Project "Inequality and Growth"

Смешкова Екатерина, Цисарук Мария

November 2023

Содержание

1 Introduction

Основываясь на данных, полученных из OWID (Our World in Data) (https://ourworldindata.org), мы проанализировали **Economic Inequality** (https://ourworldindata.org/economic-inequality) и **Economic Growth** (https://ourworldindata.org/economic-growth#all-charts). Основными нашими целями было исследовать такие экономические показатели, как коэффициент Джини и уровень ВВП, а также взаимосвязь неравномерного распределения доходов с уровнем экономического развития и составить прогноз на основе линейной регрессии.

В дальнейшем мы рассмотрим оба этих показателя отдельно и потом их зависимость друг от друга.

2 Inequality

В качестве исходного показателя был выбран Коэффициент Джини, так называемый индекс концентрации доходов, который характеризует степень отклонения фактического распределения общего объема денежных доходов населения от линии их равномерного распределения. Данный коэффициент изменяется от 0 до 1. То есть мы получаем, что чем больше значение Джини отклоняется от нуля и приближается к единице, тем в большей степени доходы сконцентрированы в руках отдельных групп населения. Мы будем рассматривать данные до выплаты налогов, так как это во многом может исказить картину из-за того, что в странах могут быть представлены разные виды налогооблажения.

Обработаем данные экономического неравенства (https://github.com/tsisarukm/Project1/blob/main/inequality.csv). Для начала удалим все ненужные столбцы, которые являются неинформативными для нас. Также почистим все строки, где не хватает определенных данных, и отсортируем по возрастанию коэффициента Джини (т.е. по увеличению разрыва в доходах населения) (https://github.com/tsisarukm/Project1/blob/main/fixed%20Inequality.numbers).

Удалим данные для регионов мира, оставим только отдельные страны и рассмотрим теперь определенный год (у нас это будет 2018, так как это ближайший год с наиболее полной информацией) (https://github.com/tsisarukm/Project1/blob/main/Inequality%202018.numbers) (Рис. 1).

