Parth Rastogi
Aastha Singh

Econometrics Project

Groundwater Quality Assessment

Project's Objectives

01.

Objective- Assess the link between State Domestic Product (SDP) and groundwater quality, specifically Residual Sodium Carbonate (RSC), across Indian districts from 2000 to 2018.

02.

Data Synthesis- Combine data on RSC levels with economic indicators like SDP and Gini coefficients to construct a comprehensive dataset.

03.

Economic Influence- Examine the influence of economic growth on groundwater quality by analyzing changes in RSC.

04.

Environmental Insights- Derive insights into the environmental impact of economic development on the quality of groundwater resources.

Residual Sodium Carbonate

Our Ground water quality variable is **Residual Sodium Carbonate (RSC).**

- Importance- RSC is a key measure in assessing water quality, especially for irrigation suitability. It gauges the concentration of sodium ions relative to carbonate and bicarbonate ions It can be negative because it is difference between Carbonate and Bicarbonate level
- **Sources-** Elevated levels often result from agricultural fertilizers or industrial alkaline waste.
- **Impact on Agriculture-**
  - Soil Degradation: High RSC water leads to sodium accumulation in soil, causing dispersion and reduced permeability, which degrades soil quality.
  - Crop Productivity: Sodium hinders nutrient uptake, diminishing crop productivity and quality.
  - Soil pH Imbalance: Carbonate and bicarbonate ions can raise soil pH, exacerbating nutrient absorption issues and potentially leading to alkaline conditions.

Data Processing for Q1-3

Groundwater Quality Data (RSC) -
Observations: 5606 valid (not null)
Handling Missing Data: Not addressed due to large number (5500) of missing values; requires individual diagnostics.

State Domestic Product (SDP) Data -
Data Integration: Merged three tables based on common years.
Scaling Method: Adjusted post-2005 figures to match prior data did same for 2011 . Example for Andhra Pradesh 2004-05: Like in refference to 2000 value was 158714 and with refference to 2005 value is 121388 Formula: *Scaled Value with refference to 2000 =*

$$\frac{\textit{Original Value} \times 158714}{121388}$$

Data Selection: Focused on the first year data due to its comprehensive coverage (covers 3/4 of the year).
Data Merging-
Variables Merged: RSC and SDP data on state and year basis.
Data Loss: Exclusion of some union territories resulted in 394 lost observations.
Methodology: Employed an inner join to eliminate records with null values, ensuring dataset completeness.

Q1-3
Regression Analysis

$$GWQ_{i,t} = \beta_0 + \beta_1 SDP_{i,t} + u_{i,t}$$

Q1. Regression Outcome

Variable: State Domestic Product (SDP)
Purpose: To determine if wealthier states have higher or lower levels of Residual Sodium Carbonate (RSC).
Number of observations= 5212
R-squared: 0.04503
SDP Coefficient: Negative

- As SDP increases, RSC decreases on average, suggesting that higher economic output is associated with better water quality.

T-value: -15.68 ,F -value : 245.8
Statistically significant; strong evidence against the null hypothesis (B_hat = 0), confirming the impact of SDP on RSC.
Safety Thresholds:

