Project 1

Manager: Brendan McDonnell
Masum Billah
Madhu Chencharapu
Gabriel Loos
Roshan Ravichandran

February 2026

1 Introduction

In this project, we will explore three different data sets using multiple linear regression, ridge regression, lasso, and transformed regression (specifically sqrt, log1p, and Box-Cox). The three data sets we will be using are:

- **Auto MPG:** This data set contains information about different cars. The goal is predict the miles per gallon (MPG) given the other characteristics of the car. The data set can be found at https://archive.ics.uci.edu/dataset.
- **House Prices:** This data set contains information about different houses. The goal is to predict the price of the house given the other characteristics. The data set can be found at https://www.kaggle.com/datasets/prokshitha/home-value-insights.
- **Insurance Charges:** This data set contains information about different people. The goal is to predict a persons insurance charges given their other characteristics. The data set can be found at https://www.kaggle.com/datasets/mirichoi0218/insurance.

The report will be organized as follows,Section 2 goes over the EDA, Section 3 will be focused on linear regression, Section 4 regularized regression, Section 5 transformed regression, and Section 6 will be comparing all of the models.

In Section 3, in-sample, out-of-sample, 5-fold cross validation, and feature selection results are given for multiple linear regression on all three data sets. The results come from both scalation in scala and statsmodels in python. Section 4 goes over in-sample and out-of-sample quality of fit metrics and plots for Ridge Regression and Lasso. Section 5 gives in-sample, out-of-sample, and feature selection results for the Sqrt, Log1p, and Box-Cox transformations.

2 Exploratory Data Analysis (EDA)

2.1 Auto MPG Dataset

2.1.1 Introduction

This analysis is conducted on the Auto MPG dataset obtained from the UCI Machine Learning Repository. This dataset contains information about various automobiles and their fuel efficiency.

The primary objective of this dataset is to predict the fuel efficiency (miles per gallon - MPG) of a car based on the technical characteristics. This dataset consists of the 398 Observations.

Predictive Variables (Features):
cylinders, displacement, horsepower, weight, acceleration, model_year, origin (If car_name exists, it is treated as non-numeric and is typically excluded for basic linear regression).

Target variable:
mpg (miles per gallon) The prediction task is a regression problem, since the target variable (mpg) is continuous.

2.1.2 Data Preprocessing

Dataset obtained from the UCI Machine Learning Repository.
Retrieved programmatically using the ucimlrepo Python package and ID used: 9 (Auto MPG dataset).
The returned object contains:

1. auto_mpg.data.features → Independent variables (pandas DataFrame)
2. auto_mpg.data.targets → Dependent variable (pandas DataFrame)
3. auto_mpg.metadata → Dataset description and source details
4. auto_mpg.variables → Column-level information

2.1.3 Dataset Dimensions

The dataset contains 398 rows (observations). The dataset contains 8 columns (features + target).

2.1.4 Column Names

The dataset includes the following variables:

1. displacement
2. cylinders
3. horsepower
4. weight
5. acceleration
6. model_year
7. origin
8. mpg (target variable)

Shape of dataset: (398, 8)

2.1.5 Numerical Distribution

Float columns: displacement, horsepower, acceleration, mpg
Integer columns: cylinders, weight, model_year, origin

2.1.6 Missing Value Analysis

horsepower column contains 6 missing values out of 398 observations.
All other variables contain complete data.
Missing percentage in horsepower:

$$\frac{6}{13} \approx 1.5\%$$

Since missing data is very small ($< 5\%$), it does not significantly affect overall dataset structure.
We dropped those records which had null values, since there are only a few records.

2.1.7 Heat Map

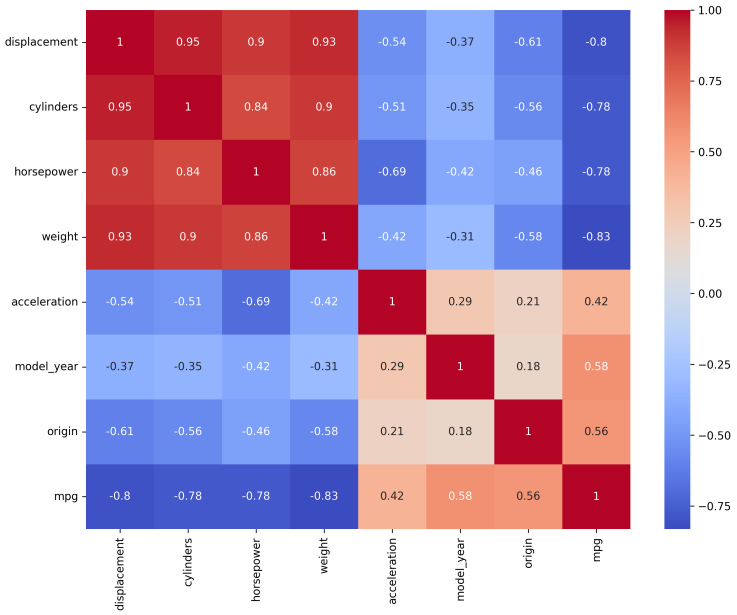


Figure 1: Auto MPG Correlation Heat Map

2.1.8 Correlation Analysis

Strong negative correlation observed between mpg and engine-related features such as weight, displacement, horsepower, and cylinders. model_year and origin show moderate positive correlation with mpg. High inter-correlation among displacement, cylinders, and weight indicates potential multicollinearity.

2.1.9 Box Plots

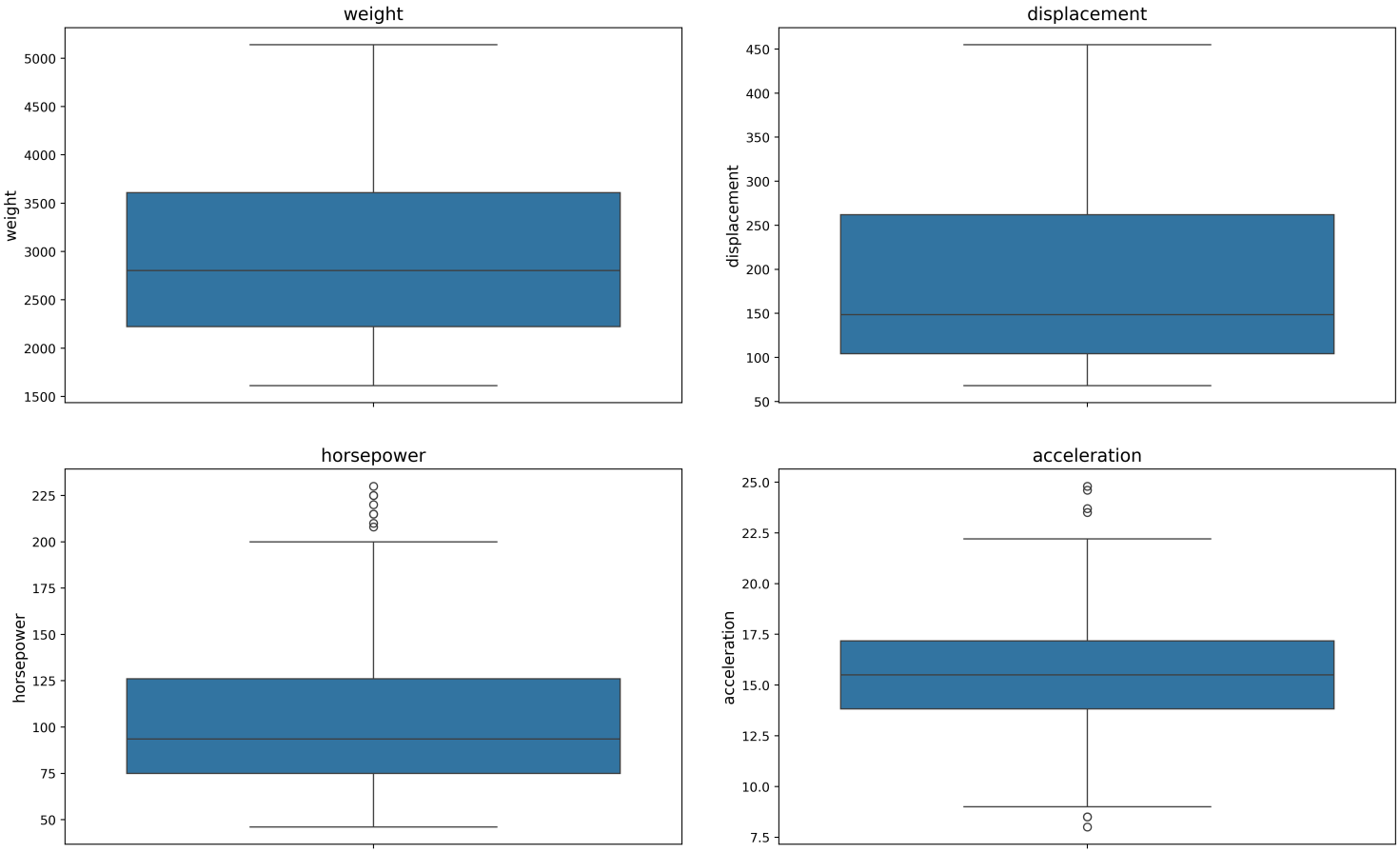


Figure 2: Auto MPG Box Plots

Weight and displacement show wide spread but no extreme outliers.

Horsepower contains a few high-value outliers and is slightly right-skewed.

Acceleration shows moderate spread with minor outliers on both ends.

Overall, variables exhibit reasonable variability; only horsepower may require attention during modeling.

2.1.10 Count Plots

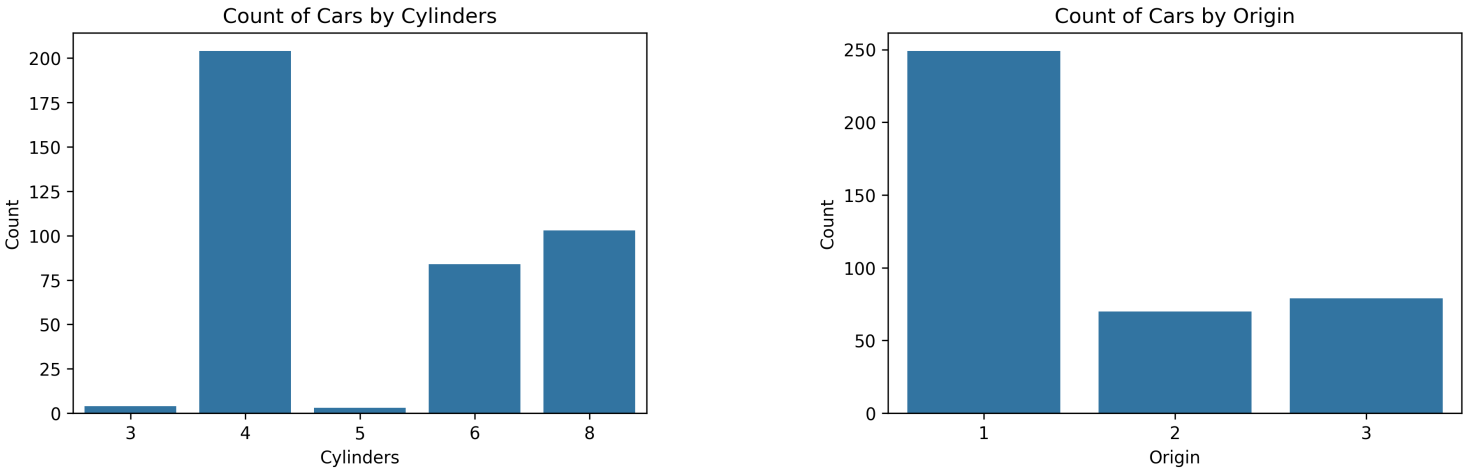


Figure 3: Auto MPG Count Plots

Most vehicles have 4 cylinders, making it the dominant engine type.

Very few vehicles have 3 or 5 cylinders, indicating class imbalance.

A substantial number of vehicles have 6 and 8 cylinders, but significantly fewer than 4-cylinder cars.

Majority of cars originate from Origin 1 (USA). Origins 2 and 3 (Europe and Japan) have fewer observations, showing dataset imbalance across regions.

2.1.11 Scatter Plot Observations

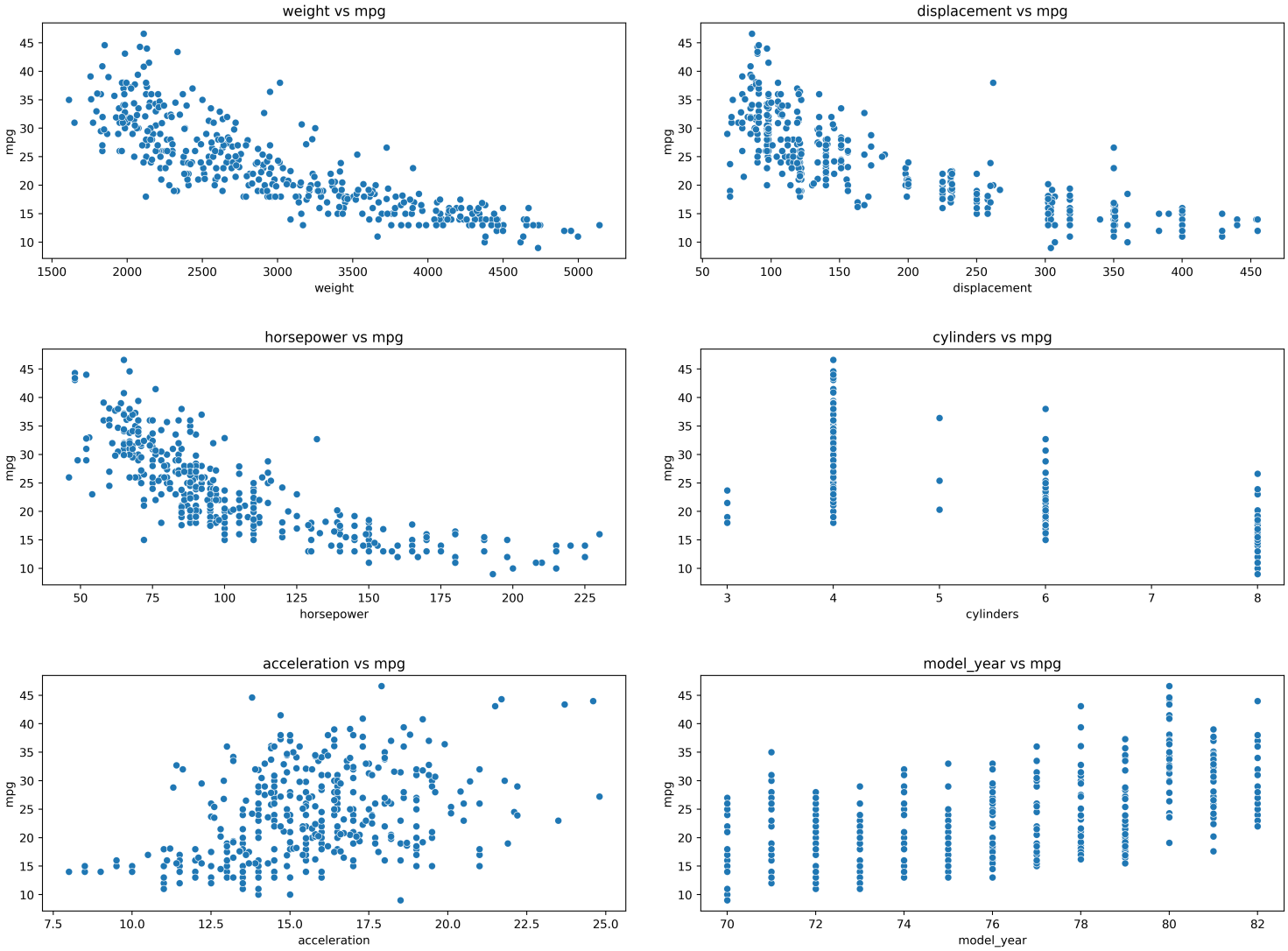


Figure 4: Auto MPG Scatter Plots

Weight, displacement, horsepower, and cylinders show strong negative relationships with mpg — higher engine size and vehicle weight correspond to lower fuel efficiency.

Model_year shows a clear positive trend with mpg — newer cars tend to have better fuel efficiency.

Acceleration shows a weak to moderate positive relationship with mpg.

Relationships appear largely linear, supporting the use of linear regression.

No extreme structural anomalies are observed, though some spread increases for lighter vehicles.

2.2 House Price Regression Dataset

2.2.1 Introduction

This analysis is conducted on the House Price Regression Dataset. The dataset contains structural and location-related attributes of houses along with their market prices. The primary objective of this dataset is to predict the House_Price based on housing characteristics. The dataset consists of 1000 observations.

Predictive Variables (Features):

Square_Footage, Num_Bedrooms, Num_Bathrooms, Year_Built, Lot_Size, Garage_Size, Neighborhood_Quality.

Target Variable:

House_Price. The prediction task is a regression problem, since the target variable (House_Price) is continuous.

2.2.2 Dataset Dimensions

The dataset contains 1000 rows (observations) and 8 columns (7 features + 1 target). Shape of dataset: (1000, 8).

2.2.3 Data Types

- 6 columns are of type int64: Square_Footage, Num_Bedrooms, Num_Bathrooms, Year_Built, Garage_Size, Neighborhood_Quality.
- 2 columns are of type float64: Lot_Size, House_Price.

No object or categorical string columns are present.

2.2.4 Missing Value Analysis

All columns contain 1000 non-null values. No missing values are present. Missing percentage in dataset: 0%. No imputation or deletion was required.

2.2.5 Heat Map

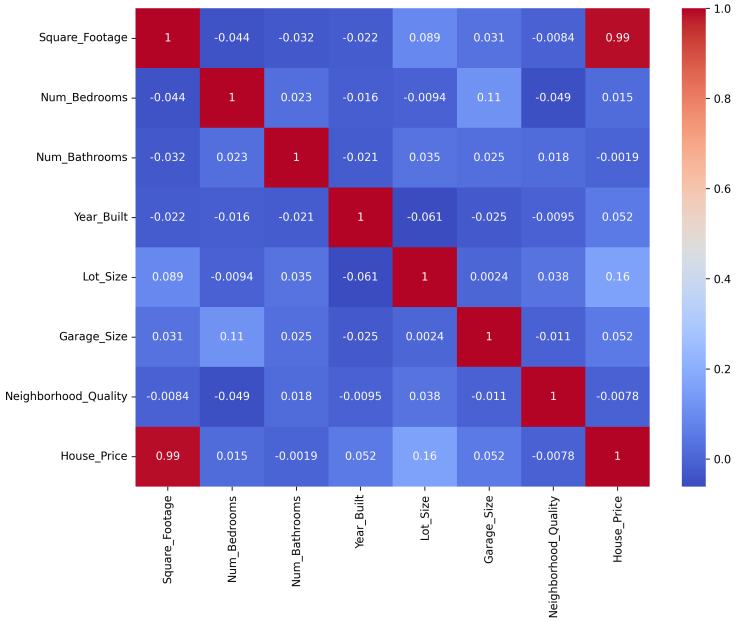


Figure 5: House Price Correlation Heat Map

Square_Footage and House_Price show extremely strong positive correlation (0.99).
Lot_Size shows weak positive correlation (0.16) with House_Price.
Year_Built shows very weak correlation (0.052) with House_Price.
Num_Bedrooms and Num_Bathrooms show near-zero correlation with House_Price.
No strong multicollinearity exists among independent variables.

2.2.6 Distribution Analysis of Continuous Variables

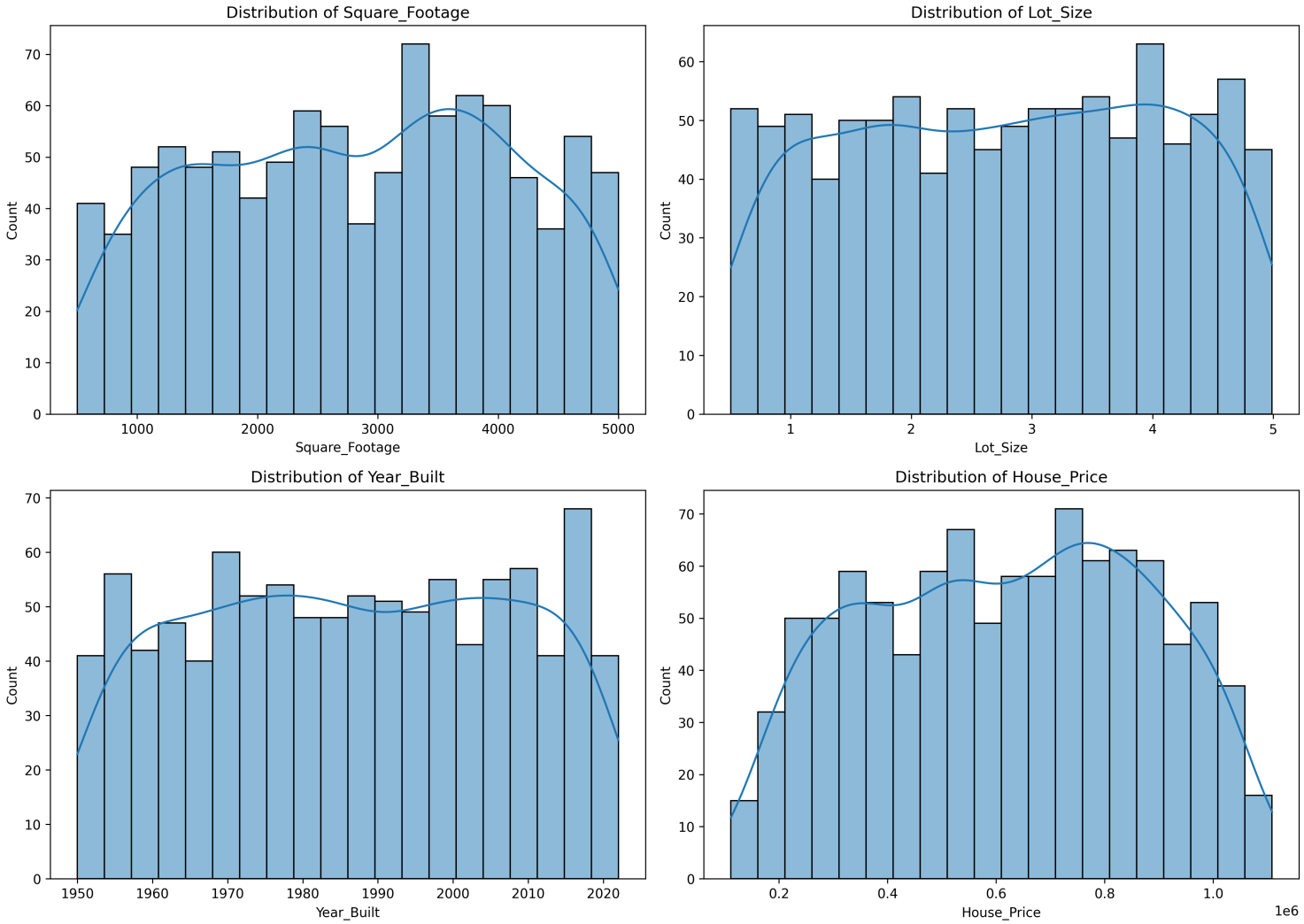


Figure 6: House Price Distribution Analysis of Continuous Variables

Square_Footage: Ranges from approximately 500 to 5000 sq.ft. The distribution appears relatively uniform across the range. No strong skewness is observed. There is substantial variability in property sizes. This wide variation suggests square footage may strongly influence house price.

Lot_Size:

Lot size ranges from approximately 0.5 to 5 units. The distribution is fairly evenly spread. No significant skewness or extreme outliers are visible. The spread indicates moderate variability across properties.

Year_Built:

Houses were constructed between 1950 and 2022. The distribution is relatively balanced across decades. No strong clustering in a specific time period is observed.

House_Price:

House prices range from approximately $111,626 to $1,108,237. The distribution shows moderate spread across the price range. No severe skewness is evident.

