МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (ФГБОУ ВО «ВГУ»)

Факультет прикладной математики, информатики и механики

**Отчет**

на тему:

**«Кластерный анализ»**

Выполнил: студент 3 к. 6 гр. ПМИ

Бакаев Илья Игоревич

Проверил: преподаватель

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Воронеж – 2020

**Теоретическая информация**

**Кластерный анализ** — многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.

Спектр применений кластерного анализа очень широк: его используют в археологии, медицине, психологии, химии, биологии и других дисциплинах. Однако универсальность применения привела к появлению большого количества несовместимых терминов, методов и подходов, затрудняющих однозначное использование и непротиворечивую интерпретацию кластерного анализа.

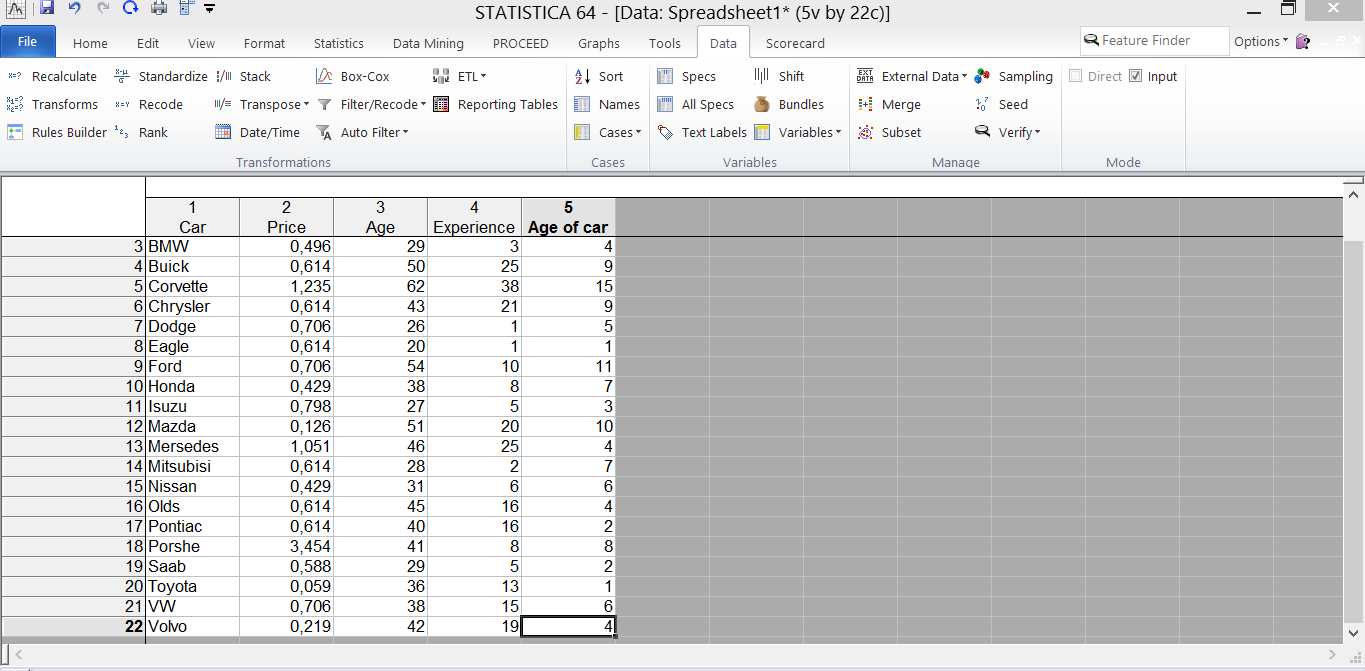
Кластерный анализ выполняет следующие основные задачи:

* Разработка типологии или классификации.
* Исследование полезных концептуальных схем группирования объектов.
* Порождение гипотез на основе исследования данных.
* Проверка гипотез или исследования для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

**Постановка задачи**

Исходный файл данных содержит следующую информацию об автомобилях и их владельцах:

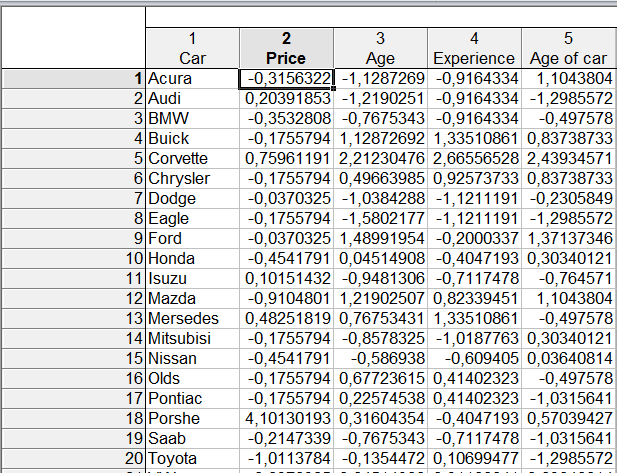
* марка автомобиля – первая переменная;
* стоимость автомобиля – вторая переменная;
* возраст водителя – третья переменная;
* стаж водителя – четвертая переменная;
* Возраст автомобиля – пятая переменная;



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.

