**CAR PRICE PREDICTION USING MULTIPLE LINEAR REGRESSION ANALYSIS IN EXCEL**

**INTRODUCTION**

The analysis aims at identifying significant variables required to predict the price of a car and to describe how well the variables describe the price of the car.

**UNDERSTANDING DATA STRUCTURE**

The dataset used for this analysis with a 250 sample size was gotten from Kaggle, containing car prices and attributes related to the predictive analysis.

The variables included in the analysis are as follows;

1. Car name
2. Wheel base
3. Car length
4. Car width
5. Curb weight
6. Engine size
7. Bore ratio
8. Stroke
9. Compression ratio
10. Horse power
11. Peak rpm
12. City mpg
13. Highway mpg
14. Price
15. Engine type
16. Cylinder number
17. Fuel type
18. Aspiration
19. Door number
20. Car body
21. Drive wheel
22. Engine location
23. Fuel system

By understanding the data structure and the variables involved, correlation analysis was further done to explore the relationships between these car attributes and the prices. This analysis will provide insights into the factors that may influence car prices and help understand the dynamics of the car market.

**DATA CLEANING AND EXPLORATION**

* Incorrect spellings; the car names with incorrect spellings were corrected.
* The Car brands were extracted from the car name column and converted to the proper text format.

**DESCRIPTIVE STATISTICS**

|  |  |
| --- | --- |
|  |  |
| *Summary statistics table* | |
|  |  |
| Mean | 13276.71 |
| Standard Error | 557.97 |
| Median | 10295.00 |
| Mode | 16500.00 |
| Standard Deviation | 7988.85 |
| Sample Variance | 63821761.58 |
| Kurtosis | 3.05 |
| Skewness | 1.78 |
| Range | 40282.00 |
| Minimum | 5118.00 |
| Maximum | 45400.00 |
| Sum | 2721725.67 |
| Count | 205.00 |

**CORRELATION AND MULTIPLE REGRESSION ANALYSIS**

**NUMERICAL VARIABLES**

The dataset contains thirteen numerical variables and correlation analysis was performed on them to determine the variables having a strong correlation with the dependent variables.

|  |  |  |
| --- | --- | --- |
| ***Independent variable*** | **Corr. coefficient (r)** | **Description** |
| *Wheelbase* | 0.58 | Moderate Positive correlation |
| *Car length* | 0.68 | Moderate Positive correlation |
| *Car width* | 0.76 | Strong Positive correlation |
| *Car height* | 0.12 | Weak Positive correlation |
| *Curb weight* | 0.84 | Strong Positive correlation |
| *Engine size* | 0.87 | Strong Positive correlation |
| *Bore ratio* | 0.55 | Moderate Positive correlation |
| *Stroke* | 0.08 | Weak Positive correlation |
| *Compression ratio* | 0.07 | Weak Positive correlation |
| *Horsepower* | 0.81 | Strong Positive correlation |
| *Peak rpm* | -0.09 | Weak Negative correlation |
| *City mpg* | -0.69 | Weak Negative correlation |
| *Highway mpg* | -0.70 | Weak Negative correlation |

**Significant Numerical Variables based on the correlation analysis**

1. Curb weight
2. Engine Size
3. Horsepower

**Regression analysis result on the significant numerical variable**s

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.90 |  |  |  |
| R Square | 0.81 |  |  |  |
| Adjusted R Square | 0.81 |  |  |  |
| Standard Error | 3472.704 |  |  |  |
| Observations | 205 |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | -13463.8 | 1333.051 | -10.1 | 0.00 |
| Curb weight | 4.262552 | 0.906525 | 4.702077 | 0.00 |
| Engine size | 84.87963 | 12.76137 | 6.651295 | 0.00 |
| Horse power | 48.74726 | 10.69684 | 4.557165 | 0.00 |

Interpretation; R square shows that 81% of variance in dependent variable is affected by the independent variables. The p-values indicates that the regression is statistically significant (p <0.05)

Calculating VIF = 5.46 which is high (> = 5.0), indicating that there is an indicative significance of multicollinearity.

***Calculating Variance Inflation Factor (VIF) of each variable; using the formula VIF = 1 / (1-r) where r is the coefficient of determination.***

* Engine size; 4.77
* Horse power; 3.02
* Curb weight; 3.77

To address multicollinearity, the engine size which has high value of VIF (approximately 5) was removed.

