РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ

Факультет физико-математических и естественных наук

Кафедра прикладной информатики и теории вероятностей

Практикум по математической статистике

Лабораторная работа №2

Тема: «Кластерный анализ»

Вариант 10

Выполнил

Студент: Феоктистов Владислав

Группа: НПМбд-01-19б

№ с/б: 1032192939

Преподаватель: Матюшенко Сергей Иванович

МОСКВА

2022 г.

Цель работы: приобрести практические навыки применения кластерного анализа для решения конкретных задач с использованием статистического пакета SPSS.

Ход работы:

1. Повторил теоретические основы кластерного анализа, используя материалы учебного пособия.
2. Разобрал пример реализации кластерного анализа в SPSS.
3. В разделе «Задачи для самостоятельного решения» выбрал задачу 3.10 «Классификация государств по социально-экономическим и демографическим признакам», соответствующую моему варианту и импортировал данные из файла «2.1.4. Данные для д.з. №2.xls» листа «Задача 10».

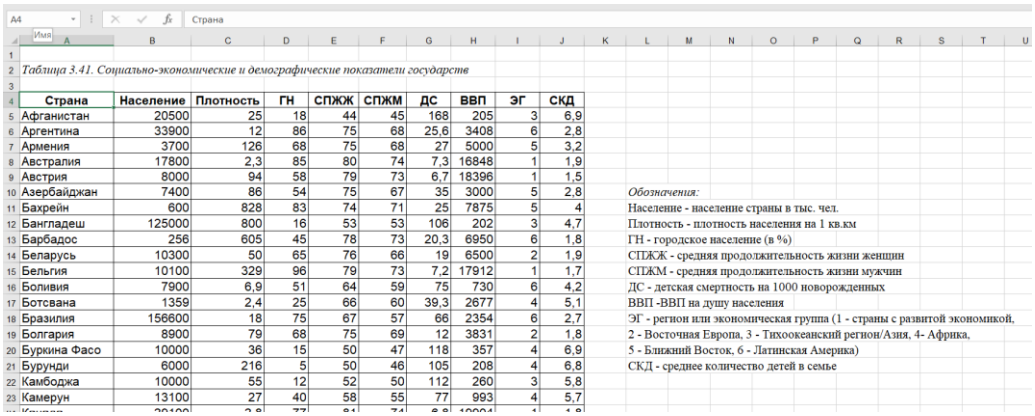
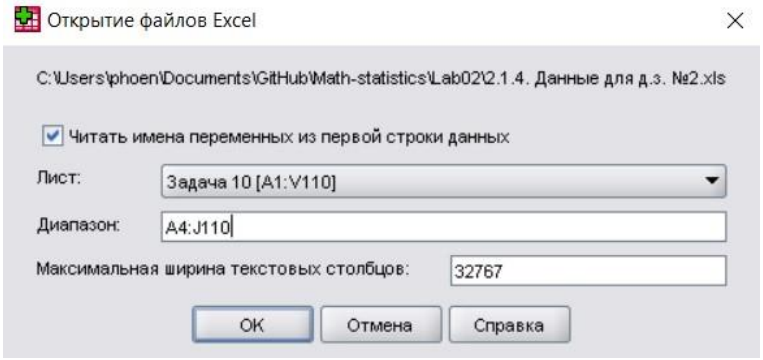


Таблица 3.41. Социально-экономические и демографические показатели государств

Страна	Население	Плотность	ГН	СПЖЖ	СПЖМ	ДС	ВВП	ЭГ	СКД
Афганистан	20500	25	18	44	45	168	205	3	6,9
Аргентина	33900	12	86	75	68	25,8	3408	6	2,8
Армения	3700	126	68	75	68	27	5000	5	3,2
Австралия	17800	2,3	85	80	74	7,3	16848	1	1,9
Австрия	8000	94	58	79	73	6,7	18396	1	1,5
Азербайджан	7400	86	54	75	67	35	3000	5	2,8
Бахрейн	600	828	83	74	71	25	7875	5	4
Бангладеш	125000	800	16	53	53	106	202	3	4,7
Барбадос	256	605	45	78	73	20,3	6950	6	1,8
Беларусь	10300	50	65	76	66	19	6500	2	1,9
Бельгия	10100	329	96	79	73	7,2	17912	1	1,7
Боливия	7900	6,9	51	64	59	75	730	6	4,2
Ботсвана	1359	2,4	25	66	60	39,3	2677	4	5,1
Бразилия	156600	18	75	67	57	66	2354	6	2,7
Болгария	8900	79	68	75	69	12	3831	2	1,8
Буркина Фасо	10000	36	15	50	47	118	357	4	6,9
Бурунди	6000	216	5	50	46	105	208	4	6,8
Камбоджа	10000	55	12	52	50	112	260	3	5,8
Камерун	13100	27	40	58	55	77	993	4	5,7

