Тема 7

Статистические методы снижения размерности признакового пространства

<u>Снижение размерности признакового пространства обусловлено</u>:

- необходимость наглядного представления данных
- стремление к лаконизму исследуемых моделей:

  отбор наиболее информативных показателей
- необходимость существенного сжатия хранимой статистических данных

Инструментарий снижения размерности признакового пространства:

- метод главных компонент
- методы факторного анализа
- многомерное шкалирование
- отбор существенных предикторов в регрессионном анализе
- отбор типообразующих признаков в дискриминантном анализе
- целенаправленное проецирование и отбор типообразующих признаков в кластер-анализе

Факторный анализ - совокупность статистических методов, которые на основе реально существующих связей признаков (или объектов) позволяют выявить *латентные* обобщающие характеристики организационной структуры и механизма развития изучаемых явлений и процессов.

$$Z_j = a_{j1}F_1 + a_{j2}F_2 + \ldots + a_{jn}F_n,$$

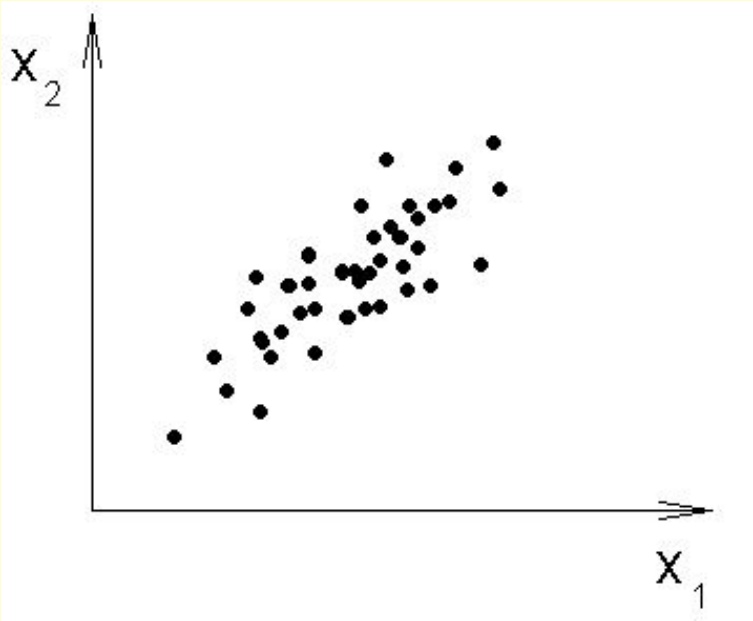
где Z_j – стандартизированные переменные

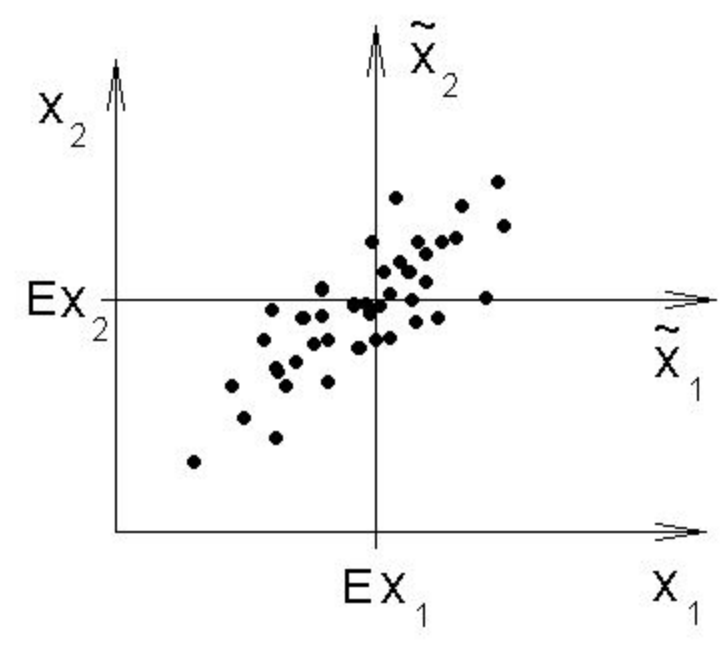
$F_1, F_2, \ldots F_n$ – общие факторы,

$$\mathrm{Var}\, F_1 \geq \mathrm{Var}\, F_2 \geq \ldots \geq \mathrm{Var}\, F_n$$

