Econometrics Project
Group - 8

Investigating Trends in Groundwater Quality

Aditya Sharma
2022038

Aryaan Bazaz
2022108

Manik Sharma
2021336

Poorvi Kumar
2021343

Overview

00 Introduction

01 Data Merging

02 Regressing GWQ indicator on SDP

03 Plotting GWQ indicator vs SDP

Plotting $\hat{u}i,t$ vs SDP

04 verifing that $\Sigma i,t \; \hat{u}i,t = 0$.

05 Hypothesis

06 Kuznets's curve

07 Kuznet's curve by Year

08 Relation b/w economic growth and GWQ in enhanced model

09 Kuznet Curve By-Region

10 Limitations

Introduction

The project aims to analyze **groundwater quality (GWQ)** data for **Indian districts** from **2000 to 2018** and its relationship with **economic indicators**. It involves **merging GWQ data with state-level economic output** and **district-level Gini index values**, followed by regression analysis to estimate the GWQ-SDP relationship. **Visualizations and statistical tests** are conducted to **validate the model.** Additionally, the model is enhanced to explore **non-linear relationships** (Kuznets curve) between **environmental quality and economic growth, considering regional variations** within India. The project provides insights into the dynamic interaction between economic development and environmental sustainability.

Merging The Data

Data we had :

- GWQ indicator(residualsodiumcarbonate) district wise from 2000 - 2018
- State Wise GDP from 1980-81 to 2021-22
- Gini Index for states/districts 2010 - 11

Merging the data:

1. Removing entries with empty residualsodiumcarbonate indicator.

	has_data.csv	missing_data.csv	District-Level_GWQ_AllYears.csv
Entries Count	5609 (47%)	6100(53%)	11709

State	Uttar Pradesh	Telangana	Karnataka	Bihar	Odisha	West Bengal	Jharkhand	Assam
missing_count	1190	469	457	421	413	386	310	257

- The main reason for data gaps in the leading states is the high population, which increases the chances of missing observations. Additionally, inadequate data collection is exacerbated by poor development.

- To standardize SDP across different base years, SDP values from 1999-2000 to 2021-2022 are utilized. This is done by multiplying all entries in table1 by a constant value, determined as c = SDP(table2) / SDP(table1). The intersection of table1 and table2 corresponds to the same year, with the base year of table2 being later than that of table1.

- Combining data from residual sodium carbonate and SDP based on ["year", "state"].

- Merging with Gini Index (2011) [left_on='state', right_on='Districts/State'].

Regress the GWQ indicator on SDP

RSCcap = β0cap +β1cap SPD +Ucap

Interpretation : $beta_1$ (cap) gdp < 0 which suggests that as residual sodium carbonate increases there is an decrease in SDP for that state.

No. Observations:	R-squared	Adj. R-squared:	F-statistic
5130	0.017	0.017	87.70

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.3593	0.205	16.355	0.000	2.957	3.762
gdp	-3.953e-06	4.22e-07	-9.365	0.000	-4.78e-06	-3.13e-06

	mean	median
gdp	364721.87	259230.1691
GWQ-I	1.91737	1.4425001

alpha	const	gdp
0.1	significant	significant
0.01	significant	significant
0.05	significant	significant

Mean Residual Sodium Carbonate	1.917373
Median Residual Sodium Carbonate	1.4425

	const	gdp
P values	1.201081e-58	1.113494e-20

Percentiles of SDP	25th Percentile (Q1)	75th Percentile (Q3)	95th Percentile
SDP	165354.2083	462181.6751	970265.197

Regress the GWQ indicator on SDP

A low R-squared (R^2) value does not always signify a lack of relationship or a weak fit between the independent and dependent variables. Here are some reasons for a low R^2:

- **Non-linearity:** The relationship might be non-linear, which a linear model may not accurately capture.
- **Noisy Data:** High variability or noise in the data can make it challenging for the model to account for the variance.
- **Missing Variables:** Excluding crucial variables from the model can result in lower explained variance.
- **Heteroscedasticity:** Varied variance across levels of independent variables can diminish R^2.
- Outliers: Outliers can significantly impact the model's fit and reduce R^2.

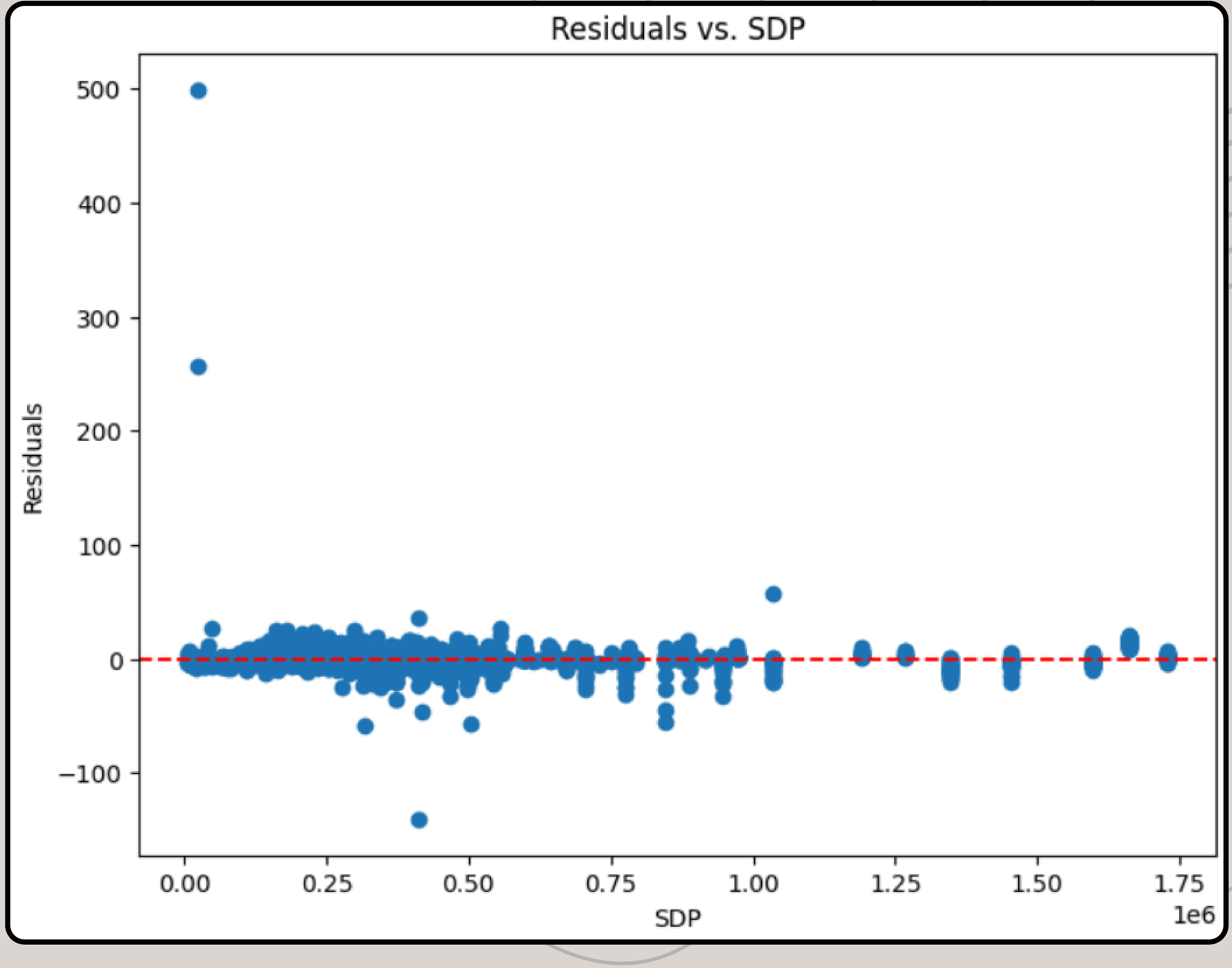
Therefore, while a low R^2 indicates limited explained variance by the model, it does not necessarily mean no relationship or poor model performance. Further diagnostics and analyses are necessary for a thorough assessment.

Plotting Residual indicator vs SDP

Interpretation :

We get the data as per what we expect this is because the residual is around the line y = 0 which shows that the covariance between residuals and SDP is 0.

Cov(x,u) = 0



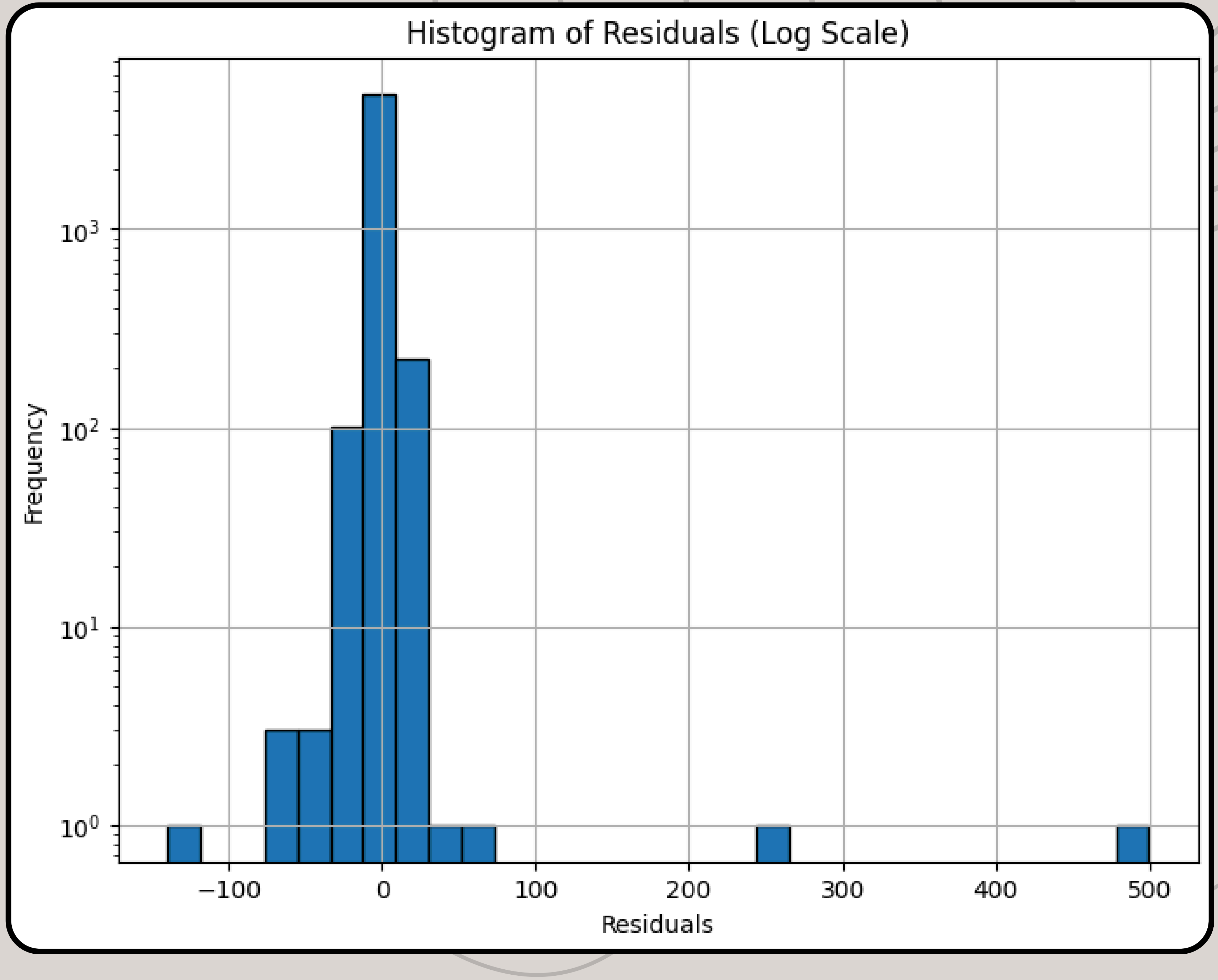
verifing that $\sum_{i,t} \hat{u}_{i,t} = 0$

We know that

Sum of Residuals= **Σ** (*yi*−*y^i*)

Sum of residuals: -1.490 × 10^{-11}

The value of SSR is almost 0 which verifies the fact that SSR



Kuzenet’s curve

Interpretation: We want to test for non-linear relationships b/w GWQ-I i.e residula sodium carbonate and GDP. We introduce new squared and cubed terms for GDP to check for this and run the regression again.

