1. Данные

Целью данного анализа является разбиение автомобилей и их владельцев на классы, каждый из которых соответствует определенной рисковой 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

Сгруппируем по признакам совокупности данных

	стоимость Авто	возраст Водителя	стаж Водителя	возраст Авто
Acura	0,521	23	1	31
Audi	0,866	26	8	4
bmw	0,496	43	20	11
buick	0,614	22	2	2
corvette	1,243	37	7	6
chrysler	0,613	54	20	2
dodge	0,781	61	40	3
eagle	0,612	22	4	21
ford	0,753	45	10	5
honda	0,511	44	20	10
isuzu	0,788	65	4	7
mazda	0,143	32	10	12
mersedes	1,051	38	9	25
mitsub	0,674	51	2	15
nissan	0,489	36	16	1
porshe	3,543	48	7	17
saab	0,544	59	12	30

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

таблица со стандартизированными переменными:

	стоимость Авто	возраст Водителя	стаж Водителя	возраст Авто
Acura	0,008015385	0,353846154	0,015384615	0,47692308
Audi	0,013323077	0,400000000	0,123076923	0,06153846
bmw	0,007630769	0,661538462	0,307692308	0,16923077
buick	0,009446154	0,338461538	0,030769231	0,03076923
corvette	0,019123077	0,569230769	0,107692308	0,09230769
chrysler	0,009430769	0,830769231	0,307692308	0,03076923
dodge	0,012015385	0,938461538	0,615384615	0,04615385
eagle	0,009415385	0,338461538	0,061538462	0,32307692
ford	0,011584615	0,692307692	0,153846154	0,07692308
honda	0,007861538	0,676923077	0,307692308	0,15384615
isuzu	0,012123077	1	0,061538462	0,10769231
mazda	0,0022	0,492307692	0,153846154	0,18461538
mersedes	0,016169231	0,584615385	0,138461538	0,38461538
mitsub	0,010369231	0,784615385	0,030769231	0,23076923
nissan	0,007523077	0,553846154	0,246153846	0,01538462
porshe	0,054507692	0,738461538	0,107692308	0,26153846
saab	0,008369231	0,907692308	0,184615385	0,46153846

3. Иерархическая классификация

Будем считать расстояния, используя формулу Евклида:
$d(\vec{x}, \vec{y})=\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_m - y_m)^2}$

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие. Алгоритм иерархической кластеризации начинает работу с формирование несвязанных кластеров, помещая каждый из объектов в отдельный кластер. Определяется способ интерпретации матрицы расстояний, чтобы объединить два или более кластера следующего уровня. Процесс повторяется, пока не получим резкого скачка в минимальном расстоянии.

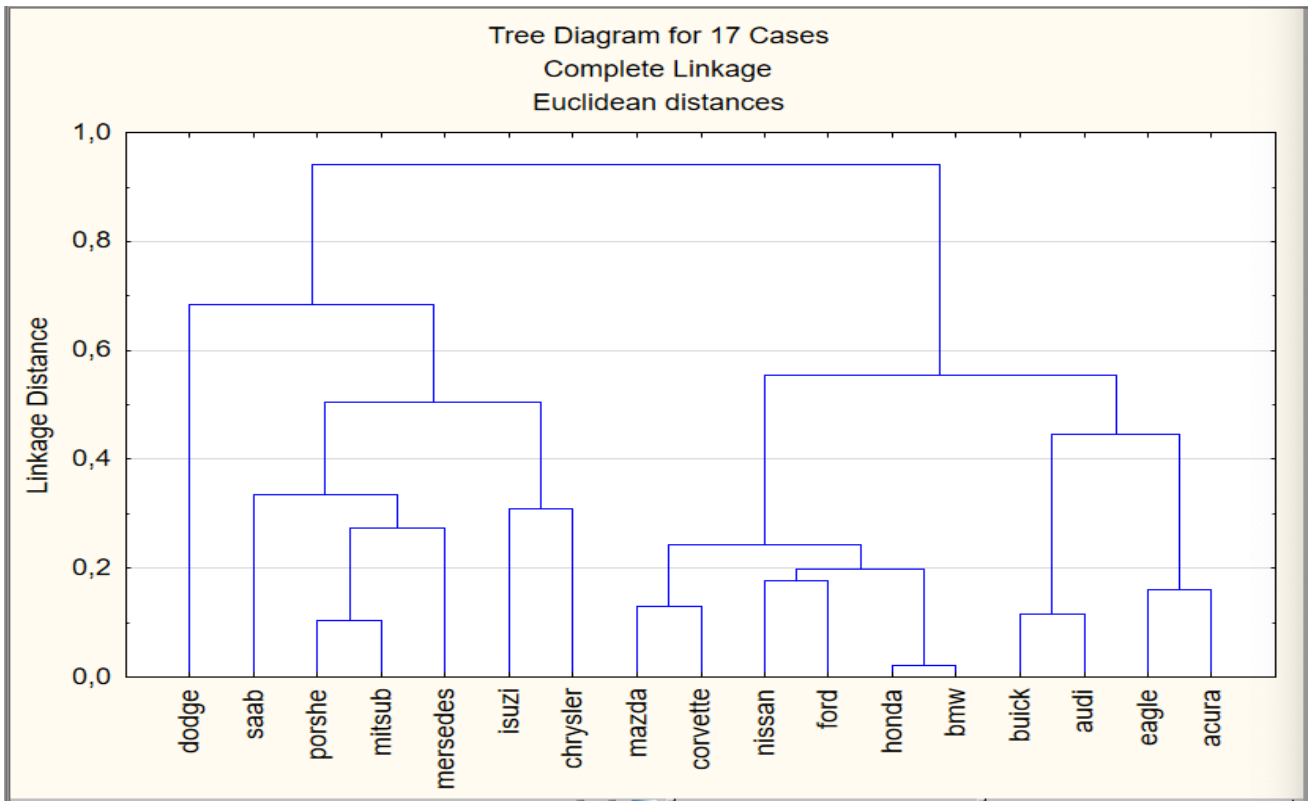
Методы поиска необходимых расстояний:

