PCA Step-by-Step (Visualization)

Using Student Dataset (6 Features)

Dataset: 10 Students × 6 Features

Student	Height_cm	Weight_kg	Shoe_Size	Age	Math_Score	English_Score
A	150	50	36	15	78	65
B	160	55	38	16	82	70
C	165	60	39	15	76	60
D	170	65	40	17	85	72
E	175	72	42	16	90	75
F	180	78	44	18	88	74
G	185	85	46	19	92	80
H	155	52	37	15	70	62
I	162	58	39	16	80	68
J	168	64	40	17	83	73

Step 1 — Standardize Features

- Put all features on the same scale (mean=0, std=1)
- Prevents large units (cm, kg) from dominating
- After standardization, every feature has similar spread

PCA Pipeline

1. Standardize → scale features
2. Correlation matrix → see relationships
3. Eigenvalues & loadings → find best directions
4. Pick top PCs (PC1, PC2) → reduce dimensions
5. Use reduced data for visualization & modeling

Standardization Calculation Example (Height)

Original Value (x)	Mean (μ)	Std (σ)	Standardized (z)
150	167.0	10.38	-1.64
185	167.0	10.38	1.73
$z = (x - \mu) / \sigma$			

Standardization: Mean and Std for Each Feature

Feature	Mean (μ)	Std (σ)
Height_cm	167.0	10.38
Weight_kg	63.9	10.84
Shoe_Size	40.1	2.95
Age	16.4	1.28
Math_Score	82.4	6.39
English_Score	69.9	5.89

Student	Height_cm	Weight_kg	Shoe_Size	Age	Math_Score	English_Score
A	-1.63734	-1.28237	-1.39083	-1.09322	-0.68851	-0.83194
B	-0.6742	-0.82109	-0.71238	-0.31235	-0.06259	0.016978
C	-0.19263	-0.3598	-0.37315	-1.09322	-1.00147	-1.68087
D	0.288943	0.101483	-0.03392	0.468521	0.406846	0.356547
E	0.770514	0.747282	0.644531	-0.31235	1.189243	0.865901
F	1.252085	1.300824	1.322984	1.24939	0.876285	0.696116
G	1.733657	1.946624	2.001438	2.030259	1.502202	1.714823
H	-1.15577	-1.09786	-1.0516	-1.09322	-1.94034	-1.3413
I	-0.48157	-0.54432	-0.37315	-0.31235	-0.37555	-0.32259
J	0.096314	0.009226	-0.03392	0.468521	0.093888	0.526332

How the Correlation Matrix Works

- A correlation matrix is a table showing relationships between features.
- Each value ranges from -1 to +1:

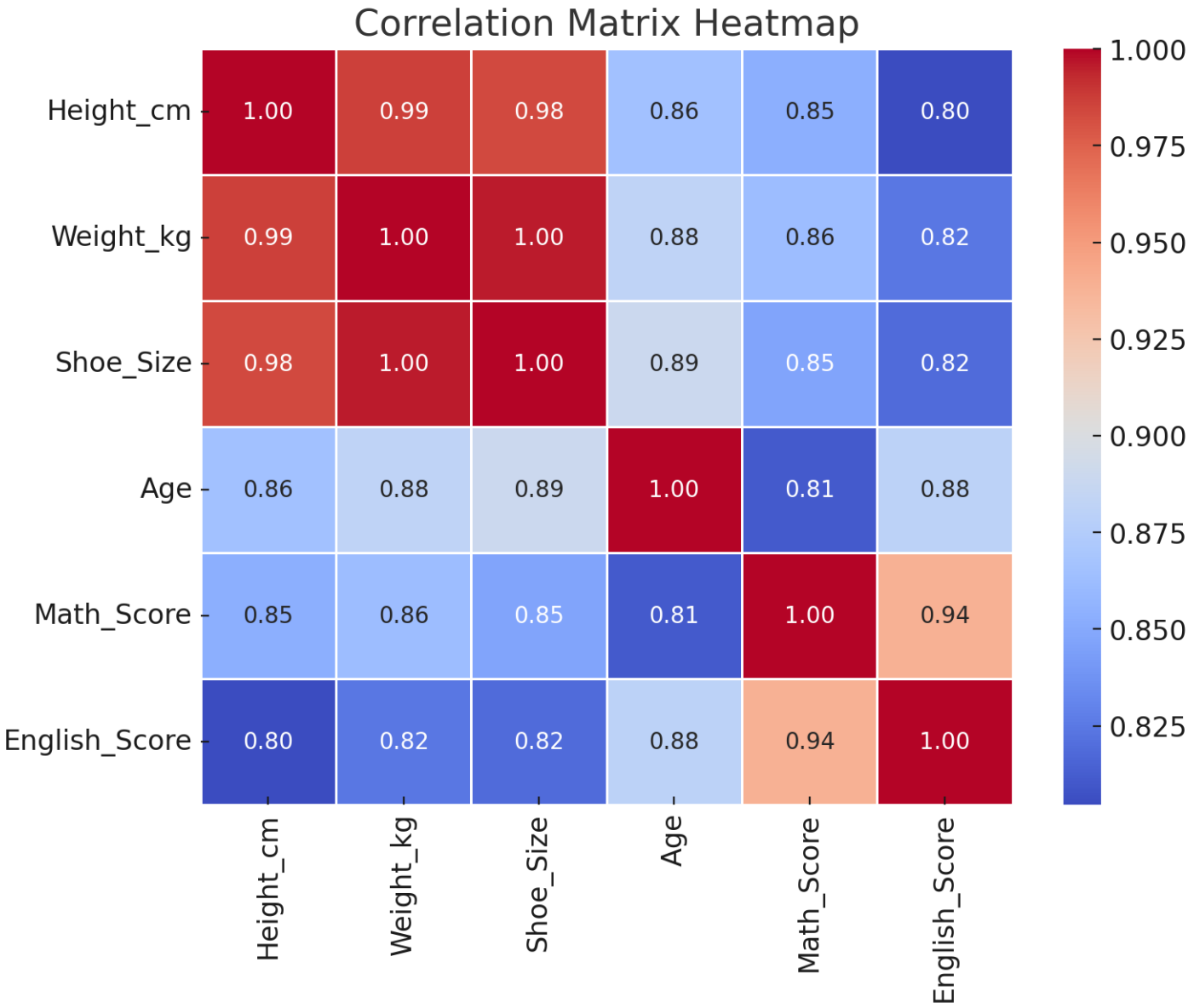
  → +1 means perfect positive relationship (both increase together).

