

# Predicting Car Prices Based on Certain Characteristics

DSCI310 Group 07: Harbor Zhang, Jiaying Liao, Ning Wang, Xiwen Wei

2023-04-09

Original Project Authors: Henry Zhang, Moira Renata, Ning Wang, Paige Wills, Xinrui Wang in STAT 301 Group 36.

## 1 Introduction

Over the past few decades, we have seen a rapid increase in demand for the car industry. The high market price of both brand new and used cars have created a large economic impact all over the world. Based on previous studies, it was found that there are multiple factors affecting the final price of a car (Balce 2016) and that while most factors do have a positive contribution or effect to the final price, there are still some factors that create a negative effect (Erdem and Şentürk 2009). Moreover, according to previous studies, the car price one of the most significant factor when people deciding whether to purchase a car(Armstrong 2022).

Therefore, in this project, we hope to create a model that allows us to predict the final price of a car given its characteristics.

## 2 Description

The sample we use is from the the Automobile Data Set that was created by Jeffrey C. Schlimmer in 1987 (Schlimmer 1987). The author created a data set that consists of 26 columns with 205 rows, where each row refers to one car sample. Out of the 25 columns predictor variables, there are 9 categorical variables and 16 numerical variables. Our response variable is the 26th column, which represents the price of a car in USD(\$).

Variable	Type	Description
symboling	Categorical	Assigned insurance risk rating
normalized-losses	Numerical	Relative average loss payment per insured vehicle year in dollars (USD)
make	Categorical	Car manufacturer/model
fuel-type	Categorical	Type of fuel to power car
aspiration	Categorical	Engine aspiration (std, turbo)
num-of-doors	Numerical	Number of doors
body-style	Categorical	Car's style (sedan, convertible, etc.)
drive-wheels	Categorical	amount and location of wheels
engine-location	Categorical	Engine location (front, back)

Variable	Type	Description
wheel-base	Numerical	Horizontal distance between the front and rear wheel in inches.
length	Numerical	Length of car in inches
width	Numerical	Width of car in inches
height	Numerical	Height of car in inches
curb-weight	Numerical	Weight of car in pounds
engine-type	Categorical	Engine type (dohc, dohc, etc.)
num-of-cylinders	Categorical	Number of cylinders in engine
Engine-size	Numerical	Engine size in cubic inches
fuel-system	Categorical	Fuel system in car (1bbl, mfi, etc.)
bore	Numerical	Diameter of each cylinder in inches
stroke	Numerical	Movement of piston in gigapascal
compression-ratio	Numerical	Ratio between the cylinder's highest and lowest volumes at the bottom and top of the piston's stroke.
horsepower	Numerical	Engine horsepower (hp)
peak-rpm	Numerical	RPM at which engine delivers peak horsepower
city-mpg	Numerical	Mileage in the city in miles per gallon
highway-mpg	Numerical	Mileage in the highway in miles per gallon
price	Numerical	Price of car in USD (\$)

### 3 Preliminary Analysis

In this section, we load and clean the data. Note that the all ? are replaced with NA.

From tables 2, we noticed that there are some NA values. Each row represents an observation, each column is a variable, and each cell is a value, which means there is not a lot of data tidying to do. We will first check the number of NA values in each column, the number of levels in columns that are categorical variables, and the summary statistics of each variable.

According to table 3, there are 45 rows that contain NA values. And the number of rows that have complete observations are 160.

### 4 Exploratory Data Analysis

Next, we will perform EDA to better understand the variables that we will be using in our analysis.

It would be beneficial to visualize the pairwise correlation coefficients of our dataset to check for multicollinearity. This can be done either by using the `ggpairs` function, or by creating a correlation heatmap. However, since our data contains multiple categorical variables with a large number of levels, this is not possible to do at this point. Therefore, our EDA is limited to checking the Coefficient of Determination of all the predictor variables and visualizing the relationship of the top 8 predictor variables based on their  $R^2$  value.

Firstly, we want to calculate the coefficient of determination of all of our predicted variables.

