Changelog

Here I noted down all the changes made in the data cleaning and preparation process of the Financial Analysis Project.

Project name: Financial Analytics
Organization: Unified Mentor
Dataset used:

- Top 500 Companies in India
- ET Top 500 Indian Companies (2009-2021)
- Nifty 500 fundamental statistics

Dated: 12.05.24 - 14.05.24

Data Cleaning in Excel:

1. Before starting cleaning the dataset, I saved a copy of the original **Top 500 Companies in India** dataset. And I also downloaded 2 other datasets which I used for joining columns in later steps.
2. **Fixing Formatting Error**:
   Some records from the quarterly_sales column were not aligned with the rest of the values, so I manually copied and pasted them in their designated place.

S.No.	Name	Mar Cap - Crore	Sales Qtr - Crore	
57	Britannia Inds.	56837.2	2567.48	
58	Tech Mahindra	56244.26	7775.96	
59	Hindalco Inds.	55854.68	11022.81	
60	Zee Entertainmen	54817.89	1838.07	
61	Cairn India	53528.57		2149.36
62	Indiabulls Hous.	52781.67		3115.89
63	Ambuja Cem.	52361.46		6170.71
64	Interglobe Aviat	48621.37		6177.88
65	Cipla	48577.43		3913.82
66	Piramal Enterp.	47483.97		2858.36
67	United Spirits	46725.05		2263.3
68	Pidilite Inds.	45855.5		1542.9
69	Siemens	44239.04		2429.5
70	Cadila Health.	41876.19		3259.6
71	NMDC	41415.33		2469.03
72	DLF	40159.35		1693.72
73	Marico	39813.84		1337.59
74	Ashok Leyland	39047.57		7113.16
75	Bharat Electron	37776.23		2512.82
76	ICICI Lombard	37219.22		2110.99
77	Lupin	36878.85		3975.62
78	Petronet LNG	36615		7757.06
79	Aditya Birla Cap	36215.92		3325.02
80	Dr Reddy's Labs	35893.55		3834.1
81	Sun TV Network	35824.26		683.28
82	S A L L	35729.04	15323.65	

3. **Fixing Typographical Errors:** Mapped the company names in original dataset from “Economic times Top 500 companies (2009-2021)” dataset filtered for 2018. Here I replaced the incomplete or incomprehensible Company names with more accurate names.
4. **Checking Null values:**
   The total number of records = 488
   Nulls in Market Cap column = 9
   Nulls in Sales Qtr column = 29

5. **Imputing the Quarterly sales values**: Imputed the null values for these companies by taking average of all the 4 quarterly net sales/income from operations from www.moneycontrol.com.

Company Name	**Average Quarterly Sales in 2018 in Crore INR**
Amber Enterprises Limited	₹ 477.2825
Bharti Infratel Limited	₹ 1,719.10
Bombay Burmah Trading Corporation Limited	₹ 56.20
BSE Limited	₹ 101.79
Colgate Palmolive (India) Limited	₹ 1,100.09
Endurance Technologies Limited	₹ 1,337.28
Force Motors	₹ 910.29
Gayatri Projects Limited	₹ 811.31
GE Power India Limited	₹ 481.14
Hindustan Construction Co. Limited	₹ 1,115.32
HMT Limited	₹ 5.29
Info Edge (India) Limited	₹ 261.57
ISGEC Heavy Engineering	₹ 897.83
Jindal Saw Limited	₹ 2,353.01
Jaiprakash Associates	₹ 1,679.67
Jaiprakash Power Ventures Limited	₹ 915.66
LT Foods	₹ 558.20
Linde India Limited	₹ 547.91

Mahanagar Gas	₹ 663.88
Mahindra CIE Automotive Limited	₹ 680.71
Max Financial Services Limited	₹ 81.23
MMTC Limited	₹ 6,243.18
National Aluminum Co. Limited	₹ 2,899.07
Prism Cement	₹ 1,439.82
Shoppers Stop Limited	₹ 884.97
Swan Energy	₹ 218.52

There are 4 companies whose quarterly sales data was unavailable. So I replaced the NaN values with 0 for computational ease.

Company Name	**Average Quarterly Sales in 2018 in Crore INR**
Standard Chartered PLC	0
Sundaram Clayton	0
Bajaj Corp Limited	0
Ujjivan Financial Services Limited	0

As for the null values in market capitalization for these companies, I used the ChatGPT to give me an approximate result in USD. In 2018, the USD to INR currency rate fluctuated between 63 - 71 INR per USD. Taking a mean value of 68 INR per USD, I did the following conversions.

