At first using function <<reading_data_set>> which its' input is the path to the file we read the data

From the csv files and create a list of numeric data.

Some had non-numeric variables to handle wrote a code to change it to number respectfully.

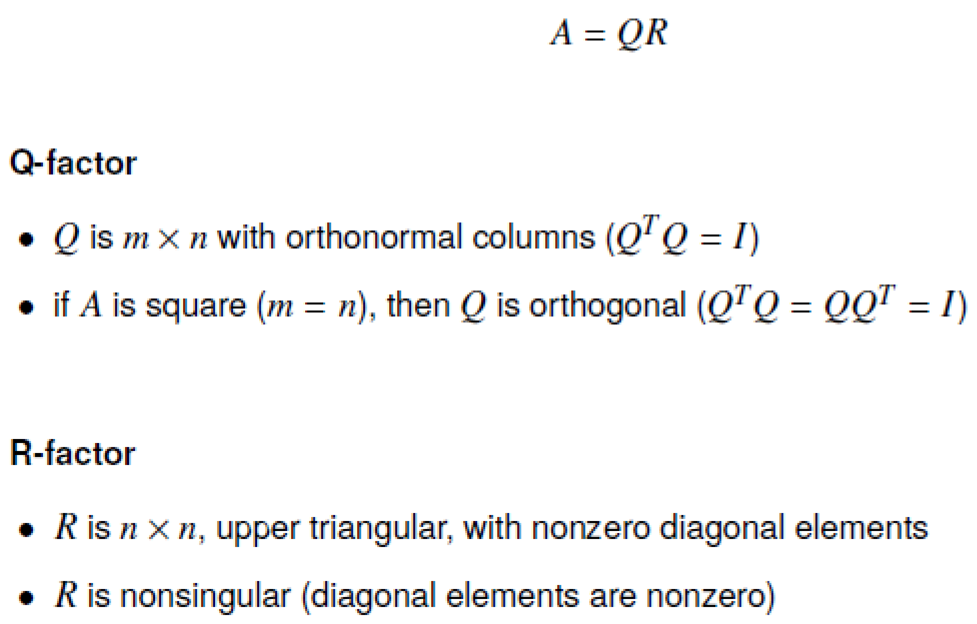
After reading data I split them to two set of target and features.

Now that I had the data ready completely, I decided to use some of the existing PCA such as scikit to have a baseline to compare my work with.

Then At this time I know the data I know the variance of data I start to write my PCA class in which I at first compute centered data (subtracting data from its average) in another word I normalized my data

I used this approach because not only it normal data but also it helps to simplify the math behind the PCA for example in covariance formula you are going to divide by two variances but since they are both equal to you don't need to compute them also since you changed the average of data to 0 you can cancel out nominator to a dot product.

After computing variance I need eigenvalues and eigenvectors of variance matrix and I used QR decomposition to do so.



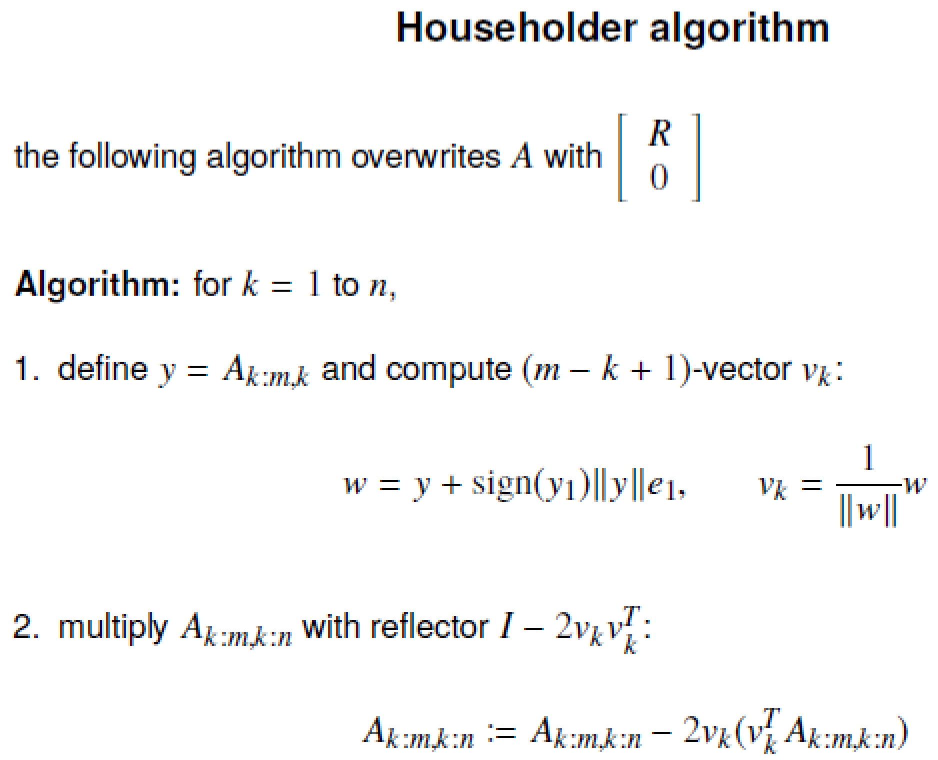
$$A = QR$$

Q-factor

- Q is $m \times n$ with orthonormal columns ($Q^T Q = I$)
- if A is square ($m = n$), then Q is orthogonal ($Q^T Q = QQ^T = I$)

