Enhanced Linear Regression Report: Weight → MPG

Dataset: auto-mpg (rows after cleaning: 392, columns: 9)

Selected numeric columns: mpg, cylinders, displacement, horsepower, weight, acceleration, model year, origin, car name

Dataset Summary (first rows and descriptive statistics shown separately in code)

Models and Performance (test set):

1) Simple Linear Regression (weight → mpg)
Equation (approx): mpg = 47.201 + (-0.007904) *
weight Test R^2: 0.6533 | MAE: 3.4641 | RMSE: 4.2064

2) Multiple Linear Regression (weight, horsepower, displacement, acceleration)
Test R^2: 0.6460 | MAE: 3.5096 | RMSE: 4.2504

3) Polynomial Regression (degree 2, weight)
Test R^2: 0.6730 | MAE: 3.1383 | RMSE: 4.0852

4) Ridge (regularized) Regression (multi-features)
Test R^2: 0.6463 | MAE: 3.4792 | RMSE: 4.2486

5) Lasso (regularized) Regression (multi-features)
Test R^2: 0.6539 | MAE: 3.4597 | RMSE: 4.2029

Cross-validation (5-fold) R^2 averages:

Simple LR mean R^2: 0.6853 (std 0.0316)
Multiple LR mean R^2: 0.6905 (std 0.0453)
Polynomial (deg2 on weight) mean R^2: 0.7074 (std 0.0306)

Figure 1

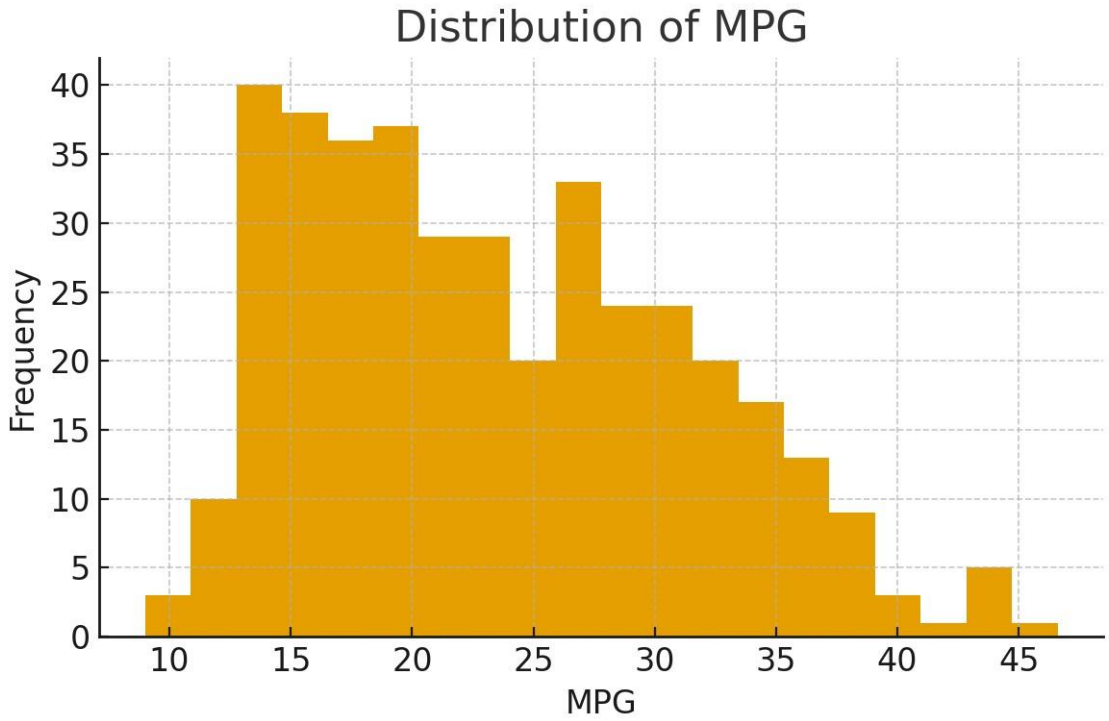

Distribution of MPG
Frequency
MPG
0
5
10
15
20
25
30
35
40
10
15
20
25
30
35
40
45

