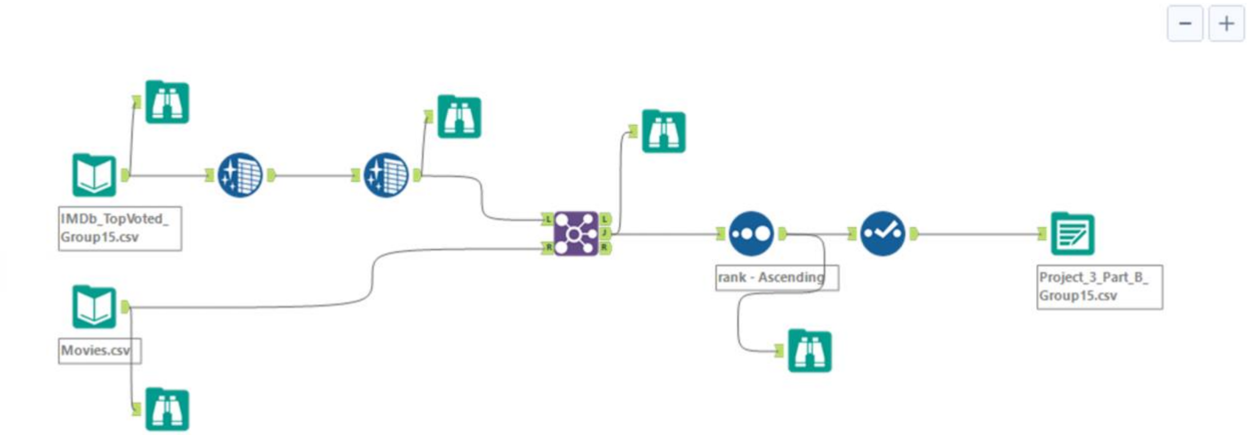
Project 3

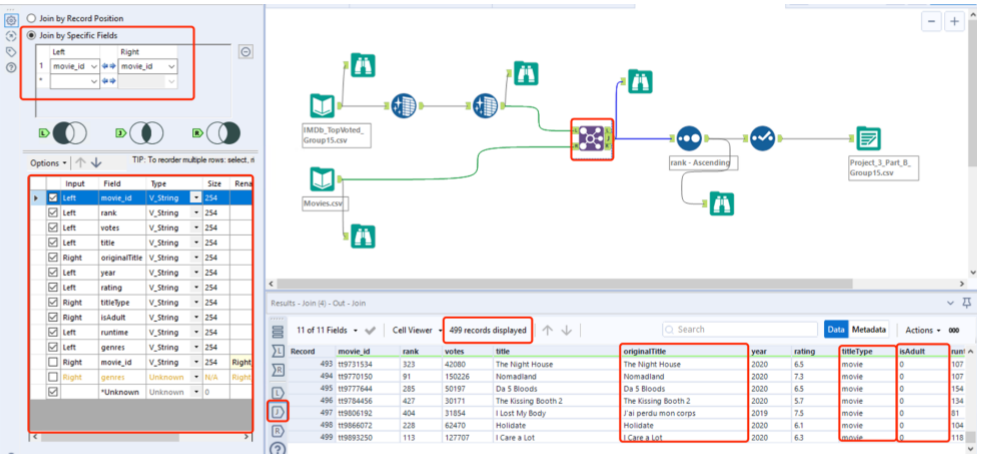
Group15 (Haozhe Liu, Shiming Zhao, Jinsheng Luo)

MISM 6213



1. What data was used to enrich the client's data?

On the basis of the IMDb_TopVoted_Group15.csv dataset we created in Part A, we have enriched it with other columns ('titleType', 'originalTitle', and 'isAdult') from Movies.csv by the inner join function. The join by specific field is 'Movie_id'. (IMDb_TopVoted_Group15.csv as the left side, Movie.csv as the right side) By checking out the output of the joined dataset, we find there are 499 rows compared to the original 500 rows, and 11 columns compared to the original 8 columns. The only one row was removed because the left table's Movie_id is not same as the right table.



