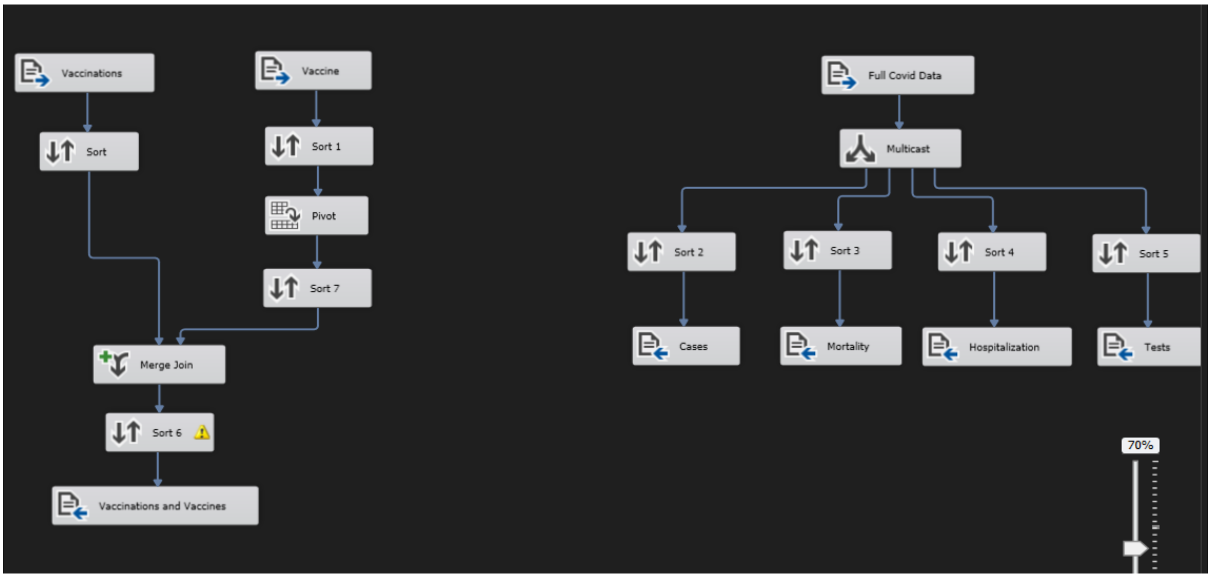
Preprocessing Stage

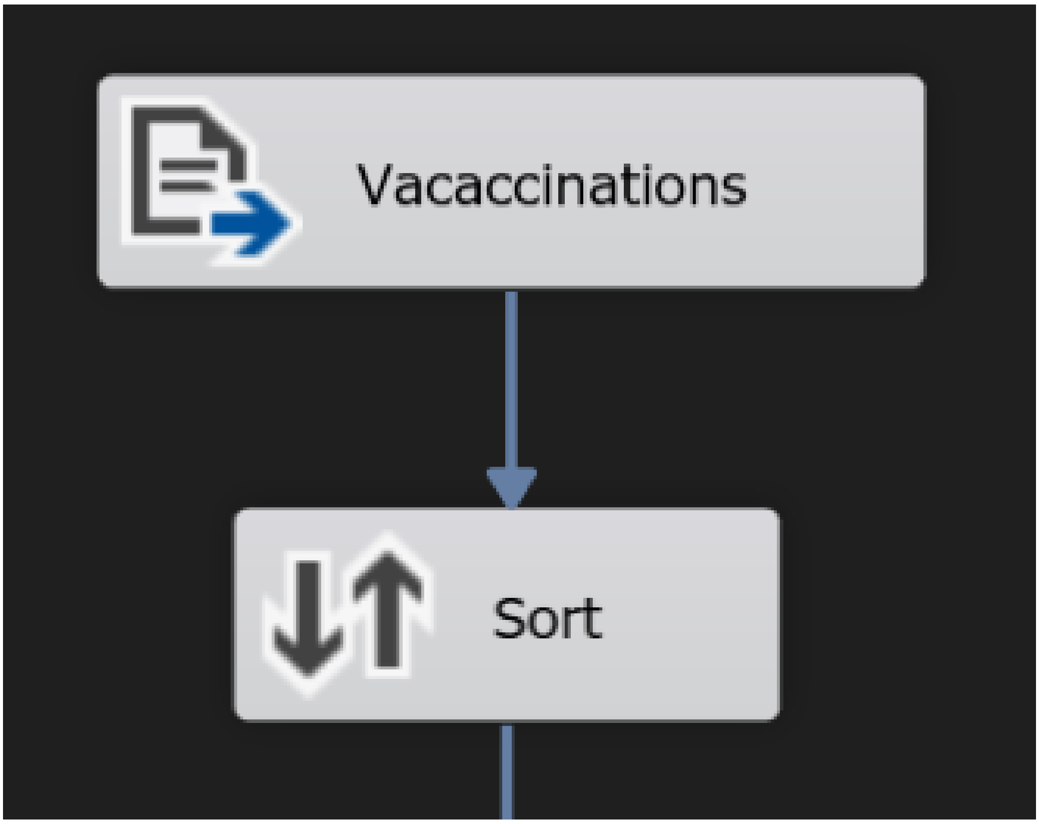
Aim of the process

The full preprocessing stage is done to obtain files ready for data cleansing done with the scripts written in python. The whole data flow process at this stage is shown below:

