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

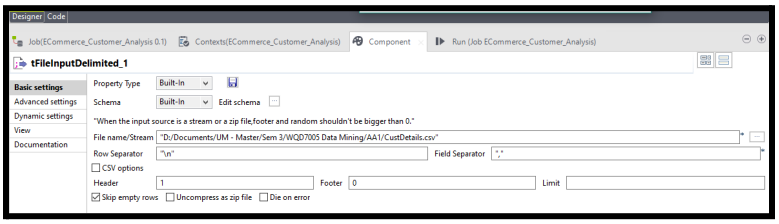
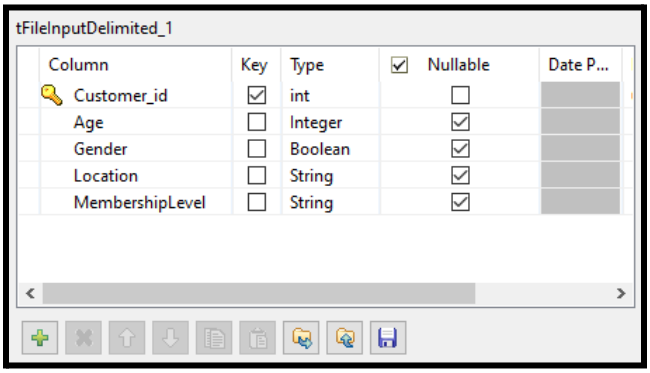
Talend Data Integration is a comprehensive open-source tool used for Extract, Transform, Load (ETL) processes. Once the data is imported into Talend, users can apply a variety of data integration tasks, including data cleansing, transformation, and loading into target systems. The importation of data 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.

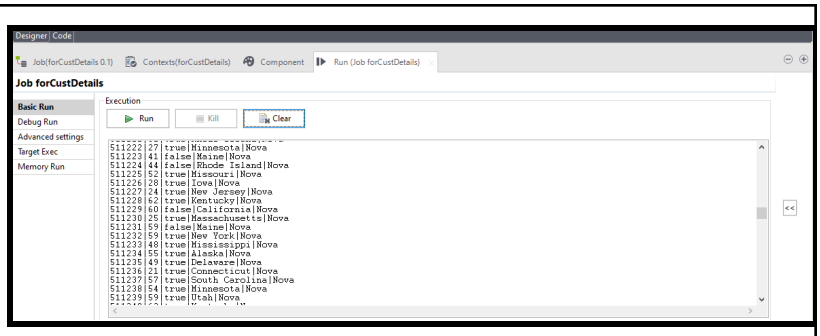
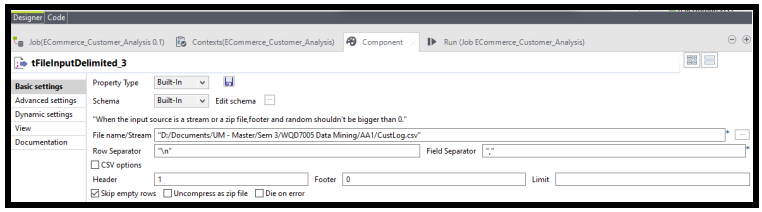
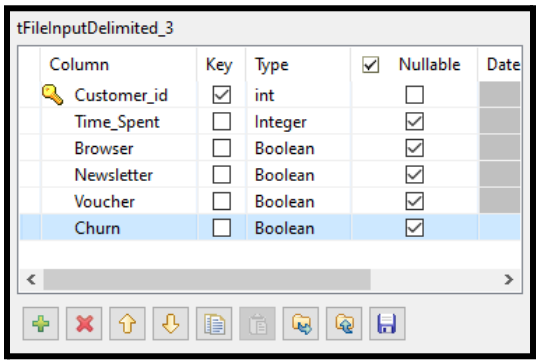
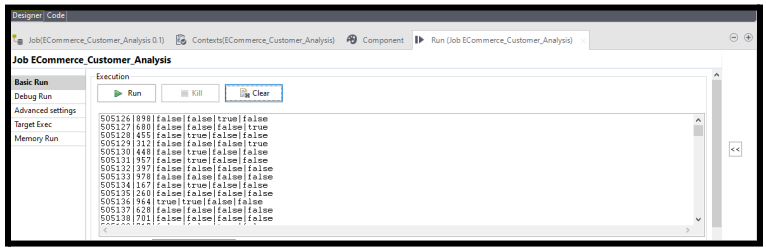
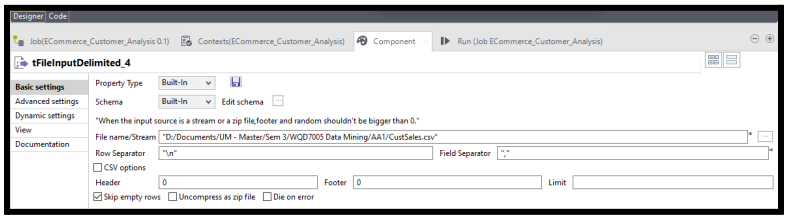
I used multiple nodes to complete this integration, which are:

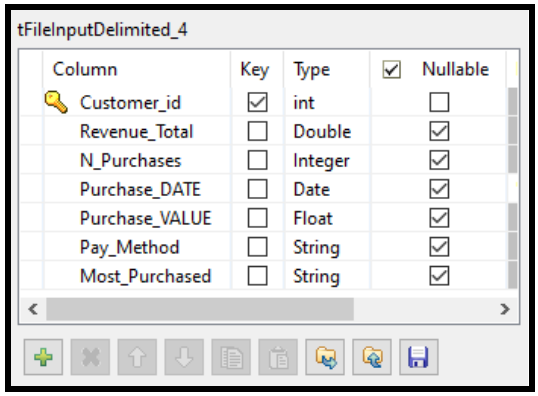
1. **tFileInputDelimited**

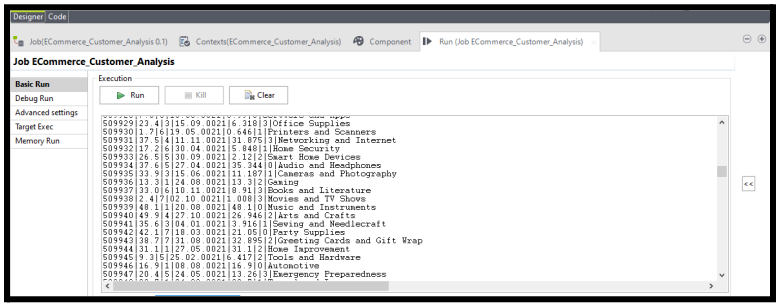
   A component used for reading data from delimited text files. It is part of the input/output (I/O) family of components and is specifically designed to handle delimited files, where data is separated by a specified delimiter, such as a comma, tab, semicolon, etc.

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

No	Dataset Name	Evidance
1	CustDetails	 

		
2	CustLog	  
3	CustSales	

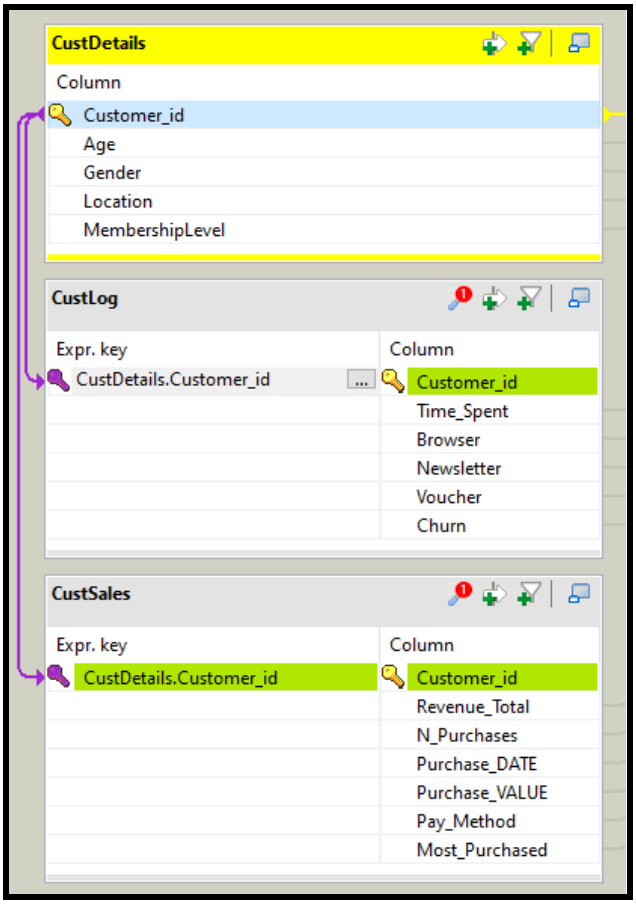




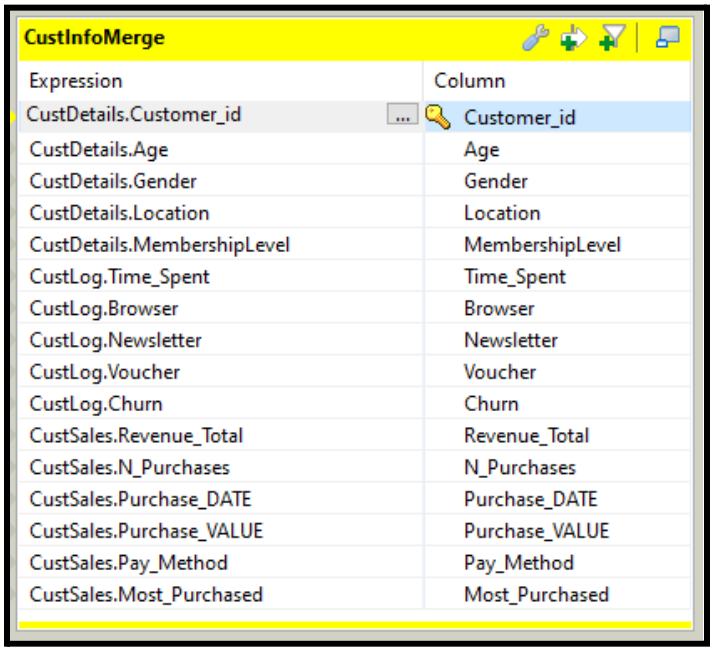
2. **tMap1**

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.