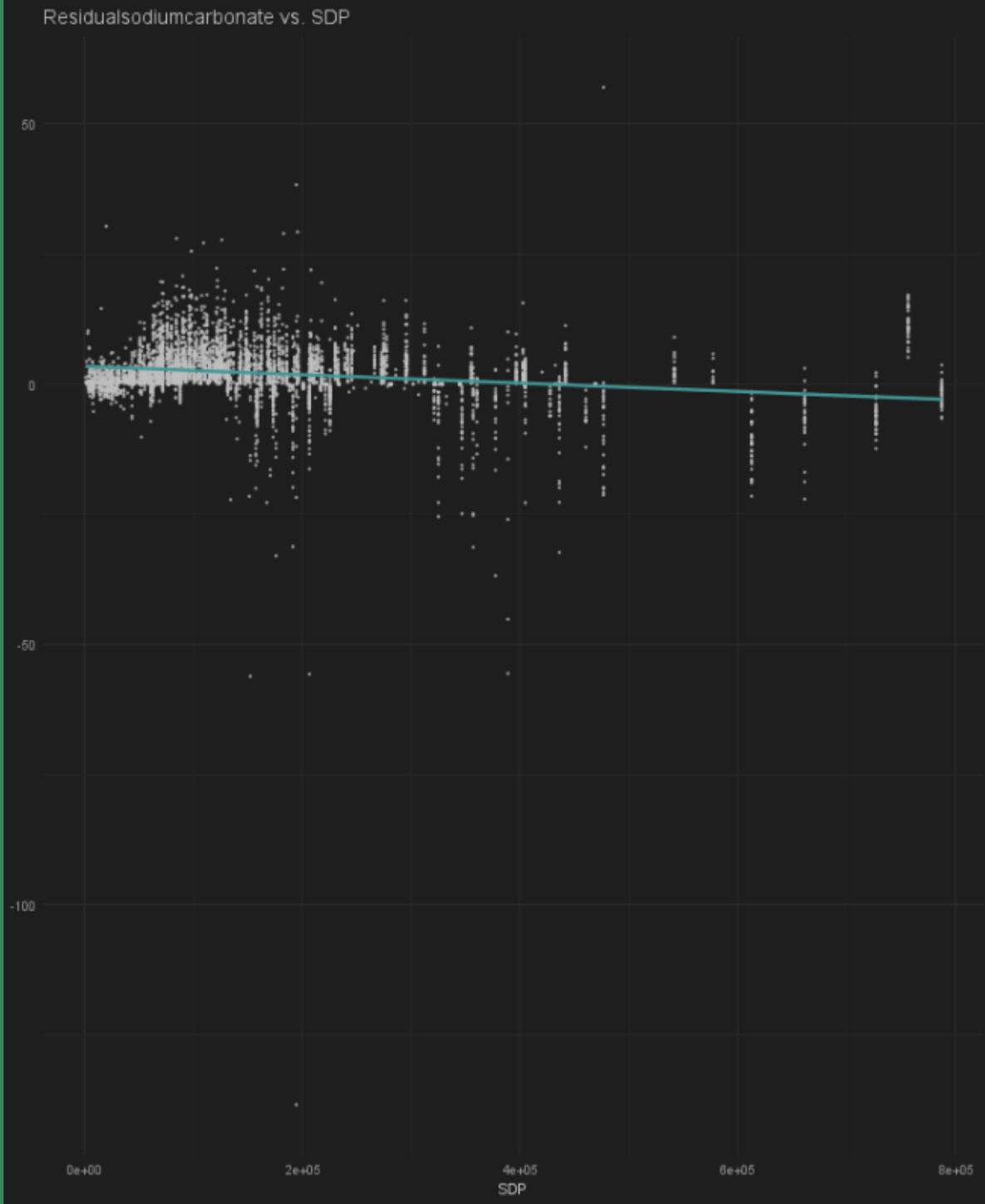
- Safe Limit: RSC < 1.25 mm / L
- Hazardous: RSC > 2.5 mm / L

