НИУ Высшая Школа Экономики

Миннеахметова Рената ***Осипян Александр***

Научно-исследовательский проект

11 ноября 2024 г.

Содержание

1 Цели

1. Для имеющегося датасета построить модель, отвечающую на вопрос, будет ли государство переживать кризис
2. Познакомиться с различными библиотеками языка Python, позволящими проводить анализ данных и визуализировать результат
3. Применить все полученные за этот курс навыки использования Github и Overleaf

2 Использованные методы

1. Библиотеки:
   - Numpy
   - Pandas
   - Scipy
   - Seaborn
   - Matplotlib
   - Ipywidgets
   - Sklearn
   - Statsmodels
2. Анализ датасета на предмет необходимости и удобства использования переменных
3. Построение различных графиков для отображения зависимостей переменных и наглядного анализа
4. Линейная регрессия - регрессионная модель зависимости одной переменной y от другой или нескольких других переменных x с линейной функцией зависимости

3 Содержание проекта

В ходе нашего проекта мы исследовали базу данных Jordà-Schularick-Taylor Macrohistory Database и документацию с описанием всех переменных.

Первый этап - предобработка данных. Мы убрали из рассмотрения те столбцы и строки, информация по которым была недостаточно представлена. Также мы убрали не репрезентативный столбец *"iso"*, содержащий код стран, поскольку будем использовать полные названия государств.

Второй этап - EDA - построение графиков, чтобы изучить данные наглядно.

3.1 "ВВП по ППС на душу населения по годам"

Мы построили **график-бар** с возможностью выбора страны. На оси абсцисс *"year"* - года, а на оси ординат *"rgdpmad"* - значение ВВП по ППС на душу населения. Видим неудивительную общую тенденцию для всех представленных стран - рост ВВП с незначительными падениями, обусловленными историческими событиями.

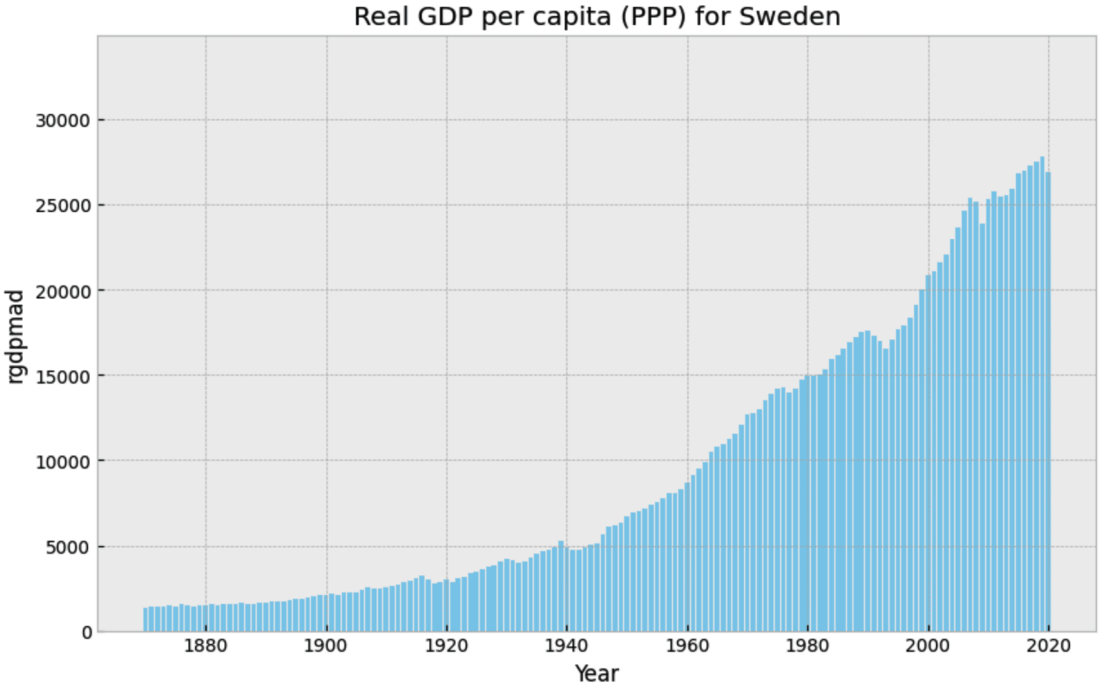

Рис. 1: ВВП по ППС на душу населения по годам

3.2 "Отображение доли экспорта и импорта"

