Shivan Ajay Iyer be22b048

# Predicting Fossil Age with a Regression Model

## Assignment 2

AIBD_as_2.ipynb

# Introduction

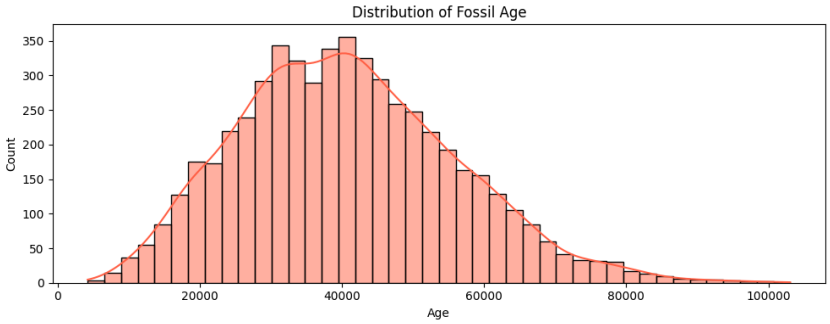
The provided code implements a regression analysis pipeline to predict fossil age using geological and biological features provided in the dataset- uranium_lead_ratio carbon_14_ratio, radioactive_decay_series, stratigraphic_layer_depth, inclusion_of_other_fossils, isotopic_composition, fossil_size, fossil_weight. The following information aids in interpreting the results from the code and plots, facilitating a deeper understanding of the data distribution

# Pipeline Outline

The code in the attached notebook follows these stages :

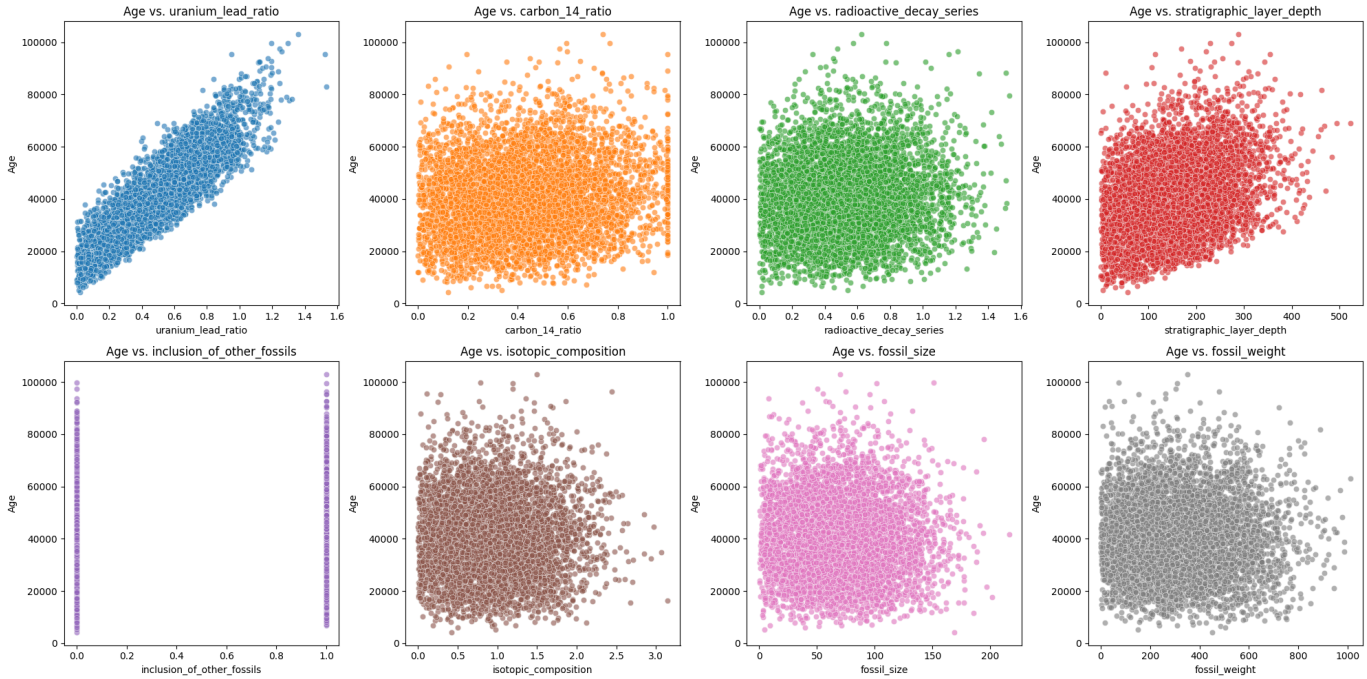
### Data Preparation

- Loading & Inspection: The dataset was loaded from CSV, and basic statistics were generated.
- Preprocessing:
    - Conversion of boolean column inclusion_of_other_fossils to numeric (0/1).
    - Separation of features (X) and target (age).
- Visualisation: Age distribution analysis using histograms and feature-age scatterplots.



