Predictive Model for Estimating Crab Age

200018630 - Group 4

March 25, 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. This project aims to develop a predictive model to estimate the age of crabs based on measurements such as size, weight, and sex, utilizing a dataset of 3893 observations with 9 dimensions [2].

2 Method

2.1 Data Preparation

The crab dataset was preprocessed to ensure data quality. This included cleaning, normalization, managing missing values, and encoding the categorical 'Sex' variable. Specifically, the 'Sex' variable was converted into binary columns for 'Female,' 'Male,' and 'Indeterminate' to facilitate analysis.

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.

A scatterplot matrix visualizes relationships between pairs of numerical variables from a random sample of 30 crabs. This confirmed the positive correlations observed in the correlation matrix, showing that larger crabs tend to have higher values for length, diameter, height, and weight (Figure 3). A parallel coordinates plot and a Chernoff faces were used to have a more creative and intuitive visualizations of the sampled crabs' data (Figure 2).

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 Principal Component Analysis (PCA) [1] to reduce the data to two dimensions.

2.3 Dimensionality Reduction and Clustering

In the process of dimensionality reduction, a stratified sample of 30 data points from our larger crab dataset was first analyzed using Multidimensional Scaling (MDS) and Principal Component Analysis (PCA). The visual comparison between the MDS (Figure 4) and the first two principal components of the PCA (Figure 5) suggests a congruent representation of the data's inherent structure, despite the mirror reflection which is a common artifact due to the arbitrary sign of eigenvectors in PCA [1]. Given the computationally intensive nature of MDS, especially with the full dataset comprising 3,800 observations, 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 6), 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 [3]. 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 7), a substantial amount that can justify the use of these components for further clustering analysis.

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. The implementation of k-means clustering on the two PCs isolated three (or two) central clusters, as revealed by the clustering plot (Figure 8). 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. Despite the heightened silhouette score in the reduced dataset, the congruence in the 2-cluster result across the majority of methods underscores a stable clustering structure within the data (Figures 9, 10, 11, 12).

Hierarchical clustering was applied to a stratified sample of our dataset to further explore its intrinsic structure (Figure 13). The structure of the dendrogram aligns well with the Silhouette sample result, further reinforcing the

validity of our two-cluster solution.

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 explore relationships between crab attributes (physical measurements and sex) and age. The model's predictive accuracy was assessed on a held-out test set using Root Mean Squared Error (RMSE).

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.

4 Conclusion

This project successfully applied a multivariate analysis approach to predict crab age from physical attributes. Through careful data preparation, insightful exploratory analysis, strategic dimensionality reduction, and effective model building, the study provided a viable predictive model for estimating crab age. These findings have the potential to benefit marine biologists in their research and crab farmers in their operations, emphasizing the importance of data-driven approaches in marine biology and aquaculture.

A Appendix: R Code

```
library(dplyr)
library(GGally)
library(aplpack)
library(corrplot)
library(rpart)
library(rpart.plot)
library(factoextra)

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

# 3. Exploratory Data Analysis (EDA)
# Correlation Matrix
Crab_Data_Corr <- Crab_Data[, sapply(Crab_Data, is.numeric)]
corrplot(cor(Crab_Data_Corr), method = 'circle')

# Scatterplot Matrix
set.seed(123)
sample_size <- 30
Crab_Data_Sample <- Crab_Data_Corr[sample(1:nrow(Crab_Data),
    sample_size), ]
pairs(Crab_Data_Sample)
# pairs(Crab_Data_Corr) # Whole Dataset

# Parallel Coordinates & Chernoff Faces
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)

# 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])

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])
```

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

# Calculate the proportion of variance explained by each principal
    component
variance_explained <- Crab_Data_PCA$sdev^2 / sum(Crab_Data_PCA$sdev^2)
cumulative_variance <- cumsum(variance_explained)

# Plot the cumulative variance to determine how many components to retain
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

# 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")

# 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")
# 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")

# Sample compare
# Hierarchical clustering
hc_result <- hclust(dist_matrix) # Apply hierarchical clustering
plot(hc_result) # Plot the dendrogram
cutree(hc_result, k=3) # Assuming you want to create 3 clusters

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

# 5. Model Building: Decision Tree
# Split Data into Training and Testing
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, ]

