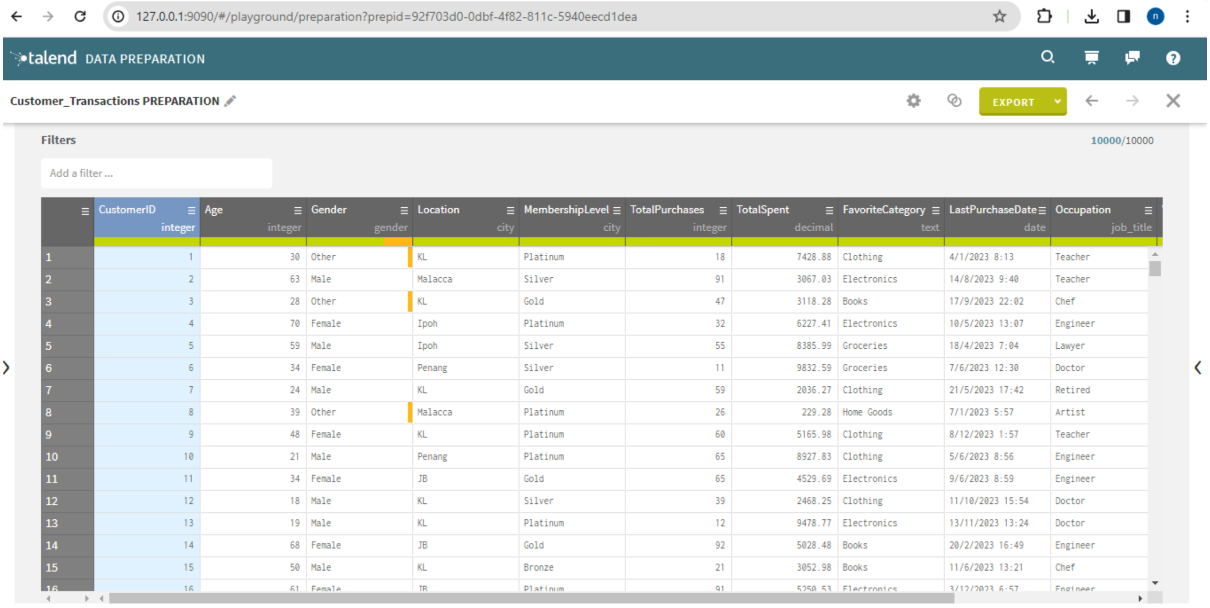
## Preprocessing using Talend Data Preparation

For the next preprocessing, I utilized Talend Data Preparation platform. I uploaded the integrated data from Talend Data Integration, which was produced in CSV delimiter format generated by the tFileOutputDelimited component. The figure below shows the uploaded CSV file in the talend data Preparation.


127.0.0.1:9090/#/playground/preparation?prepid=92f703d0-0dbf-4f82-811c-5940eecd1dea
talend DATA PREPARATION
EXPORT

Customer_Transactions PREPARATION

Filters

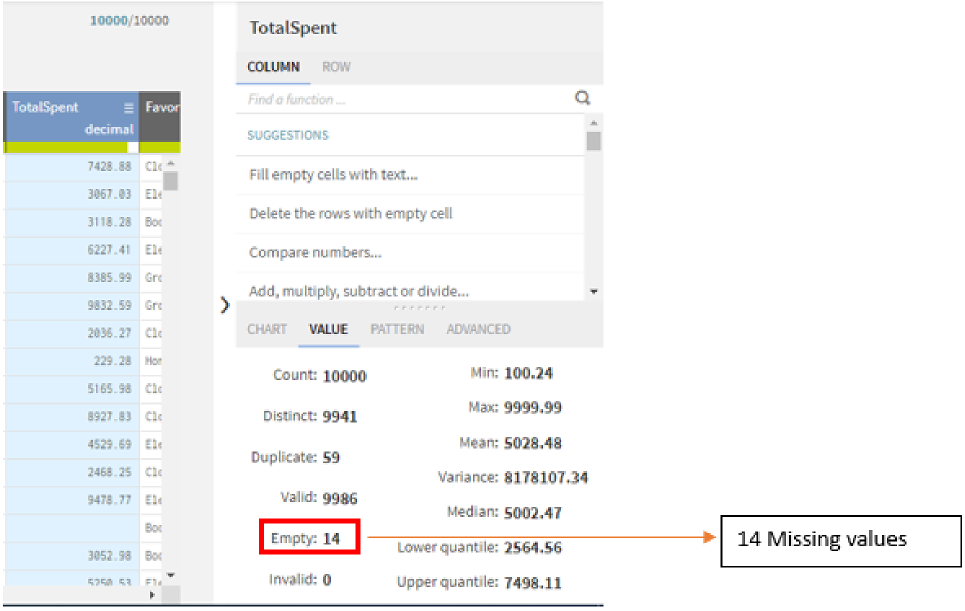
Add a filter ...

