The Effect of Vehicle Characteristics on Selling Price

Ryan Murage

10/26/2025

1 Overview

The used car market is a significant portion of the automotive industry, with prices varying widely based on vehicle characteristics. Understanding which features most strongly influence selling prices can help both buyers and sellers make more informed decisions. In this analysis, we applied linear regression modeling to data from the used auto industry to investigate the effect of vehicle characteristics on selling price.

The main question of this investigation is how do the characteristics of a used vehicle effect its selling price?

2 Data Description and Processing

The data set we used was compiled by the used auto industry which records the characteristics of used cars along with their selling price.

The data set contains several types of observations about the characteristics for several used cars. Variables in the data set include:
selling_price (numeric): Selling price of the vehicle in US dollars. Dependent Variable
km_driven (numeric): The kilometers the car has been driven.
seller_type (categorical): Individual or dealer.
transmission (categorical): Manual or automatic.
owner (categorical): The number of previous owners of the vehicle, with levels: First Owner, Second Owner, Third Owner, Fourth and Above Owner.
mileage (numeric): The fuel consumption of a vehicle in miles per gallon.
engine (numeric): The volume of an engine in CCs.
max_power (numeric): The horsepower of a vehicle in bhp (brake horsepower).
seats (categorical): The number of seats.

The data set was checked for missing values. Of the original 3,425 observations, we removed 111 observations that contained missing values leaving 3,314 complete observations.

3 Data Exploration

Table 1: Summary Statistics for Selling Price

Statistic	Value
Min.	29,999.0
1st Qu.	200,000.0
Median	340,000.0
Mean	446,538.5
3rd Qu.	535,000.0
Max.	10,000,000.0

The table above shows descriptive statistics for selling price, including the minimum, first quartile, median, mean, third quartile, and maximum values.

We then explored the distribution of our outcome variable, selling price.

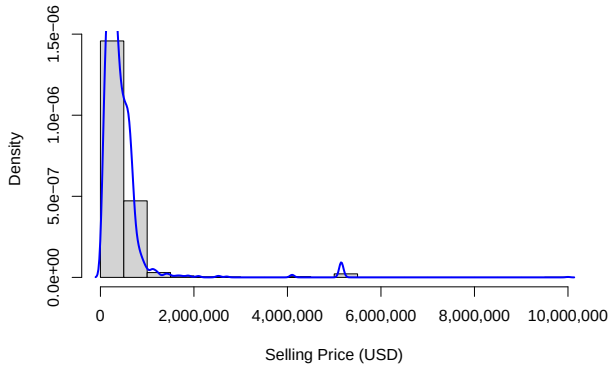

Figure 1: Distribution of Selling Price

This histogram displays that the prices for the vehicles are right-skewed, with a large majority of observations concentrated on the lower end of the price range. This makes sense as most cars available on the used car market are affordable, with few vehicles being high-end.

We also explored the distribution of categorical variables. The following tables and plots display the frequency of observations across seller type, transmission type, previous ownership, and number of seats.

Table 2: Distribution of Seller Types

Seller_Type	Count	Percentage
Dealer	411	12.4
Individual	2903	87.6

Figure 2: Distribution of Seller Types

These results show that a large majority of sellers for used vehicles are individuals, which account for 87.6% of the vehicles being sold in the data set.

Table 3: Distribution of Transmission Types

Transmission_Type	Count	Percentage
Automatic	406	12.25
Manual	2908	87.75

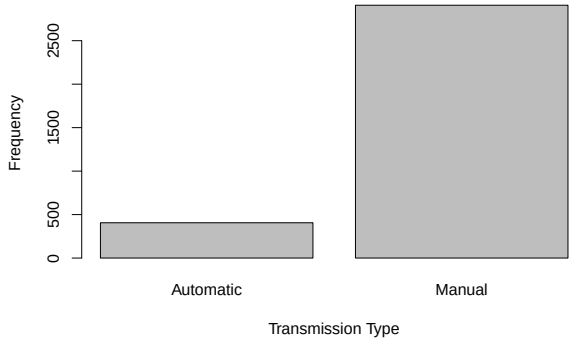

Figure 3: Distribution of Transmission Types.

