Used Car Price Analysis

Allen Chen

2024-06-21

Table of Contents:

Introduction

Background

I have always noticed that although used cars depreciate quite significantly compared to other assets, there are certain reasons why some used cars depreciate faster than others. These reasons may be due to factors such as car brand, how many kilometers are driven, and many more. Thus, I want to see if some explanatory variables have more impact on the value of a used car than others.

Proposed Model

My plan consists of using a multiple linear regression model to determine the relationship between different factors and the price of a used car.

In this analysis, we aim to predict the Price of used cars in Indian Rupees using the following independent variables:

- **Year:** The manufacturing year of the car.
- **Brand:** The brand or manufacturer of the car.
- **Engine:** The engine capacity of the car in CC (Cubic Centimeters).
- **Owner_Type:** The number of previous owners of the car.
- **Kilometers_Driven:** The total kilometers driven by the car.
- **Mileage:** The fuel efficiency of the car in kilometers per liter.
- **Power:** The maximum power output of the car in bhp (Brake Horsepower).

The dependent variable is **Price (Indian Rupees)**.

Data Source

My dataset is from a CSV file on Kaggle: https://www.kaggle.com/datasets/sujithmandala/second-hand-car-price-prediction/data

Part One: Data Summary

Number of NA values
Count: 0

Car Brands and Mean Price

Car_Brands	Mean
Volkswagen	1115000.0
Toyota	1490000.0
Tata	795454.5
Mercedes	2880000.0
Maruti	708333.3
Mahindra	940000.0
Hyundai	713636.4
Honda	808333.3
Ford	1468181.8
BMW	3030000.0
Audi	2570000.0

Owner Type and Mean Price

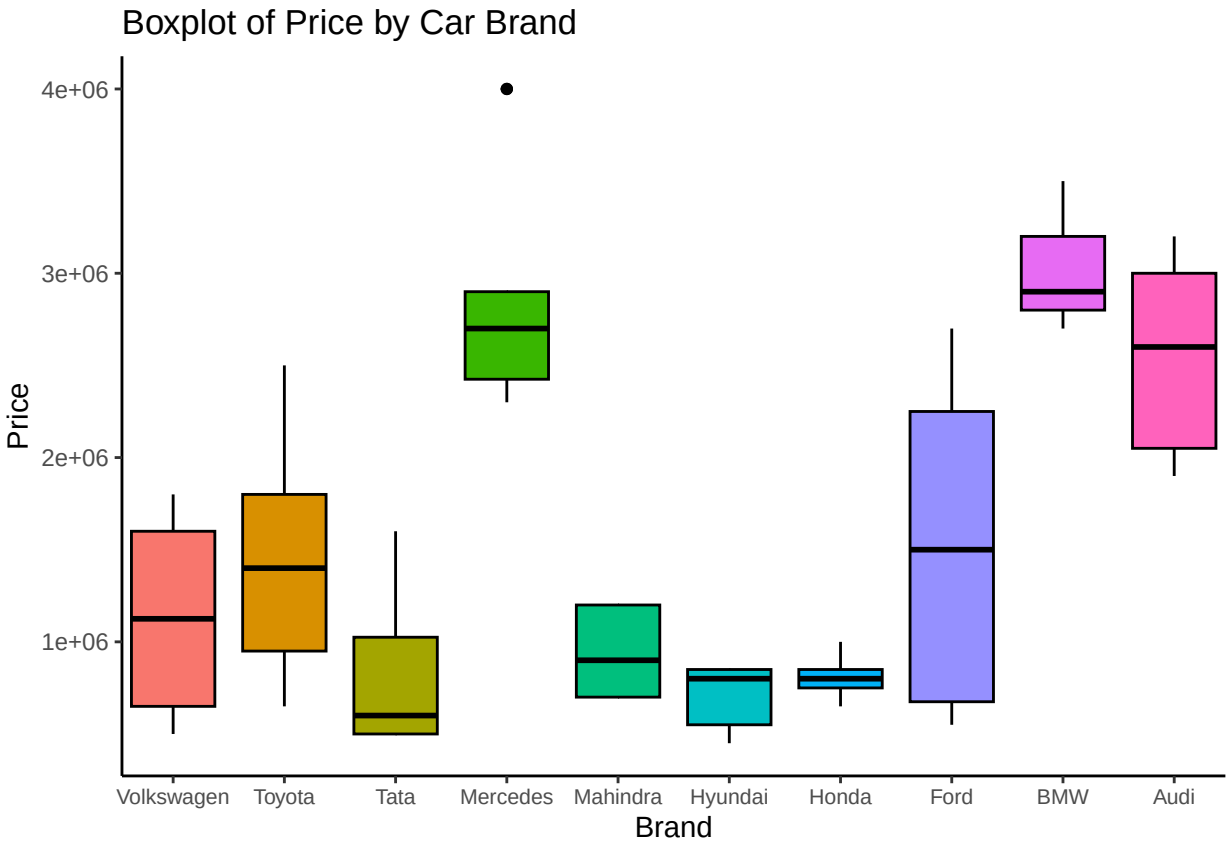
Past_Owners	Mean
First	1753409.1
Second	1687209.3
Third	592307.7

