ФИО студента Гаврилина А., Глушкова Д. (группа 2)

Дисциплина: Анализ данных

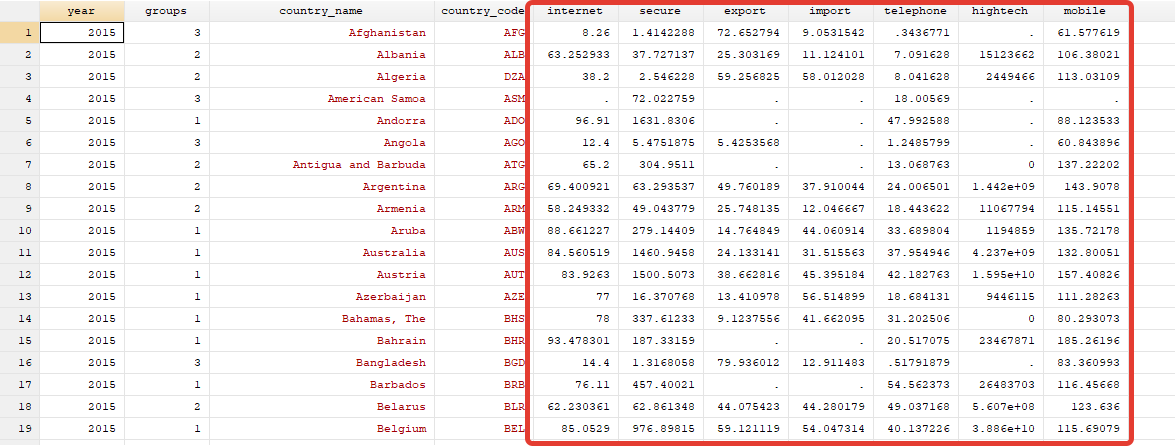
**Задание 1**

Откройте файл **Worldbank\_data.dta**

Проведите кластерный анализ данных с использованием всех переменных кроме year, groups, country\_name, country\_code. Используйте **метод Варда**.

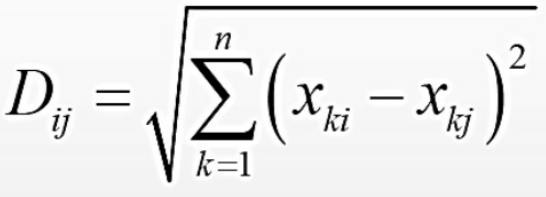
**1. Определите число кластеров, объясните своё решение.**

Изучим данные:



Имеются численные значения различных порядков.

Согласно формуле евклидовой меры, переменная, имеющая большие значения, практически полностью доминирует над переменной с малыми значениями.



Для решения этой проблемы используется *z-преобразование* (стандартизация) значений переменных, которая приводит значения всех переменных к единому диапазону значений.

// стандартизация всех переменных

egen zinternet = std(internet)

egen zsecure = std(secure)

egen zexport = std(export)

egen zimport = std(import)

egen ztelephone = std(telephone)

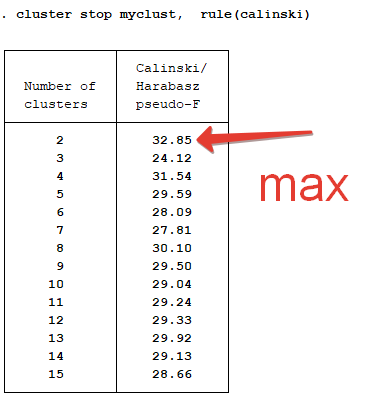
egen zhightech = std(hightech)

egen zmobile = std(mobile)

// проведение анализа методом Варда

cluster wardslinkage zinternet zsecure zexport zimport ztelephone zhightech zmobile , measure(L2) name(myclust)

