PCA

Dhruvi Mehta (202318003)

Prachi Mehta (202318008)

Riya Dave (202318010)

Simran Dalvi (202318042)

WHAT IS PRINCIPAL COMPONENT ANALYSIS ?

Principal component analysis (PCA) is a dimensionality reduction and machine learning method used to simplify a large data set into a smaller set while still maintaining significant patterns and trends.

HOW DO YOU DO A PRINCIPAL COMPONENT ANALYSIS?

Principal component analysis can be broken down into five steps. I'll go through each step, providing logical explanations of what PCA is doing and simplifying mathematical concepts such as standardization, covariance, eigenvectors and eigenvalues without focusing on how to compute them.

1. Standardize the range of continuous initial variables

2. Compute the covariance matrix to identify correlations

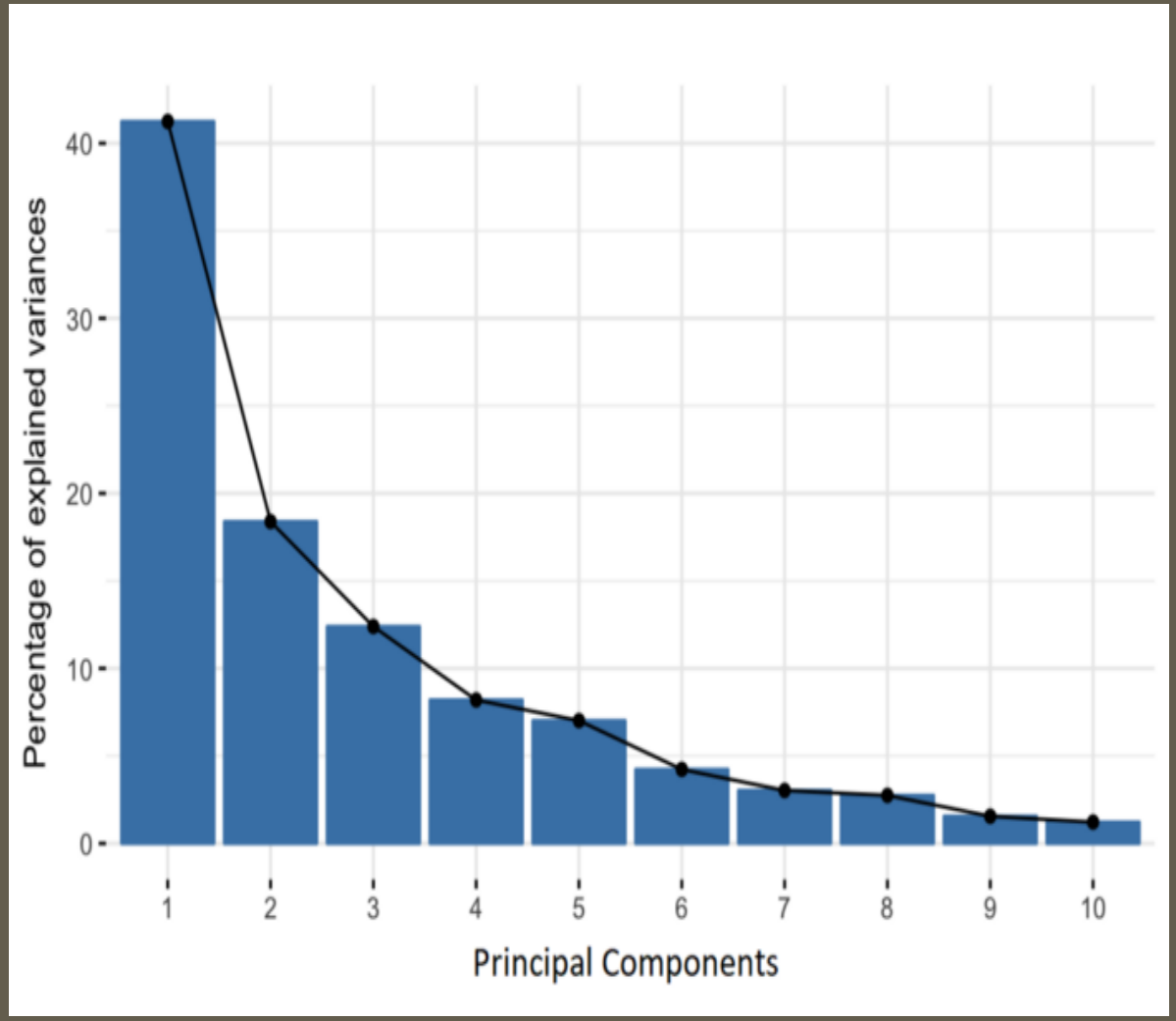
3. Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components

4. Create a feature vector to decide which principal components to keep

5. Recast the data along the principal components axes

WHAT ARE PRINCIPAL COMPONENTS?

Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components. So, the idea is 10-dimensional data gives you 10 principal components, but PCA tries to put maximum possible information in the first component, then maximum remaining information in the second and so on, until having something like shown in the scree plot below.



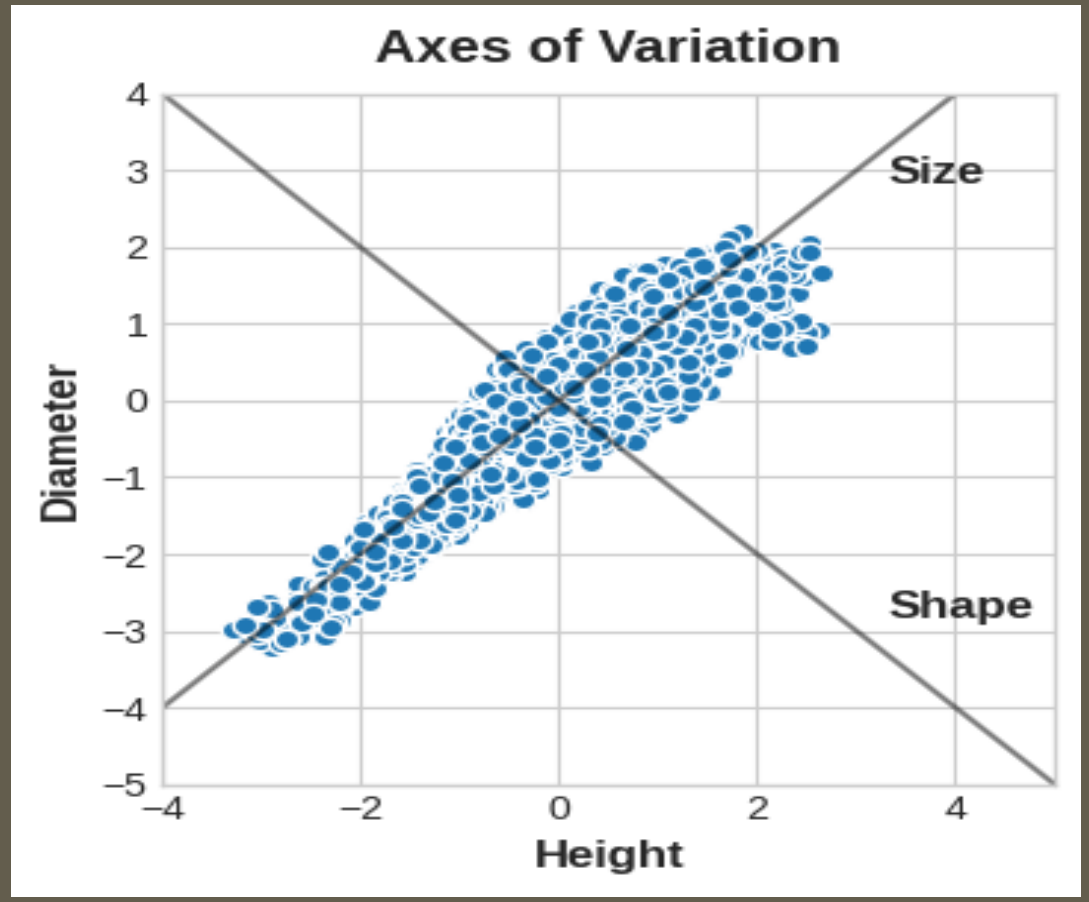
- Within this data are "axes of variation" that describe the ways the Abalone tend to differ from one another. Pictorially, these axes appear as perpendicular lines running along the natural dimensions of the data, one axis for each original feature.

