# ETL (Extract, Transform, Load) Workflow

Contents

[ETL (Extract, Transform, Load) Workflow 1](#_Toc169032984)

[Read in Data 3](#_Toc169032985)

[Input Data Tool 3](#_Toc169032986)

[Picture 1. Input Data 3](#_Toc169032987)

[Input Data – Configuration 3](#_Toc169032988)

[Picture 2. Input Data (2) – Configuration 4](#_Toc169032989)

[Set Up a Connection 4](#_Toc169032990)

[Picture 3. Set Up a Connection 4](#_Toc169032991)

[Picture 4. Input Data – Annotation 5](#_Toc169032992)

[Picture 5. Workflow – Configuration 5](#_Toc169032993)

[Check of the Data Quality 5](#_Toc169032994)

[Picture 6. Column details 6](#_Toc169032995)

[Picture 7. Select – Configuration 7](#_Toc169032996)

[How to treat the error for conversion 7](#_Toc169032997)

[Picture 8. List of errors 7](#_Toc169032998)

[Picture 9. Gross profit 8](#_Toc169032999)

[Dates and Calculated Columns 8](#_Toc169033000)

[Convert date 8](#_Toc169033001)

[Picture 10. Date errors 9](#_Toc169033002)

[How to use the Help section 9](#_Toc169033003)

[Picture 11. Help Section 10](#_Toc169033004)

[Picture 12. Data Time Tool 10](#_Toc169033005)

[How create a new column 10](#_Toc169033006)

[Picture 13. DateTime Tool – DataTimeYear 10](#_Toc169033007)

[Picture 14. DataTimeMonth, DateTimeDay, DateTimeFormat 11](#_Toc169033008)

[Conditional formula 11](#_Toc169033009)

[Picture 15. Conditional formula 11](#_Toc169033010)

[Reorder the columns 11](#_Toc169033011)

[Picture 16. Change order of columns 12](#_Toc169033012)

[Outliers 12](#_Toc169033013)

[Filter Tool 12](#_Toc169033014)

[Picture 17. Filder tool 12](#_Toc169033015)

[Picture 18. Sorting from Result Section 13](#_Toc169033016)

[Missing Values 13](#_Toc169033017)

[Picture 19. Gross Profit 13](#_Toc169033018)

[Picture 20. Imputation tool 14](#_Toc169033019)

[Data Joins 14](#_Toc169033020)

[Picture 21. Remove column and Extract a text with Left function 14](#_Toc169033021)

[Picture 22. Join tool and configuration 15](#_Toc169033022)

[Picture 23. Union Tool 15](#_Toc169033023)

[Writing Data and Organizing the Workflow 15](#_Toc169033024)

[Picture 24. Output Data 15](#_Toc169033025)

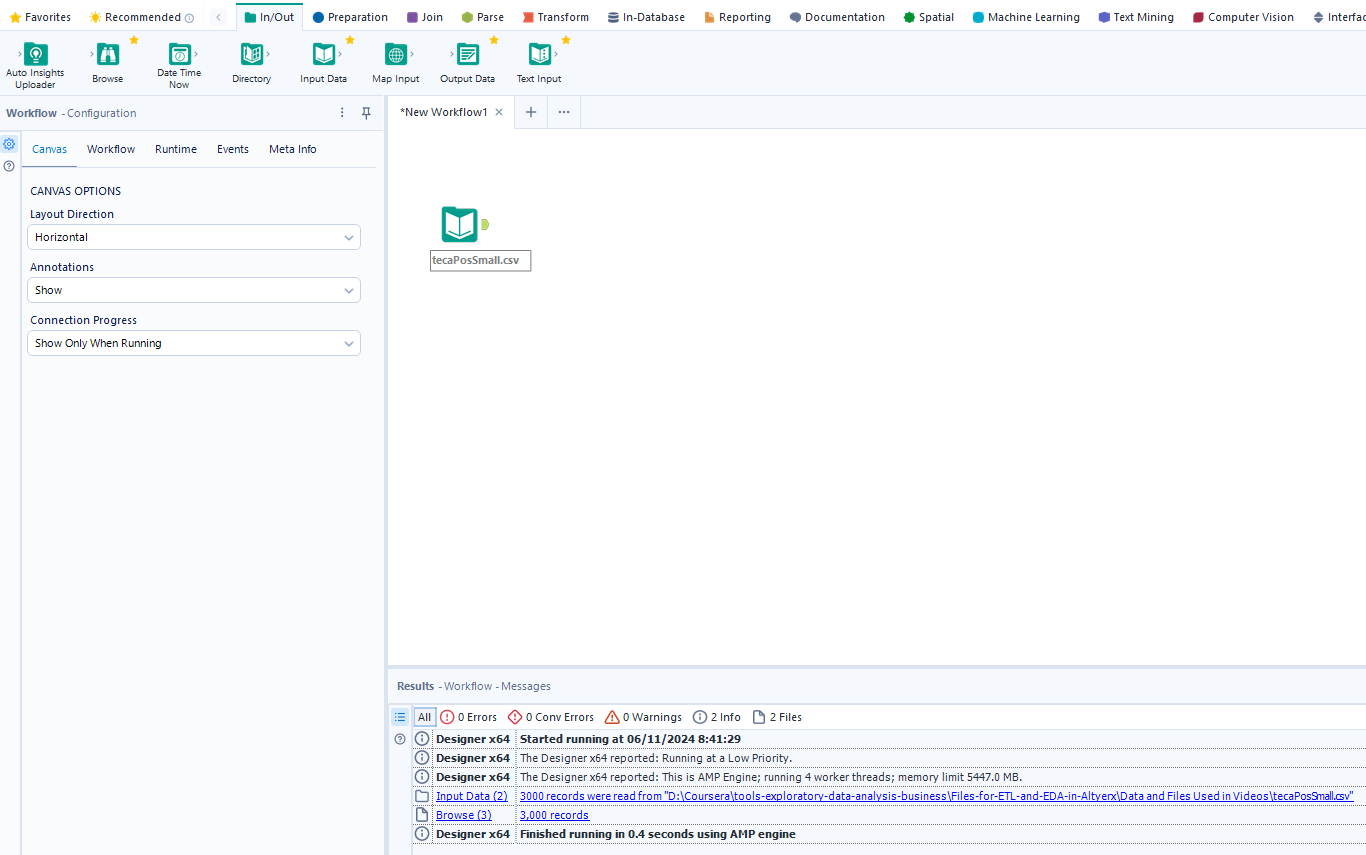
[Picture 25. Container 16](#_Toc169033026)

## Read in Data

In the first step of ETL workflow was imported the file with the TECA data from the file "tecaPosSmall.csv" that contains 3000 rows and 23 columns.

