Doing ETL on data process through various methods

B. karthikeya 2301010031

ETL

pipeline using Pentaho Data Integration (PDI).

Practical (10 Marks )

Download a dataset of your choice and build a basic ETL pipeline using Pentaho Data Integration (PDI).

Instructions:

1. Data set Selection: ● Choose a dataset that interests you from a public data repository (e.g., Kaggle, UCI Machine Learning Repository, or any government data portal). Ensure the dataset has at least two distinct types of data columns (e.g., numerical and categorical data).

2. Task Overview: ● Describe the dataset you have chosen, including the source, the nature of the data, and why you selected this particular dataset.

3. ETL Pipeline Design: 2

● Extract: Outline the steps you took to extract the dataset into Pentaho. Detail any challenges you faced during the extraction phase.

● Transform: Describe the transformations you applied to the data. This could include filtering rows, converting data types, handling missing values, or merging data from multiple sources. Explain the rationale behind each transformation.

● Load: Discuss how you loaded the data into a target system (this could be another database, a flat file, or a data warehouse schema). Describe the structure of the target system and why it was chosen.

4. Challenges and Solutions:

● Highlight any specific challenges you encountered while setting up your ETL pipeline in Pentaho. Discuss how you resolved these issues. 5. Outcome and Reflection:

● Reflect on the process of building the ETL pipeline. Discuss what you learned and how the exercise helped you understand ETL concepts better.

● Optionally, include screenshots of your Pentaho workflow to supplement your explanation.

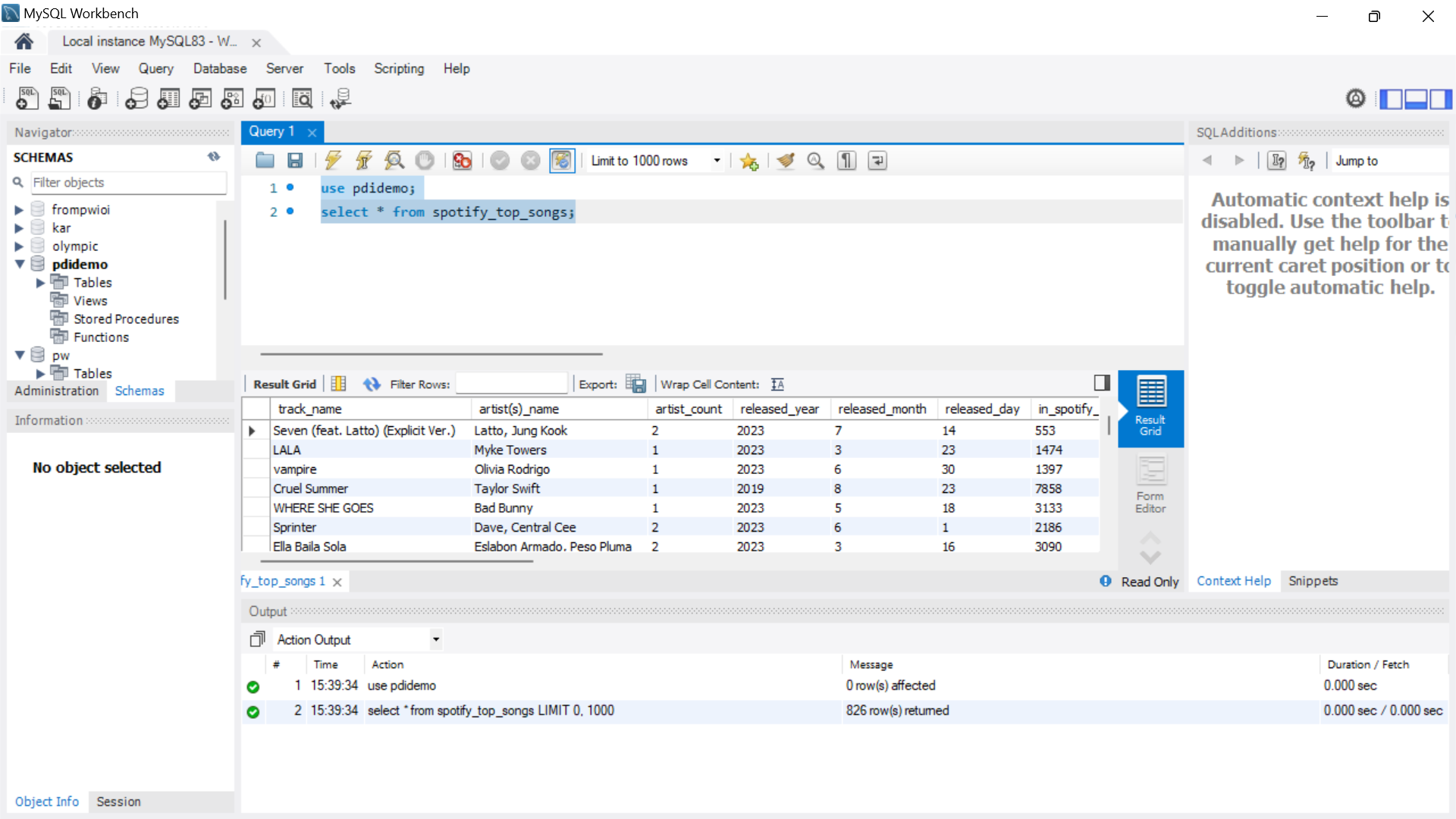
Link of dataset: <https://www.kaggle.com/datasets/nelgiriyewithana/top-spotify-songs-2023>

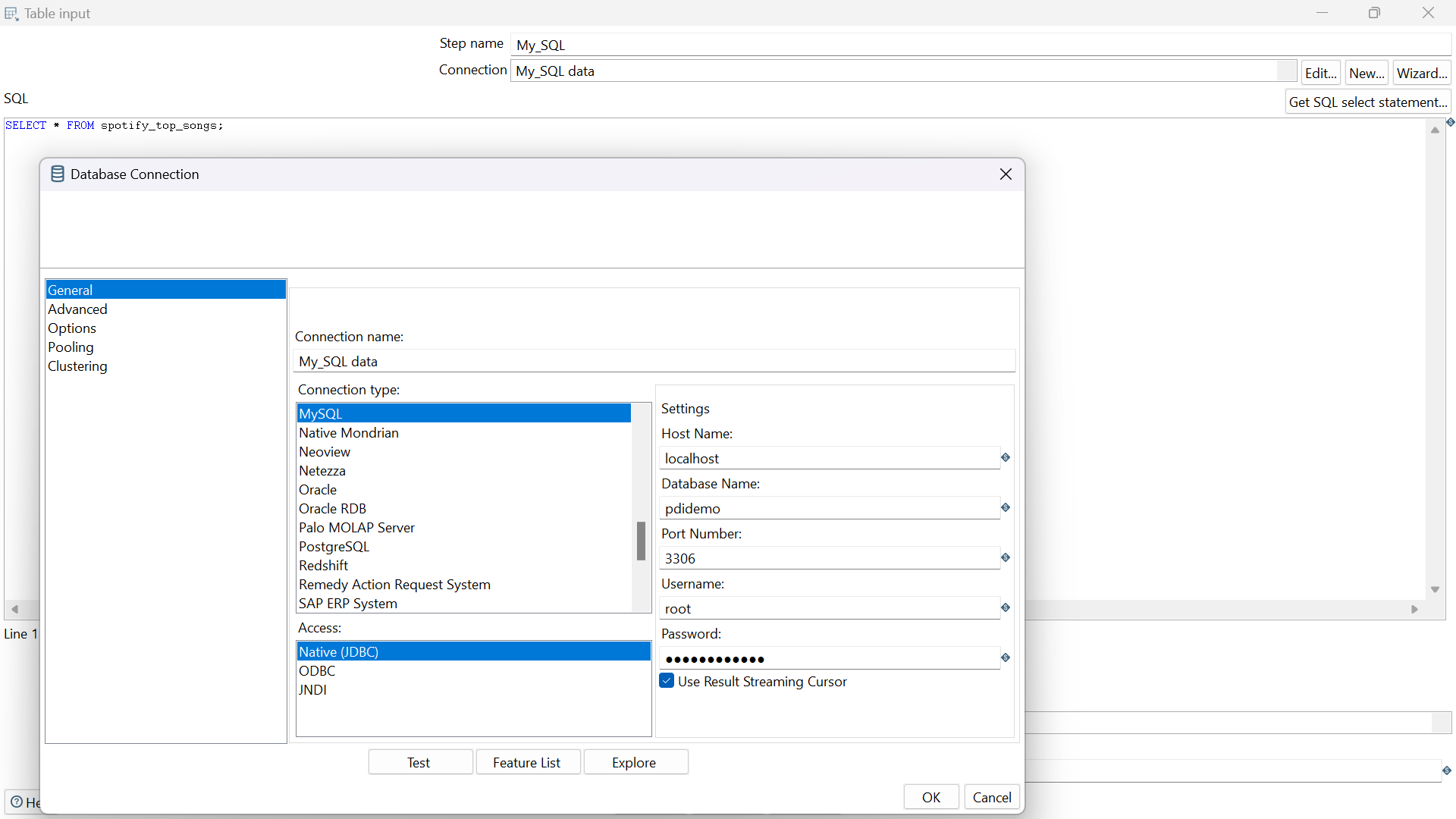
I did the dataset in two ways one by My\_SQL and one by input\_csv

