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

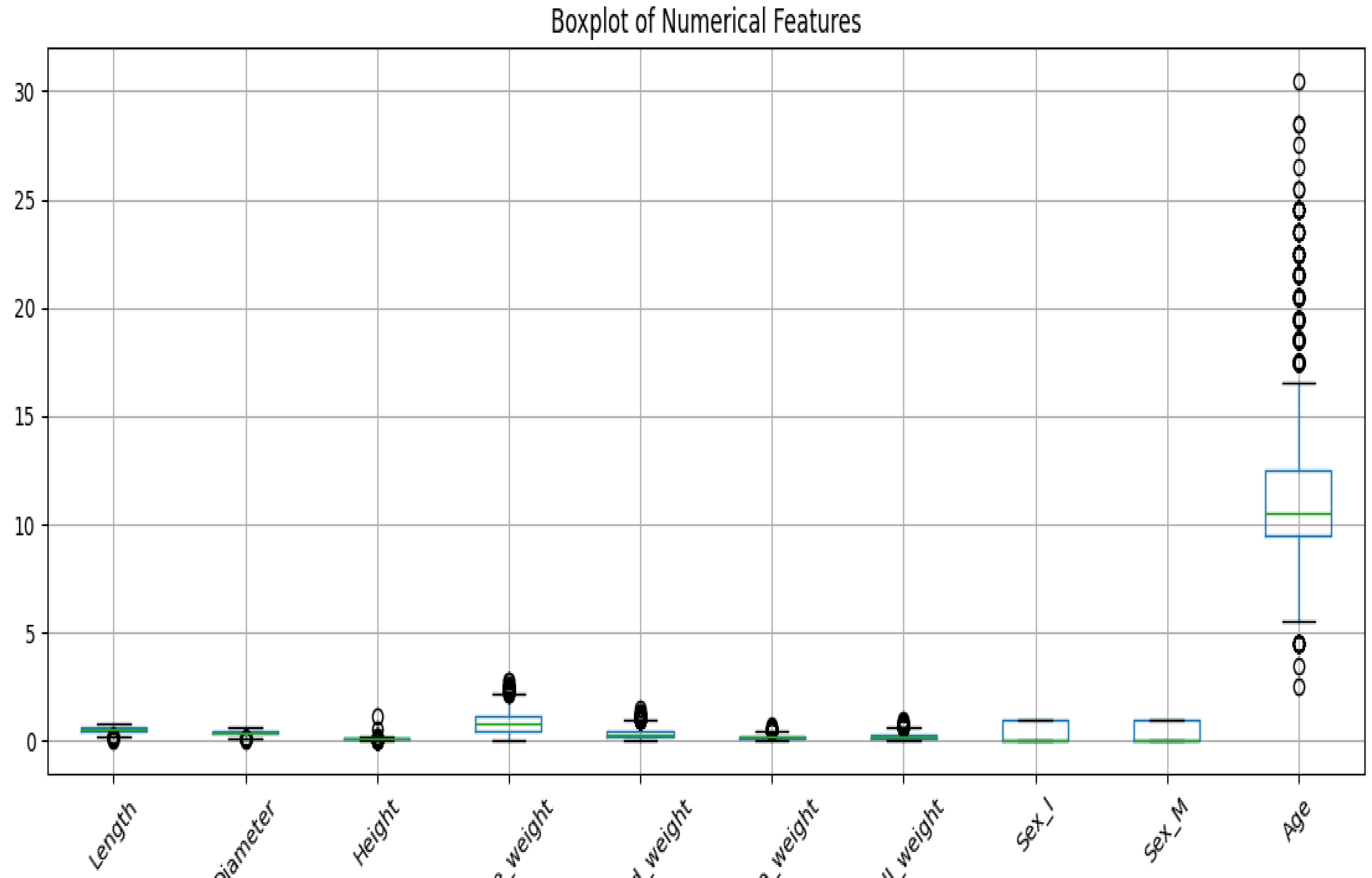
- Step 1: Loading the dataset using Pandas.
- Code Snippet: `pd.read_csv('abalone.csv')`
- Initial DataFrame: Show the first few rows of the dataset.
- Handling Missing Values: Mention that there are no missing values in the dataset.

