8/1/2022

Best Linear Regression Model to Predict Car Prices

Truong Bao Tran

CONTENTS

LINEAR REGRESSION MODEL (SIMPLE, MULTIPLE)

1.1. Introduction:

1.1.1. Dataset

Dataset: Car sales

Author: Gagan Bhatia

Data source: https://www.kaggle.com/

Link of dataset: https://www.kaggle.com/datasets/gagandeep16/car-sales

This is a Car Sales Dataset that includes information about different types of cars. This dataset is being obtained from Analytixlabs for prediction purposes.

The data set has 11 independent variables and 1 dependent variable, consisting of 157 observations. The dependent variable is the sale price of cars.

1.1.2. Problems

I need to find the best model to predict the final price of cars. For used car dealers, they can use this model to target certain types of cars, cars with more potential and higher selling prices. On the other hand, car owners can use the same model to know what to do to increase the value of their property (remodel, upgrade, etc.), depending on which variable increases the property value. theirs the most.

Furthermore, car seekers/potential buyers can select the features of the car they want to buy to estimate the required budget.

1.2. Analysis with Python

1.2.1. Data Description and Preprocessing

When examining Null data in the dataset, it can be seen that the attribute "__year_resale_value" has 36 Null values. In other attributes, Null values exist quite a bit, only ranging from 1 to 3 values.

```
In [10]: df.isnull().sum()
Out[10]:
Manufacturer            0
Model                   0
Sales_in_thousands      0
__year_resale_value    36
Price_in_thousands      2
Engine_size             1
Horsepower              1
Wheelbase               1
Width                   1
Length                  1
Curb_weight             2
Fuel_capacity           1
Fuel_efficiency         3
Latest_Launch           0
Power_perf_factor       2
dtype: int64
```

```
In [12]: df.isnull().sum()
Out[12]:
Manufacturer           0
Model                  0
Sales_in_thousands     0
__year_resale_value    0
Price_in_thousands     0
Engine_size            0
Horsepower             0
Wheelbase              0
Width                  0
Length                 0
Curb_weight            0
Fuel_capacity          0
Fuel_efficiency        0
Latest_Launch          0
Power_perf_factor      0
dtype: int64
```

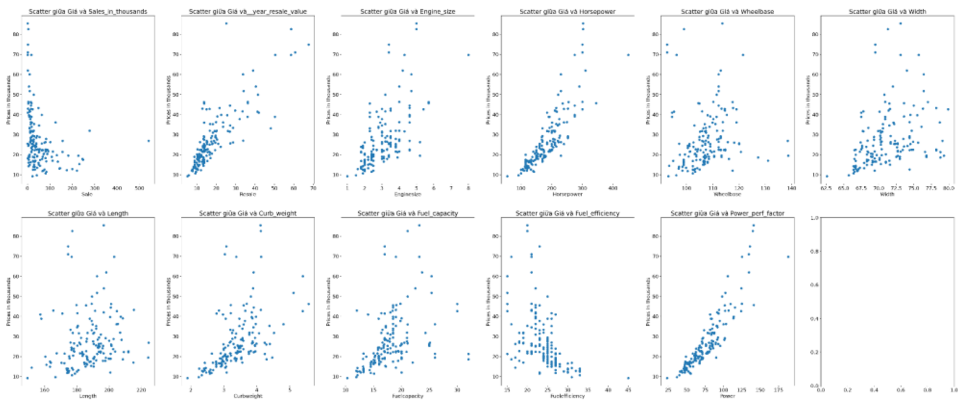
I will use the interpolate() function available in Python to fill in.

```
In [13]: print(df.describe())
       Sales_in_thousands  ...  Power_perf_factor
count          157.000000  ...         157.000000
mean            52.998076  ...          77.290632
std             68.029422  ...          25.082600
min              0.110000  ...          23.276272
25%             14.114000  ...          60.727447
50%             29.450000  ...          72.290355
75%             67.956000  ...          90.211700
max            540.561000  ...         188.144323
```

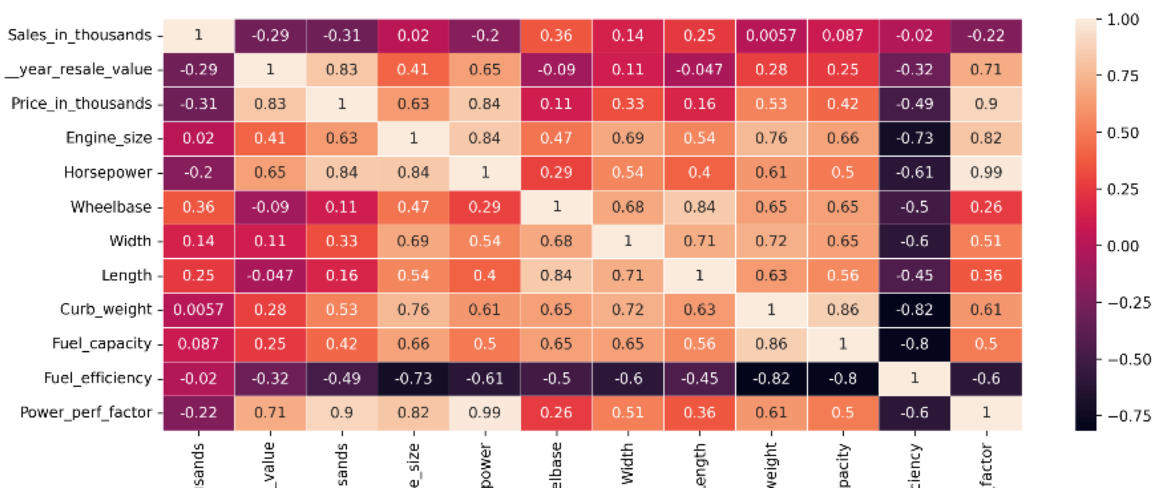
1.3.2. Simple Linear Regression

- **Correlation between variables through Scatter chart**

Comment on the correlation between variables through the Scatter chart: Scatter charts use dots to represent the values (intersection points) of two different variables. The main purpose of the Scatter chart in this data set is to observe and show the correlation between 2 variables, the price of the car (price) and 11 other attributes corresponding to 11 Scatter charts. Where the dependent variable (price) runs fixed on the vertical axis and the independent variable runs fixed on the horizontal axis. The dots in the scatter plot not only represent the value of a data point, but also the trend when we look at the entire data set as a whole.