R-factor

- R is $n \times n$, upper triangular, with nonzero diagonal elements
- R is nonsingular (diagonal elements are nonzero)



Householder algorithm

the following algorithm overwrites A with $\begin{bmatrix} R \\ 0 \end{bmatrix}$

Algorithm: for $k = 1$ to n,

1. define $y = A_{k:m,k}$ and compute $(m - k + 1)$-vector v_k:

$$w = y + \operatorname{sign}(y_1)\|y\|e_1, \qquad v_k = \frac{1}{\|w\|}w$$

2. multiply $A_{k:m,k:n}$ with reflector $I - 2v_k v_k^T$:

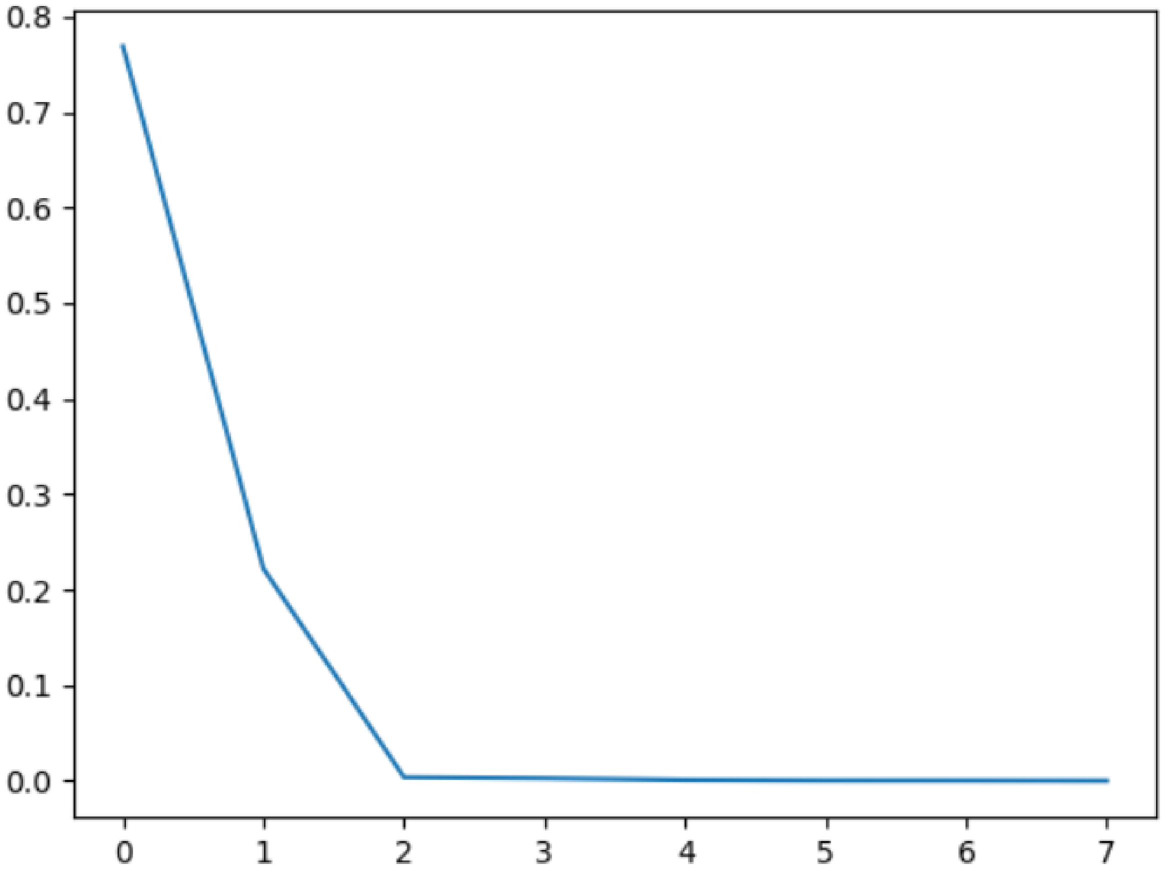
$$A_{k:m,k:n} := A_{k:m,k:n} - 2v_k(v_k^T A_{k:m,k:n})$$

After having Eigen attributes I sorted them in decreasing order and I asked to give how many reduction we are going to do.