10000/10000

| | CustomerID | Age | Gender | Location | MembershipLevel | TotalPurchases | TotalSpent | FavoriteCategory | LastPurchaseDate | Occupation |
|---|---|---|---|---|---|---|---|---|---|---|
| | integer | integer | gender | city | city | integer | decimal | text | date | job_title |
| 1 | 1 | 30 | Other | KL | Platinum | 18 | 7428.88 | Clothing | 4/1/2023 8:13 | Teacher |
| 2 | 2 | 63 | Male | Malacca | Silver | 91 | 3067.03 | Electronics | 14/8/2023 9:40 | Teacher |
| 3 | 3 | 28 | Other | KL | Gold | 47 | 3118.28 | Books | 17/9/2023 22:02 | Chef |
| 4 | 4 | 70 | Female | Ipoh | Platinum | 32 | 6227.41 | Electronics | 10/5/2023 13:07 | Engineer |
| 5 | 5 | 59 | Male | Ipoh | Silver | 55 | 8385.99 | Groceries | 18/4/2023 7:04 | Lawyer |
| 6 | 6 | 34 | Female | Penang | Silver | 11 | 9832.59 | Groceries | 7/6/2023 12:30 | Doctor |
| 7 | 7 | 24 | Male | KL | Gold | 59 | 2036.27 | Clothing | 21/5/2023 17:42 | Retired |
| 8 | 8 | 39 | Other | Malacca | Platinum | 26 | 229.28 | Home Goods | 7/1/2023 5:57 | Artist |
| 9 | 9 | 48 | Female | KL | Platinum | 60 | 5165.98 | Clothing | 8/12/2023 1:57 | Teacher |
| 10 | 10 | 21 | Male | Penang | Platinum | 65 | 8927.83 | Clothing | 5/6/2023 8:56 | Engineer |
| 11 | 11 | 34 | Female | JB | Gold | 65 | 4529.69 | Electronics | 9/6/2023 8:59 | Engineer |
| 12 | 12 | 18 | Male | KL | Silver | 39 | 2468.25 | Clothing | 11/10/2023 15:54 | Doctor |
| 13 | 13 | 19 | Male | KL | Platinum | 12 | 9478.77 | Electronics | 13/11/2023 13:24 | Doctor |
| 14 | 14 | 68 | Female | JB | Gold | 92 | 5028.48 | Books | 20/2/2023 16:49 | Engineer |
| 15 | 15 | 50 | Male | KL | Bronze | 21 | 3052.98 | Books | 11/6/2023 13:21 | Chef |
| 16 | 16 | 61 | Female | JB | Platinum | 91 | 5250.53 | Electronics | 3/12/2023 6:57 | Engineer |

## Data Cleaning

The data cleaning process involves individually clicking on each column. During this, I noticed that the 'TotalSpend' column contains 14 missing values.



10000/10000

| TotalSpent decimal | Favor |
|---|---|
| 7428.88 | Cl |
| 3067.03 | El |
| 3118.28 | Bo |
| 6227.41 | El |
| 8385.99 | Gr |
| 9832.59 | Gr |
| 2036.27 | Cl |
| 229.28 | Ho |
| 5165.98 | Cl |
| 8927.83 | Cl |
| 4529.69 | El |
| 2468.25 | Cl |
| 9478.77 | El |
|  | Bo |
| 3052.98 | Bo |
| 5250.53 | El |

TotalSpent

Fill empty cells with text...

Delete the rows with empty cell

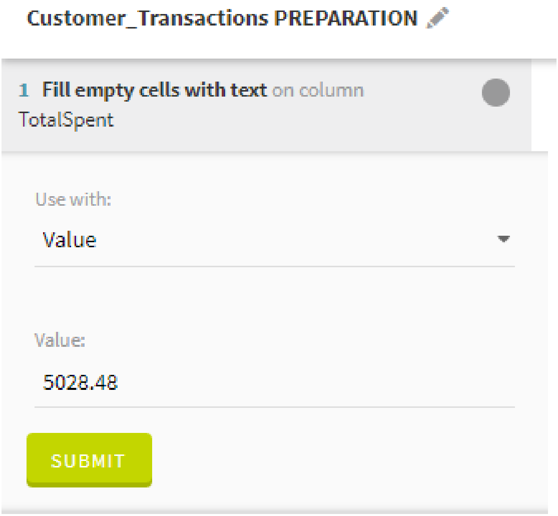
Compare numbers...

Add, multiply, subtract or divide...

