Factors Influencing CO2 Emissions

B1901922
Zhao Yao

INTRODUCTION

The pressing global challenge of climate change has prompted extensive research into understanding the factors that contribute to CO2 emissions. As societies strive to transition towards sustainable energy sources and reduce their carbon footprint, it becomes crucial to identify the key variables that influence these aspects. This research delves into the intricate relationship between CO2 emissions, energy consumption, and various socio-economic factors to shed light on potential pathways for achieving a greener and more sustainable future.

PROJECT OBJECTIVES

1

Identify Influential Factors: Determine which variables have the most significant influence on CO_2 emission per person.

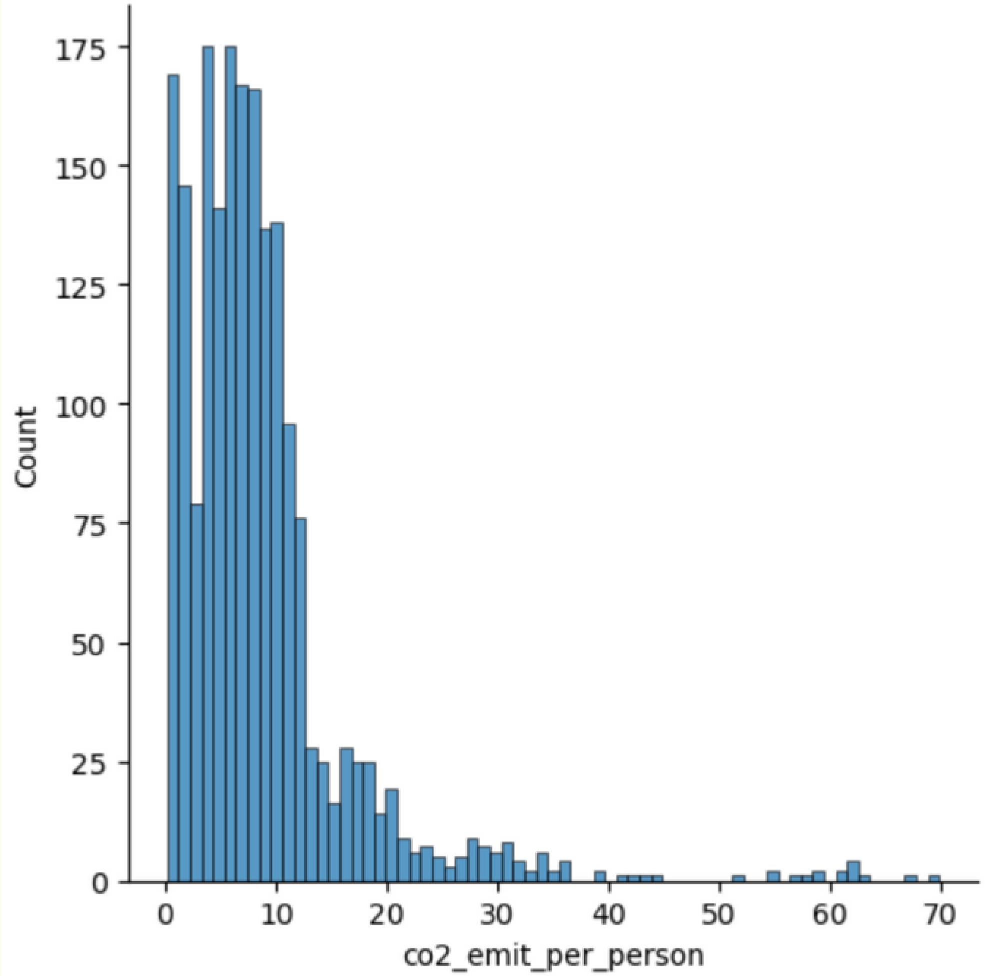
2

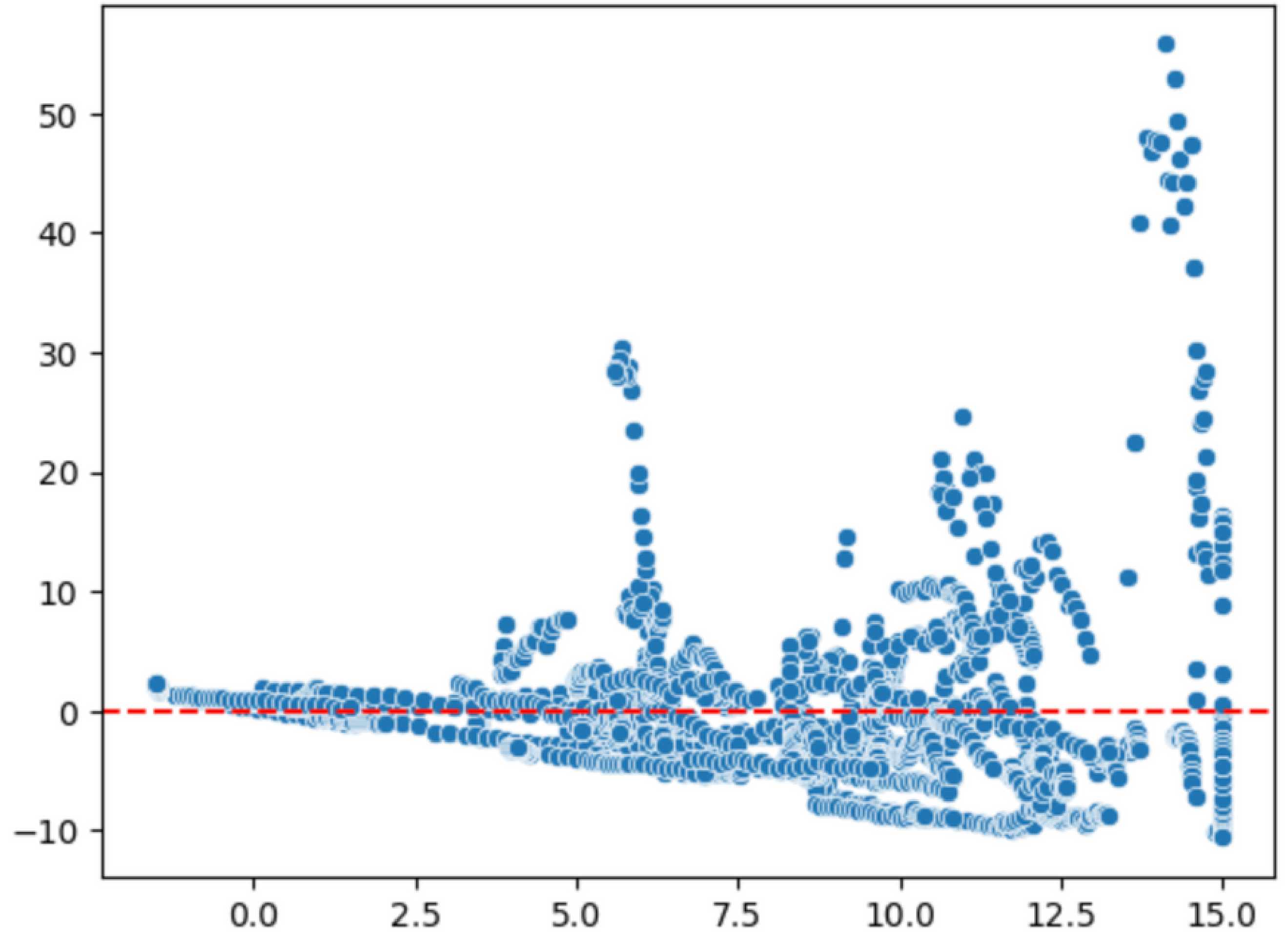
Assess Renewable Energy Impact: Evaluate the effectiveness of renewable energy sources in reducing CO_2 emissions and driving sustainable energy consumption.

3

Predictive Modeling: Develop predictive models to forecast CO_2 emission per person based on varying levels of fossil fuel usage, renewable energy usage and urbanization.

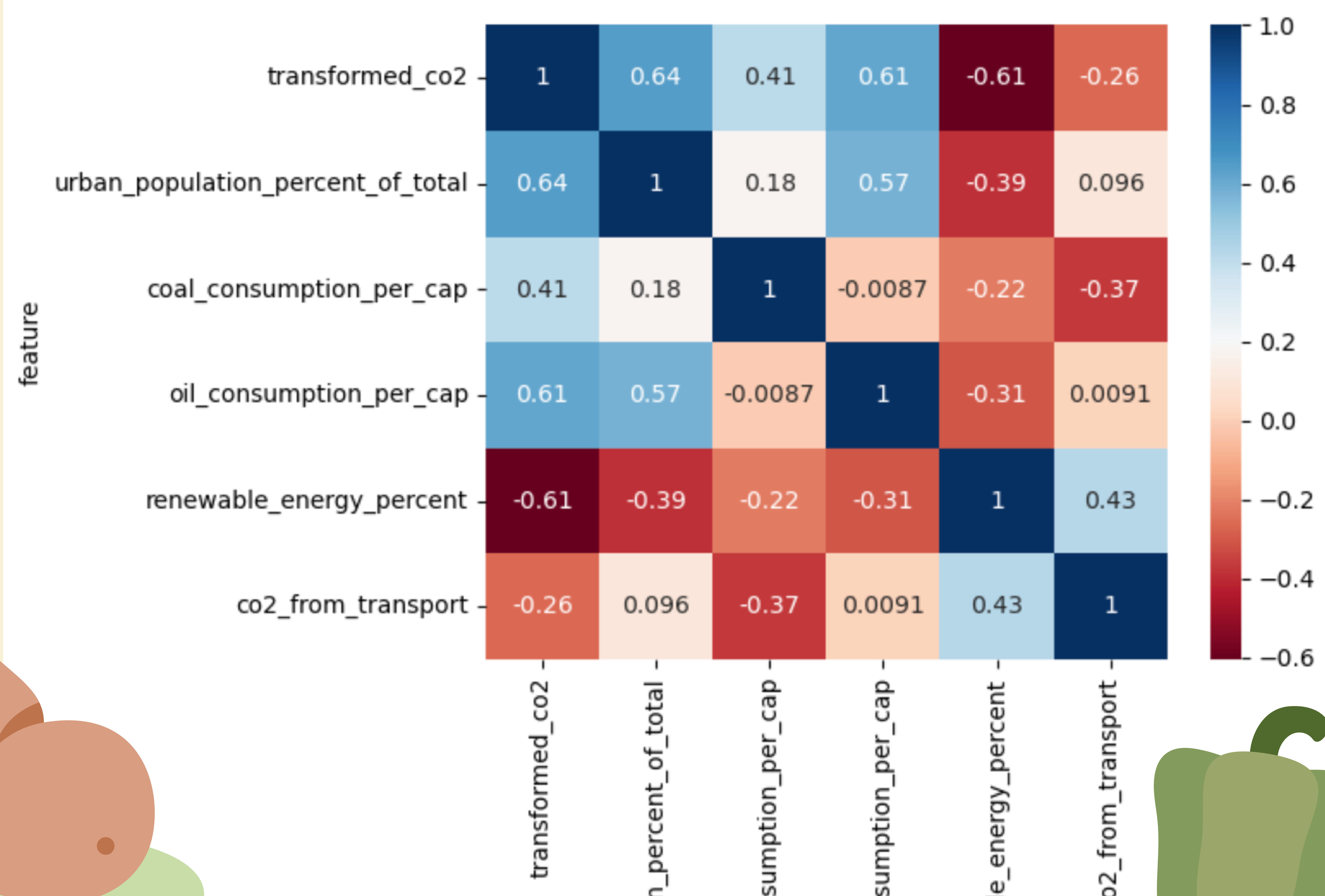
BOXCOX "Y"





Since the response variable is skewed (first graph), and the first linear model that I tried to build has severe heteroscedasticity (second graph), I decided to transform response variable co2_emit_per_person into transformed_co2 using boxcox method before building linear models.

