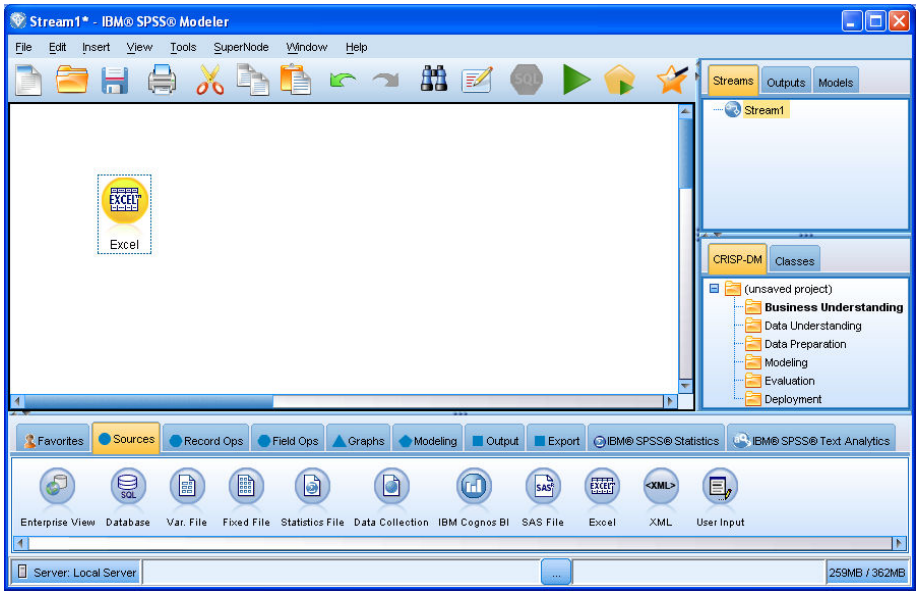
Lab Exercise One
Data Preprocessing with SPSS Modeler

Handling Missing Data

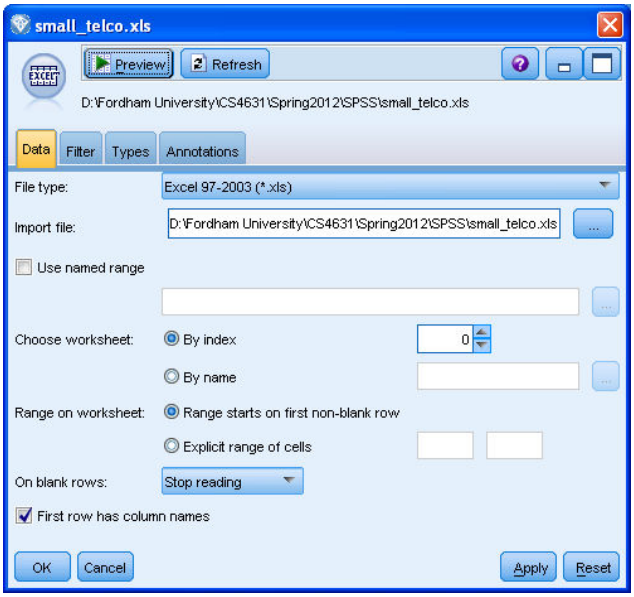
1. Download data file ***small_telco.xls*** from course website, save it on desktop or a folder of your choice.

2. Open IBM SPSS Modeler, choose Create a new stream.

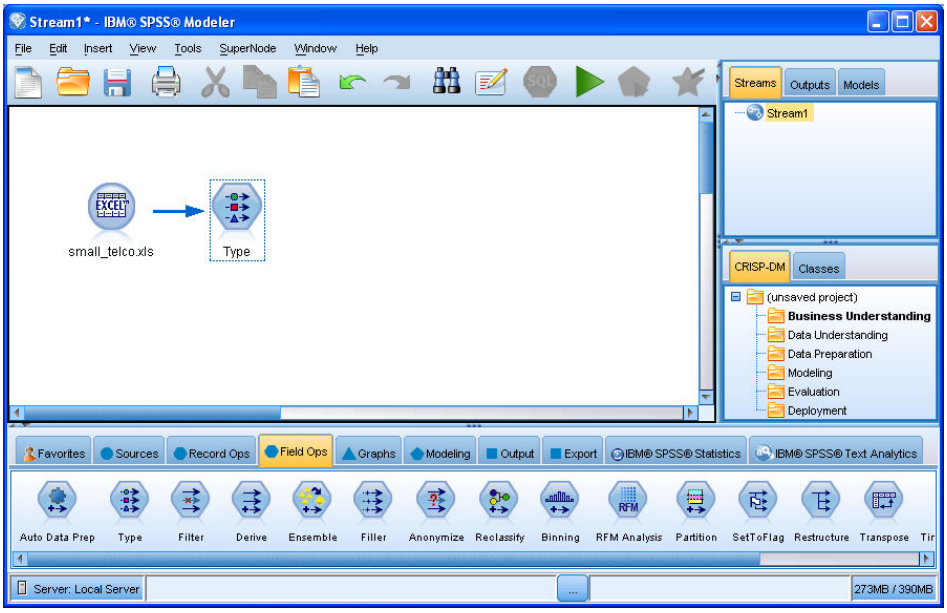
3. Put an Excel Source Node on the stream canvas.



