Final Regression Project

Team C1 - Gabriella Armada, Aditi Attavar, Naomi Baron, Ruida Liu, Yifan Pan, Haonan Pu, Yuqi Zhang

I. Introduction

Our analysis investigates the relationship between environmental impact, measured as CO2 emissions, and economic prosperity, measured as GDP Per Capita. Specifically, we aimed to test the hypothesis that higher carbon emissions are a primary driver of economic development.

By utilizing a multivariable regression model, we sought to determine if the historical link between carbon-intensive industrialization and wealth remains statistically significant in the modern era. In our model we aimed to predict GDP Per Capita by including CO2 Emissions Per Capita, Birth Rate, Share of Renewable Energy Consumption, and Share of Population Living in Extreme Poverty as dependent variables.

The results provided strong evidence supporting the alternative hypothesis, rejecting the null that there is no relationship between these variables. Our model explains approximately 58% of the global variation in economic development and identifies statistically significant positive correlation between emissions (CO2) and wealth (GDP). Even when controlling for variables such as renewable energy adoption and poverty rates, the model predicts that a single metric ton increase in per capita CO2 emissions is associated with an increase of approximately $14,922 in GDP per capita.

The results confirm that expanded manufacturing and energy use drive higher GDP, and likely living standards. Although renewables positively impact GDP, higher carbon output remains the primary indicator of GDP per capita.

- **Data Sources**

  We downloaded all of our datasets from https://ourworldindata.org/. By joining on *Country* and *Year*, we produced a final dataset of 1,308 rows containing 70 countries from 1990-2023.

  - CO2 Emissions Per Capita
  - Birth Rate
  - GDP Per Capita
  - Share of Renewable Energy Consumption
  - Share of Population Living in Extreme Poverty

- **Data Dictionary**
  - **Country**: contains the full country name where each data point originates from.
  - **Code**: contains a 3-letter abbreviation for country.
  - **Year**: contains the year (format: YYYY) where each data point originates from.
  - **Annual CO2 Emissions Per Capita**: Carbon dioxide (CO_2) emissions (measured in tonnes per capita) from burning fossil fuels and industrial processes. This includes emissions from transport, electricity generation, and heating, but not land-use change.
  - **Birth Rate**: Total number of births per 1,000 people in a given country, year.
  - **GDP Per Capita**: Gross Domestic Product divided by country population. This data is adjusted for inflation and differences in living costs between countries.
  - **Percent Renewable Energy**: Share of primary energy consumption from renewable resources. Renewables include hydropower, solar, wind, geothermal, bioenergy, wave, and tidal, but not traditional biofuels.
  - **Share Extreme Poverty**: Share of population living in extreme poverty, defined as living below the International Poverty Line of $3 per day. This data is adjusted for inflation and for differences in living costs between countries.

- **Descriptive Statistics**

	Annual CO2 Emissions Per Capita	Birth Rate	GDP Per Capita	Percent Renewable Energy	Share Extreme Poverty
Mean	7.48	13.44	37497.77	14.23	4.90
Median	6.69	11.35	35568.85	8.47	0.64
SD	4.96	5.40	24211.24	15.36	10.95
Min	0.13	5.00	1666.93	0.00	0.00
Max	39.03	42.89	138677.97	83.70	83.03

