Multivariate Crabs Age Analysis: An Integrative Approach Using Dimensionality Reduction and Clustering

200018630 - Group 4

March 30, 2024

1 Introduction

The study of marine life offers valuable insights into biodiversity, ecological balance, and the sustainability of marine food sources. Among the various aspects of marine biology, understanding the age of crabs from their physical attributes serves dual purposes: aiding marine biologists in population dynamics studies and assisting crab farmers in optimizing harvest cycles for sustainability. We employ a dataset of 3893 observations with 9 measurements features (size, weight, sex, etc.) (Sidhu, 2021). The project aims to identify suitable dimensionality reduction and clustering techniques, followed by the development of a predictive model for crab age estimation.

2 Method

2.1 Data Preparation

Data preprocessing ensured data quality. The dataset primarily comprised numerical features, with 'sex' as the sole categorical variable, and the dataset do not have missing values. Consequently, preprocessing focused on scaling numerical features and encoding the categorical "sex" variable. Specifically, one-hot encoding (Geron, 2019) was applied to create binary columns for "Female," "Male," and "Indeterminate."

2.2 Exploratory Data Analysis (EDA)

In the preliminary phase of our Exploratory Data Analysis, we constructed a correlation matrix to investigate the relationships between various physical

characteristics of crabs and their age. The correlation matrix reveals a strong positive correlation between the crab's length, diameter, height, and weight (Figure 1). This suggests that these physical attributes increase proportionally with crab growth. Consequently, for dimensionality reduction purposes, any of these variables could serve as a suitable proxy for overall crab size. Age shows a positive correlation with crab size indicators (length, diameter, height, weight), suggesting that crabs generally grow larger as they age. Sex (Female, Male, Indeterminate) shows weak correlation with other physical crab measurements.

The insights garnered from the correlation analysis, along with the subsequent visual examinations, underline the feasibility of reducing the dimensionality of our dataset. We started with Multidimensional Scaling (MDS) and Principal Component Analysis (PCA) (Everitt and Hothorn, 2011) to reduce the data to two dimensions.

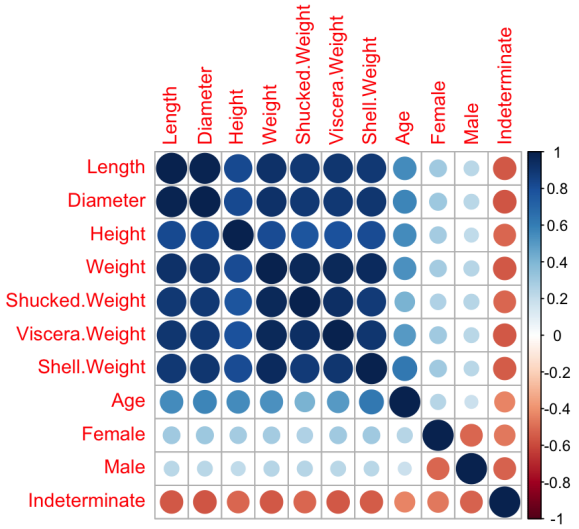
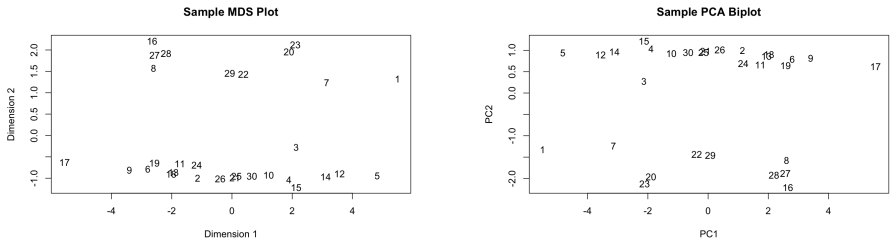


Figure 1: Correlation Matrix with Encoded Sex

2.3 Dimensionality Reduction and Clustering

To assess dimensionality reduction suitability, a stratified sample of 30 data points was analyzed using MDS and PCA. PCA is a dimensionality reduction technique that projects data onto a lower-dimensional space composed of orthogonal 'principal components' that maximize the captured variance (Everitt and Hothorn, 2011). On the other hand, MDS is a visualization

technique that maps high-dimensional data into a lower-dimensional space while attempting to preserve the pairwise distances between data points (Everitt and Hothorn, 2011). Given the metric nature of the data, Euclidean distance MDS theoretically yield results equivalent to the first two principal components of PCA. Visual inspection of the MDS plot and the first two principal components suggests a congruent representation of the underlying data structure (Figure 2a, 2b), despite the potential mirroring artifact in PCA due to the arbitrary sign of eigenvectors (Everitt and Hothorn, 2011).



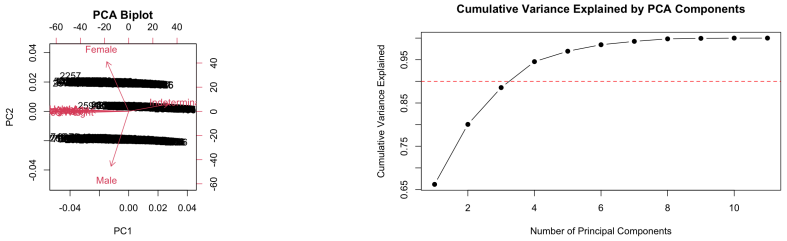
(a) MDS for Selected Samples (b) PCA for Selected Samples

Figure 2: Comparison of MDS and PCA for Selected Samples

Given the computationally intensive nature of MDS (need to compute the pairwise distance between all observations), especially with the full dataset comprising 3,800 observations (3800 x 3800 Matrix), PCA was chosen as the sole dimensionality reduction technique for the entire dataset.

We employed the `prcomp` function in R, configured to retain two principal components. Looking at the Biplot (Figure 3a), the lengths and directions of these vectors indicated the variables' contribution to each principal component, thus informing us of their significance in explaining variance within the dataset (Zelterman, 2015). Variables associated with size, such as `diameter` and various `weight` measurements, predominantly contribute to the first principal component (PC1), with the reaming variance explained by `Sex` attributes. The cumulative plot revealed that approximately 80% of the total variance in the dataset is explained by the first two components (Figure 3b), a substantial amount that can justify the use of these components for further clustering analysis.

The plot derived from the Principal Component Analysis (PCA) reduction suggests the presence of two or three distinct clusters within the data. Subsequent K-means clustering on the principal components distilled from PCA aimed to uncover natural groupings, potentially reflective of the crabs' biological and demographic distinctions. K-means clustering is an algorithm



(a) PCA Biplot of Crab Data (b) PCA % of Variance Explained

Figure 3: Combined PCA Analysis Results

that partitions data points into a predetermined number (k) of clusters by iteratively minimizing the distance between data points and their assigned cluster centroids (Everitt and Hothorn, 2011). The implementation of k-means clustering on the two PCs isolated two (or potentially three) central clusters, as revealed by the clustering plot (Figure 4). The silhouette method, applied to the PCA-reduced data, was then employed to assess the optimal clusters. This analysis, alongside the within-cluster sum of squares method, known as the elbow method, was extended to both the PCA-reduced and the full datasets to confirm the optimal number of clusters. Although silhouette score volatile in the PCA-reduced dataset, the consistent 2-cluster result observed across the PCA-reduced and the Original dataset indicates a robust clustering structure (Figures 5a and 5b. See Code Appendix for the Within Sum of Squares, i.e Elbow Plot).

