Car Prices Analysis

## Abstract:

Geely Auto, a Chinese manufacturer, needs to break into the US market by establishing an assembly plant and producing automobiles domestically in order to compete with its American and European counterparts. They've hired an automobile consulting business to help them understand the factors that influence vehicle costs. They are particularly interested in learning about the factors that influence car costs in the American market, as they may differ significantly from those in China. The data is taken from the Kaggle's website and can be found at <https://www.kaggle.com/datasets/hellbuoy/car-price-prediction/download>. Our main goal in this project is to analyze the data and find some interesting insights and get an idea about what factors affect the prices of a car. This research will help the company in making business strategies and a car buyer in deciding on a car for himself based on his budget, and he could get an idea of what type of specifications he/she will get in his/her budget.

## Introduction:

In light of a few market research, the counseling organization has accumulated a massive data collection on various types of autos across the American market (Zuboff, 2015). We should show the estimate of autos using the offered free factors. The executives will utilize it to determine how expenses fluctuate in response to various elements. They can then alter the vehicle's plan, business strategy, and other elements to meet predetermined evaluation objectives. Furthermore, the approach would aid executives in comprehending the value features of a different market.

## Literature Review:

This research article by Nakagawa et al. (2017) solved the same problem with the linear regression model, and the base model's R squared value obtained using his method is 83%. Parameter tuning is then performed, and finally, 6 variables are selected, which are used to predict the prices of the car. The final model's R square thus obtained is 88% in this case (Pan et al., 2018). The main research questions that would be answered through this project are: Which variables are important in forecasting a car's price, how well those factors accurately represent a car's pricing, and What factors affect the car prices positively or negatively. ## Theory:

In this project, We hope to find some interesting facts about whether the type of car, such as sedan, hatchback, etc., has any effect on car prices or not. Do the city and highway mileage of a car and company size of a millage affects the car prices or not? Is there any evidence of horsepower affecting car prices in a positive or a negative way? Do the number of cylinders in a car affects the car prices? What is the accuracy of our forecasting model for predicting the prices of cars? All of these hypotheses will be analyzed and answered using correlation metrics, visualizations, and linear regression.

## Data:

The data has been downloaded from Kaggle, <https://www.kaggle.com/datasets/hellbuoy/car-price-prediction/download>. We have a total of 26 attributes in the data set, which provides us with certain data points. As far as cleaning the data is concerned, a couple of data pre-processing steps are applied, which include checking for null values and data type issues, but our data is pretty much cleaned. However, the Car ID column is unnecessary and doesn't provide any useful information. Hence that is removed from the data set. A gist of our data set is attached below:

## car\_ID symboling CarName fueltype aspiration doornumber

## 1 1 3 alfa-romero giulia gas std two

## 2 2 3 alfa-romero stelvio gas std two

## 3 3 1 alfa-romero Quadrifoglio gas std two

## 4 4 2 audi 100 ls gas std four

## 5 5 2 audi 100ls gas std four

## 6 6 2 audi fox gas std two

## carbody drivewheel enginelocation wheelbase carlength carwidth carheight

## 1 convertible rwd front 88.6 168.8 64.1 48.8

## 2 convertible rwd front 88.6 168.8 64.1 48.8

## 3 hatchback rwd front 94.5 171.2 65.5 52.4

## 4 sedan fwd front 99.8 176.6 66.2 54.3

## 5 sedan 4wd front 99.4 176.6 66.4 54.3

## 6 sedan fwd front 99.8 177.3 66.3 53.1

## curbweight enginetype cylindernumber enginesize fuelsystem boreratio stroke

## 1 2548 dohc four 130 mpfi 3.47 2.68

## 2 2548 dohc four 130 mpfi 3.47 2.68

## 3 2823 ohcv six 152 mpfi 2.68 3.47

## 4 2337 ohc four 109 mpfi 3.19 3.40

## 5 2824 ohc five 136 mpfi 3.19 3.40

## 6 2507 ohc five 136 mpfi 3.19 3.40

## compressionratio horsepower peakrpm citympg highwaympg price

## 1 9.0 111 5000 21 27 13495

## 2 9.0 111 5000 21 27 16500

## 3 9.0 154 5000 19 26 16500

## 4 10.0 102 5500 24 30 13950

## 5 8.0 115 5500 18 22 17450

## 6 8.5 110 5500 19 25 15250

The data is made up of 205 records and 26 variables.