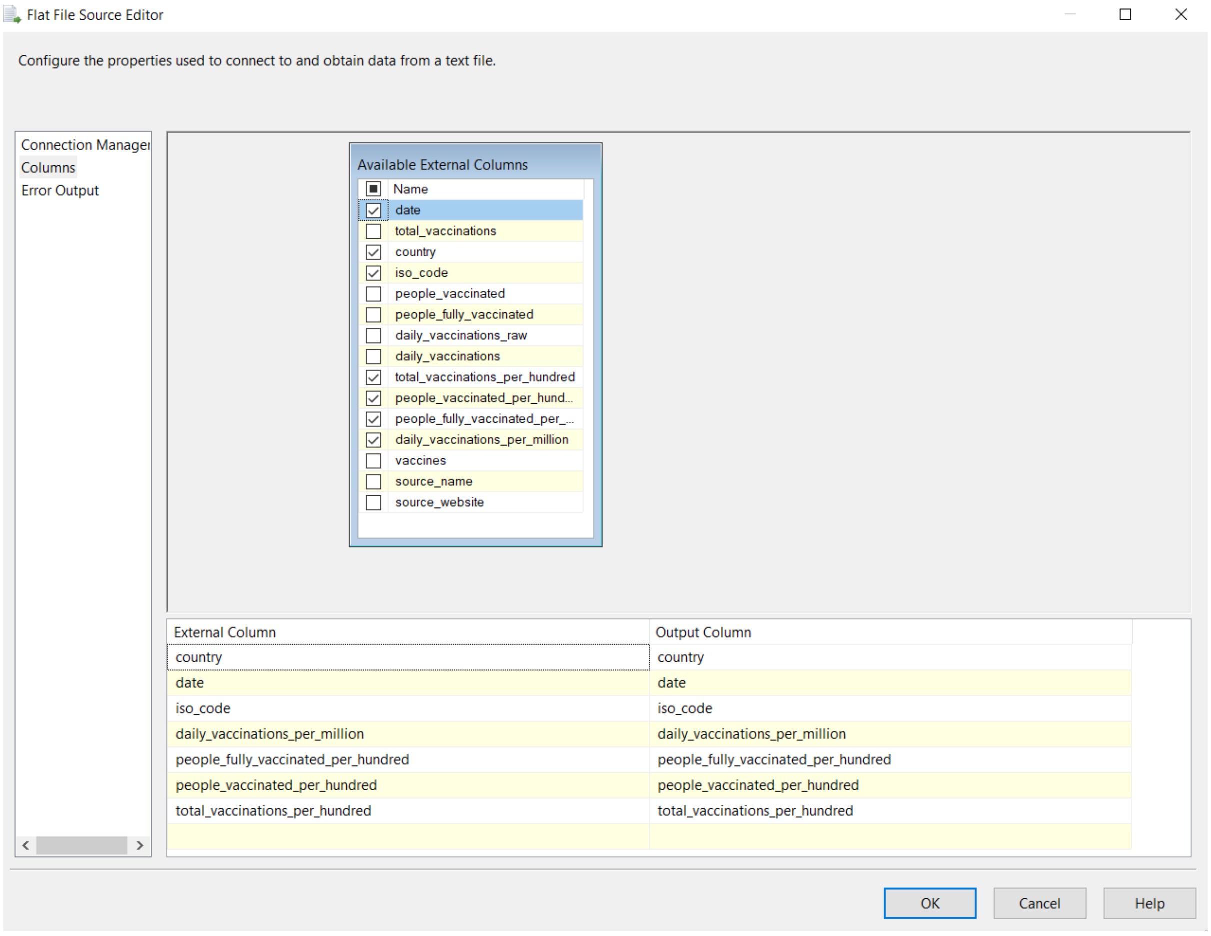


Defining Data Sources

Defining 'Vaccinations' Flat File Source (country_vaccinations.csv) and then Sorting the data:



Selecting the appropriate columns:



Flat File Source Editor

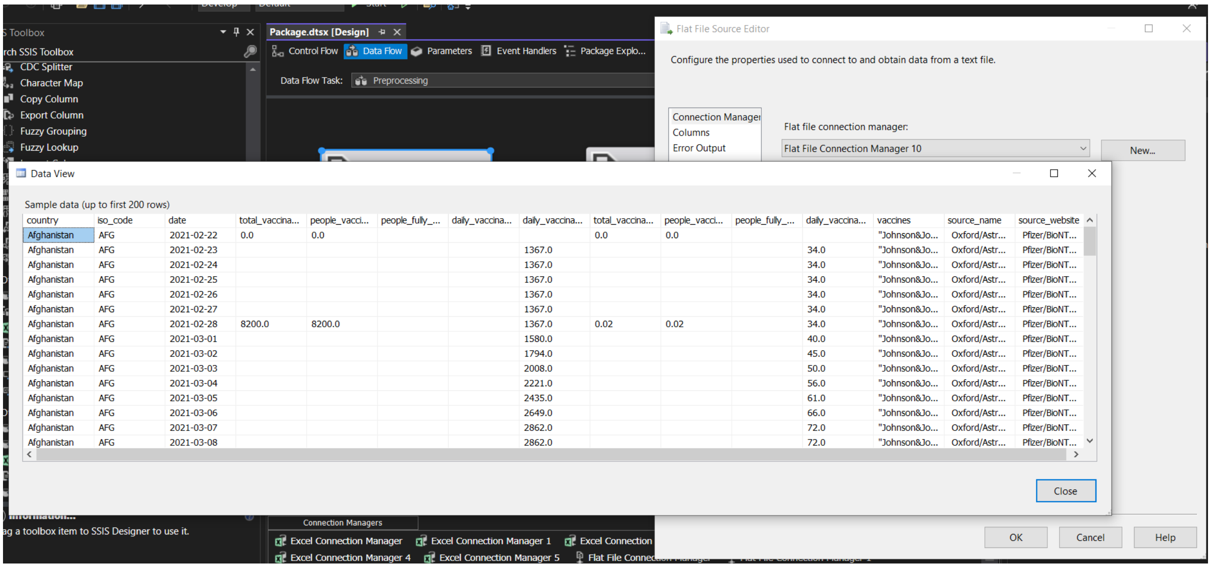
Configure the properties used to connect to and obtain data from a text file.

Available External Columns

- ■ Name
- ☑ date
- ☐ total_vaccinations
- ☑ country
- ☑ iso_code
- ☐ people_vaccinated
- ☐ people_fully_vaccinated
- ☐ daily_vaccinations_raw
- ☐ daily_vaccinations
- ☑ total_vaccinations_per_hundred
- ☑ people_vaccinated_per_hund...
- ☑ people_fully_vaccinated_per_...
- ☑ daily_vaccinations_per_million
- ☐ vaccines
- ☐ source_name
- ☐ source_website

External Column	Output Column
country	country
date	date
iso_code	iso_code
daily_vaccinations_per_million	daily_vaccinations_per_million
people_fully_vaccinated_per_hundred	people_fully_vaccinated_per_hundred
people_vaccinated_per_hundred	people_vaccinated_per_hundred
total_vaccinations_per_hundred	total_vaccinations_per_hundred

Data preview (as we can see, there is a lot of missing and not continuous data, which will be cleaned also by using python scripts):



Configure the properties used to connect to and obtain data from a text file.

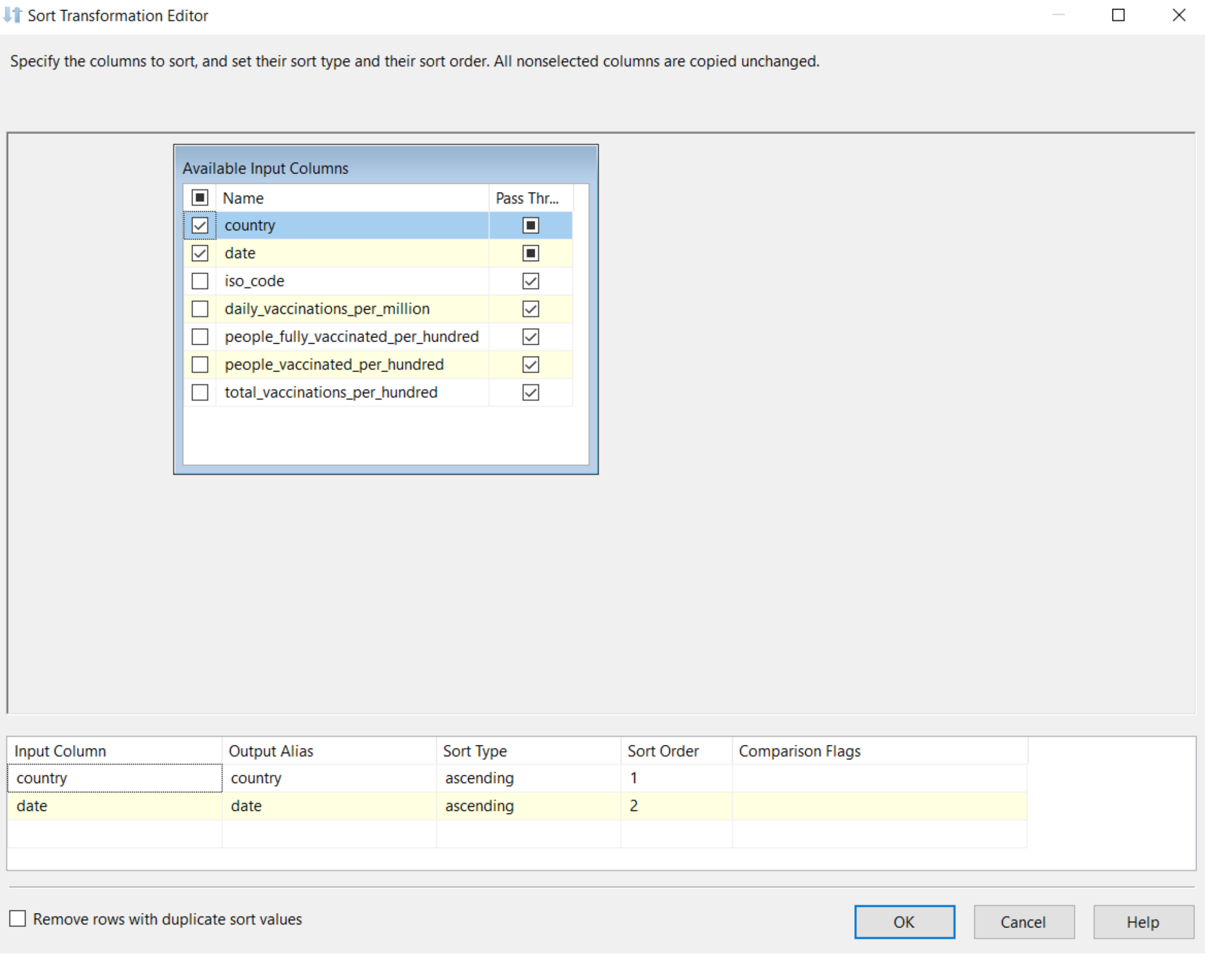
Flat file connection manager: Flat File Connection Manager 10

