CAR PRICE PREDICTION

ST 3008
Group Project

Contents

Introduction

In today's competitive market, predicting car prices has become vital for various stakeholders involved in the automobile industry such as manufacturers, dealers and customers. Therefore accurate price prediction becomes a crucial task as it is helpful for various decision-making processes, allowing individuals and organizations to make more informed decisions about investments, pricing strategies, purchases and other financial decisions. With the advent of big data and various statistical techniques, the ability to forecast has significantly improved.

The main objective of our project is to build a model using multiple linear regression to predict car prices based on the features. And also we focus on identifying important factors that influence car prices and by what amount, they influence the car price.

Background of the Dataset

The dataset we chose for the study was selected from Kaggle. It contains 8128 entries of used cars and 12 total columns. It has a rich amount of data to train a model and predict a wide range of outcomes. It includes various features related to cars such as the year of manufacture, kilometers driven, fuel type, seller type, transmission type, number of previous owners, mileage, and engine specifications. These attributes provide valuable insights into the factors influencing car prices.

Literature Review

Researchers have been studying the prices of automobiles, mainly directing efforts toward predicting the price of a second-hand car. They have used various techniques and factors that help to estimate the worth of a car.

One of the more recent studies by Galarraga et al. (2014), who observed how the energy efficiency labels for cars, which currently are A or B in Europe, may affect the price of these cars. They found that cars with higher energy efficiency scores sold for about 3% to 5.9% more than those with lower scores.

The findings of the study carried out by Prieto et al. (2015) The study reveals that individuals are not willing to pay a high price for a secondhand car if they perceive the information to be less reliable. On the other hand, they are ready to pay more if a car is more reliable. This shows that reliability is one of the main aspects playing into the second-hand car market.

Pal et al. (2018) used the Random Forest machine learning tool in predicting the prices of the cars and it gained a good test accuracy of 83.63%. Noor and Jan (2017) applied the multiple linear regression approach, and that also gave very accurate results. Overall, it can be summarized from these studies that the prices of cars depend on various factors

If we check analyses done on the same dataset that we chose, most of them have used machine learning techniques to achieve high accuracy while some of them used linear regression.

Although machine learning techniques help achieve high accuracy, the problem with these techniques is that they are less interpretable. A pricing model like this requires interpretability as it is important to know what features influence the price the most and how other features contribute to the outcome variable.

Descriptive Analysis

This analysis aims to identify patterns, relationships, and potential data quality issues by summarizing the key characteristics of the dataset, providing a foundation for the subsequent advanced modelling.

The dataset comprises 8128 records and 12 features, including both categorical and numerical variables.

Numerical Variables: Year, Selling Price, Km Driven, Mileage, Engine, Max Power, Seats

Categorical Variables: Name, Fuel, Seller Type, Transmission, Owner

Since we are interested in predicting the selling price, let's explore the relationship between the selling price and each independent variable (features)

