IE 5080 - EDA and Determination of Policy Form

```r
# run variable selection
source(path) # run the variable selection function
```

```
## Loading required package: Matrix
```

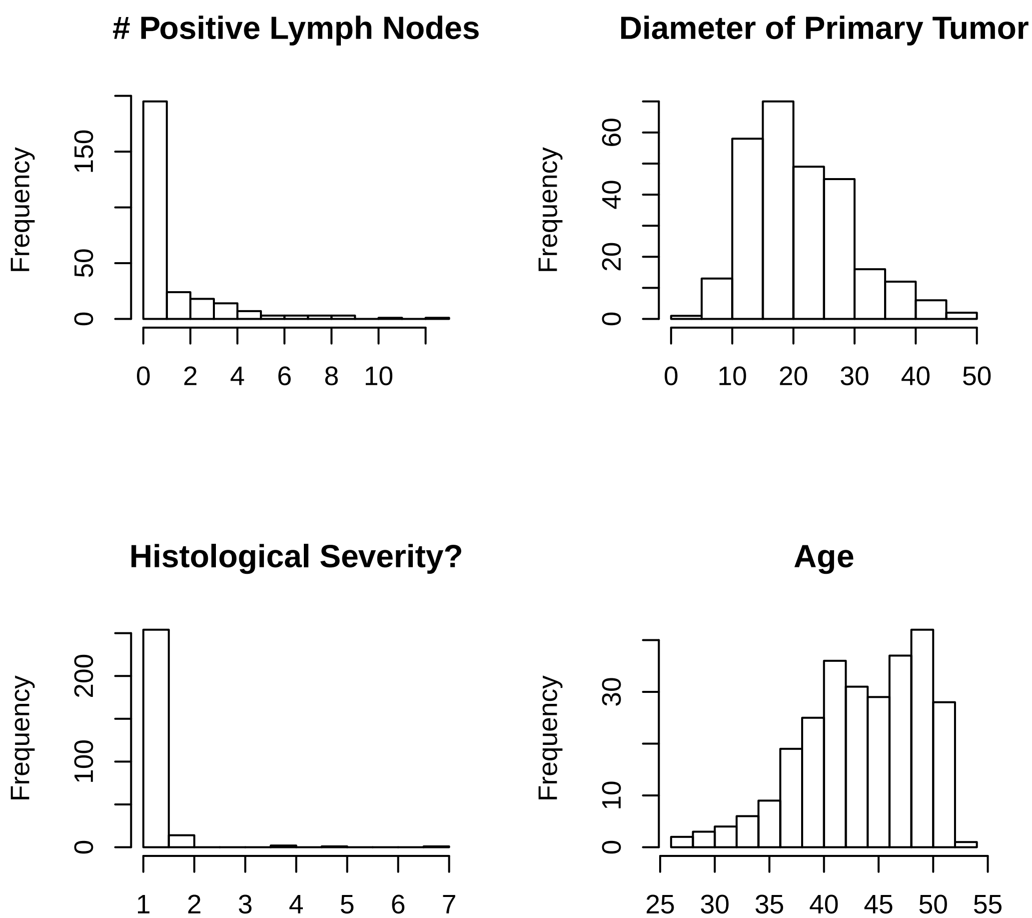
```
## Loaded glmnet 3.0-1
```

```r
set.seed(1234) # this seed results in a dataset with 28 variables
df = variable_select()$newdata
```

For practical purposes we would like to construct our policy form using only the metadata. Since the imaging data has been normalized prior to obtaining possession, a policy form using abstracted information based on characteristics of imaging is less interpretable for clinicians. So, we have five candidate variables:

- posnodes (# positive lymph nodes)
- diam (diameter of primary tumor)
- hist (a measure of the severity of the histology?)
- age
- grade (histological grade)

```r
# eda for metadata
par(mfrow=c(2,2))
hist(df$posnodes, main = "# Positive Lymph Nodes", xlab="")
hist(df$diam, main = "Diameter of Primary Tumor", xlab="")
hist(df$histtype, main = "Histological Severity?", xlab="")
hist(df$age, main = "Age", xlab="")
```




```
table(df$grade)
```

```
##
##   1   2   3
##  71  95 106
```

```
table(df$histtype)
```

```
##
##   1   2   4   5   7
## 254  14   2   1   1
```

```
table(df$posnodes)
```

