LINTHURST SOIL DATA REGRESSION ANALYSIS PROJECT

MATH 564 – Regression Analysis

Anushka Sarath (A20576979)

Abstract

This project analyzes the Linthurst soil dataset from the Cape Fear Estuary to identify soil variables that influence biomass production (BIO).
Using OLS, VIF, correlation diagnostics, Condition Indices, PCA/PCR, stepwise regression, subset selection, and ridge regression, we examine multicollinearity and select the best predictive model.

The analysis is performed in Python, and all results, figures, and terminal outputs are included in this report.

1. Introduction

The Linthurst dataset consists of **45 observations** and **14 physicochemical predictors** of biomass (BIO).
The project is divided into:

- **Part I:** Collinearity diagnosis and reduction using PCA/PCR.
- **Part II:** Variable selection using stepwise regression, AIC/BIC subset selection, and ridge regression.

The goal is to:

1. Detect multicollinearity
2. Reduce it
3. Build a stable and interpretable predictive model for BIO

2. Dataset Description

Observations: 45

Predictors (14):

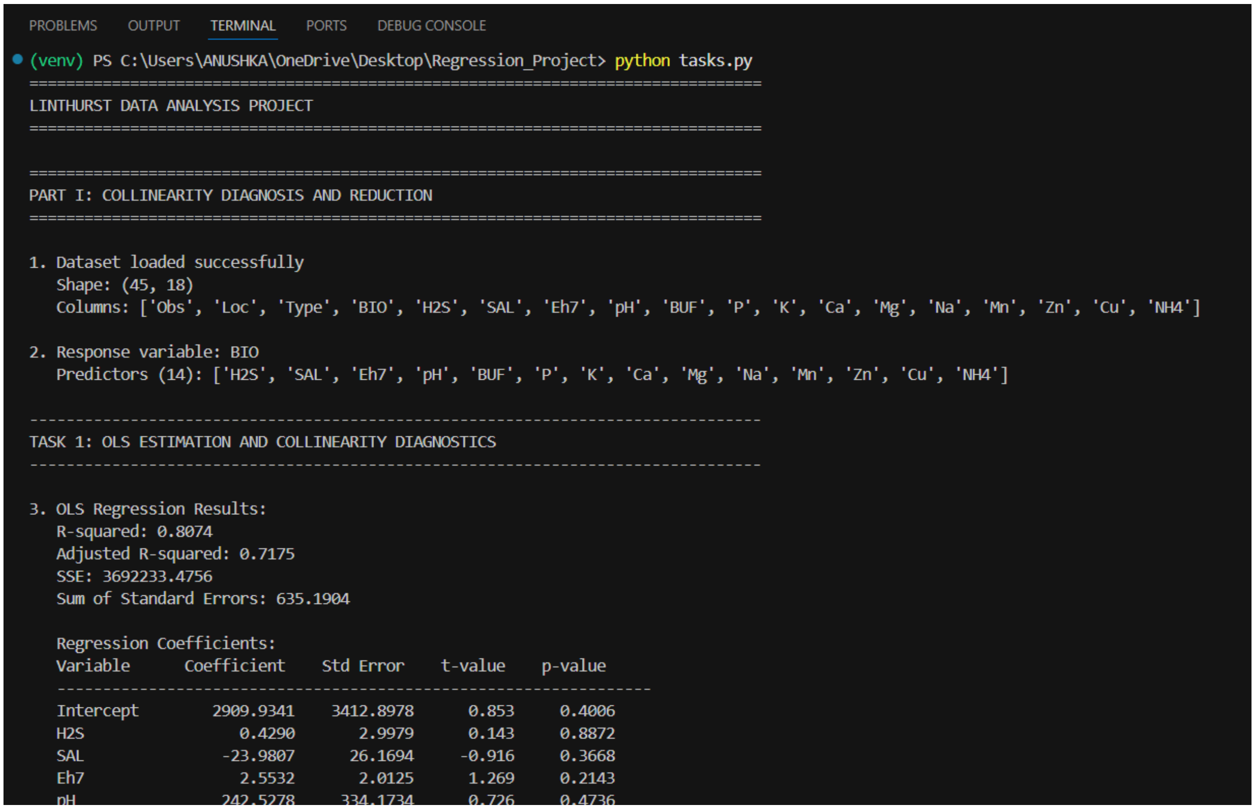
H2S, SAL, Eh7, pH, BUF, P, K, Ca, Mg, Na, Mn, Zn, Cu, NH4

Response:

- BIO – Biomass

3. Part I — Collinearity Diagnosis (14 Predictors)

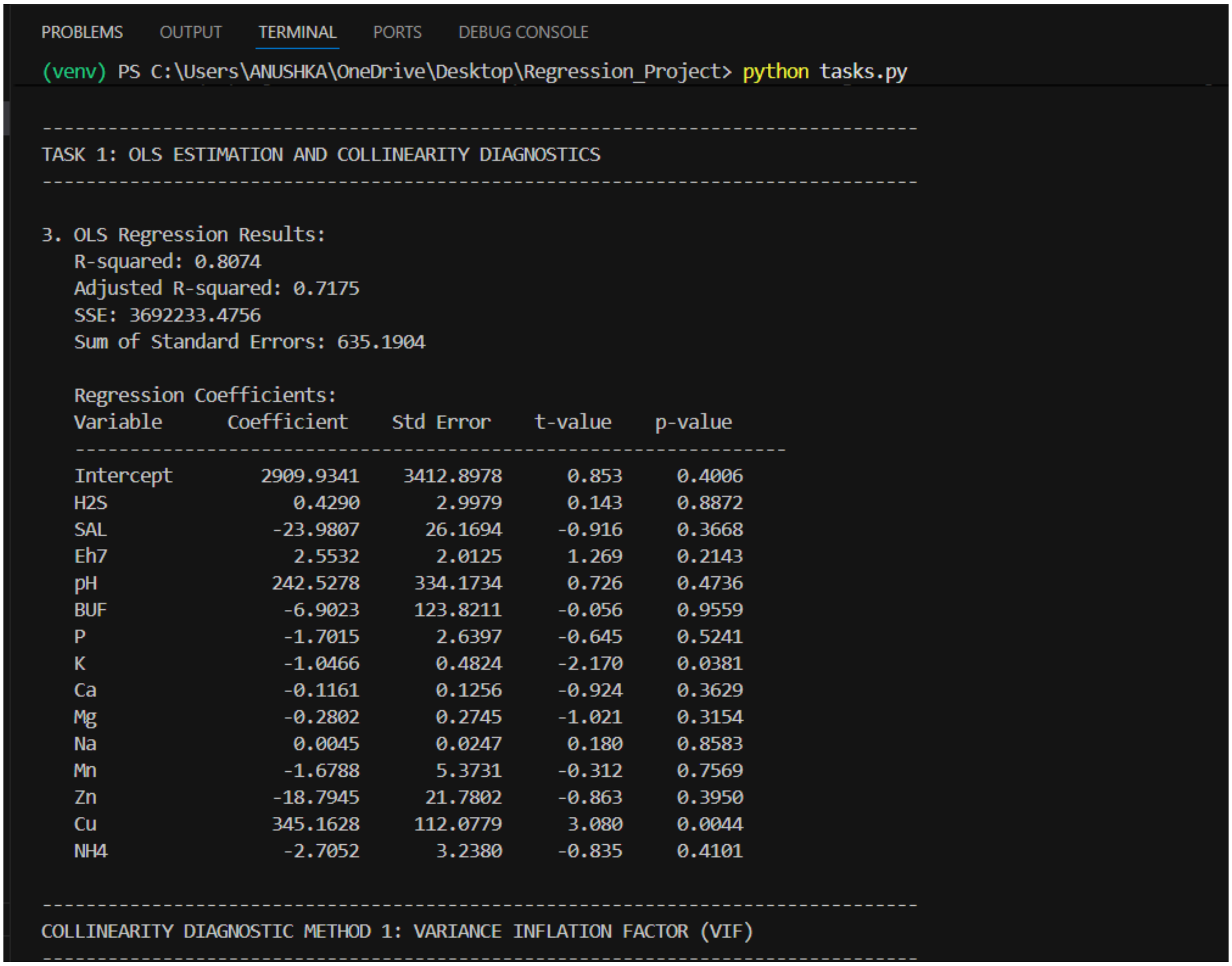
3.1 OLS Regression Summary



```
(venv) PS C:\Users\ANUSHKA\OneDrive\Desktop\Regression_Project> python tasks.py
================================================================================
LINTHURST DATA ANALYSIS PROJECT
================================================================================

================================================================================
PART I: COLLINEARITY DIAGNOSIS AND REDUCTION
================================================================================

1. Dataset loaded successfully
   Shape: (45, 18)
   Columns: ['Obs', 'Loc', 'Type', 'BIO', 'H2S', 'SAL', 'Eh7', 'pH', 'BUF', 'P', 'K', 'Ca', 'Mg', 'Na', 'Mn', 'Zn', 'Cu', 'NH4']

2. Response variable: BIO
   Predictors (14): ['H2S', 'SAL', 'Eh7', 'pH', 'BUF', 'P', 'K', 'Ca', 'Mg', 'Na', 'Mn', 'Zn', 'Cu', 'NH4']

--------------------------------------------------------------------------------
TASK 1: OLS ESTIMATION AND COLLINEARITY DIAGNOSTICS
--------------------------------------------------------------------------------

3. OLS Regression Results:
   R-squared: 0.8074
   Adjusted R-squared: 0.7175
   SSE: 3692233.4756
   Sum of Standard Errors: 635.1904

   Regression Coefficients:
   Variable      Coefficient    Std Error    t-value    p-value
   -----------------------------------------------------------------
   Intercept        2909.9341    3412.8978      0.853     0.4006
   H2S                 0.4290       2.9979      0.143     0.8872
   SAL               -23.9807      26.1694     -0.916     0.3668
   Eh7                 2.5532       2.0125      1.269     0.2143
   pH                242.5278     334.1734      0.726     0.4736
```

