Tenzin_Gyaltsen_BTC1877_Assignment_2

T.G

2024-10-07

A. Regression

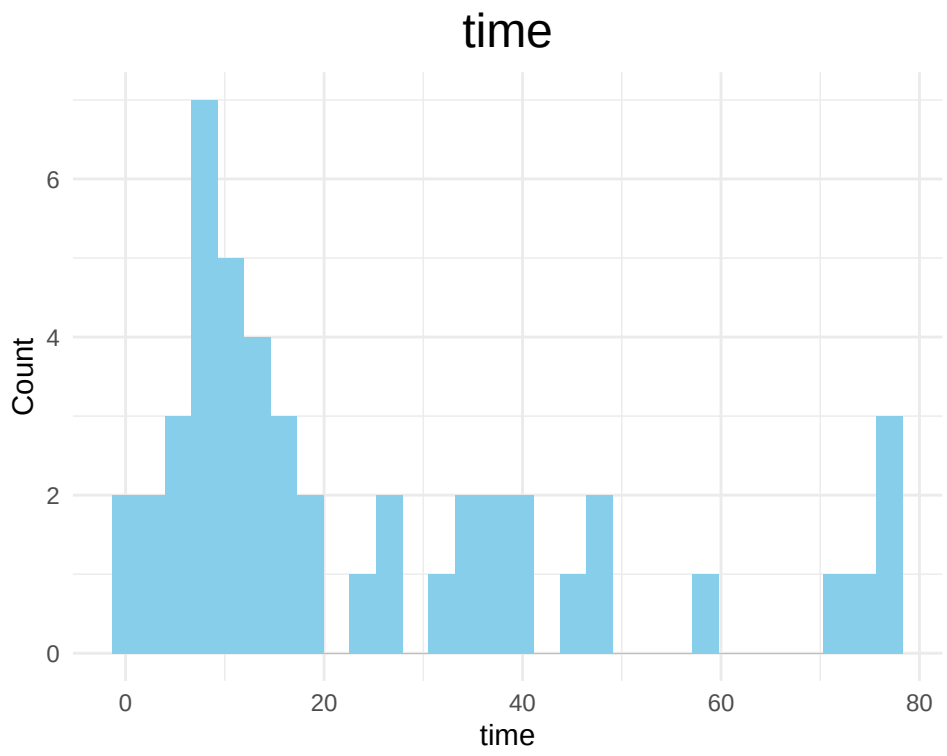
A1 - Read and explore the data, assessing and assigning NAs. Also rename variables.

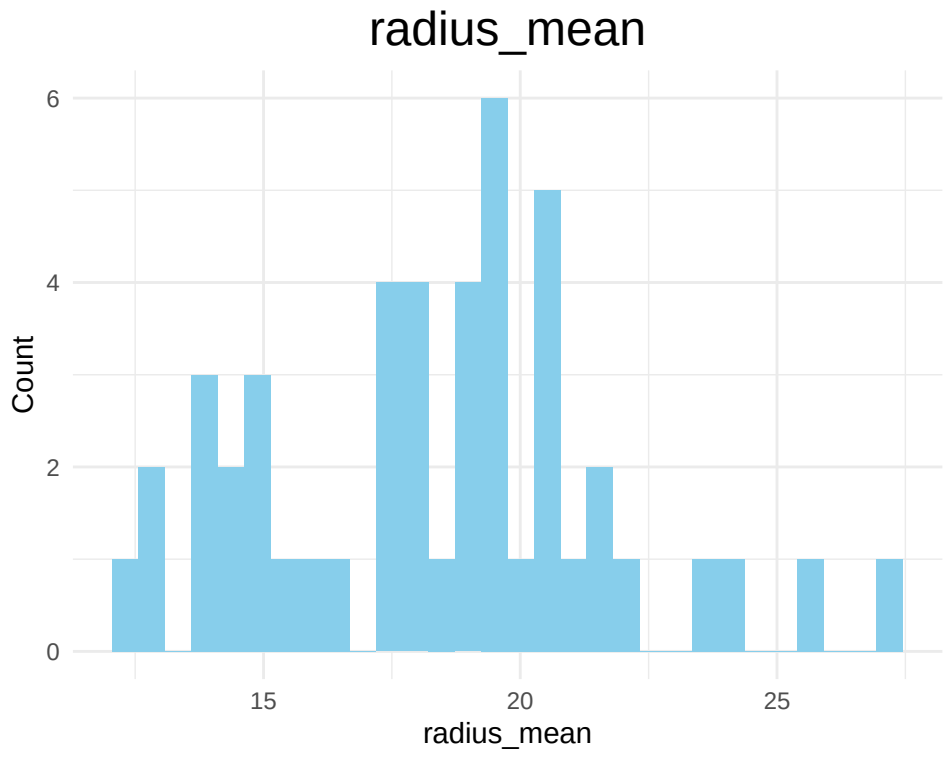
A2i - Refactor lymph nodes variable based on assignment criteria:

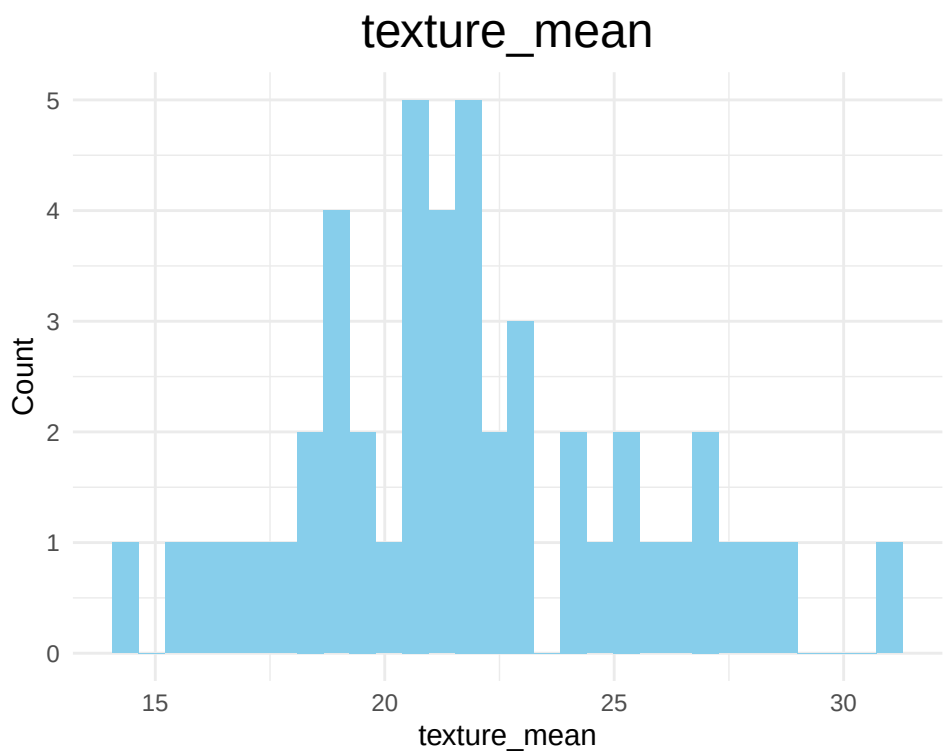
A2ii - Generate descriptive statistics for categorical and numerical variables.

