DAT603

ASSESSMENT 3

STATISTICAL ANALYSIS OF USED CAR SELLING PRICES USING REGRESSION MODELS

Abstract

The study analyses the Car dataset from Kaggle to analyses the factors that influence car selling prices. “”Simple linear regression”” and multiple linear regression are used in R Studio for regression analysis. “”Simple regression”” model is used to assess the relationship between selling price and maximum power; multiple regression model is used with the maximum power, engine size and manufacturing year as the predictors. “”Diagnostic plots”” and statistical tests are used to evaluate the assumptions of regression: “”normality””, “”homoscedasticity””, and “”multicollinearity””. All the selected independent variables are found to have a significant effect on the selling price. Model 2 worked better for predicting the data, and thus gave a more accurate model of the used car market.

Table of Contents

1. Introduction and Objectives

Important statistical methods that are used to study the relationship between variables and predict a value are considered as “”’regression analysis’””. The “”’selling price’”’ of cars can be affected by factors such as “”’engine capacity’””, “”’selling age’”’ and “”’maximum power’”’ in the automobile industry. The data of cars is analyses with the help of R Studio from the Car dataset, which is downloaded from Kaggle, to explore the factors which influence the price of cars. Simple linear regression models and multiple linear regression models are developed to examine the relationship between the dependent variable (“”’selling price’”’) and selected independent variables. The study also analyses the regression assumptions and compares the performance of the models to find the best model to predict the selling price of used cars.

1.1 Aim and Objectives

Aim

The study aims to examine the factors affecting used car selling prices with the regression analysis techniques and test simple and multiple regression models using R Studio.

Objectives

- To use simple linear regression to determine the relationship between selling price and maximum power.
- To applying multiple linear regression analysis to assess the effect of engine size, maximum power and manufacturing year on selling price.
- To test and examine the assumptions of regression models, such as “”’normality’””, “”’homoscedasticity’””, “”’linearity’”” and “”’multicollinearity’”” using appropriate diagnostic plots and statistical tests.

- To compare the performance of the simple and multiple regression models using the measures of evaluation such as “”Adjusted R^2””, “”AIC””, “”BIC””, and “”RMSE””.

2. Literature Review

“”Regression analysis”” is a very common method to find the relationship between dependent variables and independent variables in data science and predictive modelling. (Lee et al. 2022) state that regression models can be used to explain the effect that changes in the explanatory variables have on a response variable. “”Regression”” is used widely to model prices for vehicles and to determine the factors that influence the market value of the car in the field of automobiles.

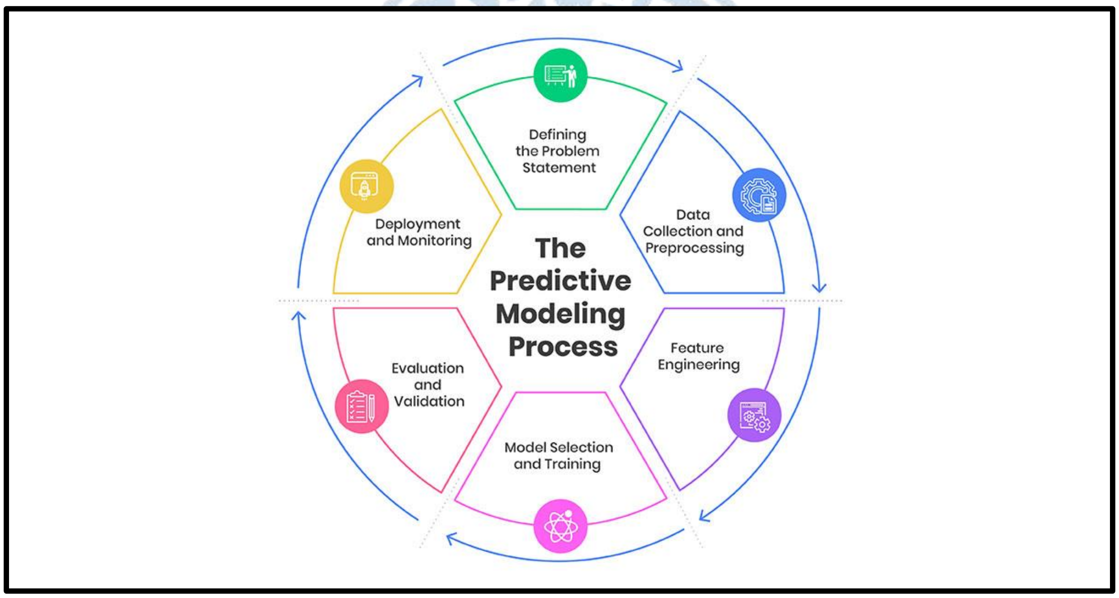


Figure 1: Predictive Modeling Process

(Source: dasca.org, 2025)

The authors have found that the selling price of a vehicle is significantly affected by various vehicle attributes, including the “”engine size””, “”horsepower””, “”mileage””, “”manufacturing year””, and “”fuel type””. Newer cars are more likely to have a higher resale value than older cars with less power (Bergmann & Feuerriegel, 2025). Strong connection between maximum power and engine capacity with the car prices in the used market, as it indicates the performance of the car and the consumer demand.

