Regression Project

Aim: Analyse and predict the impact of CO_2 emissions from the agri-food sector as well as related sectors on climate change, as measured by average temperature rise, and develop strategies for sustainable practices using FAO and IPCC datasets.

For this project, we have concentrated on a single country: France.

Presentation Structure

- Exploring The Data
- Regression Analysis & Model Comparison
- Recommendation & Conclusion

Exploring The Data

THE DATA

- The data consisted of emission and temperature data has been obtained from the Food and Agriculture Organization of the United Nations' (FAO) website (Emissions CO2 eq AR5). These have been chosen to span from 1990 to 2020 (i.e. 31 years of data)
- Populations data has been obtained from the World bank Group's data bank. These have been matched by country and by year to the FAO data.
- The data was filtered for the country of interest: France and was checked for null values.
- Fields which had all values as zero were dropped (these include: N*et Forest Conversion, Fires In Organic soils* and *Fires in Humid Tropical Forests*).
- The next slides provides a view of the fields used after exclusion. The following process is followed in the next slides:

THE DATA - FIELDS IN THE DATASET

The target variable is ***Average Temperature °C*** which is the average temperature rise in any particular country over the previous year and measured in degrees celsius.

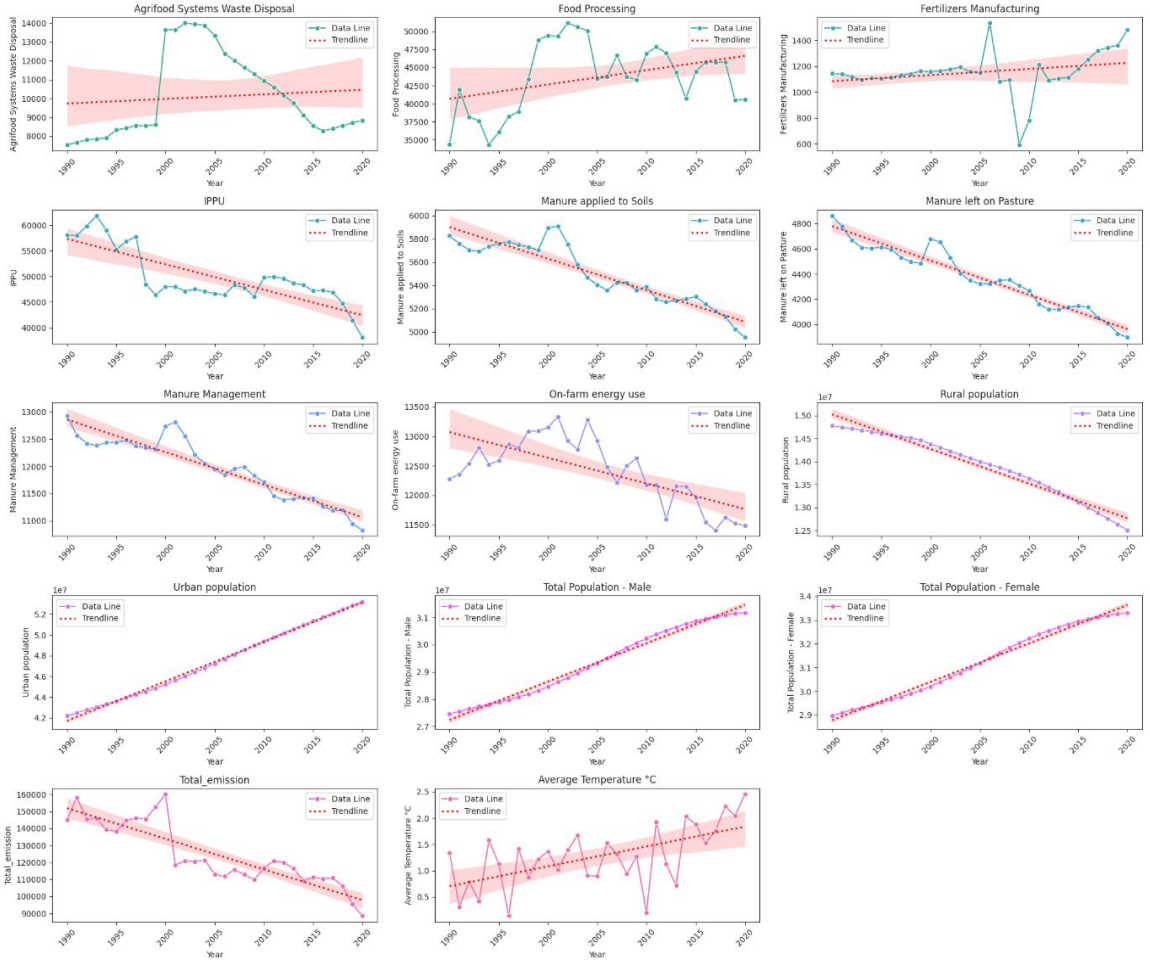
Field	Description	Units
Agrifood Systems Waste Disposal	Emissions from waste disposal in the agrifood system	kt CO_2
Area	Country/Region for which data is applicable	N/A
Average Temperature °C	Average temperature rise in land temperature over the previous year	°C
Crop Residues	Emissions from burning or decomposing leftover plant material after crop harvesting	kt CO_2
Drained Organic Soils (CO2)	Emissions from CO_2 released when draining organic soils	kt CO_2
Fertilizers Manufacturing	Emissions from the production of fertilizers	kt CO_2
Forest Fires	Emissions from fires in forested areas	kt CO_2
Forestland	Land covered by forests	kt CO_2
Food Household Consumption	Emissions from food consumption at the household level	kt CO_2
Food Packaging	Emissions from the production & disposal of food packaging materials	kt CO_2
Food Processing	Emissions from processing food products	kt CO_2
Food Retail	Emissions from the operation of retail establishments selling food	kt CO_2
Food Transport	Emissions from transporting food products	kt CO_2
IPPU	Emissions from industrial processes & product use	kt CO_2
Manure Applied to Soils	Emissions from applying animal manure to agricultural soils	kt CO_2
Manure Left on Pasture	Emissions from animal manure on pasture or grazing land	kt CO_2