dim(data)

## [1] 205 26

To ensure that accurate results will be obtained, we check for any missing values in the columns of the dataset. The result is false, and this means that the data has no missing values.

anyNA(data)

## [1] FALSE

summary(data)

## car\_ID symboling CarName fueltype

## Min. : 1 Min. :-2.0000 Length:205 Length:205

## 1st Qu.: 52 1st Qu.: 0.0000 Class :character Class :character

## Median :103 Median : 1.0000 Mode :character Mode :character

## Mean :103 Mean : 0.8341

## 3rd Qu.:154 3rd Qu.: 2.0000

## Max. :205 Max. : 3.0000

## aspiration doornumber carbody drivewheel

## Length:205 Length:205 Length:205 Length:205

## Class :character Class :character Class :character Class :character

## Mode :character Mode :character Mode :character Mode :character

##

##

##

## enginelocation wheelbase carlength carwidth

## Length:205 Min. : 86.60 Min. :141.1 Min. :60.30

## Class :character 1st Qu.: 94.50 1st Qu.:166.3 1st Qu.:64.10

## Mode :character Median : 97.00 Median :173.2 Median :65.50

## Mean : 98.76 Mean :174.0 Mean :65.91

## 3rd Qu.:102.40 3rd Qu.:183.1 3rd Qu.:66.90

## Max. :120.90 Max. :208.1 Max. :72.30

## carheight curbweight enginetype cylindernumber

## Min. :47.80 Min. :1488 Length:205 Length:205

## 1st Qu.:52.00 1st Qu.:2145 Class :character Class :character

## Median :54.10 Median :2414 Mode :character Mode :character

## Mean :53.72 Mean :2556

## 3rd Qu.:55.50 3rd Qu.:2935

## Max. :59.80 Max. :4066

## enginesize fuelsystem boreratio stroke

## Min. : 61.0 Length:205 Min. :2.54 Min. :2.070

## 1st Qu.: 97.0 Class :character 1st Qu.:3.15 1st Qu.:3.110

## Median :120.0 Mode :character Median :3.31 Median :3.290

## Mean :126.9 Mean :3.33 Mean :3.255

## 3rd Qu.:141.0 3rd Qu.:3.58 3rd Qu.:3.410

## Max. :326.0 Max. :3.94 Max. :4.170

## compressionratio horsepower peakrpm citympg

## Min. : 7.00 Min. : 48.0 Min. :4150 Min. :13.00

## 1st Qu.: 8.60 1st Qu.: 70.0 1st Qu.:4800 1st Qu.:19.00

## Median : 9.00 Median : 95.0 Median :5200 Median :24.00

## Mean :10.14 Mean :104.1 Mean :5125 Mean :25.22

## 3rd Qu.: 9.40 3rd Qu.:116.0 3rd Qu.:5500 3rd Qu.:30.00

## Max. :23.00 Max. :288.0 Max. :6600 Max. :49.00

## highwaympg price

## Min. :16.00 Min. : 5118

## 1st Qu.:25.00 1st Qu.: 7788

## Median :30.00 Median :10295

## Mean :30.75 Mean :13277

## 3rd Qu.:34.00 3rd Qu.:16503

## Max. :54.00 Max. :45400

str(data)

