1)imported the dataset, removed NA to reduce it size from.. to 2756796

2) further reduced the size by taking only 2011-1016 data

3)rounded off the score to get 6 distinct drought scores

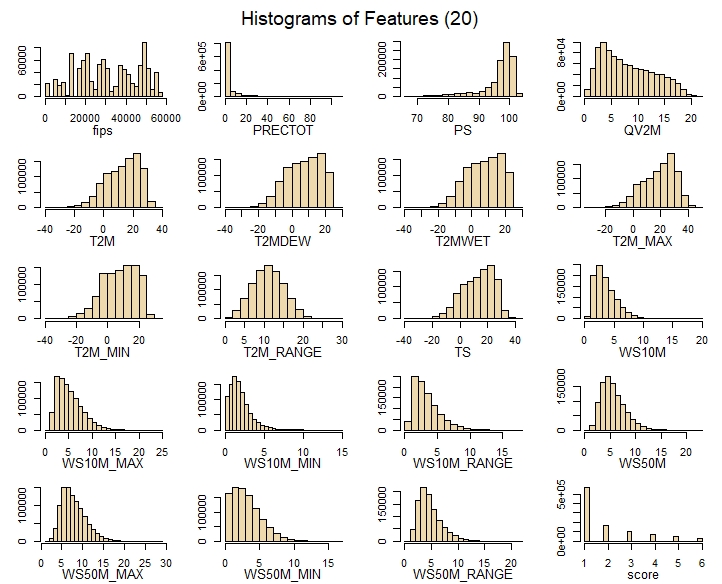
4)converted the score column to factor and fips to numeric, as c5.0 takes only factor predictor values

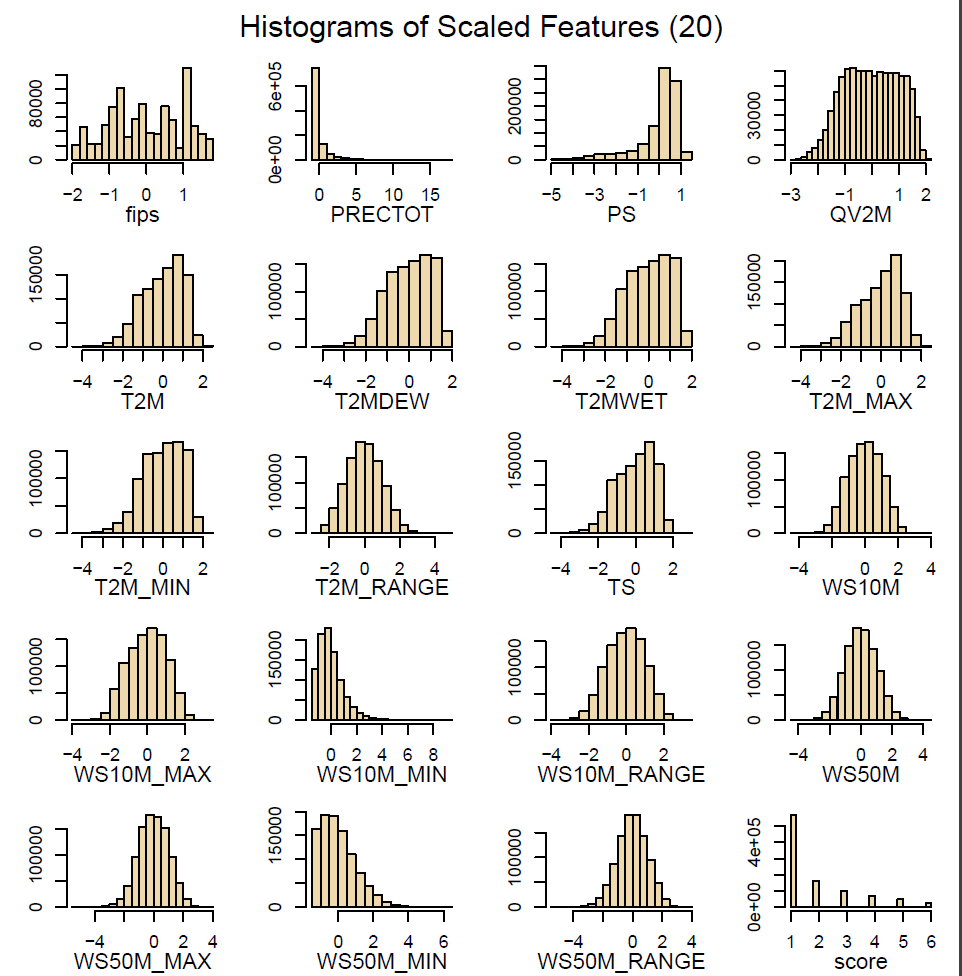
5)this factor score column is converted to numeric again while finding correlation table

6)EXPLORATORY DATA ANALYSIS:

1)Build histograms of all the features and realised that some columns are not normally distributed.

