МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
Кафедра штучного інтелекту

Звіт з виконання завдань
комп’ютерного практикуму № 4
Кореляційно-регресійний аналіз у середовищі STATISTICA
з кредитного модуля «Багатовимірний статистичний аналіз»

Звіт склав
студент гр. КІ-01
Копцов В.О.
Прийняла: *Ірина Джигирей*

Київ 2022

Мета роботи. Дослідити процедуру класифікування методом k-середніх.

Завдання. Виконати розрахунки згідно індивідуального завдання і набути вмінь кластерного аналізу сукупності об’єктів методом k-середніх.

Варіант №30

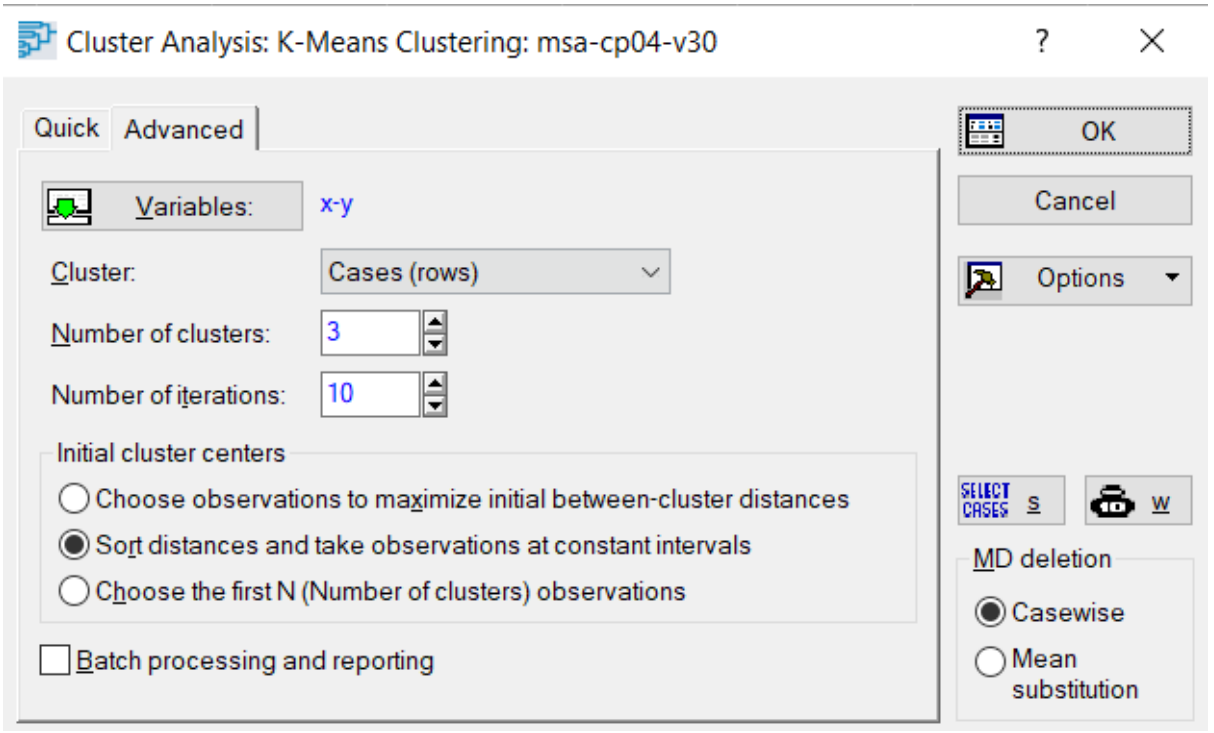
Хід виконання завдань практикуму

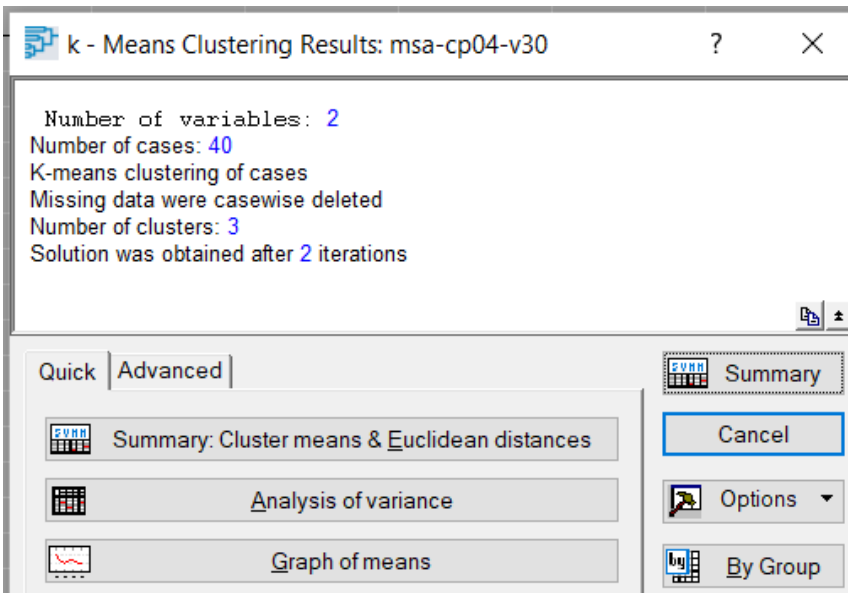
1. Завантажив данні, додав нормований варіант

Data: msa-cp04-v30* (4v by 40c)

