Breast Cancer Research

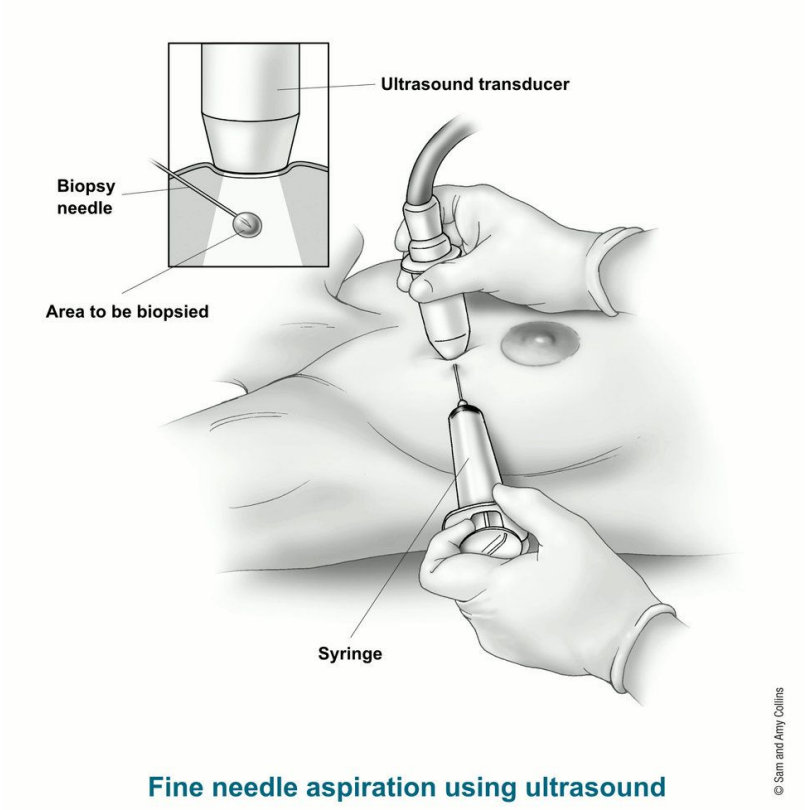
Tarek El-Hajjaoui
Sheldon Gu
Michael Strand

Background

DATA

Data Background:

- sample a small amount of breast tissue
- sample is checked for cancer cells
- digitized image
- computing variable



32 Variables:

ID, Diagnosis (M = malignant, B = benign)

mean, se, worst

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

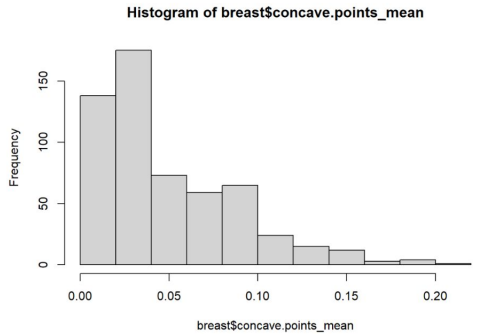
i) symmetry

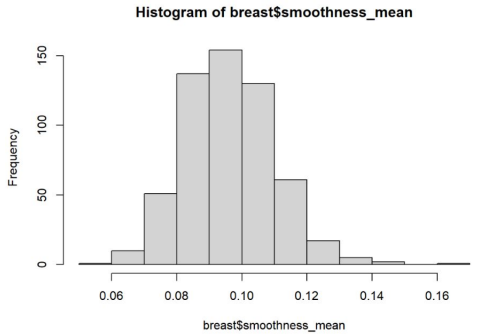
j) fractal dimension ("coastline approximation" - 1)

