Data Analytics Report

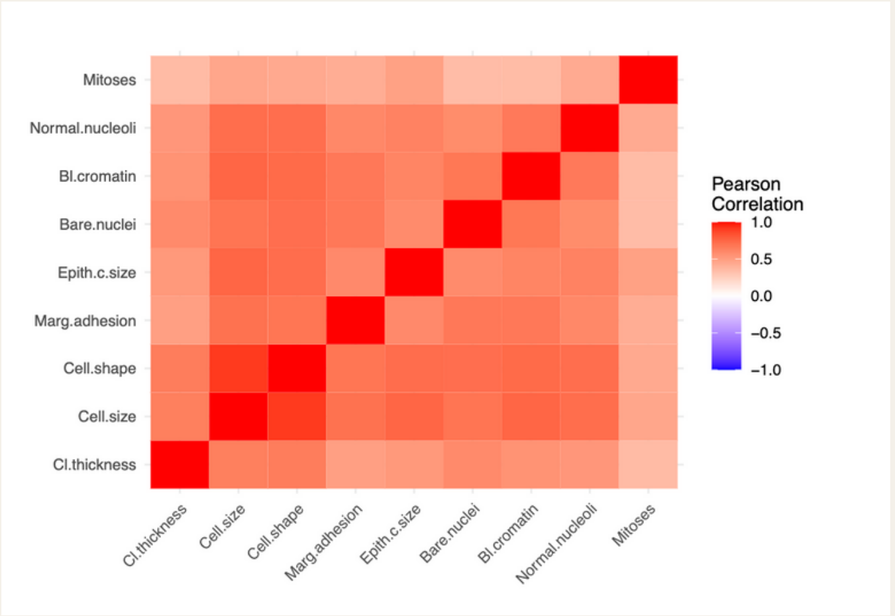
Of Breast Cancer Analysis

Davies
(Zheng Luo)

1st Chapter - Data Exploratory

Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size
Min. : 1.000	Min. : 1.000	Min. : 1.000	Min. : 1.00	Min. : 1.000
1st Qu.: 2.000	1st Qu.: 1.000	1st Qu.: 1.000	1st Qu.: 1.00	1st Qu.: 2.000
Median : 4.000	Median : 1.000	Median : 1.000	Median : 1.00	Median : 2.000
Mean : 4.442	Mean : 3.151	Mean : 3.215	Mean : 2.83	Mean : 3.234
3rd Qu.: 6.000	3rd Qu.: 5.000	3rd Qu.: 5.000	3rd Qu.: 4.00	3rd Qu.: 4.000
Max. :10.000	Max. :10.000	Max. :10.000	Max. :10.00	Max. :10.000

Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	Class
Min. : 1.000	Min. : 1.000	Min. : 1.00	Min. : 1.000	benign :444
1st Qu.: 1.000	1st Qu.: 2.000	1st Qu.: 1.00	1st Qu.: 1.000	malignant:239
Median : 1.000	Median : 3.000	Median : 1.00	Median : 1.000	
Mean : 3.545	Mean : 3.445	Mean : 2.87	Mean : 1.603	
3rd Qu.: 6.000	3rd Qu.: 5.000	3rd Qu.: 4.00	3rd Qu.: 1.000	
Max. :10.000	Max. :10.000	Max. :10.00	Max. :10.000	



- Variables like Bl.cromatin and Normal.nucleoli show skew towards higher values, indicating severe cytological abnormalities in some samples.
- Notable correlations among Cl.thickness, Cell.size, and Cell.shape suggest their combined role in malignancy progression.

