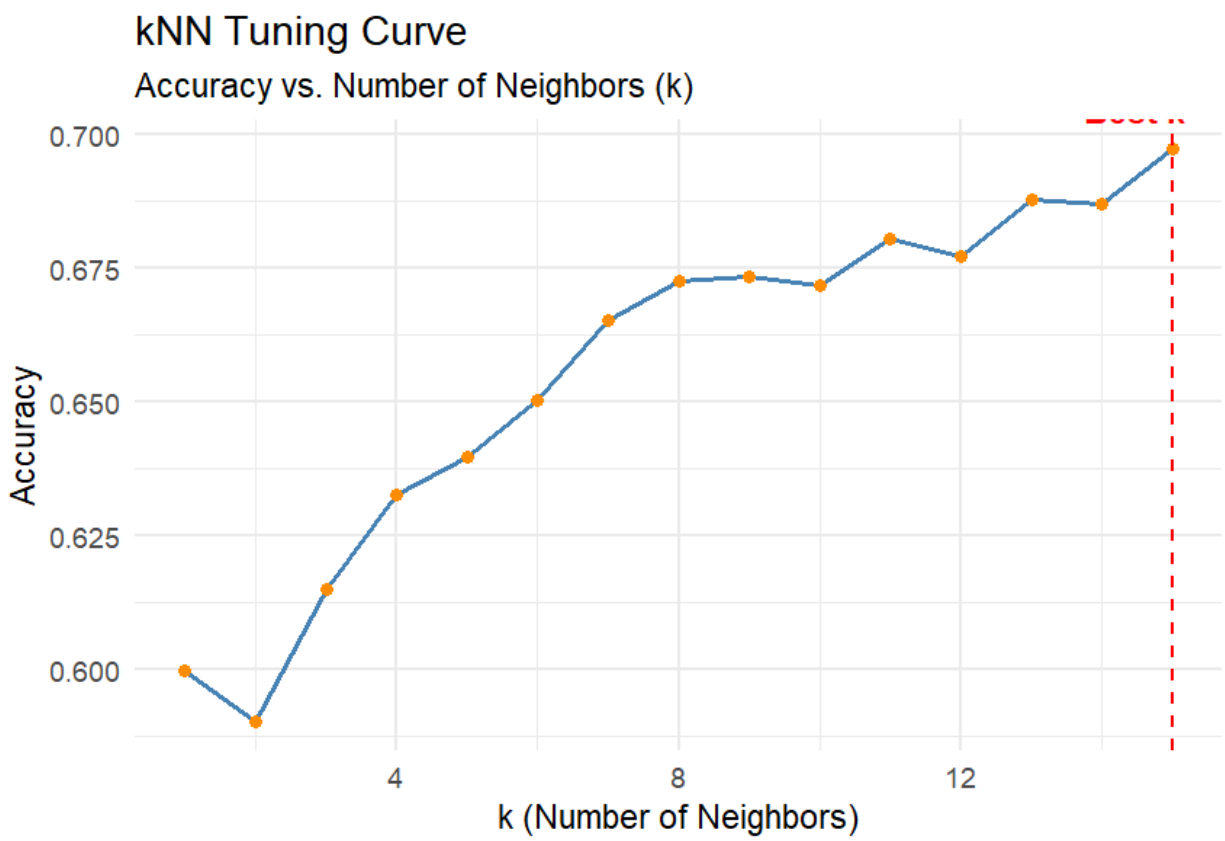
Ryan Lee
Data Analytics
October 17, 2025

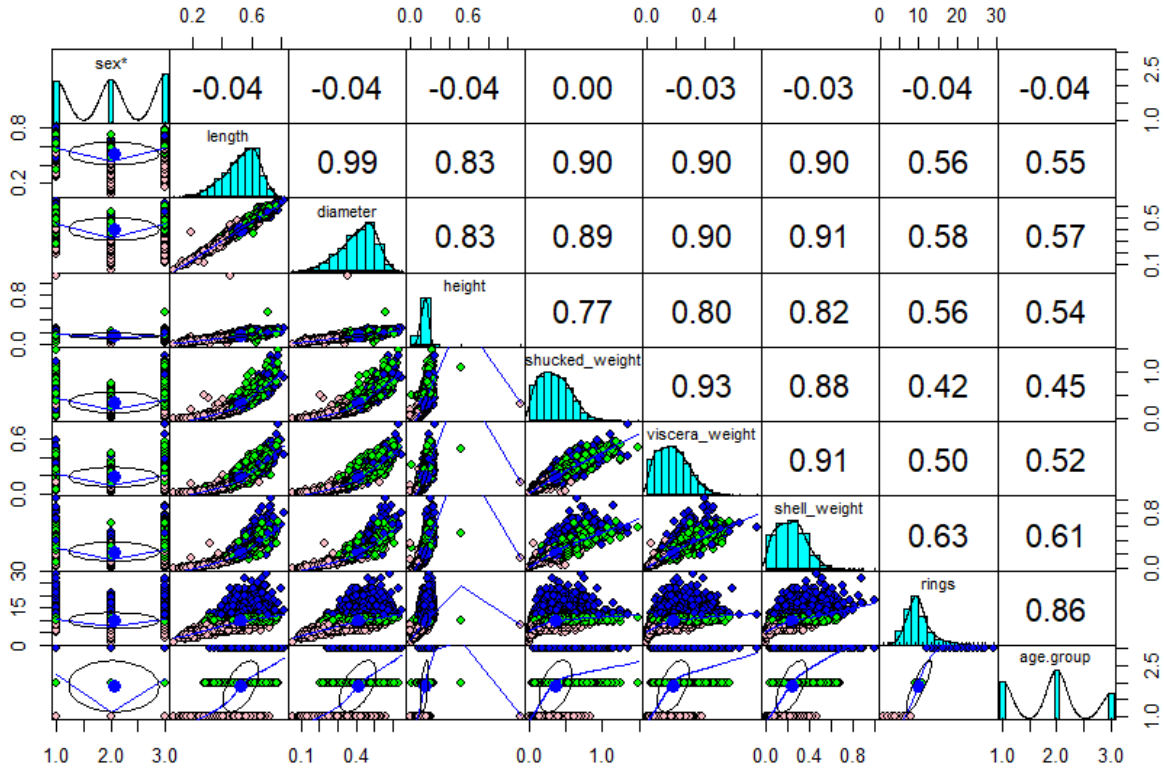
Lab 3

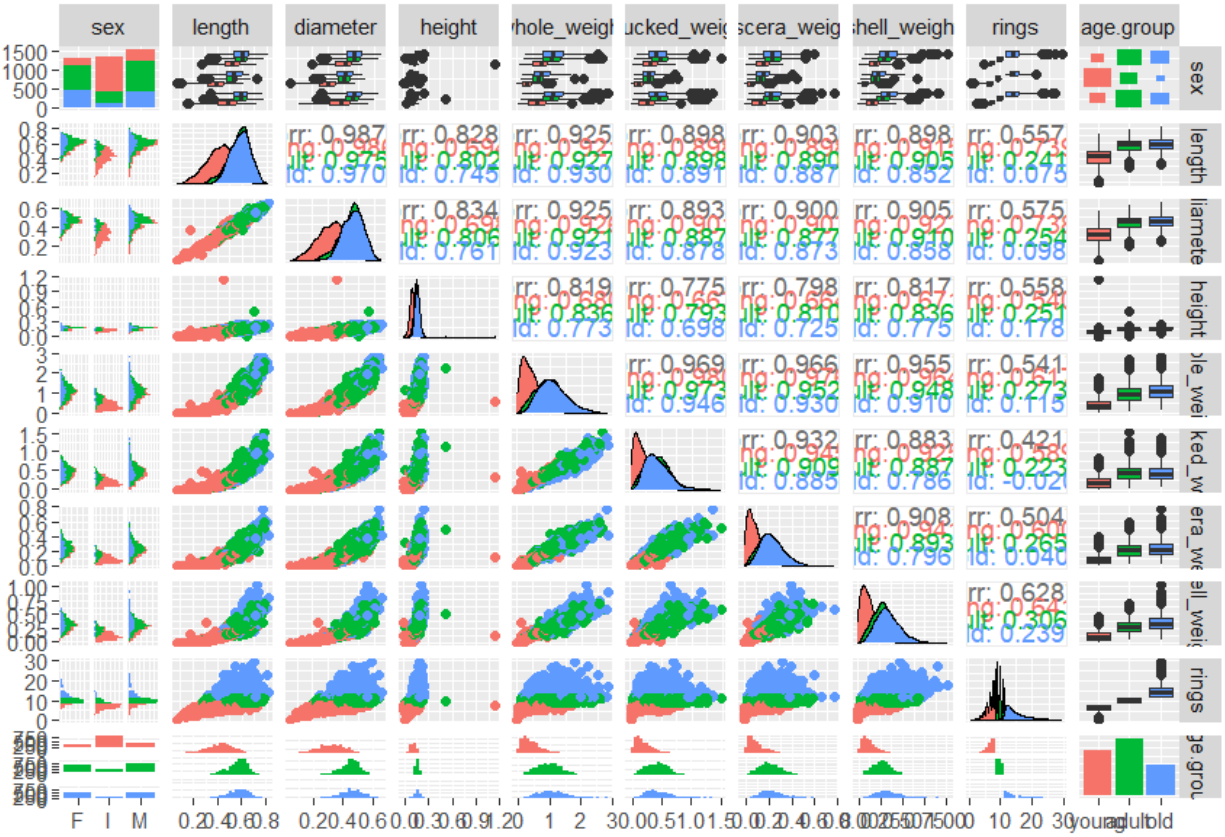
```
> # Compare model performance
> mod1.knn
k-Nearest Neighbors

2924 samples
   3 predictor
   3 classes: 'young', 'adult', 'old'

No pre-processing
Resampling: Cross-Validated (20 fold)
Summary of sample sizes: 2778, 2777, 2777, 2778, 2779, 2779, ...
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  5  0.6007982  0.3697740
  7  0.5974110  0.3624127
  9  0.6083633  0.3788473

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 9.
> mod2.knn
k-Nearest Neighbors

2924 samples
   4 predictor
   3 classes: 'young', 'adult', 'old'

No pre-processing
Resampling: Cross-Validated (20 fold)
Summary of sample sizes: 2777, 2778, 2778, 2777, 2778, 2778, ...