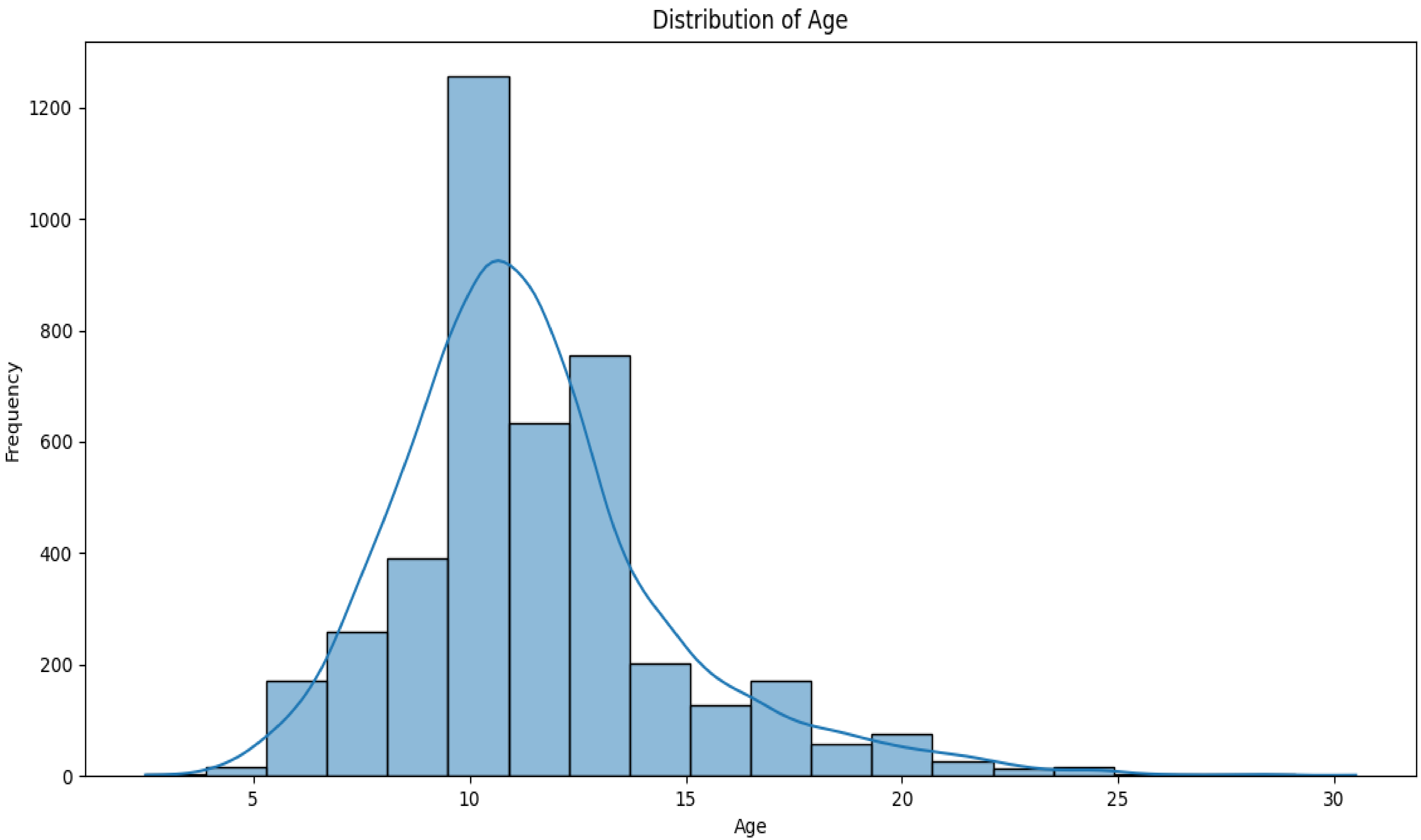
Data Preprocessing

- 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.

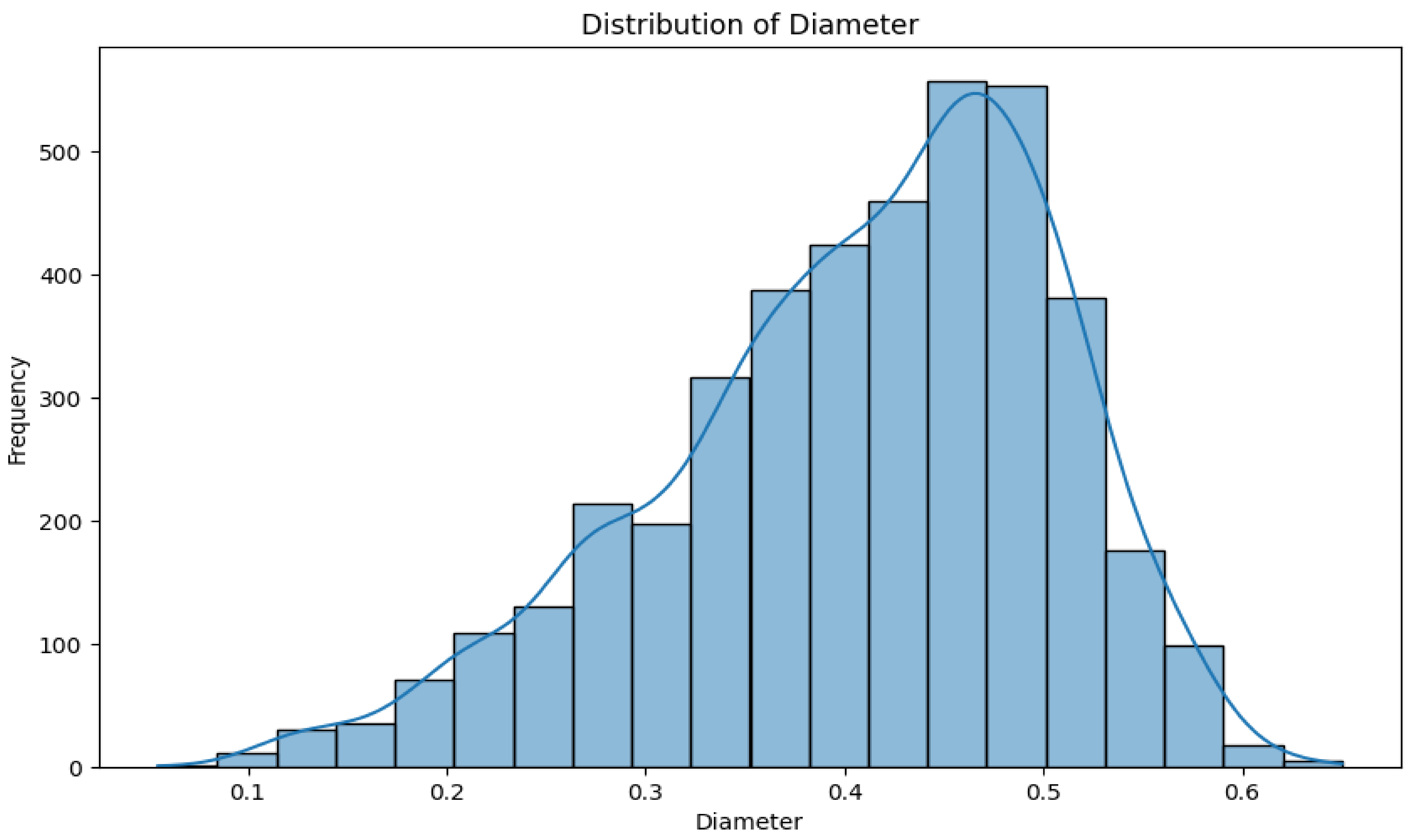
Exploratory Data Analysis (EDA)

- 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.