Table 2: Automobile Dataset

symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	
3	NA	alfa-romero	gas	std	two	convertible	
3	NA	alfa-romero	gas	std	two	convertible	
1	NA	alfa-romero	gas	std	two	hatchback	
2	164	audi	gas	std	four	sedan	
2	164	audi	gas	std	four	sedan	
2	NA	audi	gas	std	two	sedan	
drive-wheels		engine-location	wheel-base	length	width	height	curb-weight
rwd		front	88.6	168.8	64.1	48.8	2548
rwd		front	88.6	168.8	64.1	48.8	2548
rwd		front	94.5	171.2	65.5	52.4	2823
fwd		front	99.8	176.6	66.2	54.3	2337
4wd		front	99.4	176.6	66.4	54.3	2824
fwd		front	99.8	177.3	66.3	53.1	2507
engine-type	num-of-cylinders	engine-size	fuel-system	bore	stroke	compression-ratio	
dohc	four	130	mpfi	3.47	2.68	9.0	
dohc	four	130	mpfi	3.47	2.68	9.0	
ohcv	six	152	mpfi	2.68	3.47	9.0	
ohc	four	109	mpfi	3.19	3.40	10.0	
ohc	five	136	mpfi	3.19	3.40	8.0	
ohc	five	136	mpfi	3.19	3.40	8.5	
horsepower		peak-rpm	city-mpg	highway-mpg	price		
111		5000	21	27	13495		
111		5000	21	27	16500		
154		5000	19	26	16500		
102		5500	24	30	13950		
115		5500	18	22	17450		
110		5500	19	25	15250		

Table 3: Summary of Automobile Dataset

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style
	3 :27	Min. : 65	toyota : 32	gas :185	std :168	two : 89	convertible: 6
	1 :54	1st Qu.: 94	nissan : 18	diesel: 20	turbo: 37	four:114	hatchback :70
	2 :32	Median :115	mazda : 17			NA : 2	sedan :96
	0 :67	Mean :122	honda : 13				wagon :25
	-1:22	3rd Qu.:150	mitsubishi: 13				hardtop : 8
	-2: 3	Max. :256	subaru : 12				
		NA's :41	(Other) :100				
	drive-wheels	engine-location	wheel-base	length	width	height	curb-weight
	rwd: 76	front:202	Min. : 86.60	Min. :141.1	Min. :60.30	Min. :47.80	Min. :1488
	fwd:120	rear : 3	1st Qu.: 94.50	1st Qu.:166.3	1st Qu.:64.10	1st Qu.:52.00	1st Qu.:2145
	4wd: 9		Median : 97.00	Median :173.2	Median :65.50	Median :54.10	Median :2414
			Mean : 98.76	Mean :174.0	Mean :65.91	Mean :53.72	Mean :2556
			3rd Qu.:102.40	3rd Qu.:183.1	3rd Qu.:66.90	3rd Qu.:55.50	3rd Qu.:2935
			Max. :120.90	Max. :208.1	Max. :72.30	Max. :59.80	Max. :4066
	engine-type	num-of-cylinders	engine-size	fuel-system	bore	stroke	compression-ratio
	dohc : 12	four :159	Min. : 61.0	mpfi :94	Min. :2.54	Min. :2.070	Min. : 7.00
	ohcv : 13	six : 24	1st Qu.: 97.0	2bbl :66	1st Qu.:3.15	1st Qu.:3.110	1st Qu.: 8.60
	ohc :148	five : 11	Median :120.0	idi :20	Median :3.31	Median :3.290	Median : 9.00
	l : 12	three : 1	Mean :126.9	1bbl :11	Mean :3.33	Mean :3.255	Mean :10.14
	rotor: 4	twelve: 1	3rd Qu.:141.0	spdi : 9	3rd Qu.:3.59	3rd Qu.:3.410	3rd Qu.: 9.40
	ohcf : 15	two : 4	Max. :326.0	4bbl : 3	Max. :3.94	Max. :4.170	Max. :23.00
	dohcv: 1	eight : 5		(Other): 2	NA's :4	NA's :4	
		horsepower	peak-rpm	city-mpg	highway-mpg	price	
		Min. : 48.0	Min. :4150	Min. :13.00	Min. :16.00	Min. : 5118	
		1st Qu.: 70.0	1st Qu.:4800	1st Qu.:19.00	1st Qu.:25.00	1st Qu.: 7775	
		Median : 95.0	Median :5200	Median :24.00	Median :30.00	Median :10295	
		Mean :104.3	Mean :5125	Mean :25.22	Mean :30.75	Mean :13207	
		3rd Qu.:116.0	3rd Qu.:5500	3rd Qu.:30.00	3rd Qu.:34.00	3rd Qu.:16500	
		Max. :288.0	Max. :6600	Max. :49.00	Max. :54.00	Max. :45400	
		NA's :2	NA's :2			NA's :4	

Table 4: The 8 variables with highest  $R^2$ 

$\hat{R}^2$	names
0.796	make
0.761	engine-size
0.696	curb-weight
0.657	horsepower
0.63	num-of-cylinders
0.564	width
0.497	highway-mpg
0.477	length

Based on the result in table 4, the variable that has the highest  $R^2$  value is **make** with a value of 0.796. This can be interpreted as 79.6% of the variation observed in **price** is explained by the model with **make** as the explanatory variable.

Then, we created plots for the top 8 predictor variables. For the numerical variables, we created both a histogram to see the distribution, and a scatter plot to see the relationship between the variable and the car price. For the categorical variables, we created a bar graph to compare the count of each category in a variable. Analysis of the plots created are written after the code.