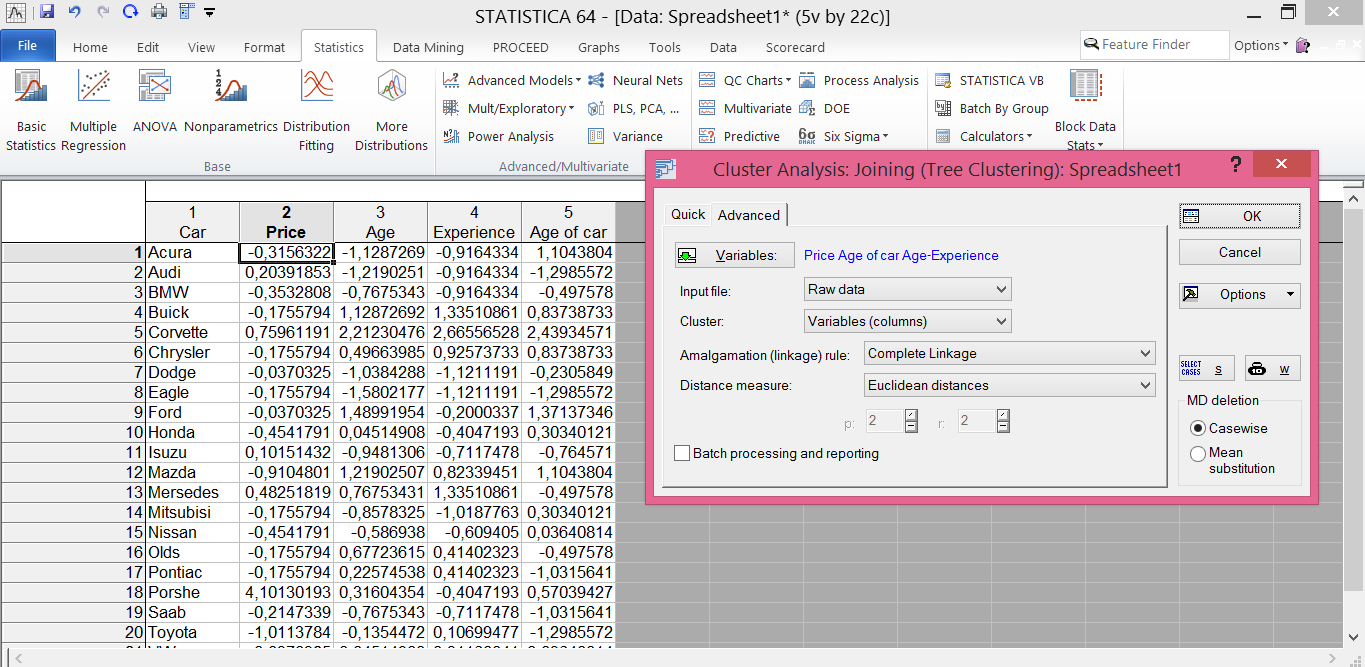
Использование кластер-анализа для решения данной задачи наиболее эффективно. В общем случае кластер-анализ предназначен для объединения некоторых объектов в классы (кластеры) таким образом, чтобы в один класс попадали максимально схожие, а объекты различных классов максимально отличались друг от друга.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