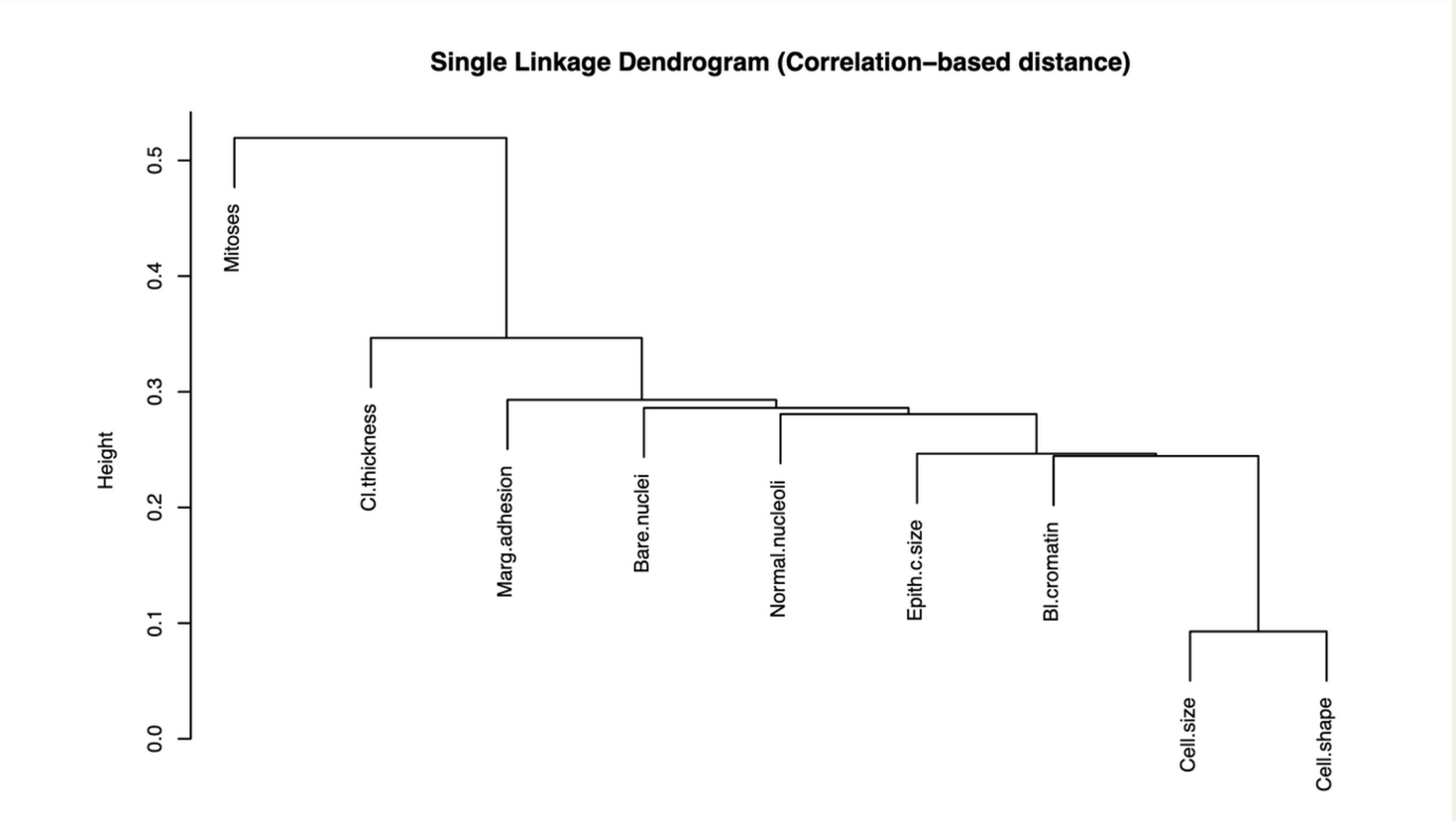
Relationships Between Variables

- Inference: Strong correlation between Cl.thickness, Cell.size, Cell.shape, and cancer classification
- Insight: Identification of potential predictors for further analysis