CORRELATION HEATMAP



number of models

THREE MODELS

1. t_co2 ~ urbanization
2. t_co2 ~ urbanization
   + oil + coal
3. t_co2 ~ urbanization
   + oil + coal
   + renewable energy

REASONS

Variables added to First model: urbanization: 0.64 - moderate correlation to transformed_co2. Highest correlation among all vars.

Variables added to Second model: oil: 0.41, coal: 0.61 - moderate corr. These two variables were added because the burning of fossil fuels directly contributes to the emission of co2

Variables added to Third model: renewable energy: -0.61 - moderate corr. This variable is added because to mitigate excess co2 emission, renewable energy should be adopted.

LINEAR MODEL - 1

Result:

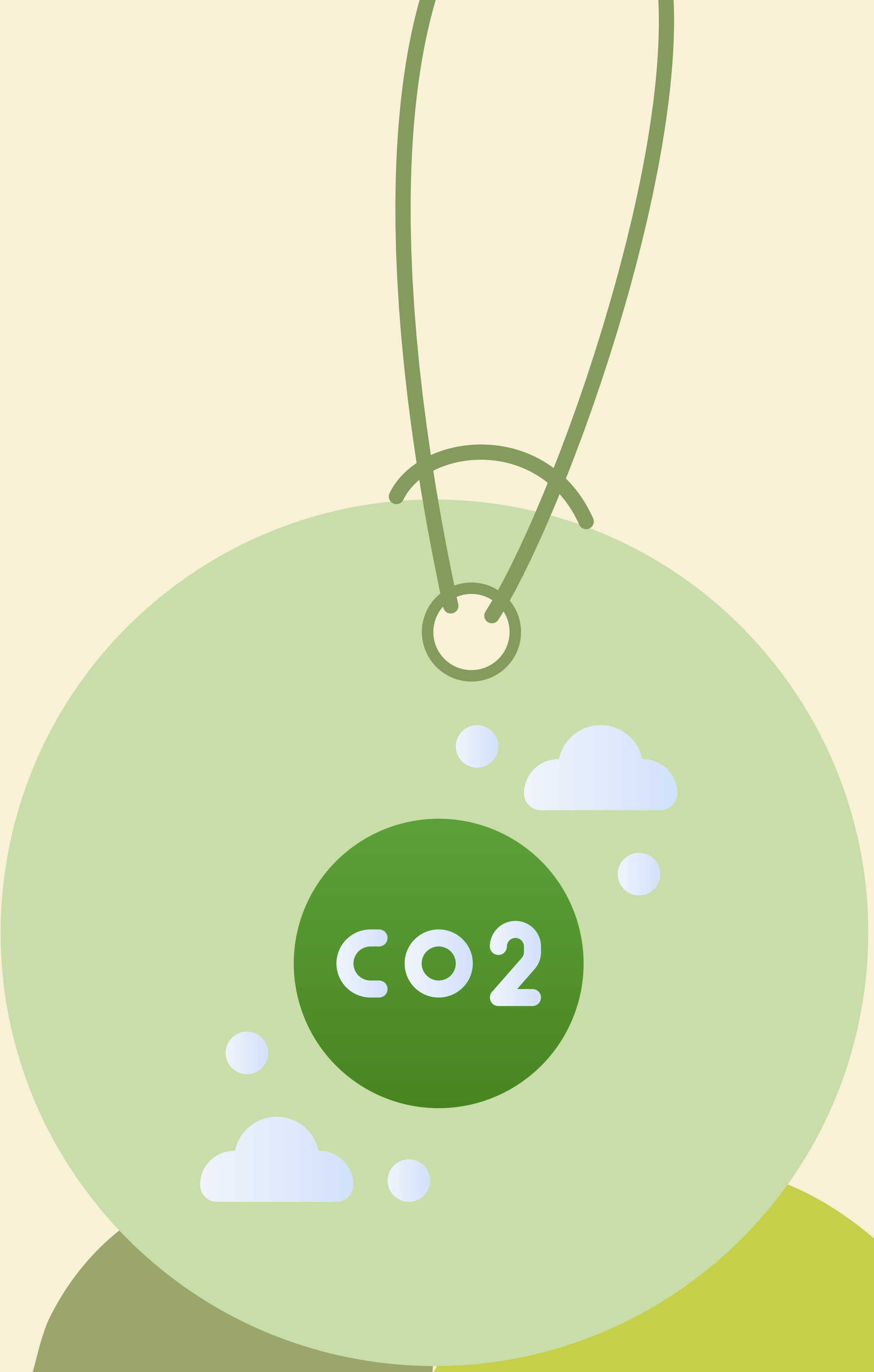
transformed_co2 = -0.7756 + 0.0423 * urban_population_percent_of_total

OLS Regression Results

Dep. Variable:	transformed_co2	R-squared:	0.414
Model:	OLS	Adj. R-squared:	0.414
Method:	Least Squares	F-statistic:	1377.
Date:	Mon, 07 Aug 2023	Prob (F-statistic):	1.79e-228
Time:	00:32:50	Log-Likelihood:	-2629.9
No. Observations:	1950	AIC:	5264.
Df Residuals:	1948	BIC:	5275.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.7756	0.080	-9.686	0.000	-0.933	-0.619
urban_population_percent_of_total	0.0423	0.001	37.111	0.000	0.040	0.045

Omnibus:	4.933	Durbin-Watson:	0.093
Prob(Omnibus):	0.085	Jarque-Bera (JB):	5.714
Skew:	0.021	Prob(JB):	0.0574
Kurtosis:	3.262	Cond. No.	266.



ANALYSE MODEL - 1

VARIABLES & COEFFICIENTS

1. coefficient of the urban population percentage is 0.0423 - as urban population increases 1 unit, transformed_co2 increases 0.0423
2. thriving of urbanization leads to increased CO2 emissions to meet people's demands in the city.

OLS RESULTS

1. R-squared value: 0.414 - 41.4% of the variation in the transformed_co2 be explained by urban_population_percent_of_total
2. P-value of urban_population_percent_of_total < 0.05 - significant variable