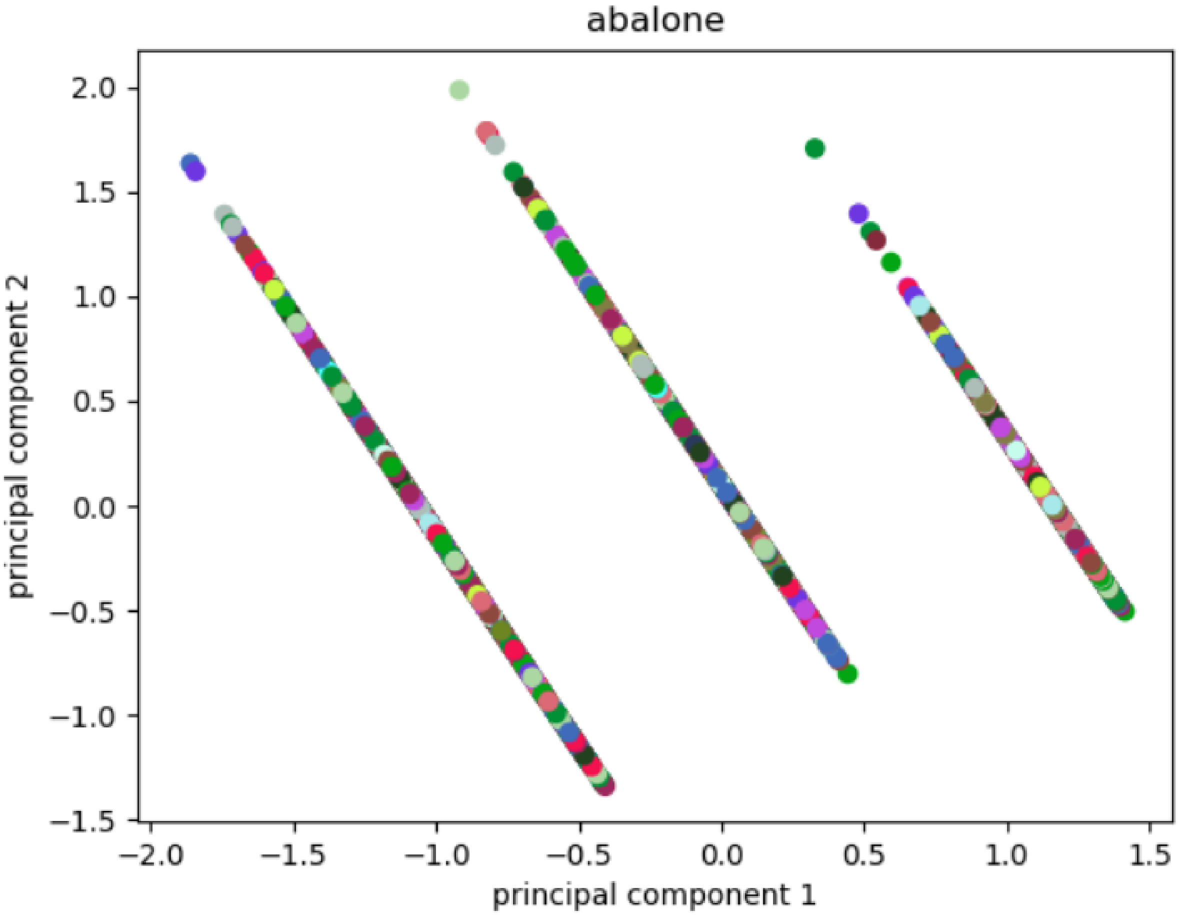
For computing the reduced matrix I computed dot product of available data and eigenvectors.

Now that I have my decreased data I need to visualized data. We are told to decrease to dimension to 2 so we could visualize the data easily. But at first I did the reduction the reduced data didn’t satisfy me so I added 3D visualization too.

For the first data set which is abalone:

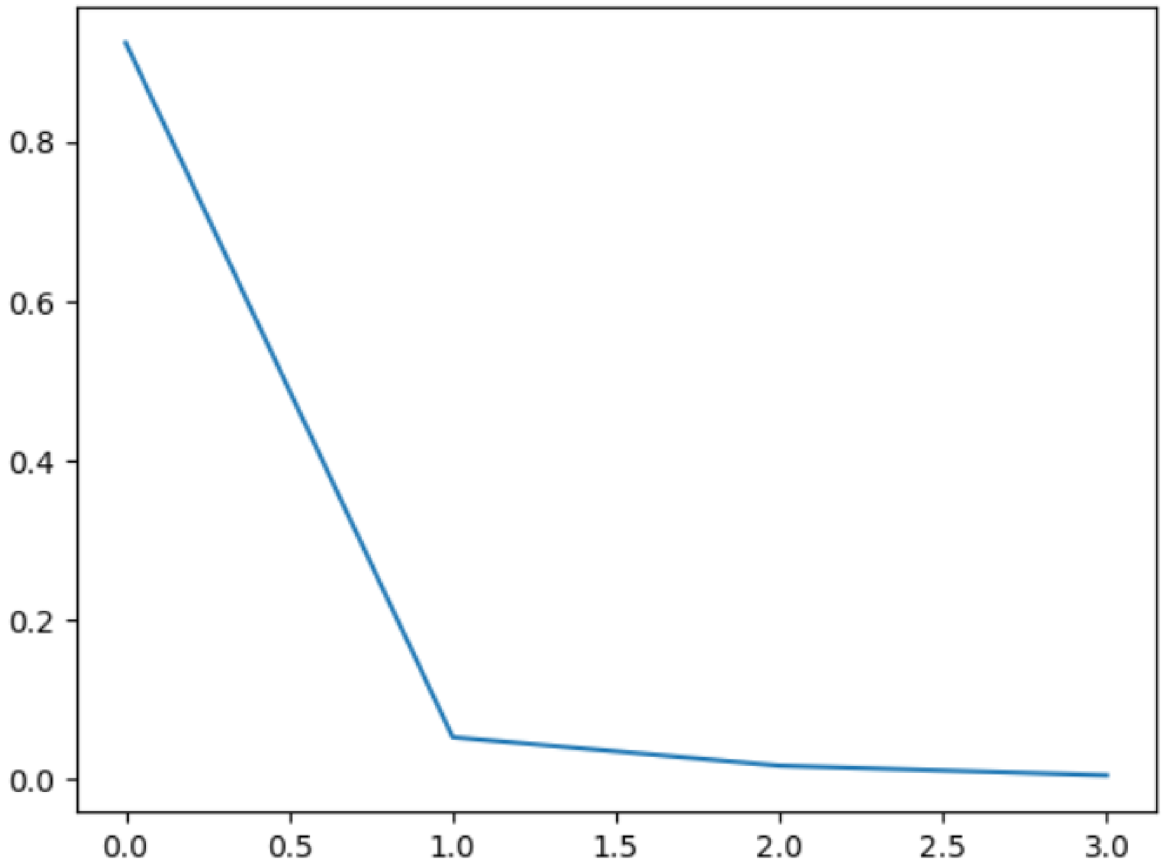


Is the figurine of how many percentage of eigenvalue is decided eigenvalue and as we can suspect the first two PC should give us pretty good reduction