Analysis on the plots:

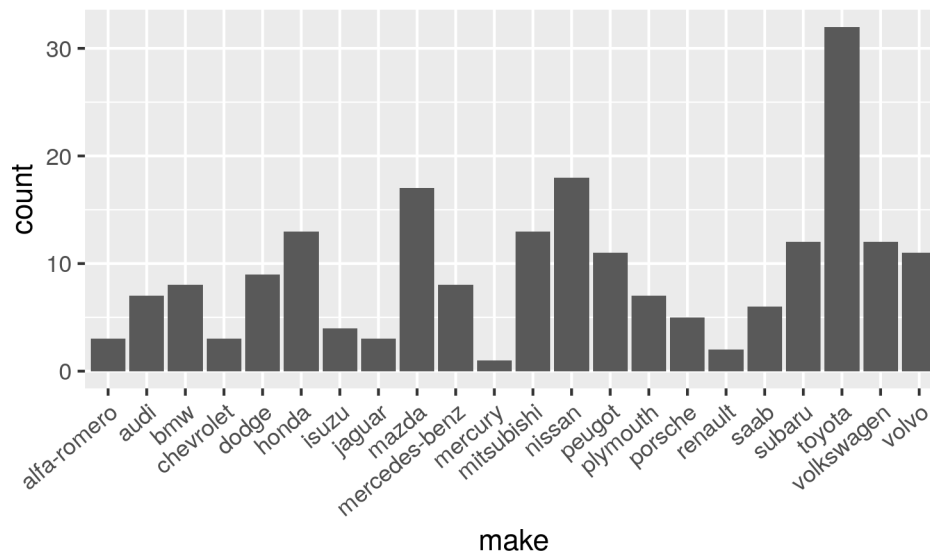


Figure 1: Distribution of make

- For the variable **make** (see Figure 1), we can see that Japanese brands, such as Toyota, Nissan and Mazda have the top 3 counts, which means they produce the most cars.

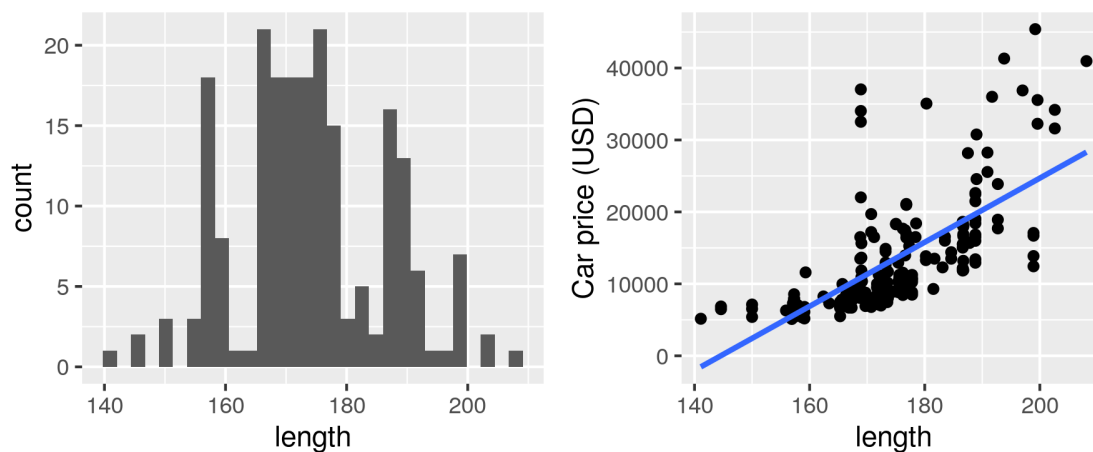


Figure 2: Distribution of length and price vs length

- For the variable **length** (see Figure 2), we can see the distribution is approximately normal and has a positive linear relationship with **price**.