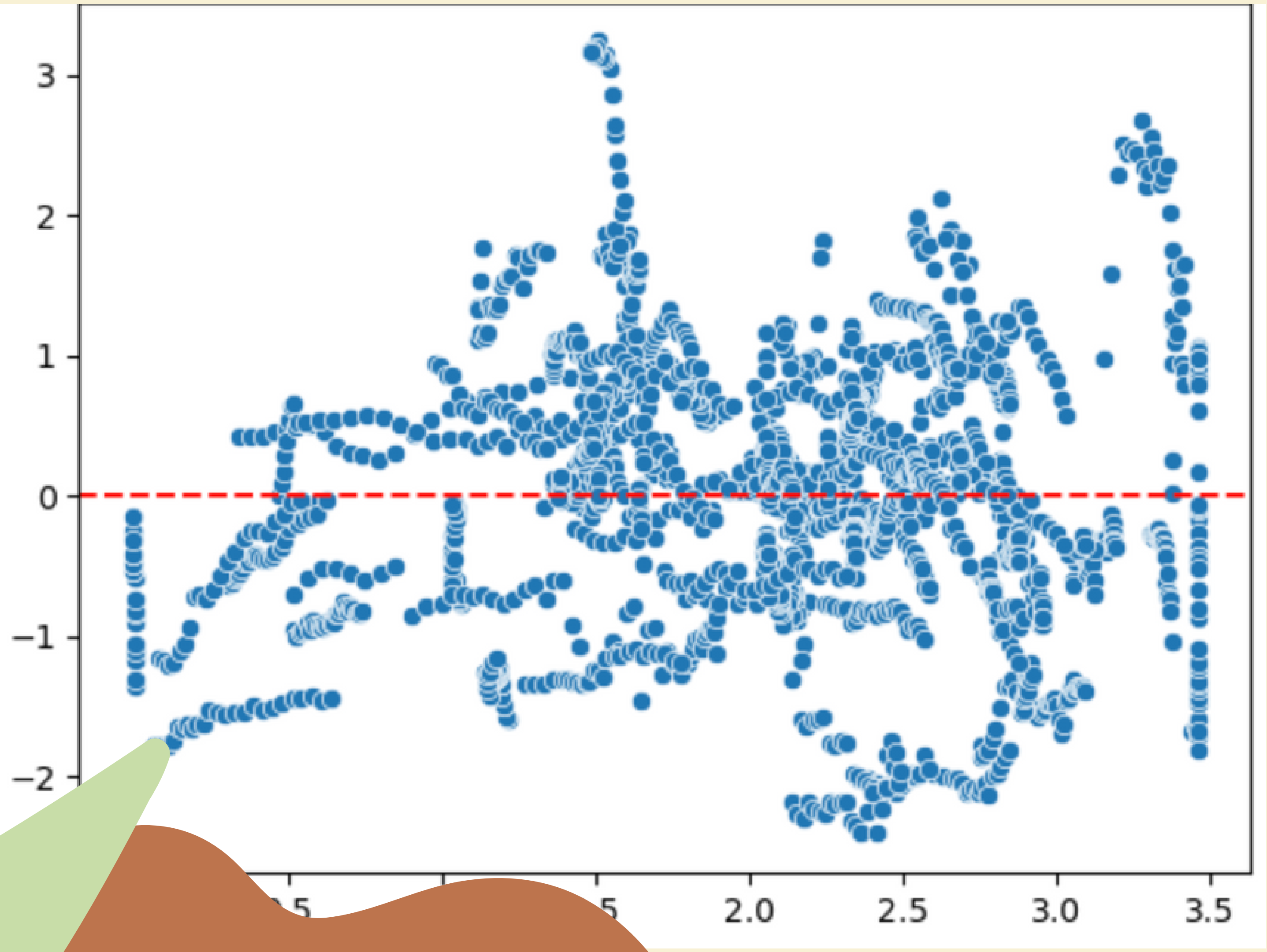
HYPOTHESIS TESTING - 1

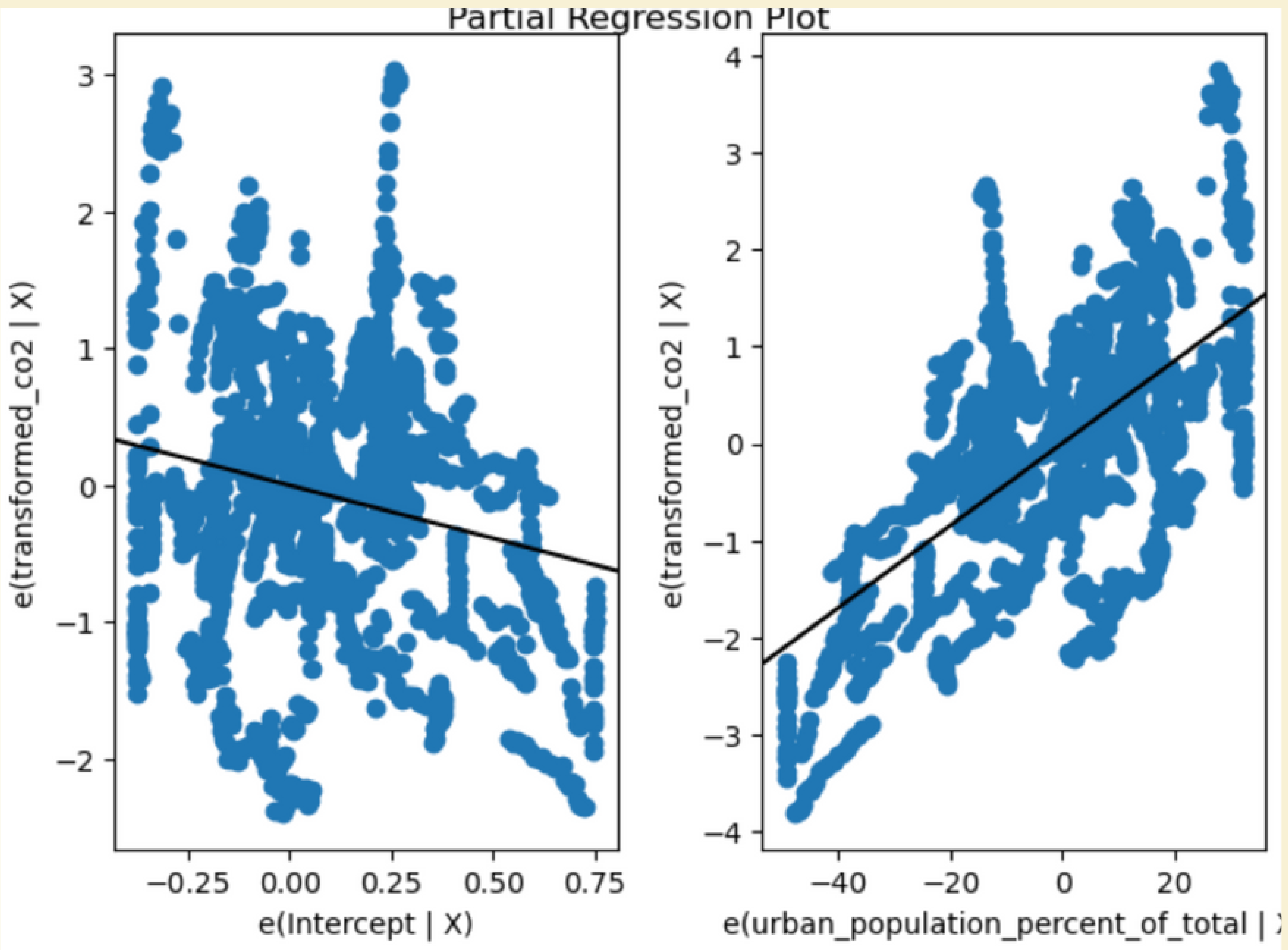
H0: There is no significant relationship between "urban_population_percent_of_total" and "transformed_co2".
H1: There is a significant relationship between "urban_population_percent_of_total" and "transformed_co2".

Since the p-value for "urban_population_percent_of_total" is smaller than 0.05, we can reject the null hypothesis in favor of the alternative hypothesis. This suggests that the "urban_population_percent_of_total" variable is statistically significant and has an effect on the transformed_co2.

MODEL ASSUMPTION - 1

1. Heteroscedasticity: In the first plot, the scatter points depicting residuals against fitted values exhibit a uniform distribution, suggesting the absence of Heteroscedasticity.
2. Partial regression plot: Most data align to the partial regression line. As the urban population goes up, transformed_co2 goes up. Hence, it becomes imperative to mitigate the negative effects of urbanization: curbing deforestation, and promoting the widespread adoption of renewable energy sources within urban areas.





LINEAR MODEL - 2

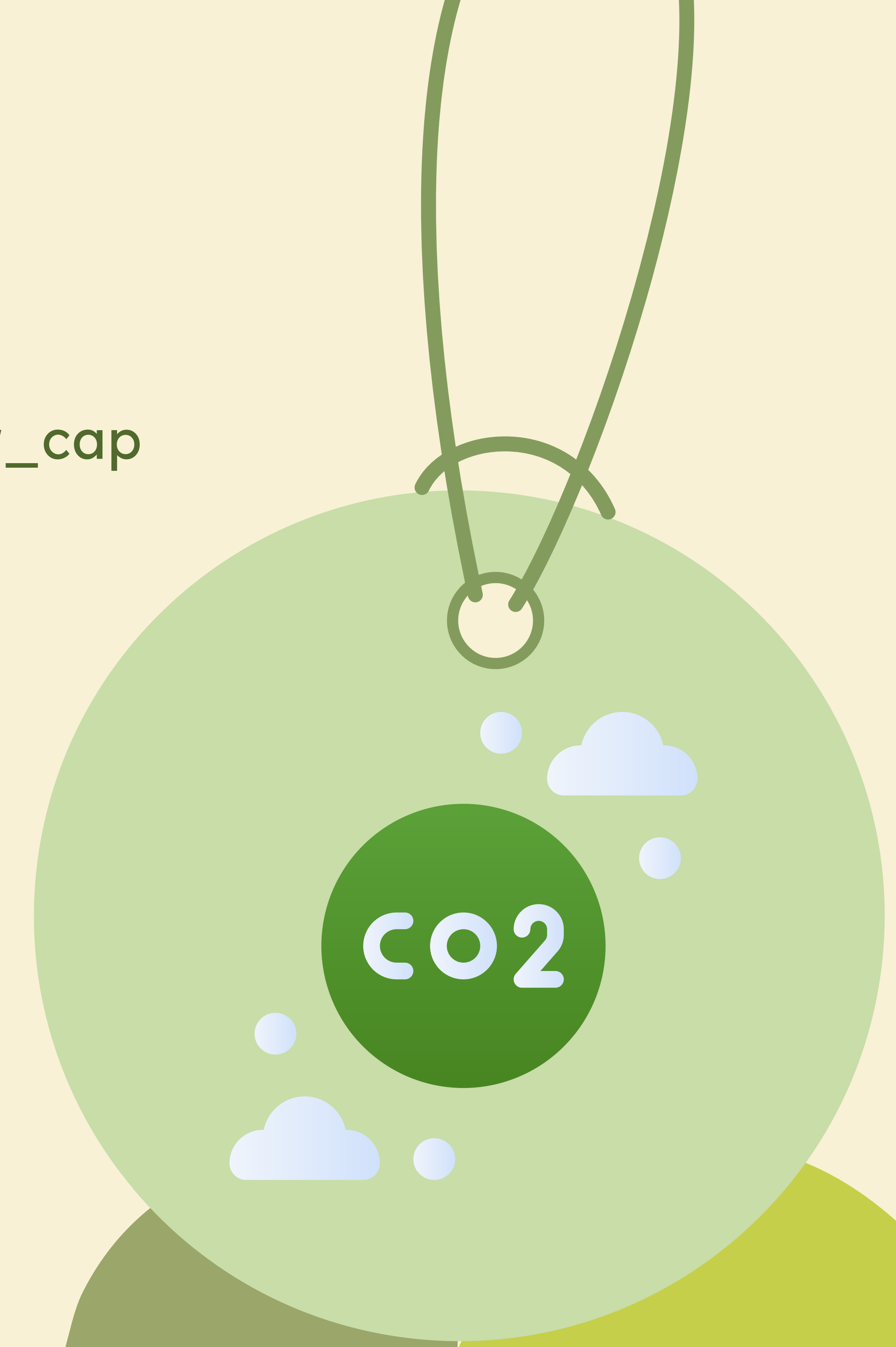
transformed_co2 = -0.2315 + 0.0223 * urban_population_percent_of_total + 0.6999 * coal_consumption_per_cap + 0.3231 * oil_consumption_per_cap

OLS Regression Results

Dep. Variable:	transformed_co2	R-squared:	0.622
Model:	OLS	Adj. R-squared:	0.622
Method:	Least Squares	F-statistic:	1068.
Date:	Mon, 07 Aug 2023	Prob (F-statistic):	0.00
Time:	00:33:18	Log-Likelihood:	-2202.5
No. Observations:	1950	AIC:	4413.
Df Residuals:	1946	BIC:	4435.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.2315	0.069	-3.364	0.001	-0.366	-0.097
urban_population_percent_of_total	0.0223	0.001	19.475	0.000	0.020	0.025
coal_consumption_per_cap	0.6999	0.028	24.844	0.000	0.645	0.755
oil_consumption_per_cap	0.3231	0.013	24.471	0.000	0.297	0.349

Omnibus:	189.142	Durbin-Watson:	0.098
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1166.998
Skew:	0.207	Prob(JB):	3.89e-254
Kurtosis:	6.767	Cond. No.	286.



ANALYSE MODEL - 2

VARIABLES & COEFFICIENTS

1. every 1 unit increase in coal, transformed_co2 increase by 0.6999
2. every 1 unit increase in oil, transformed_co2 increase 0.3231
3. coal affects co2 emission the most, indicating among these two fossil fuels, coal usage should be limited

OLS RESULTS

1. R-squared value: 0.622 - 62.2% of the variation in the transformed_co2 is explained by current variables (improved)
2. P-values of each variable < 0.05 - all variables are significant to transformed_co2

HYPOTHESIS TESTING - 2

H0: Given the variance of transformed_co2 is explained by other two variables, there is no significant relationship between "coal_consumption_per_cap" and "transformed_co2".
H1: Given the variance of transformed_co2 is explained by other two variables, there is a significant relationship between "coal_consumption_per_cap" and "transformed_co2".

Since the p-value for "coal_consumption_per_cap" is smaller than 0.05, we can reject the null hypothesis in favor of the alternative hypothesis. This suggests that the "coal_consumption_per_cap" variable is statistically significant and has an effect on the transformed_co2.

H0: Given the variance of transformed_co2 is explained by other two variables, there is no significant relationship between "oil_consumption_per_cap" and "transformed_co2".
H1: Given the variance of transformed_co2 is explained by other two variables, there is a significant relationship between "oil_consumption_per_cap" and "transformed_co2".

Since the p-value for "oil_consumption_per_cap" is smaller than 0.05, we can reject the null hypothesis in favor of the alternative hypothesis. This suggests that the "oil_consumption_per_cap" variable is statistically significant and has an effect on the transformed_co2.