Count: 10000
Distinct: 9941
Duplicate: 59
Valid: 9986
Empty: 14
Invalid: 0

Min: 100.24
Max: 9999.99
Mean: 5028.48
Variance: 8178107.34
Median: 5002.47
Lower quantile: 2564.56
Upper quantile: 7498.11

14 Missing values

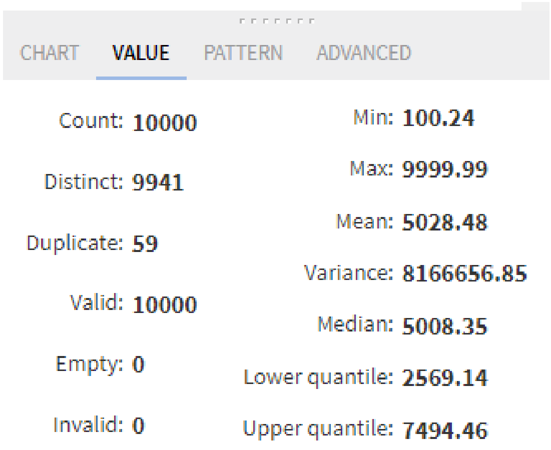


Customer_Transactions PREPARATION

1 Fill empty cells with text on column TotalSpent

Use with: Value

Value: 5028.48



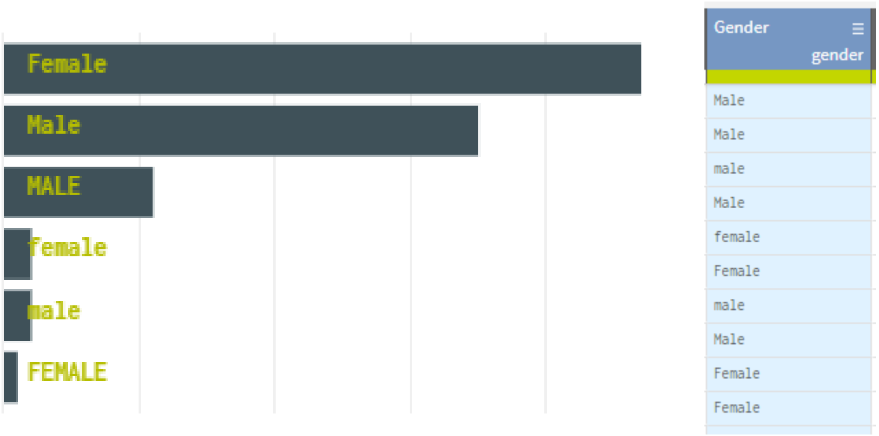
Count: 10000
Distinct: 9941
Duplicate: 59
Valid: 10000
Empty: 0
Invalid: 0

Min: 100.24
Max: 9999.99
Mean: 5028.48
Variance: 8166656.85
Median: 5008.35
Lower quantile: 2569.14
Upper quantile: 7494.46

For the "TotalSpend" column, I identified 14 missing values. These were replaced with the mean value of 5028.48 to ensure that the column is free from any empty entries.

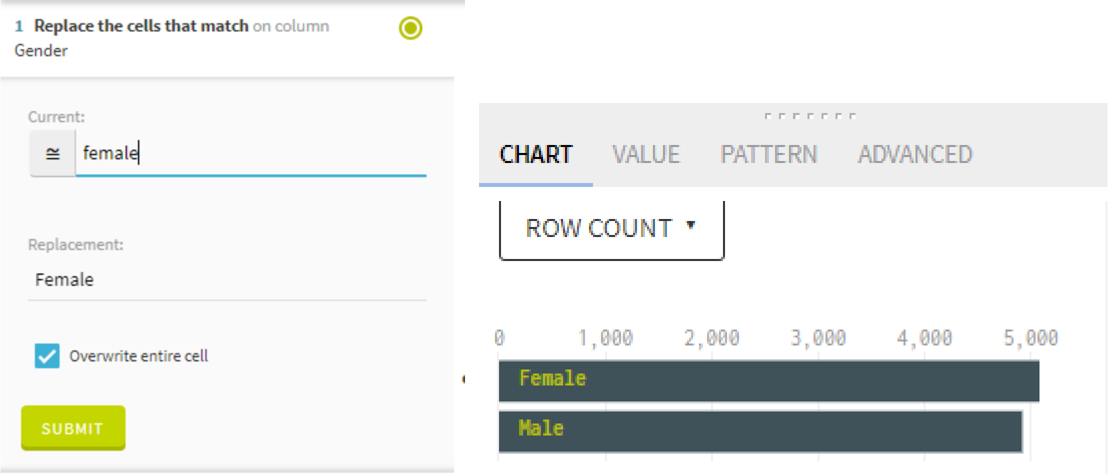
### Data Standardization

The data standardization process involves individually clicking on each column. During this, I noticed that the 'Gender' column contains varying letter cases, and the 'Date' column also lacks standardization.



| Gender<br>gender |
|---|
| Male |
| Male |
| male |
| Male |
| female |
| Female |
| male |
| Male |
| Female |
| Female |

The figure above shows the gender contains varying letter cases.