- **Correlation between variables through Heatmap**



Looking at the Scatter chart and the heatmap chart, we have commented on the correlation between variables with Price as follows:

- Sale: -0.31 negative correlation and moderate correlation
- Enginesize: 0.63 positive correlation and strong correlation
- Horsepower: 0.84 positive correlation and strong correlation
- Wheelbase: 0.11 positive correlation and weak correlation
- Width 0.33 positive correlation and moderate correlation
- Length 0.16 positive correlation, weak correlation
- Curbweight: 0.53 positive correlation, strong correlation
- Fuelcapacity: 0.42 positive correlation, moderate correlation

- Fuelefficency: -0.49 negative correlation, moderate correlation
- Power: 0.9 positive correlation, strong correlation

Of all the variables just surveyed, the Power variable (0.9) has the strongest correlation with the Price variable. Therefore, we will proceed to build a univariate regression model with the independent variable being Power and the dependent variable being Price. And then make predictions through this model.

```
Hệ số R_square: 0.8028057065086073
Hệ số chặn: [-11.95889516]
Hệ số góc: [[0.5098694]]
```

A simple linear regression predicts the price of a car (dependant variable) from the power factor of the car (independent variable) having an R^2 of 0.8028. From this R^2 value, we know that:

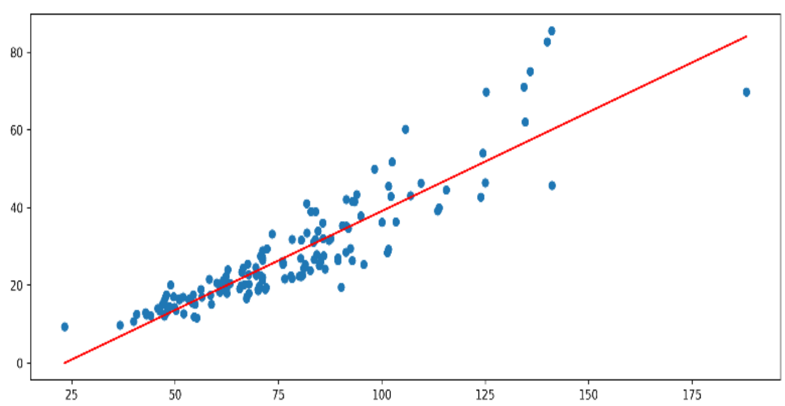
- 80.28% of the variance in car prices is predicted by the vehicle's power factor
- 19.72% of the variance in car prices is not explained by the model

The power factor of the vehicle has a great influence on the price of the car

The univariate linear regression model has the following form:

Prices = -11.9589 + 0.5099 Power

Meaning of the model: This means that for every 1 unit increase in Power Factor, the Price increases by 0.5099 units.



After building a univariate linear regression model of the form. We visualize the model using a linear regression graph as follows. Where the dependent variable is Price_of_thousands is on the vertical axis and the independent variable is Power_perf_factor is on the horizontal axis. The blue dots (intersection points) located near the red regression line show that the model has actual results that are close to the predicted results.

Dự báo

```
[[ 90.01498389]
 [ 64.52151413]
 [141.00192341]]
```

We have:

With Power Factor = 200, the price of the car is 90,01498 thousand dollars

With Power Factor = 100, the car price is 64,01498 thousand dollars

With Power Factor = 300, the price of the car is 141,01498 thousand dollars

1.3.3. Multiple Linear Regression

I will build a multivariable linear regression model with 11 independent variables and 1 dependent variable. Then remove each variable through p-value, if p-value > 0.05, then remove the variable from the model.

Before proceeding to build the model, I will divide the dataset into 2 parts training data and testing data with the ratio of 90% and 10% respectively to avoid overfitting when testing the model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Price_in_thousands   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.999
Method:                 Least Squares   F-statistic:                     9475.
Date:                Thu, 11 Aug 2022   Prob (F-statistic):          3.76e-201
Time:                        13:42:23   Log-Likelihood:                -123.19
No. Observations:                 157   AIC:                             270.4
Df Residuals:                     145   BIC:                             307.1
Df Model:                          11
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                  -0.1901      1.459     -0.130      0.897      -3.075       2.694
Sales_in_thousands     -0.0001      0.001     -0.169      0.866      -0.002       0.001
__year_resale_value    -0.0008      0.008     -0.101      0.919      -0.016       0.014
Engine_size            -0.7401      0.106     -6.966      0.000      -0.950      -0.530
Horsepower             -0.9051      0.009    -97.894      0.000      -0.923      -0.887
Wheelbase              -0.0028      0.013     -0.219      0.827      -0.028       0.023
Width                  -0.0163      0.023     -0.719      0.473      -0.061       0.028
Length                  0.0010      0.007      0.136      0.892      -0.013       0.015
Curb_weight             0.1900      0.194      0.982      0.328      -0.193       0.572
```

```
Fuel_capacity           0.0035      0.026      0.138      0.890      -0.047       0.054
Fuel_efficiency         0.0246      0.021      1.184      0.238      -0.016       0.066
Power_perf_factor       2.5697      0.022    116.121      0.000       2.526       2.613
==============================================================================
Omnibus:                      334.595   Durbin-Watson:                   2.049
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           136031.387
Skew:                          11.812   Prob(JB):                         0.00
Kurtosis:                     145.255   Cond. No.                     1.04e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.04e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model 1:

Because p-value of Resale > 0.05 => Not statistically significant => Remove Resale value

```
                             OLS Regression Results
==============================================================================
Dep. Variable:     Price_in_thousands   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.998
Method:                 Least Squares   F-statistic:                     9081.
Date:                Thu, 11 Aug 2022   Prob (F-statistic):          1.05e-179
Time:                        13:49:16   Log-Likelihood:                -117.71
No. Observations:                 141   AIC:                             257.4
Df Residuals:                     130   BIC:                             289.8
Df Model:                          10
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                  0.0123      1.614      0.008      0.994      -3.180       3.205
Sales_in_thousands -3.733e-06      0.001     -0.004      0.997      -0.002       0.002
Engine_size           -0.7520      0.113     -6.656      0.000      -0.976      -0.528
Horsepower            -0.9030      0.009   -103.643      0.000      -0.920      -0.886
Wheelbase             -0.0005      0.015     -0.037      0.970      -0.029       0.028
Width                 -0.0246      0.027     -0.911      0.364      -0.078       0.029
Length                 0.0007      0.008      0.095      0.925      -0.014       0.016
Curb_weight            0.2761      0.228      1.213      0.227      -0.174       0.726
```

```
Curb_weight            0.2761      0.228      1.213      0.227      -0.174       0.726
Fuel_capacity         -0.0095      0.033     -0.286      0.775      -0.075       0.056
Fuel_efficiency        0.0269      0.022      1.215      0.227      -0.017       0.071
Power_perf_factor      2.5657      0.019    136.288      0.000       2.528       2.603
==============================================================================
Omnibus:                      297.370   Durbin-Watson:                   2.106
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            95529.977
Skew:                          11.072   Prob(JB):                         0.00
Kurtosis:                     128.579   Cond. No.                     1.04e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.04e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model 2:

Since the p-value of Fuel capacity > 0.05 => Not statistically significant => Drop the value of Fuel capacity

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Price_in_thousands   R-squared:                       0.998
Model:                            OLS   Adj. R-squared:                  0.998
Method:                 Least Squares   F-statistic:                     9461.
Date:                Thu, 11 Aug 2022   Prob (F-statistic):          1.12e-179
Time:                        13:55:31   Log-Likelihood:                -118.06
No. Observations:                 141   AIC:                             256.1
Df Residuals:                     131   BIC:                             285.6
Df Model:                           9
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                 -0.1149      1.608     -0.071      0.943      -3.296       3.066
Sales_in_thousands    -0.0002      0.001     -0.200      0.842      -0.002       0.002
Engine_size           -0.7383      0.123     -6.007      0.000      -0.981      -0.495
Horsepower            -0.9042      0.009   -105.417      0.000      -0.921      -0.887
Wheelbase             -0.0023      0.013     -0.168      0.867      -0.029       0.024
Width                 -0.0181      0.025     -0.718      0.474      -0.068       0.032
Length                 0.0009      0.008      0.118      0.906      -0.014       0.016
Curb_weight            0.2133      0.199      1.070      0.286      -0.181       0.607
Fuel_efficiency        0.0243      0.021      1.161      0.248      -0.017       0.066
Power_perf_factor      2.5674      0.019    137.642      0.000       2.531       2.604
==============================================================================
Omnibus:                      298.523   Durbin-Watson:                   2.077
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            97497.809
Skew:                          11.156   Prob(JB):                         0.00
Kurtosis:                     129.876   Cond. No.                     1.03e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.03e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model 3

Because p-value of Sale > 0.05 => Not statistically significant => Remove Sale value

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Price_in_thousands   R-squared:                       0.998
Model:                            OLS   Adj. R-squared:                  0.998
Method:                 Least Squares   F-statistic:                     9372.
Date:                Thu, 11 Aug 2022   Prob (F-statistic):          7.57e-178
Time:                        13:56:57   Log-Likelihood:                -117.84
No. Observations:                 141   AIC:                             253.7
Df Residuals:                     132   BIC:                             280.2
Df Model:                           8
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                -0.1226      1.572     -0.078      0.938      -3.231       2.986
Engine_size          -0.7594      0.112     -6.766      0.000      -0.981      -0.537
Horsepower           -0.9016      0.010    -91.814      0.000      -0.921      -0.882
Wheelbase            -0.0042      0.014     -0.307      0.759      -0.031       0.023
Width                -0.0184      0.024     -0.760      0.448      -0.066       0.030
Length                0.0013      0.008      0.164      0.870      -0.014       0.017
Curb_weight           0.2440      0.189      1.292      0.199      -0.129       0.617
Fuel_efficiency       0.0263      0.021      1.246      0.215      -0.015       0.068
Power_perf_factor     2.5619      0.022    118.629      0.000       2.519       2.605
```

```
Power_perf_factor     2.5619      0.022    118.629      0.000       2.519       2.605
==============================================================================
Omnibus:                      297.794   Durbin-Watson:                   1.967
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            96250.385
Skew:                          11.103   Prob(JB):                         0.00
Kurtosis:                     129.055   Cond. No.                     9.90e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.9e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model 4

Because p-value of Length > 0.05 => Not statistically significant => Remove Length value

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Price_in_thousands   R-squared:                       0.998