Resampling results across tuning parameters:

  k  Accuracy   Kappa
  5  0.6446518  0.4409157
  7  0.6593971  0.4617910
  9  0.6569786  0.4568936

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 7.
```

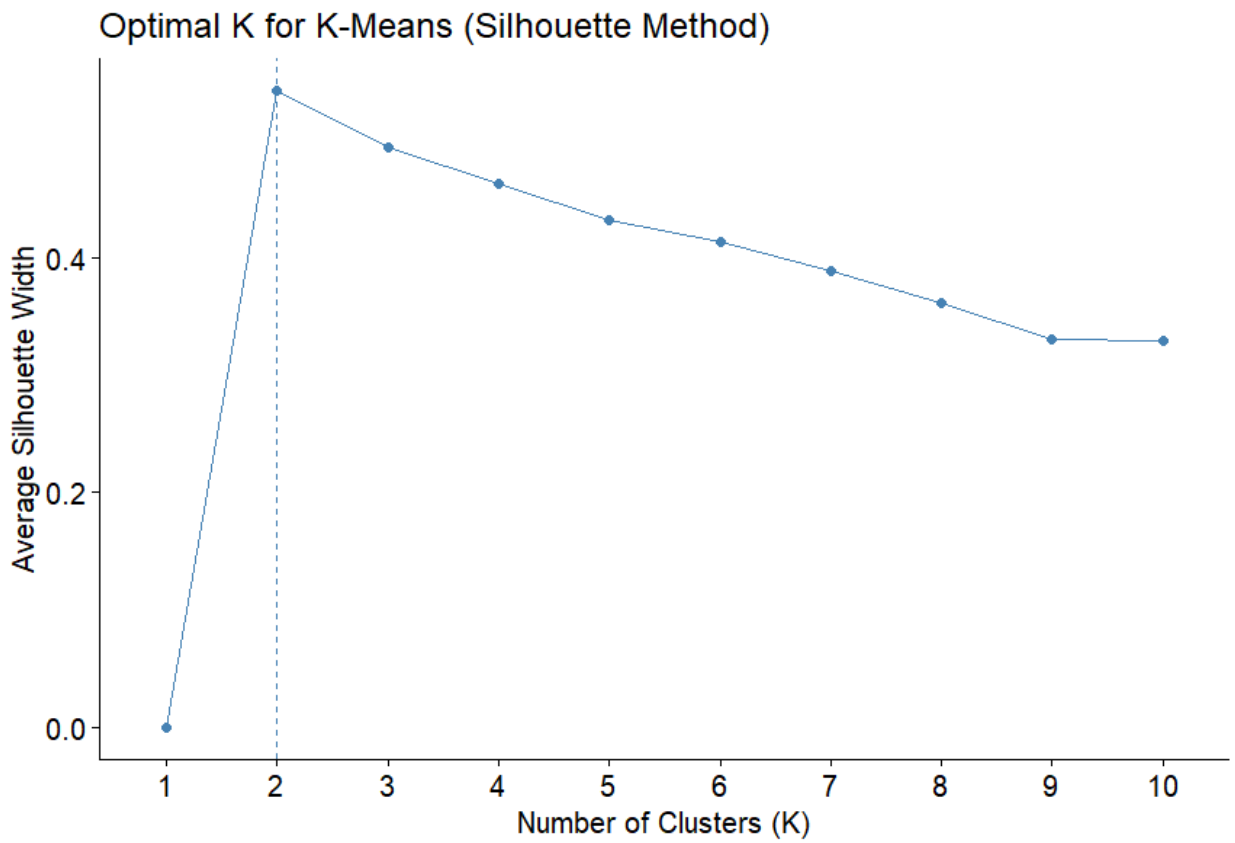
```
> # Display results
> cm1
Confusion Matrix and Statistics

          Reference
Prediction young adult old
     young   284    89  23
     adult   115   359 179
     old      23    95  85

Overall Statistics

               Accuracy : 0.5815
                 95% CI : (0.5536, 0.609)
    No Information Rate : 0.4337
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3357

 Mcnemar's Test P-Value : 2.17e-06

Statistics by Class:

                     Class: young Class: adult Class: old
Sensitivity                0.6730       0.6611    0.29617
Specificity                0.8651       0.5853    0.87772