Method\_1:

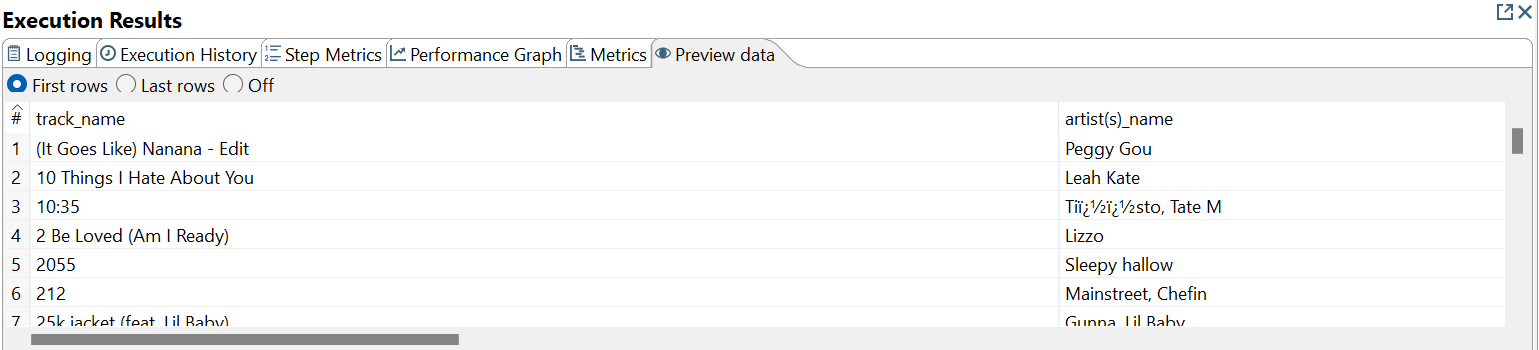
Case-1: In this, I did in two cases on by changing name and another by not changing name

* I created a data base , as pdidemo and imported a table called spotify\_top\_songs

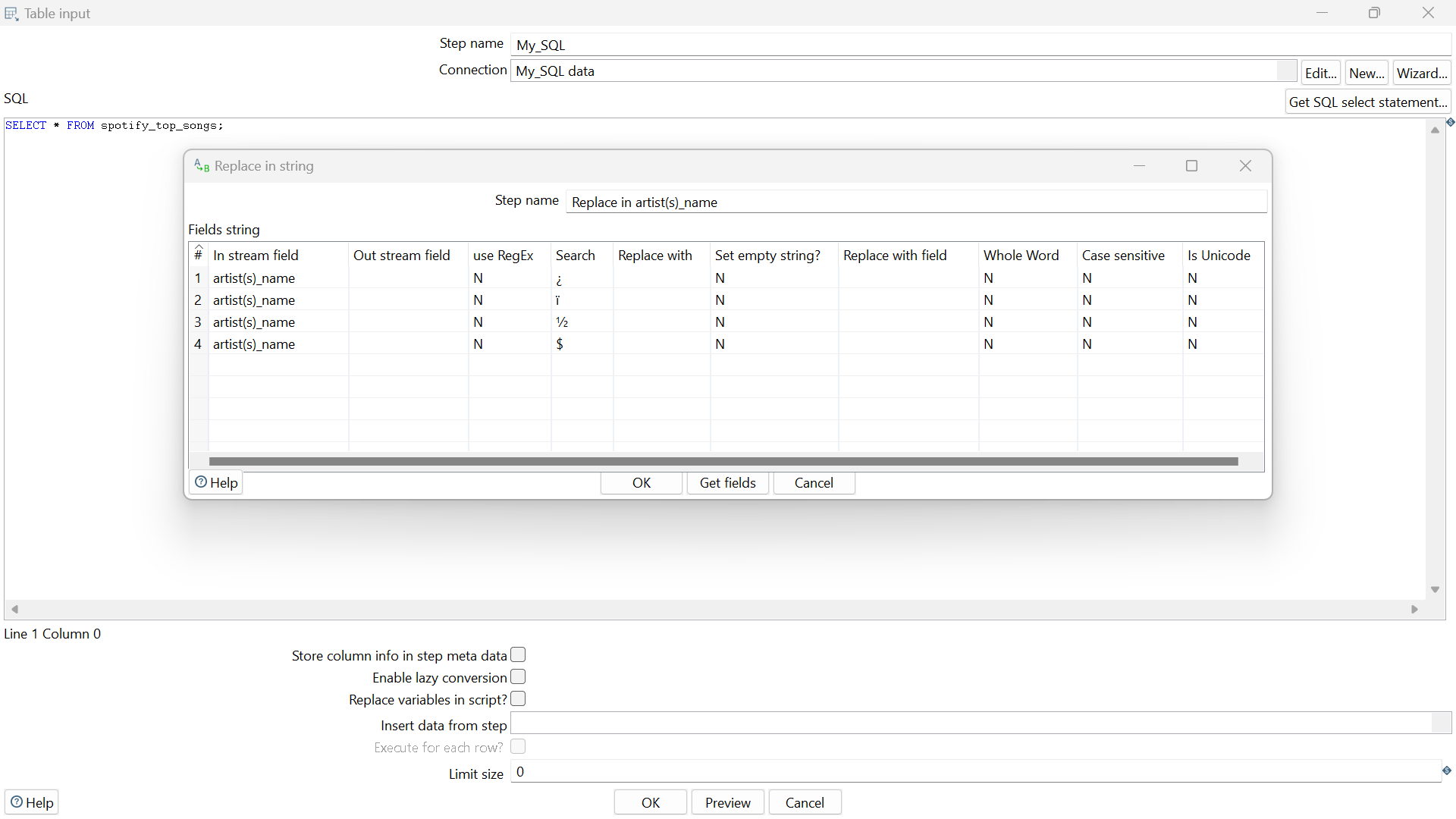


* And I open Pentaho and I connected My\_SQL and Pentaho 
* And I sorted rows by “track\_name” in ascending order
* Then, later I make unique rows by “artist\_name” in ignoring the cases
* Replacing some characters(suppose in line 3 in “artist\_name”) like that it contains in many rows

