Sai Mualagan
10/12/24

U.S. Automotive Price Prediction: Multiple Linear Regression

I. Introduction

Geely Auto, a Chinese automobile company, aims to expand into the U.S. market by establishing local manufacturing operations to compete with American and European automakers. Geely Auto has requested our assistance to analyze the factors that influence car prices in the U.S., which may differ from those in the Chinese market. Understanding these factors will help the company make informed decisions about product design and market positioning.

The primary objectives of this project are to:

1. Identify the key variables that significantly impact car pricing in the American market.
2. Assess how well these variables explain price variation.

To achieve these goals, we have sourced a dataset from Kaggle (https://www.kaggle.com/datasets/hellbuoy/car-price-prediction) that includes a wide variety of cars sold in the US market. The analysis of this dataset will enable the company to model car prices based on the identified factors. This pricing model will guide the design and strategy decisions for Geely's entry into the American automobile market. Our team has decided to implement three different modeling strategies for predicting car price.

First, we will employ Multiple Linear Regression (MLR). This method will allow us to quantify the relationships between car prices and various independent variables, such as engine size, horsepower, fuel efficiency, and manufacturing country. By using MLR, we can identify the strength and significance of each predictor, giving us a clear understanding of how these factors influence pricing in the U.S. market. However, MLR can face challenges, particularly with multicollinearity. This occurs when predictor variables are highly correlated, leading to unreliable coefficient estimates. To address this limitation, we intend to supplement MLR with Lasso and Ridge regression techniques.

Lasso regression employs an L1 penalty, which penalizes the absolute size of the regression coefficients. This allows for less significant coefficients to be shrunk to zero, effectively performing feature selection. This not only simplifies the model but also highlights the most impactful variables, giving Geely clearer insights into the features that drive pricing.

$$\hat{\beta}^{lasso} = argmin \sum_{i=1}^{n} (y_i - \sum_j x_{ij}^T \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$\widehat{\beta^{lasso}} = argmin\,||y - X\beta||_2^2 + \lambda||\beta||_1$$

Last, we will utilize Ridge regression, which employs L2 regularization to address multicollinearity without eliminating features. Ridge regression stabilizes the coefficient estimates and ensures that we capture the effects of all variables, even those that may be correlated. This approach will provide a complementary perspective to the Lasso results, allowing us to understand the dynamics of car pricing more thoroughly.

$$\widehat{\beta^{ridge}} = argmin \sum_{i=1}^{n} (y_i - x_i^T\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$\widehat{\beta^{ridge}} = argmin\,||y - X\beta||_2^2 + \lambda||\beta||_2^2$$

II. Description of Data:

From our original dataset we started with 25 predictors, 1 response variable, and 206 observations.

These included:

1. car _id (UUID): Unique numeric identifier for each car
2. Symboling (int): Risk factor value from insurance companies
3. Carname (String): A description of the car, having the model and the brand (e.g. Chevrolet monte carlo)
4. Fueltype (String): Value representing type of fuel the car takes (Gas or Fuel)
5. Aspiration (String): Represents how the engine takes in air (Std or Turbo)
6. Doornumber (String): How many doors does the car have (Four or Two)
7. Carbody (String): Type of Car (e.g. Hatchback)
8. DriveWheel (String): Which wheels are transmitting the force to allow the car to move (Fwd, Rwd, or 4wd)
9. EngineLocation (string): Where in the car the engine is located (front or rear)
10. Wheelbase (float): Distance between the front and rear tires
11. CarLength (float): Length of the car
12. CarWidth (float): Width of the car
13. CarHeight (float): Height of the car
14. CurbWeight (float): Total weight of the car
15. EngineType (String): Type of engine the car is using (e.g. Ohc)
16. CylinderNumber (String): Number of cylinders the car is using (two, three, four, five, six, eight, twelve)
17. Enginesize (float): Length dimension of the engine
18. FuelSystem (String): Fuel Delivery System that the car uses (e.g. 2bbl)
19. BoreRatio (float): Ratio of the diameter of the cylinder's bore to length of the stroke
20. Stroke (float: Length of the Stroke in the engine
21. CompressionRatio (float): The ratio between the maximum volume of a cylinder and the minimum volume of the cylinder
22. Horsepower (int): Maximum power production of the car engine
23. PeakRPM (int): Maximum rotations per minute the car can handle at peak
24. CityMPG (int): How many miles per gallon the car will get driving through a city
25. HighwayMPG (int): How many miles per gallon the car will get driving down a freeway
26. Price (int): Price in dollars (RESPONSE VARIABLE)

III. Column Transformations, EDA, Outliers, and Feature Selection

---

1. Column Transformations

- **Carname:**
  - Created a dictionary to map car brands to the country of origin.
  - Applied one-hot encoding to convert the countries into categorical variables.
- **Categorical Variables:** Transformed the following columns into categorical variables:
  - Fuel
  - Aspiration
  - Door Number
  - Engine Location
  - Drive Wheel
  - Car Body
  - Engine Type
  - Cylinder Number
  - Fuel System
  - Country of Origin

2. Exploratory Data Analysis:

We first began our analysis investigating the response variable, price. From *Figure 1* below, it can be seen the distribution of car prices is right skewed with prices ranging from roughly $5,000 to over $45,000. We then created side by side boxplots to visualize the distribution of prices between different groups in categorical variables. From our analysis in *Figure2*, the largest variations in median prices between groups can be seen with manufacturing countries, engine location, and cylinder number. We next made a distribution plot (*Figure 3*) which plots scatterplots of all combinations of variables. This helped visualize the multicollinearity between variables which would help guide our feature selection. This in combination with our correlation heat map (*Figure 4*) identified many features like curbweight which had correlations with several variables like volume, citympg, highwaympg, etc.

