

Исследование датасета world-happiness-report с kaggle

Ссылка на данные: <https://www.kaggle.com/datasets/usamabuttar/world-happiness-report-2005-present>

Описание задачи исследования: целью исследования является закрепление пройденного материала + проверить гипотезы, выдвинутые после предварительного анализа.

Описание столбцов:

1. Country Name - название страны, для которой проводилось исследование, строковые данные
2. Regional Indicator - часть света, в которой располагается страна, строковые данные
3. Year - год исследования, непрерывные данные
4. Life Ladder - показатель удовлетворенности жизнью по шкале Кантрила (лестница Кантрила), непрерывные данные
5. Log GDP Per Capita - показатель, рассчитанный по паритету покупательной способности (ППС), скорректированному с учетом постоянных международных долларов за 2017 год, взят из Показателей мирового развития (WDI) Всемирного банка
6. Social Support - средняя оценка, рассчитанная путём сбора статистики ответа на вопрос: "есть ли у Вас родственники и друзья, которые могут помочь Вам в любое время и в любой момент, когда вы в них нуждаетесь?" Ответом был 0 (нет) или 1 (да)
7. Healthy Life Expectancy At Birth - параметр, показывающий среднюю продолжительность жизни в стране. Составлен на основе данных из хранилища данных Глобальной обсерватории здравоохранения Всемирной организации здравоохранения (ВОЗ).
8. Generosity - это остаток от регрессии среднего национального показателя ответов GWP на вопрос о пожертвовании «Жертвовали ли вы деньги на благотворительность в прошлом месяце?» на Log GDP Per Capita.
9. Freedom To Make Life Choices - средняя оценка, рассчитанная путём сбора статистики ответа на вопрос: "удовлетворены Вы или нет свободой выбора принятия решений?" Ответом был 0 (нет) или 1 (да)
10. Perceptions Of Corruption - являются средними бинарными ответами на два вопроса GWP: "Широко распространена коррупция в правительстве или нет?" и "Широко распространена коррупция в бизнесе или нет?"

Какие данные взяты из таблицы: из таблицы было принято решение взять данные по первым 7 показателям после Year за 2022 год, убрав строки, содержащие хотя бы 1 пустой столбец, для корректного анализа. И ещё была удалена о 33 странах (начиная с Jamaika и заканчивая Paraguay) для уменьшения обширности вычислений

Какие данные анализируются: все столбцы, начиная с Life Ladder.

Предварительный анализ

1. Список гипотез

1. Одну из главных ролей счастья в жизни человека является наличие родных и близких, готовых прийти на помощь в любой трудный момент.
2. Дороговизна жизни при наличии средств - один из влияющих факторов, влияющих на счастье человека.
3. Люди, проживающие в странах с большей продолжительностью жизни, счастливее, чем люди в странах с меньшей продолжительностью.
4. Люди, проживающие в странах с более высокой средней продолжительностью жизни, щедрее, чем из других стран.
5. На Log GDP напрямую влияют данные из Perceptions Of Corruption
6. Чем больше люди видят коррупцию в стране и бизнесе, тем меньше удовлетворенность жизнью.
7. Чем больше покупательская способность человека, тем он добре к другим людям.
8. Чем больше покупательская способность человека, тем дольше продолжительность его жизни.

Ключевым, уникальным фактором для каждой строки является страна, поэтому при анализе каждого столбца будет также включаться страна для более наглядного представления и анализа столбцов

При нормализации данных будем пользоваться только удалением потому, что в таблице нет повторяющихся и пропущенных значений

Ответ на удаление определенного кол-ва данных при корректировки переменных таков: постепенно анализировались получавшиеся гистограммы и ящиковыe диаграммы, а затем принимались решения: нужно или нет еще удалять в столбце данные.

2. Проведем предварительный анализ по каждому столбцу

1.1. Столбец Life Ladder (в таблице лист “предварительный анализ Life Ladder”)

1. Проведём её характеристики, вычислив моду, медиану, среднее значение, коэффициент вариации, квантили и квинтили:

X-ка Life Ladder												
mean	mode	mode (окр. до целого)	median	размах вариации	коэффициент вариации	дисперсия	станд откл	квартиль 1	квартиль 2	квартиль 3	IQR	квинтиль 1
5.67875	nan	6.00000	5.92888	5.16857	21.22748361	1.453124512	1.205456143	4.87774682	5.928882122	6.611006737	1.733259917	4.3966460;

Как можно увидеть среднее значение равно примерно 5.7. Если округлить 5.7 по правилам мат. округления, то получим 6, что означает по шкале Кантрила, что люди нормально живут в странах (неплохо, но можно было бы и лучше).

Моду стандартного ряда не удалось определить, так как все значения в ряде уникальные. Однако, если округлить до целого каждое число Life Ladder, то получим значение 6, что подтверждает выводы, сделанные по среднему

Медиана получилась примерно 5.93, что близко к 6, а это подтверждает вывод, сделанный по среднему значению.

По значениям данные не слабо разбросаны, что подчёркивает размах вариации (5.16857), значения квантили и коэффициент вариации, и интерпретация по среднему значению всё ещё является корректной

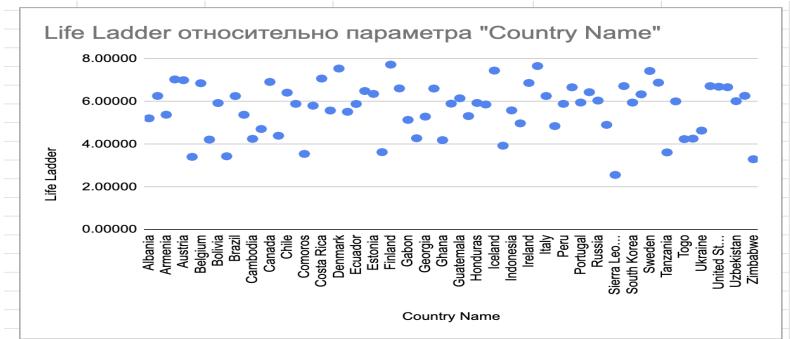
2. Z преобразование

Country name	Z преобразование
Sierra Leone	-2.586840186
Zimbabwe	-1.976457122
Bangladesh	-1.884116452
Botswana	-1.861102041
Comoros	-1.769909501
Tanzania	-1.711308018
Ethiopia	-1.701165514
India	-1.45084973
Ghana	-1.234302405
Benin	-1.212342969
Togo	-1.194377559
Cambodia	-1.185004629
Tunisia	-1.176221882
Gambia	-1.160814201
Chad	-1.063585723
Ukraine	-0.8638353058
Cameroon	-0.8016416902
Ivory Coast	-0.6886008675
Senegal	-0.6403651848
Iran	-0.5821501694
Gabon	-0.4473421542
Albania	-0.3870227295
Georgia	-0.3202081146
Guinea	-0.2996865592
Bulgaria	-0.2492028768

Можно увидеть, что разница 1-ого и 2-ого z-преобразований велика по сравнению с последующими. Это означает, что 1-ое число является выбросом. В остальных случаях разница между z-преобразованиями не такая большая, поэтому остальные данные можно считать нормальными. Но, если совместить с box plot результаты, то окажется, что можно было бы удалить все данные до Ганы.

3. Проведём графическое представление исходных данных

1. Построим точечное распределение данных



По построенному распределению видно, что данные независимы и разбросаны в совершенно случайном порядке.

Максимальное скопление данных можно увидеть на линии с Life Ladder = 6, что подтверждает выводы сделанные при характеристики Life Ladder. Лучше всего себя ощущают люди, проживающие в Финляндии. Хуже же всего в Сьерре-Леоне, находящегося в странах Африки к югу от Сахары.

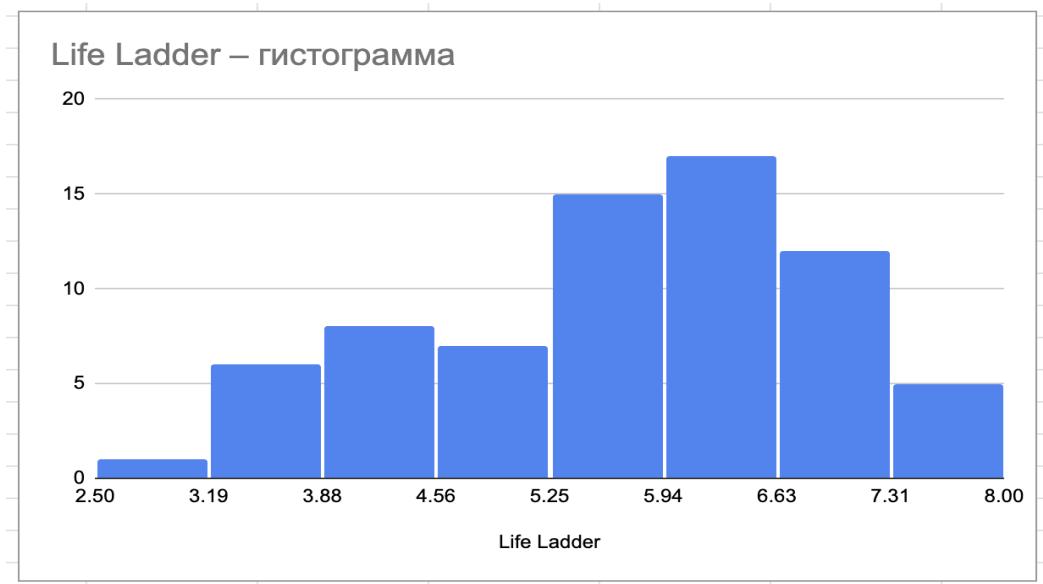
2. Построим листовую диаграмму

Для построения воспользуемся отсортированным столбцом, в котором данные округлены до десятых ("Life Ladder, округлённый до десятых" в таблице)

Steam-leaf диаграмма					
leaf					
				9	
				9	
				9	
				9	
				7	
				7	
				7	
				7	
				7	
			9	6	
			9	6	
			9	5	
			9	4	
			9	4	
			9	4	
		9	9	3	
		8	8	3	
		7	6	3	
		6	6	3	7
	9	4	5	3	7
	6	3	4	2	5
	6	3	4	0	4
	5	3	3	0	4
	4	2	2	0	1
	4	2	1	0	0
6	3	2	0	0	0
2	3	4	5	6	7
steam					

Если мысленно провести линию тренда, то ни одно из основных распределений не соответствует полученному графику. Однако также, как и в предыдущих выводах, видно что основная доля значений находится ближе к стеблю 6.

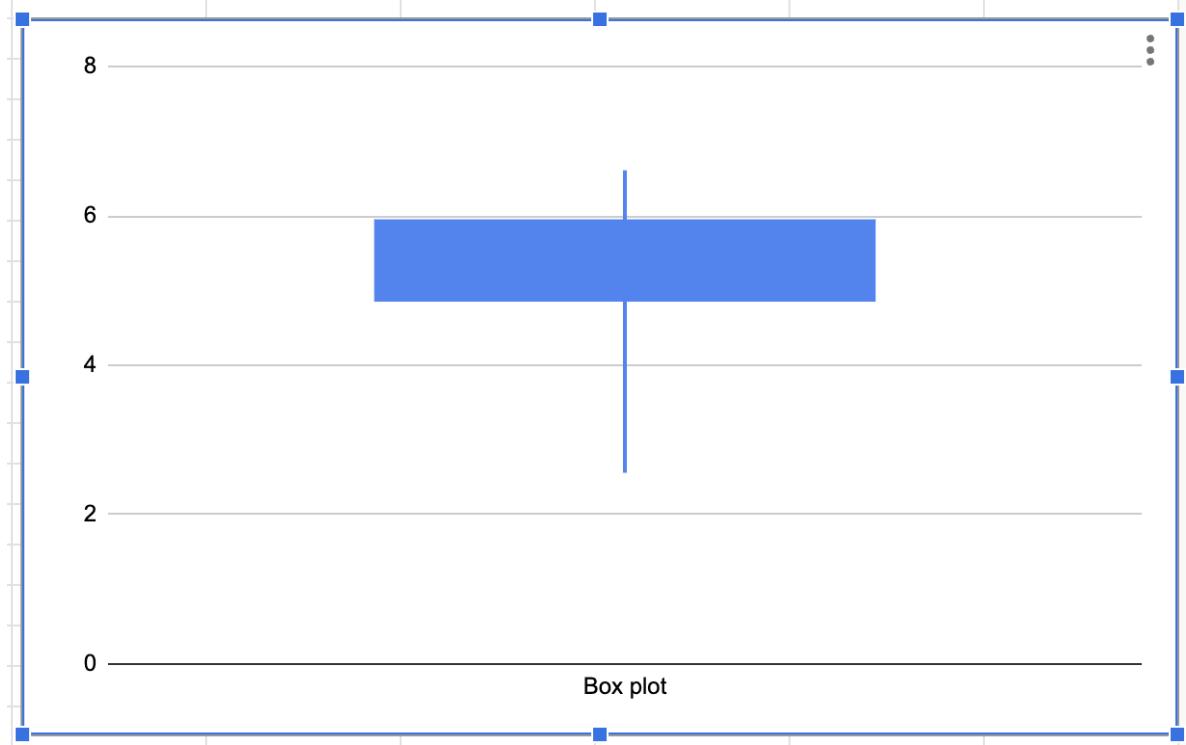
3. Построим гистограмму



Кол-во стран, в которых удовлетворенность жизнью между 7 и 8 равно 5. Среди них Australia, Costa Rica, Sweden, Iceland, Denmark, Israel, Finland. Наименьшее значение = 2.60 у страны Сьерре-Леоне. Наибольшее количество (17) стран имеют Life Ladder между 5.94 и 6.63.

4. Продиагностируем выбросы ряда

- Построим ящиковую диаграмму



Можно увидеть, что минимум довольно далеко находится от остальных данных. За счёт этого выброс между 1 квартileм и минимумом довольно высок. В остальном данные распределены нормально.

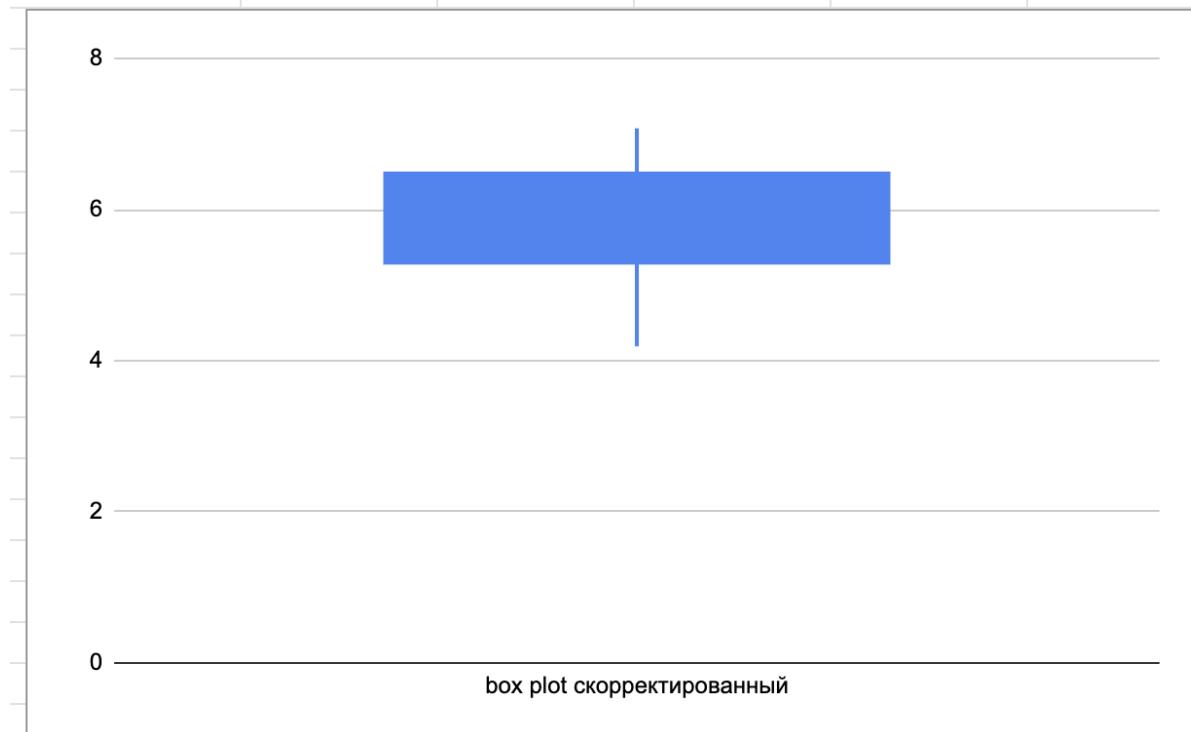
- Проверим правило 3 сигм

правило 3 сигм		
mean - 3сигма	mean	mean + 3 сигма
2.06238	5.67875	9.29512

=> оно верно, но слишком большое значение у правой границы получилось, что характеризует наличие выброса в данных

5. Корректирование столбца

Удалим из столбца первые 15 данных из сортированного списка и посмотрим, что получится



Видно, что выброс всё ещё есть небольшой, но уже значительно меньше

1.2. Столбец Log GDP Per Capita (в таблице лист “предварительный анализ Log GDP Per Capita”)

1. Проведём её характеристики, вычислив моду, медиану, среднее значение, коэффициент вариации, квантили и квинтили:

Х-ка Log GDP Per Capita															
mean	mode	mode (окр. до сотых)	median	размах вариации	коэффициент вариации	дисперсия	станд откл	квартиль 1	квартиль 2	квартиль 3	IQR	квинтиль 1	квинтиль 2		
9.622713	нан	9.63000	9.667766	4.363785	11.64793461	1.256298784	1.120847351	8.796590805	9.667765617	10.61582852	1.81923771	8.611810684	9.44880586		

В среднем покупательская способность высока (лежит ближе к максимуму в отсортированном столбце Log GDP Per Capita), и это означает, что в среднем стоимость товаров, рассмотренных в расчёте показателя, в стране высока, что может негативно повлиять на ощущение людьми удовлетворенностью жизнью.

Значение наиболее встречаемого числа в стандартном ряде (отсортированный ряд Log GDP Per Capita без округления) не удалось определить, поэтому я решил округлить каждое число до сотых. И оно оказалось равным 9.63, что подтверждает вывод, сделанный при интерпретации среднего значения

Значение медианы (9.667766) также подтверждает вывод, сделанный при анализе среднего.

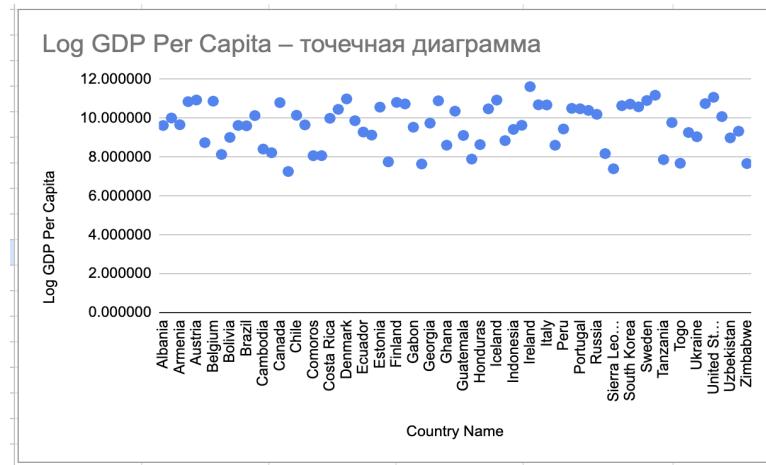
Размах вариации не сильно большой (4.363785), что означает небольшую разбросанность данных и шанс встретить выбросы небольшой.

Данные имеют среднюю вариативность, и это означает, что изначальные данные не сильно разбросаны. А также это означает, что интерпретация по среднему значению корректная (является некорректной, когда > 33 коэффициент).

Поочерёдные разницы между квантилями и квинтилями низка (< 1), что подтверждает выводы, сделанные в предыдущих 2 абзацах.

2. Проведём графическое представление исходных данных

1. Построим точечное распределение данных



По построенному распределению видно, что данные независимы и разбросаны в совершенно случайном порядке. Интересная переплетение в начале получилось в виде днк.

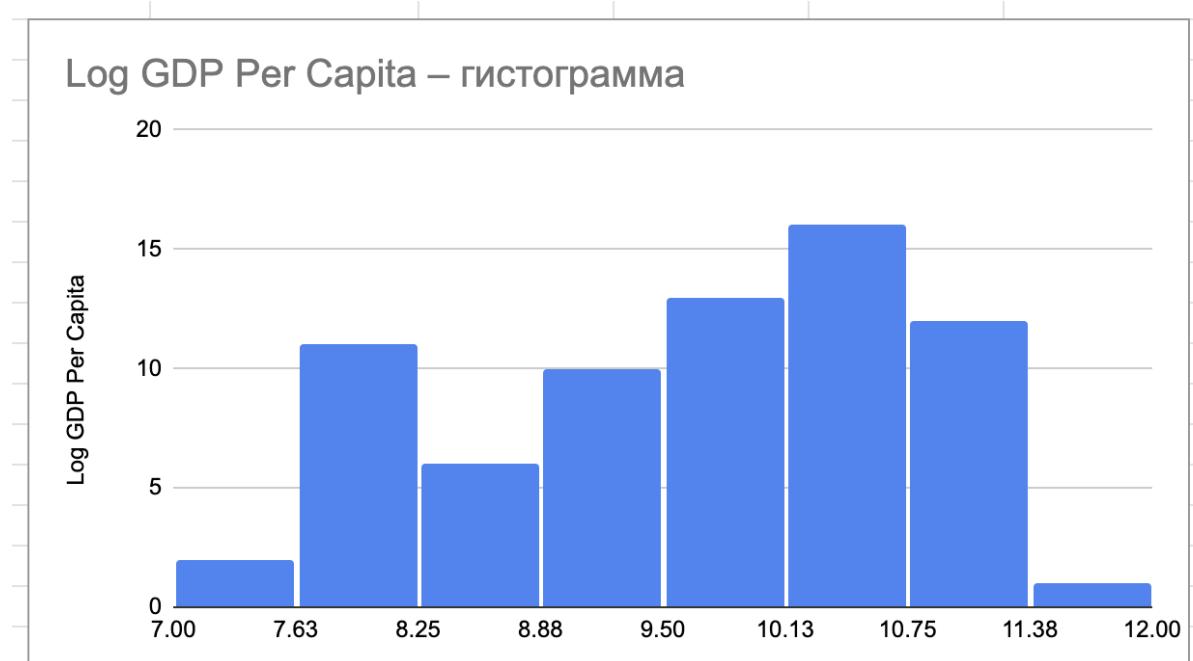
Максимальное скопление данных можно увидеть на линии с Log GDP Per Capita примерно равным 10, что подтверждает выводы сделанные при характеристики Log GDP Per Capita. Наибольший ППС в Ирландии. Наименьший же в Чаде, находящегося в странах Африки к югу от Сахары.

2. Построим листовую диаграмму

Steam-leaf диаграмма	
leaf	99
	94
	94
	91
	90
	88
	85
	81
	80
	78
	75
	74
	73
	66
	69
	64
	64
	63
	63
	57
	99
	63
	51
	85
	61
	49
	74
	54
	48
	64
	45
	46
	90
	62
	43
	40
	87
	61
	33
	36
	76
	41
	29
	20
	69
	23
	27
	15
	67
	18
	13
	14
	65
	14
	12
	8
	62
	40
	8
	5
	1
	18
	26
	7
	1
	0
	8
	9
	10
	11
7	8

Если мысленно провести линию тренда, то ни одно из основных распределений не соответствует полученному графику. Однако также, как и в предыдущих выводах, видно что основная доля значений находится ближе к стеблю 10.

3. Построим гистограмму

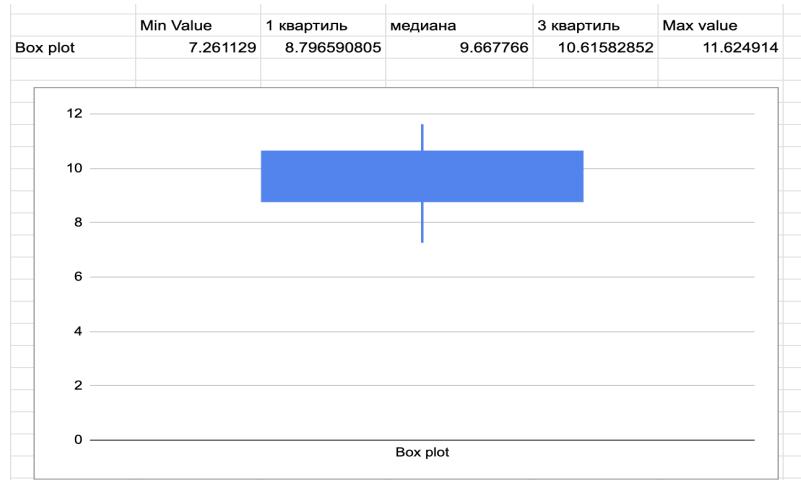


Кол-во стран, в которых ППС находится в крайнем пределе = 3. Среди них United States, Switzerland, Ireland. Наименьшее значение = 7.26 у страны Чада. Наибольшее количество (16) стран имеют Log GDP Per Capita между 10.13 и 10.75, что вновь говорит о том, что ППС в странах большой и за счёт этого жизни людей, проживающих в таких странах, могут им казаться не такими счастливыми.

Листовая и точечная диаграммы чем то напоминает графики те же по столбцу Life Ladder. Для того, чтобы выдвинуть гипотезу о взаимосвязи построим рассеянную диаграмму [в пункте 2](#).

3. Продиагностируем выбросы ряда

1. Построим ящиковую диаграмму



Можно увидеть, что данные распределены равномерно, что подтверждает правильность небольших разниц между квартилями и небольшой размах

2. Проверим правило 3 сигм

правило 3 сигм		
mean - 3сигма	mean	mean + 3 сигма
6.260171	9.622713	12.985255

=> оно верно, и у данных нет выбросов.

4. Корректировка данных

Тут корректировка данных не нужна, так как при удалении определённых значений гистограмма становится только хуже и появляется всё больше и больше скачков между промежутками. Последние корректировки можете увидеть в соответствующем листе (“предварительный анализ Log GDP Per Capita”)

1.3. Столбец Social Support (в таблице лист “предварительный анализ Social Support”)

1. Проведём её характеристики, вычислив моду, медиану, среднее значение, коэффициент вариации, квантили и квинтили:

Х-ка Social Support														
mean	mode	mode (окр. до сотых)	median	размах вариации	коэффициент вариации	дисперсия	станд откл	квартиль 1	квартиль 2	квартиль 3	IQR	квинтиль 1	квинтиль 2	
0.79966	nan	0.88000	0.86234	0.61912	18.07058057	0.02088109703	0.1445029309	0.7343912125	0.862344146	0.903216809	0.1688255965	0.666171908	0.81116921	

В среднем социальная поддержка высока (примерно 0.8 равна), и это означает, в среднем у многих людей есть близкие, друзья, готовые прийти на помощь в любое время.

Значение наиболее встречаемого числа в стандартном ряде (отсортированный ряд Social Support без округления) не удалось определить, поэтому я решил округлить каждое число до сотых. И оно оказалось равным 0.88, что подтверждает вывод, сделанный при интерпретации среднего значения

Значение медианы (0.86234) также подтверждает вывод, сделанный при анализе среднего.

Несмотря на высокие показатели предыдущих величин, значение размах вариации оказалось немаленьким (0.61912), что говорит о том, что всё таки есть страны или страны, где большая часть людей считает, что у них нет таких близких и друзей.

Данные имеют среднюю вариативность, и это означает, что изначальные данные не сильно разбросаны. А также это означает, что интерпретация по среднему значению корректная (является некорректной, когда > 33 коэффициент).

Последние 2 квартиля и 3 квинтиля имеют небольшую разницу между собой (< 0.06). В то время как разница между разницами между 1 и 2 квартилями чуть < 0.13 , а разница между 1 и 2 квинтилями чуть < 0.15 . Это означает то, что в $\leq 20\%$ данных есть низкие числа, которые и занижают разницы. Это видёт к выводу, сделанному при анализе размаха вариации.

