**МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ**

Федеральное государственное автономное образовательное

Учреждение высшего образования

**НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ**

**ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**

Инженерная школа ядерных технологий

Отделение экспериментальной физики

Направление «Прикладная математика и информатика»

Отчет по лабораторной работе №8

Вариант 2

По дисциплине

**«МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ»**

**Кластерный анализ**

Выполнил:

Студент группы 0В01 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Белясов А.А.

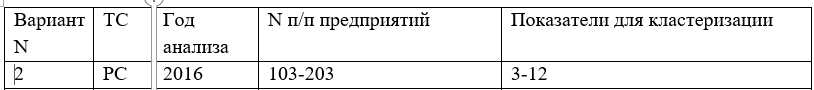
Проверил:

Старший преподаватель ОЭФ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Шинкеев М.Л.

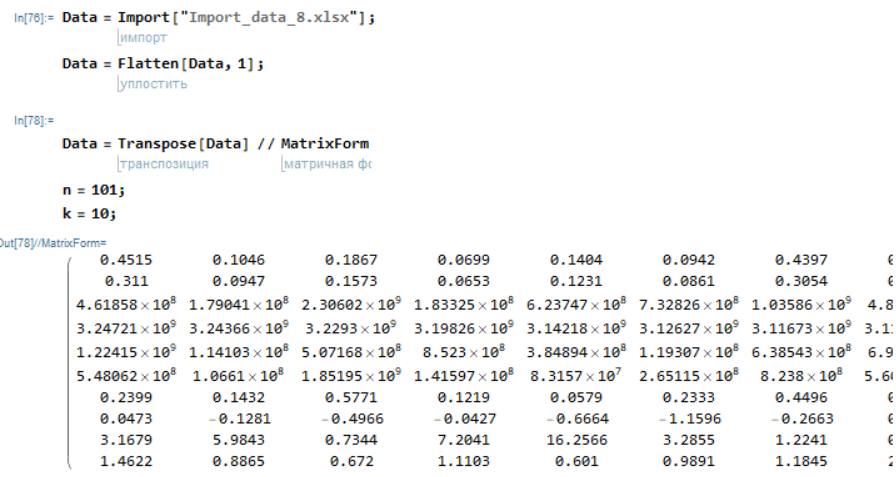
Томск 2023

В файле “Данные” на листах “РС”, “CC”, “ИС” приведены экономические показатели предприятий пищевой промышленности России за 2012 – 2016 годы различных форм собственности. Для указанных в варианте задания типа собственности предприятий, года анализа и группы экономических показателей требуется:

1. Используя метод главных компонент, перейти от пространства исходных признаков (показателей) к факторному пространству главных компонент, объясняющих не менее 70% вариации исходных признаков.
2. Провести иерархическую кластеризацию предприятий на основе построенных главных компонент, используя алгоритм Варда (выбрать в качестве меры близости объектов квадрат евклидова расстояния).
3. На основе, полученной в результате кластеризации дендрограммы, выбрать оптимальное разбиение совокупности объектов на кластеры. В качестве оптимального разбиения выбрать множество кластеров, для которого: 1) по каждому фактору (главной компоненте) средние значения фактора по кластерам для всей совокупности кластеров значимо различаются; 2) для каждой пары различных кластеров существует значимое различие средних хотя бы для одного фактора (главной компоненты); 3) для любого другого разбиения с большим числом кластеров условия 1 или 2 не выполняются. Для анализа различий средних по кластерам использовать методы дисперсионного анализа. Для получения устойчивого разбиения, соответствующего условиям 1-3, выбрать подходящий критерий множественного сравнения (Дункана, Бонферрони и др.), а также при необходимости подобрать уровень значимости критерия.
4. Провести кластеризацию той же совокупности, используя алгоритм k-средних, положив число кластеров, равным, полученному числу кластеров на предыдущем этапе. Оценить качество кластеризации, используя дисперсионный анализ (проверить для полученного разбиения выполнение условий 1-2 предыдущего раздела).
5. Сравнить полученные в пунктах 3 и 4 разбиения, используя в качестве функционала качества разбиения сумму средних квадратов расстояний до центров кластеров.



