Analysis of Factors Influencing Shoe Size

Objective: The goal of this analysis is to explore the relationship between a person's age, the price of their shoes, their sex, and their shoe size. Various regression models will be built and evaluated to determine the best predictor for shoe size based on the available features.

---

1. Data Loading and Exploration

The project begins by loading the `exploring_dataset.csv` file into a pandas DataFrame and performing an initial exploratory analysis.

1.1. Dataset Overview

The dataset contains 1000 entries and four columns: `age`, `shoe_size`, `price(£)`, and `sex`.

A preview of the first five rows is shown below:

```
   age  shoe_size  price(£) sex

0    3        27         4   m

1    4        27         4   m

2    5        28         5   m

3    6        29         5   f

4    7        29         6   f
```

1.2. Descriptive Statistics

A statistical summary provides insights into the distribution of the numerical data:

- **Age:** Ranges from 3 to 16 years, with a mean of approximately 9.5 years.
- **Shoe Size:** Ranges from 27 to 36, with a mean of about 30.7.
- **Price (£):** Ranges from £3 to £15, with a mean of roughly £7.4.

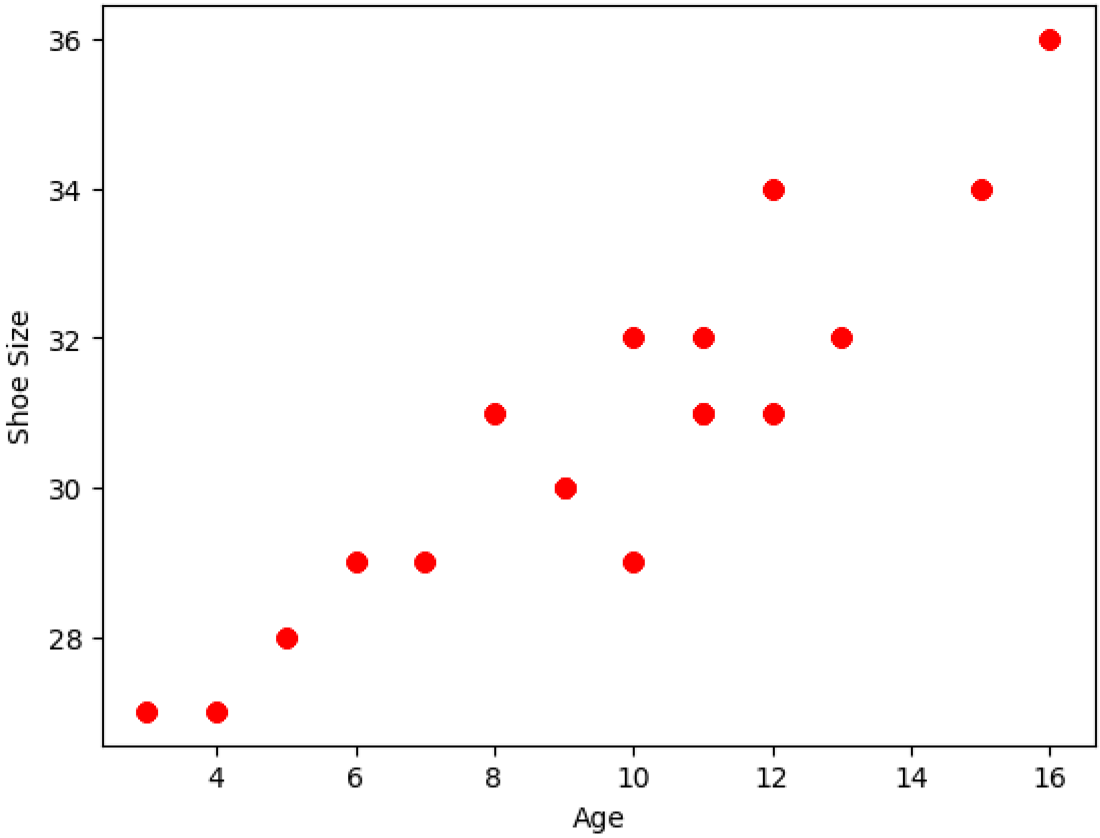
```
             age   shoe_size     price(£)

count  1000.000000  1000.00000  1000.000000

mean      9.532000    30.70700     7.427000

std       3.484985     2.35174     3.361841

min       3.000000    27.00000     3.000000
```

```
25%      7.000000    29.00000     5.000000
50%     10.000000    31.00000     6.000000
75%     12.000000    32.00000     9.000000
max     16.000000    36.00000    15.000000
```

The dataset is complete, with no missing values, and the data types are appropriate for analysis.