Running regression analysis with the rest of the variables; Horse power and curb weight against the dependent variable Price.

**Regression Analysis Result**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.88 |  |  |  |  |  |
| R Square | 0.77 |  |  |  |  |  |
| Adjusted R Square | 0.77 |  |  |  |  |  |
|  | *Coefficients* | *S.E* | *t Stat* | *P-value* |
| Intercept | -15989.9 | 1407.94 | -11.357 | 0.00 |  |  |
| Curb weight | 8.037492 | 0.78 | 10.31946 | 0.00 |  |  |
| horsepower | 83.81222 | 10.25 | 8.1725 | 0.00 |

The VIF value = 4.4 which is moderate, and the P-values associated with each variable shows that they are statistically significant.

R Square indicates that 77% of the price of the car is explained by the curb weight and horse power. The standard errors are shows that the estimate is precise and more reliable and that we can have more confidence in the estimate.

**CATEGORICAL VARIABLES**

The categorical variables which are aspiration, Car body, Engine Type, Cylinder number, Door number, and Fuel type, were converted into dummy variables, and regression analysis was performed.

**REGRESSION ANALYSIS RESULT**

This was carried out on each categorical variable to determine the relationship between each with the price of the car.

Car Name; this regression analysis was performed after removing variables with evidence of possible multicollinearity and presence of extreme values.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.62 |  |  |  |
| R Square | 0.39 |  |  |  |
| Adjusted R Square | 0.36 |  |  |  |
|  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 17893.38 | 723.20 | 24.74 | 0.00 |
| Honda | -9708.69 | 1913.42 | -5.07 | 0.00 |
| Mazda | -7240.50 | 1709.62 | -4.24 | 0.00 |
| Mitsubishi | -8653.62 | 1913.42 | -4.52 | 0.00 |
| Nissan | -7477.72 | 1670.17 | -4.48 | 0.00 |
| Plymouth | -9929.96 | 2520.12 | -3.94 | 0.00 |
| Porsche | 13507.12 | 2946.56 | 4.58 | 0.00 |
| Subaru | -9352.13 | 1980.58 | -4.72 | 0.00 |
| Toyota | -8007.57 | 1340.86 | -5.97 | 0.00 |
| Volkswagen | -7957.88 | 2145.37 | -3.71 | 0.00 |

Interpretation; 38% of the variation in the price of a car is dependent on the car model and the P-values show that the variables are statistically significant. However, the higher standard error associated with the coefficient indicates that the estimate not precise and therefore not reliable which reduces the confidence that we can have in the estimate.

**Drive Wheel**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.64 |  |  |  |
| R Square | 0.41 |  |  |  |
| Adjusted R Square | 0.40 |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 11087.46 | 2054.75 | 5.39 | 0.00 |
| rwd | 8823.346 | 2173.01 | 4.06 | 0.00 |
| fwd | -1848.15 | 2130.41 | -0.87 | 0.39 |

*Interpretation;* the adjusted R square indicates that 41% of variance in the price of car can be explained by the drive wheel. The variable "rwd" (Rear-Wheel Drive) has a low p-value of 0.00, indicating a statistically significant relationship with car prices. Including this variable in the model would be beneficial as it shows a strong association with the target variable. “fwd” on the other hand is not statistically significant and may not be needed to improve model predictive accuracy.

**Cylinder Number**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.79 |  |  |  |
| R Square | 0.63 |  |  |  |
| Adjusted R Square | 0.62 |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 36000 | 4894.79 | 7.35 | 0.00 |
| four | -25714.2 | 4910.16 | -5.24 | 0.00 |
| six | -12328.2 | 4995.73 | -2.46 | 0.01 |
| five | -14369.5 | 5112.44 | -2.81 | 0.01 |
| eight | 1400.1 | 5361.98 | 0.26 | 0.79 |
| three | -30849 | 6922.28 | -4.46 | 0.00 |
| two | -22980 | 5472.54 | -4.19 | 0.00 |

*Interpretation;* the adjusted R square indicates that 63% of variance in the price of car can be explained by the cylinder number. The variables "four," "six," "five," "three," and "two" have low p-values (all less than 0.05), suggesting statistically significant relationships with car prices. Including these variables in the model would be relevant as they provide information about the engine's cylinder configuration, which can influence car prices.

