

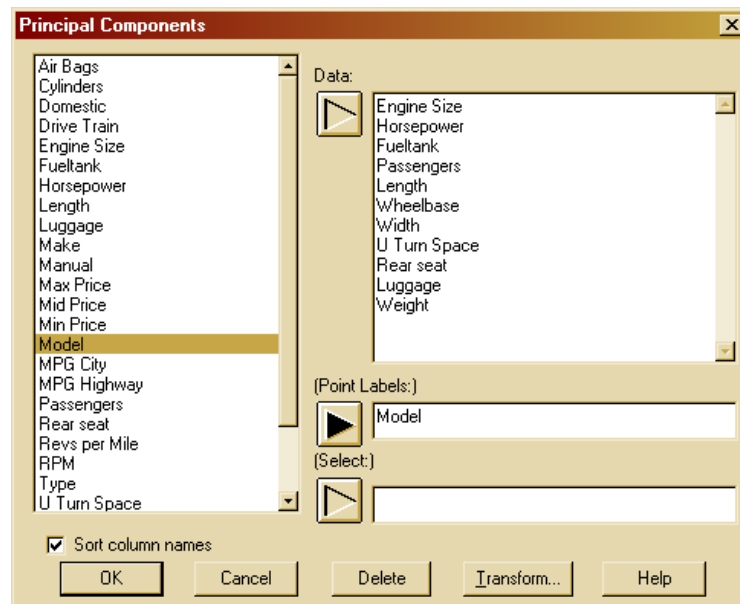
ЛАБОРАТОРНАЯ РАБОТА № 11

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

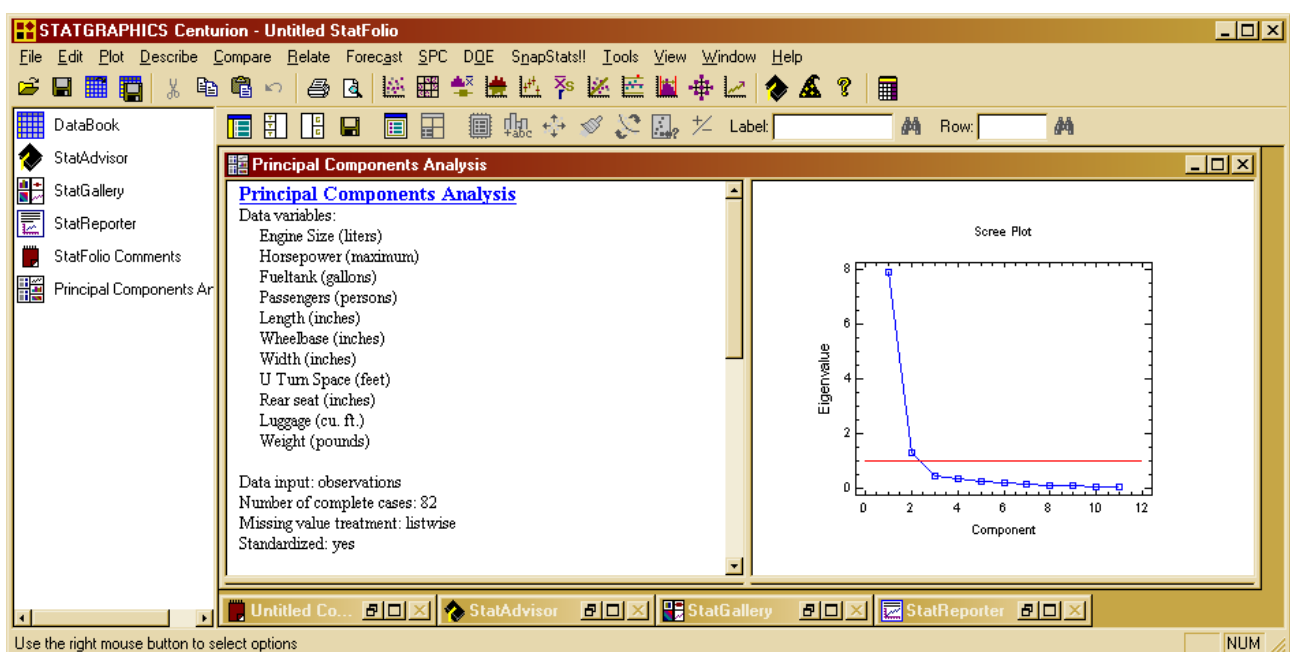
Практически ни одно исследование многомерных данных не обходится без применения метода главных компонент. Это классический метод снижения размерности данных путем определения незначительного числа линейных комбинаций исходных признаков, объясняющих большую часть изменчивости данных в целом, дающий однозначное решение.

Метод главных компонент осуществляет переход к новой системе координат y_1, \dots, y_r от исходных признаков x_1, \dots, x_p . Преобразованные координаты ищутся в системы ортонормированных линейных комбинаций исходных признаков. Геометрически такое преобразование задает некоторый поворот системы координат. Коэффициенты линейных комбинаций являются элементами соответствующих собственных (характеристических) векторов корреляционной матрицы. Первая главная компонента — это линейная комбинация, обладающая наибольшей дисперсией. Геометрически выглядит как новая ось y_1 , ориентированная вдоль направления наибольшей вытянутости эллипсоида рассеивания объектов выборки в исходном пространстве. Вторая главная компонента имеет наибольшую дисперсию среди всех оставшихся линейных преобразований, некоррелированных с первой главной компонентой. Она интерпретируется как направление наибольшей вытянутости эллипсоида рассеивания, перпендикулярное первой главной компоненте и т.д.

Рассмотрим пример, данные к которому есть в *Statgraphics*. Это пример, относящийся к сравнительному оцениванию изделий, характеризующихся одновременно несколькими параметрами. В примере рассматриваются автомобили. Откройте файл с названием *93cars.sf* В строке меню выберите **Describe**, в раскрывшемся меню выберите **Multivariate Methods**, затем **Principal Components**. Раскроется диалоговое окно:



В этом окне вы должны выбрать переменные *Engine Size* (размер двигателя), *Fuel tank* (объем горючего), *Passengers* (кол-во пассажиров), *Length* (длина), *Wheelbase*, *Width* (ширина), *U Turn Space* (радиус разворота), *Rear seat* (клиренс), *Luggage* (багаж), *Weight* (вес), *horsepower* (мощность в лошадиных силах). В поле *Point Label* введите **Model** (это необязательно). Нажмите кнопку ОК, появится окно анализа:



В сводке видно, что исследуется одиннадцать переменных, число объектов составляет 82. Раскройте окно *Principal Components Analysis*.

