Национальный исследовательский университет

“Высшая школа экономики”

Студентка второго курса

ОП “Бизнес-информатика”

Григоращенко Екатерина Андреевна

30.03.2021

ДОМАШНЕЕ ЗАДАНИЕ ПО КУРСУ “КЛАССИФИКАЦИЯ СТАТИСТИЧЕСКИХ ДАННЫХ” КОРРЕЛЯЦИОННЫЙ И КЛАСТЕРНЫЙ АНАЛИЗ ДАННЫХ

Содержание

1. Описание исследуемых данных

Задача данного исследования найти данные соответствующие требованиям, провести для них корреляционный и кластерный анализы. Сделать соответствующие выводы по проведенным анализам.

Для проведения исследования будут использоваться данные по количеству людей, болеющих Covid-19 в России, США, Германии, Индии, Индонезии и Украине с 1 января 2021 года по 28 марта 2021 года.

Данные является порядковыми, временными.

Таблица 1. Данные по количеству больных Covid-19 в разных странах в период с 1 января по 28 марта 2021 года.

№	Дата	Индонезия	Германия	Украина	Индия	США	Россия
1	28.03.2021	124 236	225 277	311 484	523 602	6 989 393	282 964
2	27.03.2021	124 517	218 790	303 098	487 840	7 006 619	282 842
3	26.03.2021	124 497	210 099	293 088	454 249	7 016 622	284 681
4	25.03.2021	125 279	199 814	282 420	422 596	7 018 080	286 799
5	24.03.2021	123 926	188 661	272 861	396 696	7 012 991	288 852
6	23.03.2021	126 439	179 697	266 415	370 101	7 040 066	290 747
7	22.03.2021	128 250	179 444	263 316	347 071	7 175 014	293 577
8	21.03.2021	129 844	179 530	258 573	336 392	7 212 208	292 444
9	20.03.2021	131 616	174 055	251 471	310 801	7 243 639	292 259
10	19.03.2021	131 828	168 413	241 629	290 152	7 261 358	294 298
11	18.03.2021	131 753	170 151	230 633	273 032	7 282 611	297 379
12	17.03.2021	131 695	152 492	222 361	253 882	7 296 873	300 097
13	16.03.2021	134 042	146 044	216 284	236 008	7 323 714	302 281
14	15.03.2021	136 524	146 170	213 108	225 072	7 304 022	303 975
15	14.03.2021	137 912	146 876	209 937	220 951	7 365 186	303 209
16	13.03.2021	138 942	143 343	203 813	212 147	7 388 461	302 933
17	12.03.2021	140 451	139 906	195 677	203 661	7 416 659	306 368
18	11.03.2021	141 070	135 950	187 591	200 401	8 591 627	310 556
9	10.03.2021	144 213	130 879	184 099	190 295	8 679 663	315 751

20	09.03.2021	144 311	128 028	183 628	185 767	8 711 254	320 488
21	08.03.2021	145 628	130 170	181 418	189 172	8 756 390	321 310
22	07.03.2021	147 740	131 823	177 686	190 501	8 821 756	321 758
23	06.03.2021	147 172	130 252	172 769	186 253	8 842 857	323 107
24	05.03.2021	148 356	129 490	167 132	181 997	8 874 812	327 553
25	04.03.2021	147 845	128 899	161 298	177 967	8 899 008	332 455
26	03.03.2021	147 197	126 785	156 272	175 044	8 921 400	337 668
27	02.03.2021	149 645	126 136	154 340	171 776	8 936 169	343 279
28	01.03.2021	153 074	129 145	154 360	169 786	8 969 496	348 121
29	28.02.2021	155 765	131 275	151 844	170 293	9 035 250	348 160
30	27.02.2021	157 039	130 388	147 802	166 078	9 045 632	349 571
31	26.02.2021	158 408	130 285	144 547	160 985	9 079 762	354 496
32	25.02.2021	157 705	130 316	140 934	157 418	9 096 024	359 560
33	24.02.2021	158 162	128 727	136 811	153 178	9 115 931	364 910
34	23.02.2021	158 604	128 349	134 758	148 584	9 170 205	365 762
35	22.02.2021	157 148	132 728	135 080	148 882	9 199 577	367 312
36	21.02.2021	157 088	135 541	133 280	151 639	9 281 124	367 988
37	20.02.2021	158 197	134 974	130 406	147 214	9 297 326	371 675
38	19.02.2021	160 142	136 841	128 723	144 654	9 292 344	376 686
39	18.02.2021	160 494	138 762	128 169	140 998	9 314 794	382 360
40	17.02.2021	162 182	140 710	127 320	138 802	9 354 485	388 123
41	16.02.2021	160 689	144 830	129 108	138 254	9 401 811	393 681
42	15.02.2021	158 498	152 121	132 304	138 579	9 462 875	398 534
43	14.02.2021	159 012	157 040	132 512	141 325	9 540 008	398 656
44	13.02.2021	161 731	159 490	131 465	139 277	9 547 775	400 095
45	12.02.2021	165 086	164 386	132 538	138 253	9 573 871	404 501
46	11.02.2021	166 492	169 102	135 092	137 578	9 584 499	410 639
47	10.02.2021	168 416	174 218	136 745	144 032	9 586 691	418 115
48	09.02.2021	169 351	181 480	141 305	143 416	9 680 957	426 732
49	08.02.2021	171 288	192 429	147 603	145 690	9 711 647	434 038