График-пай так же, как и предыдущий график, можем отобразить для конкретной страны. У всех стран примерно одинаковые проценты импорта и экспорта. Наибольшая разница наблюдается в Германии (45.5% - импорт, 54.5% - экспорт), Норвегии (41.1% - импорт, 58.9% - экспорт), Великобритании (56.3% - импорт, 43.7% - экспорт) и США (55.8% - импорт, 44.2% - экспорт)

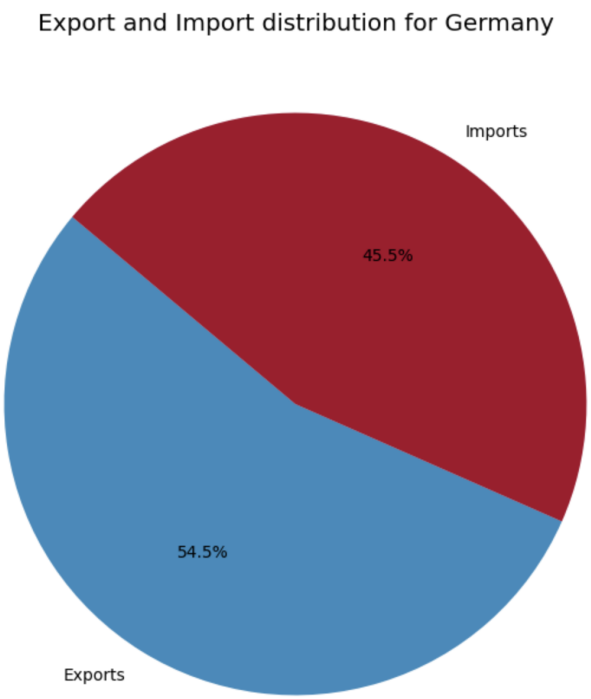

Рис. 2: Доли импорта и экспорта в Германии

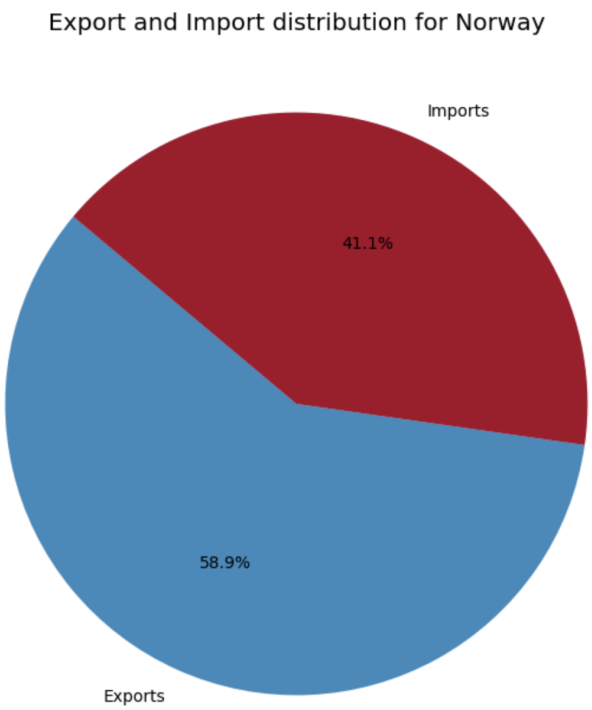

Рис. 3: Доли импорта и экспорта в Норвегии

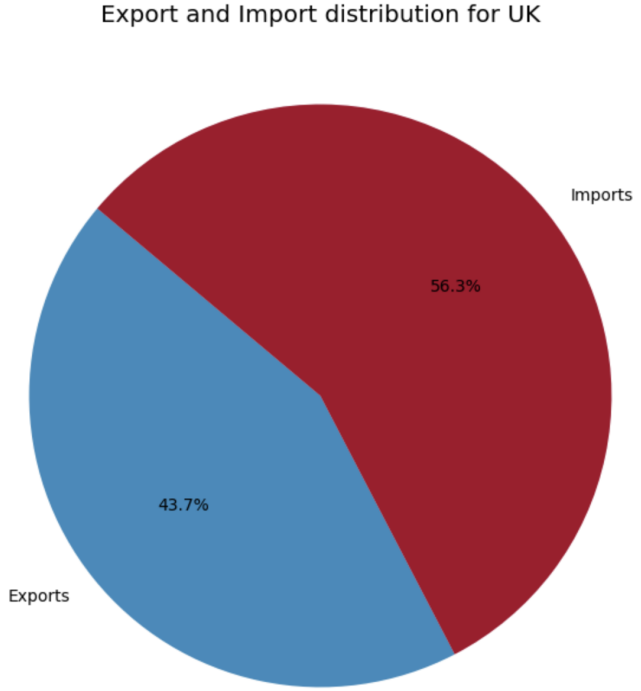

Рис. 4: Доли импорта и экспорта в Великобритании