2.2.7 Distribution Analysis of Discrete Variables

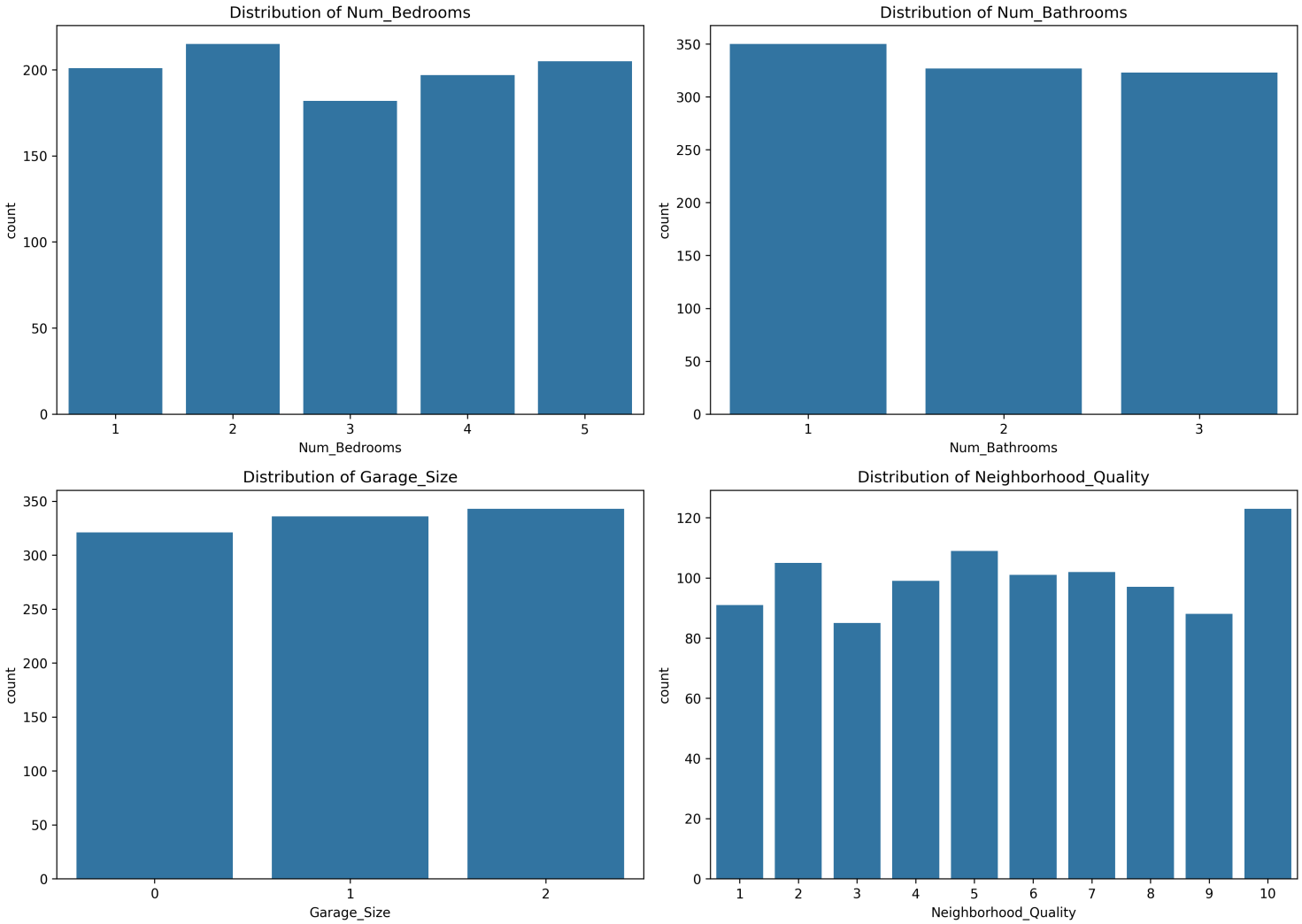


Figure 7: House Price Distribution Analysis of Discrete Variables

Num_Bedrooms:
The number of bedrooms ranges from 1 to 5. The distribution appears relatively balanced across categories. Houses with 2 and 5 bedrooms are slightly more frequent.
Num_Bathrooms:
Bathrooms range from 1 to 3. Properties with 1 bathroom appear slightly more common.
Garage_Size:
Garage size ranges from 0 to 2. The three categories are nearly evenly distributed.
Neighborhood_Quality:
Neighborhood quality ranges from 1 to 10. The distribution is fairly spread across all quality levels. Slightly higher counts appear in mid-to-high quality ranges.

2.2.8 Bivariate Analysis: Features vs House Price

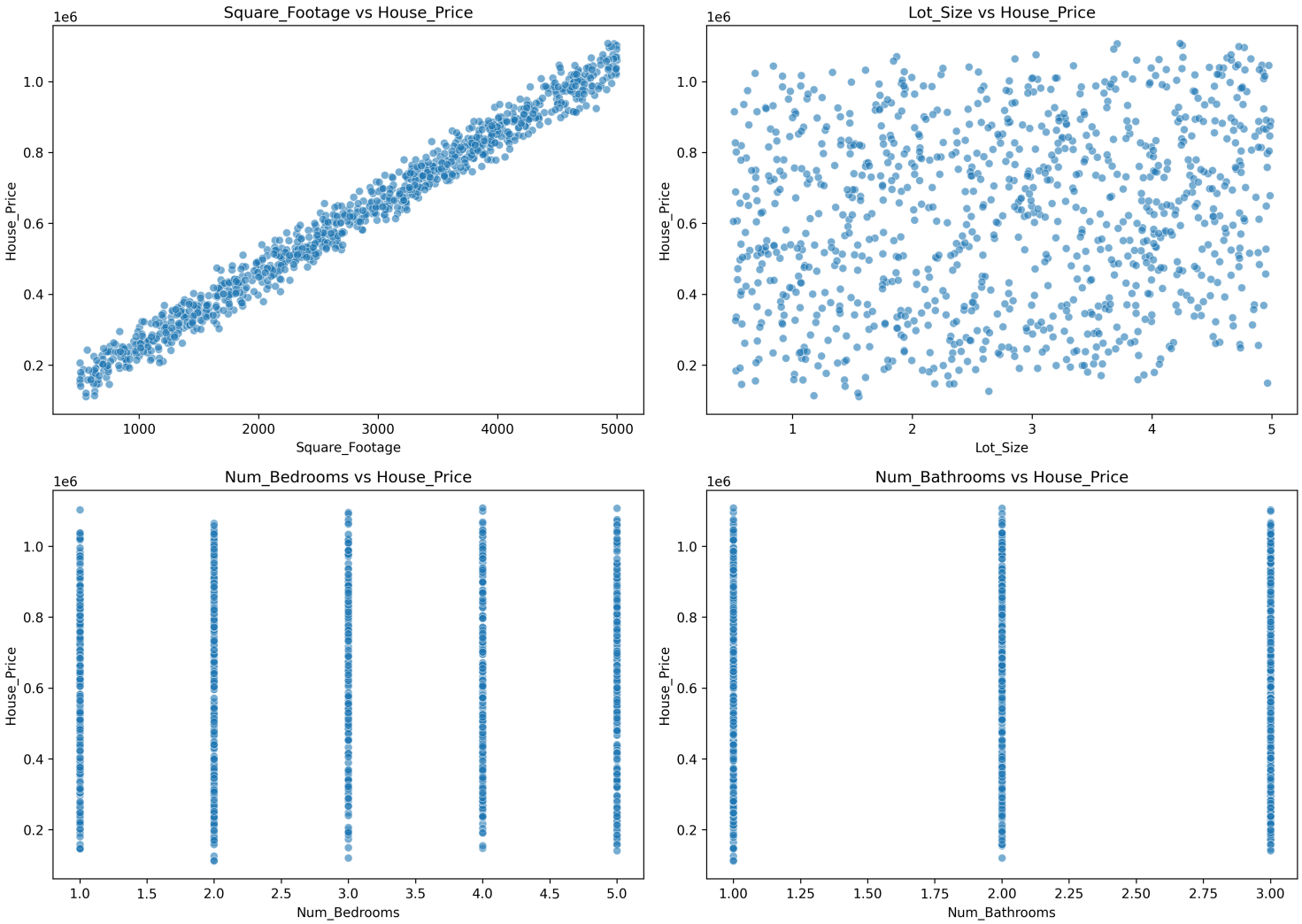


Figure 8: House Price Bivariate Analysis: Features vs House Price

Square_Footage vs House_Price:
A very strong positive linear relationship is observed. As square footage increases, house price increases proportionally. The points are tightly clustered along a straight upward trend. This indicates extremely high correlation (≈ 0.99).
Lot_Size vs House_Price:
The scatter plot shows a widely dispersed pattern. No clear linear trend is visible. The relationship appears weak.
Num_Bedrooms vs House_Price:
Vertical strip pattern is observed due to discrete values (1–5). No strong increasing price trend across bedroom categories. Prices vary significantly within each bedroom count.
Num_Bathrooms vs House_Price:
Similar vertical strip pattern due to discrete values (1–3). No clear monotonic increase in price. Price variation is high within each bathroom category.

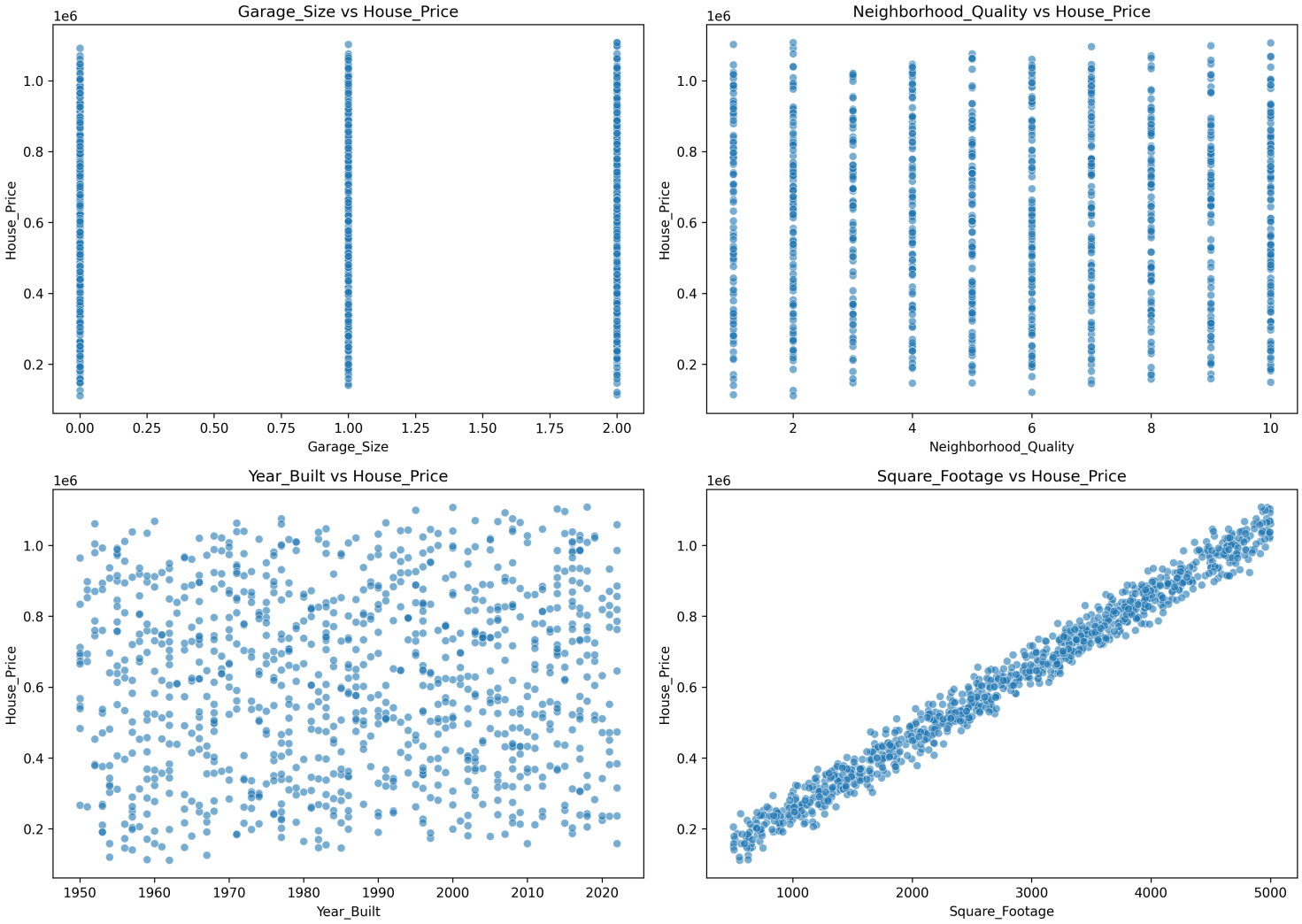


Figure 9: House Price Bivariate Analysis: Features vs House Price

Garage_Size vs House_Price:
A vertical strip pattern is observed due to discrete values (0, 1, 2). Prices vary widely within each garage category. No strong linear upward trend is visible.
Neighborhood_Quality vs House_Price:
Discrete vertical clusters are visible (values 1 to 10). Slight upward tendency is noticeable as quality increases. However, price variation remains high within each category.
Year_Built vs House_Price:
Scatter points are widely dispersed. No clear increasing or decreasing linear trend is visible.

2.3 Medical Cost Personal Dataset

2.3.1 Introduction

This analysis is conducted on the Insurance dataset obtained from Kaggle. This dataset contains demographic and health-related information of individuals along with their medical insurance charges. The primary objective is to predict the medical insurance charges (charges) based on demographic and health-related characteristics.

Predictive Variables (Features):
age, sex, bmi, children, smoker, region.

Target Variable:
charges (medical insurance cost). The prediction task is a regression problem, since the target variable (charges) is continuous.

2.3.2 Dataset Overview

Dataset obtained from Kaggle: https://www.kaggle.com/datasets/mirichoi0218/insurance. The dataset is provided as a CSV file (insurance.csv).

2.3.3 Dataset Dimensions

The dataset contains 1,338 rows (observations) and 7 columns (features + target). Shape of dataset: (1338, 7).

2.3.4 Data Types

- 3 columns are of type int64 → age, children
- 2 columns are of type float64 → bmi, charges
- 3 columns are of type object → sex, smoker, region

2.3.5 Numerical Distribution

- Continuous Variables: age, bmi, charges
- Discrete Numerical Variable: children
- Categorical Variables: sex, smoker, region

2.3.6 Missing Value Analysis

No missing values are present in the dataset. All 1,338 observations are complete. Missing percentage in all columns: 0%. Since there are no missing values, no imputation or deletion is required.

2.3.7 Histogram of Continuous Variables

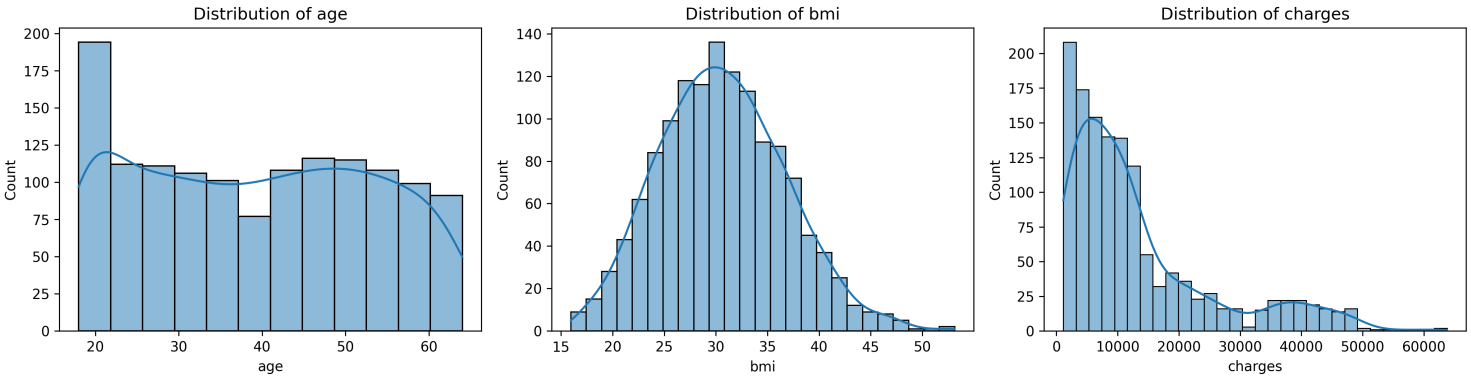


Figure 10: Insurance Charges Histogram of Continuous Variables

Charges distribution is highly right-skewed. Age distribution is fairly uniform across adult range. BMI is approximately normally distributed with slight right skew. Charges show large variance compared to other variables. Skewness in charges suggests possible log transformation.

2.3.8 Boxplots (Outliers)

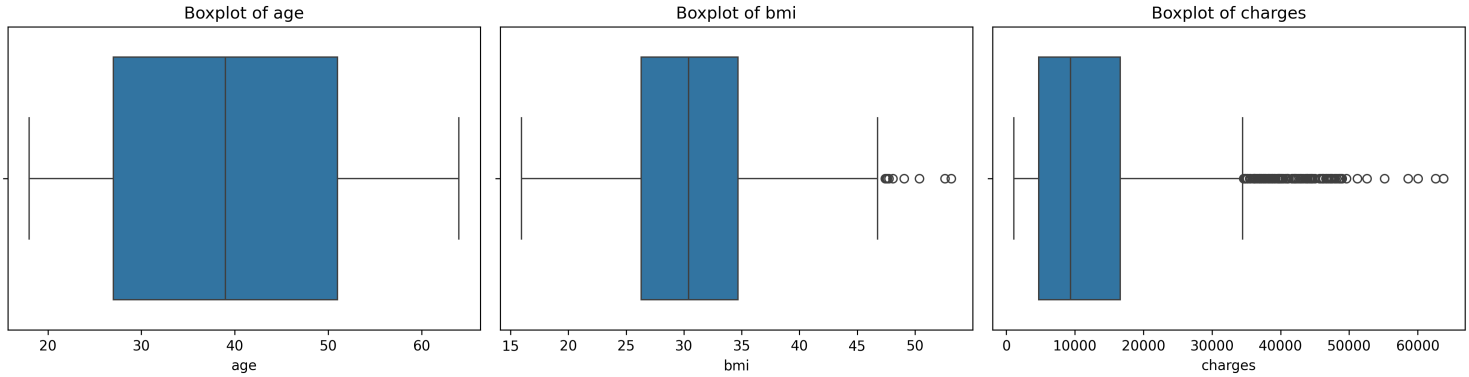


Figure 11: Insurance Charges Boxplots (Outliers)

Charges contain significant high-value outliers. BMI shows moderate outliers. Age does not show extreme outliers. Charges variability is much larger than other variables. Outliers may influence regression performance.

2.3.9 Countplots (Categorical Variables)

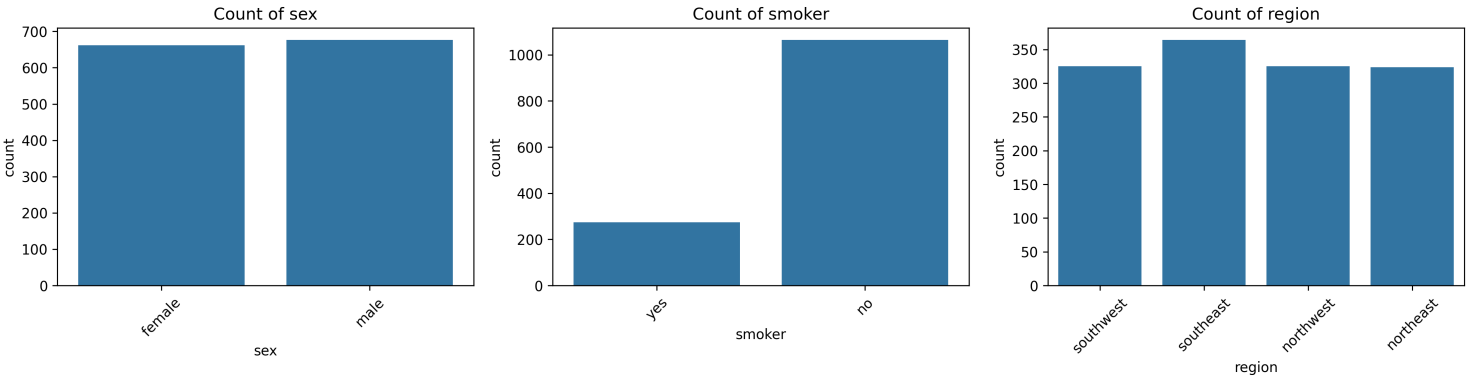


Figure 12: Insurance Charges Countplots (Categorical Variables)

Gender distribution is nearly balanced. Non-smokers are more frequent than smokers. Regions are relatively evenly distributed. No major imbalance in categorical predictors. Smoker category likely impactful despite lower frequency.

2.3.10 Children Distribution

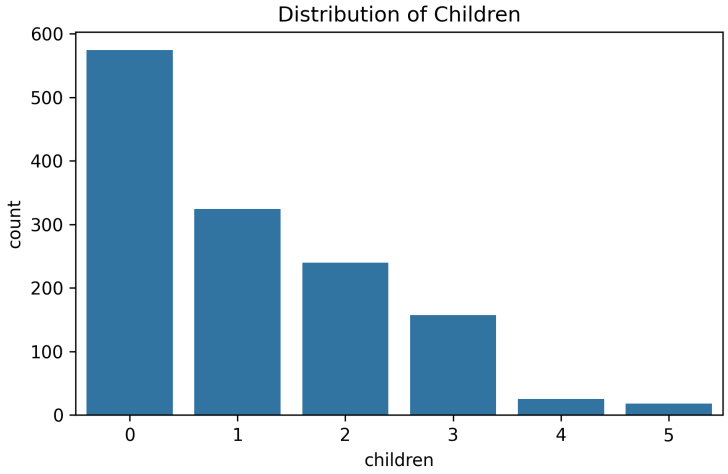


Figure 13: Insurance Charges Children Distribution

Majority have 0–2 children. Few observations for higher number of children. Distribution is right-skewed. Variable has limited range. Impact on charges appears minor but requires further testing.

2.3.11 Scatterplots (Continuous vs Charges)

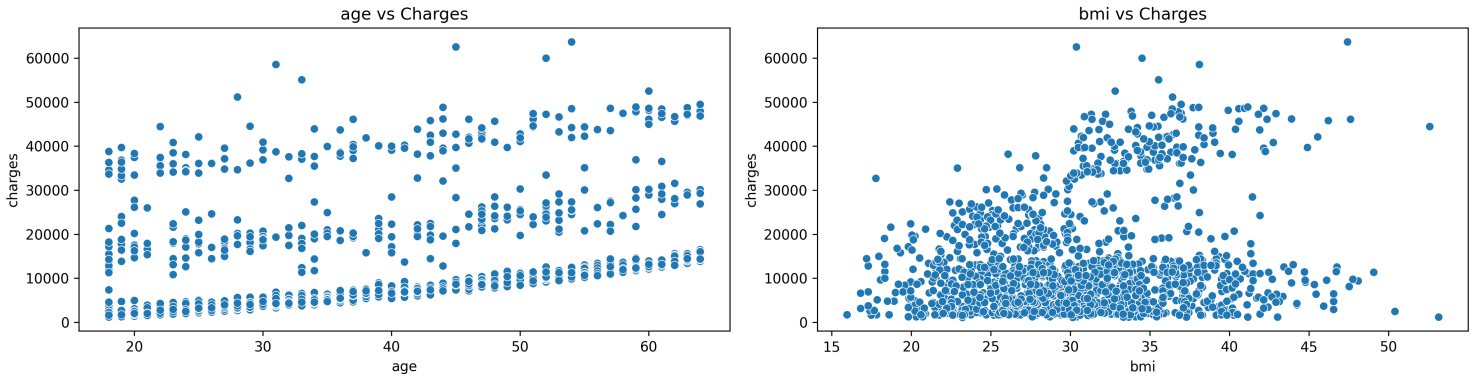


Figure 14: Insurance Charges Scatterplots (Continuous vs Charges)

Age shows positive relationship with charges. BMI shows moderate upward trend with charges. Charges variability increases at higher age and BMI. Evidence of heteroscedasticity. Possible non-linear relationship exists.