50	07.02.2021	176 291	200 182	148 206	150 653	9 782 082	434 410
51	06.02.2021	176 433	202 101	147 214	150 733	9 777 664	438 678
52	05.02.2021	176 672	206 510	150 191	150 476	9 790 450	445 379
53	04.02.2021	174 798	213 609	153 611	153 270	9 775 210	452 800
54	03.02.2021	175 236	219 135	158 602	156 873	9 779 825	461 153
55	02.02.2021	172 576	226 582	165 583	161 865	9 818 915	470 027
56	01.02.2021	175 349	238 320	174 802	165 234	9 827 632	476 295
57	31.01.2021	175 095	246 209	177 964	170 203	9 911 107	477 253
58	30.01.2021	174 083	247 922	178 992	170 670	9 876 547	479 419
59	29.01.2021	170 017	251 541	185 773	171 808	9 865 162	485 401
60	28.01.2021	166 540	254 674	192 030	173 762	9 824 791	492 901
61	27.01.2021	164 113	258 321	198 596	175 328	9 784 075	501 113
62	26.01.2021	163 526	265 678	209 215	178 200	9 808 397	511 888
63	25.01.2021	161 636	277 754	219 107	178 808	9 812 845	518 009
64	24.01.2021	162 617	287 495	222 437	185 922	9 863 221	518 178
65	23.01.2021	158 751	289 753	224 012	185 963	9 808 205	519 987
66	22.01.2021	156 683	293 041	232 338	187 260	9 743 146	527 404
67	21.01.2021	151 658	295 544	241 392	190 244	9 674 810	533 789
68	20.01.2021	149 388	298 199	250 031	193 650	9 614 267	539 416
69	19.01.2021	146 842	306 192	260 361	198 596	9 611 435	544 151
70	18.01.2021	144 798	319 509	271 651	202 202	9 666 132	546 265
71	17.01.2021	145 482	330 944	273 920	209 519	9 646 497	542 212
72	16.01.2021	143 517	333 651	276 823	210 215	9 557 138	542 547
73	15.01.2021	138 238	335 964	279 030	212 646	9 471 604	546 356
74	14.01.2021	133 149	338 319	281 325	214 472	9 342 166	549 832
75	13.01.2021	129 628	340 050	283 754	214 812	9 247 519	553 595
76	12.01.2021	126 313	344 364	292 043	215 795	9 163 873	559 969
77	11.01.2021	123 636	353 480	303 062	217 839	9 077 487	562 321
78	10.01.2021	122 873	362 620	303 661	224 103	9 050 677	561 228
79	09.01.2021	120 928	362 198	304 041	224 903	8 925 570	562 913

80	08.01.2021	117 704	360 638	307 236	225 756	8 823 543	563 754
81	07.01.2021	114 766	356 070	306 774	226 716	8 619 495	562 233
82	06.01.2021	112 593	352 129	309 241	229 403	8 463 484	562 927
83	05.01.2021	110 693	352 697	317 048	228 826	8 352 337	562 210
84	04.01.2021	110 089	359 034	325 766	232 343	8 254 416	561 114
85	03.01.2021	110 679	367 155	326 374	245 474	8 316 492	559 399
86	02.01.2021	110 400	370 581	328 171	248 633	8 184 352	555 600
87	01.01.2021	111 005	377 206	325 790	252 275	8 085 060	548 643

Визуализация данных:

Рис. 1. График изменения количества людей больных Covid-19 в Индонезии с 1 января по 28 марта 2021 года.