Regression Equation

RSCcap = β0cap +β1capSPD + β2capSPD^2 +β3capSPD^3 +β4capGINI +Ucap

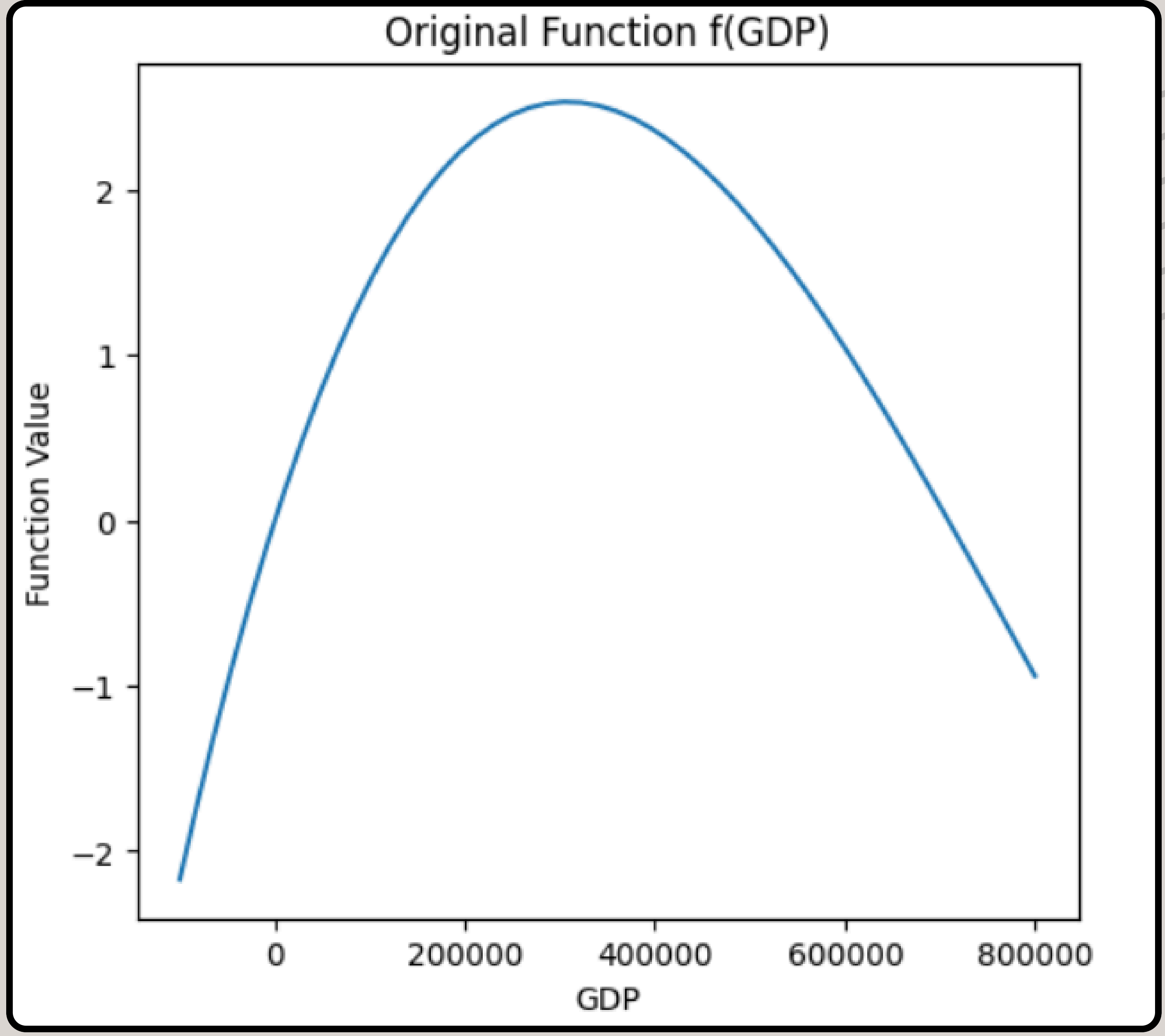
No. Observations:	R-squared	Adj. R-squared:	F-statistic
4843	0.017	0.017	41.87

Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.95e-11	6.94e-12	14.327	0.000	8.59e-11	1.13e-10
gdp	1.793e-05	1.25e-06	14.327	0.053	1.55e-05	2.04e-05
gdp_square	-3.629e-11	2.81e-12	-12.934	0.000	-4.18e-11	-3.08e-11
gdp_cube	1.551e-17	1.35e-18	11.474	0.000	1.29e-17	1.82e-17
Gini	3.015e-11	2.1e-12	14.327	0.000	2.6e-11	3.43e-11

Interpretation : The curve increases and then decreases. There is a saturation point in curve. Which can be found out by double derivative test. We get our expected inverted ‘U’ shape.

The value of GDP for which the value of kuznet curve saturates is : **3,682,215.25**



Plotting the curve

[EXTRA]Enhancing the model

We can try to induce other terms to better fit the mode. For example in this example we introuduce a new term of gini_gdp = “gini”*””gdp”

New Regression Equation

$$\text{residualsodiumcarbonate} = \hat{\beta}_0 + \hat{\beta}_1\text{gdp} + \hat{\beta}_2\text{gdp}^2 + \hat{\beta}_3\text{gdp}^3 + \hat{\beta}_4\text{gdp}\backslash\_\text{gini} + \hat{\beta}_5\text{Gini}$$

Summary results

No. Observations:	R-squared	Adj. R-squared:	F-statistic
5130	0.023	0.023	**41.02**

Summary results

There is a significant increase in R(squared) after introducing gini*gdp in the model. Which suggests presence of vairables that are not included in the model

R-squared(original)	R-sqaured(new)
0.017	0.023

Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.394e-11	1.12e-11	4.807	0.000	3.19e-11	7.59e-11
gdp	4.911e-06	2.53e-06	1.938	0.053	-5.79e-08	9.88e-06
gdp_square	4.435e-11	2.86e-12	-15.520	0.000	-5e-11	-3.88e-11
gdp_cube	1.822e-17	1.37e-18	13.321	0.000	1.55e-17	2.09e-17
gdp_gini	5.12e-05	7.71e-06	6.641	0.000	3.61e-05	6.63e-05
Gini	1.207e-10	1.49e-11	8.122	0.000	9.15e-1	1.5e-10

Interpretation: A Look at the p-value associated with the coefficient of the gdp_gini variable in the regression output. A low p-value (typically less than 0.05) indicates that the coefficient is statistically significant, suggesting that the interaction term is likely important in explaining the variation in the dependent variable.

Outliers & Influential Observations

FINDING OUTLIERS AND INFLUENTIAL OBSERVATIONS

Identifying Outliers and Influential Observations in OLS Model

1. **Residual Analysis**
    1. Plot the residuals against the fitted values. Outliers often show up as points with large residuals.
    2. Look for patterns in the residuals. Random scatter indicates model assumptions are likely met.
2. **Leverage**
    1. Calculate leverage for each observation. High leverage values indicate extreme values on independent variables.
3. **DFBETAS**
    1. Measures change in coefficients when an observation is excluded.
    2. DFBETAS values greater than threshold are considered influential.
4. **DFITS**
    1. Measures influence of each observation on predicted values.
    2. Observations with high DFITS values are influential.
5. **Studentized Residuals**
    1. Residuals divided by their standard errors.
    2. Observations with studentized residuals greater than threshold are potential outliers..

Running regression after removing outliers

Summary results

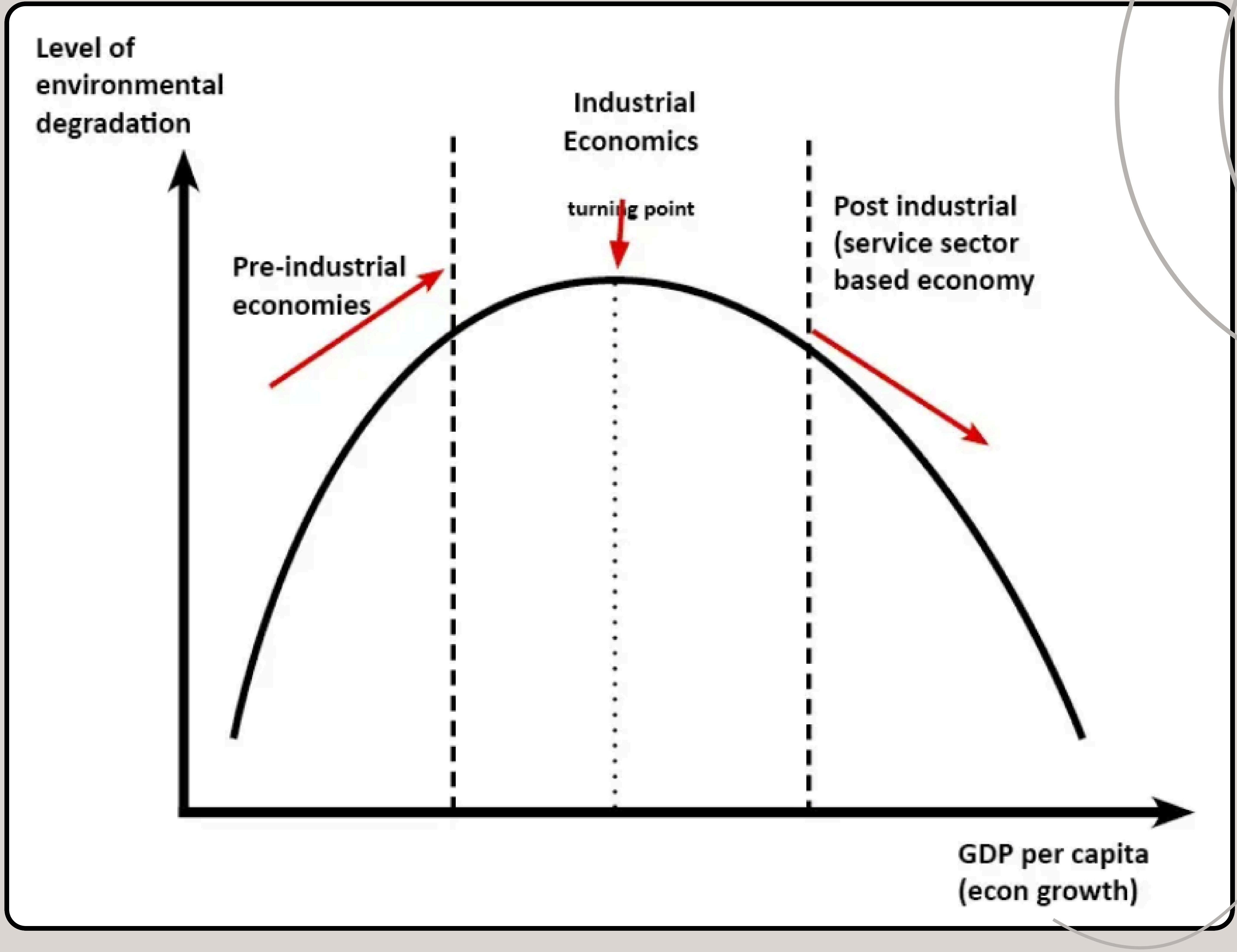
No. Observations:	R-squared	Adj. R-squared:	F-statistic
5130	0.045	0.023	**41.02**