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

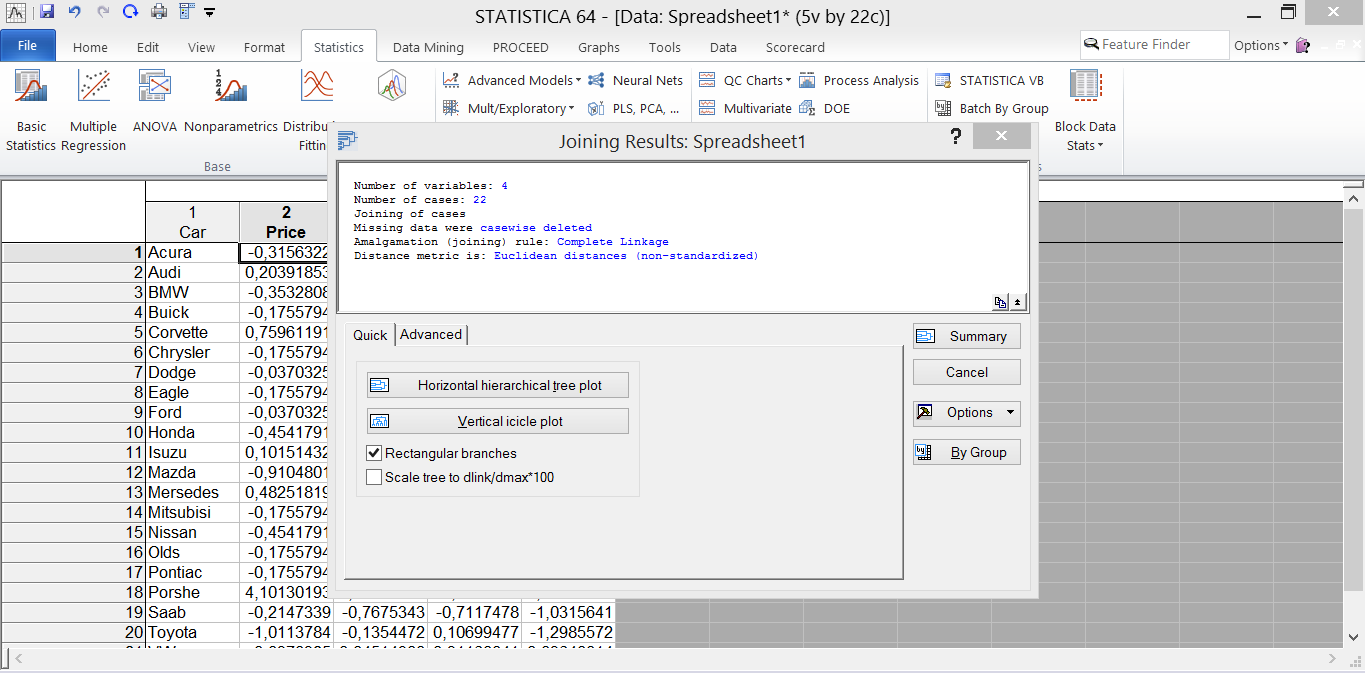
****

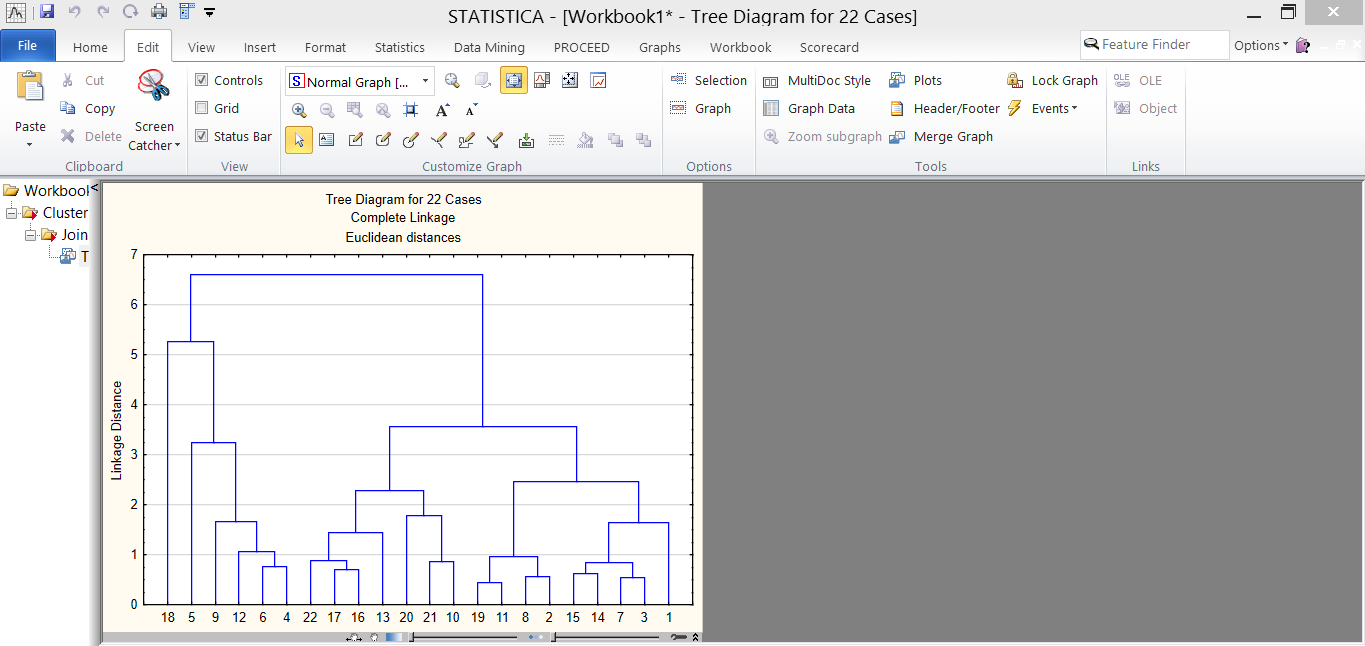
**2.** На первом этапе выясним, формируют ли автомобили естественные кластеры, которые могут быть осмыслены.

Входим в раздел Statistics → Mult/Exploratory → Cluster → Joining (Tree clustery)