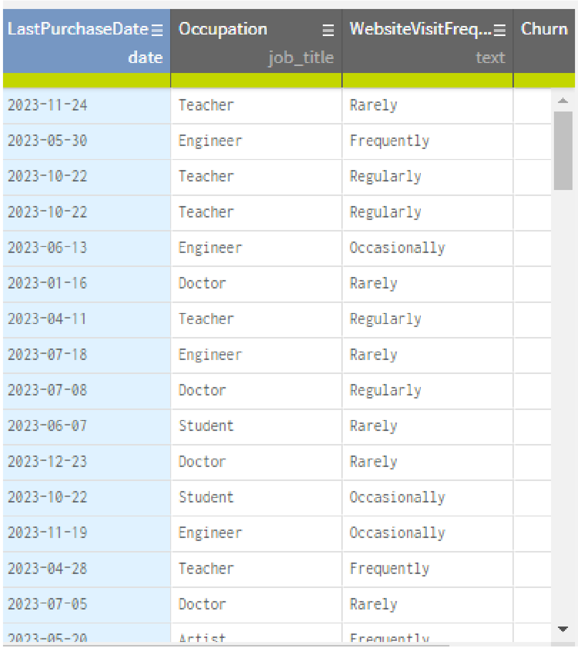

1 **Replace the cells that match** on column Gender

Current: ≅ female

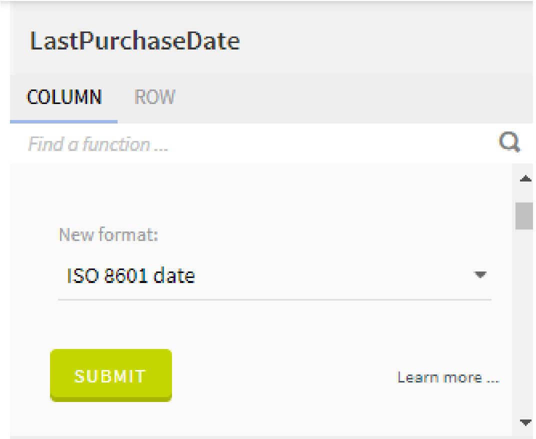
Replacement: Female

☑ Overwrite entire cell

The figure above shows after standardized the gender.



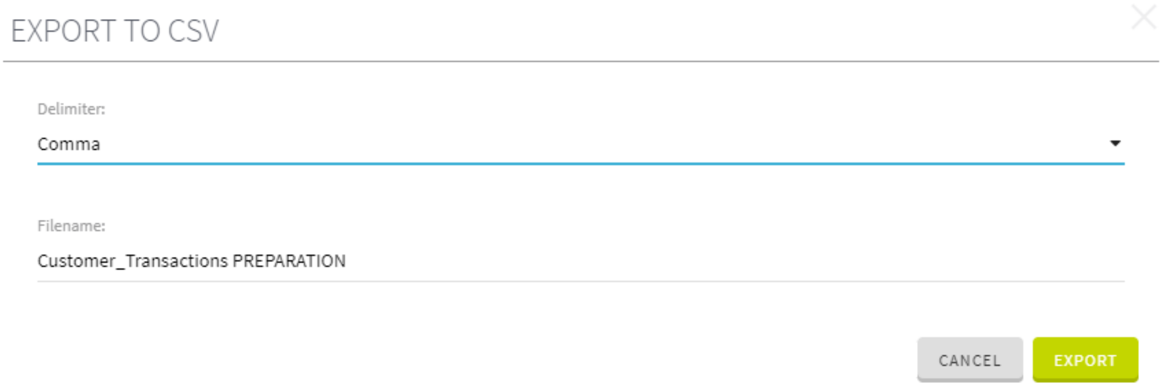
| LastPurchaseDate | Occupation | WebsiteVisitFreq... | Churn |
|---|---|---|---|
| date | job_title | text | |
| 2023-11-24 | Teacher | Rarely | |
| 2023-05-30 | Engineer | Frequently | |
| 2023-10-22 | Teacher | Regularly | |
| 2023-10-22 | Teacher | Regularly | |
| 2023-06-13 | Engineer | Occasionally | |
| 2023-01-16 | Doctor | Rarely | |
| 2023-04-11 | Teacher | Regularly | |
| 2023-07-18 | Engineer | Rarely | |
| 2023-07-08 | Doctor | Regularly | |
| 2023-06-07 | Student | Rarely | |
| 2023-12-23 | Doctor | Rarely | |
| 2023-10-22 | Student | Occasionally | |
| 2023-11-19 | Engineer | Occasionally | |
| 2023-04-28 | Teacher | Frequently | |
| 2023-07-05 | Doctor | Rarely | |
| 2023-05-20 | Artist | Frequently | |