Pos Pred Value             0.7172       0.5498    0.41872
Neg Pred Value             0.8388       0.6928    0.80744
Prevalence                 0.3371       0.4337    0.22923
Detection Rate             0.2268       0.2867    0.06789
Detection Prevalence       0.3163       0.5216    0.16214
Balanced Accuracy          0.7690       0.6232    0.58694
> cm2
Confusion Matrix and Statistics

          Reference
Prediction young adult old
     young   304    80  17
     adult   114   382 120
     old       4    81 150

Overall Statistics

               Accuracy : 0.6677
                 95% CI : (0.6409, 0.6938)
    No Information Rate : 0.4337
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4773
```

```
 Mcnemar's Test P-Value : 8.001e-05

Statistics by Class:

                     Class: young Class: adult Class: old
Sensitivity                0.7204       0.7035     0.5226
Specificity                0.8831       0.6700     0.9119
Pos Pred Value             0.7581       0.6201     0.6383
Neg Pred Value             0.8613       0.7469     0.8653
Prevalence                 0.3371       0.4337     0.2292
Detection Rate             0.2428       0.3051     0.1198
Detection Prevalence       0.3203       0.4920     0.1877
Balanced Accuracy          0.8018       0.6867     0.7173

> # Choose better model and corresponding feature set
> if (acc2 > acc1) {
+  cat("\nModel 2 performed better. Proceeding with feature subset 2.\n")
+  best_features <- features2
+ } else {
+  cat("\nModel 1 performed better. Proceeding with feature subset 1.\n")
+  best_features <- features1
+ }

Model 2 performed better. Proceeding with feature subset 2.