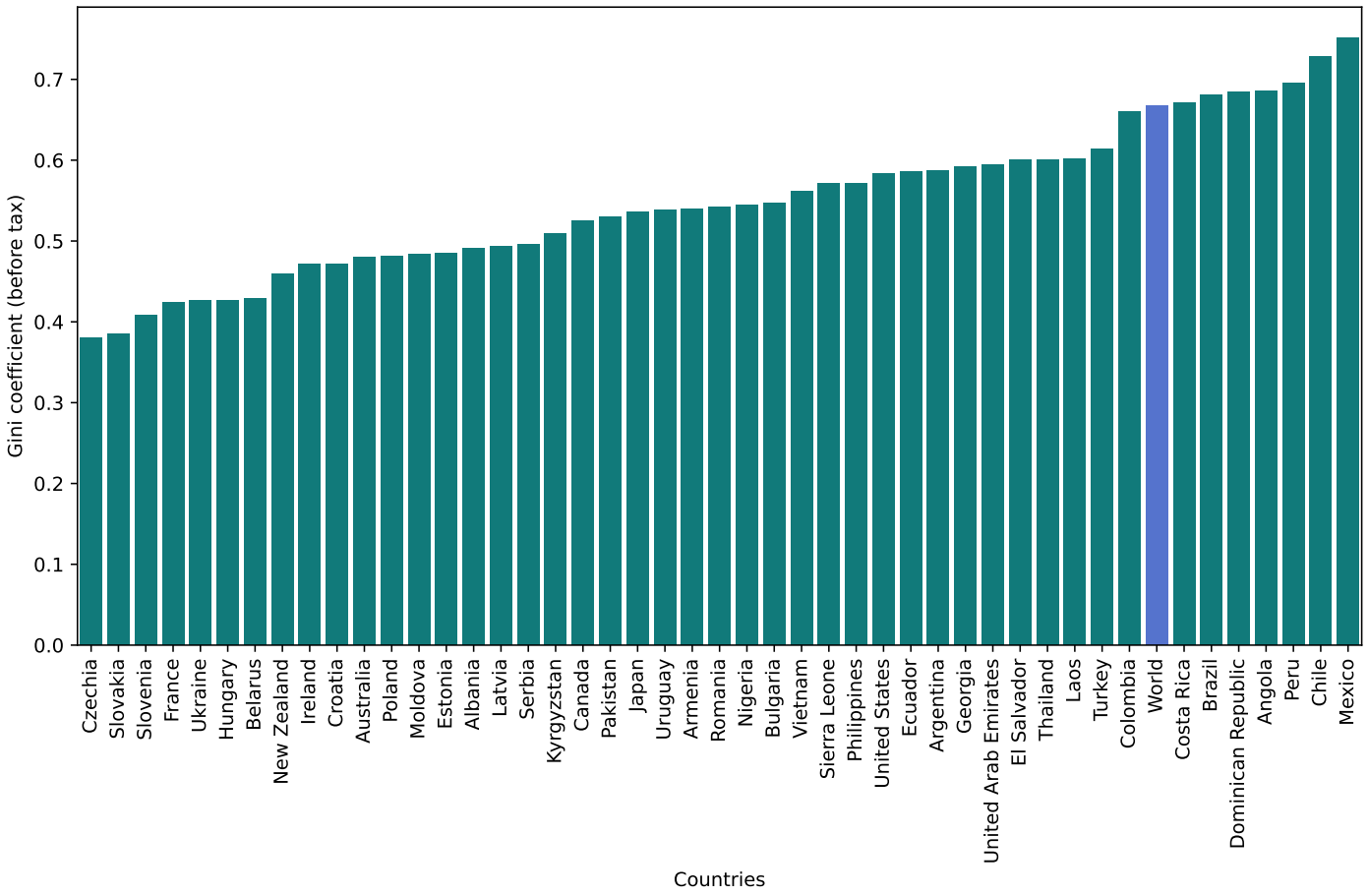

Рис. 1: Коэффициент Джини для разных стран

Теперь рассмотрим изменение коэффициента с течением времени. Для примера возьмем Россию, и опять же сравним с показателями по миру. Данные по России: (https://github.com/tsisarukm/Project1/blob/main/Inequality%20Russia.numbers). И для мира: (https://github.com/tsisarukm/Project1/blob/main/Inequality%20World.numbers). Рассматривать будем примерно с 1980-х годов, так как в России только в эти годы появляются полные данные (Рис. 2). Наши результаты показывают, что индекс концентрации доходов для России всегда ниже, чем для остального мира, следовательно, больше приближен к нулю, что определенно является хорошей тенденцией. Также можно заметить, что изменение коэффициента Джини для России соблюдает мировой тренд.