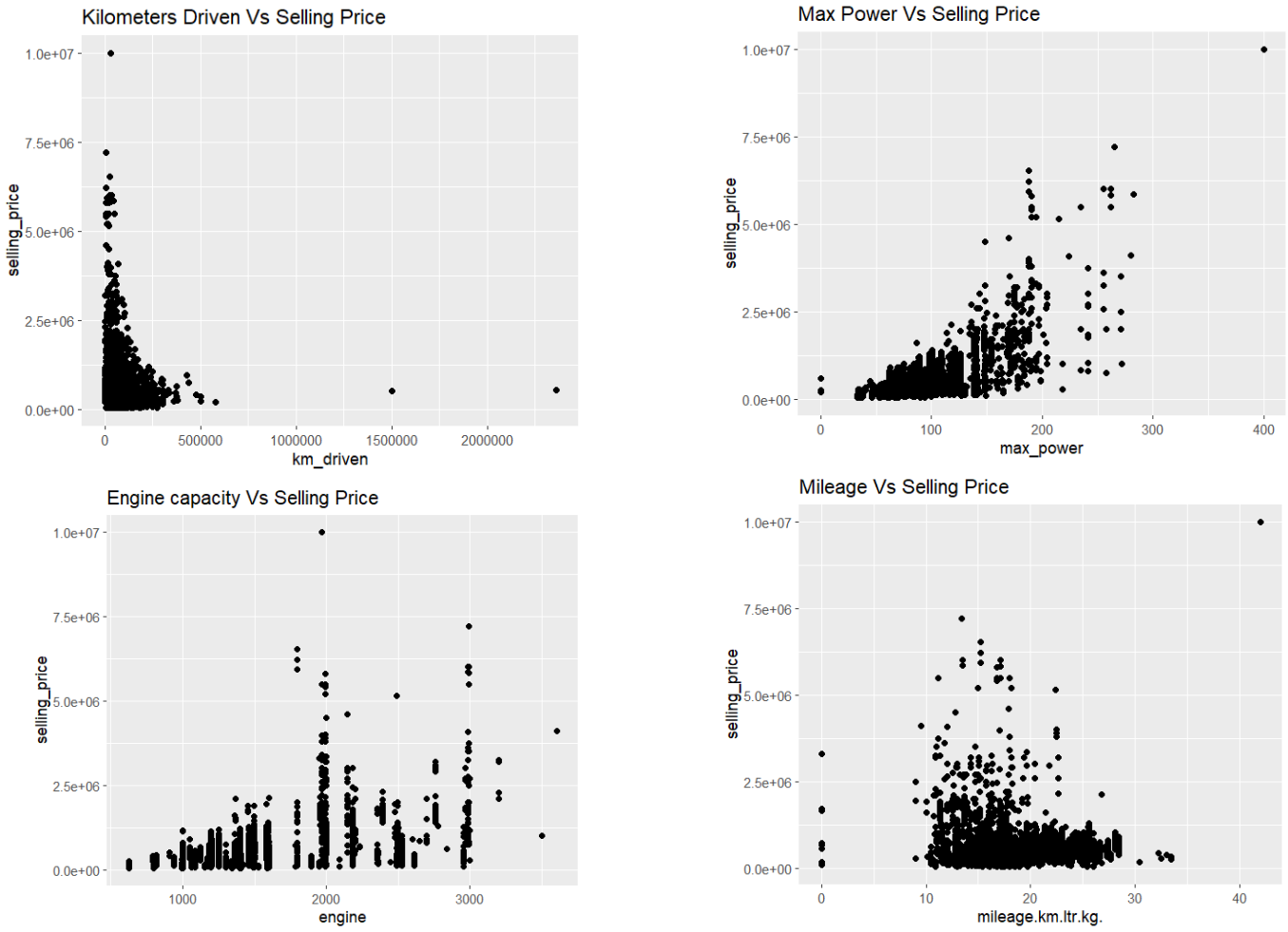
Handling missing values

In this dataset, we can find some missing values. We have to handle those missing values to improve the quality of the analysis

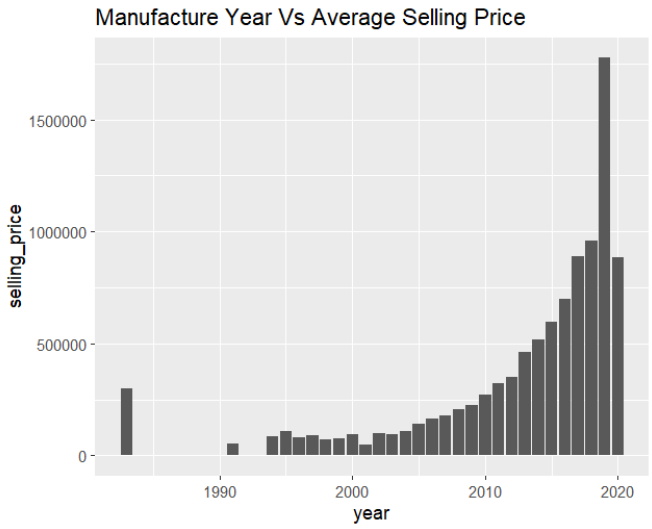
```
> colSums(is.na(cardekho))
              name               year 
                 0                  0 
     selling_price          km_driven 
                 0                  0 
              fuel        seller_type 
                 0                  0 
      transmission              owner 
                 0                  0 
mileage.km.ltr.kg.             engine 
               221                221 
         max_power              seats 
               216                221 
```

The output from the R function displays the count of missing values for various columns in the dataset. Only the columns "mileage.km.ltr.kg." and "max_power" columns show missing values. Since our dataset contains over 8000 entries, we will proceed to drop these rows with missing values to ensure the integrity of the analysis.

Relationship between numerical independent variables and response

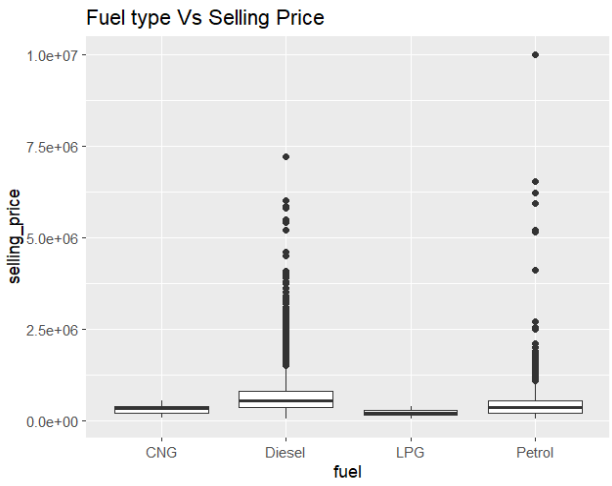


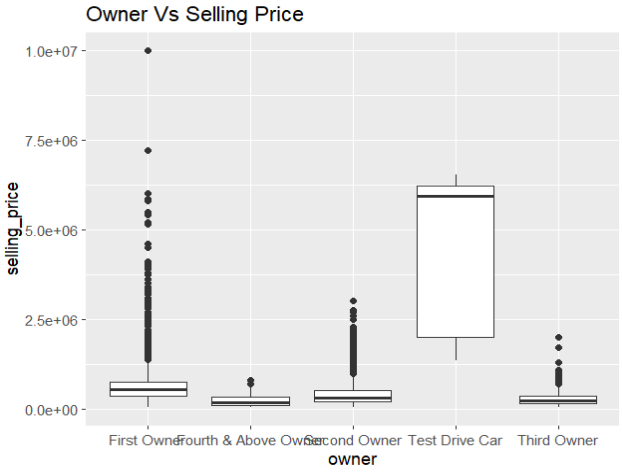
Here we can see kilometers driven, engine capacity and mileage do not exhibit a clear relationship with the selling price. But there's a positive correlation between max power and selling price, indicating an increase in max power increases the selling price. Also, we can observe that as the manufacturing year

