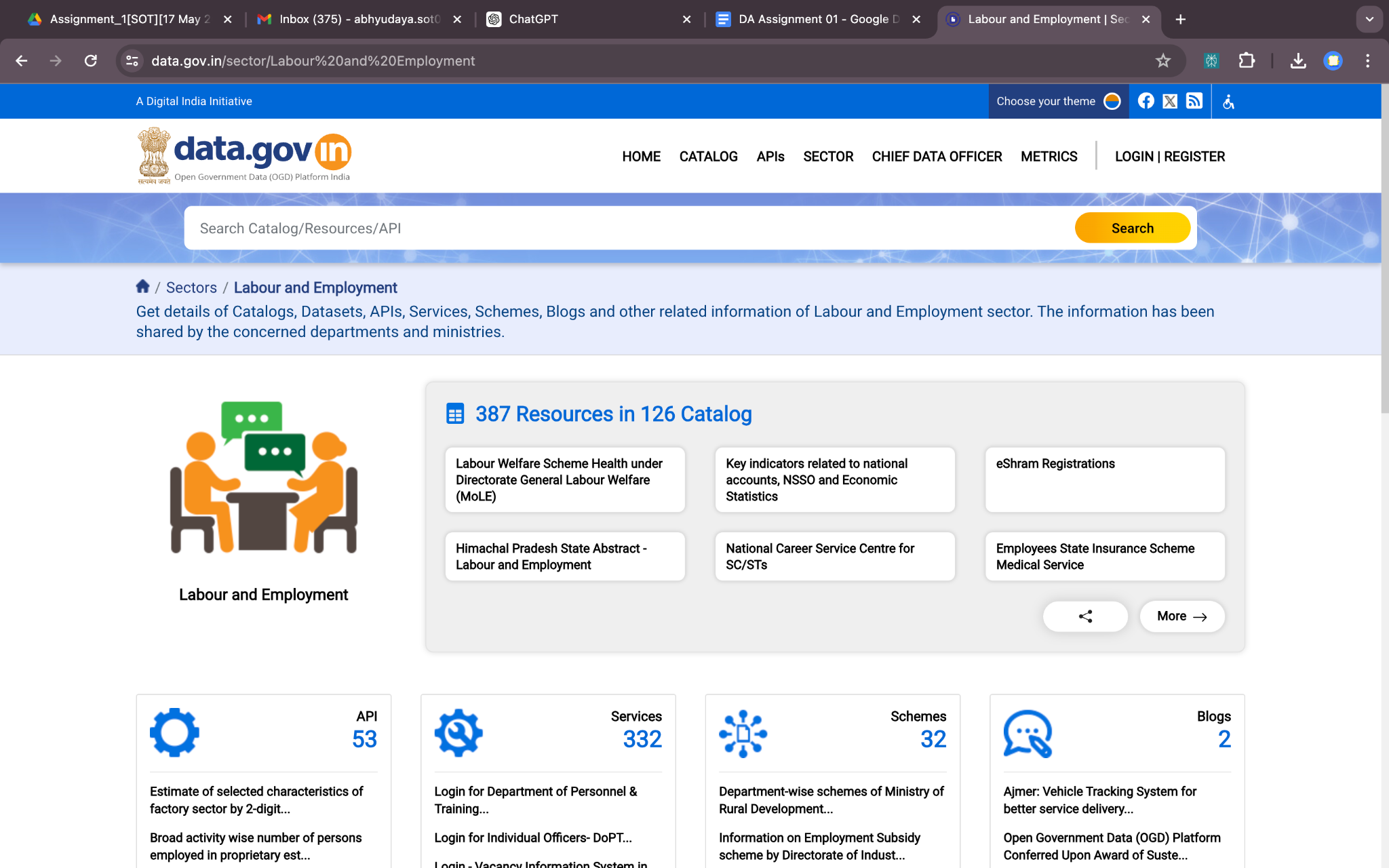
