Инжиниринг управления данными

Михайлов Роман

1 Определение области проекта

создание аналитической платформы, которая объединяет данные по странам (ВВП, сельскохозяйственный индекс, выбросы CO2) создаёт локальную SQLite базу данных для хранения данных фильтрует данные и строит визуализацию по запросам к бд

2 Сбор данных

данные для базы данных берём из csv файлов с портала https://worlddataview.com/

взяты 3 параметра

по каждому взят набор 21

csv файл по годам с 2000

по 2020

Copy Print CSV Excel PDF

Country	Crop production index (2004-2006 = 100)	Rank	Year
Singapore	390.68	1	1961
Seychelles	291.80	2	1961
Saint Kitts and Nevis	290.61	3	1961
Barbados	271.57	4	1961
Saint Lucia	247.81	5	1961
Trinidad and Tobago	221.84	6	1961
Antigua and Barbuda	195.53	7	1961
Tonga	152.15	8	1961
Grenada	130.93	9	1961
Norway	122.57	10	1961
Japan	120.81	11	1961

Исходные данные

для анализа взята статистика за год по параметрам:

1)gdp_growth 2)crop_index 3)co2_emissions

цель работы проанализировать как связана динамика параметров в разных странах

База данных

база данных создается локально с помощью библиотеки sqlite3 на питоне

2 этапа

1 первоначальный - база создается в первый раз с инпутом из папки инпут1 в которой хранится набор из 3х csv по 3м параметрам за один год - в данном случае 2000

2 этап в бд добавляются новые записи за следующие годы

1 этап

создаются 2 таблицы

Первая - countries хранит список всех стран, для которых есть данные, и присваивает каждой стране уникальный номер (ID) чтобы не хранить длинные названия стран в основной таблице данных. Вторая таблица country_indicators содержит значения показателей (ВВП, сельскохозяйственный индекс, выбросы CO2) для каждой страны по годам. Здесь используется ID страны из таблицы countries, чтобы связать данные с конкретной страной.

1 этап

Также в этой таблице создан дополнительный столбец хранящий буквенный код страны (например, RUS для России) просто как пример доп данных предполагалось например хранить там флаги стран чтобы их можно было вызывать для визуализаций и выводов по запросам бд но некоторые страны не имеют своих флагов в библиотеках питон поэтому для примера заменён на буквенный код который не дублируется у каждой строки с данными по годам а хранится отдельно в таблице стран в одном экземпляре для экономии места

2 этап

следующие данные прошлых лет хранятся в отдельной папке инпут2 и уже позже добавляются к бд база данных строится из таблицы в которой хранится название страны из исходного csv файла, страна получает свой уникальный id при появлении новых стран в следующие года новые страны получают свои id id короче чем текстовое название страны что позволяет экономить память в бд если там будут записи за миллионы лет

db_create_and_update

1 база данных успешно создана и сохранена в файл

2 из отдельной папки с новыми данными успешно обновлена

db_analysis

анализ и визуализация данных из бд

Обеспечение качества данных

```
--- Запуск автоматизированных проверок качества данных ---
✅ Проверка 1 пройдена: Таблица 'countries' имеет ожидаемую структуру.
✅ Проверка 2 пройдена: Таблица 'country_indicators' имеет ожидаемую структуру.
✅ Проверка 3 пройдена: Таблица 'countries' не пуста. Записей: 218
✅ Проверка 4 пройдена: Таблица 'country_indicators' не пуста. Записей: 4099
✅ Проверка 5 пройдена: В таблице 'country_indicators' нет дубликатов по (country_id, year).
✅ Проверка 6 пройдена: Все country_id в 'country_indicators' ссылаются на существующие в 'countries'.
✅ Проверка 7 пройдена: Диапазон лет в 'country_indicators' (2000-2020) ожидаемый (2000-2020).
✅ Проверка 8 пройдена: В столбце 'co2_emissions' нет отрицательных значений.

--- Результат проверок ---
Пройдено: 8 / 8
✅ Все проверки качества данных пройдены успешно!
```

Анализ пропущенных данных

функция вызывает бд

```
conn = sqlite3.connect(db_path)

# Загрузим данные из country_indicators
query_all_data = """
SELECT c.country_name, ci.year, ci.gdp_growth, ci.crop_index, ci.co2_emissions
FROM country_indicators ci
JOIN countries c ON ci.country_id = c.country_id;
"""
df_full = pd.read_sql_query(query_all_data, conn)
conn.close()

print("Структура датафрейма из БД:")
print(df_full.info())

print("\nКоличество пропусков по столбцам:")
print(df_full.isnull().sum())

print("\nПроцент пропусков по столбцам:")
print((df_full.isnull().sum() / len(df_full)) * 100)
```

```
Структура датафрейма из БД:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4099 entries, 0 to 4098
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   country_name   4099 non-null   object
 1   year           4099 non-null   int64
 2   gdp_growth     4005 non-null   float64
 3   crop_index     3591 non-null   float64
 4   co2_emissions  3634 non-null   float64
dtypes: float64(3), int64(1), object(1)
memory usage: 160.2+ KB
None

Количество пропусков по столбцам:
country_name       0
year               0
gdp_growth        94
crop_index       508
co2_emissions    465
dtype: int64

Процент пропусков по столбцам:
country_name      0.000000
year              0.000000
gdp_growth        2.293242
crop_index       12.393267
co2_emissions    11.344230
dtype: float64
```

