Национальный исследовательский университет «Высшая школа экономики»
Факультет: Московский институт электроники и математики им.Тихонова
Образовательная программа: Прикладная математика

Отчет по модульной работе №1 по майнору «Прикладной статистический анализ»

Работу выполнила
студентка 2 курса
Беломытцева Алена
Владимировна
Преподаватель:
Тихонова Арина Михайловна

Москва, 2023г.

Введение

Для анализа в данной модульной работе в качестве переменных были взяты 6 параметров:

- ВВП на душу населения по некоторым странам
- Государственные расходы на образование, всего (% от ВВП)
- Коэффициент рождаемости (на 1000 человек)
- Внутренние государственные расходы на здравоохранение (% от ВВП)
- Распространенность употребления табака в настоящее время (% взрослого населения)
- Смертность в результате дорожно-транспортного травматизма (на 100 000 человек)

Все данные были собраны на 2018 год. В выборке участвуют 96 стран мира.

Источником данных является World Bank Data.

Первый параметр — валовый внутренний продукт на душу населения, который отражает уровень экономической активности и уровень жизни в стране.

Второй параметр — траты государства на образование, показывает сколько процентов от ВВП страны вкладывается государством в образование. Показатель относительный и конкретное его значение зависит он общего ВВП страны, но тем не менее, чем больше страна готова потратить на образование, тем грамотнее и работоспособнее население.

Третий параметр — коэффициент рождаемости, показывающий сколько произошло живорождений на 1000 человек. Важен для демографической оценки страны.

Четвертый параметр - внутренние государственные расходы на здравоохранение показывает лишь процент от общего ВВП страны, но можно

сказать, что чем больше государство вкладывается в медицину своей страны — тем меньше смертность от заболеваний, оказание первой медицинской помощи, проведение исследований над штаммами неизлечимых болезней и т.п.

Пятый параметр — показывает распространенность употребления табака у взрослого населения.

Шестой параметр — смертность в результате дорожно-транспортного происшествия. Показывает аккуратность вождения внутри страны, удобство дорог, возможно оперативность работы экстренных служб.

Задачей этой модульной работы является изучение взаимосвязей между выбранными показателями, получение данных о возможности разбиения стран на группы.

Гипотезы:

Существует взаимосвязь между рождаемостью и высоким уровнем ВВП на душу населения — предположительно чем больше ВВП, тем люди охотнее думают о продолжении своего рода.

Существует взаимосвязь между уровнем ВВП и тратами на образование. Чем больше ВВП, тем возрастает необходимость в хороших работниках, следовательно для их подготовки тратится больше и затраты государства на образование тратится больше.

Чем больше траты на здравоохранение, тем меньше смертность от дорожно-транспортных происшествий.

Траты на здравоохранение и рождаемость связаны.

Распространенность употребления табака связана с ВВП.

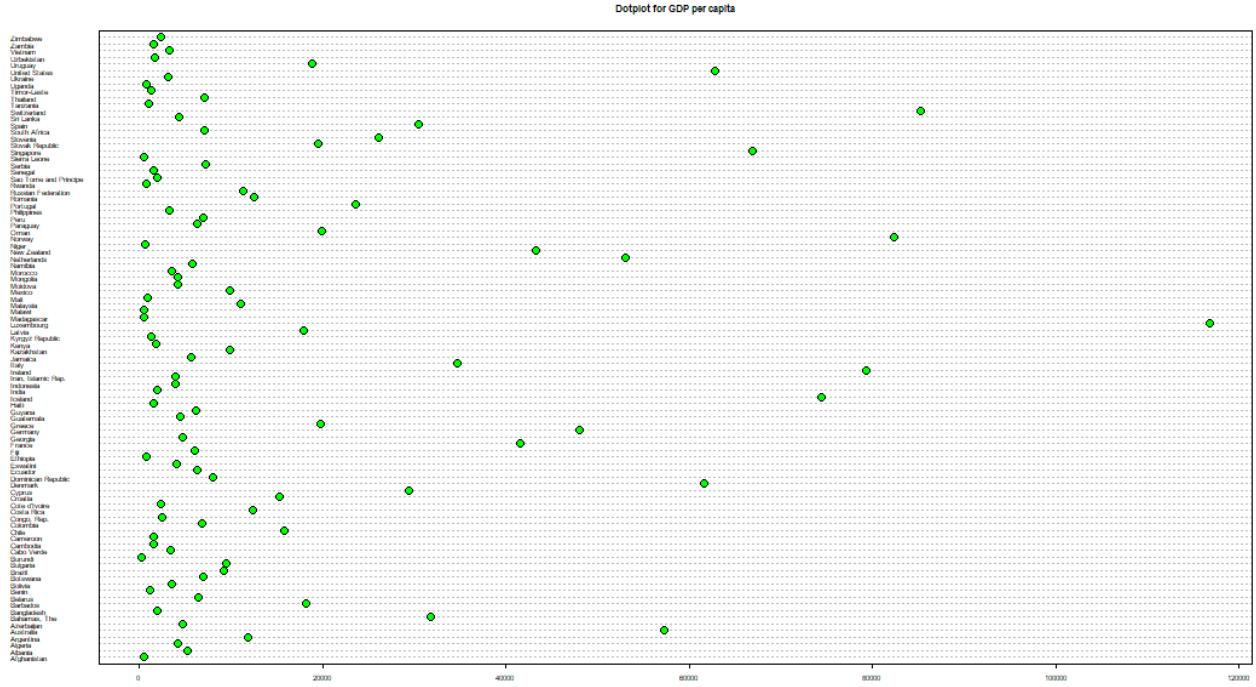
Также все страны выборки предположительно можно разделить на три группы: страны с высоким уровнем развития (образованные граждане, высокий ВВП, развитое здравоохранение), средним уровнем развития и низким уровнем.

Предварительный анализ данных

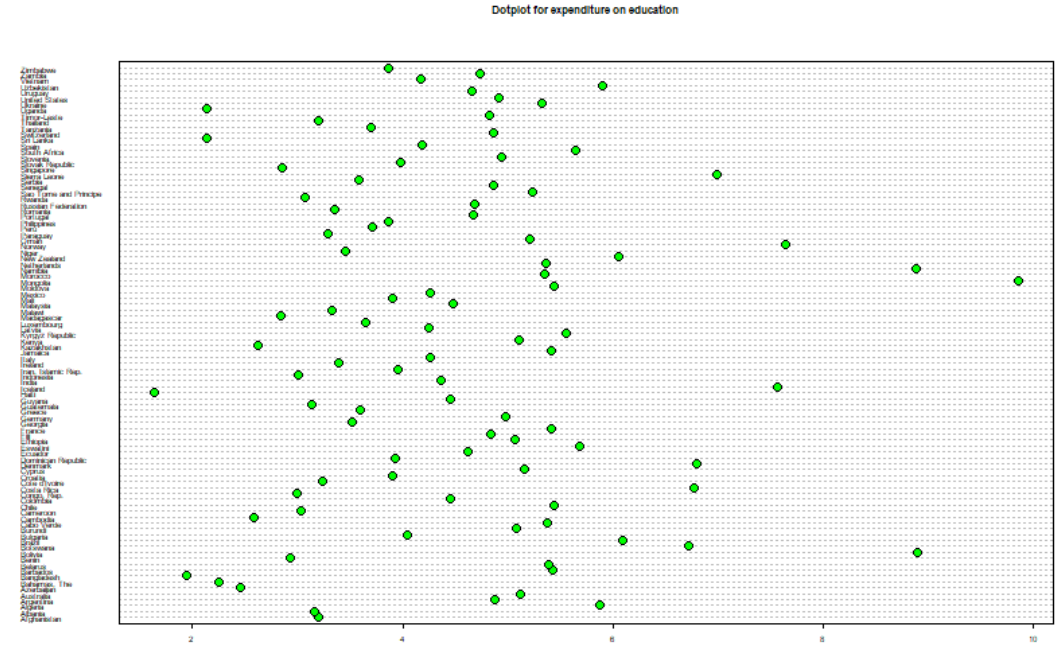
В этом разделе будет проведен первичный анализ данных, а именно будут построены точечные диаграммы, steamplot, boxplot, для каждого из параметров, исследованы основные характеристики случайных величин, характеристики их разброса, их ранговые показатели, их стандартизация и нормализация, а также будут выявлены аномальные данные.

Визуализация

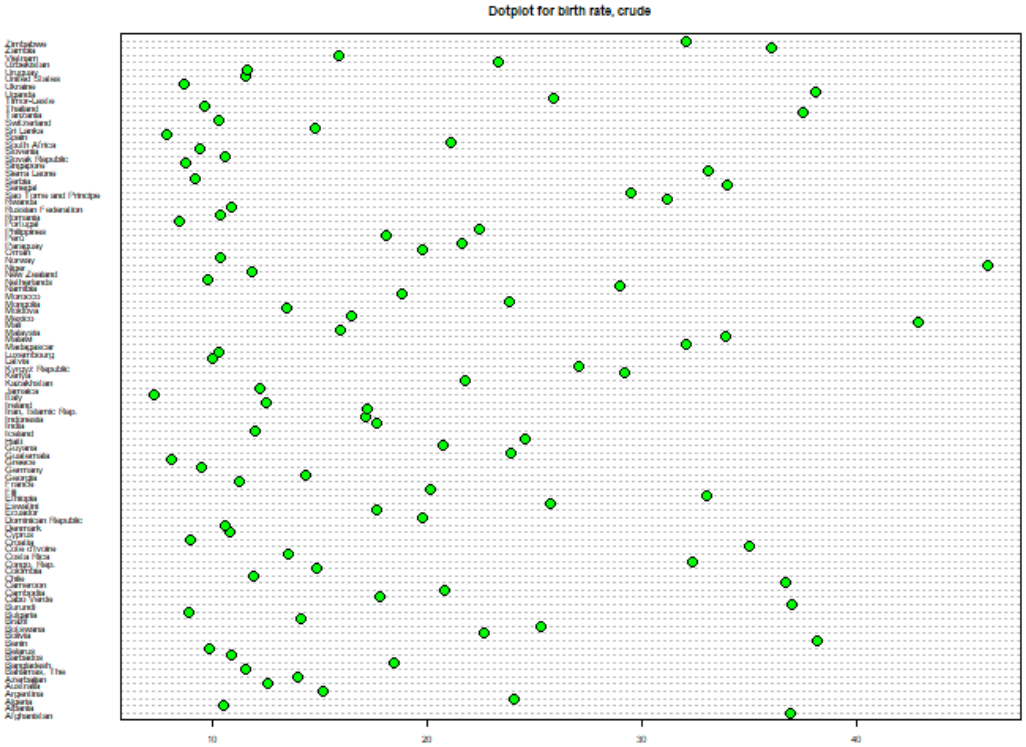
Для выявления аномальных данных и представления используются точечные диаграммы.



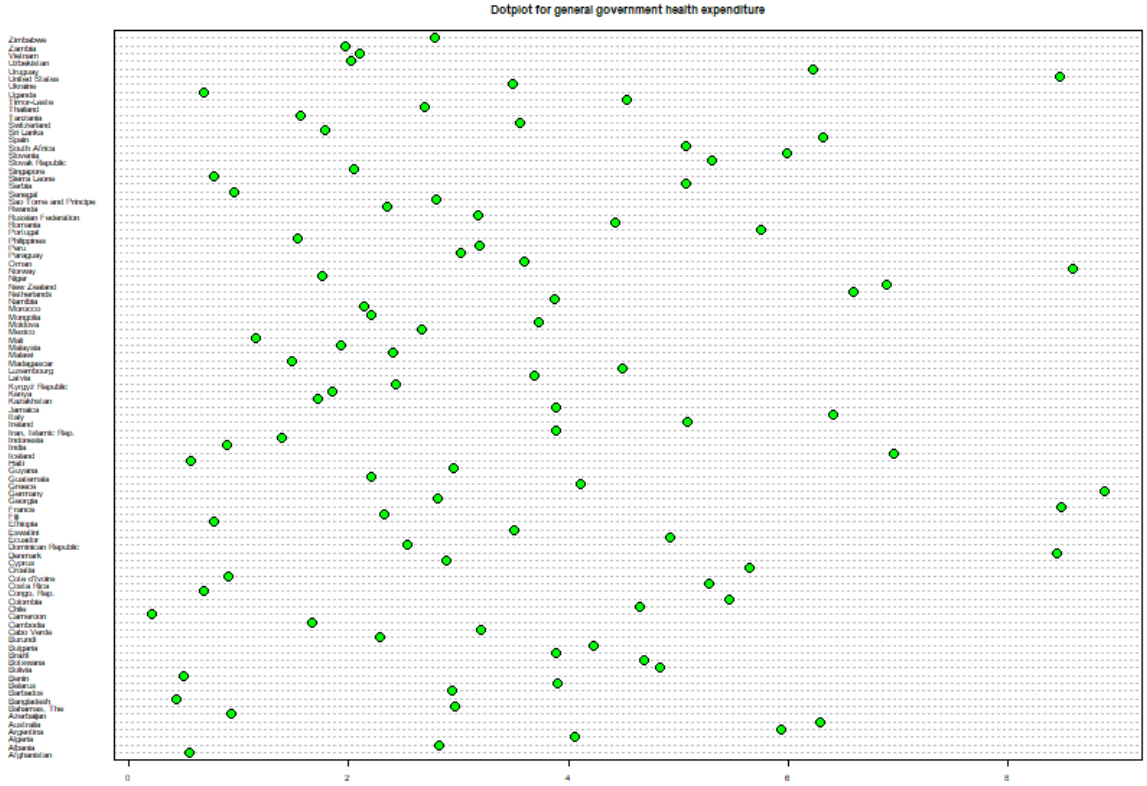
На представленной диаграмме видно, что в среднем ВВП стран колеблется от 0 до 20000 тысяч долларов. По мере увеличения ВВП, стран соответствующим значениям становится все меньше. Так же можно заметить явный выброс — крайнюю правую точку далеко отстоящую ото всех остальных.



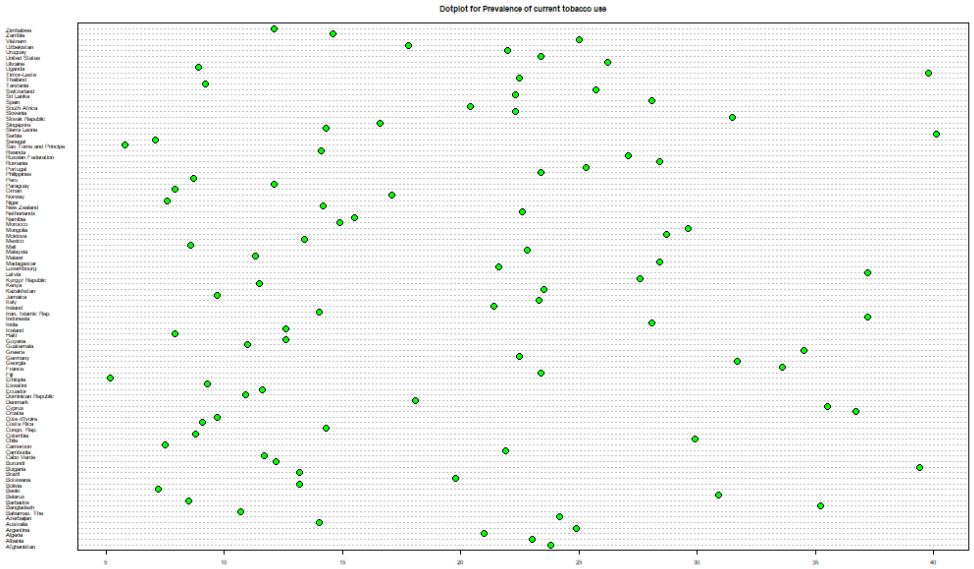
Точечная диаграмма по затратам государства на образование показывает, что в среднем на образование затрачивается от 2,5 % до 6% от ВВП, но так же есть и значения, достигающие около 10% и минимальное — около 0,5%.



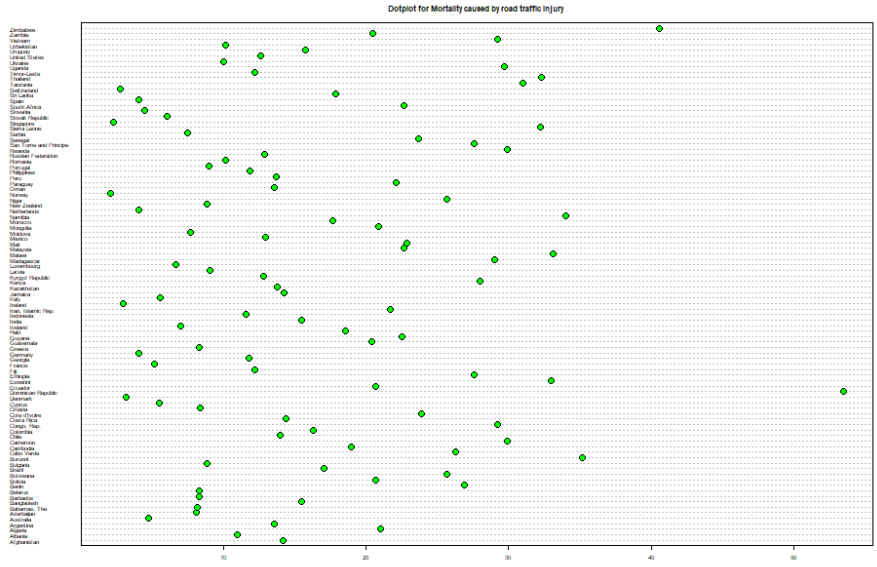
Точечная диаграмма для рождаемости показывает, что в среднем значения варьируются от 10 до 20. Гораздо меньше стран имеют коэффициент рождаемости ниже или выше.



По точечной диаграмме по затратам государства на здравоохранение можно видеть, что все значения достаточно сильно варьируются, но большинство стран тратят от 0 до 6% от ВВП на эти нужды.



Точечная диаграмма показывает, что процент употребления табака варьируется от 1 до 40% на 100000 человек, но определенную закономерность выявить трудно.



По точечной диаграмме по смертности от транспортно дорожных происшествий можно сказать, что в среднем в странах процент варьируется от 0,5 до 25%, а также присутствует явный выброс соответствующий >50%.

Выводы по точечным диаграммам о распространенности того или иного события в странах можно подтвердить диаграммами stem-plot.

```
The decimal point is 4 digit(s) to the right of the |

 0 | 011111111111111112222222223333333444444445556666666777777899001122 25
 2 | 00469025
 4 | 23837
 6 | 23749
 8 | 25
10 | 7
```

По графику для ВВП можно видеть, что большинство значений находятся на стебле «0», и количество элементов уменьшается с увеличением номера стебля. Таким образом, значению стебля «10» соответствует лишь одно значение.(Если увеличивать количество стеблей появляются разрывы)

```
The decimal point is at the |

0 | 69
2 | 1125668990001122223334556667799999
4 | 000222334455677778899999011112223344444444466799
6 | 01788066
8 | 999
```

Можно заметить, что на графике по затратам государства на образование, большинство значений находятся на стеблях «2» и «4» - самые распространенные значения.