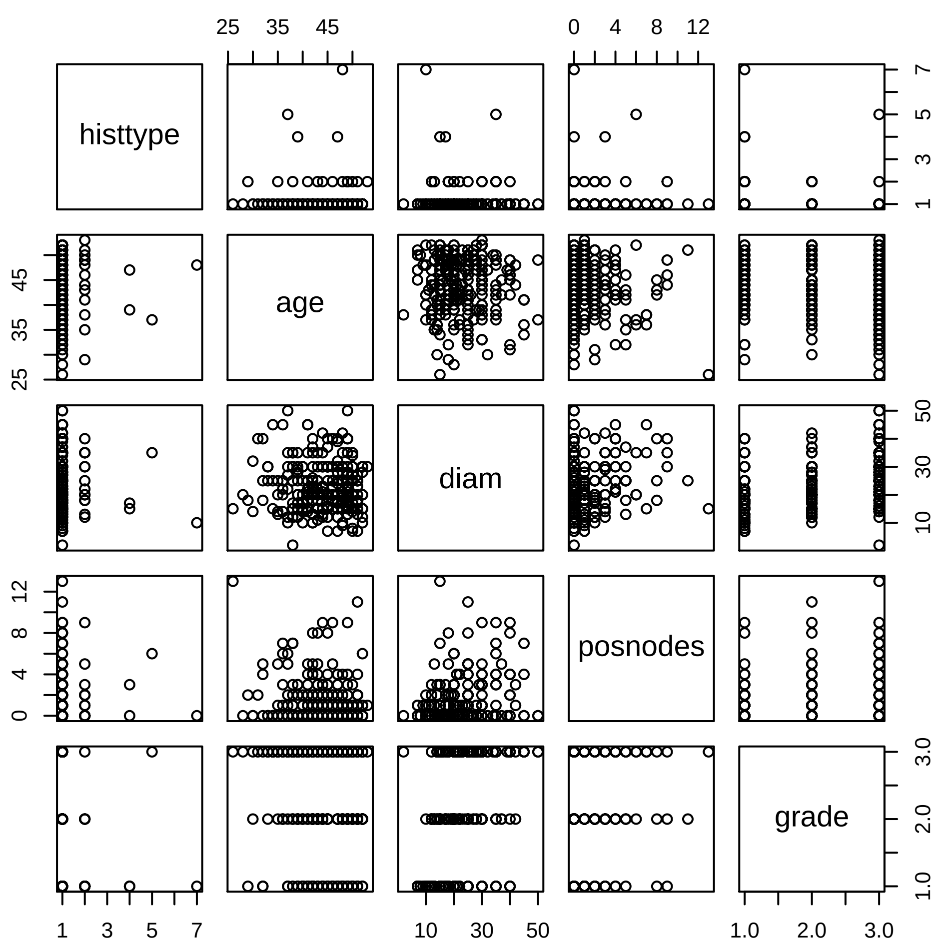
```
##
##   0   1   2   3   4   5   6   7   8   9  11  13
## 137  58  24  18  14   7   3   3   3   3   1   1
```

- posnodes contains a lot of zeros so any policy form with posnodes >0 will automatically exclude 137 of the patients
- histtype contains 254 1s, so there is not enough variation to use it to determine treatment policy
- age, diameter, and grade seem like good candidates

```r
metavars = df[8:12]
cor = cor(metavars, method="spearman")
cor
```

```
##                histtype          age        diam    posnodes       grade
## histtype  1.00000000  0.019264075 0.033572695  0.05924833 -0.22221437
## age       0.01926407  1.000000000 0.006178775 -0.03597298 -0.09107382
## diam      0.03357269  0.006178775 1.000000000  0.15083359  0.35613078
## posnodes  0.05924833 -0.035972981 0.150833586  1.00000000 -0.01546974
## grade    -0.22221437 -0.091073825 0.356130779 -0.01546974  1.00000000
```

```r
plot(metavars)
```