видим что разные параметры имеют разное число пропусков


Количество пропущенных значений по годам и параметрам
Параметр
gdp_growth
crop_index
co2_emissions
Количество пропусков
Год
0
25
50
75
100
125
150
175
200
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

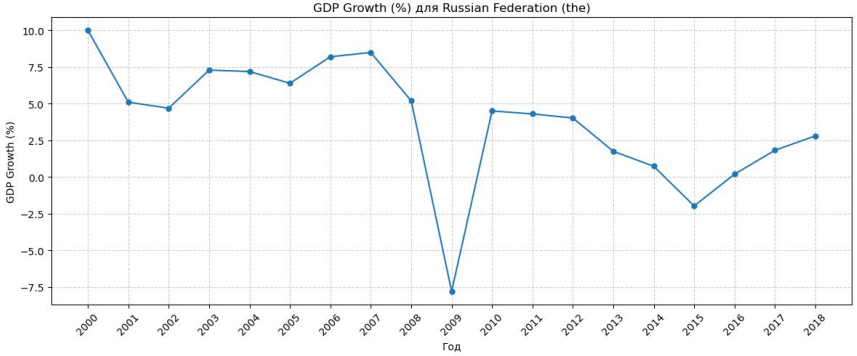
Фильтрация

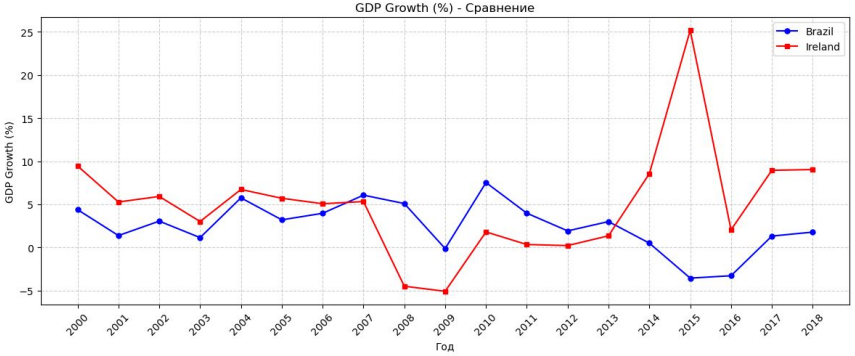
видим что есть сильные проблемы с данными с начала пандемии - 2 года можно отсеивать как недостоверные

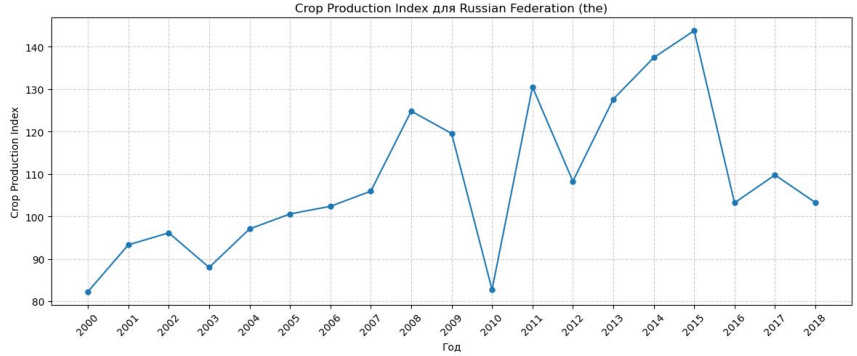
Количество стран без пропусков по всем параметрам (2000-2018): 170

