**Федеральное государственное автономное образовательное учреждение высшего образования**

**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

Название дисциплины:

«Прикладной статистический анализ»

Отчет о выполнении модульного домашнего задания №1

«Проведение предварительного, корреляционного и кластерного анализов данных разных стран мира»

Выполнила:

Зайдес Анна Александровна, ББИ226

Москва, 2024 г.

## **Содержание**

[**Содержание 1**](#_s7m4b1v1oqra)

[**Введение 2**](#_hgedk689dh3s)

[**Предварительный анализ 2**](#_w8i4ggq3uy5e)

[Графическое представление данных. 2](#_4caqju519zsa)

[Точечные графики 3](#_b7gdg98roebo)

[Листовые диаграммы 6](#_aaky702yy3vm)

[Ящик с усами 7](#_xp4h3288f6on)

[Характеристики СВ 11](#_1cxb9qj5tnlm)

[Z-преобразование 11](#_uvuycn5kumni)

[1,5 и 3 IQR 16](#_rk96uv5qnh7w)

[**Корреляционный анализ 21**](#_m305d2qhcfvo)

[Множественная регрессия 33](#_lpe3hixyuakn)

[Вывод 34](#_5ita47f87o66)

[**Кластерный анализ 34**](#_mn40qpb3ij6o)

[Иерархическая кластеризация 34](#_8p67o8o6797j)

[Метод k-средних 37](#_yfk7321fh5n)

[**Выводы 40**](#_y0lwr9d5ka9j)

## **Введение**

Данная работа ставит перед собой цель анализа того, как некоторые выбранные нами показатели влияют на выбросы углекислого газа в атмосферу. В рамках увеличивающегося тренда на охрану окружающей среды и реализации так называемой “зеленой политики” важно понимать, какие именно факторы могут быть названы предикторами для более выбросов углекислого газа. Для реализации данной цели мной были выбраны данные Всемирного банка о ВВП на душу населения, коэффициенте фертильности, населении на 2020 год для 90 стран мира, преимущественно расположенных на Евразийском континенте, а также данные организации Freedom House о том, является ли политический режим в стране демократическим или нет. На основе этих данных я хочу понять, влияет ли численность населения, его прирост, богатство государства и его политический режим на выбросы углекислого газа, для чего нами будет проведен первичный анализ данных для определения направлений дальнейшего исследования.

**Задача:** провести предварительный, корреляционный и кластерный анализы данных и принять или опровергнуть сформулированные гипотезы.

**Гипотеза:** Рабочими гипотезами являются следующие тезисы: Политический режим, численность населения, ВВП на душу населения, а также коэффициент рождаемости связаны с числом выбросов углекислого газа в атмосферу и являются предикторами данных выбросов. Также мной будут проверены соотношения между данными переменными для подтверждения или опровержения устоявшихся тезисов (например, что в богатых странах малая рождаемость и т.п).

## **Предварительный анализ**

### **Графическое представление данных.**

Для начала исследуем данные на пустые значения. С пустыми значениями можно:

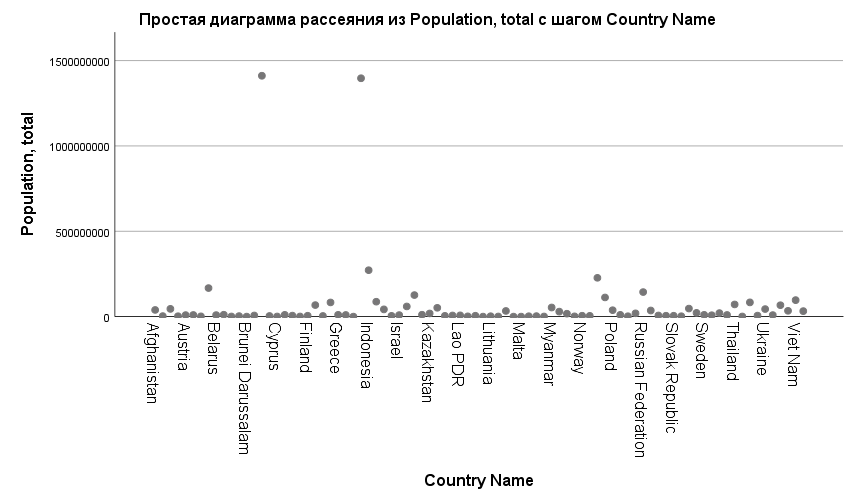
1. Присвоить среднее
2. Присвоить мин/мах
3. Удалить страну из анализа
4. Поставить просто 0

Так как исследуются реальные данные на реальных странах, мне кажется не корректным для каких-то стран ставить неверные значения, поэтому я решила удалить 4 страны из анализа. Подстановка других значений может привести к ложным результатам, а 4 страны на фоне 90 не такое большое значение.

### **Точечные графики**

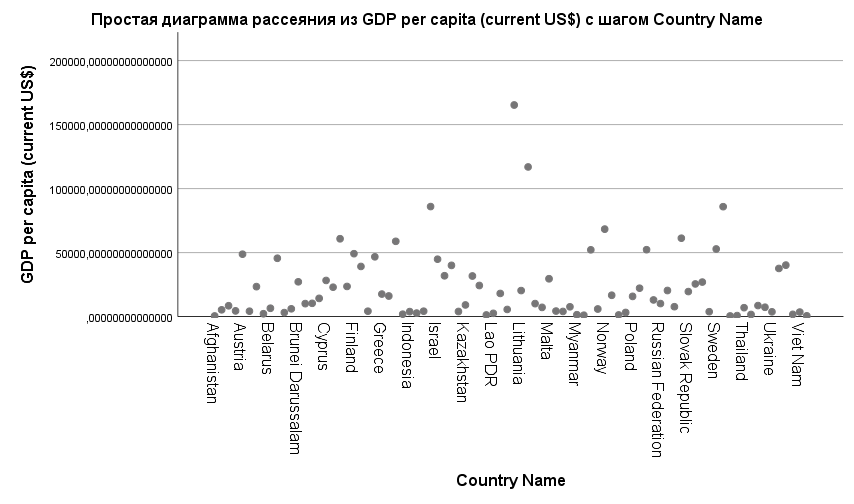
У нас имеются 5 признаков для каждой страны. Построим 5 точечных диаграмм.

1. Страна - численность населения



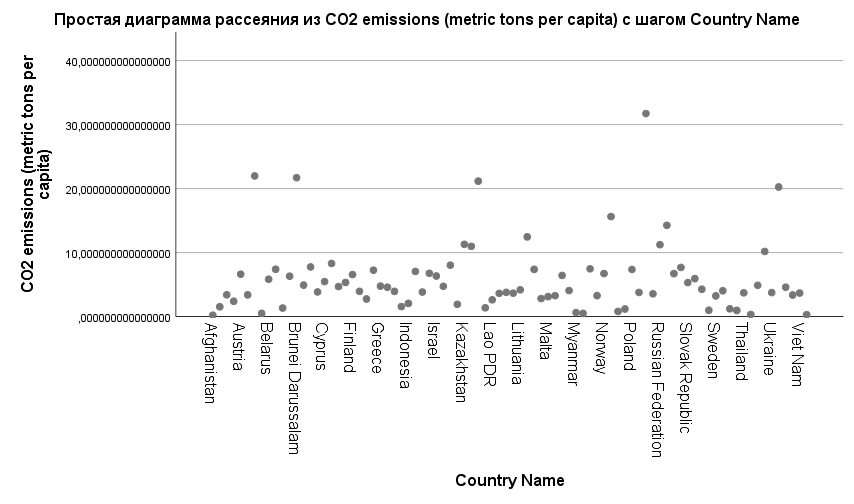
На графике уже очень хорошо видно выбросы - сразу бросаются в глаза страны, с численностью населения приближающегося к 1,5 миллиардам — это Китай и Индия, что очень очевидно, ведь это известный факт.

1. Страна - ВВП на душу населения



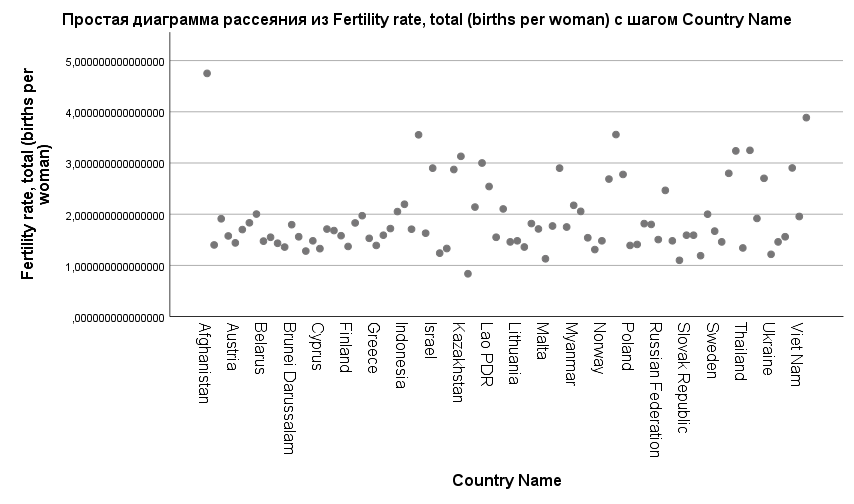
Здесь тоже бросаются в глаза выбросы - назовем пока что один, который находится выше всех, так как он явно выброс, остальные могут и не войти. Итак, страна с наибольшим ВВП на душу населения - Лихтенштейн - самая маленькая страна по населению, которая, кстати, может оказаться и выбросом в первом графике по нижней границе. Наверное, из-за такого маленького населения и выходит большое ВВП на душу населения. Проверим позже корреляционную связь между этими двумя признаками

1. Страна - CO2 выбросы



Здесь тоже видны предположительные выбросы - самый высокий это Катар.

1. Страна - коэффициент фертильности

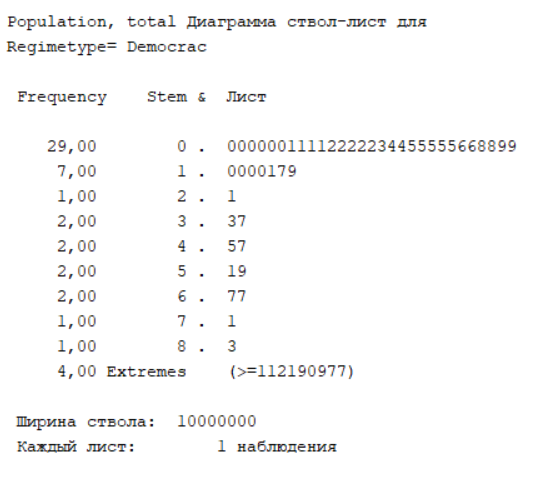
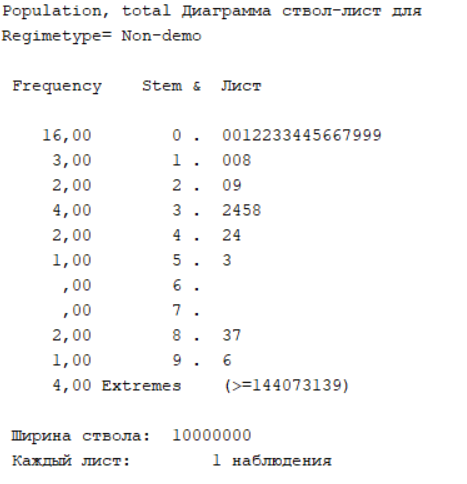


Тут пока что кажется, что выбросов вообще не будет или максимум один - Афганистан.