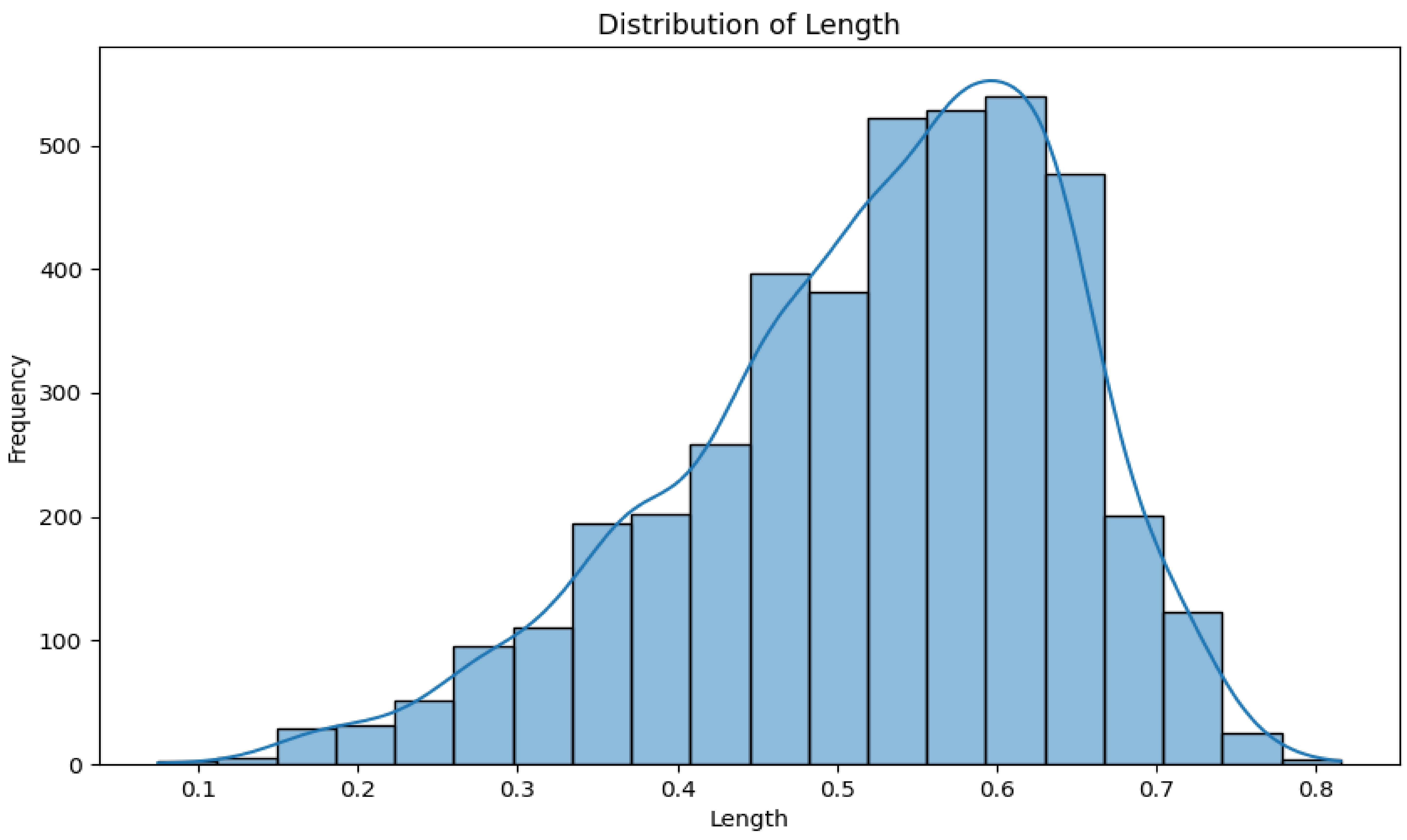

Boxplot of Numerical Features
30
25
20
15
10
5
0
Length
Height
Sex_I
Sex_M
Age

