Changelog

Here I noted down all the changes made in the data cleaning and preparation process of the AdventureWorks Sales Budget Analytics Project.

Project name: Sales Budget Analytics
Organization: Unified Mentor
Created by: Arpita Deb
Dated: 21.05.2024 - 24.05.2024
Datasets used:

- Original database: Included in GitHub
- Lookup database: AdventureWorks sample databases

DATA CLEANING IN EXCEL:

1. Made a duplicate copy of the datasets.

2. Cleared the formats from all the sheets.

3. Clearing the formats changed the dates into string format. So I changed the date columns back to a short date format.

4. Removed the extra columns from sales sheet [columns with no headers]

5. Removed these columns -
    - From Territory sheet - RegionImage
    - From Calender sheet - MonthYear, MonthYearLong, MonthYearNum, MonthNum, WeekdayNum
    - From Customer sheet- FullName
    - From Sales sheet - PromotionKey, DateKey

6. Removed 1 row with NA Values from Territory sheet.

7. Checked for duplicates. No duplicate entry found in any of the sheets.

8. Changed data types of these columns-
    - In Sales sheet
        - DateKey - Text

- In Customer sheet -
    - CustomerKey - Text

- In Product sheet -
    - ProductKey - Text

- In Territory sheet -
    - SalesTerritoryKey - Text

- In Sales sheet -
    - CustomerKey -Text
    - PromotionKey - Text
    - SalesterritoryKey - Text

9. Checked the categorical names for Typographical Errors. Changed 'AddressLine1' header to AddressLine from Customer sheet.

10. Since there were sales data from 2014 to 2016 only, I filtered the 'Date' column from Calender sheet for 2014, 2015 and 2016.

11. There are 228 Null values in these columns from Product Sheet - SubCategory, Category, Color, StandardCost, ListPrice, ProductLine, ModelName and ProductDescription.

In order to impute the correct values I've used Microsoft SQL Server where I've already installed the Original Adventureworks database. From the [Production.Product] table from Adventureworks database, I tried to impute these values.

Here is the SQL query I used to extract the relevant information of these products with null entries in excel sheet.

```
SELECT [ProductID], [Name], [ListPrice], [StandardCost], [ProductLine],
[ProductModelID], [ProductSubcategoryID]

FROM [AdventureWorks2022].[Production].[Product]

WHERE [Name] IN ('Adjustable Race','Bearing Ball','BB Ball Bearing','Headset Ball
Bearings','Blade','LL Crankarm', 'ML Crankarm','HL Crankarm','Chainring
```