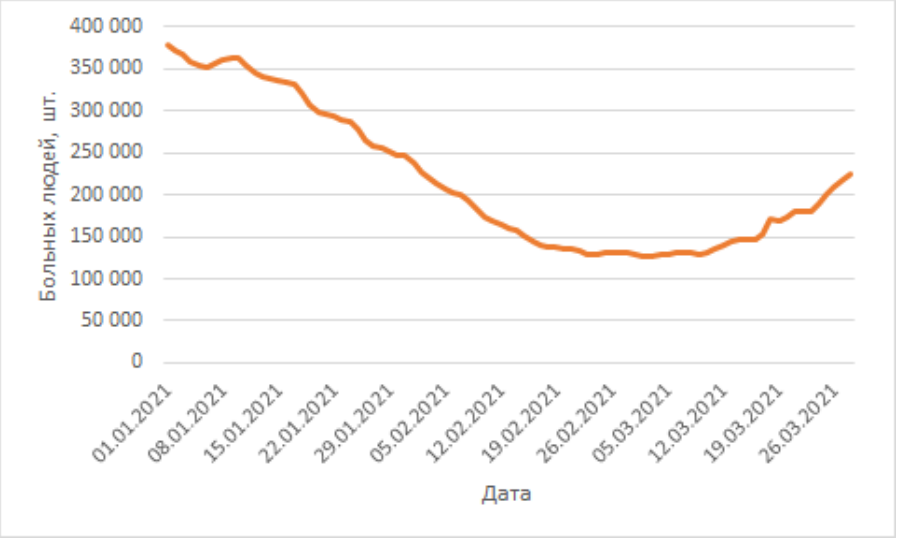

Рис. 2. График изменения количества людей больных Covid-19 в Германии с 1 января по 28 марта 2021 года.

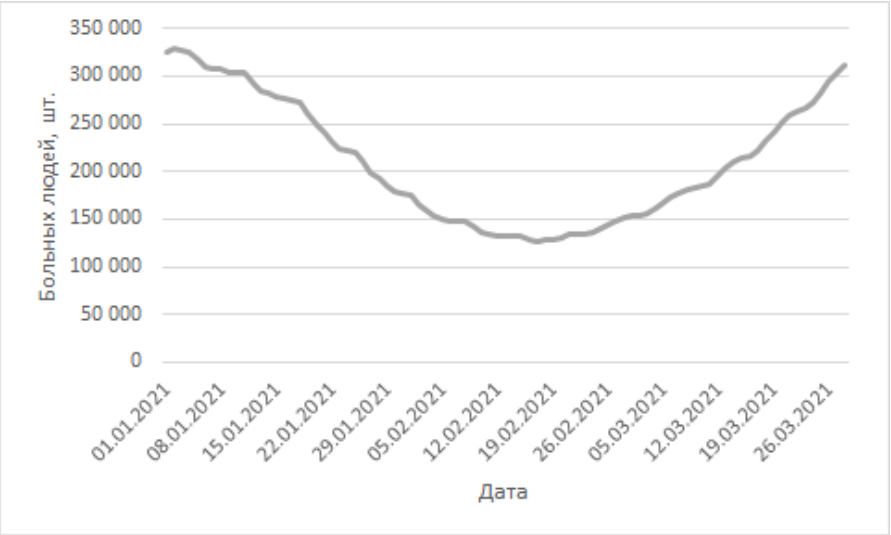

Рис. 3. График изменения количества людей больных Covid-19 в Украине с 1 января по 28 марта 2021 года.

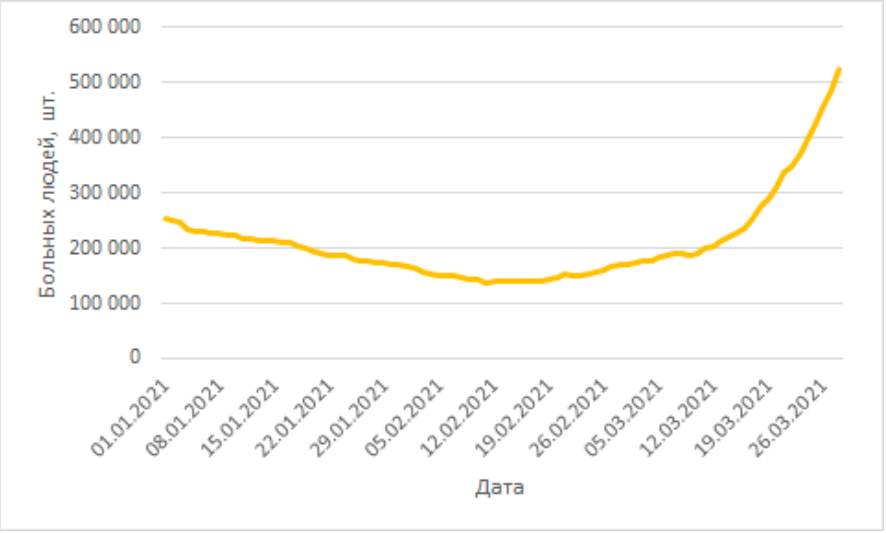

Рис. 4. График изменения количества людей больных Covid-19 в Индии с 1 января по 28 марта 2021 года.