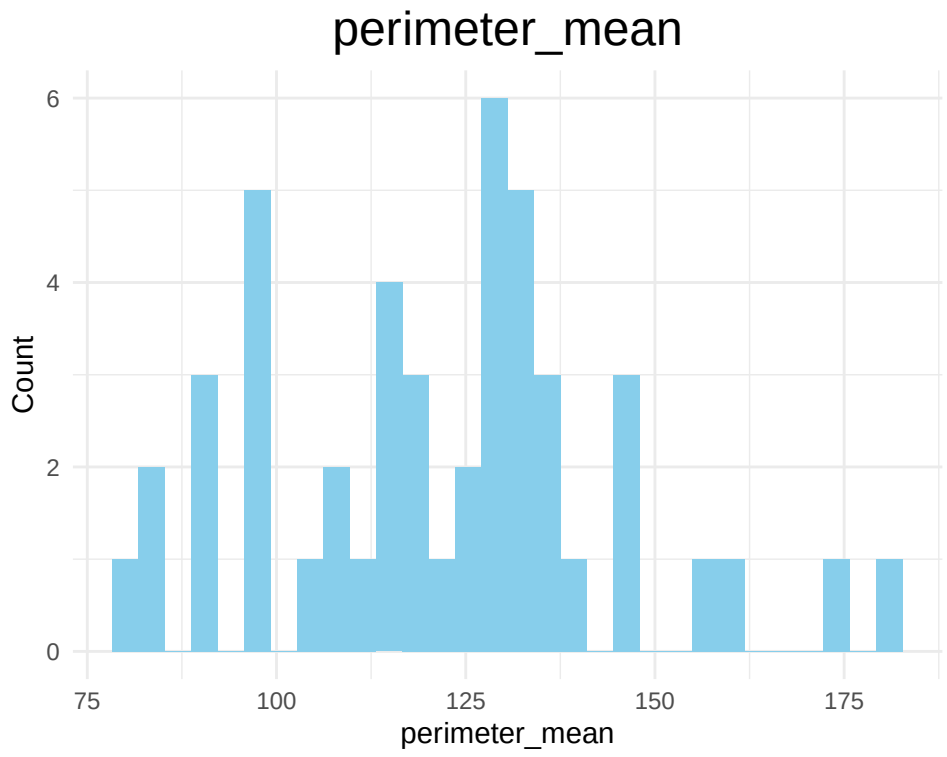
```
# A tibble: 12 x 7
   Variable                  mean    median        sd      min     max     iqr
   <chr>                    <dbl>     <dbl>     <dbl>    <dbl>   <dbl>   <dbl>
 1 time                     25.1      16       22.7       1      7.8 e+1 2.75e+1
 2 radius_mean              18.4      19        3.36     12.3    2.72e+1 4.47e+0
 3 texture_mean             21.8      21.5      3.66     14.3    3.10e+1 4.83e+0
 4 perimeter_mean          122.      124.      22.9      81.2    1.82e+2 2.88e+1
 5 area_mean              1090.     1104      396.      477.     2.25e+3 4.89e+2
 6 smoothness_mean           0.103     0.103    0.0102    0.0822 1.21e-1 1.76e-2
 7 compactness_mean          0.143     0.134    0.0410    0.0672 2.36e-1 5.31e-2
 8 concavity_mean            0.163     0.166    0.0625    0.0525 3.37e-1 1.00e-1
 9 concave_points_mean       0.0939    0.0899   0.0345    0.0333 1.91e-1 4.15e-2
10 symmetry_mean             0.188     0.187    0.0209    0.142  2.36e-1 2.60e-2
11 fractal_dimension_mean    0.0613    0.0608   0.00625   0.0503 7.45e-2 8.7 e-3
12 tumour_size               3.46      3        2.01      0.4    1   e+1 1.6 e+0
```

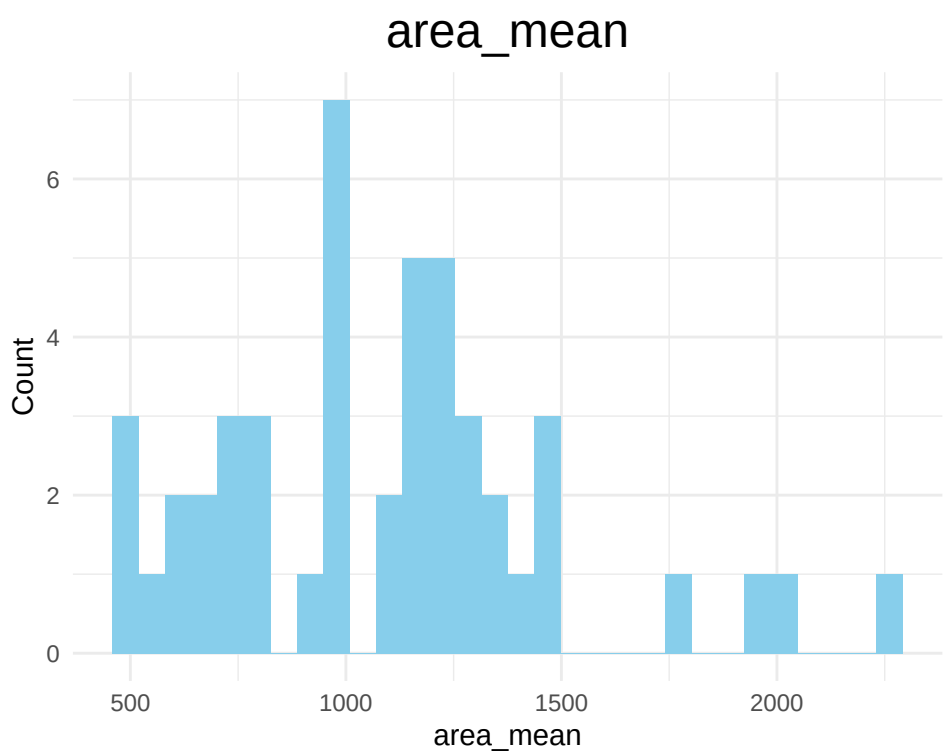
```
  lymph_nodes  n proportion
1           0 12  0.2553191
2         1-3 12  0.2553191
3   4 or more 22  0.4680851
4        <NA>  1  0.0212766
```