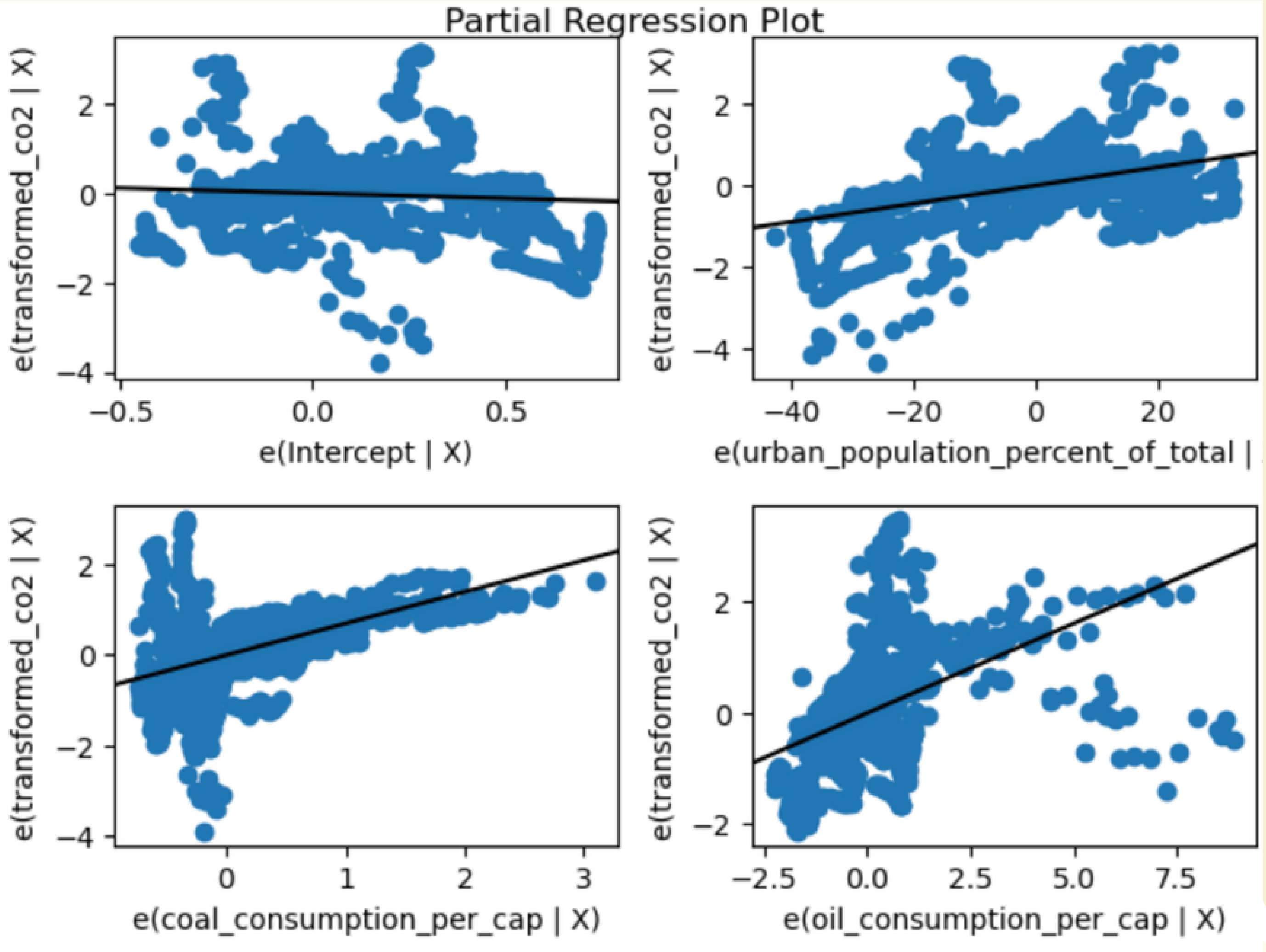
MODEL ASSUMPTION - 2

1. VIF testing for collinearity: In the first graph, the VIF values are relatively close to 1 for all independent variables, indicating that there is no significant multicollinearity among the predictors. They can be considered independently in the model without causing issues due to multicollinearity.
2. Partial regression plot: Most data align to the partial regression line. Some outliers could be removed in oil consumption graph since they blunt the steepness of the line.

```
vifs = [vif(dm, i) for i in range(4)]
```
✓ 0.0s

```
for v,n in zip(vifs, names):
    print(f"{n:<20} {v:.4f}")
```
✓ 0.0s

```
Intercept            16.4428
urban_population_percent_of_total 1.5683
coal_consumption_per_cap 1.0511
oil_consumption_per_cap 1.5202
```



LINEAR MODEL - 3

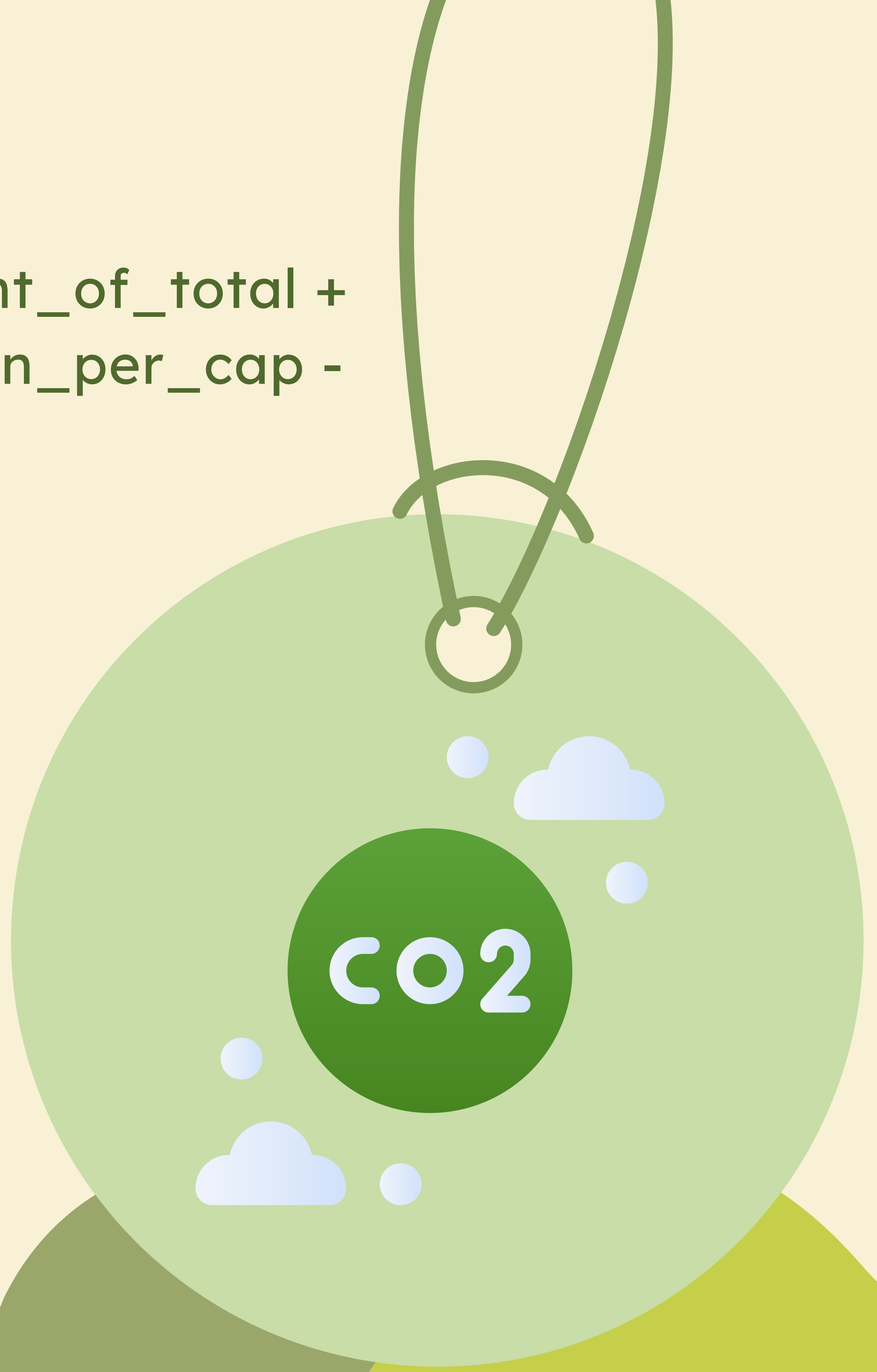
transformed_co2 = 0.6405 + 0.0166 * urban_population_percent_of_total + 0.5858 * coal_consumption_per_cap + 0.2826 * oil_consumption_per_cap - 0.0228 * renewable_energy_percent

OLS Regression Results

Dep. Variable:	transformed_co2	R-squared:	0.709
Model:	OLS	Adj. R-squared:	0.709
Method:	Least Squares	F-statistic:	1187.
Date:	Mon, 07 Aug 2023	Prob (F-statistic):	0.00
Time:	21:49:52	Log-Likelihood:	-1946.2
No. Observations:	1950	AIC:	3902.
Df Residuals:	1945	BIC:	3930.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.6405	0.070	9.109	0.000	0.503	0.778
renewable_energy_percent	-0.0228	0.001	-24.179	0.000	-0.025	-0.021
coal_consumption_per_cap	0.5858	0.025	23.287	0.000	0.536	0.635
oil_consumption_per_cap	0.2826	0.012	24.149	0.000	0.260	0.306
urban_population_percent_of_total	0.0166	0.001	16.089	0.000	0.015	0.019

Omnibus:	219.746	Durbin-Watson:	0.096
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1566.461
Skew:	0.259	Prob(JB):	0.00
Kurtosis:	7.360	Cond. No.	339.



ANALYSE MODEL - 3

VARIABLES & COEFFICIENTS

1. urban: 0.0166, coal: 0.5858, oil: 0.2826, renewable: -0.0228
2. Every 1 unit increase of renewable energy leads to 0.0228 unit of decrease in transformed_co2, indicating more renewable energy should be adopted to reduce CO2 emission

OLS RESULTS

1. R-squared value: 0.709 - 70.9% of the variation in the transformed_co2 is explained by current variables (improved)
2. P-values of each variable < 0.05 - all variables are significant to transformed_co2

HYPOTHESIS TESTING - 3

H0: Given the variance of transformed_co2 is explained by the other three variables, there is no significant relationship between "renewable_energy_percent" and "transformed_co2".

H1: Given the variance of transformed_co2 is explained by the other three variables, there is a significant relationship between "renewable_energy_percent" and "transformed_co2".

