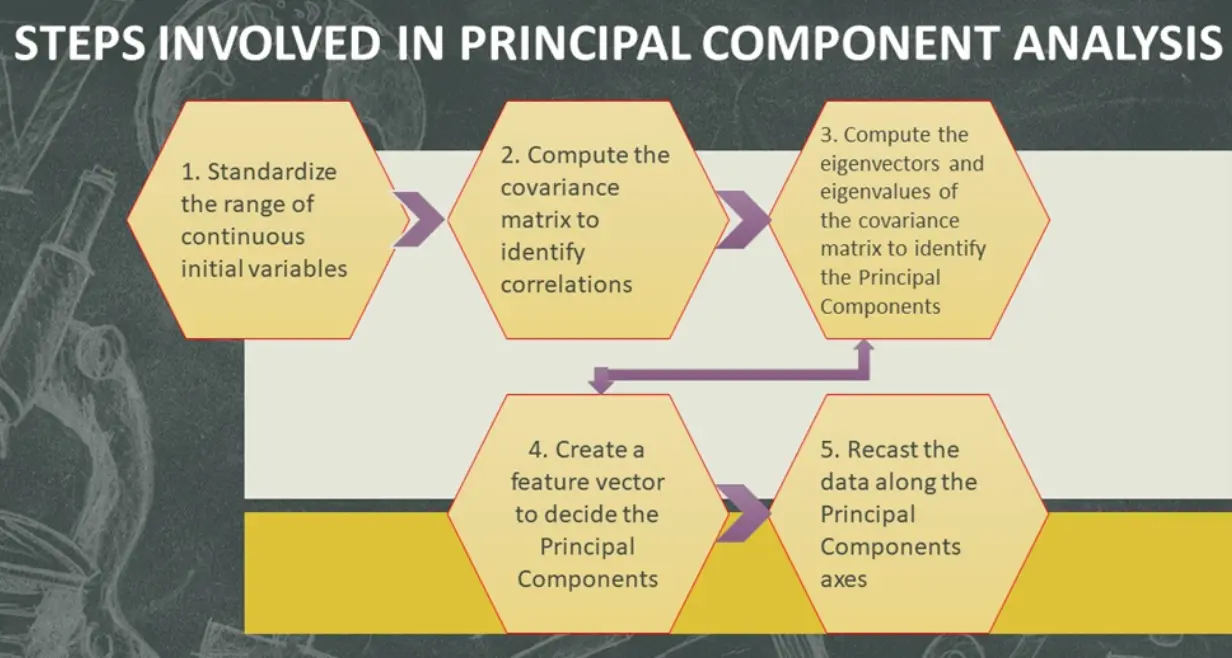
Step by step explanation of Principal Component Analysis

In this section, you will get to know about the steps involved in the Principal Component Analysis technique.