Other clustering method include Hierarchical clustering, which progressively builds (or divides) a hierarchy of data clusters based on their distance metrics. It is thereby computationally intensive similar to MDS method (Everitt and Hothorn, 2011) (Detailed result of the Hierarchical clustering dendrogram can be found in the Code Appendix).

By integrating PCA for dimensionality reduction with K-means clustering, we gain computational efficiency for handling the large datasets to quickly identify clusters within the crab data. The two-cluster solution provides a guidance for subsequent decision tree modeling, allowing us to explore the factors influencing crab age.

2.4 Decision Tree Model

A decision tree model (`rpart` package, R) was fitted using the ANOVA method (analysis of variance method suitable for continuous variable) to ex-

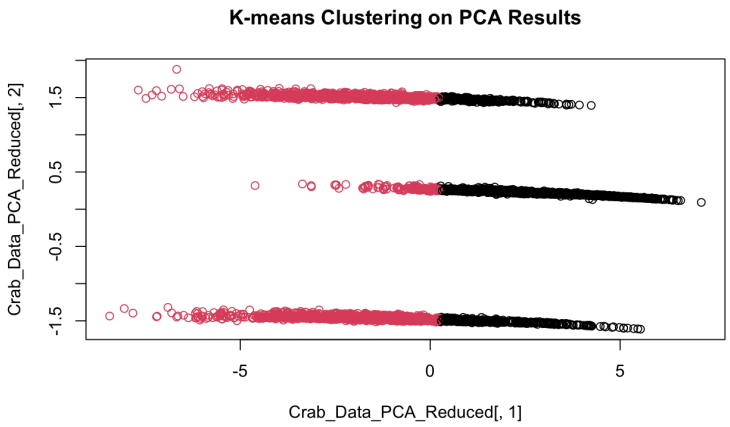
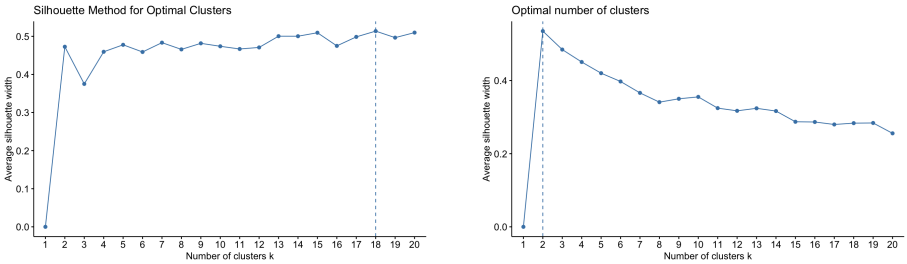


Figure 4: Kmeans Clustering on 2 PCs



(a) Silhouette Method on 2 PCs (b) Silhouette Method on Full Data

Figure 5: Comparison of Silhouette Method on PCA and Full Data

plore relationships between crab attributes (physical measurements and sex) and age. The fitting result reveals that shell weight is the most significant predictor, with the initial split occurring at a shell weight of 4.4. Crabs with a shell weight less than 4.4 are predicted to be younger. The tree further refines age predictions by considering additional splits on shell weight and shucked weight. The model's predictive accuracy was assessed on a held-out test set using Root Mean Squared Error (RMSE), with a result at 2.35.

3 Results

The exploratory analysis revealed significant correlations between crab size attributes and age, suggesting the feasibility of predicting age from these

variables. The PCA and clustering phases further refined the data structure, enabling a focused classification effort. The Decision Tree model demonstrated promising results, with an RMSE (Root Mean Squared Error) on the test data is 2.346, which means the model's predictions deviated from the actual crab ages by an average of 2.346 months.

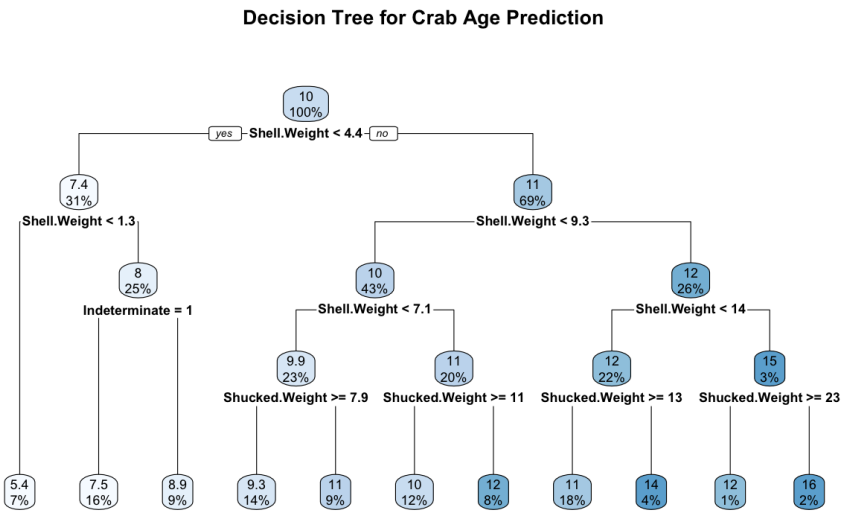


Figure 6: Decision Tree Age Prediction of Crab Data

4 Conclusion

This project applied a multivariate analysis approach to predict crab age from physical attributes. Through data preparation, exploratory analysis, dimensionality reduction, and decision tree age predictive model, the study provided a strong conclusion of the correlation between any single physical dimension (i.e weight) of crab with its age. These findings have the potential to benefit marine biologists and crab farmers to identify the right harvest time in their operations, emphasizing the importance of data-driven approaches in marine biology and aquaculture (Sidhu, 2021).

Crab Age Prediction Appendix: R Markdown

200018630 - Group 4

March 30, 2024

Introduction

This document presents an analysis for predicting the age of crabs using various statistical techniques and data visualization methods. The analysis involves data preprocessing, exploratory data analysis (EDA), dimensionality reduction, clustering, and decision tree model.

Library Import

Import the necessary libraries.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(GGally)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(aplpack)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
library(rpart)
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.2.3
```

```r
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

Data Preparation

Load data, set directory, and preprocess:

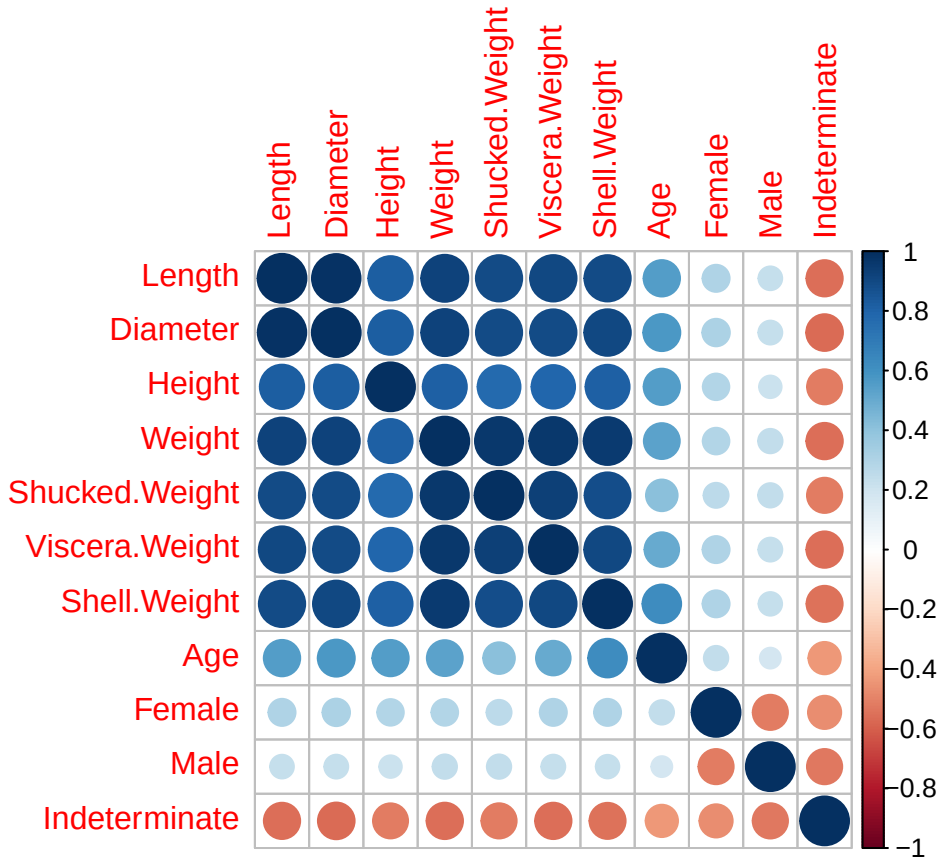
```r
folder_path <- file.path(dirname(rstudioapi::getSourceEditorContext()$path), "/Group Project ")
setwd(folder_path)

# Load and preprocess data
Crab_Data <- read.csv('CrabAgePrediction.csv')
Crab_Data <- Crab_Data %>%
  mutate(Sex = as.factor(Sex)) %>%  # binary column
  mutate(Female = as.integer(Sex == 'F'),
         Male = as.integer(Sex == 'M'),
         Indeterminate = as.integer(Sex == 'I'))
```

Exploratory Data Analysis (EDA)

Correlation Matrix

Correlation matrix of the numerical columns.

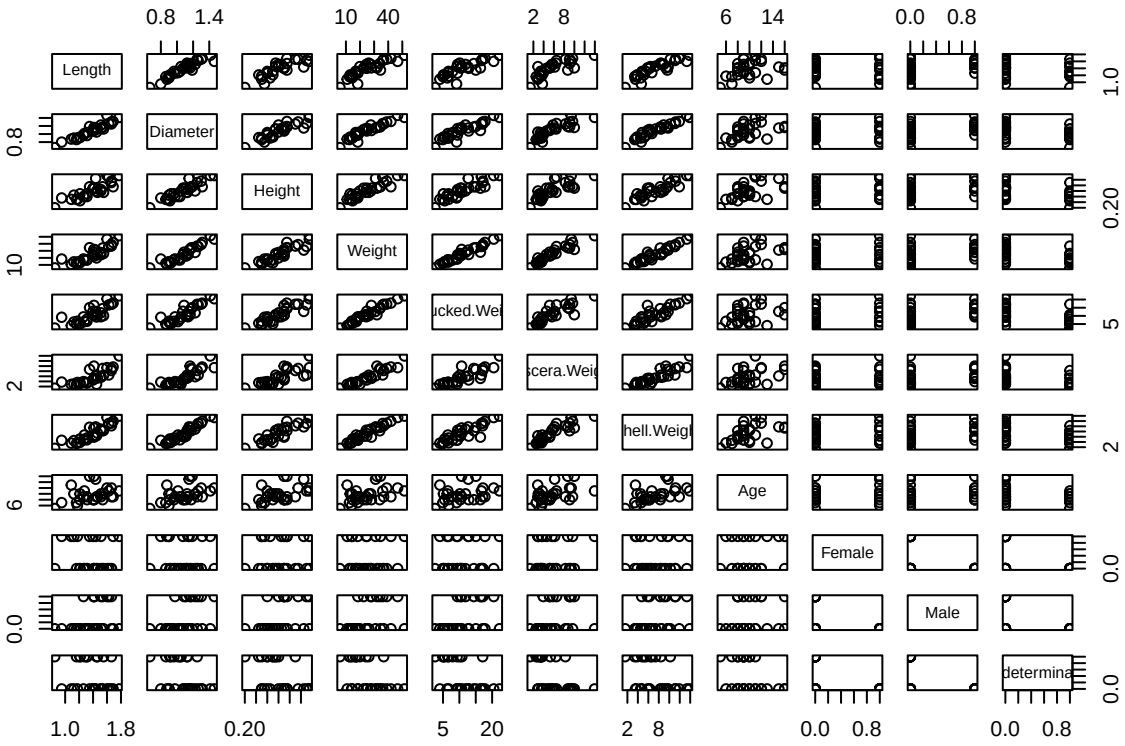
```r
Crab_Data_Corr <- Crab_Data[, sapply(Crab_Data, is.numeric)]
corrplot(cor(Crab_Data_Corr), method = 'circle')
```



Scatterplot Matrix

Scatterplot matrix to understand pairwise relationship between variables:

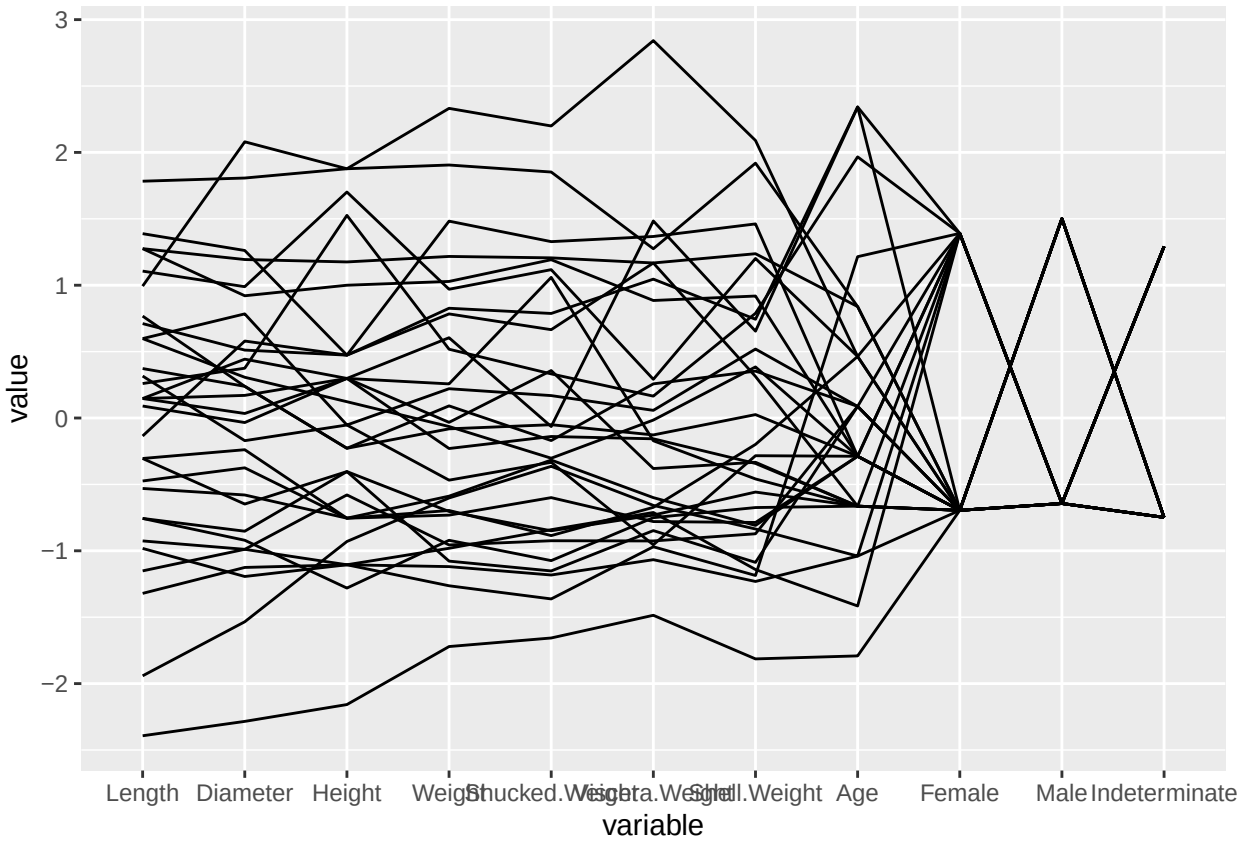
```
set.seed(123)
sample_size <- 30 # Sample
Crab_Data_Sample <- Crab_Data_Corr[sample(1:nrow(Crab_Data), sample_size), ]
pairs(Crab_Data_Sample)
```