THE DATA - FIELDS IN THE DATASET (continued)

Field	Description	Unit
Manure Management	Emissions from managing and treating animal manure	kt CO_2
On-farm Electricity Use	Electricity consumption on farms	kt CO_2
On-farm Energy Use	Energy consumption on farms	kt CO_2
Pesticides Manufacturing	Emissions from the production of pesticides	kt CO_2
Rice Cultivation	Emissions from methane released during rice cultivation	kt CO_2
Rural Population	Number of people living in rural areas	count
Savanna Fires	Emissions from fires in savanna ecosystems	kt CO_2
Total Emission	Total greenhouse gas emissions from various sources	kt CO_2
Total Population - Female	Total number of female individuals in the population	count
Total Population - Male	Total number of male individuals in the population	count
Urban Population	Number of people living in urban areas	count
Year	Year in which observation was made	N/A

The next slides show the evolution of these variables over the period considered.

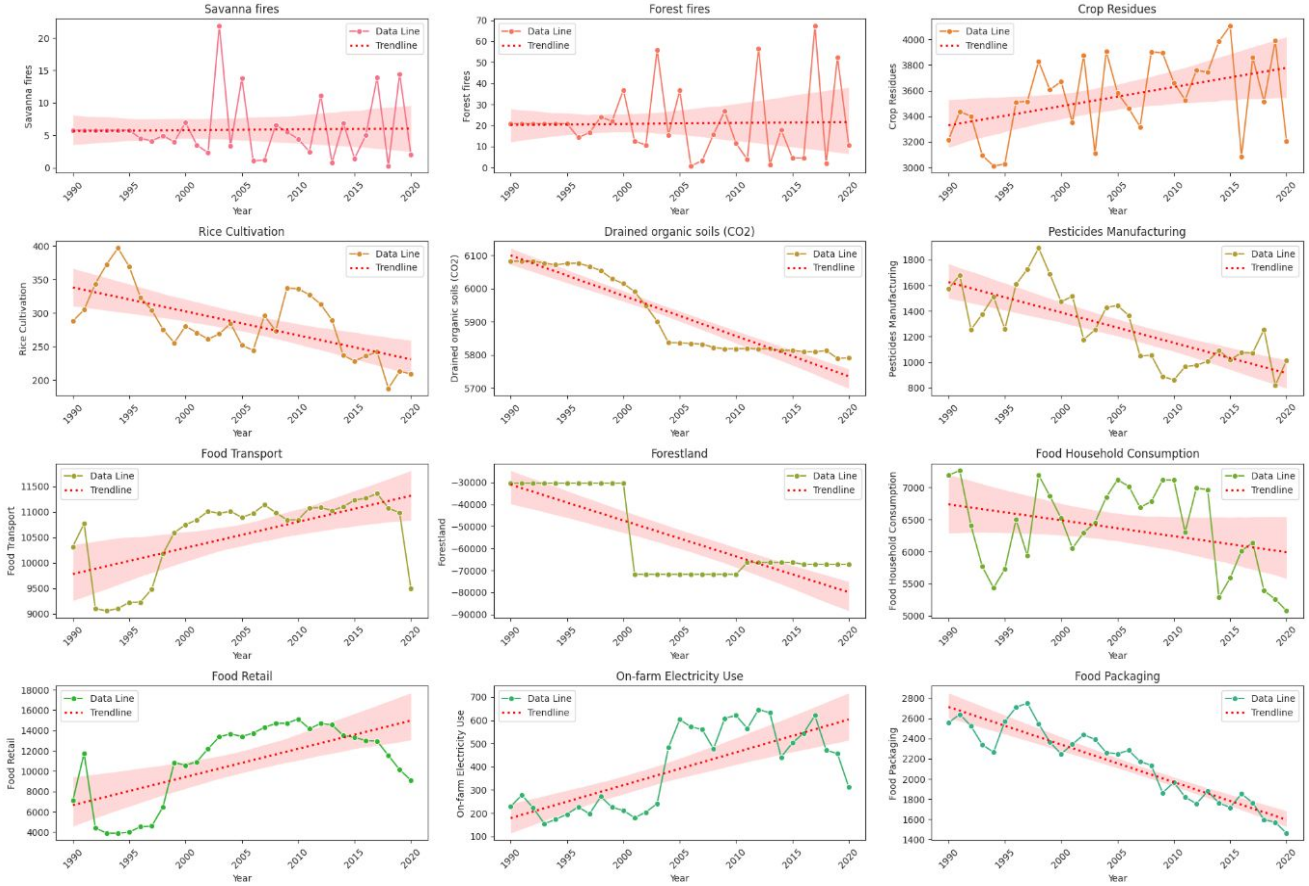
EVOLUTION OF THE VARIABLES OVER TIME



Observations

- The **average rise in temperature per year** over the period 1990 to 2020 has been increasing (i.e. temperature is rising at an increasing rate).
- ***Total emission*** has been decreasing in general over the period.
- The ***populations of male, female*** have been increasing in France, with more people seeming to choose to live in *urban* areas than ***rural*** areas.
- Emissions from ***On-farm energy*** use has been decreasing in general, after peaking in the period 2000-2005.