## Model Development

- Train-Test Splits: Two ratios tested (70-30 and 80-20).
- Baseline Model: Linear regression without regularisation.
- Regularised Model: Ridge regression (*L2 regularisation*) with lambda (default parameter 'alpha' in the code) tuning via *GridSearchCV* and *KFold cross-validation*.




## Evaluation

- Metrics: MAE, MSE, RMSE, $R^2$.
- Comparison: Regularized vs. non-regularized models.

## Results Presentation

- Actual vs. predicted age plots.

## Assumptions

1. Linearity: Relationships between features and targets are linear
2. Feature Relevance: All 8 features (e.g., uranium-lead ratio) contribute meaningfully to age prediction.
3. Normalisation: Implicitly handled by Ridge regularisation (no explicit scaling in code).
4. Independence: Observations (fossils) are independent.

### Why Ridge Over Lasso?

Multicollinearity: The  Geological features (e.g., isotopic composition and radioactive decay series) are often correlated. Ridge preserves all features while shrinking coefficients, whereas Lasso might eliminate critical variables.

Fossil dating typically benefits from collective feature contributions rather than sparse models. [1]

## Question 1: Model Building with Splits

### Splitting Strategy:

Generally, it is recommended that over 70 % of the data be used to train the model. [2]

70-30 Split: Larger test set for robust generalisation assessment.

80-20 Split: Maximizes training data for parameter learning.

With and without regularisation, the 70-30 split performs marginally better. This could be due to:

- Randomness in Data Splitting
- Overfitting vs. Generalization

- Variability due to the small size of the dataset

### Insight:

Ridge marginally outperforms simple linear regression, suggesting regularisation mitigates overfitting.

```
Model Comparison:
                        Model           MSE         RMSE          MAE        R²
0  Linear Regression (80-20)  1.864761e+07  4318.287491  3525.103738  0.919156
1  Linear Regression (70-30)  1.859305e+07  4311.966330  3547.562942  0.919900
2   Ridge Regression (80-20)  1.864706e+07  4318.224063  3525.098103  0.919159
3   Ridge Regression (70-30)  1.859314e+07  4311.976699  3547.640389  0.919900
4    Optimized Ridge (80-20)  1.864674e+07  4318.186660  3525.100799  0.919160
5    Optimized Ridge (70-30)  1.859316e+07  4311.978604  3547.652402  0.919900
```

## Question 2: Cross-Validation for Lambda Tuning

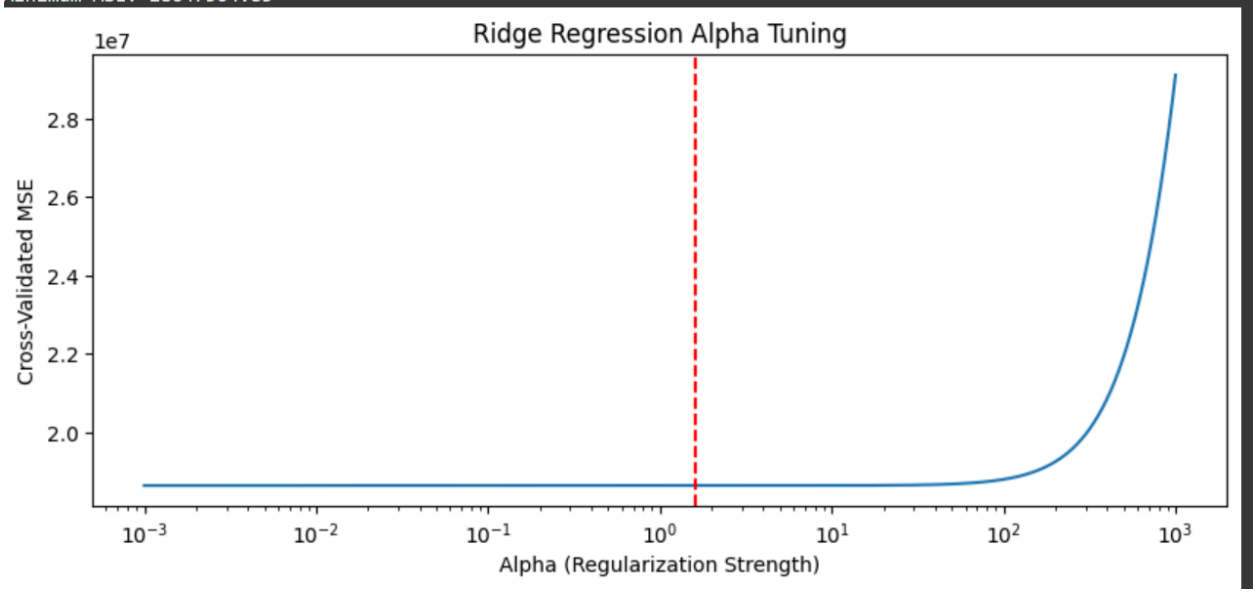
### Methods:

*KFold* (k=5): Reduces variance in lambda selection:

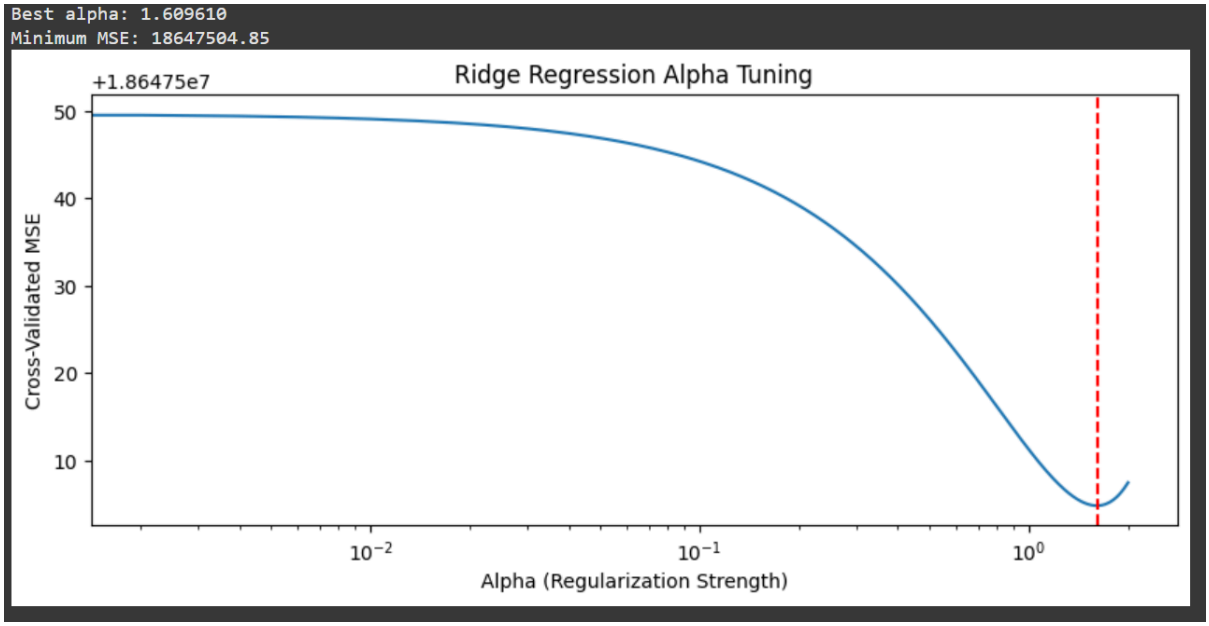
Initially, course parameter optimization was performed using logspace to explore a broad spectrum of values with a substantial step size, allowing for greater computational efficiency. Following this, a more precise approach was employed using linspace with narrower ranges and smaller step sizes to enhance accuracy in the calculations.

Presented below are the figures illustrating the optimal alpha value obtained through both coarse and fine time steps for the 80-20 split experiment, respectively.

```
Best alpha: 1.611414
Minimum MSE: 18647504.85
```




```
Best alpha: 1.609610
Minimum MSE: 18647504.85
```



*GridSearchCV*:

This was used after establishing a tight range (0,2) using K-fold validation since GridSearchCV more computationally expensive. I have described the process for finding the best value of the hyperparameter lambda as 'alpha' as that is the variable name in the code.

For each of the five folds in cross-validation, GridSearchCV does the following:

1. Splits the data into a training set (4 folds) and a validation set (1 fold).
2. For every alpha value in alpha_range, it:
    - Trains the Lasso model on the four folds.
    - Evaluate it on the one validation fold using negative MSE.
3. Repeat this for all five-fold combinations.
4. Averages the scores for each alpha across the five folds.
5. Picks the alpha with the best average score

```
Best alpha from Grid Search (80-20): 1.616162
Best CV score: 18647504.85
Best alpha from Grid Search (70-30): 1.151515
Best CV score: 18671983.15
```

## Question 3: Model Quality Assessment

Metrics:

Linear Model without Regularisation:

```
Linear Regression (80-20 split) Performance Metrics:
MSE: 18647606.85
RMSE: 4318.29
MAE: 3525.10
R²: 0.9192


Linear Regression (70-30 split) Performance Metrics:
MSE: 18593053.64
RMSE: 4311.97
MAE: 3547.56
R²: 0.9199
```

Ridge model: Lower MAE/MSE and higher $R^2$ vs. linear regression.

```
Optimized Ridge Regression (80-20 split) Performance Metrics:
MSE: 18646736.03
RMSE: 4318.19
MAE: 3525.10
R²: 0.9192


Optimized Ridge Regression (70-30 split) Performance Metrics:
MSE: 18593159.49
RMSE: 4311.98
MAE: 3547.65
R²: 0.9199
```

Conclusion: The Ridge model is superior due to better generalisation (lower error metrics).

## Question 4: Actual vs. Predicted Age

- Visualisation: Scatterplot with diagonal line (ideal predictions).
- Observation: Predictions cluster tightly around the line, indicating high accuracy