Regression Project:
Predicting Average Temperature in the Agri-Food Sector

Anton du Plooy

Sept 2025

Contents

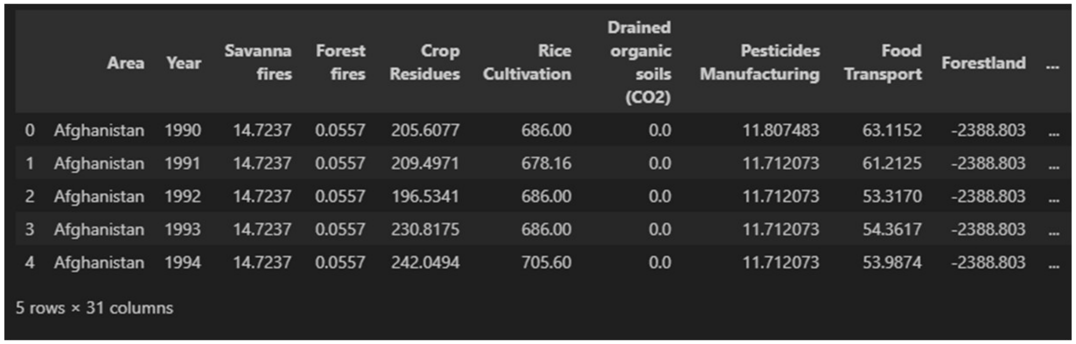
- Project Objectives
- Dataset Overview
- Data Cleaning
- Exploratory Data Analysis
- Models Tested
- Model Performance
- Predicted vs Actual
- Feature Importance
- Key Insights
- Recommendations

Project Objectives

- Analyse factors influencing **average temperature** using regression.
- Apply regression techniques in **Python** (Linear, Ridge, Lasso).
- Develop collaboration & reproducibility (GitHub, Trello, Jupyter).
  - https://github.com/ADP777/2401PTDS_Regression_Project_ADP
  - Trello – not used as individual project
  - C:\Users\aduplooy\OneDrive - Ninety One\Documents\GitHub\2401PTDS_Regression_Project_ADP
- Deliver insights & presentation-ready results.

Dataset Overview

- Source: FAO & IPCC climate/agriculture data.
- Records: ~7,000 (1990–2020).
- Features: 30+ (agriculture, population, emissions, country).
- Target: Average Temperature (°C)



	Area	Year	Savanna fires	Forest fires	Crop Residues	Rice Cultivation	Drained organic soils (CO2)	Pesticides Manufacturing	Food Transport	Forestland	...
0	Afghanistan	1990	14.7237	0.0557	205.6077	686.00	0.0	11.807483	63.1152	-2388.803	...
1	Afghanistan	1991	14.7237	0.0557	209.4971	678.16	0.0	11.712073	61.2125	-2388.803	...
2	Afghanistan	1992	14.7237	0.0557	196.5341	686.00	0.0	11.712073	53.3170	-2388.803	...
3	Afghanistan	1993	14.7237	0.0557	230.8175	686.00	0.0	11.712073	54.3617	-2388.803	...
4	Afghanistan	1994	14.7237	0.0557	242.0494	705.60	0.0	11.712073	53.9874	-2388.803	...