- In general, emissions from ***Manure Management***, ***Manure applied to Soils*** and ***Manure left on Pasture*** have been falling over the period considered. This can be linked to the shift which has occured over the years, with farms now preferring fertilizers to manure. Traditionally, manure was the primary source of nutrients for crops. With the development of commercial fertilizers, the reliance on manure decreased.
- Similarly, emissions from ***IPPU*** sector has been falling over the period - this could perhaps point towards more efficient and/or sustainable measures being adopted in industrial processes.
- Emissions from ***Fertilizers Manufacturing***, ***Food Processing***, ***on-farm Electricity Use***, ***Food Retail*** (on next slide), ***Food Transport*** (on next slide)and ***Crop Residues*** (on next slide) have been increasing.
- Emissions from ***Agrifood Systems Waste Disposal*** has increased and peak in the period 2000 to 2005 before starting to fall in susbsequent years

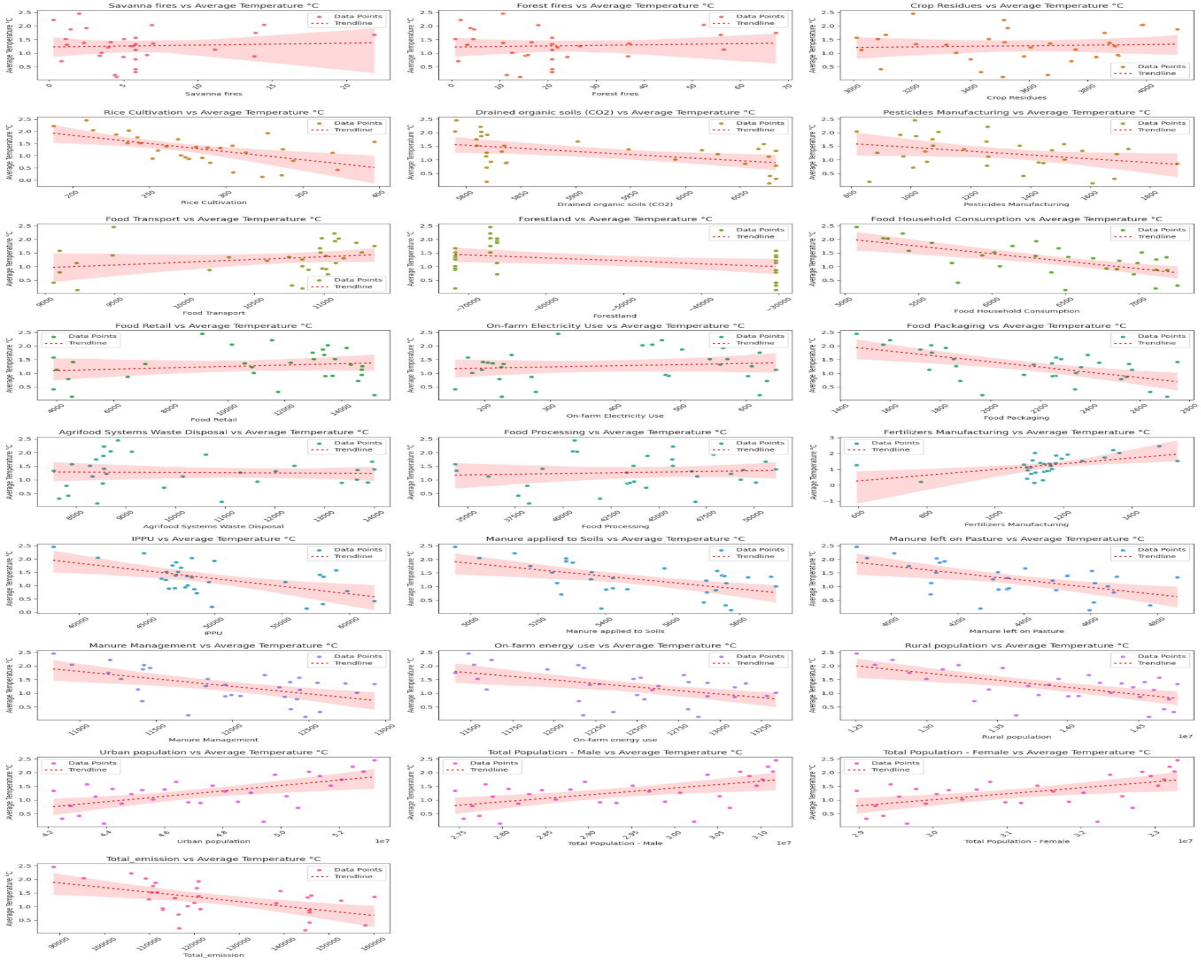
EVOLUTION OF THE VARIABLES OVER TIME (continued)



Observations

- To note that the 'carbon-sink' effect of ***Forestland*** has also been on the rise over time (it has become negative over the period).
- Emissions from ***Rice Cultivation***, ***Drained organic soils (CO2)***, ***Pesticides Manufacturing*** and ***Food Household Consumption*** show decreasing trends.
- Emissions from ***Savanna fires***, which are relatively low, and ***Forest Fires*** do not show any trend, this is expected as they are occurences of nature and not a planned process; they are sporadic.