2.3.12 Categorical vs Charges (Boxplots)

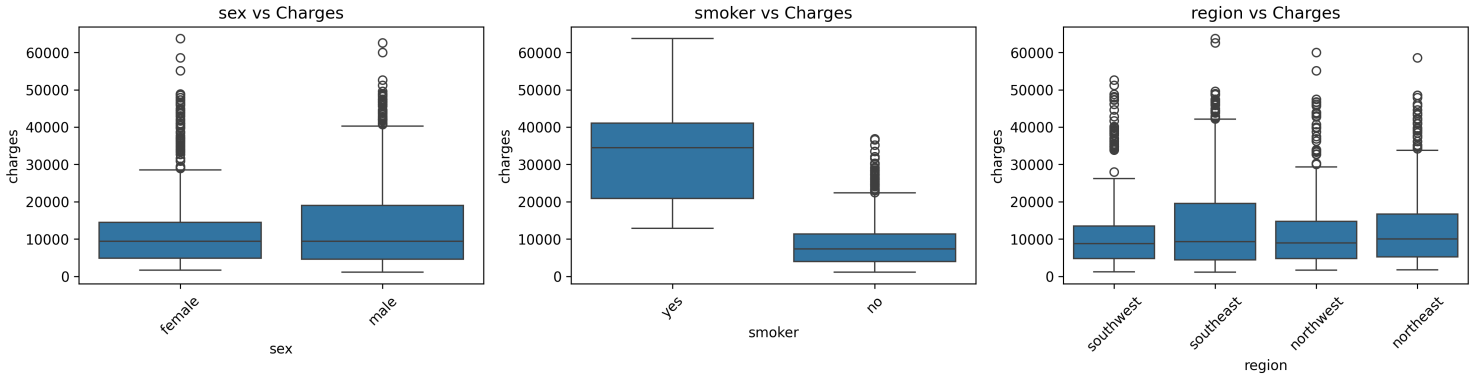


Figure 15: Insurance Charges Categorical vs Charges (Boxplots)

Smokers have significantly higher charges. Clear separation between smoker and non-smoker groups. Gender difference in charges is minimal. Regional differences are small. Smoker variable appears strongest categorical predictor.

2.3.13 Heatmap

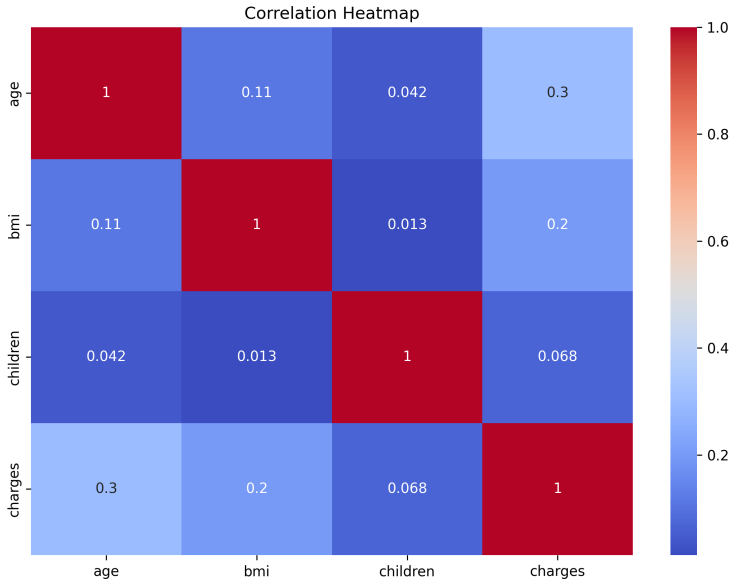


Figure 16: Insurance Charges Correlation Heat Map

Strongest correlation with charges: smoker. Moderate correlation: BMI and age. Weak correlation: children. No severe multicollinearity among numeric features. Age, BMI, and smoker likely key predictors.

3 Linear Regression

In this section, multiple linear regression is applied to the Auto MPG, House Prices, and Insurance Charges data sets. We evaluated the model using in-sample, out-of-sample, and cross validation. Finally, we applied forward selection, backward elimination, and stepwise selection to identify which features are optimal for explanation of the response variable.

3.1 Auto MPG

Table 1 presents the quality of fit metrics for the in-sample and out-of-sample evaluations using scalation, while Table 2 presents the same metrics using statsmodels. We can see that regression is performing decently, and that the main statistics are similar for both scalation and mathstats.

Table 1: Scalation - AutoMPG Linear Regression

Metric	In-Sample	80-20 Split
rSq	0.809255	0.822842
rSqBar	0.806283	0.820081
sst	23819.0	4731.23
sse	4543.35	838.174
sde	3.40878	3.29026
mse0	11.5902	10.7458
rmse	3.40443	3.27808
mae	2.61826	2.48735
smape	12.0589	11.8858
m	392.000	78.0000
dfr	6.00000	6.00000
df	385.000	385.000
fStat	272.234	298.034
aic	-1022.45	-189.284
bic	-994.656	-172.787

Table 2: Statsmodels - Auto MPG Linear Regression

Metric	In-Sample	80-20 Split
rSq	0.8093	0.7942
rSqBar	0.8063	0.7801
sst	23818.9935	4032.2061
sse	4543.3470	829.6873
sde	3.4352	3.3713
mse0	11.8009	10.5024
rmse	3.4352	3.2407
mae	2.6183	2.5039
smape	12.0589	12.3880
m	392.0000	79.0000
dfr	6.0000	6.0000
df	385.0000	73.0000
fStat	272.2341	46.9622
aic	2086.9095	197.7765
bic	2114.7083	211.9932

Figure 17 shows that plots for the predicted y-values vs. the actual y-values from scalation, while Figure 18 show the same from mathstats. Again the results are very similar.

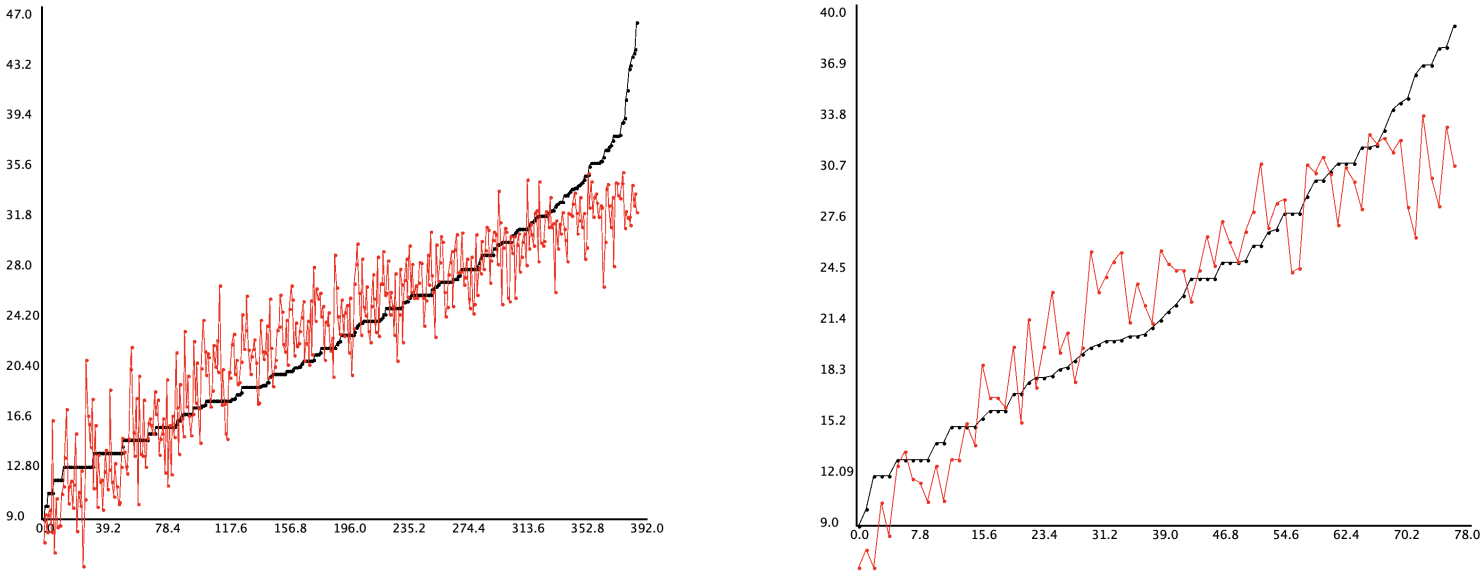


Figure 17: Scalation - Auto MPG Regression
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

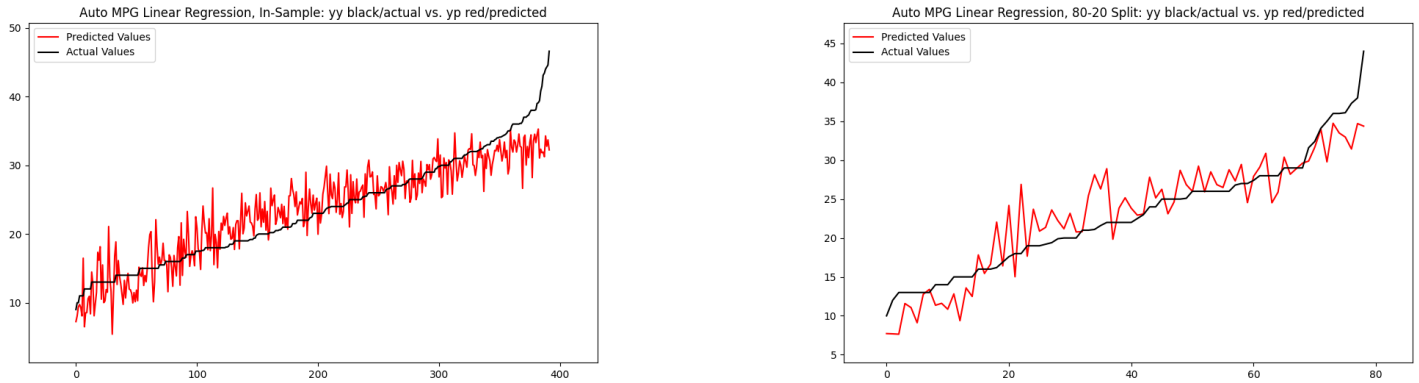


Figure 18: Statsmodels - Auto MPG Regression
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

Next, we performed 5 fold cross validation. Tables 3 and 4 show the resulting quality of fit metrics from scalation and statsmodels respectively. Again we see that the results are similar.

Table 3: Scalation - Auto MPG Linear Regression CV

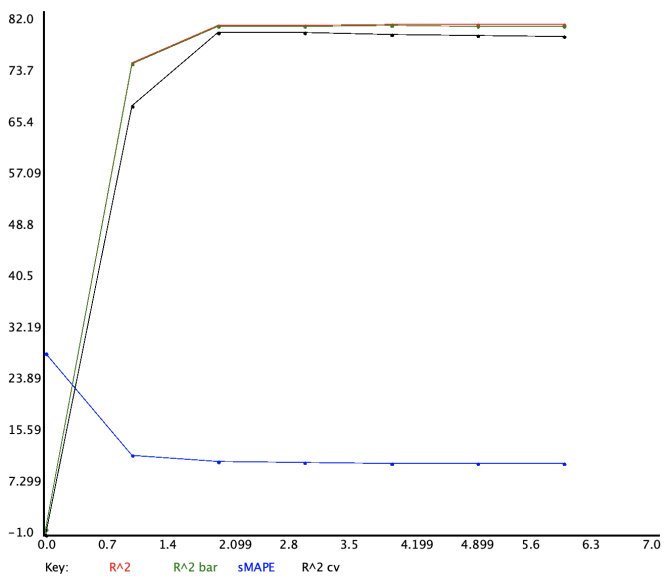
Name	num	min	max	mean	stdev	interval
rSq	5	0.788	0.823	0.798	0.014	0.018
rSqBar	5	0.785	0.820	0.795	0.014	0.018
sst	5	3962.818	5671.580	4700.481	620.767	770.935
sse	5	824.554	1176.435	950.494	142.696	177.215
sde	5	3.177	3.738	3.431	0.226	0.281
mse0	5	10.571	15.083	12.186	1.829	2.272
rmse	5	3.251	3.884	3.483	0.256	0.318
mae	5	2.487	2.850	2.689	0.151	0.188
smape	5	11.886	12.905	12.372	0.427	0.530
m	5	78.000	78.000	78.000	0.000	0.000
dfr	5	6.000	6.000	6.000	0.000	0.000
df	5	385.000	385.000	385.000	0.000	0.000
fStat	5	239.054	298.034	254.430	24.517	30.448
aic	5	-205.110	-188.647	-194.539	6.676	8.291
bic	5	-188.613	-172.150	-178.042	6.676	8.291

Table 4: Statsmodels - Auto MPG Linear Regression CV

Name	In-num folds	min	max	mean	stdev
rSq	5	0.7654	0.8282	0.8010	0.0216
rSqBar	5	0.7491	0.8163	0.7873	0.0231
sst	5	4032.2061	5792.1365	4724.2751	617.8288
sse	5	745.1709	1359.0577	947.8346	213.7052
sde	5	3.2171	4.3446	3.5980	0.3912
mse0	5	9.5535	17.4238	12.0958	2.7627
rmse	5	3.0909	4.1742	3.4576	0.3756
mae	5	2.5039	3.1904	2.6786	0.2601
smape	5	11.2379	14.1325	12.3805	0.9795
m	5	78.0000	79.0000	78.4000	0.4899
dfr	5	6.0000	6.0000	6.0000	0.0000
df	5	72.0000	73.0000	72.4000	0.4899
fStat	5	39.1425	57.8612	49.2658	6.3699
aic	5	188.0386	234.9114	205.6271	15.7223
bic	5	202.1788	249.0516	219.7980	15.7128

Finally, we apply forward selection, backward elimination, and stepwise selection to determine which variables best explain response variable. Figure 19 shows plots for R^2, $\overline{R^2}$, sMAPE, R^2 cv, and AIC vs. n (the number of variables selected) when utilizing forward selection. From the left plot, it is clear that only 2 variables are needed to explain the response variable (even just 1 variable is not too bad either). Those variables 2 variables are weight and modelyear. The order in which the variables were chosen was 1. weight, 2. modelyear, 3. cylinders, 4. acceleration, 5. horsepower, and 6. displacment.

One interesting thing to note is how the graph of AIC vs. n is always increasing.



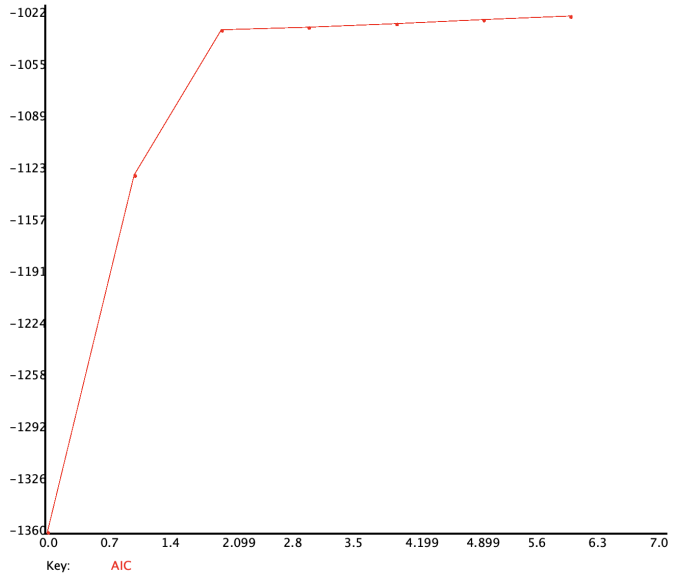
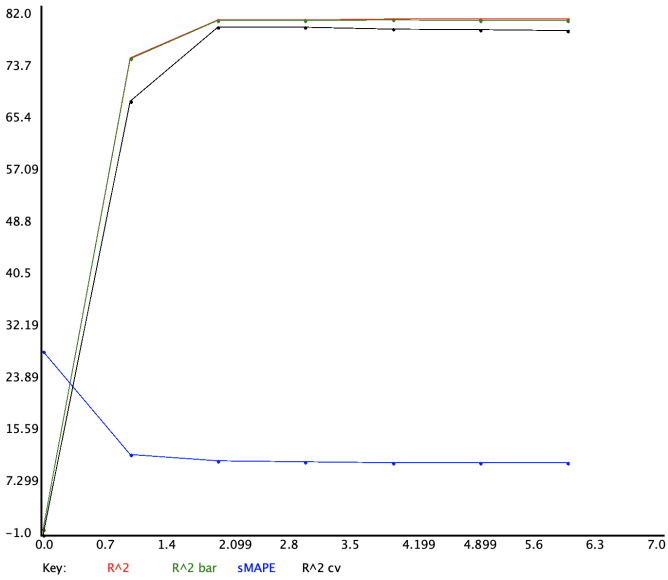


Figure 19: Scalation - Auto MPG Regression Forward Selection
Left: R^2 vs. n
Right: aic vs. n

Figure 20 shows plots for R^2, $\overline{R^2}$, sMAPE, R^2 cv, and AIC vs. n (the number of variables selected) when utilizing backward elimination. From the left plot, it is again clear that only 2 variables are needed to explain the response variable (and again even just 1 variable is not too bad either). Here, both backward and forward selection agree with each other in their selection of variables, so if you were to remove one variable at a time, in order you would remove 1. displacment, 2. horsepower, 3. acceleration, 4. cylinders, 5. modelyear, and 6. weight. This is the reverse order of forward selection, meaning that they agree.



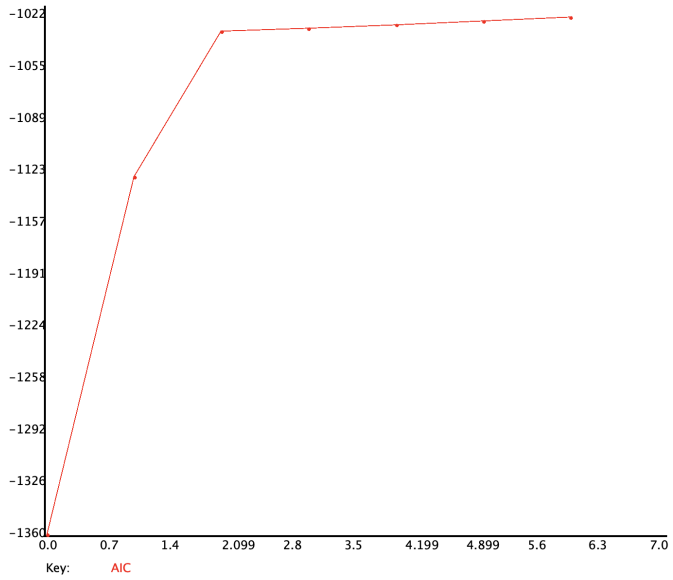


Figure 20: Scalation - Auto MPG Regression Backward Elimination
Left: R^2 vs. n
Right: aic vs. n

Last, Figure 21 shows plots for R^2, $\overline{R^2}$, sMAPE, R^2 cv, and AIC vs. n (the number of variables selected) when utilizing stepwise selection. Here, since stepwise can move forward and backwards which results in multiple different metric evaluations for models that use the same number of variables, (I believe) the evaluation that is plotted for n variables is the best evaluation that stepwise selection found for all the models with n variables that it tested. Despite this, we still find that if we want to get

the best models recommended from stepwise, we should add in the following order, 1. weight, 2. modelyear, 3. cylinders, 4. acceleration, and 5. horsepower. Note that this almost agrees with forward selection and backward elimination, but we do not add displacement as stepwise decided that moving was not worthwhile.

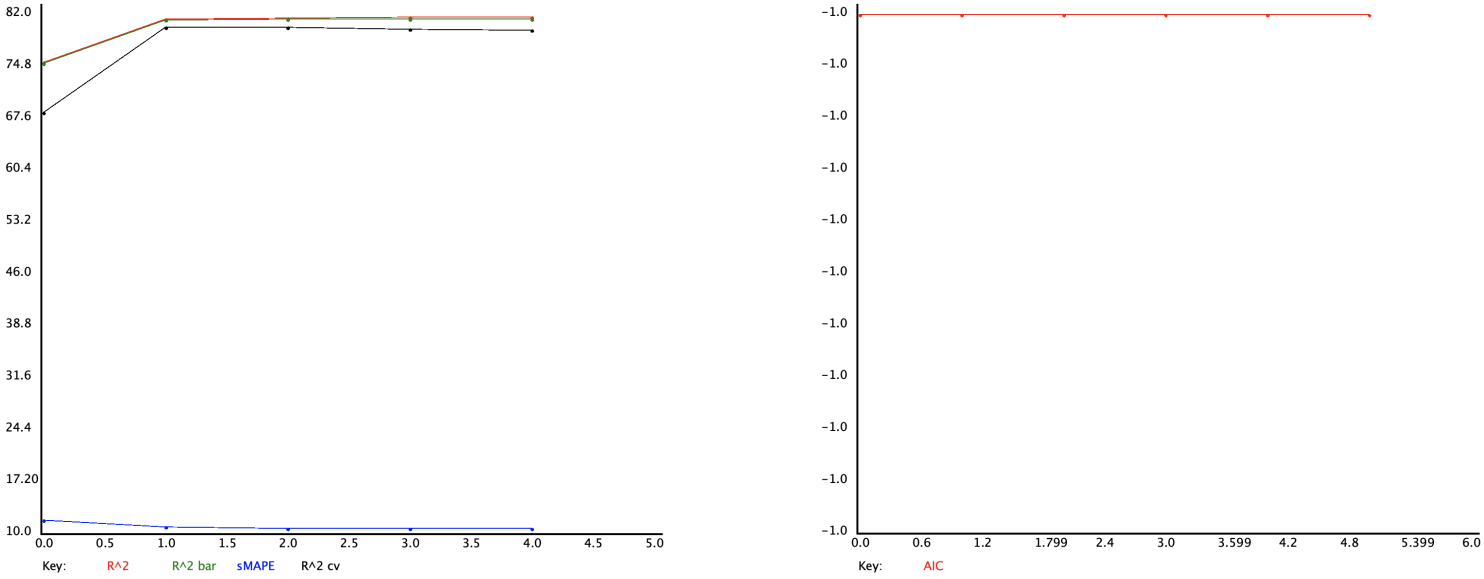


Figure 21: Scalation - Auto MPG Regression Stepwise Selection
Left: R^2 vs. n
Right: aic vs. n

3.2 House Prices

Table 5 presents the quality of fit metrics for the in-sample and out-of-sample evaluations using scalation, while Table 6 presents the same metrics using statsmodels. We can see that regression is performing extremely well, almost perfectly capturing the data. The main statistics are similar for both scalation and mathstats.

Table 5: Scalation - House Price Linear Regression

Metric	In-Sample	80-20 Split
rSq	0.998516	0.998649
rSqBar	0.998507	0.998641
sst	6.42325e+13	1.33700e+13
sse	9.53030e+10	1.80638e+10
sde	9767.21	9501.56
mse0	9.53030e+07	9.03190e+07
rmse	9762.32	9503.63
mae	7747.66	7484.48
smape	1.57791	1.60847
m	1000.00	200.000
dfr	6.00000	6.00000
df	993.000	993.000
fStat	111378	122330
aic	-10591.2	-2101.67
bic	-10556.9	-2078.59

Table 6: Statsmodels - House Price Linear Regression

Metric	In-Sample	80-20 Split
rSq	0.9985	0.9984
rSqBar	0.9985	0.9984
sst	64232463468052.5469	12891771417242.6445
sse	95249090298.3967	20286959701.1265
sde	9798.8381	10252.5012
mse0	96017228.1234	101434798.5056
rmse	9798.8381	10071.4844
mae	7740.4301	8174.5836
smape	1.5774	1.6620
m	1000.0000	200.0000
dfr	7.0000	7.0000
df	992.0000	193.0000
fStat	95425.1583	17493.2676
aic	21225.8831	3700.9854
bic	21265.1451	3724.0736

Figure 22 shows that plots for the predicted y-values vs. the actual y-values from scalation, while Figure 23 show the same from mathstats. Again the results are very similar.