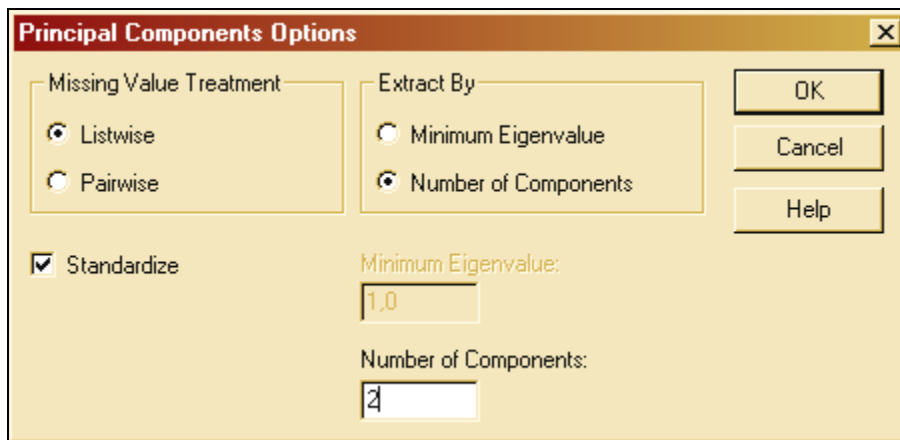
Principal Components Analysis			
Principal Components Analysis			
Component Number	Eigenvalue	Percent of Variance	Cumulative Percentage
1	7,92395	72,036	72,036
2	1,32354	12,032	84,068
3	0,47071	4,279	88,347
4	0,353248	3,211	91,559
5	0,269048	2,446	94,004
6	0,190242	1,729	95,734
7	0,172892	1,572	97,306
8	0,107148	0,974	98,280
9	0,0824071	0,749	99,029
10	0,0694689	0,632	99,660
11	0,0373497	0,340	100,000

The StatAdvisor
This procedure performs a principal components analysis. The purpose c

В таблице даны собственные значения главных компонент (*eigenvalue*); процент дисперсии, приходящийся на каждую выделенную главную компоненту (*Percent of Variance*); накопленный процент дисперсии (*Cumulative Percentage*). По данным этой сводки видно, что первые две компоненты имеют собственные значения, большие единицы, и они описывают 84,07 % дисперсии исходных данных. Третья главная компонента добавляет еще 4,2 % дисперсии. *StatGraphics* рекомендует оставить две главные компоненты.

Если данные по разным переменным обладают величинами разного порядка или используются различные единицы измерения для разных переменных, перед применением метода главных компонент данные стандартизуют, т. е. вычитают среднее и делят на стандартное отклонение. Сумма собственных значений в таблице для стандартизованных данных равна числу компонент. По умолчанию в *StatGraphics* стандартизация включена. Стандартизацию ***Standartize*** нужно отключить, если все данные имеют одну единицу измерения.

Для более детального анализа щелкните правой кнопкой в этом окне, выберите ***Analysis Options***, раскроется следующее окно:



The dialog box 'Principal Components Options' contains the following settings:

- Missing Value Treatment:** ☒ Listwise, ☐ Pairwise
- Extract By:** ☐ Minimum Eigenvalue, ☒ Number of Components
- Standardize:** ☒
- Minimum Eigenvalue:** 1,0
- Number of Components:** 2

Buttons: OK, Cancel, Help

В этом окне можно самому выбрать (если это необходимо), сколько главных компонент следует оставить, следует ли снять стандартизацию, а также выбрать, следует ли использовать объекты с неполными данными. Способ *Listwise* использует только те наблюдения, в которых нет пропущенных данных. *Pairwise* позволяет использовать все наблюдения. Выберите ***Number of Components***, в одноименном поле напишите 2. Нажмите ОК.

Нажмите кнопку ***Tables***, в появившемся окне выберите ***Component Weights*** (компонентные веса). Раскроется сводка ***Principal Components Options***, в которой можно проанализировать главные компоненты.

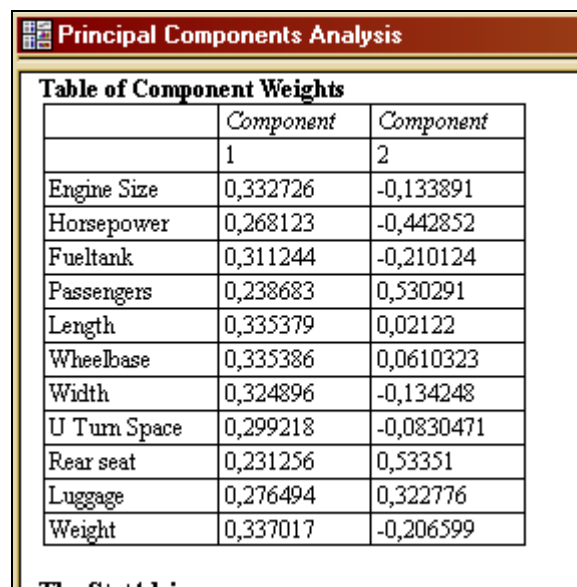


Table of Component Weights		
	Component 1	Component 2
Engine Size	0,332726	-0,133891
Horsepower	0,268123	-0,442852
Fuel tank	0,311244	-0,210124
Passengers	0,238683	0,530291
Length	0,335379	0,02122
Wheelbase	0,335386	0,0610323
Width	0,324896	-0,134248
U Turn Space	0,299218	-0,0830471
Rear seat	0,231256	0,53351
Luggage	0,276494	0,322776
Weight	0,337017	-0,206599

Заметим, что значения весов первой главной компоненты примерно одинаковы. Это означает, что первая компонента – среднее всех значений.

Во второй главной компоненте мощности машины (с отрицательным знаком), противопоставляется сочетание *Rear seat* (клиренс) и *Passengers* (с положительным знаком) т.е. ее «вместимость».