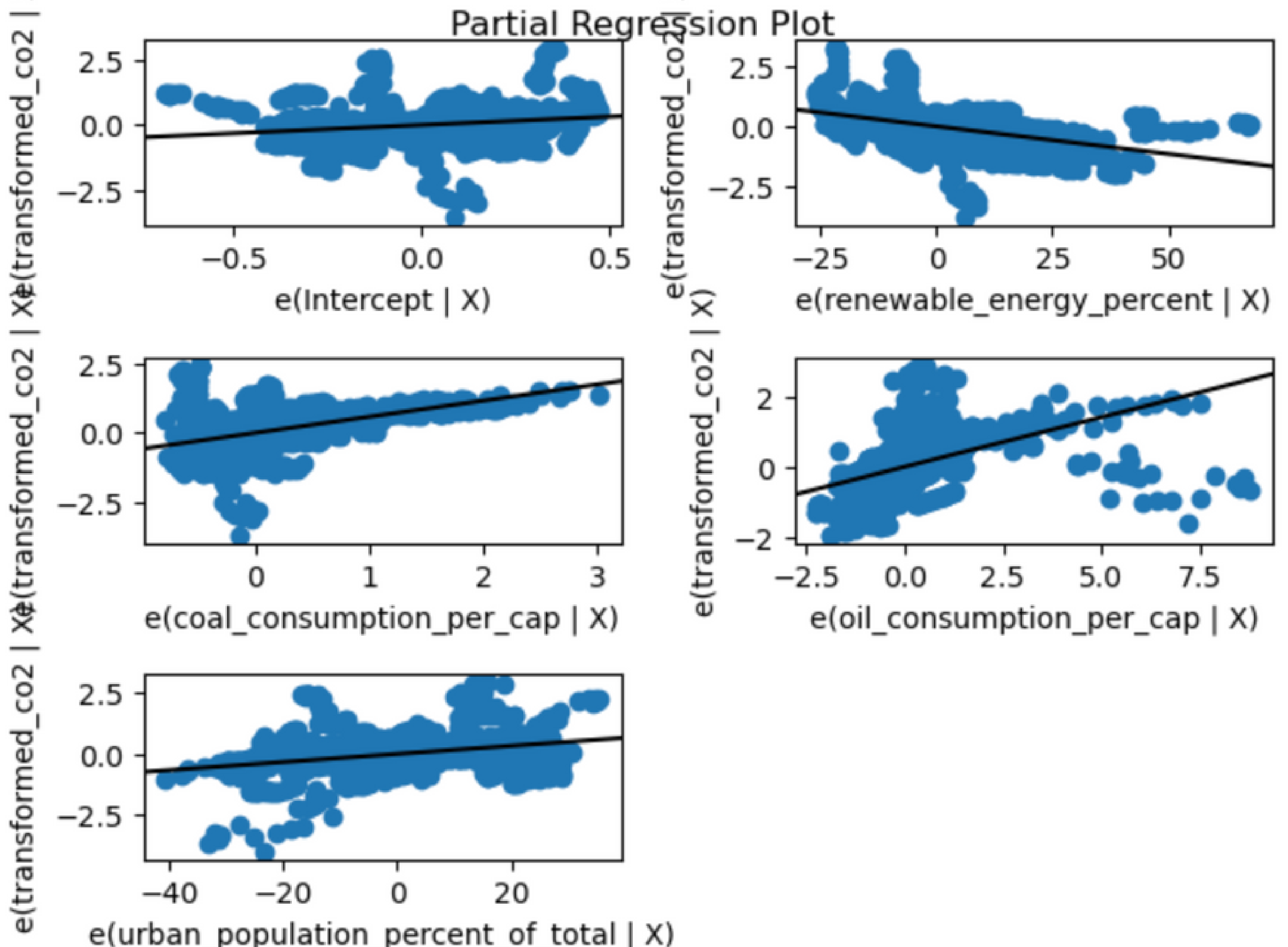
Since the p-value for "renewable_energy_percent" is smaller than 0.05, we can reject the null hypothesis in favor of the alternative hypothesis. This suggests that the "renewable_energy_percent" variable is statistically significant and has an effect on the transformed_co2.

MODEL ASSUMPTION - 3

1. VIF testing for collinearity: In the first graph, the VIF values are relatively close to 1 for all independent variables, indicating no significant multicollinearity among the predictors. They can be considered independently in the model without causing issues due to multicollinearity.
2. Partial regression plot: Most data align to the partial regression line. Some influential points could be removed in the oil consumption graph since they blunt the steepness. As more renewable energy is adopted, the co2 emission decreases.

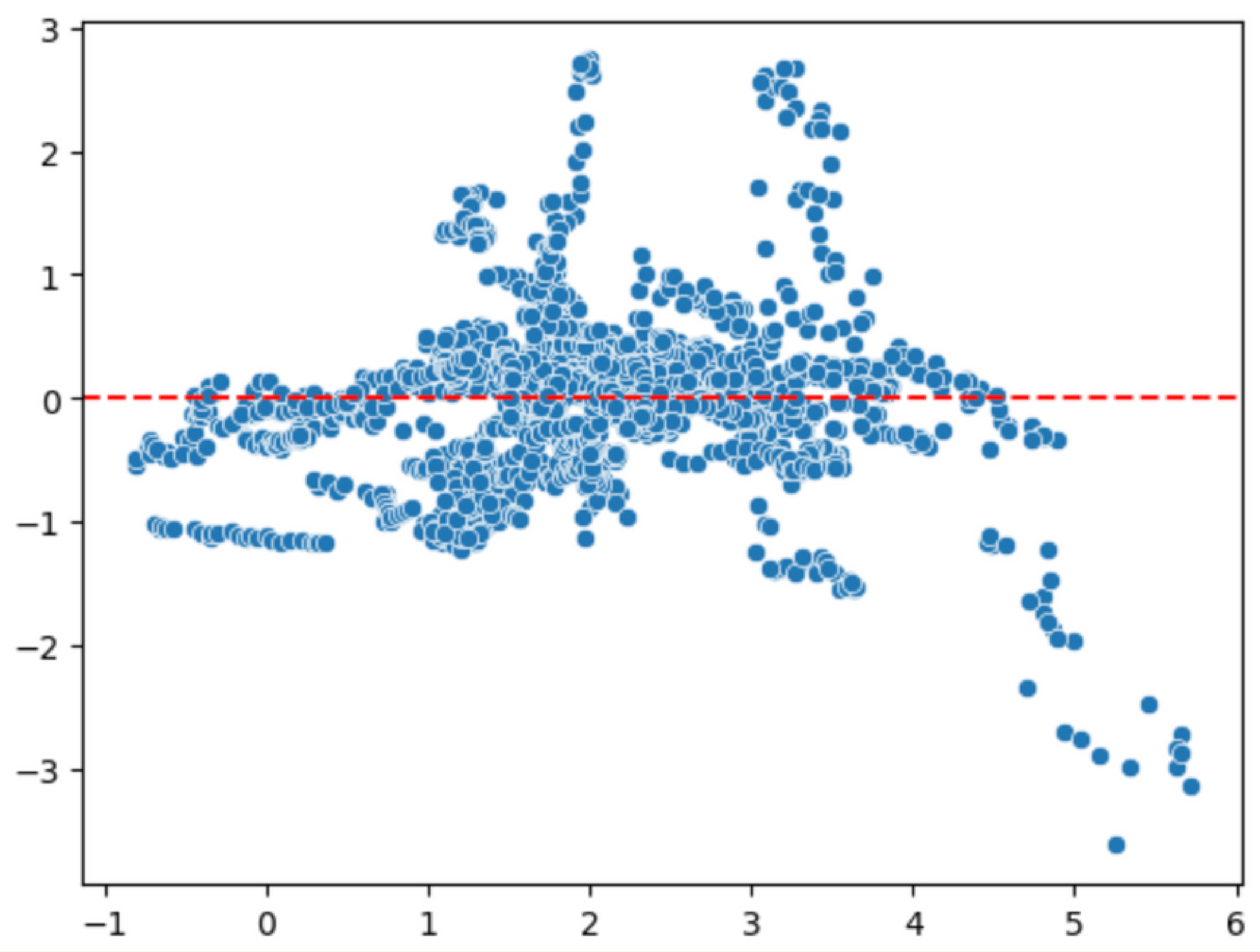
```
for v,n in zip(vifs, names):
    print(f"{n:<20} {v:.4f}")
✓ 0.0s

Intercept                22.3123
renewable_energy_percent 1.2365
coal_consumption_per_cap 1.0894
oil_consumption_per_cap 1.5520
urban_population_percent_of_total 1.6548
```



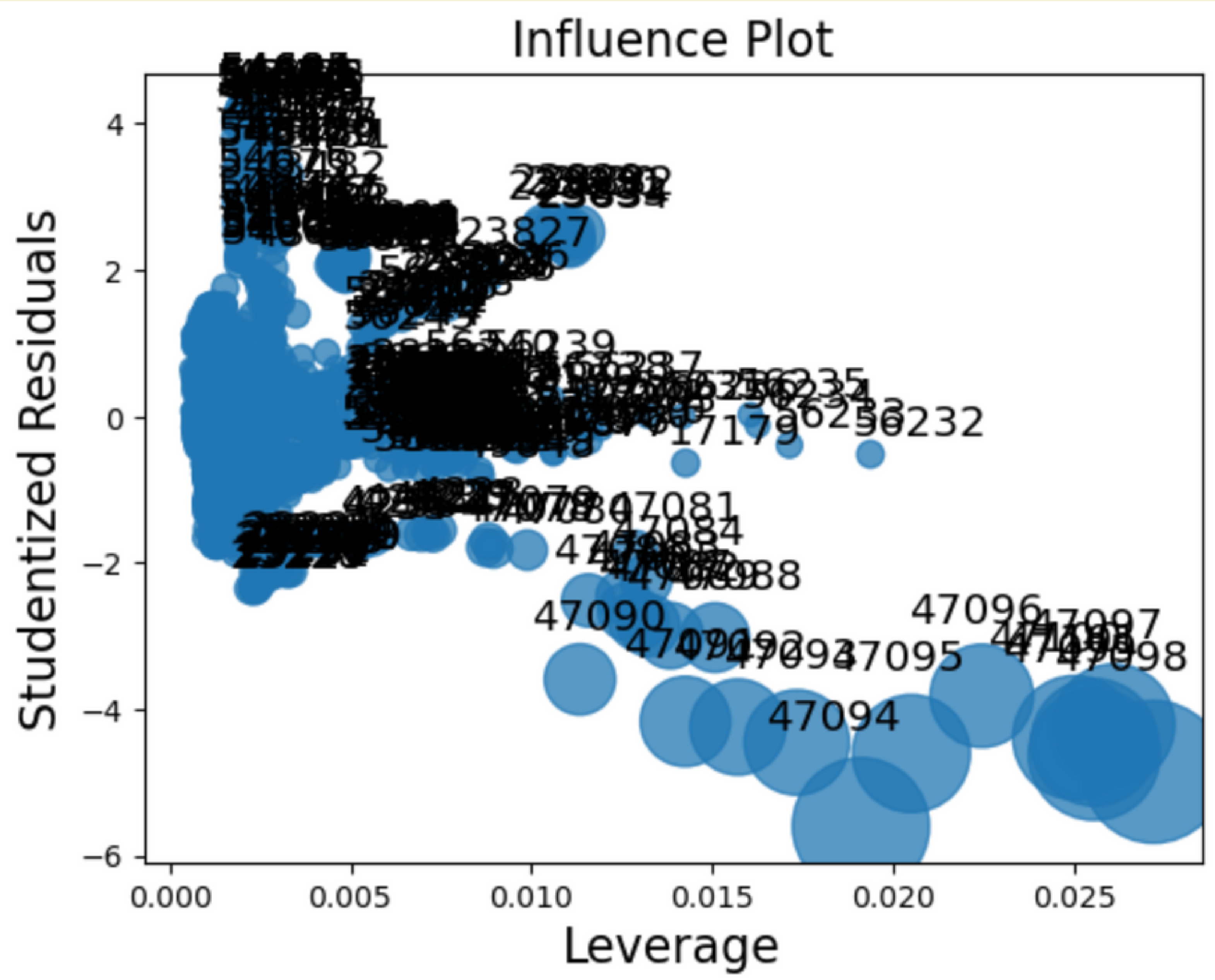
MODEL ASSUMPTION - 3

3. Heteroscedasticity is evident in the initial graph where scatter points of residuals against fitted values display a cone-shaped pattern above and below the line. This pattern, possibly due to the non-linear relationship between transformed_co2 and oil/coal consumption variables, suggests potential inaccuracies in the model's predictions. This issue could potentially be solved by constructing a polynomial model that incorporates the squared variables (oil or coal consumption).