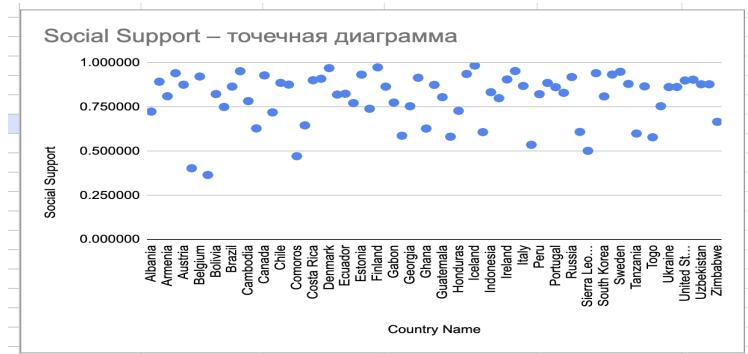
2. Z преобразование

Country Name	Z преобразован ие
Benin	-3.003241431
Bangladesh	-2.740099897
Comoros	-2.268736537
Sierra Leone	-2.058728471
Ivory Coast	-1.822543774
Togo	-1.527511717
Guinea	-1.506146461
Gambia	-1.466847383
Tanzania	-1.38044962
India	-1.326027769
Senegal	-1.318113915
Ghana	-1.186818066
Cameroon	-1.180262511
Congo (Brazzaville)	-1.060981394
Zimbabwe	-0.9237742874
Chad	-0.5535404265
Albania	-0.5229677871
Honduras	-0.4916522631
Ethiopia	-0.4117036356
Botswana	-0.3408976392
Georgia	-0.312582089
Tunisia	-0.3108599855
El Salvador	-0.1887793613
Gabon	-0.1689394592
Cambodia	-0.1111266042
Iran	0.002839291892

Можно увидеть, что поочерёдные разницы с 1 по 6 велики по сравнению со следующими. Это означает, что значения стран от Benin до Ivory Coast являются выбросными.

3. Проведём графическое представление исходных данных

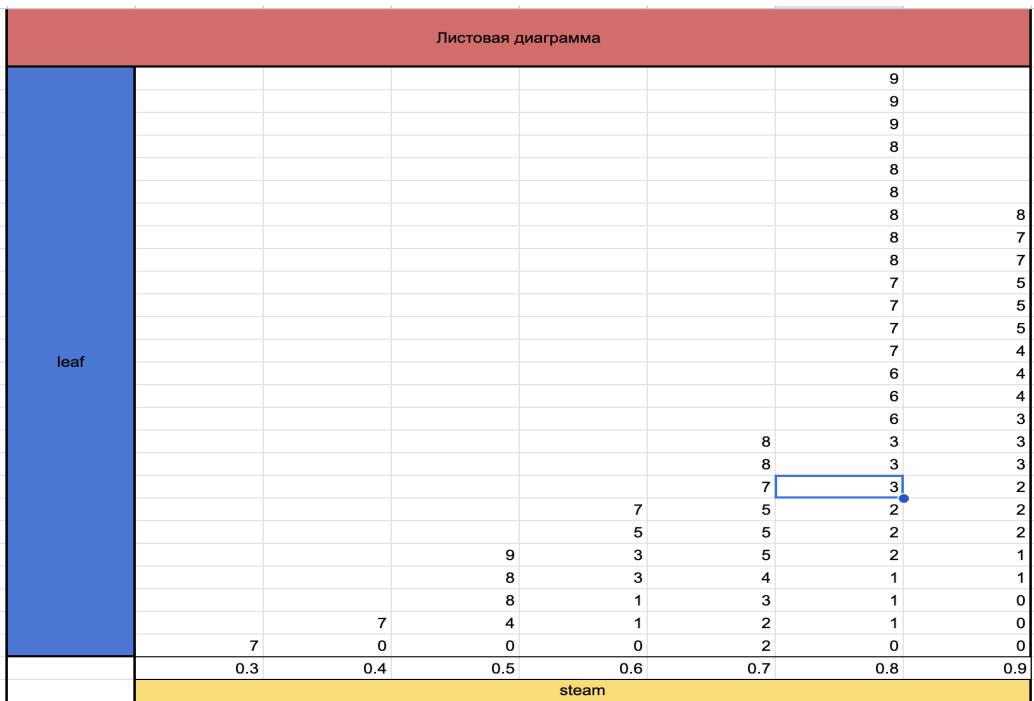
- Построим точечное распределение данных



По построенному распределению видно, что данные независимы и не сильно упорядочены.

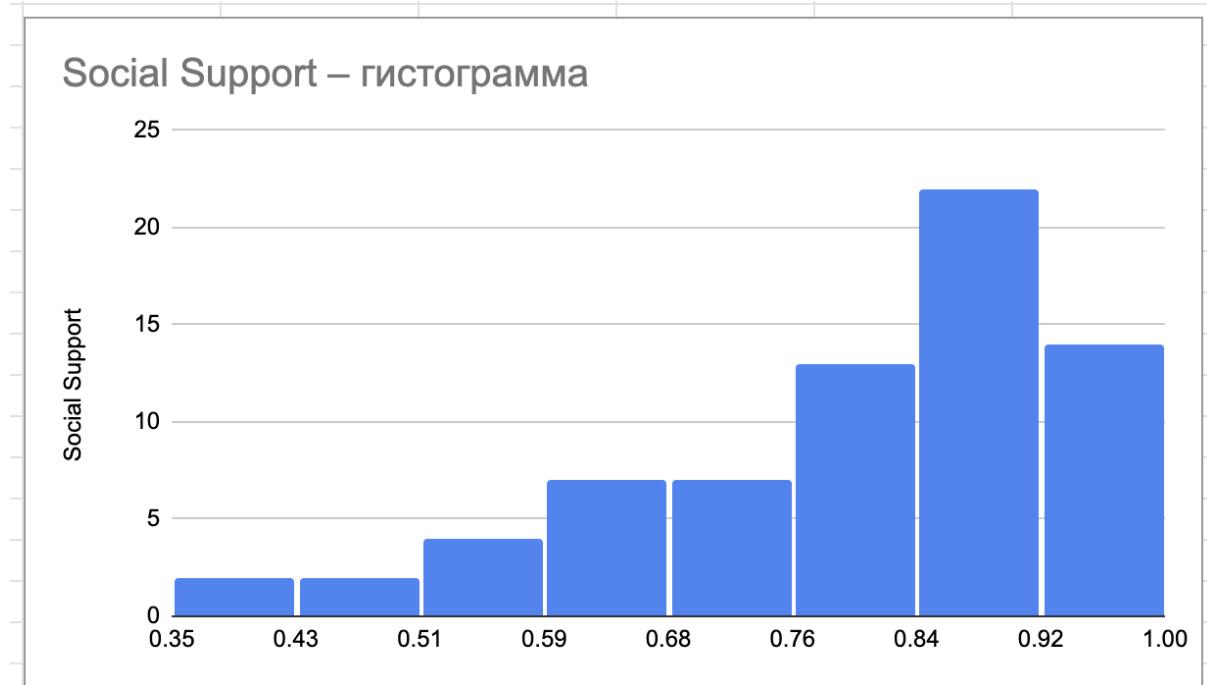
Максимальное скопление данных можно увидеть на линии с Social Support > 0.825, что подтверждает выводы сделанные при характеристики Social Support. Больше всего люди верят и надеются в сложные моменты на близких в Исландии. Хуже же всего в Бенине, находящегося в странах Африки к югу от Сахары.

2. Построим листовую диаграмму



Если мысленно провести линию тренда, то ни одно из основных распределений не соответствует полученному графику. Однако также, как и в предыдущих выводах, видно что основная доля значений находится ближе к стеблю 0.8.

3. Построим гистограмму



Кол-во стран, в которых люди могут надеяться в трудные моменты на своих близких и друзей = 14. Среди них все страны от России до Исландии в отсортированном списке. Наименьшее значение = 0.37 у страны Бенин. Наибольшее количество (22) стран имеют Social Support между 0.84 и 0.92.

Листовая и точечная диаграммы чем то напоминает графики те же по столбцу Life Ladder. Для того, чтобы выдвинуть гипотезу о взаимосвязи построим рассеянную диаграмму [в пункте 2](#).

4. Продиагностируем выбросы ряда

1. Построим ящиковую диаграмму



Можно увидеть, что минимум довольно далеко находится от остальных данных. За счёт этого выброс между 1 квартилем и минимумом довольно высок. В остальном данные распределены нормально, равномерно.

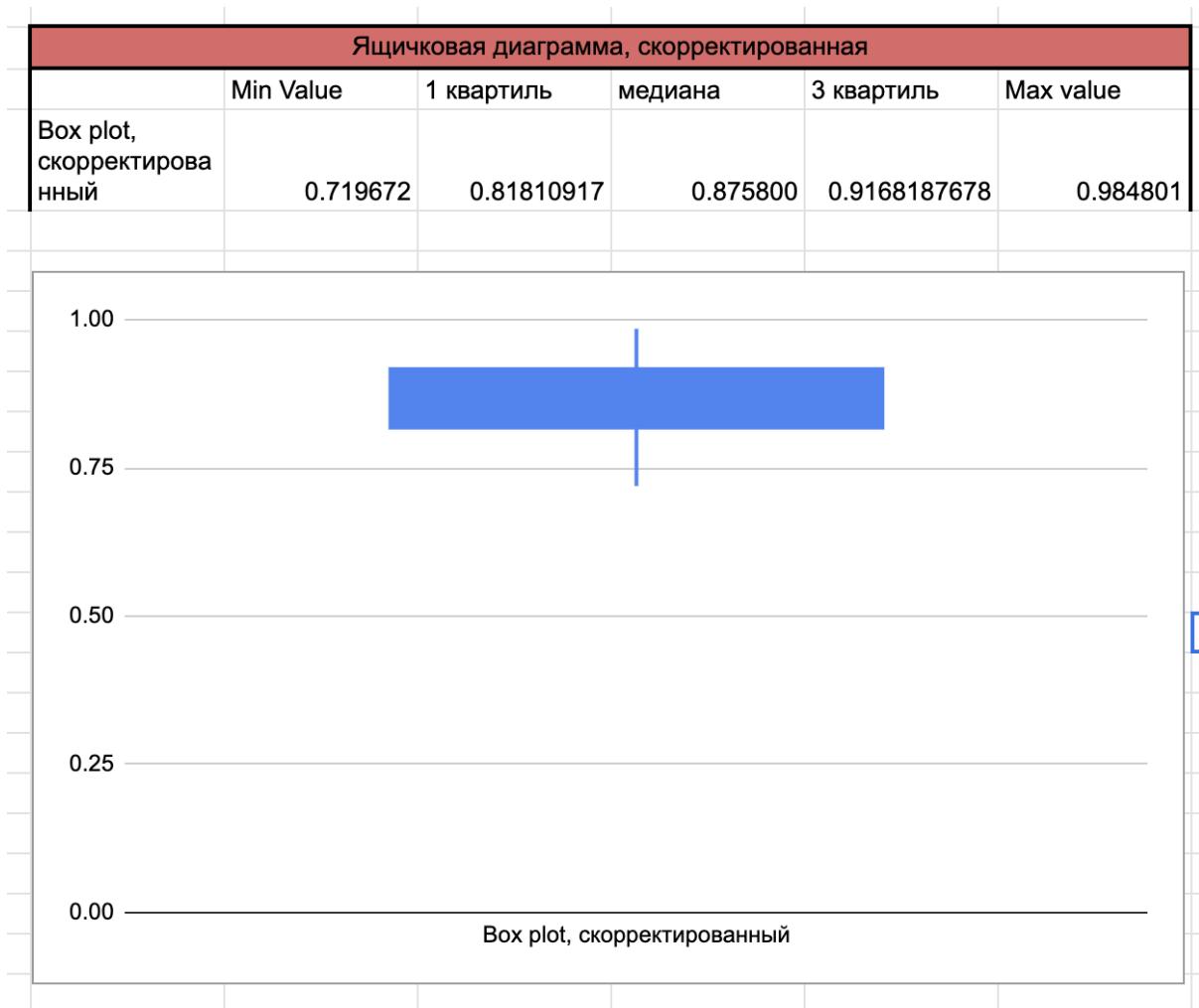
2. Проверим правило 3 сигм

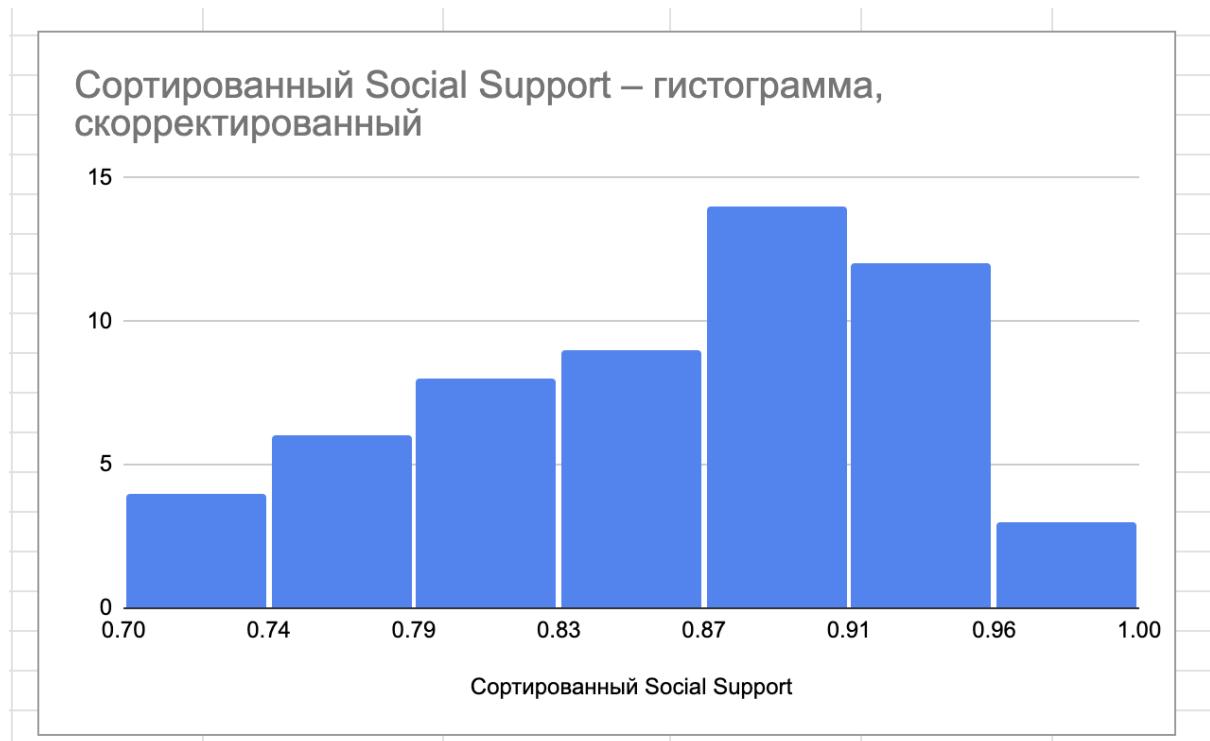
правило 3 сигм		
mean - 3сигма	mean	mean + 3 сигма
0.36615	0.79966	1.23317

=> оно верно, но слишком большое значение у правой границы получилось и минимальное значение не попало в диапазон, что характеризует наличие выброса в данных.

5. Корректировка данных

Удалим первые 17 данных и посмотрим, что получится





Ящиковая диаграмма получилась отличной, а вот гистограмма Social Support немного скосившаяся всё равно, но зато уже напоминает нормальное распределение.

1.4. Столбец Healthy Life Expectancy At Birth (в таблице лист “предварительный анализ Healthy Life Expectancy At Birth”)

1. Проведём её характеристики, вычислив моду, медиану, среднее значение, коэффициент вариации, квантили и квинтили:

Х-ка Healthy Life Expectancy At Birth													
mean	mode	median	размах вариации	коэффициент вариации	дисперсия	станд откл	квартиль 1	квартиль 2	квартиль 3	IQR	квинтиль 1	квинтиль 2	
65.80916	56.52500	67.05000	20.80000	8.683222998	32.65386096	5.714355691	61.64999962	67.05000305	71.08750153	9.43750191	59.84999847	65.59999847	

В среднем продолжительность жизни низка (всего лишь 66 лет примерно), а это означает, что люди, проживающие в странах с такой продолжительностью, могут часто болеть, иметь плохое психологическое состояние, а это явный признак, который негативно может влиять на восприятие человеком ощущения удовлетворенности его жизнью.

Значение наиболее встречаемого числа в стандартном ряде (отсортированный ряд Life Expectancy At Birth без округления) = 56.525, то есть примерно 57 лет. Сделать вывод можно такой же, как и при анализе среднего значения, но разница большая между показателями, что является признаком того, что нужно вычислять моду не в

изначальном ряде, а округлённом до 10. Тогда получается значение 71.4 - перебор наоборот, но по разнице ближе, чем предыдущий результат. Если до сотых округлить, то получится 56.53, что говорит о том, что по разнице больше, чем результат, округлённый до 10.

Значение медианы (67.05) также подтверждает вывод, сделанный при анализе среднего.

Размах вариации получился средним (20.8), что говорит с большой натяжкой о том, что в данных присутствуют все категории по продолжительности жизни.

Данные имеют низкую вариативность (8.7), и это означает, что изначальные данные почти не разбросаны. А также это означает, что интерпретация по среднему значению корректная (является некорректной, когда > 33 коэффициент).

Все разницы между квантилями и квинтилями низкие, что подчёркивает структурированность данных (данные без выбросов)

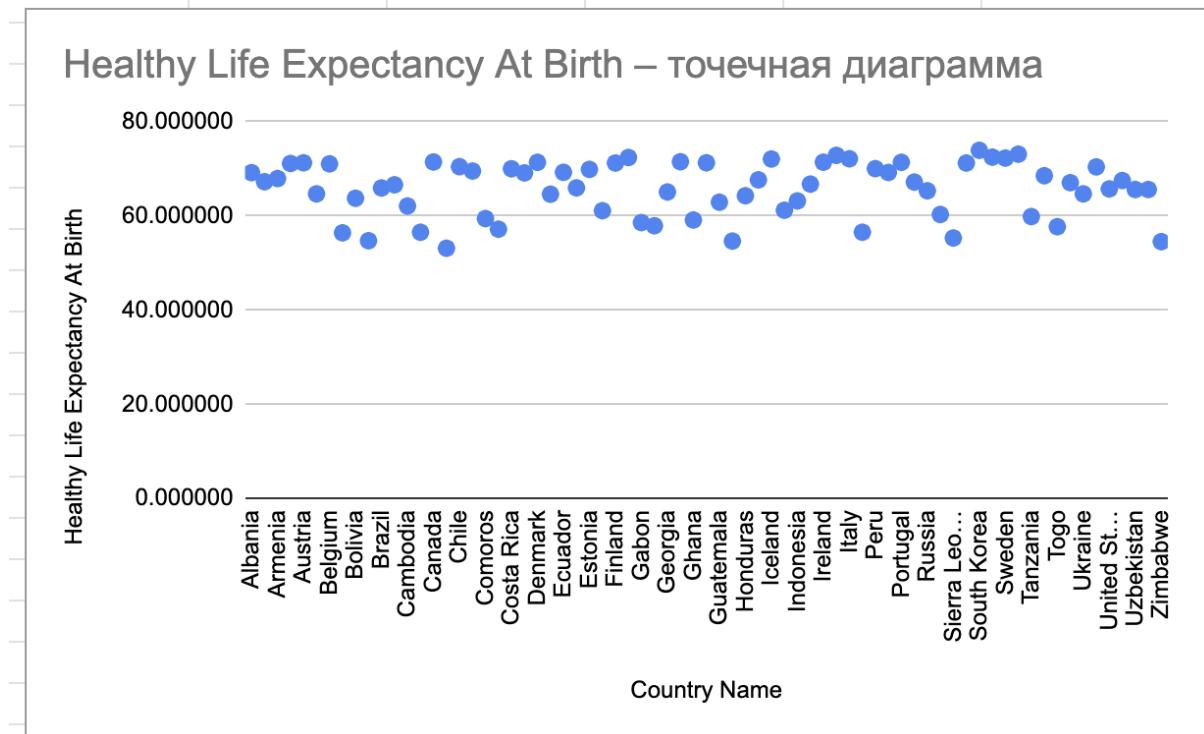
2. Z преобразование

Country Name	Z преобразован ие
Chad	-2.219699961
Zimbabwe	-1.974702693
Guinea	-1.952827961
Botswana	-1.939703657
Sierra Leone	-1.839079755
Benin	-1.646581711
Cameroon	-1.624706979
Ivory Coast	-1.624706979
Congo (Brazzavil	-1.510958773
Togo	-1.419084632
Gambia	-1.384084926
Gabon	-1.265961507
Ghana	-1.169712819
Comoros	-1.117213595
Tanzania	-1.04283964
Senegal	-0.9640904697
Ethiopia	-0.8240923188
India	-0.8065921324
Cambodia	-0.6490944618
Guatemala	-0.5090956407
Indonesia	-0.4609716308
Bolivia	-0.36034773
Honduras	-0.2684735874
Dominican Repu	-0.2115998193

По разницам Z преобразований можно увидеть, что выводы, сделанные при х-ке ряда верны

3. Проведём графическое представление исходных данных

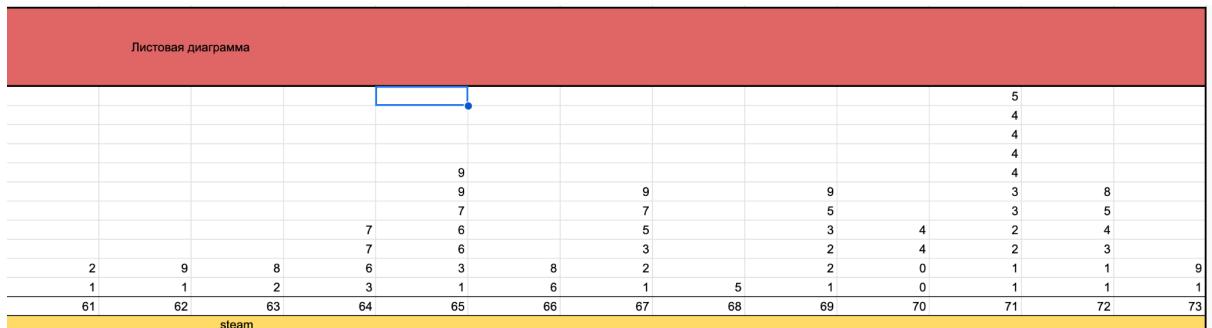
1. Построим точечное распределение данных



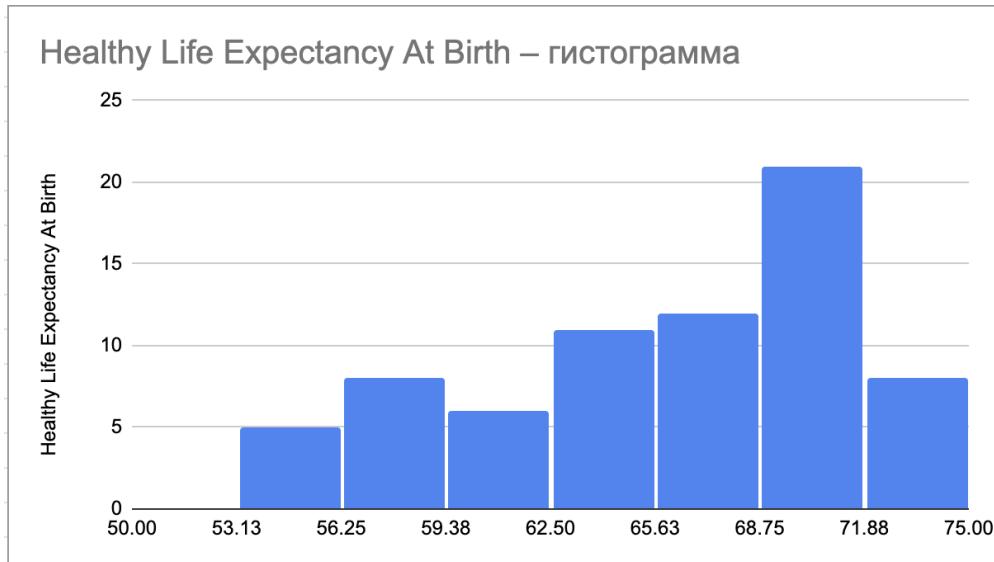
По построенному распределению видно, что данные независимы и не сильно упорядочены. Но зато данные не сильно разбросаны, что подтверждает выводы по данному поводу, сделанные при характеристики ряда.

Самая большая продолжительность жизни выборки в Южной Корее. Хуже же всего в Чаде, находящегося в странах Африки к югу от Сахары.

2. Построим листовую диаграмму



3. Построим гистограмму

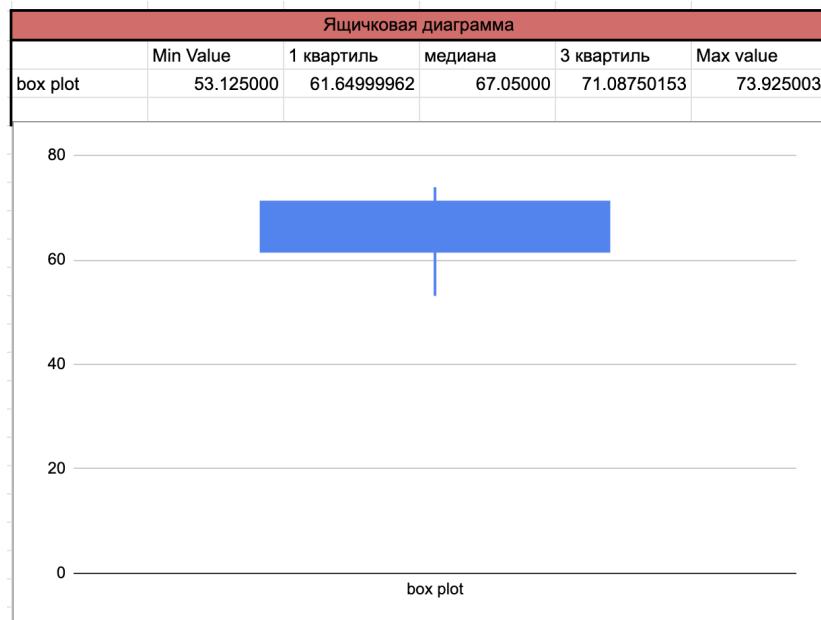


Кол-во стран с самыми высокими продолжительностями жизни = 8. Среди них все страны в упорядоченном списке от Исландии до Южной Кореи. В самый маленький промежуток по продолжительности жизни попадают 5 стран от Чада до Сьерра-Леона при чём все государства располагаются в странах Африки к югу от Сахары

Точечная диаграмма чем то напоминает тот же график по столбцу Life Ladder. Для того, чтобы выдвинуть гипотезу о взаимосвязи построим рассеянную диаграмму [в пункте 2](#).

4. Продиагностируем выбросы ряда

1. Построим ящиковую диаграмму



В данные неравномерно распределены, и выбросы есть в данных от минимального значения до 1 квартиля.