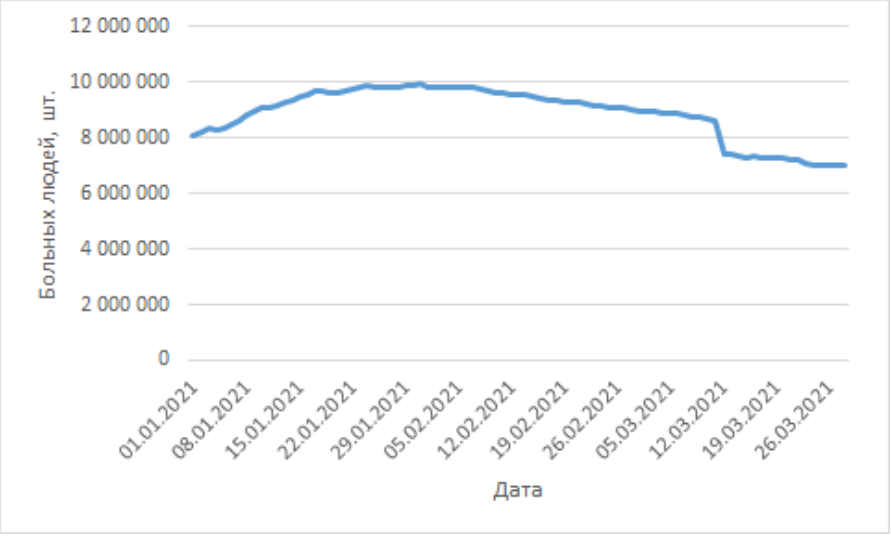

Рис. 5. График изменения количества людей больных Covid-19 в США с 1 января по 28 марта 2021 года.

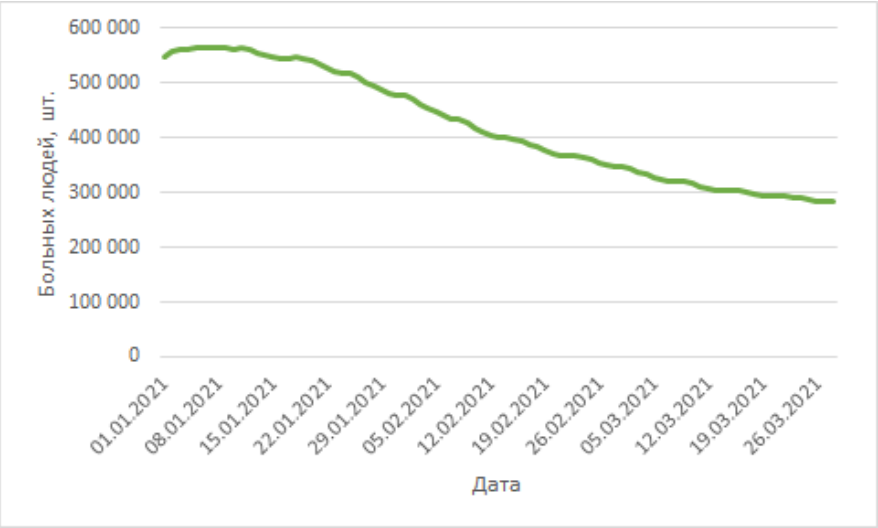

Рис. 6. График изменения количества людей больных Covid-19 в России с 1 января по 28 марта 2021 года.

Таблица 2. Описательные статистики данных с доверительным интервалом 95%.

			Статистика	Стандартная ошибка
Индонезия	Среднее		146830,97	2052,701
	95% Доверительный интервал для среднего	Нижняя граница	142750,33	
		Верхняя граница	150911,60	
	Среднее по выборке, усеченной на 5%		147222,27	
	Медиана		147845,00	
	Дисперсия		366581621.8	
	Стандартная отклонения		19146,321	
	Минимум		110089	
	Максимум		176672	
	Диапазон		66583	
	Межквартильный диапазон		30020	
	Асимметрия		-,281	,258
	Эксцесс		-,961	,511
Германия	Среднее		217076,80	9047,964
	95% Доверительный интервал для среднего	Нижняя граница	199090,05	
		Верхняя граница	235063,56	
	Среднее по выборке, усеченной на 5%		213627,05	
	Медиана		188661,00	
	Дисперсия		7122311379	
	Стандартная отклонения		84393,788	
	Минимум		126136	
	Максимум		377206	
	Диапазон		251070	
	Межквартильный диапазон		154279	
	Асимметрия		,587	,258
	Эксцесс		-1,149	,511
Украина	Среднее		210129,59	6990,626
	95% Доверительный интервал для среднего	Нижняя граница	196232,68	
		Верхняя граница	224026,49	