Figure 2

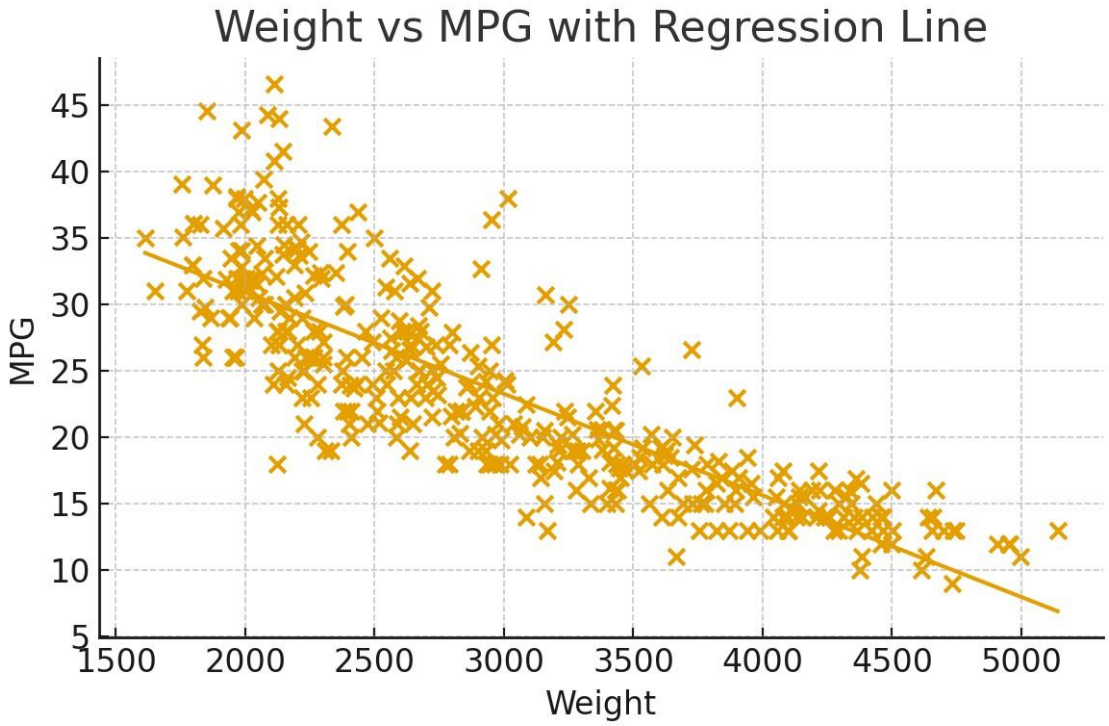

Weight vs MPG with Regression Line
MPG
Weight
45
40
35
30
25
20
15
10
5
1500
2000
2500
3000
3500
4000
4500
5000