Data View

Sample data (up to first 200 rows)

country	iso_code	date	total_vaccina...	people_vacci...	people_fully_...	daily_vaccina...	daily_vaccina...	total_vaccina...	people_vacci...	people_fully_...	daily_vaccina...	vaccines	source_name	source_website
Afghanistan	AFG	2021-02-22	0.0	0.0				0.0	0.0			"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...
Afghanistan	AFG	2021-02-23					1367.0				34.0	"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...
Afghanistan	AFG	2021-02-24					1367.0				34.0	"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...
Afghanistan	AFG	2021-02-25					1367.0				34.0	"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...
Afghanistan	AFG	2021-02-26					1367.0				34.0	"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...
Afghanistan	AFG	2021-02-27					1367.0				34.0	"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...
Afghanistan	AFG	2021-02-28	8200.0	8200.0			1367.0	0.02	0.02		34.0	"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...
Afghanistan	AFG	2021-03-01					1580.0				40.0	"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...
Afghanistan	AFG	2021-03-02					1794.0				45.0	"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...
Afghanistan	AFG	2021-03-03					2008.0				50.0	"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...
Afghanistan	AFG	2021-03-04					2221.0				56.0	"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...
Afghanistan	AFG	2021-03-05					2435.0				61.0	"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...
Afghanistan	AFG	2021-03-06					2649.0				66.0	"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...
Afghanistan	AFG	2021-03-07					2862.0				72.0	"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...
Afghanistan	AFG	2021-03-08					2862.0				72.0	"Johnson&Jo...	Oxford/Astr...	Pfizer/BioNT...

The data are sorted based on the ‘country’ and ‘date’ columns:



Specify the columns to sort, and set their sort type and their sort order. All nonselected columns are copied unchanged.

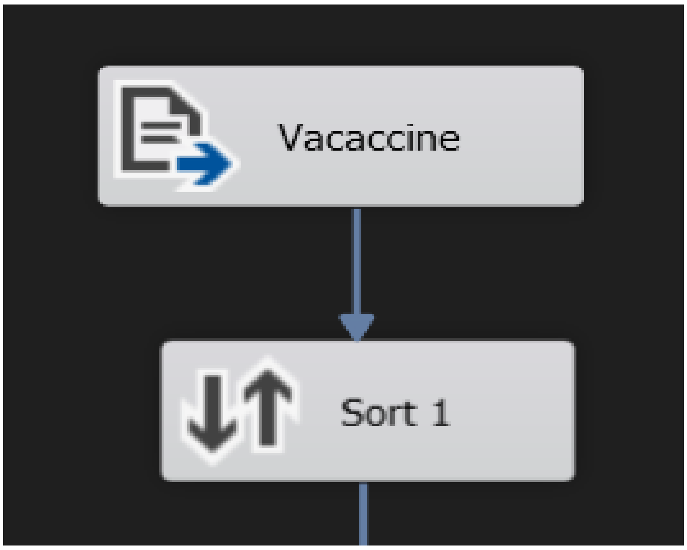
Available Input Columns

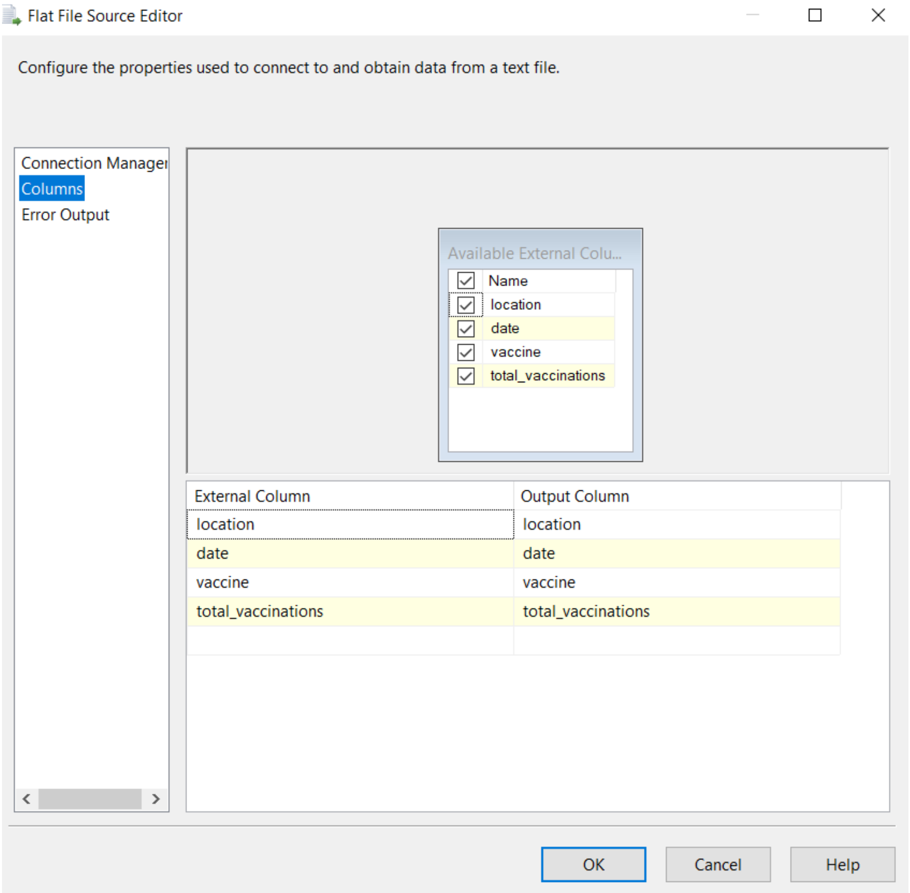
■	Name	Pass Thr...
☑	country	■
☑	date	■
□	iso_code	☑
□	daily_vaccinations_per_million	☑
□	people_fully_vaccinated_per_hundred	☑
□	people_vaccinated_per_hundred	☑
□	total_vaccinations_per_hundred	☑

