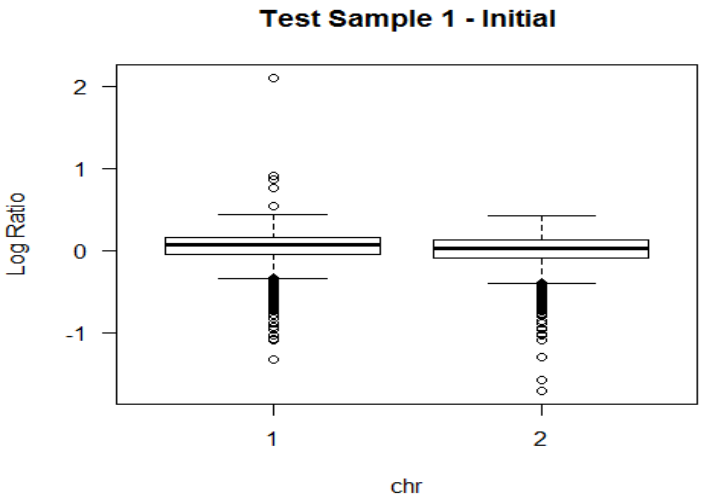
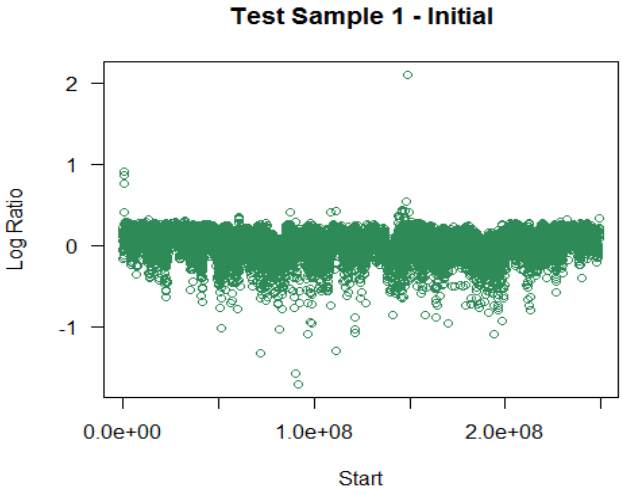
Implementation & Analysis of Basic Noise Reduction Function

Aim: To reduce the noise among the logRatio values of 2 samples in a given dataset.

Function Algorithm Details:

First a boxplot and scatter plot was plotted for Test Sample 1 to analyze the distribution of values. It looked as follows:





Clearly, the boxplot depicts a large number of negative outlier values (dots below horizontal line). But these cannot be considered as noise since they are present in large numbers.

Algorithm: Generally, a value is considered to be outlier if it greater/lesser than Quartile 1 or 3 by **1.5 times** the Inter Quartile Range (**IQR**). But as the above boxplot depicts, a large number of values are outliers and hence considering all of them as noise would result in ignoring a large set of data.

Hence, I came up with a factor of **2.5** (instead of 1.5 in normal case) after some manual regression to fit the data values efficiently. This factor of 2.5 can be modified by applying regression to suit the needs of the analyst.

As a next step, I deleted the values which were greater/lesser than Quartile 1 or 3 by 2.5 times the Inter Quartile Range (IQR) resulting in a more data set and having less noise (extreme values).

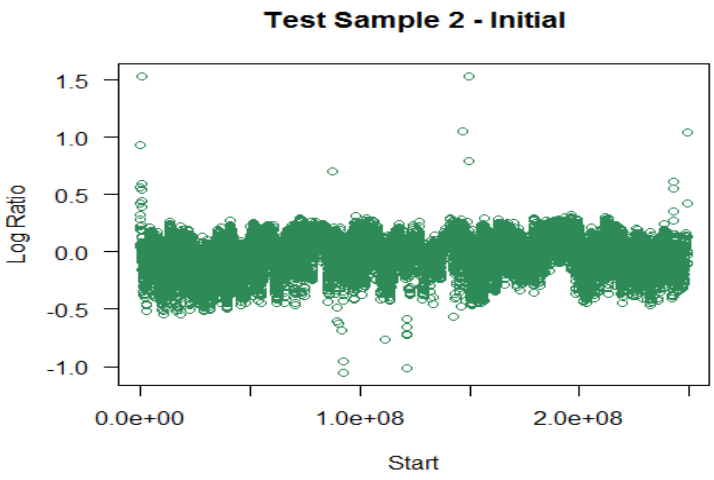
The deciding factor to choose 2.5 value was the **correlation coefficient** between Sample 1 and Sample 2.

