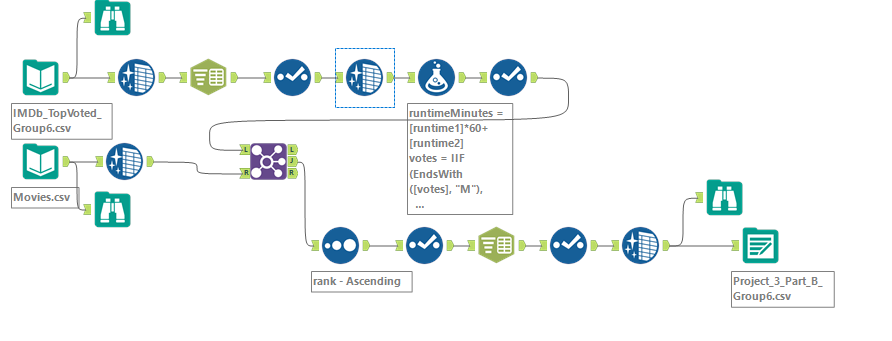
**Project 3: Data Transformation and Automation**

**Part B: Automate Data Transformation and Integration**



**Detailed Explanation of Alteryx Workflow:**

**Workflow Overview:**

This project involves automating the cleaning, transformation, and integration of two datasets using Alteryx for further analysis:

IMDb\_TopVoted\_Group6.csv: Contains top-voted movies data obtained via web scraping.

Movies.csv: Provides additional metadata, including a description column.

**Step 1: Data Cleaning and Transformation (IMDb\_TopVoted\_Group6.csv)**

Runtime Column Standardization:

Used the Data Cleansing tool to replace null values with blanks ("") or 0 to prevent errors.

Stripped leading/trailing whitespaces and removed non-numeric characters (e.g., "h," "m") for uniformity.

Split the cleaned runtime into two new columns using the Text to Columns tool:

runtime1 (hours).

runtime2 (minutes).

Converted runtime1 and runtime2 to Int64 format using the Select tool.

Calculating Total Runtime:

Used the Formula tool to calculate total runtime in minutes:

runtimeMinutes=(runtime1×60)+runtime2

Stored the result in a new column, runtimeMinutes.

Numeric Data Standardization:

Converted columns rank, year, runtimeMinutes, and votes to Int64 for consistency in numeric operations.

**Step 2: Data Cleaning (Movies.csv)**

Replaced null values in the description column with blanks ("").

Removed extra spaces to standardize text fields.

**Step 3: Dataset Integration**

Used the Join tool to merge cleaned datasets (IMDb\_TopVoted\_Group6.csv and Movies.csv) using movie\_id as the key.

Sorted the joined dataset in ascending order by rank to prioritize the highest-rated movies.

Reorganized columns using the Select tool for logical and user-friendly arrangement.

**Step 4: Genre Column Transformation**

Used the Text to Columns tool to split the genre column into three distinct columns: genre01, genre02, genre03.

Removed the original genre column after splitting.

Replaced null values in new genre columns with blanks ("").

Standardized genre data by removing extra spaces and whitespaces.

**Step 5: Output Data**

Saved the final enriched dataset using the Output Data tool as Project\_3\_Part\_B\_Group6.csv.

The output dataset is fully cleaned, transformed, and ready for advanced analysis or reporting.