Comparing from previous results

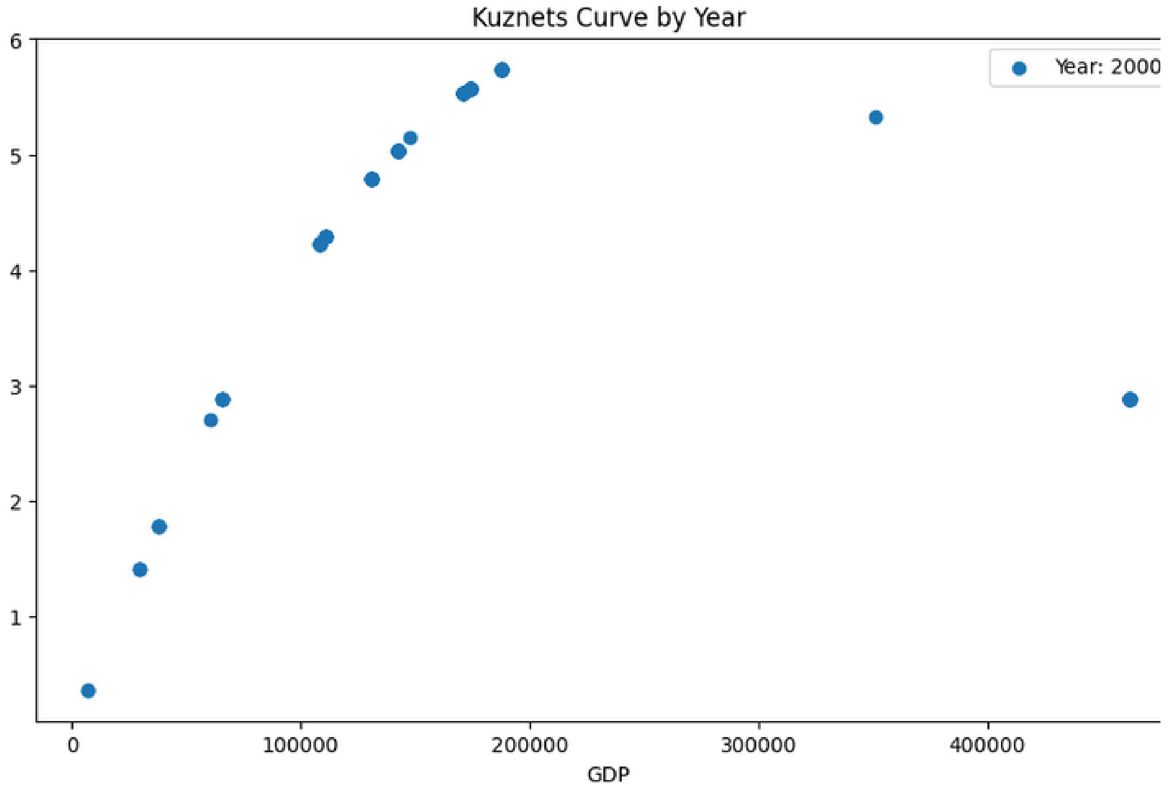
We can see a significant increase in R^2 after removing the outliers from data sets.

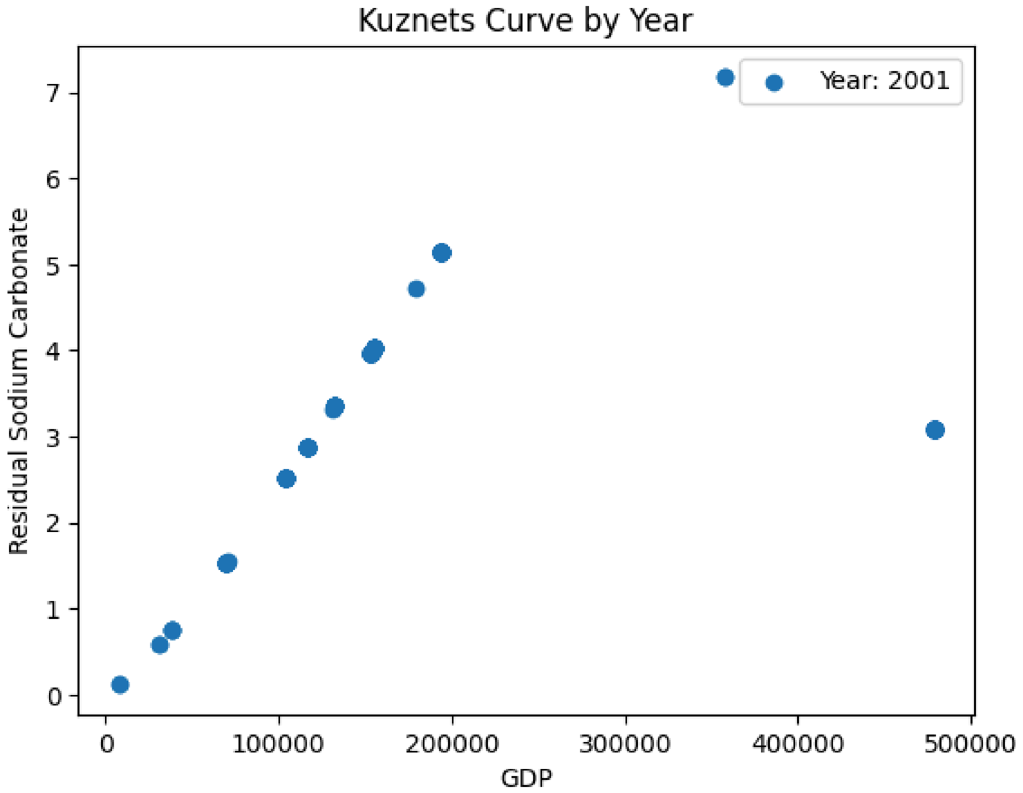
R-squared(NEW)	Adj. R-squared:
0.045	0.023

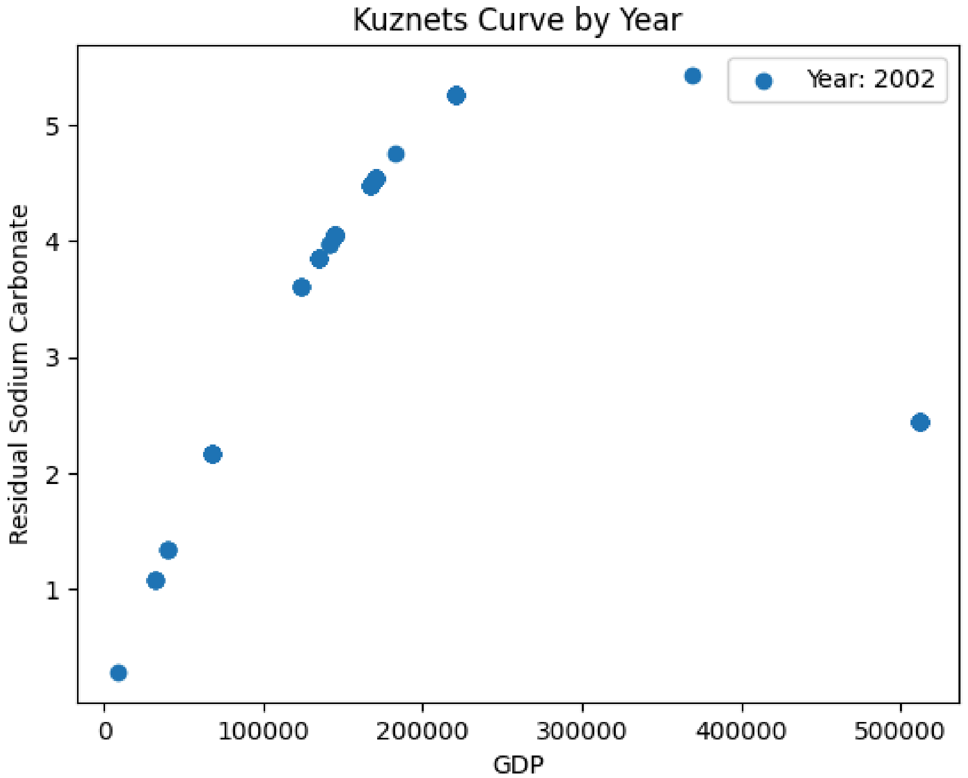
Kuznet Curve

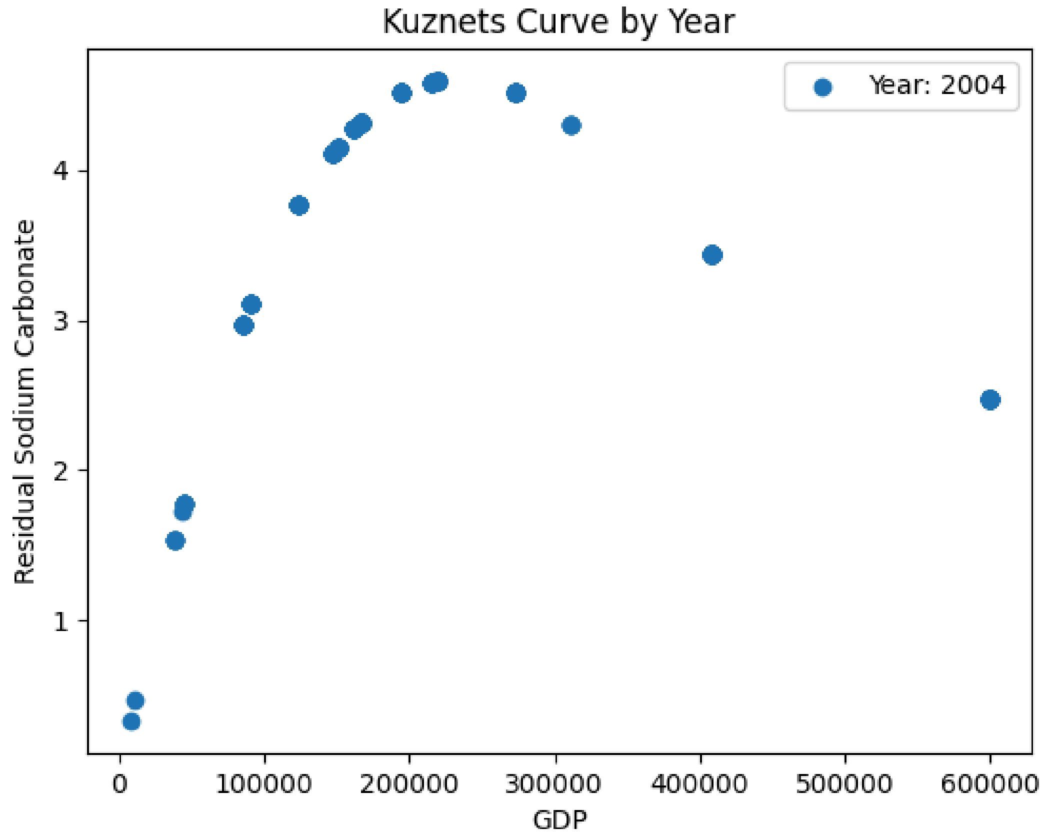

Level of environmental degradation
Industrial Economics
turning point
Pre-industrial economies
Post industrial (service sector based economy
GDP per capita (econ growth)