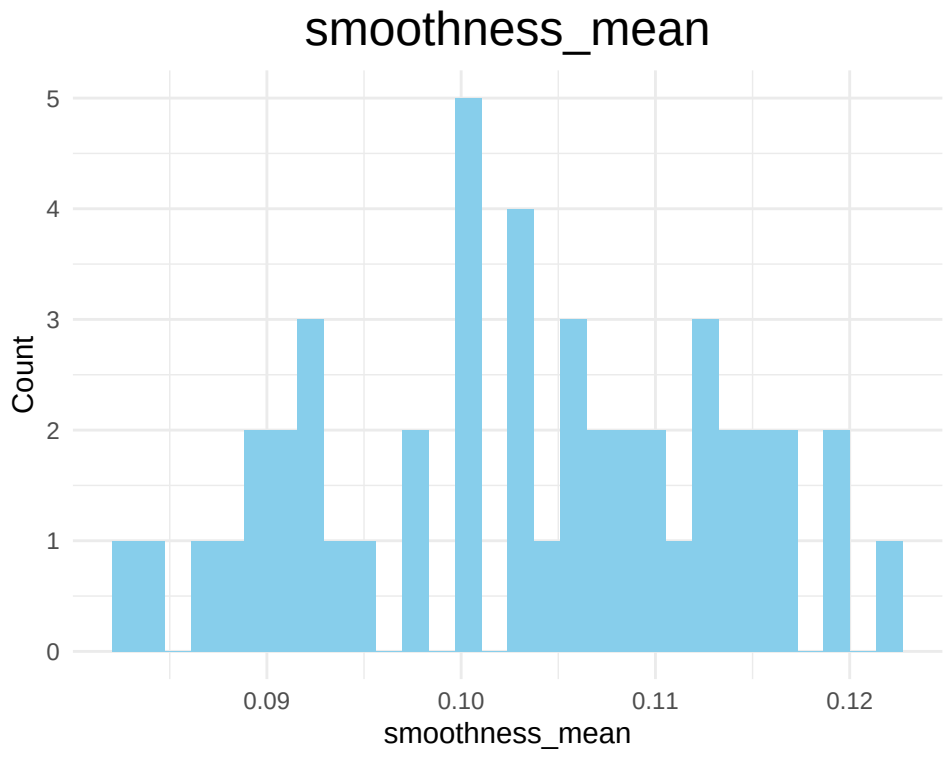

time
Count
6
4
2
0
0
20
40
60
80
time

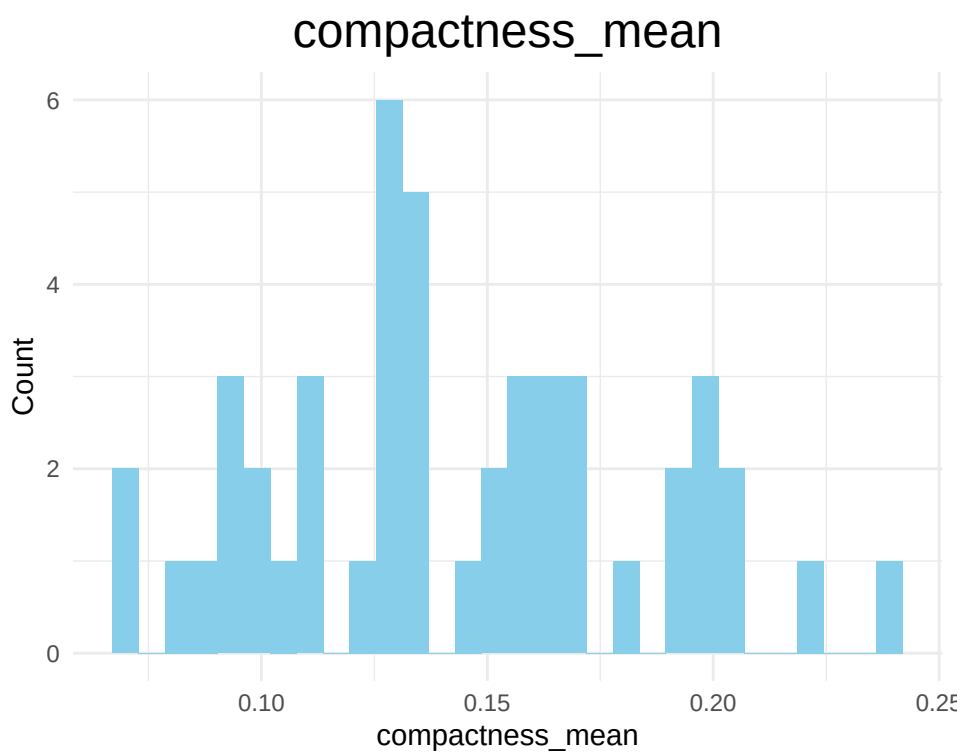

radius_mean
Count
6
4
2
0
15
20
25
radius_mean

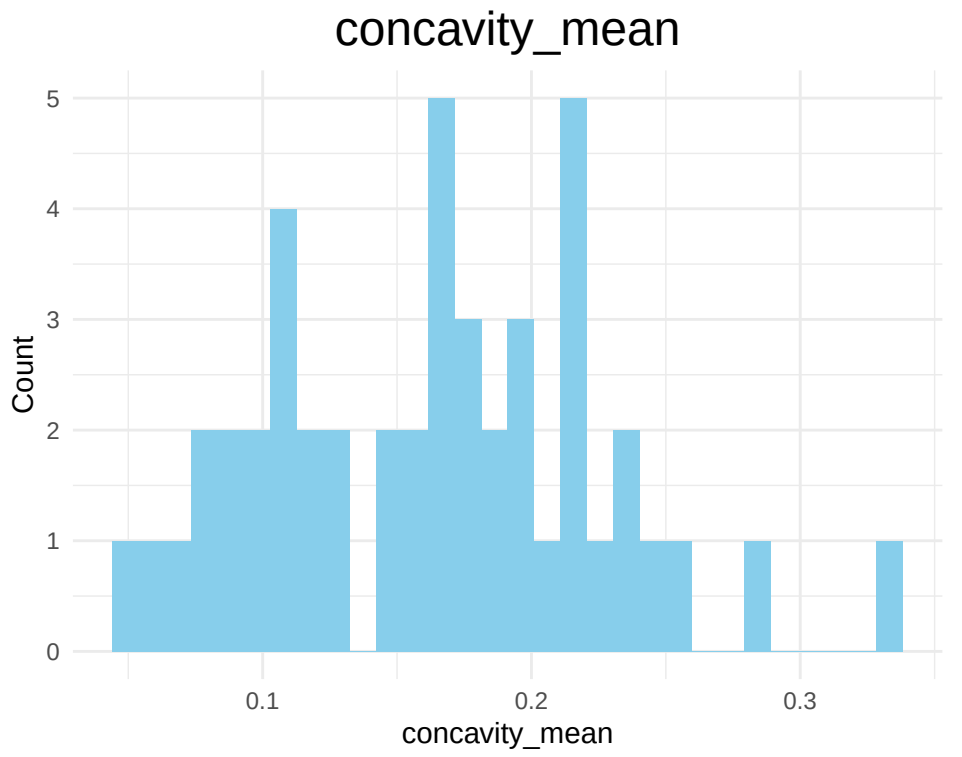

texture_mean
Count
0
1
2
3
4
5
15
20
25
30
texture_mean

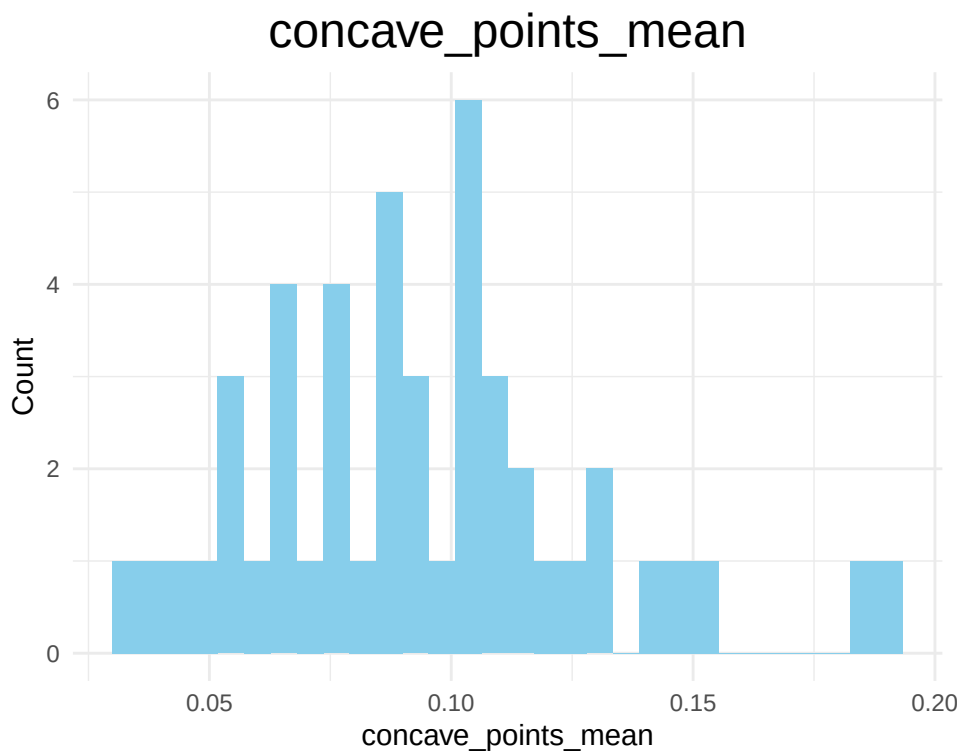

perimeter_mean
Count
0
2
4
6
75
100
125
150
175
perimeter_mean

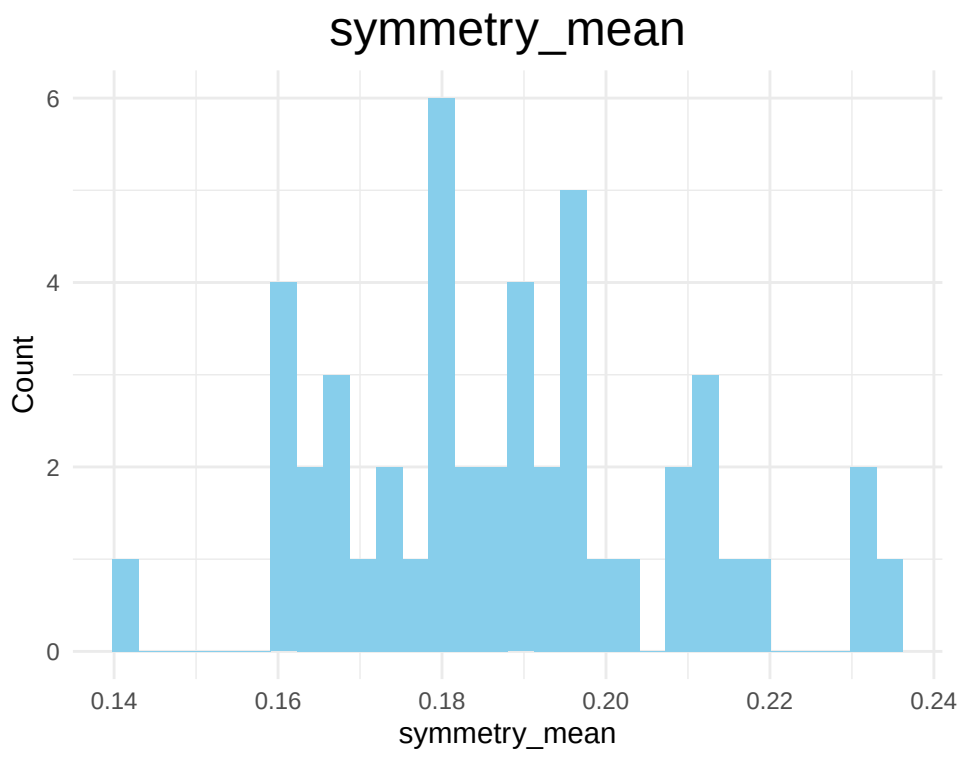

area_mean
Count
6
4
2
0
500
1000
1500
2000
area_mean

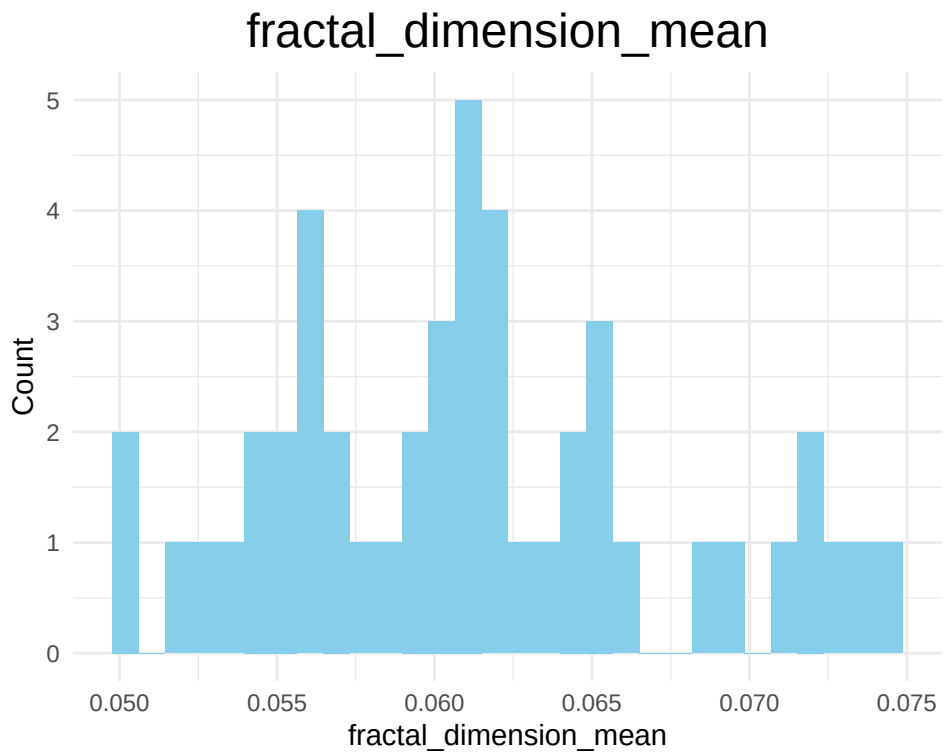

smoothness_mean
Count
5
4
3
2
1
0
0.09
0.10
0.11
0.12
smoothness_mean

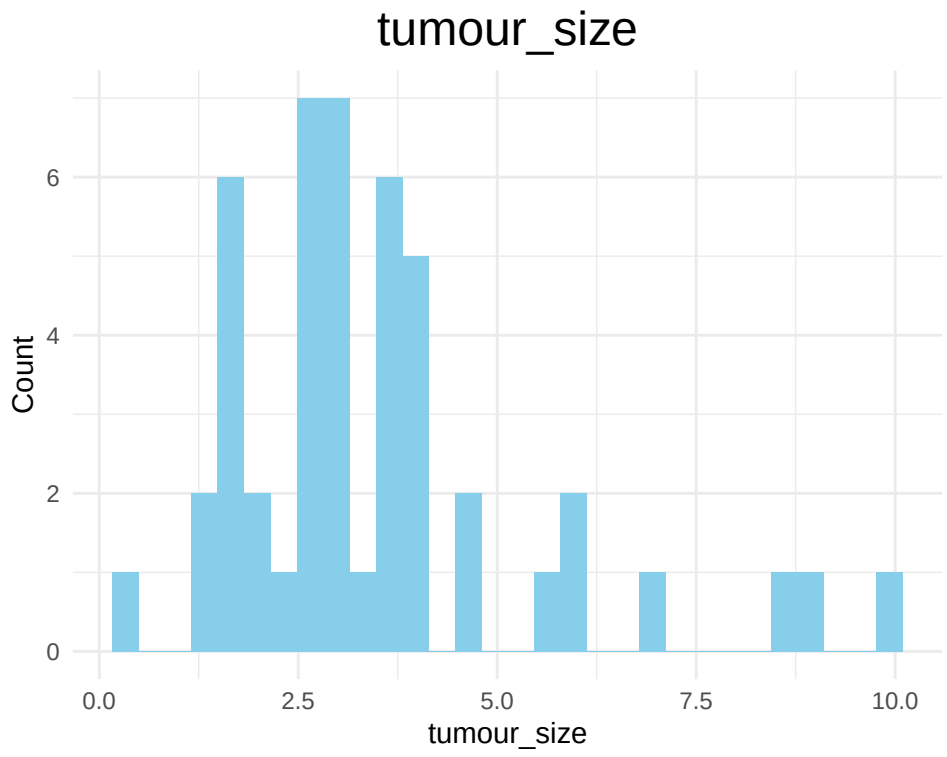

compactness_mean
Count
6
4
2
0
0.10
0.15
0.20
0.25
compactness_mean

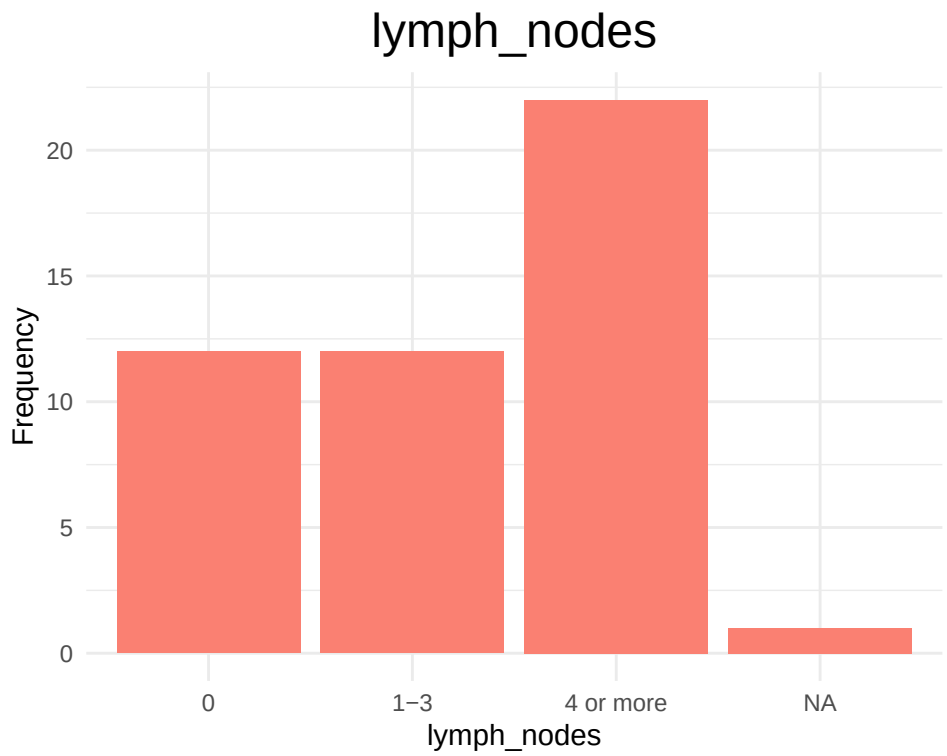

concavity_mean
Count
5
4
3
2
1
0
0.1
0.2
0.3
concavity_mean


concave_points_mean
Count
6
4
2
