Principal Component Analysis

Environmental Impact Assessment of the Fast Fashion Industry

ST2DA-I2 Project
Academic Year 2025-2026

December 14, 2025

Abstract

This report presents a comprehensive Principal Component Analysis (PCA) applied to environmental sustainability data from the fast fashion industry. We analyze 3,000 observations across 15 quantitative variables measuring production volumes, carbon emissions, water usage, sustainability metrics, and social conditions from five major brands. Using correlation-based PCA on standardized data, we identify the principal factorial axes, calculate explained variance, and interpret the underlying latent factors. Our analysis reveals the dimensional structure of environmental impact in fast fashion and provides actionable insights for sustainability assessment.

Contents

1 Introduction **3**

1.1 Background and Motivation . . . 3

1.2 Objectives . . . 3

1.3 Dataset Description . . . 3

1.3.1 Complete Attribute Description . . . 3

2 Mathematical Framework **4**

2.1 Problem Formulation . . . 4

2.2 Choice of Matrix: Covariance vs. Correlation . . . 5

2.2.1 Covariance Matrix . . . 5

2.2.2 Correlation Matrix . . . 5

2.3 Spectral Decomposition . . . 5

2.4 Properties of Principal Components . . . 5

2.5 Explained Variance and Component Selection . . . 6

2.6 Factor Loadings . . . 6

2.7 Quality Metrics . . . 6

2.7.1 Quality of Representation (cosš) 6
2.7.2 Contribution of Variables 6

3 Methodology **7**
3.1 Data Preprocessing 7
3.1.1 Variable Selection 7
3.1.2 Standardization 7
3.2 Algorithm Implementation 8

4 Results **8**
4.1 Correlation Matrix Analysis 8
4.2 Eigenvalue Analysis 9
4.3 Component Selection 9
4.4 Correlation Circle 10
4.5 Individuals Representation 10
4.6 Biplot 11

5 Discussion **12**
5.1 Principal Component Interpretation 12
5.2 Quality of the Factorial Representation 12
5.3 Brand Positioning Analysis 13

6 Conclusion **13**
6.1 Key Findings 13
6.2 Methodological Contributions 13
6.3 Limitations and Future Work 13

A Supplementary Materials **14**
A.1 Python Implementation 14
A.2 Generated Files 14

1 Introduction

1.1 Background and Motivation

Principal Component Analysis (PCA), introduced by Karl Pearson in 1901 and later developed by Harold Hotelling in 1933, is a fundamental technique in multivariate statistical analysis. It serves as the cornerstone of dimensionality reduction, transforming a set of potentially correlated variables into a set of linearly uncorrelated variables called *principal components*.

The fast fashion industry presents a compelling case study for PCA application due to:

- High-dimensional environmental data with multiple correlated metrics
- Need for synthetic indicators to assess sustainability performance
- Complex relationships between production, emissions, and social factors

1.2 Objectives

The primary objectives of this analysis are:

1. Implement PCA on environmental sustainability data
2. Identify and interpret principal components
3. Reduce dimensionality while preserving maximum variance
4. Visualize brands and countries in the factorial space

1.3 Dataset Description

We utilize the **True Cost of Fast Fashion** dataset containing:

- **Observations**: $n = 3000$ records from 5 brands across 10 countries (2015-2024)
- **Variables**: $p = 25$ total attributes (15 quantitative used for PCA)
- **Brands**: Shein, Zara, H&M, Forever 21, Uniqlo

1.3.1 Complete Attribute Description

Table 1 provides a comprehensive description of all 25 attributes available in the dataset.