Model:                            OLS   Adj. R-squared:                  0.998
Method:                 Least Squares   F-statistic:                 1.112e+04
Date:                Thu, 11 Aug 2022   Prob (F-statistic):          9.39e-181
Time:                        13:58:44   Log-Likelihood:                -118.01
No. Observations:                 141   AIC:                             252.0
Df Residuals:                     133   BIC:                             275.6
Df Model:                           7
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                -0.0892      1.558     -0.057      0.954      -3.172       2.993
Engine_size          -0.7475      0.118     -6.317      0.000      -0.981      -0.513
Horsepower           -0.9032      0.008   -106.484      0.000      -0.920      -0.886
Wheelbase            -0.0029      0.011     -0.280      0.780      -0.024       0.018
Width                -0.0168      0.024     -0.701      0.484      -0.064       0.031
Curb_weight           0.2210      0.191      1.154      0.250      -0.158       0.600
Fuel_efficiency       0.0258      0.020      1.265      0.208      -0.015       0.066
Power_perf_factor     2.5659      0.019    138.040      0.000       2.529       2.603
```

```
==============================================================================
Omnibus:                      298.349   Durbin-Watson:                   2.014
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            97198.984
Skew:                          11.143   Prob(JB):                         0.00
Kurtosis:                     129.680   Cond. No.                     7.83e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.83e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model 5

Because p-value of Wheelbase > 0.05 => Not statistically significant => Remove Wheelbase value

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Price_in_thousands   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.998
Method:                 Least Squares   F-statistic:                 1.518e+04
Date:                Thu, 11 Aug 2022   Prob (F-statistic):          3.61e-187
Time:                        14:00:04   Log-Likelihood:                -117.96
No. Observations:                 141   AIC:                             249.9
Df Residuals:                     134   BIC:                             270.6
Df Model:                           6
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                -0.3644      1.529     -0.238      0.812      -3.388       2.659
Engine_size          -0.7368      0.113     -6.542      0.000      -0.960      -0.514
Horsepower           -0.9045      0.008   -113.284      0.000      -0.920      -0.889
Width                -0.0199      0.022     -0.895      0.372      -0.064       0.024
Curb_weight           0.2335      0.167      1.394      0.166      -0.098       0.565
Fuel_efficiency       0.0321      0.023      1.396      0.165      -0.013       0.078
Power_perf_factor     2.5685      0.017    149.769      0.000       2.535       2.602
```

```
==============================================================================
Omnibus:                      298.202   Durbin-Watson:                   2.034
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            96920.090
Skew:                          11.133   Prob(JB):                         0.00
Kurtosis:                     129.496   Cond. No.                     7.20e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.2e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model 6

Because p-value of Curbweighte > 0.05 => Not statistically significant => Remove Curbweight value

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Price_in_thousands   R-squared:                       0.998
Model:                            OLS   Adj. R-squared:                  0.998
Method:                 Least Squares   F-statistic:                 1.686e+04
Date:                Thu, 11 Aug 2022   Prob (F-statistic):          7.69e-187
Time:                        14:01:42   Log-Likelihood:                -118.46
No. Observations:                 141   AIC:                             248.9
Df Residuals:                     135   BIC:                             266.6
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                -1.2274      0.899     -1.365      0.175      -3.006       0.551
Engine_size          -0.7533      0.117     -6.460      0.000      -0.984      -0.523
Horsepower           -0.9063      0.008   -119.486      0.000      -0.921      -0.891
Curb_weight           0.1401      0.162      0.862      0.390      -0.181       0.461
Fuel_efficiency       0.0247      0.020      1.226      0.222      -0.015       0.065
Power_perf_factor     2.5726      0.016    157.287      0.000       2.540       2.605
==============================================================================
Omnibus:                      299.849   Durbin-Watson:                   1.986
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            99801.447
```

```
Skew:                          11.254   Prob(JB):                         0.00
Kurtosis:                     131.378   Cond. No.                     3.94e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.94e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model 7

Because p-value of Width > 0.05 => Not statistically significant => Remove Width value

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Price_in_thousands   R-squared:                       0.998
Model:                            OLS   Adj. R-squared:                  0.998
Method:                 Least Squares   F-statistic:                 2.182e+04
Date:                Thu, 11 Aug 2022   Prob (F-statistic):          7.67e-190
Time:                        14:02:53   Log-Likelihood:                -118.86
No. Observations:                 141   AIC:                             247.7
Df Residuals:                     136   BIC:                             262.5
Df Model:                           4
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                -0.5121      0.551     -0.930      0.354      -1.601       0.577
Engine_size          -0.7365      0.104     -7.080      0.000      -0.942      -0.531
Horsepower           -0.9063      0.008   -113.826      0.000      -0.922      -0.891
Fuel_efficiency       0.0119      0.016      0.742      0.460      -0.020       0.044
Power_perf_factor     2.5729      0.017    150.933      0.000       2.539       2.607
==============================================================================
Omnibus:                      301.151   Durbin-Watson:                   2.026
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           102118.745
```