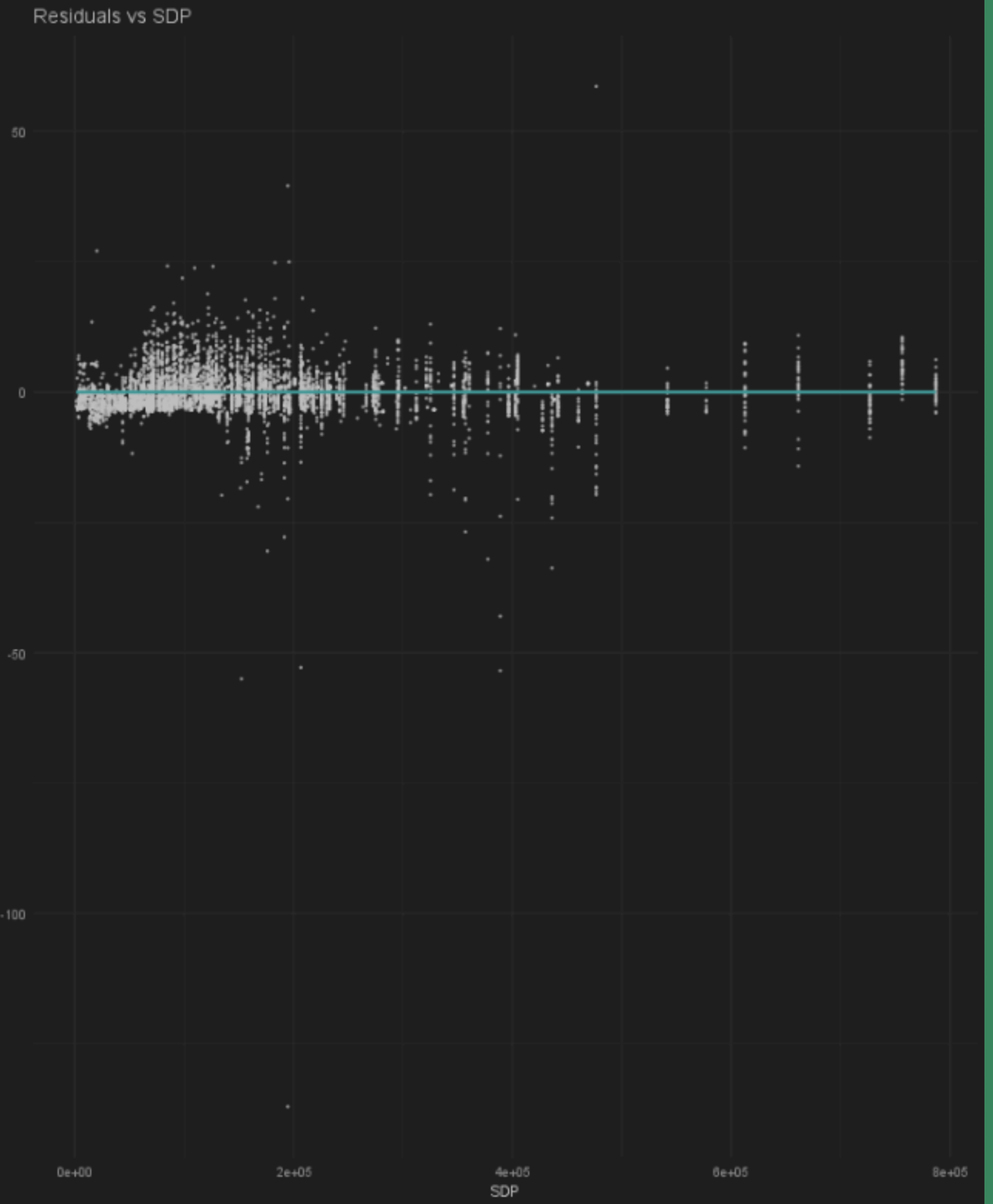
Interpretation: On average, states with SDP < 110,000 tend to have RSC levels in the hazardous range.

VARIABLE	COEFFICIENT (STD ERROR)
Intercept	3.393 (0.1192)
SDP	-8.056e-06 (5.139e-07)

RSC vs SDP Plot

Here we have Residual Sodium Carbonate(on Y) vs SDP(on X) plot and cyan line in middle is the regression line.
Inverse Relationship: As SDP increases, RSC typically decreases, indicating that higher economic output is associated with better water quality.
Outliers: Numerous outliers, particularly at lower SDP levels, suggest variance in factors affecting RSC that are not captured solely by economic output.
Data Distribution: Greater data density at lower SDP levels indicates more
datapoints for lower SDP rather than high SDP.





Residuals vs SDP Plot

Visualization: Displays residuals from RSC regression on SDP.
Regression Line: Cyan line aligns with the X-axis.
0 Slope: Indicates no effect of SDP on the residuals.
Covariance Check: Alignment with X-axis confirms ***Cov(*** SDP, u_hat ***)=0***, validating the regression model's assumptions.