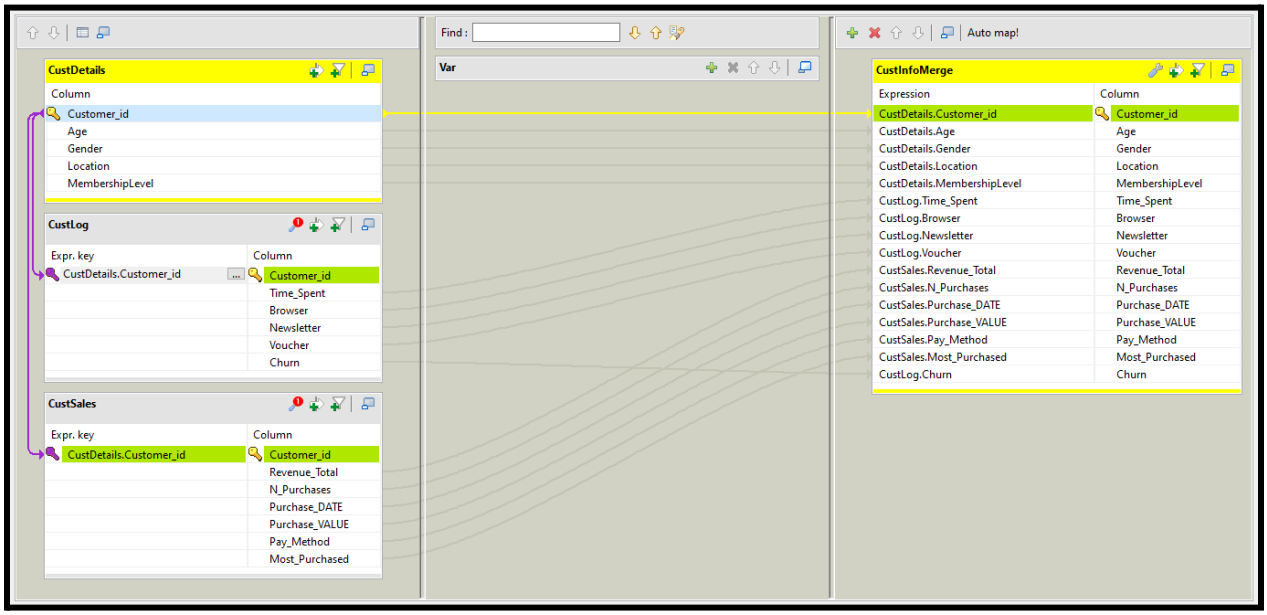
In this case study, those three datasets are being connected with Customer_id as it is the primary keys for all three datasets. Refer to image below:



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



This i s 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.

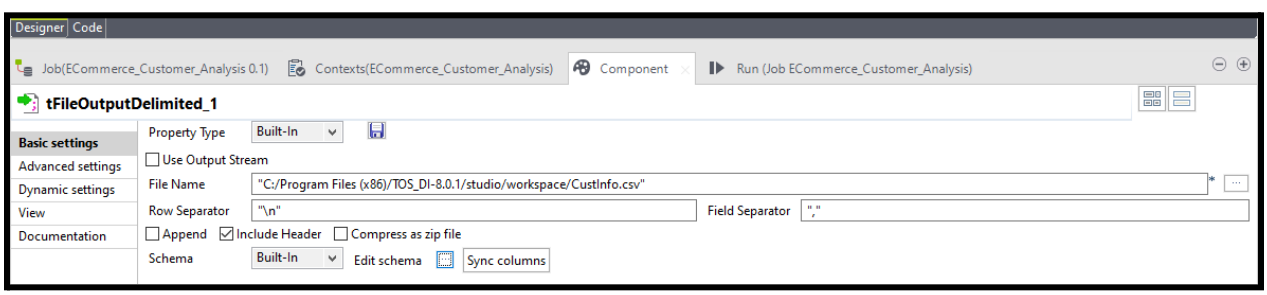
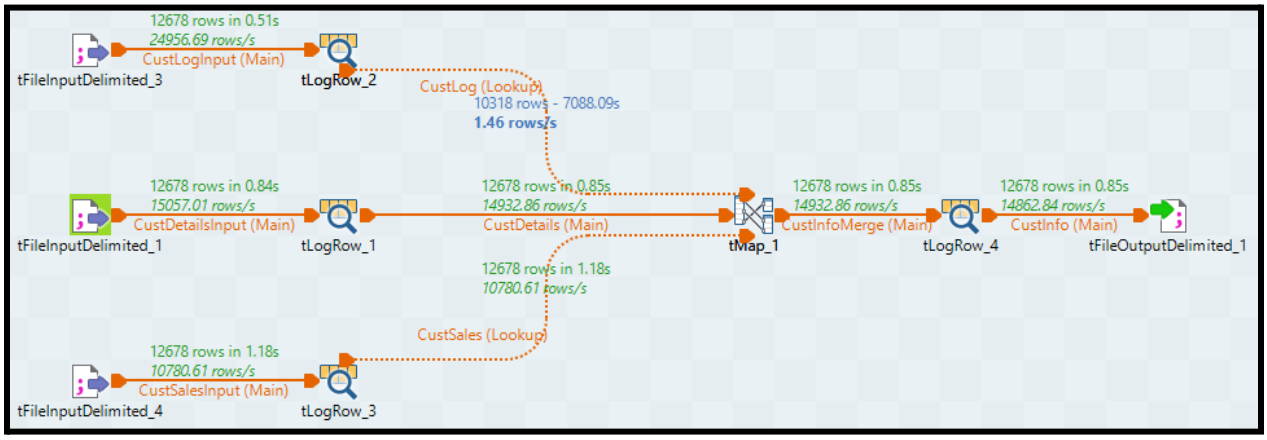
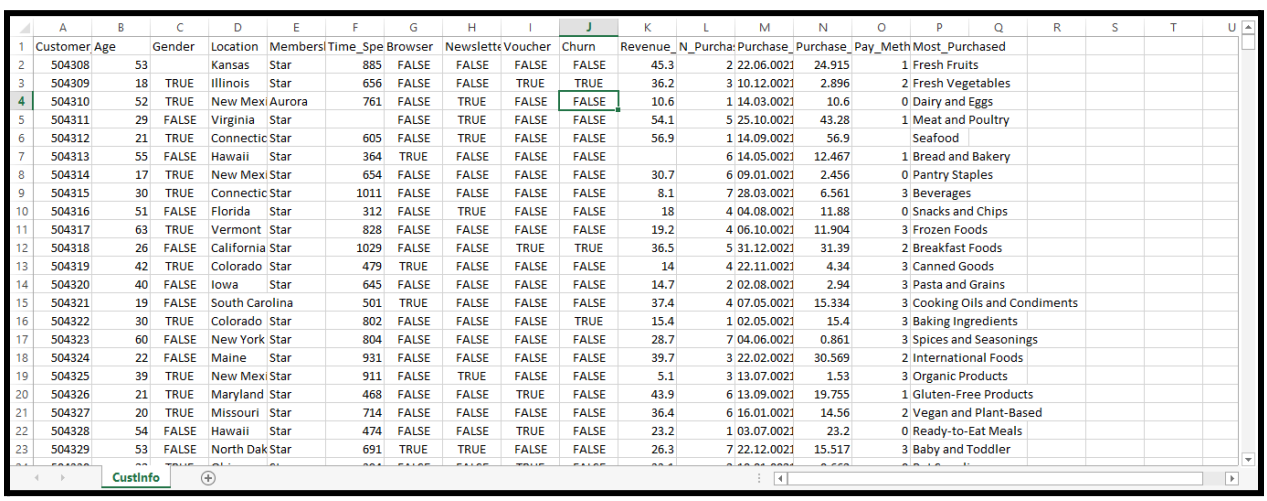


Image below shows overall view in Talend Data Integration:



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.



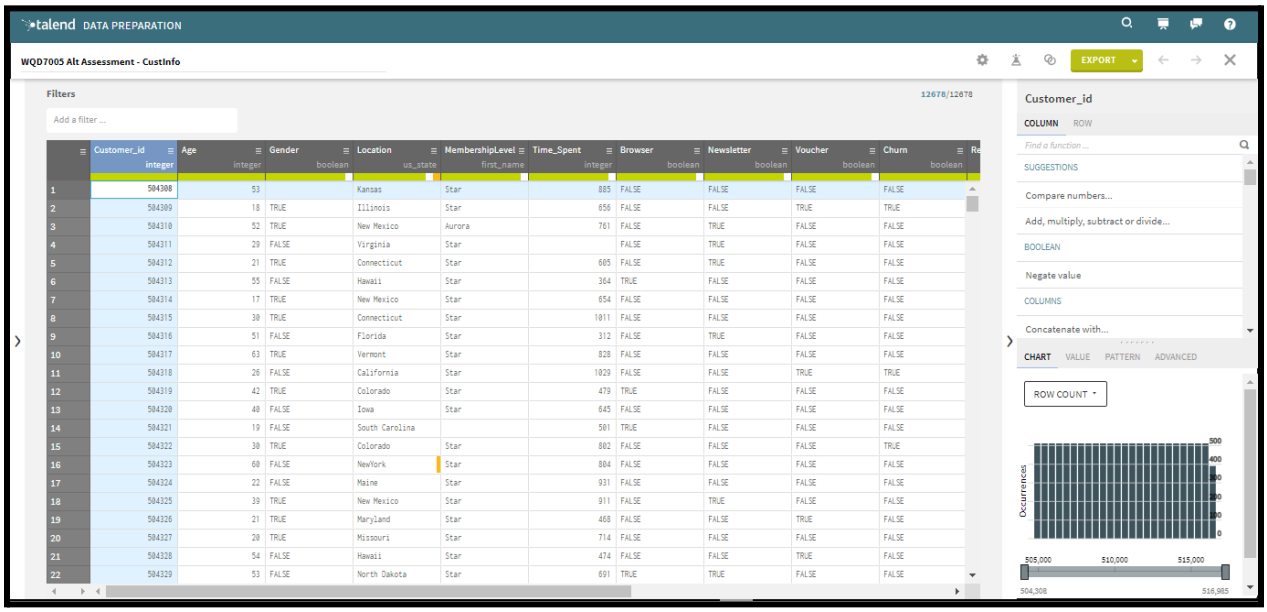
Customer	Age	Gender	Location	Members	Time_Spe	Browser	Newslette	Voucher	Churn	Revenue_	N_Purcha	Purchase_	Purchase_	Pay_Meth	Most_Purchased
504308	53		Kansas	Star	885	FALSE	FALSE	FALSE	FALSE	45.3	2	22.06.0021	24.915	1	Fresh Fruits
504309	18	TRUE	Illinois	Star	656	FALSE	FALSE	TRUE	TRUE	36.2	3	10.12.0021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

CustInfo

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.