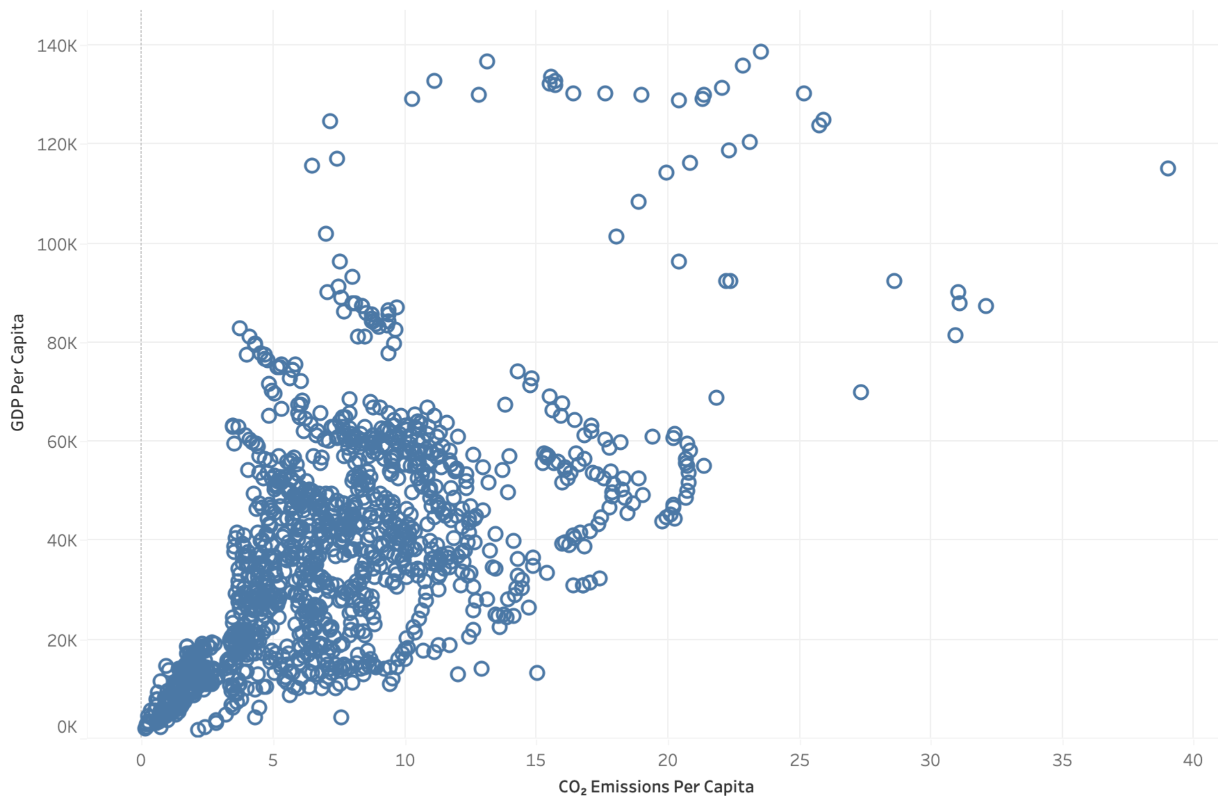
- **Correlation Matrix**

Variable	Year	CO2 Per Capita	Birth Rate	GDP Per Capita	Renewable %	Extreme Poverty
Year	1	-0.149	-0.309	0.197	0.181	-0.315
CO2 Per Capita	-0.149	1	-0.319	0.619	-0.183	-0.408
Birth Rate	-0.309	-0.319	1	-0.471	-0.053	0.617
GDP Per Capita	0.197	0.619	-0.471	1	0.207	-0.478
Renewable %	0.181	-0.183	-0.053	0.207	1	-0.035
Extreme Poverty	-0.315	-0.408	0.617	-0.478	-0.035	1

