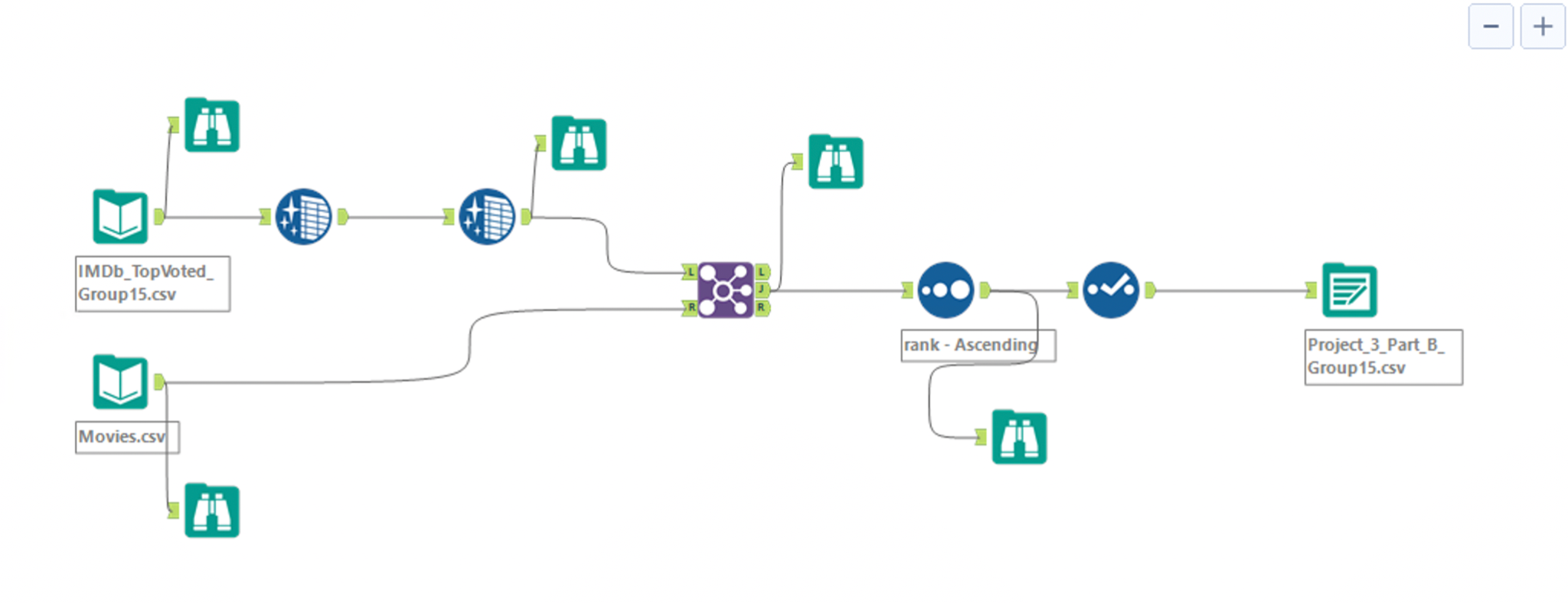
**Project 3**