○ Join by Record Position
● Join by Specific Fields

	Left	Right
1	movie_id	movie_id
*		

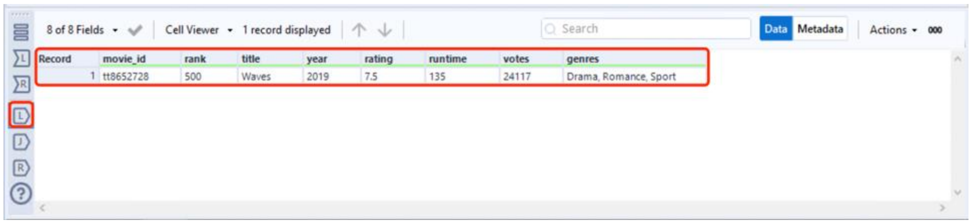
TIP: To reorder multiple rows: select, ri

	Input	Field	Type	Size	Rena
☑	Left	movie_id	V_String	254	
☑	Left	rank	V_String	254	
☑	Left	votes	V_String	254	
☑	Left	title	V_String	254	
☑	Right	originalTitle	V_String	254	
☑	Left	year	V_String	254	
☑	Left	rating	V_String	254	
☑	Right	titleType	V_String	254	
☑	Right	isAdult	V_String	254	
☑	Left	runtime	V_String	254	
☑	Left	genres	V_String	254	
☐	Right	movie_id	V_String	254	Right
☐	Right	genres	Unknown	N/A	Right
☑		*Unknown	Unknown	0	

Results - Join (4) - Out - Join

11 of 11 Fields · Cell Viewer · 499 records displayed

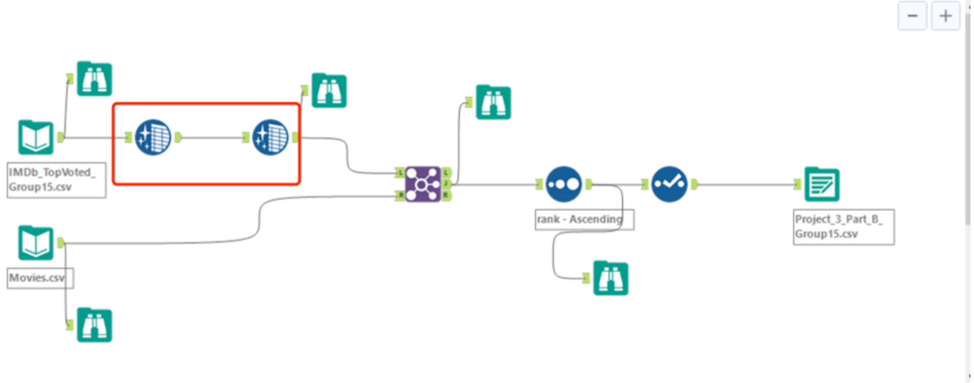
Record	movie_id	rank	votes	title	originalTitle	year	rating	titleType	isAdult	runt
493	tt9731534	323	42080	The Night House	The Night House	2020	6.5	movie	0	107
494	tt9770150	91	150226	Nomadland	Nomadland	2020	7.3	movie	0	107
495	tt9777644	285	50197	Da 5 Bloods	Da 5 Bloods	2020	6.5	movie	0	154
496	tt9784456	427	30171	The Kissing Booth 2	The Kissing Booth 2	2020	5.7	movie	0	134
497	tt9806192	404	31854	I Lost My Body	J'ai perdu mon corps	2019	7.5	movie	0	81
498	tt9866072	228	62470	Holidate	Holidate	2020	6.1	movie	0	104
499	tt9893250	113	127707	I Care a Lot	I Care a Lot	2020	6.3	movie	0	118



8 of 8 Fields · Cell Viewer · 1 record displayed · Search · Data · Metadata · Actions

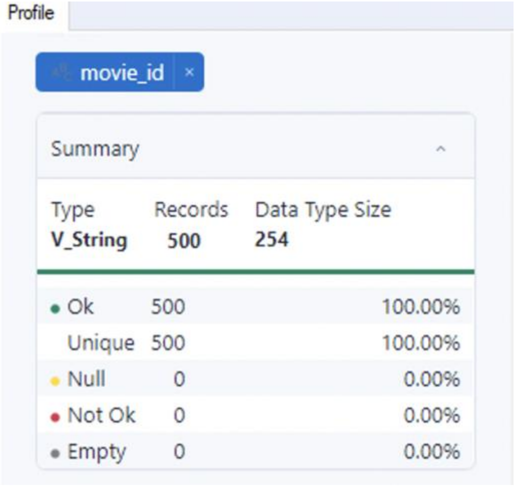
Record	movie_id	rank	title	year	rating	runtime	votes	genres
1	tt8652728	500	Waves	2019	7.5	135	24117	Drama, Romance, Sport

2. **Describe the data cleaning and transformation that was implemented.**



First: Find duplicated rows

By using the browse function of the input data, we checked the unique values of each column and found none of the columns have duplicated rows. The snap screen below is one example of checking the unique values of movie_id.