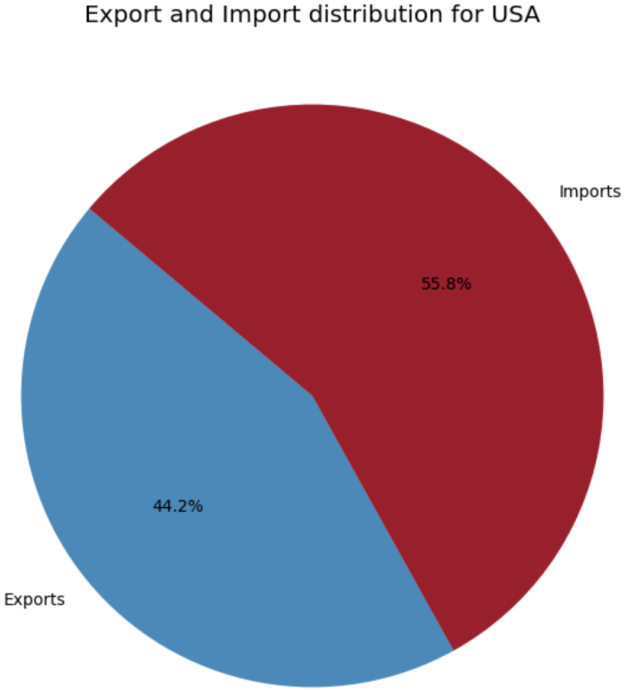

Рис. 5: Доли импорта и экспорта в США

3.3 "Зависимость размера ВВП и показателя отношения инвестиций к ВВП"

Точечный график с выбором страны. По горизонтали переменная *"rgdpmad"* - значение ВВП. По вертикали переменная *"iy"* - показатель отношения инвестиций к валовому внутреннему продукту. Хотелось бы отметить некоторое несовпадение наших предположений и реальности. Мы думали, что чем больше ВВП у страны, тем больше она будет инвестировать. Оказалось, это совсем не так. Пик инвестиций во всех государствах находится между граничными значениями ВВП и никогда в наибольшем его значении