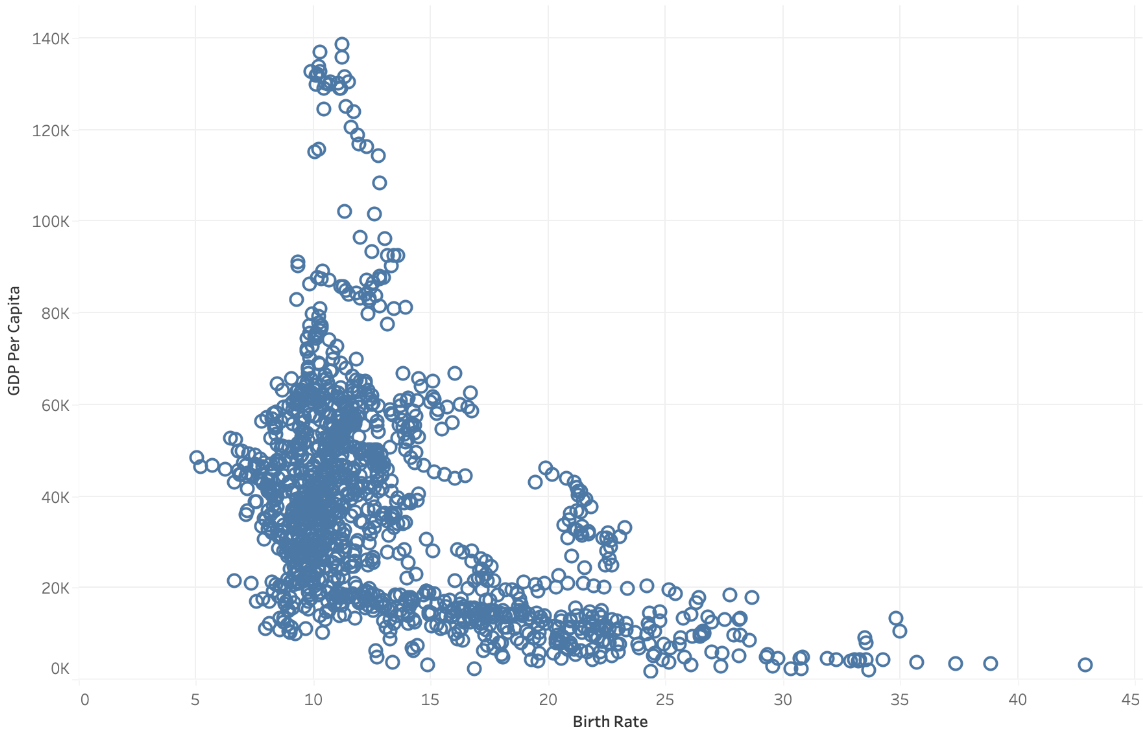
- **Scatterplots**

Main Independent Variable

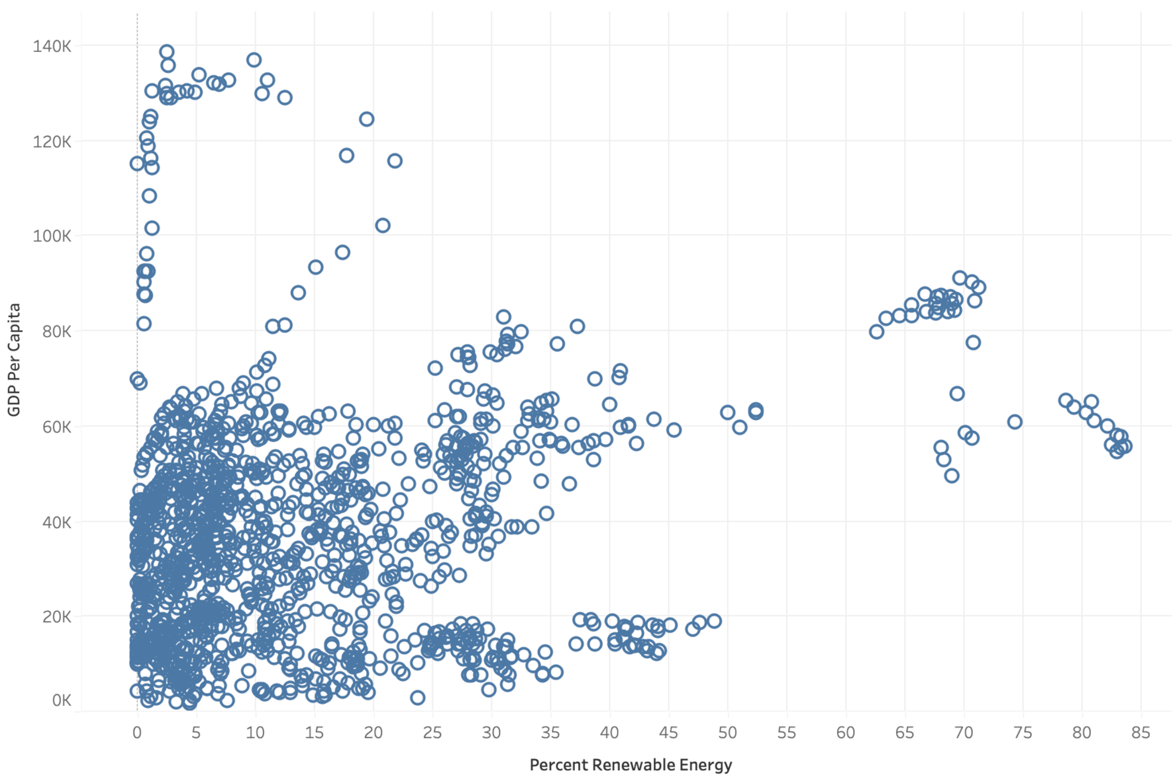
CO_2 Emissions Per Capita vs. GDP Per Capita