MODEL ASSUMPTION - 3

4. Influential points: Points with high leverage exhibit low studentized residuals, while points with high studentized residuals have low leverage. No points display both high studentized residuals and high leverage, indicating that there are no significant influential points requiring exclusion.



ANOVA TEST

```python
#anova test among three models
sm.stats.anova_lm(co2_model1,co2_model2,co2_model3)
```

✓ 0.0s

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	1948.0	1694.332221	0.0	NaN	NaN	NaN
1	1946.0	1092.974599	2.0	601.357622	695.897851	1.022449e-228
2	1945.0	840.382373	1.0	252.592226	584.605169	3.804585e-113

1. co2_model1: co2 ~ urbanization
2. co2_model2: co2 ~ urbanization + oil + coal
3. co2_model3: co2 ~ urbanization + oil + coal + renewable

MODEL SELECTION

Hypothesis test for co2_model1 and co2_model2
H0: co2_model1: transformed_co2 = b1 * urban_population_percent_of_total (b2=b3=0)
H1: co2_model2: transformed_co2 = b1 * urban_population_percent_of_total + b2 * coal_consumption_per_cap + b3 * oil_consumption_per_cap

From the ANOVA test result (row 1), the Pr(>F) value for is 1.022449e-228 < 0.05. Therefore, assuming that co2_model1, we can reject the null hypothesis that b2=b3=0, i.e. at least coal_consumption_per_cap or oil_consumption_per_cap does explain variation in transformed_co2.

Hypothesis test for co2_model2 and co2_model3
H0: co2_model2: transformed_co2 = b1 * urban_population_percent_of_total + b2 * coal_consumption_per_cap + b3 * oil_consumption_per_cap (b4 = 0)
H1: co2_model3: transformed_co2 = b1 * urban_population_percent_of_total + b2 * coal_consumption_per_cap + b3 * oil_consumption_per_cap - b4 * renewable_energy_percent

From the ANOVA test result (row 2), the Pr(>F) value is 3.804585e-113 < 0.05. Therefore, assuming that co2_model2, we can reject the null hypothesis that b4=0, i.e. renewable_energy_percent is a significant independent variable to transformed_co2.

From the above anova and hypothesis testing, we come to the conclusion that those four variables are significant to the model, meaning oil/coal consumption, urbanization and renewable energy all affect CO2 emission, thus we choose co2_model3 that includes all these four variables. As oil/coal consumption increase, co2 emission increase; as renewable energy increases, co2 emission decreases. Thus, government should regulate the use of fossil fuel and promote renewable energy usage.

REGULARIZATION

Ridge Model:
transformed_co2 = -0.3776 * renewable_energy_percent + 0.3514 * coal_consumption_per_cap + 0.4882 * oil_consumption_per_cap + 0.2856 * urban_population_percent_of_total

Coefficient Estimates: [-0.37757603 0.35139567 0.48818124 0.28558495]
R-squared on Test Data: 0.7062197647529629
Best Alpha: 20.69138081114788

This model has achieved a moderate level of prediction accuracy, as indicated by the R-squared value. The choice of higher alpha indicates that thee model is emphasizing regularization to prevent overfitting, which can be beneficial when dealing with potential multicollinearity or high-dimensionality in the data. However, previous linear models didn't exhibit multicollinearity, and the potential bias introduced by the relatively big alpha value might impact accuracy. Therefore, this model won't be the preferred choice.

POLYNOMIAL REGRESSION

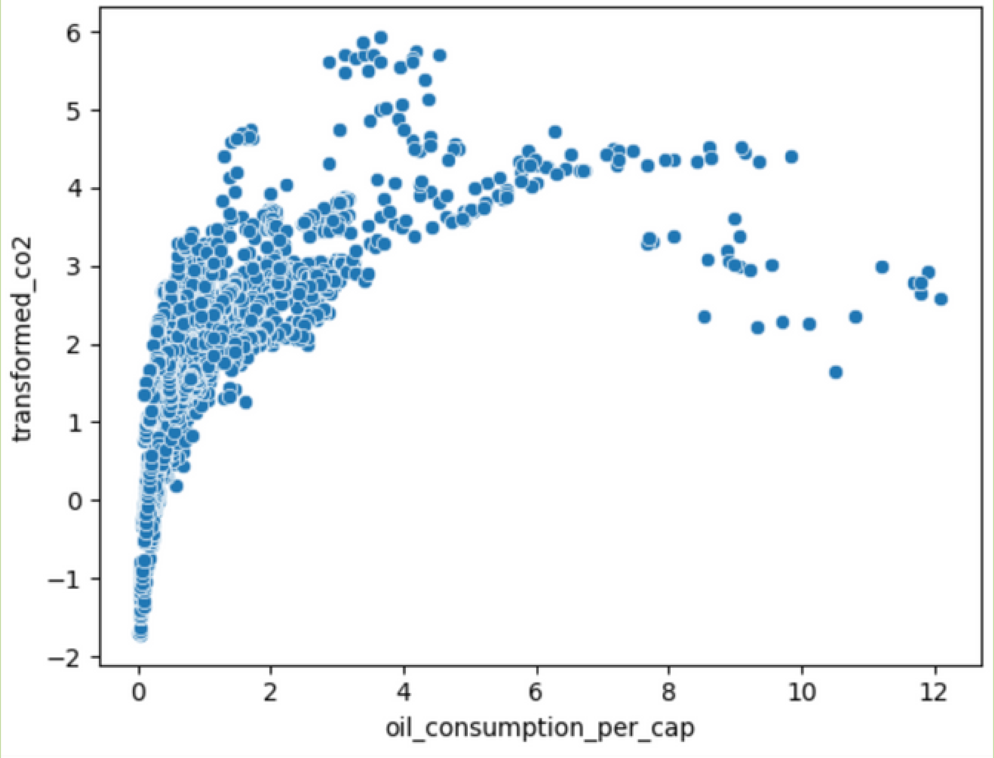
The squared of oil_consumption_per_cap is being selected to form the new polynomial model since it has a non-linear relationship with transformed_co2. Additionally, it offers best R-squared value (0.812) among all other variable's squared or cubed form.

OLS Regression Results

Dep. Variable:	transformed_co2	R-squared:	0.812
Model:	OLS	Adj. R-squared:	0.811
Method:	Least Squares	F-statistic:	1679.
Date:	Tue, 08 Aug 2023	Prob (F-statistic):	0.00
Time:	16:50:59	Log-Likelihood:	-1522.1
No. Observations:	1950	AIC:	3056.
Df Residuals:	1944	BIC:	3090.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.8305	0.057	14.600	0.000	0.719	0.942
renewable_energy_percent	-0.0214	0.001	-28.184	0.000	-0.023	-0.020
coal_consumption_per_cap	0.5085	0.020	24.949	0.000	0.469	0.548
oil_consumption_per_cap	0.9783	0.023	41.886	0.000	0.932	1.024
urban_population_percent_of_total	0.0044	0.001	4.817	0.000	0.003	0.006
I(oil_consumption_per_cap ** 2)	-0.0786	0.002	-32.550	0.000	-0.083	-0.074

Omnibus:	235.962	Durbin-Watson:	0.111
Prob(Omnibus):	0.000	Jarque-Bera (JB):	698.151
Skew:	0.629	Prob(JB):	2.50e-152
Kurtosis:	5.648	Cond. No.	343.



ANALYSE MODEL

VARIABLES & COEFFICIENTS

1. urban: 0.0166, coal: 0.5858, oil: 0.2826, renewable: -0.0228, oil^2: -0.0786
2. Every 1 unit increase in oil leads to 0.2826 increase in transformed_co2; Every 1 unit increase in oil^2 leads to 0.0786 decrease in transformed_co2.

RESULTS

1. R-squared value: 0.812 - 81.2% of the variation in the transformed_co2 is explained by current variables (improved)
2. P-values of each variable < 0.05 - all variables are significant to transformed_co2

ANOVA TEST

```
sm.stats.anova_lm(co2_model3,co2_model4)
```

✓ 0.0s

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	1945.0	840.382373	0.0	NaN	NaN	NaN
1	1944.0	543.935092	1.0	296.447281	1059.489493	6.975444e-186

1. co2_model3: co2 ~ urbanization + oil + coal + renewable

2. co2_model4: co2 ~ urbanization + oil + coal + renewable + oil^2

HYPOTHESIS TESTING

Hypothesis test for co2_model3 and co2_model4
H0: co2_model3: transformed_co2 = b1 * urban_population_percent_of_total + b2 * coal_consumption_per_cap + b3 * oil_consumption_per_cap + b4 * renewable_energy_percent (b5 = 0) (restricted model)

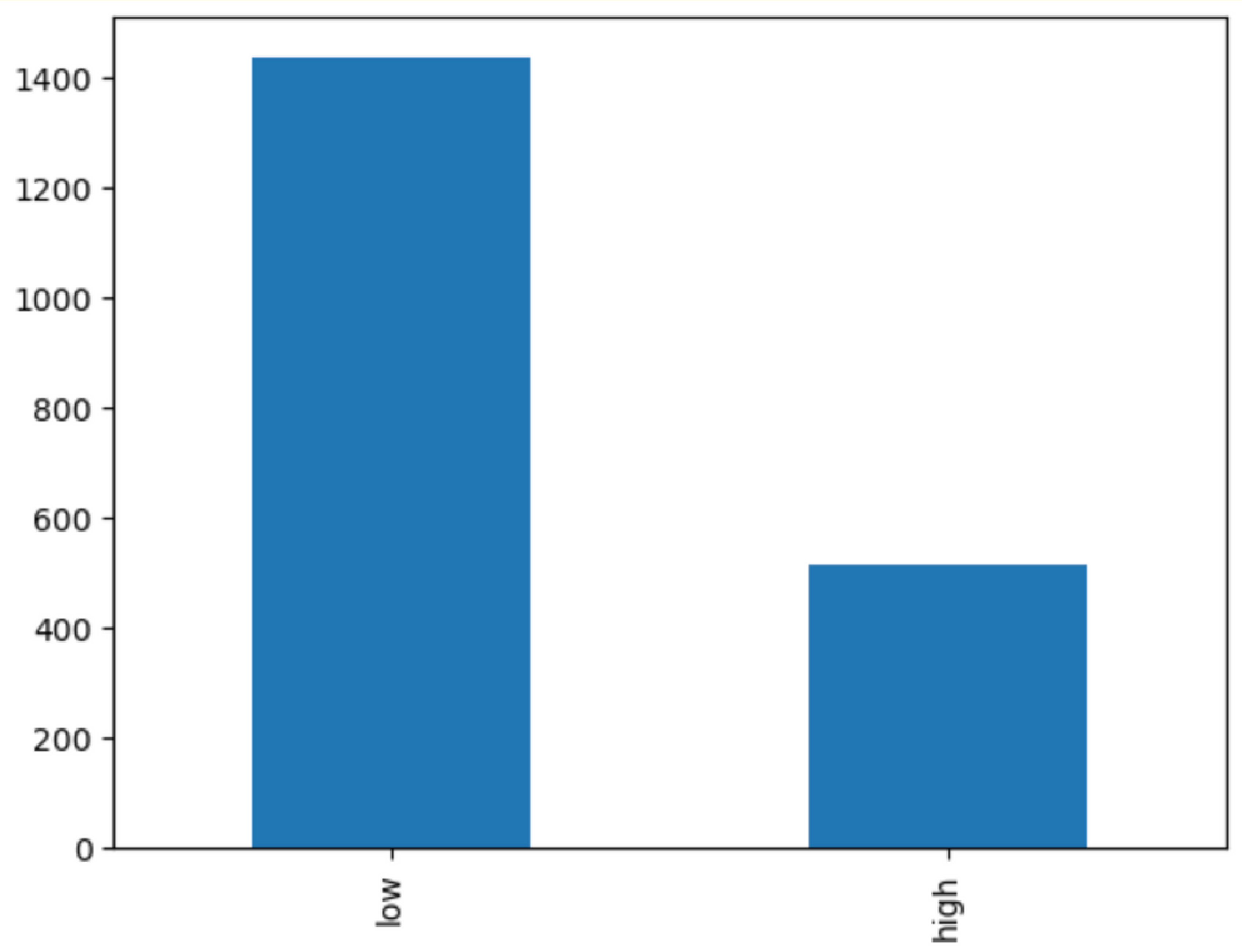
H1: co2_model4: transformed_co2 = b1 * urban_population_percent_of_total + b2 * coal_consumption_per_cap + b3 * oil_consumption_per_cap - b4 * renewable_energy_percent + b5 * oil_consumption_per_cap^2 (unrestricted model)