0
0.05
0.10
0.15
0.20
concave_points_mean


symmetry_mean
Count
6
4
2
0
0.14
0.16
0.18
0.20
0.22
0.24
symmetry_mean


fractal_dimension_mean
Count
0
1
2
3
4
5
0.050
0.055
0.060
0.065
0.070
0.075
fractal_dimension_mean

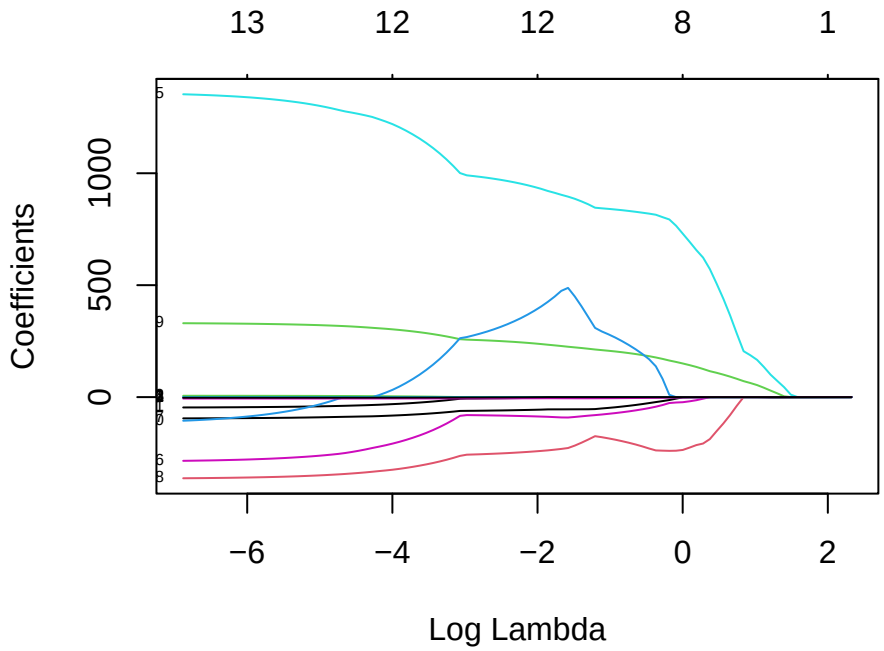

tumour_size
Count
0
2
4
6
0.0
2.5
5.0
7.5
10.0
tumour_size



Continuous variables to be used for prediction are somewhat normally distributed. Do not transform variables.

A3i - Train lasso model for prediction of time to recurrence.

A3ii - Plot predictor weights as a function of lambda.

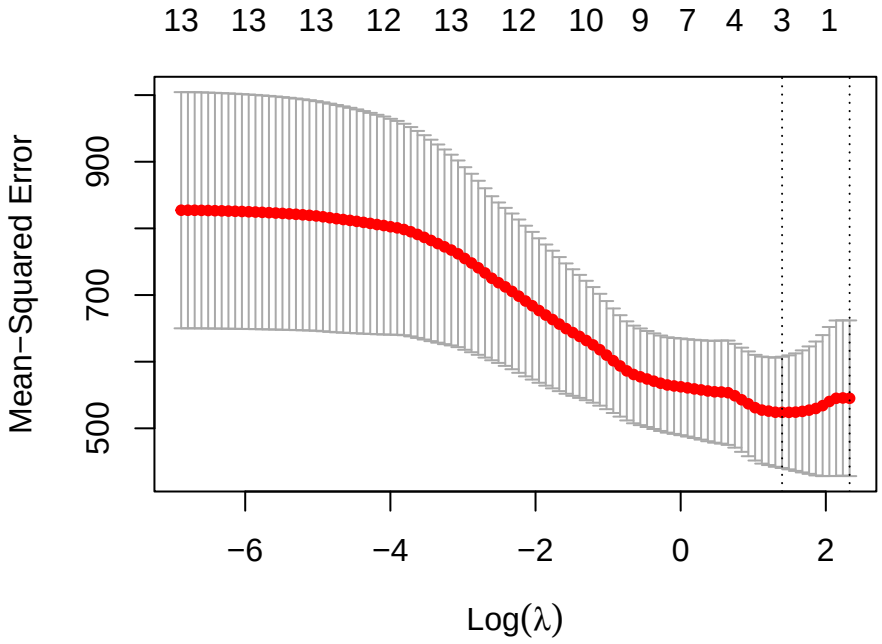



```
[1] "Note that coefficients not standardized due to differing units of predictors."
```

The plot shows the predictor weights as lambda (or log of lambda) changes in the context of lasso regularization. Note that the weights are not