Predictor Variables Relationships

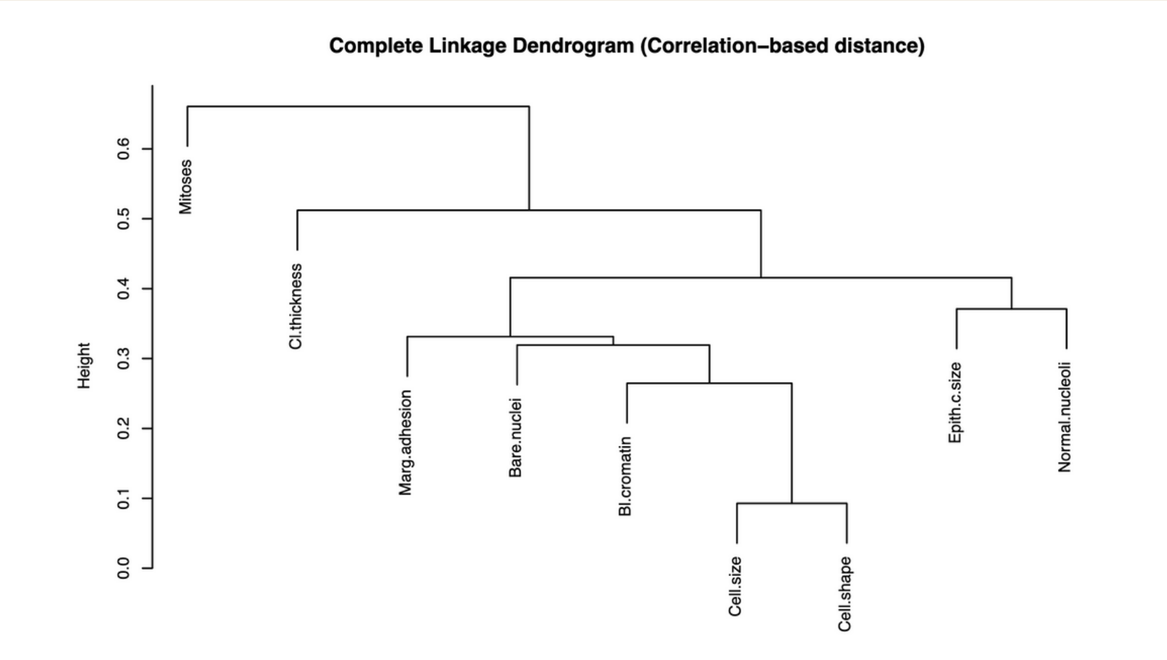
- Observation: PHeatmap reveals strong correlations between predictor variables(e.g. Cell.size and Cell.shape)
- Implication: Redundancy and impact on model performance

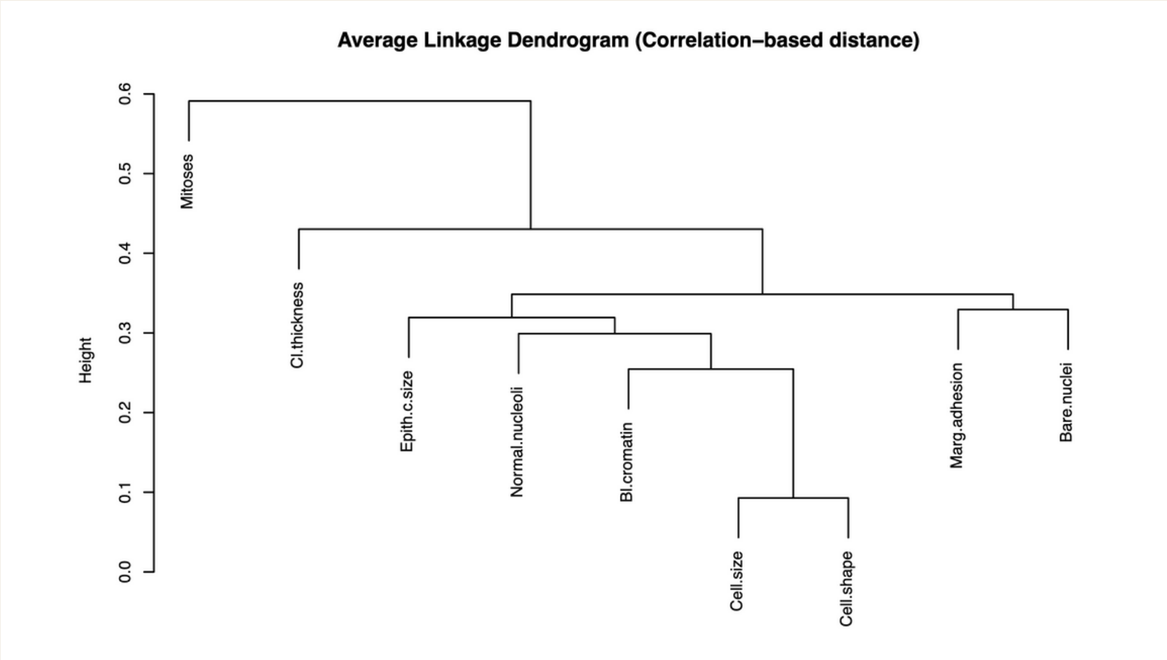
2nd Chapter: Hierarchical Clustering



- Observations: No Clear Separation of Groups
- Conclusion: Limitations of single-linkage in this context