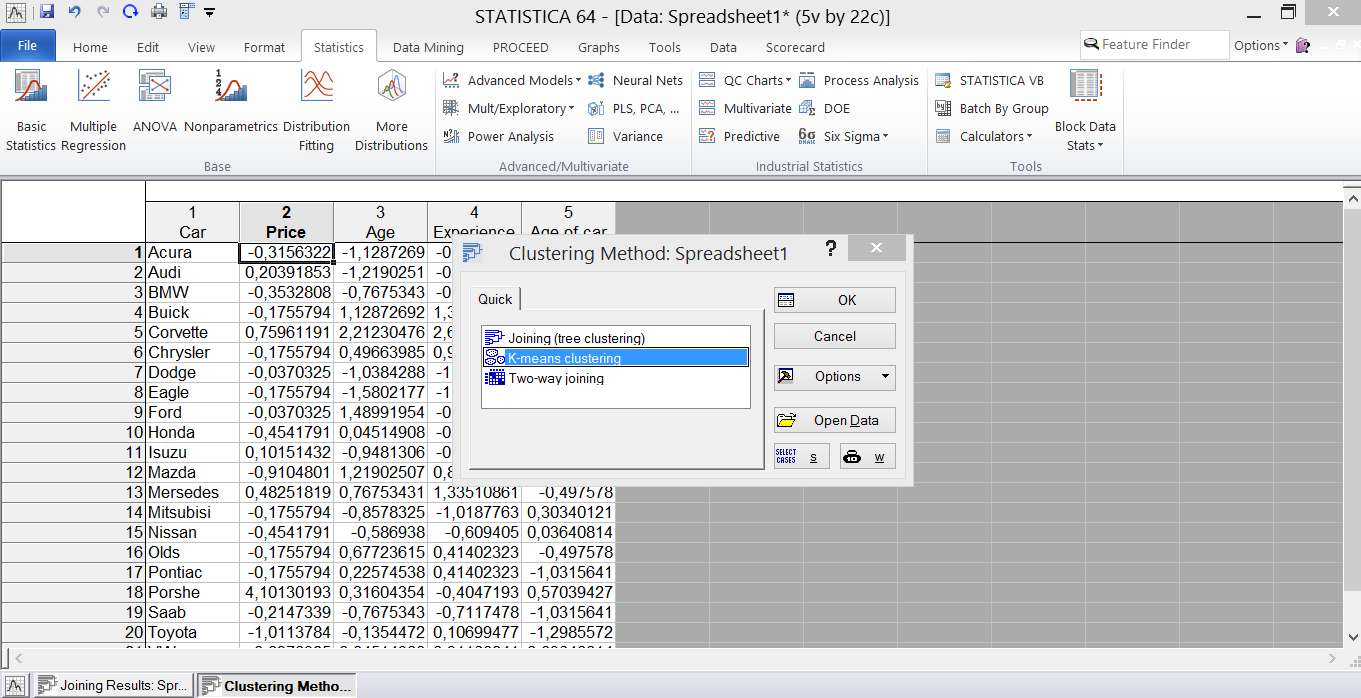
****

3. Наиболее важным результатом, получаемым в результате древовидной кластеризации, является иерархическое дерево. Нажмем на кнопку Vertical icicle plot

****

****

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



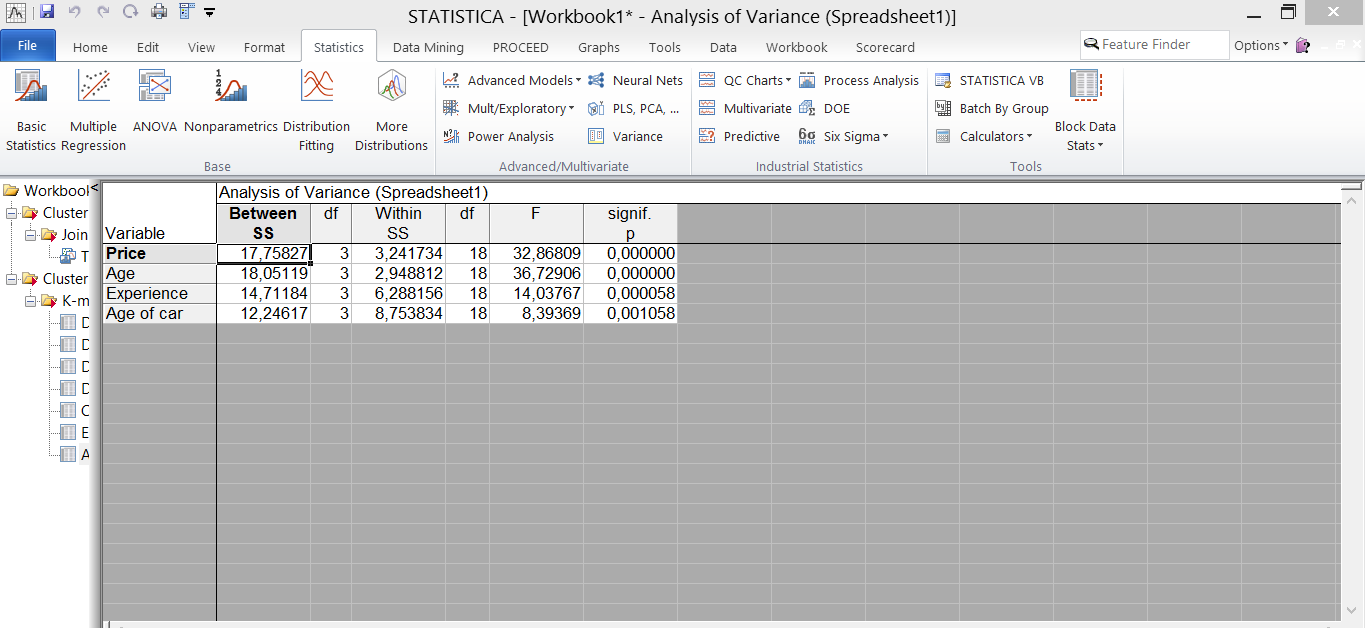
Метод *K-средних* заключается в следующем: вычисления начинаются с k случайно выбранных наблюдений (в нашем случае k=4), которые становятся центрами групп, после чего объектный состав кластеров меняется с целью минимизации изменчивости внутри кластеров и максимизации изменчивости между кластерами.

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.

