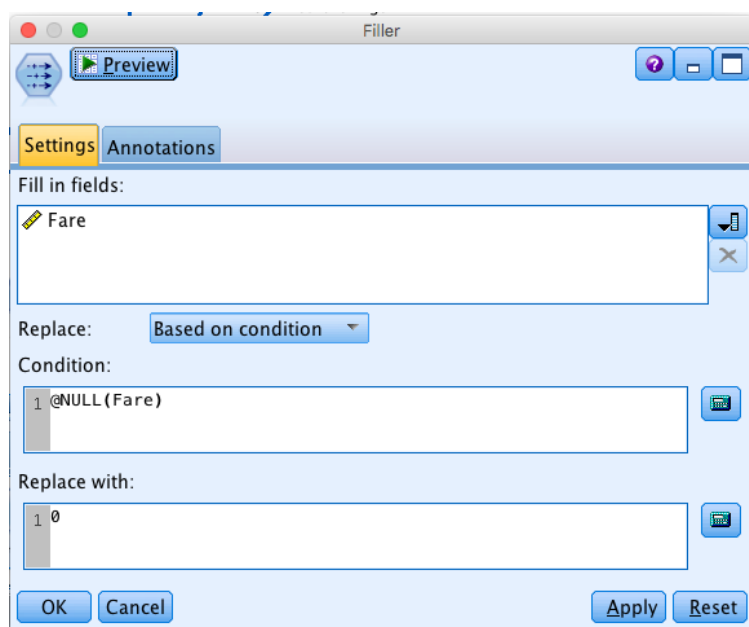


Lab #3: Data Visualization & More Modeler
 Biostat 202: Opportunities and Challenges of “Big Data”
 Due: Due Sunday 8/13 by midnight

The purpose of this lab is further familiarize you with IBM SPSS Modeler and, in particular, visualizing your data & using visualizations to examine models. 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.

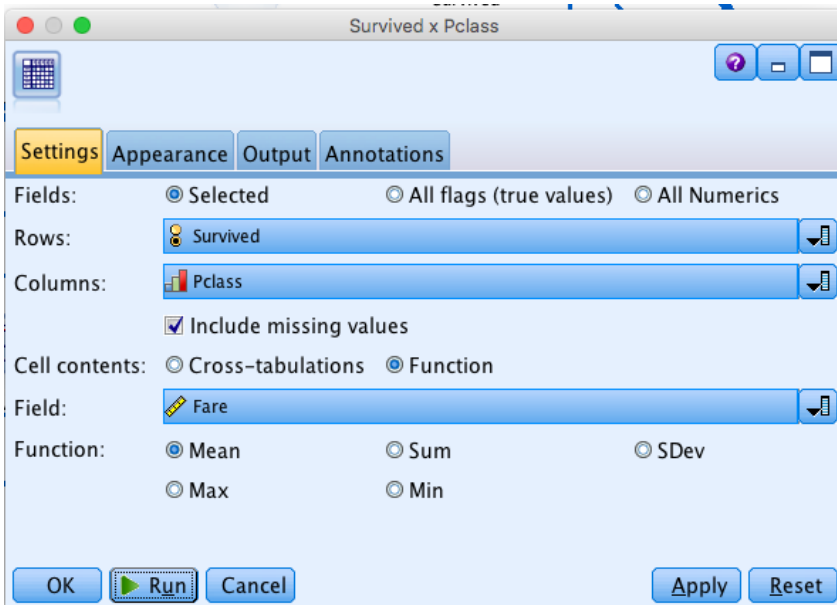
For this lab we will start with the same stream and data on survival of the passengers from the Titanic, as well as adding in information on the crew on the cruise ship. Open your stream from Lab #2 as we will be adding some nodes to the models you have already completed. Our goal will be to examine some of the data graphically as well as build a better model of survival.

1. After loading your Lab 2 stream into Modeler, this time do not filter out the ‘fare’ or ‘age’ variables. In adding in the Crew data we did not have a value for ‘fare.’ Since they are Crew, we can add a value of ‘0’ as the fare for the Crew. Create these values using the Filler node.