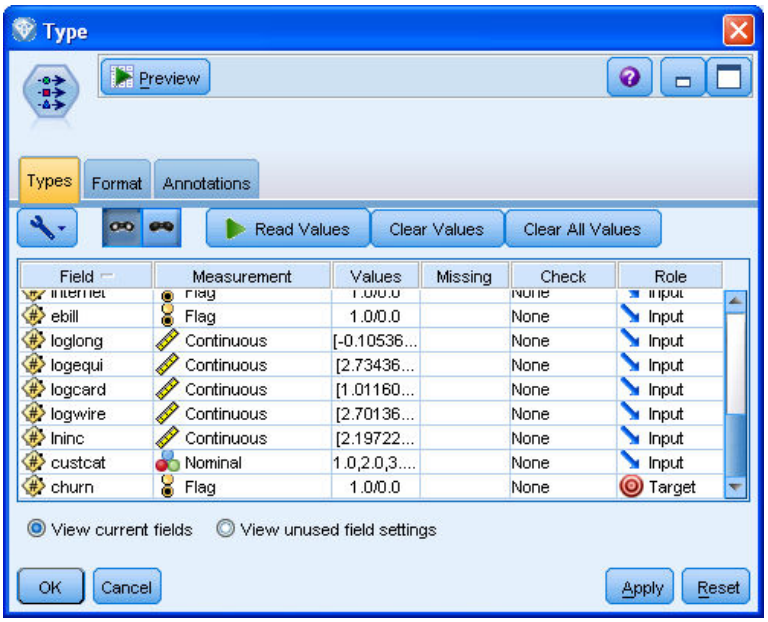
4. Import data file into the stream, keep the default settings, and click Preview to check the data.



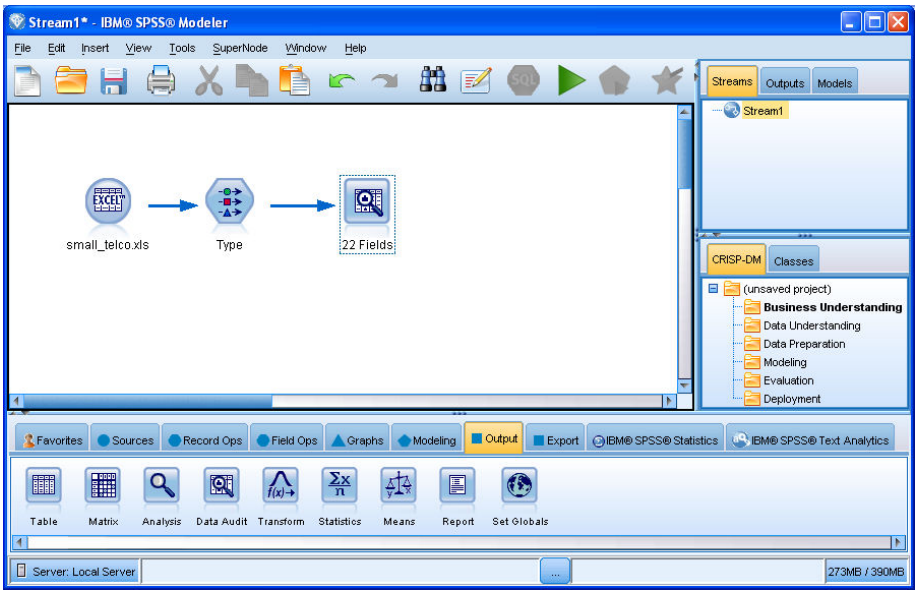
5. Add a Type Node on the stream canvas, and connect the Source Node with the Type Node.