В *StatAdviser* можно прочесть, что уравнение первой главной компоненты имеет вид

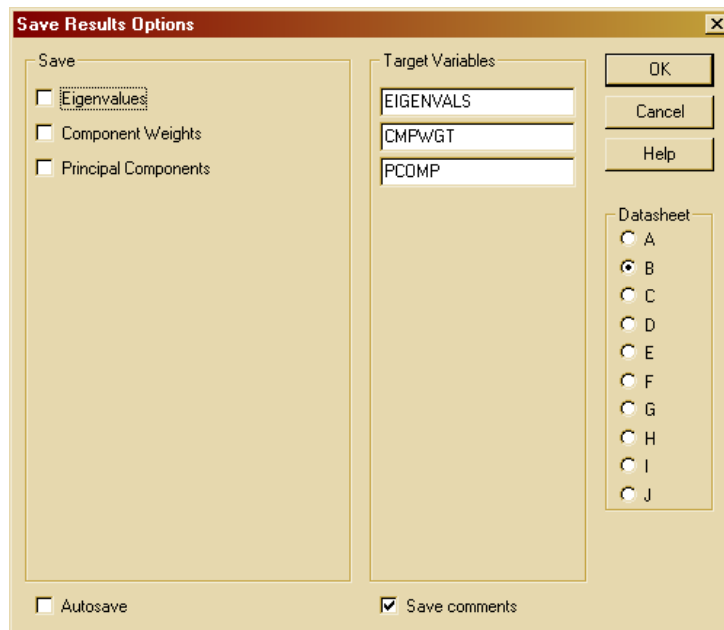
$$0,332726 * Engine \ Size + 0,268123 * Horsepower + 0,311244 * Fueltank + 0,238683 * Passengers + 0,335379 * Length + 0,335386 * Wheelbase + 0,324896 * Width + 0,299218 * U \ Turn \ Space + 0,231256 * Rear \ seat + 0,276494 * Luggage + 0,337017 * Weight$$

Самостоятельно запишите уравнения другой компоненты.

Щелкните снова по кнопке **Tables**, в раскрывшемся окне выберите **Data Table**. Перед Вами появится таблица главных компонент (*Table of Principal Components*). В ней даны значения главных компонент для соответствующих рядов данных.

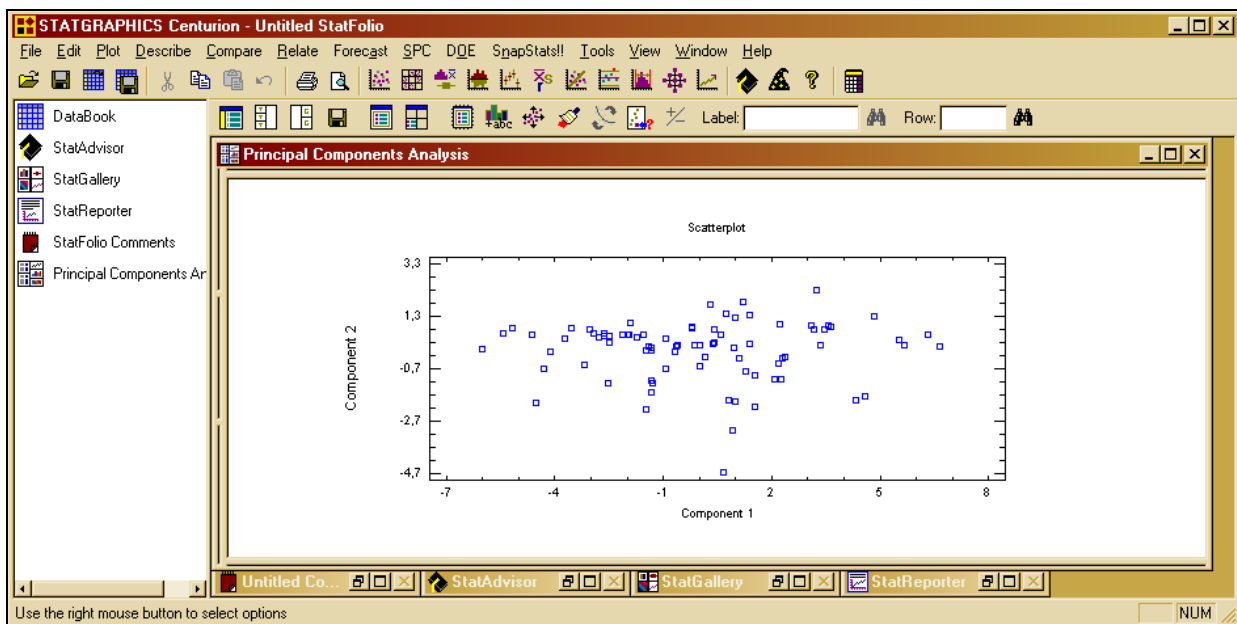
Principal Components Analysis		
Table of Principal Components		
Row	Component 1	Component 2
1	-1,49203	0,00673575
2	2,37408	-0,247278
3	0,165636	-0,261873
4	2,23212	1,01524
5	1,52815	-2,15174
6	0,723227	1,39817
7	3,46805	0,778351
8	6,6603	0,133406
9	2,24466	-1,07736
10	4,83185	1,31608
11	4,5775	-1,74611
12	-1,39935	0,130124
13	-0,590599	0,212676

Если вы хотите сохранить эти значения, нажмите кнопку **Save Results**, перед вами раскроется окно



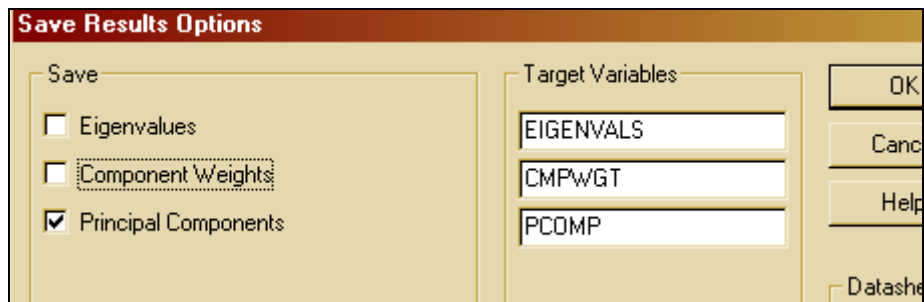
Выберите те данные, которые хотите сохранить, поставьте флажки, если имена столбцов вас не устраивают, то переименуйте. Нажмите кнопку ОК. В том же файле появятся новые столбцы с данными.