The decimal point is 1 digit(s) to the right of the |

0 | 7889999999
1 | 00000000011111112222222333444455566677888889
2 | 000111222334445566799
3 | 01222334456777888
4 | 36

График по рождаемости показывает, что самым распространенным значением является лист «2», а минимальным по распространенности лист «8».

The decimal point is at the |

0 | 245567788999024556778899
2 | 000112233344577888899000222556677999999
4 | 1124557789111335689
6 | 02334690
8 | 45569

Самый распространенный лист в графике по затратам государства на образование- «2», наименее распространенный - «8».

The decimal point is 1 digit(s) to the right of the |

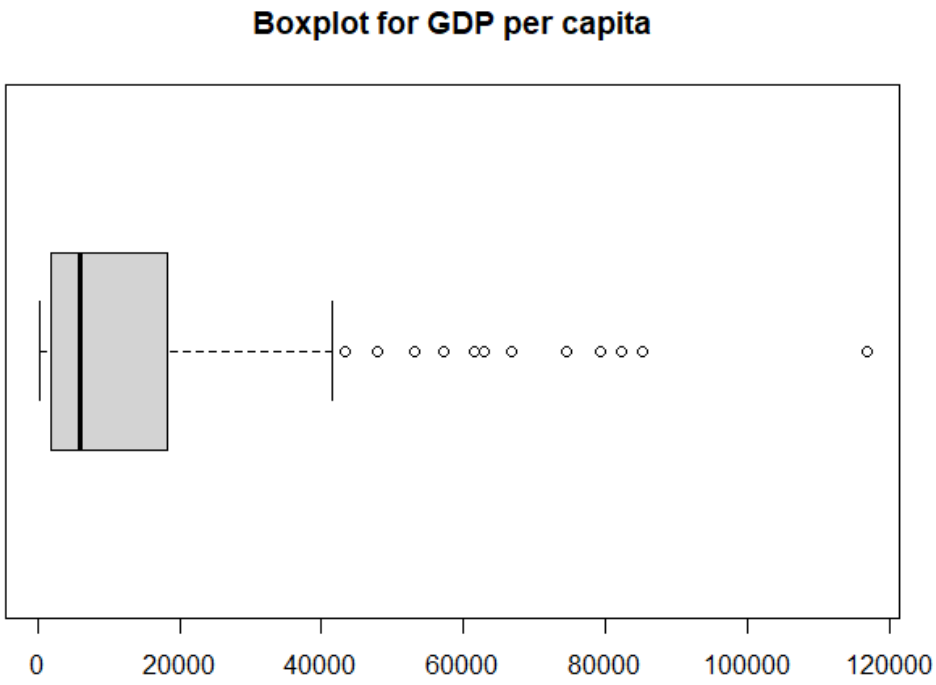
0 | 5677888899999999
1 | 00111122222233333444444556778 8
2 | 001122222333333333444555667888889
3 | 0012245567779
4 | 00

На точечном графике по употреблению табака было затруднительно выявить наиболее часто встречаемый интервал. Но на stem-plot можно сказать, что наиболее часто встречаемое значение находится на стебле «2», что говорит о том, что чаще всего встречаются страны с употреблением табака в интервале 20-30%.

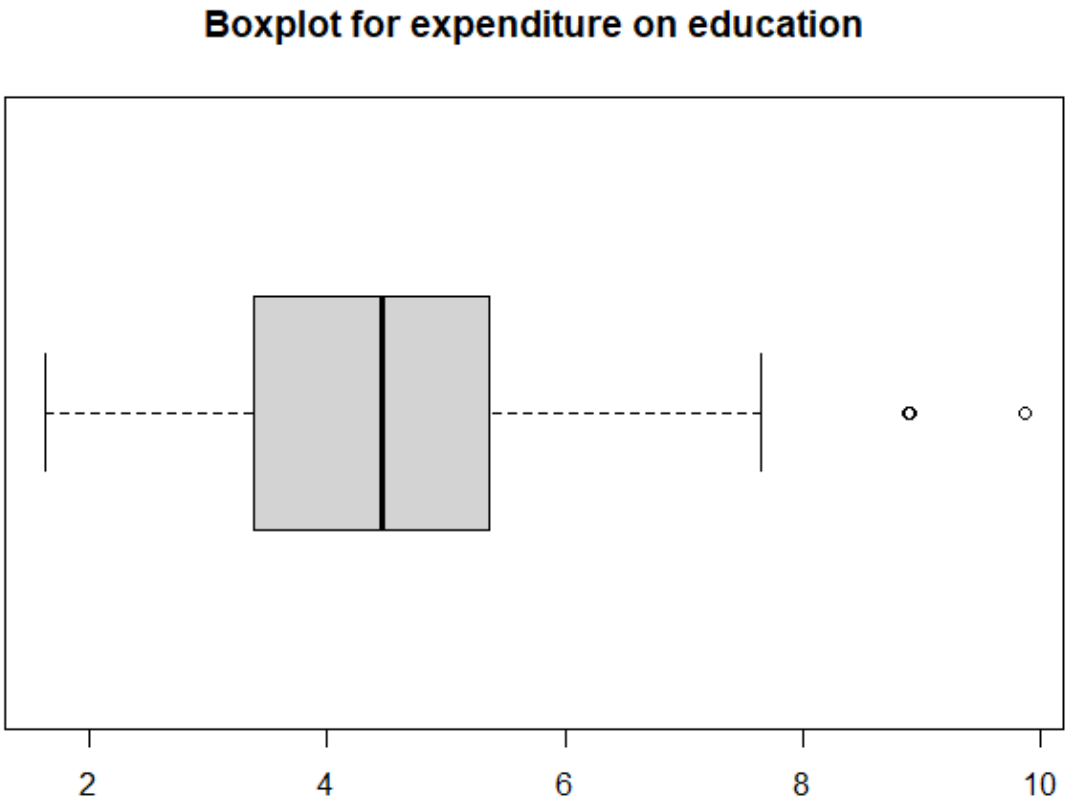
```
The decimal point is 1 digit(s) to the right of the |

0 | 22333444555666778888888889999
1 | 000122222333344444444666678899
2 | 011111223333446667888999
3 | 0001223345
4 | 1
5 | 4
```

На графике по распространенности транспортно-дорожных происшествий можно видеть, что распространенность стебля «0», «1», «2» примерно одинаковая, а самым редко встречающимся значением являются стебли «4» и «5».

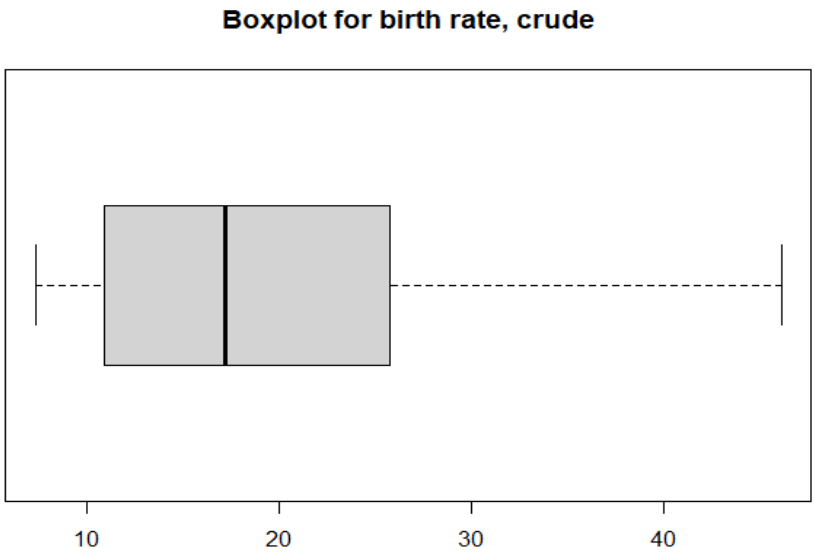
Для более качественного выявления аномальных значений используется диаграмма boxplot.



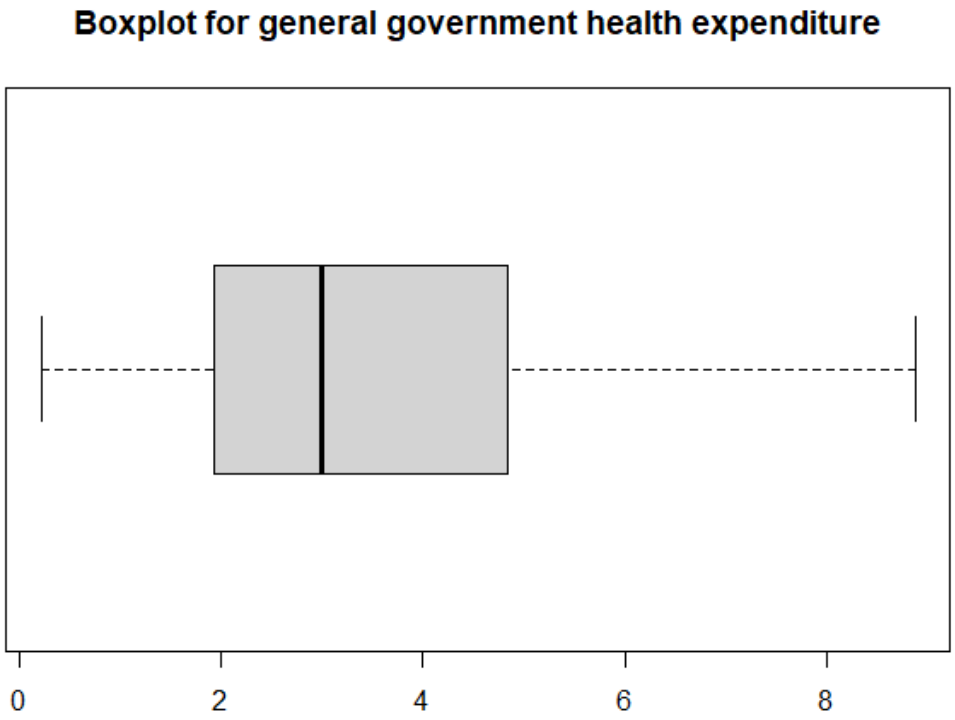
На ящечковой диаграмме по ВВП можно видеть количество выбросов и их распределение , среднее значение- чуть меньше 8000 , и интервал, в котором значения все еще являются пределом нормы — от 0 до приблизительно 40000.



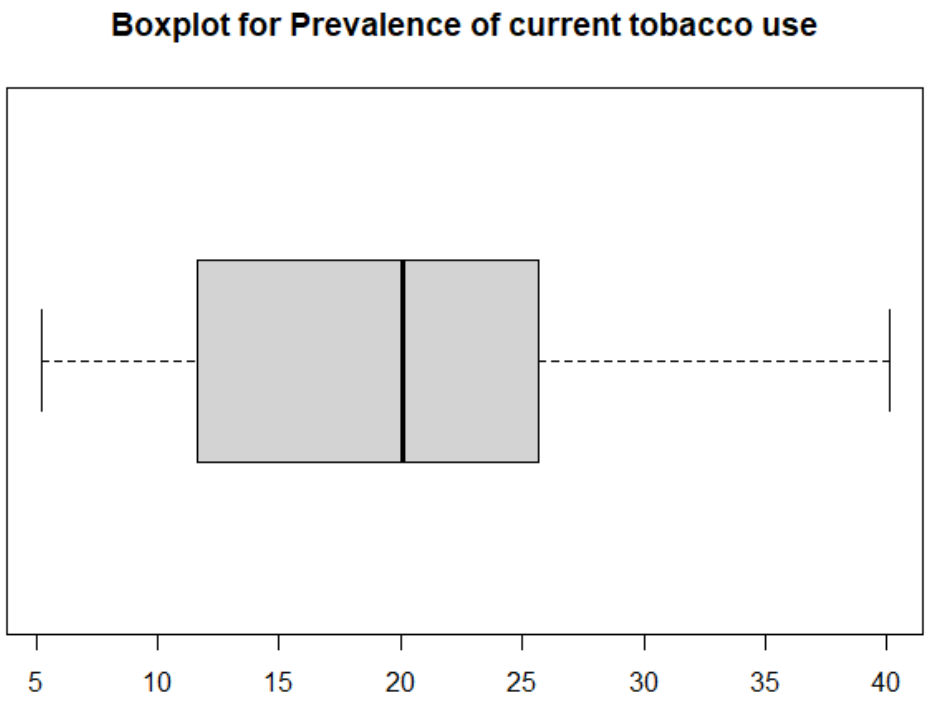
По ящечковой диаграмме можно сделать вывод, что в среднем государства тратят на образование чуть больше 4% от ВВП, допустимый интервал от 0 до 7,8% , а так же, что в выборке по тратам на образование присутствует два выброса.



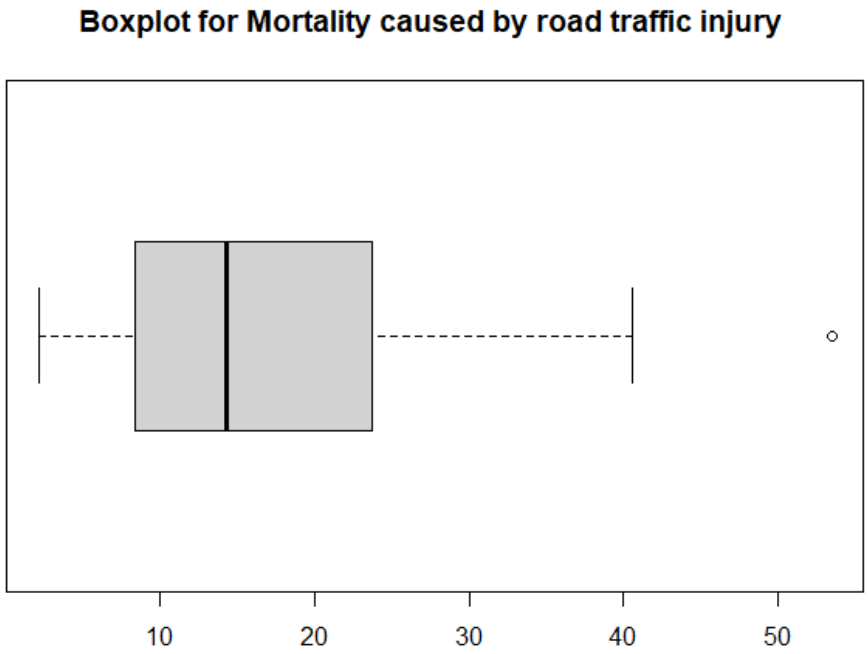
Ящичковая диаграмма по рождаемости показывает, что выбросов по этому параметру данных нет. А среднее значение находится на отметке около 18.



На ящичковой диаграмме по тратам государств на здравоохранение также нет выбросов, среднее значение находится на отметке в 3%.



На ящичковой диаграмме по распространенности употребления табака также не выявляется выбросов. Среднее значение около 20.



На диаграмме по смертности от транспортно-дорожных происшествий наблюдается один выброс и среднее значение — около 13.

Характеристики СВ

В этом разделе будут рассмотрены показатели случайных величин — параметров, изучаемых в этой модульной работе. Для визуализации была сделана таблица, содержащая показатели для всех величин.

	V1	V2	V3	V4	V5	V6
mean	15803.6991748397	4.5370822333275	19.5087978723404	3.46548947531915	19.613829787234	16.863829787234
mode	numeric	numeric	numeric	3.46548947531915	numeric	numeric
median	6084.12262356699	4.46461510658264	17.205	2.98980677	20.1	14.25
variance	116555.065178129	8.22776854038238	38.827	8.67133692	34.9	51.4
var	538792741.974586	2.30679923897566	96.8703877759094	4.45829293994664	86.6702367879204	102.885989476092
sd	23211.9094857486	1.51881507728086	9.84227553850782	2.11146701133279	9.30968510680788	10.1432731145372
koef_variance	1.46876432086882	0.334755906808187	0.504504460137045	0.609283919737871	0.474649021012064	0.601480994679851
iqr	16086.8481239483	1.94647735357285	14.71725	2.86541841	13.975	14.975
upper_3iqr	66395.4700898168	11.1948818564415	69.77225	13.395282795	67.525	68.425
lower_3iqr	-46212.4667778213	-2.43045961856844	-33.2485	-6.662646075	-30.3	-21.425
upper_1.5iqr	42265.1979038943	8.27516582608223	47.696375	9.09715518	46.5625	45.9625
lower_1.5iqr	-22082.1945918989	0.48925641179084	-11.172625	-2.36451846	-9.3375	1.0375
upper_3s	85439.4276320855	9.09352746517008	49.0356244878639	9.79989050931752	47.5428851076576	47.2936491308456
lower_3s	-53832.0292824061	-0.0193629985150796	-10.0180287431831	-2.86891155867922	-8.31522553318964	-13.5659895563776

Рассмотрим показатели для ВВП на душу населения по странам.

- Среднее значение оказалось равным 15803,69 доллара США.
- Моды не существует, что значит, что в выборке нет повторяющихся значений.
- Медиана равна 6084,1 — ниже этого значения находится ровно половина всех наблюдений
- Вариация(размах) — разница между максимальным и минимальным значениями оказалась равной 116555.
- Дисперсия равна 538792741,97 она показывает меру разброса данных вокруг среднего значения.
- Среднее квадратическое отклонение равно 2311,9 — также показывает разброс данных относительно среднего значения.
- Коэффициент вариации, который рассчитывается как частное от среднеквадратического отклонения и среднего значения, равен 1,468 — говорит о том, что данные очень сильно разбросаны относительно среднего значения
- Децили у данного показателя равны

10%	double [1]	902.7309
20%	double [1]	1595.865
30%	double [1]	3184.863
40%	double [1]	4183.877
50%	double [1]	6084.123
60%	double [1]	7226.834
70%	double [1]	12394.28
80%	double [1]	21357.57
90%	double [1]	51512.96

Итак, первый дециль равен 902,703, что значит что 10% выборки меньше этого значения, второй дециль равен 1596,865, это значит, что 20% выборки лежит ниже этого значения… и так далее.

- Квартили ВВП на душу населения соответственно равны

0%	double [1]	231.4465
25%	double [1]	2048.078
50%	double [1]	6084.123
75%	double [1]	18134.93
100%	double [1]	116786.5

Первый квартиль равен 2048,078, это значит что 25% значений данных меньше данного значения. Второй квартиль равен медиане и говорит о том, что половина значений данных меньше данного значения. В данном случае он равен 6084,123. Третий квартиль равен 116786,5, что говорит о том, что 75% данных лежит ниже данного значения.

- Интерквартильный размах равняется разнице между третьим и первым квартилями. В данном случае он равен 16086.

Далее рассчитываются границы для удаления аномальных данных по трем правилам.

1. Правило трех сигм — от среднего значения данных вычитается три среднеквадратических отклонения для нахождения нижней границы и прибавляется три среднеквадратических отклонения для нахождения

верхней границы. В данном примере эти границы по вычислениям равны соответственно -53832 и 85439,4, но так как значение не может принимать отрицательные значения, то нижнюю границу заменяем на 0 (Здесь и во всех границах ниже 0). По этому правилу получилось 1 аномальное значение.