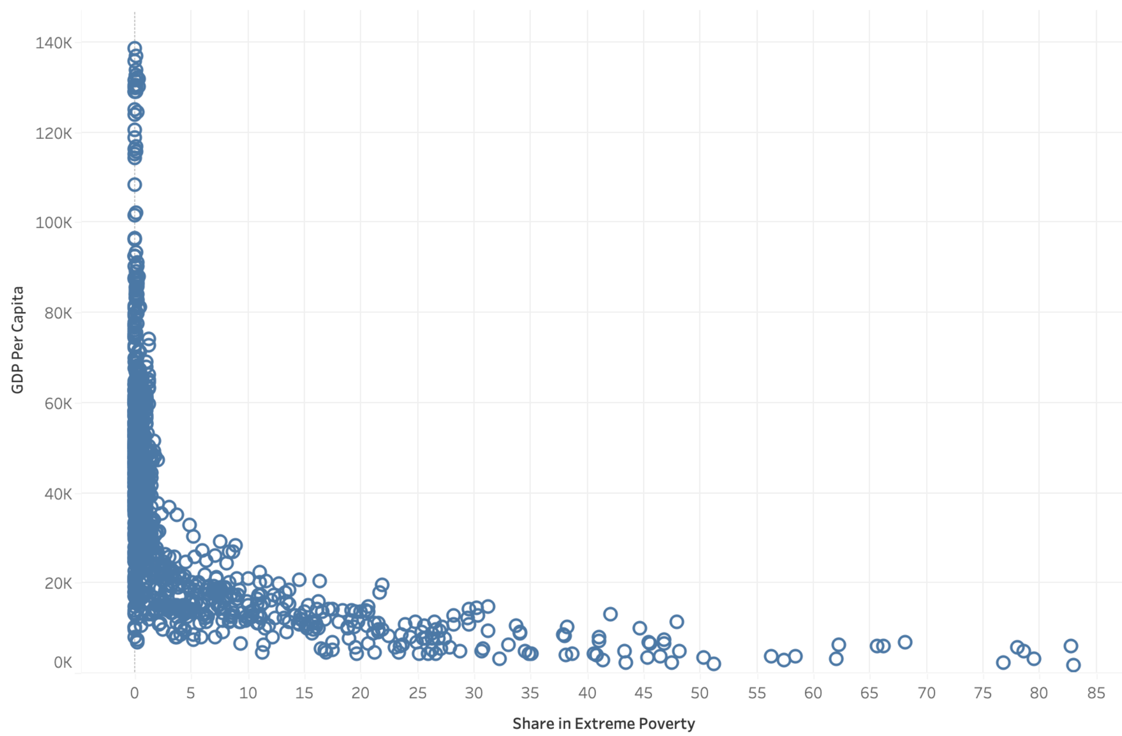
Birth Rate vs. GDP Per Capita



Percent Renewable Energy vs. GDP Per Capita



Share in Extreme Poverty vs. GDP Per Capita



Proposed Regression Objective: Our objective is to quantitatively test whether higher levels of carbon emissions are associated with greater economic prosperity. Further, we are interested in exploring if this relationship remains strong in the modern era of climate awareness.

II. Model Building

Initial regression model with all variables.

