KaitlinRMD

M. de Ferrante, K. Maciejewski, P. Batten

April 26, 2018

```
### Summary stuff for DS2 final project

library(mlbench)
data(BreastCancer)
attach(BreastCancer)
library(dplyr)
BreastCancer <- BreastCancer[,-1] # remove ID column
summary(BreastCancer) # note that everything is factor
```

```
##   Cl.thickness   Cell.size     Cell.shape  Marg.adhesion  Epith.c.size
##  1      :145   1      :384   1      :353   1      :407   2      :386
##  5      :130   10     : 67   2      : 59   2      : 58   3      : 72
##  3      :108   3      : 52   10     : 58   3      : 58   4      : 48
##  4      : 80   2      : 45   3      : 56   10     : 55   1      : 47
##  10     : 69   4      : 40   4      : 44   4      : 33   6      : 41
##  2      : 50   5      : 30   5      : 34   8      : 25   5      : 39
##  (Other):117   (Other): 81   (Other): 95   (Other): 63   (Other): 66
##   Bare.nuclei   Bl.cromatin  Normal.nucleoli    Mitoses             Class
##  1      :402   2      :166   1      :443     1      :579   benign   :458
##  10     :132   3      :165   10     : 61     2      : 35   malignant:241
##  2      : 30   1      :152   3      : 44     3      : 33
##  5      : 30   7      : 73   2      : 36     10     : 14
##  3      : 28   4      : 40   8      : 24     4      : 12
##  (Other): 61   5      : 34   6      : 22     7      :  9
##  NA's   : 16   (Other): 69   (Other): 69     (Other): 17
```

Above, we load in the data and remove the ID column as it is not needed.

The data used in this project is the `BreastCancer` data from `mlbench` library. It is from the Wisconsin Breast Cancer Database. Each variable except Class is loaded as 11 numerical factors with values ranging from 0 through 10. Class is benign or malignant, and this is the variable of interest. There are 16 missing values in bare nuclei, as seen in the summary above.

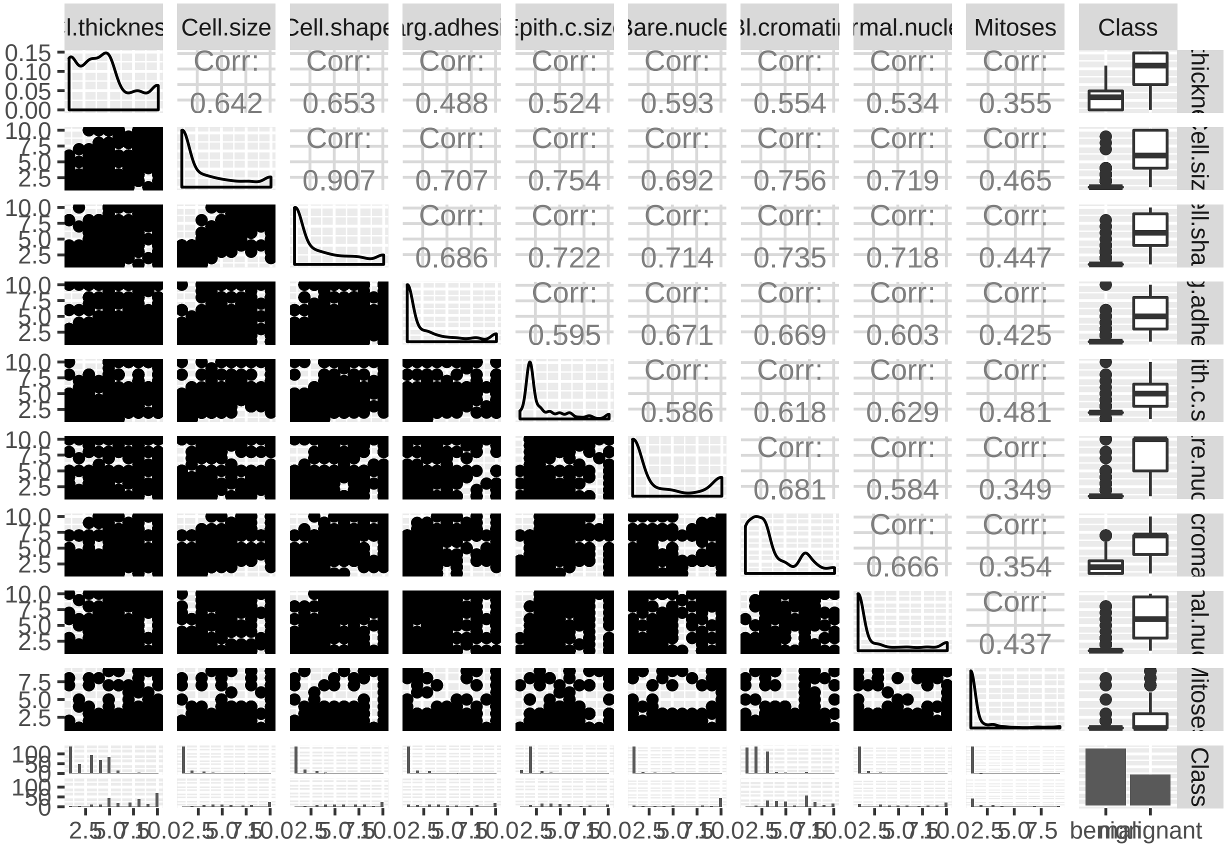
The variables included in the dataset are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, and Class. There are 458 subjects with benign growths and 241 with malignant.