Car Age and Mean Price

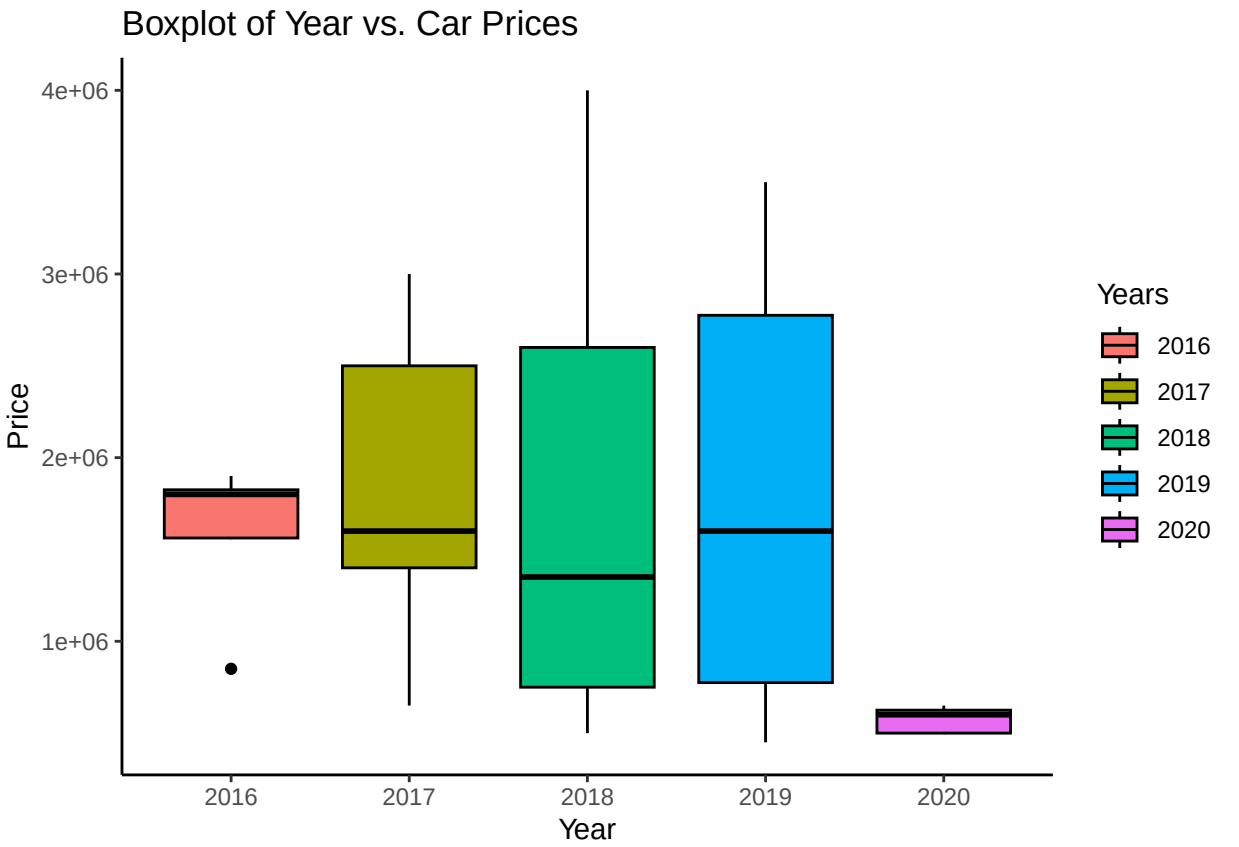
Year	Mean
2021	1200000.0
2020	867647.1
2019	1698214.3
2018	1757142.9
2017	1769047.6
2016	1587500.0

These charts depict the mean price of a used car in relation to car brand, year of manufacture, and number of previous owners. It shows a higher mean price for luxury car brands compared to common car brands, the mean used car price decreases the more previous owners it has, and a high degree of variation in the relationship between the mean used car price and the age of a used car.

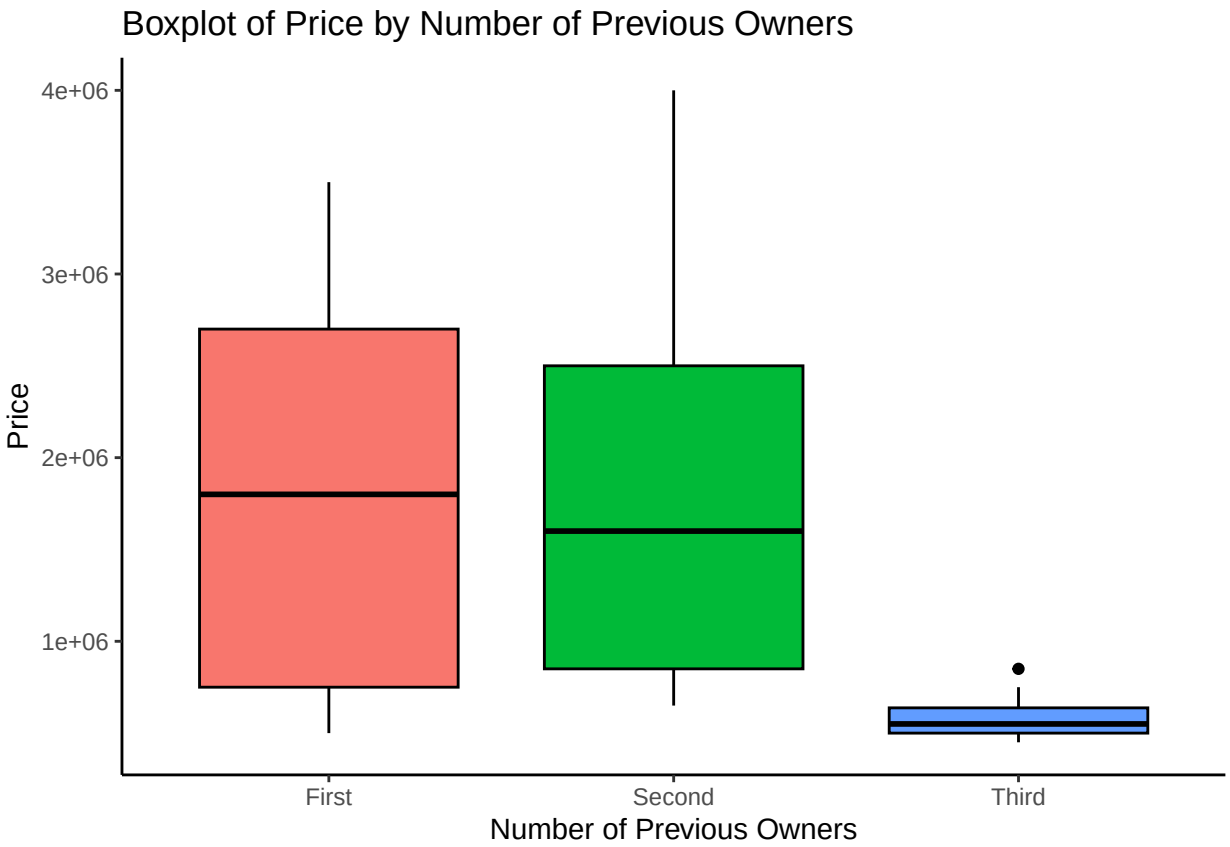
Part Two: Data Cleaning and Visualizations



