Lab #1: Introduction to IBM SPSS Modeler
Biostat 202: Opportunities and Challenges of "Big Data"
Due: Friday, 7/29 by midnight

The purpose of this lab is to familiarize you working with "nodes" and building "streams" in IBM SPSS Modeler. In Modeler each step in the data analysis is represented by a node and the sequence of actions represented as a series of linked nodes is called a stream. In this lab we will build streams to read data into the software, define their roles, do a data quality check for missing data, transform some of the variables, generate simple tabular descriptive statistics, build some predictive models, and check their performance by dividing our dataset into a "training" and "test" subset.

Please respond to the questions below (marked in bold as **Q1** to **Q8** below) and hand in your lab assignment by posting it to the CLE.

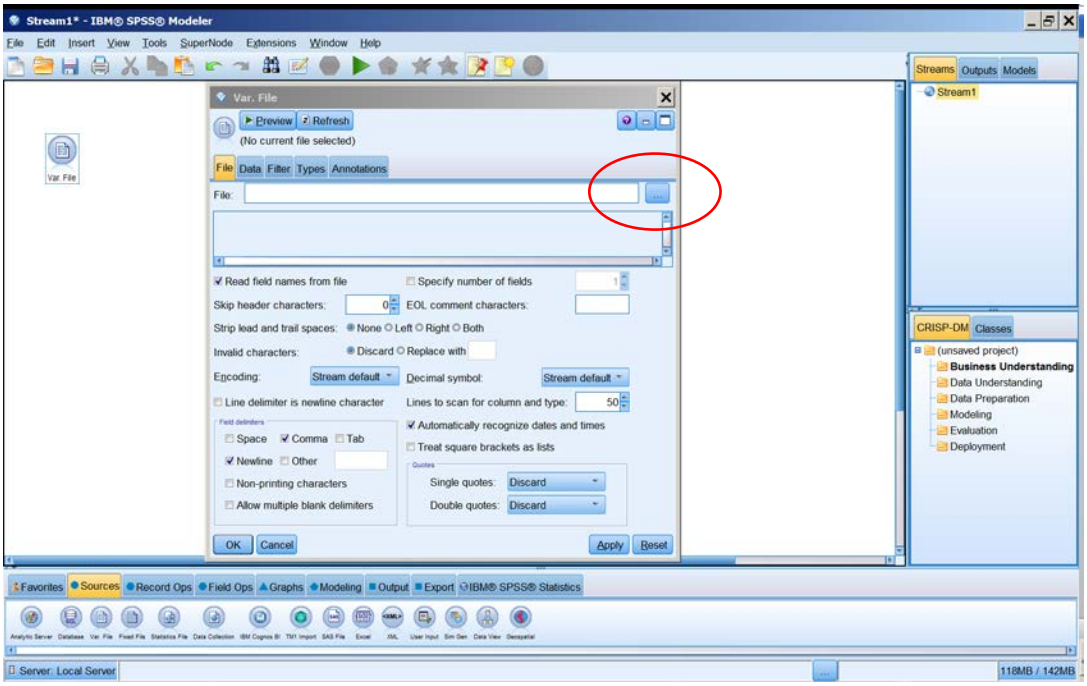
For this lab we will use the data on survival of the passengers from the famous sinking of the cruise ship, Titanic, which was also a Kaggle competition (www.kaggle.com). The dataset is available on the course website. It contains the following variables:

PassengerID:	passenger ID variable
Survived:	1=survived, 0=died
Pclass:	passenger cabin class, 1=First class, 2=Second class, 3=Third class
Name:	Name of passenger
Age:	Age of the passenger
Sibsp:	Number of siblings/spouses on board
Parch:	Number of parents/children on board
Ticket:	Ticket number
Fare:	Passenger fare
Cabin:	Cabin passenger was assigned to
Embark:	Port in which passenger boarded the Titanic (C = Cherbourgm, Q = Queenstown, S = Southampton).