# Build Decision Tree Model
model <- rpart(Age ~ ., data = train_data, method = "anova")

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

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

B Appendix: visualizations

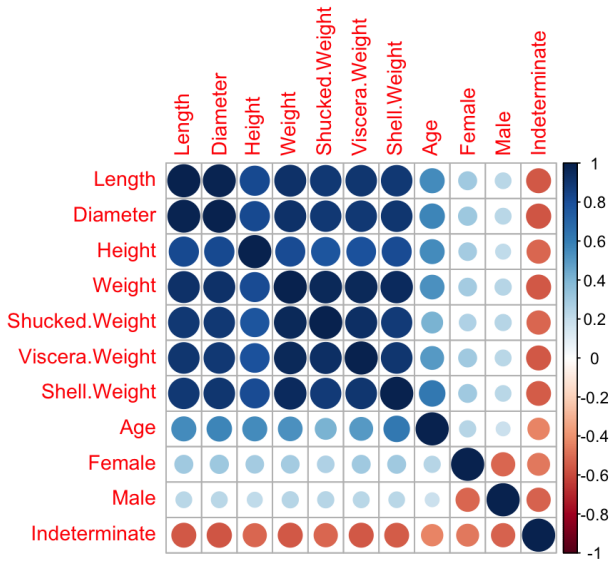

Figure 1: Correlation Matrix with Encoded Sex

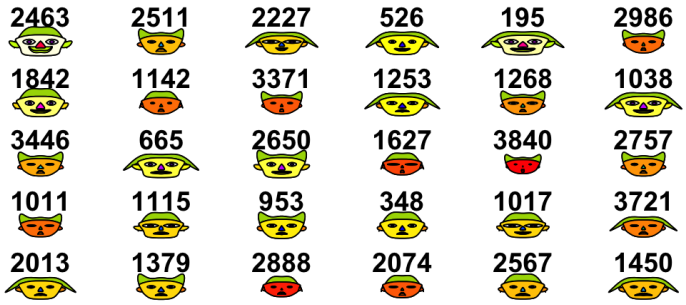