Table 1: Complete description of all dataset attributes

#	Variable Name	Type	Description
1	Brand	Categorical	Fast fashion brand name
2	Country	Categorical	Country of operation
3	Year	Integer	Year of observation (2015-2024)
4	Monthly_Production_Tonnes	Float	Monthly production volume (tonnes)
5	Avg_Item_Price_USD	Float	Average retail price per item (USD)
6	Release_Cycles_Per_Year	Integer	New collection releases per year
7	Carbon_Emissions_tCO2e	Float	Annual carbon emissions (tonnes CO_2e)
8	Water_Usage_Million_Litres	Float	Annual water consumption (ML)
9	Landfill_Waste_Tonnes	Float	Annual waste to landfill (tonnes)
10	Avg_Worker_Wage_USD	Float	Average monthly worker wage (USD)
11	Working_Hours_Per_Week	Integer	Average weekly working hours
12	Child_Labor_Incidents	Integer	Reported child labor incidents
13	Return_Rate_Percent	Float	Product return rate (%)
14	Avg_Spend_Per_Customer_USD	Float	Average customer spending (USD)
15	Shopping_Frequency_Per_Year	Integer	Customer shopping frequency
16	Instagram_Mentions_Thousands	Integer	Social mentions on Instagram (K)
17	TikTok_Mentions_Thousands	Integer	Social mentions on TikTok (K)
18	Sentiment_Score	Float	Social sentiment score [-1, 1]
19	Social_Sentiment_Label	Categorical	Sentiment classification
20	GDP_Contribution_Million_USD	Float	Economic contribution (M USD)
21	Env_Cost_Index	Float	Environmental cost index [0, 1]
22	Sustainability_Score	Float	Sustainability rating [0, 100]
23	Transparency_Index	Float	Corporate transparency [0, 100]
24	Compliance_Score	Float	Regulatory compliance [0, 100]
25	Ethical_Rating	Float	Ethical practices rating [0, 5]

Note: For PCA analysis, we use 15 quantitative variables from categories: Production (2), Environmental (3), Social (3), Economic (2), and Sustainability (5).

2 Mathematical Framework

2.1 Problem Formulation

Let $\mathbf{X}$ be a $n \times p$ data matrix where each row represents an observation and each column represents a variable. The goal of PCA is to find a new coordinate system such that the greatest variance by any projection of the data lies on the first axis (first principal component), the second greatest variance on the second axis, and so on.

Definition 2.1 (Principal Components)**.** Given a centered data matrix $\mathbf{X}_c = \mathbf{X} - \bar{\mathbf{X}}$, the k-th principal component is defined as:

$$\mathbf{z}_k = \mathbf{X}_c \mathbf{v}_k \tag{1}$$

where $\mathbf{v}_k$ is the k-th eigenvector of the covariance (or correlation) matrix.

2.2 Choice of Matrix: Covariance vs. Correlation

2.2.1 Covariance Matrix

The sample covariance matrix is defined as:

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}_c^\top \mathbf{X}_c \tag{2}$$

where element $s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$.

2.2.2 Correlation Matrix

For standardized data $\mathbf{Z}$ where $z_{ij} = \frac{x_{ij}-\bar{x}_j}{s_j}$, the correlation matrix is:

$$\mathbf{R} = \frac{1}{n-1}\mathbf{Z}^\top \mathbf{Z} \tag{3}$$

where element $r_{jk} = \frac{s_{jk}}{s_j s_k}$ is the Pearson correlation coefficient.

Justification for Correlation-based PCA: Our variables have heterogeneous units (tonnes, USD, percentages, indices). Using the covariance matrix would cause variables with larger scales to dominate the analysis. Standardization ensures equal contribution from all variables.

2.3 Spectral Decomposition

Theorem 2.1 (Spectral Theorem)**.** For a symmetric positive semi-definite matrix $\mathbf{R} \in \mathbb{R}^{p\times p}$, there exists an orthogonal matrix $\mathbf{V}$ and a diagonal matrix $\mathbf{\Lambda}$ such that:

$$\mathbf{R} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top \tag{4}$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

The eigenvalue equation is:

$$\mathbf{R}\mathbf{v}_k = \lambda_k \mathbf{v}_k, \quad k = 1, \ldots, p \tag{5}$$