Bolts','Chainring Nut','Chainring','Crown Race','Chain Stays','Decal 1','Decal 2','Down Tube','Mountain End Caps','Road End Caps','Touring End Caps','Fork End','Freewheel','Flat Washer 1','Flat Washer 6','Flat Washer 2','Flat Washer 9','Flat Washer 4','Flat Washer 3','Flat Washer 8','Flat Washer 5','Flat Washer 7','Fork Crown','Front Derailleur Cage','Front Derailleur Linkage','Guide Pulley','LL Grip Tape','ML Grip Tape','HL Grip Tape','Thin-Jam Hex Nut 9','Thin-Jam Hex Nut 10','Thin-Jam Hex Nut 1','Thin-Jam Hex Nut 2','Thin-Jam Hex Nut 15','Thin-Jam Hex Nut 16','Thin-Jam Hex Nut 5','Thin-Jam Hex Nut 6','Thin-Jam Hex Nut 3','Thin-Jam Hex Nut 4','Thin-Jam Hex Nut 13','Thin-Jam Hex Nut 14','Thin-Jam Hex Nut 7','Thin-Jam Hex Nut 8','Thin-Jam Hex Nut 12','Thin-Jam Hex Nut 11','Hex Nut 5','Hex Nut 6','Hex Nut 16','Hex Nut 17','Hex Nut 7','Hex Nut 8','Hex Nut 9','Hex Nut 22','Hex Nut 23','Hex Nut 12','Hex Nut 13','Hex Nut 1','Hex Nut 10','Hex Nut 11','Hex Nut 2','Hex Nut 20','Hex Nut 21','Hex Nut 3','Hex Nut 14','Hex Nut 15','Hex Nut 4','Hex Nut 18','Hex Nut 19','Handlebar Tube','Head Tube','LL Hub','HL Hub','Keyed Washer','External Lock Washer 3','External Lock Washer 4','External Lock Washer 9','External Lock Washer 5','External Lock Washer 7','External Lock Washer 6','External Lock Washer 1','External Lock Washer 8','External Lock Washer 2','Internal Lock Washer 3','Internal Lock Washer 4','Internal Lock Washer 9','Internal Lock Washer 5','Internal Lock Washer 7','Internal Lock Washer 6','Internal Lock Washer 10','Internal Lock Washer 1','Internal Lock Washer 8','Internal Lock Washer 2','Thin-Jam Lock Nut 9','Thin-Jam Lock Nut 10','Thin-Jam Lock Nut 1','Thin-Jam Lock Nut 2','Thin-Jam Lock Nut 15','Thin-Jam Lock Nut 16','Thin-Jam Lock Nut 5','Thin-Jam Lock Nut 6','Thin-Jam Lock Nut 3','Thin-Jam Lock Nut 4','Thin-Jam Lock Nut 13','Thin-Jam Lock Nut 14','Thin-Jam Lock Nut 7','Thin-Jam Lock Nut 8','Thin-Jam Lock Nut 12','Thin-Jam Lock Nut 11','Lock Nut 5','Lock Nut 6','Lock Nut 16','Lock Nut 17','Lock Nut 7','Lock Nut 8','Lock Nut 9','Lock Nut 22','Lock Nut 23','Lock Nut 12','Lock Nut 13','Lock Nut 1','Lock Nut 10','Lock Nut 11','Lock Nut 2','Lock Nut 20','Lock Nut 21','Lock Nut 3','Lock Nut 14','Lock Nut 15','Lock Nut 4','Lock Nut 19','Lock Nut 18','Lock Ring','Lower Head Race','Lock Washer 4','Lock Washer 5','Lock Washer 10','Lock Washer 6','Lock Washer 13','Lock Washer 8','Lock Washer 1','Lock Washer 7','Lock Washer 12','Lock Washer 2','Lock Washer 9','Lock Washer 3','Lock Washer 11','Metal Angle','Metal Bar 1','Metal Bar 2','Metal Plate 2','Metal Plate 1','Metal Plate 3','Metal Sheet 2','Metal Sheet 3','Metal Sheet 7','Metal Sheet 4','Metal Sheet 5','Metal Sheet 6','Metal Sheet 1','Metal Tread Plate','LL Nipple','HL Nipple','Paint - Black','Paint - Red','Paint - Silver','Paint - Blue','Paint - Yellow','Pinch Bolt','Cup-Shaped Race','Cone-Shaped Race','Reflector','LL Mountain Rim','ML Mountain Rim','HL Mountain Rim','LL Road Rim','ML Road Rim','HL Road Rim','Touring Rim','LL Mountain Seat Assembly','ML Mountain Seat Assembly','HL Mountain Seat Assembly','LL Road Seat Assembly','ML Road Seat Assembly','HL Road Seat Assembly','LL Touring Seat Assembly','ML Touring Seat Assembly','HL Touring Seat Assembly','LL Spindle/Axle','HL Spindle/Axle','LL Shell','HL Shell','Spokes','Seat

Lug','Stem','Seat Post','Steerer','Seat Stays','Seat Tube','Top Tube','Tension Pulley','Rear Derailleur Cage','HL Road Frame - Black, 58','HL Road Frame - Red, 58');

When I checked for ProductLine, ModelName, Category and SubCategories from the SQL database from [production.product] table, there were null values in it which corresponded to the null values in the Products sheet . So I replaced these values with NA.

Similarly when I checked for the ListPrice and StandardCost columns for null values, I found out all their values to be 0.00 in the database. So I've imputed these values with 0 in the sheet here.

However, some products had Non-Null StandardCost and ListPrice. I've Imputed these values-

Name	StandardCost	ListPrice
HL Mountain Seat Assembly	145.87	196.92
HL Road Frame - Black, 58	1059.31	1431.50
HL Road Frame - Red, 58	1059.31	1431.50
HL Road Seat Assembly	145.87	196.92
HL Touring Seat Assembly	145.87	196.92
LL Mountain Seat Assembly	98.77	133.34
LL Road Seat Assembly	98.77	133.34
LL Touring Seat Assembly	98.77	133.34
ML Mountain Seat Assembly	108.99	147.14
ML Road Seat Assembly	108.99	147.14
ML Touring Seat Assembly	108.99	147.14

12. Concatenated the FirstName and LastName columns in the customer sheet to create a new column 'CustomerName'and removed the 2 columns.

