PREPROCESSING DATA (1)

Macam-macam Preprocessing Data

- Aggregasi
- Sampling
- Pengurangan dimensi
  - Pemilihan fitur yang beririsan
  - Penciptaan fitur
- Diskritisasi dan binerisasi
- Transformasi Atribut

AGGREGASI (PENGGABUNGAN)

- Menggabungkan dua atau lebih atribut (atau objek) ke dalam satu atribut (atau objek)
- Contoh:
  - Mengganti semua data transaksi sebuah toko tunggal dengan sebuah data transaksi tunggal terpusat (OLAP)
- Tujuan:
  - Reduksi data
    - Mengurangi jumlah atribut/objek
  - Perubahan skala
    - Menggabungkan data kota ke dalam data propinsi, negara, dst
  - Data yang lebih stabil
    - Penggabungan data memungkinkan terjadinya pengurangan variabilitas
- Kerugian:
  - Potensi kehilangan data detil yang berguna

CONTOH AGGREGASI

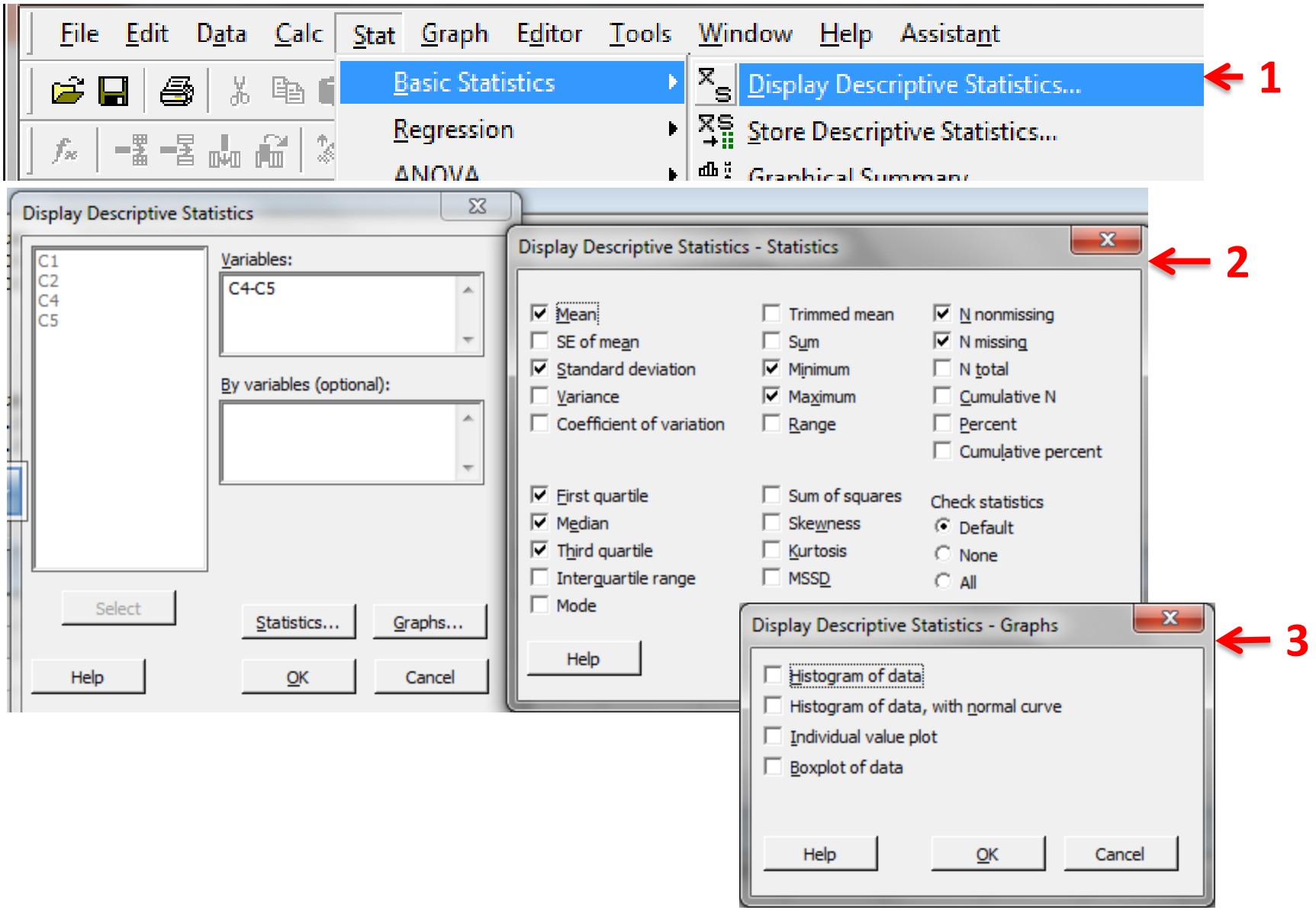
- Jika diberikan data rata-rata curah hujan bulanan dan tahunan dari beberapa lokasi di negara X sebagai berikut:

Bulan	JmlKec	CurahHujan
1	18	12
2	13	120
3	22	6
4	1	24
5	24	165
6	10	44
7	20	174
8	33	150
9	44	162
10	39	91
11	46	180
12	7	14
13	6	48
14	17	74
15	17	53
16	43	160
17	41	106
18	8	10
19	40	84
20	22	166
21	37	73
22	36	127
23	12	26
24	22	147

Hitung:

- Berapa rata-rata curah hujan bulanan dan tahunan untuk masing-masing wilayah
- Amati bagaimana trend curah hujan tahunan untuk masing-masing wilayah
- Dari hasil pengamatan Anda; bagaimana pendapat Anda tentang variability (perubahan) curah hujan untuk masing-masing wilayah

LANGKAH PENYELESAIAN (1)



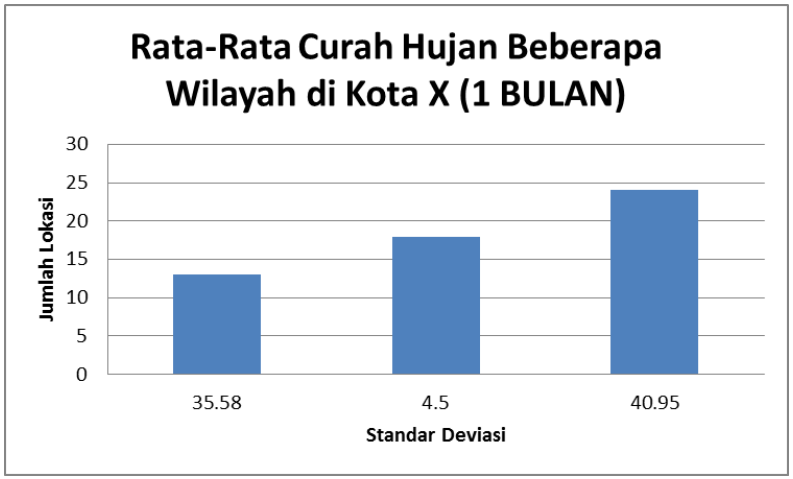
LANGKAH PENYELESAIAN (2)

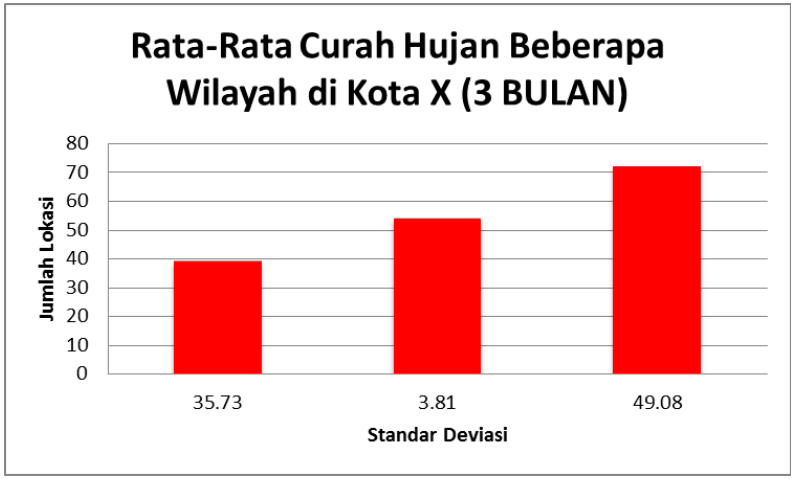
Descriptive Statistics: C10, C11, C12

Variable	N	N*	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
C10	13	0	65.69	35.58	5.00	43.50	62.00	102.50	116.00
C11	18	0	5.72	4.50	0.00	1.75	4.50	10.00	12.00
C12	24	0	110.63	40.95	7.00	89.50	121.50	141.75	160.00