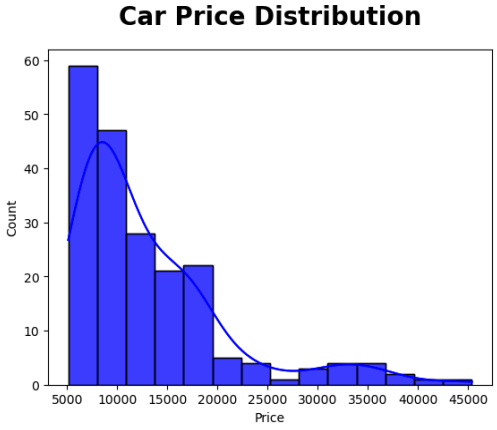


Figure 1: Distribution of Car Prices

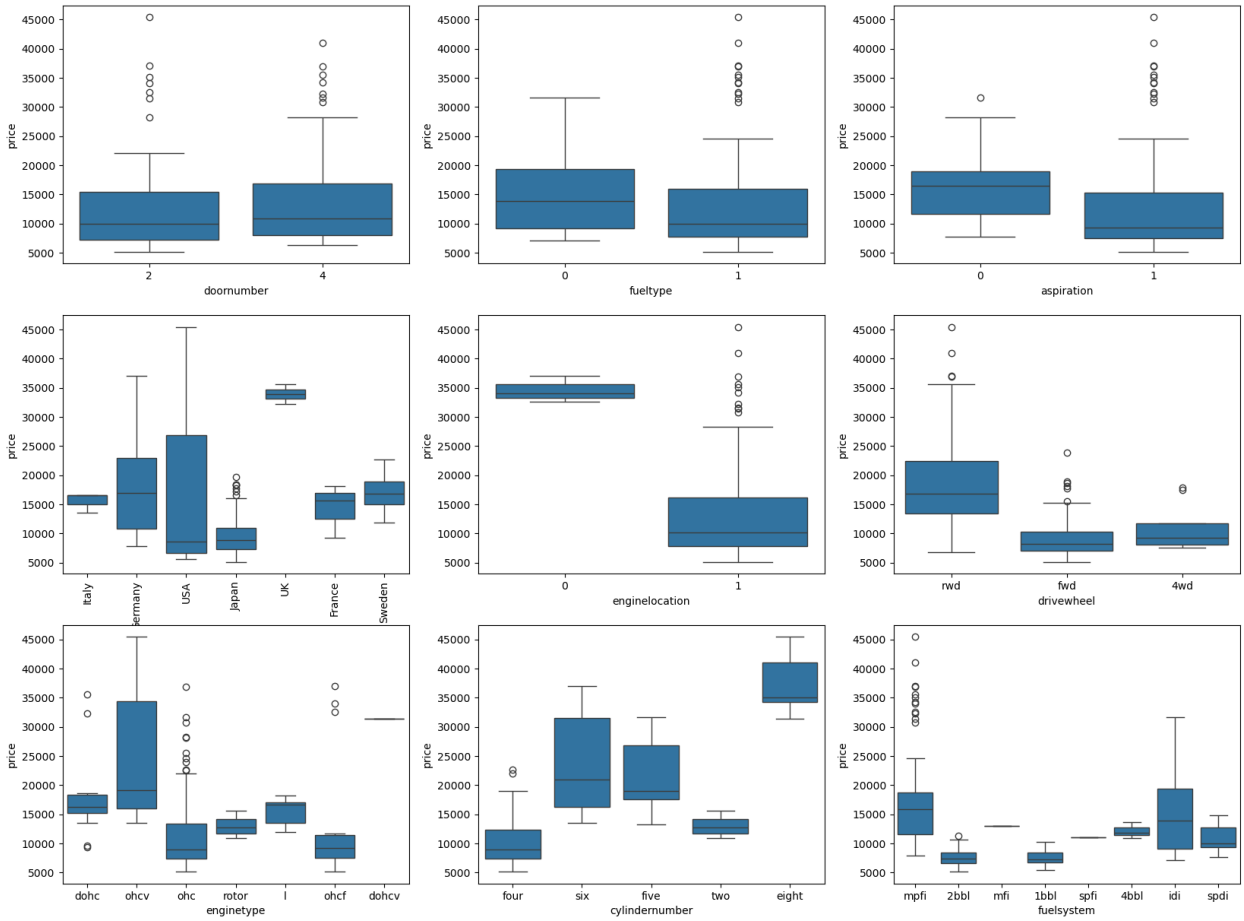


Figure 2: Categorical Variable Price Distributions

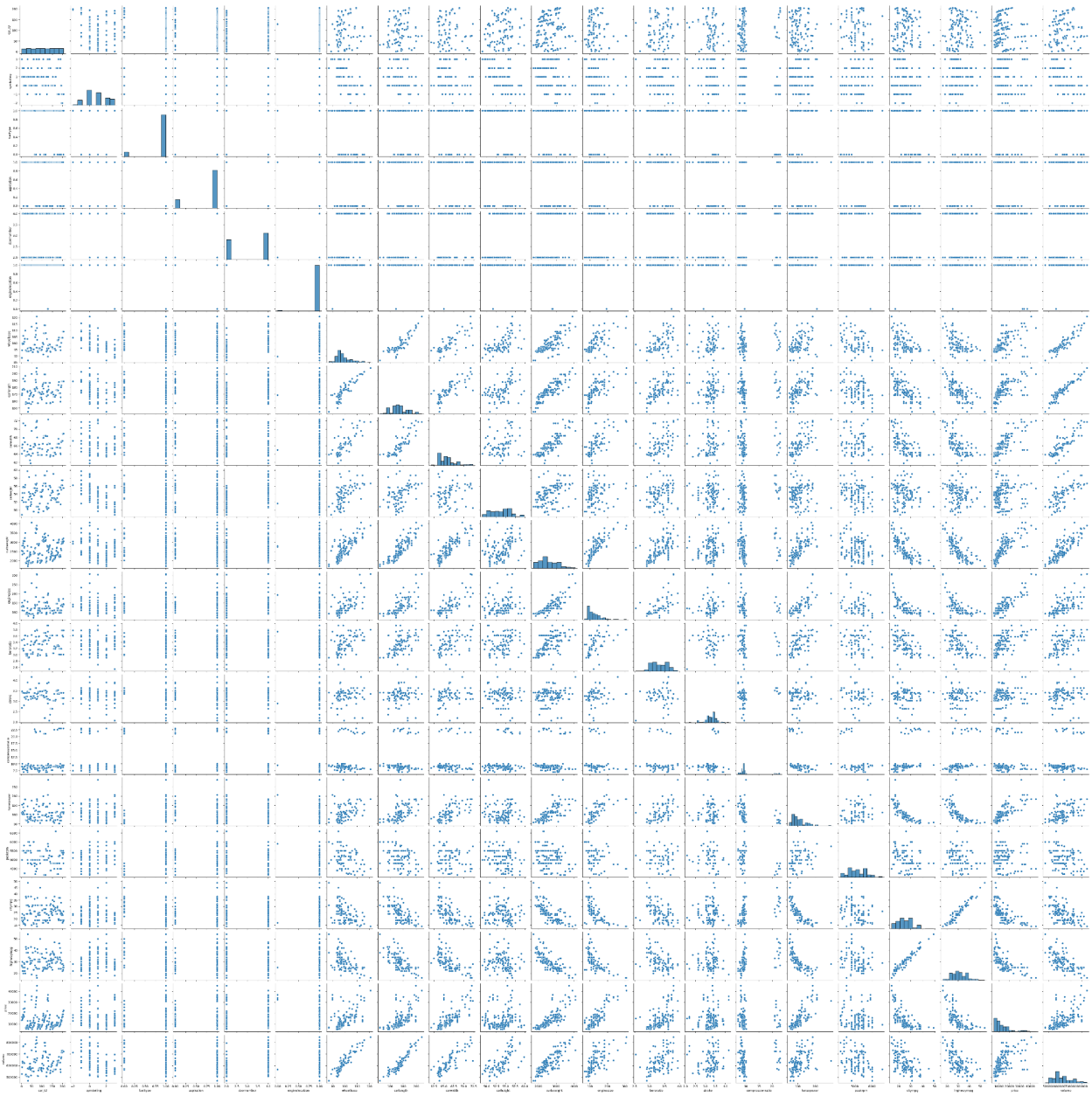


Figure 3: Distribution Plot of All Variables

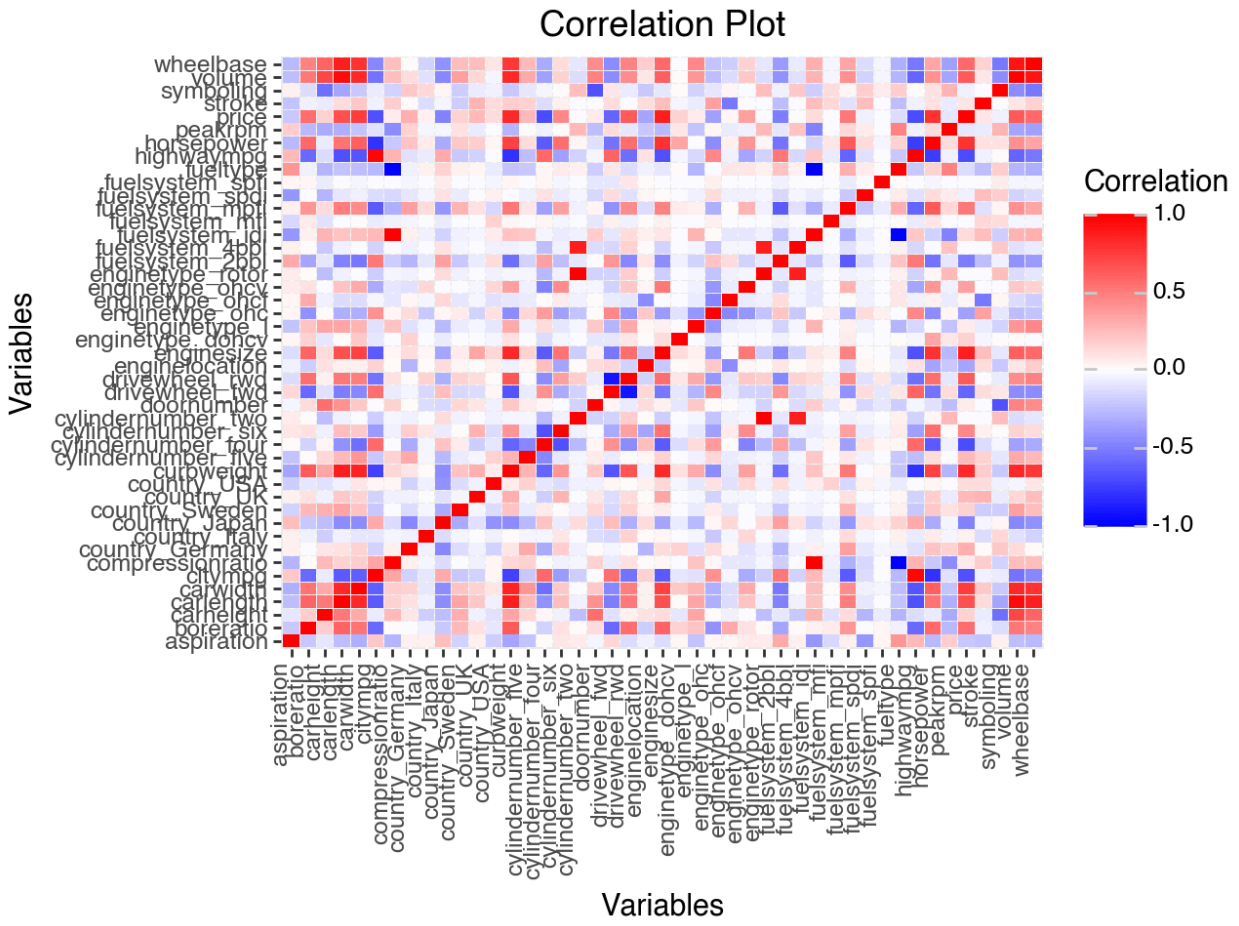


Figure 4: Correlation Heatmap of Variables

3. Handling Outliers

- **Cylinder Number:**
  - There was only one observation for both three-cylinder and twelve-cylinder cars. Also, when fitting our initial full model, the p-value for the twelve-cylinder was 0.79.
  - To prevent overfitting, both three-cylinder and twelve-cylinder values were dropped.
- **Observation #17:**
  - Later during the modeling stage, observation #17 was identified as being an extremely influential point (*Figure 5*) having a large Cook's Distance. This prompted us to remove this observation, because it made a significant change in our models.

