Car Price Prediction

Group 26

18HS20010	Avani Haritima	**18HS20020**	Kush Verma
18HS20028	Saket Mahajan	**18HS20032**	Shashank Shrivastava
18HS20036	Siddhant Vishwakarma	**18HS20039**	Suman Kumari

Abstract

Simple Linear regression or regression models in general have two main objectives. The first is to establish a relationship between two variables. More specifically, we will be going through a statistically significant relationship between the two variables. The other objective is to forecast new observations. We use the information of a known relationship to forecast unobserved values.

The two different roles played by variables in a regression model are the dependent and the independent variables. The dependent variable is the one that is to be explained or forecasted. It is called dependent because it is dependent on the other variable in the model. The independent variable explains the dependent variable.

A linear regression line has an equation of the form $\boldsymbol{Y = \beta_0 + \beta_1 X}$, where $\boldsymbol{X}$ is the explanatory variable and $\boldsymbol{Y}$ is the dependent variable. The slope of the line is $\boldsymbol{\beta_1}$, and $\boldsymbol{\beta_0}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$).

The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values.

Introduction

Problem Statement

A Chinese automobile company Geely Auto aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts.

They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market. The company wants to know:

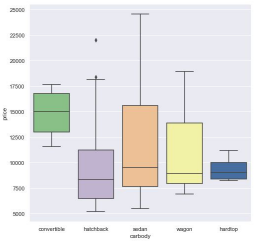
Which variables are significant in predicting the price of a car How well those variables describe the price of a car Based on various market surveys, the consulting firm has gathered a large data set of different types of cars across the American market.

Variables

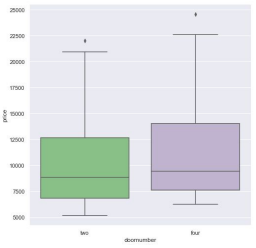
car_ID	The index number of the observation
doornumber	Number of Doors in the car
carbody	The build of the car body
wheelbase	The distance between centers of the front and rear wheels
carlength	The length of the car chassis
carwidth	The width of the car chassis
carheight	The height of the car
curbweight	The mass of a vehicle with all standard equipment and operating consumables
cylindernumber	Number of cylinders in the engine
enginesize	The total volume of fuel and air the engine pushes through cylinders.
fuelsystem	The type of fuel injection system used in the car.
boreratio	The ratio between cylinder bore diameter and piston stroke length
stroke	stroke length determined by how far the piston moves in a cylinder
compressionratio	Relative volumes of the combustion chamber and the chamber
horsepower	The power supplied by the engine
peakrpm	revolutions per minute/ rotations the engine's crankshaft makes in a minute.
citympg	miles per gallon of fuel in the city tends to be lower due to interrupted driving
highwaympg	miles per gallon of fuel on the highway.