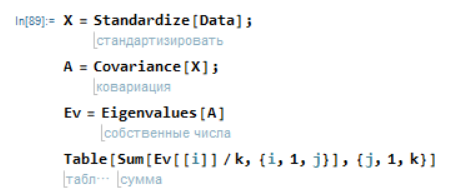
**Ход работы:**



Кластерный анализ

Построение факторной модели. Рассмотрим суммарную дисперсию первых k компонентов. Факторная модель необходима для уменьшения количества входных параметров. Сейчас в модели 10 различных факторов.

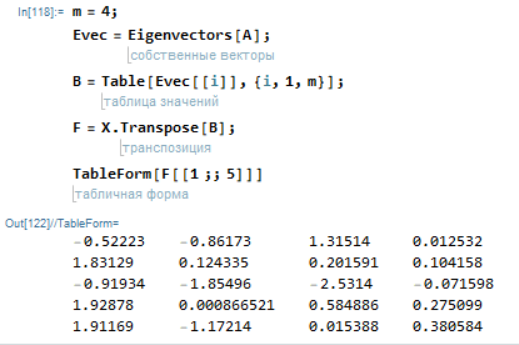
Стандартизируем данные



Получили 10 собственных значений матрицы ковариации и объясняемый первыми k компонентами процент дисперсии.

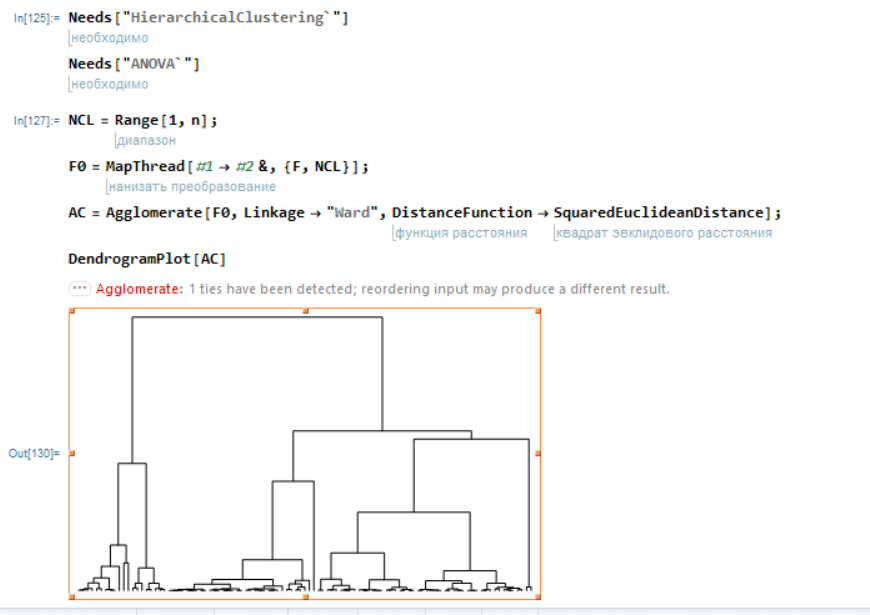
Видно, что 4 компоненты объясняют с округлением вниз 70% дисперсии.

Рассмотрим модель с четырьмя факторами



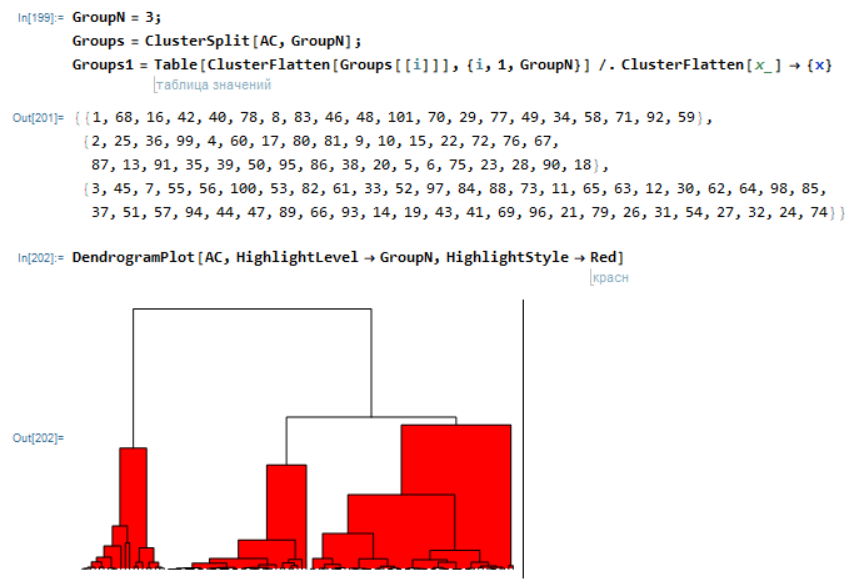
Кластеризация методом Варда и множественное сравнение средних:

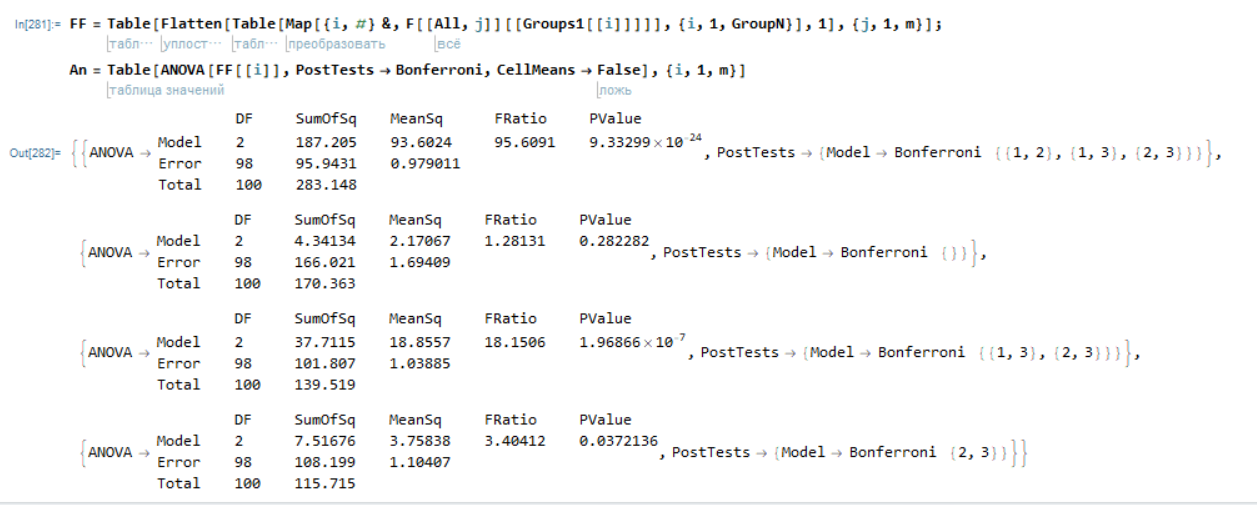
Проведем кластеризацию методом Варда с евклидовой мерой расстояния. Дадим номера каждому предприятию по порядку. Конвертировать в номера исходных предприятий можно прибавив к каждому номеру 203.



На диаграмме видно, как сгруппированы данные, оптимально взять такое число кластеров, чтобы выделить группы по основным признакам, это первые ветви иерархической кластеризации. Пусть разбиение будет проведено по группе из 3-х кластеров (можно поварьировать и посмотреть на различные результаты, хороший получается, когда выбрано 3 кластера).

