Final Paper

Naomi Golin

Feature	Response
Name	Naomi Golin
SID	805173666
Kaggle Nickname	Naomi Golin Lecture 2
Kaggle Rank	69
Kaggle R^2	0.91949
Total Number of Predictors	10
Total number of Betas	21
BIC	-6747.13
Complexity Grade	100

Abstract

The purpose of this project was to generate a multiple linear regression model that best predicts the price of cars. With a Chinese automobile company aspiring to enter the US market, we were to act as "an automobile consulting company to understand the factors on which the pricing of cars depends" (Almohalwas 2021, 1) on. Through regression techniques, a multiple linear regression model was built with a select number of variables to predict the price of cars. In total, 20 predictors were used. The model was first developed through a training dataset, and was then submitted to a class Kaggle competition, in which it placed number 69 on the leaderboard with an R^2 value of 0.9149.

Introduction

The purpose of this project was to act as an automobile consulting company to understand the factors on which the pricing of cars depends on. Utilizing techniques of multiple linear regression, we were provided a dataset that contained 23 predictors and 1500 observations of different types of cars across the American market (Kaggle 2021) to try and find the best predictors for car price levels. The final regression model was then submitted to a class Kaggle competition, in which it was used to predict the price of cars of a testing data set that contains 500 observations. Below are the set of initial predictors provided with the dataset:

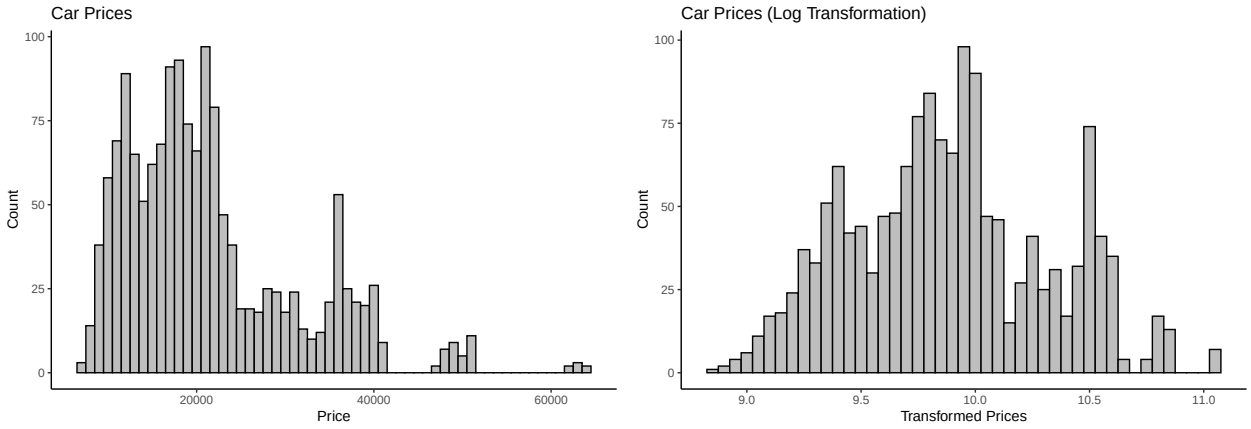
Table 2: Type of Predictors

Variables	Type
Manufacturer	Categorical
Model	Categorical
Type	Categorical
MPG.highway	Numerical
AirBags	Categorical
DriveTrain	Categorical
Cylinders	Categorical
EngineSize	Numerical
Horsepower	Numerical
RPM	Numerical
Rev.per.mile	Numerical
Man.trans.avail	Categorical
Fuel.tank.capacity	Numerical
Passengers	Numerical
Length	Numerical
Wheelbase	Numerical
Width	Numerical
Turn.circle	Numerical
Rear.seat.room	Numerical
Luggage.room	Numerical
Weight	Numerical
Origin	Categorical
Make	Categorical

Methodology

The Response Variable: Price

The first thing that we did was check the normality of the response variable, Price, to try and determine whether we would need to perform transformations on it. The histogram on the left indicates that price is right-skewed. This highlights a lack of normality. Thus, we first transformed the response variable using a log transformation in order to correct the violation of normality. This transformation can be seen on the graph on the right.



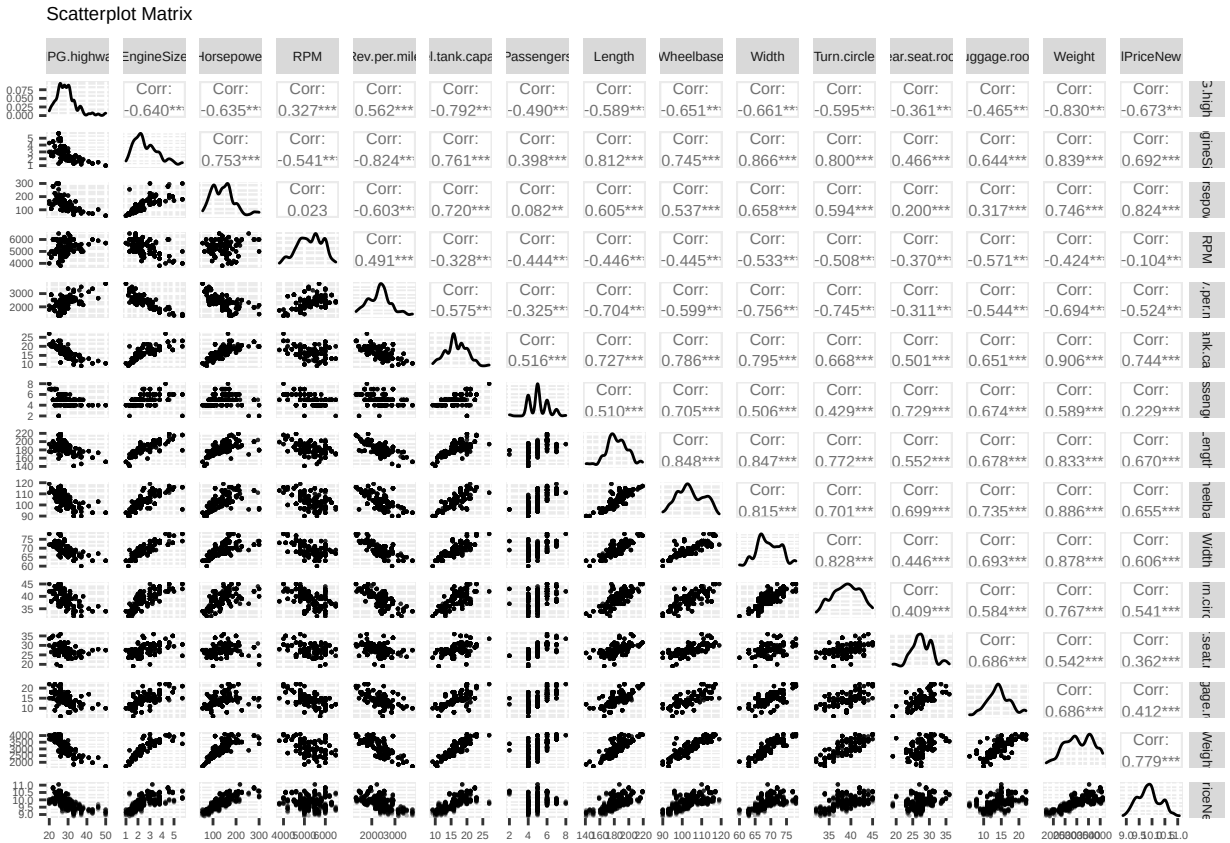
The Numerical Predictors

After transforming Price, we then investigated the relationship between the transformed Price with all the other numerical variables. The table below shows the correlation coefficient between the numerical variables with the transformed Price ordered from lowest to highest:

Table 3: Correlation with Price

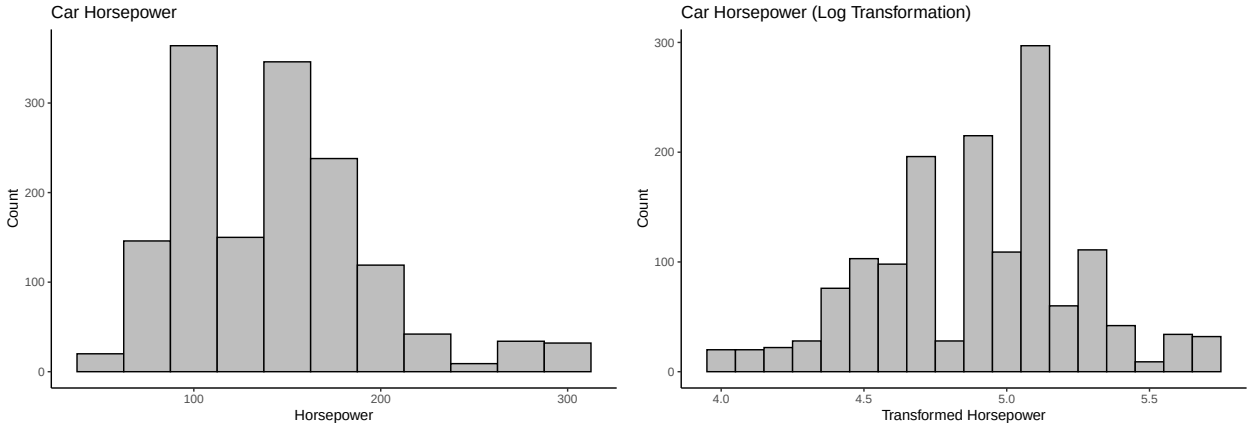
Variable	Correlation
RPM	-0.10
Passengers	0.23
Rear.seat.room	0.36
Luggage.room	0.41
Rev.per.mile	-0.52
Turn.circle	0.54
Width	0.61
Wheelbase	0.66
MPG.highway	-0.67
Length	0.67
EngineSize	0.69
Fuel.tank.capacity	0.74
Weight	0.78
Horsepower	0.82

Additionally, below is a scatterplot matrix between the numerical variables:



Based on the table and correlation matrix above, Horsepower is a clear predictor for the model. While "Weight", "Fuel.tank.capacity", "EngineSize", and "Length" have the next highest correlations with Price, they all also have a high correlation with Horsepower, creating a concern for multicollinearity. We may also consider adding "MPG.highway" and "Wheelbase".

