ECONOMETRICS PROJECT

GROUNDNUT QUALITY ASSESSMENT

PROJECT’S OBJECTIVES

01

Data Synthesis - Compile 2017 district data linking Groundnut production, inputs, rainfall, and soil quality.

02

Economic Influence- Examine the influence of economic growth on groundnut by analyzing changes in variable factors.

03

Regression Analysis - Test for returns and input complementarity in Groundnut production using a Quadratic model and Cobb Douglas model

04

Environmental Insight: To assess the impact of groundnut farming on ethe nvironment by analyzing how irrigation, fertilizer usage, rainfall, and salinity affect resource utilization.

DATA PREPROCESSION

Data Cleaning & Preprocessing for Groundnut Crop

- Season Consideration: Groundnut is a Kharif crop, so we focused on the monsoon months of June to October for 2017. These months are crucial for groundnut cultivation, as they align with the crop's growing season.
- Dataset: We used the rainfall dataset for the year 2017, filtered by the months of June, July, August, September, and October. Only districts where groundnut was grown during these months were selected.
- District Handling: All district names were standardized and renamed to ensure proper merging with other datasets.
- Null Values: Null values in the production dataset, indicating districts where groundnut was not grown, were removed.
- Final Data: After cleaning, 197 observations remained, corresponding to districts where groundnut was grown in the Kharif season.

Descriptive Statistics

	area1000hectares	production1000tonnes	irrigatedarea1000hectares	nitrogenconsumptiontonnes	phosphateconsumptiontonnes	potashconsumptiontonnes	rainfall_2017	salinity_percent	irrigated_area_new	unirrigated_area	log_production	log_nitrogen	log_phosphate	log_potash	log_irrigated	log_unirrigated
count	197	197	197	197	197	197	197	197	197	197	197	197	197	197	197	197
mean	23.4332487	45.296853	16.1544568	54856.0254	23551.4467	8848.86294	777.042116	12.2503748	8.11961419	15.3136345	1.98231191	10.5719783	9.70550756	8.22740581	0.719824191	1.38900779
std	61.0537243	132.931994	53.7171711	39660.2372	16810.0417	10548.1257	608.296547	20.1082412	24.310817	49.9678194	1.77874711	1.00771434	1.07294847	1.61286588	1.458828668	1.33718113
min	0.01	0.01	0	427	107	39	0	0	-0.49	0.5	0.00995033	6.0591232	4.68213123	3.68887945	-0.67334455	0.40546511
Q1 (25%)	0.44999999	0.52999997	0	27278	11119	1783	505.929997	0.05820487	-0.05000001	0.5	0.42526772	10.2138725	9.31650057	7.48661331	-0.0512933	0.40546511
Median (50%)	2.6600001	4	0.103	47161	20580	4865	713.678392	1.99337111	0.021	1.27	1.60943791	10.7613438	9.9321236	8.49002752	0.020782539	0.81977983
Q3 (75%)	16.440001	23.1	4.4000001	75364	31168	12903	843.75	15.8841414	2.283	6.6599998	3.18221184	11.2300983	10.3471793	9.46529262	1.188757639	2.03601196
max	450.78	977.65002	516.539	294303	93574	74158	4369.61	91.835901	220.05	390.399999	6.88617409	12.5923685	11.4465185	11.2139667	5.39838892	5.96973005

- Zero Value Handling:
  - Increased Total Area and Irrigated Area by 1.
- Irrigated Area Calculation:
  - irrigated_area_new = min(Total Area - 0.5, Irrigated Area), ensuring:
    - irrigated_area_new < Total Area.
    - Unirrigated Area > 0.5 for all observations.
- Removal of Zero Production Districts:
  - Removed districts with zero production to prevent distortion in the regression model.

```
== Top Names ==
Top State Name: uttar pradesh
Number of districts in Uttar Pradesh: 35
```

- Uttar Pradesh, contributing the most districts to the data while the average number of districts per state is 15(approx without UP) , suggests that it might influences the overall analysis.

Cobb Douglas Regression

$$\ln(Y) = \beta_0 + \beta_1 \cdot \ln(area) + \beta_2 \cdot \ln(irrigated\_area\_new) + \beta_3 \cdot \ln(unirrigated\_area) + \beta_4 \cdot \ln(nitrogen) + \beta_5 \cdot \ln(phosphate) + \beta_6 \cdot \ln(potash) + \beta_7 \cdot rainfall + \beta_8 \cdot salinity + \varepsilon$$

Variables:

- Y: Total Production (1000 tonnes)
- area: Total Area under Cultivation (1000 hectares)
- irrigated_area_new: Adjusted Irrigated Area (1000 hectares)
- unirrigated_area: Unirrigated Area (1000 hectares)
- nitrogen: Nitrogen Fertilizer (tonnes/ha)
- phosphate: Phosphate Fertilizer (tonnes/ha)
- potash: Potash Fertilizer (tonnes/ha)
- rainfall: Annual Rainfall (mm)
- salinity: Soil Salinity (dS/m)
- ε: Error term

	coef	std err
const	-0.2607	0.387
log_nitrogen	0.0441	0.049
log_potash	0.0036	0.03
log_irrigated	1.0017	0.03
log_unirrigated	0.523	0.027
rainfall_2017	0.0001	5.83E-05

ln(area): Total land under cultivation
ln(irrigated_area_new): Land with assured water supply
ln(unirrigated_area): Land reliant on rainfall
ln(nitrogen): Nutrient input for plant growth
ln(phosphate): Supports root and seed development
ln(potash): Enhances crop quality and resilience
rainfall: Climatic factor affecting water availability
salinity: Soil health indicator
ln(Y): Total groundnut output

Cobb Douglas Regression Outcome

Number of observations: 197
R-squared: 0.927
log_irrigated Coefficient: Positive (0.8398)
As irrigated area increases, production rises significantly on average, highlighting the importance of irrigation in groundnut cultivation.
F-value: 340.5
Highly statistically significant; strong evidence against the null hypothesis (β = 0), confirming the influence of irrigated land on production(in tonnes).

Variable	Coefficient	Interpretation (Effect of Increase)	t-value	Significance Level
Constant	-0.2028	Baseline level of production when all inputs (e.g., area, fertilizers, irrigation) are at log(1). Since log(1) = 0, this is a baseline reference.	-0.488	Not significant (t < 1.65)
log_nitrogen	0.0621	1% increase in nitrogen application (in tonnes) → approximately 0.06% increase in groundnut production (in tonnes).	0.646	Not significant (t < 1.65)
log_phosphate	0.0048	1% increase in phosphate application (in tonnes) → negligible increase in production (in tonnes).	0.052	Not significant (t < 1.65)
log_potash	0.0027	1% increase in potash application (in tonnes) → negligible increase in production (in tonnes).	0.081	Not significant (t < 1.65)
log_irrigated	0.8398	1% increase in irrigated area (in '000 hectares) → approximately 0.84% increase in groundnut production (in tonnes).	30.849	Highly significant (t > 1.65)
log_unirrigated	0.5939	1% increase in unirrigated area (in '000 hectares) → approximately 0.59% increase in groundnut production (in tonnes).	19.302	Highly significant (t > 1.65)
rainfall_2017	0.0001	1 mm increase in rainfall → a very small increase in production (in tonnes).	1.825	Marginally significant (t > 1.65)
salinity_percent	-0.0049	1% increase in salinity → approximately 0.49% decrease in groundnut production (in tonnes).	-2.67	Significant (t > 1.65)

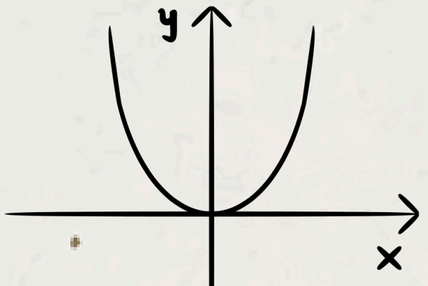
Quadratic Production Function

- Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2$$

- Y → Groundnut production (in 1000 tonnes)
- β0 → Intercept: baseline production when all inputs are zero
- Xi → Input variables (e.g. area, fertilizer, rainfall)
- βi → Coefficients on linear terms (marginal effect of each input)
- γi → Coefficients on squared terms (curvature: increasing or decreasing returns)
- ϵ → Error term (captures unobserved factors)
- Why Quadratic?
- ▪ Captures non-linear relationships between inputs and output
- ▪ Useful for testing diminishing or increasing marginal returns
- Variables Used:
- ▪ Area, Irrigated Area, Nitrogen, Phosphate, Potash, Rainfall, Salinity
- ▪ Plus squared terms for each to detect curvature in relationships

	coef	std err
const	-3.5724	7.554
irrigated_new	2.3627	0.332
unirrigated_area	2.4527	0.21
nitrogenconsumptiontonnes	0.0003	0
phosphateconsumptiontonnes	-0.0005	0
potashconsumptiontonnes	-0.0002	0
total_rainfall	-0.003	0.006
salinity_alkalinity_pct	-0.1018	0.167
irrigated_new_squared	-6.74E-05	0.002
unirrigated_area_squared	-0.0016	0.001



Quadratic Production Function- Interpretation

Final Quadratic Model - Groundnut Production

- Observations: 197
- R-squared: 0.894 (strong model fit)
- F-statistic: 174.6 (highly significant)

Key Coefficients:

- Irrigated Area: +2.36 ($p < 0.001$) → Significant positive impact on production.
- Unirrigated Area: +2.45 ($p < 0.001$) → Also contributes significantly.
- Unirrigated Area²: -0.0016 ($p = 0.012$) → Diminishing returns with increased unirrigated area.

Non-significant: Nitrogen, phosphate, potash, rainfall, salinity, and irrigated area squared.