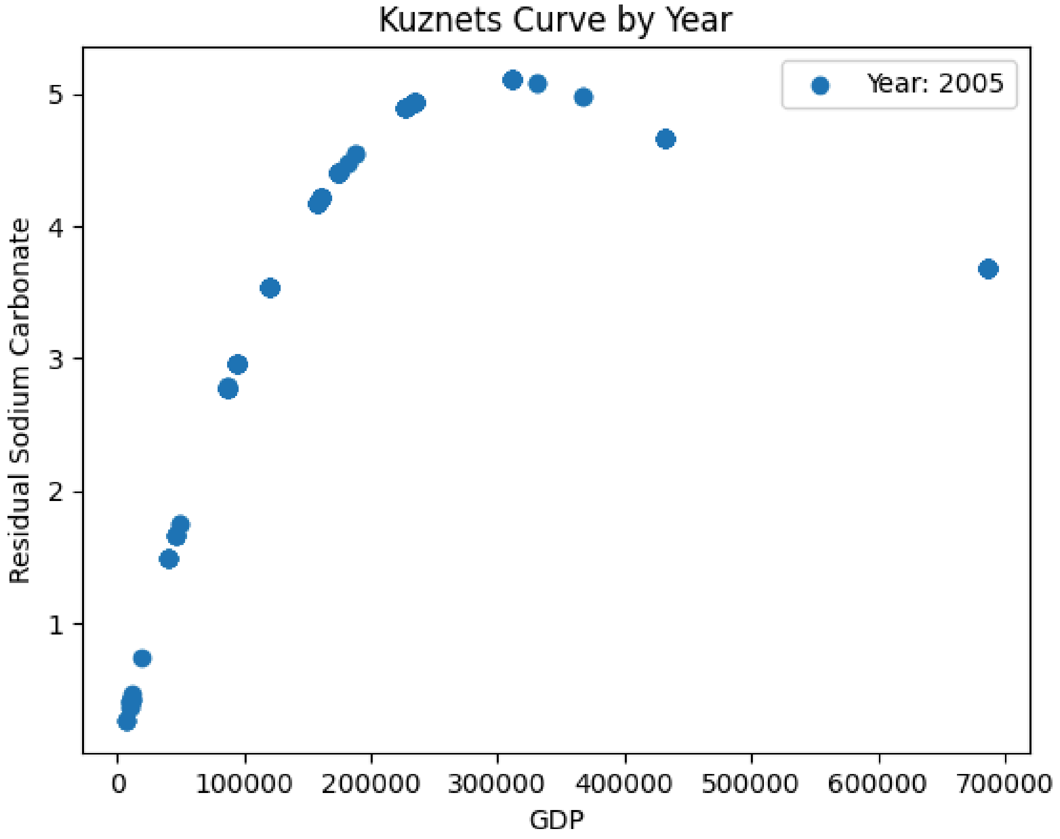
Kuznet Curve by Year

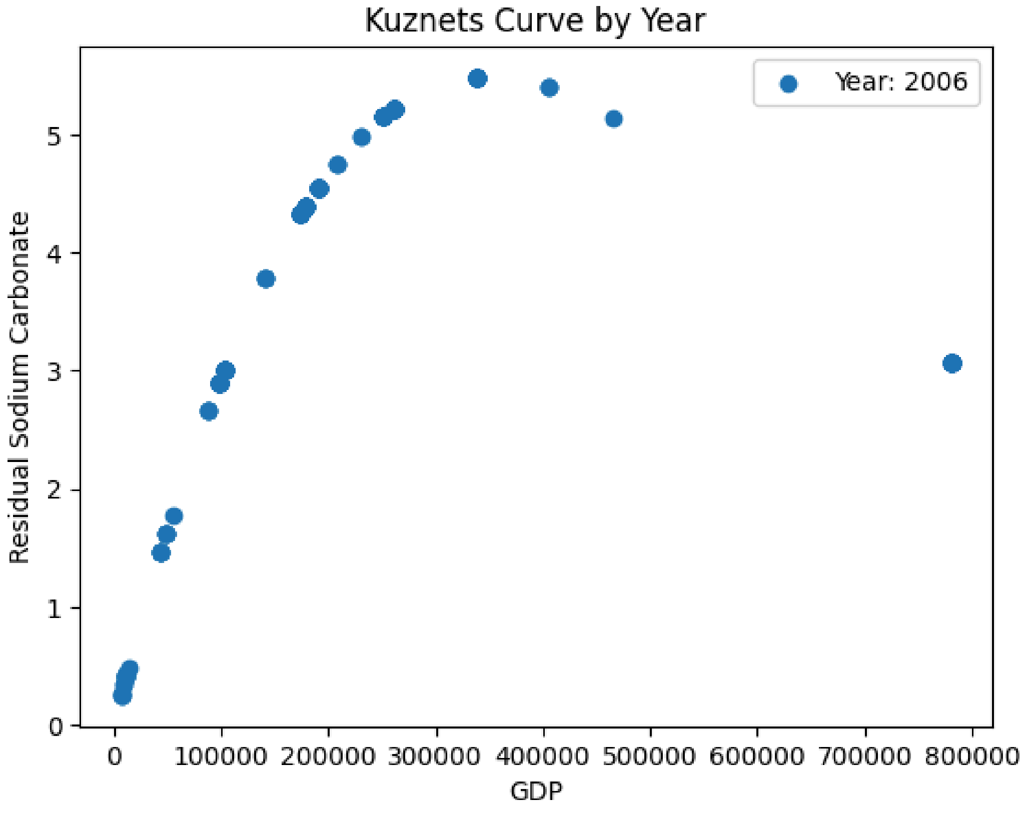




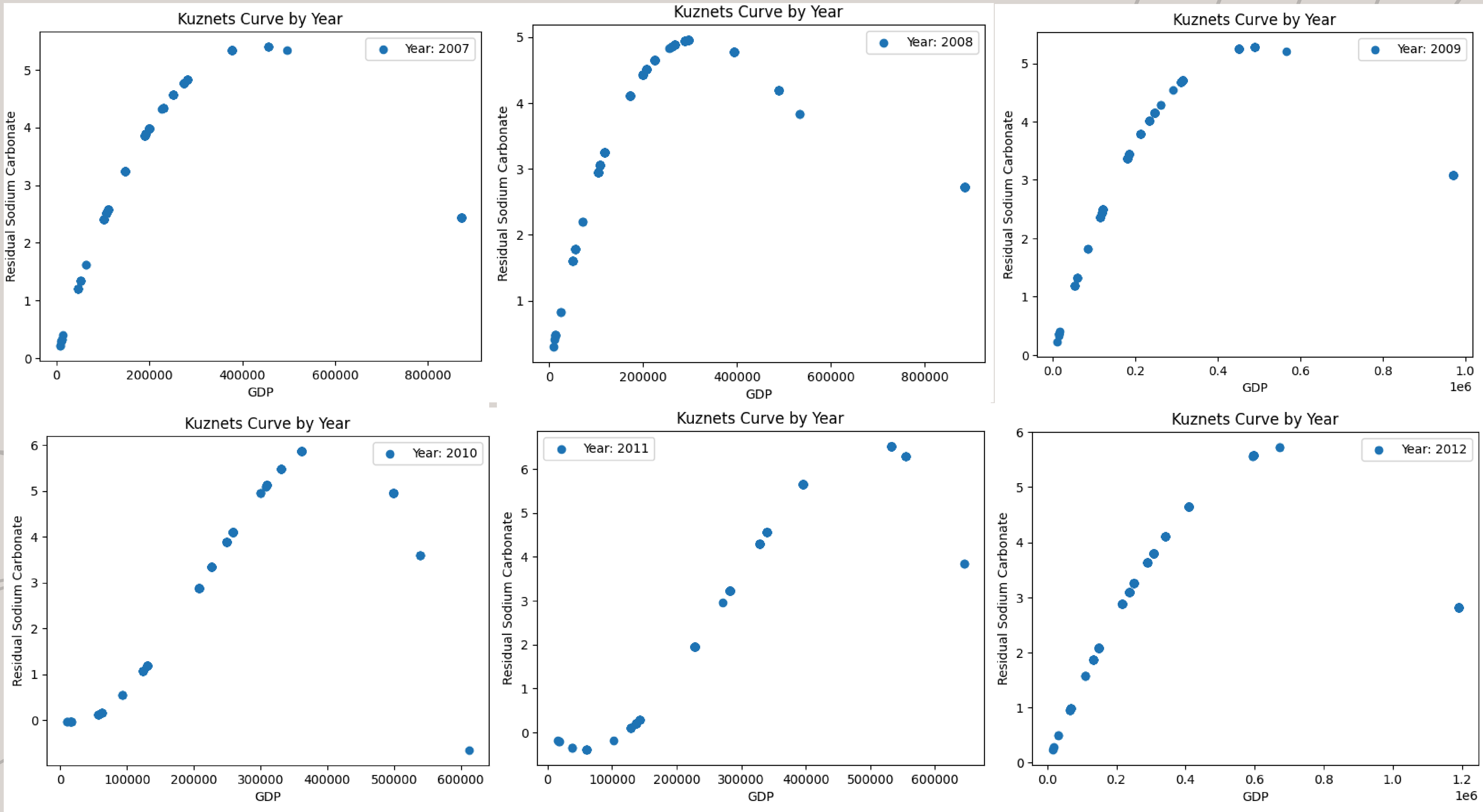




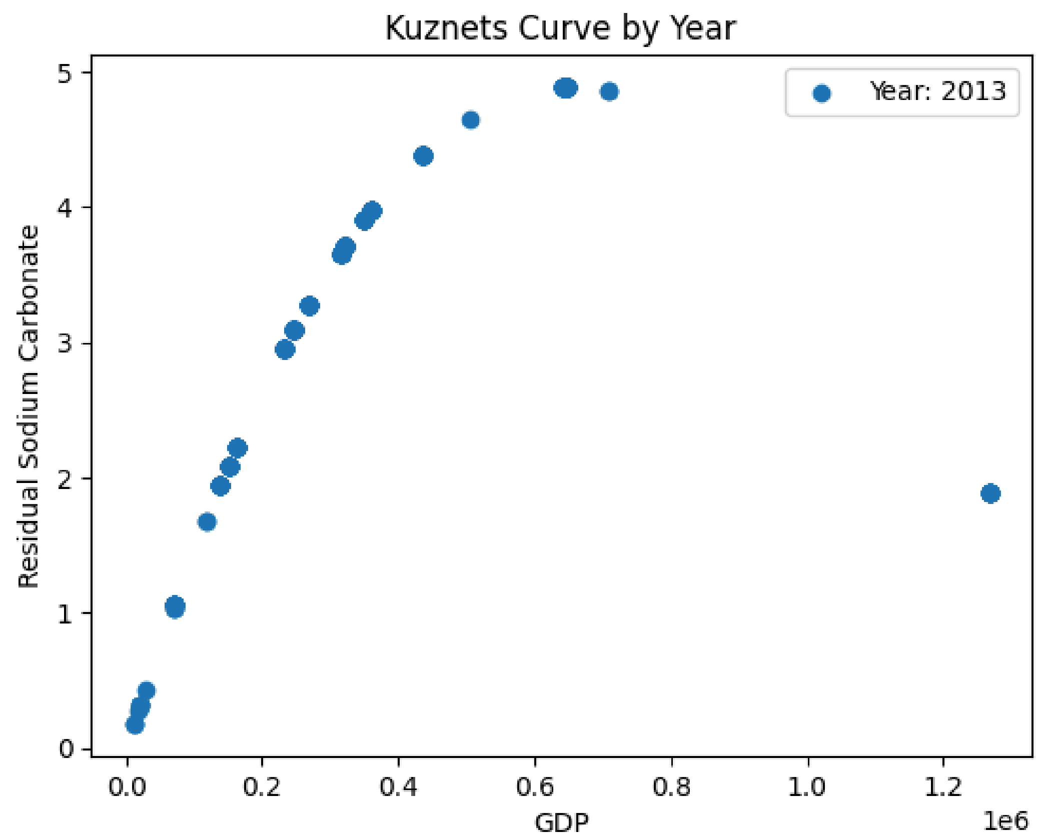


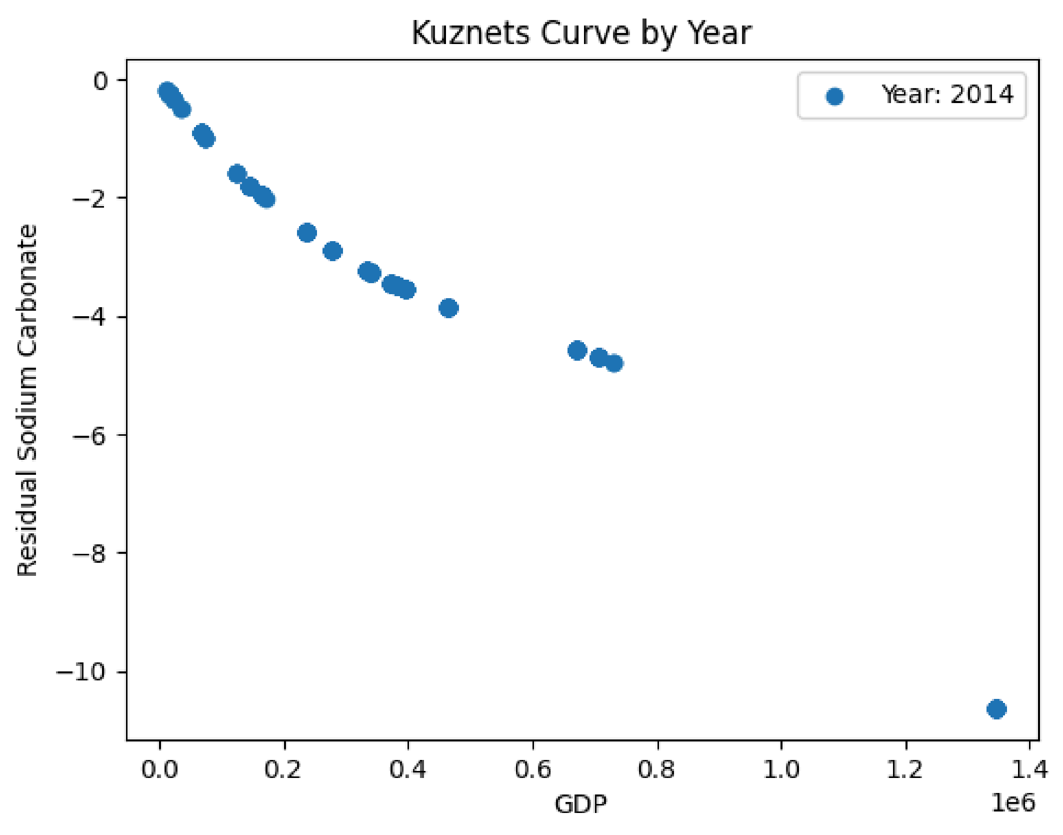


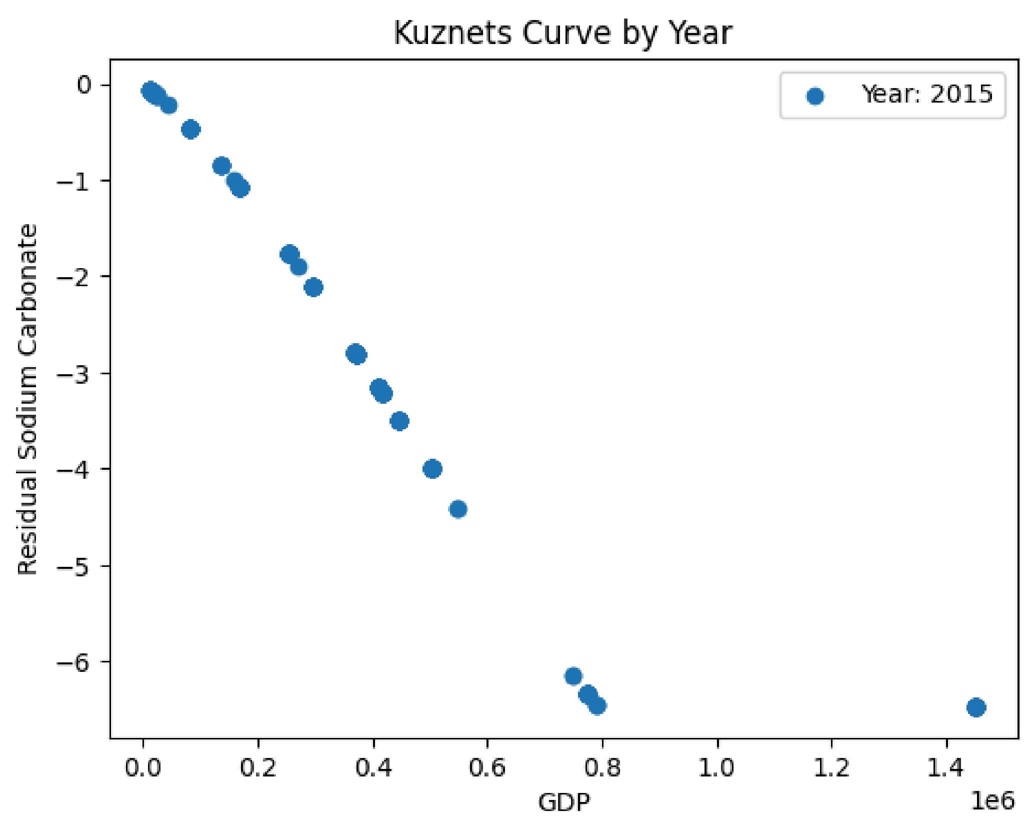
Kuznet Curve by Year

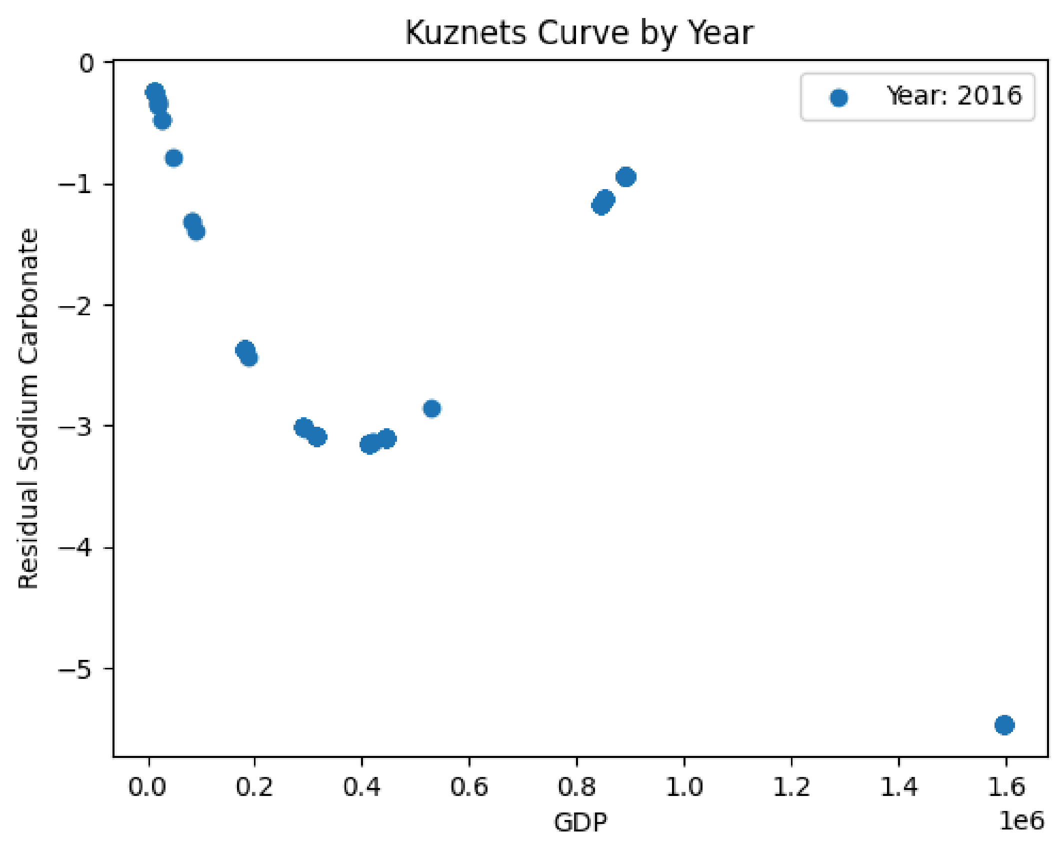


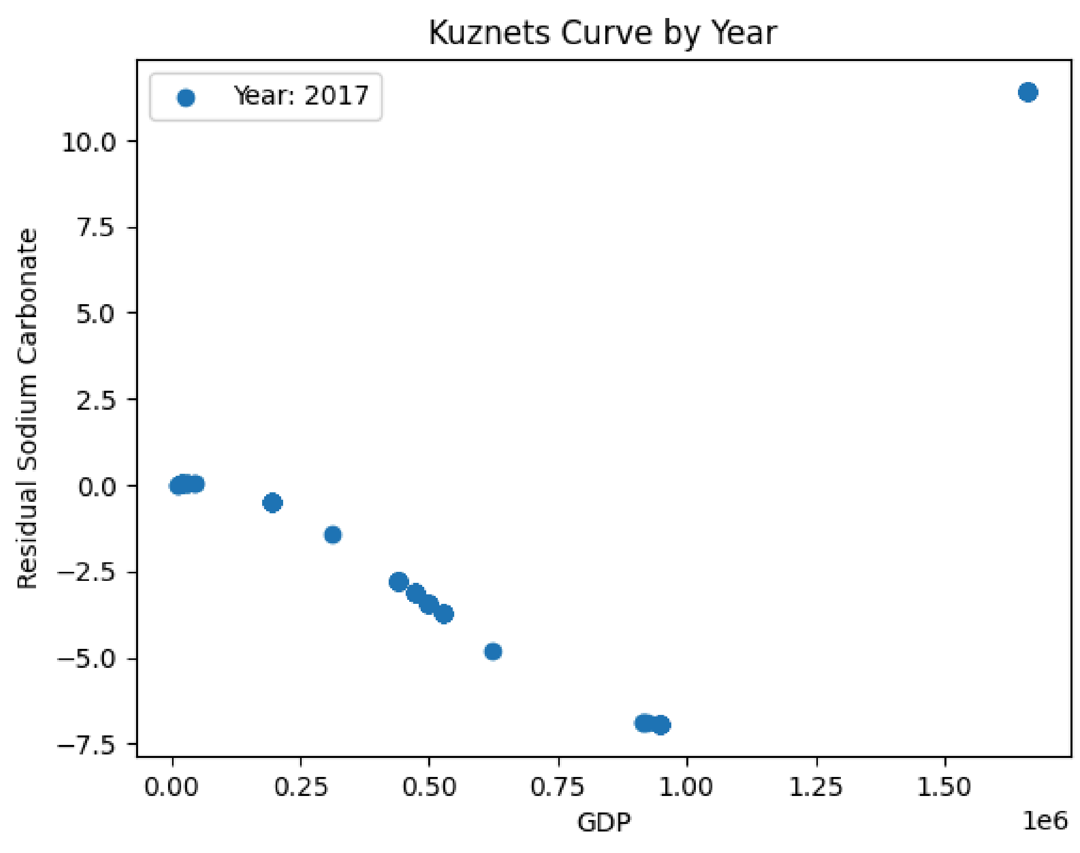
Kuznet Curve by Year

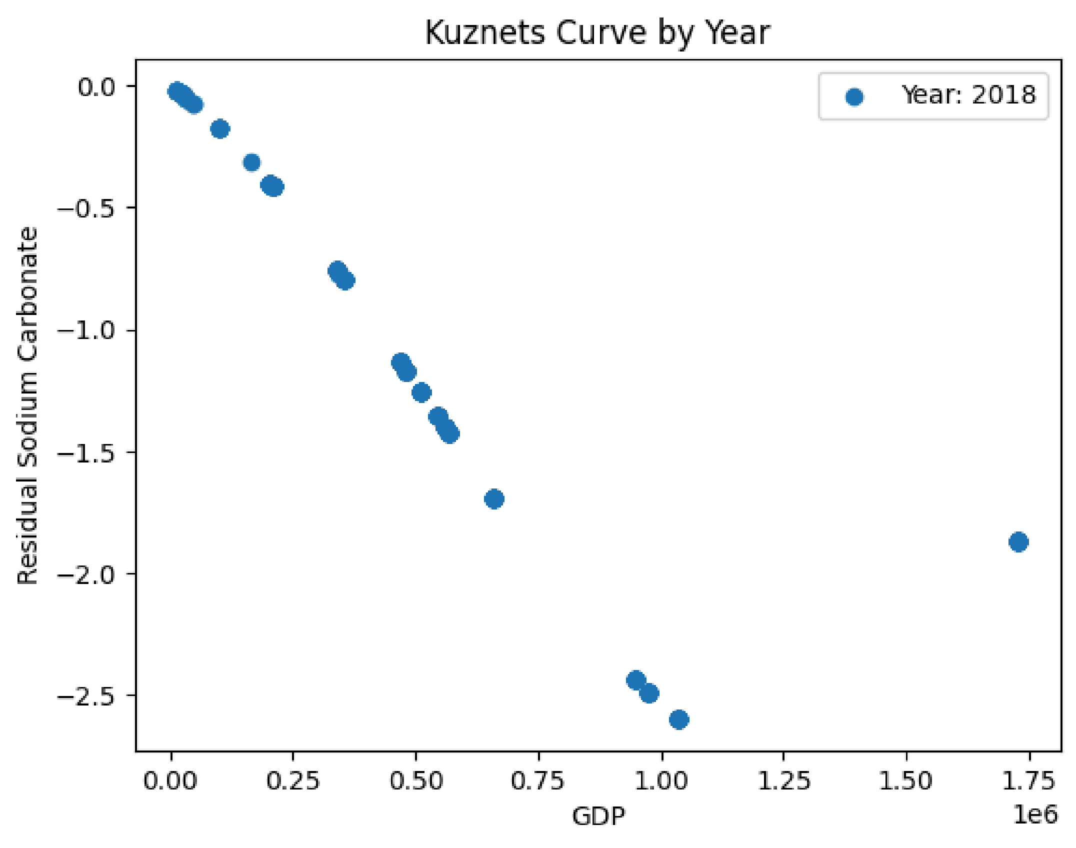












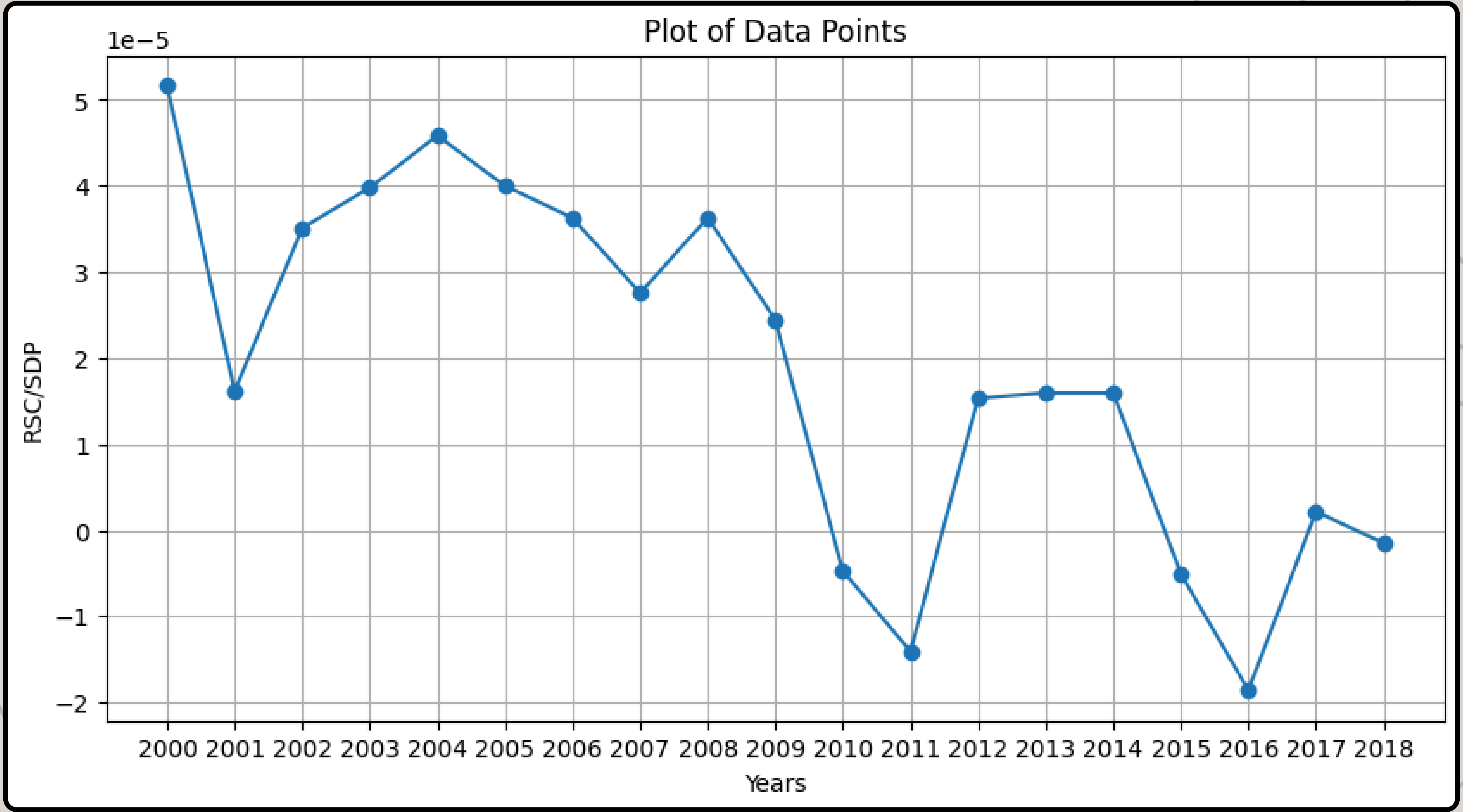
Regression Results

YEAR	B^0	B^1	B^2	B^3	B^4	R**2
2000	9.175e-10(2.15e-10)	5.166e-05(1.21e-05)	-1.217e-10(1.03e-10)	5.073e-17(1.67e-16)	3.007e-10(7.05e-11)	0.129
2001	2.79e-10(1.87e-10)	1.618e-05(1.08e-05)	1.033e-10 (8.98e-11)	-2.583e-16 (1.42e-16)	8.954e-11(5.99e-11)	0.096
2002	5.832e-10(1.83e-10)	3.511e-05(1.1e-05)	-4.482e-11 (8.57e-11)	-2.826e-17(1.27e-16)	1.964e-10(6.15e-11)	0.125
2003	8.012e-10(5.86e-10)	3.981e-05(2.91e-05)	-1.809e-10(1.72e-10)	2.587e-16 (2.27e-16)	2.292e-10(1.68e-10)	0.232
2004	5.275e-10(6.46e-11)	4.584e-05(5.61e-06)	-1.39e-10(3.17e-11)	1.158e-16(3.92e-17)	1.669e-10 (2.04e-11)	0.057