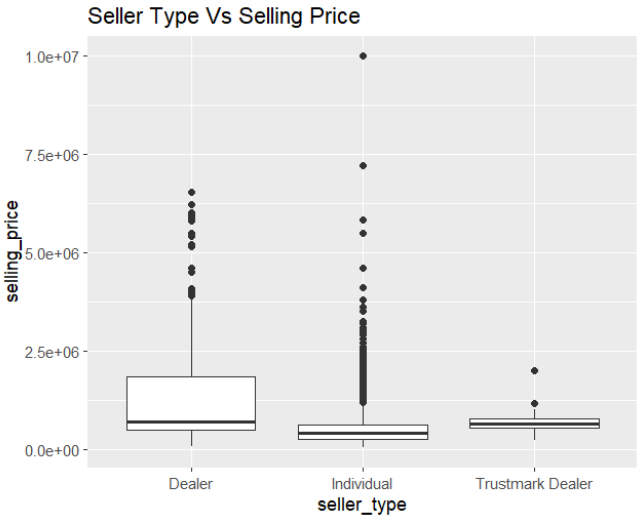


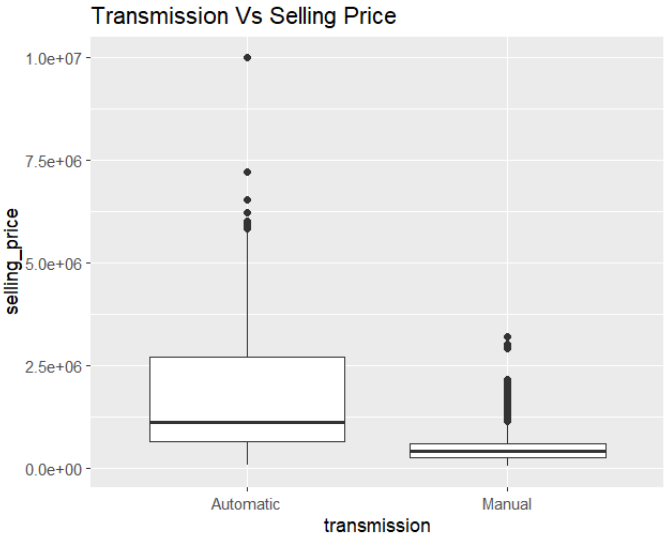
increases, the average selling price for that year also increases. This is strong evidence that recently manufactured cars have a higher selling price than old cars.

Relationship between categorical independent variables and response









From all the box plots we can see selling price has a right skewed distribution. Every plot compares different categorical variables against the selling price, indicating significant price variations among all those categories.

Advance analysis

In this section, we focus on building a predictive model. There are several methods for selecting variables for the model. Here we will use forward selection methods to determine what variables to include in our model. The typical approach for selecting variables to include in the model is to use p-values to determine the significance of variables. But in this situation, several variables have very small p-values. So we use the AIC value to determine the significance of variables.

- Null model

Variable	AIC value
year	236157
km_driven	237230
fuel	237289
seller_type	236204
transmission	234243
owner	237023
mileage.km.ltr.kg	237503
engine	235790
max_power	**231103**
seats	237616

Since 'max_power' has the minimum AIC value, it is the first variable to enter the model

➔ current AIC = 231103

- Null model + max_power

Variable	AIC value
year	**229899**
km_driven	230397
fuel	231091
seller_type	230429
transmission	230188
owner	230515
mileage.km.ltr.kg	230587
engine	230917
seats	230906

Since 'year' has the minimum AIC value and it is less than the current AIC value, it is the second variable to enter the model.

➔ current AIC = 229899

- Null model + max_power + year

Variable	AIC value
km_driven	229707
fuel	229890
seller_type	229415
transmission	**229166**
owner	229749
mileage.km.ltr.kg	229827
engine	229845
seats	229725

Since 'transmission' has the minimum AIC value and it is less than the current AIC value, it is the third variable to enter the model.

➔ current AIC = 229166

- Null model + max_power + year+transmission

Variable	AIC value
km_driven	229066
fuel	229143
seller_type	**228840**
owner	229032
mileage.km.ltr.kg	229061
engine	229151
seats	229100

Since 'seller_type' has the minimum AIC value and it is less than the current AIC value, it is the fourth variable to enter the model.