2. Правило 1,5IQR – для нахождения нижней границы от первого квартиля отнимается 1,5 значения интерквартильного размаха, а для нахождения верхнего значения к третьему квартилю прибавляется 1,5 значения интерквартильного размаха. Таким образом, проделав вычисления, границы по правилу 1,5IQR получились равными -22082 и 42265. По данному правилу в этой выборке получилось 11 выбросов.
3. Правило 3IQR аналогично предыдущему правилу, но вместо 1,5IQR используется 3IQR. В данном случае границы получились равными -2265.19 и 66395,47. По данному правилу есть 6 аномальных значений.

Далее рассмотрим данные по затратам государства на образование:

- Среднее значение оказалось равным 4,5.
- Моды не существует, что значит, что в выборке нет повторяющихся значений.
- Медиана равна 4,46 — ниже этого значения находится ровно половина всех наблюдений
- Вариация(размах) оказалась равной 8,22.
- Дисперсия - квадрат среднего квадратического отклонения равна 2,306 она показывает меру разброса данных вокруг среднего значения.
- Среднее квадратическое отклонение равно 1,51 — также показывает разброс данных относительно среднего значения.
- Коэффициент вариации, равен 0,33 — говорит о том, что слабо разбросаны относительно среднего значения.
- Децили данной величины:

10%	double [1]	2.878672
20%	double [1]	3.22388
30%	double [1]	3.642475
40%	double [1]	3.987116
50%	double [1]	4.464615
60%	double [1]	4.862252
70%	double [1]	5.158292
80%	double [1]	5.41983
90%	double [1]	6.076507

- Квантили данной величины:

0%	double [1]	1.63207
25%	double [1]	3.408972
50%	double [1]	4.464615
75%	double [1]	5.35545
100%	double [1]	9.859838

- Межквартильный размах равен соответственно 1,946.
- Границы по правилу трех сигм: 0 и 9,09 — 1 выбросов
- Границы по правилу 1,5IQR: 0,48 и 8,275 — 3 выброса
- Границы по правилу 3IQR: 0 и 11,19 — 0 выбросов.

Третий рассматриваемый параметр — рождаемость по странам.

- Среднее значение оказалось равным 19,5.
- Моды не существует, что значит, что в выборке нет повторяющихся значений.
- Медиана равна 17,205— ниже этого значения находится ровно половина всех наблюдений
- Вариация(размах) оказалась равной 38,827.
- Дисперсия равна 96,87.

- Среднее квадратическое отклонение равно 9,8 — показывает разброс данных относительно среднего значения.
- Коэффициент вариации, равен 0,5 — говорит о том, что средне разбросаны относительно среднего значения.
- Децили данной величины:

10%	double [1]	9.43
20%	double [1]	10.4702
30%	double [1]	11.6045
40%	double [1]	14.0266
50%	double [1]	17.205
60%	double [1]	20.1128
70%	double [1]	23.8453
80%	double [1]	29.322
90%	double [1]	34.6914

- квартили данной величины равны

0%	double [1]	7.3
25%	double [1]	10.90325
50%	double [1]	17.205
75%	double [1]	25.6205
100%	double [1]	46.127

- Интерквартильный размах равен 14,71
- Границы по правилу трех сигм: 0 и 49,03 — нет выбросов
- Границы по правилу 1,5IQR: 0 и 47.69 — нет выбросов
- Границы по правилу 3IQR: 0 и 69.77 — 0 выбросов.

Четвертый рассматриваемый параметр — траты стран на здравоохранение.

- Среднее значение оказалось равным 3,46.
- Мода существует и равна 3,465, это самое часто встречающееся значение.
- Медиана равна 2,98— ниже этого значения находится ровно половина всех наблюдений

- Вариация(размах) оказалась равной 8,67.
- Дисперсия равна 4,45.
- Среднее квадратическое отклонение равно 2,11 — показывает разброс данных относительно среднего значения.
- Коэффициент вариации, равен 0,6 — говорит о том, что средне разбросаны относительно среднего значения.
- Децили данной величины:

10%	double [1]	0.8934349
20%	double [1]	1.693611
30%	double [1]	2.138386
40%	double [1]	2.675656
50%	double [1]	2.989807
60%	double [1]	3.718344
70%	double [1]	4.43697
80%	double [1]	5.163769
90%	double [1]	6.311656

- квартили данной величины равны

0%	double [1]	0.2104447
25%	double [1]	1.933609
50%	double [1]	2.989807
75%	double [1]	4.799028
100%	double [1]	8.881782

- Интерквартильный размах равен 2,865.
- Границы по правилу трех сигм: 0 и 9,79 — нет выбросов
- Границы по правилу 1,5IQR: 0 и 9,09 — нет выбросов
- Границы по правилу 3IQR: 0 и 69.77 — нет выбросов.

Пятый рассматриваемый параметр — распространенность употребления табака.

- Среднее значение оказалось равным 19,6.
- Мода отсутствует.

- Медиана равна 20,1— ниже этого значения находится ровно половина всех наблюдений
- Вариация(размах) оказалась равной 34,9.
- Дисперсия равна 86,6.
- Среднее квадратическое отклонение равно 9,3 — показывает разброс данных относительно среднего значения.
- Коэффициент вариации, равен 0,47 — говорит о том, что средне разбросаны относительно среднего значения.
- Децили данной величины:

10%	double [1]	8.63
20%	double [1]	10.82
30%	double [1]	12.56
40%	double [1]	14.3
50%	double [1]	20.1
60%	double [1]	22.5
70%	double [1]	23.84
80%	double [1]	28.1
90%	double [1]	33.03

- квартили данной величины равны

0%	double [1]	5.2
25%	double [1]	11.625
50%	double [1]	20.1
75%	double [1]	25.6
100%	double [1]	40.1

- Интерквартильный размах равен 13,975.
- Границы по правилу трех сигм: 0 и 47,5 — нет выбросов
- Границы по правилу 1,5IQR: 0 и 46,5625 — нет выбросов
- Границы по правилу 3IQR: 0 и 67,525 — нет выбросов.

Шестой рассматриваемый параметр — смертность от дорожно-транспортных происшествий.

- Среднее значение оказалось равным 16,86.

- Мода отсутствует.
- Медиана равна 14,25 — ниже этого значения находится ровно половина всех наблюдений
- Вариация(размах) оказалась равной 51,4.
- Дисперсия равна 102,88.
- Среднее квадратическое отклонение равно 10,14 — показывает разброс данных относительно среднего значения.
- Коэффициент вариации, равен 0,6 — говорит о том, что средне разбросаны относительно среднего значения.
- Децили данной величины:

10%	double [1]	4.92
20%	double [1]	8.16
30%	double [1]	9.91
40%	double [1]	12.64
50%	double [1]	14.25
60%	double [1]	18.46
70%	double [1]	22.14
80%	double [1]	26.54
90%	double [1]	29.9

- квартили данной величины равны

0%	double [1]	2.1
25%	double [1]	8.525
50%	double [1]	14.25
75%	double [1]	23.5
100%	double [1]	53.5

- Интерквартильный размах равен 14,975.
- Границы по правилу трех сигм: 0 и 47,29 — 1 выброс.
- Границы по правилу 1,5IQR: 1,03 и 45,96 — 1 выброс.

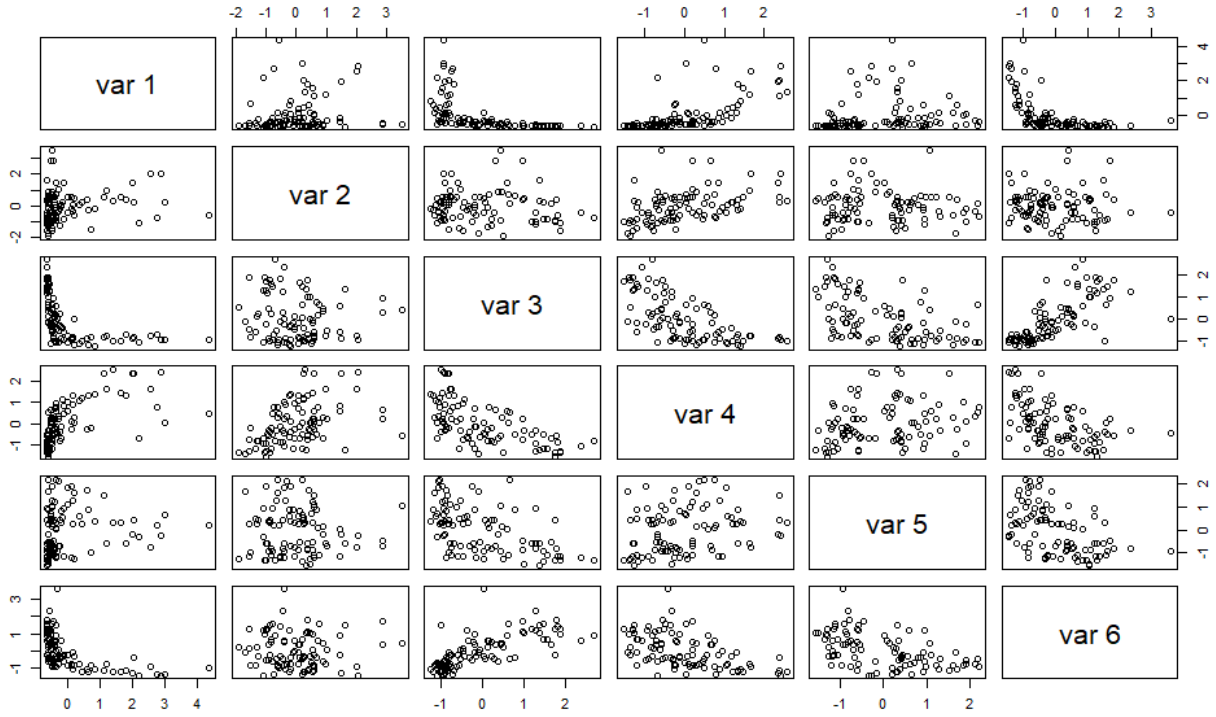
- Границы по правилу 3IQR: 0 и 68,425 — нет выбросов.

Далее необходимо стандартизировать данные(см. приложение 3) для этого была использована функция scale в R.

Выводы: в этом разделе были визуализированы начальные данные, а также были подсчитаны их основные характеристики. Для дальнейшего анализа для удаления аномальных данных будет применяться метод 1,5 IQR, так как он наиболее точно удалит выбросы.

Корреляционный анализ

Для корреляционного анализа изначально были построены корреляционные облака для всех признаков.



На данном рисунке видно, что

- между ВВП и тратами на образование связи нет
- между ВВП и рождаемостью есть связь — возможно гиперболическая, сильная и обратная по направлению
- между ВВП и затратами на здравоохранение так же есть связь — возможно гиперболическая, прямая по направлению и средняя по силе связи
- между ВВП и употреблением табака связи нет .
- между ВВП и смертностью от дорожно-транспортных происшествий есть связь — возможно гиперболическая или экспоненциальная, обратная по направлению и сильная по силе связи.
- между затратами государства на образование и рождаемостью связи нет
- между тратами на образование и затратами на здравоохранение связи нет

- между затратами государства на образование и употребление табака взаимосвязей нет
- между затратами государства на образование и смертностью от дорожных происшествий взаимосвязей нет
- между рождаемостью и тратами на затратами государства на здравоохранение есть слабая взаимосвязь, обратная по направлению
- нет взаимосвязей также между употреблением табака и смертностью от дорожно транспортных происшествий, между затратами на здравоохранение и употреблением табака, тратами на здравоохранение и смертностью от дорожно-транспортных происшествий.

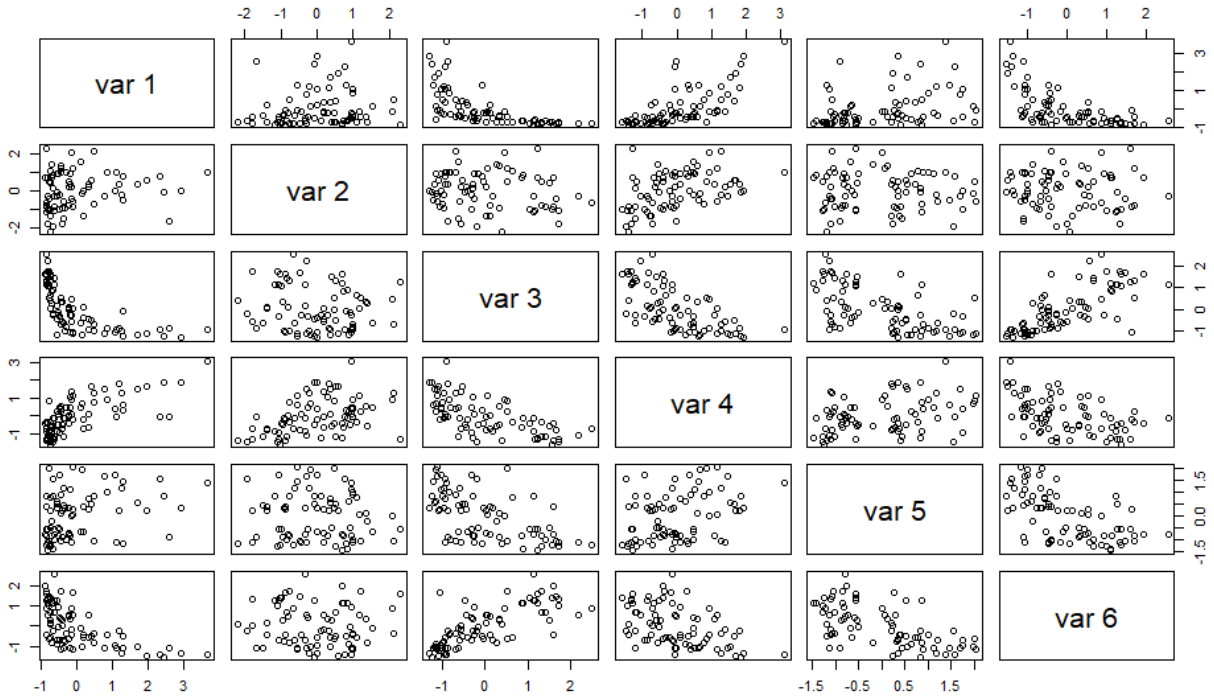
Так как выбросы не были удалены, видно, что некоторые точки сильно отстоят от других.

Далее была построена матрица парных коэффициентов корреляции.

	V1	V2	V3	V4	V5	V6
1	1.0000000	0.15839772	-0.5147299	0.6095958	0.10842737	-0.57738383
2	0.1583977	1.00000000	-0.1043873	0.4399404	-0.05317809	-0.05482441
3	-0.5147299	-0.10438729	1.0000000	-0.6311818	-0.51492279	0.71068272
4	0.6095958	0.43994040	-0.6311818	1.0000000	0.23588535	-0.52312127
5	0.1084274	-0.05317809	-0.5149228	0.2358853	1.00000000	-0.49889621
6	-0.5773838	-0.05482441	0.7106827	-0.5231213	-0.49889621	1.00000000

Корреляция не учитывает нелинейных зависимостей, поэтому в некоторых ячейках можно наблюдать значения средней корреляции, когда взаимосвязь точно есть. В других случаях корреляционная матрица подтверждает выводы, сделанные на основе графиков.

Далее производится удаление аномальных данных по правилу 1,5IQR. После удаления выборка данных сократилась до 78 значений. Так выглядят ее корреляционные поля:



Можно заметить, что связи между образованием и другими параметрами как не было, так и нет, нелинейные корреляции не изменились и теперь, может, даже более сильные. Связи между употреблением табака и другими параметрами нет.

Теперь рассмотрим матрицу корреляций после удаления аномальных данных.

	V1	V2	V3	V4	V5	V6
1	1.0000000	0.13748854	-0.6385435	0.6960709	0.3282909	-0.62267985
2	0.1374885	1.00000000	-0.1374585	0.4232827	-0.0469317	-0.05895168
3	-0.6385435	-0.13745847	1.0000000	-0.6448528	-0.5695756	0.73408029
4	0.6960709	0.42328272	-0.6448528	1.0000000	0.3522405	-0.49277744
5	0.3282909	-0.04693170	-0.5695756	0.3522405	1.0000000	-0.60070234
6	-0.6226798	-0.05895168	0.7340803	-0.4927774	-0.6007023	1.00000000

Можно заметить, что

- корреляция между ВВП и рождаемостью увеличилась на 0,12
- корреляция между ВВП и тратами на здравоохранение возросла на 0,09

- другие корреляции незначительно уменьшились или увеличились(изменения меньше 0,05)
- там где корреляция между параметрами была незначительной, она так и осталась очень маленькой

Выводы о связи в корреляциях:

- Траты государства на образование и медицину имеют среднюю положительную связь, но траты на образование в сочетании с другими параметрами имеет слабую связь.
- ВВП имеет в основном нелинейную, но достаточно сильную взаимосвязь с другими параметрами(кроме образования и табака).
- Рождаемость имеет среднюю (ближе к сильной) обратную по направлению связь с затратами государства на здравоохранение.
- Траты на здравоохранение имеют среднюю обратную связь со смертностью от транспортно-дорожных происшествий.
- Употребление табака имеет среднюю связь только со смертностью от транспортно-дорожных происшествий.

Далее была построена матрица частных коэффициентов корреляции

	V1	V2	V3	V4	V5	V6
1	1.0000000	-0.19944590	-0.12068363	0.5220391	-0.2009974	-0.35243690
2	-0.1994459	1.00000000	0.04137969	0.4759762	-0.1735443	-0.03530037
3	-0.1206836	0.04137969	1.00000000	-0.3217182	-0.2328014	0.41959434
4	0.5220391	0.47597624	-0.32171824	1.0000000	0.1282132	0.13623983
5	-0.2009974	-0.17354429	-0.23280139	0.1282132	1.0000000	-0.36384711
6	-0.3524369	-0.03530037	0.41959434	0.1362398	-0.3638471	1.00000000

Коэффициенты частных корреляций показывают взаимосвязь между параметрами без учета влияния других корреляций (они фиксируются)

Таким образом можно видеть, что существует средняя взаимосвязь между затратами на образование и медицину. ВПП имеет среднюю связь с тратами на медицину. Употребление табака и смертность от транспортно-дорожных происшествий имеет слабую (ближе к средней) отрицательную по

направлению связь. А также смертность от дорожно-транспортных происшествий имеет слабую(ближе к средней) отрицательно направленную связь с ВВП на душу населения.

Сравнивая матрицу коэффициентов частных корреляций с матрицей парных корреляций, можно заметить, что все значения корреляций уменьшились. Это значит, что каждый показатель влиял на другой, повышая его корреляцию.

Найдем множественный коэффициент корреляции для ВВП на душу населения.