**Engine Type**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.47 |  |  |  |
| R Square | 0.22 |  |  |  |
| Adjusted R Square | 0.19 |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 31400.50 | 7155.87 | 4.39 | 0.0 |
| dohc | -13284.08 | 7448.07 | -1.78 | 0.1 |
| ohcv | -6302.12 | 7426.00 | -0.85 | 0.4 |
| ohcf | -17661.90 | 7390.56 | -2.39 | 0.0 |
| L | -16772.92 | 7448.07 | -2.25 | 0.0 |
| ohc | -19826.45 | 7180.01 | -2.76 | 0.0 |
| rotor | -18380.50 | 8000.51 | -2.30 | 0.0 |

*Interpretation;* the adjusted R square indicates that 22% of variance in the price of car can be explained by the drive wheel. The variables "ohcf," "L," "ohc," and "rotor" have low p-values (all less than 0.05), indicating statistically significant relationships with car prices. Including these variables in the model would be valuable as they represent different engine types, which can impact car prices.

**Car Body**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.37 |  |  |  |
| R Square | 0.14 |  |  |  |
| Adjusted R Square | 0.12 |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 21890.50 | 3057.46 | 7.16 | 0.00 |
| hatchback | -11513.85 | 3185.80 | -3.61 | 0.00 |
| hardtop | 318.00 | 4044.63 | 0.08 | 0.94 |
| sedan | -7546.23 | 3151.55 | -2.39 | 0.02 |
| wagon | -9518.54 | 3404.64 | -2.80 | 0.01 |

*Interpretation;* the adjusted R square indicates that 14% of variance in the price of car can be explained by the drive wheel. The p-values associated with the rwd indicates that it is statistically significant but that of fwd indicates that it is not statistically significant and may not provide meaningful insights or improve model predictive accuracy.

**Door Number**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.03 |  |  |  |
| R Square | 0.00 |  |  |  |
| Adjusted R Square | -0.00 |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 12989.92 | 843.74 | 15.39 | 0.00 |
| four | 511.23 | 1126.51 | 0.45 | 0.65 |

*Interpretation;* the adjusted R square of 0% shows that the price of car cannot be explained by the door number. The p-values associated with the door number (four) is very high which indicates that it is not statistically significant and including it in the model may not be necessary as it does not provide strong evidence of influencing car price.

**Aspiration**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.18 |  |  |  |
| R Square | 0.03 |  |  |  |
| Adjusted R Square | 0.03 |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 16298.17 | 1295.58 | 12.58 | 0.00 |
| std | -3686.9 | 1431.16 | -2.58 | 0.01 |

*Interpretation;* The p-value indicates that the analysis is statistically significant but the adjusted R square indicates that the variable contributes 3.0% to the variance in the price of the car and so including this to the predictive model may not contribute valuable information or enhance the model's performance.

**Engine Location**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.32 |  |  |  |
| R Square | 0.11 |  |  |  |
| Adjusted R Square | 0.10 |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 34528 | 4372.75 | 7.89 | 0.00 |
| front | -21566.9 | 4405.10 | -4.89 | 0.00 |

*Interpretation;* the R square indicates that 11% of variance in the price of car can be explained by the engine location. The p-values shows that the analysis is statistically significant. This variable would however not be included in the model due to the low coefficient of determination.

**Fuel Type**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.10 |  |  |  |
| R Square | 0.01 |  |  |  |
| Adjusted R Square | 0.00 |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 15838.15 | 1780.72 | 8.89 | 0.00 |
| Gas | -2838.35 | 1874.51 | -1.51 | 0.13 |

*Interpretation;* the adjusted R square indicates that 10% of variance in the price of car can be explained by the fuel type. The p-values indicates that it is not statistically significant (above threshold of 0.05) and including it in the model may not provide meaningful insights or improve model predictive accuracy.