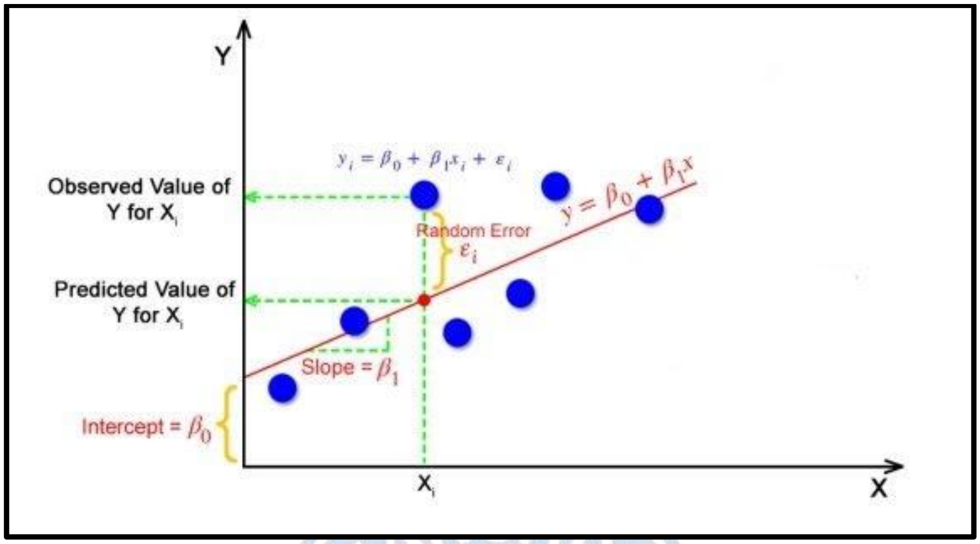


Figure 2: Simple Regression Calculation mechanism

(Source: analyticsvidhya.com, 2025)

“”Simple linear regression”” is commonly used to study the relationship between one predictor and a response variable, and “”multiple linear regression”” can be used to study the relationship between more than one predictor variable (James et al. 2023). “”Multiple regression”” models typically have higher predictive accuracy as they are able to account for a larger proportion of the variation in the dependent variable. The researchers also highlight the need to evaluate the regression assumptions such as “”normality””, “”homoscedasticity””, “”linearity””, and “”multicollinearity”” to ensure the model is valid (Lenz et al. 2026). The correct assessment of these assumptions increases the confidence and precision of the statistical results of predictive modelling.

3. Response to the Criteria

The study meets all the assessment criteria as it uses the regression analysis technique with the Car data set in R Studio. There are several continuous numerical variables, which can be used for statistical modelling and predictive analysis, in addition to the categorical variables. The relationship between “”selling price”” and “”maximum power”” is tested with a simple “”linear regression”” model, and the “”selling price”” is predicted from maximum “”power””, ””engine size””, and “”manufacturing year”” by a “”multiple linear regression model””. “”Regression”” assumptions such as “”normality””, “”homoscedasticity””, “”linearity””, and “”multicollinearity”” are tested using appropriate tests and plots (Tanni et al. 2023). The significance of independent variables is tested at the 5% level with the “”p-value”” of the regression output. Both regression models are evaluated using statistical procedures such as Adjusted “”R^2,””, “”AIC””, “”BIC””, and “”RMSE””. The analysis is implemented in a practical way to prove the use of regression in data analytics and predictive modelling.