➔ current AIC = 228840

- Null model + max_power + year+transmission+seller_type

Variable	AIC value
km_driven	228764
fuel	228822
owner	**228735**
mileage.km.ltr.kg	228762
engine	228835
seats	228796

- Null model + max_power + year+transmission+seller_type+owner

Variable	AIC value
km_driven	228664
fuel	228712
mileage.km.ltr.kg	**228654**
engine	228731
seats	228693

Since ‘owner’ has the minimum AIC value and it is less than the current AIC value, it is the fifth variable to enter the model.

➔ current AIC = 228735

- Null model + max_power + year+transmission+seller_type+owner+ mileage.km.ltr.kg

Variable	AIC value
km_driven	**228584**
fuel	228645
engine	228652
seats	228648

Since ‘km_driven’ has the minimum AIC value and it is less than the current AIC value, it is the seventh variable to enter the model.

➔ current AIC = 228584

- Null model + max_power + year+ transmission+seller_type+ owner+mileage.km.ltr.kg+km_driven+fuel

Variable	AIC value
engine	228553
seats	**228547**

Since ‘seats’ has the minimum AIC value and it is less than the current AIC value, it is the ninth variable to enter the model.

➔ current AIC = 228547

Since ‘mileage.km.ltr.kg’ has the minimum AIC value and it is less than the current AIC value, it is the sixth variable to enter the model.

➔ current AIC = 228654

- Null model + max_power + year+transmission+seller_type+owner+mileage.km.ltr.kg+km_driven

Variable	AIC value
fuel	**228555**
engine	228566
seats	228585

Since ‘fuel’ has the minimum AIC value and it is less than the current AIC value, it is the eighth variable to enter the model.

➔ current AIC = 228555

- Null model + max_power + year+transmission+seller_type+owner+ mileage.km.ltr.kg+km_driven+fuel+seats

Variable	AIC value
engine	**228534**

Since the AIC value of ‘engine’ is less than the current AIC value, it is the tenth variable to enter the model.

Therefore, the final model can be written as:

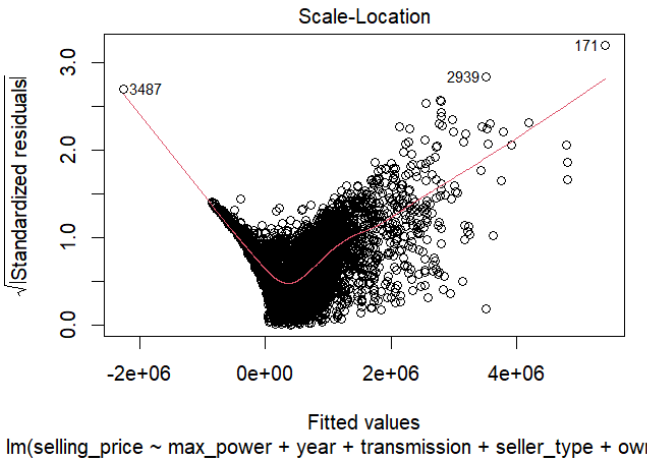
selling_price = max_power + year + transmission + seller_type + owner + mileage.km.ltr.kg + km_driven + fuel + seats + engine

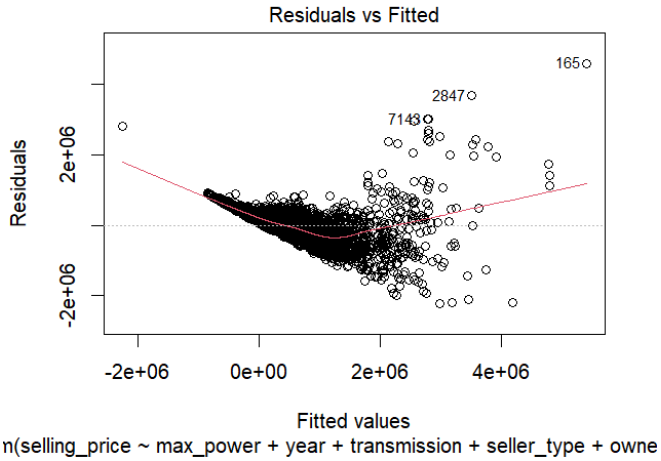
```
> summary(lm(selling_price~max_power+year+transmission+seller_type+owner+mileage.km.ltr.kg.+km_driven+fuel+seats+engine))

Call:
lm(formula = selling_price ~ max_power + year + transmission +
    seller_type + owner + mileage.km.ltr.kg. + km_driven + fuel +
    seats + engine)

Residuals:
     Min       1Q   Median       3Q      Max
-2235266  -196332     6419   149763  4594428

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -6.598e+07  3.822e+06 -17.262  < 2e-16 ***
max_power                    1.264e+04  2.602e+02  48.583  < 2e-16 ***
year                         3.280e+04  1.910e+03  17.170  < 2e-16 ***
transmissionManual          -4.595e+05  1.970e+04 -23.331  < 2e-16 ***
seller_typeIndividual       -2.454e+05  1.648e+04 -14.889  < 2e-16 ***
seller_typeTrustmark Dealer -3.506e+05  3.372e+04 -10.397  < 2e-16 ***
ownerFourth & Above Owner    1.184e+04  3.831e+04   0.309 0.757261
ownerSecond Owner           -4.337e+04  1.333e+04  -3.253 0.001145 **
ownerTest Drive Car          2.117e+06  2.052e+05  10.316  < 2e-16 ***
ownerThird Owner            -1.870e+04  2.293e+04  -0.815 0.414841
mileage.km.ltr.kg.           1.205e+04  2.135e+03   5.642 1.74e-08 ***
km_driven                   -1.038e+00  1.087e-01  -9.554  < 2e-16 ***
fuelDiesel                  -3.515e+04  6.434e+04  -0.546 0.584845
fuelLPG                      1.708e+05  1.007e+05   1.696 0.089891 .
fuelPetrol                  -8.749e+04  6.471e+04  -1.352 0.176411
seats                       -3.592e+04  7.988e+03  -4.497 7.00e-06 ***
engine                       9.021e+01  2.382e+01   3.788 0.000153 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 457300 on 7889 degrees of freedom
Multiple R-squared:  0.6847,    Adjusted R-squared:  0.6841
F-statistic:  1071 on 16 and 7889 DF,  p-value: < 2.2e-16
```