	1 x	2 y	3 x_n	4 y_n
1	5,98	39,14	-0,59	1,00
2	8,77	26,12	0,72	-1,11
3	4,67	39,41	-1,21	1,04
4	8,82	26,90	0,74	-0,98
5	10,02	24,97	1,30	-1,30
6	6,21	36,62	-0,48	0,59
7	9,55	29,77	1,08	-0,52
8	10,59	21,76	1,57	-1,82
9	4,47	40,83	-1,30	1,27
10	7,99	34,40	0,35	0,23
11	7,84	27,43	0,28	-0,90
12	4,89	36,67	-1,10	0,60
13	10,05	27,62	1,32	-0,87
14	5,93	36,83	-0,62	0,63
15	5,14	36,23	-0,99	0,53
16	5,18	34,63	-0,97	0,27
17	5,62	35,15	-0,76	0,35
18	7,73	30,94	0,23	-0,33
19	11,61	20,19	2,05	-2,07
20	4,02	42,44	-1,51	1,54

Діалогове вікно початкових налаштувань кластерного аналізу





Матриця міжкластерних евклідових(під діагоналлю) і квадратичних евклідових(над діагоналлю)

Cluster Number	Euclidean Distances between Clusters (msa-cp04-v30) Distances below diagonal Squared distances above diagonal		
	No. 1	No. 2	No. 3
No. 1	0,00	23,2	93,0
No. 2	4,81	0,0	23,7
No. 3	9,64	4,9	0,0

Описова статистика першого кластеру

Variable	Descriptive Statistics for Cluster 1 (msa-cp04-v30) Cluster contains 20 cases		
	Mean	Standard Deviation	Variance
x	5,43	0,865	0,749
y	38,11	2,483	6,167

Описова статистика другого кластеру

Variable	Descriptive Statistics for Cluster 2 (msa-cp04-v30) Cluster contains 8 cases		
	Mean	Standard Deviation	Variance
x	8,11	0,957	0,916
y	31,85	1,949	3,799

Описова статистика третього кластеру

Variable	Descriptive Statistics for Cluster 3 (msa-cp04-v30) Cluster contains 12 cases		
	Mean	Standard Deviation	Variance
x	9,68	1,111	1,233
y	25,15	2,467	6,084

Результати дисперсійного аналізу

Variable	Analysis of Variance (msa-cp04-v30)					
	Between SS	df	Within SS	df	F	signif. p
x	143,11	2	34,21	37	77,40	0,000000
y	1272,22	2	210,69	37	111,71	0,000000

Відстанями кожного об’єкту до центру для першого кластеру

	Members of Cluster Number 1 (msa-cp04-v30) and Distances from Respective Cluster Center Cluster contains 20 cases
	Distance
1	0,828
3	1,066
6	1,188
9	2,041
12	1,085
14	0,970
15	1,343
16	2,465
17	2,095
20	3,222
22	3,476
23	0,487
25	1,591
29	0,669
31	1,181
32	1,963
33	1,065
35	3,047
38	0,935
39	1,302

Відстанями кожного об’єкту до центру для другого кластеру

	Members of Cluster Number 2 (msa-cp04-v30) and Distances from Respective Cluster Center Cluster contains 8 cases
	Distance
7	1,789
10	1,803
18	0,700
26	1,878
28	1,266
30	1,812
34	1,045
37	0,236

Відстанями кожного об’єкту до центру для третього кластеру

	Members of Cluster Number 3 (msa-cp04-v30) and Distances from Respective Cluster Center Cluster contains 12 cases
	Distance
2	0,942
4	1,381
5	0,271
8	2,480
11	2,074
13	1,768
19	3,761
21	2,382
24	0,233
27	1,629
36	0,453
40	1,031

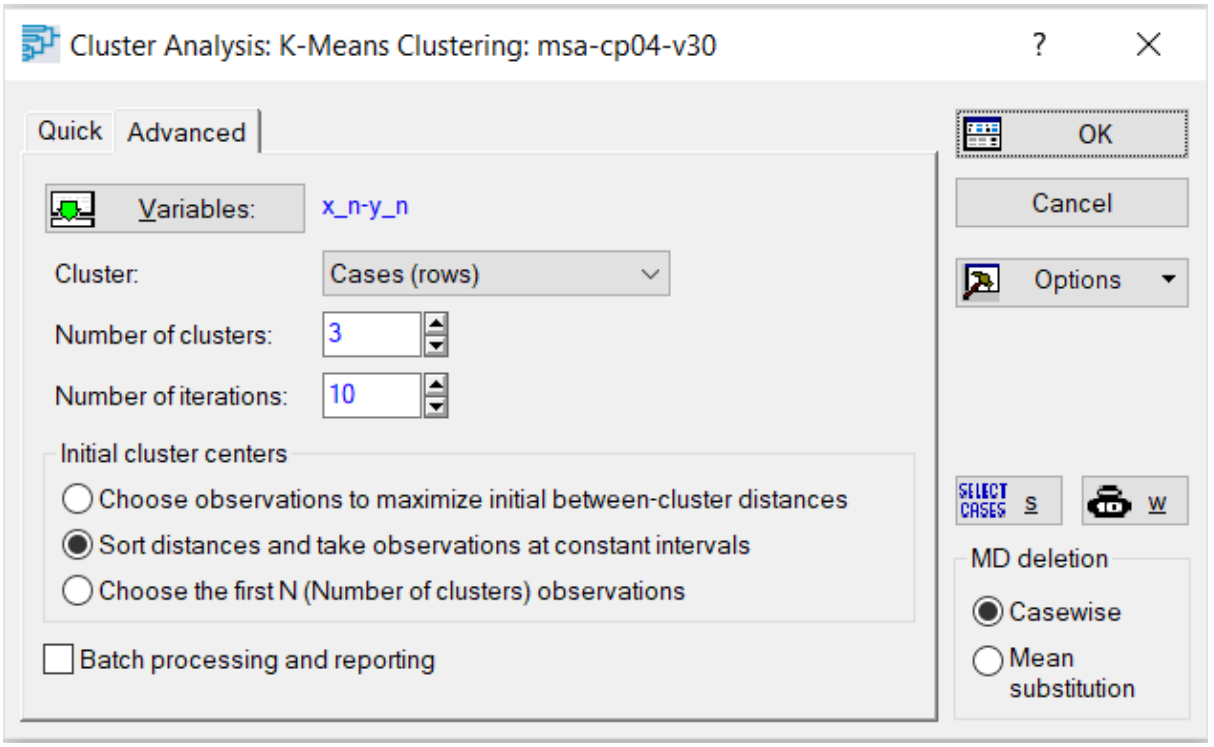
Підрахунок значення критерію якості кластеризування за допомогою Google Table