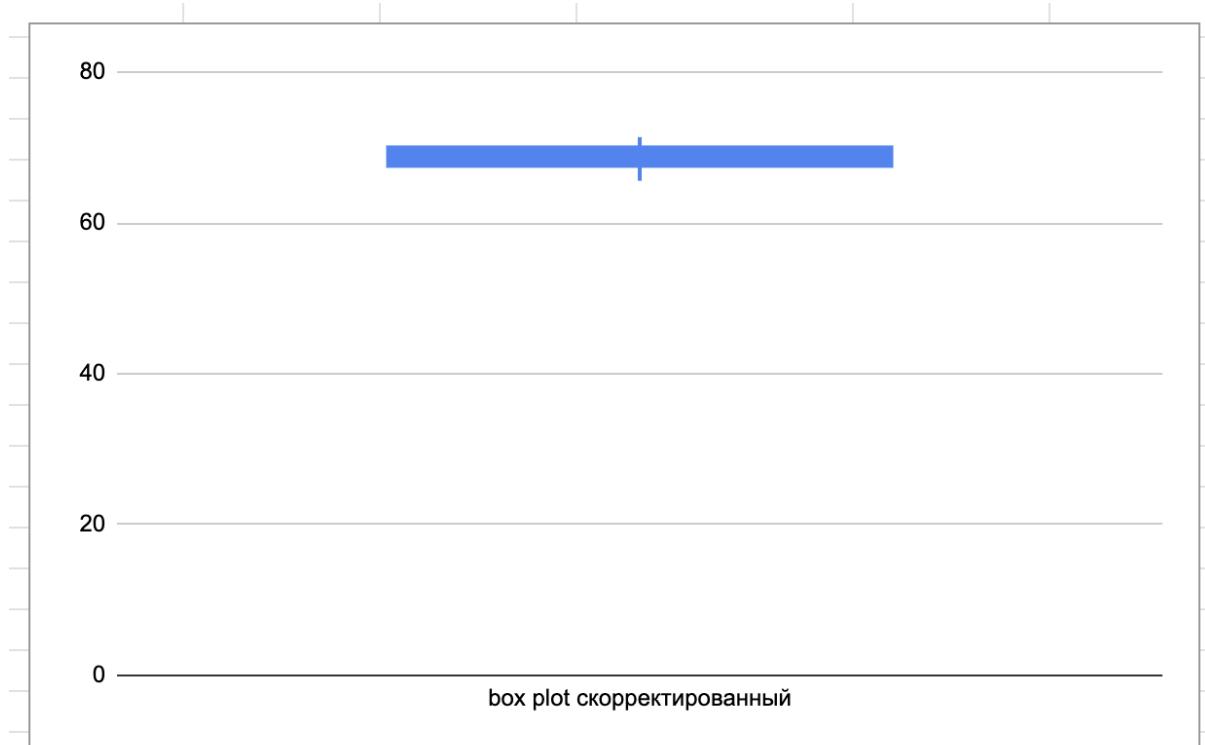
2. Проверим правило 3 сигм

правило 3 сигм		
mean - 3сигма	mean	mean + 3 сигма
48.66609	65.80916	82.95222

=> оно верно, но выброс всё же есть потому, что разница между крайним левым значением и минимумом ровно, как и разница между крайним правым значением и максимумом велика.

5. Корректировка данных

Удалим первые 30 данных и посмотрим, что получится



Ящичковая диаграмма получилась отличной, а вот гистограмма Healthy Life Expectancy At Birth немножко скосившаяся всё равно, но зато уже напомниает нормальное распределение.

1.5. Столбец Freedom To Make Life Choices (в таблице лист “предварительный анализ Freedom To Make Life Choices”)

1. Проведём её характеристики, вычислив моду, медиану, среднее значение, коэффициент вариации, квантили и квинтили:

Х-ка Healthy Freedom To Make Life Choices													
mean	mode	mode, округленный до сотых	median	размах вариации	коэффициент вариации	дисперсия	станд откл	квартиль 1	квартиль 2	квартиль 3	IQR	квинтиль 1	квинтиль 2
0.79519	нап	0.86000	0.80225	0.50122	13.08586988	0.01266490969	0.1125384809	0.726356	0.802249789	0.8792911765	0.1529351765	0.710519016	0.7814088

В среднем ответ на вопрос о свободном выборе решения высок (0.79519), а это означает, что люди, проживающие в странах с такой возможностью, могут удовлетворены политикой государства в этом плане, а значит это увеличивает удовлетворенность ими жизнью.

Значение наиболее встречаемого числа в стандартном ряде (отсортированный ряд Freedom To Make Life Choices без округления) не удалось определить, поэтому я решил округлить каждое число до сотых. И оно оказалось равным 0.86 (если округлить до десятых, то вообще получится что мода = 0.9), что в целом подтверждает вывод, сделанный при интерпретации среднего значения.

Значение медианы (0.80225) также подтверждает вывод, сделанный при анализе среднего.

Размах вариации получился немаленьким (0.50122), что говорит о том, что можно вполне встретить выбросы в данных, а также, что данные неоднородны (данные не сфокусированы на маленьком промежутке. Если бы 0.2 размах был, то можно было бы сказать спокойно, что размах однородный).

Данные имеют среднюю вариативность (13.08586988), и это означает, что изначальные данные немного разбросаны. А также это означает, что интерпретация по среднему значению корректная (является некорректной, когда > 33 коэффициент).

Все разницы между квантилями и квинтилями низкие с почти постоянной разницей, что говорит о равномерности распределения данных (в каждом $n\%$ значений данные распределены по группам равномерно (группа - цифра в десятых долях)).

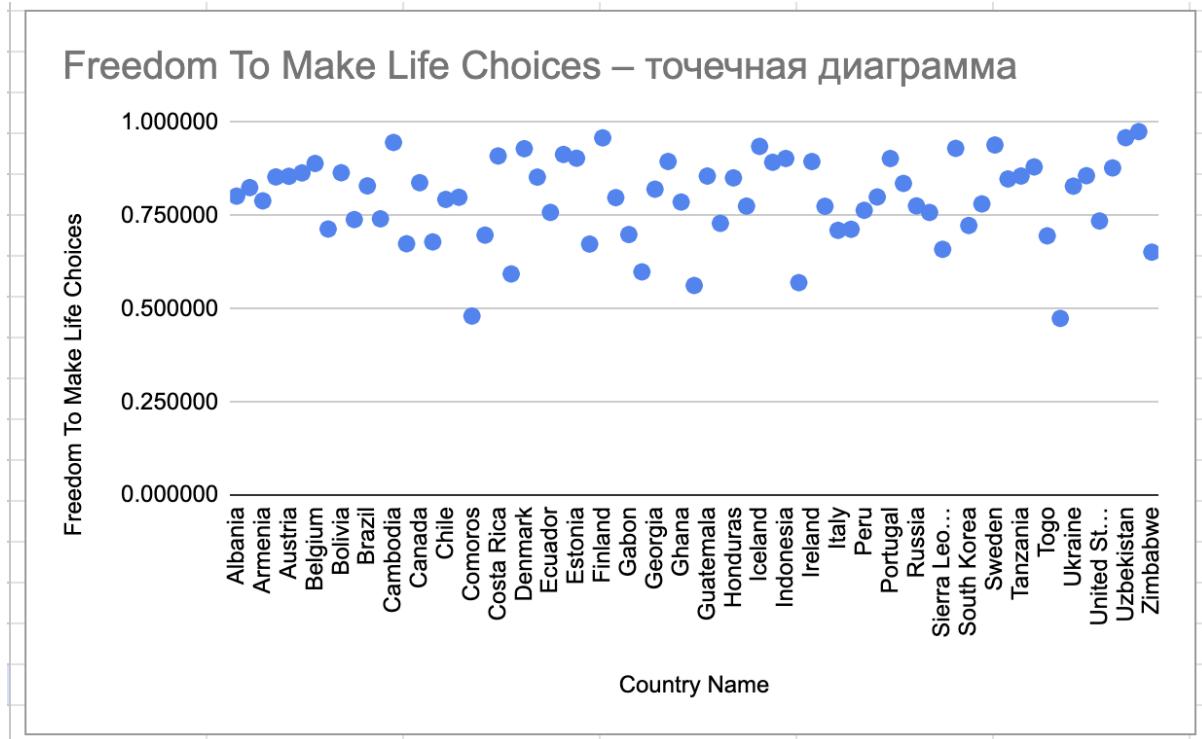
2. Z преобразование

Country Name	Z преобразован ие
Tunisia	-2.852400
Comoros	-2.795848
Greece	-2.067185
Iran	-1.999232
Croatia	-1.792347
Gambia	-1.742574
Zimbabwe	-1.272515
Sierra Leone	-1.204900
Ethiopia	-1.079946
Cameroon	-1.072412
Chad	-1.028221
Togo	-0.883889
Congo (Brazzavil)	-0.866366
Gabon	-0.852313
Italy	-0.752409
Ivory Coast	-0.727255
Benin	-0.721151
South Korea	-0.637240
Guinea	-0.586128
United States	-0.529190
Botswana	-0.495750
Bulgaria	-0.478366
Senegal	-0.326307
Ecuador	-0.325282
Peru	-0.274812

Можно увидеть, что между Comoros и Greece довольно большая разница, и это может означать, что Tunisia и Comoros являются потенциальными данными-выбросами.

3. Проведём графическое представление исходных данных

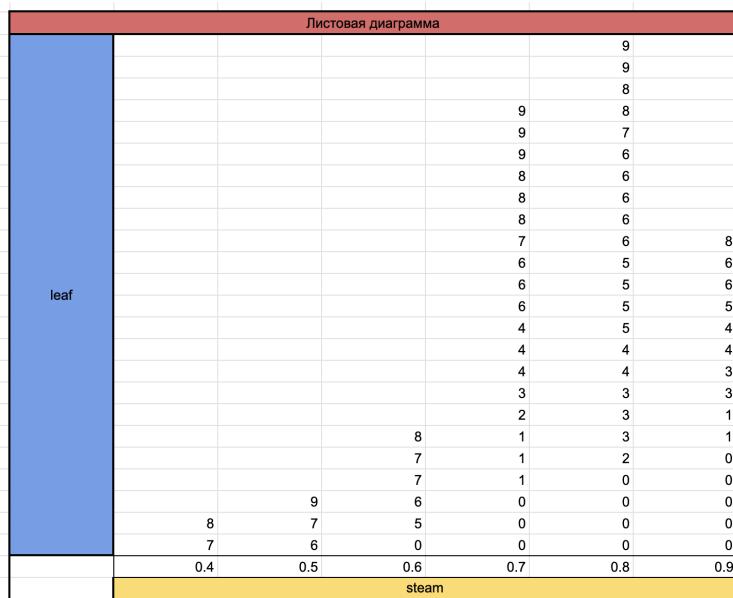
1. Построим точечное распределение данных



По построенному распределению видно, что данные независимы и не сильно упорядочены.

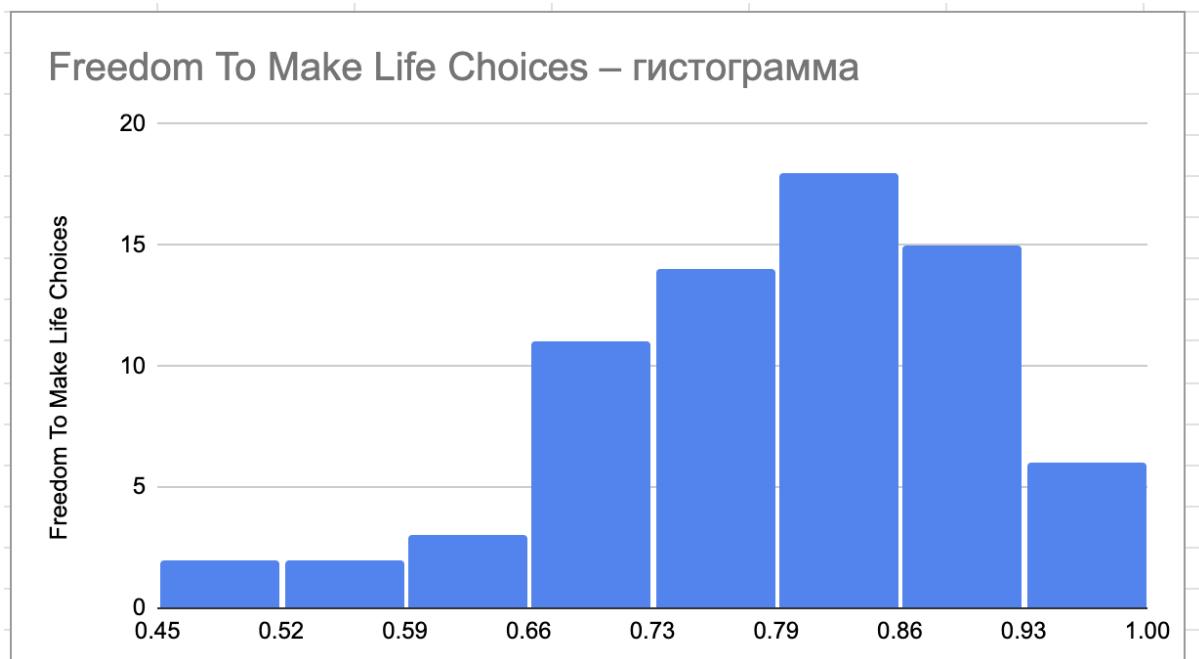
Больше всего свободы в своих решениях ощущают жители Вьетнама. Меньше же всего в Тунисе (интересные и неожиданные наблюдения).

2. Построим листовую диаграмму



Если мысленно провести линию тренда, то ни одно из основных распределений не соответствует полученному графику. Но основная доля значений всё же приходится на стебель 0.8.

3. Построим гистограмму

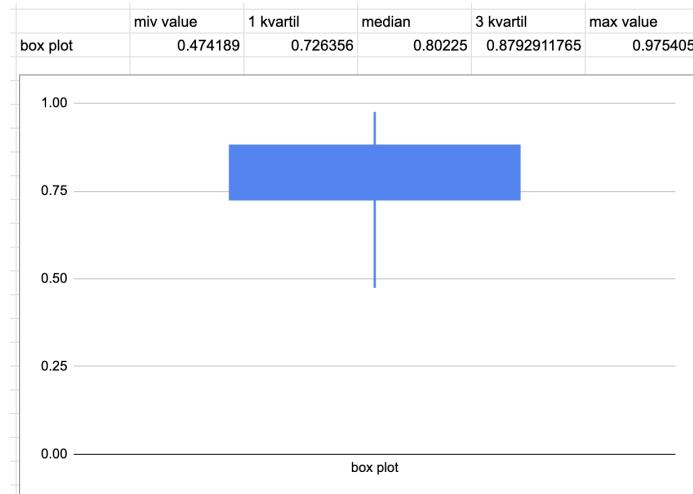


Кол-во стран, в которых люди считают себя наиболее свободными в плане принятия решений = 6. Среди них все страны от Дании до Вьетнама в отсортированном списке . Наименьшее значение = 0.47 у Туниса. Наибольшее количество (18) стран имеют Freedom To Make Life Choices между 0.79 и 0.86.

Точечная и листовая диаграммы чем то напоминает те же графики по столбцу Life Ladder. Для того, чтобы выдвинуть гипотезу о взаимосвязи построим рассеянную диаграмму [в пункте 2](#).

4. Продиагностируем выбросы ряда

1. Построим ящиковую диаграмму



Можно увидеть, что в нижней части есть выбросы так, как прямоугольник смещена вверх. Вполне вероятно, что те данные о 2 странах, выявленные при анализе Z преобразования, являются выбросами.

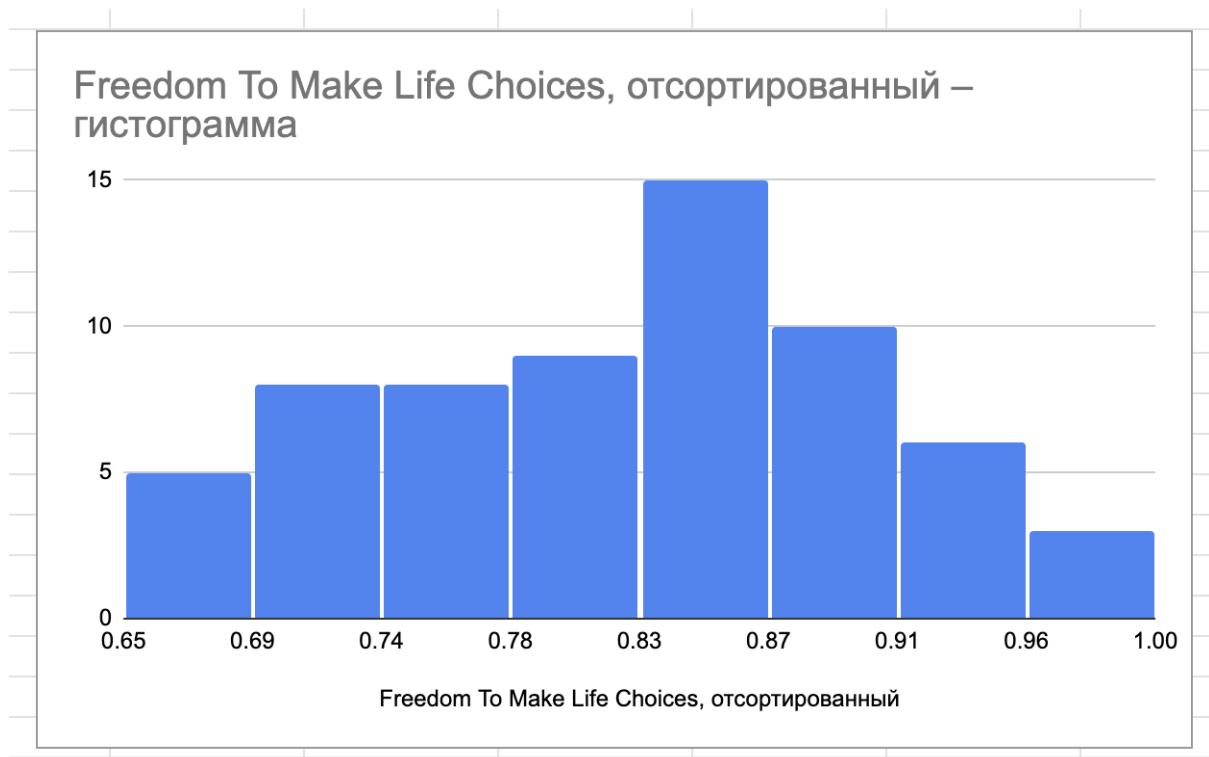
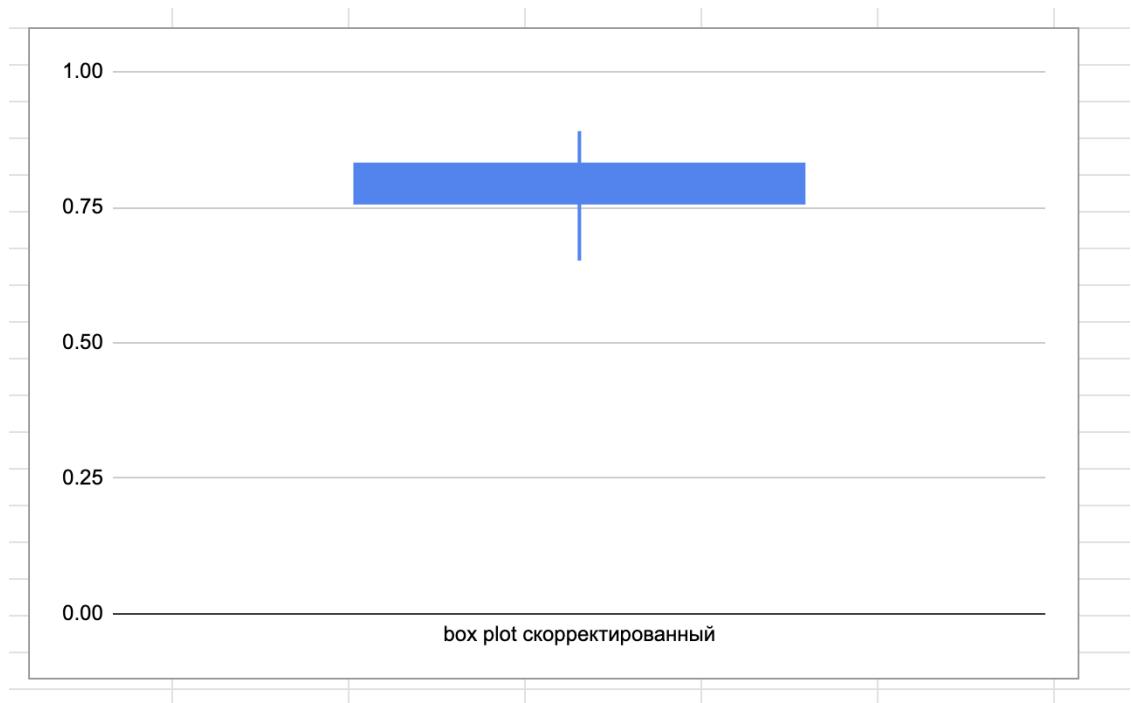
2. Проверим правило 3 сигм

правило 3 сигм		
mean - 3сигма	mean	mean + 3 сигма
0.45758	0.79519	1.13281

=> оно верно, но странно то, что правая граница смещена больше чем левая.

5. Корректировка данных

Удалим первые 6 данных из отсортированного исходного столбца и стране Togo и посмотрим, что получится



Ящичковая диаграмма получилась отличной, и гистограмма Freedom To Make Life Choices тоже почти получилась отличной, так как почти напоминает график нормального распределения.

1.6. Столбец Generosity (в таблице лист “предварительный анализ Generosity”)

1. Проведём её характеристики, вычислив моду, медиану, среднее значение, коэффициент вариации, квантили и квинтили:

Х-ка Generosity														
mean	mode	mode, округленный до сотых	median	размах вариации	коэффициент вариации	дисперсия	станд откл	квартиль 1	квартиль 2	квартиль 3	IQR	квинтиль 1	квинтиль	
0.03341	nan	0.14000	-0.17186	0.83525	522.2701557	0.03044554361	0.1744865141	-0.080116205	0.00342142	0.1391024885	0.2192186935	-0.1294170646	-0.026862	

В среднем ответ на вопрос о пожертвовании низок (0.03341), а это означает, что люди, проживающие в странах с таким показателем редко жертвовали средствами в опрашиваемый месяц. Но показатель вообще некорректный на мой взгляд потому, что люди в целом иногда могут жертвовать, но конкретно в рассматриваемом месяце не жертвовали. Мне кажется было бы корректнее взять не месяц, а полгода хотя бы.

Значение наиболее встречаемого числа в стандартном ряде (отсортированный ряд Generosity без округления) не удалось определить, поэтому я решил округлить каждое число до сотых. И оно оказалось равным 0.14, что странно по сравнению со средним значением. Но оно также подтверждает вывод, сделанный при интерпретации среднего значения.

Значение медианы (0.00342) также подтверждает вывод, сделанный при анализе среднего.

Размах вариации получился немаленьким (0.83525), что говорит о том, что можно вполне встретить выбросы в данных, а также, что данные неоднородны (данные не сфокусированы на маленьком промежутке. Если бы 0.2 размах был, то можно было бы сказать спокойно, что размах однородный).

Данные имеют крайне низкую вариативность (522.2701557), и это означает, что изначальные сильно разбросаны. А также это означает, что интерпретация по среднему значению некорректная. Но возможно такое значение получилось из-за того, что в наборе присутствуют данные, которые далеко находятся по сравнению с основной массой.

IQR получился немаленьким (0.2192186935), что подтверждает выводы, сделанные при анализе коэффициента вариации.

2. Z преобразование

Country Name	Z преобразование
Greece	-2.004936
Georgia	-1.640594
Tunisia	-1.514509
Botswana	-1.421484
Croatia	-1.410287
Poland	-1.375445
Vietnam	-1.217265
Peru	-1.209116
Romania	-1.177048
Gabon	-1.133468
Colombia	-1.118805
Armenia	-1.075927
Bulgaria	-1.023701
Portugal	-0.976979
Argentina	-0.922223
El Salvador	-0.834660
Dominican Repu	-0.671847
Bolivia	-0.651361
Ecuador	-0.649891
Russia	-0.595157
Zimbabwe	-0.589860
Albania	-0.569651
Brazil	-0.539296
Guatemala	-0.519637
Bangladesh	-0.505399
Uruguay	-0.487588
Costa Rica	-0.460845

На представленном кадре можно увидеть, что разница показателя Греции и Грузии велика по сравнению с Грузией и Тунисом. Но если пойти дальше, то можно увидеть также непостоянность разностей данных в разрядах стран от Грузии до Польши, включительно. Например, разница показателей между Туниса и Ботсваны заключается в сотых долях, а у Грузии и Туниса в десятых. Поэтому эти данные потенциально являются выбросными.

United States	0.900768
Iran	1.021534
Chad	1.076028
Canada	1.078208
Iceland	1.080320
Denmark	1.092956
Sweden	1.149670
Thailand	1.534069
Uzbekistan	1.579156
United Kingdom	1.581698
Ethiopia	1.878861
Gambia	1.895816
Ukraine	2.259046
Indonesia	2.781957

Это конечный отрывок данных после Z преобразования. Здесь можно явно по той же логике, что и в начале анализа Z преобразования считать показатели стран от Тайланда до Индонезии, включительно, выбросными.

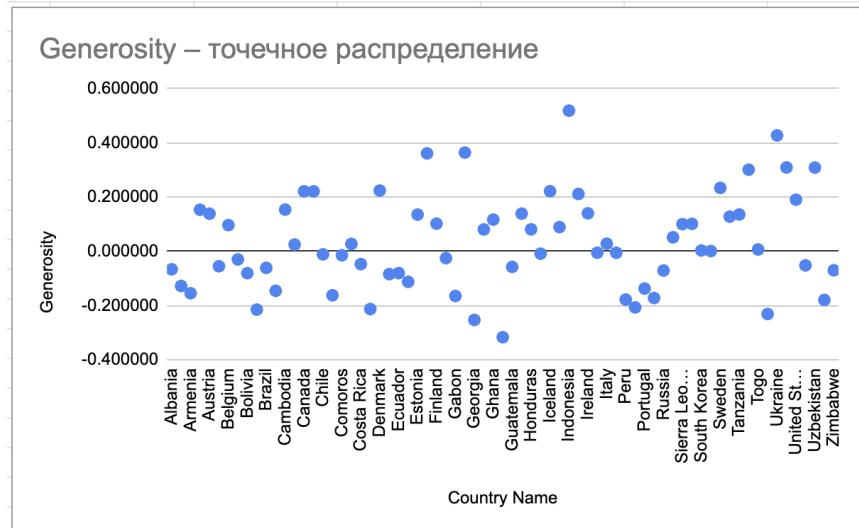
Попробуем удалить выдвинутые данные и заново для них вычислить среднее, медиану, коэффициент вариации, стандартное отклонение, 1 и 3 квартили, межквартильную разницу.

mean	median	коэффициент вариации	станд откл	квартиль 1	квартиль 3	IQR
0.020944	0.002315	578.4675132	0.1211568301	-0.06863177775	0.1254038498	0.1940356275

Ситуация чуть лучше стала, но всё равно наблюдаются выбросы. В любом случае я думаю оставшиеся данные не стоит удалять так, как сильно мало показателей останется.

3. Проведём графическое представление исходных данных

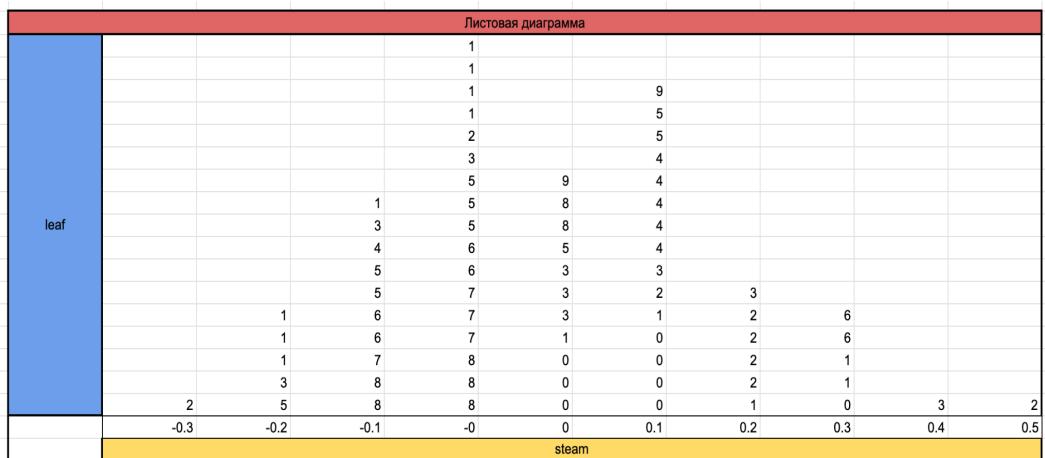
1. Построим точечное распределение данных



По построенному распределению видно, что данные независимы, не сильно упорядочены и не сильно сгруппированы так, как видно, что размах большой.