Complete & Average Linkage





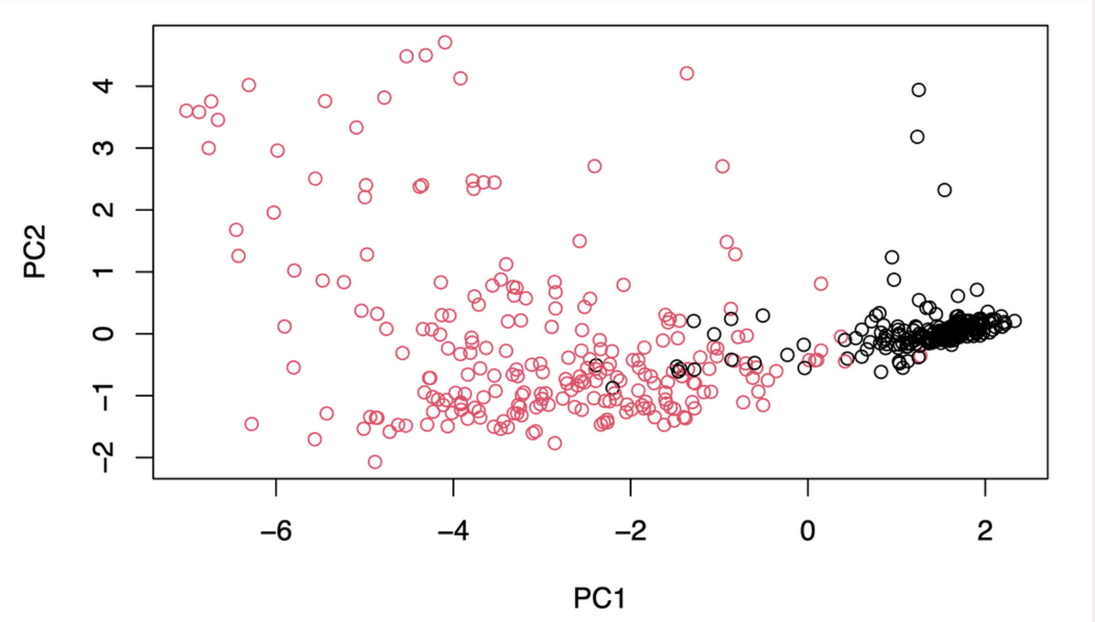
- Observations: Complete Linkage is more compact, while Average Linkage is more moderate
- Conclusion: Both methods provided unique insights into the dataset's structure, each highlighting different potential groupings effectively.

Variable Differences

Loading PCA

```
##                    PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9
## Cl.thickness    -0.302 -0.141  0.866  0.108 -0.080 -0.243 -0.009 -0.248  0.003
## Cell.size       -0.381 -0.047 -0.020 -0.204  0.146 -0.139 -0.205  0.436  0.733
## Cell.shape      -0.378 -0.082  0.034 -0.176  0.108 -0.075 -0.127  0.583 -0.667
## Marg.adhesion   -0.333 -0.052 -0.413  0.493  0.020 -0.655  0.124 -0.163 -0.046
## Epith.c.size    -0.336  0.164 -0.088 -0.427  0.637  0.069  0.211 -0.459 -0.067
## Bare.nuclei     -0.335 -0.261  0.001  0.499  0.125  0.609  0.403  0.127  0.077
## Bl.cromatin     -0.346 -0.228 -0.213  0.013 -0.228  0.299 -0.700 -0.384 -0.062
## Normal.nucleoli -0.336  0.034 -0.134 -0.417 -0.690  0.022  0.460 -0.074  0.022
## Mitoses         -0.230  0.906  0.080  0.259 -0.105  0.148 -0.132  0.054 -0.007
```

PCA Scattrt plot



- PC1 primarily reflects variations in Cell.size, Cell.shape, and Bl.cromatin, suggesting it represents general cell abnormality and chromatin patterns, vital for distinguishing between benign and malignant samples
- PC2 is significantly influenced by Mitoses and contrasts with Bare.nuclei and Bl.cromatin, highlighting its role in differentiating samples based on cell division rates and nuclear features.