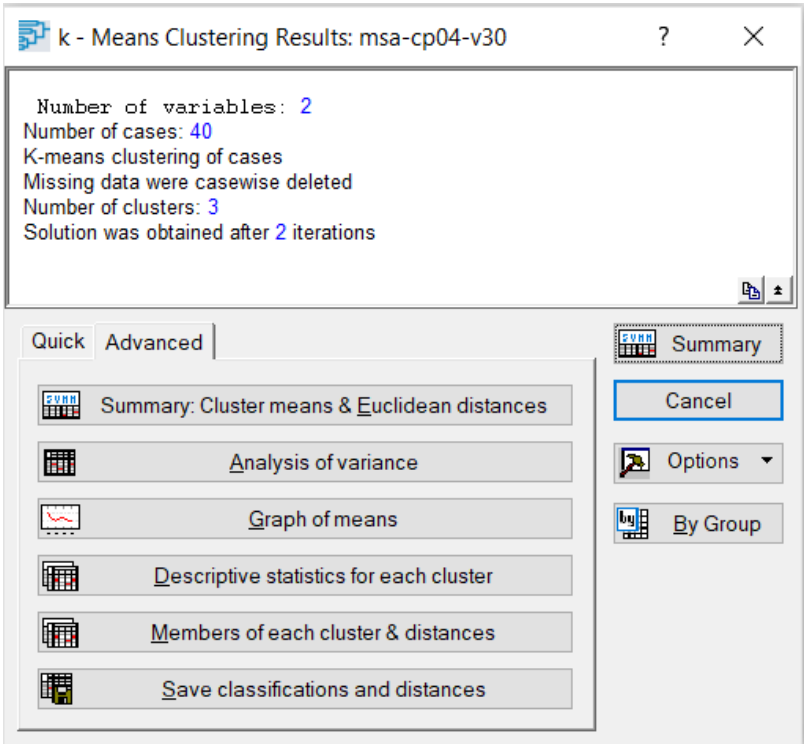
A	B	C	D	E	F	G	H	I	J	K
0,828	1,789	0,942								
1,066	1,803	1,381								
1,188	0,7	0,271		Внутрішньокластерні відстані		1	2	3	4	SUM
2,041	1,878	2,48				32,019	10,529	18,405	0	60,953
1,085	1,266	2,074				20	8	12	0	40
0,97	1,812	1,768								
1,343	1,045	3,761								
2,465	0,236	2,382								
2,095		0,233		0	23,2	93				
3,222		1,629		4,81	0	23,7				
3,476		0,453		9,64	4,9	0			SUM	
0,487		1,031							19,35	
1,591									3	
0,669										
1,181									Result	
1,963									0,236251938	
1,065										
3,047										
0,935										
1,302										

Результат: 0,236

Повторив таке саме для нормалізованих даних

Діалогове вікно початкових налаштувань кластерного аналізу





Матриця міжкластерних евклідових(під діагоналлю) і квадратичних евклідових(над діагоналлю)

Cluster Number	Euclidean Distances between Clusters (msa-cp04-v30) Distances below diagonal Squared distances above diagonal		
	No. 1	No. 2	No. 3
No. 1	0,00	2,00	5,02
No. 2	1,42	0,00	0,68
No. 3	2,24	0,83	0,00

Описова статистика першого кластеру

Variable	Descriptive Statistics for Cluster 1 (msa-cp04-v30) Cluster contains 14 cases		
	Mean	Standard Deviation	Variance
x_n	1,13	0,480	0,231
y_n	-1,16	0,456	0,208

Описова статистика другого кластеру

Variable	Descriptive Statistics for Cluster 2 (msa-cp04-v30) Cluster contains 15 cases		
	Mean	Standard Deviation	Variance
x_n	-0,239	0,441	0,195
y_n	0,297	0,368	0,135

Описова статистика третього кластеру

Variable	Descriptive Statistics for Cluster 3 (msa-cp04-v30) Cluster contains 11 cases		
	Mean	Standard Deviation	**Variance**
x_n	-1,11	0,294	0,086
y_n	1,07	0,381	0,145