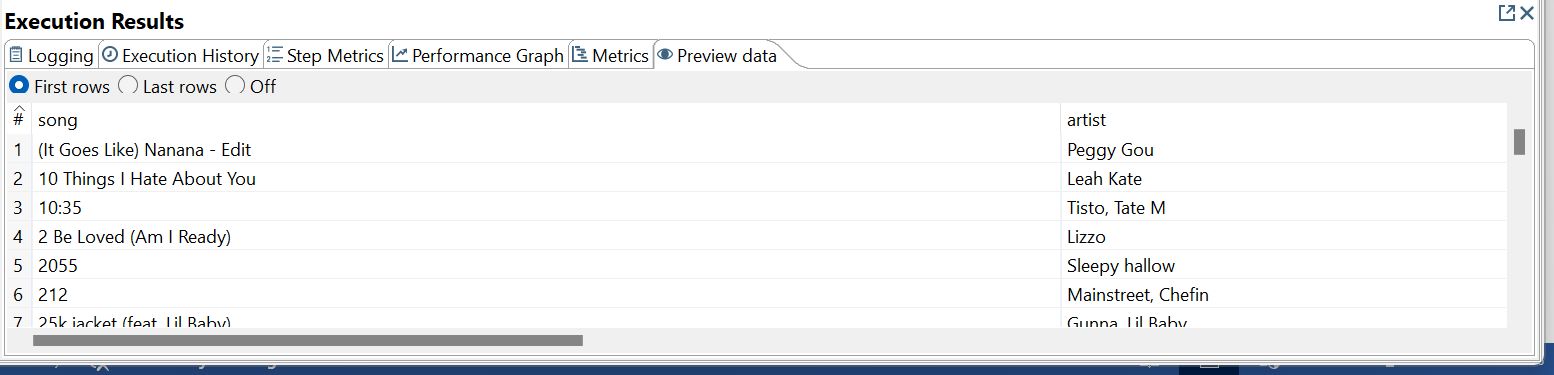
* Before



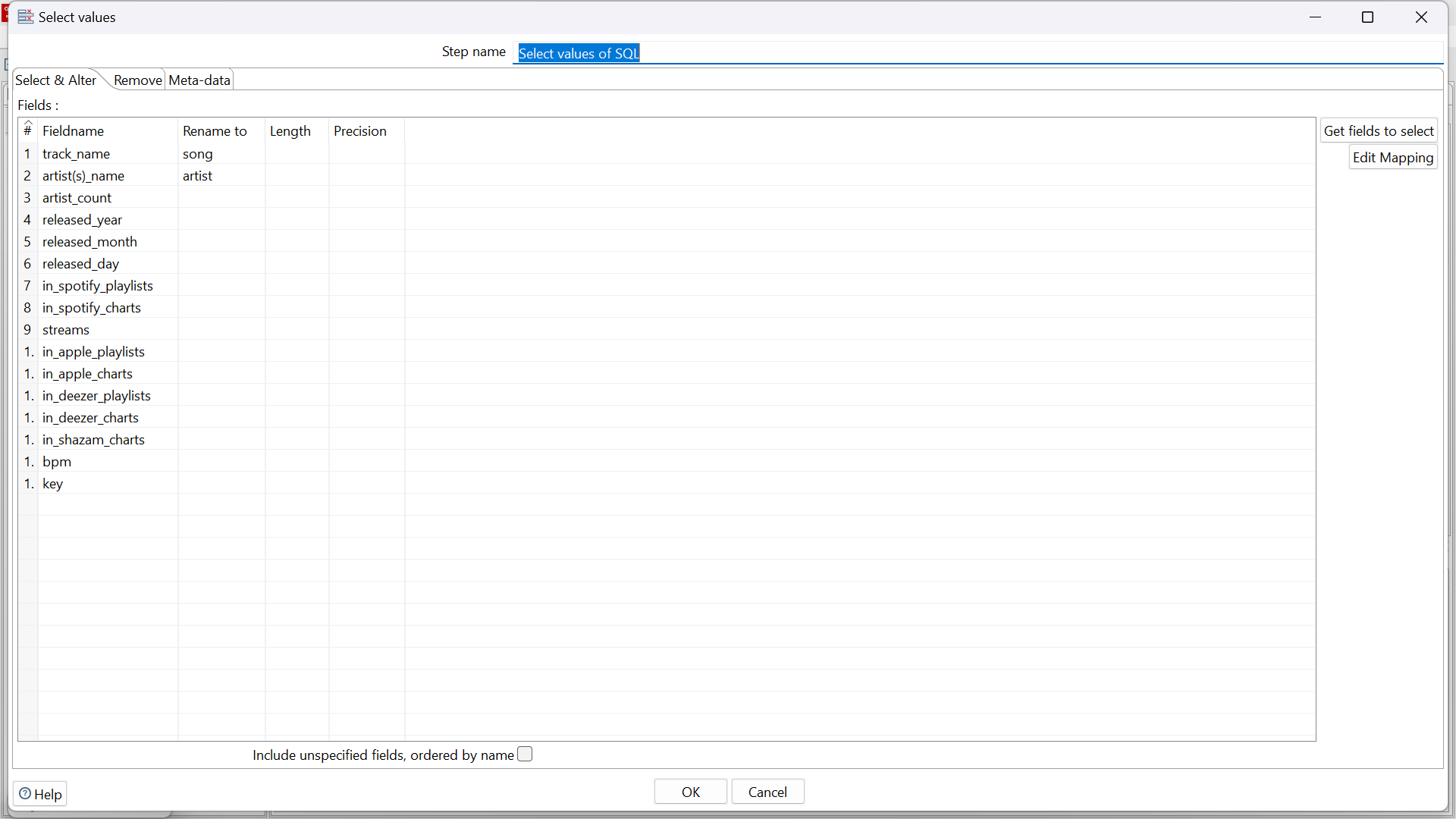
* Using this



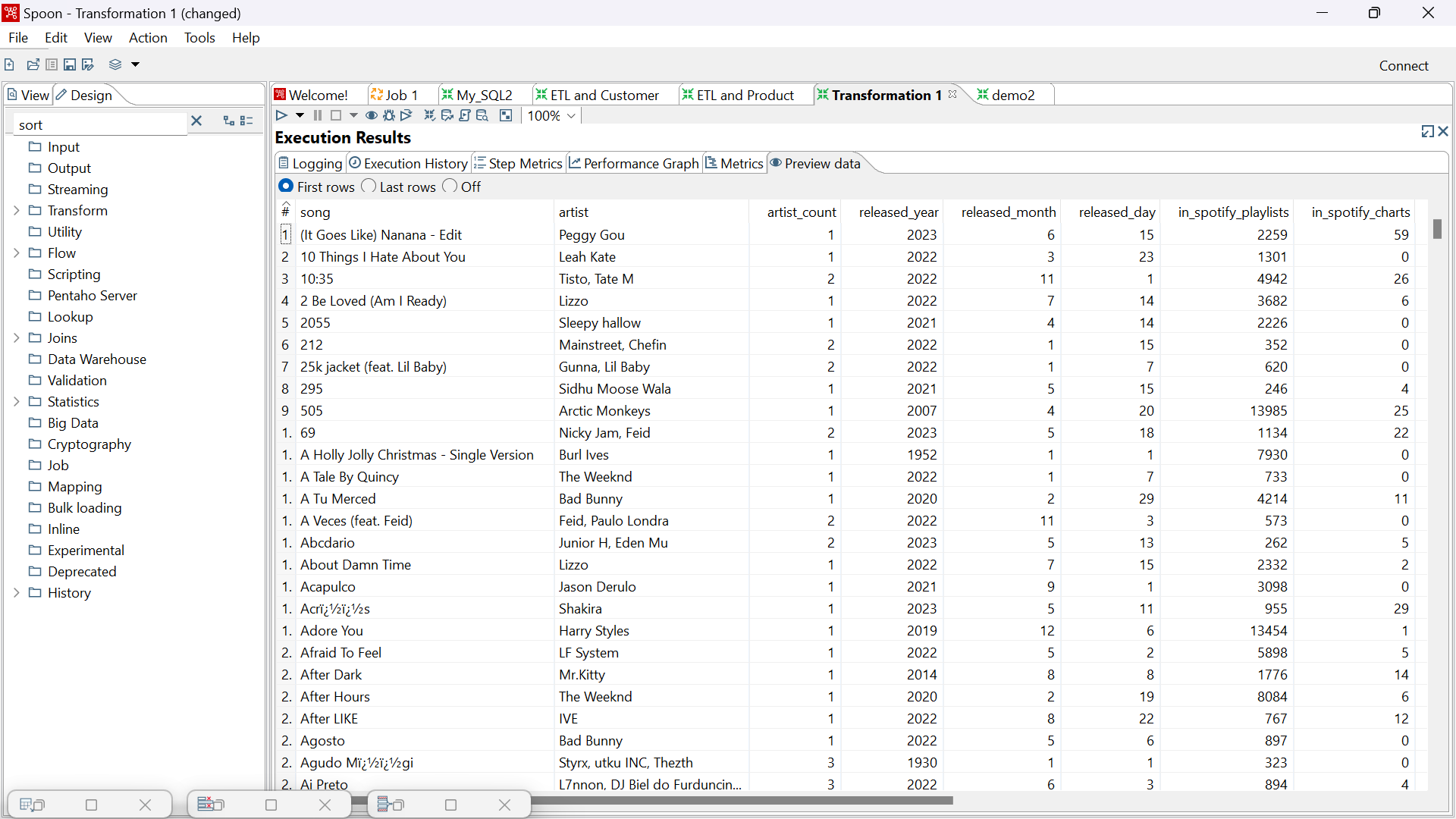
* It become



* I removed last rows and I want up “key” and I rename first two rows