Input Column	Output Alias	Sort Type	Sort Order	Comparison Flags
country	country	ascending	1	
date	date	ascending	2	

□ Remove rows with duplicate sort values

The same process is repeated for the ‘Vaccine’ (country_vaccinations_by_manufacturer.csv) and ‘Full Covid Data’ (cowid-covid-data-full.csv) Flat File Sources:



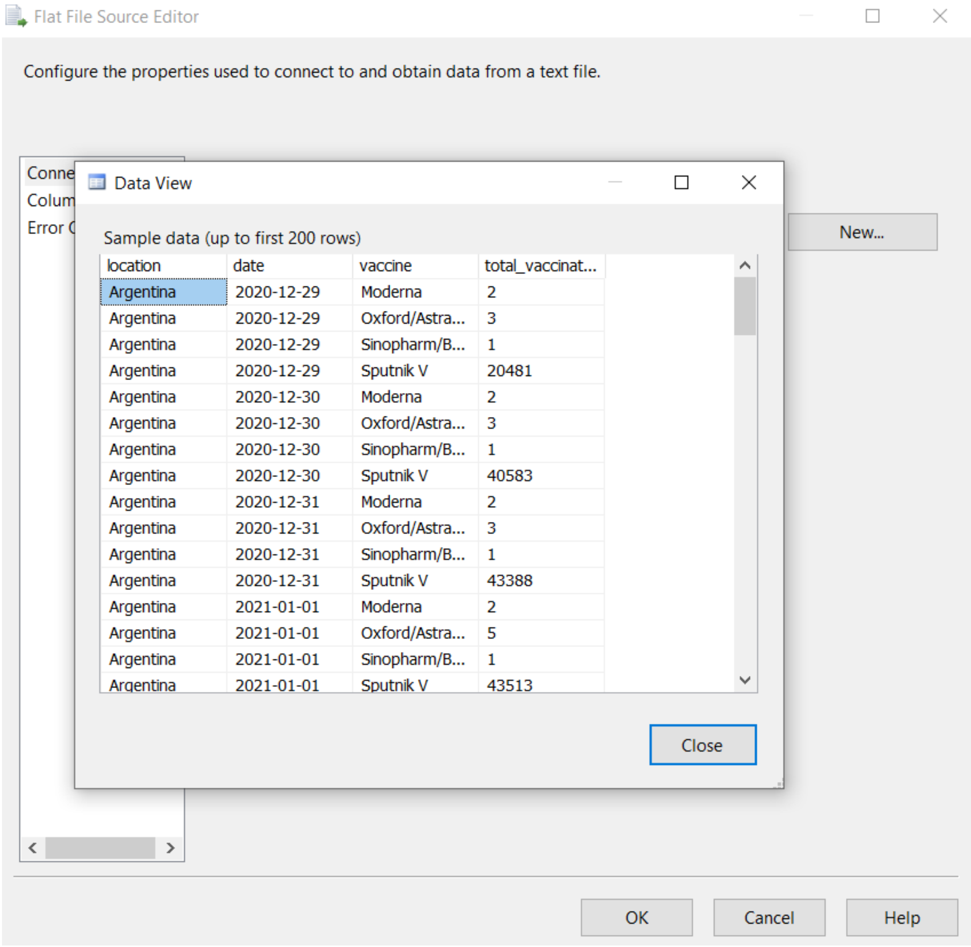


Configure the properties used to connect to and obtain data from a text file.

Available External Colu...

- ☑ Name
- ☑ location
- ☑ date
- ☑ vaccine
- ☑ total_vaccinations

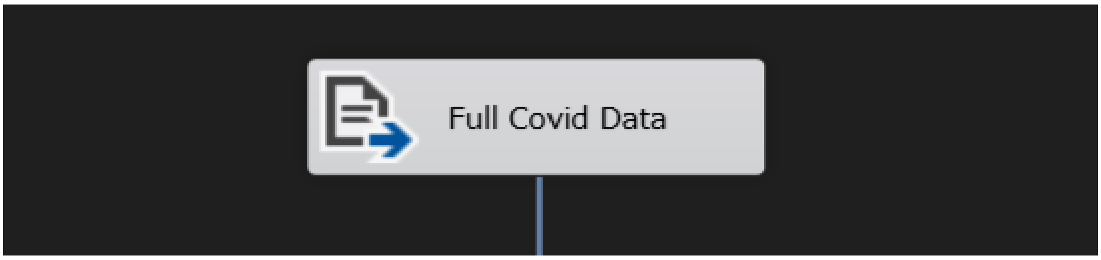
External Column	Output Column
location	location
date	date
vaccine	vaccine
total_vaccinations	total_vaccinations

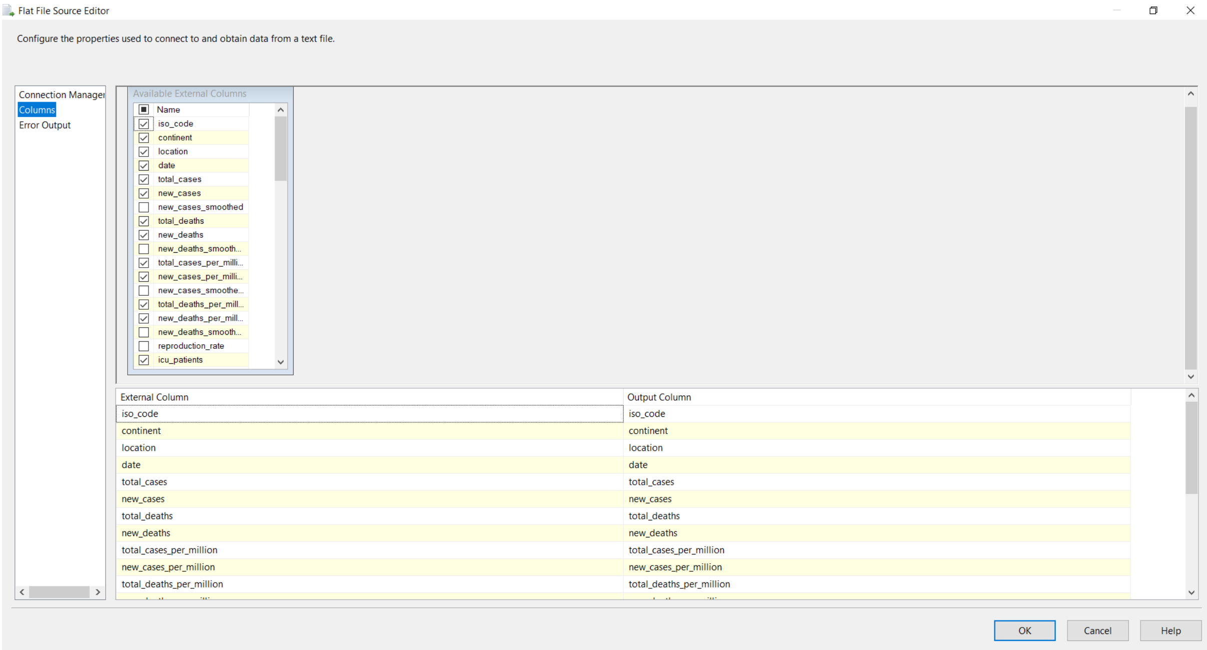


Configure the properties used to connect to and obtain data from a text file.