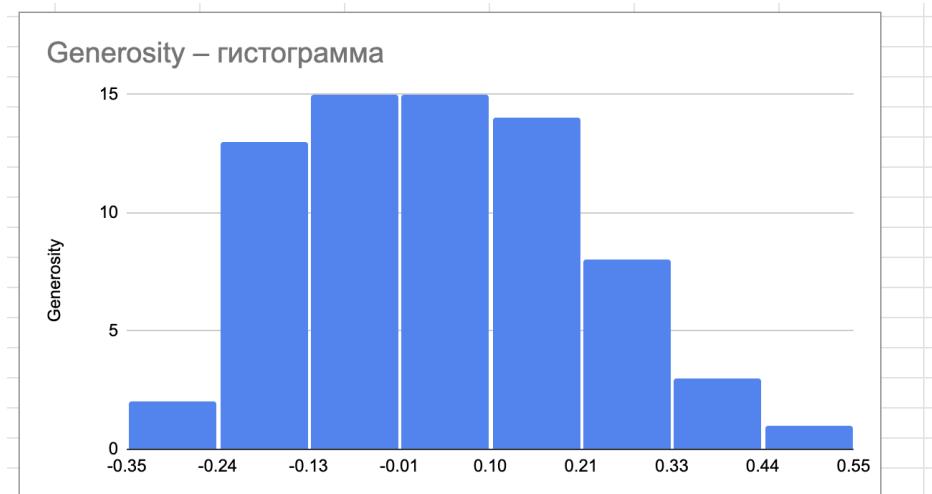
Больше всего оказывают поддержку другим Индонезия, меньше же всего Греция.

2. Построим листовую диаграмму



Если мысленно провести линию тренда, то ни одно из основных распределений не соответствует полученному графику. Но основная доля значений всё же приходится на стебель -0.0, а не на 0.1, как получилось при х-ке ряда.

3. Построим гистограмму

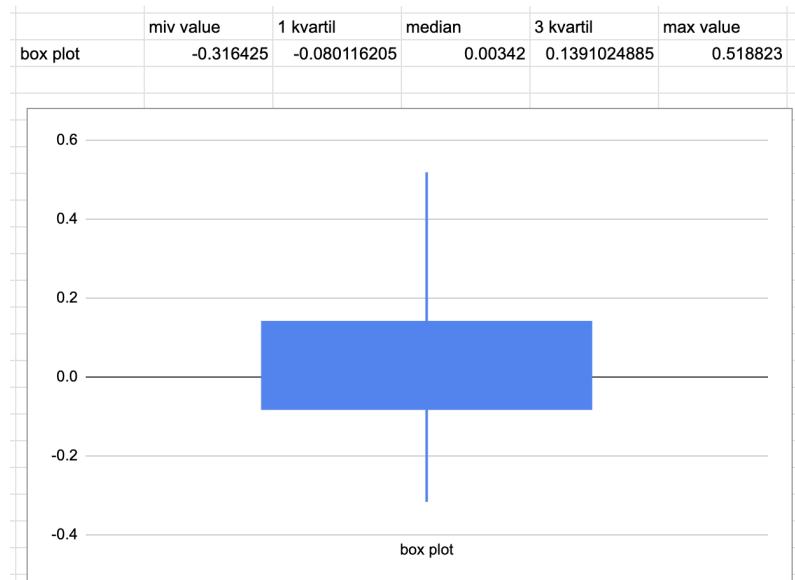


Только единственная страна находится в списке самых жертвующих. И это Индонезия. Наименьшее значение = 0.47 у Туниса. Наибольшее количество (30) стран имеют Generosity между -0.13 и 0.1

Все диаграммы не напоминают диаграммы из Life Ladder. Но небольшие перекрестья всё же наблюдаются. Для того, чтобы выдвинуть гипотезу о взаимосвязи построим рассеянную диаграмму [в пункте 2](#).

4. Продиагностируем выбросы ряда

1. Построим ящиковую диаграмму



Можно увидеть, что в верхней части есть выбросы так, как прямоугольник смещен вниз. Но диаграмма не демонстрирует выбросы в нижней части данных, поэтому по ней нельзя точно сказать про все выбросы.

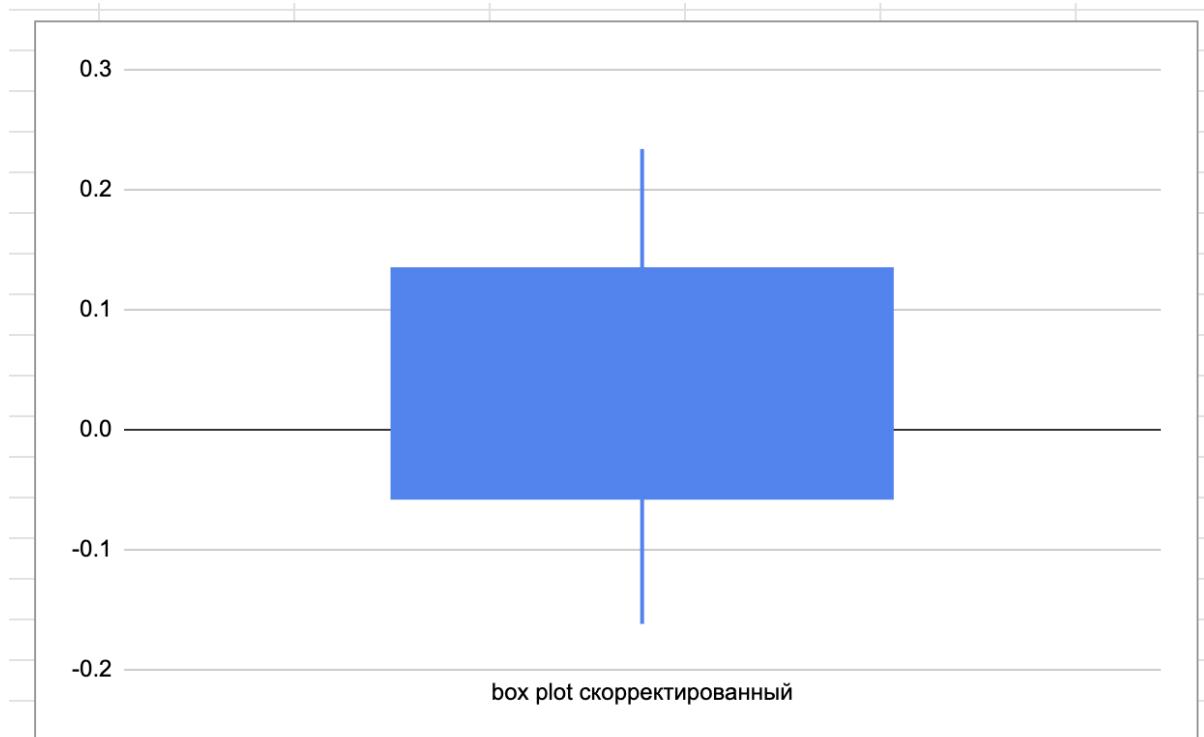
2. Проверим правило 3 сигм

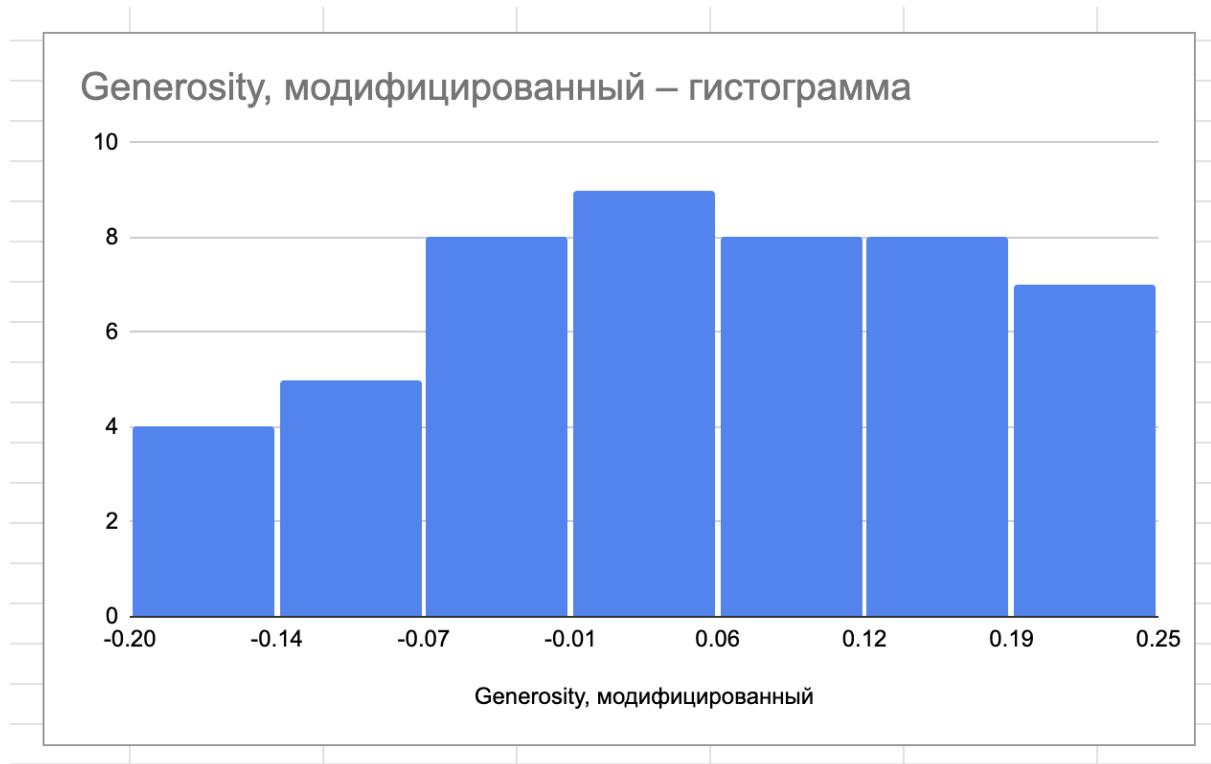
правило 3 сигм		
mean - 3сигма	mean	mean + 3 сигма
-0.49005	0.03341	0.55687

=> оно верно, но при этом видно, что выбросов больше должно быть в нижней части потому, что смещение направлено больше в левую часть чем в правую.

5. Корректировка данных

Удалим первые 19 данных и посмотрим, что получится, а также данные, начиная с России, заканчивая Гватемалой, для нормировки данных по отсортированному исходному столбцу





Ящичковая диаграмма получилась хорошей, но не отличной потому, что квартили от медианы находятся не на близком расстоянии.

А вот гистограмма Generosity почти имеет нормальное распределение. Было бы отлично, если бы в промежутки 0.12-0.19 и 0.19-0.25 были бы по количеству симметричны первым 2 промежуткам. Но это невозможно сделать, так как тогда гистограмма будет всё больше не похожа на нормальную.

1.7. Столбец Perceptions Of Corruption (в таблице лист “предварительный анализ Perceptions Of Corruption”)

1. Проведём её характеристики, вычислив моду, медиану, среднее значение, коэффициент вариации, квантили и квинтили:

X-ка Perceptions Of Corruption													
mean	mode	mode, округленный до сотых	median	размах вариации	коэффициент вариации	дисперсия	станд откл	квартиль 1	квартиль 2	квартиль 3	IQR	квинтиль 1	кв
0.70703	nan	0.85000	0.75975	0.75142	26.19780826	0.03430925668	0.1852275808	0.626217067	0.759754062	0.8487873675	0.2225703005	0.5834574222	0.7

В среднем ответ на вопрос о коррупции высок (0.70703), а это означает, что люди, проживающие в странах с таким показателем считают или наблюдают коррупцию в стране, что сказывается на ценах в стране вероятнее всего и негативно влияет на удовлетворение людьми своими жизнями.

Значение наиболее встречаемого числа в стандартном ряде (отсортированный ряд Generosity без округления) не удалось определить, поэтому я решил округлить

каждое число до сотых. И оно оказалось равным 0.85, что не сильно близко к среднему значению. Но оно также подтверждает вывод, сделанный при интерпретации среднего значения.

Значение медианы (0.75975) также подтверждает вывод, сделанный при анализе среднего.

Размах вариации получился немаленьким (0.75142), что говорит о том, что можно вполне встретить выбросы в данных, а также, что данные неоднородны (данные не сфокусированы на маленьком промежутке. Если бы 0.2 размах был, то можно было бы сказать спокойно, что размах однородный).

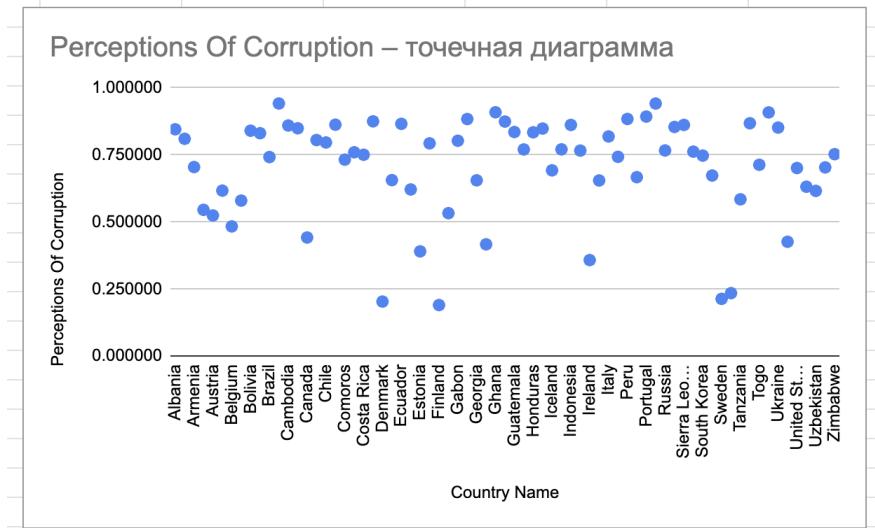
Данные имеют крайне низкую вариативность (26.19780826), и это означает, что выборка слабо устойчива. А также это означает, что интерпретация по среднему значению корректная (является некорректной, когда > 33 коэффициент). IQR получился немаленьким (0.2225703005), что подтверждает выводы, сделанные при анализе коэффициента вариации.

2. Z преобразование

Country Name	Z преобразован ие
Finland	-2.790232
Denmark	-2.720406
Sweden	-2.665900
Switzerland	-2.550455
Ireland	-1.885369
Estonia	-1.709971
Germany	-1.568111
United Kingdom	-1.516944
Canada	-1.430859
Belgium	-1.207435
Austria	-0.987016
France	-0.940777
Australia	-0.873615
Benin	-0.687887
Tanzania	-0.661983
Uzbekistan	-0.492314
Bangladesh	-0.486765
El Salvador	-0.463955
Uruguay	-0.408675
Israel	-0.282549
Georgia	-0.279994
Dominican Repu	-0.275652
Poland	-0.216004
Spain	-0.183141
Iceland	-0.078825
United States	-0.031891

Можно увидеть, что между Швейцарией и Ирландией довольно большая разница, и это может означать, что данные от Финляндии до Швейцарии, включительно, являются выбросами.

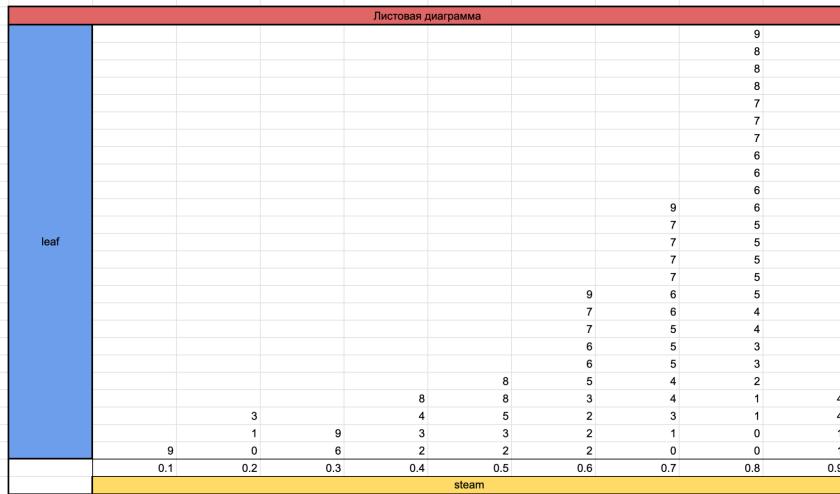
3. Проведём графическое представление исходных данных
1. Построим точечное распределение данных



По построенному распределению видно, что данные независимы и не сильно упорядочены.

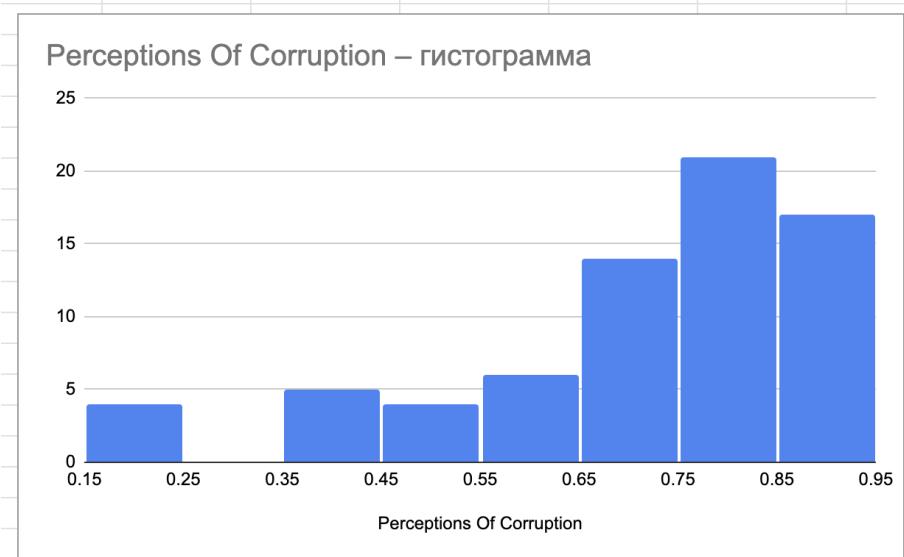
Больше всего люди видят коррупцию в Болгарии, меньше же всего в Финляндии.

2. Построим листовую диаграмму



Если мысленно провести линию тренда, то ни одно из основных распределений не соответствует полученному графику. Но основная доля значений всё же приходится на стебель 0.8.

3. Построим гистограмму

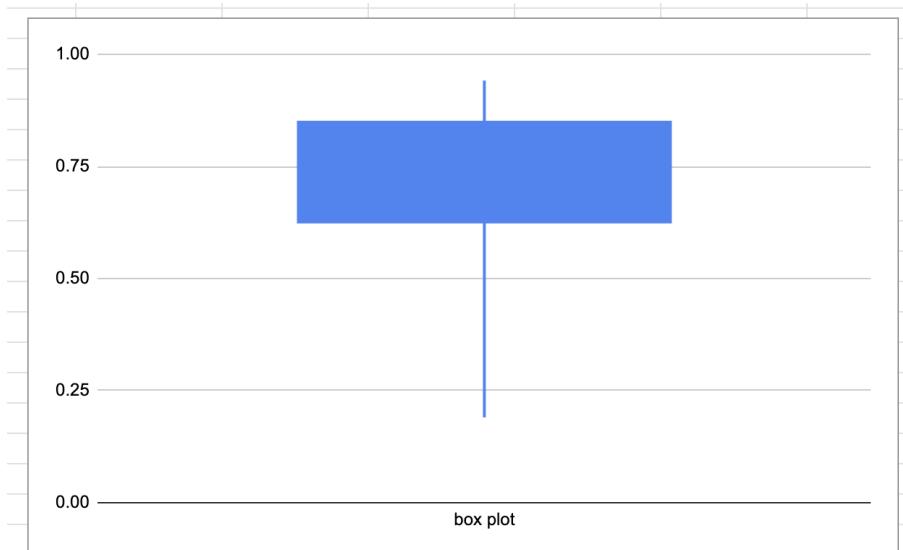


В 17 странах люди наблюдают высокую коррупцию в разных сферах страны. Среди них все страны от Венгрии до Болгарии, включительно. Наименьшее значение = 0.19 у Финляндии и по идеи это должно сказываться лучшем удовлетворением жизнью по сравнению со странами, в которых высокая коррупция. Наибольшее количество (21) стран имеют Perceptions Of Corruption между 0.75 и 0.85.

Точечная и листовая диаграммы напоминают диаграммы из Life Ladder. Для того, чтобы проверить взаимосвязь между данными построим рассеянную диаграмму [в пункте 2](#).

4. Продиагностируем выбросы ряда

1. Построим ящиковую диаграмму



Можно увидеть, что в нижней части есть выбросы так, как прямоугольник смещена вверх. Вполне вероятно, что те данные, начиная с Финляндии, выявленные при анализе Z преобразования, являются выбросами.

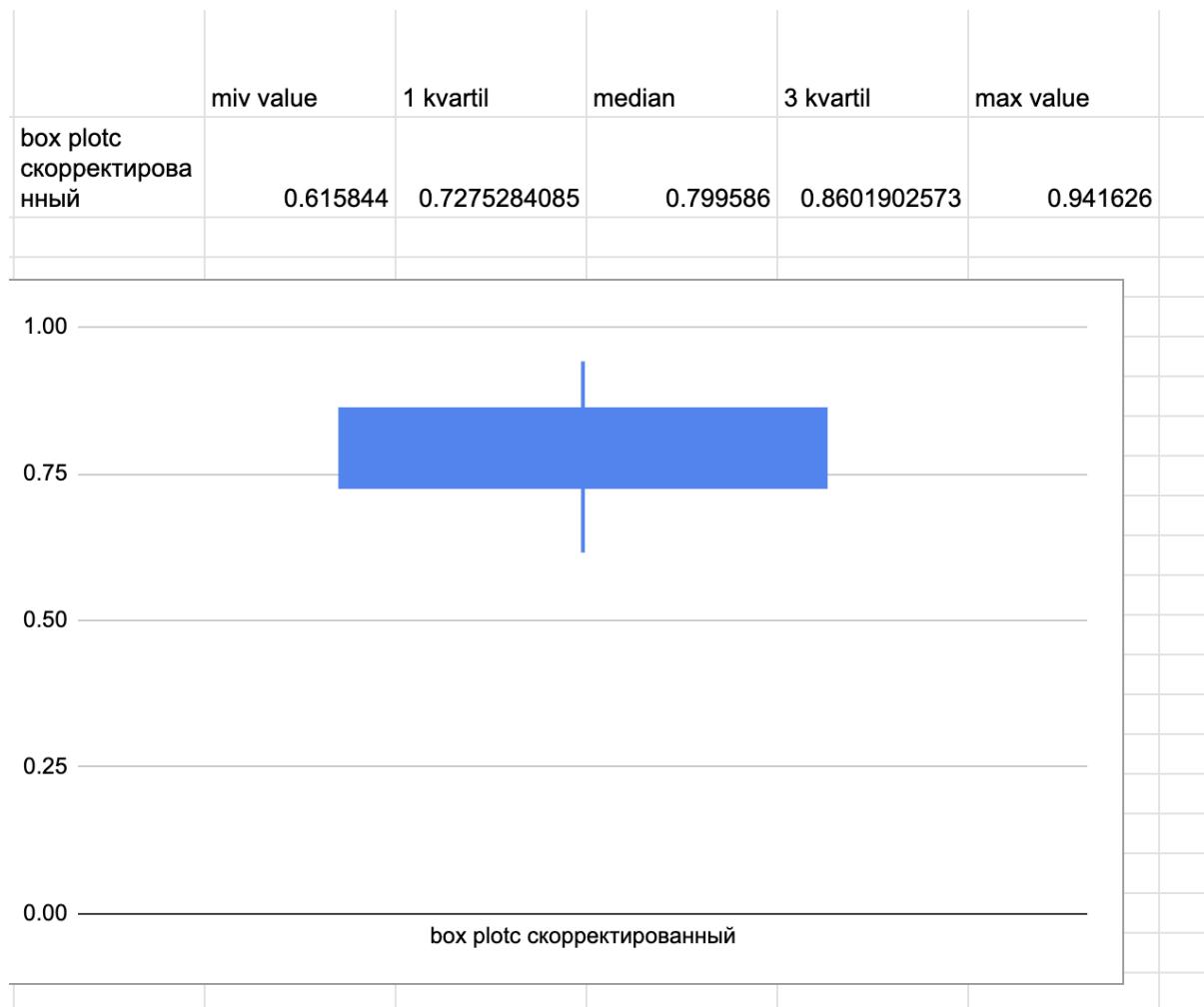
2. Проверим правило 3 сигм

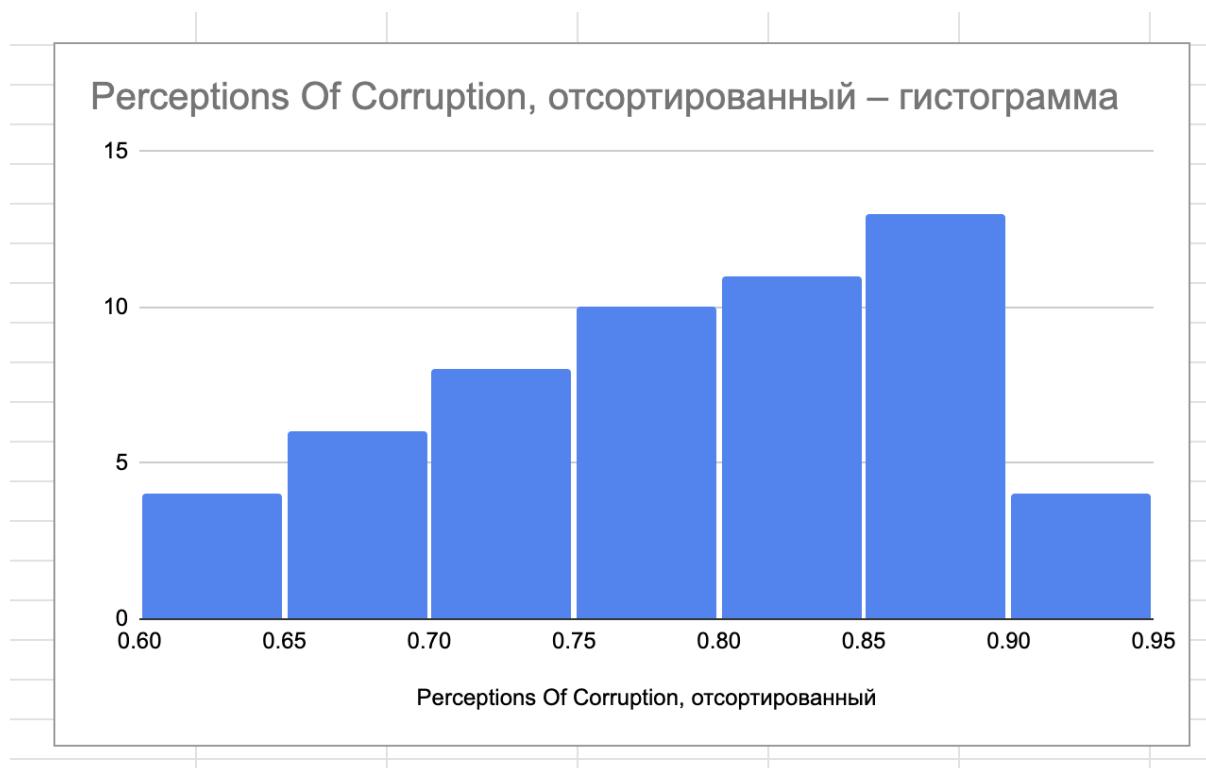
правило 3 сигм		
mean - 3сигма	mean	mean + 3 сигма
0.15135	0.70703	1.26272

=> оно верно, но всё же выбросы имеются потому, что правая граница сильно сдвинута от максимума и стандартное отклонение сильно большое, за счёт чего образуются такие большие интервалы.