Обозначения:
Население - население страны в тыс. чел.
Плотность - плотность населения на 1 кв.км
ГН - городское население (в %)
СПЖЖ - средняя продолжительность жизни женщин
СПЖМ - средняя продолжительность жизни мужчин
ДС - детская смертность на 1000 новорожденных
ВВП -ВВП на душу населения
ЭГ - регион или экономическая группа (1 - страны с развитой экономикой, 2 - Восточная Европа, 3 - Тихоокеанский регион/Азия, 4- Африка, 5 - Ближний Восток, 6 - Латинская Америка)
СКД - среднее количество детей в семье

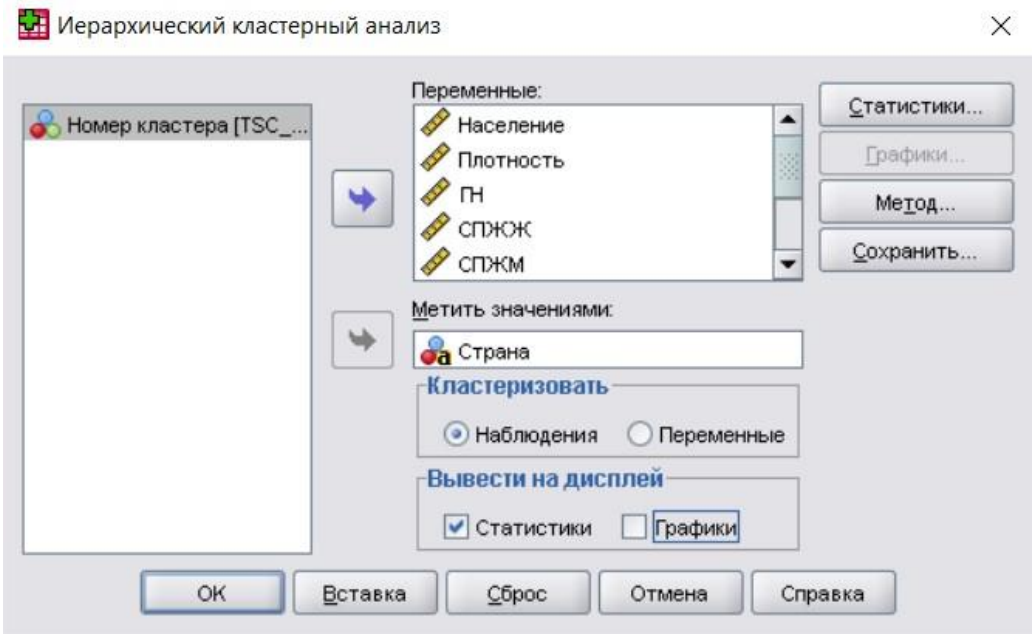


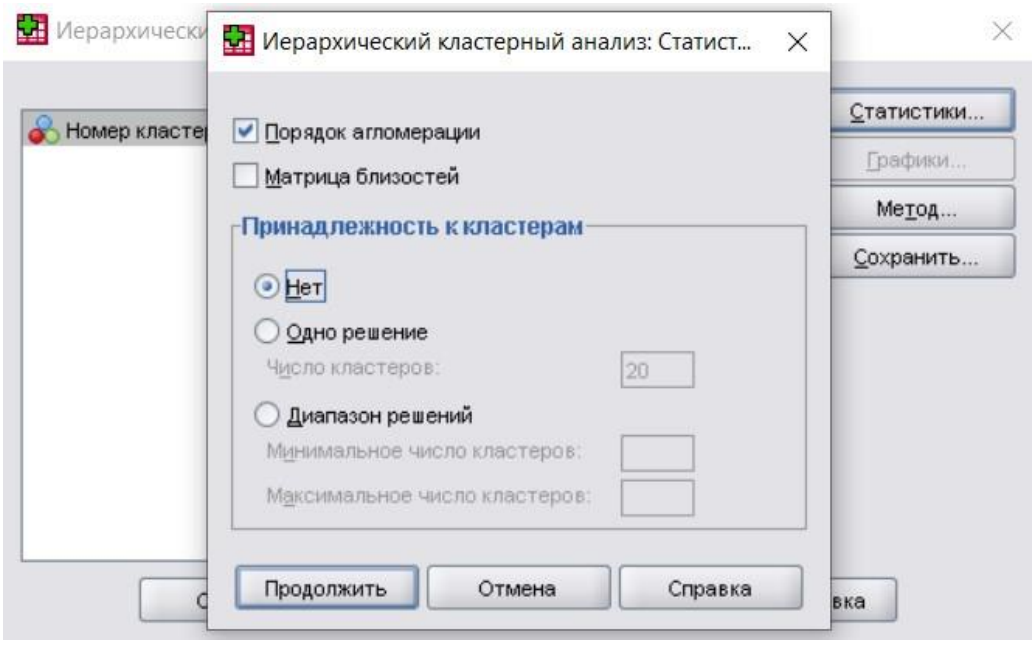
	Страна	Население	Плотность	ГН	СПЖЖ	СПЖМ	ДС	ВВП	ЭГ	СКД
1	Афганистан	20500	25	18	44	45	168	205	3	6,9
2	Аргентина	33900	12	86	75	68	26	3408	6	2,8
3	Армения	3700	126	68	75	68	27	5000	5	3,2
4	Австралия	17800	2	85	80	74	7	16848	1	1,9
5	Австрия	8000	94	58	79	73	7	18396	1	1,5
6	Азербайджан	7400	86	54	75	67	35	3000	5	2,8
7	Бахрейн	600	828	83	74	71	25	7875	5	4,0
8	Бангладеш	125000	800	16	53	53	106	202	3	4,7
9	Барбадос	256	605	45	78	73	20	6950	6	1,8
10	Беларусь	10300	50	65	76	66	19	6500	2	1,9
11	Бельгия	10100	329	96	79	73	7	17912	1	1,7
12	Боливия	7900	7	51	64	59	75	730	6	4,2
13	Ботсвана	1359	2	25	66	60	39	2677	4	5,1
14	Бразилия	156600	18	75	67	57	66	2354	6	2,7