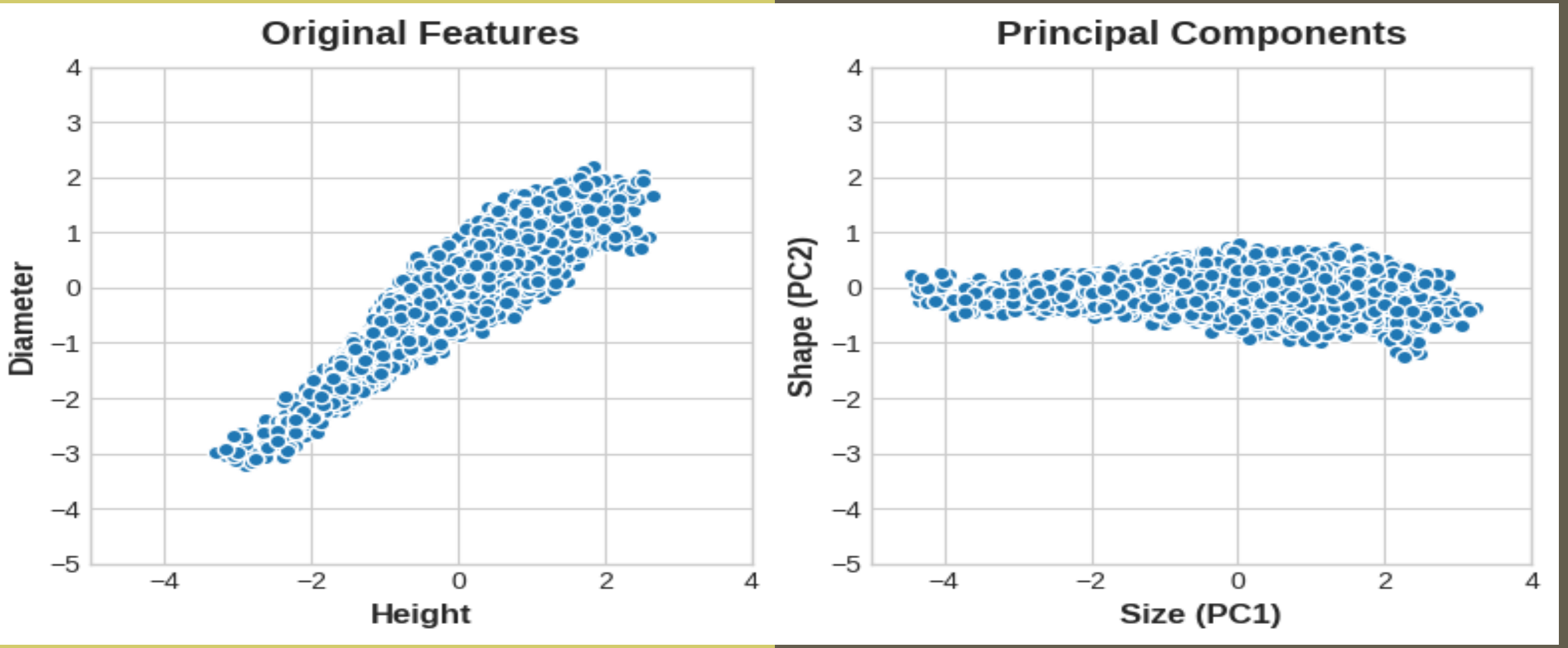
- The "size" column maximizes the variance or the average of the squared distances from the projected points (red dots) to the origin

- The PCA component is directly perpendicular to the first and in this case the "shape" column.

- The second principal component is calculated in the same way, with the condition that it is uncorrelated with (i.e., perpendicular to) the first principal component and that it accounts for the next highest variance.

- This continues until a total of *p* principal components have been calculated, equal to the original number of variables.





THE PRINCIPAL COMPONENTS BECOME THE NEW FEATURES BY A ROTATION OF THE DATASET IN THE FEATURE SPACE.

STEP-BY-STEP CALCULATION

STEP 1: STANDARDIZATION

- It is critical to perform standardization prior to PCA, is that the latter is quite sensitive regarding the variances of the initial variables. That is, if there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges.
- Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{value - mean}{standard\ deviation}$$

STEP 2: COVARIANCE MATRIX COMPUTATION

To understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them. Because sometimes, variables are highly correlated in such a way that they contain redundant information. So, in order to identify these correlations, we compute the covariance matrix.

The covariance matrix is a $p \times p$ symmetric matrix (where p is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables. For example, for a 3-dimensional data set with 3 variables x, y, and z, the covariance matrix is a 3×3 data matrix of this from:

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

STEP 3: COMPUTE THE EIGENVECTORS AND EIGENVALUES OF THE COVARIANCE MATRIX TO IDENTIFY THE PRINCIPAL COMPONENTS

- It is eigenvectors and eigenvalues who are behind all the magic of principal components because the eigenvectors of the Covariance matrix are actually *the directions of the axes where there is the most variance* (most information) and that we call Principal Components. And eigenvalues are simply the coefficients attached to eigenvectors, which give the *amount of variance carried in each Principal Component*.

- By ranking your eigenvectors in order of their eigenvalues, highest to lowest, you get the principal components in order of significance.