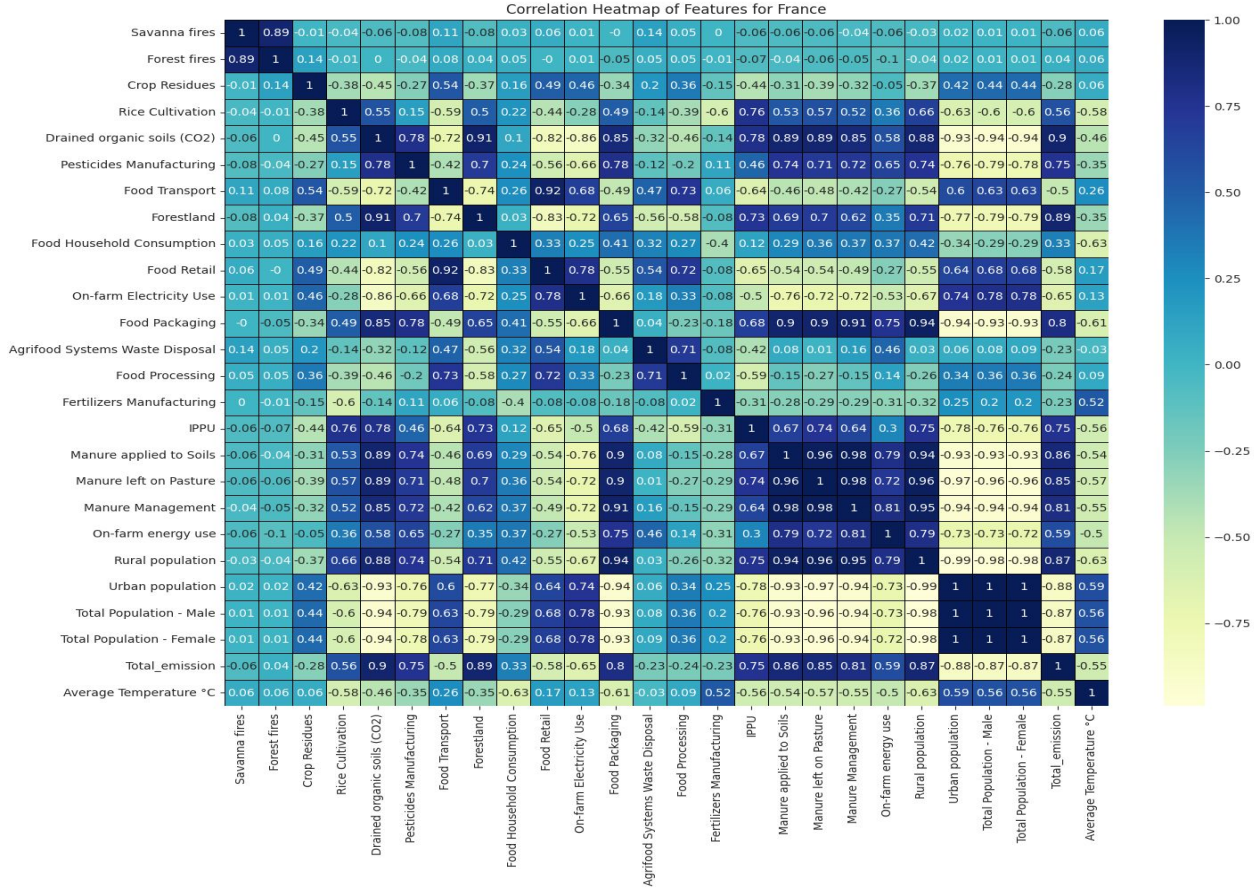
RELATIONSHIP WITH THE TARGET VARIABLE



Observations

- For certain such as ***Fertilizers Manufacturing***, the data points are clustered around a certain value, with a few outliers.
- For ***Forestland*** the observations are clustered around very high or very low values.
- At first glance, most of the relationships seem somewhat linear which does not suggest the need to transform the variables before modelling.

CORRELATION BETWEEN THE VARIABLES



The correlations show most notably:

- Certain factors such as ***Savanna fires***, ***Forest Fires***, ***Crop Residues***, ***Agrifood Systems Waste Disposal*** and ***Food Processing*** have a very low (absolute) correlation with ***Average Temperature °C***.
- Relatively, ***Urban population***, ***Total Population - Male***, ***Total Population - Female*** & F***ertilizers Manufacturing*** have the highest positive correlation with ***Average Temperature °C***.
- On the other hand, ***Rural population***, ***Food Household Consumption***, ***Food Packaging***, ***Rice Cultivation*** have a relatively strong negative correlation with ***Average Temperature °C***.
- Important to note that certain factors show strong correlation with each other; this could cause multicollinearity down the line.

Top 5 Absolute Correlations

Factor	Absolute Correlation Coefficient with *Average Temperature*
Food Household Consumption	0.63
Rural Population	0.63
Food Packaging	0.61
Urban Population	0.59
Rice Cultivation	0.58

Regression Analysis & Model Comparison

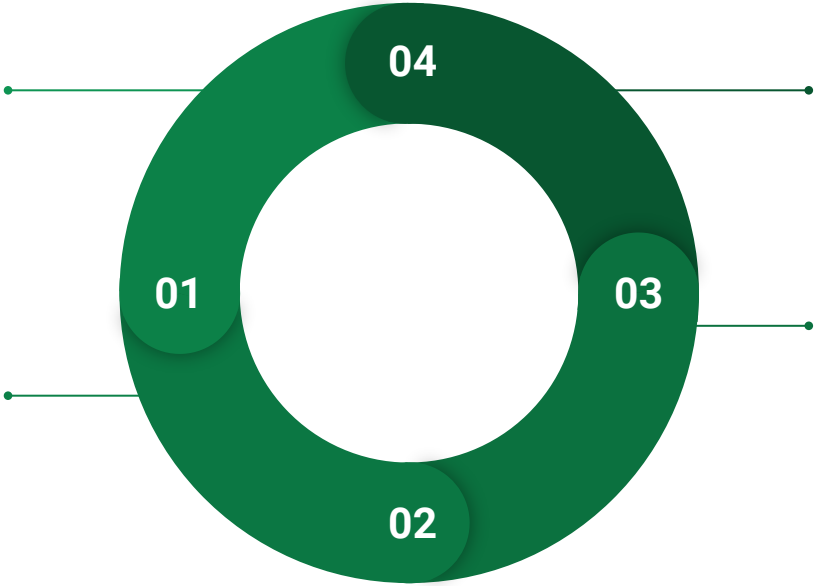
METHODOLOGY APPLIED

- Below is described the process followed for the regression analysis.