4.1 Dataset Selection and Preparation

The Car dataset is collected from Kaggle, as it has relevant numerical variables that can be analyses using regression. Data preprocessing is carried out in R Studio, which includes managing missing values, renaming variables, categorizing variables into factors, and excluding incomplete observations to ensure the accurate and reliable analysis of data and the development of models.

```
> # Import Dataset
> car_data <- read.csv("cardekho.csv")
> # Viewing Dataset
> head(car_data)
                        name year selling_price km_driven   fuel
1        Maruti Swift Dzire VDI 2014        450000    145500 Diesel
2  Skoda Rapid 1.5 TDI Ambition 2014        370000    120000 Diesel
3      Honda City 2017-2020 EXi 2006        158000    140000 Petrol
4     Hyundai i20 Sportz Diesel 2010        225000    127000 Diesel
5        Maruti Swift VXI BSIII 2007        130000    120000 Petrol
6 Hyundai Xcent 1.2 VTVT E Plus 2017        440000     45000 Petrol
  seller_type transmission        owner mileage.km.ltr.kg. engine
1  Individual       Manual  First Owner              23.40   1248
2  Individual       Manual Second Owner              21.14   1498
3  Individual       Manual  Third Owner              17.70   1497
4  Individual       Manual  First Owner              23.00   1396
5  Individual       Manual  First Owner              16.10   1298
6  Individual       Manual  First Owner              20.14   1197
  max_power seats
1     74.00     5
2    103.52     5
3     78.00     5
4     90.00     5
5     88.20     5
6     81.86     5
```

```
> str(car_data)
'data.frame':	8128 obs. of  12 variables:
 $ name              : chr  "Maruti Swift Dzire VDI" "Skoda Rapid 1.5 TDI Ambi
tion" "Honda City 2017-2020 EXi" "Hyundai i20 Sportz Diesel" ...
 $ year              : int  2014 2014 2006 2010 2007 2017 2007 2001 2011 2013
...
 $ selling_price     : int  450000 370000 158000 225000 130000 440000 96000 45
000 350000 200000 ...
 $ km_driven         : int  145500 120000 140000 127000 120000 45000 175000 50
00 90000 169000 ...
 $ fuel              : chr  "Diesel" "Diesel" "Petrol" "Diesel" ...
 $ seller_type       : chr  "Individual" "Individual" "Individual" "Individua
l" ...
 $ transmission      : chr  "Manual" "Manual" "Manual" "Manual" ...
 $ owner             : chr  "First Owner" "Second Owner" "Third Owner" "First
Owner" ...
 $ mileage.km.ltr.kg.: num  23.4 21.1 17.7 23 16.1 ...
 $ engine            : num  1248 1498 1497 1396 1298 ...
 $ max_power         : num  74 103.5 78 90 88.2 ...
 $ seats             : num  5 5 5 5 5 5 5 4 5 5 ...
```

Figure 3: Dataset Loading and Structure Analysis

The function "'"*read.csv()*"" is used to easily import the dataset into R Studio. The output for the "'head(car_data)"" is shown in the output, the first 6 observations, and an overview of the variables and values of the dataset. The "'str(car_data)"" function revealed there are 12 variables including numerical and categorical types needed for regression and 8,128 observations in the data set.

```
     name                year       selling_price        km_driven
 Length:8128        Min.   :1983   Min.   :   29999   Min.   :      1
 Class :character   1st Qu.:2011   1st Qu.:  254999   1st Qu.:  35000
 Mode  :character   Median :2015   Median :  450000   Median :  60000
                    Mean   :2014   Mean   :  638272   Mean   :  69820
                    3rd Qu.:2017   3rd Qu.:  675000   3rd Qu.:  98000
                    Max.   :2020   Max.   :10000000   Max.   :2360457

     fuel           seller_type        transmission
 Length:8128        Length:8128        Length:8128
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character



    owner           mileage.km.ltr.kg.     engine        max_power
 Length:8128        Min.   : 0.00      Min.   : 624   Min.   :  0.00
 Class :character   1st Qu.:16.78      1st Qu.:1197   1st Qu.: 68.05
 Mode  :character   Median :19.30      Median :1248   Median : 82.00
                    Mean   :19.42      Mean   :1459   Mean   : 91.52
                    3rd Qu.:22.32      3rd Qu.:1582   3rd Qu.:102.00
                    Max.   :42.00      Max.   :3604   Max.   :400.00
                    NA's   :221        NA's   :221    NA's   :216
     seats
 Min.   : 2.000
 1st Qu.: 5.000
 Median : 5.000
 Mean   : 5.417
 3rd Qu.: 5.000
 Max.   :14.000
```

Figure 4: Summary Statistics

Summary statistics provide an overview of the data variables, such as minimum, maximum, mean, and median. The average selling price is around 638,272, and the average kilometers driven are around 69,820. The data also revealed differences in mileage, engine size, max power, and seating capacity. The distribution and nature of numeric variables are identified before regression analysis by these statistics.

```
> # Checking Missing Values
> colSums(is.na(car_data))
         name          year selling_price     km_driven          fuel 
            0             0             0             0             0 
  seller_type  transmission         owner       mileage        engine 
            0             0             0           221           221 
    max_power         seats 
          216           221 
> # Converting Variables to Correct Data Types
> car_data$fuel <- as.factor(car_data$fuel)
> car_data$seller_type <- as.factor(car_data$seller_type)
> car_data$transmission <- as.factor(car_data$transmission)
> car_data$owner <- as.factor(car_data$owner)
> # Removing Missing Values
> car_data <- na.omit(car_data)
> # Scatterplot Matrix
> numeric_data <- car_data %>%
+   select(selling_price, km_driven, mileage, engine, max_power, seats, year)
> ggpairs(numeric_data)
> cor_matrix <- cor(numeric_data)
```

Figure 5: Checking missing values and removing missing values

Missing value analysis is used, and it is found that some variables, such as “”mileage””, “”engine””, “”max_power””, and “”seats””, “”has missing”” values. The number of missing values per variable is calculated using the “”colSums(is.na(car_data))”” function. The “”na.omit()”” function is used to remove incomplete observations for accurate regression analysis. The data is preprocessed to enhance data quality and minimize the possibility of data errors in model development.