Using these coefficients, we can write our model as,

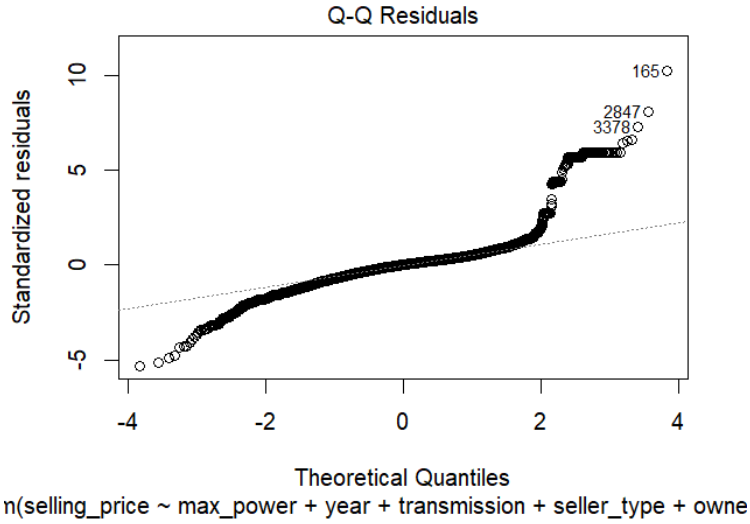
selling_price = $-6.598*10^7 + 1.264*10^4$ (max_power) + $3.280*10^4$ (year) - $4.595*10^5$ (transmissionManual) - $2.454*10^5$ (seller_typeIndividual) - $3.506*10^5$ (seller_typeTrustmark Dealer) + $1.184*10^4$(ownerFourth & Above Owner) - $4.337*10^4$ (ownerSecond owner) + $2.117*10^6$ (ownerTest DriveCar) - $1.870*10^4$(ownerThird owner) + $1.205*10^4$ (mileage.km.ltr.kg) - 1.038 (km_driven) $-3.515*10^4$ (fuelDiesel) + $1.708*10^5$(fuelLPG) - $8.749*10^4$(fuelPetrol) - $3.592*10^4$ (seats) + 90.21 (engine)

Checking the assumptions of the Multiple Linear Regression Model





- The scatterplots of fitted values Vs residuals and fitted values Vs standardized residuals show a non-random pattern, indicating that the linearity assumption is violated. It can be observed that as the size of the fitted values increases, the spread of the residuals increases, suggesting that the assumption of homoscedasticity is violated. Since there's a curve shape present in the plot, we can say the model is not adequate.



- This Q-Q plot exhibits significant deviation from the straight line, indicating that the assumption of residuals normally distributed is violated

Since this model violates many of the linear regression assumptions, some kind of transformation for the response variable is essential to reduce heteroscedasticity, normalize residuals and improve linearity. In here, to achieve this we apply log transformation to the response variable.