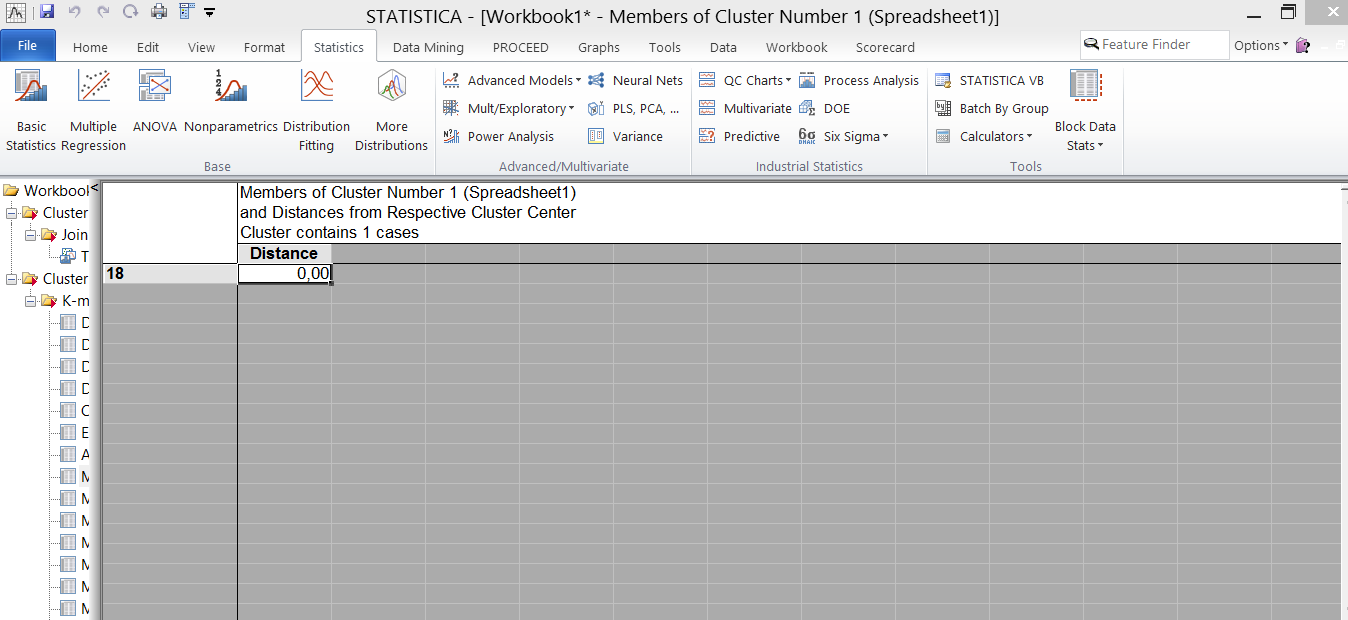
4. В окне Результаты метода К среднихвыберем Дисперсионный анализ для определения значимости различия между полученными кластерами



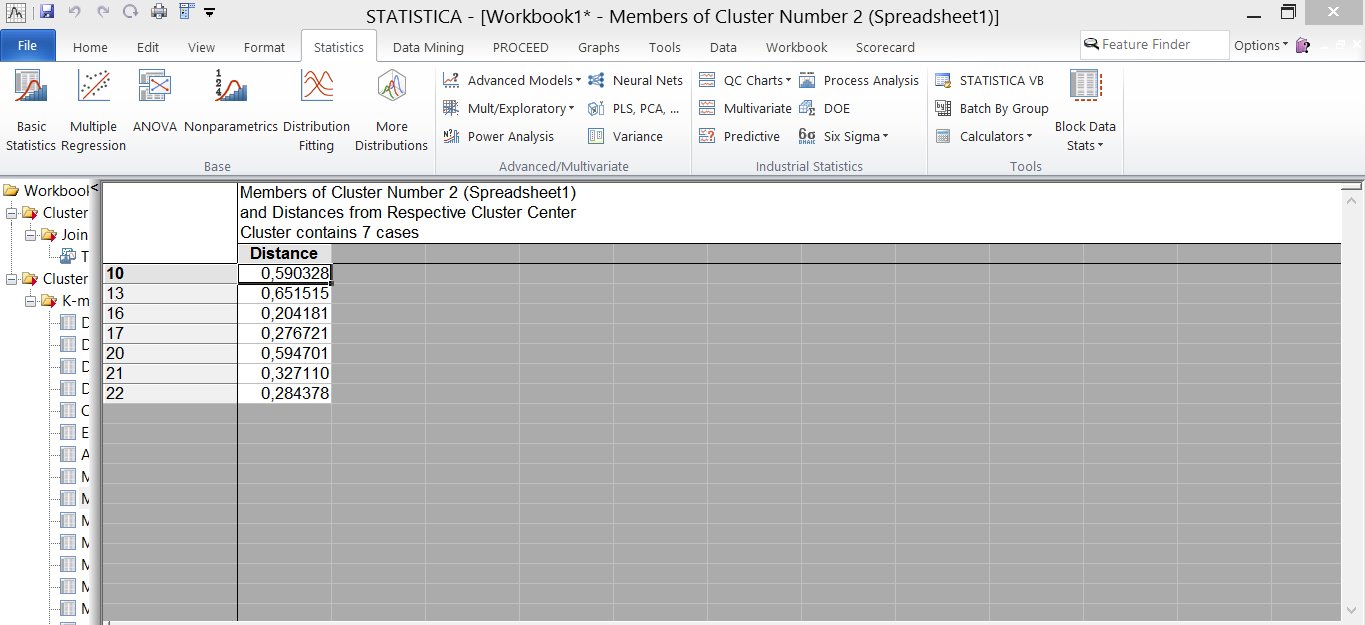
Значение р<0.05, что говорит о значимом различии.