One preliminary issue to consider with Horsepower is the issue of normality. The histogram on the left below indicates that the predictor is positively skewed. Thus, we will use a log transformation to fix this problem, as shown with the graph on the right.



Below is a first attempt at a model using just the numeric predictors considered above.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.163	0.2724	26.29	2.016e-125	* * *
MPG.highway	-0.003722	0.001955	-1.904	0.05708	
EngineSize	0.03334	0.01448	2.303	0.02141	*
lHorsepower	0.7974	0.04908	16.25	9.425e-55	* * *
RPM	-7.236e-05	1.723e-05	-4.201	2.821e-05	* * *
Rev.per.mile	0.0001165	1.932e-05	6.033	2.034e-09	* * *
Fuel.tank.capacity	0.01133	0.003866	2.931	0.003427	* *
Passengers	-0.0921	0.009592	-9.602	3.182e-21	* * *
Length	0.005176	0.0007651	6.765	1.916e-11	* * *
Wheelbase	0.01402	0.002145	6.539	8.511e-11	* * *
Width	-0.05254	0.003764	-13.96	1.029e-41	* * *
Turn.circle	-0.01315	0.00294	-4.473	8.286e-06	* * *
Rear.seat.room	0.007388	0.002837	2.604	0.009303	* *
Luggage.room	0.001828	0.002753	0.664	0.5068	
Weight	0.0002362	3.825e-05	6.174	8.586e-10	* * *

Table 5: Model with Numeric Predictors

Observations	Residual Std. Error	R^2	Adjusted R^2
1500	0.1853	0.8167	0.815

The model indicates that perhaps we should remove MPG.highway, EngineSize, Rear.seat.room, and Luggage.room. Below is a summary of the reduced model:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	6.979	0.2582	27.03	3.111e-131	* * *
lHorsepower	0.8739	0.04244	20.59	4.266e-83	* * *
RPM	-0.0001045	1.431e-05	-7.306	4.453e-13	* * *
Rev.per.mile	0.0001045	1.834e-05	5.698	1.458e-08	* * *
Fuel.tank.capacity	0.01553	0.003644	4.26	2.168e-05	* * *
Passengers	-0.07676	0.008215	-9.343	3.294e-20	* * *
Length	0.005539	0.000751	7.375	2.719e-13	* * *
Wheelbase	0.01577	0.002001	7.881	6.21e-15	* * *
Width	-0.05525	0.003497	-15.8	4.333e-52	* * *
Turn.circle	-0.01286	0.002943	-4.368	1.338e-05	* * *
Weight	0.0002357	3.509e-05	6.716	2.646e-11	* * *

Table 7: Model with Numeric Predictors 2

Observations	Residual Std. Error	R^2	Adjusted R^2
1500	0.1862	0.8145	0.8132

Following this, we then did multiple rounds of checking the VIF and summary models and ended up with the model below:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.062	0.1917	21.19	2.39e-87	* * *
lHorsepower	0.9264	0.03785	24.47	1.843e-111	* * *
RPM	-7.565e-05	1.399e-05	-5.409	7.366e-08	* * *
Rev.per.mile	0.0001457	1.976e-05	7.376	2.7e-13	* * *
Fuel.tank.capacity	0.01673	0.003426	4.883	1.158e-06	* * *
Passengers	-0.05981	0.008563	-6.984	4.293e-12	* * *
Length	0.003018	0.000798	3.782	0.0001617	* * *
Wheelbase	0.01692	0.001978	8.555	2.883e-17	* * *
Turn.circle	-0.02468	0.003077	-8.022	2.089e-15	* * *

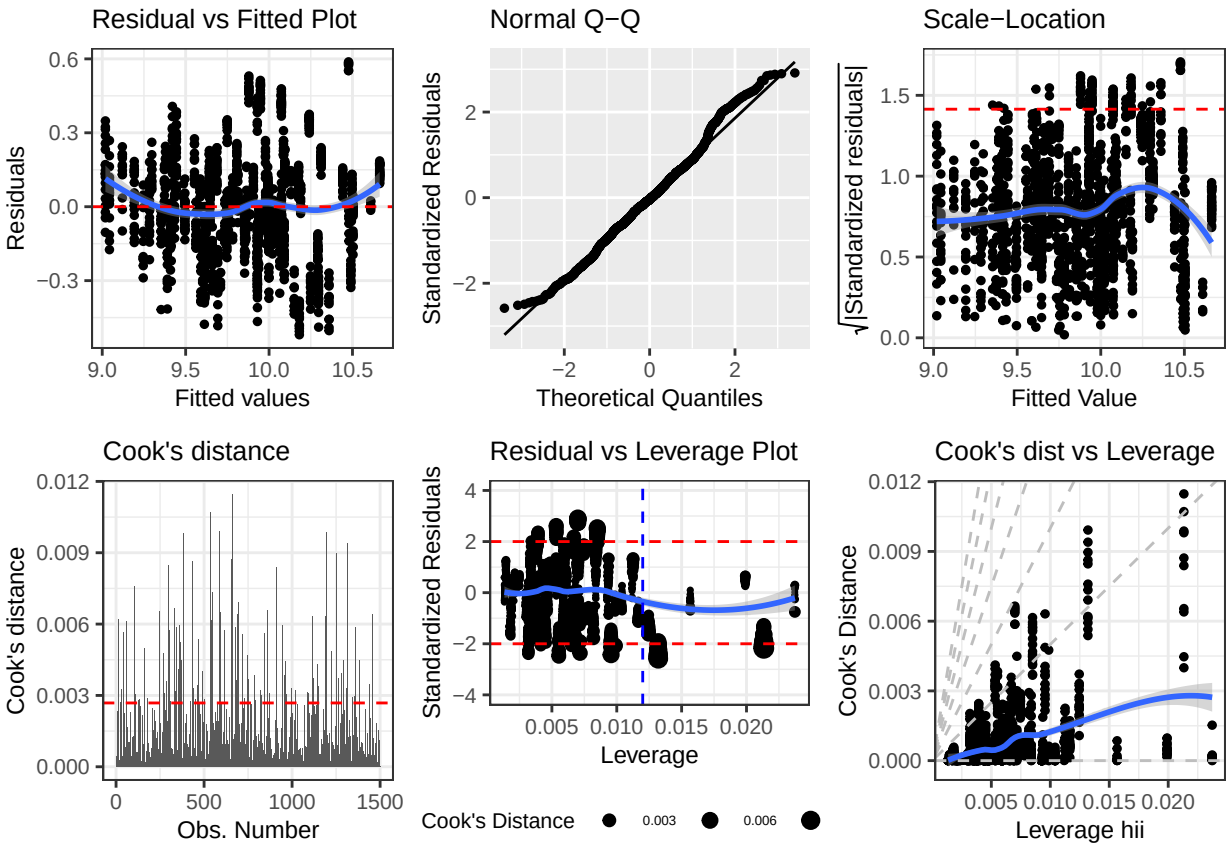
Table 9: Model with Numeric Predictors 3

Observations	Residual Std. Error	R^2	Adjusted R^2
1500	0.2024	0.7804	0.7792

Below are the diagnostic plots for the model. There appear to be a few violations, however we will work on fixing them later as we continue to develop the model.

Note: The code for the diagPlot was taken from the Professor's Chapter 5 notes (Almohalwas 2020, 2).