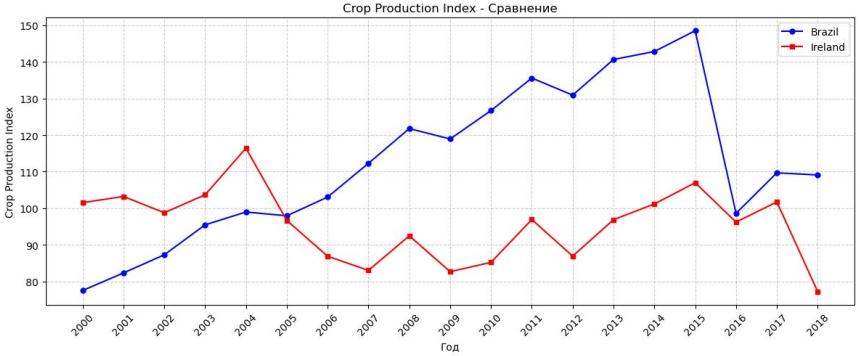
Анализ данных и визуализация

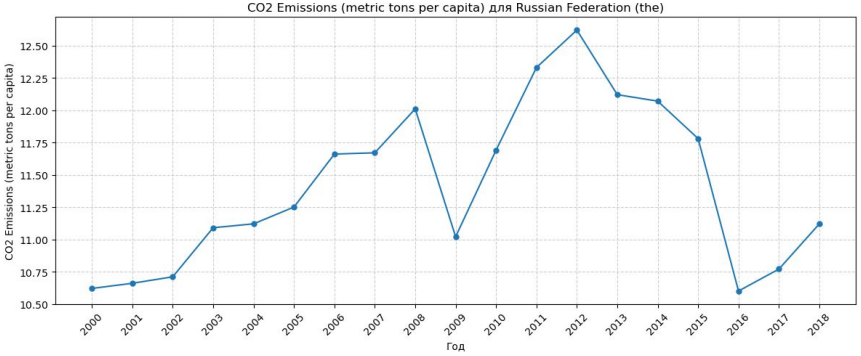
написана функция для визуализации уровней параметров по годам для одной страны по выбору и для двух стран VS по выбору

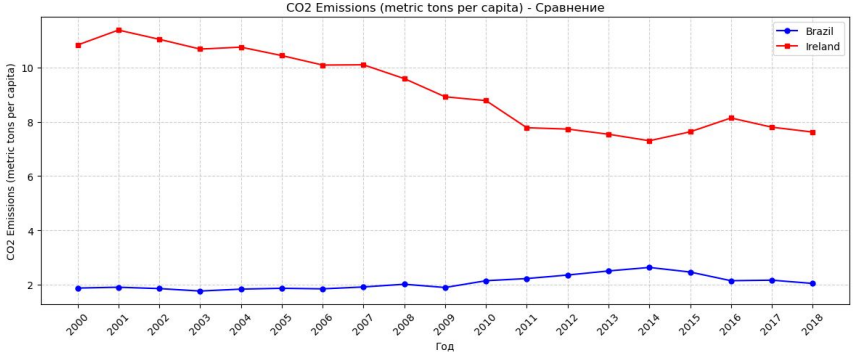

GDP Growth (%) для Russian Federation (the)
GDP Growth (%)
Год


GDP Growth (%) - Сравнение
Brazil
Ireland
GDP Growth (%)
Год


Crop Production Index для Russian Federation (the)
Crop Production Index
Год


Crop Production Index - Сравнение
Brazil
Ireland
Crop Production Index
Год


CO2 Emissions (metric tons per capita) для Russian Federation (the)
CO2 Emissions (metric tons per capita)
Год


CO2 Emissions (metric tons per capita) - Сравнение
Brazil
Ireland
CO2 Emissions (metric tons per capita)
Год