5. Нажмем кнопку Cluster’s members and distances для просмотра наблюдений, входящих в каждый из кластеров:

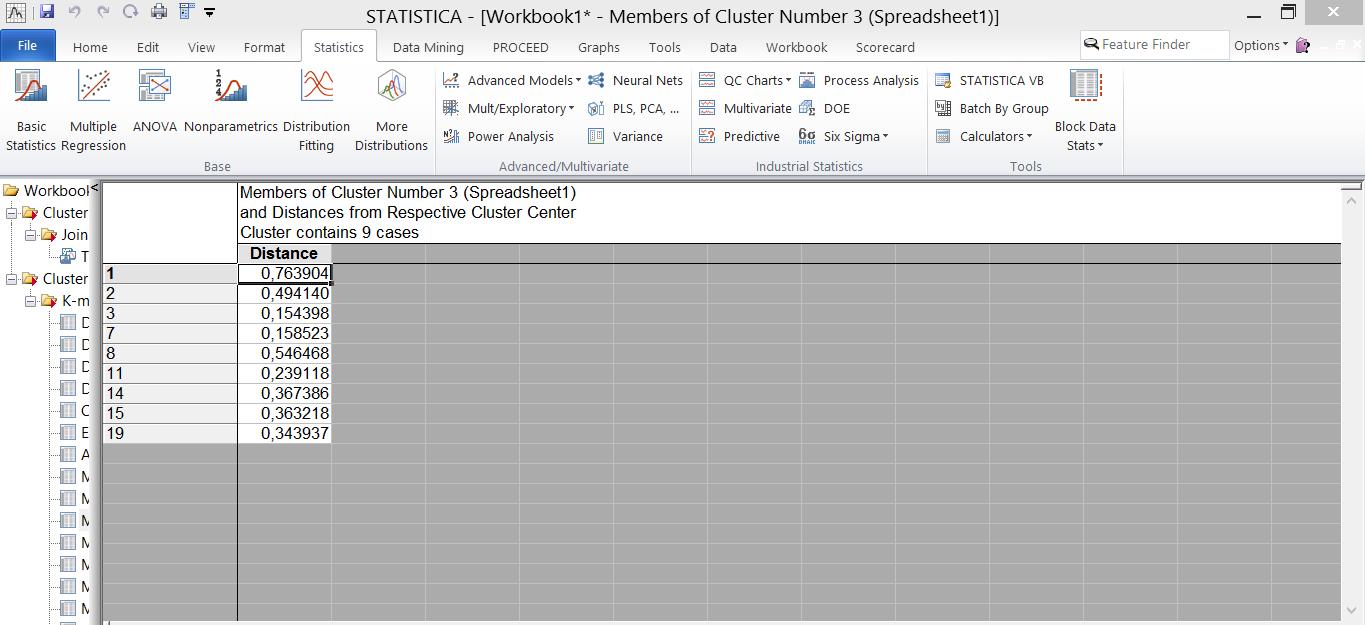
**Первый кластер**



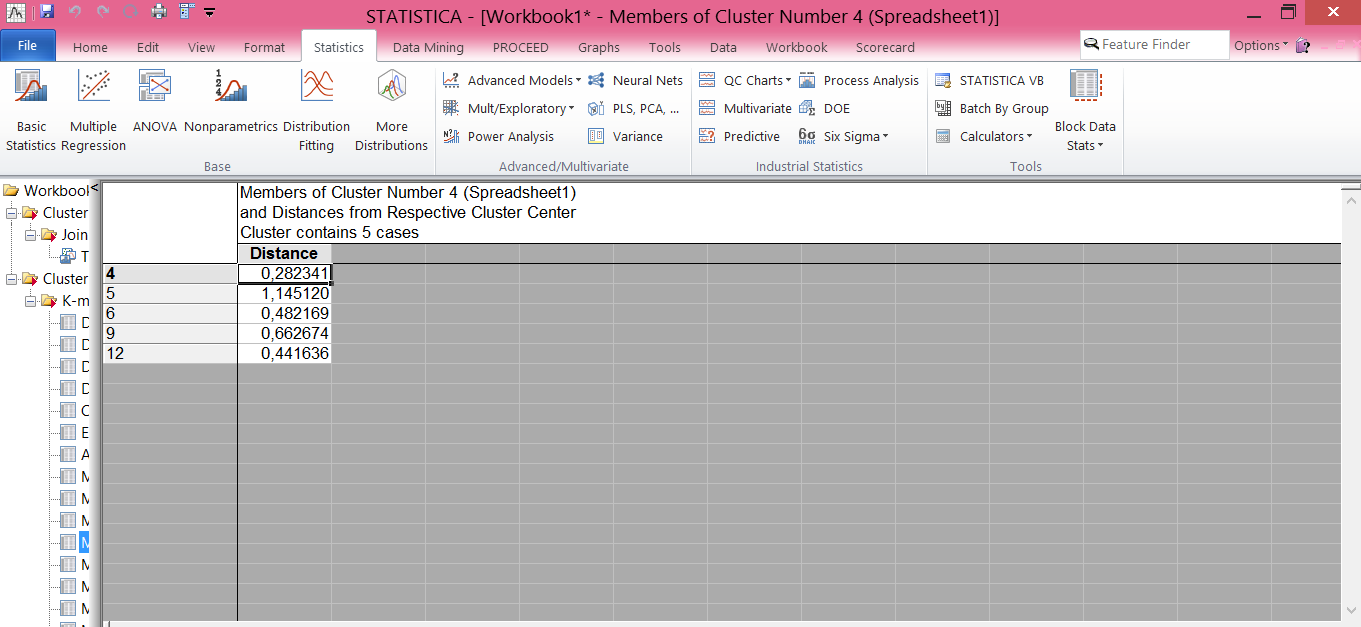
**Второй кластер**

****

**Третий кластер**

****

**Четвёртый кластер**

****

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