Делать точечный график на политический режим не имеет смысла, так как это бинарные данные. Для листовой диаграммы и ящика с усами сделаем разделение на политические режимы, чтобы более подробно ознакомиться с данными.

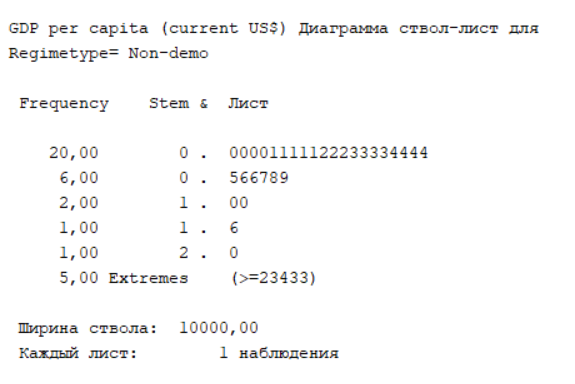
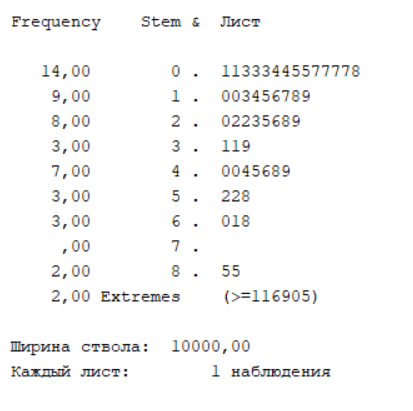
### **Листовые диаграммы**

1. Страна - численность населения



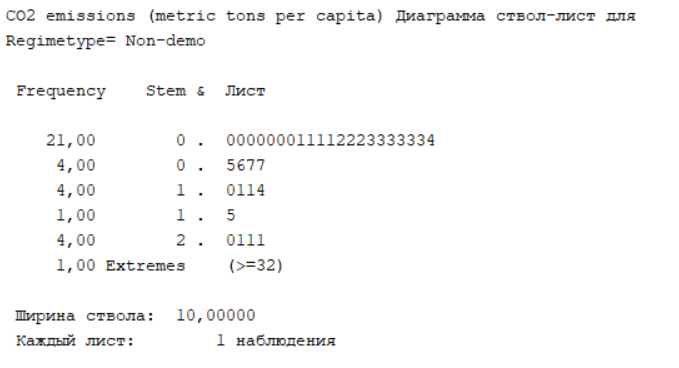
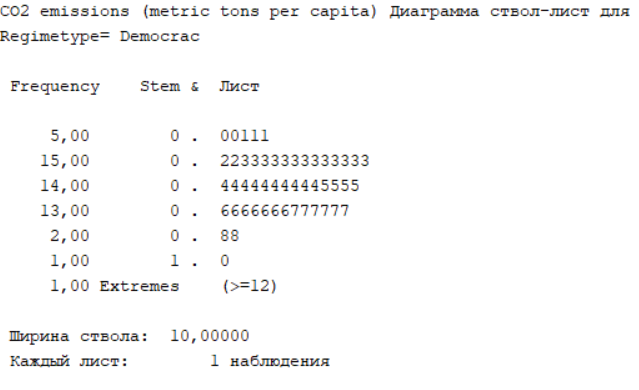
Можем заметить, что для демократических и недемократических стран, выбросы определились после 144 млн и 112 млн соответственно. Довольно близкие значения и они гораздо меньше значений в Китае и Индии.

1. Страна - ВВП на душу населения



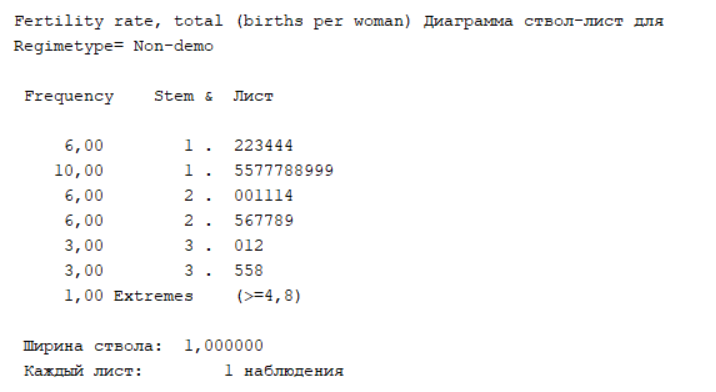
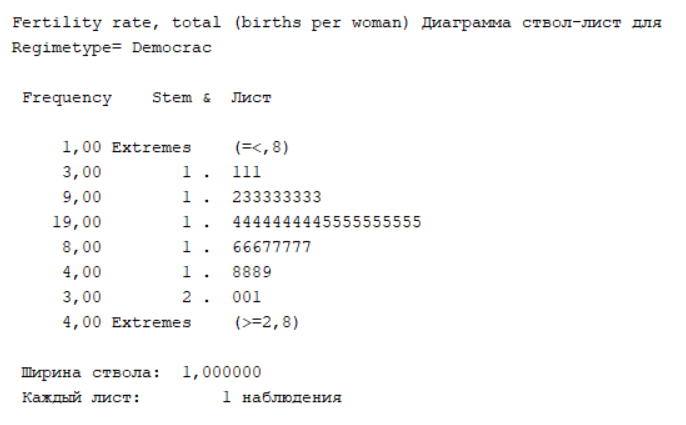
Видим, что у демократических стран выбросы начинаются с гораздо большего значения, чем у недемократических, что может быть очередным доказательством факта того, что демократические страны являются более зажиточными и богатыми по сравнению с не демократиями, но при этом мы не можем говорить о каузальности без дополнительных проверок

1. Страна - CO2 выбросы



Тут же наоборот у недемократических значение выше.

1. Страна - коэффициент фертильности

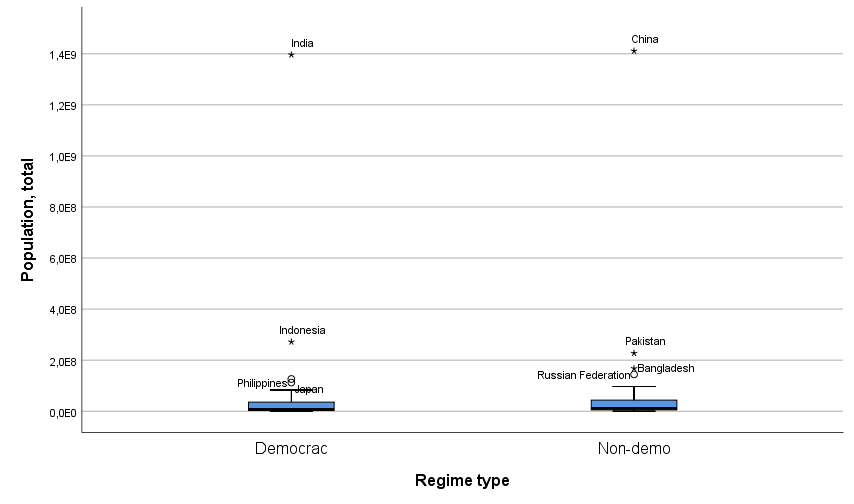


Наконец можем заметить выбросы и по нижней границе - у демократических стран есть значения меньшие 0,8 — это значит, что большинство женщин просто не рожают. Зато у недемократических верхние выбросы начинаются с большего значения. Можно сделать вывод, что в демократических странах рождаемость ниже.

### **Ящик с усами**

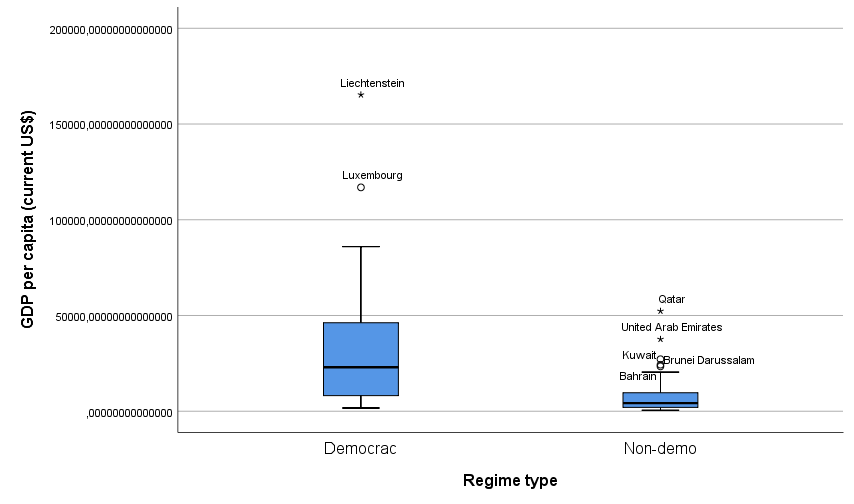
Теперь рассмотрим ящики с усами, где уже наглядно говорится о точных выбросах, рассчитанных через медиану и квартили.

1. Страна - численность населения



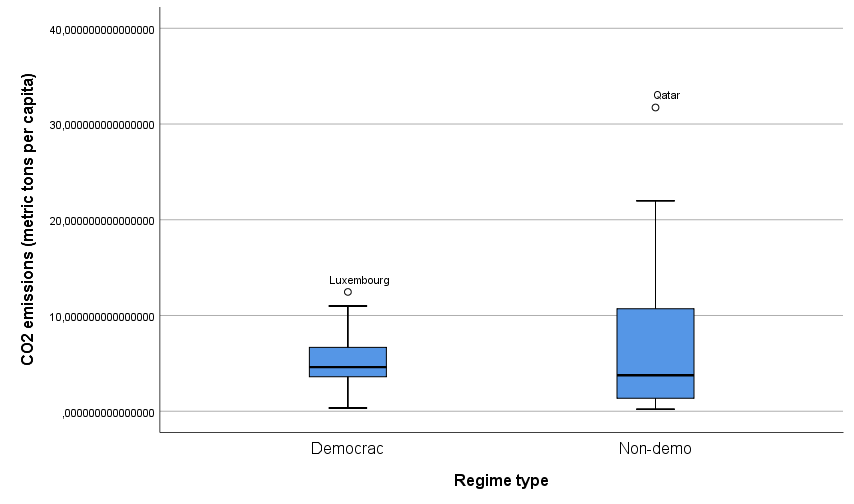
Как и заметили выше, Китай и Индия очень сильно отклоняются от других значений в своих группах, также уже можно и выделить другие выбросы - Индонезия, Филиппины, Япония, Россия, Бангладеш, Пакистан. Данные страны являются довольно важными наблюдениями по причине того, что они играют важную роль в международной экономике и политике, что вынуждает нас оставить эти государства в выборке, но при этом стоит помнить, что данные наблюдения не являются типичными.