Below, we convert the factors into numeric values and remove the NA's. Since there are few missing values compared to the number in the dataset, removing them should not effect our overall analysis power.

```
BreastCancer = BreastCancer %>%
  mutate(Cl.thickness=as.numeric(Cl.thickness)) %>%
  mutate(Cell.size=as.numeric(Cell.size)) %>%
  mutate(Cell.shape=as.numeric(Cell.shape)) %>%
  mutate(Marg.adhesion=as.numeric(Marg.adhesion)) %>%
  mutate(Epith.c.size=as.numeric(Epith.c.size)) %>%
  mutate(Bare.nuclei=as.numeric(Bare.nuclei)) %>%
  mutate(Bl.cromatin=as.numeric(Bl.cromatin)) %>%
  mutate(Normal.nucleoli=as.numeric(Normal.nucleoli)) %>%
  mutate(Mitoses=as.numeric(Mitoses)) %>% na.omit()
```

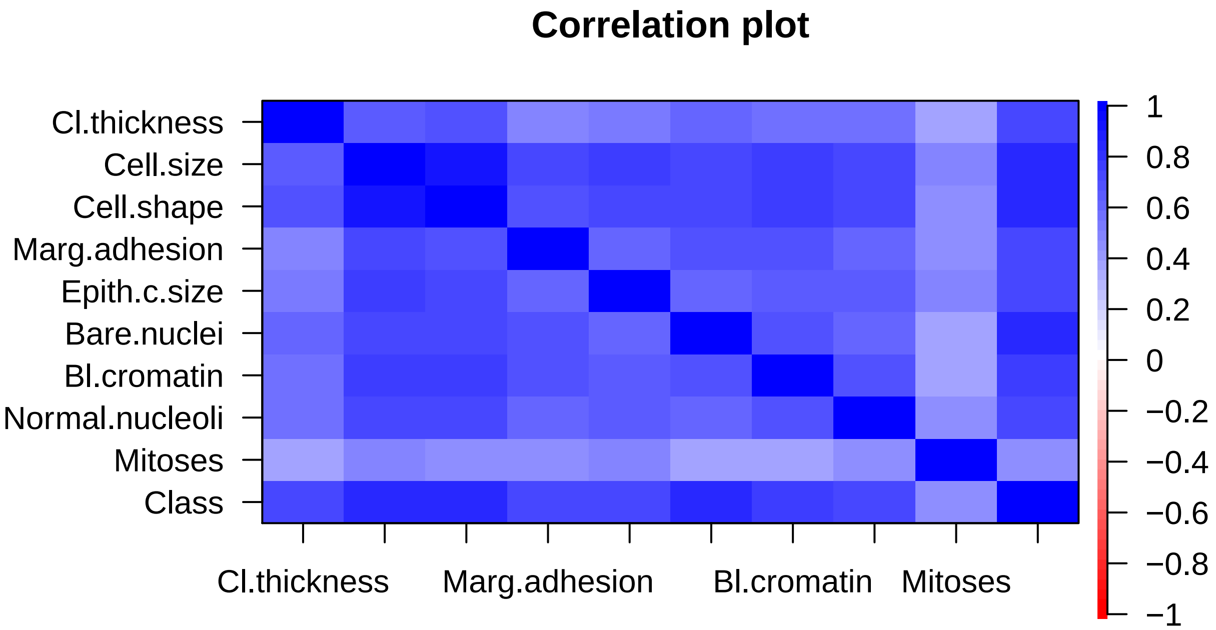
Summary plots

```
library(GGally)
ggpairs(BreastCancer) # all but response are numeric
```