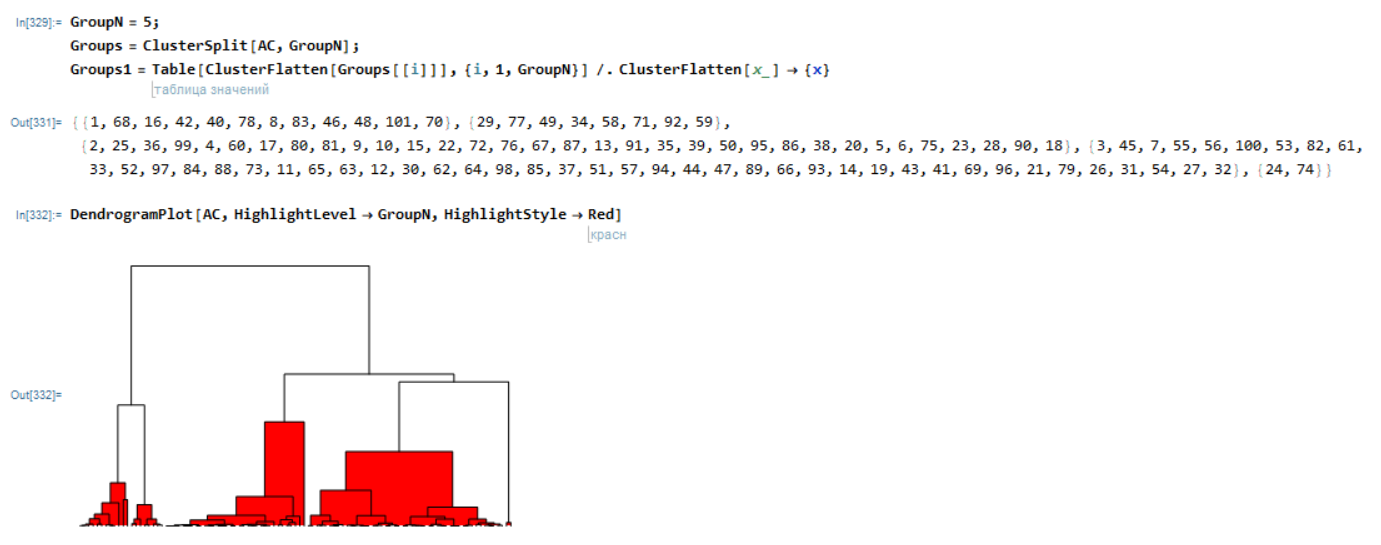
Разбиваем на 3 кластера

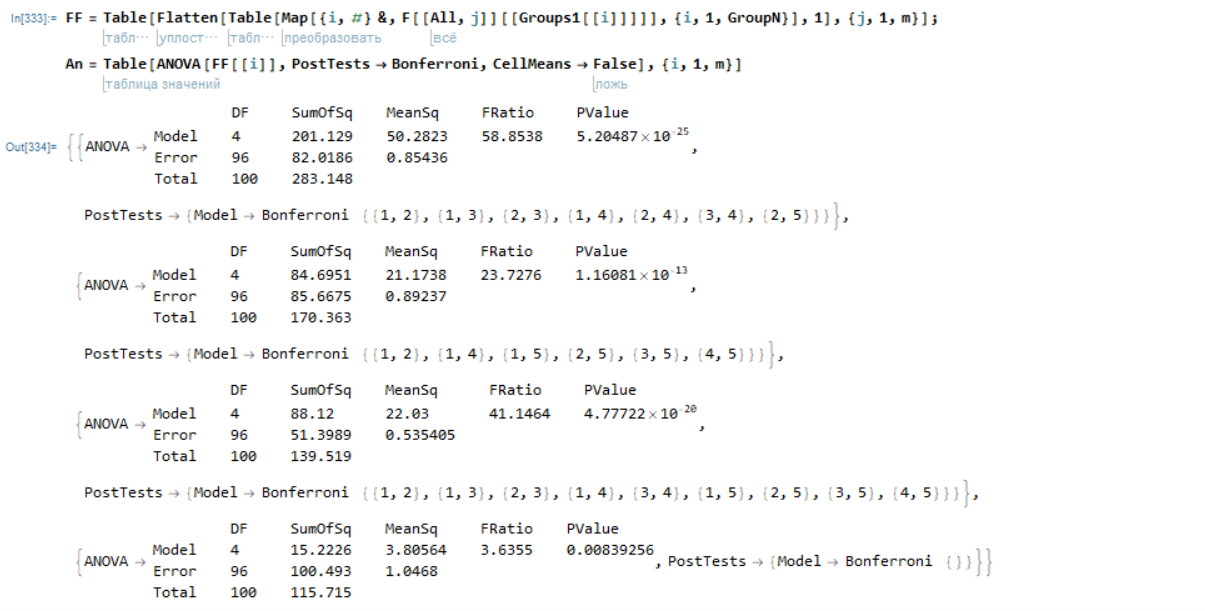




Из статистики видно, что PValue для фактора F2 слишком большое и этот фактор не значим. Следует увеличить количество кластеров

