

## Постановка задачи и описание данных:

Задача работы: Кластеризация выборки.

Описание данных: Целью данного анализа является разбиение автомобилей и их владельцев на классы, каждый из которых соответствует определенной рискованной группе. Наблюдения, попавшие в одну группу, характеризуются одинаковой вероятностью наступления страхового случая, которая впоследствии оценивается страховщиком.

Использование кластер-анализа для решения данной задачи наиболее эффективно. В общем случае кластер-анализ предназначен для объединения некоторых объектов в классы (кластеры) таким образом, чтобы в один класс попадали максимально схожие, а объекты различных классов максимально отличались друг от друга. Количественный показатель сходства рассчитывается заданным способом на основании данных, характеризующих объекты.

## Иерархический алгоритм

Задачей кластерного анализа является классификация многофакторной выборки. Выборка представляется в виде набора векторов  $\{X_i\}_{i=1}^N$  размерности  $m$ , где  $m$  - число факторов. Близость двух произвольных векторов в выборке характеризуется функционалом  $\rho: \mathbb{R}^2 \rightarrow \mathbb{R}$ , называемым метрикой. В качестве метрики выбирают обычно евклидово расстояние (для выборок, чьи показатели могут принимать действительные значения), либо расстояние Хэмминга (для дискретных показателей)

Евклидовым расстоянием (евклидовой метрикой) между векторами  $X_i$  и  $X_j$  называется величина

$$\rho(X_i, X_j) = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \dots + (x_{mi} - x_{mj})^2}$$

Эта функция удовлетворяет всем аксиомам метрики, а именно:

1.  $\rho(X_i, X_j) \geq 0 \quad \forall X_i, X_j \in \mathbb{R}^m$
2.  $\rho(X_i, X_j) = 0 \Leftrightarrow i = j$
3.  $\rho(X_i, X_j) = \rho(X_j, X_i)$
4.  $\rho(X_i, X_k) \leq \rho(X_i, X_j) + \rho(X_j, X_k) \quad \forall X_i, X_j, X_k \in \mathbb{R}^m$

Не все факторы могут иметь одинаковую размерность. Для того, чтобы исключить неоднородность влияния факторов разной размерности, могут производиться следующие действия:

1. Приведение единиц измерения факторов (если возможно)
2. Нормирование величин

$$X_j^H = (x_{1j}^H, x_{2j}^H, \dots, x_{mj}^H), \quad j = \overline{1, N}$$

$$x_{ij}^H = \frac{x_{ij} - \bar{x}_{iB}}{\sigma_{jB}}, \quad i = \overline{1, m}$$

### 3. Использование весовых коэффициентов

$$\varrho(X_i, X_j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + \dots + w_m(x_{im} - x_{jm})^2}$$

$$\sum_{k=1}^m w_k = 1$$

## 1. Сгруппируем по признакам совокупности данных

	1 цена авто	2 возраст водителя	3 опыт водителя	4 возраст авто
acura	0,545	25	5	10
audi	0,875	23	1	2
bmw	0,432	30	10	15
volvo	0,233	49	5	12
toyota	0,566	71	40	7
saab	0,635	28	9	2
corvette	1,269	33	7	8
chrysler	0,677	56	2	5
dodge	0,705	22	5	3
eagle	0,63	38	4	10
mazda	0,128	26	5	12
mersedes	1,125	57	15	20
porsche	2,958	31	11	1
nissan	0,502	20	2	2
olds	0,724	19	1	1

## 2. Масштаб измерений. Расстояние между элементами

Все кластерные алгоритмы нуждаются в оценках расстояний между кластерами или объектами, и ясно, что при вычислении расстояния необходимо задать масштаб измерений.

Поскольку различные измерения используют абсолютно различные типы шкал, данные необходимо стандартизовать

Каждая переменная будет иметь среднее 0 и стандартное отклонение 1.