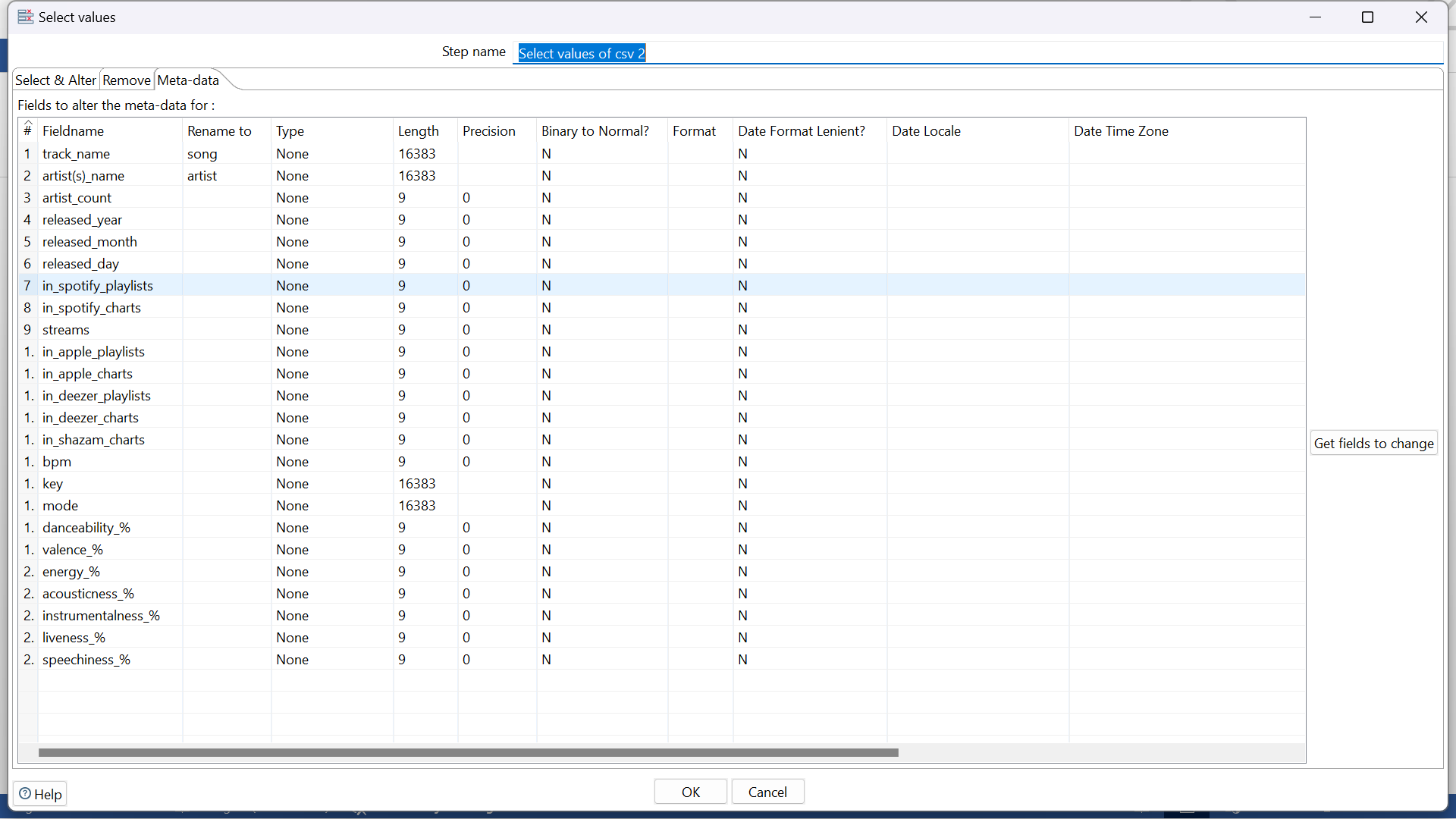


* I got out put as



Case-2:

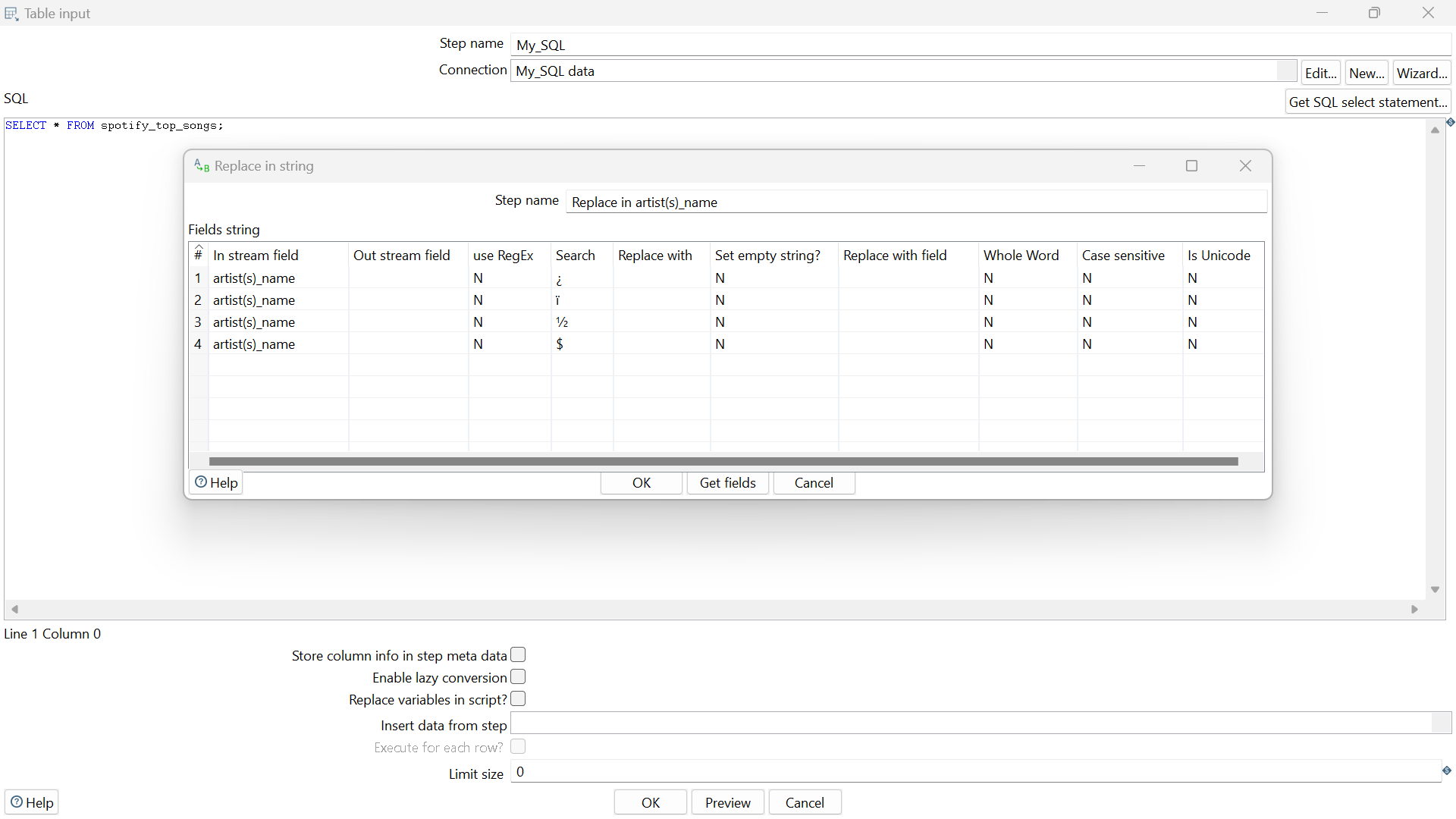
I change the “title\_name” to “song” and “artict(s)\_name” to “artist”