Descriptive Statistics: C14, C15, C16

Variable	N	N*	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
C14	39	0	62.26	35.73	1.00	38.00	63.00	98.00	119.00
C15	54	0	5.722	3.814	0.000	2.000	6.000	9.250	12.000
C16	72	0	93.15	49.08	1.00	49.00	107.50	136.00	165.00





SAMPLING

- Merupakan teknik utama yang sering diaplikasikan untuk proses pemilihan data
  - Sering digunakan untuk investigasi data di awal dan analisis data akhir
- Penggunaan sample merupakan cara paling efektif untuk memperoleh sekumpulan data yang diinginkan
- Sampling digunakan dalam data mining karena mengambil data keseluruhan dapat menghabiskan banyak biaya (mahal) dan memakan waktu yang lama
- Kunci utama untuk pengambilan sample secara efektif:
  - Penggunaan data sample akan menghasilkan output sama bagusnya dengan penggunaan data set secara penuh jika sample yang diambil representatif
  - Sebuah sample dikatakan representatif jika memiliki hampir semua properti data set asli

PENDEKATAN SAMPLING

- Jika jumlah record tidak cukup; maka:
  - Gunakan sample random
    - Sampling tanpa replacement
      - Setelah digunakan sebagai sample; data di-remove dari populasi
    - Sampling dengan replacement
      - Setelah digunakan sebagai sample; data dibiarkan tetap berada di dalam populasi sehingga memungkinkan penggunaan ulang record
  - Gunakan sample bertingkat
    - Pecah data ke dalam beberapa partisi dan gunakan random sample untuk masing-masing partisi

PENGURANGAN DIMENSI

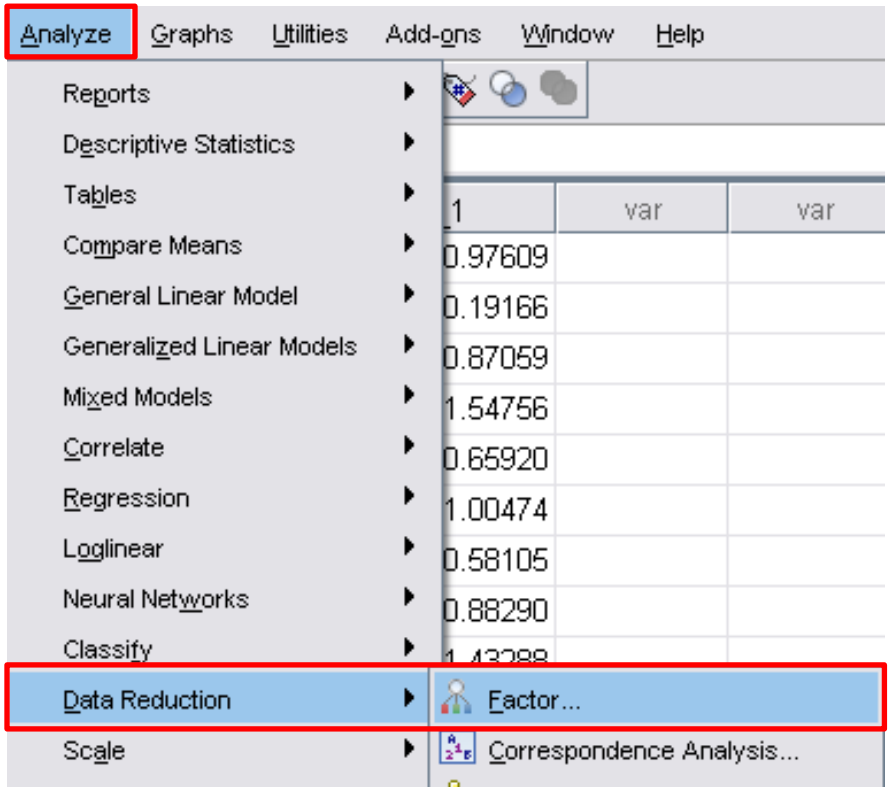
- Manfaat
  - Kebanyakan algoritma data mining bekerja lebih baik pada data dengan ukuran dimensi rendah
  - Reduksi dapat mempermudah pemahaman terhadap model karena berisi sedikit atribut
  - Data lebih mudah divisualisasikan
  - Jumlah waktu dan memori yang dibutuhkan oleh algoritma berkurang seiring dengan pengurangan dimensi
- Metode
  - Menciptakan fitur baru yang dapat merepresentasikan fitur2 lama
  - Memilih atribut2 baru yang merupakan irisan dari atribut lama

