alteryx

DATA PREPARATION

Presented by:ANGELO MAMMANA, JOGINDER SINGH E LEONARDO FANTIN

Index

1

ALTERYX

- Overview
- How it works
- Learning Datasets
- Pros and cons

2

PIPELINE

- Data Ingestion and discovery
- Data Validation
- Data Structuring
- Data Enrichment
- Data Filtering
- Data Cleaning

3

RESULTS

- Final pipeline
- Datasets obtained
- Conclusion

1 Overview

Alteryx is a **data science and business intelligence platform** that provides tools for data analysis, report creation, and predictive modeling. It features capabilities such as data processing, cleaning, and preparation, statistical analysis, and the creation of interactive dashboards.

It is designed to be used by non technical users, making it a popular option for businesses looking to analyze their data.

1 How it works

Alteryx includes a *drag and drop interface* that allows users to create workflow for data processing, analysis, and report creation. Users *can import* data from various sources, such as excel format and CSV files

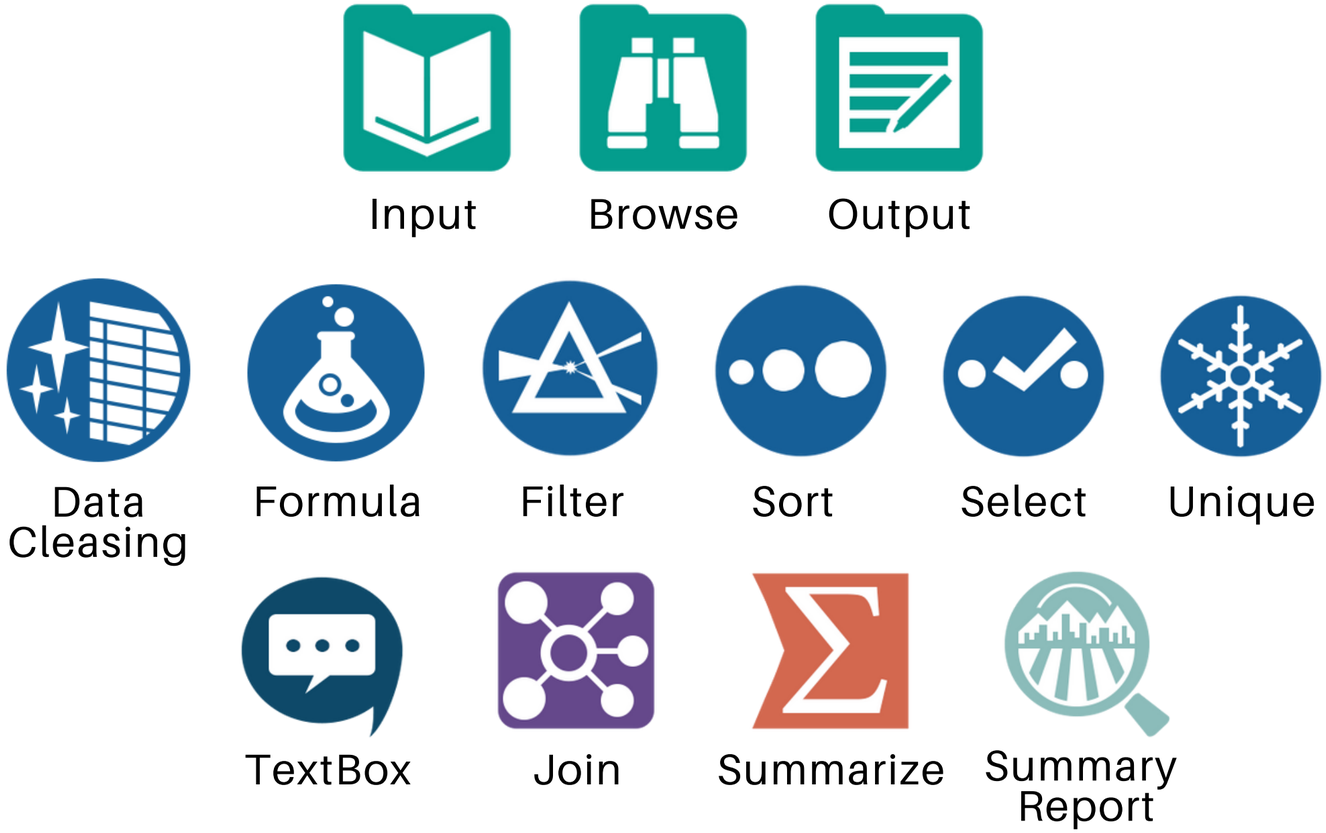
Once the data is processed, users can utilize predictive models and statistical analysis features to generate insights that can inform business decisions

1 Learning Datasets

Dataset about **crimes**

Dataset about **retail sales**

1 Tools used


Input
Browse
Output
Data Cleasing
Formula
Filter
Sort
Select
Unique
TextBox
Join
Summarize
Summary Report

1 Pros and Cons

- Intuitive and easy to use interface, allowing even non expert users to create complex data analysis workflows

- Offers proprietary formats like 'yxdb' that we used to write and read data

- Advanced data preparation tools, including join, union, cleaning, and data transformation

- Support for a wide variety of file formats, including Excel, CSV, JSON, and SQL

- High pricing for some organizations

- it is not possible to define global variables

- Complexity in transitioning from SQL to Alteryx for some advanced functionality

Index

1

ALTERYX

- Overview
- How it works
- Learning Datasets
- Pros and cons

2

PIPELINE

- Data Ingestion and discovery
- Data Validation
- Data Structuring
- Data Enrichment
- Data Filtering
- Data Cleaning