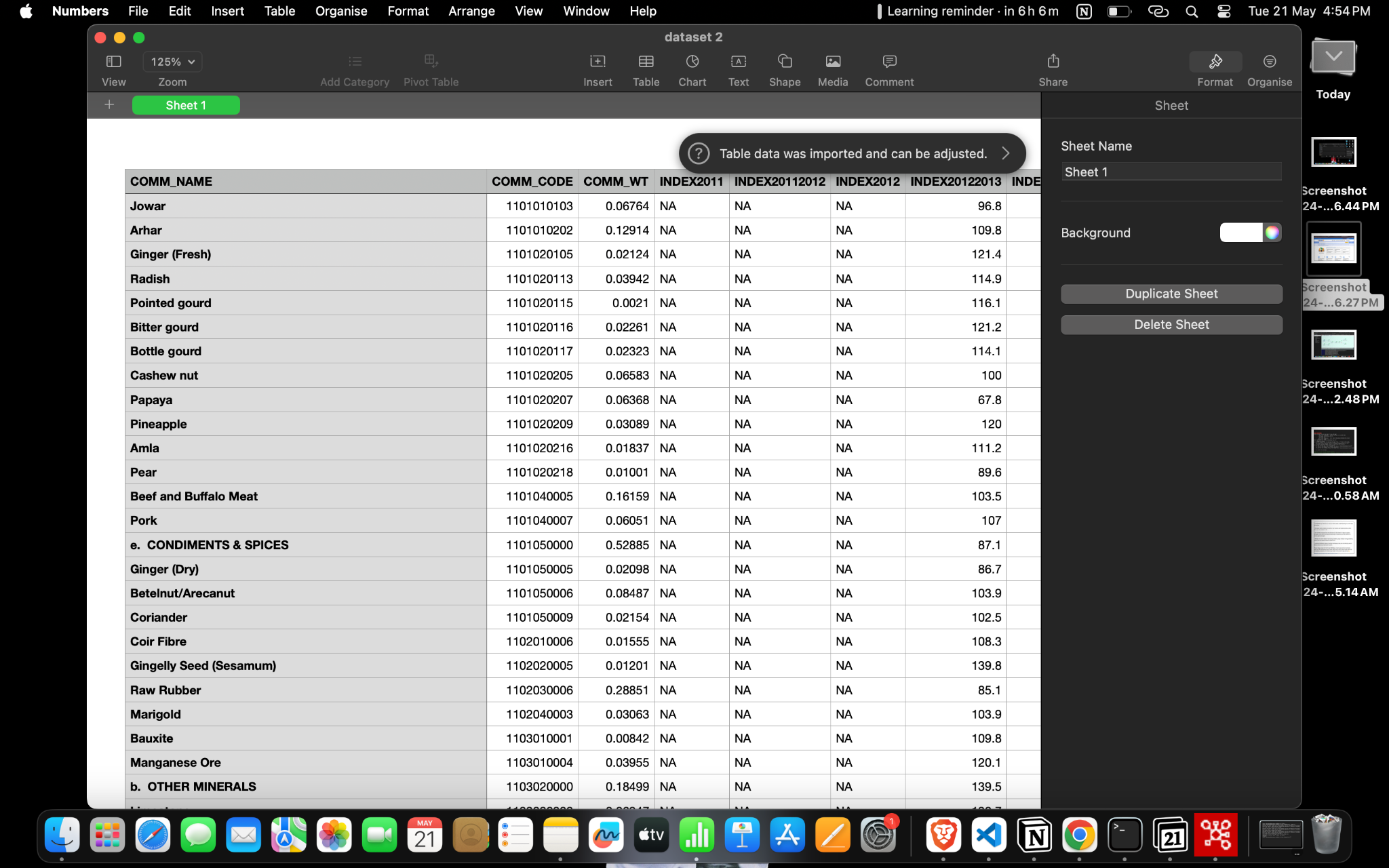
**1.Data Selection**