**Fuel System**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.59 |  |  |  |
| R Square | 0.36 |  |  |  |
| Adjusted R Square | 0.33 |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 11048 | 6517.62 | 1.69 | 0.09 |
| mpfi | 6706.603 | 6552.19 | 1.02 | 0.31 |
| 2bbl | -3569.85 | 6566.81 | -0.54 | 0.59 |
| 1bbl | -3492.45 | 6807.43 | -0.51 | 0.61 |
| Spdi | -57.5556 | 6870.18 | -0.01 | 0.99 |
| 4bbl | 1097 | 7525.90 | 0.14 | 0.88 |
| Idi | 4790.15 | 6678.57 | 0.71 | 0.47 |
| mfi | 1916 | 9217.30 | 0.21 | 0.84 |

*Interpretation;* the adjusted R square indicates that 36% of variance in the price of car can be explained by the fuel system. However, none of the fuel system variables (mpfi, 2bbl, 1bbl, Spdi, 4bbl, Idi, mfi) show statistically significant relationships with car prices, as their respective p-values are above the typical threshold of 0.05. Including these variables in the model may not contribute significant predictive power and could introduce noise or unnecessary complexity.

**Categorical variables with higher significance after interpretation of the result from regression analysis and intepretation**

1. Cylinder Number
2. Engine type
3. Drive wheel

***Running regression analysis on the significant categorical variables***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.88 |  |  |  |
| R Square | 0.78 |  |  |  |
| Adjusted R Square | 0.77 |  |  |  |
|  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 30969.87 | 1717.78 | 18.02 | 0.00 |
| dohc | 6391.51 | 1870.00 | 3.42 | 0.00 |
| ohcf | 9778.98 | 1887.97 | 5.18 | 0.00 |
| L | 9132.36 | 2231.64 | 4.09 | 0.00 |
| ohc | 8459.95 | 1747.28 | 4.84 | 0.00 |
| rotor | -24146.7 | 2458.68 | -9.82 | 0.00 |
| four | -30810 | 2465.63 | -12.49 | 0.00 |
| six | -18008 | 2075.21 | -8.68 | 0.00 |
| five | -20052.8 | 2685.29 | -7.47 | 0.00 |
| three | -34951.2 | 4858.99 | -7.19 | 0.00 |
| rwd | 6196.88 | 729.89 | 8.49 | 0.00 |

Interpretation; the adjusted R square indicates that 78% of variance in the price of car can be explained by these categorical variables. The p-values associated with the all the variables shows evidence of statistical significance (below 0.05)

**Multiple Regression analysis on the categorical and numerical variables with dependent variable (price)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.97 |  |  |  |
| R Square | 0.94 |  |  |  |
| Adjusted R Square | 0.94 |  |  |  |
|  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 0.845824 | 0.076117 | 11.11 | 0.00 |
| Six | -0.69517 | 0.028323 | -24.54 | 0.00 |
| Five | -0.89285 | 0.036619 | -24.38 | 0.00 |
| Three | -1.20345 | 0.113319 | -10.62 | 0.00 |
| Rwd | -0.02136 | 0.022784 | -0.94 | 0.35 |
| curbweight | -0.000075 | 0.000028 | -2.59 | 0.01 |
| horsepower | -0.00102 | 0.00034 | -3.00 | 0.00 |
| Dohc | 0.390427 | 0.041763 | 9.35 | 0.00 |
| Ohcf | 0.389569 | 0.04257 | 9.15 | 0.00 |
| L | 0.517554 | 0.050269 | 10.29 | 0.00 |
| Ohc | 0.413285 | 0.036699 | 11.26 | 0.00 |
| Rotor | -0.53299 | 0.065225 | -8.17 | 0.00 |

The adjusted R square indicates that 94% of the variance in car price is explained by the model.

***Calculating VIF of each of the variables to check for multicollinearity***

Six; 6.66

five; 5.42

Three; 1.92

rwd; 2.37

Curb weight; 4.5

Horsepower; 3.6

Dohc; 2.72

Ohcf; 3.43

L; 4.21

Ohc; 8.74

Rotor; 2.13

*Removing six, five, ohc due to the high VIF values and running another regression analysis with the rest of the variables.*