STEP 1: STANDARDIZATION

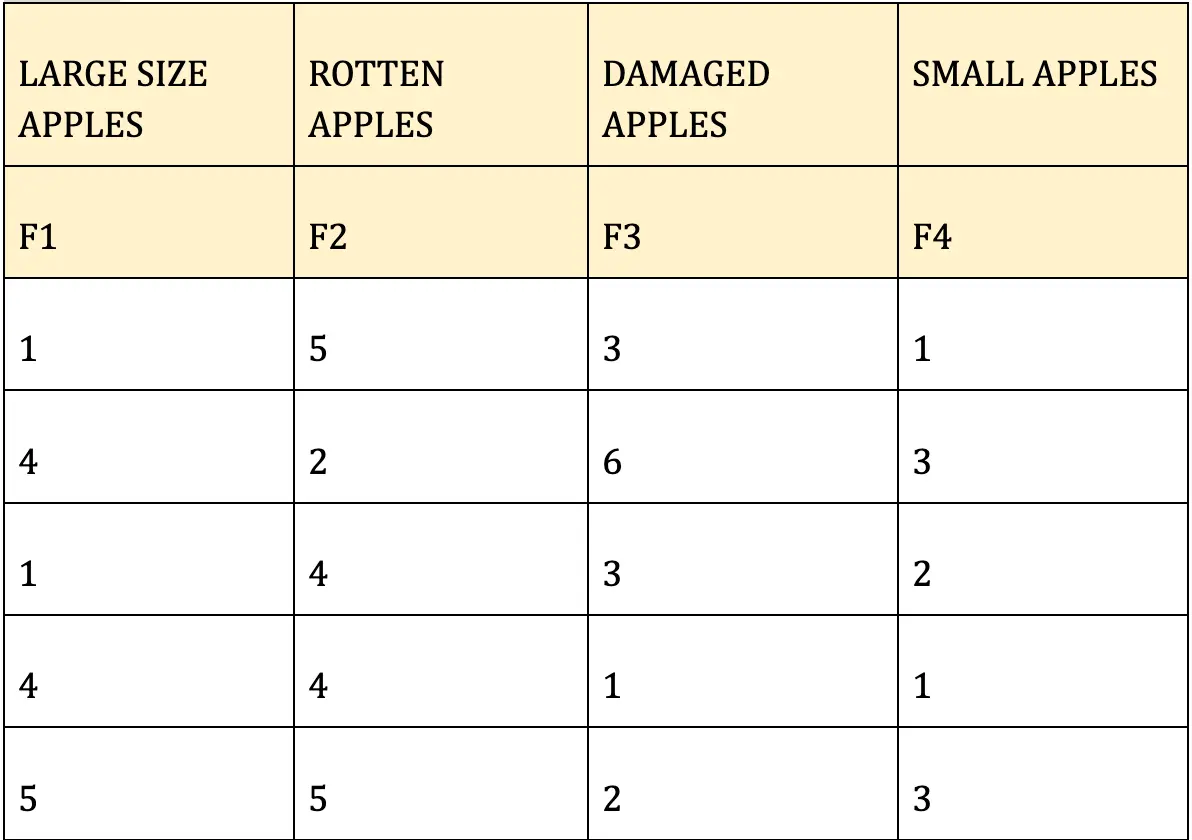
* The range of variables is calculated and standardized in this process to analyze the contribution of each variable equally.
* Calculating the initial variables will help you categorize the variables that are dominating the other variables of small ranges.
* This will help you attain biased results at the end of the analysis.
* To transform the variables of the same standard, you can follow the following formula.

Principal Component Analysis Standardization

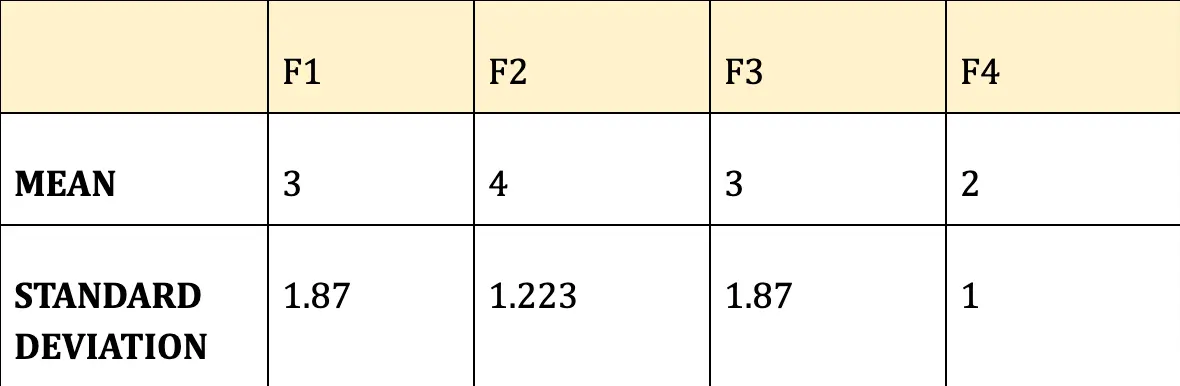
You can refer to the Standard Deviation Formula if you have any doubts about calculating Standard Deviation.