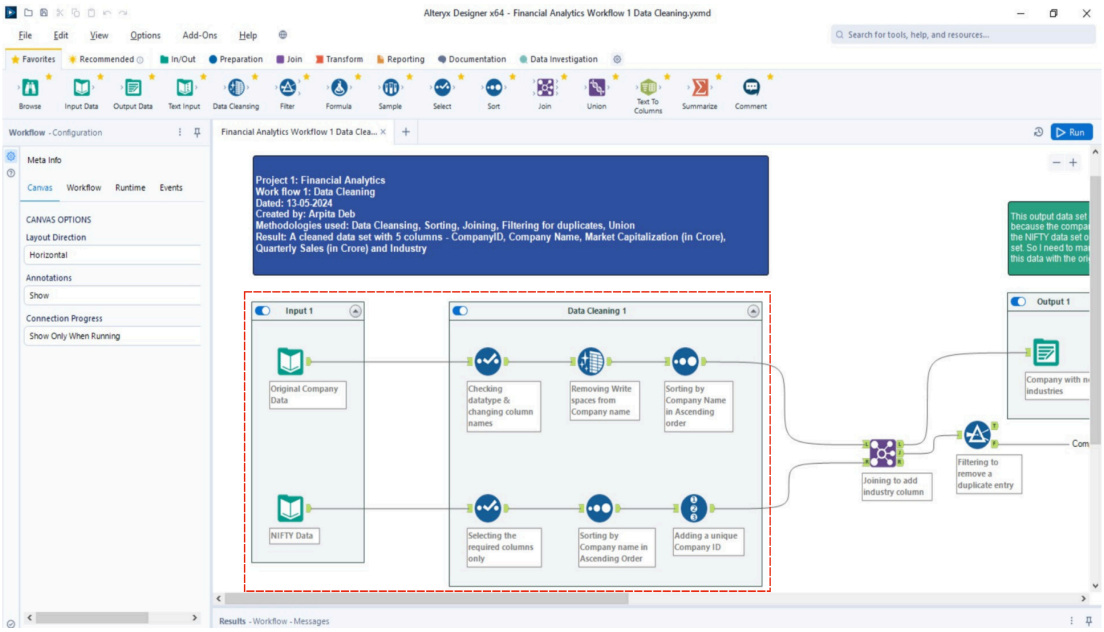
Company Name	**Market Capitalization in 2018 in Crore INR**

Bajaj Corp Limited	$14.75 B = ₹100300 cr
BSE Limited	$1.3B = ₹8840 Cr
Colgate Palmolive (India) Limited	$64B = ₹435200 Cr
Endurance Technologies Limited	$1.7B = ₹11560 Cr
Force Motors	$585M = ₹3978 Cr
ISGEC Heavy Engineering	$240M = ₹1632 Cr
L T Foods	$570M = ₹3876 Cr
Mahanagar Gas	$2.4B = ₹16320 Cr
National Aluminum Co. Limited	$2.2B = ₹14960 Cr

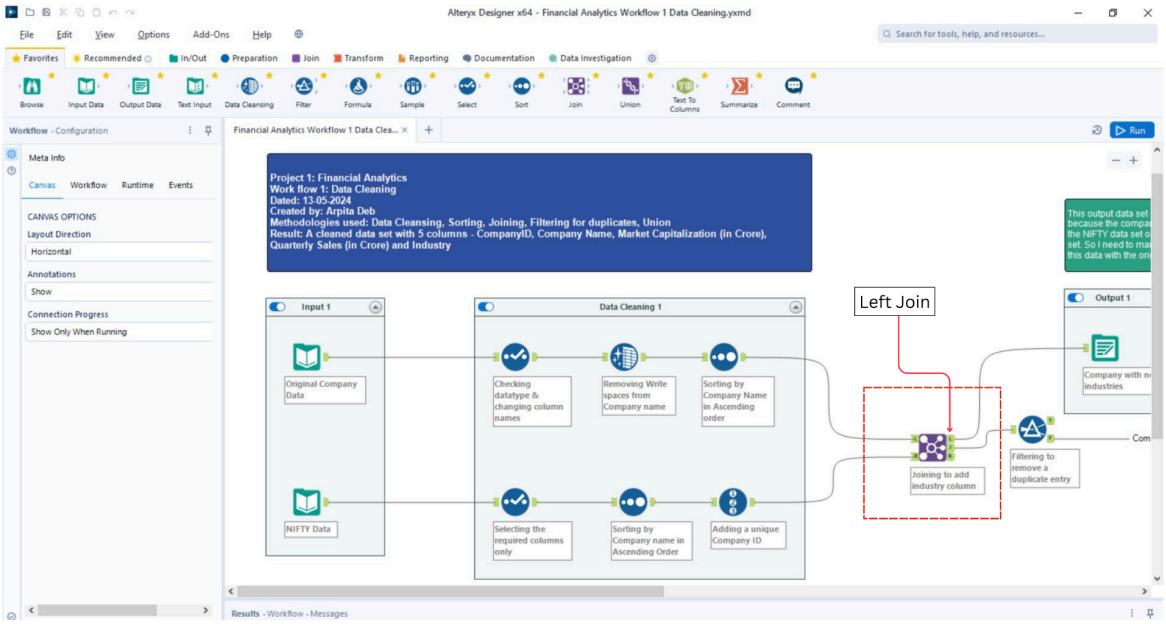
6. **Fixing the number formats:** Before saving the file as .xlsx, I removed all the formats from the sheet and changed the format of the numerical columns from general to numeric.