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

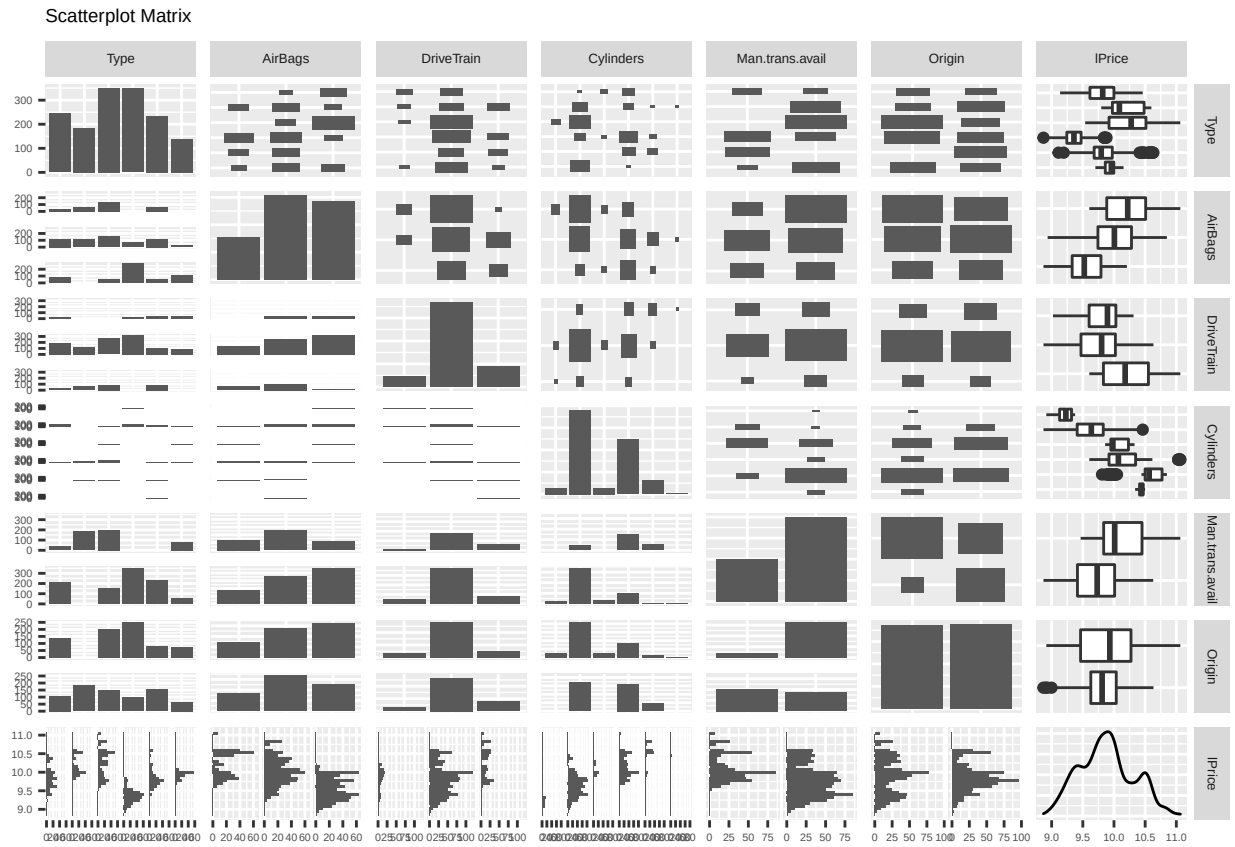


The Categorical Predictors

Below is a table of a list of categorical predictors along with the number of levels they each have. This is followed by a scatterplot matrix of price along with most of the categorical predictors. The scatterplot matrix suggests that Type, AirBags, Man.trans.avail, and Cylinders would be good predictors of price. However, to avoid having too many predictors, the next section will first show the categories being regrouped. Additionally, we are also going to transform some of the variables with a large number of categories into new predictors. We will also be taking some of the numerical predictors from earlier and transforming them into categorical predictors.

Table 10: Categorical Variables and Levels

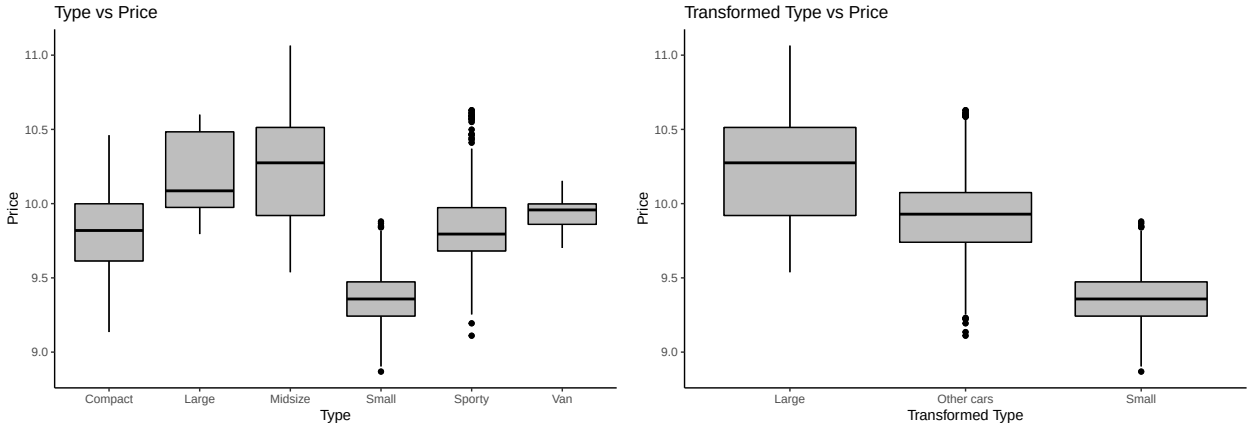
Variable	Number of Categories
Man.trans.avail	2
Origin	2
AirBags	3
DriveTrain	3
Type	6
Cylinders	6
Manufacturer	32
Model	93
Make	93



Transforming Categorical Predictors

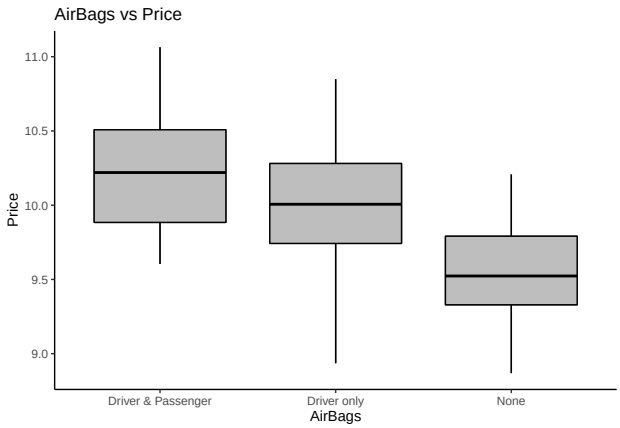
Type

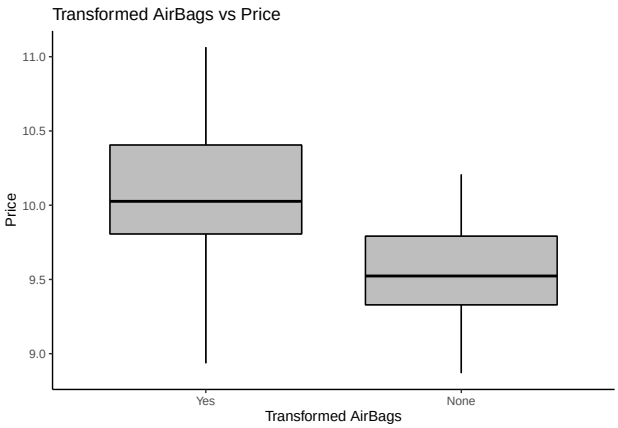
The graph on the left shows the initial categories of Type compared to the transformed price. The graph on the right shows "Type" after the categories are regrouped into "Small", "Other Cars", and "Large".



AirBags

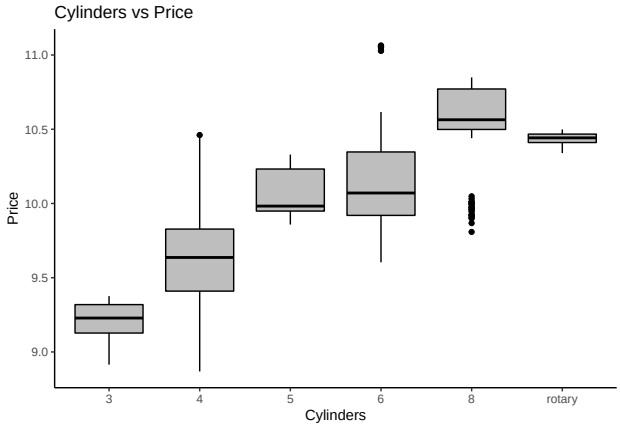
The graph on the left shows the initial categories of AirBags compared to the transformed price. The graph on the right shows "AirBags" after the categories are regrouped into "None" and "Yes".

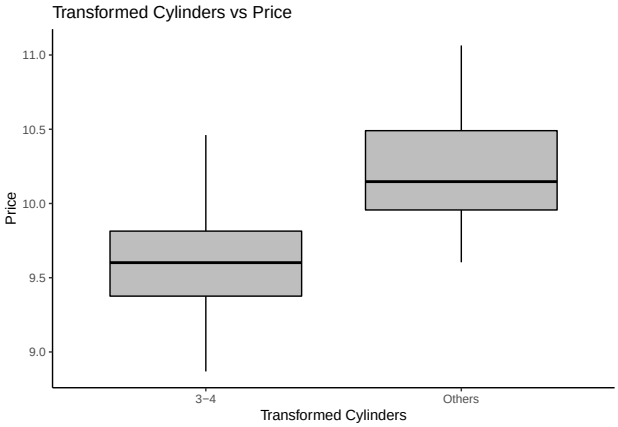




Cylinders

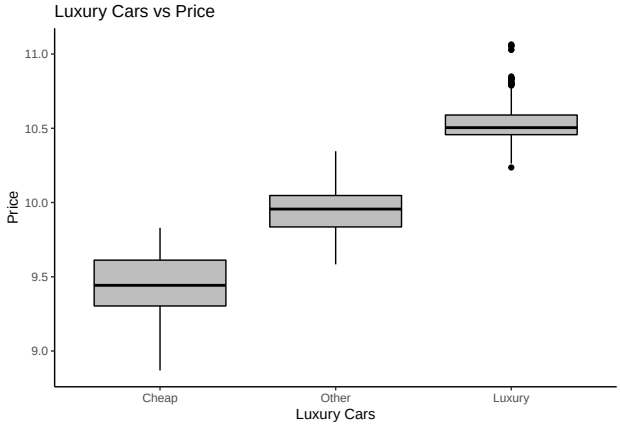
The graph on the left shows the initial categories of Cylinders compared to the transformed price. The graph on the right shows "Cylinders" after the categories are regrouped into "3-4" and "Others".