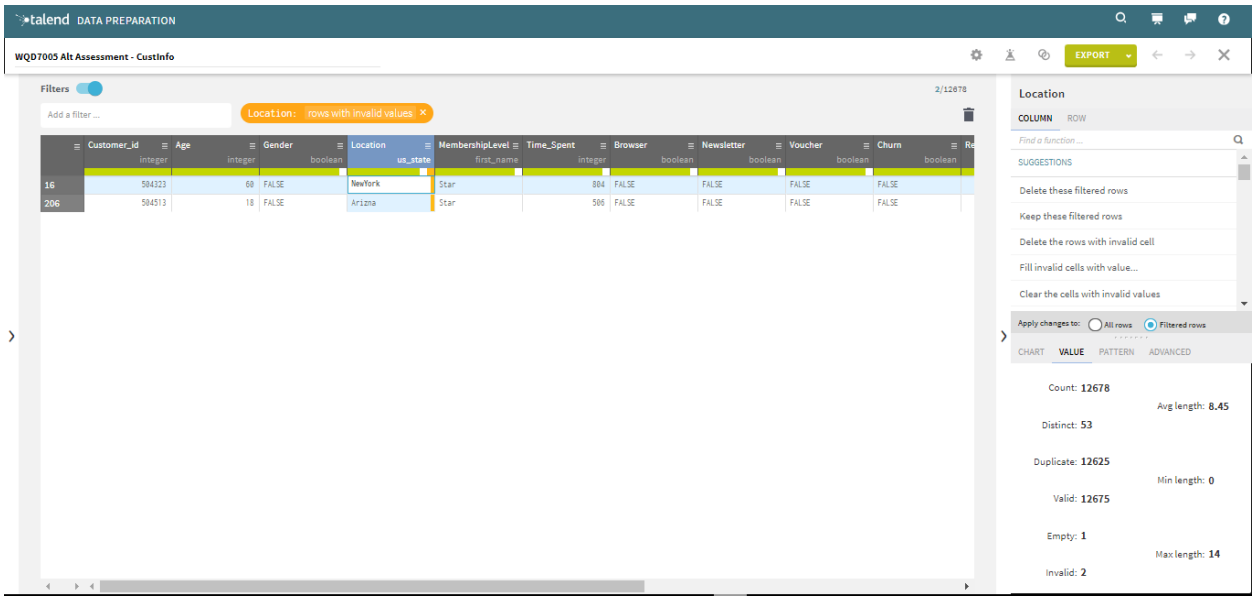
Here is the view after we upload the CSV file from Talend Data Integration to Talend Data Preprocessing:

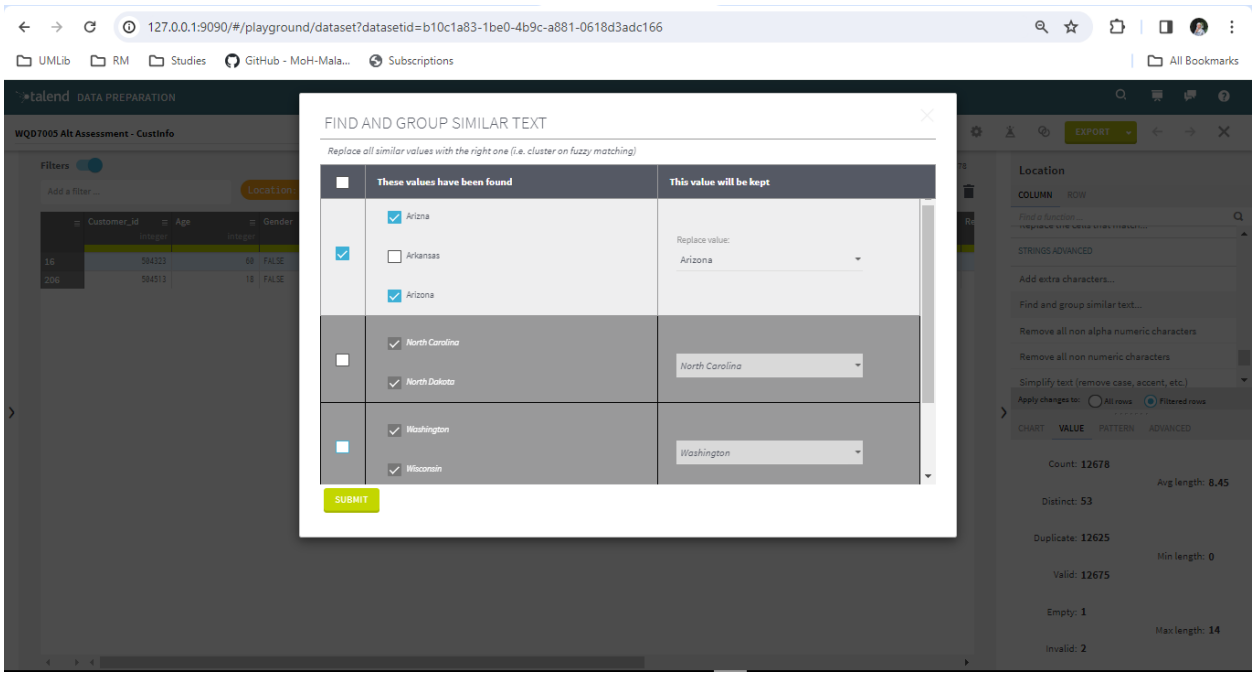


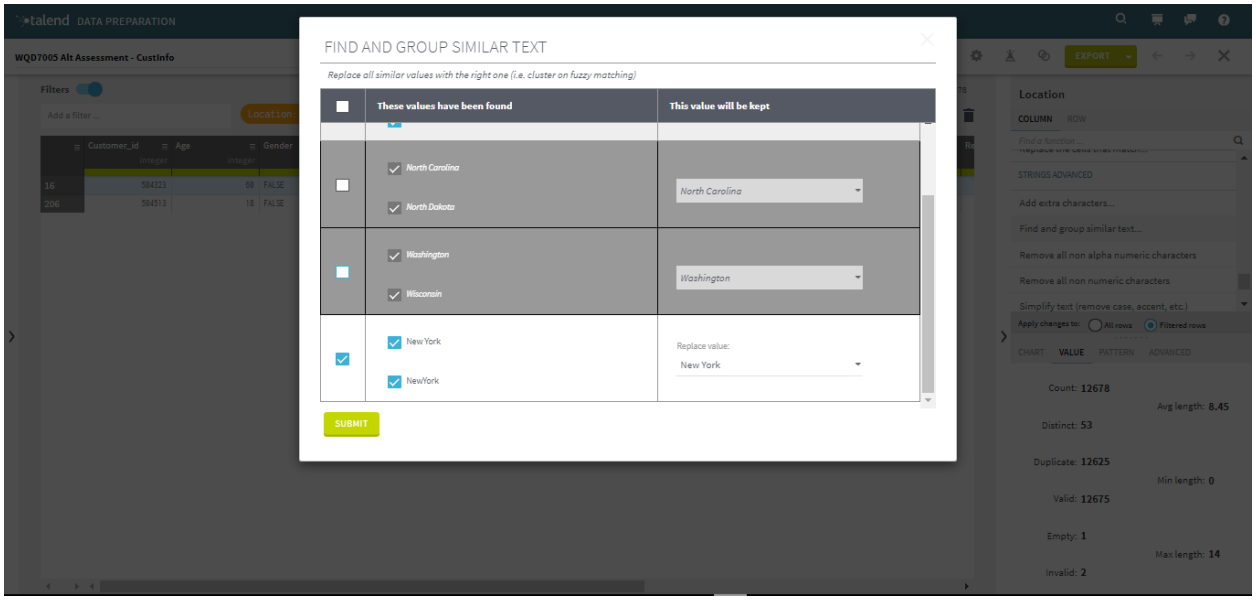
	Customer_id	Age	Gender	Location	MembershipLevel	Time_Spent	Browser	Newsletter	Voucher	Churn
	integer	integer	boolean	us_state	first_name	integer	boolean	boolean	boolean	boolean
1	504308	53		Kansas	Star	885	FALSE	FALSE	FALSE	FALSE
2	504309	18	TRUE	Illinois	Star	656	FALSE	FALSE	TRUE	TRUE
3	504310	52	TRUE	New Mexico	Aurora	761	FALSE	TRUE	FALSE	FALSE
4	504311	29	FALSE	Virginia	Star		FALSE	TRUE	FALSE	FALSE
5	504312	21	TRUE	Connecticut	Star	605	FALSE	TRUE	FALSE	FALSE
6	504313	55	FALSE	Hawaii	Star	364	TRUE	FALSE	FALSE	FALSE
7	504314	17	TRUE	New Mexico	Star	654	FALSE	FALSE	FALSE	FALSE
8	504315	30	TRUE	Connecticut	Star	1011	FALSE	FALSE	FALSE	FALSE
9	504316	51	FALSE	Florida	Star	312	FALSE	TRUE	FALSE	FALSE
10	504317	63	TRUE	Vermont	Star	828	FALSE	FALSE	FALSE	FALSE
11	504318	26	FALSE	California	Star	1029	FALSE	FALSE	TRUE	TRUE
12	504319	42	TRUE	Colorado	Star	479	TRUE	FALSE	FALSE	FALSE
13	504320	40	FALSE	Iowa	Star	645	FALSE	FALSE	FALSE	FALSE
14	504321	19	FALSE	South Carolina		501	TRUE	FALSE	FALSE	FALSE
15	504322	30	TRUE	Colorado	Star	802	FALSE	FALSE	FALSE	TRUE
16	504323	60	FALSE	NewYork	Star	804	FALSE	FALSE	FALSE	FALSE
17	504324	22	FALSE	Maine	Star	931	FALSE	FALSE	FALSE	FALSE
18	504325	39	TRUE	New Mexico	Star	911	FALSE	TRUE	FALSE	FALSE
19	504326	21	TRUE	Maryland	Star	468	FALSE	FALSE	TRUE	FALSE
20	504327	20	TRUE	Missouri	Star	714	FALSE	FALSE	FALSE	FALSE
21	504328	54	FALSE	Hawaii	Star	474	FALSE	FALSE	TRUE	FALSE
22	504329	53	FALSE	North Dakota	Star	691	TRUE	TRUE	FALSE	FALSE