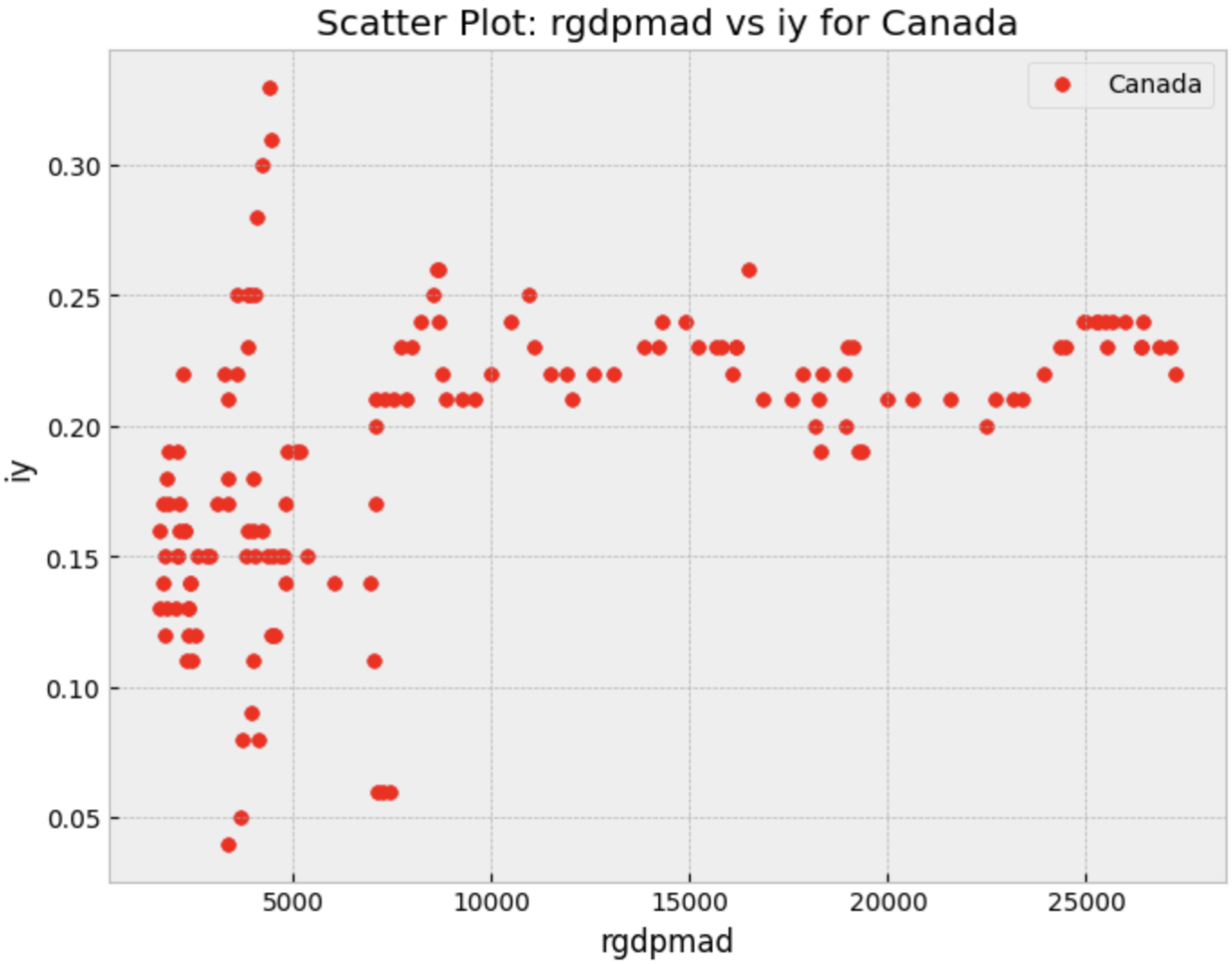

Рис. 6: График для Канады

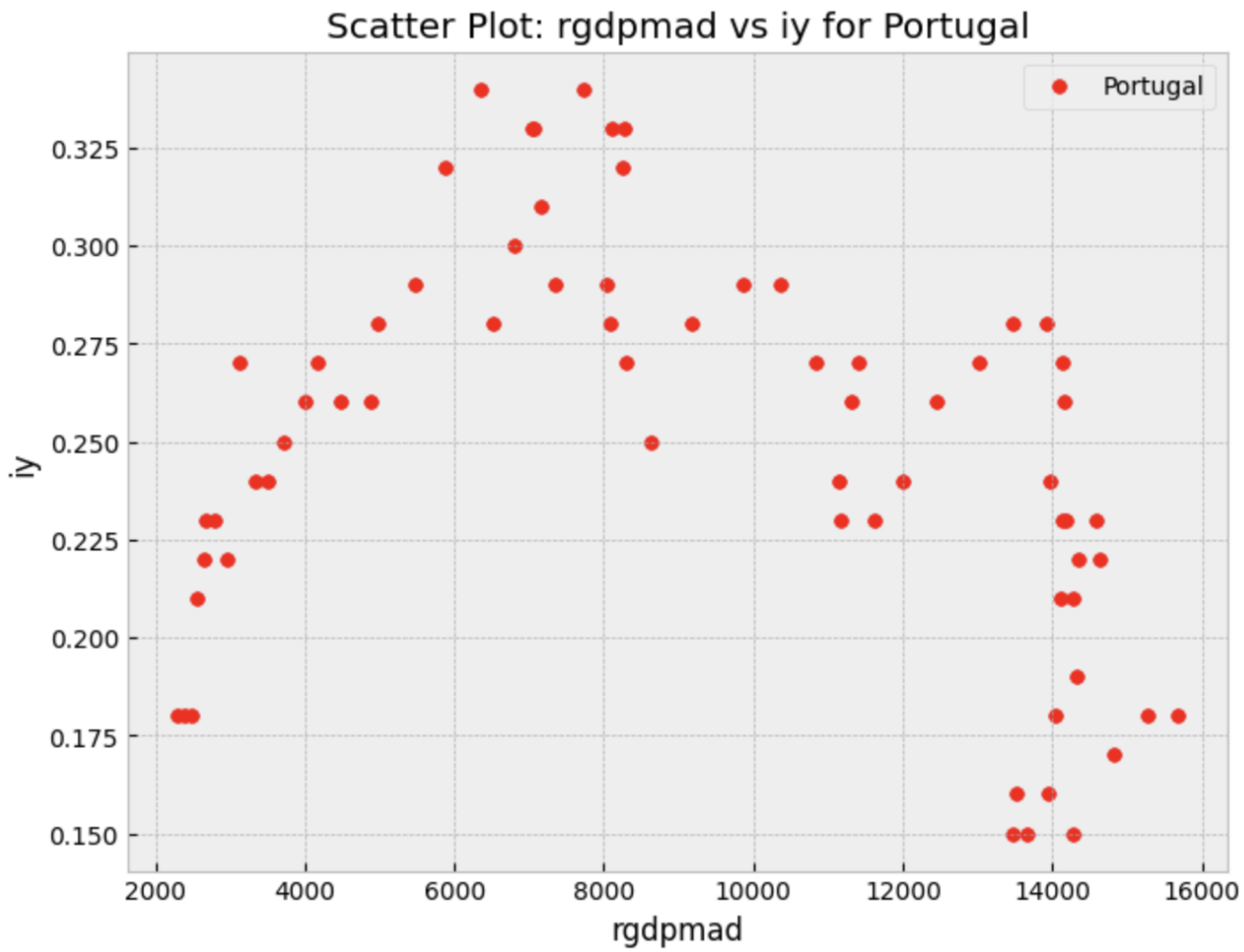

Рис. 7: График для Португалии

3.4 "Размах показателя отношения инвестиций к ВВП"

Бокс-плот (ящик с усами) для каждой страны. Красная линия - это медианное значение показателя отношения инвестиций к ВВП. Коробка (межквартильный размах, IQR) представляет собой разницу между 25-м и 75-м процентилями данных. Она включает в себя средние 50% данных. Усы представляют собой диапазон данных в 1.5 раза выше и ниже межквартильного размаха. А любые точки данных за пределами этого диапазона считаются потенциальными выбросами. Самое низкое медианное значение наблюдается в Великобритании. Оно равно 0.1.