3

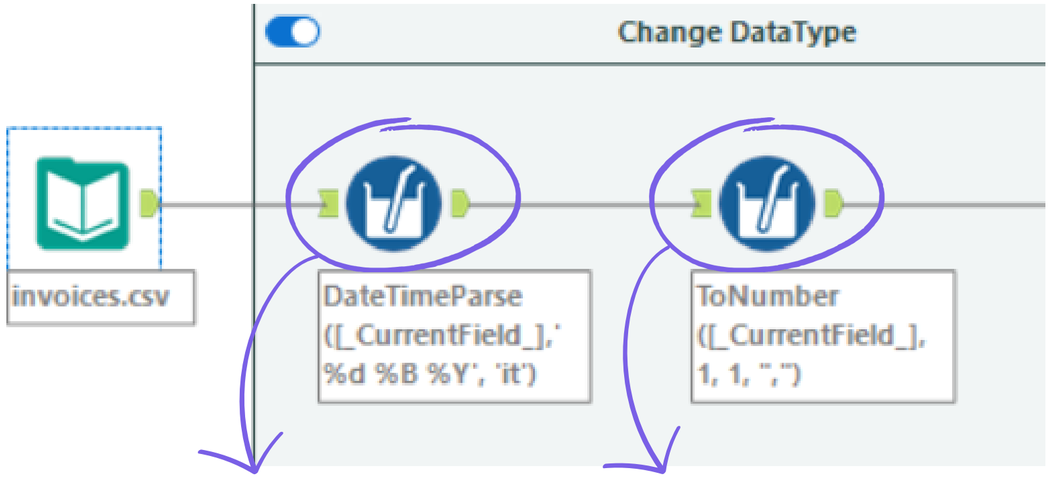
RESULTS

- Final pipeline
- Datasets obtained
- Conclusion

2 Data ingestion and discovery

- Column Cleaning

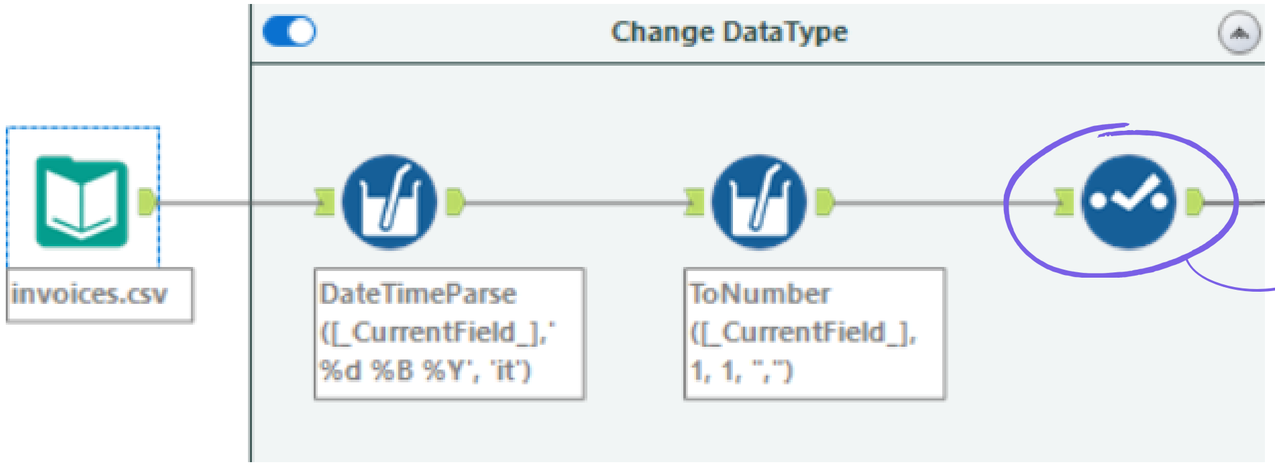
date	gas_amount	gas_average_cost	howmuch_pay	total_amount	average_gas_bill_cost
16 Dicembre 2020	383,66	0,32 €/smc	383,66	383,66	0,86 €/smc
21 Novembre 2020	386,63	0,37 €/smc	197,77	197,77	0,86 €/smc
12 Dicembre 2020	15,69	0,22 €/smc	31,63	31,63	0,65 €/smc
5 Dicembre 2020	67,55	0,29 €/smc	114,12	114,12	0,73 €/smc



date	gas_amount	gas_average_cost	average_gas_bill_cost	total_amount	howmuch_pay
2020-12-16	383.66	0.32	0.86	383.66	383.66
2020-11-21	386.63	0.37	0.86	197.77	197.77
2020-12-12	15.69	0.22	0.65	31.63	31.63

2 Data ingestion and discovery

- Column Casting



Field	Type	Size	Rename	Description
bill_id	Int64	8		Invoice identifier
user_code	V_WString	254		(Anonymized) code for the customer that owns thi...
customer_code	V_WString	254		Combined with user_code provides a unique iden...
city	V_WString	254		City where the utility is located
address	V_WString	254		(Anonymized) address of the utility location
nominative	V_WString	254		(Anonymized) customer name
sex	V_WString	254		Sex of the customer It could be 'M', 'F', 'P', with '...
age	Int32	4		Age of the customer, set to null for commercial ac...
supply_type	V_WString	254		Supply type ('light', 'gas', 'gas and light')
date	V_WString	254		
light_start_date	V_WString	254		
light_end_date	V_WString	254		
New_light_start_date	Date	10	start_date	Start date of invoice
New_light_end_date	Date	10	end_date	End date of invoice
New_emission_date	Date	10	emission_date	Emission date
New_gas_amount	Double	8	gas_amount	Gas fee to pay
New_gas_average_cost	Double	8	gas_average_cost	Average cost of gas
gas_consumption	Float	4		Consumed gas
gas_offer	V_WString	254		Name of the subscribed gas plan (anonymized)
New_average_gas_bill_cost	Double	8	average_gas_bill_cost	Average cost for the gas invoice
gas_system_charges	Float	4		Extra gas fees
gas_material_cost	Float	4		Costs for gas
gas_transport_cost	Float	4		Extra gas fees
F1_kWh	Float	4		kWh of electricity consumed in the F1 time slot
F2_kWh	Float	4		kWh of electricity consumed in the F2 time slot
F3_kWh	Float	4		kWh of electricity consumed in the F3 time slot
light_average_cost	Float	4		Average cost of electricity
light_consumption	Float	4		Consumed electricity
light_offer_type	V_WString	254		Kind of plan for the electricity ('single zone', 'bizo...
light_offer	V_String	254		Name of the subscribed electricity plan (anonymiz...
New_light_amount	Double	8	light_amount	Amount to pay for the electricity
New_average_unit_light_cost	Double	8	average_unit_light_cost	Average cost for the electricity
New_average_light_bill_cost	Double	8	average_light_bill_cost	Average cost for the electricity invoice
light_system_charges	Float	4		Extra electricity fees
light_transport_cost	Float	4		Extra electricity fees
light_material_cost	Float	4		Costs for electricity

2 Data ingestion and discovery

- Locate Missing Values

emission_date

2021-03-23	438,455
2021-05-21	410,418
2021-03-19	390,737
2021-04-23	372,241
2021-04-20	357,738

78 more >

Null 35.12%
Ok 64.88%

gas_amount

0	74,913
45.46	3,622
56.55	2,620
45.22	2,545
22.73	2,331

995 more >

gas_average_cost

0.27	231,511
0.28	228,707
0.26	223,796
0.29	217,429
0.25	212,499

995 more >

gas_consumption

0	250,073
1	47,985
2	39,219
3	36,869
4	35,042

995 more >

Summary

Type	Records	Data Type Size
Double	10,497,143	8

Ok	6,810,389	64.88%
Unique	> 10,000	
Null	3,686,754	35.12%
Not Ok	0	0.00%
Empty	0	0.00%

Value Statistics

Grouped Values

-66,102.43-387,143.35

Max	387,143.35
Min	-66,102.43
Lower quartile	54.255
Upper quartile	201.464
Average	159.1595
Standard deviation	354.1522
Sum	1,083,938,631.6500
Median	106.896
Variance	125,423.7973