2005	4.843e-10(6.32e-11)	4.002e-05(5.22e-06)	-9.687e-11 (2.85e-11)	6.757e-17 (3.15e-17)	1.477e-10 (1.93e-11)	0.111
2006	4.397e-10 (7.24e-11)	3.628e-05(5.98e-06)	-7.308e-11(3.1e-11)	4.053e-17 (3.05e-17)	1.326e-10 (2.18e-11)	0.149
2007	2.941e-10 (4.51e-11)	2.757e-05(4.23e-06)	-4.113e-11(2.03e-11)	1.458e-17(1.8e-17)	8.828e-11(1.35e-11)	0.152
2008	3.641e-10(4.63e-11)	3.63e-05(4.62e-06)	-8.048e-11(2.08e-11)	4.852e-17(1.8e-17)	1.065e-10 (1.35e-11)	0.132
2009	2.186e-10(2.72e-11)	2.444e-05(3.04e-06)	-3.397e-11(1.24e-11)	1.241e-17(9.83e-18)	6.432e-11 (7.99e-12)	0.141
2010	-5.869e-11(9.55e-11)	-4.745e-06(7.72e-06)	1.331e-10 (4.41e-11)	-2.076e-16(5.85e-17)	-1.705e-11 (2.77e-11)	0.175
2011	-1.534e-10(9.11e-11)	-1.408e-05(8.36e-06)	1.363e-10(4.35e-11)	-1.632e-16 (5.31e-17)	-4.384e-11(2.6e-11)	0.216
2012	1.131e-10(1.91e-11)	1.535e-05(2.59e-06)	-9.189e-12(8.81e-12)	-1.448e-18 (5.74e-18)	3.431e-11 (5.79e-12)	0.135
2013	1.068e-10(1.21e-11)	1.596e-05(1.8e-06)	-1.463e-11(5.36e-12)	2.529e-18(3.29e-18)	3.008e-11(3.4e-12)	0.206
2014	-9.209e-11(2.03e-11)	1.596e-05(3.24e-06)	1.831e-11(8.98e-12)	-9.874e-18(5.03e-18)	-2.615e-11(5.77e-12)	0.187
2015	-2.952e-11(2.4e-11)	-5.036e-06(4.09e-06)	-9.124e-12(1.09e-11)	6.558e-18(5.68e-18)	-8.621e-12(7e-12)	0.086