Займемся теперь графическим представлением метода главных компонент. Иногда возникает потребность проанализировать графически двумерное соотношение компонент, допустим первой и второй. Щелкните по кнопке **Graphs**, выберите в окне *2D Scatterplots*. Раскроется следующее окно:

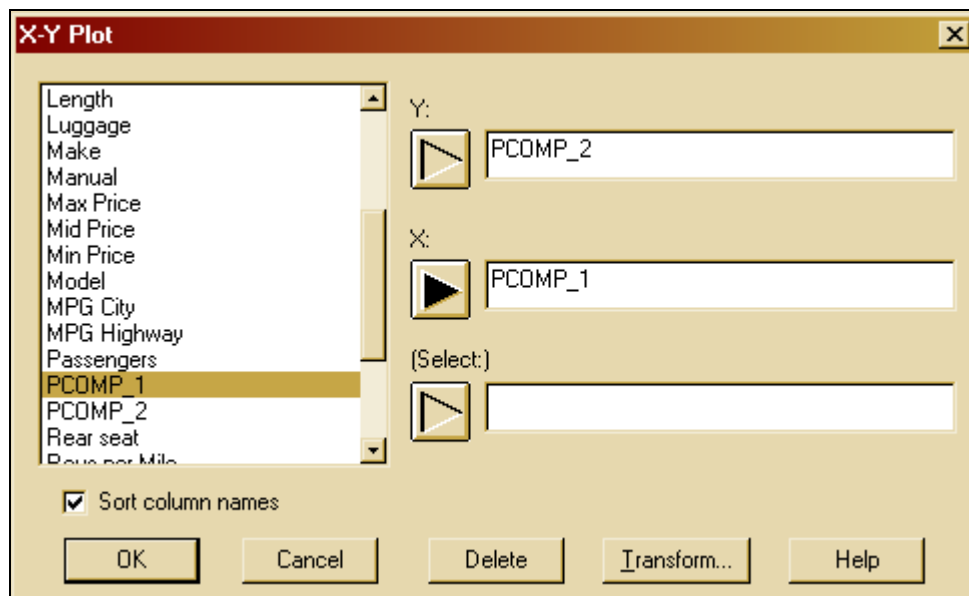


Можно видеть, что одна из машин имеет очень маленькую вторую компоненту. Если щелкнуть по этой точке, появится подсказка, что это 28 наблюдение (*Dodge Stealth*).

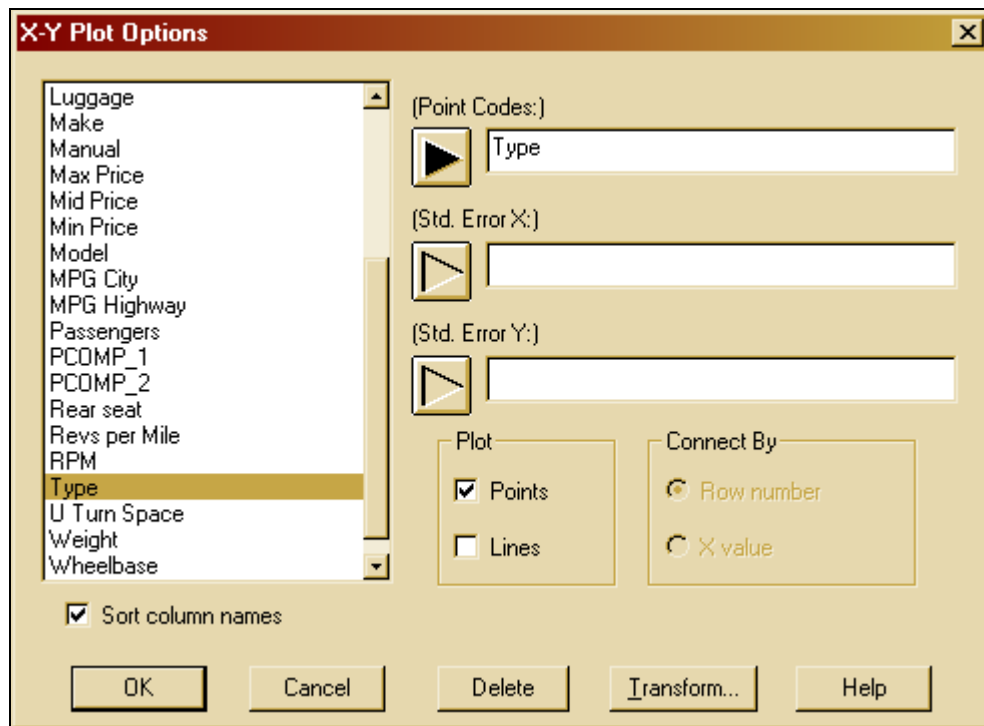
Сохраните главные компоненты, нажав кнопку Сохранить и выбрав *Principal Components*.