Figure 3

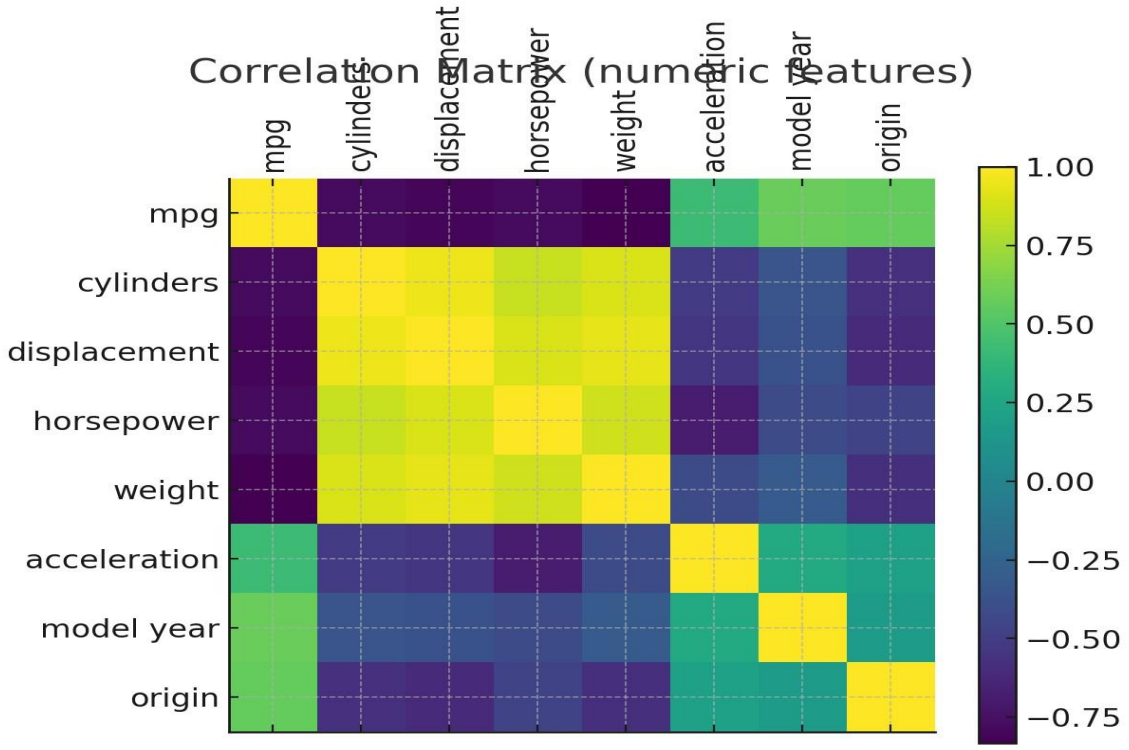

Correlation Matrix (numeric features)
mpg
cylinders
displacement
horsepower
weight
acceleration
model year
origin
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Figure 4

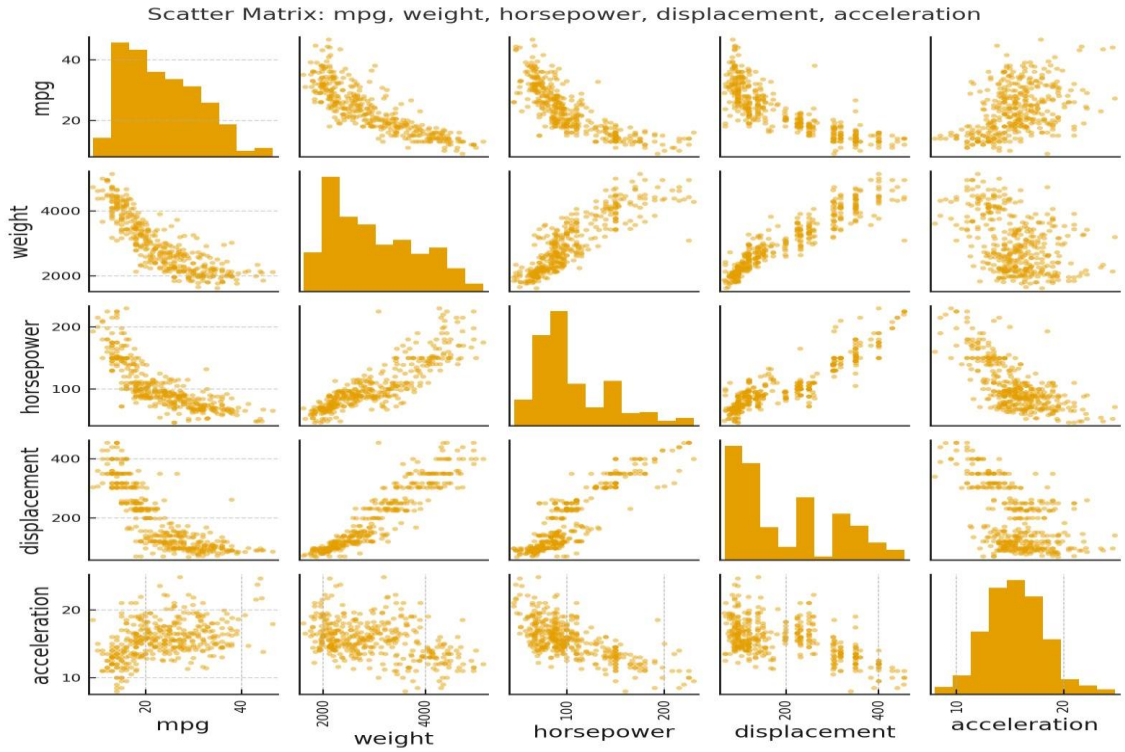

Scatter Matrix: mpg, weight, horsepower, displacement, acceleration
mpg
weight
horsepower
displacement
acceleration

