**Some exploration on the raw data :**

Features 4-6 and 26-29 have about 70% of invalid data.

Features 23-25 have about 40% of invalid data

**Unfiltered data**

The radviz plot on the unfiltered dataset (no processing, just took a random subset of the data because otherwise it takes ages to plot) shows 3 clusters of points. 2 of them have an almost complete overlap of signal and background (color) but one seems to show a higher tendency for background data to have higher values of features number 16 and 17.

The parallel coordinates plot seems to show higher values of feature number 5 when the point corresponds to signal. I’m not sure why we don’t see that in the radviz plot.

**Removing rows with -999 in them**

However, when we remove all rows in which there is at least one -999, the plots change completely.

The radviz plot shows an almost complete overlap between signal and background while we still see higher values of feature 5 when the data corresponds to signal.

**Skewness transformation**

Applying the skewness correction transformation works quite well, as seen by comparing the symmetry of the histograms before and after the correction.

However both Radviz and parallel coordinate plots show no clear structure.

**Conclusion:**

Features 4-6 and 26-29 have about 70% of invalid data and features 23-25 have about 40% of invalid data

In principle we could remove features 4-6 and 26-29 but feature 5 seems to have some prediction power so we might want to think carefully before removing this particular feature.

Skewness correction transformation works well but removes all apparent structure in the data as seen in the parallel coordinate plots. This could be due just to the change of scale when comparing the y-axis on the parallel coordinate plots before and after transformation. Essentially maybe we see more of the outliers before the transformation because the distributions are so skewed. This needs to be checked too.

To do: Check more about how the parallel coordinates and radviz plots work

Check in more detail the effect of feature 5 on signal vs background