

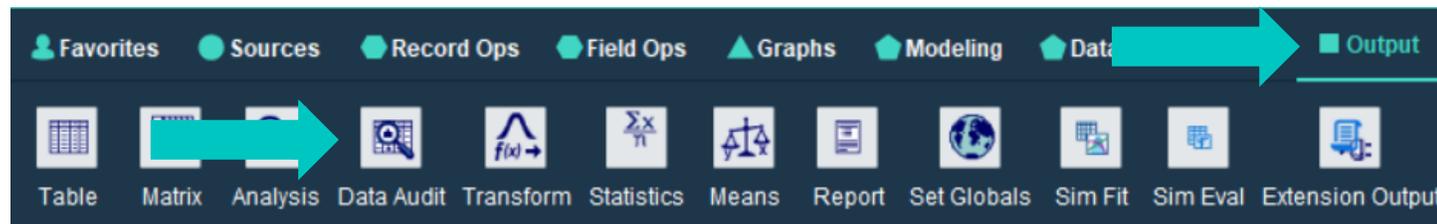


Quickly Audit Data

Tech Tips - IBM SPSS Modeler

Tech Tips – Quickly Audit Data

- Here’s a quick tip to audit your data file in IBM SPSS Modeler.
- Users can obtain a quick audit and overview of a data file by using the **Data Audit** node. The **Data Audit** node is located on the **Output** palette and provides visualizations of fields within the data file. The **Data Audit** also identifies outliers, extreme cases and missing data.



Data Audit of [127 Fields] #3

File Edit Generate

Audit Quality Annotations

Field	Sample Graph	Measurement	Min	Max	Mean	Std Dev	Skewness	Median	Mode	Unique	Valid
region		Nominal	1,000	5,000					5,000	5	5000
townsize		Ordinal	1,000	5,000					1,000	5	4998
age		Continuous	18,000	79,000	47,026	17,770	0.991	47,000	18,000		5000
agecat		Ordinal	2,000	6,000					4,000	5	5000
birthmonth		Nominal							September	12	5000
ed		Continuous	6,000	23,000	14,543	3,281	0.004	14,000	14,000		5000
edcat		Ordinal	1,000	5,000					2,000	5	5000
jobcat		Nominal	1,000	6,000					2,000	6	5000

* Indicates a multimode result * Indicates a sampled result

OK

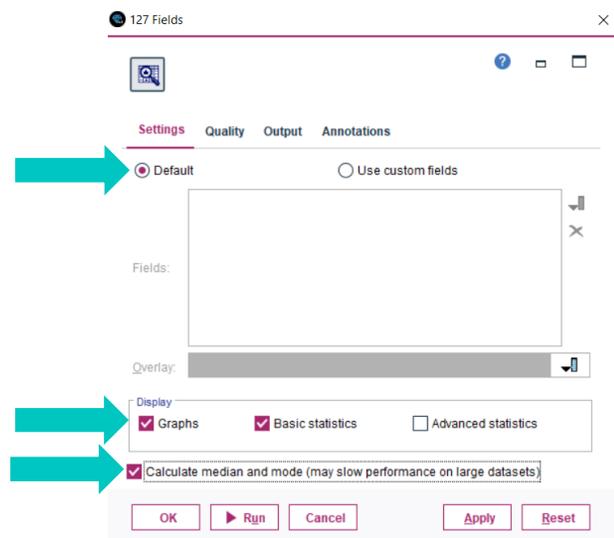
Tech Tips – Quickly Audit Data

- To audit data go to the **Output** palette. Select the **Data Audit** 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.
- When you connect the node to your stream (data) you will see the number of fields in your data file. In this case, there are 127 fields in our data file. We want to get an overview of all fields.



Tech Tips – Quickly Audit Data

- Double click to open the node. If we wanted to audit a subset we could click the **Use custom fields** button. In this case, we will select **Default** to audit all fields. We have also asked for **Graphs**, and **Basic Statistics**. There is also the option to **Calculate the median and mode**.