Check For Multicollinearity
The independent variables are checked for multicollinearity using the Variance Inflation Factor (VIF).

Select Variables & Models
Select variables for inclusion in the regression model based on the VIF and absolute correlation with the **Average Temperature** (target variable).

Variables are dropped if they produce coefficients with p-values less than 5% (not statistically significant).



Add (Back) Additional Variables & Compare Performance
Using the best model in step 4 , add back variables and compare and contrast the results.

Model Diagnostic
Diagnose the best model after dropping statistically insignificant coefficients based on statistical tests and other metrics.

CHECKING FOR MULTICOLLINEARITY

- The VIF was tested using all (independent) variables at first, this produced very high values.
- Generally for the VIF:

VIF	Interpretation
< 5	Generally acceptable, low multicollinearity.
5 < VIF < 10	Moderate multicollinearity; some caution is needed.
> 10	High multicollinearity; consider removing or combining these features.

- Based on the previous correlation results, the factors which had low correlation (less than 0.20) with ***Average Temperature*** were dropped.
- Given that the VIF was still high after dropping the factors, the remaining factors were removed successively (with the one with the highest VIF first).
- This was done until all the VIFs for each factor remaining was less than 5. This left us with ***Forestland*** and ***Pesticides Manufacturing*** as independent variables.

Python Output

```
                      feature         VIF
0  Food Household Consumption  118.160885
1            Rice Cultivation   35.276616
2   Fertilizers Manufacturing   43.467718
3                  Forestland   23.962857
4    Pesticides Manufacturing   54.807393
Dropping feature: Food Household Consumption with VIF: 118.16088474576448
Updated VIF Data:
                      feature         VIF
0           Rice Cultivation  20.208372
1  Fertilizers Manufacturing  42.737880
2                  Forestland  12.465461
3   Pesticides Manufacturing  32.580025
Dropping feature: Fertilizers Manufacturing with VIF: 42.73788005209956
Updated VIF Data:
                     feature         VIF
0          Rice Cultivation  20.200187
1                 Forestland   5.598495
2  Pesticides Manufacturing  15.526882
Dropping feature: Rice Cultivation with VIF: 20.200186805769818
Updated VIF Data:
                     feature       VIF
0                 Forestland  4.303195
1  Pesticides Manufacturing  4.303195
Final VIF Data:
                     feature       VIF
0                 Forestland  4.303195
1  Pesticides Manufacturing  4.303195
Selected Features for Regression:
 Index(['Forestland', 'Pesticides Manufacturing'], dtype='object')
```

VARIABLE & MODEL SELECTION

Model 1

- Despite the results from the VIF, we started fitting regression models inclusive of a few more variables (***Forestland***, ***Pesticides Manufacturing***, ***Food Household Consumption***, ***Rice Cultivation*** & ***Fertilizers Manufacturing***). To note that the values were standardised/scaled before applying the regression model, in each case.
- A regression model was fit (using Ordinary Least Squares) with the target variable being (***Average Temperature***).
- Many of the coefficients were not statistically significant; their p-values were were higher than 0.05.

Python Output

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.2731	0.088	14.469	0.000	1.088	1.458
Forestland	-0.0169	0.160	-0.106	0.917	-0.354	0.320
Q('Pesticides Manufacturing')	-0.1526	0.138	-1.108	0.282	-0.442	0.137
Q('Food Household Consumption')	-0.2514	0.101	-2.487	0.023	-0.464	-0.039
Q('Rice Cultivation')	-0.1918	0.143	-1.343	0.196	-0.492	0.108
Q('Fertilizers Manufacturing')	0.1268	0.126	1.009	0.326	-0.137	0.391

Model 2

- ***Forestland*** was dropped and the model refit (it had the highest p-value). Certain coefficients were still not statistically significant.

Python Output

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.2731	0.086	14.861	0.000	1.094	1.452
Q('Pesticides Manufacturing')	-0.1626	0.097	-1.677	0.110	-0.366	0.040
Q('Food Household Consumption')	-0.2482	0.094	-2.644	0.016	-0.445	-0.052
Q('Rice Cultivation')	-0.2005	0.113	-1.772	0.092	-0.437	0.036
Q('Fertilizers Manufacturing')	0.1264	0.122	1.034	0.314	-0.129	0.382

VARIABLE & MODEL SELECTION (continued)

Model 3

- The process is repeated and ***Fertilizers Manufacturing*** is dropped next.
- A few coefficients were not still statistically significant.
- To note that the test Adjusted R2 it is positive for the first time here, showing that the model now has some predictive power.
- For the previous models, this value was negative.