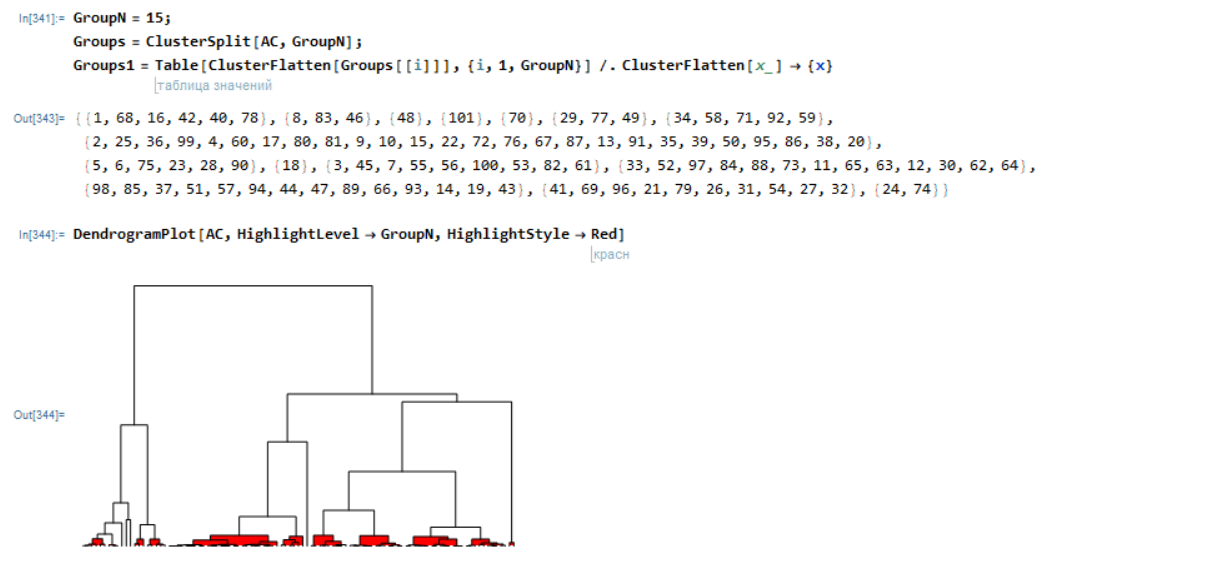
Разбиваем на 5 кластеров (тк при разбиении на 4 фактор F4 становится не значим):