### Input Data Tool

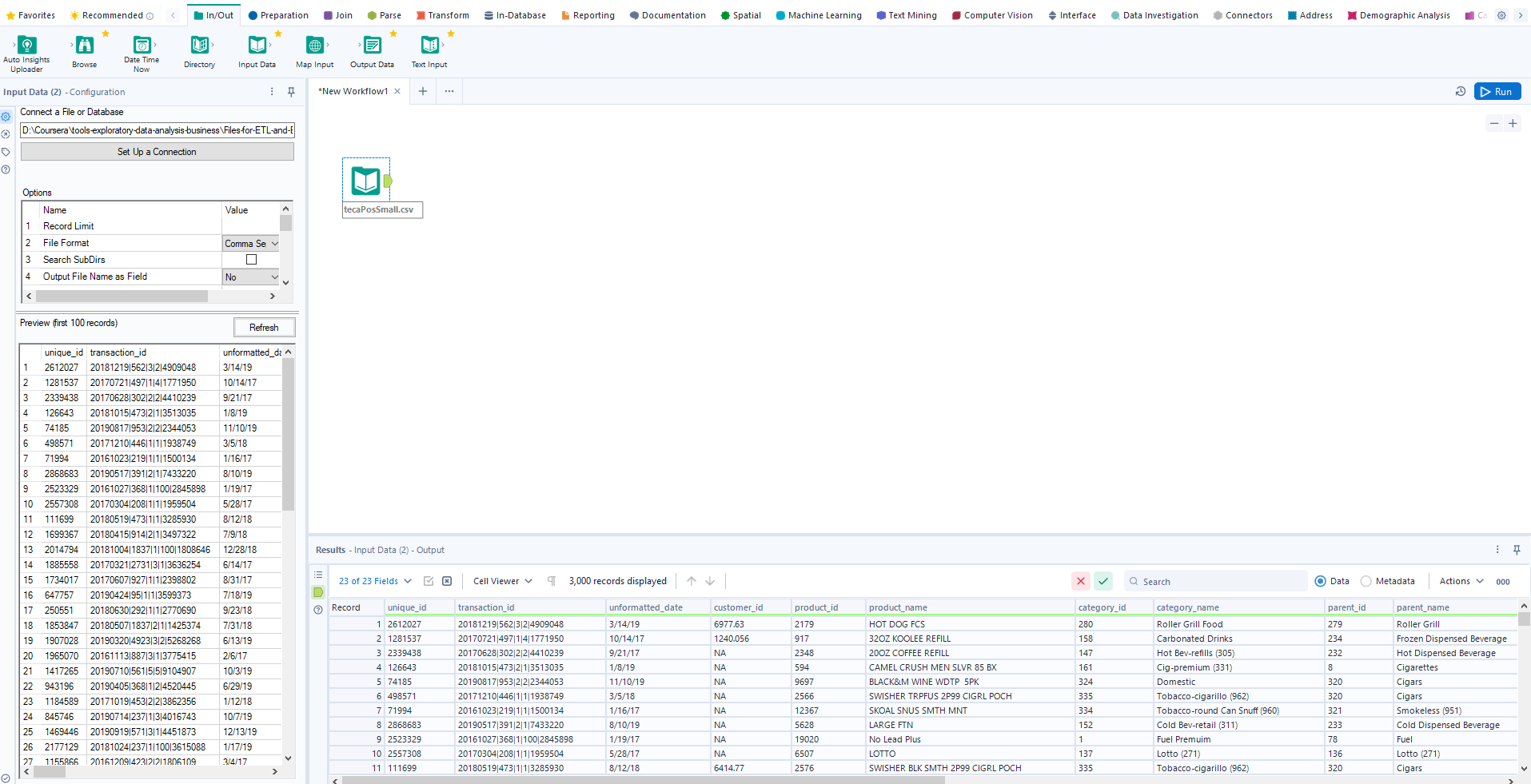
For import the input file on the Alteryx was drag on the canvas the step "Input Data" from the submenu "In/Out".



#### Picture 1. Input Data

### Input Data – Configuration

To configurate the step was necessary to set the connection to the input file pressing the button “*Set Up a Connection*” from the left menu “*Input Data (2) – Configuration*” ([Picture 2. Input Data (2) – Configuration)](#_Picture_2._Input)

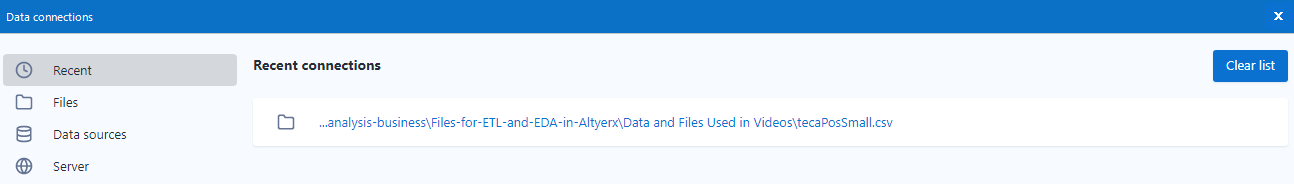


#### Picture 2. Input Data (2) – Configuration

### Set Up a Connection

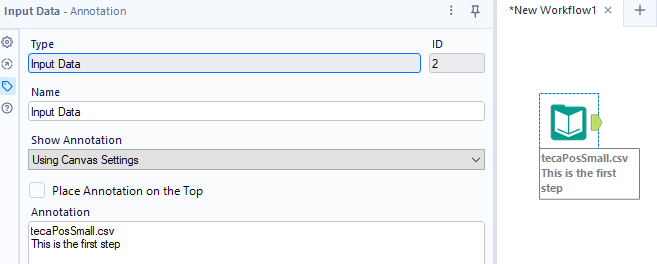
After the button “*Set Up a Connection*” ([*Picture 3. Set Up a Connection*](#_Picture_3._Set))*,* Alteryx app open a new window “*Data Connections*” where it could be found the type the connection that contain:

1. Recent – here could be found all the recent connections that exist
2. Files – offer the possibility to drag and drop the input file or to go in the folder where it’s the input file; Also, in the tab Alteryx show some examples of supported files like: *JSON, Microsoft Excel, SAS, SQLite, QlikView, Text, Zip, etc.*
3. Data sources – offer the possibility to connect to various type of Data bases like: *Oracle, MySQL, Microsoft SQL Server, Amazon, IBM, Microsoft Azure Data Lake Store, SAP IQ, SAP Hana, etc.*
4. Servers – offer to possibility to connect directly to a server through a URL



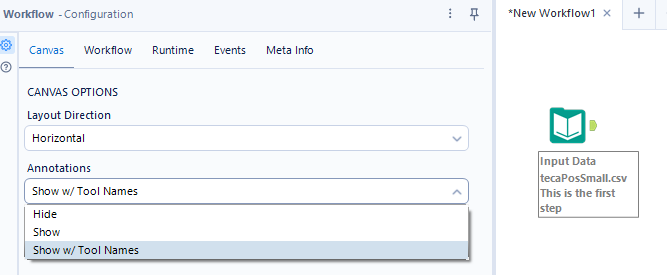
#### Picture 3. Set Up a Connection

In the left site of the menu “[*Input Data (2) – Configuration*](#_Picture_2._Set)*”* we have “Input Data – Annotation” menu that allow us to write a short description about the step that we did. Under the name of the input file was added “*This is the first step*” to be used to identify which is the first step of the workflow ([Picture 4. Input Data - Annotation](#_Picture_4._Input))



#### Picture 4. Input Data – Annotation

For a better and useful insight was added also the name of the tool that was used. To do this was need to go in the Workflow – Configuration where exists the Annotations option which allow us to “*Hide*” the name, to “*Show*” only the name of the input file and also the annotation of it, and, also we can show Tool names if choose “*Show w/ Tool Names*” ([Picture 5. Workflow - Configuration](#_Picture_5._Workflow))



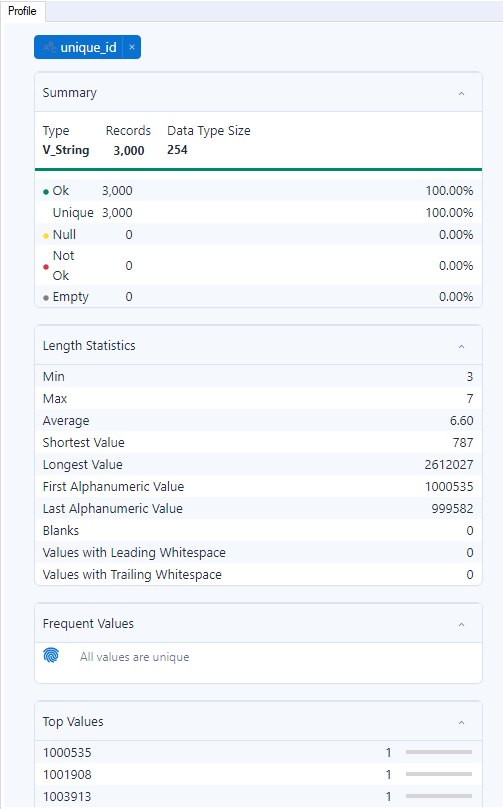
#### Picture 5. Workflow – Configuration