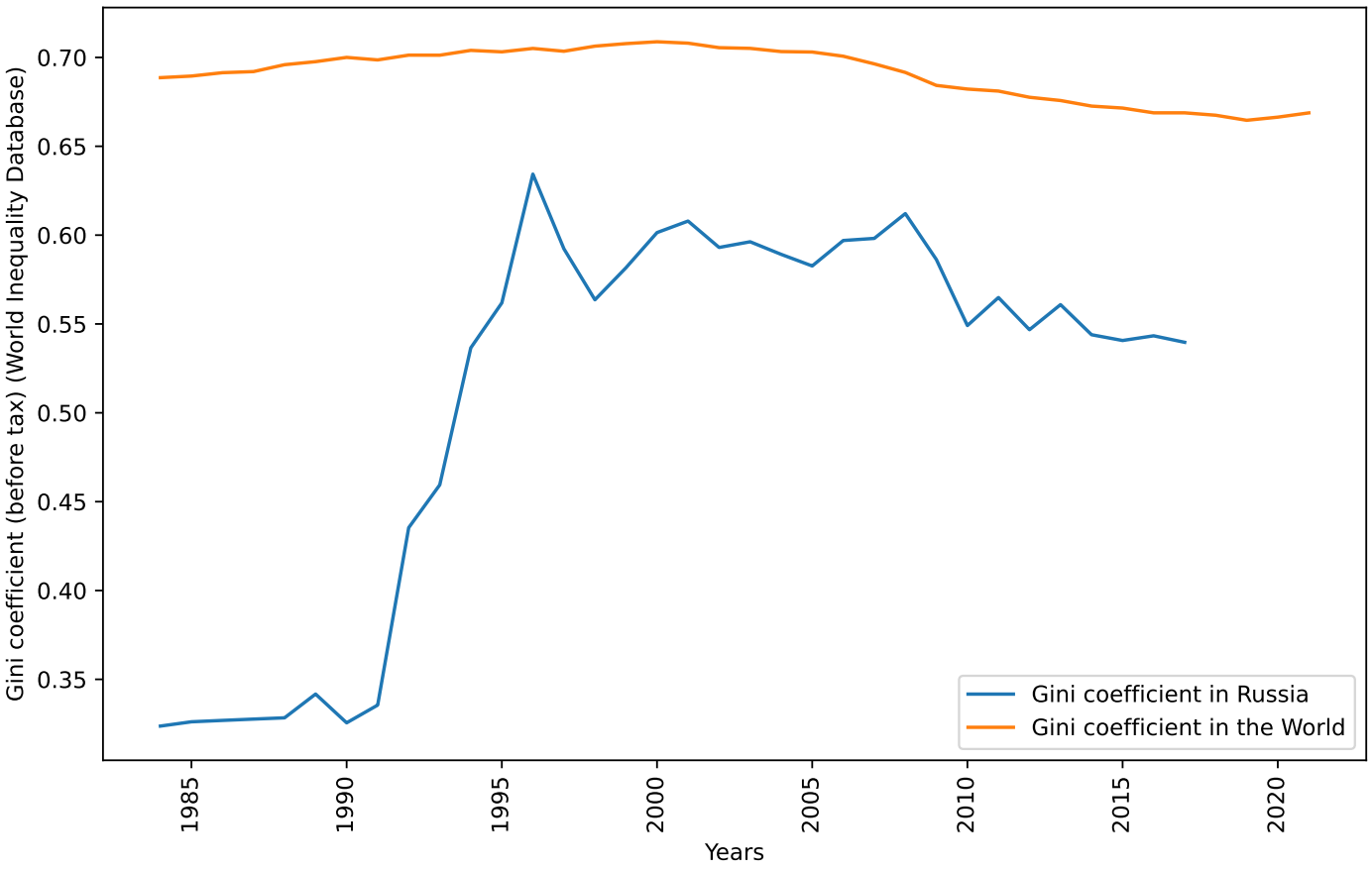

Рис. 2: Изменение коэффициента Джини с течением времени

3 Growth

Теперь рассмотрим данные по уровню экономического развития, для этого мы будем оценивать ВВП на душу населения (GDP (Gross Domestic Product) — это общая стоимость всех произведенных товаров и услуг в определенной стране за определенный период времени). Он является одним из основных показателей экономического развития страны и имеет прямое влияние на уровень жизни населения. Чем выше уровень ВВП, тем больше денег общество имеет для инвестиций в образование, здравоохранение, социальную защиту и другие общественные блага. Данные скорректированы с учетом инфляции и различий в стоимости жизни между странами.

Теперь начнем обрабатывать: в данном датафрейме всего 4 столбца, причем информация в каждом из них важна для анализа, поэтому в данном случае не будем удалять никакие колонки. В исходных данных переименуем колонку 'Entity' на 'Country', чтобы было нагляднее и красивее, и отсортируем наши строки в возрастающем порядке по значению ВВП. Потом по году, а затем в алфавитном порядке по стране. (https://github.com/tsisarukm/Project1/blob/main/GDP%20per%20capita%20.numbers)

Посмотрим, как изменялось среднее значение ВВП по всем странам с течением времени (https://github.com/tsisarukm/Project1/blob/main/mean%20GDP%20per%20capita.numbers) (Рис. 3). Из графика видно, что среднее значение ВВП имеет возрастающий рост с течением времени, исключая некоторые спады, связанные с историческими особенностями.