## 'data.frame': 205 obs. of 26 variables:

## $ car\_ID : int 1 2 3 4 5 6 7 8 9 10 ...

## $ symboling : int 3 3 1 2 2 2 1 1 1 0 ...

## $ CarName : chr "alfa-romero giulia" "alfa-romero stelvio" "alfa-romero Quadrifoglio" "audi 100 ls" ...

## $ fueltype : chr "gas" "gas" "gas" "gas" ...

## $ aspiration : chr "std" "std" "std" "std" ...

## $ doornumber : chr "two" "two" "two" "four" ...

## $ carbody : chr "convertible" "convertible" "hatchback" "sedan" ...

## $ drivewheel : chr "rwd" "rwd" "rwd" "fwd" ...

## $ enginelocation : chr "front" "front" "front" "front" ...

## $ wheelbase : num 88.6 88.6 94.5 99.8 99.4 ...

## $ carlength : num 169 169 171 177 177 ...

## $ carwidth : num 64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4 67.9 ...

## $ carheight : num 48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52 ...

## $ curbweight : int 2548 2548 2823 2337 2824 2507 2844 2954 3086 3053 ...

## $ enginetype : chr "dohc" "dohc" "ohcv" "ohc" ...

## $ cylindernumber : chr "four" "four" "six" "four" ...

## $ enginesize : int 130 130 152 109 136 136 136 136 131 131 ...

## $ fuelsystem : chr "mpfi" "mpfi" "mpfi" "mpfi" ...

## $ boreratio : num 3.47 3.47 2.68 3.19 3.19 3.19 3.19 3.19 3.13 3.13 ...

## $ stroke : num 2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 3.4 ...

## $ compressionratio: num 9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...

## $ horsepower : int 111 111 154 102 115 110 110 110 140 160 ...

## $ peakrpm : int 5000 5000 5000 5500 5500 5500 5500 5500 5500 5500 ...

## $ citympg : int 21 21 19 24 18 19 19 19 17 16 ...

## $ highwaympg : int 27 27 26 30 22 25 25 25 20 22 ...

## $ price : num 13495 16500 16500 13950 17450 ...

Convert variables to categorical variables

data$fueltype<- as.factor(data$fueltype)

data$fuelsystem<- as.factor(data$fuelsystem)

data$enginelocation<- as.factor(data$enginelocation)

data$enginetype<- as.factor(data$enginetype)

data$carbody<- as.factor(data$carbody)

data$drivewheel<- as.factor(data$drivewheel)

## Methodology:

For the exploratory data analysis of exploring and insights generation of data points on the target variable price, box plots, correlation metrics, and scatter plots are used. In order to have a look at the distribution of car prices, a histogram is created. Finally, for forecasting the price of a car based on input features, a linear regression model is created and evaluated on the data set to check the performance of the forecasting model. The evaluation measure that is used in this case is RMSE, MSE and R squared. The higher the R square of the model, the better the performance. The lower the RMSE and MSE, the better is the model because the error is low. The first exploratory data analysis is the analysis of car price by horsepower.

ggplot(data, aes(x = horsepower, y = price)) +

geom\_point() +

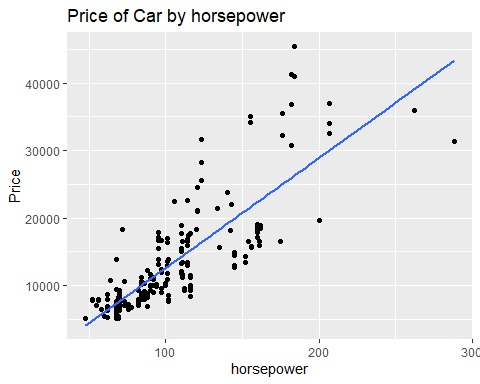
geom\_smooth(method = "lm", se = FALSE) +

xlab("horsepower") +

ylab("Price") +

ggtitle("Price of Car by horsepower")

## `geom\_smooth()` using formula 'y ~ x'



ggplot(data, aes(x = carbody, y = price, fill = carbody)) +

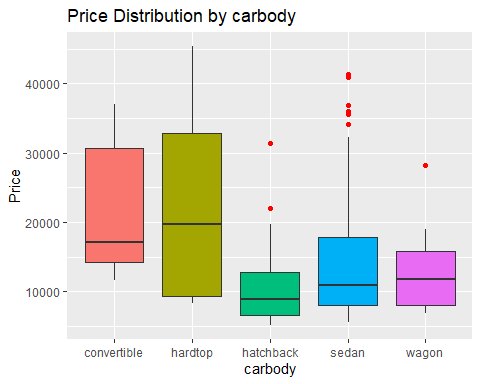
geom\_boxplot(outlier.color = "red") +

theme(legend.position = "none") +

xlab("carbody") +

ylab("Price") +

ggtitle("Price Distribution by carbody")

 Next, is the exploratory analysis of the price of cars by engine location.