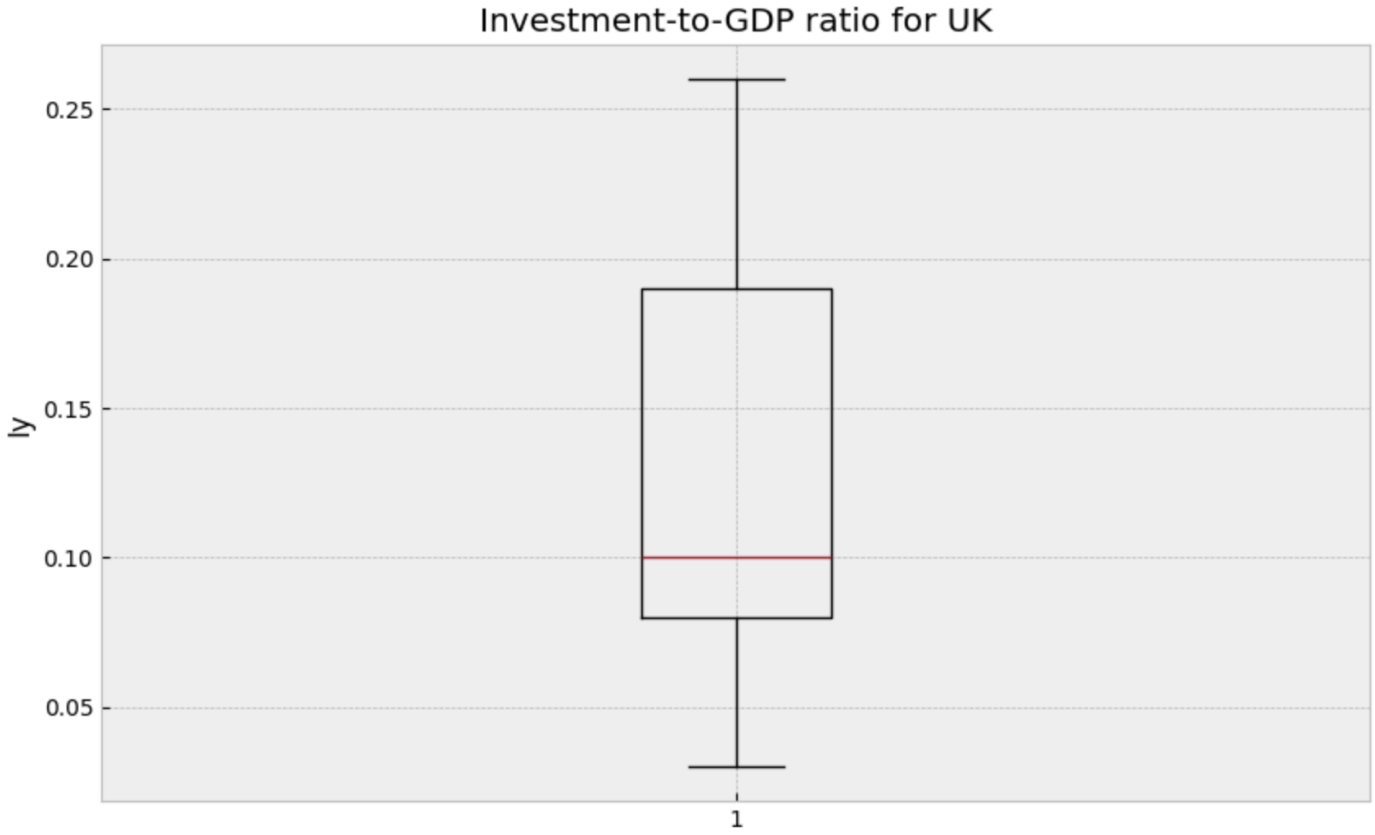

Рис. 8: Бокс-плот для Великобритании

3.5 "Показатель отношения инвестиций к ВВП по годам"

Графики-бары. Горизонтальная ось - *"year"* - года. Вертикальная ось - *"iy"* - показатель отношения инвестиций к ВВП. В этом случае мы решили вывести сразу все графики для всех государств, чтобы показать, отстутвие данных у 10-ти стран: Австралии, Бельгии, Швейцарии, Германии, Дании, Франции, Японии, Нидерландов, Норвегии, Португалии.