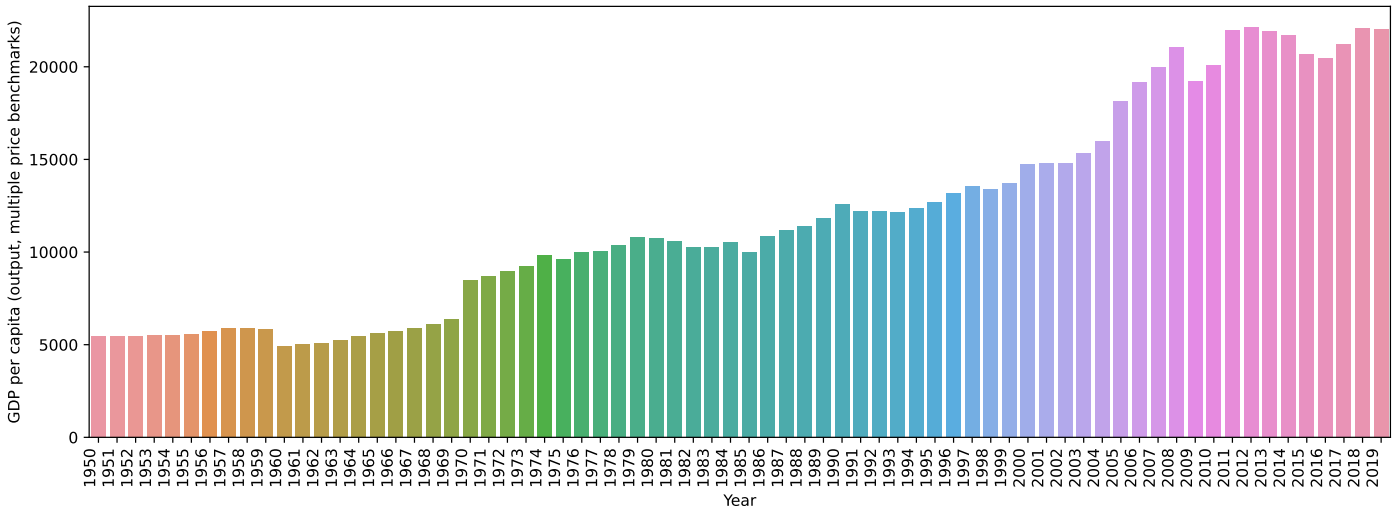

Рис. 3: Средний ВВП с течением времени

Далее посмотрим на среднее значение ВВП по странам за все годы с 1950 по 2019 (https://github.com/tsisarukm/Project1/blob/main/Countries%20GDP%20per%20capita.numbers) (Рис. 4). В формате картинки график не очень нагляден, поскольку в датасете представлена информация по многим странам. Однако при более детальном рассмотрении в подходящей среде (например, в Google Colab) видно, что самый высокий показатель ВВП наблюдается в ОАЭ, Катаре и Брунее. Самый низкий - в государствах Мозамбик, Бурунди и Эфиопия. Кроме того, наиболее близкое к среднему по миру значение ВВП наблюдается в Черногории и государстве Сент-Китс и Невис.