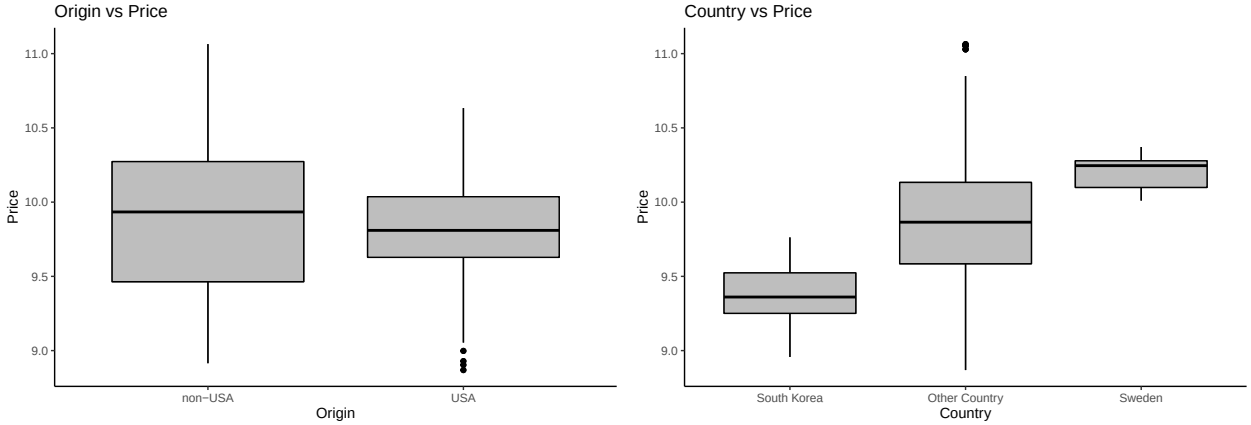
Luxury Cars

Considering the large number of categories present in "Model", "Make", and "Manufacturer", we created a new categorical variable based on the Price of "Make". This variable is named "Luxury Cars", and is split into three categories: "Luxury", which is anything greater than $30,000; "Cheap", which is anything less than $17,000; and "Other", which is anything in between.



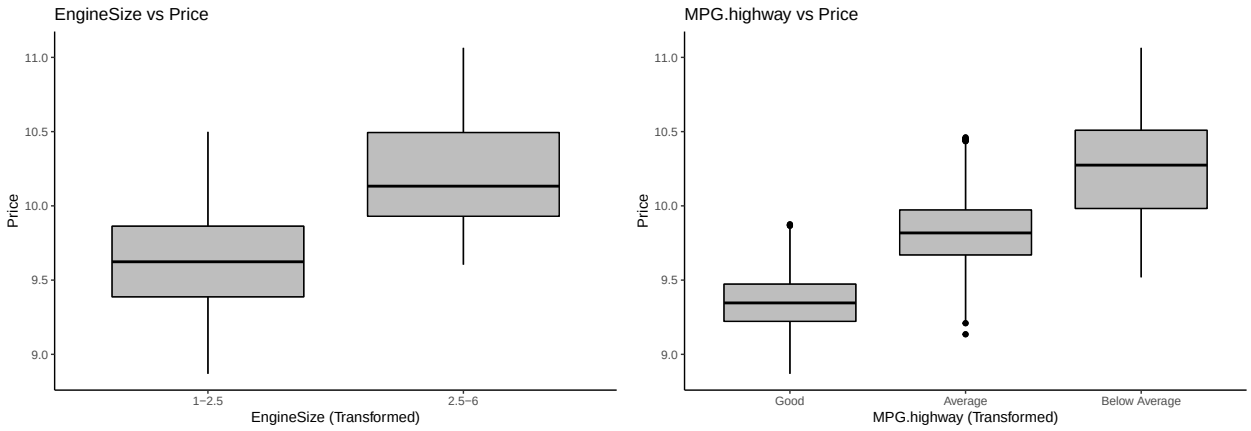
Country

While the dataset itself comes with an "Origin" variable, we did not think that it contained enough categories to truly be able to predict Price. As a result, we came up with a new variable called "Country". A comparison of these two variables with Price can be seen below:



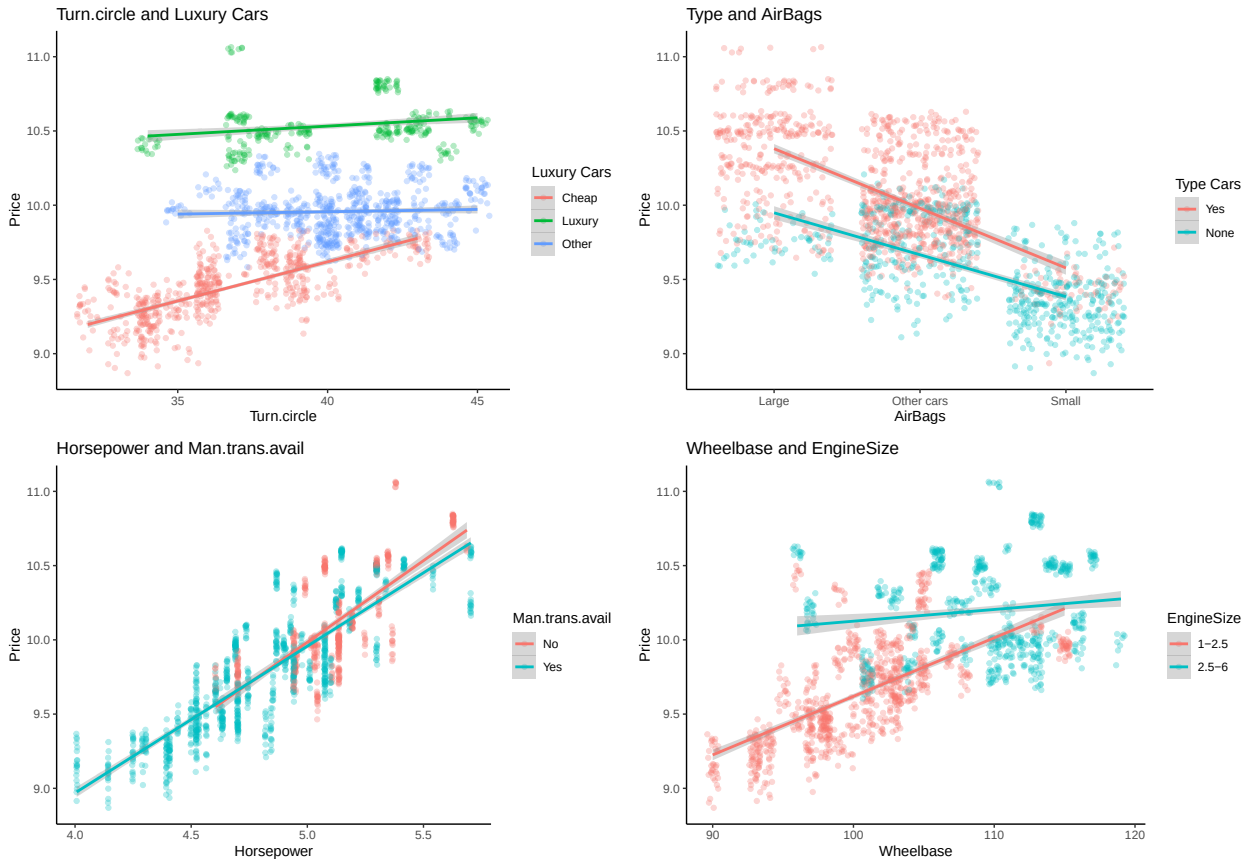
Transforming Numerical Predictors to Categorical Predictors

Both EngineSize and MPG.highway were re-categorized from numerical to categorical predictors:



Exploring Interactions

Below is a list of the most prominent interactions that were found along with their plots