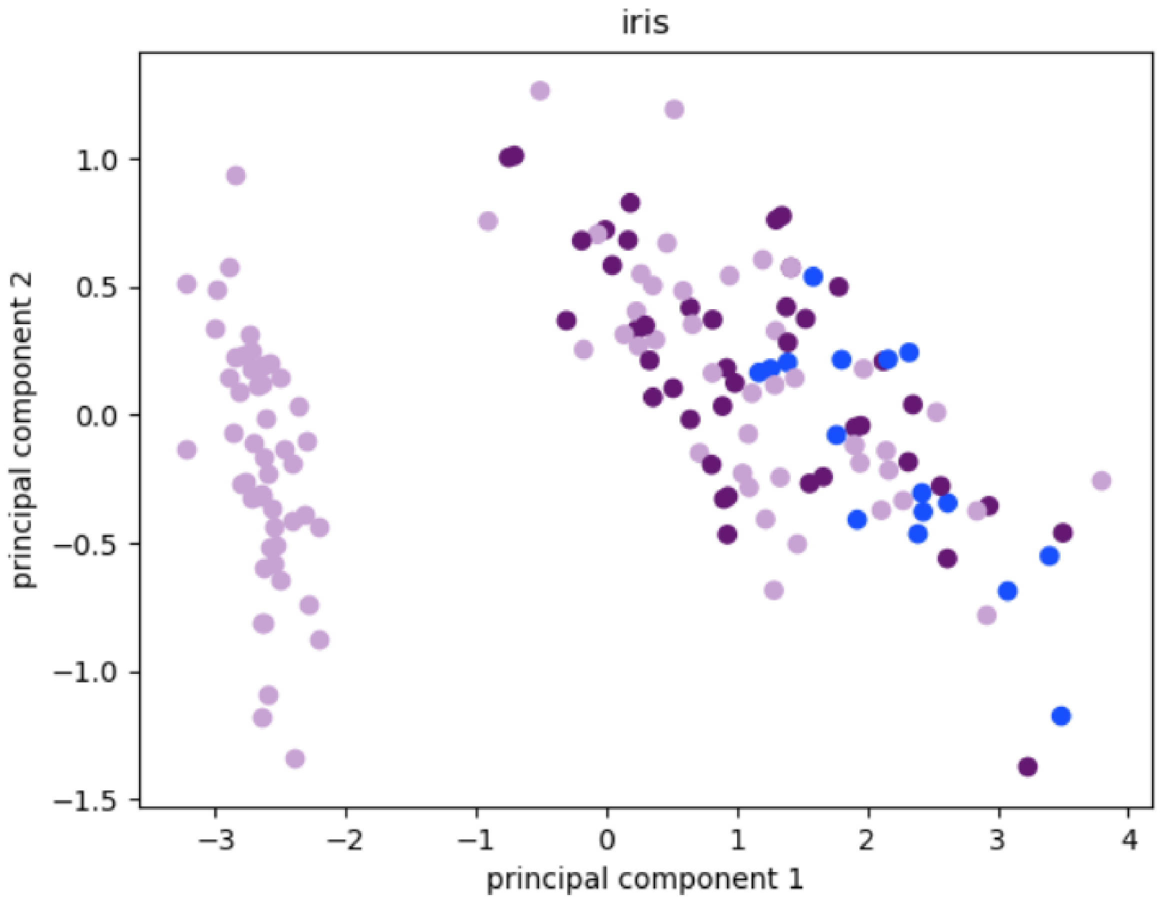


which we can see it seperates data but not as nicely as we want since the color are not separated.

For next data set which is iris:

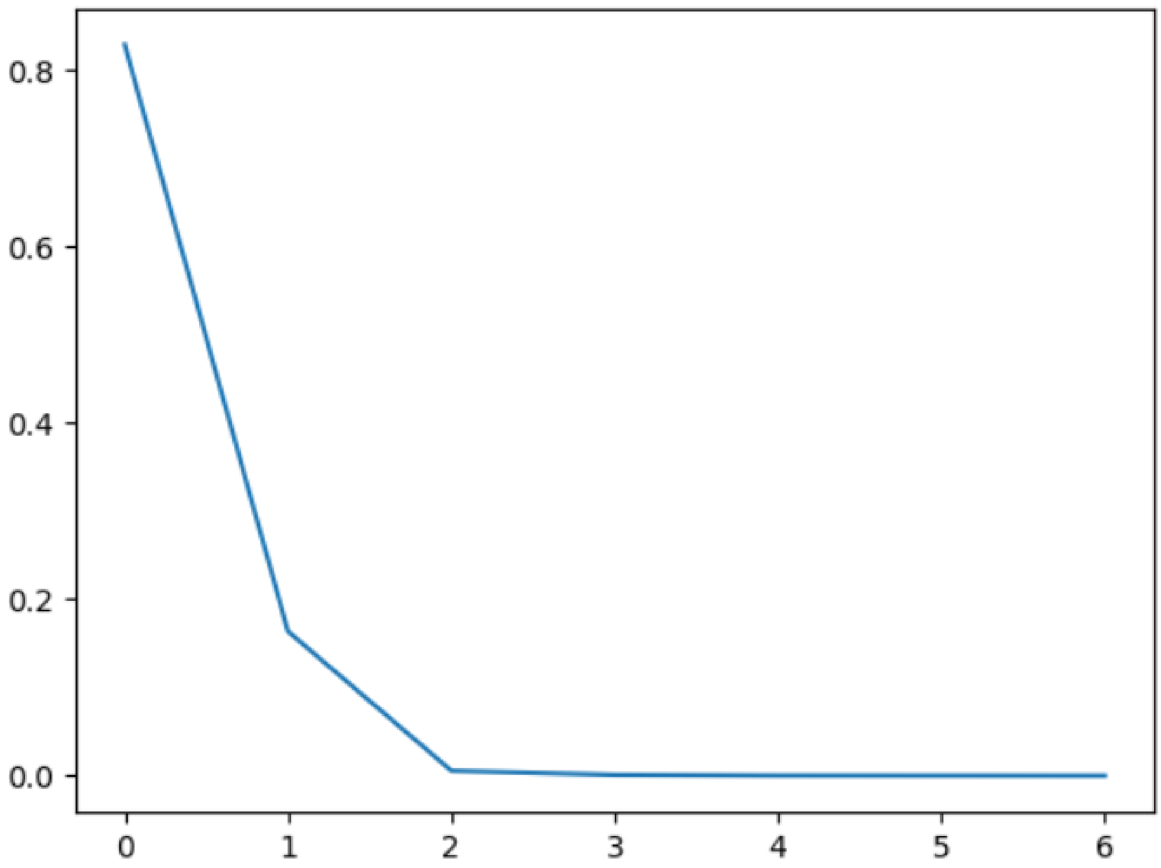


Can make us assume that two principle components should be pretty good separator:

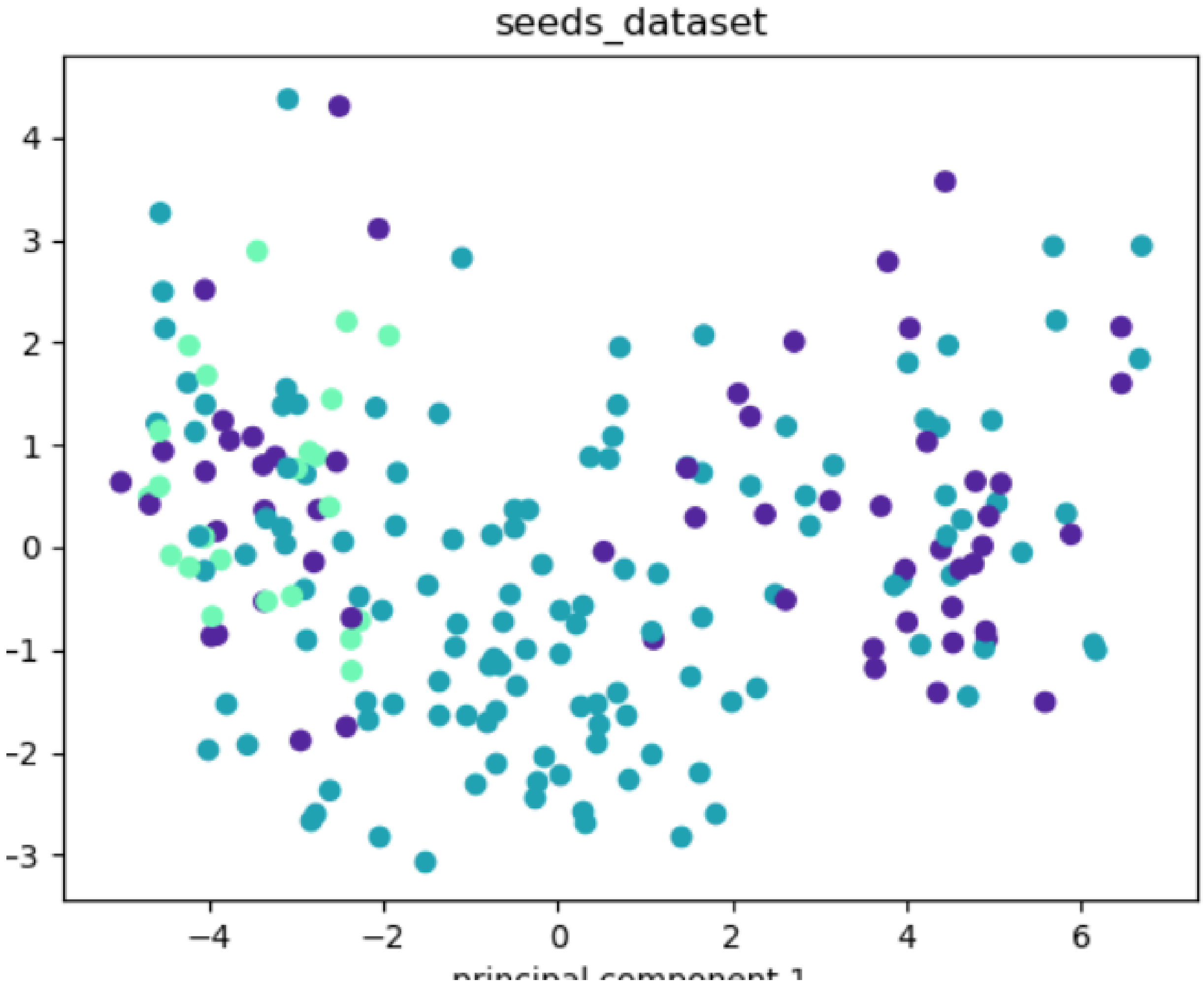


Which it separates one of the data completely so it can use PCA.

And for the last data set which is seeds:

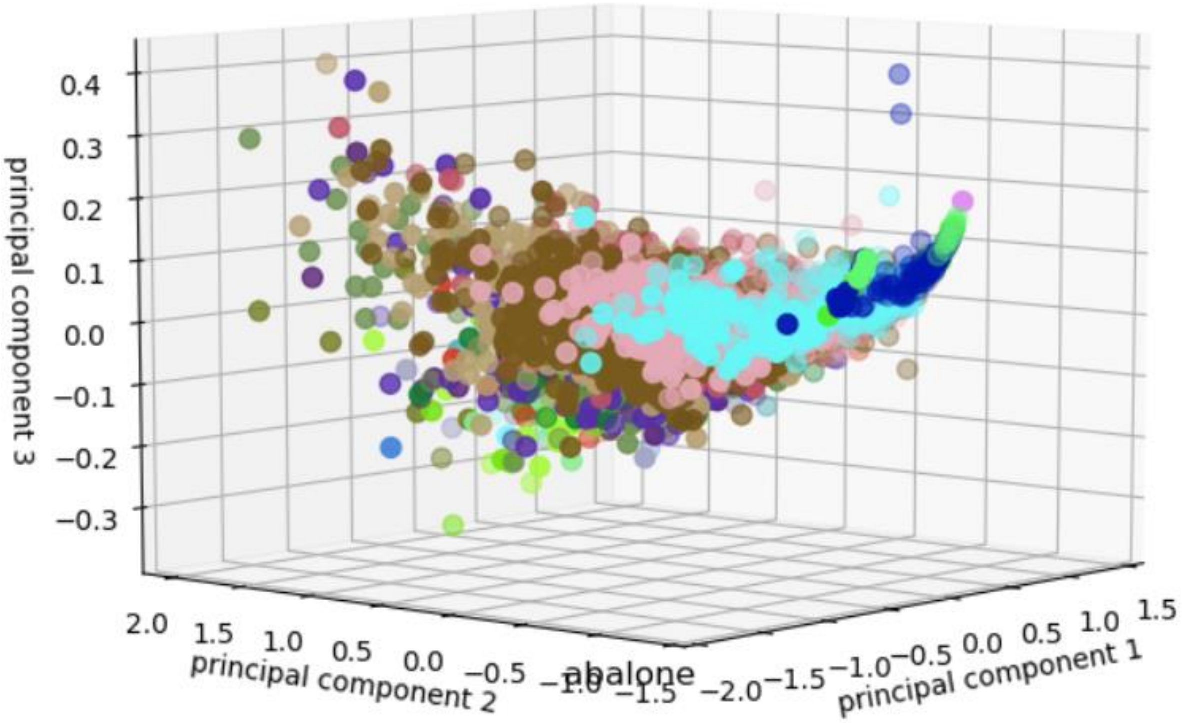


So we again suspect that we will get a pretty good result using two component which it does almost separate two of the data which is desirable



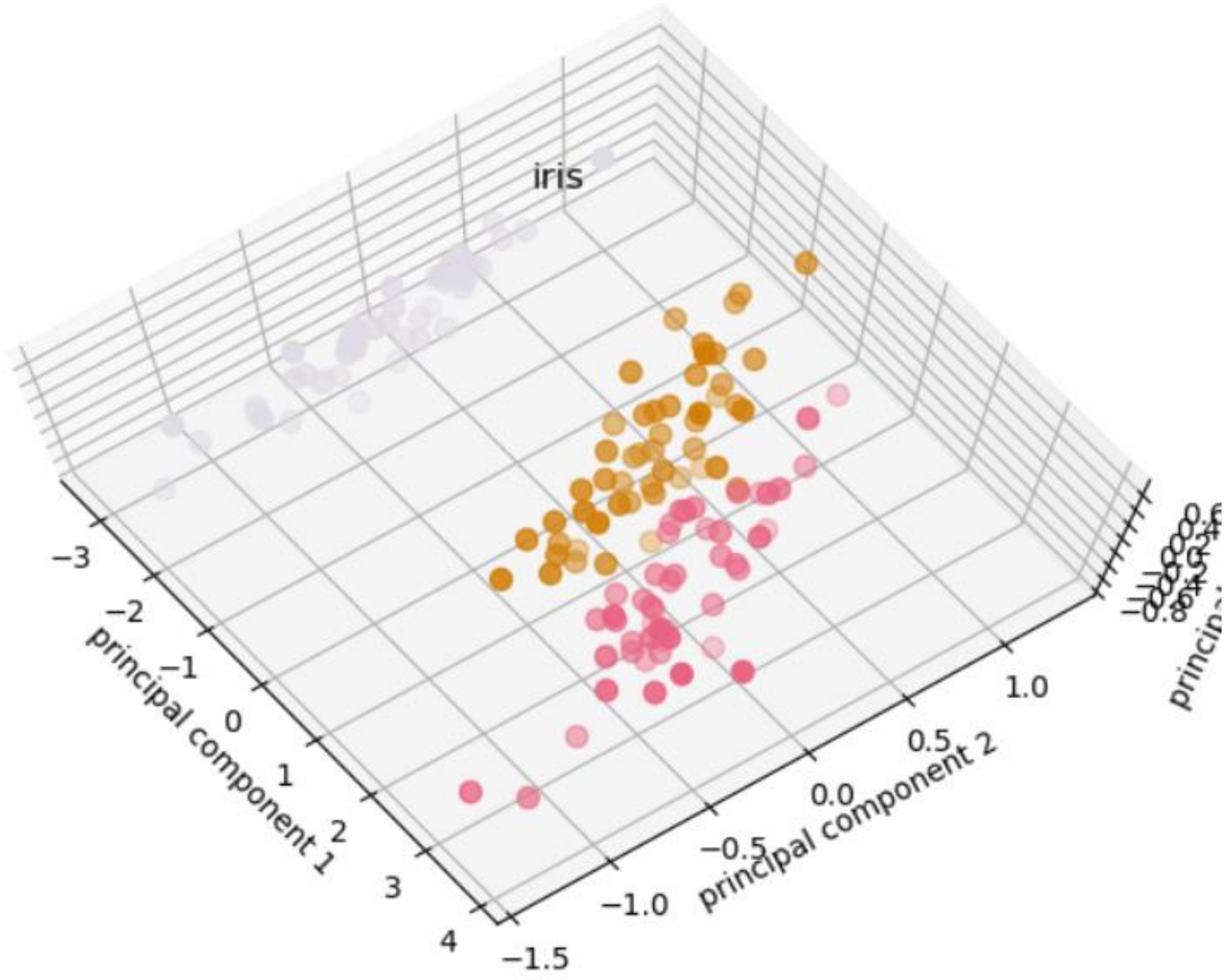
And now I use 3 as the number of component:

Abalone:



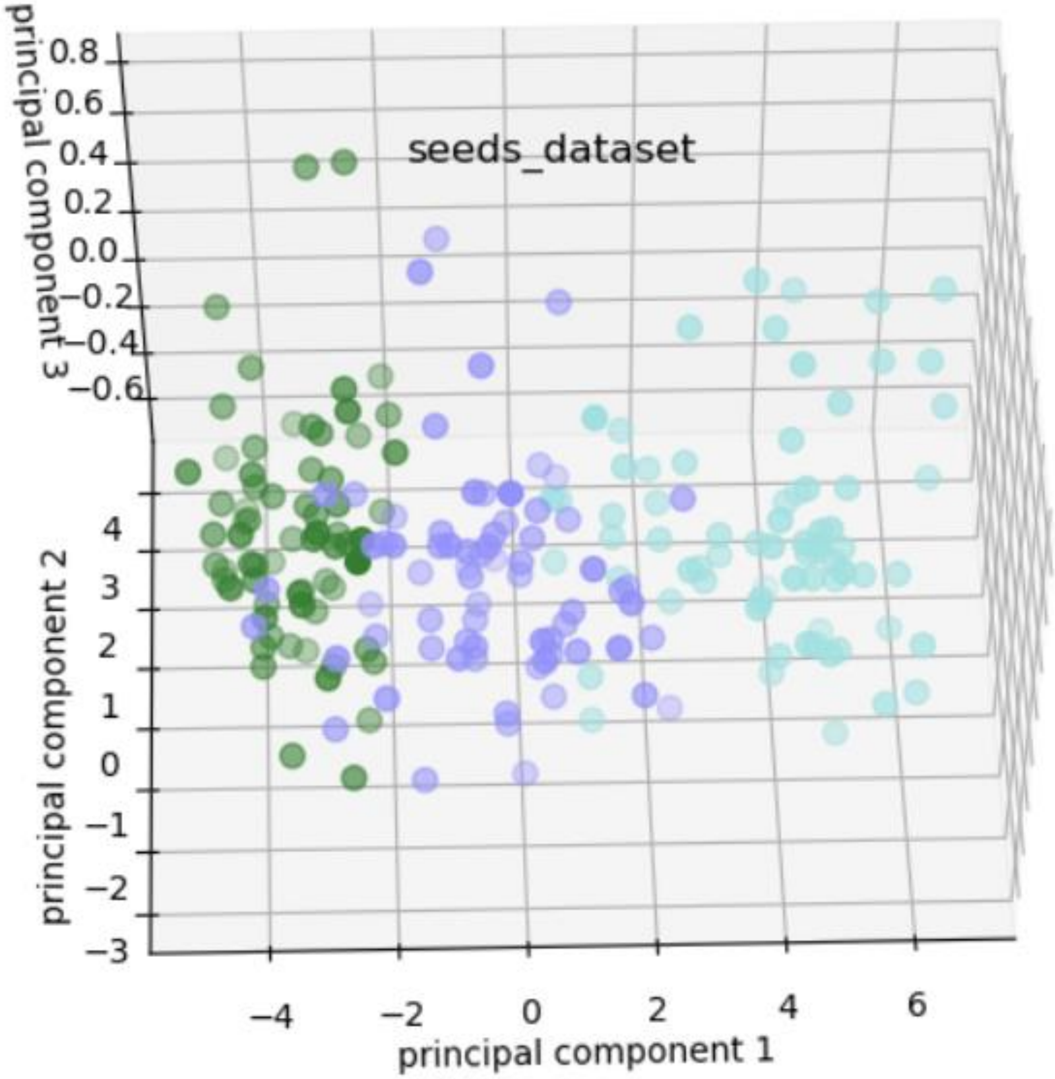
Which we clearly see some separation point

Iris:



Which they are fully separated.

Seeds:



Which again we can see the line clearly.