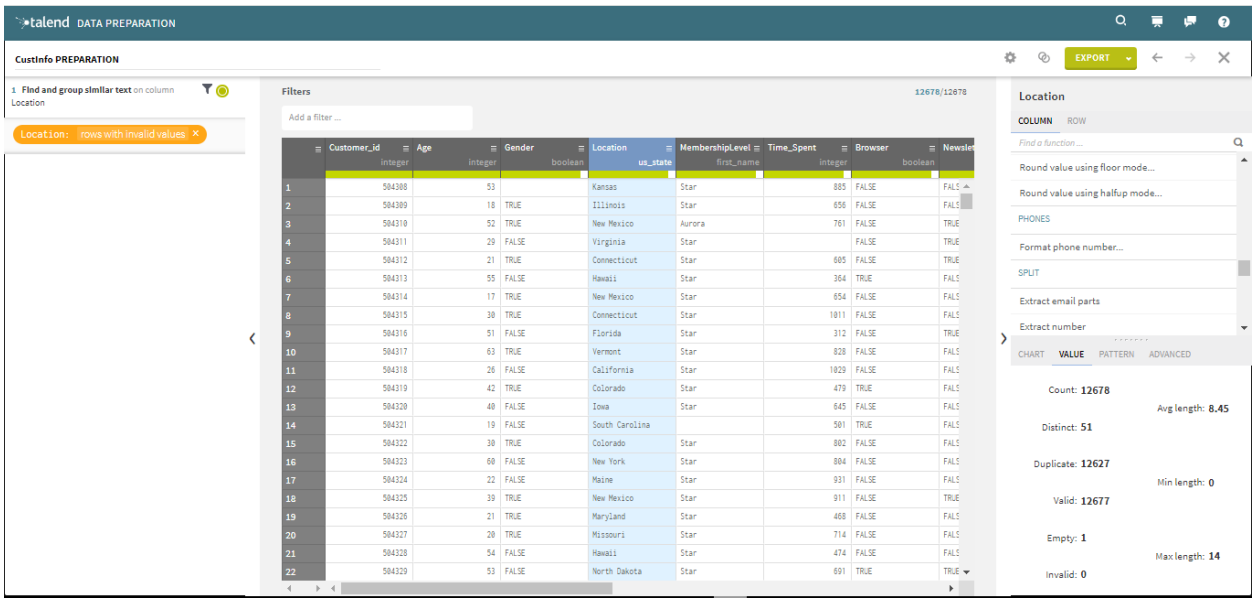
To make the dataset become more valuable and precise, there are few steps need to be done for some field, which are:

1. Location



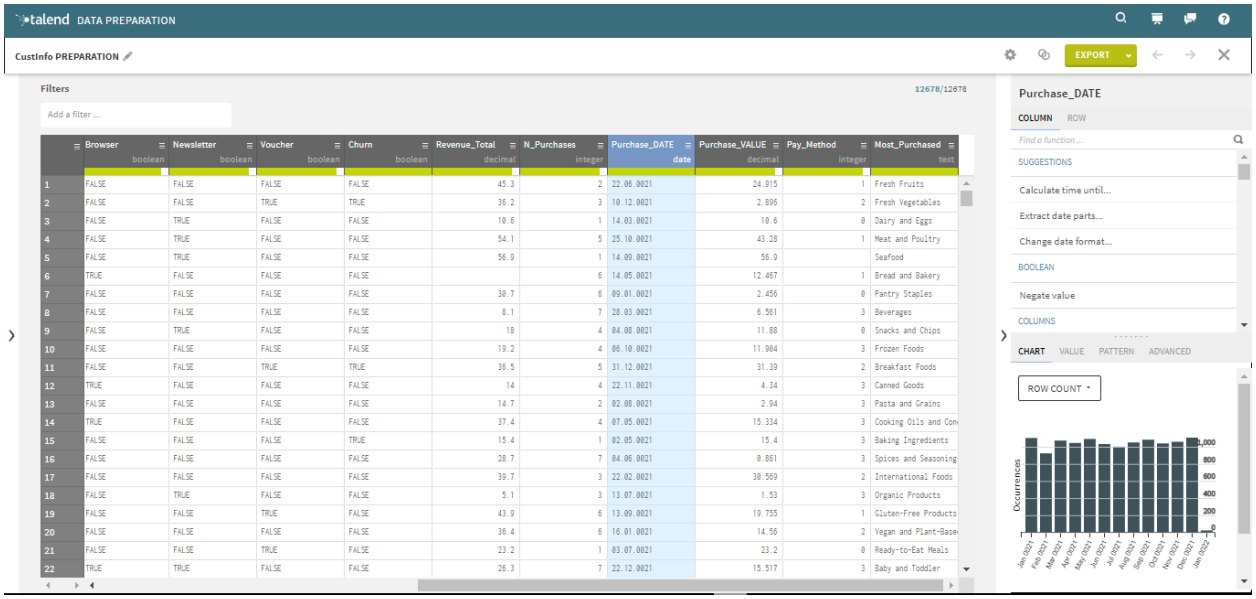


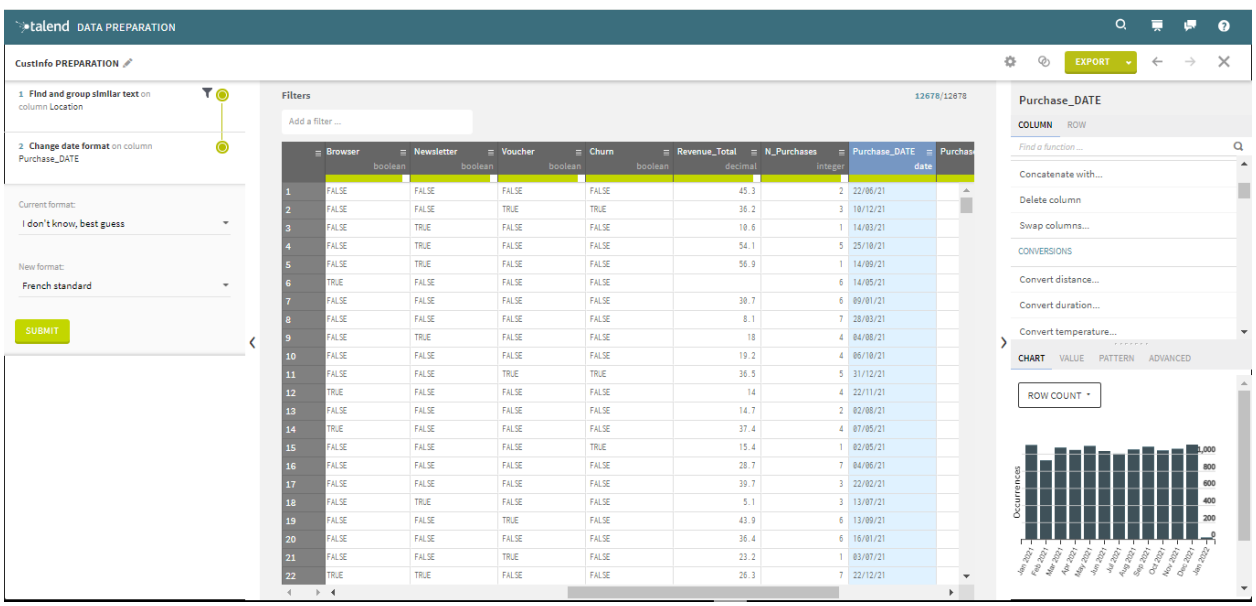




To rectify the problem of invalid location data, a systematic solution is applied by grouping and modifying the dataset within Talend. Initial identification of records with invalid location entries is followed by grouping the data based on relevant criteria, facilitating a focused approach. Talend's 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.