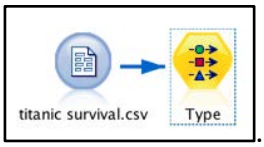
Our goal will be to build a model to accurately predict who survived and who did not.

1. Because Modeler depends on a graphical interface to select nodes and build streams, it is important to get an idea of how a stream is built up by watching its construction. Watch the YouTube video available at https://www.youtube.com/watch?v=316XuQHs7oE . This video covers the analysis of the same dataset. However it moves very quickly through the steps without much explanation so just watch it to get a rough idea of the flow. We will walk through a simpler set of steps below.

2. Next download the dataset “titanic survival.csv” from the course website. Open it with Excel or a spreadsheet program to get a sense of the data. The extension “.csv” is short for comma separated values, so named because if you look at the file in its raw format (e.g., using Word) the data values are separated by commas.
3. Now we will start building a stream. A stream represents the sequence of operations from the data source to the final model or output. Start Modeler and click on “Create a new stream.”
4. Along the bottom, click on the “Sources” tab and double-click on the “Var. File” node. A node will appear in your main window titled “Var. File”. Hovering over your selection will show a short description of the operation of the node.
5. Double-click on the node to open it up. Your screen should look something like the below. Click on the browse button (circled in red) and find the file you downloaded. Click on the “OK” button to import the file into Modeler.