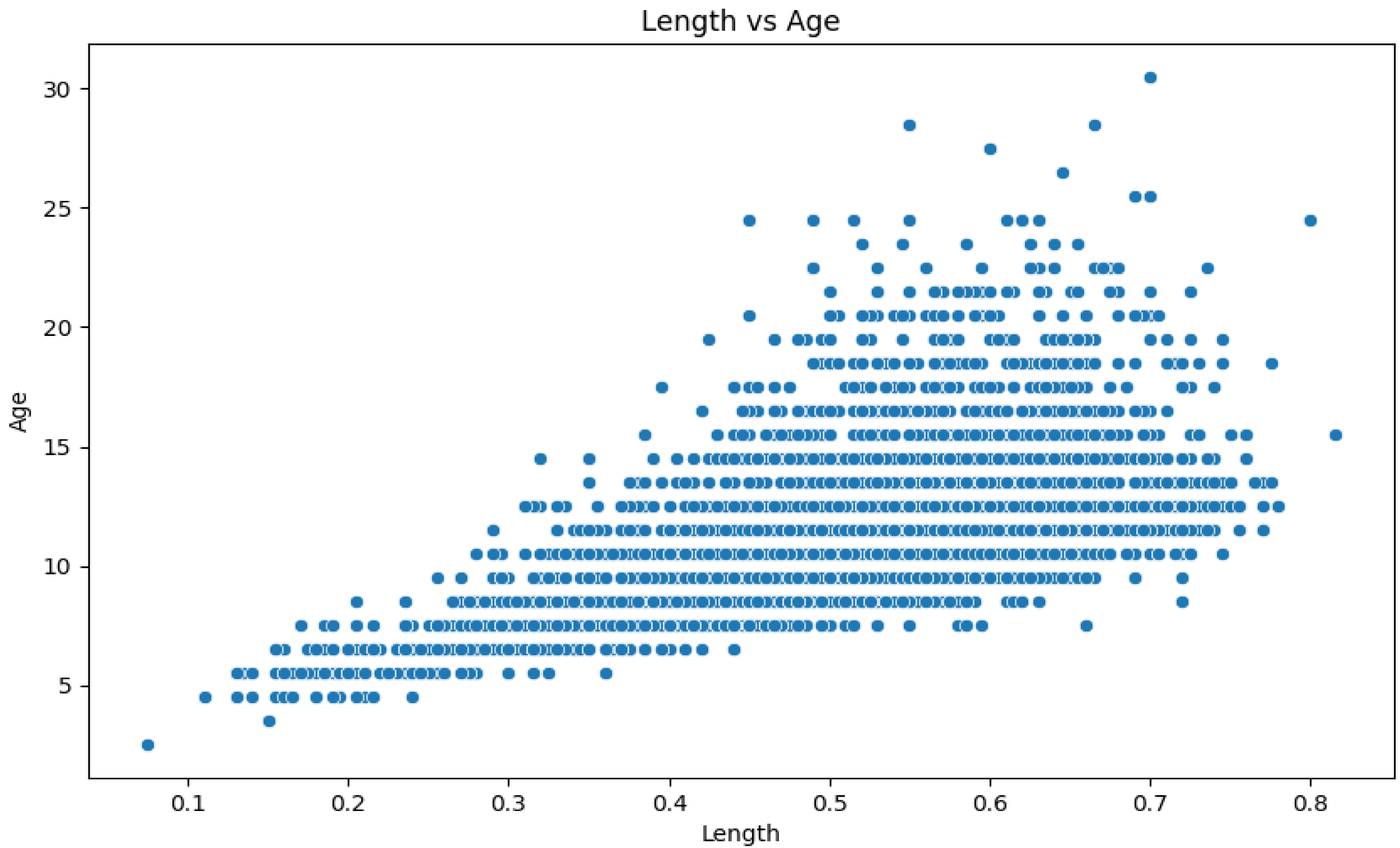


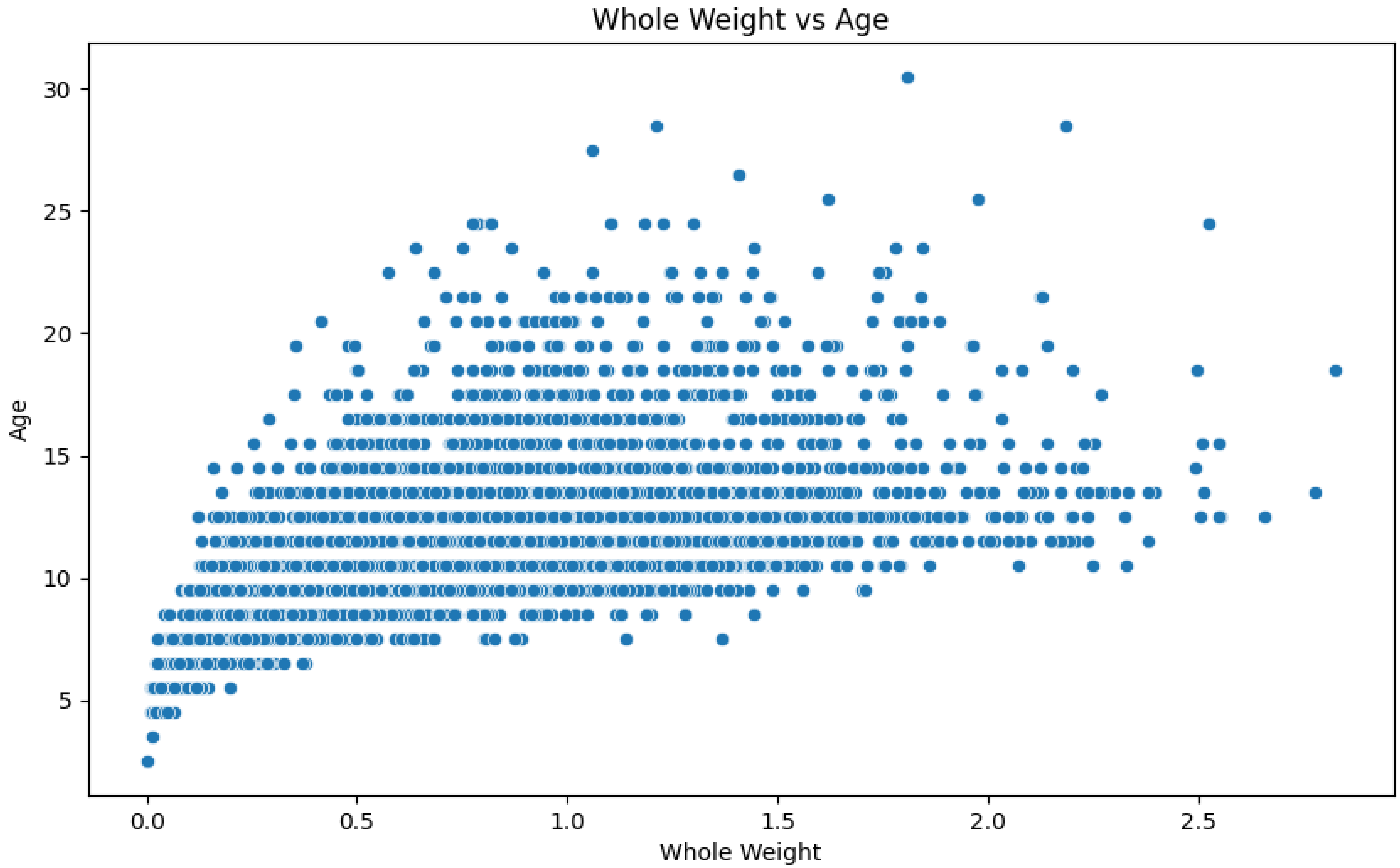
Distribution of Age
Frequency
0
200
400
600
800
1000
1200
5
10
15
20
25
30
Age