Figure 2: Chernoff Faces for Selected Samples

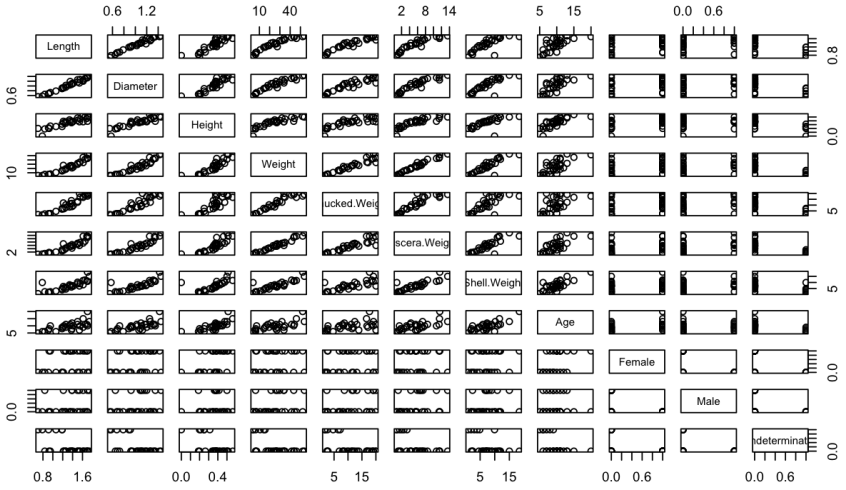

Figure 3: Scatterplot Matrix for Selected Samples

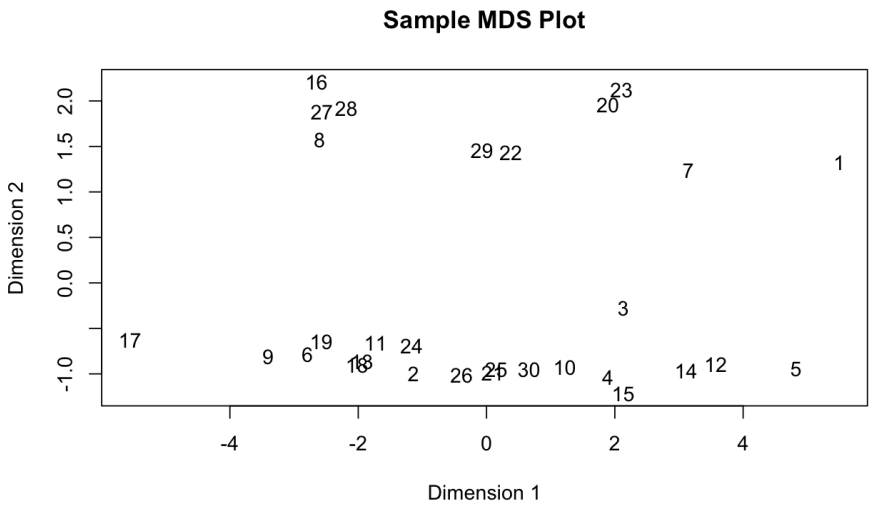

Figure 4: Multidimensional Scaling for Selected Samples

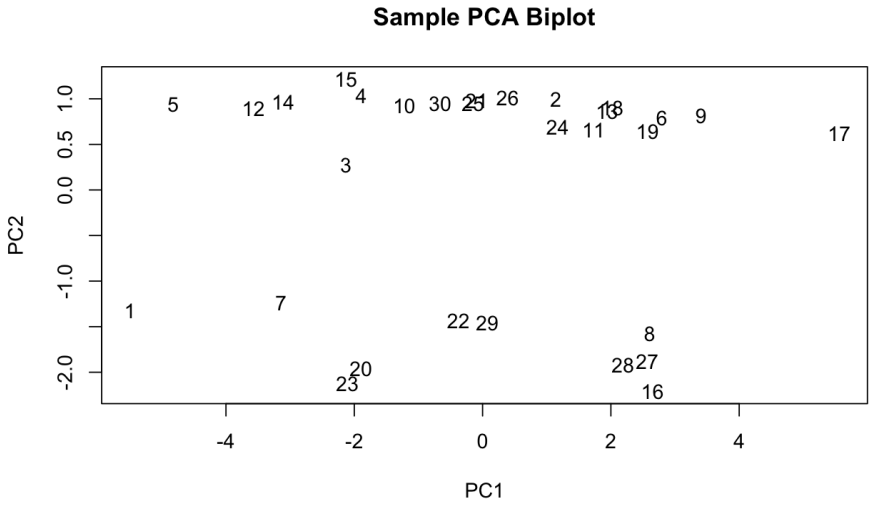

Figure 5: PCA for Selected Samples

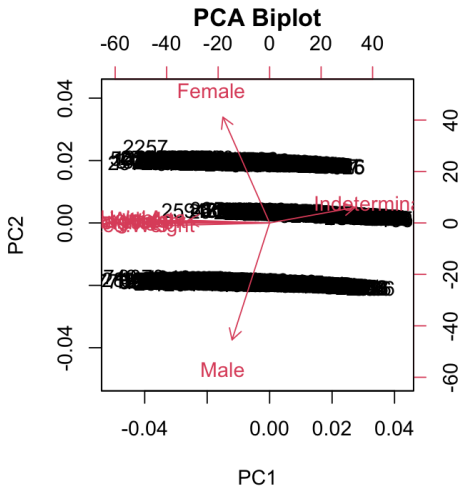