Profile

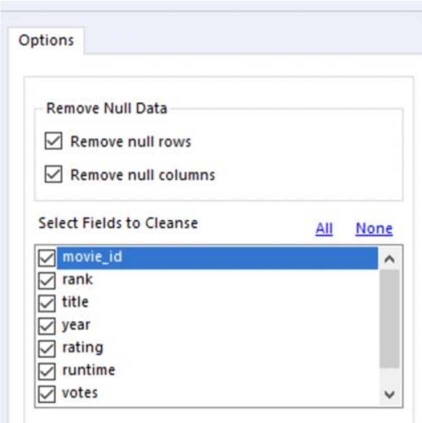
movie_id ×

Summary

Type	Records	Data Type Size
V_String	500	254

● Ok	500	100.00%
Unique	500	100.00%
● Null	0	0.00%
● Not Ok	0	0.00%
● Empty	0	0.00%

Second: Find and replace the null value



Options

Remove Null Data

☑ Remove null rows

☑ Remove null columns

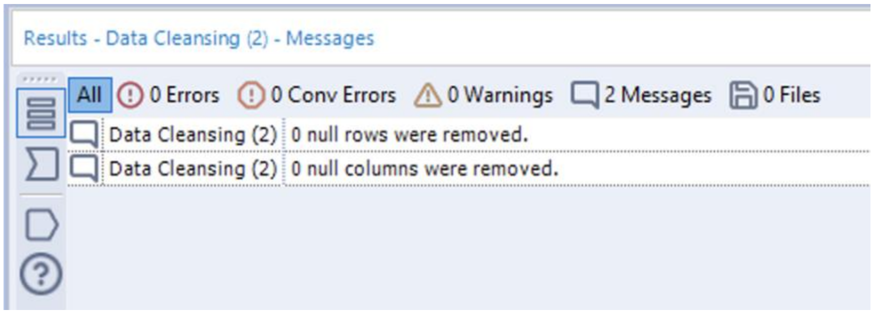
Select Fields to Cleanse

☑ movie_id
☑ rank
☑ title
☑ year
☑ rating
☑ runtime
☑ votes

By using the data cleansing function, we selected these two options:

1. Remove null rows
2. Remove null columns

We found no rows and columns were removed, which means this data doesn't have any null rows and columns.

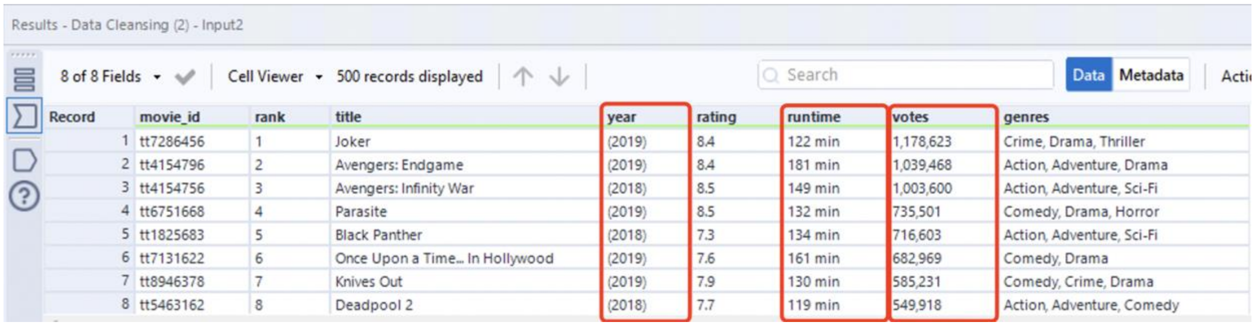


Results - Data Cleansing (2) - Messages

All | 0 Errors | 0 Conv Errors | 0 Warnings | 2 Messages | 0 Files

Data Cleansing (2)	0 null rows were removed.
Data Cleansing (2)	0 null columns were removed.