Data Analysis

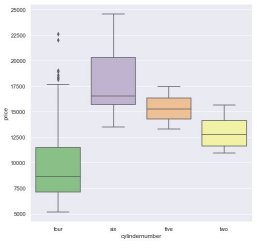
Car Body vs Price



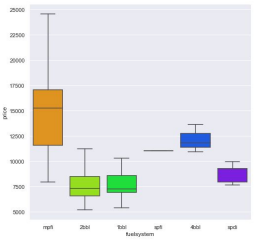
Door Number vs Price



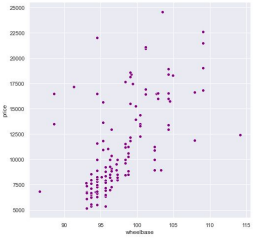
Cylinder Number vs Price



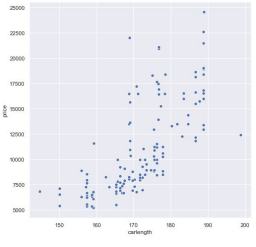
Fuel System vs Price



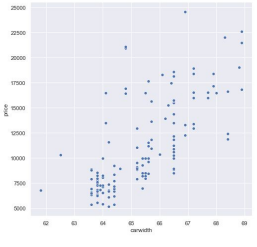
Wheel Base vs Price



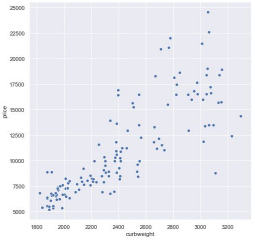
Car Length vs Price



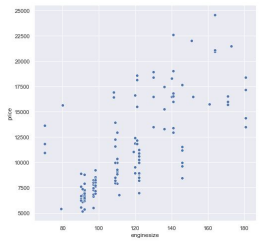
Car Width vs Price



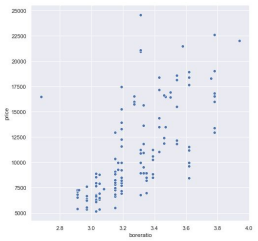
Curb Weight vs Price



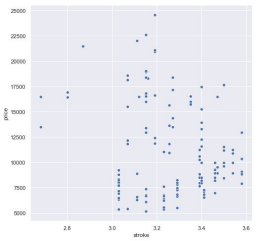
Engine Size vs Price



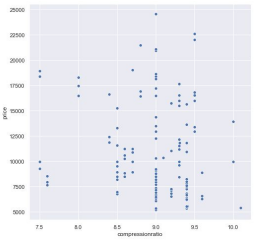
Bore Ratio vs Price



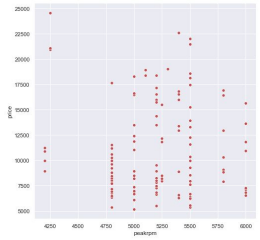
Stroke vs Price



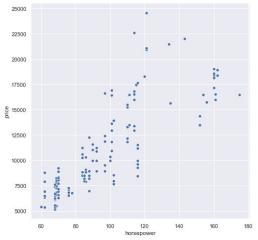
Compression Ratio vs Price



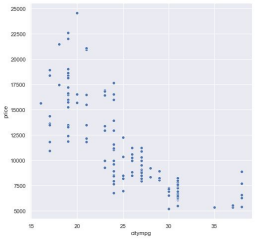
Horsepower vs Price



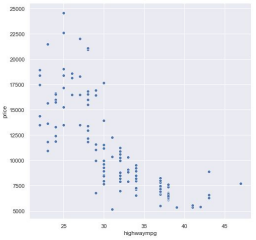
Peak RPM vs Price



City MPG vs Price



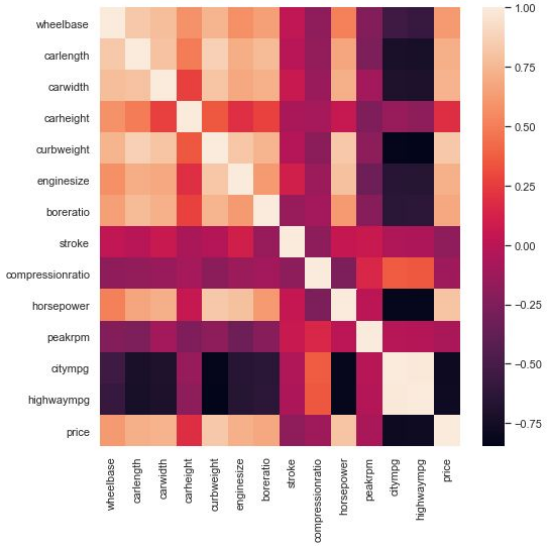
Highway MPG vs Price



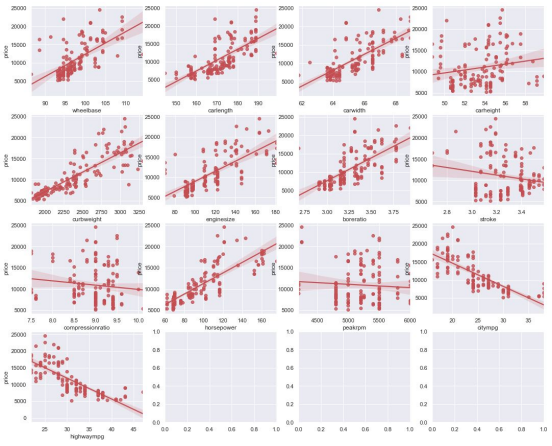
Some observations from these graphs:

- The price of the hardtop is very high compared to others
- Cars having cylinder number eight have a higher price range.
- The price range is high for the car having a mpfi fuel system.
- Car length is also scattered but less scattered than the wheelbase.
- points are scattered after curbweight of 2900, initially, it is increasing as curbweight increases as we can see in the joint plot color becomes lighter after curbweight of 2900 .
- Very weak correlation between stroke vs price.
- No relation between compression ratio and price & peakrpm and price.
- A negative correlation is seen between citympg and price.
- A negative correlation between highwaympg and price.

Heat Map of the Variables



Regression Plot of Variables with Price



Model Fitting

Firstly, we applied the Multiple Linear Regression Model on 17 independent variables and the dependent variable (Price).

Hypothesis Testing and p-value

Conventional hypothesis test

Null Hypothesis	**Alternative Hypothesis**
There is no relationship between Price and variable X:	There is a relationship between Price and variable X
β **equals** zero	β is **not equal** to zero

Testing hypothesis

Reject the null	**Fail to reject the null**
There is a relationship	There is no relationship
If the 95% confidence interval **does not include zero**	If the 95% confidence interval **includes zero**

The p-values for the independent 17 variables came out to be as follows:

Feature	P Values
Door Number	0.50573
Car Body	0.15206
Wheel Base	0.01981
Car Length	0.95139
Car Width	0.71430
Car Height	0.92296
Curb Weight	0.54815
Cylinder Number	0.35269
Engine Size	0.05508
Fuel System	0.52818
Bore Ratio	0.08810

Stroke	0.00000
Compression Ratio	0.02453
Horsepower	0.00388
Peak RPM	0.36990
City MPG	0.03653
Highway MPG	0.38506

Results

We've found that the **r-squared value** of the model is **0.72367** if we've considered all the variables. Then we've extended our analysis to find the p-Values of each variable having a significant value of less than 0.05.

According to the **null hypothesis**, if the P-values is less than 0.05(level of significance), then we can reject the null hypothesis and say that the variable is having some relation with the response variable: **'Price of the Car'.**

After seeing the P-values, we've dropped the variables for which the P-values are not meeting the significance mark, calculating the **r-squared value** later shows an increment having a value **0.781**.

The variables used in the final model are **Wheelbase, Stroke, Compression Ratio, Horsepower, and City MPG.** The rest of them are not significant for the determination of the **Price of the Car.**

Conclusion

After studying and analyzing the Linear Regression model for the above Problem statement, we can say that the linear Regression is not a suitable fit for the given dataset hence more advanced technology must be required to properly examine and predict the car price with the variables given.

Resources

Github Repo

Dataset