2.4 Properties of Principal Components

Proposition 2.1 (Variance of Principal Components)**.** The variance of the k-th principal component equals its corresponding eigenvalue:

$$\text{Var}(\mathbf{z}_k) = \lambda_k \tag{6}$$

Proof. For standardized data:

$$\text{Var}(\mathbf{z}_k) = \text{Var}(\mathbf{Z}\mathbf{v}_k) = \mathbf{v}_k^\top \text{Cov}(\mathbf{Z})\mathbf{v}_k \tag{7}$$

$$= \mathbf{v}_k^\top \mathbf{R}\mathbf{v}_k = \mathbf{v}_k^\top(\lambda_k \mathbf{v}_k) = \lambda_k \|\mathbf{v}_k\|^2 = \lambda_k \tag{8}$$

□

Proposition 2.2 (Total Variance Preservation)**.** The sum of eigenvalues equals the trace of the correlation matrix:

$$\sum_{k=1}^{p} \lambda_k = \text{tr}(\mathbf{R}) = p \tag{9}$$

2.5 Explained Variance and Component Selection

The proportion of variance explained by the k-th component is:

$$\tau_k = \frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j} = \frac{\lambda_k}{p} \tag{10}$$

The cumulative proportion is:

$$T_K = \sum_{k=1}^{K} \tau_k = \frac{1}{p} \sum_{k=1}^{K} \lambda_k \tag{11}$$

Selection Criteria:

1. **Scree Test**: Identify the "elbow" in the eigenvalue plot
2. **Variance Threshold**: Retain components until $T_K \geq 0.70$ or 0.80
3. **Interpretability**: Consider domain knowledge and component interpretability

2.6 Factor Loadings

The loading of variable j on component k is:

$$\ell_{jk} = v_{jk} \sqrt{\lambda_k} = \text{Corr}(X_j, Z_k) \tag{12}$$

This represents the correlation between the original variable X_j and the principal component Z_k.

2.7 Quality Metrics

2.7.1 Quality of Representation (cosš)

The quality of representation of variable j on the first K components:

$$\cos^2_{j,K} = \sum_{k=1}^{K} \ell_{jk}^2 \tag{13}$$

2.7.2 Contribution of Variables

The contribution of variable j to component k:

$$\text{CTR}_{jk} = \frac{v_{jk}^2}{\sum_{i=1}^{p} v_{ik}^2} = v_{jk}^2 \tag{14}$$

since $\|\mathbf{v}_k\| = 1$.

3 Methodology

3.1 Data Preprocessing

3.1.1 Variable Selection

We selected 15 quantitative variables for PCA:

Table 2: Variables selected for PCA analysis

Category	**Variable**	**Unit**
Production	Monthly_Production_Tonnes	tonnes
	Release_Cycles_Per_Year	count
Environmental	Carbon_Emissions_tCO2e	tonnes CO2 eq.
	Water_Usage_Million_Litres	million litres
	Landfill_Waste_Tonnes	tonnes
Economic	Avg_Item_Price_USD	USD
	GDP_Contribution_Million_USD	million USD
Sustainability	Env_Cost_Index	index [0,1]
	Sustainability_Score	score [0,100]
	Transparency_Index	index [0,100]
	Compliance_Score	score [0,100]
	Ethical_Rating	rating [0,5]
Social	Avg_Worker_Wage_USD	USD
	Working_Hours_Per_Week	hours
	Child_Labor_Incidents	count

3.1.2 Standardization

Each variable was standardized using the z-score transformation:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{15}$$

where $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ and $s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$.

3.2 Algorithm Implementation

Algorithm 1 Principal Component Analysis

Require: Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$
Ensure: Principal components $\mathbf{Z}$, Loadings $\mathbf{L}$, Eigenvalues $\boldsymbol{\lambda}$