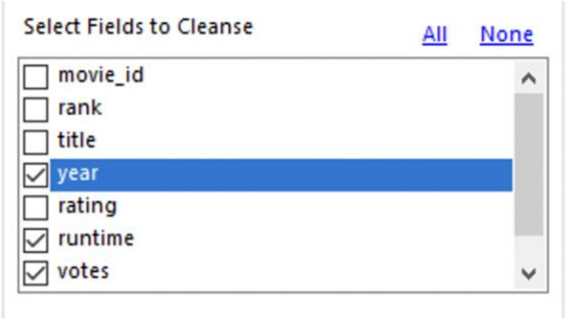
Third: correct data format



Results - Data Cleansing (2) - Input2

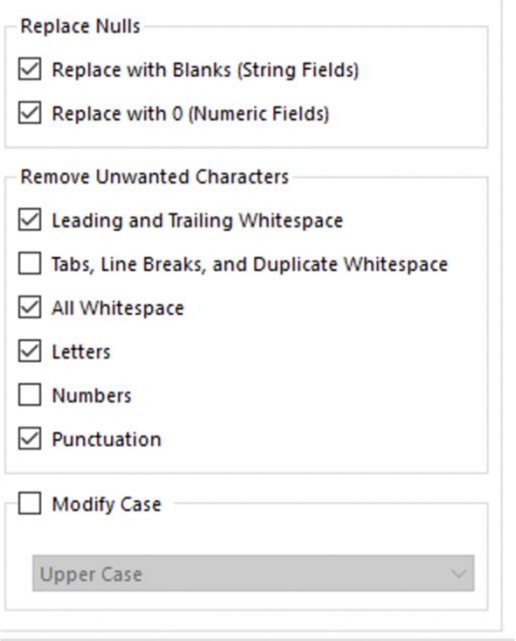
Record	movie_id	rank	title	year	rating	runtime	votes	genres
1	tt7286456	1	Joker	(2019)	8.4	122 min	1,178,623	Crime, Drama, Thriller
2	tt4154796	2	Avengers: Endgame	(2019)	8.4	181 min	1,039,468	Action, Adventure, Drama
3	tt4154756	3	Avengers: Infinity War	(2018)	8.5	149 min	1,003,600	Action, Adventure, Sci-Fi
4	tt6751668	4	Parasite	(2019)	8.5	132 min	735,501	Comedy, Drama, Horror
5	tt1825683	5	Black Panther	(2018)	7.3	134 min	716,603	Action, Adventure, Sci-Fi
6	tt7131622	6	Once Upon a Time... In Hollywood	(2019)	7.6	161 min	682,969	Comedy, Drama
7	tt8946378	7	Knives Out	(2019)	7.9	130 min	585,231	Comedy, Crime, Drama
8	tt5463162	8	Deadpool 2	(2018)	7.7	119 min	549,918	Action, Adventure, Comedy

we selected ‘year’, ‘runtime’, and ‘votes’ because we think that the brackets of the ‘year’ column, the letter 'min' of ‘runtime’ column, the comma of votes are unnecessary, and these unwanted characters: Leading and trailing Whitespace, All Whitespace, Letters, and Punctuation.



Select Fields to Cleanse

- ☐ movie_id
- ☐ rank
- ☐ title
- ☑ year
- ☐ rating
- ☑ runtime
- ☑ votes



Replace Nulls

- ☑ Replace with Blanks (String Fields)
- ☑ Replace with 0 (Numeric Fields)

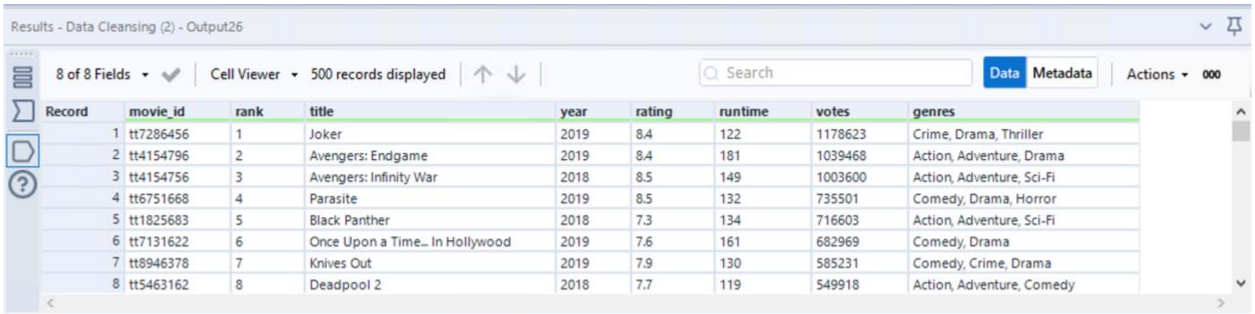
Remove Unwanted Characters

- ☑ Leading and Trailing Whitespace
- ☐ Tabs, Line Breaks, and Duplicate Whitespace
- ☑ All Whitespace
- ☑ Letters
- ☐ Numbers
- ☑ Punctuation