LastPurchaseDate

New format: ISO 8601 date

After identifying the date columns, a function within Talend was selected to standardize the dates. The ISO 8601 format was specified within this function to transform all date entries to this uniform standard. Following the transformation, the data was reviewed to ensure that the dates were correctly standardized, with any anomalies or errors requiring further attention being identified and possibly corrected.



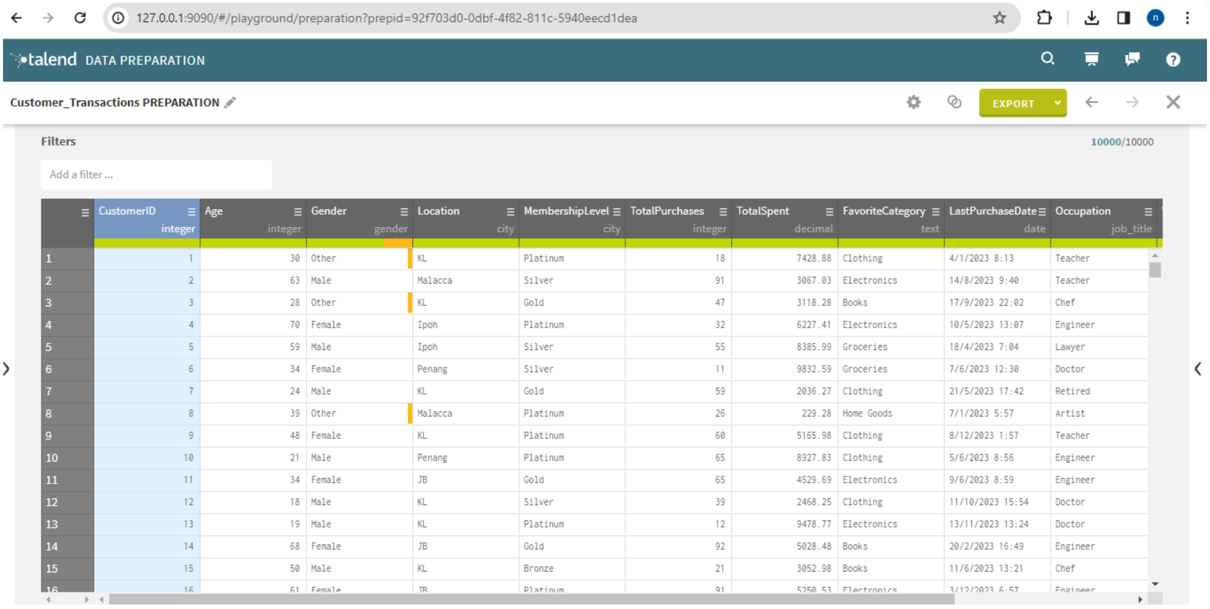
EXPORT TO CSV

Delimiter: Comma

Filename: Customer_Transactions PREPARATION

The final step involves exporting the file from Talend Data Preparation in CSV format, ensuring that the delimiter is set to a comma.

# APPENDIX



127.0.0.1:9090/#/playground/preparation?prepid=92f703d0-0dbf-4f82-811c-5940eecd1dea

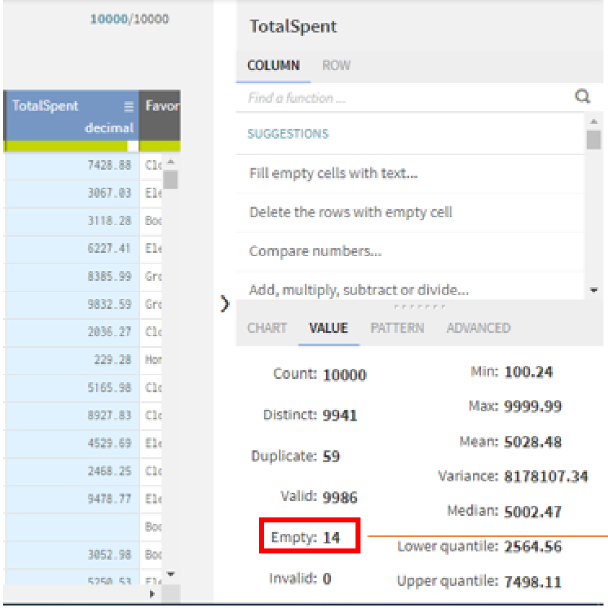
Customer_Transactions PREPARATION

Filters

Add a filter ...