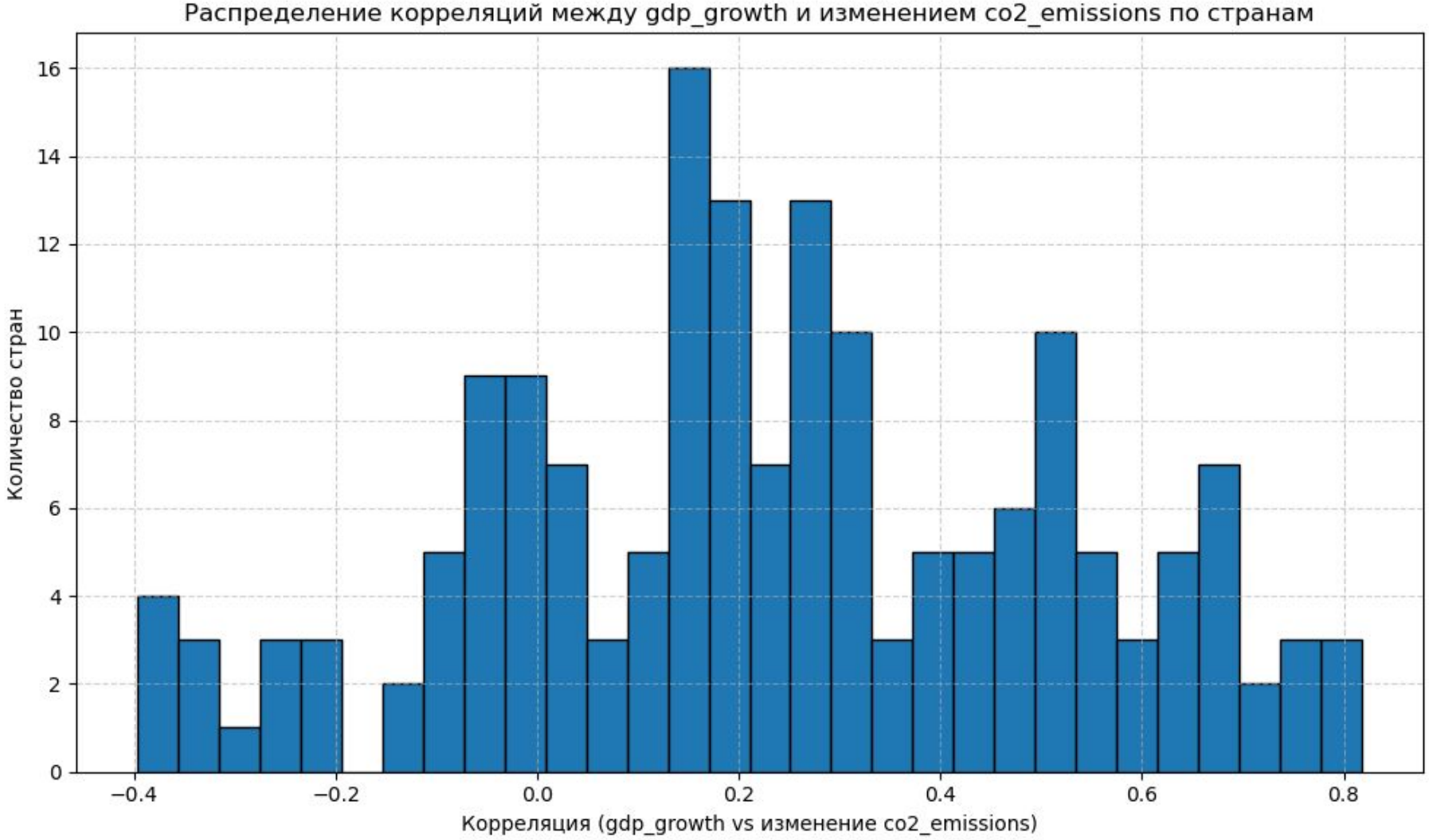
Инсайты

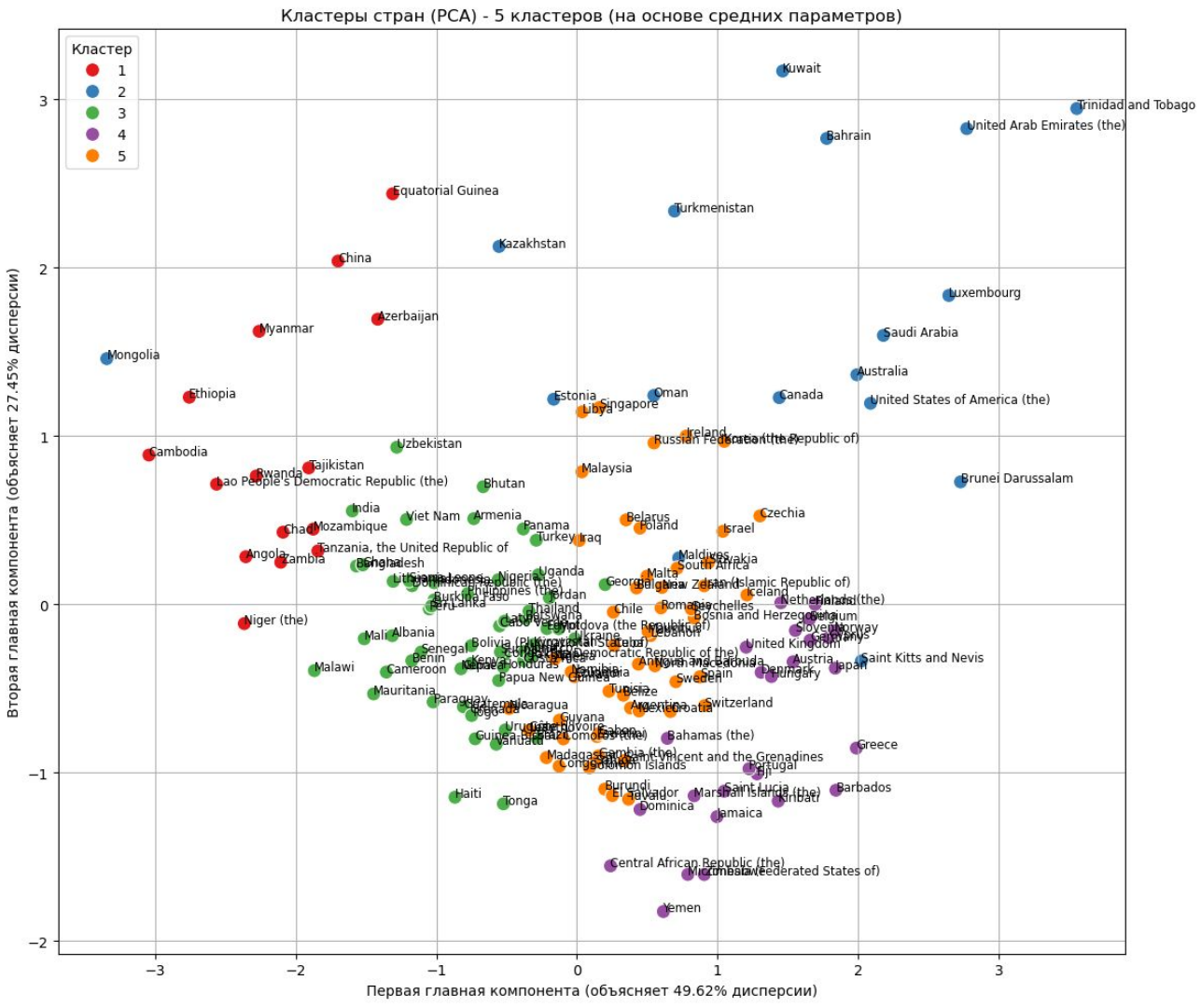
видим что в разных странах синхронизация отличается вероятно из за того что кредиты в банковской сфере берутся не на ближайший сезон посева и тд