***Regression analysis result***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.895853 |  |  |  |
| R Square | 0.802552 |  |  |  |
| Adjusted R Square | 0.794493 |  |  |  |
|  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | -16184.3 | 1671.25 | -9.68391 | 0.00 |
| horsepower | 63.61599 | 11.15 | 5.704264 | 0.00 |
| dohc | -3045.5 | 1137.93 | -2.67634 | 0.01 |
| ohcf | 1439.22 | 995.54 | 1.445667 | 0.15 |
| L | -5291.13 | 1335.94 | -3.9606 | 0.00 |
| rotor | -1331.28 | 1921.94 | -0.69268 | 0.49 |
| three | 10627.21 | 3944.95 | 2.693878 | 0.01 |
| rwd | 2591.68 | 776.85 | 3.336121 | 0.00 |
| curbweight | 8.700031 | 0.9165 | 9.49211 | 0.00 |

The adjusted R square indicates that Removing ohcf and rotor because the P-values associated with them shows that they are not statistically significant.

**Final Regression Analysis Result**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |
| Multiple R | 0.894362 |  |  |  |
| R Square | 0.799883 |  |  |  |
| Adjusted R Square | 0.793819 |  |  |  |
| Standard Error | 3627.508 |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | -16142.9 | 1627.53 | -9.919 | 0.00 |
| horsepower | 66.41401 | 10.9870 | 6.044 | 0.00 |
| dohc | -3097.06 | 1132.57 | -2.73 | 0.01 |
| L | -5156.62 | 1334.28 | -3.86 | 0.00 |
| Three | 10413.67 | 3947.98 | 2.64 | 0.01 |
| Rwd | 2345.715 | 742.51 | 3.16 | 0.00 |
| Curbweight | 8.635081 | 0.89 | 9.62 | 0.00 |

***Interpretation;*** The multiple R-value 0.89 shows that there is a strong relationship between the variables and the car prices. The R square shows that 79% of variance in the dependent variable is explained by the independent variables. The P-values associated with the variables shows that they are statistically significant.

**Linear Regression Model Equation**

Dependent variable (Price of car) = -16142.9 + 66.41401 \* horsepower + -3097.06 \* dohc + -5156.62 \* L + 10413.67 \* three + 2345.715 \* rwd + 8.635081 \* curb weight

**Testing the Model Using Prediction**

Expected car price; 13495

Engine Type; dohc

Horse Power; 111

Cylinder number; four

Curb weight; 2548

Drive wheel; rwd

***Using the regression model equation;***

-16142.9 + 66.41401 \* horsepower + -3097.06 \* dohc + -5156.62 \* L + 10413.67 \* three + 2345.715 \* rwd + 8.635081 \* curb weight

Y (price of the car) = -16142.9 + 66.41401 \* 111 + -3097.06 \* 1 + -5156.62 \*0 + 10413.67 \* 0 + 2345.715 \* 1 + 8.635081 \* 2548

Predicted price: 12,479.9

Actual price: 10,661.05 (79% of the expected price)

Deviation: The predicted price is 1,818.85 higher than the actual price.

Assessment: Based on this evaluation, the prediction is slightly overestimating the actual price.

**CONCLUSION**

The analysis reveals that the variables Horsepower, dohc engine type, cylinder number (three), Curb weight, and drive wheel (rwd) have huge impact in the variance in price of a car.

The analysis also revealed that the numerical variables of wheelbase, car length, car width, car height, engine size, bore ratio, peak rpm, city mpg and highway mpg had varying degrees of correlation with the price of the car, although they were not included in the final regression model due to multicollinearity issues or lack of statistical significance.

Regarding the categorical variables, the analysis showed that the drive wheel, cylinder number, engine type, and car body had statistically significant relationships with the car prices. Specifically, rear-wheel drive (rwd) in drive wheel, and four, six, five, three, and two cylinder numbers were found to be significant as well as the engine type ohcf, L, ohc, and rotor were found to have a significant relationship with car prices. Variables associated with aspiration, door number, fuel type, and fuel system did not show significant relationships with car prices and may not contribute valuable information to the predictive model.

**RECOMMENDATION**

The regression model developed explains approximately 79% of the variance in car prices, 21% can is dependent on other unknown variables. However, further improvements or alternative modelling approaches can be considered.

Overall, this analysis provides insights into the significant variables and their relationships with car prices, helping to understand the dynamics of the car market and providing a foundation for predicting car prices using multiple linear regression analysis.