1.3. Exploratory Data Analysis (EDA)

To visualize the relationship between age and shoe size, a scatter plot was generated. The plot shows a clear **positive linear relationship**, suggesting that as age increases, shoe size also tends to increase.



2. Model 1: Simple Linear Regression

A simple linear regression model was built to predict `shoe_size` using only `age` as the predictor variable.

2.1. Model Training

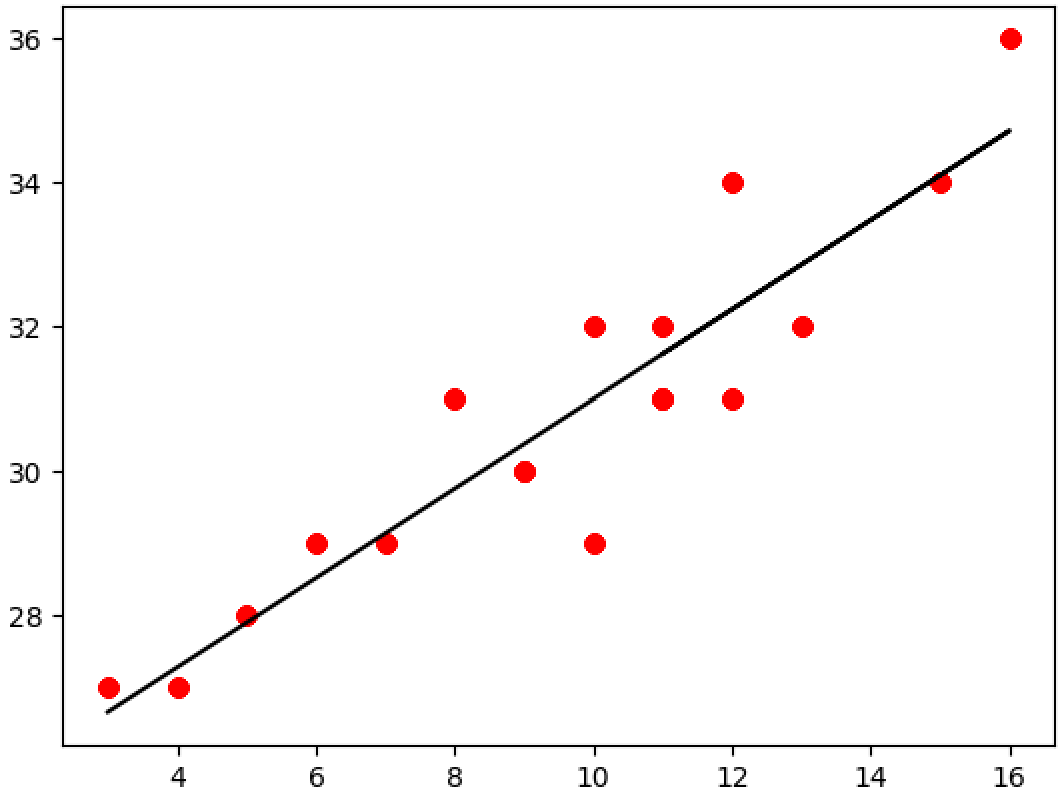
The model was trained using scikit-learn's `LinearRegression`.

- **Intercept (b_0):** 24.786
- **Coefficient (b_1):** 0.621

This means the model's equation is: **Shoe Size = 24.786 + 0.621 * Age**.

2.2. Visualization and Evaluation

The regression line was plotted over the data, visually confirming a good fit.