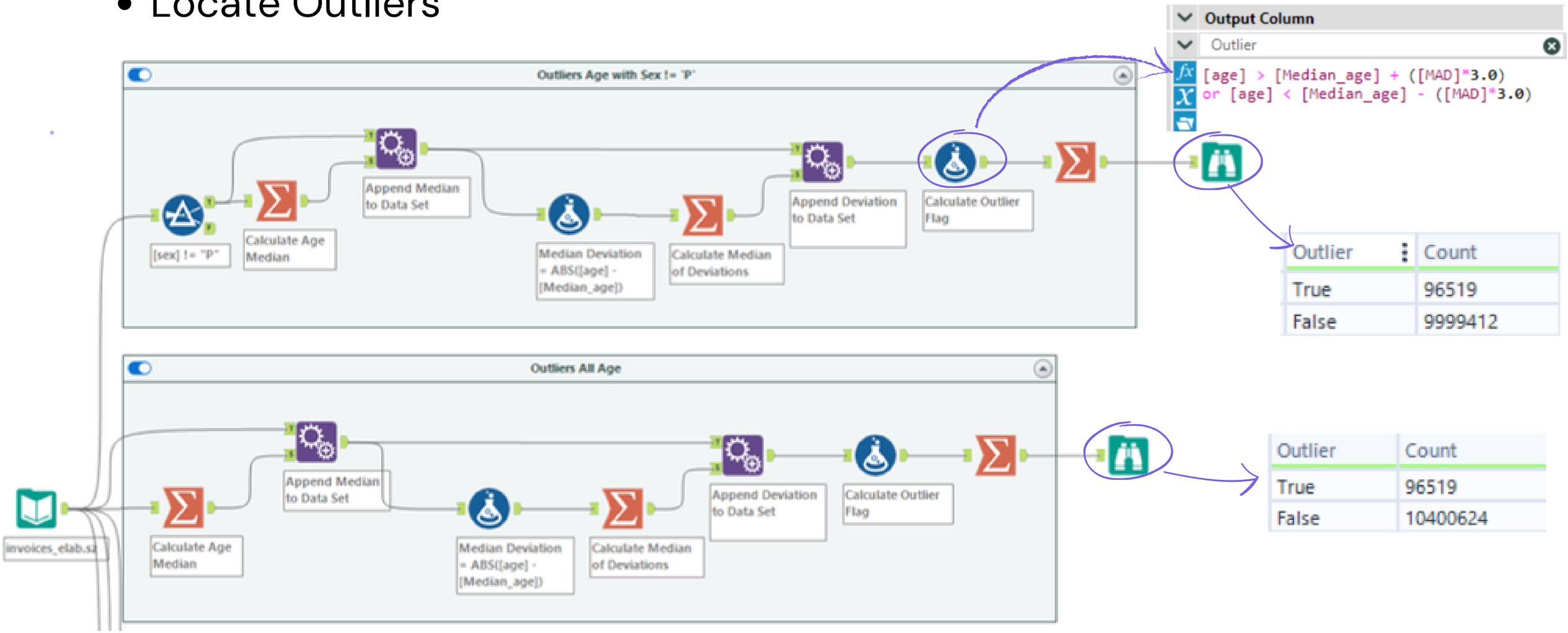
Top Values

0	74,913
45.46	3,622
56.55	2,620
45.22	2,545
22.73	2,331

995 more >

2 Data ingestion and discovery

- Locate Outliers



Outlier	Count
True	96519
False	9999412

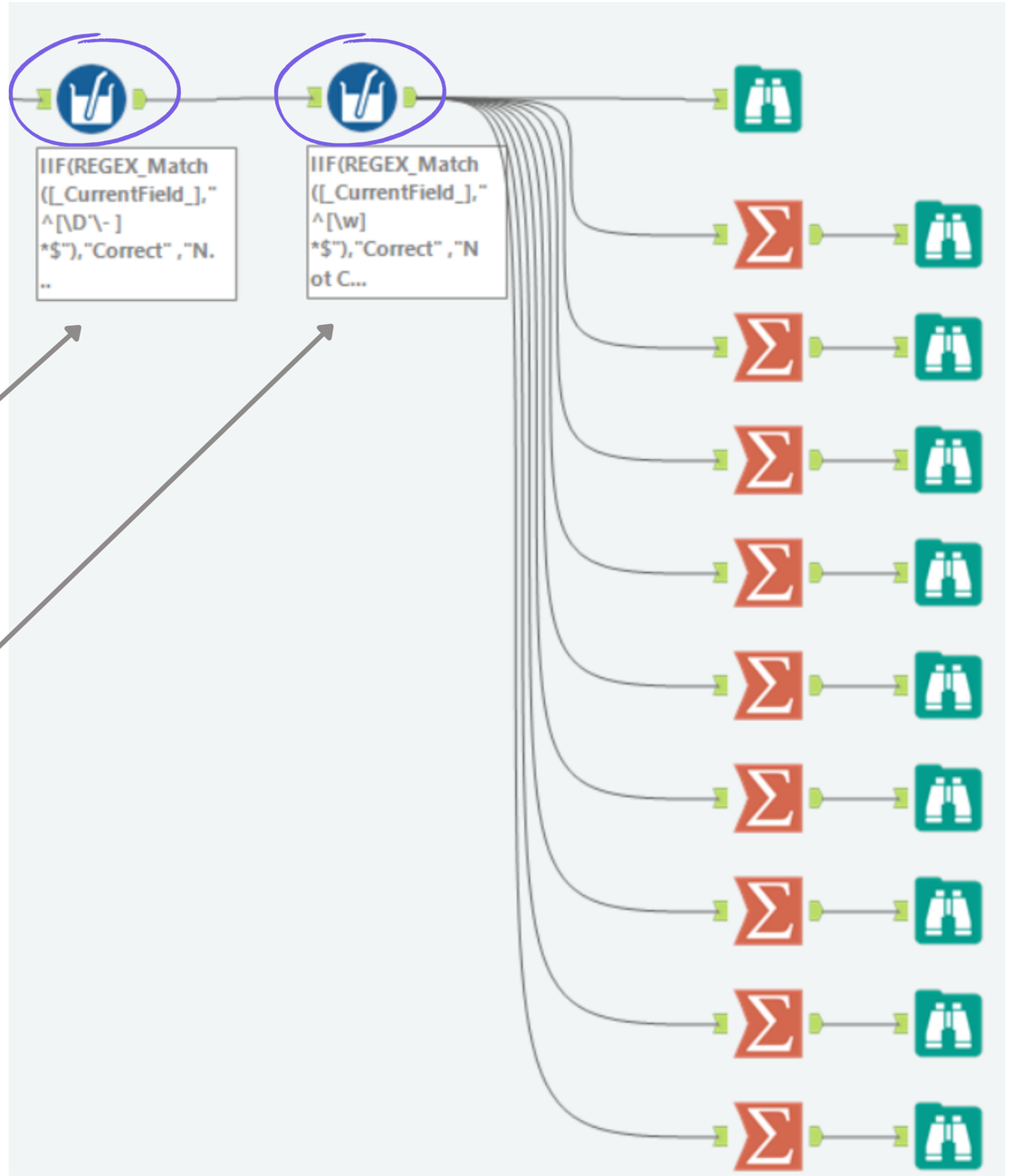
Outlier	Count
True	96519
False	10400624

2 Data Validation

- Check permitted characters

```
IIF(REGEX_Match([_CurrentField_],"^[\D'\- ]*$"),"Correct" ,"Not Correct")
```

```
IIF(REGEX_Match([_CurrentField_],"^[\w]*$"),"Correct" ,"Not Correct")
```



2 Data Validation

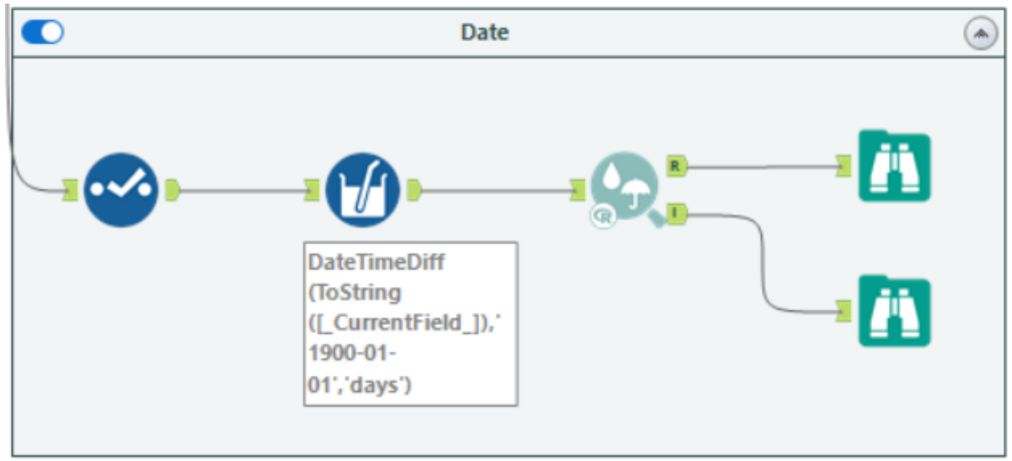
Check data range

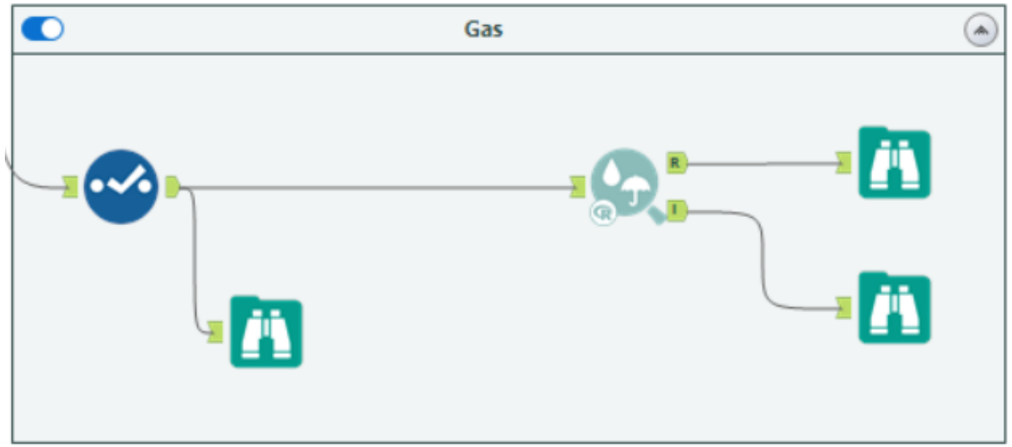
- «**age**» column have 111689(1,06% of the dataset) rows with value < 18
- The column '**total_amount**' defined as the sum of 'light_amount', 'gas_amount' and 'extra_fees' is correct
- The column '**howmuch_pay**' is defined as the sum of 'tv' and total_amount'. incorrect values: 257825 (2,46% of the dataset) and will be correct in Data Enrichment step
- For Numeric attributes having values less than zero