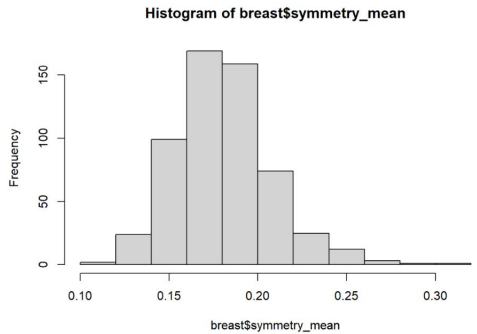
EDA

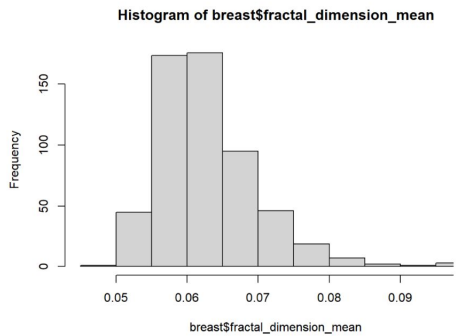
Preliminary Screening

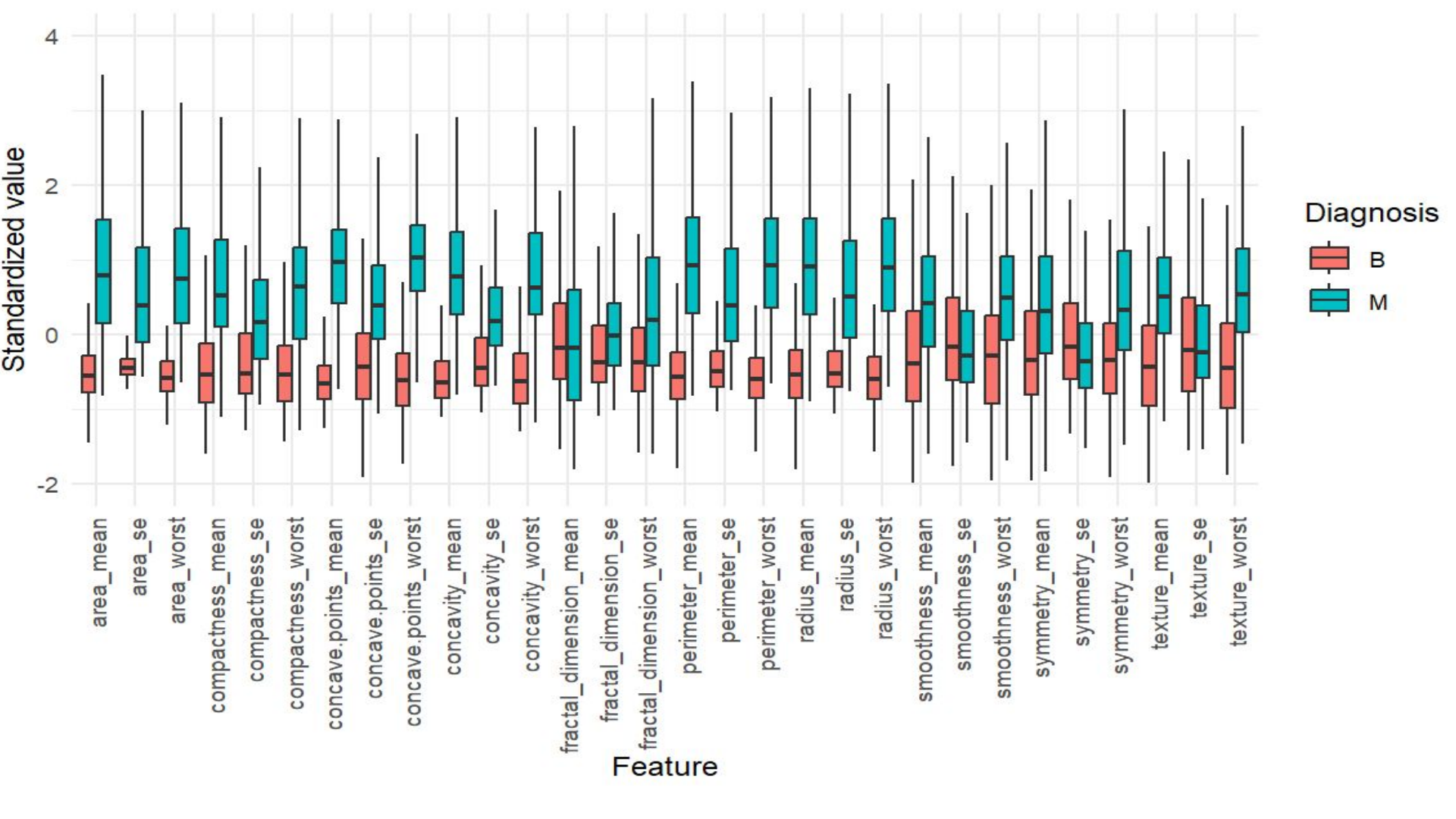
- raw data
- missing value
- remove unused column (id, X)
- distribution