u_hat Histogram

Residual Sum-

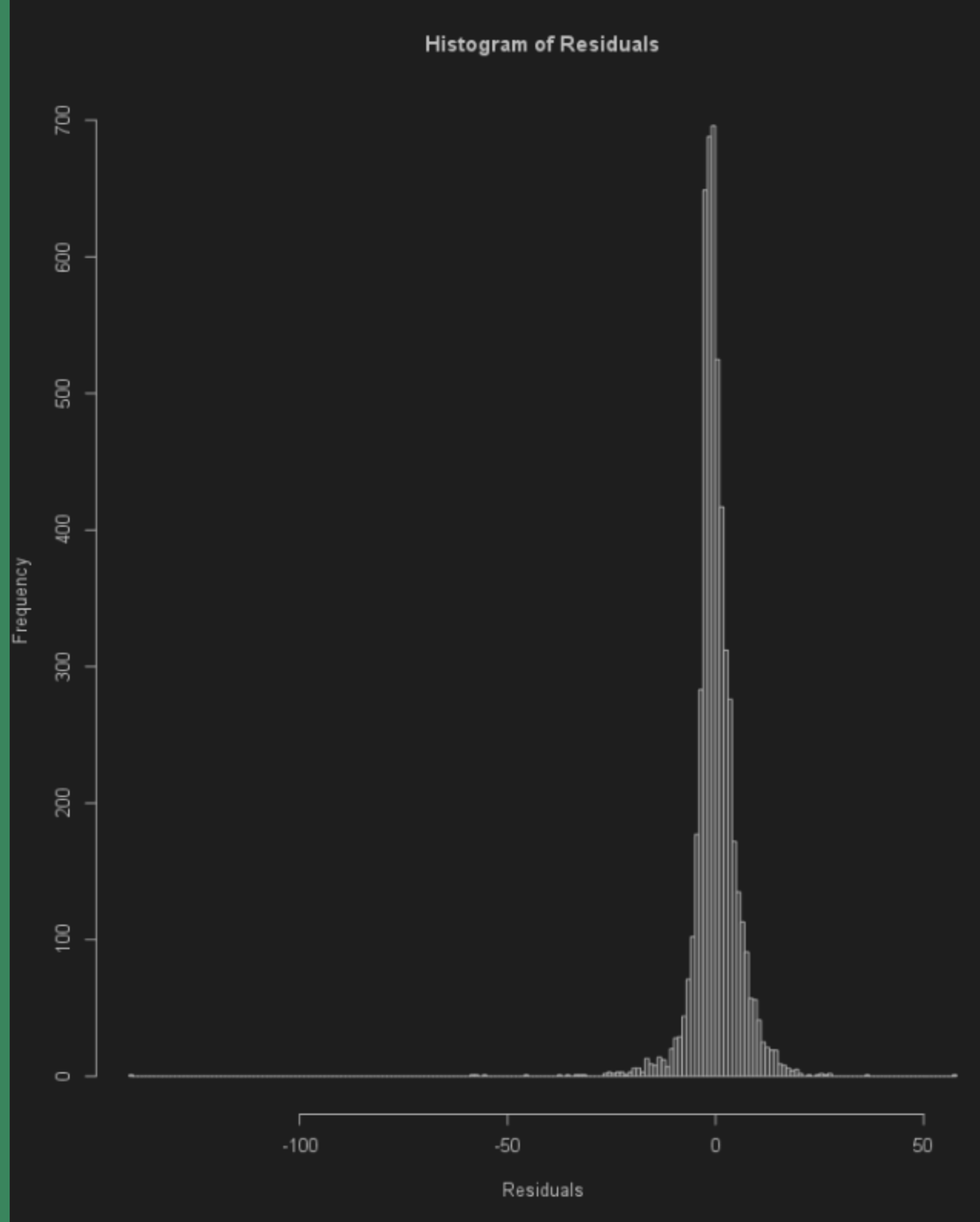
- Value: Evaluated at approximately *9.8605e-12*, which is effectively zero, indicating a small rounding error.

Histogram Overview-

- Axes Information: X-axis represents the residuals; Y-axis shows frequency.
- Distribution Characteristics: The histogram reveals a distribution that is nearly normal, with a mean close to zero and symmetric about the mean.

Key Takeaway-

- Model Fit: The negligible sum of residuals and the normal distribution confirm a good fit of the model, with unbiased error terms.