1. Страна - ВВП



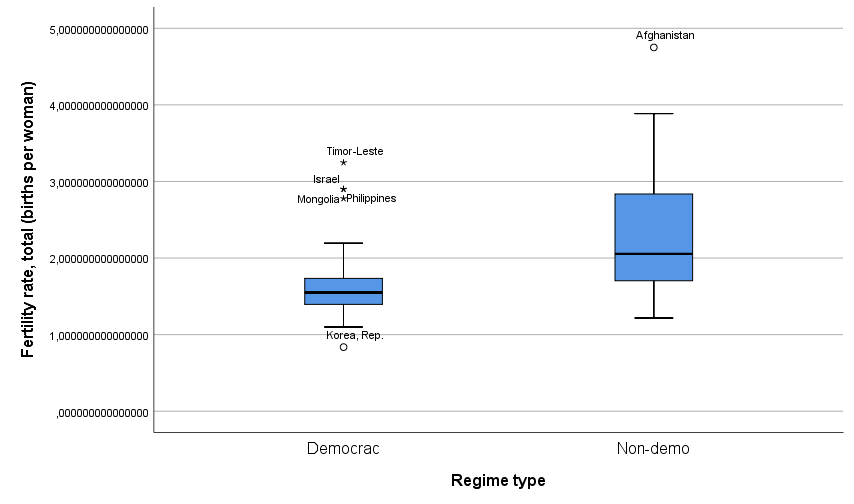
Также видим очень много выбросов. В недемократических странах это в арабские государства, в основном небольшие по населению, а в демократических - небольшие европейские государства. Можно сделать вывод, что чем меньше население, тем больше ВВП, что мы уже затрагивали ранее.

1. Страна - CO2 выбросы



Уже не так много выбросов, всего 2 страны, которые и были замечены еще на точечном графике. Эти страны входят в выбросы и по ВВП. Нам кажется, что правильным будет удалить их из анализа, так как это довольно маленькие государства, но они отклоняются сильно и в выбросах, и в ВВП, это может привести к неверным результатам, хотя страны не являются ведущими в мировой экономике и политике.

1. Страны - коэффициент фертильности

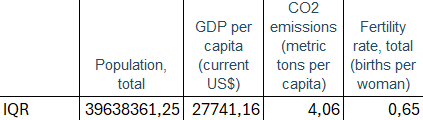
В данном случае видно, что в странах, относимых Freedom House к недемократическим, рождаемость в среднем заметно выше, чем в демократических, но при этом стоит отметить, что среди демократий есть заметное число выбросов, а именно стран с высокой рождаемостью, к которым относятся Монголия, Филиппины, Израиль и Восточный Тимор.

### **Характеристики СВ**

****

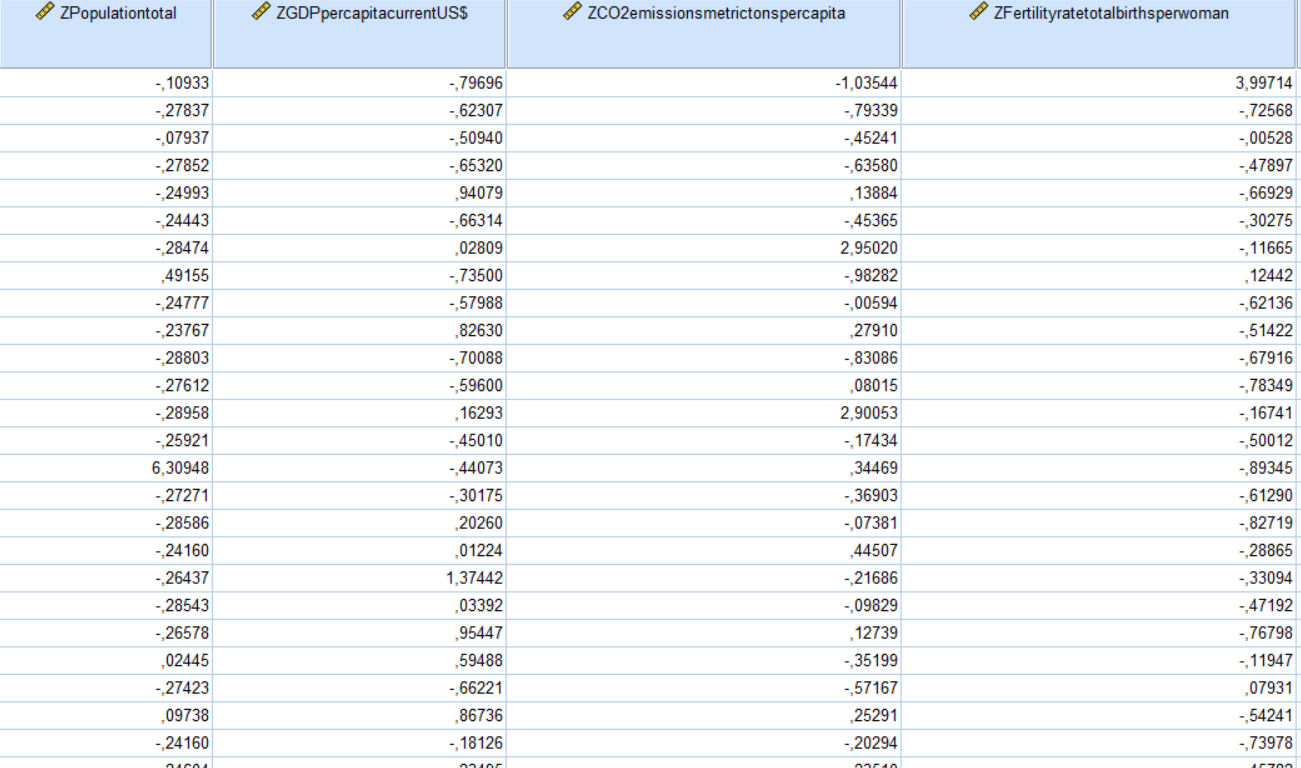
Дескриптивная статистика для набора данных

Рассчитаем межквартальную разницу IQR по данным выше - нужно отнять из 75%ого значения 25%ое значения, то есть Q3 - Q1.



### **Z-преобразование**

С помощью функционала SPSS рассчитаем z-оценку для всех числовых признаков. Часть результата можно видеть ниже



Теперь, чтобы понять, где начинаются выбросы, нужно построить график отсортированных z - score.

1. Численность населения.

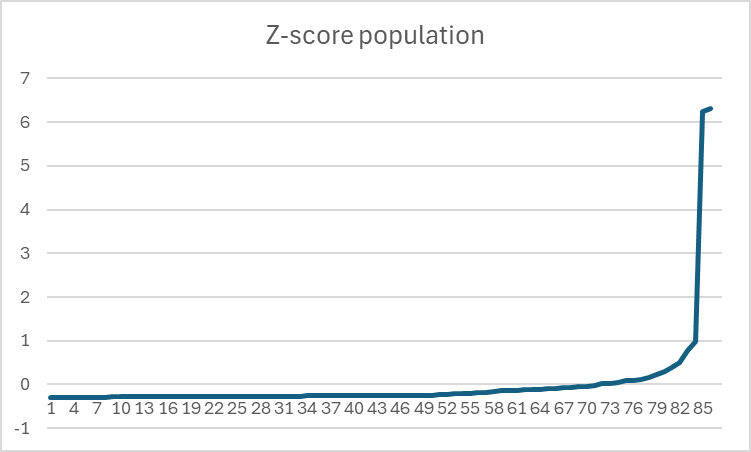
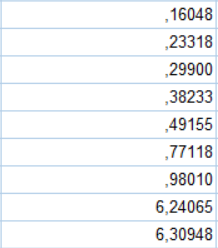
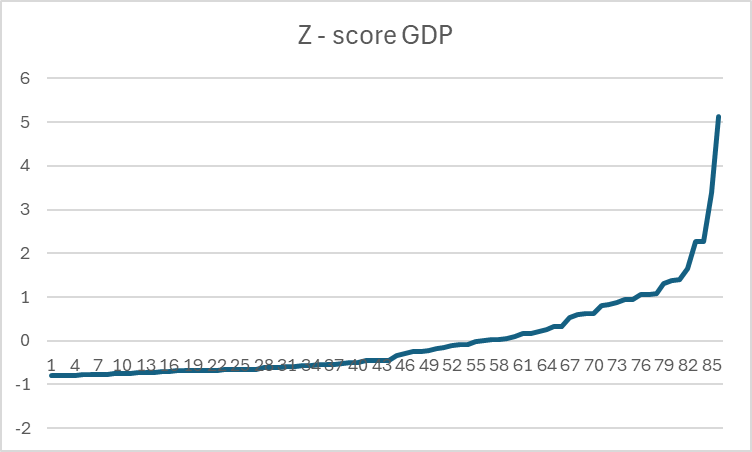


График показывает перегиб на примерно 79 стране в отсортированном списке по Zpopulation.



Можно заметить, что начиная с 79 страны как раз начинаются выбросы, которые были найдены по графикам. Именно в этом месте и происходит перегиб, а также начинаются большие скачки в значениях.

1. ВВП на душу населения

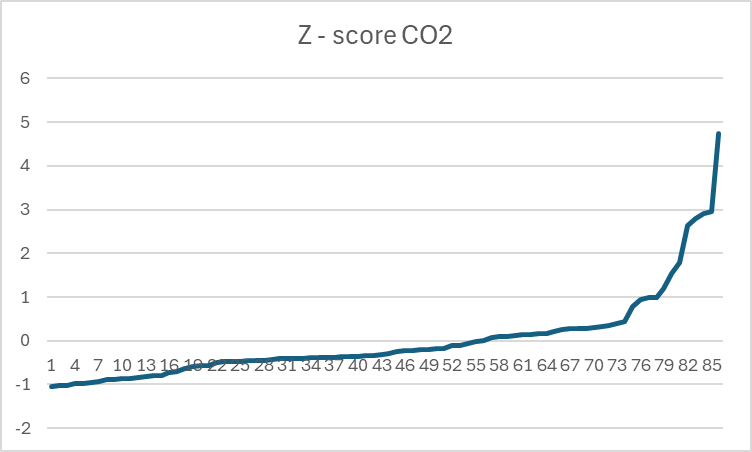


Сильный перегиб начинается где-то с 79 страны.