We remove the data values with the car brand Maruti since it contains multiple outliers and a small sample size (six data values). It shows that luxury car brands (Mercedes, Audi, BMW) have **higher** median used car prices compared to other brands.



We remove all data values with the year 2021 since it contains a small sample size, which does not fulfill the assumptions of linear regression (only two data values with a 2021 car model). The years **2016** and **2019** show a small degree of variability, whereas the years **2017 - 2019** show a high degree of variability.



These boxplots show clearly that the more previous owners a used car has, the median price of the used car will decrease. There is less variation in price for used cars that have had three previous owners compared to one or two previous owners.

Part Three: Model Selection

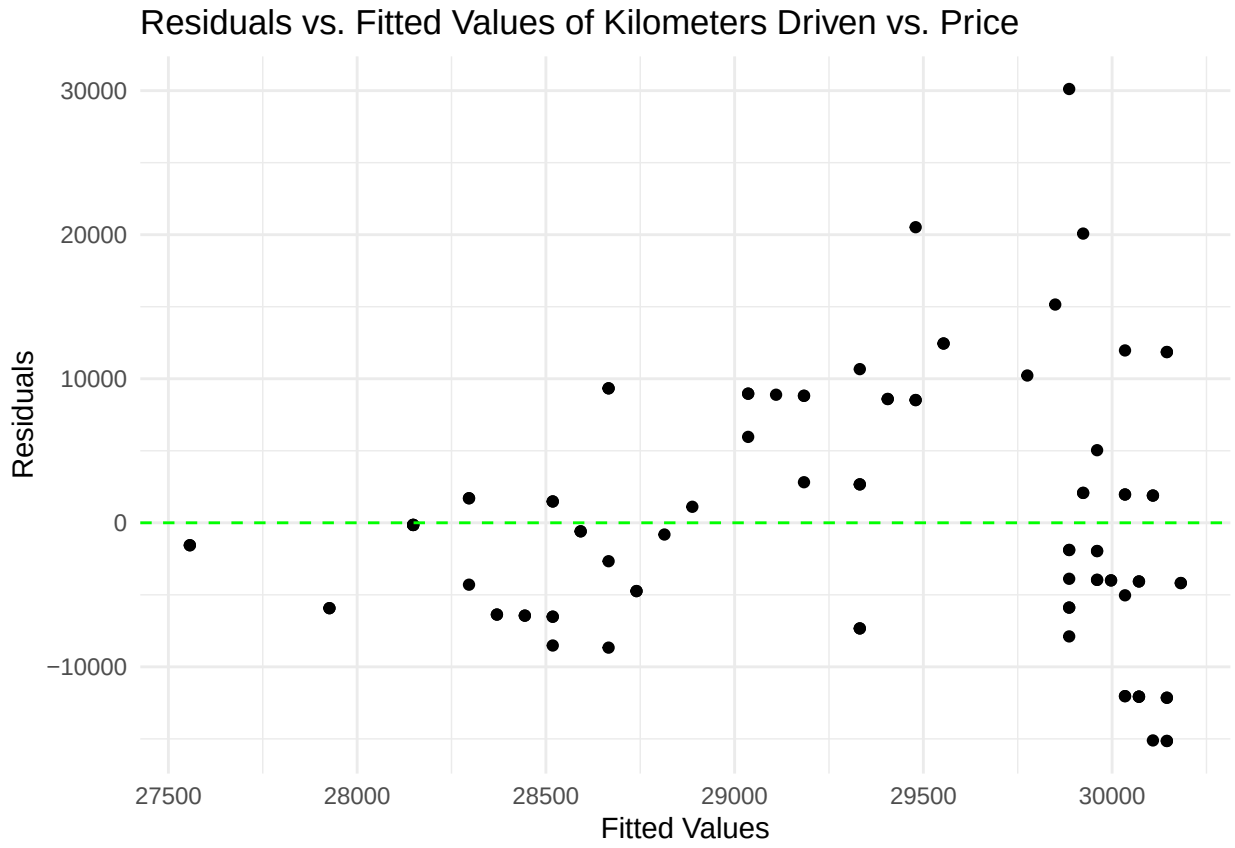
Base Model

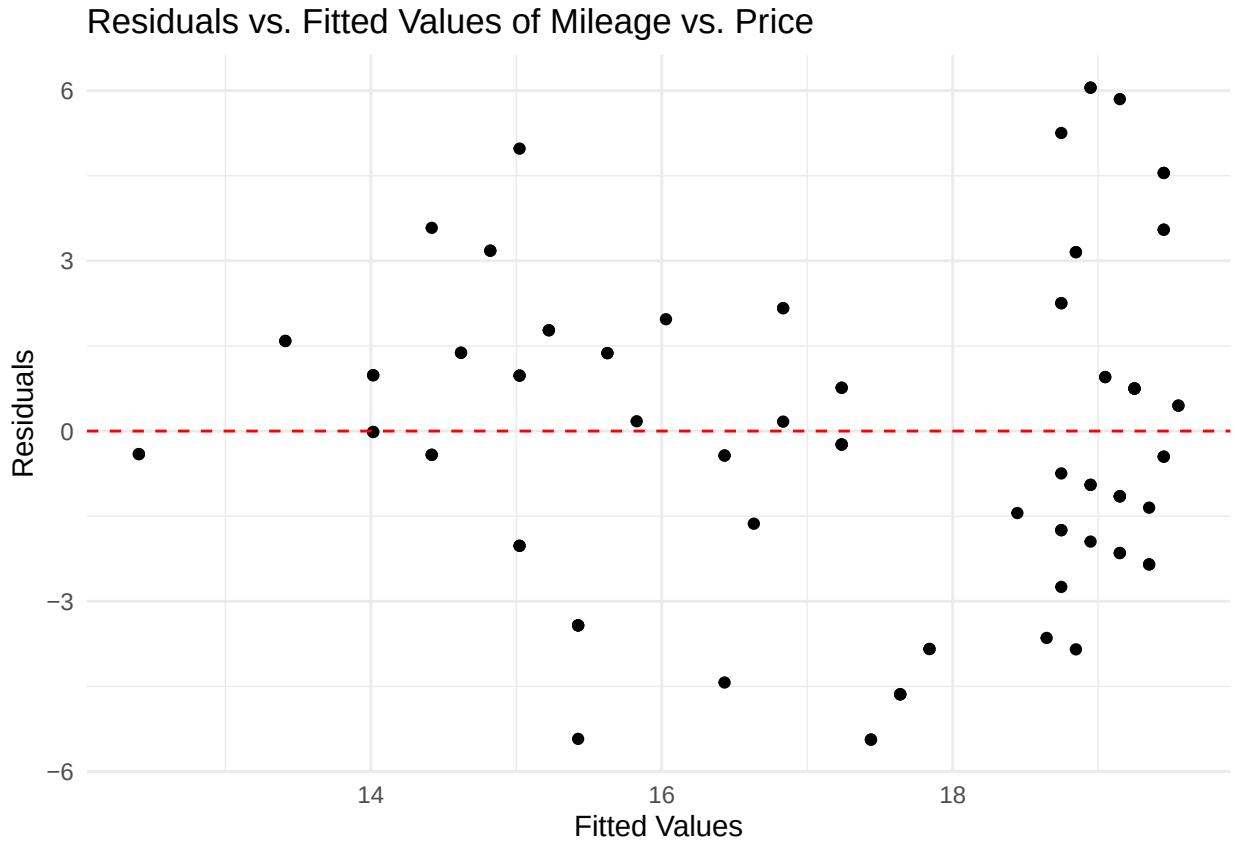
```
Call:
lm(formula = Price ~ factor(Brand) + factor(Year) + factor(Owner_Type) +
    Kilometers_Driven + Mileage + Engine + Power, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-597194 -116322   34107  150136  622230

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               2.075e+06  5.893e+05   3.521  0.00076 ***
factor(Brand)BMW         -6.467e+03  1.588e+05  -0.041  0.96764
