Assignment 4 Solution 1 Output

"""

a) Which variable would you add next? Why?
After BMI and S5, we added all other attributes of the dataset namely Bp, S1, S2, S3, S4, and S6. We focused the analysis on Root Mean Squarred Error (RMSE),
and R^2 Score. We calculated the values and found that addition of BP resulted in lowering of RMSE. Thus, we decided to select BP as the next attribute to be added.

Feature Added	RMSE	R^2 Score
bp	53.768366	0.454331
s3	53.705538	0.455605
s6	53.810247	0.453481
s1	54.221504	0.445095
s2	54.090240	0.447778
s4	53.801874	0.453651

b) How does adding it affect the model's performance? Compute metrics and compare to having just bmi and s5.
We observed that addition of BP reduced the Root Mean Squarred Error (RMSE). With less error, the prediction power of the model will improve.

Model	RMSE	R^2 Score
Base (bmi, s5)	53.868701	0.452293
Base + bp	53.768366	0.454331

c) Does it help if you add even more variables?
To test this, we considered bmi + s5 + bp as the new base and then further added other features one by one.

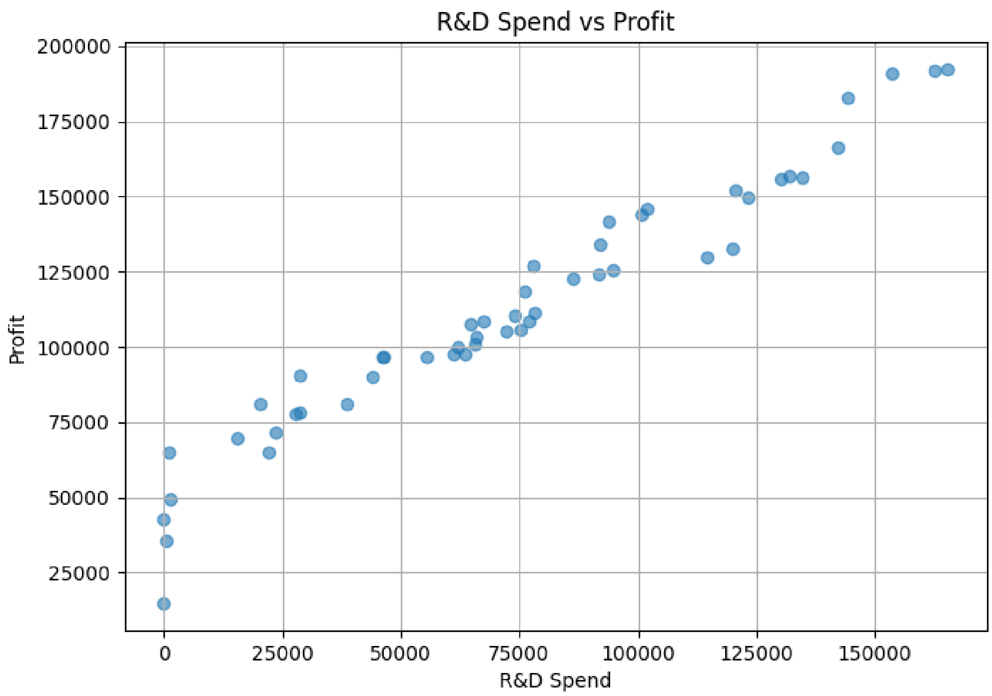
Model	RMSE	R^2 Score
s3	54.127467	0.447018
s6	53.779654	0.454102
s1	54.180881	0.445926
s2	53.886656	0.451927
s4	53.854365	0.452584

