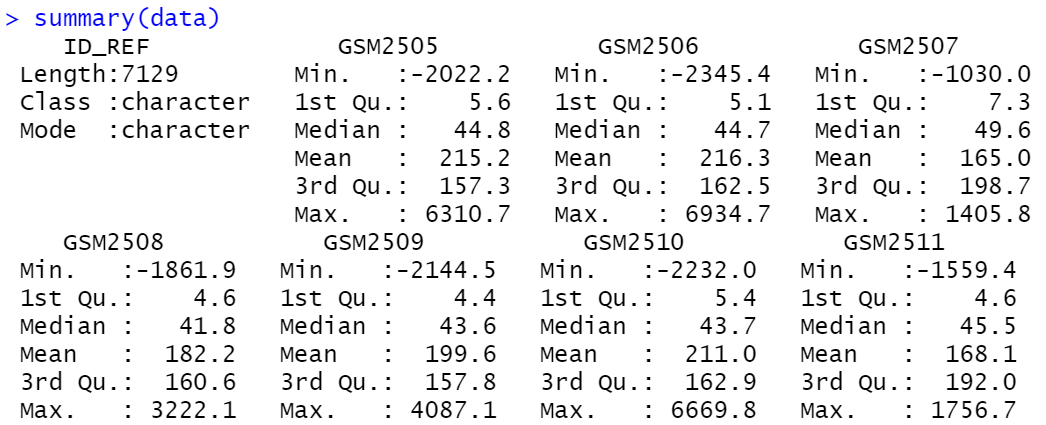
**Introduction:**

The data consists of bladder cancer data from 40 samples. Each of these samples record more than 5000 individual genes. The task at hand is classify the data to find which tumor stage of bladder cancer it is.

The data looks like so:



**Number of data samples:**

The data has only 40 samples so we cannot use 80-20 as it will give poor accuracy for training and testing of models. We can use “**k-fold cross validation**” since the number of samples is very less. In this case we can use 10-fold cross validation for model training and testing.

**Number of columns**:

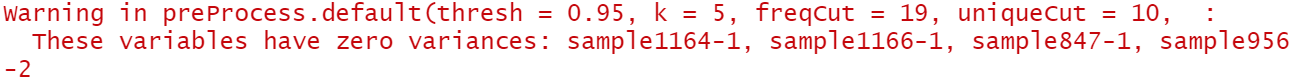
The data has more than 5000 individual genes. This effectively means there are many predictors variables which indicate which tumor stage is present in a sample. But we cannot effectively use all of them as some might be highly correlated and some might not be strong predictor variables. So we need to do **feature selection** to shortlist some of them.

**Random Forest**:

**Original data**

Tried to train random forest model with original data having 5000+ individual genes. This was problematic because the model:

1. Was taking more than 25 minutes training time,
2. Had a lot of variables have zero variance.



Since there are way too many variables, we need to remove variables which contribute less to the output variable.

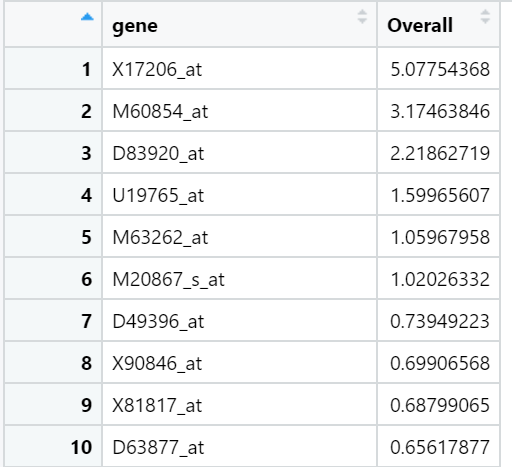
**Variable importance**

Used **‘Cor’** function to find the correlation between all the predictor variables. Created correlation matrix and removed the variables from the dataset that had more than 50% correlation.

We are then left with only 188 columns out of the original 5000+ columns.

Then trained Random forest model using 10-fold cross validation on this new data with 188 columns.

After a model is trained we can see the variable importance using Random Forest model training:



We observe that X17206\_at, M60854\_at and D83920\_at have high variable importance.

We summarize the model and we can observe an accuracy of 88%. This is when mtry is 2. Mtry is the number of variables for splitting at each node in each tree of the random forest. The function tries to create model with the value of parameter mtry as 2 to 188.