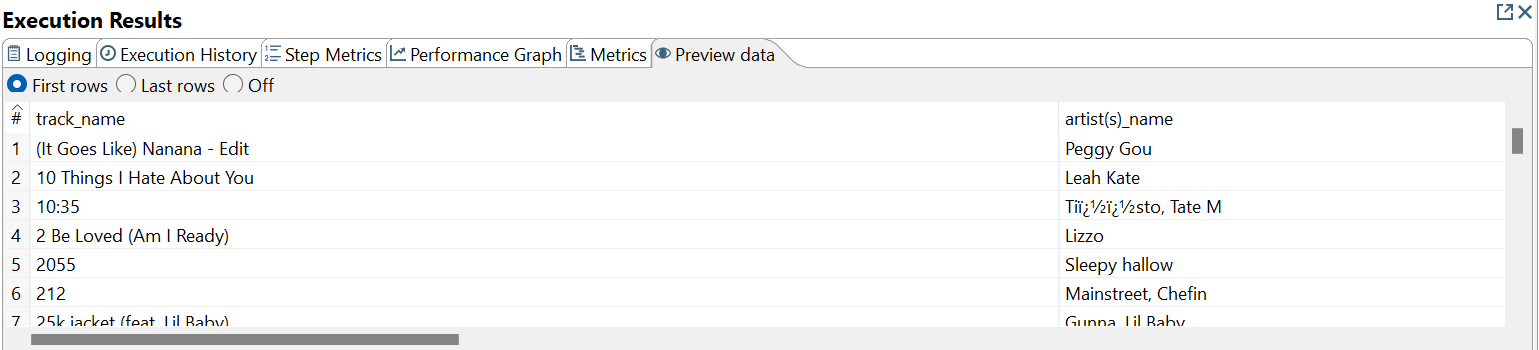


Method\_2:

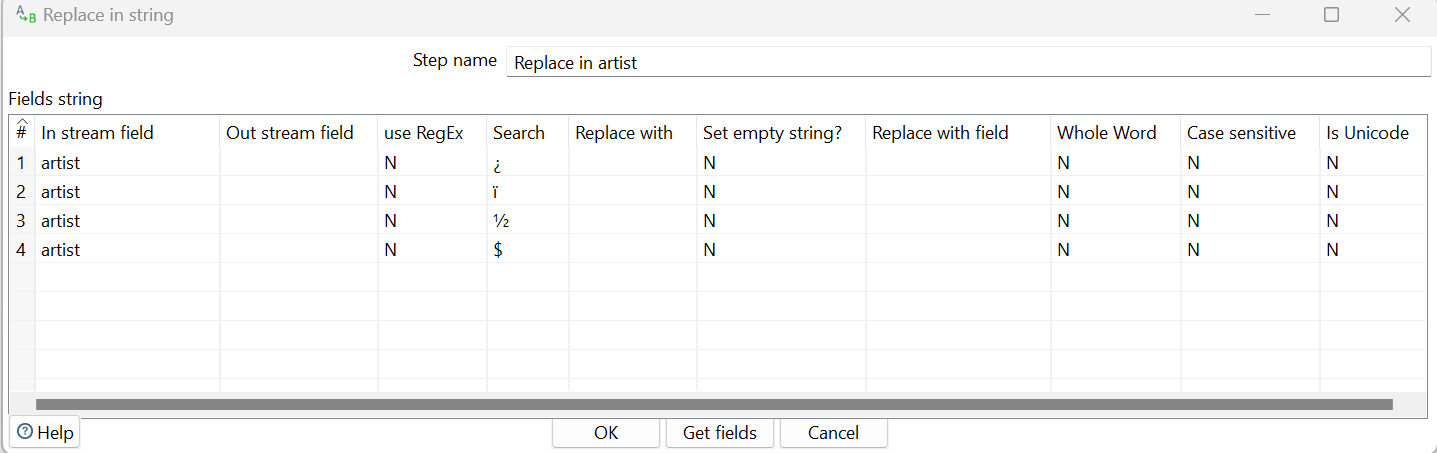
In this, I did in two cases on by sort by song name and another by sort by sporify\_charts

Case-1:

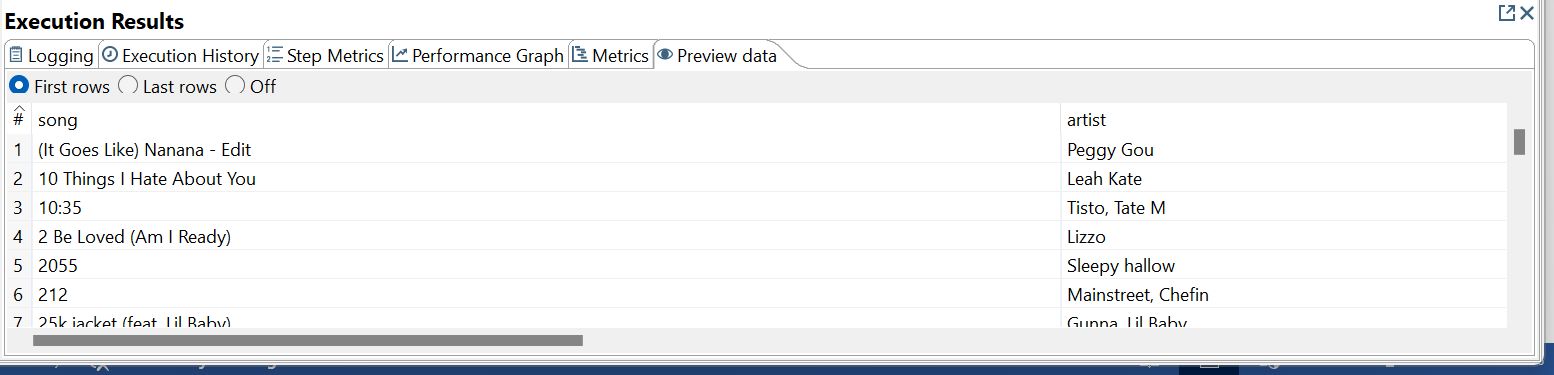
* I taken input from csv file spotify\_top\_songs , and sorted by song name
* Then, later I make unique rows by “artist” in ignoring the cases
* Replacing some characters(suppose in line 3 in “artist”) like that it contains in many rows
*  Before