Python Output

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.1031	0.947	6.445	0.000	4.128	8.078
Q('Pesticides Manufacturing')	-0.0005	0.000	-1.358	0.190	-0.001	0.000
Q('Food Household Consumption')	-0.0004	0.000	-3.192	0.005	-0.001	-0.000
Q('Rice Cultivation')	-0.0052	0.002	-3.048	0.006	-0.009	-0.002

Model 4

- ***Pesticides Manufacturing*** was dropped next and this model produced statistically significant parameters.
- The test Adjusted R2 is 0.76 for this model.
- So far this is the best model.

Python Output

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.2731	0.088	14.547	0.000	1.091	1.455
Q('Food Household Consumption')	-0.2839	0.090	-3.155	0.005	-0.471	-0.097
Q('Rice Cultivation')	-0.2921	0.090	-3.246	0.004	-0.479	-0.105

MODEL DIAGNOSTICS

- We proceed with analysing the full set of results for Model 4 (which includes ***Food Household Consumption*** & **Rice Cultivation** as independent variables.

Python Output

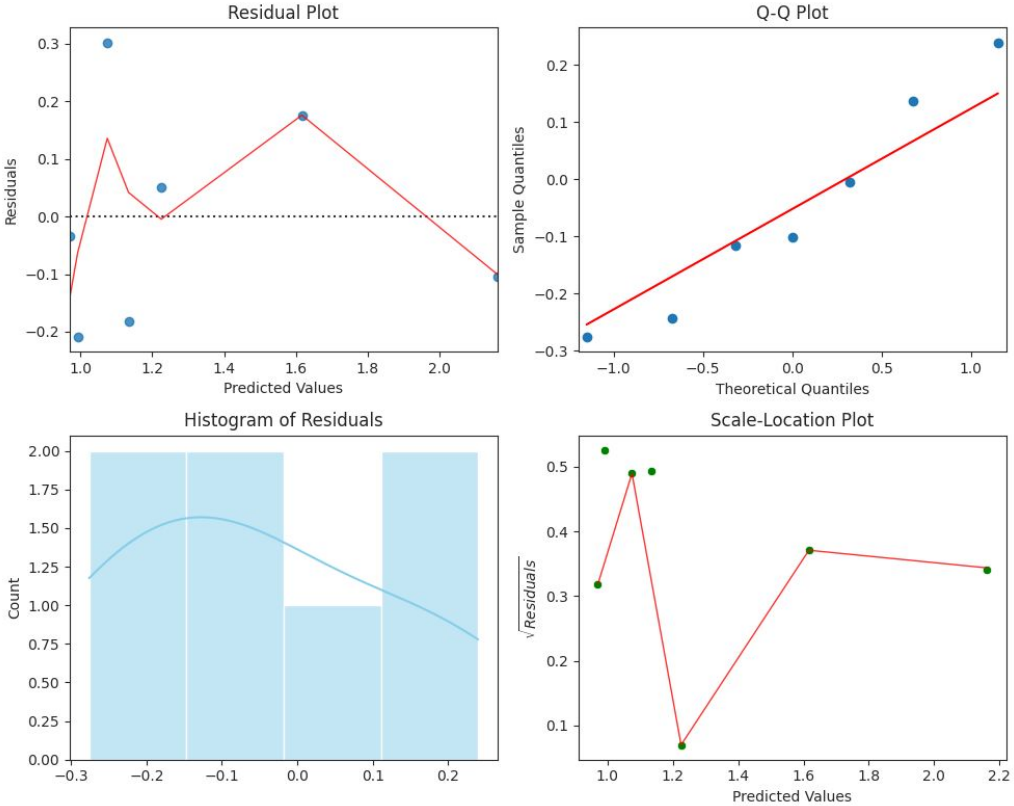
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q("Average Temperature °C")   R-squared:                       0.560
Model:                                    OLS   Adj. R-squared:                  0.518
Method:                         Least Squares   F-statistic:                     13.35
Date:                        Mon, 11 Nov 2024   Prob (F-statistic):           0.000181
Time:                                04:24:16   Log-Likelihood:                -12.127
No. Observations:                          24   AIC:                             30.25
Df Residuals:                              21   BIC:                             33.79
Df Model:                                   2
Covariance Type:                    nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                        1.2731      0.088     14.547      0.000       1.091       1.455
Q('Food Household Consumption')  -0.2839     0.090     -3.155      0.005      -0.471      -0.097
Q('Rice Cultivation')            -0.2921     0.090     -3.246      0.004      -0.479      -0.105
==============================================================================
Omnibus:                        0.047   Durbin-Watson:                   2.481
Prob(Omnibus):                  0.977   Jarque-Bera (JB):                0.083
Skew:                          -0.047   Prob(JB):                        0.959
Kurtosis:                       2.727   Cond. No.                         1.27
==============================================================================
Linear Regression Equation:
Average Temperature °C = 1.27 + (-0.28 * Food_Household_Consumption) + (-0.29 * Rice_Cultivation)
Train MSE: 0.16, Train RMSE: 0.40
Test MSE: 0.03, Test RMSE: 0.18
Train Adjusted R-squared: 0.52
Test Adjusted R-squared: 0.76
```

Key takeaways:

- **Significant Predictors:** Both ***Food Household Consumption*** and ***Rice Cultivation*** are statistically significant predictors of ***Average Temperature °C*** with p-values less than 0.05.
- **Interpretation:** The intercept (1.27) represents the baseline value of the ***average temperature*** when both ***Food Household Consumption*** and ***Rice Cultivation*** are zero. Since the data was scaled before the regression, this value pertains to the standardised data.
- ***Food Household Consumption:*** The coefficient (-0.28) indicates that for each unit increase in standardised ***Food Household Consumption***, the ***average temperature*** change decreases by 0.28 units, holding ***Rice Cultivation*** constant.
- ***Rice Cultivation:*** The coefficient (-0.29) shows that for each unit increase in standardised ***Rice Cultivation***, the ***average temperature*** change decreases by 0.29 units, holding ***Food Household Consumption*** constant.
- **Model Fit:** The model shows a good fit, especially on the test set, with a high adjusted R-squared value of 0.76, indicating that a significant portion of the variability in ***Average Temperature °C*** is explained by the model for the test data. The variability explained by the model for the training data stands at 0.52 (52%)

MODEL DIAGNOSTICS (continued)

Python Output



Model Diagnostics (statistical tests):

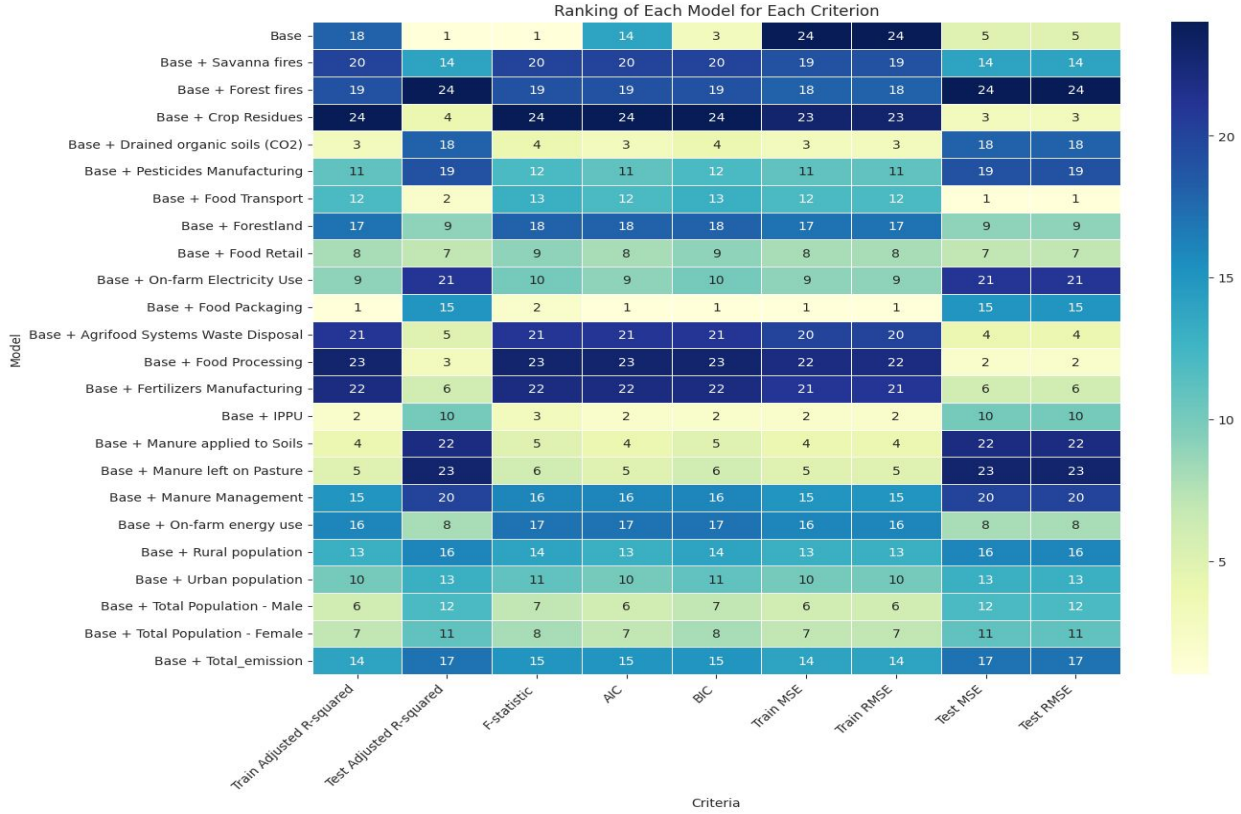
- **Durbin-Watson Statistic:** 2.481 (indicating no significant autocorrelation in the residuals, ideal value should be close to 2)
- **Omnibus Test**: Prob > chi-squared = 0.977 (indicating normal distribution of residuals)
- **Jarque-Bera Test:** Prob > chi-squared = 0.959 (supporting the normality of residuals)