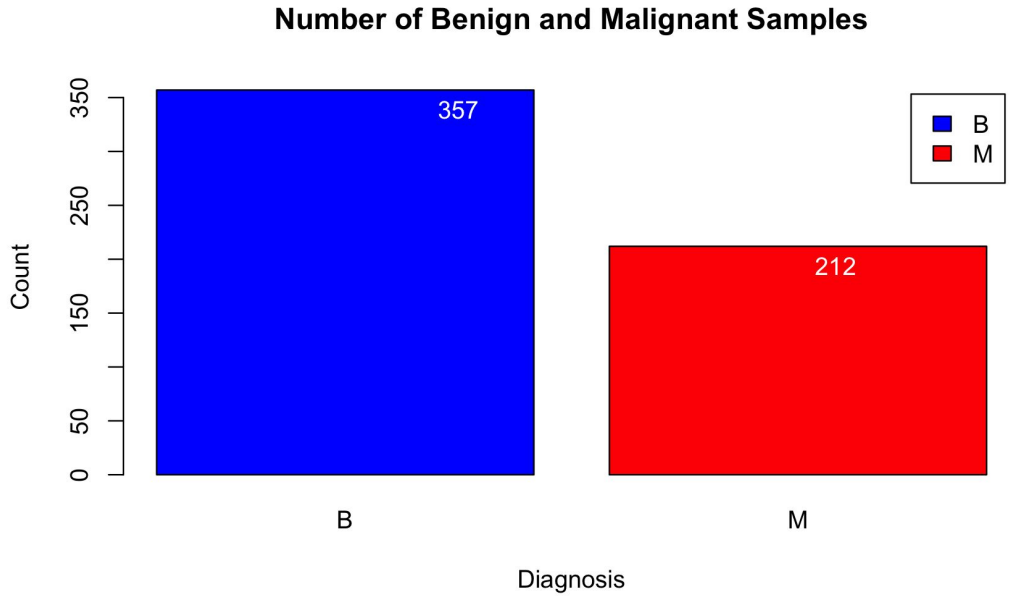





Standardized value
4
2
0
-2
area_mean
area_se
area_worst
compactness_mean
compactness_se
compactness_worst
concave.points_mean
concave.points_se
concave.points_worst
concavity_mean
concavity_se
concavity_worst
fractal_dimension_mean
fractal_dimension_se
fractal_dimension_worst
perimeter_mean
perimeter_se
perimeter_worst
radius_mean
radius_se
radius_worst
smoothness_mean
smoothness_se
smoothness_worst
symmetry_mean
symmetry_se
symmetry_worst
texture_mean
texture_se
texture_worst
Feature
Diagnosis
B
M

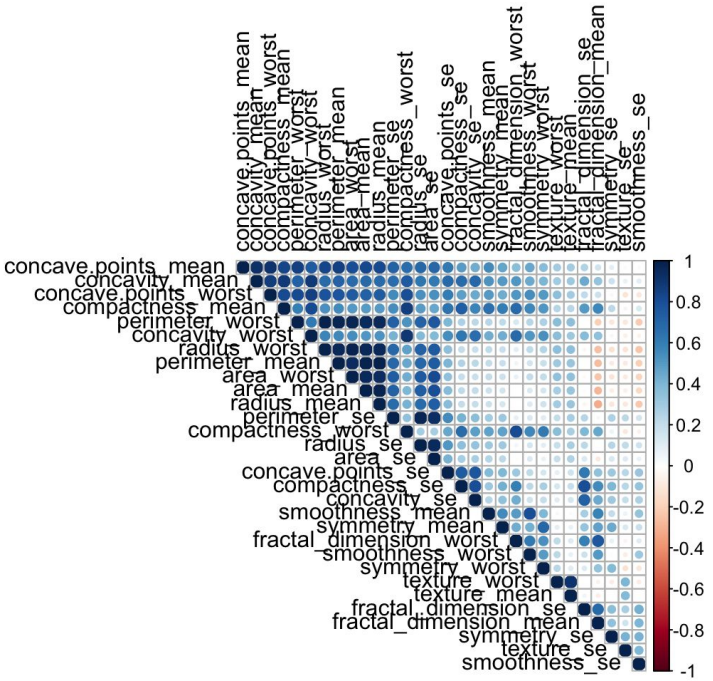
Response Variable

- distribution
- as.factor



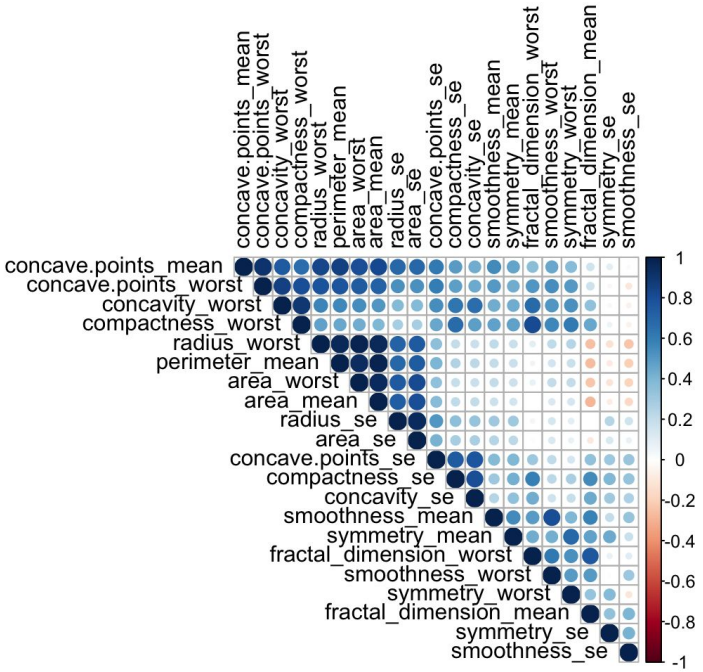
Correlation

- all correlations
- arrange with order = “FPC”
- upper half



Correlation

- too many variables can cause: increased computation, complex for visualization, and interpretation
- remove highly correlated variables (cut off = 0.9)
- library(caret)
- absolute value is considered