6. Нажмем кнопку Save classifications and distances

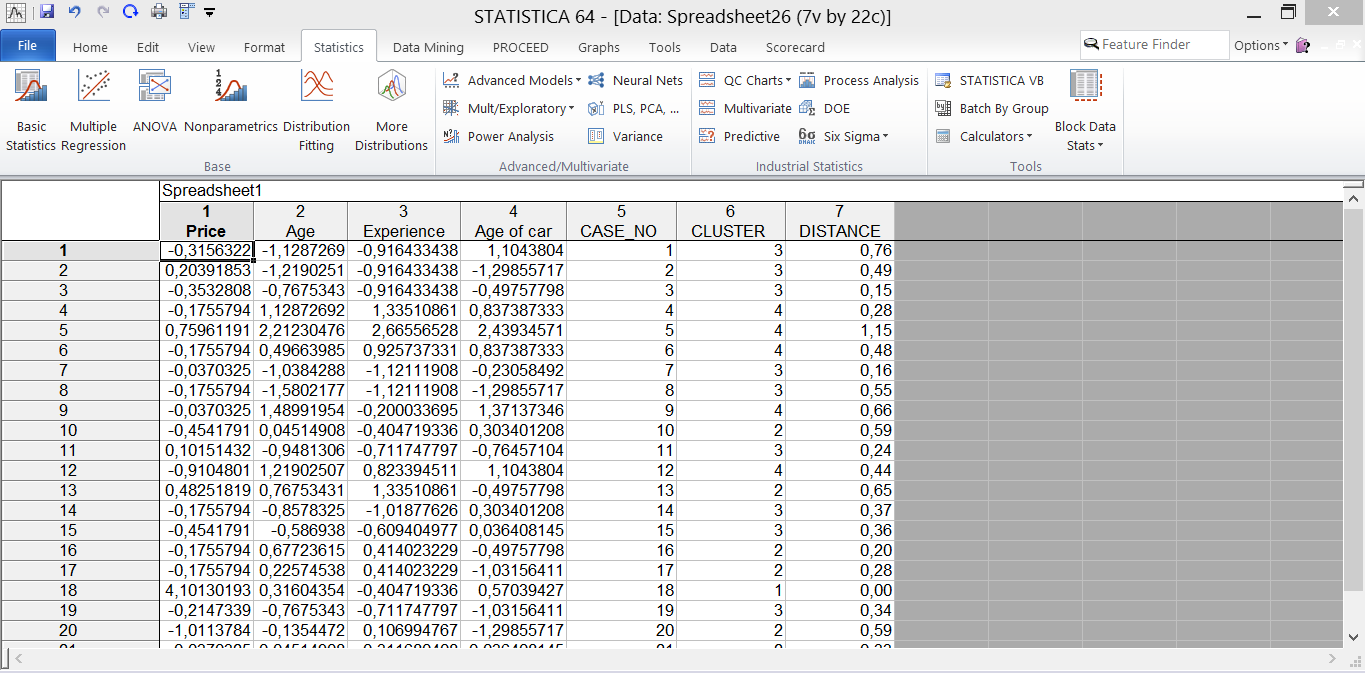
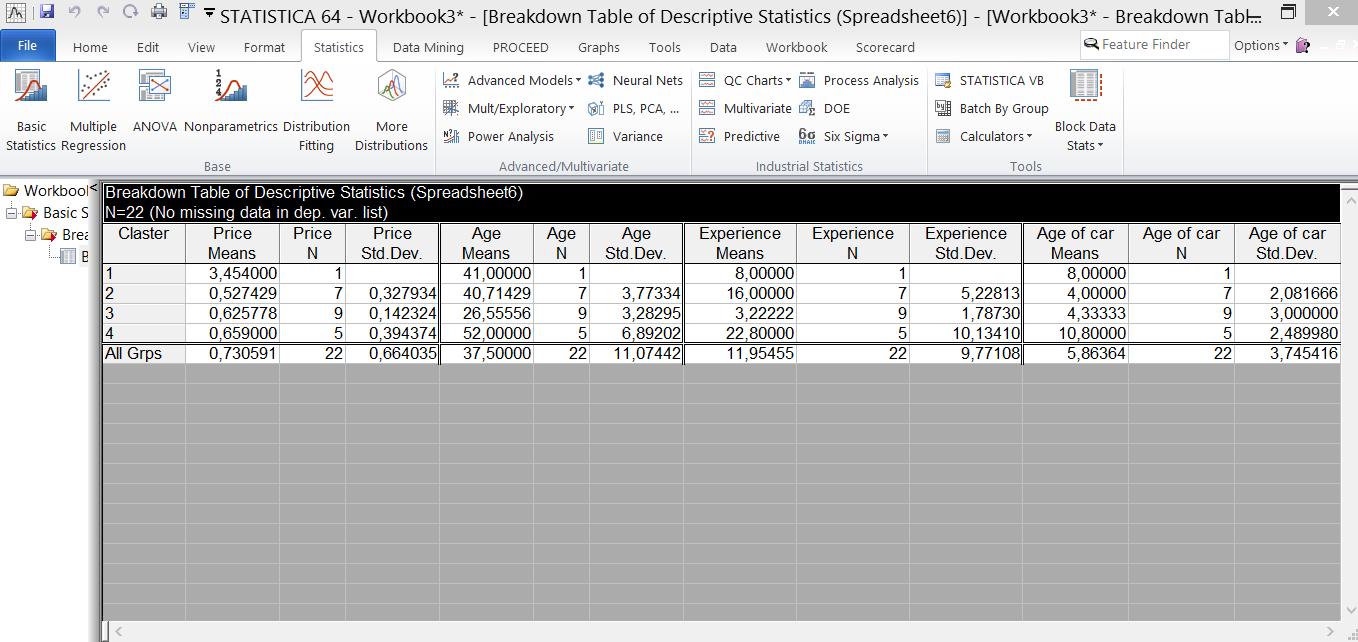
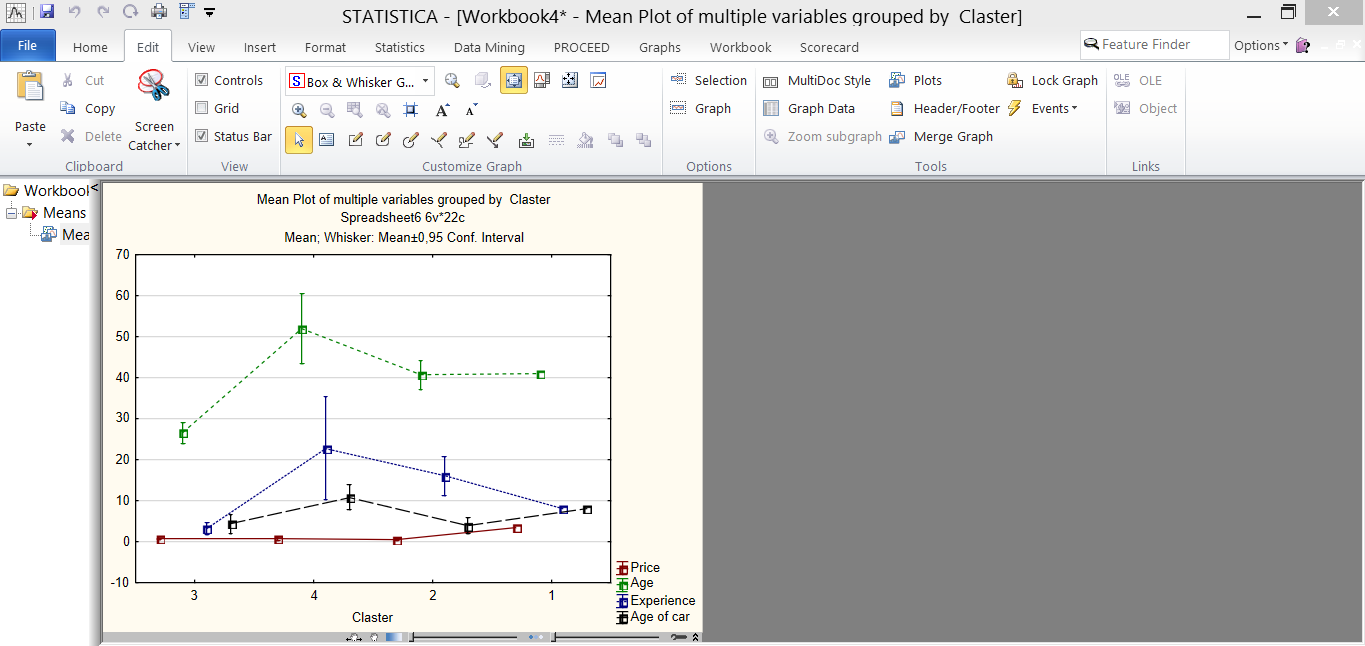


Таблица стандартизованных данных дополнилась информацией о кластере, к которому принадлежит наблюдение, евклидовом расстоянии и номере наблюдения.

7. Скопируем переменную Cluster в исходную таблицу данных и для каждого кластера вычислим описательные статистики, нажав на кнопку Breakdown & one-way ANOVA:



8. Построим график средних и доверительных интервалов для переменных в каждом кластере



Итак, для каждого кластера может быть определена вероятность наступления страхового случая