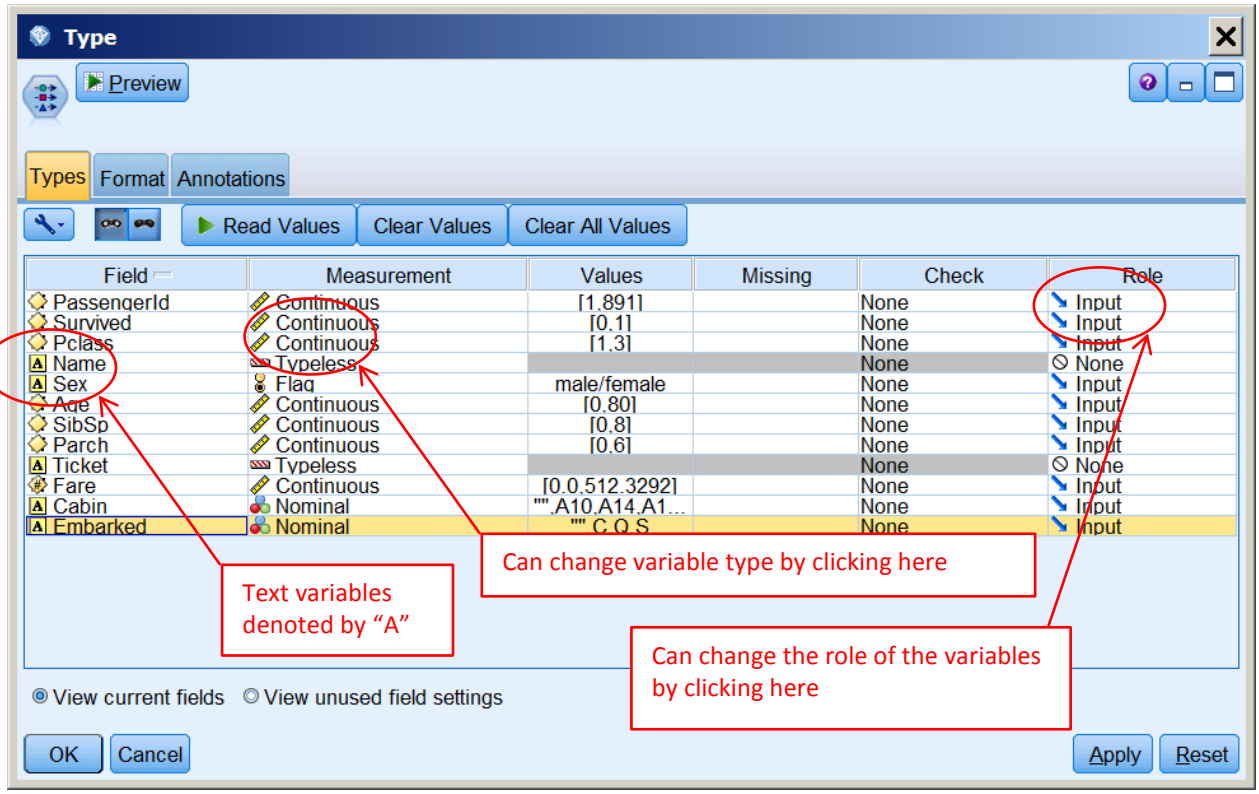


6. The menu will display the first few lines of the dataset. Click on the “Preview” button to see how it did at reading in the data. What do you notice about the first 3 variables? **Q1. Why do you think Modeler was unable to read in the dataset using its default values?**
7. To fix this error, close the preview window and select the drop down menu in the lower right corner title “Double quotes” and select the option “Pair and discard” and again click on the “Preview” button.
8. **Q2. For passenger number 6, what do you think happened to the value of “age”?** Close the menu for the node.