10000/10000

| | CustomerID<br>integer | Age<br>integer | Gender<br>gender | Location<br>city | MembershipLevel<br>city | TotalPurchases<br>integer | TotalSpent<br>decimal | FavoriteCategory<br>text | LastPurchaseDate<br>date | Occupation<br>job_title |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 30 | Other | KL | Platinum | 18 | 7428.88 | Clothing | 4/1/2023 8:13 | Teacher |
| 2 | 2 | 63 | Male | Malacca | Silver | 91 | 3067.03 | Electronics | 14/8/2023 9:40 | Teacher |
| 3 | 3 | 28 | Other | KL | Gold | 47 | 3118.28 | Books | 17/9/2023 22:02 | Chef |
| 4 | 4 | 70 | Female | Ipoh | Platinum | 32 | 6227.41 | Electronics | 10/5/2023 13:07 | Engineer |
| 5 | 5 | 59 | Male | Ipoh | Silver | 55 | 8385.99 | Groceries | 18/4/2023 7:04 | Lawyer |
| 6 | 6 | 34 | Female | Penang | Silver | 11 | 9832.59 | Groceries | 7/6/2023 12:30 | Doctor |
| 7 | 7 | 24 | Male | KL | Gold | 59 | 2036.27 | Clothing | 21/5/2023 17:42 | Retired |
| 8 | 8 | 39 | Other | Malacca | Platinum | 26 | 229.28 | Home Goods | 7/1/2023 5:57 | Artist |
| 9 | 9 | 48 | Female | KL | Platinum | 60 | 5165.98 | Clothing | 8/12/2023 1:57 | Teacher |
| 10 | 10 | 21 | Male | Penang | Platinum | 65 | 8927.83 | Clothing | 5/6/2023 8:56 | Engineer |
| 11 | 11 | 34 | Female | JB | Gold | 65 | 4529.69 | Electronics | 9/6/2023 8:59 | Engineer |
| 12 | 12 | 18 | Male | KL | Silver | 39 | 2468.25 | Clothing | 11/10/2023 15:54 | Doctor |
| 13 | 13 | 19 | Male | KL | Platinum | 12 | 9478.77 | Electronics | 13/11/2023 13:24 | Doctor |
| 14 | 14 | 68 | Female | JB | Gold | 92 | 5028.48 | Books | 20/2/2023 16:49 | Engineer |
| 15 | 15 | 50 | Male | KL | Bronze | 21 | 3052.98 | Books | 11/6/2023 13:21 | Chef |
| 16 | 16 | 61 | Female | JB | Platinum | 91 | 5250.53 | Electronics | 3/12/2023 6:57 | Engineer |



10000/10000

| TotalSpent<br>decimal | Favor |
|---|---|
| 7428.88 | Clo |
| 3067.03 | Ele |
| 3118.28 | Boo |
| 6227.41 | Ele |
| 8385.99 | Gro |
| 9832.59 | Gro |
| 2036.27 | Clo |
| 229.28 | Hor |
| 5165.98 | Clo |
| 8927.83 | Clo |
| 4529.69 | Ele |
| 2468.25 | Clo |
| 9478.77 | Ele |
| | Boo |
| 3052.98 | Boo |
| 5250.53 | Ele |

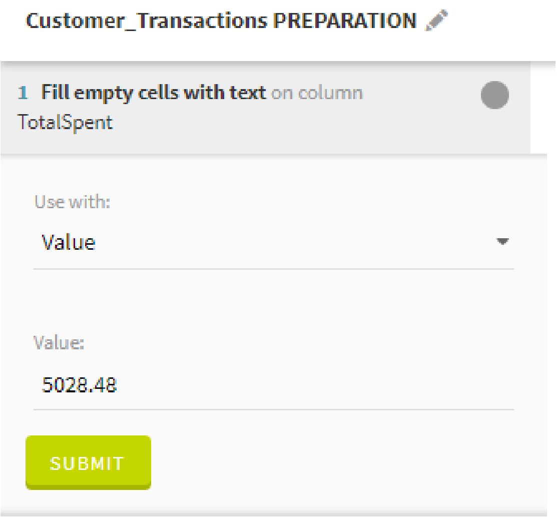
TotalSpent

COLUMN
ROW
Find a function ...
SUGGESTIONS
Fill empty cells with text...
Delete the rows with empty cell
Compare numbers...
Add, multiply, subtract or divide...
CHART
VALUE
PATTERN
ADVANCED