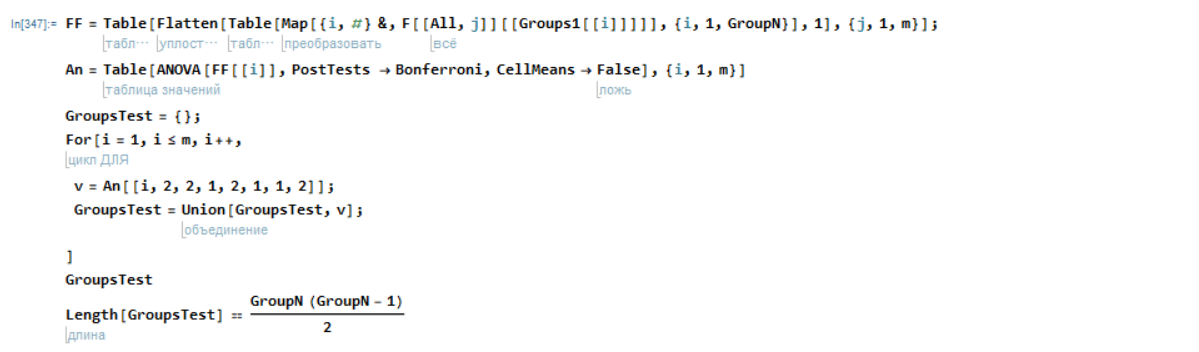


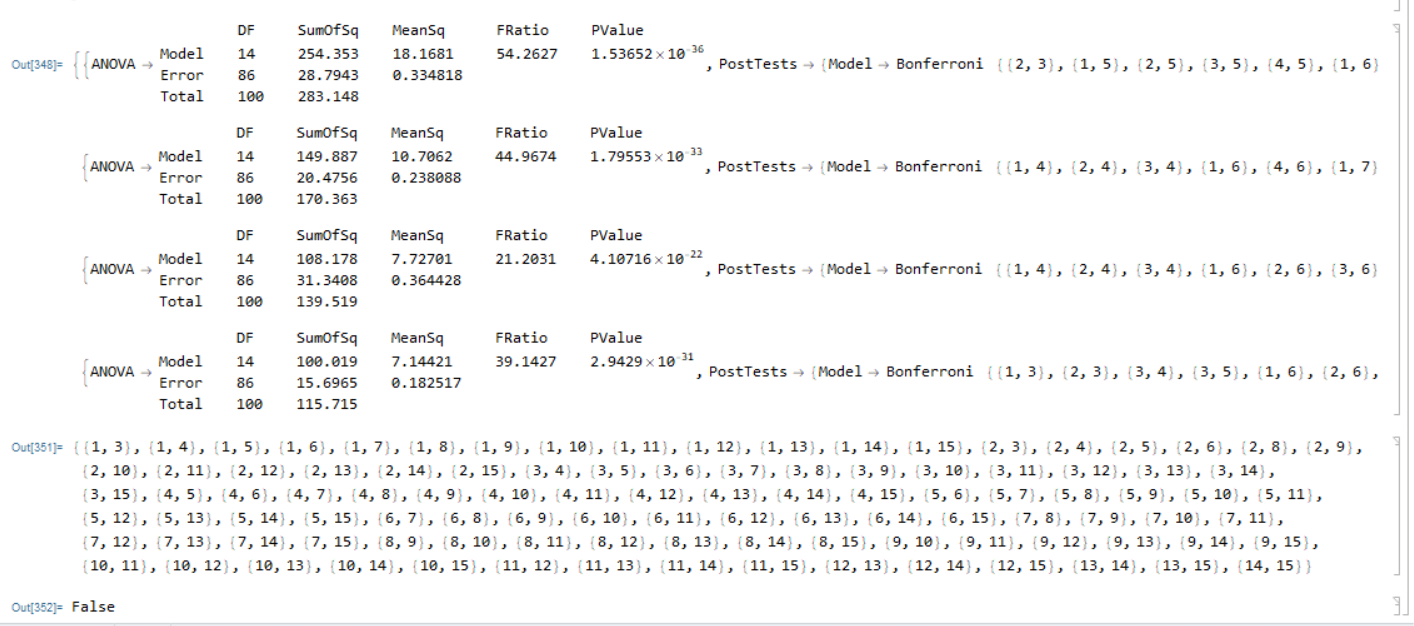


В данном случае фактор F2 значим, и мы имеем достаточно разные суммы квадратов средних, такую модель можно использовать, но пока рано говорить, что она оптимальна. Для оптимальной модели следует минимизировать сумму квадратов средних внутри кластеров и максимизировать суммы квадратов средних между кластерами.