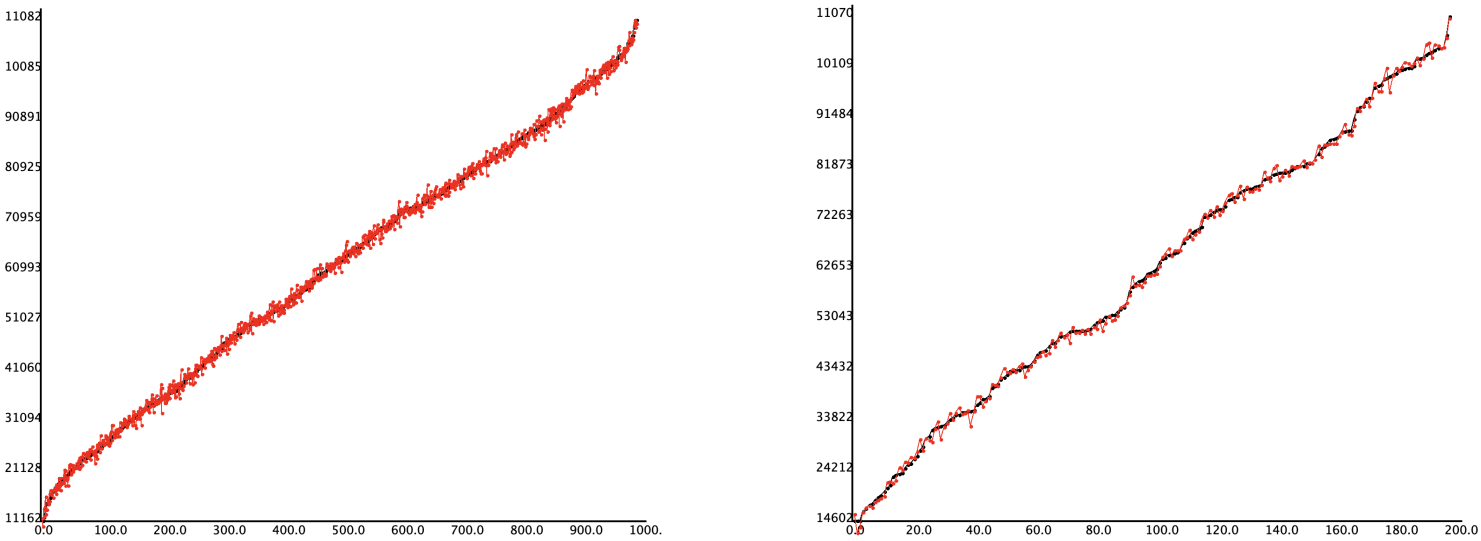


Figure 22: Scalation - House Price Regression
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

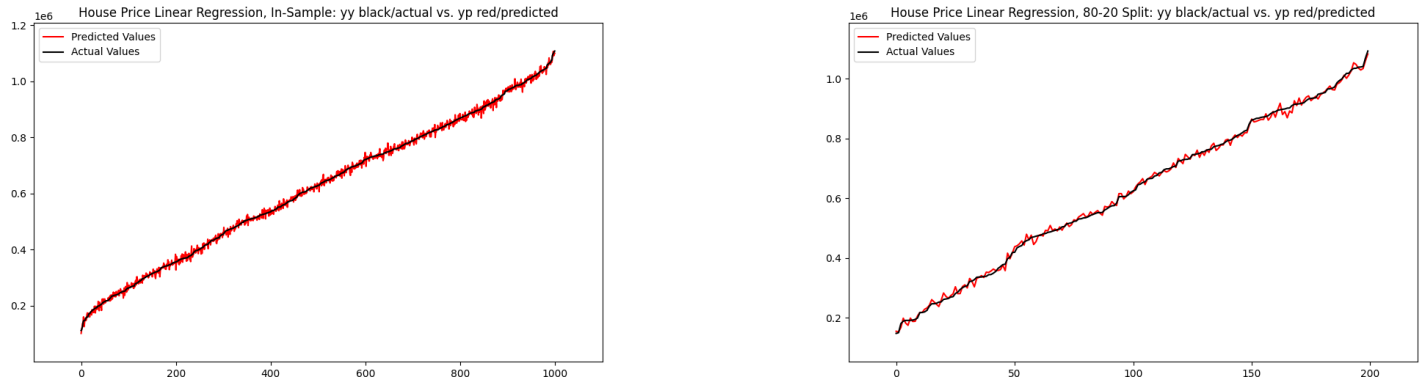


Figure 23: Statsmodels - House Price Regression
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

Next, we performed 5 fold cross validation. Tables 7 and 8 show the resulting quality of fit metrics from scalation and statsmodels respectively. Again we see that the results are similar.

Table 7: Scalation - House Price Linear Regression CV

Name	num folds	min	max	mean	stdev	interval
rSq	5	0.998	0.999	0.998	0.000	0.000
rSqBar	5	0.998	0.999	0.998	0.000	0.000
sst	5	11805100457351.200	13369989427686.710	12829460198984.865	606159535374.424	752793908917.133
sse	5	17790794374.705	22663619750.593	19394433560.564	2117948592.845	2630295668.134
sde	5	9365.942	10608.530	9818.348	533.068	662.021
mse0	5	88953971.874	113318098.753	96972167.803	10589742.964	13151478.341
rmse	5	9431.541	10645.097	9836.093	528.486	656.330
mae	5	7418.536	8609.308	7817.189	534.033	663.219
smape	5	1.408	1.804	1.592	0.141	0.176
m	5	200.000	200.000	200.000	0.000	0.000
dfr	5	6.000	6.000	6.000	0.000	0.000
df	5	993.000	993.000	993.000	0.000	0.000
fStat	5	96001.466	122329.999	110142.703	10843.190	13466.236
aic	5	-2127.278	-2100.155	-2109.082	11.789	14.641
bic	5	-2104.190	-2077.067	-2085.993	11.789	14.641

Table 8: Statsmodels - House Price Linear Regression CV

Name	In-num folds	min	max	mean	stdev
rSq	5	0.9984	0.9986	0.9985	0.0001
rSqBar	5	0.9984	0.9986	0.9985	0.0001
sst	5	12384264865084.2500	13564576490393.4668	12842056251568.3359	395911885307.9127
sse	5	17238729843.0851	20996608313.5931	19277864710.9665	1296174615.0581
sde	5	9450.9176	10430.2788	9988.5569	337.6978
mse0	5	86193649.2154	104983041.5680	96389323.5548	6480873.0753
rmse	5	9284.0535	10246.1232	9812.2003	331.7354
mae	5	7516.9137	8174.5836	7794.6342	224.6310
smape	5	1.5104	1.6620	1.5879	0.0540
m	5	200.0000	200.0000	200.0000	0.0000
dfr	5	7.0000	7.0000	7.0000	0.0000
df	5	193.0000	193.0000	193.0000	0.0000
fStat	5	17493.2676	20276.9346	18399.0661	1001.2402
aic	5	3668.4214	3707.8619	3690.3225	13.5977
bic	5	3691.5096	3730.9501	3713.4107	13.5977

Finally, we apply forward selection, backward elimination, and stepwise selection to determine which variables best explain response variable. Figure 24 shows plots for R^2, $\overline{R^2}$, sMAPE, R^2 cv, and AIC vs. n (the number of variables selected) when utilizing forward selection. From the left plot, it is clear that only 1 variable is needed to explain the response variable. The order in which the variables were chosen was 1. Square Footage, 2. Year Built, 3. Lot Size, 4. Num Bedrooms, 5. Num Bathrooms, and 6. Garage Size.

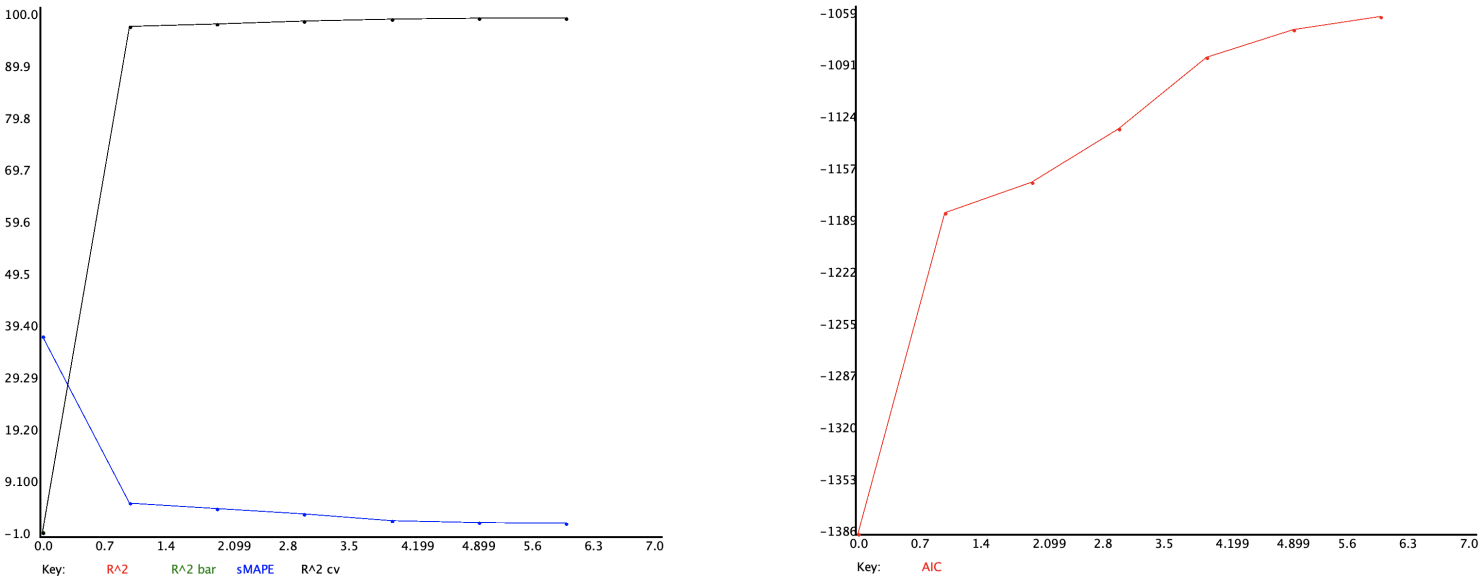
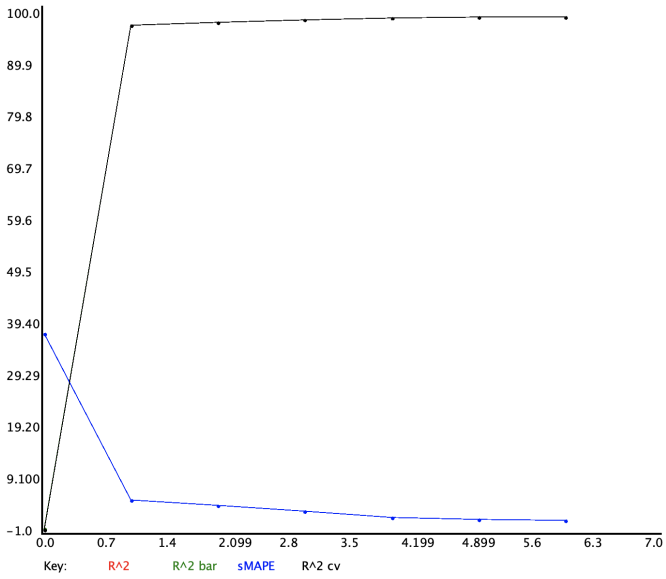


Figure 24: Scalation - House Prices Regression Forward Selection
Left: R^2 vs. n
Right: aic vs. n

Figure 25 shows plots for R^2, $\overline{R^2}$, sMAPE, R^2 cv, and AIC vs. n (the number of variables selected) when utilizing backward elimination. From the left plot, it is again clear that only 1 variable is needed to explain the response variable. Here, both backward and forward selection agree with each other in their selection of variables, so if you were to remove one variable at a time, in order you would remove 1. Garage Size, 2. Num Bathrooms, 3. Num Bedrooms, 4. Lot Size, 5. Year Built, and 6. Square Footage. This is the reverse order of forward selection, meaning that they agree.



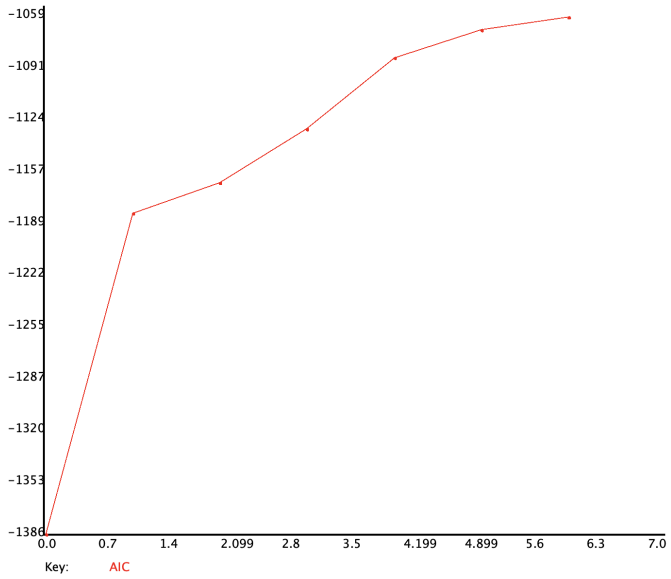
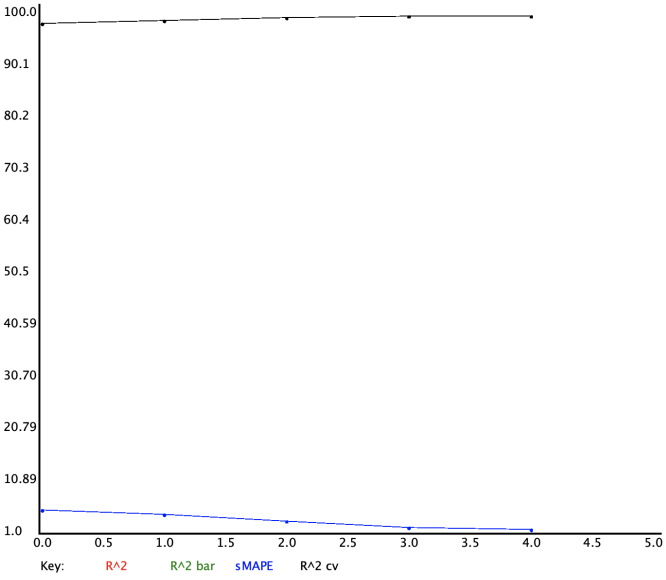


Figure 25: Scalation - House Prices Regression Backward Elimination
Left: R^2 vs. n
Right: aic vs. n

Last, Figure 26 shows plots for R^2, $\overline{R^2}$, sMAPE, R^2 cv, and AIC vs. n (the number of variables selected) when utilizing stepwise selection. We still find that if we want to get the best models recommended from stepwise, we should add in the following order, 1. Square Footage, 2. Year Built, 3. Lot Size, 4. Num Bedrooms, and 5. Num Bathrooms. Note that this almost agrees with forward selection and backward elimination, but we do not add Garage Size as stepwise decided that moving was not worthwhile.



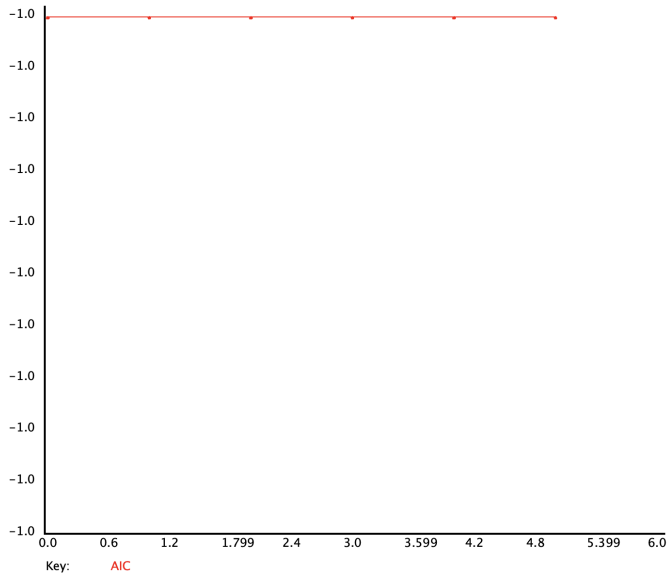


Figure 26: Scalation - House Prices Regression Stepwise Selection
Left: R^2 vs. n
Right: aic vs. n

3.3 Insurance Charges

Table 9 presents the quality of fit metrics for the in-sample and out-of-sample evaluations using scalation, while Table 10 presents the same metrics using statsmodels. We can see that regression is performing okay, and that the main statistics are similar for both scalation and mathstats.

Table 9: Scalation - Insurance Charges Linear Regression

Metric	In-Sample	80-20 Split
rSq	0.750157	0.720005
rSqBar	0.748842	0.718531
sst	1.96074e+11	4.06432e+10
sse	4.89878e+10	1.13799e+10
sde	6053.11	6540.46
mse0	3.66127e+07	4.26213e+07
rmse	6050.84	6528.50
mae	4179.54	4430.66
smape	37.9722	40.0602
m	1338.00	267.000
dfr	7.00000	7.00000
df	1330.00	1330.00
fStat	570.477	488.584
aic	-13533.8	-2708.17
bic	-13492.2	-2679.47

Table 10: Statsmodels - Insurance Charges Linear Regression

Metric	In-Sample	80-20 Split
rSq	0.7509	0.7836
rSqBar	0.7494	0.7778
sst	196074221568.3671	41606660039.7953
sse	48839532843.9219	9003973448.1649
sde	6062.1023	5884.7827
mse0	36749084.1564	33596915.8514
rmse	6062.1023	5796.2847
mae	4170.8869	4181.1945
smape	37.8059	40.0220
m	1338.0000	268.0000
dfr	8.0000	8.0000
df	1329.0000	260.0000
fStat	500.8107	117.6800
aic	27113.5058	4660.4252
bic	27160.2962	4689.1531

Figure 27 shows that plots for the predicted y-values vs. the actual y-values from scalation, while Figure 28 show the same from mathstats. Again the results are very similar.

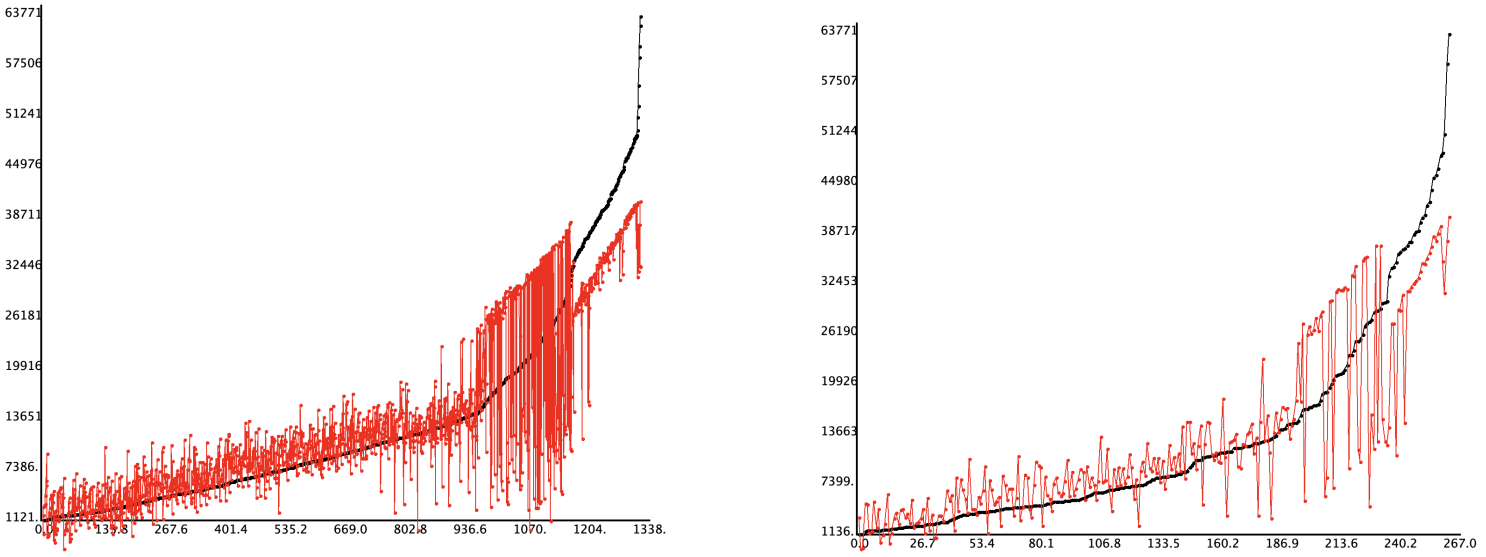


Figure 27: Scalation - Insurance Charges Regression
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

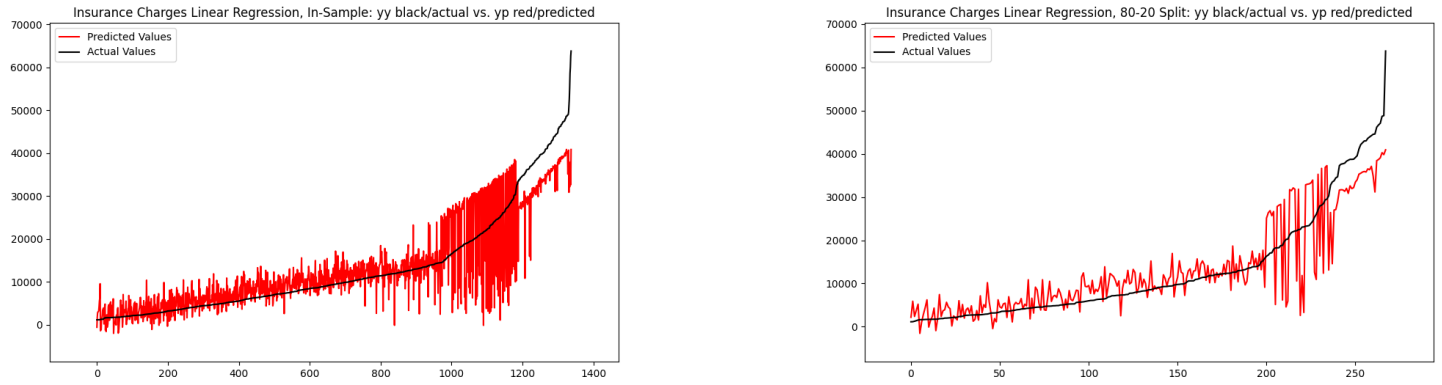


Figure 28: Statsmodels - Insurance Charges Regression
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

Next, we performed 5 fold cross validation. Tables 11 and 12 show the resulting quality of fit metrics from scalation and statsmodels respectively. Again we see that the results are similar.