Результати дисперсійного аналізу

Variable	Analysis of Variance (msa-cp04-v30)					
	Between SS	df	Within SS	**df**	F	signif. p
x_n	32,41	2	6,59	37	91,0	0,000
y_n	32,95	2	6,05	37	100,7	0,000

Відстанями кожного об’єкту до центру для першого кластеру

	Members of Cluster Number 1 (msa-cp04-v30) and Distances from Respective Cluster Center Cluster contains 14 cases
	Distance
2	0,295
4	0,304
5	0,154
7	0,457
8	0,558
11	0,630
13	0,247
19	0,914
21	0,654
24	0,113
27	0,395
30	0,448
36	0,132
40	0,141

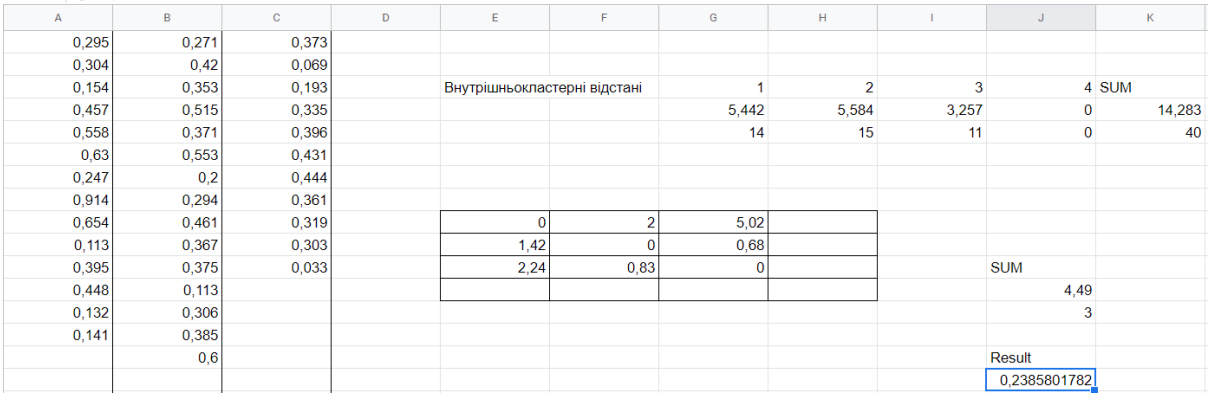
Відстанями кожного об’єкту до центру для другого кластеру

	Members of Cluster Number 2 (msa-cp04-v30) and Distances from Respective Cluster Center Cluster contains 15 cases
	Distance
6	0,271
10	0,420
14	0,353
16	0,515
17	0,371
18	0,553
25	0,200
26	0,294
28	0,461
29	0,367
31	0,375
32	0,113
33	0,306
34	0,385
37	0,600

Відстанями кожного об’єкту до центру для третього кластеру

	Members of Cluster Number 3 (msa-cp04-v30) and Distances from Respective Cluster Center Cluster contains 11 cases
	Distance
1	0,373
3	0,069
9	0,193
12	0,335
15	0,396
20	0,431
22	0,444
23	0,361
35	0,319
38	0,303
39	0,033