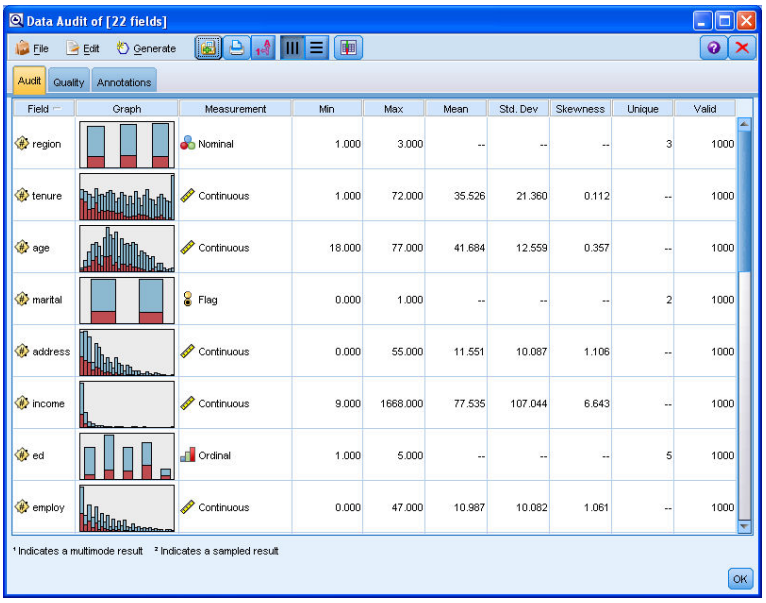
6. Double click the Type Node, set the appropriate measurement level for every field of the data. Set the role of the last field Churn as Target.



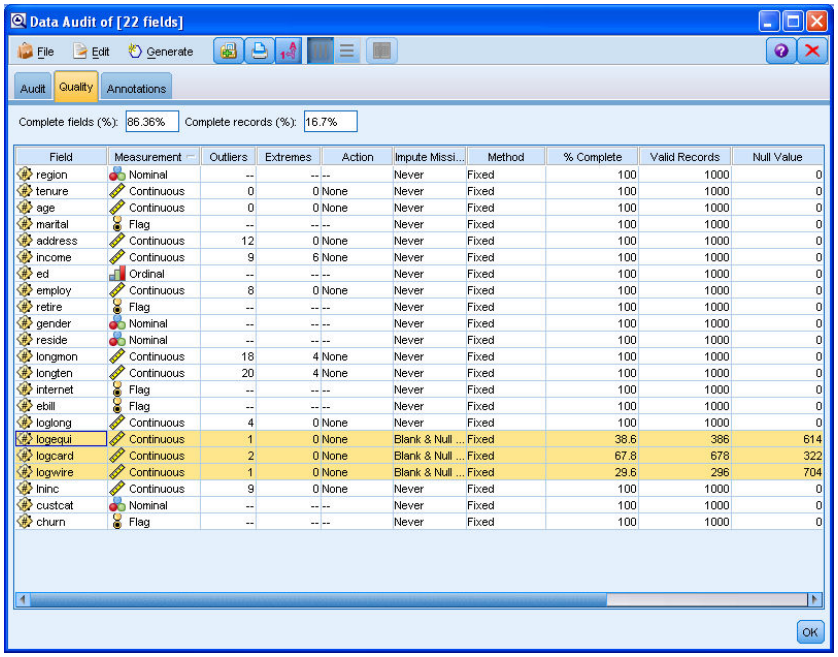
7. Add a Data Audit Node on the stream canvas, connect it with the Type Node.



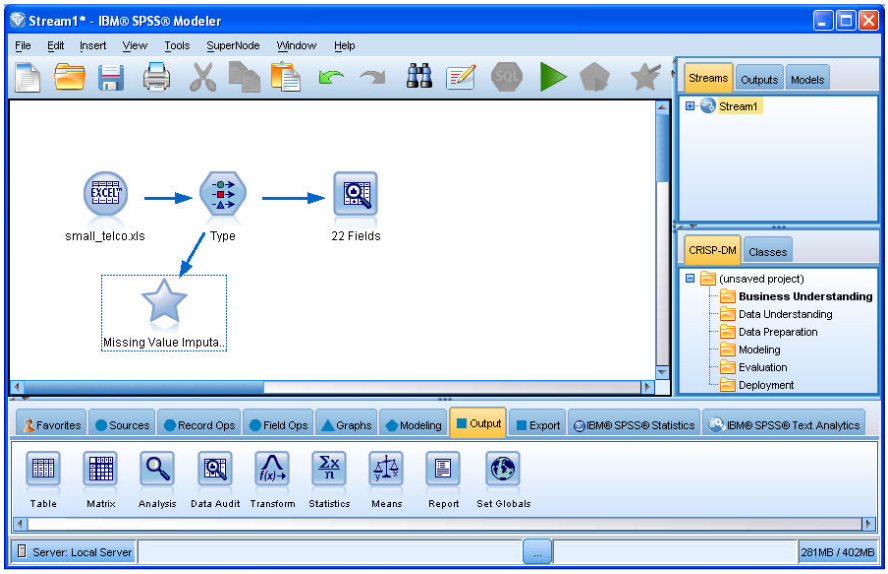
8. Double click the Data Audit Node, keep the default settings, click Run button. The statistics and charts are shown below.