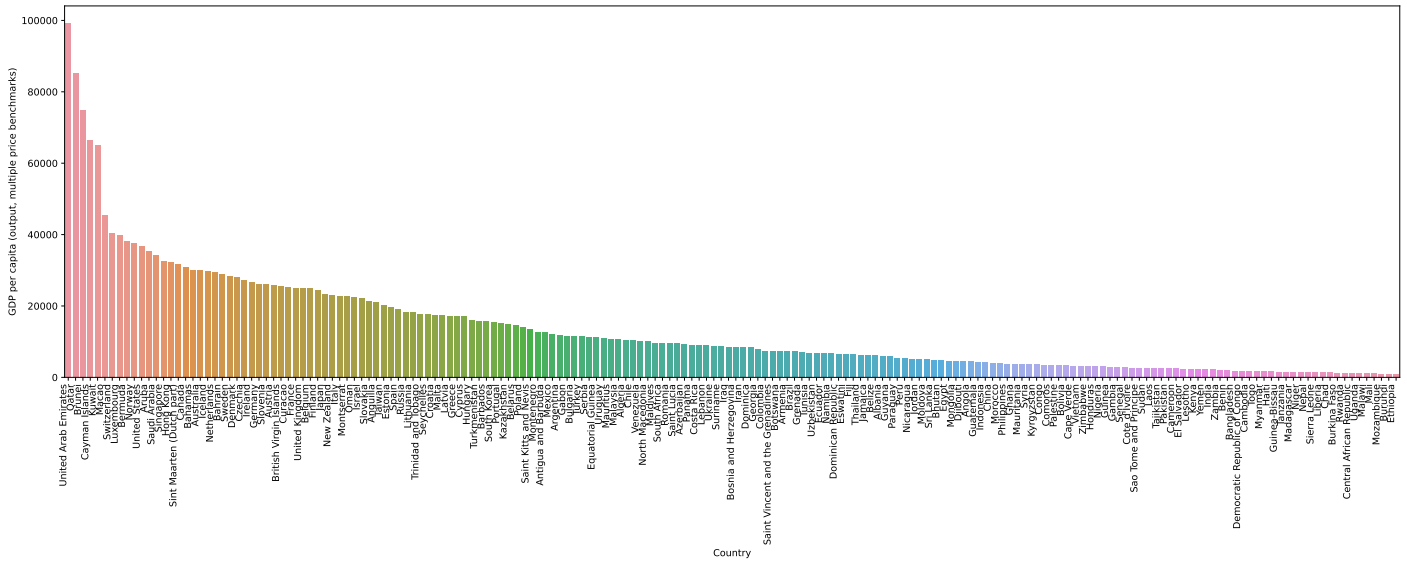

Рис. 4: Средний ВВП в разных странах

Сравним значение ВВП по России и всем остальным странам с течением времени. Данные для России: (https://github.com/tsisarukm/Project1/blob/main/Russia%20GDP%20per%20capita.numbers). И для остальных стран: (https://github.com/tsisarukm/Project1/blob/main/%20World%20GDP%20per%20capita.numbers) (Рис. 5). Из графика видно, что, во-первых, в анализируемом датафрейме данные о ВВП по России представлены начиная с 1990-го года. Во-вторых, до 1993 года уровень ВВП в России превышает среднее значение по миру. Однако с 1993 по 2007 год наблюдается спад: ВВП становится ниже среднего уровня. Несмотря на это, с 2008-го года анализируемый показатель для России вновь начинает преобладать над средним значением по остальным странам.

