Vehicle Price Modeling and Prediction

Alex Gilbert, Nicole Norelli, and Mingyang Nick YU

Introduction

Variables which contribute to vehicle pricing are of interest to both consumers and retailers. While it might seem straightforward to build a simple model to both explain and predict vehicle price given sufficient data, different models are best suited to each of these tasks. This analysis will explore a data set of automobile information with the goal of better understanding and predicting the retail price of a vehicle. We will use regression to build an interpretable model, and we will compare a more complex regression model with a k-nearest neighbors (KNN) regression model to obtain the most accurate pricing predictions.

Data Description

The data set contains 16 variables describing 11,914 vehicles. It was provided for the purposes of this analysis, and its origin is unclear. Original variable names, types, and descriptions are available in Appendix Table1.

Exploratory Data Analysis (EDA)

Missing Data

The variables with missing data were: EngineHP (69), EngineCylinders (30), NumberOfDoors (6), EngineFuelType (3), and MarketCategory (3742). Through research into basic data regarding car models, we were able to supply the correct values for EngineFuelType and NumberOfDoors. All vehicles without EngineCylinders data were either electric vehicles or Mazda RX models with a rotary engine. As neither of these car types have cylinders, we corrected the missing values to zero. For EngineHP, we researched the 11 models missing data and imputed the mean horsepower for each model.

The MarketCategory variable contained entries with multiple categories listed for single vehicles. To split MarketCategory into a usable form, we created separate variables for each category value (such as Exotic, Luxury, Hybrid, etc.) and assigned a yes or no value based on the original MarketCategory entry. Next, we corrected 18 values in the newly formed FlexFuel category for consistency with the indicated EngineFuelType. Additionally, we created a “N_A” category to represent the missing values in MarketCategory to address the possibility that the vehicles missing information were systematically different than those with categories assigned.

Finally, our examination of summary statistics uncovered a data entry error. One entry had a listed HighwayMPG of 354, which we corrected to 34 after further research.

Data Sets: Training, Test, and Validate

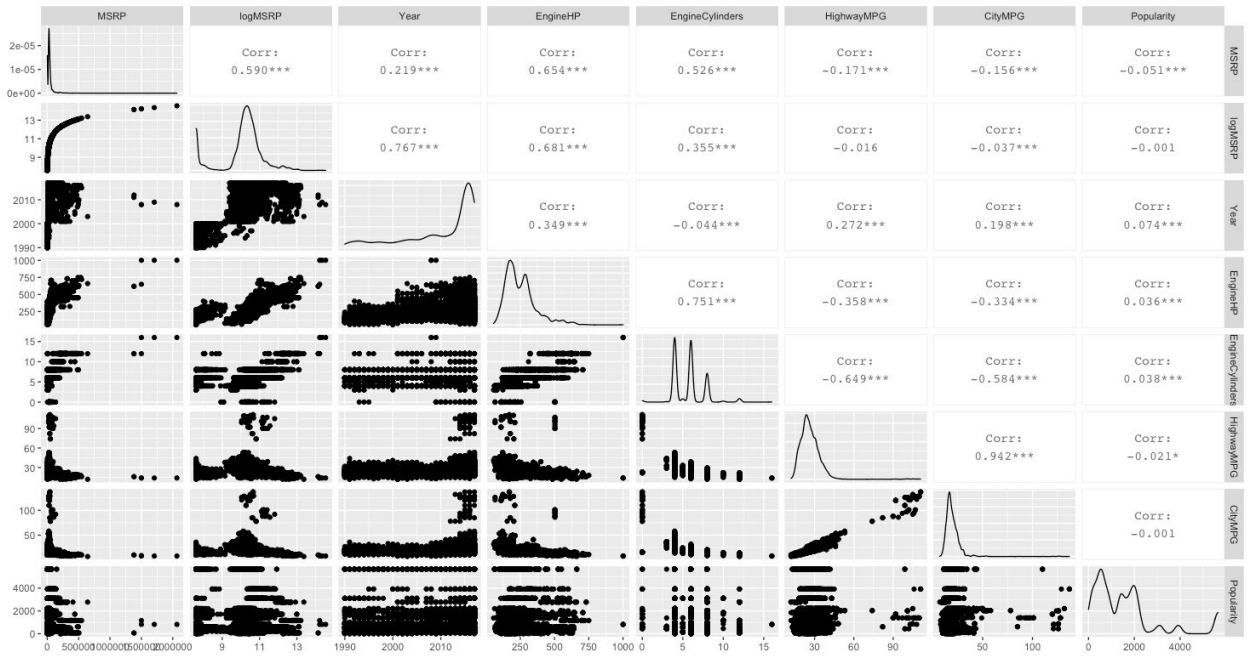
The Model variable contains 912 levels, and some of the levels only contain a small number of entries. We anticipated problems with potential test and validation sets not containing data from each Model level. Because we wanted to preserve the possibility of using the Model variable in our analysis, we first identified Model levels with five entries or less. We assigned these 836 entries to our training set. We then split the remainder of the data randomly into a 10% validation set and a 10% test set, and we assigned the rest of the data to the training set. In this way, we could use all types of vehicles in our training set but the test and validation sets had a very small probability of containing Model levels that were not in the training set, which we verified to be true after the split. See Appendix Table 2 for the training data set summary statistics.