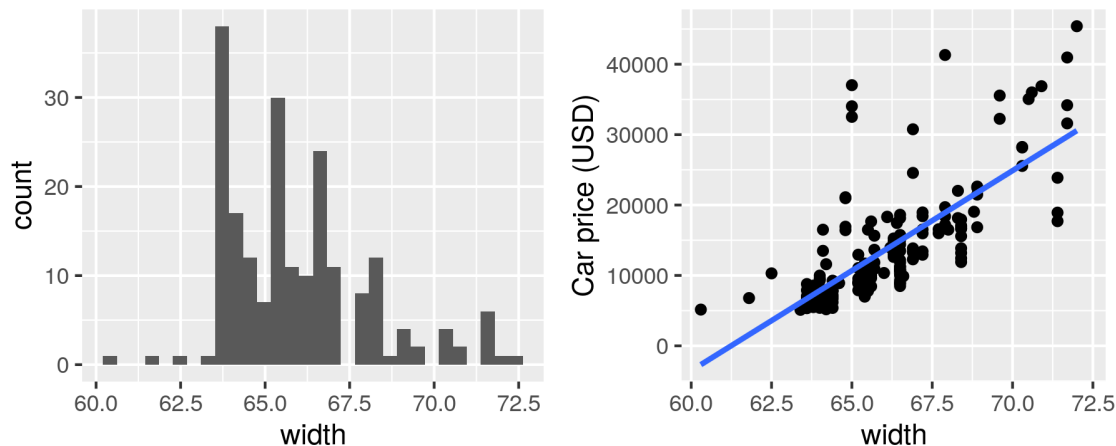


Figure 3: Distribution of width and price vs width

- For the variable **width** (see Figure 3), we can see the distribution is approximately normal and has a positive linear relationship with **price**.

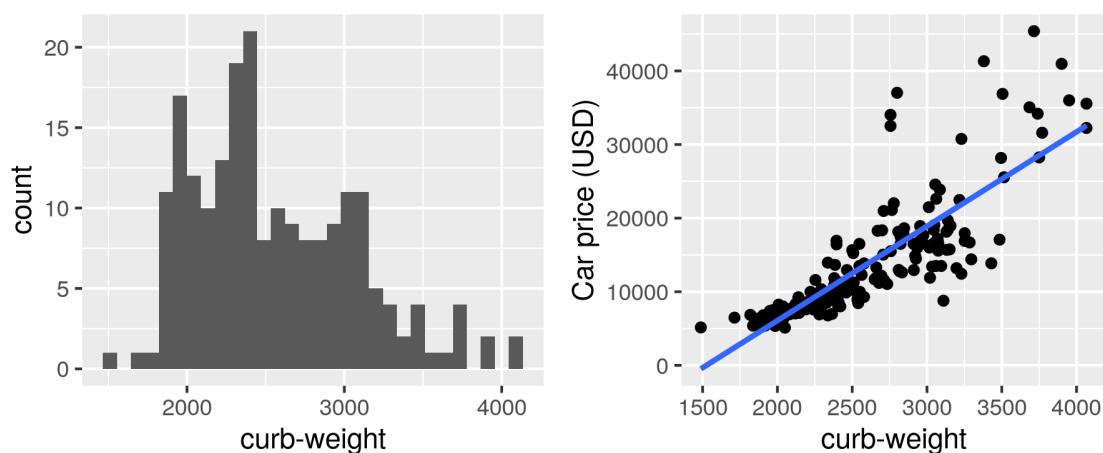


Figure 4: Distribution of car weight and price vs weight

- For the variable **curb-weight** (see Figure 4), we can see the distribution is skewed to right and has a positive linear relationship with **price**.
- For the variable **num-of-cylinders** (see Figure 5), we can see that most cars have 4-cylinders.
- For the variable **engine-size** (see Figure 6), we can see the distribution is skewed to right and has a positive linear relationship with **price**.
- For the variable **horse-power** (see Figure 7), we can see the distribution is skewed to right and has a positive linear relationship with **price**.
- For the variable **highway-mpg** (see Figure 8), we can see the distribution is approximately normal and has a negative linear relationship with **price**.