5 rows × 31 columns

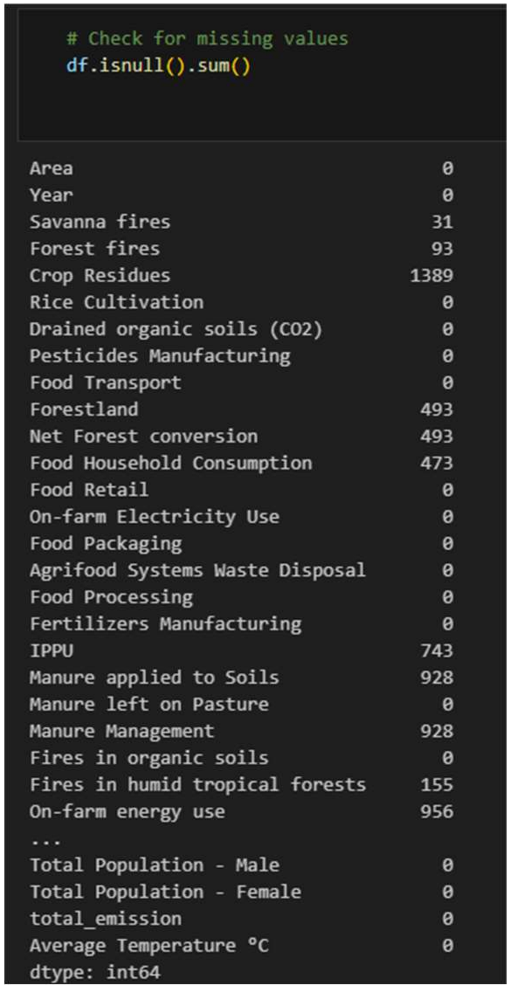
Data Cleaning

- Checked missing values (0–20%).

- Dropped features with high missingness & weak correlation: *Crop Residues, Manure Mgmt, Forestland, etc.*

- Imputed remaining missing values (median)

- Created cleaned dataset
  - co2_emissions_from_agri_clean.csv.



```
# Check for missing values
df.isnull().sum()
```

```
Area                               0
Year                               0
Savanna fires                     31
Forest fires                      93
Crop Residues                   1389
Rice Cultivation                   0
Drained organic soils (CO2)        0
Pesticides Manufacturing           0
Food Transport                     0
Forestland                       493
Net Forest conversion            493
Food Household Consumption       473
Food Retail                        0
On-farm Electricity Use            0
Food Packaging                     0
Agrifood Systems Waste Disposal    0
Food Processing                    0
Fertilizers Manufacturing          0
IPPU                             743
Manure applied to Soils          928
Manure left on Pasture             0
Manure Management                928
Fires in organic soils             0
Fires in humid tropical forests  155
On-farm energy use               956
...
Total Population - Male            0
Total Population - Female          0
total_emission                     0
Average Temperature °C             0
dtype: int64
```

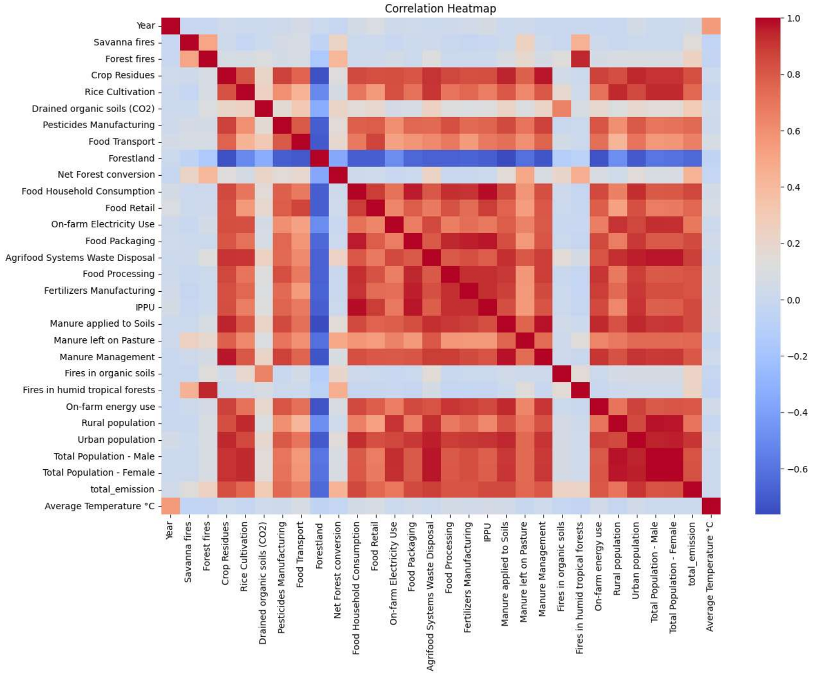
Exploratory Data Analysis

- Distributions: Most features right-skewed, population extremely large.