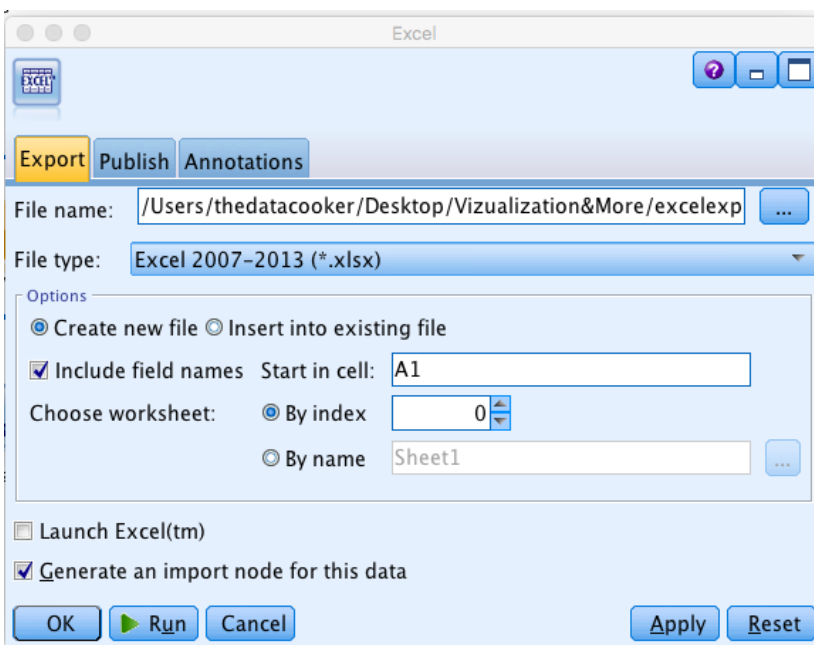


Run an audit to make sure the data are correct and that you have data for Sex, age, Adult, PClass, Survived, and Fare. **Q1. Examine the quality of the data using the Audit node. Where are there missing values?**

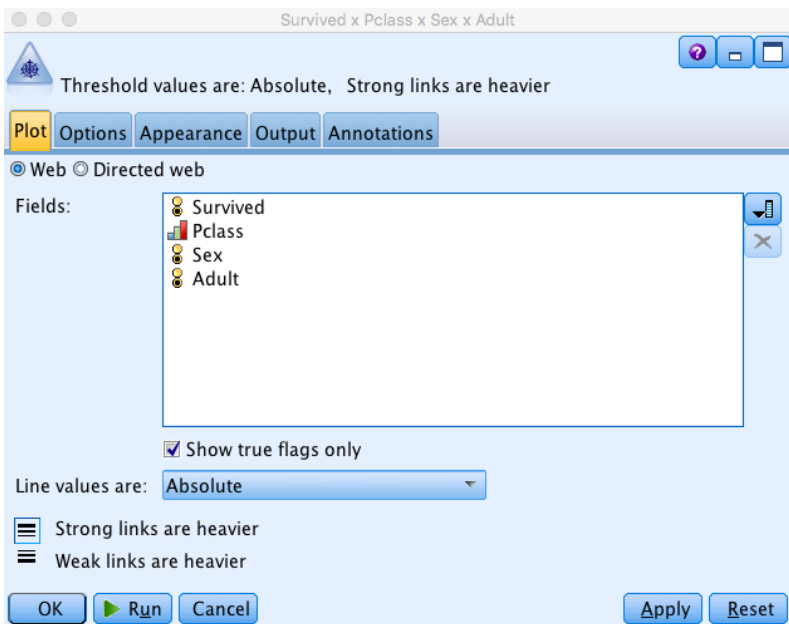
2. Using the Matrix node, look at the Fare paid for Survivors vs. non-Survivors by Sex and then by Class. **Q2. Does it appear from these tables that the higher the fare, the more likely to survive? Is this true within the Pclass?**



3. You can export your data (after modifying it) into several formats. Export the titanic data with an Excel node from the Export tab. Check 'Generate an import node for this data' and you will have a new Excel data node. Start a new analysis stream connecting a Type node to this new Excel node (labeled 'excelexp.xlsx').