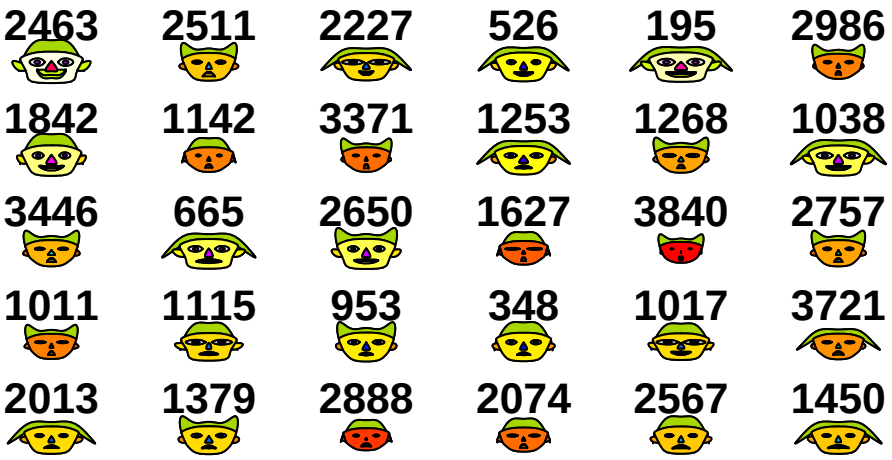
Parallel Coordinates & Chernoff Faces

Parallel Coordinates and Chernoff Faces are used for multivariate data visualization:

```
Crab_Data_Sample_Norm <- as.data.frame(scale(Crab_Data_Sample))

ggparcoord(data = Crab_Data_Sample_Norm %>% select_if(is.numeric),
           groupColumn = NULL)
```




```
faces(Crab_Data_Sample_Norm)
```



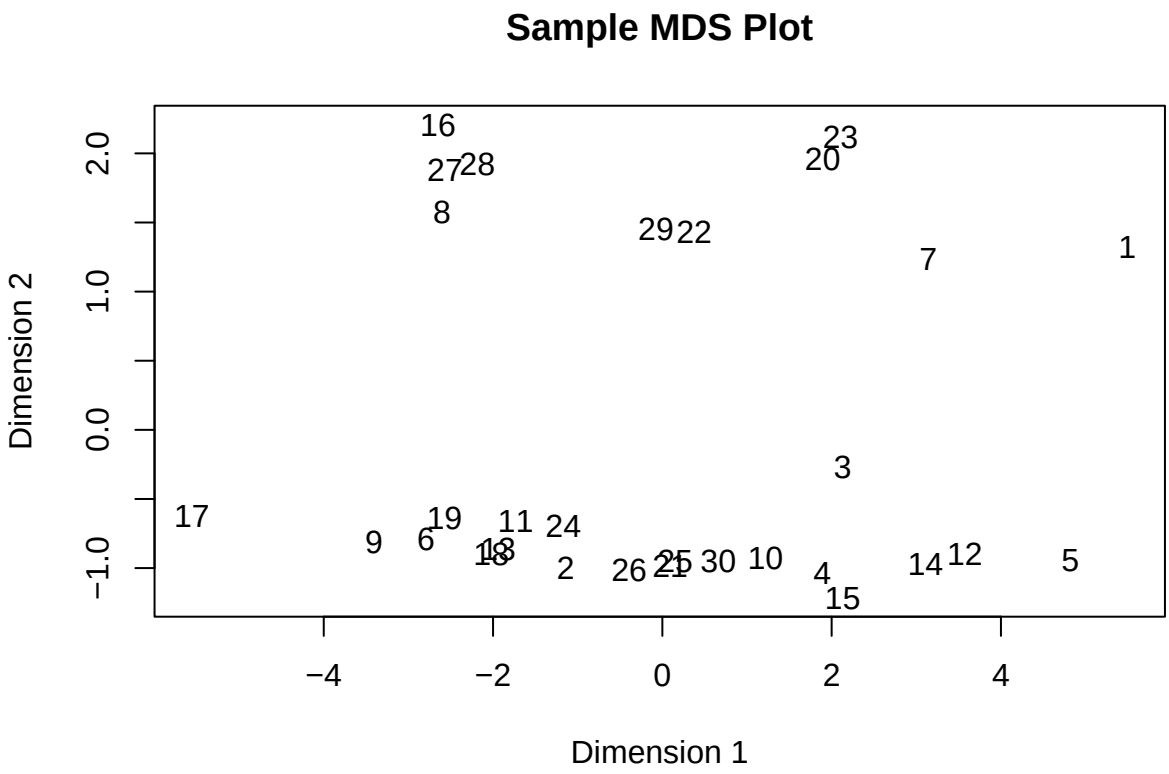

```
## effect of variables:
##  modified item       Var
##  "height of face   " "Length"
##  "width of face    " "Diameter"
##  "structure of face" "Height"
##  "height of mouth  " "Weight"
##  "width of mouth   " "Shucked.Weight"
##  "smiling          " "Viscera.Weight"
##  "height of eyes   " "Shell.Weight"
```

```
##  "width of eyes    " "Age"
##  "height of hair   " "Female"
##  "width of hair    "  "Male"
##  "style of hair    "  "Indeterminate"
##  "height of nose   "  "Length"
##  "width of nose    "  "Diameter"
##  "width of ear     "  "Height"
##  "height of ear    "  "Weight"
```