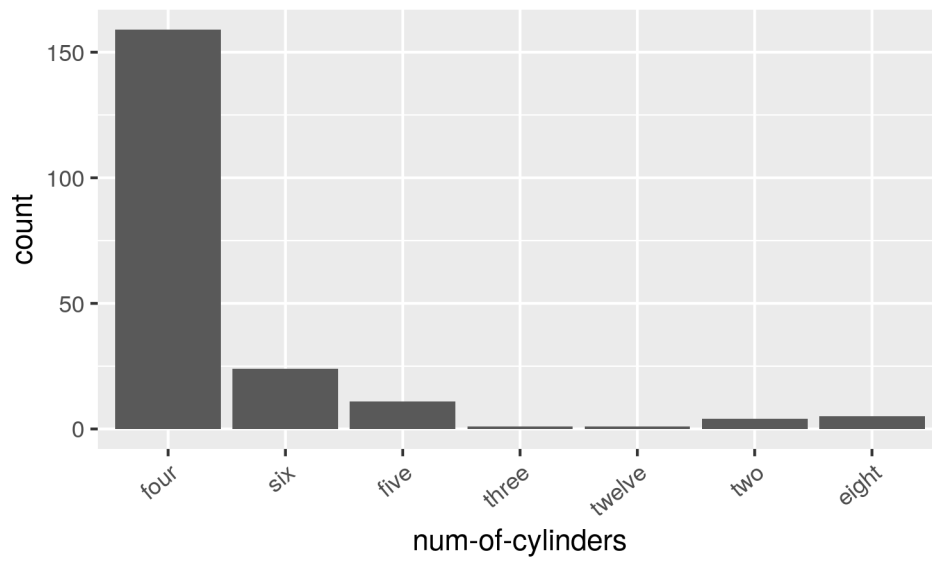


Figure 5: Distribution of number of cylinders

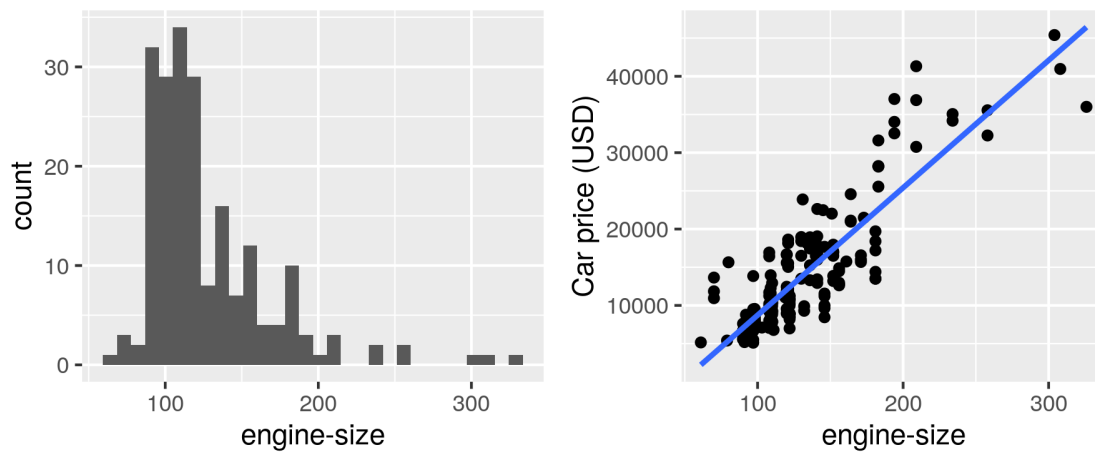


Figure 6: Distribution of engine size and price vs engine size

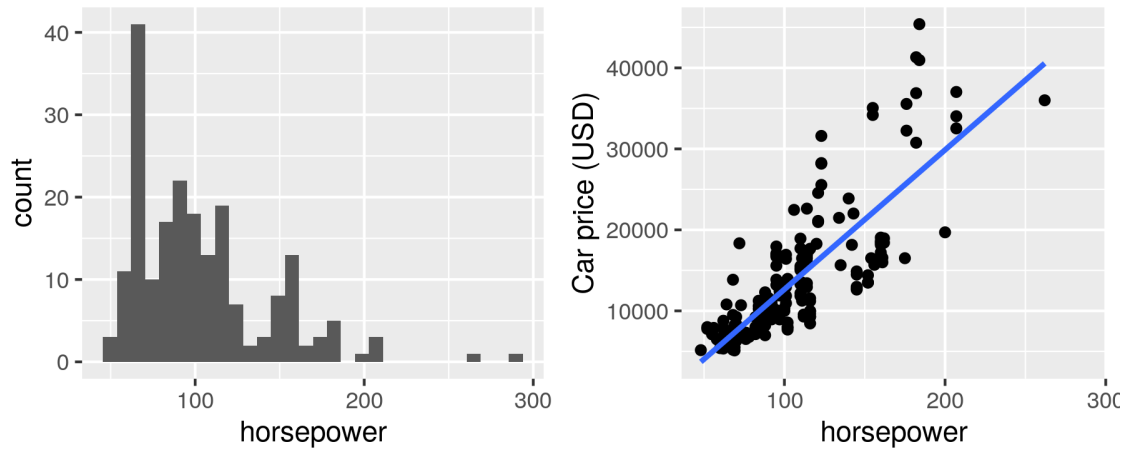


Figure 7: Distribution of horse power and price vs horse power

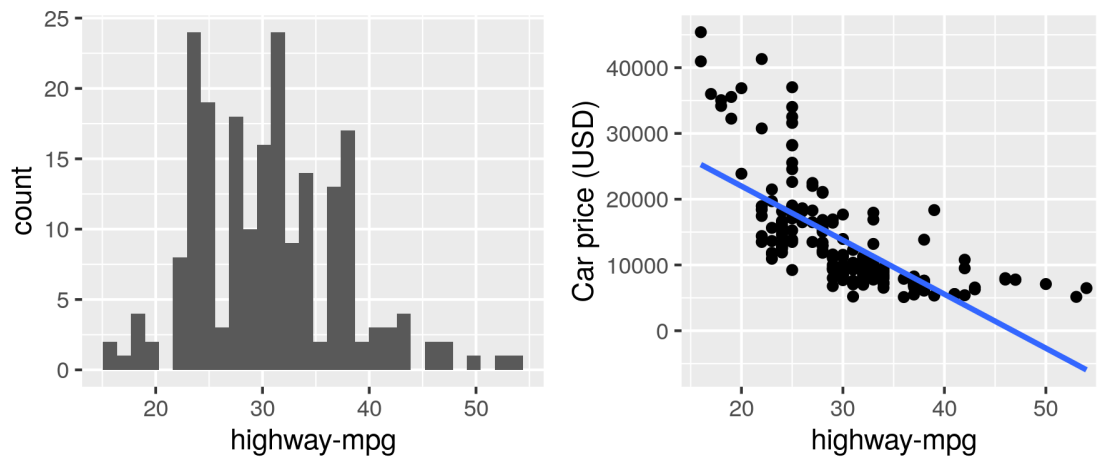


Figure 8: Distribution of highway mpg and Price vs highway mpg



Table 5: Levels of all variables

num_unique	type
6	factor
51	numeric
18	factor
2	factor
2	factor
2	factor
5	factor
3	factor
1	factor
40	numeric
56	numeric
33	numeric
39	numeric
136	numeric
5	factor
5	factor
32	numeric
6	factor
33	numeric
31	numeric
29	numeric
48	numeric
20	numeric
25	numeric
28	numeric
145	numeric

## 5 Methods

Based on table 5, we noticed that the variables: symboling, make, num-of-doors, body-style, drive-wheels, engine-type, num-of-cylinders, fuel-system have more than 2 levels. Since the shrinkage methods we are planning to use to perform model selection (LASSO and Ridge) is not possible when there are more than 2 levels in a categorical variable, the variables listed above are all dropped because of their high levels.

Apart from that, after we omit NA, the levels of engine-location appears to be 1. This will cause contrasts since we need a categorical variables to be factors with 2 or more levels. Thus we need to remove `engine-location`.

Then, the data set is split into two data sets - training and testing using a 70-30% basis and the ID variables are removed.

We then create new training and testing datasets that excludes the variables listed. We call them:

1. `training_df_sub`
2. `testing_df_sub`

This code prepares the dataset(s) for `glmnet()` which only takes matrices (hence `model.matrix`). The `glmnet()` function has an argument `object`, which is the formula of the model and therefore needs clear x and y variables, explaining why the training and testing datasets are split into subsets of x and y.

Now our data is prepared for the `glmnet()` function, we will use `cv.glmnet` to obtain the optimal value of lambda using the training set. Since this is a LASSO model, we will use the argument `alpha=1` and `n.folds=10` to find the optimal value of lambda using cross-validation by defining a sequence of values.

Then the plot function will be used to visualise the MSE of different lambdas.

`lasso_mod` provides the  $\hat{\lambda}_{\min}$  for LASSO (explained below) and `lasso_mod_1se` provides the  $\hat{\lambda}_{1SE}$  for LASSO (explained below).