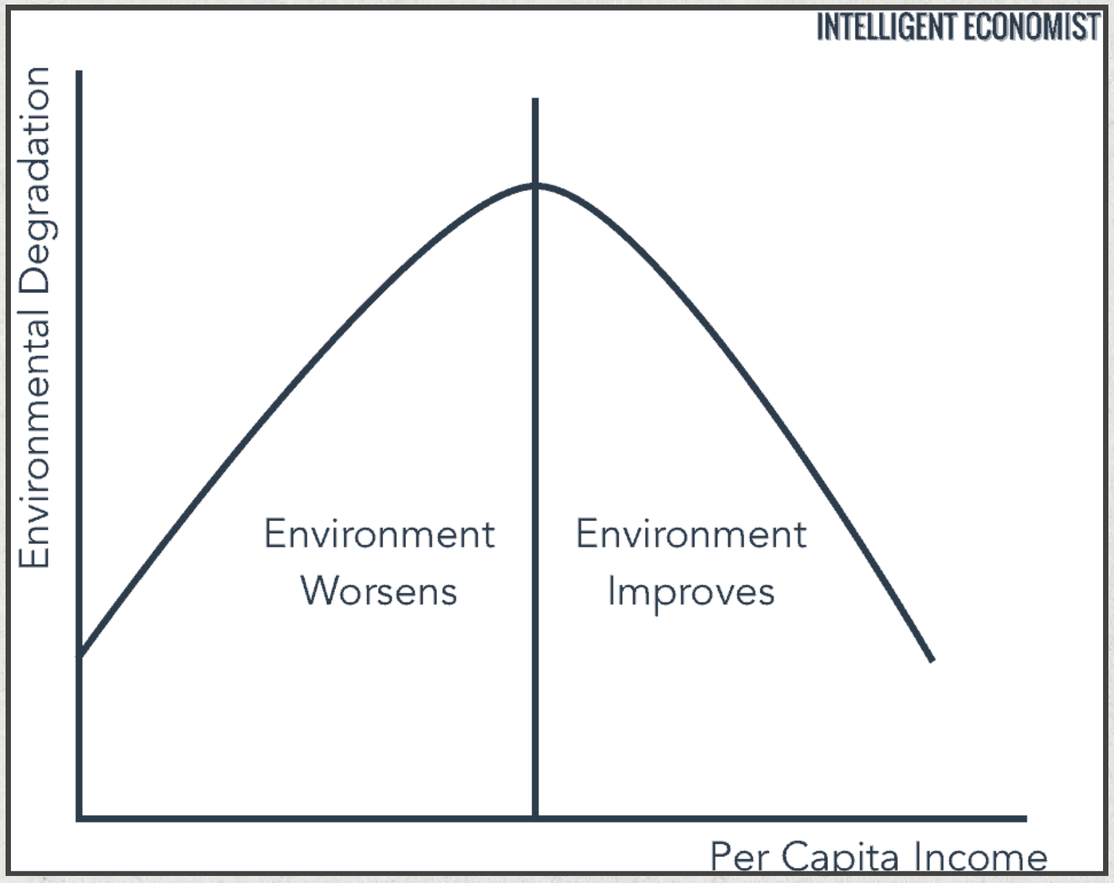
Data Processing for Q4-7

During the process of merging our previously consolidated dataset with district-level Gini values, we encountered challenges due to inconsistencies in district naming conventions.
For example, our dataset lists 'Prayagraj,' whereas the Gini value dataset refers to the same district as 'Allahabad.' This discrepancy led to the loss of some observations.
However, it's important to note that these missing observations are considered **Missing Completely At Random (MCAR)**, indicating that their absence is unlikely to introduce bias into the analysis.
Another note is that we are not taking data points of missing observations.

Q4 -7
Regression Analysis

$$GWQ_{i,t} = \beta_0 + \beta_1 SDP_{i,t} + \beta_2 SDP\text{^}2_{i,t} + \beta_3 SDP\text{^}3_{i,t} + \beta_4 GINI + u_{i,t}$$

Kuznet's Curve



01. **Relationship Basis:** The Environmental Kuznets Curve theorizes a link between economic development and environmental quality.

02. **Initial Impact:** As economies grow and industrialize, environmental degradation tends to worsen up to a certain point of average income.

03. **Reversal Point:** Beyond this income threshold, the trend reverses, and environmental conditions start to improve as the economy continues to develop.