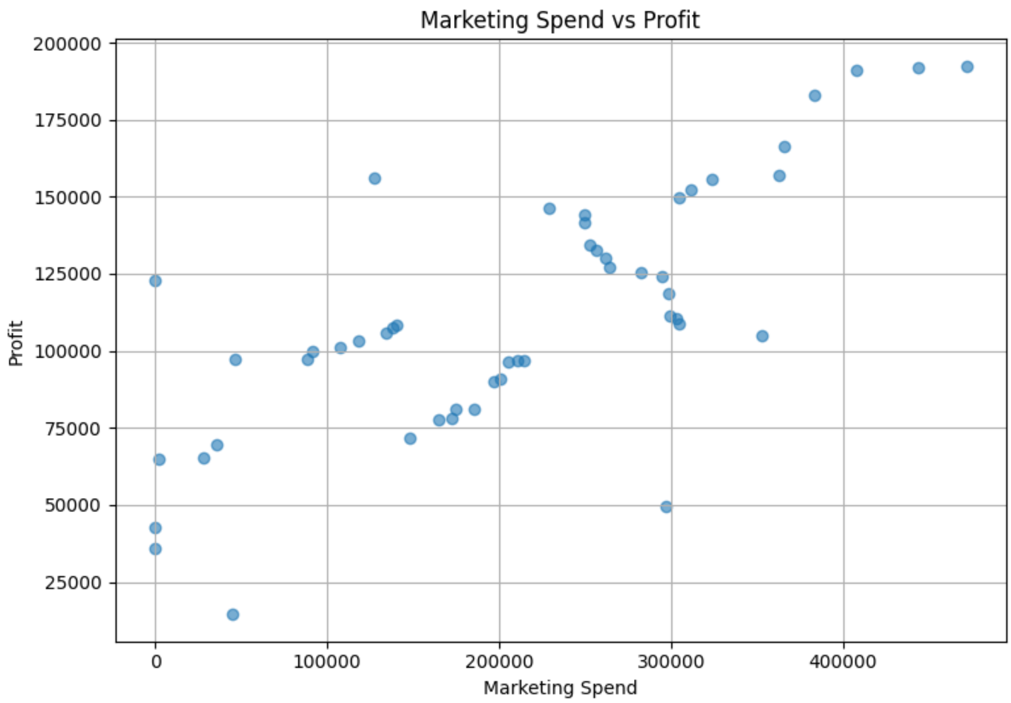
What we found that the next attribute that can be added is S6. However, S6 causes the RMSE to increase. Thus, we will
not add any more attributes after bp.

Model	RMSE	R^2 Score
bmi, s5, bp	53.768366	0.454331

bmi, s5, bp, s6   53.779654   0.454102
"""

Assignment 4 Solution 2 Output





<class 'pandas.core.frame.DataFrame'>

RangeIndex: 50 entries, 0 to 49

Data columns (total 6 columns):

#	Column	Non-Null Count	Dtype
0	R&D Spend	50 non-null	float64
1	Administration	50 non-null	float64
2	Marketing Spend	50 non-null	float64
3	Profit	50 non-null	float64
4	State_Florida	50 non-null	bool
5	State_New York	50 non-null	bool

dtypes: bool(2), float64(4)

memory usage: 1.8 KB

None

Correlation Matrix:

	R&D Spend	Administration	...	State_Florida	State_New York
R&D Spend	1.000000	0.241955	...	0.105711	0.039068
Administration	0.241955	1.000000	...	0.010493	0.005145
Marketing Spend	0.724248	-0.032154	...	0.205685	-0.033670
Profit	0.972900	0.200717	...	0.116244	0.031368
State_Florida	0.105711	0.010493	...	1.000000	-0.492366
State_New York	0.039068	0.005145	...	-0.492366	1.000000

[6 rows x 6 columns]

Selected Predictors: ['R&D Spend', 'Marketing Spend']

Training Data Metrics:

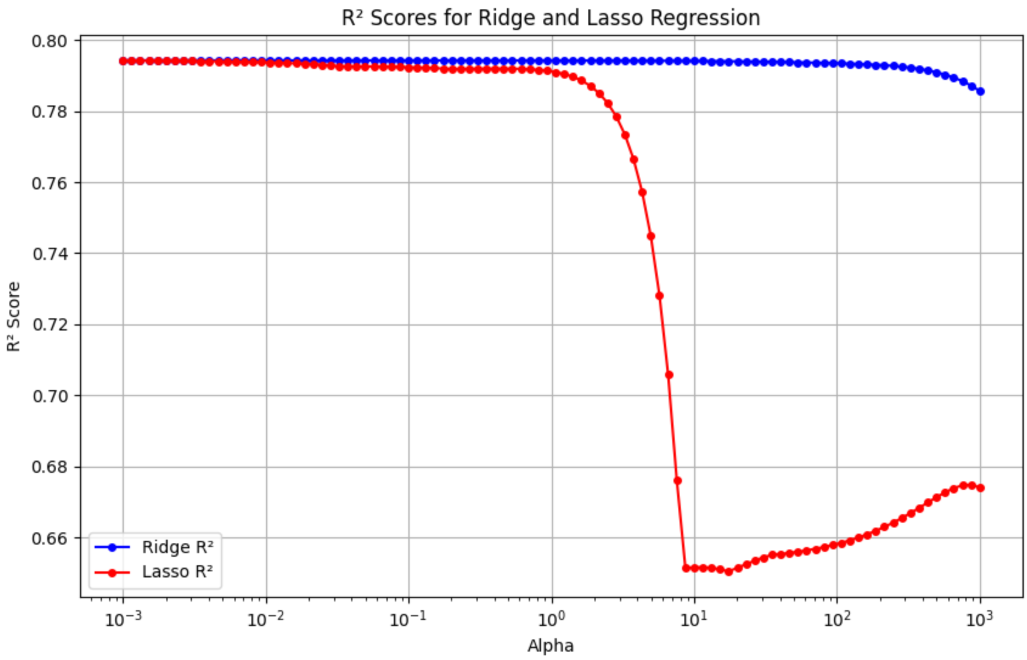
RMSE: 9101.191468669913, R^2: 0.9518828286863577