Figure 5

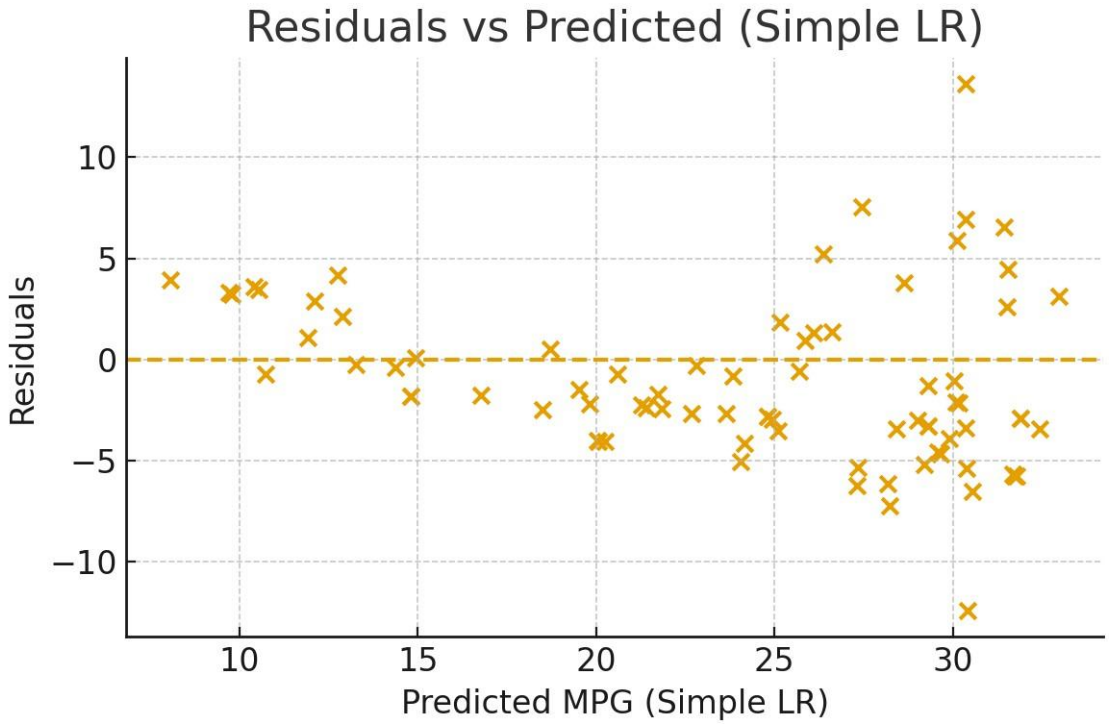

Residuals vs Predicted (Simple LR)
Residuals
10
5
0
−5
−10
10
15
20
25
30
Predicted MPG (Simple LR)