Data Cleaning in Alteryx:

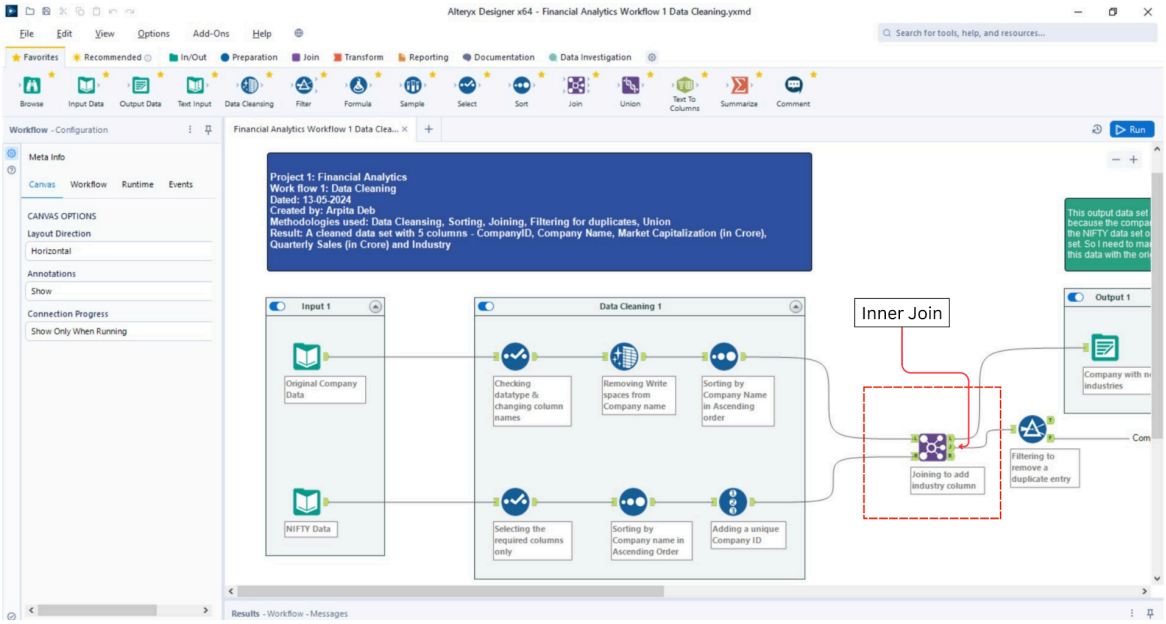
1. First connected the finance data with an input tool and selected the 4 columns. I checked for the data types and changed the column names with more comprehensive names. The reason I removed the index column is because it contains inconsistent indices.
2. Using a data Cleansing tool I removed white spaces from the company name. Then sorted the data by company name in alphabetical order.
3. I did the same for the NIFTY dataset. I first selected only 2 columns - Company name and Industry, sorted the results in alphabetical order and added a unique id for each of the records.



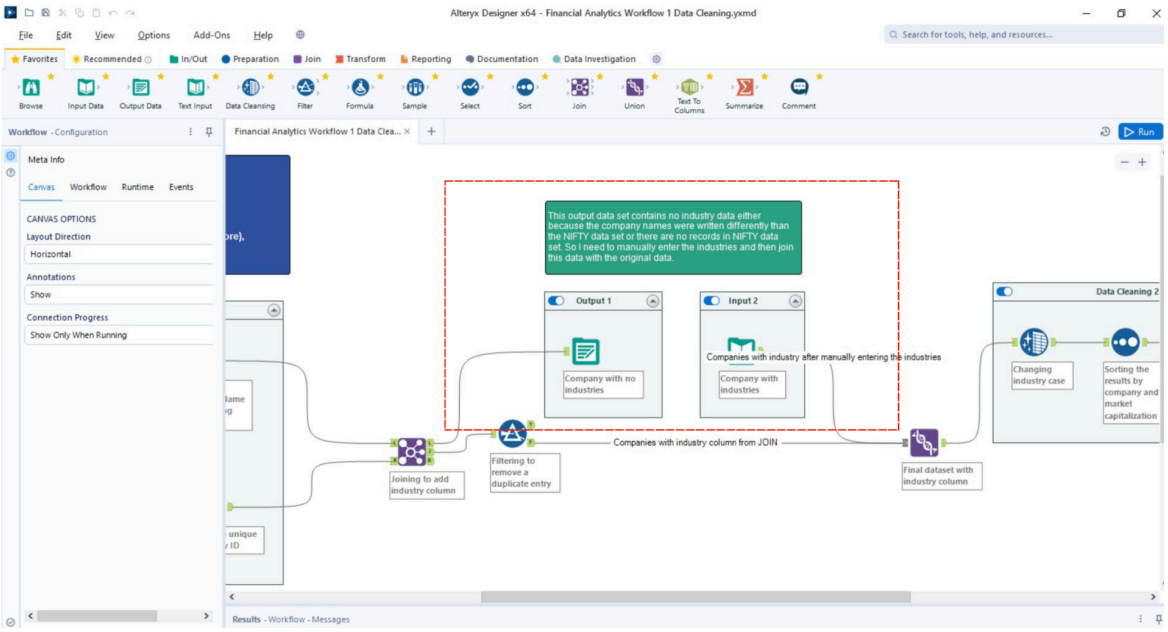
4. I joined the two datasets on the Company Name column and selected the 5 columns - companyID, Company Name, Market Capitalization (in Crore), Quarterly Sales (in Crore) and Industry. I wanted to identify the industry for each company.
5. Now the Left (L) output of the Join Tool shows only 166 companies that are present in our original dataset with no corresponding industry. This is either because the companies' names do not match with the names in the NIFTY dataset or they don't exist in the NIFTY dataset.