standardized since the predictor units are different. An optimal value of lambda that allows the model to make accurate predictions without overfitting can be calculated using cross-validation. From the plot, we can see a few predictors whose weights do not shrink to zero as quickly as the other predictors (with increasing lambda).

A4i - Decide on an optimal value for lambda using cross-validation.

A4ii - Create plot of MSE as a function of lambda and print optimal lambda value and significant predictors of time to recurrence.




```
[1] 4.04503
```

```
[1] "radius_mean"     "smoothness_mean" "symmetry_mean"
```

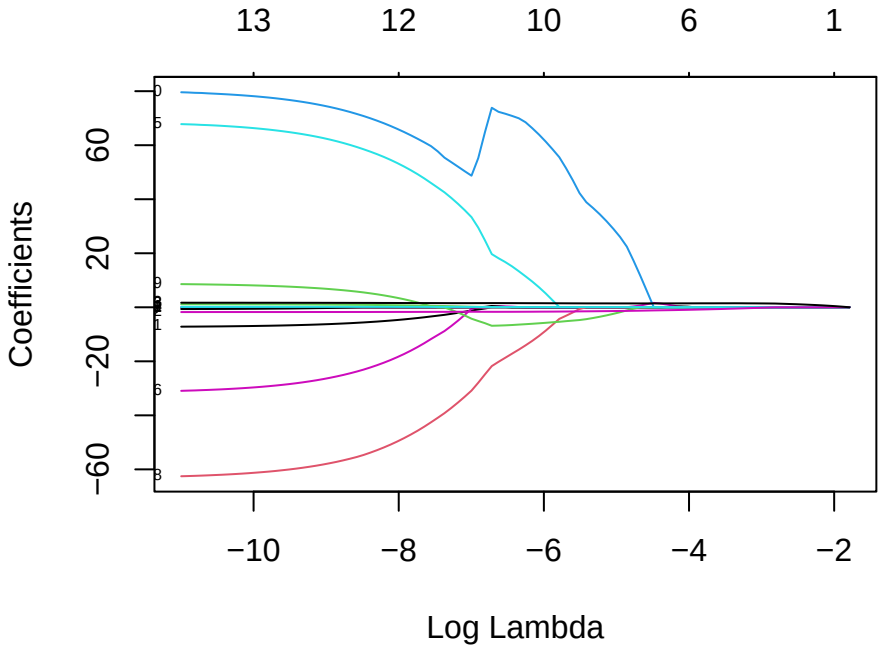
The coefficients for the optimal lambda value are -1.847190, 44.457887 and 3.481176 for the predictors radius_mean, smoothness_mean and symmetry_mean, respectively.

