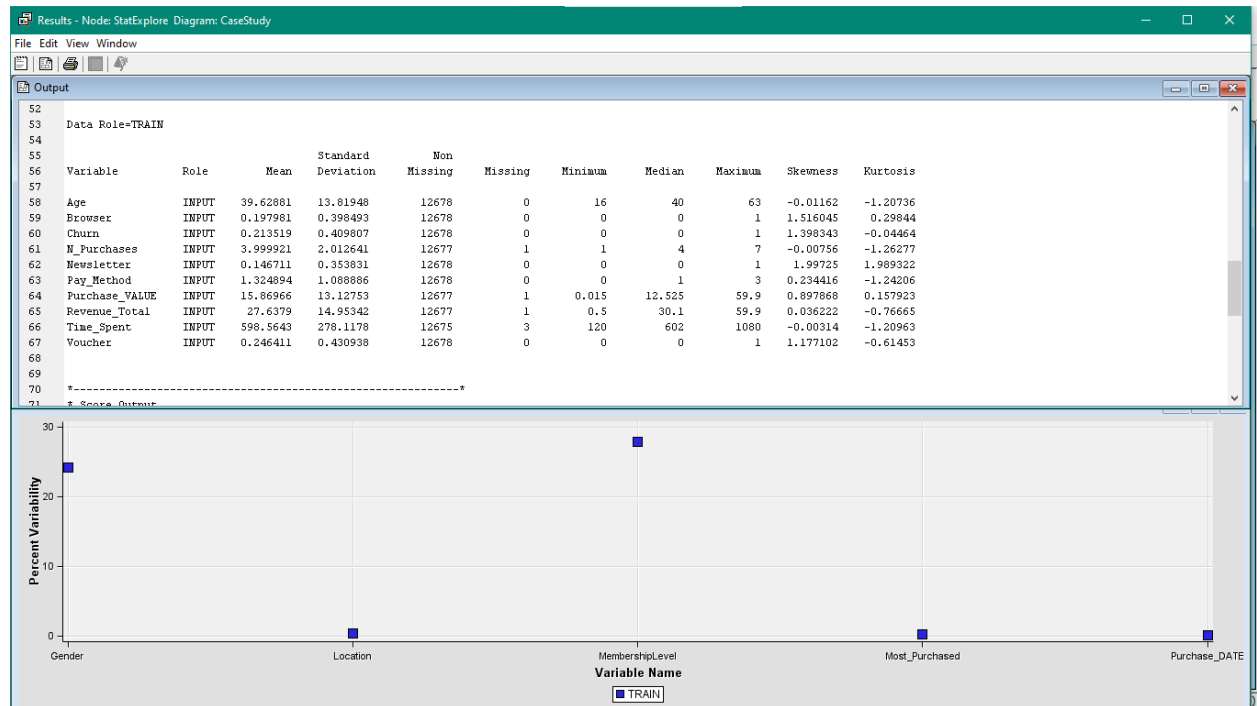
View File Import Types Preview OK Cancel

Results - Node: File Import Diagram: CaseStudy

File Edit View Window

Output

```
40
41 Data Set Page Size 131072
42 Number of Data Set Pages 16
43 First Data Page 1
44 Max Obs per Page 818
45 Obs in First Data Page 792
46 Number of Data Set Repairs 0
47 Filename /home/u62569858/WQD7005AltAsmnt1/Workspaces/ENWS1/fimport_data.sas7bdat
48 Release Created 9.0401M7
49 Host Created Linux
50 Inode Number 224789474
51 Access Permission rw-r--r--
52 Owner Name u62569858
53 File Size 2MB
54 File Size (bytes) 2282224
55
56
57 Alphabetic List of Variables and Attributes
58
59 # Variable Type Len Format Informat Label
60
61 2 Age Num 8 BEST. Age
62 7 Browser Num 8 BEST. Browser
63 10 Churn Num 8 BEST. Churn
64 1 Customer_id Num 8 BEST. Customer_id
65 3 Gender Char 5 $5. Gender
66 4 Location Char 14 $14. $14. Location
67 5 MembershipLevel Char 6 $6. $6. MembershipLevel
68 16 Most_Purchased Char 32 $32. $32. Most_Purchased
69 12 N_Purchases Num 8 BEST. N_Purchases
70 8 Newsletter Num 8 BEST. Newsletter
71 15 Pay_Method Num 8 BEST. Pay_Method
72 13 Purchase_DATE Char 8 $8. $8. Purchase_DATE
73 14 Purchase_VALUE Num 8 BEST. Purchase_VALUE
74 11 Revenue_Total Num 8 BEST. Revenue_Total
75 6 Time_Spent Num 8 BEST. Time_Spent
76 9 Voucher Num 8 BEST. Voucher
77
78
```



Variables - FIMPORT

(none)

☐ not

Equal to

...

Apply

Reset

Columns:

☐ Label

☐ Mining

☐ Basic

☐ Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Browser	Input	Interval	No		No	.	.
Churn	Input	Interval	No		No	.	.
Customer_id	ID	Nominal	No		No	.	.
Gender	Input	Nominal	No		No	.	.
Location	Input	Nominal	No		No	.	.
MembershipLevel	Input	Nominal	No		No	.	.
Most_Purchased	Input	Nominal	No		No	.	.
N_Purchases	Input	Interval	No		No	.	.
Newsletter	Input	Interval	No		No	.	.
Pay_Method	Input	Interval	No		No	.	.
Purchase_DATE	Input	Nominal	No		No	.	.
Purchase_VALUE	Input	Interval	No		No	.	.
Revenue_Total	Target	Interval	No		No	.	.
Time_Spent	Input	Interval	No		No	.	.
Voucher	Input	Interval	No		No	.	.

Explore...

OK

Cancel