Table 11: Scalation - Insurance Charges Linear Regression CV

Name	num	min	max	mean	stdev	interval
rSq	5	0.701	0.814	0.743	0.046	0.057
rSqBar	5	0.699	0.813	0.742	0.046	0.057
sst	5	31902777173.848	43430486230.657	38949749086.422	4343029725.651	5393640012.108
sse	5	7539480548.420	11454966761.392	9918152392.401	1670500799.560	2074607019.047
sde	5	5320.957	6562.018	6084.654	526.821	654.263
mse0	5	28237754.863	42902497.234	37146638.174	6256557.302	7770063.742
rmse	5	5313.921	6550.000	6076.688	525.006	652.009
mae	5	3810.754	4511.498	4216.703	292.688	363.491
smape	5	36.128	40.060	38.143	1.833	2.277
m	5	267.000	267.000	267.000	0.000	0.000
dfr	5	7.000	7.000	7.000	0.000	0.000
df	5	1330.000	1330.000	1330.000	0.000	0.000
fStat	5	444.882	830.519	573.015	157.534	195.643
aic	5	-2709.047	-2663.110	-2691.017	19.599	24.340
bic	5	-2680.349	-2634.412	-2662.319	19.599	24.340

Table 12: Statsmodels - Insurance Charges Linear Regression CV

Name	In-num folds	min	max	mean	stdev
rSq	5	0.6324	0.7956	0.7402	0.0578
rSqBar	5	0.6225	0.7901	0.7332	0.0593
sst	5	30189024179.7055	43857198758.4016	39154186092.5516	4823002859.8713
sse	5	8965018845.2774	11096336332.7613	9899546225.2180	821555419.0107
sde	5	5872.0390	6545.4562	6170.1628	260.9758
mse0	5	33451562.8555	41559312.1077	36998683.9155	3128575.4360
rmse	5	5783.7326	6446.6512	6077.2269	256.8986
mae	5	4054.1099	4427.9335	4203.4121	129.0554
smape	5	35.6194	40.0220	38.1279	1.5723
m	5	267.0000	268.0000	267.6000	0.4899
dfr	5	8.0000	8.0000	8.0000	0.0000
df	5	259.0000	260.0000	259.6000	0.4899
fStat	5	55.7054	126.4912	97.8508	24.6138
aic	5	4659.2632	4699.8828	4678.3124	16.0624
bic	5	4687.9911	4728.5808	4707.0283	16.0528

Finally, we apply forward selection, backward elimination, and stepwise selection to determine which variables best explain response variable. Figure 29 shows plots for R^2, $\overline{R^2}$, sMAPE, R^2 cv, and AIC vs. n (the number of variables selected) when utilizing forward selection. From the left plot, it is clear that only s variables are needed to explain the response variable. The order in which the variables were chosen was 1. smoker_yes, 2. age, 3. bmi, 4. children, 5. region_southeast, 6. sex_male, and 7. region_northwest

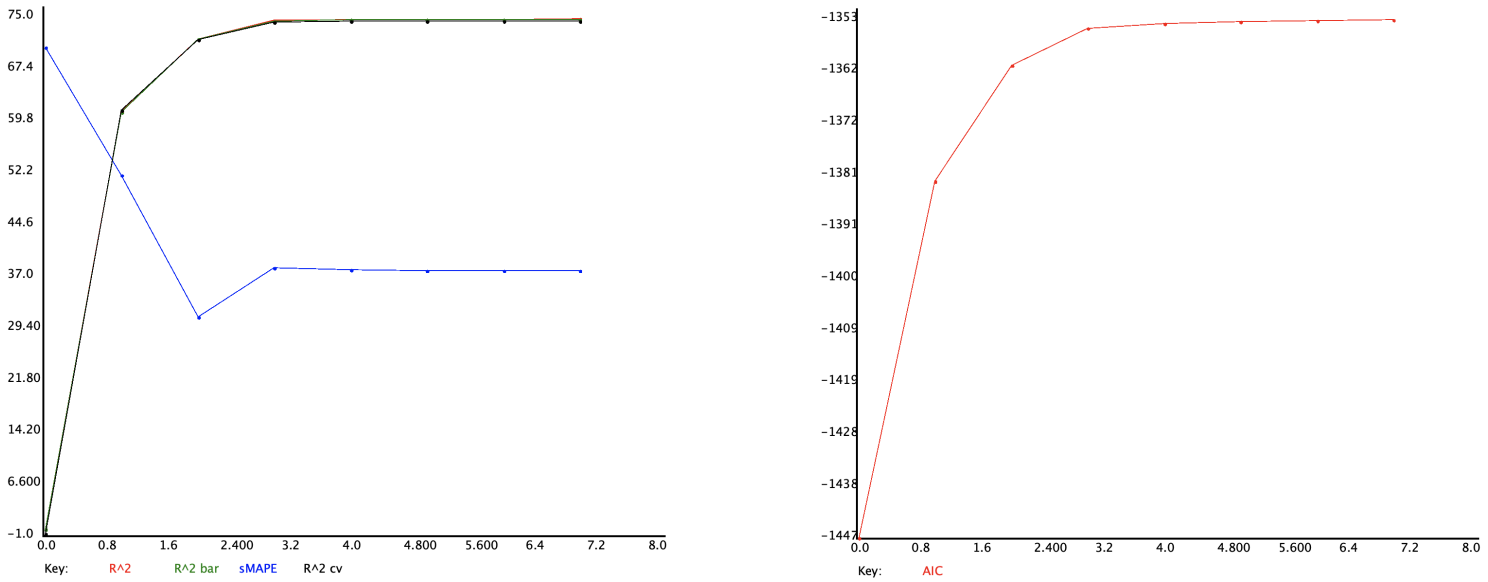


Figure 29: Scalation - Insurance Charges Regression Forward Selection
Left: R^2 vs. n
Right: aic vs. n

Figure 30 shows plots for R^2, $\overline{R^2}$, sMAPE, R^2 cv, and AIC vs. n (the number of variables selected) when utilizing backward elimination. From the left plot, it is again clear that only 2 variables are needed to explain the response variable. Here, both backward and forward selection agree with each other in their selection of variables, so if you were to remove one variable at a time, in order you would remove 1. region_northwest, 2. sex_male, 3. region_southeast, 4. children, 5. bmi, 6. age, and 7. smoker_yes. This is the reverse order of forward selection, meaning that they agree.

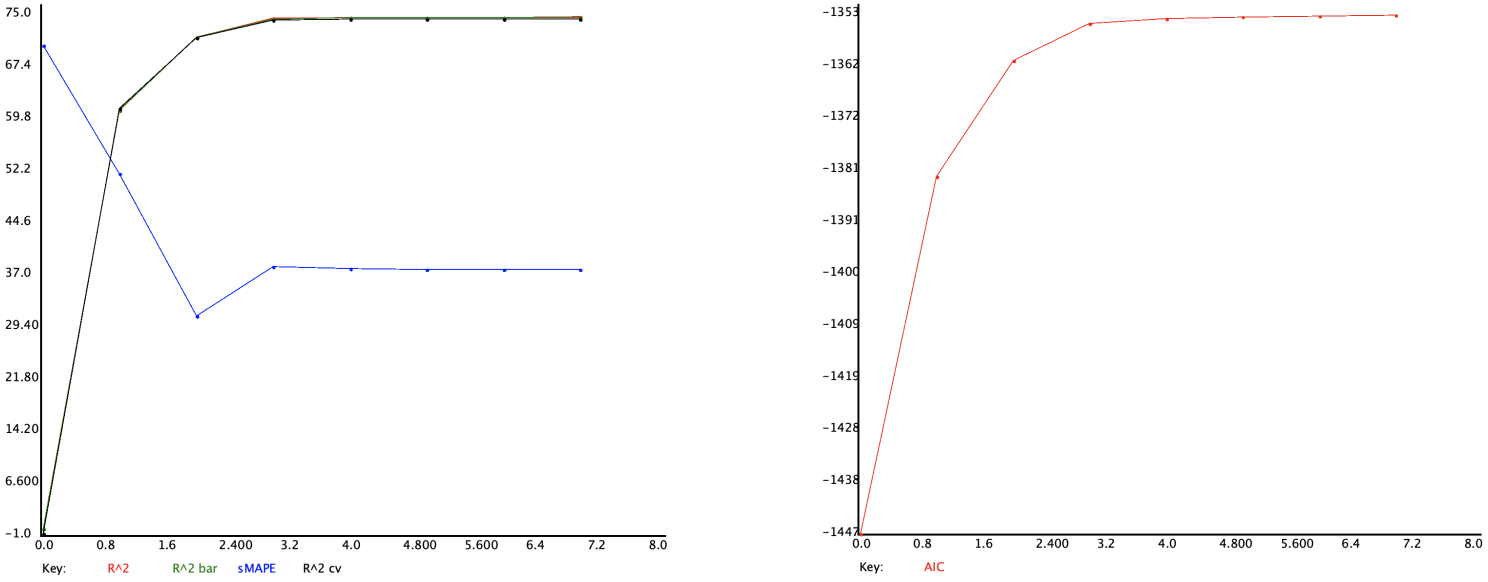


Figure 30: Scalation - Insurance Charges Regression Backward Elimination
Left: R^2 vs. n
Right: aic vs. n

Last, Figure 31 shows plots for R^2, $\overline{R^2}$, sMAPE, R^2 cv, and AIC vs. n (the number of variables selected) when utilizing stepwise selection. We still find that if we want to get the best models recommended from stepwise, we should add in the following order, 1. smoker_yes, 2. age, 3. bmi, 4. children, 5. region_southeast, and 6. sex_male. Note that this almost agrees with forward selection and backward elimination, but we do not add region_northwest as stepwise decided that moving was not worthwhile.

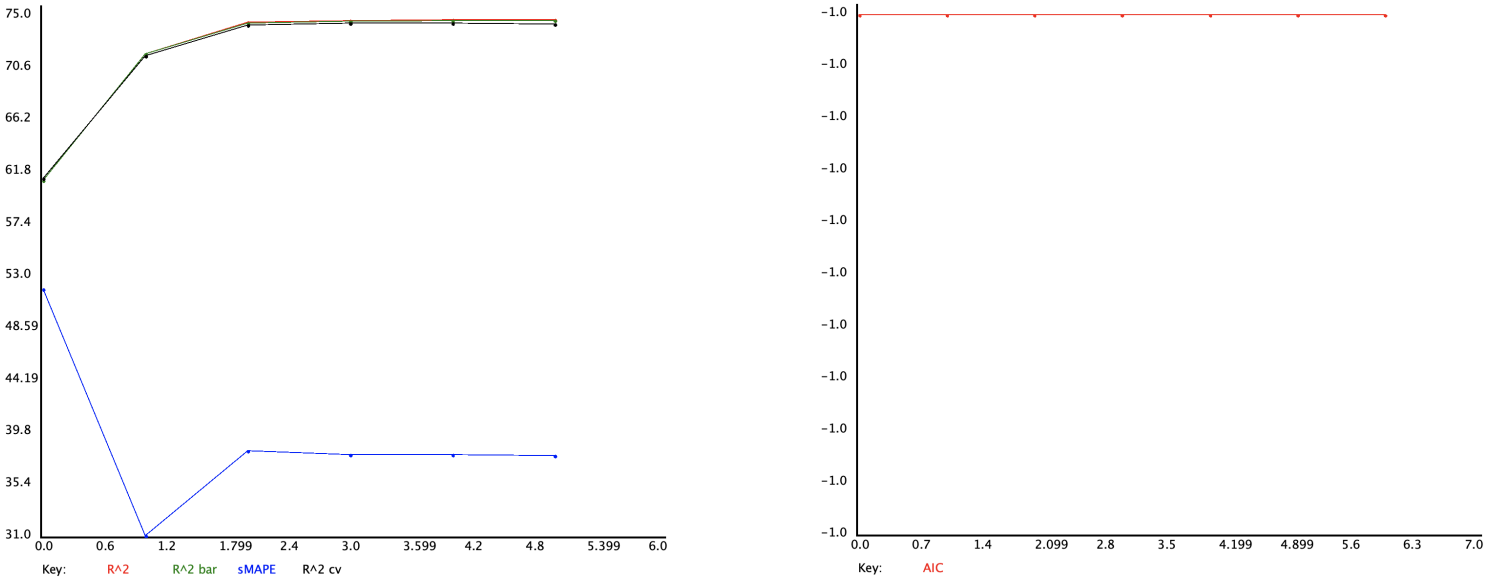


Figure 31: Scalation - Insurance Charges Regression Stepwise Selection
Left: R^2 vs. n
Right: aic vs. n

4 Regularized Regression

For regularized regression, we applied both lasso and ridge regression to each dataset and compared the performance from each method among each dataset. More specifically, in scala we applied all forms of feature selection alongside cross validation to ridge regression, while applying forward selection to lasso regression. In statsmodels we split each dataset into in-sample and 80–20 splits and tested them on the lasso and ridge models.

4.1 Auto MPG

Table 13: Lasso and Ridge Regression AutoMPG results

Metric	Ridge	Lasso
rSq	0.809255	0.804643
rSqBar	0.806283	0.802112
sst	23819.0	23819.0
sse	4543.35	4653.21
sde	3.40878	3.44970
mse0	11.5902	11.8704
rmse	3.40443	3.44535
mae	2.61826	2.62953
smape	64.8110	12.0246
m	392.000	392.000
dfr	6.00000	5.00000
df	385.000	386.000
fStat	272.234	317.974
aic	-1022.45	-1029.14
bic	-994.656	-1005.31

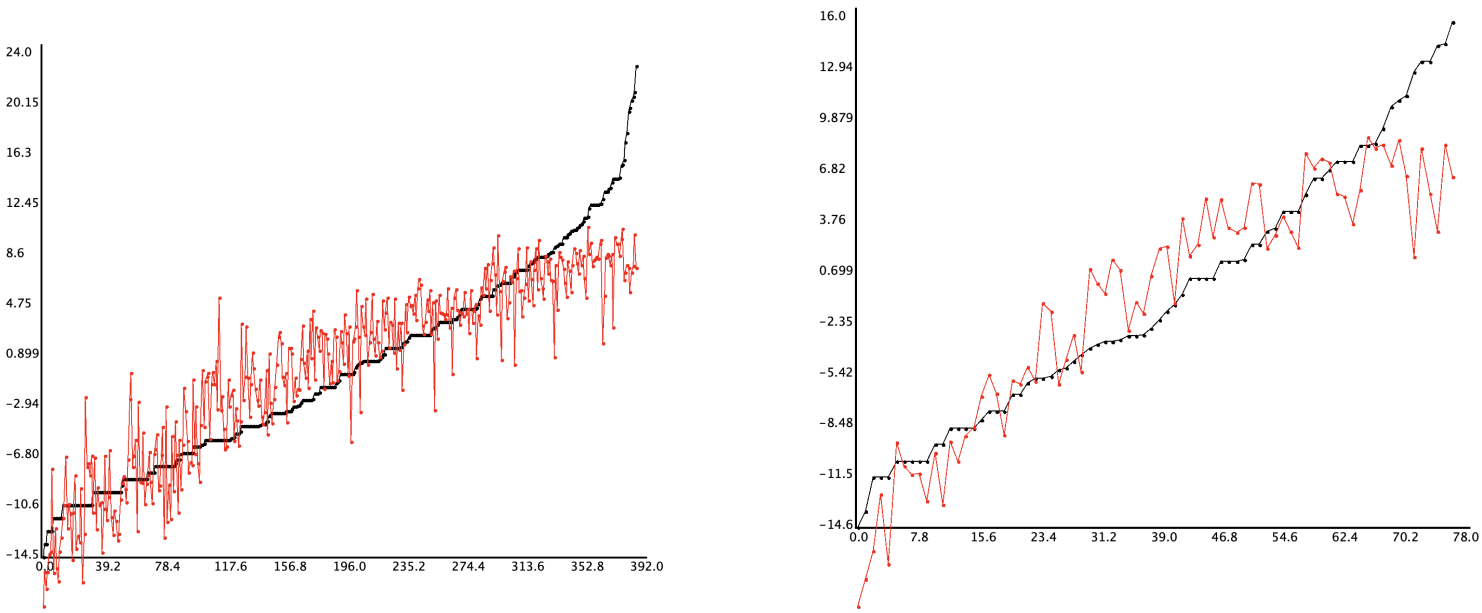


Figure 32: Scalation - Auto MPG Ridge
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

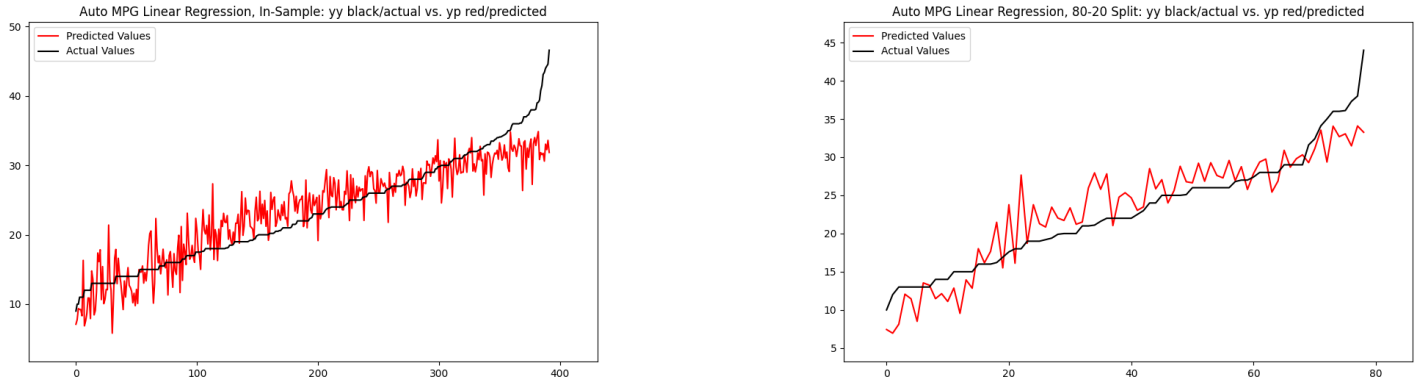


Figure 33: Statsmodels - Auto MPG Ridge
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

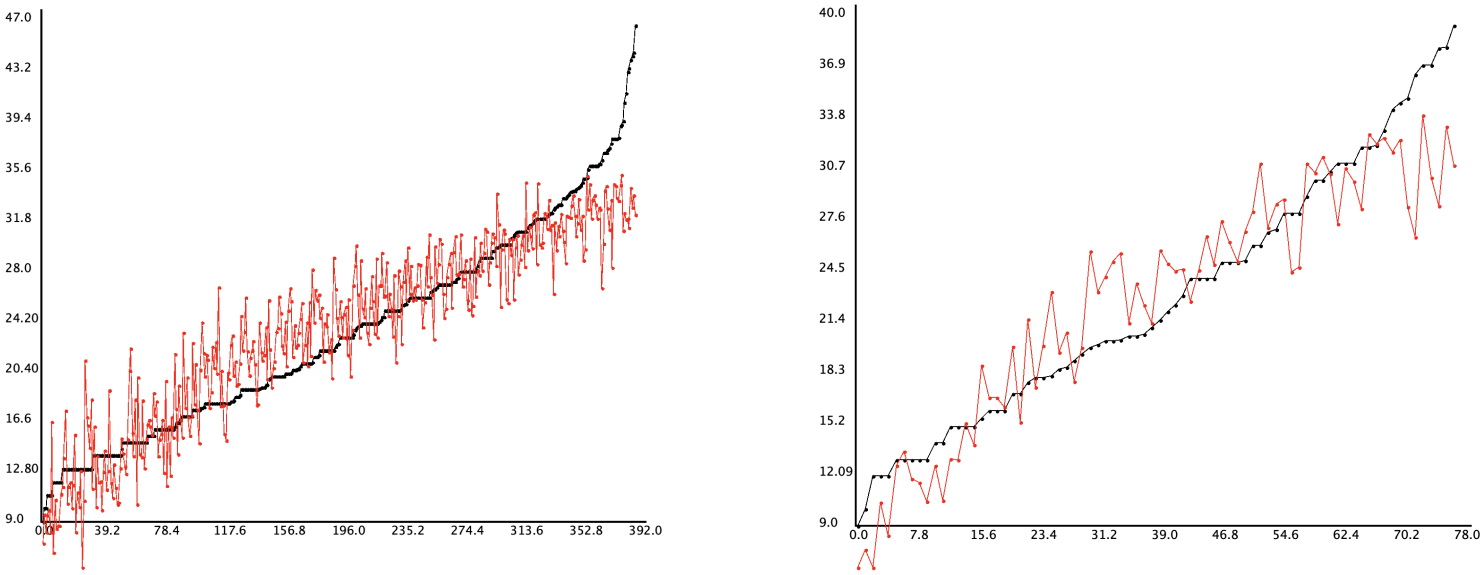


Figure 34: Scalation - Auto MPG Lasso
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

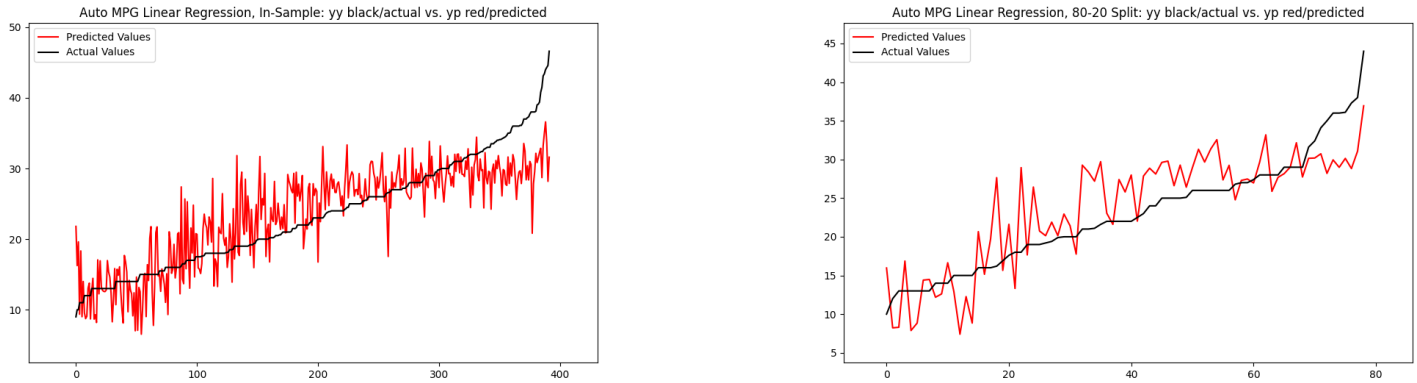


Figure 35: Statsmodels - Auto MPG Lasso
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

4.2 House Prices

Table 14: Lasso and Ridge Regression House Prices results

Metric	Ridge	Lasso
rSq	0.998517	0.981186
rSqBar	0.998507	0.981072
sst	6.42325e+13	6.42325e+13
sse	9.52491e+10	1.20848e+12
sde	9764.45	22644.0
mse0	9.52491e+07	1.20848e+09
rmse	9759.56	34763.2
mae	7740.46	28672.9
smape	9.75718	5.78522
m	1000.00	1000.00
dfr	7.00000	6.00000
df	992.000	993.000
fStat	95425.2	8631.07
aic	-10588.9	-12111.7
bic	-10549.7	-12077.3

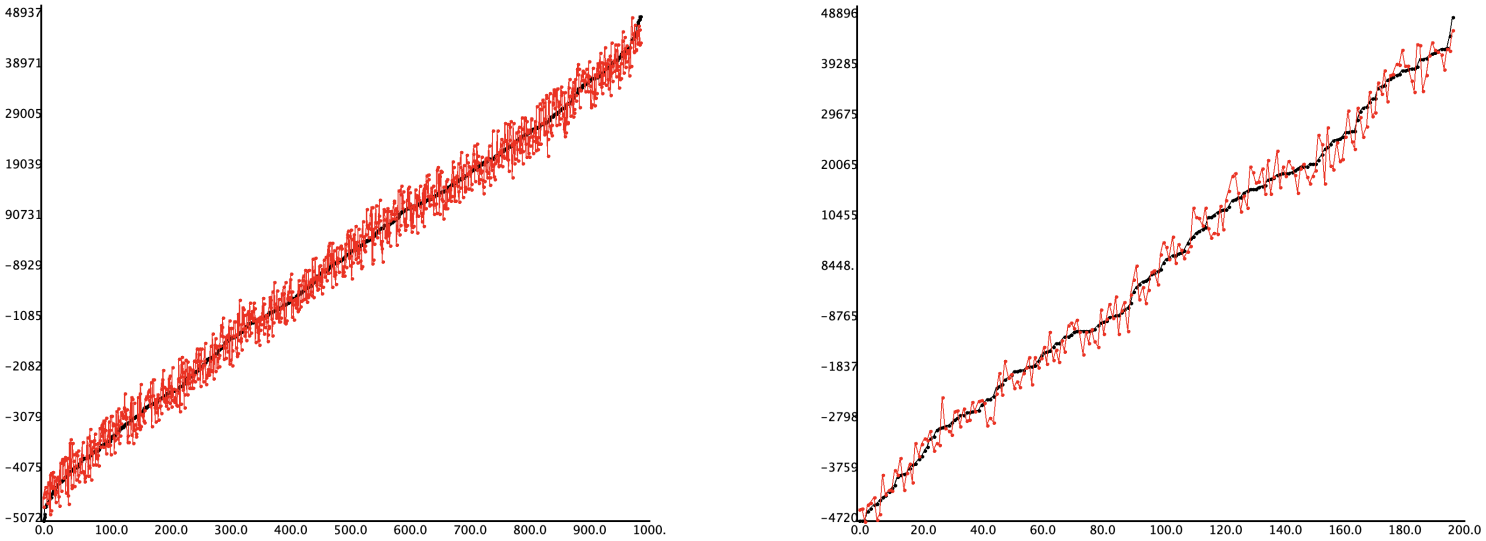
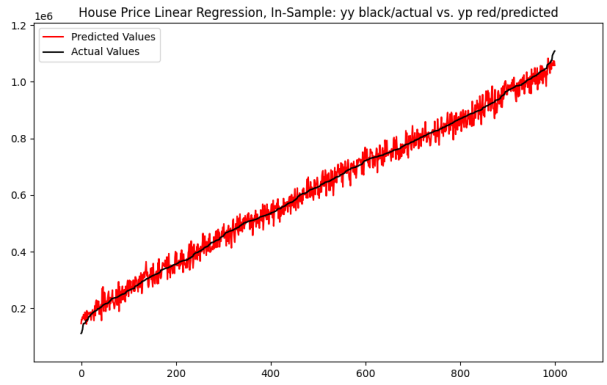