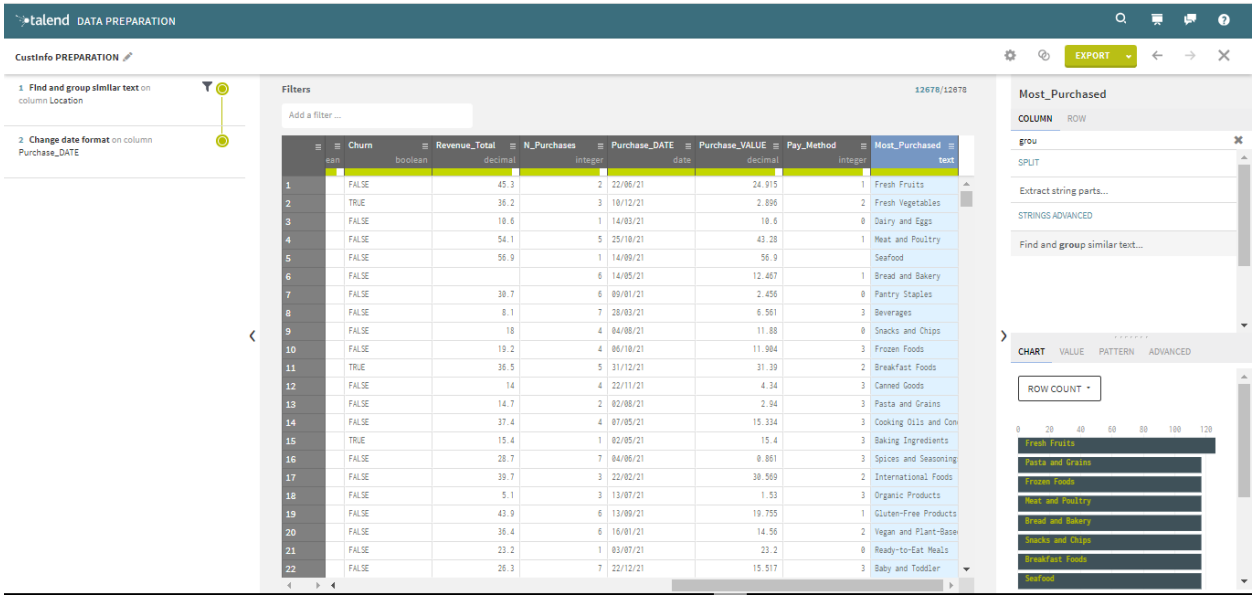
2. Date

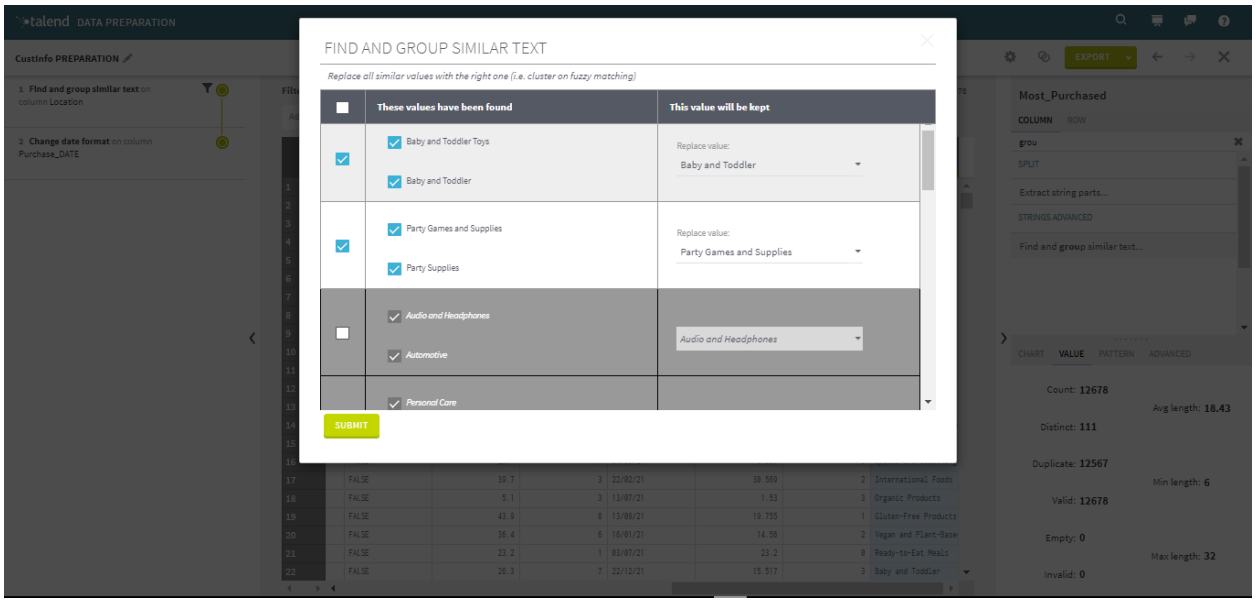


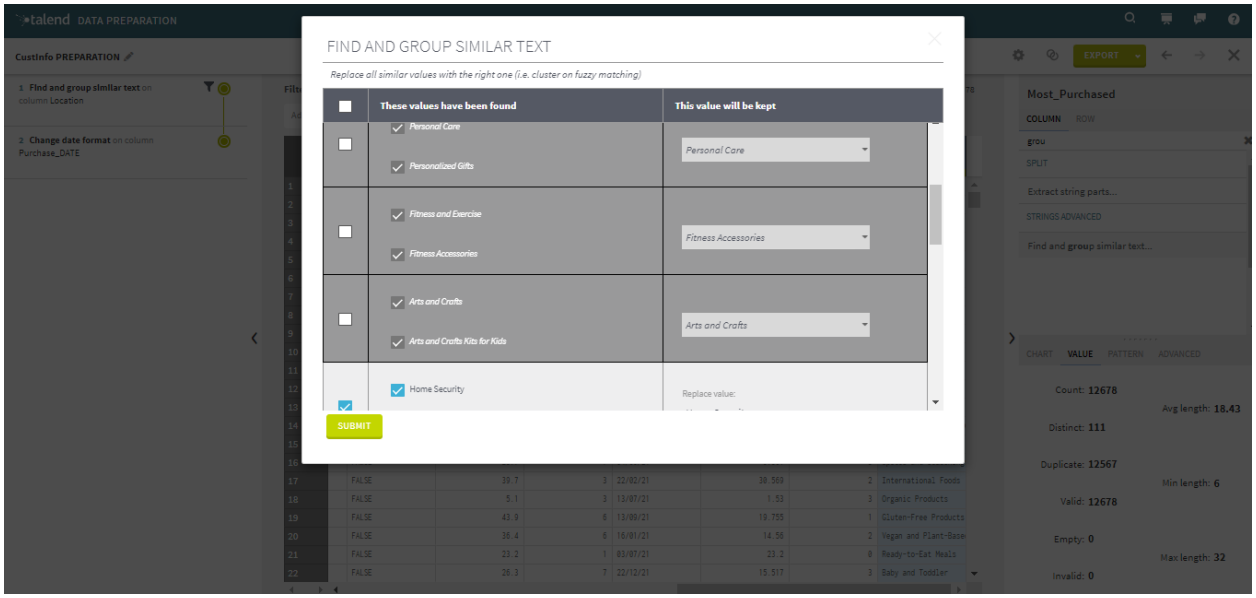


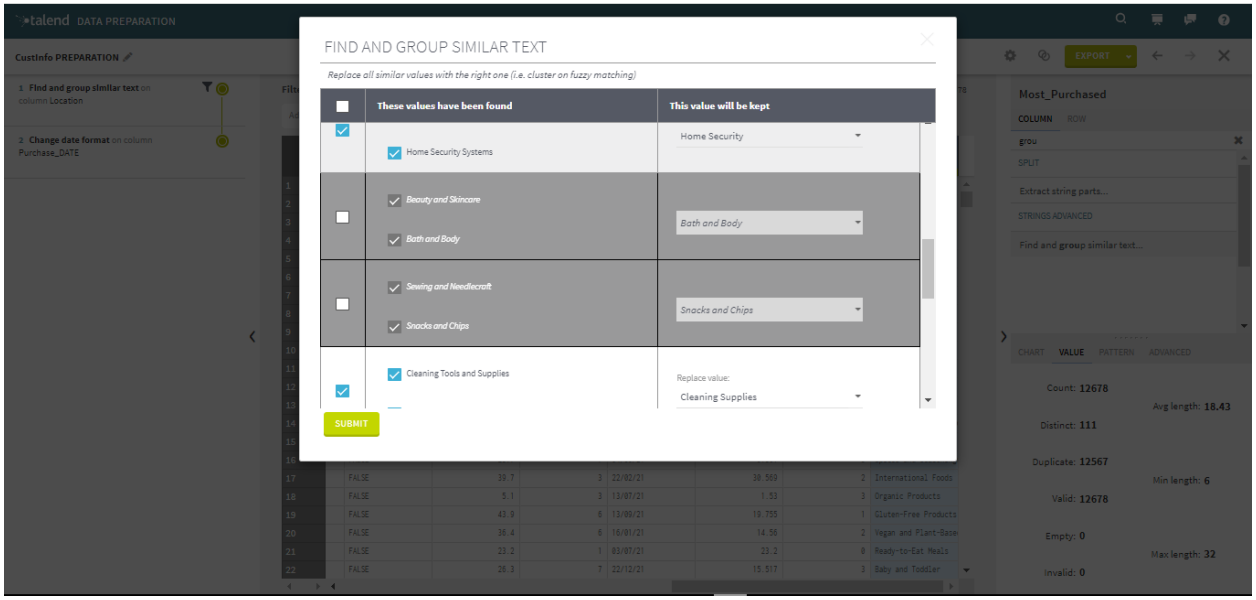
Following the resolution of invalid location data, the subsequent step involves standardizing the format of the date column within the dataset using Talend. This crucial data preparation step ensures consistency and coherence in date representations. Utilizing these components, users can define the desired 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.