TEKNIK-TEKNIK PENGURANGAN DIMENSI (ALJABAR LINIEAR)

- Principle Component Analysis (PCA)
- Singular Value Decomposition
- Supervised and Non-Liniear Techniques

PENGURANGAN DIMENSI: PCA

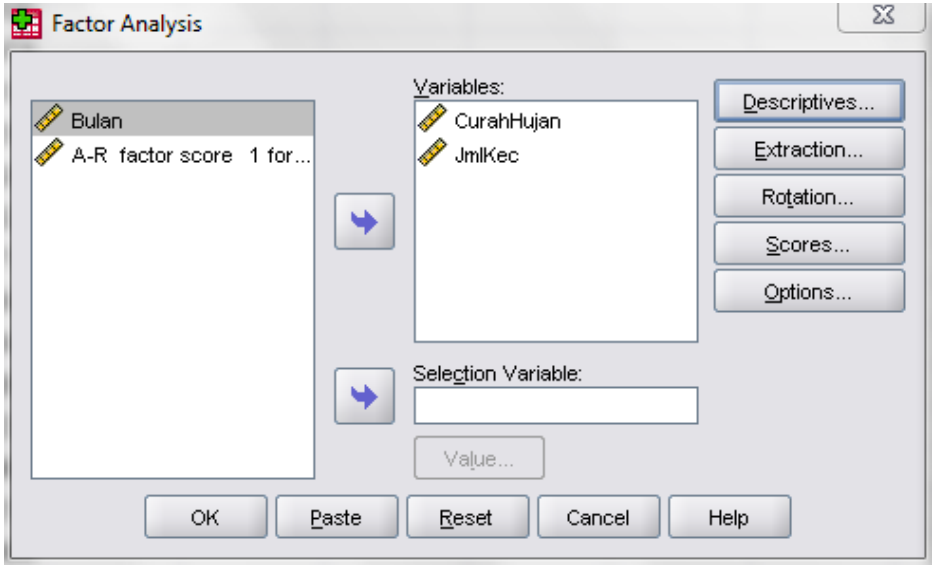
- Temukan nilai eigenvector dari matrix covariance
- Nilai eigenvector mendefinisikan space baru
- Gunakan fitur “factor analysis”