Sample data (up to first 200 rows)

location	date	vaccine	total_vaccinat...
Argentina	2020-12-29	Moderna	2
Argentina	2020-12-29	Oxford/Astra...	3
Argentina	2020-12-29	Sinopharm/B...	1
Argentina	2020-12-29	Sputnik V	20481
Argentina	2020-12-30	Moderna	2
Argentina	2020-12-30	Oxford/Astra...	3
Argentina	2020-12-30	Sinopharm/B...	1
Argentina	2020-12-30	Sputnik V	40583
Argentina	2020-12-31	Moderna	2
Argentina	2020-12-31	Oxford/Astra...	3
Argentina	2020-12-31	Sinopharm/B...	1
Argentina	2020-12-31	Sputnik V	43388
Argentina	2021-01-01	Moderna	2
Argentina	2021-01-01	Oxford/Astra...	5
Argentina	2021-01-01	Sinopharm/B...	1
Argentina	2021-01-01	Sputnik V	43513

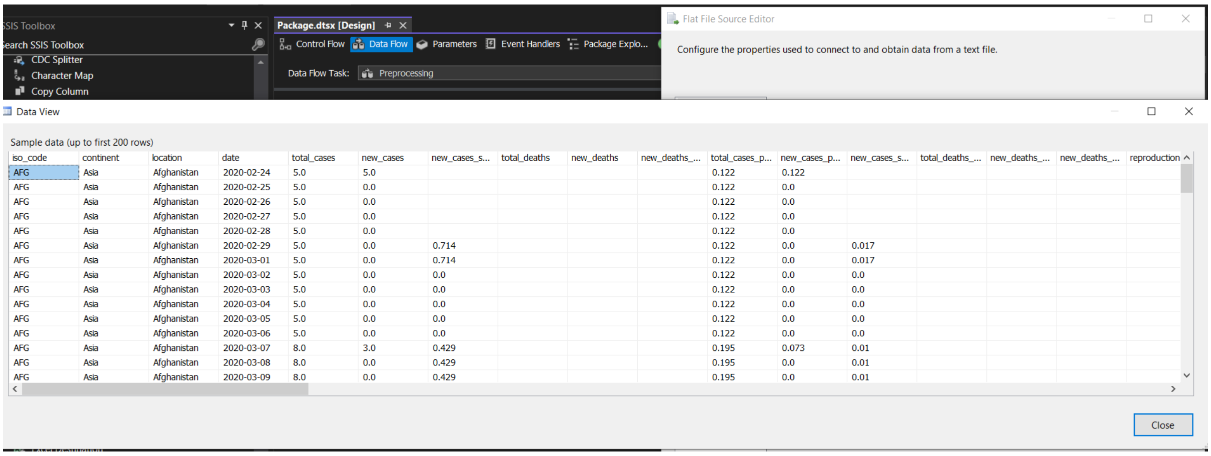




Configure the properties used to connect to and obtain data from a text file.

■	Name
☑	iso_code
☑	continent
☑	location
☑	date
☑	total_cases
☑	new_cases
□	new_cases_smoothed
☑	total_deaths
☑	new_deaths
□	new_deaths_smooth...
☑	total_cases_per_milli...
☑	new_cases_per_milli...
□	new_cases_smoothe...
☑	total_deaths_per_mill...
☑	new_deaths_per_mill...
□	new_deaths_smooth...
□	reproduction_rate
☑	icu_patients

External Column	Output Column
iso_code	iso_code
continent	continent
location	location
date	date
total_cases	total_cases
new_cases	new_cases
total_deaths	total_deaths
new_deaths	new_deaths
total_cases_per_million	total_cases_per_million
new_cases_per_million	new_cases_per_million
total_deaths_per_million	total_deaths_per_million



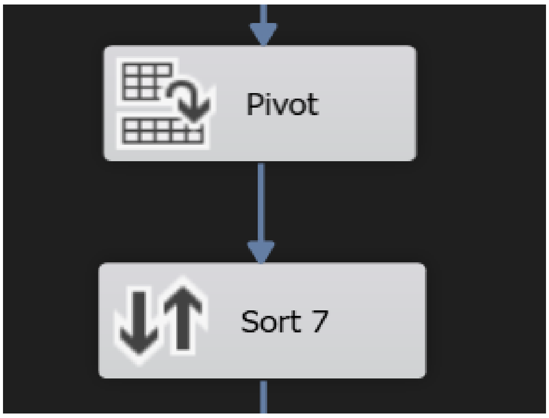
Configure the properties used to connect to and obtain data from a text file.

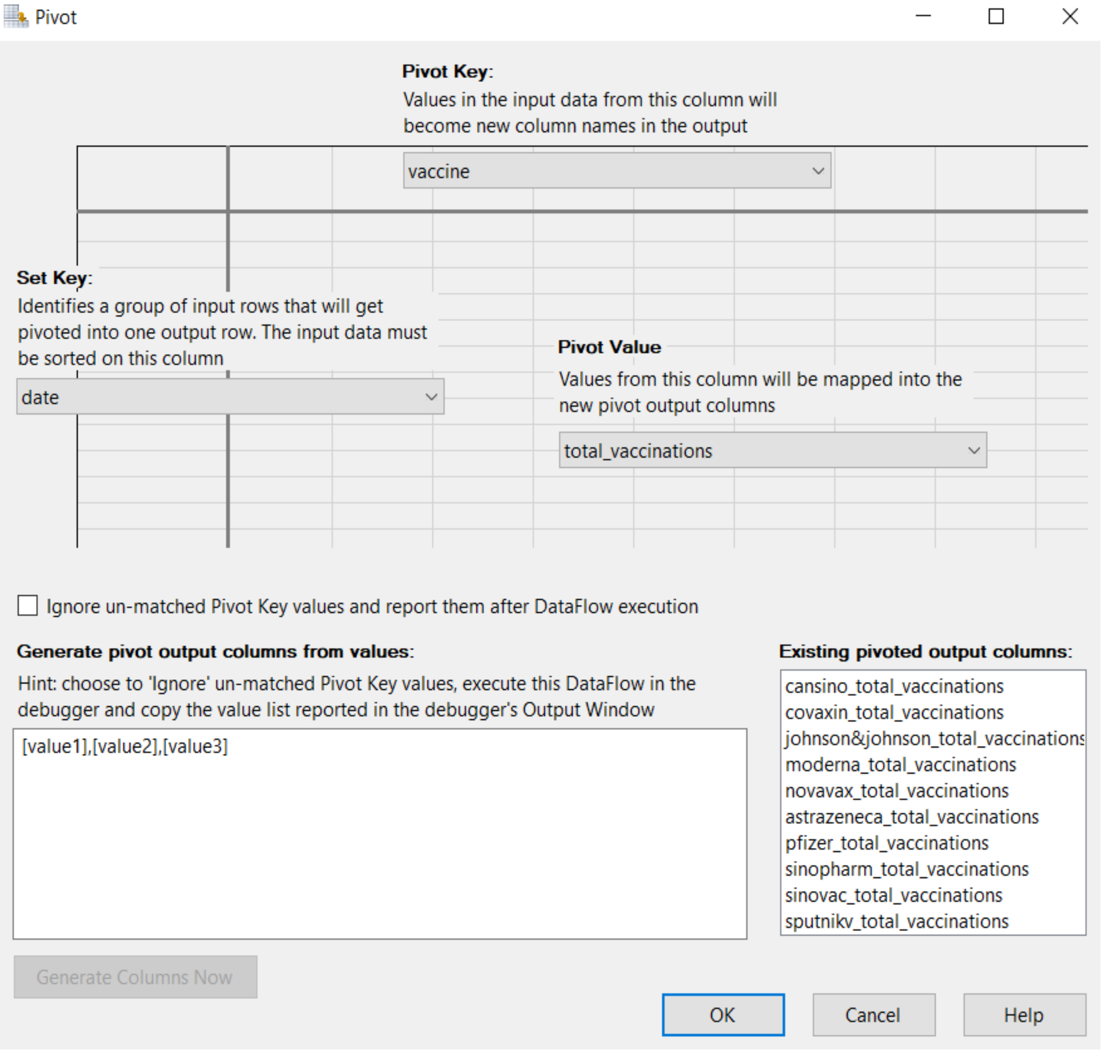
Sample data (up to first 200 rows)