$a_{j1}, a_{j2}, \ldots, a_{jn}$ – факторные нагрузки

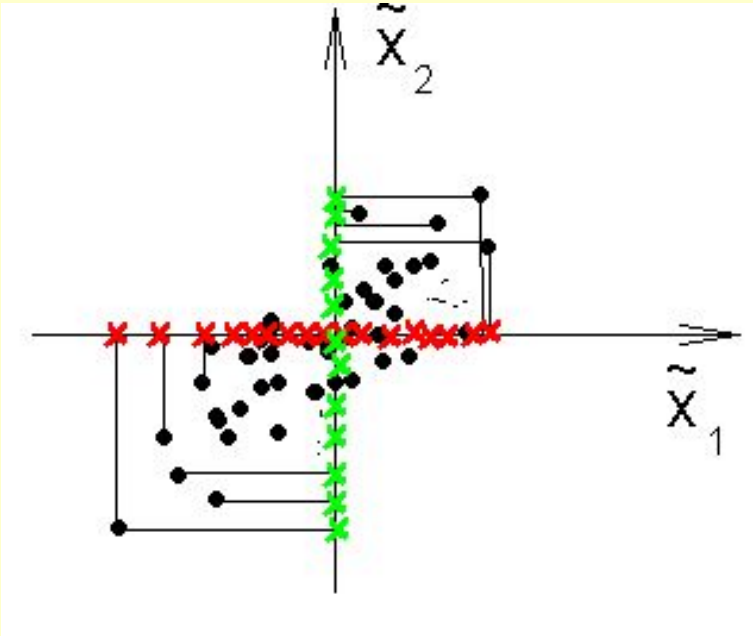
Метод главных компонент

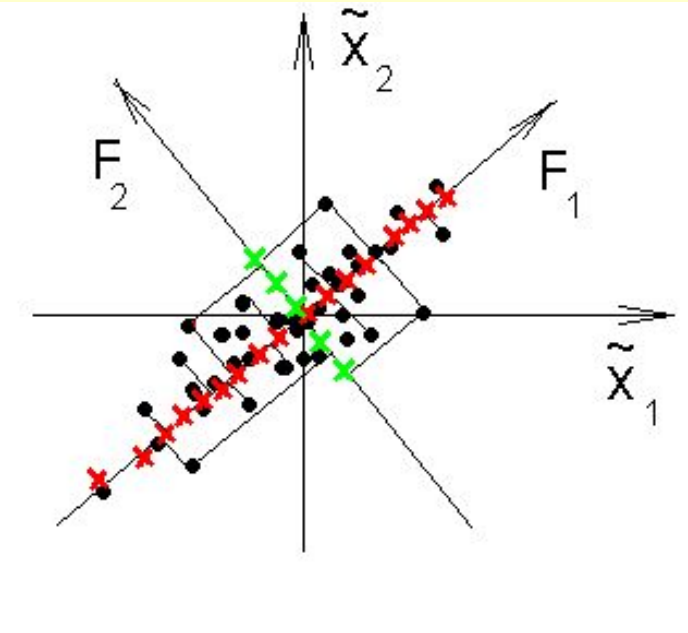




Стандартизация

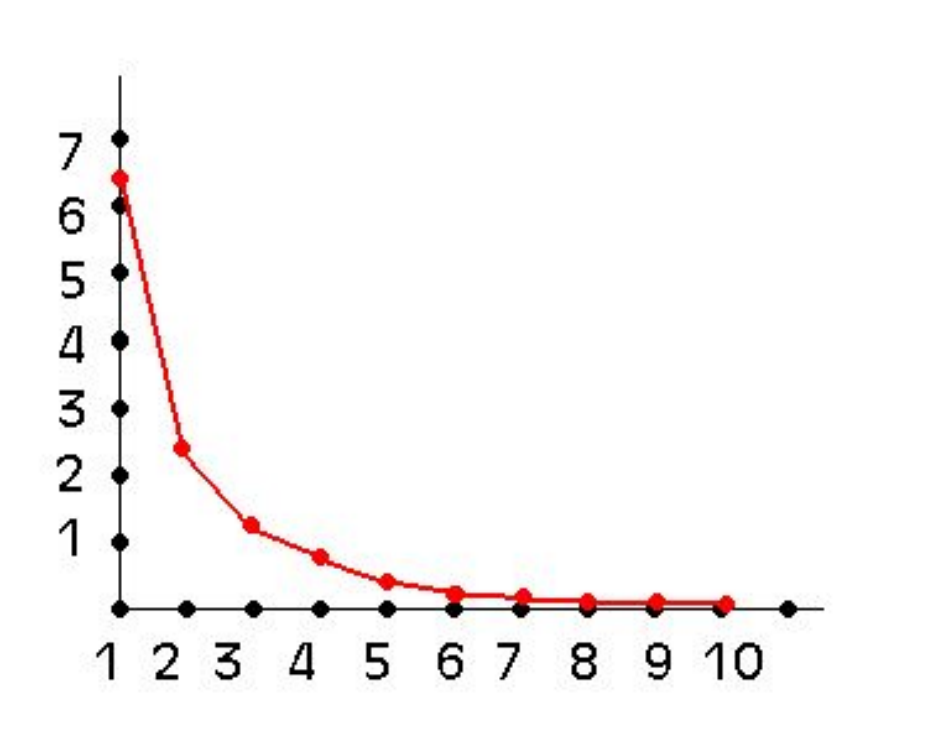
Метод главных компонент





Вращение

Объяснение общей дисперсии



$$\frac{\lambda_1 + \lambda_2 + \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_{10}} = \frac{\mathrm{Var}\ F_1 + \mathrm{Var}\ F_2 + \mathrm{Var}\ F_3}{\text{общая ддисперси}}$$

Взаимосвязи | Классификация | Снижение размерно

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

19 : возраст

	марка	цена	т_сост	возраст	пробег	var	var	var	var	var	var	var
1	Audi	16350	7	3	66000							
2	BMW	14500	8	4	92500							
3	Buick	8950	9	8	92500							
4	Chrysler	8950	6	6	92500							
5	Dodge	8450	5	6	92500							
6	Honda	9850	7	4	118500							
7	Mazda	12650	7	8	58000							
8	Mercedes	17250	8	5	92500							
9	Mitsub.	8950	3	5	136000							
10	Nissan	9850	6	4	150000							
11	Pontiac	8950	7	6	110000							
12	Saab	14950	7	4	101000							
13	Toyota	11700	8	7	57500							
14	VW	8450	5	6	110000							
15	Volvo	13100	8	4	75000							
16												
17												
18												
19												
20												
21												