The correlation plot below does not give much information. As expected, most measures have correlation with the outcome. Mitoses, however, has almost no corellation.

```
library(psych)
BreastCancer_num <- BreastCancer %>%
  mutate(Class = as.numeric(Class)-1) # numeric response
cor.plot(BreastCancer_num[,])
```

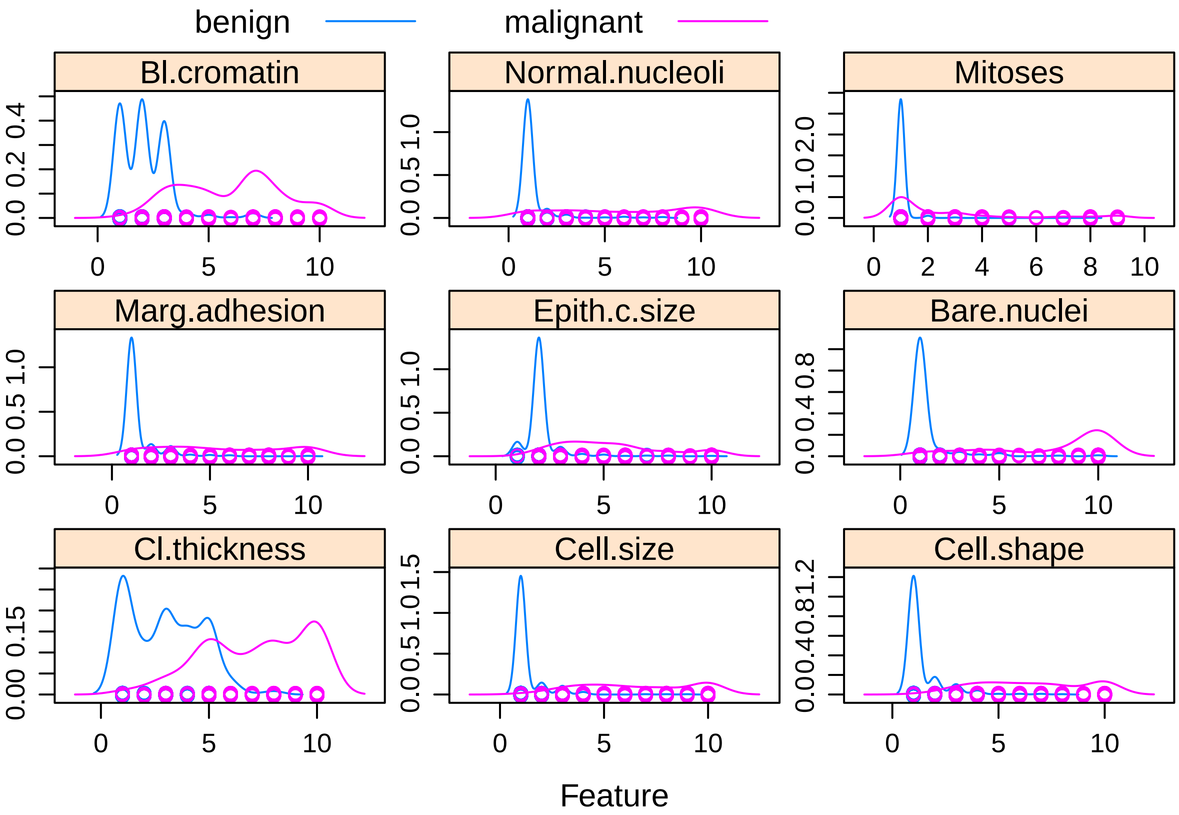