13. Created a new column 'Hemisphere' in the Territory sheet.

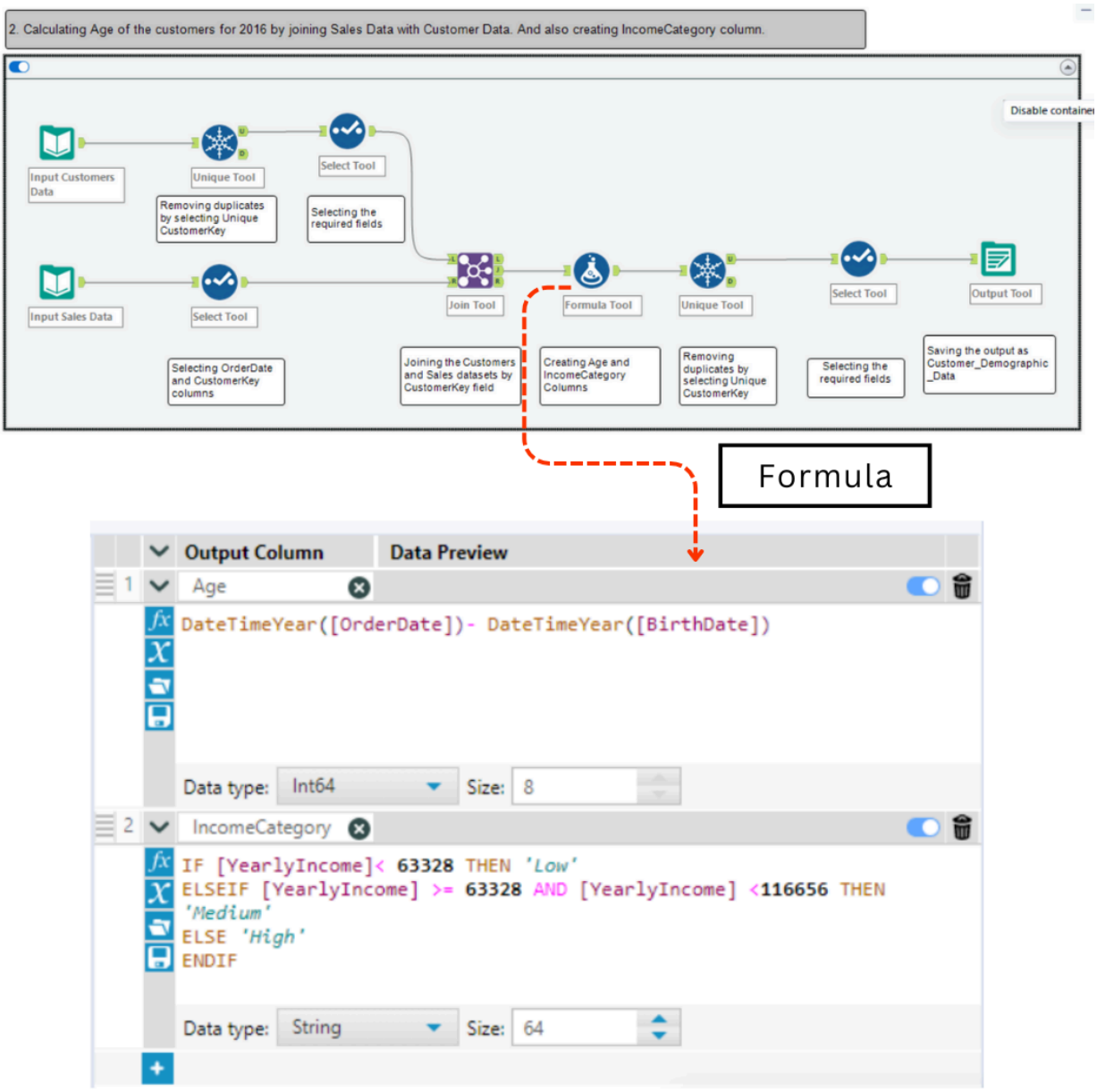
14. Created a new table 'Season' with 3 columns (Month, Hemisphere and Season) and 24 rows.

Similarly I took the following data cleaning steps for the Budget data -
1. Cleared all formattings.
2. Removed the top 3 rows along with the Subtotal Clothing, Subtotal Accessories, Subtitle Bikes, Grand total row and Grand Total column.
3. Changed the headers starting from Jan,2016 to Dec, 2016 to only Jan, Feb, etc using LEFT function.