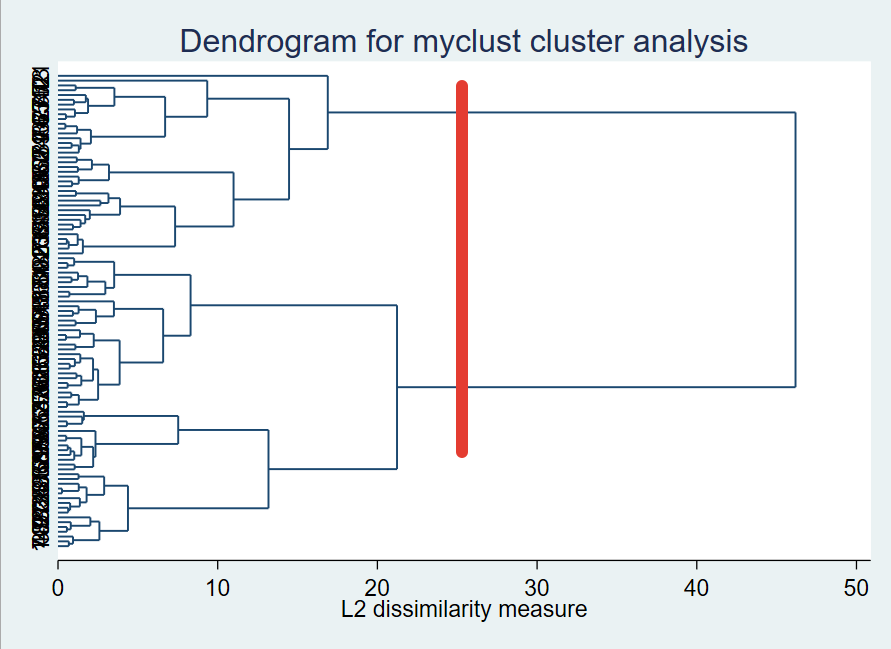
cluster stop myclust, rule(calinski)



**2. Создайте дендрограмму, скопируйте её в этот файл и опишите её.**

// дендрограмма

cluster dendrogram myclust, horizontal



Оптимально выбрать **2 кластера**. Хотя можно было бы выбрать 4 кластера (см. результат вычисления индекса), так как значения отличаются немного. Оставим 2 кластера.

**3. Опишите получившиеся кластеры с помощью описательной статистики. Нужно не просто скопировать таблицы, но детально описать полученные результаты.**

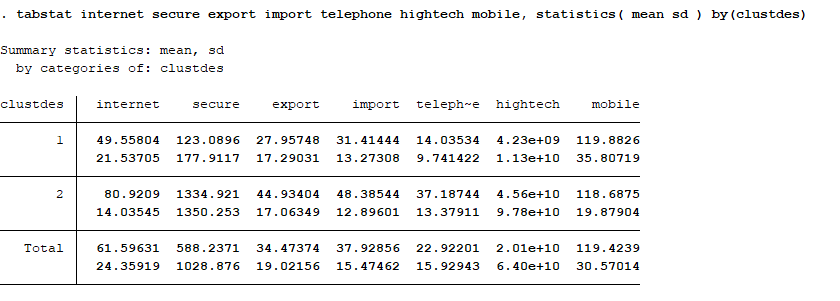
// сохранение кластерного решения

cluster generate clustdes = groups (2), name (myclust)

// смотрим среднее и стандартное отклонение по двум кластерам

tabstat internet secure export import telephone hightech mobile, statistics( mean sd ) by(clustdes)

Результат:



* Переменная internet (*процент пользователей интернета*): для первого кластера среднее значение **50%**, для второго **81%**.
* Переменная secure (*защищенные интернет-серверы на 1 млн чел.*): в первом кластере среднее значение **123**, во втором кластере **1334**.
* Переменная export (*экспорт услуг по коммуникации, % от экспорта услуг*): в первом кластере среднее значение **28%**, во втором кластере **45%**.
* Переменная import (*импорт услуг по коммуникации, % от импорта услуг*): в первом кластере среднее значение **31%**, во втором **48%**.
* Переменная telephone (*процент стац. телефонных линий*): в первом кластере среднее значение **14%**, во втором кластере **37%**.
* Переменная hightech (*экспорт высоких технологий, US$*): в первом кластере среднее значение **4 230 000 000**, во втором **45 600 000 000**.
* Переменная mobile (*процент абонентов мобильной сотовой связи*): в первом кластере **120**, во втором **119** (примерно одинаково).

Можно сделать вывод о том, что страны первого кластера находятся на низком уровне развития современных информационно-коммуникационных технологий. Во втором же кластере все показатели достаточно высокие, в разы превосходят значения для первого кластера. Исключение составляет только мобильная связь, показатели практически одинаковые.