> cat("\nBest k:", best_k, "with accuracy:", round(max(accuracies), 4), "\n")

Best k: 15 with accuracy: 0.6973
```

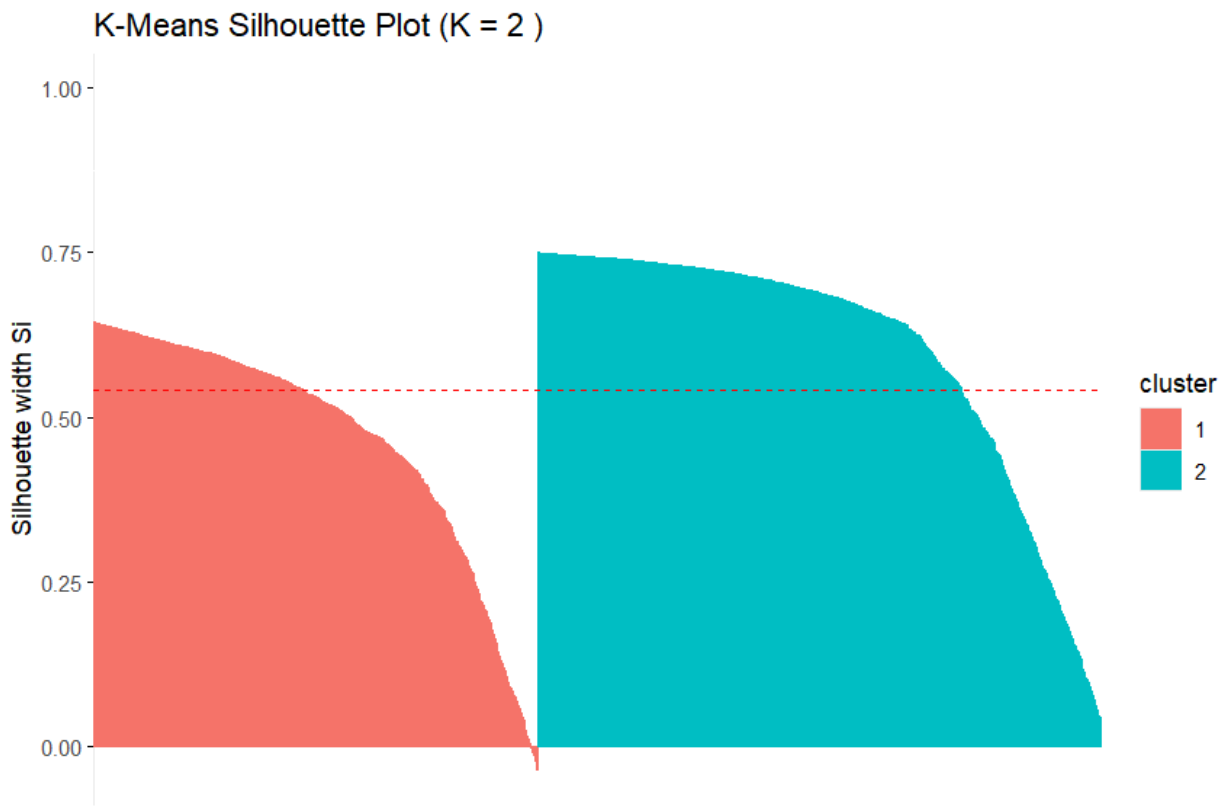

kNN Tuning Curve
Accuracy vs. Number of Neighbors (k)
0.700
0.675
0.650
0.625
0.600
Accuracy
4
8
12
k (Number of Neighbors)

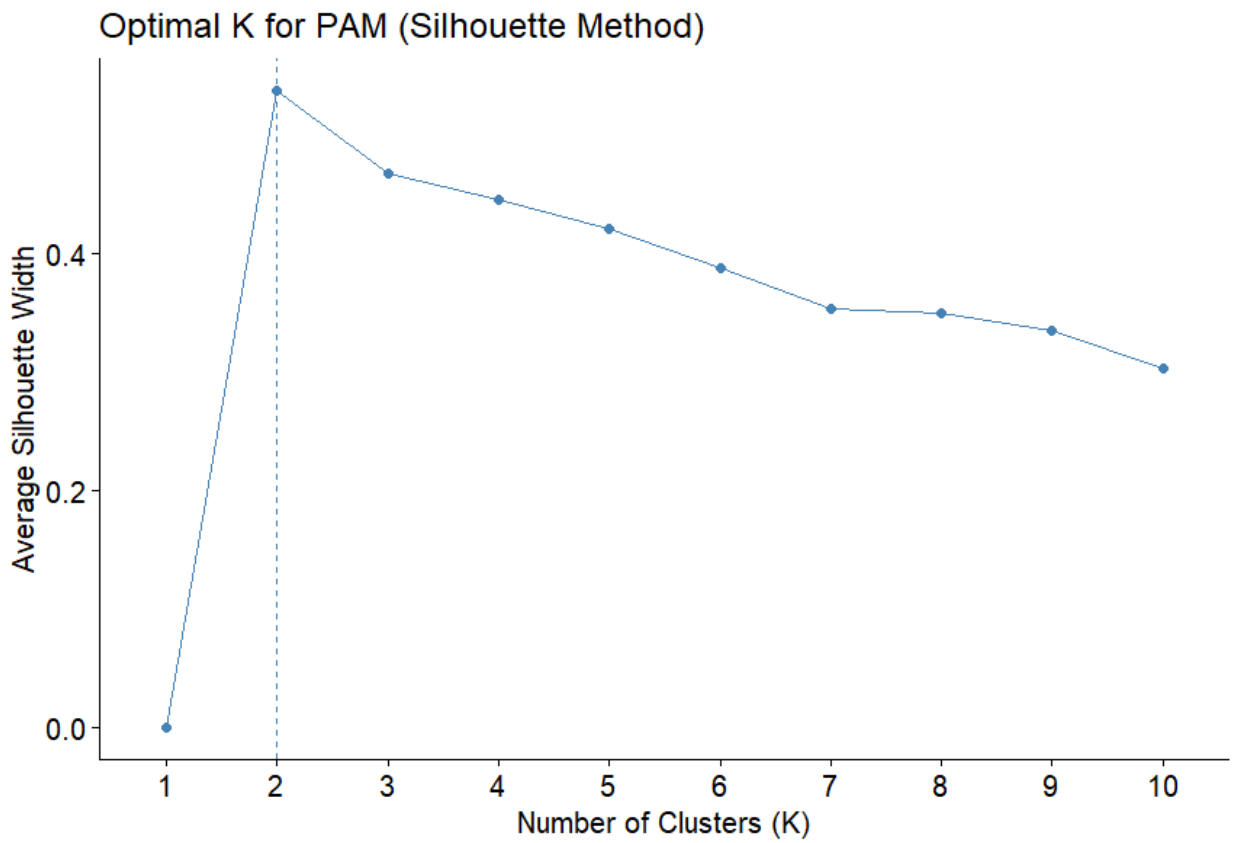

sex*
length
diameter
height
shucked_weight
viscera_weight
shell_weight
rings
age.group
-0.04 -0.04 -0.04 0.00 -0.03 -0.03 -0.04 -0.04