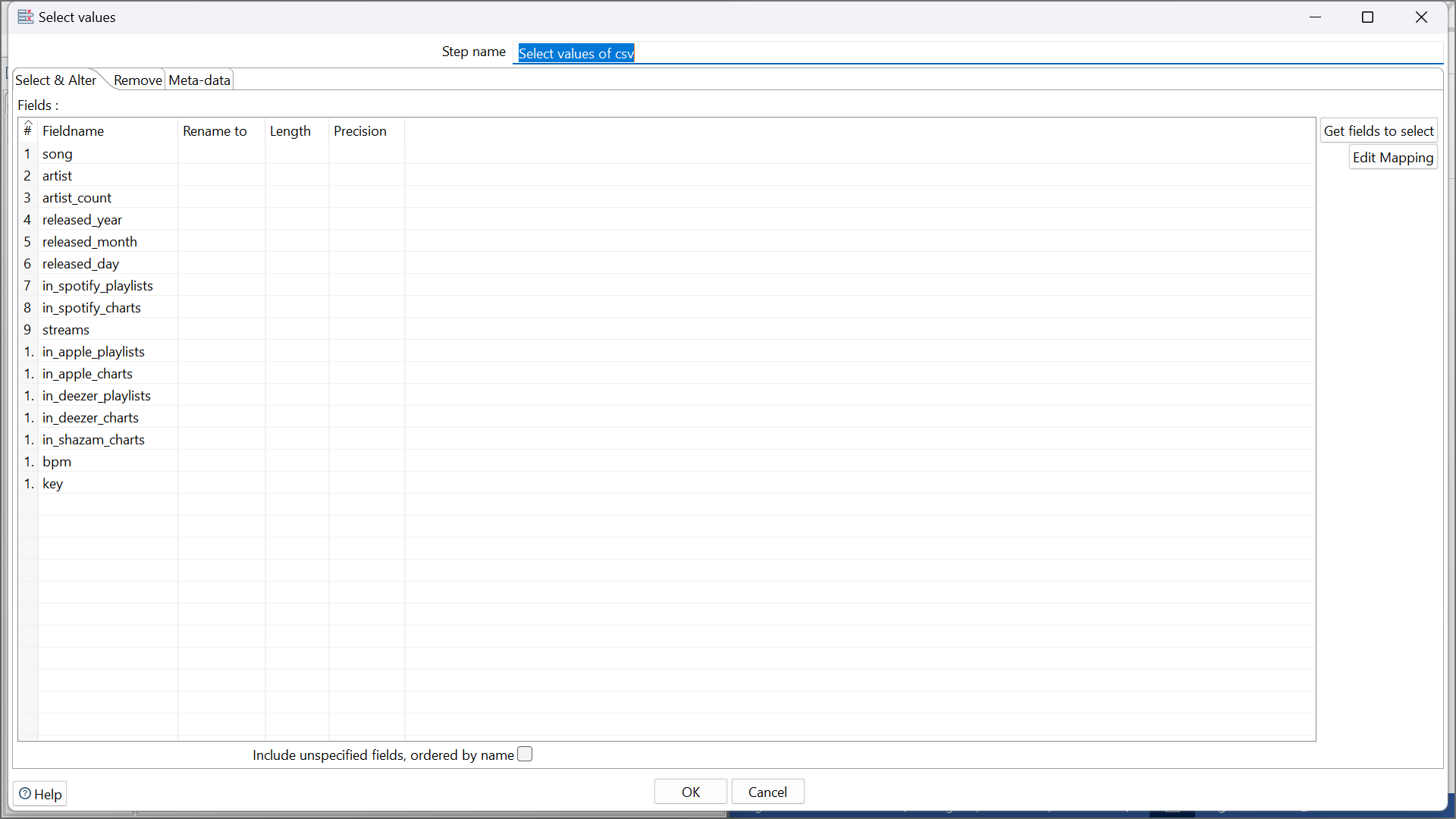


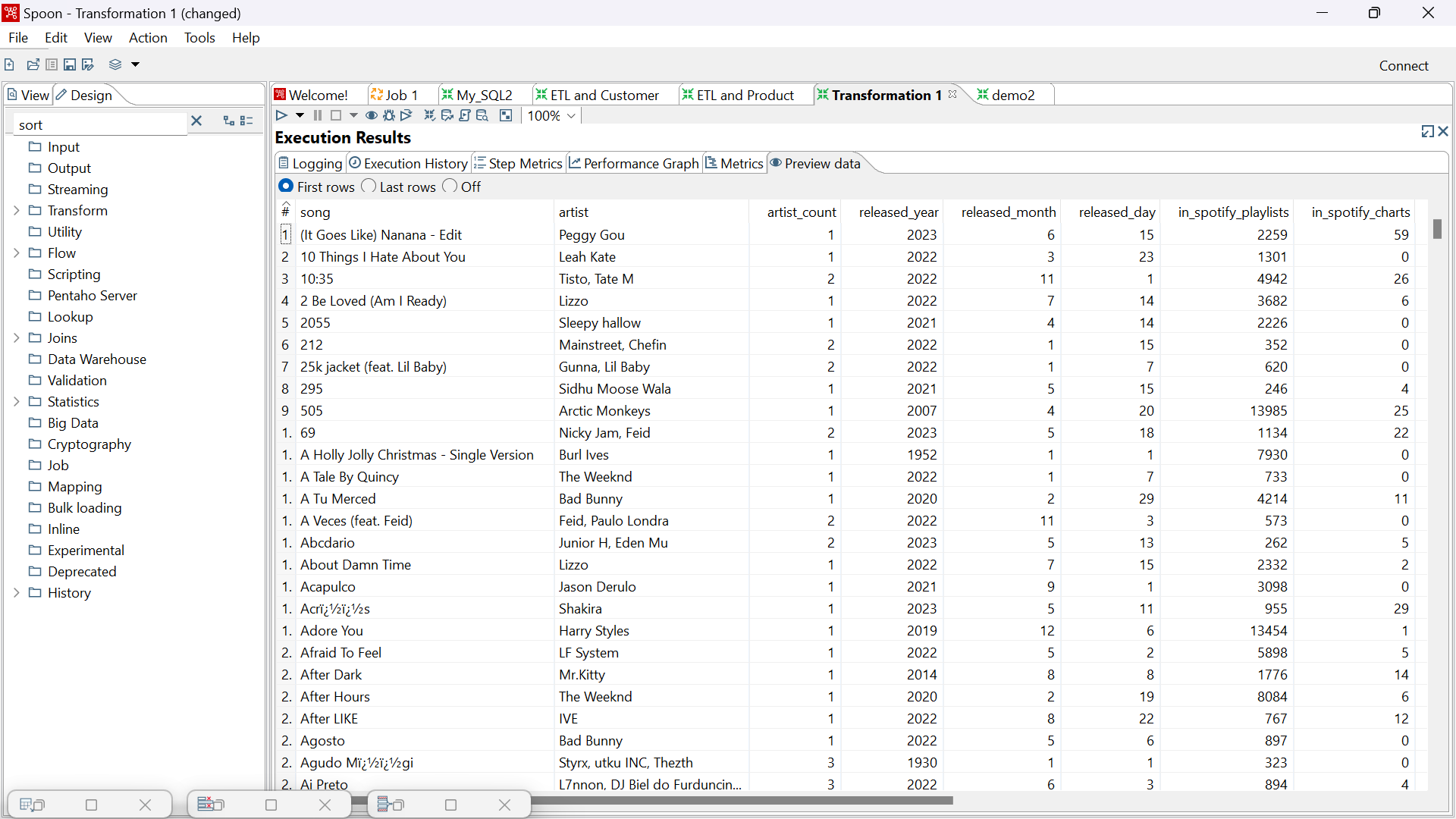
* Using this



* It become

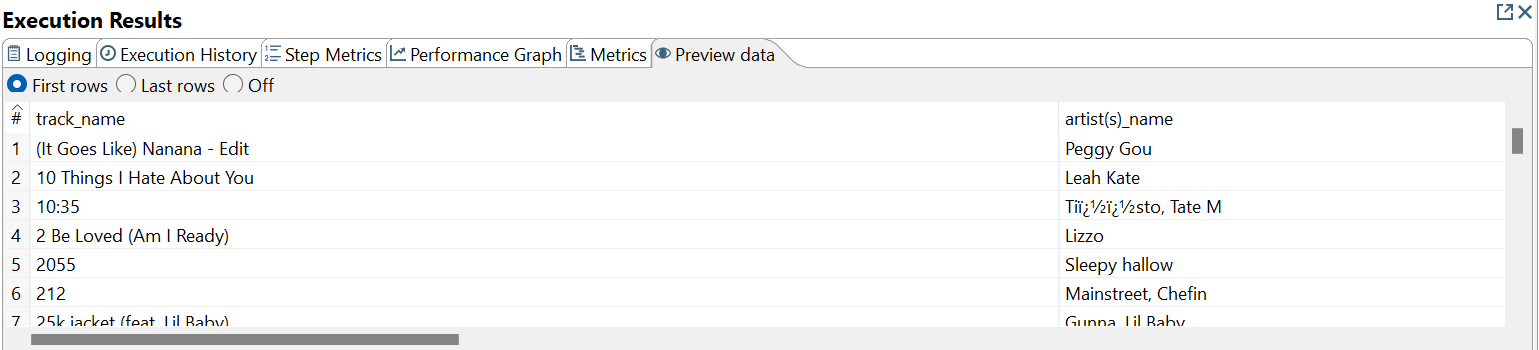


* I removed last rows and I want up “key”
* I got output as

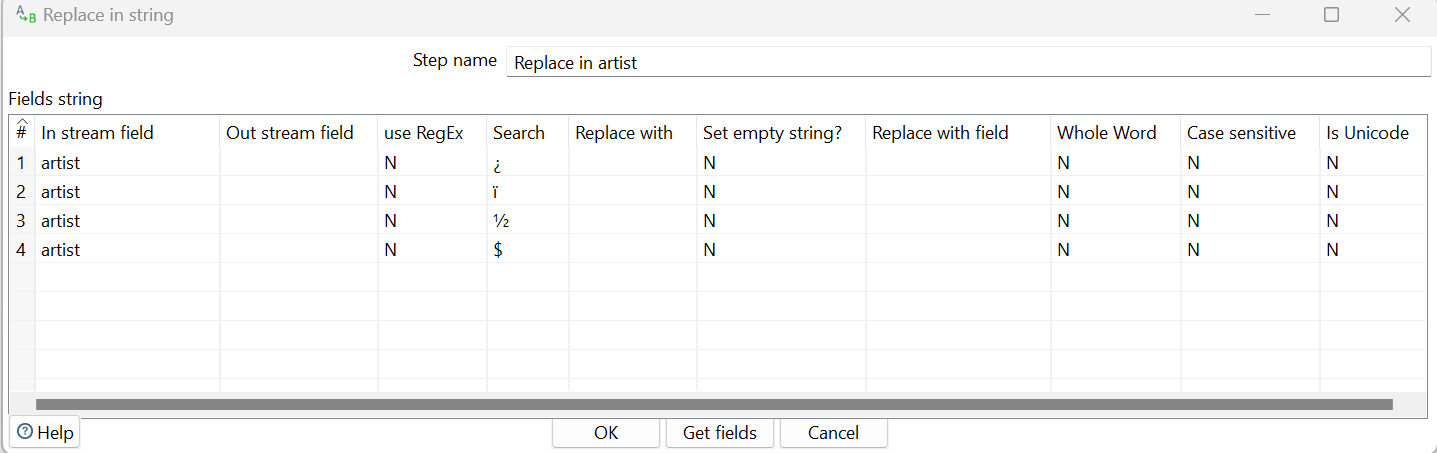


Case-2:

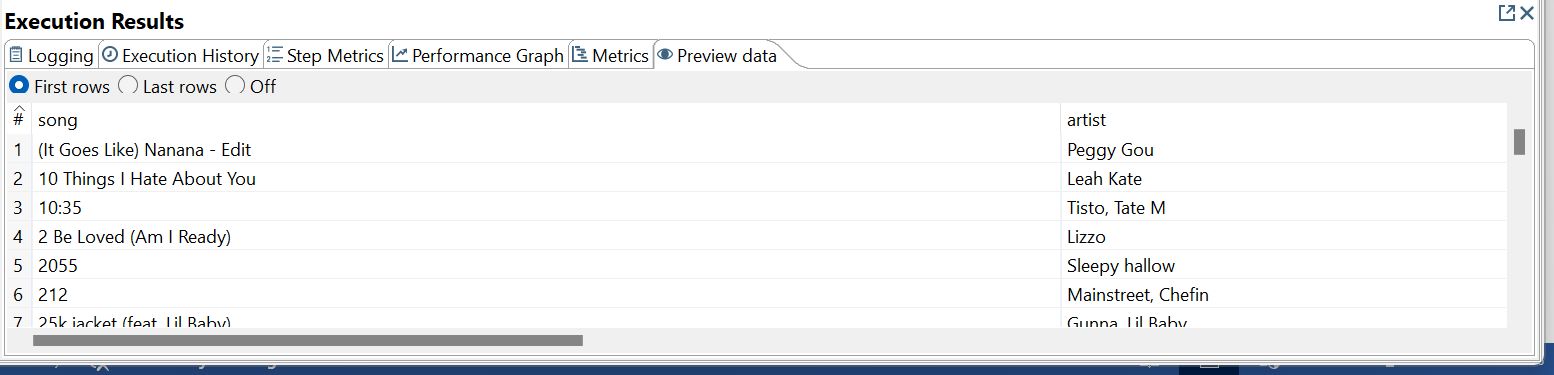
* I taken input from csv file spotify\_top\_songs , and sorted by song name
* Then, later I make unique rows by “artist” in ignoring the cases
* Replacing some characters(suppose in line 3 in “artist”) like that it contains in many rows
* Before



