Principal Component Analysis (PCA)

Introduction

In the world of data science and machine learning, data is often collected with many features or variables, leading to what is known as the **"Curse of Dimensionality."**
The Curse of Dimensionality makes data **processing** and **visualization difficult**. It can lead to model **overfitting** during training, **feature redundancy**, and more complex models that try to consider all features, memorizing the data instead of generalizing from it.

That was the motivation to use Dimensionality Reduction techniques

Dimensionality Reduction:
i) **What?**
Dimensionality reduction is the process of reducing the number of features (dimensions) in your data while preserving as much relevant information as possible.

ii) **Why?**

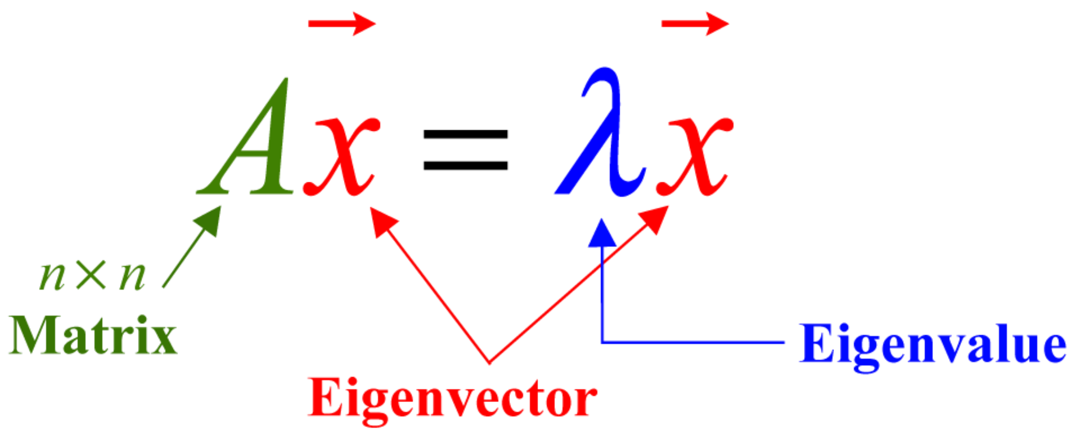
- To overcome the *curse of dimensionality*.
- To reduce computational cost.
- To help in data visualization and understanding.
- To remove noise and redundant features.

iii) **How?**
There are many techniques for dimensionality reduction, but today we will focus on PCA and the math behind it.

Firstly: Eigenvalues and Eigenvectors

- Now we deal with a matrix **A** as if it's a *transformation matrix*, which transforms the entire vector space.
- "Eigen" means *characteristics*, but characteristics of what?
  It's the characteristics of the transformation matrix **A**.
- Eigenvectors are specific vectors resulting from transforming the entire vector space.
- So, which vectors can be called eigenvectors?
  They are the vectors that are only *scaled* and not rotated.
  That means **no rotation is allowed**, only a change in magnitude.
- In other words, on the same line (still using the same span)

- These vectors are scaled by corresponding **eigenvalues**.
- Eigenvalues can be any number, even negative.
  Negative values just reverse the direction of the vector but still keep it in the same span (on the same line).
- Given a transformation matrix of size $n \times n$, the maximum number of independent eigenvectors is n.
- Final formula:



How to get eigenvalues?

- For small n, we can get them using algebraic formulas.

  $$Ax - \lambda x = 0$$

  $$\text{or, } (A - \lambda I)x = 0 \text{ [where I is the identity matrix]}$$

  - Take Determent for both sides
- For $n > 4$, we use **numerical methods**.