Grade and age are moderately correlated (r = 0.36).

```r
# run glm to determine which vars most influence death
# don't want to control for treatment
df_for_fit = df[, c(2, 8:12, 14:ncol(df))]
m1 = glm(eventdeath ~ ., data=df_for_fit, family=binomial)
summary(m1) # none of the metadata significant
```

```
##
```

```
## Call:
## glm(formula = eventdeath ~ ., family = binomial, data = df_for_fit)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2167  -0.5844  -0.2719   0.4581   2.6667
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.287896   1.934703  -0.666 0.505615
## histtype       -0.168745   0.509058  -0.331 0.740278
## age            -0.038230   0.033489  -1.142 0.253641
## diam            0.022732   0.021692   1.048 0.294672
## posnodes        0.036908   0.081527   0.453 0.650760
## grade           0.497738   0.317052   1.570 0.116440
## NM_000926      -0.770172   0.696014  -1.107 0.268490
## NM_003258      -0.770176   0.857845  -0.898 0.369290
## NM_012067      -0.810431   0.512377  -1.582 0.113716
## NM_003430      -2.794452   1.119411  -2.496 0.012548 *
## AL117418       -0.591907   0.860750  -0.688 0.491664
## NM_006096       0.609843   0.804132   0.758 0.448219
## Contig23211_RC  3.049363   1.200824   2.539 0.011105 *
## NM_016109       1.386623   0.736864   1.882 0.059865 .
## AL049265        0.941092   0.859317   1.095 0.273445
## Contig55725_RC -1.099418   0.652872  -1.684 0.092187 .
## NM_016359       3.814783   1.132880   3.367 0.000759 ***
## Contig48913_RC  1.682878   1.204535   1.397 0.162378
## NM_001109       0.343163   0.878435   0.391 0.696054
## NM_001124      -0.004492   0.728294  -0.006 0.995079
## NM_001333       0.343926   0.771656   0.446 0.655815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 324.14  on 271  degrees of freedom
## Residual deviance: 206.49  on 251  degrees of freedom
## AIC: 248.49
##
## Number of Fisher Scoring iterations: 6
```

Indeed, histtype and posnodes have the largest pvalues of the 5 metadata variables.

Next try a model with only the metadata:

```r
m2 = glm(eventdeath ~ histtype + age + diam + posnodes + grade, data=df, family=binomial)
summary(m2)
```

```
##
## Call:
## glm(formula = eventdeath ~ histtype + age + diam + posnodes +
##     grade, family = binomial, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5807  -0.7595  -0.4980   1.0345   2.4505
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.44220    1.39115  -1.756   0.0792 .
## histtype     0.25291    0.30125   0.840   0.4012
## age         -0.04391    0.02627  -1.671   0.0947 .
## diam         0.02613    0.01769   1.477   0.1396
## posnodes     0.04632    0.06658   0.696   0.4866
## grade        1.08274    0.22537   4.804 1.55e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 324.14  on 271  degrees of freedom
## Residual deviance: 278.79  on 266  degrees of freedom
## AIC: 290.79
##
## Number of Fisher Scoring iterations: 4
```

Once again, age, diam, and grade have the smallest p-values. Surprisingly, older ages correspond to higher death rates.

```
car::vif(m2)
```

```
## histtype      age     diam posnodes    grade
## 1.056702 1.016003 1.101816 1.041062 1.132890
```

```
m3 = glm(eventdeath ~ age + diam + grade, data=df, family=binomial)
summary(m3)
```

```
##
## Call:
## glm(formula = eventdeath ~ age + diam + grade, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5688  -0.7449  -0.4764   1.0152   2.4287
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.95285    1.30328  -1.498   0.1340
## age         -0.04644    0.02605  -1.783   0.0746 .
## diam         0.02871    0.01748   1.642   0.1005
## grade        1.04169    0.21979   4.739 2.14e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 324.14  on 271  degrees of freedom
## Residual deviance: 279.98  on 268  degrees of freedom
## AIC: 287.98
##
## Number of Fisher Scoring iterations: 4
```

```r
car::vif(m3)
```

```
##      age     diam    grade
## 1.005466 1.071803 1.075243
```

Based on these results, I think we should define the policy form based on three different sets of variables:

- age, diam, grade
- age, diam, grade, NM_003430, Contig23211_RC, NM_016359
- NM_003430, Contig23211_RC, NM_016359

The coefficient values for the variables were:

- positive for diam, grade, Contig23211_RC, and NM_016359
- negative for age (surprisingly), and NM_003430