EDA Highlights

Our exploration of the data called our attention to the six entries with MSRP values above $1 million (Figure 3). Because these vehicles were outliers in early exploratory model fits, we decided to remove the six entries and restrict the range of our models to vehicles under $1 million. Although this changes the population to which our models can be applied, we feel that we would require more data to accurately model vehicle prices over a broader range.

Figure 3

Correlation matrix of continuous variables



The distribution of MSRP and its relationship with other potentially important variables in the data set indicated a log transformation should be considered. We built models with both logged and unaltered MSRP and ultimately let the fit of model residuals drive our decision regarding which response variable to use.

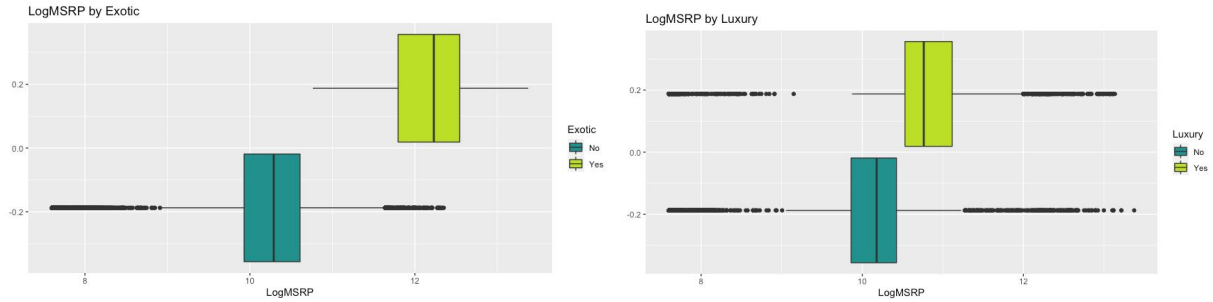
Correlations highlighted some variable relationships to explore. Year, EngineHP, and EngineCylinders all have notable correlations with MSRP. This matches our intuition regarding these variables. Year, for example, is clearly associated with mileage and condition, two characteristics that are important to purchase price but are unavailable in our data set. Similarly, EngineHP and EngineCylinders are measures of performance and engine quality, both of which contribute to pricing.

The high correlation between HighwayMPG and CityMPG was unsurprising, as was the correlation between EngineHP and EngineCylinders. We noted the relationships and delayed any actions to reduce multicollinearity and variance inflation rates until we built each model.

An exploration of the categorical variables also brought our attention to a few variables. Exotic (Figure 4) shows a clear relationship to MSRP, as does Luxury. We hypothesized that these variables would be important, especially when Make and Model were not included as predictors, as they can assist in differentiating premium brand pricing.