Building the Model

Attempt 1

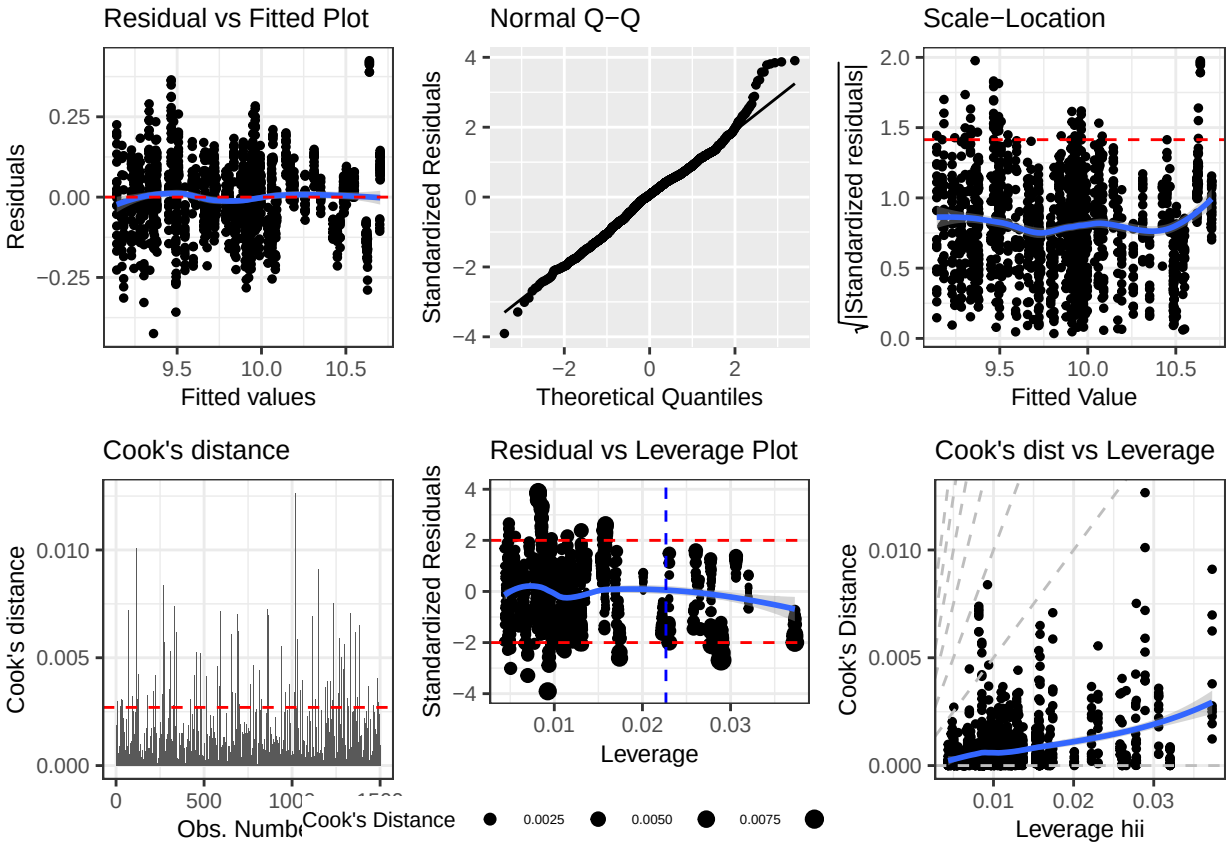
Below was a first attempt at producing a final model created by combining the most prominent categorical variables alongside the model that we created with just the numerical predictors earlier. Based on the most prominent interactions, we removed a few of the numerical predictors from the model first. While the diagnostic plots did not show serious violations, we can see that Passengers and Length are both not statistically significant. As a result, we will remove them.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.965	0.1085	73.44	0	* * *
lHorsepower	0.3547	0.0178	19.92	1.88e-78	* * *
Passengers	-0.000762	0.005094	-0.1496	0.8811	
Length	0.0001979	0.0004564	0.4337	0.6646	
Wheelbase	0.004323	0.001126	3.841	0.0001279	* * *
Turn.circle	-0.007894	0.00172	-4.59	4.797e-06	* * *
LuxuryLuxury	0.5643	0.01455	38.78	9.615e-228	* * *
LuxuryOther	0.1804	0.01006	17.92	3.779e-65	* * *
type_newOther cars	-0.09718	0.007726	-12.58	1.505e-34	* * *
type_newSmall	-0.209	0.01321	-15.82	3.412e-52	* * *
new_AirBagsNone	-0.08665	0.008018	-10.81	2.947e-26	* * *
countrySouth Korea	-0.1052	0.01652	-6.372	2.491e-10	* * *
countrySweden	0.09635	0.01938	4.972	7.401e-07	* * *

	Estimate	Std. Error	t value	Pr(>|t|)	
MPG_catAverage	-0.06307	0.01013	-6.228	6.152e-10	* * *
MPG_catGood	-0.1097	0.01515	-7.241	7.102e-13	* * *
EngineSize_cat2.5-6	-0.0905	0.01413	-6.404	2.03e-10	* * *
cylinders_newOthers	0.07572	0.01454	5.208	2.177e-07	* * *

Table 12: Model 1

Observations	Residual Std. Error	R^2	Adjusted R^2
1500	0.1092	0.9365	0.9358



Attempt 2

Removing Passengers and Length, the variable Cylinders seems to cause an issue of multicollinearity. Thus, we will now remove that. We also ran a boxcox transformation to try and see whether any additional transformations needed to be made. In this case, we decided to transform the numerical predictor "Wheelbase" by taking its inverse.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.961	0.1063	74.86	0	* * *
lHorsepower	0.3568	0.01646	21.68	8.893e-91	* * *
Wheelbase	0.004486	0.0007055	6.358	2.708e-10	* * *

	Estimate	Std. Error	t value	Pr(>|t|)	
Turn.circle	-0.007646	0.001611	-4.745	2.285e-06	* * *
LuxuryLuxury	0.5644	0.01452	38.87	1.355e-228	* * *
LuxuryOther	0.1793	0.009752	18.39	3.534e-68	* * *
type_newOther cars	-0.09769	0.007625	-12.81	1.026e-35	* * *
type_newSmall	-0.2098	0.01296	-16.19	2.068e-54	* * *
new_AirBagsNone	-0.08712	0.007812	-11.15	8.523e-28	* * *
countrySouth Korea	-0.1053	0.01643	-6.413	1.913e-10	* * *
countrySweden	0.09699	0.01898	5.109	3.66e-07	* * *
MPG_catAverage	-0.06172	0.009545	-6.465	1.367e-10	* * *
MPG_catGood	-0.1083	0.01465	-7.398	2.3e-13	* * *
EngineSize_cat2.5-6	-0.09046	0.01376	-6.573	6.815e-11	* * *
cylinders_newOthers	0.07613	0.01433	5.311	1.257e-07	* * *

Table 14: Model 2

Observations	Residual Std. Error	R^2	Adjusted R^2
1500	0.1091	0.9364	0.9358

Table 15: VIF Model

	GVIF	Df	GVIF^(1/(2*Df))
lHorsepower	4.3	1	2.074
Wheelbase	2.983	1	1.727
Turn.circle	3.34	1	1.828
Luxury	4.941	2	1.491
type_new	3.274	2	1.345
new_AirBags	1.817	1	1.348
country	1.404	2	1.088
MPG_cat	5.885	2	1.558
EngineSize_cat	5.852	1	2.419
cylinders_new	6.419	1	2.534

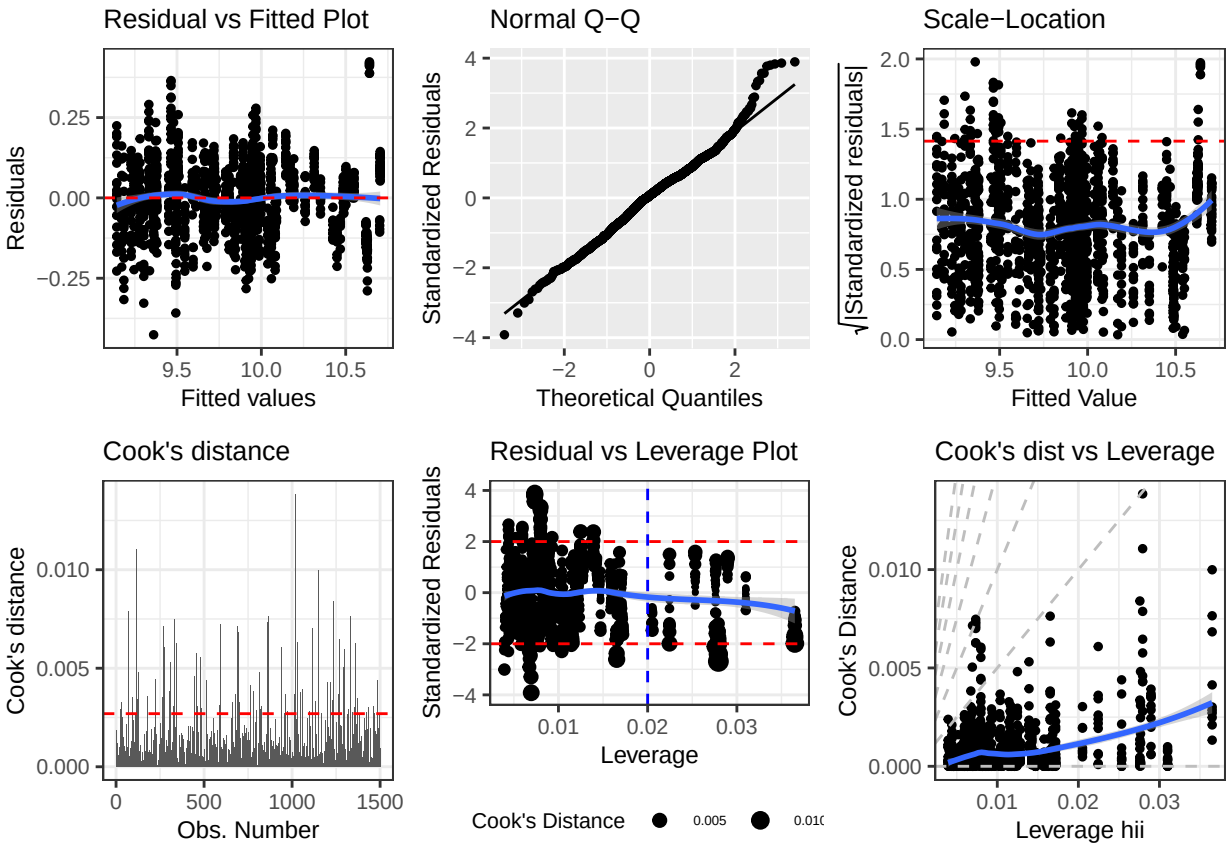


Table 16: Boxcox Transformation

	Est Power	Rounded Pwr	Wald Lwr Bnd	Wald Upr Bnd
lPriceNew	-4.3487099	-4.35	-5.0548485	-3.6425713
Luxury	0.0637545	0.00	-0.0897484	0.2172574
Turn.circle	0.1402411	0.00	-0.3725160	0.6529982
type_new	1.3940472	1.39	1.2642169	1.5238775
new_AirBags	-2.0893421	-2.00	-2.3554569	-1.8232273
lHorsepower	0.7644021	1.00	0.3519286	1.1768756
country	-18.2482520	-18.25	-19.1718752	-17.3246289
MPG_cat	0.5487797	0.50	0.4202244	0.6773350
EngineSize_cat	-1.2024773	-1.00	-1.4597787	-0.9451758
Wheelbase	-1.3063072	-1.00	-1.9553950	-0.6572193
cylinders_new	-0.7770153	-1.00	-1.0317639	-0.5222668

