“

Predicting the Age of Abalone Using Regression Models

A Comparative Study of Linear, Ridge, and Lasso Regression

”

MUHAMMAD MUBASHAR SHAHZAD

Registration No. SM3600012

Scientific and Data Intensive Computing (SDIC)

University of Trieste

Introduction

- Objective: To predict the age of abalone using physical measurements.
- Dataset: Abalone dataset with 4177 instances and 8 features.
- Target Variable: Rings (predicting Age by adding 1.5).

Dataset Overview

- Features:
  - Sex (Categorical)
  - Length (Continuous)
  - Diameter (Continuous)
  - Height (Continuous)
  - Whole weight (Continuous)
  - Shucked weight (Continuous)
  - Viscera weight (Continuous)
  - Shell weight (Continuous)
- Target: Rings (Integer, Age = Rings + 1.5)

Data Loading and Initial Exploration

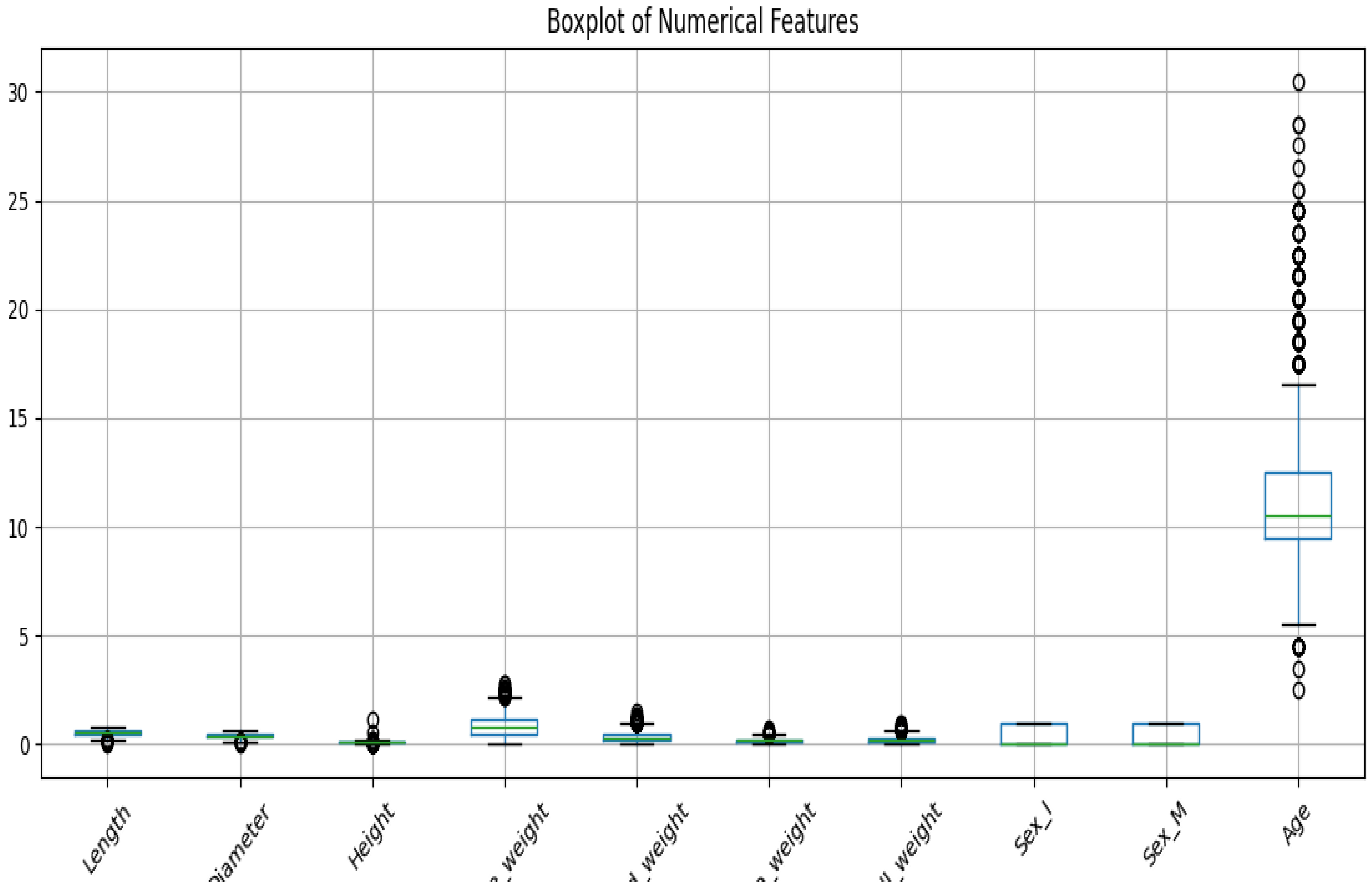
- Step 1: Loading the dataset using Pandas.
- Code Snippet: `pd.read_csv('abalone.csv')`
- Initial DataFrame: Show the first few rows of the dataset.
- Handling Missing Values: Mention that there are no missing values in the dataset.