Below are the density plots for the variables. The blue line signifies benign and pink are malignant. We see that there are few malignant subjects with the following: normal nuclei, mitoses, marginal adhesion, single epithelial cell size, uniformity of cell size, uniformity of cell shape.

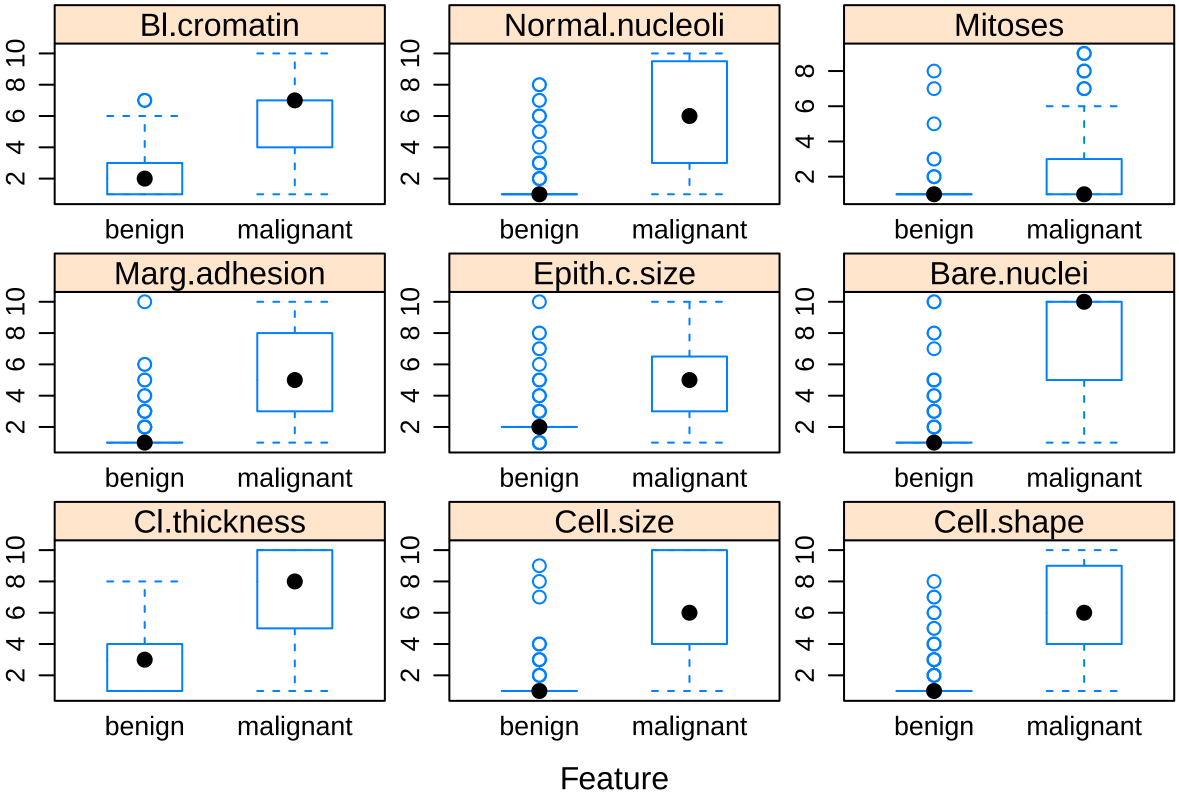
There are differences in the density plots of the following varaiables: bland chromatin, bare nucleoli, clump thickness.