Model 8

Because p-value of Fuel_efficiency > 0.05 => Not statistically significant => Remove Fuel_efficiency value

```
                           OLS Regression Results
==============================================================================
Dep. Variable:     Price_in_thousands   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.999
Method:                 Least Squares   F-statistic:                 3.119e+04
Date:                Thu, 11 Aug 2022   Prob (F-statistic):          5.91e-194
Time:                        14:04:23   Log-Likelihood:                -119.24
No. Observations:                 141   AIC:                             246.5
Df Residuals:                     137   BIC:                             258.3
Df Model:                           3
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                -0.1161      0.177     -0.655      0.514      -0.467       0.235
Engine_size          -0.7623      0.085     -9.018      0.000      -0.929      -0.595
Horsepower           -0.9069      0.007   -121.841      0.000      -0.922      -0.892
Power_perf_factor     2.5739      0.016    161.133      0.000       2.542       2.606
==============================================================================
Omnibus:                      302.381   Durbin-Watson:                   2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           104331.621
Skew:                          11.442   Prob(JB):                         0.00
```

Looking at the model, it can be seen that these 3 variables all have p-values less than 0.05 (significant level of 5%)

It follows that these three variables are statistically significant for this model

Price = -0.7623 * Enginesize - 0.9069 * Horsepower + 2.539 * Power - 0.1161

Testing multivariable regression model

There is no perfect multicollinearity between the independent variables

According to Gujarati and Porter (2009), there are some signs of multicollinearity in the model when:

(1) VIF >= 10

(2) The correlation coefficient r of any variable in the model is greater than 0.8

As we can see, there is a very large multicollinearity between the Horsepowwer variable and the Power variable

```
            feature        VIF
0             const  14.100151
1       Engine_size   3.455221
2        Horsepower  80.604595
3  Power_perf_factor  73.144607
```

Build two more models between the variable

- Enginesive and Horsepower with Price

- Enginesive and Power with Price

Model: Enginesive and Power with Price

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Price_in_thousands   R-squared:                       0.838
Model:                            OLS   Adj. R-squared:                  0.836
Method:                 Least Squares   F-statistic:                     357.3
Date:                Thu, 11 Aug 2022   Prob (F-statistic):           2.68e-55
Time:                        14:07:11   Log-Likelihood:                -446.66
No. Observations:                 141   AIC:                             899.3
Df Residuals:                     138   BIC:                             908.2
Df Model:                           2
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const               -10.2740      1.608     -6.391      0.000     -13.453      -7.095
Engine_size          -4.1547      0.826     -5.031      0.000      -5.788      -2.522
Power_perf_factor     0.6531      0.034     19.020      0.000       0.585       0.721
==============================================================================
Omnibus:                       41.563   Durbin-Watson:                   2.008
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              115.931
Skew:                           1.131   Prob(JB):                     6.70e-26
```

Check VIF

```
            feature        VIF
0             const  11.093502
1       Engine_size   3.019280
2  Power_perf_factor   3.019280
```

Model: Enginesive and Horsepower with Price

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Price_in_thousands   R-squared:                       0.703
Model:                            OLS   Adj. R-squared:                  0.698
Method:                 Least Squares   F-statistic:                     163.1
Date:                Thu, 11 Aug 2022   Prob (F-statistic):           4.45e-37
Time:                        14:09:39   Log-Likelihood:                -474.06
No. Observations:                 141   AIC:                             954.1
Df Residuals:                     138   BIC:                             963.0
Df Model:                           2
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const          -8.7573      2.069     -4.232      0.000     -12.849      -4.665
Horsepower      0.2233      0.020     10.905      0.000       0.183       0.264
Engine_size    -1.8927      1.077     -1.757      0.081      -4.022       0.237
==============================================================================
Omnibus:                       68.561   Durbin-Watson:                   1.402
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              300.626
Skew:                           1.740   Prob(JB):                     5.25e-66
```

Check VIF

```
       feature        VIF
0        const  11.973957
1   Horsepower   3.327215
2  Engine_size   3.327215
```

<u>Conclusion</u>

Both models have VIF < 10, so there is no multicollinearity between these variables

Corrected R square is 83.6% VIF is equal to 3.01928 both less than 10.

A multiple linear regression predicts the vehicle price (dependent variable) from the vehicle's power factor (independent variable) and cylinder capacity (Enginesize) with an R^2 of 0.8368. From this R value, we know that:

- 83.6% of variance in vehicle price is predicted by vehicle's power factor and cylinder capacity

- 16.4% of variance in vehicle prices is not explained by the model

The power factor and cylinder capacity of the vehicle have a great influence on the price of the vehicle

Hence choose Enginesive and Power model with Price

Conclusion: We have the following multivariable regression model:

Price = -4.4157 * Enginesize + 0.6531* Power - 10.27

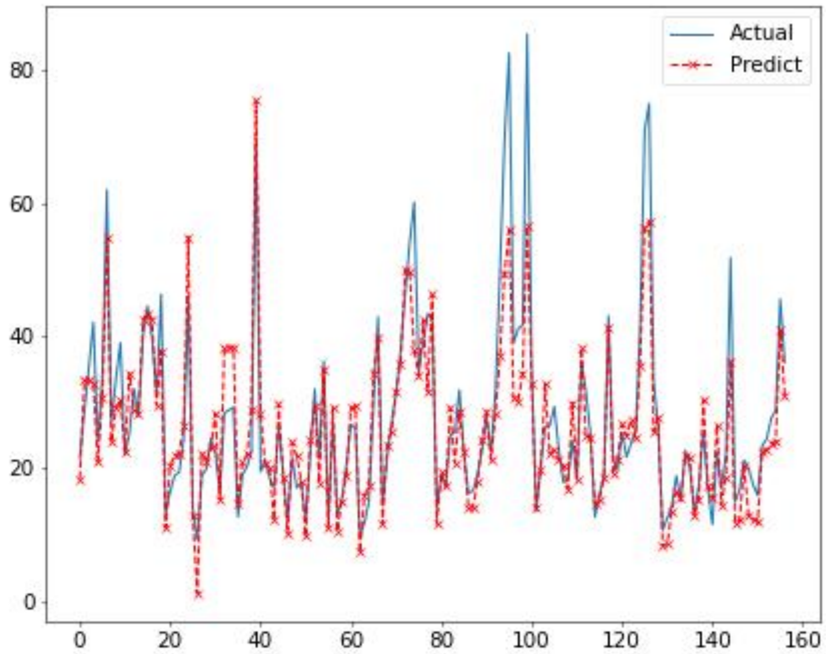
Model Meaning: This means that for every 1 unit increase in Cylinder Capacity, the Price (price) decreases by 4,4157 units. Meanwhile, for every 1 unit increase in the Power Factor, the Price (Price) increases by 0.6531 units.