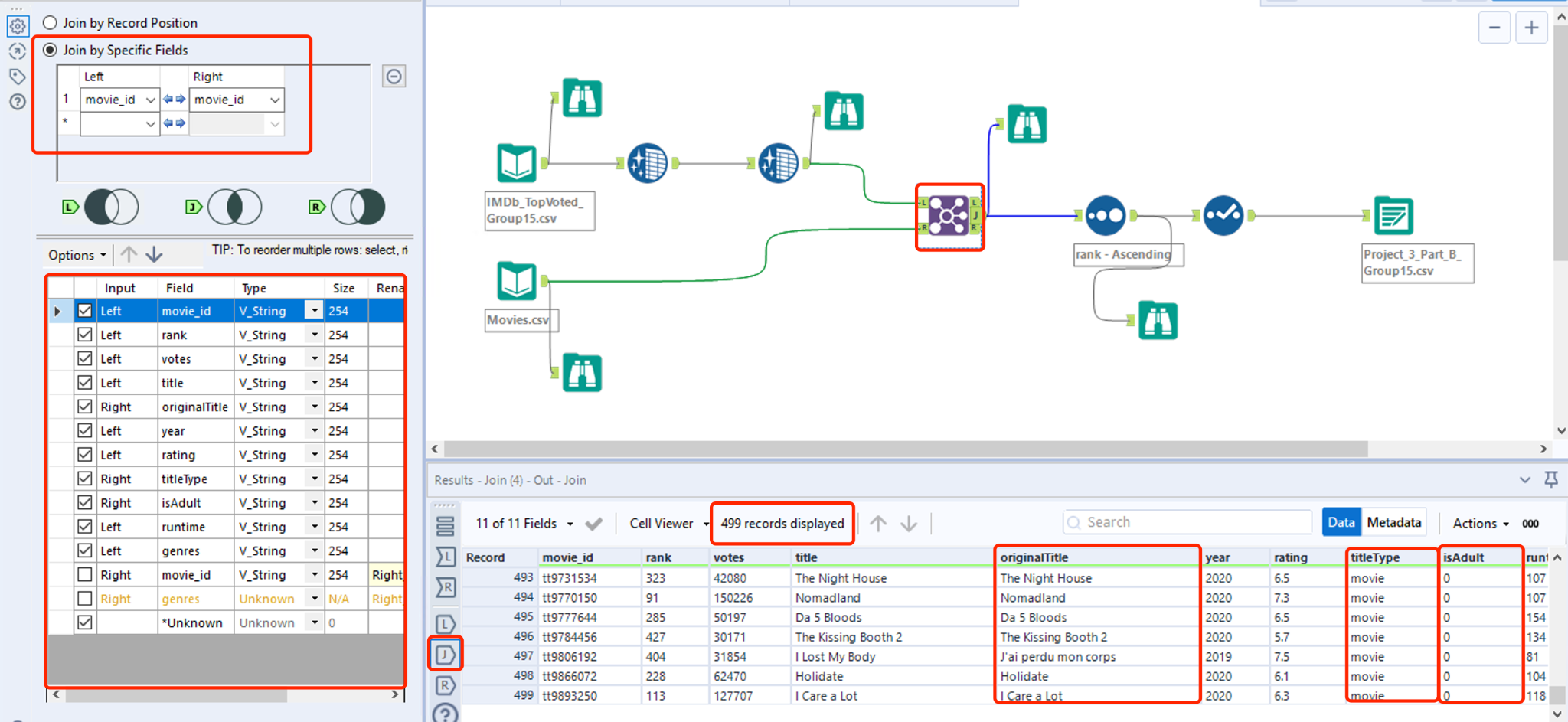
**Group15 (Haozhe Liu, Shiming Zhao, Jinsheng Luo)**

