Implentation of Principal Component Analysis using Singular Value Decomposition

Arvind Murugan | MUR0071

*VSB – Technical University of Ostrava*

# Aim

The main objective is to implement Principal Component Analysis (PCA) using Singular Value Decomposition (SVD).

# Introduction

A matrix may have several eigenvectors such that when the matrix multiplies the eigenvector, the result is a constant multiple of the eigenvector. That constant is the eigenvalue associated with this eigenvector. Together the eigenvector and its eigenvalue are called an eigenpair.

In general, the eigenvectors are called principal axes or principal directions of the data. Projections of the data on the principal axes are called principal components, also known as PC scores; these can be seen as new, transformed, variables.

By representing the matrix of points by a small number of its eigenvectors, we can approximate the data in a way that minimizes the root-mean-square error for the given number of columns in the representing matrix.

The singular-value decomposition of a matrix consists of three matrices, U, Σ, and V. The matrices U and V are column-orthonormal, meaning that as vectors, the columns are orthogonal, and their lengths are 1. The matrix Σ is a diagonal matrix, and the values along its diagonal are called singular values. The product of U, Σ, and the transpose of V equals the original matrix.

The principal components are the columns of V, the coordinates of the data in the basis defined by the principal components are UΣ.

Hence we’ve decided to implement the PCA with SVD.

## Dataset

### Dimension: 671 x 9064

### Attributes

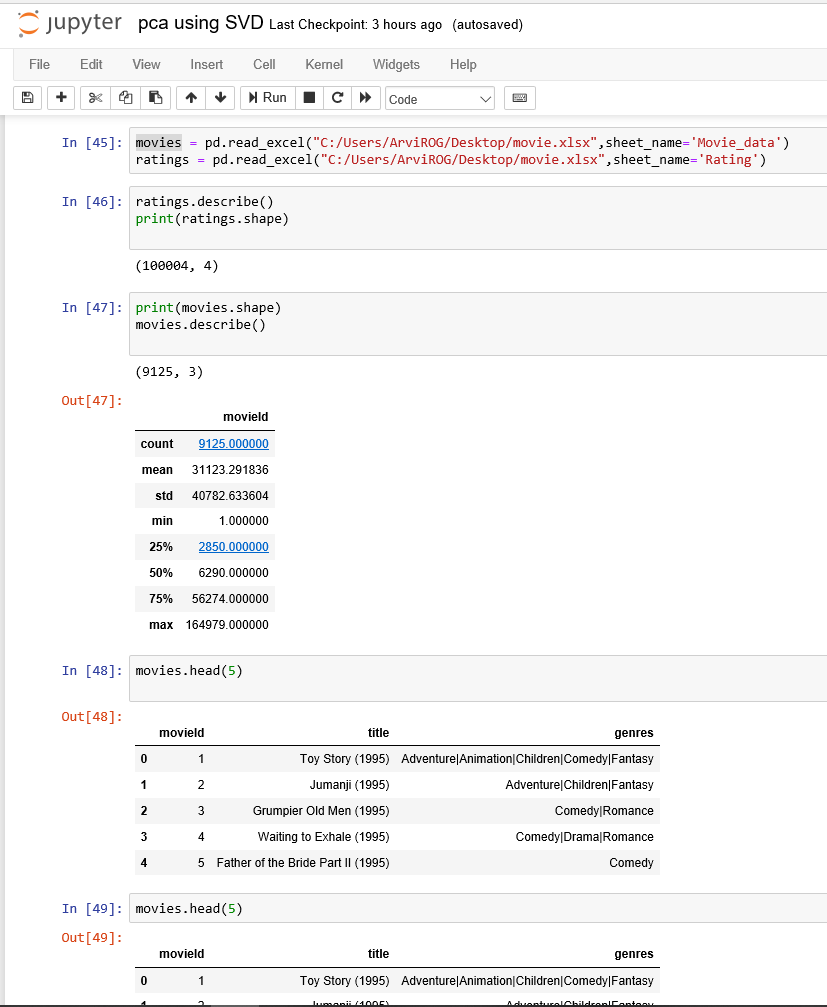
* + - **userId**: Contains unique identification of the users who has voted more than 2 movies.
    - **movieId**: Contains unique identification of the movies
    - **ratings**: A scale of ‘0’ to ‘5’. ‘NaN’ if the user didn’t rate the movie.

### Preprocessing of the dataset

* + - All ‘NaN’ values are changed to ‘0’.
    - Movie ids were replaced with their respective titles.
    - Standardized features by removing the mean and scaling to unit variance.