  → -1 means perfect negative relationship (one increases, other decreases).

  → 0 means no relationship.
- Example: Height & Weight usually have strong positive correlation.
- PCA uses the correlation matrix to detect redundancy and group features that carry similar information.

Extended Correlation Matrix (Example Values)

	Height	Weight	Shoe Size	Age	Math	English
Height	1.0	0.95	0.9	0.4	0.1	0.05
Weight	0.95	1.0	0.88	0.42	0.12	0.08
Shoe Size	0.9	0.88	1.0	0.38	0.15	0.1
Age	0.4	0.42	0.38	1.0	0.2	0.18
Math	0.1	0.12	0.15	0.2	1.0	0.85
English	0.05	0.08	0.1	0.18	0.85	1.0


Correlation Matrix Heatmap
Height_cm
Weight_kg
Shoe_Size
Age
Math_Score
English_Score
1.00 0.99 0.98 0.86 0.85 0.80
0.99 1.00 1.00 0.88 0.86 0.82
0.98 1.00 1.00 0.89 0.85 0.82
0.86 0.88 0.89 1.00 0.81 0.88
0.85 0.86 0.85 0.81 1.00 0.94
0.80 0.82 0.82 0.88 0.94 1.00
Height_cm
Weight_kg
Shoe_Size
Age
Math_Score
English_Score
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825

Correlation Matrix: Academic & Age Features

- Age has weaker correlation with physical traits, but can still show trends.

  → Older students may be slightly taller/heavier, but not always.

- Math Score and English Score usually have strong positive correlation.

  → Students good at Math are often good at English too.

- These academic features are not strongly tied to Height/Weight/Shoe Size.
- PCA identifies this and may form a separate component for 'Academic Factor'.

Eigenvectors = Principal Components

- PCA creates new 'axes' (principal components) that summarize data.
- Each axis is an eigenvector: a direction combining original features.
- PCs are ranked:
  → PC1 explains the most variation
  → PC2 explains the next biggest variation
  → Later PCs explain much less

Interpreting Our Dataset's PCs

- PC1 ≈ 'Overall Size Factor'
- – Combines Height, Weight, Shoe Size
- – Captures differences in body size among students

- PC2 ≈ 'Academic Performance Factor'
- – Combines Math & English scores
- – Captures variation in student performance

- PC3–PC6
- – Represent smaller details (e.g., slight age variation)
- – Explain very little variance; often ignored

Eigenvalues (Importance of Each PC)

PC	Eigenvalue	Explained Variance (%)
PC1	4.36	72.7
PC2	1.3	21.6
PC3	0.22	3.7
PC4	0.09	1.5
PC5	0.02	0.4
PC6	0.01	0.1

Eigenvectors (Principal Components)

PC	Height_cm	Weight_kg	Shoe_Size	Age	Math_Score	English_Score
PC1	0.45	0.46	0.44	0.34	0.29	0.29
PC2	0.05	0.08	0.12	0.1	0.7	0.7
PC3	0.12	0.2	-0.05	0.94	-0.15	-0.09

- **PC1**: Height, Weight, Shoe Size have the biggest coefficients → this direction = **Size Factor**.
- **PC2**: Math and English dominate → this direction = **Academic Factor**.
- **PC3**: Age has the strongest weight (0.94) → this direction = **Age Factor**.

Practical Rule of Thumb

- A feature is considered **important** for a component if its coefficient (loading) is relatively large compared to others in that row.
- You can also **square the loadings** to see the proportion of variance each feature contributes to the component.