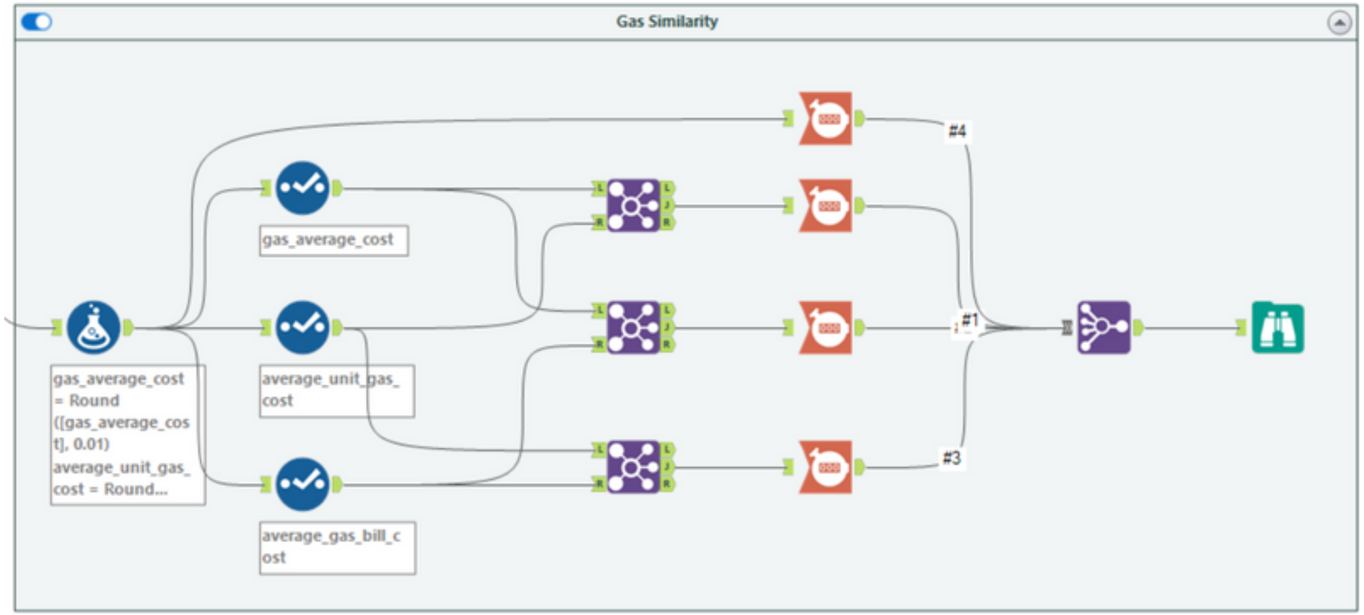
F1_kWh	103454 (0.99%)	**total_amount**	31006 (0.3%)
F2_kWh	96042 (0.91%)	**light_amount**	30359 (0.3%)
F3_kWh	93080 (0.89%)	**gas_system_charges**	2900681 (27,63%)
tv	384 (0.004%)	**light_system_charges**	56722 (0.54%)
gas_amount	152185 (1.45%)	**gas_material_cost**	140490 (1.34%)
extra_fees	351630 (3.35%)	**light_material_cost**	33499 (0.32%)
gas_consumption	338653 (3.23%)	**gas_transport_cost**	144439 (1.38%)
light_consumption	96921 (0.92%)	**light_transport_cost**	121300 (1.16%)
howmuch_pay	17 (0.00006%)		

2 Data Validation

- Check column uniqueness







OLD DUPLICATE COLUMN	REPLACEMENT COLUMN
date, light_start_date, gas_end_date	start_date
light_end_date, gas_start_date	end_date
gas_average_cost, average_unit_gas_cost	

2 Data Validation

- Find data-mismatched data types

gas_offer	Float	Name of the subscribed gas plan (anonymized)
light_offer	String	Name of the subscribed electricity plan (anonymized)

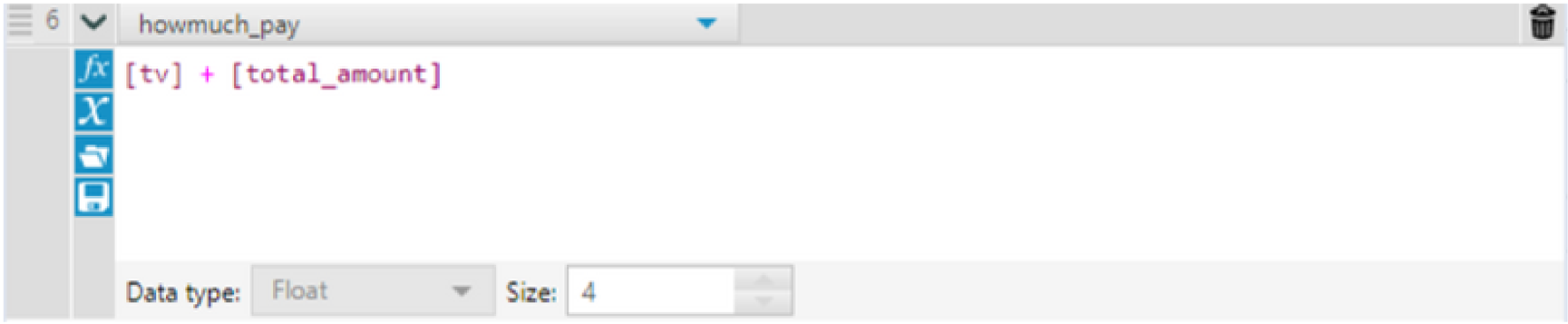
☑	gas_offer	V_WString	254		Name of the subscribed gas plan (anonymized)
☑	light_offer	V_String	254		Name of the subscribed electricity plan (anonymiz...

2 Data Structuring

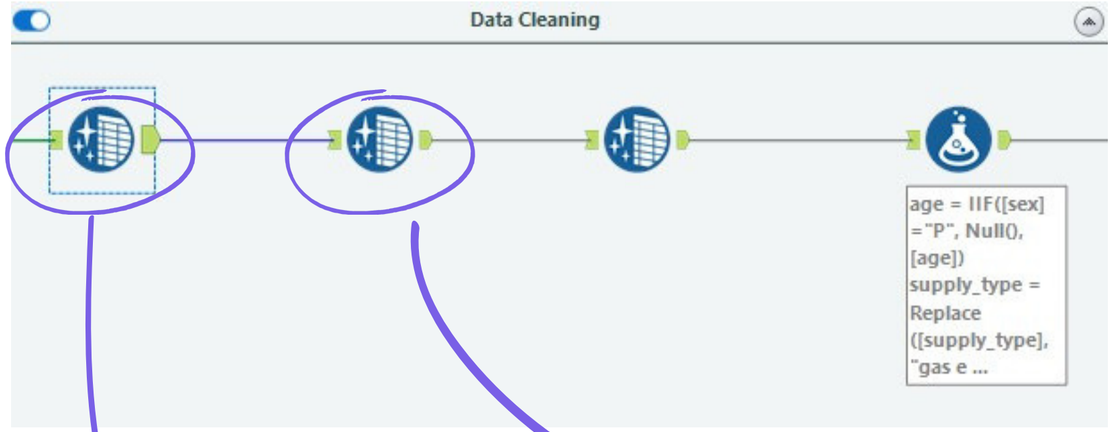
- **Change column datatype** — gas_offer [float] -> gas_offer[string]
- **Delete column** — date, gas_start_date, gas_end_date
- **Rename column** — F1_kWh -> f1_kwh
  F2_kWh -> f2_kwh
  F3_kWh -> f3_kwh
  light_start_date -> start_date
  light_end_date -> end_date

2 Data Enrichment

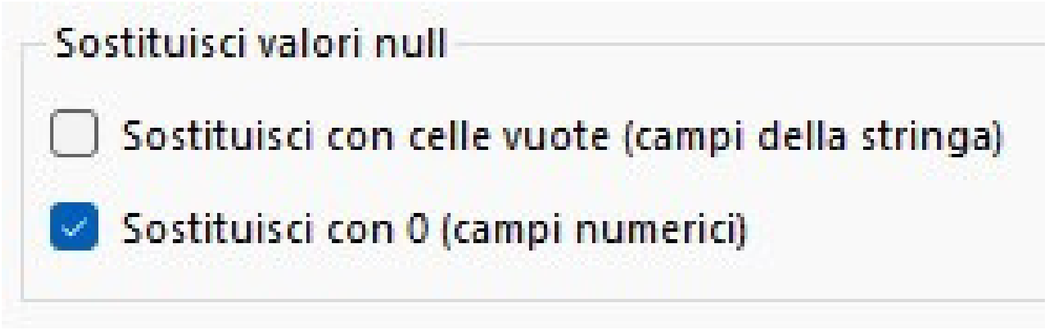
Since the data in the howmuch_pay column, calculated as the sum of 'total_amount' and 'tv', are wrong in 257825 (2,46% of the dataset), so we decided to re-calculate them to obtain the correct results.