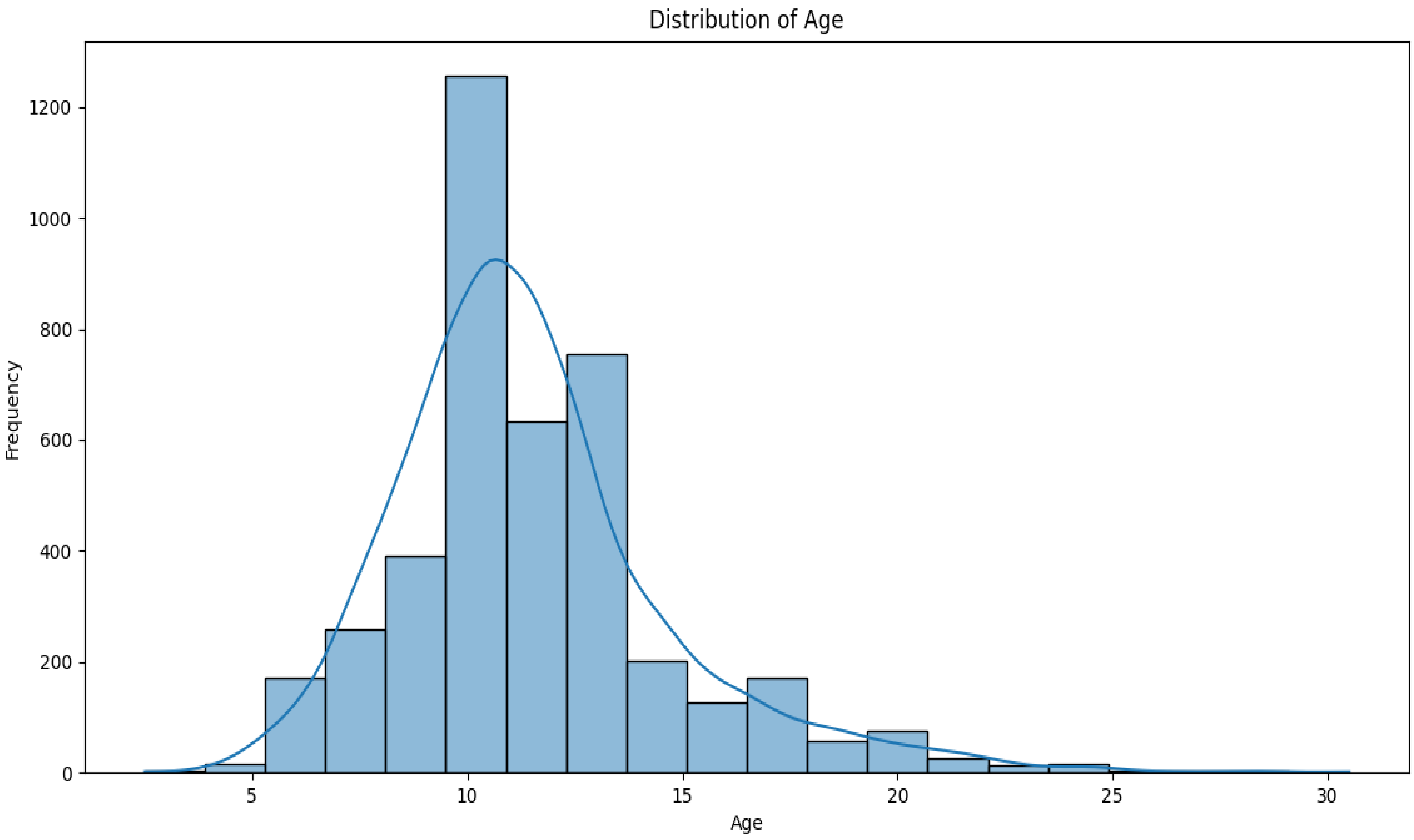
Data Preprocessing

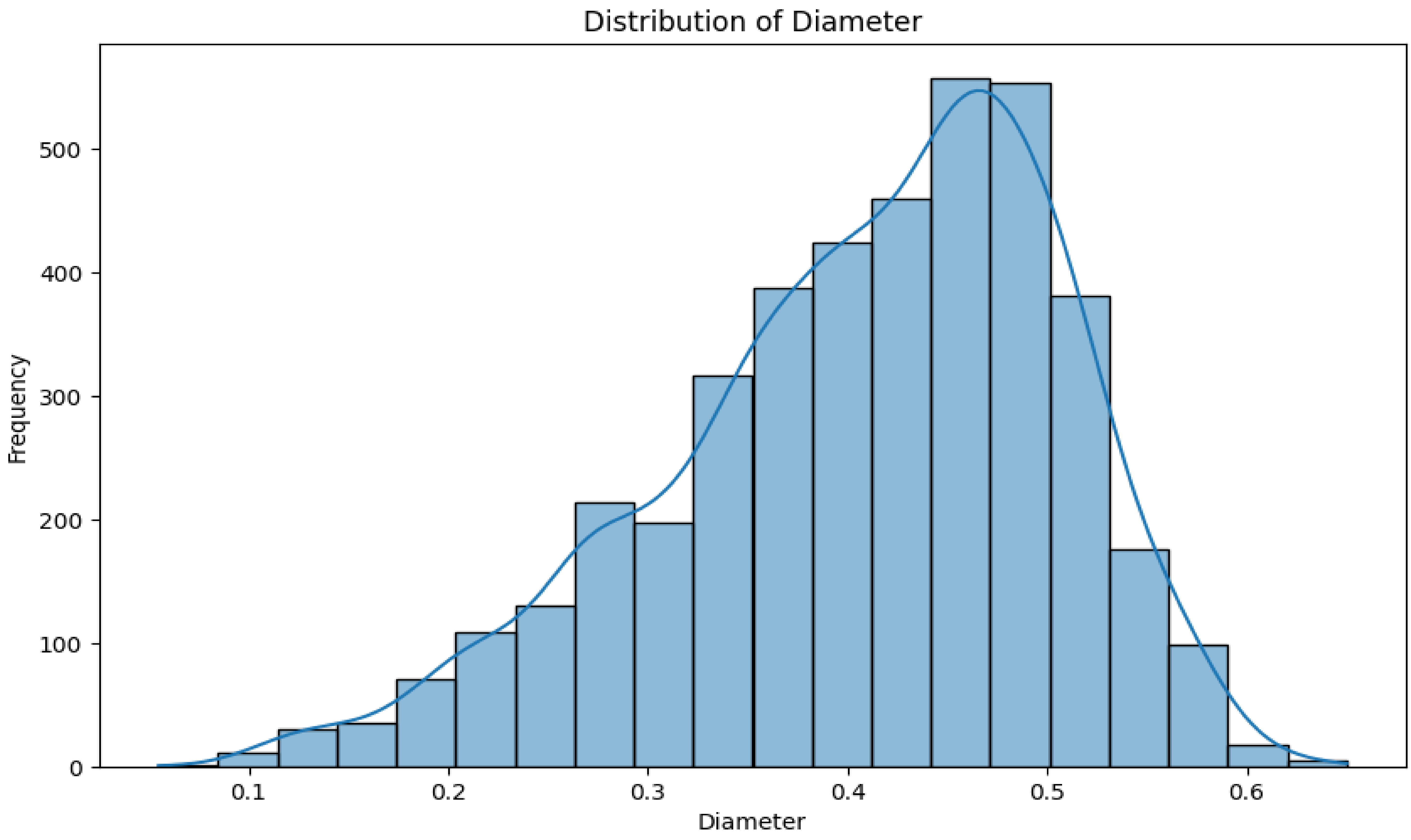
- Step 2: Converting Categorical Variable
- Sex column converted to numerical using one-hot encoding.
- Step 3: Creating Age Column
- Adding 1.5 to the Rings column to get the age.