Form of the log-transformed model,

log(selling_price) =

max_power + year + transmission + seller_type + owner + mileage.km.ltr.kg

+ km_driven + fuel + seats + engine

```
> summary(lm(log(selling_price)~max_power+year+transmission+seller_type+owner+mileage.km.ltr.kg.+km_driven+fuel+seats+eng
ine))

Call:
lm(formula = log(selling_price) ~ max_power + year + transmission +
    seller_type + owner + mileage.km.ltr.kg. + km_driven + fuel +
    seats + engine)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77155 -0.17416  0.02515  0.19318  2.07744

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -2.104e+02  2.481e+00 -84.777  < 2e-16 ***
max_power                   1.005e-02  1.689e-04  59.495  < 2e-16 ***
year                        1.102e-01  1.240e-03  88.858  < 2e-16 ***
transmissionManual         -1.959e-01  1.279e-02 -15.325  < 2e-16 ***
seller_typeIndividual      -1.182e-01  1.070e-02 -11.050  < 2e-16 ***
seller_typeTrustmark Dealer -1.764e-03  2.189e-02  -0.081  0.93579
ownerFourth & Above Owner  -1.277e-01  2.487e-02  -5.135 2.89e-07 ***
ownerSecond Owner          -6.930e-02  8.653e-03  -8.009 1.32e-15 ***
ownerTest Drive Car         6.001e-01  1.332e-01   4.504 6.75e-06 ***
ownerThird Owner           -1.023e-01  1.489e-02  -6.872 6.82e-12 ***
mileage.km.ltr.kg.          1.100e-02  1.386e-03   7.934 2.42e-15 ***
km_driven                  -3.219e-07  7.055e-08  -4.563 5.11e-06 ***
fuelDiesel                  2.123e-01  4.177e-02   5.082 3.82e-07 ***
fuelLPG                     2.131e-01  6.537e-02   3.260  0.00112 **
fuelPetrol                  5.952e-02  4.201e-02   1.417  0.15658
seats                       2.549e-02  5.185e-03   4.916 8.99e-07 ***
engine                      2.372e-04  1.546e-05  15.341  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2968 on 7889 degrees of freedom
Multiple R-squared:  0.8716,    Adjusted R-squared:  0.8714
F-statistic:  3347 on 16 and 7889 DF,  p-value: < 2.2e-16
```

Using these coefficients, we can write our final model as,

log(selling_price) = -210.4 + 0.01005(max_power) + 0.1102(year) - 0.1959(transmissionManual) +

0.1182(seller_typeIndividual) - $1.764*10^{-3}$(seller_typeTrustmark Dealer) -

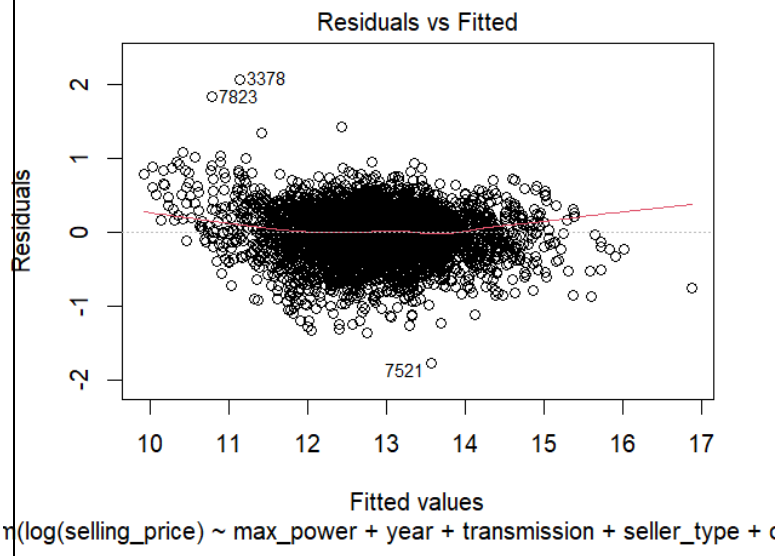
0.1277(ownerFourth & Above Owner) - 0.0693(ownerSecond Owner) +

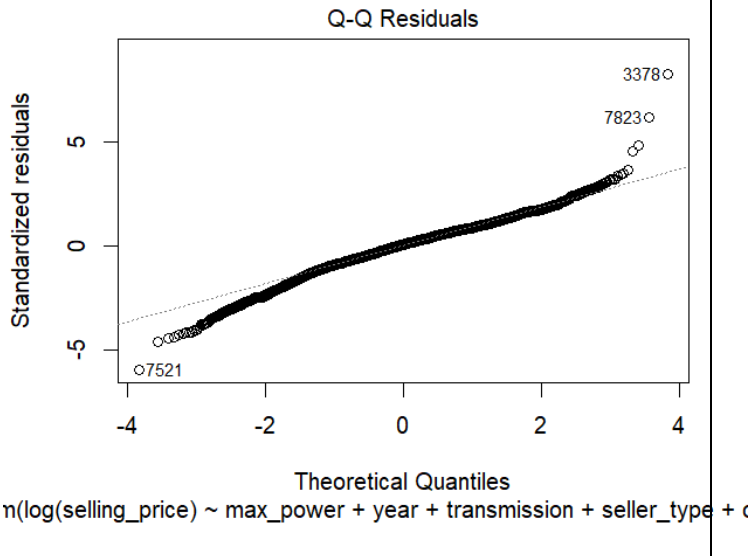
0.6001(ownerTest Drive Car) - 0.1023(ownerThird Owner) + 0.011(mileage.lm.ltr.kg.) -