Attempt 3

Below is the model after transforming Wheelbase and removing Cylinders. As we can see, there doesn't seem to be an issue of multicollinearity. Furthermore, the assumptions of linearity, constant variance, and normality all seem to be met.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.592	0.104	82.6	0	* * *
LuxuryLuxury	0.5766	0.01403	41.11	2.195e-247	* * *

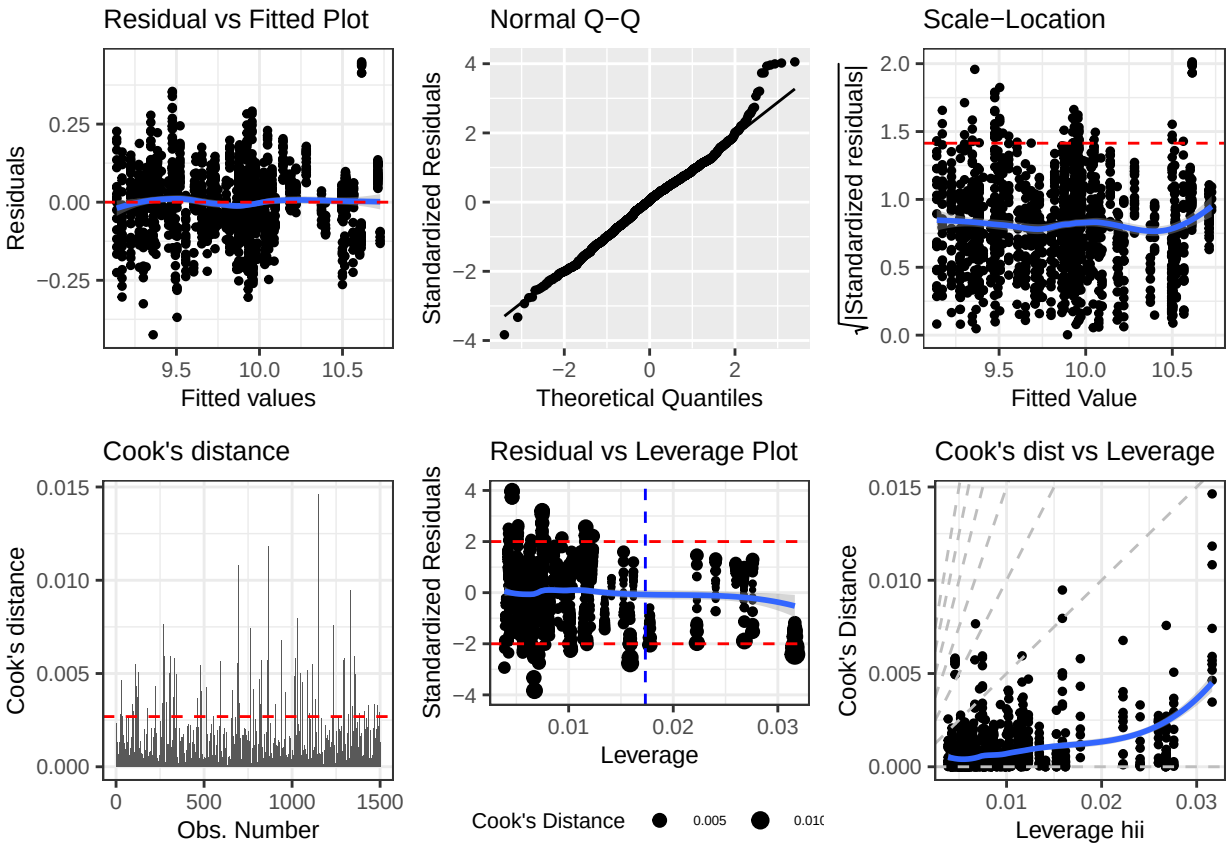
	Estimate	Std. Error	t value	Pr(>|t|)	
LuxuryOther	0.1869	0.009806	19.06	1.218e-72	* * *
type_newOther cars	-0.1041	0.007699	-13.52	2.232e-39	* * *
type_newSmall	-0.2039	0.01314	-15.51	2.093e-50	* * *
new_AirBagsNone	-0.07374	0.007708	-9.566	4.421e-21	* * *
lHorsepower	0.3529	0.01634	21.6	3.336e-90	* * *
countrySouth Korea	-0.1088	0.01671	-6.508	1.04e-10	* * *
countrySweden	0.1335	0.01855	7.2	9.493e-13	* * *
MPG_catAverage	-0.08327	0.008937	-9.317	4.171e-20	* * *
MPG_catGood	-0.1175	0.01446	-8.124	9.355e-16	* * *
EngineSize_cat2.5-6	-0.05396	0.009777	-5.519	4.014e-08	* * *
tWheelbase	-43.8	7.012	-6.246	5.468e-10	* * *

Table 18: Model 3

Observations	Residual Std. Error	R^2	Adjusted R^2
1500	0.1111	0.9341	0.9335

Table 19: VIF Model 2

	GVIF	Df	GVIFˆ(1/(2*Df))
Luxury	4.357	2	1.445
type_new	3.099	2	1.327
new_AirBags	1.708	1	1.307
lHorsepower	4.092	1	2.023
country	1.292	2	1.066
MPG_cat	4.89	2	1.487
EngineSize_cat	2.851	1	1.689
tWheelbase	2.45	1	1.565



Attempt 4: Adding Interaction Terms

Next, we added interaction terms to the existing model based on the interactions we discovered in the previous sections:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.907	0.2215	35.7	7.77e-202	* * *
LuxuryLuxury	1.035	0.1444	7.166	1.216e-12	* * *
LuxuryOther	1.116	0.1401	7.964	3.293e-15	* * *
Turn.circle	0.002878	0.002819	1.021	0.3074	
type_newOther cars	-0.1214	0.008173	-14.86	1.219e-46	* * *
type_newSmall	-0.251	0.01723	-14.56	5.244e-45	* * *
new_AirBagsNone	-0.1537	0.01663	-9.243	8.096e-20	* * *
lHorsepower	0.5743	0.02529	22.71	3.343e-98	* * *
Man.trans.availYes	1.68	0.1426	11.78	1.091e-30	* * *
countrySouth Korea	-0.07978	0.01551	-5.144	3.042e-07	* * *
countrySweden	0.1058	0.01709	6.194	7.591e-10	* * *
MPG_catAverage	-0.08936	0.008898	-10.04	5.263e-23	* * *
MPG_catGood	-0.1174	0.0135	-8.696	8.962e-18	* * *
EngineSize_cat2.5-6	-1.073	0.1537	-6.983	4.352e-12	* * *
tWheelbase	-96.27	11.09	-8.678	1.046e-17	* * *
LuxuryLuxury:Turn.circle	-0.01279	0.003652	-3.502	0.0004764	* * *
LuxuryOther:Turn.circle	-0.02366	0.003536	-6.693	3.102e-11	* * *
type_newOther cars:new_AirBagsNone	0.06032	0.01815	3.323	0.0009118	* * *

	Estimate	Std. Error	t value	Pr(>|t|)	
type_newSmall:new_AirBagsNone	0.09641	0.02192	4.398	1.169e-05	* * *
lHorsepower:Man.trans.availYes	-0.338	0.02845	-11.88	3.755e-31	* * *
EngineSize_cat2.5-6:tWheelbase	111.1	16.13	6.889	8.306e-12	* * *

Table 21: Model 4

Observations	Residual Std. Error	R^2	Adjusted R^2
1500	0.1009	0.9458	0.9451

Checking BIC

To ensure that we are using the most appropriate model, we are also comparing the BICs of all the models we've created above. As we can see, the most recent attempt has the highest BIC.

Table 22: Model BICs

Model.Attempt	BIC
1	-6536.738
2	-6551.155
3	-6510.932
4	-6747.130

Results and Discussion

Final Model

Based on the attempts above, our final model is given by:

$$\hat{Price} = 7.907 + 1.035(LuxuryLuxury) + 1.116(LuxuryOther)$$
$$+0.003(Turn.circle) - 0.121(type.newOthercars) - 0.251(type.newSmall)$$
$$-0.154(new.AirBagsNone) + 0.574(lHorsepower) + 1.680(Man.trans.availYes)$$
$$-0.080(countrySouthKorea) + 0.106(countrySweden) - 0.089(MPG.catAverage)$$
$$-0.117(MPG.catGood) - 1.073(EngineSize.cat2.5-6) - 96.267(tWheelbase)$$
$$-0.013(LuxuryLuxury * Turn.circle) - 0.024(LuxuryOther * Turn.circle)$$
$$+0.060(type.newOthercars * new.AirBagsNone) + 0.096(type.newSmall * new.AirBagsNone)$$
$$-0.338(lHorsepower * Man.trans.availYes) + 111.129(EngineSize.cat2.5-6 * tWheelbase)$$

In order to ensure the model is accurate, we split the initial training set into a separate training and testing dataset based on setting the seed to "123456". The results as shown below indicate that the model is relatively accurate. This can be seen by how most of the predictors are statistically significant, how the R^2 values are relatively similar, and how the correlation coefficient between the actual price and predicted price is 0.972.