Figure 4

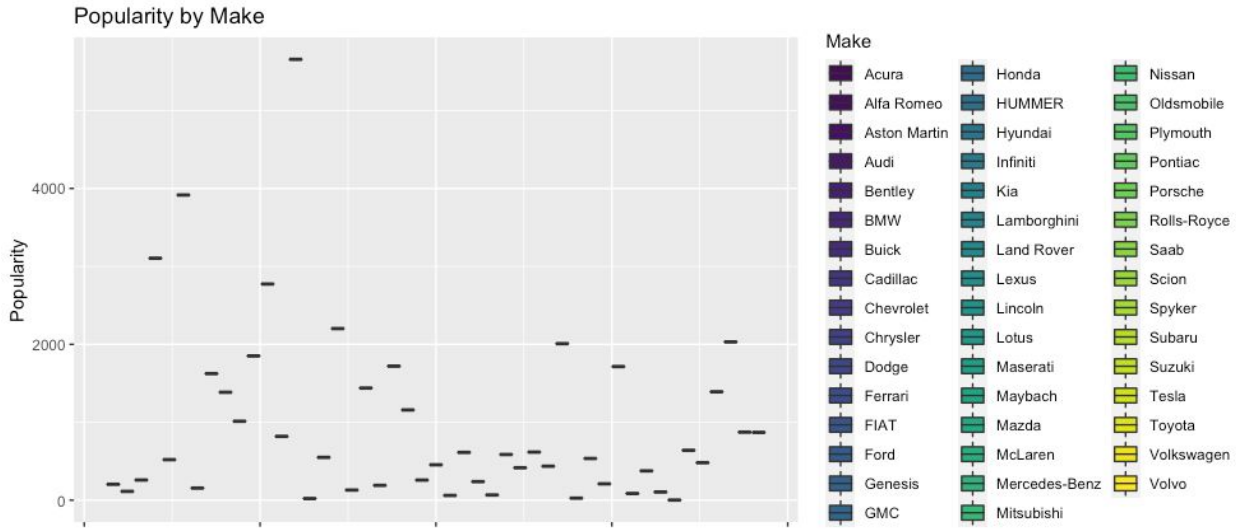
Boxplots: LogMSRP vs Exotic and LogMSRP vs Luxury



Another observation of interest was the relationship between Popularity and Make, as illustrated in the boxplots of Figure 5. Popularity score was the same for each Make, with no variation between Models or individual vehicles.

Figure 5

Popularity vs Make



Objective 1: Interpretable Linear Regression Model

Our goal for objective one was to build an interpretable regression model that identified key relationships with vehicle retail price. Part of this interpretation included an analysis of the importance of the Popularity variable. To achieve our goal, we used a variety of model selection approaches to find a model that balanced simplicity and interpretability with accuracy.

Model Selection

Before beginning model selection, we removed the Make and Model variables from our training set. Both variables have too many levels to be considered easily interpretable. Because our EDA indicated a log transform of MSRP might be beneficial, we built our models using a logged MSRP value and confirmed this was the best option when examining the final model residuals.

First we custom-built manual models using key variables identified through correlation, visualizations, and intuition to create a baseline parsimonious model. To see if variable selection could improve our models, we built options using Forward, Stepwise, and LASSO methods with our training set. We further customized the Stepwise and LASSO models, removing variables with high variance inflation rates (VIF > 10) and nonsignificant p-values ($p > .05$). We compared the test set root mean squared error (RMSE) for the final two models, ultimately choosing the modified stepwise model for Model 1 because it had the better RMSE. Model 1 has an adjusted R^2 of 0.8419, and its test set RMSE is 23,903.2.

Checking Assumptions

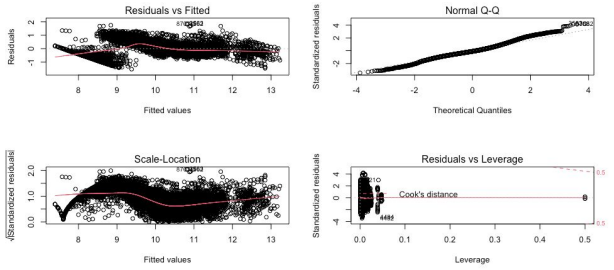
An examination of residual plot (Figure 6) indicates that Model 1 meets the necessary assumptions, and the residuals confirm our use of the logged MSRP response variable. Specifically, the Normal QQ appears to be a straight line in the logged model, indicating normality. The Residuals vs. Fitted values are closer to a random cloud shape in the logged

model, indicating improved constant variance when contrasted with the unlogged model. Finally, the influential points analysis does not highlight any high leverage points.