DATA TRANSFORMATION IN ALTERYX:

1. I loaded the 6 datasets (except for Budget data) into Alteryx.

2. Next, I created a separate dataset that contains customer demographic data such as name, age, gender, income etc and named it Customer_Demographic_Data. In order to calculate age, I needed to subtract birth date from order date. So I joined Sales Data with Customer data on unique CustomerKey. I wanted to create CustomerAge and IncomeCategory columns.
   Next, using the Formula tool I created 2 new columns Age and IncomeCategory. To create the IncomeCategory column from the YearlyIncome column I did the following calculations:
   Range of income = Max Income - Min Income = 170000 -10000 = $160000
   Number of categories = 3 (High, Medium, Low)
   Therefore, I need to divide $160000 into 3 equal parts, each being 33.33% of the total.
   Low income < 10000 + 33.33% of 160000 < $63328
   Mid Income >= $63328 and < 100000 + 66.66 % of 1600000 = $116656
   High Income >= $116656



3. Next, I created 2 new columns CustomerLifetimeDuration and CustomerType based on Customer Lifetime Duration (High Value, Medium Value and Low Value) and saved the data as Customer_Order_Details.

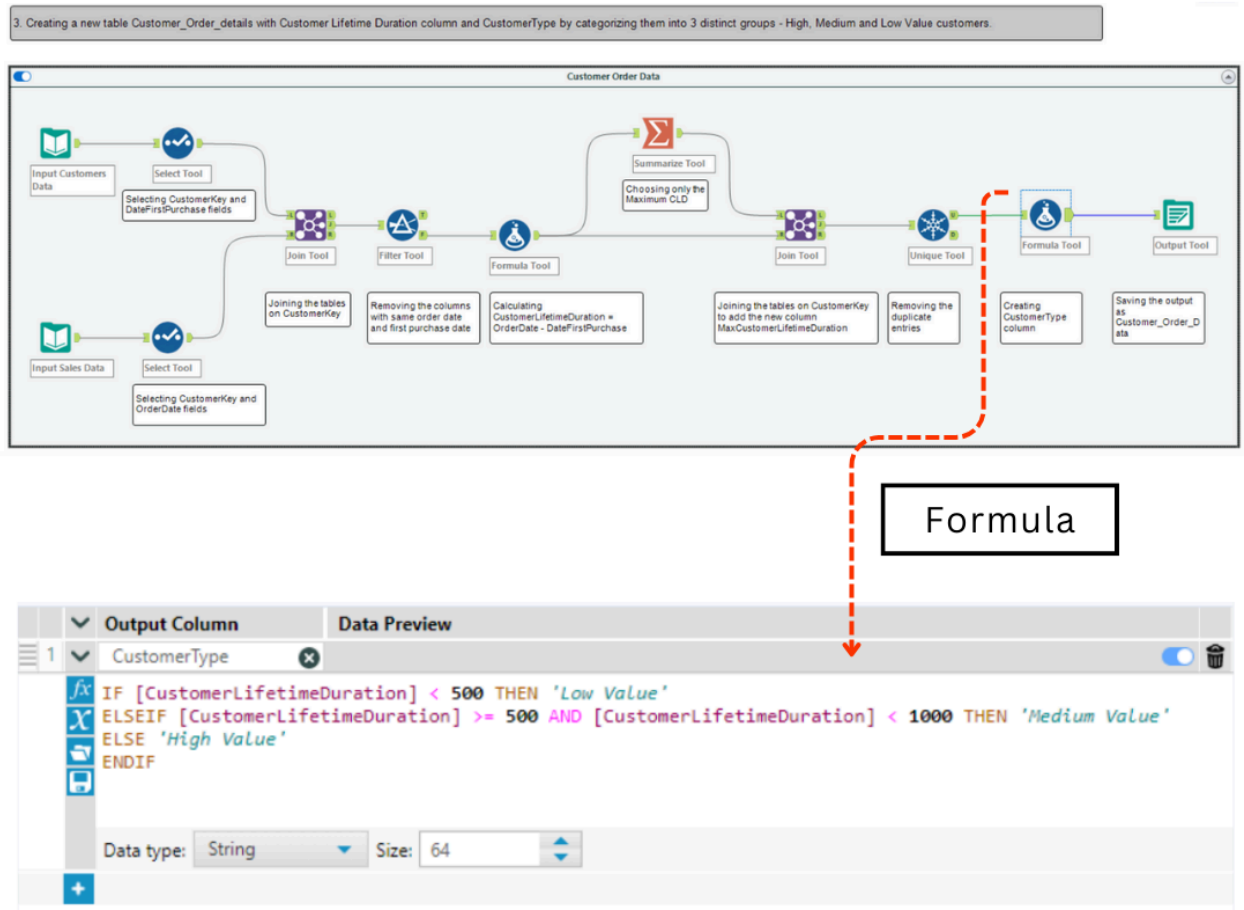
   The formula for CustomerLifetimeDuration = OrderDate - DateFirstPurchase. So I joined the Customer data with Sales data on unique CustomerKey. I filtered out the rows with the same OrderDate and DateFirstPurchase because they'll lead to 0 Customer Lifetime Duration which are not helpful in our analysis.