По значениям видим сильный разрыв после 78 страны. Я бы отнесла к выбросам страны, начиная с 79 номера. Но анализ по графикам показал нам другое — это легко объясняется, так ка на графиках мы смотрели распределение по группам, а тут в целом.

1. CO2 выбросы

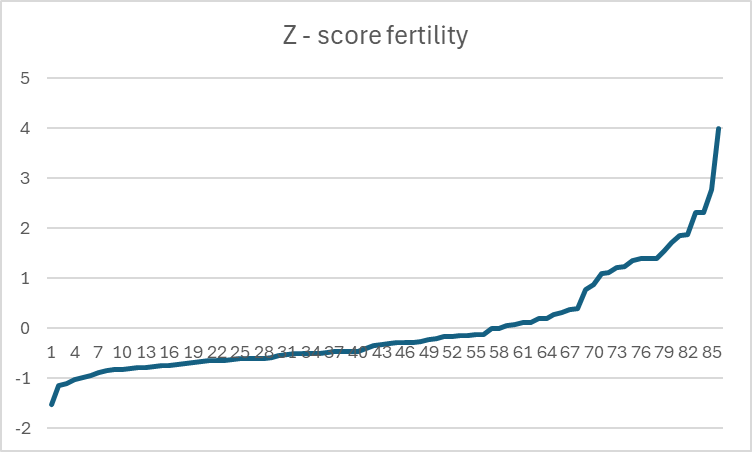


Видим перегиб на 77 номере.

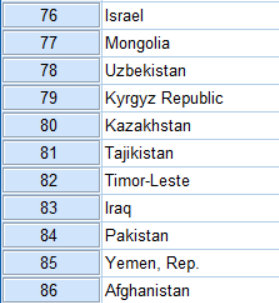


Сильное расхождение в значениях происходит после 81 номера. Так что правильным будет выбрать значения с 82 включительно как выбросы. Оно как раз находится на перегибе.

1. Коэффициент фертильности



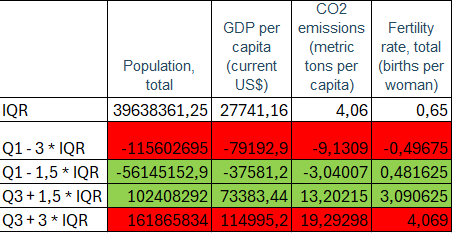
Видим перегиб где-то с 80 номера, посмотрим на значения



Видим большое расхождение в значениях с 82 номера. Выделим как выбросы - страны с 83 номера включительно.

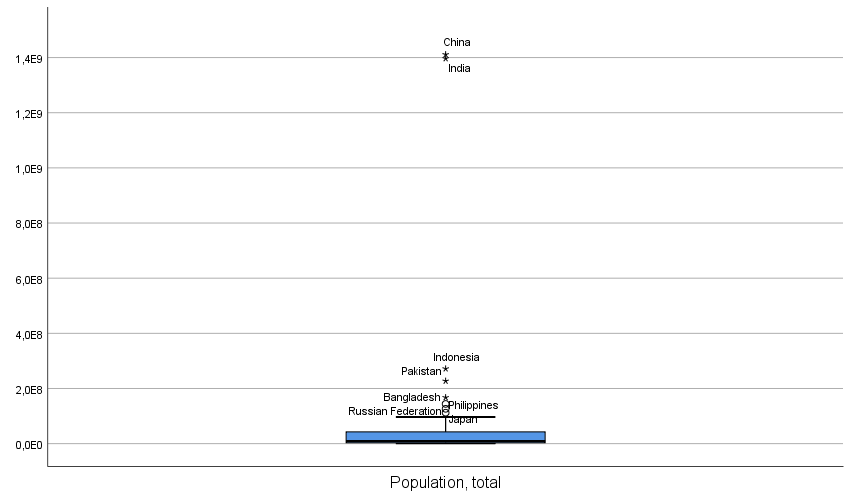
### **1,5 и 3 IQR**

В SPSS значения 1,5 и 3 IQR как раз используются в диаграмме ящик с усами. Круглыми точками показаны значения, которые выше Q3 + 1,5 \* IQR, но ниже Q3 + 3 \* IQR. Называются они мягкими выбросами, а все что выше Q3 + 3 \* IQR называется экстремальными выбросами. Аналогично с нижней границей (берем только Q1 и знак минус). Давайте рассчитаем граничные значения, а также еще раз посмотрим на ящики с усами, чтобы понять, какие из выбросов экстремальны.



Можем сразу заметить, что экстремальных выбросов по нижней границе нет нигде, так как значения не могут быть отрицательными, по этой же причине в первых трех признаках нет и мягких выбросов.

По данным границам можно выявить, что экстремальными выбросами по численности населения являются:

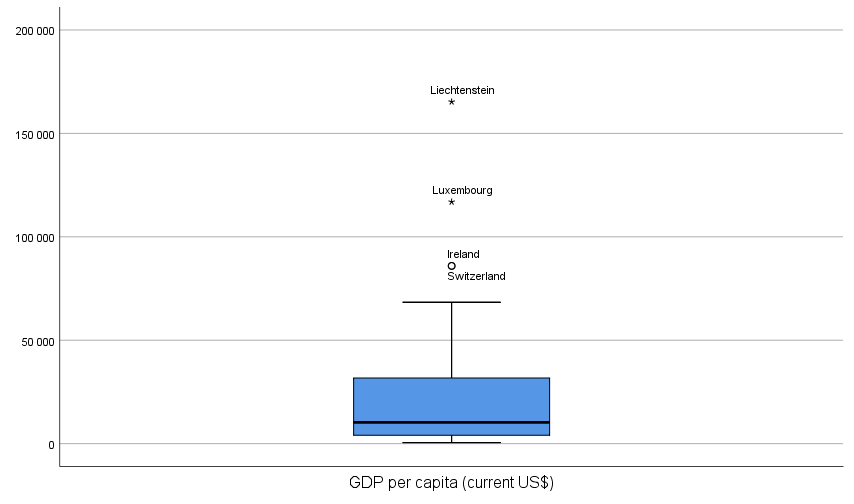


Китай, Индия, Пакистан, Бангладеш, Индонезия. Мягкими - Россия, Филиппины и Япония. Это значит, что экстремальных выбросов почти 6%. Так как эти страны важны для нас в экономическом плане, то мы решили не удалять данные выбросы из исследования.



Можем заметить, что при удалении экстремальных выбросов, медиана почти не изменятся, поэтому если мы оставляем выбросы, то просто будем опираться на медиану, а не среднее.

Теперь рассмотрим экстремальные выбросы в ВВП

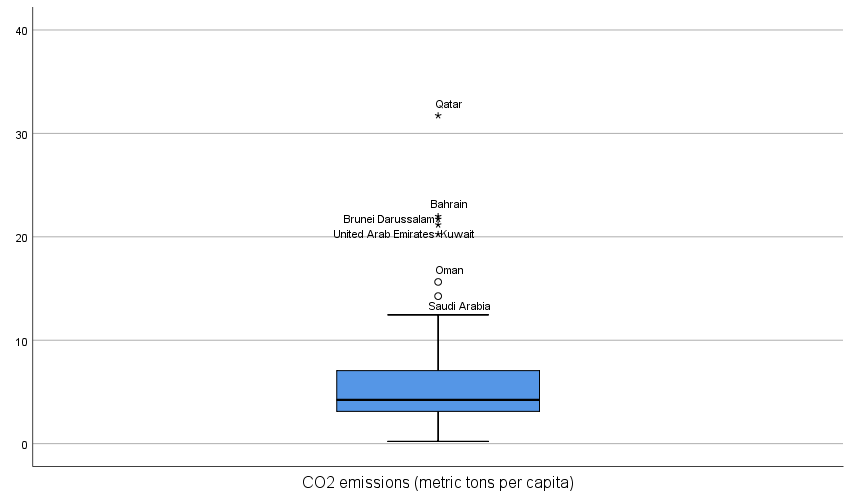


Здесь экстремальными выбросами являются две страны - Лихтенштейн и Люксембург. Это всего 2,3% стран, что не так много в общей массе. Давайте посмотрим на описательную статистику с ними и без них, чтобы точнее определиться.



Видим как сильно меняются стандартное отклонение и среднее. Медиана тут также несильно зависит от выбросов, но так как их количество не такое большое, а на мир в целом эти страны оказывают небольшое влияние, так как их размеры небольшие, то мы удалим данные экстремальные выбросы, чтобы получить более точную аналитику.

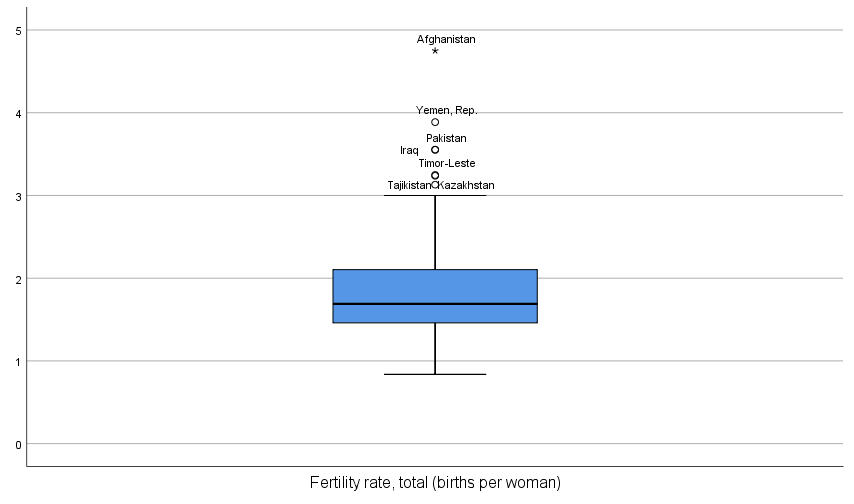
Перейдем к CO2 выбросам.



Довольно много стран оказались за границей: Катар, Бахрейн, Бруней, ОАЭ и Кувейт. Все страны арабские. Можно сделать вывод, что в арабских странах люди больше всего доставляют проблем воздуху и природе. Данный факт может быть объяснен особенностями энергопотребления в данных странах, которые в силу географических условий используют лишь “грязные” виды топлива для выработки электроэнергии.

Так как в данной ситуации мы уже сделали первый вывод, то в дальнейшем анализе данные выбросы могут предоставить ложные результаты, поэтому было принято решение удалить данные страны из анализа.

И последнее - коэффициент фертильности.



Здесь в экстремальных только Афганистан. Это общеизвестный факт, что там высокая рождаемость. Вызвано это плохими медициной и уровнем жизни. Женщина может много рожать, но выживают не все, а также Афганистан одно из государств с самой низкой продолжительностью жизни. В мягких выбросах мы также можем заметить страны с похожей ситуацией. Оставим мягкие выбросы, так как их довольно много, но удалим Афганистан из анализа, так как его данные могут испортить результаты, а большинство стран имеют более благоприятный уровень жизни.