Exploratory Data Analysis (EDA)

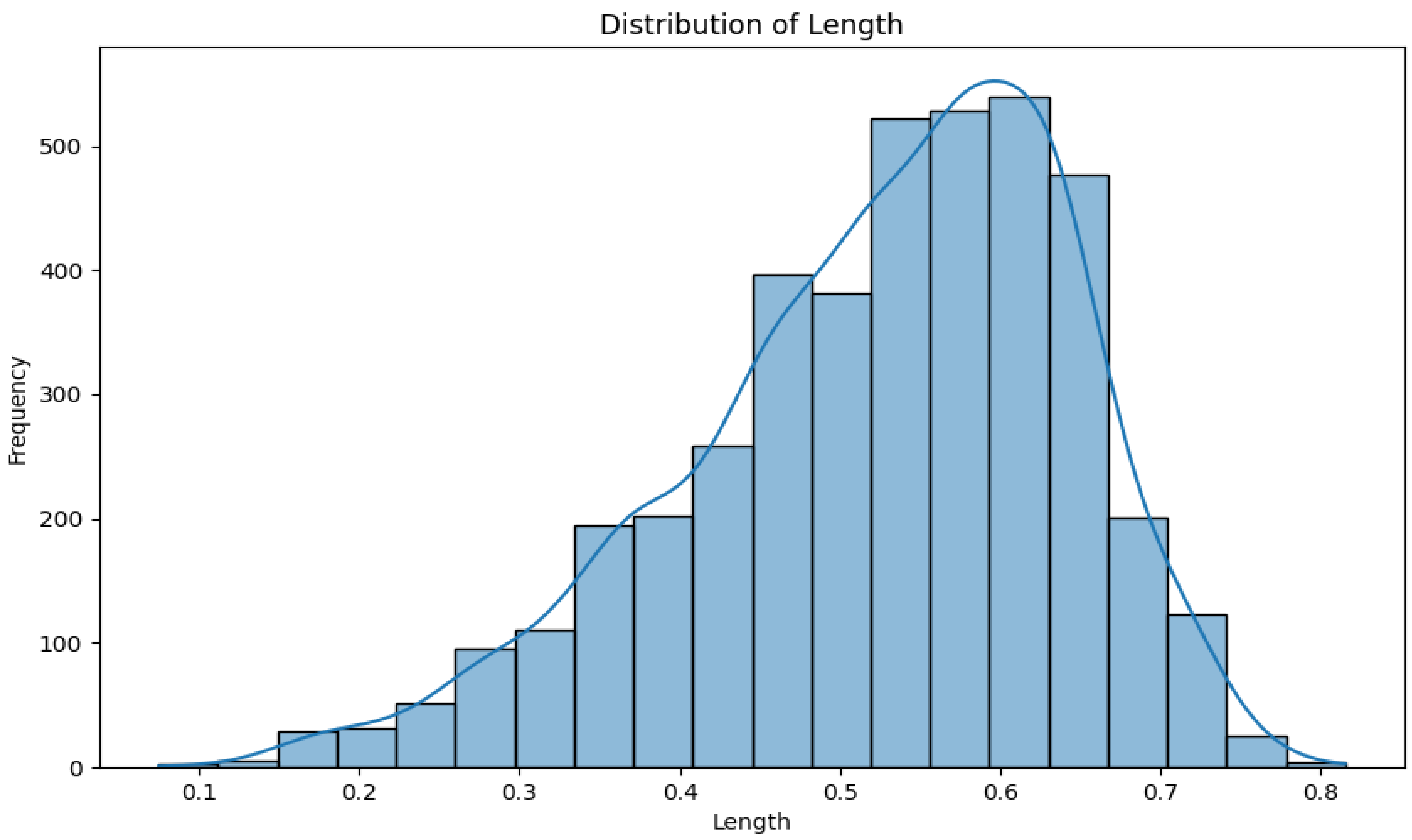
- Step 4: Visualizations
- Histograms of numerical features.
- Box plots to check for outliers.
- Step 5: Correlation Analysis
- Heatmap of correlation matrix to understand relationships between features and target.