	95% Доверительный интервал для среднего	Нижняя граница	196232,68	
		Верхняя граница	224026,49	
	Среднее по выборке, усеченной на 5%		208240,67	
	Медиана		195677,00	
	Дисперсия		4251590298	
	Стандартная отклонения		65204,220	
	Минимум		127320	
	Максимум		328171	
	Диапазон		200851	
	Межквартильный диапазон		124655	
	Асимметрия		,352	,258
	Эксцесс		-1,300	,511
Индия	Среднее		208779,21	8425,794
	95% Доверительный интервал для среднего	Нижняя граница	192029,28	
		Верхняя граница	225529,13	
	Среднее по выборке, усеченной на 5%		198408,89	
	Медиана		186253,00	
	Дисперсия		6176478300	
	Стандартная отклонения		78590,574	
	Минимум		137578	
	Максимум		523602	
	Диапазон		386024	
	Межквартильный диапазон		68199	
	Асимметрия		2,168	,258
	Эксцесс		4,985	,511
США	Среднее		8870438,67	100608,563
	95% Доверительный интервал для среднего	Нижняя граница	8670435,46	
		Верхняя граница	9070441,87	
	Среднее по выборке, усеченной на 5%		8918141,80	
	Медиана		9115931,00	
	Дисперсия		8.806E+11	
	Стандартная отклонения		938414,200	
	Минимум		6989393	
	Максимум		9911107	
	Диапазон		2921714	
	Межквартильный диапазон		1294160	
	Асимметрия		-,890	,258
	Эксцесс		-,553	,511
Россия	Среднее		419771,43	10722,991
	95% Доверительный интервал для среднего	Нижняя граница	398454,83	
		Верхняя граница	441088,03	
	Среднее по выборке, усеченной на 5%		419323,54	
	Медиана		400095,00	
	Дисперсия		1.000E+10	
	Стандартная отклонения		100017,398	
	Минимум		282842	
	Максимум		563754	
	Диапазон		280912	
	Межквартильный диапазон		205646	
	Асимметрия		,152	,258
	Эксцесс		-1,515	,511

2. Корреляционный анализ

Таблица 3. Корреляция и ее значимость

		Индонезия	Германия	Украина	Индия	США	Россия
Индонезия	Корреляция Пирсона	1	-,432**	-,844**	-,662**	,714**	-,054
Германия	Корреляция Пирсона	-,432**	1	,780**	,165	,173	,882**
Украина	Корреляция Пирсона	-,844**	,780**	1	,673**	-,452**	,408**
Индия	Корреляция Пирсона	-,662**	,165	,673**	1	-,795**	-,300**
США	Корреляция Пирсона	,714**	,173	-,452**	-,795**	1	,586**
Россия	Корреляция Пирсона	-,054	,882**	,408**	-,300**	,586**	1

**. Корреляция значима на уровне 0,01 (односторонняя).

Корреляционный анализ был проведен с помощью программы SPSS. Значимость коэффициентов корреляций была проверена на уровне 0,01. С помощью анализа было выявлено, что выборки по Индонезии и Украине, Индонезии и США, Германии и Украине, Германии и России, Индии и США имеют сильную, близкую к линейной (или же обратной линейной, в случае отрицательных значений) , значимую корреляционную связь. Данные по Индонезии и Индии, Украине и Индии, а также США и России имеют умеренную, стремящуюся к (обратной) линейной, значимую связь. Коэффициенты корреляции между Индонезий и Германией, Украиной и США, Украиной и Россией и Индией и Россией показывает наличие слабой, однако значимой связи. Корреляционная связь между Индонезией и Россией, Германией и Индией, Германией и США практически отсутствует и не является значимой.

Таким образом, для дальнейшего кластерного анализа будут использованы переменные: Индонезия, Украина, Индия и США, так как все их парные корреляции имеют значимую связь, стремящуюся к (обратной) линейной.

Нормализация и стандартизация данных проведена в Excel с помощью встроенной функции “Нормализация”.

Таблица 4. Нормализованные и стандартизированные данные по количеству больных Covid-19 в разных странах в период с 1 января по 28 марта 2021 года.

№	Дата	Индонезия	Украина	Индия	США
1	28.03.2021	-1,18	1,55	4,01	-2,00
2	27.03.2021	-1,17	1,43	3,55	-1,99
3	26.03.2021	-1,17	1,27	3,12	-1,98
4	25.03.2021	-1,13	1,11	2,72	-1,97

5	24.03.2021	-1,20	0,96	2,39	-1,98
6	23.03.2021	-1,07	0,86	2,05	-1,95
7	22.03.2021	-0,97	0,82	1,76	-1,81
8	21.03.2021	-0,89	0,74	1,62	-1,77
9	20.03.2021	-0,79	0,63	1,30	-1,73
10	19.03.2021	-0,78	0,48	1,04	-1,71
11	18.03.2021	-0,79	0,31	0,82	-1,69
12	17.03.2021	-0,79	0,19	0,57	-1,68
13	16.03.2021	-0,67	0,09	0,35	-1,65
14	15.03.2021	-0,54	0,05	0,21	-1,67
15	14.03.2021	-0,47	0,00	0,15	-1,60
16	13.03.2021	-0,41	-0,10	0,04	-1,58
17	12.03.2021	-0,33	-0,22	-0,07	-1,55
18	11.03.2021	-0,30	-0,35	-0,11	-0,30
ë9	10.03.2021	-0,14	-0,40	-0,24	-0,20
20	09.03.2021	-0,13	-0,41	-0,29	-0,17
21	08.03.2021	-0,06	-0,44	-0,25	-0,12
22	07.03.2021	0,05	-0,50	-0,23	-0,05
23	06.03.2021	0,02	-0,57	-0,29	-0,03
24	05.03.2021	0,08	-0,66	-0,34	0,00
25	04.03.2021	0,05	-0,75	-0,39	0,03
26	03.03.2021	0,02	-0,83	-0,43	0,05
27	02.03.2021	0,15	-0,86	-0,47	0,07
28	01.03.2021	0,33	-0,86	-0,50	0,11
29	28.02.2021	0,47	-0,89	-0,49	0,18
30	27.02.2021	0,53	-0,96	-0,54	0,19
31	26.02.2021	0,60	-1,01	-0,61	0,22
32	25.02.2021	0,57	-1,06	-0,65	0,24
33	24.02.2021	0,59	-1,12	-0,71	0,26
34	23.02.2021	0,61	-1,16	-0,77	0,32