These results show that a large majority of transmission types in these used vehicles are manual, which account for 87.75% of vehicles being sold in the data set.

Table 4: Distribution of Owner Types

Owner_Type	Count	Percentage
First Owner	2204	66.51
Second Owner	801	24.17
Third Owner	233	7.03
Fourth & Above Owner	76	2.29

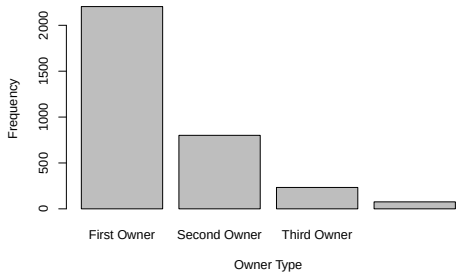

Figure 4: Distribution of Owner Types.

These results show that a large majority of the used vehicles in this data set are sold by their first owner, which account for 66.51% of used vehicles being sold in this data set.

Table 5: Distribution of Number of Seats

Seat_Number	Count	Percentage
4	117	3.53
5	3062	92.40
6	21	0.63
7	82	2.47
8	32	0.97

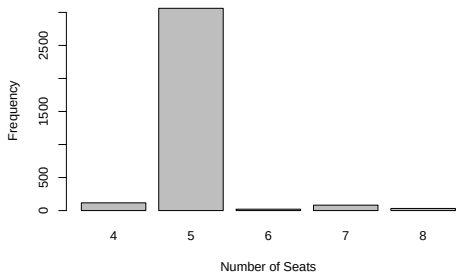

Figure 5: Distribution of Number of Seats.

These results show that a large majority of used vehicles being sold have 5 seats, which account for 92.4% of used vehicles being sold in the data set.

4 Modeling the Effect of Vehicle Characteristics on Selling Price

What factors most strongly influence the selling price of a used vehicle? The variables mileage, engine size, kilometers driven, and max power represent technical attributes that capture a vehicle's performance and efficiency, while seller type, transmission type, number of previous owners, and number of seats represent market and usage characteristics. Technical features serve as indicators of a car's quality and capability, which are expected to increase its market value. In contrast, market-related variables such as high usage, manual transmission, or multiple ownership histories typically reduce perceived value. By examining how these groups of variables affect selling price, we can infer the relative importance of technical condition versus market perception in determining a vehicle's selling price.

4.1 Hypothesis

We hypothesize that characteristics like higher engine power and better fuel efficiency will increase selling price, while higher kilometers driven and multiple previous owners will decrease selling price. We also expect automatic transmission vehicles to have higher selling prices than manual transmission vehicles.

4.2 Model 1:

We fit a linear regression model with the technical characteristics of a vehicle: km_driven, mileage, engine, max_power, and an interaction term engine:max_power, with selling_price as the response variable. The fit results are displayed below:

Table 6: Technical Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-396213.410	76889.3865	-5.1530	0
km_driven	-0.974	0.1016	-9.5916	0
mileage	41837.819	2212.9330	18.9060	0
engine	-1232.284	63.2312	-19.4885	0
max_power	10642.105	672.4361	15.8262	0
engine:max_power	6.826	0.2326	29.3450	0

The adjusted R-squared for this model was 0.7265, which is a reasonably good fit.

A plot of the fit residuals and a QQ plot show the fit errors are reasonably normally distributed around zero with a mean skewed to the left.