Count: 10000
Distinct: 9941
Duplicate: 59
Valid: 9986
Empty: 14
Invalid: 0

Min: 100.24
Max: 9999.99
Mean: 5028.48
Variance: 8178107.34
Median: 5002.47
Lower quantile: 2564.56
Upper quantile: 7498.11

14 Missing values

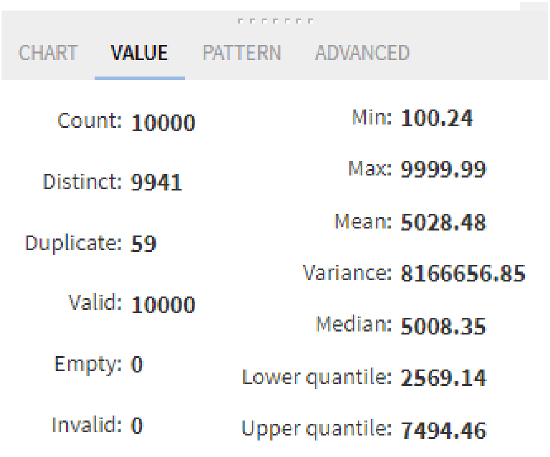


Customer_Transactions PREPARATION

1 **Fill empty cells with text** on column TotalSpent

Use with: Value

Value: 5028.48



Count: 10000

Distinct: 9941

Duplicate: 59

Valid: 10000

Empty: 0

Invalid: 0

Min: 100.24

Max: 9999.99

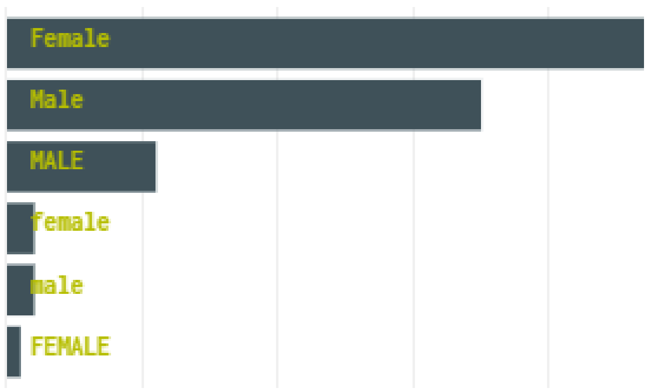
Mean: 5028.48

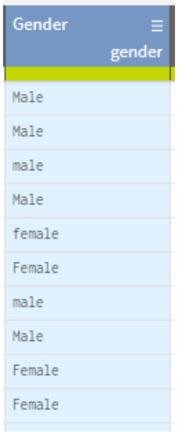
Variance: 8166656.85

Median: 5008.35

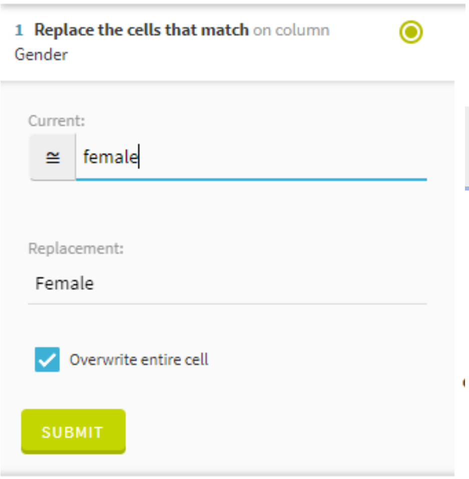
Lower quantile: 2569.14

Upper quantile: 7494.46





| Gender<br>gender |
|---|
| Male |
| Male |
| male |
| Male |
| female |
| Female |
| male |
| Male |
| Female |
| Female |