factor(Brand)Ford        -1.073e+06  1.317e+05  -8.149 9.73e-12 ***
factor(Brand)Honda       -1.197e+06  1.693e+05  -7.068 9.38e-10 ***
factor(Brand)Hyundai     -1.279e+06  1.428e+05  -8.957 3.17e-13 ***
factor(Brand)Mahindra    -1.556e+06  2.106e+05  -7.389 2.43e-10 ***
factor(Brand)Mercedes     9.911e+03  1.207e+05   0.082  0.93478
factor(Brand)Tata        -1.067e+06  1.500e+05  -7.114 7.73e-10 ***
factor(Brand)Toyota      -9.364e+05  1.529e+05  -6.124 4.73e-08 ***
factor(Brand)Volkswagen  -9.164e+05  1.325e+05  -6.915 1.79e-09 ***
factor(Year)2017         -1.336e+05  1.647e+05  -0.811  0.42019
factor(Year)2018          1.379e+05  1.734e+05   0.795  0.42918
factor(Year)2019          1.188e+05  2.018e+05   0.589  0.55805
factor(Year)2020         -3.445e+05  2.682e+05  -1.284  0.20326
factor(Owner_Type)Second  8.022e+04  8.174e+04   0.981  0.32981
factor(Owner_Type)Third  -2.497e+05  1.303e+05  -1.916  0.05950 .
Kilometers_Driven        -1.103e+01  6.506e+00  -1.695  0.09451 .
Mileage                  -2.674e+04  1.521e+04  -1.758  0.08312 .
Engine                    2.278e+02  1.000e+02   2.277  0.02582 *
Power                     4.024e+03  8.389e+02   4.796 8.82e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 253900 on 70 degrees of freedom
Multiple R-squared:  0.9505,    Adjusted R-squared:  0.937
F-statistic: 70.68 on 19 and 70 DF,  p-value: < 2.2e-16
```