Для этого необходимо найти определитель матрицы парных коэффициентов, он равен 0,047, и алгебраическое дополнение элемента, для которого ищется коэффициент множественной корреляции (то есть определитель матрицы при вычеркивании соответствующей строки и столбца). Для такой матрицы определитель получился равным 0,303. Далее по формуле необходимо поделить детерминант целой матрицы на алгебраическое дополнение. Получается 0,1561531.Далее по формуле $\sqrt{1 - 0{,}1561531} = 0{,}9186$. Получили коэффициент множественной корреляции для ВВП на душу населения. Связь между ВВП на душу населения и другими параметрами достаточно высокая, о направлении связи не можем говорить, так как этот коэффициент измеряется от 0 до 1.

Выводы: был проведен корреляционный анализ, включающий в себя изучение полей корреляции и матриц парных коэффициентов корреляции до и после удаления выбросов, также была построена матрица частных коэффициентов и рассчитан коэффициент парной корреляции. Между многими параметрами, взятыми для анализа, даже после удаления аномальных значений отсутствует даже средняя связь. Так же при построении облаков корреляции было обнаружено присутствие нелинейной связи между параметрами.

Кластерный анализ

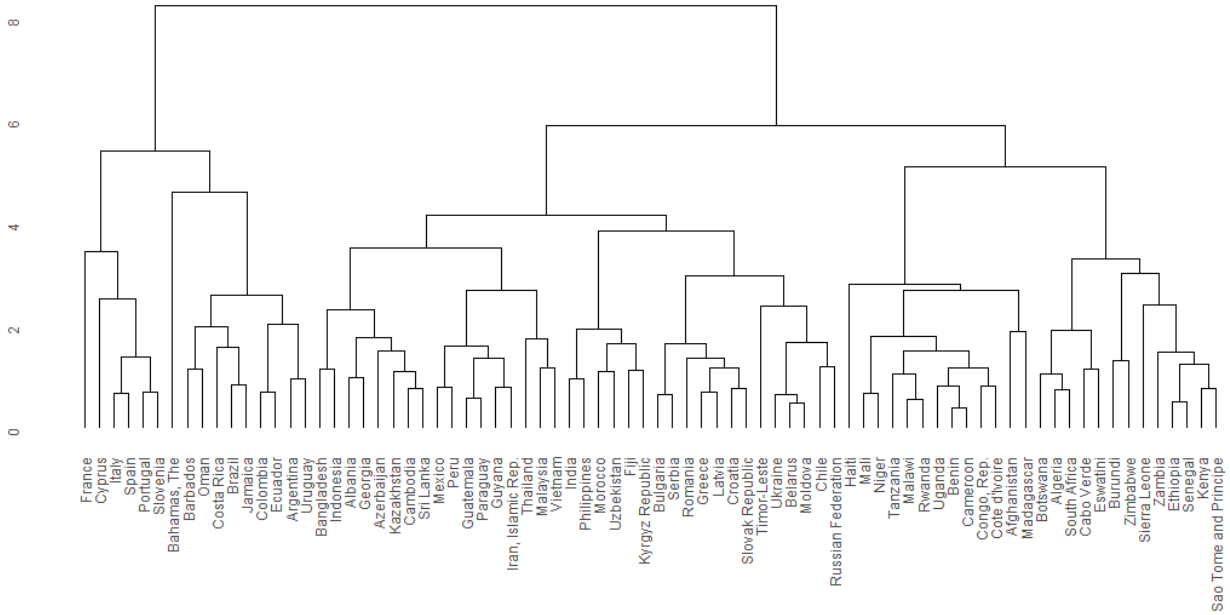
Для кластерного анализа будет использоваться евклидова мера расстояния, он подходит для дальнейшего анализа так как ведется работа со стандартизованными данными.

Иерархические методы кластеризации

Изначально была построена дендрограмма по методу ближнего соседа.

В этом методе расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами в различных кластерах, но получился не самый удачный пример для анализа(см приложение).

Далее была построена дендрограмма по методу дальнего соседа.



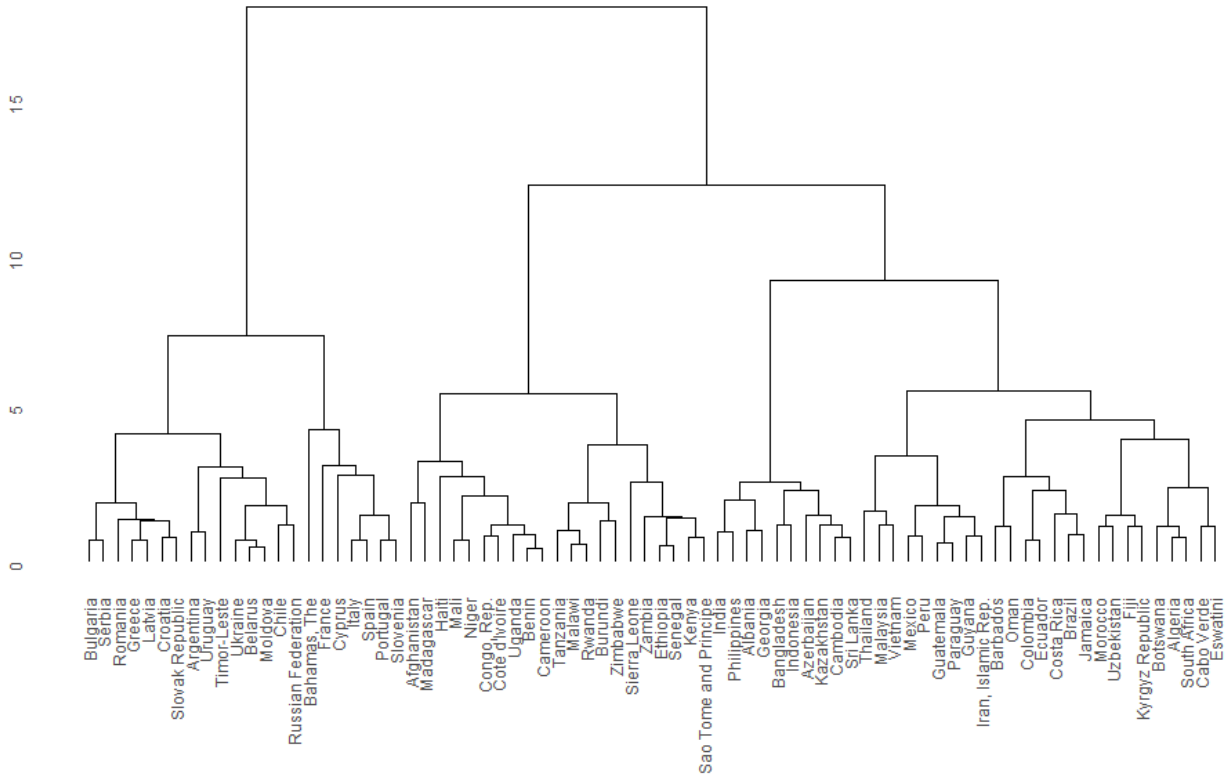
На ней отчетливо видно, что все объекты разбиваются на три кластера, расстояние друг до друга у них достаточно большое, чтоб сказать что это показательный пример. Так же можно заметить, что кластеры образуют некую структуру (хоть и не до конца однозначную) — самый правый кластер отражает бедные африканские страны, средний кластер - средне развитые страны, а левый страны побогаче.

Далее была построена дендрограмма по методу среднего

На данной дендрограмме, хоть и не так отчетливо заметны три кластера, а так же есть странное значение Франции, похожее на выброс. Эта

дендрограмма не так хорошо отображает наличие трех кластеров из-за того, что расстояние между ними крайне мало.

Затем была построена дендрограмма по методу Варда.



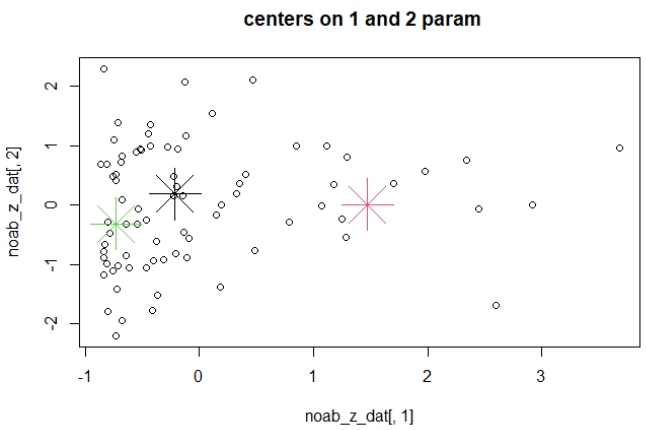
На этой дендрограмме можно выделить три основных кластера, которые объединяются в себя примерно на одном расстоянии, но до друг другу них достаточно большое расстояние. Можно также рассмотреть приблизительную структуру их разбиения — слева по большей части страны, с хорошо развитой экономикой и инфраструктурой, средний — бедные страны, левый — средние страны.

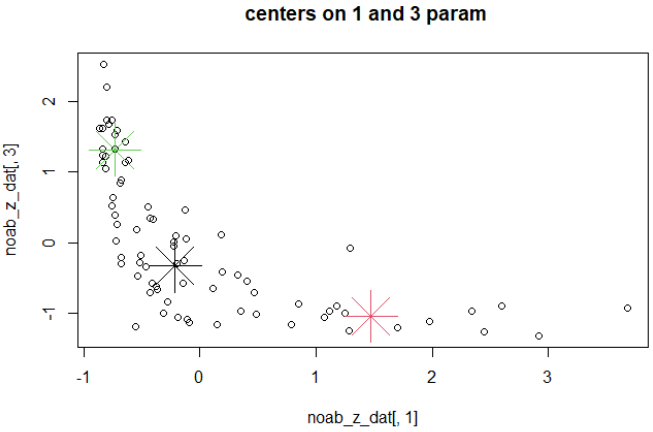
На основе интерпретируемых дендрограмм, а также по логическом размышлении о возможном распределении стран в реальном мире можно сделать предположить, что оптимальное количество кластеров — 3.

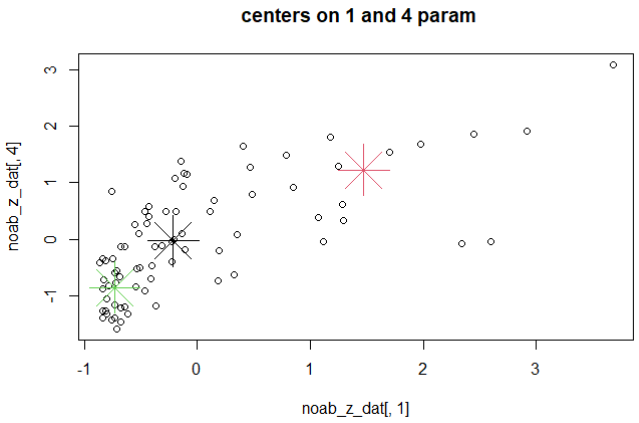
Метод k-средних

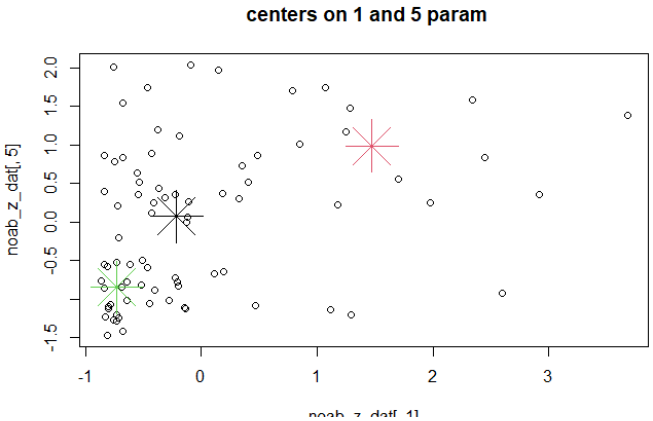
Для метода к- средних использовалась евклидова мера расстояния, потому что данные либо нормированные, либо стандартизированные. После построения графиков по нормированным и стандартизированным данным, можно сделать вывод, что они отличаются незначительно.

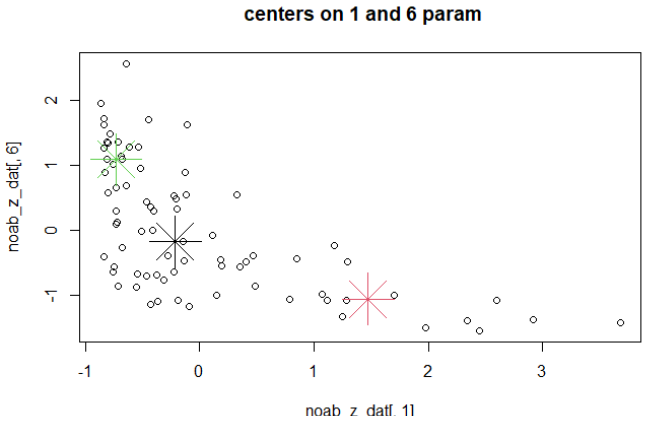
Получившиеся графики средних значений кластеров:

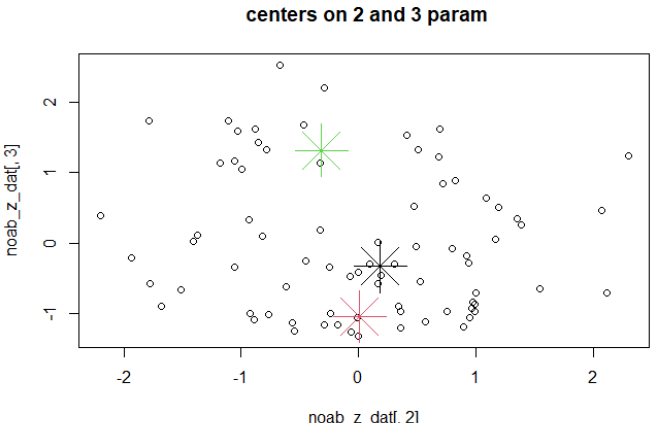


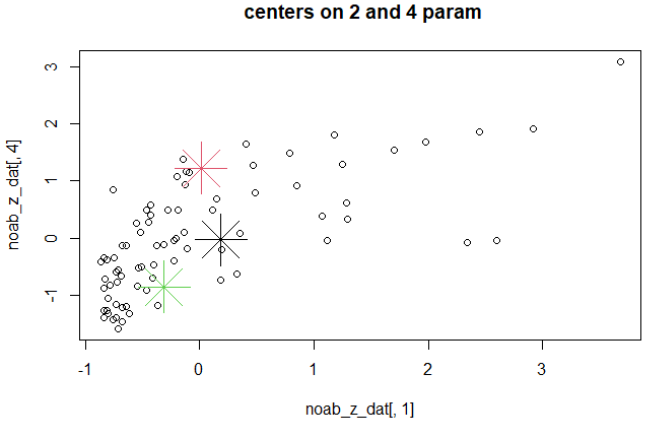


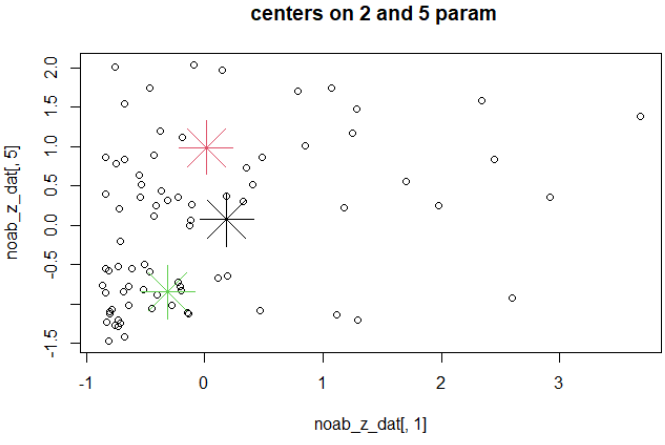




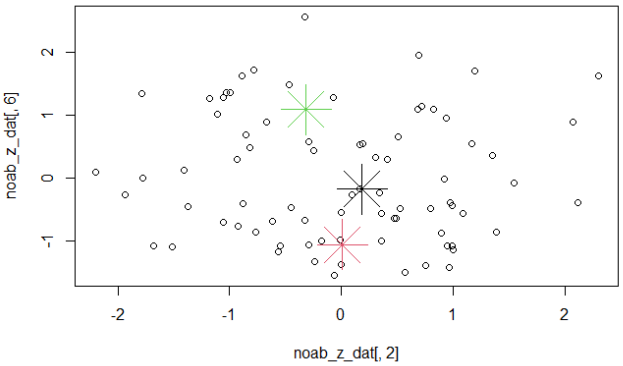




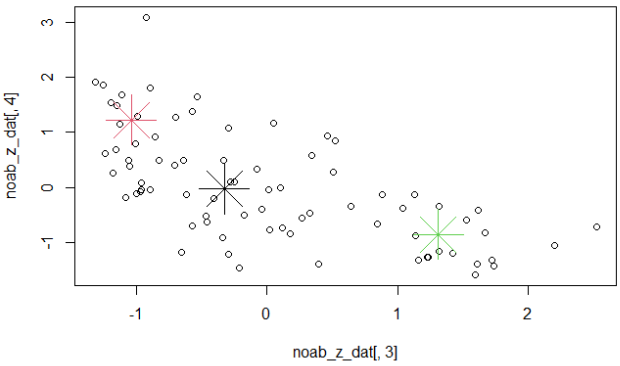




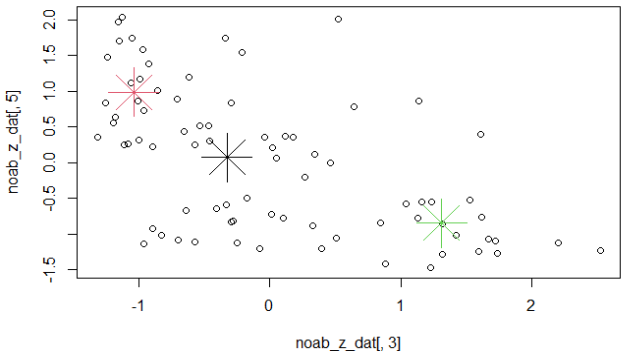
centers on 2 and 6 param



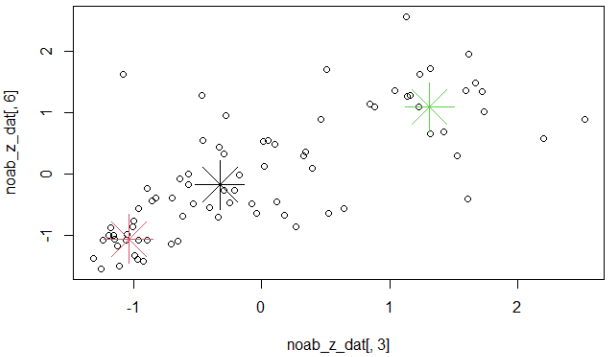
centers on 3 and 4 param



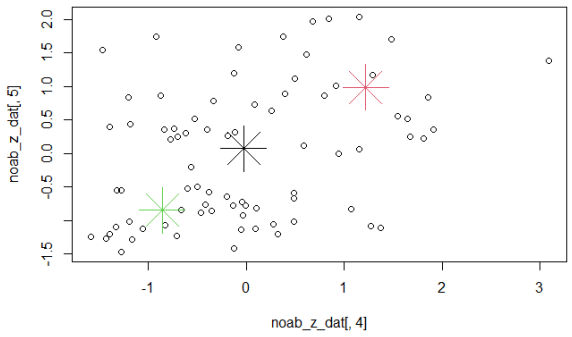
centers on 3 and 5 param



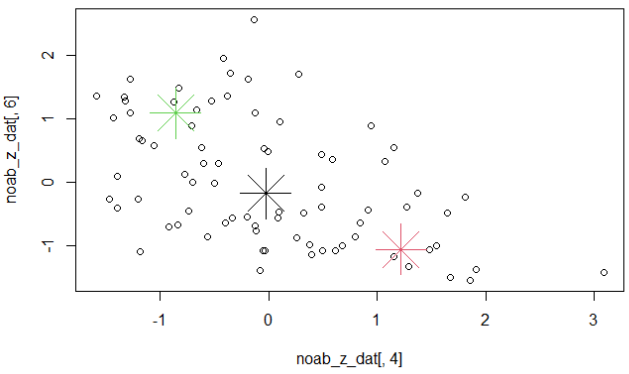
centers on 3 and 6 param



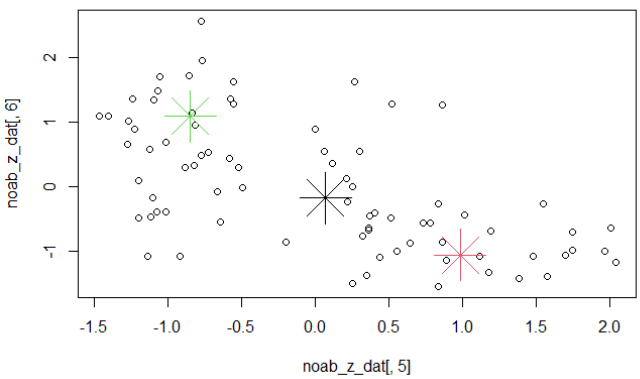
centers on 4 and 5 param



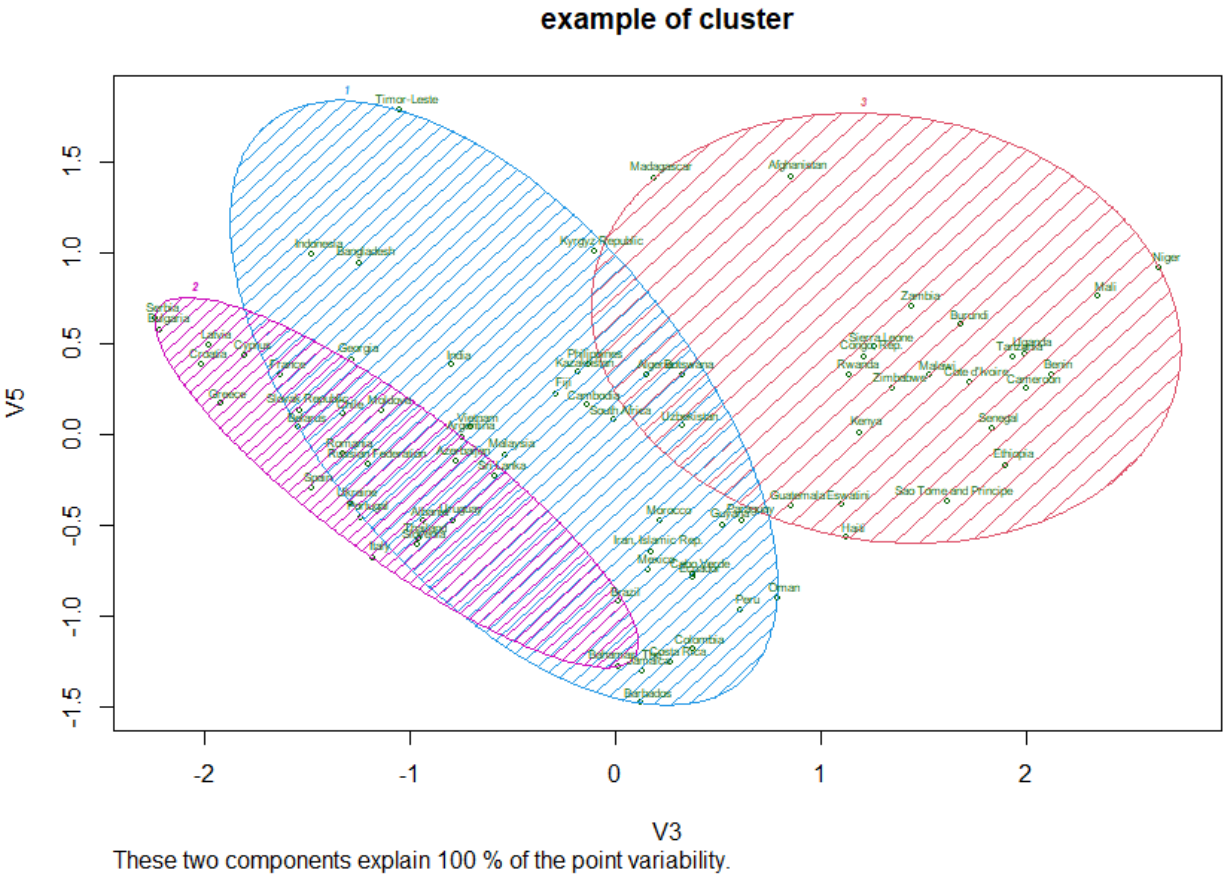
centers on 4 and 6 param



centers on 5 and 6 param



На некоторых графиках понятно, что средние значения кластеров иногда лежат слишком близко.
Пример получившегося кластера:



These two components explain 100 % of the point variability.

На данном кластере видно три основных разбиения:

1. Страны с высоким развитием - розовый кластер (например, Франция, Испания)
2. Страны со средним развитием — синий кластер (например, Индия)
3. Страны с низким развитием — красный кластер(например, Зимбабве, Нигер)

Видно, что кластеры имеют точки пересечения — не удалось точно определить все страны в один кластер. Но в среднем получилось такое разбиение по содержанию стран внутри кластеров: в 1 -38 стран, во 2 -17, в 3 — 23.

Вывод

Для анализа были подобраны данные по шести параметрам, в выборке было 94 страны. Для первичного анализа были визуализированы эти данные, подсчитаны их характеристики, найдены аномальные значения. Далее был проведен корреляционный анализ, включающий в себя нахождение различных матриц корреляций, и в ходе их построения были доказаны и опровергнуты некоторые гипотезы наличии связей между параметрами. Так, например, связь между ВВП и рождаемостью действительно есть связь, но обратная по силе направлению. Далее было выявлено, что параметры оказывают среднее влияние друг на друга в парной корреляции.

Затем кластерный анализ показал, что данные страны можно разбить на три кластера по уровню развития, как и предполагалось.

Приложение 1

Данные для анализа

Country Name	GDP per capita	expenditure on education	Birth rate, crude	general government health expenditure	Prevalence of current tobacco use	Mortality caused by road traffic injury
Afghanistan	502,056770622973	3,19979000091553	36,927	0,54922014	23,8	14,2
Albania	5287,66369446913	3,15490126609802	10,517	2,82499456	23	11
Algeria	4171,79501086373	5,8663501739502	24,074	4,05819511	21	21
Argentina	11795,1593866287	4,87773990631104	15,187	5,93580294	24,9	13,6
Australia	57207,8715094806	5,12344980239868	12,6	6,29709959	14	4,8
Azerbaijan	4739,84171028393	2,45543003082275	14	0,92868197	24,2	8,1
Bahamas, The	31738,2671569969	2,24967002868652	11,566	2,96176958	10,7	8,2
Bangladesh	1963,41185467584	1,94795548915863	18,462	0,42589214	35,2	15,5
Barbados	18224,8905923744	5,42886018753052	10,913	2,93843102	8,5	8,3
Belarus	6360,06247301284	5,38111019134521	9,9	3,9000001	30,9	8,3
Benin	1194,43821427144	2,93161988258362	38,186	0,4908835	7,2	26,9
Bolivia	3471,00695064499	8,89999961853027	22,634	4,835742	13,2	20,7
Botswana	6947,80081217637	6,72131013870239	25,328	4,68888426	19,8	25,7
Brazil	9121,08340257783	6,08851003646851	14,133	3,88923883	13,2	17,1
Bulgaria	9446,71709419731	4,04530000686646	8,9	4,23249435	39,4	8,9
Burundi	231,446476548394	5,07864999771118	37,011	2,28904033	12,2	35,2
Cabo Verde	3442,74200302219	5,36885023117065	17,821	3,20918369	11,7	26,3
Cambodia	1533,31598466695	2,57805943489075	20,805	1,66258287	21,9	19
Cameroon	1594,05998970315	3,03096008300781	36,716	0,21044466	7,5	29,9
Chile	15795,7084630954	5,43316984176636	11,94	4,65451193	29,9	14
Colombia	6782,03792033195	4,44910001754761	14,841	5,46089888	8,8	16,3
Congo, Rep.	2512,38397871635	2,99548006057739	32,337	0,68406534	14,3	29,2
Costa Rica	12383,1499737386	6,76991987228394	13,562	5,28166151	9,1	14,4
Cote d'Ivoire	2275,4959502948	3,23993992805481	35,007	0,90654147	9,7	23,9
Croatia	15244,4310483476	3,90755009651184	9	5,64517641	36,7	8,4
Cyprus	29418,935546875	5,15285396575928	10,843	2,889117	35,5	5,5
Denmark	61591,9288698958	6,79277992248535	10,6	8,44737244	18,1	3,2
Dominican Republic	7947,15820732109	3,92848992347717	19,762	2,54002523	10,9	53,5
Ecuador	6321,34940071717	4,6214599609375	17,658	4,92990112	11,6	20,7
Eswatini	4020,27305022394	5,67916011810303	25,718	3,51297188	9,3	33
Ethiopia	758,299059582623	5,06867980957031	32,996	0,77089483	5,2	27,6
Fiji	6073,33638842434	4,83571004867554	20,187	2,31982136	23,4	12,2
France	41557,8548588876	5,40716981887817	11,3	8,49391079	33,6	5,2
Georgia	4722,0424231414	3,5210599899292	14,395	2,80701351	31,7	11,8
Germany	47939,2782884504	4,97576999664307	9,5	8,88178158	22,5	4,1
Greece	19756,990456255	3,59734010696411	8,1	4,1116519	34,5	8,3
Guatemala	4485,73125489372	3,13704991340637	23,929	2,20523524	11	20,4
Guyana	6094,90885870963	4,45058012008667	20,77	2,95149136	12,6	22,5
Haiti	1494,2249622283	1,63206994533539	24,579	0,55767161	7,9	18,6
Iceland	74461,479998678	7,56170988082886	12	6,9625082	12,6	7
India	1974,37778792366	4,36373996734619	17,651	0,8878178	28,1	15,5
Indonesia	3902,6616681577	3	17,179	1,395854	37,2	11,6
Iran, Islamic Rep.	3873,99534020792	3,9553599357605	17,231	3,88674212	14	21,7
Ireland	79250,3878517684	3,39286994934082	12,5	5,08517361	21,4	3
Italy	34622,1696664741	4,25614023208618	7,3	6,40597105	23,3	5,6
Jamaica	5594,49393466948	5,41380977630615	12,22	3,88880515	9,7	14,3
Kazakhstan	9812,62637077396	2,61595010757446	21,77	1,71429646	23,5	13,8
Kenya	1845,78294001398	5,10761976242065	29,18	1,85613954	11,5	28
Kyrgyz Republic	1308,14016549619	5,55267000198364	27,1	2,4247191	27,6	12,8
Latvia	17865,0310947642	4,24356985092163	10	3,69703722	37,2	9,1
Luxembourg	116786,511654677	3,64749002456665	10,3	4,48882246	21,6	6,7
Madagascar	512,543984587107	2,84409999847412	32,088	1,48056126	28,4	29
Malawi	537,932204050438	3,32295989990234	33,925	2,40220785	11,3	33,1
Malaysia	11074,0640947671	4,47865009307861	15,978	1,92385709	22,8	22,7
Mali	856,356531797333	3,9055700302124	42,901	1,15118265	8,6	22,9
Mexico	9857,02882925559	4,2542200088501	16,465	2,6718502	13,4	13
Moldova	4232,20688845801	5,43973016738892	13,443	3,72367048	28,7	7,7
Mongolia	4165,02273851435	9,85983848571777	23,836	2,21058345	29,6	20,9
Morocco	3492,67333984375	5,34884977340698	18,813	2,14312124	14,9	17,7
Namibia	5687,3990933188	8,88111972808838	29,008	3,8711977	15,5	34
Netherlands	53044,5324352253	5,35764980316162	9,8	6,59804344	22,6	4,1
New Zealand	43250,440973659	6,04850006103516	11,84	6,89220285	14,2	8,9
Niger	567,330806700397	3,457279920578	46,127	1,76169312	7,6	25,7
Norway	82267,809316159	7,64411020278931	10,4	8,59087181	17,1	2,1
Oman	19887,5743113149	5,20723009109497	19,816	3,59855247	7,9	13,6
Paraguay	6242,96145417898	3,28320002555847	21,664	3,01784396	12,1	22,1
Peru	6912,11029696274	3,71429991722107	18,092	3,19640088	8,7	13,7
Philippines	3194,67452128933	3,86405324935913	22,445	1,53632784	23,4	11,9

Portugal	23562,5545228191	4,67515993118286	8,5	5,75663471	25,3	9
Romania	12494,4774670309	3,34474992752075	10,4	4,43120813	28,4	10,2
Russian Federation	11287,3603515625	4,67819976806641	10,9	3,18402338	27,1	12,9
Rwanda	769,437310706602	3,07375001907349	31,158	2,34513021	14,1	29,9
Sao Tome and Principe	1950,62935180937	5,23510980606079	29,535	2,80085683	5,8	27,6
Senegal	1484,2396733777	4,85822010040283	33,955	0,96029741	7,1	23,7
Serbia	7252,40185773992	3,58231997489929	9,2	5,06508875	40,1	7,5
Sierra Leone	519,650015592018	6,98963022232056	33,099	0,76867908	14,3	32,2
Singapore	66859,3383447804	2,85597991943359	8,8	2,04291058	16,6	2,3
Slovak Republic	19486,3936845505	3,97256994247437	10,6	5,3113966	31,5	6,1
Slovenia	26123,7471277911	4,93587017059326	9,4	5,9950242	22,3	4,5
South Africa	7048,52221139611	5,64401006698608	21,137	5,07550335	20,4	22,7
Spain	30379,7211126422	4,18157005310059	7,9	6,31789398	28,1	4,1
Sri Lanka	4360,58363413213	2,13539004325867	14,805	1,7875824	22,3	17,9
Switzerland	85217,3691512274	4,86325979232788	10,3	3,55291438	25,7	2,8
Tanzania	1010,93762207031	3,69643998146057	37,525	1,56340396	9,2	31
Thailand	7124,56454354255	3,19860005378723	9,662	2,69087815	22,5	32,3
Timor-Leste	1239,36612167153	4,81987905502319	25,921	4,52639627	39,8	12,2
Uganda	793,128081123392	2,13052010536194	38,062	0,67621309	8,9	29,7
Ukraine	3096,56176757813	5,3199200630188	8,7	3,49084067	26,2	10
United States	62823,309438197	4,91233015060425	11,6	8,47234249	23,4	12,6
Uruguay	18825,2838068916	4,66279983520508	11,605	6,23086023	22	15,8
Uzbekistan	1597,0683366109	5,90000009536743	23,3	2,02401519	17,8	10,2
Vietnam	3267,22500852051	4,16744995117188	15,873	2,09576464	25	29,2
Zambia	1475,20453821416	4,73974990844727	36,04	1,96286535	14,6	20,5
Zimbabwe	2269,17701232332	3,86611008644104	32,074	2,78293462	12,1	40,6

Приложение 2

Данные после выброса аномальных данных по правилу 1,5IQR

Country Name	GDP per capita	expenditure on education		Birth rate, crude	general government health expenditure	Prevalence of current tobacco use	Mortality caused by road traffic injury
1	Afghanistan	502.0568	3.199790	36.927	0.5492201	23.8	14.2
2	Albania	5287.6637	3.154901	10.517	2.8249946	23.0	11.0
3	Algeria	4171.7950	5.866350	24.074	4.0581951	21.0	21.0
4	Argentina	11795.1594	4.877740	15.187	5.9358029	24.9	13.6
5	Azerbaijan	4739.8417	2.455430	14.000	0.9286820	24.2	8.1
6	Bahamas, The	31738.2672	2.249670	11.566	2.9617696	10.7	8.2
7	Bangladesh	1963.4119	1.947955	18.462	0.4258921	35.2	15.5
8	Barbados	18224.8906	5.428860	10.913	2.9384310	8.5	8.3
9	Belarus	6360.0625	5.381110	9.900	3.9000001	30.9	8.3
10	Benin	1194.4382	2.931620	38.186	0.4908835	7.2	26.9
11	Botswana	6947.8008	6.721310	25.328	4.6888843	19.8	25.7
12	Brazil	9121.0834	6.088510	14.133	3.8892388	13.2	17.1
13	Bulgaria	9446.7171	4.045300	8.900	4.2324943	39.4	8.9
14	Burundi	231.4465	5.078650	37.011	2.2890403	12.2	35.2
15	Cabo Verde	3442.7420	5.368850	17.821	3.2091837	11.7	26.3
16	Cambodia	1533.3160	2.578059	20.805	1.6625829	21.9	19.0
17	Cameroon	1594.0600	3.030960	36.716	0.2104447	7.5	29.9
18	Chile	15795.7085	5.433170	11.940	4.6545119	29.9	14.0
19	Colombia	6782.0379	4.449100	14.841	5.4608989	8.8	16.3
20	Congo, Rep.	2512.3840	2.995480	32.337	0.6840653	14.3	29.2
21	Costa Rica	12383.1500	6.769920	13.562	5.2816615	9.1	14.4
22	Cote d'Ivoire	2275.4960	3.239940	35.007	0.9065415	9.7	23.9
23	Croatia	15244.4310	3.907550	9.000	5.6451764	36.7	8.4
24	Cyprus	29418.9355	5.152854	10.843	2.8891170	35.5	5.5
25	Ecuador	6321.3494	4.621460	17.658	4.9299011	11.6	20.7
26	Eswatini	4020.2731	5.679160	25.718	3.5129719	9.3	33.0
27	Ethiopia	758.2991	5.068680	32.996	0.7708948	5.2	27.6
28	Fiji	6073.3364	4.835710	20.187	2.3198214	23.4	12.2
29	France	41557.8549	5.407170	11.300	8.4939108	33.6	5.2
30	Georgia	4722.0424	3.521060	14.395	2.8070135	31.7	11.8
31	Greece	19756.9905	3.597340	8.100	4.1116519	34.5	8.3
32	Guatemala	4485.7313	3.137050	23.929	2.2052352	11.0	20.4
33	Guyana	6094.9089	4.450580	20.770	2.9514914	12.6	22.5
34	Haiti	1494.2250	1.632070	24.579	0.5576716	7.9	18.6
35	India	1974.3778	4.363740	17.651	0.8878178	28.1	15.5
36	Indonesia	3902.6617	3.000000	17.179	1.3958540	37.2	11.6
37	Iran, Islamic Rep.	3873.9953	3.955360	17.231	3.8867421	14.0	21.7
38	Italy	34622.1697	4.256140	7.300	6.4059710	23.3	5.6
39	Jamaica	5594.4939	5.413810	12.220	3.8888052	9.7	14.3
40	Kazakhstan	9812.6264	2.615950	21.770	1.7142965	23.5	13.8
41	Kenya	1845.7829	5.107620	29.180	1.8561395	11.5	28.0
42	Kyrgyz Republic	1308.1402	5.552670	27.100	2.4247191	27.6	12.8