PCA

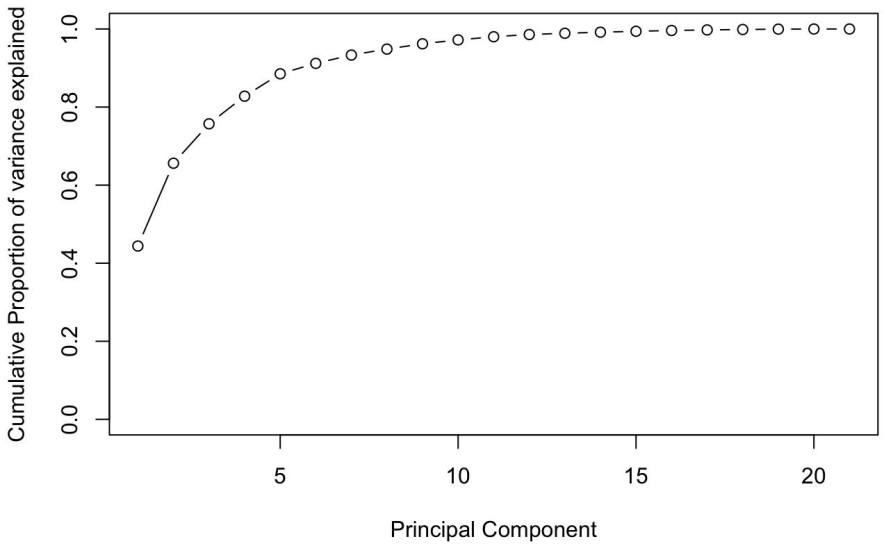
PCA

- standardization is typically recommended before PCA
- summary

```
## Importance of components:
##                           PC1    PC2   PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.053 2.1105 1.456 1.21994 1.09673 0.75004 0.66893
## Proportion of Variance 0.444 0.2121 0.101 0.07087 0.05728 0.02679 0.02131
## Cumulative Proportion  0.444 0.6561 0.757 0.82791 0.88519 0.91197 0.93328
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.56454 0.53543 0.45639 0.41367 0.34423 0.26012 0.24137
## Proportion of Variance 0.01518 0.01365 0.00992 0.00815 0.00564 0.00322 0.00277
## Cumulative Proportion  0.94846 0.96211 0.97203 0.98018 0.98582 0.98904 0.99182
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.22045 0.20547 0.17791 0.15094 0.13695 0.08384 0.02885
## Proportion of Variance 0.00231 0.00201 0.00151 0.00108 0.00089 0.00033 0.00004
## Cumulative Proportion  0.99413 0.99614 0.99765 0.99873 0.99963 0.99996 1.00000
```

PCA

- variance explained
- 6 PCs



PCA

- what variable?
- get_pca_var

  contrib: contributions of the individuals/variables

```
##                              Dim.1      Dim.2        Dim.3        Dim.4
## perimeter_mean           6.52835229  7.6741170 3.713654e-02  0.008162087
## area_mean                6.08218595  8.7868389 3.523792e-02  0.073009228
## smoothness_mean          3.57123398  4.1008720 5.569493e-01 16.129061050
## concave.points_mean      9.49181386  0.6196138 5.268898e-03  0.323929042
## symmetry_mean            3.29916722  4.2690418 7.396811e-04  7.693565172
## fractal_dimension_mean   0.98961096 15.8499742 6.566177e-02  0.079450777
## radius_se                5.43189229  2.8083614 8.518307e+00  3.437975050
## area_se                  5.20087098  4.7028132 5.271109e+00  2.793003983
## smoothness_se            0.06680881  4.5148553 1.642257e+01  7.320387078
## compactness_se           4.45307501  4.4692656 5.761075e+00 10.694968762
## concavity_se             3.63423824  2.9124347 7.506142e+00 15.280775265
## concave.points_se        4.98886476  0.8719787 1.059585e+01  4.246613061
## symmetry_se              0.34136964  3.6538722 1.428490e+01  4.762308037
## radius_worst             6.62893811  7.7001559 5.199705e-01  0.067623666
## area_worst               6.42998687  7.8098453 1.171750e-01  0.307135795
## smoothness_worst         3.02028633  4.2091834 6.831158e+00 10.799498587
## compactness_worst        6.86784733  2.0953073 5.515787e+00  3.910123154
## concavity_worst          7.88869866  0.6996878 2.484502e+00  6.434395806
## concave.points_worst     9.17119726  0.1282285 2.729402e+00  0.422373485
## symmetry_worst           2.75482705  3.1245773 7.138170e+00  3.010585616
## fractal_dimension_worst  3.15873438  8.9989756 5.602894e+00  2.205055300
```