I selected the data from a public repository on the government website Data.gov.

**2.Task Overview:**

****

- The wholesale data provided has been sourced from government records, detailing prices of daily essentials such as oils, crops, and vegetables.

- This dataset includes both numerical and categorical columns, offering a comprehensive view of wholesale prices across various categories.

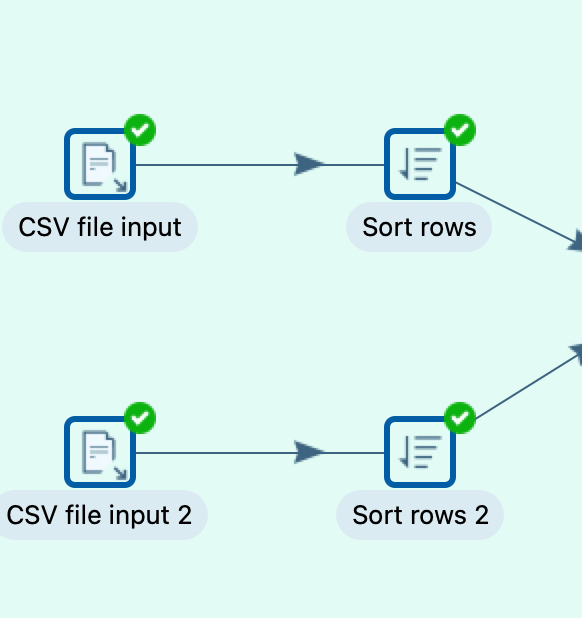
- The provided dataset has been edited to address missing values and remove duplicate entries.

- Additionally, the dataset has been divided into two parts.

- When these two parts are merged together, access to complete rows is ensured, allowing for comprehensive data analysis.

**The task I am currently undertaking involves:**

1. **Initially, I'll organize the rows in ascending order for both segments of the incomplete dataset.**

****

The **"Sort Rows"** step allows you to sort the rows of data based on one or more fields within those rows.

Here's how it typically works:

1. **Select Fields to Sort**: You specify which fields or columns you want to sort your data by.

2. **Ascending or Descending Order**: You define whether the sorting should be in ascending or descending order for each field.

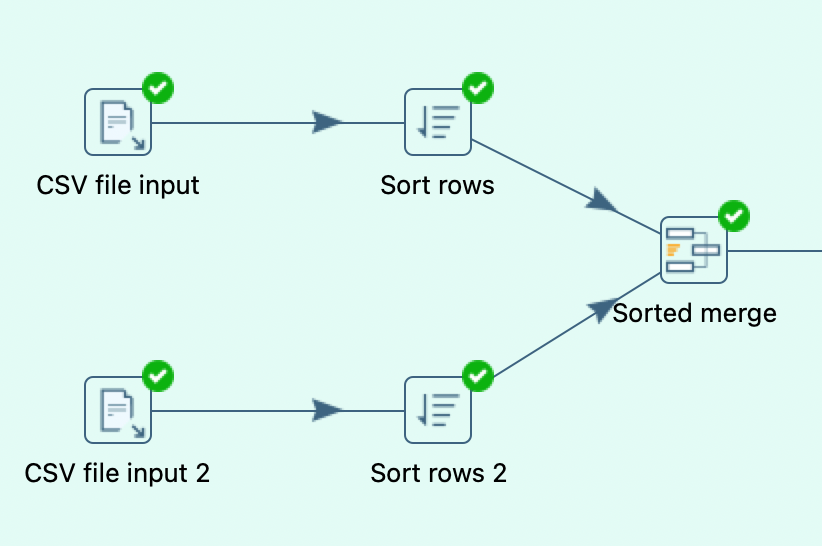
3. **Sort Algorithm**: You may also have options for the sorting algorithm to be used, depending on the version of Pentaho you're using. Common options include quicksort or mergesort.

