Name : Shashank Barai

Data Profiling and Cleaning Recipe: YouTube Trend Videos Data set

Ingredients:

- Five CSV files of 5 different countries YouTube Trending Videos and 1 text file
- Join function
- Union
- Split function
- Calculated fields
- Aggregate functions
- Relevant column renaming

Recipe:

1. Combine all five CSV files of YouTube Trend Videos using a Join function and remove irrelevant fields like thumbnails and links, resulting in 18 fields and 170K rows.
2. Address the challenge of improving the Trending-Date field:
    - Convert the Trending-Date field to date format.
    - Correct discrepancies in date distribution using the Split function to separate days, months, and years, and correct the format with a calculated field.
3. Clean and extract main keywords from the Tag fields:
    - Eliminate duplicate tags with a calculated column and rename it to Tags.
    - Preserve the Tag field, cleaning only special characters and excess whitespace, renaming it to 'main tag'.
4. Prepare for sentiment analysis:
    - Extract the first character from the Description column to capture sentiment.
    - Preserve the Original Description field, cleaning only special characters and excess whitespace.
5. Rename all fields/columns for clarity in user story identification.
6. Enrich and integrate data with category information:
    - Combine the cleaned CSV with a text file containing categories using a left outer join, utilizing 'category_id' from the combined CSV and 'id' from the text file.
7. Utilize aggregate functions for business queries:
    - Examine views, likes, and dislikes over days or months for each Video-id to assess popularity.
    - Generate an output CSV named final 2 with 68K rows and 5 columns.
8. Perform final data cleaning:
    - Further clean the main CSV by removing null values, duplicates, and unnecessary fields like category_id and id.
    - Ensure the category field from the text file takes precedence.
    - The resulting output CSV, named Final 1, comprises 170K rows and 17 fields.

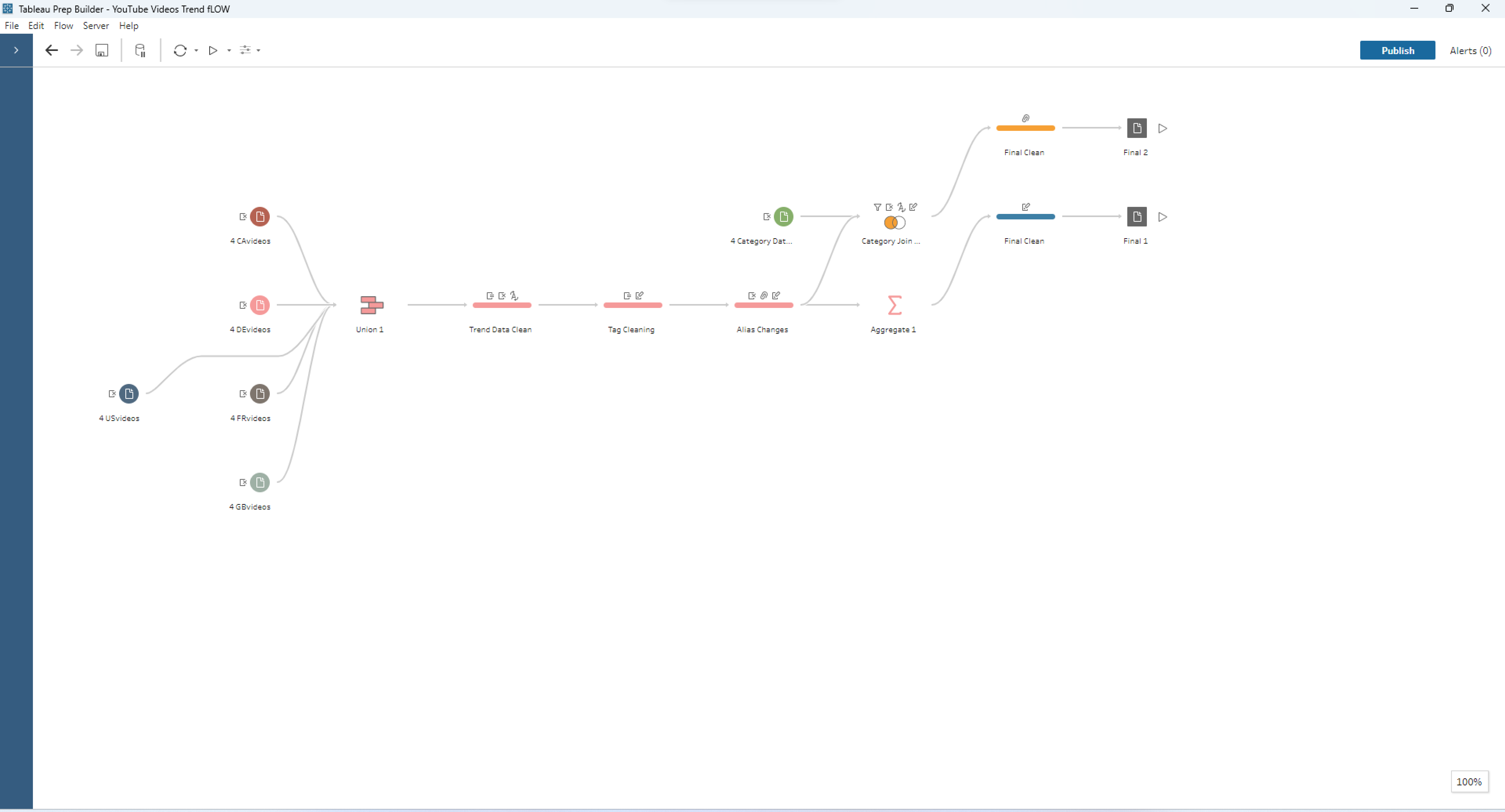
Tableau Prep Builder - YouTube Videos Trend fLOW
File
Edit
Flow
Server
Help
Publish
Alerts (0)
4 CAvideos
4 DEvideos
4 USvideos
4 FRvideos
4 GBvideos
Union 1
Trend Data Clean
Tag Cleaning
Alias Changes
4 Category Dat...
Category Join ...
Aggregate 1
Final Clean
Final 2
Final Clean
Final 1
100%