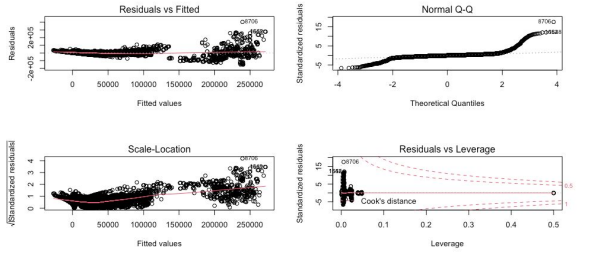
Figure 6

Model 1 residual analysis

Model 1 - log(MSRP)



Model 1 - MSRP (for comparison)



Parameter Interpretation

Model 1:

Predicted logMSRP = Year + EngineHP + Exotic + EngineFuelType + VehicleStyle + Hybrid + EngineCylinders + Luxury + VehicleSize + DrivenWheels + FlexFuel + Crossover + FactoryTuner + Popularity

See Appendix Table 7 for Model 1 coefficients, t-values, p-values, and 95% confidence intervals for coefficient estimates.

Interpretation Example: Year variable

A one unit increase in Year is associated with a multiplicative change of $e^{.09645} = 1.1013$ in the median of MSRP holding all other explanatory variables constant. In other words, a one unit increase in Year is associated with a 10.13% increase in the median of MSRP holding all other variables constant. A 95% confidence interval for this multiplicative increase in the median of MSRP is $(e^{.094752}, e^{.098142}) = [1.0994, 1.1031]$.

Interpretation Example: Exotic variable

The Exotic variable is categorical with two levels (yes/no). No is the reference category. When all other explanatory variables are negligible, the predicted median of MSRP for exotic cars is $e^{.8925} = 2.4412$ times more than the median for non-exotic cars. A 95% confidence interval for this multiplicative increase in the median of MSRP is $(e^{.83396}, e^{.95102}) = [2.3024, 2.5883]$. The Exotic variable is both statistically and practically significant in determining MSRP.

Importance of Popularity variable:

The Popularity variable is still statistically significant (p-value < .0001) in Model 1, after accounting for all other explanatory variables. A 1000 unit increase in Popularity is associated with a multiplicative change of $e^{-.00001803*1000} = 0.982$ in the median of MSRP (a 1.8% decrease) holding all other explanatory variables constant. The Popularity variable ranges in value from 2

to 5657. While statistically significant in the model, this variable does not hold much practical significance in its relationship to MSRP. It is also important to note that the Popularity values are solely determined by Make, with no variation for Model or individual vehicle. If Make is included in the subsequent models, it is unlikely that the Popularity variable would add any value at all.

Objective 2: Multiple Linear Regression & Nonparametric Models

Our goal for objective two was to build two models: one multiple linear regression model incorporating additional complexity and one nonparametric regression model. The purpose of both models is to predict future MSRP values as accurately as possible.

Complex Regression Model (Model 2)

Because we were focused on prediction rather than interpretation, we included Make and Model as variables in the training set to build our complex regression model. We used Forward, Backward, and LASSO selection techniques with explorations of logged and unaltered MSRP to build three candidate models. Lowest BIC value was the criterion used to select the optimal number of predictors in the Forward and Backward models. The LASSO model produced the lowest test set RMSE.

Next we tried adding some possible polynomial and interaction terms to the LASSO model. We noted during our EDA that Year, EngineHP, and the interaction of EngineFuelType and EngineCylinders could be worth exploration. While $Year^2$ and $Year^3$ were not selected during our subsequent LASSO iterations, and the EngineFuelType*EngineCylinders interaction was redundant according to the adjusted R^2 and test RMSE, we did include an $EngineHP^2$ variable. Model 2 has an adjusted R^2 value of 0.9775 and a test set RMSE of 4957.52.

Model 2:
$$Predicted\ logMSRP = Make + Model + Year + EngineHP + EngineHP^2 + EngineFuelType + EngineCylinders + TransmissionType + DrivenWheels + NumberOfDoors + VehicleSize + VehicleStyle + HwyMPG + CityMPG + Hybrid + EngineCylinders + Crossover + Diesel + Luxury + HighPerformance + Exotic + FlexFuel + Performance + FactoryTuner + FlexFuel + N\_A$$

Nonparametric Model (Model 3)

We created our final model using a k-nearest neighbors (KNN) regression approach. This is a nonparametric technique that averages the K closest training observations to estimate the prediction point. When K is a small number, the model will have low bias and high variance. This is because the prediction is based on one or few observations and can lead to an overfit model.