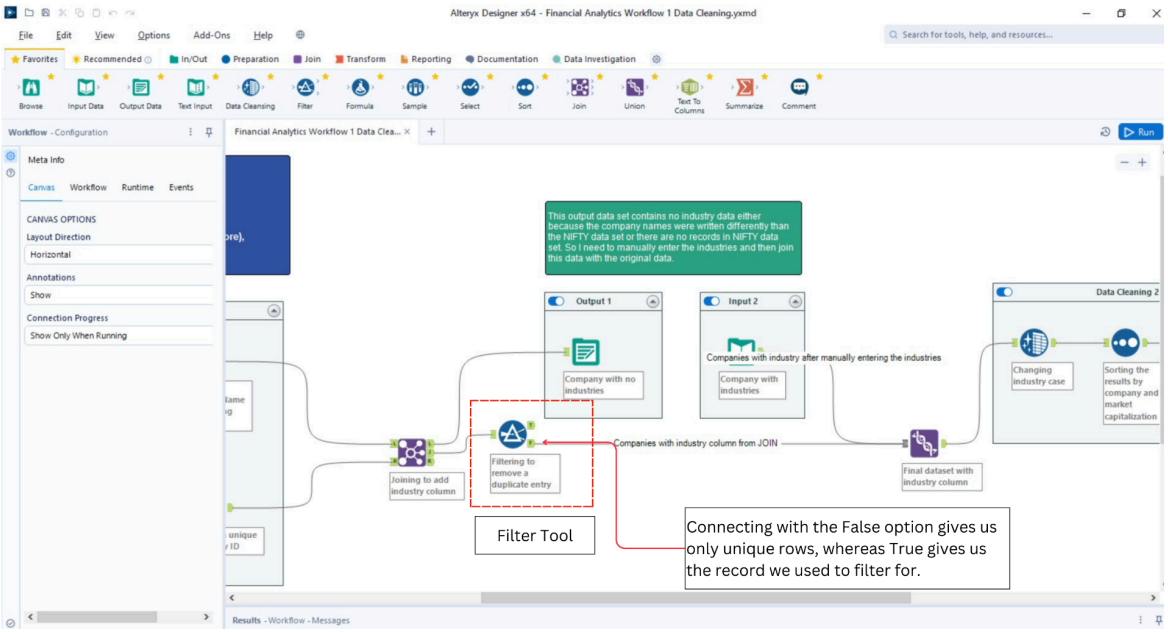
6. The Join (J) output shows 322 companies that are present in both the datasets.