### Check of the Data Quality

To understand better the data from the input file and also to see the quality of them was choose to use the tool “Browse”. This tool allows us to see for every column from the file a summery that contains:

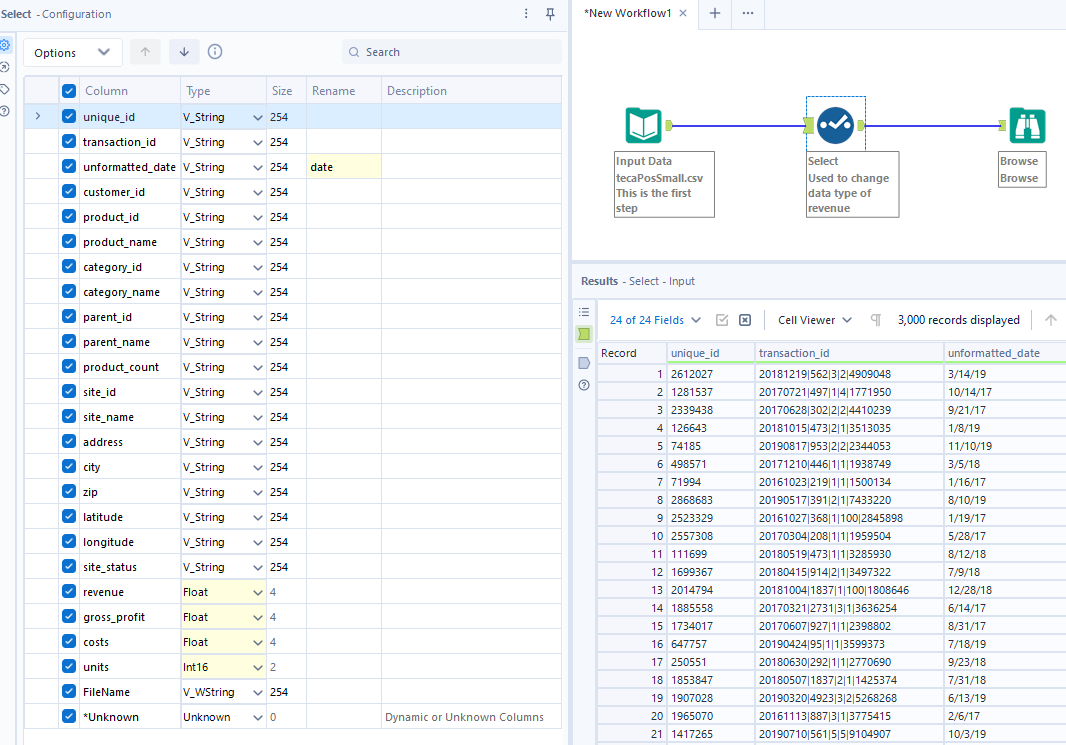
* Total records;
* Total fields;
* The dimension of file;
* For each column the distinct values, how many times appear every distinct value, what type are the data from columns, the quality of them with also the percentage;

To see more details about the one column we can click on it and the Alteryx made us a Summary regarding the quality, statistics, frequent values and top values ([Picture 6. Column details](#_Picture_6._Column))



#### Picture 6. Column details

To change data type for a column was used the “Select” tool that offer us the possibility to keep a column or not, to change data type of not. In our case was change the data type for columns: *revenue, gross\_profit, costs, units* and rename the column *unformatted\_date* in *date.* ([Picture 7. Select – Configuration](#_Picture_7._Select))

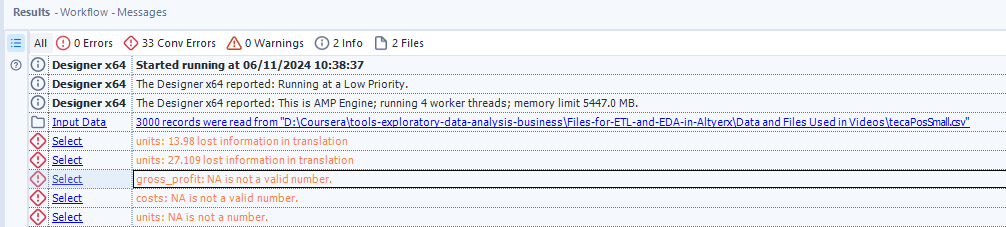


#### Picture 7. Select – Configuration

### How to treat the error for conversion

In the “Select” tool step we observed a set of errors like (Picture 8. List of errors):

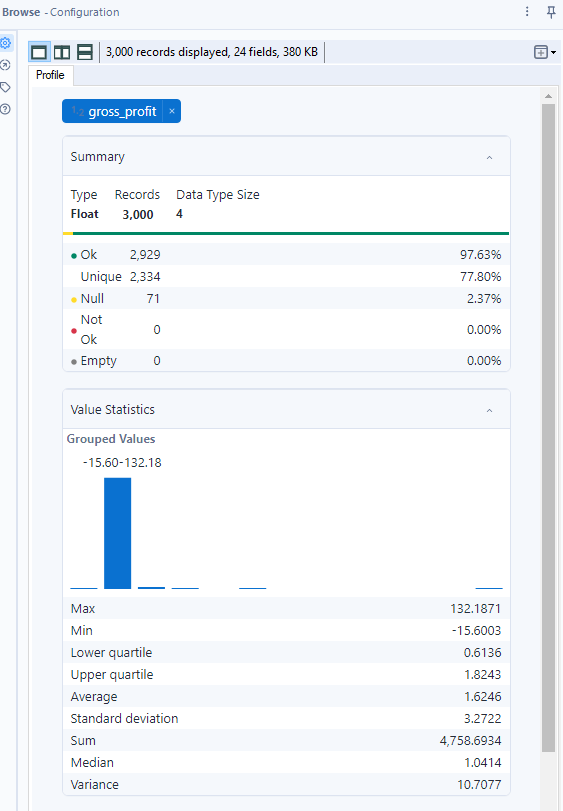
* ConvError: Select: units: 13.98 lost information in translation
* ConvError: Select: gross\_profit: NA is not a valid number.
* ConvError: Select: costs: NA is not a valid number.



#### Picture 8. List of errors

To solve the “ConvError: Select: units: 13.98 lost information in translation“ was needed to change data type from “*Int16*” to “*Float*” with the same technic like in [Picture 7. Select – Configuration](#_Picture_7._Select)

In the case of the *gross\_profit* and *costs* the values that weren’t convert was transform in *null* values by default. In the Browser – Configuration it can be see the overview of the data after the conversion ([Picture 9. Gross profit](#_Picture_9._Gross))