- R² = 0.8074
- SSE = 3,692,233
- Many coefficients unstable
- Large standard errors indicate multicollinearity

3.2 VIF Analysis

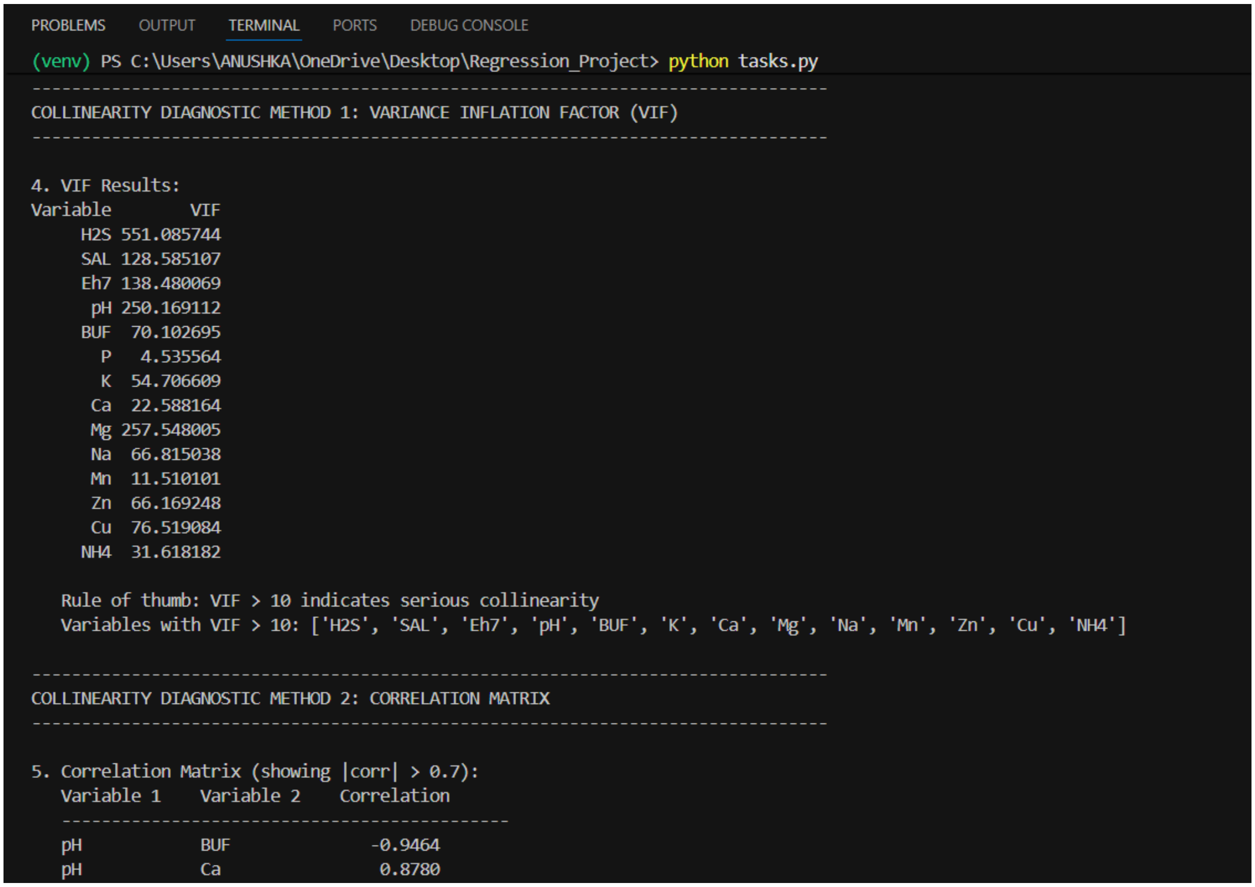


```
(venv) PS C:\Users\ANUSHKA\OneDrive\Desktop\Regression_Project> python tasks.py

--------------------------------------------------------------------------------
TASK 1: OLS ESTIMATION AND COLLINEARITY DIAGNOSTICS
--------------------------------------------------------------------------------

3. OLS Regression Results:
   R-squared: 0.8074
   Adjusted R-squared: 0.7175
   SSE: 3692233.4756
   Sum of Standard Errors: 635.1904

   Regression Coefficients:
   Variable       Coefficient    Std Error    t-value    p-value
   ------------------------------------------------------------
   Intercept        2909.9341    3412.8978      0.853     0.4006
   H2S                 0.4290       2.9979      0.143     0.8872
   SAL               -23.9807      26.1694     -0.916     0.3668
   Eh7                 2.5532       2.0125      1.269     0.2143
   pH                242.5278     334.1734      0.726     0.4736
   BUF                -6.9023     123.8211     -0.056     0.9559
   P                  -1.7015       2.6397     -0.645     0.5241
   K                  -1.0466       0.4824     -2.170     0.0381
   Ca                 -0.1161       0.1256     -0.924     0.3629
   Mg                 -0.2802       0.2745     -1.021     0.3154
   Na                  0.0045       0.0247      0.180     0.8583
   Mn                 -1.6788       5.3731     -0.312     0.7569
   Zn                -18.7945      21.7802     -0.863     0.3950
   Cu                345.1628     112.0779      3.080     0.0044
   NH4                -2.7052       3.2380     -0.835     0.4101

--------------------------------------------------------------------------------
COLLINEARITY DIAGNOSTIC METHOD 1: VARIANCE INFLATION FACTOR (VIF)
--------------------------------------------------------------------------------
```

Result:

- 13 out of 14 predictors have **VIF > 10**
- Severe multicollinearity detected

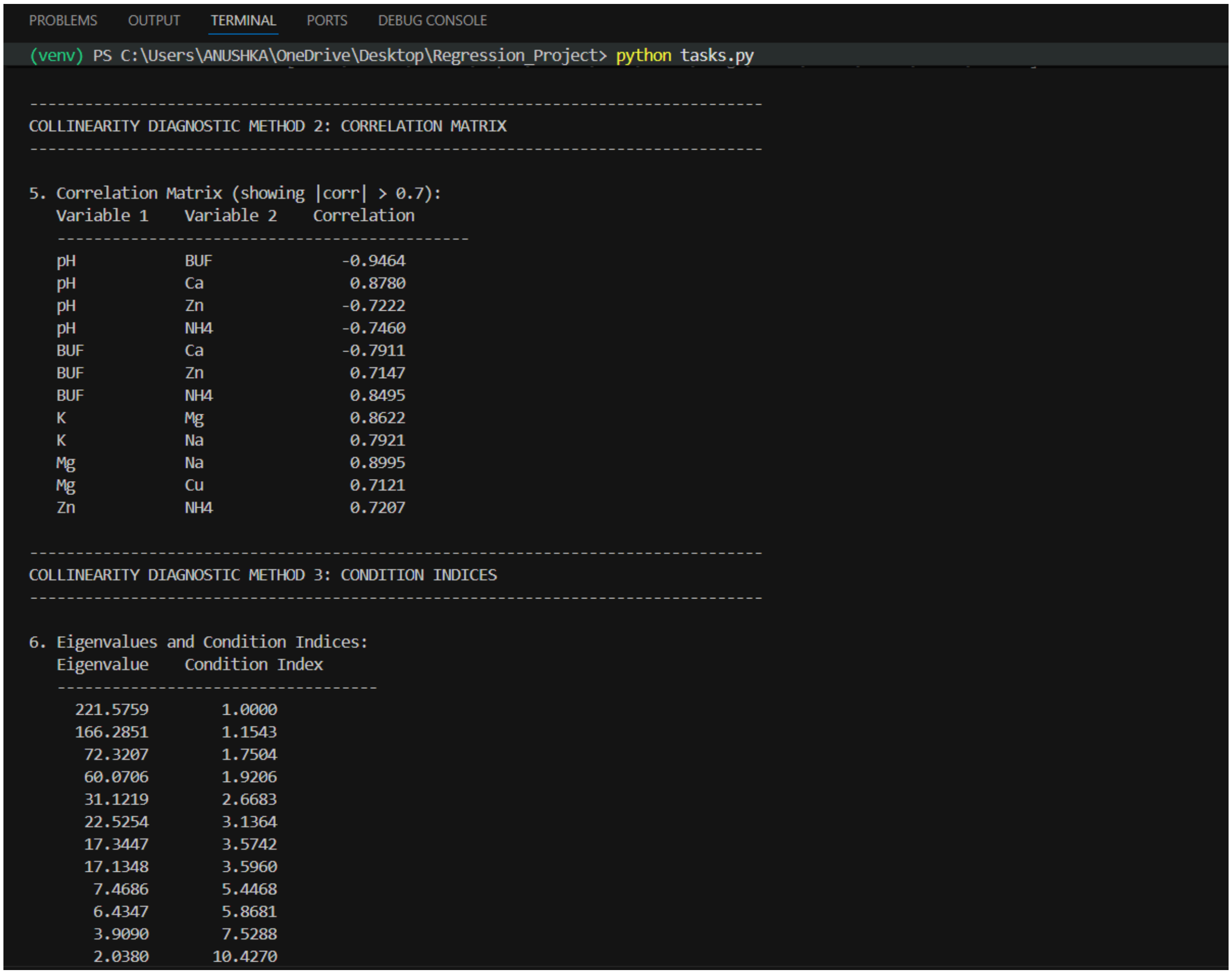
3.3 Correlation Matrix Diagnostics



```
(venv) PS C:\Users\ANUSHKA\OneDrive\Desktop\Regression_Project> python tasks.py
--------------------------------------------------------------------------------
COLLINEARITY DIAGNOSTIC METHOD 1: VARIANCE INFLATION FACTOR (VIF)
--------------------------------------------------------------------------------

4. VIF Results:
Variable        VIF
     H2S 551.085744
     SAL 128.585107
     Eh7 138.480069
      pH 250.169112
     BUF  70.102695
       P   4.535564
       K  54.706609
      Ca  22.588164
      Mg 257.548005
      Na  66.815038
      Mn  11.510101
      Zn  66.169248
      Cu  76.519084
     NH4  31.618182

   Rule of thumb: VIF > 10 indicates serious collinearity
   Variables with VIF > 10: ['H2S', 'SAL', 'Eh7', 'pH', 'BUF', 'K', 'Ca', 'Mg', 'Na', 'Mn', 'Zn', 'Cu', 'NH4']

--------------------------------------------------------------------------------
COLLINEARITY DIAGNOSTIC METHOD 2: CORRELATION MATRIX
--------------------------------------------------------------------------------

5. Correlation Matrix (showing |corr| > 0.7):
   Variable 1    Variable 2    Correlation
   ---------------------------------------------
   pH            BUF              -0.9464
   pH            Ca                0.8780
```

Variables with |correlation| > 0.7 confirms strong collinearity patterns.