9. Next we need to tell the software what the nature of the variables will be in the analysis and what type of variables they are.
10. Click on the “Field Ops” tab (this is where you will find nodes that manipulate the variables or fields) and double-click on a “Type node” which will now appear in your

main window and look something like this: 

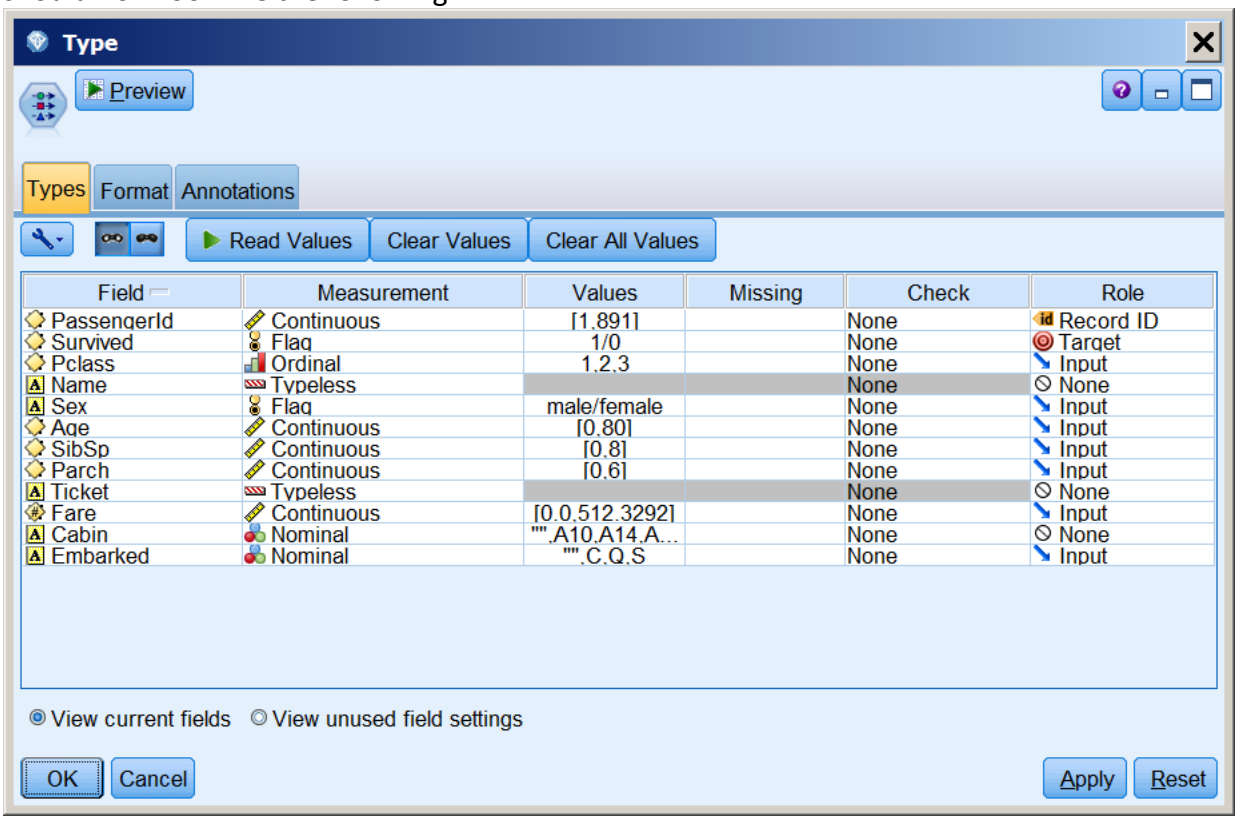
. If the "Var. File" node was selected then the new "Type" node will be connected to it with an arrow. If not connect them using the steps below.