43	Latvia	17865.0311	4.243570	10.000	3.6970372	37.2	9.1
44	Madagascar	512.5440	2.844100	32.088	1.4805613	28.4	29.0
45	Malawi	537.9322	3.322960	33.925	2.4022078	11.3	33.1
46	Malaysia	11074.0641	4.478650	15.978	1.9238571	22.8	22.7
47	Mali	856.3565	3.905570	42.901	1.1511826	8.6	22.9
48	Mexico	9857.0288	4.254220	16.465	2.6718502	13.4	13.0
49	Moldova	4232.2069	5.439730	13.443	3.7236705	28.7	7.7
50	Morocco	3492.6733	5.348850	18.813	2.1431212	14.9	17.7
51	Niger	567.3308	3.457280	46.127	1.7616931	7.6	25.7
52	Oman	19887.5743	5.207230	19.816	3.5985525	7.9	13.6
53	Paraguay	6242.9615	3.283200	21.664	3.0178440	12.1	22.1
54	Peru	6912.1103	3.714300	18.092	3.1964009	8.7	13.7
55	Philippines	3194.6745	3.864053	22.445	1.5363278	23.4	11.9
56	Portugal	23562.5545	4.675160	8.500	5.7566347	25.3	9.0
57	Romania	12494.4775	3.344750	10.400	4.4312081	28.4	10.2
58	Russian Federation	11287.3604	4.678200	10.900	3.1840234	27.1	12.9
59	Rwanda	769.4373	3.073750	31.158	2.3451302	14.1	29.9
60	Sao Tome and Principe	1950.6294	5.235110	29.535	2.8008568	5.8	27.6
61	Senegal	1484.2397	4.858220	33.955	0.9602974	7.1	23.7
62	Serbia	7252.4019	3.582320	9.200	5.0650888	40.1	7.5
63	Sierra Leone	519.6500	6.989630	33.099	0.7686791	14.3	32.2
64	Slovak Republic	19486.3937	3.972570	10.600	5.3113966	31.5	6.1
65	Slovenia	26123.7471	4.935870	9.400	5.9950242	22.3	4.5
66	South Africa	7048.5222	5.644010	21.137	5.0755033	20.4	22.7
67	Spain	30379.7211	4.181570	7.900	6.3178940	28.1	4.1
68	Sri Lanka	4360.5836	2.135390	14.805	1.7875824	22.3	17.9
69	Tanzania	1010.9376	3.696440	37.525	1.5634040	9.2	31.0
70	Thailand	7124.5645	3.198600	9.662	2.6908782	22.5	32.3
71	Timor-Leste	1239.3661	4.819879	25.921	4.5263963	39.8	12.2
72	Uganda	793.1281	2.130520	38.062	0.6762131	8.9	29.7
73	Ukraine	3096.5618	5.319920	8.700	3.4908407	26.2	10.0
74	Uruguay	18825.2838	4.662800	11.605	6.2308602	22.0	15.8
75	Uzbekistan	1597.0683	5.900000	23.300	2.0240152	17.8	10.2
76	Vietnam	3267.2250	4.167450	15.873	2.0957646	25.0	29.2
77	Zambia	1475.2045	4.739750	36.040	1.9628653	14.6	20.5
78	Zimbabwe	2269.1770	3.866110	32.074	2.7829346	12.1	40.6

Приложение 3

Стандартизованные данные без выбросов

V1	V2	V3	V4	V5	V6
Afghanistan	-0.83894087	-0.8849749125	1.61216622	-1.39665434	0.4016989584	-0.414583075
Albania	-0.31210886	-0.9226903900	-0.99879071	-0.11307055	0.3212818425	-0.775266015
Algeria	-0.43495124	1.3554667131	0.34148733	0.58248004	0.1202390530	0.351868173
Argentina	0.40428034	0.5248371937	-0.53710314	1.64148965	0.5122724926	-0.482211126
Azerbaijan	-0.37241681	-1.5103855364	-0.65445285	-1.18263002	0.4419075163	-1.102134930
Bahamas, The	2.59975270	-1.6832649180	-0.89508402	-0.03592661	-0.9151313132	-1.090863588
Bangladesh	-0.67806501	-1.9367652216	-0.21332866	-1.46621388	1.5476428588	-0.268055631
Barbados	1.11210869	0.9878879927	-0.95964119	-0.04909004	-1.1362783817	-1.079592246
Belarus	-0.19405194	0.9477684859	-1.05978884	0.49325479	1.1154008613	-1.079592246
Benin	-0.76271884	-1.1102912215	1.73663403	-1.42955741	-1.2669561949	1.016877344
Botswana	-0.12934972	2.0738033600	0.46546083	0.93820175	-0.0003866207	0.881621241
Brazil	0.10989995	1.5421252437	-0.64130415	0.48718521	-0.6638278263	-0.087714161
Bulgaria	0.14574791	-0.1745781132	-1.15865127	0.68078841	1.9698327169	-1.011964195
Burundi	-0.86873149	0.6936416843	1.62047067	-0.41535980	-0.7643492211	1.952398719
Cabo Verde	-0.51521033	0.9374676776	-0.27669948	0.10362006	-0.8146099184	0.949249292
Cambodia	-0.72541288	-1.4073524138	0.01830603	-0.76869475	0.2107083083	0.126441335
Cameroon	-0.71872576	-1.0268256683	1.59130625	-1.58773071	-1.2367997765	1.355017600
Chile	0.84468787	0.9915089606	-0.85810947	0.91881505	1.0148794665	-0.437125759
Colombia	-0.14759802	0.1646943179	-0.57130954	1.37363395	-1.1061219633	-0.177884896
Congo, Rep.	-0.61763044	-1.0566359534	1.15838764	-1.32059886	-0.5532542920	1.276118207
Costa Rica	0.46901033	2.1146452175	-0.69775460	1.27254037	-1.0759655449	-0.392040391
Cote d'Ivoire	-0.64370868	-0.8512409773	1.42235034	-1.19511772	-1.0156527080	0.678737087
Croatia	0.78399952	-0.2903154726	-1.14876503	1.47757028	1.6984249510	-1.068320904
Cyprus	2.34442497	0.7559877962	-0.96656156	-0.07690417	1.5777992773	-1.395189818
Ecuador	-0.19831373	0.3095109980	-0.29281406	1.07414024	-0.8246620579	0.318054147
Eswatini	-0.45163180	1.1981897849	0.50401718	0.27496288	-1.0558612659	1.704429198
Ethiopia	-0.81073199	0.6852647406	1.22353799	-1.27162523	-1.4679989845	1.095776737
Fiji	-0.22561668	0.4895237425	-0.04279096	-0.39799867	0.3614904005	-0.640009913
France	3.68075941	0.9696637631	-0.92138143	3.08431533	1.3868086272	-1.429003844
Georgia	-0.37437628	-0.6150441366	-0.61540219	-0.12321223	1.1958179771	-0.685095280
Greece	1.28077262	-0.5509536461	-1.23774122	0.61263078	1.4772778825	-1.079592246
Guatemala	-0.40039101	-0.9376890815	0.32715228	-0.46262761	-0.8849748948	0.284240121
Guyana	-0.22324184	0.1659378988	0.01484584	-0.04172374	-0.7241406631	0.520938301
Haiti	-0.72971628	-2.2021719912	0.39141286	-1.39188753	-1.1965912186	0.081355968
India	-0.67685781	0.0929748751	-0.29350610	-1.20567826	0.8339409559	-0.268055631
Indonesia	-0.46457926	-1.0528383041	-0.34016917	-0.91913535	1.7486856484	-0.707637964
Iran, Islamic Rep.	-0.46773504	-0.2501456856	-0.33502832	0.48577701	-0.5834107105	0.430767566
Italy	2.91723222	0.0025696676	-1.31683117	1.90667418	0.3514382610	-1.383918476
Jamaica	-0.27833090	0.9752426498	-0.83042799	0.48694060	-1.0156527080	-0.403311733
Kazakhstan	0.18602968	-1.3755167023	0.11370828	-0.73952722	0.3715425399	-0.459668442
Kenya	-0.69101440	0.7179820564	0.84627893	-0.65952479	-0.8347141974	1.140862104
Kyrgyz Republic	-0.75020176	1.0919128952	0.64064506	-0.33883418	0.7836802585	-0.572381861
Latvia	1.07249292	-0.0079919561	-1.04990259	0.37877953	1.7486856484	-0.989421511
Madagascar	-0.83778637	-1.1838253549	1.13377089	-0.87135870	0.8640973743	1.253575523
Malawi	-0.83499146	-0.7814876665	1.31538119	-0.35153099	-0.8548184764	1.715700540
Malaysia	0.32489729	0.1895222666	-0.45890295	-0.62133070	0.3011775636	0.543480985
Mali	-0.79993716	-0.2919791226	2.20277042	-1.05713504	-1.1262262423	0.566023668
Mexico	0.19091780	0.0009562977	-0.41075694	-0.19944713	-0.6437235473	-0.549839178
Moldova	-0.42830069	0.9970209407	-0.70951923	0.39380124	0.8942537928	-1.147220297
Morocco	-0.50971355	0.9206633097	-0.17862794	-0.49766119	-0.4929414552	-0.020086109
Niger	-0.83175507	-0.6686320975	2.52170063	-0.71279453	-1.2267476370	0.881621241
Oman	1.29514818	0.8016745712	-0.07946892	0.32323210	-1.1965912186	-0.482211126
Paraguay	-0.20694321	-0.8148938796	0.10322886	-0.00429950	-0.7744013605	0.475852933
Peru	-0.13327877	-0.4526841115	-0.24990776	0.09641029	-1.1161741028	-0.470939784
Philippines	-0.54251928	-0.3268614871	0.18044042	-0.83990520	0.3614904005	-0.673823938
Portugal	1.69971489	0.3546296681	-1.19819625	1.54043507	0.5524810505	-1.000692853
Romania	0.48126601	-0.7631797035	-1.01035762	0.79286709	0.8640973743	-0.865436750
Russian Federation	0.34837838	0.3571837365	-0.96092640	0.08942913	0.7334195611	-0.561110519

Rwanda	-0.80950581	-0.9908735998	1.04182883	-0.38372395	-0.5733585710	1.355017600
Sao Tome and Principe	-0.67947219	0.8250990843	0.88137509	-0.12668472	-1.4076861476	1.095776737
Senegal	-0.73081553	0.5084366691	1.31834706	-1.16479826	-1.2770083344	0.656194403
Serbia	-0.09581717	-0.5635735485	-1.12899254	1.15038886	2.0401976932	-1.169762981
Sierra Leone	-0.83700409	2.2992456697	1.23372082	-1.27287496	-0.5532542920	1.614258463
Slovak Republic	1.25098350	-0.2356858524	-0.99058513	1.28931157	1.1757136981	-1.327561767
Slovenia	1.98166831	0.5736781931	-1.10922006	1.67489166	0.2509168662	-1.507903237
South Africa	-0.11826163	1.1686567413	0.05112836	1.15626291	0.0599262161	0.543480985
Spain	2.45019475	-0.0600841344	-1.25751371	1.85699689	0.8339409559	-1.552988605
Sri Lanka	-0.41416811	-1.7792828661	-0.57486859	-0.69819244	0.2509168662	0.002456575
Tanzania	-0.78291982	-0.4676900145	1.67128596	-0.82463371	-1.0659134054	1.479002361
Thailand	-0.10989037	-0.8859747051	-1.08331810	-0.18871497	0.2710211452	1.625529805
Timor-Leste	-0.75777287	0.4762225548	0.52408625	0.84655517	2.0100412748	-0.640009913
Uganda	-0.80689777	-1.7833745839	1.72437509	-1.32502769	-1.0960698238	1.332474916
Ukraine	-0.55332019	0.8963565913	-1.17842376	0.26248042	0.6429503058	-0.887979434
Uruguay	1.17820404	0.3442447258	-0.89122838	1.80790806	0.2207604478	-0.234241605
Uzbekistan	-0.71839459	1.3837393489	0.26496781	-0.56483946	-0.2014294103	-0.865436750
Vietnam	-0.53453243	-0.0719478322	-0.46928351	-0.52437129	0.5223246321	1.276118207
Zambia	-0.73181018	0.4088981129	1.52447524	-0.59932923	-0.5230978736	0.295511463
Zimbabwe	-0.64440431	-0.3251333343	1.13238682	-0.13679322	-0.7744013605	2.561051181

Приложение 4

Нормализованные данные без выбросов

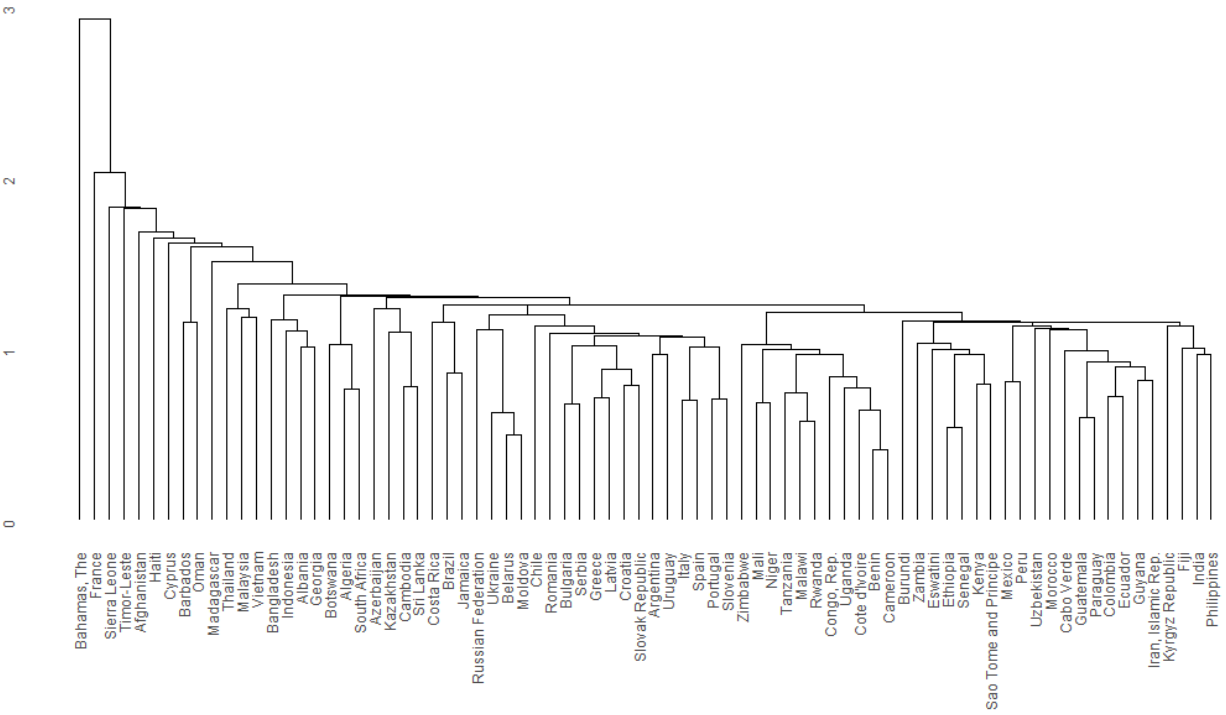
Afghanistan	0.006548120	0.29261828	0.76305148	0.04089779	0.53295129	0.27671233
Albania	0.122348334	0.28423970	0.08285471	0.31563477	0.51002865	0.18904110
Algeria	0.095346987	0.79033739	0.43201896	0.46450971	0.45272206	0.46301370
Argentina	0.279814128	0.60581119	0.20313184	0.69117905	0.56446991	0.26027397
Azerbaijan	0.109092355	0.15368191	0.17256033	0.08670734	0.54441261	0.10958904
Bahamas, The	0.762389521	0.11527637	0.10987200	0.33214658	0.15759312	0.11232877
Bangladesh	0.041909410	0.05896071	0.28748036	0.02600934	0.85959885	0.31232877
Barbados	0.435398207	0.70867896	0.09305380	0.32932909	0.09455587	0.11506849
Belarus	0.148297813	0.69976632	0.06696371	0.44541203	0.73638968	0.11506849
Benin	0.023302091	0.24256375	0.79547737	0.03385525	0.05730659	0.62465753
Botswana	0.162519672	0.94991749	0.46431607	0.54064802	0.41833811	0.59178082
Brazil	0.215107900	0.83180400	0.17598578	0.44411290	0.22922636	0.35616438
Bulgaria	0.222987455	0.45043451	0.04120844	0.48555153	0.97994269	0.13150685
Burundi	0.000000000	0.64331148	0.76521493	0.25093308	0.20057307	0.85205479
Cabo Verde	0.077705652	0.69747797	0.27097123	0.36201500	0.18624642	0.60821918
Cambodia	0.031502121	0.17657095	0.34782497	0.17530563	0.47851003	0.40821918
Cameroon	0.032971980	0.26110581	0.75761712	0.00000000	0.06590258	0.70684932
Chile	0.376617824	0.70948337	0.11950447	0.53649851	0.70773639	0.27123288
Colombia	0.158508607	0.52580464	0.19422052	0.63384749	0.10315186	0.33424658
Congo, Rep.	0.055193219	0.25448339	0.64483478	0.05717663	0.26074499	0.68767123
Costa Rica	0.294042090	0.95899060	0.16127952	0.61220952	0.11174785	0.28219178
Cote d'Ivoire	0.049461097	0.30011235	0.71360136	0.08403448	0.12893983	0.54246575
Croatia	0.363278232	0.42472320	0.04378396	0.65609392	0.90257880	0.11780822
Cyprus	0.706267257	0.65716181	0.09125093	0.32337578	0.86819484	0.03835616
Ecuador	0.147361050	0.55797599	0.26677312	0.56974416	0.18338109	0.45479452
Eswatini	0.091680519	0.75539797	0.47436063	0.39868905	0.11747851	0.79178082
Ethiopia	0.012748569	0.64145053	0.66180751	0.06765890	0.00000000	0.64383562
Fiji	0.141359730	0.59796623	0.33190821	0.25464904	0.52148997	0.22191781
France	1.000000000	0.70463041	0.10302109	1.00000000	0.81375358	0.03013699
Georgia	0.108661655	0.35258400	0.18273366	0.31346405	0.75931232	0.21095890
Greece	0.472471350	0.36682185	0.02060422	0.47096314	0.83954155	0.11506849
Guatemala	0.102943492	0.28090771	0.42828444	0.24081593	0.16618911	0.44657534
Guyana	0.141881732	0.52608091	0.34692353	0.33090577	0.21203438	0.50410959

