DATE __/__/____
PAGE NO _____

Assignment - 09 mycompanion 43304

* Aim → Principal Component Analysis
Finding principal components, variance and S.D calculations of principal components (using R)

* Theory →

- Principal Component Analysis :

It is a method of extracting important variables (in form of components) from a large set of variables available in a dataset. It extracts low dimensional set of features from a high dimensional dataset with a nature to capture as much information as possible. It is useful for dealing with 3 or higher dimensional data.

It is always performed on a symmetric correlation or covariance matrix. This means that the matrix should be numeric and have standardized data.

The first principal component gives the direction of the maximum spread of data. The second gives the direction of the maximum spread ⊥ᵃⁿ to the first direction. Each column gives a direction.

- Normalization :

The principal components are supplied with normalized version of original predictors. This is because the original predictors may have different scales.

Performing PCA on unnormalised variables will lead to insanely large loadings for variables with high variance. In turn this will lead to dependence of a principal component on the variable with high variance, which is undesirable.

Variance and Co-variance are a measure of ~~"spread"~~ 'spread' of a set of points around their center of mass (mean).

Variance is measure of the deviation from the mean for points in one dimension.

$\therefore$ Variance = (Standard Deviation)^2

Co-variance is measure of how much each of the dimensions varys from the mean w.r.t each other.

$\therefore$ Covariance $(X, Y) = \sum_{i=1}^{n} \frac{(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})}{(n-1)}$

- Standard Deviation:

Measure of spread of the data points.

$\therefore \sigma = \sqrt{\frac{(x_i - \text{mean})^2}{n}}$

* Conclusion → Thus in this assignment, we studied and implemented the concept of PCA on Big Mart dataset using R.