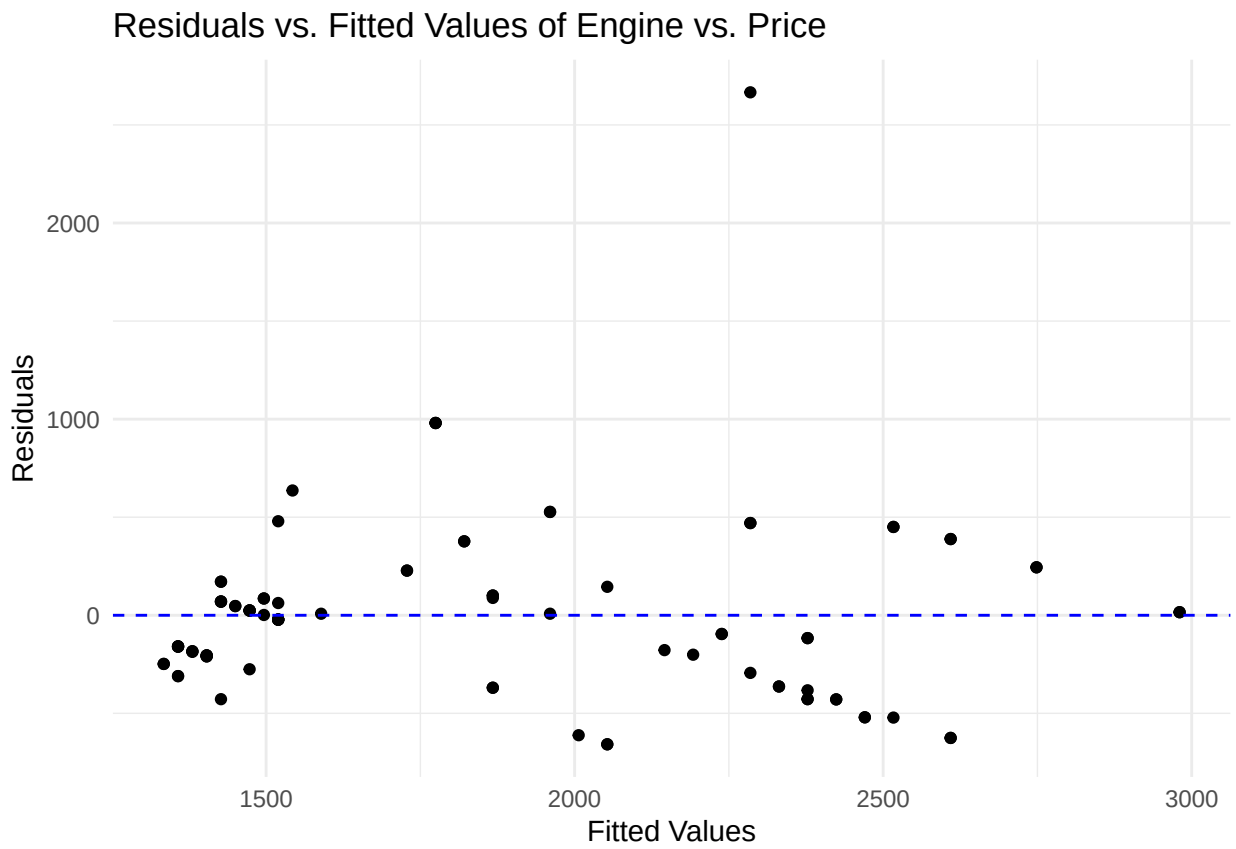
Correlations between Numeric Predictors and Car Price

	Numeric Predictor Correlations
Kilometers_Driven	-0.0867403
Mileage	-0.5995611
Engine	0.7177621
Power	0.8452140

Non-constant Variance Verification







We determine the strength of correlation between the numeric predictors and car price. Several predictors must be transformed to adequately capture the trend in the residual plots, which can increase their statistical significance to the models.

- **Kilometers_Driven** has a weak negative correlation with used car price and its residuals show some outliers and a high degree of variance. There is randomness in the residual plot but a log transformation can help reduce skewness.
- **Mileage** has a moderate negative correlation with price, however its residual plot shows a quadratic trend, so a squared term can more accurately capture its relationship with Price.
- **Engine** has a moderate positive correlation with price, however its residual plot shows a square root relationship, so a square root term can more accurately capture its relationship with Price.
- **Power** does not need any transformation as it has a strong positive correlation with Price.

The transformed predictors will constitute the revised base model.