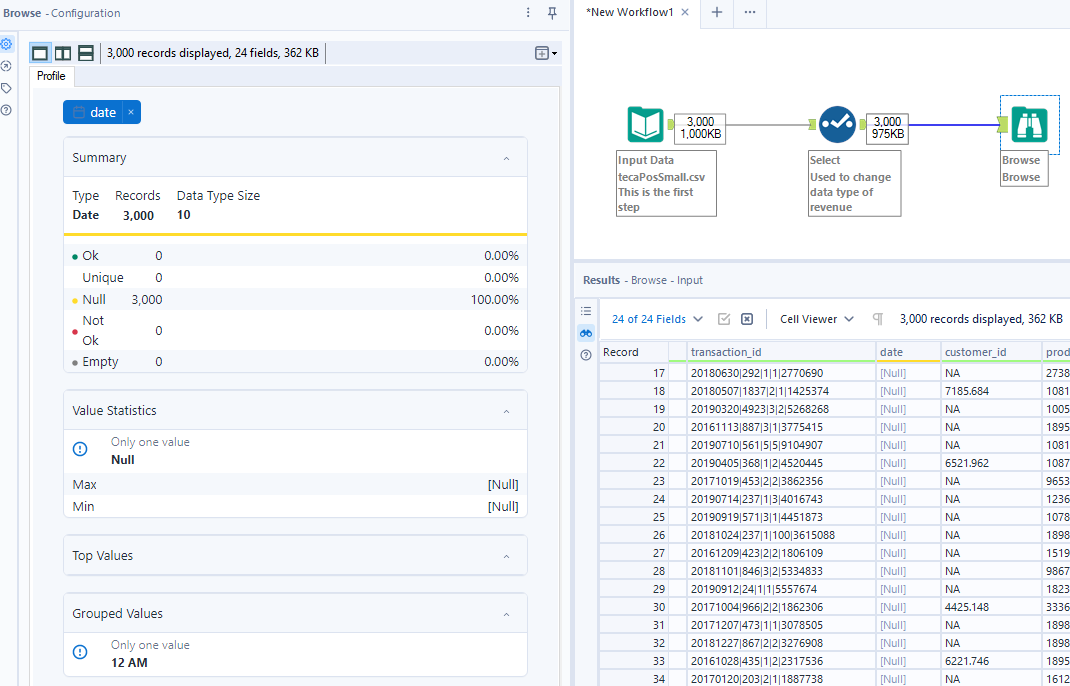


#### Picture 9. Gross profit

## Dates and Calculated Columns

### Convert date

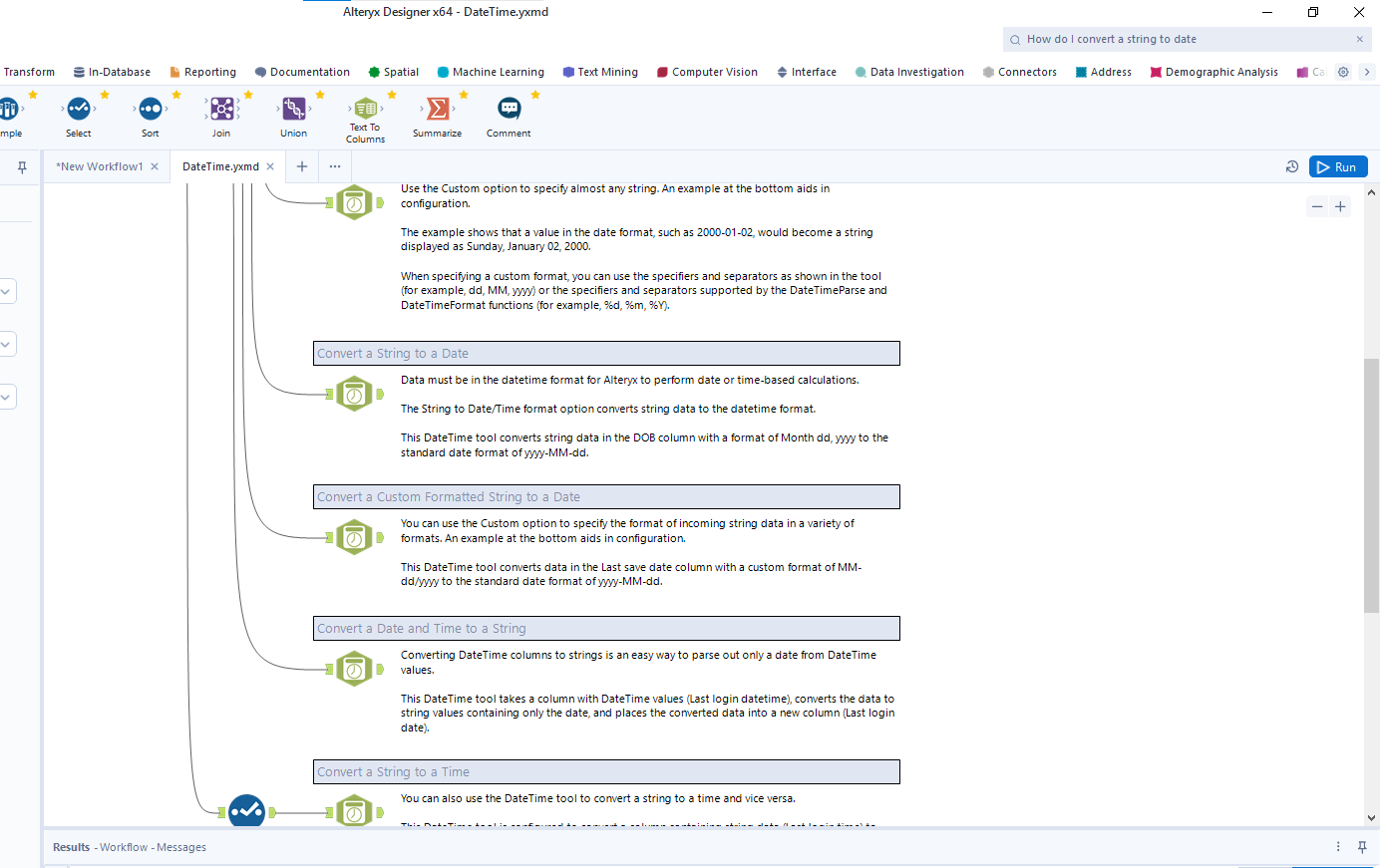
First step when we need to work with date column it’s to verify it’s converted correctly. In our case the date is “V\_String” type and to converted was used the same steps like in [Picture 7. Select – Configuration](#_Picture_7._Select). In this case because the dates weren’t in the correct format (ISO 8601 Format: YYYY-MM-DD) we observed that all the dates had *null* value. ([Picture 10. Date errors](#_Picture_10._Date))