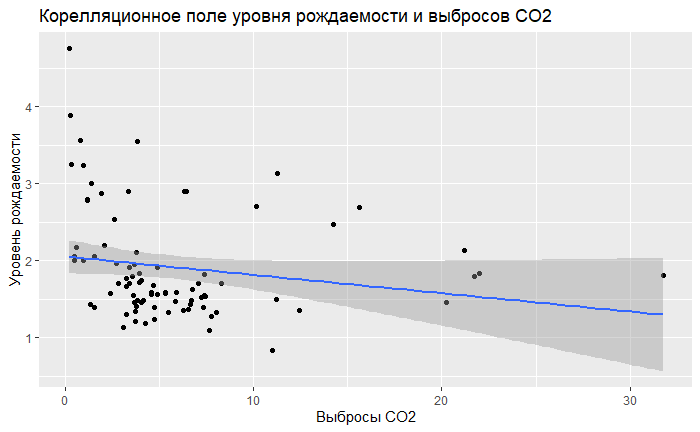
После проведения анализа выбросов можно сделать вывод, что из дальнейшего анализа стоит исключить монархии Персидского залива, а также Бруней (в целом с ними схожий) и Афганистан. Данные случаи могут быть проанализированы в рамках отдельного сравнительного исследования так как опыт и условия данных государств сильно отличается от средних тенденций, зафиксированных нами в ходе предварительного анализа. Таким образом количественное исследование помогло нам избавить выборку от экстремальных выбросов и сформировать пул стран, не сильно отклоняющихся от общих тенденций.

## **Корреляционный анализ**

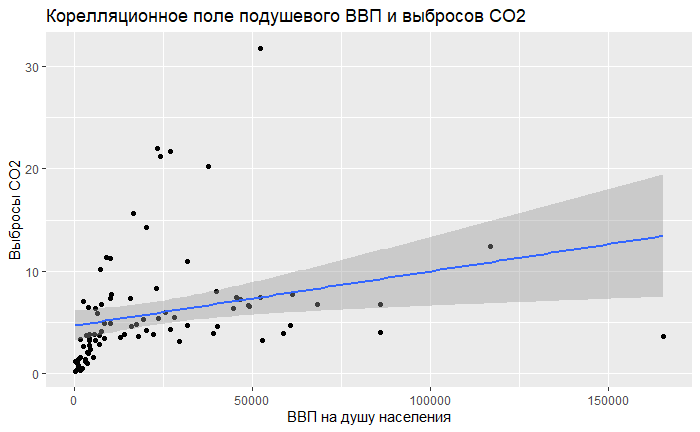
В следующей фазе работы будет выполнен корреляционный анализ, который направлен на изучение связей между двумя или более переменными. Эта фаза будет исследована через два сценария: анализ данных, включая выбросы, и анализ данных после исключения выбросов.

В качестве переменных для корреляционного анализа были выбраны данные о населении страны, уровне ВВП в столице, уровне рождаемости, выбросах углекислого газа в столице. Представленные данные могут быть полезны для понимания общей информации о стране, так как они отображают ее густонаселенность, чистоту воздуха, состояние экономики, а также привлекательность для построения семьи.

В начале процесса строятся корреляционные карты для каждой из переменных. Первоначально представляются графики, в которых учтены выбросы, с нормализованными данными.



На диаграмме представлено корреляционное поле, иллюстрирующее взаимосвязь между выбросами CO2 и коэффициентом рождаемости в различных странах. На графике рассеяния отображены точки, каждая из которых соответствует отдельной стране. Линия тренда, проходящая среди точек, показывает общее направление корреляции между уровнем выбросов CO2 и коэффициентом рождаемости: страны с более высокими выбросами углекислого газа имеют тенденцию к более низкому коэффициенту рождаемости.



На представленной диаграмме рассеяния изображено корреляционное поле, которое демонстрирует отношение между выбросами CO2 и уровнем валового внутреннего продукта (ВВП) на душу населения. Точки на графике представляют конкретные значения каждой страны. Из графика можно наблюдать распределение точек, которое предоставляет визуальное представление о потенциальной корреляции между двумя рассматриваемыми переменными. Благодаря линии тренда можно увидеть тенденцию в данных. Исходя из графика можно определить характер корреляции. На графике заметна средняя положительная связь между уровнем подушевого ВВП и выбросами углекислого газа.

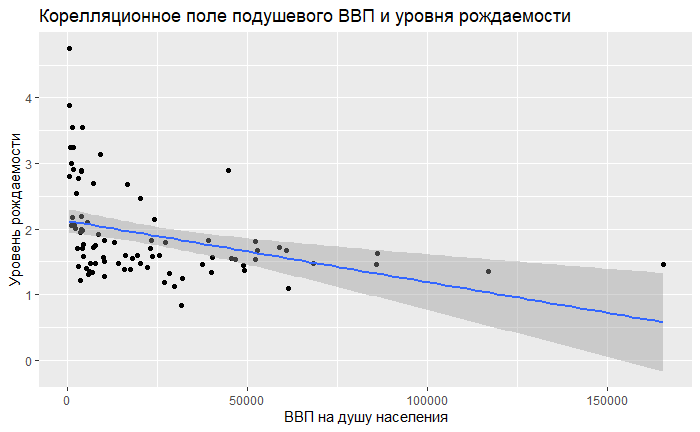
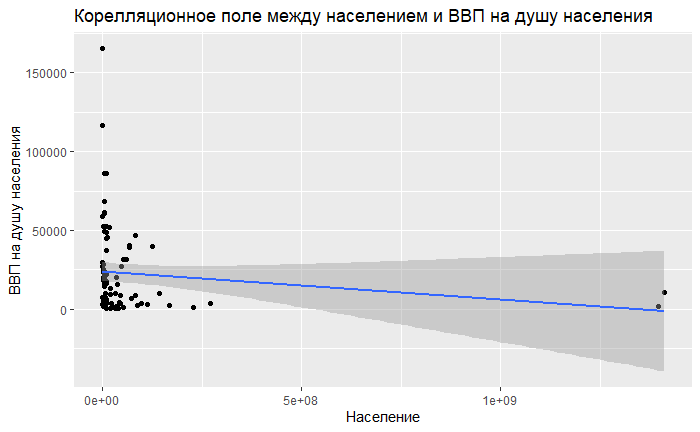


График представляет собой корреляционное поле, показывающее взаимосвязь между уровнем ВВП и уровнем рождаемости в различных странах. На графике разбросаны точки, каждая из которых соответствует определенной стране, с координатами, основанными на уровне ВВП и уровне рождаемости этой страны. Линия, проходящая через точки, представляет собой линию тренда, которая помогает визуализировать общую тенденцию корреляции между двумя переменными. Линия тренда направлена вниз, что может указывать на среднюю отрицательную корреляцию между уровнем ВВП и уровнем рождаемости: страны с более высоким уровнем ВВП имеют тенденцию к более низкой рождаемости. Однако точки достаточно равномерно распределены вокруг линии тренда, что указывает на то, что взаимосвязь не является сильной.



Этот график является корреляционным полем, показывающим взаимосвязь между населением столиц различных стран и уровнем ВВП этих стран. На графике разбросаны точки, каждая из которых соответствует отдельной стране, масштаб данных предполагает, что рассматриваются общие экономические показатели страны в целом. Линия тренда указывает на общую тенденцию в данных, которая, судя по визуальной оценке, показывает отрицательную корреляцию между населением страны и её ВВП - страны с большим населением часто имеют более низкий ВВП(Китай).  
Стоит отметить,что на графике присутствуют выбросы - две точки, находящиеся вдалеке от основной массы данных, с существенно большим населением и сравнительно низким ВВП,а также точки с слишком высоким ВВП относительно других стран и низким уровнем населения. Включение или исключение этих выбросов может существенно повлиять на расчетные показатели корреляции.

## 

В данной матрице парных коэффициентов корреляций представлены значения корреляций для четырех различных переменных: население, ВВП на душу населения, уровень CO2 выбросов и уровень рождаемости. Коэффициент корреляции может варьироваться от -1 до 1, где -1 указывает на идеальную обратную корреляцию, 0 означает отсутствие корреляции, а 1 указывает на идеальную прямую корреляцию между двумя переменными.  
Согласно матрице можно выделить следующие зависимости между переменными:

* Население и ВВП на душу населения: -0.1359068 (небольшая обратная корреляция)
* Население и уровень CO2 выбросов: -0.07121432 (очень слабая обратная корреляция)
* Население и уровень рождаемости: -0.01887726 (очень слабая обратная корреляция)
* ВВП на душу населения и уровень CO2 выбросов: 0.2700497 (слабая прямая корреляция)
* ВВП на душу населения и уровень рождаемости: -0.36686421 (умеренная обратная корреляция)
* Уровень CO2 выбросов и уровень рождаемости: -0.18300673 (слабая обратная корреляция)

Значения по диагонали равны 1, что означает идеальную прямую корреляцию переменной с самой собой, что является логичным, так как любая переменная полностью коррелирует сама с собой

## 

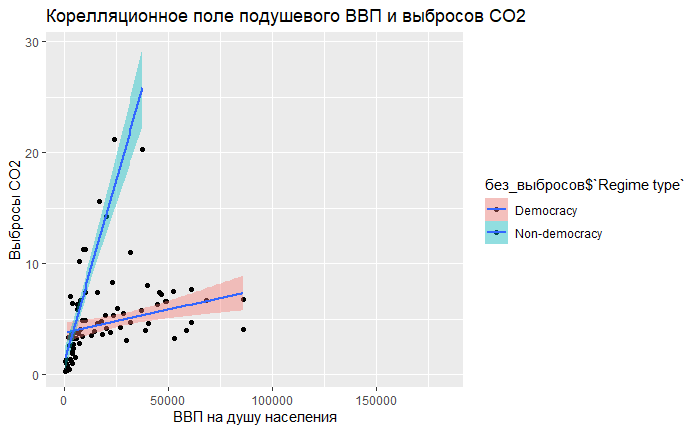
С помощью данного графика можно более наглядно увидеть корреляцию переменных и понять ее масштабность. Чем больше круг и ярче его цвет, тем более существенно корреляцию между двумя признаками.

## 

## 

На первом графике изображено корреляционное поле уровня рождаемости и выбросов CO2. На графике присутствуют две группы данных, разделённые по типу политического режима: демократия и другие политические режимы. Каждая группа данных имеет свою линию тренда и доверительный интервал, представленный цветной заливкой вокруг линии тренда. Точки представляют отдельные данные, а линии тренда показывают общую тенденцию. В группе "Демократия" виден негативный тренд (синий цвет), то есть уровень рождаемости снижается с увеличением выбросов CO2. В группе иных политических режимов тренд также негативный (красный цвет), но с меньшим углом наклона, что может свидетельствовать о более слабой связи между уровнем рождаемости и выбросами CO2 в недемократических странах.