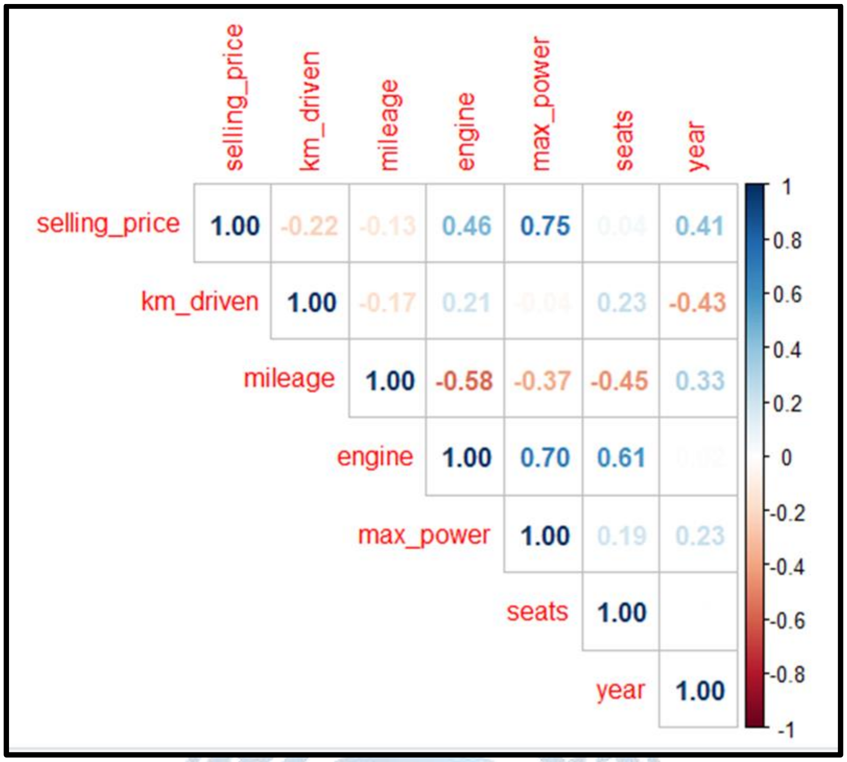


Figure 6: Correlation Matrix

Relationships among numerical variables in the data set are explored by using the correlation matrix. “”Selling price’”’ has a strong positive correlation with maximum “”power’”’ and “”engine size’”’, which is logical, since a larger engine size and higher power are often expected to correspond to a higher “”selling price’”’ for the vehicle. There are negative correlations between “”selling price’”’ and “”kilometres’”’ driven. The correlation analysis helped to choose suitable independent variables to be used in regression modelling and to establish meaningful relationships between variables.

4.2 Simple Linear Regression (Model 1)

```
> # Model 1: Simple Linear Regression
> model1 <- lm(selling_price ~ max_power, data = car_data)
> summary(model1)

Call:
lm(formula = selling_price ~ max_power, data = car_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2739158  -193351     5008   185683  4232623

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -912856.8    16656.7   -54.8   <2e-16 ***
max_power     17062.1      169.4   100.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 538500 on 7904 degrees of freedom
Multiple R-squared:  0.562,	Adjusted R-squared:  0.562
F-statistic: 1.014e+04 on 1 and 7904 DF,  p-value: < 2.2e-16
```

Figure 6: Simple Linear Regression Model

“”Simple linear”” regression model is developed to investigate the relationship between “”selling price”” and “”maximum power”” of cars. “”Selling_price”” is the dependent variable, and “”max_power”” is the independent variable in the model. Statistically significant is found when maximum power has a positive effect on “”selling price””, with a “”p-value”” less than 0.05 in the regression results. The result for the “”max_power”” coefficient is that the selling price increases with the increase in the maximum power. The maximum power gives an adjusted “”R^2“”of 0.562, which implies that about 56.2% of the variation in selling price can be explained by the model, and the overall model is found to have a moderate linear relationship.