Regression Outcome

Explanatory Variables- The model includes State Domestic Product (SDP), its squared term (SDP^2), its cubic term (SDP^3), and the Gini index value.
Observations: The analysis was conducted using a total of 4593 observations.
Model Fit (R-squared)- The R-squared value is 0.086, nearly double that of the model with just SDP. This enhancement indicates that the inclusion of SDP's higher powers and the Gini index significantly improves the model's ability to explain variations in Residual Sodium Carbonate.
Significance (F-value)- The model achieved an F-value of 107.8 with a p-value less than $2 \times 10^{(-16)}$. This exceptionally low p-value allows us to confidently reject the null hypothesis that all coefficients of the explanatory variables are zero, confirming their significant impact on the model.

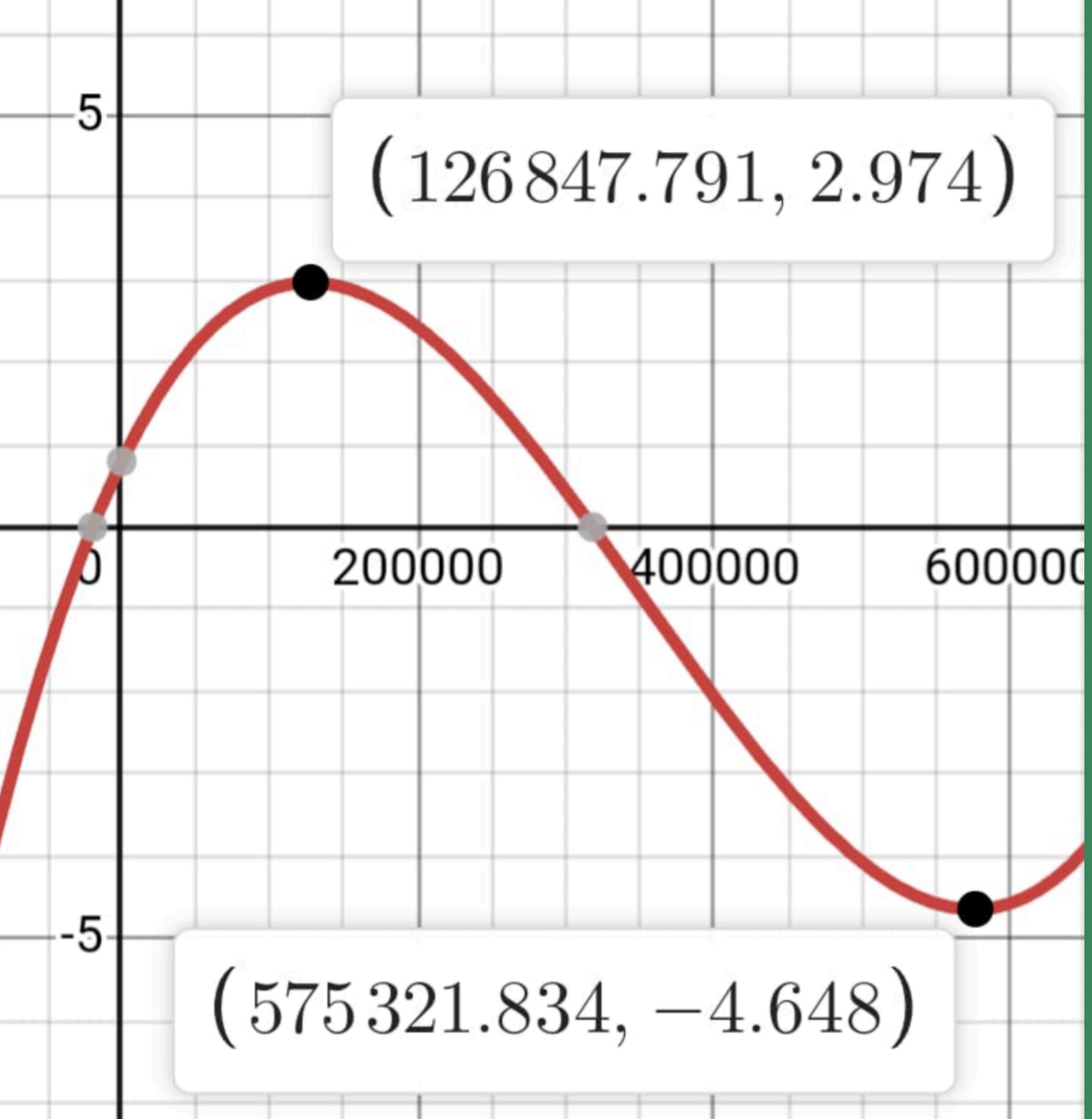
VARIABLE	COEFFICIENT (STD ERROR)
Intercept	8.621e-01 (4.518e-01)
SDP	3.784e-05 (3.479e-06)
SDP^2	-1.786e-10 (1.235e-11)
SDP^3	1.696e-16 (1.146e-17)
Gini Value	-3.643e-02 (1.440e+00)