Сравнивая представленный график с тем, который содержит в себе выбросы, можно заметить, что обе диаграммы показывают отрицательную корреляцию между уровнем рождаемости и выбросами CO2. Вдобавок, большая часть значений на диаграмме без выбросов относится именно к странам с демократическим политическим режимом.



На графике изображено корреляционное поле подушевого ВВП и выбросов CO2 с разделением данных на две категории в зависимости от типа политического режима стран: демократии (синий цвет) и недемократическим (красный цвет). Ось X представляет подушевой ВВП (Валовой внутренний продукт на душу населения), а ось Y отображает выбросы CO2 на душу населения.

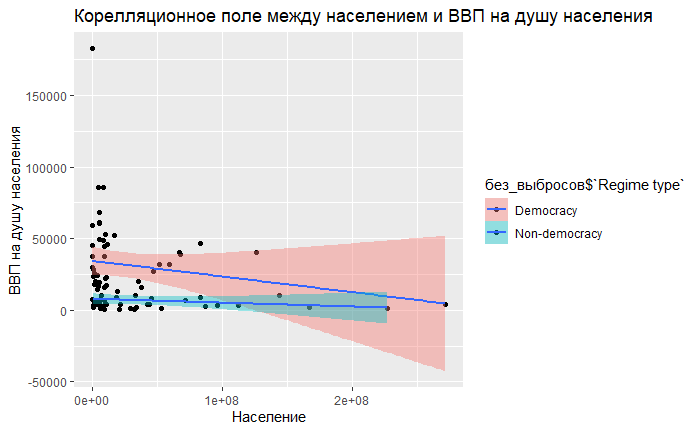
Обе группы показывают положительную корреляцию между подушевым ВВП и выбросами CO2, что указывает на то, что более высокий уровень экономического развития (измеряемый подушевым ВВП) связан с более высоким уровнем выбросов CO2 на душу населения. Тренды отражаются в виде линий тренда для каждой категории, с доверительным интервалом, представленным цветной областью вокруг линий.

Синяя линия тренда для демократических стран показывает более крутой наклон, что может свидетельствовать о более сильной зависимости между экономическим ростом и увеличением выбросов CO2 в этих странах по сравнению с недемократическими, чья красная линия тренда имеет меньший наклон.

На графике без учета выбросов наблюдалась тенденция более высоких выбросов CO2 в странах с более высоким подушевым ВВП независимо от типа политического режима. Однако демократические страны показали более крутой наклон, что может указывать на более сильное увеличение выбросов CO2 по мере роста ВВП в демократических странах по сравнению с недемократическими.

На графике без выбросов можно было интерпретировать, что демократические страны, возможно, сталкиваются с более высокими выбросами CO2 на единицу ВВП, что может отражать их более интенсивный промышленный рост или более высокий уровень потребления энергии на душу населения. На новом графике такая специфика отсутствует, и мы видим более обобщенную картину.

.



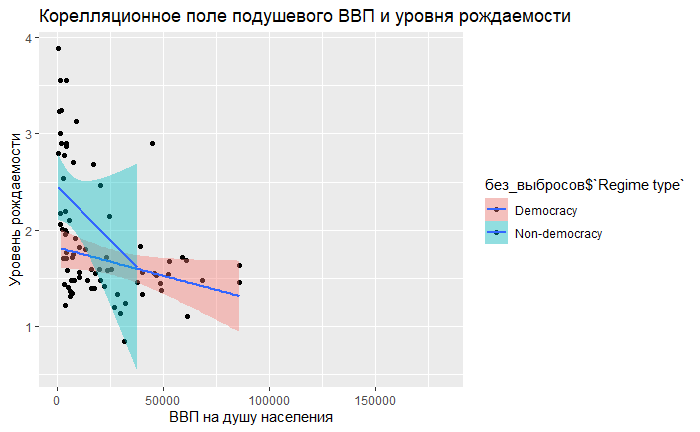
На этом графике изображено корреляционное поле между населением страны и ВВП на душу населения с разделением на две категории: демократии (синий цвет) и недемократическими (красный цвет).

Линии тренда для обеих категорий показывают негативную корреляцию: по мере увеличения населения ВВП на душу населения уменьшается. Однако наклон линии тренда для недемократических стран (красная линия) кажется более крутым, что может указывать на более сильную отрицательную зависимость между этими двумя переменными в недемократических странах по сравнению с демократическими (синяя линия).

Доверительные интервалы для каждой категории представлены цветными областями вокруг соответствующих линий тренда, показывая возможный диапазон значений тренда в зависимости от неопределенности данных.

Этот график может указывать на то, что в странах с большим населением сложнее достичь высокого ВВП на душу населения, и что этот эффект может быть более выраженным в недемократических странах.

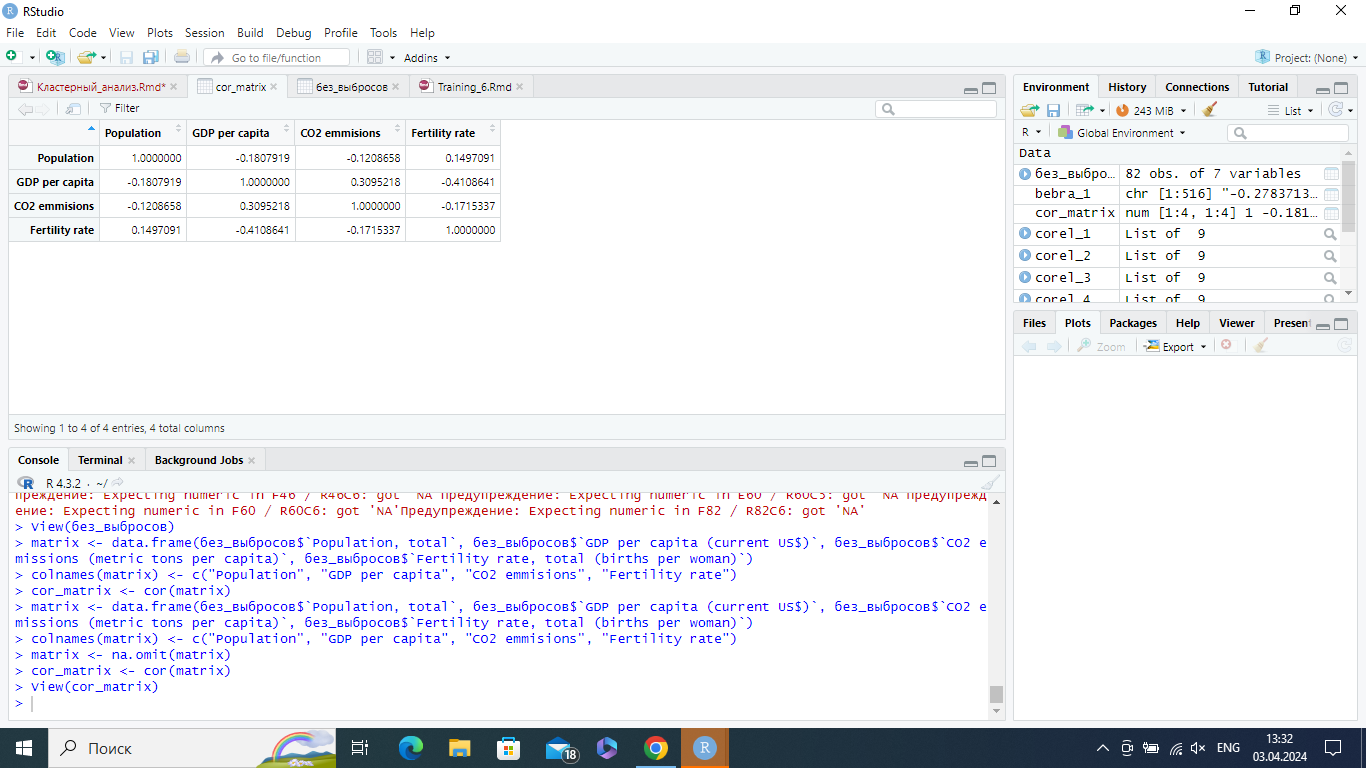
График Корреляционного поля без выбросов может использоваться для анализа взаимосвязи в контексте политического устройства, в то время как график корреляционного поля с выбросами — для общего понимания тенденций вне зависимости от политической системы страны.



На графике представлена корреляция между подушевым ВВП и уровнем рождаемости с разделением на страны с демократическим режимом (розовый цвет) и недемократическим (голубой цвет). Обе категории показывают отрицательную корреляцию: с увеличением ВВП на душу населения уровень рождаемости уменьшается.

На обоих графиках с учетом и без учета выбросов можно заметить отрицательную корреляцию между населением страны и подушевым ВВП.

График без выбросов предоставляет более детальное представление о том, как уровень рождаемости может различаться в зависимости от политической системы страны при одном и том же уровне подушевого ВВП, тогда как график с выбросами представляет обобщенный вид, объединяя все страны в один набор данных.  
На графике без выбросов можно интерпретировать, что взаимосвязь между подушевым ВВП и уровнем рождаемости может быть различной в зависимости от политического устройства страны. На графике с выбросами такое сравнение сделать нельзя.  
Таким образом, график корреляционного поля без выбросов может быть полезен для анализа влияния политических систем на демографические показатели, в то время как график корреляционного поля с выбросами полезен для общего понимания тенденции между экономическим развитием и рождаемостью.



Исходя из данной матрицы парных значений корреляций можно сделать следующие выводы о зависимости переменных друг с другом

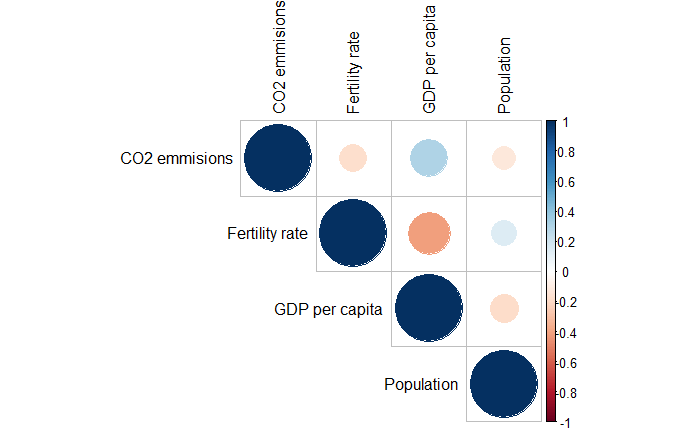
* Корреляция между населением и ВВП на душу населения: -0.1807919 (слабая обратная связь).
* Корреляция между населением и CO2 выбросами: -0.1208658 (очень слабая обратная связь).
* Корреляция между населением и уровнем рождаемости: 0.1497091 (слабая прямая связь).
* Корреляция между ВВП на душу населения и CO2 выбросами: 0.3095218 (умеренная прямая связь).
* Корреляция между ВВП на душу населения и уровнем рождаемости: -0.4108641 (умеренная обратная связь).
* Корреляция между CO2 выбросами и уровнем рождаемости: -0.1715337 (слабая обратная связь).