Так выглядит таблица со стандартизированными переменными:

	1 цена авто	2 возраст водителя	3 опыт водителя	4 возраст авто
acura	-0,368794269	-0,645621347	-0,323950889	0,461550654
audi	0,0535869508	-0,772213767	-0,737505216	-0,923101308
bmw	-0,513427839	-0,329140294	0,192992019	1,32695813
volvo	-0,768136514	0,873487704	-0,323950889	0,807713644
toyota	-0,341915465	2,26600433	3,29464947	-0,0576938317
saab	-0,253599391	-0,455732715	0,0896034374	-0,923101308
corvette	0,557884529	-0,139251663	-0,117173726	0,115387663
chrysler	-0,199841781	1,31656118	-0,634116634	-0,403856822
dodge	-0,164003375	-0,835509978	-0,323950889	-0,750019813
eagle	-0,259999107	0,177229389	-0,427339471	0,461550654
mazda	-0,902530539	-0,582325136	-0,323950889	0,807713644
mersedes	0,373572724	1,37985739	0,709934927	2,19236561
porsche	3,35072036	-0,265844084	0,296380601	-1,0961828
nissan	-0,423831822	-0,962102399	-0,634116634	-0,923101308
olds	-0,139684456	-1,02539861	-0,737505216	-1,0961828

## Иерархический алгоритм

Иерархический алгоритм заключается в нахождении такого разбиения на кластеры, при котором сумма квадратов попарных расстояний между векторами внутри каждого кластера будет минимальной совокупности. Опишем данный алгоритм:

1. На первом шаге векторы разбиваются на N кластеров, по одному на каждый кластер
2. Строится матрица расстояний

$$\begin{pmatrix} 0 & \varrho(X_1, X_2) & \dots & \varrho(X_1, X_N) \\ \varrho(X_1, X_2) & 0 & \dots & \varrho(X_2, X_N) \\ \vdots & & \ddots & \vdots \\ \varrho(X_1, X_N) & \varrho(X_2, X_N) & \dots & 0 \end{pmatrix},$$

на главной диагонали которой строят нули в силу аксиомы 2.

3. Выбирается наименьший ненулевой элемент матрицы. Соответствующие векторы  $X_i$  и  $X_j$  объединяются к один кластер
4. Строится матрица расстояний между кластерами

$$\begin{pmatrix} 0 & \varrho(S_1, S_2) & \dots & \varrho(S_1, S_{N-1}) \\ \varrho(S_1, S_2) & 0 & \dots & \varrho(S_2, S_{N-1}) \\ \vdots & & \ddots & \vdots \\ \varrho(S_1, S_{N-1}) & \varrho(S_2, S_{N-1}) & \dots & 0 \end{pmatrix}$$

Расстояние между кластерами может вычисляться одним из способов:

1. По принципу ближайшего соседа (метод одиночной связи)

$$\varrho(S_l, S_k) = \min_{X_i \in S_l, X_j \in S_k} \varrho(X_i, X_j)$$

2. По принципу самого дальнего соседа (метод полной связи)

$$\varrho(S_l, S_k) = \max_{X_i \in S_l, X_j \in S_k} \varrho(X_i, X_j)$$

3. По центру тяжести (центроидный метод)

Центром кластера  $S_l$  называется вектор

$$\bar{X}_l = \begin{pmatrix} \frac{1}{n_l} \sum_{X_i \in S_l} x_{1i} \\ \frac{1}{n_l} \sum_{X_i \in S_l} x_{2i} \\ \vdots \\ \frac{1}{n_l} \sum_{X_i \in S_l} x_{mi} \end{pmatrix},$$

, где  $n_l$  - размер l-ого кластера

4. Другие методы (метод средней связи, метод Уорда)

5) Выбирается минимальный ненулевой элемент матрицы. Соответствующие кластеры объединяются

б) Алгоритм продолжается до тех пор, пока минимальное расстояние относительно предыдущих шагов изменится сильно (резкий скачок), либо на усмотрение исследователя

### 3. Иерархическая классификация

Назначение этого [алгоритма](#) состоит в объединении объектов (например, животных) в достаточно большие кластеры, используя некоторую меру сходства или расстояние между объектами. Типичным результатом такой кластеризации является иерархическое дерево.