iso_code	continent	location	date	total_cases	new_cases	new_cases_s...	total_deaths	new_deaths	new_deaths_...	total_cases_p...	new_cases_p...	new_cases_s...	total_deaths_...	new_deaths_...	new_deaths_...	reproduction
AFG	Asia	Afghanistan	2020-02-24	5.0	5.0					0.122	0.122					
AFG	Asia	Afghanistan	2020-02-25	5.0	0.0					0.122	0.0					
AFG	Asia	Afghanistan	2020-02-26	5.0	0.0					0.122	0.0					
AFG	Asia	Afghanistan	2020-02-27	5.0	0.0					0.122	0.0					
AFG	Asia	Afghanistan	2020-02-28	5.0	0.0					0.122	0.0					
AFG	Asia	Afghanistan	2020-02-29	5.0	0.0	0.714				0.122	0.0	0.017				
AFG	Asia	Afghanistan	2020-03-01	5.0	0.0	0.714				0.122	0.0	0.017				
AFG	Asia	Afghanistan	2020-03-02	5.0	0.0	0.0				0.122	0.0	0.0				
AFG	Asia	Afghanistan	2020-03-03	5.0	0.0	0.0				0.122	0.0	0.0				
AFG	Asia	Afghanistan	2020-03-04	5.0	0.0	0.0				0.122	0.0	0.0				
AFG	Asia	Afghanistan	2020-03-05	5.0	0.0	0.0				0.122	0.0	0.0				
AFG	Asia	Afghanistan	2020-03-06	5.0	0.0	0.0				0.122	0.0	0.0				
AFG	Asia	Afghanistan	2020-03-07	8.0	3.0	0.429				0.195	0.073	0.01				
AFG	Asia	Afghanistan	2020-03-08	8.0	0.0	0.429				0.195	0.0	0.01				
AFG	Asia	Afghanistan	2020-03-09	8.0	0.0	0.429				0.195	0.0	0.01				

Pivot

In the ‘Vaccine’ data values from the ‘vaccine’ column are transferred to separate columns and then sorted:





Pivot Key:
Values in the input data from this column will become new column names in the output

vaccine

Set Key:
Identifies a group of input rows that will get pivoted into one output row. The input data must be sorted on this column

date

Pivot Value
Values from this column will be mapped into the new pivot output columns

total_vaccinations

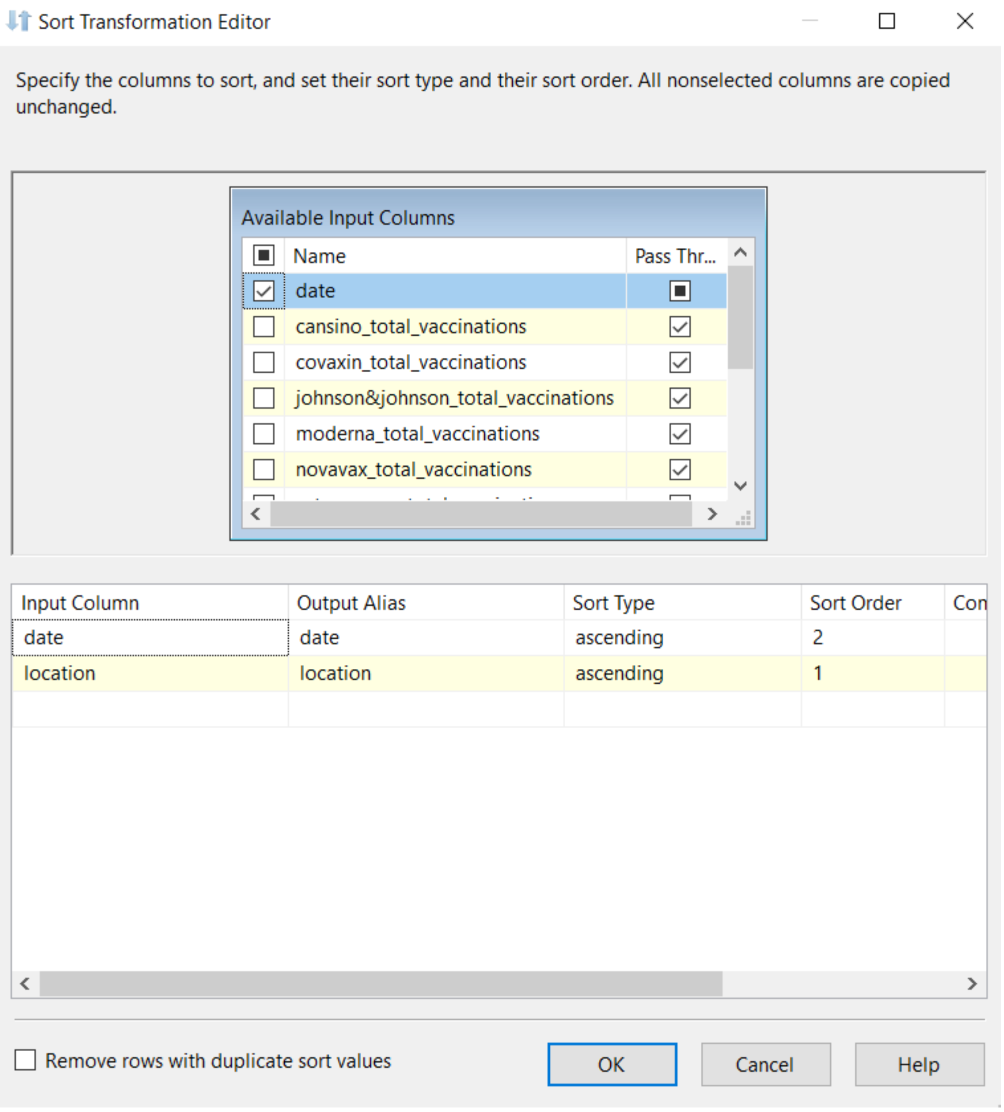
□ Ignore un-matched Pivot Key values and report them after DataFlow execution

Generate pivot output columns from values:
Hint: choose to 'Ignore' un-matched Pivot Key values, execute this DataFlow in the debugger and copy the value list reported in the debugger's Output Window

[value1],[value2],[value3]

Existing pivoted output columns:

cansino_total_vaccinations
covaxin_total_vaccinations
johnson&johnson_total_vaccinations
moderna_total_vaccinations
novavax_total_vaccinations
astrazeneca_total_vaccinations
pfizer_total_vaccinations
sinopharm_total_vaccinations
sinovac_total_vaccinations
sputnikv_total_vaccinations



Sort Transformation Editor

Specify the columns to sort, and set their sort type and their sort order. All nonselected columns are copied unchanged.