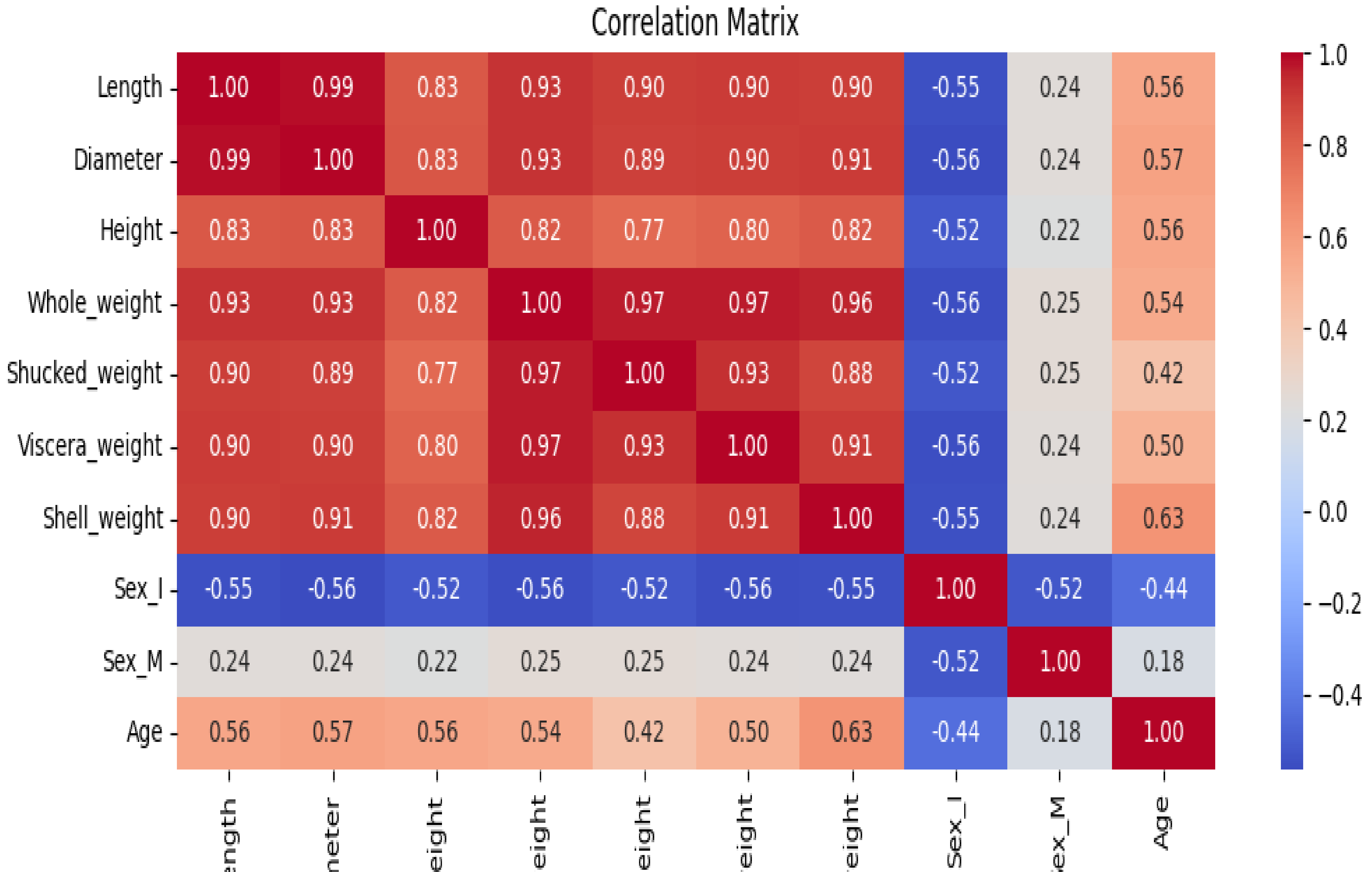


Distribution of Diameter
Frequency
Diameter
0
100
200
300
400
500
0.1
0.2
0.3
0.4
0.5
0.6


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


Correlation Matrix
Length
Diameter
Height
Whole_weight
Shucked_weight
Viscera_weight
Shell_weight
Sex_I
Sex_M
Age
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