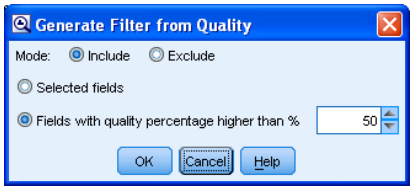
9. Click Quality Tab, then specify impute method for missing values in fields *logequi, logcard, logwire*.

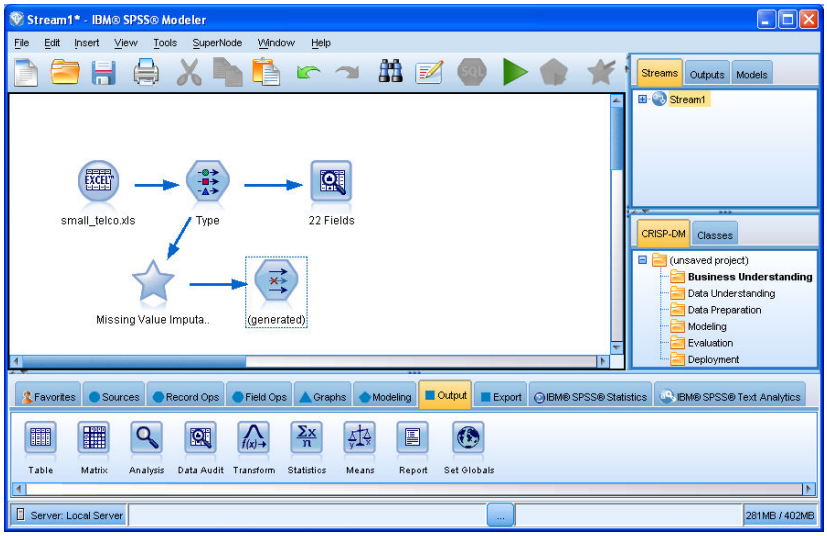


10. Generate Missing Values SuperNode with all fields, then connect it with the Type Node.

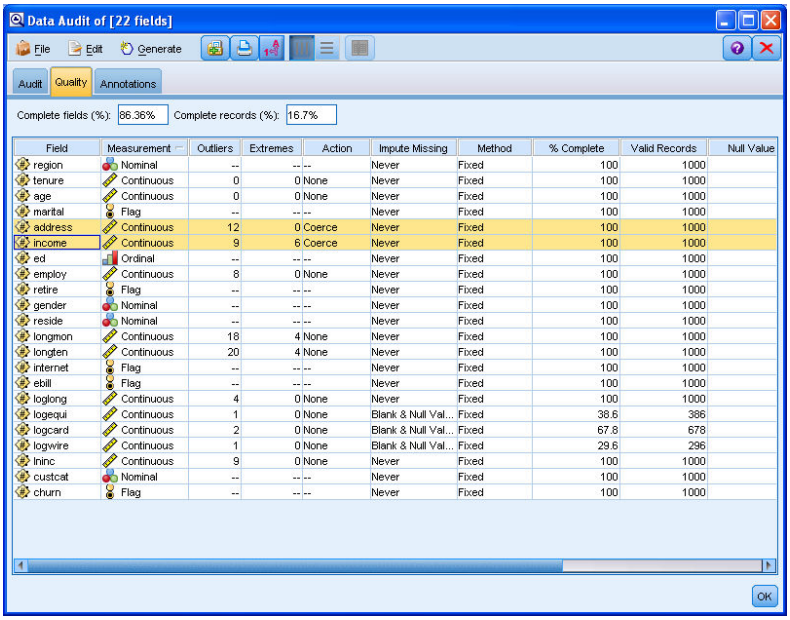


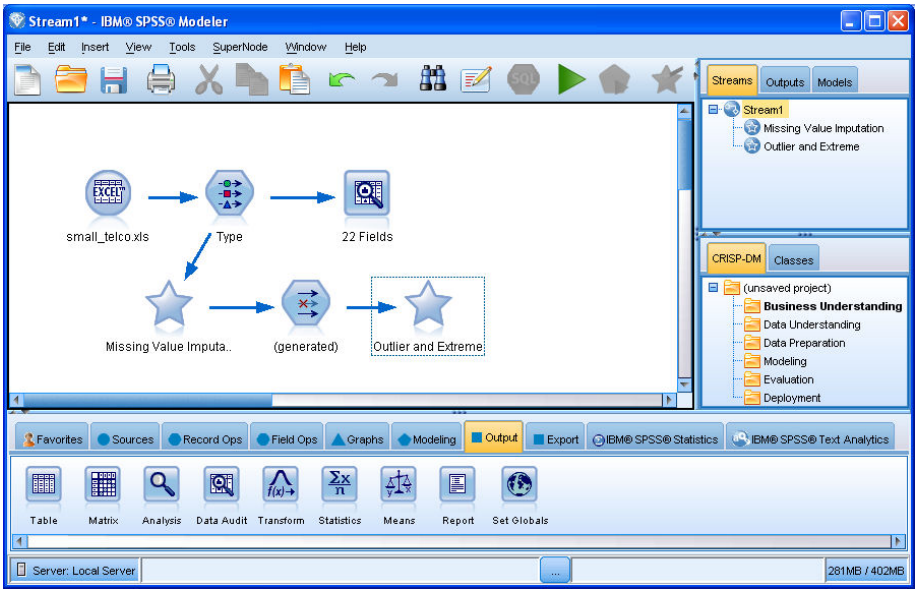