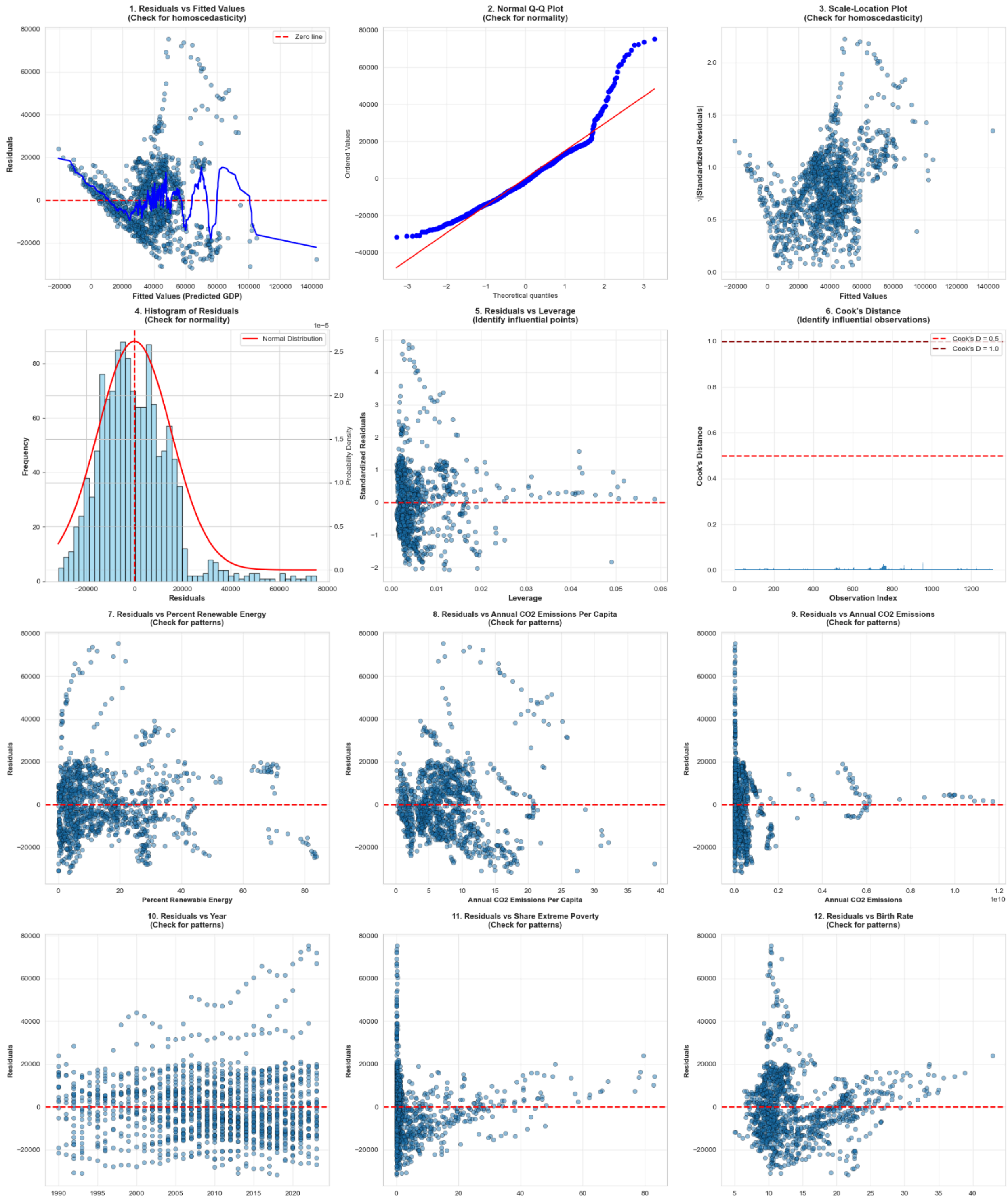
```
================================================================================
COEFFICIENT ANALYSIS
================================================================================

                        Variable  Coefficient  Std Error  t-statistic  p-value  95% CI Lower  95% CI Upper Significant
                       Intercept 37497.772106 422.018726    88.853337 0.000000  36670.615404  38324.928808         YES
                            Year  4483.553632 485.495119     9.235013 0.000000   3531.983200   5435.124065         YES
             Annual CO2 Emissions -3947.074609 367.002512   -10.754898 0.000000  -4666.399534  -3227.749685         YES
Annual CO2 Emissions Per Capita  8102.216547        NaN          NaN      NaN           NaN           NaN          NO
                      Birth Rate -4335.020196 175.594518   -24.687674 0.000000  -4679.185452  -3990.854939         YES
         Percent Renewable Energy  6398.796544 431.111881    14.842543 0.000000   5553.817258   7243.775831         YES
            Share Extreme Poverty  -553.002709 393.389402    -1.405739 0.160041  -1324.045937    218.040519          NO
Annual CO2 Emissions Per Capita  8102.216547        NaN          NaN      NaN           NaN           NaN          NO
```

```
================================================================================
MODEL PERFORMANCE METRICS
================================================================================

R-squared: 0.604720
Adjusted R-squared: 0.602591
Standard Error: 15262.848509

F-statistic: 469.829539
F-statistic p-value: 1.110223e-16

Model Significance: YES - Model is statistically significant
```

```
================================================================================
RESIDUAL DIAGNOSTICS SUMMARY
================================================================================

1. NORMALITY TESTS:
   Shapiro-Wilk Test:
     • Statistic: 0.933594
     • p-value: 1.124156e-23
     • Result: ✗ Not Normal

   Jarque-Bera Test:
     • Statistic: 913.707370
     • p-value: 3.899111e-199
     • Result: ✗ Not Normal

2. HOMOSCEDASTICITY:
   Correlation test (fitted vs |residuals|):
     • Correlation: 0.433537
     • p-value: 4.587927e-61
     • Result: ✗ Heteroscedastic

3. INFLUENTIAL POINTS:
   Cook's Distance > 0.5: 0 observations
   Cook's Distance > 1.0: 0 observations

4. OUTLIERS:
   |Standardized Residuals| > 3: 23 observations (1.76%)
```

```
5. RESIDUAL STATISTICS:
   Mean: 0.000000 (should be ≈ 0)
   Std Dev: 15221.921433
   Min: -31640.35
   Max: 75335.96
   Range: 106976.31
```