5. Корректировка данных

Удалим первые 17 данных и посмотрим, что получится

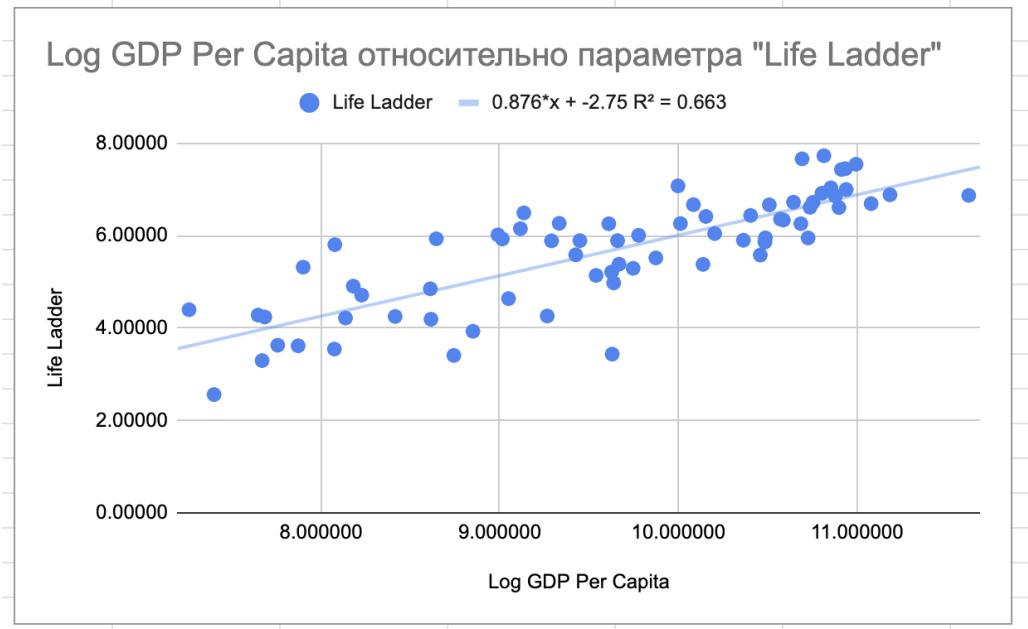




Ящичковая диаграмма получилась отличной, а вот гистограмма Perceptions Of Corruption немного скосившаяся всё равно, но зато уже напоминает нормальное распределение.

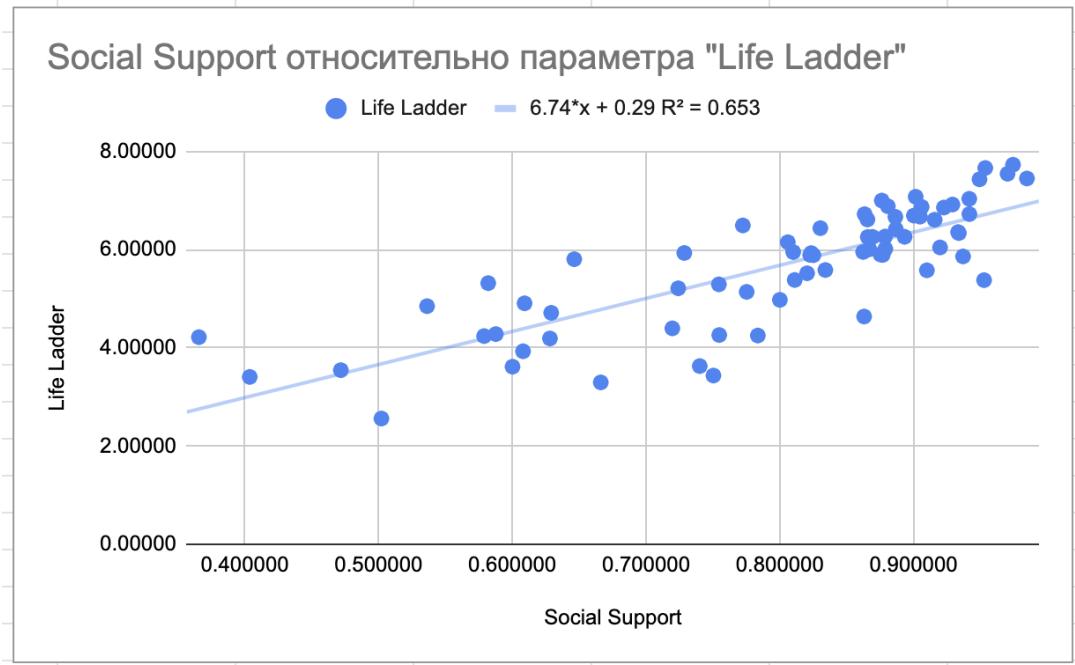
3. Построение диаграмм рассеяния между столбцами, у которых похожи графики точечные

1. Между Life Ladder и Log GDP Per Capita (всю информацию можно увидеть в листе "рассеянная диаграмма между Life Ladder и Log GDP..")



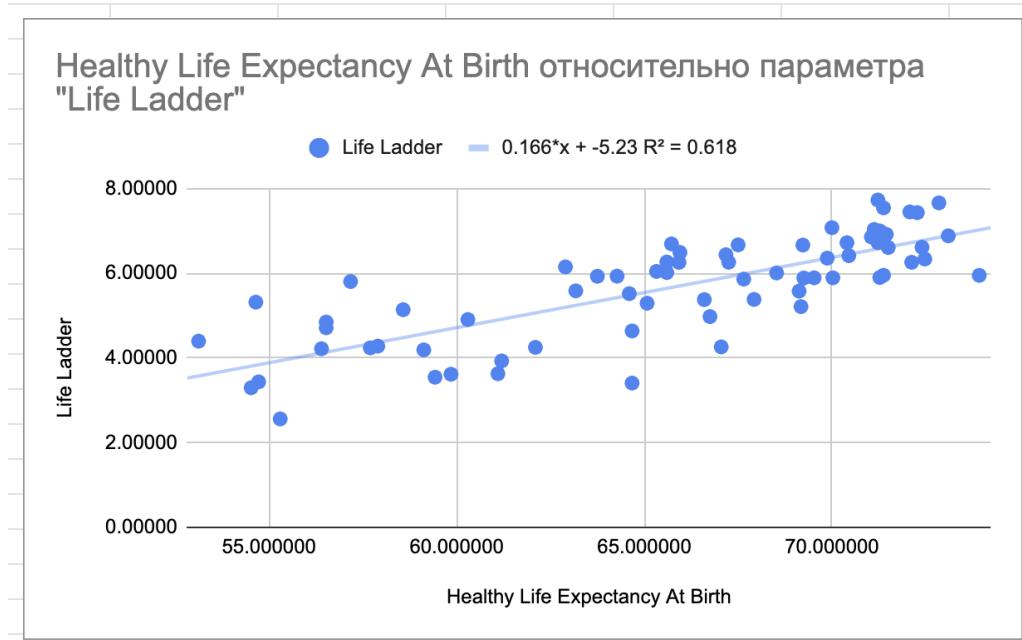
=> 66,3% дисперсии Life Ladder можно объяснить логарифмированным ППС стран. То есть присутствует значительная связь между рассмотренными параметрами. А значит можно предполагать гипотезы с этими параметрами

2. Между Life Ladder и Social Support (всю информацию можно увидеть в листе "рассеянная диаграмма между Life Ladder и Social...")



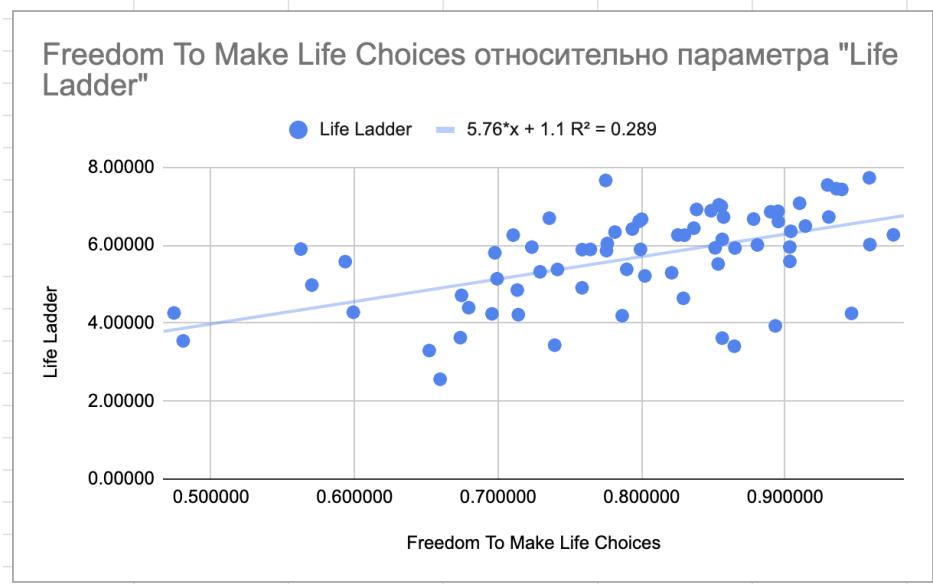
=> 65,3% дисперсии Life Ladder можно объяснить ощущением людей наличия близких и друзей, готовых помочь в трудные моменты. То есть присутствует значительная связь между рассмотренными параметрами. А значит можно предполагать гипотезы с этими параметрами

3. Между Life Ladder и Healthy Life Expectancy At Birth (всю информацию можно увидеть в листе "рассеянная диаграмма между Life Ladder и Healthy..")



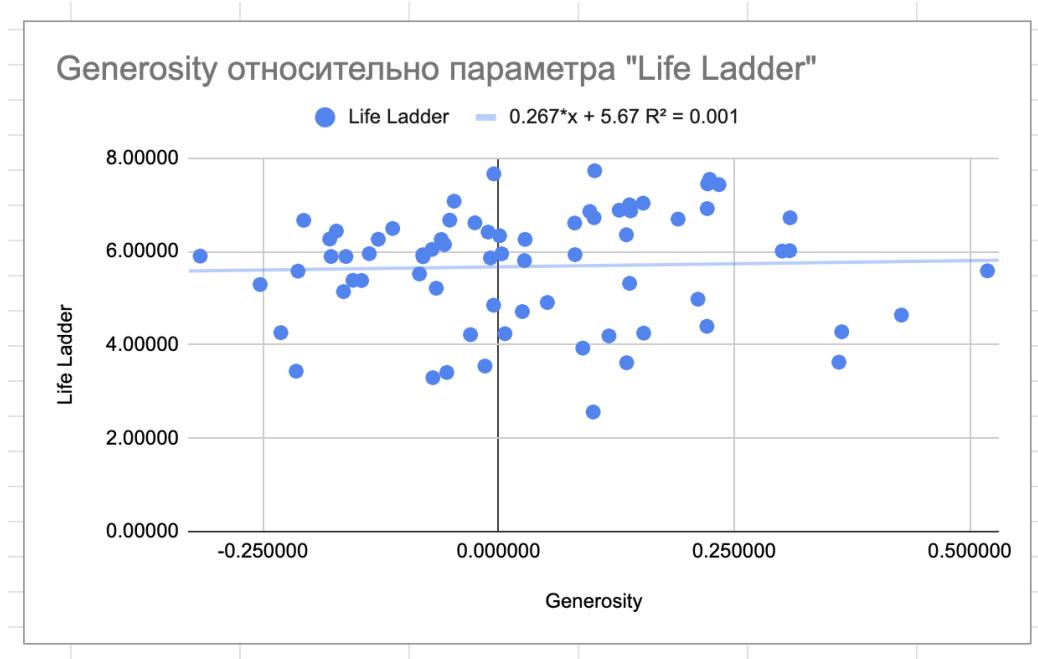
=> 61,8% дисперсии Life Ladder можно объяснить средним возрастом людей, проживающих в стране. То есть присутствует значительная связь между рассмотренными параметрами. А значит можно предполагать гипотезы с этими параметрами

4. Между Life Ladder и Freedom To Make Life Choices (всю информацию можно увидеть в листе "рассеянная диаграмма между Life Ladder и Freed.."")



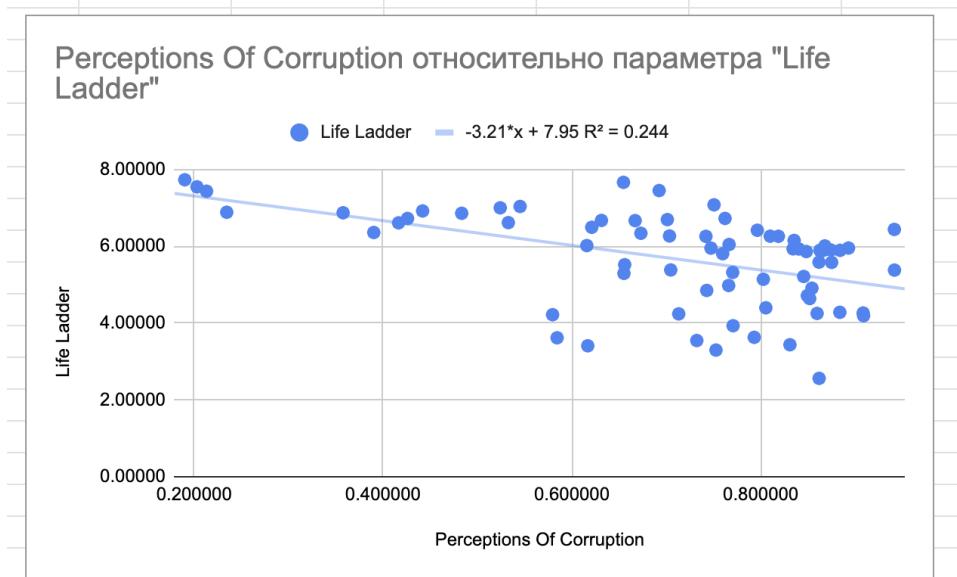
=> 28,9% дисперсии Life Ladder можно объяснить возможностью свободно принимать решения в стране. То есть присутствует умеренная связь между рассмотренными параметрами. А значит не стоит предполагать гипотезы между данными величинами.

5. Между Life Ladder и Generosity (всю информацию можно увидеть в листе "рассеянная диаграмма между Life Ladder и Gener..")



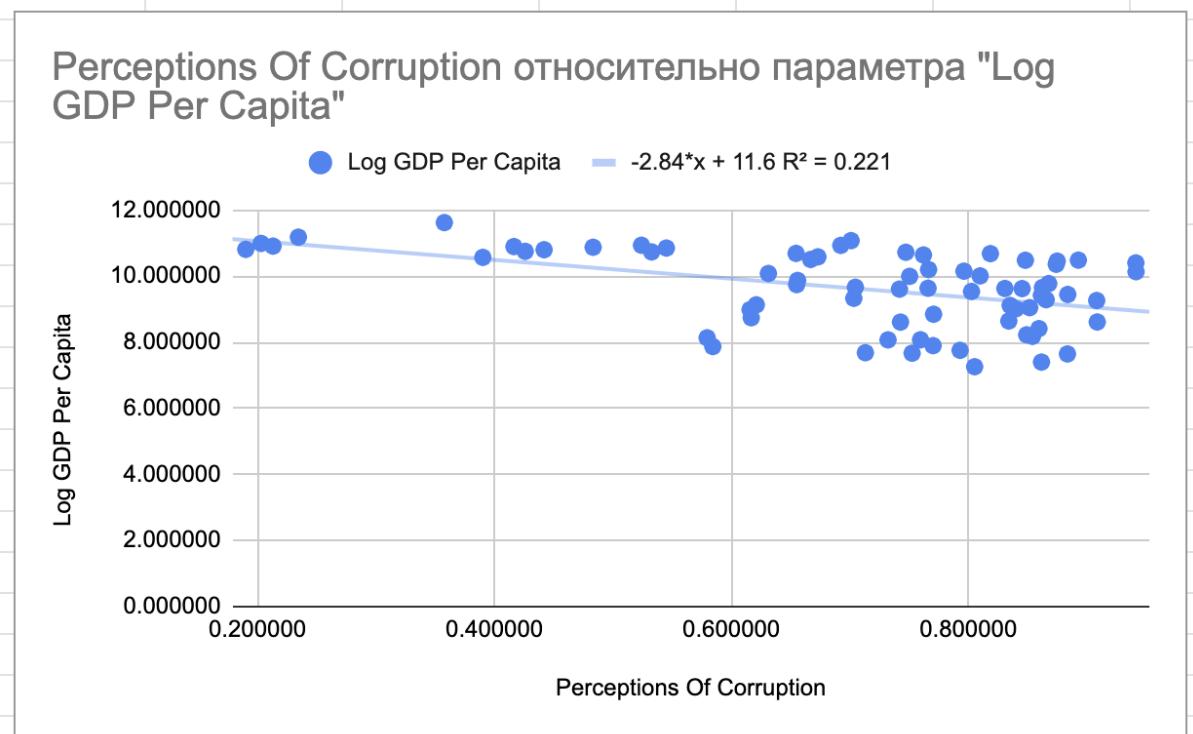
=> 1% дисперсии Life Ladder можно объяснить тем, что люди жертвуют материально на что-то или кому-то. То есть присутствует слабая связь между рассмотренными параметрами. А значит нельзя предполагать гипотезы между данными величинами потому, что получится некорректный вывод.

6. Между Life Ladder и Perceptions Of Corruption (всю информацию можно увидеть в листе "рассеянная диаграмма между Life Ladder и Perce..")



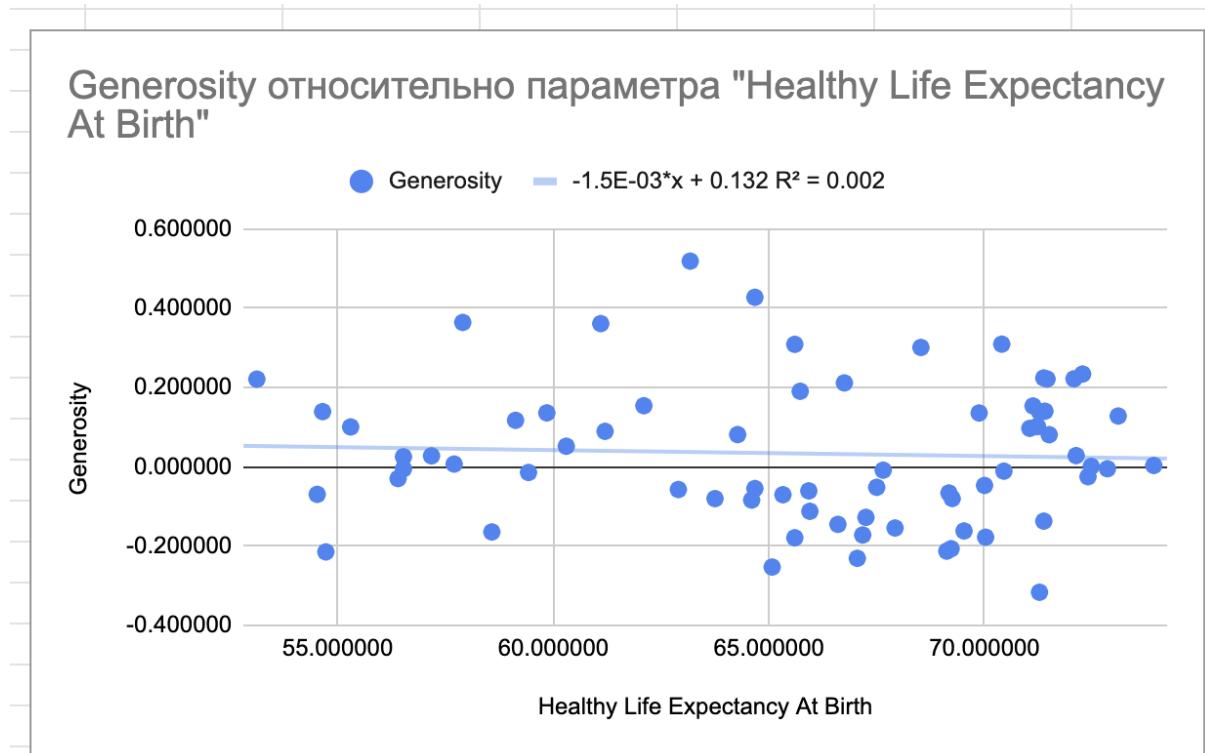
=> 24,4% дисперсии Life Ladder можно объяснить тем, что, чем больше люди видят коррупцию в стране и бизнесе, тем меньше удовлетворенность жизнью. То есть присутствует слабая обратная связь между рассмотренными параметрами. А значит нельзя предполагать гипотезы между данными величинами потому, что получится некорректный вывод. И тут же можно опровергнуть 6 гипотезу, ссылаясь на этот фактор.

7. Между Log GDP Per Capita и Perceptions Of Corruption (всю информацию можно увидеть в листе "рассеянная диаграмма между Log GDP и Perce..")



=> 22,1% дисперсии Log GDP Per Capita можно объяснить тем, что, чем больше люди видят коррупцию в стране и бизнесе, тем меньше ППС страны становится, что странно так, как должно быть наоборот. То есть присутствует слабая обратная связь между рассмотренными параметрами. А значит нельзя предполагать гипотезы между данными величинами потому, что получится некорректный вывод. И тут же можно опровергнуть 5 гипотезу, ссылаясь на этот фактор.

8. Между Healthy Life Expectancy At Birth и Generosity (всю информацию можно увидеть в листе "рассеянная диаграмма между Healthy и Gen..")



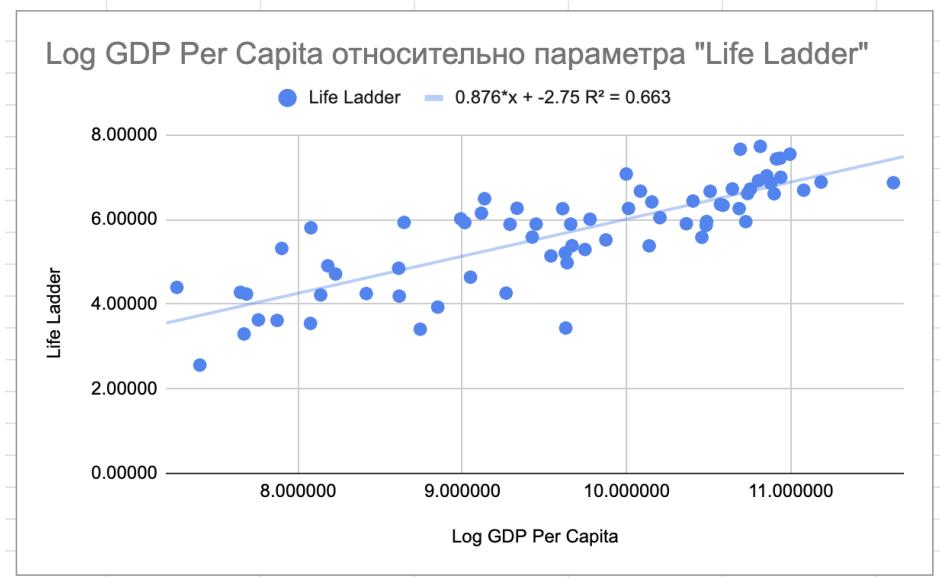
=> 2% дисперсии Generosity можно объяснить тем, что некоторые люди живут в странах с более продолжительной жизнью. То есть присутствует слабая связь между рассмотренными параметрами. А значит нельзя предполагать гипотезы между данными величинами потому, что получится некорректный вывод. Следовательно, 4 гипотеза неверна

Корреляционный анализ

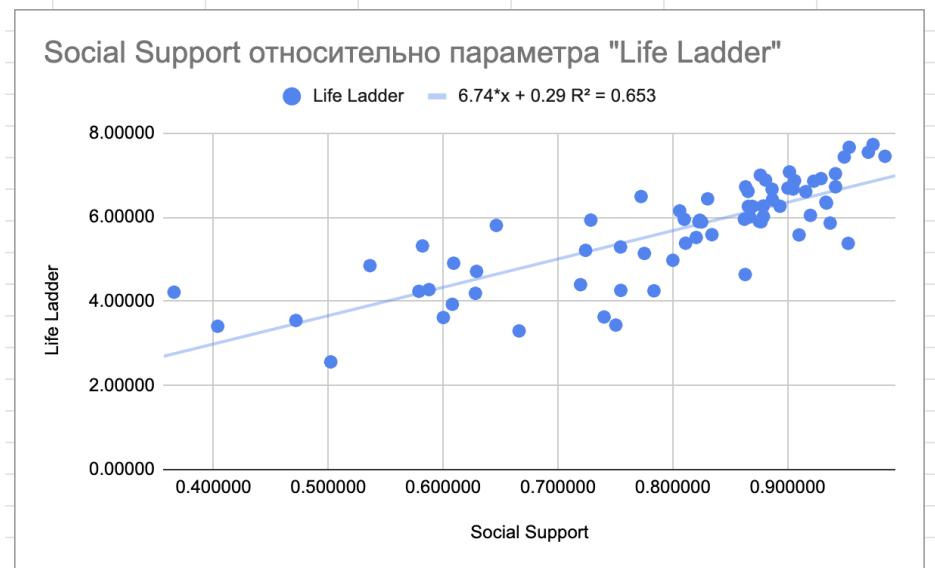
1. Построение полей корреляции для исследования связи между переменными

Поскольку последние 3 гипотезы были опровергнуты, то воспользуемся диаграммами, построенными в пункте [3 предварительного анализа](#). И также добавим диаграммы для оставшихся исследуемых признаков.