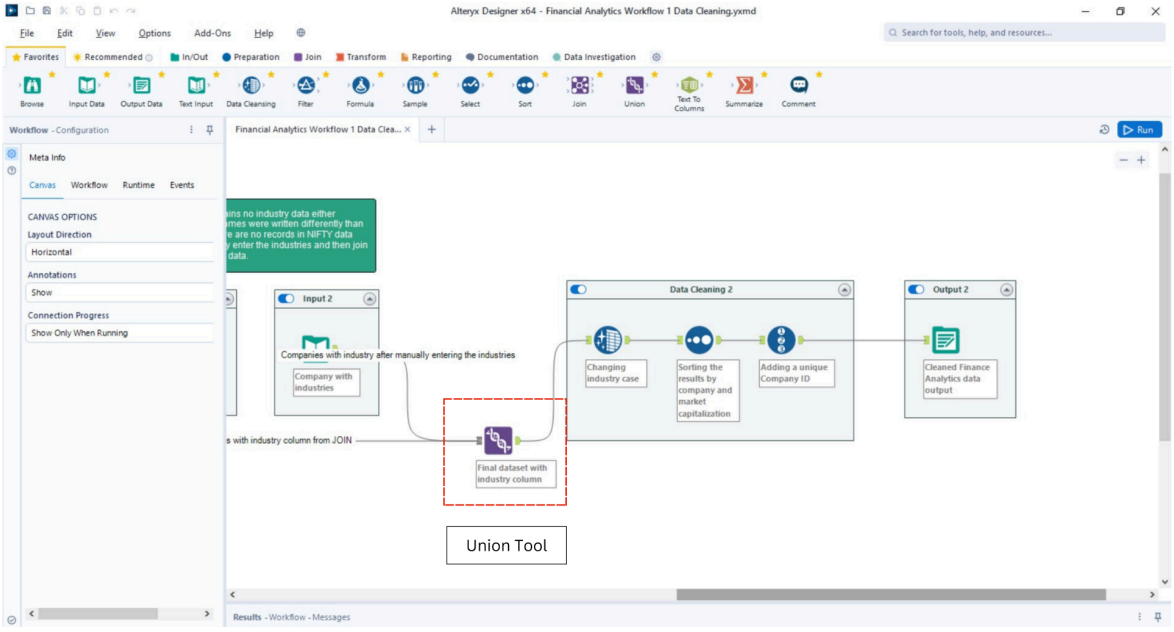
7. In order to update the industry names for the 166 companies, I used an Output tool to select only those data and saved it as an excel file. Then I manually updated the industries from information collected from the web and the NIFTY dataset. I added the updated table with an input tool.



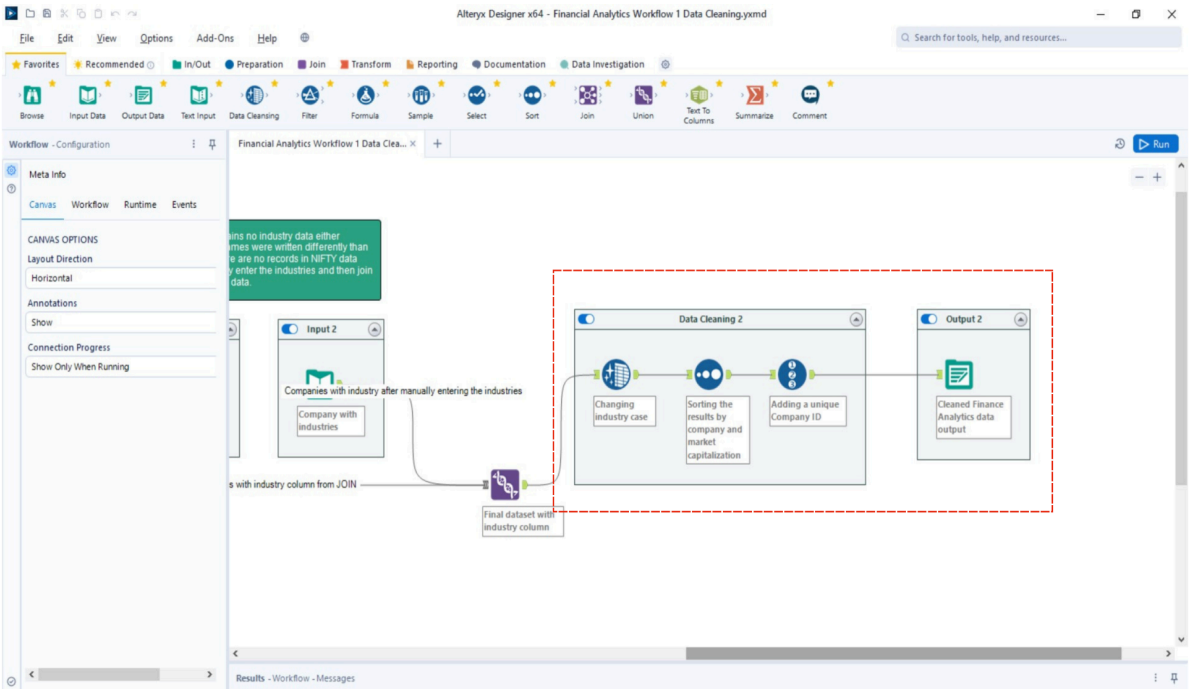
8. Before combining all the data from Left and Inner Join tables, I used a filter tool to remove a duplicate entry. There were 2 entries for HDFC Bank with 2 different market capitalization values. So I set a filter for Market Capitalization = 289497.37, connected the False option with the union tool because it contains all the unique information. It contained 321 rows.