Variable	Coefficient	Interpretation	t-value	Significance level
constant	-3.574		-0.473	Not significant (t < 1.65)
irrigated_new	2.3627	1 unit ↑ in irrigated area (1000 ha) → ~2.36 unit ↑ in production (1000 tonnes).	7.34	☑ Highly significant (t > 1.65)
unirrigated_new	2.4527	1 unit ↑ in unirrigated area → ~2.45 unit ↑ in production.	11.682	☑ Highly significant (t > 1.65
nitrogrenconsumptiontonnes	0.0003	1 tonne ↑ in nitrogen → ~0.0003 tonne ↑ in production. Small effect.	1.555	✗ Not significant (t < 1.65)
phosphateconsumptiontonnes	-0.0005	Slight ↓ in production. Likely due to multicollinearity or overuse.	-1.2	✗ Not significant (t < 1.65)
potashconsumptiontonnes	-0.0002	Negligible Negative effect	-0.519	✗ Not significant (t < 1.65)
total_rainfall	-0.003	1 mm ↑ in rainfall → ~0.003 tonne ↓ in output. Weak, unclear effect.	-0.543	✗ Not significant (t < 1.65)
salinity_alkalinity_pct	-0.1018	1% in ↑ salinity → ~0.10 tonne ↓ in yield. Suggests salinity harms production	-0.61	✗ Not significant (t < 1.65)
irrigated_new_squared	-6.74E-05	Very small negative curvature → no strong nonlinear effect.	-0.035	✗ Not significant (t < 1.65)
unirrigated_new_squared	-0.0016	Diminishing returns from adding more unirrigated land → yield gains taper of	-2.535	☑ Significant (t > 1.65)

Hypothesis Testing– Constant Returns to Scale Q5

Null and Alternate hypothesis CRS test
Null Hypothesis (Ho): The production function exhibits constant returns to scale (CRS)
→ βnitrogen+βpotash+βirrigated+βunirrigated=1

H_1 (Alternative):
The production function does not exhibit CRS
→ The sum of elasticities is not equal to 1

Estimated sum of elasticities: 1.5741
Standard error: 0.0420
t-statistic: 13.6832
p-value: 0.0000
❌ Reject H0: Evidence against CRS.

Interpretation

The null hypothesis of constant returns to scale (CRS) — that the sum of elasticities for farmer-controlled inputs equals 1 — is rejected at the 1% significance level (p-value = 0.0000).
The estimated elasticity sum is 1.5741 with a standard error of 0.0420, implying a significant deviation from CRS.
Therefore, the data provides strong evidence of increasing returns to scale: farmers who scale up inputs such as irrigation and fertilizer see more than proportionate increases in groundnut yield.

In groundnut cultivation, super-linear returns to input use—especially irrigation and fertilizers—are observed due to the nature of fixed costs and mechanization. Once farmers invest in irrigation infrastructure like tube wells, sprinklers, or drip systems, each additional hectare planted benefits from pre-established systems, reducing per-unit cost and amplifying output. Mechanized sowing and harvesting tools also become more efficient when used over larger areas, further boosting marginal returns with scale.

Hypothesis Testing – Insights & Significance Q6

Variable: irrigated_new_squared
Null Hypothesis (Ho): The squared term for irrigated area does not affect production.
Beta irrigated_new_squared=0
Alternative Hypothesis (H1): The squared term for irrigated area significantly affects production.
irrigated_new_squared not 0
F-statistic: 0.0012 T-statistic :0.0012
p-value: 0.972

.**Conclusion: Fail to reject Ho - Irrigated area is not significant**

Variable	Coefficient	p-value	Interpretation
irrigated_new_squared	-0.00007	0.972	❌ Not significant – no non-linear irrigation effect
unirrigated_area_squared	-0.0016	0.012	✅ Significant – shows **diminishing returns**

Variable: unirrigated_new_squared
Null Hypothesis (Ho): The squared term for unirrigated area has no effect on production.
Beta unirrigated_area_squared = 0
Alternative Hypothesis (H1): The squared term for unirrigated area significantly affects production.
Beta unirrigated_area_squared not 0
F-statistic: 6.4247 T statistic: 6.4247
p-value: 0.0121

Conclusion: Reject Ho - Unirrigated area is Significant

Variable	Coefficient	p-value	Interpretation
irrigated_new_squared	0.0012	0.972	❌ Not significant – no non-linear irrigation effect
unirrigated_area_squared	6.4247	0.012	✅ Significant – shows **diminishing returns**

Hypothesis Testing: Input complementary

Purpose:
To check if irrigation and individual fertilizers (nitrogen, phosphate, potash) interact significantly — i.e., complement each other in contributing to groundnut production.
Hypothesis (Individual Interactions):
H_0: $\beta_{(irrigation \times fertilizer)} = 0$
H_1: $\beta_{(irrigation \times fertilizer)} \neq 0$

Interpretation: All p-values $\gg 0.05 \Rightarrow$ no evidence of synergy.
Fertilizer effectiveness does not depend on irrigation.