22												
23												

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Label	Values	Missing	Columns	Align	Measure
1	марка	String		None	None	8	Left	Nominal
2	цена	Numeric		None	None	8	Right	Scale
3	т_сост	Numeric		None	None	8	Right	Scale
4	возраст	Numeric		None	None	8	Right	Scale
5	пробег	Numeric		None	None	8	Right	Scale

Analyze:
- Reports
- Descriptive Statistics
- Tables
- Compare Means
- General Linear Model
- Mixed Models
- Correlate
- Regression
- Loglinear
- Classify
- Data Reduction
  - Factor...
  - Correspondence Analysis...
  - Optimal Scaling...
- Scale
- Nonparametric Tests
- Time Series
- Survival
- Multiple Response
- Missing Value Analysis...
- Complex Samples

Data View | Variable View

Factor | SPSS Processor is ready

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	марка	String	8	0		None	None	8	Left	Nominal
2	цена	Numeric	8	0		None	None	8	Right	Scale
3	т_сост	Numeric	8	0		None	None	8	Right	Scale
4	возраст	Numeric	8	0		None	None	8	Right	Scale
5	пробег	Numeric	8	0		None	None	8	Right	Scale

Factor Analysis

марка

Variables:

цена
т_сост
возраст
пробег

OK
Paste
Reset
Cancel
Help

Selection Variable:

Value...

Descriptives... Extraction... Rotation... Scores... Options...