0.99 0.83 0.90 0.90 0.90 0.56 0.55
0.83 0.89 0.90 0.91 0.58 0.57
0.77 0.80 0.82 0.56 0.54
0.93 0.88 0.42 0.45
0.91 0.50 0.52
0.63 0.61
0.86


sex
length
diameter
height
rings
age.group
F
I
M
young
adult
old

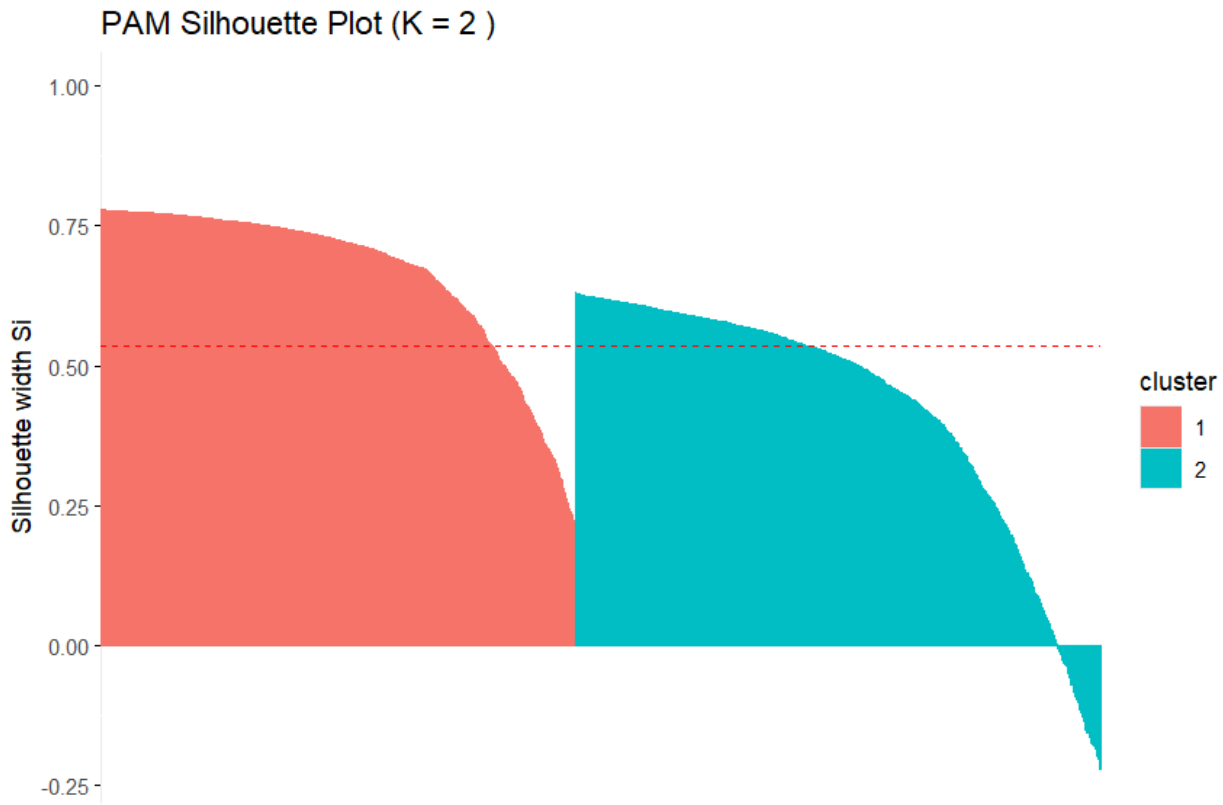



```
> fviz_silhouette(silhouette(kmeans_model$cluster, dist(train.scaled))) +
+  labs(title = paste("K-Means Silhouette Plot (K =", best_K_kmeans, ")"))
  cluster size ave.sil.width
1       1 1290          0.47
2       2 1634          0.60
```


K-Means Silhouette Plot (K = 2)
Silhouette width Si
1.00
0.75
0.50
0.25
0.00
cluster
1
2




```
> fviz_silhouette(pam_model) +
+  labs(title = paste("PAM Silhouette Plot (K =", best_K_pam, ")"))
  cluster size ave.sil.width
1       1 1388          0.66
2       2 1536          0.42
```


PAM Silhouette Plot (K = 2)
Silhouette width Si
1.00
0.75
0.50
0.25
0.00
-0.25
cluster
1
2