

Lab #2: Data Management & Cleaning

Biostat 202: Opportunities and Challenges of “Big Data”

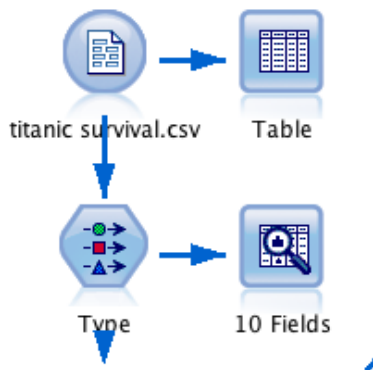
Due: Due Sunday 8/7 by midnight

The purpose of this lab is further familiarize you with IBM SPSS Modeler and, in particular, cleaning, imputing, & merging data. Please respond to the questions below (marked in bold) and hand in your lab assignment by posting it to the CLE. Make sure to add your name before posting.

For this lab we will use the same data on survival of the passengers from the Titanic, as well as adding in information on the crew on the cruise ship. You can use your final stream from Lab #1 or open the dataset is available on the course website.

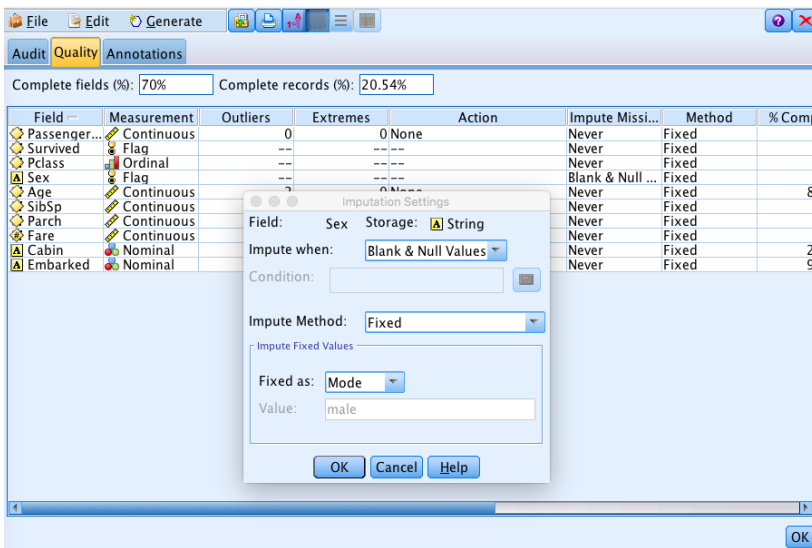
Our goal will be to add the crew information into the passenger dataset and build a model to accurately predict who survived and who did not.

1. After loading your data or stream into Modeler, run an audit and a table to ensure the data are correct.



Examine the audit results, where are the most missing values? **Q1. What are some ways we can impute values for the missing data?**

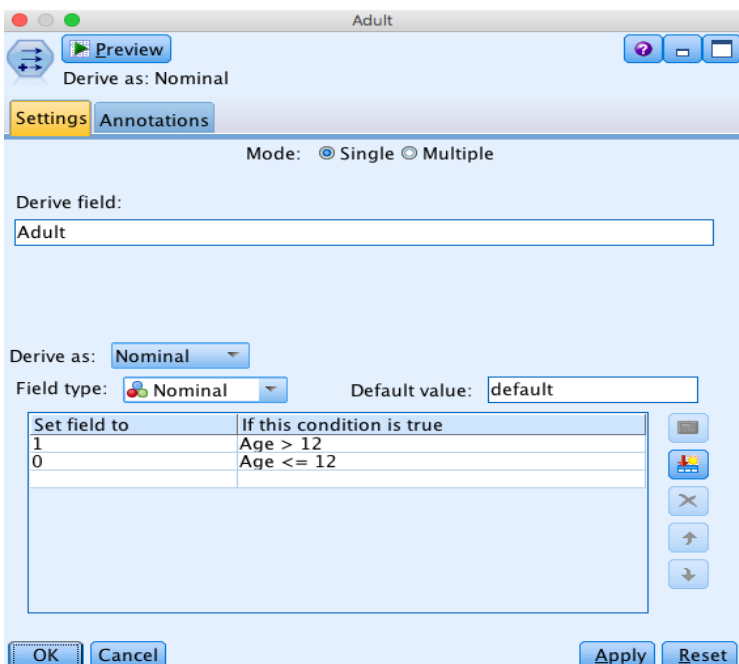
2. Open your Audit node and click on the ‘Quality’ tab and see that only 80% of the Age variable is complete. You can highlight the Age row and use the pull down menu in the ‘Impute Missing’ column to have Modeler impute Ages for those with Blank and Null Values. Use the pull down menu in the ‘Method’ column to identify the method you would like to use for imputation. Select ‘Specify’ and select either Mean or Mid-Range as the ‘Fixed as:’ method of imputation. Modeler offers a wide selection including identifying your own algorithm for imputing missing data. Finally, in the Generate menu, create a ‘Missing Values Supernode.’ **Q2. You now have a new star in your main window, where does it go in your stream? And what node will you attach after the Star node?**