Example:

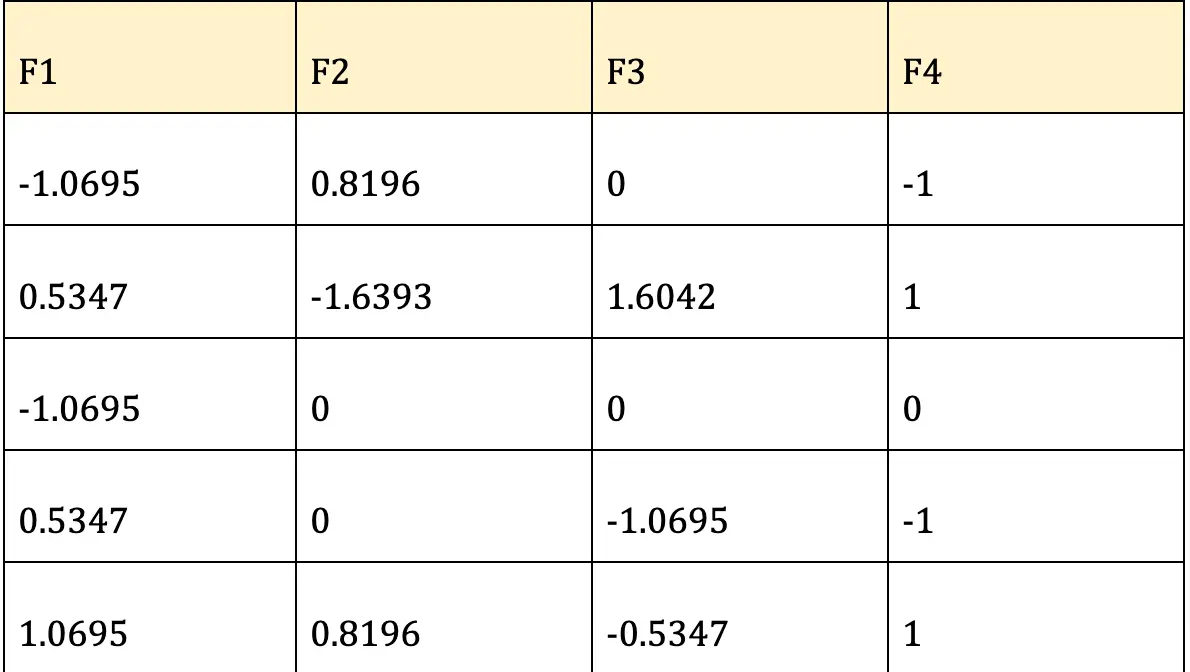
Let us consider the same scenario that we have taken as an example previously. Let us assume the following features of dimensions as F1, F2, F3, and F4.



Calculate the Mean and Standard Deviation for each feature and then, tabulate the same as follows.



Then, after the Standardization of each variable, the results are tabulated below.



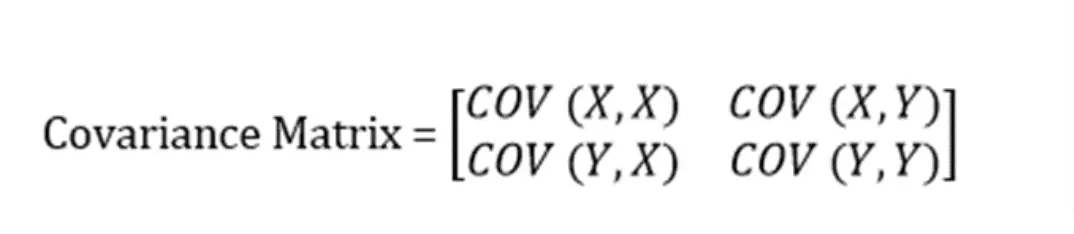
This is the Standardized data set.

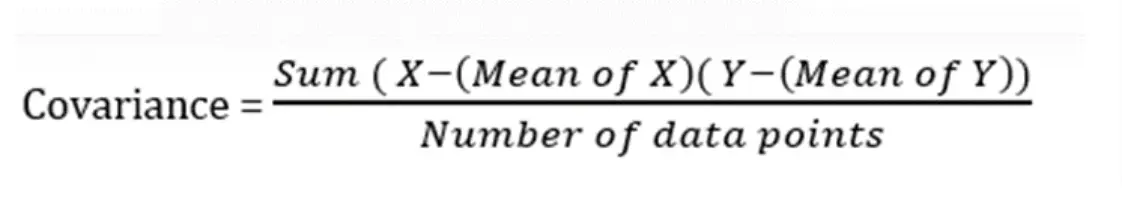
STEP 2: COVARIANCE MATRIX COMPUTATION

* In this step, you will get to know how the variables of the given data are varying with the mean value calculated.
* Any interrelated variables can also be sorted out at the end of this step.
* To segregate the highly interrelated variables, you calculate the covariance matrix with the help of the given formula.