Figure 6: PCA Biplot of Crab Data

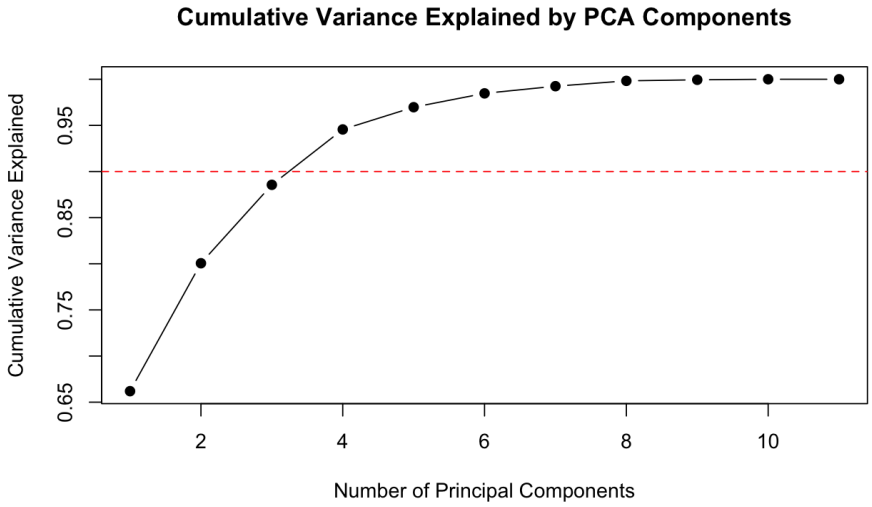

Figure 7: PCA Proportion of Variance Explained

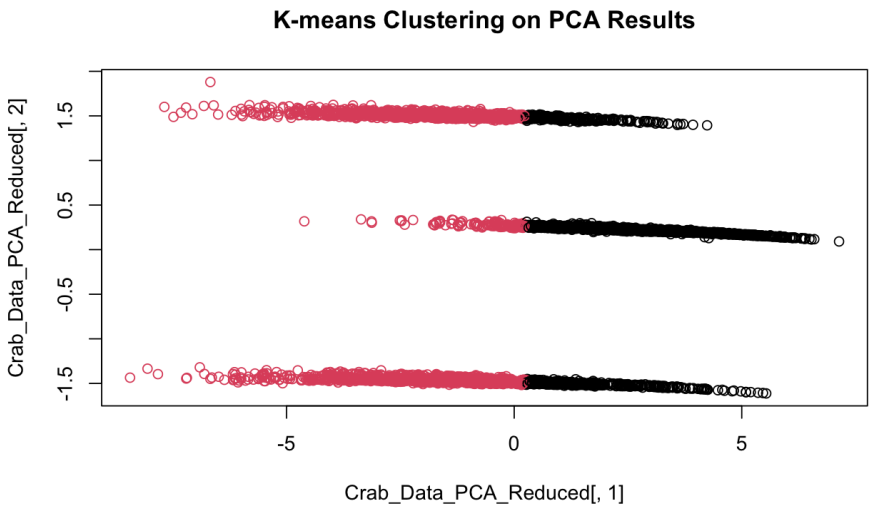

Figure 8: Kmeans Clustering on 2 PCs

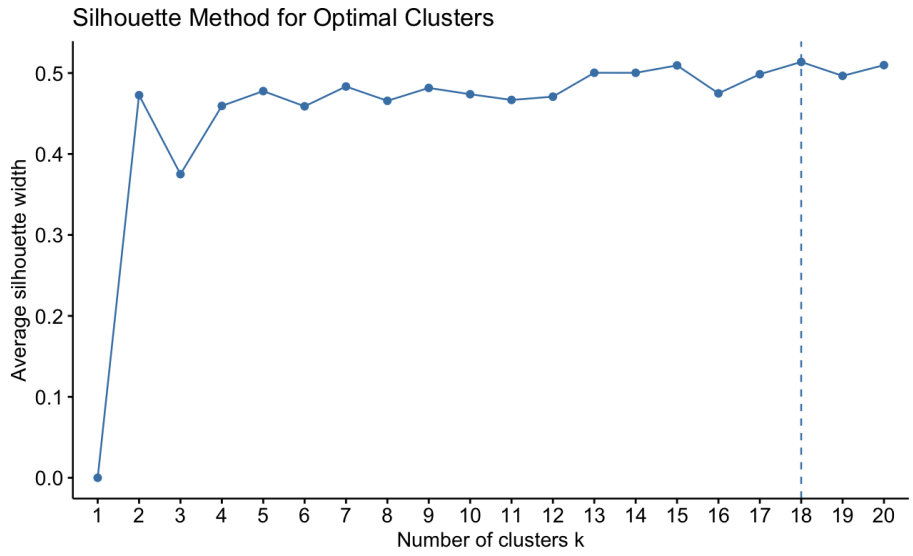

Figure 9: Silhouette Method on 2 PCs

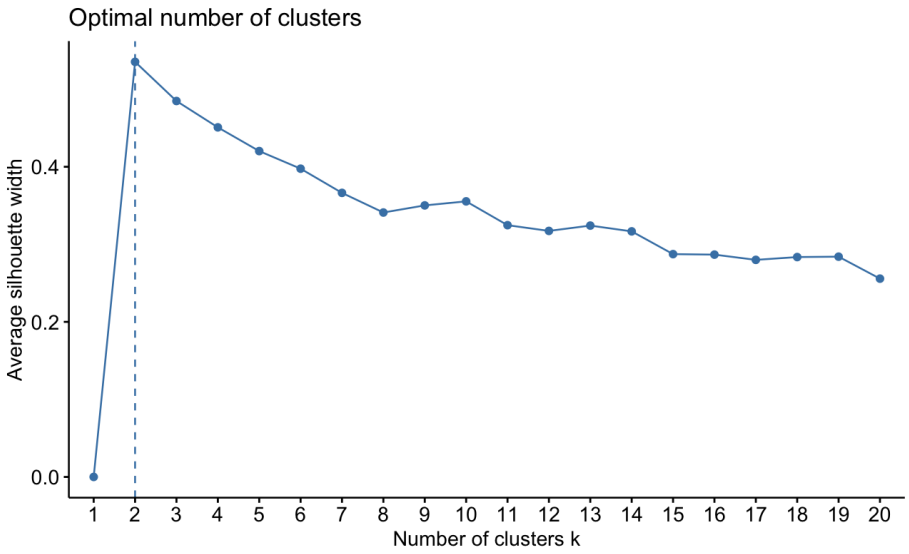