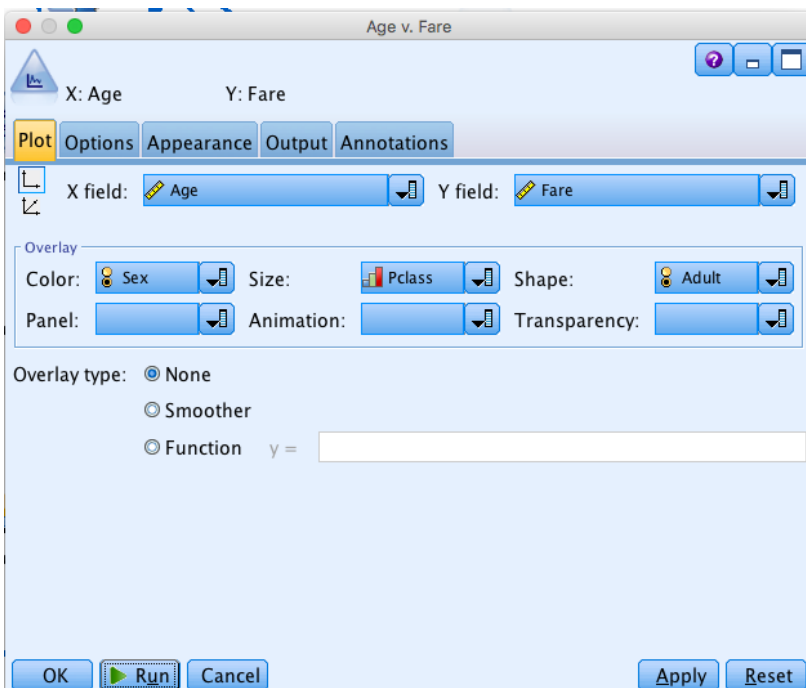


4. From the Graphics tab attach a Web node. This node only uses non-continuous data and graphically shows relationships between variables as the number of pairs of variable values that occur in the data. Open the node & check 'Show true flags only.' This eliminates all the pairs that do not occur together.



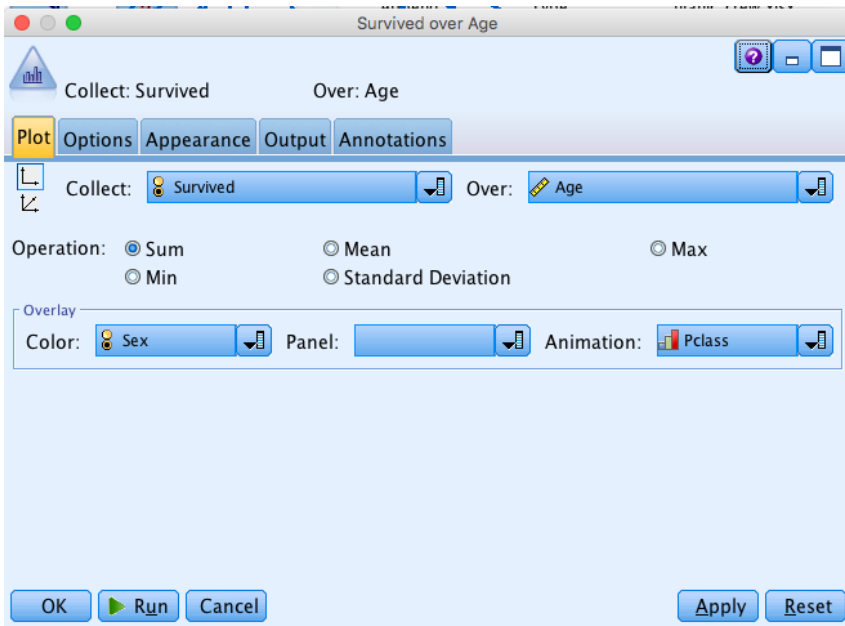
In the resulting web you will see stronger lines between pairs of variables. You can see what values & variables the lines represent by putting your cursor over them. **Q3. Which values of variables have the strongest connections?**