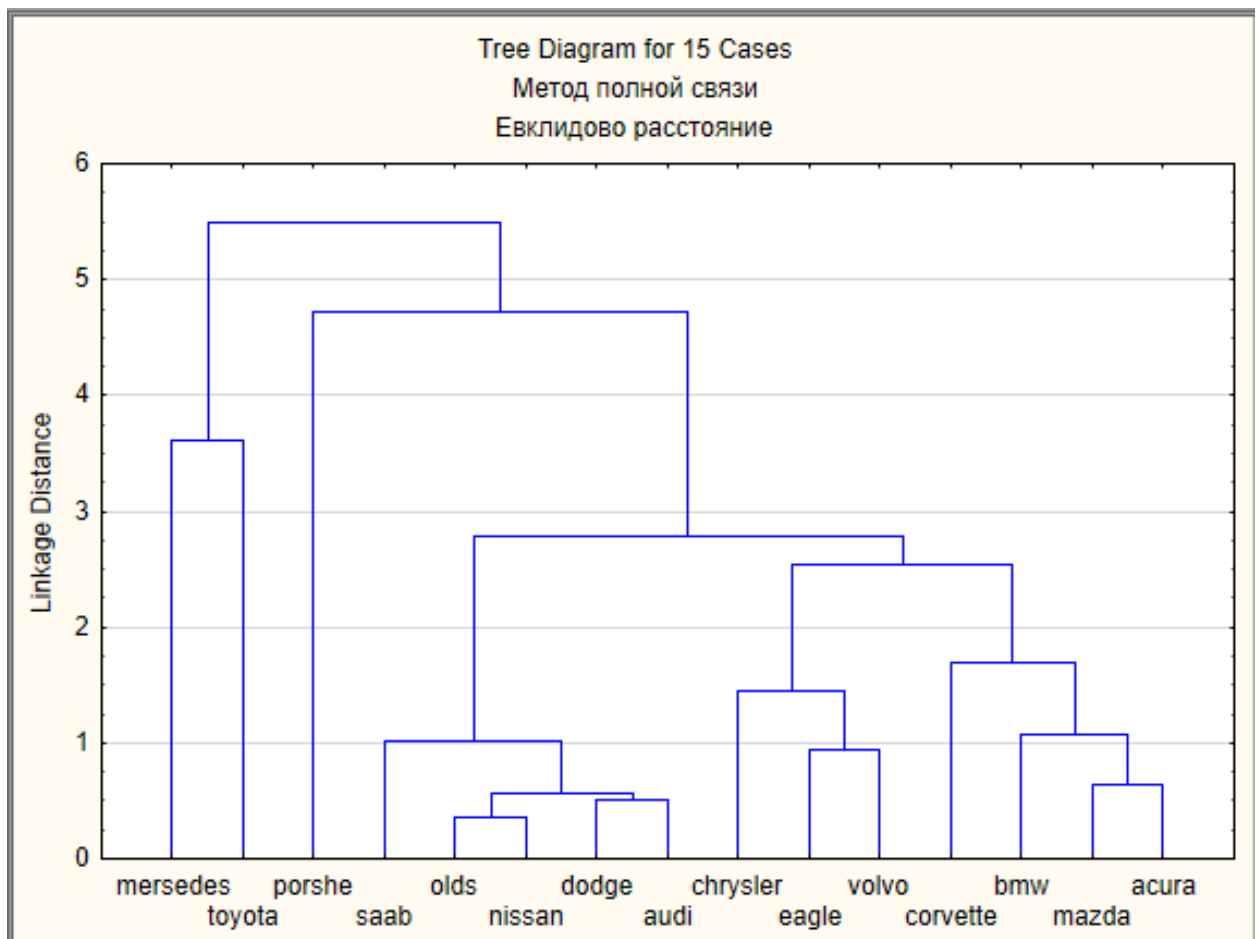
На первом этапе выясним, формируют ли автомобили "естественные" кластеры, которые могут быть осмыслены.

Метод полной связи определяет расстояние между кластерами как наибольшее расстояние между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями").

Мера близости, определяемая евклидовым расстоянием, является геометрическим расстоянием в n- мерном пространстве и вычисляется следующим образом:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Наиболее важным результатом, получаемым в результате древовидной кластеризации, является иерархическое дерево



Вначале древовидные диаграммы могут показаться немного запутанными, однако после некоторого изучения они становятся более понятными.

Диаграмма начинается сверху (для вертикальной дендрограммы) с каждого автомобиля в своем собственном кластере.

Как только вы начнете двигаться вниз, автомобили, которые "теснее соприкасаются друг с другом" объединяются и формируют кластеры. Каждый узел диаграммы, приведенной выше, представляет объединение двух или более кластеров, положение узлов на вертикальной оси определяет расстояние, на котором были объединены соответствующие кластеры.

### Алгоритм К-средних

Алгоритм К средних применяется для разделения выборки на некоторое заданное число кластеров К.

1. Векторы исходной выборки произвольным образом распределяются между К кластерами.
2. Для каждого кластера  $S_i$  произвольным образом задается центр  $x_i$
3. Векторы перераспределяются между кластерами. Для этого вычисляются расстояния от векторов  $X_i$  до центров кластеров.  $i$ -ый вектор определяется в тот кластер, расстояние до центра которого для него минимально
4. Алгоритм повторяется до тех пор, пока происходит перемещение векторов между кластерами

Результат выполнения алгоритма сильно зависит от числа кластеров  $K$ , а также от начального положения центров. Если на каком-то шаге алгоритма один из кластеров оказался пустым, стоит либо уменьшить  $K$ , либо пересмотреть начальное положение центров. Если один из кластеров оказался слишком большим по сравнению с другими, возможно стоит увеличить  $K$ .