\*\*Note: \*\*A covariance matrix is a N x N symmetrical matrix that contains the covariances of all possible data sets.

The covariance matrix of two-dimensional data is, given as follows:



Where,

4. Make a note that, the covariance of a number with itself is its variance (COV(X, X)=Var(X)), the values at the top left and bottom right will have the variances of the same initial number.

5. Likewise, the entries of the Covariance Matrix at the main diagonal will be symmetric concerning the fact that covariance is commutative (COV(X, Y)=COV(Y, X)).

6A. If the value of the Covariance Matrix is positive, then it indicates that the variables are correlated. ( If X increases, Y also increases and vice versa)

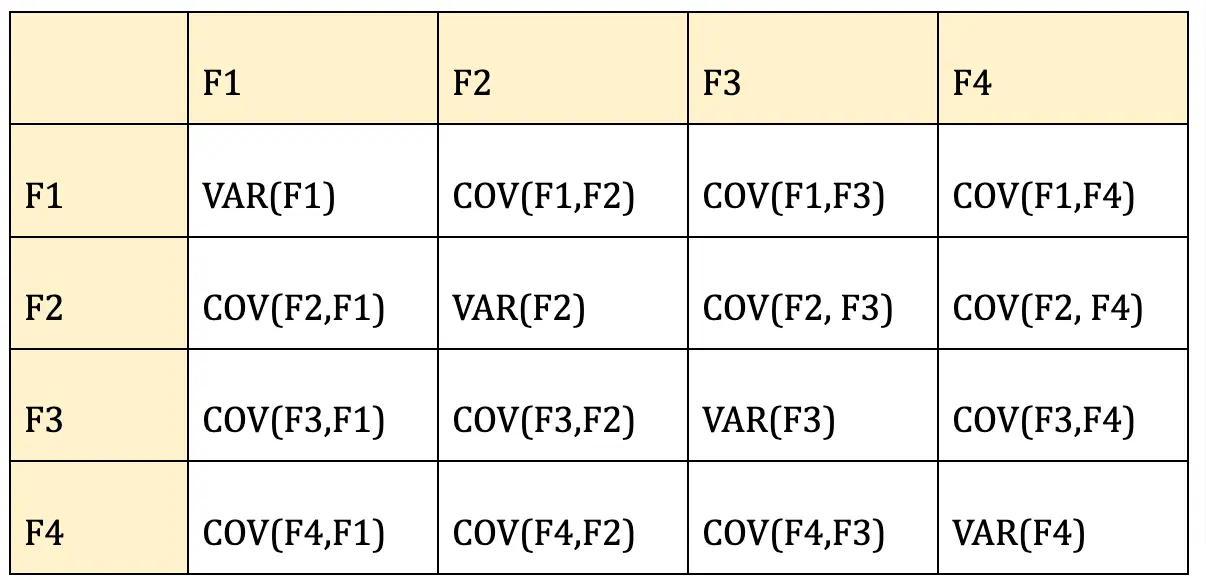
6B. If the value of the Covariance Matrix is negative, then it indicates that the variables are inversely correlated. ( If X increases, Y also decreases and vice versa).

7. As a result, at the end of this step, you will come to know which pair of variables are correlated with each other, so that you might categorize them much easier.

Example:

So, continuing with the same example,

The formula to calculate the covariance matrix of the given example will be:



Since you have already standardized the features, you can consider Mean = 0 and Standard Deviation=1 for each feature.