**MISM 6213**

****

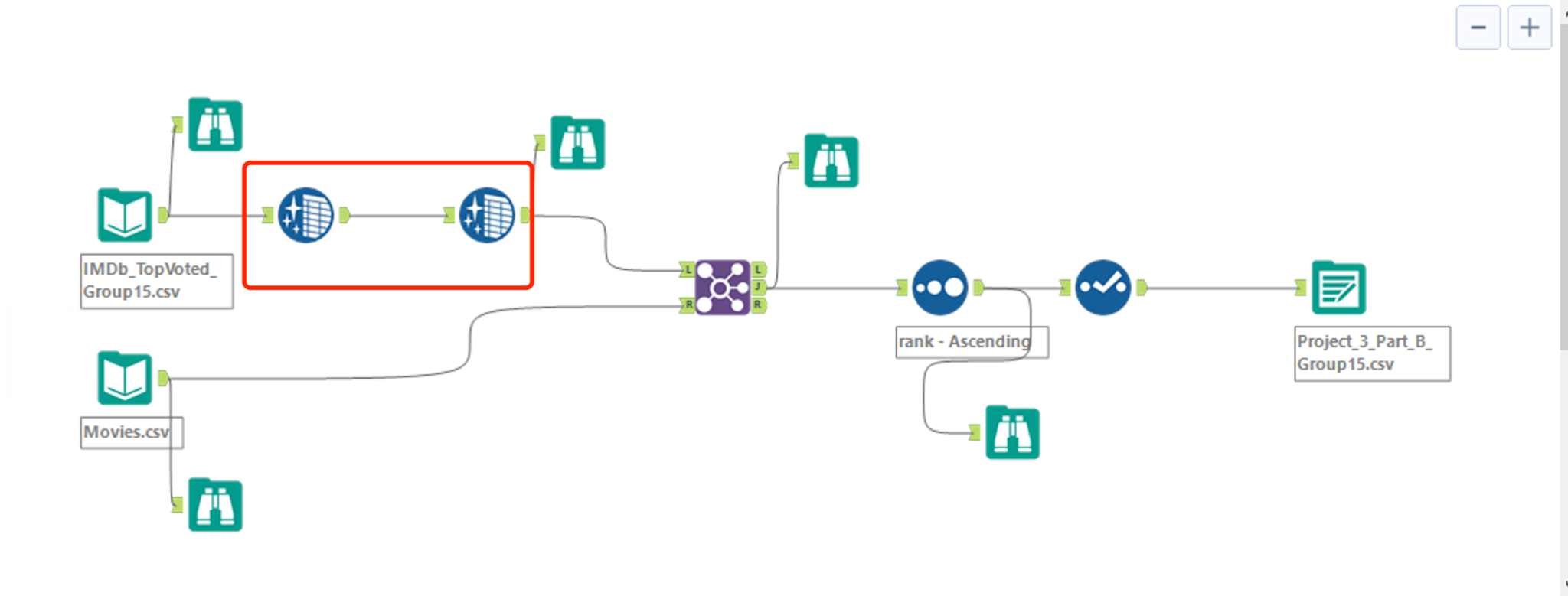
1. **What data was used to enrich the client’s data?**

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.