15	Болгария	8900	79	68	75	69	12	3831	2	1,8
16	Буркина Фасо	10000	36	15	50	47	118	357	4	6,9
17	Бурунди	6000	216	5	50	46	105	208	4	6,8
18	Камбоджа	10000	55	12	52	50	112	260	3	5,8
19	Камерун	13100	27	40	58	55	77	993	4	5,7
20	Канада	29100	3	77	81	74	7	19904	1	1,8
21	ЦАР	3300	5	47	44	41	137	457	4	5,4
22	Чили	14000	18	85	78	71	15	2591	6	2,5
23	Китай	1205200	124	26	69	67	52	377	3	1,8
24	Колумбия	35600	31	70	75	69	28	1538	6	2,5
25	Коста Рика	3300	64	47	79	76	11	2031	6	3,1
26	Хорватия	4900	85	51	77	70	9	5487	2	1,7
27	Куба	11100	99	74	78	74	10	1382	6	1,9

	Имя	Тип	Ширина	Десятич...	Метка	Значения	Пропуски	Столбцы	Выравнивание	Шкала
1	Страна	Текстовая	12	0		Нет	Нет	12	По левом...	Номинальная
2	Население	Числовая	11	0		Нет	Нет	11	По право...	Количестве...
3	Плотность	Числовая	11	0		Нет	Нет	11	По право...	Количестве...
4	ГН	Числовая	11	0		Нет	Нет	11	По право...	Количестве...
5	СПЖЖ	Числовая	11	0		Нет	Нет	11	По право...	Количестве...
6	СПЖМ	Числовая	11	0		Нет	Нет	11	По право...	Количестве...
7	ДС	Числовая	11	0		Нет	Нет	11	По право...	Количестве...
8	ВВП	Числовая	11	0		Нет	Нет	11	По право...	Количестве...
9	ЭГ	Числовая	11	0		Нет	Нет	11	По право...	Количестве...
10	СКД	Числовая	11	1		Нет	Нет	11	По право...	Количестве...