- Correlations:
  - Strongest: Year vs Temperature (r ≈ 0.55).
  - Most other features weakly correlated.

- Heatmap: Clear multicollinearity among agricultural & population variables.

- Scatterplots: Upward trend over time; weak drifts for agri features.

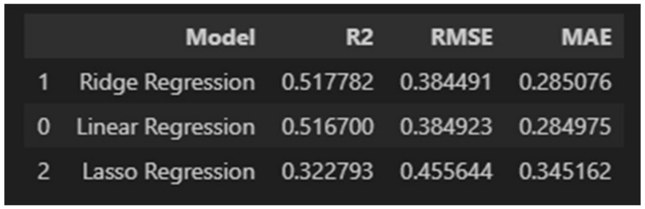


Models Tested

- Linear Regression – baseline.
- Ridge Regression – handles multicollinearity
- Lasso Regression – feature selection.

Model Performance

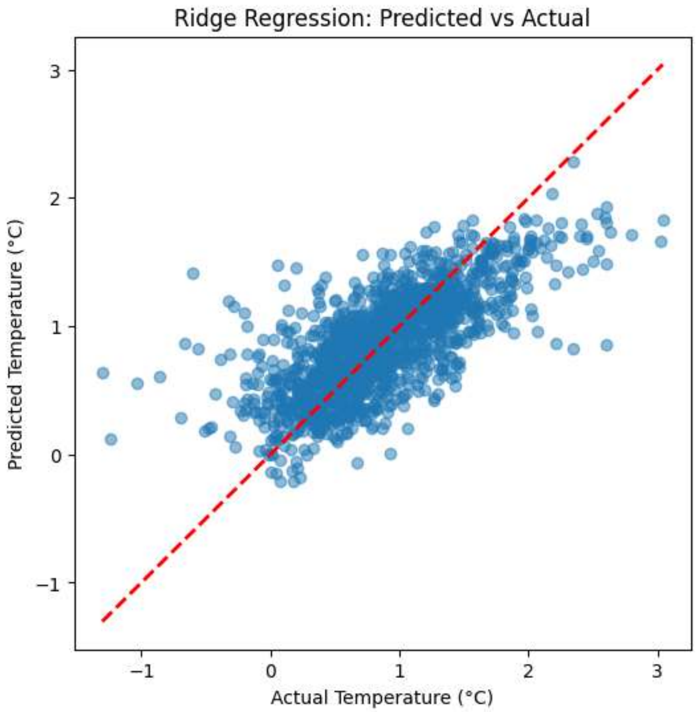
- Linear Regression: $R^2 \approx 0.52$, RMSE ≈ 0.38
- Ridge Regression: $R^2 \approx 0.52$, RMSE ≈ 0.38 (best performer)
- Lasso Regression: $R^2 \approx 0.32$, RMSE ≈ 0.46 (underperformed)
- Residuals: Normal, centered at 0 → good fit assumptions.



	Model	R2	RMSE	MAE
1	Ridge Regression	0.517782	0.384491	0.285076
0	Linear Regression	0.516700	0.384923	0.284975
2	Lasso Regression	0.322793	0.455644	0.345162