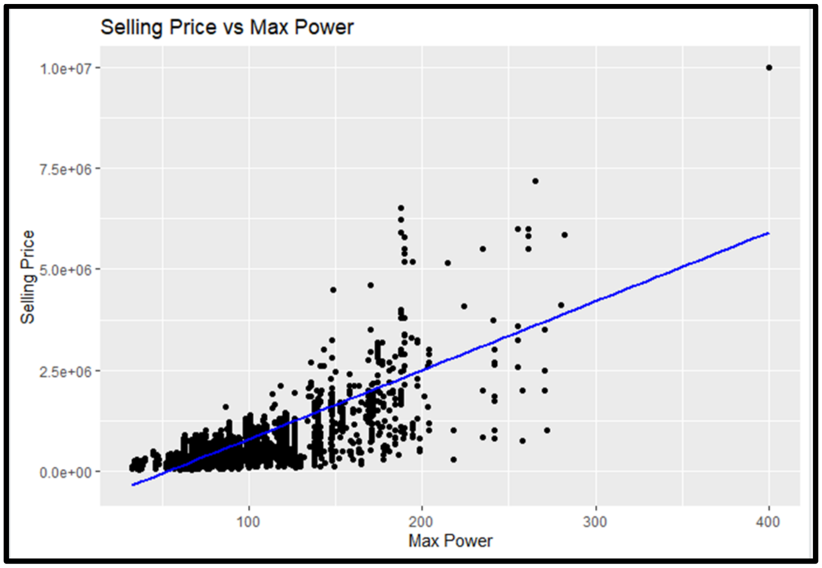


Figure 7: Selling Price vs Max Power

The scatter plot shows a positive linear association between the “”selling price”” and “”maximum power””. The higher the maximum power of the car, the higher its price. The regression line is a good fit and shows the same trend of increasing, which suggests that maximum power is indeed an important predictor of selling price for the regression model.

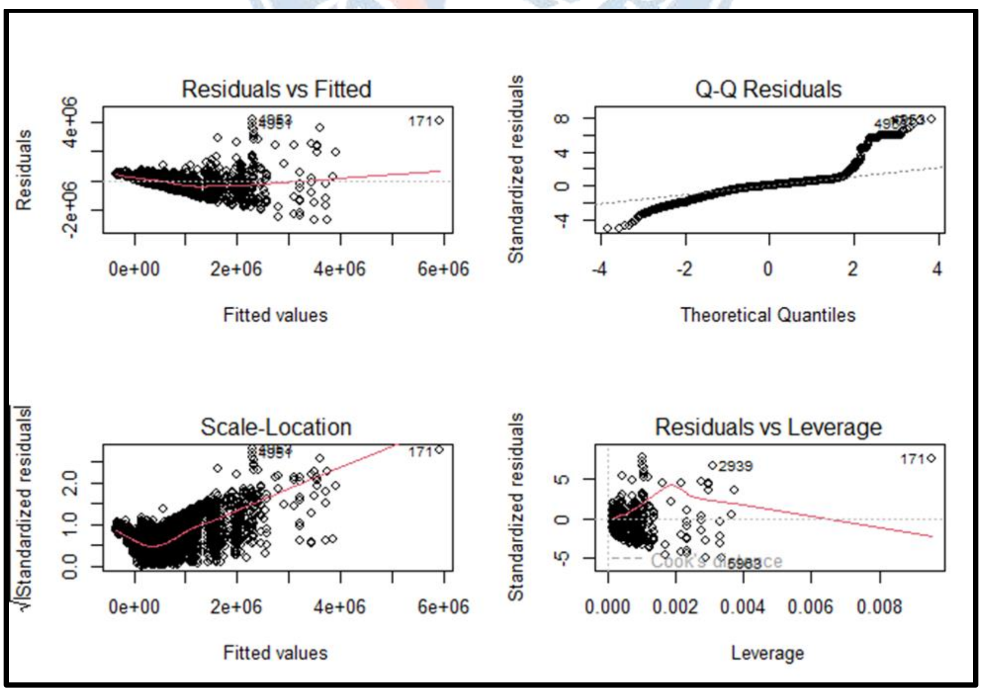


Figure 8: Residual Plots

Assumptions of the “”simple linear regression”” model are assessed with the residual plots. Plots showed some nonnormality and unequal variance of residuals and the residuals are mostly scattered around the regression line, and it indicating that there is a linear relationship between the variables.

```
> # Homoscedasticity Test
> bptest(model1)

        studentized Breusch-Pagan test

data:  model1
BP = 2247.4, df = 1, p-value < 2.2e-16
```