Revised Base Model

```
Call:
lm(formula = Price ~ factor(Brand) + factor(Year) + factor(Owner_Type) +
    I(Mileage^2) + I(sqrt(Engine)) + Power + I(log(Kilometers_Driven)),
    data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-556048 -150850   24852  138141  621496
```

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                3020240.8  2609740.5   1.157  0.25109
factor(Brand)BMW            -83687.0   157141.8  -0.533  0.59603
factor(Brand)Ford         -1086550.5   129936.1  -8.362 3.93e-12 ***
factor(Brand)Honda        -1243212.0   169313.2  -7.343 2.95e-10 ***
factor(Brand)Hyundai      -1310397.3   140248.5  -9.343 6.22e-14 ***
factor(Brand)Mahindra     -1630065.4   206684.3  -7.887 2.95e-11 ***
factor(Brand)Mercedes       -11253.3   119813.0  -0.094  0.92544
factor(Brand)Tata         -1095960.3   148257.7  -7.392 2.39e-10 ***
factor(Brand)Toyota       -1014762.7   153498.1  -6.611 6.35e-09 ***
factor(Brand)Volkswagen    -946324.9   129905.8  -7.285 3.77e-10 ***
factor(Year)2017            -71702.1   156788.0  -0.457  0.64886
factor(Year)2018            217757.1   165347.6   1.317  0.19215
factor(Year)2019            229189.0   198803.1   1.153  0.25290
factor(Year)2020           -193608.4   291959.9  -0.663  0.50942
factor(Owner_Type)Second     97319.3    79804.0   1.219  0.22676
factor(Owner_Type)Third    -207687.9   131069.0  -1.585  0.11757
I(Mileage^2)                  -467.6      403.6  -1.159  0.25060
I(sqrt(Engine))              30651.8     9829.3   3.118  0.00264 **
Power                         3790.0      828.9   4.572 2.02e-05 ***
I(log(Kilometers_Driven))  -243649.4   226534.3  -1.076  0.28582
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 250500 on 70 degrees of freedom
Multiple R-squared:  0.9518,    Adjusted R-squared:  0.9387
F-statistic:  72.7 on 19 and 70 DF,  p-value: < 2.2e-16
```

Alternative Models

Table 6: VIF of Revised Base Model

	GVIF	Df	GVIF^(1/(2*Df))
factor(Brand)	33.4966	9	1.2154
factor(Year)	20.7570	4	1.4610
factor(Owner_Type)	5.0299	2	1.4976
I(Mileage^2)	3.4177	1	1.8487
I(sqrt(Engine))	6.9051	1	2.6277
Power	5.9343	1	2.4360
I(log(Kilometers_Driven))	6.1000	1	2.4698

Predictor variables with a VIF larger than 10 indicate a significant level of multicollinearity. Thus, removing them will improve the model's predictive ability.

We create a new model, MLR_2 to contain all predictor variables with a VIF of less than 10.

Transformed Revised Base Model

```
Call:
```

```
lm(formula = Price ~ factor(Owner_Type) + I(Mileage^2) + I(sqrt(Engine)) +
    Power + I(log(Kilometers_Driven)), data = cars)

Residuals:
     Min       1Q   Median       3Q      Max
-1898895  -360724   -59859   361693  1006841

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 3081448    2313135   1.332   0.1865
factor(Owner_Type)Second    -107487     124438  -0.864   0.3902
factor(Owner_Type)Third     -455590     195671  -2.328   0.0223 *
I(Mileage^2)                  -1075        642  -1.674   0.0979 .
I(sqrt(Engine))               17624      15598   1.130   0.2618
Power                          7969       1369   5.819 1.07e-07 ***
I(log(Kilometers_Driven))   -299175     219753  -1.361   0.1771
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 527200 on 83 degrees of freedom
Multiple R-squared:  0.7468,    Adjusted R-squared:  0.7284
F-statistic: 40.79 on 6 and 83 DF,  p-value: < 2.2e-16
```

Multicollinearity Verification

Table 7: VIF of Transformed Revised Base Model

	GVIF	Df	GVIF^(1/(2*Df))
factor(Owner_Type)	1.4790	2	1.1028
I(Mileage^2)	1.9530	1	1.3975
I(sqrt(Engine))	3.9268	1	1.9816
Power	3.6575	1	1.9125
I(log(Kilometers_Driven))	1.2963	1	1.1385

The results from MLR_2 show that **Owner_Type, I(sqrt(Engine))**, **I(log(Kilometers_Driven))**, and **I(Mileage^2)** have p-values > 0.05, meaning that they are not significant in adding to the model's predictive ability and can be removed.

Next, we will utilize the residual sum of squares (RSS), adjusted R-squared, Akaike information criterion (AIC), and analysis of variance (ANOVA) test to determine which model is a better fit. AIC is a measure that has a penalty for model complexity, and determines if adding a predictor improves model fit. Similarly, an ANOVA test determines if a model with added complexity would have better fit given a null hypothesis that states a simpler model would be the best fit.

We will perform the ANOVA test on nested models which will start with a model with all the statistically significant predictors from MLR_2 and individually add the predictors that are not statistically significant.

Nested Models ANOVA Test

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
88	2.602018e+13	NA	NA	NA	NA
84	2.496415e+13	4	1.056032e+12	1.1159149	0.3549991
82	2.342369e+13	2	1.540456e+12	3.2556179	0.0437879
81	2.222504e+13	1	1.198650e+12	5.0664807	0.0271724
80	2.200685e+13	1	2.181856e+11	0.9222317	0.3398198
79	1.869016e+13	1	3.316691e+12	14.0190625	0.0003420

Metric Verification for Nested Models

Model	Adjusted_R_Squared	AIC
Model 1	0.7111410	2636.516
Model 2	0.7096675	2640.788
Model 3	0.7209386	2639.055
Model 4	0.7319500	2636.328
Model 5	0.7312638	2637.440
Model 6	0.7688764	2624.738

The ANOVA test shows that Model 6, a complex model with more predictors, is a better fit than the simple base model (Model 1) with just Power as the predictor variable.

The p-value for Model 6 was statistically significant for the F-test, meaning that having additional predictors do have some effect with increasing model fit. Models 3 and 4 are also significant.

To determine the best model, we take the results from the AIC test, RSS, and Adjusted R-squared into consideration. Model 6 has the lowest AIC of all the models and has a higher adjusted R-squared of all the models. It also has a lower RSS than all of the other models. Therefore, **Model 6** has the best fit for determining used car prices.

Part 4: Final Results

Final Model