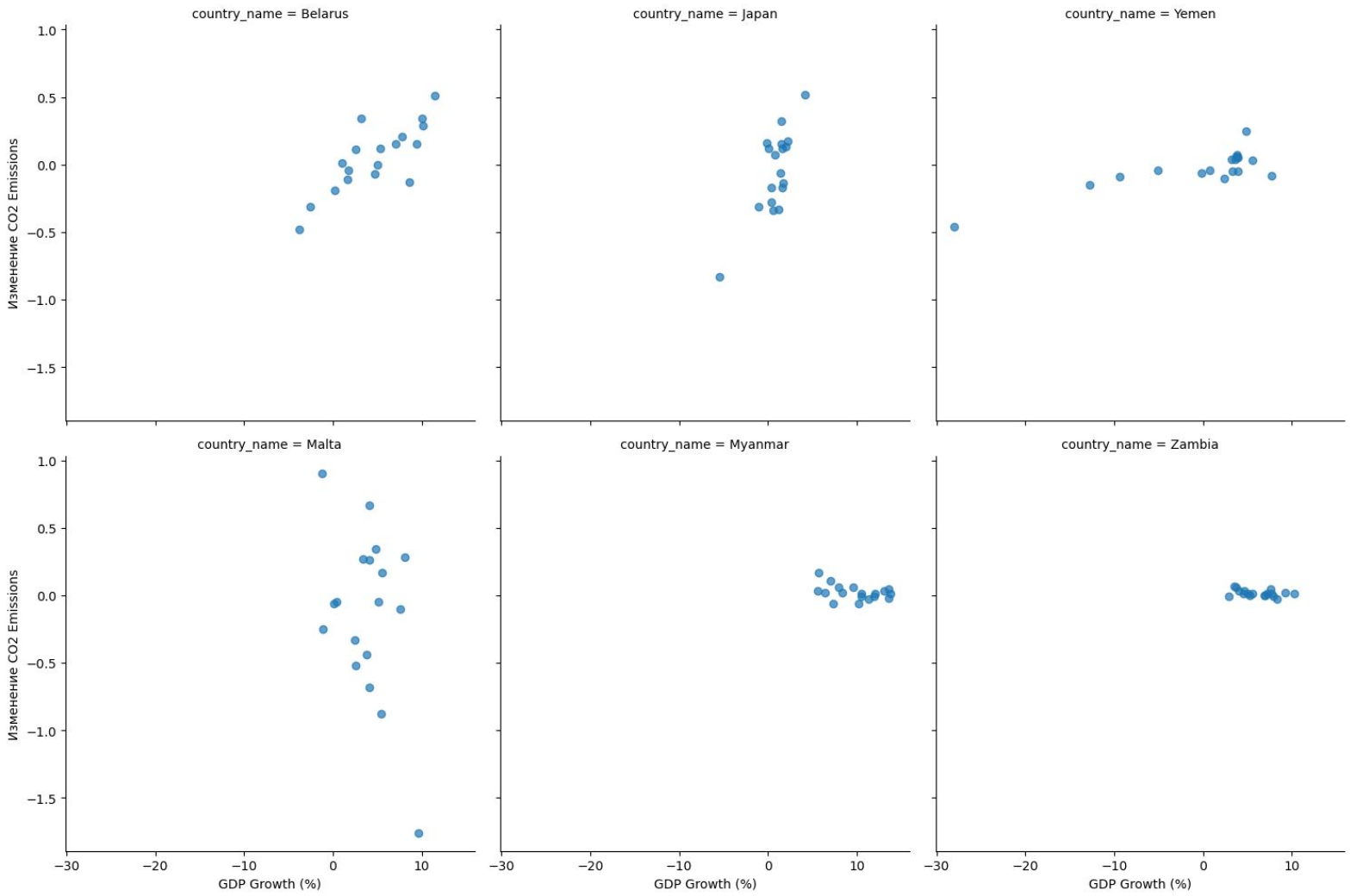
наглядно видно что например в ирландии резкий экономический спад кризиса 2008 начался на год раньше чем в бразилии при этом бразилия не сокращала производство сх а наращивала

Поиск и визуализация паттернов и корреляций

в ноутбуке db_analysis находятся функции для анализа кореляций и визуализации


Распределение корреляций между gdp_growth и изменением co2_emissions по странам
Количество стран
Корреляция (gdp_growth vs изменение co2_emissions)
0
2
4
6
8
10
12
14
16
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8