Predicted vs Actual

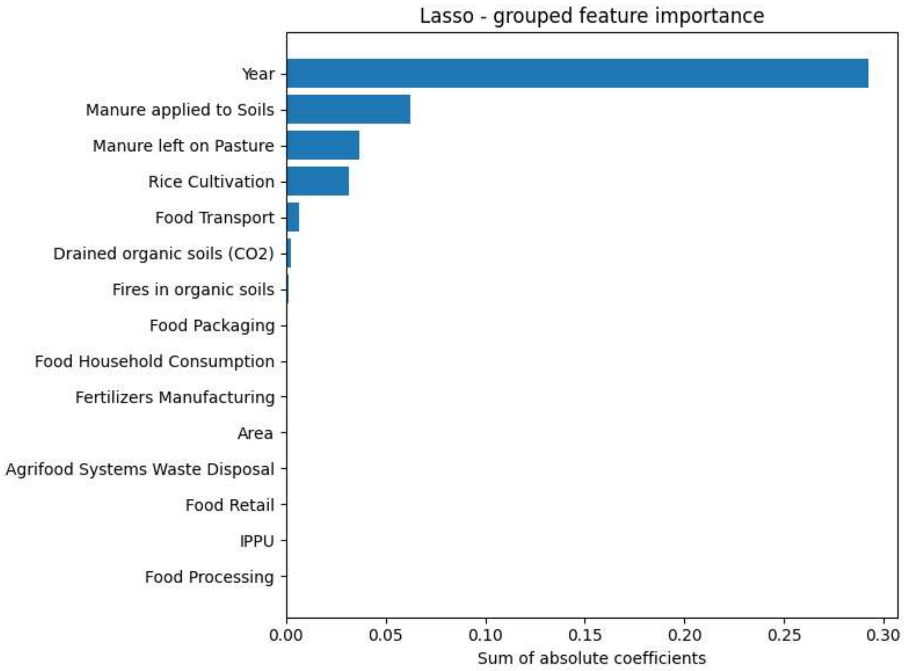
- Linear & Ridge ⟶ follow 45° line, small bias at extremes
- Lasso ⟶ underestimates variation.



Feature Importance

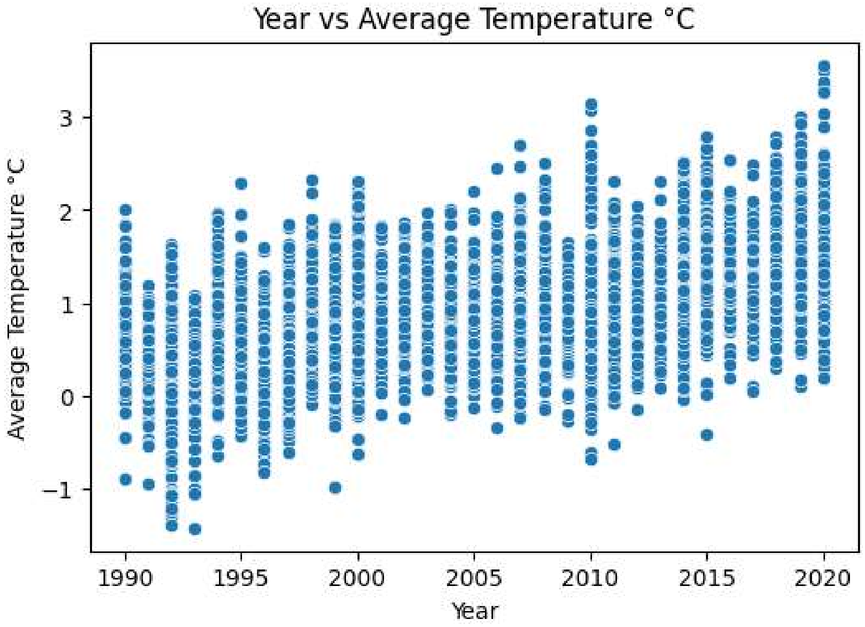
- Ridge: Area dominates (country effect), then Year & weak agri features

- Lasso: Year dominates, small contributions from:
  - Manure applied to Soils
  - Rice Cultivation
  - Food Transport