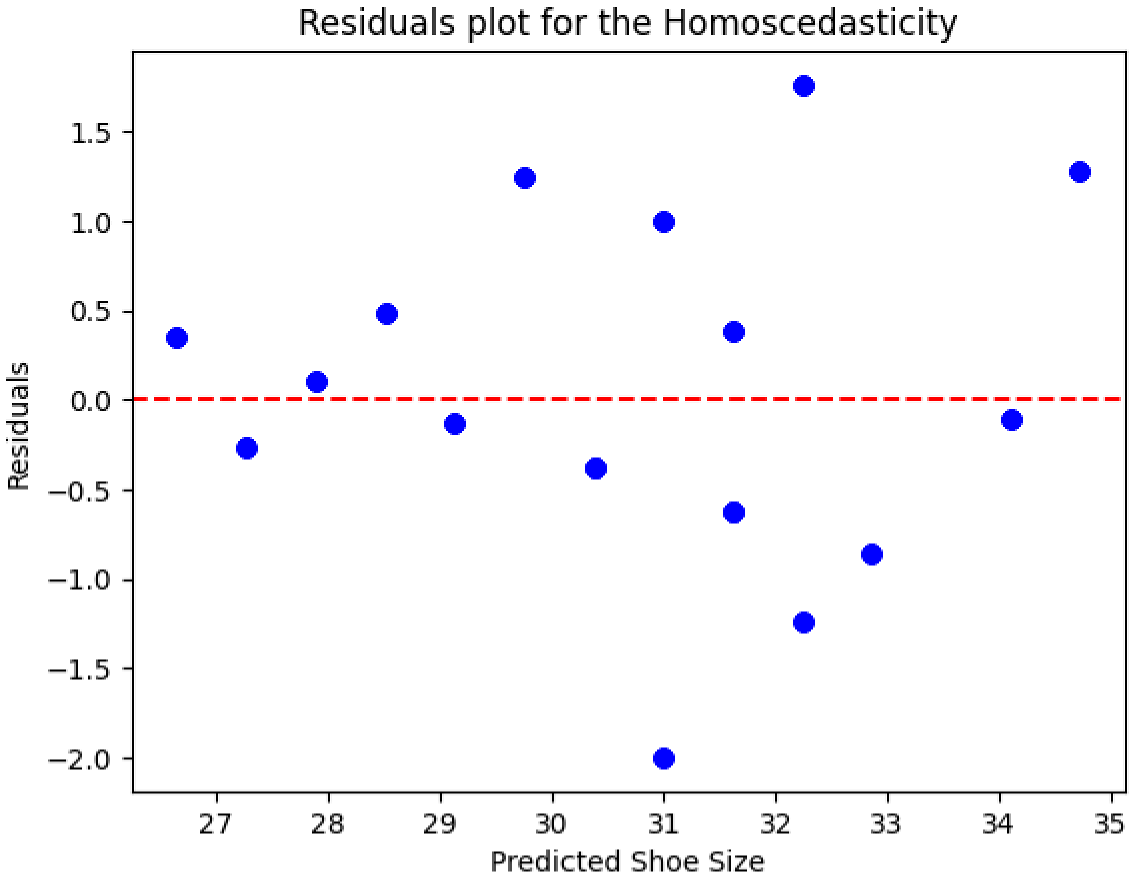


To evaluate the model's performance robustly, 5-fold cross-validation was used.

