Principal Component Analysis

Principal Component analysis

- It **reduces** the **number of dimensions in large datasets** to **principal components** that **retain** most of the **original information**.
- It does this by **transforming** potentially correlated variables into a **smaller set of variables**, called **principal components**.

Principal Component analysis

- Karl Pearson is credited with the development of PCA in 1901.
- Is very effective for visualizing and exploring high-dimensional datasets, or data with many features.
- Commonly used for data preprocessing for use with machine learning algorithms.
- PCA also minimizes, or altogether eliminates, common issues such as ***multicollinearity*** and ***overfitting***.

PCA VS K-MEANS CLUSTERING

- PCA and k-means clustering are both **unsupervised** machine learning techniques used for data analysis, but they have different goals and methods.

- PCA is used to **reduce the dimensionality of the data**, while k-means clustering groups data points together based on similarity.

PCA vs K-MEANS CLUSTERING

- PCA creates new variables, i.e. principal components, that are linear combinations of the original variables. PCA takes a dataset with multiple variables as input, and it produces a dataset into a lower subspace, that is, a **reduced dataset with fewer variables.**
- K-means is a clustering algorithm that assigns data points to clusters based on their distance from the cluster centers.

HOW PCA WORKS?

Here are the steps on how to calculate PCA:

1. Calculate the MEAN and STANDARD DEVIATION. (**Original Data**)
2. **STANDARDIZE** the data.
3. Compute **COVARIANCE MATRIX** to identify correlations.
4. Compute the **Eigenvectors** and **Eigenvalues** of the COVARIANCE MATRIX.
5. Select the **Principal Components**
6. **Transform** data into the new Coordinate system

SIMULATION:

REFER TO EXCEL FILE.