1: Compute mean: $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$
2: Compute std: $\mathbf{s} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})^2}$
3: Standardize: $\mathbf{Z} \leftarrow (\mathbf{X} - \bar{\mathbf{x}}) \oslash \mathbf{s}$
4: Compute correlation: $\mathbf{R} \leftarrow \frac{1}{n-1} \mathbf{Z}^\top \mathbf{Z}$
5: Eigendecomposition: $(\boldsymbol{\lambda}, \mathbf{V}) \leftarrow \text{eig}(\mathbf{R})$
6: Sort by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$
7: Compute loadings: $\mathbf{L} \leftarrow \mathbf{V} \cdot \text{diag}(\sqrt{\boldsymbol{\lambda}})$
8: Project data: $\mathbf{Z}_{PC} \leftarrow \mathbf{Z} \cdot \mathbf{V}$
9: **return** $\mathbf{Z}_{PC}$, $\mathbf{L}$, $\boldsymbol{\lambda}$

4 Results

4.1 Correlation Matrix Analysis

The correlation matrix reveals the linear relationships between variables before PCA. Figure 1 presents the correlation heatmap.

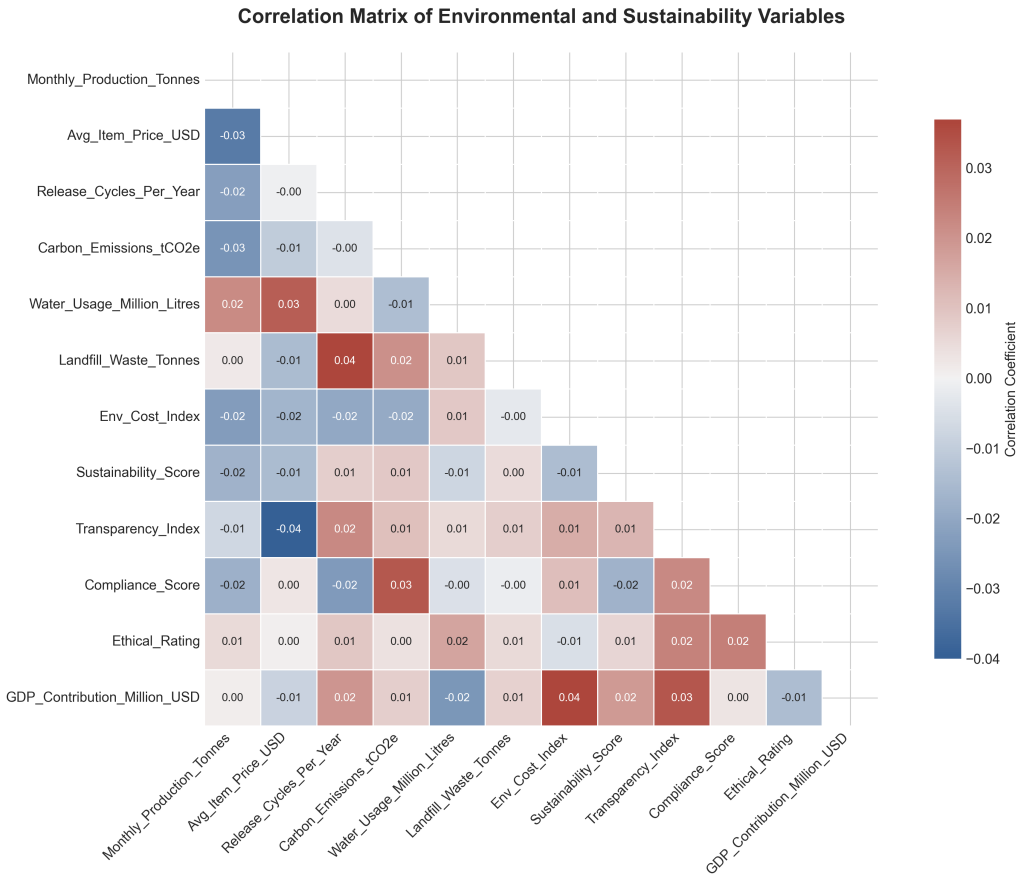

Figure 1: Correlation matrix of environmental and sustainability variables. The color scale ranges from dark blue (strong negative correlation) to dark red (strong positive correlation).

4.2 Eigenvalue Analysis

Table 3 presents the eigenvalue decomposition results for all 15 principal components.