2016	-1.009e-10(4.82e-11)	-1.861e-05(8.89e-06)	3.262e-11(1.68e-11)	-1.448e-17(7.22e-18)	-3.063e-11(1.46e-11)	0.022
2017	9.15e-12 (1.46e-11)	2.134e-06 (3.42e-06)	-2.701e-11 (7.38e-12)	1.798e-17 (3.29e-18)	2.679e-12 (4.29e-12)	0.587
2018	-6.462e-12 (5.06e-11)	-1.483e-06 (1.16e-05)	-2.816e-12 (2.31e-11)	1.763e-18 (9.85e-18)	-1.876e-12 (1.47e-11)	0.003

Regression Results

	Mean	Median
GDP (B^1)	364721.88	259230.17
GDP^2 (B^2)	236725497696.05	67200280571.62
GDP^3 (B^3)	2.32e+17	1.742e+16
Gini (B^4)	0.314	0.31

Kuznet Curve by Year



Interpretation: Kuznet Curve by Year

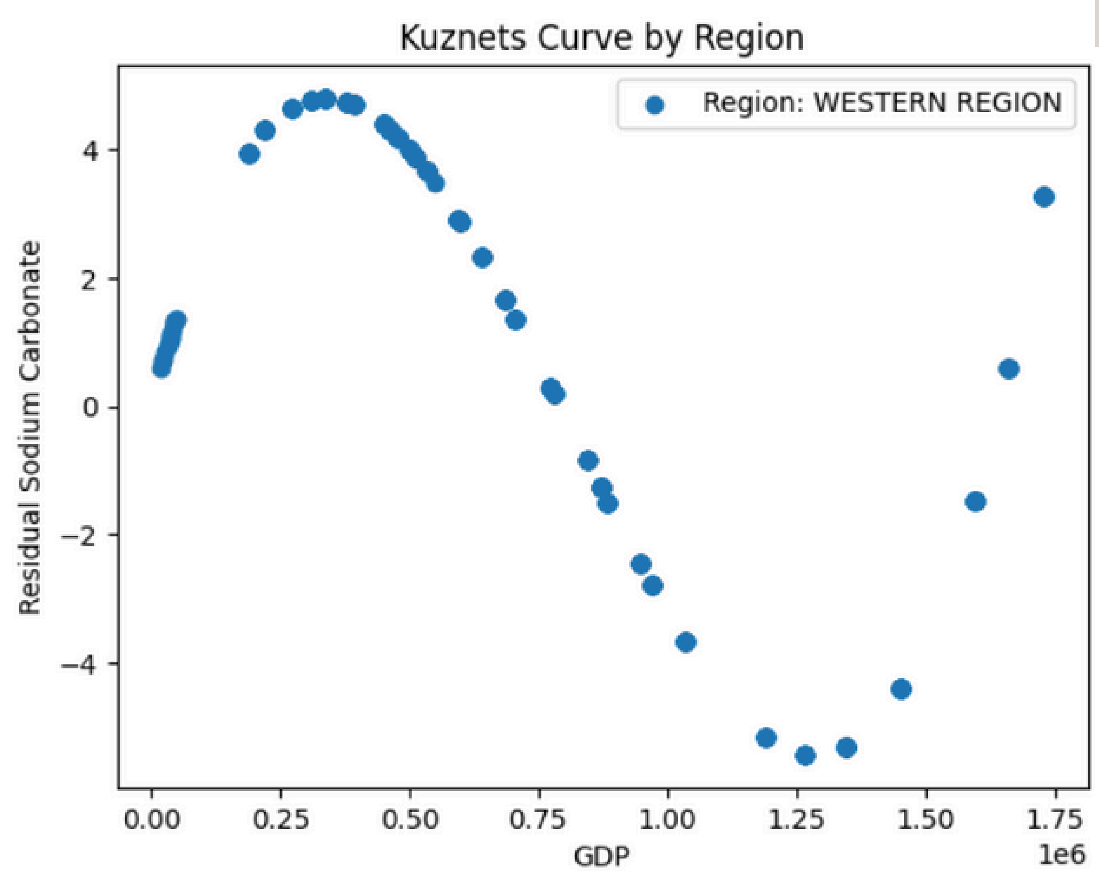
- From the plots in the previous slides, we see that the general trend is that as the years increase, the GDP turning point value also increases.
- This indicates that in the given span of time, the residual sodium carbonate value increases by such an amount that for the turning point to arrive, the GDP value is observed to be higher.
- For some years, it can be seen that the turning point cannot be identified easily this is mainly because the data available for that year was not sufficient.

