Model Building

- Normalized version of price (price_log) is used for modelling the dataset.
- Categorical variable 'Name' can be removed as it does not have any significance to our model
- Kilometers_Driven, Price and price_log are the target dependent variables so they can be dropped from train X dataset
- The entire dataset of 7253 used car entries can be split into train and test dataset on a 70% - 30% ratio

Model Performance Comparisons

	Model	Train_r2	Test_r2	Train_RMSE	Test_RMSE
0	Linear Regression	0.704073	0.657405	5.852769	6.329361
1	Decision Tree	0.368083	0.379616	8.552623	8.517250
2	Ridge Regression	0.695206	0.671818	5.939807	6.194791
3	Random Forest Regressor	0.406375	0.411432	8.289439	8.295979

Linear Regression Model

R-sqaure on training set : 0.7040732008196231
R-square on test set : 0.6574047216841202
RMSE on training set : 5.8527691291856705
RMSE on test set : 6.329361466562765

- The RMSE of test data is greater than the train data, that means the model over fitted the data.
- The R- squared values of Train and Test data > 0.6 and hence better in predicting the target value.

Drawbacks - There are no significant categorical variables identified by Linear Regression model.

Ridge/Lasso Regression Model

Ridge Model has a good prediction score of **0.8046030265770266**. Hence t
his model is slightly better than Linear Regression. But Lasso was helpful in i
mportant feature selection as there are only few parameters with more signif
icances. Price and Year are identified as the significant features.

Decision Tree

The R-squared values of train and test data are closer. The RMSE values look
underfitting the dataset.

The R-squared value for the model is less than 0.5 so its performing weaker
than Linear Regression

Power is identified as the most important feature that can influence the price
of the used cars. Year comes as the second most influencer. But the gap
between their importance is higher.

Random Forest Regressor

The model has performed weaker with the R-squared and RMSE values.

RMSE value is also higher than linear regression so its weaker

But the RMSE values of Train and Test data are much closer than decision
tree and there is no overfitting or underfitting of data.

Power and Year are identified as the most important features that can
influence the price of the used cars. Both features have significant
importance.

Refined Insights

- Overall, the Engine Power and Year has a significant impact on the price of the Used Cars. More the power of the engine, the car is priced higher. More the car is recently made, it’s priced expensive and it’s more from the first hand owner.
- Cities with more cosmopolitan population and cities with higher growth of IT and other technological job sectors have more demand for cars.

More first-hand owners tend to switch to a new car with modern features.

- Overall, the Linear Regression and Ridge Regressor models performed better than a Decision Tree and Random Forest. So these models can be best fit. But feature importances are not identified by Linear Regression, so we can use Lasso for retrieving the significant features.