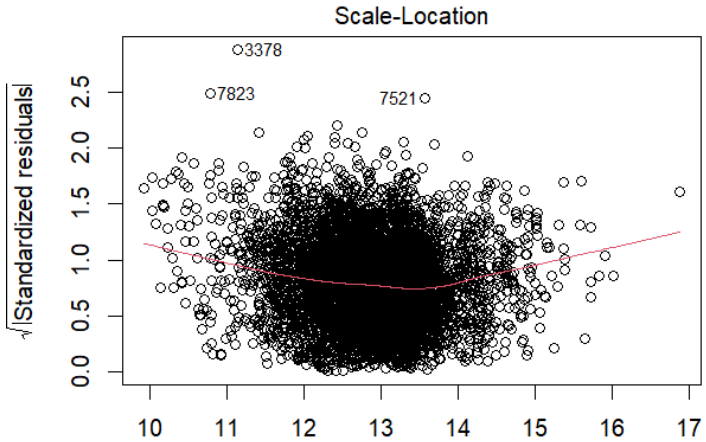
$3.219*10^{-7}$(km_driven) + 0.2123(fuelDiesel) + 0.2131(fuelLPG) + 0.05952(fuelPetrol) +

0.02549(seats) + $2.372*10^{-4}$(engine)

Checking the assumptions of the log-transformed Multiple Linear Regression Model







- Both the scatterplots of fitted values Vs residuals and fitted values Vs standard residuals now exhibit a random pattern. Thus, the linearity assumption and homoscedasticity assumption are satisfied.

- Since most of the data points in the Q-Q plot lie closely to the straight line, residuals are normally distributed.

Since all of the assumptions of linear regression are satisfied, the final model is the log-transformed model.

Check for multicollinearity

```
> vif(mlr)
                      GVIF Df GVIF^(1/(2*Df))
max_power         3.270681  1        1.808502
year              2.059160  1        1.434977
transmission      1.677035  1        1.295004
seller_type       1.324112  2        1.072707
owner             1.422323  4        1.045020
mileage.km.ltr.kg. 2.808439  1        1.675840
km_driven         1.440177  1        1.200074
fuel              2.168595  3        1.137705
seats             2.219208  1        1.489701
engine            5.444341  1        2.333311
```

In order to detect multicollinearity, we use Variance Inflation Factor (VIF) values. A VIF value greater than 10 typically indicates a high level of multicollinearity. Since none of the VIF values in our model exceed this threshold, we can conclude that there is no significant multicollinearity among the predictors.

Interpretation of the model coefficients

In this model,

The dependent variable is the natural logarithm of the selling price (log(selling_price))).

The independent variables are the factors max_power, year etc.

Each coefficient of a variable represents the expected change in the dependent variable (log(selling_price)) for a one-unit change in that independent variable, while keeping all other variables constant.

(1) Intercept (-210.4)

This is the baseline value for log(selling_price) when all the independent variables are zero.

(2) max_power (0.01005)

For each additional one unit increment in max power, the logarithm of the selling price will be increased by 0.01005, while keeping all other variables constant.(For each additional one unit increment in max power, the selling price will be increased by 1.005%, while keeping all other variables constant.)

(3) year (0.1102)

For each additional year increment, the logarithm of selling price will be increased by 0.1102, while keeping all other variables constant.(For each additional year increment, the selling price will be increased by 11.02%, while keeping all other variables constant.)

(4) transmissionManual (-0.1959)

If the car has a manual transmission instead of an automatic transmission, the log of selling price will be decreased by 19.59%, while keeping all other variables constant.

(5) seller_typeIndividual (0.1182)

If the car is sold by an individual instead of a dealer, the log of selling price will be increased by 11.82%, while keeping all other variables constant.

(6) seller_typeTrustmark Dealer (-1.764×10^{-3})

If the car is sold by a trustmark dealer instead of a dealer, the log of selling price will be decreased by 0.1764%, while keeping all other variables constant.

(7) ownerFourth & Above Owner (-0.1277)

If the car has had four or more owners instead of having the first owner, the log of selling price will be decreased by 12.77%, while keeping all other variables constant.

(8) ownerSecond Owner (-0.0693)

If the car has had a second owner instead of having the first owner, the log of selling price will be decreased by 6.93%, while keeping all other variables constant.

(9) ownerTest Drive Car (0.6001)

If the car has had a test drive owner instead of having the first owner, the log of selling price will be increased by 60.01%, while keeping all other variables constant.

(10) ownerThird Owner (-0.1023)

If the car has had a third owner instead of having the first owner, the log of selling price will be decreased by 10.23%, while keeping all other variables constant.

(11) mileage.km.ltr.kg. (0.011)

If we increase mileage by one unit, the log of selling price will be increased by 1.1%, while keeping all other variables constant.

(12) km_driven (-3.219×10^{-7})

If we increase kilometers driven by one, the log of selling price will be decreased by $3.219 \times 10^{-}$5%, while keeping all other variables constant.

(13) fuelDiesel (0.2123)

If the car uses diesel instead of using CNG, the log of selling price will be increased by 21.23%, while keeping all other variables constant.

(14) fuelLPG (0.2131)

If the car uses LPG instead of using CNG, the log of selling price selling price will be increased by 21.31%, while keeping all other variables constant.

(15) fuelPetrol (0.05952)

If the car uses Petrol instead of using CNG, the log of selling price will be increased by 5.952%, while keeping all other variables constant.

(16) seats (0.02549)

For each additional one seat increment in the car, the log of selling price will be increased by 0.02549, while keeping all other variables constant.(For each additional one seat increment in the car, the selling price will be increased by 2.549%, while keeping all other variables constant.)