Boxplot of Numerical Features
30
25
20
15
10
5
0
Length
Height
Sex_I
Sex_M
Age

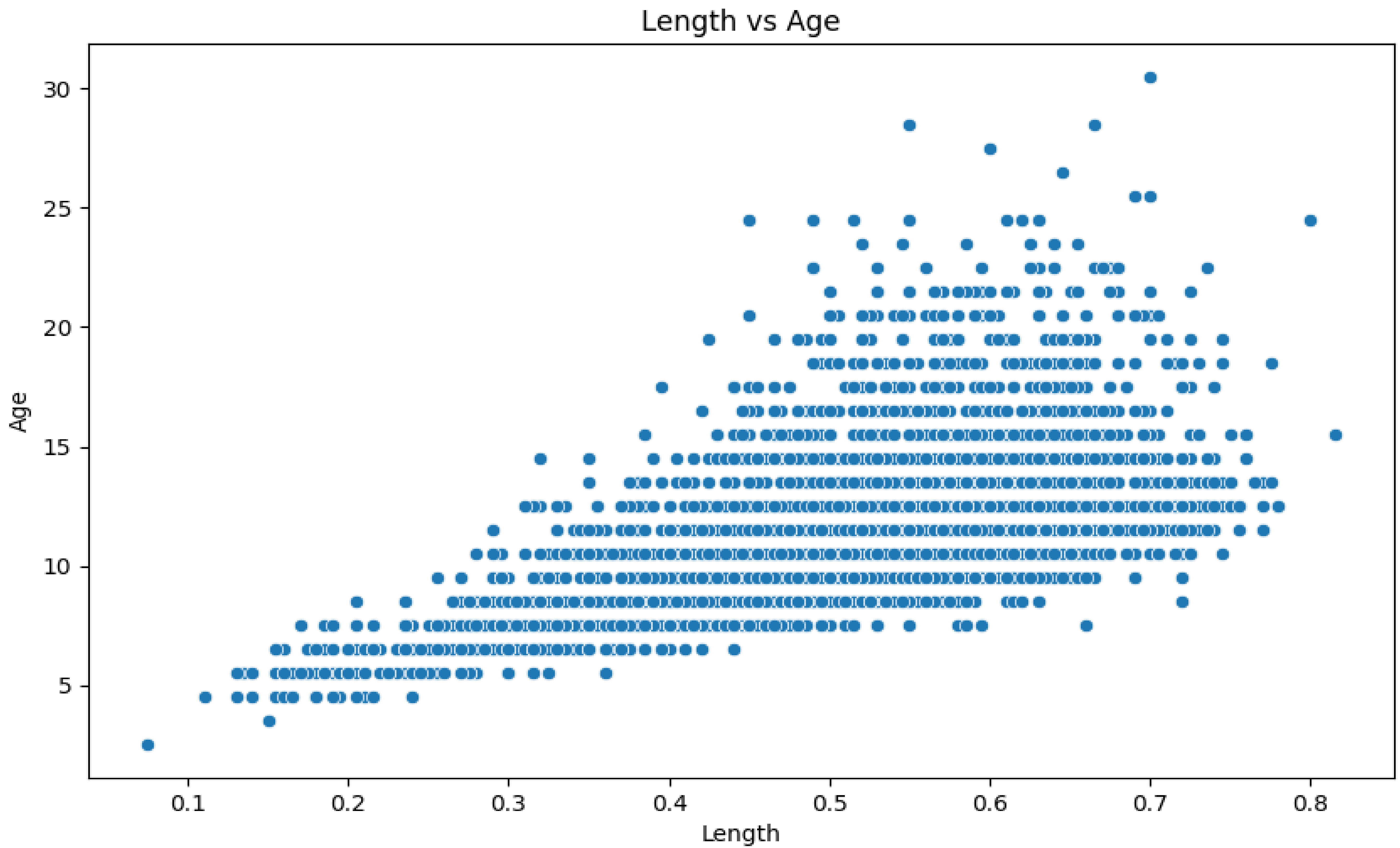

Distribution of Age
Frequency
0
200
400
600