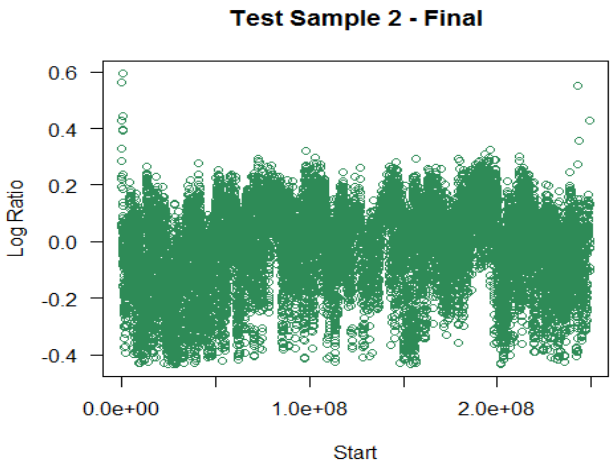
Correlation (Sample1, Sample2) Initially: **-0.6611315**

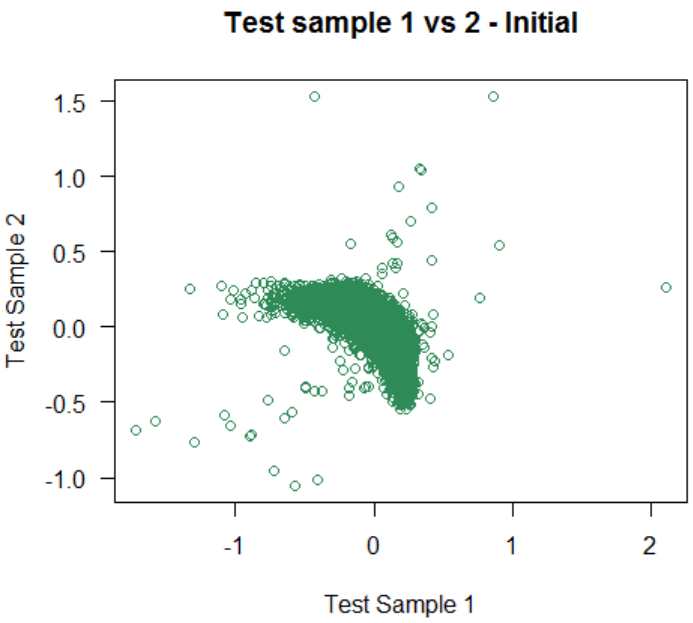
Correlation (Sample1, Sample2) Finally: **- 0.7562598**

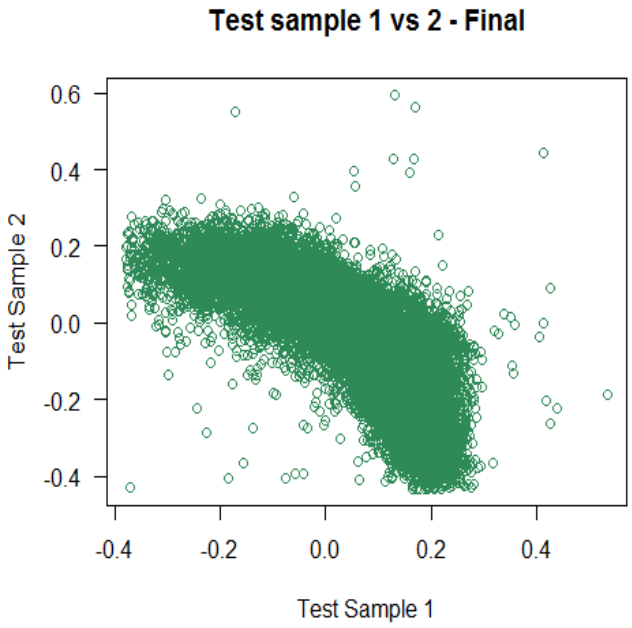
The negative sign indicating that an increase in one variable reliably predicts a decrease in the other one. As we can see, the **magnitude** of the correlation has **increased** after modifying the outlier factor (2.5) depicting an increased relation.

The results of this algorithm can be depicted by following plots:









As we can see,

- The data has become more compact
- The max distance of farthest points has reduced i.e. the magnitude of noise has reduced.
- The number of outlier (extreme) values has reduced.