**For Linear Model:**

For the model\_1: multiple regression we used 13 variables, the R-squared is 0.857, which indicate a pretty well fit to the data. From the output, we can tell which of those 13 variables are statistically more significant to the car price (Y) variable and which are not.

For the model\_3: single variable least squared regression we use engine size as X and car Price as Y. R-squared is 0.7899 which indicate a good fit.

The bootstrap method is used to quantify the uncertainty associated with given statistical estimator or with a predictive model. The bootstrap procedure consists of randomly selecting a sample of n observations from the original data set. The subset, called bootstrap data set is then sued to evaluate the model. The bootstrap process, repeated large number of times and the standard error of the bootstrap estimate is then calculated. The results provide an indication of the variance of the model's performance.

For model\_1, we resampled 500 and test its model prediction performance, R-squared is 0.821, slightly lower than original 0.857. but still a good fit.

For model\_3. we also resampled 500 and test the prediction performance, R-squared is 0.787, almost identical to original R-square 0.789, which indicate a good model prediction performance.

**For Polynomial Model:**

They Polynomial model and log transformation also works on the data set(with only “engines size” as explanatory variable ), However, the R-squared is slightly lower than the OLS method. In this case, we say we prefer the OLS and multiple regression method over the polynomial model.

**For Lasso/Ridge Model:**

Ridge regression shrinks the regression coefficients, so that variables, with minor contribution to the outcome, have their coefficients close to zero. Ridge regression works better if responsive variables are affected by explanatory variables of roughly equal coefficients. From the model output, we can see that “peakrpm”, “curbweight” were shrink to roughly to zero. However, this model only told us which of those 13 variables have minimal impact on car price but did not tell us which have the most.

Lasso stands for Least Absolute Shrinkage and Selection Operator. It shrinks the regression coefficients toward zero by penalizing the regression model with a penalty term called L1-norm, which is the sum of the absolute coefficients. Lasso regression works better if some of the explanatory variables affect responsive variables significantly while other variables have very small coefficient. From the model output, we can tell “highwaympg”, “citympg”, “bore”, “length”, “wheelbase” have been dropped compared to 8 other variables.

In this case, Lasso regression works better, it helped us reduce the complexity of the model by dropping irrelevant variables. But it still didn’t show which of those variables have the most impact on car price.

**For Random Forest Model:**

The random forest method is slightly more complicated than other method. According to the output, the number of tress providing the lowest rate is at 97 tress with an average car-sales-price error of $2013.53. With the caret workflow, which invoked the randomforest() function to automatically select the optimal number(mtry) [which is 7] of predictor variables randomly sampled as candidates at each split. From the variable importance chart, 13 variables are ranked from the most important, such as “enginesize”, “curbweight”, “citympg” to the least important, such as “compressionratio”, “height”, “stroke”.

**Summary:**

Multiple Regression Model/Simple Linear Regression model is easy to use and the result is clean to interpret.

Random Forest Method is also preferable as it showed importance of variables but it did not tell us the specific variable coefficient.

Polynomial Regression Model works well on single variable but we still prefer OLS method.

Lasso/Rdige helped us reduce model complexity and eliminate irrelevant variables.