- **Kurtosis:** 2.727. This is close to 3. This suggests that the residuals of the model are approximately normally distributed, indicating a good fit of the model to the data with a reasonable amount of outliers.
- **Condition Number:** 1.27 (indicating no significant multicollinearity) and which has also been previously supported by the VIF.

Residual Analysis:

- Diagnostic plots (residual plots, Q-Q plots, histogram of residuals, and scale-location plots) suggest that the residuals are well-behaved, with no evident patterns indicating potential issues with the model.
- For the residuals plot the, points plot roughly equally above and below the zero line.
- For the QQ plot, the points do not deviate significantly from the red line and more or less evenly distributed on both sides of it, exhibiting normality to a certain degree.
- The histogram of residuals does show that the values are close to zero although a slight skew is present.
- It is slightly more difficult to interprete this scale-location plot given the lack of points. Comfort is however taken from the below diagnostic tests that the residuals are not heteroskedastic.

ADDING BACK VARIABLES

- Using the previous best model as **Base,** additional variables are added to see if any model improvement can be obtained.
- **In each case, the models did not produce statistically significant coefficients at 5% level of significance.**
- The models were also ranked on several criteria as per below>
- The observations are summarised on the next slide



ADDING BACK VARIABLES (continued)

- The following observations were made and the **Base** model was still deemed as the best.

Metric	Explanation	General Comments	Observations
Train Adjusted R-squared	Proportion of variance explained by the model on the training data, adjusted for the number of predictors.	Higher values indicate better fit on training data. Reflects model performance on the data it was trained on.	The *base* model performs poorly here with a rank of 18th. This indicates underfitting to the test data compared to the other models.
Test Adjusted R-squared	Proportion of variance explained by the model on the testing data, adjusted for the number of predictors.	Higher values indicate better generalization to new data. Reflects model's ability to generalise to unseen data.	The *base* model has the best performance here, demonstrating that it might have the best predictiive power.
F-statistic	Measure of the overall significance of the model.	Higher values indicate a more significant model. Tests if the overall regression model is a good fit for the data and here if in combination the coefficients are statistically significant.	The *base* model has the performance best here which is also in line with the low p-values obtained for its coefficients.
AIC (Akaike Information Criterion)	Measure of the relative quality of the model, with lower values indicating better fit.	Lower values are better, balancing model fit and complexity. Indicates how parsimonious the model is, penalizing for complexity.	The *base* is ranked 14th here; it is not as parsimonious as other models.
BIC (Bayesian Information Criterion)	Similar to AIC but with a stronger penalty for models with more parameters.	Lower values indicate a better model considering both fit and parsimony. Penalises more heavily for additional predictors, preferring simpler models.	The *base* model performs better here (relative to the AIC) implying that the other models are more heavily penalised for their complexity brought about by their additional terms.
Train MSE (Mean Squared Error)	Average squared difference between observed and predicted values on the training data.	Lower values indicate better fit on training data. Measures the average squared error on the training data.	The *base* model has the worst performance in this criteria. This has been also by the poor train adjusted R^2
Train RMSE (Root Mean Squared Error)	Square root of MSE, providing error magnitude on the training data.	Lower values indicate smaller errors on training data. Provides an easy-to-interpret measure of error magnitude.	The *base* model has the worst performance in this criteria.
Test MSE (Mean Squared Error)	Average squared difference between observed and predicted values on the testing data.	Lower values indicate better generalisation to new data. Measures the average squared error on the testing data	The *base* model is ranked 5th in this criteria showing that the model's predictions are closer to the actual values in the test dataset, suggesting better accuracy.