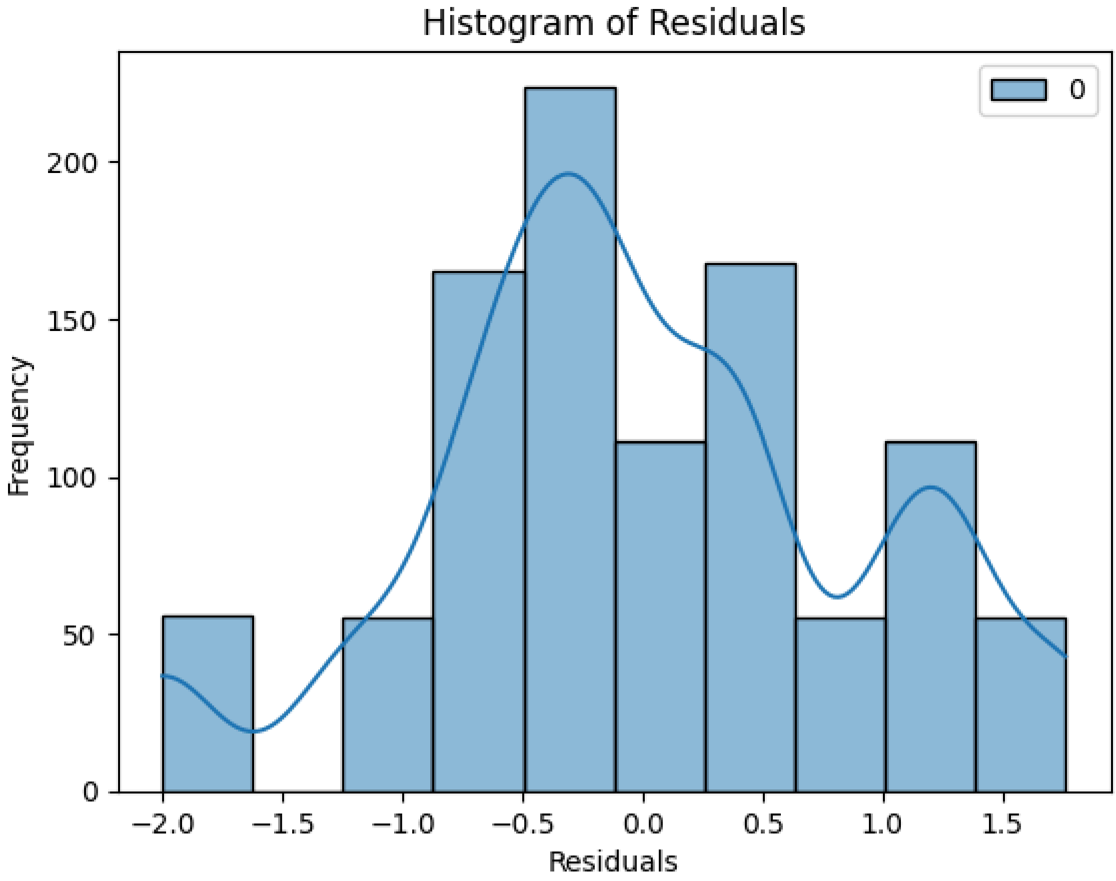
- **Average Mean Squared Error (MSE): 0.843**

2.3. Assumption Checking

- **Homoscedasticity:** A residuals plot shows that the errors are randomly scattered around the horizontal line at 0, confirming that the variance of the errors is constant. This assumption holds.



- **Normality of Residuals:** A histogram of the residuals shows a roughly bell-shaped curve, suggesting that the errors are normally distributed. This assumption also holds.



3. Model 2: Linear Regression with Gradient Descent

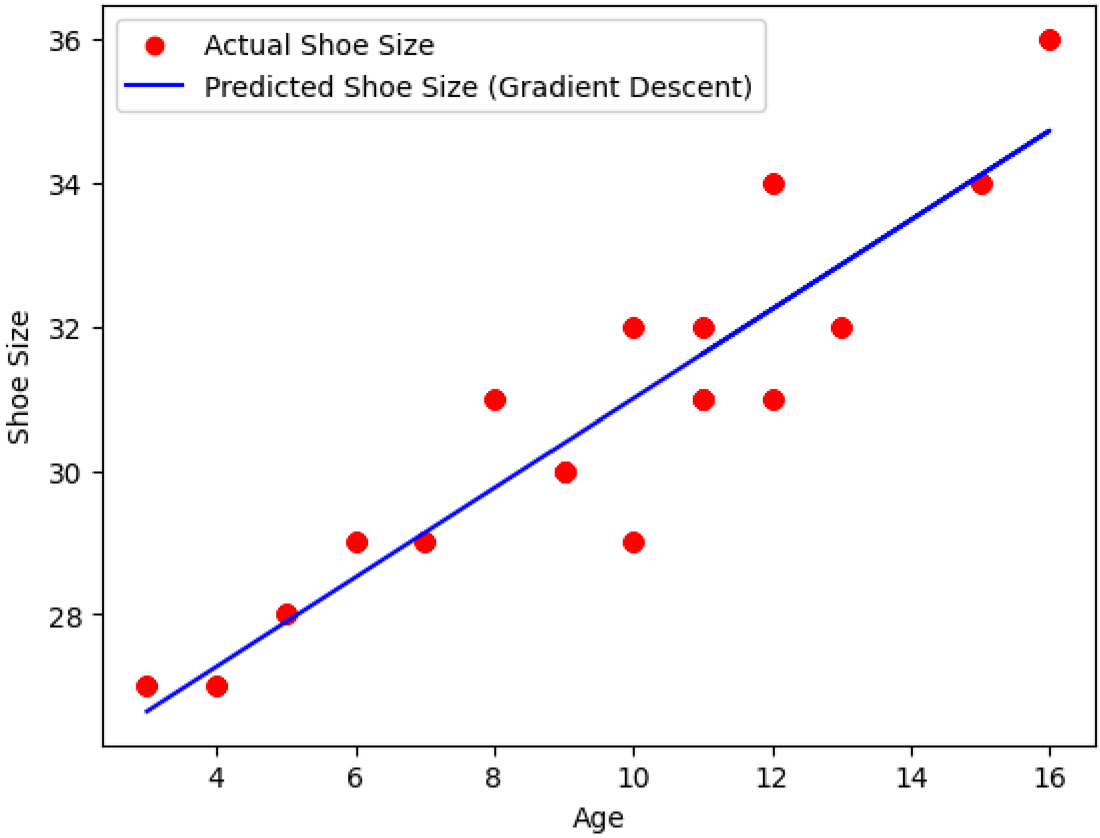
As an alternative to the standard linear regression solver, a model was trained using the gradient descent optimization algorithm to find the optimal bias (intercept) and weight (coefficient).