3rd Chapter: K-means Clustering

```r
my_kmeans <- function(X, K, max.iter = 100) {
  # Randomly assign initial cluster_center
  cluster_center <- X[sample(nrow(X), K), ]
  clusters <- rep(0, nrow(X))
  for (i in 1:max.iter) {
    # Assign clusters based on closest cluster_center
    clusters <- apply(X, 1, function(x) {
      which.min(colSums(t(cluster_center - x)^2))})

    # Recalculate cluster_center
    new_cluster_center <- matrix(ncol = ncol(X), nrow = K)
    for (k in 1:K) {
      new_cluster_center[k, ] <- colMeans(X[clusters == k, , drop=FALSE], na.rm = TRUE)}

    # Check for convergence
    if (all(cluster_center == new_cluster_center)) {
      message("Convergence reached after ", i, " iterations.")
      break}

    cluster_center <- new_cluster_center}

  # Calculate within-cluster sum-of-squares
  SSW <- sum(sapply(1:K, function(k) {
    sum(rowSums((X[clusters == k, , drop=FALSE] - cluster_center[k, ])^2))}))

  return(list(cluster = clusters, centers = cluster_center, ssw = SSW))}
```

- The custom my_kmeans function iteratively assigns data points to the nearest of K initial centroids and updates these cluster-centers as the mean of their clusters until convergence.

Comparing to R's kmeans function

1. This is the cluster of the best run I found:

	Length	Class	Mode
cluster	683	-none-	numeric
centers	18	-none-	numeric
ssw	1	-none-	numeric

2. The SSw value of it:

```
## [1] 30166.39
```

3. Check if the two partitions are the same:

```
## [1] FALSE
```

4. A contingency table to compare the cluster assignments:

```
##       1   2
##   1 452   5
##   2   1 225
```

- Conclusion: The implementation of the K-means algorithm is consistent with the built-in R function in identifying clusters within the breast cancer dataset.

Equivalence of Cluster Partitions

- For a given dataset and choice of K, even with the same stopping conditions for R's kmeans function and my own K-means implementation, the procedures will **NOT** always give two equivalent cluster partitions.

- Main Reasons:
  - Random cluster-center initialisation: Different initial starting points in K-means can lead to different solutions.
  - Local Minima: K-means may converge to different local minima, affecting the final clusters.

4th Chapter: Classification

Summary Statistics for Training and Test Data

	Mean	Median	Min	Max	Variance	SD
Training.Id	1086817.980	1173511.5	76389	13454352	458360949439.561	677023.596
Training.Cl.thickness	4.394	4.0	1	10	7.758	2.785
Training.Cell.size	3.201	1.0	1	10	9.644	3.105
Training.Cell.shape	3.267	1.5	1	10	9.220	3.036
Training.Marg.adhesion	2.897	1.0	1	10	8.456	2.908
Training.Epith.c.size	3.231	2.0	1	10	4.956	2.226
Training.Bare.nuclei	3.599	1.0	1	10	13.507	3.675
Training.Bl.cromatin	3.469	3.0	1	10	6.224	2.495
Training.Normal.nucleoli	2.868	1.0	1	10	9.293	3.048
Test.Id	1036476.628	1145420.0	63375	1368882	92804948617.015	304639.046
Test.Cl.thickness	4.635	4.0	1	10	8.763	2.960
Test.Cell.size	2.949	1.0	1	10	8.416	2.901
Test.Cell.shape	3.007	1.0	1	10	7.787	2.790
Test.Marg.adhesion	2.562	1.0	1	10	7.174	2.679
Test.Epith.c.size	3.248	2.0	1	10	4.923	2.219
Test.Bare.nuclei	3.328	1.0	1	10	12.399	3.521
Test.Bl.cromatin	3.350	3.0	1	9	5.141	2.267
Test.Normal.nucleoli	2.876	1.0	1	10	9.492	3.081