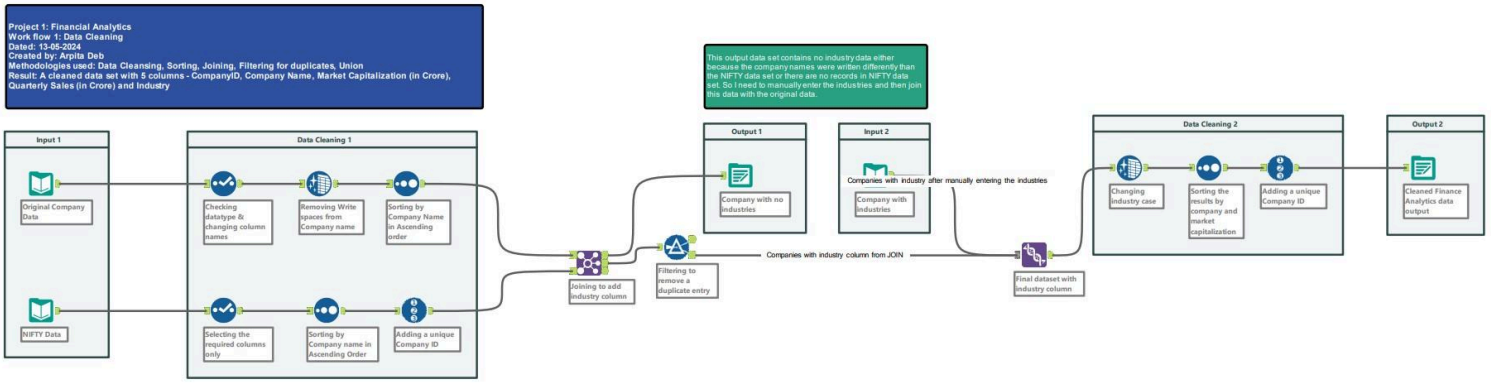
9. With the Union tool I joined these 321 rows from the Join Tool and 166 manually updated rows. This gave us 487 unique rows.



10. Finally changed the cases of the industry records to Title case and sorted the data in alphabetical order. I added a unique ID column named CompanyID that will uniquely identify each company.



11. Finally using containers from the Documentation tool I organized the workflow into different containers based on their action. And using an Output tool I collected **the cleaned dataset with 5 columns and 487 rows** into an Excel file which I've used for data transformation. This is how the entire workflow looks like:

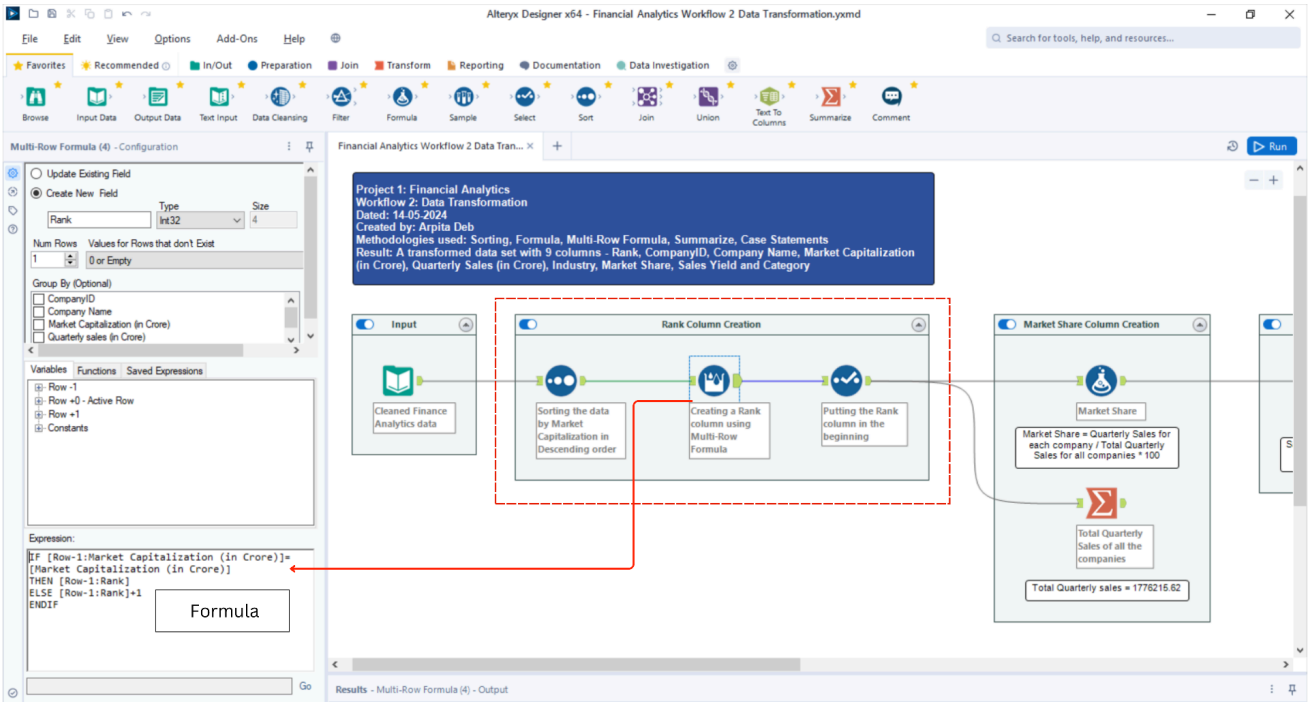


Data Transformation in Alteryx:

In this phase I've created new columns based on existing columns. I've added Market Share, Market Cap Classification (Small, Mid and Large), Rank and Sales Yield columns based on capitalization.