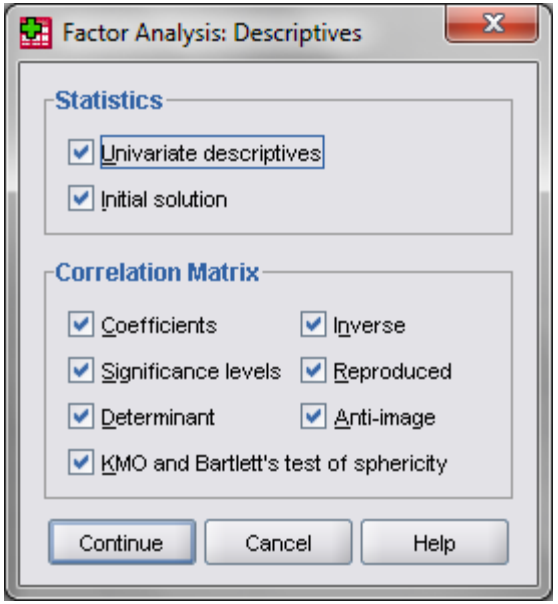


Bulan	JmlKec	CurahHujan	FAC1_1
1.00	18.00	12.00	-0.97609
2.00	13.00	120.00	-0.19166
3.00	22.00	6.00	-0.87059
4.00	1.00	24.00	-1.54756
5.00	24.00	165.00	0.65920
6.00	10.00	44.00	-1.00474
7.00	20.00	174.00	0.58105
8.00	33.00	150.00	0.88290
9.00	44.00	162.00	1.43288
10.00	39.00	91.00	0.58528
11.00	46.00	180.00	1.67710
12.00	7.00	14.00	-1.39842
13.00	6.00	48.00	-1.12848
14.00	17.00	74.00	-0.45085
15.00	17.00	53.00	-0.64232
16.00	43.00	160.00	1.37460

STEP BY STEP DATA REDUCTION PROCESS USING PCA (1)

- Input semua variabel yang akan dianalisis





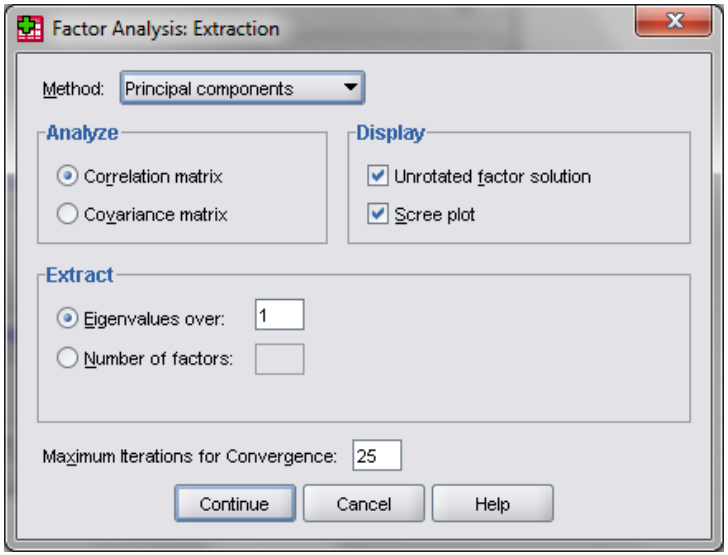
- Klik Descriptive, cek semua analisis.
  - Coefficient → menghitung R-Matrix
  - Sign.level → menghitung signifikansi korelasi antar matrix

STEP BY STEP DATA REDUCTION PROCESS USING PCA (2)

- Determinan → menguji multikolinieritas atau singularitas. Nilai det R-Matrix harus > 0.00001; jika kurang dari itu lihat matrix korelasi dan eliminasi satu atau beberapa variabel berdasarkan masalah
- KMO and Bartlett's test → syarat kecukupan model (syarat cukup model jika nilai KMO > 0.5)

STEP BY STEP DATA REDUCTION PROCESS USING PCA (3)

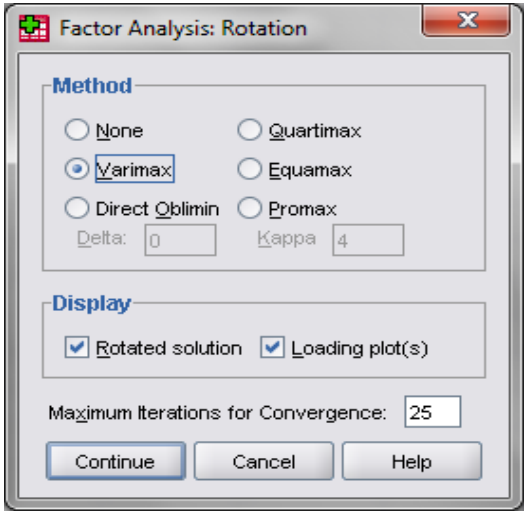
- Klik Extraction
  - Pilih metode sesuai dengan tujuan analisis
  - Set tampilan ke dalam scree plot untuk melihat faktor apa saja yang harus tetap ada di dalam dataset
  - Fitur unrotated solution berguna untuk meningkatkan interpretasi korelasi selama rotasi berlangsung