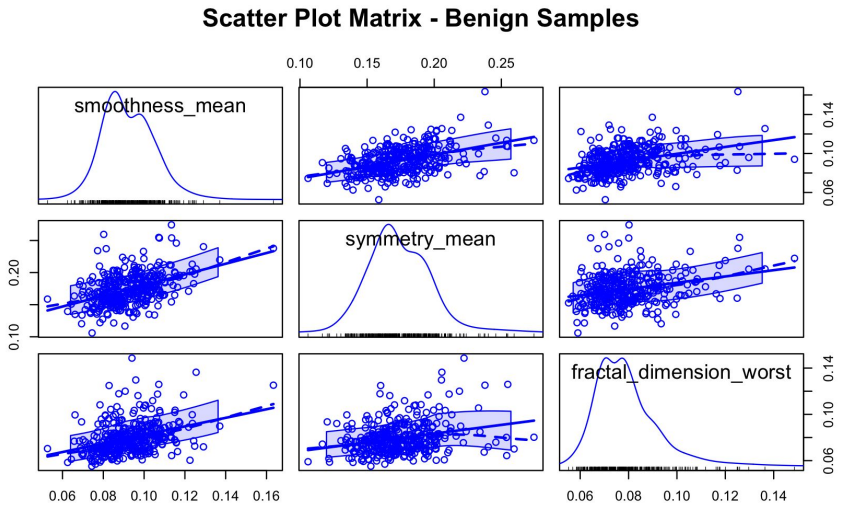
Hotelling T^2

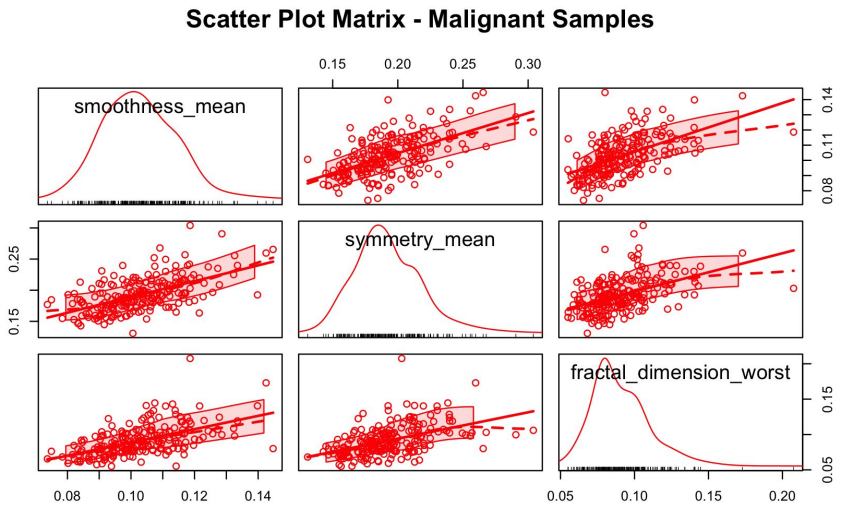
Hotelling T^2: Research Question

- Is there a significant difference between the mean vectors of the Benign and Malignant groups?
  - Null Hypothesis (H_0)
    - No significant difference between the mean vectors of the Benign and Malignant groups.
    - $\mu_M = \mu_B$
  - Alternative Hypothesis (H_A)
    - The Null Hypothesis is not true.
    - $\mu_M \neq \mu_B$

Hotelling T^2:
Multivariate Normality Assumption

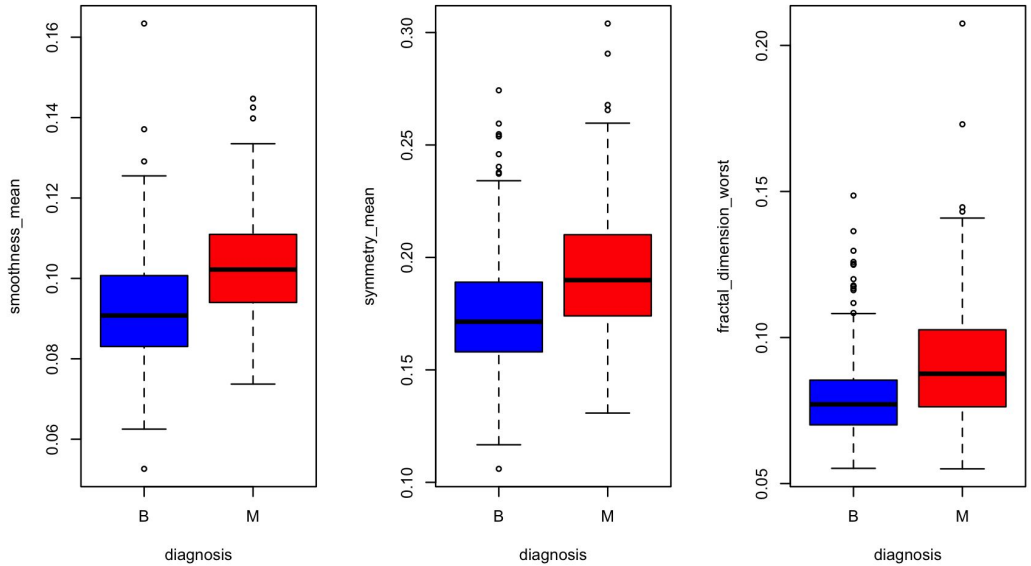
- Variables are marginally and jointly Normally distributed.





Hotelling T^2: Homoscedasticity Assumption

- The variances of the variables within each population should be equal.

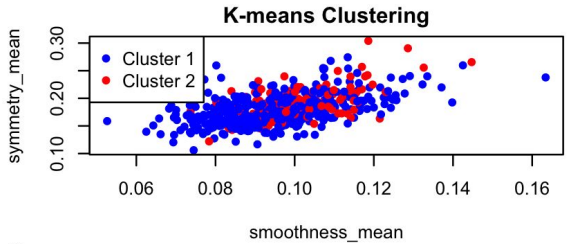


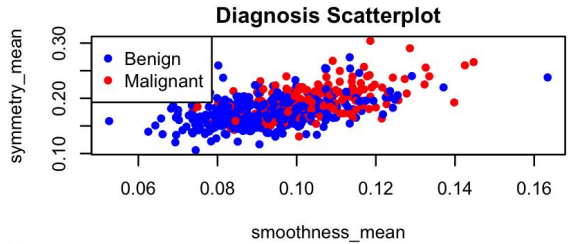
Hotelling T^2 Test: Conclusion

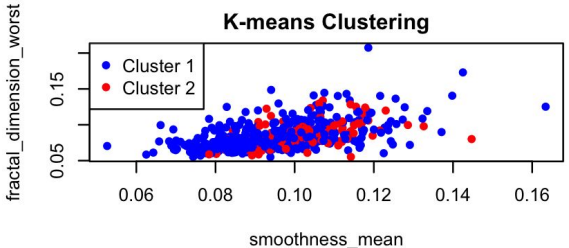
- Results:
  - P-value ≈ 0 ⇒ Reject H_0
- Conclusion: There is a significant difference between mean vectors of Benign population and Malignant population.
  - $\mu_M \neq \mu_B$