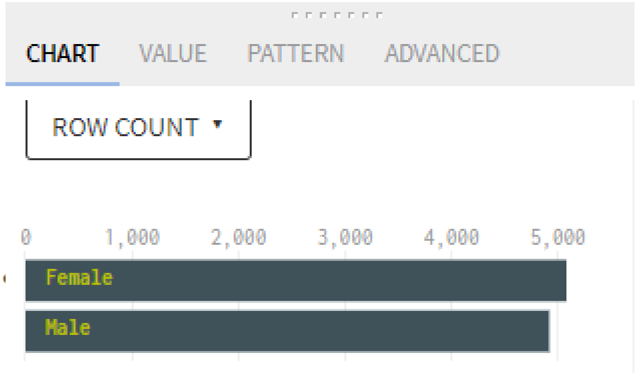


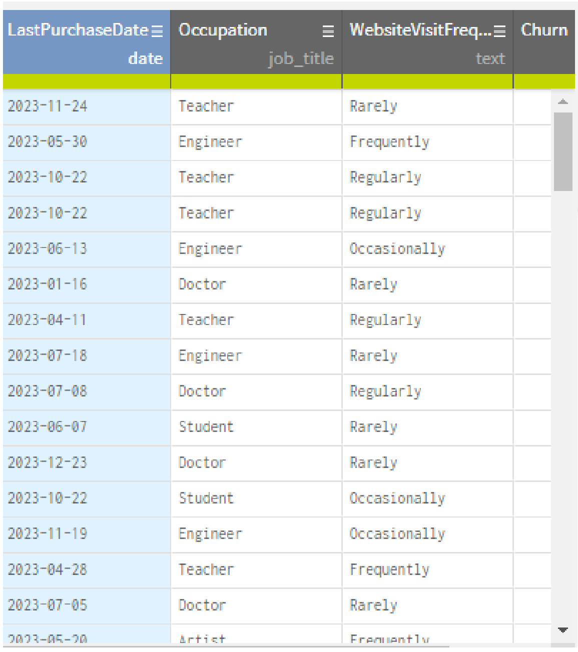
1 **Replace the cells that match** on column Gender

Current: ≅ female

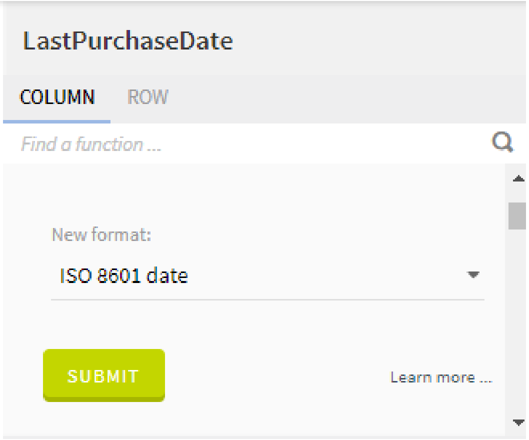
Replacement: Female

☑ Overwrite entire cell



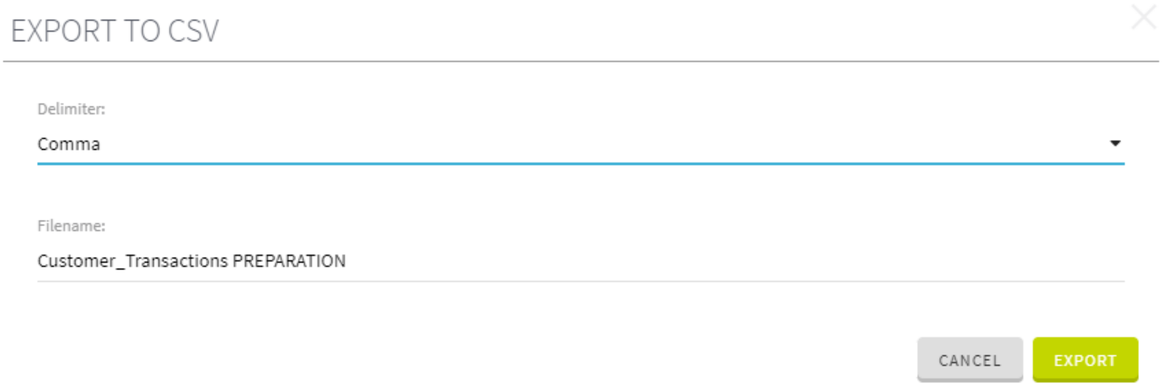


| LastPurchaseDate<br>date | Occupation<br>job_title | WebsiteVisitFreq...<br>text | Churn |
|---|---|---|---|
| 2023-11-24 | Teacher | Rarely | |
| 2023-05-30 | Engineer | Frequently | |
| 2023-10-22 | Teacher | Regularly | |
| 2023-10-22 | Teacher | Regularly | |
| 2023-06-13 | Engineer | Occasionally | |
| 2023-01-16 | Doctor | Rarely | |
| 2023-04-11 | Teacher | Regularly | |
| 2023-07-18 | Engineer | Rarely | |
| 2023-07-08 | Doctor | Regularly | |
| 2023-06-07 | Student | Rarely | |
| 2023-12-23 | Doctor | Rarely | |
| 2023-10-22 | Student | Occasionally | |
| 2023-11-19 | Engineer | Occasionally | |
| 2023-04-28 | Teacher | Frequently | |
| 2023-07-05 | Doctor | Rarely | |
| 2023-05-20 | Artist | Frequently | |



LastPurchaseDate

New format: ISO 8601 date



EXPORT TO CSV

Delimiter: Comma

Filename: Customer_Transactions PREPARATION