Available Input Columns

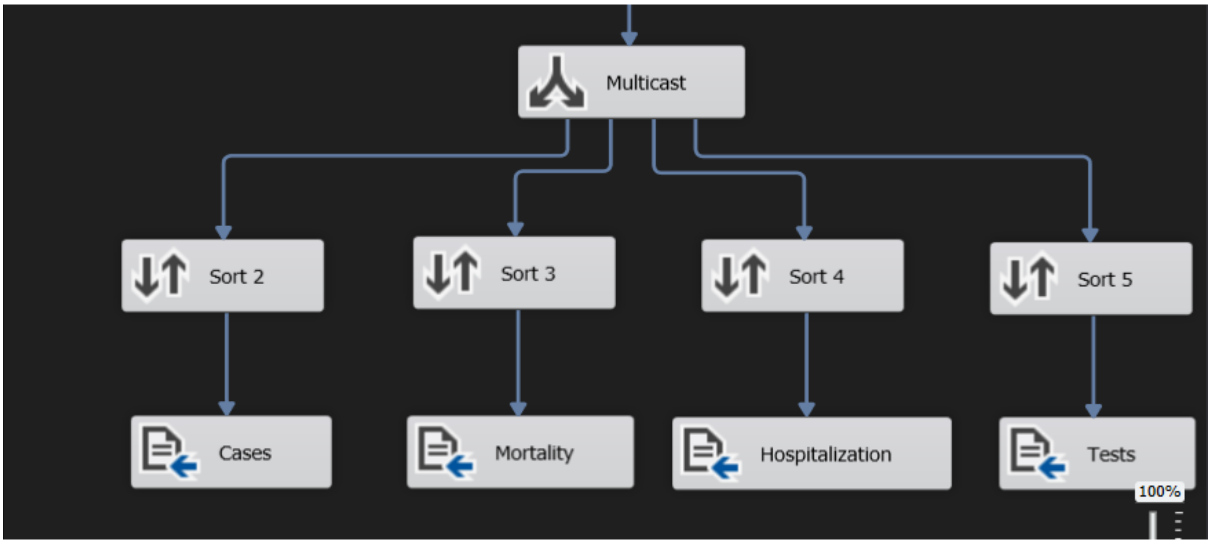
■	Name	Pass Thr...
☑	date	■
□	cansino_total_vaccinations	☑
□	covaxin_total_vaccinations	☑
□	johnson&johnson_total_vaccinations	☑
□	moderna_total_vaccinations	☑
□	novavax_total_vaccinations	☑

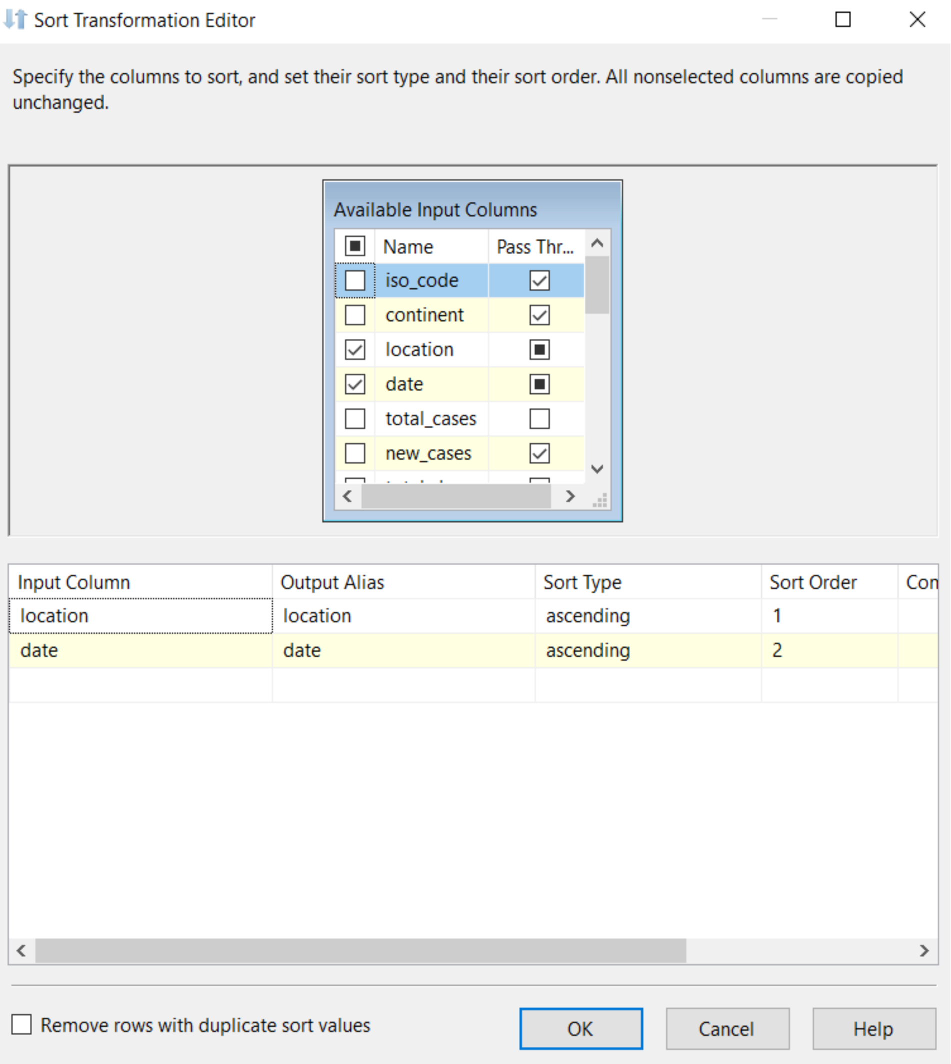
Input Column	Output Alias	Sort Type	Sort Order	Con
date	date	ascending	2	
location	location	ascending	1	

□ Remove rows with duplicate sort values

Multicast

Data from the ‘Full Covid Data’ are separated, sorted (location and date) and then sent to separate files, which will be processed by python scripts:





Specify the columns to sort, and set their sort type and their sort order. All nonselected columns are copied unchanged.