B. Classification

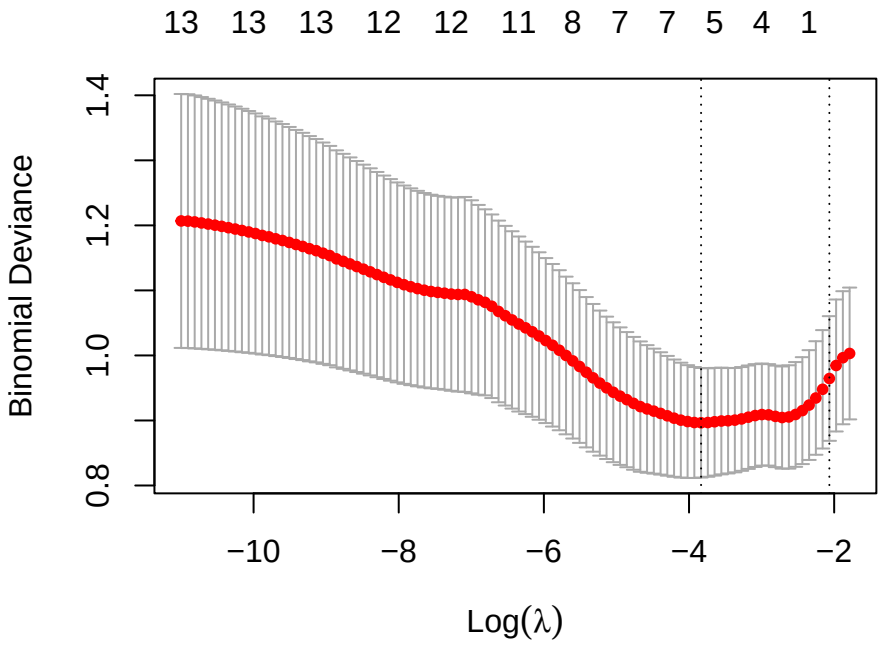
B1i - Clean data as before, but do not remove non-recurrence patients, recode outcome variable values and remove observations with NAs for use in matrix creation and classification.

B1ii - Create training/test sets and lasso model.

B1iii - Plot coefficients of predictors based on different values of lambda.



B1iv - Cross-validate based on deviance to find optimal lambda and extract predictors for that lambda.

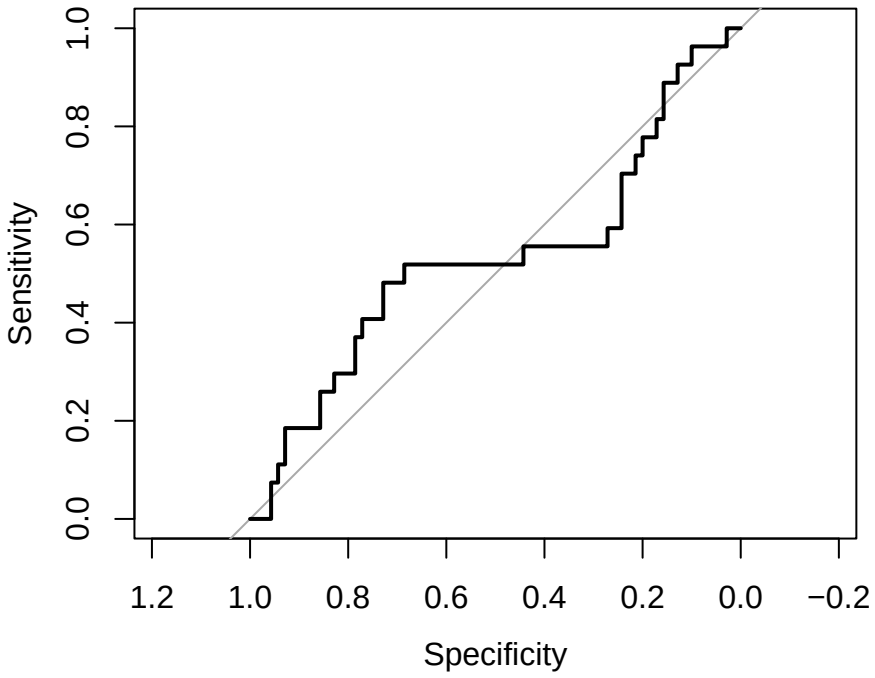


B1v - Test model on test set using optimal lambda

B1vi - Print optimal lambda and significant predictors and plot ROC with AUC for lasso model for predicting recurrence.

```
[1] 0.02162332
```

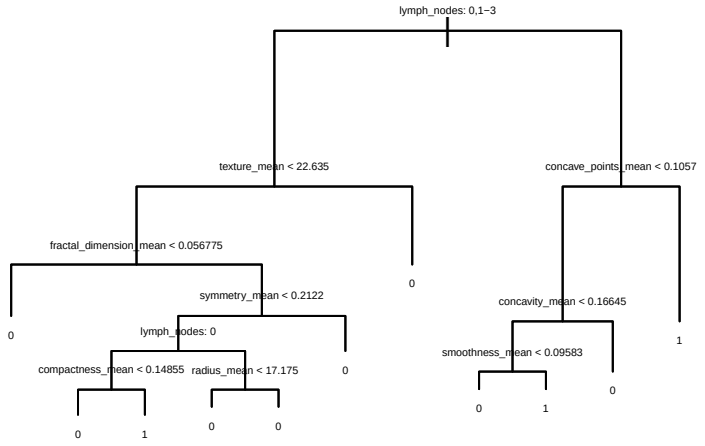
```
[1] "texture_mean"         "area_mean"            "tumour_size"
[4] "lymph_nodes1-3"       "lymph_nodes4 or more"
```

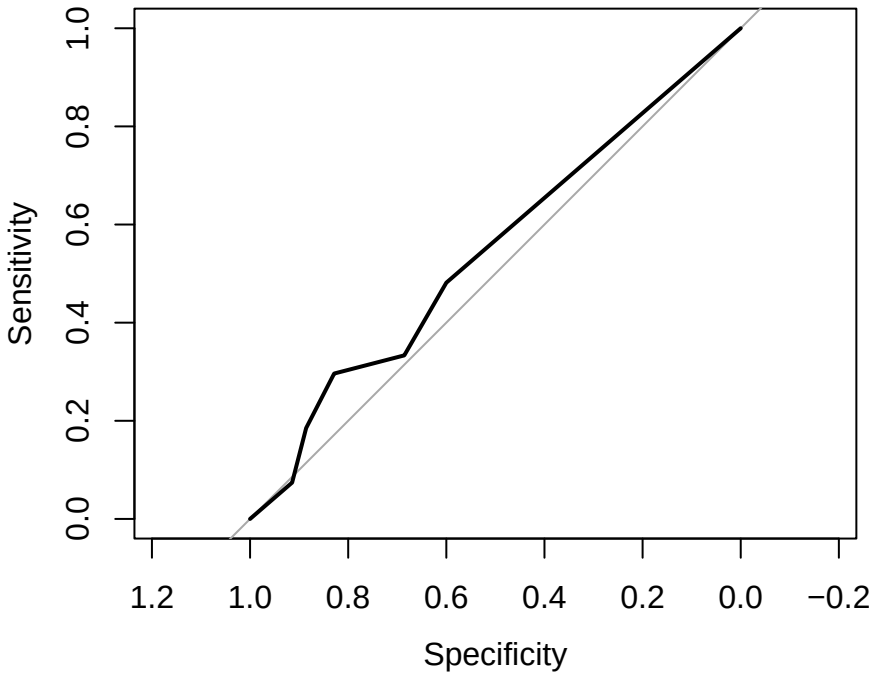



```
Area under the curve: 0.5312
```

B1vii - Create unpruned tree model, get probabilites for test set and obtain ROC with associated AUC.