Key Insights

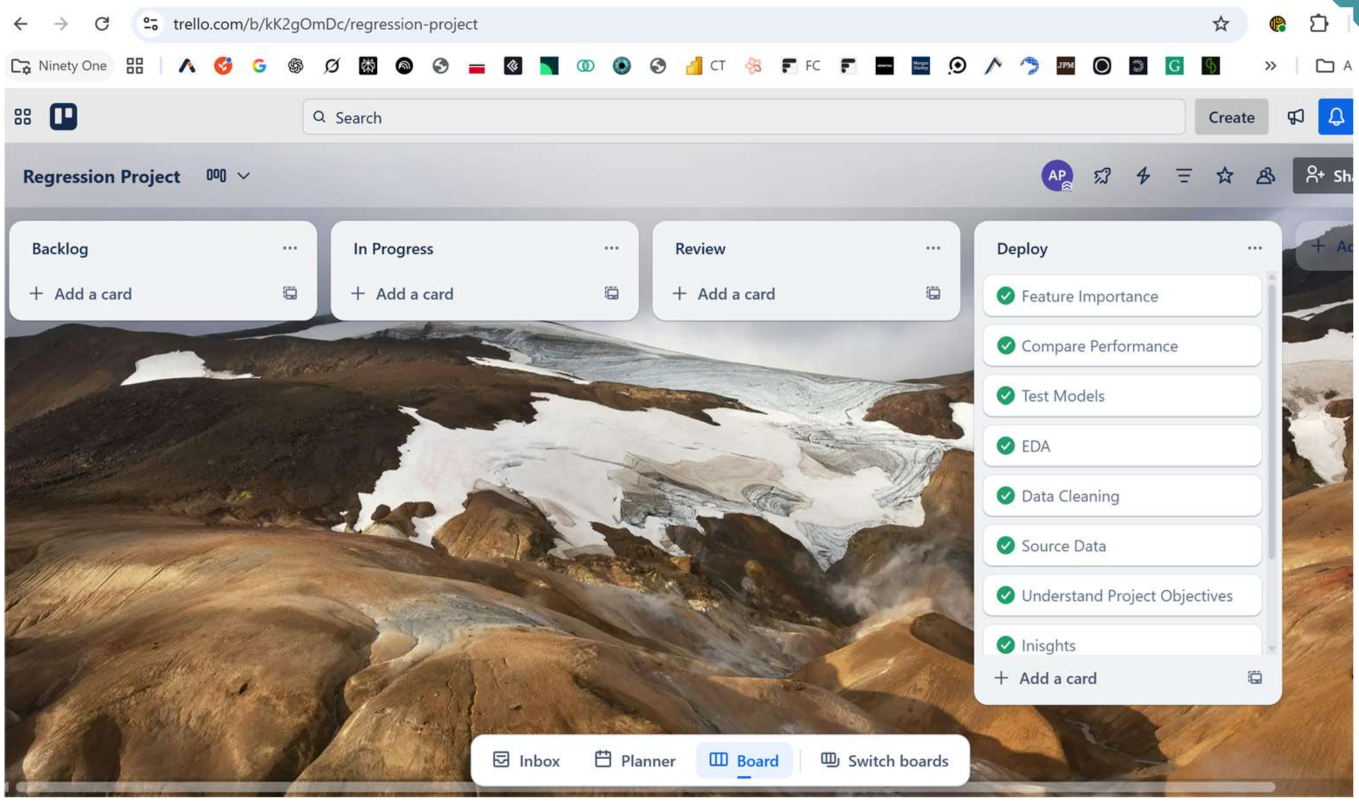
- Year is the strongest predictor of average temperature (time trend)
- Agricultural features provide weak additional signals
- Evidence of global warming pattern: clear upward trend from 1990–2020
- Ridge Regression balances accuracy & stability best.



Recommendations

- Use Ridge Regression as final model
- Consider non-linear models (Polynomial features, Random Forest, Gradient Boosting)
- Improve encoding of Area (collapse rare countries into "Other")
- Explore feature interactions for deeper insights.

Appendix: Trello



trello.com/b/kK2gOmDc/regression-project

Regression Project

Backlog

In Progress

Review

Deploy

- ✓ Feature Importance
- ✓ Compare Performance
- ✓ Test Models
- ✓ EDA
- ✓ Data Cleaning
- ✓ Source Data
- ✓ Understand Project Objectives
- ✓ Inisghts