11. Note on connecting nodes into a stream: The easiest way to build a stream of nodes is to first select the starting node you want to add to and then double-click to add the ending node to the main window. They will then automatically be connected. If you added a node and then want to connect them afterwards there are several methods. You can select the starting node and press F2 and then click on the ending node. Or you press the Alt key and then click and drag from the starting node to the ending node. Finally, you can right click on the starting node, select connect and then click on the ending node.
12. Next open the "Type" node and click on "Read Values." Modeler will read the values in each of the variables and try to figure out what type of variable it is. Continuous is a numeric value, Flag is a binary categorical variable, Nominal is a multiple value categorical variable. To the left of the "Field" column, an "A" designates a text field. Your node menu should look something like the following:



13. The Survived variable is binary, so change it to a "Flag" measurement type. The Pclass variable is ordinal, not continuous so change that as well. While this node is open, identify the role of the variables by clicking in the "Role" column and selecting new

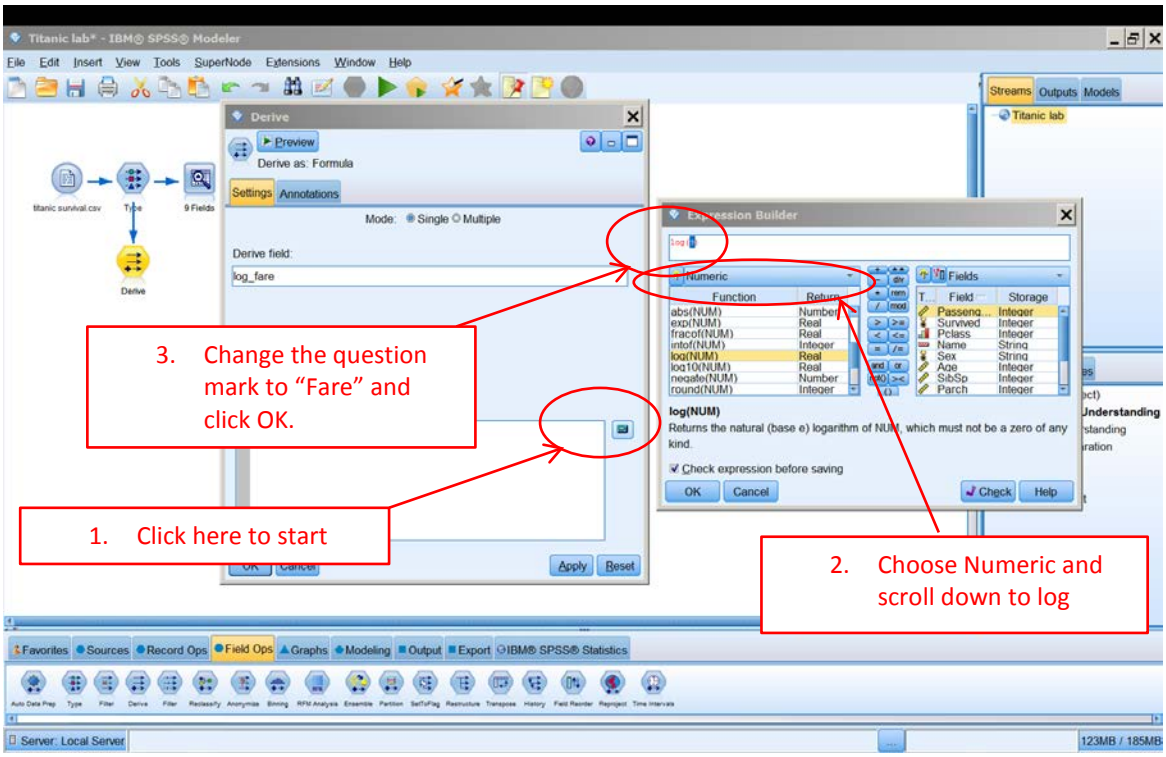
types of roles. PassengerID should have a role of "Record ID", Survived should be "Target" since it is the target of our prediction modeling. For this first exercise we will just use a subset of the variables. Variables designated as "Input" will be used in the model but those designated as "None" will not be used. Modeler has already designated Name and Ticket as "None." Also change Cabin to "None". Your node menu should now look like the following:



14. Close the "Type" node menu and next do some quality control checks of the data. From the "Output" tab at the bottom add a "Data Audit" node to the stream. Open the node and click on Run.
15. Modeler will display each of the variables in a histogram (for continuous measurements) or a bar chart (for categorical measurements) with color coding for survival or not (since we have designated that as the target variable). Double-click on the Fare variable figure to enlarge the histogram. Note that it is highly skewed to the right, which we will consider below. **Q3. Looking at the figure, which fares are likely to fare better with regard to survival?**
16. Close the histogram window and click on the "Quality" tab in the output window from the "Data Audit" node. Modeler marks some of these as outliers (you can control the definition of an outlier in the "Quality" tab in the "Data Audit" node). **Q4. Which variables have missing data?**
17. The fare variable is skewed right and has some outliers but is likely to be an important variable for predicting survival. Sometimes with a highly skewed variable, predictive models will work better if they are transformed so as not to be so skewed. The

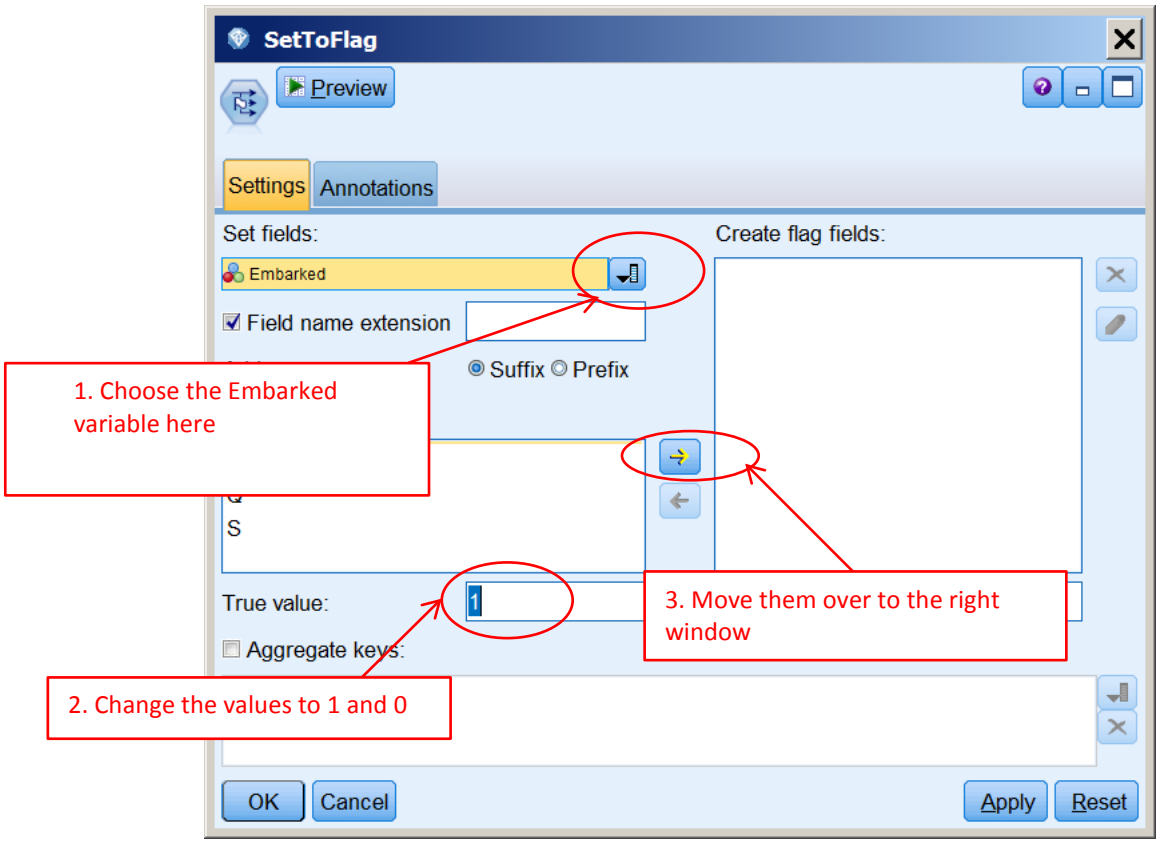
logarithm of right skewed variables tends to be much less skewed, so we will create a log transformed version of the variable. To do so, add a "Derive" node and connect it to the Type node.

18. Give the new variable a name in the "Derive field" box. I called it "log_fare." Click on the little calculator on the right and select Numeric, scroll down to "log(NUM)" and double-click on it. It will enter "log(?)" in the upper window. Change the "?" to the "Fare" and click OK. Modeler *does* distinguish upper and lower case so don't use "fare". If you already knew the name of the function you wanted, you could have just entered it in the lower box in the "Derive" node and not used the menus. You can also double-click on the variable name and it will insert it for you. Here is what the node should look like:

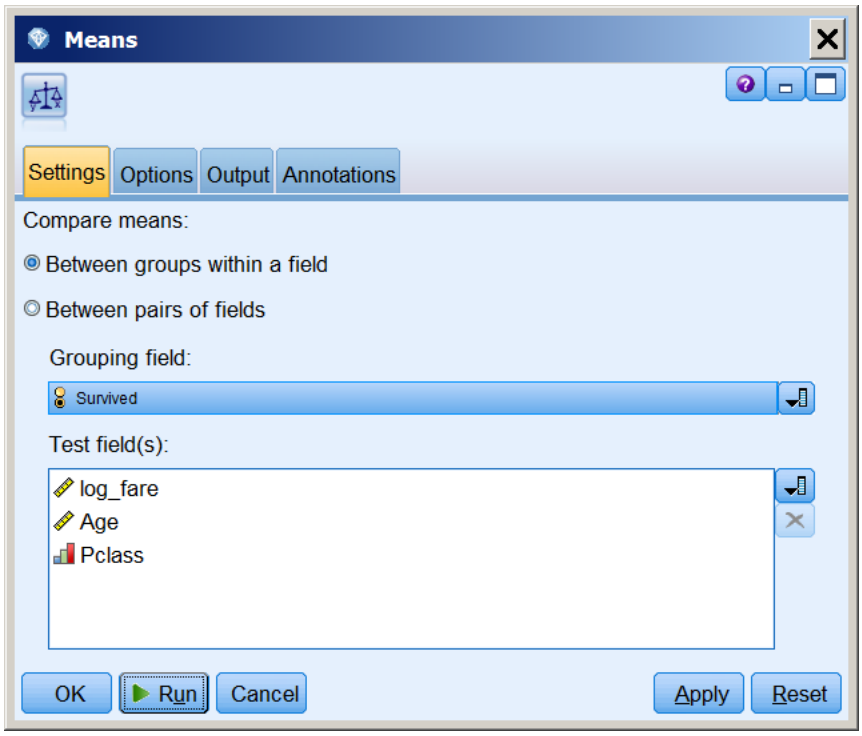


19. Disconnect the "Data Audit" node from the "Type" node and attach it to the "Derive" node. You can do this with the F3 key, or by right-clicking on it and selecting "Disconnect." Or, if all else fails, delete it and add a new "Data Audit" node. Now check the effect of the transformation on the skewness and data quality. **Q5. Is the log_fare variable less skewed? Does it have fewer outliers?**
20. Another common variable transformation is to change a categorical variable into a series of dummy or indicator variables. We will next create a set of dummy variables for the categorical variable Embark. When we are done we will have a set of variables that are coded 0 or 1 (with 1 indicating it is that category and 0 indicating it is not).
21. We can do this using a "SetToFlag" node (under "Field Ops"). Disconnect the "Data Audit" node and add such a node to the "Derive" node that is now titled "log_fare".

22. Select the “Embarked” variable by clicking on the variable list (see below). Also change the “True value” to 1 and the “False value” to 0. Click the right arrow to move it from the left to right windows and click OK.