Підрахунок значення критерію якості кластеризування за допомогою Google Table

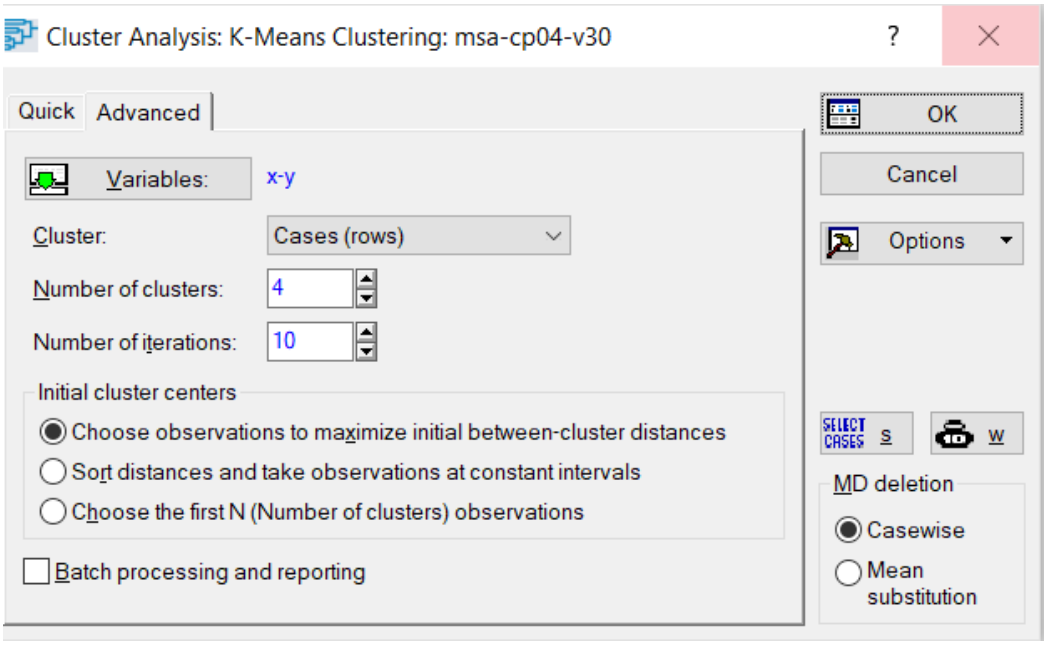


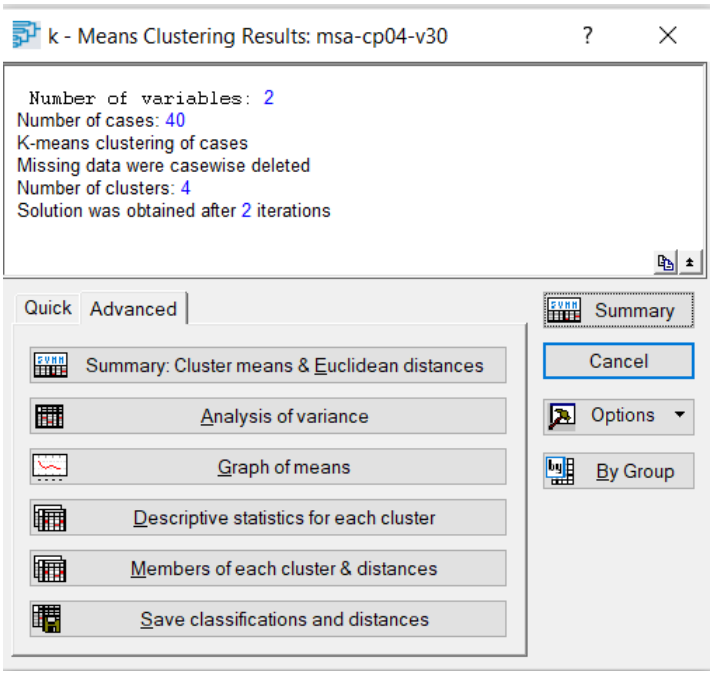
A	B	C	D	E	F	G	H	I	J	K
0,295	0,271	0,373								
0,304	0,42	0,069								
0,154	0,353	0,193		Внутрішньокластерні відстані		1	2	3	4	SUM
0,457	0,515	0,335				5,442	5,584	3,257	0	14,283
0,558	0,371	0,396				14	15	11	0	40
0,63	0,553	0,431								
0,247	0,2	0,444								
0,914	0,294	0,361								
0,654	0,461	0,319		0	2	5,02				
0,113	0,367	0,303		1,42	0	0,68				
0,395	0,375	0,033		2,24	0,83	0			SUM	
0,448	0,113								4,49	
0,132	0,306								3	
0,141	0,385									
	0,6								Result	
									0,2385801782	

Результат: 0,239

Повторив те саме для 4-ох кластерів

Діалогове вікно початкових налаштувань кластерного аналізу





Матриця міжкластерних евклідових(під діагоналлю) і квадратичних евклідових(над діагоналлю)

Cluster Number	Euclidean Distances between Clusters (msa-cp04-v30) Distances below diagonal Squared distances above diagonal			
	No. 1	No. 2	No. 3	No. 4
No. 1	0,00	88,34	166,4	15,52
No. 2	9,40	0,00	12,3	29,81
No. 3	12,90	3,50	0,0	80,32
No. 4	3,94	5,46	9,0	0,00

Описова статистика першого кластеру

Variable	Descriptive Statistics for Cluster 1 (msa-cp04-v30) Cluster contains 6 cases		
	Mean	Standard Deviation	Variance
x	10,47	0,822	0,676
y	23,33	2,266	5,136

Описова статистика другого кластеру

Variable	Descriptive Statistics for Cluster 2 (msa-cp04-v30) Cluster contains 15 cases		
	Mean	Standard Deviation	Variance
x	6,20	1,063	1,131
y	35,92	1,266	1,602

Описова статистика третього кластеру

Variable	Descriptive Statistics for Cluster 3 (msa-cp04-v30) Cluster contains 8 cases		
	Mean	Standard Deviation	Variance
x	4,87	0,732	0,536
y	40,69	1,675	2,806

Описова статистика четвертого кластеру

Variable	Descriptive Statistics for Cluster 4 (msa-cp04-v30) Cluster contains 11 cases		
	Mean	Standard Deviation	Variance
x	8,63	0,976	0,953
y	28,59	1,971	3,885

Результати дисперсійного аналізу

Variable	Analysis of Variance (msa-cp04-v30)					
	Between SS	df	Within SS	df	F	signif. p
x	144,8	3	32,49	36	53,49	0,0000
y	1376,3	3	106,60	36	154,93	0,0000

Відстанями кожного об’єкту до центру для першого кластеру

	Members of Cluster Number 1 (msa-cp04-v30) and Distances from Respective Cluster Center Cluster contains 6 cases
	Distance
5	1,203
8	1,112
19	2,362
21	0,985
24	1,571
36	1,682

Відстанями кожного об’єкту до центру для другого кластеру

	Members of Cluster Number 2 (msa-cp04-v30) and Distances from Respective Cluster Center Cluster contains 15 cases
	Distance
6	0,497
10	1,659
12	1,069
14	0,673
15	0,782
16	1,162
17	0,680
25	0,640
26	1,434
29	1,086
31	0,500
32	0,340
33	1,168
34	2,164
38	1,434

Відстанями кожного об’єкту до центру для третього кластеру

	Members of Cluster Number 3 (msa-cp04-v30) and Distances from Respective Cluster Center Cluster contains 8 cases
	Distance
1	1,349
3	0,917
9	0,299
20	1,375
22	1,611
23	1,642
35	1,194
39	0,581

Відстанями кожного об’єкту до центру для четвертого кластеру

	Members of Cluster Number 4 (msa-cp04-v30) and Distances from Respective Cluster Center Cluster contains 11 cases
	Distance
2	1,748
4	1,201
7	1,061
11	0,990
13	1,217
18	1,781
27	1,010
28	2,074
30	0,952
37	2,102
40	1,515