Variable	t-stat	p-value	Interpretation
Irrigation × Nitrogen	0.557	0.5782	❌ Fail to reject H_0
Irrigation × Phosphate	-0.4144	0.679	❌ Fail to reject H_0
Irrigation × Potash	0.6518	0.5154	❌ Fail to reject H_0

Purpose:
To test if all irrigation-fertilizer interaction terms jointly have a significant impact on production.
Hypothesis (Joint Test):
H_0: All $\beta_{(interaction\ terms)} = 0$
H_1: At least one $\beta \neq 0$
Test: F-test for joint significance
• F-statistic = 0.3653, p = 0.7781 → Fail to reject H_0
Interpretation: No joint complementarity found between irrigation and fertilizer inputs.

Variable	Coefficient	p-value	Interpretation
irrigated_npk	0.3653	0.7781	Fail to reject H_0

Hypothesis Testing: Regional Differences

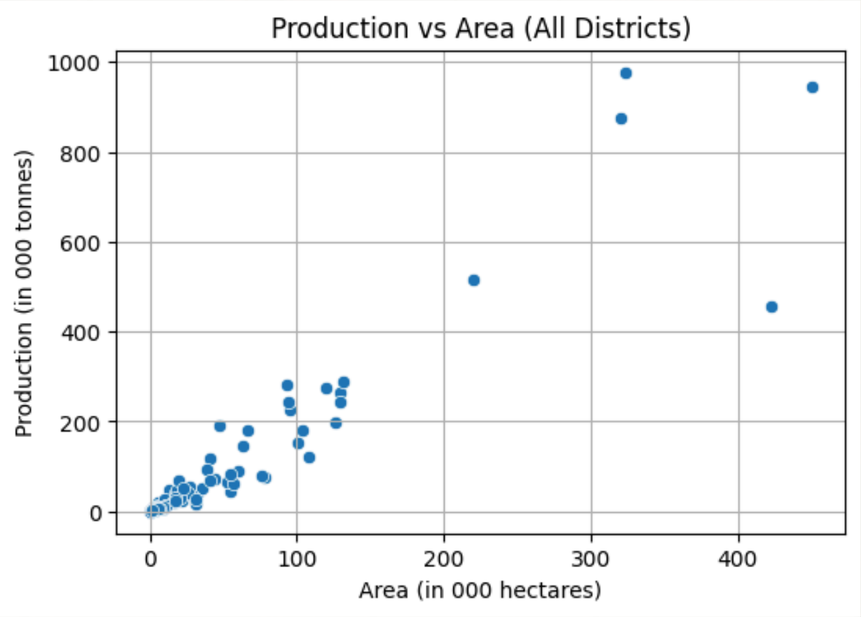
H_0: All region–input interaction terms = 0
H_1: At least one ≠ 0
Test: F-test on joint significance of dummy–interaction block
Result: Reject H_0 ($p<0.01$).

Interpretation: Production functions differ significantly across regions; soil, climate, and practices vary regionally.

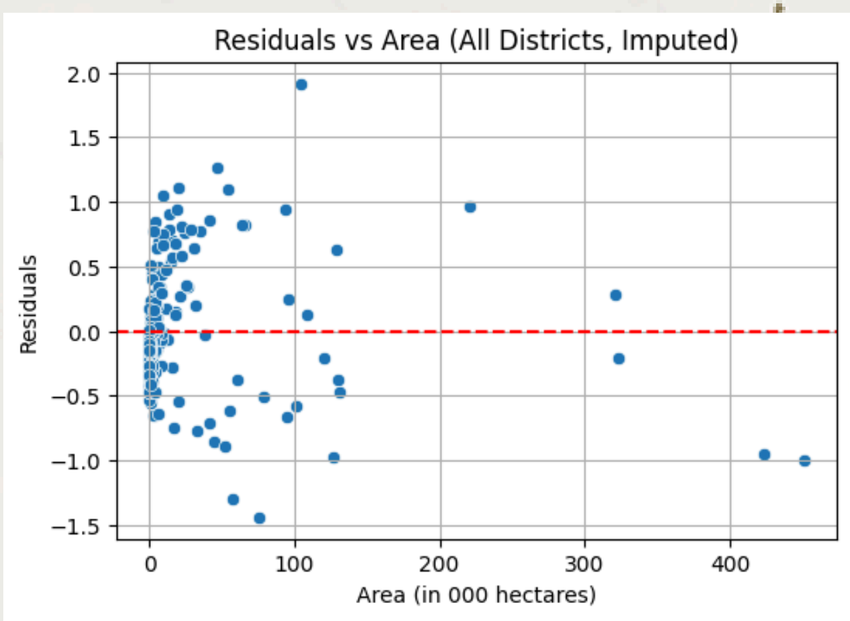
We added dummy variables for all regions along with their interactions with the other terms. The null was rejected, indicating significant structural differences in production functions across North, South, East, and West zones.
Groundnut, grown mainly in the Kharif season, is influenced by distinct regional factors. The agro-ecological zones differ in monsoon timing, rainfall distribution, and soil types, affecting returns to irrigation and fertilizer use. For instance, in the humid South, rainfall boosts productivity, while in the arid West, reliance on irrigation infrastructure plays a larger role.