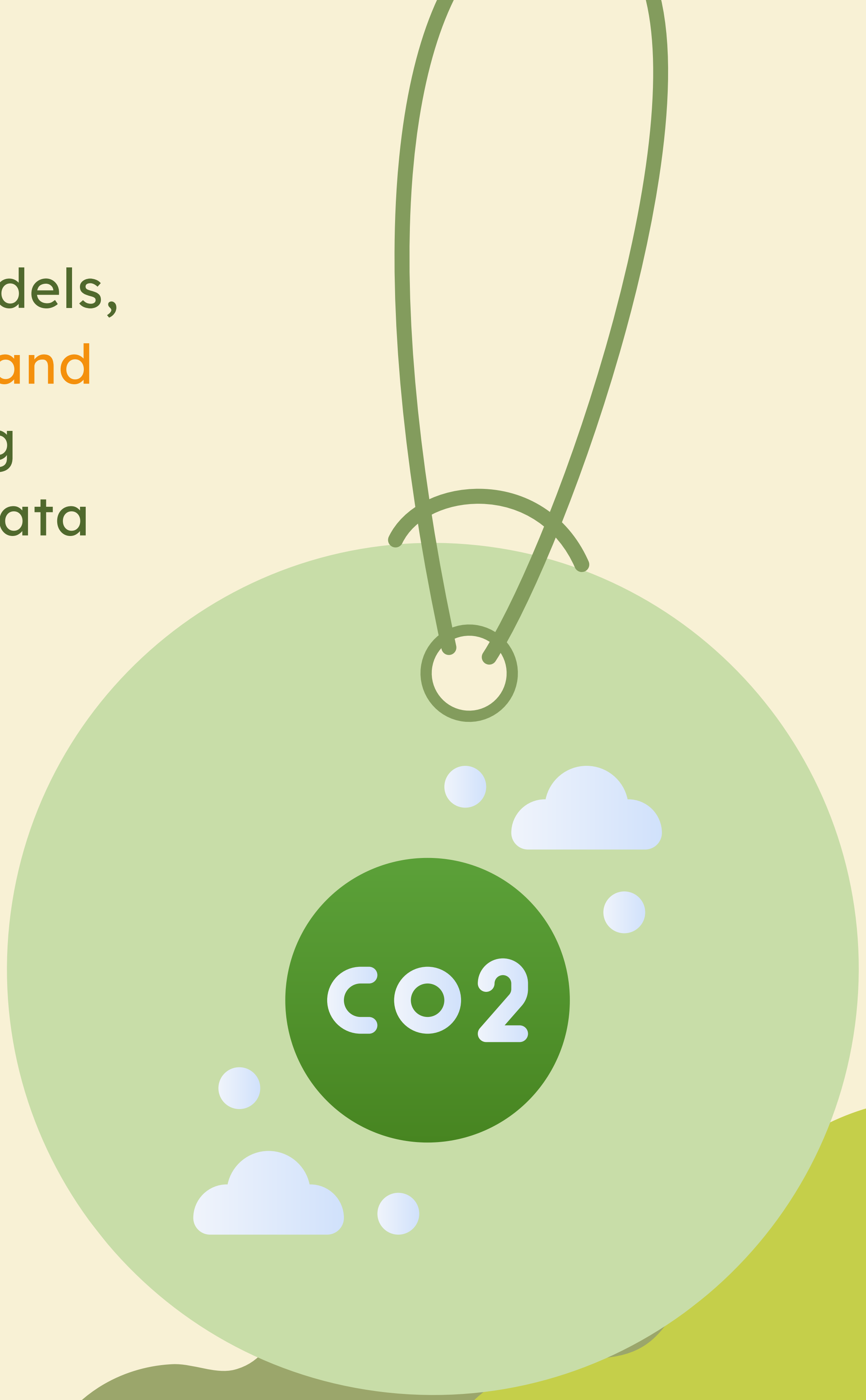
From the ANOVA test result (row 2), the Pr(>F) value is 6.975444e-186 < 0.05. Therefore, assuming that co2_model3, we can reject the null hypothesis that b5=0, i.e. oil_consumption_per_cap^2 is a significant independent variable to transformed_co2.

From the above anova and hypothesis testing, we come to the conclusion that all five variables are significant to the model, thus we choose co2_model4 that includes all these five variables.

CLASSIFICATION MODELS

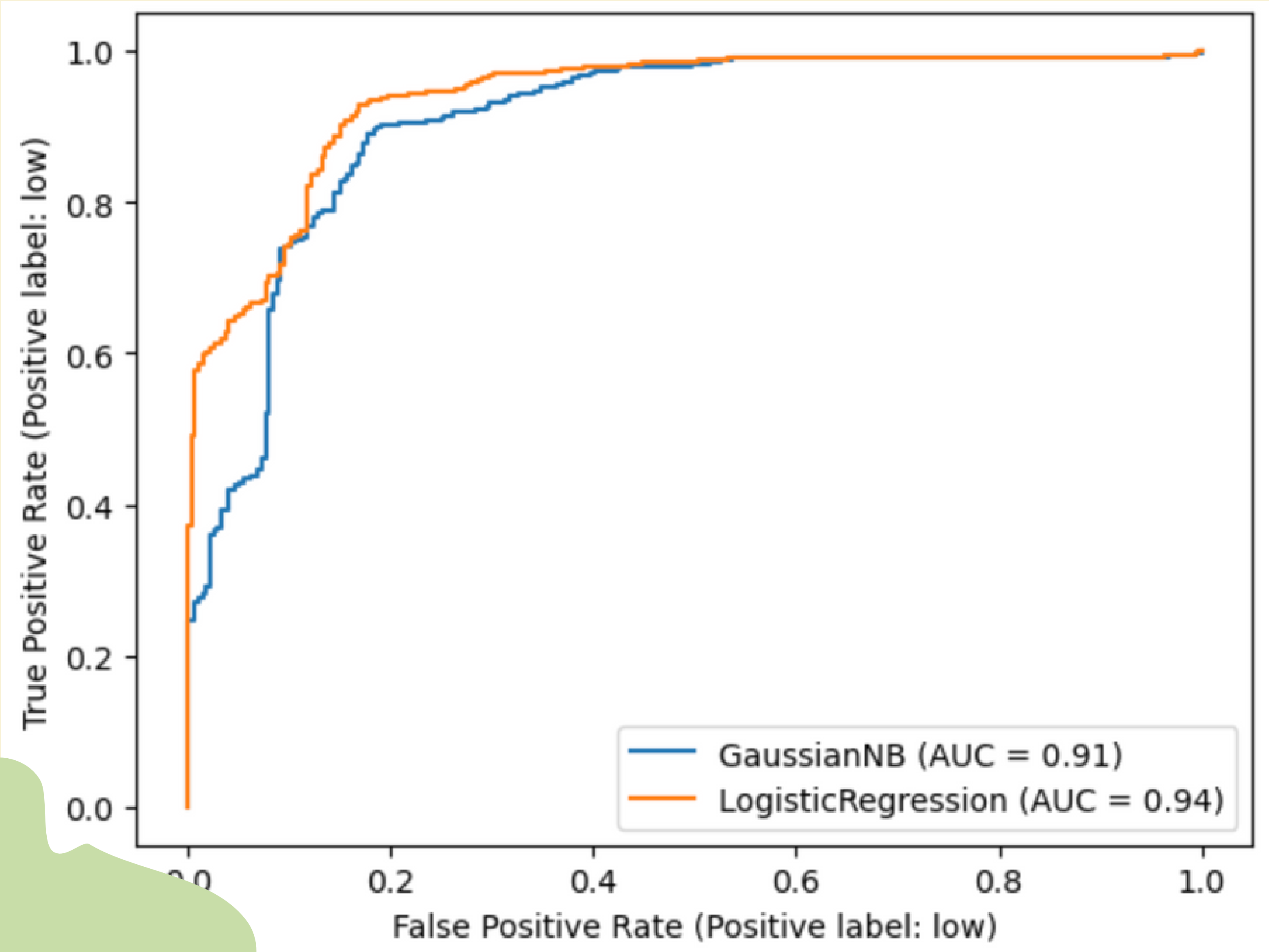
For both the Gaussian Naive Bayes and Logistic regression models, the data has been categorized into two distinct groups: "high" and "low" based on co2_emission_per_person variable. Data falling within the range of 0.0672 to 10 is categorized as "low," while data within the range of 10 to 69.9 is categorized as "high."





COMPARISON - ROC CURVE

The Logistic Regression model has a higher TPR and a lower FPR than the GaussianNB model. This means that the Logistic Regression model is better at correctly classifying positive samples as positive. The AUC (area under the curve) for both models is above 0.9, which is considered to be a good AUC. This means that both models are relatively accurate at classifying positive and negative samples. Overall, the Logistic Regression model is the better choice for this task, as it has a higher TPR and a lower FPR.