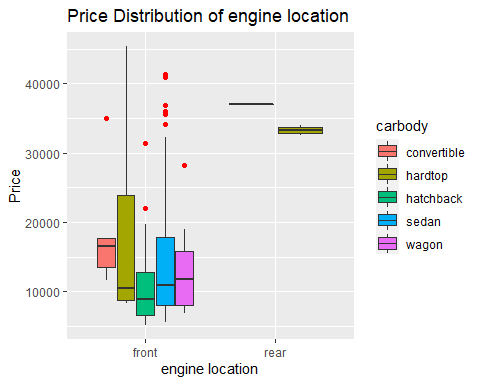
ggplot(data, aes(x = enginelocation, y = price, fill = carbody)) +

geom\_boxplot(outlier.color = "red") +

xlab("engine location") +

ylab("Price") +

ggtitle("Price Distribution of engine location")



The next step includes the creation of the logistic regression model. First, split data into training and testing sets.

set.seed(444)

data <- data %>%

select(-c(carbody,horsepower,enginelocation,enginesize))

data$price <- log(data$price)

colnames(data) <- make.names(colnames(data))

train\_split <- createDataPartition(y = data$price, p = 0.8, list = FALSE)

training <- data[train\_split,]

testing <- data[-train\_split,]

Create the logistic model

options(warn = -1)

lm <- trainControl(method = "repeatedcv", number = 10, repeats = 10)

lm\_model <- train(price~., data=training,

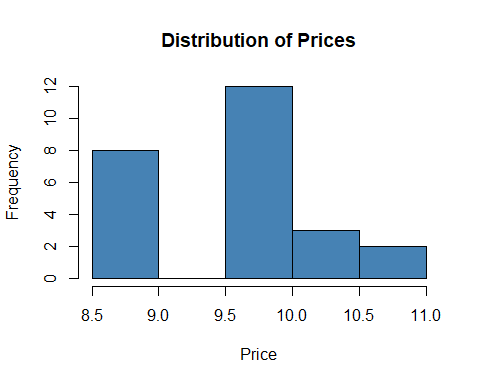
trControl = lm,

method = "lm",

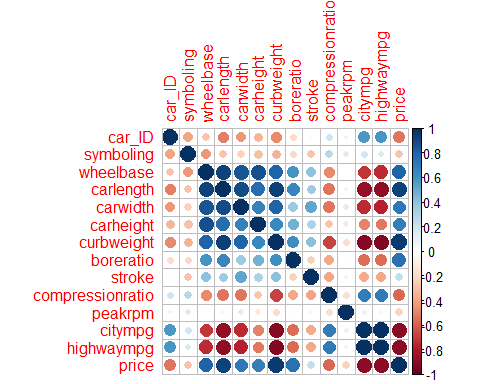
metric = "Rsquared")

## Results:

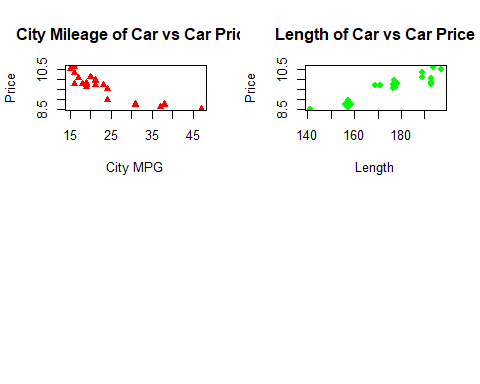
Below is the histogram [4] of our model attached. It can be seen that the distribution of our target variable is skewed and not normally distributed in this case. This tells us that most of the cars that are present in the data set have a price up to 20,000. There is a very less number of cars having prices above 20,000 in this case.



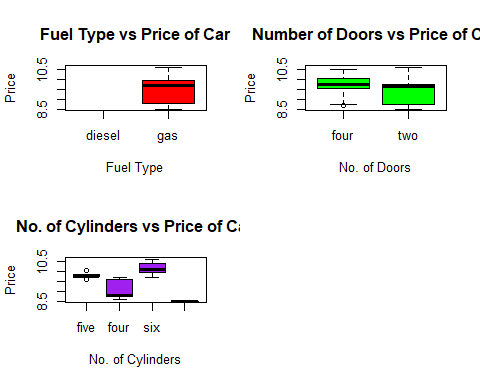
Below is the correlation matrix plot [5] attached. From the correlation matrix plot, it can be observed that, if we look at the target variable price, most of the columns have a strong positive correlation with the target variable. A strong positive correlation means with the increase in one variable, the other variable also increases, and with the decrease in one variable, the other variable decreases. A strong negative correlation indicates that with an increase in one value, the other value decreases and vice versa. The pairs having a strong positive correlation with the price are WheelBase, Car length, car width, curb weight, engine size, and horsepower. The pairs having a strong negative correlation with price are city mileage and highway mileage. Car height, stroke, compression ratio, and peak rpm have no correlation with the target variable.