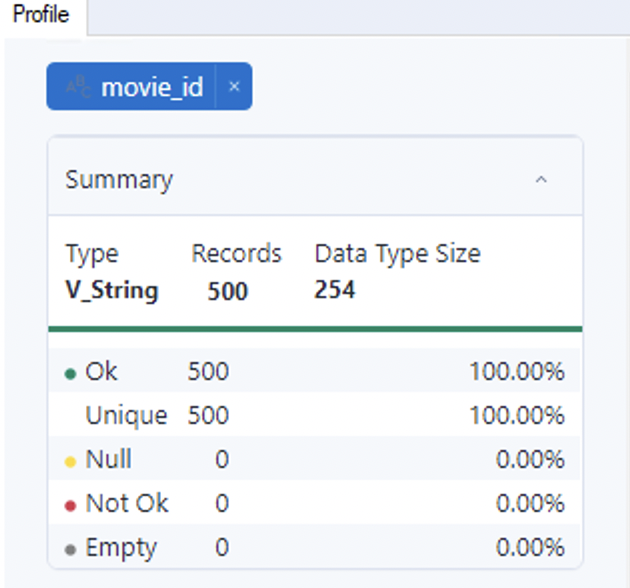


1. **Describe the data cleaning and transformation that was implemented.**

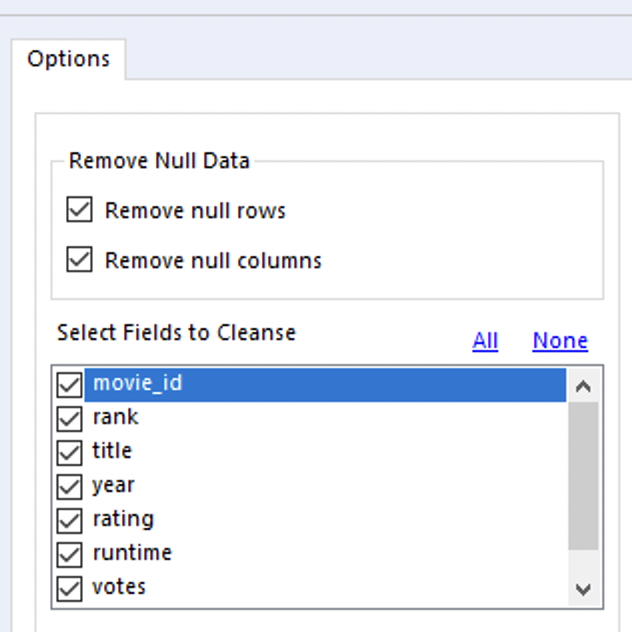


**First: Find duplicated rows**

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.



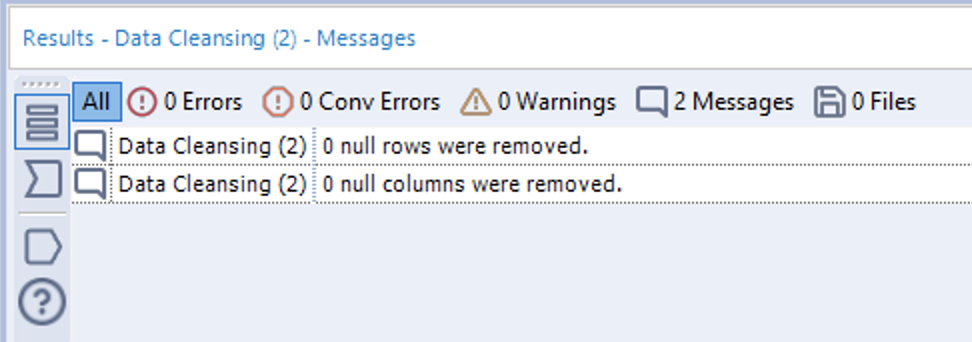
**Second: Find and replace the null value**



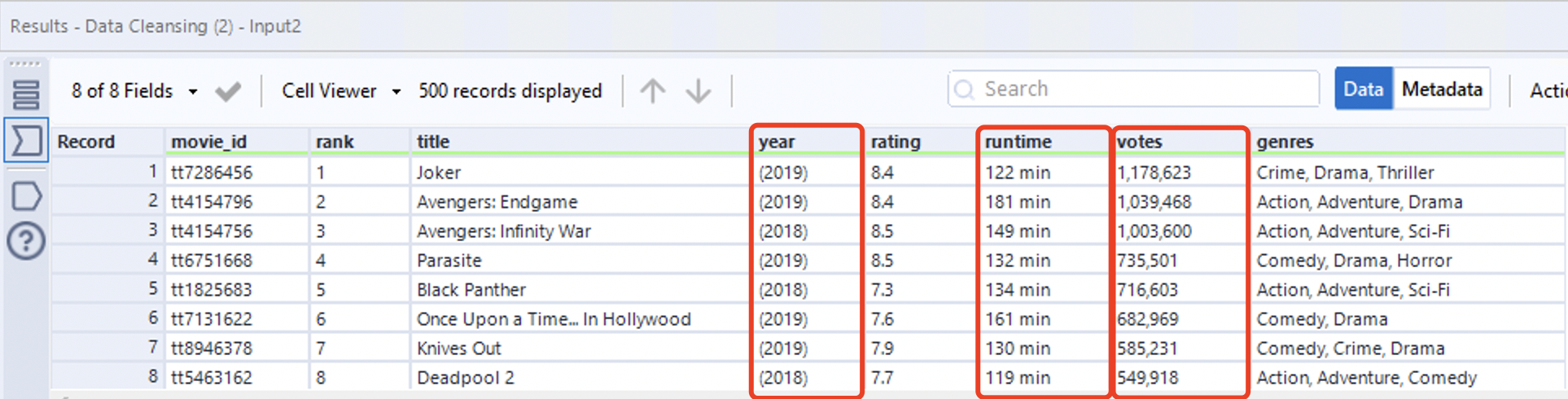
By using the data cleansing function, we selected these two options:

1. Remove null rows
2. Remove null columns

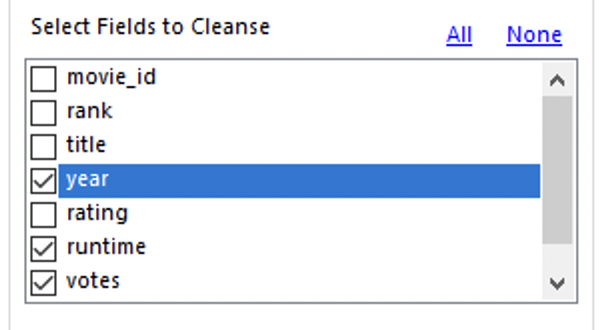
We found no rows and columns were removed, which means this data doesn't have any null rows and columns.

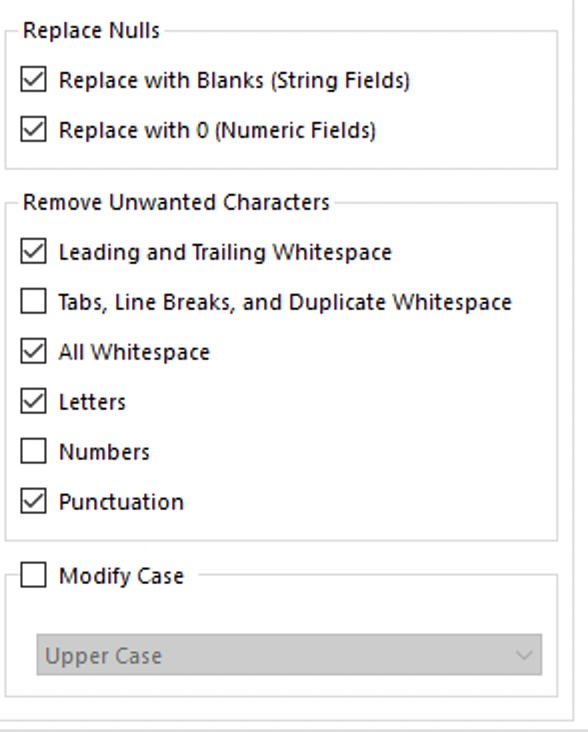


**Third: correct data format**

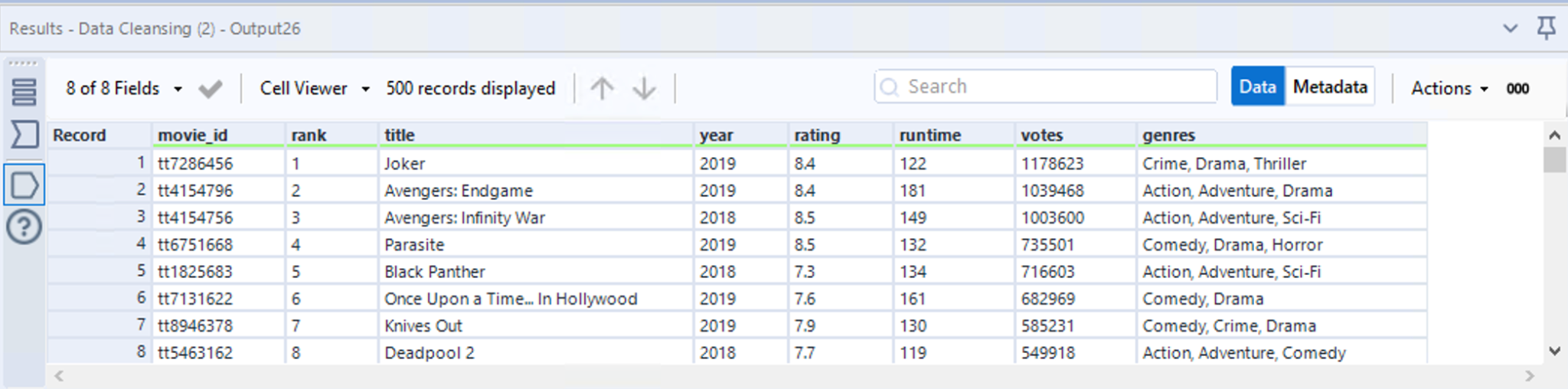


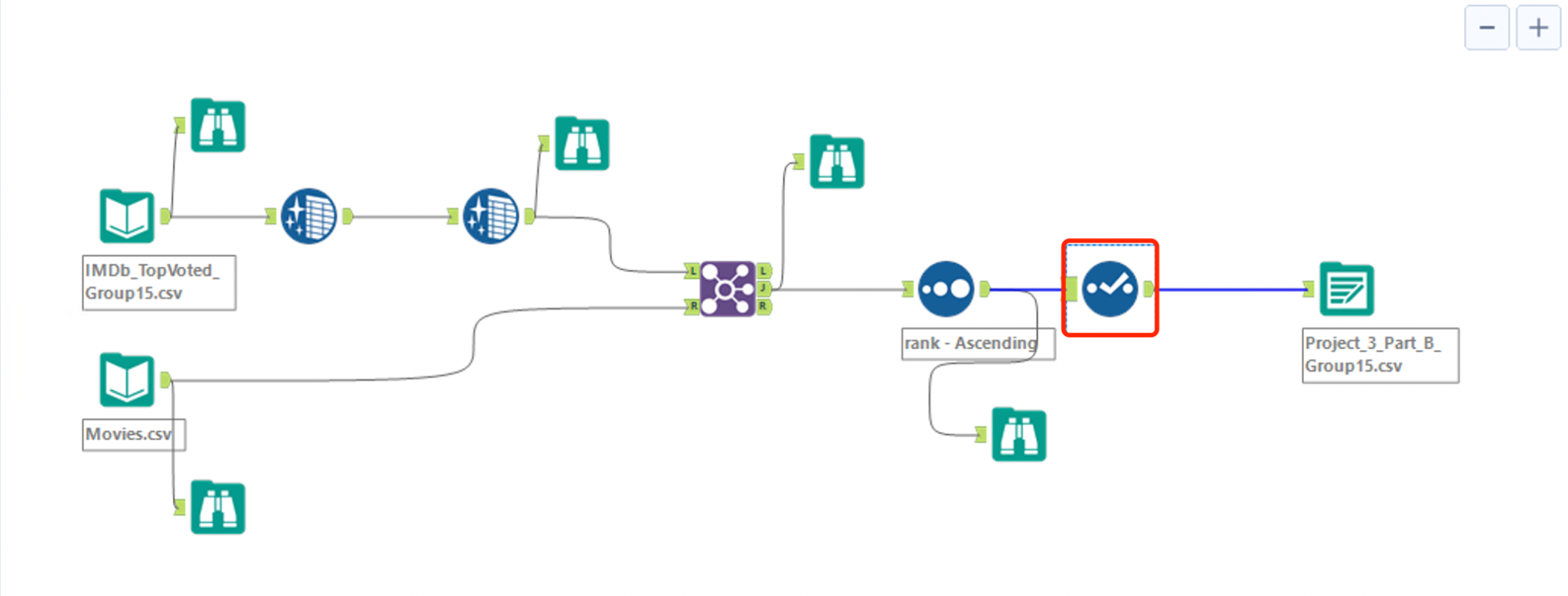
we selected ‘year’, ‘runtime’, and ‘votes’ because we think that the brackets of the ‘year’ column, the letter 'min' of ‘runtime’ column, the comma of votes are unnecessary, and these unwanted characters: Leading and trailing Whitespace, All Whitespace, Letters, and Punctuation.





After cleaning we get the result.



**Fourth: transform the incorrect data type**

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