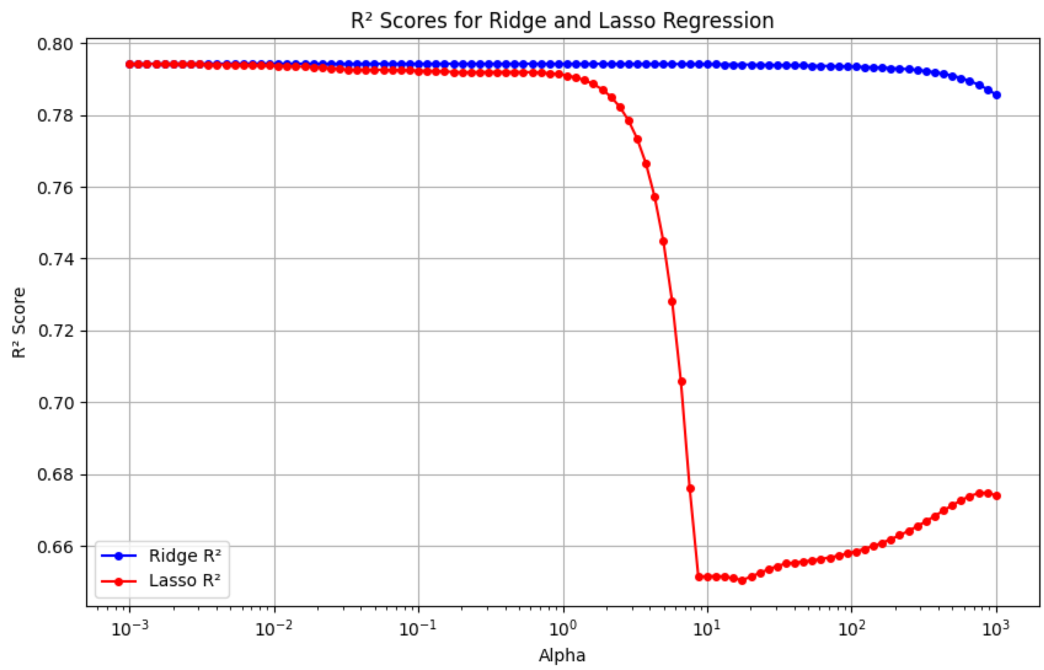
Testing Data Metrics:

RMSE: 8206.32881316585, R^2: 0.9168381183550247

Process finished with exit code 0

Assignment 4 Solution 3 Output






```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 392 entries, 0 to 391
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   mpg           392 non-null    float64
 1   cylinders     392 non-null    int64
 2   displacement  392 non-null    float64
 3   horsepower    392 non-null    int64
 4   weight        392 non-null    int64
 5   acceleration  392 non-null    float64
 6   year          392 non-null    int64
 7   origin        392 non-null    int64
 8   name          392 non-null    object
dtypes: float64(3), int64(5), object(1)
memory usage: 27.7+ KB
```

Optimal Ridge Regression Alpha and R^2:

Alpha: 0.001, Best R^2 Score: 0.7942348920666245

Optimal Lasso Regression Alpha and R^2:

Alpha: 0.001, Best R^2 Score: 0.7941834683177982

Comparison Table:

	Regressor	Optimal Alpha	Best R^2 Score
0	Ridge	0.001	0.794235
1	Lasso	0.001	0.794183

Process finished with exit code 0