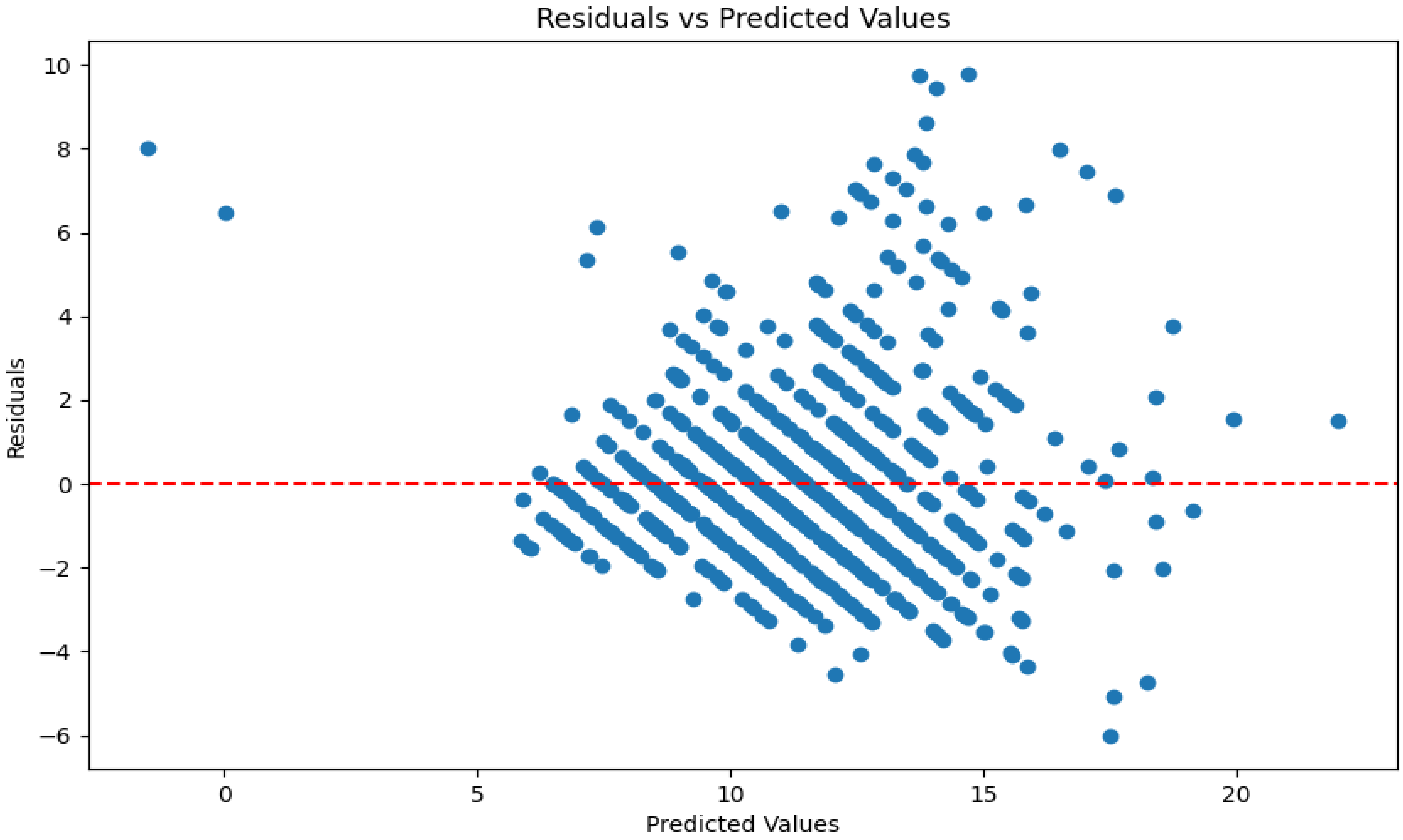
- Linear Regression:
- MSE: 4.8912
- R2: 0.5482
- Lasso Regression:
- MSE: 7.6826
- R2: 0.2903
- Ridge Regression:
- MSE: 4.8911
- R2: 0.5482