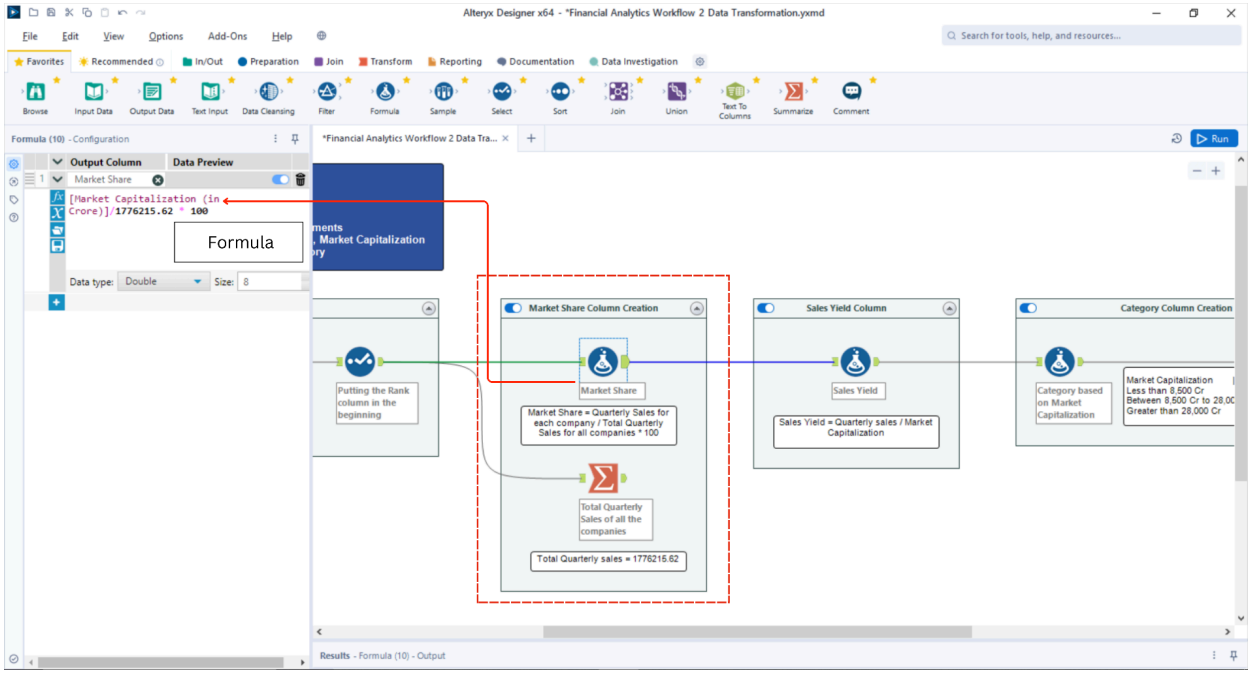
1. In the first step I loaded the cleaned dataset into Alteryx using the Input tool.
2. Then based on Market Capitalization, I wanted to rank the companies. To create the rank column I first sorted the data in descending order of market capitalization. Then using a Multi-Row formula tool, I created a new column named Rank and placed it in the beginning. The formula I used is given as -

   IF [Row-1:Market Capitalization (in Crore)]= [Market Capitalization (in Crore)]
   THEN [Row-1:Rank]
   ELSE [Row-1:Rank]+1
   ENDIF