Data View / Variable View

SPSS Processor is ready

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	марка	String	8	0		None	None	8	Left	Nominal
2	цена	Numeric	8	0		None	None	8	Right	Scale
3	т_сост	Numeric	8	0		None	None	8	Right	Scale
4	возраст	Numeric	8	0		None	None	8	Right	Scale
5	пробег	Numeric	8	0		None	None	8	Right	Scale

Factor Analysis

марка

Variables:
цена
т_сост
возраст
пробег

OK
Paste
Reset
Cancel
Help

Selection Variable:

Value...

Descriptives... Extraction... Rotation... Scores... Options...

Factor Analysis: Descriptives

Statistics
- [ ] Univariate descriptives
- [x] Initial solution

Continue
Cancel
Help

Correlation Matrix
- [x] Coefficients
- [x] Significance levels
- [x] Determinant
- [x] KMO and Bartlett's test of sphericity
- [ ] Inverse
- [ ] Reproduced
- [ ] Anti-image

Data View / Variable View

SPSS Processor is ready

Взаимосвязи | Классификация | **Снижение размерности**

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	марка	String	8	0		None	None	8	Left	Nominal
2	цена	Numeric	8	0		None	None	8	Right	Scale
3	т_сост	Numeric	8	0		None	None	8	Right	Scale
4	возраст	Numeric	8	0		None	None	8	Right	Scale
5	пробег	Numeric	8	0		None	None	8	Right	Scale

Factor Analysis

марка

Variables:
цена
т_сост
возраст
пробег

OK
Paste

Selection Variable

Descriptives... Extraction... Rotation...

Factor Analysis: Extraction

Method: Principal components

Analyze
Correlation matrix
Covariance matrix

Display
Unrotated factor solution
Scree plot

Extract
Eigenvalues over: 1
Number of factors:

Maximum Iterations for Convergence: 25

Continue
Cancel
Help

Data View / Variable View

SPSS Processor is ready

Взаимосвязи | Классификация | **Снижение размерности**

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	марка	String	8	0		None	None	8	Left	Nominal
2	цена	Numeric	8	0		None	None	8	Right	Scale
3	т_сост	Numeric	8	0		None	None	8	Right	Scale
4	возраст	Numeric	8	0		None	None	8	Right	Scale
5	пробег	Numeric	8	0		None	None	8	Right	Scale
6										
7										
23										
24										

Factor Analysis

марка

Variables:
цена
т_сост
возраст
пробег

Selection Variable:

OK
Paste
Reset
Cancel
Help
Value...

Descriptives... Extraction... Rotation... Scores... Options...

Factor Analysis: Extraction

Method: Principal components

Analyze
Correlation matrix
Covariance matrix

Display
Unrotated factor solution
Scree plot

Extract
Eigenvalues over: 1
Number of factors:

Maximum Iterations for Convergence: 25

Continue
Cancel
Help

Data View / Variable View

SPSS Processor is ready

Взаимосвязи | Классификация | **Снижение размерности**

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	марка	String	8	0		None	None	8	Left	Nominal
2	цена	Numeric	8	0		None	None	8	Right	Scale
3	т_сост	Numeric	8	0		None	None	8	Right	Scale
4	возраст	Numeric	8	0		None	None	8	Right	Scale
5	пробег	Numeric	8	0		None	None	8	Right	Scale
6										
7										
23										
24										

Factor Analysis

марка

Variables:
цена
т_сост
возраст
пробег

Selection Variable:

OK
Paste
Reset
Cancel
Help
Value...

Descriptives... Extraction... Rotation... Scores... Options...

Factor Analysis: Rotation

Method
None
Varimax
Direct Oblimin
Quartimax
Equamax
Promax
Delta: 0
Kappa: 4

Display
Rotated solution
Loading plot(s)

Maximum Iterations for Convergence: 25

Continue
Cancel
Help

Data View / Variable View

SPSS Processor is ready

Взаимосвязи | Классификация | **Снижение размерности**

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	марка	String	8	0		None	None	8	Left	Nominal
2	цена	Numeric	8	0		None	None	8	Right	Scale
3	т_сост	Numeric	8	0		None	None	8	Right	Scale
4	возраст	Numeric	8	0		None	None	8	Right	Scale
5	пробег	Numeric	8	0		None	None	8	Right	Scale
6										
7										
23										
24										

Factor Analysis

марка

Variables:
цена
т_сост
возраст
пробег

Selection Variable:

OK
Paste
Reset
Cancel
Help
Value...