Initial Residual Plots



- **Data Analysis Steps**

Phase 1: Problem Statement	We began our analysis by defining our problem statement. With the rise in renewable energy and a shift towards sustainability, we wanted to explore the relationship between carbon emissions and economic development. We defined our dependent variable as economic development (measured in GDP Per Capita) and our main independent variable as carbon emissions (measured in metric tonnes per capita).
Phase 2: Data Collection	After defining our problem statement, we began researching potential factors of economic growth. We identified our other independent variables as year, birth rate, percent renewable energy, and share in extreme poverty. We downloaded all of our datasets from https://ourworldindata.org/.
Phase 3: Data Cleaning	The data cleaning phase involved joining 6 datasets by country and year. We produced a final dataset of 1,308 rows containing 70 countries from 1990-2023. Datasets: - Annual CO_2 Emissions - Annual CO_2 Emissions Per capita - GDP Per Capita - Birth Rate - Percent Renewable Energy - Share in Extreme Poverty
Phase 4: Initial Model Development	We built our initial multilinear regression model based on: - **Dependent variable**: GDP Per Capita - **Main Independent Variable**: Annual CO_2 Emissions - **Other Independent Variable**s: Annual CO_2 Emissions Per capita, GDP Per Capita, Birth Rate, Percent Renewable Energy, and Share in Extreme Poverty
Phase 5: Model Refinement	While refining our initial model, we decided to remove *Annual CO_2 Emissions* and keep *Annual CO_2 Emissions Per Capita* to avoid multicollinearity. Since Annual CO_2

	Emissions Per Capita is calculated by dividing Annual CO_2 Emissions by population, including both variables in our model adds redundant information, making it difficult to interpret a variable's true impact.
Phase 6: Interpretation	We interpreted our null and alternative hypothesis based on our final model. (See Part III: Final Model Analysis, Interpretation Section)

III. Final Model Analysis

Final fitted regression model.

```
================================================================================
COEFFICIENT ANALYSIS
================================================================================

                    Variable  Coefficient  Std Error  t-statistic      p-value  95% CI Lower  95% CI Upper Significant
                   Intercept 37497.772106 434.304736    86.339773 0.000000e+00  36646.534824  38349.009388         YES
                        Year  4040.432642 495.379899     8.156231 8.881784e-16   3069.488040   5011.377243         YES
Annual CO2 Emissions Per Capita 14921.544575 515.955451    28.920219 0.000000e+00  13910.271892  15932.817258         YES
                  Birth Rate -4160.605085 565.246931    -7.360686 3.230749e-13  -5268.489071  -3052.721099         YES
       Percent Renewable Energy  6723.601139 448.591905    14.988236 0.000000e+00   5844.361005   7602.841272         YES
          Share Extreme Poverty -1404.458451 596.884908    -2.352980 1.877172e-02  -2574.352872   -234.564031         YES
```