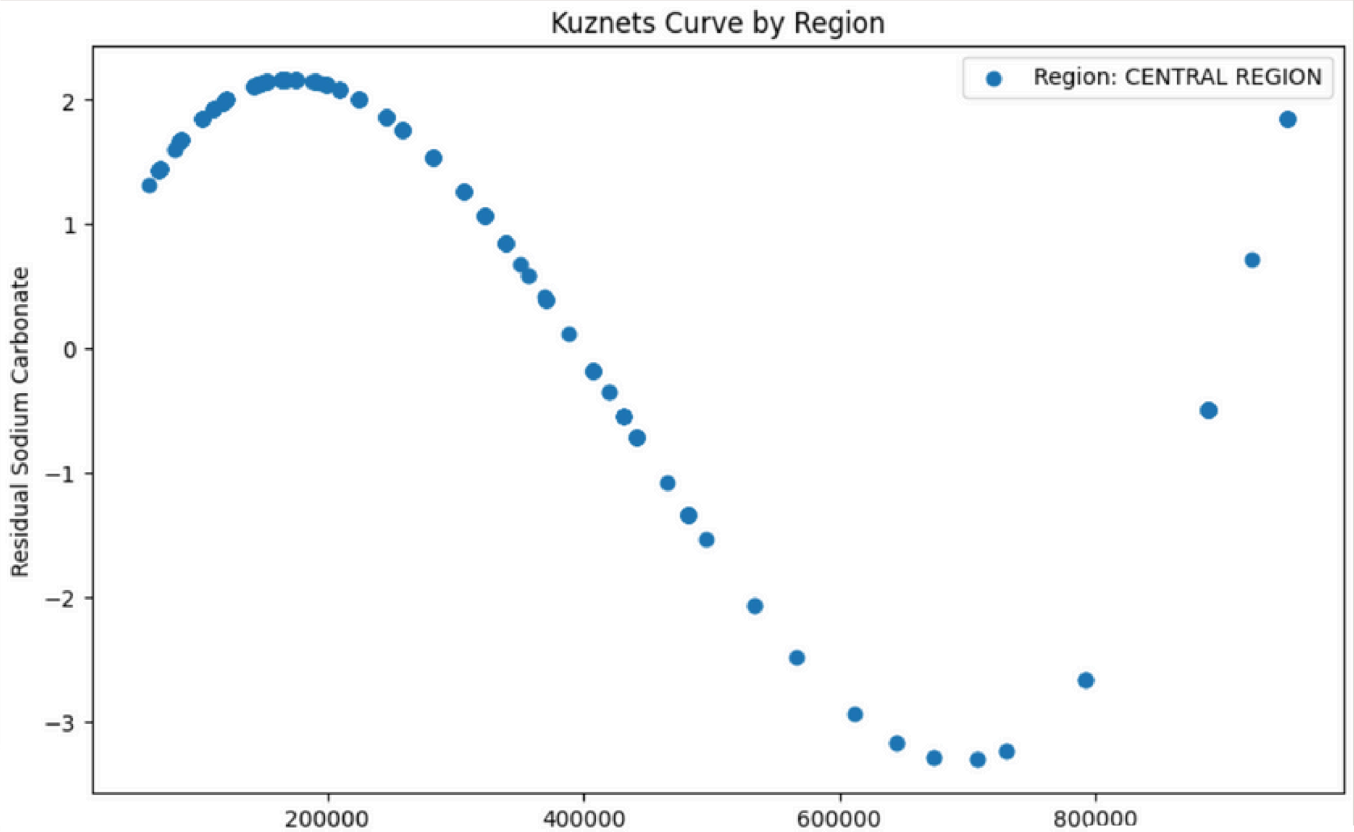
Approach for Regression

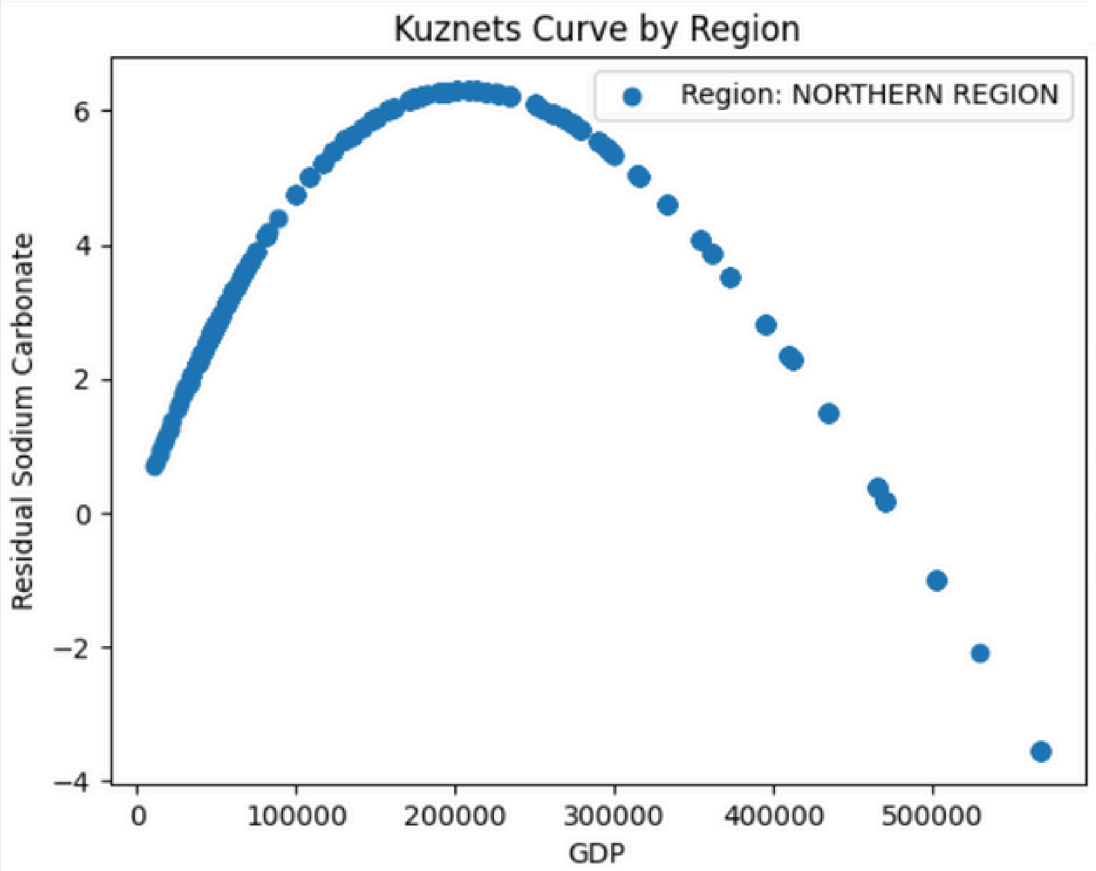
- Using dummy variables AND by articulating the data points according on the given year in the dataset
- But the restriction of using the dummy variable is that we had to incorporate various interaction terms between the dummy variables(18 dummy variables) and 4 other variables(SDP,SDP2,SDP3) Hence we would have got a total of 94 variables which would not have been feasible. Hence we had grouped the data points according to the years and then ran a regression individually for each year which helped us get a better graphical understanding of the kuznet curve and its shifting.

