Pearl Ayem – 34404160

**EOSC 410/510 Assignment 1**

**Problem 1:**

The following ranking were derived:

|  |  |  |
| --- | --- | --- |
| **Graph** | **Pearson** | **Spearman** |
| x-y | 0.580098 | 0.572420 |
| x2-y2 | 0.339721 | 0.572420 |
| x3-y3 | -0.901029 | 0.431895 |

Table 1: Pearson and Spearman ranking for 3 pair of datasets (1. x and y, x2 and y2, x3 and y3)

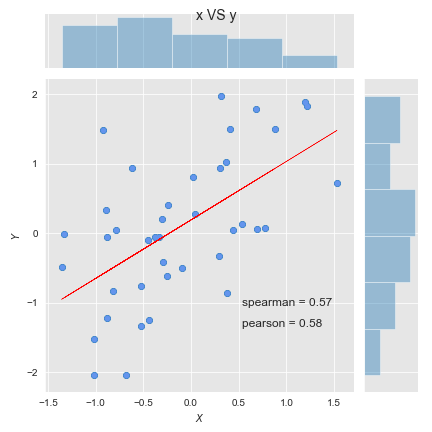


Figure 1: Regression between the x and y datasets. The Spearman and Pearson ranking is given on the graph.

When there is a linear relationship and no outliers the Pearson and Sprearman rank are almost the same. As seen in the first graph for x vs y (Figure 1), this is true for distributions that are generally normal. However, the Pearson correlation scored 0.01 higher than the Sprearman in this example for Figure 1.

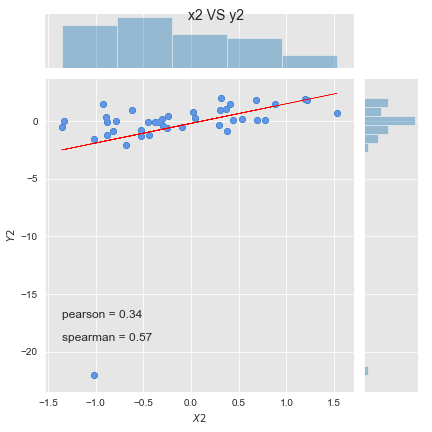


Figure 2: Regression between the x2 and y2 datasets. The Spearman and Pearson ranking is given on the graph.

In Figure 2 we notice that an outlier is added in the y2 dataset. This seems to throw off the Pearson correlation more than the Spearman. The Spearman raking was the same as that in Figure 1 (0.57 in both cases) whereas the Pearson ranking drops from 0.58 in Figure 1 to 0.34 in Figure 2 when an outlier in y2 is introduced. This would suggest that the Spearman ranking performs better when there is an outlier in the dependant variable; that is the Spearman is more robust to outliers.

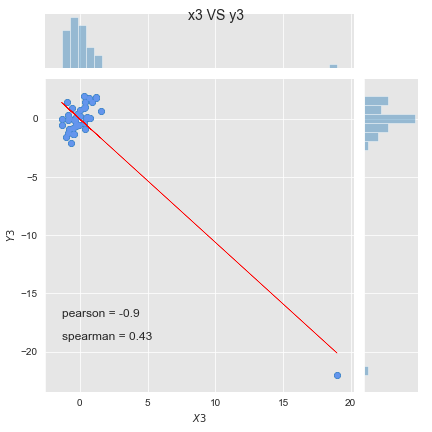


Figure 3: Regression between the x3 and y3 datasets. The Spearman and Pearson ranking is given on the graph.

In Figure 3, outliers exist in both the x3 and y3 datasets. In this case both rankings show lower correlation, the Spearman still showing one stronger than the Pearson. The Spearman ranking only fell from 0.57 in Figure 1 to 0.43 in Figure 3. The Pearson ranking however fell from 0.58 to -0.9, suggesting that an outlier in the dependent variable greatly affects the correlation. This would mean the Spearman ranking is usually more robust to outliers overall. The Pearson correlation coefficient measures the strength of the linear relationship between normally distributed variables. When the variables are not normally distributed or the relationship between the variables is not linear, it may be more appropriate to use the Spearman rank correlation method.

**Problem 2:**