35	22.02.2021	0,54	-1,15	-0,76	0,35
36	21.02.2021	0,54	-1,18	-0,73	0,44
37	20.02.2021	0,59	-1,22	-0,78	0,45
38	19.02.2021	0,70	-1,25	-0,82	0,45
39	18.02.2021	0,71	-1,26	-0,86	0,47
40	17.02.2021	0,80	-1,27	-0,89	0,52
41	16.02.2021	0,72	-1,24	-0,90	0,57
42	15.02.2021	0,61	-1,19	-0,89	0,63
43	14.02.2021	0,64	-1,19	-0,86	0,71
44	13.02.2021	0,78	-1,21	-0,88	0,72
45	12.02.2021	0,95	-1,19	-0,90	0,75
46	11.02.2021	1,03	-1,15	-0,91	0,76
47	10.02.2021	1,13	-1,13	-0,82	0,76
48	09.02.2021	1,18	-1,06	-0,83	0,86
49	08.02.2021	1,28	-0,96	-0,80	0,90
50	07.02.2021	1,54	-0,95	-0,74	0,97
51	06.02.2021	1,55	-0,96	-0,74	0,97
52	05.02.2021	1,56	-0,92	-0,74	0,98
53	04.02.2021	1,46	-0,87	-0,71	0,96
54	03.02.2021	1,48	-0,79	-0,66	0,97
55	02.02.2021	1,34	-0,68	-0,60	1,01
56	01.02.2021	1,49	-0,54	-0,55	1,02
57	31.01.2021	1,48	-0,49	-0,49	1,11
58	30.01.2021	1,42	-0,48	-0,48	1,07
59	29.01.2021	1,21	-0,37	-0,47	1,06
60	28.01.2021	1,03	-0,28	-0,45	1,02
61	27.01.2021	0,90	-0,18	-0,43	0,97
62	26.01.2021	0,87	-0,01	-0,39	1,00
63	25.01.2021	0,77	0,14	-0,38	1,00
64	24.01.2021	0,82	0,19	-0,29	1,06

65	23.01.2021	0,62	0,21	-0,29	1,00
66	22.01.2021	0,51	0,34	-0,27	0,93
67	21.01.2021	0,25	0,48	-0,24	0,86
68	20.01.2021	0,13	0,61	-0,19	0,79
69	19.01.2021	0,00	0,77	-0,13	0,79
70	18.01.2021	-0,11	0,94	-0,08	0,85
71	17.01.2021	-0,07	0,98	0,01	0,83
72	16.01.2021	-0,17	1,02	0,02	0,73
73	15.01.2021	-0,45	1,06	0,05	0,64
74	14.01.2021	-0,71	1,09	0,07	0,50
75	13.01.2021	-0,90	1,13	0,08	0,40
76	12.01.2021	-1,07	1,26	0,09	0,31
77	11.01.2021	-1,21	1,43	0,12	0,22
78	10.01.2021	-1,25	1,43	0,19	0,19
79	09.01.2021	-1,35	1,44	0,21	0,06
80	08.01.2021	-1,52	1,49	0,22	-0,05
81	07.01.2021	-1,67	1,48	0,23	-0,27
82	06.01.2021	-1,79	1,52	0,26	-0,43
83	05.01.2021	-1,89	1,64	0,26	-0,55
84	04.01.2021	-1,92	1,77	0,30	-0,66
85	03.01.2021	-1,89	1,78	0,47	-0,59
86	02.01.2021	-1,90	1,81	0,51	-0,73
87	01.01.2021	-1,87	1,77	0,55	-0,84

3. Иерархическая кластеризация

Для построения дендрограммы будет использовано евклидово расстояние, так как мы ищем расстояние в двухмерном пространстве и все анализируемые данные имеют одинаковые единицы измерения.
Построим дендрограмму с использованием метода межгрупповых связей в программе SPSS, чтобы узнать на какое количество кластеров возможно поделить выборку.