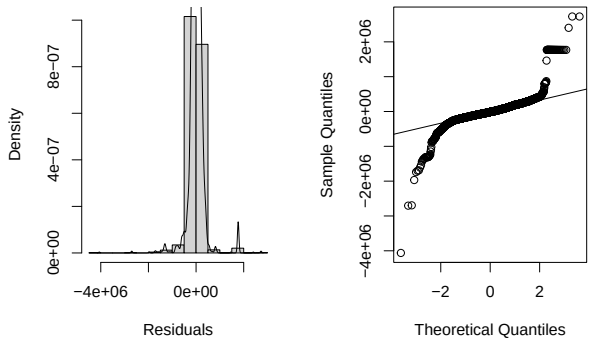

Figure 6: Distribution of Residuals: Model 1

In this model, each predictor was significant. The variables that measured mileage, max power, and the interaction between engine size and max power had positive coefficients, meaning that an increase in the miles per gallon, and an increase in the max power and engine size of a vehicle increase the price. The variables that measured kilometers driven and engine size had negative coefficients, meaning that an increase in the kilometers a used vehicle has driven, and the increase in the size of an engine without a proportionate increase in power reduce the price.

4.3 Model 2: Including Market Factors

The data contains the technical aspects of a car, as well as characteristics that are more so market factors such as the number of previous owners or seller type. There may be significant differences in the selling price of a vehicle when considering these factors as well as the vehicles technical characteristics.

We fit a second linear model that include the technical characteristics as well as the following market factors: seller_type, transmission, owner, and seats, with selling_price as the response variable

Table 7: Comprehenive Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1094763.2643	75043.0567	-14.5885	0.0000
km_driven	-0.5635	0.1108	-5.0865	0.0000
seller_typeIndividual	-138762.2207	19179.9114	-7.2348	0.0000
transmissionManual	-161836.4651	20421.3501	-7.9249	0.0000
ownerSecond Owner	-42210.4160	14648.3645	-2.8816	0.0040
ownerThird Owner	-78761.8066	23854.4015	-3.3018	0.0010
ownerFourth & Above Owner	-245.4997	39637.7464	-0.0062	0.9951
mileage	58917.6232	2368.8674	24.8716	0.0000
engine	-364.8915	59.7999	-6.1019	0.0000
max_power	20224.4732	617.4868	32.7529	0.0000
seats5	-472701.9059	32985.8691	-14.3304	0.0000
seats6	-403042.8466	79633.4182	-5.0612	0.0000
seats7	-341232.0589	49391.6736	-6.9087	0.0000
seats8	117458.5369	66765.8318	1.7593	0.0786

The adjusted R-squared for this model was 0.7006, which is less than the value for the first model, indicating that the inclusion of market factors didn't help the model explain more of the variance.

A plot of the fit residuals and a QQ plot show the fit errors are reasonably normally distributed around zero.

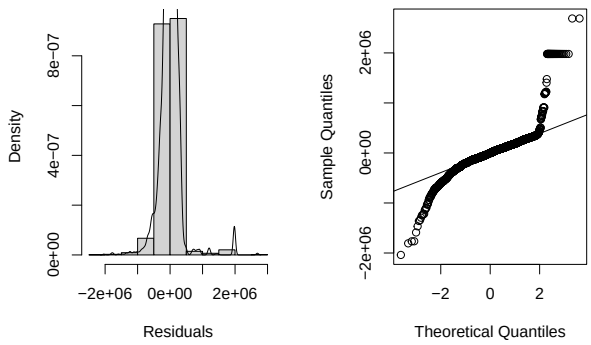

Figure 7: Distribution of Residuals: Model 2.

All variables that were significant in Model 1 remain significant in Model 2. In addition, seller type, transmission type, number of previous owners (up to the third owner), and number of seats (up to seven) were significant predictors of selling price. Each of these variables has a negative coefficient, indicating that vehicles with manual transmission, individual sellers, more previous owners, and a higher seat count tend to have lower selling prices compared to their respective baseline categories.

5 Predicting with Models 1 and 2

We evaluated the ability of the first and second models to predict selling price on test data. We used 10-fold cross validation using 80% of the data for training and 20% for testing on each fold.

We extracted and plotted the RMSE values over the 10 folds from each model.