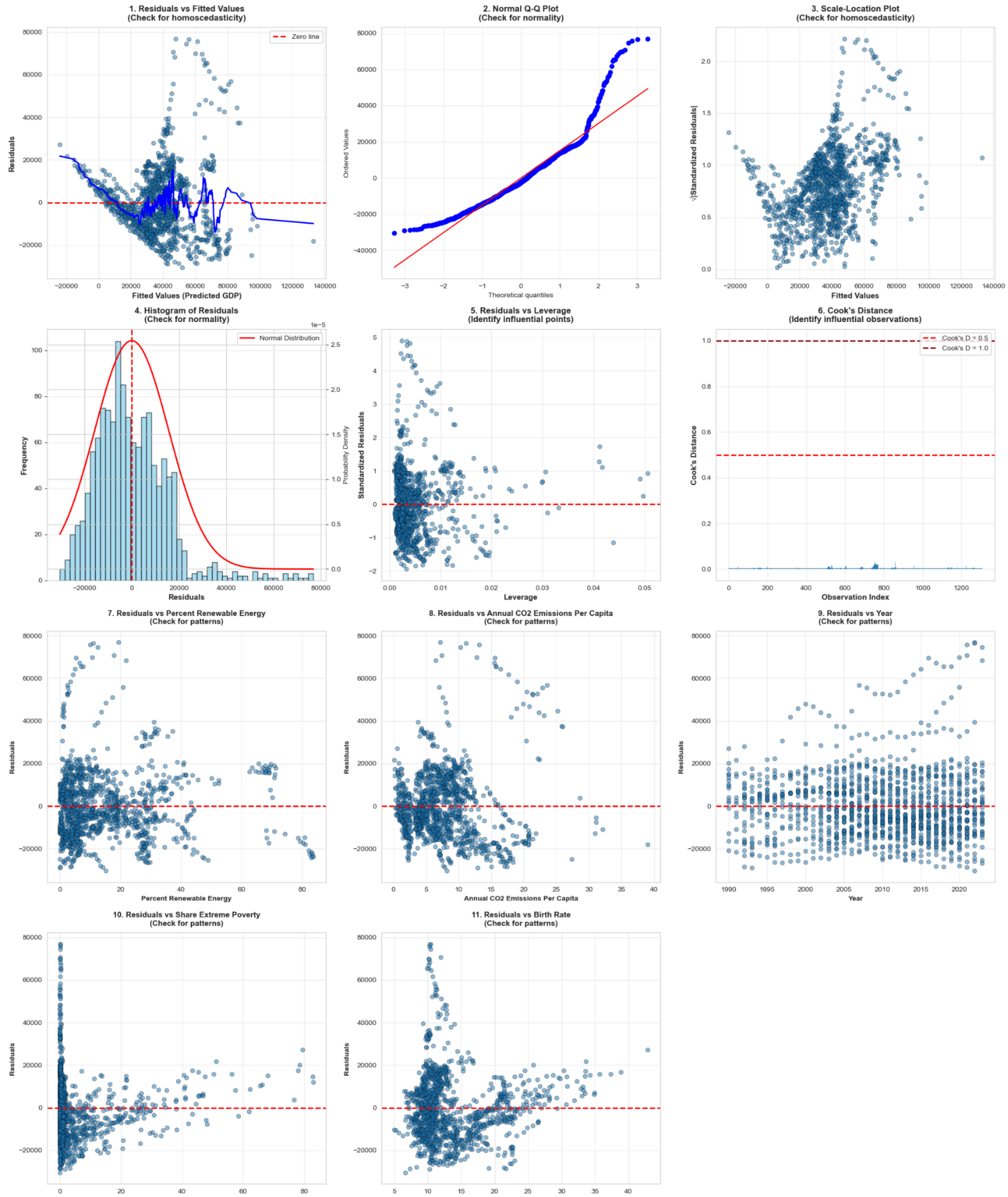

```
================================================================================
COEFFICIENT ANALYSIS
================================================================================

                    Variable  Coefficient  Std Error  t-statistic      p-value  95% CI Lower  95% CI Upper Significant
                   Intercept 37497.772106 434.304736    86.339773 0.000000e+00  36646.534824  38349.009388         YES
                        Year  4040.432642 495.379899     8.156231 8.881784e-16   3069.488040   5011.377243         YES
Annual CO2 Emissions Per Capita 14921.544575 515.955451    28.920219 0.000000e+00  13910.271892  15932.817258         YES
                  Birth Rate -4160.605085 565.246931    -7.360686 3.230749e-13  -5268.489071  -3052.721099         YES
       Percent Renewable Energy  6723.601139 448.591905    14.988236 0.000000e+00   5844.361005   7602.841272         YES
          Share Extreme Poverty -1404.458451 596.884908    -2.352980 1.877172e-02  -2574.352872   -234.564031         YES
```


```
================================================================================
MODEL PERFORMANCE METRICS
================================================================================

R-squared: 0.580726
Adjusted R-squared: 0.579116
Standard Error: 15707.187827

F-statistic: 621.073025
F-statistic p-value: 1.110223e-16

Model Significance: YES - Model is statistically significant
```

```
================================================================================
RESIDUAL DIAGNOSTICS SUMMARY
================================================================================

1. NORMALITY TESTS:
   Shapiro-Wilk Test:
     • Statistic: 0.919550
     • p-value: 7.999495e-26
     • Result: ✗ Not Normal

   Jarque-Bera Test:
     • Statistic: 1138.230521
     • p-value: 6.860926e-248
     • Result: ✗ Not Normal

2. HOMOSCEDASTICITY:
   Correlation test (fitted vs |residuals|):
     • Correlation: 0.382586
     • p-value: 7.545922e-47
     • Result: ✗ Heteroscedastic

3. INFLUENTIAL POINTS:
   Cook's Distance > 0.5: 0 observations
   Cook's Distance > 1.0: 0 observations

4. OUTLIERS:
   |Standardized Residuals| > 3: 24 observations (1.83%)
```

```
5. RESIDUAL STATISTICS:
   Mean: 0.000000 (should be ≈ 0)
   Std Dev: 15677.114685
   Min: -30387.77
   Max: 76837.56
   Range: 107225.33
```