11. Generate Missing Values Filter Node with quality percentage higher than 50%, then connect the node with the Missing Value Impute SuperNode.

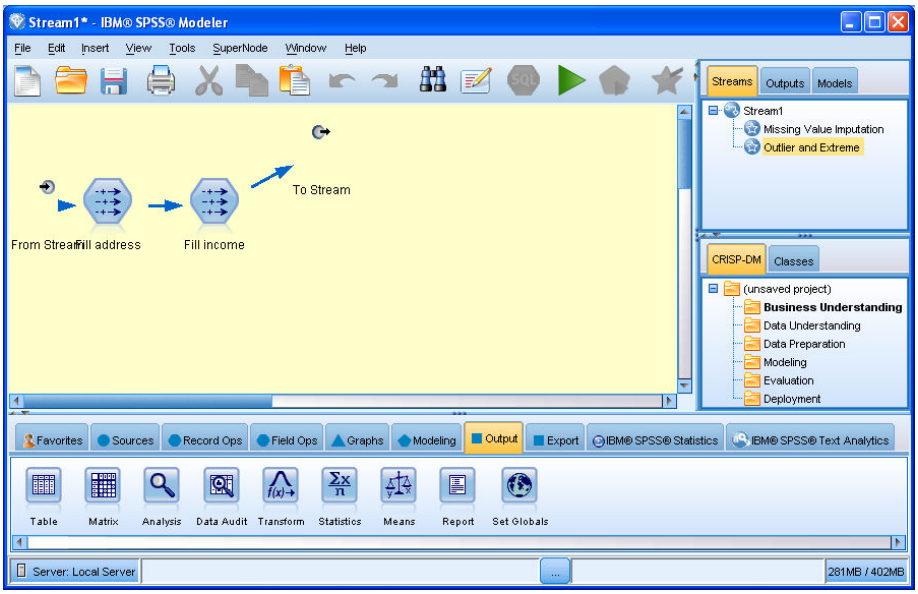




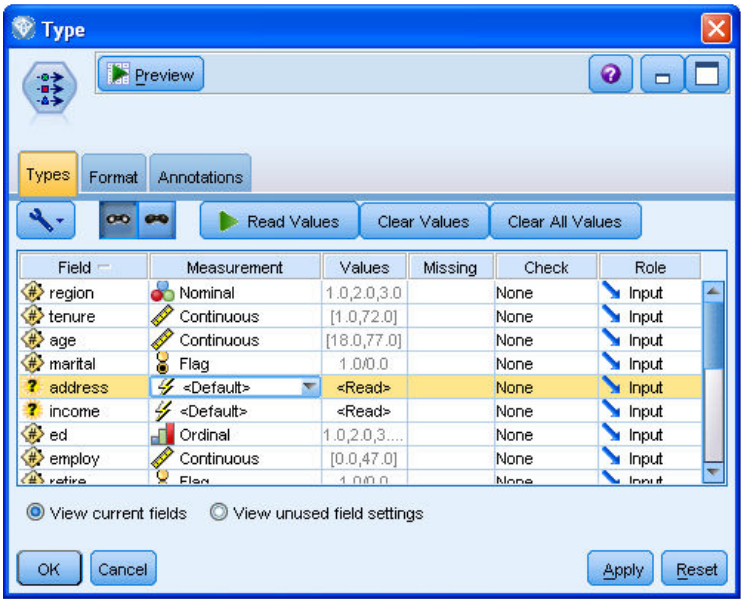
12. Select two fields with Outliers and Extreme values, choose appropriate Actions, and generate Outlier and Extreme SuperNode. Then connect it with the Filter Node just created. You could Zoom In the SuperNode to have a look at its details.