5. Choose the Plot node from the Graphs tab and plot Age vs. Fare. Overlay Sex (Color), Pclass (Size) and Adult (Shape). Note that you can move your cursor over the points in the plot to identify the data from the plot point. Under options choose Jitter and Use all data. **Q4. Which points are outliers and who are they?**



Next re-create the plot by adding Survived to the Panel. You now have two plots, one for those who survived and one for those who did not. **Q5. What are the differences between these plots?**

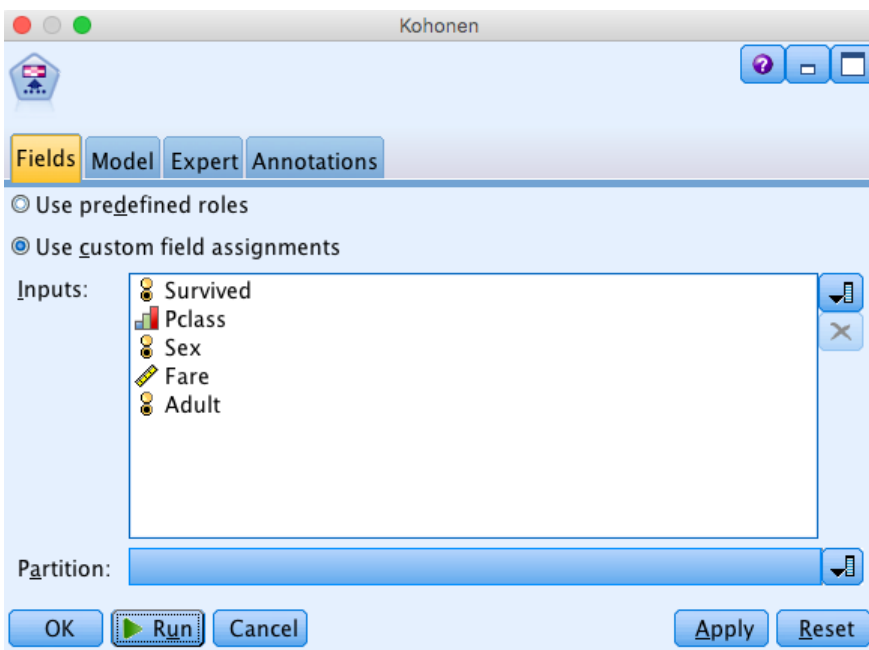
6. Use the Collection node to create an interactive plot. Slide the bar at the bottom over the 4 Pclasses.



Q6. How does the histogram change and why are there no values for Pclass = 4.

7. Using the Auto Classifier node, run a model choosing Survived as your Target and Adult, Sex, Pclass, and Fare as your predictors. Open the result & choose the Graph tab. **Q7. Which predictors are most important?** Attach an Evaluation plot from the Graph panel. This shows the Gain your model improves predicting survival vs. the 45 degree line representing chance.

8. Attach a Kohonen model from the Modeling tab. This is a graphical cluster analysis that groups the data into clusters and displays them in a matrix format.



Open the resulting Node and use the tab at the bottom to examine the clusters and the tab on the bottom right to examine predictor importance. **Q8. Which variable was not important to the analysis?**

9. Attach a Plot node to the Kohonen result node and choose \$KX-Kohonen as your X field and \$KY-Kohonen as your Y field. Choose Survived for Color, Pclass for Size, and Sex for Shape. Under options choose Jitter and Use all data. Note that in the resulting plot you can put your cursor over a data point to see the characteristics of that cluster. Re-create the plot by changing the overlays to Panel by Survived, Color by Pclass, Size by Fare, and Shape by Sex. **Q9. What clusters show up in one plot but not the other?**