Final Residual Plots



Interpreting Main Input Variable Coefficient: Annual CO2 Emissions Per Capita

Coefficient: 14,921.54

Interpretation:
For every 1 unit increase in Annual CO2 Emissions Per Capita, GDP Per Capita increases by ~ $14,921.54, holding all other variables (Year, Birth Rate, Renewable Energy, and Extreme Poverty) constant

P-Value of Main Input Variable (Annual CO2 Emissions Per Capita) is significantly less than alpha (0.05). This Indicates that the relationship is statistically significant, and we can reject the null hypothesis that there is no relationship between these variables

95% Confidence Interval of Main Input Variable (Annual CO2 Emissions Per Capita): (13910.27, 15932.81). This interval does not include 0, meaning we can reject our Null Hypothesis and we are 95% confident that the true population parameter for the effect of CO2 on GDP is positive and falls within this range.

Additionally, the Positive T-statistic (28.92) and positive coefficient (14921.54) aligns with our hypothesis that the correlation is positive, meaning that as emissions go up, GDP goes up.

We expected our relatively low R^2 of 0.58. In physical sciences, this R^2 is considered low or 'moderate', but for our context of complex macroeconomic modeling, it is considered good. We expected a lower R^2 since it is extremely difficult to be able to identify all the variables that affect something as complex as GDP. In our case, we were able to identify all of the core drivers of GDP but our model has ~42% of variation that still needs to be explained.

Our model output has a huge economic significance. We found that developing nations have a GDP per capita of ~7000. So an increase of almost ~$15,000 (main input variable coefficient) is transformative and could mean the difference between a low-income and high-income economy. An increase of 1 metric ton of CO2 per person is associated with an increase of roughly $14,921 in GDP per capita.

Relevant Example: Predicting an observation from our data set

```
================================================================================
EXAMPLE 1: Predicting for Row 1
================================================================================

Country: Algeria
Year: 1995

Input Features:
  Year: 1995
  Annual CO2 Emissions Per Capita: 3.3686354
  Birth Rate: 24.56
  Percent Renewable Energy: 0.1712212
  Share Extreme Poverty: 11.807

==================================================
PREDICTED GDP Per Capita: $2,751.31
ACTUAL GDP Per Capita:    $10,588.44
==================================================
Prediction Error: $7,837.13
Percentage Error: 74.02%
```

```
================================================================================
EXAMPLE 2: Predicting for Row 10
================================================================================

Country: Australia
Year: 2016

Input Features:
  Year: 2016
  Annual CO2 Emissions Per Capita: 16.837934
  Birth Rate: 12.727
  Percent Renewable Energy: 6.7850986
  Share Extreme Poverty: 0.49874178

==================================================
PREDICTED GDP Per Capita: $66,584.95
ACTUAL GDP Per Capita:    $56,341.52
==================================================
Prediction Error: $10,243.43
Percentage Error: 18.18%
```

```
================================================================================
EXAMPLE 3: Custom Hypothetical Scenario
================================================================================

Hypothetical Scenario:
  Year: 2025
  Annual CO2 Emissions Per Capita: 10.0
  Birth Rate: 12.0
  Percent Renewable Energy: 25.0
  Share Extreme Poverty: 0.5

==================================================
PREDICTED GDP Per Capita: $58,739.33
==================================================
```

Post-regression recommendations

Given the complexity of economic development, our current set of variables likely does not fully capture the drivers of GDP per capita. To improve this model, we would suggest incorporating additional variables to improve the model's explanatory power, while ensuring they do not decrease the Adjusted R^2. Furthermore,we would recommend exploring the effect of transforming our GDP per capita variable into the natural log of GDP. Because economic growth is exponential and not linear, this transformation might be able to better reflect the reality that the marginal increase in wealth has a larger impact on developing nations. By doing this, the model might improve the model validity and allow coefficients to be interpreted more intuitively as percentage growth.

Recap

Yes, we were satisfied with our results. While an Adjusted R^2 of 0.58 might be considered low in physical sciences, it is a strong result for complex macroeconomic modeling, like that of GDP. We knew that capturing every driver of GDP would be difficult, so explaining nearly 60% of the variation with our selected variables is a success. Furthermore, the results we found demonstrated huge economic significance. The model predicts that a 1-unit increase in per capita CO2 is associated with a ~$15,000 increase in GDP. For developing nations—where GDP per capita often averages ~$7,000—this is transformative and validates our hypothesis that carbon intensity is a primary driver of economic development.