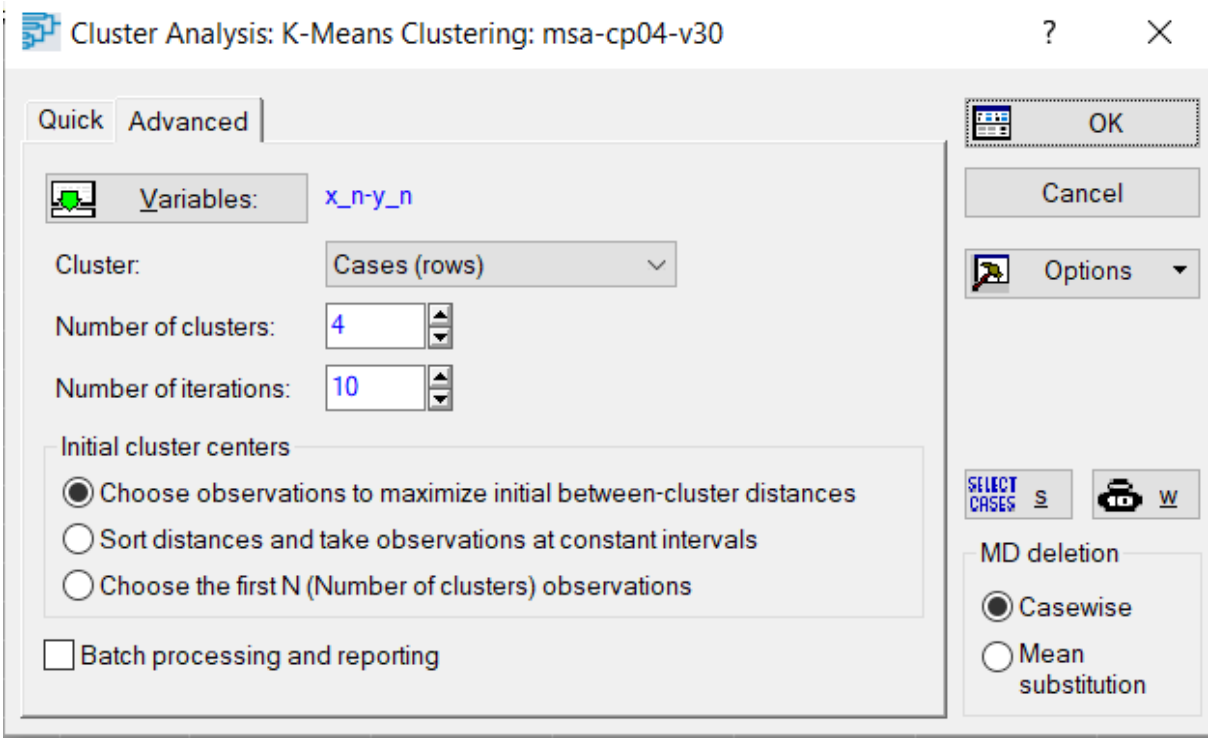
Підрахунок значення критерію якості кластеризування за допомогою Google Table

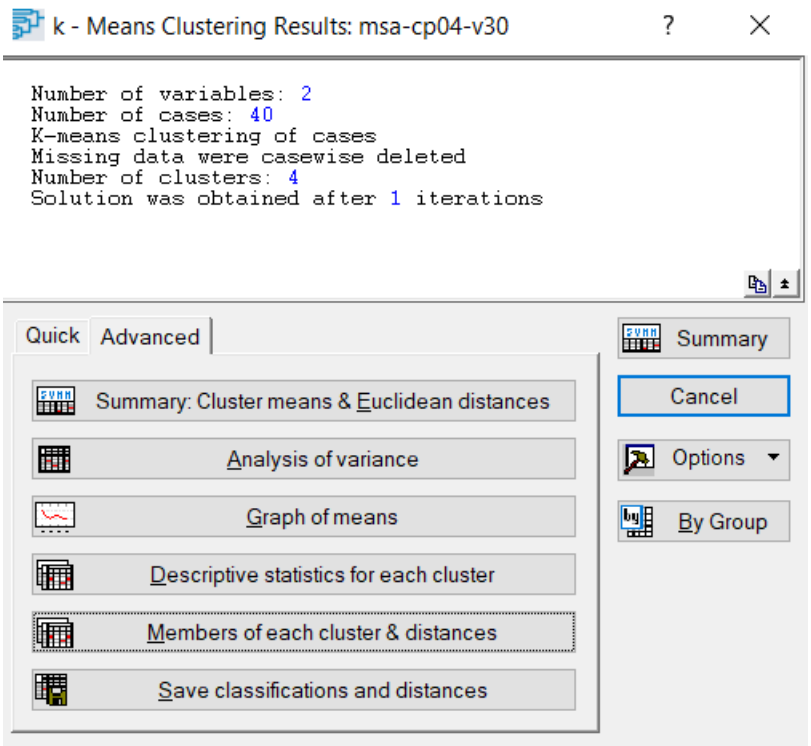
A	B	C	D	E	F	G	H	I	J	K
1,203034	0,4969042	1,349306	1,747542							
1,112403	1,658917	0,916816	1,20084							
2,361816	1,068727	0,2985407	1,060724	Внутрішньокластерні відстані		1	2	3	4	SUM
0,9850839	0,6731348	1,374501	0,9897168			8,9148249	15,2864856	8,9679404	15,6485322	48,8177831
1,570671	0,7819124	1,611099	1,216505			6	15	8	11	40
1,681817	1,161677	1,642419	1,780509							
	0,6804315	1,194127	1,010355							
	0,6402321	0,5811317	2,073635							
	1,433534		0,9516484	0	88,34	166,4	15,52			
	1,085654		2,10218	9,4	0	12,3	29,81			
	0,4998137		1,514877	12,9	3,5	0	80,32		SUM	
	0,3395739			3,94	5,46	9	0		44,2	
	1,167719								6	
	2,163988									
	1,434267								Result	
									0,1656712096	