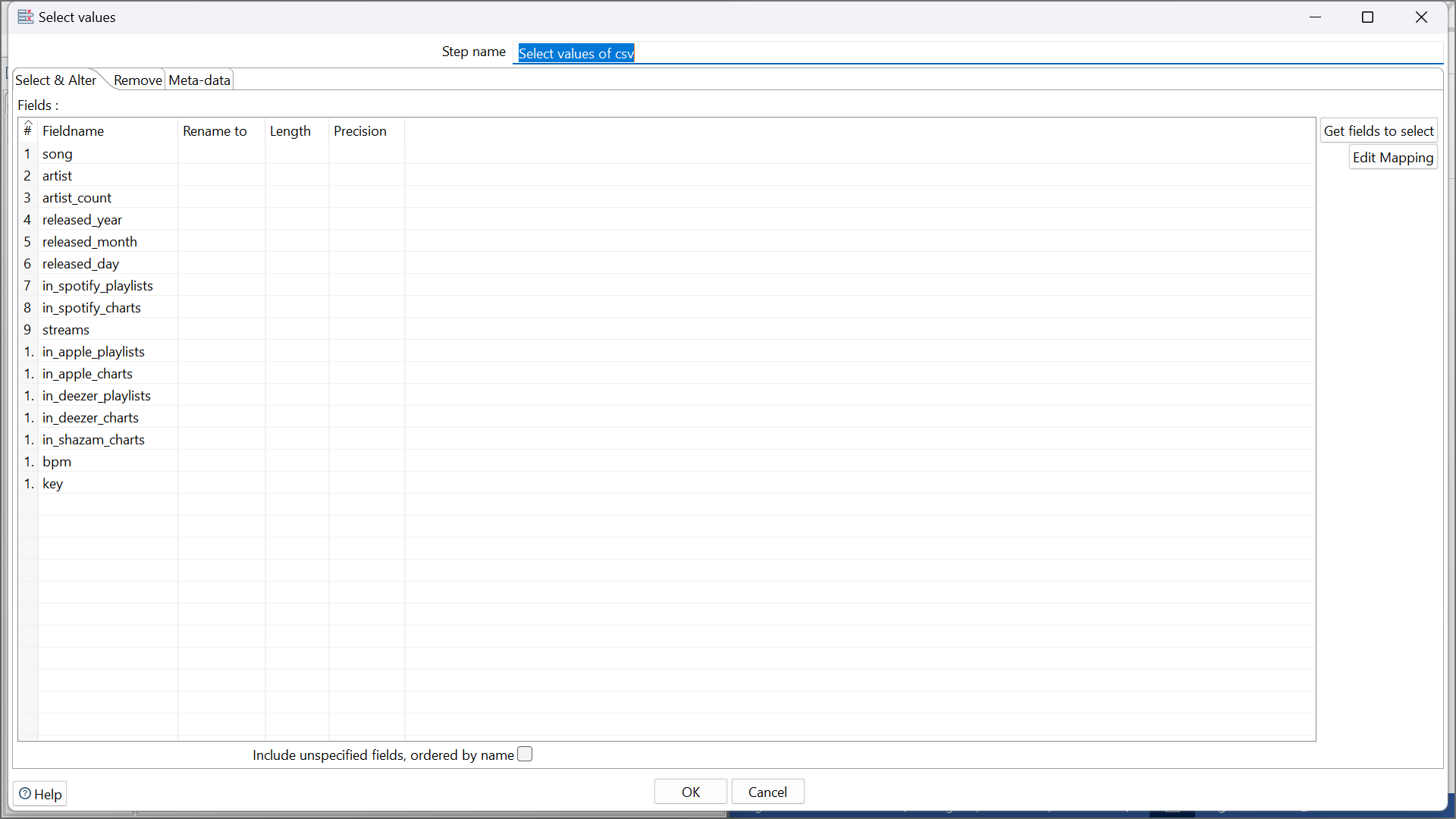
* Using this



* It become

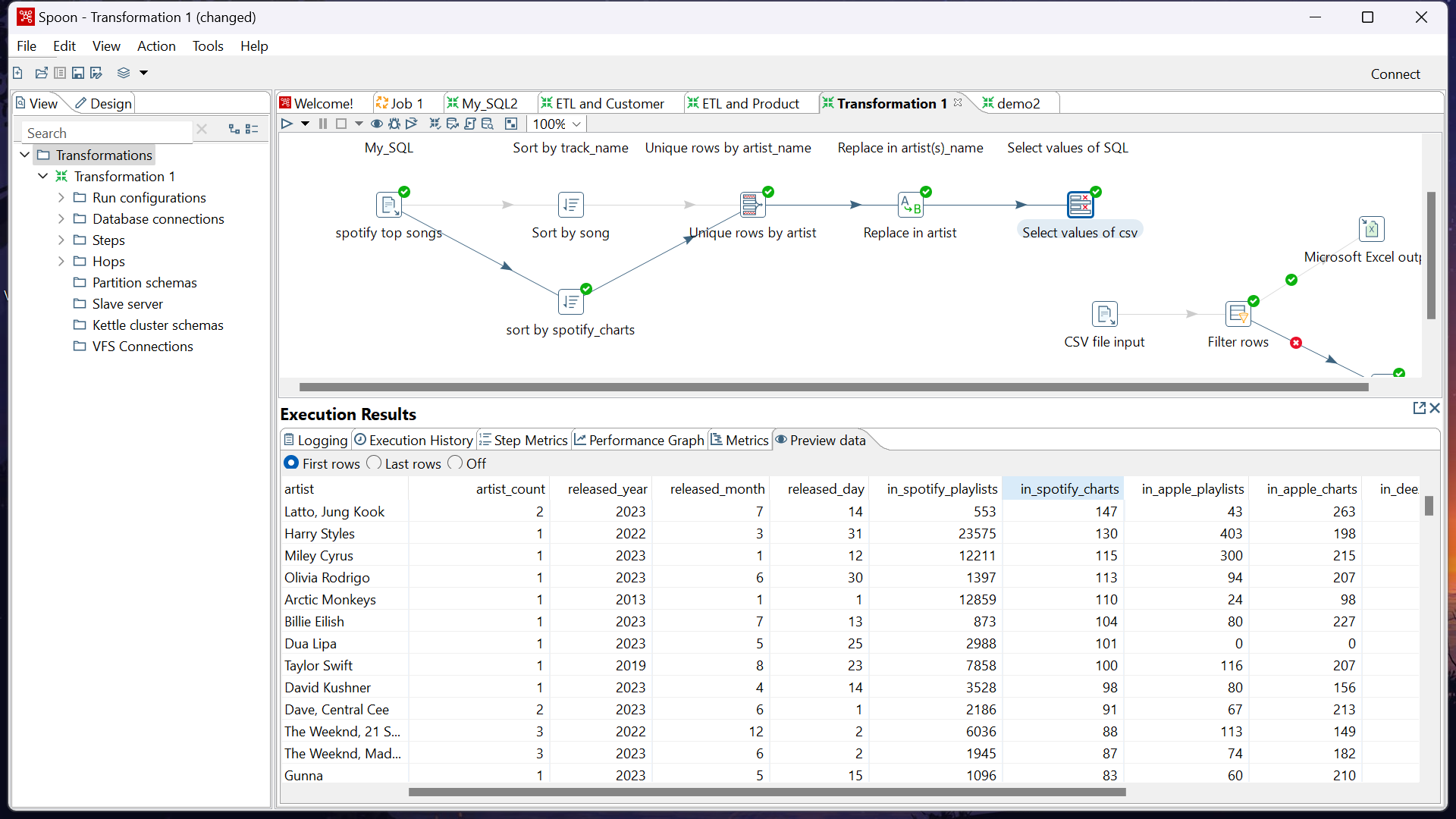


* I removed last rows and I want up “key”

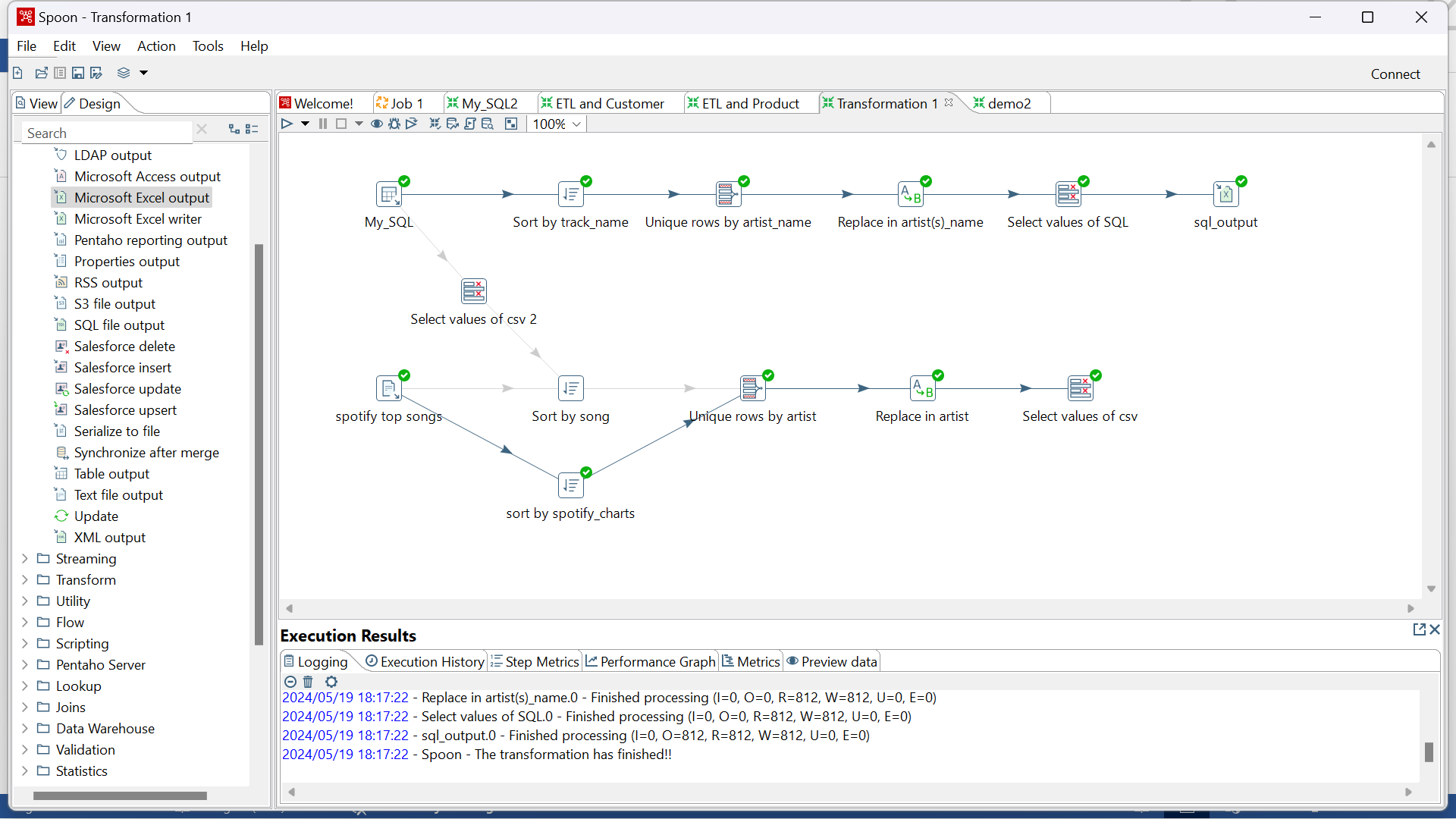


* I got output as some different than the other

It is order ‘spotify\_ charts’

****

**My Trasformation looks like this**

****