Test RMSE (Root Mean Squared Error)	Square root of MSE, providing error magnitude on the testing data.	Lower values indicate smaller errors on testing data. Provides an easy-to-interpret measure of error magnitude on new data.	Similar to the Test MSE.

Recommendation & Conclusion

RECOMMENDATION & CONCLUSION

Interpretation

- The regression results suggest that increases in ***Food Household Consumption*** and ***Rice Cultivation*** are associated with decreases in the rise in the yearly average temperature in France within the context of your dataset. However, any real-world implications should be interpreted cautiously and within the broader context of climate science.
- This appears counterintuitive as one might expect that increasing ***Rice Cultivation*** and ***Food Household Consumption*** would increase CO2 emissions which would in turn cause average temperatures rises to increase. However, the unique agricultural and consumption practices in France may offer some insights into these findings.
- One possible explanation lies in the way rice is cultivated and how ***Food Household Consumptio***n are in France.
- Most of the rice cultivated in France originates from the Camargue Region. In 2021, rice cultivation covered an area of 11,800 hectares, producing around 70,000 tons of rice (a figure which has been rising over the years). There are close to 200 rice farmers operating in the area, 90% of which dedicated to the cultivation of PGI Camargue Rice. The rice is produced both conventionally and organically, with as much as 25% of the area devoted to organic agriculture. The farmers are committed to the 'Rice Farmers Charter for Environmental Protection and Production Quality', emphasizing sustainable practices.
- Similarly, the emissions from ***Food Household Consumption*** have been decreasing over the years, despite a rising population. This trend points towards improved and sustainable practices in food consumption. Households are likely adopting more environmentally friendly habits, such as reducing food waste, choosing sustainable products, and supporting local produce.
- By promoting such practices, increases in these two factors might lead to lesser increases in average temperature in France over the years. However, it is essential to approach these findings with caution and conduct further research to validate these observations; i.e. take the results with a grain of salt.

Based on the results however, the recommendations are:

Recommendations for Policymakers

- **Promote Sustainable Household Consumption:**
  Given that increased Food Household Consumption is associated with a decrease in average temperature rises in the dataset, encourage sustainable practices in household food consumption. This includes promoting consumption of locally sourced and seasonal foods to reduce the carbon footprint

  Launching public awareness campaigns to educate consumers about the environmental impacts of their food choices and encouraging them to adopt more sustainable habits is also recommended.

- **Support Eco-friendly Rice Cultivation:**
  Since Rice Cultivation is also negatively associated with average temperature rises, policies that support eco-friendly rice farming techniques could be beneficial. Consider encouraging the adoption of System of Rice Intensification (SRI) methods, which use less water and reduce methane emissions.

- **Further Research:**
  Extend the research to other countries within the same region to see if similar patterns can be uncovered. This will help in understanding whether these findings are unique to France or if they can be observed more broadly.
  Use more powerful regression techniques to better address multicollinearity (such as Principal Component Analysis) and test out additional models.
  Ensure continuous data collection and analysis to refine policies based on real-world outcomes, ensuring they remain effective and relevant over time.

Conclusion:

These recommendations aim to leverage the insights from your regression analysis to promote sustainable practices and effective climate action policies. By integrating these findings into policy decisions, policymakers can contribute to both climate mitigation and sustainable development. However, the results should still be taken with a grain of salt, and further research is essential to fully understand the broader implications.


Thank You!