Результат: 0,166

Повторив таке саме для 4-ох кластерів використовуючи нормалізовані дані

Діалогове вікно початкових налаштувань кластерного аналізу





Матриця міжкластерних евклідових(під діагоналлю) і квадратичних евклідових(над діагоналлю)

Cluster Number	Euclidean Distances between Clusters (msa-cp04-v30) Distances below diagonal Squared distances above diagonal			
	No. 1	No. 2	No. 3	No. 4
No. 1	0,000	1,017	0,779	3,622
No. 2	1,008	0,000	3,513	0,803
No. 3	0,883	1,874	0,000	7,675
No. 4	1,903	0,896	2,770	0,000

Описова статистика першого кластеру

Variable	Descriptive Statistics for Cluster 1 (msa-cp04-v30) Cluster contains 9 cases		
	Mean	Standard Deviation	Variance
x_n	0,051	0,290	0,084
y_n	0,146	0,393	0,154

Описова статистика другого кластеру

Variable	Descriptive Statistics for Cluster 2 (msa-cp04-v30) Cluster contains 11 cases		
	Mean	Standard Deviation	Variance
x_n	0,944	0,328	0,108
y_n	-0,966	0,258	0,066

Описова статистика третього кластеру

Variable	Descriptive Statistics for Cluster 3 (msa-cp04-v30) Cluster contains 17 cases		
	Mean	Standard Deviation	Variance
x_n	-0,959	0,332	0,110
y_n	0,880	0,418	0,174

Описова статистика четвертого кластеру

Variable	Descriptive Statistics for Cluster 4 (msa-cp04-v30) Cluster contains 3 cases		
	Mean	Standard Deviation	Variance
x_n	1,82	0,240	0,057
y_n	-1,88	0,166	0,028

Результати дисперсійного аналізу

Variable	Analysis of Variance (msa-cp04-v30)					
	Between SS	df	Within SS	df	F	signif. p
x_n	35,37	3	3,626	36	117,1	0,000
y_n	34,26	3	4,742	36	86,7	0,000

Відстанями кожного об’єкту до центру для першого кластеру

	Members of Cluster Number 1 (msa-cp04-v30) and Distances from Respective Cluster Center Cluster contains 9 cases
	Distance
10	0,220
18	0,359
25	0,325
26	0,109
28	0,419
32	0,339
33	0,452
34	0,154
37	0,376

Відстанями кожного об’єкту до центру для другого кластеру

	Members of Cluster Number 2 (msa-cp04-v30) and Distances from Respective Cluster Center Cluster contains 11 cases
	Distance
2	0,190
4	0,144
5	0,345
7	0,331
11	0,471
13	0,273
24	0,204
27	0,231
30	0,310
36	0,306
40	0,056

Відстанями кожного об’єкту до центру для третього кластеру

	Members of Cluster Number 3 (msa-cp04-v30) and Distances from Respective Cluster Center Cluster contains 17 cases
	Distance
1	0,273
3	0,210
6	0,393
9	0,369
12	0,222
14	0,302
15	0,249
16	0,432
17	0,398
20	0,606
22	0,620
23	0,234
29	0,322
31	0,289
35	0,470
38	0,257
39	0,196

Відстанями кожного об’єкту до центру для четвертого кластеру

	Members of Cluster Number 4 (msa-cp04-v30) and Distances from Respective Cluster Center Cluster contains 3 cases
	Distance
8	0,182
19	0,210
21	0,089

Підрахунок значення критерію якості кластеризування за допомогою Google Table

A	B	C	D	E	F	G	H	I	J	K
0,22	0,19	0,273	0,182							
0,359	0,144	0,21	0,21							
0,325	0,345	0,393	0,089	Внутрішньокластерні відстані		1	2	3	4	SUM
0,109	0,331	0,369				2,753	2,861	5,842	0,481	11,937
0,419	0,471	0,222				9	11	17	2	39
0,339	0,273	0,302								
0,452	0,204	0,249								
0,154	0,231	0,432								
0,376	0,31	0,398		0	1,016926	0,779219	3,621548			
	0,306	0,606		1,008428	0	3,513466	0,8034819			
	0,056	0,62		0,8827338	1,874424	0	7,674842		SUM	
		0,234		1,903037	0,8963715	2,77035	0		9,3353443	
		0,322							6	
		0,289								
		0,47							Result	
		0,257							0,1967213505	
		0,196								

Результат: 0,197

Визначення ступеня впливу ознак

Рівень значущості 0.01

	$p = 0{,}01$						
df_1 / df_2	**14**	**16**	**20**	**24**	**30**	**40**	
1	6142	6169	6208	6234	6261	6286	6
2	99,43	99,44	99,45	99,46	99,47	99,48	9
3	26,92	26,83	26,69	26,60	26,50	26,41	2

Для 3 кластрів

$$F_{\text{кр}} = F(0.01, 37, 2) = 99,48$$

$$K_x = \frac{F_x}{F_{\text{кр}}} = \frac{77,4}{99,48} = 0,778 \qquad K_y = \frac{F_y}{F_{\text{кр}}} = \frac{111,71}{99,48} = 1,123$$

Таким чином, змінна у має вищий ступінь впливу на результати кластеризування ніж змінна х.

