**Lab Sheet 5: Answers**

**Visual Analytics**

Ian Nabney

* **Looking at the graph, what can you tell about the distribution of the data? Do you think it is likely to come from a normal distribution?** There are no negative Z-scores bigger than 0.7 in magnitude, while the largest positive value is 2.455. Also, there are 5 positive Z-scores and 10 negative ones. These facts suggest that the underlying distribution is positively skewed and is unlikely to be a Gaussian.
* Note the values of 1.000 down the diagonal. **Why is this?** The diagonal entries represent the correlation of each variable with itself and this is always equal to 1.
* **What happens to the correlation matrix? Why does this happen?** The main diagonal entries become blank rectangles. It happens because the filter excludes precisely the main diagonal entries.
* **Find the three largest positive correlations in the matrix. How do you interpret these results?** We count only once the symmetric entries above and below the diagonal. The three largest are 0.7724 (Labels and Bookcases), 0.7163 (Paper and Envelopes), and 0.6788 (Storage and Binders). The middle of these is reasonable enough: envelopes are not much use without paper to put in them. The other two correlations are not easily explained (other than by noting that this data is made up, and probably not enough care was put into the data generation process). As a bonus, it is noteworthy that there are no negatively correlated sub-categories.