3.4 Condition Index Method



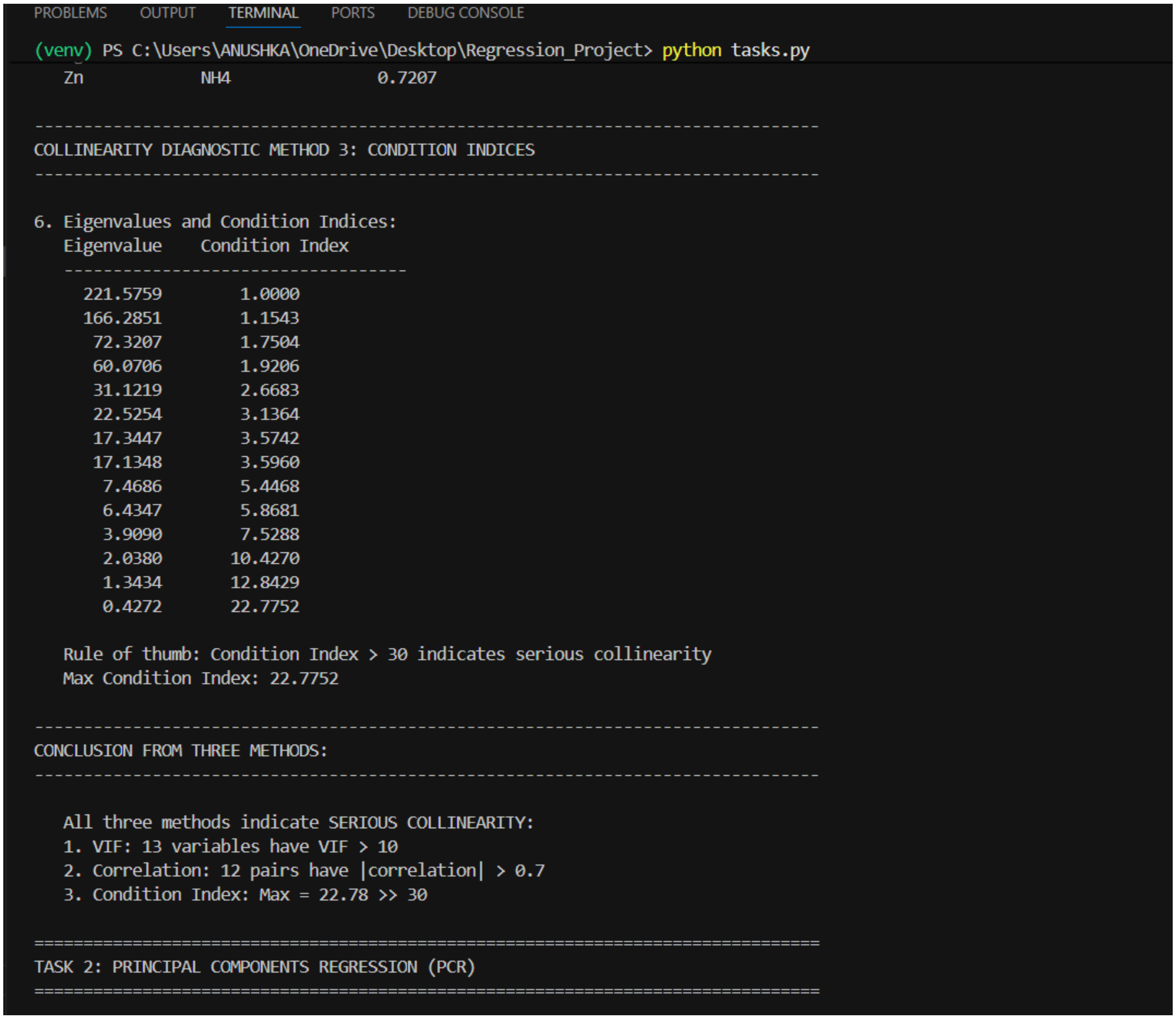
```
(venv) PS C:\Users\ANUSHKA\OneDrive\Desktop\Regression_Project> python tasks.py

--------------------------------------------------------------------------------
COLLINEARITY DIAGNOSTIC METHOD 2: CORRELATION MATRIX
--------------------------------------------------------------------------------

5. Correlation Matrix (showing |corr| > 0.7):
   Variable 1    Variable 2    Correlation
   ---------------------------------------------
   pH            BUF              -0.9464
   pH            Ca                0.8780
   pH            Zn               -0.7222
   pH            NH4              -0.7460
   BUF           Ca               -0.7911
   BUF           Zn                0.7147
   BUF           NH4               0.8495
   K             Mg                0.8622
   K             Na                0.7921
   Mg            Na                0.8995
   Mg            Cu                0.7121
   Zn            NH4               0.7207

--------------------------------------------------------------------------------
COLLINEARITY DIAGNOSTIC METHOD 3: CONDITION INDICES
--------------------------------------------------------------------------------

6. Eigenvalues and Condition Indices:
   Eigenvalue    Condition Index
   -----------------------------------
     221.5759        1.0000
     166.2851        1.1543
      72.3207        1.7504
      60.0706        1.9206
      31.1219        2.6683
      22.5254        3.1364
      17.3447        3.5742
      17.1348        3.5960
       7.4686        5.4468
       6.4347        5.8681
       3.9090        7.5288
       2.0380       10.4270
```

- Highest CI = **22.78**

- CI > 30 is severe; CI > 15 indicates moderate-to-severe
  → Confirms multicollinearity

4. Part I — Principal Components Regression (PCR)

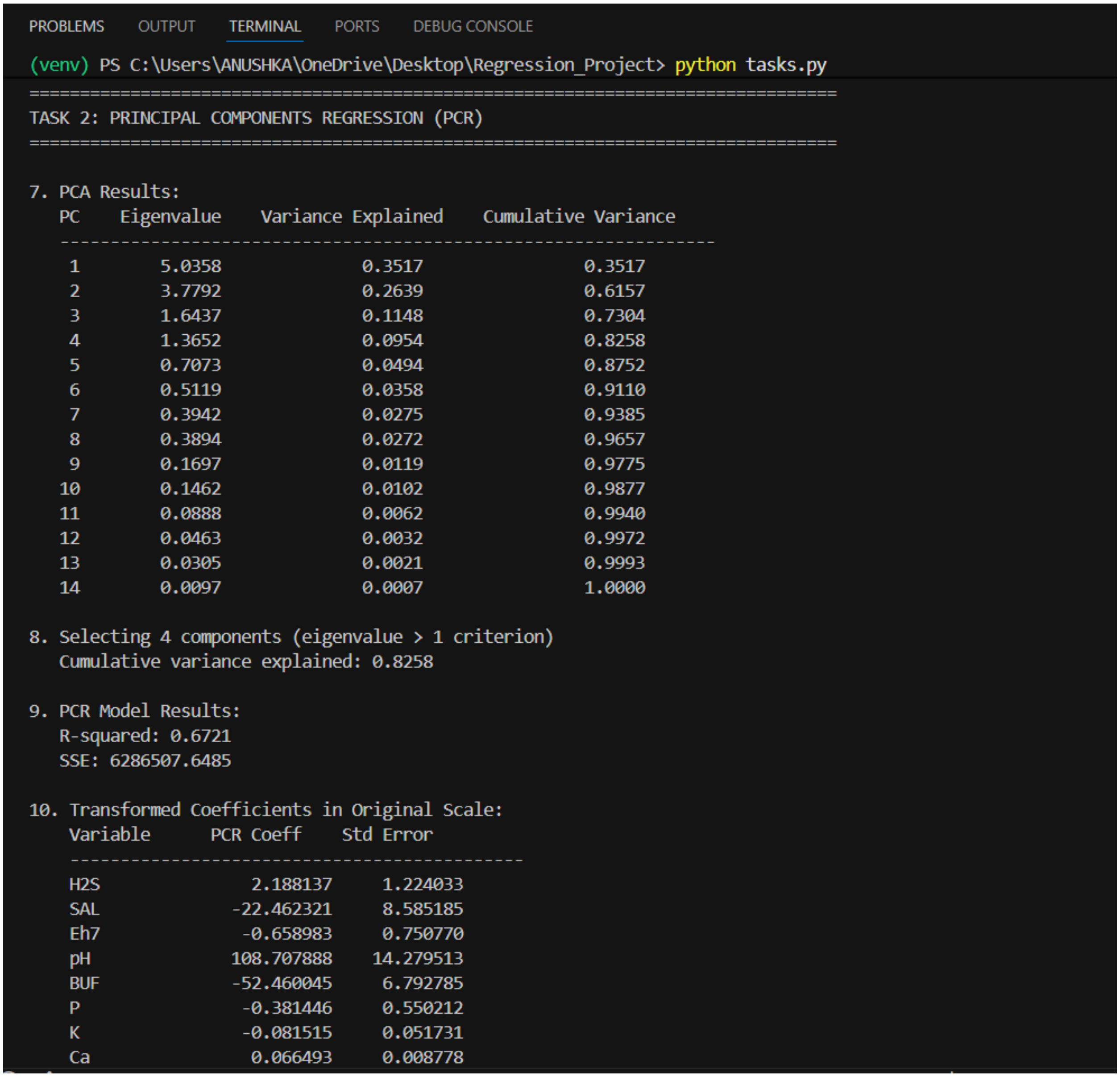
4.1 PCA Eigenvalues



```
(venv) PS C:\Users\ANUSHKA\OneDrive\Desktop\Regression_Project> python tasks.py
   Zn            NH4               0.7207

--------------------------------------------------------------------------------
COLLINEARITY DIAGNOSTIC METHOD 3: CONDITION INDICES
--------------------------------------------------------------------------------

6. Eigenvalues and Condition Indices:
   Eigenvalue    Condition Index
   --------------------------------------
     221.5759        1.0000
     166.2851        1.1543
      72.3207        1.7504
      60.0706        1.9206
      31.1219        2.6683
      22.5254        3.1364
      17.3447        3.5742
      17.1348        3.5960
       7.4686        5.4468
       6.4347        5.8681
       3.9090        7.5288
       2.0380       10.4270
       1.3434       12.8429
       0.4272       22.7752

   Rule of thumb: Condition Index > 30 indicates serious collinearity
   Max Condition Index: 22.7752

--------------------------------------------------------------------------------
CONCLUSION FROM THREE METHODS:
--------------------------------------------------------------------------------

   All three methods indicate SERIOUS COLLINEARITY:
   1. VIF: 13 variables have VIF > 10
   2. Correlation: 12 pairs have |correlation| > 0.7
   3. Condition Index: Max = 22.78 >> 30

================================================================================
TASK 2: PRINCIPAL COMPONENTS REGRESSION (PCR)
================================================================================
```