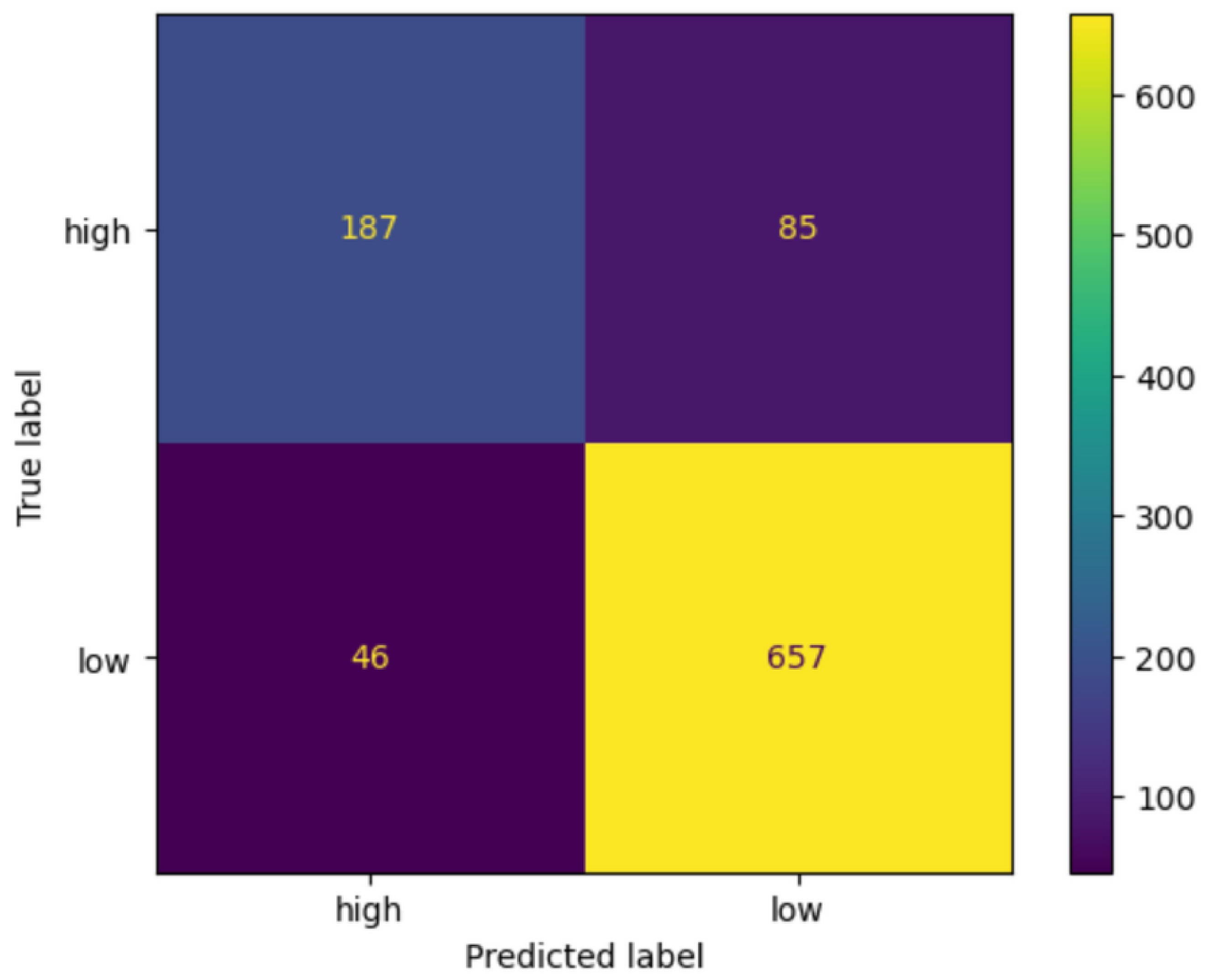


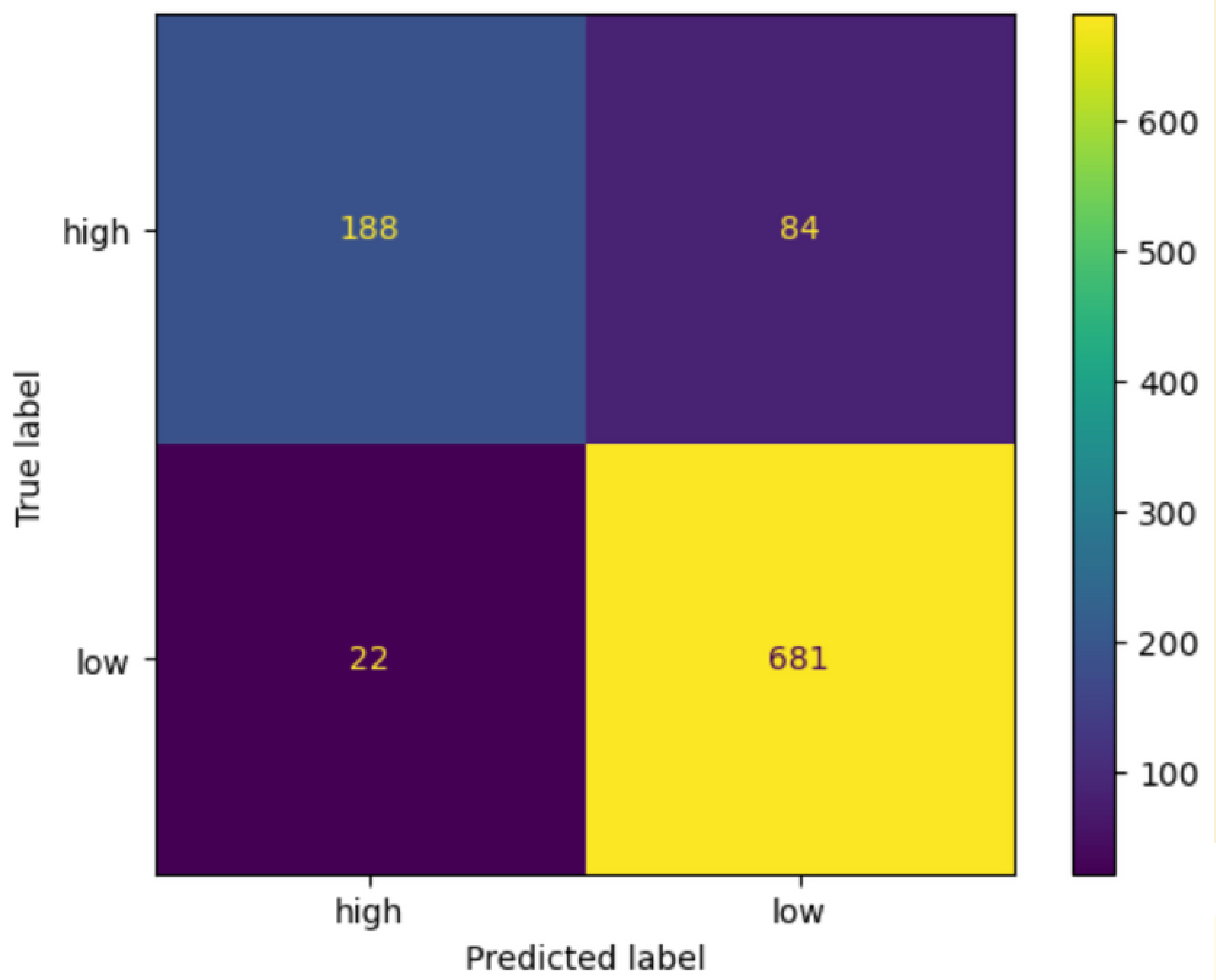
COMPARISON - CONFUSION MATRIX

The GaussianNB model correctly classifies 187 high-value samples and 657 low-value samples. However, it also incorrectly classifies 46 low-value samples as high-value samples and 85 high-value samples as low-value samples.

The Logistic Regression model correctly classifies 188 high-value samples and 681 low-value samples. However, it also incorrectly classifies 22 low-value samples as high-value samples and 84 high-value samples as low-value samples.

Overall, the Logistic Regression model has a slightly higher accuracy than the GaussianNB model. It correctly classifies more high-value samples and low-value samples. However, the difference in accuracy is not very significant.





GAUSSIAN NAIVE BAYES - REPORT

Precision: For the "high" class, the precision is 0.80, indicating that when the model predicts a data point as "high," it is correct 80% of the time. For the "low" class, the precision is 0.89, indicating that when the model predicts a data point as "low," it is correct 89% of the time.

Recall: For the "high" class, the recall is 0.69, meaning that the model correctly identifies 69% of the actual "high" class instances. For the "low" class, the recall is 0.93, meaning that the model correctly identifies 93% of the actual "low" class instances.

F1-Score: The F1-score is the harmonic mean of precision and recall. It gives a balanced measure of the model's accuracy in both precision and recall. For the "high" class, the F1-score is 0.74, and for the "low" class, it is 0.91.

Support: The support indicates the number of actual occurrences of each class in the test data.

	precision	recall	f1-score	support
high	0.80	0.69	0.74	272
low	0.89	0.93	0.91	703
accuracy			0.87	975
macro avg	0.84	0.81	0.82	975
weighted avg	0.86	0.87	0.86	975

LOGISTIC REGRESSION - REPORT

Precision: For the "high" class, the precision is 0.90, indicating that when the model predicts a data point as "high," it is correct 90% of the time. For the "low" class, the precision is 0.89, indicating that when the model predicts a data point as "low," it is correct 89% of the time.

Recall: For the "high" class, the recall is 0.69, meaning that the model correctly identifies 69% of the actual "high" class instances. For the "low" class, the recall is 0.97, meaning that the model correctly identifies 97% of the actual "low" class instances.

F1-Score: The F1-score is the harmonic mean of precision and recall. It gives a balanced measure of the model's accuracy in both precision and recall. For the "high" class, the F1-score is 0.78, and for the "low" class, it is 0.93.

Support: The support indicates the number of actual occurrences of each class in the test data.

	precision	recall	f1-score	support
high	0.90	0.69	0.78	272
low	0.89	0.97	0.93	703
accuracy			0.89	975
macro avg	0.89	0.83	0.85	975
weighted avg	0.89	0.89	0.89	975

SUMMARY

In conclusion, both cat_model1 and cat_model2 were trained to classify co2 emission per person into "high" and "low" categories. Although both models showcased commendable performance, cat_model2 (Logistic Regression) exhibited a marginal performance advantage over cat_model1 (Gaussian Naive Bayes).

For cat_model1, it achieved an accuracy of 0.87, indicating that it correctly predicted the category for 87% of instances. The precision and recall values for both categories were satisfactory, with "low" category prediction showing slightly higher accuracy. The F1-scores for both categories were well-balanced, implying that the model provided a reliable prediction for both classes.

On the other hand, cat_model2 exhibited an improved performance with an accuracy of 0.89. While its precision and recall for the "high" category were commendable, its recall was comparatively lower for this category. However, the "low" category showed exceptional precision and recall values, resulting in a high F1-score and overall accuracy.

Analyzing these two models reveals that all four variables play a pivotal role in constructing meaningful models, underscoring their practical relevance. This underscores the importance of considering these factors in everyday life. To mitigate CO_2 emissions, it's advisable to reduce coal and oil consumption while embracing renewable energy sources like windmill-generated power.

CONCLUSION

Among all the regression models I've constructed, the most effective model for predicting the dependent variable transformed_co2 is co2_model4, which is derived from polynomial regression. For the categorical models, the logistic regression model outperforms the naive bayes model, achieving an accuracy of 89%, in predicting whether co2_emit_per_person is categorized as high or low. Both the regression and categorical models have underscored the significance of all four variables in relation to the response variables transformed_co2 and co2_emit_per_person. Notably, the models reveal that increased coal and oil consumption, along with higher levels of urbanization, correspond to elevated co2 emissions. Conversely, greater adoption of renewable energy is associated with decreased co2 emissions. Consequently, governmental efforts should prioritize the promotion of clean and renewable energy sources, such as nuclear and wind energy, while also imposing limits on the use of fossil fuels and advocating for reduced air conditioning and car usage in urban areas, or promote the usage of electic electc cars cars . These measures collectively serve to mitigate the impact of urbanization on co2 emissions.

Thank you for listening

Any questions?