In order to make sure that the conclusion drawn above is true and to support our conclusion, a set of scatter plots is created to see the trend. It can be seen from below attached figures below that, except for car mileage, all other variables have a linear relationship with the target variable price, and a linear positive trend line could be fitted to the graphs. This means with an increase in these variables, the price of cars also increases. For city mileage and price, the linear negative trend line could be fitted, which represents a negative relationship with the target variable price of the car.



In order to see the effect of categorical variables on car prices, a series of box plots have been created. It can be seen from below attached figures below that the average car prices of cars having fuel type “Diesel” is higher as compared to cars having fuel type “Gas." This means Diesel cars tend to be expensive as compared to Gas cars. The same goes for the number of doors. Four-door cars tend to have a higher average price as compared to the average price of two-door cars. The most expensive car having the highest average price according to the box plots attached below is a sedan, and the least expensive car having the lowest average price is a hatchback. Finally, four-cylinder cars are inexpensive and have the lowest average prices, while 12-cylinder cars have the highest average car prices.



Finally, a regression model is created, and the summary of our model is attached below. It can be seen that the R square of our model is 85% which means the model was able to explain 85% variability in our data set. The p-value of our overall model is less than the significance level alpha = 0.05. We can say that the model is significant in this case.

##

## Call:

## lm(formula = price ~ ., data = data\_num)

##

## Residuals:

## Min 1Q Median 3Q Max

## -0.119801 -0.047432 0.004114 0.026240 0.141715

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 6.567e+00 1.615e+00 4.066 0.00186 \*\*

## car\_ID -6.850e-03 5.738e-03 -1.194 0.25767

## symboling -8.415e-03 4.350e-02 -0.193 0.85015

## wheelbase -7.727e-03 2.441e-02 -0.317 0.75754

## carlength 1.160e-02 1.242e-02 0.935 0.37003

## carwidth -5.324e-02 3.312e-02 -1.607 0.13628

## carheight 3.919e-02 3.530e-02 1.110 0.29054

## curbweight 7.751e-04 2.056e-04 3.769 0.00310 \*\*

## boreratio 3.876e-01 2.295e-01 1.689 0.11935

## stroke -2.983e-02 1.664e-01 -0.179 0.86095

## compressionratio 2.491e-02 6.059e-02 0.411 0.68890

## peakrpm -4.111e-06 6.575e-05 -0.063 0.95127

## citympg -2.079e-02 2.142e-02 -0.971 0.35264

## highwaympg 1.507e-02 2.436e-02 0.619 0.54882

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.09292 on 11 degrees of freedom

## Multiple R-squared: 0.9897, Adjusted R-squared: 0.9776

## F-statistic: 81.42 on 13 and 11 DF, p-value: 5.358e-09

The accuracy of our forecasting model is calculated below. The Root means square error of our model is 0.189, while the mean absolute error is 0.144. This error could further be reduced if we perform parameter tuning and have more meaningful data.

## [1] "The root mean squared error is : 0.0616372754137951"

## [1] "The Mean Absolute Error is : 0.0474681988557009"

## Implications:

A future study that could be done in this area is to gather more information about cars, such as the number of airbags and automatic or manual transmission, and having more data about cars and types of cars would definitely improve the results of this analysis and will provide the results at the granular level.

## Conclusion:

In this project, we have analyzed the car's data set and provided a forecasting model. We have explored that engine size, the number of cylinders, horsepower, and length of the car affect the prices of a car positively. The greater the length of the car, the greater the number of cylinders, the greater the engine size, and the greater the price of a car. We further built a forecasting model using a regression model to predict the prices of cars and found that the model was able to explain 85% variability in the target variable and the RMSE and MAE of our model are 0.188 and 0.144, respectively, which indicates that the model can predict data more accurately.

## References:

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. Journal of the Royal Society Interface, 14(134), 20170213.

Pan, F., Zhu, P., & Zhang, Y. (2010). Metamodel-based lightweight design of B-pillar with TWB structure via support vector regression. Computers & structures, 88(1-2), 36-44.

Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. Journal of information technology, 30(1), 75-89.