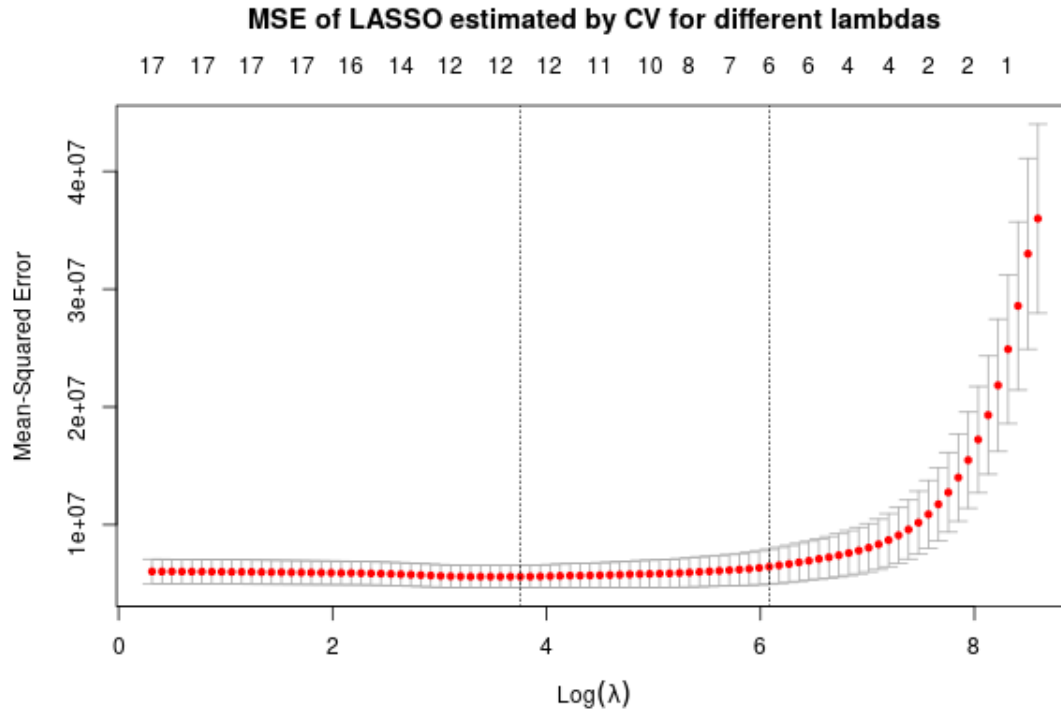


Figure 9: Lasso Regression

Figure 9 shows the estimated test MSE on the y-axis for a grid of values of  $\lambda$  on the x-axis (on a natural log-scale). The two vertical dotted lines show us where lambda is minimized, in other words, how many variables are needed for the best model. The numbers on the top x-axis indicate the number of input variables whose estimated coefficients are different from 0 for different values of lambda. The error bars represent the variation across the different sets of the CV folds. The left line shows  $\hat{\lambda}_{\min}$  - which is the minimum MSE in the grid and the right line represents  $\hat{\lambda}_{1SE}$  - which is the largest values of lambda such that the corresponding MSE is still within 1 standard error of that of the minimum (more penalization at low cost).

A similar method is followed for Ridge, except `alpha = 0`. `ridge_mod` provides the  $\hat{\lambda}_{\min}$  for Ridge and `ridge_mod_1se` provides the  $\hat{\lambda}_{1SE}$  value for Ridge (explained above).

Figure 10 for Ridge regression shows the estimated test MSE for each value of lambda, just like that of LASSO. However the main difference here is that the top x-axis is all the same value - 17. This is because the Ridge estimator never shrinks estimates to 0, unlike LASSO. The two vertical lines represent  $\hat{\lambda}_{\min}$  and  $\hat{\lambda}_{1SE}$  with the x and y axis being the same as LASSO.

For explanatory analysis purposes, we will be using both  $\hat{\lambda}_{\min}$  and  $\hat{\lambda}_{1SE}$  for both LASSO and Ridge to create four different regression models. Additionally, we also be creating an OLS model for comparison. The 5 models we will be creating are listed below:

1. `mod_lasso`: LASSO regression using  $\lambda = \hat{\lambda}_{\min}$  from LASSO

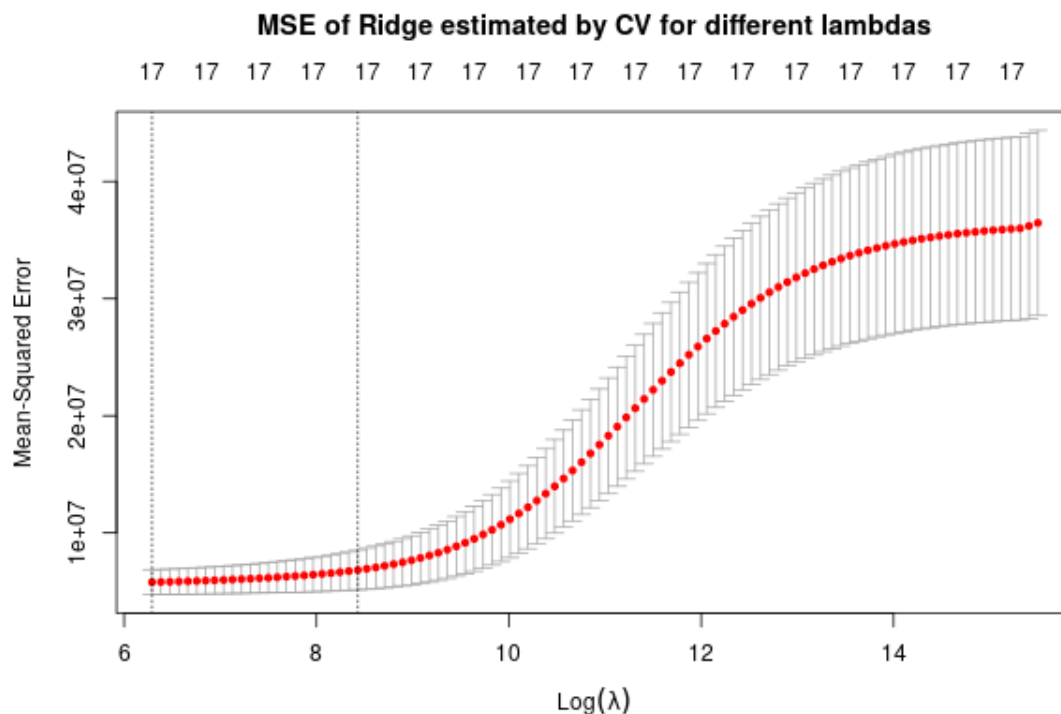


Figure 10: Ridge Regression

Table 6: Cross-Validation Result

Model	R_MSE
LASSO Regression with minimum MSE	671.366
LASSO Regression with 1SE MSE	810.066
Ridge Regression with minimum MSE	658.405
LASSO Regression with 1SE MSE	845.964
OLS Full Regression	1052.463