Available Input Columns

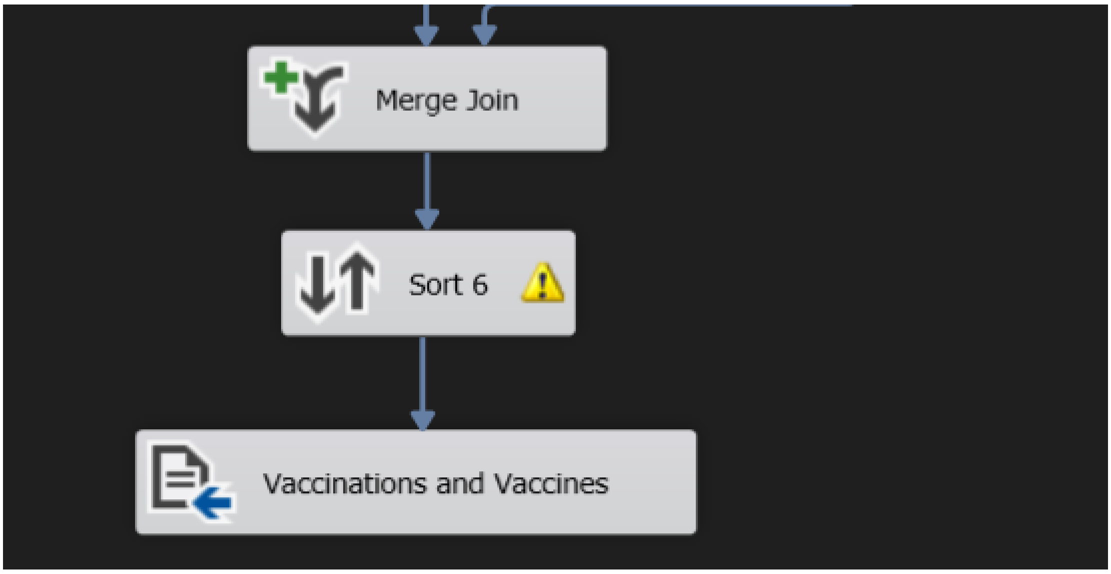
■	Name	Pass Thr...
□	iso_code	☑
□	continent	☑
☑	location	■
☑	date	■
□	total_cases	□
□	new_cases	☑

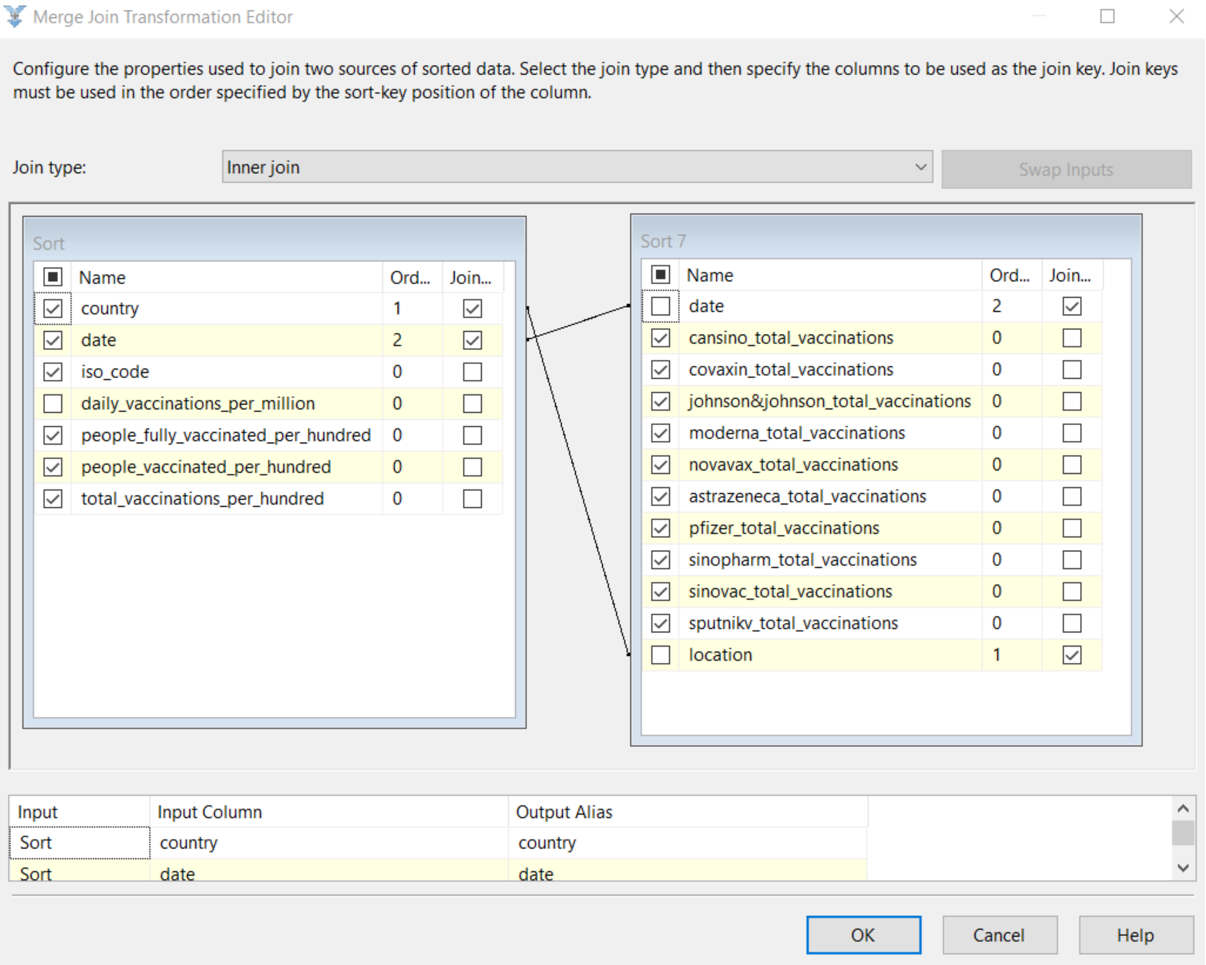
Input Column	Output Alias	Sort Type	Sort Order	Con
location	location	ascending	1	
date	date	ascending	2	

□ Remove rows with duplicate sort values

Merge Join

Data from the ‘Vaccinations’ and ‘Vaccine’ are merged using an inner join, then sorted and transferred to separate file:





Configure the properties used to join two sources of sorted data. Select the join type and then specify the columns to be used as the join key. Join keys must be used in the order specified by the sort-key position of the column.

Join type: Inner join

Sort

■	Name	Ord...	Join...
☑	country	1	☑
☑	date	2	☑
☑	iso_code	0	☐
☐	daily_vaccinations_per_million	0	☐
☑	people_fully_vaccinated_per_hundred	0	☐
☑	people_vaccinated_per_hundred	0	☐
☑	total_vaccinations_per_hundred	0	☐

Sort 7

■	Name	Ord...	Join...
☐	date	2	☑
☑	cansino_total_vaccinations	0	☐
☑	covaxin_total_vaccinations	0	☐
☑	johnson&johnson_total_vaccinations	0	☐
☑	moderna_total_vaccinations	0	☐
☑	novavax_total_vaccinations	0	☐
☑	astrazeneca_total_vaccinations	0	☐
☑	pfizer_total_vaccinations	0	☐
☑	sinopharm_total_vaccinations	0	☐
☑	sinovac_total_vaccinations	0	☐
☑	sputnikv_total_vaccinations	0	☐
☐	location	1	☑

Input	Input Column	Output Alias
Sort	country	country
Sort	date	date