## 4. Кластеризация методом $K$ средних

Исходя из визуального представления результатов, можно сделать предположение, что автомобили образуют четыре естественных кластера. Проверим данное предположение, разбив исходные данные методом  $K$  средних на 4 кластера, и проверим значимость различия между полученными группами.

В чем заключается метод?

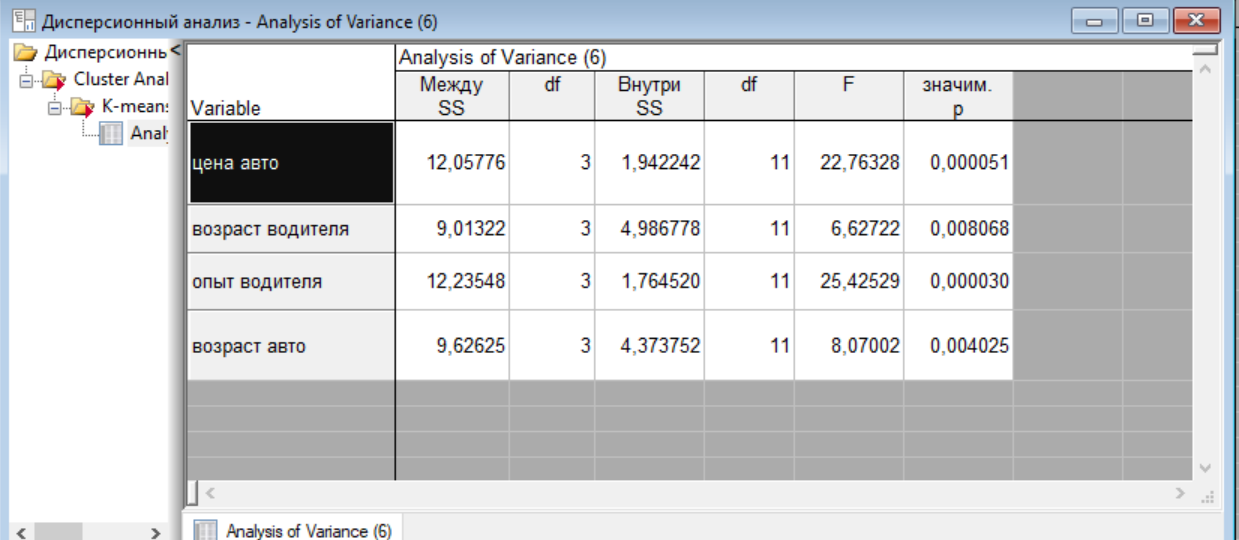
Метод  $K$ -средних заключается в следующем: вычисления начинаются с  $k$  случайно выбранных наблюдений (в нашем случае  $k=4$ ), которые становятся центрами групп, после чего объектный состав кластеров меняется с целью минимизации изменчивости внутри кластеров и максимизации изменчивости между кластерами.

Каждое следующее наблюдение ( $K+1$ ) относится к той группе, мера сходства с центром тяжести которого минимальна.

После изменения состава кластера вычисляется новый центр тяжести, чаще всего как вектор средних по каждому параметру. Алгоритм продолжается до тех пор, пока состав кластеров не перестанет меняться.

Когда результаты классификации получены, можно рассчитать среднее значение показателей по каждому кластеру, чтобы оценить, насколько они различаются между собой.

Выберем *Дисперсионный анализ* для определения значимости различия между полученными кластерами.



Variable	Между SS	df	Внутри SS	df	F	значим. p
цена авто	12,05776	3	1,942242	11	22,76328	0,000051
возраст водителя	9,01322	3	4,986778	11	6,62722	0,008068
опыт водителя	12,23548	3	1,764520	11	25,42529	0,000030
возраст авто	9,62625	3	4,373752	11	8,07002	0,004025

Variable	Analysis of Variance (6)					
	Между SS	df	Внутри SS	df	F	значим. р
цена авто	12,05776	3	1,942242	11	22,76328	0,000051
возраст водителя	9,01322	3	4,986778	11	6,62722	0,008068
опыт водителя	12,23548	3	1,764520	11	25,42529	0,000030
возраст авто	9,62625	3	4,373752	11	8,07002	0,004025

Variable	Между SS	df	Внутри SS	df	F	значим. р
цена авто	12,05776	3	1,942242	11	22,76328	0,000051
возраст водителя	9,01322	3	4,986778	11	6,62722	0,008068
опыт водителя	12,23548	3	1,764520	11	25,42529	0,000030
возраст авто	9,62625	3	4,373752	11	8,07002	0,004025

значение  $p > 0.05$ , что говорит о слабой схожести

Что означает эта таблица?