**4. Дайте названия кластерам.**

Кластер 1: высокий уровень развития информационно-коммуникационных технологий (high\_level)

Кластер 2: невысокий/средний уровень развития информационно-коммуникационных технологий (low\_mid\_level)

**5. Сохраните переменную принадлежности к кластеру в файле.**

// сохранение кластерного решения

cluster generate clustdes = groups (2), name (myclust)

//перечисление принадлежности к кластерам

list country\_name clustdes

gsort clustdes

by clustdes: list country\_name

**Задание 2**

Откройте файл **video\_games\_sales.dta**

Проведите кластерный анализ данных с использованием не менее пяти интервальных переменных. Используйте метод **k-средних**.

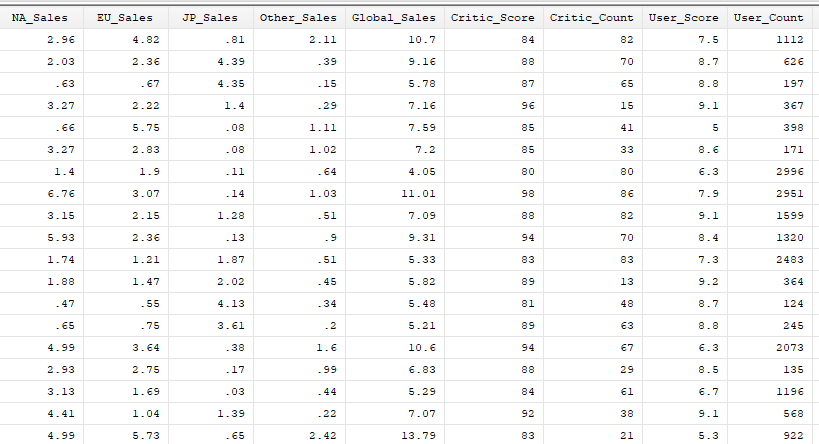
Список переменных:



Интервальные переменные: NA\_Sales, EU\_Sales, Other\_Sales, Global\_Sales, Critic\_Score, Critic\_Count, User\_Score, User\_Count

**1. Определите число кластеров, объясните своё решение.**

Проведем стандартизацию переменных



// стандартизация всех переменных

egen zna\_sales = std(NA\_Sales)

egen zeu\_sales = std(EU\_Sales)

egen zjp\_sales = std(JP\_Sales)

egen zother\_sales = std(Other\_Sales)

egen zglobal\_sales = std(Global\_Sales)

egen zcritic\_score = std(Critic\_Score)

egen zcritic\_count = std(Critic\_Count)

egen zuser\_score = std(User\_Score)

egen zuser\_count = std(User\_Count)

Чтобы определить оптимальное число кластеров, попробуем разное количество и оценим индекс Цалиньски-Харабаша

*2 кластера*

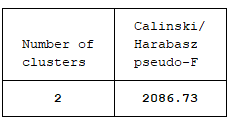
// проведение анализа методом k-средних

cluster kmeans zna\_sales zeu\_sales zjp\_sales zother\_sales zglobal\_sales zcritic\_score zcritic\_count zuser\_score zuser\_count , k(2) measure(L2) start(krandom)

cluster name: \_clus\_1

// индекс Цалиньски-Харабаша

cluster stop, rule(calinski)



3 кластера

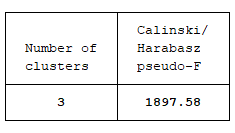
// проведение анализа методом k-средних

cluster kmeans zna\_sales zeu\_sales zjp\_sales zother\_sales zglobal\_sales zcritic\_score zcritic\_count zuser\_score zuser\_count , k(3) measure(L2) start(krandom)

cluster name: \_clus\_2

// индекс Цалиньски-Харабаша

cluster stop, rule(calinski)



4 кластера

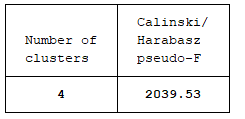
// проведение анализа методом k-средних

cluster kmeans zna\_sales zeu\_sales zjp\_sales zother\_sales zglobal\_sales zcritic\_score zcritic\_count zuser\_score zuser\_count , k(4) measure(L2) start(krandom)

cluster name: \_clus\_3

// индекс Цалиньски-Харабаша

cluster stop, rule(calinski)