Figure 36: Scalation - House Price Ridge
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted



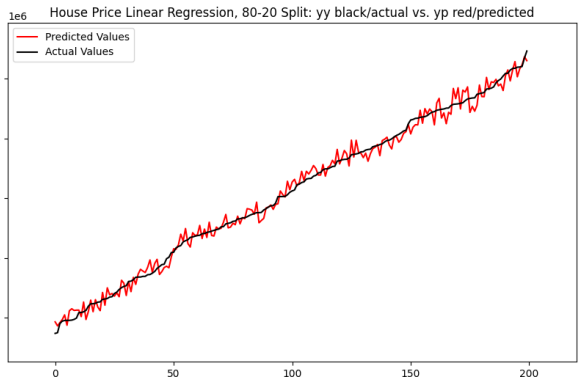
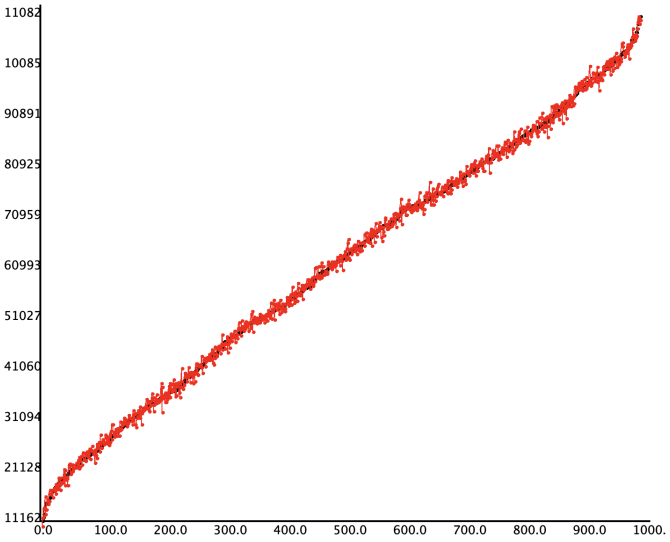


Figure 37: Statsmodels - House Price Ridge
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted



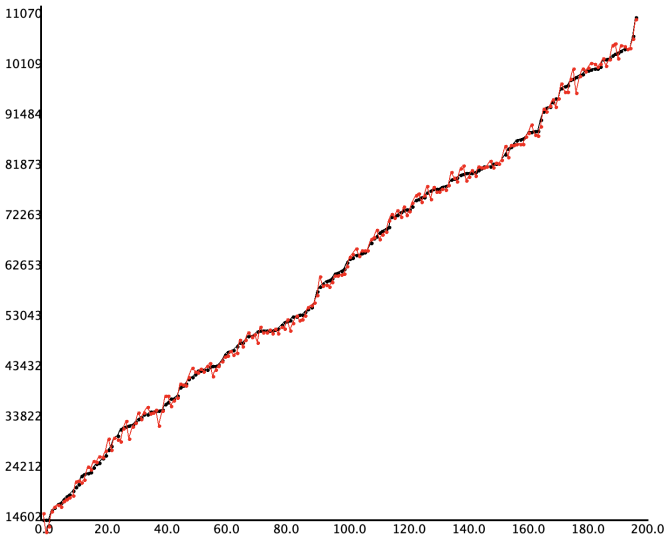
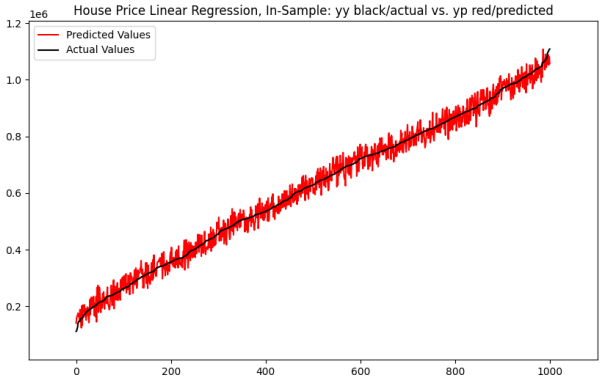


Figure 38: Scalation - House Price Lasso
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted



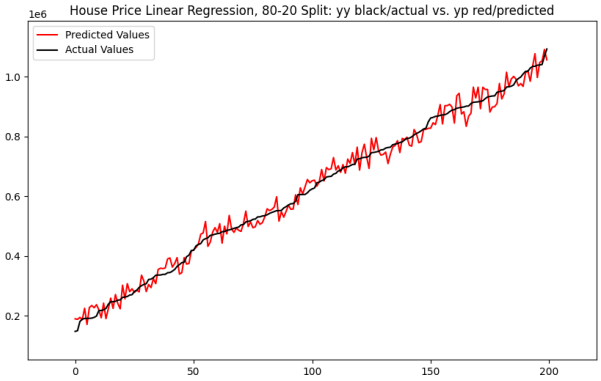


Figure 39: Statsmodels - House Price Lasso
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

4.3 Insurance Charges

Table 15: Lasso and Ridge Regression for Insurance Charges results

Metric	Ridge	Lasso
rSq	0.750913	0.723537
rSqBar	0.749414	0.722082
sst	1.96074e+11	1.96074e+11
sse	4.88396e+10	5.42073e+10
sde	6043.94	6358.54
mse0	3.65019e+07	4.05137e+07
rmse	6041.68	6365.04
mae	4171.66	4367.97
smape	68.9369	37.6686
m	1338.00	1338.00
dfr	8.00000	7.00000
df	1329.00	1330.00
fStat	500.810	497.252
aic	-13529.8	-13601.5
bic	-13483.0	-13559.9

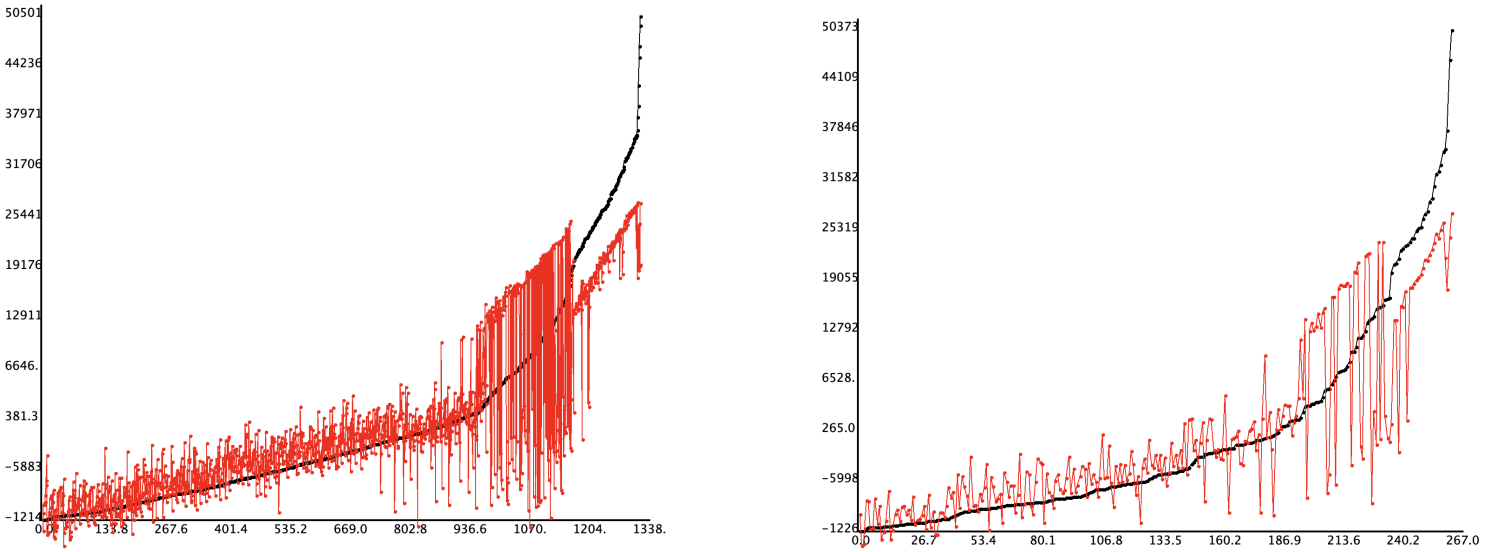


Figure 40: Scalation - Insurance Charges Ridge
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

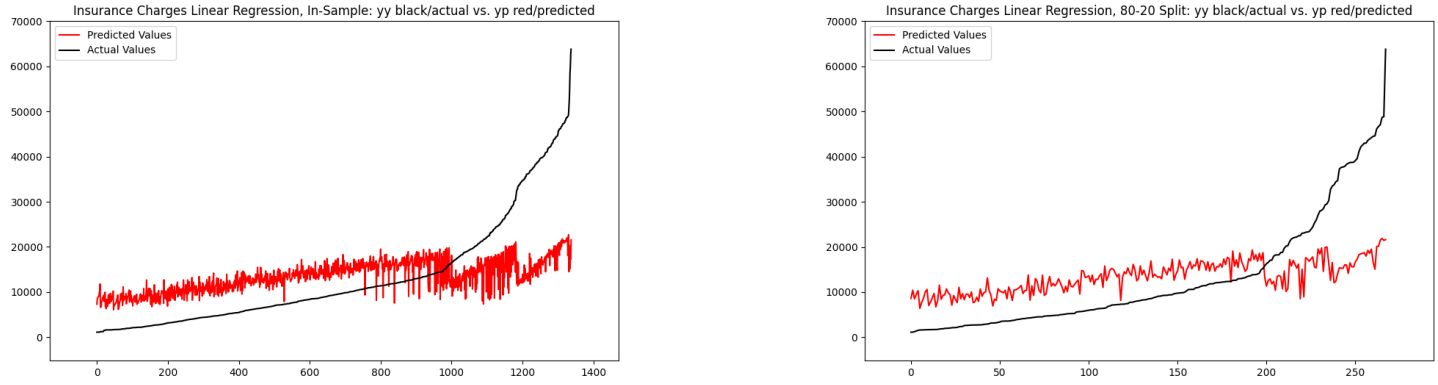


Figure 41: Statsmodels - Insurance Charges Ridge
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

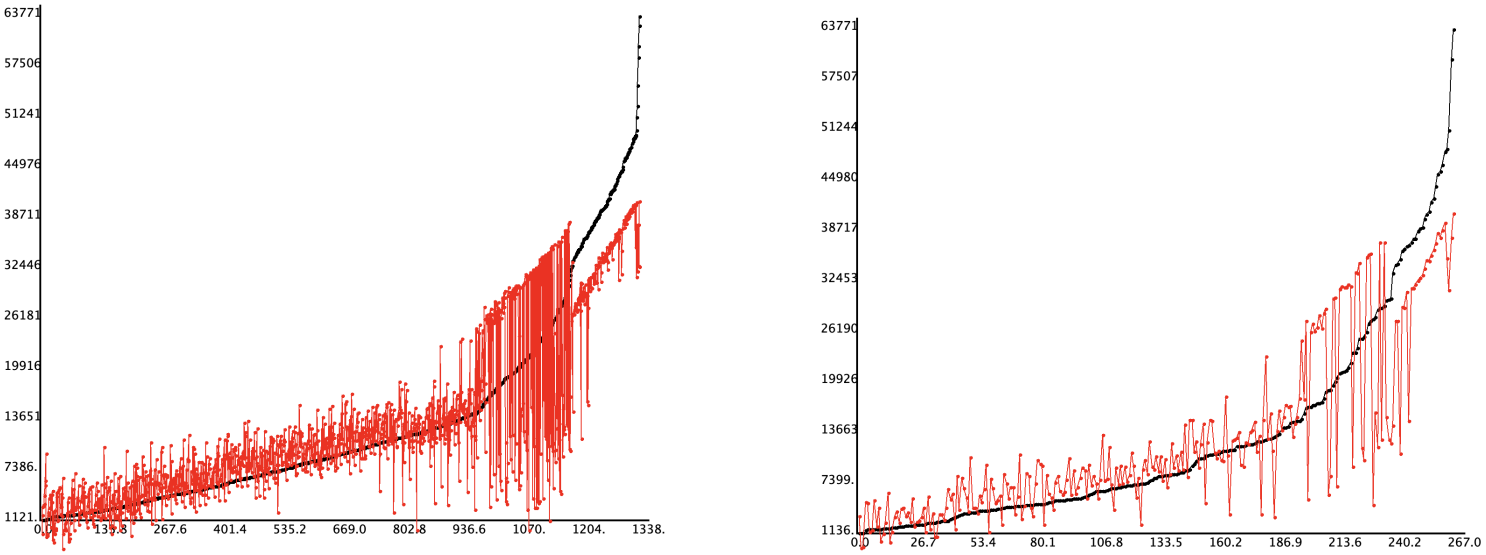


Figure 42: Scalation - Insurance Charges Lasso
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

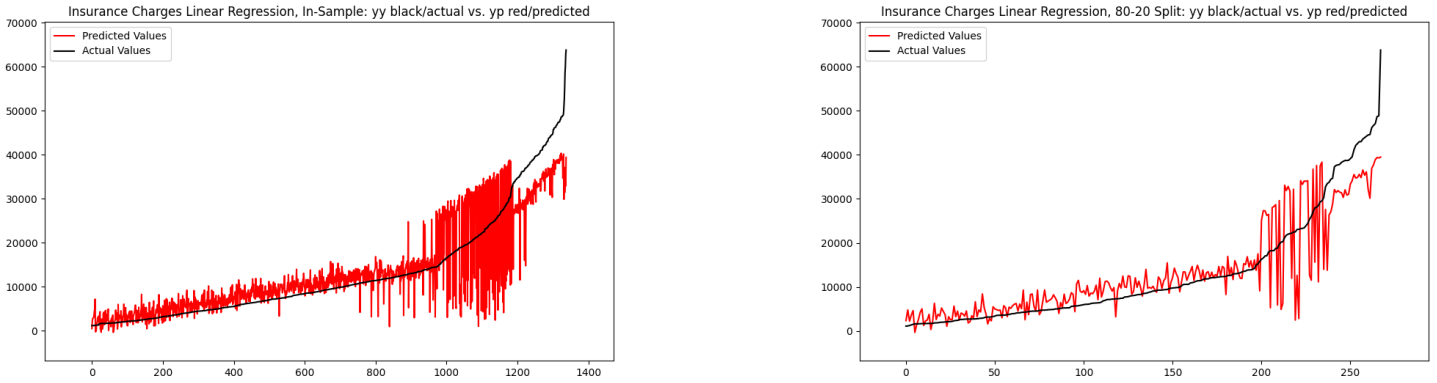


Figure 43: Statsmodels - Insurance Charges Lasso
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

5 Transformed Regression: Sqrt, Log1p, Box-Cox

In this section, we applied three different transformation techniques to the response variables (mpg, house price, and insurance charge) across the Auto MPG, Boston House Price, and Medical Cost datasets. The distributions for mpg, house price, and medical charge exhibit right-skewed patterns. Therefore, these transformations could significantly improved model accuracy. To measure and compare the quality of fit for the different transformation models, we extracted 15 metrics, including R^2, adjusted R^2, MSE, and MAE. Each model was evaluated using both in-sample validation and a validation set with an 80-20% split. Finally, we applied feature selection to identify the optimal set of features that best describe the response variable. For box-cox transformation we reqired a optimal lamda parameter which was 0.19, 0.85, and 0.04 for mpg, house price, and insurance cost respectively.

Transformed Regression using Scala

We used TranRegression function form the modeling to perfrom Sqrt, Log1p, Box-Cox on the Auto MPG, Boston House Price, and Medical Cost datasets. The function takes predictor variables, a response variable, feature names, a factorization method, and transformation and inverse transformation methods as input, and fits the model based on the specified transformation. An example of the code used to fit the square-root (sqrt) transformation model is provided below.

```
f = ("sqrt, sq, "sqrt")
val mod = new modeling.TranRegression (ox, y, ox_fname,
                                       modeling.Regression.hp,
                                       f._1, f._2, f._3)
```

Transformed Regression on the Auto MPG Dataset using Scala: In-Sample and Validation Results

Tables 16 and 17 present the quality of fit for the in-sample and validation (80-20% split) evaluations using the Auto MPG data. For the in-sample case, the models utilized all available data. The results indicate that the log1p and Box-Cox transformations outperform the sqrt transformation, which is further confirmed by the adjusted R^2 values. The Mean Square Error (MSE) for the sqrt, log1p, and Box-Cox transformations reached 10.02, 9.17, and 9.40, respectively, with the log1p transformation achieving the lowest error. Similarly, in the validation accuracy results, where 80% of the data was randomly used for model training and 20% for testing, the log1p transformation achieved the highest R^2 and adjusted R^2 values, as well as the lowest MSE.

Table 16: Auto MPG (In-Sample): Sqrt, Log1p, and Box–Cox($\lambda = 0.19$)

Metric	sqrt	log1p	box-cox(λ=0.19)
rSq	0.835138	0.849102	0.845366
rSqBar	0.832569	0.846751	0.842956
sst	23819.0	23819.0	23819.0
sse	3926.85	3594.23	3683.22
sde	3.16757	3.02514	3.06455
mse0	10.0175	9.16895	9.39596
rmse	3.16504	3.02803	3.06528
mae	2.34675	2.18422	2.22945
smape	10.2046	9.32001	9.54480
m	392.000	392.000	392.000
dfr	6.00000	6.00000	6.00000
df	385.000	385.000	385.000
fStat	325.048	361.067	350.793
aic	-993.873	-976.527	-981.320
bic	-966.074	-948.728	-953.521

Table 17: Auto MPG (Validation): Sqrt, Log1p, and Box–Cox($\lambda = 0.19$)

Metric	sqrt	log1p	box-cox(λ=0.19)
rSq	0.846480	0.852864	0.851819
rSqBar	0.844088	0.850571	0.849510
sst	4731.23	4731.23	4731.23
sse	726.337	696.133	701.080
sde	3.05724	2.97969	2.99552
mse0	9.31202	8.92478	8.98820
rmse	3.05156	2.98744	2.99803
mae	2.11846	1.98665	2.01582
smape	9.00068	8.28831	8.41880
m	78.0000	78.0000	78.0000
dfr	6.00000	6.00000	6.00000
df	385.000	385.000	385.000
fStat	353.804	371.939	368.862
aic	-183.700	-182.048	-182.322
bic	-167.203	-165.551	-165.825

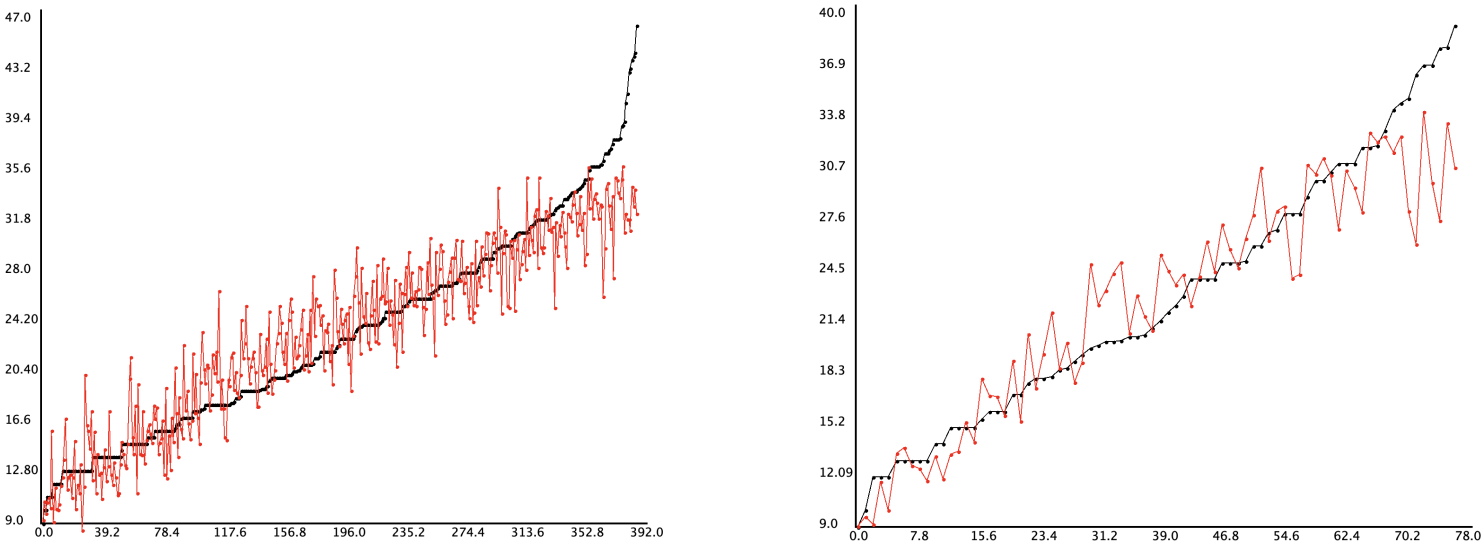
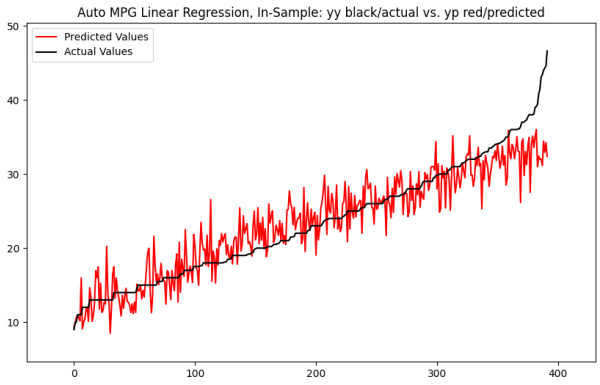


Figure 44: Scalation - Auto MPG Sqrt
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted



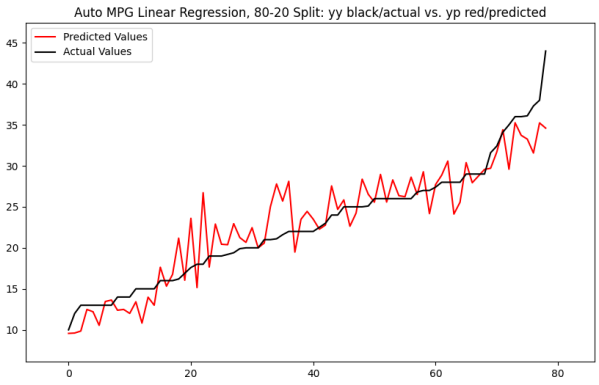
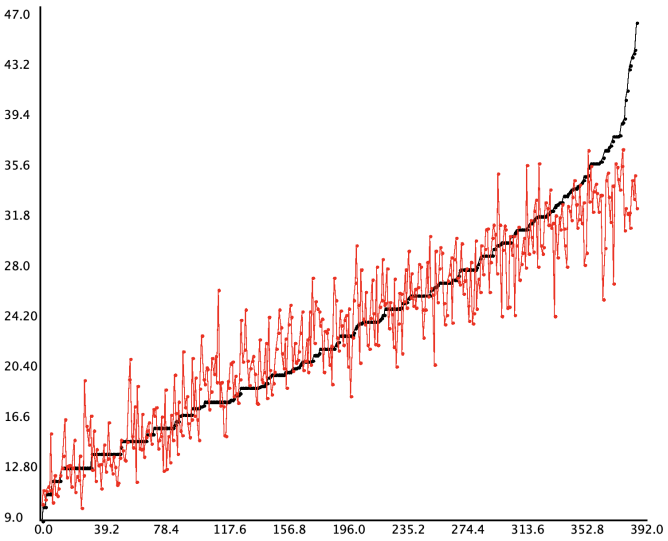


Figure 45: Statsmodels - Auto MPG Sqrt
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted



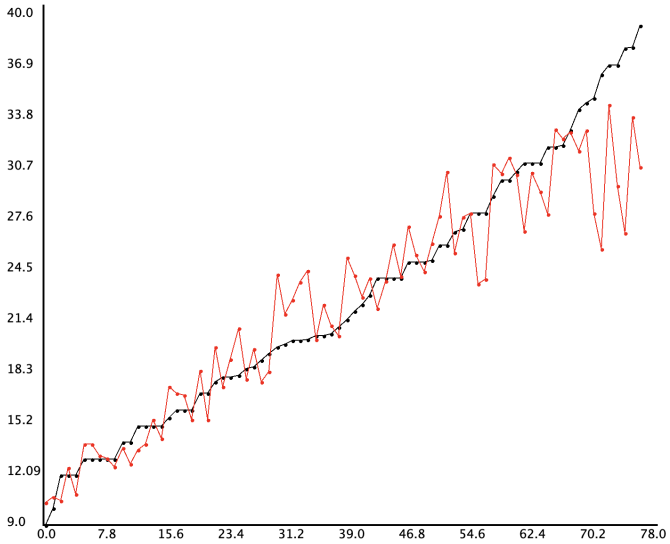
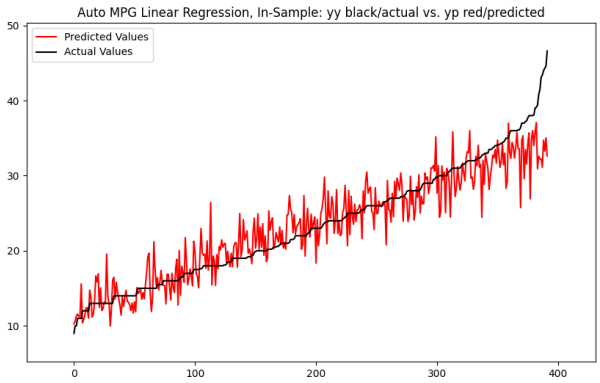


Figure 46: Scalation - Auto MPG log1p
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted



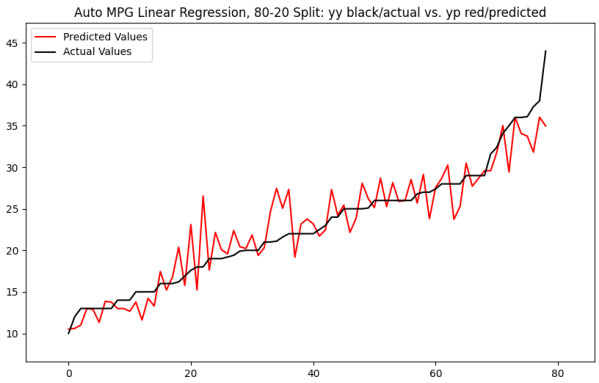


Figure 47: Statsmodels - Auto MPG log1p
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

Transformed Regression on the Boston House Price Dataset: In-Sample and Validation Results

For the house price prediction, the log1p transformed model performed significantly worse than the sqrt and Box-Cox transformations. The Box-Cox and sqrt models performed similarly, although Box-Cox achieved highest R^2 and adjusted R^2 of 0.998 in the in-sample evaluation. A similar trend was observed in the validation, where Box-Cox and sqrt outperformed the log1p transformation, with Box-Cox achieving the highest R^2 and adjusted R^2.