(17) engine (2.372×10^{-4})

For each additional one unit increment in engine size, the log of selling price will be increased by 2.372×10^{-4}, while keeping all other variables constant.(For each additional one unit increment in engine size, the selling price will be increased by $2.372 \times 10^{-}$6%, while keeping all other variables constant.)

Categorize the response variable

Apart from directly predicting the price of the car, we can categorize the response variable into price ranges, and predict the price range based on the features. This will help simplify the prediction and enhance interpretability.

Categorization method

The selling price was categorized into three distinct groups:

Low Price: Cars priced below $200,000

Middle Price: Cars priced between $200000 and $1,000,000

High Price: Cars priced above $1,000,000

These categories were chosen based on the distribution of the price variable, as well as industry standards for car pricing.

As previously, the forward selection method based on AIC value is used for the variable selection.

```
> summary(polr_model)

Re-fitting to get Hessian

Call:
polr(formula = selling_price_cat ~ max_power + year + transmission +
    owner + mileage.km.ltr.kg. + log(km_driven) + fuel + seats +
    engine, method = "logistic")

Coefficients:
                            Value Std. Error  t value
max_power                0.066030  0.0024561   26.884
year                     0.687423        NaN      NaN
transmissionManual      -0.935868  0.0039918 -234.449
ownerFourth & Above Owner -1.023506  0.0048076 -212.891
ownerSecond Owner        -0.527539  0.0887468   -5.944
ownerTest Drive Car       9.183922        NaN      NaN
ownerThird Owner         -0.854441  0.0261129  -32.721
mileage.km.ltr.kg.        0.105547        NaN      NaN
log(km_driven)           -0.403291  0.0199153  -20.250
fuelDiesel                0.811064  0.0301242   26.924
fuelLPG                   1.276708        NaN      NaN
fuelPetrol                0.544496  0.0171125   31.819
seats                     0.116906  0.0522745    2.236
engine                    0.002281  0.0001747   13.054

Intercepts:
             Value      Std. Error t value
low|middle   1386.9774        NaN        NaN
middle|high  1397.8904     0.1249 11190.3729

Residual Deviance: 3839.76
AIC: 3871.76
```

Using these coefficients we can write our model as,

For "Low | Middle" Category:

logit(P(selling_price_cat ≤ middle)) = 1386.98 + 0.066(max_power) + 0.687(year) -

0.936(transmissionManual) - 1.024(ownerForth and Above) - 0.528(ownerSecond Owner) +

9.184(ownerTest Drive Car) - 0.854(ownerThird Owner) + 0.106(mileage.km.ltr.kg) –

0.403(log(km_driven)) + 0.811(fuelDiesel) + 1.277(fuelLPG) + 0.544(fuelPetrol) +

0.117(seats) + 0.002(engine)

logit(P(selling_price_cat ≤ high)) = 1397.89 + 0.066(max_power) + 0.687(year) -

0.936(transmissionManual) - 1.024(ownerForth and Above) - 0.528(ownerSecond Owner) +

9.184(ownerTest Drive Car) - 0.854(ownerThird Owner) + 0.106(mileage.km.ltr.kg) –

0.403(log(km_driven)) + 0.811(fuelDiesel) + 1.277(fuelLPG) + 0.544(fuelPetrol) +

0.117(seats) + 0.002(engine)

Discussion and Conclusion

The objective of this project was to predict car selling prices using a dataset which includes various car features. The analysis employed both descriptive statistics and advanced modeling techniques, the multiple linear regression and proportional odds model for ordinal logistic regression.

Descriptive analysis helped in understanding the distribution of car prices and other characteristics of the dataset. In multiple linear regression it was shown that all the variables are significant in predicting the car price. Here we applied log transformation to the response variable to better meet the assumptions of linear regression, such as normality of residuals and homoscedasticity, thereby enhancing the model's reliability and interpretability.

In advance analysis, we obtain a multiple linear regression model to predict car prices as well as we implemented a proportional odds logistic regression model to predict the car price category (low, middle, high). The model's performance, measured by the Akaike Information Criterion (AIC) and residual deviance, suggests a good fit for the data, with significant coefficients for most predictors. Accessing the goodness of the fit, the multiple linear regression model has a R^2 value of 87.16 %, suggesting it is a good model.

The study successfully demonstrated the application of statistical and machine learning techniques to predict car prices based on various features. The compatibility of the results with market expectations is due to the fact that such factors as vehicle age, power, and previous ownership history have considerable significance on selling prices. The proportional odds logistic regression model worked well in classifying cars based on price to be used by potential buyers, sellers, and market analysts.

Dataset information

• Kaggle link to the data set:

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

Contribution to the project

Index	Name	Task Completed
s16341	T.P.T.D.Panditawatta	Descriptive Analysis, Advanced Analysis
s16344	P.G.D.Perera	Descriptive Analysis, Advanced Analysis
s16365	A.V.A.Silva	Literature Review, Advanced Analysis
s16370	E.M.V.N.Thismalpola	Introduction, Discussion and Conclusion