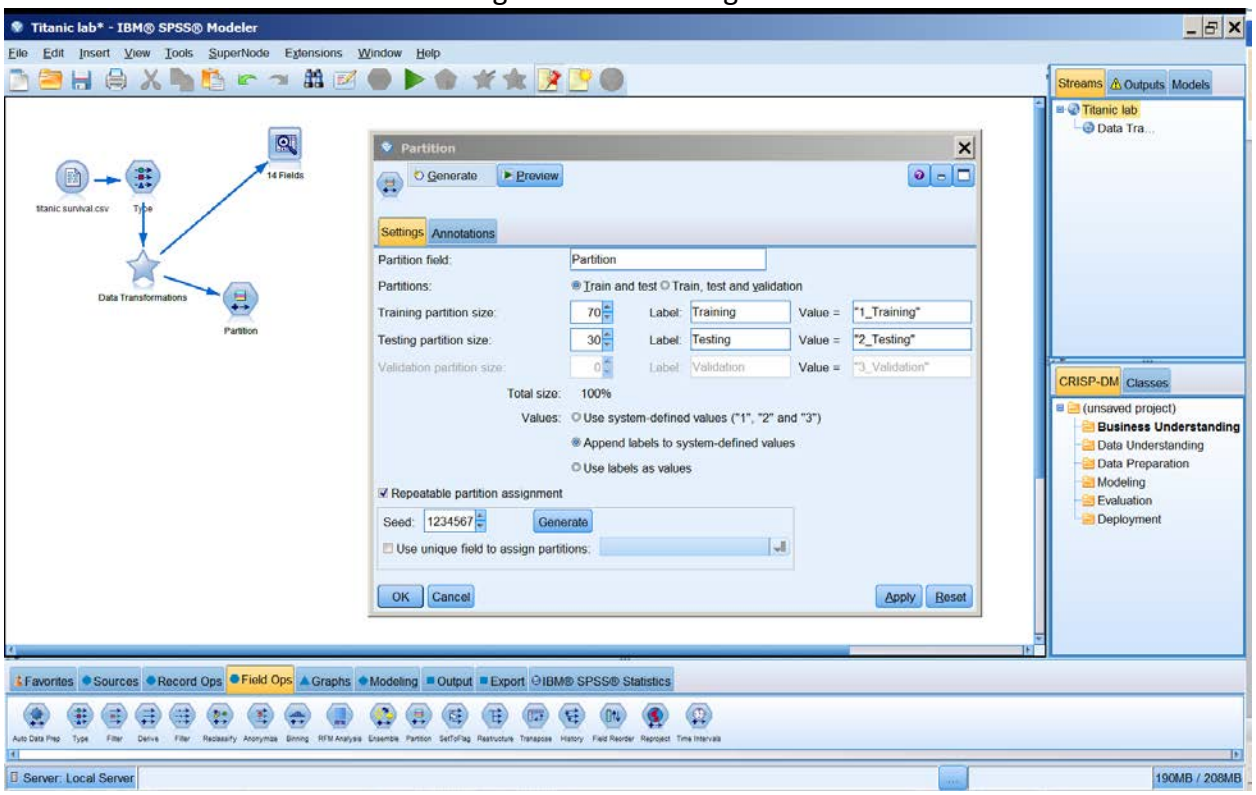


23. Reconnect the “Data Audit” node and view the results of creating the dummy variables.
24. Next we will calculate some simple descriptive statistics. Let’s create a table that gives the average value of age and the logarithm of fare separately for those that did and did not survive. Under the “Output” tab connect a “Means” node to the “SetToFlag” node. Open up the node and select “Survived” as the “Grouping Field” and select log_fare, Age and Pclass for the “Test Field” and click on “Run”. **Q6. Do there seem to be differences between those that did and did not survive? Do they make sense?**



25. Now we are ready to start building a predictive model. Since we created new variables we need to make sure Modeler knows what type they are and what their role will be. Add another "Type" node after the "Derive" nodes and make sure they are read in correctly and specified as "Input".
26. Housekeeping: When the streams get complicated three things can help make them easier to read. First, you should name the nodes. Go back to the "log_fare" node and open it. Click on "Annotations" and "Custom" and enter a better name such as "Calc log of fare". Also add a name to the "SetToFlag" node. Second, you can group nodes into what is called a "SuperNode". We will group the two data transformation nodes and the "Type" nodes. Select all of those nodes (you can also click and drag a rectangle around them). Then right click one of them and select "Create SuperNode" or select "SuperNode" from the top menu and select "Create SuperNode". You can likewise give this a better title. If you want to look inside a SuperNode later you can right click on it and "Zoom in" or you can expand it back out like it was. Finally, you can easily drag the nodes around in the main window to arrange them more artfully to make the stream easier to read.
27. A very good idea when building a predictive model (as will be discussed later in the course) is to reserve some of the data (called a "test" data set) to evaluate the performance of the model we developed (typically called "trained" in the predictive modeling literature). If you try to use the same dataset to train and test a model, it usually appears over-optimistically good since it may adhere too closely to fluctuations in the data that just represent random variation that would not be present when using the model to predict additional cases. The data can be randomly divided into training and testing subsets using a "Partition" node (under "Field Ops"). Add such a node, open

it up and change the “Training partition size” to 70 and the “Testing partition size” to 30. Your screen should look something like the following. Then click OK.



28. For our first model we will choose to run a logistic regression (more detail later in the course). It is a relatively standard statistical modeling method that takes multiple inputs (or predictors) to predict a binary target variable (in this case survived or not). Add a “Logistic” node (from the “Modeling” tab at the bottom) to the “Partition” node and click on “Run”. Another node will appear (Modeler calls these “model nuggets”). You can open the model nugget if you want to see the details of the model that was fit to the training data.
29. Next we will calculate some summary statistics to see how well the model works. Add an “Analysis” node (under the “Output” tab) to the model nugget node, open it and click on “Run.” **Q7. How well did it work? Did it do better on the training or test dataset?**
30. Next try another method to see if you can do better than the logistic regression (so add “Modeling” and “Analysis” nodes). Good bets would be some of the modern machine learning methods such a Classification and Regression Tree (“C&R Tree”), C5.0, a neural net, or a support vector machine (SVM). More details on some of these algorithms later in the course. **Q8. What additional method did you try? Was it better, worse or about the same as the logistic regression?**