The box plots below also show differences in the distributions for bland chromatin, bare nuclei, clump thickness.

```
featurePlot(x=BreastCancer[,-10], y=BreastCancer[,10],
            plot="density",
            scales=list(x=list(relation="free"),
                        y=list(relation="free")),
            auto.key=list(columns=3),
            layout=c(3,3))
```



```r
featurePlot(x=BreastCancer[,-10], y=BreastCancer[,10],
            plot="box",
            scales=list(x=list(relation="free"),
                        y=list(relation="free")),
            auto.key=list(columns=3),
            layout=c(3,3))
```



Logistic Analysis

Here is a generalized linear model with all variables included.

```
glm1 = glm(Class ~., data=BreastCancer,family=binomial)

summary(glm1)
```

```
##
## Call:
## glm(formula = Class ~ ., family = binomial, data = BreastCancer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4855  -0.1152  -0.0619   0.0222   2.4702
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -10.110096   1.173774  -8.613  < 2e-16 ***
## Cl.thickness      0.535256   0.141938   3.771 0.000163 ***
## Cell.size        -0.005943   0.209158  -0.028 0.977332
## Cell.shape        0.322136   0.230644   1.397 0.162510
## Marg.adhesion     0.330694   0.123462   2.679 0.007395 **
## Epith.c.size      0.096797   0.156568   0.618 0.536415
## Bare.nuclei       0.383015   0.093865   4.080 4.49e-05 ***
## Bl.cromatin       0.447401   0.171392   2.610 0.009044 **
## Normal.nucleoli   0.213074   0.112894   1.887 0.059109 .
## Mitoses           0.538551   0.325615   1.654 0.098138 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 884.35  on 682  degrees of freedom
## Residual deviance: 102.90  on 673  degrees of freedom
## AIC: 122.9
##
## Number of Fisher Scoring iterations: 8
```

At $\alpha = 0.05$ the following appear significant :

Cl.thickness

Marg.adhesion

Bare.nuclei

Bl.cromatin

We rerun the generalized linear model with only the significant values:

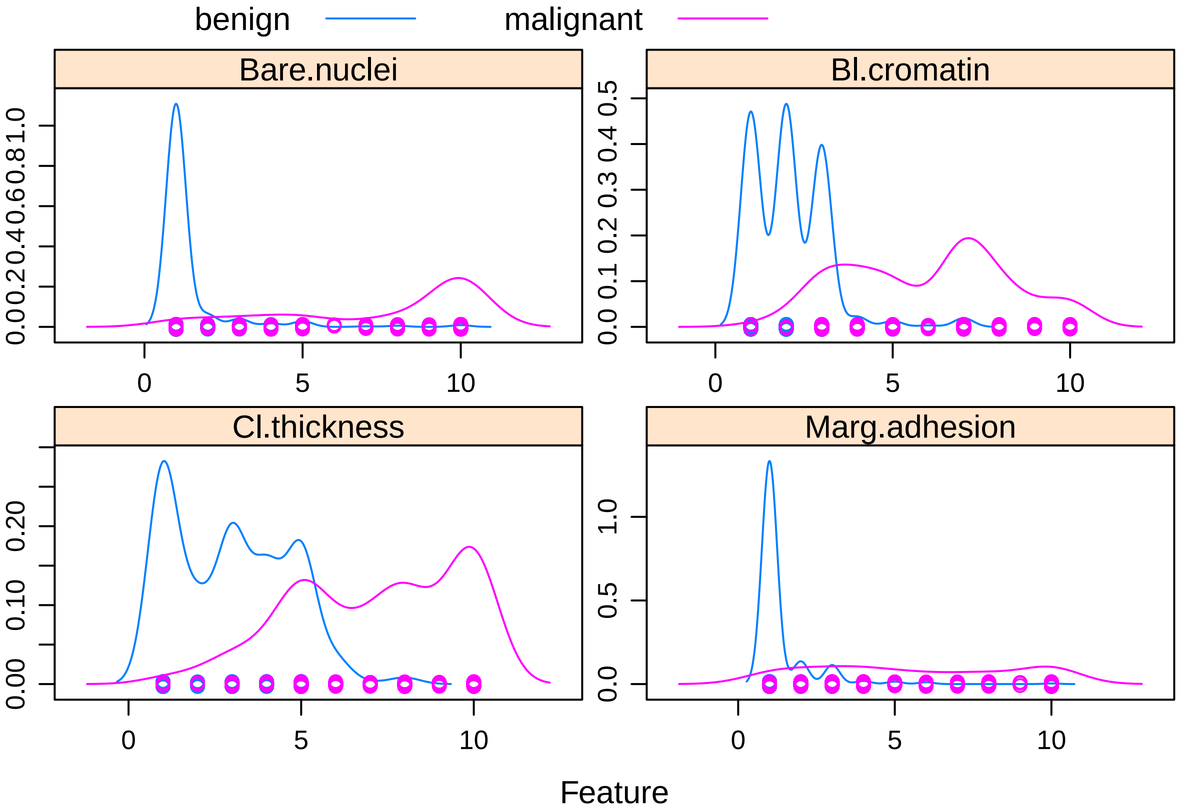
```r
glm2 = glm(Class ~ Cl.thickness + Marg.adhesion +
             Bare.nuclei + Bl.cromatin,
           data=BreastCancer,family=binomial)

summary(glm2)
```

```
##
## Call:
## glm(formula = Class ~ Cl.thickness + Marg.adhesion + Bare.nuclei +
##     Bl.cromatin, family = binomial, data = BreastCancer)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6964  -0.1451  -0.0609   0.0232   2.4476
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -10.11370    1.03264  -9.794  < 2e-16 ***
## Cl.thickness    0.81166    0.12585   6.450 1.12e-10 ***
## Marg.adhesion   0.43412    0.11403   3.807 0.000141 ***
## Bare.nuclei     0.48136    0.08816   5.460 4.76e-08 ***
## Bl.cromatin     0.70154    0.15196   4.616 3.90e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 884.35  on 682  degrees of freedom
## Residual deviance: 125.77  on 678  degrees of freedom
## AIC: 135.77
##
## Number of Fisher Scoring iterations: 8
```