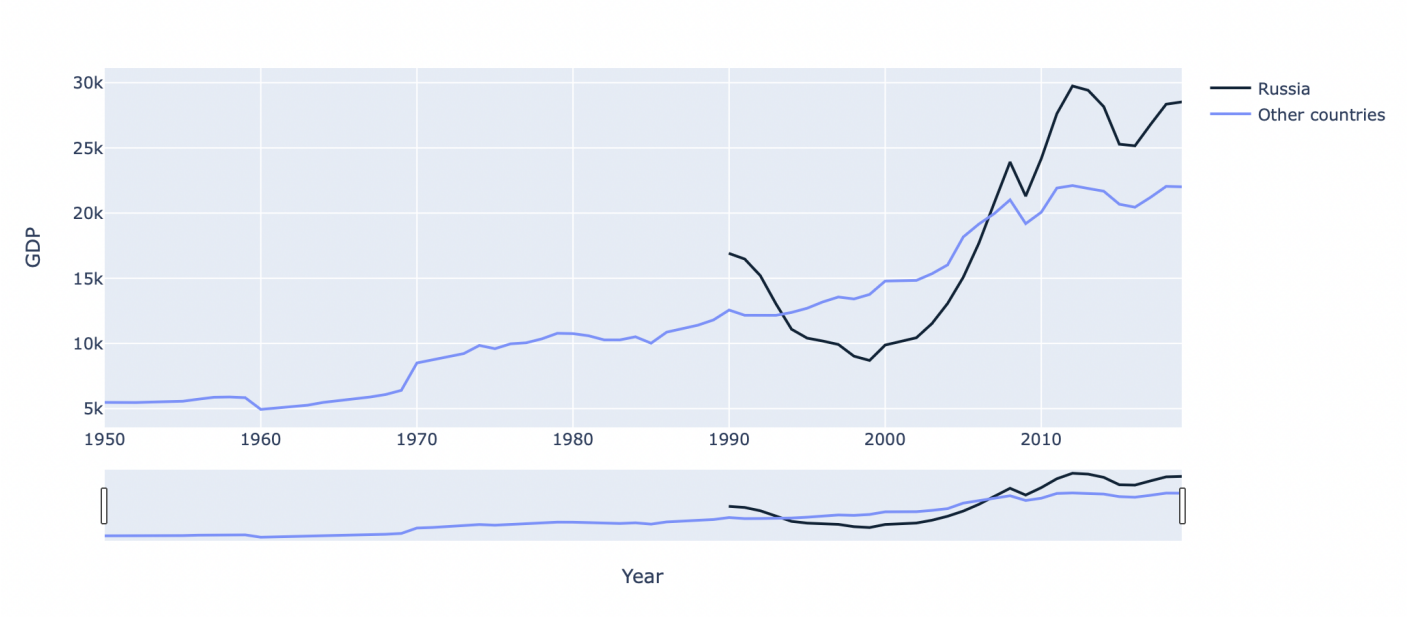

Рис. 5: Средний ВВП

4 Linear regression

Для построения линейной регрессии нужно объединить два уже проанализированных по-отдельности датафрейма Growth и Inequality по общим столбцам (Year и Country). Для начала построим краткую сводку по полученной таблице (https://github.com/tsisarukm/Project1/blob/34c2343dba713b631fb
Description%20of%20merged_table.numbers). Затем построим корреляционную таблицу для выявления взаимосвязей между столбцами, которые в дальнейшем будут мешать построению линейной регрессии (https://github.com/tsisarukm/Project1/blob/4d72afd760c377c1e5cd2310e801eaccf3b37
Correlation%20of%20Inequality%20and%20Growth.numbers).

Из (Рис. 6) мы видим, что такие показатели, как "Доля доходов 10 % самых богатых "Доля доходов 1 % самых богатых "Доля доходов 0.1 % самых богатых "Доля доходов 50 % самых богатых "Коэффициент Пальма напрямую взаимосвязаны с основным показателем (коэффициентом Джини), по которому мы оценивали распределение доходов. Данный вывод является весьма логичным, так как первые четыре показателя из ранее перечисленных по определению и будут составляющими индекса концентрации доходов, а коэффициент Пальма является альтернативой Коэффициента Джини, фокусирующейся на растущей пропасти между богатыми и бедными в обществе. Вследствие этого при построении линейной регрессии мы не стали учитывать эти показатели.