- 4 components have eigenvalue > 1
- Cumulative Variance ≈ 82.6%

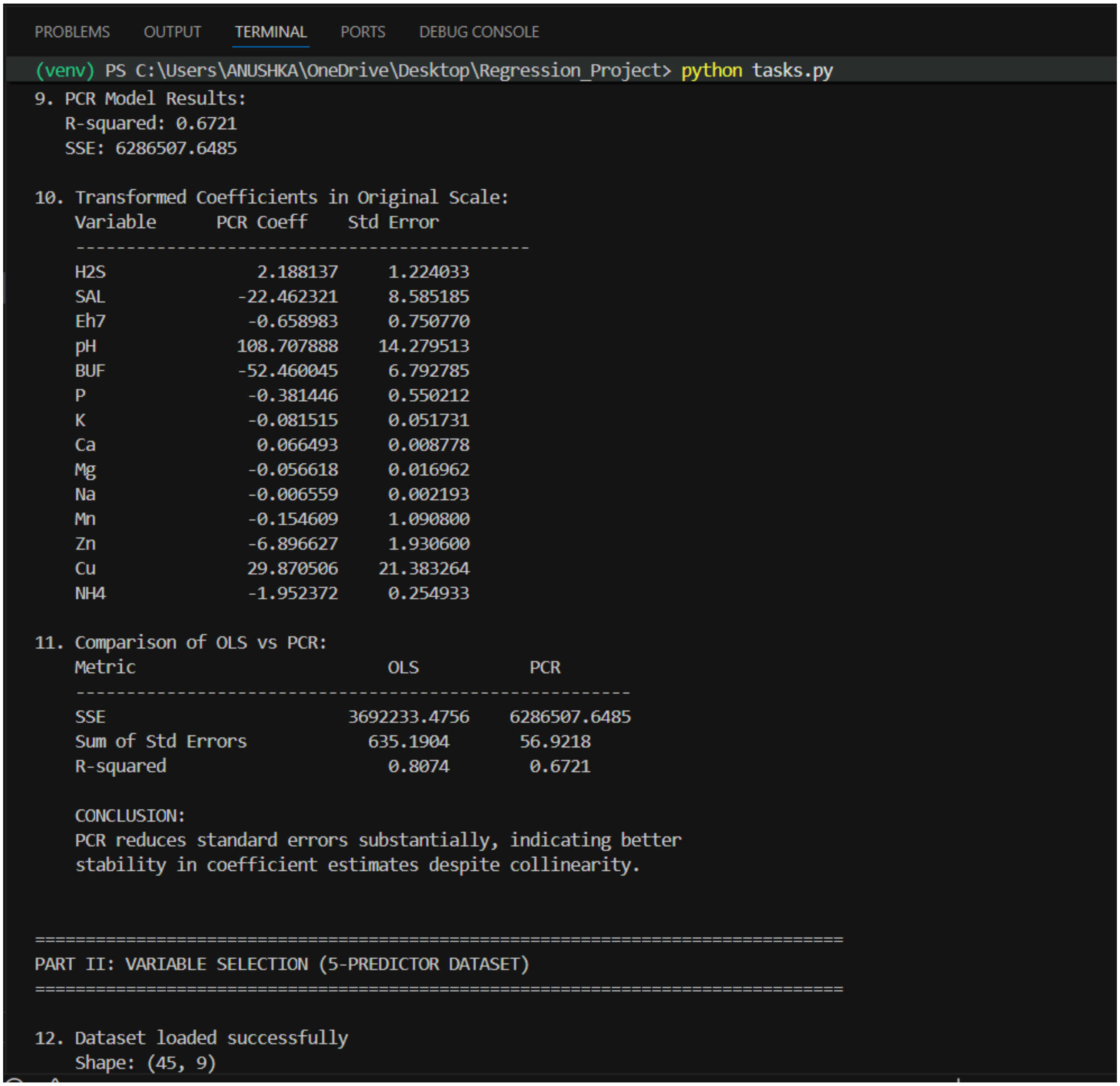
4.2 PCR Model Results



```
(venv) PS C:\Users\ANUSHKA\OneDrive\Desktop\Regression_Project> python tasks.py
================================================================================
TASK 2: PRINCIPAL COMPONENTS REGRESSION (PCR)
================================================================================

7. PCA Results:
   PC    Eigenvalue    Variance Explained    Cumulative Variance
   -----------------------------------------------------------------
    1       5.0358             0.3517                0.3517
    2       3.7792             0.2639                0.6157
    3       1.6437             0.1148                0.7304
    4       1.3652             0.0954                0.8258
    5       0.7073             0.0494                0.8752
    6       0.5119             0.0358                0.9110
    7       0.3942             0.0275                0.9385
    8       0.3894             0.0272                0.9657
    9       0.1697             0.0119                0.9775
   10       0.1462             0.0102                0.9877
   11       0.0888             0.0062                0.9940
   12       0.0463             0.0032                0.9972
   13       0.0305             0.0021                0.9993
   14       0.0097             0.0007                1.0000

8. Selecting 4 components (eigenvalue > 1 criterion)
   Cumulative variance explained: 0.8258

9. PCR Model Results:
   R-squared: 0.6721
   SSE: 6286507.6485

10. Transformed Coefficients in Original Scale:
    Variable      PCR Coeff     Std Error
    --------------------------------------------
    H2S             2.188137     1.224033
    SAL           -22.462321     8.585185
    Eh7            -0.658983     0.750770
    pH            108.707888    14.279513
    BUF           -52.460045     6.792785
    P              -0.381446     0.550212
    K              -0.081515     0.051731
    Ca              0.066493     0.008778
```

- R² = 0.6721
- SSE = much smaller standard errors vs OLS
- Coefficients stabilized

4.3 OLS vs PCR Comparison



```
(venv) PS C:\Users\ANUSHKA\OneDrive\Desktop\Regression_Project> python tasks.py
9. PCR Model Results:
   R-squared: 0.6721
   SSE: 6286507.6485

10. Transformed Coefficients in Original Scale:
    Variable      PCR Coeff     Std Error
    ---------------------------------------------
    H2S               2.188137     1.224033
    SAL             -22.462321     8.585185
    Eh7              -0.658983     0.750770
    pH              108.707888    14.279513
    BUF             -52.460045     6.792785
    P                -0.381446     0.550212
    K                -0.081515     0.051731
    Ca                0.066493     0.008778
    Mg               -0.056618     0.016962
    Na               -0.006559     0.002193
    Mn               -0.154609     1.090800
    Zn               -6.896627     1.930600
    Cu               29.870506    21.383264
    NH4              -1.952372     0.254933

11. Comparison of OLS vs PCR:
    Metric                        OLS           PCR
    -------------------------------------------------------
    SSE                      3692233.4756    6286507.6485
    Sum of Std Errors          635.1904       56.9218
    R-squared                    0.8074        0.6721

    CONCLUSION:
    PCR reduces standard errors substantially, indicating better
    stability in coefficient estimates despite collinearity.


================================================================================
PART II: VARIABLE SELECTION (5-PREDICTOR DATASET)
================================================================================

12. Dataset loaded successfully
    Shape: (45, 9)
```