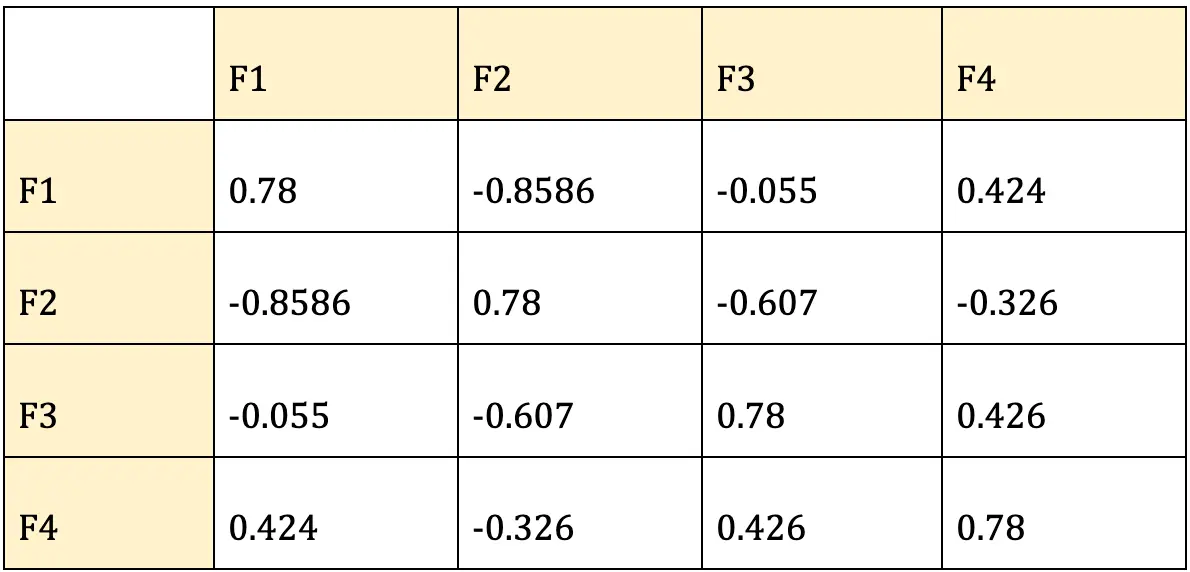
VAR(F1) = ((-1.0695-0)² + (0.5347-0)² + (-1.0695-0)² + (0.5347–0)² +(1.069–0)²)/5

On solving the equation, you get, VAR(F1) = 0.78

COV(F1,F2) = ((-1.0695–0)*(0.8196-0) + (0.5347–0)*(-1.6393-0) + (-1.0695–0)\* (0.0000-0) + (0.5347–0)*(0.0000-0)+ (1.0695–0)*(0.8196–0))/5

On solving the equation, you get, COV(F1,F2 = -0.8586)

Similarly solving all the features, the covariance matrix will be,



STEP 4: FEATURE VECTOR

1. To determine the principal components of variables, you have to define eigen value and eigen vectors for the same. Let A be any square matrix. A non-zero vector v is an eigenvector of A if

Av = λv

for some number λ, called the corresponding eigenvalue.

2. Once you have computed the eigen vector components, define eigen values in descending order ( for all variables) and now you will get a list of principal components.

3. So, the eigen values represent the principal components and these components represent the direction of data.

4. This indicates that if the line contains large variables of large variances, then there are many data points on the line. Thus, there is more information on the line too.

5. Finally, these principal components form a line of new axes for easier evaluation of data and also the differences between the observations can also be easily monitored.

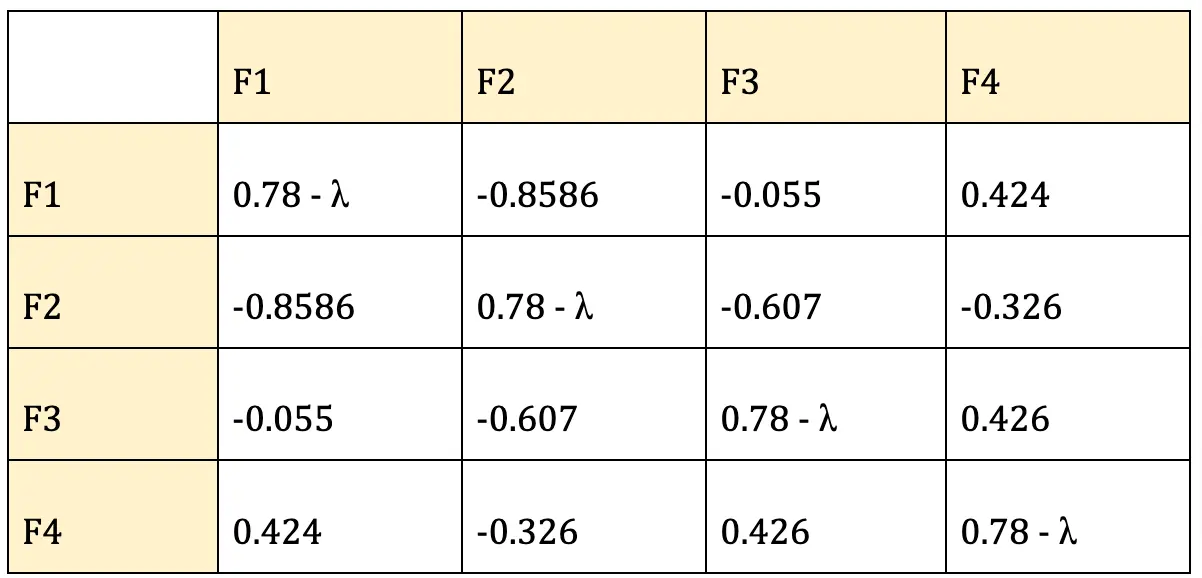
Example:

Let ν be a non-zero vector and λ a scalar.

As per the rule,

Aν = λν, then λ is called eigenvalue associated with eigenvector ν of A.

Upon substituting the values in det(A- λI) = 0, you will get the following matrix.



When you solve the following the matrix by considering 0 on right-hand side, you can define eigen values as

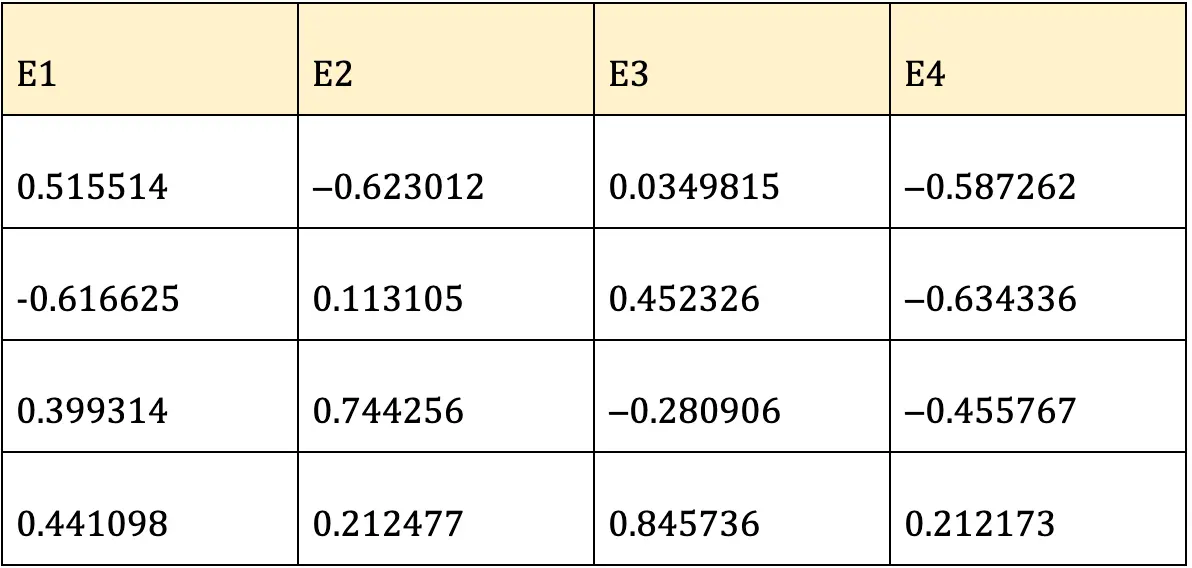
λ = 2.11691 , 0.855413 , 0.481689 , 0.334007

Then, substitute each eigen value in (A-λI)ν=0 equation and solve the same for different eigen vectors v1, v2, v3 and v4.