В первую очередь, нужно провести кластеризацию государств на оптимально число кластеров. Для этого построим таблицу «Шаги агломерации» и определим шаг, на котором происходит резкий скачок коэффициента, исключая первый шаги, чтобы не получить слишком больше число кластеров, содержащих 1-2 страну.

Выбираем Анализ – Классификация – Иерархическая кластеризация. Графики строить нет необходимости, поэтому галочку перед «Графики» можно убрать, в панели «Статистики…» нужно установить галочку перед «Порядок агломерации» и выбрать пункт «Нет» для принадлежности к кластерам (см. рис. ниже).





В итоге, после нажатия на кнопку «ОК» получаем такую таблицу:

Шаги агломерации

Этап	Кластер объединен с		Коэффициенты	Этап первого появления кластера		Следующий этап
	Кластер 1	Кластер 2		Кластер 1	Кластер 2	
1	16	18	9830,210	0	0	34
2	32	44	82323,210	0	0	12
3	12	29	126270,620	0	0	26
4	21	61	163799,960	0	0	30
5	86	106	190640,360	0	0	14
6	30	42	230089,890	0	0	11
7	56	60	281175,090	0	0	23
8	17	43	281464,810	0	0	33
9	67	77	344043,180	0	0	19
10	58	73	447087,530	0	0	22
11	27	30	449165,265	0	6	34
12	32	76	488141,145	2	0	33
13	1	100	510204,010	0	0	55
14	83	86	515271,330	0	5	26
15	26	62	549834,420	0	0	36
16	69	97	560128,860	0	0	54
17	25	75	624948,290	0	0	25
18	57	66	707831,410	0	0	39
19	67	103	778139,070	9	0	41
20	22	93	839347,200	0	0	49
21	41	80	895246,000	0	0	66
22	58	59	901098,565	10	0	37
23	56	70	907831,875	7	0	30
24	7	9	1025181,930	0	0	46
25	25	102	1062127,495	17	0	47
26	12	83	1070613,617	3	14	50
27	28	72	1094470,180	0	0	51
28	64	104	1241202,920	0	0	29
29	51	64	1250363,580	0	28	55
30	21	56	1261139,250	4	23	47
31	6	88	1316712,250	0	0	52
32	31	95	1472217,250	0	0	65
33	17	32	1546283,885	8	12	52
34	16	27	1551780,405	1	11	50