800
1000
1200
5
10
15
20
25
30
Age

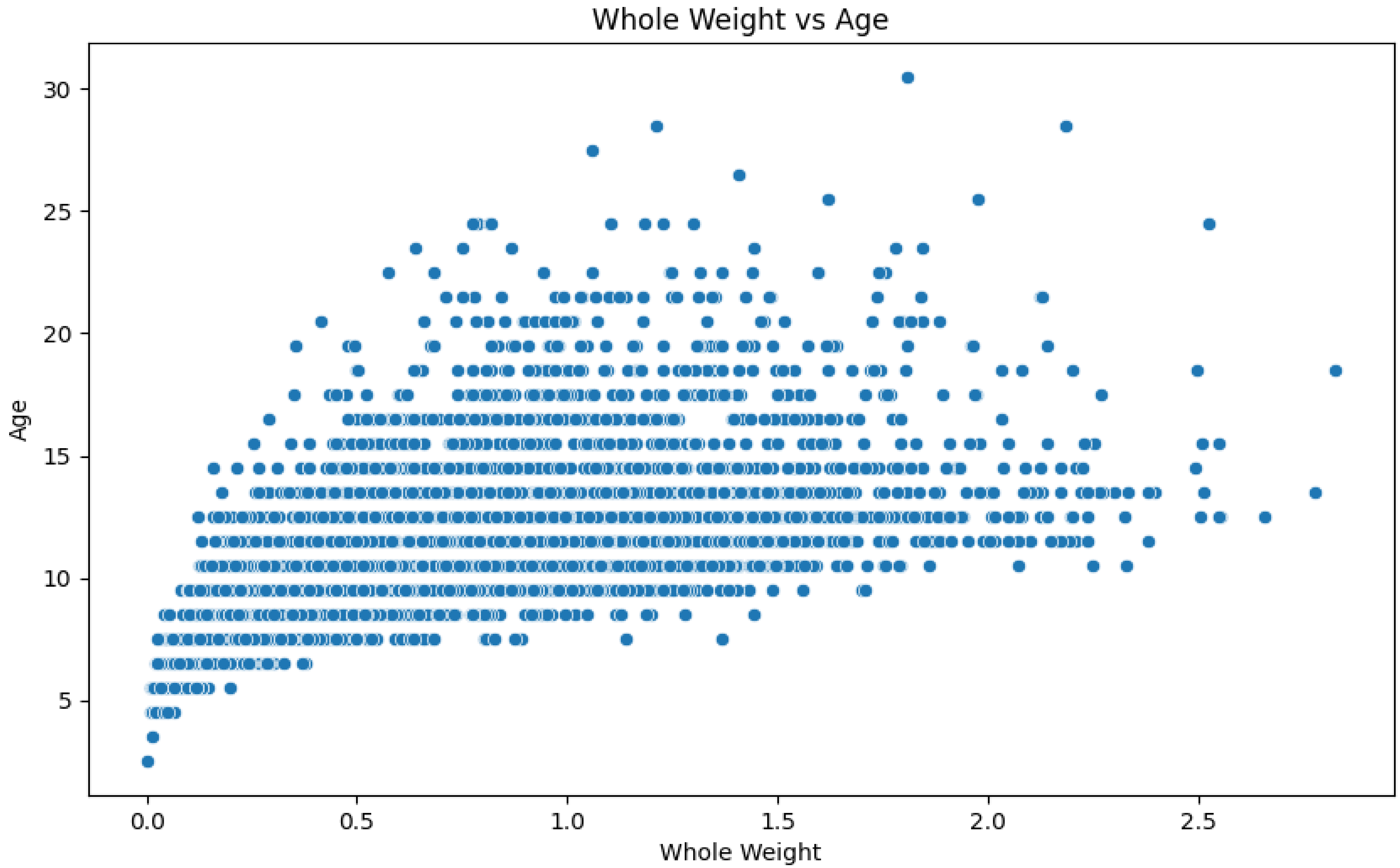


Distribution of Diameter
Frequency
0
100
200
300
400
500
0.1
0.2
0.3
0.4
0.5
0.6
Diameter