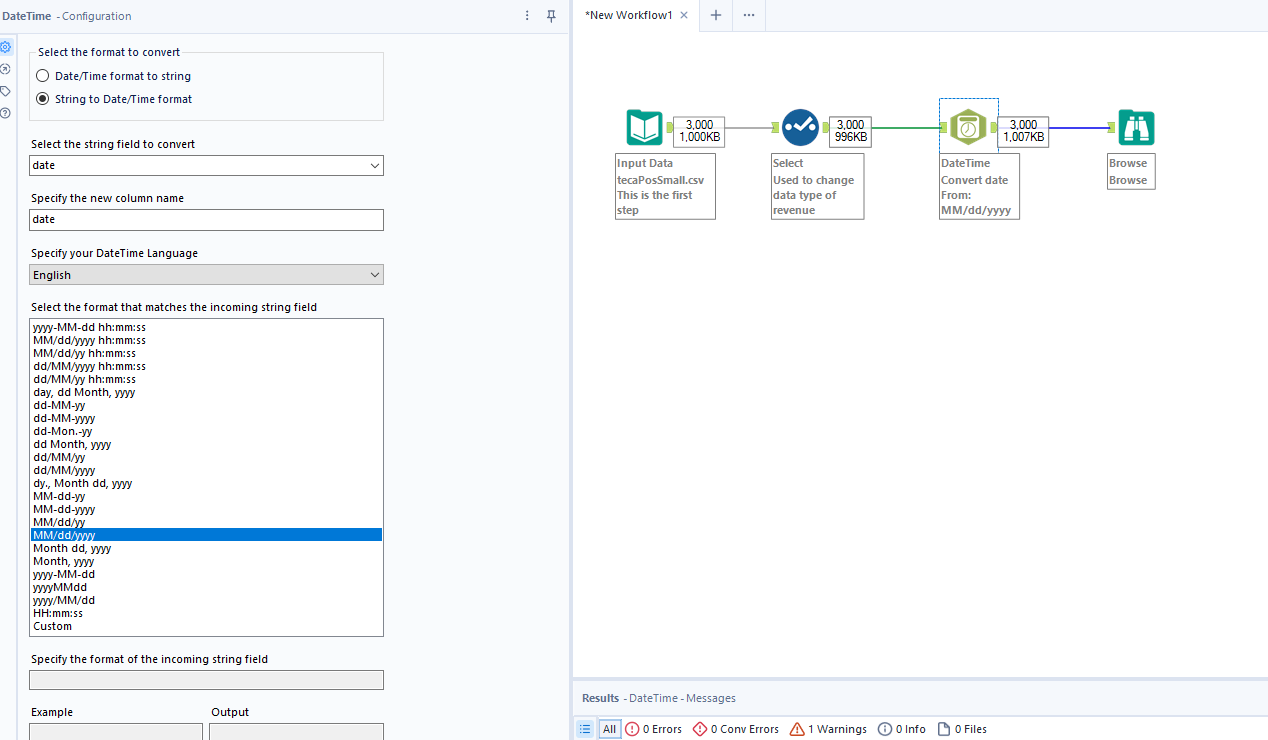
#### Picture 10. Date errors

### How to use the Help section

To solve this problem, we search it on the *Help search bar* and the app will open a new workflow where it was found *“Convert a String to Date”* solution ([Picture 11. Help Section](#_Picture_11._Help)). There is a *Date Time* tool that could be found in *Parse* menu. On the *Date Time* we can change the name of the column if we want and also we need to identify which is the format of the date ([Picture 12. Data Time Tool](#_Picture_12._Data))

**

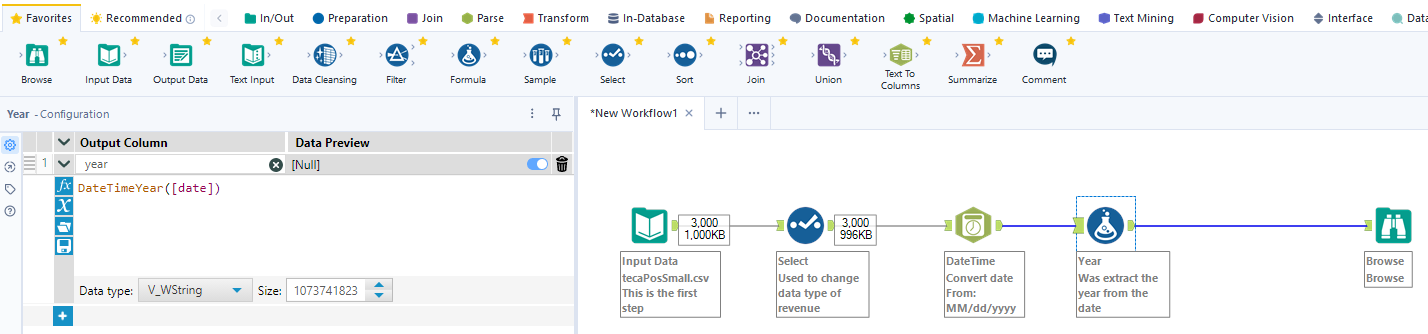
#### Picture 11. Help Section



#### Picture 12. Data Time Tool

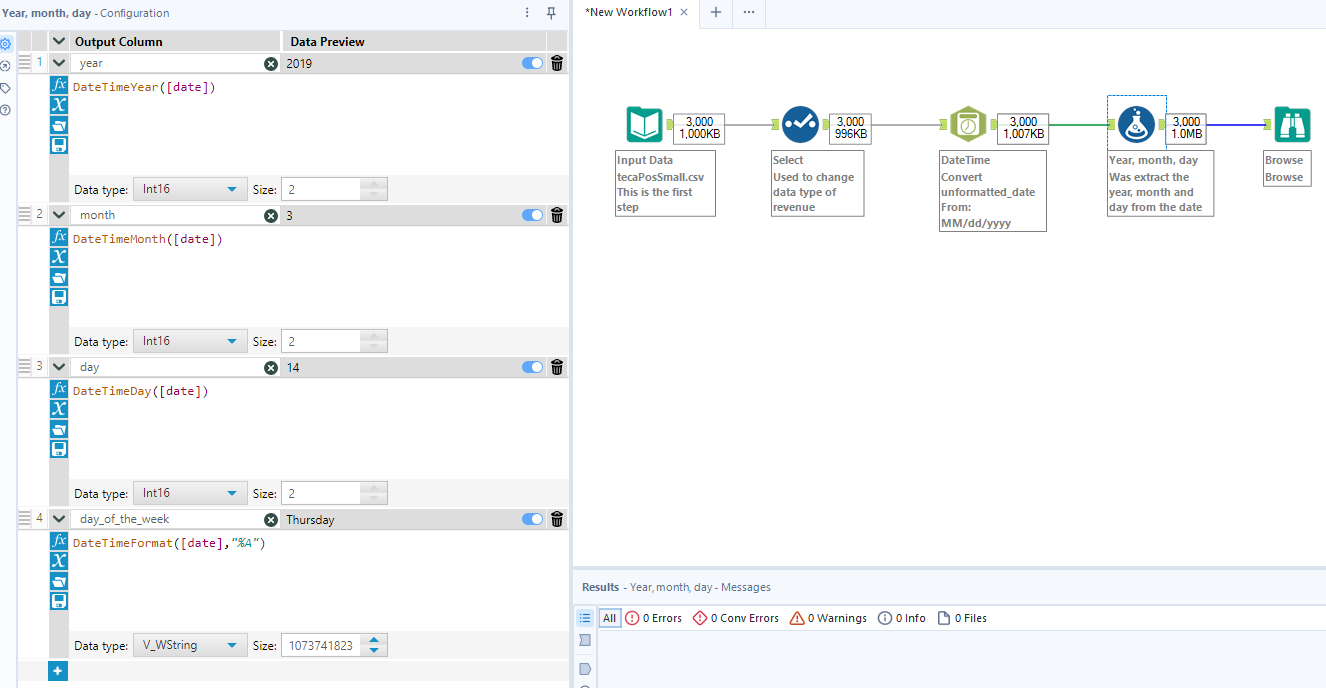
### How create a new column

To create a new column that contains the *year* of the *date* column in Alteryx was inserted a new step “*Year*” with the help of “*Formula*” tool. The formula “*DataTimeYear*” was used to extract the year. Also, this “*Formula”* tool had the possibility to choose the final data type; In this case the new column, year, has “*V\_WString”* as data type. ([Picture 13. DateTime Tool - DataTimeYear](#_Picture_13._DateTime)).