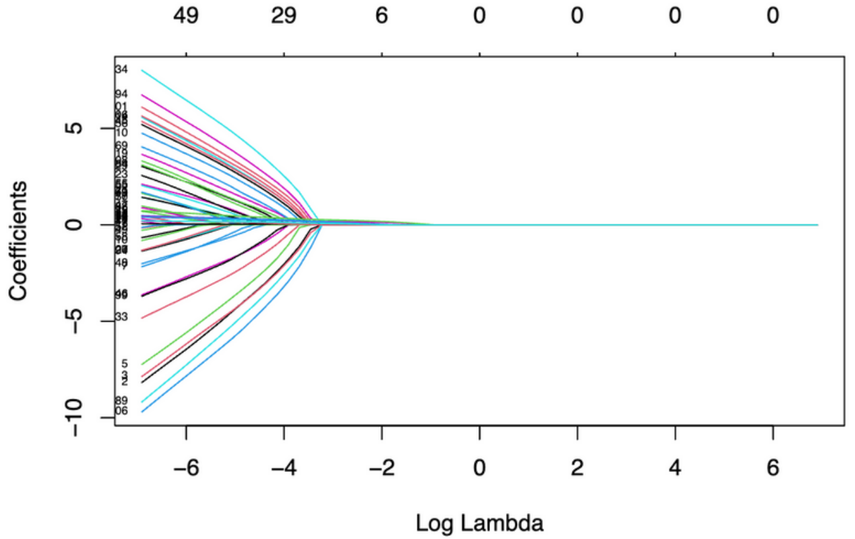
- Data split strategy: 80% training, 20% testing
- Goal: Balanced approach for model training and validation

Building Classifiers

Coefficients from Subset Selection Logistic Regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.914	1.257	-7.885	0.000
Cl.thickness	0.415	0.155	2.687	0.007
Cell.shape	0.388	0.183	2.116	0.034
Marg.adhesion	0.326	0.134	2.427	0.015
Bare.nuclei	0.419	0.103	4.075	0.000
Bl.cromatin	0.576	0.200	2.876	0.004
Normal.nucleoli	0.178	0.122	1.463	0.144
Mitoses	0.525	0.348	1.509	0.131

LASSO Coefficient Paths



- The Subset Selection model includes key variables like Cl.thickness and Cell.shape, with positive coefficients indicating significant role in predicting malignancy in breast tissue.
- The LASSO model demonstrates variable importance and coefficient shrinkage, revealing the diminishing impact of variables like Mitoses while underscoring the consistent relevance of Cl.thickness, Cell.shape, and others.

Model Interpretation

Comparison of Coefficients from Subset Selection and LASSO Logistic Regression

Variable	Subset_Coefficient	LASSO_Coefficient
(Intercept)	-9.914	-8.255
Bare.nuclei	0.419	0.464
Bl.cromatin	0.576	0.302
Cell.shape	0.388	0.304
Cell.size	NA	0.264
Cl.thickness	0.415	0.330
Epith.c.size	NA	0.141
Marg.adhesion	0.326	0.033
Mitoses	0.525	0.000
Normal.nucleoli	0.178	0.169

- Observation: In the subset selection model, Cell.size and Epith.c.size were dropped, whereas the LASSO model only reduced Mitoses to zero, highlighting its unique approach to variable selection.
- Conclusion: The retained variables across both models, particularly Cl.thickness, Cell.shape, and Bare.nuclei, are crucial predictors for classifying breast cancer tissue, demonstrating their strong association with malignancy.

Model Comparison

Test Error Camparison Between Two Models

Model	Test_Error
Subset Selection	0.0583942
LASSO	0.0656934

- he Subset Selection model exhibited slightly superior performance despite including one less variable than the LASSO model.
- the inclusion of additional variables in the LASSO model does not necessarily translate to increased accuracy.

Final Model Selection

Choose **Subset Selection model** as the final classifier due to slightly better accuracy.

- Important to consider the nature of misclassification errors, especially in medical diagnostics.
- Decision balances accuracy with model complexity and practical application needs.

Thanks

Davies
(Zheng Luo)