3. One hopes that the clinical and medical data we receive does not include actual names. However, if it does and you want to, it is easy to anonymize your data in Modeler. Attach a new Type node after the star and then to the Anonymize node. Open the node & highlight Name, open the 'Anonymized Values' tab, see that Name is in the 'Anonymized field:' and click the 'Anonymized Values' green arrow. **Q3. Check that the data are now anonymized. What has replaced the name field?**

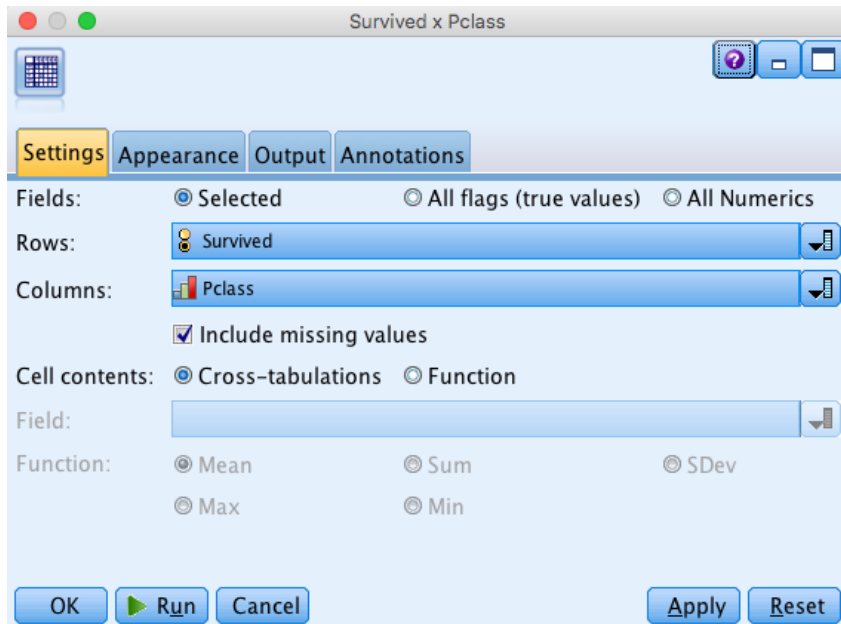
4. Next read in the titanic crew dataset (titanic_crew.xlsx). It is an Excel file so you do this with the Excel node in the Sources tab. You can open and read this data into the same palette but do not connect it to the passenger dataset. **Q4. Connect a type node and examine the variables, which are comparable to your passenger data? To make the two datasets comparable, what do you need to do in the passenger dataset?**

5. Use a Derive node to convert the Age variable in the passenger data to a variable called 'Adult' to make it comparable to the crew dataset. We now usually use 18 years to separate children from adults but there were no children's labor laws in 1912. **Q5. For my age cutoff I used 12 years old but feel free to use a different cutoff (older or younger). Attach a 'Means' node from the Output tab. What is the mean age of children vs. adults?**



6. Next you are ready to combine the datasets. You can choose to merge them or, better, to just append one dataset to the other. Connect both datasets to an 'Append' node in the Record Ops tab and open the node. **Q6. Which variables do the two datasets have in common? Attach a 'Filter' node to your 'Append' node and filter out unneeded variables. How many variables are left? What do you need to attach to your 'Filter' node before analyses?**

7. Attach 'Matrix' nodes to your stream and create separate tables of survival by Pclass, Gender, and Adult. Use the 'Appearance' tab to include row or column percentages. **Q7. Adding the crew information, do you see any trends by Pclass? Gender? Adult? What percent of the crew were female? Construct the appropriate table to answer this.**



8. Using the 'Logistic' node, run a survival model and attach an analysis node. **Q8. Is your model better or worse than without the crew (i.e.; compare to last week's lab)? Add Pclass to the panel at the bottom of the 'Analysis' node. You now have results by passenger class & crew. Are any more accurate than overall?**

9. Finally, attach an 'Auto Classifier' node (from the Models tab) and run this analysis of survival. You will now have several models with predicted accuracy. **Q9. Are the model results the same? Which model is best? Any ideas why they are different?**