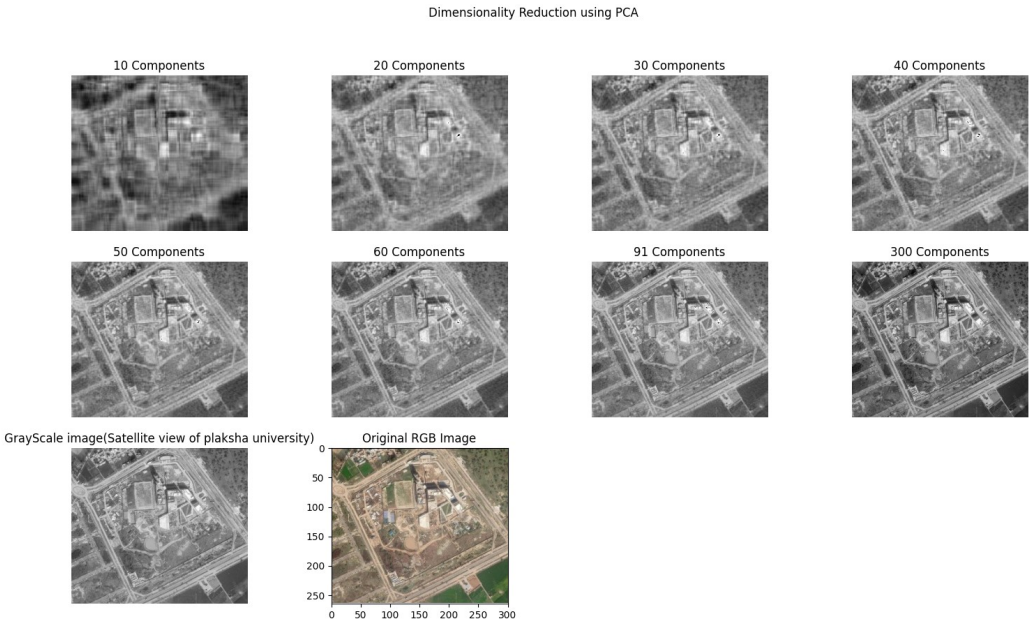
USMAN AKINYEMI
PLAKSHA UNIVERSITY
U20220090
LAB3
s



1. What is the main idea behind PCA in terms of transforming data?

PCA is a technique which we can use for dimensionality reduction. The main idea used in the technique is to linearly transformed the data unto a new coordinate system such that the direction which is known as the principal components capture the largest variation in the data. The eigenvalue of the covariance of the data(matrix) represent the extent of variation it's corresponding eigenvectors(principal component) represent in the data. The higher the value of the eigenvalue, the higher the variability it's eigenvector represent. PCA project the data from higher dimensionality to lower sub-space. The first principal components will account for the highest variability in the dataset. The second one will be orthogonal to the first one and has the second highest variability and so on.

2. Describe the concept of variance in the context of PCA

From what I understand as the meaning of variance in the context of PCA is that, it is the parameter which actually shows how a components(features) contribute to the important information in the dataset. We can simply say variance is the parameter which shows the important information in a dataset. We can also say that variance quantifies the amount of information or variability contained along a particular direction (principal components).

3. What is the significance of eigenvalues and eigenvectors in PCA?
They both are significance in PCA, eigenvectors represent the directions of the principal component axes while the eigenvalues is used to determine the importance of each principal component in

capturing the variance of the original data along the direction of each respective eigenvectors. In a simple way the eigenvalue of the covariance of the data(matrix) represent the extent of variation it's corresponding eigenvectors(the principal component) represent in the data. The higher the value of the eigenvalue the higher the variability it's principal component(eigenvectors) represents.

4. How do you choose the number of principal components to retain?
It is quite simple, what someone can do is to plot the cumulative explained variance against the number of principal component and choose the number of component that captures a significant portion of the total variance. The threshold variance which we used in this assignment is 95% and the first 91 component represent that out of 300 component.

5. Can PCA be sensitive to outliers in the dataset? Why or why not? PCA is sensitive to outliers as it can affect the calculations of the covariance matrix which is used to determine the principal components. Process like standardization can help to remove the outliers and minimize their impact.