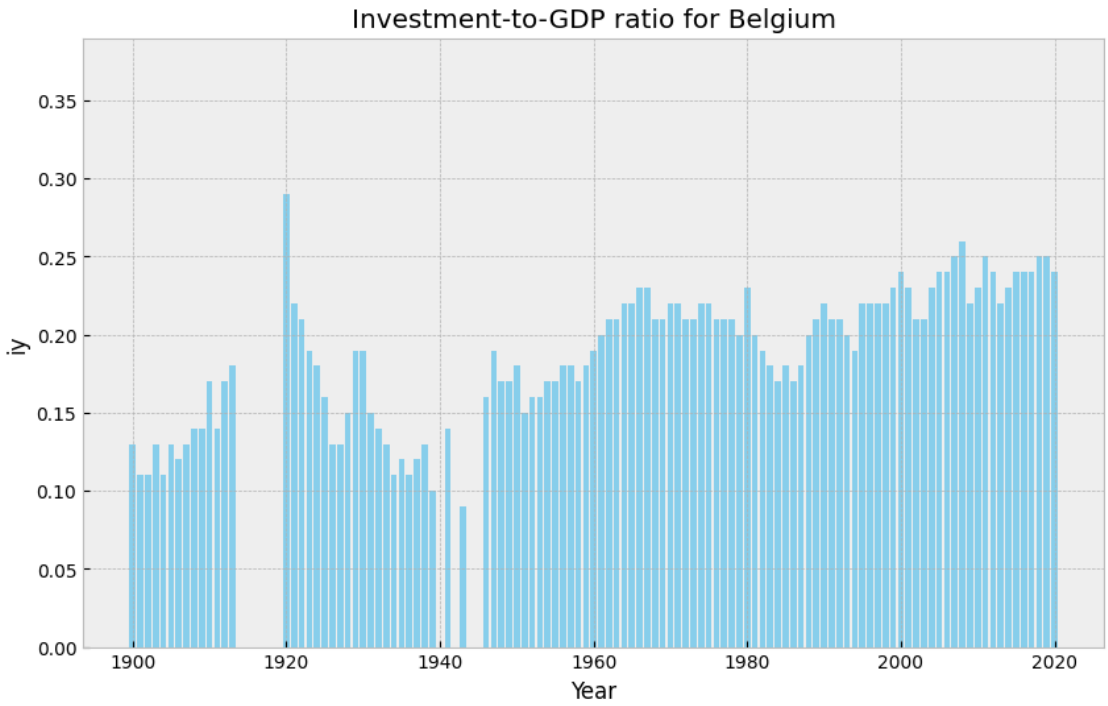

Рис. 9: Показатель отношения инвестиций к ВВП в Бельгии

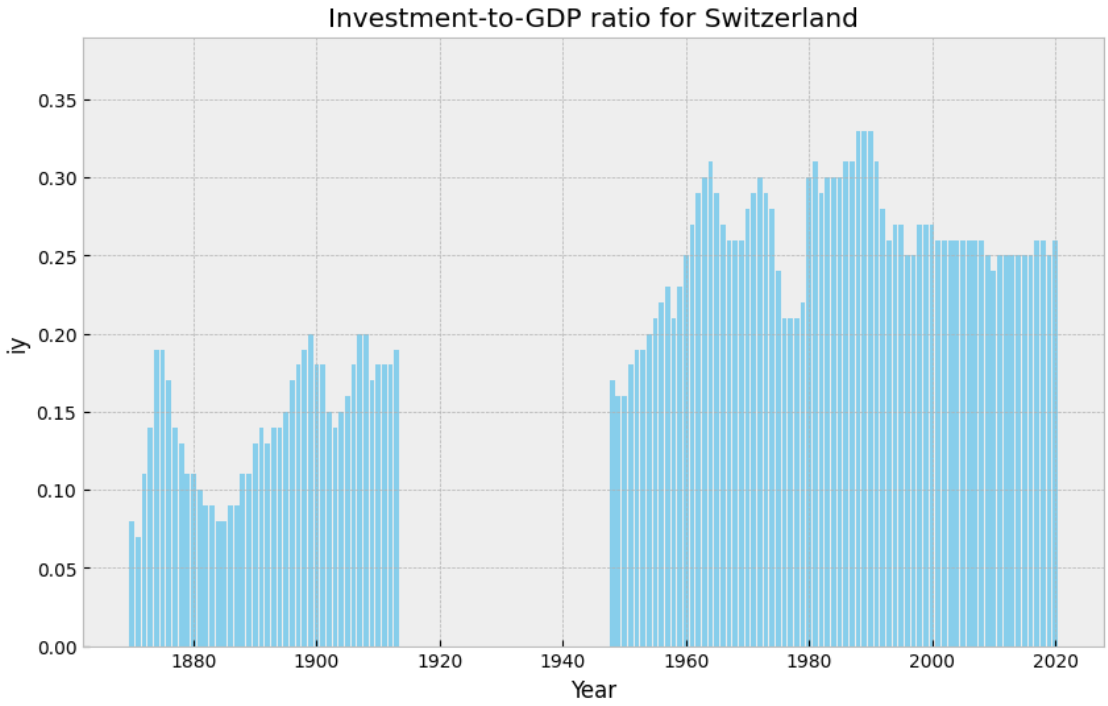

Рис. 10: Показатель отношения инвестиций к ВВП в Швейцарии

Третий этап - заполнение пропусков в данных

После того, как мы визуализировали все разведывательные данные, можем заметить, что некоторые стобцы имеют пропуски. Такие ячейки заполним средним значением этого показателя по стране (если пропуск во всех строках одной страны, то берется средний во всем столбце), чтобы не спровоцировать вброс. Данные в этом столбце никак не помешают обучению модели, но зато другие значения этих строк помогут точнее определить коэффициенты с помощью линейной регрессии.

Четвертый этап - тепловая карта

Как мы видим на левой картинке, многие столбцы коррелируют друг с другом, что испортит нашу модель. Необходимо удалить ненужные столбцы, зависимые друг от друга, чтобы этого избежать.

На правой картинке мы видим, что корреляция между переменными встречается довольно редко. Очистив всю таблицу от корреляции, мы будем иметь маленький набор критериев, что не позволит хорошо обучить модель, поэтому лучше не будем ничего удалять.