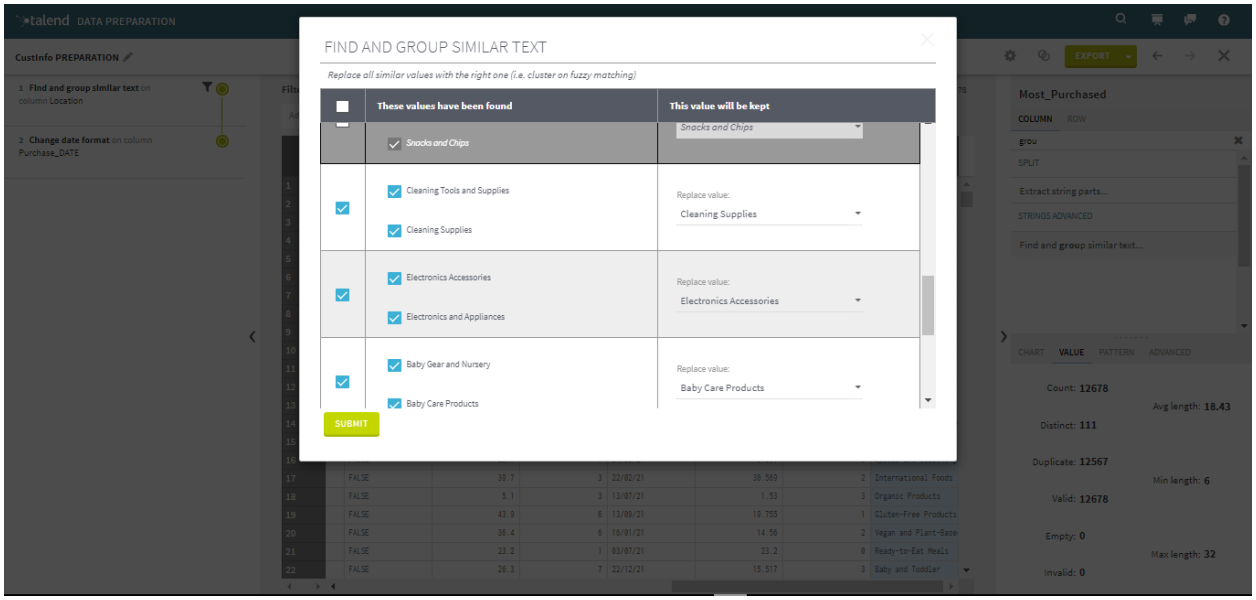
3. Most Purchased

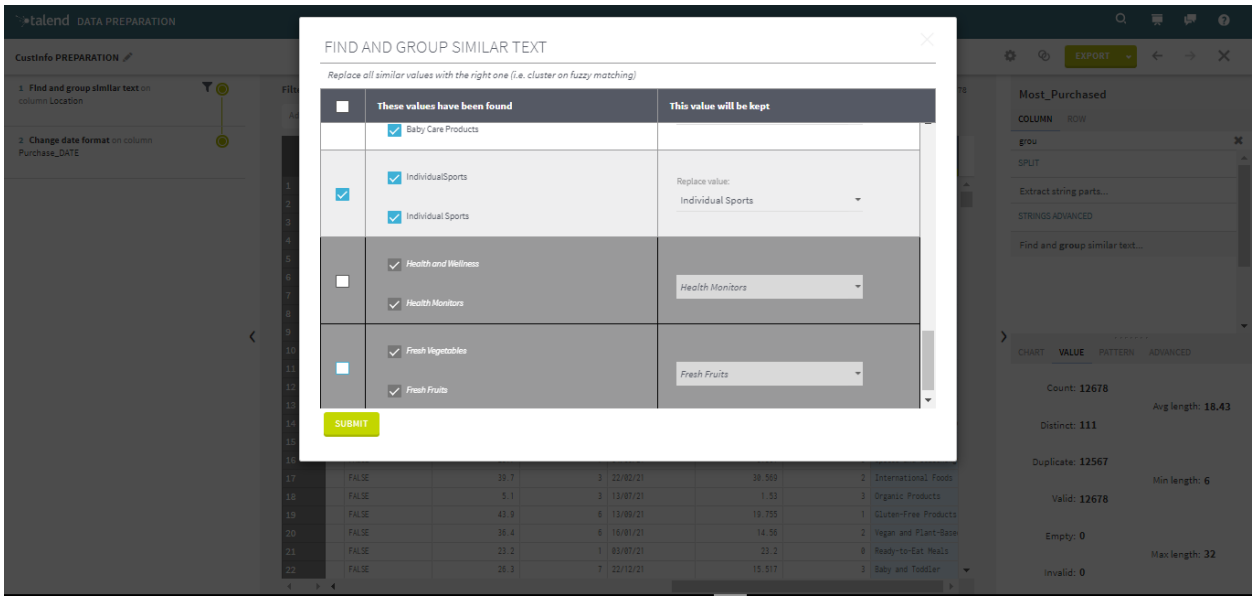





talend DATA PREPARATION
CustInfo PREPARATION
1 Find and group similar text on column Location
2 Change date format on column Purchase_DATE
FIND AND GROUP SIMILAR TEXT
Replace all similar values with the right one (i.e. cluster on fuzzy matching)
These values have been found
This value will be kept
Personal Care
Personalized Gifts
Personal Care
Fitness and Exercise
Fitness Accessories
Fitness Accessories
Arts and Crafts
Arts and Crafts Kits for Kids
Arts and Crafts
Home Security
Replace value:
SUBMIT
EXPORT
Most_Purchased
COLUMN
ROW
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
Max length: 32
Invalid: 0

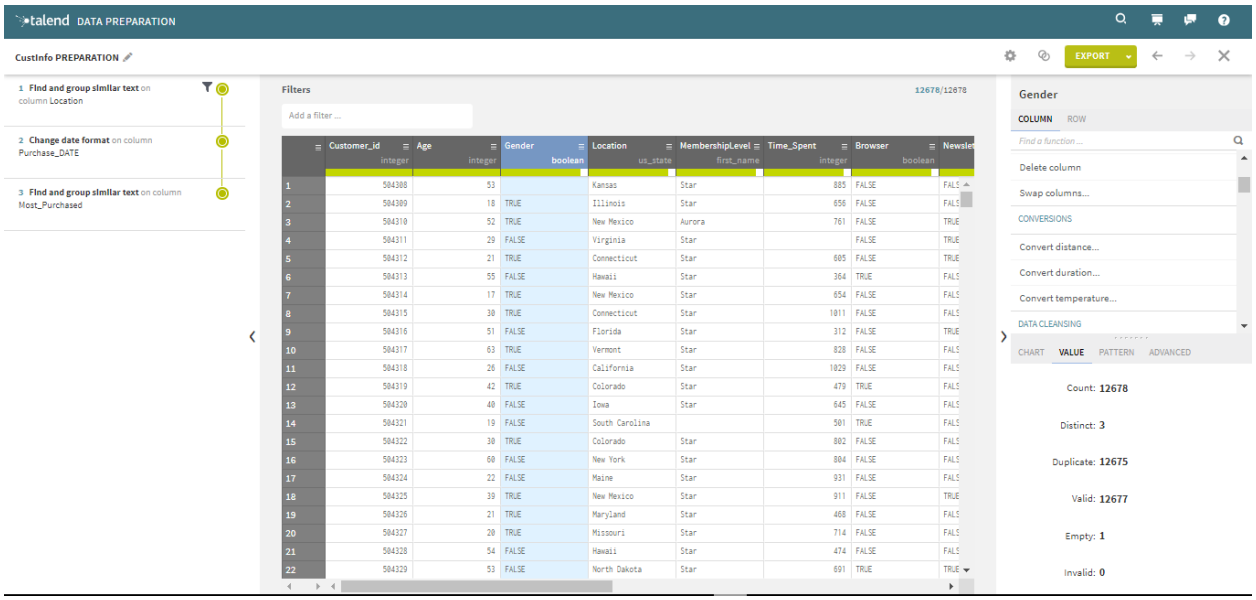

talend DATA PREPARATION
CustInfo PREPARATION
1 Find and group similar text on column Location
2 Change date format on column Purchase_DATE
FIND AND GROUP SIMILAR TEXT
Replace all similar values with the right one (i.e. cluster on fuzzy matching)
These values have been found
This value will be kept
Home Security Systems
Home Security
Beauty and Skincare
Bath and Body
Bath and Body
Sewing and Needlecraft
Snacks and Chips
Snacks and Chips
Cleaning Tools and Supplies
Replace value:
Cleaning Supplies
SUBMIT
EXPORT
Most_Purchased
COLUMN
ROW
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
Max length: 32
Invalid: 0

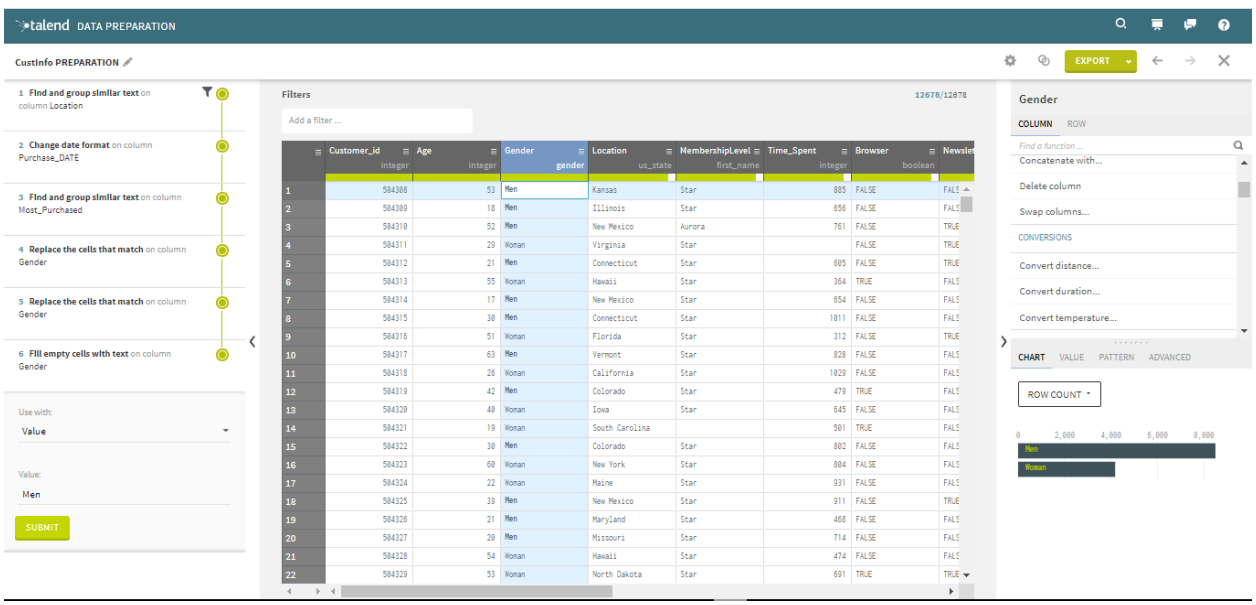




The subsequent step involves grouping and modifying the data within the "most_purchased" column to ensure uniformity in format throughout the dataset using Talend. This process addresses any inconsistencies, variations, or irregularities in the "most_purchased" column, fostering standardized representations 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.

