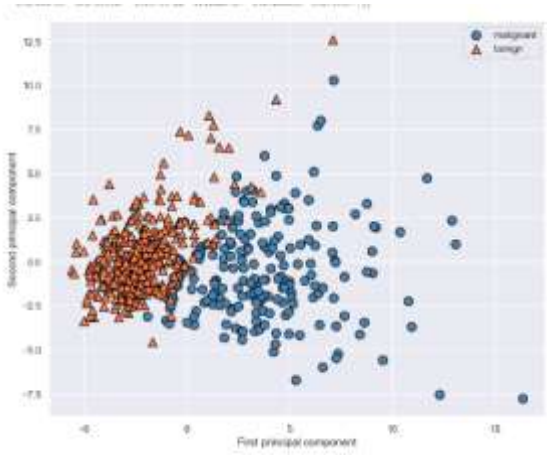
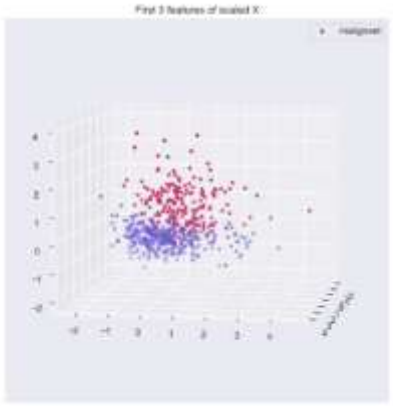
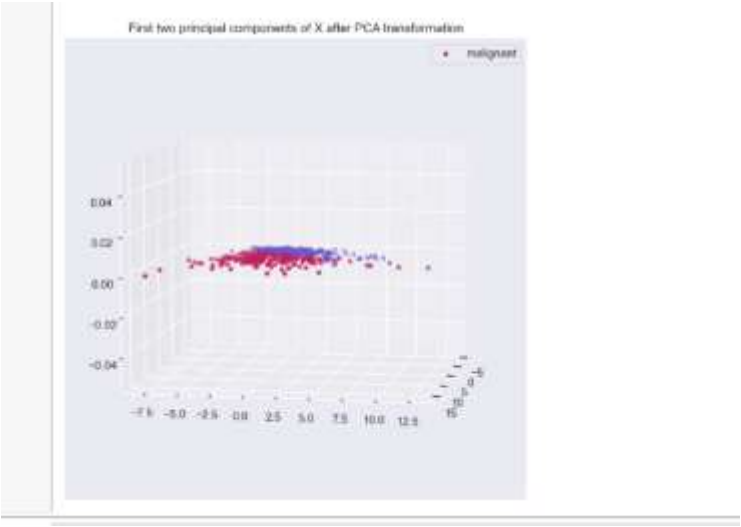
8.1P - PCA dimensionality reduction

PCA (Principle Component Analysis) is a dimensionality reduction technique that projects the data into a lower dimensional space. It can be used to reduce high dimensional data into 2 or 3 dimensions so that we can visualize and hopefully understand the data better.

In this task, you use PCA to reduce the dimensionality of a given dataset and visualize the data.







This code uses the Scikit-learn library's Breast Cancer dataset to conduct Principal Component Analysis (PCA) and displays the findings in both 2D and 3D graphs.

First, the load_breast_cancer function is used to load the dataset. This method returns information about cases of breast cancer, including characteristics that can be used to identify a tumour as benign or malignant. Next, to guarantee that every feature contributes equally to the analysis—a crucial component of PCA—the data is standardised using Scikit-learn's StandardScaler.

In order to visualise the data in two dimensions while retaining as much variation as feasible, PCA is used to reduce the dataset's dimensionality from its initial 30 features down to 2 primary components. To determine the success of the reduction, the code prints the principle components, the PCA component matrix, and the shapes of the original and reduced datasets.

To visualise these two main components, a 2D scatter plot is made, with various colours denoting benign and malignant cases. Benign instances are displayed with coral-colored triangular markers, whereas malignant cases are represented with blue circular markers. Based on the first two main components, which account for the majority of the variance in the data, this figure facilitates the observation of the differences between the two groups of tumours.

Subsequently, the algorithm generates two 3D scatter plots and specifies a custom colour map for visualisation. A more thorough understanding of the distribution of the data in three dimensions is made possible by the first 3D graphic, which displays the first three aspects of the standardised dataset. The data is once more visualised using the two principle components that were obtained using PCA in the second 3D graphic. By changing the viewing angle and viewpoint, these 3D visualisations aim to make the structure and separation of data points easier to interpret.

Overall, this code shows how PCA can effectively analyse and interpret a complicated dataset by reducing its dimensionality while preserving important information. It does this by combining 2D and 3D visualisations.