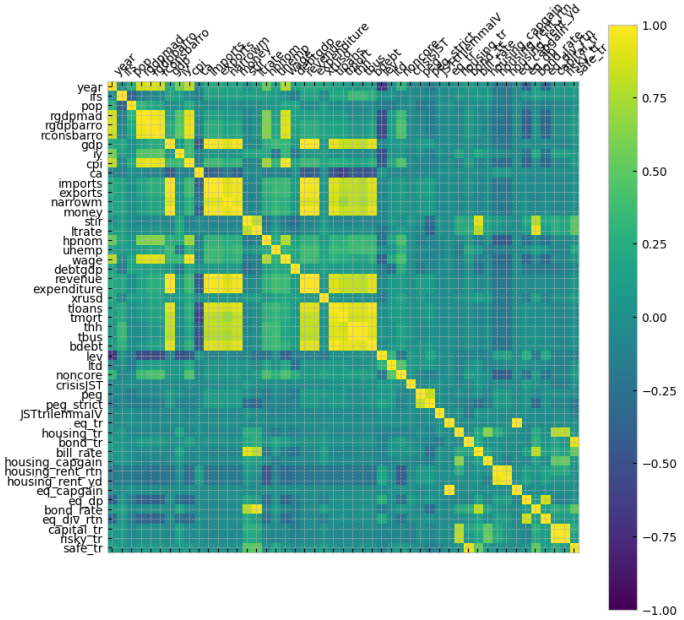

Рис. 11: Матрица корреляции

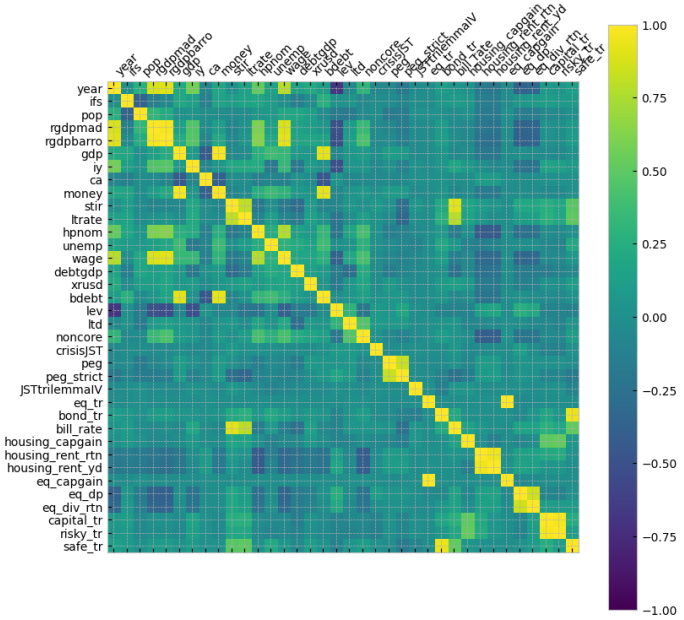

Рис. 12: Скорректированная матрица корреляции

Пятый этап - линейная регрессия

Таргетом выберем переменную *"crisisJST"*, которая показывает наличие кризиса в данном году в стране с помощью 0 и 1.

Следующим шагом мы создаем тренировочную и тестовую выборку, поделив данные в пропорции 80/20.

Создадим регрессию и сделаем прогноз значений *"crisisJST"* на тестовом наборе данных: оценим производительность модели, вычислив среднеквадратичную ошибку (MSE), и коэффициент детерминации (R^2) на тестовом наборе данных.

```
Mean Squared Error = 0.030008732921429488
R² = 0.061641945447983515
```

Такой результат может быть, если модель способна уловить общую тенденцию или закономерность в данных, но не может уловить конкретные взаимосвязи между независимыми и зависимыми переменными.

Среднеквадратическая ошибка говорит нам о приближении фактических значений и предсказанных - чем ниже MSE, тем точнее модель.

Коэффициент детерминации показывает, насколько хорошо линия регрессии соответствует наблюдаемым точкам данных. R^2 находится в диапазоне от 0 до 1, где 0 указывает на отсутствие линейной зависимости, а 1 указывает на идеальное соответствие.

Коэффициенты линейной регрессии (всех значений не видно, в ноутбуке мы вывели их отдельно)**:**

OLS Regression Results

Dep. Variable:	crisisJST	R-squared (uncentered):	0.111
Model:	OLS	Adj. R-squared (uncentered):	0.095
Method:	Least Squares	F-statistic:	7.169
Date:	Fri, 10 Nov 2023	Prob (F-statistic):	4.17e-32
Time:	13:52:31	Log-Likelihood:	675.31
No. Observations:	2053	AIC:	-1281.
Df Residuals:	2018	BIC:	-1084.
Df Model:	35		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
year	-1.592e-05	2.77e-05	-0.576	0.565	-7.02e-05	3.83e-05
ifs	-0.0001	0.000	-0.511	0.610	-0.001	0.000
pop	6.759e-08	1.1e-07	0.613	0.540	-1.48e-07	2.84e-07
rgdpmad	-2.976e-06	3.05e-06	-0.975	0.330	-8.96e-06	3.01e-06
rgdpbarro	0.0006	0.001	0.757	0.449	-0.001	0.002
gdp	-1.216e-09	1.36e-09	-0.897	0.370	-3.87e-09	1.44e-09
iy	-0.0342	0.093	-0.369	0.712	-0.216	0.147
ca	-5.574e-09	6.48e-09	-0.860	0.390	-1.83e-08	7.13e-09
money	1.048e-09	1.42e-09	0.739	0.460	-1.73e-09	3.83e-09
stir	0.0139	0.004	3.814	0.000	0.007	0.021
ltrate	-0.0044	0.002	-1.772	0.076	-0.009	0.000

...

[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.

[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[3] The condition number is large, 2.42e+10. This might indicate that there are strong multicollinearity or other numerical problems.

В итоге при умножении критериев на коэффициенты мы получим число от нуля до единицы, характеризующее вероятность кризиса в стране при данных условиях.