Haiti	0.030556212	0.00000000	0.44502537	0.04191807	0.07736390	0.39726027
India	0.042174759	0.50987201	0.26659283	0.08177412	0.65616046	0.31232877
Indonesia	0.088834606	0.25532705	0.25443635	0.14310547	0.91690544	0.20547945
Iran, Islamic Rep.	0.088140949	0.43364701	0.25577562	0.44381149	0.25214900	0.48219178
Italy	0.832173047	0.48978829	0.00000000	0.74793888	0.51862464	0.04109589
Jamaica	0.129772890	0.70586977	0.12671595	0.44406055	0.12893983	0.27945205
Kazakhstan	0.231841582	0.18364332	0.37267881	0.18154861	0.52435530	0.26575342
Kenya	0.039063072	0.64871875	0.56352538	0.19867225	0.18051576	0.65479452
Kyrgyz Republic	0.026053406	0.73178832	0.50995441	0.26731255	0.64183381	0.23835616
Latvia	0.426690470	0.48744200	0.06953924	0.42090986	0.91690544	0.13698630
Madagascar	0.006801886	0.22622798	0.63842172	0.15333154	0.66475645	0.68219178
Malawi	0.007416220	0.31560820	0.68573415	0.26459494	0.17478510	0.79452055
Malaysia	0.262365350	0.53132023	0.22350426	0.20684728	0.50429799	0.50958904
Mali	0.015121325	0.42435362	0.91691349	0.11356816	0.09742120	0.51506849
Mexico	0.232916015	0.48942988	0.23604708	0.29714681	0.23495702	0.24383562
Moldova	0.096808810	0.71070786	0.15821464	0.42412509	0.67335244	0.09863014
Morocco	0.078913871	0.69374485	0.29652046	0.23331738	0.27793696	0.37260274
Niger	0.008127595	0.34067932	1.00000000	0.18727045	0.06876791	0.59178082
Oman	0.475631167	0.66731123	0.32235300	0.40902054	0.07736390	0.26027397
Paraguay	0.145464249	0.30818693	0.36994875	0.33891601	0.19770774	0.49315068
Peru	0.161656047	0.38865264	0.27795091	0.36047183	0.10028653	0.26301370
Philippines	0.071703014	0.41660442	0.39006362	0.16006381	0.52148997	0.21369863
Portugal	0.564556877	0.56799921	0.03090633	0.66954943	0.57593123	0.13424658
Romania	0.296735949	0.31967535	0.07984135	0.50954074	0.66475645	0.16712329
Russian Federation	0.267526608	0.56856660	0.09271898	0.35897759	0.62750716	0.24109589
Rwanda	0.013018088	0.26909265	0.61446931	0.25770439	0.25501433	0.70684932
Sao Tome and Principe	0.041600104	0.67251504	0.57266850	0.31272080	0.01719198	0.64383562
Senegal	0.030314592	0.60216778	0.68650681	0.09052403	0.05444126	0.53698630
Serbia	0.169890287	0.36401831	0.04893502	0.58606434	1.00000000	0.09315068
Sierra Leone	0.006973835	1.00000000	0.66446030	0.06739140	0.26074499	0.76986301
Slovak Republic	0.465923557	0.43685929	0.08499240	0.61579921	0.75358166	0.05479452
Slovenia	0.626531597	0.61666133	0.05408607	0.69832839	0.48997135	0.01095890
South Africa	0.164956888	0.74883714	0.35637572	0.58732161	0.43553009	0.50958904
Spain	0.729515964	0.47586961	0.01545316	0.73730601	0.65616046	0.00000000
Sri Lanka	0.099915219	0.09394576	0.19329333	0.19039587	0.48997135	0.37808219
Tanzania	0.018861817	0.38531905	0.77845314	0.16333251	0.11461318	0.73698630

Thailand	0.166796931	0.29239617	0.06083396	0.29944391	0.49570201	0.77260274
Timor-Leste	0.024389239	0.59501134	0.47958895	0.52103208	0.99140401	0.22191781
Uganda	0.013591348	0.09303678	0.79228372	0.05622869	0.10601719	0.70136986
Ukraine	0.069328921	0.68834505	0.03605738	0.39601731	0.60171920	0.16164384
Uruguay	0.449926283	0.56569217	0.11087645	0.72679908	0.48137536	0.32054795
Uzbekistan	0.033044775	0.79661822	0.41208437	0.21893861	0.36103152	0.16712329
Vietnam	0.073458562	0.47323406	0.22079996	0.22760037	0.56733524	0.68767123
Zambia	0.030095963	0.58005506	0.74020656	0.21155645	0.26934097	0.44931507
Zimbabwe	0.049308193	0.41698834	0.63806114	0.31055719	0.19770774	1.00000000