Кластеры стран (PCA) - 5 кластеров (на основе средних параметров)
Кластер
1
2
3
4
5
Вторая главная компонента (объясняет 27.45% дисперсии)
Первая главная компонента (объясняет 49.62% дисперсии)
Kuwait
Trinidad and Tobago
United Arab Emirates (the)
Bahrain
Equatorial Guinea
Turkmenistan
Kazakhstan
China
Luxembourg
Azerbaijan
Myanmar
Saudi Arabia
Mongolia
Australia
Ethiopia
Estonia
Oman
Canada
Singapore
Libya
United States of America (the)
Ireland
Russian Federation (the)
Cambodia
Uzbekistan
Tajikistan
Rwanda
Malaysia
Lao People's Democratic Republic (the)
Bhutan
Brunei Darussalam
India
Viet Nam
Armenia
Belarus
Czechia
Panama
Poland
Israel
Chad
Mozambique
Turkey
Iraq
Tanzania, the United Republic of
Maldives
Angola
Zambia
South Africa
Uganda
Malta
Iceland
Niger (the)
Albania
Bolivia (Plurinational State of)
Chile
Bosnia and Herzegovina
Belgium
Norway
Senegal
United Kingdom
Benin
Malawi
Cameroon
Saint Kitts and Nevis
Austria
Japan
Papua New Guinea
Spain
Sweden
Hungary
Mauritania
Paraguay
Tunisia
Belize
Switzerland
Argentina
Croatia
Guyana
Bahamas (the)
Greece
Madagascar
Gambia (the)
Saint Vincent and the Grenadines
Portugal
Fiji
Burundi
Haiti
Tonga
El Salvador
Marshall Islands (the)
Saint Lucia
Barbados
Kiribati
Dominica
Jamaica
Central African Republic (the)
Micronesia (Federated States of)
Yemen


Сравнение: Высокая vs Низкая корреляция (gdp_growth vs изменение co2_emissions)
country_name = Belarus
country_name = Japan
country_name = Yemen
country_name = Malta
country_name = Myanmar
country_name = Zambia
Изменение CO2 Emissions
GDP Growth (%)
1.0
0.5
0.0
−0.5
−1.0
−1.5
−30
−20
−10
0
10

Анализ

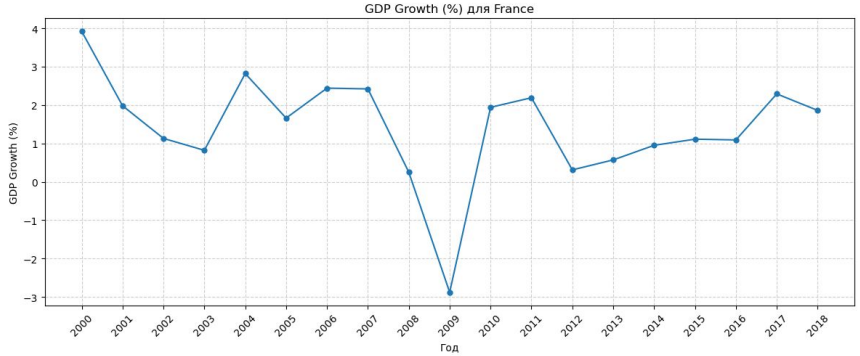
наглядно видно что в разных странах параметры изменяются не одинаково

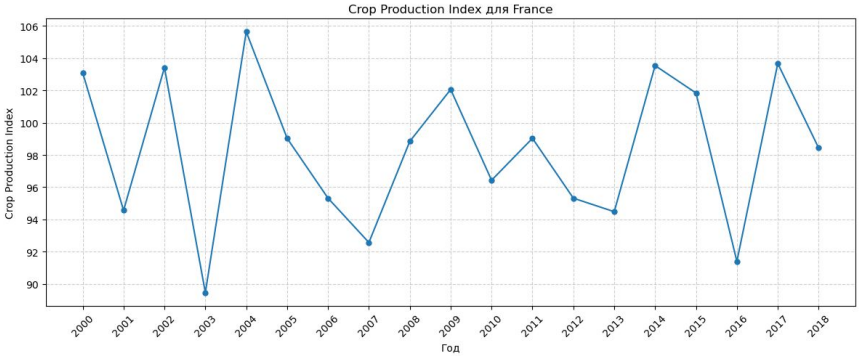
кластеризация по корреляциям не даёт однозначного объединения богатейших стран или преимущественно сельскохозяйственных стран что указывает на отсутствие единой модели волшебной палочки

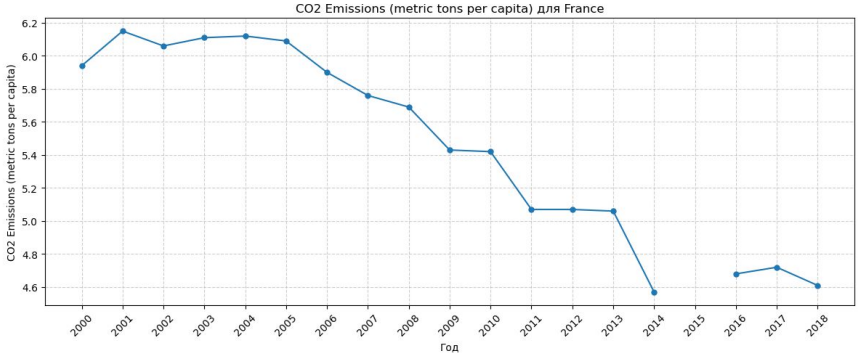
Работа с пропусками

написаны функции для поиска стран имеющих пропуски в данных

для примера взята франция с 1 пропуском по отчётам выбросов углекислого газа за 2015й год