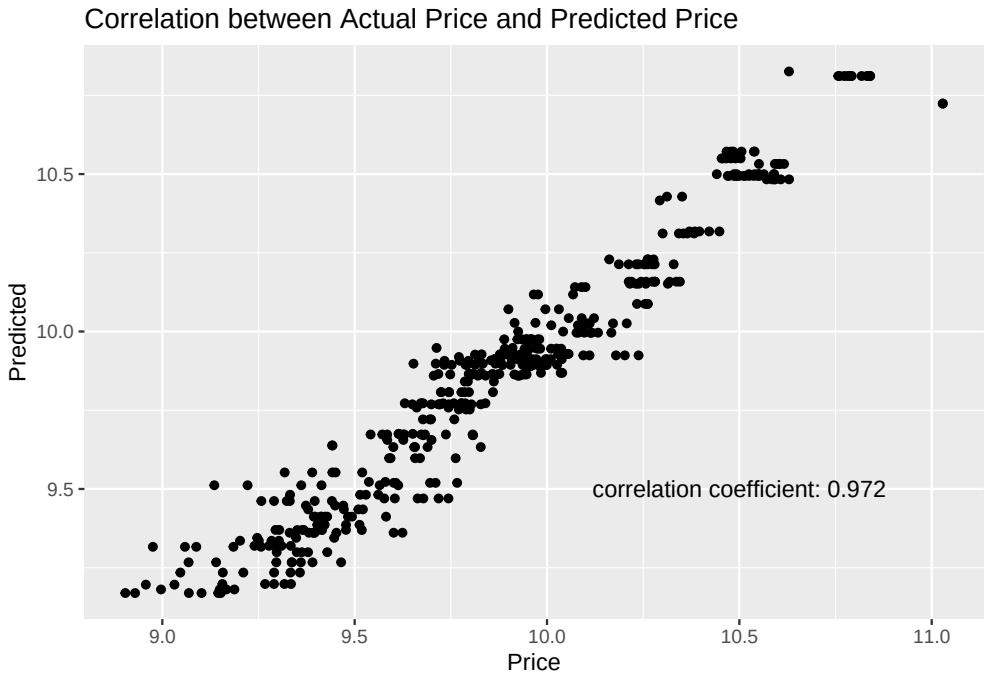


Table 23: Training Model (Adj R^2: 0.9451)

Predictor	Estimate	SE	t-statistic	p-value
(Intercept)	7.96	0.266	29.96	<0.001
LuxuryLuxury	1.10	0.175	6.31	<0.001
LuxuryOther	1.18	0.172	6.85	<0.001
Turn.circle	0.00	0.003	1.17	0.241
type_newOther cars	-0.12	0.010	-12.29	<0.001
type_newSmall	-0.25	0.021	-11.79	<0.001
new_AirBagsNone	-0.15	0.019	-7.96	<0.001
lHorsepower	0.57	0.031	18.63	<0.001
Man.trans.availYes	1.72	0.172	10.02	<0.001
countrySouth Korea	-0.09	0.018	-4.79	<0.001
countrySweden	0.10	0.021	4.78	<0.001
MPG_catAverage	-0.09	0.010	-8.15	<0.001
MPG_catGood	-0.11	0.016	-7.07	<0.001
EngineSize_cat2.5-6	-1.09	0.183	-5.97	<0.001
tWheelbase	-104.83	13.131	-7.98	<0.001
LuxuryLuxury:Turn.circle	-0.01	0.004	-3.25	0.001
LuxuryOther:Turn.circle	-0.03	0.004	-5.83	<0.001
type_newOther cars:new_AirBagsNone	0.06	0.021	2.78	0.006
type_newSmall:new_AirBagsNone	0.11	0.026	4.11	<0.001
lHorsepower:Man.trans.availYes	-0.35	0.034	-10.11	<0.001
EngineSize_cat2.5-6:tWheelbase	112.70	19.226	5.86	<0.001

Table 24: Testing Model (Adj R^2: 0.9449)

Predictor	Estimate	SE	t-statistic	p-value
(Intercept)	7.76	0.415	18.71	<0.001

Predictor	Estimate	SE	t-statistic	p-value
LuxuryLuxury	1.01	0.273	3.68	<0.001
LuxuryOther	1.00	0.247	4.04	<0.001
Turn.circle	0.00	0.005	0.15	0.883
type_newOther cars	-0.12	0.016	-7.83	<0.001
type_newSmall	-0.26	0.031	-8.39	<0.001
new_AirBagsNone	-0.15	0.034	-4.48	<0.001
lHorsepower	0.58	0.046	12.47	<0.001
Man.trans.availYes	1.55	0.263	5.87	<0.001
countrySouth Korea	-0.07	0.030	-2.18	0.030
countrySweden	0.11	0.030	3.71	<0.001
MPG_catAverage	-0.10	0.018	-5.56	<0.001
MPG_catGood	-0.13	0.025	-5.12	<0.001
EngineSize_cat2.5-6	-0.94	0.288	-3.26	0.001
tWheelbase	-73.01	21.152	-3.45	<0.001
LuxuryLuxury:Turn.circle	-0.01	0.007	-1.76	0.078
LuxuryOther:Turn.circle	-0.02	0.006	-3.30	0.001
type_newOther cars:new_AirBagsNone	0.06	0.037	1.67	0.096
type_newSmall:new_AirBagsNone	0.07	0.042	1.71	0.087
lHorsepower:Man.trans.availYes	-0.31	0.052	-5.95	<0.001
EngineSize_cat2.5-6:tWheelbase	97.81	30.248	3.23	0.001

Additional Diagnostic Plots

Leverage Points

There appears to be only one bad leverage point:

Leverage	Outliers	
	No	Yes
No	1389	75
Yes	35	1

Final Diagnostics

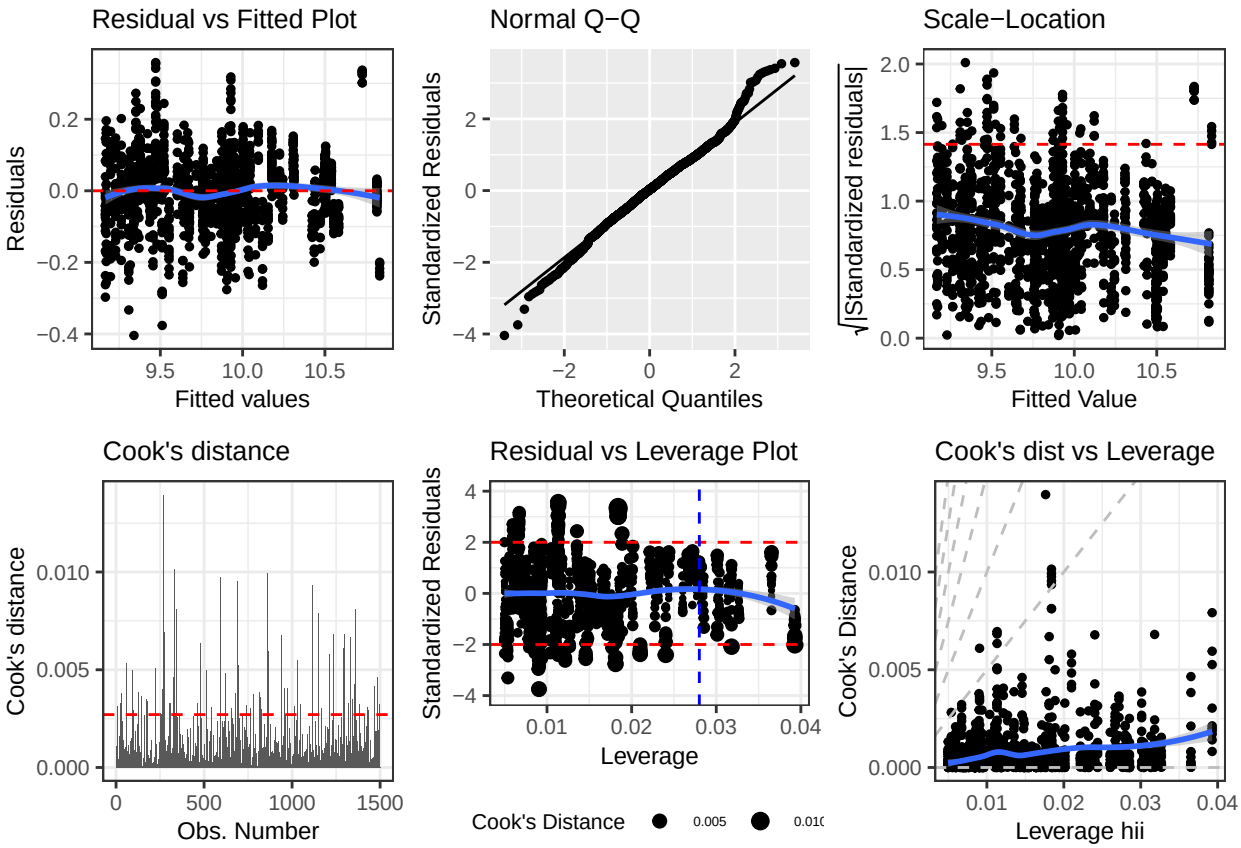
Considering the accuracy of the training model as shown above, we will continue to use this as the final model. Below shows a summary of the final model as well as the diagnostic plots, marginal density plot, and leverage plots. The diagnostic plots indicate that the assumptions are met, and that the predictors are relatively accurate in predicting the price of a car. One issue we may consider is how the predictor "Turn.circle" is not statistically significant.