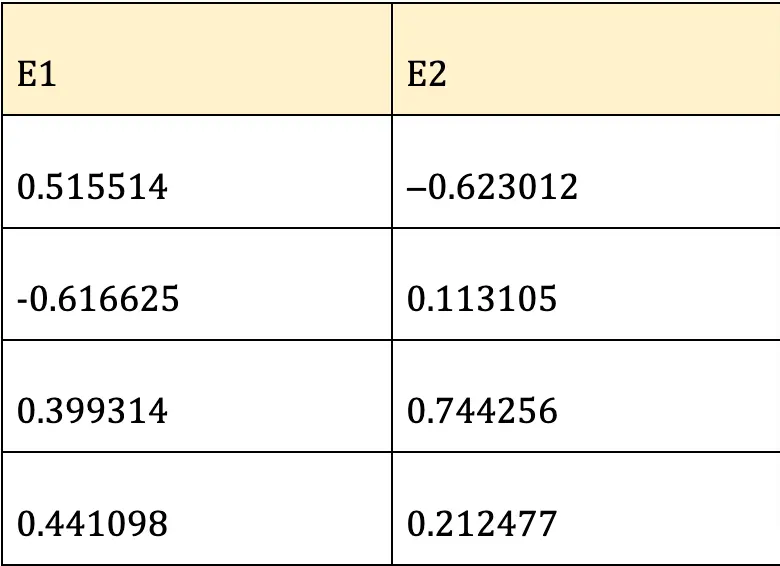
For instance,

For λ = 2.11691, solving the above equation using Cramer's rule, the values for the v vector are v1 = 0.515514 v2 = -0.616625 v3 = 0.399314 v4 = 0.441098

Follow the same process and you will form the following matrix by using the eigen vectors calculated as instructed.



Now, calculate the sum of each Eigen column, arrange them in descending order and pick up the topmost Eigen values. These are your Principal components.



STEP 5: RECAST THE DATA ALONG THE PRINCIPAL COMPONENTS AXES

* Still now, apart from standardization, you haven’t made any changes to the original data. You have just selected the Principal components and formed a feature vector. Yet, the initial data remains the same on their original axes.
* This step aims at the reorientation of data from their original axes to the ones you have calculated from the Principal components.

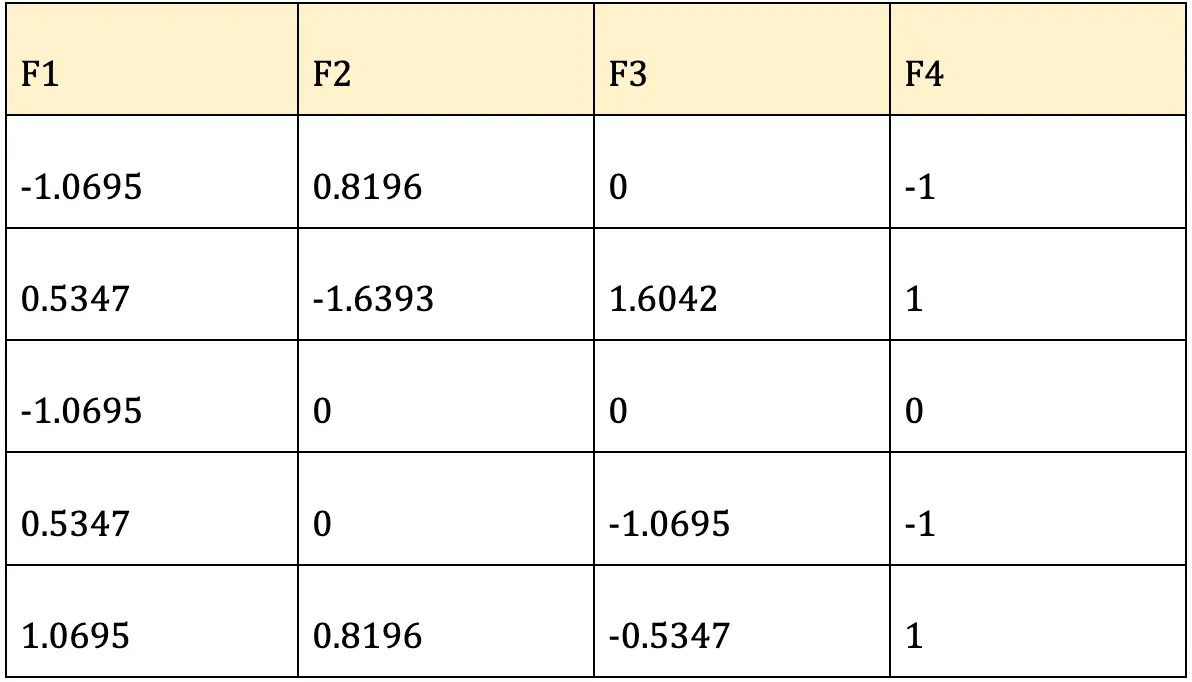
This can be done by the following formula.