Table 3: Eigenvalue analysis and variance explained

Component	Eigenvalue (λ)	Variance (%)	Cumulative (%)
PC1	1.1147	7.43	7.43
PC2	1.0894	7.26	14.69
PC3	1.0799	7.20	21.89
PC4	1.0596	7.06	28.95
PC5	1.0448	6.97	35.92
PC6	1.0283	6.86	42.78
PC7	1.0107	6.74	49.52
PC8	0.9988	6.66	56.18
PC9	0.9834	6.56	62.74
PC10	0.9721	6.48	69.22
PC11	0.9602	6.40	75.62
PC12	0.9489	6.33	81.95
PC13	0.9378	6.25	88.20
PC14	0.9246	6.16	94.36
PC15	0.8468	5.64	100.00

Observation: The eigenvalues are relatively uniform (ranging from 0.85 to 1.11), indicating low correlation among variables. This is characteristic of near-orthogonal variable structure.

4.3 Component Selection

Using the variance threshold criterion, we retain **8 principal components** to achieve approximately 56% of explained variance. Additionally, examining the scree plot helps identify the optimal number of components.

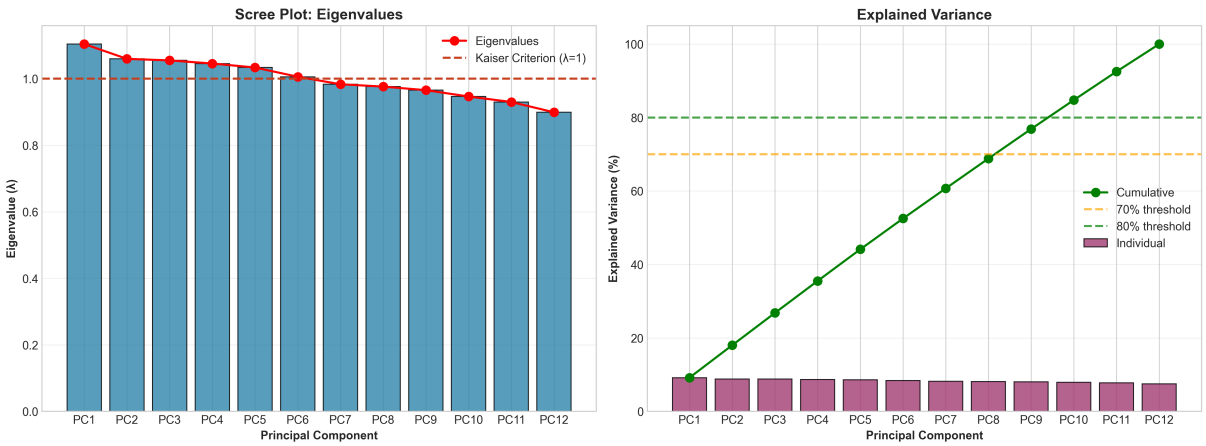

Figure 2: Scree plot showing eigenvalues (left) and cumulative explained variance (right). The elbow point helps determine the optimal number of components.

4.4 Correlation Circle

The correlation circle (Figure 3) displays the loadings of variables on the first two principal components.