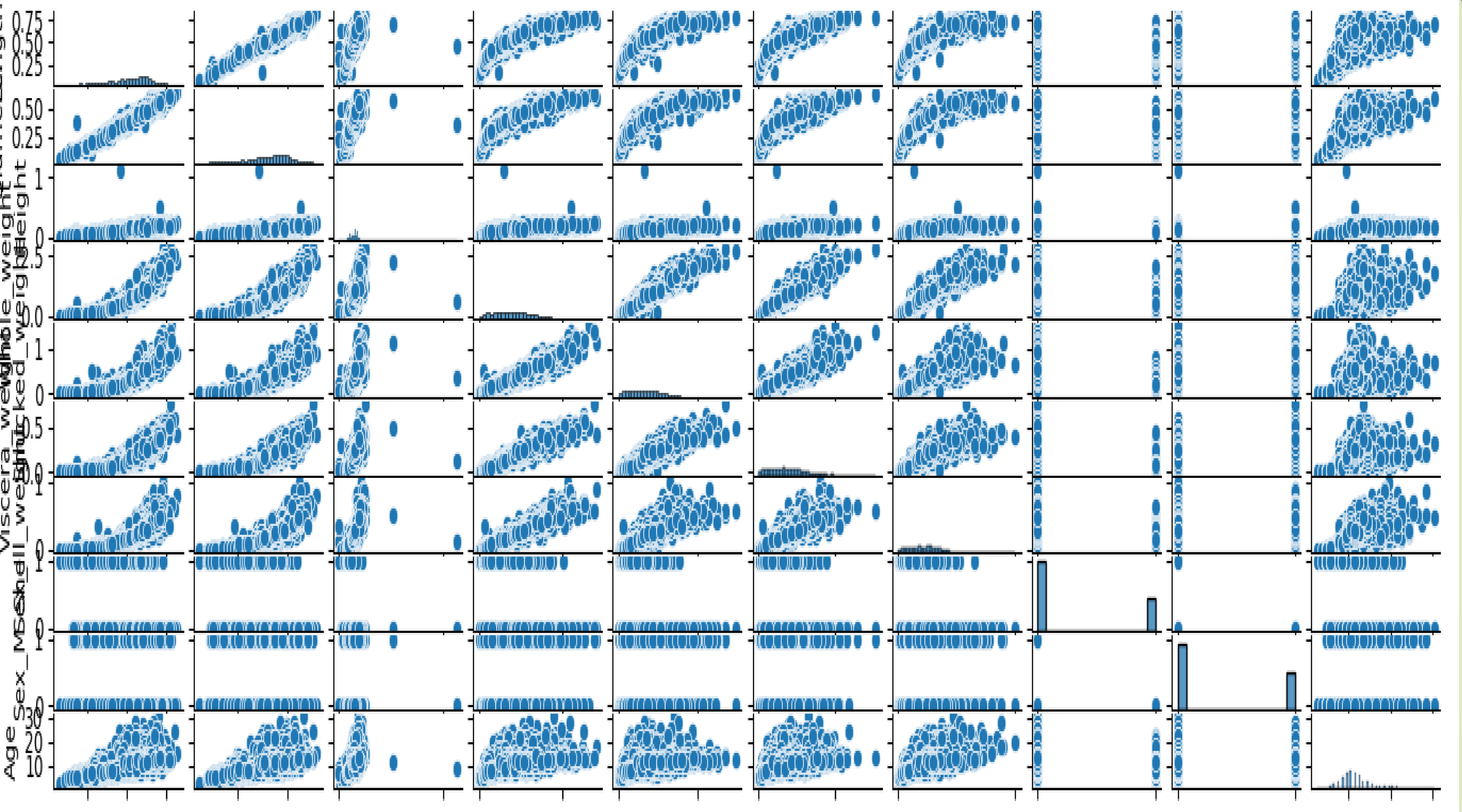


Distribution of Length
Frequency
0
100
200
300
400
500
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Length