Figure 6

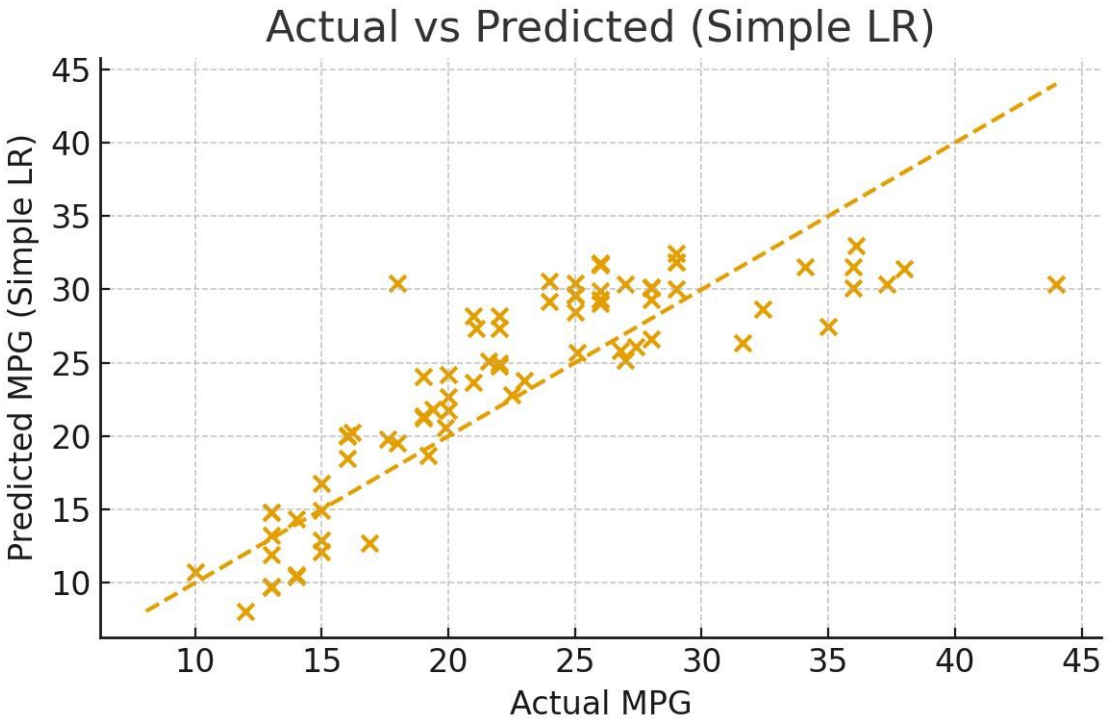
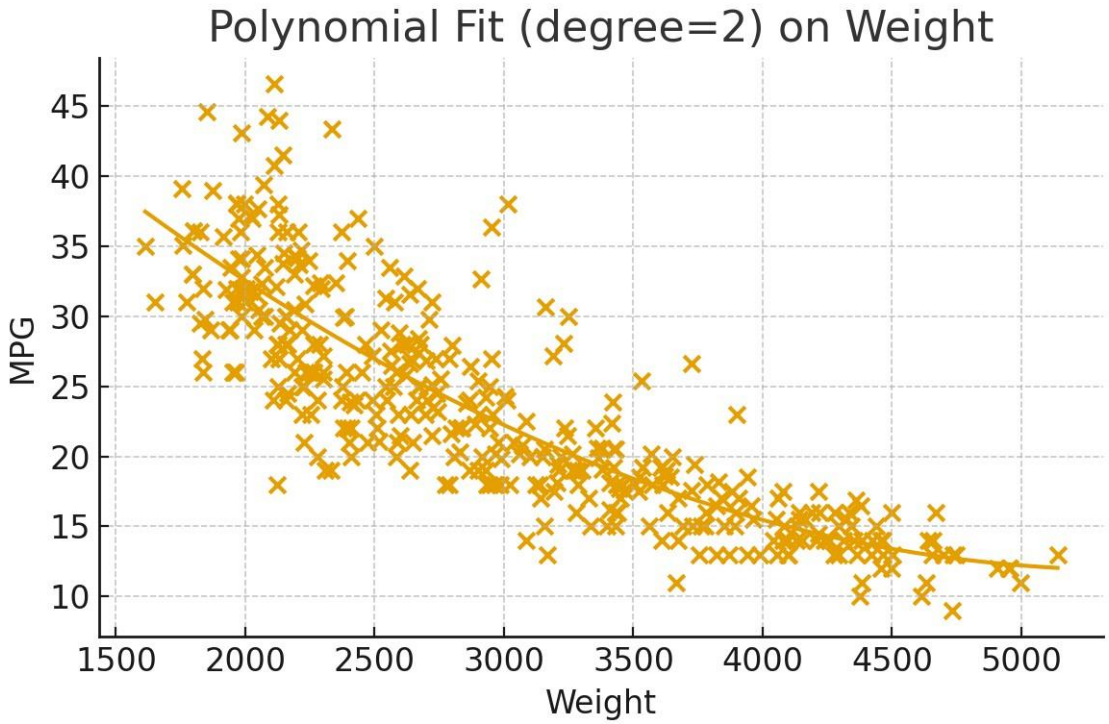

Actual vs Predicted (Simple LR)
Predicted MPG (Simple LR)
Actual MPG
10
15
20
25
30
35
40
45

Figure 7



Conclusions

Weight is strongly negatively correlated with MPG. Simple linear regression explains a large portion of variance,but using multiple features and polynomial terms improves performance. Regularized models (Ridge/Lasso) help when using multiple correlated predictors. For a top-grade report: include EDA, diagnostics (residuals, assumption checks), model comparisons, and clear interpretation of coefficients.