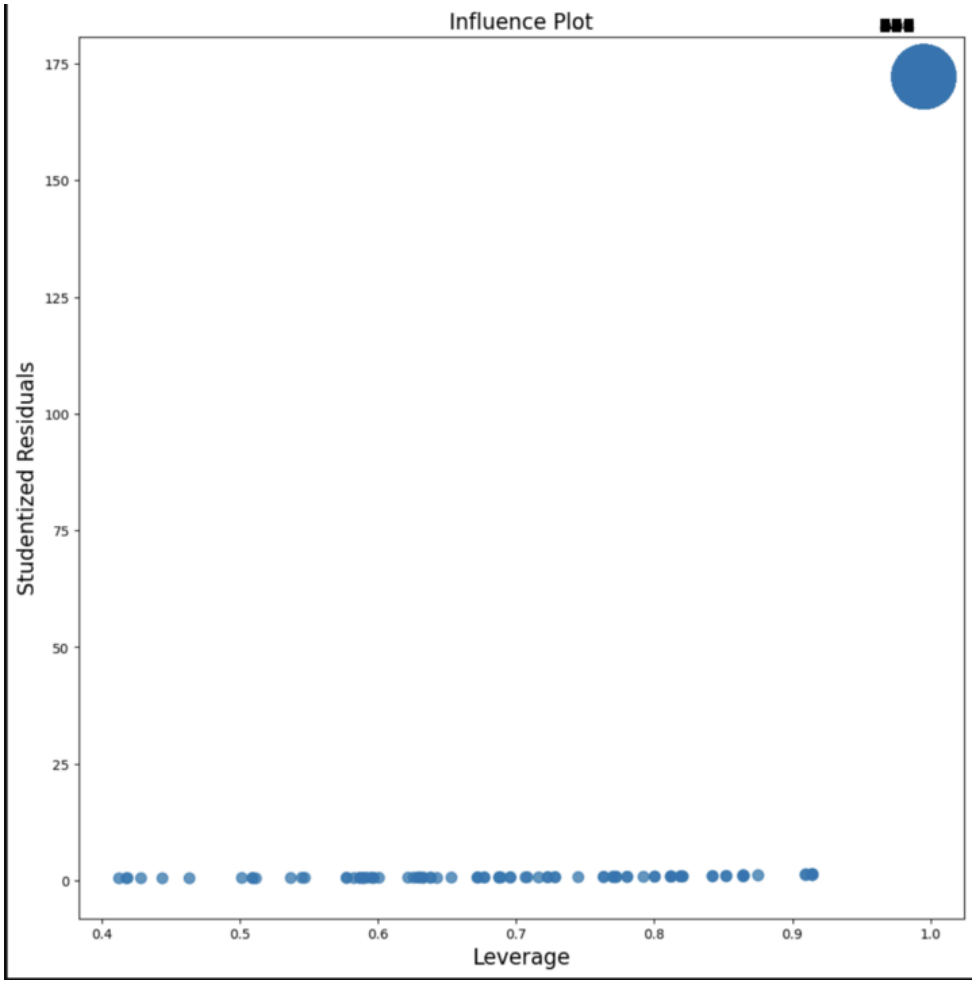


Figure 5: Influence Plot

4. Feature Selection:

Initially, we began by removing variables which had high correlations identified from our exploratory data analysis and external research. We then calculated the VIF's of our remaining variables to check for additional multicollinearity. The resulting multicollinearity was reduced, with only Cylinder Number 4 showing multicollinearity (VIF > 10). Based on research, Cylinder Number 4 was kept, as four-cylinder engines are commonly associated with fuel-efficient cars, a strong predictor of initial car price. Our decision making is outlined below:

VIF's Pre-Predictor Selection VIFs:

	Feature	VIF
0	const	2798.775900
1	symboling	3.245495
2	aspiration	2.489702
3	doornumber	2.584708
4	enginelocation	2.396418
5	enginesize	15.280761
6	peakrpm	3.512499
7	citympg	7.462751
8	volume	8.580830
9	drivewheel_fwd	8.323307
10	drivewheel_rwd	11.014111
11	enginetype_dohcv	1.819026
12	enginetype_l	10.391195
13	enginetype_ohc	9.970917
14	enginetype_ohcf	5.177896
15	enginetype_ohcv	3.478961
16	enginetype_rotor	inf
17	cylindernumber_five	11.958592
18	cylindernumber_four	44.240487
19	cylindernumber_six	15.158593
20	cylindernumber_two	inf
21	fuelsystem_2bbl	6.866177
22	fuelsystem_4bbl	4.558890
23	fuelsystem_idi	6.527923
24	fuelsystem_mfi	1.363551
25	fuelsystem_mpfi	10.326849
26	fuelsystem_spdi	2.704534
27	fuelsystem_spfi	1.220753
28	country_Germany	15.693334
29	country_Italy	3.288521
30	country_Japan	28.742387
31	country_Sweden	10.003894
32	country_UK	2.839905
33	country_USA	14.724981

Columns Exhibiting Multicollinearity	**Action taken based on external research/intuition/VIF analysis**
Car Width, Car Height, Car Length	Combined into a new column representing **Volume**.
City MPG, Highway MPG	Kept **City MPG** over Highway MPG, as city was more significant