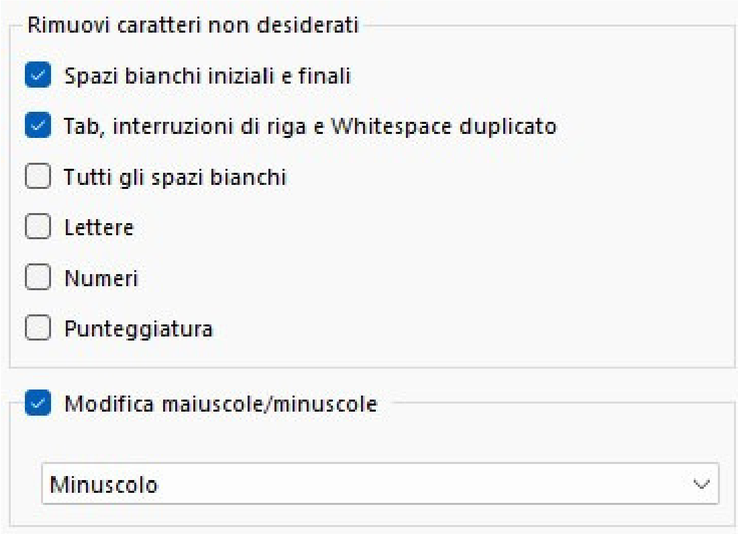
2 Data Cleaning



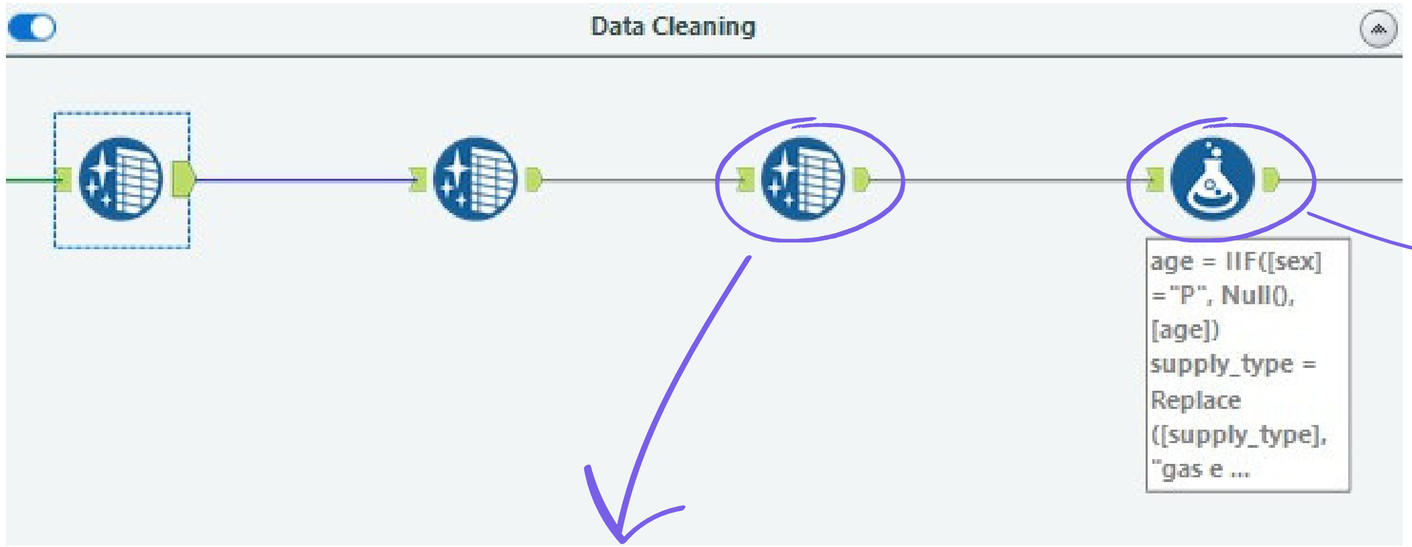
- Converted null values to 0



- Adjusted the String characters

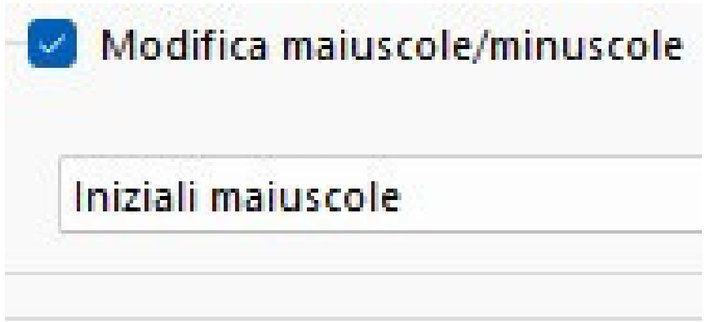


2 Data Cleaning

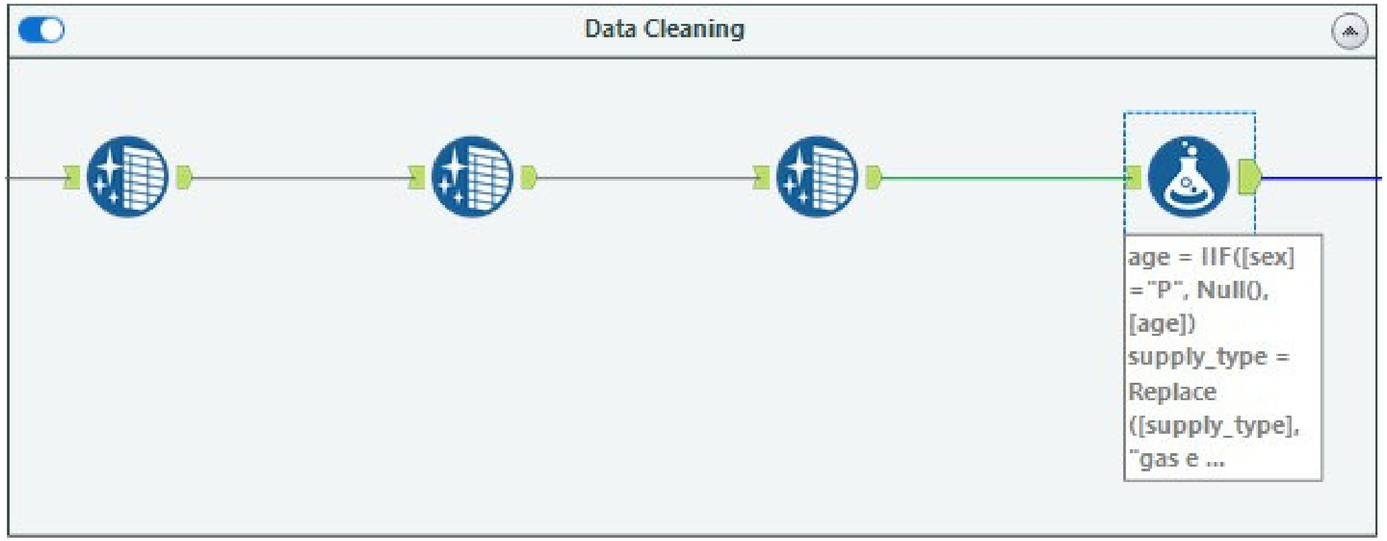


- Set sex=P
- Transalate the supply tupe

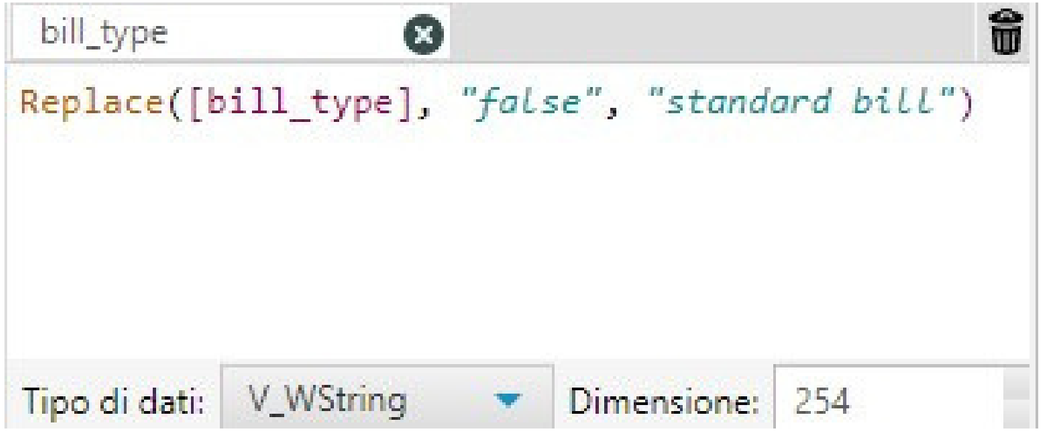
- Set Title Case



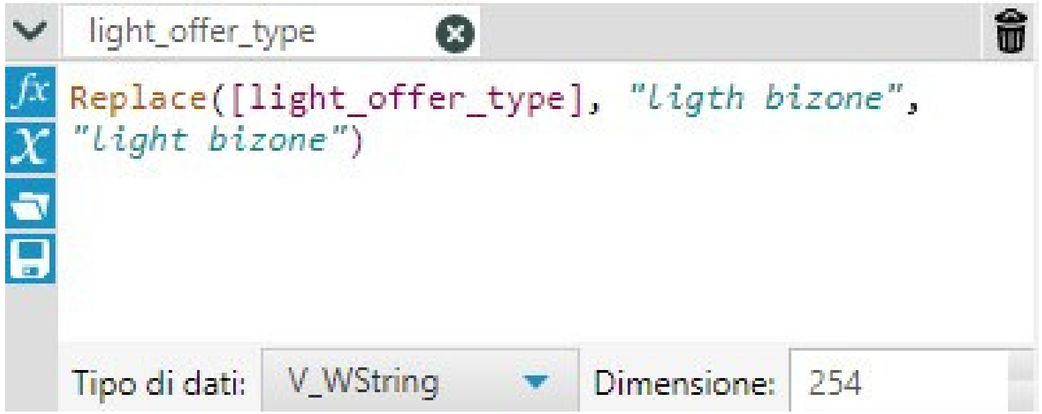
2 Data Cleaning



- Set Null Supply Type to "standard bill"

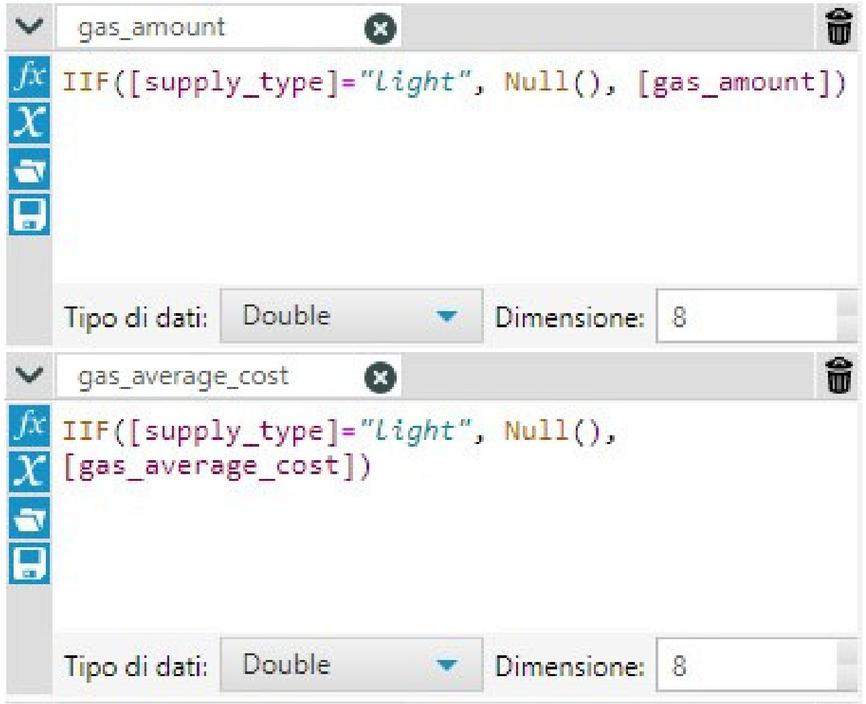


- Correct typo "light offer type"