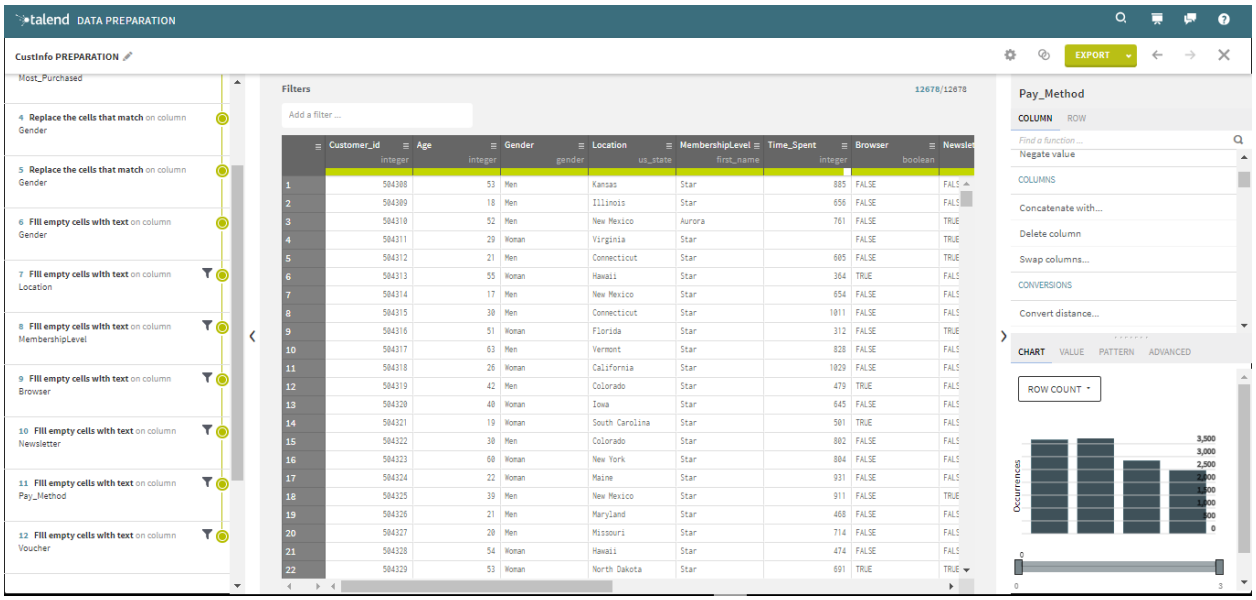
4. Gender

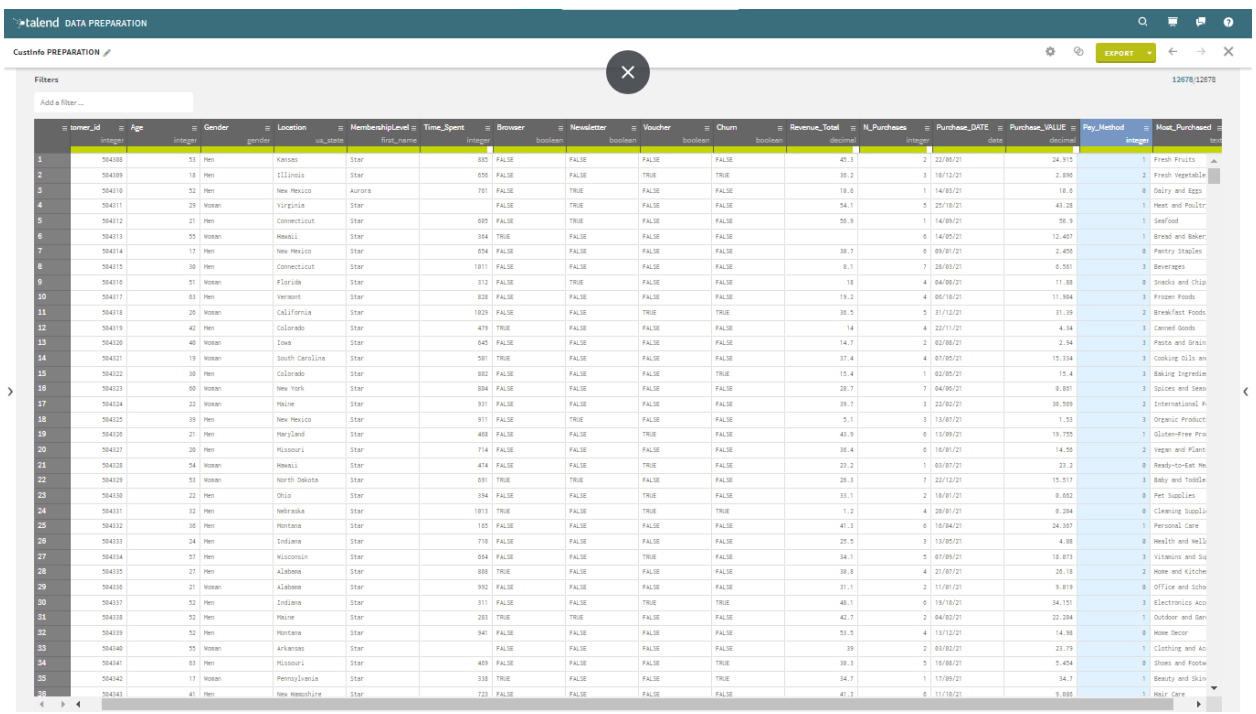




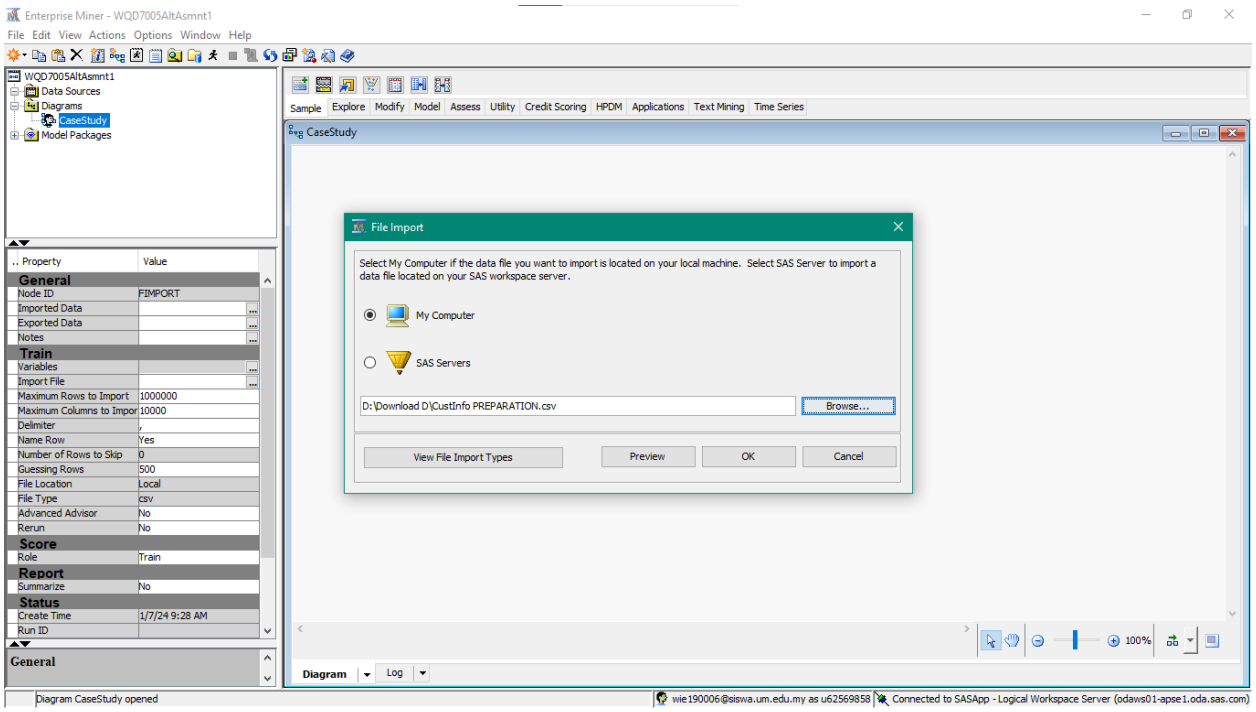
To uniformly represent gender in the dataset, Talend's tMap component is utilized. A new output column, "Gender," is created using a ternary expression, assigning "Men" for true and "Women" for false in the specified boolean column. Executing the Talend job applies this transformation across all rows,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