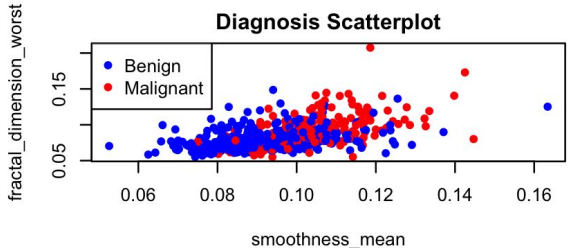
Clustering: KMeans

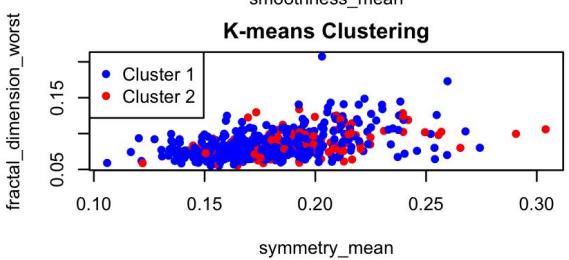
Accuracy Score: 85.413 %

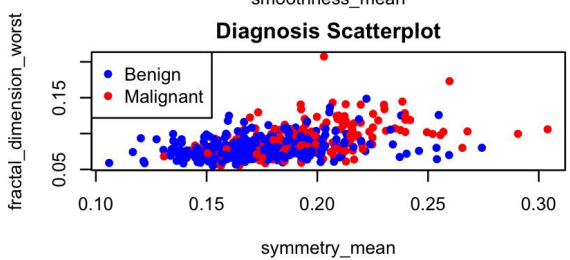












Discriminant Analysis

FLDA classification

- Supervised learning
  - Maximizes class separability
  - Project data onto line
- Assume equal class covariance
  - Relaxed in QDA

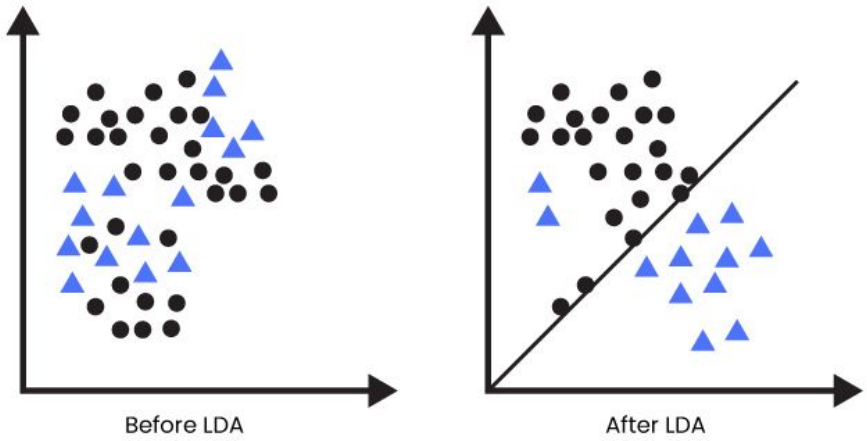


Image source: *Analytics Steps*

- Feature selection informed by PCA
- Want a model which avoids predicting Benign for a Malignant tumor

FLDA decision boundary

Let $\bar{X}_M$ and $\bar{X}_B$ be the two class means. Then LDA seeks a discriminant function

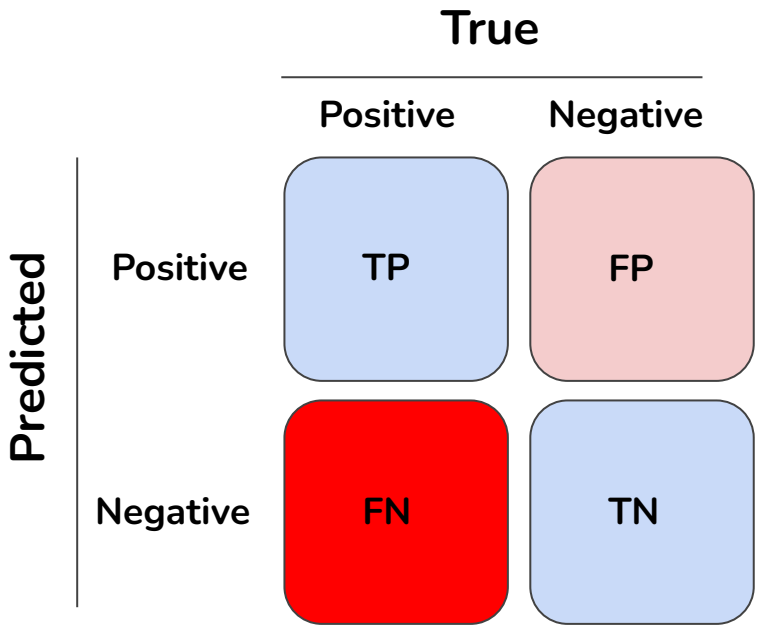
$$f(x) = (\bar{X}_M - \bar{X}_B)^T S_p^{-1} x$$

and uses the decision rule for new observation x_0

$$\text{Malignant if } f(x_0) > (\bar{X}_M - \bar{X}_B)^T S_p^{-1} \frac{\bar{X}_M + \bar{X}_B}{2}$$

$$\text{Benign if } f(x_0) < (\bar{X}_M - \bar{X}_B)^T S_p^{-1} \frac{\bar{X}_M + \bar{X}_B}{2}$$