Application of Eigenvalues and Eigenvectors in PCA

After understanding the role of eigenvalues and eigenvectors, it becomes clear why they are at the heart of Principal Component Analysis (PCA). Let's now explain how PCA uses them:

Step-by-Step Concept:

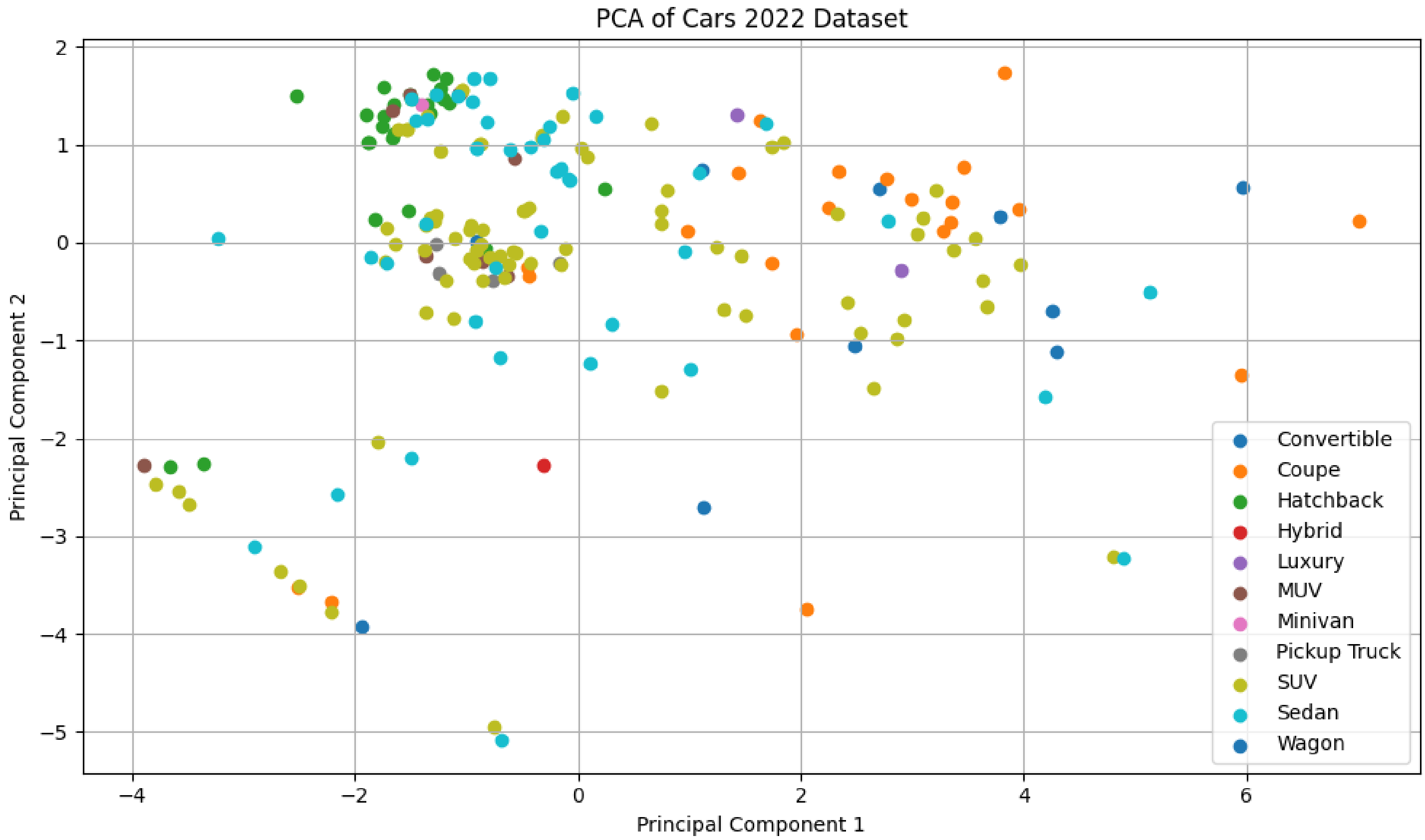
1. **Data Transformation**
   PCA starts by centering the data (subtracting the mean) and forming a covariance matrix. This matrix represents how features vary with respect to each other.
2. **Covariance Matrix as a Transformation Matrix**
   Think of the covariance matrix as a transformation matrix (like matrix A we mentioned before). It transforms the feature space to **align with directions of maximum variance**.
3. **Eigenvectors of the Covariance Matrix**
   The eigenvectors of this covariance matrix give us the principal directions (called *principal components*).
   Each eigenvector shows a new axis in a transformed feature space where the data has the most spread (variance).
4. **Eigenvalues Represent the Variance Along Each Eigenvector**
   The corresponding eigenvalue tells us how much variance (or information) is captured along that new direction.
   - A large eigenvalue means that the direction holds a lot of variance.
   - A small eigenvalue means less variance and can often be discarded without losing much information.