Figure 9: Homoscedasticity Test

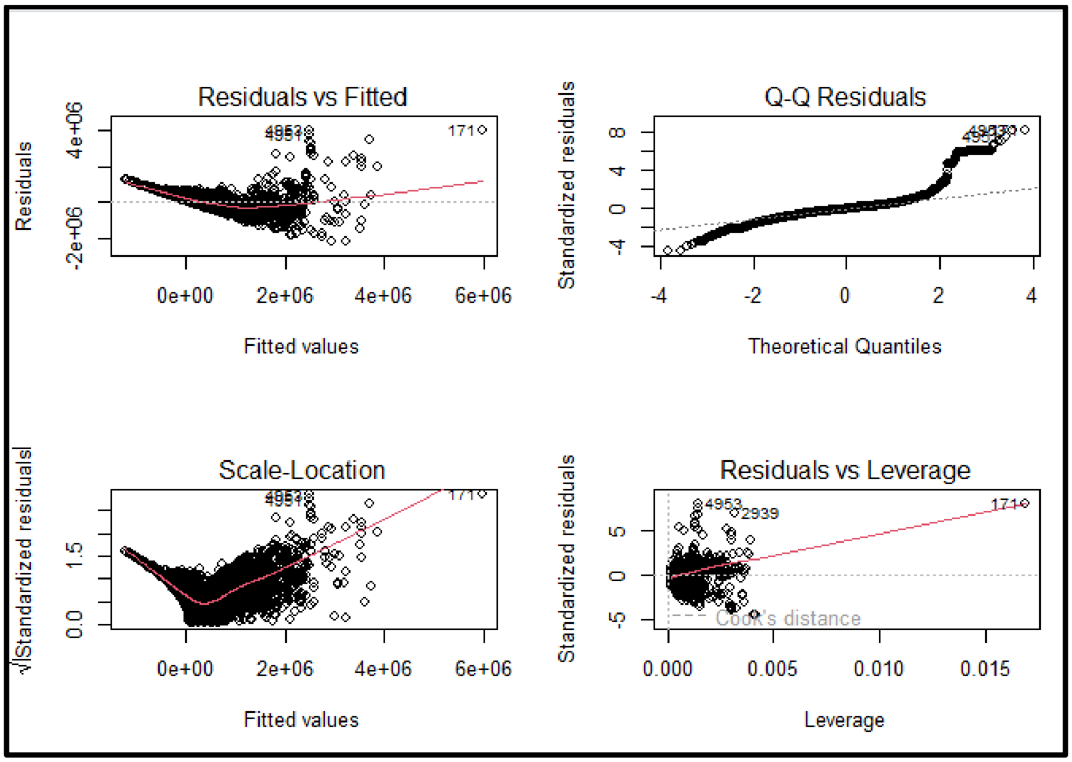
“”Breusch-Pagan”” test is used to test for the homoscedasticity of the regression model using R code. “”p-value”” smaller than 0.05, which shows that the residuals are not homoscedastic. It indicates heteroscedasticity of the errors, a frequent problem in many big real-world data sets.

4.3 Multiple Linear Regression (Model 2)

```
> # Model 2: Multiple Linear Regression
> model2 <- lm(selling_price ~ max_power + engine + year,
+              data = car_data)
> summary(model2)

Call:
lm(formula = selling_price ~ max_power + engine + year, data = car_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2214086  -215294   -33774   137456  4021758

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.044e+08  3.056e+06 -34.167  < 2e-16 ***
max_power    1.699e+04  2.310e+02  73.570  < 2e-16 ***
engine      -1.202e+02  1.596e+01  -7.527 5.77e-14 ***
year         5.148e+04  1.518e+03  33.910  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 497200 on 7902 degrees of freedom
Multiple R-squared:  0.6266,    Adjusted R-squared:  0.6265
F-statistic:  4421 on 3 and 7902 DF,  p-value: < 2.2e-16
```

Figure 8: Multiple Linear Regression Model

“”’Selling_price’”” is predictable from “”’max_power’””, “”’engine’””, and “”’year’”” as of the “”’multiple linear regression’”” model, which is developed. The model has an “”’Adjusted R^2“”’of 0.6265, meaning that 62.65% of the “”’selling price’”” variation is accounted for by the predictors. The ““‘p-values’”” of all independent variables are significant at the 0.05 level, which indicates that all the independent variables have contributed significantly to the prediction of the selling prices of cars.




```
> # Multicollinearity Test
> vif(model2)
max_power    engine      year
 2.179982  2.068737  1.099993
>
```

Figure 9: Diagnostic Plots with Multicollinearity Test for Model 2

Assumptions of regression are assessed using the diagnostic plots, such as normality, linearity, and influential observations. The plots showed minor deviations from normality and unequal variances of the residuals. The “”’*Variance Inflation Factor (VIF)*’”” values for all the independent variables (“”max_power””, “”engine, year””) in the regression model do not exceed 5, suggesting that there are no serious multicollinearity problems among these variables.

```
> # Homoscedasticity Test
> bptest(model2)

	studentized Breusch-Pagan test

data:  model2
BP = 2050.6, df = 3, p-value < 2.2e-16
```

Figure 10: Homoscedasticity Test

Model 2's homoscedasticity has been tested using the “”Breusch-Pagan”” test. The test yielded a “”p-value” of “< *0.05*””, which suggests that there is heteroscedasticity in the residuals. This indicates that the spread of the residuals is not equal for all fitted values. It occurs frequently in large data sets and does not have a major impact on the interpretation of the regression.