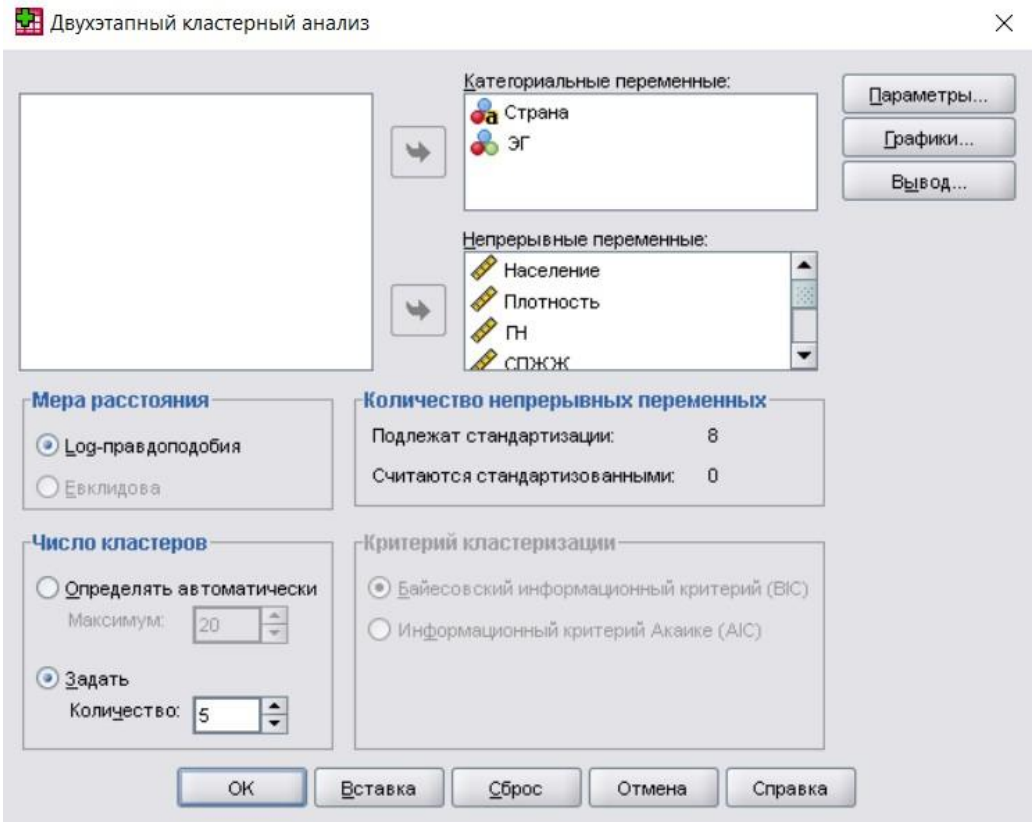
35	10	46	1608766,260	0	0	58
36	26	39	1665552,310	15	0	45
37	33	58	2043511,427	0	22	46
38	36	54	2102043,060	0	0	53
39	57	94	2325271,425	18	0	84
40	13	37	2586282,540	0	0	62
41	67	81	2693655,800	19	0	64
42	11	91	2810398,410	0	0	44
43	3	63	2938839,440	0	0	45
44	5	11	3792625,825	0	42	74
45	3	26	3878973,007	43	36	69
46	7	33	3943217,058	24	37	69
47	21	25	4027173,312	30	25	59
48	52	53	4079041,080	0	0	54
49	19	22	4348095,000	0	20	78
50	12	16	4936720,982	26	34	67
51	28	35	4972544,370	27	0	71
52	6	17	5206572,817	31	33	67
53	36	98	5710031,330	38	0	93
54	52	69	5750614,710	48	16	71
55	1	51	5999778,558	13	29	64
56	4	68	6049905,780	0	0	88
57	2	24	6387523,850	0	0	72
58	10	15	6828119,630	35	0	66
59	21	38	8142412,715	47	0	62
60	45	87	9609286,260	0	0	77
61	8	74	1,008E7	0	0	95
62	13	21	1,021E7	40	59	75
63	78	105	1,130E7	0	0	90
64	1	67	1,203E7	55	41	83
65	31	96	1,261E7	32	0	76
66	10	41	1,336E7	58	21	79
67	6	12	1,348E7	52	50	79
68	84	89	1,362E7	0	0	87
69	3	7	1,509E7	45	46	75
70	34	101	1,649E7	0	0	86
71	28	52	1,885E7	51	54	73
72	2	79	2,026E7	57	0	84