GDP Growth (%) для France
GDP Growth (%)
Год


Crop Production Index для France
Crop Production Index
Год


CO2 Emissions (metric tons per capita) для France
CO2 Emissions (metric tons per capita)
Год

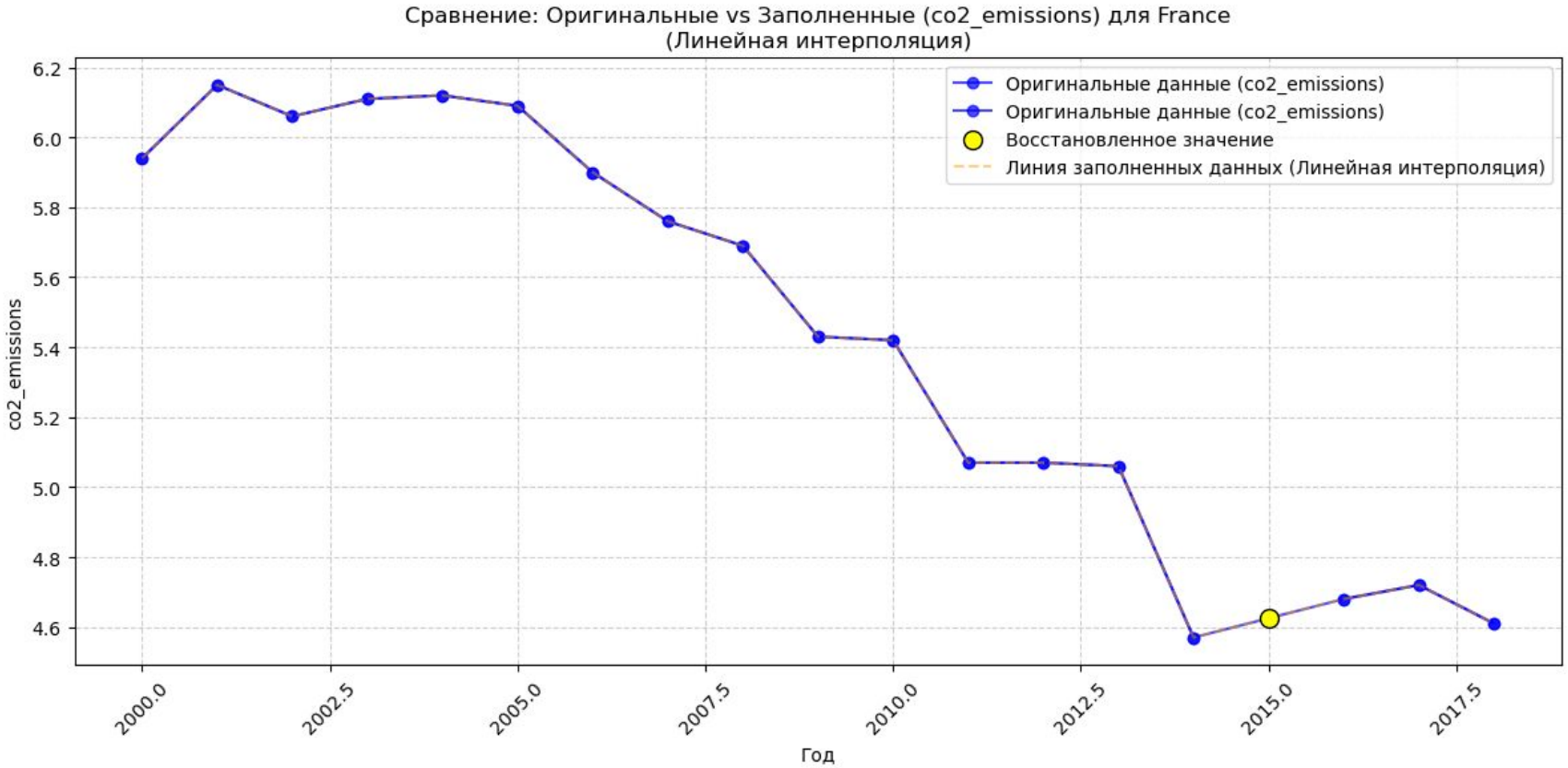
Заполнение пропусков

написаны 2 разные функции которые восстанавливают пропущенные значения разными способами