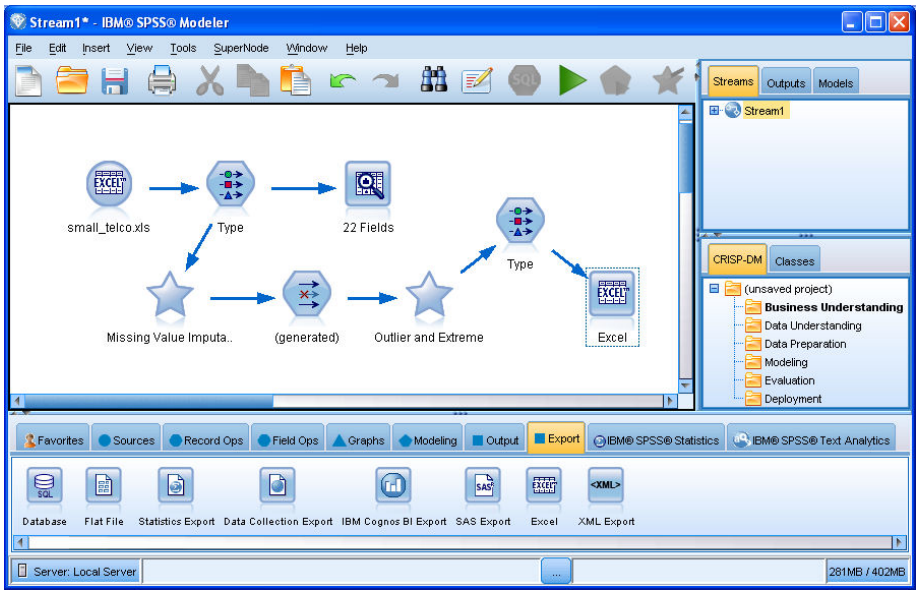




13. Add another Type Node on the stream canvas, connect it with the SuperNode. Then reset the measurement levels of these two fields you just processed.



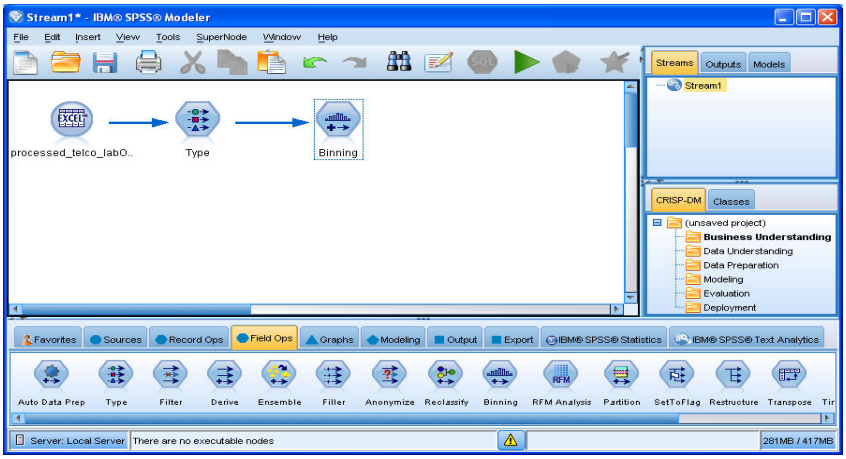
14. Then you could save the preprocessed data into an Excel file by adding an Excel Export Node on the stream canvas, then connect it with the Type Node. Double click the Excel Export Node to choose a location for the export data file. Save it as ***processed_telco.xls***.