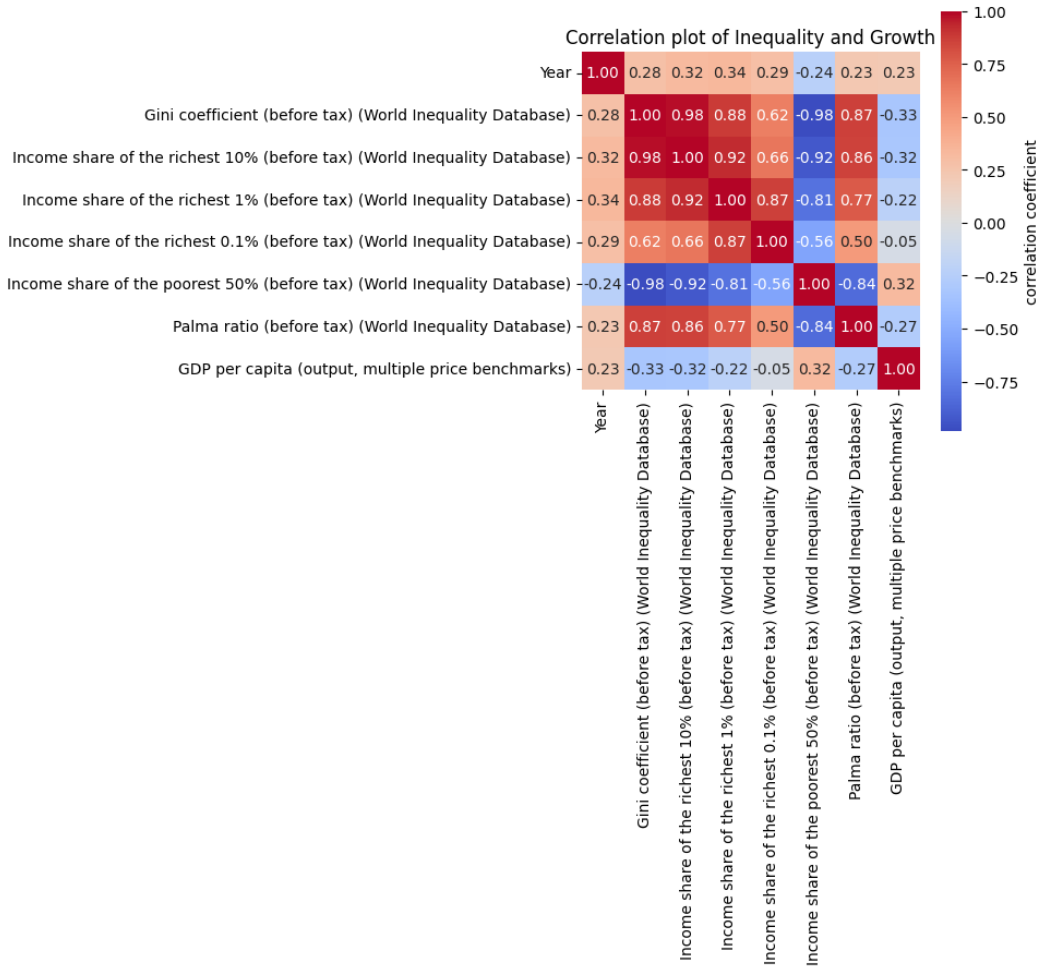

Рис. 6: Корреляция между таблицами Inequality и Growth

Удалим зависимые столбцы, а также столбцы со строковыми данными. Получим итоговую таблицу (https://github.com/tsisarukm/Project1/blob/9a268af9313ade321bd8c3f6b061790afaa657e6/Linear%20regression.numbers), по которой с помощью библиотеки sklearn построим линейную регрессию.

```
                                            OLS Regression Results
==============================================================================================================================
Dep. Variable:     GDP per capita (output, multiple price benchmarks)   R-squared (uncentered):                   0.685
Model:                                                            OLS   Adj. R-squared (uncentered):              0.685
Method:                                                 Least Squares   F-statistic:                              732.2
Date:                                                Fri, 10 Nov 2023   Prob (F-statistic):                        0.00
Time:                                                        14:56:20   Log-Likelihood:                         -22362.
No. Observations:                                                2022   AIC:                                  4.474e+04
Df Residuals:                                                    2016   BIC:                                  4.477e+04
Df Model:                                                           6
Covariance Type:                                            nonrobust
==============================================================================================================================
                                                                                coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------------------------------
Gini coefficient (before tax) (World Inequality Database)                 4.992e+04   1.59e+04      3.144      0.002    1.88e+04    8.11e+04
Income share of the richest 10% (before tax) (World Inequality Database)   -691.1815    263.893     -2.619      0.009   -1208.713    -173.650
Income share of the richest 1% (before tax) (World Inequality Database)    -911.2634    417.492     -2.183      0.029   -1730.025     -92.502
Income share of the richest 0.1% (before tax) (World Inequality Database)  3118.2559    472.064      6.606      0.000    2192.471    4044.041
Income share of the poorest 50% (before tax) (World Inequality Database)   1116.1179     70.993     15.722      0.000     976.891    1255.345
Palma ratio (before tax) (World Inequality Database)                        503.3171    205.333      2.451      0.014     100.629     906.005
==============================================================================
Omnibus:                      586.865   Durbin-Watson:                   2.068
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2280.301
Skew:                           1.373   Prob(JB):                         0.00
Kurtosis:                       7.418   Cond. No.                     2.14e+03
==============================================================================

Notes:
[1] R² 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, 2.14e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Рис. 7: Результаты линейной регрессии

5 Conclusion

Итак, на (Рис. 6) видно, что коэффициент корреляции между уровнем ВВП и коэффициентом Джини равен -0.33, то есть зависимость умеренная. В используемых датасетах нет дополнительной информации, которую можно было бы использовать для построения линейной регрессии и прогнозирования уровня ВВП, который мы использовали, как показатель экономического развития. Вследствие этого при использовании метода наименьших квадратов коэффициент детерминации (R-squared (uncentered)) равен 0.685, однако как правило требуется более высокий уровень дисперсии для того, чтобы модель считалась хорошей.

Таким образом, мы рассмотрели такие экономические показатели, как коэффициент Джини и уровень ВВП, проанализировали их значения в разных странах, а также их изменение с течением времени. Кроме того, было установлено, что прямая взаимосвязь между данными характеристиками отсутствует. Ввиду отсутствия дополнительной информации о других экономических показателях в датасете не удалось построить точную линейную регрессию. Для использования метода наименьших квадратов требуется больше информации.