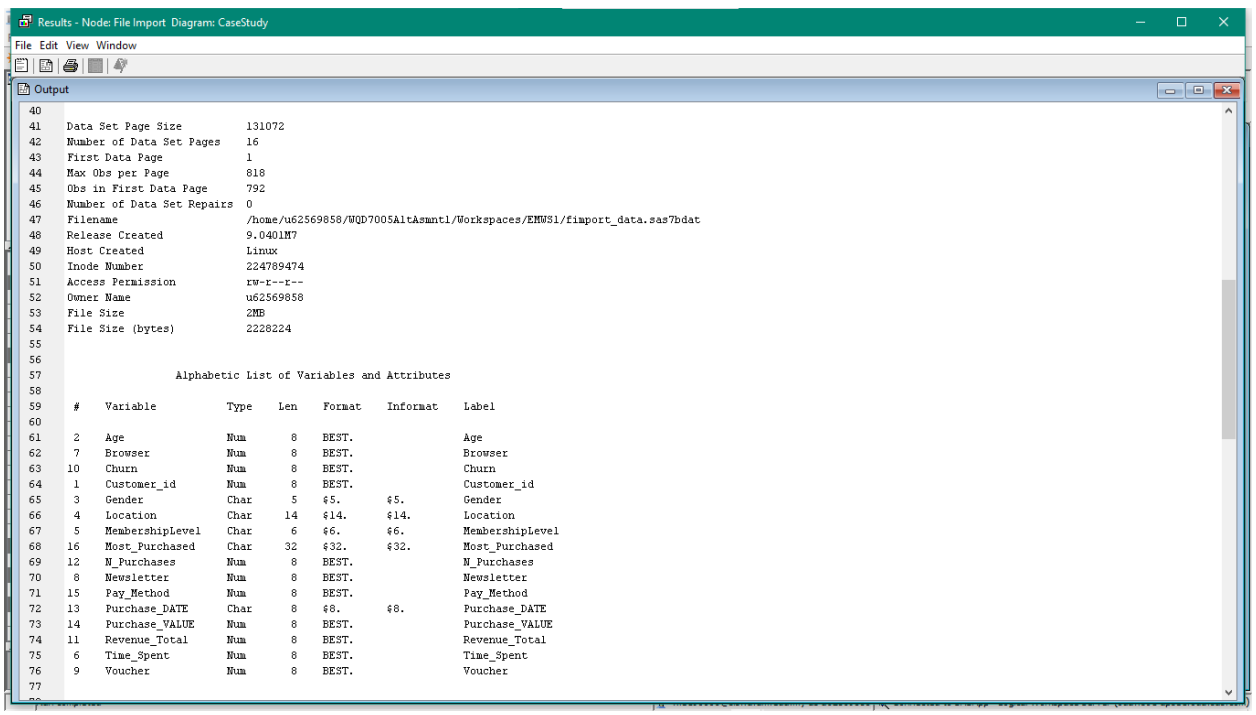
5. Fill empty cells





SAS Enterprise Miner






```
Data Set Page Size          131072
Number of Data Set Pages    16
First Data Page             1
Max Obs per Page            818
Obs in First Data Page      792
Number of Data Set Repairs  0
Filename                    /home/u62569858/WQD7005AltAsmnt1/Workspaces/EMWS1/fimport_data.sas7bdat
Release Created             9.0401M7
Host Created                Linux
Inode Number                224789474
Access Permission           rw-r--r--
Owner Name                  u62569858
File Size                   2MB
File Size (bytes)           2228224


                  Alphabetic List of Variables and Attributes

 #    Variable           Type    Len    Format    Informat    Label

 2    Age                Num       8    BEST.                 Age
 7    Browser            Num       8    BEST.                 Browser
10    Churn              Num       8    BEST.                 Churn
 1    Customer_id        Num       8    BEST.                 Customer_id
 3    Gender             Char      5    $5.       $5.         Gender
 4    Location           Char     14    $14.      $14.        Location
 5    MembershipLevel    Char      6    $6.       $6.         MembershipLevel
16    Most_Purchased     Char     32    $32.      $32.        Most_Purchased
12    N_Purchases        Num       8    BEST.                 N_Purchases
 8    Newsletter         Num       8    BEST.                 Newsletter
15    Pay_Method         Num       8    BEST.                 Pay_Method
13    Purchase_DATE      Char      8    $8.       $8.         Purchase_DATE
14    Purchase_VALUE     Num       8    BEST.                 Purchase_VALUE
11    Revenue_Total      Num       8    BEST.                 Revenue_Total
 6    Time_Spent         Num       8    BEST.                 Time_Spent
 9    Voucher            Num       8    BEST.                 Voucher
```

Results - Node: StatExplore Diagram: CaseStudy

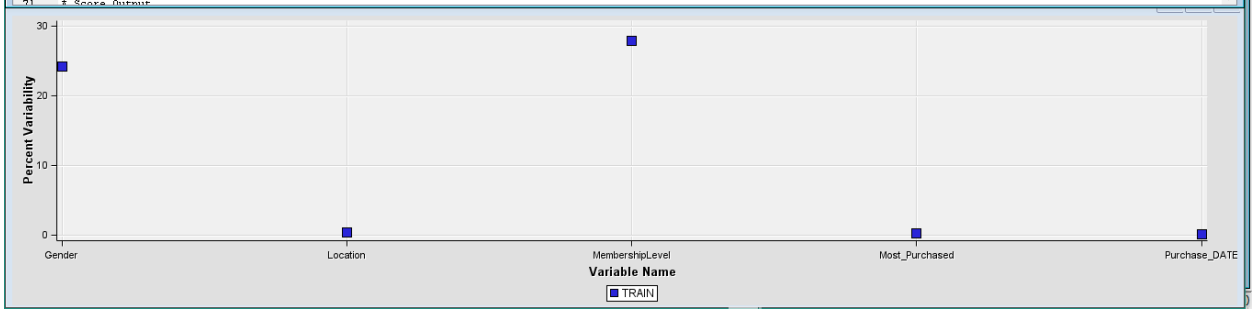
File Edit View Window

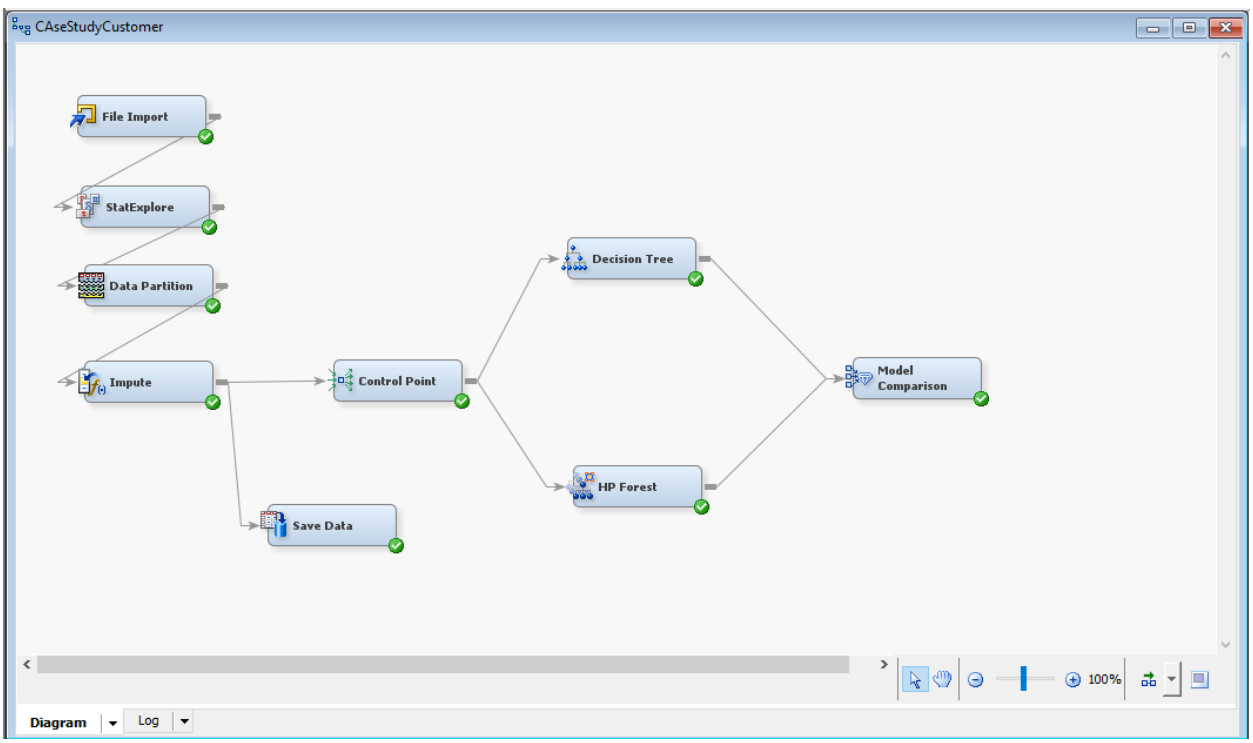
Output

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Age	INPUT	39.62881	13.81948	12678	0	16	40	63	-0.01162	-1.20736
Browser	INPUT	0.197981	0.398493	12678	0	0	0	1	1.516045	0.29844
Churn	INPUT	0.213519	0.409807	12678	0	0	0	1	1.398343	-0.04464
N_Purchases	INPUT	3.999921	2.012641	12677	1	1	4	7	-0.00756	-1.26277
Newsletter	INPUT	0.146711	0.353831	12678	0	0	0	1	1.99725	1.989322
Pay_Method	INPUT	1.324894	1.088886	12678	0	0	1	3	0.234416	-1.24206
Purchase_VALUE	INPUT	15.86966	13.12753	12677	1	0.015	12.525	59.9	0.897868	0.157923
Revenue_Total	INPUT	27.6379	14.95342	12677	1	0.5	30.1	59.9	0.036222	-0.76665
Time_Spent	INPUT	598.5643	278.1178	12675	3	120	602	1080	-0.00314	-1.20963
Voucher	INPUT	0.246411	0.430938	12678	0	0	0	1	1.177102	-0.61453

```
*-----------------------------------------------------------*
* Score Output
```





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	.	.
MembershipLeve	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