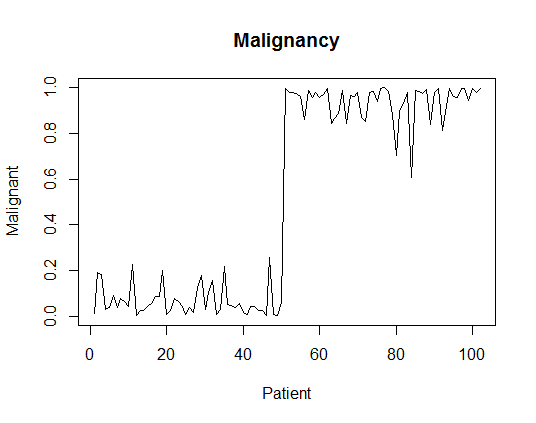
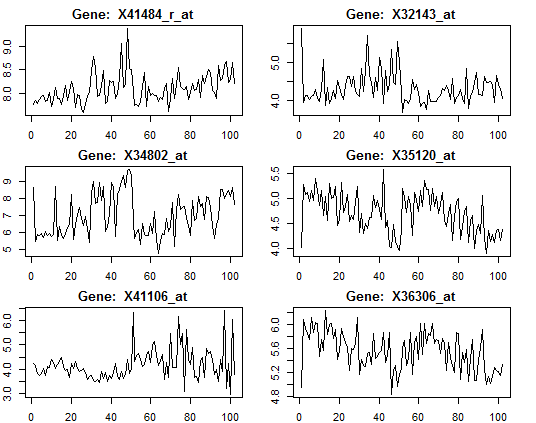
David Lowe

EDA #3

1. Malignant cancer can be tracked by gene expressions. Of the thousands of genes in humans, there is a subset that can be tracked to have triggered in people with malignant cancer. This study is to find those genes that are triggered when malignant cancer is present, and thereby understand what genes go into the uncontrolled cell growth of cancer. If these can be tracked, treatments can be developed to target these genes and stop their function in the overproduction of cells.
2. The data is structured with 5,149 genes (columns), and 102 observations from patients. Each has a rating between 0 and 1 as to how malignant their cancer is. The general trend of the data shows that half of the patients have fairly benign cancer, and the other have are rated as malignant (see plot). There are very few patients in the middleground between benign and malignant. With this pattern we can see if there are genes that jump at a similar point, that may give us understanding as to which genes affect malignant growth. The set of gene plots show a random sample of genes which seem to spike at the same point when patients switch from benign to malignant. There may be something to these genes that give insight into malignant cancer.
3. Since we are trying to see the relationship between malignancy and these genes, I would think something like a binomial regression could help, though there are far too many variables to successfully build a regression model. A binomial regression would produce estimates between 0 and 1, and would estimate the relationship between specific genes and malignancy. This would accomplish our goal, if there were not more genes than observations. This would be an overparameterized model.
4. I don’t know how to reduce the variables to a reasonable number. The number of genes far exceeds the number of patients.