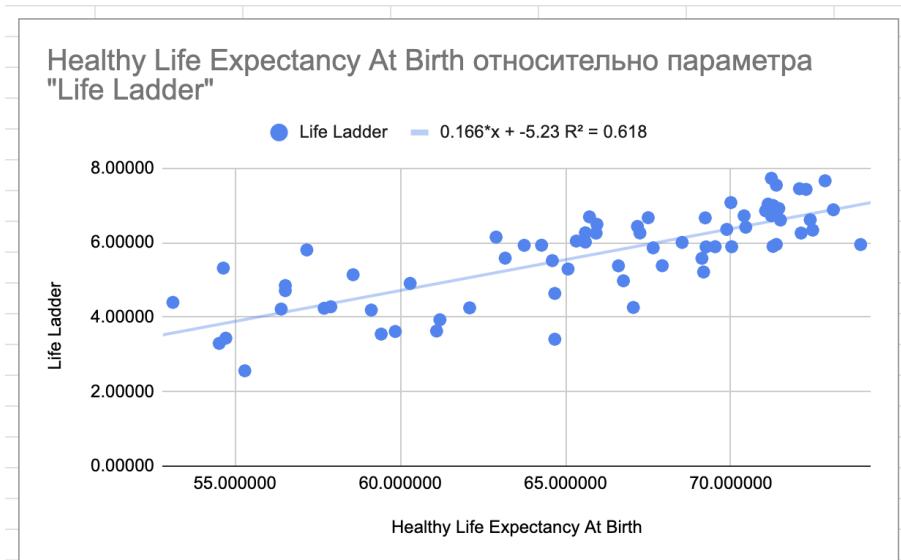
1. Life Ladder и Log GDP Per Capita



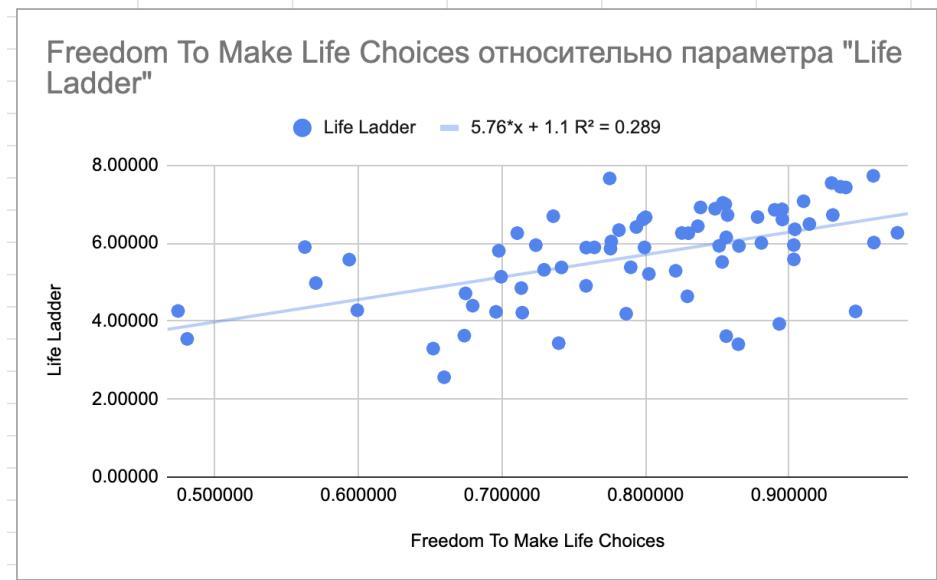
2. Life Ladder и Social Support



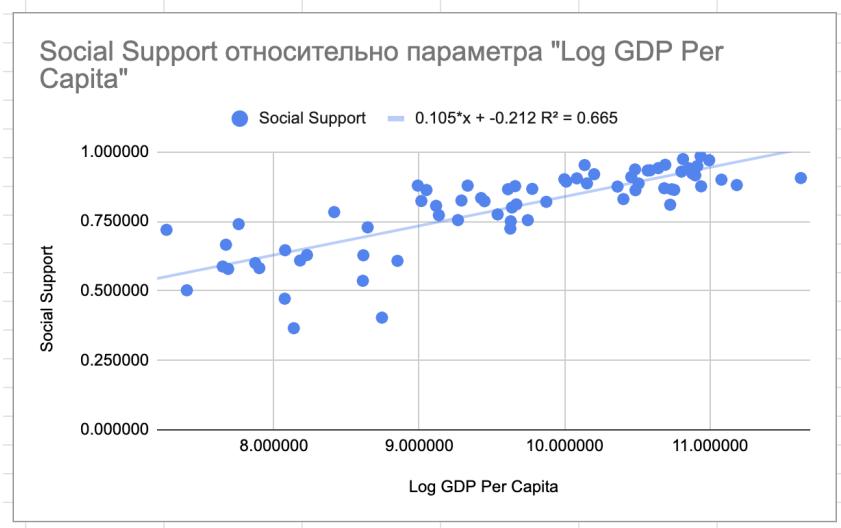
3. Life Ladder и Healthy Life Expectancy At Birth



4. Life Ladder и Freedom To Make Life Choices

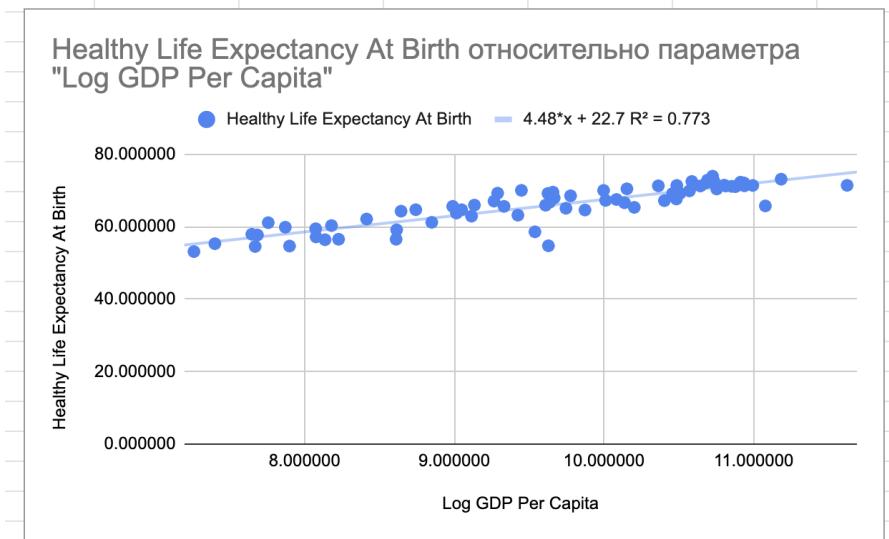


5. Между Log GDP Per Capita и Social Support (в таблице лист "поле Log и Soc")



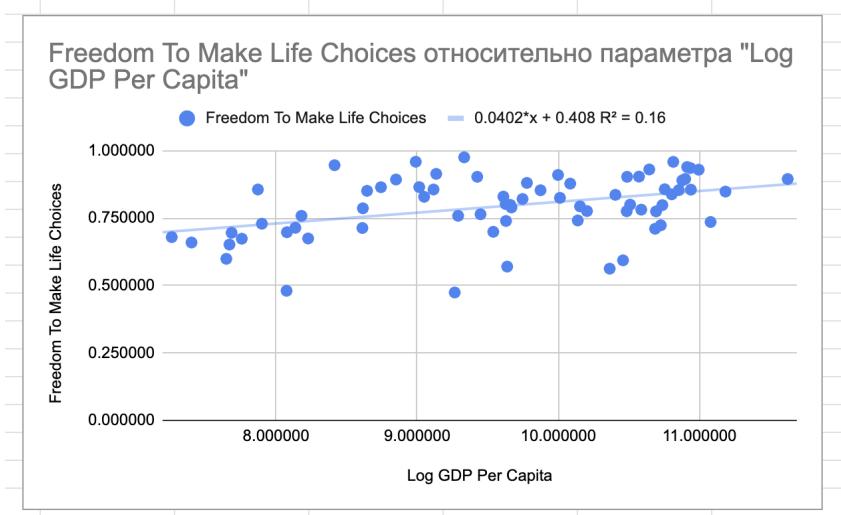
=> 66,5% дисперсии Social Support можно объяснить тем, что люди живут в странах с высоким ППС. То есть присутствует значительная связь между рассмотренными параметрами.

6. Между Log GDP Per Capita и Healthy Life Expectancy At Birth (в таблице лист "поле Log и Healthy...")



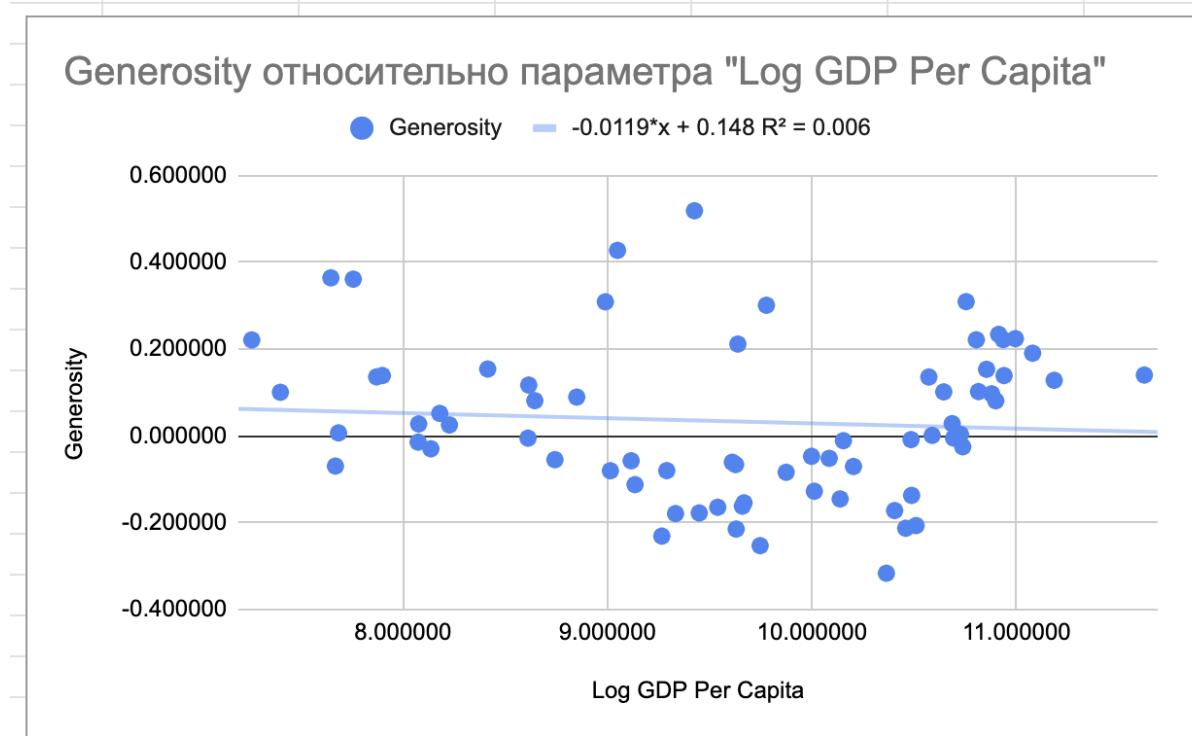
=> 77,3% дисперсии Healthy Life Expectancy At Birth можно объяснить тем, что люди живут в странах с высоким ППС. То есть присутствует сильная связь между рассмотренными параметрами.

7. Между Log GDP Per Capita и Freedom To Make Life Choices (в таблице лист "поле Log и Free...")



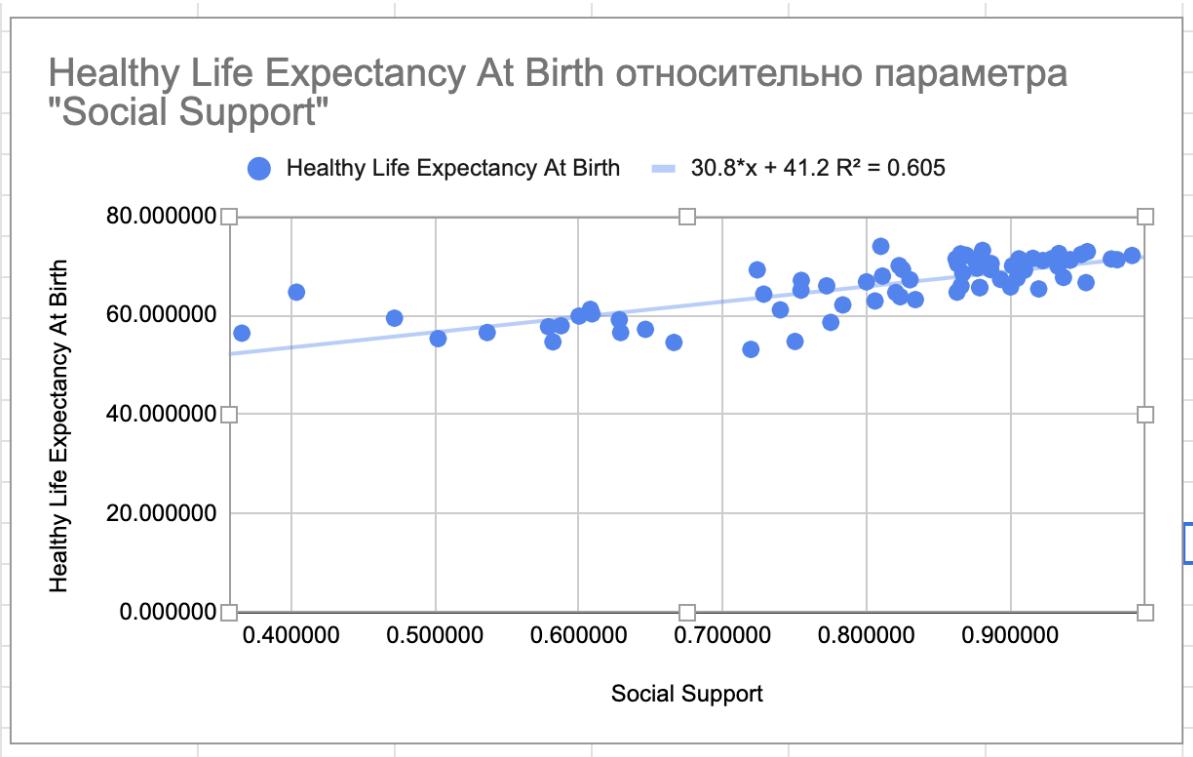
=> 16% дисперсии Freedom To Make Life Choices можно объяснить тем, что люди живут в странах с высоким ППС. То есть присутствует слабая связь между рассмотренными параметрами. А значит в дальнейшем нет смысла рассматривать между ними взаимосвязь.

8. Между Log GDP Per Capita и Generosity (в таблице лист "поле Log и Gen..")



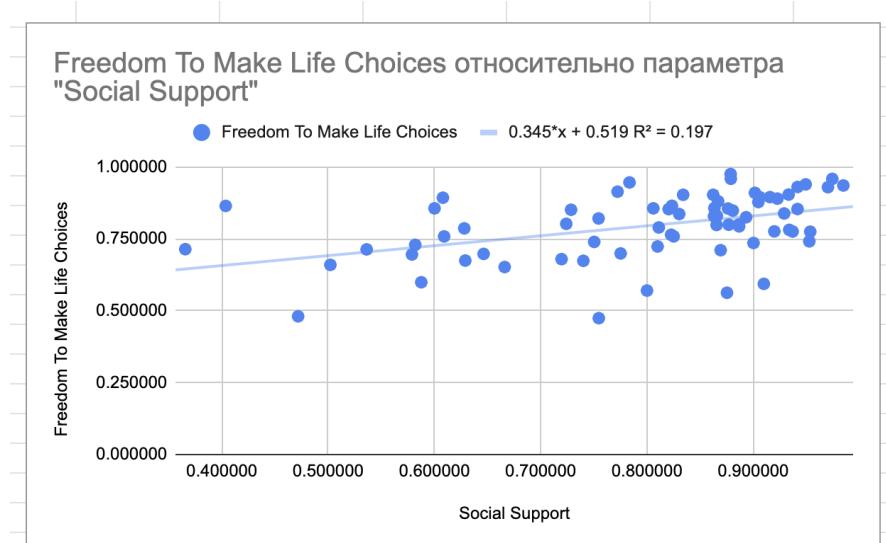
=> 6% дисперсии Generosity можно объяснить тем, что люди живут в странах с высоким ППС. То есть присутствует слабая связь между рассмотренными параметрами. А значит в дальнейшем нет смысла рассматривать между ними взаимосвязь.

9. Между Social Support и Healthy Life Expectancy At Birth (в таблице лист "поле Soc и health..")



=> 60,5% дисперсии Healthy Life Expectancy At Birth можно объяснить тем, что у людей есть близкие и родственники, готовые помочь в любой момент. То есть присутствует значительная связь между рассмотренными параметрами.

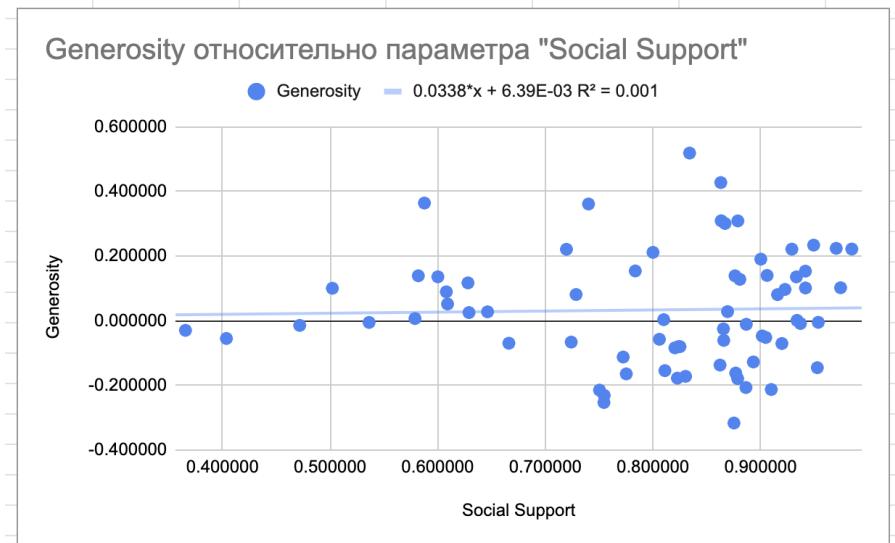
10. Между Social Support и Freedom To Make Life Choices (в таблице лист "поле Soc и free..")



=> 19,7% дисперсии Freedom To Make Life Choices можно объяснить тем, что у людей есть близкие и родственники, готовые помочь в любой момент. То есть присутствует

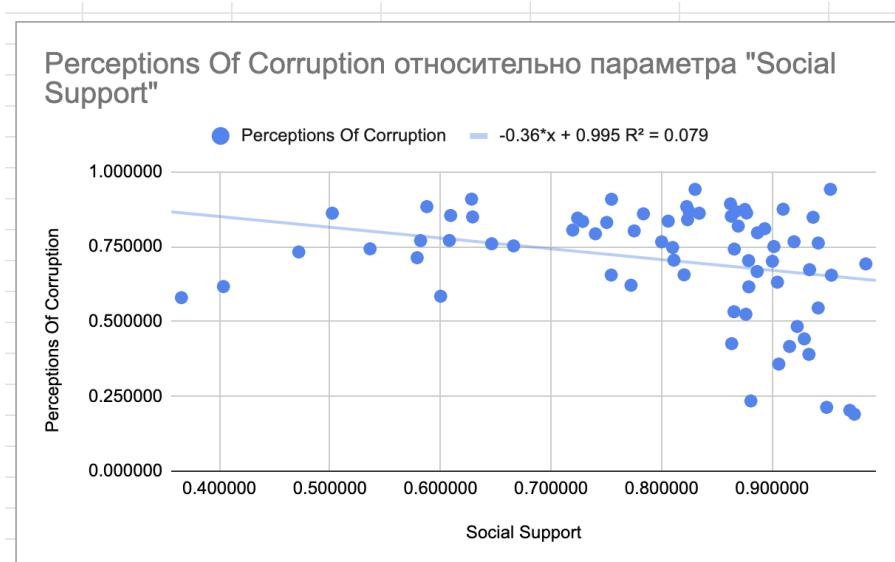
слабая связь между рассмотренными параметрами. А значит в дальнейшем нет смысла рассматривать между ними взаимосвязь.

11. Между Social Support И Generosity (в таблице лист "поле Soc и gener..")



=> 1% дисперсии Generosity можно объяснить тем, что у людей есть близкие и родственники, готовые помочь в любой момент. То есть присутствует слабая связь между рассмотренными параметрами. А значит в дальнейшем нет смысла рассматривать между ними взаимосвязь.

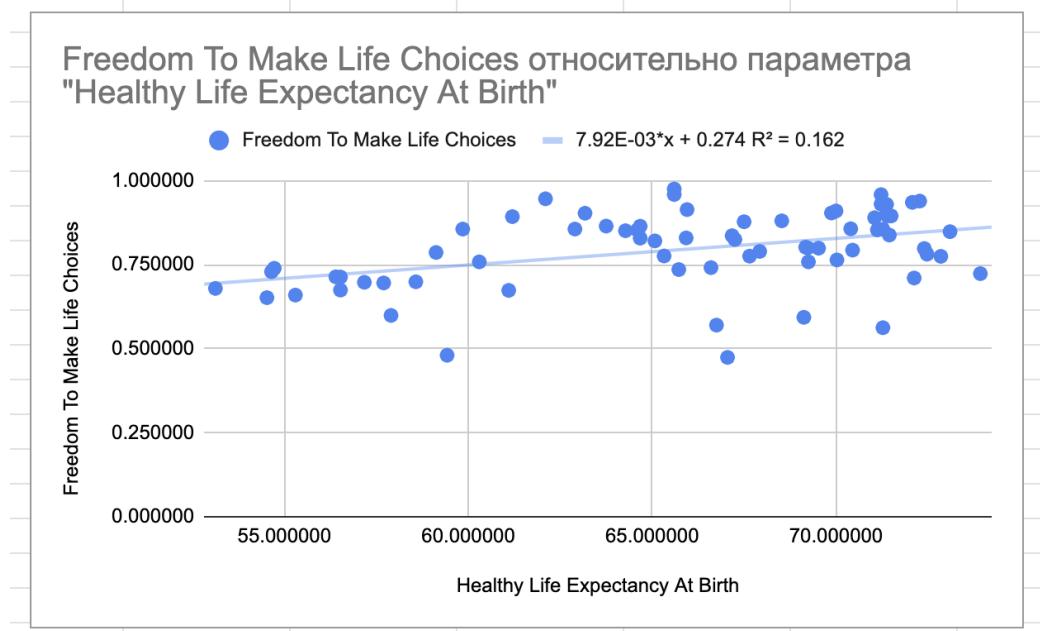
12. Между Social Support И Perceptions Of Corruption (в таблице лист "поле Soc и pre..")



=> 7,9% дисперсии Perceptions Of Corruption можно объяснить тем, что у людей есть близкие и родственники, готовые помочь в любой момент. То есть присутствует слабая

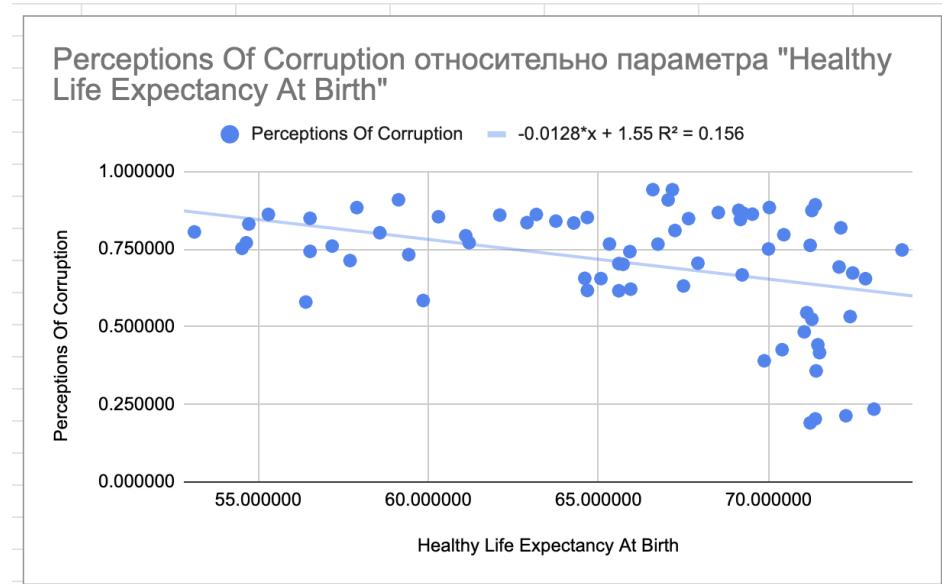
обратная связь между рассмотренными параметрами. А значит в дальнейшем нет смысла рассматривать между ними взаимосвязь.

13. Между Healthy Life Expectancy At Birth и Freedom To Make Life Choices (в таблице лист "поле Healthy и free..")



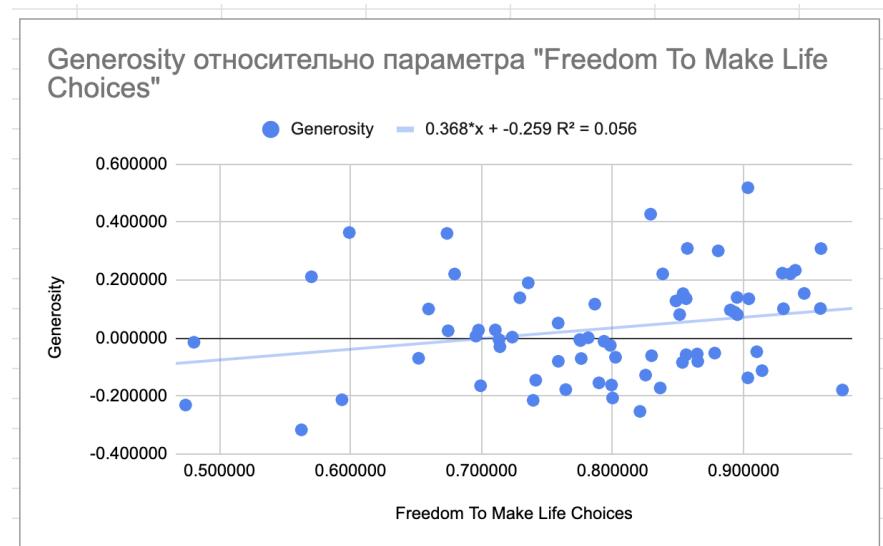
=> 16,2% дисперсии Freedom To Make Life Choices можно объяснить тем, что люди проживают в странах с высокой средней продолжительностью жизни. То есть присутствует слабая связь между рассмотренными параметрами. А значит в дальнейшем нет смысла рассматривать между ними взаимосвязь.

14. Между Healthy Life Expectancy At Birth и Perceptions Of Corruption (в таблице лист "поле life и prep..")



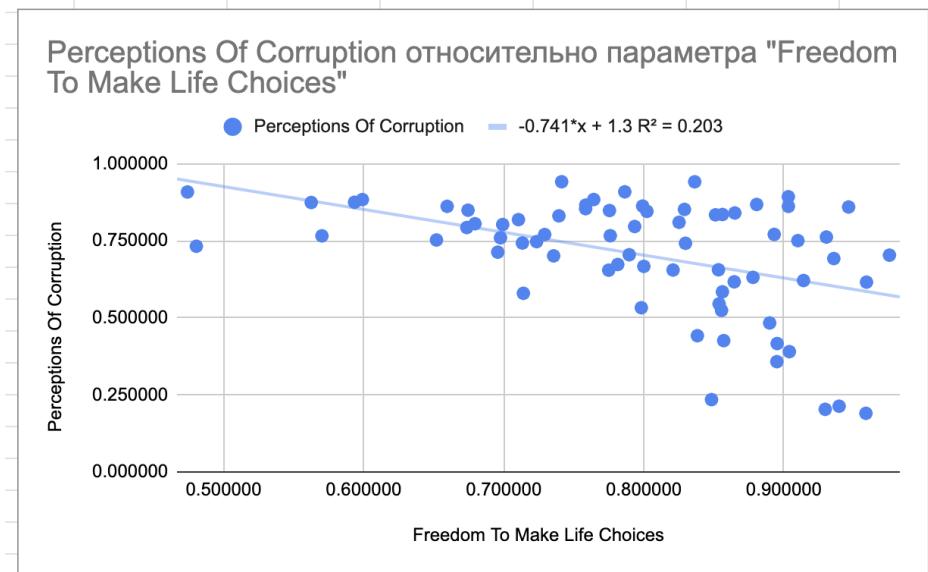
=> 15,6% дисперсии Perceptions Of Corruption можно объяснить тем, что люди проживают в странах с высокой средней продолжительностью жизни. То есть присутствует слабая связь между рассмотренными параметрами. А значит в дальнейшем нет смысла рассматривать между ними взаимосвязь.

15. Между Freedom To Make Life Choices и Generosity (в таблице лист "поле free и gen..")



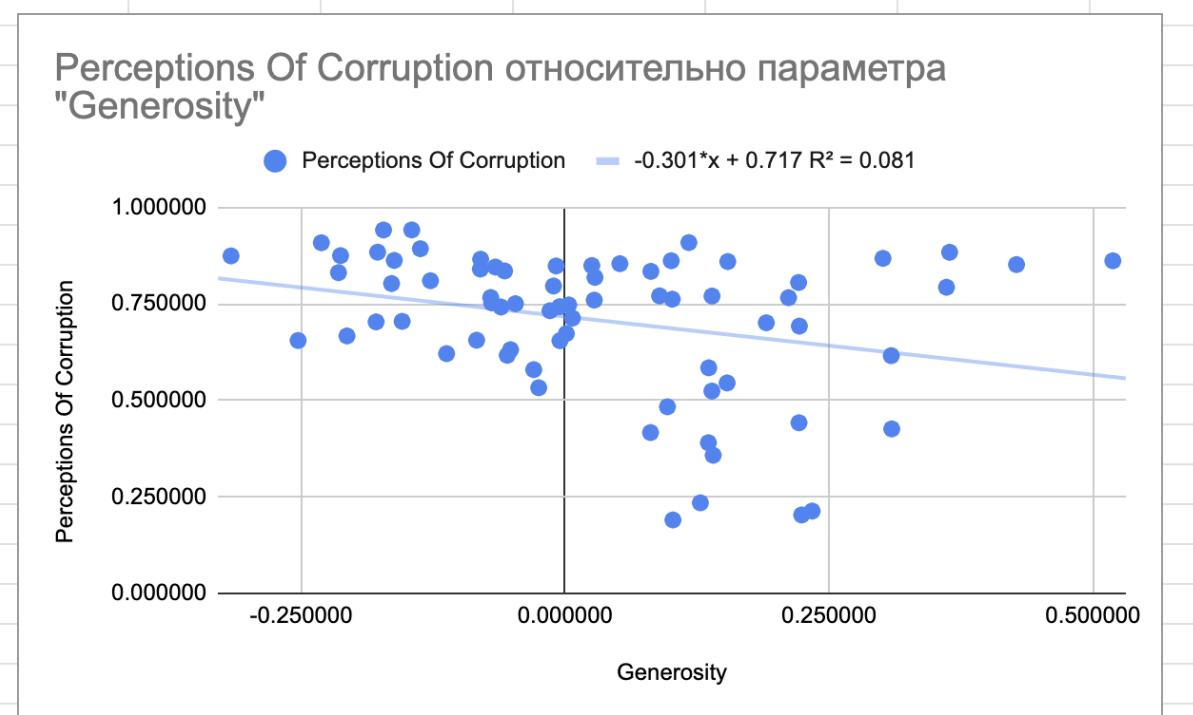
=> 5,6% дисперсии Generosity можно объяснить тем, что люди чувствуют свободу в принятии решения с своей стране. То есть присутствует слабая связь между рассмотренными параметрами. А значит в дальнейшем нет смысла рассматривать между ними взаимосвязь.