5 кластеров

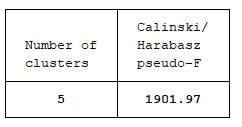
// проведение анализа методом k-средних

cluster kmeans zna\_sales zeu\_sales zjp\_sales zother\_sales zglobal\_sales zcritic\_score zcritic\_count zuser\_score zuser\_count , k(5) measure(L2) start(krandom)

cluster name: \_clus\_4

// индекс Цалиньски-Харабаша

cluster stop, rule(calinski)



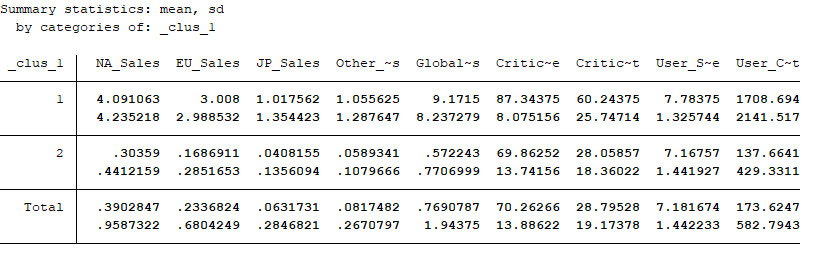
Оптимально выбрать 2 кластера, так как индекс Цалиньски-Харабаша для этого числа максимальный. Для 3 кластеров индекс заметно меньше. Для 4 кластеров индекс чуть меньше, чем для 2, но мы возьмем 2 кластера.

**2. Опишите получившиеся кластеры с помощью описательной статистики. Нужно не просто скопировать таблицы, но детально описать полученные результаты.**

// смотрим среднее и дисперсию по двум кластерам

tabstat NA\_Sales EU\_Sales JP\_Sales Other\_Sales Global\_Sales Critic\_Score Critic\_Count User\_Score User\_Count , statistics( mean sd ) by(\_clus\_1)

Результат:



* Переменная NA\_Sales (*продажи игр в Северной Америке*): для первого кластера среднее значение **4.1**, для второго **0.3**.
* Переменная EU\_Sales (*продажи игр в ЕС*): в первом кластере среднее значение **3**, во втором кластере **0.17**.
* Переменная JP\_Sales (*продажи игр в Японии*): в первом кластере среднее значение **1**, во втором кластере **0.04**.
* Переменная Other\_Sales (*продажи игр в остальном мире, в т.ч. Африка, Азия, кроме Японии и т.д.*): в первом кластере среднее значение **1.1**, во втором **0.06**.
* Переменная Global\_Sales (*продажи в мире*): в первом кластере среднее значение **9.2**, во втором кластере **0.6**.
* Переменная Critic\_Score (*совокупный балл/рейтинг, составленный Metacritic*): в первом кластере среднее значение **87**, во втором **70**.
* Переменная Critic\_Count (*число критиков*): в первом кластере среднее значение **60**, во втором **28**.
* Переменная User\_Score (*балл/рейтинг от подписчиков Metacritic*): в первом кластере среднее значение **7.8**, во втором **7.2**.
* Переменная User\_Count (*число пользователей, кто ставил баллы*): в первом кластере среднее значение **1708**, во втором **137**.

Игры из первого кластера лучше продаются в Северной Америке, Европе и Японии, да и в целом в мире. Также они имеют высокий рейтинг критиков (по этому критерию игры из другого кластера немного отстают). Игры из второго кластера имеют примерно такие же оценки пользователей, что и игры из первого кластера, однако количество проголосовавших пользователей в первом кластере намного больше, поэтому нельзя сказать, что эти оценки можно сравнивать (выборки не соизмеримы).

Такая ситуация может быть связана с тем, что игры из первого кластера выпускают какие-то более крупные студии, известные всему миру. Эти игры имеют локализации на многих языках, в них играет большое количество пользователей. Игры из второго кластера, скорее всего, выпускаются локальными студиями, и не исключено, что для немногочисленного числа языков.

**3. Дайте названия кластерам.**

Кластер 1: популярные видеоигры (popular\_games)

Кластер 2: менее популярные видеоигры (unpopular\_games)

**4. Сохраните переменную принадлежности к кластеру в файле.**

Сохраняется автоматически при кластеризации

//перечисление принадлежности к кластерам

list Name \_clus\_1

gsort \_clus\_1

by \_clus\_1: list Name

Загрузите, пожалуйста, этот файл с ответами и изменённые файлы с данными в MS Teams в раздел «Задания».