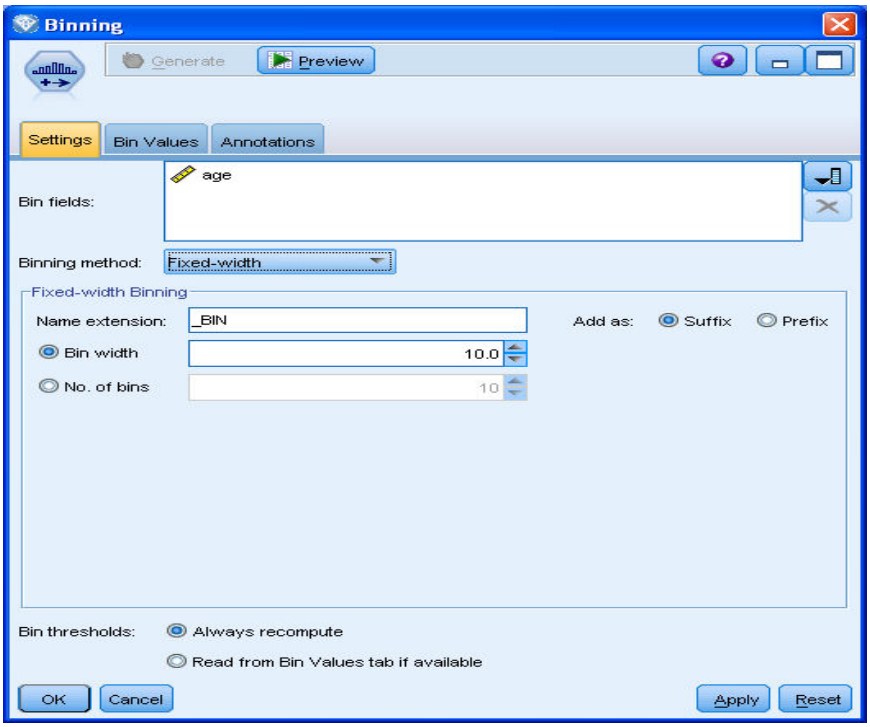
15. Open ***processed_telco.xls***, compare it with the original ***small_telco.xls***, what are the differences?

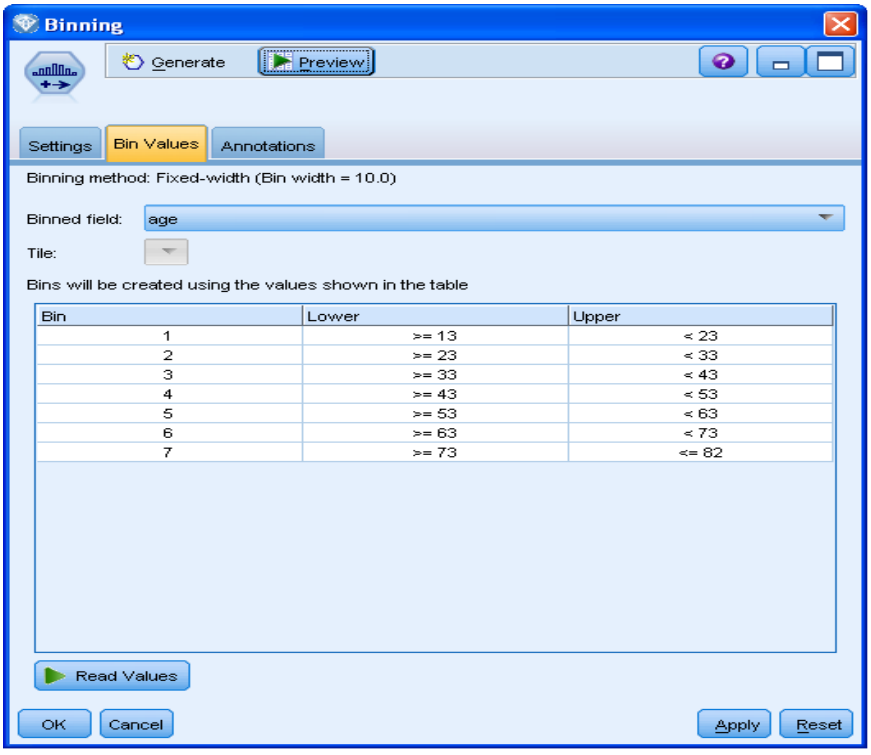
Binning

16. Create a new stream and load processed_telco.xls into SPSS Modeler, connect it with a Type node, defining the measurement levels for fields.
17. Create a Binning node and add it on the stream.