Variable Statistics

	mean	median	25th percentile	75th percentile	95th percentile
SDP	175552	129132	72088	217963	465810
SDP^2	5.390e+10	1.667e+10	5.197e+09	4.751e+10	2.169e+11
SDP^3	2.388e+16	2.153e+15	3.746e+14	1.035e+16	1.01096e+17
Ginivalue	0.2743	0.2700	0.2400	0.3100	0.3700
Residual Sodium Carbonate	1.93	1.58	0	4.58	10.36747
Year	2010	2010	2005	2014	2018

Outliers Detection and Handling

In refining our regression model, we employed the studentize statistic with a threshold set at +-3 to identify outliers. This process revealed a total of 59 outliers within our dataset. Upon excluding these outliers and conducting a subsequent regression analysis, we observed that the R-squared value increased slightly to 0.088, compared to the previous model that included outliers. This suggests that the outliers contribute minorly to the explanatory power of our model so removing them would just lead to loss of data so we decided to keep them as we already have loss of observation and we are performing more diagnostics and modifying model in future. Given the complexity introduced by having multiple explanatory variables, we decided against winsorizing the outliers. Therefore, we retained the outliers in our analysis to maintain the integrity and robustness of our model.
We also detected the Influential observation by calculating df beta and setting threshold at 2 / root n where we got 32 influential observation and we did regression after removing them which gave a R square of 0.081 which means those observation are positively influencing our model. Hence, no need of their removal.



Environmental Kuznets Curve & Pollution

Environmental Kuznets Curve (EKC)- A representation of the relationship between economic prosperity and environmental health.

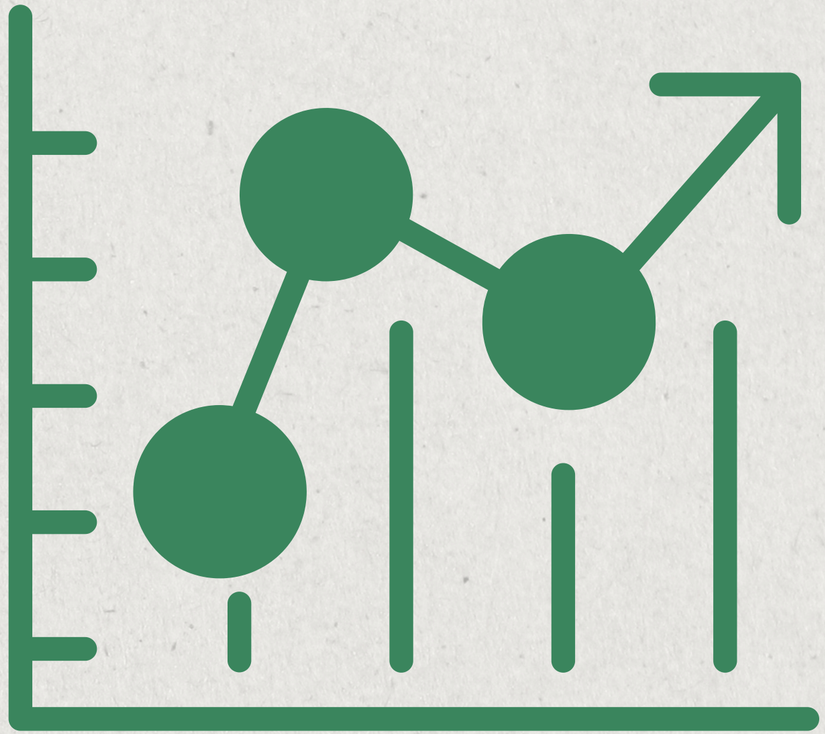
Residual Sodium Carbonate (RSC)- Chosen as the pollution indicator, where higher levels indicate poorer soil quality.