4. **Sorting Options:** You might have additional options for handling null values or case sensitivity during sorting.

5. **Output**: The step outputs the sorted rows of data, ready for further processing or output to a destination.

This step is often used in data transformation processes where you need to arrange data in a specific order before performing further operations, such as joining with another dataset or aggregating values. Sorting data is a fundamental operation in data processing, and the "Sort Rows" step in Pentaho provides a convenient way to achieve this within your data pipelines.

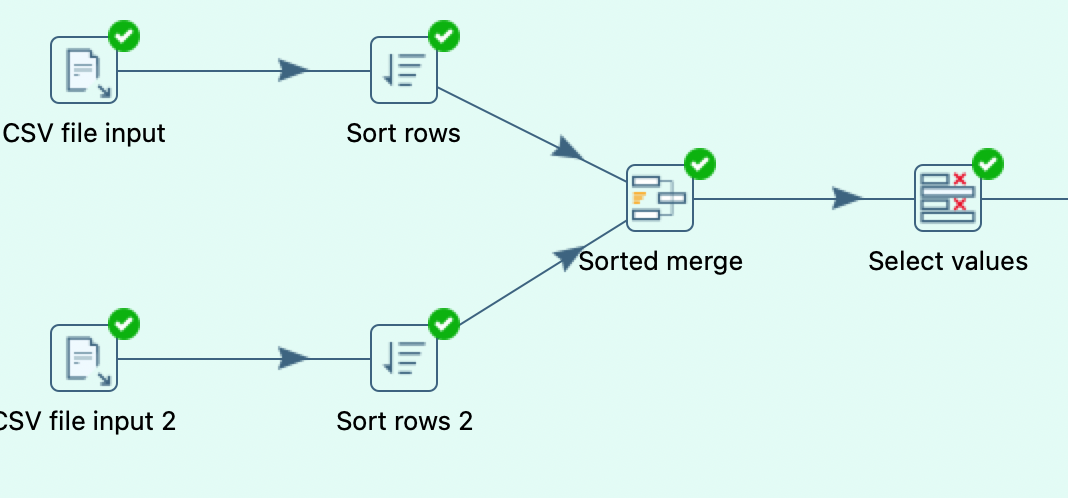
2.**Merging the two sorted dataset:-**

****

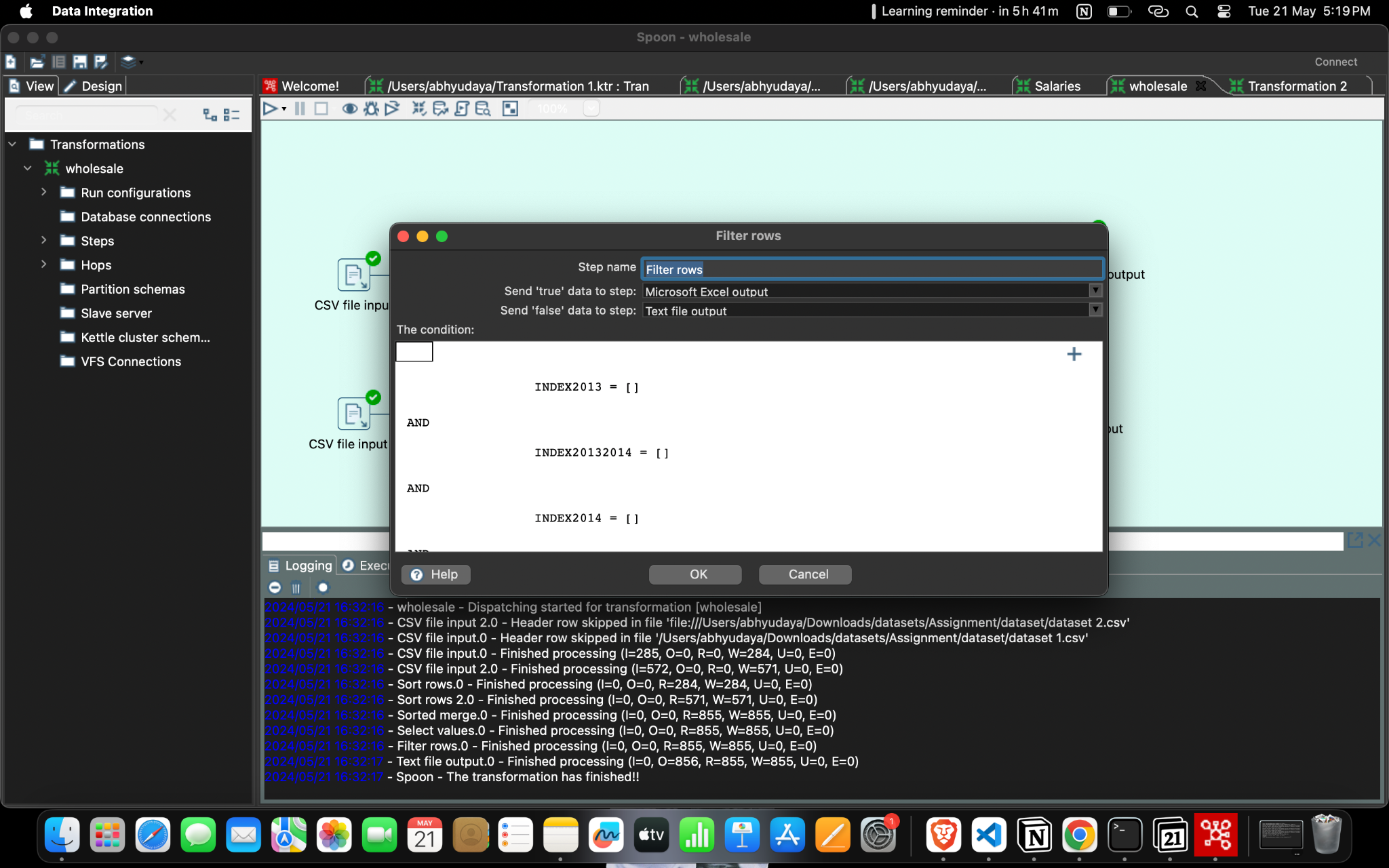
The "**Sorted Merge"** step in Pentaho, also known as "Merge Join" in some versions, is used to merge two sorted datasets based on a common key field or fields. This step is similar to a database join operation, but it requires that both input streams are sorted based on the join key beforehand.

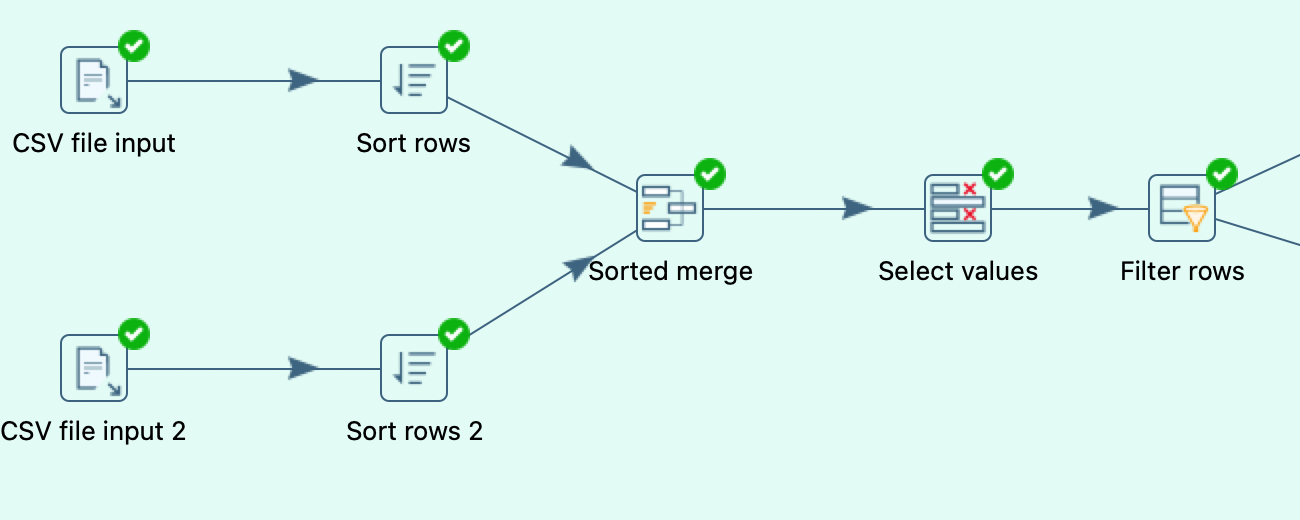
integration workflows to combine data from different sources before further processing or analysis**.**