```
Call:
lm(formula = Price ~ Power + factor(Year) + factor(Owner_Type) +
    I(Mileage^2) + I(sqrt(Engine)) + I(log(Kilometers_Driven)),
    data = cars)

Residuals:
     Min       1Q   Median       3Q      Max
-1810493  -305350    10351   350746   869483

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               15941114.7  4294833.4   3.712 0.000382 ***
Power                        5508.3     1391.9   3.957 0.000164 ***
factor(Year)2017          -362259.3   290565.6  -1.247 0.216177
factor(Year)2018          -211547.4   292260.7  -0.724 0.471308
factor(Year)2019          -474489.4   340171.8  -1.395 0.166970
factor(Year)2020         -1501999.9   480513.0  -3.126 0.002481 **
factor(Owner_Type)Second   -158431.3   127086.7  -1.247 0.216213
factor(Owner_Type)Third    -765933.0   195893.2  -3.910 0.000194 ***
I(Mileage^2)                -1951.7      635.8  -3.069 0.002938 **
I(sqrt(Engine))              8925.4    14959.7   0.597 0.552458
I(log(Kilometers_Driven)) -1401223.3   374238.0  -3.744 0.000342 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 486400 on 79 degrees of freedom
Multiple R-squared:  0.7948,    Adjusted R-squared:  0.7689
F-statistic: 30.61 on 10 and 79 DF,  p-value: < 2.2e-16
```

Significance and Interpretation of Final Model Predictors

-**Power:** The coefficient is positive and statistically significant ($p < 0.05$), which indicates a strong relationship with used car prices. Every unit increase in brake horsepower increases the used car price by 5508.3 Indian Rupees.

-**Year:** Only the level for the year 2020 is statistically significant. As year increases by one, the used car price decreases by a differing amount based on the year ranging from -1501999.9 Indian Rupees in 2020 to -211547.4 Indian Rupees in 2018 relative to the used car price in 2016.

-**Owner_Type:** The levels for used cars with either one or three previous owners are statistically significant. As the number of previous owners for a used car increases by one, the used car price decreases by -158431.3 Indian Rupees for two previous owners or -765933.0 Indian Rupees for three previous owners relative to the used car price for one previous owner.

-**I(Mileage^2):** The coefficient is negative and is statistically significant. It suggests that as mileage increases by one km/liter, the used car price may decrease in a non-linear way.

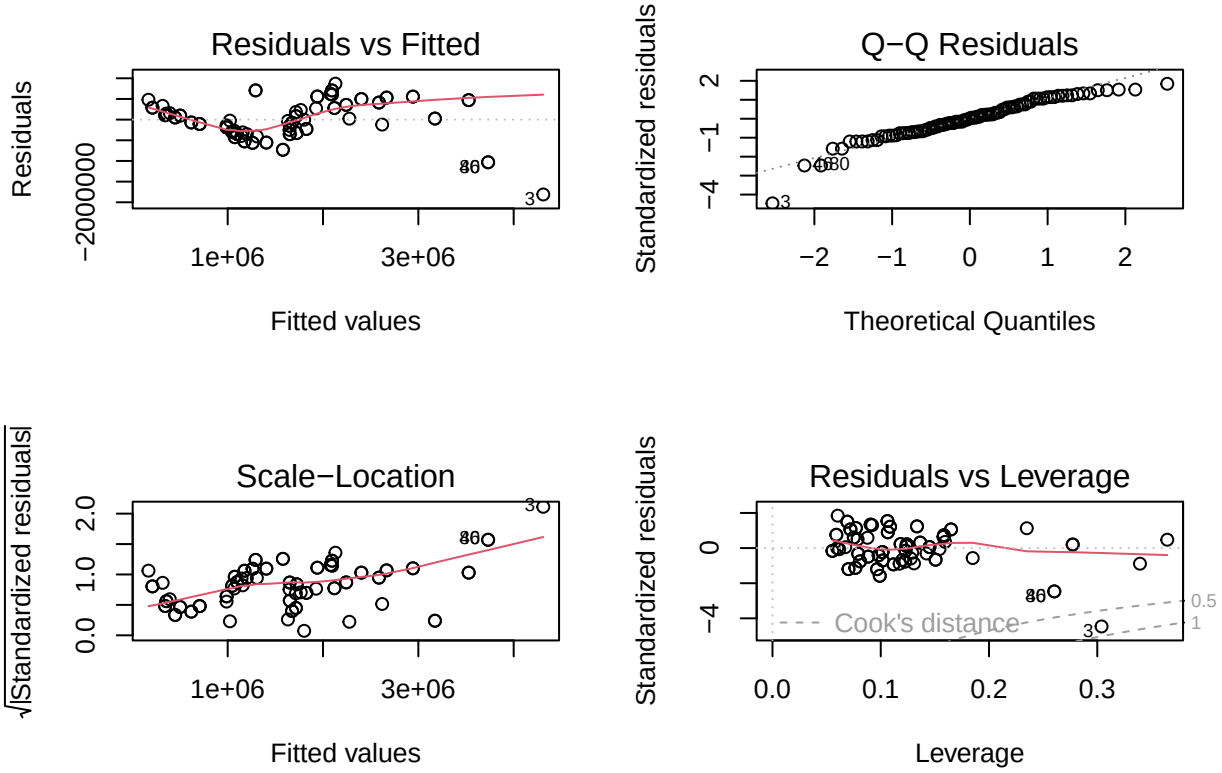
-**I(sqrt(Engine)):** The coefficient is positive and not statistically significant. This means that the engine capacity of a used car may not sufficiently capture predictions for used car price reliably.

-I(log(Kilometers_Driven)): The coefficient is negative and is statistically significant. This shows that as the kilometers already driven in a used car increases by one kilometer, the used car price may decrease in a non-linear fashion.

Adjusted R-Squared

The adjusted R-squared for the final model is 0.7689. This means that the independent variables in the model can explain 76.8% of the variation in the dependent variable, which is the price of a used car. This is a high value and shows that the model is powerful in its predictions for used car prices.

Outliers, Residual Plots, and Heteroscedasticity

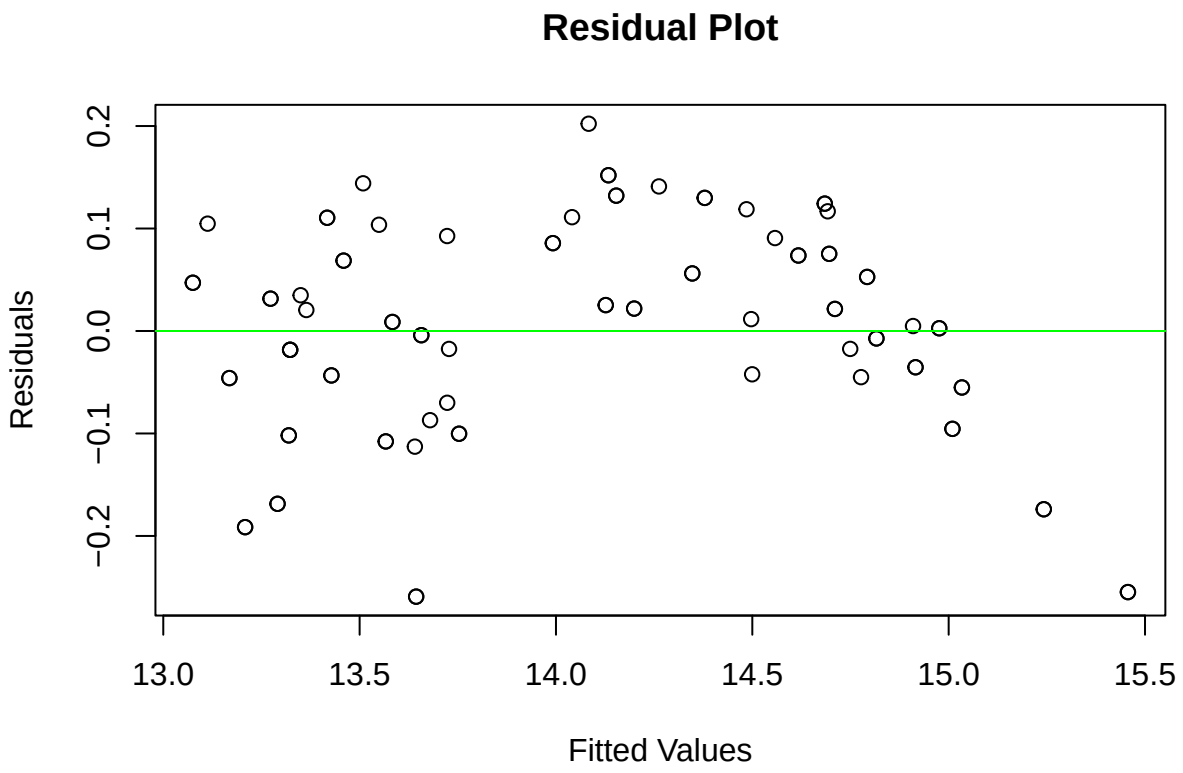