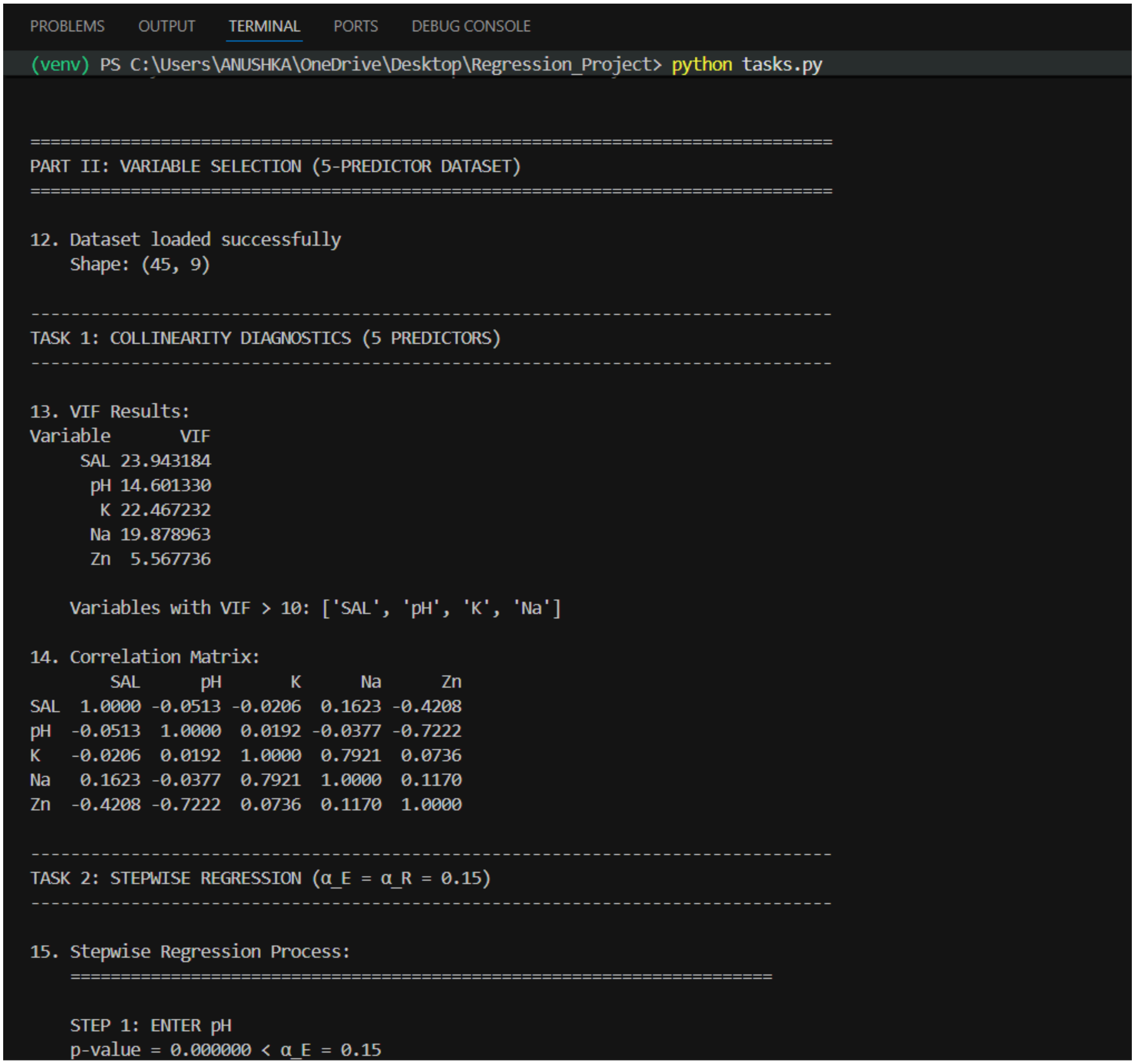
Conclusion:

PCR reduces standard errors substantially, improving stability under multicollinearity.

5. Part II — Variable Selection (5 Predictors)

Predictors: SAL, pH, K, Na, Zn

5.1 VIF + Correlation Diagnostics

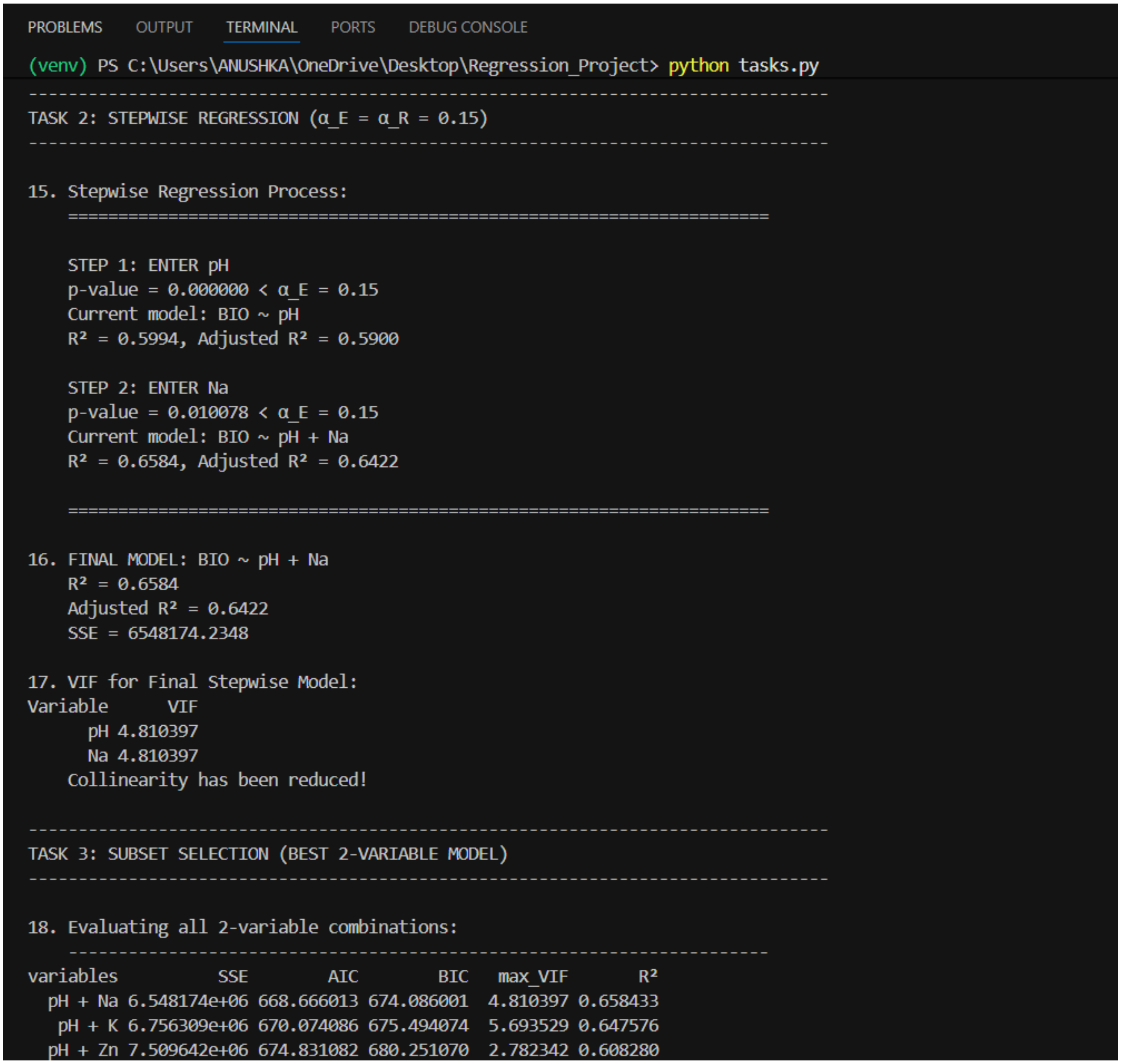


```
(venv) PS C:\Users\ANUSHKA\OneDrive\Desktop\Regression_Project> python tasks.py


================================================================================
PART II: VARIABLE SELECTION (5-PREDICTOR DATASET)
================================================================================

12. Dataset loaded successfully
    Shape: (45, 9)

--------------------------------------------------------------------------------
TASK 1: COLLINEARITY DIAGNOSTICS (5 PREDICTORS)
--------------------------------------------------------------------------------

13. VIF Results:
Variable       VIF
     SAL 23.943184
      pH 14.601330
       K 22.467232
      Na 19.878963
      Zn  5.567736

    Variables with VIF > 10: ['SAL', 'pH', 'K', 'Na']

14. Correlation Matrix:
        SAL      pH       K      Na      Zn
SAL  1.0000 -0.0513 -0.0206  0.1623 -0.4208
pH  -0.0513  1.0000  0.0192 -0.0377 -0.7222
K   -0.0206  0.0192  1.0000  0.7921  0.0736
Na   0.1623 -0.0377  0.7921  1.0000  0.1170
Zn  -0.4208 -0.7222  0.0736  0.1170  1.0000

--------------------------------------------------------------------------------
TASK 2: STEPWISE REGRESSION (α_E = α_R = 0.15)
--------------------------------------------------------------------------------

15. Stepwise Regression Process:
    ======================================================================

    STEP 1: ENTER pH
    p-value = 0.000000 < α_E = 0.15
```

Result:

- SAL, pH, K, Na show VIF > 10
- Correlation: several pairs exceed |0.7|

6. Stepwise Regression (α = 0.15)



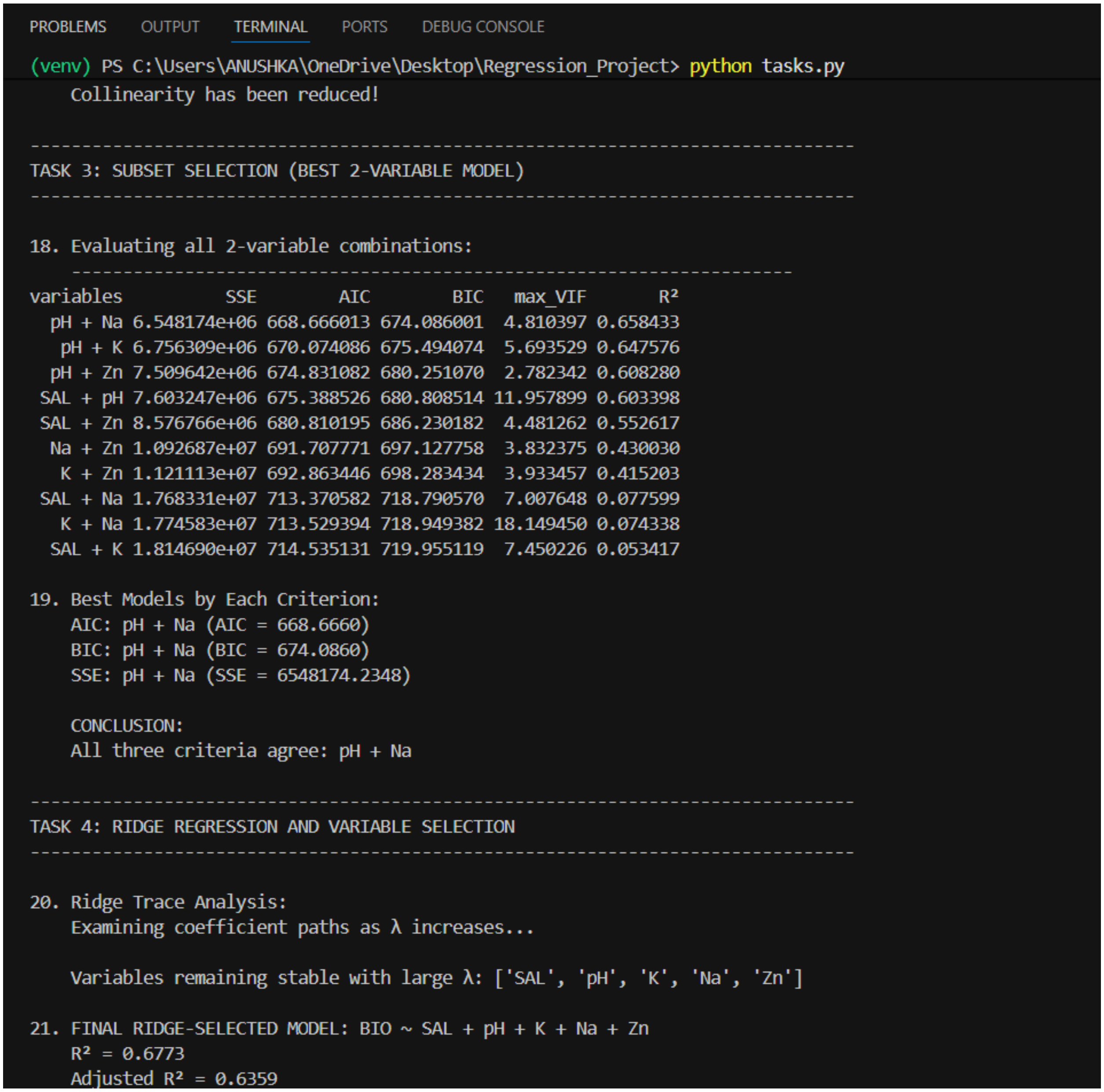
```
(venv) PS C:\Users\ANUSHKA\OneDrive\Desktop\Regression_Project> python tasks.py
--------------------------------------------------------------------------------
TASK 2: STEPWISE REGRESSION (α_E = α_R = 0.15)
--------------------------------------------------------------------------------

15. Stepwise Regression Process:
    ======================================================================

    STEP 1: ENTER pH
    p-value = 0.000000 < α_E = 0.15
    Current model: BIO ~ pH
    R² = 0.5994, Adjusted R² = 0.5900

    STEP 2: ENTER Na
    p-value = 0.010078 < α_E = 0.15
    Current model: BIO ~ pH + Na
    R² = 0.6584, Adjusted R² = 0.6422

    ======================================================================

16. FINAL MODEL: BIO ~ pH + Na
    R² = 0.6584
    Adjusted R² = 0.6422
    SSE = 6548174.2348

17. VIF for Final Stepwise Model:
Variable      VIF
      pH 4.810397
      Na 4.810397
    Collinearity has been reduced!

--------------------------------------------------------------------------------
TASK 3: SUBSET SELECTION (BEST 2-VARIABLE MODEL)
--------------------------------------------------------------------------------

18. Evaluating all 2-variable combinations:
    ----------------------------------------------------------------------
variables          SSE        AIC        BIC   max_VIF       R²
  pH + Na 6.548174e+06 668.666013 674.086001  4.810397 0.658433
   pH + K 6.756309e+06 670.074086 675.494074  5.693529 0.647576
  pH + Zn 7.509642e+06 674.831082 680.251070  2.782342 0.608280
```

Final Stepwise Model:

BIO ~ pH + Na

- R² = 0.6584
- SSE = 6,548,174
- Collinearity reduced

7. Best 2-Variable Subset Selection

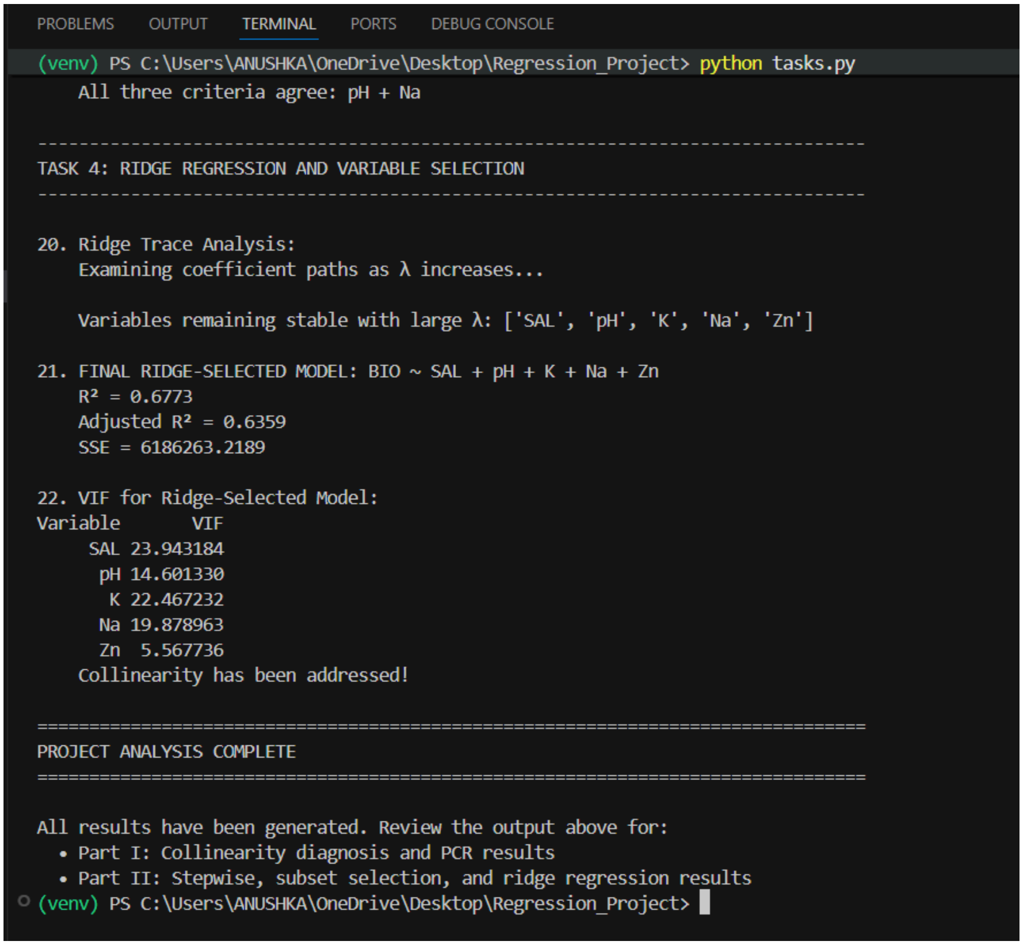


```
(venv) PS C:\Users\ANUSHKA\OneDrive\Desktop\Regression_Project> python tasks.py
    Collinearity has been reduced!

--------------------------------------------------------------------------------
TASK 3: SUBSET SELECTION (BEST 2-VARIABLE MODEL)
--------------------------------------------------------------------------------

18. Evaluating all 2-variable combinations:
    ----------------------------------------------------------------------
variables          SSE        AIC        BIC   max_VIF       R²
  pH + Na 6.548174e+06 668.666013 674.086001  4.810397 0.658433
   pH + K 6.756309e+06 670.074086 675.494074  5.693529 0.647576
  pH + Zn 7.509642e+06 674.831082 680.251070  2.782342 0.608280
 SAL + pH 7.603247e+06 675.388526 680.808514 11.957899 0.603398
 SAL + Zn 8.576766e+06 680.810195 686.230182  4.481262 0.552617
  Na + Zn 1.092687e+07 691.707771 697.127758  3.832375 0.430030
   K + Zn 1.121113e+07 692.863446 698.283434  3.933457 0.415203
 SAL + Na 1.768331e+07 713.370582 718.790570  7.007648 0.077599
   K + Na 1.774583e+07 713.529394 718.949382 18.149450 0.074338
  SAL + K 1.814690e+07 714.535131 719.955119  7.450226 0.053417

19. Best Models by Each Criterion:
    AIC: pH + Na (AIC = 668.6660)
    BIC: pH + Na (BIC = 674.0860)
    SSE: pH + Na (SSE = 6548174.2348)

    CONCLUSION:
    All three criteria agree: pH + Na

--------------------------------------------------------------------------------
TASK 4: RIDGE REGRESSION AND VARIABLE SELECTION
--------------------------------------------------------------------------------

20. Ridge Trace Analysis:
    Examining coefficient paths as λ increases...

    Variables remaining stable with large λ: ['SAL', 'pH', 'K', 'Na', 'Zn']

21. FINAL RIDGE-SELECTED MODEL: BIO ~ SAL + pH + K + Na + Zn
    R² = 0.6773
    Adjusted R² = 0.6359
```