   After filtering the data, I calculated the Customer Lifetime Duration. Now, there are many instances where one customer has placed several orders between 2014 and 2016. This led to multiple non-zero Customer Lifetime Duration for a unique customer. Therefore,

using the Summarize tool, I selected the Maximum Customer Lifetime Duration for each customer as it will account for the most recent purchase.

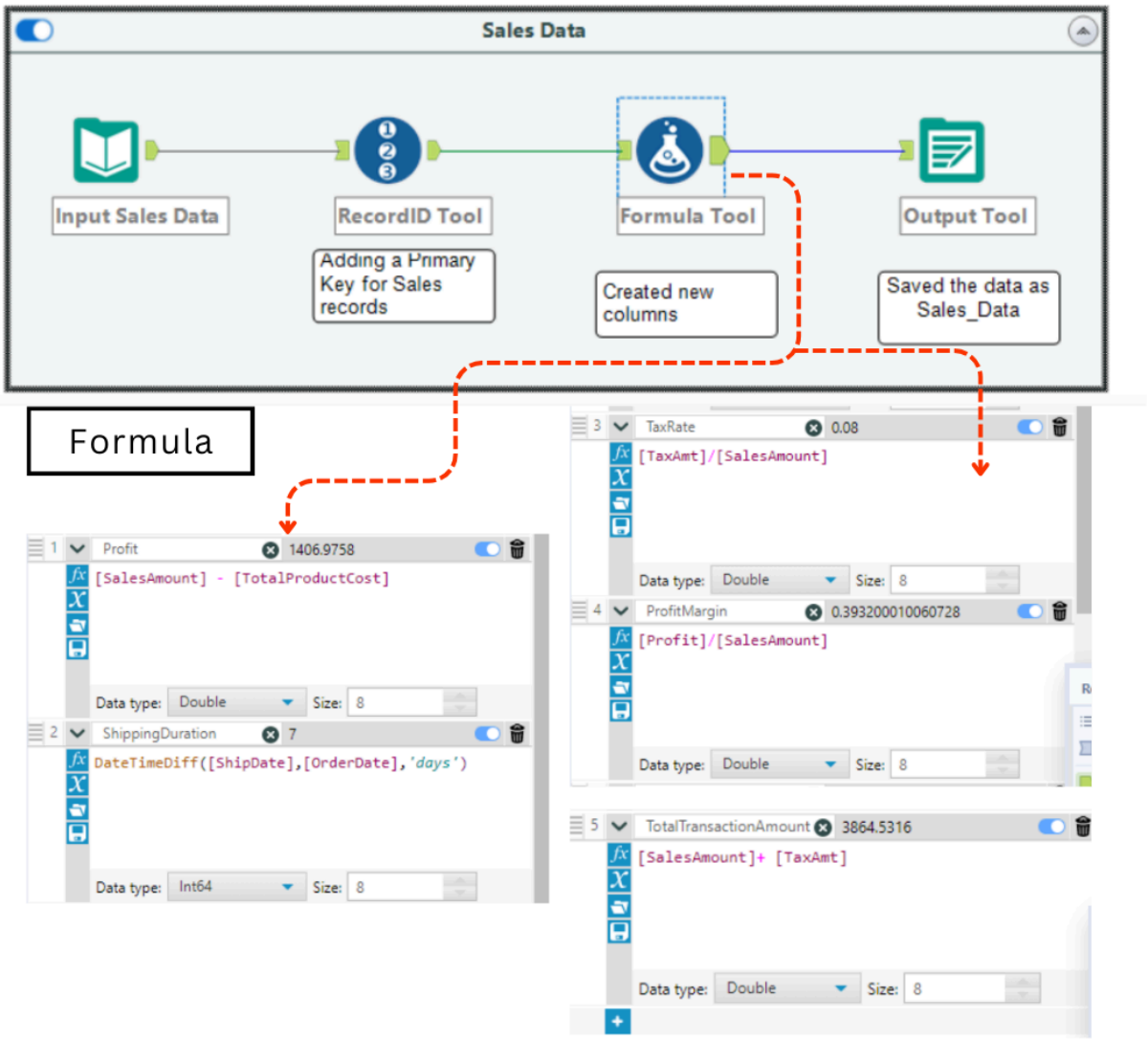
Then I joined these columns with the previously filtered data using the Join tool. This created some duplicate entries, so I removed them using Unique tools. This gave us about 6600 unique customers with non zero CustomerLifetimeDuration.