B1viii - Plot tree, ROC and AUC.

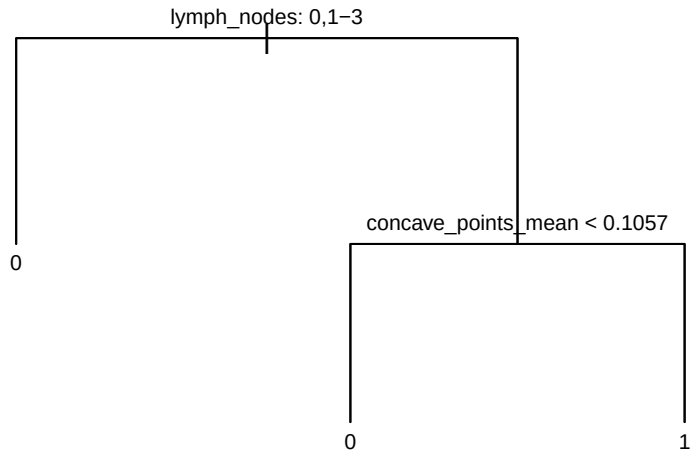


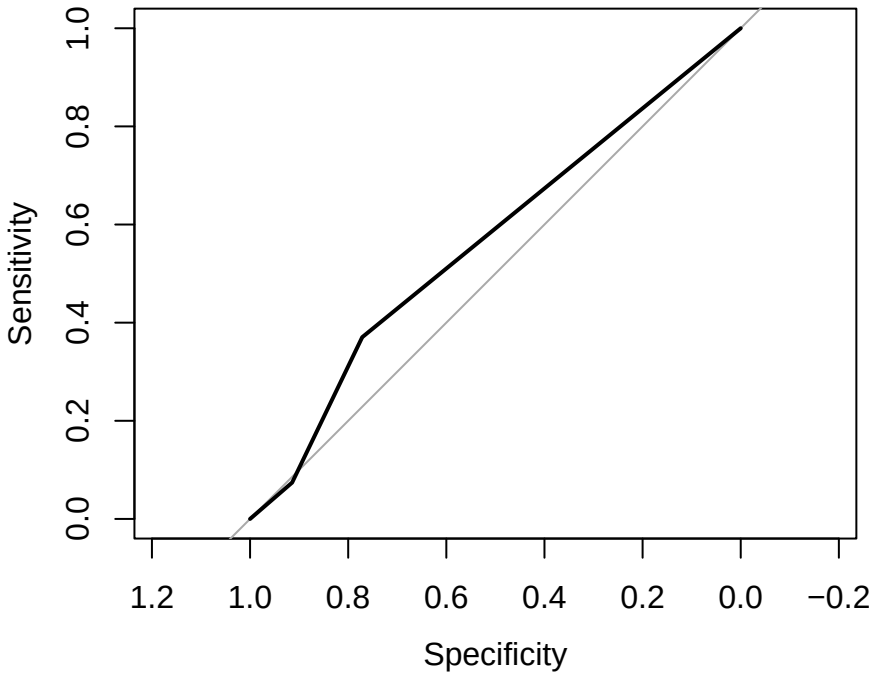


```
Area under the curve: 0.545
```

B1ix - Prune tree by determining best size via cross validation, and obtain ROC and AUC for pruned tree.

B1x - Plot pruned tree, ROC and AUC.






```
Area under the curve: 0.5635
```

B2 - Create table summarizing AUC values for each model assuming different splits via random seed change.

```
  set_seed Lasso_AUC Unpruned_AUC Pruned_AUC
1      123    0.5312       0.5450     0.5635
2      234    0.5571       0.5841     0.4997
3      345    0.6172       0.5161     0.5419
4      456    0.6475       0.5605     0.5996
5      567    0.6598       0.5185     0.4398
```

The 5 repeats of this approach do not come to a consensus on which of models are the best, according to AUC. However, in 3 of the repeats, the lasso regression model performed the best according to AUC, meaning it was better in terms of discrimination and optimizing sensitivity and specificity. This approach would benefit from more repeats (using more random splits) to better support the use of any particular model.

C. Survival Analysis

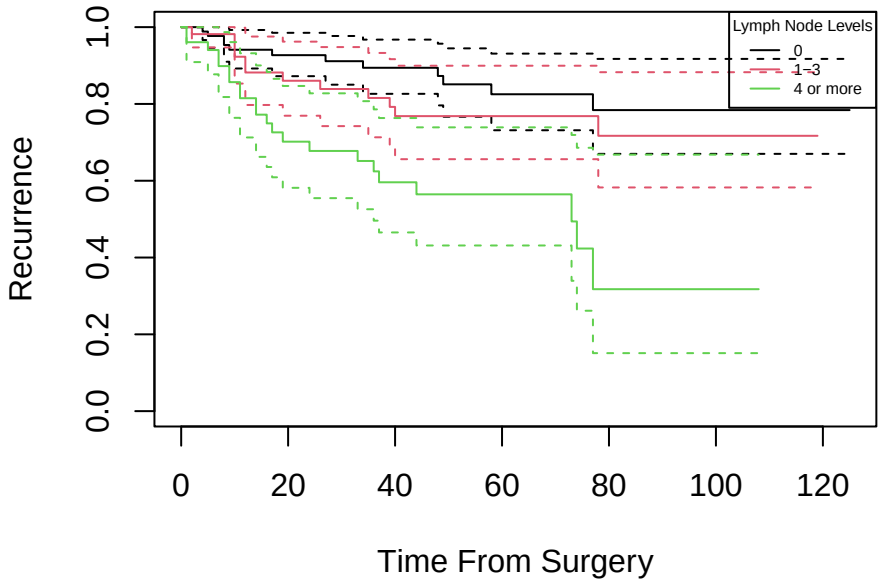
C1 - Censoring:

In this data set, we have censoring because for some patients we cannot ascertain the time of recurrence (event of interest). We know that they have non-recurrence at some time point, but no information on recurrence after that time point. The censored observations would be the patients that have non-recurrence at their given time point while the "event" observations are the patients that have recurrence at their given time points.