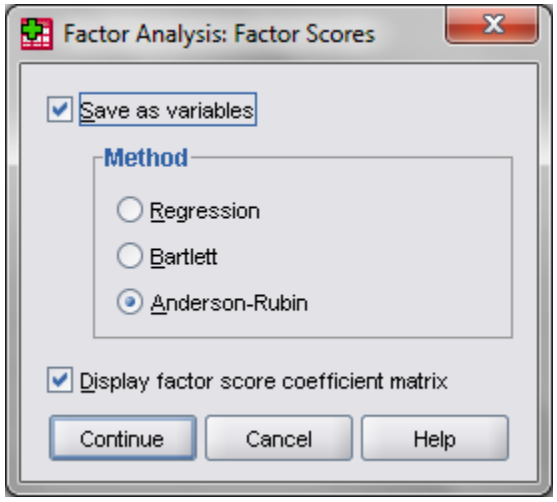


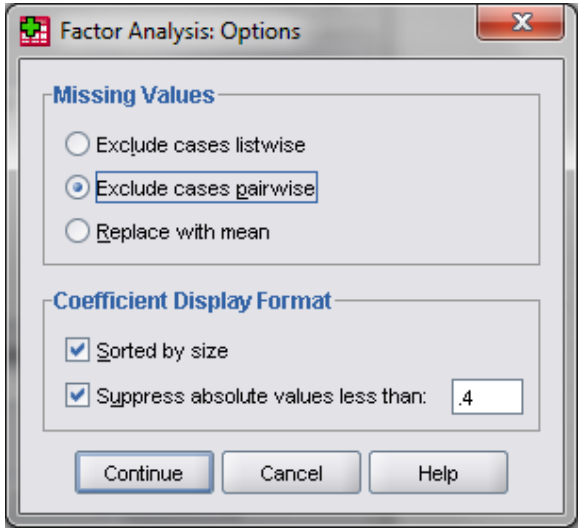
STEP BY STEP DATA REDUCTION PROCESS USING PCA (4)

- Klik rotation
  - Orthogonal rotation (Varimax, quartimax, equamax) → jika Anda yakin bahwa variabel yang Anda analisis seharusnya masuk ke dalam fitur dataset
- Klik scores
  - Pilih metode:
    - regression method → jika mean = 0; var = (estimated factor x true vactor values)^2
    - Bartlett → jika mean = 0; SS = min
    - Anderson-Rubin → mean = 0; std. dev = 1; uncorrelated

STEP BY STEP DATA REDUCTION PROCESS USING PCA (5)







STEP BY STEP DATA REDUCTION PROCESS USING PCA (6)

Correlation Matrix[a]

		CurahHujan	JmlKec
Correlation	CurahHujan	1.000	.618
	JmlKec	.618	1.000
Sig. (1-tailed)	CurahHujan		.001
	JmlKec	.001	

a. Determinant = .618

- Korelasi antara faktor CurahHujan dan JmlKec = 0.618 dengan tingkat signifikansi (P-Value = 0.001); dengan nilai eigenvalue= $\det(A - \lambda I) = 0.618$.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.500
Bartlett's Test of Sphericity	Approx. Chi-Square	10.358
	df	1
	Sig.	.001

STEP BY STEP DATA REDUCTION PROCESS USING PCA (7)

Anti-image Matrices

		CurahHujan	JmlKec
Anti-image Covariance	CurahHujan	.618	-.382
	JmlKec	-.382	.618
Anti-image Correlation	CurahHujan	.500[a]	-.618
	JmlKec	-.618	.500[a]

a. Measures of Sampling Adequacy(MSA)