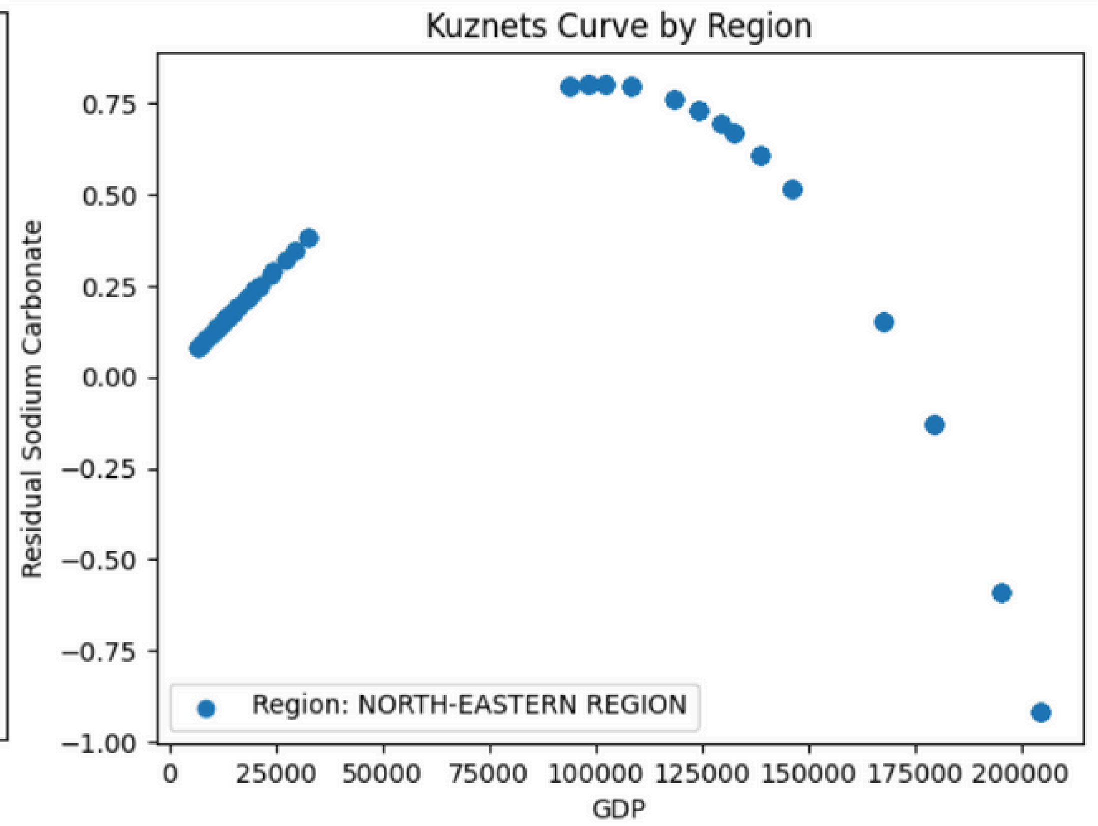
Kuznet Curve By-Region

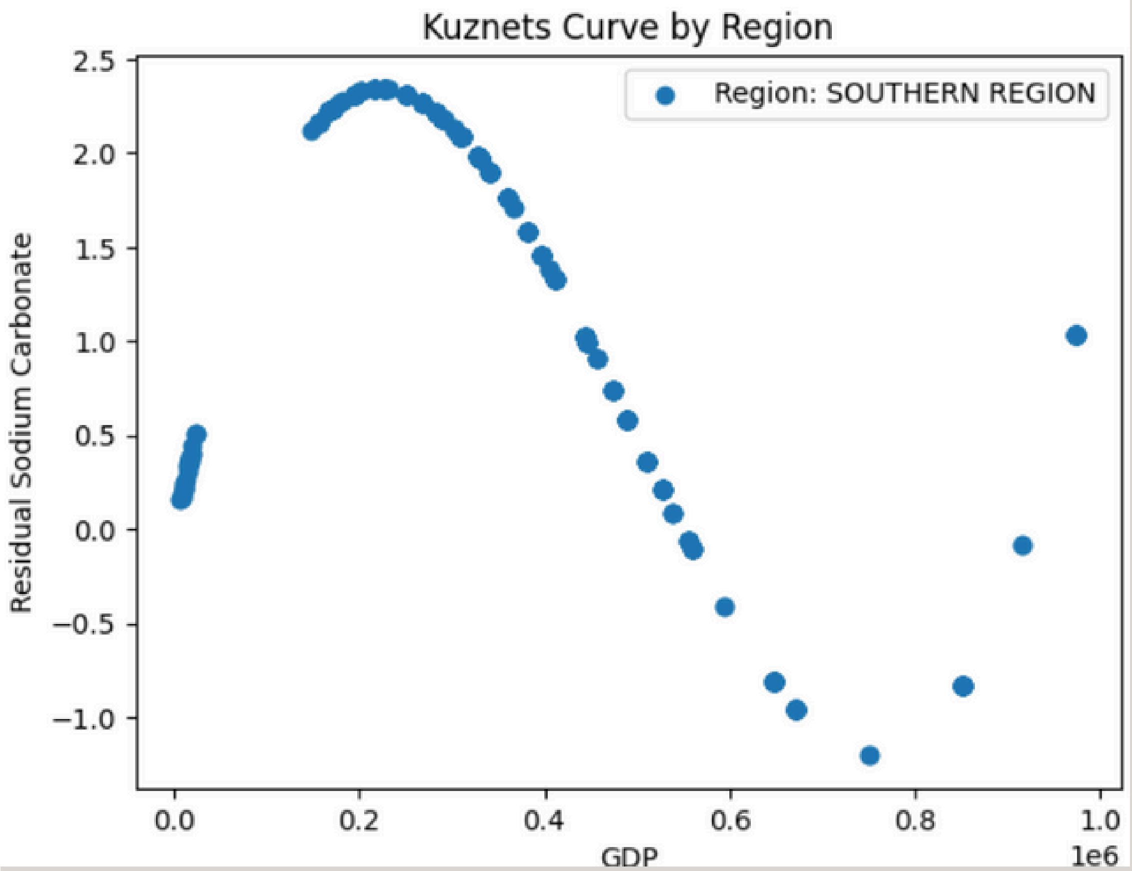
residualsodiumcarbonate= beta0 + beta1* gdp+ beta2*gdp^2+beta3*gdp*3+beta4*gini











Kuznet Curve By-Region(Grouping)

Region	B^0	B^1	B^2	B^3	B^4	R**2
Central	2.497e-10(1.36e-11)	2.474e-05(1.5e-06)	-9.964e-11(6.07e-12)	7.666e-17(5e-18)	8.749e-11(4.63e-12)	0.131
Eastern	4.206e-10(4.46e-11)	3.268e-05(3.46e-06)	-1.395e-10(2.19e-11)	1.369e-16 (2.6e-17)	1.197e-10(1.27e-11)	0.056
North- Eastern	6.426e--10(4.31e-10)	1.217e-05(8.19e-05)	-3.058e-12 (1.13e-10)	-3.828e-16(3.79e-16)	1.659e-10(1.11e-10)	0.089
Northern	8.591e-10(7.18e-11)	6.658e-05(5.56e-06)	-2.009e-10(3.23e-11)	1.28e-16 (4.38e-17)	2.607e-10(2.18e-11)	0.129
Southern	1.582e-10(1.13e-10)	2.338e-05(1.67e-05)	-6.803e-11(5.57e-11)	4.63e-17(4.23e-17)	5.259e-11 (3.75e-11)	0
Western	1.237e-10(8.14e-12)	3.115e-05(2.05e-06)	-5.828e-11 (3.94e-12)	2.393e-17 (1.7e-18)	3.955e-11 (2.6e-12)	0.188

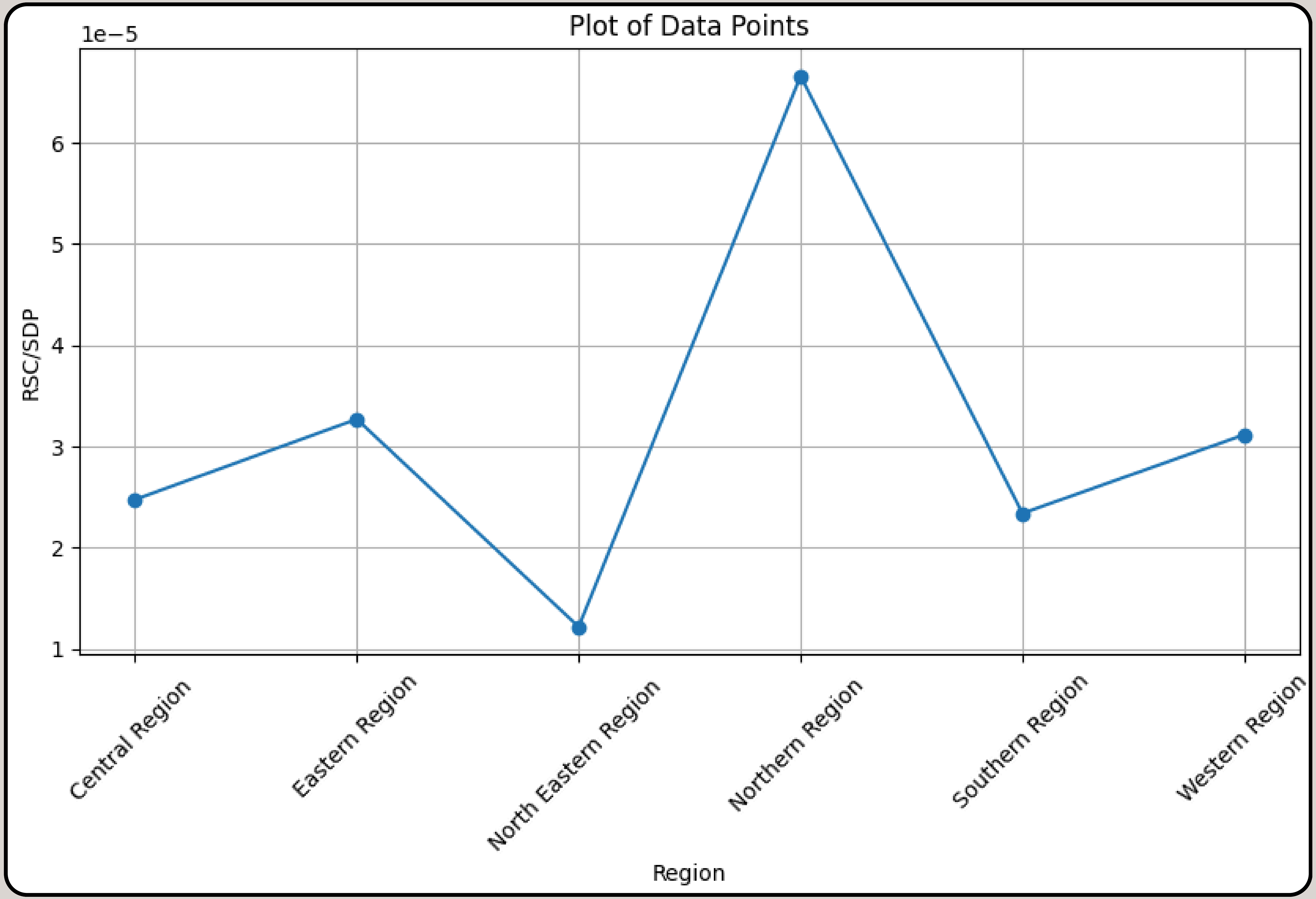
Regress the GWQ indicator on SDP, Gini, SDP*Gini and Region Dummy Variables

No. Observations:	R-squared	Adj. R-squared:	F-statistic
5130	0.045	0.044	30.37

- **Interpretation :** R-squared (0.045): This value indicates that approximately 4.5% of the variability in residual sodium carbonate can be explained by the model. It’s a relatively low value, suggesting that there are other factors not included in the model that influence the dependent variable.
- Adjusted R-squared (0.044): This is a modified version of the R-squared that takes into account the number of predictors in the model. It’s very close to the R-squared, which indicates that the number of predictors is appropriate for the number of observations.
- F-statistic (30.37): This value tests the null hypothesis that all regression coefficients are equal to zero. The associated Prob (F-statistic) is very small (7.67e-47), indicating that we can reject the null hypothesis and conclude that at least one of the predictors is significantly related to the dependent variable.
- Coefficients:
    - const: The constant term (intercept) is not statistically significant (p-value: 0.194), suggesting that when all other variables are zero, the average effect on the dependent variable is not significantly different from zero.
    - Gini: The coefficient for the Gini coefficient is positive (9.3500) and statistically significant (p-value: 0.030), indicating that as income inequality increases, the residual sodium carbonate also increases.
    - gdp: The coefficient for GDP is negative (-5.956e-06) and highly significant (p-value: 0.000), suggesting that an increase in GDP is associated with a decrease in residual sodium carbonate.
    - Regions: The coefficients for the regions indicate that compared to the baseline region (not shown), the Southern, Northern, Eastern, and Western regions have higher levels of residual sodium carbonate, with the Northern Region having the highest increase. The North-Eastern Region is not significantly different from the baseline.
- Standard Error: The standard errors give us an estimate of the standard deviation of the coefficients. For example, the standard error for the Gini coefficient is 4.319, which is relatively high compared to the coefficient itself, indicating some uncertainty in this estimate.
- Confidence Intervals: The 95% confidence intervals provide a range within which we can be 95% confident that the true coefficient lies. For example, for the Gini coefficient, we can be 95% confident that the true coefficient is between 0.882 and 17.818.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.5043	1.158	-1.299	0.194	-3.774	0.765
Gini	9.3500	4.319	2.165	0.030	0.882	17.818
gdp	-5.956e-06	5.56e-07	-10.716	0.000	-7.05e-06	-4.87e-06
Region Southern	3.0201	0.801	3.768	0.000	1.449	4.591
Region Northern	4.2313	0.619	6.837	0.000	3.018	5.445
Region Northern-Eastern	-0.0310	0.722	-0.043	0.966	-1.447	1.385
Region Eastern	1.9232	0.685	2.809	0.005	0.581	3.266
Region Western	4.2388	0.744	5.694	0.000	2.779	5.698
Region Central	1.4208	0.693	2.051	0.040	0.063	2.779

	mean	Median	Standardeviation	25%	50%	75%
Gini	0.31	0.31	0.04	0.29	0.31	0.35
gdp	364721.87	259230.1691	322061.59	1.65e+10	2.59e+05	4.62e+05
Region Southern	0.1083	0	0.3108	0	0	0
Region Northern	0.260	0	0.43	0	0	1
Region North Eastern	0.06	0	0.25	0	0	0
Region Eastern	0.090	0	0.286	0	0	0
Region Western	0.1742	0	0.37	0	0	0
Region Central	0.2231	0	0.4164	0	0	0


Plot of Data Points
1e−5
RSC/SDP
6
5
4
3
2
1
Central Region
Eastern Region
North Eastern Region
Northern Region
Southern Region
Western Region
Region

Limitations

Missing data points

- 53% missing data points in GWQ indicator
- 19 entries missing for year-wise SDP data points. This inturn results in lot of missing data points in merged_data of GWQ indicator and SDP yearwise

Insufficient Explanatory variables

- Absence of all relevant explanatory variables result in low values of R(squared). For example, in Kuznet’s curve there can be many other indicators of environmental quality other than proportion of residula sodium carbonate dissolved in water

References

- Lecture slides and class notes
- Geek for Geeks
- https://www.researchgate.net/publication/346629758_High_residual_sodium_carbonate_water_in_the_Indian_subcontinent_concerns_challenges_and_remediation
- https://www.sciencedirect.com/science/article/abs/pii/S0378377401001032
- ChatGPT

Thank You!