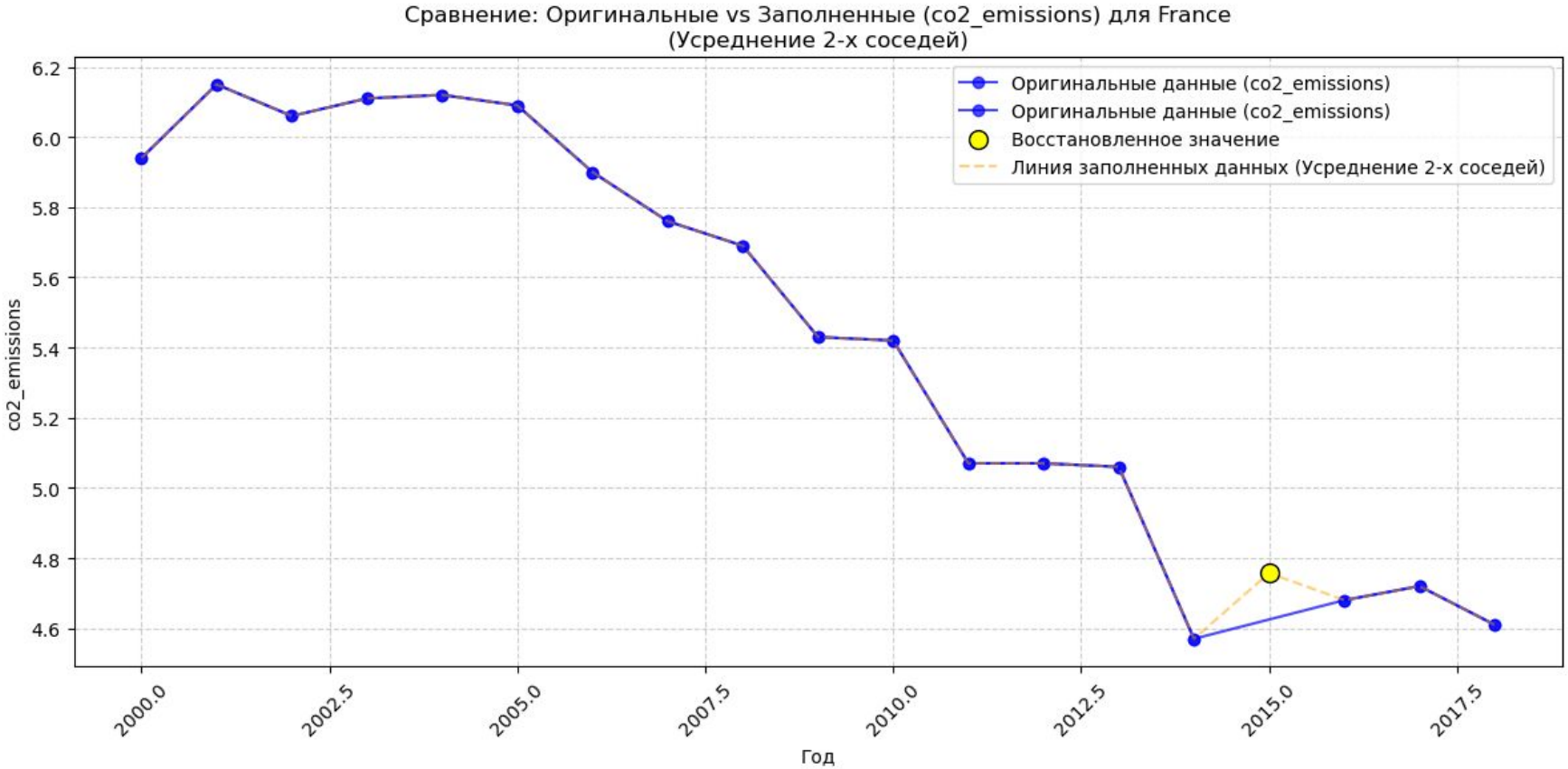
1 метод - усредняется 1 предыдущий и 1 следующий

2 метод - усредняется 2 предыдущих и 2 следующих

метод 1


Сравнение: Оригинальные vs Заполненные (co2_emissions) для France
(Линейная интерполяция)
Оригинальные данные (co2_emissions)
Оригинальные данные (co2_emissions)
Восстановленное значение
Линия заполненных данных (Линейная интерполяция)
co2_emissions
6.2
6.0
5.8
5.6
5.4
5.2
5.0
4.8
4.6
2000.0
2002.5
2005.0
2007.5
2010.0
2012.5
2015.0
2017.5
Год

метод 2


Сравнение: Оригинальные vs Заполненные (co2_emissions) для France
(Усреднение 2-х соседей)
Оригинальные данные (co2_emissions)
Оригинальные данные (co2_emissions)
Восстановленное значение
Линия заполненных данных (Усреднение 2-х соседей)
co2_emissions
6.2
6.0
5.8
5.6
5.4
5.2
5.0
4.8
4.6
2000.0
2002.5
2005.0
2007.5
2010.0
2012.5
2015.0
2017.5
Год

сравнение

наглядно видно что разные подходы возвращают разные значения

это важно учитывать для анализа и прогнозов

Итоги

Создана БД

В бд добавляются новые записи

Созданы функции анализа качества данных

Протестированы методы заполнения пропущенных данных

Созданы функции визуализации и анализа корреляций в данных

Проанализированы корреляции и тенденции на основе данных

Выводы

нет единой модели зависимости между ввп , уровнем производств сх культур и выбросами CO2
различные страны демонстрируют разные корреляции и различные паттерны по годам что указывает на невозможность единого курса на эко френдли производства в странах с разной структурой экономики и культурами потребления

нельзя просто воспроизводить производственные практики и ожидать пропорционального снижения выбросов и так далее