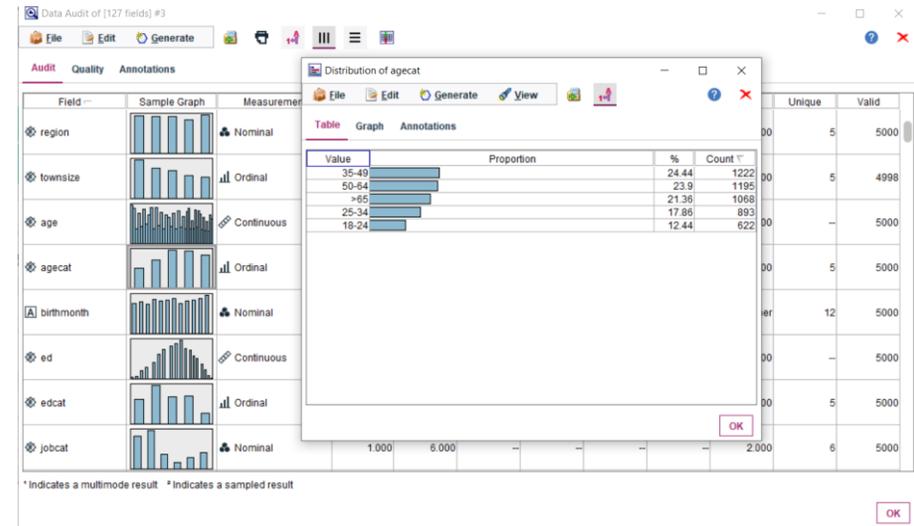
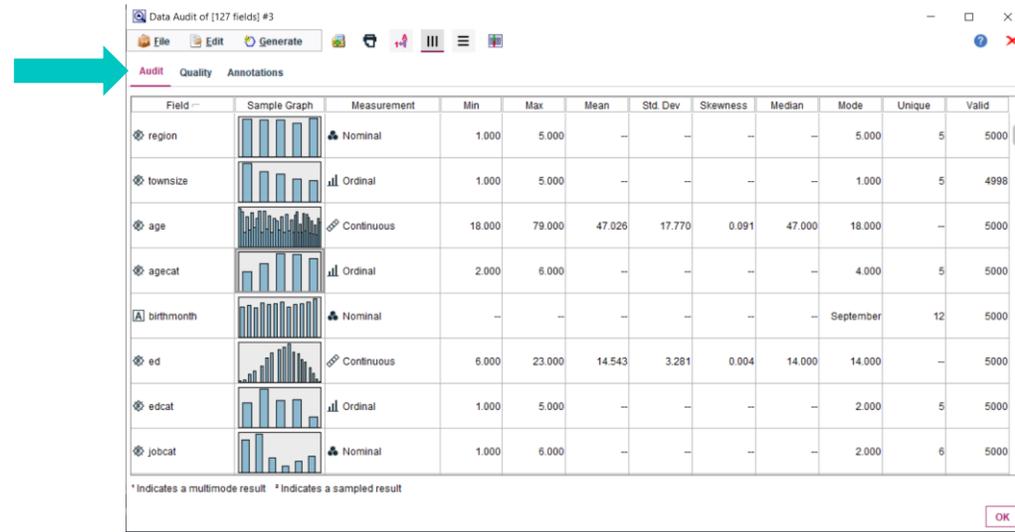


Field	Sample Graph	Measurement	Min	Max	Mean	Std Dev	Skewness	Median	Mode	Unique	Valid
region		Nominal	1.000	5.000	--	--	--	--	5.000	5	5000
townsize		Ordinal	1.000	5.000	--	--	--	--	1.000	5	4998
age		Continuous	18.000	79.000	47.026	17.770	0.091	47.000	18.000	--	5000
agecat		Ordinal	2.000	6.000	--	--	--	--	4.000	5	5000
birthmonth		Nominal	--	--	--	--	--	--	September	12	5000
ed		Continuous	6.000	23.000	14.543	3.281	0.004	14.000	14.000	--	5000
edcat		Ordinal	1.000	5.000	--	--	--	--	2.000	5	5000
jobcat		Nominal	1.000	6.000	--	--	--	--	2.000	6	5000

* Indicates a multimode result # Indicates a sampled result

Tech Tips – Quickly Audit Data

- On the **Audit** tab each field is listed with a thumbnail graph and statistics. Users can click on a thumbnail graph to see distributions or histograms.



One more tip...

- You can also look at fields by a target field you are interested in. For example, we want to build a model to predict churn. We can use the Overlay option and chose churn.
- Select Use custom fields and use Overlay to select churn.
- On the Audit tab we can see 'agecat' by "churn."

The image shows two screenshots of the SPSS Data Audit interface. The left screenshot shows the '13 Fields' dialog box with 'Use custom fields' selected and 'churn' in the 'Overlay' field. The right screenshot shows the 'Data Audit of [13 fields]' window with the 'Audit' tab active, displaying a distribution of 'agecat' by 'churn'.

Table: Distribution of agecat

Value	Proportion	%	Count
35-49	24.44	1222	
50-64	23.9	1195	
>65	21.36	1068	
25-34	17.86	893	
18-24	12.44	622	

Table: Switched providers within last month

Value	Proportion	%	Count
No	0.000	109 073	1 857
Yes	3.416	11 120	



Thank You

For more information
please visit spssanalyticspartner.com