4.4 Comparison of Models

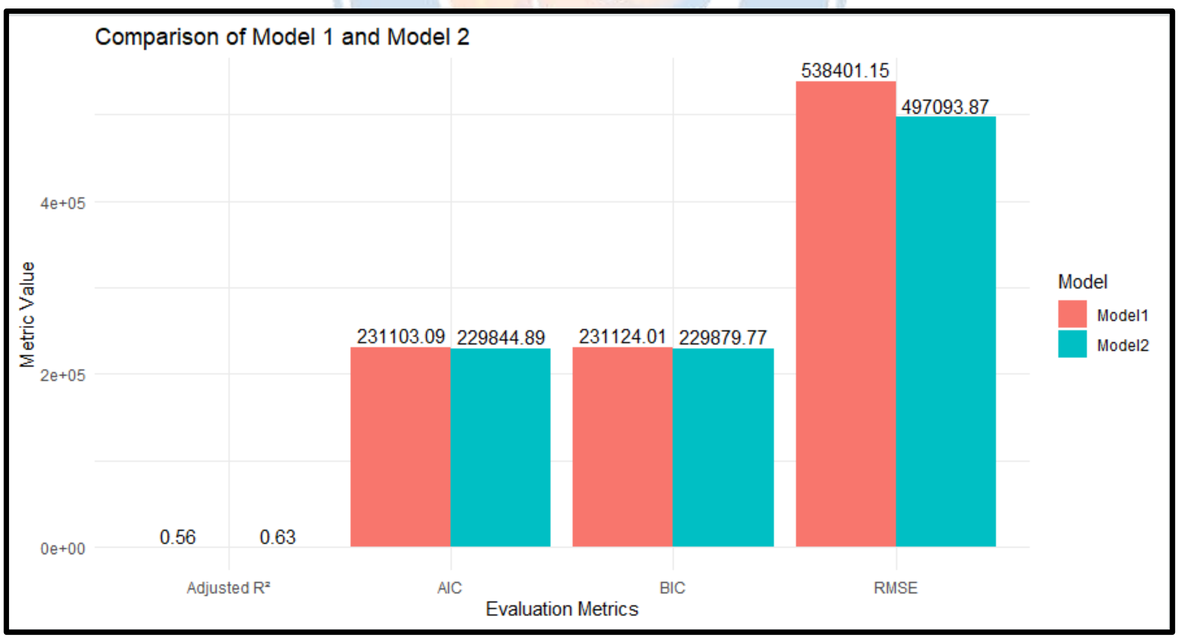


Figure 11: Comparison Graph of Models

Comparison bar graph is used and compared on the basis of “”Adjusted R^{2}’””, “”AIC’””, “”BIC’””, and “”RMSE”” values to assess the performance of “”Model 1’”” and “”Model 2’””. “”Model 2’”” has a higher “”Adjusted R^{2}“”and lower “”AIC’””, “”BIC’””, and “”RMSE”” values than Model 1. These results show that the multiple linear regression model is more accurate and has a better overall fit in predicting car selling prices.

```
> # Extract p-value for max_power
> p_value_mp <- summary(model1)$coefficients["max_power", "Pr(>|t|)"]
> # Set significance level
> alpha <- 0.05
> # Display p-value
> cat("P-value for max_power in Model 1:", p_value_mp, "\n")
P-value for max_power in Model 1: 0
> # Decision and Conclusion
> if (p_value_mp < alpha) {
+
+   cat("Conclusion: At 0.05 significance level, max_power significantly
predicts selling_price.\n")
+
+ } else {
+
+   cat("Conclusion: At 0.05 significance level, max_power does NOT signi
ficantly predict selling_price.\n")
+
+ }
Conclusion: At 0.05 significance level, max_power significantly predicts
selling_price.
```