Разбиваем на 15 кластеров

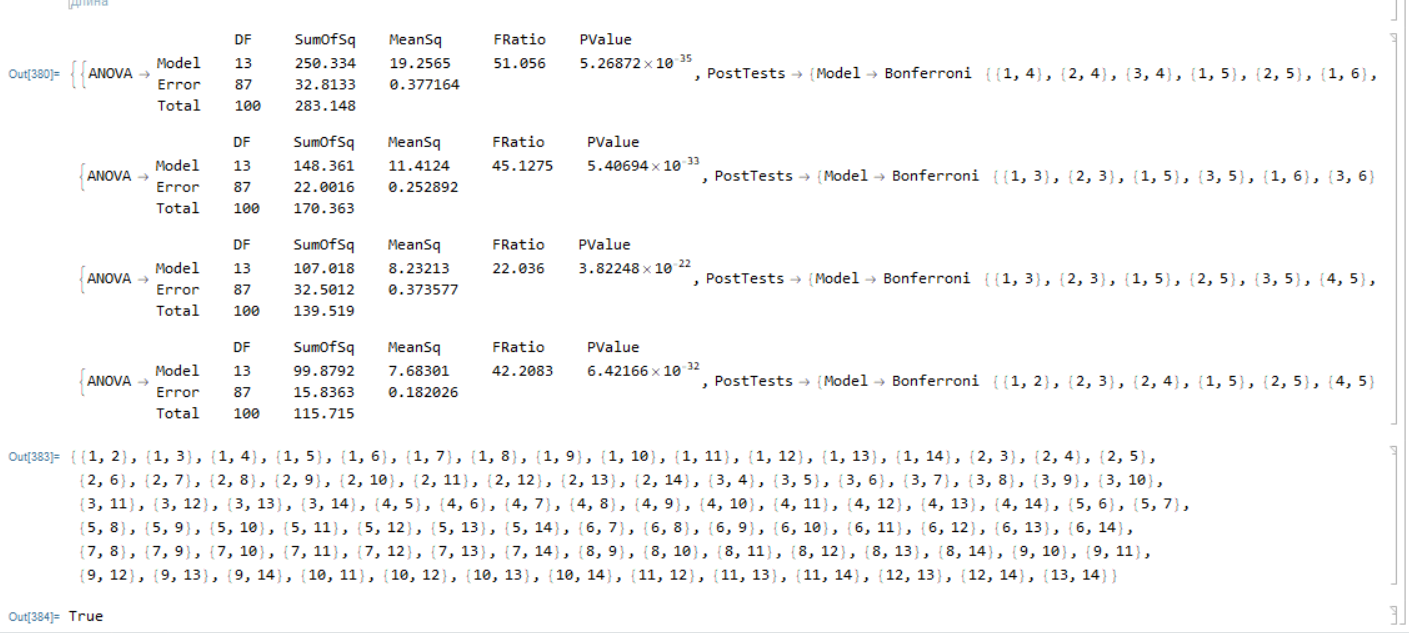






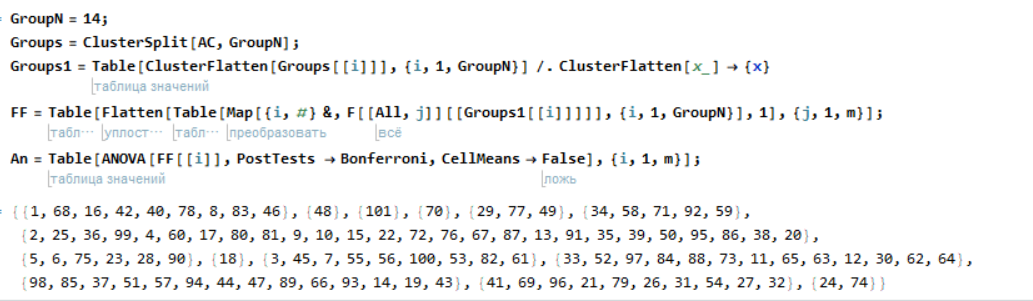
Видим, что для 15 кластеров, не существует значимого различия средних ни по одному фактору для одной пары.

Разбиваем на 14 кластеров

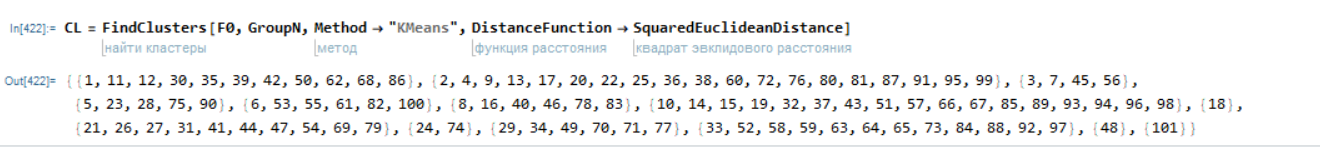


Здесь условие значимости различия средних для всех пар выполняется. Значит разбиение на 14 кластеров является оптимальным.