В ней приведена межгрупповая и внутригрупповая дисперсии.

Строки - переменные (наблюдения)

Столбцы - показатели для каждой переменной:

1. дисперсия между кластерами
2. число степеней свободы для межклассовой дисперсии
3. дисперсия внутри кластеров
4. число степеней свободы для внутриклассовой дисперсии
5. F - критерий, для проверки гипотезы о неравенстве дисперсий. Проверка данной гипотезы похожа на проверку гипотезы в дисперсионном анализе, когда делается предположение о том, что уровни фактора не влияют на результат.

Отобразим евклидовы расстояния объектов от центров (средних значений) соответствующих им кластеров.



Members of Cluster Number 1 (6) and Distances from Respective Cluster Center Cluster contains 1 cases	
Distance	
porsche	0,00

Members of Cluster Number 2 (6) and Distances from Respective Cluster Center Cluster contains 8 cases	
Distance	
acura	0,479839
bmw	0,473201
volvo	0,410675
corvette	0,546392
chrysler	0,809634
eagle	0,191911
mazda	0,536594
mercedes	1,069818

Members of Cluster Number 3 (6) and Distances from Respective Cluster Center Cluster contains 5 cases	
Distance	
audi	0,180878
saab	0,332406
dodge	0,114030
nissan	0,163738
olds	0,194055

Members of Cluster Number 4 (6) and Distances from Respective Cluster Center Cluster contains 1 cases	
Distance	
toyota	0,00

Первый кластер:

Members of Cluster Number 1 и расстояния до центра кластера Cluster contains 1 cases	
Distance	
porsche	0,00

Второй кластер:

Members of Cluster Number 2 и расстояния до центра кластера Cluster contains 8 cases	
Distance	
acura	0,479839
bmw	0,473201
volvo	0,410675
corvette	0,546392
chrysler	0,809634
eagle	0,191911
mazda	0,536594
mercedes	1,069818

Третий кластер:

Members of Cluster Number 3 и расстояния до центра кластера Cluster contains 5 cases	
Distance	
audi	0,180878
saab	0,332406
dodge	0,114030
nissan	0,163738
olds	0,194055

Четвертый кластер:

	Members of Cluster Number 4 и расстояния до центра кластера Cluster contains 1 cases		
	Distance		
toyota	0.00		

**Cluster Means&Euclidean Distances** позволяет вывести таблицы, в первой из которых указаны средние для каждого кластера (усреднение производится внутри кластера), во второй указаны евклидовы расстояния и квадраты евклидовых расстояний между кластерами.

В строках таблиц указано расстояние от каждой машины до центра кластера.

В каждом из четырех кластеров находятся объекты со схожим влиянием на процесс убытков.

### Оценка качества разбиения

Для оценки качества того или иного способа разбиения на кластеры вводятся функционалы качества разбиения  $Q(S)$ , определенные на множестве всех разбиений. Пусть в результате работы алгоритма получилось разбиение

$$S = S_1 \cup S_2 \dots \cup S_p$$

Оценка качества разбиения производится по следующим функционалам:

1. Сумма внутриклассовых дисперсий

$$Q_1(S) = \sum_{l=1}^K \sum_{X_i \in S_l} \varrho^2(X_i, \bar{X}_l),$$

2. Сумма попарных внутриклассовых расстояний

$$Q_2(S) = \sum_{l=1}^K \sum_{X_i, X_j \in S_l} \varrho^2(X_i, X_j)$$

3. Обобщенная внутриклассовая дисперсия

$$Q_3(S) = \det \left( \sum_{l=1}^K n_l W_l \right)$$

или

$$Q_4(S) = \prod_{l=1}^K (\det W_l)^{n_l},$$

, где  $W_l$  - ковариационная матрица.