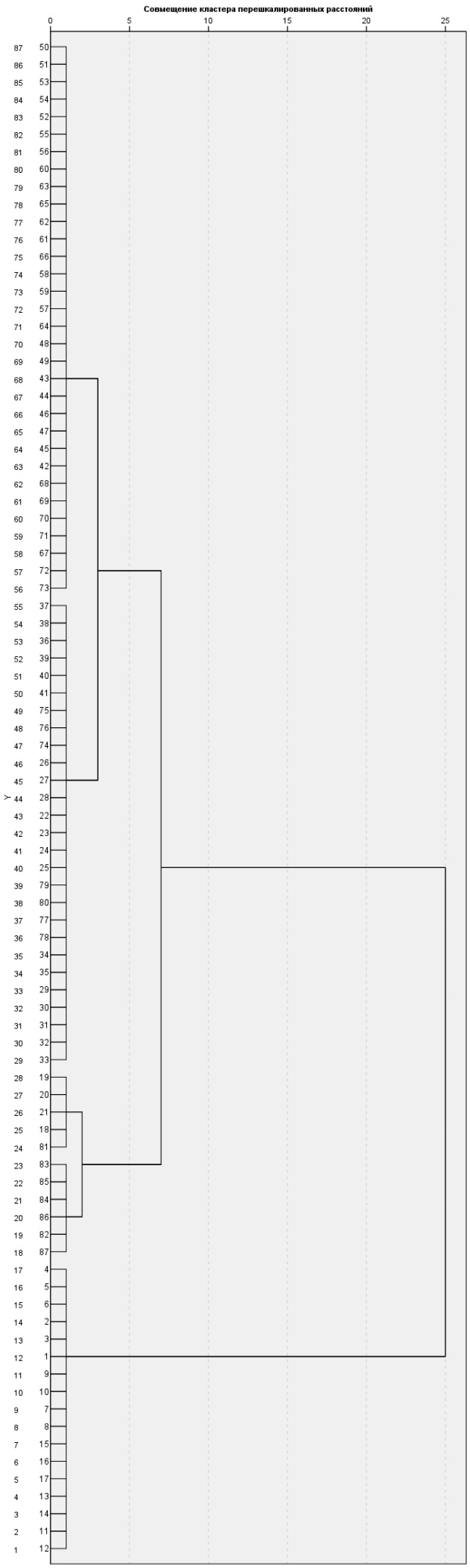

Рис. 7. Дендрограмма с использованием метода межгрупповых связей

Судя по получившейся дендрограмме, исследуемую выборку можно разделить на 2-5 кластеров. Однако, деление на 3 кластера кажется самым логичным, так как при таком делении будет более равное количество наблюдений в каждом кластере. Чтобы сделать более точные выводы, исследуем дополнительные параметры.

4. Определение количества кластеров

При помощи SPSS найдем квадраты расстояний до центров кластеров при k = 2, 3, 4, 5. Построим график зависимости сумм квадратов расстояний до центров кластеров от количества кластеров.

Таблица 5. Суммы квадратов расстояний до центров кластеров при разном количестве возможных кластеров

Количество кластеров	Сумма квадратов расстояний
2	32153560,04
3	18163140,65
4	13796952,96
5	12178330,47

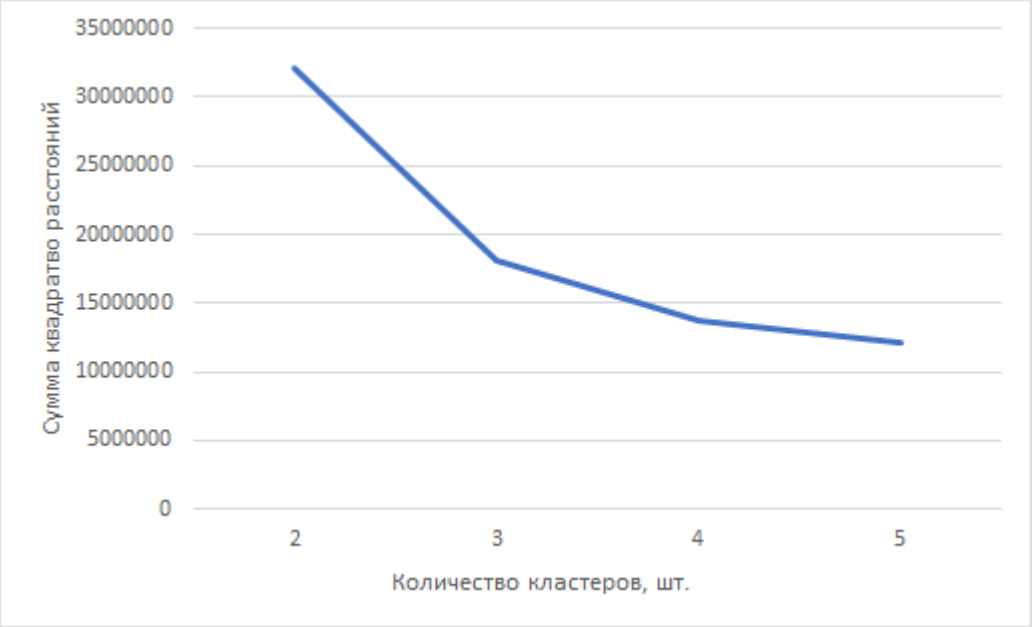

Рис. 8. График зависимости сумм квадратов расстояний до центров кластеров от количества кластеров

Проанализировав график, можно заметить, что наибольшее изменение в значении расстояния происходит в кластере 3, поэтому можно сделать вывод о том, что оптимальное количество кластеров для кластерного анализа методом k-средних будет 3 кластера.