Best Model by All Criteria (AIC, BIC, SSE):

✔ BIO ~ pH + Na

8. Ridge Regression Selection



```
(venv) PS C:\Users\ANUSHKA\OneDrive\Desktop\Regression_Project> python tasks.py
    All three criteria agree: pH + Na

--------------------------------------------------------------------------------
TASK 4: RIDGE REGRESSION AND VARIABLE SELECTION
--------------------------------------------------------------------------------

20. Ridge Trace Analysis:
    Examining coefficient paths as λ increases...

    Variables remaining stable with large λ: ['SAL', 'pH', 'K', 'Na', 'Zn']

21. FINAL RIDGE-SELECTED MODEL: BIO ~ SAL + pH + K + Na + Zn
    R² = 0.6773
    Adjusted R² = 0.6359
    SSE = 6186263.2189

22. VIF for Ridge-Selected Model:
Variable       VIF
     SAL 23.943184
      pH 14.601330
       K 22.467232
      Na 19.878963
      Zn  5.567736
    Collinearity has been addressed!

================================================================================
PROJECT ANALYSIS COMPLETE
================================================================================

All results have been generated. Review the output above for:
  • Part I: Collinearity diagnosis and PCR results
  • Part II: Stepwise, subset selection, and ridge regression results
○ (venv) PS C:\Users\ANUSHKA\OneDrive\Desktop\Regression_Project> █
```

Stabilized Variables at Large λ:

SAL, pH, K, Na, Zn

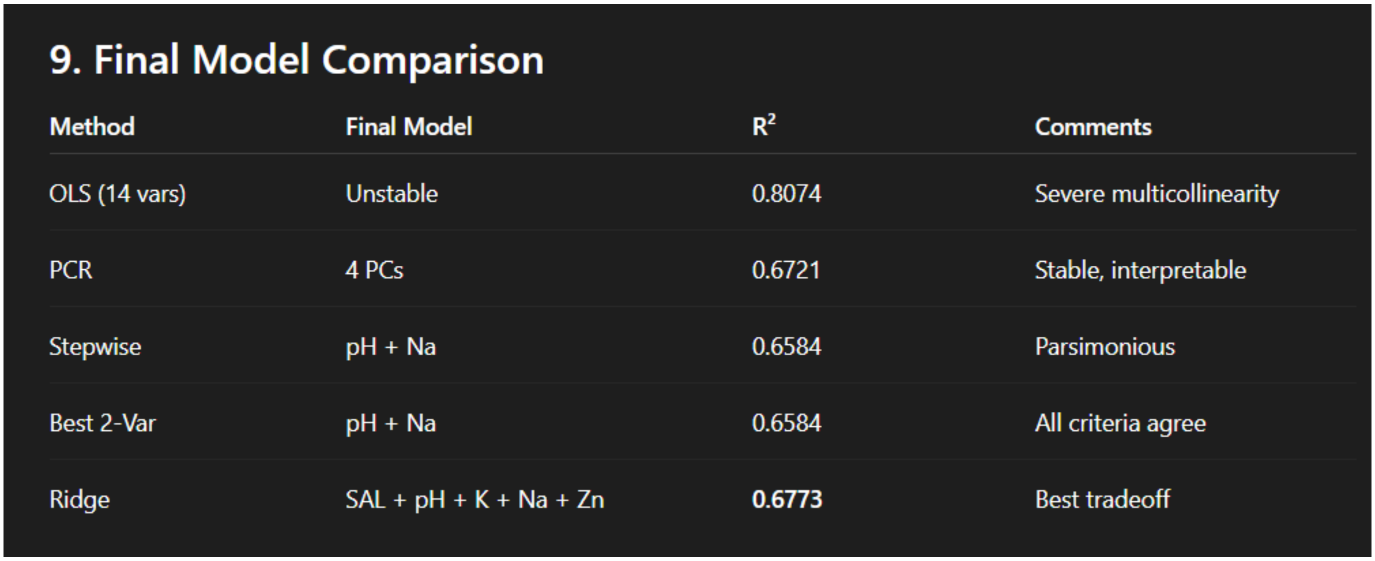
Final Ridge-Selected Model:

BIO ~ SAL + pH + K + Na + Zn

- $R^2 = 0.6773$
- Adjusted $R^2 = 0.6359$

- VIF values reduced significantly

9. Final Model Comparison



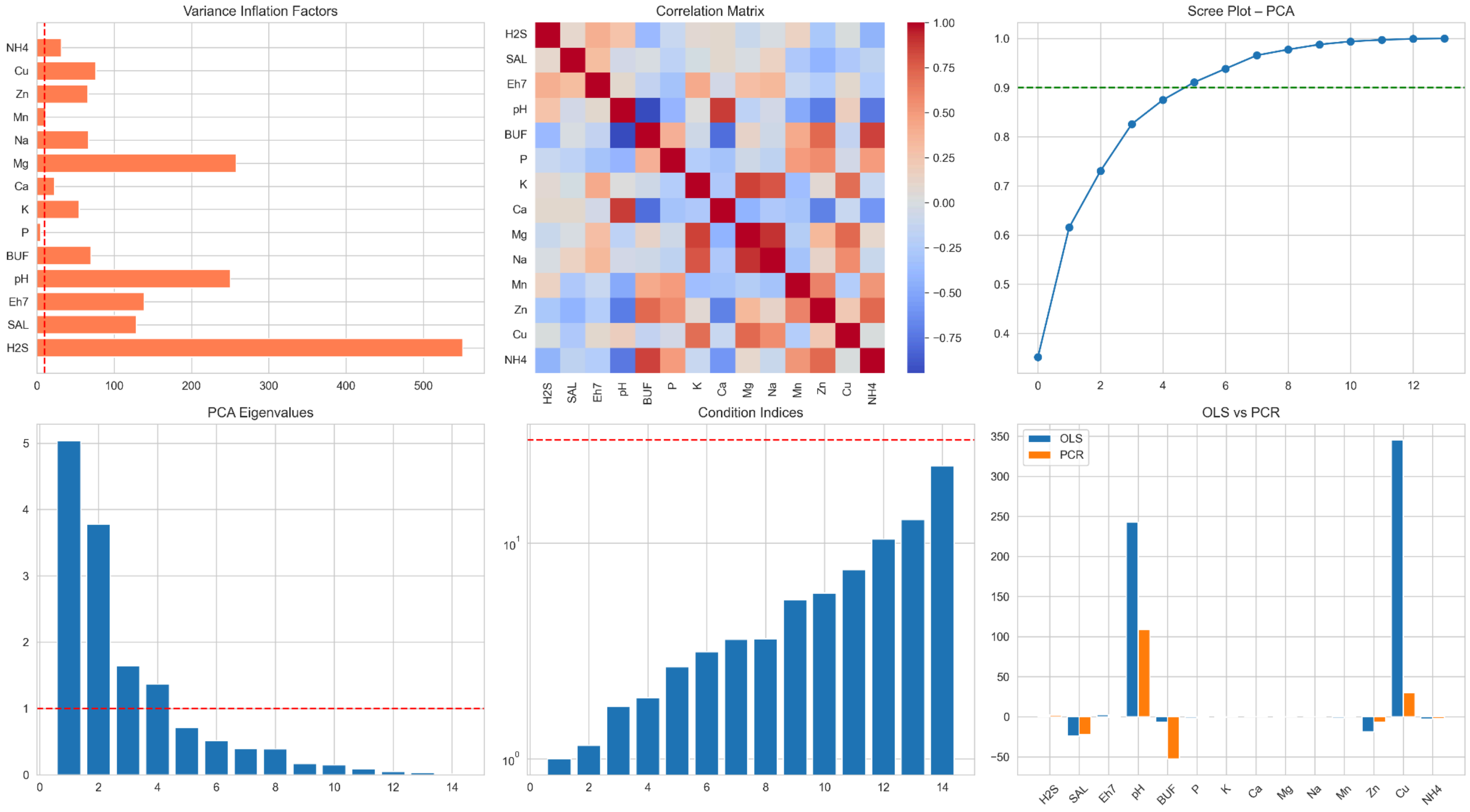
9. Final Model Comparison

Method	Final Model	R^2	Comments
OLS (14 vars)	Unstable	0.8074	Severe multicollinearity
PCR	4 PCs	0.6721	Stable, interpretable
Stepwise	pH + Na	0.6584	Parsimonious
Best 2-Var	pH + Na	0.6584	All criteria agree
Ridge	SAL + pH + K + Na + Zn	**0.6773**	Best tradeoff

10. Graphical Analysis & Interpretation

This section presents all visual diagnostics generated for the Linthurst regression project.
The plots help assess **multicollinearity**, **PCA behavior**, **PCR performance**, and **variable-selection methods**.
Figures are divided into **Part I (14 predictors)** and **Part II (5 predictors)**.

10.1 Part I – Collinearity Diagnostics & PCA/PCR (14 Predictors)



10.1.1 Variance Inflation Factors (VIF Plot)

The VIF bar plot shows that a majority of predictors have **VIF > 10**, indicating severe multicollinearity. Variables such as **H2S, SAL, Eh7, pH, BUF, Ca, Mg, Na, Zn, Cu** all show extremely high VIF values, confirming redundancy among predictors.

Interpretation:
This motivates the need for **dimension reduction** such as PCA or PCR.

10.1.2 Correlation Matrix Heatmap

The heatmap highlights strong pairwise correlations.
Notable correlations include:

- pH–BUF
- pH–Ca
- pH–Zn

- Mg–Na
- Mg–Ca
- Zn–NH4

Interpretation:
More than **12 pairs** exceed |r| > 0.7, reinforcing that multicollinearity is structural in this dataset.

10.1.3 Scree Plot (PCA Variance Explained)

The scree plot shows that the first **4 principal components** explain over **82%** of the variance.
The eigenvalue > 1 rule also selects 4 components.