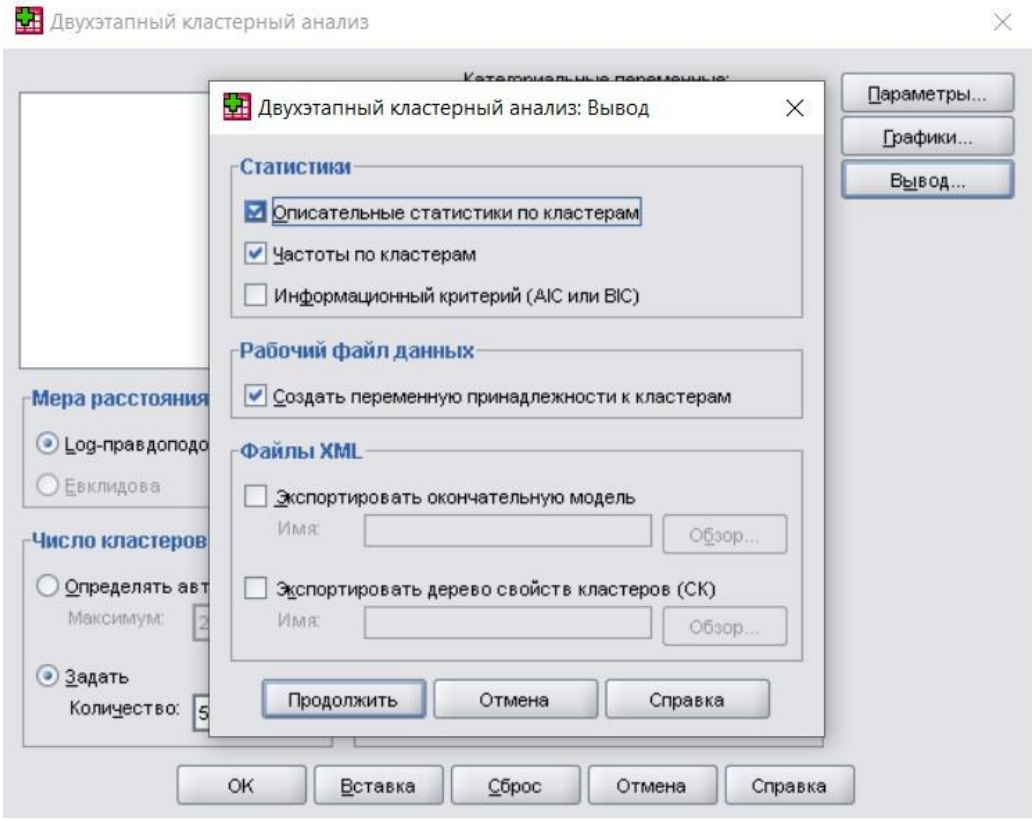
73	28	47	2,608E7	71	0	77
74	5	92	2,663E7	44	0	81
75	3	13	2,884E7	69	62	82
76	31	50	2,899E7	65	0	86
77	28	45	3,416E7	73	60	81
78	19	85	3,863E7	49	0	83
79	6	10	4,252E7	67	66	82
80	65	71	5,073E7	0	0	96
81	5	28	5,075E7	74	77	88
82	3	6	5,440E7	75	79	91
83	1	19	5,547E7	64	78	91
84	2	57	6,381E7	72	39	89
85	14	82	7,348E7	0	0	97
86	31	34	8,062E7	76	70	90
87	84	90	9,774E7	68	0	89
88	4	5	1,484E8	56	81	92
89	2	84	1,789E8	84	87	94
90	31	78	1,799E8	86	63	93
91	1	3	2,234E8	83	82	92
92	1	4	2,665E8	91	88	99
93	31	36	3,211E8	90	53	98
94	2	20	3,566E8	89	0	98
95	8	55	3,861E8	61	0	97
96	40	65	4,450E8	0	80	100
97	8	14	8,237E8	95	85	100
98	2	31	8,300E8	94	93	99
99	1	2	1,957E9	92	98	102
100	8	40	2,501E9	97	96	102
101	49	99	4,253E9	0	0	103
102	1	8	1,151E10	99	100	103
103	1	49	4,425E10	102	101	105
104	23	48	8,620E10	0	0	105
105	1	23	1,083E12	103	104	0

По таблице наблюдаем резкий скачок с 101 шага по 102 (более чем в 2 раза). Как уже было сказано ранее, начальные шаги исключаем из выбора. В итоге, получаем 106 – 101 = 5 кластеров, где 106 – число наблюдений, т.е. 106 стран; 101 – шаг, с которого начинается резкий скачок.

Далее проведем двухэтапный кластерный анализ для того, чтобы разбить страны на кластеры, дать содержательную интерпретацию полученных сегментов на основе исследования кластерных профилей (они будут получены автоматически при двухэтапном кластерном анализе) и определить характерные особенности экономических групп по их принадлежности у полученным сегментам.

Выбираем Анализ – Классификация – Двухэтапный кластерный анализ. В качестве категориальный переменных выберем: «Страна», «ЭГ»; а в качестве непрерывных – все остальные. Также нужно в разделе «Число кластеров» установить пункт перед «Задать» и указать оптимальное число кластеров, которое мы нашли. В панели «Вывод…» стоит поставить галочку поставить галочку перед «Создать переменную принадлежности к кластерам», чтобы увидеть, какое итоговое разбиение по кластерам мы получили.





В итоге, получаем вот такие результаты:

Распределение по кластерам

		N	% объединенных	% от итога
Кластер	1	24	22,6%	22,6%
	2	11	10,4%	10,4%
	3	17	16,0%	16,0%
	4	22	20,8%	20,8%
	5	32	30,2%	30,2%
	Объединенный	106	100,0%	100,0%
Итог		106		100,0%