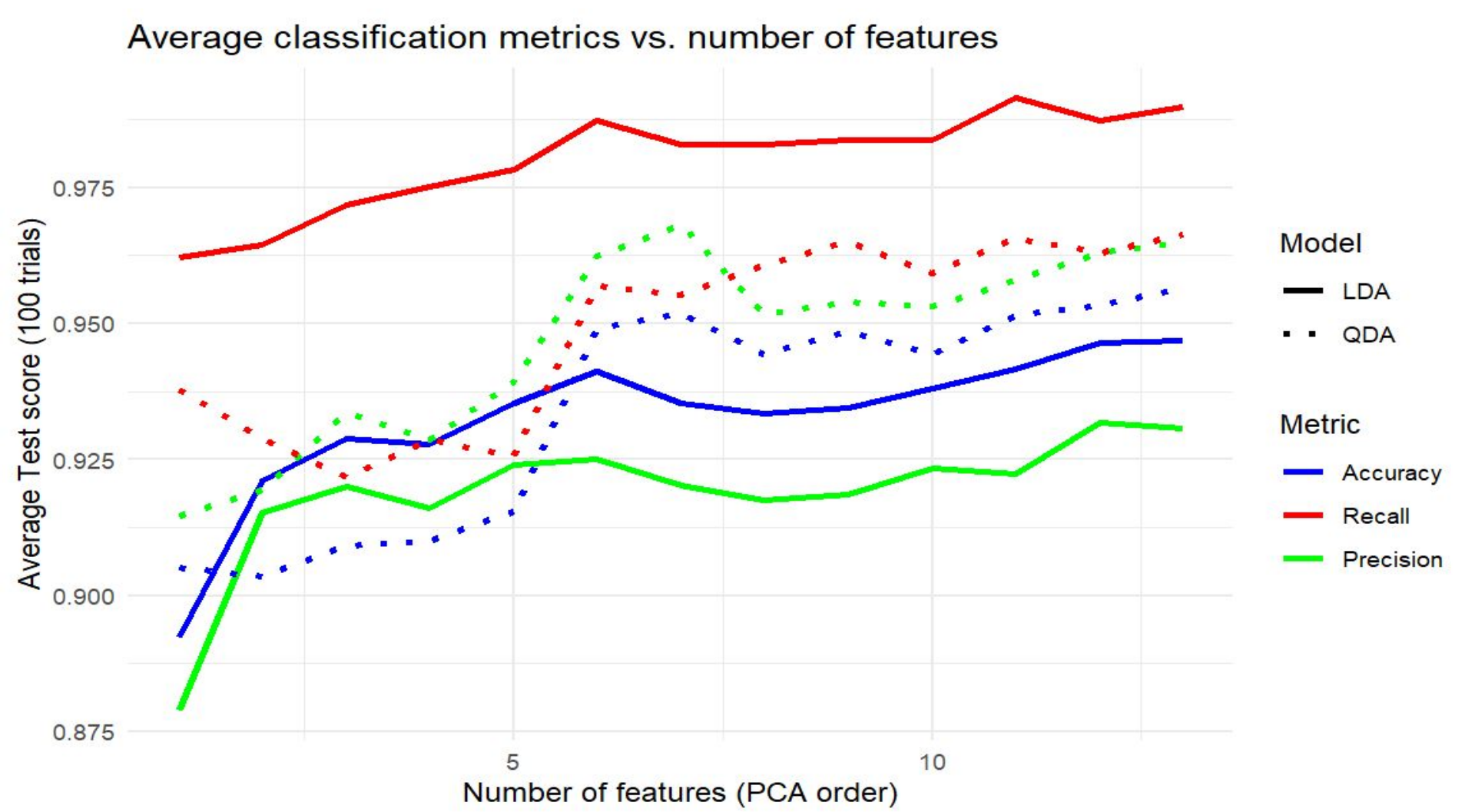
Classification metrics

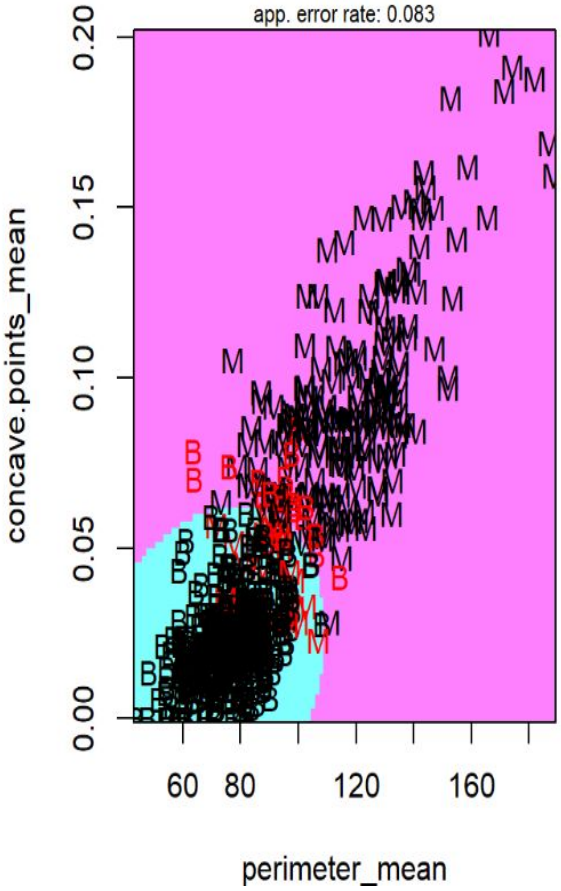


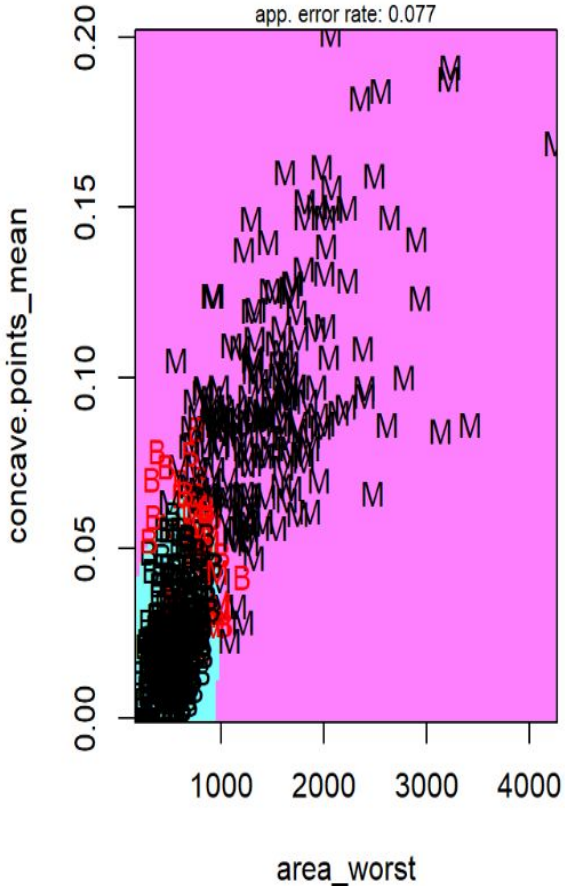
$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

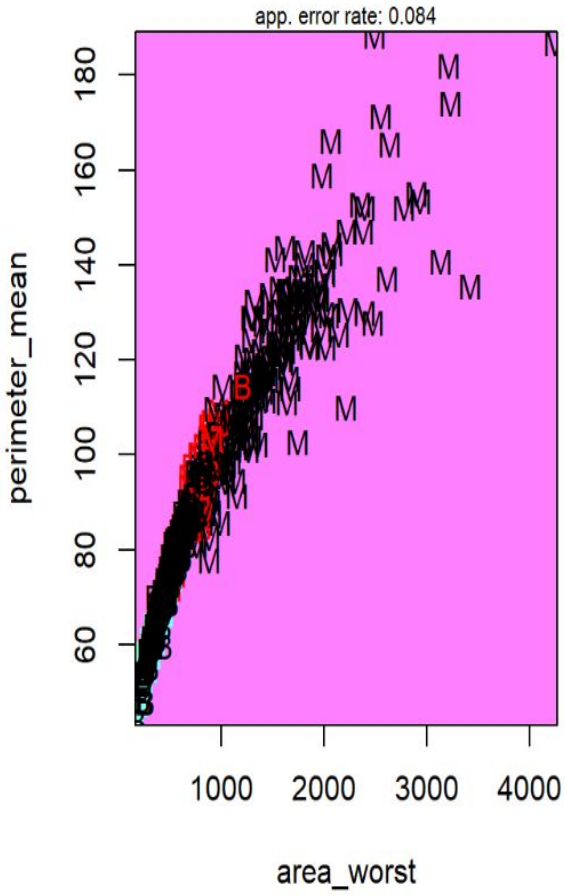
$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$


Average classification metrics vs. number of features
Average Test score (100 trials)
0.975
0.950
0.925
0.900
0.875
5
10
Number of features (PCA order)
Model
LDA
QDA
Metric
Accuracy
Recall
Precision


app. error rate: 0.083
concave.points_mean
perimeter_mean
0.00
0.05
0.10
0.15
0.20
60
80
120
160


app. error rate: 0.077
concave.points_mean
area_worst
0.00
0.05
0.10
0.15
0.20
1000
2000
3000
4000


app. error rate: 0.084
perimeter_mean
area_worst
60
80
100
120
140
160
180
1000
2000
3000
4000

Other Models

Comparison to logistic regression & random forest

	Random forest (all 14 predictors)	**Logistic regression** (all 14 predictors)	**Discriminant analysis** (any # predictors)
Accuracy	0.971	0.965	0.956 (QDA)
Precision	0.968	0.955	0.965 (QDA)
Recall	0.952	0.991	0.992 (LDA)

Thank You