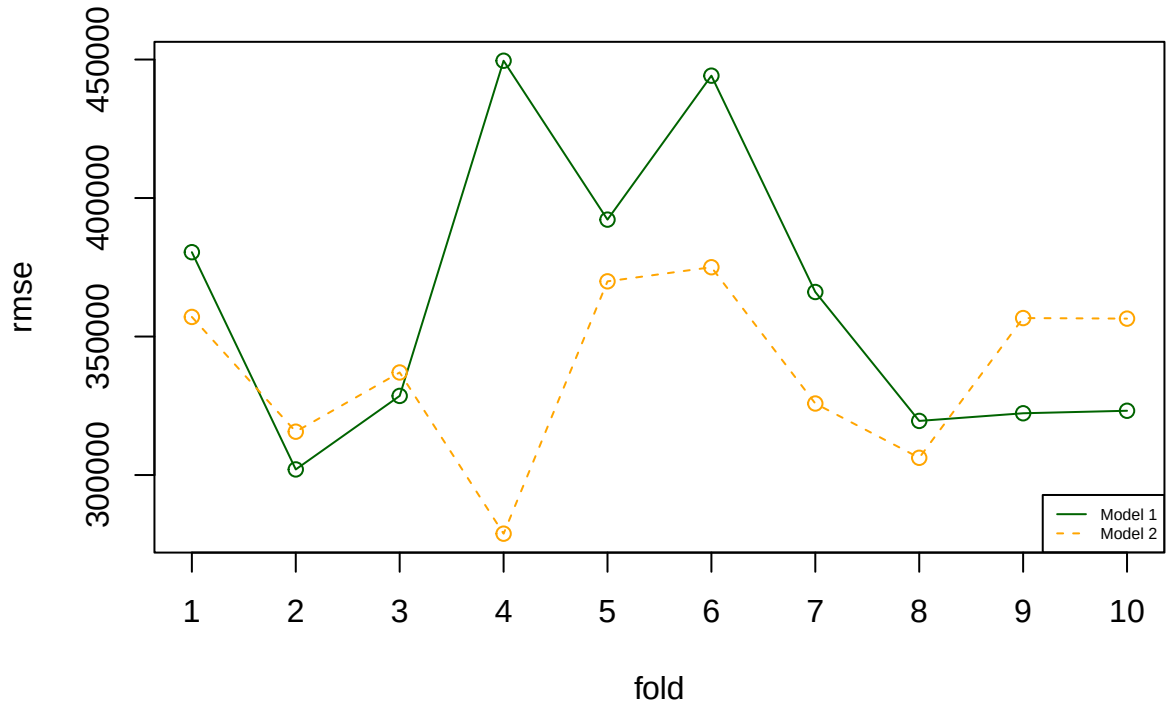

Figure 8: RMSE Over 10 Folds.

From the plot, it's clear that the RMSE for model 2 was consistently lower than for model 1 on most folds.

We calculated the average root mean squared error (RMSE) for both models' predictions over the 10 folds. The results are shown in the table below.

Table 8: Mean RMSE for Models 1 and 2.

label	rmse
Model 1	362809.8
Model 2	337870.23

6 Summary and Conclusions

Both model 1 and model 2 fit the data well, and shows a positive, significant correlation with higher selling prices for vehicles with better mileage, engine size and power, and a negative correlation with lower selling prices for vehicles with higher kilometers driven, individual sellers, multiple previous owners, and a higher seat count.

Model 1 had a better fit, but a higher prediction error, suggesting potential overfitting. Model 2, while having a slightly worse fit, has a lower prediction error. This suggests that Model 2 is a better predictor of selling price. Model 2 included all predictor variables, which intuitively makes sense as to why it is the better model. The inclusion of both the technical aspects of a vehicle, as well as the market factors of a vehicle, allowed Model 2 to fit the data better and be a better predictor.

We recognize that this is a relatively small data sample of the broader used car market, and that there may be additional factors outside of the variables we analyzed that influence the selling price of a vehicle.

With these caveats in mind, the results suggest that vehicles with higher engine capacity, greater power output, and better fuel efficiency tend to sell for higher prices, while those with higher mileage (km driven), more previous owners, or manual transmission tend to sell for less.