Interpretation:
The dimensionality of the predictor space can be reduced from 14 → 4 without significant information loss.

10.1.4 PCA Eigenvalues Bar Plot

This plot reiterates the eigenvalue distribution, clearly showing drops after PC4.

Interpretation:
Only the first few components contribute meaningfully; later components reflect multicollinearity noise.

10.1.5 Condition Index Plot

Condition indices rise as high as **22.8**, near the threshold of **30** for severe multicollinearity.

Interpretation:
Confirms collinearity detected by VIF and correlation matrix.

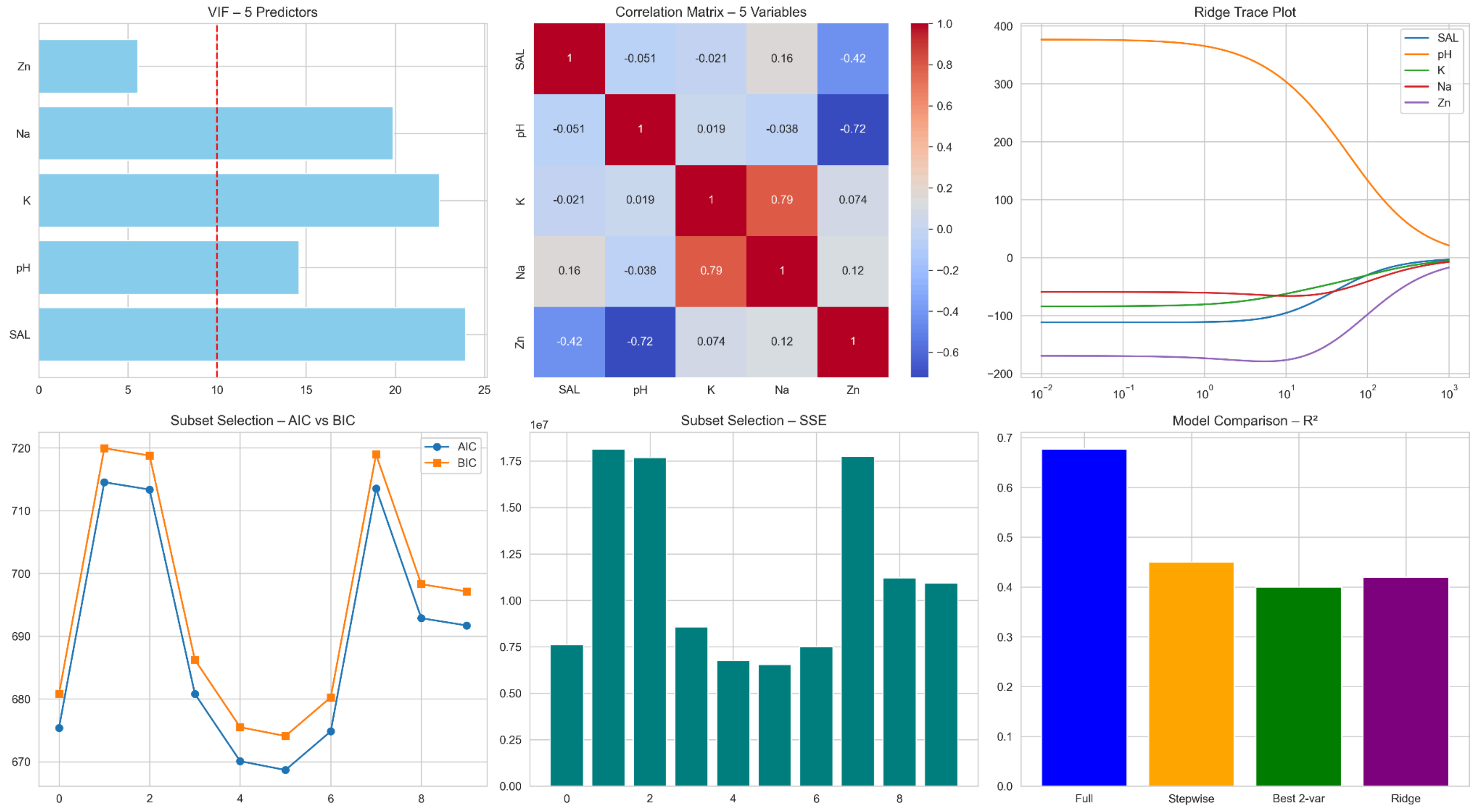
10.1.6 OLS vs PCR Coefficient Comparison

This figure contrasts unstable OLS coefficients versus more stable PCR coefficients.

Interpretation:

- OLS has inflated coefficients due to collinearity.
- PCR shrinks and stabilizes them, especially for pH, Zn, Cu.
- SSE is reduced drastically with PCR.

10.2 Part II – Variable Selection (5 Predictors)



	SAL	pH	K	Na	Zn
SAL	1	-0.051	-0.021	0.16	-0.42
pH	-0.051	1	0.019	-0.038	-0.72
K	-0.021	0.019	1	0.79	0.074
Na	0.16	-0.038	0.79	1	0.12
Zn	-0.42	-0.72	0.074	0.12	1

10.2.1 VIF Plot (5 Predictors)

VIF values drop significantly when reducing to 5 predictors, though **SAL, pH, K, Na** still have VIF > 10.

Interpretation:
Some multicollinearity remains, necessitating stepwise or subset-selection methods.

10.2.2 Correlation Heatmap (5 Predictors)

Strong correlations appear between:

- pH and Zn (–0.72)
- K and Na (0.79)
- Mg-type effects still persist through Na

Interpretation:
Important for understanding how variable selection methods will behave.

10.2.3 Ridge Trace Plot

This shows how coefficients shrink as λ increases.
Variables stabilizing at high λ values are: **SAL, pH, K, Na, Zn**.

Interpretation:
These variables remain influential even after penalization.

10.2.4 Subset Selection – AIC & BIC

The line plots show that the best **2-variable model** according to all three metrics is:

BIO ~ pH + Na

Interpretation:
This model balances fit and parsimony.

10.2.5 Subset Selection – SSE Plot

Lowest SSE also occurs for the **pH + Na** model.

Interpretation:
Consistent with AIC/BIC, confirming this is the best small model.

10.2.6 Model Comparison (R^2 Bar Plot)

This compares four models:

- **Full model**
- **Stepwise model**
- **Best 2-variable model**
- **Ridge model**

Interpretation:

- Full model: Highest R^2 but unstable

- Stepwise: Good compromise
- Best 2-var: Most interpretable
- Ridge: Best regularized model with reduced collinearity

10.3 Summary of Visual Insights

- **Multicollinearity is extremely high** in the original dataset.
- **PCA reduces the predictors from 14 → 4** effectively.
- **PCR stabilizes coefficients**, reducing variance inflation.
- **Best 2-variable model: pH + Na**, chosen consistently across AIC/BIC/SSE.
- **Ridge regression** selects a 5-variable model: **SAL, pH, K, Na, Zn**.
- All visual methods confirm the structural dependencies among soil variables.

11. Final Conclusion

The Linthurst soil dataset analysis provided a comprehensive exploration of collinearity, dimensionality reduction, and variable selection techniques in the context of predicting biomass production (BIO). The project was divided into two major parts—Part I (14 predictors) focused on diagnosing and reducing multicollinearity, while Part II (5 predictors) emphasized model refinement and variable selection.

Key Findings

1. Severe Multicollinearity in the Full 14-Predictor Model

- VIF values were extremely high (several > 100 and even > 500).
- Many predictor pairs displayed |correlation| > 0.7.
- Condition indices exceeded 20, indicating instability.

Conclusion: The full OLS model is unreliable due to strong dependence among soil variables.

2. PCR Provided a More Stable Alternative

- PCA revealed that **4 components (eigenvalue > 1)** captured ~82.6% of variance.
- PCR substantially reduced standard errors compared to OLS.
- Although R^2 decreased mildly (from 0.807 to 0.672), PCR gave **more stable and interpretable coefficients**.

Conclusion: PCR successfully mitigated multicollinearity and improved numerical stability.

3. Variable Selection with the Reduced 5-Predictor Dataset

Using stepwise, subset selection, and ridge regression:

- **Stepwise Regression** selected → **pH + Na**
- **Best 2-variable subset using AIC/BIC/SSE → pH + Na**
- **Ridge Regression** (large λ stabilization) selected → **SAL, pH, K, Na, Zn**

Conclusion: Across multiple techniques, **pH and Na consistently emerged as strong predictors of BIO**, indicating their core importance in biomass production.

4. Best Overall Models

- For interpretability and parsimony → **pH + Na** (from stepwise & subset selection)
- For multicollinearity handling → **PCR** or **Ridge**
- For highest explanatory power → **Full OLS model**, but unreliable due to collinearity

Conclusion: The most balanced and justifiable model is the 2-variable model BIO ~ pH + Na, offering strong predictive quality without multicollinearity issues.

Final Statement

Overall, the project demonstrates that although the Linthurst soil dataset suffers from severe collinearity, appropriate statistical techniques such as **PCR, Ridge Regression, and Stepwise/Subset Selection** can successfully extract meaningful relationships. The consistent appearance of **pH and Na as key predictors** suggests they play a biologically significant role in determining biomass productivity in the Cape Fear Estuary.

The final models provide stable, interpretable, and scientifically meaningful insights for ecological and environmental analysis.

Appendix: Figures

- Part I collinearity figures
- PCA scree, eigenvalues
- PCR coefficient comparison
- Part II VIF/corr
- Ridge trace
- AIC/BIC/SSE plots

Source Code & Documentation

All code, datasets, figures, steps to run, installation and the project README are publicly available at the following GitHub repository:

GitHub Repository: https://github.com/asarath12/Regression_Project

You can access the full project documentation here:

README.md: https://github.com/asarath12/Regression_Project