Final Data Set= Standardized Original Data Set \* FeatureVector

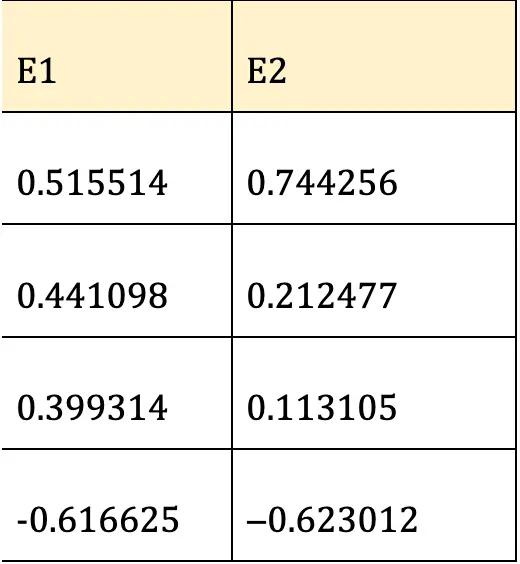
Example:

So, in our guide, the final data set becomes

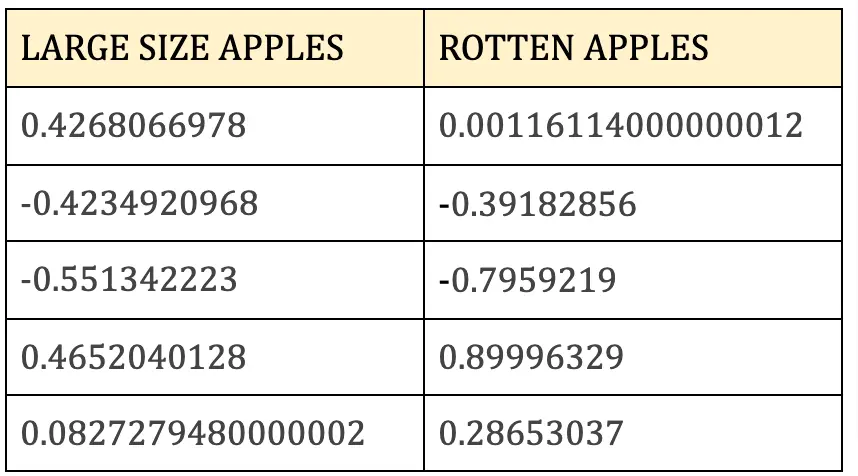
Standardized Original Data Set =



FeatureVector =



By solving the above equations, you will get the transformed data as follows.



Did you notice something? Your large dataset is now compressed into a small dataset without any loss of data! This is the significance of Principal Component Analysis.

Applications of PCA Analysis

* PCA in machine learning is used to visualize multidimensional data.
* In healthcare data to explore the factors that are assumed to be very important in increasing the risk of any chronic disease.
* PCA helps to resize an image.
* PCA is used to analyze stock data and forecasting data.
* You can also use Principal Component Analysis to analyze patterns when you are dealing with high-dimensional data sets.

Advantages of Principal Component Analysis

* Easy to calculate and compute.
* Speeds up machine learning computing processes and algorithms.
* Prevents predictive algorithms from data overfitting issues.
* Increases performance of ML algorithms by eliminating unnecessary correlated variables.
* Principal Component Analysis results in high variance and increases visualization.
* Helps reduce noise that cannot be ignored automatically.

Disadvantages of Principal Component Analysis

* Sometimes, PCA is difficult to interpret. In rare cases, you may feel difficult to identify the most important features even after computing the principal components.
* You may face some difficulties in calculating the covariances and covariance matrices.
* Sometimes, the computed principal components can be more difficult to read rather than the original set of components.