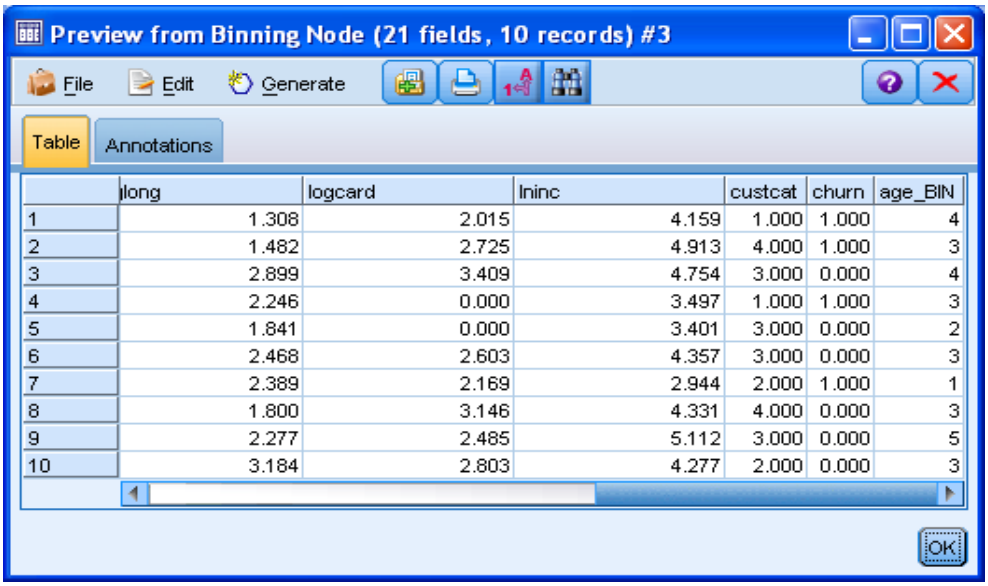


18. Choose age field to performing binning.





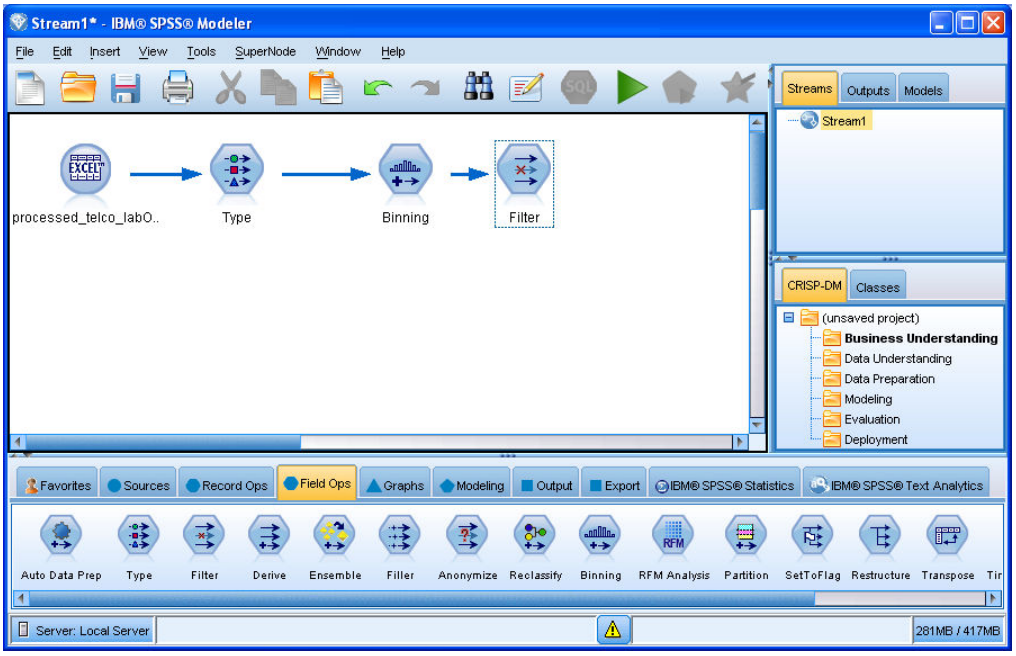
19. Click preview button to see the result.



	llong	logcard	lninc	custcat	churn	age_BIN
1	1.308	2.015	4.159	1.000	1.000	4
2	1.482	2.725	4.913	4.000	1.000	3
3	2.899	3.409	4.754	3.000	0.000	4
4	2.246	0.000	3.497	1.000	1.000	3
5	1.841	0.000	3.401	3.000	0.000	2
6	2.468	2.603	4.357	3.000	0.000	3
7	2.389	2.169	2.944	2.000	1.000	1
8	1.800	3.146	4.331	4.000	0.000	3
9	2.277	2.485	5.112	3.000	0.000	5
10	3.184	2.803	4.277	2.000	0.000	3

20. Export the modified data file with the new field added.

21. Could you remove the age field with the age_BIN field only? Which node should you add? Adding a Filter Node.



22. Discard the old age field, save the new age_BIN field.