Table 18: House Price (In-Sample): Sqrt, Log1p, and Box–Cox($\lambda = 0.85$)

Metric	sqrt	log1p	box-cox(λ=0.85)
rSq	0.986164	0.922376	0.997549
rSqBar	0.986067	0.921828	0.997531
sst	6.42325e+13	6.42325e+13	6.42325e+13
sse	8.88711e+11	4.98598e+12	1.57458e+11
sde	29823.1	70573.3	12554.0
mse0	8.88711e+08	4.98598e+09	1.57458e+08
rmse	29811.3	70611.5	12548.2
mae	24296.5	52989.3	10078.9
smape	4.85788	9.21218	2.12365
m	1000.00	1000.00	1000.00
dfr	7.00000	7.00000	7.00000
df	992.000	992.000	992.000
fStat	10100.8	1683.94	57668.5
aic	-11705.6	-12567.9	-10840.3
bic	-11666.3	-12528.6	-10801.0

Table 19: House Price (Validation): Sqrt, Log1p, and Box–Cox($\lambda = 0.85$)

Metric	sqrt	log1p	box-cox(λ=0.85)
rSq	0.984017	0.906459	0.997398
rSqBar	0.983904	0.905799	0.997380
sst	1.33700e+13	1.33700e+13	1.33700e+13
sse	2.13694e+11	1.25064e+12	3.47839e+10
sde	32755.5	78812.3	13217.8
mse0	1.06847e+09	6.25321e+09	1.73920e+08
rmse	32687.5	79077.2	13187.9
mae	26140.1	58024.9	10655.9
smape	5.18595	9.88665	2.22264
m	200.000	200.000	200.000
dfr	7.00000	7.00000	7.00000
df	992.000	992.000	992.000
fStat	8724.77	1373.28	54329.4
aic	-2346.74	-2523.43	-2165.20
bic	-2320.35	-2497.04	-2138.81

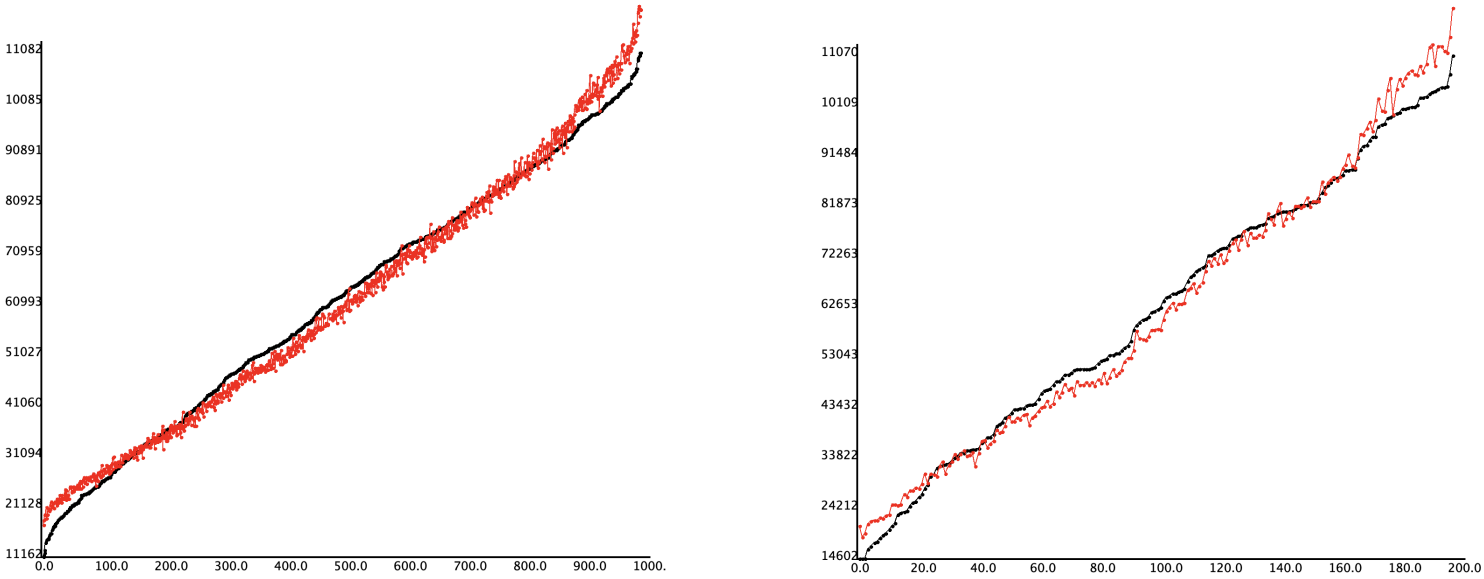
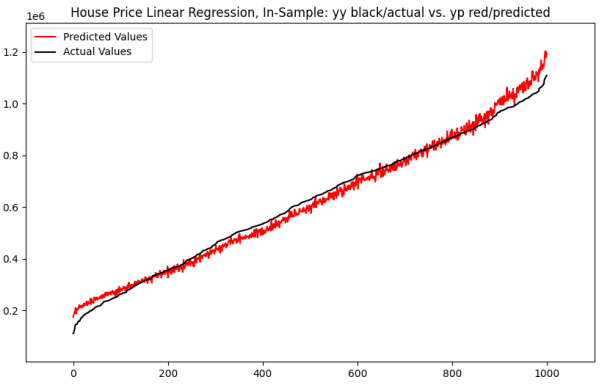


Figure 48: Scalation - House Price Sqrt
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted



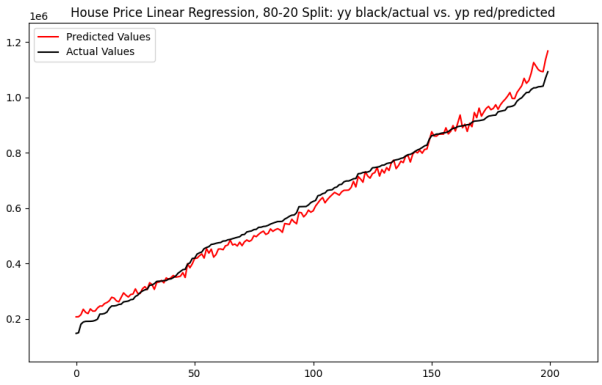
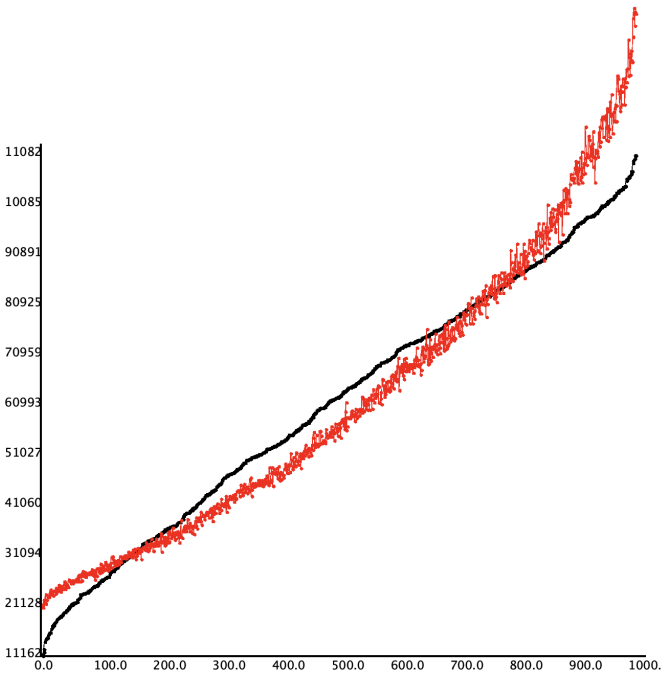


Figure 49: Statsmodels - House Price Sqrt
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted



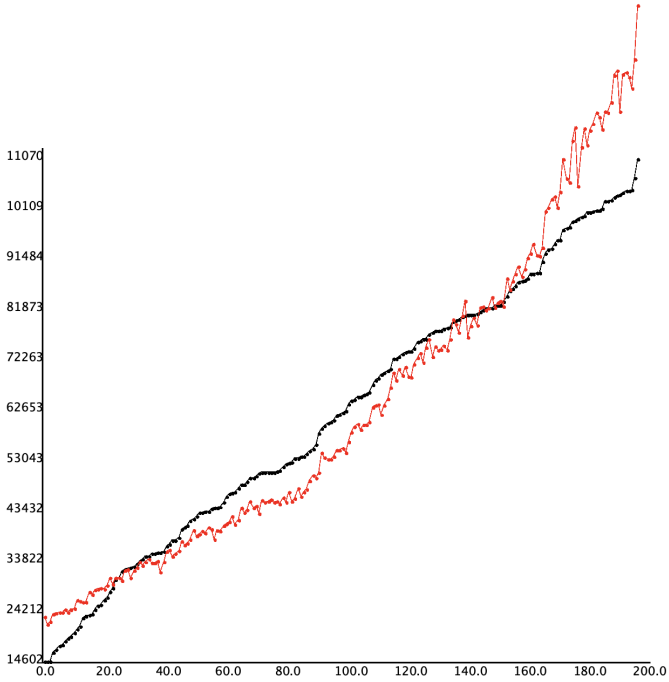
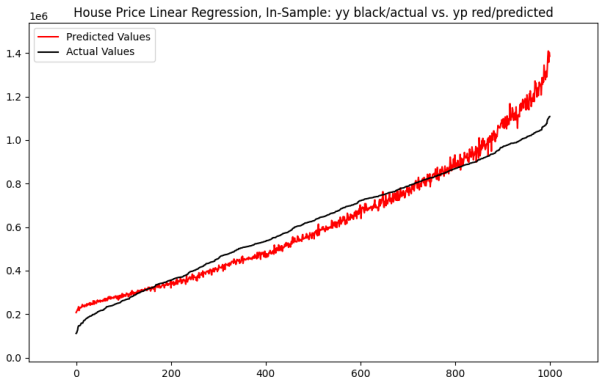


Figure 50: Scalation - House Price log1p
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted



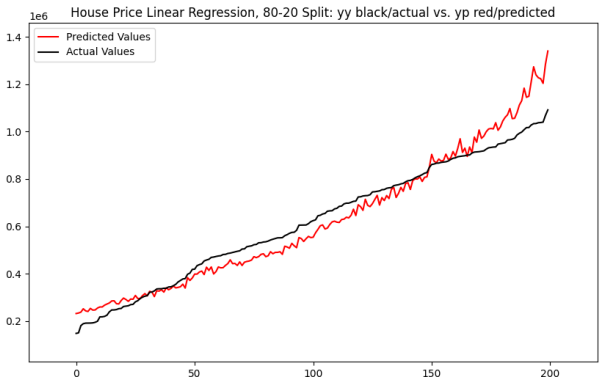


Figure 51: Statsmodels - House Price log1p
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

Transformed Regression on the Medical Cost Dataset: In-Sample and Validation Results

In the Medical Cost dataset, the sqrt transformation performed significantly better than the log1p and Box-Cox transformations. The sqrt transformation achieved an R^2 of 0.753 and an adjusted R^2 of 0.751 in the in-sample evaluation, compared to 0.730 and 0.729, respectively, for the validation set. The adjusted R^2 also shows a similar trend, where sqrt, log1p, and Box-Cox achieved 0.751, 0.520, and 0.574 for in-sample evaluation and 0.729, 0.569, and 0.607 for validation, respectively.

Table 20: Medical Cost (In-Sample): Sqrt, Log1p, and Box–Cox($\lambda = 0.04$)

Metric	sqrt	log1p	box-cox(λ=0.04)
rSq	0.752657	0.522782	0.576265
rSqBar	0.751168	0.519909	0.573715
sst	1.96074e+11	1.96074e+11	1.96074e+11
sse	4.84975e+10	9.35702e+10	8.30835e+10
sde	6001.71	8358.44	7870.39
mse0	3.62463e+07	6.99329e+07	6.20953e+07
rmse	6020.49	8362.59	7880.06
mae	3613.90	4219.51	4052.90
smape	27.6903	26.2889	26.0851
m	1338.00	1338.00	1338.00
dfr	8.00000	8.00000	8.00000
df	1329.00	1329.00	1329.00
fStat	505.514	181.986	225.924
aic	-13525.1	-13964.7	-13885.2
bic	-13478.3	-13917.9	-13838.4

Table 21: Medical Cost (Validation): Sqrt, Log1p, and Box–Cox($\lambda = 0.04$)

Metric	sqrt	log1p	box-cox(λ=0.04)
rSq	0.730154	0.571048	0.609124
rSqBar	0.728530	0.568466	0.606771
sst	4.06432e+10	4.06432e+10	4.06432e+10
sse	1.09674e+10	1.74340e+10	1.58865e+10
sde	6405.50	8088.13	7716.02
mse0	4.10764e+07	6.52957e+07	5.94998e+07
rmse	6409.08	8080.58	7713.61
mae	3839.79	4116.83	3998.83
smape	30.1151	27.3599	27.2984
m	267.000	267.000	267.000
dfr	8.00000	8.00000	8.00000
df	1329.00	1329.00	1329.00
fStat	449.505	221.156	258.882
aic	-2701.24	-2763.11	-2750.71
bic	-2668.95	-2730.83	-2718.42

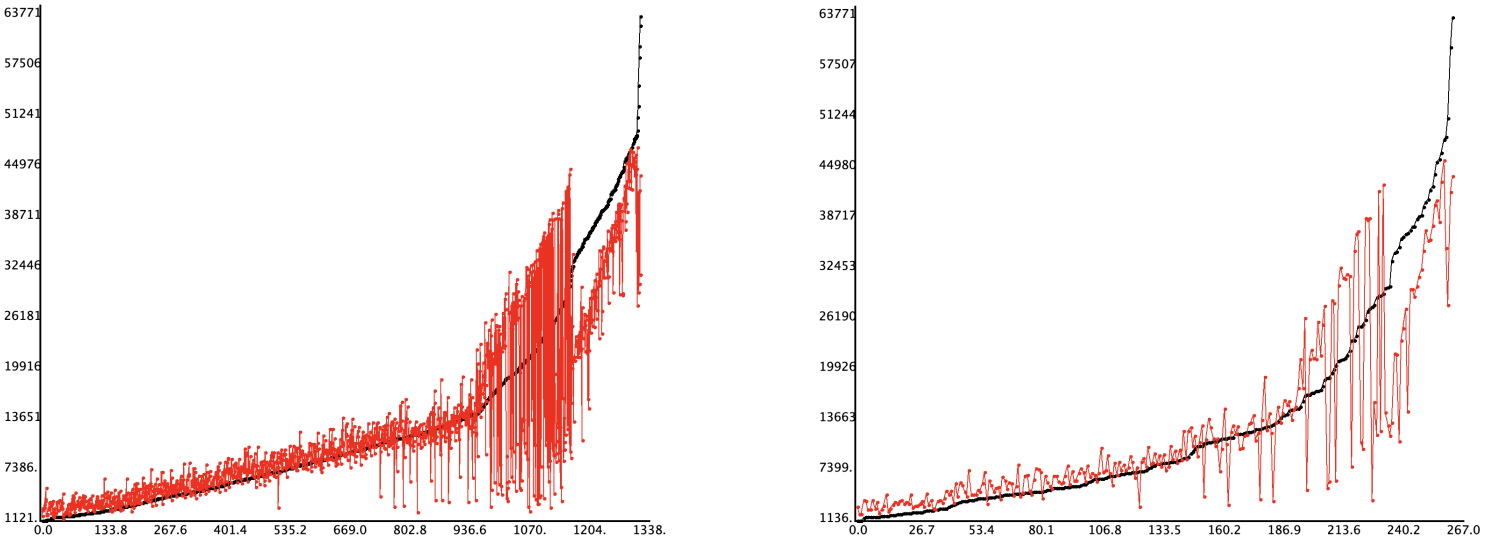
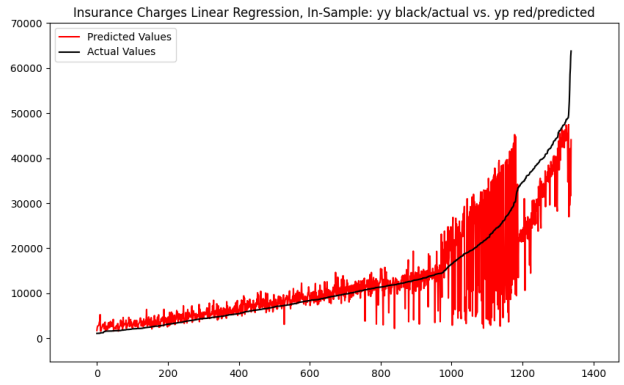


Figure 52: Scalation - Insurance Charges Sqrt
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted



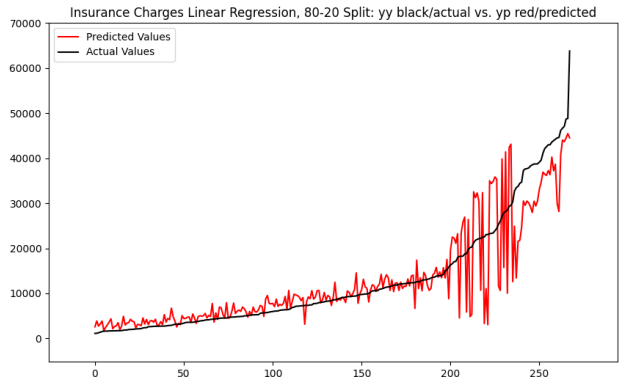
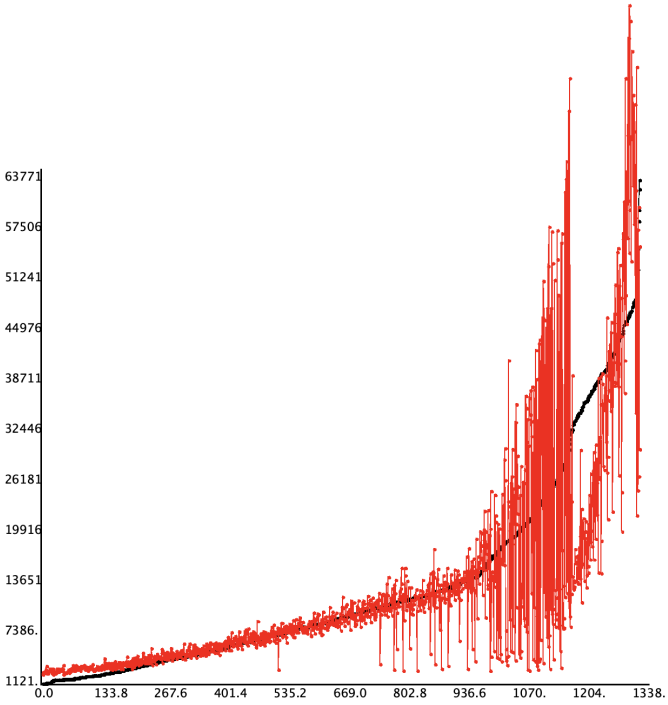


Figure 53: Statsmodels - Insurance Charges Sqrt
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted



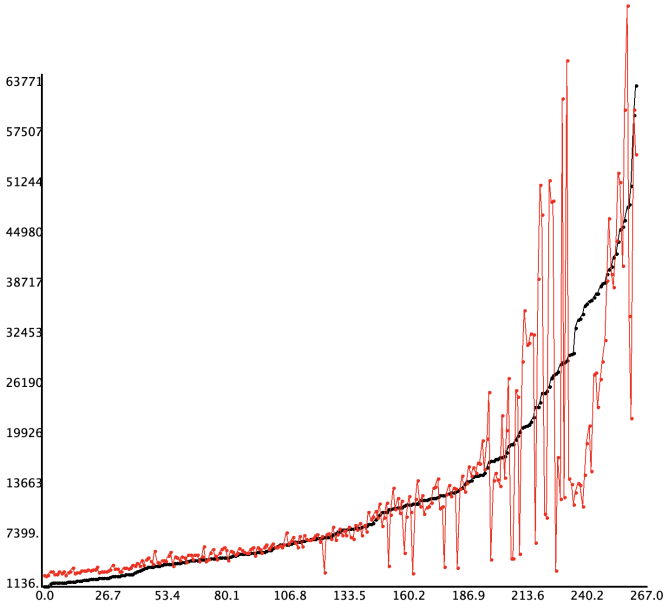
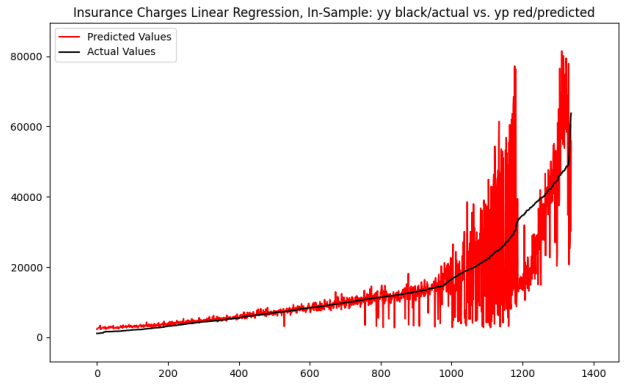


Figure 54: Scalation - Insurance Charges log1p
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted



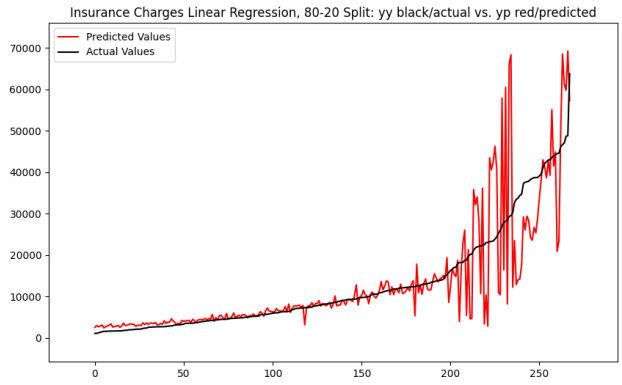


Figure 55: Statsmodels - Insurance Charges log1p
Left: In Sample Predictions
Right: 80-20 Out of Sample Predictions
yy black/actual vs. yp red/predicted

Transformed Regression using statsmodels

Here, we used Python statsmodels library to reproduce all results generated by the TranRegression package.