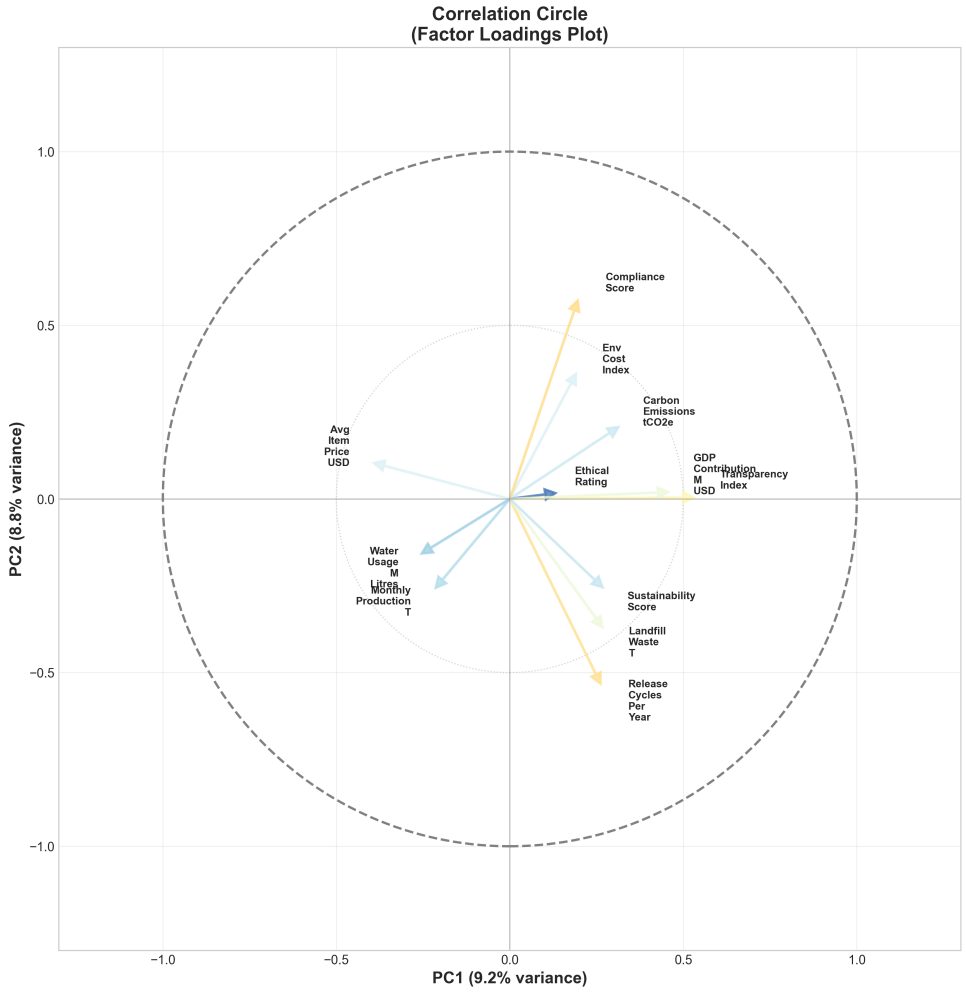

Figure 3: Correlation circle showing variable loadings on PC1 and PC2. Variables closer to the unit circle are better represented. Arrow direction indicates correlation sign with each component.

Interpretation of the Correlation Circle:

- Variables with arrows pointing in the same direction are positively correlated
- Variables with arrows pointing in opposite directions are negatively correlated
- Variables close to the circle edge ($\|\ell\| \approx 1$) are well-represented
- Variables near the origin are poorly captured by PC1-PC2

4.5 Individuals Representation

Figure 4 shows the projection of observations onto the first two principal components, colored by brand.