**3.Utilizing the "Select Values" method offers a versatile approach for refining data, encompassing crucial tasks such as meticulous data type management, precise field selection, and efficient removal of duplicate values.**

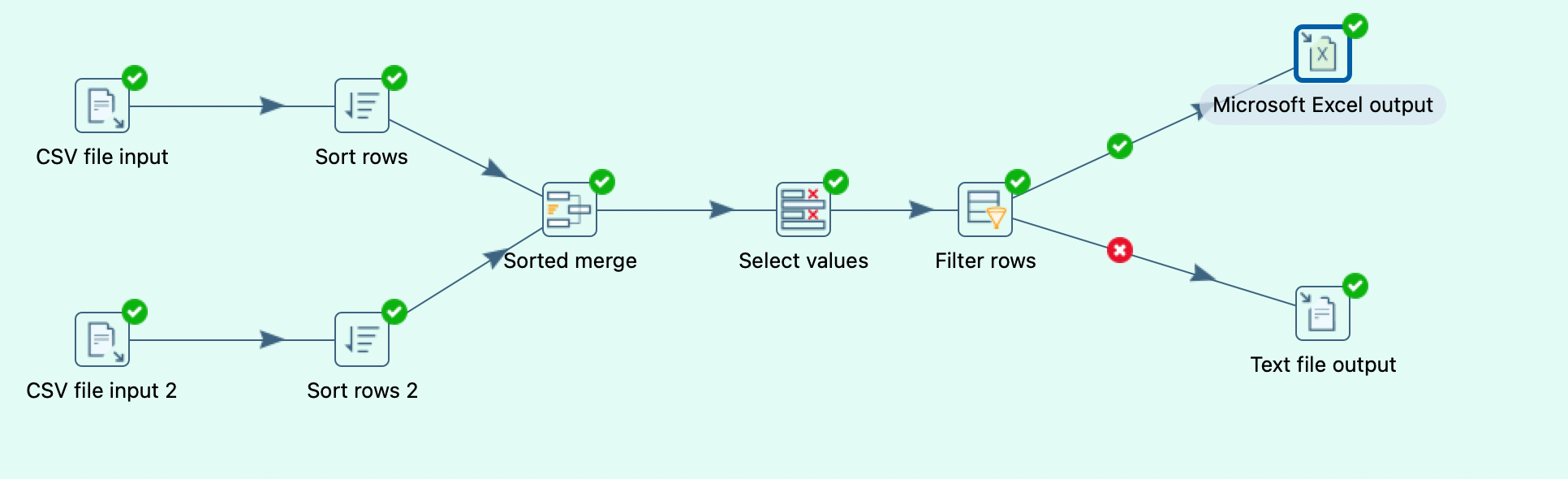
****

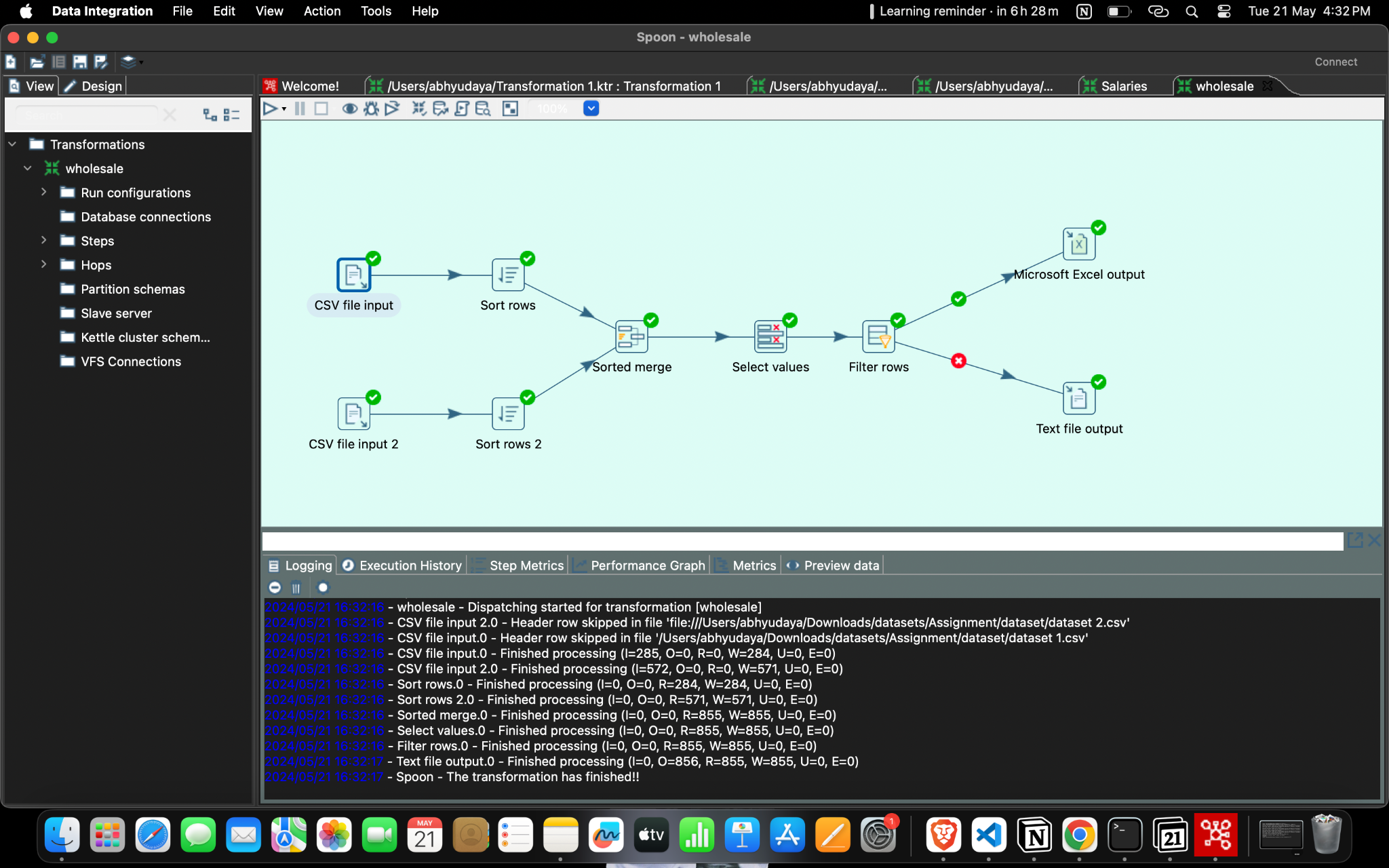
**4.Utilizing the "Filter Rows" method to meticulously refine the dataset by systematically eliminating duplicate entries and null values, ensuring the data integrity remains pristine.**

****

****

**5. Effectively isolating accurate data from erroneous entries post-application of the 'Filter Rows' method.**

****

****