Communalities

	Initial	Extraction
CurahHujan	1.000	.809
JmlKec	1.000	.809

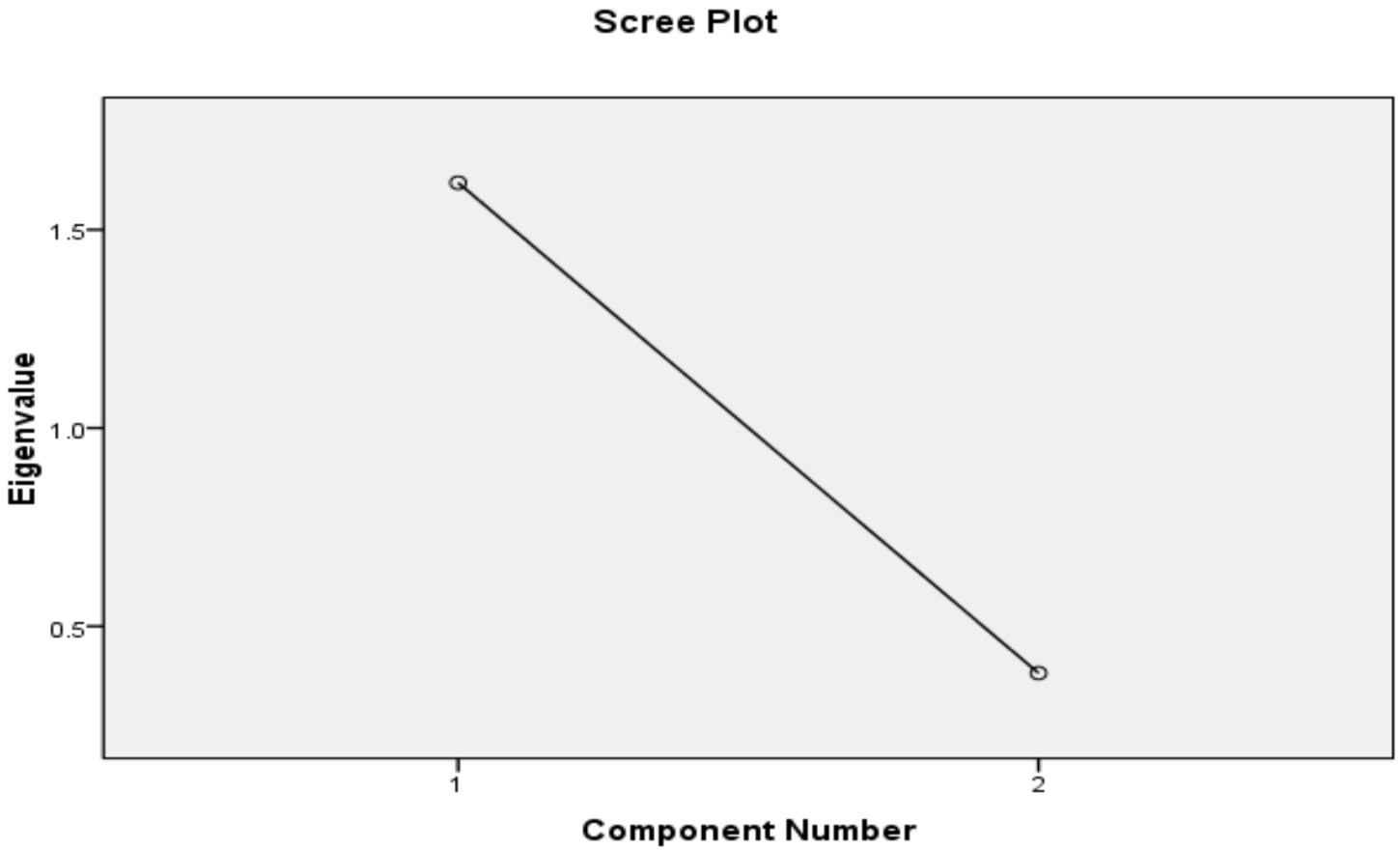
Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.618	80.916	80.916	1.618	80.916	80.916
2	.382	19.084	100.000			

Extraction Method: Principal Component Analysis.

STEP BY STEP DATA REDUCTION PROCESS USING PCA (8)



STEP BY STEP DATA REDUCTION PROCESS USING PCA (9)

Component Matrix[a]

	Component
	1
JmlKec	.900
CurahHujan	.900

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

Component Score Coefficient Matrix

	Component
	1
CurahHujan	.556
JmlKec	.556

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Component Scores.

Reproduced Correlations

		CurahHujan	JmlKec
Reproduced Correlation	CurahHujan	.809[a]	.809
	JmlKec	.809	.809[a]
Residual[b]	CurahHujan		-.191
	JmlKec	-.191	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 1 (100.0%) nonredundant residuals with absolute values greater than 0.05.