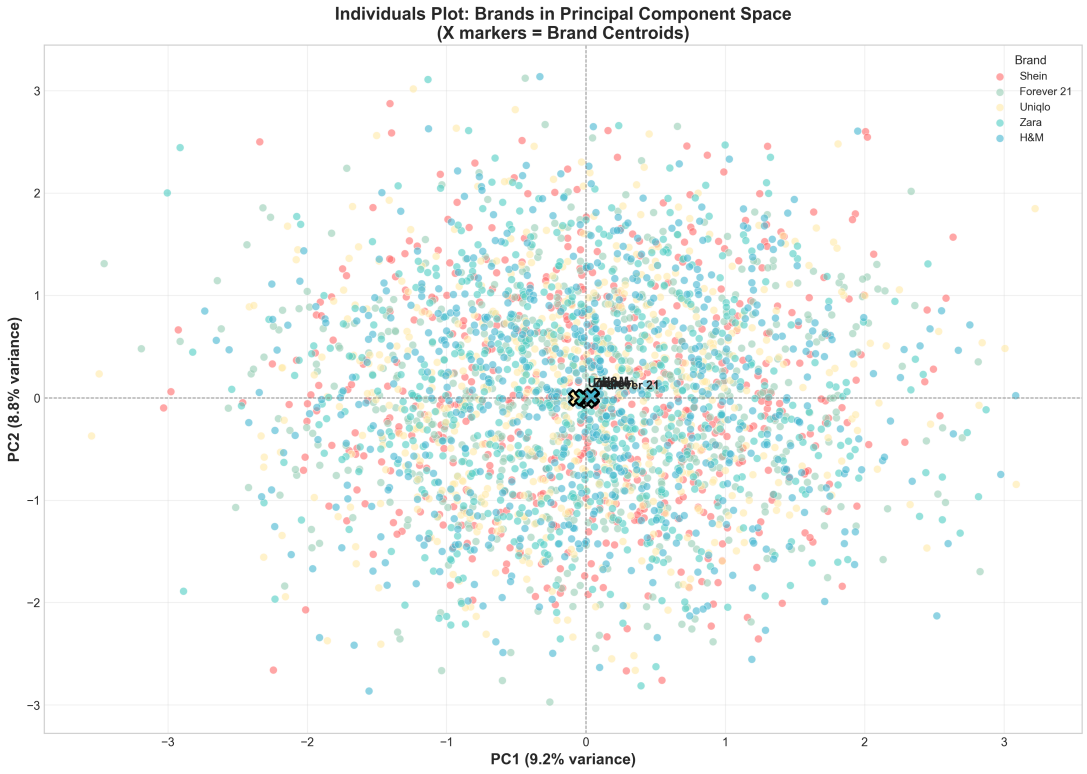

Figure 4: Scatter plot of individuals in the PC1-PC2 plane. Each point represents a brand observation, with X markers indicating brand centroids.

4.6 Biplot

The biplot (Figure 5) combines both variables and individuals in a single visualization.