Descriptives... Extraction... Rotation... Scores... Options...

Factor Analysis: Factor Scores

Save as variables

Method
Regression
Bartlett
Anderson-Rubin

Continue
Cancel
Help

Display factor score coefficient matrix

Data View / Variable View

SPSS Processor is ready

Взаимосвязи | Классификация | **Снижение размерности**

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	марка	String	8	0		None	None	8	Left	Nominal
2	цена	Numeric	8	0		None	None	8	Right	Scale
3	т_сост	Numeric	8	0		None	None	8	Right	Scale
4	возраст	Numeric	8	0		None	None	8	Right	Scale
5	пробег	Numeric	8	0		None	None	8	Right	Scale
6										
7										
23										
24										

Factor Analysis

марка

Variables:
цена
т_сост
возраст
пробег

Selection Variable:

OK
Paste
Reset
Cancel
Help
Value...

Descriptives... Extraction... Rotation... Scores... Options...

Factor Analysis: Options

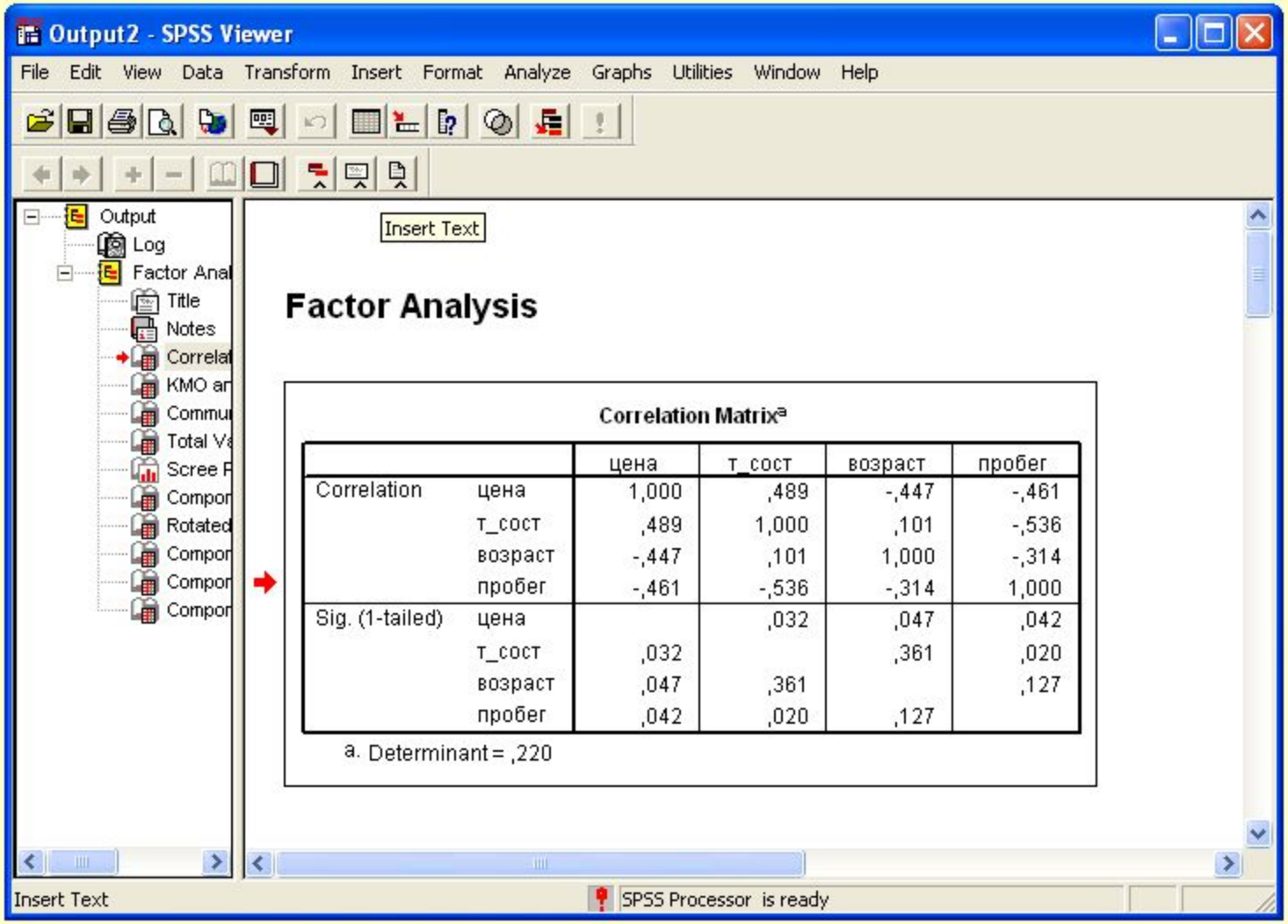
Missing Values
- Exclude cases listwise
- Exclude cases pairwise
- Replace with mean