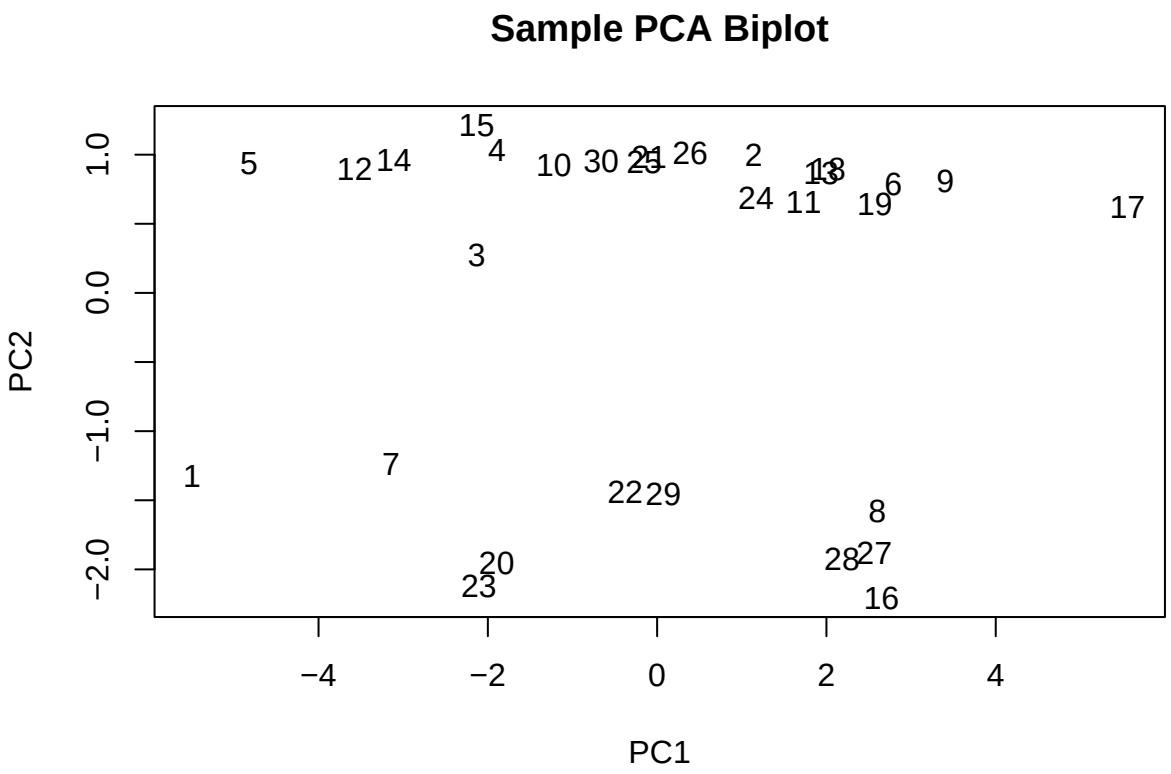
Multidimensional Scaling (MDS) and PCA

Multidimensional Scaling (MDS) and Principal Component Analysis (PCA) to visualize the data in lower dimensions:

```r
# MDS
dist_matrix <- dist(Crab_Data_Sample_Norm)  # Distance matrix
mds_result <- cmdscale(dist_matrix)  # Apply MDS
plot(mds_result[,1], mds_result[,2], xlab="Dimension 1", ylab="Dimension 2",
     main="Sample MDS Plot", type="n")
text(mds_result[,1], mds_result[,2])
```



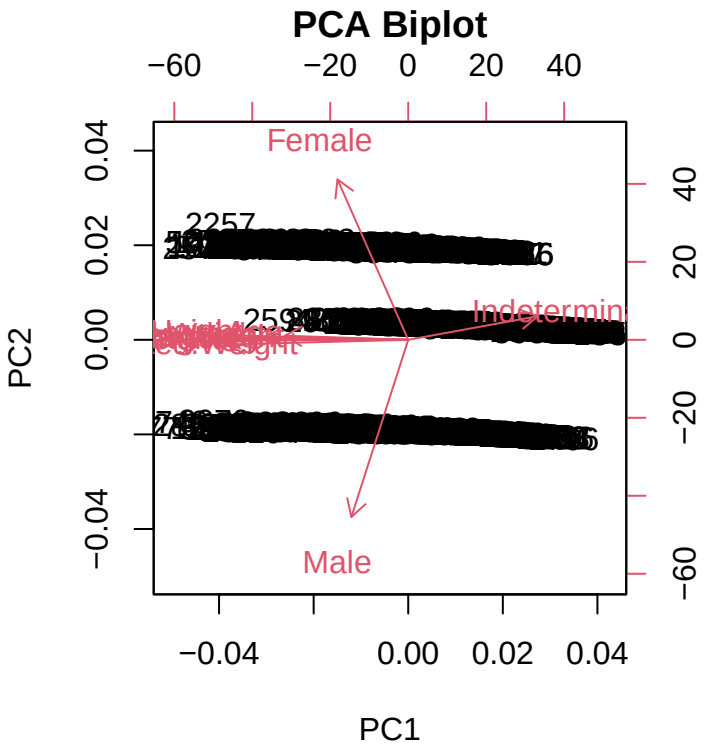

```r
Crab_Data_PCA_Sample <- prcomp(Crab_Data_Sample, scale. = TRUE)
pca_scores <- Crab_Data_PCA_Sample$x[, 1:2]
plot(pca_scores, main = 'Sample PCA Biplot', type="n")
text(pca_scores[,1], pca_scores[,2])
```



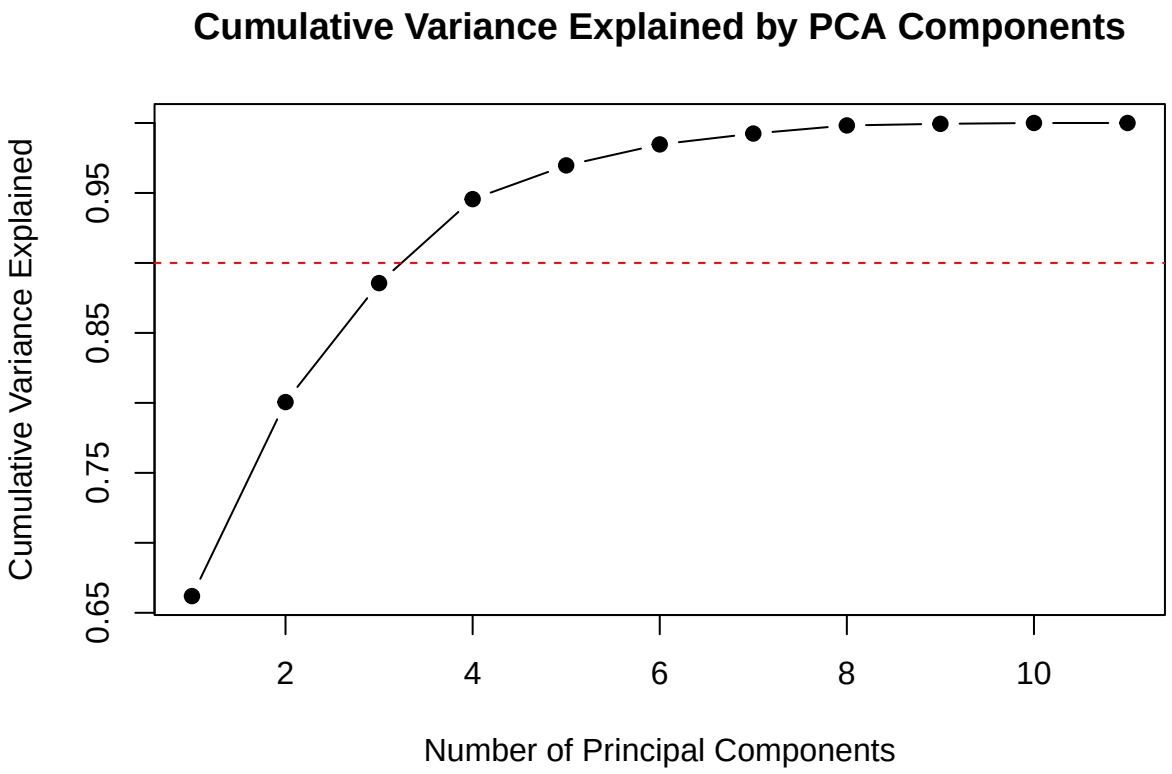
Dimensionality Reduction and Clustering

Perform PCA on the full dataset and explore clustering techniques to identify data patterns:

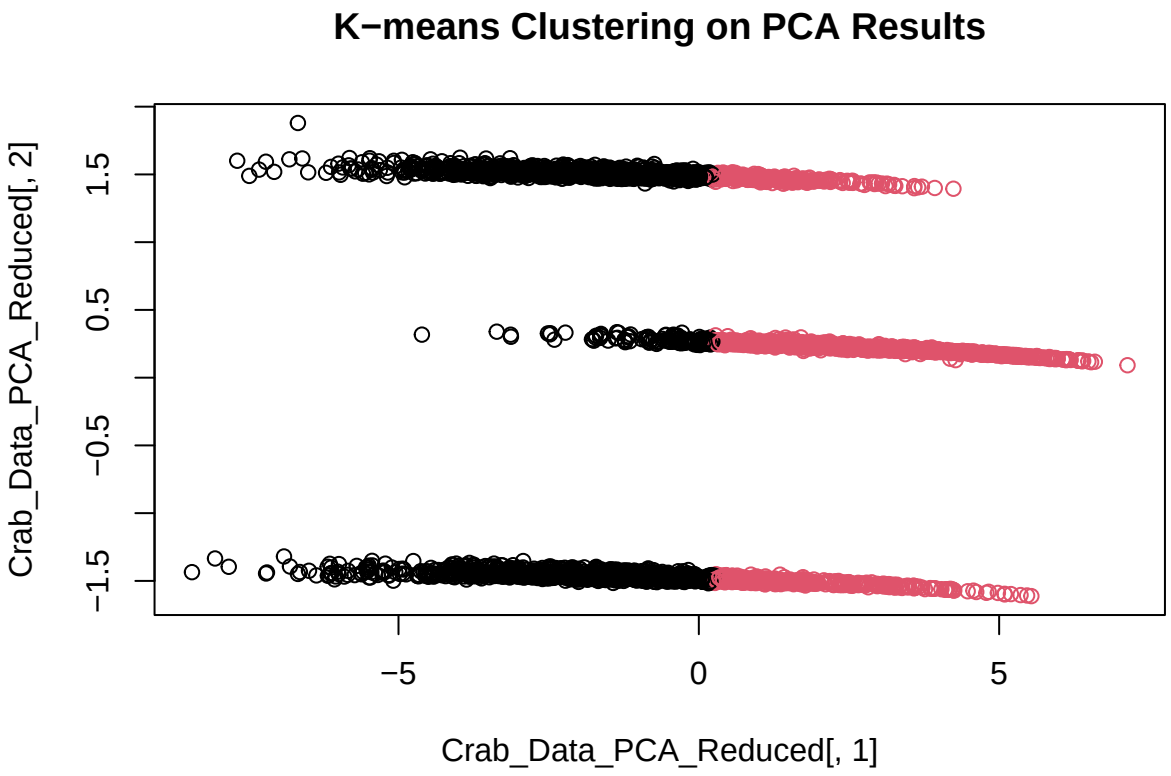
```
# PCA on Full Data
Crab_Data_PCA <- prcomp(Crab_Data_Corr, scale. = TRUE)
biplot(Crab_Data_PCA, main = 'PCA Biplot')
```




```
# Proportion of variance explained by each PCs
variance_explained <- Crab_Data_PCA$sdev^2 / sum(Crab_Data_PCA$sdev^2)
cumulative_variance <- cumsum(variance_explained)
# Cumulative variance plot
plot(cumulative_variance, xlab = "Number of Principal Components",
     ylab = "Cumulative Variance Explained",
     type = "b", pch = 19, main = "Cumulative Variance Explained by PCA Components")
abline(h = 0.9, col = "red", lty = 2)  # Add a line at 90% variance explained for reference
```



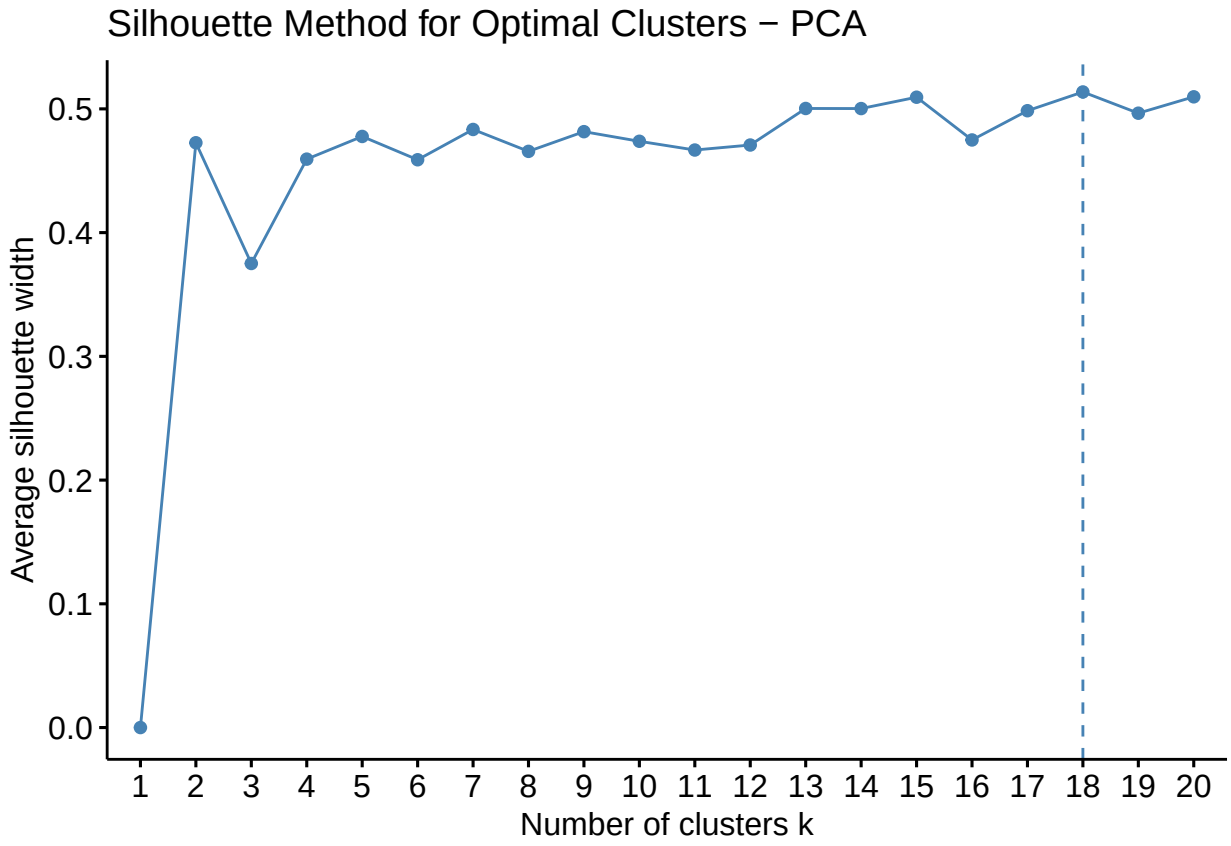

```r
# K-means Clustering based on PCA
Crab_Data_PCA_Reduced <- Crab_Data_PCA$x[, 1:2] # 2 PC
kmeans_result <- kmeans(Crab_Data_PCA_Reduced, centers = 2)
plot(Crab_Data_PCA_Reduced[, 1], Crab_Data_PCA_Reduced[, 2], col = kmeans_result$cluster,
     main = "K-means Clustering on PCA Results")
```