As a result, the same bundle of inputs—say, a fixed irrigation dose and standard NPK ratio—can produce very different outcomes depending on the region. This regional heterogeneity justifies the statistical finding that input–region interaction terms are jointly significant, confirming that production functions must be adapted to local agro-climatic realities.

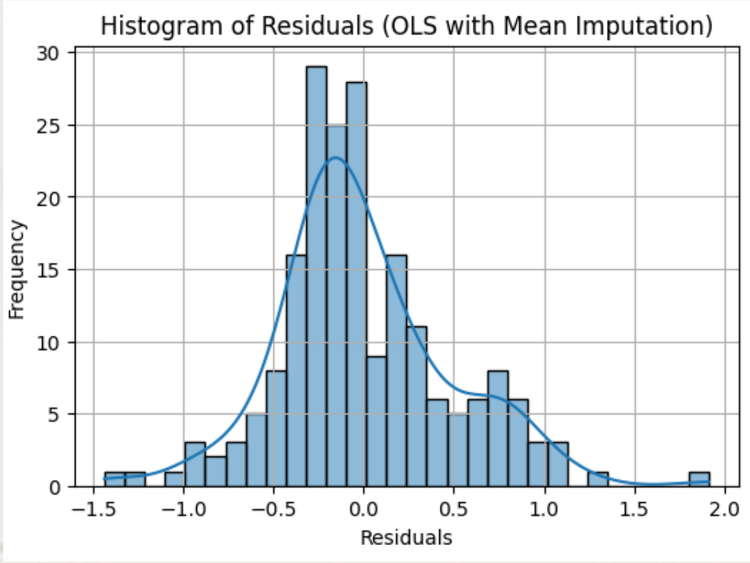
Residual Diagnostics



Production vs. area should rise steadily, with residuals randomly colored—no clear pattern.



A plot of residuals (Y-axis) against crop area (X-axis) should not show any distinct structure or trend, reflecting constant variance (homoscedasticity).



The residuals' histogram should approximate the standard result

Robustness Checks

What We Tested:

- Outlier Removal & Influential Observations

13 outliers with standardized residuals > 2 were identified

- 10 influential points flagged using DFBETA and DFITS
- Removed districts:

→ Ahmedabad (GJ) – Hybrid varieties, unusually high yield

→ Ujjain (MP) – Potential data quality issues

→ Dehradun (UK) – High-altitude climate anomaly

- Multicollinearity Test (VIF Analysis)

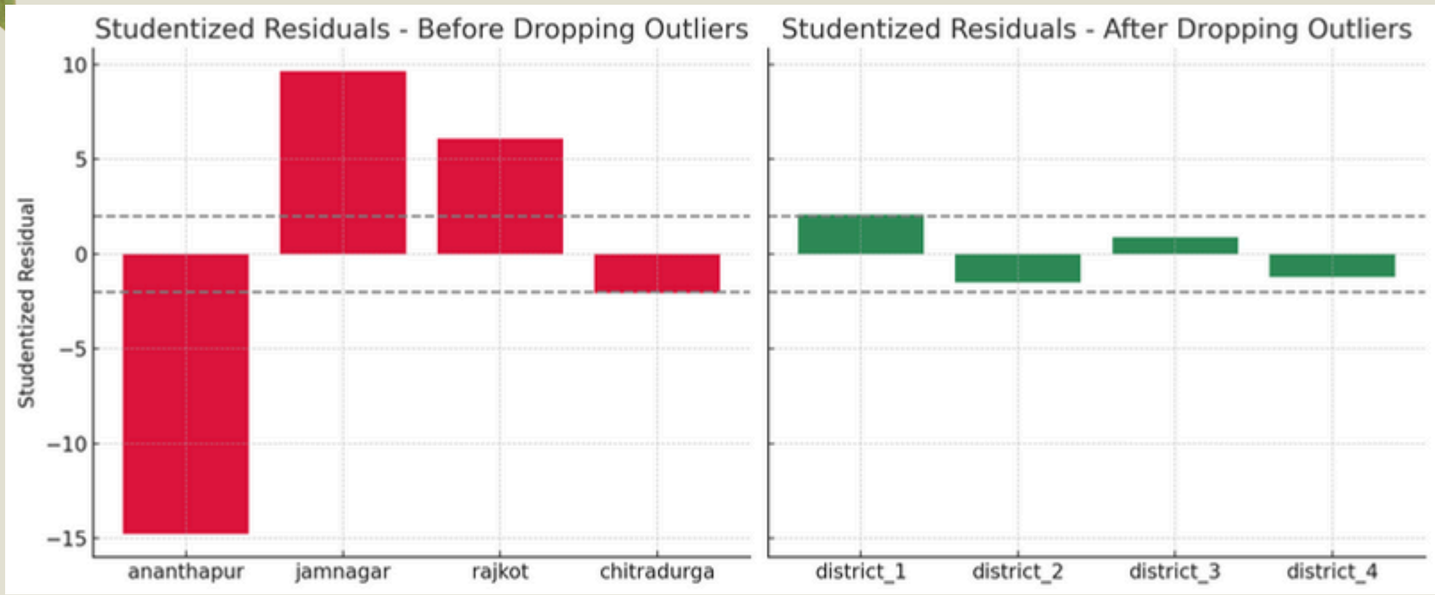
→ Nitrogen & Phosphate → VIF > 10 → Highly collinear