Continue
Cancel
Help

Coefficient Display Format
- [x] Sorted by size
- [ ] Suppress absolute values less than: .10

Data View / Variable View

SPSS Processor is ready



Factor Analysis

Correlation Matrix[a]

		цена	т_сост	возраст	пробег
Correlation	цена	1,000	,489	-,447	-,461
	т_сост	,489	1,000	,101	-,536
	возраст	-,447	,101	1,000	-,314
	пробег	-,461	-,536	-,314	1,000
Sig. (1-tailed)	цена		,032	,047	,042
	т_сост	,032		,361	,020
	возраст	,047	,361		,127
	пробег	,042	,020	,127	

a. Determinant = ,220

Output2 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Output
Log
Factor Anal
Title
Notes
Correlat
KMO an
Commu
Total Va
Scree P
Compon
Rotated
Compon
Compon
Compon

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,427
Bartlett's Test of Sphericity	Approx. Chi-Square	17,922
	df	6
	Sig.	,006

Communalities

	Initial	Extraction
цена	1,000	,898
т_сост	1,000	,711
возраст	1,000	,946
пробег	1,000	,809

Extraction Method: Principal Component Analysis.

SPSS Processor is ready

Output2 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

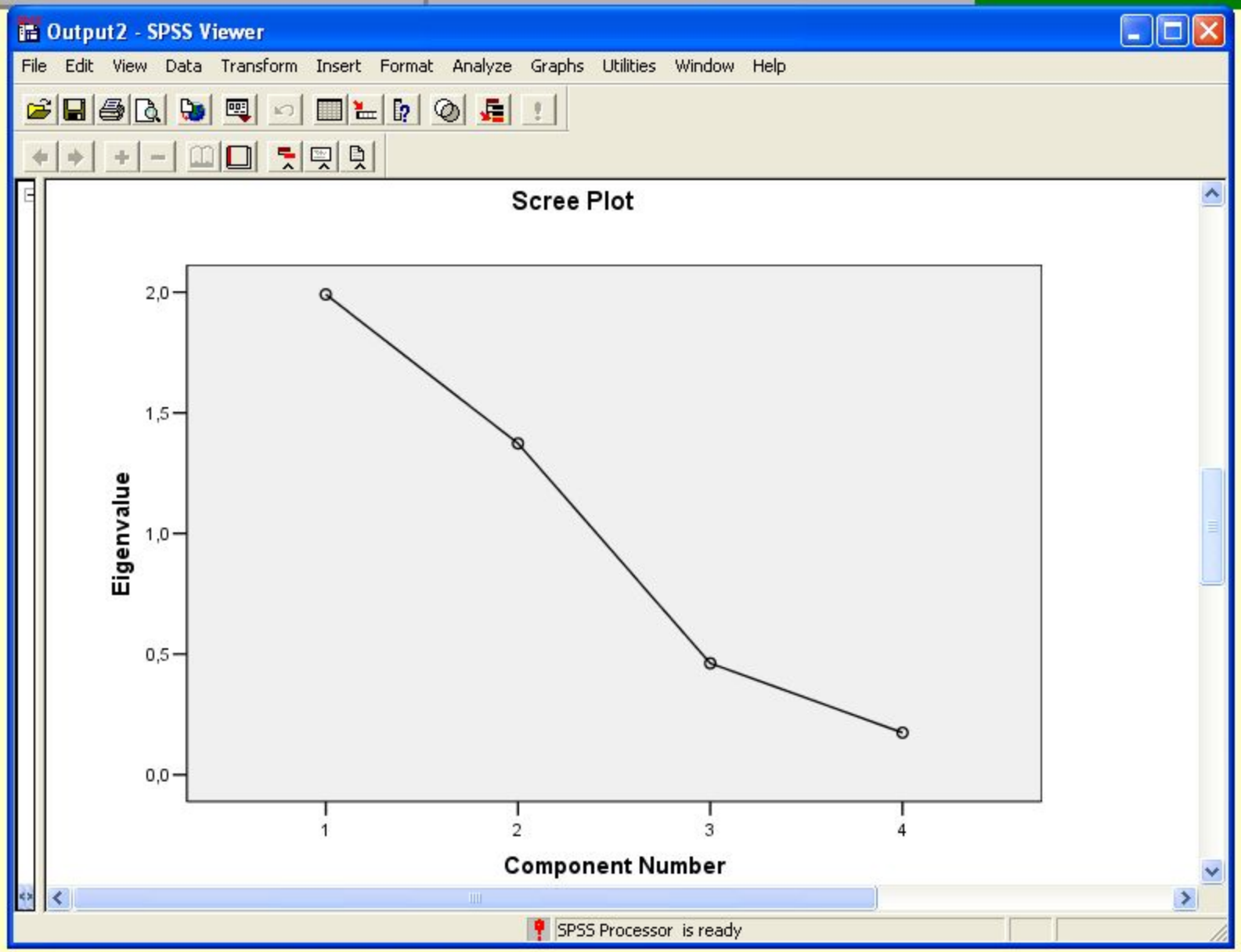
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1,991	49,772	49,772	1,991	49,772	49,772	1,962	49,053	49,053
2	1,373	34,329	84,101	1,373	34,329	84,101	1,402	35,048	84,101
3	,462	11,544	95,645						
4	,174	4,355	100,000						