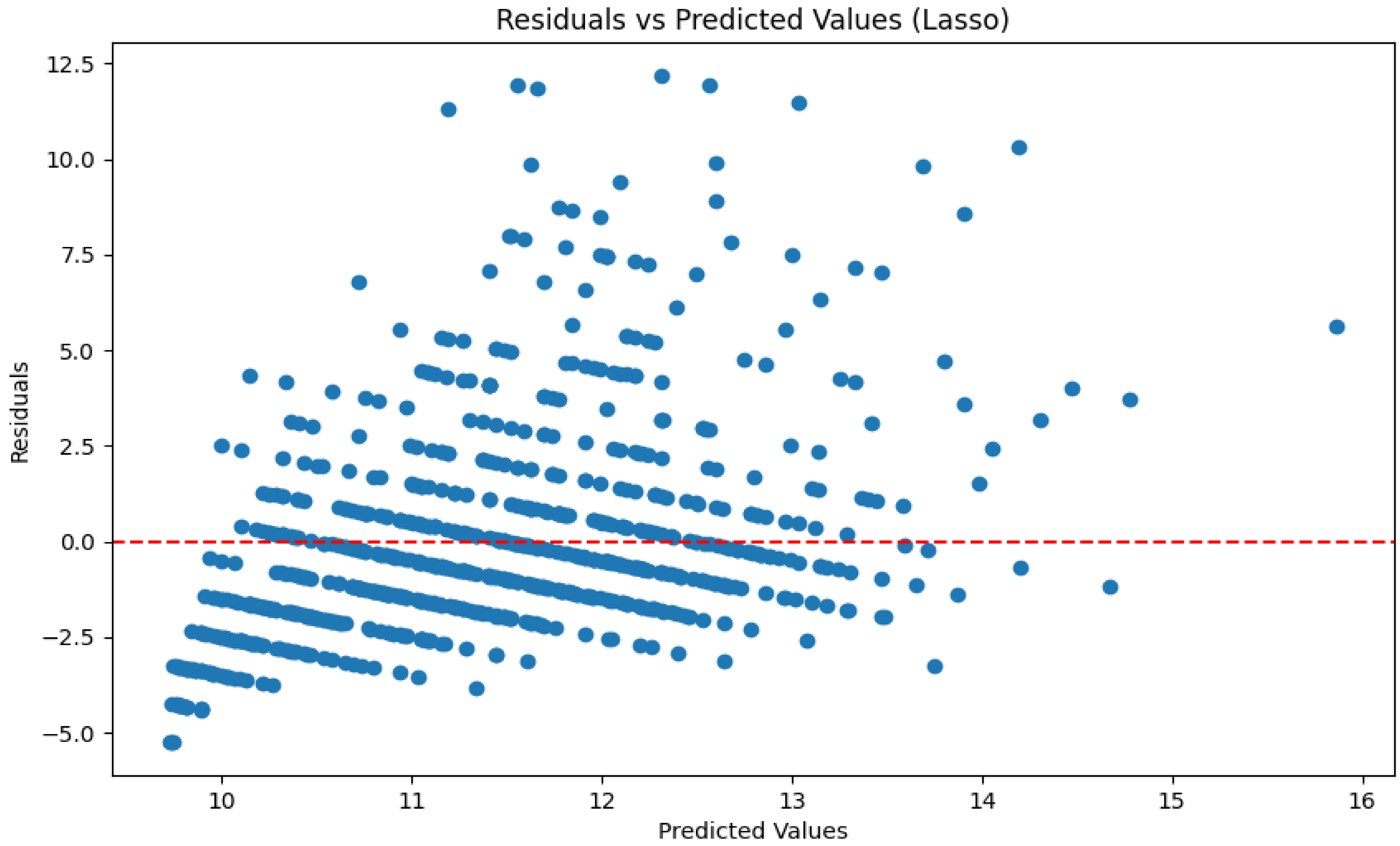
Residual Analysis - Summary

- **Linear Regression**:
  - Mean: -0.009
  - Std: 2.21
  - Min: -6.01
  - Max: 9.78
- **Lasso Regression**:
  - Mean: -0.023
  - Std: 2.77
  - Min: -5.24
  - Max: 12.18
- **Ridge Regression**:
  - Mean: -0.008
  - Std: 2.21
  - Min: -6.02
  - Max: 9.77