☐ Modify Case

Upper Case

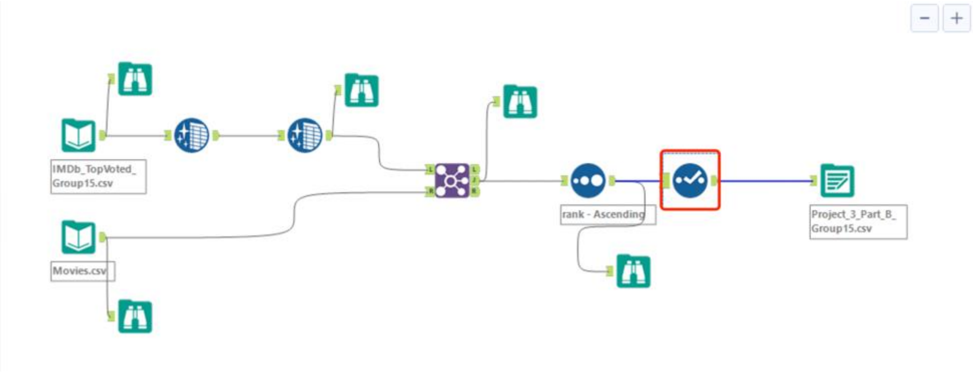
After cleaning we get the result.



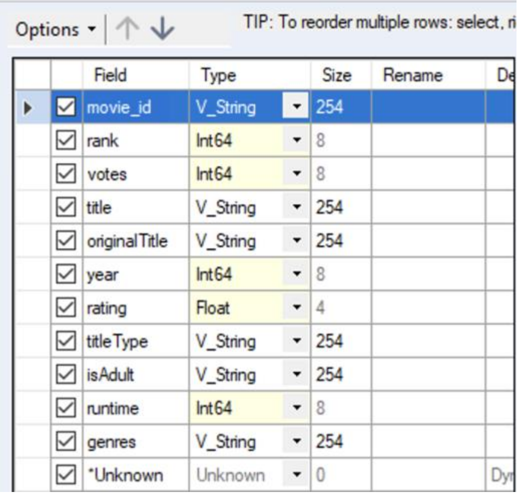
Results - Data Cleansing (2) - Output26

Record	movie_id	rank	title	year	rating	runtime	votes	genres
1	tt7286456	1	Joker	2019	8.4	122	1178623	Crime, Drama, Thriller
2	tt4154796	2	Avengers: Endgame	2019	8.4	181	1039468	Action, Adventure, Drama
3	tt4154756	3	Avengers: Infinity War	2018	8.5	149	1003600	Action, Adventure, Sci-Fi
4	tt6751668	4	Parasite	2019	8.5	132	735501	Comedy, Drama, Horror
5	tt1825683	5	Black Panther	2018	7.3	134	716603	Action, Adventure, Sci-Fi
6	tt7131622	6	Once Upon a Time… In Hollywood	2019	7.6	161	682969	Comedy, Drama
7	tt8946378	7	Knives Out	2019	7.9	130	585231	Comedy, Crime, Drama
8	tt5463162	8	Deadpool 2	2018	7.7	119	549918	Action, Adventure, Comedy

Fourth: transform the incorrect data type



By using select function, we correct the wrong data type so that every column has right type.



Options ▾ | ↑ ↓ TIP: To reorder multiple rows: select, ri

		Field	Type	Size	Rename	De
▶	☑	movie_id	V_String ▾	254		
	☑	rank	Int64 ▾	8		
	☑	votes	Int64 ▾	8		
	☑	title	V_String ▾	254		
	☑	originalTitle	V_String ▾	254		
	☑	year	Int64 ▾	8		
	☑	rating	Float ▾	4		
	☑	titleType	V_String ▾	254		
	☑	isAdult	V_String ▾	254		
	☑	runtime	Int64 ▾	8		
	☑	genres	V_String ▾	254		
	☑	*Unknown	Unknown ▾	0		Dy