Likely applied together by farmers → Hard to distinguish effects

Outlier analysis removed 13 districts with abnormal yield patterns and 10 with high influence. Notable exclusions include Ahmedabad (GJ) with hybrid groundnut use, Ujjain (MP) due to data issues, and Dehradun (UK) where high-altitude climate skews typical groundnut conditions. Groundnut needs warm temperatures (25–30°C) and well-drained soils—conditions inconsistent in hilly regions like Uttarakhand. VIF analysis showed high multicollinearity between nitrogen and phosphate, likely due to joint application in standard NPK fertilizers. This makes it hard to isolate their individual effects in regression.

Robustness Checks

- Y-Outliers (|Residual| > 2):
  → Extreme deviations in production found in Ananthapur, Jamnagar, Rajkot, Chitradurga.
- High Leverage Points:
  → Districts like Thanjavur and Ananthapur had high leverage, indicating potential distortion of model estimates.
- Influential Observations (DFFITS):
  → Ananthapur, Rajkot, Jamnagar showed strong influence on model predictions.
- Action Taken:
→ Top 5 outliers dropped based on residuals and influence.
  → Post-cleaning, residuals stabilized, improving model reliability.



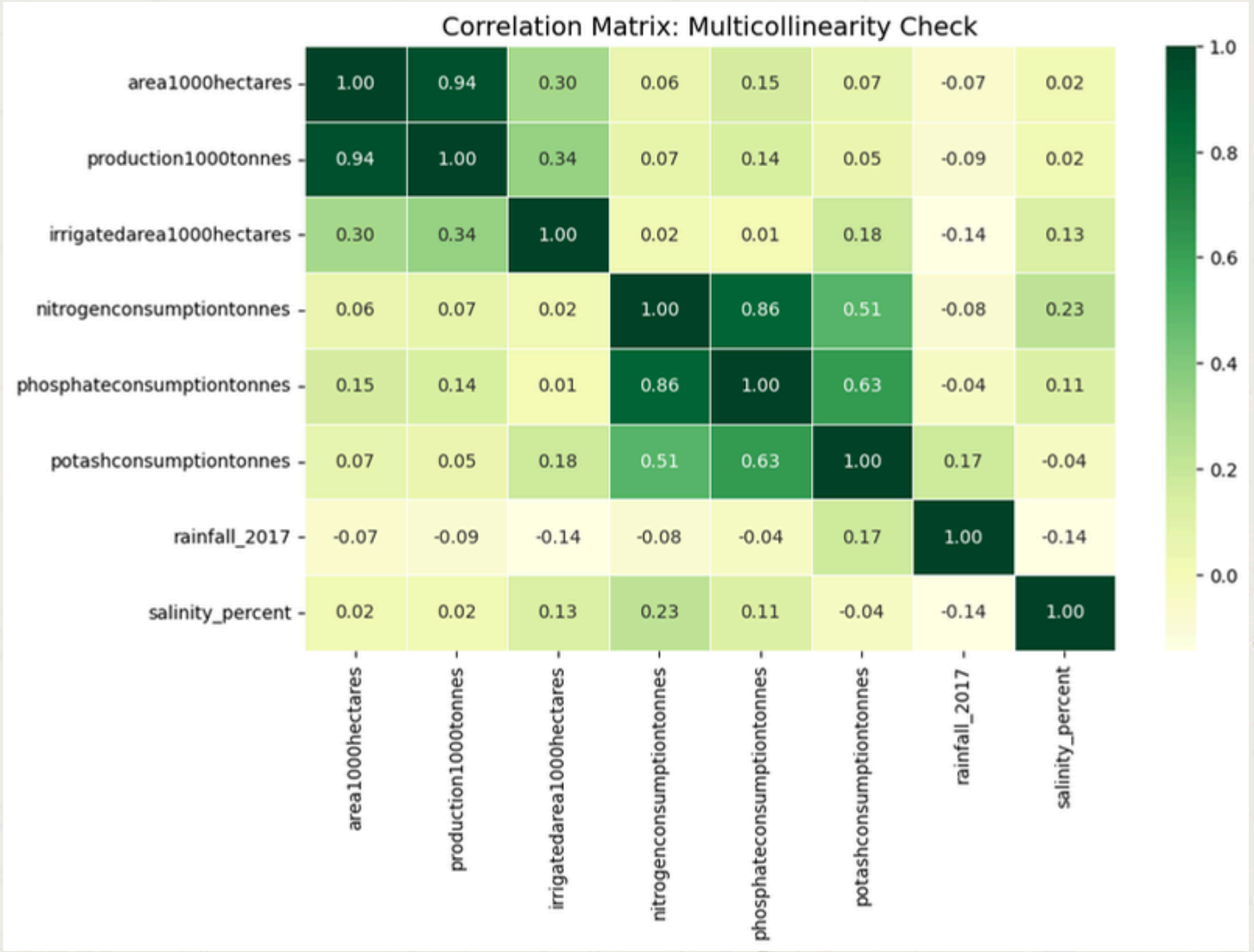
	Metric	District	Value
0	Studentized Residual (Y-Outlier)	ananthapur	-14.77
1	Studentized Residual (Y-Outlier)	jamnagar	9.65
2	Studentized Residual (Y-Outlier)	rajkot	6.09
3	Studentized Residual (Y-Outlier)	chitradurga	-2.08