3.1. Model Training and Results

After normalizing the data and running the algorithm for 1000 iterations, the model converged with the following parameters (on the normalized scale):

- **Bias (Intercept):** 9.02e-10 (effectively 0)
- **Weight (Coefficient):** 0.921

The resulting regression line is visually very similar to the one from the standard linear regression model, demonstrating the effectiveness of gradient descent.



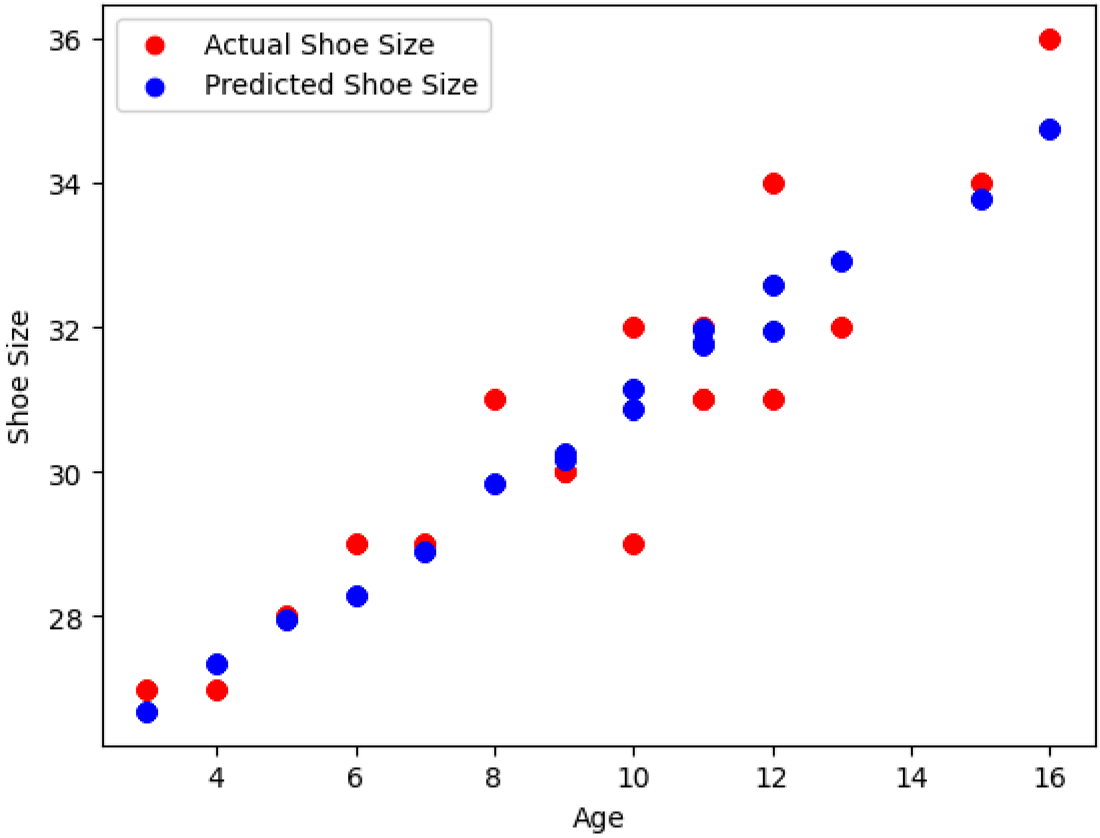
4. Model 3: Multiple Linear Regression

To improve predictive accuracy, a multiple linear regression model was created using `age`, `price(£)`, and `sex` as predictors. The categorical `sex` variable was converted into a numerical format using one-hot encoding.