```
    studentized Breusch-Pagan test

data:  model6
BP = 31.93, df = 10, p-value = 0.0004114
```

The residual plots and Breusch-Pagan test show indications of **outliers** and minor **heteroscedasticity** present within the model. To resolve this, we log transform the dependent variable Price so that homoscedasticity is present in the model.

```
    studentized Breusch-Pagan test
```

```
data:  model6_transformed
BP = 11.101, df = 11, p-value = 0.4348
```



The Breusch-Pagan test now fails to reject the null hypothesis of homoscedasticity, which resolves the issue of the minor heteroscedasticity present in the final model. The residual plot also contains fewer outliers. The combination of these changes result in a final model that is much more reliable in its predictive ability.

Conclusions

The analysis in this report utilized a multiple linear regression model to predict used car prices. It revealed that the **power output** of a car is a core component of the price of a used car, and that other attributes of a used car can affect a used car's price in a nonlinear manner, such as I(sqrt(Engine)) or I(Mileageˆ2). Therefore, the results of this analysis shows that the **condition** that a car is in is much more important in determining the price of a used car in comparison to other factors such as the year it was manufactured or the number of previous owners.

A limitation of this model mainly involves its small sample size of 100 rows. A primary assumption of linear regression is that large sample sizes are necessary to ensure the stability in the estimation of regression parameters and to avoid overparameterization, so increasing the sample size is a way the model can be improved upon.

The used car market is constantly evolving due to high demand in the midst of an increasingly complex economic situation. The results of this analysis can be applied to the used car market to assist companies in India to determine a reasonable price for a used car. This can allow companies to take advantage of a data-driven decision making approach in which used cars can be priced in accordance with what consumers can afford.