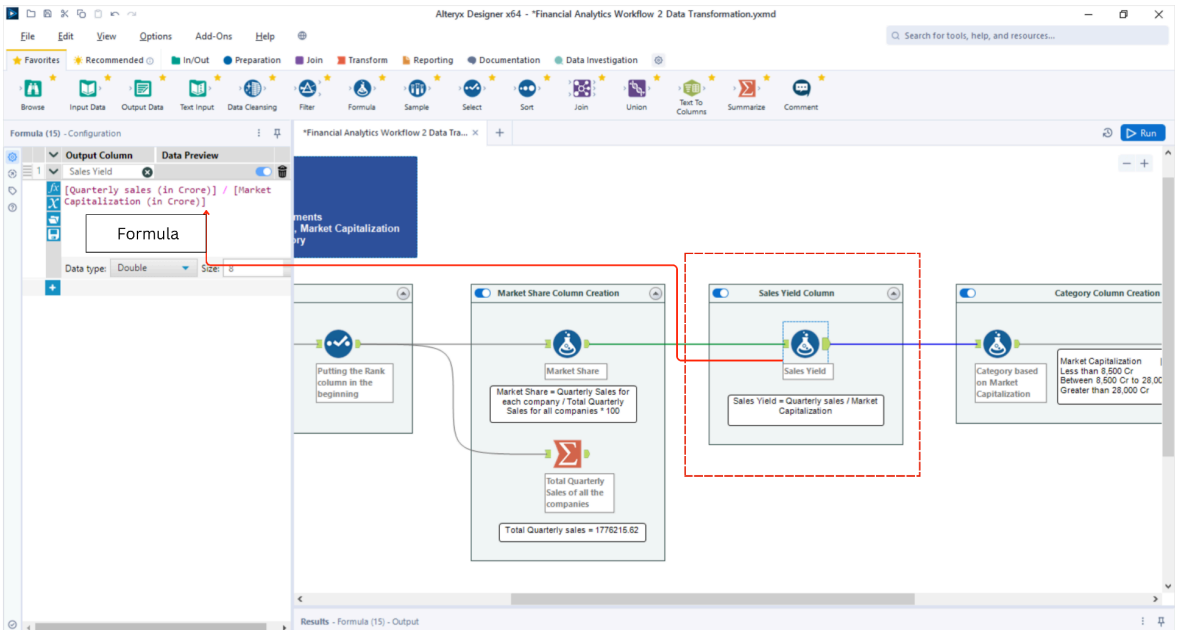


3. Then I created a Market Share Column. The formula for
   Market Share = Quarterly Sales for each company / Total Quarterly Sales for all companies * 100

   To do that I first summarized the quarterly sales using the Summarize tool. Then I created a Market Share column using Formula tool where I divided the quarterly sales for each company with 1776215.62 (which is the total quarterly sales for all the companies)



4. Another column named Sales Yield is created by using the Formula tool where I divided Quarterly sales by Market capitalization for each company.



5. Finally I created a Category column where I categorized the company into 3 groups - Small Cap, Mid Cap and Large Cap based on their market capitalization. I used the following case statement in the formula tool -

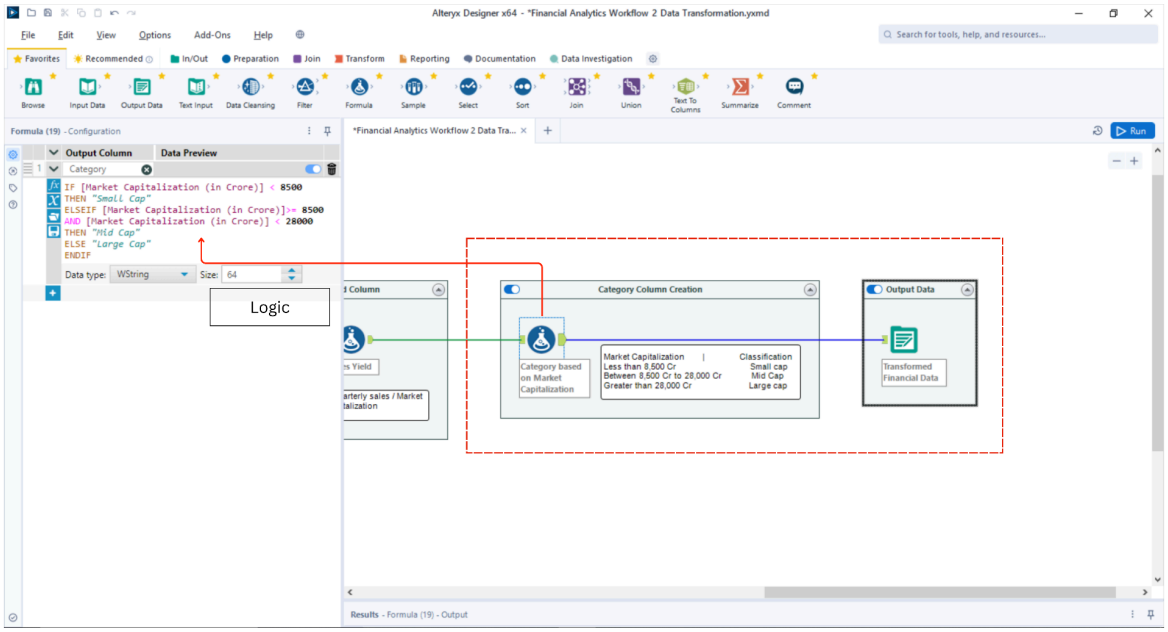
   IF [Market Capitalization (in Crore)] < 8500

   THEN "Small Cap"

ELSEIF [Market Capitalization (in Crore)]>= 8500 AND [Market Capitalization (in Crore)] < 28000 THEN "Mid Cap"

ELSE "Large Cap"

ENDIF



6. Lastly, I collected the transformed data as an Excel file using an Output tool. This is how the entire workflow looks like.