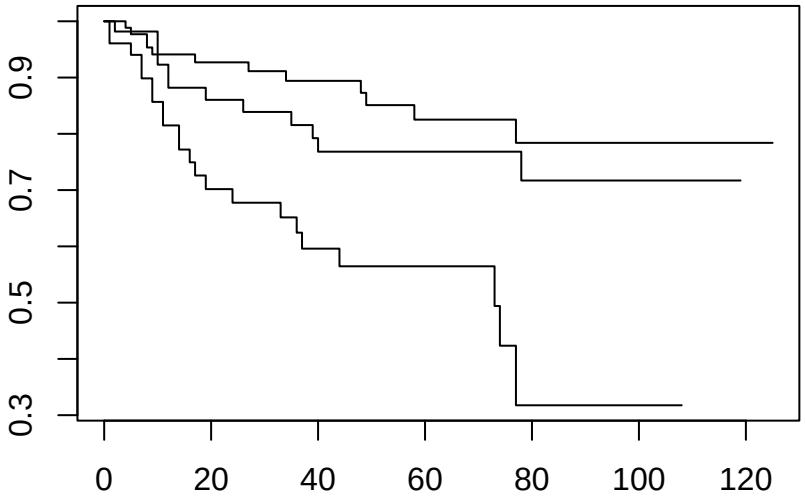
C2i - Report Median Time to Recurrence:

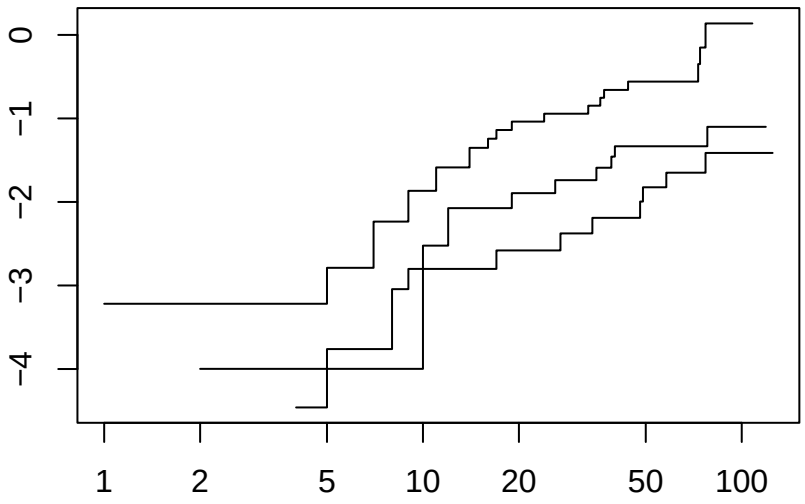
```
                  Statistic Value
1 Median Time to Recurrence    NA
```

C2ii - Create KM "survival" curves for recurrence stratified by lymph nodes. Plot curves.



C2iii - Check PH assumption via different plots and proceed to log rank test.






```
[1] "Curves are somewhat parallel, but do not cross over. Proceed."
```

```
Call:
survdiff(formula = Surv(time, status == 1) ~ lymph_nodes, data = workingc)

                       N Observed Expected (O-E)^2/E (O-E)^2/V
lymph_nodes=0         87       12     21.4     4.125     7.762
lymph_nodes=1-3       56       12     14.4     0.391     0.574
lymph_nodes=4 or more 51       22     10.2    13.523    17.682

 Chisq= 18.3  on 2 degrees of freedom, p= 1e-04
```

C2iv - Interpretation of survival curve findings:

The median time to recurrence was 16.5 time units, likely days or weeks. The result of the log rank test reveals a significant p-value (< 0.05). We reject the null hypothesis that there is no difference in survival between the lymph node groups which provides evidence towards the alternate hypothesis that there is a difference in survival between the lymph node groups.

C3i - Create Cox PH model to predict recurrence and test for PH assumption. Then examine significant predictors.

```
                          chisq df    p
radius_mean            8.37e-01  1 0.36
texture_mean           4.38e-01  1 0.51
perimeter_mean         6.94e-01  1 0.40
area_mean              3.96e-01  1 0.53
smoothness_mean        1.12e+00  1 0.29
compactness_mean       5.68e-04  1 0.98
concavity_mean         2.10e-02  1 0.88
concave_points_mean    1.05e-01  1 0.75
symmetry_mean          5.33e-01  1 0.47
fractal_dimension_mean 1.21e+00  1 0.27
tumour_size            2.49e-01  1 0.62
lymph_nodes            1.27e+00  2 0.53
GLOBAL                 1.93e+01 13 0.11
```

```
Call:
coxph(formula = Surv(time, status == 1) ~ radius_mean + texture_mean +
    perimeter_mean + area_mean + smoothness_mean + compactness_mean +
    concavity_mean + concave_points_mean + symmetry_mean + fractal_dimension_mean +
    tumour_size + lymph_nodes, data = workingc)

  n= 194, number of events= 46

                           coef  exp(coef)   se(coef)      z Pr(>|z|)
radius_mean           -3.775e+00  2.293e-02  1.587e+00 -2.380 0.017333 *
texture_mean          -6.005e-02  9.417e-01  4.457e-02 -1.347 0.177908
perimeter_mean         5.605e-01  1.752e+00  2.493e-01  2.248 0.024574 *
area_mean              2.236e-03  1.002e+00  3.400e-03  0.658 0.510782
smoothness_mean        5.695e+01  5.397e+24  3.096e+01  1.839 0.065844 .
compactness_mean       1.358e+00  3.890e+00  1.353e+01  0.100 0.920009
concavity_mean        -6.922e+00  9.854e-04  7.292e+00 -0.949 0.342462
concave_points_mean   -1.596e+01  1.174e-07  1.803e+01 -0.885 0.376168
```

```
symmetry_mean          -8.838e+00  1.451e-04  8.979e+00 -0.984 0.324956
fractal_dimension_mean -1.359e+02  9.449e-60  6.579e+01 -2.066 0.038857 *
tumour_size            -3.043e-02  9.700e-01  8.272e-02 -0.368 0.712998
lymph_nodes1-3          4.944e-01  1.640e+00  4.224e-01  1.171 0.241771
lymph_nodes4 or more    1.457e+00  4.293e+00  4.360e-01  3.342 0.000833 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                       exp(coef) exp(-coef)  lower .95 upper .95
radius_mean            2.293e-02  4.362e+01  1.023e-03 5.139e-01
texture_mean           9.417e-01  1.062e+00  8.629e-01 1.028e+00
perimeter_mean         1.752e+00  5.709e-01  1.074e+00 2.855e+00
area_mean              1.002e+00  9.978e-01  9.956e-01 1.009e+00
smoothness_mean        5.397e+24  1.853e-25  2.400e-02 1.214e+51
compactness_mean       3.890e+00  2.571e-01  1.190e-11 1.272e+12
concavity_mean         9.854e-04  1.015e+03  6.118e-10 1.587e+03
concave_points_mean    1.174e-07  8.517e+06  5.262e-23 2.620e+08
symmetry_mean          1.451e-04  6.892e+03  3.303e-12 6.375e+03
fractal_dimension_mean 9.449e-60  1.058e+59 9.372e-116 9.526e-04
tumour_size            9.700e-01  1.031e+00  8.249e-01 1.141e+00
lymph_nodes1-3         1.640e+00  6.099e-01  7.165e-01 3.752e+00
lymph_nodes4 or more   4.293e+00  2.329e-01  1.827e+00 1.009e+01

Concordance= 0.738  (se = 0.041 )
Likelihood ratio test= 42.33  on 13 df,   p=6e-05
Wald test            = 43.98  on 13 df,   p=3e-05
Score (logrank) test = 50.65  on 13 df,   p=2e-06
```

C3ii - Comparison of findings (significant predictors) to Parts A and B:

In Part A, the significant predictors for time to recurrence were determined to be mean radius, mean smoothness and mean symmetry of breast mass cell nuclei. In Part B, the significant predictors (based on lasso regression, since repeated splits showed mostly best AUC) for recurrence were determined to be mean texture and mean area of breast mass cell nuclei as well as tumour size and number of positive axillary lymph nodes observed at time of surgery. In Part C (Cox PH model), the significant predictors for recurrence were determined to be mean radius, mean perimeter and mean fractal dimension of breast mass cell nuclei as well as number of positive axillary lymph nodes observed at time of surgery.

Since the outcome for Part A was different from Parts B and C, it can be argued that similarities/differences/patterns in predictors between Part A/Part B and Part A/Part C could be due to the different outcomes measured, rather than the actual validity of the predictors. Regardless, between Parts A and C, the only common significant predictor was mean radius of breast mass cell nuclei. Between Parts B and C which measured the same outcome (recurrence), only the number of lymph nodes observed at time of surgery was chosen as a significant predictor between both models used. Given constraints on model size, it can be argued that these common significant predictors are part of the "true" models for prediction of time to recurrence and recurrence.