- Принцип «ближнего соседа»

- Принцип «дальнего соседа»
- Принцип средних связей

Получим матрицу 17 на 17. Далее найдем минимальное расстояние. В нашем случае оно равно 0,02. И объединим 3 и 10 в один кластер. Матрица стала 16 на 16.
Теперь пересчитаем матрицу. Расстояние до нового кластера будем брать по принципу «дальнего соседа». Повторяем весь алгоритм, пока не получим резкого скачка в минимальном расстоянии.

	Euclidean distances (Spreadsheet11)																
Case No.	acura	audi	bmw	buick	corvette	chrysler	dodge	eagle	ford	honda	isuzi	mazda	mersedes	mitsub	nissan	porshe	saab
acura	0,00	0,43	0,52	0,45	0,45	0,72	0,94	0,16	0,54	0,54	0,75	0,35	0,28	0,50	0,55	0,45	0,58
audi	0,43	0,00	0,34	0,12	0,17	0,47	0,73	0,28	0,29	0,35	0,60	0,16	0,37	0,43	0,20	0,40	0,65
bmw	0,52	0,34	0,00	0,45	0,23	0,22	0,43	0,43	0,18	0,02	0,42	0,23	0,28	0,31	0,20	0,24	0,40
buick	0,45	0,12	0,45	0,00	0,25	0,56	0,84	0,29	0,38	0,45	0,67	0,25	0,44	0,49	0,30	0,47	0,73
corvette	0,45	0,17	0,23	0,25	0,00	0,34	0,63	0,33	0,13	0,24	0,43	0,13	0,29	0,27	0,16	0,24	0,51
chrysler	0,72	0,47	0,22	0,56	0,34	0,00	0,33	0,62	0,21	0,20	0,31	0,40	0,46	0,34	0,28	0,32	0,45
dodge	0,94	0,73	0,43	0,84	0,63	0,33	0,00	0,86	0,52	0,42	0,56	0,66	0,68	0,63	0,53	0,59	0,60
eagle	0,16	0,28	0,43	0,29	0,33	0,62	0,86	0,00	0,44	0,45	0,70	0,23	0,27	0,46	0,42	0,41	0,60
ford	0,54	0,29	0,18	0,38	0,13	0,21	0,52	0,44	0,00	0,17	0,32	0,23	0,33	0,22	0,18	0,20	0,44
honda	0,54	0,35	0,02	0,45	0,24	0,20	0,42	0,45	0,17	0,00	0,41	0,24	0,30	0,31	0,20	0,24	0,40
isuzi	0,75	0,60	0,42	0,67	0,43	0,31	0,56	0,70	0,32	0,41	0,00	0,52	0,51	0,25	0,49	0,31	0,39
mazda	0,35	0,16	0,23	0,25	0,13	0,40	0,66	0,23	0,23	0,24	0,52	0,00	0,22	0,32	0,20	0,27	0,50
mersedes	0,28	0,37	0,28	0,44	0,29	0,46	0,68	0,27	0,33	0,30	0,51	0,22	0,00	0,27	0,39	0,20	0,34
mitsub	0,50	0,43	0,31	0,49	0,27	0,34	0,63	0,46	0,22	0,31	0,25	0,32	0,27	0,00	0,38	0,10	0,30
nissan	0,55	0,20	0,20	0,30	0,16	0,28	0,53	0,42	0,18	0,20	0,49	0,20	0,39	0,38	0,00	0,34	0,57
porshe	0,45	0,40	0,24	0,47	0,24	0,32	0,59	0,41	0,20	0,24	0,31	0,27	0,20	0,10	0,34	0,00	0,28
saab	0,58	0,65	0,40	0,73	0,51	0,45	0,60	0,60	0,44	0,40	0,39	0,50	0,34	0,30	0,57	0,28	0,00