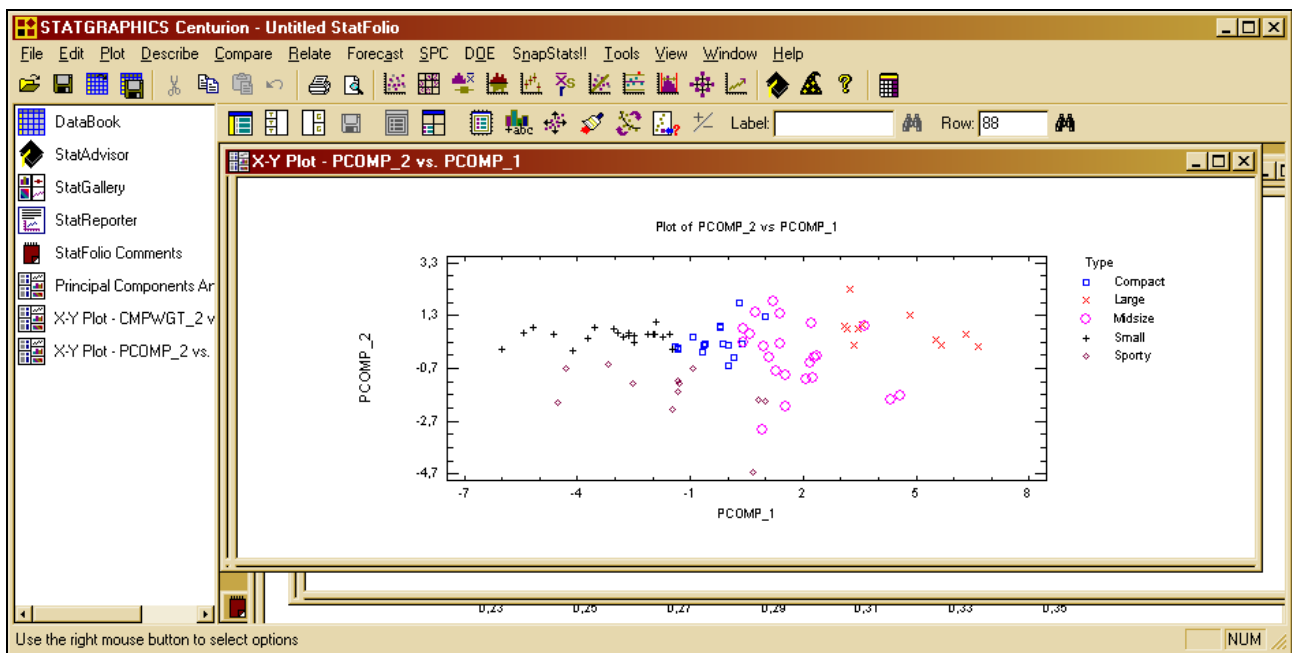
Нажмите кнопку  *X-Y Scatterplot*.



Нажмите кнопку ОК. Затем щелкните правой кнопкой, выберите *Pane Options*.
В поле *Point Codes* введите *Type*.



Нажмите ОК, раскроется окно:



В этом окне можно видеть, что вторая компонента отделяет одни типы машин от других.

Таким образом, произведенный анализ данных с помощью метода главных компонент позволяет получить более «объемное» видение современного автомобильного рынка, что может способствовать лучшей ориентации как потребителей этой продукции, так и производителей с позиции оценки существующих тенденций.

Также сейчас можно вспомнить методы кластерного анализа, воспользоваться им уже для главных компонент.

Еще одна возможность графического анализа главных компонент — *2D Bip-plot*. Рассмотрите этот график самостоятельно.

С помощью регрессионных методов попытайтесь установить, от чего зависит цена машины, результаты покажите преподавателю.

Попробуем теперь рассмотреть третью компоненту. Щелкните правой кнопкой, выберите *Analysis Options*, укажите число компонент – 3. Получим:

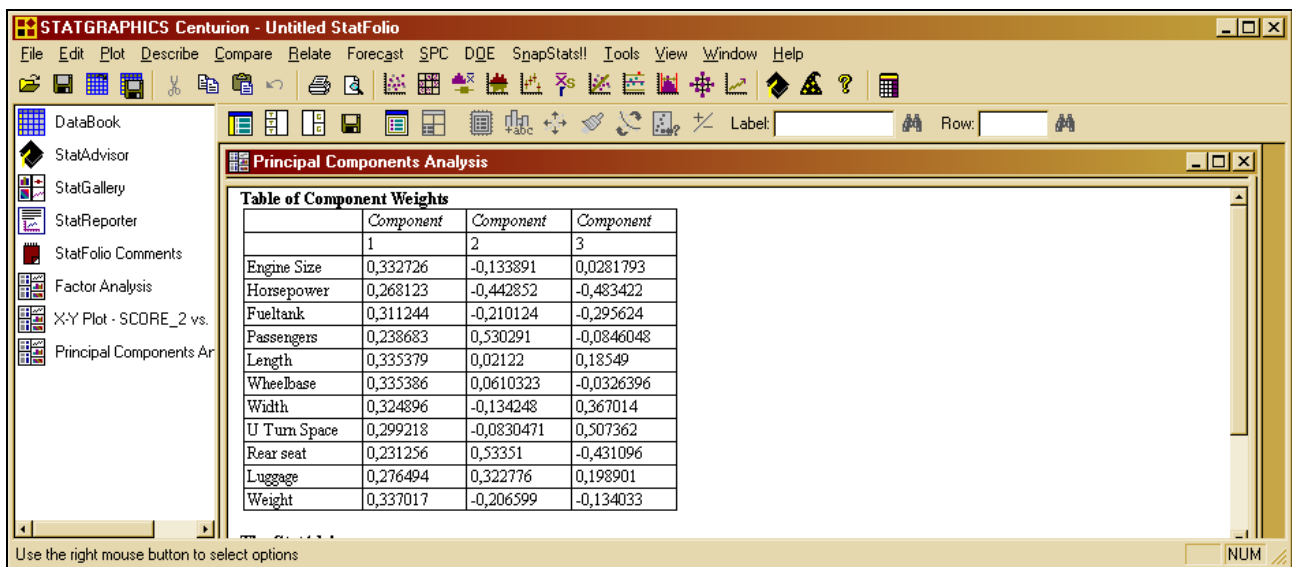


Table of Component Weights

	Component 1	Component 2	Component 3
Engine Size	0,332726	-0,133891	0,0281793
Horsepower	0,268123	-0,442852	-0,483422
Fuel tank	0,311244	-0,210124	-0,295624
Passengers	0,238683	0,530291	-0,0846048
Length	0,335379	0,02122	0,18549
Wheelbase	0,335386	0,0610323	-0,0326396
Width	0,324896	-0,134248	0,367014
U Turn Space	0,299218	-0,0830471	0,507362
Rear seat	0,231256	0,53351	-0,431096
Luggage	0,276494	0,322776	0,198901
Weight	0,337017	-0,206599	-0,134033

Видно, что третья главная компонента определяется разницей между *U Turn Spase* (положительная) и *Horsepower* и *Rear seat* (отрицательные).

ФАКТОРНЫЙ АНАЛИЗ

Факторный анализ основан не на дисперсионном критерии, а ориентирован на объяснение корреляций, имеющихсся между признаками. Поэтому он применяется в более сложных случаях совместного проявления в структуре экспериментальных данных действия латентных факторов.