Центроиды

Кластер	Население		Плотность		ГН		СПЖЖ		СПЖМ		ДС		ВВП		СКД	
	Среднее	Стд. отклонение	Среднее	Стд.	Среднее	Стд.	Среднее	Стд.	Среднее	Стд.	Среднее	Стд.	Среднее	Стд.	Среднее	Стд.
1	27728,54	37439,921	101,48	168,480	27,63	14,185	53,54	7,553	50,54	6,871	99,39	29,708	851,29	1083,643	6,054	1,0632
2	237736,36	414538,568	1064,27	1950,253	48,64	28,939	70,82	6,014	66,00	4,940	38,14	23,699	4043,91	5637,260	2,618	,9621
3	17640,06	22509,394	126,66	200,569	65,94	21,614	71,59	4,501	67,35	3,920	41,39	19,179	4957,41	4057,448	4,724	1,5356
4	37285,77	59149,030	118,07	115,482	74,82	14,539	80,18	1,220	73,82	1,220	6,80	1,262	16758,55	3701,571	1,741	,2423
5	23912,69	38608,150	77,43	108,870	63,53	13,870	74,03	4,076	67,34	4,240	28,97	17,567	3231,13	2133,757	2,716	,9579
Объединенный	48735,46	148677,126	201,62	684,161	56,58	24,309	69,94	10,644	64,75	9,343	43,25	38,192	5861,08	6563,660	3,581	1,9031

По этим данным можно провести некоторый анализ кластеров.

К первому сегменту можно отнести 22,6% стран мира и для них характерна средняя численность населения со средним стандартным отклонением, ниже среднего плотность населения, низкий процент городского населения, короткая продолжительность жизни с сильным стандартным отклонением, низкий ВВП, высокая рождаемость и смертность детей.

Ко второму сегменту относятся 10,4% - наименьшая группа стран. В первую очередь, стоит отметить, что у этой группы самые большие стандартные отклонения по всем параметрам. Таким странам свойственна большая численность населения с высокой плотностью, половина граждан проживает в городах, средняя продолжительность жизни, околосредняя смертность среди детей, чуть ниже среднего ВВП, немного ниже среднего рождаемость, но все еще положительная динамика роста населения.

К третьему сегменту относится 16,0% государств. Для них характерна довольно низкая численность населения, ниже среднего плотность населения, выше среднего процент городского населения, средняя продолжительность жизни, средняя смертность среди детей, средний ВВП, но с сильным достаточно сильным разбросом, выше среднего рождаемость, достаточно быстрый естественный прирост населения.

К четвертому сегменту относится 20,8% процентов стран – третья по размеру группа. Также в первую очередь стоит заметить, что для них характерна низкое стандартное отклонение по почти всем показателям. Средняя численность населения, ниже среднего плотность населения, высокий процент населения, высокая продолжительность жизни, низкая детская смертность и рождаемость, очень высокий ВВП. Отрицательный естественный прирост населения – вырождаемость коренного населения.

К пятому сегменту относится самая большая группа стран – 30,2%. Всем им характерна средняя или выше среднего стандартное отклонение. Численность населения маленькая, как и плотность, чуть выше среднего процент городского населения, выше среднего продолжительность жизни, значительно ниже среднего детская смертность, ниже среднего ВВП, немного ниже среднего рождаемость, но все еще положительный естественный прирост населения.

Таким образом, можно кратко охарактеризовать группы этих стран:

1 группа – бедные страны с низким уровнем жизни и нестабильной обстановкой и высокой рождаемость, однако эти нивелируется высокой смертностью детей. С большой вероятность низкий уровень образования населения.

2 группа – быстро развивающиеся страны с очень сильными изменениями, уровень жизни распределен неравномерно, но в целом все хорошо, сохраняется естественный рост населения.

3 группа – страны середнячки, которые скорее относятся к развивающимся странам. Быстрый естественный прирост населения. Сложно сказать, что население там живет в достатке.

4 группа - богатые развитый страны с высокими уровнем жизни и положительными показателями почти по всем критериям. Однако таким странам характерно вырождение коренного населения.

5 группа – развивающейся страны, которые переходят постепенно уже в 4 группу. Средние почти во всем, низкая плотность населения. Стабильная жизнь, примерно обеспеченная жизнь, возможны кризисы и стагнации экономики, это и отличает ее от 2ой группы.

Стоит отметь, что 3 и 5 группу отличает то, что 5 группа имеет более однородные показатели среди населения и меньшие стандартные отклонения.

Можно прийти к некоторому выводу: развивающие страны мира пытаются поднимать в целом показатели и уменьшать разрыв среди населения, чтобы перейти в категорию развитых стран. Однако это приводит к уменьшению рождаемости детей.

Вывод: приобрёл практические навыки применения кластерного анализа для решения конкретных задач с использованием статистического пакета SPSS.