Summary:

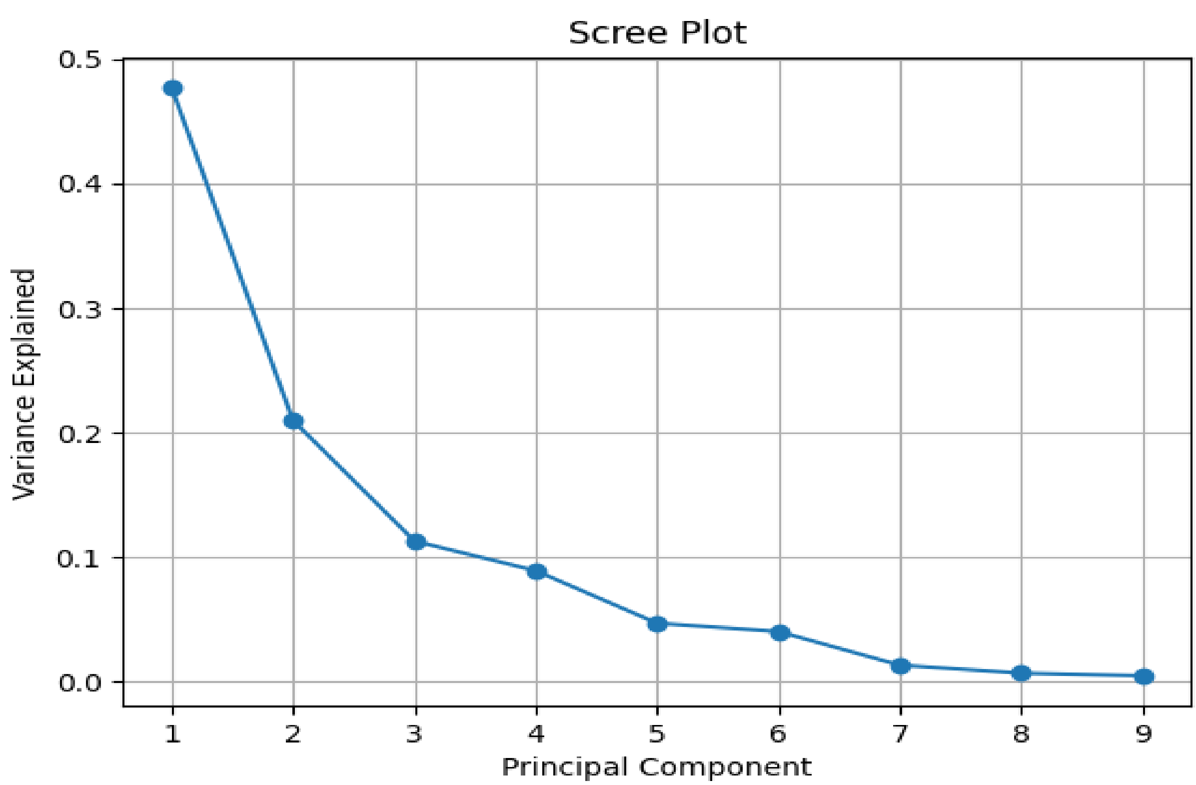
- PCA is all about changing the coordinate system to a new axes (the eigenvectors) where:
  - The axes are orthogonal (eliminate multicolinearity).
  - The data is most spread out in as few dimensions as possible.
- Once you have these, you can:
  - Project the original data onto the top-n eigenvectors (those with the highest eigenvalues(variance)).

Task Visualizations

i) Scatter Plot using top 2 Principle Components



ii) Scree Plot



Resources Used

1. StatQuest with Josh Starmer – *PCA Clearly Explained*
   YouTube Video
2. Dr. Hatem's Master Linear Algebra for Artificial Intelligence Playlist

   YouTube Playlist
3. CalcWorkshop – Eigenvalues and Eigenvectors Explained
   Article & Examples