Эти коэффициенты корреляции позволяют увидеть, как каждая из этих переменных связана с другими. Например, отрицательная корреляция между ВВП на душу населения и уровнем рождаемости может указывать на то, что в странах с более высоким уровнем дохода на душу населения обычно наблюдается более низкий уровень рождаемости.

Говоря о различиях парных коэффициентов корреляции матрицы без учета выбросов по сравнению с матрицей, учитывающей выбросы, можно сделать следующие выводы:

* Отрицательная корреляция между населением и ВВП на душу населения (-0.1359068) менее выражена по сравнению с матрицей, учитывающей выбросы. Это говорит о том, что выбросы оказывали влияние на отношение населения к ВВП
* Корреляция между населением и CO2 выбросами (-0.07121432) также менее выражена. Выбросы оказали влияние, но на незначительном уровне
* Отрицательная корреляция между населением и уровнем рождаемости (-0.01887726) указывает на отсутствие сильной связи. Более того, она стала отрицательной, изменив характер изменения населения на уровень рождаемости.
* Положительная корреляция между ВВП на душу населения и CO2 выбросами (0.27004968) немного ниже, чем в матрице, учитывающей выбросы.
* Отрицательная корреляция между ВВП на душу населения и уровнем рождаемости (-0.36686421) тоже менее сильная, но по-прежнему умеренная.
* Отрицательная корреляция между CO2 выбросами и уровнем рождаемости (-0.18300673) более сильная, чем в матрице, учитывающей выбросы.

Изменение коэффициентов корреляции между двумя матрицами может быть связано с различными выборками данных, изменениями в данных со временем или использованием разных методов расчета коэффициентов корреляции.



Данный график даёт возможность наглядно представить масштабность значение корреляции между переменными попарно.

### **Множественная регрессия**

С помощью функционала SPSS мы провели множественный регрессионный анализ. Получились следующие данные:

****

Множественный коэффициент корреляции равен 0,127, что говорит о слабой связи, то есть все признаки разом имеют слабую линейную связь с выбросами CO2.

Также были рассчитаны коэффициенты для итогового уравнения.



Уровень значимости был выбран 0,05. Взглянем на значимость коэффициента корреляции - 0,033. Это значит, что наше уравнение в целом значимо. Рассмотрим значимость отдельно для коэффициентов. Видим, что значим только коэффициент при ВВП, остальные коэффициенты оказались незначимы.

### **Вывод**

По итогу проведенного анализа мной замечены следующие тенденции. Между большинством из выбранных показателей если и присутствует корреляционная связь, то незначительная (с небольшим уровнем значимости большинства коэффициентов). Категоризация на политические режимы помогла мне определить разный характер корреляции для этих групп стран и тренды зачастую сильно разнились даже после исключения аномальных наблюдений. Множественный регрессионный анализ в данном случае не является полезным инструментом, так как он направлен на выявление каузации при наличии значимой корреляции, в то время как полученные результаты таковыми не являются. Тем не менее выявленные тенденции помогли задать направление для дальнейших исследований, которые будут использовать если не количественные методы, то качественные.

## **Кластерный анализ**

### **Иерархическая кластеризация**

Построение и анализ дендрограмм.

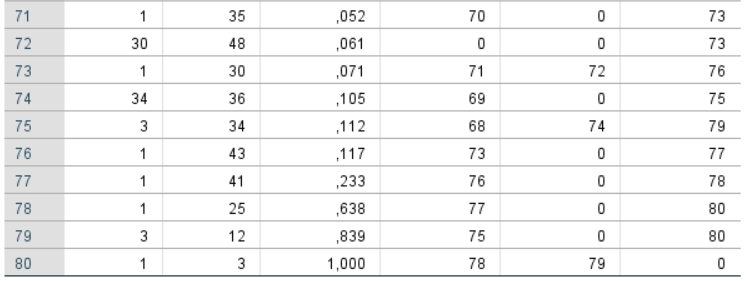
Для начала была выбрана меру расстояния - квадрат евклидового расстояния, так как имеем и интервальные переменные, и бинарные.

Рассмотрим 4 метода иерархической кластеризации, чтобы определить оптимальное количество кластеров.

1. Первый выбранный нами метод - метод ближайшего соседа.

Расставим все нужные, вышеописанные настройки в SPSS, и выберем стандартизацию - от 0 до 1, так как мы удалили не все выбросы и, например, при Z преобразовании численности населения появляются значения гораздо большие 3. Перед этим мы преобразовали Democracy и Non-democracy в 1 и 0 соответственно.

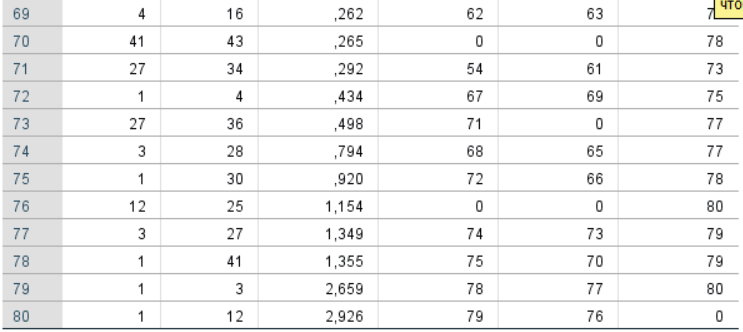
По результатам агломерации можно определить оптимальное количество кластеров. Посмотрим на коэффициенты агломерации под конец процесса. Обратимся к 3 столбцу



Видим большой скачок между 0,233 и 0,638, то есть на 77 шаге. Тогда оптимальное количество это 81 - 77 = 4 кластера.

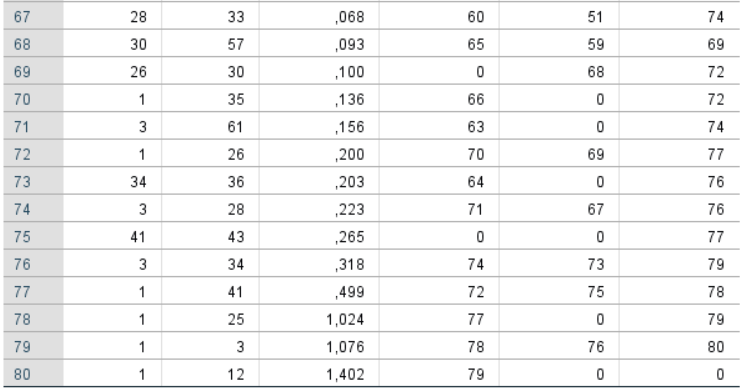
1. Теперь рассмотрим метод дальнего соседа.

Остальные настройки оставим такими же. Рассмотрим коэффициенты агломерации.



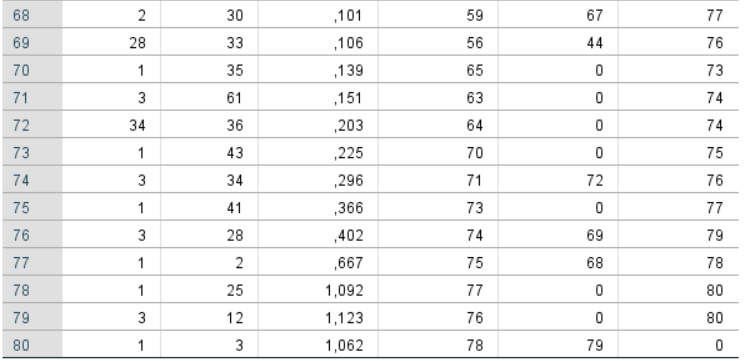
Тут видим скачок на 78 шаге в два раза. То есть оптимальное количество кластеров = 81 - 78 = 3.

1. При методе центроидной кластеризации получаем такие коэффициенты:



Скачок на 77 шаге. То есть 4 кластера

1. Последний метод, который был рассмотрен — это метод медианной кластеризации. Напомним, что в некоторых признаках, мы не удаляли выбросы, а значит договорились использовать именно медиану.

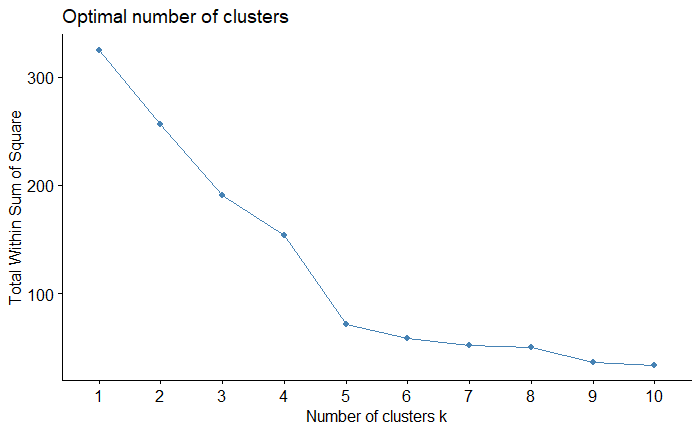


Видим скачок на 77 шаге, что опять же говорит о 4 кластерах.

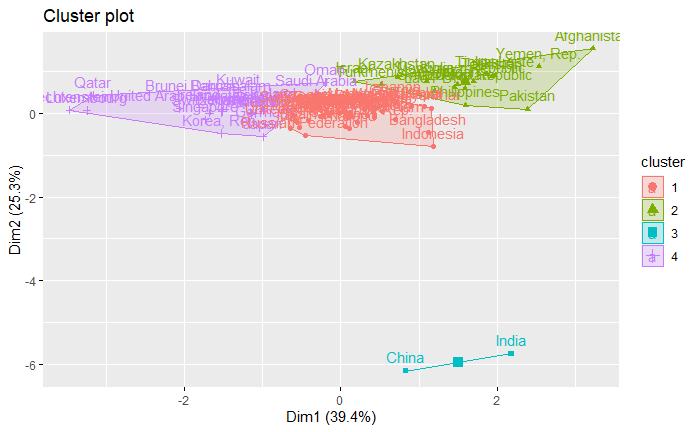
4 полученные дендрограммы, на которых также можно заметить скачки и как именно страны собирались в группы более наглядно, лежат в приложении расчетов с SPSS.