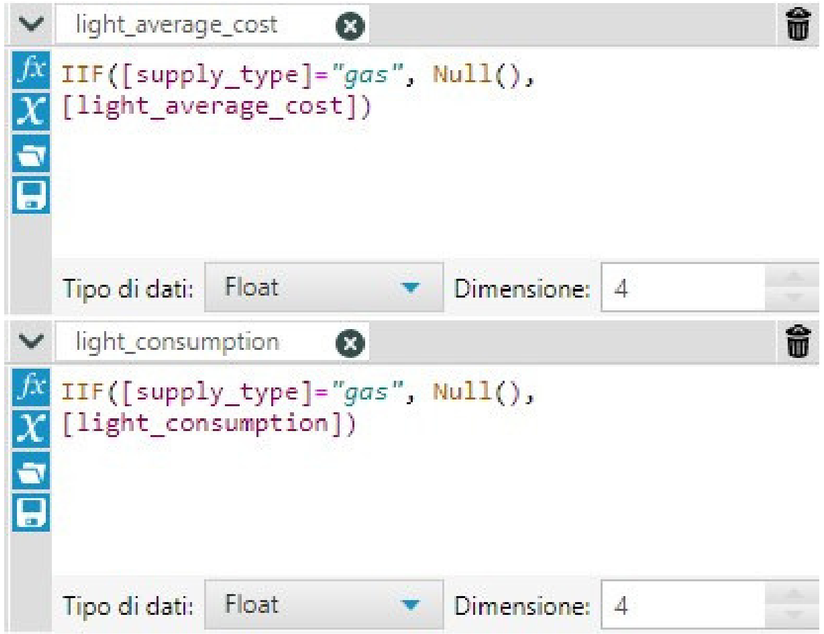
2 Data Cleaning

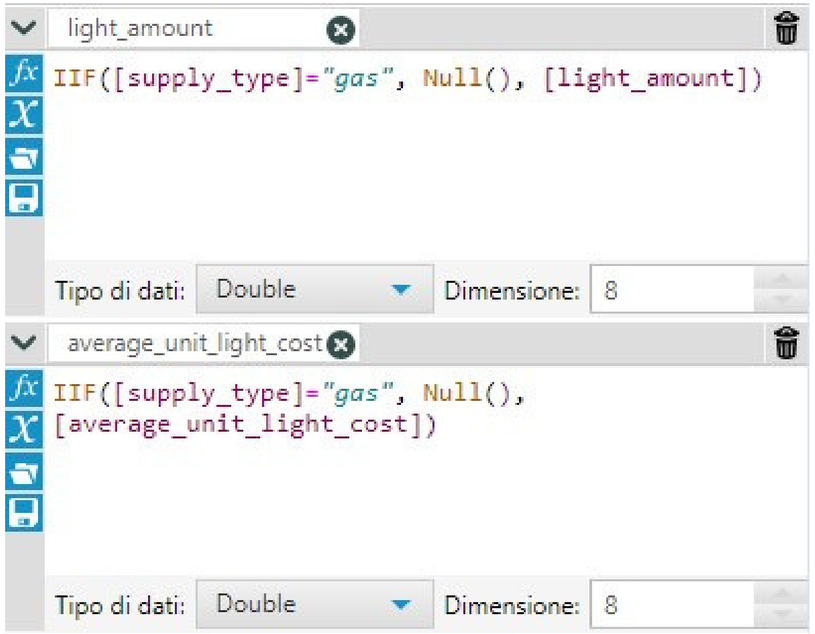
- Light invoces set null gas fields

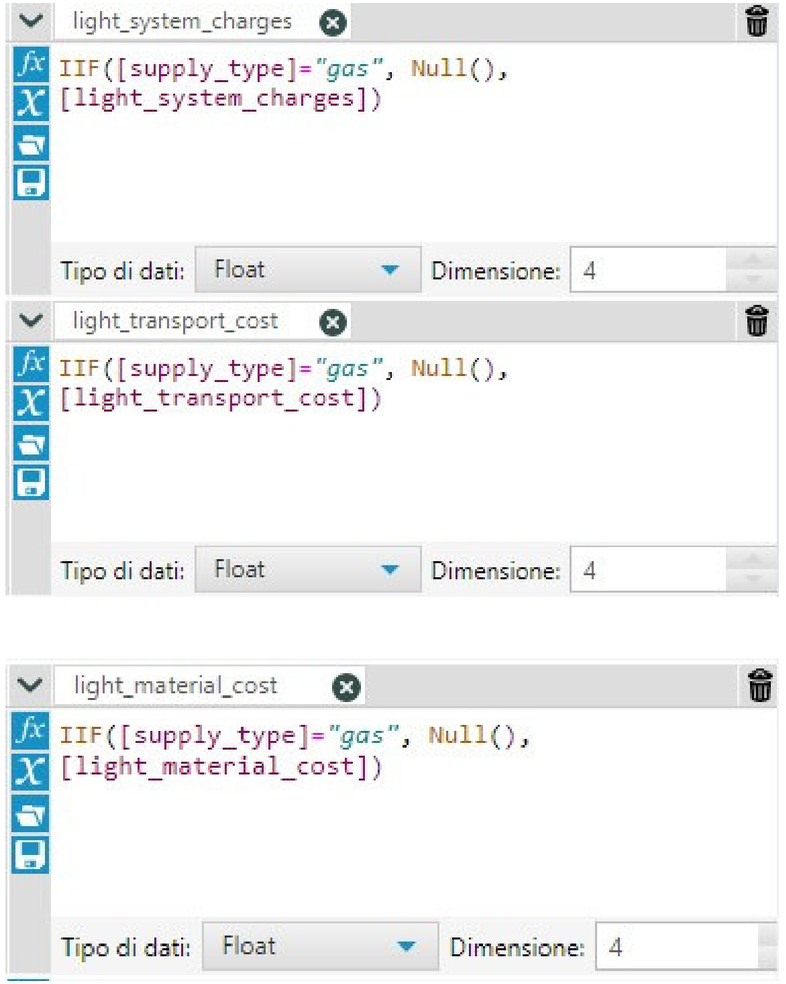


2 Data Cleaning

- Gas invoces set null light fields







Index

1

ALTERYX

- Overview
- How it works
- Learning Datasets
- Pros and cons

2

PIPELINE

- Data Ingestion and discovery
- Data Validation
- Data Structuring
- Data Enrichment
- Data Filtering
- Data Cleaning