	Metric	District	Value
4	Leverage (X-Outlier)	thanjavur	0.502
5	Leverage (X-Outlier)	ananthapur	0.249
6	Leverage (X-Outlier)	junagadh	0.262
7	Leverage (X-Outlier)	hissar	0.275
8	Leverage (X-Outlier)	jalgaon	0.301

	Metric	District	Value
9	DFFITS (Influence)	ananthapur	-8.5
10	DFFITS (Influence)	jamnagar	3.83
11	DFFITS (Influence)	rajkot	2.55
12	DFFITS (Influence)	thanjavur	-1.79
13	DFFITS (Influence)	jalgaon	0.52

Multicollinearity Check

	variable	VIF
0	area1000hectares	10.306009
1	production1000tones	10.322209
2	irrigatedarea1000hectares	1.380057
3	nitrogenconsumptiontonne	11.94412
4	phosphateconsumptionton	14.212834
5	potashconsumptiontonnes	3.286735
6	rainfall_2017	1.737473
7	salinity_2017	1.517177

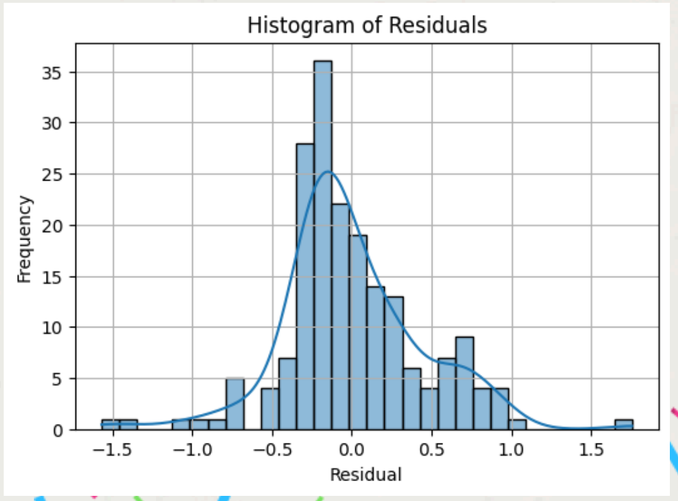


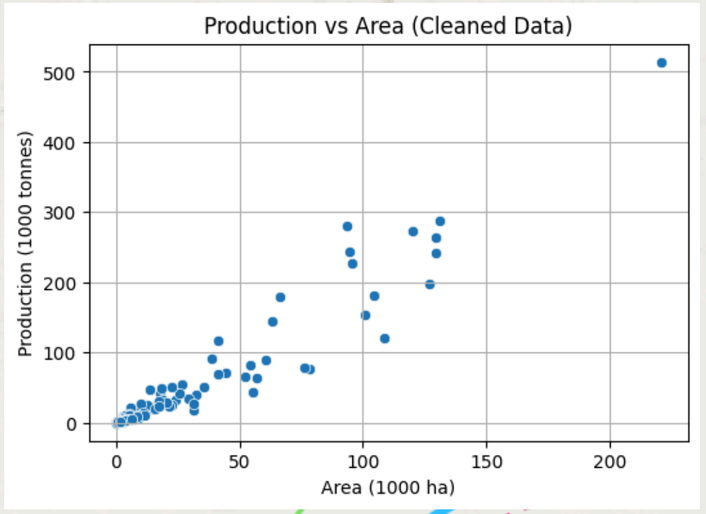
The multicollinearity check shows severe overlap among predictors. area1000hectares and production1000tonnes are highly correlated (0.94), as are nitrogen and phosphate (0.86), leading to VIFs above 10. This inflates standard errors and weakens coefficient reliability. In contrast, rainfall and salinity show low correlation with inputs. Conclusion: Multicollinearity is present and can distort inference.

