VERSION 1

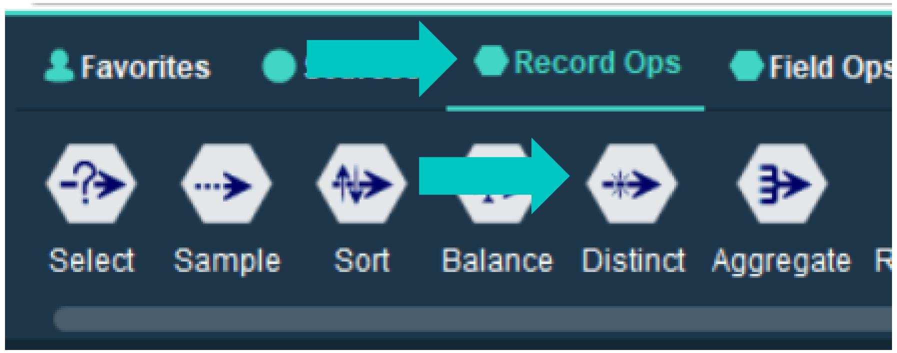
SPSS ANALYTICS PARTNER

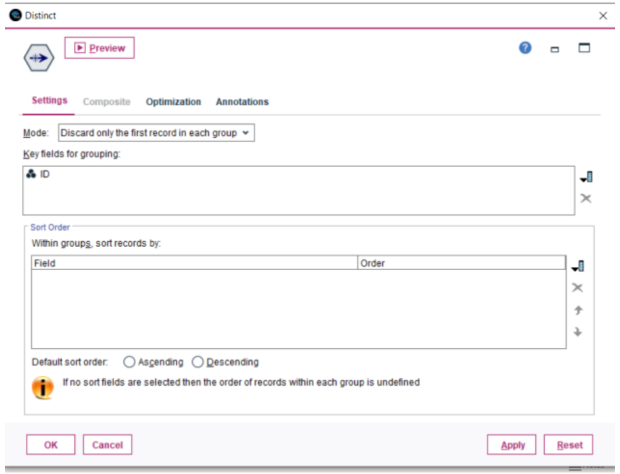
Find Duplicate Records

Tech Tips – IBM SPSS Modeler

Tech Tips – Find Duplicate Records

- Here’s a quick tip to find duplicate records in IBM SPSS Modeler.
- The **Distinct** node makes identifying duplicates quick and easy and is located on the **Record Ops** palette.





Mode: Discard only the first record in each group

Key fields for grouping:

ID

Sort Order

Within groups, sort records by:

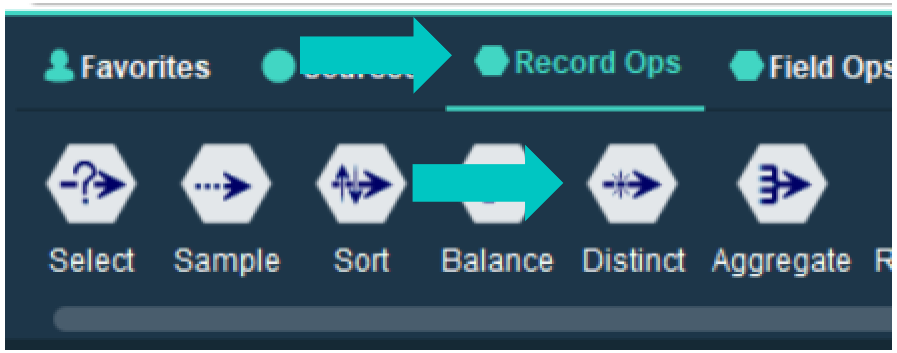
Field	Order

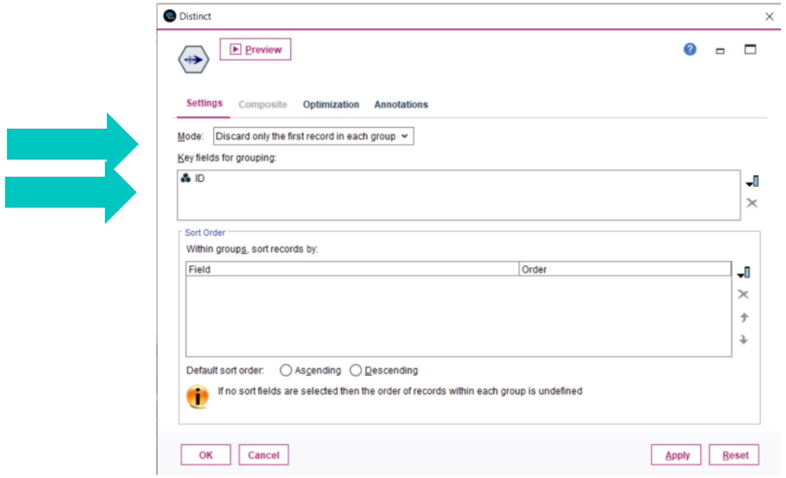
Default sort order: ○ Ascending ○ Descending

If no sort fields are selected then the order of records within each group is undefined

Tech Tips – Find Duplicate Records

- To identify duplicates go to the **Record Ops** palette. Select the **Distinct** node and drag it onto the stream canvas. You can also double click the node to drop it onto the stream canvas. Once it is on the canvas you can connect it to your stream.
- Double click to open the node. We want to identify records that are duplicates so that we can discard them. We will set the **Mode** to **Discard only the first record in each group**. Use the field chooser button to select the field. Here we are looking for to identify duplicate records for customers.





Mode: Discard only the first record in each group

Key fields for grouping:
ID

Sort Order
Within groups, sort records by:

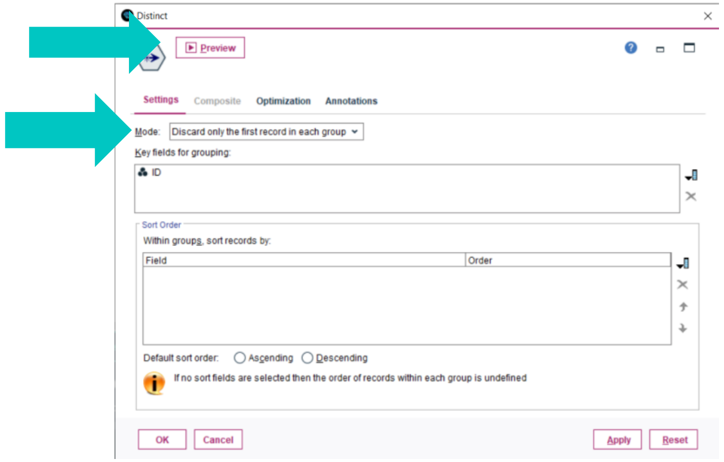
Field	Order

Default sort order: ○ Ascending ○ Descending

If no sort fields are selected then the order of records within each group is undefined

Tech Tips – Find Duplicate Records

- On the Distinct node, the **Include** and **Discard** options control whether the first distinct record is passed (**Include**) or all but the first distinct record are passed (**Discard**).
- To remove duplicates in a database set the **Mode** to **Include only the first record in each group**; to identify duplicates set the **Mode** to **Discard only the first record in each group**. Fields that provide the basis for identification of duplicates are selected in the **Key fields for grouping** list. To check your data, click on the **Preview** button.



Mode: Discard only the first record in each group

Key fields for grouping:
ID

Sort Order

Within groups, sort records by:

Field	Order

Default sort order: ○ Ascending ○ Descending

If no sort fields are selected then the order of records within each group is undefined

VERSION 1

SPSS ANALYTICS PARTNER

Thank You

For more information

please visit spssanalyticspartner.com