Testing the model with the set of Testing data

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Price_in_thousands   R-squared:                       0.862
Model:                            OLS   Adj. R-squared:                  0.841
Method:                 Least Squares   F-statistic:                     40.64
Date:                Thu, 11 Aug 2022   Prob (F-statistic):           2.55e-06
Time:                        14:13:14   Log-Likelihood:                -54.018
No. Observations:                  16   AIC:                             114.0
Df Residuals:                      13   BIC:                             116.4
Df Model:                           2
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                -7.9927      5.272     -1.516      0.153     -19.382       3.397
Engine_size          -4.8565      2.247     -2.161      0.050      -9.711      -0.002
Power_perf_factor     0.6587      0.099      6.673      0.000       0.445       0.872
==============================================================================
Omnibus:                        1.285   Durbin-Watson:                   1.012
Prob(Omnibus):                  0.526   Jarque-Bera (JB):                0.645
Skew:                           0.488   Prob(JB):                        0.725
```

The adjusted R square of 84.1% is not too big of a difference from the experimental set. So this model is good for this dataset

Visualize with graphs



Forecast

```
[90.34147408 61.73901257]
```

We have:

With Power Factor = 200 and Enginesize = 4.2, then Price = 90,3414 thousand dollars

With Power Factor = 150 and Enginesize = 4.8, then Price = 61,73901 thousand dollars

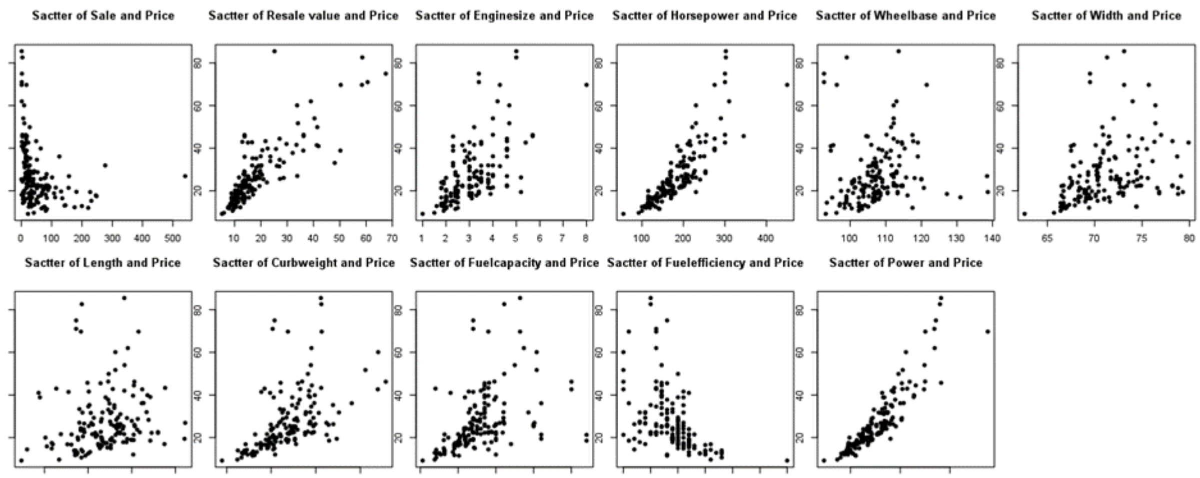
1.3. Analysis with Python

1.3.1. Simple Linear Regression

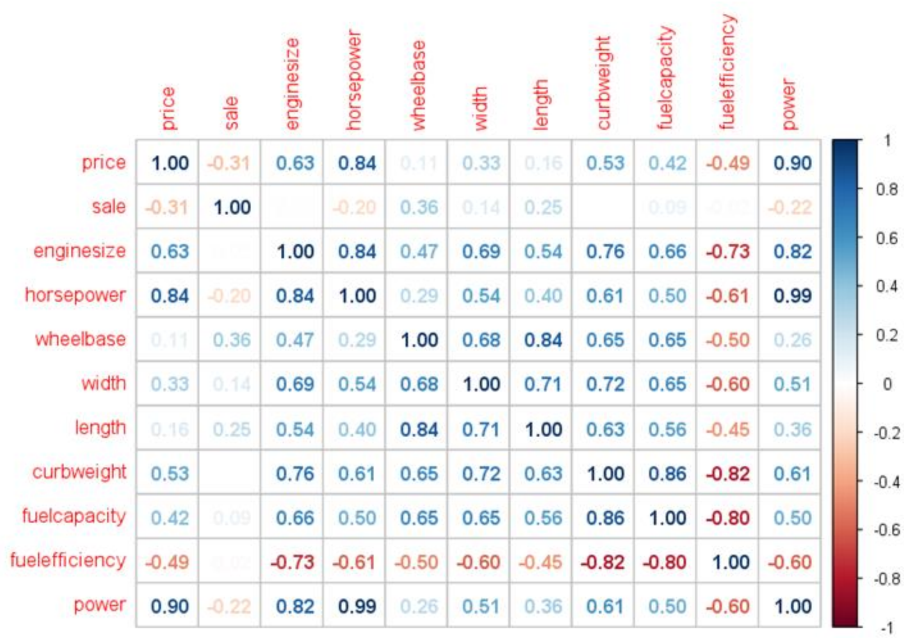
After cleaning the data with Python, I proceed to export to a CSV file and then re-import the processed file to facilitate analysis in R.

We will also perform the same modeling steps as in Python. So in this section I will do a quick analysis and only illustrate the final model.

Correlation between variables through Scatter chart



Correlation between variables through Heatmap



Of all the variables just surveyed, the Power variable (0.9) has the strongest correlation with the Price variable. Therefore, we will proceed to build a univariate regression model with the independent variable being Power and the dependent variable being Price. And then make predictions through this model.

After determining the correlation between variables: we can build a model to predict car price (Price) based on the variable Power by univariate regression model as follows:

```
Call:
lm(formula = price ~ power, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-14.5773  -4.5539   0.0283   2.6507  25.5158