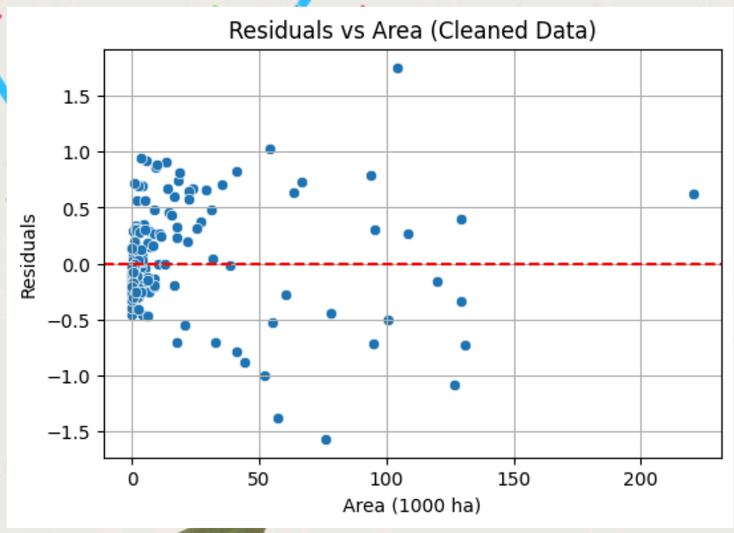
Residual Diagnostics - Cleaned Data

The following are not included in the analysis:

- Districts: Ananthapur, Jamnagar, Rajkot, Thanjavur, Hissar, South Arcot / Cuddalore, Jalgaon, Junagadh.
- Variables: area1000hectares, phosphateconsumptiontonnes (due to multicollinearity with other predictors).





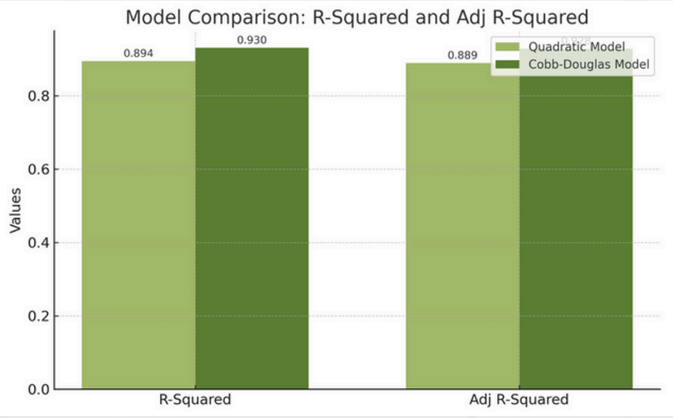


- Improvements in the Cleaned Data Model:
- R-squared: Increased from 0.927 to 0.930.
- Adjusted R-squared: Increased from 0.924 to 0.928.
- F-statistic: Increased from 340.5 to 402.1.
- Log-Likelihood: Increased from -135.31 to -111.45.
- P-values:
  - log_nitrogen and log_potash became less significant ($p > 0.05$).
  - log_irrigated and log_unirrigated coefficients' significance remained strong ($p = 0.000$).
- Coefficient Values:
  - log_irrigated: Increased from 0.8398 to 0.9996.

Model Fit Comparison

Metric	Quadratic Model	Cobb-Douglas Model	Better
R-Squared	0.894	0.930	Cobb-Douglas
Adj R-Squared	0.889	0.928	Cobb-Douglas
AIC	2063	249.0	Cobb-Douglas
BIC	2096	271.8	Cobb-Douglas
F-Stastistic	174.6	412.1	Cobb-Douglas

The Cobb-Douglas model is better
- statistically and practically.
-Higher R^2
-More significant coefficients.
-Easier to interpret for policy or optimization.



LIMITATIONS

1. Data Quality Issues: Missing values and potential errors in the dataset may affect the accuracy of the analysis.
2. Regional Bias: Limited regional coverage and agro-ecological variations may skew the results.
3. Omitted Variables: Exclusion of important factors like pest control, farm management, and socio-economic influences.
4. Model Assumptions: Assumes linear relationships, which may not fully capture complex agricultural dynamics.
5. External Influences: Market dynamics and policy changes are not considered but could significantly affect production.
6. Selection Bias: Focusing only on Kharif season groundnut may exclude regions with differing production patterns.
7. Generalizability: The model may not generalize well to other regions or crops, limiting broader application.
8. Simplified Input Relationships: Irrigation and fertilizer use are treated too simplistically, ignoring regional differences in practices.

Thank You

Team Members

- Sanskriti Gupta-2023487
- Arnav Singh-2023128
- Yash Goel-2023606
- Mansi Garg-2023310

🥜 **"We dug deep into the data—just like groundnuts, all the insights were hiding below the surface!"**