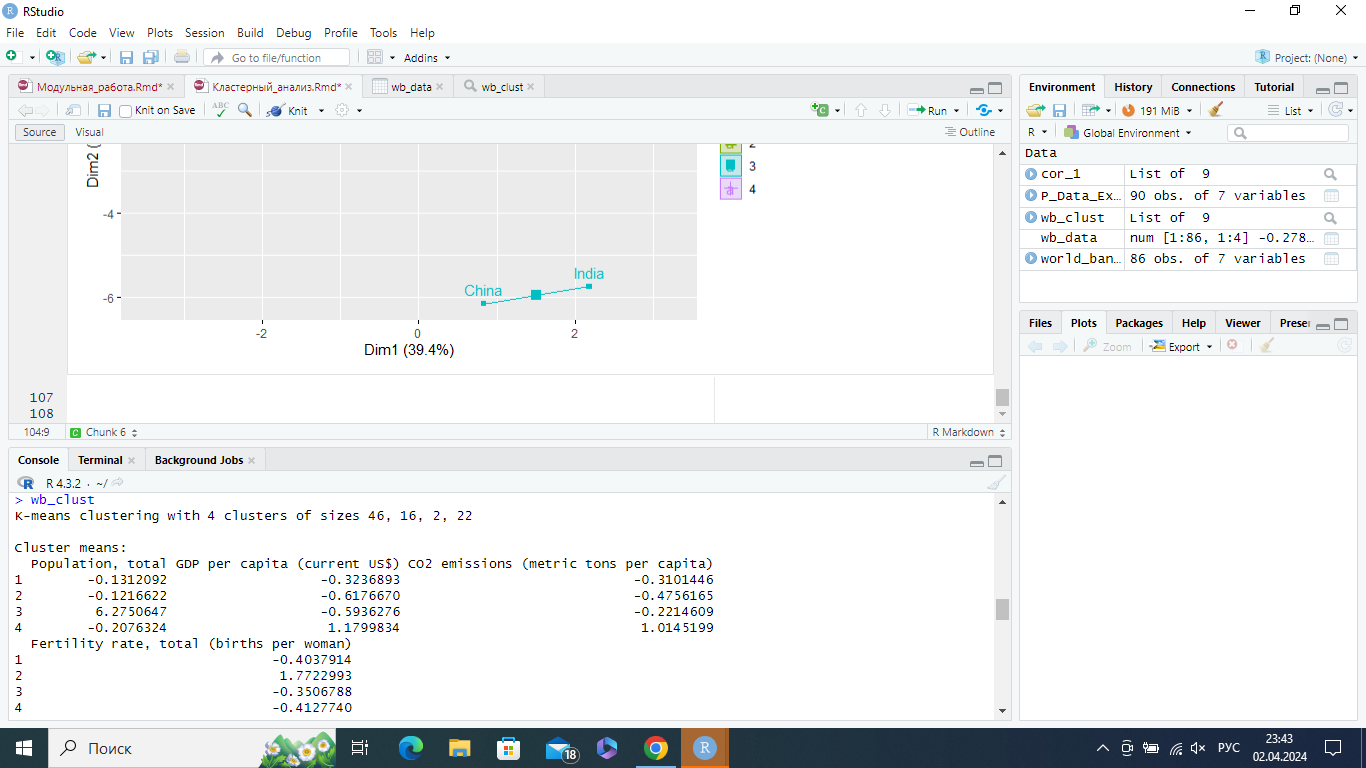
**Вывод:** большинство методов показали, что оптимальное количество кластеров - 4. Теперь с помощью этой информации можно провести анализ k средних.

### **Метод k-средних**

Для проверки построенных кластерных моделей было проведено ещё одно кластерное деление с помощью языка R метода K-средних.****

Несмотря на то, что предыдущие построения показали оптимальное число кластеров равное четырем мы провели дополнительный тест для выявления оптимального числа кластеров, с помощью “метода локтя” нами было обнаружено то же оптимальное число кластеров равное четырем

****



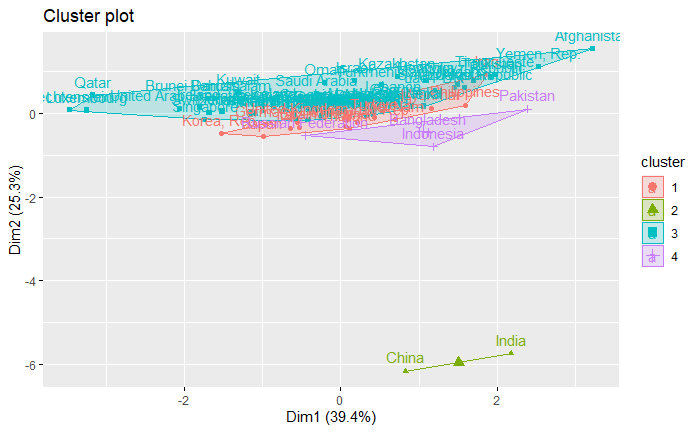
На данном графике видно распределение наблюдений по кластерам, а также средние значения кластеров по каждой из переменных, участвовавших в анализе (стандартизированные данные).

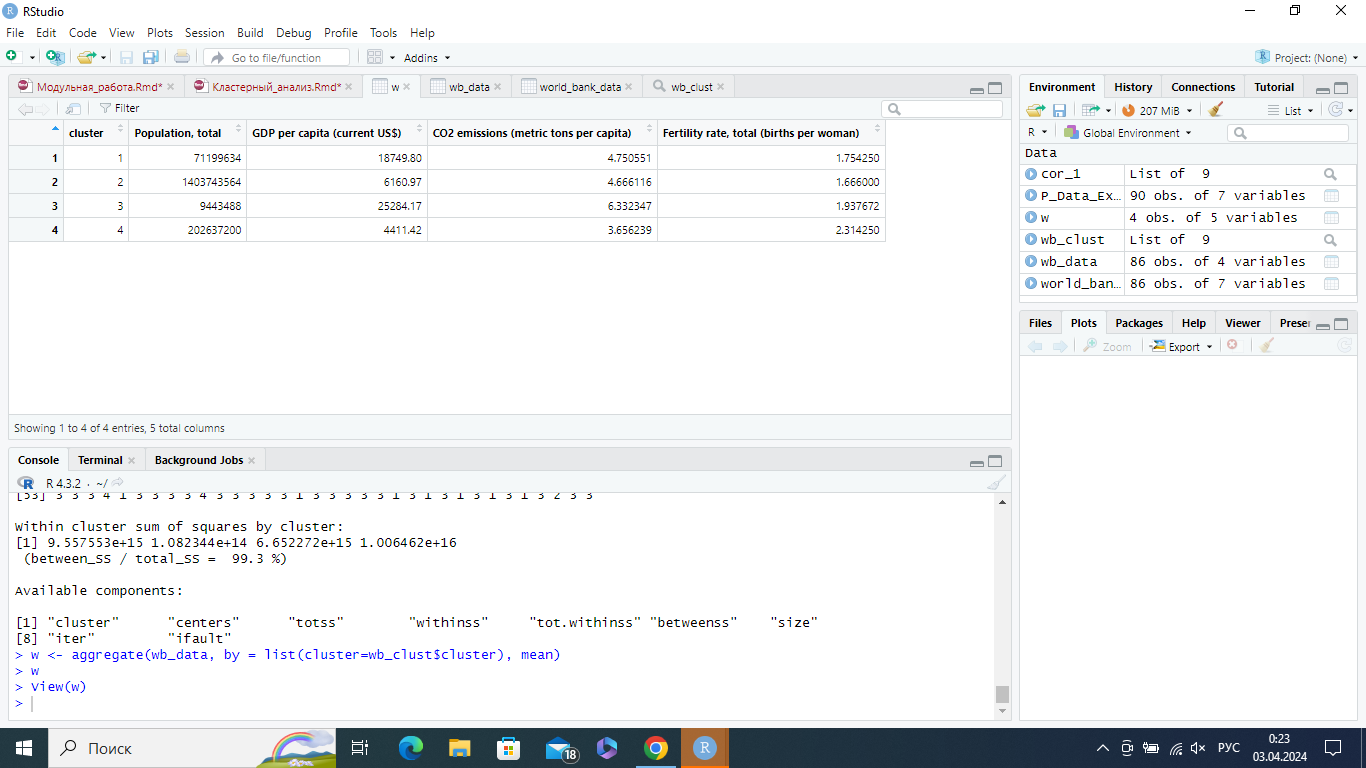
Первый кластер - 46 стран

Второй кластер - 16 стран

Третий кластер - 2 страны

Четвертый кластер - 22 страны

Попробуем аналитически взглянуть на полученные результаты. Первый кластер включает в себя развивающиеся страны с относительно высоким уровнем ВВП на душу населения и низким уровнем рождаемости (условная категория стран среднего достатка). Второй кластер включат в себя самые бедные развивающиеся страны с высокой рождаемостью и низким ВВП на душу населения. Третий кластер состоит из “стран-выбросов”: Китая и Индии, которые сложно анализировать вместе с другими кейсами. Четвертый же кластер включает в себя развитые страны Западной Европы и монархии Персидского залива с высоким подушевым ВВП и низкой рождаемостью.



Первый кластер - 16 стран

Второй кластер - 2 страны

Третий кластер - 64 стран

Четвертый кластер - 4 страны

На графике представлена кластеризация нормализованных данных, несмотря на некоторые сходства (например, отдельный кластер со “странами-выбросами”) видно, что кластеризация прошла несколько иначе. В отличие от кластеризации стандартизированных данных у нас получился перевес в пользу третьего кластера, в который вошло большее число государств. Преимуществом данной модели является возможность её практической интерпретации, в ходе которой мы можем описать средние величины для каждого кластера. Например, страны в первом кластере имеют в среднем население равное 71199634 человек, а также характеризуются относительно низким уровнем выбросов углекислого газа и низким уровнем рождаемости. Но проблема возникает при более детальном рассмотрении стран, входящих в кластер. Видно, что в кластер входят, как и развитые Соединенное королевство и Испания, так и развивающиеся страны типа Вьетнама и Мьянмы. Но аналитически, это скорее помогает сделать вывод о некотором сходстве между крайними случаями (деиндустриализованные страны Запада схожи со слабо индустриализованными развивающимися странами)

Анализ проведенных кластеризаций показал, что кластеризация стандартизированных данных лучше подходит для аналитической интерпретации и как следствие именно она выделяется нами как наиболее подходящая для дальнейшего исследования

## **Выводы**

Подводя итоги проделанной работы, стоит отметить, что предварительное исследование набора данных позволило понять направления для дальнейшей работы в этой сфере. При анализе данных на выбросы нами было замечено, что некоторые страны могут стать частью отдельных case-studies, направленных на понимание того, как выбранные предикторы влияют на выбросы углекислого газа. примерами таких стран могут быть Китай и Индия, как показали анализ выбросов и кластерный анализ, а также монархии Персидского залива. При этом были подтверждены гипотезы по связи между разными показателями. Например, в богатых странах рождаемость меньше, а также найдены новые тенденции, которые могут получить развитие в будущем (например, недемократические страны в среднем выбрасывают большего углекислого газа в воздух чем демократические). Поэтому предварительное исследование в целом дало множество направлений для развития анализа в сторону более подробных исследований взаимосвязей, например регрессионного анализа для выявления каузальности в проведенных корреляционных анализах. Результаты данного исследования могут быть полезны не только в теоретической области, но и практической, к примеру стать элементом принятия решений в сфере международных отношений и в целом в выработке политик, связанных с защитой окружающей среды.