Кластерный анализ

1 Постановка задачи

Классифицировать страны, используя социально-экономические и медико-санитарные факторы. И проанализировать с точки зрения семантики разбиение на кластеры.

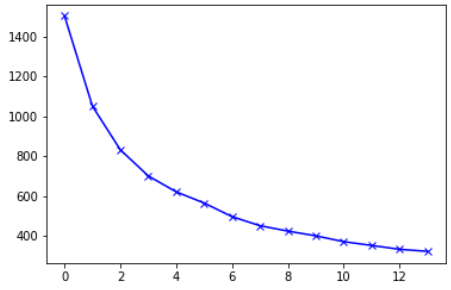
2 Решение

Был использован набор данных с сайта kaggle.com.
Используемые данные:
child_mort - детская смертность в возрасте до 5 лет на 1000 рожденных.
exports - экспорт товаров и услуг на душу населения.
health - общие расходы на здравоохранение на душу населения.
import - импорт товаров и услуг на душу населения.
income - чистый доход на человека.
inflation - измерение годового темпа роста ВВП.
life_expec - среднее число лет, которое прожил бы новорожденный ребенок, если бы текущие показатели смертности оставались прежними.
total_fer - число детей, которые родились бы у каждой женщины, если бы нынешние коэффициенты рождаемости оставались прежними.
gdpp - ВВП на душу населения.

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

Рис. 1: Данные

Затем с помощью локтевого метода определили примерное количество кластеров в наборе данных, используя K-means.



Получили оптимальное значение - 3. И использовали его для основной кластеризации и получения данных для анализа(см. рис. 2).

3 Анализ данных

У последнего кластера наблюдается низкая детская смертность, высокий чистый доход, высокое ВВП на душу населения, высокая предполагаемая продолжительность жизни, низкая инфляция. Все это указывает на высокое развитие данных стран.
У первого же все наоборот: высокая детская смертность, низкий доход, низкое ВВП на душу населения, высокая инфляция, что указывает на слабое развитие представителей кластера.
У среднего кастера показатели лучше, чем у первого, но хуже, чем у последнего, что соответствует развивающимся странам.
В итоге приходим к такому предположению о разбиении:
Cluster 0 - слаборазвитые страны
Cluster 1 - развивающиеся страны
Cluster 2 - развитые страны

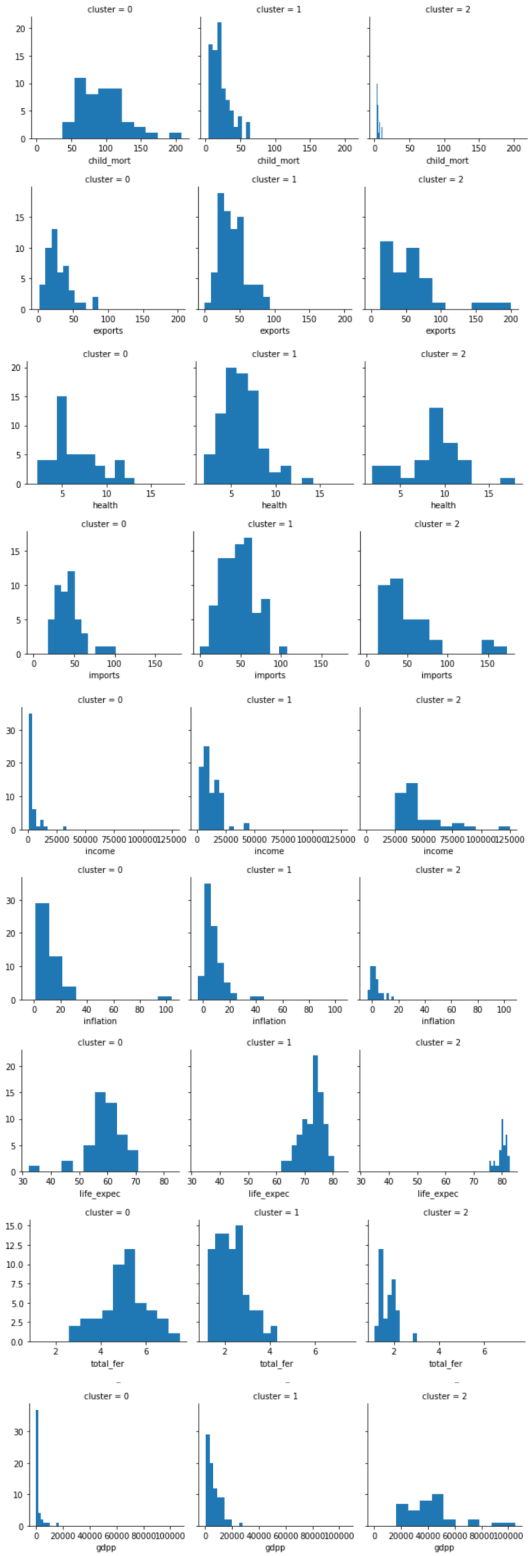


Рис. 2: Данные для анализа