Если выбрать кол-во кластеров = 2 , то получим след. Кластеры:

Кластер 1:

Acura	0,00801538	0,353846154	0,01538462	0,47692308
Audi	0,01332308	0,400000000	0,12307692	0,06153846

bmw	0,00763077	0,661538462	0,30769231	0,16923077
buick	0,00944615	0,338461538	0,03076923	0,03076923
corvette	0,01912308	0,569230769	0,10769231	0,09230769
eagle	0,00941538	0,338461538	0,06153846	0,32307692
ford	0,01158462	0,692307692	0,15384615	0,07692308
honda	0,00786154	0,676923077	0,30769231	0,15384615
mazda	0,0022	0,492307692	0,15384615	0,18461538
nissan	0,00752308	0,553846154	0,24615385	0,01538462

Кластер 2:

chrysler	0,00943077	0,830769231	0,30769231	0,03076923
dodge	0,01201538	0,938461538	0,61538462	0,04615385
isuzu	0,01212308	1	0,06153846	0,10769231
mersedes	0,01616923	0,584615385	0,13846154	0,38461538
mitsub	0,01036923	0,784615385	0,03076923	0,23076923
porshe	0,05450769	0,738461538	0,10769231	0,26153846
saab	0,00836923	0,907692308	0,18461538	0,46153846

4. Кластеризация методом К средних

Исходя из визуального представления результатов, можно сделать предположение, что автомобили образуют четыре естественных кластера. Проверим данное предположение, разбив исходные данные методом К средних на 2 кластера, и проверим значимость различия между полученными группами.

Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике.

Алгоритм завершается, когда на какой-то итерации не происходит изменения внутрикластерного расстояния. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное отклонение V уменьшается, поэтому зацикливание невозможно.
Так как в прошлом пункте получили 6 классов, значит, здесь возьмем k=6.

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.

Начальные значения центров выбирались случайным образом.

Центр кластера – среднее геометрическое место точек в пространстве переменных.

$$x_{kj} = \frac{\sum_{j=1}^{n} w_j x_{ij}}{I_k}$$

	Кластер 1	Кластер 2
центры:	(0,00378 ; 0,0944 ; 0,2307 ; 0,17)	(0,0106; 0,4538 ; 0,109; 0,1961)

Первый кластер. Элементы и расстояния до центра кластера:

	расстояния
bmw	0,01351917
chrysler	0,56778082
dodge	0,17193507
ford	0,02359665
honda	0,01486619
isuzu	0,04146958
mitsub	0,05239673
porshe	0,03136845

saab	0,09527594

Второй кластер. Элементы и расстояния до центра кластера:

0,09771775	Acura
0,02120906	Audi
0,04688385	buick
0,02417309	corvette
0,031736	eagle
0,00364022	mazda
0,05348088	mersedes
0,06133006	nissan

5. Сравнение результатов с помощью многочлена качества

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 $S = S_1 \cup S_2 \cup ... \cup S_K$. Оценка качества разбиения производится по следующим функционалам:

1. Сумма внутриклассовых дисперсий

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

2. Сумма попарных внутриклассовых расстояний

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

Оценим качество разбиения для кластеров из алгоритма К-средних:

Q1(S) = 1,01
Q2(S) = 0,34

Оценим качество разбиения для кластеров из Иерархического алгоритма:

Q1(S) = 0,287
Q2(S) = 0,529

Q характеризует дисперсию кластеров.
Дисперсия кластера – это мера рассеяния точек в пространстве относительно центра кластера.

Вывод

Итак, после проведения кластеризации мы выяснили, что наиболее успешная кластеризация получилась иерархическим алгоритмом.