Wheelbase, Fuel Type, Curb Weight	External research indicated that curb weight is the "total weight of a vehicle when it's ready to drive, including all standard equipment and a full tank of fuel." This made curb weight highly correlated with other size-related columns, which were then removed.
Compression Ratio and Horsepower	Both columns exhibited multicollinearity and were removed based on external research. Horse Power is indicative of engine size and city MPG, while Compression Ratio is the ratio between the volume of the cylinder and combustion chamber, which explains the high correlations.
door number and peakrpm	After running the OLS model Dropped door number and peakrpm based on their extremely high p-values when running the full OLS model. The adjusted R-squared increased from 0.883 to 0.884 after removing those predictors, allowing us to reduce the number of parameters while maintaining the same level of variance explained in the dependent variable (y).
Engintype_rotor and cylindertype_two	Both variables had VIF's of INF (extremely large) so levels were removed.
country_Japan & enginesize	Both removed for having VIFs.

Post-Predictor Selection VIFs:

	Feature	VIF
0	const	1959.639054
1	symboling	3.093180
2	aspiration	2.468506
3	doornumber	2.554264
4	enginelocation	2.107685
5	peakrpm	3.092902
6	citympg	6.388578
7	volume	6.754233
8	drivewheel_fwd	8.129230
9	drivewheel_rwd	9.745500
10	enginetype_dohcv	1.494620
11	enginetype_l	3.804936
12	enginetype_ohc	9.676452
13	enginetype_ohcf	5.125719
14	enginetype_ohcv	2.780800
15	cylindernumber_five	6.352480
16	cylindernumber_four	19.730605
17	cylindernumber_six	7.628052
18	fuelsystem_2bbl	6.408211
19	fuelsystem_4bbl	2.695832
20	fuelsystem_idi	5.962103
21	fuelsystem_mfi	1.336545
22	fuelsystem_mpfi	9.699978
23	fuelsystem_spdi	2.694762
24	fuelsystem_spfi	1.194701
25	country_Germany	3.052139
26	country_Italy	1.489752
27	country_Sweden	2.279577
28	country_UK	1.366485
29	country_USA	1.637692

IV. Modeling

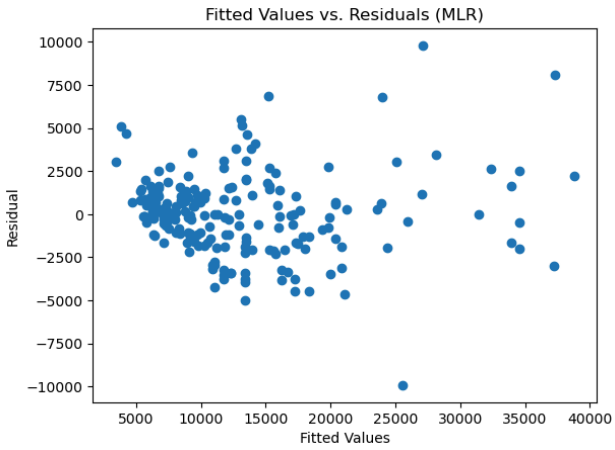
1. MLR Results:

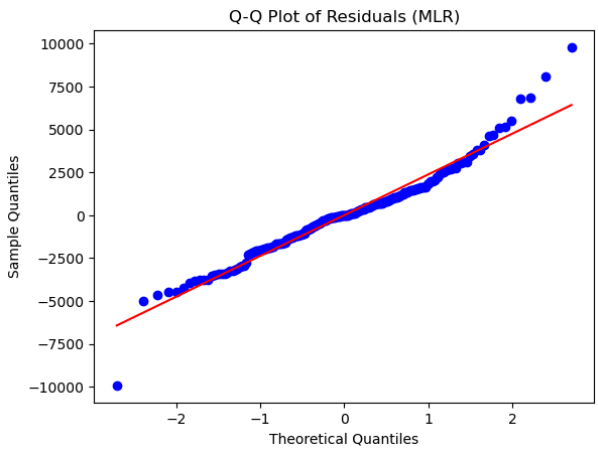
OLS Regression Results

Dep. Variable:	price	R-squared:	0.899
Model:	OLS	Adj. R-squared:	0.884
Method:	Least Squares	F-statistic:	57.65
Date:	Thu, 10 Oct 2024	Prob (F-statistic):	1.13e-72
Time:	23:03:29	Log-Likelihood:	-1859.5
No. Observations:	202	AIC:	3775.
Df Residuals:	174	BIC:	3868.
Df Model:	27		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.967e+04	5255.350	3.743	0.000	9296.716	3e+04
symboling	406.3839	211.506	1.921	0.056	-11.064	823.832
aspiration	-2480.4587	730.832	-3.394	0.001	-3922.895	-1038.022
enginelocation	-1.428e+04	2088.717	-6.835	0.000	-1.84e+04	-1.02e+04
citympg	-109.4887	72.071	-1.519	0.131	-251.735	32.757
volume	0.0368	0.005	6.718	0.000	0.026	0.048
drivewheel_fwd	528.9241	1055.318	0.501	0.617	-1553.949	2611.797
drivewheel_rwd	4424.8707	1165.143	3.798	0.000	2125.238	6724.503
enginetype_dohcv	1799.0515	3158.498	0.570	0.570	-4434.849	8032.952
enginetype_l	-1073.1145	1528.877	-0.702	0.484	-4090.646	1944.417
enginetype_ohc	3099.8915	1127.167	2.750	0.007	875.212	5324.571
enginetype_ohcf	3354.1130	1388.917	2.415	0.017	612.819	6095.407
enginetype_ohcv	1781.2977	1273.887	1.398	0.164	-732.963	4295.558
cylindernumber_five	-1.236e+04	1944.455	-6.355	0.000	-1.62e+04	-8519.577
cylindernumber_four	-1.581e+04	1931.172	-8.187	0.000	-1.96e+04	-1.2e+04
cylindernumber_six	-1.076e+04	1579.407	-6.811	0.000	-1.39e+04	-7639.767
fuelsystem_2bbl	-1717.9498	880.788	-1.950	0.053	-3456.353	20.454
fuelsystem_4bbl	-1.48e+04	2464.864	-6.006	0.000	-1.97e+04	-9939.584
fuelsystem_idi	-1210.3610	1301.722	-0.930	0.354	-3779.559	1358.837
fuelsystem_mfi	-2144.1830	2954.260	-0.726	0.469	-7974.980	3686.614
fuelsystem_mpfi	-1487.0367	1052.602	-1.413	0.160	-3564.548	590.475
fuelsystem_spdi	-1772.0951	1414.318	-1.253	0.212	-4563.522	1019.332
fuelsystem_spfi	-3037.4262	2807.477	-1.082	0.281	-8578.520	2503.668
country_Germany	1613.1115	807.860	1.997	0.047	18.646	3207.577
country_Italy	4308.4372	1762.666	2.444	0.016	829.479	7787.396
country_Sweden	162.9887	917.661	0.178	0.859	-1648.192	1974.169
country_UK	1.347e+04	2134.321	6.312	0.000	9260.225	1.77e+04
country_USA	1557.0983	671.450	2.319	0.022	231.863	2882.334

Omnibus:	21.037	Durbin-Watson:	1.547
Prob(Omnibus):	0.000	Jarque-Bera (JB):	60.991
Skew:	0.351	Prob(JB):	5.70e-14
Kurtosis:	5.599	Cond. No.	1.97e+07