#### Picture 13. DateTime Tool – DataTimeYear

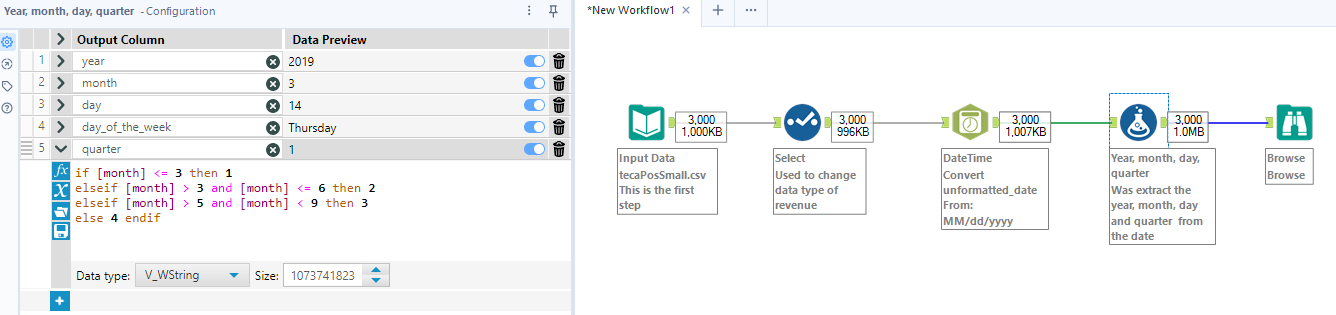
To extract the day of the month we used the formula “*DataTimeDay*”. Also, in this step were change the data type for column “year from “V\_WString” as data type in “Int16”. In the same manner it was created the column for month of year and column for day of the week. ([Picture 14. DataTimeMonth, DateTimeDay, DateTimeFormat](#_Picture_14._DataTimeMonth,))



#### Picture 14. DataTimeMonth, DateTimeDay, DateTimeFormat

### Conditional formula

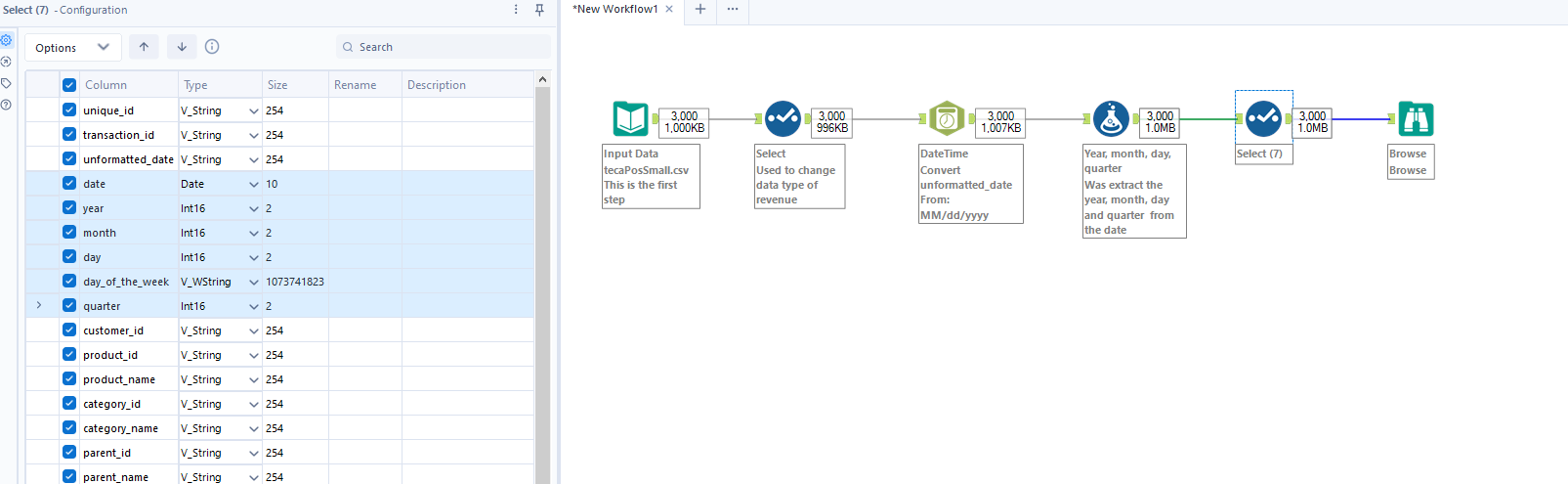
To obtain the column that contain the quarter was inserted a conditional formula that help us to identify it. ([Picture 15. Conditional formula](#_Picture_15._Conditional))



#### Picture 15. Conditional formula

### Reorder the columns

Because the columns: *day, day of the week, month, quarter, year* are on the end of the data set and we want them neither original date, the order of the column was change inserting a new “Select” tool and inside of it by selected them and using the “Right click” were dragging on the new position. ([Picture 16. Change order of columns](#_Picture_16._Change))

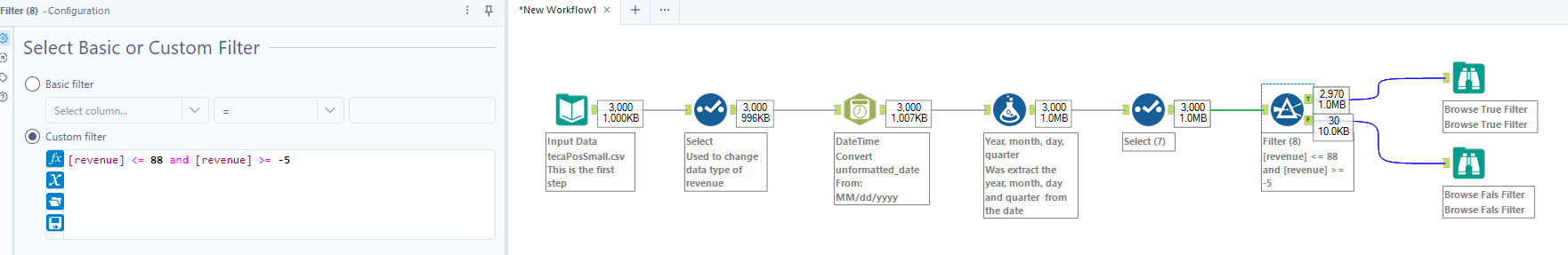


#### Picture 16. Change order of columns

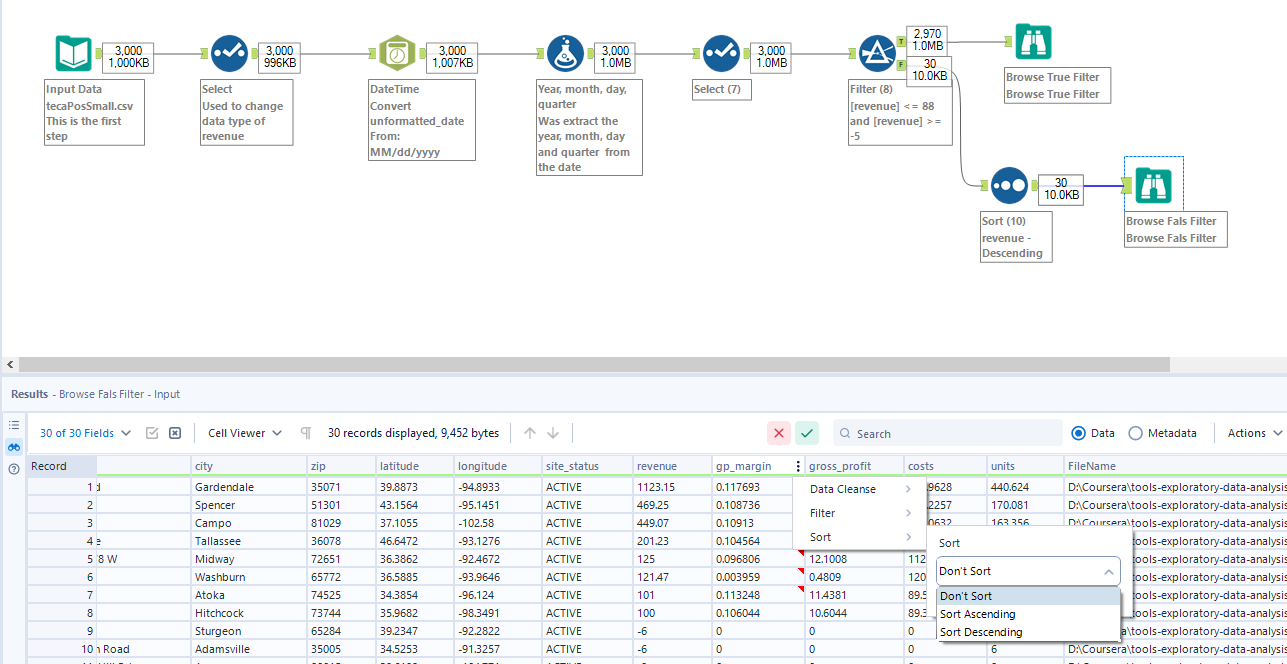
## Outliers

### Filter Tool

To eliminate the revenue values that are not relevant for our analysis it was used Filter tool that help us to write a formula that exclude the outlier values. ([Picture 17. Filder tool](#_Picture_17._Filder)) To see the values and verify if the Filter works properly was added also a new Browser tool that would show the outlier values. Also, to see much easy if the values are in the interval that it was set with the formula, it could be sorting the data from the Result section from the bottom side ([Picture 18. Sorting from Result Section](#_Picture_18._Sorting)). To keep that manipulation data it could be done only if you click and confirm the changes by clicking the validation bottom from the center, near “Search bar”. After you confirm de changes, Alteryx will insert a new step that used “Sort” tool ([Picture 18. Sorting from Result Section](#_Picture_18._Sorting))