Обновим переменные. Разбиваем на 14 кластеров



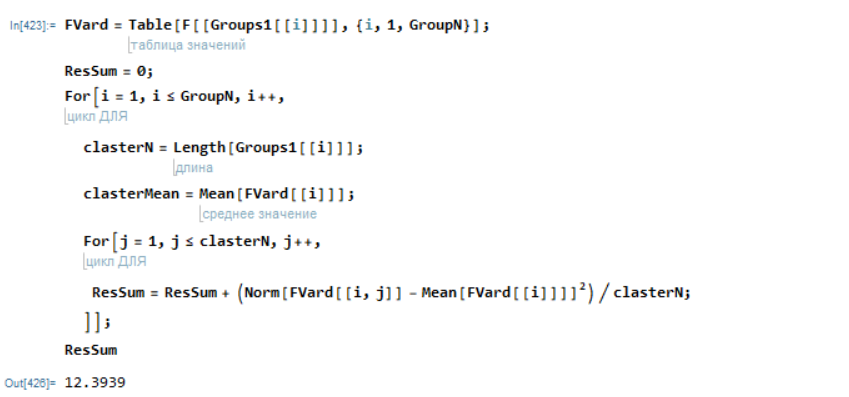
**Кластеризация методом k-средних. Разбиение на кластеры**



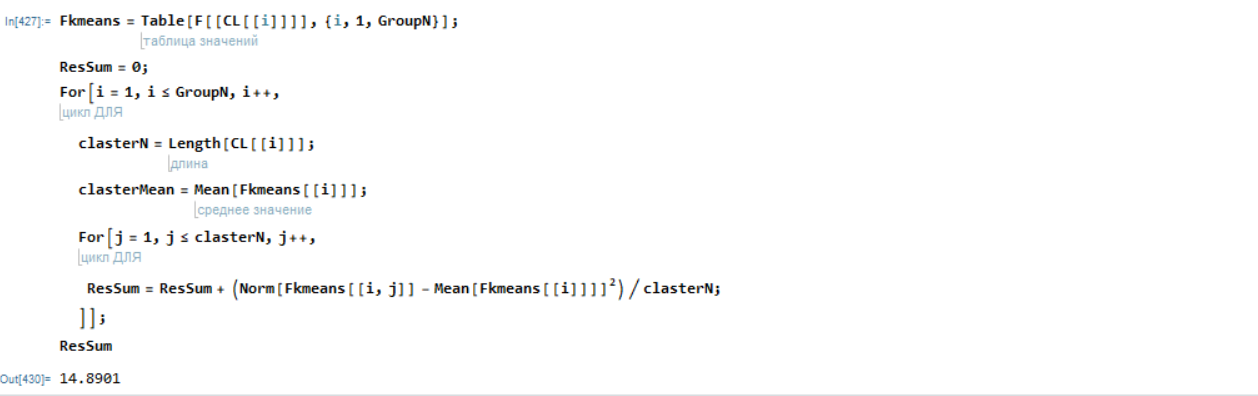
В данном случае разбиение на кластеры отличается от разбиения методом Варда.

Сравним кластеризации

Рассчитаем сумму средних квадратов расстояний до центров кластеров для разбиения методом Варда.



Рассчитаем сумму средних квадратов расстояний до центров кластеров для разбиения методом k-средних.



Сумма средних квадратов расстояний до центров кластеров для разбиения методом k-средних оказалась выше, чем для разбиения методом Варда. Метод k-средних минимизирует расстояние от объектов до центров кластеров, а метод Варда минимизирует квадрат расстояния. Расстояние до центров кластеров оказалось велико, и метод Варда сработал более оптимально, чем метод k-средних в данном случае, для оптимального количества кластеров, равного 14.

**Вывод:**

В ходе лабораторной работы проведен кластерных анализ двумя различными методами: методом Варда и методом k-средних. Перед кластеризацией получена факторная модель с меньшим числом признаков.

Для модели полученной методом Варда число кластеров определено из приблизительного вида иерархического дерева, а также из суммы квадратов расстояний и значимости пар кластеров.

В конечном итоге, произведено сравнение двух методов на основе суммы квадратов расстояний до центров кластеров, кластеризация методом Варда показала лучший результат.