Optimal Clusters Visualization

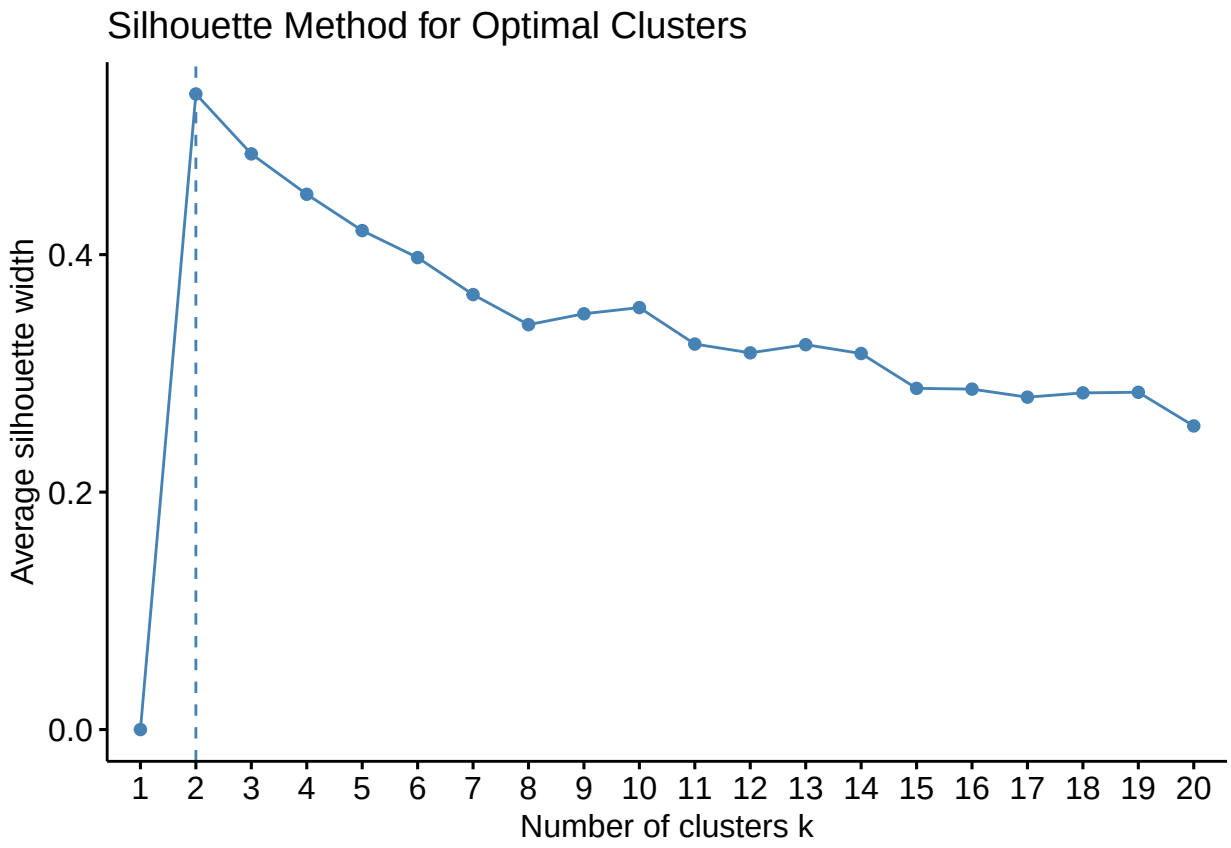
Use silhouette and elbow methods to determine the optimal number of clusters:

```
# Visualize the average silhouette method
fviz_nbclust(Crab_Data_PCA_Reduced, kmeans, method = "silhouette", k.max = 20) +
  ggtitle("Silhouette Method for Optimal Clusters - PCA")
```

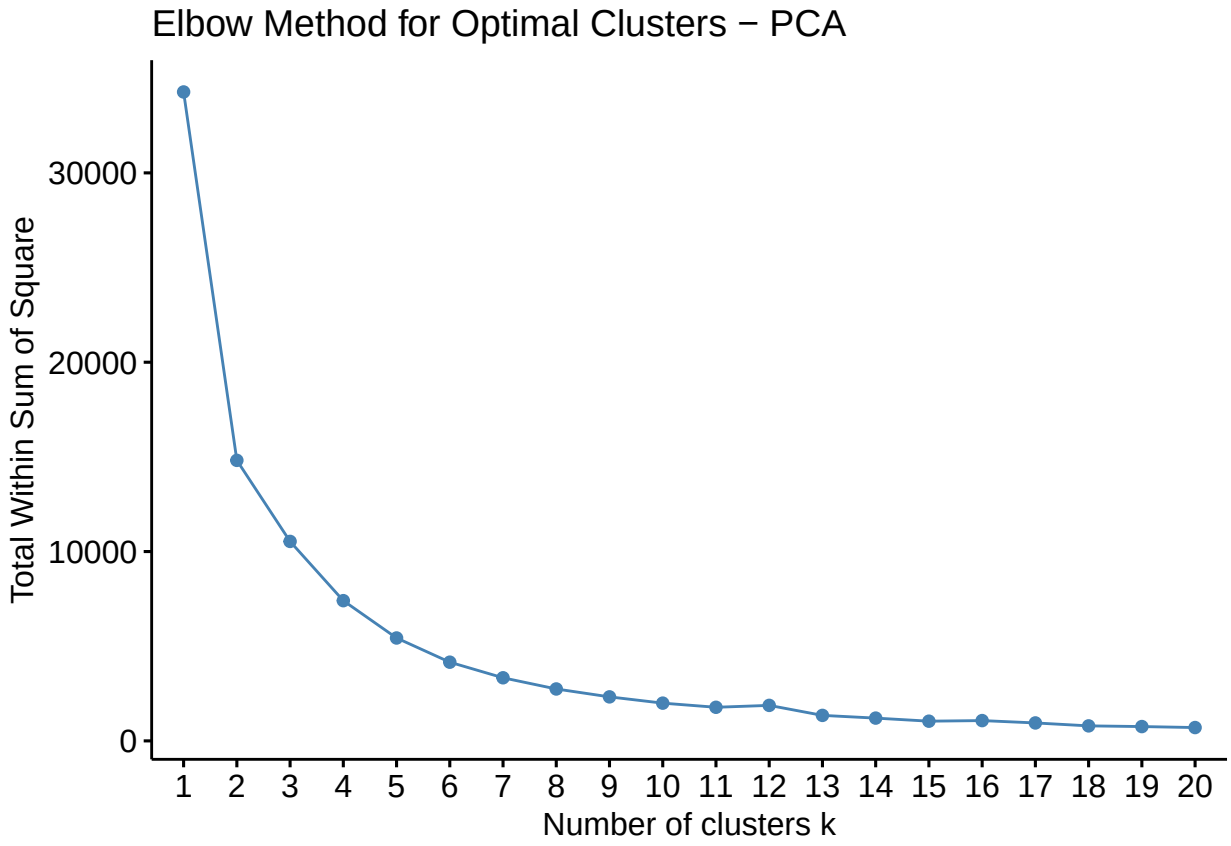



```
fviz_nbclust(Crab_Data_Corr, kmeans, method = "silhouette", k.max = 20) +
  ggtitle("Silhouette Method for Optimal Clusters")
```

```
## Warning: did not converge in 10 iterations
```