Positive GDP Coefficient- Suggests initial pollution increase with economic growth.

Negative GDP Squared Coefficient- Indicates eventual pollution decrease, supporting the EKC hypothesis.

The EKC illustrated with RSC data shows an inverted parabolic trend, turning upwards after an SDP of 575k. However, this affects a small portion (1%) of observations.

Economic Growth, Inequality, and Pollution



Inequality's Role-

- Gini Index Impact: Analysis reveals that greater economic inequality correlates with reduced pollution levels Which is counter intuitive, like more inequality are thought to have more industrialists hence more industries leading to more pollution but it’s going opposite.

Conclusive Insights-

- Majority Trend: Since 99% of the data points fall below the 575k SDP threshold, the general trend supports the EKC hypothesis for Indian districts during 2000-2019.
- Policy Implications: These findings highlight the complex interplay between economic growth, policy measures, industrial changes, and environmental outcomes.
- SDP relation was expected because it was almost like kuznet’s curve.

Yearly Impact on Groundwater Quality

The analysis investigates year-over-year changes in Residual Sodium Carbonate (RSC) levels, controlling for State Domestic Product (SDP) and Gini coefficient, with the year 2000 as the reference.
Observations reveal a downward trend in RSC levels over the years, supporting the hypothesis that environmental quality improves with economic development over time.

The R-squared value is **0.2572**, indicating a good predictive power of the model for RSC levels. While the F-statistic has decreased to **71.92**, the model remains statistically significant due to the very low p-value.

Analytical Insights and Model Considerations

01.

The decline in RSC levels suggests an improvement in groundwater quality, consistent with increased SDP and the Environmental Kuznets Curve principle.

02.

However, certain years show less impact on RSC, as indicated by higher p-values, like 2001 ($p = 0.075$), 2002 ($p = 0.17$), and 2005 ($p = 0.293$).

03.

The possibility of multicollinearity is noted, as SDP may be correlated with time, potentially affecting the model's coefficients.

04.

Despite some years showing less significance, the overall model supports the notion that economic growth over time contributes to better environmental outcomes.

Regression Analysis in Regional Dummies

We conducted a regression analysis to assess the influence of economic development on groundwater quality, using Residual Sodium Carbonate as an indicator.
Num of Observation = 4593
The model includes regional dummy variables representing the North East, East, South, Center, and West of India, with the North as the reference category. Incorporating these regional dummies improved our model significantly, increasing the Multiple R-squared from 0.086 to 0.12, which indicates a better fit.

F statistics = 69.8 which shows that the model is significant in explaining the RSC level in India.

Which in turn show that geography has a role to play in the RSC level.

	COEFFICIENT (STD ERROR)
Intercept	3.776 (0.5380)
SDP	2.808e-05 (3.734e-06)
SDP^2	-1.541e-10 (1.269e-11)
SDP^3	1.492e-16 (1.160e-17)
Gini Value	-3.097 (1.515)
North East	-3.482 (0.3683)
East	-2.187 (0.3248)
South	-1.102 (0.2740)
Center	-2.277 (0.2277)
West	-0.02638 (0.3110)

Regional Comparisons and Insights

01. Keeping the State Domestic Product and Gini coefficient constants, we found the RSC levels in the West are on par with the North.

02. However, when compared to the North, the NorthEast, East, Central, and South regions exhibit lower RSC levels by 3.482, 2.187, 2.277, and 1.102 units respectively.

03. These findings suggest that north and west have more RSC for same SDP which can be case because of more industrialization or more lime content in soil of the given region .

Limitations

Our dataset experienced attrition, notably with significant loss of observations due to the exclusion of Union Territories and inconsistencies in district names across different data sources.

The absence of data for approximately 5,500 observations suggests a potential underrepresentation of variables influencing RSC levels.

The analysis does not encompass certain geographical and industrial factors such as soil lime content, overall soil quality, and factory prevalence, which could provide additional explanatory power.

Furthermore, the model does not incorporate interaction terms or logarithmic transformations, which might yield better interpretations of the economic-environmental relationship.

Presented by Group 11

Thank you very much!