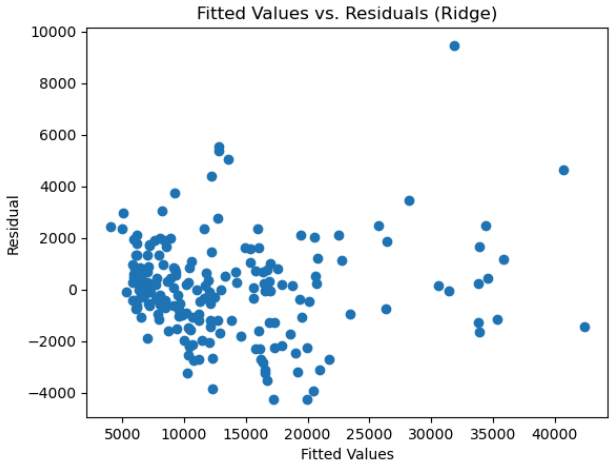


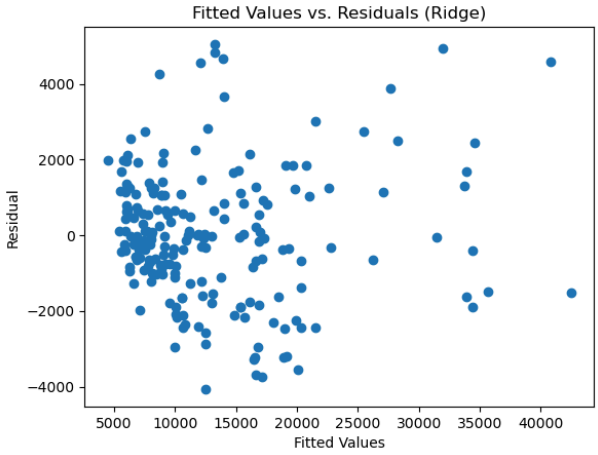
The MLR model achieved an adjusted R-squared of 0.884, meaning 88.4% of the variance in car prices is explained by the MLR model. Additionally, the model achieved an F-statistic of 57.65, garnering a near zero p-value indicating this model as a whole is statistically significant. Some of the significant variables at the 5% significance level include aspiration, enginelocation, volume, the OHC and OHCF engine types, 5 and 6 cylinder engines, and fuel system types 4bbl, idi, spdi. For example, holding all variables constant, rear-engine cars cost $14,280 more on average than front engine cars. This follows, because typically only sports cars are rear engine vehicles. There are variables and categorical levels that are insignificant, which include drivewheel_fwd, symboling, enginetype_dohcv, drivewheel_rwd, fuelsystem_mpi, and fuelsystem_spfi. When analyzing the fitted values and residuals to assess the adherence to linear assumptions, it can be seen the residuals appear to be somewhat scattered, but there's a slight pattern where the residuals spread out as the fitted values increase. This suggests that the model may suffer from heteroscedasticity, meaning the variance of the errors increases with the fitted values. Also, when examining the QQ-plot of residuals, it can be seen the points mostly follow the red line, but there are deviations at the tails, indicating some potential non-normality of residuals, particularly for the larger residuals.

RIDGE/LASSO

Ridge and Lasso regression are regularization techniques aimed at improving model performance by addressing unnecessary predictors. Both methods add a penalty term to the regression coefficients, which helps to control the influence of less relevant features. Ridge regression minimizes the risk of overfitting by shrinking coefficients, which leads to better generalization across different datasets. Lasso regression goes a step further by not only shrinking coefficients but also forcing some to exactly zero. This feature selection capability allows Lasso to exclude irrelevant predictors, resulting in a more simplified and interpretable model. Given this, we applied these techniques to a different dataset subset—one with less focus on feature selection to manage multicollinearity compared to our MLR model.

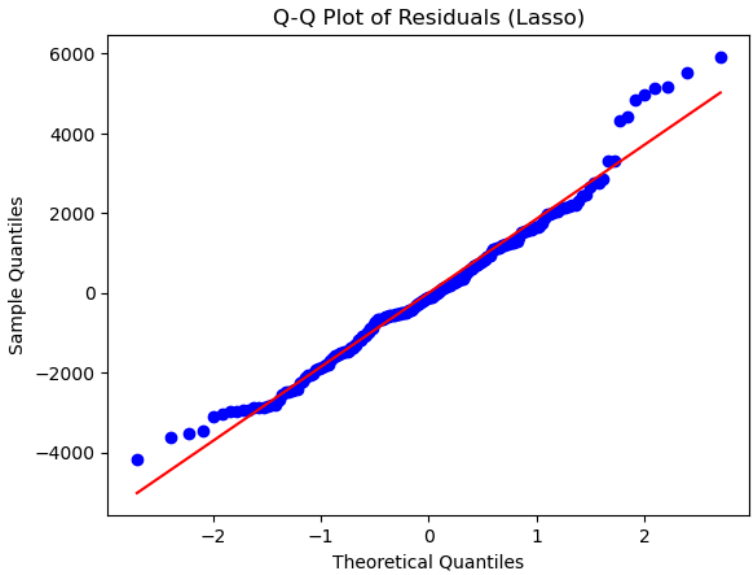
After fitting both models one of our initial observations we noticed was an outlier point, which was skewing the Fitted Values vs Residuals plot. This was a discrepancy point in our fitted values, and made the plot seem to have more constant variance than there was. Below are the plots, from the Ridge model, before and after removing this instance from the dataset.





However, when we re-fit the Ridge regression model again, none of the coefficients shrank towards zero, with the smallest predictor (cylindernumber_two) having an absolute value of around 75. Now, this is a good thing, but it can be rare for every predictor to be significant, when there are so many in a model of this size. This goes even further considering the multicollinearity mess was found during our initial data analysis.

To confirm this, no variables were snapped to zero when running the Lasso model. The Lasso model showed to have an MSE of $3,448,576.61. This may seem like a lot at first, but in the context of predicting car prices, with over 200 observations, it is a low MSE. The model also protected the assumptions of regression, that our residuals follow a normal distribution, and that we are not skewed in any direction nor heavy or light tailed. Below is the QQ plot generated from the model.