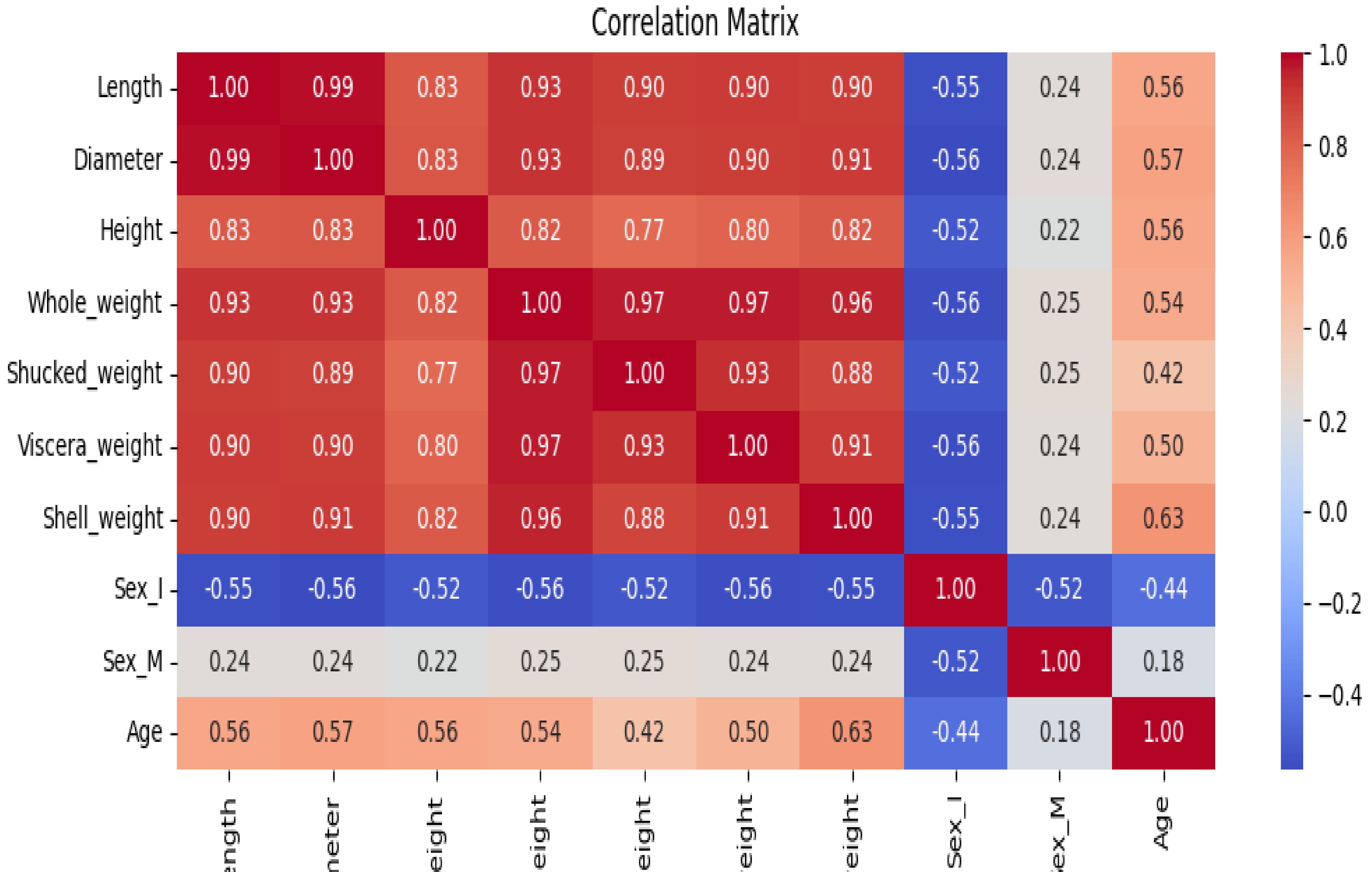

Length vs Age
Age
Length
30
25
20
15
10
5
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8



Whole Weight vs Age
Age
Whole Weight
0.0
0.5
1.0
1.5
2.0
2.5
5
10
15
20
25
30



0.25
0.50
0.75
Age
10
20
30


Correlation Matrix
Length 1.00 0.99 0.83 0.93 0.90 0.90 0.90 -0.55 0.24 0.56
Diameter 0.99 1.00 0.83 0.93 0.89 0.90 0.91 -0.56 0.24 0.57
Height 0.83 0.83 1.00 0.82 0.77 0.80 0.82 -0.52 0.22 0.56
Whole_weight 0.93 0.93 0.82 1.00 0.97 0.97 0.96 -0.56 0.25 0.54
Shucked_weight 0.90 0.89 0.77 0.97 1.00 0.93 0.88 -0.52 0.25 0.42
Viscera_weight 0.90 0.90 0.80 0.97 0.93 1.00 0.91 -0.56 0.24 0.50
Shell_weight 0.90 0.91 0.82 0.96 0.88 0.91 1.00 -0.55 0.24 0.63
Sex_I -0.55 -0.56 -0.52 -0.56 -0.52 -0.56 -0.55 1.00 -0.52 -0.44
Sex_M 0.24 0.24 0.22 0.25 0.25 0.24 0.24 -0.52 1.00 0.18
Age 0.56 0.57 0.56 0.54 0.42 0.50 0.63 -0.44 0.18 1.00
Sex_I
Age
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Feature Scaling and Train-Test Split

- Step 6: Scaling the Features
- Normalization using StandardScaler.
- Step 7: Splitting the Data
- Train-Test split (80-20) using `train_test_split`.

Model Training - Linear Regression

- Model: Linear Regression
- Training: Fit the model on training data.
- Evaluation: MSE and R-squared on test data.
- Results:
- MSE: 4.8912
- -R2: 0.5482

Model Training - Ridge Regression

- Model: Ridge Regression
- Training: Fit the model on training data.
- Evaluation: MSE and R-squared on test data.
- Results:
- MSE: 4.8911
- R2: 0.5482

Model Training - Lasso Regression

- Model: Lasso Regression
- Training: Fit the model on training data.
- Evaluation: MSE and R-squared on test data.
- Results:
- MSE: 7.6826
- R2: 0.2903

Model Comparison

- Linear Regression:
- MSE: 4.8912
- R2: 0.5482
- Lasso Regression:
- MSE: 7.6826
- R2: 0.2903
- Ridge Regression:
- MSE: 4.8911
- R2: 0.5482

Conclusion

- Summary: Linear and Ridge Regression perform similarly and better than Lasso Regression.
- Recommendation: Choose Ridge Regression if multicollinearity is a concern; otherwise, Linear Regression is also a good choice.
- Future Work: Consider other regression techniques or feature engineering for improved performance.

Questions & Answers

Open the floor for any questions from the audience.


Thank You