4.1. Model Training and Evaluation

- **Coefficients:**
  - Age: 0.671
  - Price(£): -0.061
  - Sex (Male): 0.341
- **Intercept:** 24.574

The model's performance was evaluated using 5-fold cross-validation.



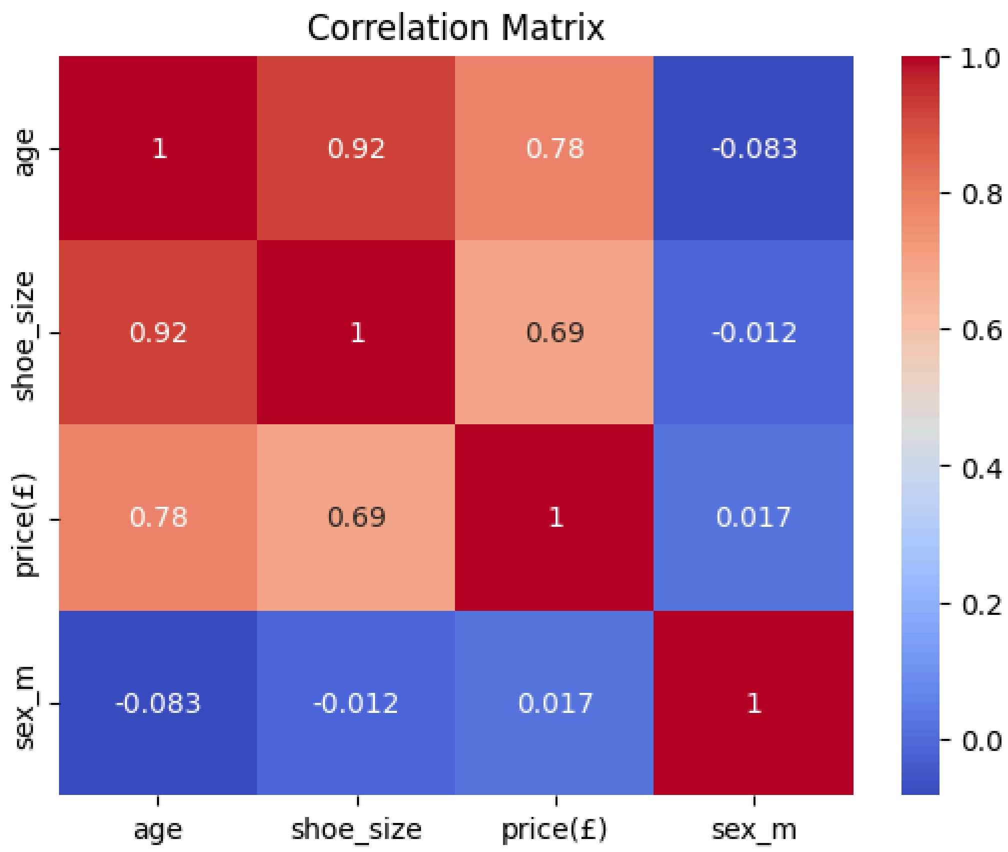
- **Average Mean Squared Error (MSE): 0.804**

This is a notable **improvement** over the simple linear regression model's MSE of 0.843, indicating that including price and sex adds predictive power.

4.2. Multicollinearity Check

A correlation matrix and Variance Inflation Factor (VIF) scores were used to check for multicollinearity.

- **Correlation Matrix:** The heatmap shows a very high correlation (0.98) between `age` and `price(£)`.



- **VIF Scores:**
  - Age: 15.10
  - Price(£): 15.21
  - Sex (Male): 1.90

The VIF scores for **age** and **price** are well above the common threshold of 10, confirming a high degree of multicollinearity. This can make the model's coefficients unstable and difficult to interpret.

---

5. Model 4: Regularized Regression (Ridge & Lasso)

To address the multicollinearity identified in the multiple regression model, Ridge (L2) and Lasso (L1) regularization techniques were applied.