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.9589     1.6488  -7.253 1.82e-11 ***
power         0.5099     0.0203  25.120  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.359 on 155 degrees of freedom
Multiple R-squared:  0.8028,	Adjusted R-squared:  0.8015
F-statistic:   631 on 1 and 155 DF,  p-value: < 2.2e-16
```

A simple linear regression predicts the price of a car (dependant variable) from the power factor of the car (independent variable) having an R^2 of 0.8028. From this R^2 value, we know that:

- 80.28% of the variance in car prices is predicted by the vehicle's power factor
- 19.72% of the variance in car prices is not explained by the model

The power factor of the vehicle has a great influence on the price of the car

The univariate linear regression model has the following form:

Prices = -11.9589 + 0.5099 Power

Meaning of the model: This means that for every 1 unit increase in Power Factor, the Price increases by 0.5099 units.

Forecast

1	2	3
90.01498	64.52151	141.00192

We have:

With Power Factor = 200, the car price is 90,01498 thousand dollars

With Power Factor = 100, the car price is 64,01498 thousand dollars

With Power Factor = 300, the price of the car is 141,01498 thousand dollars

1.3.2. Multiple Linear Regression

We will also perform the same modeling steps as in Python. So in this section I will do a quick analysis and only illustrate the final model.

Before proceeding to build the model, I will divide the dataset into 2 parts training data and testing data with the ratio of 90% and 10% respectively to avoid overfitting when testing the model.

Model 1:

```
lm(formula = price ~ sale + resale + enginesize + horsepower +
    wheelbase + width + length + curbweight + fuelcapacity +
    fuelefficiency + power, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2407 -0.1001 -0.0357  0.0257  6.5169

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.1900753  1.4594615  -0.130    0.897
sale           -0.0001325  0.0007859  -0.169    0.866
resale         -0.0007752  0.0076390  -0.101    0.919
enginesize     -0.7400776  0.1062471  -6.966 1.06e-10 ***
horsepower     -0.9051248  0.0092459 -97.894  < 2e-16 ***
wheelbase      -0.0028325  0.0129353  -0.219    0.827
width          -0.0162775  0.0226259  -0.719    0.473
length          0.0009517  0.0069765   0.136    0.892
curbweight      0.1899974  0.1935285   0.982    0.328
fuelcapacity    0.0035297  0.0255759   0.138    0.890
fuelefficiency  0.0245525  0.0207317   1.184    0.238
power           2.5697159  0.0221297 116.121  < 2e-16 ***
```

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

Residual standard error: 0.5518 on 145 degrees of freedom
Multiple R-squared:  0.9986,	Adjusted R-squared:  0.9985
F-statistic:  9475 on 11 and 145 DF,  p-value: < 2.2e-16
```

Model 2:

```
Call:
lm(formula = price ~ sale + enginesize + horsepower + wheelbase +
    width + length + curbweight + fuelcapacity + fuelefficiency +
    power, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2417 -0.1017 -0.0372  0.0250  6.5174

Coefficients:
                 Estimate Std. Error  t value Pr(>|t|)
(Intercept)    -0.2217612  1.4208305   -0.156    0.876
sale           -0.0001371  0.0007820   -0.175    0.861
enginesize     -0.7386766  0.1049886   -7.036 7.12e-11 ***
horsepower     -0.9046574  0.0079902 -113.220  < 2e-16 ***
wheelbase      -0.0026979  0.0128234   -0.210    0.834
width          -0.0161447  0.0225114   -0.717    0.474
length          0.0009622  0.0069520    0.138    0.890
curbweight      0.1937453  0.1893269    1.023    0.308
fuelcapacity    0.0030710  0.0250879    0.122    0.903
fuelefficiency  0.0246675  0.0206304    1.196    0.234
power           2.5683332  0.0173787  147.787  < 2e-16 ***
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.55 on 146 degrees of freedom
Multiple R-squared:  0.9986,    Adjusted R-squared:  0.9985
F-statistic: 1.049e+04 on 10 and 146 DF,  p-value: < 2.2e-16
```

Model 3:

```
Call:
lm(formula = price ~ sale + enginesize + horsepower + wheelbase +
    width + length + curbweight + power, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2284 -0.0988 -0.0404  0.0128  6.5852

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  0.6106293  1.2471202    0.490    0.625
sale        -0.0001983  0.0007790   -0.255    0.799
enginesize  -0.7563129  0.1036553   -7.296 1.67e-11 ***
horsepower  -0.9062204  0.0078399 -115.591  < 2e-16 ***
wheelbase   -0.0037852  0.0123737   -0.306    0.760
width       -0.0156815  0.0224552   -0.698    0.486
length       0.0029083  0.0067017    0.434    0.665
curbweight   0.0863785  0.1429762    0.604    0.547
power        2.5714565  0.0171072  150.314  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.5491 on 148 degrees of freedom
Multiple R-squared:  0.9986,    Adjusted R-squared:  0.9985
F-statistic: 1.316e+04 on 8 and 148 DF,  p-value: < 2.2e-16
```

Model 4

```
Call:
lm(formula = price ~ enginesize + horsepower + wheelbase + width +
    length + curbweight + power, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2221 -0.0949 -0.0383  0.0082  6.5880

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  0.663404   1.225906    0.541    0.589
enginesize  -0.763242   0.099704   -7.655 2.25e-12 ***
horsepower  -0.906207   0.007815 -115.956  < 2e-16 ***
wheelbase   -0.004754   0.011737   -0.405    0.686
width       -0.015656   0.022384   -0.699    0.485
length       0.002957   0.006678    0.443    0.658
curbweight   0.094716   0.138738    0.683    0.496
power        2.571719   0.017022  151.079  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.5473 on 149 degrees of freedom
Multiple R-squared:  0.9986,    Adjusted R-squared:  0.9985
F-statistic: 1.513e+04 on 7 and 149 DF,  p-value: < 2.2e-16
```