KNN regression can only be applied to numerical variables, so we removed all categorical variables from our training and test sets, leaving Year, EngineHP, EngineCylinders, NumberOfDoors, HighwayMPG, CityMPG, and Popularity as predictors. Next we explored the

optimal number for k, achieving the lowest test RMSE with a value of k = 2. The final version for Model 3 produced a test set RMSE of 6861.36

Model Comparisons

A comparison of our three models (Table 8) indicates that Model 2 has the smallest validation RMSE, although Model 3 performs almost as well with far fewer predictors. While Model 1 is easy to interpret, its prediction accuracy leaves something to be desired. Overall it is not surprising that including Make and Model in Model 2 leads to the most accurate MSRP predictions, as a regression model that contains only the variables Make and Model explains 97.2% of the variance in logMSRP alone.

We were surprised that Model 3, the nonparametric KNN regression model, performed almost as well as Model 2, our complex regression model. We were concerned with the potential for overfitting Model 3, as the optimal K value was only two. Such a small K can produce low bias and high variance, but Model 3 has a smaller validation RMSE than its test RMSE. With only seven predictors and no categorical variables to include Make and Model, Model 3 has a validation RMSE approximately $200 away from Model 2, our best prediction model. It is possible that the nonparametric Model 3 was able to model some nonlinear complexity we could not identify through our EDA or incorporate into our traditional regression models. Because Model 3 performed surprisingly well, we hypothesize a nonparametric method that can include categorical predictors, such as random forest, might produce even higher prediction accuracy than our best regression models.

Table 8
Model Comparisons

	Test RMSE	Validation RMSE	Adjusted R^2
Regression Model 1	23,903.2	21,896.2	0.8419
Regression Model 2	4,957.53	5,521.82	0.9775
KNN Regression Model 3	6,861.36	5,718.33	N/A

Final Summary

Model 1 is an interpretable linear regression model that allows for the identification and communication of key relationships with predicted price. The relationships of the variables with MSRP make intuitive sense. Model 1 also illustrated that although the Popularity variable was statistically significant, it does not hold much practical value for predicting retail price. The interpretability of Model 1 came at the cost of prediction accuracy, with a much larger validation RMSE than the other models.

Model 2 is a much more complex linear regression, using Make, Model, and most of the additional variables as well as a polynomial term. While it would be difficult to interpret, it provides the most accurate MSRP predictions. Model 3 used nonparametric KNN regression and contained only the numeric variables in our data set. While it had a slightly larger validation RMSE than Model 2, it is far better than Model 1 in terms of prediction error.

Before building our models, we restricted the range of vehicles to those under $1 million, so the models we built can only be applied to vehicles in that price range. Additionally, the origins of this data set are unclear and it is unknown if they were from a random sample, so the scope of inference for these models cannot be extended beyond this data set, nor can causal inferences be made.

If Model 2 is applied to additional data, the Make and Model predictors could pose a problem. The inclusion of a vehicle with a Make or Model not used in our training set would mean Model 2 could not predict the MSRP for that vehicle. Model 1 and Model 3 do not suffer from this restriction, however.

Although this data set contained a large amount of information, it might have been useful to have variables measuring mileage and condition. Also, more explanation about how the various MarketCategories were assigned would have been helpful. As previously mentioned, future exploration avenues could include building a model using random forest methods. If interpretability is not essential, the capacity for random forest to use categorical variables as well as variable selection methods could produce better predictions than we obtained using regression and KNN regression models.

References

For project instructions, please see:
https://github.com/nickmingyang/MSDS6372Project1/blob/main/Project1_Requirements_2021.docx

For original data set:
https://github.com/nickmingyang/MSDS6372Project1/blob/main/data1.csv

For cleaned data set:
https://github.com/nickmingyang/MSDS6372Project1/blob/main/CleanedCarData.csv

For principle code:
https://github.com/nickmingyang/MSDS6372Project1/blob/main/MSDS6372Project1.Rmd