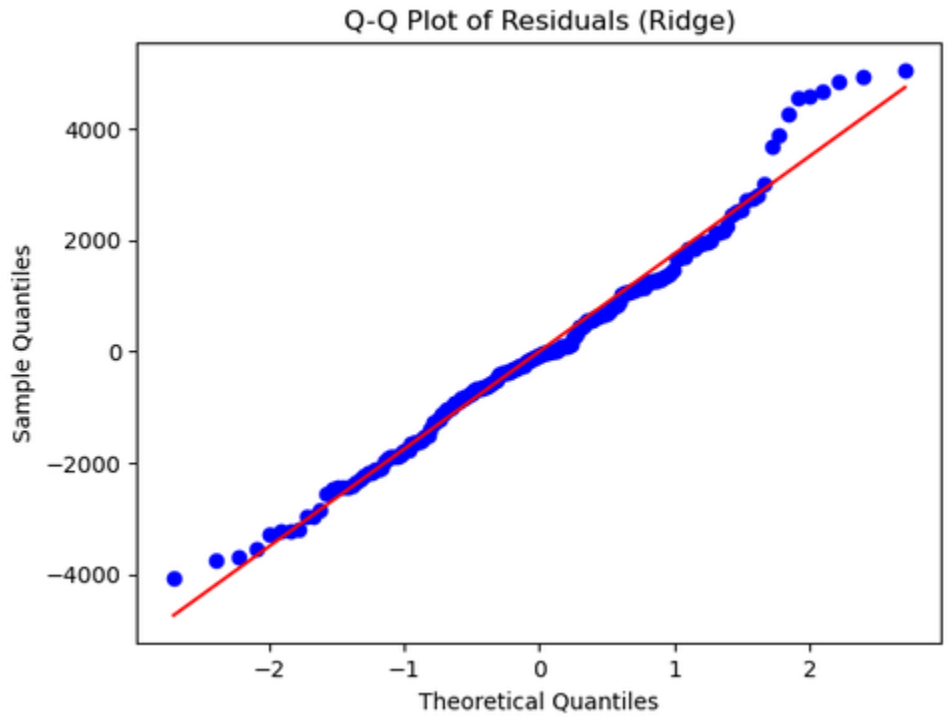


Having both the Ridge and Lasso model agree that all coefficients have significance shows that these are a good set of predictors to use for the response. We can also conclude that the data processing was effective, and any insignificant predictors were taken out in that procedure. Once we had all three models, it was time to test them against one another. Now, we had to decide which model fit our data the best, and gave a clear understanding to our problem statements we defined from the beginning. As previously mentioned, our criterion for the models were AIC, BIC, MSE, and adjusted R^2. When running our benchmarks across the models, Ridge and Lasso outperformed the OLS. Originally, our hypothesis was that it would be a close call between our Ridge and Lasso models, and these values agreed with the prediction. Ridge scored perfect, having an AIC and BIC with just a difference of under 25 against Lasso, while having an adjusted R^2 not even 0.01 greater than Lasso.

	Model	AIC	BIC	MSE	Adjusted R2
0	OLS	3774.906405	3867.537900	6.806060e+06	0.883845
1	Lasso	3130.801370	3279.673416	3.448577e+06	0.923395
2	Ridge	3106.708038	3255.580084	3.060835e+06	0.932008

We truly understand the difference when looking at MSE. This mean squared error was $3,060,834.67, around $400,000 under our mean squared error for Lasso. Though our metrics show the difference between the two models is miniscule, in the context of the problem, our Ridge was the most dependable. We also noticed, through a QQ-Plot, that our residuals did not

violate our initial model assumptions. In fact, we saw less variance in this plot than the one from Lasso, strengthening the claim that Ridge fits the data better.



V. Conclusions

Here, we aimed to model U.S. car prices based off of various factors to assist Geely Auto in assessing their entrance into the American market. Using the dataset we sourced, we employed multiple linear regression models, and regularization techniques, like Ridge and Lasso regression to identify key predictors influencing these prices. The data included over 200 observations containing 25 statistics per car, along with the price of the vehicle. Through exploratory data analysis, transforming categorical variables, removing influential points, as well as addressing multicollinearity issues, our analysis helped result in an interpretable, accurate model that provides valuable insights for Geely Auto.

Key features such as cylinder number and engine showed significant impacts on car pricing, having the lowest p-values in our final model. When considering the context of the problem, horsepower was also a massive contributor to the price of the car. The coefficient value we saw here was just under 640, and with the predictor having an average value of over 100,

horsepower was consistently influencing almost $60,000 towards the final price. When doing data cleaning, multiple predictors were combined into one, 'Volume` variable, which ended up being another significant predictor. In the final model, we observed a coefficient value of -3531.17. Finally, the country the car was manufactured in also played a role, as the difference of a German car compared to a Japanese car was around $100 dollars, when all other factors considered remained constant. These factors in our final Ridge regression model also explain the variation in price extremely well, factoring in for 93% of the total variation.

Our final selected Ridge regression model explained 93% of the total variation in car prices, with our initial MLR model and Lasso Regression model achieving adjusted R-squareds of 0.884 and 0.923 respectively. These models provided valuable guidance into the key factors driving car prices, offering Geely Auto a clear framework for understanding market dynamics. By identifying both significant and insignificant predictors, the model helps Geely focus on critical aspects of vehicle design and positioning, ultimately putting them in a strong position as they enter the U.S. market.