5.1. Ridge Regression

A Ridge model with `alpha=0.1` was trained.

- **Average MSE: 0.804**
  - This result is almost identical to the standard multiple regression model, showing that Ridge effectively manages multicollinearity without sacrificing predictive accuracy.

5.2. Lasso Regression

A Lasso model with alpha=0.1 was trained.

- **Average MSE: 0.840**
    - Lasso regression increased the MSE slightly but performed feature selection by driving the coefficient for sex_m to zero. This suggests that sex might be the least important feature among the three.

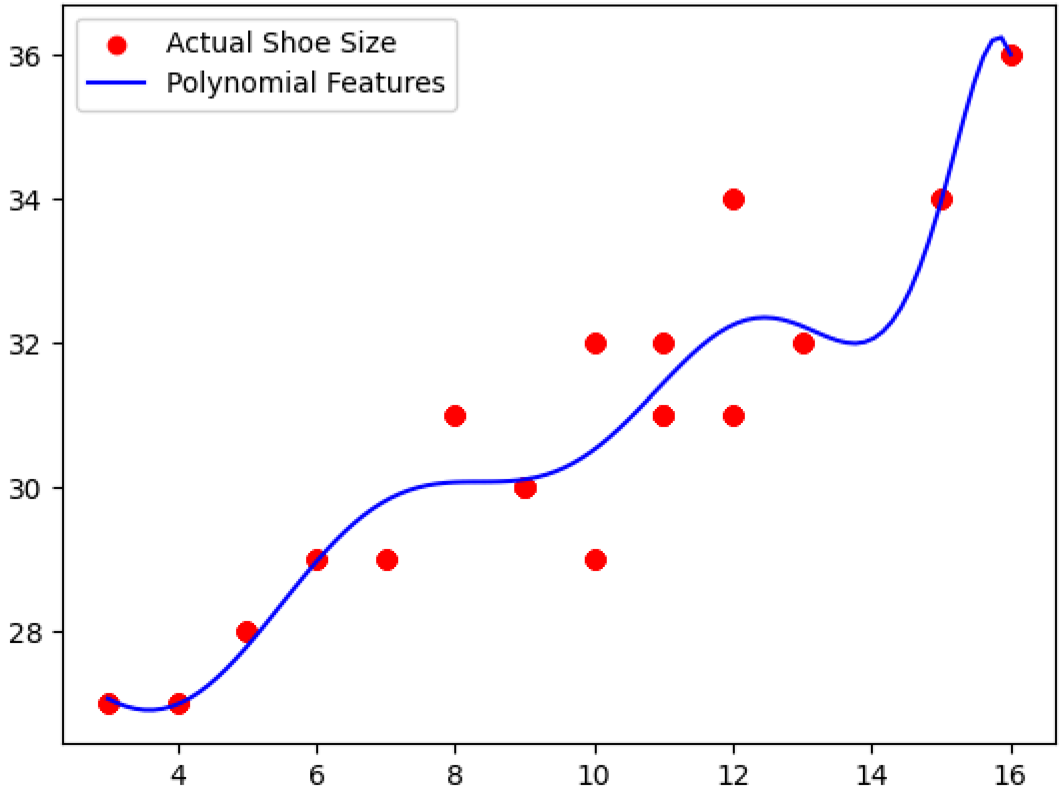
---

6. Model 5: Polynomial Regression

Finally, to check for any non-linear relationships, a polynomial regression model was fitted using age as the predictor.

6.1. Model Training and Evaluation

A model with polynomial features up to degree 10 was created.



- **Average MSE: 0.804**
    - The MSE is identical to the best-performing multiple regression model. While it fits the data well, the resulting curve is complex, and the multiple linear regression model offers a simpler, more interpretable solution with the same accuracy.

---

7. Conclusion

This analysis explored several regression models to predict shoe size. The key findings are:

- The **Multiple Linear Regression model** provided the best performance, achieving the lowest Average Mean Squared Error of **0.804**.
- This model, however, suffered from **multicollinearity** between the `age` and `price` features.
- **Ridge Regression** effectively managed this multicollinearity and achieved the same low MSE, making it an excellent and more robust alternative.
- Simple linear regression using only `age` was a strong baseline (MSE=0.843), confirming that age is the most significant predictor of shoe size in this dataset. The polynomial model did not offer any performance improvement over the multiple linear model.