For additional EDA, visualizations, and data exploration:
https://github.com/nickmingyang/MSDS6372Project1/blob/main/EDA%20Project%201%20NN%206372.Rmd

https://github.com/nickmingyang/MSDS6372Project1/blob/main/SASedaAttempt

https://github.com/nickmingyang/MSDS6372Project1/blob/main/Model1NN

Appendix

Table 1

Original variable descriptions

Variable Name	Data Type	Description
MSRP	Numeric	The response variable
Car Make	Factor	The company that made the car. Ex: Honda, Toyota, etc.
Car Model	Factor	The model of the car. Ex: 4Runner, Accord, etc.
Year	Numeric	Year the car was produced
Engine Fuel Type	Factor	Type of fuel the car accepts. Ex: Regular unleaded, Premium unleaded, Diesel
Engine HP	Numeric	Horsepower of the car's engine.
Engine Cylinders	Numeric	Number of cylinders in the car's engine.
Transmission Type	Factor	Type of transmission in the car. Usually manual or automatic, but there are a few specialty transmission types in the data.
Driven_Wheels	Factor	The wheels that are powered by the engine. Ex: Front Wheel, Rear Wheel, Four Wheel Drive
Number of Doors	Numeric	The number of doors that the car has. Usually 2 or 4
Market Category	Factor	Various special factors for each car. Ex: Exotic, Luxury, High-Performance, Flex Fuel.
Vehicle Size	Factor	The size of the vehicle. Ex: Midsize, Large, Compact
Vehicle Style	Factor	Body type of the vehicle. Ex: Coupe, Convertible, etc.
Highway MPG	Numeric	Fuel efficiency on the highway in MPG
City MPG	Numeric	Fuel efficiency in the city in MPG
Popularity	Numeric	A popularity score for each car. The dataset does not detail how the popularity score is calculated.

Table 2

Training set summary statistics

Characteristic	N = 8,690
Year, Mean (SD) Maximum Minimum	2,011 (7) 2,017 1,990
EngineFuelType, n (%)	
diesel	125 (1.4%)
electric	26 (0.3%)
flex-fuel (premium unleaded recommended/E85)	22 (0.3%)
flex-fuel (premium unleaded required/E85)	33 (0.4%)
flex-fuel (unleaded/E85)	690 (7.9%)
flex-fuel (unleaded/natural gas)	2 (<0.1%)
natural gas	2 (<0.1%)
premium unleaded (recommended)	1,135 (13%)
premium unleaded (required)	1,311 (15%)
regular unleaded	5,344 (61%)
EngineHP, Mean (SD) Maximum Minimum	246 (103) 750 55
EngineCylinders, Mean (SD) Maximum Minimum	6 (2) 12 0
TransmissionType, n (%)	
AUTOMATED_MANUAL	423 (4.9%)
AUTOMATIC	6,040 (70%)
DIRECT_DRIVE	26 (0.3%)
MANUAL	2,187 (25%)
UNKNOWN	14 (0.2%)
DrivenWheels, n (%)	
all wheel drive	1,752 (20%)
four wheel drive	1,063 (12%)
front wheel drive	3,519 (40%)
rear wheel drive	2,356 (27%)
NumberOfDoors, Mean (SD) Maximum Minimum	3 (1) 4 2
VehicleStyle, n (%)	
2dr Hatchback	370 (4.3%)
2dr SUV	103 (1.2%)
4dr Hatchback	505 (5.8%)
4dr SUV	1,882 (22%)
Cargo Minivan	56 (0.6%)
Cargo Van	66 (0.8%)
Convertible	548 (6.3%)
Convertible SUV	18 (0.2%)
Coupe	811 (9.3%)
Crew Cab Pickup	537 (6.2%)
Extended Cab Pickup	495 (5.7%)
Passenger Minivan	306 (3.5%)
Passenger Van	92 (1.1%)
Regular Cab Pickup	305 (3.5%)
Sedan	2,154 (25%)
Wagon	442 (5.1%)
HighwayMPG, Mean (SD) Maximum Minimum	27 (7) 109 12
CityMPG, Mean (SD) Maximum Minimum	20 (7) 128 8
Popularity, Mean (SD) Maximum Minimum	1,585 (1,452) 5,657 26
MSRP, Mean (SD) Maximum Minimum	37,515 (42,123) 548,800 2,000
Crossover, n (%)	1,561 (18%)
Diesel, n (%)	163 (1.9%)
Exotic, n (%)	264 (3.0%)
Luxury, n (%)	2,252 (26%)
HighPerformance, n (%)	889 (10%)
FactoryTuner, n (%)	342 (3.9%)
FlexFuel, n (%)	926 (11%)
Hatchback, n (%)	875 (10%)
Hybrid, n (%)	235 (2.7%)
N_A, n (%)	2,805 (32%)