Задача факторного анализа не имеет однозначного решения. Представление корреляционной матрицы факторами (как говорят, ее факторизацию) можно произвести бесконечно большим числом способов. Известно много методов факторного анализа. Если удалось произвести факторизацию корреляционной матрицы с помощью некоторой матрицы нагрузок F , то любое линейное преоб-

разование F (ортогональное вращение) приведет к такой же факторизации. Поэтому нередко в одном и том же пакете программ анализа данных реализовано сразу несколько версий таких методов. Выбор среди группы методов наилучшего производится в основном с точки зрения вычислительных удобств.

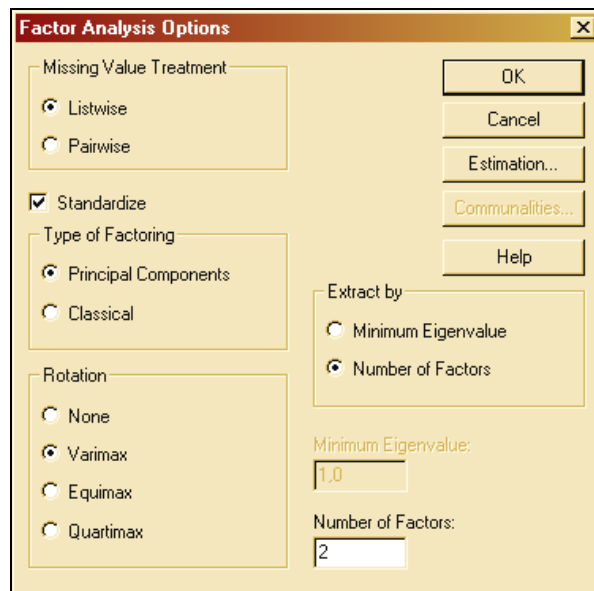
В *Statgraphics* реализовано три метода вращения факторов: варимакс, кватримакс и эквимакс. Вращение методом варимакс ставит целью упростить столбцы факторной матрицы, сводя все значения к величинам, близким к 1 или 0. Вращение методом кватримакс ставит целью аналогичное упрощение только по отношению к строкам факторной матрицы. Эквимакс занимает промежуточное положение — при вращении факторов по этому методу одновременно делается попытка упростить и столбцы, и строки.

Факторный анализ широко применяется в экономике, социологии, медицине, психологии для выявления скрытых закономерностей в данных. Разберем этот метод на этом же примере (файл *93cars.sf*).

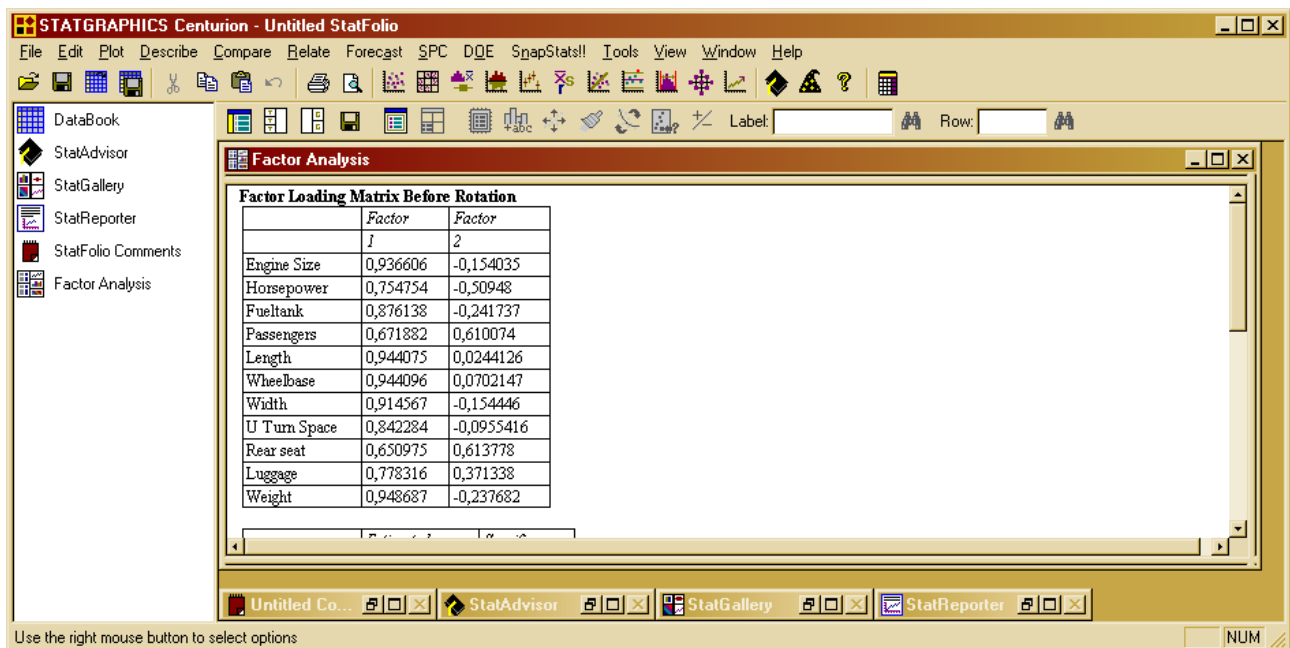
Откройте файл *93cars.sf*. В строке меню выберите **Describe**, в раскрывшемся меню выберите **Multivariate Methods**, затем выберите **Factor Analysis**. Перед вами раскроется окно **Factor Analysis**. В поле **Data** введите переменные (*accel, weight, cylinders, displace, horsepower, year, mpg*), нажмите кнопку ОК. Перед вами раскроется окно с первичной сводкой факторного анализа.

Внимательно рассмотрите полученный результат, прочитайте *StatAdvisor*. Там написано, что цель анализа — получить небольшое количество факторов, которые вбирают в себя большую часть общей изменчивости наблюдаемых данных, а потому передают большую часть информации, заключенной в первоначальных наблюдениях. У вас на первые три фактора приходится 91,52 % дисперсии, причем на первый — 64,68 %, на второй — 14,91 %, на третий — 11,92, на четвертый же всего 5,02 %.