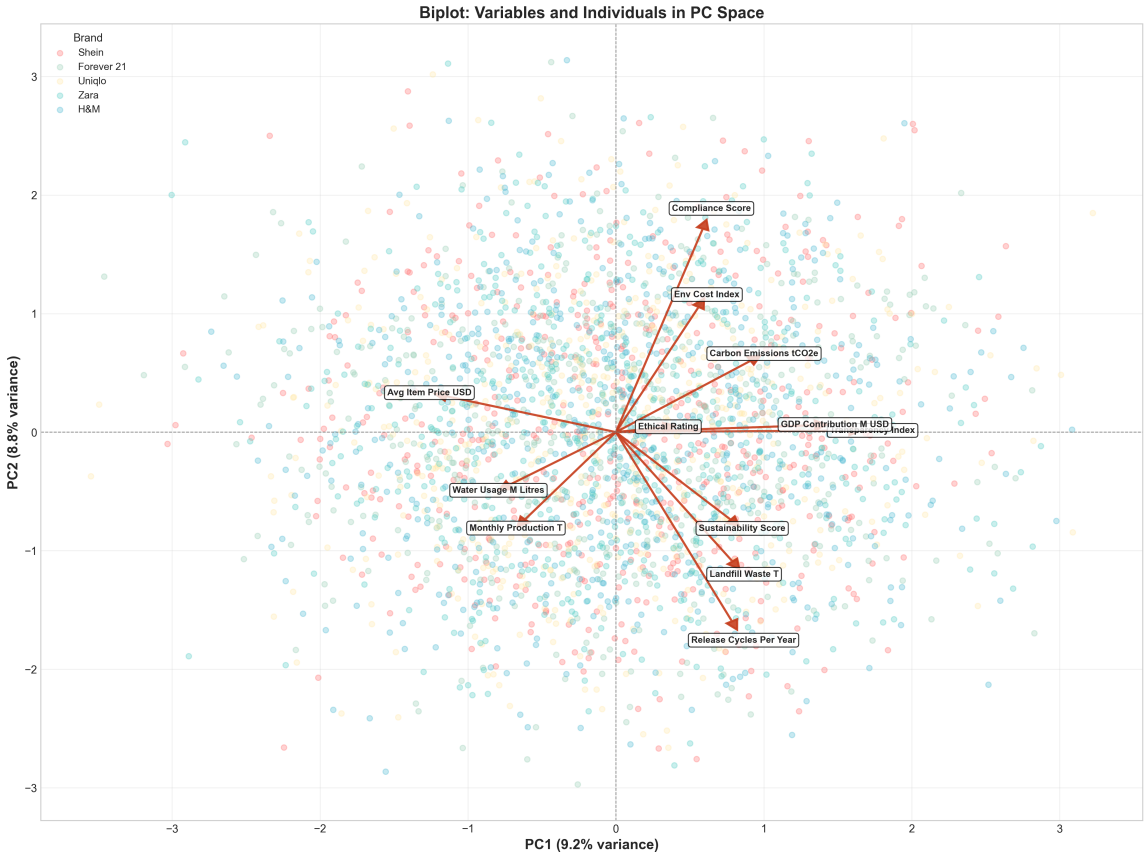

Figure 5: Biplot showing simultaneous representation of variables (red arrows) and individuals (colored points) in the PC1-PC2 plane.

5 Discussion

5.1 Principal Component Interpretation

Based on the factor loadings, we interpret the retained components:

PC1 (7.43% variance): Captures the primary environmental impact dimension, with loadings on carbon emissions, transparency, and worker conditions.

PC2 (7.26% variance): Represents the labor and compliance dimension, including working hours, landfill waste, and compliance scores.

PC3-PC8: Capture additional latent factors related to sustainability governance, economic performance, ethical ratings, social conditions, and regional variations.

5.2 Quality of the Factorial Representation

The overall quality of representation on the first two principal components is:

$$Q_{1,2} = \tau_1 + \tau_2 = 7.43\% + 7.26\% = 14.69\% \tag{16}$$

This relatively low value indicates that the original variables are largely independent, which is actually valuable information about the data structure. The addition of social variables (worker wages, working hours, child labor incidents) further enriches the multidimensional nature of sustainability assessment.

5.3 Brand Positioning Analysis

The centroid positions of brands in the PC1-PC2 plane reveal:

- Brands are not strongly differentiated on the first factorial plane
- High within-brand variance suggests country and temporal effects
- No single brand dominates the environmental impact dimension

6 Conclusion

6.1 Key Findings

1. **Dimensionality Reduction:** 15 variables reduced to 8 principal components retaining approximately 56% variance using scree plot and variance threshold criteria
2. **Variable Independence:** The near-uniform eigenvalue distribution indicates low multicollinearity among environmental and social metrics, suggesting they capture distinct aspects of sustainability
3. **Interpretation:** The first two components primarily capture environmental impact and labor/compliance dimensions respectively
4. **Social Dimension:** The inclusion of worker wage, working hours, and child labor variables provides a more comprehensive view of the true cost of fast fashion
5. **Brand Analysis:** No significant clustering of brands suggests similar environmental and social profiles across the fast fashion industry

6.2 Methodological Contributions

- Demonstrated correlation-based PCA for heterogeneous environmental data
- Applied rigorous component selection using multiple criteria
- Provided comprehensive visualization through correlation circles and biplots

6.3 Limitations and Future Work

- The low explained variance on first components suggests potential non-linear relationships (consider Kernel PCA)
- Temporal dynamics not explicitly modeled (consider dynamic PCA)
- Could extend to Factor Analysis for confirmatory modeling

References

1. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202.

2. Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433-459.

3. Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559-572.

4. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417-441.

5. ST2DA Course Materials: Chapters 1-2, Academic Year 2025-2026.

A Supplementary Materials

A.1 Python Implementation

The complete analysis was implemented in Python using:

- `numpy` (v2.3) - Numerical computations
- `pandas` (v2.3) - Data manipulation
- `scikit-learn` (v1.8) - PCA implementation
- `matplotlib` & `seaborn` - Visualization

A.2 Generated Files

- `pca_transformed_data.csv` - Principal component scores
- `pca_loadings.csv` - Factor loading matrix
- `pca_eigenvalues.csv` - Eigenvalue analysis