Lastly, I categorized these customers into 3 types (High value, Medium Value and Low Value) based on their Customer Lifetime Value by creating a new column CustomerType.
Customer Lifetime Value < 500 days >> ‘Low Value’
Customer Lifetime Value >= 500 days and <1000 >> ‘Medium Value’
Customer Lifetime Value >= 1000 days >> ‘Low Value’



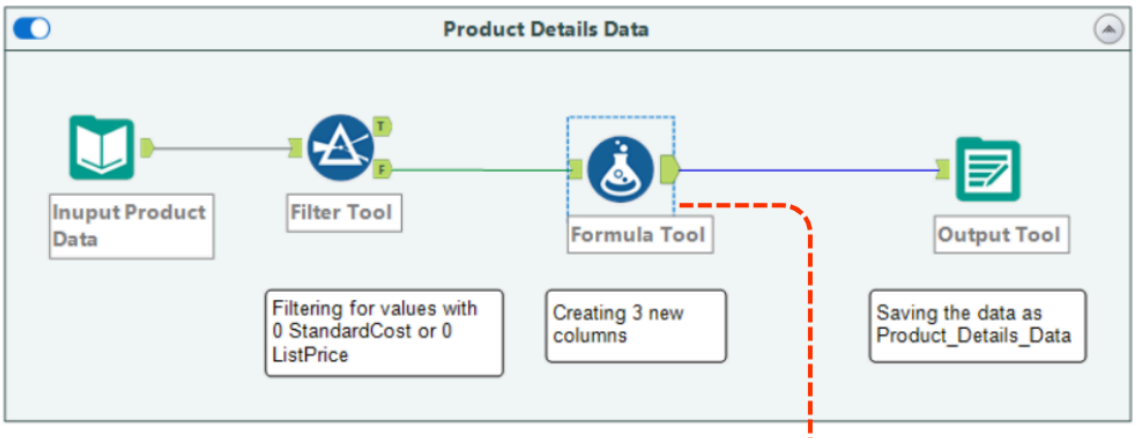
4. For the Sales data, I added a unique SalesKey column using the RecordID tool. Then using the Formula tool, I added 5 new numerical columns Profit, TaxRate, ShippingDuration, TotalTransactionCost and ProfitMargin.

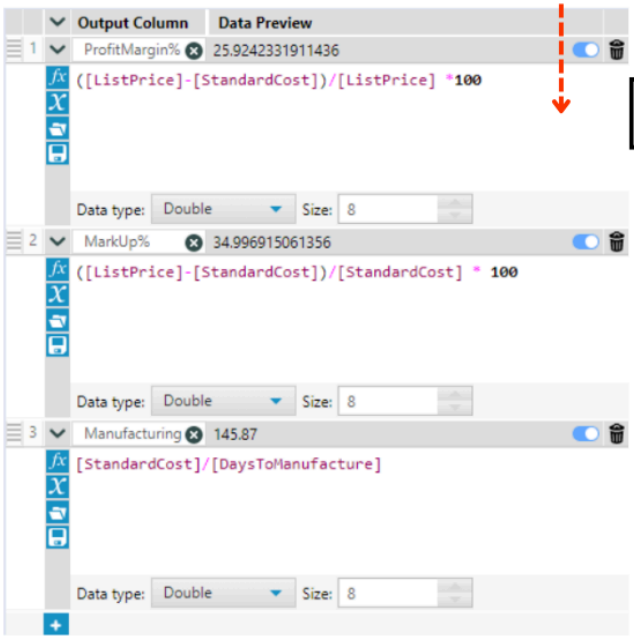
4. Creating 5 new columns - Profit, TaxRate, ShippingDuration, TotalTransactionCost, ProfitMargin in the sales sheet.



5. Similarly for the Products data, using the Formula tool, I added 3 new numerical columns ProfitMargin%, Markup% and ManufacturingEfficiency. I saved the data as Product_Details.

5. Creating 3 new columns - ProfitMargin%, Markup% and ManufacturingEfficiency in the Product sheet.





Formula