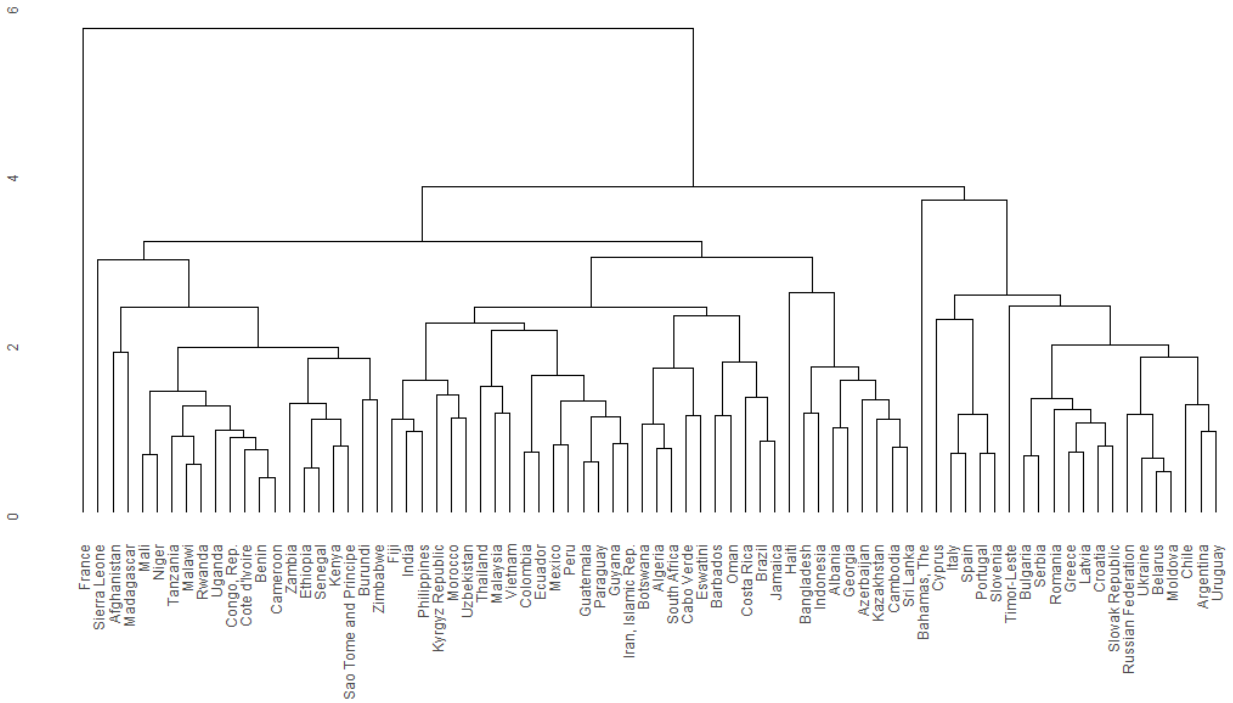
Приложение 5

Дендрограмма по методу ближайшего соседа



Приложение 6

Дендрограмма по методу среднего



Приложение 7

Код программы на R

```
                #uploading data
library(readxl)
data <- read_excel("мр21/dataa.xlsx",sheet = "Data",range = "A1:G95")
#View(data)

#steamplots

# install.packages("aplpack")
#install.packages("aplpack")
#install.packages("tcltk")
library(aplpack)
library(tcltk)


#1
#picture
plot.new()
out <- capture.output(stem(data$`GDP per capita`,scale = 1))
text(0, 1, paste(out, collapse = "\n"), adj = c(0, 1))


#2

#View(edu)

plot.new()
out <- capture.output(stem(data$`expenditure on education`, scale = 0.5))
text(0, 1, paste(out, collapse = "\n"), adj = c(0, 1))


#3
#View(birth)

plot.new()
out <- capture.output(stem(data$`Birth rate, crude`,scale = 0.5))
text(0, 1, paste(out, collapse = "\n"), adj = c(0, 1))

#4
#View(health)
plot.new()
out <- capture.output(stem(data$`general government health expenditure`, scale = 0.5))
text(0, 1, paste(out, collapse = "\n"), adj = c(0, 1))


#5
#View(tobacco)
plot.new()
out <- capture.output(stem(data$`Prevalence of current tobacco use`,scale = 0.5))
text(0, 1, paste(out, collapse = "\n"), adj = c(0, 1))

#6
plot.new()
out <- capture.output(stem(data$`Mortality caused by road traffic injury`))
text(0, 1, paste(out, collapse = "\n"), adj = c(0, 1))


#do matrix of our data
dat =  matrix( ,94,6)
rownames(dat) <- data$`Country Name`
dat[,1]<-data$`GDP per capita`
dat[,2]<-data$`expenditure on education`
dat[,3]<-data$`Birth rate, crude`
dat[,4]<-data$`general government health expenditure`
dat[,5]<-data$`Prevalence of current tobacco use`
dat[,6]<-data$`Mortality caused by road traffic injury`

#dotplots

#dotchart(data$`GDP per capita`, labels = data$name, pch = 21, bg = "green", pt.cex = 1.5)
dotchart(dat[,1],  pch = 21, bg = "green",cex = 0.4, pt.cex = 1.0, main ="Dotplot for GDP per capita")
dotchart(dat[,2], pch = 21, bg = "green",cex = 0.4, pt.cex = 1.0, main ="Dotplot for expenditure on education")
dotchart(dat[,3], pch = 21,cex = 0.4, bg = "green", pt.cex = 1.0, main ="Dotplot for birth rate, crude")
dotchart(dat[,4], cex = 0.4, pch = 21, bg = "green", pt.cex = 1.0, main ="Dotplot for general government health expenditure")
dotchart(dat[,5], cex = 0.4, pch = 21, bg = "green", pt.cex = 1.0, main ="Dotplot for Prevalence of current tobacco use")
```

```
dotchart(dat[,6],cex = 0.4,pch = 21,bg ="green", pt.cex = 1.0,main="Dotplot for Mortality caused by road traffic injury ")

#boxplots

boxplot(dat[,1], horizontal = TRUE, main ="Boxplot for GDP per capita")
boxplot(dat[,2], horizontal = TRUE, main ="Boxplot for expenditure on education" )
boxplot(dat[,3],horizontal = TRUE, main ="Boxplot for birth rate, crude")
boxplot(dat[,4], horizontal = TRUE, main ="Boxplot for general government health expenditure")
boxplot(dat[,5], horizontal = TRUE, main ="Boxplot for Prevalence of current tobacco use")
boxplot(dat[,6],horizontal = TRUE,main="Boxplot for Mortality caused by road traffic injury ")


#mean, median, mode
ch_rv = matrix(,14,6)
rownames(ch_rv)<-data.frame(c("mean"),c("mode"),c("median"),c("variance"),c("var"),c("sd"), c("koef_variance"),c("iqr"),
                            c("upper_3iqr"),c("lower_3iqr"),c("upper_1.5iqr"),c("lower_1.5iqr"),
                            c("upper_3s"),c("lower_3s"))

ch_rv[1,1] <-mean(dat[,1])
ch_rv[2,1] <-mode(dat[,1])
ch_rv[3,1]<- median(dat[,1])

ch_rv[1,2] <-mean(dat[,2])
ch_rv[2,2] <-mode(dat[,2])
ch_rv[3,2]<- median(dat[,2])

ch_rv[1,3] <-mean(dat[,3])
ch_rv[2,3] <-mode(dat[,3])
ch_rv[3,3] <-median(dat[,3])

ch_rv[1,4] <-mean(dat[,4])
ch_rv[2,4] <-mean(dat[,4])
ch_rv[3,4] <-median(dat[,4])

ch_rv[1,5] <-mean(dat[,5])
ch_rv[2,5] <-mode(dat[,5])
ch_rv[3,5] <-median(dat[,5])

ch_rv[1,6] <-mean(dat[,6])
ch_rv[2,6] <-mode(dat[,6])
ch_rv[3,6] <-median(dat[,6])


#range of variation and coefficient of variation
#variance and mean square deviation
ch_rv[4,1] <- max(dat[,1])-min(dat[,1])
ch_rv[4,2] <- max(dat[,2])-min(dat[,2])
ch_rv[4,3] <- max(dat[,3])-min(dat[,3])
ch_rv[4,4] <- max(dat[,4])-min(dat[,4])
ch_rv[4,5] <- max(dat[,5])-min(dat[,5])
ch_rv[4,6] <- max(dat[,6])-min(dat[,6])


ch_rv[5,1] <-var(dat[,1])
ch_rv[5,2] <-var(dat[,2])
ch_rv[5,3] <-var(dat[,3])
ch_rv[5,4] <-var(dat[,4])
ch_rv[5,5] <-var(dat[,5])
ch_rv[5,6] <-var(dat[,6])

ch_rv[6,1] <-sd(dat[,1])
ch_rv[6,2] <-sd(dat[,2])
ch_rv[6,3] <-sd(dat[,3])
ch_rv[6,4] <-sd(dat[,4])
ch_rv[6,5] <-sd(dat[,5])
ch_rv[6,6] <-sd(dat[,6])

ch_rv[7,1] <-as.numeric(ch_rv[6,1])/as.numeric(ch_rv[1,1])
ch_rv[7,2] <-as.numeric(ch_rv[6,2])/as.numeric(ch_rv[1,2])
ch_rv[7,3] <-as.numeric(ch_rv[6,3])/as.numeric(ch_rv[1,3])
ch_rv[7,4] <-as.numeric(ch_rv[6,4])/as.numeric(ch_rv[1,4])
ch_rv[7,5] <-as.numeric(ch_rv[6,5])/as.numeric(ch_rv[1,5])
ch_rv[7,6] <-as.numeric(ch_rv[6,6])/as.numeric(ch_rv[1,6])

#decile

dec_gdp <-quantile(dat[,1],probs = seq (.1, .9, by = .1 ))
dec_edu <- quantile(dat[,2],probs = seq (.1, .9, by = .1 ) )
dec_birth <- quantile(dat[,3],probs = seq (.1, .9, by = .1 ) )
```

```
dec_health <- quantile(dat[,4],probs = seq (.1, .9, by = .1 ) )
dec_tobacco <-quantile(dat[,5],probs = seq (.1, .9, by = .1 ) )
dec_road <-quantile(dat[,6], probs = seq (.1, .9, by = .1 ))

#quantile

quan_gdp <-quantile(dat[,1], probs = c(0,0.25,0.5,0.75,1))
quan_edu <- quantile(dat[,2],probs = c(0,0.25,0.5,0.75,1) )
quan_birth <- quantile(dat[,3],probs = c(0,0.25,0.5,0.75,1))
quan_health <- quantile(dat[,4],probs = c(0,0.25,0.5,0.75,1) )
quan_tobacco <-quantile(dat[,5],probs = c(0,0.25,0.5,0.75,1))
quan_road <-quantile(dat[,6], probs = c(0,0.25,0.5,0.75,1))

#IQR
ch_rv[8,1]<-as.numeric(quan_gdp[c(4)])-as.numeric(quan_gdp[c(2)])
ch_rv[8,2]<-as.numeric(quan_edu[4])-as.numeric(quan_edu[2])
ch_rv[8,3]<-as.numeric(quan_birth[4])-as.numeric(quan_birth[2])
ch_rv[8,4]<-as.numeric(quan_health[4])-as.numeric(quan_health[2])
ch_rv[8,5]<-as.numeric(quan_tobacco[4])-as.numeric(quan_tobacco[2])
ch_rv[8,6]<-as.numeric(quan_road[4])-as.numeric(quan_road[2])

#counting the lower and upper bounds at 3IQR

ch_rv[9,1]<-as.numeric(quan_gdp[c(4)])+3*as.numeric(ch_rv[8,1])
ch_rv[9,2]<-as.numeric(quan_edu[c(4)])+3*as.numeric(ch_rv[8,2])
ch_rv[9,3]<-as.numeric(quan_birth[c(4)])+3*as.numeric(ch_rv[8,3])
ch_rv[9,4]<-as.numeric(quan_health[c(4)])+3*as.numeric(ch_rv[8,4])
ch_rv[9,5]<-as.numeric(quan_tobacco[c(4)])+3*as.numeric(ch_rv[8,5])
ch_rv[9,6]<-as.numeric(quan_road[c(4)])+3*as.numeric(ch_rv[8,6])

ch_rv[10,1]<-as.numeric(quan_gdp[c(2)])-3*as.numeric(ch_rv[8,1])
ch_rv[10,2]<-as.numeric(quan_edu[c(2)])-3*as.numeric(ch_rv[8,2])
ch_rv[10,3]<-as.numeric(quan_birth[c(2)])-3*as.numeric(ch_rv[8,3])
ch_rv[10,4]<-as.numeric(quan_health[c(2)])-3*as.numeric(ch_rv[8,4])
ch_rv[10,5]<-as.numeric(quan_tobacco[c(2)])-3*as.numeric(ch_rv[8,5])
ch_rv[10,6]<-as.numeric(quan_road[c(4)])-3*as.numeric(ch_rv[8,6])

#counting the lower and upper bounds at 1.5IQR

ch_rv[11,1]<-as.numeric(quan_gdp[c(4)])+1.5*as.numeric(ch_rv[8,1])
ch_rv[11,2]<-as.numeric(quan_edu[c(4)])+1.5*as.numeric(ch_rv[8,2])
ch_rv[11,3]<-as.numeric(quan_birth[c(4)])+1.5*as.numeric(ch_rv[8,3])
ch_rv[11,4]<-as.numeric(quan_health[c(4)])+1.5*as.numeric(ch_rv[8,4])
ch_rv[11,5]<-as.numeric(quan_tobacco[c(4)])+1.5*as.numeric(ch_rv[8,5])
ch_rv[11,6]<-as.numeric(quan_road[c(4)])+1.5*as.numeric(ch_rv[8,6])

ch_rv[12,1]<-as.numeric(quan_gdp[c(2)])-1.5*as.numeric(ch_rv[8,1])
ch_rv[12,2]<-as.numeric(quan_edu[c(2)])-1.5*as.numeric(ch_rv[8,2])
ch_rv[12,3]<-as.numeric(quan_birth[c(2)])-1.5*as.numeric(ch_rv[8,3])
ch_rv[12,4]<-as.numeric(quan_health[c(2)])-1.5*as.numeric(ch_rv[8,4])
ch_rv[12,5]<-as.numeric(quan_tobacco[c(2)])-1.5*as.numeric(ch_rv[8,5])
ch_rv[12,6]<-as.numeric(quan_road[c(4)])-1.5*as.numeric(ch_rv[8,6])

#counting by the 3sigma rule

ch_rv[13,1]<-as.numeric(ch_rv[1,1])+3*as.numeric(ch_rv[6,1])
ch_rv[13,2]<-as.numeric(ch_rv[1,2])+3*as.numeric(ch_rv[6,2])
ch_rv[13,3]<-as.numeric(ch_rv[1,3])+3*as.numeric(ch_rv[6,3])
ch_rv[13,4]<-as.numeric(ch_rv[1,4])+3*as.numeric(ch_rv[6,4])
ch_rv[13,5]<-as.numeric(ch_rv[1,5])+3*as.numeric(ch_rv[6,5])
ch_rv[13,6]<-as.numeric(ch_rv[1,6])+3*as.numeric(ch_rv[6,6])

ch_rv[14,1]<-as.numeric(ch_rv[1,1])-3*as.numeric(ch_rv[6,1])
ch_rv[14,2]<-as.numeric(ch_rv[1,2])-3*as.numeric(ch_rv[6,2])
ch_rv[14,3]<-as.numeric(ch_rv[1,3])-3*as.numeric(ch_rv[6,3])
ch_rv[14,4]<-as.numeric(ch_rv[1,4])-3*as.numeric(ch_rv[6,4])
ch_rv[14,5]<-as.numeric(ch_rv[1,5])-3*as.numeric(ch_rv[6,5])
ch_rv[14,6]<-as.numeric(ch_rv[1,6])-3*as.numeric(ch_rv[6,6])


#all lower bounds should be 0

#z-transformation
z_dat =  matrix( ,94,6)
rownames(z_dat) <- data$`Country Name`
z_dat[,1]<-scale(data$`GDP per capita`)
z_dat[,2]<-scale(data$`expenditure on education`)
z_dat[,3]<-scale(data$`Birth rate, crude`)
z_dat[,4]<-scale(data$`general government health expenditure`)
```

```
z_dat[,5]<-scale(data$`Prevalence of current tobacco use`)
z_dat[,6]<-scale(data$`Mortality caused by road traffic injury`)



#corelation with abnormal data
corel<-cor(z_dat)
pairs(z_dat)


#plots of correlation with abnormal data
with(data, plot(data$`GDP per capita`, data$`Birth rate, crude`, pch = 21, bg = "green",main="correlation cloud between 1 and 3 parameter "))
with(data, plot(data$`GDP per capita`, data$`general government health expenditure`, pch = 21, bg = "green",main="correlation cloud between 1 and 4 parameter "))
with(data, plot(data$`GDP per capita`, data$`Prevalence of current tobacco use`, pch = 21, bg = "green",main="correlation cloud between 1 and 5 parameter "))
with(data, plot(data$`GDP per capita`, data$`Mortality caused by road traffic injury`, pch = 21, bg = "green",main="correlation cloud between 1 and 6 parameter "))

with(data, plot(data$`Birth rate, crude`, data$`general government health expenditure`, pch = 21, bg = "green",main="correlation cloud between 3 and 4 parameter "))
with(data, plot(data$`Birth rate, crude`, data$`Prevalence of current tobacco use`, pch = 21, bg = "green",main="correlation cloud between 3 and 5 parameter "))
with(data, plot(data$`Birth rate, crude`, data$`Mortality caused by road traffic injury`, pch = 21, bg = "green",main="correlation
                cloud between 3 and 6 parameter "))

with(data, plot(data$`general government health expenditure`, data$`Prevalence of current tobacco use`, pch = 21, bg = "green",main="correlation cloud between 4 and 5 parameter "))
with(data, plot(data$`general government health expenditure`,data$`Mortality caused by road traffic injury`, pch = 21, bg = "green",main="correlation cloud between 4 and 6 parameter "))

with(data, plot(data$`Prevalence of current tobacco use`, data$`Mortality caused by road traffic injury`, pch = 21, bg = "green",main="correlation cloud between 5  and 6 parameter "))


#using 1.5IQR to delete abnormal data
no_outliers_15 <- subset(data, data$`GDP per capita` > (as.numeric(ch_rv[12,1])) & data$`GDP per capita` < as.numeric(ch_rv[11,1]))
no_outliers_15 <- subset(no_outliers_15, no_outliers_15$`expenditure on education` > (as.numeric(ch_rv[12,2])) & no_outliers_15$`expenditure on education` < as.numeric(ch_rv[11,2]))
no_outliers_15 <-  subset(no_outliers_15, no_outliers_15$`Birth rate, crude` > (as.numeric(ch_rv[12,3])) & no_outliers_15$`Birth rate, crude` < as.numeric(ch_rv[11,3]))
no_outliers_15 <-  subset(no_outliers_15, no_outliers_15$`general government health expenditure` > (as.numeric(ch_rv[12,4])) & no_outliers_15$`general government health expenditure` < as.numeric(ch_rv[11,4]))
no_outliers_15 <-  subset(no_outliers_15, no_outliers_15$`Prevalence of current tobacco use` > (as.numeric(ch_rv[12,5])) & no_outliers_15$`Prevalence of current tobacco use` < as.numeric(ch_rv[11,5]))
no_outliers_15 <- subset(no_outliers_15, no_outliers_15$`Mortality caused by road traffic injury` > (as.numeric(ch_rv[12,6])) & no_outliers_15$`Mortality caused by road traffic injury` < as.numeric(ch_rv[11,6]))


noab_z_dat =  matrix( ,78,6)
rownames(noab_z_dat) <- no_outliers_15$`Country Name`
noab_z_dat[,1]<-scale(no_outliers_15$`GDP per capita`)
noab_z_dat[,2]<-scale(no_outliers_15$`expenditure on education`)
noab_z_dat[,3]<-scale(no_outliers_15$`Birth rate, crude`)
noab_z_dat[,4]<-scale(no_outliers_15$`general government health expenditure`)
noab_z_dat[,5]<-scale(no_outliers_15$`Prevalence of current tobacco use`)
noab_z_dat[,6]<-scale(no_outliers_15$`Mortality caused by road traffic injury`)
pairs(noab_z_dat)


#plots without abnormal data(1.5IQR)

#with(no_outliers_15, plot(no_outliers_15$`GDP per capita`, no_outliers_15$`Birth rate, crude`, pch = 21, bg = "green",main="correlation cloud between 1 and 3 parameter after 1.5IQR "))
#with(no_outliers_15, plot(no_outliers_15$`GDP per capita`, no_outliers_15$`general government health expenditure`, pch = 21, bg = "green",main="correlation cloud between 1 and 4 parameter after 1.5IQR"))
#with(no_outliers_15, plot(no_outliers_15$`GDP per capita`, no_outliers_15$`Prevalence of current tobacco use`, pch = 21, bg = "green",main="correlation cloud between 1 and 5 parameter after 1.5IQR"))
#with(no_outliers_15, plot(no_outliers_15$`GDP per capita`, no_outliers_15$`Mortality caused by road traffic injury`, pch = 21, bg = "green",main="correlation cloud between 1 and 6 parameter after 1.5IQR"))
#
#with(no_outliers_15, plot(no_outliers_15$`Birth rate, crude`, no_outliers_15$`general government health expenditure`, pch = 21, bg = "green",main="correlation cloud between 3 and 4 parameter after 1.5IQR"))
#with(no_outliers_15, plot(no_outliers_15$`Birth rate, crude`, no_outliers_15$`Prevalence of current tobacco use`, pch = 21, bg = "green",main="correlation cloud between 3 and 5 parameter after 1.5IQR"))
#with(no_outliers_15, plot(no_outliers_15$`Birth rate, crude`, no_outliers_15$`Mortality caused by road traffic injury`, pch = 21, bg = "green",main="correlation cloud between 3 and 6 parameter after 1.5IQR"))
```

```
#with(no_outliers_15, plot(no_outliers_15$`general government health expenditure`, no_outliers_15$`Prevalence of current tobacco use`, pch = 21, bg = "green",main="correlation cloud between 4 and 5 parameter after 1.5IQR "))
#with(no_outliers_15, plot(no_outliers_15$`general government health expenditure`,no_outliers_15$`Mortality caused by road traffic injury`, pch = 21, bg = "green",main="correlation cloud between 4 and 6 parameter after 1.5IQR"))

#with(no_outliers_15, plot(no_outliers_15$`Prevalence of current tobacco use`, no_outliers_15$`Mortality caused by road traffic injury`, pch = 21, bg = "green",main="correlation cloud between 5  and 6 parameter after 1.5IQR"))


#correlation after 1.5IQR
corel_no<- cor(noab_z_dat)


install.packages("ppcor")
library(ppcor)

#partial correlation

#without abnormal data
p_cor<-pcor(noab_z_dat)

View(p_cor[["estimate"]])



#multiply correlation for data without abnormal data
det <- det(corel_no)
det_1 <- (-1)^(1+1)*det(corel_no[2:5,2:5])


R_1 <-sqrt(1-det/det_1)


#View(corel_no_outliers_15[c(1,3,4,5),c(1,3,4,5)])


#-----------------------------
#cluster analysis
library("ape")
library("ggplot2")
library("ggdendro")


dd <- dist(noab_z_dat, method = "euclidean")
hc <- hclust(dd, method = "ward.D2")
ggdendrogram(hc, hang = -1, cex = 0.5, main = "ward.D2")


#method of nearest neighbor
hc1 <- hclust(dd, method = "single")
ggdendrogram(hc1, hang = -1, cex = 0.5)

#the far neighbor method

hc_far<-hclust(dd, method = "complete")
ggdendrogram(hc_far, hang = -1, cex = 0.6)

#average
hc_avr<-hclust(dd, method = "average")
ggdendrogram(hc_avr, hang = -1, cex = 0.6)

install.packages("ClusterR")
install.packages("cluster")
install.packages("factoextra")
?hclust

library(ClusterR)
library(cluster)
library(factoextra)
set.seed(1)


km <- kmeans(noab_z_dat, centers = 3, nstart = 100)
y_kmeans <- km$cluster
clusplot(noab_z_dat[, c(3,5)],
         y_kmeans,
         lines = 0,
         shade = TRUE,
         color = TRUE,
```

```
    labels = 2,
    plotchar = FALSE,
    span = TRUE,
    cex=0.5,
    main = paste("example of cluster"),
    xlab = 'V3',
    ylab = 'V5')



plot(noab_z_dat[,1],noab_z_dat[,2],main = " centers on 1 and 2 param")
points(km$centers[, c(1, 2)],col = 1:3, pch = 8, cex = 5)

plot(noab_z_dat[,1],noab_z_dat[,3],main = " centers on 1 and 3 param")
points(km$centers[, c(1, 3)],col = 1:3, pch = 8, cex = 5)

plot(noab_z_dat[,1],noab_z_dat[,4],main = " centers on 1 and 4 param")
points(km$centers[, c(1, 4)],col = 1:3, pch = 8, cex = 5)

plot(noab_z_dat[,1],noab_z_dat[,5],main = " centers on 1 and 5 param")
points(km$centers[, c(1, 5)],col = 1:3, pch = 8, cex = 5)

plot(noab_z_dat[,1],noab_z_dat[,6],main = " centers on 1 and 6 param")
points(km$centers[, c(1, 6)],col = 1:3, pch = 8, cex = 5)

plot(noab_z_dat[,2],noab_z_dat[,3],main = " centers on 2 and 3 param")
points(km$centers[, c(2, 3)],col = 1:3, pch = 8, cex = 5)

plot(noab_z_dat[,1],noab_z_dat[,4],main = " centers on 2 and 4 param")
points(km$centers[, c(2, 4)],col = 1:3, pch = 8, cex = 5)

plot(noab_z_dat[,1],noab_z_dat[,5],main = " centers on 2 and 5 param")
points(km$centers[, c(2, 5)],col = 1:3, pch = 8, cex = 5)

plot(noab_z_dat[,2],noab_z_dat[,6],main = " centers on 2 and 6 param")
points(km$centers[, c(2, 6)],col = 1:3, pch = 8, cex = 5)

plot(noab_z_dat[,3],noab_z_dat[,4],main = " centers on 3 and 4 param")
points(km$centers[, c(3, 4)],col = 1:3, pch = 8, cex = 5)

plot(noab_z_dat[,3],noab_z_dat[,5],main = " centers on 3 and 5 param")
points(km$centers[, c(3, 5)],col = 1:3, pch = 8, cex = 5)

plot(noab_z_dat[,3],noab_z_dat[,6],main = " centers on 3 and 6 param")
points(km$centers[, c(3, 6)],col = 1:3, pch = 8, cex = 5)

plot(noab_z_dat[,4],noab_z_dat[,5],main = " centers on 4 and 5 param")
points(km$centers[, c(4, 5)],col = 1:3, pch = 8, cex = 5)

plot(noab_z_dat[,4],noab_z_dat[,6],main = " centers on 4 and 6 param")
points(km$centers[, c(4, 6)],col = 1:3, pch = 8, cex = 5)

plot(noab_z_dat[,5],noab_z_dat[,6],main = " centers on 5 and 6 param")
points(km$centers[, c(5, 6)],col = 1:3, pch = 8, cex = 5)



tmp_norm =  matrix( ,78,6)
rownames(tmp_norm) <- no_outliers_15$`Country Name`

min_gdp<-min(no_outliers_15$`GDP per capita`)
max_gdp<-max(no_outliers_15$`GDP per capita`)
min_edu<-min(no_outliers_15$`expenditure on education`)
max_edu<-max(no_outliers_15$`expenditure on education`)
min_birth<-min(no_outliers_15$`Birth rate, crude`)
max_birth<-max(no_outliers_15$`Birth rate, crude`)
min_health<-min(no_outliers_15$`general government health expenditure`)
max_health<-max(no_outliers_15$`general government health expenditure`)
min_tobacco<-min(no_outliers_15$`Prevalence of current tobacco use`)
max_tobacco<-max(no_outliers_15$`Prevalence of current tobacco use`)
min_road<-min(no_outliers_15$`Mortality caused by road traffic injury`)
max_road<-max(no_outliers_15$`Mortality caused by road traffic injury`)

tmp_norm[,1]<-(no_outliers_15$`GDP per capita`-min_gdp)/(max_gdp-min_gdp)
tmp_norm[,2]<-(no_outliers_15$`expenditure on education`-min_edu)/(max_edu-min_edu)
tmp_norm[,3]<-(no_outliers_15$`Birth rate, crude`-min_birth)/(max_birth-min_birth)
tmp_norm[,4]<-(no_outliers_15$`general government health expenditure`-min_health)/(max_health-min_health)
tmp_norm[,5]<-(no_outliers_15$`Prevalence of current tobacco use`-min_tobacco)/(max_tobacco-min_tobacco)
tmp_norm[,6]<-(no_outliers_15$`Mortality caused by road traffic injury`-min_road)/(max_road-min_road)
```

```
#--------------------

km_norm <- kmeans(tmp_norm, centers = 3, nstart = 100)
y_kmeans_norm <- km$cluster


clusplot(tmp_norm[, c(3,5)],
        y_kmeans_norm,
        lines = 0,
        shade = TRUE,
        color = TRUE,
        labels = 2,
        plotchar = FALSE,
        span = TRUE,
        cex=0.5,
        main = paste("Cluster on norm data"),
        xlab = 'V3',
        ylab = 'V5')
plot(tmp_norm[,1],tmp_norm[,2],main = " centers on 1 and 2 param in norm data")
points(km_norm$centers[, c(1, 2)],col = 1:3, pch = 8, cex = 5)

plot(tmp_norm[,1],tmp_norm[,3],main = " centers on 1 and 3 param in norm data")
points(km_norm$centers[, c(1, 3)],col = 1:3, pch = 8, cex = 5)

plot(tmp_norm[,1],tmp_norm[,4],main
                        = " centers on 1 and 4 param in norm data")
points(km_norm$centers[, c(1, 4)],col = 1:3, pch = 8, cex = 5)

plot(tmp_norm[,1],tmp_norm[,5],main = " centers on 1 and 5 param in norm data")
points(km_norm$centers[, c(1, 5)],col = 1:3, pch = 8, cex = 5)

plot(tmp_norm[,1],tmp_norm[,6],main = " centers on 1 and 6 param in norm data")
points(km$centers[, c(1, 6)],col = 1:3, pch = 8, cex = 5)

plot(tmp_norm[,2],tmp_norm[,3],main = " centers on 2 and 3 param in norm data")
points(km_norm$centers[, c(2, 3)],col = 1:3, pch = 8, cex = 5)

plot(tmp_norm[,1],tmp_norm[,4],main = " centers on 2 and 4 param in norm data")
points(km_norm$centers[, c(2, 4)],col = 1:3, pch = 8, cex = 5)

plot(tmp_norm[,1],tmp_norm[,5],main = " centers on 2 and 5 param in norm data")
points(km_norm$centers[, c(2, 5)],col = 1:3, pch = 8, cex = 5)

plot(tmp_norm[,2],tmp_norm[,6],main = " centers on 2 and 6 param in norm data")
points(km_norm$centers[, c(2, 6)],col = 1:3, pch = 8, cex = 5)

plot(tmp_norm[,3],tmp_norm[,4],main = " centers on 3 and 4 param in norm data")
points(km_norm$centers[, c(3, 4)],col = 1:3, pch = 8, cex = 5)

plot(tmp_norm[,3],tmp_norm[,5],main = " centers on 3 and 5 param in norm data")
points(km_norm$centers[, c(3, 5)],col = 1:3, pch = 8, cex = 5)

plot(tmp_norm[,3],tmp_norm[,6],main = " centers on 3 and 6 param in norm data")
points(km_norm$centers[, c(3, 6)],col = 1:3, pch = 8, cex = 5)

plot(tmp_norm[,4],tmp_norm[,5],main = " centers on 4 and 5 param in norm data")
points(km_norm$centers[, c(4, 5)],col = 1:3, pch = 8, cex = 5)

plot(tmp_norm[,4],tmp_norm[,6],main = " centers on 4 and 6 param in norm data")
points(km_norm$centers[, c(4, 6)],col = 1:3, pch = 8, cex = 5)

plot(tmp_norm[,5],tmp_norm[,6],main = " centers on 5 and 6 param in norm data")
points(km_norm$centers[, c(5, 6)],col = 1:3, pch = 8, cex = 5)
 means<-data.frame(c("rich"),c("poor"),c("median"))
l <- table( km$cluster)
l
```