Figure 12: Significance of Independent Variable in Simple Linear Regression

The significance test for the independent variable, “”max_power””, is carried out at the 0.05 level of significance. The “”p-value”” from the regression model is <0.05, which means that “”max_power”” is a significant factor in predicting “”selling_price””. The null hypothesis is rejected, and it is concluded that there is a statistically significant positive relationship between maximum power and the selling price of a car.

```
> # Extract p-values for independent variables
> p_value_max_power <- summary(model2)$coefficients["max_power", "Pr(>|t|)"]
> p_value_engine <- summary(model2)$coefficients["engine", "Pr(>|t|)"]
> p_value_year <- summary(model2)$coefficients["year", "Pr(>|t|)"]
> # Set significance level
> alpha <- 0.05
> # Display p-values
> cat("P-value for max_power:", p_value_max_power, "\n")
P-value for max_power: 0
> cat("P-value for engine:", p_value_engine, "\n")
P-value for engine: 5.767785e-14
> cat("P-value for year:", p_value_year, "\n\n")
P-value for year: 1.903419e-235

> # Conclusion for max_power
> if (p_value_max_power < alpha) {
+
+   cat("Conclusion: max_power significantly predicts selling_price.\n")
+
+ } else {
+
+   cat("Conclusion: max_power does NOT significantly predict selling_price.\n")
+
+ }
Conclusion: max_power significantly predicts selling_price.
> # Conclusion for engine
> if (p_value_engine < alpha) {
+
+   cat("Conclusion: engine significantly predicts selling_price.\n")
+
+ } else {
```

```
+
+   cat("Conclusion: engine does NOT significantly predict selling_price.\n")
+
+ }
Conclusion: engine significantly predicts selling_price.
> # Conclusion for year
> if (p_value_year < alpha) {
+
+   cat("Conclusion: year significantly predicts selling_price.\n")
+
+ } else {
+
+   cat("Conclusion: year does NOT significantly predict selling_price.\n")
+
+ }
Conclusion: year significantly predicts selling_price.
>
```

Figure 13: Significance of Independent Variables in Multiple Linear Regression

The significance tests for “”max_power””, “”engine””, and “”year”” are performed at the 0.05 level of significance. All the variables had a “”p-value”” less than 0.05, which means that the variable is useful for predicting “”selling_price””. The null hypotheses for all the predictors are rejected, and the results showed that selling prices for used cars are significantly affected by the following factors, such as “”maximum power’”’, “”engine size’”’, and “‘‘manufacturing year’’” (statistically).

4. Discussion / Analysis

The regression analysis showed that the selling prices of used cars are significantly affected by the car characteristics. “”Simple linear regression’’” indicated that there is a strong positive correlation between maximum “”power’”’ and “”selling price’”’ (Zhou et al. 2023). The multiple linear regression model also enhanced the prediction accuracy by adding the engine size and manufacturing year. Model 2 has a higher value of Adjusted “‘‘R^2’”’ and lower “‘‘RMSE’’” as compared to Model 1, which shows that the model performed better. “”Heteroscedasticity’’” is found, but models are still statistically significant and appropriate for predictive analysis. The results demonstrate that both vehicle “”age’”’ and “”engine performance’’” are important factors in estimating market value.

5. Conclusion

“”Simple’’” and “”multiple linear regression’’” methods are used in the study to analyze the factors that influence the selling price of used cars. The results indicated that the selling price is significantly affected by the factors of “”maximum power’”’, “”engine size’”’, and “”manufacturing year’”’. Model 2 showed a greater predictive ability, suggesting the use of multiple regression analysis for estimating vehicle market value is more effective.

6. References

Dataset Link: https://www.kaggle.com/datasets/sukhmandeepsinghbrar/car-price-prediction-dataset/data

analyticsvidhya.com, 2025, Retrieve From: https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/ , [Retrieve on 9th May, 2026]

Bergmann, S. and Feuerriegel, S., 2025. Machine learning for predicting used car resale prices using granular vehicle equipment information. Expert Systems with Applications, 263, p.125640. https://doi.org/10.1016/j.eswa.2024.125640

dasca.org, 2025, Retrieve From: https://www.dasca.org/world-of-data-science/article/predictive-modeling-types-and-algorithms-for-data-success , [Retrieve on 9th May, 2026]

James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J., 2023. Linear regression. In An introduction to statistical learning: With applications in python (pp. 69-134). Cham: Springer international publishing. https://doi.org/10.1007/978-3-031-38747-0_3

Lee, S.W., 2022. Regression analysis for continuous independent variables in medical research: statistical standard and guideline of Life Cycle Committee. Life cycle, 2. https://doi.org/10.54724/lc.2022.e3

Lenz, A. S., Haktanir, A., & Flores, D. (2026). Statistical Power, Reliability, and Model Assumptions in Multiple Linear Regression: Assessing and Addressing Common Issues. MeasureMent and evaluation in Counseling and developMent, 59(1), 94-107. https://doi.org/10.1080/07481756.2025.2589207

Tanni, K. F., Akter, Y., Jamal, M. K., Akter, S., & Patwary, M. J. (2023, December). Discovering Hidden knowledge and optimizing the model by analyzing linear regression assumptions. In 2023 26th International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE. https://doi.org/10.1109/ICCIT60459.2023.10441559

Zhou, J., Tuo, J., Wang, L., Shi, Y. and Zha, W., 2023. Calculation of electricity sales based on multi-factor correlation analysis. Frontiers in Energy Research, 11, p.1326100. https://doi.org/10.3389/fenrg.2023.1326100