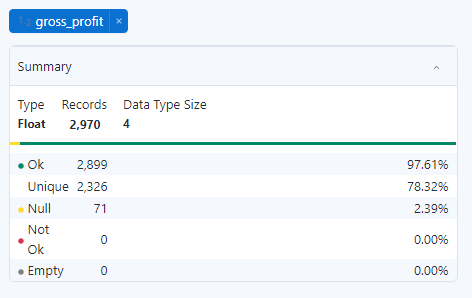
#### Picture 17. Filder tool



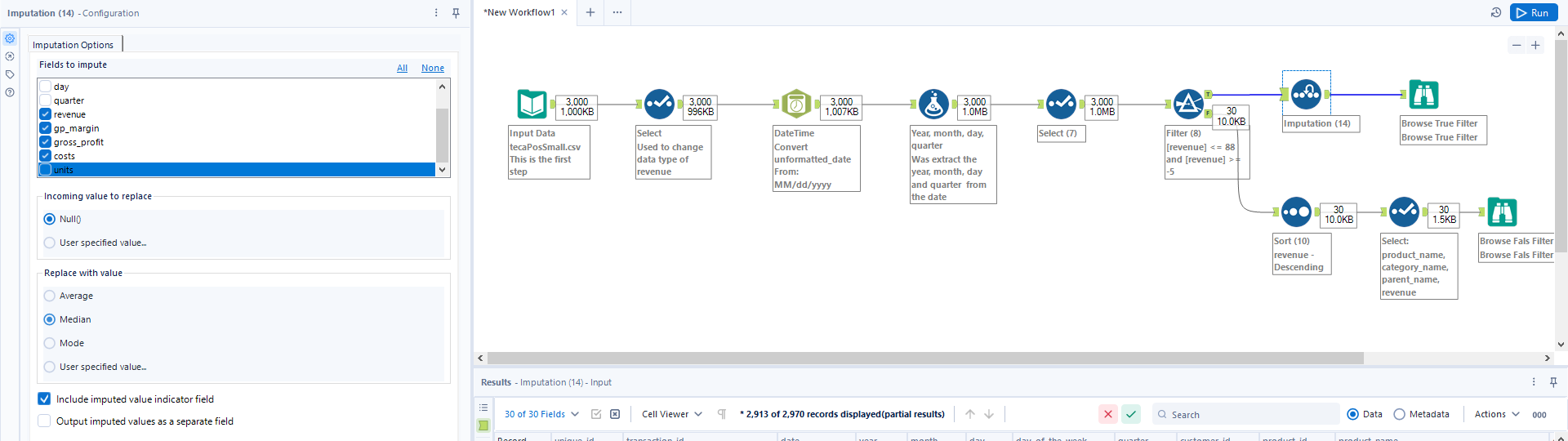
#### Picture 18. Sorting from Result Section

## Missing Values

To handle the missing values from the column gross\_profit where the summary of the column shows us that 71 (2.39%) are the null values from the total of 2899 records (Picture 19. Gross Profit) it was necessary to remove the missing them with the “*Filter*” tool or “*Imputation*” tool that offer us the possibility to replace the null values with Average, Median, Mode or a Specified value ([*Picture 20. Imputation tool*](#_Picture_20._Imputation)). The best approach in this case it is to use the median because average could be influence by the outline values.