**Note: Make and Model are not included in these summary statistics.*

Make: Categorical variable - 47 levels

Model: Categorical variable - 912 levels

Table 7
Model 1 coefficients

	Est.	2.5%	97.5%	t val.	p
(Intercept)	-184.577883	-187.989240	-181.166525	-106.061039	0.000000
Year	0.096447	0.094752	0.098142	111.546138	0.000000
EngineHP	0.001432	0.001217	0.001648	13.022035	0.000000
ExoticYes	0.892490	0.833963	0.951017	29.891583	0.000000
EngineFuelTypeelectric	0.385612	0.230073	0.541152	4.859752	0.000001
EngineFuelTypeflex-fuel (premium unleaded recommended/E85)	-0.168565	-0.381040	0.043909	-1.555124	0.119950
EngineFuelTypeflex-fuel (premium unleaded required/E85)	-0.292848	-0.459447	-0.126249	-3.445675	0.000572
EngineFuelTypeflex-fuel (unleaded/E85)	-0.450200	-0.557788	-0.342612	-8.202454	0.000000
EngineFuelTypeflex-fuel (unleaded/natural gas)	-0.390461	-1.012625	0.231703	-1.230203	0.218652
EngineFuelTypenatural gas	-0.016649	-0.634867	0.601568	-0.052790	0.957900
EngineFuelTypepremium unleaded (recommended)	-0.115386	-0.197376	-0.033396	-2.758659	0.005815
EngineFuelTypepremium unleaded (required)	0.182665	0.099475	0.265855	4.304149	0.000017
EngineFuelTyperegular unleaded	-0.279229	-0.358200	-0.200259	-6.931064	0.000000
VehicleStyle2dr SUV	-0.065073	-0.166159	0.036012	-1.261877	0.207024
VehicleStyle4dr Hatchback	-0.078489	-0.137071	-0.019907	-2.626303	0.008646
VehicleStyle4dr SUV	0.217063	0.155188	0.278938	6.876612	0.000000
VehicleStyleCargo Minivan	0.189618	0.068794	0.310441	3.076308	0.002102
VehicleStyleCargo Van	-0.260179	-0.380458	-0.139899	-4.240176	0.000023
VehicleStyleConvertible	0.222637	0.163873	0.281400	7.426670	0.000000
VehicleStyleConvertible SUV	0.318468	0.139473	0.497464	3.487610	0.000490
VehicleStyleCoupe	0.036994	-0.018828	0.092816	1.299059	0.193955
VehicleStyleCrew Cab Pickup	0.102396	0.032977	0.171814	2.891411	0.003844
VehicleStyleExtended Cab Pickup	-0.052026	-0.120101	0.016049	-1.498092	0.134143
VehicleStylePassenger Minivan	0.301850	0.234044	0.369657	8.726159	0.000000
VehicleStylePassenger Van	-0.110203	-0.218466	-0.001941	-1.995350	0.046033
VehicleStyleRegular Cab Pickup	-0.125155	-0.200034	-0.050275	-3.276333	0.001055
VehicleStyleSedan	0.019800	-0.030719	0.070319	0.768270	0.442346
VehicleStyleWagon	0.183919	0.122651	0.245187	5.884321	0.000000
HybridYes	0.225074	0.171290	0.278859	8.202995	0.000000
EngineCylinders	0.097193	0.086256	0.108130	17.419990	0.000000
LuxuryYes	0.132260	0.106208	0.158313	9.951550	0.000000
VehicleSizeLarge	-0.031722	-0.064738	0.001294	-1.883401	0.059677
VehicleSizeMidsize	-0.048570	-0.073640	-0.023499	-3.797594	0.000147
DrivenWheelsfour wheel drive	-0.007107	-0.052153	0.037940	-0.309261	0.757130
DrivenWheelsfront wheel drive	-0.031640	-0.061845	-0.001435	-2.053330	0.040068
DrivenWheelsrear wheel drive	-0.084238	-0.116493	-0.051982	-5.119240	0.000000
FlexFuelYes	0.242604	0.174265	0.310943	6.958796	0.000000
CrossoverYes	-0.063356	-0.105321	-0.021390	-2.959369	0.003090
FactoryTunerYes	-0.092465	-0.138497	-0.046433	-3.937505	0.000083
Popularity	-0.000018	-0.000025	-0.000011	-5.339876	0.000000