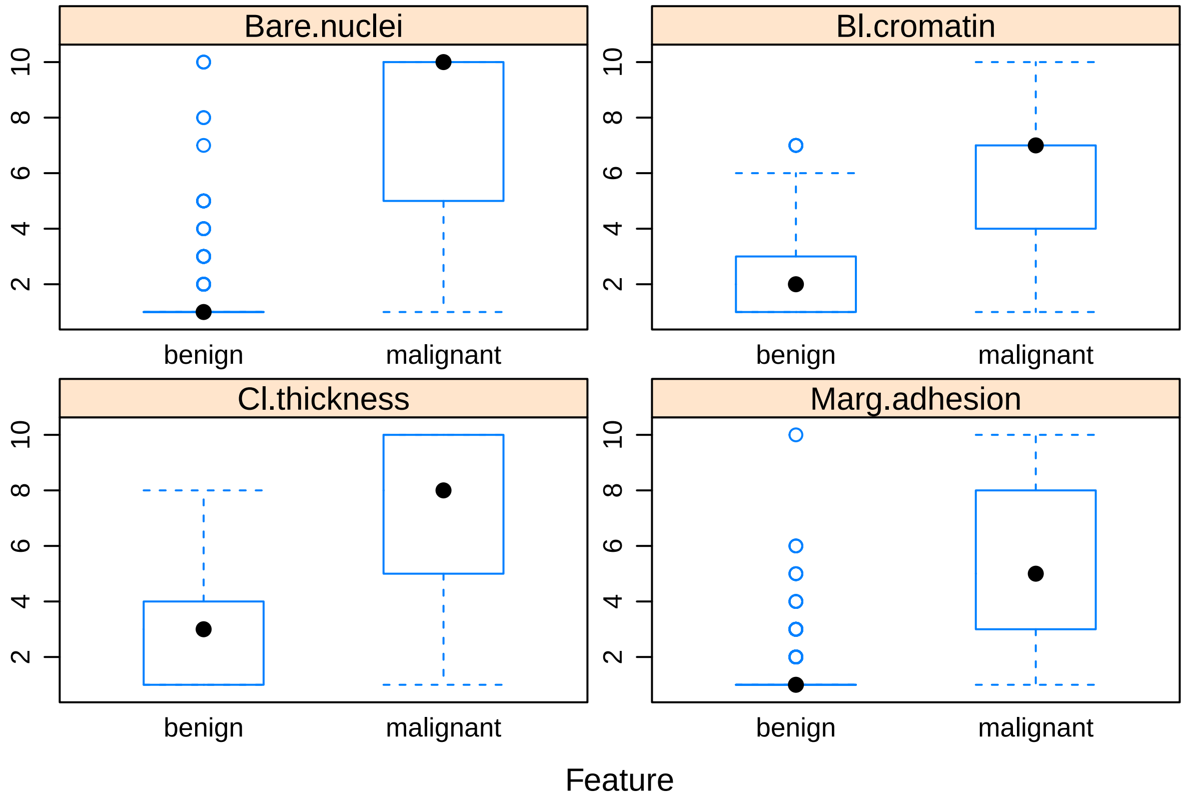
(conclusion/ explanation here)

Below are the feature plots for only the significant predictors, so that we may better see the difference in distributions between the variable and the outcome.

```
featurePlot(x=BreastCancer[,c(1,4,6,7)], y=BreastCancer[,10],
            plot="density",
            scales=list(x=list(relation="free"),
                        y=list(relation="free")),
            auto.key=list(columns=3))
```



```
featurePlot(x=BreastCancer[,c(1,4,6,7)], y=BreastCancer[,10],
            plot="box",
            scales=list(x=list(relation="free"),
                        y=list(relation="free")),
            auto.key=list(columns=3))
```



Now we split our data into a training and test set so we can make predictions

```
set.seed(1)
BreastCancer.train <- sample(1:nrow(BreastCancer), 410)

BreastCancer.test=BreastCancer[-BreastCancer.train,] # test

Class.test=BreastCancer$Class[-BreastCancer.train]

glm.fits=glm(Class ~ Cl.thickness + Marg.adhesion +
             Bare.nuclei + Bl.cromatin,
           data=BreastCancer,family=binomial, subset=BreastCancer.train)

glm.probs = predict(glm.fits, BreastCancer.test, type = "response")
glm.pred=rep("benign",273)
glm.pred[glm.probs >.5]="malignant"
table(glm.pred, Class.test)
```

```
##              Class.test
## glm.pred     benign malignant
##   benign        176         8
##   malignant       5        84
```