16. Между Freedom To Make Life Choices и Perceptions Of Corruption (в таблице лист "поле free и prep..")



=> 20,3% дисперсии Perceptions Of Corruption можно объяснить тем, что люди чувствуют свободу в принятии решения с своей стране. То есть присутствует слабая связь между рассмотренными параметрами. А значит в дальнейшем нет смысла рассматривать между ними взаимосвязь.

17. Между Generosity и Perceptions Of Corruption (в таблице лист "поле gen и prep..")



=> 8,1% дисперсии Perceptions Of Corruption можно объяснить тем, что люди пожертвовали за опрошенный месяц. То есть присутствует слабая обратная связь

между рассмотренными параметрами. А значит в дальнейшем нет смысла рассматривать между ними взаимосвязь.

2. Описание связи между переменными (сила, направление связи, наличие аномальных наблюдений)

Все подробности были описаны в предыдущих разделах, но вкратце опишем их

1. Life Ladder и Log GDP Per Capita

Данные связаны между собой значительно, направлены линейно (смотреть [тут](#)), аномальные данные имеются в Life Ladder в нижней части данных (смотреть [тут](#))

2. Life Ladder и Social Support

Данные связаны между собой значительно, направлены линейно (смотреть [тут](#)), аномальные данные имеются в Life Ladder (смотреть [тут](#)) и Social Support (смотреть [тут](#)) в нижних частях данных

3. Life Ladder и Healthy Life Expectancy At Birth

Данные связаны между собой значительно, направлены линейно (смотреть [тут](#)), аномальные данные имеются в Life Ladder (смотреть [тут](#)) и Healthy Life Expectancy At Birth (смотреть [тут](#)) в нижних частях данных

4. Life Ladder и Freedom To Make Life Choices

Данные связаны между собой умеренно, направлены линейно (смотреть [тут](#)), аномальные данные имеются в Life Ladder (смотреть [тут](#)) и Freedom To Make Life Choices (смотреть [тут](#)) в нижних частях данных

5. Log GDP Per Capita и Social Support

Данные связаны между собой значительная, направлены линейно (смотреть [тут](#)), аномальные данные имеются в Social Support (смотреть [тут](#)) в нижней части данных

6. Log GDP Per Capita и Healthy Life Expectancy At Birth

Данные связаны между собой сильно, направлены линейно (смотреть [тут](#)), аномальные данные имеются в Healthy Life Expectancy At Birth (смотреть [тут](#)) в нижней части данных.

7. Log GDP Per Capita и Freedom To Make Life Choices

Данные связаны между собой слабо, направлены линейно (смотреть [тут](#)), аномальные данные имеются в Freedom To Make Life Choices (смотреть [тут](#)) в нижней части данных.

8. Log GDP Per Capita и Generosity

Данные связаны между собой слабо, направлены обратно (смотреть [тут](#)), аномальные данные имеются в Generosity (смотреть [тут](#)) в нижней части данных.

9. Social Support и Healthy Life Expectancy At Birth

Данные связаны между собой значительно, направлены линейно (смотреть [тут](#)), аномальные данные имеются в Social Support (смотреть [тут](#)) и Healthy Life Expectancy At Birth (смотреть [тут](#)) в нижних частях данных

10. Social Support и Freedom To Make Life Choices

Данные связаны между собой слабо, направлены линейно (смотреть [тут](#)), аномальные данные имеются в Social Support (смотреть [тут](#)) и Freedom To Make Life Choices (смотреть [тут](#)) в нижних частях данных

11. Social Support и Generosity

Данные связаны между собой слабо, направлены обратно (смотреть [тут](#)), аномальные данные имеются в Social Support (смотреть [тут](#)) и Generosity (смотреть [тут](#)) в нижних частях данных

12. Social Support и Perceptions Of Corruption

Данные связаны между собой слабо, направлены линейно (смотреть [тут](#)), аномальные данные имеются в Social Support (смотреть [тут](#)) и Perceptions Of Corruption (смотреть [тут](#)) в нижних частях данных

Дальнейшие данные имеют слабую связь, поэтому их описывать и работать дальше с ними в корреляционном анализе смысла нет.

3. Построение и интерпретация матрицы парных коэффициентов корреляции наблюдений) без удаления аномальных данных (в таблице лист "попарные корреляции")

	Life Ladder	Log GDP Per Capita	Social Support	Healthy Life Expectancy At Birth	Freedom To Make Life Choices	Generosity	Perceptions Of Corruption
Life Ladder	1	0.8141827442	0.807808518	0.7861371492	0.5373562055	0.03866142183	-0.4936263965
Log GDP Per Capita	0.8141827442	1	0.815666241	0.8790280644	0.4002573452	-0.0764542749	-0.4697301206
Social Support	0.807808518	0.815666241	1	0.777814788	0.4436019948	0.02797741604	-0.2805480838
Healthy Life Expectancy At Birth	0.7861371492	0.8790280644	0.777814788	1	0.4021836877	-0.04906143535	-0.3945244837
Freedom To Make Life Choices	0.5373562055	0.4002573452	0.4436019948	0.4021836877	1	0.2373291594	-0.4504988265
Generosity	0.03866142183	-0.0764542749	0.02797741604	-0.04906143535	0.2373291594	1	-0.2838738709
Perceptions Of Corruption	-0.4936263965	-0.4697301206	-0.2805480838	-0.3945244837	-0.4504988265	-0.2838738709	1

Несмотря на то, что при построении полей корреляции зависимость между Life Ladder и Log GDP Per Captia, Life Ladder и Social Support, Life Ladder и Healthy life Expectancy at birth, Log GDP Per Captia и Social Support, Social Support и Healthy life Expectancy at birth имели значительную связь, а тут у них всех сильная взаимосвязь.

Также можно отметить, что имеет место дальше проверять 6 гипотезу так, как коэффициент корреляции между Life Ladder и Perceptions Of Corruption с натяжкой переходит в среднюю зависимость.

Ещё был неверный вывод при построении полей корреляции между Life Ladder и Freedom To Make Life Choices. При построении было замечено, что между параметрами слабая зависимость, но на деле она оказалась средней. А значит можно предположить новые гипотезы на основе этих столбцов.

Можно также дальше поработать и выдвинуть гипотезы по столбцам Healthy life Expectancy at birth и Generosity.

Все остальные результаты результаты совпали с выводами из построения полей корреляции.

4. Построение и интерпретация матрицы парных коэффициентов корреляции (в таблице лист "попарные корреляции скоррект")

Стандартная матрица корреляции

	Life Ladder	Log GDP Per Capita	Social Support	Healthy Life Expectancy At Birth	Freedom To Make Life Choices	Generosity	Perceptions Of Corruption
Life Ladder	1	0.8141827442	0.807808518	0.7861371492	0.5373562055	0.03866142183	-0.4936263965
Log GDP Per Capita	0.8141827442	1	0.815666241	0.8790280644	0.4002573452	-0.0764542749	-0.4697301206
Social Support	0.807808518	0.815666241	1	0.777814788	0.4436019948	0.02797741604	-0.2805480838
Healthy Life Expectancy At Birth	0.7861371492	0.8790280644	0.777814788	1	0.4021836877	-0.04906143535	-0.3945244837
Freedom To Make Life Choices	0.5373562055	0.4002573452	0.4436019948	0.4021836877	1	0.2373291594	-0.4504988265
Generosity	0.03866142183	-0.0764542749	0.02797741604	-0.04906143535	0.2373291594	1	-0.2838738709
Perceptions Of Corruption	-0.4936263965	-0.4697301206	-0.2805480838	-0.3945244837	-0.4504988265	-0.2838738709	1

Матрица корреляции с скорректированными данными

	Life Ladder	Log GDP Per Capita	Social Support	Healthy Life Expectancy At Birth	Freedom To Make Life Choices	Generosity	Perceptions Of Corruption
Life Ladder	1	0.7997667139	0.8051153208	0.7287752747	0.4965766265	-0.02591072873	-0.5774004218
Log GDP Per Capita	0.7997667139	1	0.8015826255	0.8873948293	0.3409461618	-0.1106488354	-0.5471312015
Social Support	0.8051153208	0.8015826255	1	0.766882929	0.3693032035	-0.00901464123	-0.3758551138
Healthy Life Expectancy At Birth	0.7287752747	0.8873948293	0.766882929	1	0.3368137191	-0.1358500665	-0.4518924714
Freedom To Make Life Choices	0.4965766265	0.3409461618	0.3693032035	0.3368137191	1	0.2159734366	-0.4849010156
Generosity	-0.02591072873	-0.1106488354	-0.00901464123	-0.1358500665	0.2159734366	1	-0.2744890446
Perceptions Of Corruption	-0.5774004218	-0.5471312015	-0.3758551138	-0.4518924714	-0.4849010156	-0.2744890446	1

Как можно увидеть корреляция между переменными Life Ladder и всеми другими переменными кроме последней; Log GDP Per Capita и всеми другими переменными кроме 4-ой; Social Support и всеми другими переменными; Healthy Life Expectancy At Birth и всеми другими переменными кроме 2-ой; Freedom To Make Life Choices и всеми другими переменными; Generosity и всеми другими переменными кроме последней; Perceptions Of Corruption и всеми другими переменными кроме предпоследней ухудшилось

По итогу ситуация только ухудшилась в основной массе, но это можно объяснить тем, что принцип удаления данных был неправильно избран (удалялись все страны, которые встречались в хотя бы в 3 столбцах исследуемых). Но это не означает, что корректировка данных при предварительном анализе была некорректной, так как во всех столбцах получилось разное количество выбросных данных и сложно было выбрать метод для корректировки всей таблицы.

5. Частные коэффициенты корреляции

Расчеты частных коэффициентов корреляции находятся в таблице (лист “Частные коэффициенты корреляции”).

Частные коэффициенты корреляции							
R12/3	0.4553585429		R23/1	0.4615589094		R34/1	0.391884387
R12/4	0.4179138211		R23/4	0.440335662		R34/2	0.2205088855
R12/5	0.7751820901		R23/5	0.7769507803		R34/5	0.7304958035
R12/6	0.8201504592		R23/6	0.820527095		R34/6	0.7804323541
R12/7	0.758458135		R23/7	0.8070799534		R34/7	0.7563997568
R45/1	-0.03885366198		R56/1	0.2569690452		R67/1	-0.304697698
R45/2	0.1152316462		R56/2	0.2932302924		R67/2	-0.3633006621
R45/3	0.1014493938		R56/3	0.2510604778		R67/3	-0.2876864468
R45/6	0.4265120965		R56/4	0.2811078304		R67/4	-0.3303952806
R45/7	0.2736005252		R56/7	0.1278480944		R67/5	-0.204039589

Анализ частных и парных коэффициентов:

1. Все переменные усиливают связь между Life Ladder и Log GDP Per Capita, кроме Generosity.
2. Все переменные усиливают связь между Social Support и Log GDP Per Capita, кроме Generosity.
3. Все переменные усиливают связь между Social Support и Healthy life expectancy at birth, кроме Generosity.
4. Все переменные усиливают связь между Freedom to make life choices и Healthy life expectancy at birth, кроме Generosity.
5. Все переменные ослабляют связь между Freedom to make life choices и Generosity, кроме Perceptions of corruption.
6. Все переменные усиливают связь между Generosity и Perceptions of corruption, кроме Freedom to make life choices.

По частным и парным коэффициентам корреляции можно заметить, что переменная Generosity практически всегда ослабляет связь между другими переменными, что показывает ее независимость.

6. Множественный коэффициент корреляции

Расчеты множественного коэффициента корреляции находится в таблице (лист “Частные коэффициенты корреляции”).

Множественный коэффициент корреляции(зависимая переменная - Life Ladder)									
R1/2,3	0.8512331598	R1/3,4	0.8459334599	R1/4,5	0.8224199566	R1/5,6	0.5450878436	R1/6,7	0.5048413376
R1/2,4	0.8274821072	R1/3,5	0.8321359338	R1/4,6	0.7899307245	R1/5,7	0.6067453225		
R1/2,5	0.8462540829	R1/3,6	0.8079682917	R1/4,7	0.8110984392				
R1/2,6	0.8204487019	R1/3,7	0.8543607712						
R1/2,7	0.8238653887								

В качестве зависимой переменной был выбран параметр Life Ladder.

По таблице можно увидеть, что наиболее коррелируемые для рассматриваемого параметра переменные являются Log GDP Per Capita, Social Support и Perceptions of corruption. Это говорит о том, что счастье людей очень сильно зависит от социальной поддержки, уровня коррупции, а также паритета покупательской способности.

Также сильную зависимость показывает Healthy life expectancy at birth и Freedom to make life choices, то есть на уровень счастья также сильно оказывают влияние продолжительность жизни и свобода выбора.

7. Вывод

Факторы, наиболее влияющие на уровень счастья человека, являются социальная поддержка, уровень покупательской способности, свобода выбора, а также уровень коррупции и средняя продолжительность жизни в стране.

В ходе анализа подтвердились гипотезы 1, 2, 3, 6, 7 и 8, то есть человеку очень важна поддержка от других людей; чем больше покупательская способность человека, тем он добре к людям и тем больше продолжительность его жизни. Опровергаются гипотезы 4 и 5, то есть средняя продолжительность жизни человека не связана с его щедростью, а также покупательская способность человека не зависит от уровня коррупции в стране.

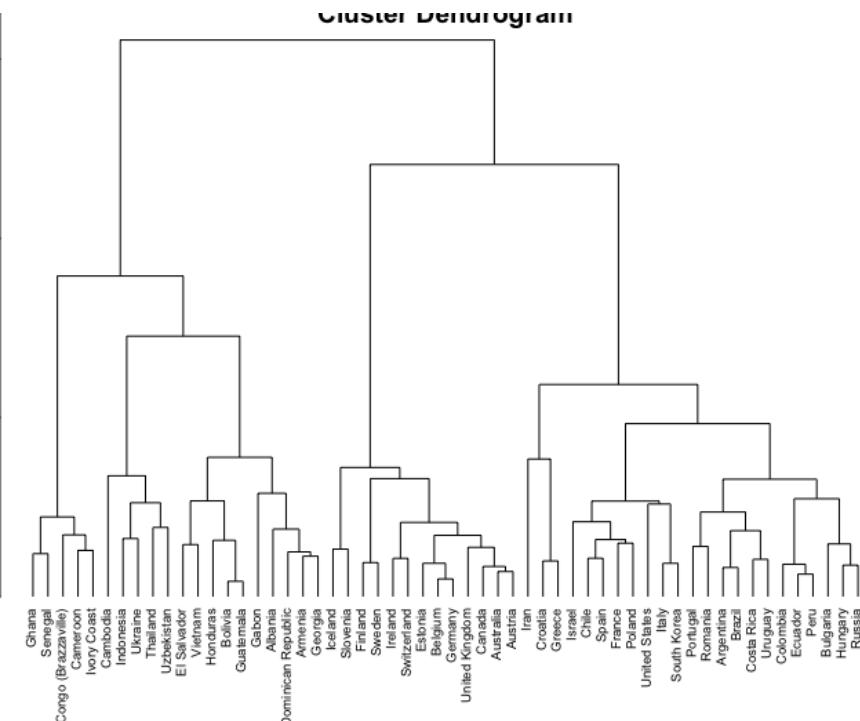
Кластерный анализ

Для проведения иерархической кластеризации было решено выбрать в качестве меры расстояния метрику Канберры (взвешенное версия манхэттенской метрики), которая рассчитывается по следующей формуле:

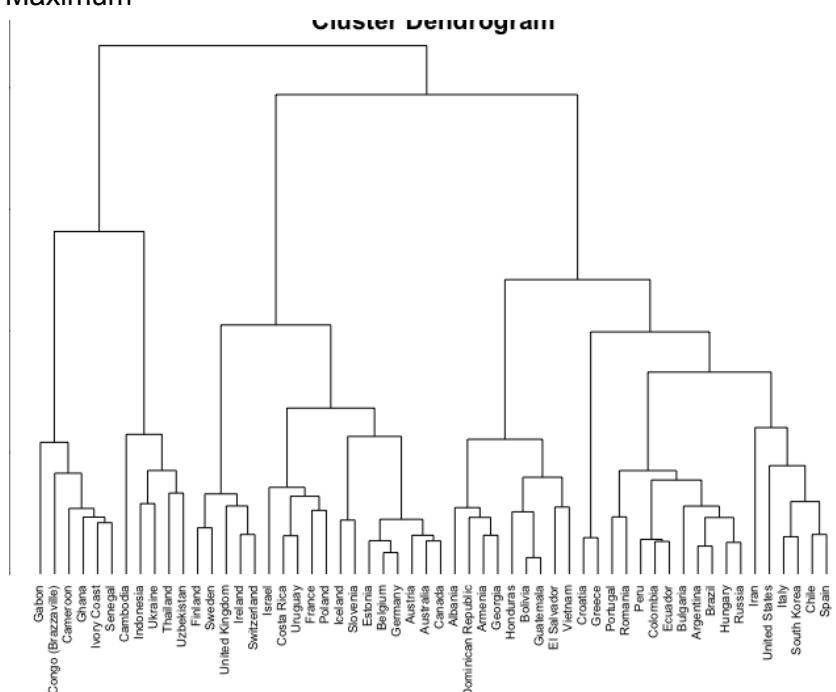
$$\sum_i |x_i - y_i| / (|x_i| + |y_i|)$$

Она наиболее равномерно разбивает рассматриваемые параметры. Это можно понять, посмотрев на дендрограммы:

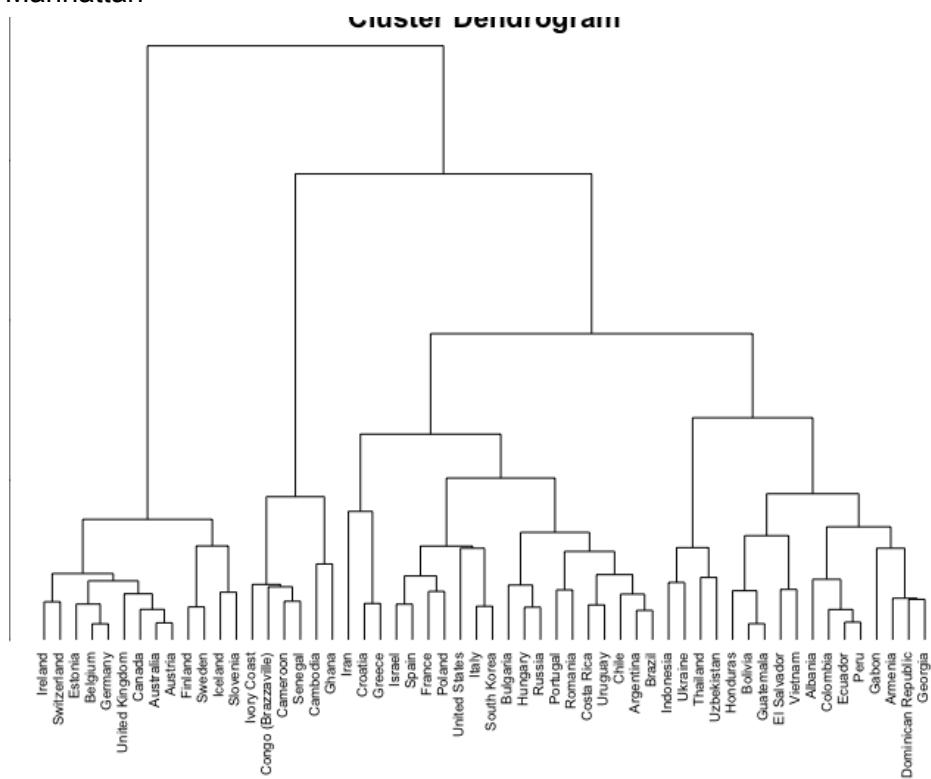
1. Евклидово расстояние



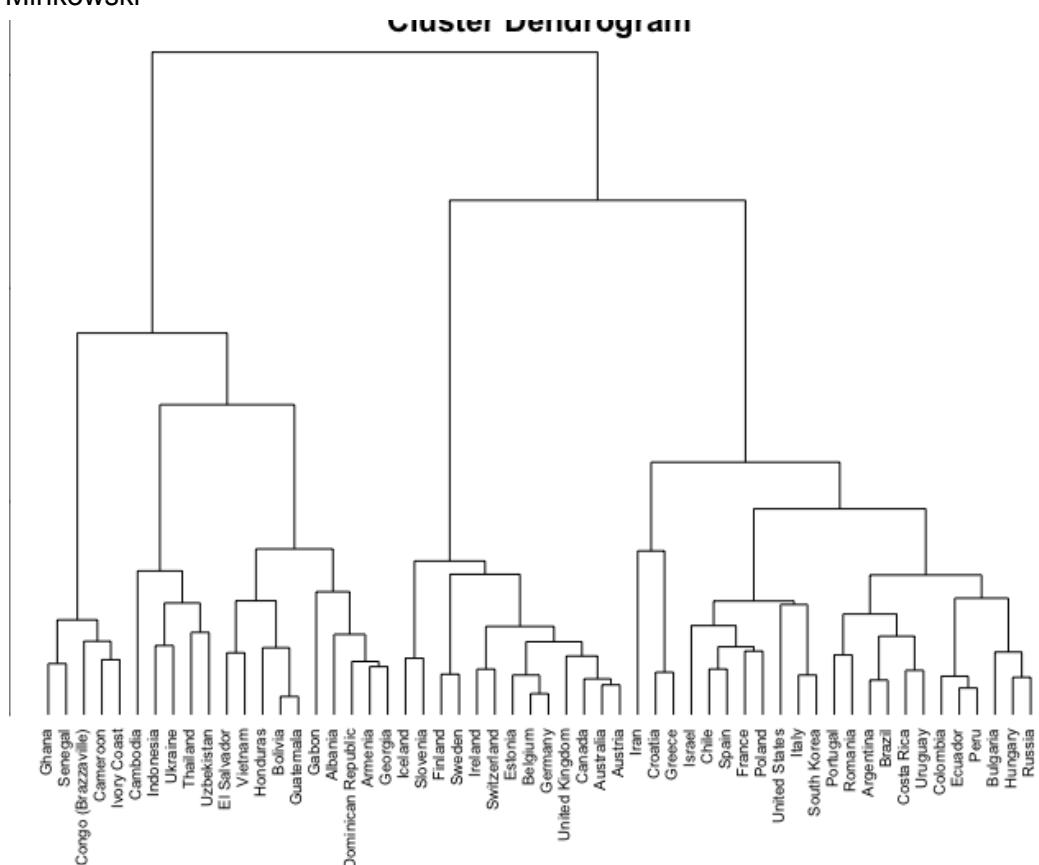
2. Maximum



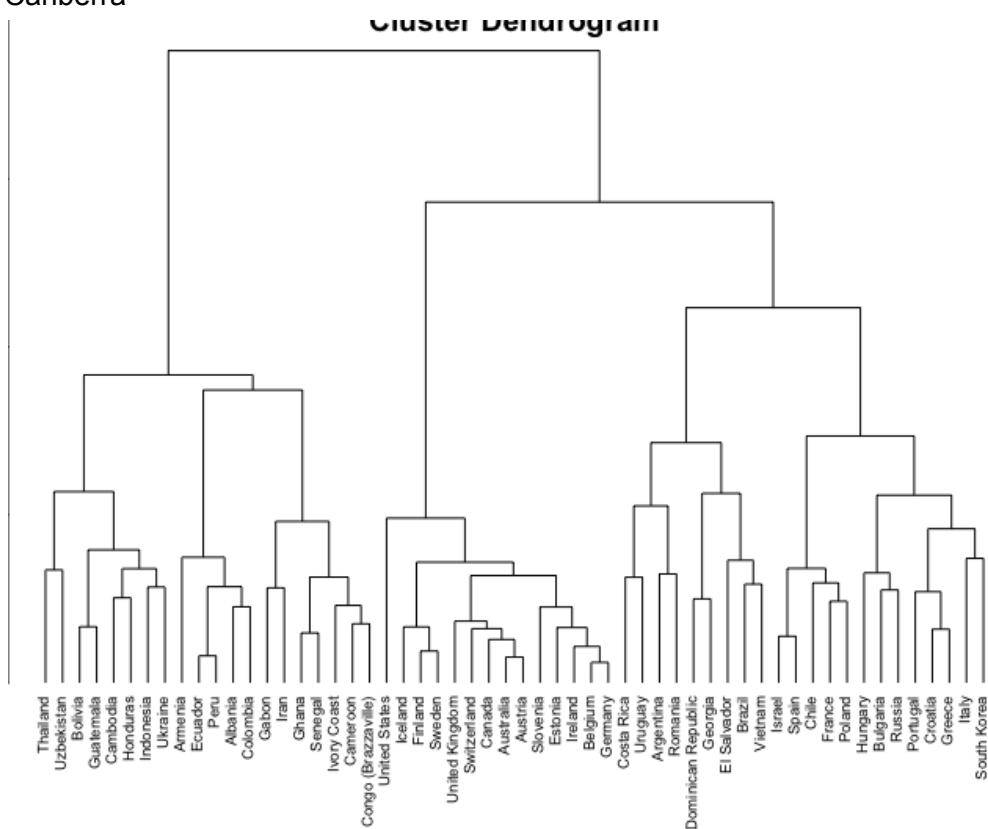
3. Manhattan



4. Minkowski

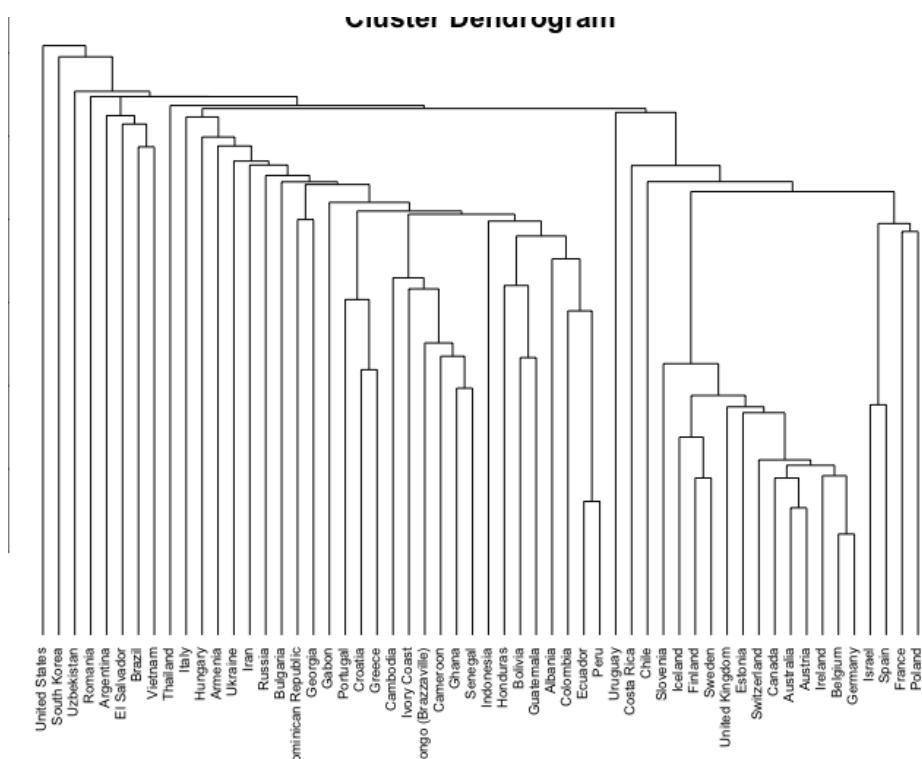


5. Canberra

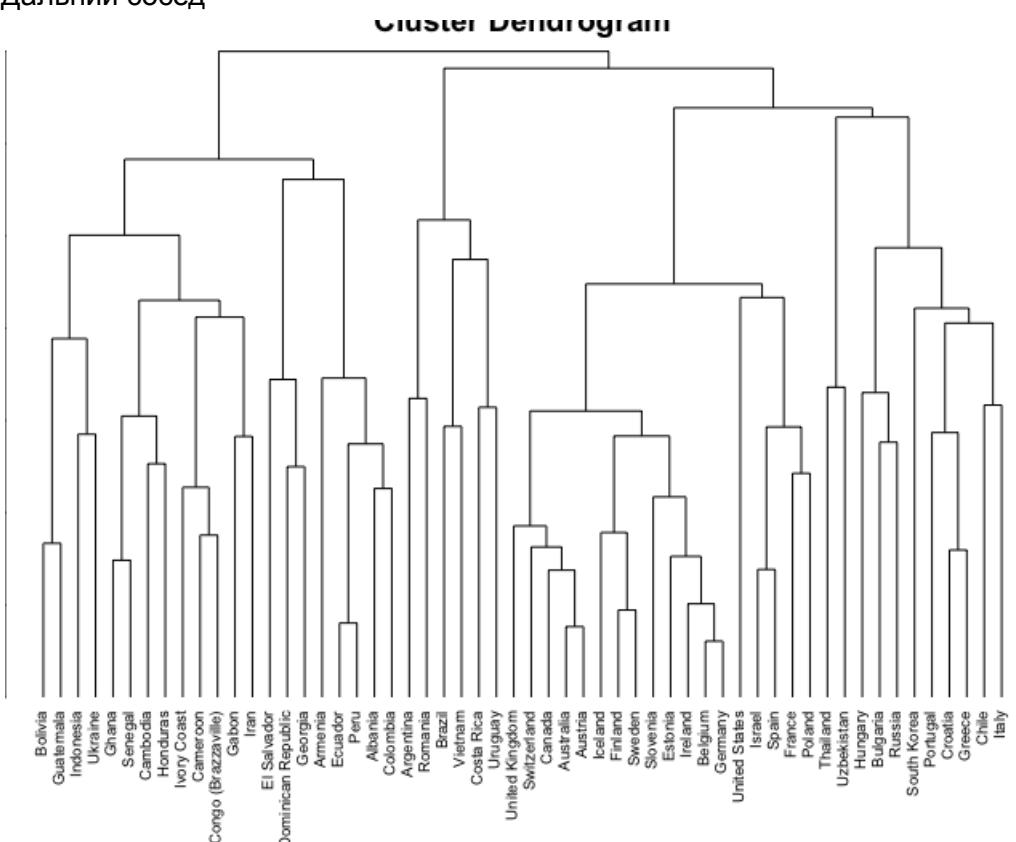


Было рассмотрено 5 вариантов разбиения на кластеры:

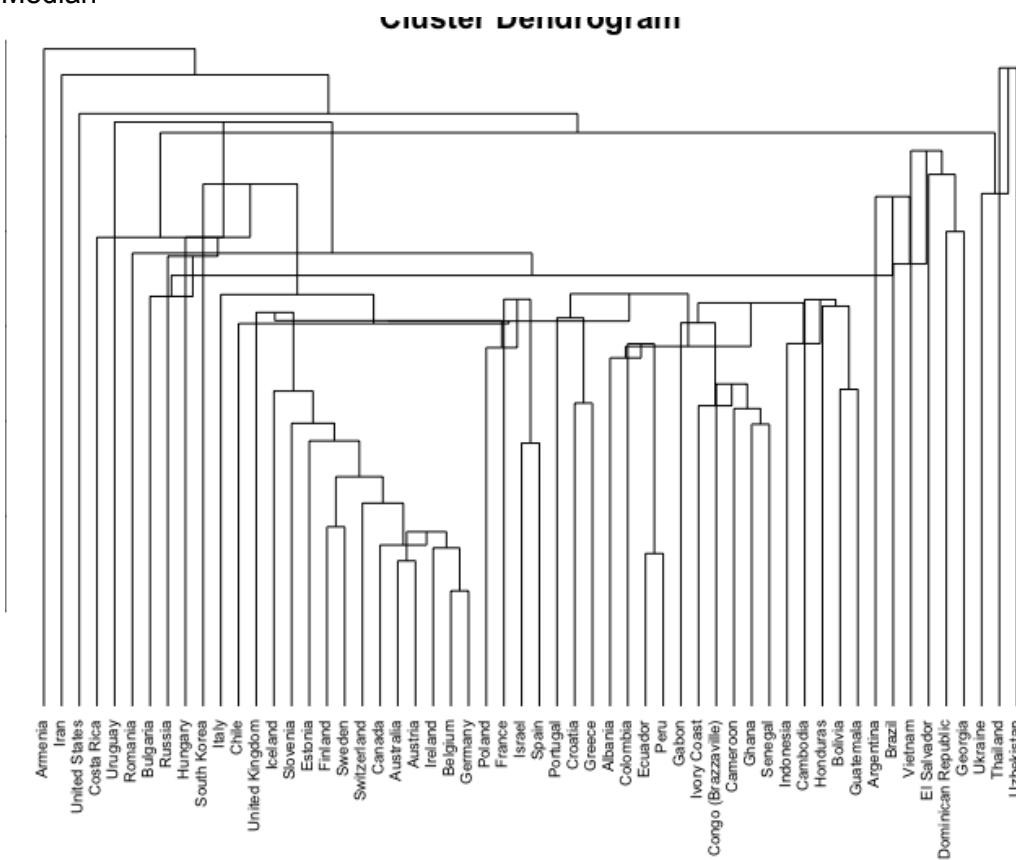
1. Ближний сосед



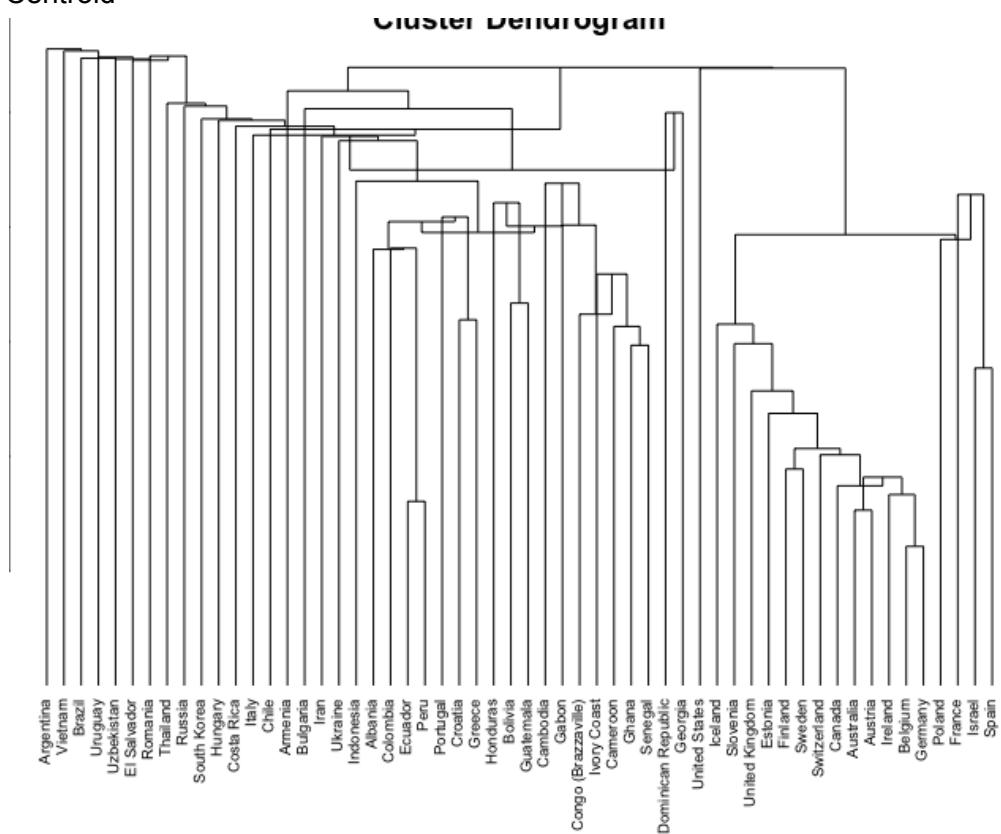
2. Дальний сосед



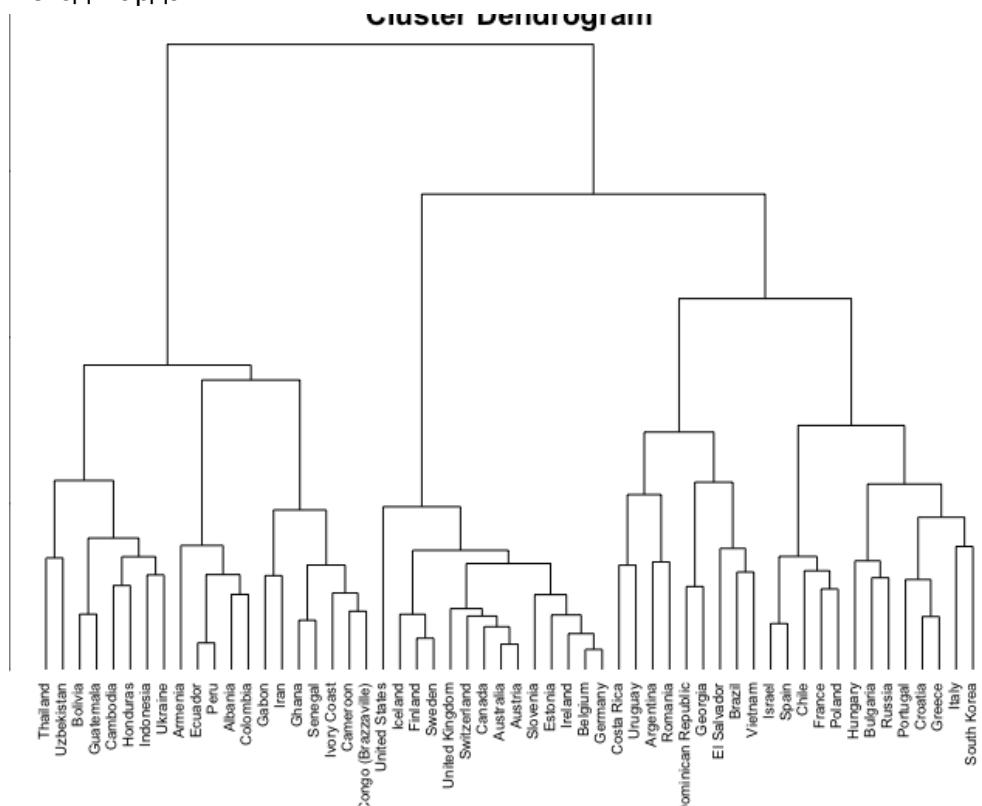
3. Median



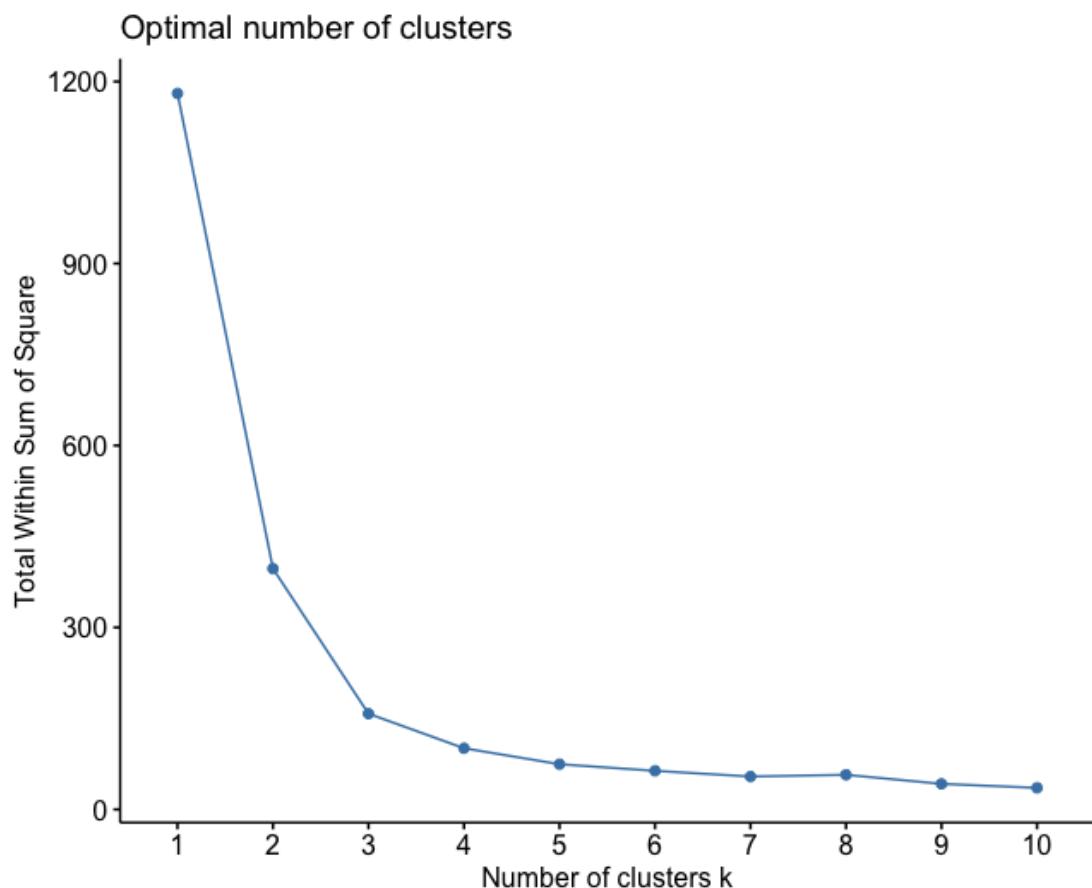
4. Centroid



5. Метод Варда

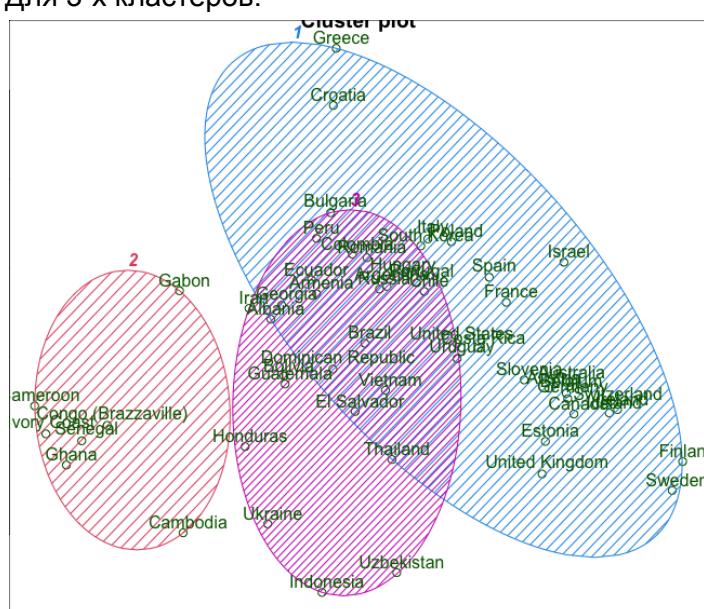


Также был построен график количества кластеров по сравнению с общей суммой квадратов для определения оптимального числа кластеров:

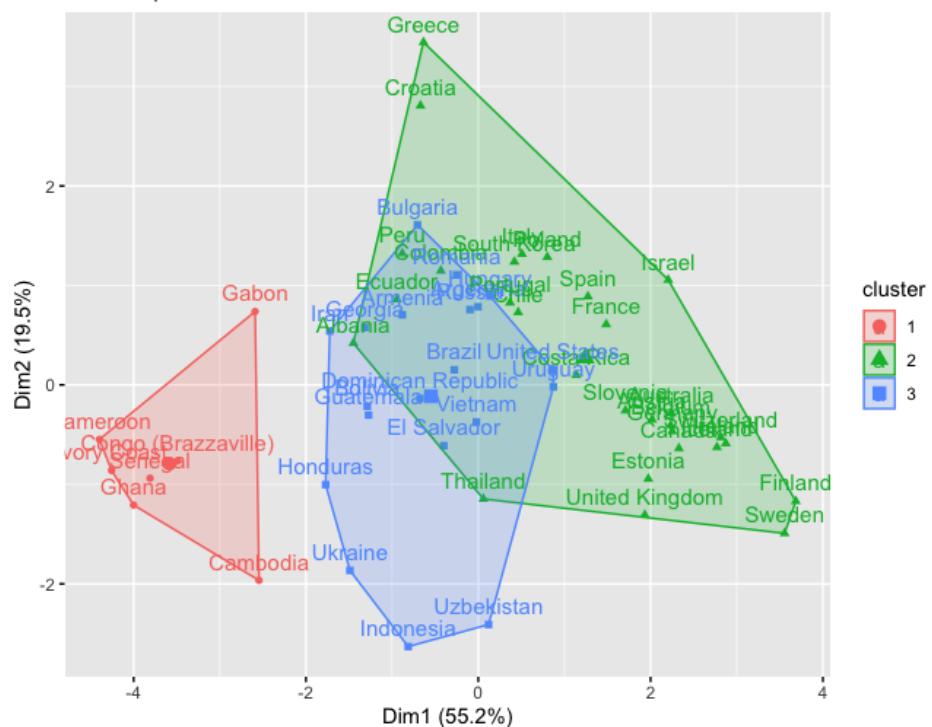


По графику видно изгибы при 3 и 4 кластерах, но изгиб при трех кластерах больше, поэтому выдвинем гипотезу о том, что разбиение на 3 кластера оптимальнее, чем на 4. Приведем по 2 графика для каждого количества кластеров и сравним результаты:

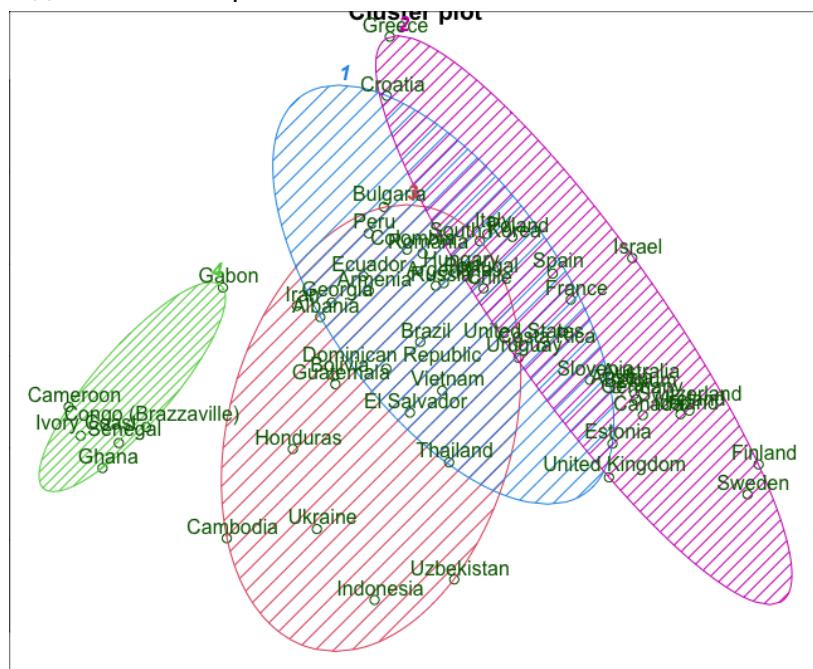
Для 3-х кластеров:



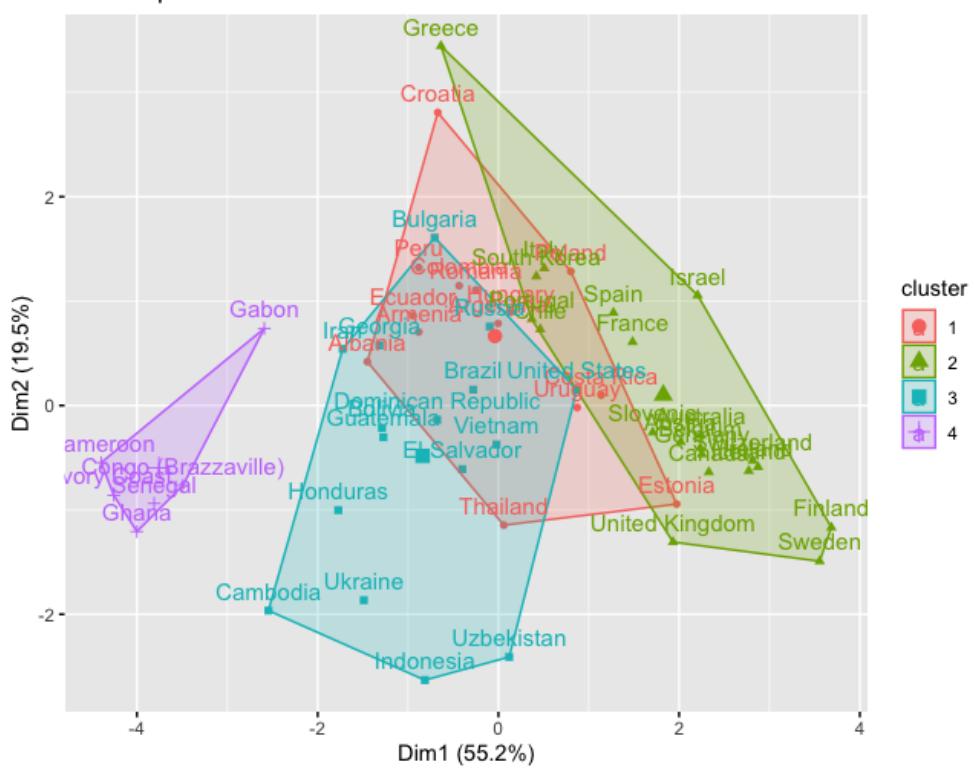
Cluster plot



И для 4-х кластеров:



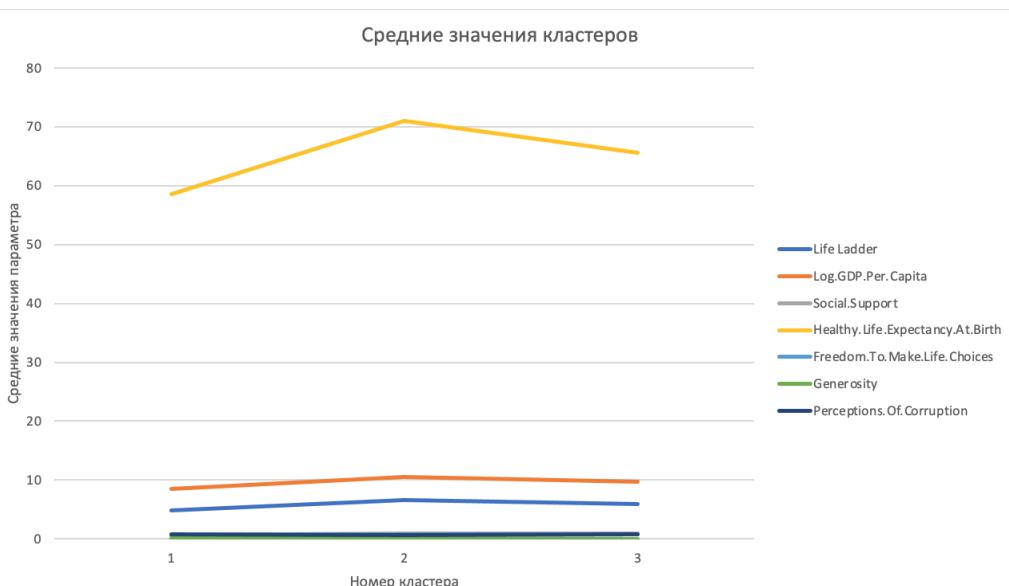
Cluster plot



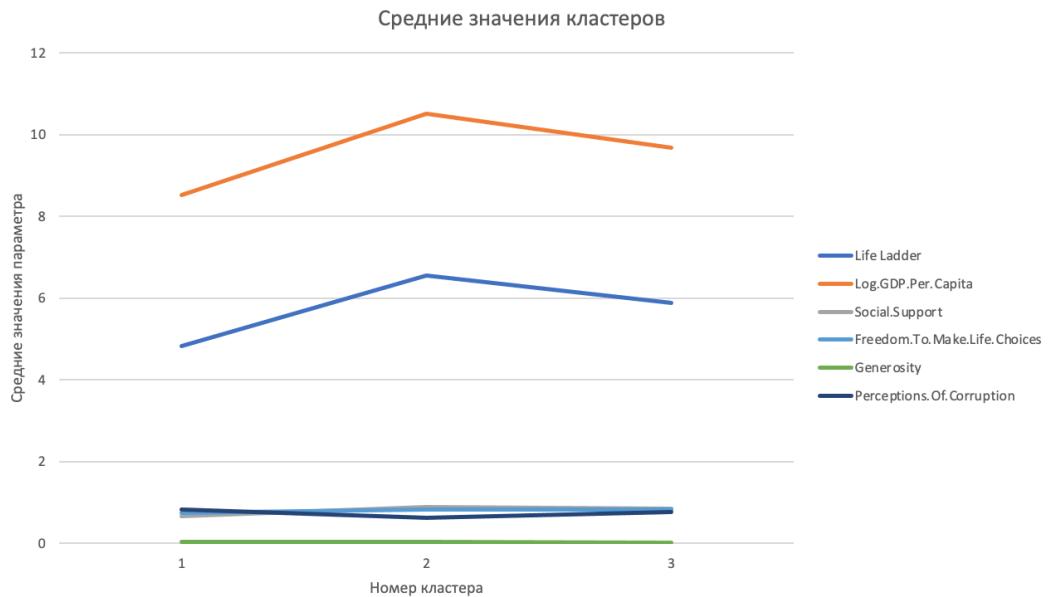
При разбиении на 4 кластера можно увидеть, что первый кластер почти полностью перекрыт вторым и третьим, и занимает небольшую непокрытую область.

При разбиении на 3 кластера классификация выглядит намного лучше, как и предполагалась исходя из графика количества кластеров по сравнению с общей суммой квадратов.

Также был построен график средних значений кластера:



И также еще один график средних значений кластера без параметра Healthy life expectancy at birth, так как этот параметр принимает значения намного большие, чем остальные, и для удобства рассмотрения остальных параметров:



Структура кластеров:

K-means clustering with 3 clusters of sizes 29, 7, 20

Cluster means:

	Life.Ladder	Log.GDP.Per.Capita	Social.Support	Healthy.Life.Expectancy.At.Birth
1	6.547702	10.517665	0.8927716	71.02414
2	4.836207	8.523667	0.6582787	58.61786
3	5.889023	9.682321	0.8487860	65.67125

	Freedom.To.Make.Life.Choices	Generosity	Perceptions.Of.Corruption
1	0.8261193	0.03286421	0.6305891
2	0.7537117	0.02951929	0.8253569
3	0.8292496	0.00915865	0.7664057

Clustering vector:

	Albania	Argentina	Armenia	Australia
	1	3	3	1
Austria		Belgium	Bolivia	Brazil
	1	1	3	3
Bulgaria		Cambodia	Cameroon	Canada
	3	2	2	1
Chile		Colombia	Congo (Brazzaville)	Costa Rica
	1	1	2	1
Croatia		Dominican Republic	Ecuador	El Salvador
	1	3	1	3
Estonia		Finland	France	Gabon
	1	1	1	2
Georgia		Germany	Ghana	Greece
	3	1	2	1
Guatemala		Honduras	Hungary	Iceland
	3	3	3	1
Indonesia		Iran	Ireland	Israel
	3	3	1	1
Italy		Ivory Coast	Peru	Poland
	1	2	1	1
Portugal		Romania	Russia	Senegal
	1	3	3	2
Slovenia		South Korea	Spain	Sweden
	1	1	1	1
Switzerland		Thailand	Ukraine	United Kingdom
	1	1	3	1
United States		Uruguay	Uzbekistan	Vietnam
	3	3	3	3

Within cluster sum of squares by cluster:

[1] 72.51909 29.51544 55.79379
(between_SS / total_SS = 86.6 %)

К первому кластеру отнеслось 7 стран, во вторую - 29, в третью - 20.

Вывод

Второй кластер получилось определить довольно точно, так как он не пересекается с другими. Первый и третий кластеры имеют большую общую часть.

Если посмотреть на средние значения по кластерам можно увидеть, что наименьший средний показатель Life Ladder заключен во второй кластер, то есть в нем находятся страны с наименьшим показателем счастья, это такие страны, как: Cambodia, Cameroon, Congo, Gabon, Ghana, Ivory Coast, Senegal, то есть африканские страны.

Наибольший же показатель счастья заключен в первой кластере, это такие страны, как: Греция, Израиль, Испания, Франция, Англия, Эстония, Финляндия и Швеция.

И средний показатель счастья заключен в третьем кластере: Россия, Украина, Узбекистан, Индонезия и Гватемала.