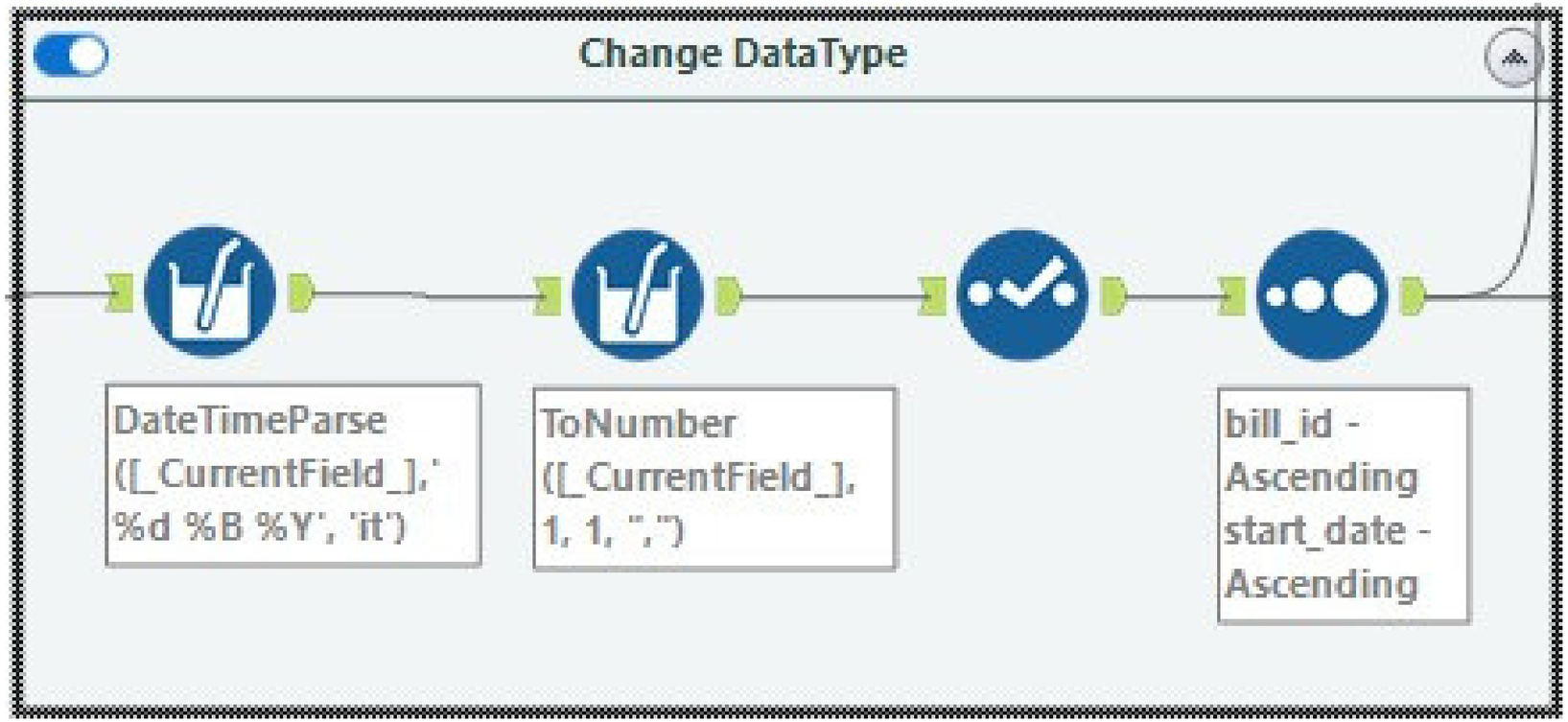
3

RESULTS

- Final pipeline
- Datasets obtained
- Conclusion

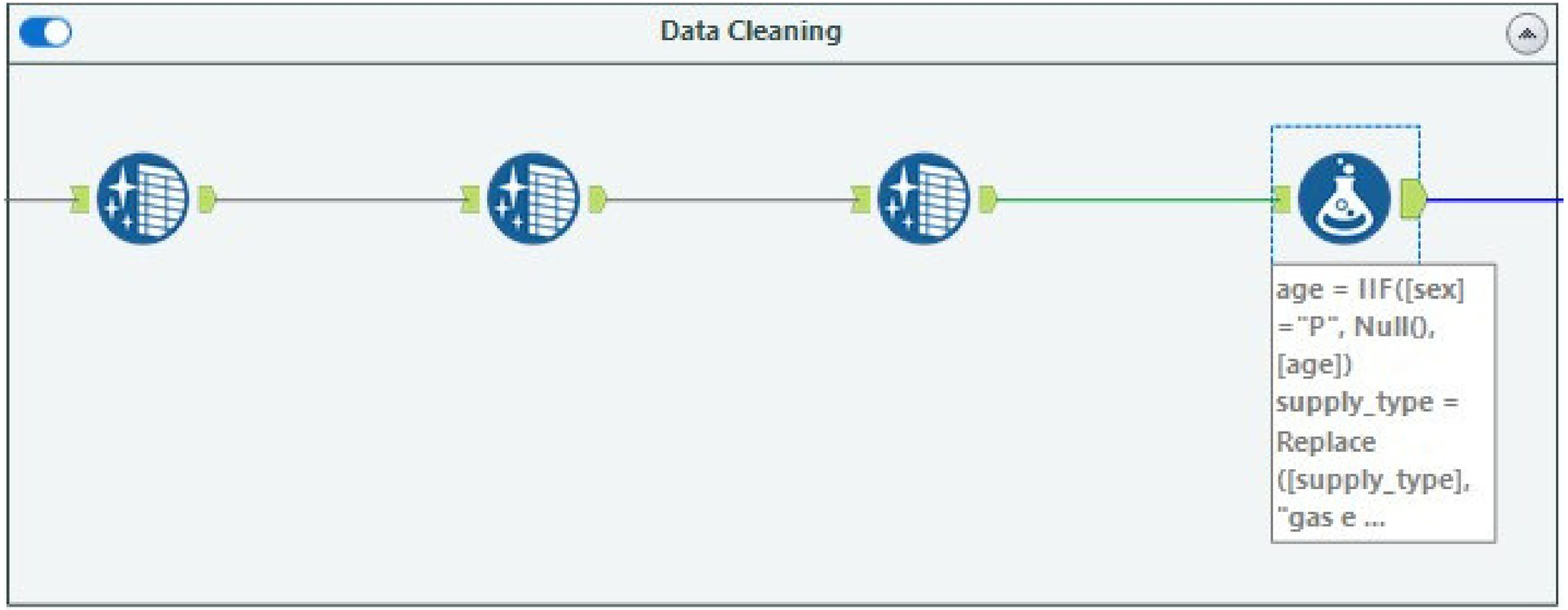
3 Final Pipeline

- Data format correction



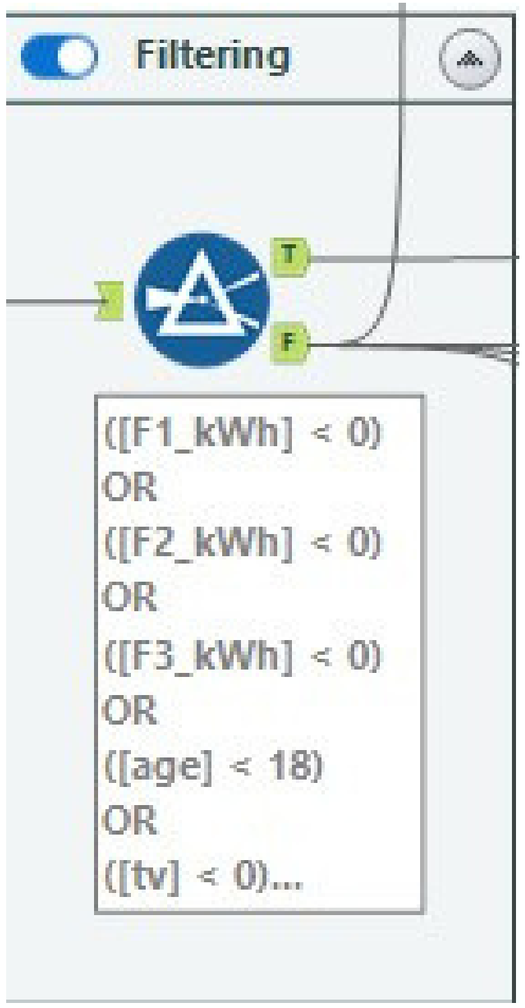
3 Final Pipeline

- Data cleaning



3 Final Pipeline

- Check the condition

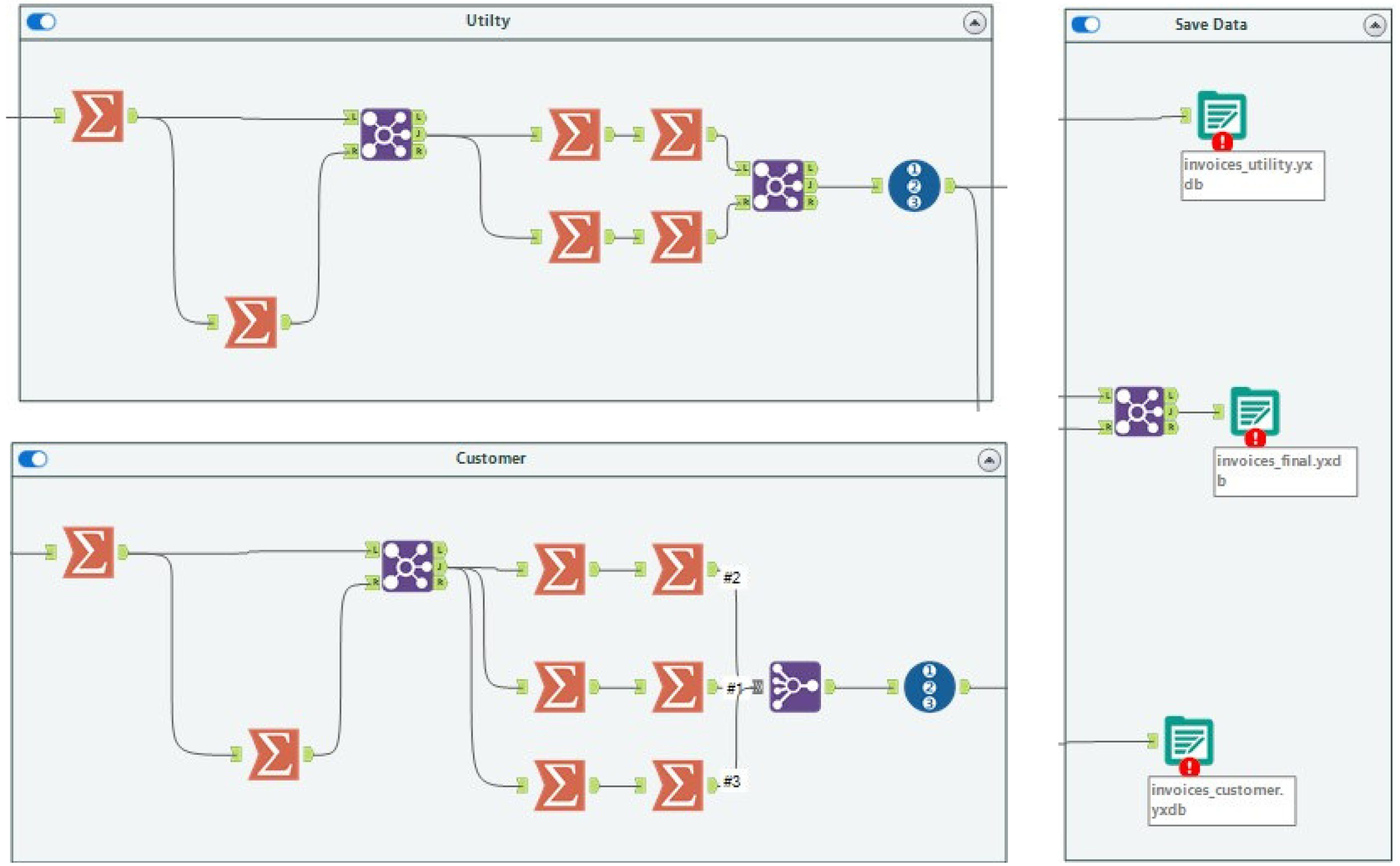


Filtro personalizzato

```
([F1_kWh] < 0) OR
([F2_kWh] < 0) OR
([F3_kWh] < 0) OR
([age] < 18) OR
([tv] < 0) OR
(IsNull([address])) OR
(IsNull([nominative]))OR
(IsNull([city]))OR
(IsNull([start_date])) OR
(IsNull([end_date])) OR
(([gas_consumption] < 0 OR
IsNull([gas_consumption])) and [supply_type] IN
('gas', 'gas and light')) OR
(([light_consumption] < 0 OR
IsNull([light_consumption])) and [supply_type]
IN ('light', 'gas and light')) OR
((Round(IIF(IsNull([light_consumption]), 0,
[light_consumption]), 0.01) !=
Round(IIF(IsNull([F1_kWh]), 0, [F1_kWh]) +
IIF(IsNull([F2_kWh]), 0, [F2_kWh]) +
IIF(IsNull([F3_kWh]), 0, [F3_kWh]), 0.01)) and
[supply_type] IN ('light', 'gas and light'))
```

3 Final Pipeline

- Creation of the 3 DBs



3 Datasets obtained

- **Customer** 3.898.716 of records

- **Utilty** 4.967.519 of records

- **Invoices** 9.781.634 of records

Pipeline execution time invoices is 15-30 minutes

3 Conclusion

- Offers the possibility of caching the intermediate result

- Execution times depend on hardware

- Use the maximum resources available

- AMP allows parallel pipeline execution

- Great community

- Does not offer complex tools.

- Saves temporary files between steps on disk

- Learning the interface and features may take time

Thanks for your attention!