Transformed Regression with Sqrt, Log1p, and Box–Cox Transformations on the Auto MPG Dataset: In-Sample, Validation, Forward, and Backward Results

The following three tables present the results for the sqrt, log1p, and Box-Cox transformations, including in-sample evaluation, validation, and forward and backward feature selection. From these tables, we can see that, similar to the Scala results, the log1p and Box-Cox transformations perform similarly, with log1p being slightly better than Box-Cox. For the in-sample case, the sqrt, log1p, and Box-Cox transformations achieved R^2 values of 0.835, 0.849, and 0.845, respectively, with adjusted R^2 values of 0.833, 0.847, and 0.843. Similarly, for the validation case, the R^2 values were 0.822, 0.832, and 0.830, with adjusted R^2 values of 0.807, 0.818, and 0.816.

Table 22: Auto MPG Regression with Square-Root Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.835	0.822	0.823	0.822
rSqBar	0.833	0.807	0.818	0.814
sst	23818.993	5372.801	5372.801	5372.801
sse	3926.849	955.390	950.390	958.806
mse	4.017	6.094	10.030	9.137
rmse	3.165	3.478	3.468	3.484
mae	2.347	2.376	2.375	2.350
m	392.000	79.000	79.000	79.000
dfr	6.000	6.000	2.000	3.000
df	385.000	72.000	76.000	75.000
fStat	825.235	120.822	220.454	161.034

Table 23: Auto MPG Regression with Log1p Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.849	0.832	0.806	0.804
rSqBar	0.847	0.818	0.801	0.794
sst	23818.993	5372.801	5372.801	5372.801
sse	3594.229	900.453	1044.122	1052.182
mse	3.169	5.398	11.217	9.319
rmse	3.028	3.376	3.635	3.649
mae	2.184	2.221	2.710	2.713
m	392.000	79.000	79.000	79.000
dfr	6.000	6.000	2.000	4.000
df	385.000	72.000	76.000	74.000
fStat	1063.694	138.083	192.956	115.912

Table 24: Auto MPG Regression with Box-Cox Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.845	0.830	0.829	0.829
rSqBar	0.843	0.816	0.825	0.823
sst	23818.993	5372.801	5372.801	5372.801
sse	3683.218	914.477	917.735	916.540
mse	3.396	5.576	9.617	8.602
rmse	3.065	3.402	3.408	3.406
mae	2.229	2.262	2.356	2.234
m	392.000	79.000	79.000	79.000
dfr	6.000	6.000	2.000	3.000
df	385.000	72.000	76.000	75.000
fStat	988.221	133.268	231.627	172.688

Transformed Regression with Sqrt, Log1p, and Box–Cox Transformations on the Boston House Price Dataset: In-Sample, Validation, Forward, and Backward Results

For the house price prediction dataset, the ScalaTion and statsmodels summaries are consistent, as the Box–Cox transformation was identified as the best-performing model under both approaches.

Table 25: Boston House Price Regression with Square-Root Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.986	0.987	0.281	0.981
rSqBar	0.986	0.986	0.277	0.981
sst	64232463468052.547	11927062037792.469	11927062037792.469	11927062037792.469
sse	888710772890.930	155868297872.986	8579357295076.382	223122191452.058
mse	888710765.891	779341482.365	42896786474.382	1115610953.260
rmse	29811.252	27916.688	207115.394	33400.763
mae	24296.468	22632.355	177100.990	27022.336
m	1000.000	200.000	200.000	200.000
dfr	7.000	7.000	1.000	4.000
df	992.000	192.000	198.000	195.000
fStat	10182.286	2157.718	78.041	2622.765

Table 26: Boston House Price Regression with Log1p Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.922	0.934	-81398.531	0.877
rSqBar	0.922	0.931	-81809.639	0.875
sst	64232463468052.547	11927062037792.469	11927062037792.469	11927062037792.469
sse	4985983480190.929	789371506083.466	970857251871500928.000	1461933128041.826
mse	4985983473.191	3946857523.417	4854286259357504.000	7309665636.209
rmse	70611.497	62824.020	69672708.139	85496.583
mae	52989.293	48059.016	20766229.670	61057.028
m	1000.000	200.000	200.000	200.000
dfr	7.000	7.000	1.000	4.000
df	992.000	192.000	198.000	195.000
fStat	1697.515	403.131	-199.998	357.921

Table 27: Boston House Price Regression with Box-Cox Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.998	0.997	0.924	0.990
rSqBar	0.998	0.997	0.924	0.989
sst	64232463468052.531	11927062037792.469	11927062037792.469	11927062037792.469
sse	157457720227.919	32215998366.736	903559871079.215	124824257960.295
mse	157457713.228	161079984.834	4517799354.396	624121285.801
rmse	12548.216	12691.729	67214.577	24982.420
mae	10078.898	10289.124	56408.842	20247.645
m	1000.000	200.000	200.000	200.000
dfr	7.000	7.000	1.000	4.000
df	992.000	192.000	198.000	195.000
fStat	58133.527	10549.192	2440.016	4727.542

Transformed Regression with Sqrt, Log1p, and Box–Cox Transformations on the Medical Cost Dataset: In-Sample, Validation, Forward, and Backward Results

Finally, for the medical cost dataset, the sqrt transformation explained the insurance costs better than the other methods, despite the relatively lower R^2 values. For the in-sample case, the R^2 values were 0.753, 0.523, and 0.576 for the sqrt, log1p, and Box-Cox transformations, respectively, with adjusted R^2 values of 0.751, 0.520, and 0.574. These results were verified by the validation set, where the sqrt transformation achieved an R^2 of 0.706 and an adjusted R^2 of 0.697, outperforming the log1p (R^2 of 0.505, Adjusted R^2 of 0.490) and Box-Cox ((R^2 of 0.546, Adjusted R^2 of 0.532) models.

Table 28: Medical Cost Regression with Square-Root Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.753	0.706	0.702	0.704
rSqBar	0.751	0.697	0.699	0.699
sst	196074221568.367	35797001122.000	35797001122.000	35797001122.000
sse	48497528289.829	10509483431.295	10649651869.890	10582381967.672
mse	36246276.223	39214482.415	39737503.977	39486494.879
rmse	6020.489	6262.147	6303.769	6283.828
mae	3613.896	3611.073	3687.410	3606.738
m	1338.000	268.000	268.000	268.000
dfr	8.000	8.000	3.000	5.000
df	1329.000	259.000	264.000	262.000
fStat	508.937	80.606	210.946	127.713

Table 29: Medical Cost Regression with Log1p Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.523	0.505	-415.056	-53.303
rSqBar	0.520	0.490	-419.784	-54.340
sst	196074221568.367	35797001122.000	35797001122.000	35797001122.000
sse	93570168847.798	17725546380.471	14893568669580.479	1943894999500.171
mse	69932853.620	66140090.435	55573017420.808	7253339545.374
rmse	8362.587	8132.656	235739.300	85166.540
mae	4219.512	4231.425	66672.183	25327.152
m	1338.000	268.000	268.000	268.000
dfr	8.000	8.000	3.000	5.000
df	1329.000	259.000	264.000	262.000
fStat	183.219	34.154	-89.119	-52.613

Table 30: Medical Cost Regression with Box-Cox Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.576	0.546	-61.221	-10.125
rSqBar	0.574	0.532	-61.928	-10.337
sst	196074221568.367	35797001122.000	35797001122.000	35797001122.000
sse	83083467025.754	16248889297.412	2227326392657.103	398225509782.019
mse	62095258.835	60630175.946	8310919372.586	1485916076.276
rmse	7880.055	7786.539	91164.244	38547.582
mae	4052.895	4083.531	31806.787	14527.915
m	1338.000	268.000	268.000	268.000
dfr	8.000	8.000	3.000	5.000
df	1329.000	259.000	264.000	262.000
fStat	227.454	40.302	-87.898	-48.782

6 Comparing Different Models on the Same Data Set

In summary, the linear regression model for the Auto MPG dataset achieved an R^2 and adjusted R^2 of 0.81, with consistent results of 0.80 across cross-validation in both Scala and Python. For the House Price dataset, the model performed exceptionally well, with R^2 values nearly reaching 1.0 across in-sample, validation, and cross-validation sets. The Insurance dataset showed lower performance, with an R^2 of 0.75 and an adjusted R^2 of 0.72.

When applying regularized regression, on the Auto MPG dataset, Lasso performed slightly better than Ridge when using scalation, but in statsmodels Lasso took a significant dip, resulting in Ridge performing better. They both performed comparably on the House Price dataset. On the Insurance dataset, using scalation they again performed comparably, but in statsmodels Ridge performed far worse.

Regarding data transformations, the Log1p method was most effective for the Auto MPG dataset, whereas the Box-Cox transformation yielded the best results for House Price prediction. For the Insurance dataset, the Sqrt transformation achieved the highest performance, reaching an R^2 and adjusted R^2 of 0.75 for in-sample data and 0.73 for validation. These findings remained consistent when implemented through both ScalaTion and statsmodels.

6.1 Auto MPG

Table 31: Scalation - Auto MPG In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.809255	0.776580	0.809163	0.835138	0.849102
rSqBar	0.806283	0.772507	0.806189	0.832569	0.846751
sst	23819.0	23819.0	23819.0	23819.0	23819.0
sse	4543.35	5321.63	4545.54	3926.85	3594.23
sde	3.40878	3.68883	3.40888	3.16757	3.02514
mse0	11.5902	13.5756	11.5958	10.0175	9.16895
rmse	3.40443	3.68451	3.40526	3.16504	3.02803
mae	2.61826	2.79509	2.61703	2.34675	2.18422
smape	12.0589	65.4181	11.9861	10.2046	9.32001
m	392.000	392.000	392.000	392.000	392.000
dfr	6.00000	7.00000	6.00000	6.00000	6.00000
df	385.000	384.000	385.000	385.000	385.000
fStat	272.234	190.677	272.072	325.048	361.067
aic	-1022.45	-1051.45	-1022.55	-993.873	-976.527
bic	-994.656	-1019.68	-994.750	-966.074	-948.728

Table 32: Scalation - Auto MPG Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.822842	0.797491	0.822903	0.846480	0.852864
rSqBar	0.820081	0.793799	0.820143	0.844088	0.850571
sst	4731.23	4731.23	4731.23	4731.23	4731.23
sse	838.174	958.118	837.889	726.337	696.133
sde	3.29026	3.51709	3.28969	3.05724	2.97969
mse0	10.7458	12.2836	10.7422	9.31202	8.92478
rmse	3.27808	3.50479	3.27752	3.05156	2.98744
mae	2.48735	2.62052	2.48643	2.11846	1.98665
smape	11.8858	61.9272	11.8808	9.00068	8.28831
m	78.0000	78.0000	78.0000	78.0000	78.0000
dfr	6.00000	7.00000	6.00000	6.00000	6.00000
df	385.000	384.000	385.000	385.000	385.000
fStat	298.034	216.030	298.158	353.804	371.939
aic	-189.284	-192.500	-189.271	-183.700	-182.048
bic	-172.787	-173.646	-172.774	-167.203	-165.551

Table 33: Statsmodels - Auto MPG In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	Log1p
rSq	0.8093	0.8037	0.6416	0.8477	0.8725
rSqBar	0.8063	0.8012	0.6369	0.8453	0.8705
sst	23818.9935	23818.9935	23818.9935	252.1610	41.1728
sse	4543.3470	4675.2421	8537.0473	38.4071	5.2483
sde	3.4352	3.4802	4.7028	0.3158	0.1168
mse0	11.8009	11.9266	21.7782	0.0998	0.0136
rmse	3.4352	3.4535	4.6667	0.3158	0.1168
mae	2.6183	2.6269	3.6246	2.3467	2.1842
smape	12.0589	12.0433	16.2905	10.2046	9.3200
m	392.0000	392.0000	392.0000	392.0000	392.0000
dfr	6.0000	6.0000	6.0000	6.0000	6.0000
df	385.0000	386.0000	386.0000	385.0000	385.0000
fStat	272.2341	263.4262	115.1614	357.1178	439.2179
aic	2086.9095	983.6796	1219.7162	215.8245	-564.3874
bic	2114.7083	1007.5071	1243.5438	243.6234	-536.5886

Table 34: Statsmodels - Auto MPG Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	Log1p
rSq	0.7942	0.7854	0.5925	0.8373	0.8604
rSqBar	0.7801	0.7707	0.5646	0.8261	0.8508
sst	4032.2061	4032.2061	4032.2061	4032.2061	4032.2061
sse	829.6873	865.2680	1643.1024	656.0919	563.0196
sde	3.3713	3.4428	4.7443	2.9979	2.7772
mse0	10.5024	10.9528	20.7988	8.3050	7.1268
rmse	3.2407	3.3095	4.5606	2.8818	2.6696
mae	2.5039	2.5877	3.7495	2.1509	1.9533
smape	12.3880	12.5974	17.6830	9.8913	8.7248
m	79.0000	79.0000	79.0000	79.0000	79.0000
dfr	6.0000	6.0000	6.0000	6.0000	6.0000
df	73.0000	73.0000	73.0000	73.0000	73.0000
fStat	46.9622	44.5308	17.6906	62.6072	74.9680
aic	197.7765	201.0937	251.7566	179.2314	167.1455
bic	211.9932	215.3104	265.9733	193.4481	181.3621

6.2 House Prices

Table 35: Scalation - Housing Prices In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.998516	0.987398	0.998516	0.986152	0.922307
rSqBar	0.998507	0.987309	0.998507	0.986068	0.921838
sst	6.42325e+13	6.42325e+13	6.42325e+13	6.42325e+13	6.42325e+13
sse	9.53030e+10	8.09438e+11	9.53030e+10	8.89504e+11	4.99039e+12
sde	9767.21	28463.5	9767.21	29836.4	70604.5
mse0	9.53030e+07	8.09438e+08	9.53030e+07	8.89504e+08	4.99039e+09
rmse	9762.32	28450.6	9762.32	29824.5	70642.7
mae	7747.66	24147.7	7747.66	24309.3	53070.5
smape	1.57791	23.5576	1.57791	4.86164	9.22319
m	1000.00	1000.00	1000.00	1000.00	1000.00
dfr	6.00000	7.00000	6.00000	6.00000	6.00000
df	993.000	992.000	993.000	993.000	993.000
fStat	111378	11103.9	111378	11785.5	1964.69
aic	-10591.2	-11658.9	-10591.2	-11708.0	-12570.3
bic	-10556.9	-11619.6	-10556.9	-11673.7	-12536.0

Table 36: Scalation - Housing Prices Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.998649	0.989193	0.998649	0.984018	0.906860
rSqBar	0.998641	0.989117	0.998641	0.983921	0.906297
sst	1.33700e+13	1.33700e+13	1.33700e+13	1.33700e+13	1.33700e+13
sse	1.80638e+10	1.44485e+11	1.80638e+10	2.13678e+11	1.24528e+12
sde	9501.56	26880.4	9501.56	32757.9	78673.0
mse0	9.03190e+07	7.22424e+08	9.03189e+07	1.06839e+09	6.22641e+09
rmse	9503.63	26877.9	9503.63	32686.3	78907.6
mae	7484.48	22419.4	7484.48	26186.8	58148.8
smape	1.60847	19.6544	1.60847	5.19788	9.91629
m	200.000	200.000	200.000	200.000	200.000
dfr	6.00000	7.00000	6.00000	6.00000	6.00000
df	993.000	992.000	993.000	993.000	993.000
fStat	122330	12971.9	122330	10189.9	1611.39
aic	-2101.67	-2307.60	-2101.67	-2348.73	-2525.00
bic	-2078.59	-2281.21	-2078.59	-2325.64	-2501.91

Table 37: Statsmodels - House Price In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	Log1p
rSq	0.9985	0.9906	0.9875	0.9855	0.9415
rSqBar	0.9985	0.9905	0.9875	0.9854	0.9411
sst	64232463468052.5469	64232463468052.5469	64232463468052.5469	29445342.0661	241.7492
sse	95249090298.3967	604344645977.1542	800169699033.4265	425598.9981	14.1461
sde	9798.8381	24669.9185	28386.7992	20.7131	0.1194
mse0	96017228.1234	604344645.9772	800169699.0334	429.0312	0.0143
rmse	9798.8381	24583.4222	28287.2710	20.7131	0.1194
mae	7740.4301	20205.2499	24002.5173	24296.4685	52989.2927
smape	1.5774	4.1035	4.9163	4.8579	9.2122
m	1000.0000	1000.0000	1000.0000	1000.0000	1000.0000
dfr	7.0000	7.0000	7.0000	7.0000	7.0000
df	992.0000	993.0000	993.0000	992.0000	992.0000
fStat	95425.1583	14935.3572	11245.5195	9662.8803	2280.1137
aic	21225.8831	20233.6552	20514.3344	8907.3747	-1404.4424
bic	21265.1451	20268.0095	20548.6887	8946.6367	-1365.1804

Table 38: Statsmodels - House Price Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	Log1p
rSq	0.9984	0.9913	0.9884	0.9875	0.9304
rSqBar	0.9984	0.9910	0.9880	0.9871	0.9283
sst	12891771417242.6445	12891771417242.6445	12891771417242.6445	12891771417242.6445	12891771417242.6445
sse	20286959701.1265	112341107192.6465	149469076702.9745	160935342144.9142	897031269775.1045
sde	10252.5012	24126.2984	27828.9629	28876.6666	68174.9984
mse0	101434798.5056	561705535.9632	747345383.5149	804676710.7246	4485156348.8755
rmse	10071.4844	23700.3278	27337.6185	28366.8241	66971.3099
mae	8174.5836	19629.7981	23012.0161	23553.0274	51521.6937
smape	1.6620	3.9691	4.6112	4.7594	9.0567
m	200.0000	200.0000	200.0000	200.0000	200.0000
dfr	7.0000	7.0000	7.0000	7.0000	7.0000
df	193.0000	193.0000	193.0000	193.0000	193.0000
fStat	17493.2676	3136.4045	2350.4760	2181.0457	368.6740
aic	3700.9854	4043.2977	4100.4076	4115.1902	4458.8078
bic	3724.0736	4066.3859	4123.4958	4138.2785	4481.8961

6.3 Insurance Charges

Table 39: Scalation - Insurance Charges In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.750157	0.749962	0.750157	0.752656	0.527431
rSqBar	0.748842	0.748457	0.748842	0.751354	0.524943
sst	1.96074e+11	1.96074e+11	1.96074e+11	1.96074e+11	1.96074e+11
sse	4.89878e+10	4.90260e+10	4.89878e+10	4.84977e+10	9.26587e+10
sde	6053.11	6055.42	6053.11	6001.45	8316.89
mse0	3.66127e+07	3.66413e+07	3.66127e+07	3.62464e+07	6.92516e+07
rmse	6050.84	6053.20	6050.84	6020.50	8321.76
mae	4179.54	4150.46	4179.54	3623.20	4215.04
smape	37.9722	67.8472	37.9722	27.9341	26.5034
m	1338.00	1338.00	1338.00	1338.00	1338.00
dfr	7.00000	8.00000	7.00000	7.00000	7.00000
df	1330.00	1329.00	1330.00	1330.00	1330.00
fStat	570.477	498.274	570.477	578.161	212.057
aic	-13533.8	-13532.3	-13533.8	-13527.1	-13960.2
bic	-13492.2	-13485.5	-13492.2	-13485.5	-13918.6

Table 40: Scalation - Insurance Charges Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.720005	0.719714	0.720005	0.730926	0.583107
rSqBar	0.718531	0.718027	0.718531	0.729510	0.580913
sst	4.06432e+10	4.06432e+10	4.06432e+10	4.06432e+10	4.06432e+10
sse	1.13799e+10	1.13917e+10	1.13799e+10	1.09360e+10	1.69438e+10
sde	6540.46	6543.82	6540.46	6396.98	7972.86
mse0	4.26213e+07	4.26655e+07	4.26213e+07	4.09588e+07	6.34601e+07
rmse	6528.50	6531.89	6528.50	6399.91	7966.18
mae	4430.66	4429.38	4430.66	3868.65	4087.97
smape	40.0602	62.9178	40.0602	30.5895	27.5909
m	267.000	267.000	267.000	267.000	267.000
dfr	7.00000	8.00000	7.00000	7.00000	7.00000
df	1330.00	1329.00	1330.00	1330.00	1330.00
fStat	488.584	426.574	488.584	516.126	265.753
aic	-2708.17	-2706.31	-2708.17	-2702.86	-2761.31
bic	-2679.47	-2674.02	-2679.47	-2674.16	-2732.61

Table 41: Statsmodels - Insurance Charges In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	Log1p
rSq	0.7509	0.2718	0.7469	0.7795	0.7680
rSqBar	0.7494	0.2680	0.7456	0.7782	0.7666
sst	196074221568.3671	196074221568.3671	196074221568.3671	3051091.5131	1130.1100
sse	48839532843.9219	142780871945.3538	49623455821.6671	672635.9707	262.2315
sde	6062.1023	10361.1794	6108.2624	22.4972	0.4442
mse0	36749084.1564	106712161.3941	37087784.6201	506.1219	0.1973
rmse	6062.1023	10330.1579	6089.9741	22.4972	0.4442
mae	4170.8869	8134.7714	4115.6884	3613.8958	4219.5115
smape	37.8059	66.1576	34.1997	27.6903	26.2889
m	1338.0000	1338.0000	1338.0000	1338.0000	1338.0000
dfr	8.0000	8.0000	8.0000	8.0000	8.0000
df	1329.0000	1330.0000	1330.0000	1329.0000	1329.0000
fStat	500.8107	62.0533	490.6438	587.4216	549.8054
aic	27113.5058	24749.7939	23335.7320	12137.4774	1634.5362
bic	27160.2962	24791.3854	23377.3235	12184.2678	1681.3266

Table 42: Statsmodels - Insurance Charges Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	Log1p
rSq	0.7836	0.2951	0.7797	0.7926	0.6067
rSqBar	0.7778	0.2761	0.7738	0.7871	0.5961
sst	41606660039.7953	41606660039.7953	41606660039.7953	41606660039.7953	41606660039.7953
sse	9003973448.1649	29328394351.2134	9164643207.2381	8627220450.5265	16363971889.4475
sde	5884.7827	10620.8058	5937.0555	5760.3487	7933.3696
mse0	33596915.8514	109434307.2806	34196429.8778	32191121.0841	61059596.6024
rmse	5796.2847	10461.0854	5847.7714	5673.7220	7814.0640
mae	4181.1945	8328.6585	4069.5309	3556.9640	3888.4432
smape	40.0220	69.4557	35.5690	29.1245	25.7136
m	268.0000	268.0000	268.0000	268.0000	268.0000
dfr	8.0000	8.0000	8.0000	8.0000	8.0000
df	260.0000	260.0000	260.0000	260.0000	260.0000
fStat	117.6800	13.6061	115.0471	124.2384	50.1338
aic	4660.4252	4976.9038	4665.1653	4648.9699	4820.5327
bic	4689.1531	5005.6317	4693.8932	4677.6978	4849.2606