# Implementation

## Loading Dataset:

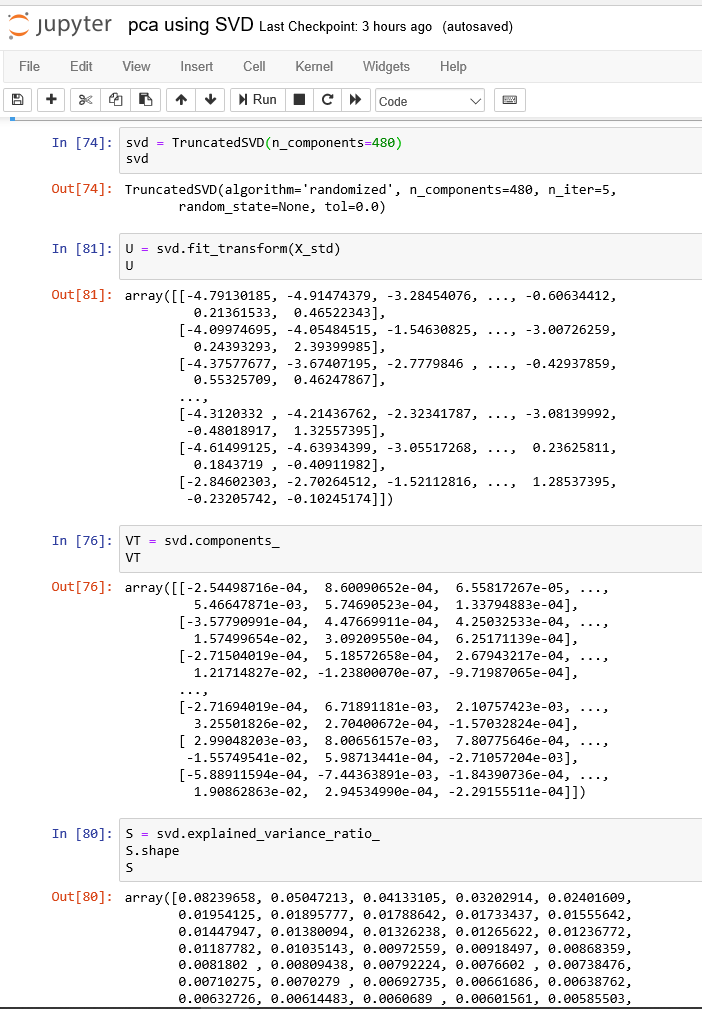


## Replacing Nan Values:

A screenshot of a social media post

Description automatically generated

## `Calculating SVD:



## Components vs cumulative explained variance:

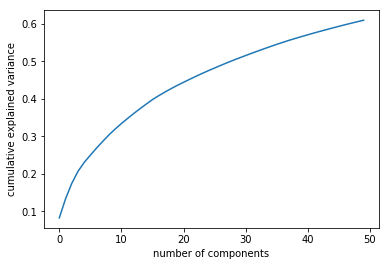
### For number of components = 3

A close up of a logo

Description automatically generated

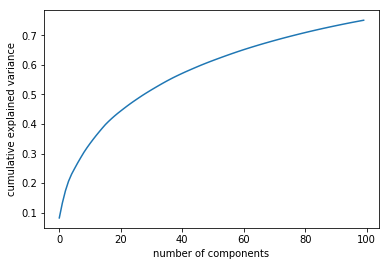
Sum of Explained Variance Ratio = 0.1741997364066073

### For number of components = 50



Sum of Explained Variance Ratio = 0.609631993318398

### For number of components = 100



Sum of Explained Variance Ratio = 0.7503687577394819

### For number of components = 480

## A close up of a logo Description automatically generated

Sum of Explained Variance Ratio = 0.9892265052302065

### For number of components = 240

A close up of a logo

Description automatically generated

Sum of Explained Variance Ratio = 0.9080436305687265

# Observation

With all the K values, it is clearly evident that the minimum number of important components is 240 with the sum of 0.9080436305687265 Explained Variance Ratio.

# Conclusion

In general, PCA looks for such a direction that the data projected to it has the maximal variance.

PCA Using SVD continues by seeking the next direction that is orthogonal to all previously found directions. All directions are orthogonal.

A useful rule of thumb is to retain enough singular values to make up 90% of the energy in Σ. That is, the sum of the squares of the retained singular values should be at least 90% of the sum of the squares of all the singular values.