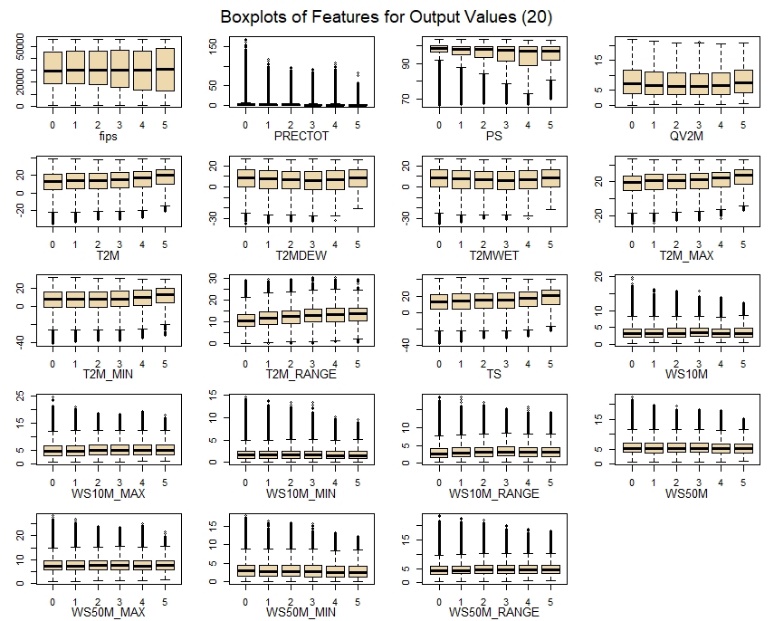
To solve this issue, we preprocessed our data using preProcess function's 'BoxCox' method along with center scaling.





2) Drew boxplots of each feature for score values and noticed outliers spread across each feature.

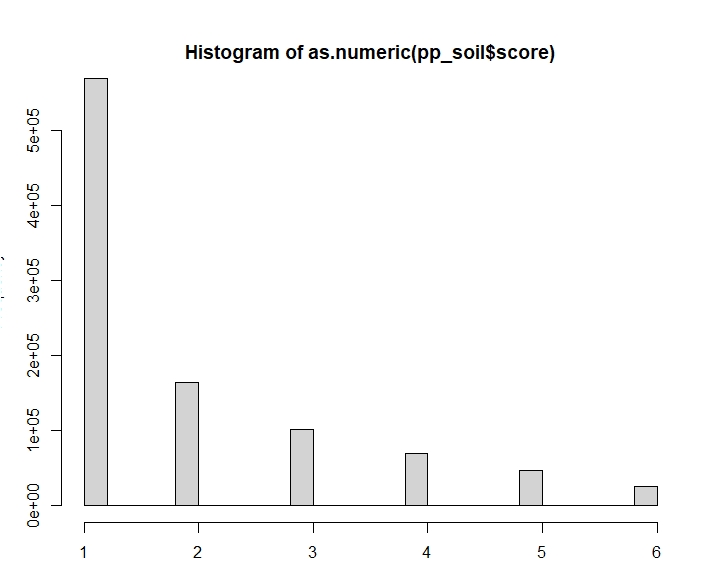
These outliers will not be removed because they are seen key features to be related to predict drought conditions

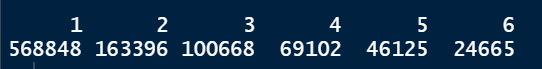


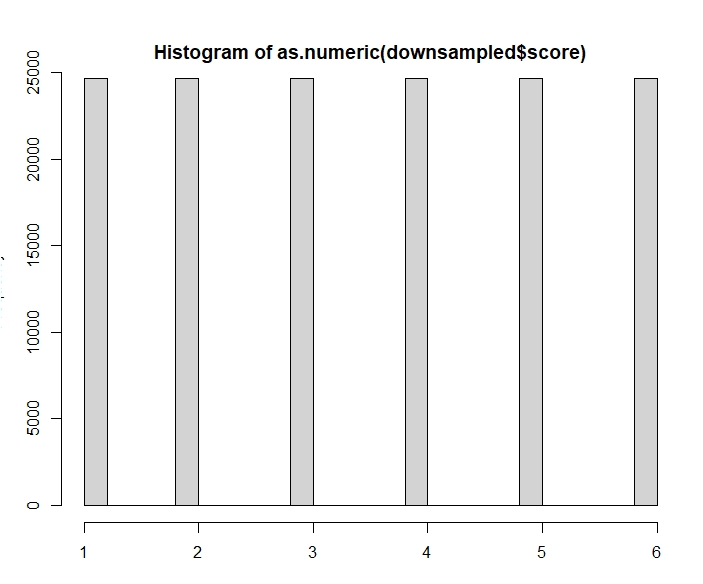
Other explorations to be added by Luke.

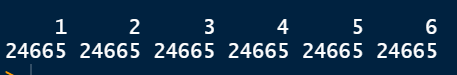
7) There is an imbalance in the dataspread, this is solved by using downsampling

downsampled <- downSample(pp\_soil[, -20], y = pp\_soil$score, list = FALSE)









8)MODELLING 🡪 TO BE ADDED