Extraction Method: Principal Component Analysis.

SPSS Processor is ready H: 172 , W: 346 pt

Взаимосвязи | Классификация | **Снижение размерности**



Output2 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Component Matrix[a]

	Component	
	1	2
т_сост	,832	,138
пробег	-,815	-,381
цена	,797	-,513
возраст	-,017	,972

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Rotated Component Matrix[a]

	Component	
	1	2
пробег	-,878	-,197
т_сост	,842	-,045
возраст	,193	,953
цена	,667	-,673

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

a. Rotation converged in 3 iterations.

SPSS Processor is ready

Взаимосвязи | Классификация | Снижение размерности

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

Component	1	2
1	1,000	,000
2	,000	1,000

Component Score Covariance Matrix

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

	Component	
	1	2
пол	,581	-,034
возраст	,669	,441
т_сост	,800	,034
цена	-,142	,031

Component Score Coefficient Matrix

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

Component	1	2
1	,976	,156
2	-,156	,976

Component Transformation Matrix

Простые методы факторного анализа:

- **однофакторная модель Ч. Спирмена:** выделяется один латентный и один характерный факторы. Для возможно существующих других латентных факторов делается предположение об их незначительности
- **бифакторная модель Г.Хользингера:** допускается влияние на вариацию элементарных признаков не одного, а нескольких латентных факторов (обычно двух) и одного характерного фактора

- **центроидный метод Л. Тэрстоуна:** корреляция между переменными рассматривается как пучок векторов, а латентный фактор геометрически представляется как уравновешивающий вектор, проходящий через центр этого пучка

Аппроксимирующие методы факторного анализа:

- **главных факторов**
- **групповой**
- **минимальных остатков**
- **канонический**
- **распознавания образов**
- **максимального правдоподобия**
- **...**

Основные стадии факторного анализа:

- **Вычисление корреляционной матрицы для всех переменных, участвующих в анализе;** если переменная слабо коррелированна с остальными, следует подумать о ее исключении при следующем запуске программы (учитывая при этом ее общность и нагрузки).
- **Извлечение факторов;** оцениваются нагрузки факторов; выбирается либо метод главных компонент, либо один из методов факторного анализа.
- **Вращение факторов для создания упрощенной структуры;** вращение увеличивает, либо уменьшает нагрузки на каждый фактор; просмотр результатов позволяет уменьшить первоначальное число факторов.
- **Интерпретация факторов**.