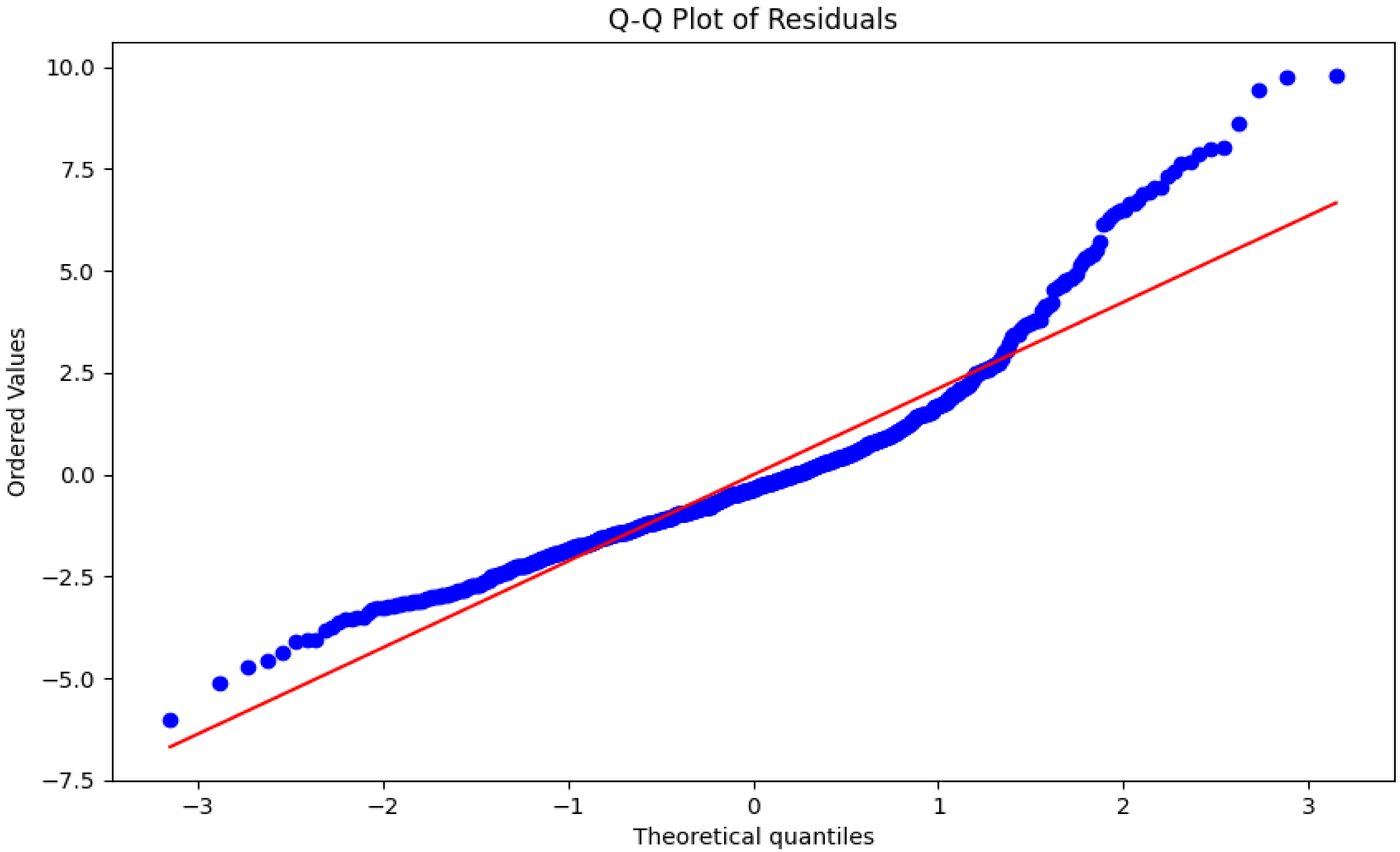
Residual Analysis - Visualizations

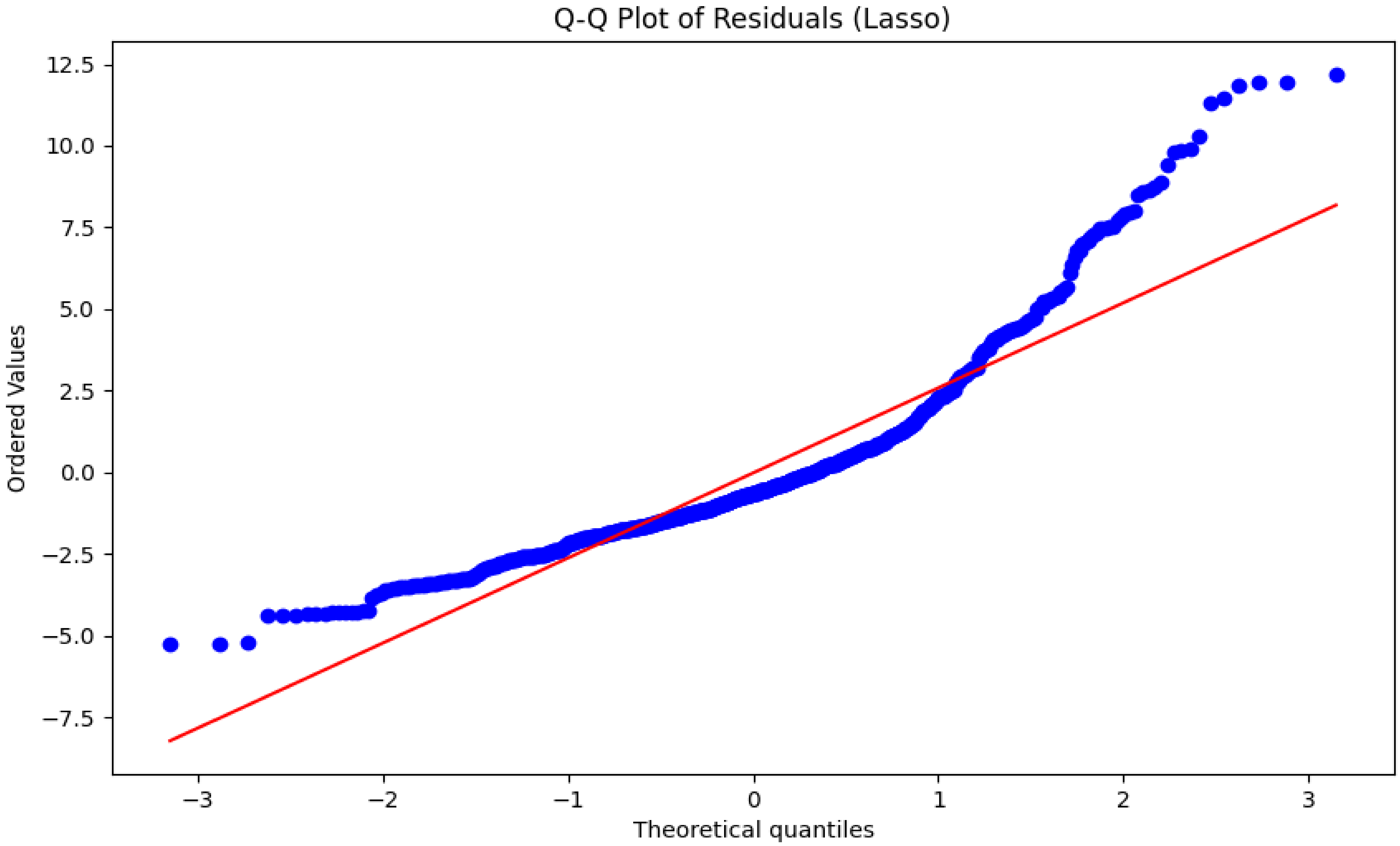
- Residual vs Predicted Values for Ridge Regression
- Residual vs Predicted Values for lasso Regression
- Q-Q Plot of Residuals for Ridge Regression
- Q-Q Plot of Residuals for lasso Regression


Residuals vs Predicted Values
Residuals
Predicted Values
10
8
6
4
2
0
−2
−4
−6
0
5
10
15
20



Residuals vs Predicted Values (Lasso)
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
Residuals
10
11
12
13
14
15
16
Predicted Values


Q-Q Plot of Residuals
Ordered Values
Theoretical quantiles
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−3
−2
−1
0
1
2
3


Q-Q Plot of Residuals (Lasso)
Ordered Values
Theoretical quantiles
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−3
−2
−1
0
1
2
3

Conclusion

- **Model Performance**:
- Linear and Ridge Regression models performed similarly, better than Lasso.
- Ridge Regression is preferred due to handling multicollinearity effectively.
- Ridge Regression: MSE (4.8911), R^2 (0.5482).
- **Residuals Analysis**:
- Similar residual patterns for Linear and Ridge, capturing data patterns well.
- Some outliers indicate underestimation or overestimation.
- Lasso showed higher residual variability, indicating less generalization.

Future Work and Acknowledgements

- **Future Work**:
- Explore advanced regression techniques and feature engineering.
- Consider a Generalized Linear Model (GLM) for better handling different distributions of the target variable.
- **Acknowledgements**:
- Thanks to the UCI Machine Learning Repository for the Abalone dataset.

Questions & Answers

Open the floor for any questions from the audience.


Thank You