Hwk #2: Data Management & Cleaning
Biostat 202: Opportunities and Challenges of "Big Data"
Due Wednesday, August 9th by Midnight

For this homework we will also use the data from the National Electronic Injury Surveillance System (NEISS) that you used for Hwk #1 (http://www.cpsc.gov/en/Research--Statistics/NEISS-Injury-Data/). We will be interested in examining age, and injury location (body_part) along with the disposition variable you created in Hwk #1.

Open your stream from Hwk #1. I have included the data as an Excel file in Hwk #2.

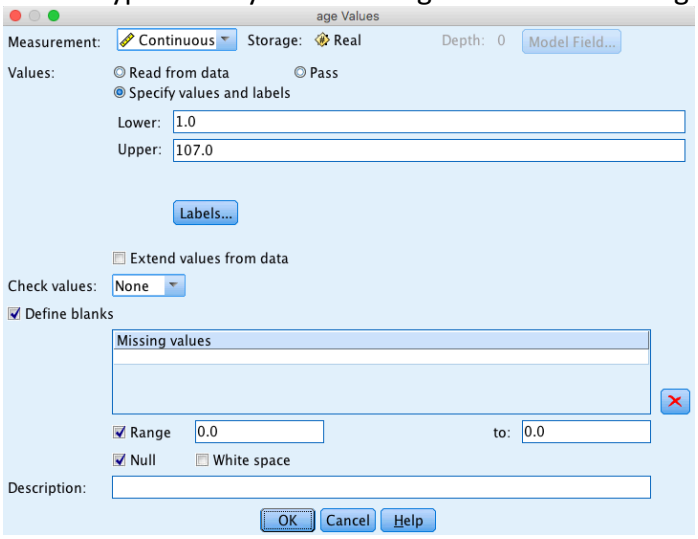
Now that you have already audited the data, we will do additional cleaning and recoding.

1. (4 points). Examine the race_other variable in the Excel file. If you were to use this variable in your analyses suggest how it could be recoded given its inconsistencies.
2. (6 points). Based on your 'Audit' node from Hwk #1 and an examination of the codebook coding for age, what is wrong with the age variable and how would you recode the information? Recode the variable into a usable variable for analyses. How have you changed the '000' used for 'not recorded' vs. the '2xx' variables used for less than 2 years old? You can use the 'Filler' node in the Field Ops tab to change the codes 200 and above to a number less than 2. [An example is given below]
3. (2 points). Should you replace the 'not recorded' observations with missing value imputation here or replace them with a specific missing value code? In your 'Type' node and, if keeping the missing data as missing or 'not recorded,' create a missing value code for this that matches your recoding from #2 in the 'Type' node.
4. (5 points). Recode body_part into two Flag variables to identify Head (body_part=75) or Other as the body_part injured, and another variable to identify Toe (body_part=93) or Other as the body_part injured. Note that you can use 'Default value' as 0 for unspecified values.
5. (3 points). Use the 'Filter' node to eliminate variables you will not use in an analysis (i.e.; text fields). Which variables have you eliminated?
6. (2 points). Audit your new dataset and examine the quality and that the recodes worked. In general, what is the quality of your recoded and existing variables?
7. (4 points). Run a 'Means' node from the Output tab for age by your new body_part variables and admitted or not (by selectively changing the target to examine age by admitted or Head or Toe). What is the mean age for head injuries, toe injuries, and hospital admissions. What does this tell you? Tabulate admitted by Head or Toe. Which body_part is more likely to require admission?
8. (3 points). Predict Hospital admission based on your new body_part variables, your new age variable, and any others that you wish to include in a Logistic model and Neural Net. Which model had higher prediction & how good were the predictions? Did the model predict Head or Toe injuries better?*

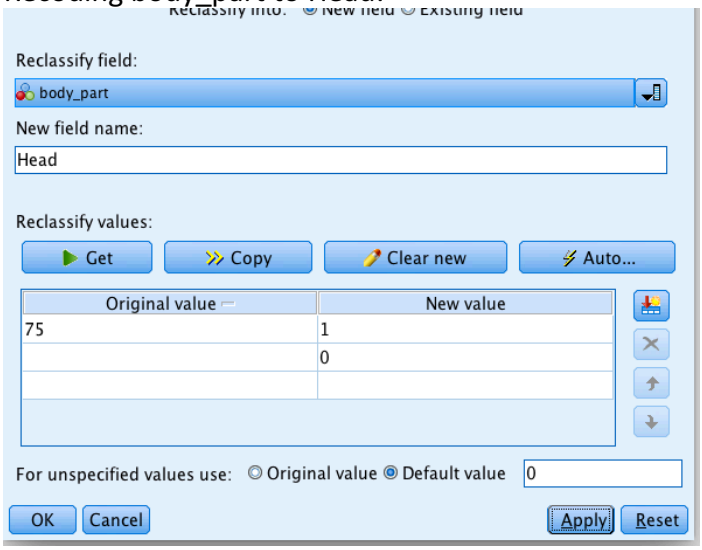
*Reminder: you want to partition your data to have a training & testing subset here

Visual Aids:

In the ‘Type’ node you can change ‘0’ to be a missing value by choosing ‘Specify’ in the Missing field:



Recoding body_part to Head:



Using the ‘Filler’ node to recode ages >= 200 to 1