STEP 4: CREATE A FEATURE VECTOR

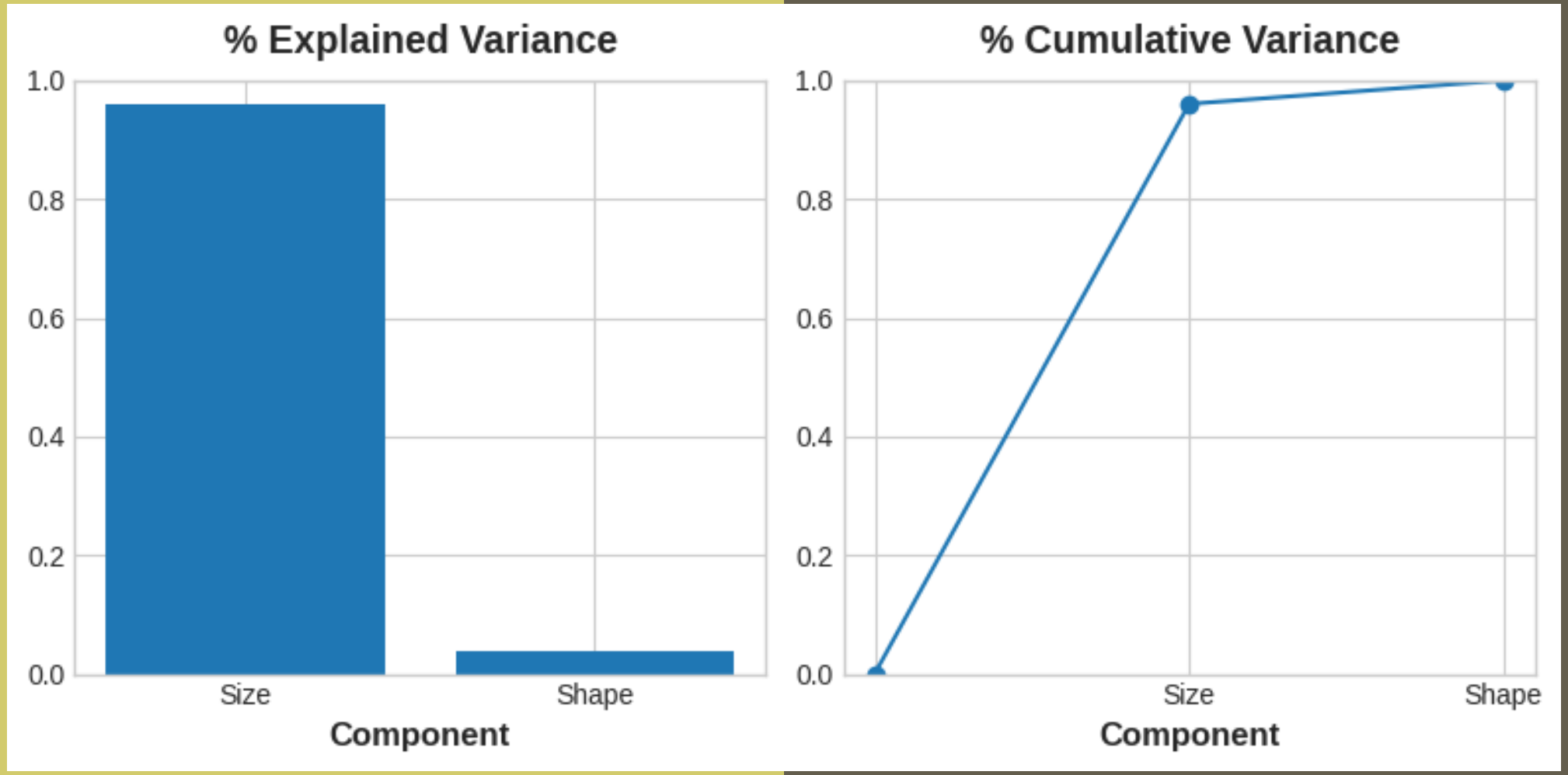
Choose whether to keep all these components or discard those of lesser significance (of low eigenvalues), and form with the remaining ones a matrix of vectors that we call *Feature vector*.

So, the feature vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep. This makes it the first step towards dimensionality reduction, because if we choose to keep only ***p*** eigenvectors (components) out of ***n***, the final data set will have only ***p*** dimensions.

STEP 5: RECAST THE DATA ALONG THE PRINCIPAL COMPONENTS AXES

- The aim is to use the feature vector formed using the eigenvectors of the covariance matrix, to reorient the data from the original axes to the ones represented by the principal components (hence the name Principal Components Analysis). This can be done by multiplying the transpose of the original data set by the transpose of the feature vector.

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$



SIZE ACCOUNTS FOR ABOUT 96% AND THE SHAPE FOR ABOUT 4% OF THE VARIANCE BETWEEN HEIGHT AND DIAMETER.

REFERENCES

- https://www.kaggle.com/code/ryanholbrook/principal-component-analysis
- https://builtin.com/data-science/step-step-explanation-principal-component-analysis

THANK YOU