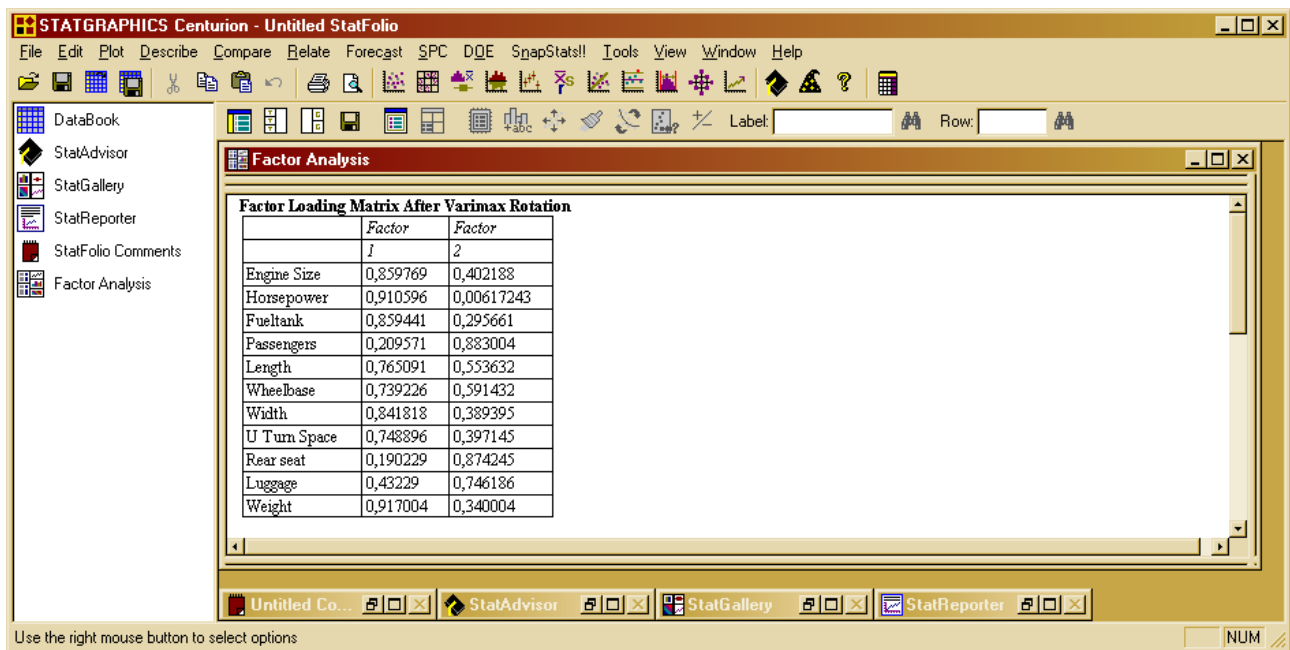
Получим новую сводку данных факторного анализа. Щелкните правой кнопкой по сводке, выберите **Analysis Options**, раскроется окно **Factor Analysis Options**:



Выберите *Number of Factors* (количество факторов), в соответствующем поле поставьте 2, нажмите ОК. Появится новая сводка факторного анализа. Нажмите кнопку *Tabs*, выберите в окне *Extraction Statistics*:

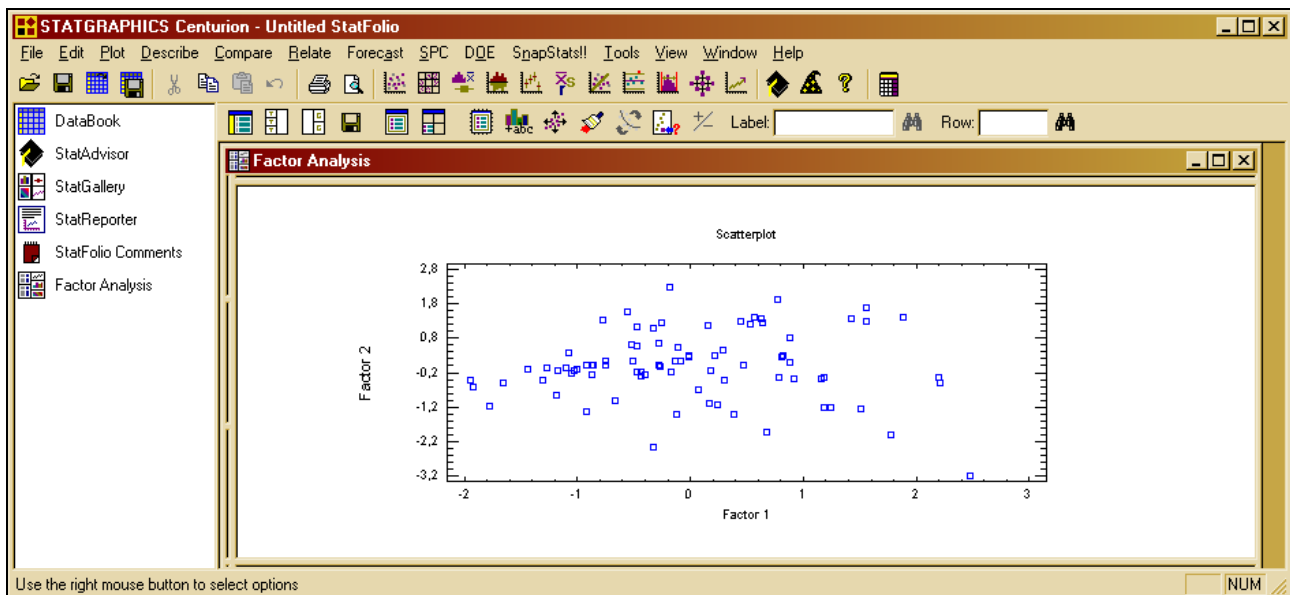


Раскрылось окно с результатами факторизации до вращения факторов. Обычно факторы, полученные методом главных компонент, не поддаются достаточно наглядной интерпретации, поэтому следующим этапом факторного анализа служит преобразование (вращение) факторов таким образом, чтобы облегчить их интерпретацию. Щелкните по кнопке *Tables*, выберите *Rotation Statistics* (нагрузки после вращения), раскроется окно:



Из анализа видно, что из первого фактора «ушли» *Passengers*, *Rear seat*, *Luggage*. Зато они стали определяющими во втором факторе. Получается, что второй фактор выделяет большие «семейные» машины от малых.