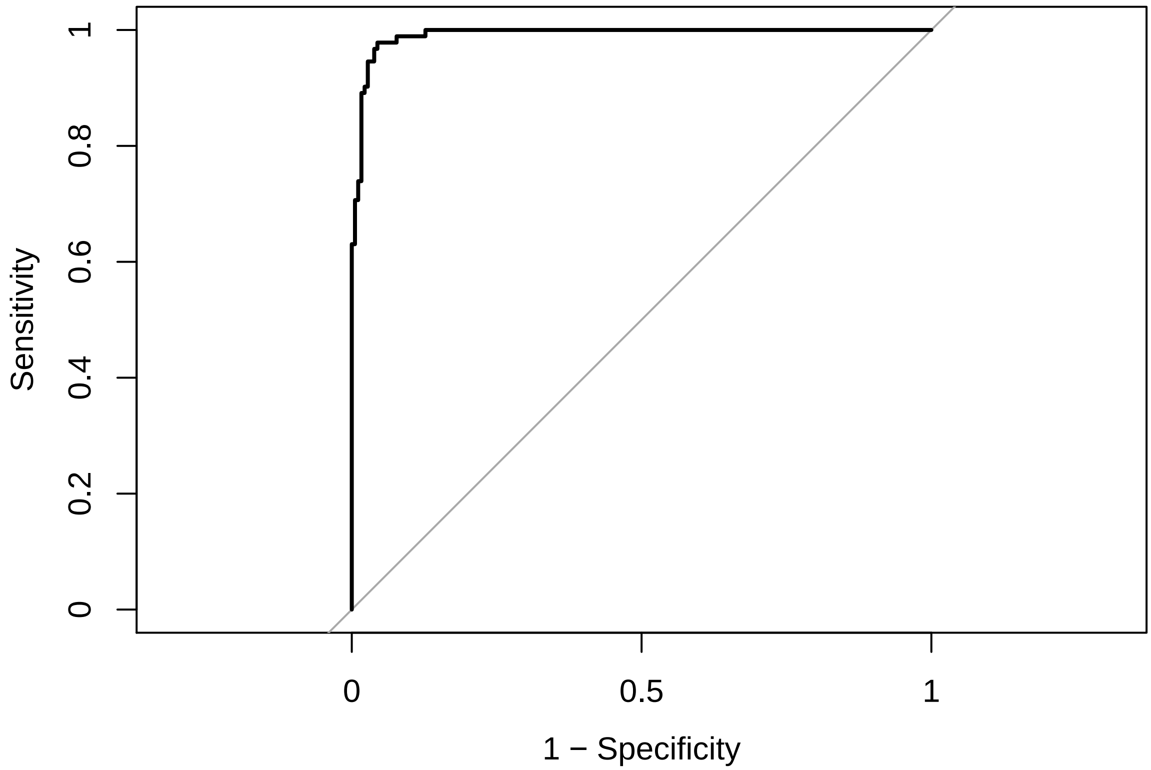
```
library(caret)
confusionMatrix(glm.pred, Class.test, positive = "malignant")
```

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction  benign malignant
##    benign       176         8
##    malignant      5        84
##
##                Accuracy : 0.9524
##                  95% CI : (0.9199, 0.9744)
##     No Information Rate : 0.663
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8926
##  Mcnemar's Test P-Value : 0.5791
##
##             Sensitivity : 0.9130
##             Specificity : 0.9724
##          Pos Pred Value : 0.9438
##          Neg Pred Value : 0.9565
##              Prevalence : 0.3370
##          Detection Rate : 0.3077
##    Detection Prevalence : 0.3260
##       Balanced Accuracy : 0.9427
##
##        'Positive' Class : malignant
##
```

```r
mean(glm.pred == Class.test)
```

```
## [1] 0.952381
```

```r
library(pROC)

roc.glm.train <- roc(BreastCancer.test$Class, glm.probs,
               levels = c("benign", "malignant"))
plot(roc.glm.train, legacy.axes = TRUE)
```



```
auc(roc.glm.train)
```

```
## Area under the curve: 0.9917
```

There were only 13 incorrect predictions; 5 benign were predicted to be malignant and 8 that were truly malignant were predicted to be benign. Logistic has 95% correct response for test, which is pretty good.

The area under the ROC curve is 99%.

Sensitivity is 91%

Specificity is 97%

PPV is 94% and NPV is 96%