Model 5

```
lm(formula = price ~ enginesize + horsepower + width + length +
    curbweight + power, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.2376 -0.0991 -0.0361  0.0086  6.5953 

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)    
(Intercept)  0.615684   1.216826    0.506    0.614    
enginesize  -0.761891   0.099371   -7.667 2.05e-12 ***
horsepower  -0.906159   0.007792 -116.288  < 2e-16 ***
width       -0.017100   0.022037   -0.776    0.439    
length       0.001196   0.005053    0.237    0.813    
curbweight   0.078068   0.132139    0.591    0.556    
power        2.571886   0.016970  151.556  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5458 on 150 degrees of freedom
Multiple R-squared:  0.9986,	Adjusted R-squared:  0.9985 
F-statistic: 1.776e+04 on 6 and 150 DF,  p-value: < 2.2e-16
```

Model 6

```
Call:
lm(formula = price ~ enginesize + horsepower + width + curbweight +
    power, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2391 -0.1003 -0.0357  0.0119  6.5978

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  0.659707   1.198761    0.550    0.583
enginesize  -0.762756   0.098992   -7.705 1.61e-12 ***
horsepower  -0.905597   0.007398 -122.411  < 2e-16 ***
width       -0.015062   0.020221   -0.745    0.458
curbweight   0.087946   0.124981    0.704    0.483
power        2.570586   0.016006  160.599  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5441 on 151 degrees of freedom
Multiple R-squared:  0.9986,    Adjusted R-squared:  0.9985
F-statistic: 2.144e+04 on 5 and 151 DF,  p-value: < 2.2e-16
```

Model 7

```
Call:
lm(formula = price ~ enginesize + horsepower + width + power,
    data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2502 -0.0810 -0.0426  0.0025  6.6194

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  0.418161   1.146654    0.365    0.716
enginesize  -0.732582   0.089074   -8.224 8.13e-14 ***
horsepower  -0.907335   0.006962 -130.331  < 2e-16 ***
width       -0.008315   0.017774   -0.468    0.641
power        2.574343   0.015065  170.888  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5432 on 152 degrees of freedom
Multiple R-squared:  0.9986,    Adjusted R-squared:  0.9986
F-statistic: 2.689e+04 on 4 and 152 DF,  p-value: < 2.2e-16
```

Model 8

```
Call:
lm(formula = price ~ enginesize + horsepower + power, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2101 -0.0753 -0.0392 -0.0034  6.6289

Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept) -0.11283    0.16237   -0.695    0.488
enginesize  -0.75305    0.07739   -9.730   <2e-16 ***
horsepower  -0.90784    0.00686 -132.335   <2e-16 ***
power        2.57558    0.01479  174.131   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5418 on 153 degrees of freedom
Multiple R-squared:  0.9986,    Adjusted R-squared:  0.9986
F-statistic: 3.604e+04 on 3 and 153 DF,  p-value: < 2.2e-16
```

=> **Price = -0.75305 Enginesize - 0.90784 Horsepower + 2.57558 Power - 0.11283**

Check VIF

```
enginesize horsepower      power
  3.455221  80.604595  73.144607
```

Build two more models between the variable

- Enginesive and Horsepower with Price
- Enginesive and Power with Price

Model: Enginesive and horsepower with Price

```
Call:
lm(formula = price ~ enginesize + horsepower, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-15.092  -4.212  -0.432   2.251  34.260

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.09205    2.10485  -5.270 4.54e-07 ***
enginesize   -3.34694    1.06834  -3.133  0.00207 **
horsepower    0.26181    0.01961  13.353  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.622 on 154 degrees of freedom
Multiple R-squared:  0.7185,    Adjusted R-squared:  0.7149
F-statistic: 196.6 on 2 and 154 DF,  p-value: < 2.2e-16
```

Check VIF

```
enginesize horsepower
  3.327215   3.327215
```

Model: Enginesive and Power with Price

```
Call:
lm(formula = price ~ enginesize + power, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-13.0356  -2.8685  -0.3174   1.8345  24.5001

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.03509    1.54252  -6.506 1.03e-09 ***
enginesize   -4.39093    0.77485  -5.667 6.96e-08 ***
power         0.65903    0.03219  20.476  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.803 on 154 degrees of freedom
Multiple R-squared:  0.8368,    Adjusted R-squared:  0.8347
F-statistic: 394.9 on 2 and 154 DF,  p-value: < 2.2e-16
```

Check VIF

```
enginesize      power
   3.01928    3.01928
```

<u>Conclusion</u>

Both models have VIF < 10, so there is no multicollinearity between these variables

Corrected R square is 83.6% VIF is equal to 3.01928 both less than 10.

A multiple linear regression predicts the vehicle price (dependent variable) from the vehicle's power factor (independent variable) and cylinder capacity (Enginesize) with an R^2 of 0.8368. From this R value, we know that:

- 83.6% of variance in vehicle price is predicted by vehicle's power factor and cylinder capacity

- 16.4% of variance in vehicle prices is not explained by the model

The power factor and cylinder capacity of the vehicle have a great influence on the price of the vehicle

Hence choose Enginesive and Power model with Price

Conclusion: We have the following multivariable regression model:

Price = -4.4157 * Enginesize + 0.6531* Power - 10.27

Model Meaning: This means that for every 1 unit increase in Cylinder Capacity, the Price (price) decreases by 4,4157 units. Meanwhile, for every 1 unit increase in the Power Factor, the Price (Price) increases by 0.6531 units.

Forecast

```
        1         2         3
 99.81612  69.06012 161.32812
```

We have:

With Power Factor = 200 and Enginesize = 5, then Price = 99,81612 thousand dollars

With Power Factor = 150 and Enginesize = 5, then Price = 69,06012 thousand dollars

With Power Factor = 300 and Enginesize = 5, then Price = 161,32812 thousand dollars