2. `mod_lasso_1se`: LASSO regression using  $\lambda = \hat{\lambda}_{1SE}$  from LASSO
3. `ridge_mod`: Ridge regression using  $\lambda = \hat{\lambda}_{min}$  from Ridge
4. `ridge_mod_1se`: Ridge regression using  $\lambda = \hat{\lambda}_{1SE}$  from Ridge
5. `ols_fs`: Ordinary least squares full regression using  $\lambda = 0$

After creating the 5 models, we will then obtain the out-of-sample predictions for the test sets of all five different models above, shown by `preds_1`, `preds_2`, `preds_3`, `preds_4` and `preds_5`.

Finally, we are able to compute the RMSE (root mean squared error) to evaluate the predicted models, which is clearly summarised in the tibble below.

Through the 10 fold cross validation error (root mean squared error) in table 6, RMSE are in the same scale across all models. Therefore, due to model simplicity and robustness to outliers, we decide to use the LASSO Regression model with 1se MSE for our final predictions.

Beside that, we obtained a root mean squared prediction error of 578.754 when using the LASSO Regression model on the test set.

Table 7: Kept variables

kept_variables	coefficient
(Intercept)	-15273.682
width	113.379
'curb-weight'	7.310
horsepower	13.261

Taking a look at the coefficients of our model in table 7, we noticed that the LASSO model had selected only three variables, which are **width**, **curb-weight**, and **horsepower**, while all the coefficients of other input variables were reduced to 0.

## 6 Discussion

Our goal requires generating a prediction model with potential independent variables that can predict the price of the car. Based on our exploratory data analysis and regression model comparisons, we chose the LASSO model with  $\hat{\lambda}_{1SE}$ , which we expect to have good prediction performance.

### 6.1 Summary

Based on the results above, the variables **width**, **curb-weight**, and **horsepower** were chosen by the LASSO model. With `lasso_mod_1se` (more penalization at a low cost) to penalize, all of the other regression coefficients of the input variables were shrunk to 0. If we refer back to our EDA, we notice that the three variables selected by LASSO are included in the list of top 8 variables with the highest coefficient of determination. However, it was surprising to see that there were only 3 predictor variables in our final LASSO model, which means all the other variables were shrunk to 0. Some variables that we thought were going to be important, like **city-mpg**, **length**, and **height** were surprisingly not included in the final model.

LASSO penalizes the residual sum of squares with  $L_1$  penalty, and the penalty parameter  $\hat{\lambda}_{1SE}$  that we chose was selected through a process called tuning in order to avoid using the test set when creating our model. Although this shrinkage method (LASSO) might lead to bias of the estimated coefficients, we sacrifice this for a lower variance to gain better prediction performance and robustness in our model.

We hope that this fitted LASSO model will allow users to predict the price of a car in USD based on the 3 variables that were selected. Although we initially expected to have more predictor variables, we believe that the 3 predictor variables can give a rough prediction of the price of a car (USD). Moreover, we believe such a model could not only provide the expected price of a new car to customers, but also help sellers of second-hand cars set ideal prices.

### 6.2 Further Questions and Improvements

There are two main problems that need to be improved:

1. Using high-level ( $N > 2$ ) categorical variables in the LASSO model. The LASSO model interprets N-1 dummy variables as its own separate variable, which may exclude certain levels. To deal with this issue, we dropped the variables with high levels. However by doing so, we might have dropped a statistically significant variable. In future research, maybe by using another regression model, including those categorical variables with more than 2 levels, we can improve our model's performance.
2. Another problem involves LASSO's biased estimators. In the future, if we want to generate an inference model, we can apply post-lasso. This is done by fitting a regular least squares model to the variables selected by LASSO.

## References

- Armstrong, M. 2022. “Most Important Factors When Buying a Car.” Digital image. <https://www.statista.com/chart/13075/most-important-factors-when-buying-a-car/>.
- Balce, Andim Oben. 2016. “Factors Affecting Prices in an Used Car e-Market.” *Journal of Internet Applications and Management* 7 (2): 5–20. <https://doi.org/10.5505/iuyd.2016.30974>.
- Erdem, Cumhur, and İsmail Şentürk. 2009. “A Hedonic Analysis of Used Car Prices in Turkey.” *International Journal of Economic Perspectives* 3 (January): 141–49.
- Schlimmer, Jeffrey C. 1987. “Automobile Data Set.” Machine Learning Repository. <https://archive.ics.uci.edu/ml/datasets/Automobile>.