5. Метод k-средних по нормализованным и стандартизированным данным. Сравнение.

Выяснив, что оптимальное деление на кластеры - это деление на 3 кластера, приступим к кластерному анализу ранее нормализованных и стандартизованных данных методом k-средних.
Для расчета расстояния будет использовано евклидово расстояние, так как мы ищем расстояние в двухмерном пространстве и все анализируемые данные имеют одинаковые единицы измерения.

№	Дата	Кластеры начальных данных	Кластеры норм. и станд. данных
1	28.03.2021	1	1
2	27.03.2021	1	1
3	26.03.2021	1	1
4	25.03.2021	1	1
5	24.03.2021	1	1
6	23.03.2021	1	1
7	22.03.2021	1	1
8	21.03.2021	1	1
9	20.03.2021	1	1
10	19.03.2021	1	3
11	18.03.2021	1	3
12	17.03.2021	1	3
13	16.03.2021	1	3
14	15.03.2021	1	3
15	14.03.2021	1	3
16	13.03.2021	1	3
17	12.03.2021	1	3
18	11.03.2021	2	2
ё9	10.03.2021	2	2

20	09.03.2021	2	2
21	08.03.2021	2	2
22	07.03.2021	2	2
23	06.03.2021	2	2
24	05.03.2021	2	2
25	04.03.2021	2	2
26	03.03.2021	2	2
27	02.03.2021	2	2
28	01.03.2021	2	2
29	28.02.2021	2	2
30	27.02.2021	2	2
31	26.02.2021	2	2
32	25.02.2021	2	2
33	24.02.2021	2	2
34	23.02.2021	2	2
35	22.02.2021	2	2
36	21.02.2021	3	2
37	20.02.2021	3	2
38	19.02.2021	3	2
39	18.02.2021	3	2
40	17.02.2021	3	2
41	16.02.2021	3	2
42	15.02.2021	3	2
43	14.02.2021	3	2
44	13.02.2021	3	2
45	12.02.2021	3	2
46	11.02.2021	3	2
47	10.02.2021	3	2
48	09.02.2021	3	2
49	08.02.2021	3	2

50	07.02.2021	3	2
51	06.02.2021	3	2
52	05.02.2021	3	2
53	04.02.2021	3	2
54	03.02.2021	3	2
55	02.02.2021	3	2
56	01.02.2021	3	2
57	31.01.2021	3	2
58	30.01.2021	3	2
59	29.01.2021	3	2
60	28.01.2021	3	2
61	27.01.2021	3	2
62	26.01.2021	3	2
63	25.01.2021	3	2
64	24.01.2021	3	2
65	23.01.2021	3	2
66	22.01.2021	3	2
67	21.01.2021	3	2
68	20.01.2021	3	2
69	19.01.2021	3	2
70	18.01.2021	3	3
71	17.01.2021	3	3
72	16.01.2021	3	3
73	15.01.2021	3	3
74	14.01.2021	3	3
75	13.01.2021	3	3
76	12.01.2021	2	3
77	11.01.2021	2	3
78	10.01.2021	2	3
79	09.01.2021	2	3

80	08.01.2021	2	3
81	07.01.2021	2	3
82	06.01.2021	2	3
83	05.01.2021	2	3
84	04.01.2021	2	3
85	03.01.2021	2	3
86	02.01.2021	2	3
87	01.01.2021	2	3

Распределение на кластеры начальных данных совпадает с распределением на кластеры нормализованных и стандартизованных данных лишь на 38%. Также распределение данных на кластеры в начальных данных является более равномерным, чем деление на кластеры нормализованных и стандартизованных данных.
В начальных данных: 1 кластер - 20% данных, 2 кластер - 34% данных, 3 кластер - 46% данных. В нормализованных: 1 кластер - 10%, 2 кластер - 60% данных, 3 кластер - 30% данных.
Наиболее подходящим вариантом кластеризации является кластеризация начальных данных методом k-средних на 3 кластера.
Так как кластеризация по начальным данным оказалась более рациональной, то рассмотрим график средних значений этих кластеров.

Рис. 9. График средних значений кластеров.

Как можно заметить по графику средних значений кластеров, значения в выборке достаточно близки к друг другу, поэтому координальной разницы между тем какой день (по данным о количестве больных людей) в какой кластер попал нет. Однако, смотря на распределение данных в США, можно сделать вывод о том, что в Кластер 1 попали те дни, когда больных было наименьшее количество, а в Кластер 3 - наименьшее.

Источники

1. https://exceltable.com/funkcii-excel/normalizovannoe-znachenie-normalizaciya
2. https://nafi.ru/upload/spss/Lection_9.pdf