🗒️ PCA Analysis Report: With vs Without Scaling

📌 Dataset Context

This synthetic dataset mimics real-world health and lifestyle metrics, designed for a Placement Coordinator exploring data-driven insights. Features include:

- Physical metrics: age, height_cm, weight_kg
- Activity/lifestyle habits: daily_steps, water_intake_liters, sleep_hours, alcohol_units_weekly, smoking_frequency
- Health indicators: resting_heart_rate, cholesterol_mg_dl

🎯 Objective

To perform Principal Component Analysis (PCA) on the dataset:

- First without feature scaling, then
- With standardization (mean = 0, std = 1),
- And compare how the results change in terms of variance explained and insightful patterns.

🔁 PCA Without Scaling

🔢 Results

- PC1 explains over 99% of the total variance.
- This is because daily_steps has a much higher numeric range than other features.

📉 Interpretation

- PCA fixates on the variance in daily_steps due to its scale.
- Other features are largely ignored by the algorithm.

⚠️ Insight

PCA without scaling is dominated by high-magnitude features — leading to biased, misleading results.

📊 PCA With Scaling (Standardized Data)

🔢 Results

- PC1 explains ~13% of the variance.
- PC2 explains ~11%, and subsequent components contribute similarly.
- No single PC dominates — variance is evenly spread, indicating complex, multidimensional structure.

🧠 Interpretation

PC1: “Age & Sedentary Lifestyle”

- **High positive weight for age, sleep_hours**
- **High negative weight for daily_steps, height_cm**
- **Captures a pattern of aging with decreasing activity and stature**

PC2: “Wellness vs Risk Habits”

- **High positive weight for water_intake, height_cm**
- **High negative weight for sleep_hours, smoking_frequency, daily_steps**
- **Contrasts healthier habits vs fatigue and smoking-related patterns**

✅ Insight

After scaling, PCA captures richer, fairer structure across all features.
Though no component dominates, combinations of PCs now reflect interpretable lifestyle and health factors.

🧠 Key Learnings

Aspect	Without Scaling	With Scaling
Dominant Feature	daily_steps (due to scale)	None (balanced contributions)
PC1 Theme	Pure step-count variance	Age, activity, and sleep pattern
PC2 Theme	Not meaningful	Health-conscious vs. risky behaviors
Interpretability	❌ Biased and misleading	✅ Subtle but realistic
Usefulness	Low for real-world pattern discovery	High for understanding hidden relationships

📌 Conclusion

- Standardization before PCA is essential for fair dimensionality reduction.
- A lower explained variance per component isn't a limitation — it reflects natural complexity.
- PCA with scaling offered valuable insights into lifestyle patterns that would've otherwise been hidden.