Figure 10: Silhouette Method on Full Data

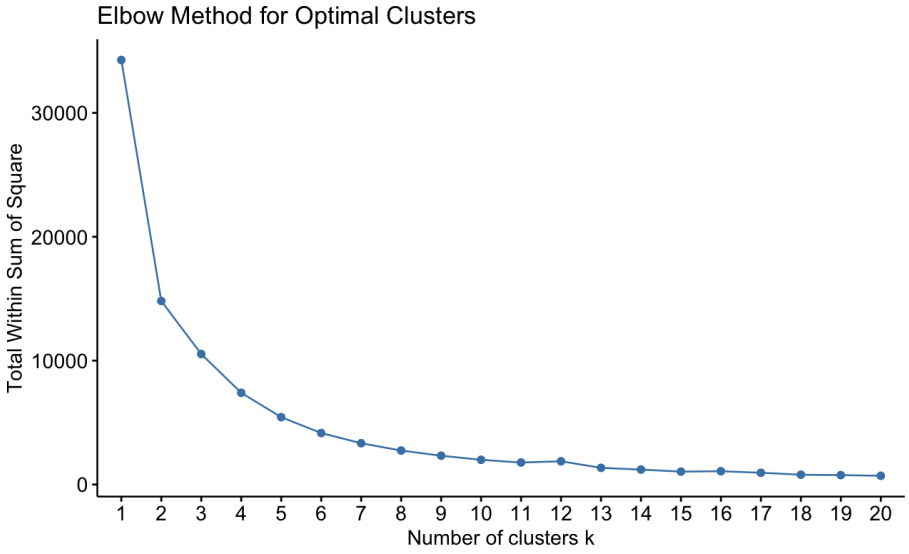

Figure 11: Elbow Method on 2 PCs

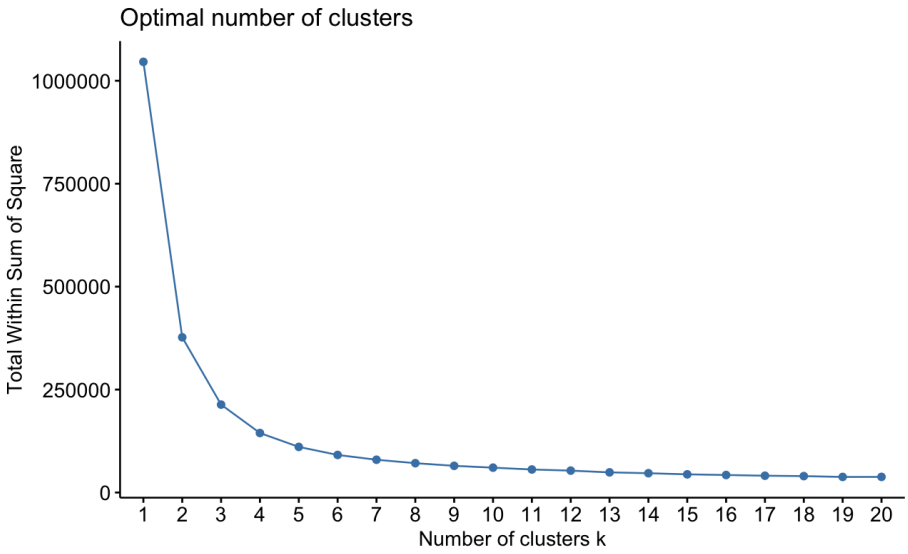

Figure 12: Elbow Method on Full Data

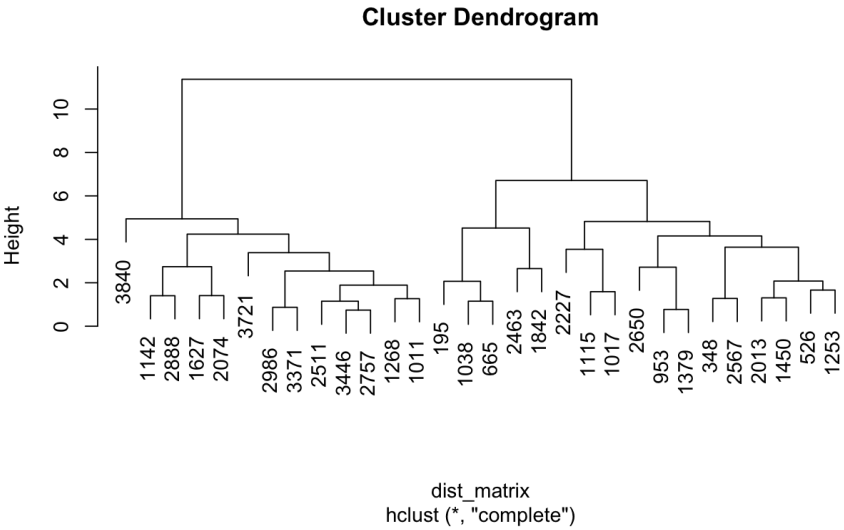

Figure 13: Hierarchical Clustering Dendrogram for Selected Samples

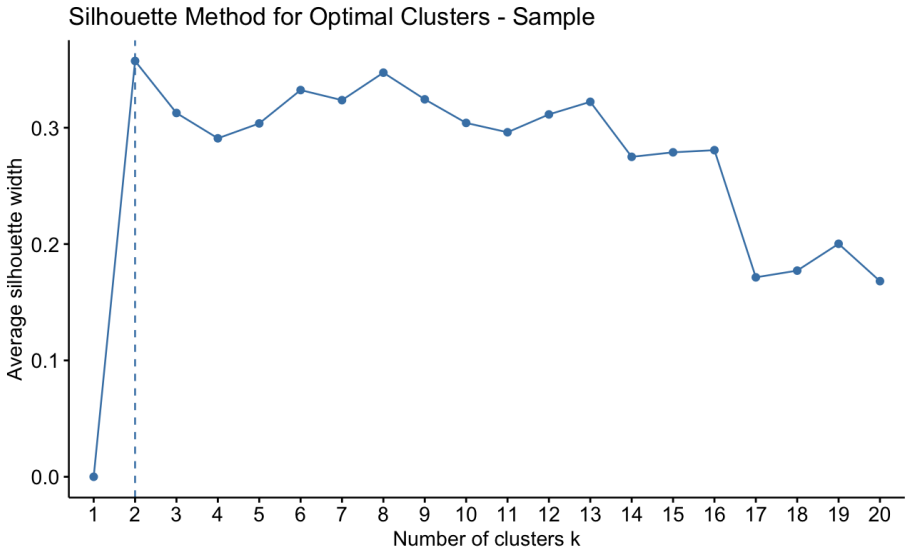

Figure 14: Silhouette Optimal Clusters under PCA for Selected Samples

References

[1] Brian Everitt and Torsten Hothorn. *Principal Components Analysis*, pages 61–103. Springer New York, New York, NY, 2011.

[2] Gursewak Singh Sidhu. Crab age prediction, 2021.

[3] Daniel Zelterman. *Clustering*, pages 287–313. Springer International Publishing, Cham, 2015.