Для 3 кластерів з нормалізованими даними

$$F_{\text{кр}} = F(0.01, 37, 2) = 99,48$$

$$K_x = \frac{F_x}{F_{\text{кр}}} = \frac{91}{99,48} = 0,915 \qquad K_y = \frac{F_y}{F_{\text{кр}}} = \frac{100,7}{99,48} = 1,012$$

Таким чином, змінна у має вищий ступінь впливу на результати кластеризування ніж змінна х.

Для 4 кластерів

$$F_{\text{кр}} = F(0.01, 36, 3) = 26,45$$

$$K_x = \frac{F_x}{F_{\text{кр}}} = \frac{53,49}{26,45} = 2,022 \qquad K_y = \frac{F_y}{F_{\text{кр}}} = \frac{154,93}{26,45} = 5,857$$

Таким чином, змінна у має вищий ступінь впливу на результати кластеризування ніж змінна х.

Для 4 кластерів з нормалізованими даними

$$F_{\text{кр}} = F(0.01, 36, 3) = 26,45$$

$$K_x = \frac{F_x}{F_{\text{кр}}} = \frac{117,1}{26,45} = 4,427$$

$$K_y = \frac{F_y}{F_{\text{кр}}} = \frac{86,7}{26,45} = 3,278$$

Таким чином, змінна х має вищий ступінь впливу на результати

кластеризування ніж змінна y.

Аналіз результатів

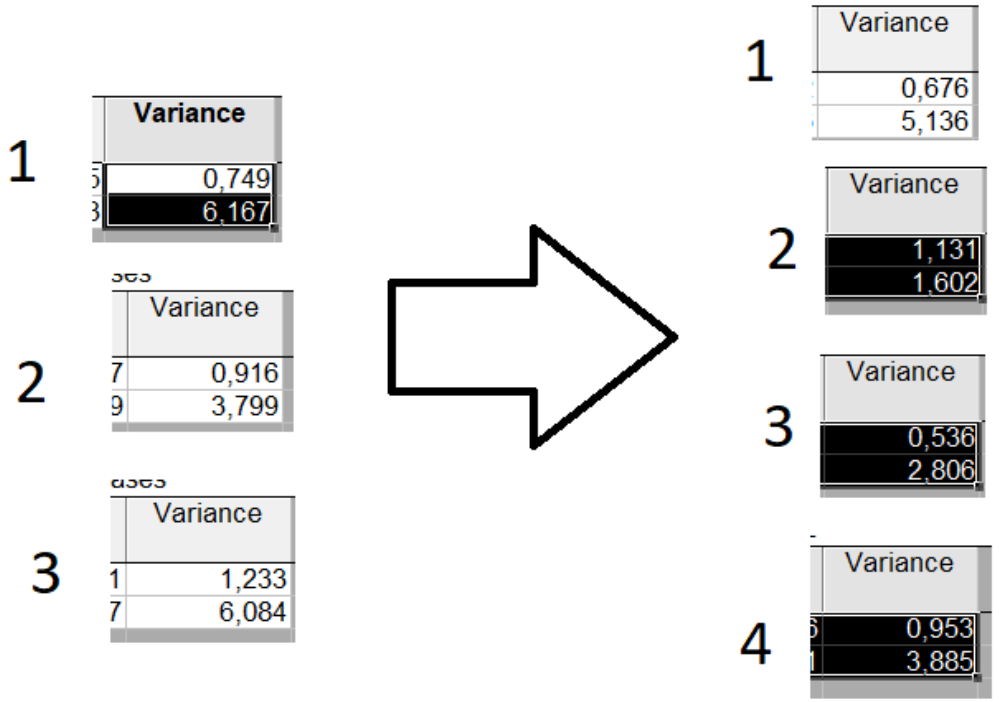
За допомогою результатів отриманих в попередньому пункті можна подивитися на якість кластеризування і отримати таку послідовність:

1. Розбиття на 4 кластери з початковими даними(0,166)
2. Розбиття на 4 кластери з нормалізованими даними(0,197)
3. Розбиття на 3 кластери з початковими даними(0,236)
4. Розбиття на 3 кластери з нормалізованими даними(0,239)

Відповідно отримано, що розбиття на чотири кластери є якіснішим, а частка залишкової дисперсії є меншою ніж для розбиття на три кластери.

Розбиття з нормалізованими даними в обох випадках виявилися гірше.

Також можна подивитись на описові статистики кожного кластеру і побачити, що дисперсія дуже значно нища для 4-ох кластерів.



Висновки:

В процесі виконання комп’ютерного практикуму №4 я виконав розрахунки згідно індивідуального завдання і набув вмінь кластерного аналізу сукупності об’єктів методом k-середніх в Statistica. Було дуже зручно отримувати результати кластеризування за допомогою додатку. Присутньо багато зручних функцій, але мені не вистачило(або я просто не знайшов) відображення кластерів на графіку. Щоб ще раз візуально проаналізувати результати кластеризації. Але оскільки таке відображення потрібне тільки для випадків з 2-3 змінними, то це не є проблемою.