Ковариационная матрица для кластера  $S_l$  составляется следующим образом:

$$W_l = \begin{pmatrix} cov(X_1, X_1) & cov(X_1, X_2) & \dots & cov(X_1, X_{n_l}) \\ cov(X_1, X_2) & cov(X_2, X_2) & \dots & cov(X_2, X_{n_l}) \\ \vdots & & \ddots & \vdots \\ cov(X_1, X_l) & cov(X_2, X_l) & \dots & cov(X_l, X_{n_l}) \end{pmatrix},$$

где

$$cov(X_i, X_j) = \frac{1}{n_l} \sum_{k=1}^m (X_{ki} - \bar{X}_{kl})(X_{kj} - \bar{X}_{kl})$$

при

$$i \neq j, cov(X_i, X_i) = D_i.$$

Разбиение, минимизирующее эти функционалы, считается оптимальным. Алгоритм К средних стремится минимизировать функционал  $Q_1(S)$ , иерархический алгоритм - функционал  $Q_1(S)$ . Результаты этих алгоритмов могут стремиться только к локальному, а не к глобальному минимуму функционалов, поэтому обычно требуются дополнительные исследования.

## 5. Сравнение результатов с помощью многочлена качества разбиения

Сравнение способов кластеризации.

$$S = S_1 \cup S_2 \dots \cup S_p$$

### 5.1. Описательная статистика

Cluster Number	Euclidean Distances between Clusters			
	Distances below diagonal			
	Squared distances above diagonal			
	No. 1	No. 2	No. 3	No. 4
No. 1	0,00000	346,1420	685,3314	987,5202
No. 2	18,60489	0,0000	146,4046	305,4047
No. 3	26,17883	12,0998	0,0000	37,1296
No. 4	31,42483	17,4758	6,0934	0,0000

**Табл. Расстояния между кластерами**

На экране появится таблица, в которой даны евклидовы расстояния между средними кластеров (по каждому из параметров внутри кластера вычисляется среднее, получается 3 точки в пятимерном пространстве, и между ними находится расстояние) (рисунок 11).

Из таблицы видим, что расстояние между первым и вторым кластером 346,142 , например

Над диагональю в таблице даны квадраты расстояний между кластерами.

Построим графики средних значений характеристик машин для каждого кластера

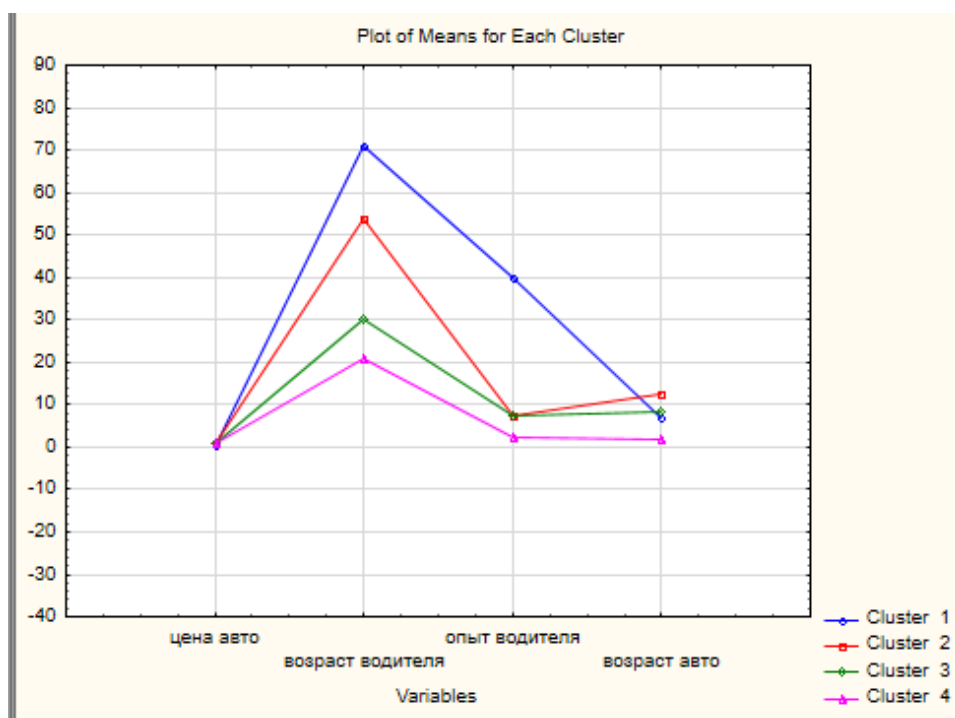


Табл. График средних для каждого кластера

Здесь строятся следующие графики средних значений характеристик машин для каждого кластера

## 6. Вывод