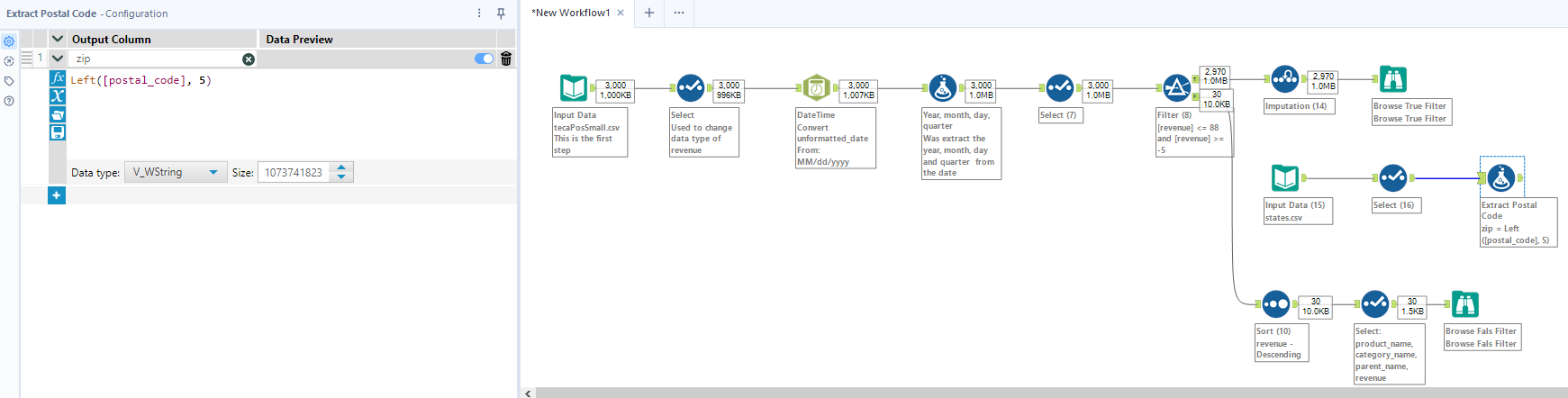
#### Picture 19. Gross Profit



#### Picture 20. Imputation tool

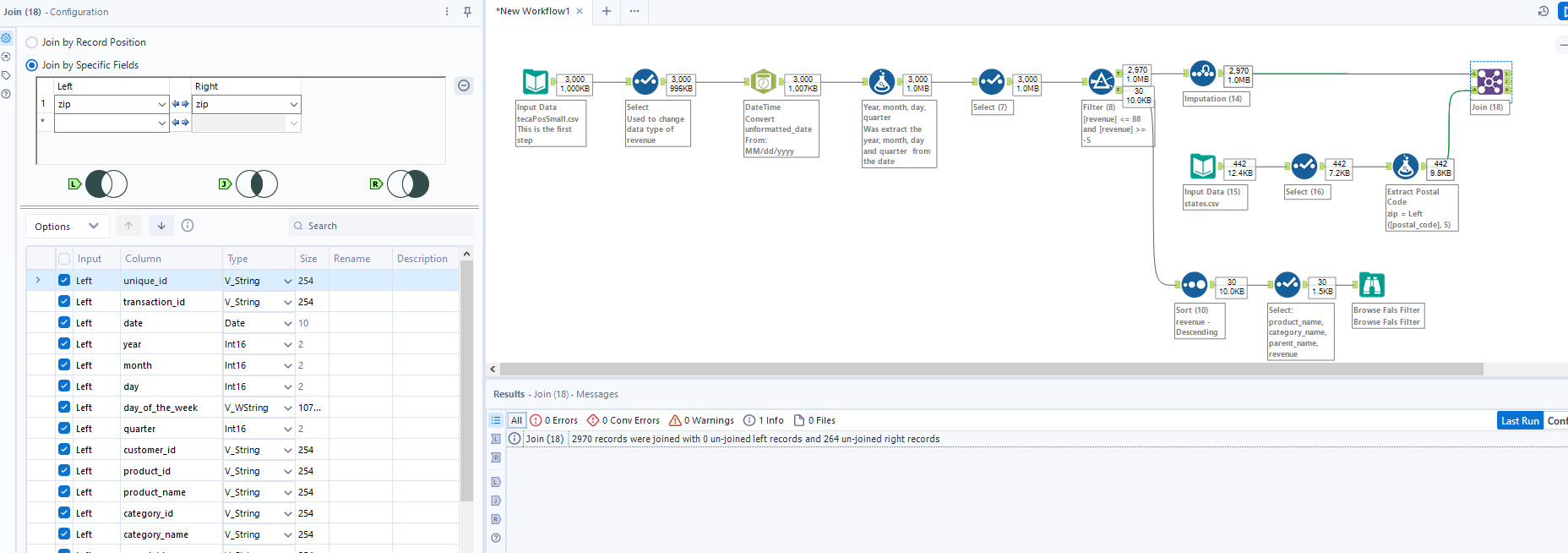
## Data Joins

To use a join was need to add a new input data set and for do this was used the steps like in the case [Input Data Tool](#_Input_Data_Tool), inset a “Input Data” tool that connect us to the *states.csv* file where it could be found information about the states province from USA. On the file *states.csv* are some observations from *postal\_code* that have more then 5 digits and need to be transform in the correct format, also to remove the columns that aren’t relevant to current analysis. To remove the column was used “Select” tool and for extraction of zip code from *postal\_code* column was used a “Formula” tool and function *Left* that offer us the possibility to extract n character from the start of the observation. ([Picture 21. Remove column and Extract a text with Left function)](#_Picture_21._Remove)

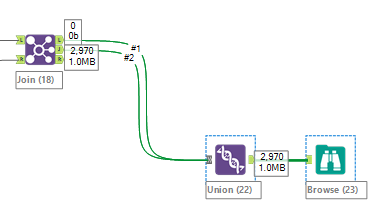


#### Picture 21. Remove column and Extract a text with Left function

Now, because it was clear properly the input from *states.csv* file, it can be done the join between the 2 input and for do this it was used the “Join” tool that need to be configurate by choose relation columns. Primary key for the *states.csv* and foreign key for *tecaPosSmall.csv* file. ([Picture 22. Join tool and configuration](#_Picture_22._Join)). To keep the observation of *tecaPasSmall.csv* file and also the observations that are common it was necessary to insert a “Union” tool that confer the possibility to union the Left Unjoined and the Inner Join. ([Picture 23. Union Tool](#_Picture_23._Union))



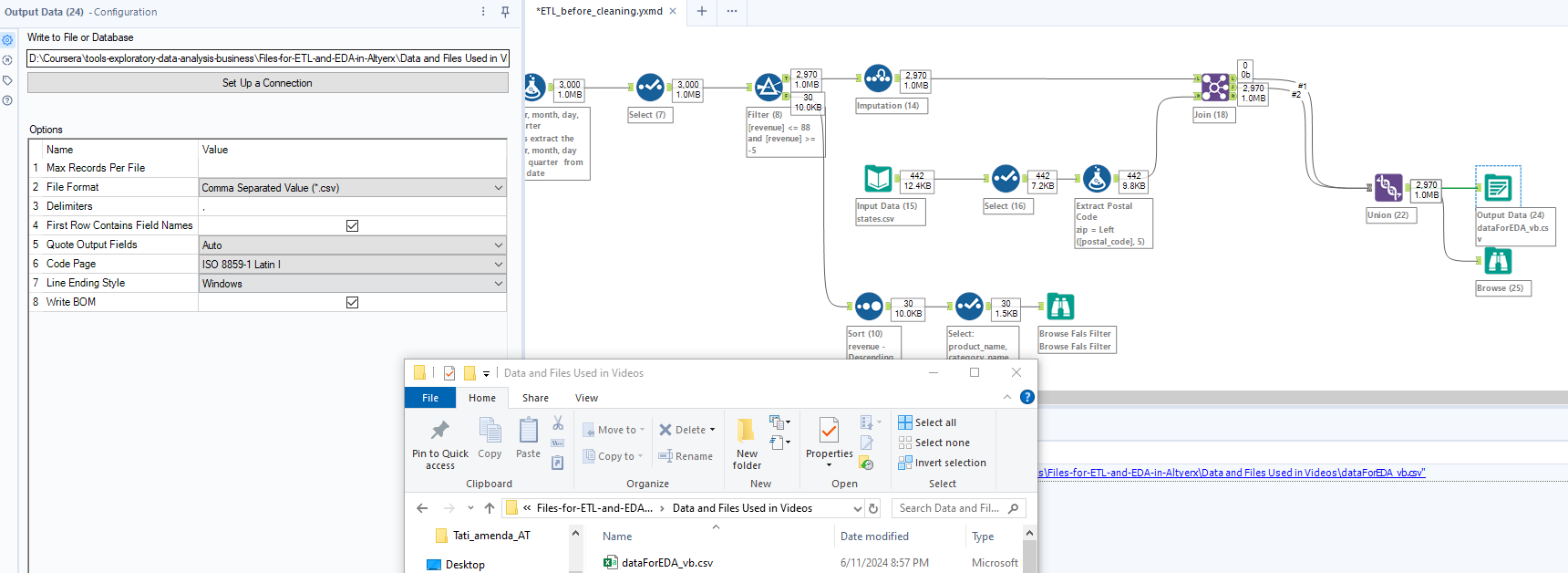
#### Picture 22. Join tool and configuration



#### Picture 23. Union Tool

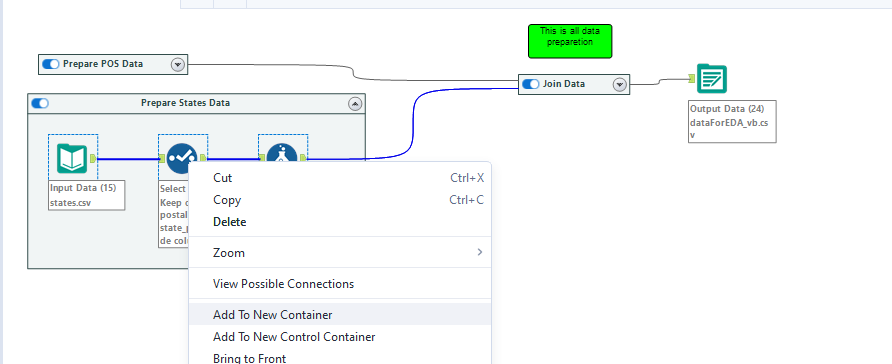
## Writing Data and Organizing the Workflow

After the data was clean, organized and prepare for export to a new file that would be a input for the EDA (Exploratory Data Analysis) process, it was insert a new step by “*Output Data*” tool. ([Picture 24. Output Data](#_Picture_24._Output)). To configurate the “*Output Data”* tool was needed only to “Set Up a Connection” that it moaned only to select the target file where would be saved.



#### Picture 24. Output Data

To transform all the complex flow in a simpler workflow it was added 2 Containers that group several steps. A Container could be added selected all the steps that you want to add insight of it, click right and select the option “*Add to new Container*” ([Picture 25. Container](#_Picture_25._Container))



#### Picture 25. Container