Principal Component Analysis

Why

Used to reduce the dimensionality of a dataset by identifying the main sources of variation within the data set e.g get rid of collinear variables and variables which do not affect the value of other variables.

How

Centre your data by removing mean of each column

Find covariance matrix

Find eigenvalues

Find eigenvectors

Eigenvectors represent the main sources of variation e.g point in direction of main source of variation and give good description of what variables impact each other the most

Eigenvalues gives the variance each component represents in the data

Can multiply original data by matrix which contains Principal Components to the components as the axis. This is called a score plot.

Think of components as weighted combinations variables.

Can use this weighted combination instead of the actual variables to reduce dimensions of the data e.g. intoruce new variable which is easier to work with and has maximal variance