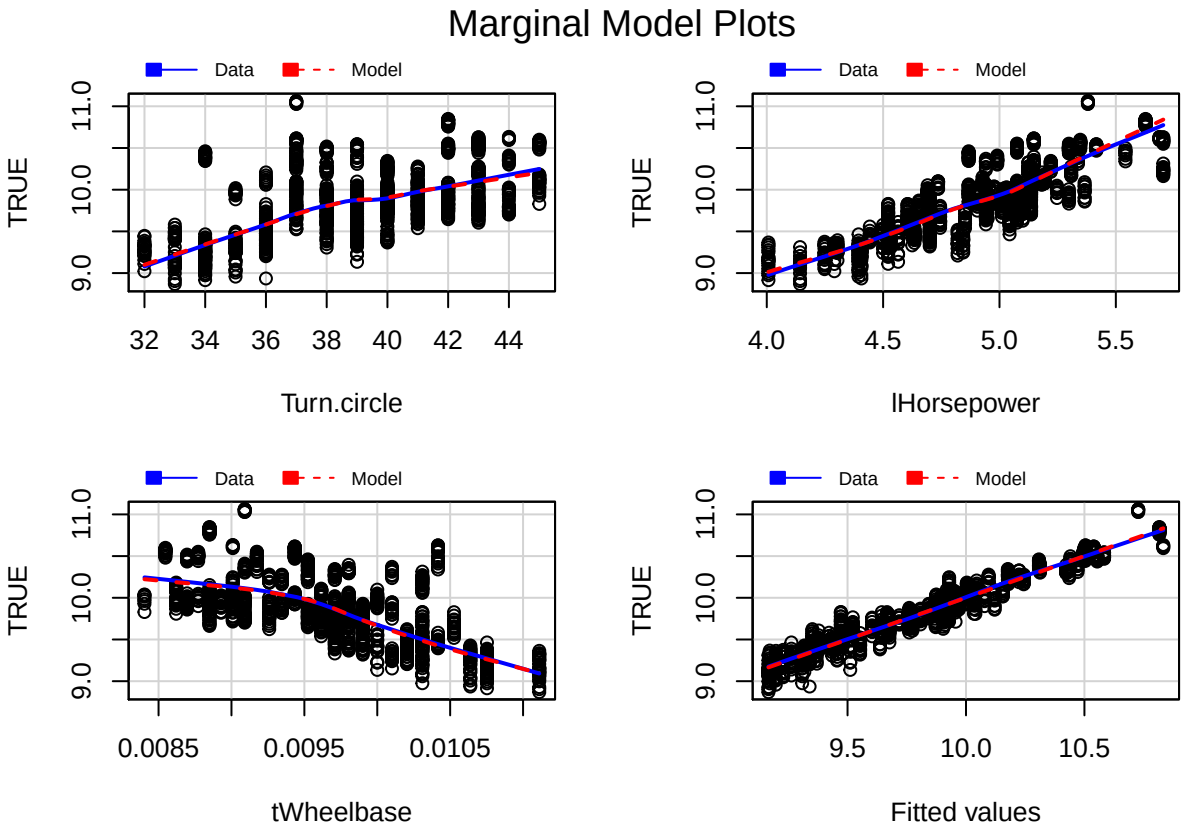
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.907	0.2215	35.7	7.77e-202	* * *
LuxuryLuxury	1.035	0.1444	7.166	1.216e-12	* * *
LuxuryOther	1.116	0.1401	7.964	3.293e-15	* * *
Turn.circle	0.002878	0.002819	1.021	0.3074	
type_newOther cars	-0.1214	0.008173	-14.86	1.219e-46	* * *
type_newSmall	-0.251	0.01723	-14.56	5.244e-45	* * *
new_AirBagsNone	-0.1537	0.01663	-9.243	8.096e-20	* * *
lHorsepower	0.5743	0.02529	22.71	3.343e-98	* * *
Man.trans.availYes	1.68	0.1426	11.78	1.091e-30	* * *

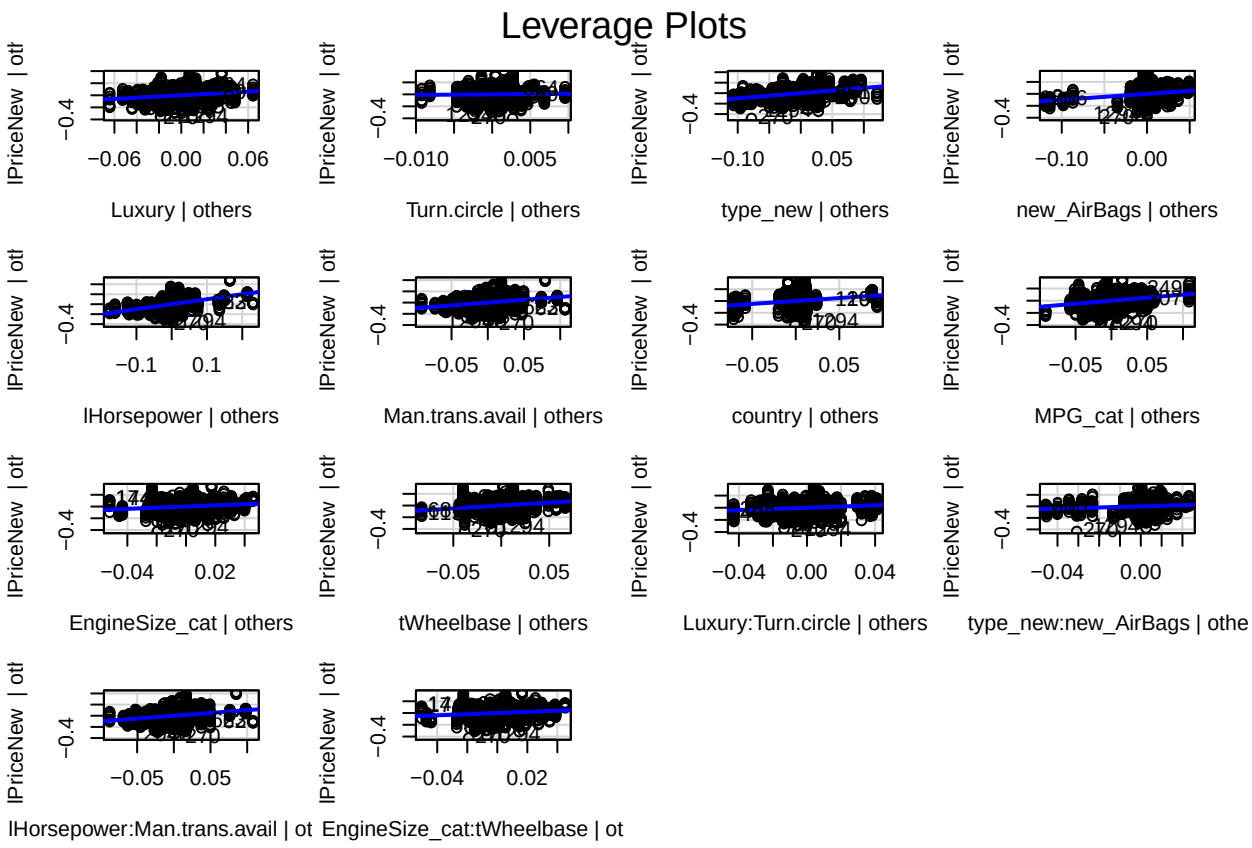
	Estimate	Std. Error	t value	Pr(>|t|)	
countrySouth Korea	-0.07978	0.01551	-5.144	3.042e-07	* * *
countrySweden	0.1058	0.01709	6.194	7.591e-10	* * *
MPG_catAverage	-0.08936	0.008898	-10.04	5.263e-23	* * *
MPG_catGood	-0.1174	0.0135	-8.696	8.962e-18	* * *
EngineSize_cat2.5-6	-1.073	0.1537	-6.983	4.352e-12	* * *
tWheelbase	-96.27	11.09	-8.678	1.046e-17	* * *
LuxuryLuxury:Turn.circle	-0.01279	0.003652	-3.502	0.0004764	* * *
LuxuryOther:Turn.circle	-0.02366	0.003536	-6.693	3.102e-11	* * *
type_newOther	0.06032	0.01815	3.323	0.0009118	* * *
cars:new_AirBagsNone					
type_newSmall:new_AirBagsNone	0.09641	0.02192	4.398	1.169e-05	* * *
lHorsepower:Man.trans.availYes	-0.338	0.02845	-11.88	3.755e-31	* * *
EngineSize_cat2.5-6:tWheelbase	111.1	16.13	6.889	8.306e-12	* * *

Table 27: Final Model

Observations	Residual Std. Error	R^2	Adjusted R^2
1500	0.1009	0.9458	0.9451




Marginal Model Plots
Data
Model
TRUE
9.0
10.0
11.0
32
34
36
38
40
42
44
Turn.circle
4.0
4.5
5.0
5.5
lHorsepower
0.0085
0.0095
0.0105
tWheelbase
9.5
10.0
10.5
Fitted values



Limitations and Conclusion

In regards to the diagnostic plots shown above, there looks to be a slight violation to the constant variance assumption, with a slight decreasing trend as the fitted value increases.

In general, perhaps the biggest limitation is the inability of this model to accurately predict the price of the testing dataset. Despite my own training and testing splitting indicating that these were good predictors, my final Kaggle R^2 was significantly lower compared to what was achieved through the training dataset. This potential issue was not a complete surprise, as my own testing dataset from earlier did indicate generally more predictors that were not statistically significant compared to the training dataset.

Another problem with this model is the fact that there are 20 predictors. Considering the number of people that got a significantly higher ranking compared to me, there are definitely other predictors that I could consider in the future.

References

Almohalwas, Akram. 2020. *Chapter 5 Updated Winter 2020.*

———. 2021. *STAT 101 a Winter 2021 Kaggle Competition: Predicting Car Prices.*

Kaggle. 2021. "CarsTrain.csv." https://www.kaggle.com/c/stat101a-us-car-prices/data.

Sheather, Simon J. 2009. *A Modern Approach to Regression with R.* New York: Springer.