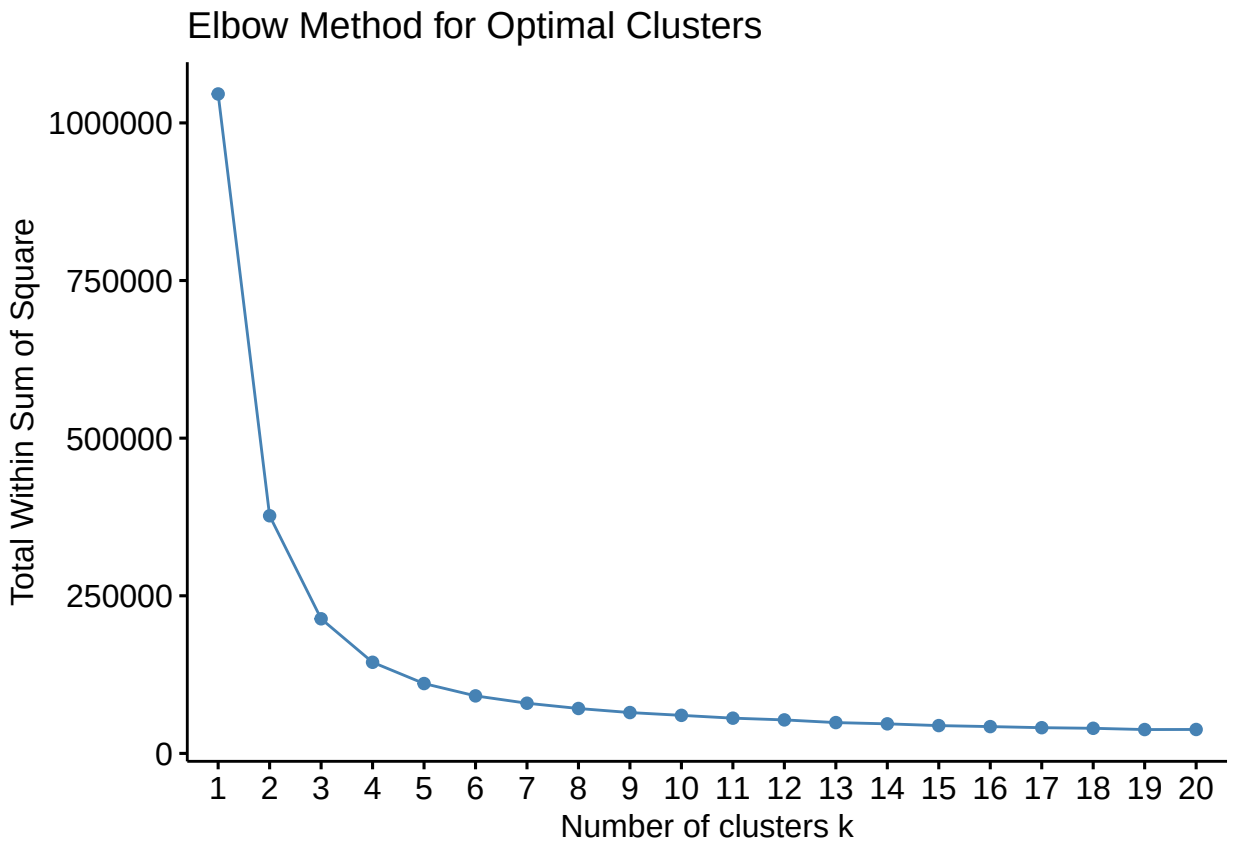

```
# Visualize the within-cluster sum of squares method (elbow method)
fviz_nbclust(Crab_Data_PCA_Reduced, kmeans, method = "wss", k.max = 20) +
  ggtitle("Elbow Method for Optimal Clusters - PCA")
```




```
fviz_nbclust(Crab_Data_Corr, kmeans, method = "wss", k.max = 20) +
  ggtitle("Elbow Method for Optimal Clusters")
```

```
## Warning: did not converge in 10 iterations
```

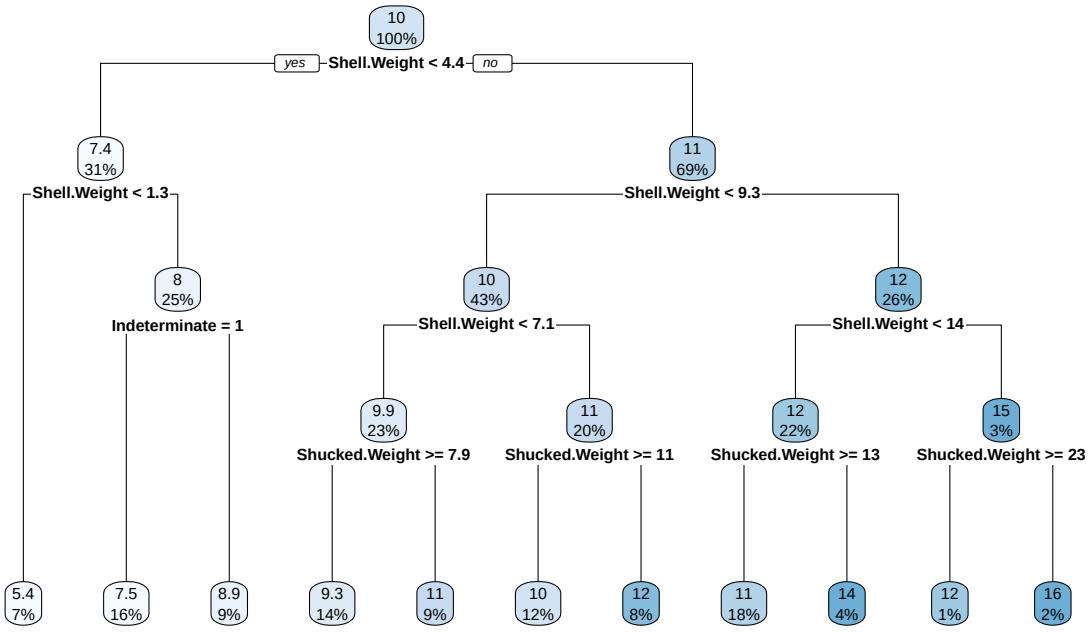
```
## Warning: did not converge in 10 iterations
```



Model Building: Decision Tree

Build a decision tree model to predict crab age:

```
set.seed(123)
train_indices <- sample(1:nrow(Crab_Data), 0.7 * nrow(Crab_Data))
train_data <- Crab_Data[train_indices, ]
test_data <- Crab_Data[-train_indices, ]

model <- rpart(Age ~ ., data = train_data, method = "anova")

# Decision Tree Result
rpart.plot(model, main="Decision Tree for Crab Age Prediction")
```




```
predictions <- predict(model, test_data)
rmse <- sqrt(mean((predictions - test_data$Age)^2))
cat("RMSE on test data:", rmse, "\n")
```

```
## RMSE on test data: 2.346292
```

References

[1] Brian Everitt and Torsten Hothorn. "Principal Components Analysis". In: *An Introduction to Applied Multivariate Analysis with R*. New York, NY: Springer New York, 2011, pp. 61–103. ISBN: 978-1-4419-9650-3. DOI: 10.1007/978-1-4419-9650-3_3. URL: https://doi.org/10.1007/978-1-4419-9650-3_3.

[2] Aurelien Geron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd. O'Reilly Media, Inc., 2019. ISBN: 1492032646.

[3] Gursewak Singh Sidhu. *Crab Age Prediction*. 2021. DOI: 10.34740/KAGGLE/DSV/2834512. URL: https://www.kaggle.com/dsv/2834512.

[4] Daniel Zelterman. "Clustering". In: *Applied Multivariate Statistics with R*. Cham: Springer International Publishing, 2015, pp. 287–313. ISBN: 978-3-319-14093-3. DOI: 10.1007/978-3-319-14093-3_11. URL: https://doi.org/10.1007/978-3-319-14093-3_11.