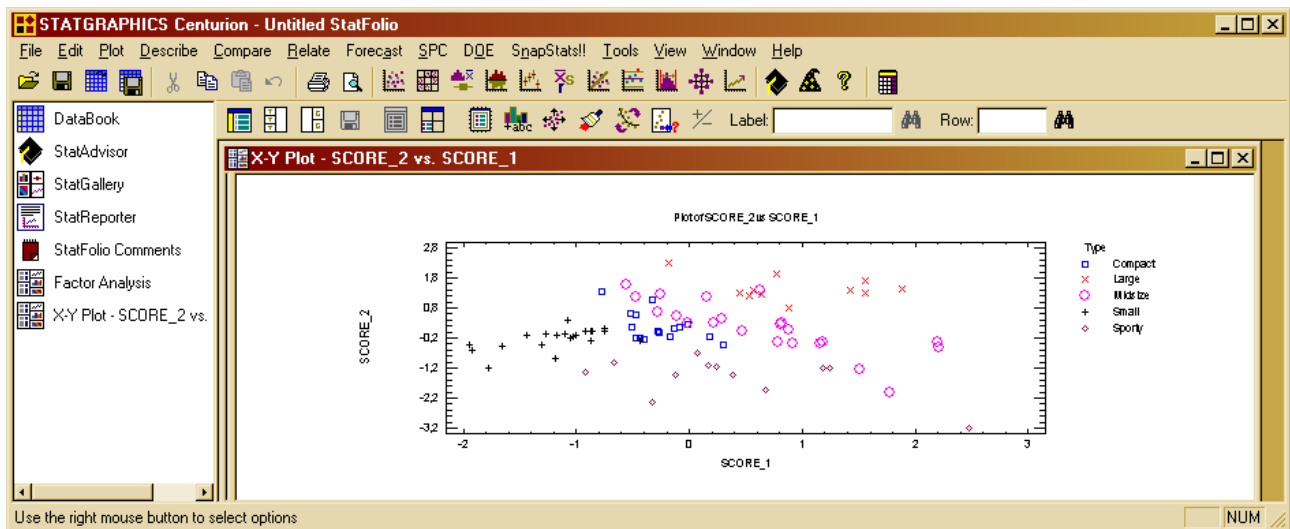
Займемся графическим анализом результатов. Щелкните по кнопке **Graphs**, выберите *2D Scatterplot* (двумерная диаграмма рассеивания).



Щелкая по наиболее далеко находящимся точкам, можно посмотреть, какие машины наиболее отличаются друг от друга.

Можно, как и в методе главных компонент, посмотреть, как отличаются машины по типу. Сохраните **Factor Score**, с помощью кнопки *X-Y Scatterplot*

постройте график, щелкните правой кнопкой, выберите *Pane Options*, введите *Type*.



Попробуйте ввести третью компоненту и проделать весь анализ.

ЗАДАНИЕ 1. Условия жизни населения 10 стран характеризуются тремя показателями: $x^{(1)}$ – оценка ВВП по паритету покупательской способности в 1994 г. на душу населения (в % к США); $x^{(2)}$ – расходы на здравоохранение (в % от ВВП); $x^{(3)}$ – численность врачей на 10000 населения, значения которых приводится в таблице:

№ п/п	Страна	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$
1	Россия	20,4	3,2	44,5
2	Австралия	71,4	8,5	32,5
3	Австрия	78,7	9,2	33,9
4	Азербайджан	12,1	3,3	38,8
5	Армения	10,9	3,2	34,4
6	Белоруссия	20,4	5,4	43,6
7	Бельгия	79,7	8,9	41,0
8	Болгария	17,3	5,4	36,4
9	Великобритания	69,7	7,1	17,9
10	Венгрия	24,5	6,0	32,1

1. По трем показателям ($x^{(1)}, x^{(2)}, x^{(3)}$) выделить главные компоненты и дать им содержательную интерпретацию. Графически представить страны в пространстве двух главных компонент.
2. Провести кластерный анализ по исходным переменным и по главным переменным и сравнить результаты. Как вы можете объяснить различия в результатах?
3. Примените факторный анализ.

Задание 2. В файле *SobakiVolki.sf* указаны размеры челюстей и зубов 30 собак (№ 1-30) и 12 волков (№ 31-42). Смысл переменных можно посмотреть в комментариях к ним в файле. Найти и проинтерпретировать главные компоненты для данного примера.

Задание 3. В файле *Razmer.sf* представлены данные, которые используются модельерами при конструировании одежды: рост, длина рук, длина предплечий, длина ног, вес, окружность бедер, окружность груди, ширина груди. Данные приведены в таблице

№ пп	Рост	Размах рук	Длина предпл	Длина ног	Окр. бедер	Окр. груди	Шири- на гру-	F1	F2
1.	163	158	25	98	92	84	75		
2.	180	172	23	115	88	89	80		
3.	153	143	18	90	95	93	73		
4.	161	155	19	95	98	100	85		
5.	173	162	20	110	90	92	80		
6.	164	158	21	110	99	88	75		
7.	157	139	24	100	84	91	70		
8.	165	149	19	107	98	90	75		
9.	171	165	23	110	90	86	70		
10.	162	154	21	97	86	82	68		
11.	159	150	22	99	90	85	69		
12.	168	167	25	105	88	89	70		

Методами кластерного анализа разбейте эти 12 человек на три группы, постройте дендрограмму. Посмотрите полученные кластеры на двумерных графиках, попробуйте проинтерпретировать их.

Проведите компонентный анализ, выделите две главные компоненты $F1$ и $F2$, проинтерпретируйте их. Вычислите компоненты $F1$ и $F2$ для каждого из 12 объектов и занесите их в таблицу.

Снова проведите кластерный анализ для признаков $F1$ и $F2$. Будут ли совпадать полученные кластеры с первоначальными. Разбейте всех людей на 3 кластера и посмотрите результаты на плоскости $F1, F2$.

Примените факторный анализ.

ВОПРОСЫ

1. Объясните, что характеризует коэффициент факторной нагрузки a_{jl} ?
2. Объясните, что характеризует квадрат коэффициента факторной нагрузки a_{jl}^2 ?
3. Как связаны собственные значения корреляционной матрицы с коэффициентами факторной нагрузки?
4. Как выбираются векторы главных компонент в k -мерном пространстве?
5. Каким требованиям должны удовлетворять «новые» показатели (факторы), полученные методом главных компонент?